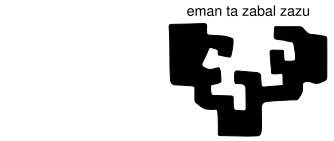
Universidad del País Vasco
UPV/EHU



Tesis Doctoral

Contribuciones de Inteligencia Artificial Aplicada en Sistemas Industriales

---

Autora:
Izaskun Mendia

Supervisores:
Prof. Dr. Manuel Maria Vélez
Dr. Sergio Gil-López

A 4 de Abril de 2022

Resumen

La dinámica de la sociedad moderna empuja al sector industrial hacia una creciente necesidad de sistemas cada vez más complejos y autónomos, destinada a liberar a los seres humanos de tareas mecánicas, repetitivas y poco gratificantes. Las tecnologías habilitadoras que harán posible esta revolución están disponibles. Y es un hecho que, la Inteligencia Artificial abre un universo de posibilidades para transformar en valor la ingente cantidad de datos existentes. En este campo de investigación, además de las técnicas ya conocidas y ampliamente utilizadas para entrenar modelos, se puede encontrar en la literatura un sinnúmero de variaciones algorítmicas. Sin embargo, esta apuesta por la Inteligencia Artificial no es todavía tangible dentro del sector industrial. Quizás porque estas potentes técnicas han de aterrizarse a la realidad de problemas concretos en industrias reales. Y sin género de dudas, la Inteligencia Artificial Aplicada es clave para ayudar a transformar el ecosistema industrial actual. Urge centrar los esfuerzos en promover estas tecnologías a través de la creación de nuevas herramientas que ejemplifiquen la aplicación de la tecnología del dato y de la Inteligencia Artificial.

Este trabajo de Tesis doctoral está centrado, no en la definición de nuevas aportaciones analíticas, sino en la investigación estratégica de las técnicas de Inteligencia Artificial aplicadas al ámbito industrial. Sencillas y entendibles técnicas, capaces de abstraer a la audiencia de las complejas fórmulas matemáticas y de las oscuras cajas negras, aplicadas a la realidad de 3 casos de investigación científica industrial no-supervisados.

Inicialmente, se propone la creación de una herramienta para la correcta y equilibrada asignación de consumidores a Fases en la red de Baja Tensión de la Red Eléctrica. En la resolución del problema se aplican algoritmos de optimización ávaros (greedy) y algoritmos meta-heurísticos (agnósticos al problema y de propósito general) y se describen métricas provenientes de diferentes dominios para medir la calidad de la solución. El concepto común en dichas métricas es el estudio de la complementariedad entre las curvas de carga (patrones de consumo) de cada consumidor telegestionado de la Línea eléctrica.

Posteriormente, se propone un procedimiento para el Control y Supervisión de procesos industriales, donde ciertas variables críticas del proceso son difícilmente medibles. En la resolución del problema, se aplican algoritmos predictivos para inferir la relación entre las variables conocidas y medibles del proceso, y su relación con las variables críticas. El sistema de inferencia propuesto, a través de la correcta secuenciación de técnicas (técnicas de selección de variables relevantes, técnicas de limpieza de datos probabilísticas, técnicas de eliminación de ruidos y redundancias y técnicas de adecuación dinámica a los cambios de comportamiento del proceso), consigue obtener el valor de las variables críticas en tiempo real.

Y finalmente, se propone una metodología para la modelización energética de una planta industrial en términos de tasa de producción y de consumos eléctricos individuales (a nivel de máquina) y consumos eléctricos agregados (a nivel de planta). En la resolución del problema se aplican sencillos algoritmos descriptivos y regresivos que permiten reconocer aquellos patrones de comportamiento que justifican el funcionamiento energético de la planta y que permiten detectar las ineficiencias energéticas que no se corresponden con los patrones identificados y descubrir la causa raíz de tales ineficiencias. Se trata de la resolución de un problema de caracterización energética no-supervisado.

Asimismo, con objeto de difundir los resultados obtenidos en los casos de investigación industrial se han realizado diversas tareas de diseminación científica (2 artículos de revista y 3 congresos internacionales) y diseminación tecnológica (3 patentes y 1 registro de software).Como reconocimiento a la innovación y calidad de los resultados y aportaciones obtenidas, estas investigaciones aplicadas también han recibido 2 premios de reconocimiento industrial ("Best use of Data Science for Industry 4.0" y "Research and development of artificial intelligence applied to industrial plants") y el reconocimiento de Innobasque como "Caso industrial de referencia". Todos ellos fruto de las diversas innovaciones en el ámbito industrial relacionadas con los resultados de las investigaciones.

Esta tesis está dedicada a todos y cada uno de los miembros de mi querida familia. Porque entre todos (y una pandemia) la hemos hecho posible.

Reconocimientos

A mi director de Tesis en UPV/EHU Manuel Vélez por su completa disponibilidad y empátia. A los socios, empresas industriales y especialmente a TECNALIA, por su comprensión y apoyo para aprovechar todo ese conocimiento y permitirme canalizarlo en formato de investigación científica. A Sandra, por animarme y apoyarme en los momentos en los que me he encontrado más desmotivada y por su inestimable ayuda con las correcciones.

Especial mención a Sergio Gil-López, a quien quiero expresar mi más sincero agradecimiento por su excelente orientación, apoyo y estímulo a lo largo de esta Tesis. Durante nuestras reuniones y discusiones ha sabido liderar, supervisar y animar en la búsqueda de soluciones. Estoy muy agradecida por está experiencia, por poderla realizar de la mano de un "gigante". Sin su apoyo incondicional nada de lo que ha ocurrido estos años hubiera sido posible.

A mi madre y a mi padre, por confiar siempre en mi, y demostrarme que pase lo que pase, ellos están siempre cerca para sostenerme y animarme en cada uno de mis nuevos retos. A mi hermana Edurne, que siempre ha defendido el lema "si quieres, puedes".

A Ibon, mi compañero de vida, porque juntos aprendemos y crecemos para ser cada día nuestra mejor versión. Porque en esta vida se trata de disfrutar en el camino, aunque a veces haya que ir saltando piedras. A mis hijos, Iñigo y Xabier, por todo lo que aprendo a vuestro lado, por vuestro afecto siempre genuino, especialmente cuando más lo he necesitado. Por estar siempre a mi lado.

A todos los que de una u otra manera, confiaron en mí.

Reconocimientos Industriales

En el marco de esta tesis de investigación industrial es muy destacable el agradecimiento a las instituciones que han apoyado dicha investigación:

Al Departamento de Desarrollo Económico, Sostenibilidad y Medio Ambiente (SPRI) del Gobierno Vasco por impulsar la industria vasca a través de diferentes programas y proyectos. Gran parte del conocimiento adquirido en mi formación doctoral lo he logrado por medio de mi participación en proyectos como NAIA4.0 (Proyecto Hazitek ZL-2019/00879), OILTWIN (Proyecto Hazitek KK-2020/00052), 3KIA (Proyecto Hazitek KK-2020/00049) y EXPERTIA (Proyecto Elkartek KK-2021/00048).

Al Programa de investigación e innovación Horizonte 2020 de la Unión Europea, por darme la oportunidad de aportar mis conocimientos y de aprender nuevos en el proyecto UPGRID (acuerdo de subvención 646.531) y en el proyecto TOPREF (acuerdo de subvención 604.140).

A mis colaboradores industriales, Iñaki Grau de la empresa GESTAMP e Itziar Landa y Lucia Orbe de la empresa PETRONOR por su disponibilidad y apoyo a mis investigaciones doctorales.

Si he logrado ver más lejos,
ha sido porque he subido a
hombros de gigantes.

Isaac Newton

¿Después? No hay "después".
Porque después el té se enfría,
después el interés se pierde,
después el día se vuelve noche,
después la gente crece,
después la gente envejece,
después la vida se termina; y uno
después se arrepiente por no hacerlo antes,
cuando tuvo oportunidad.

Antoine de Saint-Exupéry

Índice general

Índice de figuras

Índice de tablas

Lista de Acronimos

AT Alta Tensión
MT Media Tensión
BT Baja Tensión
CPS Sistemas Ciberfísicos
CT Centro de Transformación
IoT Internet de las Cosas
IoS Internet de Servicios
BD Big Data
EMS Sistemas de Monitorización Energética
MES Sistemas de Monitorización de la Producción
ML Machine Learning
DL Deep Learning
RL Reinforcement Learning
EW Expanding Window
MW Moving Window
JITL Just In Time Learning
RMSE Root Mean Square Error
NRMSE Normalized Root Mean Square Error
HS Harmony Search
GA Genetic Algorithm
CF Factor de Cresta
RF Random Forest
PCA Principal Component Analysis
IA Inteligencia Artificial
LCT Low Carbon Technologies
FP Flash Point Temperature

Parte I

Introducción

Capítulo 1

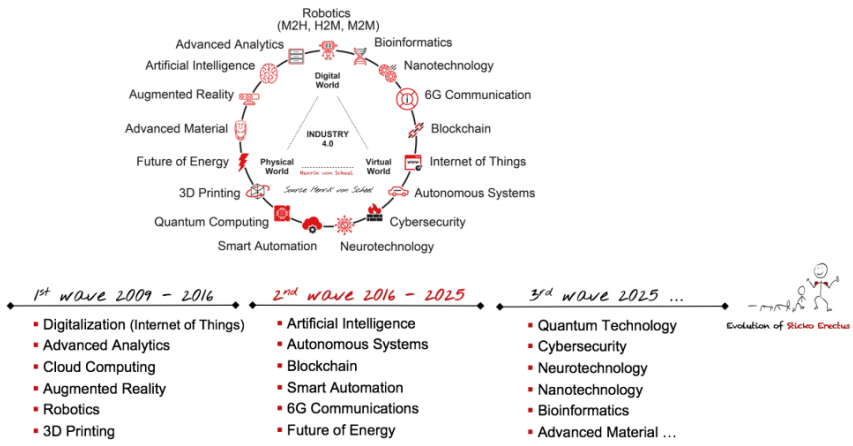
Contexto, Motivación y Objetivos

Este capítulo introduce el contexto y los proyectos de investigación científica que se abordaran en los siguientes capítulos, describe las carencias de los sistemas actuales, así como la motivación a la que ha dado lugar. Asimismo, se definen los objetivos que se van a llevar a cabo en la presente Tesis.

1.1. Contexto

El término Industria 4.0 surge en 2011 [1] como un concepto general acuñado por el gobierno alemán para denominar un nuevo paradigma industrial que define la transición de la fabricación tradicional y basada en maquinaria industrial hasta la fabricación digital. La Industria 4.0 es la fusión de los mundos digital, virtual y físico, que se manifiesta a través, por ejemplo, de los Sistemas Ciberfísicos (CPS). Abarca un conjunto de tecnologías que ya existían hace años aunque de manera inconexa o singular, como Internet de las Cosas (IoT), Internet de Servicios (IoS), Big Data (BD) o Inteligencia Artificial (IA) [2], [3], y que trabajando coordinadas ofrecen oportunidades inigualables de crecimiento y abren la puerta a nuevos modelos de producción como la fabricación bajo demanda [4] y la servitización industrial [5], así como a la optimización de procesos y recursos.

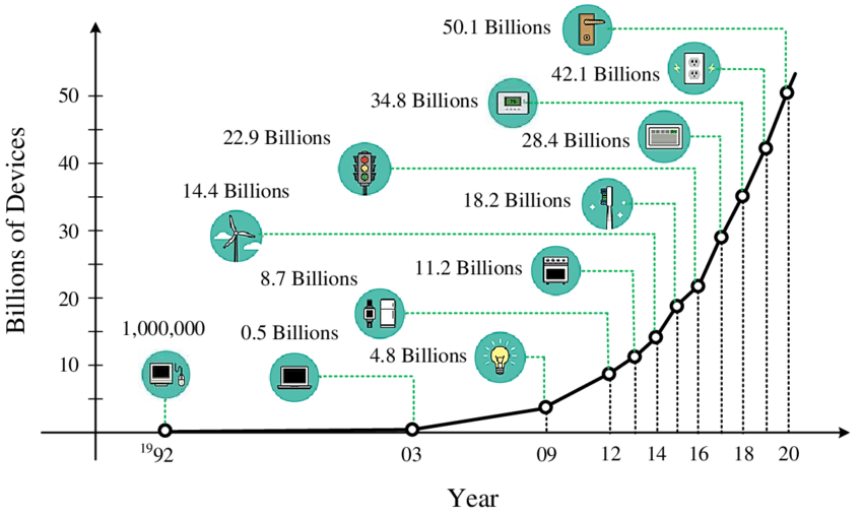
Figura 1.1: La innovación exponencial llega en tres olas de disrupción. Fuente: [6]



Henrik von Scheel, una de las principales autoridades en materia de estrategia y competitividad industrial, identifica en su estudio [7] tres olas de avance digital y automatización (lo denomina "three waves of disruption") asentadas sobre 17 pilares tecnológicos relevantes en el futuro de las industrias, véase Figura 1.1. Cada ola desbloquea el cambio de paradigma de la siguiente ola y así sucesivamente. Según von Scheel actualmente, hemos entrado en la 2ª ola de la 4ª Revolución Industrial, cuya principal característica es el cambio radical de la industria manufacturera [6]: en lugar de supervisar los procesos de forma lineal y operar de forma reactiva, las industrias deben ir aprendiendo y amoldando su comportamiento in-situ a las nuevas circunstancias, alterando su forma de operar, así como la relación con sus proveedores, clientes y otros terceros. La característica más relevante de este cambio de paradigma es la cantidad de datos que se generan. Esta tendencia anticipa un crecimiento exponencial de las

tecnologías relacionadas con IoT [8] como viene ocurriendo desde décadas pasadas, véase Figura 1.2.

Figura 1.2: Tendencia del número de dispositivos conectados a Internet en las últimas 2 décadas. Fuente: [8]



Paralelamente, el término Inteligencia Artificial (Inteligencia Artificial (IA)) surge tras algunos trabajos publicados en la década de 1940 que no tuvieron gran repercusión, pero que a partir del influyente trabajo en 1950 de Alan Turing [9] (test de Turing, examen para evaluar la capacidad de una máquina de exhibir comportamiento inteligente similar a la capacidad humana), se abre una nueva disciplina científica más amplia, la Ciencia de Datos. El objetivo de la Ciencia de Datos es el desarrollo de sistemas computacionales que muestren algún tipo de comportamiento inteligente tomando como referencia a la inteligencia humana [10], [11]. La Ciencia de Datos agrupa el conjunto de técnicas que van desde el análisis numérico clásico y el análisis estadístico, hasta las acuñadas recientemente como Aprendizaje Máquina (Machine Learning (ML)), Aprendizaje Profundo (Deep Learning (DL)), o Aprendizaje por Refuerzo (Reinforcement Learning (RL)).

Esta evolución, marcada por la clara complejidad de los métodos utilizados, viene condicionada porque en los últimos años se ha producido la confluencia de tres factores que han impulsado de manera exponencial las técnicas de aprendizaje artificial basadas en datos. Estos factores son:

- La disponibilidad de grandes cantidades de datos;
- La disponibilidad de grandes capacidades de computación a bajo coste; y
- El desarrollo de modelos de aprendizaje inspirados en aquellos primeros esquemas de los años 40 pero mucho más complejos y que

entrenados con esas grandes cantidades de datos y con esas grandes capacidades de computación permiten resolver problemas inimaginables hace 10 o 15 años [11].

El resto de la sección describe el contexto de los tres casos de estudio que se exponen en esta tesis.

1.1.1. En entornos de Gestión en Redes Eléctricas

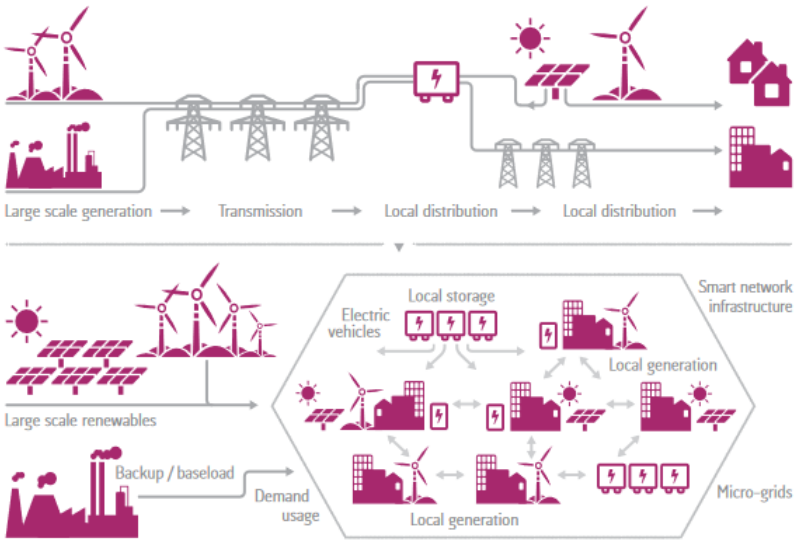
El esquema tradicional de los sistemas de suministro eléctrico se subdivide en tres actividades. Dichas actividades son generación, transporte y distribución, tal y como se ilustra en la Figura 1.3.

El objetivo del subsistema de generación consiste en generar la potencia eléctrica necesaria. Está constituido por el conjunto de todas las centrales generadoras de energía eléctrica.

El subsistema de transporte lleva la energía desde las centrales de generación eléctrica hasta las estaciones de distribución. La energía eléctrica se transporta en Alta Tensión (AT) para disminuir la corriente sin afectar a la cantidad de potencia transportada. De esta manera se minimizan las pérdidas óhmicas, proporcionales al cuadrado de la corriente. Asimismo, la reducción de la corriente permite disminuir la sección de los conductores.

Por último, el subsistema de distribución se ubica en las proximidades del punto de consumo. En esta fase del suministro eléctrico se reducen los niveles de AT de las líneas de transporte hasta niveles de Media Tensión (MT) o Baja Tensión (BT) [12]. Esta transformación adecúa la potencia transportada a la tensión de suministro para que resulte menos peligrosa en el punto de consumo, aunque ello implique aumentar las perdidas técnicas [12].

Figura 1.3: Presente y futuro de la red eléctrica. Fuente: [13]



Esta estructura de la red eléctrica se ha mantenido prácticamente sin alterar desde mediados del siglo pasado. Sin embargo, durante la última década ha comenzado un profundo proceso de cambio, principalmente provocado por el aumento del consumo, la descarbonización, la transición energética y por el avance de la tecnología [13].

La necesidad creciente de consumo de energía de la sociedad actual [14] tiene unas fuertes implicaciones medioambientales. El aumento de la generación necesario para cubrir el consumo creciente tiene como consecuencia, un aumento de las emisiones. La transición energética persigue un cambio estructural que descarbonice la producción de energía. Para ello, es necesario conectar y gestionar nuevos recursos energéticos limpios, generalmente distribuidos en la red. Este hecho exige un alto grado de digitalización para lograr tanto la interacción operativa con dichos recursos, como la supervisión y actuación en tiempo real de la red inteligente (en inglés, Smart Grid) [15].

La aparición de los contadores inteligentes ha sido un hito sin precedentes en esta digitalización de la red de distribución. Antes de su instalación, toda la información que se conocía de la red de BT era a nivel de subestación [16]. Sin embargo, los contadores inteligentes, instalados prácticamente en la totalidad de industrias, empresas u hogares españoles [17], permiten conocer con precisión el consumo de los clientes, ofreciendo un conjunto de datos desconocidos hasta la fecha [18]. Asimismo, habilitan el intercambio de información bidireccional entre el contador y el sistema de telegestión, permitiendo no sólo registrar datos de consumo y monitorizar a los clientes, sino capacidad de actuación [19].

Esta capacidad de actuación amplia el universo de la telegestión, permitiendo actualizar tarifas y contratos, cambios de límites de potencia, efectuar ordenes de desconexión del cliente si procede, etc. [20]. Además, los nuevos mercados más dinámicos y volátiles [21] propician nuevos desarrollos tecnológicos y modelos de negocio, apareciendo nuevos roles, como prosumers o agregadores [22]. De esta manera se propicia que los consumidores se integren en los distintos mercados de flexibilidad que ya se comienzan a desarrollar [15]. Conjuntamente, se ofertan nuevos servicios a los usuarios finales [22], como el aplanamiento de la curva de la demanda, facilitar la penetración de renovables, el establecimiento de la conectividad óptima, el balanceo de las fases de distribución, el descubrimiento de perdidas no técnicas, y/o detección de fallos en el retorno por el neutro entre otros [23].

Por tanto, la mejora en la monitorización de la red de distribución de BT optimiza el uso de los recursos energéticos, ofrece conocimiento en tiempo real de las necesidades de generación y demanda, y capacidad de actuación sobre ellas. En este contexto, la energía se entrega a los consumidores finales a través de un conjunto de Líneas de acometida de MT/BT. Estas líneas proveen de energía eléctrica a los consumidores a través de redes de cuatro hilos [24]. Al conjunto de esos cuatro hilos se llama línea trifásica de cuatro hilos o línea repartidora, donde cada uno de los tres primeros hilos recibe el nombre de fase. Las fases se designan por las letras R, S y T respectivamente. El cuarto hilo se denomina hilo neutro, y se

designa con la letra N [24]. La línea trifásica se conecta a la unidad de embarrado general, que consta de cuatro barras metálicas de cobre donde se conectan cada uno de los cuatro hilos. Cada barra representa una fase, y cada contador se conecta en modo monofásico o bifásico a las barras del embarrado a través de la caja general de protección (CGP).

La CGP es una caja que contiene los elementos de protección necesarios para conectar los puntos de consumo eléctrico a las líneas de la red de distribución. La Figura 1.4 muestra una Caja General de Protección (CGP) y las derivaciones individuales para conectar los contadores de forma equilibrada entre las fases, con el objetivo de evitar desequilibrios en la red. Sin embargo, en la práctica no hay una normativa expresa al respecto de cómo ha de realizarse [25].

Figura 1.4: Instalación eléctrica

Para el mantenimiento o restauración, los técnicos se aproximan físicamente al CGP y reequilibran manualmente las cargas. Habitualmente se tiende a conectar un consumidor a una fase, en base a su perfil estadístico de consumo de potencia o a su potencia contratada. Estos datos no reflejan cuándo y cuánta potencia se consume [26]. Se tarda entre 10 y 15 minutos en cambiar una carga, por lo que el trabajo total puede durar una hora, además del tiempo de desplazamiento hasta el lugar. El coste total se puede elevar hasta varios miles de euros [27].

En la decisión de reequilibrar una línea se tienen en cuenta tres factores: el coste económico del cambio de cargas, el desequilibrio previsto de la Línea y la interrupción temporal del suministro al consumidor [28]. El reequilibrado se admite bajo dos supuestos: conexión de un nuevo cliente o desequilibrios severos entre fases [27].

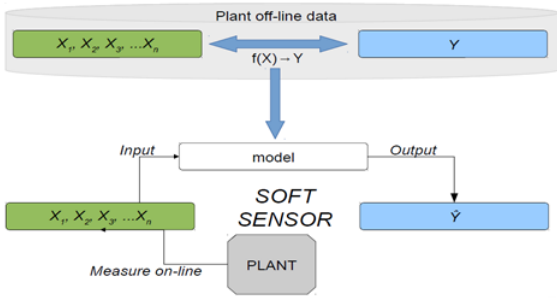
Los desafíos mencionados plantean la evolución hacia una nueva red eléctrica basada en dispositivos inteligentes distribuidos y capaces de ejecutar algoritmos de control dirigidos a una amplia gama de actuaciones [15]. Por tanto, la digitalización del CT para convertirlo en el Centro de Transformación inteligente (CTi) es un punto clave para afrontar todos los retos de la distribución energética [22].

1.1.2. En entornos de Control y Supervisión en Procesos Industriales

En los entornos de control y supervisión de procesos industriales modernos, casi todo se puede medir a través de diferentes sensores físicos. Pese a todo, algunas mediciones son difíciles de obtener. [29], [30].

En casos donde la utilización de sensores físicos está condicionada, se utiliza el concepto de **sensado virtual**. Un sensor virtual crea un modelo de inferencia, a partir de un conjunto de datos históricos, capaz de aprender ciertas relaciones de causalidad (causa-efecto) multi-paramétricas y altamente no lineales. Kadlec, uno de los pioneros del concepto de sensórica virtual, lo definió como [30]: *"los sensores virtuales son modelos inferenciales que utilizan variables fácilmente medibles para estimar variables del proceso que son difíciles de medir debido a limitaciones tecnológicas, grandes retrasos en las mediciones o altos costes de inversión"*. Se utiliza, el poder de los algoritmos de ML, para calcular con precisión las mediciones que faltan o que no se pueden medir. Este concepto se representa esquemáticamente en la Figura 1.5.

Figura 1.5: Esquema básico de un sensor virtual



Instalar y mantener una red de medición dedicada a la monitorización de una planta es complejo en términos de instalación de sensórica y costoso en términos económicos. El presupuesto necesario puede afectar significativamente a los costes totales del funcionamiento de la planta. A continuación, se describen cinco escenario típicos para resolver con sensores virtuales [29], [30]:

- **La medición es costosa en términos económicos y se opta por renunciar a sensórica en tiempo real para monitorizar variables del proceso**.
  En ocasiones, el coste de instalación y/o el coste del sensor es elevado. Cuando la alternativa es no instalar el sensor, el proceso se vuelve más difícil de controlar. La suplantación de sensores físicos por sensores virtuales proporciona, no solo mejores estimaciones económicas, sino que, además, proporciona valores igualmente válidos para el control del proceso [31]. En este escenario, es importante prestar mucha atención al análisis del rendimiento del modelo, entendiendo por modelo a las variables inferidas por el sensor virtual. Se debe

realizar una validación periódica del modelo mediante la inserción temporal de dispositivos de medición y, eventualmente, proceder al restablecimiento de los sensores virtuales [32]. Es decir, es necesario un recalibrado o ajuste periódico de los sensores virtuales (término conocido en la literatura como soft-sensor retuning [29]). Esta necesidad de recalibración es, en realidad, un requisito común a cualquier aplicación de sensor virtual. La necesidad de dicho reajuste se debe a la posible aparición de nuevas relaciones entre las variables medidas y la variable inferida, no consideradas en la fase de diseño del sensor virtual. La aparición de nuevas relaciones, normalmente, es debida a cambios en la materia prima, cambios de comportamiento parcial de las máquinas o cambios en los productos a generar, entre otros. Si el gestor de planta tiene intención de eliminar alguno de los sensores físicos de medición, la disponibilidad de dispositivos de medición físicos para la recalibración del sensor virtual debe planificarse adecuadamente.

- **La medición es muy costosa en términos económicos y, por lo tanto, la monitorización de variables de proceso se realiza sólo periódicamente**.
  Semejante al escenario anterior, pero en este caso, no se opta por renunciar completamente a la medida de los sensores físicos, sino que, con objeto de reducir costes, únicamente se obtienen medidas de manera periódica. Por ejemplo, para variables que a través del análisis en laboratorio se obtienen con un retardo considerable. En este caso, el sensor virtual infiere aquellas características que hasta ahora solo podían ser medidas - a posteriori - vía laboratorio. Esto conlleva claros beneficios económicos y de control, y además libera de tareas al laboratorio, haciendo más eficiente su labor. El modelo se reajusta/actualiza cada vez que se dispone de nuevas medidas de laboratorio, permitiendo al sensor virtual proporcionar estimaciones en tiempo real de la variable en los intervalos en los que se desconoce la medida real, mientras que las mediciones retardadas (medidas de laboratorio) permiten reajustar el modelo y mejorar el rendimiento del sensor virtual al evitar el efecto de propagación de errores.

- **No existe medida real debido a limitaciones en la tecnología de medición**.
  Por un lado, no existe instrumentación que pueda medir directamente las variables de proceso de interés. En este caso, el sensor virtual proporciona estimaciones de las variables utilizando las mediciones disponibles. Por ejemplo, para calcular la carga de calor en un sistema se trata de combinar el conocimiento de la dinámica del sistema y las mediciones de las tasas de flujo másico de refrigerante y la temperatura del sistema para calcular la carga de calor.

  Por otro lado, no se puede instalar instrumentación en el lugar de interés debido a las duras condiciones o simplemente a la falta de espacio. En este caso, el sensor virtual proporciona estimaciones de las variables, utilizando mediciones disponibles tomadas en diferentes

ubicaciones que están correlacionadas con la variable de interés. Por ejemplo, si no hay lugar para colocar el termómetro en el lugar de interés, se trata de estimar la medida en el lugar adecuado mediante el uso de otras temperaturas alrededor de dicho punto.

- **Apoyo en los procesos de mantenimiento de dispositivos de medición físicos**.
  En el entorno industrial a pesar de todas las precauciones, ciertos fallos imprevistos de medición son ineludibles y requieren de operaciones de mantenimiento no planificadas. Hasta ahora, la única solución adoptada era replicar los sensores, consiguiendo así cierta redundancia física. En estos casos, es especialmente recomendable la utilización de sensores virtuales diseñados para sustituir momentáneamente los dispositivos de medición no disponibles y evitar la detección del proceso industrial.

- **Detección de fallos en las medidas de los dispositivos de medición físicos**.
  Cuando funcionan en paralelo dispositivos de medición físicos y sensores virtuales, los fallos de medida pueden detectarse como cualquier desviación entre ambas medidas, y en caso de necesidad, la medida del sensor virtual puede aprovecharse para proporcionar una estimación de la salida del sensor en caso de fallo.

Los problemas más comunes en la búsqueda de soluciones mediante la implementación de sensores virtuales son [29]:

- las relaciones no lineales y desconocidas entre las variables, ya que en ocasiones sólo se conocen (parcialmente) las relaciones causa-efecto, pero las leyes físicas exactas que rigen los fenómenos son demasiado complicadas de construir con un modelo eficaz;

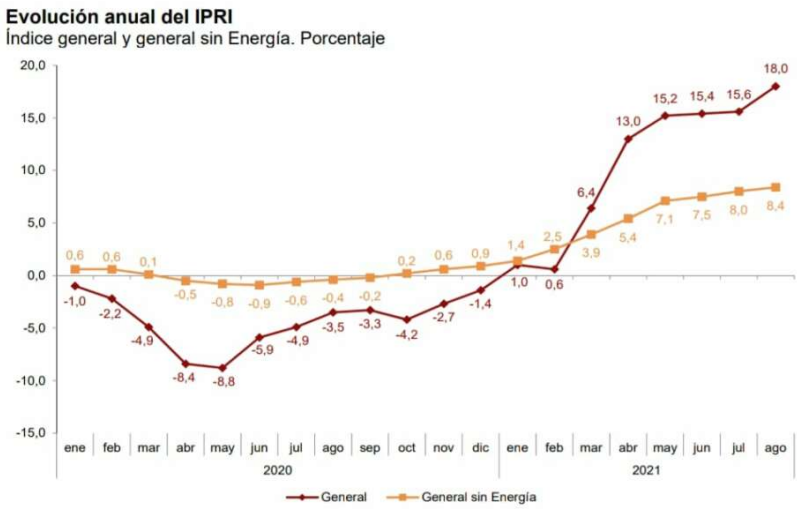
- la solución, habitualmente se ha de diseñar a partir de sensores ya existentes, sin modificar la frecuencia de muestreo de mediciones o formato de los datos y sin modificar su actual ubicación;

- a menudo es imposible reconstruir los escenarios específicos que causaron un determinado fallo, ya que el único recurso con el que se cuenta son los datos de mediciones pasadas, y que normalmente no consta de información adicional útil como registro de aparición de fallos, paradas de mantenimiento, fecha de recalibración de la sensórica o cambios en la operativa del proceso. Este tipo de conocimiento queda, a menudo, en manos de las personas expertas del dominio [33].

1.1.3. En entornos de Modelización en Producción Industrial

La digitalización de las fábricas y sus modelos de producción ofrecen un enorme abanico de posibilidades a la hora de mejorar la competitividad y la sostenibilidad de las empresas [1], [34]-[36].

Figura 1.6: Evolución de los precios industriales, con y sin coste energético asociado. Fuente: Instituto Nacional de Estadística (INE)[37]



En este escenario, los retos típicos en la digitalización de las fábricas, son: los costes energéticos, la ingente cantidad de datos y la complejidad de los sistemas productivos.

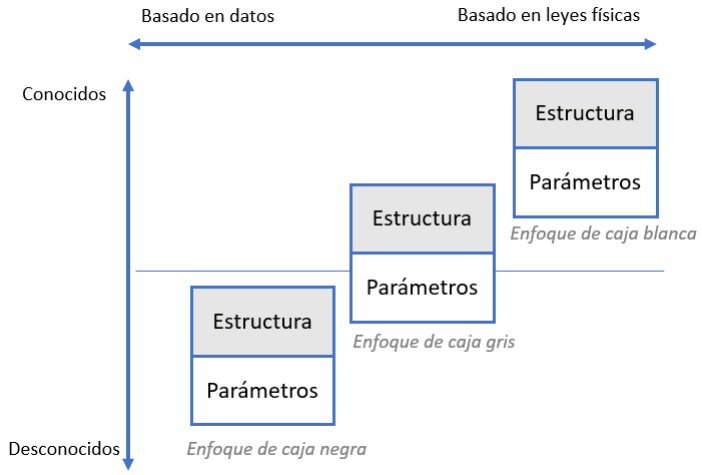
- El aumento progresivo de los costes de la energía [38] supone una preocupación creciente en el sector industrial, poniendo en muchos casos en peligro su propia competitividad. El problema radica en que, el coste de la electricidad ha crecido en mayor proporción que lo que ha crecido el valor de la producción industrial [39]. Y con ello, la competitividad de las empresas baja, en muchos casos porque sus costes de producción suben. La Figura 1.6 muestra la evolución de los precios industriales a través del Índice de PRecios Industriales (IPRI), considerando y sin considerar el coste energético. La Figura muestra como el índice IPRI acumula en 2021 meses de subida y se intensifica hasta elevarse al 18 % en agosto de 2021 por la influencia del coste de la electricidad [37]. Se estima que, dependiendo del sector, la factura energética en industria supone entre un 10 % y un 45 % del coste operacional total [40].

- Bajo la norma ISO 50.001 [41], la reducción del uso de la energía se considera esencial para minimizar la emisión de gases de efecto invernadero. Cada vez se implantan más sistemas de gestión energética que monitorizan el consumo industrial (desde nivel de planta de planta de producción, a proceso y máquina) generando ingente cantidad de datos [42]. Con la explosión de los datos en los entornos industriales, la capacidad humana para la toma de decisiones se

decrementa enormemente, disminuyendo su capacidad de inferencia (la tiranía del dato [43]).

- Y finalmente, la irrupción y el continuo desarrollo de nuevas o más diversas materias primas, nuevos y más diversos productos, nuevas y más diversas maneras de producir han generado ecosistemas industriales cada vez más complejos. Dichos ecosistemas introducen datos de un mayor número de variables de decisión y/o actuación, que exigen nuevos sistemas para la optimización y control de los procesos y bienes industriales.

Figura 1.7: Conocimientos requeridos en base al enfoque del modelo del proceso. Fuente [44]



Este contexto ha estimulado el desarrollo de nuevas estrategias de monitorización del consumo de energía en los entornos industriales hiperconectados [45]-[48]. Según el grado de conocimiento que se tenga a priori del dominio, se puede aplicar tres enfoques diferentes en la modelización de procesos (véase la Figura 1.7). En la Tabla 1.1 se presenta una comparativa de estos tres enfoques [49]:

- Enfoque de modelos físicos o de primer principio (modelos de caja blanca). Modelos que se basan en la observación de un fenómeno físico, y que tratan de establecer las leyes físicas que lo explican mediante ecuaciones matemáticas que se resuelven por métodos como diferencias finitas, balances de energía o de masa. El uso de herramientas de modelado físico suele ser la primera etapa en el desarrollo de soluciones de control industrial tradicionales. Sin embargo, ciertos estudios[50], [51] han subrayado que los enfoques que se basan exclusivamente en la modelización física consumen mucho tiempo, son difíciles de analizar e interpretar, no integran fácilmente la información ni los resultados de otras fuentes y escenarios similares.

Limitaciones relacionadas con este tipo de aproximaciones basadas en ecuaciones físicas paramétricas y su resolución [52], son:

1. requieren de conocimientos avanzados en el área que se desea modelizar, pero en muchos casos, se desconocen las relaciones de causalidad con todas las magnitudes que afectan al sistema;
2. suele estar limitado por especificaciones técnicas incompletas, interrelaciones extremadamente complejas u ocultas que son difíciles de caracterizar, por la falta de disponibilidad de toda la información necesaria y por el coste de una modelización completa [52]. En estos casos, se suele tender a la modelización simplificada y, en consecuencia, la capacidad de generalización resulta difícil [53].
3. se basan en simulación de situaciones ideales, lo que introduce mucha desviación;
4. requieren una gran capacidad de computación y tiempo, lo que les limita a la hora de poder ser aplicados como sistemas de monitorización en tiempo real;
5. es complicado introducir en el modelado clásico la dependencia de muchas de las magnitudes implicadas con el tiempo.

- Enfoque de modelos basados en datos o data-driven (modelos de caja negra). Modelos que ignoran cualquier conocimiento del dominio y que desconocen las leyes físicas inherentes a los procesos industriales. Estiman una solución exclusivamente basada en datos previamente observados, y representados mediante series de tiempo de datos históricos monitorizados. En este contexto industrial, los datos monitorizados relativos a consumos energéticos los proporcionan los EMS [54] (en inglés, Energy Management Systems) y los datos relativos a producción los proporcionan los MES [55] (Manufacturing Execution Systems). Estos datos permiten crear esquemas numéricos con gran capacidad de generalización, con aplicación en múltiples casuísticas, y con adaptación específica a cada uno de ellos a través de sus datos particulares [44]. Estos esquemas también presentan problemas:

  1. requieren grandes cantidades de datos;
  2. sólo son capaces de clasificar, predecir y optimizar aquello que está recabado en los datos históricos;
  3. son considerados como sistemas opacos (modelos de caja negra);
  4. son sistemas estocásticos, no deterministas, sometidos a incertidumbre; y
  5. su implantación y generalización suponen un claro reto.

- Enfoque de modelos híbridos (modelos de caja gris). Los métodos híbridos combinan ecuaciones matemáticas propias del dominio de proceso, con técnicas basadas en modelado de datos y que permiten evolucionar de la ciencia experimental (qué hacer) hacia la ciencia

aplicada (cómo hacerlo) [56]. El análisis y modelado de sistemas híbridos combina la interpretabilidad, el conocimiento exhaustivo y la comprensión de un enfoque de modelado basado en la física, con la capacidad de inferencia y de identificación automática de patrones de los algoritmos avanzados basados en datos [34]. Esta hibridación permite una toma de decisiones más rápida y eficaz, simplifica y evita procesos que no aportan valor, aumenta la agilidad en los procesos gracias a la información disponible en tiempo real, mejora la autonomía, las decisiones tomadas o desencadenadas por la propia infraestructura, previene o anticipa acciones futuras y ofrece una precisión en determinadas tareas que evita el error humano [57].

La integración de los sistemas basados en la IA y los sistemas de modelización física allana el camino para afrontar los retos en entornos inciertos [50]. La hibridación de ambos enfoques permite extraer conocimientos basados en datos y explicar la lógica que subyace a los algoritmos de IA a partir del conocimiento del dominio y del entorno en el que se aplican. Abordar un problema desde la doble perspectiva del conocimiento del dominio y del modelado basado en la IA añade potencialidad e interpretabilidad al resultado final.

Se basa en trabajar con datos de calidad, utilizar los datos disponibles de forma más eficaz, ya sea trabajando con bajos volúmenes de datos o extrayendo más valor de fuentes de datos no estructuradas y diversas. Para 2025, Gartner [58] estima que el 70 % de las organizaciones se vean obligadas a cambiar su enfoque de "Big Data" a "Small and Wide Data".

Los modelos físicos son tanto más útiles cuando se dispone de pocos datos y por ello, especialmente adecuados en etapas tempranas del diseño de los procesos. Los modelos basados exclusivamente en datos o de caja negra no codifican leyes físicas (por ejemplo, la ley de la conservación de la masa) y por ello, deben lidiar con el escepticismo inicial de los expertos de dominio. Requieren de grandes cantidades de datos y existen múltiples preocupaciones, principalmente con la escasa integración con el conocimiento previo, la interpretabilidad y la fiabilidad de las soluciones entre otros [59]. La comunidad científica reconoce que la adopción de enfoques puros -modelos físicos o en modelos de datos- permite abordar una serie de retos, pero los investigadores apoyan un enfoque mixto para transformar la realidad de problemas concretos en industria reales [60]. Esto implica explotar el conocimiento específico del dominio e integrar los modelos físicos u otras formas de razonamiento con el procesamiento impulsado por los datos [61]. Para abordar complejos problemas de industria es necesario personalizar los algoritmos de Ciencia de Datos para incorporar el conocimiento del dominio junto con los métodos basados en datos [61]. Tres son las claves del éxito:

- La generalización, flexibilidad y dinamismo de las herramientas basadas en modelos de datos. Crear sistemas capaces de adaptarse a

Tabla 1.1: Características de los modelos de caja blanca, negra y gris

	Modelos caja blanca	Modelos caja negra	Modelos caja gris
Dependencia de los datos	Muchos datos	Muchos datos	Datos de calidad
Dependencia de la experiencia en el dominio	Alta dependencia del conocimiento del dominio	Puede proporcionar resultados útiles con poco conocimiento del dominio	Complementa las inferencias de los modelos basados en datos con el conocimiento de dominio
Fidelidad y robustez	Las leyes físicas permiten manejar relaciones altamente no lineales y complejas	Capacidad de generalización limitada asociada al alcance y complejidad del conjunto de entrenamiento	Las leyes físicas y la capacidad de inferencia permiten manejar relaciones no lineales y complejas
Adaptabilidad y capacidad de despliegue	Adaptación lenta y compleja ante nuevos comportamientos	Fácilmente adaptables a nuevos comportamientos	Fácilmente adaptables a nuevos comportamientos
Interpretabilidad	Vínculo físicamente significativo e interpretable entre los parámetros	Limitada a la rigidez del algoritmo seleccionado para la modelización	Vinculo interpretables entre los parámetros
Gestión de la incertidumbre	Las incertidumbres pueden acotarse y estimarse	No es posible acotar las incertidumbres	Las incertidumbres pueden acotarse y estimarse
Gestión de sesgos	Pocos sesgos	Los sesgos en los datos se reflejan en la predicción	Pocos sesgos, aunque algunos se reflejan en la predicción
Complejidad de modelado	Muy complejo	Poco complejo	Poco complejo

nuevos comportamientos o nuevos datos cuyo patrón no estaba recogido en el entrenamiento, proporciona herramientas dinámicas y versátiles frente a cambios en la operativa del proceso [62].

- La imbricación del conocimiento experto en la concepción de los modelos de datos. Las técnicas conocidas como Human-in-the-Loop (HitL [63]) pretenden diseñar algoritmos/modelos que, aunque capaces de tomar decisiones de forma autónoma, estas decisiones sean sometidas a la intervención del ser humano en cualquier momento del ciclo de la toma de decisiones. Con esto se persigue incorporar el conocimiento externo en los modelos IA y favorecer la capacitación de habilidades digitales en dicho experto.

- Apostar por soluciones capaces de simplificar la interpretación de los modelos hasta un punto que sean explicados, entendidos y confiables para los expertos de dominio [64].

En este contexto, el máximo representante de la simbiosis entre la captación de información sensorizada, las tecnologías de análisis de datos y el conocimiento del conocimiento experto, es el concepto de sistema ciberfísico (en inglés, CPS). Entendido los sistemas ciberfísicos como el conjunto de máquinas conectadas bidireccionalmente y capaces de interactuar entre sí autónomamente en función del funcionamiento de la planta industrial. O en palabras de los autores [65]: "sistema de entidades computacionales que colaboran entre sí y que están intensamente conectadas con el mundo físico circundante y sus procesos en curso, proporcionando y utilizando, al mismo tiempo, servicios de procesamiento de datos y de acceso a los mismos

disponibles en Internet". En los sistemas ciberfísicos, donde el ser humano también participa en la toma de decisiones, es fundamental garantizar que comprende el proceso deductivo que la máquina sigue [64].

1.2. Retos y Motivación

La motivación de la presente Tesis surge ante la necesidad de consolidar la dinámica de colaboración científica y tecnológica entre el mundo académico y la realidad industrial, a través de contribuciones tangibles - casos de uso, artículos, congresos, jornadas de difusión, patentes y registro de software - con el objetivo de favorecer y apoyar la innovación en el tejido productivo.

La realidad es que la Inteligencia Artificial, más allá de su propio hype, aún ha de resolver grandes retos y ha de superar la barrera de impacto en la industria. De hecho, tras 8 años atascado en la cima del Hype Cycle de IA de Gartner [66], [67], parece que Machine Learning avanza, aunque sigue estando a una distancia entre 2 y 5 años de su llegada a mercado, como se puede observar en la Figura 1.8. Esta necesidad queda reflejada en los siguientes indicadores:

- El 85 % de los proyectos de Big Data fracasan. (Gartner, 2017 [68])
- El 87 % de los proyectos de Ciencia de Datos nunca llegan a la producción. (VentureBeat, 2019 [69])
- Hasta 2022, sólo el 20 % de los conocimientos analíticos ofrecen resultados empresariales. (Gartner, 2019 [70])

Pero ahora ha llegado el momento de introducirlo correctamente en los procesos de negocio [71]. De hecho, Gartner identifica la necesidad urgente de las industrias de acelerar la velocidad a la que las pruebas de concepto (PoC) pasan a producción [67].

Y es que, a pesar de las grandes expectativas depositadas [72], la adopción de la IA por parte de las empresas se encuentra aún en una fase muy temprana que varía en función de la madurez digital de la empresa. Según un informe del MIT [73] - realizado a partir de una encuesta mundial a más de 3.000 directivos y 30 entrevistas a expertos en tecnología - en 2017, el 80 % de los ejecutivos estaban de acuerdo en que la IA era una oportunidad estratégica para su organización, pero la realidad es que apenas el 5 % lo ha incorporado en sus procesos. La Figura 1.9 resume que el 54 % de las empresas no habían adoptado aún ninguna solución basada en IA, bien porque aún no lo contemplaban o porque aún no se había dado la oportunidad, el 23 % de las empresas habían implementado algún piloto y aún estaban en fase de análisis de resultados y sopesando la viabilidad de llevar tal desarrollo a producción. El 18 % de las empresas integraban IA en modo experimental en alguno de sus procesos y únicamente el 5 % apostaban activamente por la IA como elemento diferenciador.

En mayo de 2020 otro informe realizado por el instituto de investigación Capgemini [74] vuelve a examinar el ritmo de adopción de la IA. Y

Figura 1.8: Ciclo de Gartner para tecnologías de Inteligencia Artificial. Fuente: [67]

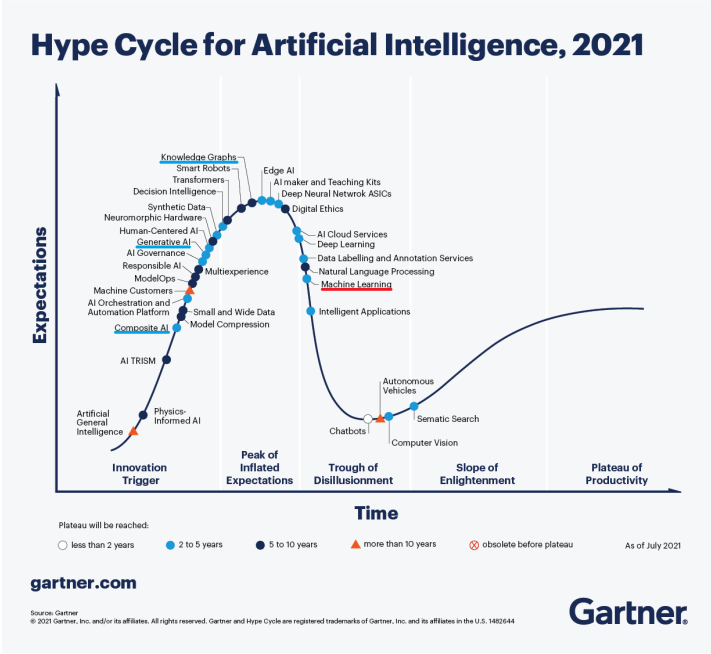
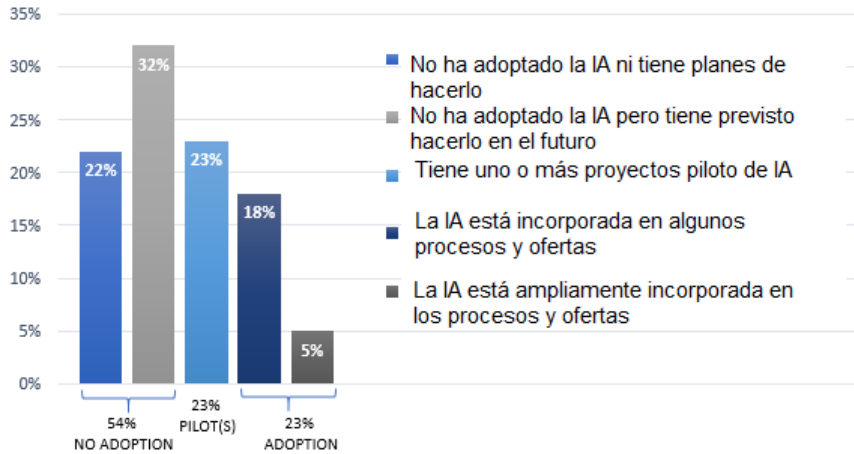


Figura 1.9: Nivel de adopción de IA en entornos empresariales e industriales. Fuente: [73]



aunque se percibe una leve mejoría (en 3 años se incrementa del 5 % al 13 % el porcentaje de empresas que han implementado satisfactoriamente desarrollos de IA en producción), el mensaje sigue siendo el mismo, la penetración de la IA en el mundo industrial sigue siendo insuficiente. Estas diferencias tienen menos que ver con las limitaciones tecnológicas y mucho más con el negocio, donde las prioridades de inversión y la falta de claridad de los casos de negocio son obstáculos más importantes para la implantación de la IA que las propias capacidades tecnológicas. Esto se debe a que, cuando se piensa en producir una solución de Ciencia de Datos, no se trata sólo de desarrollar la solución tecnológica. También se trata de integrarse estrechamente en los procesos de negocio, y si es necesario, transformarlos. Es un hecho que las empresas no invierten en IA, siempre invierten en sus problemas de negocios [73]. Pero el mensaje ha de ser positivo, y la realidad es que la Ciencia de Datos tiene un impacto transversal y transformador en los procesos industriales. Ahora bien, el reto de desplegar e integrar una solución basada en datos con éxito es, prácticamente, una labor de concienciación y convencimiento de que incorporar esta tecnología en la cultura de la empresa acarrea beneficios a medio y largo plazo. A continuación, se analiza esta necesidad en tres situaciones:

1.2.1. Situación en la Gestión de Redes Eléctricas

En la literatura, han sido varios los enfoques y metodologías de optimización que se han propuesto para dar solución al desequilibrio de fases en Líneas de Baja Tensión: desde diferentes técnicas meta-heurísticas hasta técnicas basadas en redes neuronales. Pero, según el conocimiento de la autora, ninguna de las soluciones propuestas tiene en consideración datos reales en horizontes temporales grandes (capaces de reflejar el comportamiento estacional de los consumidores) sobre topologías de red reales, que permita aprovechar (i) la información proporcionada por los contadores inteligentes en relación a la curva de consumo horaria de cada consumidor, (ii) ni tampoco contempla la existencia de consumidores conectados a dos fases (las soluciones propuestas solo hacen mención a conexiones monofásicas), (iii) ni que se aproveche el conocimiento de dominio para abordarlo de manera conjunto y global.

Los condicionantes anteriores motivan la necesidad de búsqueda de desarrollos de **nuevos procedimientos de optimización** que aborden el problema de intercambio de contadores entre fases teniendo en cuenta la varianza temporal de los patrones de carga en conexiones monofásicas y bifásicas.

1.2.2. Situación en Control y Supervisión de Procesos Industriales

Hasta la fecha, cuando se trataba de sensorizar un proceso para monitorizar procesos industriales, solo se conocían aquellos valores medibles mediante sensores físicos. Pero ahora, con la ayuda de los sensores virtuales y a través de **técnicas predictivas de inferencia** se es capaz de estimar

variables difíciles de medir a partir de otras fáciles de medir mediante modelos matemáticos de inferencia. Este campo de estudio proporciona a la industria una valiosa herramienta que permite un mejor control en las fases de los distintos procesos. Este tipo de modelos permite predecir, en tiempo real, variables de proceso que normalmente exigen sensores costosos, lentos o inexistentes.

Ante esta situación se ve necesario desarrollar una metodología centrada en la búsqueda de datos de calidad y en estrategias de aprendizaje adaptativas, que generen modelos de inferencia capaces de averiguar la relación no lineal entre variables de entrada y variables de salida.

1.2.3. Situación en el Modelado de Producción Industrial

Con intención de suplir las carencias de los sistemas actuales para la detección de ineficiencias energéticas en plantas de producción industrial y el análisis de la causa raíz que origina dicha ineficiencia, se identifica la necesidad de proponer una nueva metodología. Esta metodología permite al experto de dominio, a través de la formulación de sencillas preguntas, encontrar las respuestas adecuadas para extraer valor en su interacción con el conocimiento generado por la herramienta. No se trata de una herramienta que extrae la inteligencia del dato o que aporta conclusiones, es el humano quien, a través de la interacción con los datos, aumenta sus capacidades y conocimiento sobre el funcionamiento de la planta de producción. La metodología está concebida para facilitar la interpretación de grandes volúmenes de datos ayudando al humano a centrar el foco en lo que realmente aporta conocimiento. A través de las **técnicas descriptivas y regresivas más sencillas**, proporciona herramientas al humano que - de manera automática - despiertan alertas, generan informes y que además le aportan capacidad de análisis y de interpretación, en esta interacción con los datos.

La metodología integra sencillas técnicas estadísticas, técnicas de agrupación y técnicas regresivas multi-paramétricas para modelar la normalidad y la detección de ineficiencias energéticas y también pone a disposición del humano la información necesaria para analizar los comportamientos de las máquinas individuales que le permitan dar una heurística para averiguar la causa raíz de la ineficiencia.

1.3. Objetivos

El objetivo principal de esta Tesis es contribuir con soluciones realistas a los Sistemas Industriales mediante la aplicación de técnicas de Inteligencia Artificial. La consecución de dicho objetivo se organiza mediante su división en tres objetivos secundarios, capaces de guiar las tareas de investigación específicas a desarrollar:

- **Objetivo 1. Desarrollo de soluciones de optimización heurística y meta-heurísticas para establecer sistemas equilibrados en la Red Eléctrica de Baja Tensión.**

  Se propondrán métricas de contexto y algoritmos de optimización combinatoria complejos NP-hard para la asignación dinámica y óptima de los consumidores de la red de Baja Tensión, a cada una de las fases (a nivel de Centro de Transformación (CT) y Línea) en una red eléctrica trifásica. Se trabajará con las curvas de carga individuales de consumo, proporcionadas por los contadores inteligentes de cada usuario, con el objetivo de lograr un sistema de cargas equilibrado.

- **Objetivo 2. Desarrollo de soluciones predictivas de inferencia y modelado de sensores virtuales para control y supervisión en plantas industriales.**

  Se formularán técnicas de pre-procesamiento de los datos y estrategias adaptativas de aprendizaje basadas en algoritmos de analítica predictiva que permitirán estimar variables de proceso difíciles de medir, a partir de otras variables fáciles de medir. Para ello, se establecerá una metodología y se validará su generalización a través de la realidad de tres casos de uso, que, aunque dispares en concepto son afines en tecnología. Se concretará el concepto de sensor virtual como herramienta para el mejor control en la monitorización de procesos industriales en tiempo real. Se estimará también la afección económica de la aplicación de esta herramienta en un entorno real.

- **Objetivo 3. Desarrollo de soluciones descriptivas para el modelado de la eficiencia energética en plantas de producción para la detección de ineficiencias energéticas y descubrimiento de la causa-raíz de tales ineficiencias.**

  En el contexto de plantas de producción basadas en sistemas ciberfísicos, las máquinas, procesos y/o líneas productivas y no-productivas intercambian continuamente datos entre sí. Se propondrá una metodología basada en el conocimiento del dominio y en el estudio de varias técnicas de analítica descriptiva, que permitirán modelar el comportamiento dinámico de dichas plantas industriales de producción, detectar ineficiencias y analizar la causa raíz de tales ineficiencias energéticas. Para demostrar la validez del estudio, se aplicará la metodología en una planta industrial del sector del automóvil, que permitirá comprender las diferentes formas de operar de la planta en términos de producción de material y consumo energético, además de identificar y cuantificar las situaciones de ineficiencias energéticas.

1.4. Enfoque industrial de la Tesis

Esta Tesis se ha desarrollado en el marco del Programa de Doctorado de Tecnologías de la Información y Comunicaciones en Redes Móviles de la UPV/EHU. El planteamiento empleado es un enfoque claramente de un proyecto de investigación industrial dada su inherente capacidad de

aplicar los resultados obtenidos del estudio a la industria (y a la sociedad). Ha sido realizada a través de varios programas de colaboración públicos entre el Centro Tecnológico Tecnalia -con el que la doctoranda mantiene una relación laboral de más de 21 años- y diversas empresas privadas, que han favorecido el desarrollo de los proyectos de investigación industrial que en la presente se exponen.

El enfoque esencial de la Tesis es que, a partir de conceptos tecnológicos formulados (TRL2), se ha conseguido evolucionar hasta la demostración de prototipos en entornos relevantes (TRL7). Inicialmente, el punto de partida de la presente Tesis han sido publicaciones científicas que describen y proporcionan análisis que respaldan supuestos de un aplicativo concreto. Posteriormente, y a través de la investigación activa en la bibliografía de supuestos semejantes, se procede al desarrollo de metodologías y herramientas que no consta que se hubiesen integrado y/o probado con antelación. Todos los modelos analíticos se van a validar con datos reales en un entorno simulado. Finalmente, cuando esta validación culmine con éxito, se va a aplicar e integrar en el entorno real de producción: desde la dinámica de adquisición de datos en tiempo real, hasta la incorporación del conocimiento al sistema.

1.5. Estructura del documento

Esta Tesis está estructurada en tres grandes bloques. La Parte I consta del **Capítulo 1** que proporcionará una breve introducción de conceptos generales relativos a Industria 4.0 y situación actual de las técnicas analíticas aplicadas a diversos sectores. El capítulo tratará los obstáculos que encuentra la Ciencia de Datos para llevar a cabo sus desarrollos a entornos productivos operativos y el reto que supone que una Ciencia que proporciona empíricamente una ventaja competitiva realmente lo logre. Estos condicionantes sentaran las bases para aclarar la motivación y la definición de los objetivos de la presente Tesis.

La Parte II, con los capítulos 2, 3 y 4, recogerá tres metodologías y herramientas dispares en cuanto a tecnología y dominio, con objeto de aportar claridad en los casos de aplicación industrial. El **Capítulo 2** presentará un estudio basado en modelos/algoritmos meta-heurísticos para la optimización en la red inteligente de distribución de BT para la mejora de la operación en el ámbito energético. El **Capítulo 3** introducirá un estudio metodológico basado en técnicas de inferencia predictivas para la sensorización virtual en entornos hostiles. Demostrará la capacidad de generalización del método en la realidad de tres escenarios industriales. El **Capítulo 4** introducirá un estudio metodológico basado en la concatenación de modelos/algoritmos descriptivos para la detección de ineficiencias energéticas e identificación de su causa raíz, en plantas de producción industrial.

Finalmente, la Parte III, a través del **Capítulo 5** presentará el resumen final de las conclusiones de cada capítulo, las principales aportaciones de la Tesis e indicará varias direcciones de investigación que derivaran de las conclusiones presentadas.

Parte II

Metodología, análisis y resultados

Capítulo 2

Contribuciones meta-heurísticas para la Optimización en el equilibrado de cargas en Redes Eléctricas

Este capítulo explora la resolución de problemas de optimización combinatoria (NP-hard) para el equilibrado energético entre fases de Línea de las Redes de Distribución de BT a través del diseño de esquemas numéricos basados en Ciencia de Datos. Se persigue ajustar las necesidades del problema, a la complejidad de las técnicas mediante (i) un algoritmo Greedy, que basándose en cierto conocimiento del problema, plantea un proceso iterativo y determinista, como proceso de búsqueda basado en la concatenación de decisiones óptimas parciales; y su contrapartida, mediante (ii) métodos de optimización estocásticos globales a través de esquemas meta-heurísticos agnósticos al problema (Harmony Search (HS) y Genetic Algorithm (GA)).

2.1. Introducción

2.1.1. Estado del arte

Tradicionalmente, la planificación y despliegue de la red eléctrica se ha basado en estimaciones estadísticas como modo de aproximación más efectiva ante el desconocimiento de la topología de la red o el desconocimiento de los patrones de consumo energético. Sin embargo, los desafíos actuales de la red eléctrica implican la necesidad de un mejor control de la red de Baja Tensión, más precisa y en tiempo real. Desafíos como el desequilibrio de consumos entre fases están estrechamente ligados con la creciente complejidad en la gestión de la red y el crecimiento exponencial de los consumos energéticos.

Por una parte, surgen nuevos roles en las redes de distribución (por ejemplo, el rol de prosumer o individuo capaz de aprovechar su autogeneración de energía para consumir, almacenar, compartir o vender energía a otros individuos de la red). Por otra parte, la creciente adopción de tecnologías monofásicas de baja emisión de carbono (Low Carbon Technologies (LCT)) conectadas [75] (por ejemplo, los sistemas fotovoltaicos y los vehículos eléctricos en la red de BT) complica aún más el problema del desequilibrio de fases. El aumento de las LCT provoca que el desequilibrio de fase sea incierto y poco previsible, y por tanto, dificulta su gestión. Además, el creciente incremento de dispositivos electrónicos y del número de viviendas conectadas, ha dado lugar a redes de distribución de BT densamente pobladas.

Hasta ahora, las causas más habituales por las que ocurren los desequilibrios entre fases son:

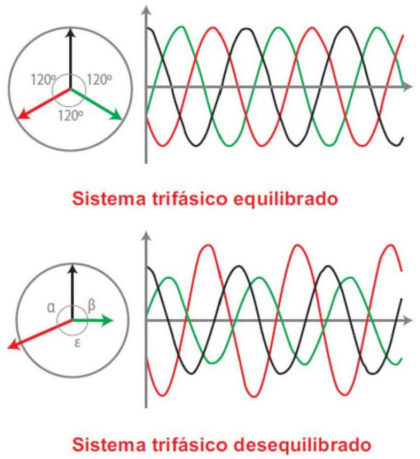
- El desconocimiento de la topología de red [22],
- El desconocimiento de los patrones de consumo en BT (especialmente los consumos residenciales que son impredecibles y que presentan grandes incertidumbres y comportamientos aleatorios) hace imposible una asignación óptima a las diferentes fases de Línea que conduce a desequilibrios sistemáticos (Systematic ImBalance) entre fases [76], [77] y principalmente,
- El reparto desigual de cargas entre las tres fases.

Actualmente, esta asignación de usuarios a fases en el embarrado del cuadro de control se realiza de forma subjetiva, y muchas veces dependiendo del criterio del técnico. En consecuencia, la distribución de consumos entre las tres fases no es uniforme [78], [79]. Se consideran diferentes escenarios [76]: una fase sobre-saturada (fase con mayor corriente) respecto al resto de fases, una fase sub-saturada (fase con menor corriente) o dos fases desiguales (dos fases sobre-saturadas o dos fases sub-saturadas). Con todo ello, surge la necesidad de crear herramientas que, basadas en datos, constituyan sistemas expertos para el apoyo a la toma de decisiones.

En este contexto, la situación española es privilegiada con respecto a otros países europeos. Aunque en 2019 la implantación de contadores era

escasa, la lectura de los contadores era mensual, las comunicaciones requeridas eran limitadas, los protocolos de comunicaciones eran simples y basados en radio, desde entonces, gracias al protocolo de comunicación PoweRline Intelligent Metering Evolution (PRIME) [80] y a la extensa implantación de los contadores inteligentes, ahora en el año 2022 se proporciona una frecuencia y volumen de datos mucho mayor. De esta manera se genera una red hiperconectada con comunicaciones bidireccionales que permiten no solo conocer el estado de la red sino también actuar sobre ella permitiendo la telegestión. Donde la curva de carga, que se obtiene a través de la lectura de los consumos de los contadores, permite conocer ciertos patrones y/o comportamientos de consumo, permitiendo caracterizar estadísticamente a los usuarios. La frecuencia por defecto es horaria, e incluso bajo determinadas circunstancias puede llegar a ser cada 5 minutos o menor. Además la explotación del conocimiento implícito en dichas curvas de carga permite a los operadores del sistemas de distribución (DSO) plantearse la creación de herramientas basadas en IA: para una mejor comprensión de cómo utilizan la energía los consumidores finales, para detectar posibles problemas en la red de manera anticipada, para ofrecer capacidades de planificación al sistema eléctrico, para estimar el proceso de automatización de la distribución e incluso para conocer qué energía se consumirá en un futuro próximo. A fin de cuentas, ayudar a transformar esos datos en información útil y luego en conocimiento [75].

Figura 2.1: Sistema trifásico equilibrado y Sistema trifásico desequilibrado. Fuente: [81]



Precisamente, por este valor añadido que se genera a través de los datos de los contadores inteligentes, es posible conocer la topología de la red de BT [82], y con ello afrontar el desafío del equilibrado de consumos entre fases de Línea. Se dice que una Línea de distribución de Baja Tensión (BT) está desequilibrada cuando entre las tres fases que la forman, o bien las

magnitudes de corriente (o tensión) no son las mismas o sus ángulos de fase no son de 120^{o} entre sí [83]. La Figura 2.1 muestra un ejemplo de Línea equilibrada y otro ejemplo de Línea desequilibrada. Los desequilibrios de tensión y de corriente provocan desequilibrios de la potencia de las fases [76], [84], lo que significa que los flujos de potencia en las tres fases no son iguales entre si [77]. Esto acarrea ciertas consecuencias negativas como: utilización ineficiente de la red, mayor corriente por el neutro, mayores pérdidas en transformadores, cables y líneas aéreas, tensiones fase-neutro fuera de los límites legales, dificultad para mantener la regulación de la tensión, propagación de desequilibrios de tensión y corriente, problemas de caídas de tensión y en última instancia, baja calidad de servicio [27], [76], [77], [85], [86]. En estos casos de desequilibrio de fases, conocer a través de los datos de los contadores el valor de las curvas de carga permite caracterizar estadísticamente los desequilibrios y tomar medidas correctivas, con los consiguientes beneficios:

- Reducir la corriente por el conductor neutro evitando inestabilidades y caídas de tensión de la red, y la intervención física del operario en el punto de conflicto [87],
- Minimizar las pérdidas técnicas [77], [88]-[90],
- Favorecer la integración de renovables [27],
- Disponer de una infraestructura de red optimizada, al aumentar la capacidad de los CTs de distribución. Así se evita el despliegue de CTs innecesarios y, en consecuencia, se reducen los costes de distribución y se utilizan de forma óptima los recursos de la red [91],
- Menor complejidad de supervisión de la red de distribución de BT, debido a una tendencia hacia el aplanamiento de la curva agregada de la demanda y a la reducción de las inestabilidades generadas por la presencia de corrientes elevadas en el conductor neutro; y
- Garantizar una mejor calidad de servicio (QoS) [92].

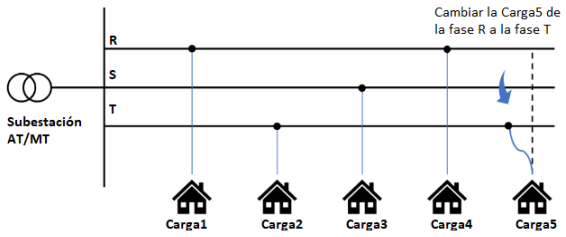
El reto que se expone en este capítulo consiste en formular el equilibrado de consumos entre fases como un problema de optimización combinatoria NP-hard [93] (Non deterministic Polynomial time Hardness problems). En este caso, la complejidad del problema crece exponencialmente con el número de cargas a asignar. Por lo tanto y como ejemplo, para el caso de N cargas que deben conectarse a 3 posibles conexiones entre fases (R, S, T) - 3 conexiones monofásicas - con un total de 3^N combinaciones (posibles soluciones). Si el número de consumidores fuese 100, el número de combinaciones posibles asciende a 3^{100}. Este número de posibles soluciones dificulta la evaluación de todas las posibles combinaciones mediante técnicas de búsqueda exhaustivos tradicionales. El capítulo se plantean diferentes opciones de optimización, desde un algoritmo Greedy o voraz, hasta métodos de computación evolutiva a partir de diferentes métricas. Además, para evaluar la calidad de las soluciones propuestas, estas se analizan en la realidad una red de distribución de Baja Tensión parcialmente monitorizada.

2.1.2. Trabajos relacionados

En la literatura el problema del equilibrado de fases en redes de distribución se plantea desde la perspectiva de la reconfiguración del CT a nivel de sistema eléctrico y desde la perspectiva del intercambio de cargas o contadores inteligentes entre fases [94].

Por una parte, la reconfiguración del CT a nivel de sistema [87], [95] consiste en el cambio de la estructura topológica de los sistemas de distribución mediante la alteración del estado abierto/cerrado de los tramos monofásicos y de los interruptores de enlace. Se basa en la transferencia de cargas de las secciones de red más cargadas a las menos cargadas. Se trata de realizar modificaciones de sistemas físicos a un nivel muy alto de la red eléctrica, en la que cualquier cambio ha de realizarse con el consentimiento del gestor de la red de transporte de energía eléctrica, por ello este estudio escapa del enfoque de este capítulo.

Figura 2.2: Intercambio de fases en red de distribución de BT. Fuente: [75]



Por otra parte, el intercambio de contadores entre fases en la red de distribución de BT [76] (en inglés, Phase Swapping) consiste en determinar a qué fase debe conectarse cada contador para reducir el desequilibrio entre las fases, es decir consiste en permutar cargas de una fase a otra hasta lograr un sistema equilibrado. Se trata de conectar cargas desde las fases más sobre-saturadas a fases sub-saturadas, tal y como se muestra en la Figura 2.2. Desde una perspectiva de operabilidad esta estrategia es más sencilla de abordar que la reconfiguración física del CT [96]. Debido a la naturaleza combinatoria de la búsqueda óptima del equilibrado de fases, la solución tiende a resolverse mediante técnicas heurísticas o meta-heurísticas.

En relación a la aplicación de técnicas heurísticas, los autores de [89] formularon el balanceo de fases mediante un algoritmo de búsqueda por recurrencia hacia atrás (Heuristic Backtracking Search algorithm), que minimizaba el desequilibrio de fases del sistema de forma que se reducía la corriente del neutro. Se obtenía un sistema fiable capaz de disminuír la probabilidad de actuación de los dispositivos de protección de sobrecorriente en el neutro. En [27] se comparaban seis algoritmos meta-heurísticos (Algoritmo Genético, Recocido Simulado, Algoritmo Greedy, Algoritmo de Búsqueda Exhaustiva, Algoritmo de búsqueda hacia atrás y Programación Dinámica) para solucionar el problema del desequilibrio de fases monofásicas. Los autores identificaron que los dos factores que más afectaban

al resultado eran: la cuantificación del desequilibrio entre fases (Phasing Unbalance Index) y el número de permutaciones de cargas entre fases (cada permutación tiene ciertas implicaciones económicas). En este trabajo, los autores llegaron a la conclusión de que técnicas heurísticas de búsqueda combinatoria como algoritmos genéticos o de recocido simulado eran superados por métodos de programación dinámica cuando se trataba de encontrar una solución capaz de balancear el tiempo de convergencia y la calidad de la solución proporcionada.

En relación a la aplicación de técnicas meta-heurísticas, en [94] se formuló el problema de equilibrado de fases como un problema de flujo de potencia óptimo, en el que las variables de control eran las opciones de reajuste en cada uno de los nodos candidatos. El objetivo era minimizar la suma ponderada del coste de desequilibrio, el coste de desequilibrio entre fases monofásicas y el porcentaje de la diferencia monofásica. Se adoptó el algoritmo de Recocido Simulado (Simulated Annealing algorithm) para resolver el problema no lineal de los flujos de potencia. En [97] se desarrolló un enfoque de equilibrio de fases que re-asignaba de forma óptima a los clientes monofásicos mediante un algoritmo genético discreto (Discrete Genetic algorithm) como método meta-heurístico. Se consideraba como función objetivo, la minimización de los índices de desequilibrio de tensiones y corrientes y las pérdidas de red, siendo la fase de cada cliente una variable de decisión. En [98] se presentó una versión discreta del algoritmo de Búsqueda de Vórtices (Discrete Vortex Search algorithm) para determinar las conexiones a fase a través de la magnitud y el ángulo de tensión por nodo. La función objetivo se centraba en la minimización de pérdidas de potencia activa. En [99] se desarrolló una solución para el equilibrado de fases a través de la optimización de colonias de hormigas (Ant Colony Optimization algorithm), donde se minimizó la pérdida de energía para los sistemas de distribución en un intervalo de 24 horas. Los autores conseguían reducir la corriente del neutro y el precio del consumo de energía por parte del usuario final. En [100] se desarrolló un nuevo modelo de reconfiguración con dos objetivos de optimización, minimizar el factor de desequilibrio trifásico y el tiempo de computación, a partir del análisis del flujo de potencia trifásico. Se diseñó un algoritmo evolutivo diferencial multi-objetivo que, mediante un método de discriminación de la conectividad de red, eliminaba soluciones inviables y favoreciendo el tiempo de convergencia.

Asimismo, respecto al cuándo se realiza la reconfiguración de las redes en [100], los autores diferencian entre reconfiguración estática (se realiza en periodos de paradas programadas) y reconfiguración dinámica (se realiza en periodos de funcionamiento). En [101] se indicó una estrategia de reconfiguración dinámica, que ajustaba automáticamente las cargas en cualquier momento del día, cuando el grado de desequilibrio superaba un determinado umbral. El estudio se centraba en la reducción del espacio de búsqueda empleando un método de análisis de sensibilidad (basado en el factor de desbalanceo de tensión o Voltage Unbalance Factor) para mejorar los tiempos de convergencia. El esquema planteado proponía un algoritmo heurístico poblacional de salto de rana barajado modificado (Modified

Tabla 2.1: Resumen de algoritmos presentados para el Equilibrado de Fases

Ref.	Método / Algoritmo	Función Objetivo a Minimizar
[94]	Recocido Simulado	Costes operativos
[99]	Optimización de colonia de hormigas	Pérdidas de energía
[27]	Programación Dinámica	Desequilibrio del sistema
[89]	Búsqueda de recurrencia hacia atrás	Desequilibrio del sistema
[100]	Diferencial Evolutivo	Desequilibrio del sistema
[101]	Optimización de salto de rana	Desequilibrio del sistema
[97]	Genético	Pérdidas de potencia
[98]	Búsqueda de vórtices	Pérdidas de potencia

Shuffled Frog Leaping Algorithm). Las métricas propuestas proponían minimizar la pérdida de potencia y la corriente por el conductor neutro.

La Tabla 2.1 recoge las referencias anteriores que resuelven el desafío del equilibrado de fases y sus correspondientes funciones objetivo, a riesgo de presentar las siguientes carencias:

- No aplican conocimiento de la conectividad de red ni información de la topología por desconocimiento;
- No consideran la información de la curva de carga, pese a que proporciona un conocimiento más profundo y fino de los patrones de carga;
- No han tenido en cuenta que las cargas pueden estar conectadas a dos fases cuando se trata de sistemas bifásicos. La mayoría de las referencias solo contempla conexiones monofásicas. La conexión bifásica incrementa la complejidad computacional del problema de intercambio de fases al aumentar el número de posibilidades con las que se puede conectar la carga. Es decir, en conexiones monofásicas la complejidad es de 3^N para $\{R, S, T\}$ y aumenta hasta 6^N para conexiones bifásicas para $\{RT, R, RS, S, ST, T\}$ en N cargas.

La contribución de esta tesis para afrontar el problema de la conectividad óptima de cargas a fase en redes de distribución de BT se centra en:

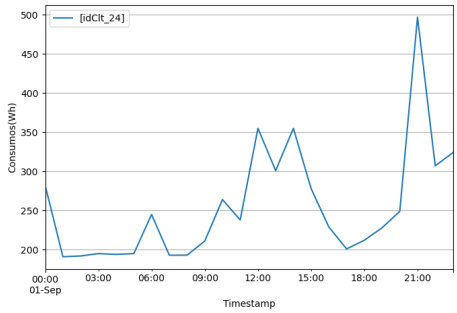
- La definición de una codificación específica que considere conexiones bifásicas, siento está una casuística más realista que considerar únicamente conexiones monofásicas.
- La definición de algoritmos que proporcionen un buen balanceo entre coste computacional y la calidad de las soluciones aportadas.
- El establecimiento de métricas específicas que cuantifiquen los desequilibrios.

- Estudiar todos los puntos anteriores en la realidad de un CT de BT real, empleando valores de cargas aportados por los contadores inteligentes.

2.2. Enfoque propuesto

En general, el término curva de carga [102]-[105] es un término comúnmente utilizado para referirse a la serie temporal del consumo de energía eléctrica. La curva de carga supone el último eslabón de sensado en la red eléctrica permitiendo conocer cómo consume cada usuario hora a hora y la constituyen medidas en kWh. Mientras que un determinado consumidor residencial de electricidad dispone de una pluralidad de cargas individuales eléctricas desagregadas (frigorífico, horno, lavadora, etc.), el contador inteligente proporciona la suma agregada de las curvas individuales. La Figura 2.3 muestra una curva de carga de 24 horas de un consumidor cualquiera. Dicha curva de carga indica, la variación del consumo durante las diferentes horas del día y el área bajo la curva de carga representa la energía total consumida en este día.

Figura 2.3: Ejemplo de curva de carga horaria de consumo individual para el periodo de 1 día



Para los estudios de este capítulo, se utiliza información de N curvas de carga que pertenecen a:

- Una Línea trifásica ubicada en entorno urbano.
  Se trata de una Línea urbana que da servicio a 152 consumidores eléctricos en BT. De los cuales, se descartan 50 consumidores por no disponer de contador inteligente tele-gestionado que proporcione la trazabilidad horaria de los consumos. Con ello, el estudio se realiza a partir de la información de los N=102 consumidores restantes, de los que sí se conoce su conectividad. Se sabe qué línea les da servicio, qué tipo de consumidores son, si son consumidores monofásicos o bifásicos y a qué fase/s concreta/s está/n conectado/s cada uno de ellos. Esta información se muestra con más detalle en la Tabla 2.2: se trata de 82 consumidores residenciales (56 % son consumidores bifásicos y el resto son monofásicos), 15 comerciales y 5 industriales

Tabla 2.2: Conectividad de consumidores por fase y por sector en Línea urbana

	Conectividad bifásica			Conectividad monofásica			Número Consumidores
	RS	RT	TS	R	S	T	N= 102
Residencial	29 %	27 %	0 %	12 %	27 %	5 %	82
Servicios	20 %	20 %	0 %	0 %	0 %	60 %	15
Industria	0 %	0 %	0 %	20 %	0 %	80 %	5

Tabla 2.3: Conectividad de consumidores por fase y por sector en Línea rural

	Conect. monofásica			Núm.Consumidores
	R	S	T	N= 14
Residencial	33 %	33 %	33 %	3
Servicios	22 %	33 %	44 %	9
Industria	50 %	50 %	0 %	2

(ambos 100 % consumidores monofásicos). La profundidad histórica es de datos horarios de un año.

- Una Línea trifásica ubicada en entorno rural.
  Se trata de una Línea rural que da servicio a N=14 consumidores, todos ellos con contadores inteligentes tele-gestionados. Todos son consumidores monofásicos, como se indica en la Tabla 2.3.

A partir de un escenario real con datos reales proporcionados por contadores inteligentes, este capítulo demuestra la virtud e importancia de la aplicación de técnicas basadas en Ciencia de Datos. Se trata de contrastar las soluciones proporcionadas por los esquemas que a continuación se proponen, con el despliegue actual de la Línea de distribución urbana y la Línea rural de BT. Como contribución adicional respecto al estado del arte se amplía el problema al caso de conexiones bifásicas, no contemplado en la literatura, y se sugiere una codificación al problema que incluya dicha casuística. Los resultados proporcionados demuestran la reducción de pérdidas técnicas, al considerar las soluciones sugeridas por los esquemas conformados respecto a las topologías de red existentes hoy en día.

2.2.1. Especificación del problema

Los algoritmos de optimización planteados se basan en tres claves de diseño: (i) la codificación del alfabeto del problema capaz de indicar soluciones específicas por medio de una formulación matemática; (ii) los esquemas o algoritmos capaces de balancear la capacidad de exploración frente a la explotación y lograr el equilibrio entre calidad de la solución y coste

computacional; y (iii) el análisis de las funciones de aptitud o métricas específicas (desde las utilizadas para medir la calidad de una señal de radio - aplicada a las series temporales proporcionadas por los contadores- hasta métricas específicas de negocio).

Codificación específica al contexto

En la mayoría de los problemas de optimización, los algoritmos trabajan con codificaciones de parámetros y no directamente con los parámetros en sí. La información ha de transcribirse de manera óptima para que el algoritmo sea capaz de procesarlo de manera adecuada. El esquema de codificación elegido juega un papel importante y debe estar muy relacionado con el dominio del problema [106]. Para la codificación del problema del desequilibrio de fases en redes de Distribución de BT, considerando conexiones monofásicas y bifásicas, se sigue un tipo de codificación octal $x_n^k \in \{1, 2, 3, 4, 5, 6\}$. Se utilizan números enteros como variables de decisión con objeto de evitar el uso excesivo de vectores binarios de gran tamaño, reducir el tamaño del espacio de solución y, en consecuencia, el tiempo de procesamiento. Esta codificación equivale al alfabeto que representa las conexiones mono/bifásicas RT, R, RS, S, ST, T. Cada uno de los valores de la codificación representa respectivamente, una de las posibles opciones de conexión. El orden establecido para las conexiones da sentido a los valores vecinos, suavizando las transiciones entre valores del alfabeto donde sólo se produce de una a otra el cambio de una fase. Este hecho es especialmente relevante sobre todo en métodos de búsqueda diferenciales. Es decir, el mapeo de la codificación a la secuencia de fases está diseñado para que cuando los valores deban ajustarse a cualquier otro valor en su proximidad, los cambios en el mapeo de fase no sean drásticos, ya que la codificación vecina en el alfabeto corresponde a fases con, al menos, una fase en común con la del valor original. Por ejemplo, si la búsqueda de nuevas codificaciones se realiza a partir de la fase R, la nueva búsqueda de valores vecinos ha de ser en la fase RT y en la fase RS, que siempre serán más semejante que fases como ST. En este caso, se presupone que, en las conexiones bifásicas, la corriente se distribuye en el mismo porcentaje por cada fase individual.

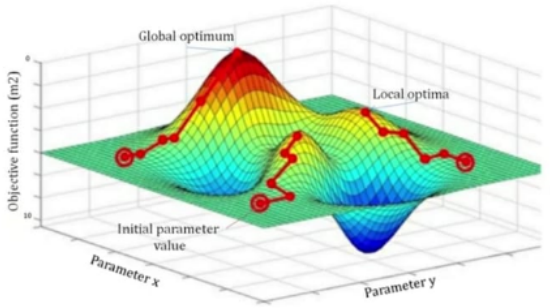
Algoritmos de optimización

A continuación se explica en qué consiste la heurística voraz o Greedy, y la meta-heurísticos de los Algoritmos Genéticos (GA) y de Búsqueda en Armonía (HS) como procedimientos iterativos que basados en información contextual (algoritmo Greedy) o en información generalista (GA y HS) guían una heurística subordinada de forma inteligente sobre diversos operadores para combinar adecuadamente dos estrategias:

- Explorar adecuadamente el espacio de búsqueda para evitar caer en regiones sub-óptimas.
- Explotar adecuadamente el espacio de búsqueda para evitar buscar innecesariamente en dichas regiones sub-óptimas.

Algoritmo voraz o Greedy La heurística voraz o Greedy [107] es una estrategia de búsqueda consistente en elegir la opción óptima en cada iteración como estrategia de búsqueda de la solución global óptima. Pese a tomar decisiones parciales óptimas, este esquema no garantiza llegar al óptimo global y entonces, se limita a proponer óptimos locales, ver Figura 2.4. Este esquema está condicionado por su inicialización, siendo en su mayor parte algoritmos deterministas [108].

Figura 2.4: Ilustración de óptimo local y óptimo global



Un algoritmo Greedy, normalmente se basa en un proceso iterativo definido en base a las métricas consideradas y las decisiones a tomar. Se trata de un algoritmo fácil de explicar a un interlocutor no experto, favoreciendo la explicabilidad mediante la fácil interpretación de las reglas que definen la búsqueda de soluciones. Además, siempre proporciona un resultado candidato. Por ello es uno de algoritmos más utilizados en aplicaciones reales para dimensiones del espacio solución no excesivamente grandes [108].

En este caso, se trata de conceptualizar un algoritmo Greedy para resolver, mediante el conocimiento de las curvas de carga horarias de los consumidores, cuál es la mejor asignación de consumidores a fase de Línea de distribución de BT. El algoritmo Greedy propuesto se resume en los siguientes pasos:

1. Inicialización. Se selecciona un horizonte temporal que contenga el conjunto de comportamientos históricos que se quiere optimizar y se ordenan los consumos agregados de los consumidores, de mayor a menor. Se asignan los tres consumidores de mayor consumo agregado a cada una de las fases R, S y T, respectivamente.

2. Se calcula el valor de la función de aptitud o fitness encargada de medir la calidad de las soluciones aportadas. La especificidad en el diseño de la función de aptitud es fundamental a la hora de poder diferenciar soluciones en términos de calidad cuantitativa. Lo ideal es lograr unicidad entre los valores de esa función de aptitud y los diferentes candidatos a la solución. La sección 2.2.1.C detalla las funciones de aptitud definidas para la optimización de la conectividad en redes de BT.

3. Actualización. Se calcula el valor de la función de aptitud correspondiente a asignar el siguiente consumidor procedente de la lista ordenada a cada una de las posibles opciones de conectividad. El consumidor evaluado se asocia a la codificación de la fase que obtiene un menor valor de la función aptitud. El algoritmo itera tantas veces como consumidores hay para asignar.

4. Criterio de parada. El algoritmo finaliza cuando todos los consumidores están asignados a una de las 6 codificaciones de fase.

En la búsqueda del equilibrio entre calidad de la solución frente a complejidad (exploración vs explotación), los algoritmos Greedy normalmente proporcionan buenas soluciones en espacios de búsqueda acotados o no muy complejos. Se tratan de algoritmos deterministas, cuando la inicialización, y las variables del análisis permanecen constantes. Sin embargo, también sufren de algunas limitaciones que se reflejan en la Tabla 2.4 y que incluyen:

- Por esta dependencia de la inicialización suelen verse abocados a terminar en regiones sub-óptimas al favorecer la explotación frente a la exploración en el proceso de búsqueda. Y es por este motivo, por el que no siempre alcanzan la solución óptima global, al quedar limitado su rendimiento a regiones óptimas locales.

- Una vez tomada una decisión, nunca reconsideran decisiones tomadas anteriormente.

- No existe una base teórica establecida, sino que se trata de algoritmos generados a medida del problema a solucionar.

Si por cualquier motivo se tuviese que añadir un consumidor (no considerado en el histórico inicial de cargas), al partir de la solución inicial propuesta por el algoritmo Greedy, asignar deterministicamente un nuevo consumidor, no provocaría cambios bruscos en la conectividad.

Tabla 2.4: Ventajas y desventajas de los algoritmos Greedy

Ventajas	Desventajas
Fácilmente implementable.	No siempre alcanzan la solución óptima global.
Requieren pocos recursos computacionales. Rápida ejecución.	No se reconsideran decisiones tomadas anteriormente.
Fácilmente interpretables.	No existe una base teórica establecida: algoritmos a medida.
Ejecución rápida	Capacidad de búsqueda limitada: capacidad explotativa y poca capacidad explorativa.

Algoritmo Genético Los algoritmos Genéticos son una técnica de optimización inspirada en el principio de la selección natural de Darwin. Se trata de una técnica habitual en problemas de Ciencia de Datos introducida en la década de 1970. Los algoritmos Genéticos son algoritmos basados

en búsquedas basadas en los conceptos biológicos de selección natural y genética.

En los Algoritmos Genéticos, se trabaja a partir de una población de posibles soluciones a un problema dado. Estas soluciones se someten a operaciones genéticas como "Mutación", "Cruce" y "Selección", produciendo nuevos hijos y el proceso se repite a lo largo de varias generaciones. A cada individuo, o solución candidata, se le asigna un valor de aptitud o fitness basado en un valor de función de evaluación y a los individuos más aptos se les da una mayor oportunidad de aparearse y producir más individuos más aptos. Esto está en línea con la teoría Darwiniana de la "supervivencia de los más fuertes". De esta manera, los mejores individuos o soluciones evolucionan a lo largo de generaciones, hasta que se llega a un criterio de terminación o parada.

La terminología básica en los Algoritmos Genéticos incluye los siguientes términos:

- Población, es un subconjunto de todas las soluciones posibles, codificadas, para un problema dado. Es decir, es un conjunto de individuos donde cada individuo es una solución al problema que se quiere resolver.
- Cromosomas, es una de las soluciones al problema en cuestión.
- Gen, es la posición de un elemento de un cromosoma. Los individuos de una población se caracterizan por un conjunto de parámetros, variables, denominados genes. Básicamente, los genes se unen en una cadena para formar un cromosoma o solución.
- Alelo, es el valor que toma un gen para un cromosoma en particular.

El proceso de selección natural comienza con la selección de los individuos más aptos de una población. Producen descendencia que hereda las características de los padres y que será añadida a la siguiente generación. Si los padres tienen un mejor estado físico, sus hijos serán mejores que los padres y tendrán una mejor oportunidad de sobrevivir. Esto proceso continúa iterando y al final, la generación con los individuos más aptos será encontrada. De forma esquemática, el procedimiento se resume en los siguientes pasos [109]:

- Inicialización. El proceso comienza con un conjunto de individuos o población. Cada individuo es una solución al problema que se desea resolver. Un individuo se caracteriza por un conjunto de parámetros, variables, conocidos como genes. Los genes se unen en una cadena para formar un cromosoma o solución. En un Algoritmo Genético, el conjunto de genes de un individuo se representa mediante una cadena, en términos de un alfabeto. Normalmente se utilizan valores binarios, cadena de 1 y 0. Decimos que codificamos los genes en un cromosoma.
- Se calcula el valor de la función aptitud o fitness. En este paso, el algoritmo debe ser capaz de determinar qué es lo que hace que una

solución sea más adecuada que otra solución. Esto se determina por la función fitness o de aptitud. El objetivo de la función es evaluar la viabilidad genética de las soluciones dentro de la población, colocando a aquellos con los rasgos genéticos más viables, favorables y superiores a la cabeza de la lista. Su función principal es desempeñar el papel de diferenciador en la población que separa a los individuos más fuertes de los más débiles. La sección 2.2.1.C detalla las funciones de aptitud definidas.

- Operadores genéticos o conjunto de operaciones a realizar sobre los individuos de una población.

  - Selección. Durante cada generación sucesiva, una proporción de la población existente es seleccionada para criar una nueva generación. Las soluciones individuales se seleccionan a través de un proceso basado en la función fitness, en el que las soluciones más adecuadas suelen ser las que más se seleccionan. Ciertos métodos de selección califican la idoneidad de cada solución y seleccionan las mejores soluciones. Otros métodos clasifican solo una muestra aleatoria de la población, ya que este proceso puede llevar mucho tiempo. La mayoría de las funciones son estocásticas y están diseñadas para que se seleccione una pequeña proporción de soluciones menos adecuadas. Esto ayuda a mantener la diversidad de la población a un nivel elevado, lo que impide la convergencia prematura hacia soluciones deficientes o sub-óptimas. Los métodos de selección más populares incluyen la selección de la rueda, de la ruleta y la selección de torneos [106]. Se busca seleccionar cromosomas padres que producirán descendientes.
  - Cruce. En términos biológicos, el cruce no es más que la reproducción de los cromosomas padres más adecuados. El tipo más común es el cruce de punto único. En el cruce de un solo punto, se elige un lugar en el que se intercambian los alelos restantes de un progenitor al otro. Los nuevos individuos toman una sección del cromosoma de cada padre. El punto en el que se rompe el cromosoma depende del punto de cruce seleccionado al azar. Este método en particular se llama cruce de punto único porque solo existe un punto de cruce. Sin embargo, el cruce no siempre ocurre. A veces, en base a un conjunto de probabilidades, no se produce ningún cruce y los padres se copian directamente a la nueva población.
  - Mutación. Después de la Selección y el Cruce, se tiene una nueva población llena de individuos. Algunos se copian directamente y otros se producen por el cruce. Para asegurar que los individuos no son todos exactamente iguales, se permite una pequeña posibilidad de mutación. Se hace un bucle a través de todos los alelos de todos los individuos, y si ese alelo es seleccionado para

la mutación, puede cambiarlo por una pequeña cantidad o reemplazarlo con un nuevo valor. La mutación es vital para asegurar la diversidad genética en la población.

- Criterio de parada. Este proceso generacional se repite hasta que se alcanza una condición de terminación. Las condiciones comunes de terminación son: (i) se encuentra una solución que satisface los criterios mínimos y/o (ii) número fijo de generaciones alcanzadas. La solución óptima es aquella cuyo valor de fitness es el mayor y por tanto, ha alcanzado un nivel tal que las iteraciones sucesivas ya no producen mejores resultados.

Como cualquier técnica científica, los Algoritmos Genéticos también sufren de algunas limitaciones. Estos incluyen:

- No apto para todos los problemas, especialmente los problemas que son simples y para los que se dispone de información contextual. Son algoritmo de propósito general y agnósticos al problema a resolver y además funcionan bien en alfabetos discretos de cardinalidad baja.
- El valor del módulo fitness se calcula repetidamente, lo que puede resultar costoso para algunos problemas. Se requiere de máquinas computacionales con alta potencia y capacidad de procesamiento, ya que se necesita mucho espacio para almacenar el aumento de la población cuando se ejecuta. Además de tiempos de procesamiento elevados.
- Al ser soluciones estocástico, no hay garantías sobre el óptimo o la calidad de la solución.
- Si no se implementan correctamente, los Algoritmos Genéticos pueden no converger hacia la solución óptima.

Algoritmo Harmony Search Los algoritmos Harmony Search pertenecen al grupo de algoritmos poblacionales. Se trata de un método de búsqueda estocástica iterativa basado en cómo una orquesta de jazz improvisa la nota musical más adecuada de una armonía. Fue introducido en 2009 por Dr. Prof. Zong Woo Geem [110].

En los Algoritmos HS, se trabaja a partir de un conjunto de armonías de posibles soluciones a un problema dado. Estas soluciones se someten a operaciones de improvisación como "probabilidad de consideración de búsqueda de armonía" (en inglés, Harmony Search Considering Rate, HMCR $\in \mathbb{R}[0,1]$), "probabilidad de ajuste de tono" (en inglés, Pitch Adjustment Rate, PAR $\in \mathbb{R}[0,1]$]) y "tasa de selección aleatoria" (en inglés, Random Selection Rate, RSR $\in \mathbb{R}[0,1]$]), produciendo nuevas armonías y el proceso se repite iterativamente. A cada armonía, se le asigna un valor de aptitud o fitness basado en un valor de función de evaluación y a las armonías más estéticas se les da una mayor oportunidad de improvisación y producir nuevas armonías aún más estéticas. De esta manera,

las mejores armonías o soluciones evolucionan a lo largo de las sucesivas improvisaciones, hasta que se llega a un criterio de terminación o parada.

La terminología básica en los Algoritmos HS incluye los siguientes términos:

- Memoria de Armonías (HM, de sus siglas en inglés, Harmony Memory), es el subconjunto de todas las armonías pasadas utilizadas anteriormente, es decir, el conjunto de soluciones potenciales.
- Armonía, es el conjunto de notas y una de las soluciones al problema en cuestión.
- Nota, es un elemento de una armonía. Básicamente, las notas se unen en una cadena para formar una armonía o solución.
- Medida de calidad estética o función de aptitud o fitness.

El proceso de composición musical que siguen los músicos de jazz se basa en tocar notas y comprobar su calidad estética recordando experiencias tocadas previamente (memoria) y una pequeña improvisación de tonos al azar. De forma esquemática, el procedimiento se resume en los siguientes pasos [110]:

1. Inicialización. Este primer paso sólo se considera en la primera iteración. Se inicializa el valor de cada nota de todas las armonías incluidas en HM aleatoriamente de entre los posibles valores definidos en el alfabeto. La asignación es aleatoria y no asume ningún conocimiento a priori sobre el espacio de soluciones.
2. Improvisación. Este proceso iterativo define el conjunto de reglas que permiten modificar el conjunto de melodía de HM de manera que se balancee el carácter explorativo frente al carácter explotativo de la búsqueda mediante los siguientes operadores de improvisación:

   a) Primer paso de generación de una nueva armonía. Cada nota es generada de forma independiente, de manera que HMCR es la probabilidad de que dicha nota sea obtenida de HM. En caso contrario, se genera aleatoriamente. Por ejemplo, si HMCR = 0,9 implica que el 90 % de las nuevas notas se extraen del conjunto de armonías en cada iteración. Este operador refuerza el carácter explotativo del algoritmo.

   En base a estudios anteriores [111]-[113] se adopta una variación lineal del parámetro HMCR a lo largo de las iteraciones para mejorar el balance entre exploración y explotación en el proceso de búsqueda y con ello, mejorar el rendimiento. La formulación matemática para la actualización dinámica del operador HMCR, es:

$$\mathrm{HMCR}(iter) = \mathrm{HMCR}_{min} + (\mathrm{HMCR}_{max} - \mathrm{HMCR}_{min}) \cdot \xi(iter) \tag{2.1}$$

, donde HMCR_{min} y HMCR_{max} son los valores mínimo y máximo de la tasa de ajuste y $\text{HMCR}(iter)$ es la tasa de ajuste del paso para la iteración $iter \in \{1, 2, \ldots, niter\}$. El coeficiente $\xi(iter)$ se calcula en función de la iteración $iter$ y del número máximo de iteraciones $numMaxIter$ como:

$$\xi(iter) = \frac{iter}{numMaxIter} \tag{2.2}$$

La estrategia consiste en partir de valores de HMCR altos, puesto que HM se inicializa aleatoriamente y por lo tanto, conviene potenciar esa aleatoriedad, tendiendo a lo largo del proceso iterativo a valores de HMCR menores para prevenir la convergencia prematura en regiones sub-óptimas. Valores mayores de HMCR explotan el contenido de información almacenado en memoria HM, provocando una minimización de los valores de las notas dentro de las melodías y así favorecer la rápida convergencia en las regiones del espacio explorado de mejores valores de la métrica almacenada.

Este enfoque prioriza valores altos del operador HMCR en las primeras iteraciones, permitiendo acelerar la velocidad de convergencia, y valores bajos en las últimas iteraciones, lo que le permite escapar de óptimos locales. El objetivo es priorizar la capacidad de exploración del proceso de búsqueda en lugar de su comportamiento de explotación.

b) Cada nota es ajustada en tono en base al valor PAR que establece la probabilidad de que el valor de una nota dada se ajuste a uno de sus valores vecinos dentro del alfabeto de notas. Por ejemplo, PAR = 0,2 implica que el 20 % de las nuevas notas se extraen del conjunto de notas vecinas (inferiores o superiores con igual probabilidad) definidas en el alfabeto de números enteros en el que están codificadas. Este operador refuerza el carácter de búsqueda local del algoritmo. Al igual que HMCR este valor se ajusta dinámicamente, de manera lineal, a lo largo del proceso iterativo. En este caso se parte de un valor inicial menor para incrementarlo a medida que se itera. Este hecho favorece la búsqueda local en las últimas iteraciones una vez la búsqueda global ha explorado lo suficiente el espacio de búsqueda. La formulación matemática para la actualización dinámica del operador PAR es análoga a la Expresión (2.1).

c) Algunas notas son reemplazadas por otras de forma aleatoria, con una probabilidad que viene determinada por el valor RSR. Se trata de un operador de carácter explorativo que intenta, mediante la aleatoriedad, aliviar la convergencia prematura a través del salto hacia regiones sub-óptimas no exploradas hasta ese momento. La nueva armonía se compara con la peor de HM, y si es mejor que ésta última la sustituye. Se modifica linealmente.

3. Actualización de HM. Se evalúan las HM melodías improvisadas empleando la función de aptitud. Se concatenan, junto con los valores de la métrica, con las melodías de la iteración anterior. Se ordenan los 2 * HM valores de métrica de menor a mayor (proceso de minimización) y se eligen las HM melodías para la siguiente interacción.

4. Criterio de parada. Este proceso de improvisación se repite hasta ue se alcanza un determinado número máximo de improvisaciones. De lo contrario se vuelve al punto 2.

Funciones de aptitud, métricas de contexto

El diseño o la elección de la métrica para evaluar la calidad de las soluciones en una técnica de optimización es un factor de diseño clave. Básicamente lo ideal es identificar una métrica unívoca que sea capaz de asignar valores semejantes de métrica a soluciones de semejante calidad y diferenciable respecto a soluciones realmente diferentes en términos de dominio. En este caso, las métricas de contexto de dominio que se contemplan son: (i) suma cuadrática de residuos de consumos energéticos entre fases (retorno por el neutro), (ii) factor de cresta y (iii) coeficiente de simultaneidad. Cualquiera de estas tres funciones de aptitud puede calcularse estadísticamente utilizando las curvas de carga diarias. A continuación, se muestra la formulación matemática asociada a cada métrica.

Suma cuadrática de residuos de consumos energéticos entre fases. A la suma de las cargas individuales asociadas a una fase, se le denomina carga agregada y se representa como una serie temporal de valores discretos. El residuo entre pares de fases se define como la distancia euclídea entre las series temporales discretas de cargas agregadas. La fórmula matemática de la función de aptitud basada en el residuo entre fases es la siguiente:

$$Q(\mathbf{x},t) = \sum_{(\phi,\theta)\in\mathcal{P}} \sqrt{\left(\sum_{n=1}^{N} \mathbb{I}(x_n,\phi)E_n(t)\right)^2 - \left(\sum_{n'=1}^{N} \mathbb{I}(x_{n'},\theta)E_{n'}(t)\right)^2} \tag{2.3}$$

,donde $\mathcal{P} \triangleq \{(R,S),(S,T),(R,T)\}$, $E_n(t)$ es el consumo energético de la carga n en tiempo t, $\mathbf{x} \triangleq \{x_n\}_{n=1}^{N}$ es el mapeo de las cargas a las fases tal que, suponiendo una red eléctrica monofásica/bifásica, $x_n \in \{RT, R, RS, S, ST, T\}\ \forall n \in \{1,\ldots,N\}$, y $\mathbb{I}(x_n,\phi)$ es una función indicadora que toma valor 1 si $\phi \in x_n$ y 0 en caso contrario. Es decir, cada fase está representada por la suma de los consumos energéticos horarios de las cargas asignadas a esa fase en un horizontal temporal histórico (preestablecido por el experto en distribución). Esta métrica logra su valor óptimo con la minimización de la función de aptitud. Cuando $Q(\mathbf{x},t) = 0$, el retorno de energía por el conductor neutro es nulo y la energía entre fases

está perfectamente equilibrada en el momento t, es decir a valores más bajos de $Q(\mathbf{x}, t)$ mejor equilibrio de las cargas x en el instante t.

Factor de Cresta (CF). Métrica comúnmente utilizada en técnicas de procesamiento de señal en sistemas de telecomunicaciones [114]. El Factor de Cresta (Factor de Cresta (CF)) es una medida relacionada con la forma de la onda, y que formula la relación entre los valores de pico y el valor efectivo, véase Figura 2.5. La formulación matemática es:

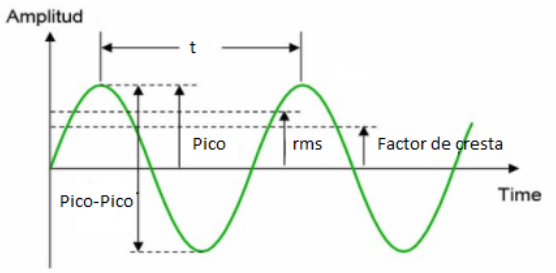
$$CF = \sum_{(\phi \in \mathcal{P})} \left(\sum_{d=1}^{D} CF(\phi)_d \right) \tag{2.4}$$

, donde

$$CF(\phi)_d = \frac{|P^t|_{max}}{P^t_{rms}} \tag{2.5}$$

, donde el CF es la amplitud máxima de pico de una onda, dividida por su valor cuadrático medio (rms) en el horizonte temporal considerado.

Figura 2.5: Factor de Cresta

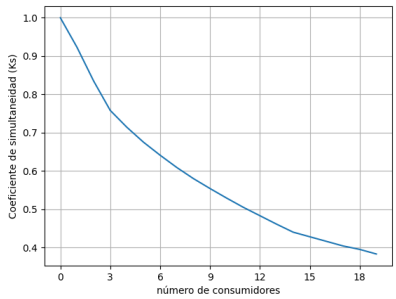


En el estudio del equilibrado de cargas entre fases el valor de CF se calcula diariamente, siendo el valor de la métrica CF global el sumatorio de los CFs individuales de cada fase. Donde cada CF individual se calcula como el valor máximo de la carga agregada de cada fase, dividido entre el valor rms de cada una de las cargas agregadas para $\forall t \in \{1, \ldots, 24\}$ horas del día. Con esta métrica asumimos que la capacidad real de un sistema eléctrico es la diferencia entre su capacidad máxima de consumo y el consumo punto a punto cuantificando la diferencia estadística: pico (máximo consumo) - valle (mínimo consumo). Este valor de CF global indica la medida de la capacidad de la red de distribución para aceptar nuevas cargas que no simultaneen el momento de aparición de sus picos máximos. Es decir, CF permite cuantificar el aplanamiento de la curva de la demanda agregada por fase a través de la complementariedad de las cargas individuales de los diferentes consumidores. Bajo este contexto, un Factor de Cresta $CF = 1$ implica que el valor pico coincide con el valor cuadrático medio y por lo tanto, implica una curva de carga agregada aplanada (las amplitudes rms y pico son iguales). Por el contrario, cuanto

más extremos son los valores de CF, mayor es la diferencia entre picos y valles.

Coeficiente de Simultaneidad (K_s). El coeficiente de simultaneidad de una red eléctrica es un coeficiente estadístico utilizado en el contexto de las redes inteligentes, que mide la afección de los picos máximos de potencia individuales sobre el perfil de carga agregado. Históricamente este coeficiente se calculaba en base a estudios estadísticos a escala espacial y temporal realizadas por las empresas distribuidoras de energía, de manera que consideraban un valor de Ks constante solo dependiente del número de consumidores de la red. Calculado de esta forma (y sin considerar, por desconocimiento, la información de las curvas de carga de los contadores inteligentes) consumidores con diferentes comportamientos o diferente distribución temporal de sus curvas de carga, indicaban el mismo coeficiente de simultaneidad. La Figura 2.6 muestra la distribución típica de valores de K_s (eje y) proporcionados por una distribuidora sólo dependiente del número de consumidores (eje x). Se observa que a medida que el número de usuario aumenta, decrece el valor de K_s, debido a la afección individual del pico máximo diario de su curva de carga sobre la forma de la curva agregada resultante de todos los consumidores asignados a la misma fase de carga.

Figura 2.6: Estimación estadística del Coeficiente de Simultaneidad para consumidores domésticos



En [115] los autores presentan la primera metodología estadística para calcular el coeficiente de simultaneidad en tiempo real K_s basado en valores agregados de carga. Los autores defienden la idoneidad de calcularlo persé para cada circunstancia en base a los consumidores asignados a cada subestación y la curva de carga de dichos consumidores. La fórmula matemática que define K_s según [115] es:

$$K_s = \sum_{(\phi \in \mathcal{P})} \left(\sum_{d=1}^{D} \frac{P_{max}^{t_d}}{\sum_{i=1}^{N} P_{max,i}^{t_d}} \right) \tag{2.6}$$

, donde $P_{max}^{t_d}$ es la potencia máxima agregada en un intervalo de tiempo dado (en este caso diario t, $\forall t \in \{1, \ldots, 24\}$ horas del día), N es el número

de consumidores a considerar (asociados a la misma fase de línea), D es el número de días y $P^{t_d}_{max,i}$ es la potencia máxima asociada a las carga individuales diarias. Si el consumo individual máximo no coincide con el máximo de la señal agregada, el valor de K_s tiende a disminuir.

2.2.2. Escenarios

El reequilibrado de fases se contempla bajo dos supuestos: desequilibrios severos entre fases o conexión de nuevos clientes [27]. Los desequilibrios severos, implican la completa asignación de consumidores a fases (presuponiendo que todos los consumidores han de ser asignados), bien (i) cuando se trata de un espacio de soluciones amplio, principalmente en entornos urbanos o porque bien (ii) cuando se trata de un espacio de soluciones de tamaño menor, principalmente en entornos rurales. En ambos casos el objetivo es aplicar las metodologías, codificaciones y métricas propuestas anteriormente para optimizar la conectividad real aportada por la empresa distribuidora de electricidad.

Pero, en que la realidad de las empresas distribuidoras la asignación completa de cargas a menudo no minimiza costes. Por ello, y con objeto de minimizar el número de cambios de cargas entre fases, además, se propone el (iii) estudio del coste de los cambios y (iv) la conexión de nuevos clientes a un sistema de desequilibrado de fases ya existente.

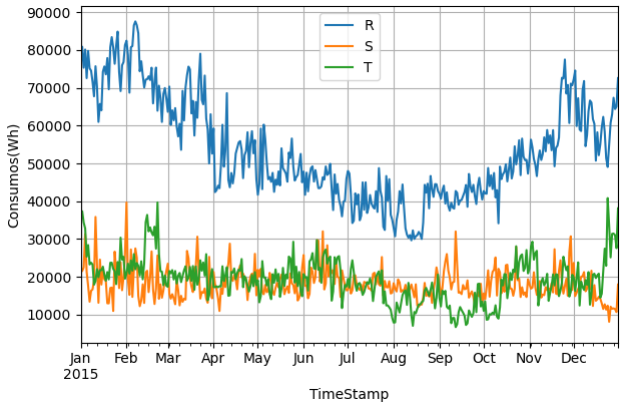
E1. Problema de optimización en entorno rural

Dada una Línea trifásica en un entorno rural, este estudio consiste en evaluar el desequilibrio entre fases y proponer una solución para la asignación de 14 consumidores a fases monofásicas (R, S, T) y/o bifásicas (RS, RT, ST). Se trata de un problema de optimización combinatoria complejo NP-hard de 6^{14} opciones.

La Figura 2.7 representa el escenario original de cargas agregadas por fase con el que opera la empresa distribuidora eléctrica en el contexto del entorno rural. Cada fase se representa por su carga agregada a través de una serie temporal de valores discretos. En este caso, se trata de un sistema desequilibrado, con la fase R sobre-saturada en comparación con el resto de fases. Con un valor medio de $53kW/dia$, la fase con mayor consumo agregado es la fase R, mientras que los consumos de fase S y fase T son más parecidos entre si, con valores medios de $18kW/dia$ y $19kW/dia$ respectivamente.

En este escenario de búsqueda en entornos rurales en espacio de solución reducido, el estudio demuestra que la convergencia a una solución sub-óptima adecuada aun no siendo la mejor, es la más óptima en el binomio: calidad y tiempo de ejecución. De esta forma, la utilización del Factor de Cresta como función de aptitud aporta calidad suficiente en el proceso de búsqueda, y la implementación del algoritmo Greedy aporta velocidad en el tiempo de computación. El resto de soluciones candidatas son computacionalmente más costosa en relación a la mejora de la calidad que obtienen. La Figura 2.9 muestra la calidad de las soluciones propuestas y la Tabla 2.5 recoge los valores de las métricas y los tiempos de ejecución

Figura 2.7: Carga agregada en cada fase de la distribuidora eléctrica en entorno rural



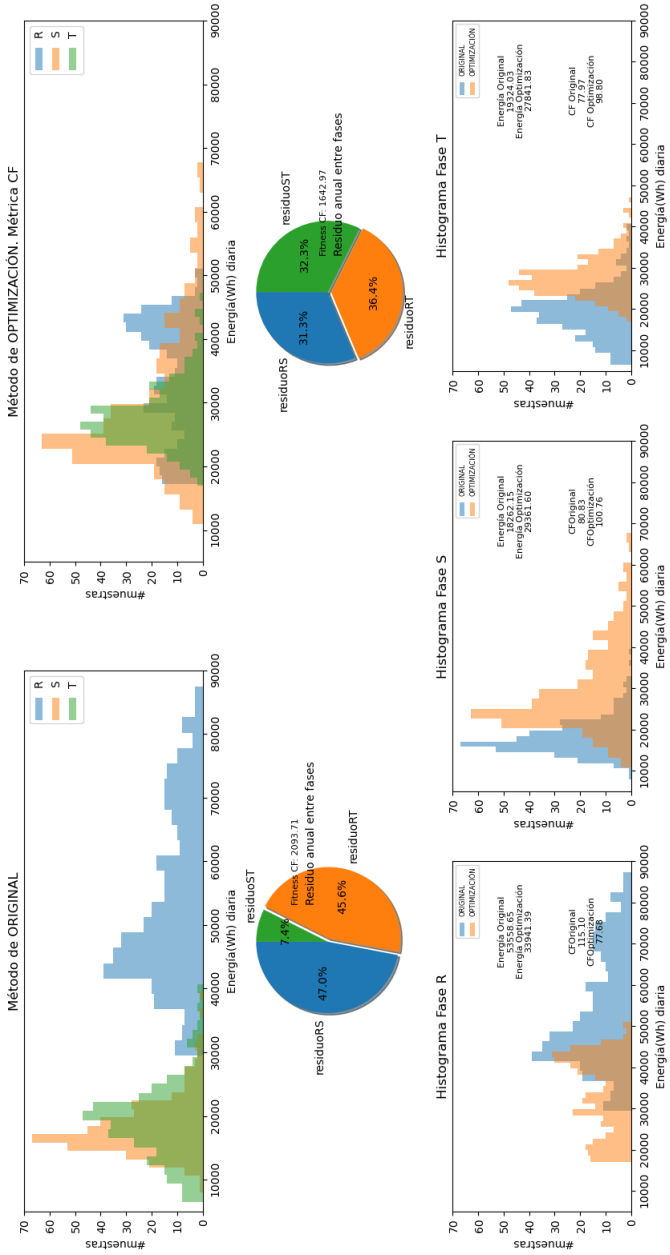
de dichas soluciones candidatas. En resumen, (i) cualquiera de los tres experimentos basados en el algoritmo Greedy es computacionalmente menos costoso que el resto de implementaciones, que (ii) la métrica en el algoritmo Greedy mejora sustancialmente con la introducción de la métrica Factor de Cresta basada en contexto y (iii) que el éxito de Greedy está estrechamente relacionado con el reducido tamaño del espacio de soluciones (la afección a caer en mínimos locales es menor).

Tabla 2.5: Tiempo de ejecución y valores de la función de aptitud de las soluciones candidatas

	Distribuidora eléctrica	Suma cuadrática y Algoritmo GA	Suma cuadrática y Algoritmo HS	Suma cuadrática y Algoritmo Greedy	Coef. Simultaneidad y Algoritmo Greedy	Factor de Cresta y Algoritmo Greedy
Fitness	1552k	487k	439k	594k	467k	442k
Tiempo de ejecución	-	8h20'	4h30'	0h0'3"	0h0'3"	0h0'3"

La mejor solución candidata, basada en la combinación de la métrica del Factor de Cresta con la implementación del algoritmo Greedy, es la que se muestra en la Figura 2.8. La figura muestra la comparativa entre la solución propuesta por el método de optimización sugerido en este capítulo y la solución de la distribuidora eléctrica. Los operadores de improvisación utilizados en los algoritmos meta-heurísticos GA y HS han sido seleccionados bajo el criterio que se explica en el siguiente apartado.

Figura 2.8: Solución candidata óptima: algoritmo de optimización Greedy y Factor de Cresta como función de aptitud



E2. Problema de optimización en entorno urbano

Dada una Línea trifásica en un entorno urbano, este estudio consiste en evaluar el desequilibrio entre fases y proponer una solución para la asignación de 102 consumidores a conexiones monofásicas (R, S, T) y/o bifásicas (RS, RT, ST). Se trata de un problema de optimización combinatoria complejo NP-hard de 6^{102} opciones.

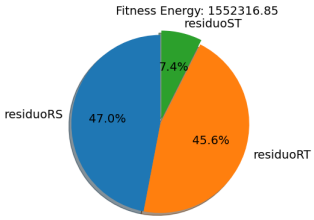
La Figura 2.10 representa el escenario original de cargas agregadas a nivel de fase en la operativa real de la empresa distribuidora de electricidad en el contexto de un sistema desequilibrado de entorno urbano. Ante la situación de desequilibrio entre fases que se muestra en la figura, el reto es encontrar aquella combinación de cargas a fase que minimice la función de aptitud, mediante algoritmos de optimización combinatoria. El equilibrio perfecto se logra cuando se logra el porcentaje equitativo del consumo de las cargas para cada fase, es decir se asigna el $33,3\,\%$ de las cargas de consumo para cada fase.

En este escenario de búsqueda en entornos urbanos en espacio de solución amplio, el estudio compara la distribución porcentual a fases de la compañía eléctrica frente a las soluciones candidatas proporcionadas en base a la codificación, algoritmos y funciones de aptitud propuestos anteriormente. Para comparar la complejidad computacional frente a la calidad de las soluciones propuestas se propone la Tabla 2.6 que aporta información del tiempo de ejecución y el valor de la métrica de la solución candidata de cada esquema. La mejor solución candidata es cuando se utiliza la métrica de contexto CF y el algoritmo HS, y que se muestra en la Figura 2.12. El estudio muestra que la debilidad de los algoritmos Greedy por caer en regiones sub-óptimas aumenta a medida que aumenta el tamaño del espacio de soluciones, y que el comportamiento de los algoritmos meta-heurísticos GA y HS mejora con el tamaño del espacio de soluciones. En ambos se observa que, para este problema NP-hard, llegar a la solución candidata es costoso en términos de tiempo (y recursos); ambos invierten varios días en la búsqueda de los operadores idóneos (6 días y 5 días respectivamente) y varias horas en el cálculo de la mejor solución candidata (14 horas y 11 horas respectivamente). Pero el carácter explotativo añadido al carácter diferencial de HS (a través del operador PAR) logra ajustar mejor el balance en términos, no solo de computación sino también en términos de calidad de la solución. Por todo ello, se puede concluir que la mejor solución candidata - en el binomio calidad y tiempo de computación - es la se obtiene a partir del algoritmo HS. La mejor melodía propuesta por el algoritmo HS es la que se muestra en la Figura 2.11.

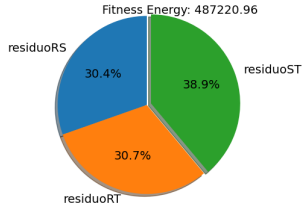
	Distribuidora eléctrica	Factor de Cresta y Algoritmo Greedy	Factor de Cresta y Algoritmo GA	Factor de Cresta y Algoritmo HS
Fitness	622k	590k	400k	380k
Tiempo de ejecución mejor solución 50MC	-	0h 0 3'	0d 14h42'	0d 11h30'
Tiempo optimización operadores 50MC	-	-	6d 20h10'	5d 14h29'

Tabla 2.6: Tiempo de ejecución y valores de la función de aptitud de las soluciones candidaras en entorno urbano

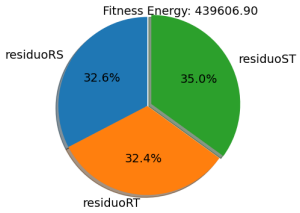
Figura 2.9: Comparativa entre diferentes algoritmos y funciones de aptitud para el equilibrado de fases en entorno rural



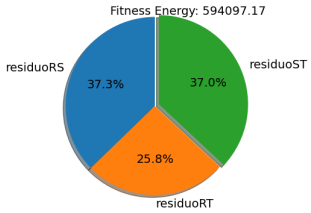
(a) Situación original de la distribuidora eléctrica



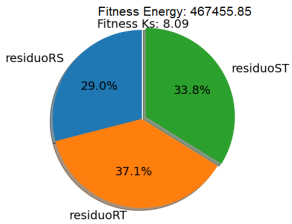
(b) Suma cuadrática de residuos con alg. GA



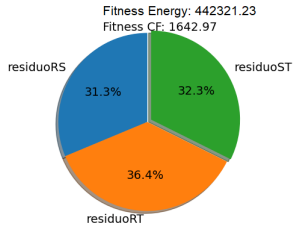
(c) Suma cuadrática de residuos con alg. HS



(d) Suma cuadrática de residuos con alg. Greedy

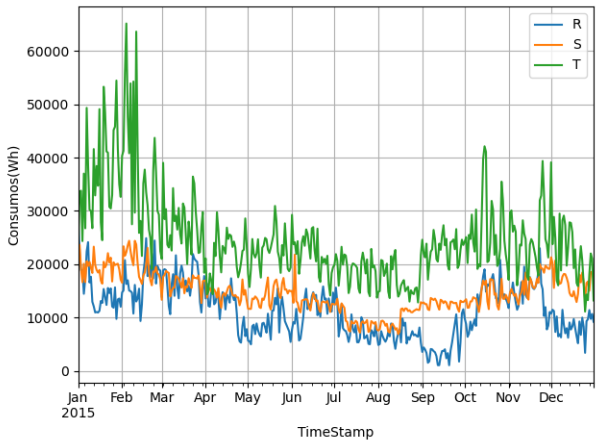


(e) Coeficiente de Simultaneidad con alg. Greedy



(f) Factor de Cresta con algoritmo Greedy

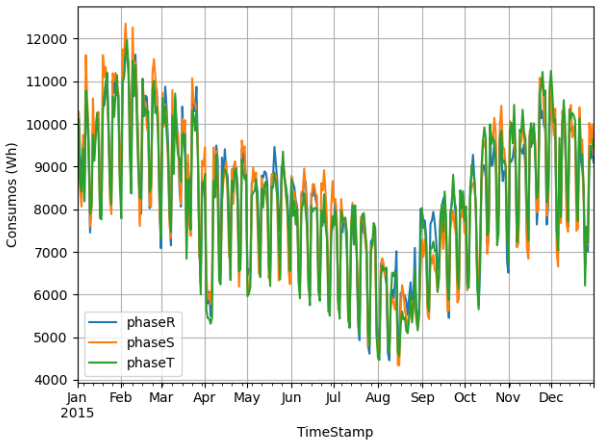
Figura 2.10: Carga agregada por fases de la distribuidora eléctrica en entorno urbano



Optimización de los hiper-parámetros que controlan los operadores de improvisación para GA y HS. Los algoritmos meta-heurísticos GA y HS, ambos esquemas tienen un conjunto de operadores de improvisación que han de ser ajustados para, dada una formulación matemática del problema, definida mediante una codificación acorde a la caustica de los datos tratados, balancear su carácter explorativo frente al explotativo. De esto depende un buen balance entre su complejidad computacional y la calidad de las soluciones aportadas. Debido al carácter estocástico de HS y GA, dicho proceso de elección óptima de los hiper-parámetros necesita proveer de resultados estadísticos. Para ello se ejecutan realizaciones consecutivas mediante MonteCarlos con diferentes inicializaciones, para elegir los hiper-parámetros que estadísticamente proporcionan los mejores valores de las métricas consideradas. Es decir, ambos esquemas, necesitan realizar determinado número de iteraciones Monte Carlo para, con la suficiente estadística, determinar el valor de los operadores de improvisación que generan la mejor solución candidata. Para garantizar una comparación justa en términos de establecer la misma complejidad computacional se establecen un mismo número de iteraciones máximas (evaluaciones métricas) para HS y GA. El conjunto de valores a optimizar para los diferentes operadores de improvisación (Mutación y Cruce para GA y HMCR, PAR y RSR para HS) se explican a continuación.

En el caso de GA, la Figura 2.13 muestra los valores medios (sobre los 50 Monte Carlo) de la métrica obtenidos en el estudio de optimización de los operadores de improvisación (Cruce y Mutación). El eje y contiene los valores medios (iteraciones Monte Carlo) de la métrica y en el eje x diferentes combinaciones de los operadores de Cruce y Mutación. La zona sombreada indica la variación estocástica del algoritmo en los 50 Monte Carlo, y representa el valor medio respecto a la desviación estándar. La parametrización resultante prima el carácter explotativo de búsqueda, a

Figura 2.11: Conectividad urbana propuesta por alg. HS



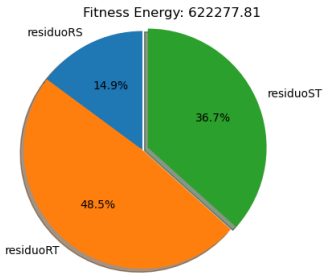
través de los siguientes valores:

- Operador $CRUCE = 0{,}7$. El algoritmo explota el 70 % del espacio de búsqueda de los progenitores y apenas introduce nuevos alelos.
- Operador $MUTACION = 0{,}2$. El algoritmo combina el 20 % de la información de los progenitores para descubrir nuevas áreas prometedoras.

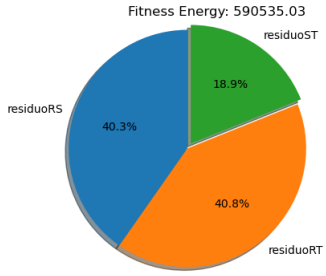
En el caso de HS, la Figura 2.14 muestra los valores medios (sobre los 50 Monte Carlo) de la métrica obtenidos en el estudio de optimización de los operadores de improvisación (HMCR-PAR-RSR). El eje y contiene los valores medios (iteraciones Monte Carlo) de la métrica y en el eje x diferentes combinaciones de los operadores de HMCR, PAR y RSR, respectivamente. La zona sombreada indica la variación estocástica del algoritmo en los 50 Monte Carlo (desviación estándar de la métrica sobre esos 50 Monte Carlo), y se representa a través de una línea de trazo continuo. Al igual que GA, esta parametrización prima el carácter explotativo de búsqueda con un valor alto de HMCR, una búsqueda local diferencial media (PAR) y baja aleatoriedad (RSR) a través de los siguiente valores de inicialización de los operadores:

- Operador $HMCR = 0{,}9$. El algoritmo elige un valor de la memoria de armonías (HM) con un 90 % de probabilidad de explotar la información contenida en la memoria HM para la generación de nuevas melodías.
- Operador $PAR = 0{,}2$ y Operador $RSR = 0{,}1$. El algoritmo ajusta la capacidad de seleccionar tono a partir de tonos adyacentes con un 20 % de probabilidad, y la capacidad de aleatoriedad en la selección de tonos es del 10 % de probabilidad.

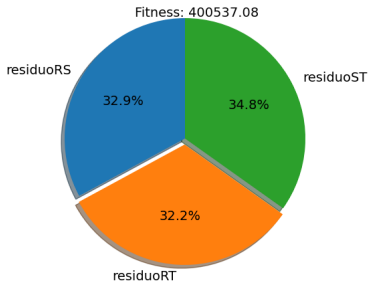
Figura 2.12: Comparativa entre diferentes algoritmos y Factor de Cresta como función de aptitud para el equilibrado de fases en entorno urbano



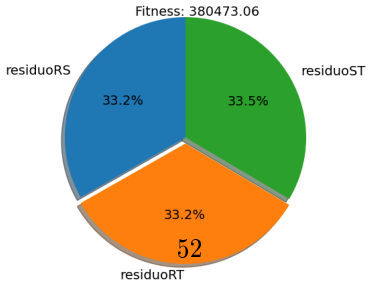
(a) Asignación original de la compañía eléctrica



(b) Asignación mediante alg. Greedy

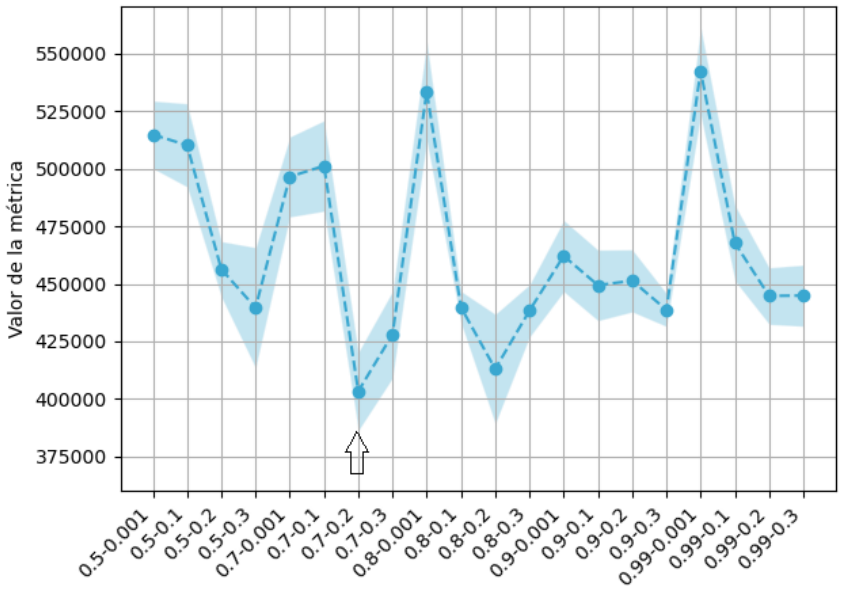


(c) Asignación mediante alg. genético



(d) Asignación mediante alg. HS

Figura 2.13: Optimización de los operadores de improvisación [CRUCE-MUTACIÓN] para algoritmo GA



Además mencionar que, en el caso del algoritmo HS y acorde a la sugerencia de autores como [111], [112] los valores de HMCR y PAR no se han preconfigurado como estáticos y como tal, se actualizan dinámicamente en las sucesivas iteraciones.

En ambos casos, para GA y HS, la inicialización se fija constante. Aplicar una inicialización aleatoria implicaría aplicar numerosos Monte Carlo y en cada caso una búsqueda del rango de soluciones mucho más amplia, condicionando el rendimiento computacional del algoritmo en términos de tiempo necesario en la búsqueda de la solución óptima, inviable en un proceso en producción.

Los valores de los operadores que logran el mejor equilibrio entre el comportamiento exploratorio y el explotativo de los esquemas propuestos, se enumeran en la Tabla 2.7. En ambos casos, predomina el carácter explotativo a través de los operadores CRUCE/HMCR sobre otros valores más exploratorios como MUTACIÓN/PAR,RSR en GA y HS, respectivamente. La tabla también recoge los valores de inicio y fin para los operadores HMCR y PAR. En la tabla no se incluye el algoritmo Greedy ya que carece de operadores de improvisación.

E3. Estudio del coste de cambios de cargas entre fases

La disposición de reequilibrar una Línea es una decisión planificada con antelación (puntual de mantenimiento o de restauración), manual y costoso en términos económicos y de recursos humanos. Una vez que se reequilibra

Figura 2.14: Optimización de los operadores de improvisación [HMCR-PAR-RSR] para algoritmo HS

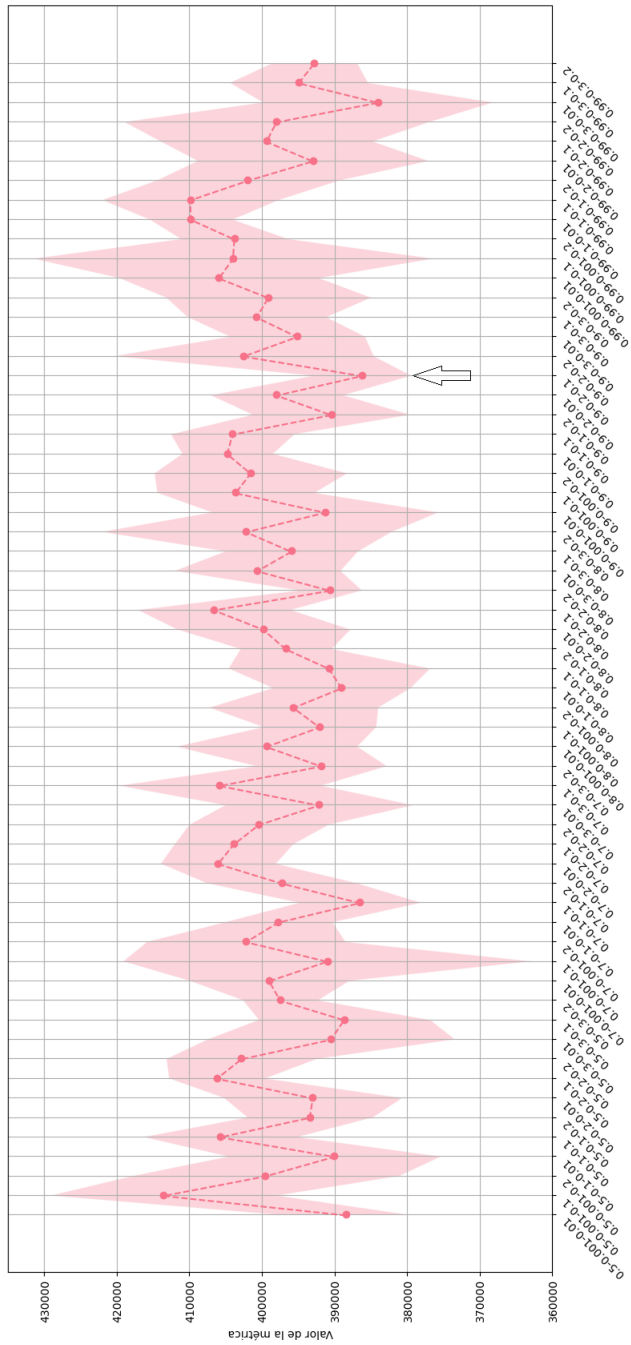


Tabla 2.7: Especificación de los operadores que parametrizan los algoritmos HS y GA

Operador HS	Valor	Operador GA	Valor
Tamaño HS	50	Tamaño de Población	100
HMCR	[0.9-0.4]	Ratio de Cruce	0.7
PAR	[0.2-0.5]	Ratio de Mutación	0.2
RSR	0.1		
Iteracciones	200	Generación máx.	100

una Línea, al principio, estará equilibrada, pero con el tiempo (aparición, desaparición o cambios de comportamiento de las cargas) la Línea tenderá a desequilibrarse nuevamente [27]. En cada nueva actualización puntual para el reequilibro de fases, lo ideal es proporcionar una lista ordenada de cargas sugeridas para su reasignación, con vistas a lograr el equilibrio de la Línea con el menor número posible de reasignaciones. El primer cambio de carga de la lista ordenada es aquella carga con mayor afección sobre la métrica a minimizar; el segundo valor tendrá menor afección sobre la métrica que el anterior valor, y así sucesivamente. Este sistema, le permite a la compañía eléctrica conocer la afección de cada reasignación sobre la métrica y además, monetizar el beneficio de reequilibrar la Línea mediante el cálculo de las pérdidas de energía en el sistema eléctrico de distribución trifásico.

El concepto pérdida de energía, indica la diferencia entre energía generada y energía vendida o bien, energía que se pierde en los diferentes equipos y elementos de la red y no es aprovechada. Dentro del sistema energético (Generación, Transporte y Distribución), las pérdidas más cuantiosas corresponden a las redes de BT; y se asumen, no pueden ser eliminadas, únicamente pueden reducirse a través de la optimización de la red [116]. En [86] los autores, indican que, en el caso extremo donde la carga monofásica sea del mismo factor de potencia que una carga trifásica equilibrada, las pérdidas técnicas en la Línea se multiplica por seis respecto a la que se pierde cuando la misma potencia se entrega a una carga trifásica equilibrada. La formulación matemática de los autores para la potencia perdida en un sistema desequilibrado se expresa como la corriente que se pierde por el neutro y la potencia que se pierde en las tres fases. Por una parte, la corriente del neutro se calcula como el vector suma de las tres corrientes de fase [117]; cuando las redes presentan un gran desequilibrio en las cargas, pueden circular corrientes mayores por el neutro que por las fases [118]. Y, por otra parte, las pérdidas de las fases se calculan con la ecuación [86]:

$$P_{pe} = 2R\frac{P^2}{U^2 \cos\alpha^2} \tag{2.7}$$

, donde P_pe indica la potencia que se pierde en las fases, R es la resistencia equivalente de la Línea, U es el valor de la tensión entre fases, y $\cos\alpha$ es el factor de potencia o ángulo entre el vector de fase y tensión.

Para los cálculos que se realizan a continuación, se presupone que:

- la resistencia equivalente (R) de la línea eléctrica y que depende de la longitud de la línea es $R = 0{,}64\Omega/km$.
- el coste de las pérdidas es de $coste = 0{,}06€/kWh$.
- el factor de potencia es $\cos\alpha = 0{,}9$.

Con objeto de mostrar la validez del método, se comparan las pérdidas técnicas obtenidas a través del método propuesto y las pérdidas técnicas de la empresa eléctrica en el instante del cambio. Se cuantifica cómo afectan las sucesivas re-asignaciones o cambios, en términos económicos y de potencia con 10 cambios: donde $Cambio\#0$ se corresponde a la situación original que la distribuidora tiene en la asignación de consumidores a fases para dicha Línea de distribución. Para el estudio se ha considerado la Línea rural, de la Sección anterior, con 14 consumidores monofásicos residenciales a 230 voltios, con tres fases desequilibradas un 38 % (entre la fase sobre-saturada, sobre la fase sub-saturada: $((consumo_{fase-sobresaturada} - consumo_{fase-subsaturada}) * 100/consumo_{fases})$). El valor de la métrica (suma cuadrática de residuos) originalmente es $1552k$, las pérdidas técnicas son de $363{,}58W/kmh$, lo que equivale a 21,81 €/km. Para los siguientes cambios, se aplica un método Greedy y se estudia la afección de cada cambio o nueva re-asignación en relación a la suma cuadrática de residuos entre fases y a lo que supone en términos de pérdidas técnicas (W/kmh) y en pérdidas económicas (€/km) cada cambio.

La Figura 2.15 muestra el valor del residuo en las sucesivas iteraciones. Los mejores/menores resultados se obtienen con 7, 9, y 8 cambios, respectivamente. En la Figura, también se muestra el porcentaje de asignación a fase de cada una de las mejores soluciones propuestas, donde la mejor solución (7 cambios) aporta residuos porcentuales entre fases de $38,7\,\%$, $32,1\,\%$ y $29,2\,\%$ respectivamente. Si calculamos el porcentaje de desequilibrio (en términos de la fase sobre-saturada y la fase subsaturada) es de $9,5\,\%$ de desequilibrio entre fase (cuando anteriormente era 38 %).

La Tabla 2.8 cuantifican las pérdidas técnicas y económicas con el número de cambios realizados. Por ejemplo, al realizar el 7 cambio, el valor de la métrica es $434k$, el coste de ese cambio es de 16,9 €/km, el coste total incluyendo todos los cambios anteriores es 144,92 €/km y 284,75 W/kmh.

Figura 2.15: Valores de la métrica y residuos de las tres soluciones con menor valor de métrica

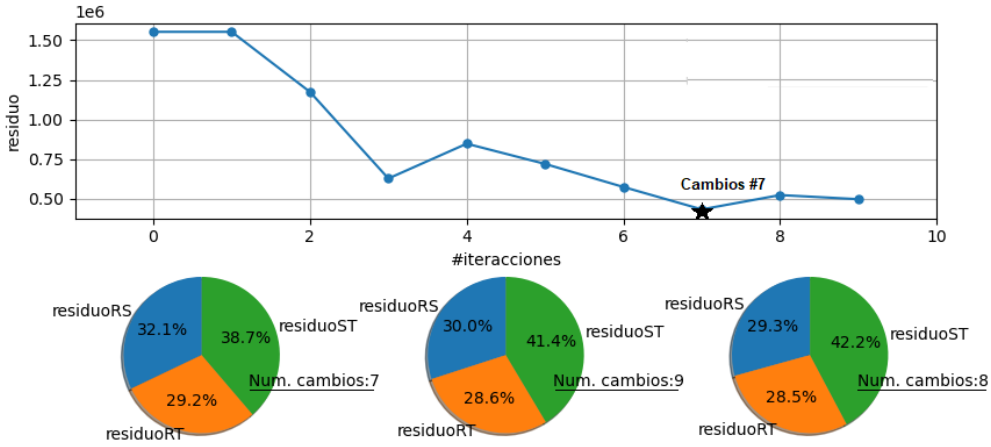


Tabla 2.8: Residuos y pérdidas técnicas, estimados en base al número de cambios

# CAMBIOS	PÉRDIDAS W/kmh	PÉRDIDAS €/km	PÉRDIDAS AGG €/km	RESIDUO
0	363.58	21.81	21.81	1552316.85
1	323.54	19.41	41.22	1552316.85
2	287.66	17.26	58.48	1173713.31
3	298.48	17.91	76.39	628752.53
4	290.57	17.43	93.82	847971.52
5	285.10	17.11	110.93	720481.81
6	281.67	16.90	127.83	575291.81
7	284.75	17.09	144.92	**434335,59**
8	283.77	17.03	161.95	**524470,22**
9	281.65	16.90	178.85	**498540,72**

E4. Conexión de nuevos clientes

Ante la necesidad de tener que incorporar una nueva conexión de carga al sistema, la carga se incorporará en aquella fase que minimice el valor de la métrica propuesta. Por tanto, la carga se añadirá en aquella fase que con cuya aportación el desequilibrio sea lo menor posible.

El enfoque más adecuado es que inicialmente, se recurra a técnicas meta-heurísticas (algoritmo HS) - dado su carácter exploratorio- para conseguir la mejor configuración inicial. Y para continuar, mediante técnicas heurísticas (algoritmo Greedy) que favorecen el carácter explotativo y que buscan la mejor solución en cada momento, asignar la carga a aquella fase que mejore la métrica.

2.3. Conclusiones

En este capítulo se ha propuesto una herramienta de telegestión, que permite el envío de indicaciones en remoto, a la Línea, al CT y/o a la Caja General de Protección, para actuar en tiempo y forma ante un imprevisto como es el desequilibrio de fases. Estos desequilibrios conllevan problemas relacionados con la eficiencia y seguridad en la red de distribución de BT: fases saturadas y propensas a sobrecargas, sobrecalentamiento de cables, mayores pérdidas técnicas, peor calidad de la electricidad suministrada a clientes debido a una tensión fluctuante, elevado riesgo de que se produzcan fallos la red y demás.

Desde un punto de vista funcional, el problema de equilibrado de fases (en entornos urbanos y/o rurales) y la conexión de nuevos consumidores, son problemas de optimización combinatoria NP-hard, que permiten varias soluciones plausibles dependiendo de la complejidad del espacio de solución, el tiempo de procesamiento y la calidad de la solución, que se esté dispuesto a asumir. Para los casos de reasignaciones completas de cargas a fases:

- En el caso de un entorno urbano, donde el espacio de soluciones es grande. Se ha optado por el desarrollo de un método meta-heurístico a través del algoritmo Harmony Search. Esta familia de algoritmos es capaz de explorar el espacio de soluciones iterativamente, a través de mecanismos de búsqueda explorativa y explotativa, generando soluciones de calidad a un coste computacional menor que lo que haría un algoritmo de búsqueda exhaustiva, pero mayor que lo haría un algoritmo basado en método gradiente como Greedy. Pese a ello, y debido a la inicialización aleatoria del algoritmo, es necesario para realizar 50 iteraciones Monte Carlo. Estas iteraciones proveen de robustez a la solución, pero aumentan considerablemente el coste computacional de la solución. La asignación de consumidores contempla conexiones monofásicas y bifásicas.

- En el caso de un entorno rural, donde el espacio de soluciones es pequeño. Se ha optado por el desarrollo de un método heurístico a través del algoritmo Greedy. En este caso, se prioriza la convergencia prematura y el carácter explotativo de la solución, a través de un algoritmo de búsqueda limitada, aunque se limite a alcanzar óptimos locales. La implementación de este algoritmo está condicionada por su conocimiento del dominio. Por ello, se establece una métrica de contexto cuya minimización de la función objetivo, permite acercarse a la solución global. La asignación de consumidores solo contempla conexiones monofásicas. Este algoritmo es capaz de ejecutarse en tiempo real, mejorando notablemente la calidad del servicio ya que permite tomar medidas correctivas de manera inmediata.

- En otras ocasiones, la compañía eléctrica puede necesitar un sistema equilibrado con el menor número posible de cambios. Se propone la generación de una lista priorizada de cambios en la búsqueda del equilibrado de fases.

En el caso de conexión de nuevos clientes, lo que se propone como mejor opción es combinar, inicialmente, mecanismos que potencien el carácter exploratorio de la búsqueda mediante técnicas meta-heurísticas (algoritmo HS); y posteriormente, mecanismos que potencien el carácter explotativo de la búsqueda mediante técnicas heurísticas basadas en el contexto (algoritmo Greedy).

Dado que las cargas y las condiciones de funcionamiento de la red varían en el tiempo, la asignación de consumidores a fase en un periodo no tiene por qué ser la más adecuada para otros múltiples periodos. La periodicidad con la que se ejecuta el algoritmo es una decisión estratégica de la empresa eléctrica, que debe considerar el coste operativo de la reasignación de fases, el coste del desequilibrio, y el coste de la interrupción temporal del suministro.

Capítulo 3

Contribuciones predictivas para el Control y Supervisión de procesos industriales mediante sensores virtuales

Este capítulo se centra en la dificultad del cálculo, en tiempo real, de variables críticas de control y supervisión de procesos industriales que son difíciles de medir mediante sensórica u otro tipo de dispositivo. Se propone un procedimiento basado en el diseño de esquemas numéricos regresivos de Ciencia de Datos capaces de inferir el valor de dichas variables críticas, a través de la secuenciación de técnicas de preprocesamiento del dato (selección de características, reducción dimensional y procesamiento de señal), técnicas de adaptación dinámica a los cambios del proceso y métricas que evalúan la calidad de la solución propuesta.

3.1. Introducción

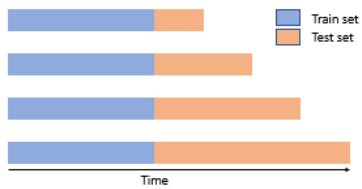
3.1.1. Estado del arte

Un sensor virtual basado en datos se define como un esquema de inferencia capaz de aprender ciertas relaciones de causalidad multiparamétricas y altamente no lineales a partir de un conjunto de datos históricos [119]. Su principal requisito es la existencia de datos [33]. Las principales ventajas competitivas de estos métodos de detección suave son [30]:

- no requieren de un conocimiento específico de las ecuaciones paramétricas que rigen las relaciones físicas del problema a tratar,
- son esquemas con alta capacidad de inferencia en relaciones multiparamétricas altamente no lineales,
- son sistemas que ofrecen, con un coste de diseño relativamente bajo, alta capacidad de generalización.

Cuando el proceso industrial a modelar responde a comportamientos estables y poco volátiles en el tiempo, una estrategia no-adaptativa es suficiente para el modelado del algoritmo de inferencia. Los métodos no-adaptativos modelan todos los datos del histórico mediante un *único* modelo global, ver Figura 3.1. Dicho modelo global es representativo de todas las realidades del proceso y la relación de inferencia aprendida con los datos históricos es aplicable también sobre datos futuros. Los métodos no-adaptativos funcionan bien cuando la dinámica del proceso está bien representada en dicho conjunto de datos de entrenamiento. Sin embargo, estos métodos no son eficientes, ni suficientes, si se producen nuevas dinámicas en el procesos, resultándoles difícil adaptarse rápidamente a cambios bruscos [120], [121]. Cuando surge un cambio brusco de comportamiento, tienen serias limitaciones para gestionarlo con propiedad. Además, no manejan bien las situaciones de no linealidad [122]-[124]. Se trata de modelos que se degradan en el tiempo [125].

Figura 3.1: Estrategia no-adaptativa



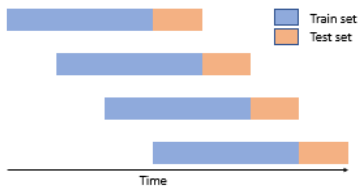
La generalidad es que el comportamiento no lineal del proceso y los cambios - bruscos o, graduales y lentos - en el modo de funcionamiento son los dos principales factores que provocan un mal rendimiento de los sensores virtuales basados en datos en los procesos industriales reales [119]-[121], [125]. Para resolver estos problemas relativos a los sensores virtuales no lineales basados en datos, los autores [126] introducen el término

sensor virtual adaptativo (adaptive soft sensor) como concepto para asimilar los cambios de comportamiento relacionados con la transformación dinámica del proceso.

Según [127], [128] son varios los mecanismos adaptativos que los sensores virtuales pueden adoptar para mantener un rendimiento satisfactorio durante un largo periodo de funcionamiento. Algunas estrategias comúnmente adoptadas incluyen:

- **Estrategia de ventana deslizante (Moving Window (MW)).** Este esquema utiliza el conjunto de datos de aprendizaje para generar sucesivos modelos locales de tamaño fijo, en lo que se conoce como ventana de instancias. El conjunto de instancias de la ventana, aunque fijo, se actualiza iterativamente incorporando nuevas y descartando las instancias más antiguas [129]. Al utilizar un enfoque basado en ventana, los conjuntos de entrenamiento y validación se desplazan iterativamente en el tiempo para incorporar las instancias más recientes, refrescando así la ventana después de cada nueva iteración [128]. Este enfoque refuerza que el funcionamiento actual tiene mayor relación de dependencia y correlación con las nuevas instancias que con las anteriores [130]-[132]. El esquema se representa en la Figura 3.2. Con cada nueva actualización de la ventana, se obtiene la información más reciente del proceso y sobre estos datos, se genera un nuevo modelo local que describe el estado actual, incluso cuando el proceso comienza a cambiar gradualmente [133]. Cuando los cambios son bruscos, el modelo está influenciado por el conjunto de datos restante de la ventana antes del cambio. El tamaño de la ventana condiciona la capacidad de generalización del modelo y el nivel de redundancia de los datos, así como la rápida o lenta adaptación a cambios bruscos del modelo: cuanto más pequeña es la ventana, menor es la afección de los datos antes del cambio y mejor la adaptación [134]. Aunque, por otra parte, puede producir estimaciones del modelo con alta varianza (especialmente en presencia de un gran número de variables de proceso colineales). La gran ventaja es que mediante la composición de un conjunto de modelos lineales locales, se puede describir aproximadamente un proceso no lineal [119], [126].

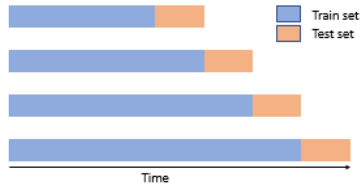
Figura 3.2: Estrategia adaptativa de ventana deslizante (MW)



- **Estrategia de ventana expansiva (Expanding Window (EW)).** Se trata de un esquema de avance de una ventana de aprendizaje incremental que se expande después de cada iteración con cada nueva

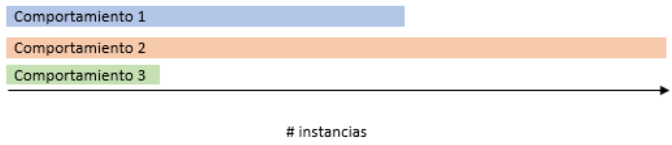
instancia. El conjunto de entrenamiento crece con cada nueva instancia y el conjunto de validación se desplaza iterativamente en el tiempo para incorporar las instancias más recientes [128]. Es decir, se amplía el tamaño de las instancias de entrenamiento desde un tamaño inicial hasta un tamaño máximo [129]. La adaptación a los nuevos comportamientos es lenta precisamente por el peso de los comportamientos anteriores que no se olvidan [134], véase Figura 3.3.

Figura 3.3: Estrategia adaptativa de ventana expansiva (EW)



- **Estrategia "Justo-A-Tiempo" (Just In Time Learning (JITL)).** Los esquemas basados en estrategias de aprendizaje Justo-A-Tiempo agrupan las instancias de entrenamiento bajo una métrica de similitud y con cada nueva instancia de validación se selecciona el grupo de instancias más similar para construir el modelo local [135]. Este enfoque de aprendizaje en base a modelos locales, al igual que los modelos MW, permite abordar bien los procesos no lineales, así como los cambios graduales [136], [137]. El principal inconveniente es el elevado tiempo de cálculo que implica entrenar continuamente un nuevo modelo desde cero cada vez que se dispone de una nueva instancia [138]. La mayor parte del éxito reside en los criterios de selección del conjunto de entrenamiento más similar [139]. La Figura 3.4 muestra una representación esquemática de esta estrategia de aprendizaje.

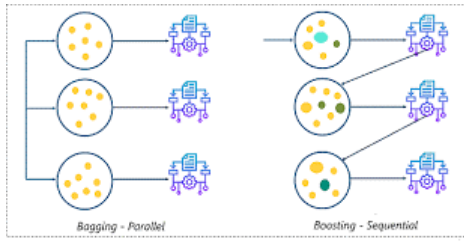
Figura 3.4: Estrategia adaptativa de aprendizaje Just-In-Time



La naturaleza de los datos es la que determina la relación de similitud más idónea en cada caso, en algunos casos la distancia euclídea puede ser la similitud correcta para determinar conjuntos de datos [135] o la distancia euclídea ponderada [140], [141], mientras que la correlación puede ser la medida de similitud real para otros [142]. Por otra parte, una estrategia JITL no es recomendable cuando se

trabaja con pocos datos [143], porque puede dar lugar a comportamientos sin muestras suficientes que los representen, muestras de entrenamiento insuficientes y no representativas de todos los tipos de comportamiento, y características en las muestras de entrenamiento diferentes a las muestras de validación [140], [141].

Figura 3.5: Estrategia adaptativa de ensamblaje de modelos: Bagging y Boosting



- **Estrategia basada en ensamblaje de modelos (Ensemble techniques)**.
  Estos esquemas basados en ensamblaje combinan varios modelos individuales con objeto de generar predicciones sólidas. La base de esta estrategia es que la combinación de varios modelos es más precisa que cualquiera de los modelos individuales que componen el modelo ensamblado [144]. Un buen método de ensamblaje de modelos es aquel en el que los modelos individuales son precisos y diversos [145]: (i) modelos precisos, en términos de baja tasa de error en los valores a predecir (disminuye el sesgo); y (ii) modelos diversos, en referencia a que los diferentes modelos sean complementarios y cometan diferentes errores (disminuye la varianza). El objetivo es dar robustez al modelo y encontrar la solución óptima que minimice el equilibrio entre sesgo y varianza y por tanto, minimice el error de generalización [146]. La Figura 3.5 esquematiza los métodos más populares de ensamblaje, que son:

  - Métodos Bagging [147] que consideran que una forma de reducir la varianza de las estimaciones es promediando (en el caso de modelos regresivos) o mediante votación (para los modelos de clasificación) de las estimaciones de distintos modelos individuales. Un ejemplo muy popular de Bagging es el algoritmos Random Forest Random Forest (RF).
  - Métodos Boosting [148] que consideran que una forma de reducir el sesgo de las estimaciones es secuenciar la salida de cada modelo individual y aprovechar la dependencia entre modelos simples para mejorar en los errores cometidos por el modelo anterior, con objeto de aprovecharse de ese conocimiento e intentar no volver a cometerlos. Un algoritmo representativo de Boosting, es el algoritmo Extreme Gradient Boosting (XGBoost [149]).

3.1.2. Trabajos relacionados

Varios son los campos de aplicación industrial de los sensores virtuales, los ejemplos más típicos son la industria química [150]-[152], la industria del papel [153]-[155], refinerías [156], [157], y recientemente, son muchos los estudios en relación al tratamiento de aguas [158]-[160]. Son ámbitos muy dispares pero que comparten la dificultad de sensorizar en entornos hostiles y que, como tal, comparten características comunes como ruido en las mediciones, valores erróneos, valores anómalos o atípicos, características colineales y/o variación en la frecuencia de muestreo de los diferentes analizadores [30]. Desde la perspectiva de cuál es el enfoque más apropiado para cada problema, se han publicado una cantidad considerable de documentos:

- La estrategia MW, se basa en la afección de la proximidad temporal entre los valores y es eficaz frente a cambios graduales de comportamiento. Por ejemplo, en [126] los autores con intención de adaptar sus modelos a las relaciones no lineales y a los cambios graduales del proceso proponen incorporar un índice de similitud en la varianza del ruido de las variable de procesos al aplicar el análisis de factores latentes supervisado (SLFA). El modelo - conocido como MW-WSLFA - sigue el enfoque de ventana deslizante y se actualiza constantemente de forma local. Otras investigaciones denotan el protagonismo de los métodos de modelado sencillos frente a modelos más complejo. Por ejemplo, en [132] los autores defienden la utilización de modelos sencillos, especialmente al trabajar con pocos datos históricos. En dicho artículo se hace una completa comparativa entre los siguientes métodos: regresión por mínimos cuadrados parciales en ventana móvil, regresión por mínimos cuadrados parciales regresivos en ventana móvil, la regresión Random Forest en ventana móvil, estimador medio de ventana móvil y un nuevo método de regresión de mínimos cuadrados parciales en Random Forest (RF-PLS). Básicamente, el estudio concluye que, para las estrategias de ventana móvil, los tamaños de ventana pequeños dan lugar a errores de predicción, también, pequeños. Finalmente, el artículo propone una regresión de mínimos cuadrados recursiva (R-PLS) por su capacidad para adaptarse a los nuevos datos y por el concepto de lo que denominan "función de olvido".

- La estrategia JITL, se basa en la proximidad espacial y son especialmente eficaces contra los cambios abruptos y recurrentes. Como estrategia basada en métodos JITL mencionar el enfoque de los autores [161] que proponen la construcción de un índice de similitud basado en la función kernel gaussiana para la selección de las variables relevantes y posteriormente, sugieren un modelo regresivo basado en mínimos cuadrados parciales (PLS) mediante técnicas de Bagging. Los métodos JITL, combinados con enfoques probabilísticos, se han aplicado reiteradamente para predecir variables de calidad de procesos [162]. Sin embargo, uno de los problemas en los enfoques JITL es

la selección poco flexible de los comportamientos locales. Para resolver este problema los autores [163] proponen un método denominado "PJITL de escala variable"(VS-PJITL) que permite determinar el tamaño de dichos comportamientos de forma variable. Los autores [164] proponen diseñar una solución JITL para seleccionar los datos de modelización basándose en el algoritmo de datos de vectores soporte (SVDD) y posteriormente la utilización de un modelo local utilizando el concepto de máquina vectorial relevante (RVM) como modelo predictivo.

- La estrategia de ensamblaje de modelos, normalmente se basa en variaciones de los enfoques de MW y JITL, con el objetivo de potenciar sus fortalezas individuales contra los diferentes tipos de concept-drift, aunque son bastante escasos en la literatura. En [165] los autores defienden la validez de métodos híbridos, combinando los métodos de adaptación habituales, como MW y JITL y utilizando la inferencia transductiva (MT_{tr}). Este aprendizaje transductivo o semisupervisado utiliza las predicciones JITL para obtener predicciones para las muestras de entrenamiento, que se utilizan posteriormente para entrenar el modelo MW en un entorno de regresión Lasso. En este caso, el estudio demuestra mejor rendimiento y precisión del modelo basado en MT_{tr} que los modelos MW y JTIL tratados de forma individual. En [139] se propone el estudio de un nuevo algoritmo de aprendizaje adaptativo basado en maquina vectorial de relevancia (RVM) llamado MWAdp-JITL, que combina el aprendizaje activo basado en MW capaz de adaptar el tamaño de la ventana frente a los casos concretos de concept-drift, que se integra con un modelo JITL en base a unos datos de históricos semejantes, y un sistema de pesos para los modelos MW y JITL. El estudio busca el equilibrio entre el aprendizaje de nuevas instancias y el olvido de instancias antiguas, en lo que se denomina dilema del aprendizaje y el olvido [166]. En[127] los autores desarrollan un enfoque de localización adaptativa para tratar la no linealidad del proceso mediante un enfoque de aprendizaje selectivo de conjuntos para salidas múltiples (SEL-MO). Además, propone una estrategia de insensibilidad evitando la búsqueda innecesaria del conjunto de datos histórico y mejorando la eficiencia computacional. En [167] los autores proponen una estrategia de aprendizaje de ensamblado de conjuntos justo a tiempo (E-JITL), donde se adoptan diferentes medidas de similitud para la selección de muestras para cada modelo local. Es difícil conocer la similitud real de la mayoría de los datos del proceso, por lo que los autores proponen utilizar diferentes medidas de similitud. En base a cada una de estas medidas de similitud, se seleccionan diferentes grupos de muestras relevantes (uno por medida de similitud), se construye cada uno de los modelos locales y se entrena para predecir el resultado de cada muestra de consulta. La predicción final se obtiene combinando todos los valores predichos de cada uno de los modelos locales entrenados mediante la suma ponderada. Se valida el funcionamiento propuesto en dos aplicaciones industriales: proceso de

hidro craqueo para gasóleos pesados para convertirlos en destilados ligero y el proceso de fabricación de hierro. Los autores [143] proponen una solución basada en la estrategia JITL cuando las muestras de entrenamiento no son representativas de los comportamientos que se aprecian en las muestras de validación. Este trabajo propone técnicas de aprendizaje por refuerzo como alternativa a la ausencia de datos como opción para aprender de datos relevantes en otros dominios, entendiendo que, aunque se trata de datos etiquetados en otras condiciones de funcionamiento están sujetos a un mecanismo de proceso similar. Como método de aprendizaje de transferencia simple, proponen la utilización de máquina de aprendizaje extremo (ELM) con adaptaciones de dominio.

Pero la realidad es que en muy pocas de las referencias anteriores se presta apenas atención a la fase de pre-procesamiento. En [168] y con el objetivo de atenuar la influencia de la correlación común y el ruido, se propone un método de análisis de causalidad y selección de las características basándose en la ortogonalidad y los factores causales más eficaces. Utiliza como métodos de aprendizaje de características el análisis de componentes principales Principal Component Analysis (PCA), por su capacidad de obtener las variables latentes que más información de la varianza aportan; el análisis de características lentas (SFA [169]) por su capacidad de obtener las variables latentes con cambios lentos; y sensores de base variable con análisis de causalidad (CA). El estudio concluye con la verificación de las técnicas propuestas en un algoritmo de ajuste por mínimos cuadrados (PLS), en una red neuronal recurrente (RNN) y en una red neuronal artificial (ANN). PCA se vuelve a utilizar en [170] para reducir la dimensión de los datos de entrada. En este caso, el algoritmo genético (GA) se utiliza para estimar los retrasos temporales del sistema mediante la optimización de un modelo lineal de retrasos temporales. En [171] los autores coinciden en la importancia de la selección de características más relevantes antes del modelado de los sensores en distribuciones no-lineales, y proponen un análisis de componentes principales probabilístico ponderado (WPPCA). El método consiste en asignar diferentes pesos en base a una métrica de similitud respecto a una muestra de prueba. En [172] proponen utilizar el análisis de componentes principales de la función del núcleo (KPCA) como alternativa a una distribución no-lineal.

Recientemente, la línea de investigación de los sensores virtuales está evolucionando hacia soluciones basadas en aprendizaje profundo, donde la selección o reducción de características se tiende a obviar en pro de la modelización automática de las redes. En [173] los autores han desarrollado recientemente una metodología basada en la técnica de Información Mutua (IM) normalizada como método de iterativo para selección de variables características. El método considera inicialmente la primera variable y mide el error del modelo, sucesivamente va añadiendo - una a una - el resto de variable. En cada iteración evalúa el modelo, de forma que solo considera aquellas variables que o mejoran su precisión o finalmente, se cumple la condición de parada. Los autores consideran IM como una

excelente métrica de evaluación, ya que no se ve afectada ni por la transformación de los datos a lo largo del tiempo, ni por el ruido. Además, es una medida aleatoria que no hace suposiciones ni de la distribución, ni de las dependencias entre variables. Este enfoque permite reducir la complejidad causada por las variables candidatas redundantes, y con aquellas variables de mayor significancia estadística, crea un modelo basado en LSTMs para manejar la dependencia de la secuencia. Otro estudio basado en aprendizaje profundo es el análisis que proponen los autores [174] para extraer representaciones latentes de las variables mediante modelos de mezcla gaussianos (GMM) mediante autoencoders variacionales (Gaussian mixture Variational Autoencoder, GMVAE). Tras la aplicación del algoritmo GMVAE, cada variable se transforma en una variable latente que se describe mediante una distribución de mezcla gaussiana. Con cada nueva muestra, se calcula la relación de divergencia simétrica entre dos distribuciones de probabilidad de mezclas gaussianas (la distribución de la nueva muestra y la distribución de las muestras históricas). Basándose en los valores calculados de MSKL (Mixture symmetric Kullback-Leibler), se asignan pesos a cada muestra y se establece el número de muestras que forman el modelo local. Sobre el modelo local se aplica un modelo de regresión probabilística de componente principales (MPPCR). Otro aspecto a considerar, y es la falta de disponibilidad de datos en la fase de modelado que en ocasiones se suple mediante técnicas de aprendizaje profundo. En estos casos, los modelos además de la falta de interpretabilidad, tienden a estar sobre-ajustados. En [152] se propone el desarrollo de sensores virtuales mediante la combinación de simulaciones de primer principio (FPM) y aprendizaje de transferencia para la generación de gran cantidad de datos mediante simulaciones dinámicas, posteriormente se ajusta el modelo trabajando únicamente con los datos reales. Este enfoque permite mejorar la precisión de la predicción en el dominio objetivo y garantizar que los modelos incorporan un conocimiento correcto del dominio. Los ejemplos anteriores, basados en aprendizaje profundo, presentan la ventaja de ser capaces de abordar problemas de gran complejidad, pero pueden resultar poco convenientes para aplicaciones industriales cuando el coste de tiempo computacional es alto, y la falta de interpretabilidad suponga un obstáculo [30].

No hay una única solución ideal que funcione siempre [175], [176], pero es realmente esencial ampliar la investigación y el desarrollo de sensores virtuales sobre la base de la compatibilidad industrial [177].

Este capítulo centra sus estudios en técnicas de pre-procesamiento, en la capacidad de encadenar técnicas estadísticas sobre las variables relevantes del proceso y en la selección de la estrategia más adecuada en cada situación, demostrando la validez del método en la realidad de tres casos de uso industriales singulares para la industria petroquímica, industria química y del sector del reciclaje industrial.

3.2. Enfoque propuesto

La metodología básica en el diseño de sensores virtuales consta de cinco pasos [178]: pre-procesamiento de los datos, selección de los características relevantes, análisis del desfase temporal, selección de la estrategia y del modelo adecuado, y por último, validación del modelo [179].

3.2.1. Pre-procesamiento de datos

El éxito de cualquier técnica de inferencia depende en gran medida de la calidad de los datos introducidos al modelo [180]. Los datos industriales del mundo real suelen estar sucios, son ruidosos y contienen valores atípicos, características irrelevantes o innecesarias, valores nulos o no estandarizados [181]. A menudo, cuando se utilizan datos erróneos o sin procesar y el modelo resultante tiende a estar sesgado o a no funcionar adecuadamente [182]. Por esta razón, la transformación de los datos originales en la fase de pre-procesamiento, es esencial.

En general un método o secuencia de métodos de pre-procesamiento son adecuados cuando tras el tratamiento de los datos son capaces de conservar la naturaleza y/o características de las variables originales (por ejemplo, la distribución o la frecuencia), incorporar el conocimiento intrínseco del proceso y capaces de mantener el coste computacional dentro de un rango manejable [183]. Dadas las propias características distintivas de cada conjunto de datos y de cada proceso, es difícil establecer un único criterio aplicable universalmente para dicho pre-procesamiento [175]. De hecho, encontrar la combinación perfecta en cada caso es la clave del éxito para la particularidad de cada caso de uso [175]. Estas técnicas que pueden utilizarse individualmente o combinadas para mejorar los resultados de etapas posteriores. A continuación, se mencionan tres de las técnicas que mayores ventajas aportan:

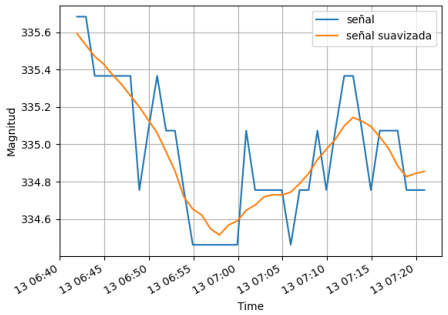
- **Limpiar variables de entrada, modelar la normalidad del proceso.** Modelar la normalidad de un proceso requiere eliminar valores inconsistentes con la distribución de los datos. De no hacerlo, los modelos estarían implementando una visión errónea de la realidad [184] aunque igualmente valiosa sobre procesos atípicos que puede interesar estudiar [185]. Lo importante es reconocer y determinar qué es lo que se quiere hacer con dichas anomalías. En las estrategias basadas en ventana móvil, la desviación se estudia en cada ventana. De esta manera, se estudia individualmente cada variable de entrada y se eliminan todas aquellas medidas que no sean representativas del proceso. A esta técnica se le conoce como "limpieza probabilística de datos" (probabilistic data cleaning) [181] y es especialmente adecuada, cuando los datos siguen una distribución normal [184].

- **Normalizar las variables de entrada.** Si las variables de entrada al sensor virtual tienen distinto rango de magnitudes, es recomendable la normalización de los datos [119]. De lo contrario, las variables importantes del proceso que tienen magnitudes pequeñas se verán

afectadas por las variables menos importantes que tienen magnitudes mayores. Para evitar esto, la normalización de los datos se realiza de forma que todas las variables numéricas se transformen en una escala común y por tanto, tenga un peso semejante. Su influencia sobre el modelo será similar, mejorando la estabilidad y el rendimiento del algoritmo. Un método común de normalización es el filtro MinMax [186] que escala y traduce cada entrada individualmente de forma que se encuentre en el rango máximo y mínimo en el conjunto de entrenamiento.

- **Eliminar ruido en las variables de entrada.** La mayoría de los datos proceden de sensores que generan señales por medios electrónicos o electro-mecánicos. A menudo, las señales de los sensores se degradan con ruido de alta frecuencia, principalmente por el entorno en el que se ubican [183]. Para eliminar el ruido, la práctica habitual es el filtrado de la señal. En este sentido, el filtro de Savitzky-Golay [187]) es un filtro adaptativo basado en datos muy utilizado. El filtro Savitzky-Golay, también conocido como filtro de suavizado polinómico de mínimos cuadrados, reduce el ruido de alta frecuencia gracias a sus propiedades de suavizado y reduce la señal de baja frecuencia mediante la diferenciación. Un ejemplo de la aplicación del filtro Savitzky-Golay se muestra en la Figura 3.6.

Figura 3.6: Aplicación del filtro de suavizado polinómico de mínimos cuadrados



3.2.2. Selección de variables

En los procesos industriales, los procesos de supervisión y control pueden llegar a estar formados por miles de variables. Lo ideal es que la selección del conjunto de variables más relevantes sea realizada por expertos del dominio [178]. Aunque en procesos físicamente grandes con miles de variables la selección es compleja, realizar esta tarea es altamente recomendable [188]. La selección óptima y reducida de variables, implica información únicamente de los sensores esenciales, permitiendo reducir costes en la adquisición y mantenimiento de sensores superfluos y reducir información

co-lineal y redundante. Además de mitigar el efecto de la *"maldición de la dimensionalidad"* [189], [190] que explica la mala escalabilidad de los algoritmos y la reducción drástica de su eficiencia cuando se trabaja con datos de alta dimensionalidad severa. Por todo ello, reducir la dimensión de las variables implica mejorar el rendimiento y la rapidez del modelo y mejorar la eficiencia de la predicción [191], [192].

Los enfoques más habituales en las técnicas de pre-procesamiento para reducir el número de variables con los que trabajar, son dos: (i) centrar los esfuerzos en la identificación de las variables más relevantes y que realmente aportan conocimiento en la obtención de la variable objetivo, descartando variables superfluas o redundantes; (ii) reducir la dimensional de las variables identificadas por el experto de dominio mediante técnicas y métodos de reducción de la dimensionalidad.

Técnicas de selección de variables relevantes

Durante el modelado de sensores virtuales, se utilizan datos históricos para aprender las relaciones entre las entradas y las salidas. Algunas de las técnicas más utilizadas en entornos de producción real son:

- **Importancia de la permutación** [193], [194]. Se trata de un método de eliminación hacia atrás. Comienza con un conjunto que incluye todas las entradas candidatas, luego se elimina iterativamente una sola entrada. Se calcula cuánto se deteriora la función de pérdida al eliminar cada entrada. Este deterioro del rendimiento mide la importancia de la entrada que se acaba de eliminar. Se sustituye la entrada eliminada y se elimina una nueva hasta que se calcula la importancia de cada entrada. Finalmente, se eliminan las entradas menos significativas.

- **Secuenciación de algoritmos para la selección de la importancia de características**. Para identificar las variables más importantes se aplica una serie de algoritmos de selección de características (por ejemplo, Random Forest y Gradient Boosting) que determinan de entre todas las características cuáles son aquellas que tienen más relevancia. Así mismo, se establece un proceso de ponderación entre las entradas elegidas por cada algoritmo y se elige como relevantes aquellas entradas que sean seleccionadas con mayor asiduidad por los distintos algoritmos. En procesos industriales y ante la posibilidad, siempre se recomienda verificar las características seleccionadas con técnico experto de dominio desde un sentido físico del proceso [195].

Pero estos datos históricos contienen información limitada, por lo que cuando el sensor virtual se enfrenta a una situación nunca vista antes, el rendimiento del modelo suele disminuir. Esta degradación del rendimiento del modelo indica que posiblemente las propiedades estadísticas del proceso -debido a cambios en el propio proceso, a cambios en la materia prima o a alguna otra condición externa del proceso- cambian con el tiempo. Esto se conoce como deriva conceptual (concept drift) [196] y afecta a la relación entre los datos de entrada y los de salida [197]. Existen dos

métodos principales para minimizar en lo posible los efectos de la deriva de conceptos:

- **Consistencia en el tiempo** [198]. Se trata de identificar y minimizar el uso de aquellas entradas que son más volátiles a lo largo del tiempo, estudiando la distribución de cada entrada al principio y al final. Se eliminan las entradas con distribuciones diferentes.
- **Validación adversaria** [199]: El objetivo es identificar si las entradas de entrenamiento son diferentes a las entradas de validación. Consiste en construir un clasificador para intentar predecir qué datos son del conjunto de entrenamiento y cuáles son del conjunto de validación. Si los dos conjuntos provienen de la misma distribución, esto debería ser imposible. Pero si hubiese diferencias sistemáticas, un clasificador sería capaz de diferenciar las entradas de entrenamiento de las de validación. Cuanto mejor sea el clasificador para diferenciarlos, mayor será la diferencia de distribución entre ambos conjuntos.

Técnicas de reducción de la dimensionalidad

Las técnicas de reducción de la dimensionalidad consisten en la transformación de entradas de alta dimensión en otro conjunto de entradas de baja dimensión obtenido mediante la combinación de las entradas originales. Si los datos de alta dimensión están dentro de un subespacio lineal, pueden representarse mediante una transformación lineal de entradas de baja dimensión sin pérdida de información. Una técnica lineal popular es el análisis de componentes principales (PCA) [200], [201]. El objetivo del PCA es encontrar los componentes principales en el espacio de baja dimensión que describen la mayor parte posible de la varianza de los datos de alta dimensión. Sin embargo, si los datos no son lineales las relaciones entre las variables son desconocidas [202].

3.2.3. Análisis de desfase temporal

En procesos reales, en ocasiones ocurre que las entradas utilizadas por el modelo para la inferencia, no se corresponden con el mismo instante de tiempo en el que se obtuvo el valor real de la salida. Es decir, a veces ocurre que el valor de la salida objetivo en t_0 está afectado por valores de las entradas en momentos previos a t_0. Una forma habitual, y poco eficiente, de introducir la dinámica del proceso en un modelo es incorporar todas las entradas del proceso y sus valores previos. Pero, este enfoque aumenta exponencialmente la complejidad del modelo.

La técnica de "correlación cruzada" [203] consiste en el estudio de la correlación entre la salida objetivo y cada una de las variables de entrada. A mayor correlación mayor es la afección de la variable de entrada con la salida. E incluso, aún más eficaz es la aplicación de la variante conocida como **correlación cruzada retardada en el tiempo** (TLCC) [204] donde en cada desplazamiento de la variable de entrada se calcula la correlación.

El procedimiento consiste en medir la correlación entre la salida objetivo y cada variable de entrada cada instante del proceso.

3.2.4. Selección de la estrategia y entrenamiento del modelo

El aprendizaje adaptativo surge para mitigar los efectos de la deriva conceptual, dando lugar a lo que se conoce como aprendizaje adaptativo del sensor virtual [205]. Pero, entre las estrategias propuestas en la Sección 3.1.1 y que tienen como objetivo fomentar dicha adaptabilidad al entorno, según el teorema *"no free lunch theorem"* [175], una determinada estrategia de aprendizaje que tiene éxito en un contexto con un determinado conjunto de datos puede no tenerlo en otro.

Toda estrategia de aprendizaje adaptativo requiere definir el tamaño óptimo de la ventana de entrenamiento. Si una ventana es más corta que la adecuada, puede que no contenga suficiente información sobre el estado actual del proceso o que se adapte al ruido y esto produciría un modelo con poca generalización [119]. Por contra, si una ventana es más larga que la adecuada, puede contener información redundante y capacidad de adaptación limitada [119], [206]. Esto también deteriora el rendimiento de la predicción del sensor virtual adaptativo junto con una baja eficiencia de cálculo. Por cada una de estas ventanas, se realiza un entrenamiento y una validación con un algoritmo concreto. El algoritmo elegido será aquel que permita obtener el resultado de inferencia óptimo. Tal y como se explica a continuación, cuatro son los algoritmos seleccionados:

- Ridge Regression, algoritmo lineal basado en el modelo clásico de regresión lineal pero que regulariza el impacto sobre las características no relevantes;
- Random Forest Regression, algoritmo no lineal que tiene la capacidad de actuar como un algoritmo de bagging mediante el ensamblaje de árboles individuales;
- XGBoost Regression[149], algoritmo lineal o no lineal dependiendo del kernel utilizado mediante el ensamblaje boosting de árboles individuales; y
- Support Vector Regression (SVR) [207], algoritmo no lineal que transforma los datos en un espacio de características de mayor dimensión para hacer posible la separación lineal.

3.2.5. Validación

La elección de la métrica para la evaluación del rendimiento del modelo es una cuestión subjetiva, estrechamente relacionada con: la estrategia de aprendizaje elegida, con el algoritmo y con los aspectos que los expertos del dominio quieran priorizar. A continuación, se exponen las métricas más populares en clasificación y regresión.

- Clasificación. Pese a que no es habitual trabajar con métodos binarios de clasificación desbalanceados en sensores virtuales, la forma más común de representarlo sería mediante una matriz de confusión, o ciertas métricas específicas del tipo balanced Accuracy (bAcc), y/o Matthews Correlation Coefficient (MCC), y

- Regresión. En el caso de la evaluación de la predicción de variables continuas, la función de pérdida del error cuadrático medio (MSE) o la raíz del error cuadrático medio (Root Mean Square Error (RMSE)) se utiliza habitualmente para este tipo de modelos. Sin embargo, si se dispone de un proceso con valores muy estables y lo que se desea es una métrica comprendida entre un rango de valores, que sea fácilmente interpretable y que permita comparar esquemas entre si, utilizar el valor de Normalized Root Mean Square Error (NRMSE) es una opción nada desdeñable. NRMSE representa el error cuadrático medio normalizado que a menudo expresa un porcentaje, y que como tal, toma valores entre 0 y 1, y que ofrece un análisis más intuitivo del rendimiento del modelo. Un valor de $NRMSE = 0$ indica una predicción perfecta, mientras que un valor de $NRMSE = 1$ corresponde a una predicción de la media estadística. La fórmula NRMSE es:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{3.1}$$

, donde y e $\hat{y}$ son los valores reales y los valores de la predicción de los datos de validación respectivamente, y_{max} e y_{min} es el valor máximo y mínimo de entre todos los valores reales y n es el número de elementos de los datos de validación. La formula de RMSE es:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{3.2}$$

Otra medida comúnmente utilizada para evaluar un modelo es el coeficiente de determinación R^2, y que al igual que NRMSE oscila entre valores de $-\infty$ y 1 (cuando el valor de R^2 es negativo, significa que el modelo es peor que predecir la media).

A continuación, se aplica el procedimiento anteriormente explicado sobre diferentes casos reales.

3.3. Caso Investigación Industrial: Planta Petro-Química

Las refinerías producen múltiples subproductos del petróleo que se utilizan para diversas aplicaciones. Por consideraciones de seguridad y para ser comercializados, estos productos tienen que cumplir una serie de especificaciones de calidad. Una de las especificaciones de calidad crítica para

el gasóleo de automoción es la temperatura del punto de inflamación. Se suele controlar después de que el gasóleo haya sido procesado en la unidad industrial de desulfuración. En esta, la unidad de desulfuración, tiene lugar el proceso de acondicionamiento de las corrientes de carga (procedentes de las unidades de destilación atmosférica y al vacío y de las unidades térmicas y catalíticas) a las especificaciones del gasóleo comercial. Este proceso consiste principalmente en una hidrogenación catalítica que elimina el azufre, el nitrógeno, los compuestos de oxígeno y otras impurezas metálicas presentes en las corrientes de carga, dando como resultado un producto más refinado que cumple los requisitos de manipulación, transporte y combustibilidad del cliente. Una de las propiedades críticas que debe cumplir la unidad, es la Temperatura de Punto de Flash (FP). La Temperatura de Flash es una propiedad del gasóleo que indica la temperatura más baja a la que hay suficiente vapor inflamable para prender cuando se aplica una fuente de ignición. Está determinado por el número de hidrocarburos ligeros presentes en el gasóleo y dicta la inflamabilidad del combustible.

FP es una propiedad relevante, no solo en los combustibles, sino también en otros productos químicos inflamables como los disolventes y las materias primas. Incluso para la manipulación segura de productos químicos inflamables, el punto de inflamación está establecido de forma oficial (norma europea UNE-EN 590 [208], [209]). Por ello, en la industria de procesos la búsqueda de estrategias para el manejo de la inflamabilidad ha sido un tema de investigación constante [210].

Los métodos habituales para la predicción de la FP se basan en el estudio de la correlación empírica entre propiedades de inflamabilidad a partir de la premisa de que compuestos estructuralmente similares tienen actividades inflamables similares [211]. Los primeros métodos de estimación de FP se basaban en modelos empíricos de correlación de propiedades físicas, principalmente relativas al punto normal de ebullición (PNB) y la entalpia de vaporización (Hv) dada su capacidad para representar la volatilidad y, por tanto, la inflamabilidad del combustible [212]. Las mediciones se realizan de forma indirecta tomando muestras del proceso y analizándolas en el laboratorio mediante métodos de medición experimentales, por ejemplo, método de la "copa cerrada" de Pensky-Martens [212], [213]. Hoy en día, los nuevos desarrollos se centran en métodos avanzados basados en el modelo de contribución de grupos (group contribution model o GCM) que utilizan para predecir las propiedades termodinámicas de los compuestos orgánicos a partir de su estructura molecular y modelos de relación cuantitativa estructura-propiedad (quantitative structure-property relationship o QSRP) [214], [215] que utilizan información de los descriptores moleculares para representar las características de los numerosos compuestos orgánicos. Estos descriptores representan numéricamente varias propiedades estructurales químicas como las características constitucionales, topológicas, geométricas, termodinámicas, químicas cuánticas y relacionadas con la carga [216]. Habitualmente, los descriptores más informativos para la predicción de una propiedad objetivo se seleccionan a través de múltiples procedimientos de selección de características [216]. Ambos enfoques utilizan las informaciones moleculares de los compuestos orgánicos como

variables predictoras de entrada a los modelos [217].

Paralelamente, se han desarrollado nuevos modelos GCM y QSPR mediante el uso de diversos algoritmos de aprendizaje máquina, aumentado la predictibilidad de estos modelos [216]. Algoritmos como, por ejemplo, la regresión lineal múltiple (MLR), la red neuronal artificial (ANN), los mínimos cuadrados parciales (PLS), la máquina de vectores de apoyo (SVM), los vecinos más cercanos (KNN), el bosque aleatorio (RF) y las regresiones no lineales [216]. Se puede encontrar una buena revisión bibliográfica en las siguientes revisiones literarias [218]-[220].

Sin embargo, en la industria de procesos, estos métodos tienen varias limitaciones operativas. Los métodos experimentales requieren tareas de análisis de muestras en el laboratorio que pueden llevar horas desde la recogida de la muestra hasta la medición de la temperatura del Flash [221]. Y, en los métodos basados en GCM y QSPR, donde la información molecular es indispensable, el problema es el mismo, la falta de información en tiempo real de las propiedades estructurales químicas [222].

El estudio presentado en esta Tesis es el primer enfoque conocido capaz de inferir la FP utilizando mediciones experimentales y datos de funcionamiento de la planta. La autora no tiene constancia de que existan trabajos previos en la literatura que estudien en tiempo real la variabilidad natural del proceso, en términos de temperatura, presión y caudal hidráulico, para el proceso de inferencia de la temperatura de Flash. Por ello, propone un estudio que evalúe la relación entre la complejidad de los métodos de aprendizaje automático utilizados y la calidad de la predicción a través de un estudio estadístico de cómo afectan las técnicas de selección de características relevantes, las técnicas de reducción dimensional y las técnicas de procesamiento de señal, en combinación con diferentes algoritmos regresivos. Desde algoritmos regresivos como Ridge -algoritmo regresivo lineal regularizado- hasta algoritmos de ensamblaje de árboles como RandomForest y XGBoost en Bagging o Boosting y sin necesidad de implementar complejas técnicas profundas.

Dataset

El proceso de desulfuración de gasóleo es un proceso continuo, constituido por secuencias de pequeños paquetes de diferentes mezclas de crudo que alimentan las unidades de destilación. La composición molecular de estas secuencias de crudo está muy condicionada por el origen geográfico de la extracción del crudo. Secuencias de crudo de un mismo origen geográfico tienden a mantenerse constantes en su composición molar, en la calidad de los múltiples subproductos que de él se obtienen y por ende, en la afección de las variables del proceso sobre el valor de FP. Pero estas relaciones varían dinámicamente y de forma imprevisible, cuando se opera con crudos de diferente origen.

Los datos utilizados para el estudio contemplan un histórico de tres años (2017, 2018 y 2019). En opinión de los técnicos expertos de la refinería este periodo es suficiente para recoger toda la variabilidad del proceso. Se trata de aproximadamente 730 paquetes, 1,600,000 instancias minutales con 200 variables de entradas por instancia (variables de proceso) y 854 mediciones experimentales diarias de "copa cerrada" (valores de FP). La

frecuencia media de muestreo de los sensores del proceso es de 10 minuto para las variables de entrada y diaria para la variable salida. Todas las variables de entrada están formadas por información de variables físicas del proceso (temperatura, presión y flujo hidráulico) de las unidades de desulfuración involucradas. Indirectamente estas variables son capaces de caracterizar los componentes más ligeros de cada una de las unidades que afectan a la FP.

A nivel de arquitectura, el acceso a los datos se realiza a través de servicios REST, que proporcionan información de cada una de las corrientes. Todos los datos se almacenaron en un único objeto serializable para el posterior estudio que se describe en los siguientes apartados.

3.3.1. Metodología

A continuación, se describen las técnicas utilizadas en la implementación del proceso, y la afección de cada una de ellas en la búsqueda de la solución más óptima.

- **Técnicas de pre-procesamiento.**
  Inicialmente, los datos en crudo recogidos de las unidades de desulfuración han de ser acondicionados para su uso en el diseño del sensor virtual. El tratamiento realizado sobre los mismos se resume en los siguientes pasos:

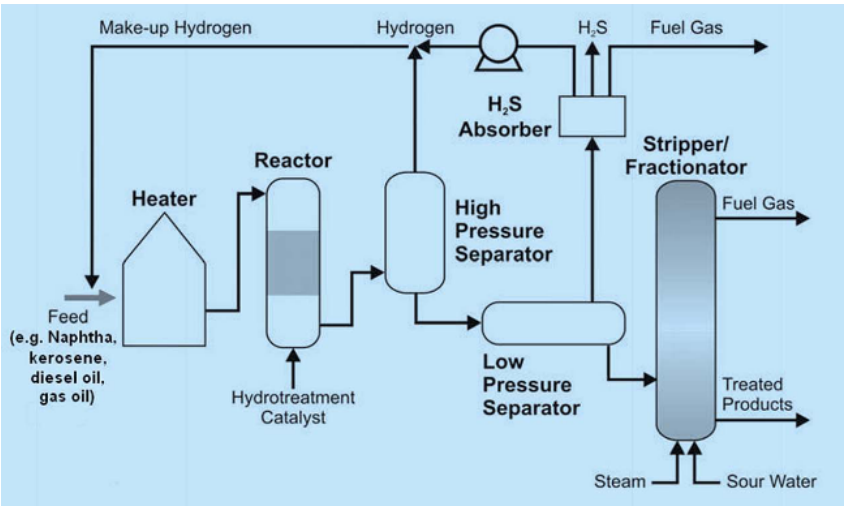
  - Se define el "Modo de operación" del proceso que se desea estudiar. Si la unidad opera en un modo de funcionamiento diferente al deseado, los datos se descartan. En este caso concreto, se eliminan las instancias en los que la unidad opera en "Modo de Arranque" o en "Modo Recirculación". Únicamente se consideran las instancias en las que la alimentación principal es "Modo Queroseno".
  - De las 200 variables de proceso se descartan las variables con un único valor constante (variables que no aportan información sobre la dinámica del proceso e innecesarias) y variables con alto porcentaje de valores nulos ($> 90\,\%$).

- **Técnicas de selección de entradas.**
  Para determinar la relevancia de cada variable de entrada se aplican métodos que permiten descartar aquellas variables sin efecto sobre FP:

  - Métodos agnósticos del modelo como la técnica de Importancia de la Permutación que permite medir los cambios de varianza relacionados con la permutación de las variables relevantes; y
  - Métodos específicos del modelo, como Random Forest y Gradient Boosting que a través del índice Gini, permiten cuantificar el efecto de cada variable de entrada.

A partir de la combinación de métodos (agnósticos y específicos) se calculan las entradas más relevantes. Asimismo, se confirma que las variables de entrada seleccionadas tienen importancia desde el punto de vista físico. Para ello se cuenta con la ayuda de los expertos del dominio y su conocimiento del proceso.

Figura 3.7: Diagrama de flujo simplificado de una planta de desulfuración de gasóleo. Fuente:[223]



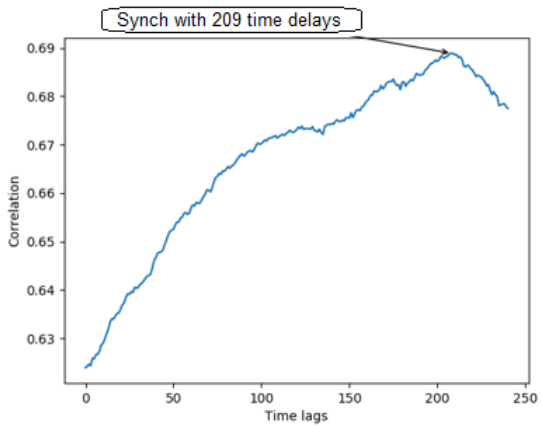
La Figura 3.7 muestra un diagrama simplificado del flujo del proceso de desulfuración, donde se ubican el conjunto de variables relevantes y que está formado por información de caudal, presión y temperatura de las unidades de desulfuración o de puntos de consigna y salidas del controlador. Las siguientes variables son las más relevantes:

- Temperatura de retirada del producto intermedio del reactor de desulfuración (tanto variable medida como salida del controlador),
- Temperatura de alimentación del separador,
- Temperatura del flujo superior del separador,
- Temperatura del flujo inferior del separador,
- Presión superior del separador (tanto variable medida como salida del controlador),
- Flujo de alimentación del separador,
- Flujo de reflujo del producto inferior del separador (tanto variable medida como salida del controlador)
- Flujo de reflujo del producto superior del separador (tanto variable medida como salida del controlador)
- Flujo de alimentación de craqueado a la unidad,
- Flujo de alimentación de nafta pesada a la unidad, y

- Flujo de alimentación combinado a la unidad.

- **Análisis del desfase temporal.**
  El proceso de desulfuración, desde que se introduce la mezcla de crudo en el sistema hasta que se obtienen los derivados del gasóleo, no excede de 4 horas. Es por ello, que es necesario encontrar para cada una de las variables de entrada relevantes el instante retardado de mayor afección sobre FP. La correlación cruzada retardada (TLCC) estudia la sincronía entre cada entrada y la salida objetivo calculando el instante en que la correlación era máxima. Se muestra un ejemplo en la Figura 3.8 donde se indican los valores de correlación entre una de las variables de entrada y FP, siendo el minuto 209 ($3\frac{1}{2}$ horas) el instante temporal de correlación máxima.

Figura 3.8: Análisis de correlación cruzada retardada para una entrada de temperatura



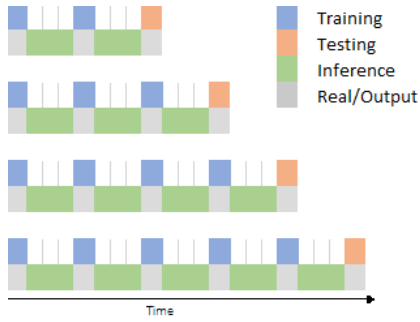
- **Selección de la estrategia y entrenamiento del modelo.**
  Paralelamente, se procede al estudio de la linearidad del proceso y se confirma que el dinamismo del proceso de desulfuración está condicionado de tal forma por la composición molar de la mezcla del crudo que ésta determina el conjunto de variables determinantes en cada momento. A través de técnicas de validación adversaria y técnicas de consistencia de las características en el tiempo se concluye que las variables relevantes no son fijas, sino que van cambiando con el tipo de crudo. Es decir, no todos los datos históricos son representativos del mismo comportamiento entre las variables de entrada y la variable de salida. O lo que es lo mismo, un único modelo global no será nunca representativo de todas las realidades del proceso. Se trata de encontrar la mejor solución a un problema predictivo adaptativo a cada nueva realidad (comportamiento).

  Se utiliza la técnica de Ventana Expansiva (EW) como estrategia de aprendizaje adaptativo, re-entrenando la ventana automáticamente con cada nueva instancia diaria. Se utiliza el algoritmo de Random

Forest como método regresivo para calcular la inferencia en cada re-entrenamiento de cada nueva ventana.

- **Validación.**
  La validez de la solución se prueba con el método de validación cruzada de series temporales divididas (split time series cross validation [224]). Y se utiliza la métrica RMSE como medida de éxito para cada ventana. Con cada nueva validación, cuando el valor del RMSE está por debajo de un determinado umbral (basado en la reproducibilidad establecida por el método estándar ASTM D97 [213]) dicha instancia se incluye en el conjunto de entrenamiento de la siguiente ventana de entrenamiento. En caso contrario, se descarta.

Figura 3.9: Inferencia mínima y validación cruzada de series temporales



Para el proceso de entrenamiento y validación de cada ventana se utilizan datos diarios (recordemos que las pruebas de laboratorio solo proporcionan valores reales diarios de FP). Con el modelo ya entrenado y conocida la relación entre las variables de entrada y salida, se aplica dicho modelo para la predicción sobre variables de entrada 10 minutales. La Figura 3.9 ilustra este procedimiento.

3.3.2. Comparación de métodos y discusión

En la búsqueda de la solución óptima, se comparan diferentes técnicas de pre-procesamiento (limpieza probabilística de los datos o probabilistic-data-cleaning, normalización entre los valores máximos y mínimos de cada característica o minmax-scaler, eliminación del ruido mediante el suavizado de la entrada con el filtro Savitzky-Golay o savgol-filter) y reducción de la dimensionalidad con PCA o pca-decomposition), diferentes algoritmos de regresión (Ridge, RF, XGB y SVR) y diferentes estrategias de aprendizaje adaptativo (MW y EW). A continuación, en la Tabla 3.1, se presentan los resultados de validación calculados en base a las diferentes técnicas y algoritmos aplicados a partir del conjunto de variables relevantes diarias.

La observación de la tabla indica que de mayor/peor a menor/mejor valor de la métrica RMSE, las soluciones óptimas en la aplicación de cada algoritmo en cada ventana se pueden ordenar de la siguiente forma:

Tabla 3.1: Comparación de la métrica RMSE: técnicas de pre-procesamiento, entradas relevantes y estrategia EW.

Técnicas de pre-procesamiento				Algoritmos			
probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	Ridge	RF	XGB	SVR
✓	✓	✓	✓	4,529	4,345	4,985	4,448
✓	✓	✓		4,462	4,319	4,982	4,451
✓	✓		✓	4,001	3,705	3,808	3,726
✓	✓			5,627	3,638	**3,653**	3,721
✓		✓	✓	4,512	4,399	4,851	4,741
✓		✓		4,462	4,319	4,978	4,744
✓			✓	4,038	3,565	3,728	4,748
✓				16,01	3,639	3,655	4,748
	✓	✓	✓	4,43	4,372	4,649	4,35
	✓	✓		4,352	4,352	4,967	4,357
	✓		✓	**3,905**	3,743	3,897	**3,646**
	✓			12,37	3,583	3,676	3,648
		✓	✓	4,443	4,309	4,554	4,631
		✓		4,352	4,353	4,961	4,634
			✓	3,944	3,653	3,822	4,653
				15,978	3,979	4,077	4,652

- $RMSE_{relevantfeatures,Ridge,minmax,pca} = 3,905$.
  Aplicación del algoritmo lineal Ridge sobre las variables relevantes cuando se aplica normalización *minmax-scaler* y reducción dimensional *pca-decomposition*,

- $RMSE_{relevantfeatures,XGB,probabilistic,minmax} = 3{,}653$.
  Aplicación del algoritmo no-lineal XGBoost sobre las variables relevantes cuando se aplica limpieza probabilística (*probabilistic-data-cleaning*) y normalización *minmax-scaler*,

- $RMSE_{relevantfeatures,SVR,minmax,pca} = 3{,}646$.
  Aplicación del algoritmo no-lineal SVR sobre las variables relevantes cuando se aplica normalización *minmax-scaler* y reducción dimensional *pca-decomposition*, y

- $RMSE_{relevantfeatures,RF,probabilistic,pca} = 3{,}565$.
  Aplicación del algoritmo no-lineal RF sobre las variables relevantes cuando se aplica limpieza probabilística (*probabilistic-data-cleaning*) y reducción dimensional *pca-decomposition*.

Los resultados muestran que las mejores soluciones se obtienen en la aplicación de algoritmos de ensamblaje no lineales (XGB, SVR y RF). Concretamente el modelo basado en arboles RandomForest es la solución óptima y la única que no precisa del escalado de las variables de entrada mediante técnicas de escalado *minmax-scaler*. Esto se debe a que, para RF, al contrario que el resto de algoritmos de ensamblaje, no asume que las variables de entrada siguen una distribución normal. Es por esta asunción que el resto de los algoritmos de ensamblaje requiere la transformación monotónica de las variables mediante técnicas *minmax-scaler*. El caso de XGBoost es un poco peculiar, ya que al ser un algoritmo de boosting basado en árboles no debería requerir ningún escalado. Sin embargo, al optimizar la función objetivo mediante el método gradiente, el escalado tiende a mejorar el resultado, como evidencia la Tabla.

Asimismo, se cuantifica la idoneidad de utilizar técnicas de selección de características relevantes, frente utilizar todas las características del proceso. La tabla 3.2 compara ambos enfoques, y muestra que la correcta selección de las entradas minimiza el valor de la métrica de error RMSE. El valor de la métrica RMSE es mejor/menor al aplicar técnicas de pre-procesamiento sobre las variables relevantes, que la mejor de las soluciones sobre la totalidad de las variables del proceso, siendo:

$$RMSE_{relevantfeatures,RF,probabilistic,pca} = 3,565$$
$$<$$
$$RMSE_{allfeatures,SVR,probabilistic,pca} = 4,836$$

Técnicas	Técnicas de pre-procesamiento				Algoritmos			
	probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	Ridge	RF	XGB	SVR
Relevant features	✓			✓	4,038	3,565	3,728	4,748
All features	✓			✓	4,888	4,928	5,162	**4,836**

Tabla 3.2: Comparación de la métrica RMSE: técnicas de pre-procesamiento, entradas todas/relevantes y estrategia EW.

El estudio replica el análisis anterior realizado con la estrategia EW, y lo aplica sobre la estrategia MW. La Tabla 3.3 recoge los resultados de aplicar las técnicas de pre-procesamiento y reducción dimensional, en la aplicación del algoritmo RandomForest con la estrategia MW, cuando se elige seleccionar las variables relevantes y cuando se elige trabajar sobre todas las variables de entradas. La mejor solución se obtiene cuando se aplica el algoritmo RandomForest sobre todas las características, normalización *minmax-scaler* y reducción dimensional *pca-decomposition* ($RMSE_{allfeatures,RF,minmax,pca} = 4,599$). Está solución es aún mejor que cuando sólo se aplican las técnicas sobre las variables relevantes, siendo $RMSE_{relevantfeatures,RF,minmax,pca} = 4,653$.

Pre-processing techniques				Algorithms	
probabilistic-data-cleaning	minmax-scaler	savgol-filter	pca-decomposition	RF relevant features	RF all features
✓	✓	✓	✓	5,216	5,026
✓	✓	✓		4,996	4,948
✓	✓		✓	4,743	4,619
✓	✓			4,742	4,804
✓		✓	✓	5,140	5,004
✓		✓		4,996	4,948
✓			✓	4,944	5,070
✓				4,742	4,804
	✓	✓	✓	5,288	4,933
	✓	✓		4,898	4,852
	✓		✓	**4,653**	**4,599**
	✓			4,682	4,812
		✓	✓	5,028	4,938
		✓		4,898	4,851
			✓	4,864	4,996
				5,079	5,012

Tabla 3.3: Comparación de la métrica RMSE: técnicas de pre-procesamiento, entradas relevantes y estrategia de MW.

3.3.3. Solución óptima

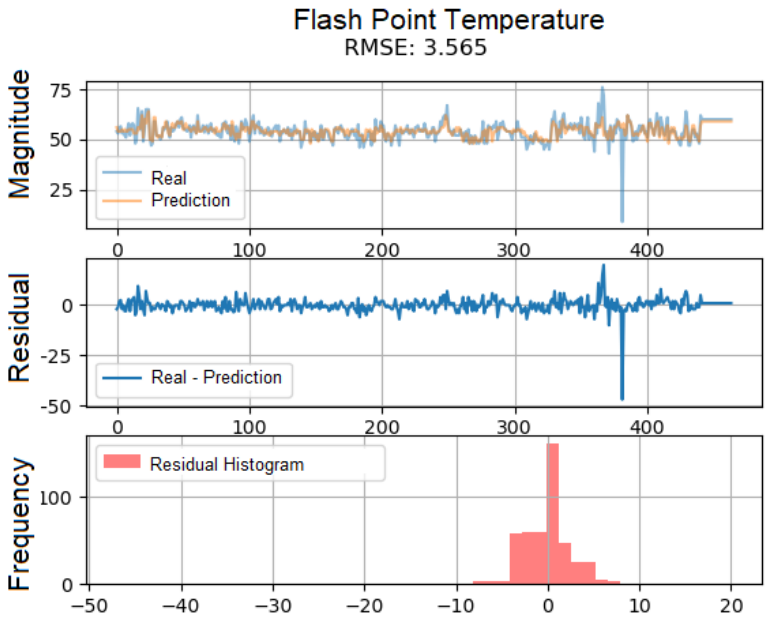
De los estudios anteriores se concluye que la solución óptima es la que se obtiene al aplicar una estrategia EW, donde en cada ventana se trabaja con el algoritmo no-lineal RandomForest. El algoritmo actúa únicamente sobre las variables relevantes del proceso y previamente se aplican técnicas de pre-procesamiento de limpieza probabilística (*probabilistic-data-cleaning*) y reducción dimensional (*pca-decomposition*). En este caso, el valor de la métrica RMSE es el siguiente:

$$RMSE_{relevantfeatures,RF,probabilistic,pca} = 3{,}565.$$

La Figura 3.10 muestra visualmente esta solución óptima:

- los valores diarios de FP (color azul) y los valores estimados de FP en esos mismos instantes (color naranja),
- el error residual, calculado como $Residual = Real - Prediccin$, y
- el histograma del error residual residual representado por una distribución prácticamente gaussiana (valor medio 0 y varianza mínima).

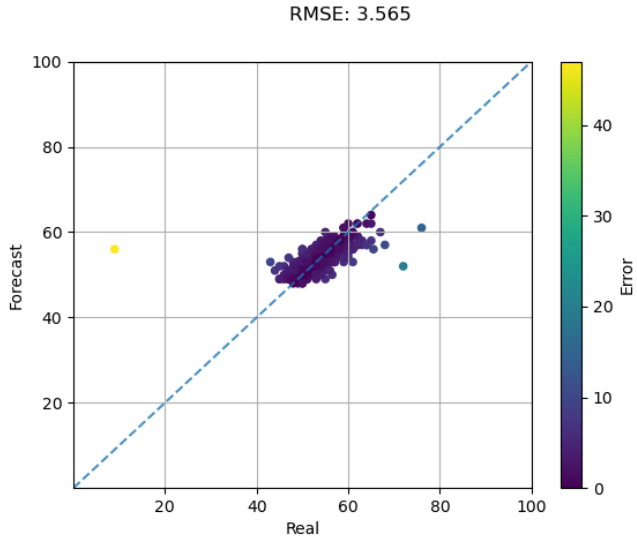
Figura 3.10: Estimaciones y residuos de la estimación de FP. **a)** Valor real diario y estimado. **b)** Error residual. **c)** Histograma del residuo,



Asimismo, la Figura 3.11 muestra el gráfico de dispersión entre el valor real FP y el valor de la predicción. Los puntos se distribuyen en torno a

la línea de regresión, normalmente distribuidos. Pero dado que la tipificación diaria de la variable de laboratorio es manual, es también propensa a errores humanos. Diariamente se entrena el modelo con la llegada de cada nueva medición de laboratorio, y se comprueba el error entre la predicción de salida y la medición real de laboratorio. Si entre estos valores el error supera un determinado umbral, ese día no se vuelve a entrenar el modelo y se elimina esa instancia. Este es el caso de los tres puntos de la Figura 3.11 que se identifican en el diagrama de dispersión, un punto amarillo y dos puntos verdes que destacan sobre el resto. El punto amarillo tiene un residuo negativo, donde el valor real es inferior al valor de predicción, lo que sugiere que se trata de un fallo en la medición de la variable real, básicamente porque no hay ninguna otra medición real con valores en ese rango (aprox. 20ºC). Los dos puntos verdes tienen un pequeño residuo positivo: su valor real es mayor que el valor de predicción, y al igual que en el caso anterior, esto bien podría deberse a un fallo de medición de la variable de laboratorio, ya que no se conocen casos de mediciones reales con temperaturas superiores a 70ºC. Si eliminamos estos puntos del conjunto de entrenamiento y evitamos que el modelo aprenda de ellos, el ajuste del modelo será más preciso y representativo de la realidad. Estos puntos se consideran valores atípicos.

Figura 3.11: Gráfico de dispersión: FP vs valor de predicción)



De los estudios anteriores se extraen las siguientes conclusiones:

- Se observa que un modelo de estrategia no adaptativo no es adecuado debido a la aparición de nuevos comportamientos y a su incapacidad para adaptarse a estos nuevos comportamientos. Por lo tanto, se concluye que es necesario trabajar con una estrategia de EW que permita el aprendizaje de nuevos comportamientos.

- La relación entre las variables relevantes y FP es multivariable y evoluciona con el tiempo. En este estudio, un enfoque de estrategia EW es la mejor manera de generar el modelo de inferencia del proceso: la diferencia entre la estrategia EW y los diferentes algoritmos de conjunto es mínima ($RMSE_{relevantfeatures,Ridge,minmax,pca} = 3{,}905$, $RMSE_{relevantfeatures,XGB,probabilistic,minmax} = 3{,}653$ y $RMSE_{relevantfeatures,SVR,minmax,pca} = 3{,}646$), obteniendo el mejor resultado con RandomForest ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3{,}565$).

- El estudio muestra mejores resultados cuando se realiza una selección de las características relevantes con la ayuda de los expertos del dominio que cuando se trabaja con la totalidad de las variables.

- Los resultados obtenidos son mejores en la estrategia EW que en la estrategia MW ($RMSE_{relevantfeatures,RF,probabilistic,pca} = 3{,}565$ vs $RMSE_{allfeatures,RF,minmax,pca} = 4{,}599$).

3.3.4. Cuantificación económica del ahorro de costes

La variable de calidad temperatura Flash indica la calidad del producto a la salida del proceso de desulfuración. Actualmente, se obtiene diariamente en un proceso de laboratorio [225]. Una vez conocido el valor de la temperatura de inflamación, el operador modifica el resto de las variables del proceso para obtener un valor de temperatura que cumpla las especificaciones. Si el valor de la temperatura es superior a las especificaciones, repercute en un sobrecoste innecesario para la refinería: se produce un producto de mayor calidad que no se traduce en mayores beneficios, ya que se mantiene el precio de venta. Este producto de mayor calidad implica un sobrecoste indirecto en el proceso, por ejemplo, operar a temperaturas más altas acorta el ciclo de vida de los catalizadores utilizados en la reacción de desulfuración. En este proceso de desulfuración y, en general, en los procesos de gran volumen de las refinerías, la limitación de no disponer de información en tiempo real provoca grandes impactos económicos. Cuanto más ajustadas estén las variables del proceso, menos se desviarán de la calidad especificada. El sensor virtual proporciona a los operadores de la refinería información en tiempo real para ajustar las condiciones de funcionamiento, maximizando la estabilidad de la unidad de desulfuración y produciendo gasóleo conforme a las especificaciones.

Según las estimaciones de la refinería y con el procedimiento actual, el 50 % de las veces, a pesar del retraso entre la toma de muestras y la obtención de las medidas experimentales en el laboratorio, se proporciona información realista sobre la temperatura del proceso en curso. Pero en el 12 % de los casos, existe una ligera incertidumbre en la información del valor de la temperatura, y en el 38 % de los casos, no se sabe si la información de la temperatura es suficientemente realista. Cuando la información no es suficientemente realista (38 %), deja de generar beneficios por valor de 497306$/*semestre*, y cuando la información es ligeramente incierta (38 % + 12 %), deja de generar beneficios por valor de

654350$/*semestre*. El sensor virtual propuesto en este estudio proporciona mediciones fiables de la temperatura del Flash 94 % del tiempo de funcionamiento. Así, se estima una reducción considerable de las pérdidas generadas, de 497306$/*semestre* a 29838$/*semestre*, en el primer escenario donde la realidad representada no es suficientemente realista, mientras que en el segundo escenario la reducción pasa de 654350$/*semestre* a 39261$/*semestre*.

3.4. Otros Casos Investigación Industrial

Esta sección trata de demostrar la capacidad de generalización del procedimiento definido, a través de otros dos casos de investigación industrial real. Ambos son casos de uso singulares, muy diferentes entre sí en términos de ámbito de aplicación, pero muy semejantes en relación a las técnicas que se aplican para su resolución: técnicas de pre-procesamiento, técnicas de selección de variables relevantes, técnicas de reducción dimensional, estrategias de aprendizaje dinámico y validación.

El primer caso de uso se trata de un proceso continuo en una planta química y de las relaciones que se producen en el proceso del cracking del etileno de un horno de nafta. Consiste en, dados unos determinados componentes químicos, inferir el valor de cuatro salidas del proceso (dihidrogeno, metano, etileno y propileno) en tiempo real.

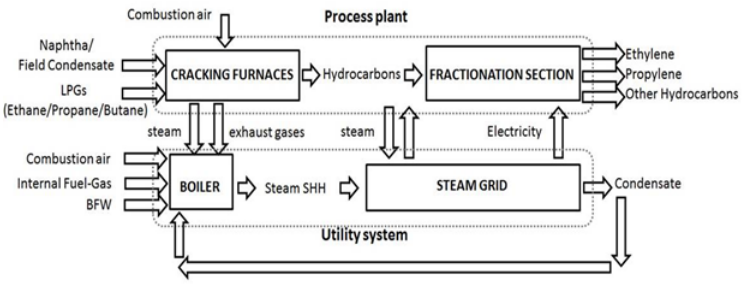
El segundo caso de uso se trata de una planta de reciclaje y de las emisiones de partículas contaminantes a la atmósfera que se provocan por el movimiento de cargas en exteriores. Consiste en dadas unas determinadas condicionantes meteorológicas predecir cuales de esas emisiones afectaran a las áreas aledañas en forma de contaminación de partículas en suspensión.

3.4.1. Planta Química: Inferencia de componentes químicos

En el proceso del craqueo al vapor, una alimentación de hidrocarburos gaseosos o líquidos como la nafta, se diluye con vapor y se calienta brevemente en un horno en ausencia de oxígeno. La reacción se produce rápidamente. Una vez alcanzada la temperatura de craqueo, el gas se apaga rápidamente para detener la reacción en un intercambiador de calor de la línea de transferencia o en el interior de un cabezal de enfriamiento con aceite de enfriamiento [226]. Los productos producidos en la reacción dependen de la composición de la alimentación, de la relación hidrocarburo/vapor, de la temperatura de craqueo y el tiempo de permanencia en el horno. Los hidrocarburos ligeros alimentados, como el etano o la nafta ligera, dan principalmente alquenos ligeros, como el etileno, el propileno, el metano y el dihidrógeno. El proceso también da lugar a la lenta deposición de coque, una forma de carbono, en las paredes del reactor. Esto degrada la eficiencia del reactor, por lo que las condiciones de reacción se diseñan para minimizarlo. No obstante, un horno de craqueo al vapor sólo puede funcionar durante unos meses entre descoque y descoque. La formación de coque durante el procesamiento de hidrocarburos a gran escala es en

gran medida inoportuna, y no solo representa una pérdida de producto, sino también provoca la disminución de la transferencia de calor [227]. Las descoquizaciones requieren que el horno se aísle del proceso y que se haga pasar un flujo de vapor o una mezcla de vapor y aire a través de los serpentines del horno. Una vez completada esta reacción, el horno puede volver a funcionar [228]. La Figura 3.12 muestra este proceso de craqueo al vapor en un sistema alimentado por nafta ligera.

Figura 3.12: Proceso de craqueo al vapor en un sistema alimentado por nafta ligera



Dataset

El estudio se centra en un horno de nafta para el proceso de cracking del etileno, donde a partir de las variables minútales de entrada se ha de predecir el valor de las variables minutales de salida. El conjunto de variables de entrada está formado por 30 variables de temperaturas, presiones y densidades seleccionadas por los expertos de dominio y con profundidad histórica de 1 año. Los datos proceden de 2 secciones de proceso de almacenamiento diferentes de la planta petroquimica. El conjunto de las variables de salida está formado por 4 variables de proceso: dihidrógeno (H2), metano (CH4), etileno (C2H4) y propileno (C3H6). Inicialmente los valores reales de estos 4 componentes químicos se obtienen en un proceso de laboratorio -con el consiguiente retardo- a través de un dispositivo capaz de determinar las masas moleculares y la composición química de las muestras (espectógrafo).

Validación cruzada

En el presente estudio se utiliza la técnica de validación cruzada de series temporales divididas (split time series cross validation) en las sucesivas ventanas móviles, y el valor de NRMSE (RMSE normalizado) como medida de la precisión de las técnicas de inferencia del sensor virtual, donde los valores de NRMSE más bajo indican una mayor precisión.

Metodología

Se demuestra la capacidad de generalización del procedimiento definido, a través de la realidad de una importante empresa química internacional.

La empresa dispone en una de sus instalaciones de una planta de cracker de nafta. El estudio persigue desarrollar:

- un sistema de monitorización y control no invasivo dedicado a la predicción de variables de proceso que, aunque se conozcan sus medidas, éstas tienen cierto retardo temporal asociado, y
- una herramienta de detección de faltas en el proceso, en el caso de que se produzca un fallo ocasional en un sensor.

El pre-procesamiento de datos implica la transformación del conjunto de datos en bruto en un formato adecuado para la interpretación de los algoritmos y para la mejora de su eficacia. La tarea de pre-procesamiento de los datos originales consiste en:

- Eliminar medidas atípicas. Se estudia individualmente el histórico de cada uno de los 30 analizadores que aporta información de las variables de entrada para la inferencia del modelo, y se eliminan todas aquellas medidas que no representen el 95 % de las medidas del analizador a través de las técnicas de limpieza probabilística de datos [181].
- Eliminar ruido de las variables de entrada. Se utilizan técnicas de procesamiento de señal para minimizar el efecto no deseado del ruido aleatorio y eliminar perturbaciones, a través del suavizado con la derivación de Savitzky-Golay [187].
- Evitar problemas de redundancias y colinealidad entre variables de entrada. Se simplifica la complejidad del espacio muestral mediante la técnica de análisis de componentes principales (PCA) para evitar problemas con las variables de entrada redundantes. Se aplica iterativamente la técnica PCA hasta que las nuevas entradas ficticias sean capaces de justificar el 95 % de la varianza explicada de la muestra original. En el caso que aplica, el número óptimo de componentes es de 4, véase 3.13.

Los métodos de pre-procesamiento de datos afectan directamente a los resultados de cualquier algoritmo analítico. Pero, es importante recordar que tan importante como esta tarea de pre-procesamiento, es comprender la operativa del proceso. En este caso, la relación estrecha con los técnicos expertos de planta y su ayuda permite reconocer los periodos de deposición de coque en los que la información que proporcionan los sensores de las variables de entrada es información no esencial y como tal, debe ser eliminada de la muestra. Una vez explicado por parte de los técnicos y conocido el efecto de descoque, es sencillo identificarlo en la Figura 3.14, ya que se corresponde con todos aquellos instantes en los que las medidas pierden su dinámica de normalidad y cambian su forma de operar, para caer hasta valores inferiores a 10 unidades de magnitud - en el caso del metano, etano y propileno - y cercanos a 20 unidades en el caso del dihidrógeno.

El resto de la metodología consiste en procedimentar un conjunto de pruebas que permita establecer la configuración más eficiente en términos

Figura 3.13: Estudio de varianza acumulada en base a componentes principales

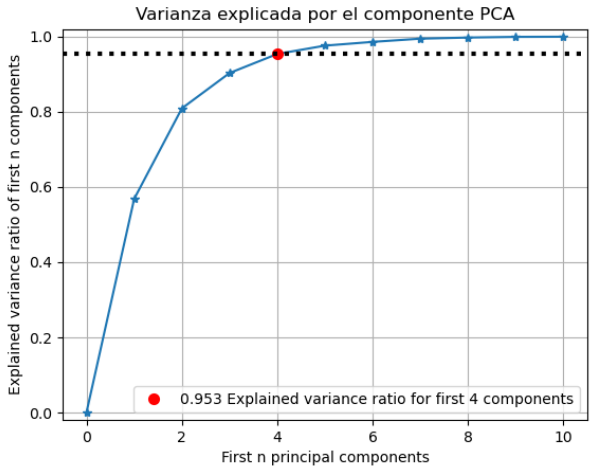
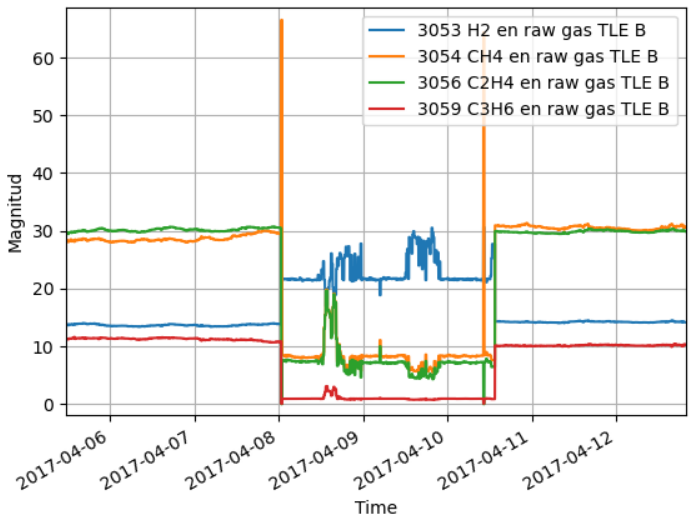


Figura 3.14: Proceso continuo de sensorización de variables de entrada con comportamiento de decoque



de minimización del error de la predicción del sensor virtual para el proceso de inferencia.

- Pruebas para definir el algoritmo de inferencia regresivo más adecuado.

  La Tabla 3.4 muestra los resultados de aplicar diferentes algoritmos de inferencia regresivos sobre el concepto de estrategia de aprendizaje de ventana móvil (MW). El tamaño de la ventana por defecto es de 50 muestras, y los algoritmos utilizados para la comparativa de los métodos son:

  - Algoritmo SVR, como algoritmo regresivo habitual en la resolución de sensores virtuales,
  - Algoritmo XGBoost, como algoritmo de ensamblaje de técnicas de boosting, y
  - Algoritmo RandomForest, como algoritmo de ensamblaje de árboles aleatorios con técnicas de bagging.

  Los valores de la tabla estiman que los mejores resultados se obtienen con el algoritmo de inferencia RandomForest para la totalidad de los 4 componentes químicos.

Tabla 3.4: Comparativa del error NRMSE para diferentes algoritmos de inferencia regresivos con MW y tamaño de ventana 50

NRMSE (MW=50)	RF	XGB	SVR
Dihidrógeno	**0,0037**	0,13	0,19
Metano	**0,005**	0,29	0,34
Etileno	**0,0038**	0,34	0,28
Propileno	**0,0089**	0,41	0,42

- Pruebas para definir el tamaño de ventana más adecuado. La calidad de la predicción varía en función del tamaño de la ventana de entrenamiento. La Tabla 3.5 muestra el valor de las comparaciones de aplicar el algoritmo regresivo RandomForest sobre diferentes tamaños de ventana en la estrategia de aprendizaje MW. Se presentan por separado para cada una de las 4 salidas del modelo junto con el error medio cuadrático normalizado (NMRSE). A modo de experimentación, se seleccionan cuatro tamaños de ventana diferentes (50, 100, 150 y 200) para la realización de los estudios. Cada ventana está formada por valores 10 minutales, de manera que una ventana de tamaño, donde $MW = 50$, implica 50 muestras tomadas cada 10 minutos y por tanto, en un rango de $50 * 10$ minutos de tiempo.

  Los valores de la tabla estiman que los mejores resultados se obtienen con tamaño de ventana de 150 para el propileno y con tamaño de

Tabla 3.5: Comparativa del error NRMSE para diferentes tamaños de ventana en algoritmo RF con MW

NRMSE	MW=50	MW=100	MW=150	MW=200
Dihidrógeno	**0,0037**	0,0054	0,005	0,0086
Metano	**0,005**	0,0067	0,0076	0,0092
Etileno	**0,0038**	0,0079	0,008	0,0075
Propileno	0,0089	0,0084	**0,007**	0,015

ventana 50 para el resto de los componentes químicos de salida del proceso (dihidrógeno, metano y etileno). Esta diferencia está relacionada con la temperatura de craqueo de los diferentes componentes en el proceso físico, pero que las técnicas analíticas abstraen e infieren de forma transparente.

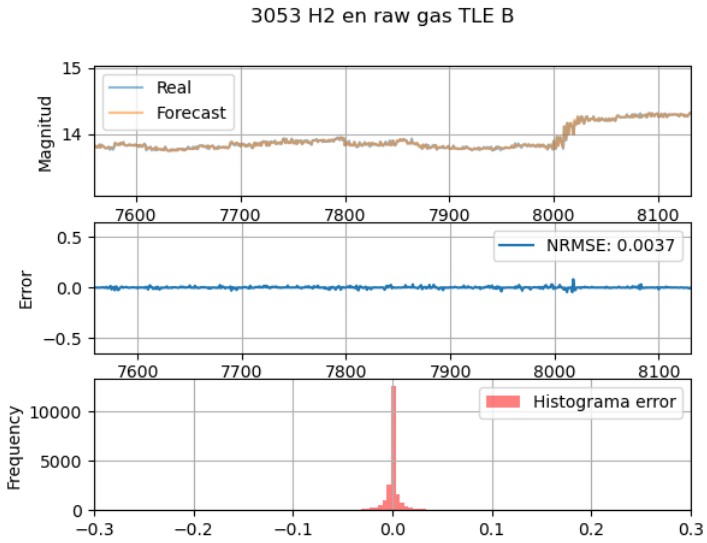
En base al error NRMSE estimado en la Tabla 3.4 y en la Table 3.5 el mejor valor de inferencia para cada componente químico se obtiene con la siguiente configuración:

- Dihidrógeno: algoritmo de inferencia RF mediante estrategia de aprendizaje MW y tamaño de ventana 50 muestras.
- Metano: algoritmo de inferencia RF mediante estrategia de aprendizaje MW y tamaño de ventana 50 muestras.
- Etileno: algoritmo de inferencia RF mediante estrategia de aprendizaje MW y tamaño de ventana 50 muestras.
- Propileno: algoritmo de inferencia RF mediante estrategia de aprendizaje MW y tamaño de ventana 150 muestras.

El resultado es un sensor virtual por cada uno de los cuatro componentes químicos, que infiere la relación entre las variables de entrada y la correspondiente salida en el ámbito de cada ventana. El modelo de cada sensor se reconstruye con cada nueva instancia. A cada paso que se mueve la ventana, se procesan las variables de entrada con las técnicas de preprocesamiento y se actualiza el conjunto de datos de entrenamiento. Como ejemplo, la Figura 3.15 compara los valores reales de Dihidrógeno y los valores de la predicción, el error punto a punto de cada uno de los valores de validación y el histograma del error. El valor $NRMSE = 0{,}0037$, un valor próximo a cero y extremadamente pequeño, es indicador del buen hacer del sensor virtual.

El análisis de las correspondencias entre los valores reales del espectógrafo y los valores de inferencia de los sensores virtuales se obtiene a través de la representación gráfica del diagrama de dispersión. La Figura 3.16 representa la correspondencia de los valores en forma de puntos por cada sensor virtual. El color de los puntos cuantifica la magnitud del error y la recta discontinua muestra la situación ideal en la que predicción y realidad coinciden. Cuanto mayor sea el número de puntos sobre la diagonal

Figura 3.15: Estudio del error del sensor virtual para Dihidrógeno con tamaño de ventana 50 en algoritmo RF con MW



mejor es la predicción. El valor NRMSE es próximo a cero para todos los componentes (véase la Tabla 3.5) y por ello, la densidad de puntos sobre la diagonal es alta.

Especial atención merece la gráfica de metano (CH4) que representa varios puntos desviados (en color amarillo y verde). Dado que el error cometido en este sensor virtual es relativamente pequeño y próximo a cero ($NRMSE = 0{,}005$) esta desviación hace suponer que puede tener dos motivos de ser: (i) funcionamiento sesgado del sensor virtual, o (ii) una medida fallida en el dispositivo de medición real. No es de extrañar que en las condiciones de trabajo extremas en las que funcionan determinados sensores reales ocurran fallos de medición [29]. En este caso, la definición de un determinado umbral entre el valor de medición real y el valor de la predicción permite identificar medidas reales anómalas e implementar el sensor virtual como herramienta de detección de faltas. Esta herramienta de detección de faltas permite a la planta petroquímica corregir valores erróneos del analizador del proceso, en este caso, del metano.

La predicción fiable de valores en tiempo real para este problema multivariante no-lineal es extremadamente beneficiosa para la empresa petroquímica. La empresa reveló que le había permitido identificar los puntos más ineficientes del proceso (45 % de las pérdidas económicas totales y el 55 % de las pérdidas energéticas totales). Esta identificación y actuación sobre estos puntos tiene un potencial de ahorro estimado en 3 M€por año en el proceso del cracker de la planta mediterránea. Finalmente, el modelo de inferencia del sensor virtual se encapsula como servicio (Model as a Service, MaaS) y se alberga en un servidor de producción, véase Figura 3.17.

Figura 3.16: Diagrama de dispersión de componentes químicos en horno de nafta

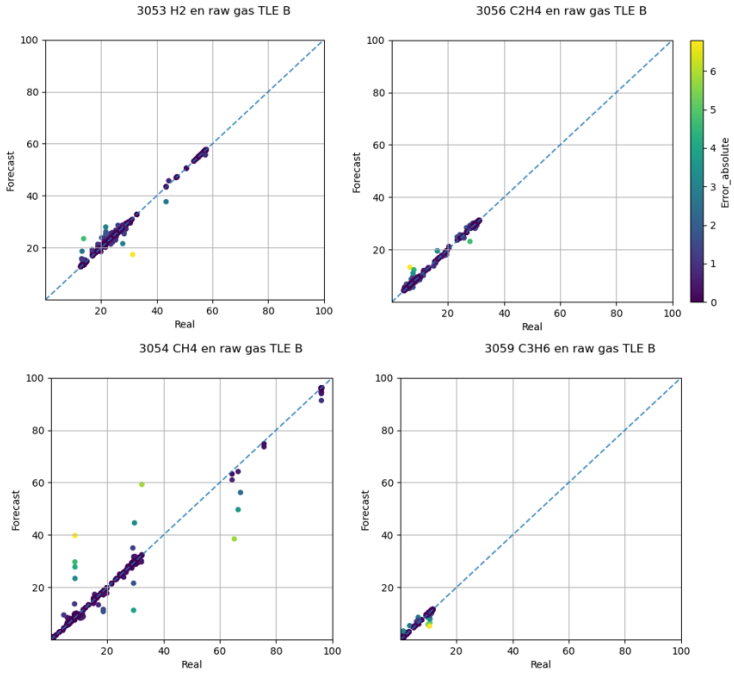
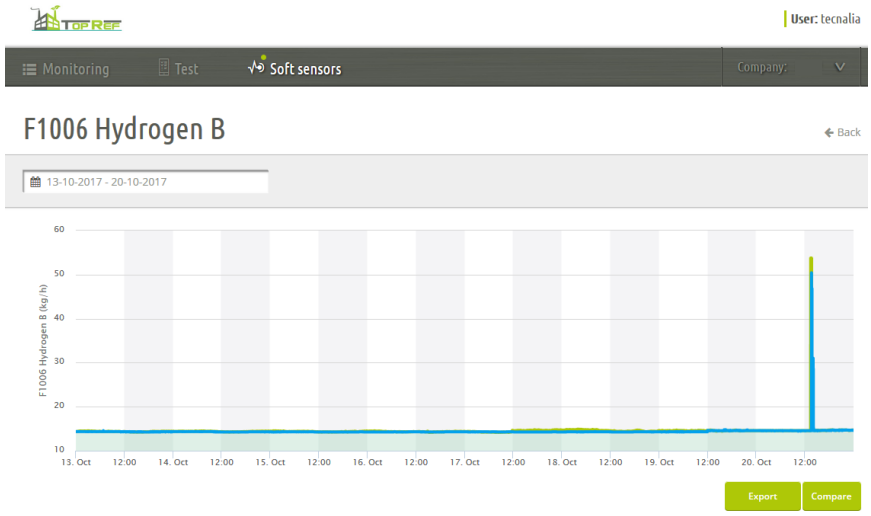


Figura 3.17: Modelo de inferencia para dihidrógeno en proceso de cracking del etileno en producción



3.4.2. Planta de Reciclaje: Inferencia de emisiones contaminantes

La lucha contra el cambio climático y la descarbonización de la economía son retos importantes que abordar en las próximas décadas. El dióxido de carbono puede ser el mayor impulsor del calentamiento global y del cambio climático, pero no es el único [229]. Otros compuestos gaseosos y de partículas afectan con igual o mayor afección sobre muchos de los procesos climáticos [230]. En lo que a las emisiones de partículas difusas se refiere, tradicionalmente, se les ha prestado poca atención, tanto por la ausencia de normativa legal como por la escasez de estudios técnicos existentes, pero en determinados casos provocan un impacto significativo en áreas aledañas a los focos de emisión, como pueden ser áreas urbanas u otras industrias. En ocasiones, estos impactos acarrean denuncias y quejas por parte de la población y otras empresas debido a los problemas generados: dificultades para respirar, la suciedad por depósito de partículas, afección a procesos industriales, salud laboral de los trabajadores, daños sobre bienes (corrosión de vehículos, por ejemplo), etc.

Históricamente la evaluación de la calidad del aire se ha basado en el análisis de los datos procedentes de las estaciones fijas, las cuales, en muchos casos, no aportan una resolución espacial suficiente. Medir los distintos contaminantes en un foco confinado es una técnica que lleva muchos años implantada y no reviste mayores complicaciones. Ahora bien, determinar las emisiones difusas en una gran superficie con procesos de emisión discontinuos requiere de un tratamiento más complejo [231]. Plantear mediciones experimentales de calidad del aire para conocer en detalle las variaciones espaciales con los aparatos/sistemas adecuados sería una tarea costosa y que se alargaría excesivamente en el tiempo. Para evitar esta tarea existen modelos de dispersión de contaminantes especialmente diseñados para valorar la calidad del aire y que ofrecen una serie de ventajas. Entre ellas se enumeran las más importantes:

- La cobertura espacial de estaciones para evaluar la calidad del aire es extremadamente limitada. La modelización puede proporcionar una cobertura espacial mucho más extensa y detallada.
- La modelización da resultados que se puede aplicar como pronóstico. Además, puede ser utilizado para predecir la calidad del aire como resultado de cambios en emisiones o condiciones meteorológicas.
- La modelización proporciona un mayor conocimiento de las causas y relaciones que determinan la calidad del aire.

Además, con respecto a las directivas, también la utilización de modelos presenta una serie de ventajas, así:

- Los modelos pueden proporcionar información de la calidad del aire en zonas donde no se realizan medidas.
- El número de estaciones de medida se puede reducir perceptiblemente con la considerable reducción de coste que ello significa.

- Los modelos se pueden utilizar para desarrollar políticas para mejorar la calidad del aire.

La modelización, sin embargo, no proporciona todas las respuestas y hay un número de limitaciones inherente a ella entre las que se puede destacar:

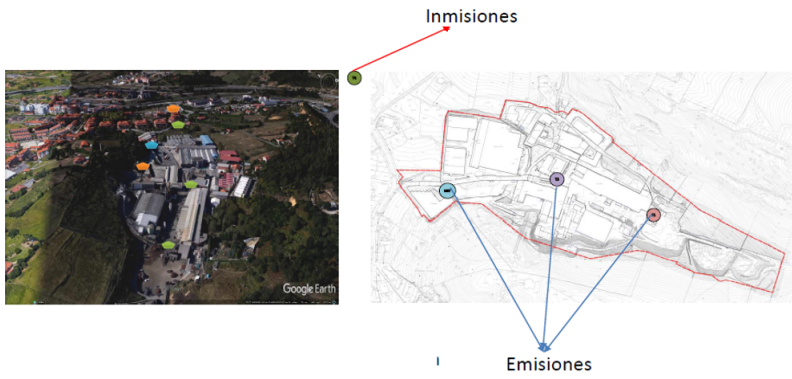
- Los modelos requieren del conocimiento de las emisiones y de la meteorología como datos de entrada, que no son siempre confiables o se pueden adquirir fácilmente.
- Los modelos siguen siendo inciertos en sus predicciones y se requiere una extensa validación antes de que puedan ser aplicados y tengan un margen aceptable de error.
- La capacidad de modelos para representar el mundo real es limitada, ya que siguen siendo una aproximación a la realidad.

Por todo ello, uno de los problemas que se plantea en el control de las emisiones difusas de partículas es la cuantificación de las mismas y el análisis de los posibles impactos que se pudieran generar por la actividad industrial. Ante este escenario, la transformación digital contribuye a mejorar sustancialmente el control de las emisiones difusas de partículas y actúa como catalizadora de nuevas dinámicas y modelos de negocio para la actividad industrial. Los sensores virtuales, explicados anteriormente, son capaces de completar la información que se obtiene de los equipos físicos instalados en estas redes de vigilancia, aumentando la resolución espacial.

En los últimos tiempos se han empezado a utilizar métodos de inferencia máquina para la predicción de contaminantes[232], [233], ya que son métodos con los que se pueden obtener muy buenos resultados para realizar predicciones de corto y largo plazo en tiempo real, en donde se cuenta con información de concentraciones y sus tendencias, aunque poseen limitaciones por cuanto no se pueden establecer relaciones causa-efecto. Estos esquemas están basados en técnicas no paramétricas capaces de analizar tendencias, relaciones de la calidad del aire y las mediciones atmosféricas, y para predecir la evolución de situaciones de contaminación.

Este caso de uso hace referencia a la realidad de una empresa de reciclaje y refinado, en la que la actividad que en ella se realiza - reciclar materiales complejos no férricos - genera emisiones por el movimiento de cargas a granel. La solución aborda la problemática de la dispersión de partículas PM_{10} (partículas con un tamaño inferior a 10 micras) , con idea de proporcionar un sistema automático no-supervisado de identificación de eventos de inmisión (instante puntual de contaminación en zona aledañas debido a emisiones) dadas las condiciones de los focos de emisión y ciertas variables meteorológicas. Se trata de lograr un sensor virtual que, instalado en el foco de emisión permita, en tiempo real y sin sensores de inmisión, determinar si una determinada emisión se refleja o no en inmisión. La Figura 3.18 muestra dos imágenes, la primera es una imagen fotográfica de la planta de reciclaje y de la ubicación de los sensores físicos para la medida de concentraciones. Y la segunda, el plano que detalla los puntos

Figura 3.18: Imagen fotográfica de la Planta de Reciclaje y plano de la ubicación de los sensores físicos



de medida de emisión e inmisión. Los sensores de emisión, por definición, están dentro del perímetro de la planta de reciclaje y el sensor de inmisión en zonas aledañas.

El sistema propuesto en comparación con las metodologías tradicionales utilizadas en la actualidad aporta numerosos beneficios, entre los que cabe destacar:

- Sistema totalmente autónomo.
- Sistema cuyo tiempo de procesado y necesidades computacionales son reducidas.
- Creación de una red de sensores económica, independiente y con mantenimientos reducidos.
- Sistema de fácil implementación en cualquier instalación.
- No requiere de un conocimiento experto por parte del usuario.
- No requiere de un conjunto de datos históricos etiquetados para los valores de inmisión.

Por último, a partir del impacto de los procesos propios de su actividad que se generan en las zonas aledañas a sus instalaciones, la empresa:

- Dispone de una evaluación continua de inmisiones (un resultado cada 10 minutos), lo que permitirá disponer de suficiente tiempo de reacción para evitar, y/o aplicar mejores técnicas disponibles de reducción de emisiones, si fuera el caso, en el proceso que esté provocando una alarma.
- Justifica ante un tercero que su instalación no es la responsable de los malestares que generados por el polvo en suspensión.

Dataset

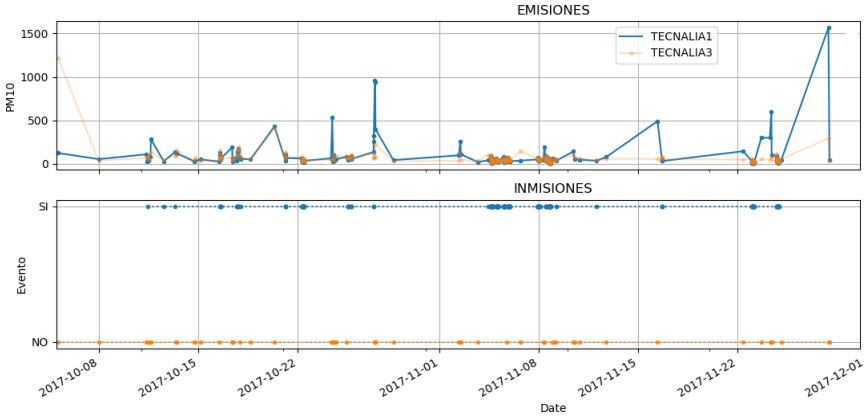
El estudio se centra en el análisis de los datos de los sensores físicos en los

puntos de captación de emisiones e inmisiones. Se trata de datos relativos a la concentración de partículas PM_{10} y a las condiciones meteorológicas (temperatura ambiente, humedad ambiente, presión ambiente, punto de rocío, dirección y velocidad del viento). Medidas obtenidas in-situ por 2 sensores físicos ubicados en el interior de la planta de reciclaje (sensores de emisiones) y un sensor físico ubicado en el exterior de la planta de reciclaje (sensor de inmisión). El proceso de captación de información era continuo durante 100 días consecutivos tanto en los sensores de emisión como de inmisión, captándose la misma información en ambos casos.

Metodología

A priori se desconoce bajo qué condiciones ciertas emisiones de partículas alcanzan las zonas exteriores aledañas a la planta de reciclaje y cuáles de las inmisiones medidas se corresponden con emisiones realmente originadas en planta. El primer reto es convertir los datos de entrenamiento en datos etiquetados. Para ello, el técnico experto de proceso define un umbral que caracteriza el incremento de concentración de partículas que se considera emisión y el umbral de incremento para inmisión. La diferenciación de las series de las concentraciones de cada sensor físico en su primera derivada permite contrastar el incremento de partículas contra el umbral establecido. Aquellos valores que están por encima del valor umbral se etiquetan como evento emisión o evento inmisión.

Figura 3.19: Plano de puntos de medida de emisión-inmisión



La observación de los datos etiquetados y el estudio de correlación entre eventos de emisión y eventos de inmisión, aporta información del comportamiento del proceso. La Figura 3.19 muestra 3 comportamientos diferentes que se identificaron durante el estudio de correlación de emisiones e inmisiones: (i) eventos de emisión que SÍ se corresponden con eventos de inmisión de la fábrica de reciclaje, (ii) eventos de inmisión que NO se corresponden con eventos de emisión de la fábrica y que se presupone que

se asocian con terceras empresas, y (iii) eventos de emisión que NO tienen su correspondiente inmisión.

El estudio de las correlaciones de los eventos de emisiones en los sensores del interior de la planta de reciclaje con los eventos de inmisiones en el sensor del exterior de la planta permite crear un repositorio de datos de instantes de eventos emisión-inmisión (eventos de emisión que SÍ se corresponden con eventos de inmisión). Esta base de datos representa la presencia o no de partículas en inmisión tras una emisión, además de representar la salida esperada del sensor virtual. Las entradas están formadas por información de las condiciones meteorológicas de los dos sensores físicos del interior de la planta. El sensor virtual infiere la relación entre entradas y salida mediante la generación de un algoritmo de clasificación basado en técnicas de emsamblaje (algoritmo RandomForestClassifier).

Para evaluar los resultados y garantizar que la partición de datos de entrenamiento y validación son independientes, se aplican técnicas de validación cruzada. Estas técnicas permiten aprender la relación entre las variables de entrada y las de salida en un conjunto de datos balanceados de entrenamiento. Y aplicar esta relación en las variables de entrada de un conjunto de datos de validación, con objeto de inferir las variables de salida. La ventaja de este método es que es computacionalmente es muy rápido.

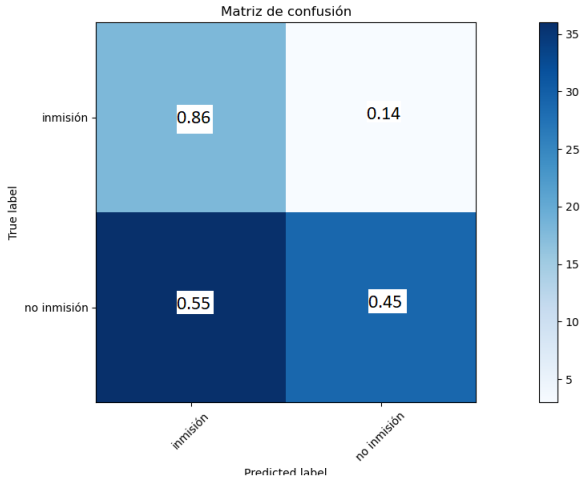
El desempeño del método de clasificación binaria se evalúa con la matriz de confusión y que describe cómo se distribuyen los valores reales de los de la predicción del sensor virtual. La diagonal principal, contiene las predicciones correctas y la otra refleja los errores cometidos por el clasificador. Según la Figura 3.20 el 86 % de los eventos de inmisión y el 45 % de los eventos de no inmisión se han clasificado correctamente. Por contra, el 14 % de las inmisiones han sido clasificadas erróneamente como no-inmisiones y el 55 % de las no-inmisiones han sido clasificadas erróneamente como inmisiones. En esta caso aunque la tasa de falsos positivos (eventos de no-inmisiones clasificadas erróneamente como inmisiones) es el 55 % no supone un problema en la realidad de la planta de reciclaje. La solución prioriza la clasificación correcta de las inmisones, y asume la tasa de falsos positivos como daños colaterales. En el caso que aplica, los falsos positivos no tienen más consecuencia que la puesta en marcha de los aspersores de regadío para depositar las partículas en suspensión en superficie, mediante emisión de agua.

No obstante, dada la orografía del terreno donde está ubicada la planta de reciclaje, ciertas situaciones excepcionales afectan a la relación entre emisiones e inmisiones. Estas situaciones son básicamente la aparición de condicionantes meteorológicos excepcionales como intrusiones saharianas de viento cálido del sur y la combustión de biomasa como forma operativa de la planta. Los instantes puntuales en los que estas situaciones ocurrieron son identificadas con ayuda de técnicos expertos y eliminadas del conjunto de datos de entrenamiento.

Los aspectos claves en la implementación del sensor virtual para la detección de contaminación urbana de partículas PM_{10}, son:

- Definición de evento de emisión, a partir de umbral parametrizable.

Figura 3.20: Matriz de confusión para clasificador binario de eventos de inmisión



- Definición de evento de inmisión, a partir de umbral parametrizable.
- Creación de base de datos etiquetada con eventos emisión-inmisión (aproximadamente 490 eventos).
- Generación de modelo de clasificación basado en técnicas de ensamblaje.
- Generación de un esquema de validación cruzada para generar resultados estadísticos de la validez del esquema.

Se crea un panel de control de vigilancia ambiental para monitorizar: (i) el nivel de concentración de las partículas PM_{10} medido por los sensores físicos del interior de la planta, (ii) las condiciones meteorológicas y (iii) un semáforo que indica el riesgo a sufrir inmisiones. Es decir, los resultados del modelo de inferencia del sensor virtual se encapsulan en un servicio con interfaz WEB que proporciona, al responsable de la planta de reciclaje, información de la concentración de partículas en el aire y la propensión a sufrir inmisiones en el corto plazo en tiempo real, véase Figura 3.21.

3.5. Conclusiones

En este capítulo se ha propuesto un procedimiento para crear modelos de inferencia, a partir de conjuntos de datos históricos, capaces de aprender las relaciones de causalidad multi-paramétricas y altamente no lineales de ciertos procesos industriales. A este concepto de inferencia de variables críticas difícilmente medibles en tiempo real, pero capaces de ser estimadas a partir de variables fácilmente medibles se ha hecho referencia en el presente Capítulo, que ha descrito y ejemplificado el procedimiento en los sensores virtuales adaptativos, analizando los siguientes aspectos clave:

Figura 3.21: Panel de control y de vigilancia ambiental en planta de reciclaje

- Técnicas de pre-procesamiento sobre los datos originales como: la eliminación de valores anómalos o no representativos de la normalidad del proceso, reducción del ruido de la señal de los sensores conocidos mediante filtros de suavizado de la señal y/o homogeneización del rango de magnitudes de las variables relevantes mediante normalización.

- Técnicas de selección de variables relevantes del proceso a través del conocimiento de los técnicos expertos de dominio y de técnicas de ponderación de pesos de variables, técnicas de permutación y/o secuenciación de algoritmos para la identificación de la importancia de las variables. Asimismo, para minimizar el efecto de la deriva de concepto sobre los cambios dinámicos del proceso y su afección sobre las variables relevantes en el tiempo se estudian técnicas de consistencia en el tiempo y validación adversaria.

- Técnicas de reducción de la dimensionalidad sobre las variables relevantes para mitigar el efecto de la redundancia y colinealidad de la información.

- Selección de la estrategia de aprendizaje, estudio del tamaño de ventana más adecuado en cada caso, así como de la selección del algoritmo óptimo.

- Selección de las métricas adecuadas para la correcta validación del modelo y estimación de la calidad de la solución propuesta.

Se ha validado la generalidad del procedimiento en la realidad de tres casos de uso singulares, muy diferentes entre si en términos de aplicación, pero muy semejantes con relación a las técnicas que se aplican para su resolución:

- Inferencia de la temperatura del punto de inflamación Flash en una Planta Petroquímica.

  Se ha implementado un sensor virtual adaptativo en una planta petroquímica para la inferencia de la temperatura del punto de inflamación en el proceso de desulfuración del gasóleo de automoción. El método ha incluido la aplicación de técnicas de selección de las variables más relevantes y técnicas de limpieza de datos probabilísticos y de descomposición PCA. La mejor solución se ha obtenido con la implementación del algoritmo RandomForestRegressor con estrategia de aprendizaje de ventana extendida EW. Se ha utilizado RMSE como métrica de la calidad de la solución, siendo $RMSE = 3,565$. Se ha conseguido valores 10 minutales óptimos de la temperatura de Flash el 94 % del tiempo de funcionamiento del estudio, cuando hasta entonces sólo se disponía de valores puntuales con frecuencia diaria y con varias horas de retraso.

- Inferencia de componentes químicos en una Planta Química.

  Se ha implementado un sensor virtual adaptativo en una planta química para la inferencia de 4 componentes químicos en el proceso del cracking del etileno. El método ha incluido la eliminación de medidas anómalas, eliminación del ruido de las variables representativas, y para evitar problemas de redundancia y colinealidad se ha implementado la técnica PCA. La mejor solución se ha obtenido con la implementación del algoritmo RandomForestRegressor con estrategia de aprendizaje de ventana móvil MW de tamaño 50 y 150 muestras respectivamente. Se ha utilizado NRMSE, como métrica de la calidad de la solución, al tratarse de un sistema altamente estable. Este estudio ha permitido a la Planta Química identificar el 55 % de las perdidas energéticas y reducir un 45 % las pérdidas económicas (se ha estimado un ahorro de 3M€/ año).

- Inferencia de emisiones contaminantes en Planta de Reciclaje.

  Se ha implementado un sensor virtual adaptativo en una planta de reciclaje para la inferencia de emisiones de partículas PM_{10} a áreas aledañas. La particularidad de este caso de uso ha sido la implementación de un método que estudia la correlación de eventos de emisiones en los sensores del interior de la planta de reciclaje con los eventos de inmisiones en el sensor exterior de la planta que ha permitido crear un repositorio de instantes de eventos emisión-inmisión. La mejor solución se ha obtenido con la implementación del algoritmo RandomForestClassifier con estrategia de aprendizaje de ventana

extendida EW. Respecto a una situación inicial en la que la medida de las emisiones contaminantes se obtenía de forma puntual a través de un sensor físico, este estudio ha proporcionado un método capaz de identificar cuándo y porqué ocurren el 86 % emisiones, capaz de proponer la actuación en tiempo real con medios mecánicos (principalmente sistemas de nebulización de agua) para depositar dichas partículas en superficie y evitar así su dispersión, y capaz, en caso de necesidad, de inferir la cuantificación de partículas PM_{10} en las zonas aledañas.

El capítulo ha demostrado que este procedimiento permite, no sólo inferir el valor de las variables críticas en tiempo real en los procesos de control y supervisión anteriores, sino también disponer de una herramienta de detección de anomalías basada en la diferencia sustancial entre los valores de la inferencia y los valores reales.

Capítulo 4

Contribuciones descriptivas en la Gestión de la Eficiencia Energética en plantas de Producción Industrial

Este capítulo propone una metodología para la movilización energética no supervisada del rendimiento nominal de una Planta de Producción Industrial. El análisis de los consumos energéticos y su relación respecto a los niveles de producción, mediante sencillos algoritmos descriptivos y regresivos, permite identificar los patrones de consumo nominales de la planta. La agregación de los consumos globales a nivel de planta permite detectar anomalías e ineficiencias en el régimen de trabajo de la fábrica. Mientras que, la desagregación de los consumos individuales hasta el nivel de máquina permite descubrir la causa raíz de tales ineficiencias. El estudio evalúa el rendimiento del método diseñado sobre un caso de estudio real del sector de la automoción, comparándolo con un extenso benchmark que comprende algoritmos de detección de anomalías no supervisados y semi-supervisados del estado del arte, desde algoritmos clásicos hasta modernos homólogos neuronales generativos.

4.1. Introducción

4.1.1. Estado del arte

Se considera eficiencia energética en Industria a cualquier acción emprendida por una organización que reduzca el uso de energía por unidad de producción sin afectar al nivel de servicio prestado [234]. Las mejoras en eficiencia energética pueden ser aplicables a todas las etapas involucradas en los entornos industriales [235].

En la literatura científica, los argumentos para la mejora de la eficiencia energética en entornos industriales se centran en la reducción de los costes energéticos para las organizaciones, la seguridad del suministro energético, la mejora del confort, la reducción de las emisiones de gases de efecto invernadero y una importante contribución al objetivo del desarrollo sostenible [236].

Pero la realidad de las empresas - según una encuesta realizada a más de 500 responsables de eficiencia energética [237] - es que invierten en medidas de eficiencia no por una iniciativa empresarial ecológica, sino porque lo consideran más como una operativa inteligente: (i) el 60 % de ellos creen que los gastos energéticos aumentaran en los próximos 12 meses (de ahí la motivación y a la vista de la Figura 1.6 no van mal encaminados); (ii) el 90 % entienden que el análisis de los datos permite establecer una línea base que guía una serie de estrategias e iniciativas basadas en eficiencia, y (iii) el 43 % demostró interés en que dichas iniciativas identificadas, sean de bajo coste o sin coste. Siendo una prioridad comprender cómo consumen los recursos (consumo a nivel más granular) a través de submedidores y contadores inteligentes. Además, la dificultad de cuantificar los costes y los beneficios en este tipo de soluciones y, por lo tanto, valorar el incremento de productividad hace que muchos de ellos perseveren en su escepticismo [238].

Existen aplicaciones de software comercial para la gestión de la energía que miden el consumo, analizan los resultados, revelan las deficiencias y elaboran recomendaciones. En mayor o menor medida todas estas soluciones comerciales de gestión de la energía integran técnicas descriptivas, predictivas y/o prescriptivas que proponen técnicas de optimización que operan o bien sobre la planta de producción o bien sobre máquinas individuales del proceso [239]-[241]. Pero, ninguna de estas aplicaciones actúa conjuntamente sobre ambos elementos conjuntamente (a nivel agregado de planta global y a nivel individual de máquina).

4.1.2. Trabajos relacionados

La literatura relacionada con la detección de anomalías de consumos energéticos es extensa en los últimos años, y conlleva desafíos comunes [242]-[245]:

- El comportamiento normal es extremadamente difícil de caracterizar, ya que muchas veces el límite entre lo normal y lo anormal, es a menudo borroso [246];

- El concepto de anomalía varía en diferentes ámbitos de aplicación, y no existe un único algoritmo puede pueda tratarlos todos por igual (teorema No-Free-Lunch [176]). Las anomalías son a veces una cuestión subjetiva [247];
- Cuando los datos contienen mucho ruido, es difícil distinguir los casos ruidosos de las anomalías [248]; y
- El comportamiento normal puede cambiar en cualquier momento, y entonces una noción actual de normalidad puede no ser válida en el futuro.

Existe un debate en curso sobre si la detección de anomalías energéticas en entornos industriales se ha de centrar en máquinas individuales o a nivel agregado de planta de producción [249], [250].

Por una parte, los científicos partidarios del estudio individual a nivel de máquina sostienen que la sensorización particular permite analizar específicamente las anomalías y con ello, proporcionar información precisa sobre la causa raíz que la ha ocasionado [251]. En [252] los autores proponen la monitorización de una turbina en una central hidroeléctrica para la detección de anomalías energéticas. A partir de los datos de la turbina proporcionados por su sistema de control de supervisión y adquisición de datos (SCADA) los autores proponen la composición de patrones de referencia mediante el algoritmo K-Means para cada modo de funcionamiento, basándose en mapas auto-organizativos (SOMs) que permiten la extracción de datos en forma de clústeres. El estudio analiza la transición entre los patrones de comportamiento y la evaluación de las desviaciones cuantificadas numéricamente a través de un conjunto de indicadores definidos. El indicador de similitud mide la similitud de cada nueva observación con respecto a la distribución de probabilidad de su patrón de referencia, y el indicador de distancia, mide las desviaciones con respecto a la desviación estándar de este patrón de referencia. La evaluación de ambos indicadores permite la identificar fallos y detectar degradaciones o cambios de comportamiento de la turbina. Los autores [253] proponen el estudio de ineficiencias energéticas en un tanque de agua. Para ello, presentan la implementación de una red vanilla LSTM que permite descubrir y modelar todos los estados de la máquina y que utiliza los datos supervisados del tanque para asignar los pesos de la red. Este sistema permite generar predicciones de consumo en tiempo real que se comparan con los valores reales, y en caso de sobrepasar un determinado umbral (que se calcula de forma dinámica), informar de dicha anomalía. Además, el estudio se complementa ajustando el problema a una distribución normal multivariante, utilizando la inferencia variacional y un esquema MonteCarlo para manejar la incertidumbre del modelo. Los autores proponen que la función de perdida tenga en cuenta la predicción de ruido. Con todo ello, se consigue emparejar la detección de anomalías con la probabilidad de eventos, y reforzar la hipótesis de la aparición de anomalías o deterioro del sistema. Los autores en [254] exponen un método de detección de anomalías en series temporales de bombas de agua. El estudio está basado en una red neuronal que aprende la dinámica de los estados de funcionamiento de la

bomba y posteriormente aplica un algoritmo de filtrado bayesiano para la detección de anomalías mediante el análisis de la incertidumbre. Se valida la efectividad del algoritmo propuesto en tres casos de uso - sistema de bombeo de agua de una pequeña ciudad, sistema de distribución de agua de área urbana, planta de tratamiento de aguas.

Pero, por otra parte, existe un conjunto de científicos partidarios del estudio global a nivel de planta de producción y que sostienen que detectar ineficiencias a nivel de máquina individual, es insuficiente y que los procesos industriales han de considerar la globalidad del sistema para poder analizar las relaciones entre máquinas [249]. En [46] se expone un sistema de almacenamiento en estanterías altas que dispone de cuatro transportadores horizontales y dos transportadores verticales. De cada transportador se conoce su consumo. Los autores explican que su estudio se basa en la estrecha relación entre las técnicas de reducción de la dimensionalidad de las variables del proceso y la detección de anomalías. Sostienen que cuando se detecta un desajuste en el espacio latente al utilizar PCA y un autoencoder no lineal (AE) es posible su detección. Sus estudios demuestran que, al aumentar el tamaño del espacio latente, el rendimiento del autoencoder aumenta ya que es capaz de trabajar con más comportamientos relevantes. En [255] los autores diseñan un proceso de diagnóstico de ahorro energético en una planta petroquímica basándose en la máquina vectores soporte Curvelet doble. Para mejorar la precisión y optimizar los hiper-parámetros del algoritmo SVM, construyen un algoritmo de optimización de enjambre de luciérnagas híbrido (HGSO) basado en la simulación de recocido simulado (Simulated Annealing Simulation). En [47] se propone un estudio de detección de ineficiencias energéticas en un centro de logística. Los autores describen la solución implementada mediante una red profunda de neuronas que permite encontrar representaciones binarias de patrones recurrentes en las series temporales del sistema industrial. El estudio se centra en detectar y diferenciar clases de anomalías creadas artificialmente respecto al comportamiento modelado de las observaciones de normalidad de la planta de producción.

Los trabajos anteriormente citados a menudo presentan las siguientes limitaciones:

- Las ineficiencias energéticas dependen de la dinámica global del sistema ciber-físico. Afrontar su detección desde el nivel individual de máquina o desde el nivel agregado de planta de producción captura sólo parcialmente el estado del entorno [253]. Para la detección de ineficiencias, en la literatura, se han utilizado un gran número de algoritmos de aprendizaje automático heterogéneos: extracción de características significativas mediante SOMs [252], utilización de algoritmos de agrupación para la identificación de patrones [47], tratamiento de series temporales para la predicción de valores y el estudio de los residuos [253], [254], modelos que reconstruyen los datos comprimidos [46] o algoritmos de máquina de vectores soporte [255]. Sin embargo, todos ellos se centran solo en detectar las anomalías para las que fueron entrenados y que corresponden al uso por exceso

o por defecto de energía [251], [252], [254] y no acostumbran a proporcionar información de cuál fue la causa que originó tal anomalía [47], [249], [255]. Esta falta de información deja al gestor de red en una posición muy vulnerable, y sin conocimiento de cómo reaccionar ante una anomalía de la que desconoce las causas que la ocasionaron [256].

- Las ventajas de aplicar soluciones de IA avanzadas se ven parcialmente compensadas con su falta de interpretabilidad. Siendo la interpretabilidad de los modelos de gran importancia en muchas aplicaciones de explicación de causa raíz [64]. Este concepto de interpretabilidad está alineado con el concepto de Daniel Kahneman [10] (premio Nobel de Economía) de que las razones que llevan a los humanos a tomar decisiones no son necesariamente racionales, sino basadas en aspecto culturales, y en ocasiones inexplicables. En el mundo del dato es un hecho que la interpretabilidad de cualquier solución propuesta crece inversamente proporcional a la complejidad de las técnicas analíticas utilizadas. Estas técnicas disminuyen su interpretación a medida que aumentan en complejidad [257]: desde las técnicas descriptivas y la regresión lineal uniparamétrica hasta los sistemas regresivos más complejos y los sistemas DL. Se trata de que es más difícil comprender los mecanismos de inferencia de información en complejos modelos no lineales (con ininteligibles parámetros) que en sencillos modelos lineales (con parámetros con sentido y catalogados por importancia) [64]. En este aspecto, existen multitud de trabajos en los que se discuten los retos de la interpretación de modelos intrínsecamente interpretables [258], estudios de cómo se comportan dichos modelos en la toma de decisiones críticas [259] y sugerencias que determinan que los modelos interpretables por si mismos no apuntan a conclusiones causales [260]. La autora desconoce que, hasta la fecha, existan estudios o metodologías que consigan la hibridación de la interpretabilidad de los modelos y de los sistemas ciberfísicos.

Pero, más allá de los estudios que enfocan las soluciones a nivel de máquina o a nivel agregado de planta, existen -escasas- publicaciones que propongan metodologías teóricas sobre cómo se ha de diseñar y desarrollar sistemas CPS inteligentes dado un entorno concreto. En [48] los autores sostienen que el CPS debe ajustarse a las necesidades específicas del entorno y resumen la metodología en cinco pasos: (i) determinar objetivos y prioridades en base a las variables de entorno de la planificación, aunque en ocasiones implique modificaciones en los sistemas de planificación y control de producción; (ii) especificar los requisitos del sistema e identificar indicadores de rendimiento; (iii) identificar fuentes de datos, y algoritmos de ML adecuados; (iv) diseñar la arquitectura de los datos y del CPS e (v) implementar un CPS considerando técnicas de integración e innovación continua y adaptabilidad de la arquitectura al dinamismo de los procesos. Los autores indican que el diseño y la arquitectura del CPS deben ser escalables y adaptables a nuevos volúmenes de datos, productos, patrones de consumo, y demás entidades susceptibles de variar. El diseño debe tener

en cuenta en la fase de planificación una estimación del error de los modelos. Por ello siempre debe considerarse una fase previa de entrenamiento de los modelos ML y estimar su precisión. En [261] los autores proponen otra metodología para detectar anomalías en el consumo de energía basada en la extracción de características en determinados momentos relevantes (momentos de toma de decisiones de los usuarios o cuando comparten preferencias) mediante un modelo basado en reglas. Los autores citan la dificultad de obtener datos reales para validar el rendimiento y eficiencia de la metodología principalmente por tratarse de un sistema supervisado.

La realidad es que, pese a que cada vez son más los científicos que coinciden que los CPS serán los pilares de las fábricas inteligentes del futuro [262]-[264], aún se encuentra en fase temprana de conceptualización [265]. El número de aplicaciones comerciales, metodologías y manuscritos científicos que integran ambos enfoques (a nivel individual de máquina y a nivel agregado de planta global) en soluciones aplicadas en el campo de la fabricación de CPS, es escaso, muy conceptual y con limitadas contribuciones prácticas hasta el momento [45]. Con la esperanza de que el desarrollo de nuevos modelos de negocio y nuevos servicios puedan cambiar muchos aspectos cotidianos, sin lugar a dudas, las expectativas en los CPS son altas: diagnósticos a distancia, control en tiempo real, eficiencia, robustez a todos los niveles, etc. [65]. En este escenario, los desafíos típicos son: la creciente complejidad de las plantas de producción, la versatilidad con la que las plantas se modifican y evolucionan en base a nuevas materias primas, nuevos productos o nuevas condiciones de operación y las causalidades y dependencias temporales no triviales [47], [266].

El objetivo de este trabajo es describir una metodología que, a través de las técnicas adecuadas -sencillas técnicas estadísticas y algoritmos básicos de Ciencia de Datos- se capte la dinámica no lineal de los complejos sub-sistemas de interrelación de la fábrica, y proporcione a los técnicos expertos una herramienta de apoyo a la toma de decisiones que les permita abstraerse de complejas técnicas matemáticas. Los retos de esta metodología son: la detección de anomalías de consumos energéticos subóptimos, el estudio de la causa raíz y proporcionar mecanismos para que el experto sea capaz de interpretar mejor su proceso y/o negocio. La metodología basa su estudio en que la caracterización del comportamiento energético y la caracterización de la relación consumo energético y cantidad de producción a nivel agregado de planta, permite deducir los patrones que definen el funcionamiento nominal de la planta de producción industrial. Asimismo el análisis de los consumos energéticos de las máquinas individuales permite conocer la causa raíz de tales ineficiencias energéticas no-supervisadas.

4.2. Enfoque propuesto

La metodología propuesta en este capítulo permite a las personas expertas de dominio conocer el comportamiento energético de su planta de producción, a través de los siguientes pasos:

- Transformar la ingente cantidad de datos energéticos recabados en planta, en información útil para el gestor de eficiencia mediante la identificación de patrones estadísticos basados en consumos energéticos (normalmente complicados de identificar en una inspección visual);
- Identificar la relación entre el consumo energético y el nivel de producción. Mediante el análisis de los patrones estadísticos y de su relación con los niveles de producción se analiza el comportamiento energético global de la planta de producción.
- Crear un cuadro de mando basado en herramientas visuales que permite al experto tener visibilidad completa del comportamiento energético de la planta y extraer conclusiones acerca de cómo y cuándo se consume la energía.
- Detectar ineficiencias energéticas o alarmas. Conocido el comportamiento energético global de planta es posible identificar comportamientos anómalos o comportamientos ineficientes que no tienen correspondencia con el comportamiento global.
- Analizar la causa-raíz de dichas ineficientes. El análisis de los comportamientos energéticos desde la perspectiva de línea, proceso, puesto y/o máquina permite averiguar el origen de dichas ineficiencias.
- Generar informes de frecuencia y contenido variable dependiendo de la audiencia a la que va dirigido.

4.2.1. Caracterización del comportamiento energético

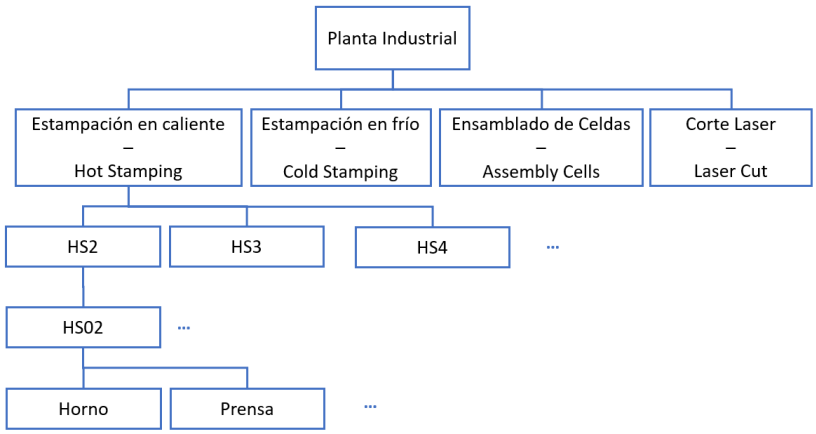
Jerárquica de monitorización energética

Toda empresa de producción industrial se puede representar mediante su distribución jerárquica de monitorización energética. La Figura 4.1 muestra de forma esquemática los niveles en los que se divide la infraestructura de una planta industrial.

- En un nivel superior se representa la **planta industrial**. La empresa a la que pertenece la planta industrial puede tener otras plantas industriales en diferentes países o ubicaciones. Por simplificar, se representa una única planta industrial.
- La planta está compuesta por varias **líneas industriales**, formando un segundo nivel. A modo de ejemplo, la figura muestra cuatro líneas de la planta que podrían ser Estampación en Caliente, Estampación en Frío, Corte Láser y Ensamblado de Celdas.
- Un tercer nivel de despliegue está compuesto por los **procesos** que forman parte de cada línea industrial. Por ejemplo, en la Figura se esquematizan tres procesos HS2, HS3 y HS4 pertenecientes a la línea Estampación en caliente (Hot Stamping) y que comprenden uno o varios **puestos** como la unidad de mecanizado HS02, etc.

- Cada puesto puede tener, a su vez, varias **máquinas**, por ejemplo, la unidad de mecanizado puede comprender Hornos para fundir metales y una Prensa para estampar troqueles y diseñar piezas.

Figura 4.1: Diagrama esquemático de la infraestructura de una planta industrial



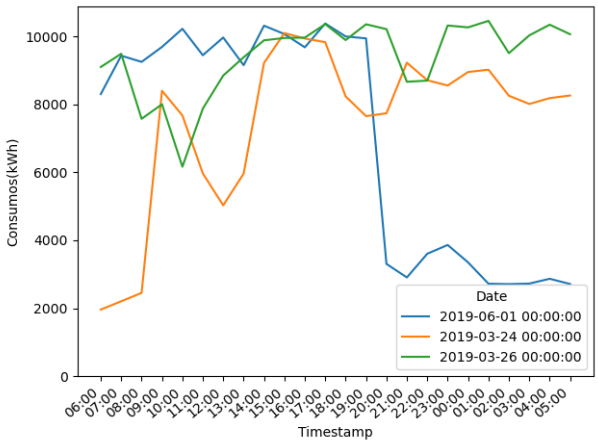
Descripción de los datos

Se utiliza un sistema de gestión de la energía (Sistemas de Monitorización Energética (EMS)) para recoger las mediciones captadas por los sensores a diferentes niveles de la planta industrial: consumo total en planta industrial, consumo en línea de producción, consumo en proceso, consumo en puesto, consumo en máquina, etc. El método también considera las mediciones indirectas, es decir, las basadas en inferencias indirectas. Por ejemplo, según el principio de conservación de la energía, la energía asociada a las cargas no monitorizadas puede deducirse a partir de la diferencia entre la energía total de un transformador (al que están conectadas tanto las cargas monitorizadas como las no monitorizadas) y la energía de las cargas monitorizadas asociadas a dicho transformador. Es decir, restando la energía de las cargas monitorizadas, de la energía total del transformador se obtiene la energía de las cargas no monitorizadas. Con ello, el método engloba tanto las cargas directas como las indirectas (deducidas de las anteriores). A modo de ejemplo, si existe un contador que monitoriza la energía total consumida por una planta que tiene 40 máquinas, pero únicamente se monitoriza el consumo energético de 30 de ellas, es posible inferir el consumo agregado de las otras 10 máquinas no monitorizadas, restando del consumo medido por el contador de la planta la suma agregada de los consumos monitorizados de esas 30 máquinas.

Con los datos de consumo de energía captados por los sensores en diferentes instantes de tiempo (por ejemplo, en determinado intervalo de tiempo Δt, en este caso cada 1 hora) para un determinado periodo de tiempo T (por ejemplo, en este caso un día), se obtiene una pluralidad

de J curvas de consumo de energía x_j. En otras palabras, una curva de consumo de energía x_j representa un perfil de consumo de energía medido con cierta discretización temporal Δt de una máquina, puesto, proceso, línea o planta.

Figura 4.2: Curvas diarias de consumo energético de una planta industrial



De este modo, la Figura 4.2 muestra tres curvas diarias de consumo energético a nivel de global de planta, en intervalos de discretización temporal de 1 hora. Es decir, cada curva tiene 24 valores correspondientes al periodo de 1 día. Los valores de la curva de consumo están comprendidos entre las 6h de la mañana y las 6h de la mañana del día siguiente; esto es así para una correcta correspondencia entre el consumo energético y el nivel de producción diario.

Dependiendo del elemento de la planta industrial al que corresponda la curva de consumo energético, la curva puede representar el consumo energético de diferentes componentes (mediciones de consumo energético directas, por ejemplo, de una máquina de bajo nivel, como una prensa) o mediciones de consumo energético indirectas, que corresponden a las mediciones agregadas de diferentes componentes comprendidos en un nivel superior de la planta industrial (por ejemplo, a nivel de Línea de Estampación en Caliente como la suma de máquinas que forman esta Línea). En otras palabras, las curvas de consumo de energía pueden obtenerse a diferentes niveles de submedición a través de la agregación de niveles inferiores. En este contexto, la expresión "nivel de submedición" se entiende como cada uno de los diferentes niveles jerárquicos de monitorización de la energía en la planta industrial, tal y como se muestra en la Figura 4.1.

Para el estudio se ha trabajado con datos históricos de medidas horarias de 2 años (2018 y 2019). El consumo de energía se representa en kWh (kilovatios hora).

Descripción de la metodología, de los modelos empleados y resultados

Una vez que, para un determinado horizonte temporal H, se ha obtenido una pluralidad de J curvas de consumo energético x_j para un determinado elemento de la planta industrial -o lo que es lo mismo, para un determinado nivel de submedición, también llamado nivel de agregación- se crea una pluralidad de K patrones de consumo energético C_k en ese nivel.

Los patrones de consumo energético C_k se obtienen a partir de la asociación por similitud de un conjunto de curvas de consumo energético x_j. Por lo tanto, los patrones de consumo de energía se extraen para un mismo nivel de submedición o agregación, en el que cabe esperar curvas de consumo de energía similares. El subíndice k es un número entero positivo que denota un patrón de consumo energético específico. En otras palabras, $k = 0, 1, 2, ..., K-1$, siendo K un número natural. Generalmente, $K \leq J$, y en la práctica $K < J$. Los patrones de consumo de energía C_k se obtienen a partir de un conjunto de curvas de consumo de energía x_j. Los patrones de consumo de energía C_k se obtienen comparando las curvas de consumo de energía de un conjunto de ellas siguiendo una determinada métrica matemática, en este caso distancia euclídea e identificando las curvas de consumo de energía similares x_j según dicha métrica de similitud. En otras palabras, en este contexto, similitud significa proximidad en términos de distancia euclídea entre curvas. Así, un patrón de consumo energético representa un comportamiento o modelo energético que representa un conjunto de curvas de consumo energético (curvas de consumo energético similares). Normalmente, cada patrón de consumo de energía agrupa un grupo de curvas de consumo de energía que cumplen una restricción de similitud impuesta por o según una determinada distancia euclídea. En resumen, se organizan los datos capturados y se buscan relaciones -en términos de similitud entre las curvas de consumo energético existentes- creando/obteniendo así diferentes patrones de comportamiento. Esta etapa puede considerarse como una fase de entrenamiento.

En la fase de entrenamiento se analizan los datos históricos, es decir, las curvas de consumo energético definidas por los valores de consumo energético en determinados intervalos de discretización temporal Δt, dentro de un periodo de tiempo T y fijando un horizonte temporal de entrenamiento H. Durante esta fase de entrenamiento, cada patrón tiene un valor que se calcula a partir de las diferentes curvas de consumo de energía pertenecientes al conjunto representado por el patrón. Cada comportamiento tipo o centroide de un patrón se obtiene calculando, por ejemplo, el valor medio de todos los valores correspondientes a todas las curvas de consumo energético del conjunto, en el instante de tiempo correspondiente. El objetivo de este cómputo es obtener un patrón de consumo energético, es decir, la curva más parecida en cantidad (valor de consumo energético) y forma (también denominada tendencia) al conjunto de curvas de consumo energético que han participado en el cómputo de un horizonte temporal definido.

La inferencia de los patrones de consumo anteriores se realiza aplicando una técnica de agrupamiento. Se recurre al conocido algoritmo de clustering K-Means. Teniendo en cuenta la notación introducida en esta sección, el algoritmo de clustering utilizado para calcular estos patrones de consumo energético implementa un proceso iterativo para minimizar la distancia entre los elementos que forman un cluster. La siguiente formulación matemática representa el cálculo de la agrupación que se utiliza para obtener K patrones de consumo de energía C_k a partir de J curvas de consumo de energía:

$$argmin \sum_{i=1}^{I} \left\| x_{ij} - C_{ik}^{t} \right\| \forall k = 1, ..., K \tag{4.1}$$

donde x_{ij} es el i-ésimo componente de la j-ésima curva de consumo de energía x_j, C_{ik}^{t} es el i-ésimo componente del k-ésimo patrón de consumo de energía, K es el número total de patrones de consumo de energía e I es el número de mediciones dentro de todo el período de tiempo (una medición tomada cada hora, por ejemplo, 24 mediciones tomadas a lo largo de 1 día). Cada centroide se calcula como el valor medio de las curvas de consumo de energía pertenecientes a dicho patrón, en cada componente i. La formulación anterior representa el proceso iterativo para identificar las curvas pertenecientes a un patrón, minimizando la distancia euclídea entre los elementos que forman un cluster y su correspondiente patrón.

K-Means viene acompañado de la necesidad de establecer el número K de patrones de consumo que permiten definir el comportamiento energético de la planta. En la literatura, se han definido muchas heurísticas y variantes algorítmicas para este fin, entre ellas el método clásico del codo [267] (pareto entre las distancias inter-cluster y el número de clusters). Cuando la dispersión en la varianza acumulada intra-cluster es alta, el método del codo no consigue definir un número óptimo de patrones K [268]. Esto es debido a que tiende a desvirtuar la conformación del número de patrones atendiendo al punto de máxima curvatura [268]. En ocasiones como esta, si el punto de máxima curvatura implica un cambio disruptivo, se determina el número K aunque no sea el valor óptimo.

Como alternativa al método del codo, se define un procedimiento iterativo donde la varianza entre patrones es independiente de la varianza inherente a cada uno de ellos. En cada iteración el procedimiento identifica el patrón más disperso - entendiendo como dispersión la suma de las diferencias al cuadrado entre cada curva de consumo energético a su correspondiente centroide - y lo fragmenta hasta conseguir patrones bien definidos, con mayor sensibilidad a los comportamientos estadísticos y con la misma significación estadística. Inicialmente, el método de clustering identifica el número mínimo de patrones que definen el comportamiento energético de la planta industrial. A continuación, se fragmenta iterativamente los patrones más dispersos en función de su densidad, hasta conseguir patrones bien definidos y representativos de la muestra. El objetivo es dar la misma significación estadística a todos los patrones.

El siguiente paso consiste en que para que el sistema sea confiable, cada comportamiento debe estar formado por las curvas estadísticamente más

representativas. Por ello en cada comportamiento se eliminan aquellas curvas menos representativas. Por ello, se utiliza iterativamente el estudio del rango intercuartil [269] para la detección y eliminación de perfiles extremos en cada cluster. El estudio del rango intercuartil considera que cada cluster ha de responder a una distribución gaussiana, siendo perfiles atípicos los extremos de la distribución. Es decir, se supone que se produce un perfil atípico cuando una curva de consumo está por encima de un umbral superior a $(Q3 + 1{,}5 * |Q3 - Q1|)$ o por debajo de un umbral inferior $(Q1 - 1{,}5 * |Q3 - Q1|)$, donde $Q1$ y $Q3$ denotan los cuartiles primero y tercero de la distribución de la distancia euclídea de cada curva a su centroide. Las curvas que caen fuera de esos umbrales se consideran anomalías o ineficiencias energéticas.

El proceso de asignación de curvas a patrones finaliza cuando todas las curvas de consumo energético dentro de un mismo patrón son más similares entre sí que con cualquier otra curva perteneciente a un patrón diferente. La Figura 4.3 presenta los cinco patrones energéticos identificados tras la aplicación de las técnicas mencionadas anteriormente a partir de una colección de curvas de consumo energético:

- Un 1er patrón de energía (a unos $10000kWh$ línea parcialmente continua y parcialmente a unos $4000kWh$) que agrupa 14 curvas de consumo de energía (líneas finas). La colección de curvas de consumo energético del cluster#0 consume energía al comienzo del día y paulatinamente, de 19:00h a 21:00h, el consumo decrece. La planta funciona en modo-parada.

- Un 2o patrón de energía (línea continua parcialmente ascendente desde $3600kWh$ hasta $10000kWh$) que agrupa 10 curvas de consumo de energía (líneas finas). La colección de curvas de consumo energético del cluster #1 consume energía de forma ascendente desde primera hora de la mañana (aproximadamente 06:00h) hasta primera hora de la tarde (aproximadamente 16:00h), a partir de entonces el consumo de consolida en $10000kWh$. La planta funciona en modo-arranque.

- Un 3er patrón de energía (línea continua a unos $9000kWh$) que agrupa 16 curvas de consumo de energía (líneas finas). La planta funciona a rendimiento medio.

- Un 4o patrón de energía (línea continua parcialmente a unos $8000kWh$ y parcialmente a unos $10000kWh$) que agrupa 16 curvas de consumo de energía (líneas finas). La planta funciona a bajo rendimiento.

- Un 5o patrón de energía (línea continua a unos $10000kWh$) que agrupa 16 curvas de consumo de energía (líneas finas). La planta funciona a alto rendimiento.

Cada patrón, está representado por un comportamiento tipo o centroide del patrón calculado como el valor medio -en cada instante de tiempo- de los valores correspondientes de las curvas representadas por dicho patrón. Se representa el centroide de cada patrón en color negro. Es decir,

a través de las curvas de consumo energético se identifican los patrones que representan la planta de producción en periodos de tiempo diarios. En este caso, el comportamiento energético global de planta está representado mediante 72 curvas de consumo repartidas en $C_k = 5$ patrones de consumo de energía obtenidos en la fase de entrenamiento.

4.2.2. Caracterización de la relación consumo energético vs. producción

Descripción de los datos

Hoy en día, en las plantas industriales los datos de producción se miden y almacenan mediante Sistemas de Monitorización de la Producción (MES), normalmente independientes de los EMS. Estos niveles de producción se miden habitualmente en kilogramos (k) de producto producido para un determinado período de tiempo, en este caso, por día. En función de la relación de las curvas de consumo energético con los niveles de producción, las cargas se clasifican como:

- Medición de **cargas productivas**, haciendo referencia a cargas asociadas a la producción,
- Medición de **cargas auxiliares**, haciendo referencia a cargas no asociadas directamente a la producción; y
- Medición de **cargas no monitorizadas** o cargas no medidas directamente. Dependiendo del nivel industrial y del tipo de elemento o carga, el consumo puede obtenerse directamente de un sensor de medida (puede ser el caso, por ejemplo, de una máquina), o indirectamente, bien a partir de diferentes sensores de medida (puede ser el caso, por ejemplo, de una Línea de producción o planta industrial) o bien a partir de inferencias indirectas como ya se ha explicado.

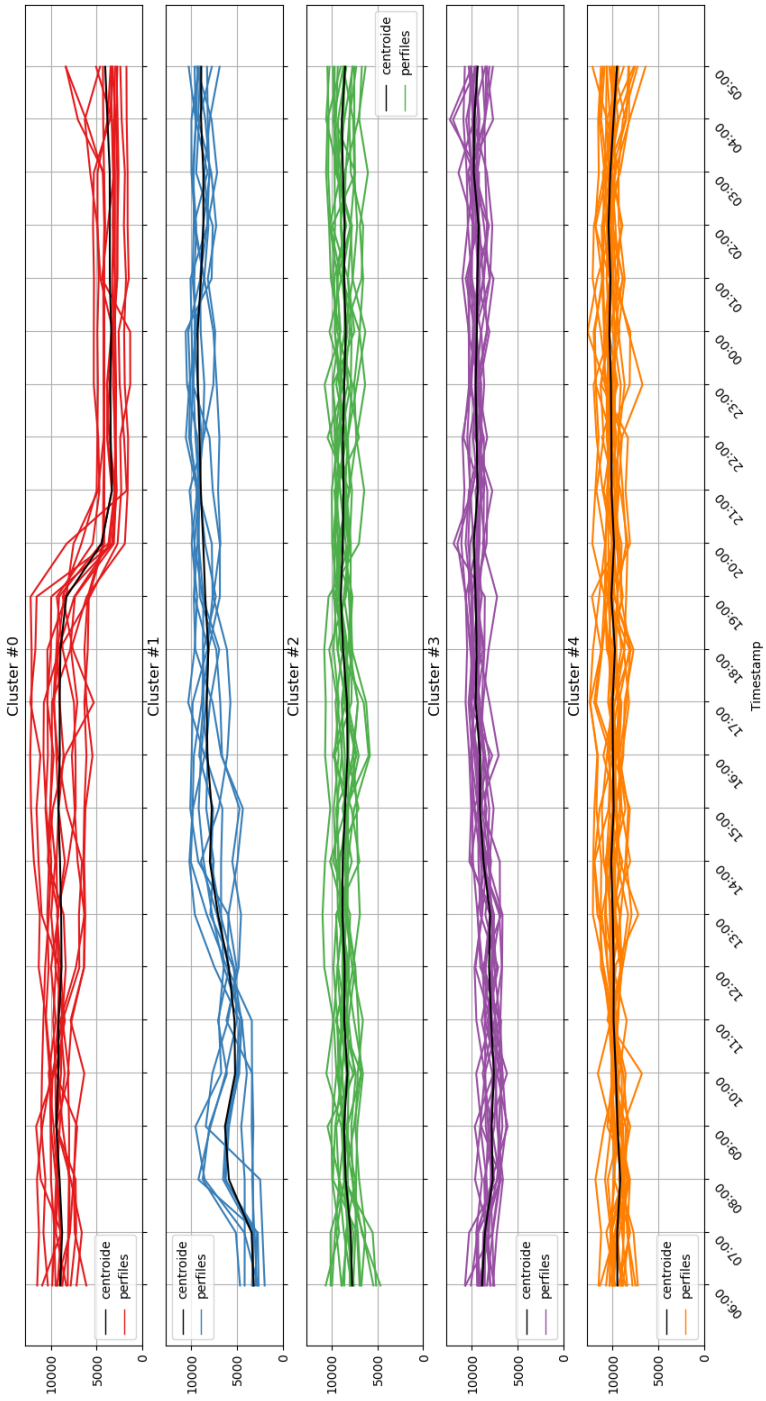
El consumo energético de estas mediciones es almacenado en un servidor central (aunque también podría ser localmente o en la nube), para su posterior análisis, evaluación, seguimiento y/o procesamiento.

Esta diferenciación entre cargas productivas y cargas auxiliares permite centrar el foco de la productividad únicamente en las cargas asociadas a producción. Y descartar para los estudios de producción, las cargas auxiliares y las cargas no monitorizadas. De no hacerlo así se estaría asignando producción a cargas que nada tienen que ver y añadiendo incertidumbre en la detección de ineficiencias. Ello provocaría un deterioro del espacio ocupado por los patrones característicos y su variabilidad natural, pudiendo incluso solaparse patrones e invalidando las técnicas de clustering al no poder distinguirse los distintos comportamientos. Por ejemplo, sería un error si en el estudio de eficiencia energética asociada a producción se incluyese información de consumo de luminarias.

Descripción de la metodología, de los modelos empleados y resultados

En la fase de entrenamiento y una vez que se dispone de un conjunto de K patrones de consumo energético C_k, y utilizando los datos de producción correspondientes a dicho periodo de tiempo T, se compara el nivel de producción con las curvas de consumo energético de los patrones para detectar posibles ineficiencias energéticas.

Figura 4.3: Curvas de consumo energético y patrones de los 5 comportamientos que definen el funcionamiento de la planta en la fase de entrenamiento



Es una práctica habitual que la captación de datos de producción se haga con menor frecuencia que los datos de consumos energéticos. En este caso, mientras que los sensores están configurados para medir la energía consumida en un determinado componente (máquina, línea, etc.) de una planta industrial cada 1 hora (intervalo de tiempo discretizado Δt), los datos de producción asociados a dicho componente están disponibles sólo cada día. Por esta razón, para comparar el consumo de energía con el nivel de producción, respecto a un mismo periodo de tiempo T, se unifica la frecuencia de muestreo a un día. En otras palabras, con los valores agregados de consumo de energía obtenidos para un determinado período de tiempo, y los valores de producción obtenidos para el mismo período de tiempo, se representa la energía frente a la producción y se infiere una relación de linealidad entre el consumo de energía y la producción diaria. Dependiendo de la configuración específica que se analice, esta suposición de linealidad podría no ser suficiente para caracterizar con precisión dicha correlación. Sin embargo, las ventajas derivadas de la simplicidad del modelo compensan su potencial falta de poder de modelización. Por ello, es posible explicar dicha relación a través de una regresión lineal simple, donde se supone que estas dos variables se ajustan a una regresión lineal dada por la siguiente recta de regresión:

$$\mathit{consumo} = \mathit{pendiente} \cdot \mathit{produccion} + \mathit{offset} \tag{4.2}$$

donde *consumo* es la variable dependiente, que denota el total de energía diaria consumida en kWh; *produccion* es la variable explicativa, que representa la producción de la planta industrial (en unidades según el tipo de producto fabricado por la planta, por ejemplo kilogramos); *pendiente* representa la parte de la producción que contribuye a la energía diaria consumida; y *offset* es el término de intercepción o error (el valor de Y cuando $X = 0$), que denota la parte de la energía diaria que no depende de las unidades de producción. Los parámetros de este modelo lineal (*pendiente* y *offset*) se aprenden a partir del histórico de datos de consumo de energía y producción dados.

La Figura 4.4 representa bidimensionalmente la relación entre energía y producción. En esta representación gráfica, hay tantos puntos como días históricos se utilizan en el análisis. Cada patrón se representa por un color diferente. La relación entre el conjunto de pares energía - producción $(E_i \mid P_i)_{i=1}^{N}$, donde E_i y P_i son respectivamente, valores diarios de energía y producción, y N es el conjunto de días (horizonte temporal) utilizados en la fase de entrenamiento, puede expresarse como sigue:

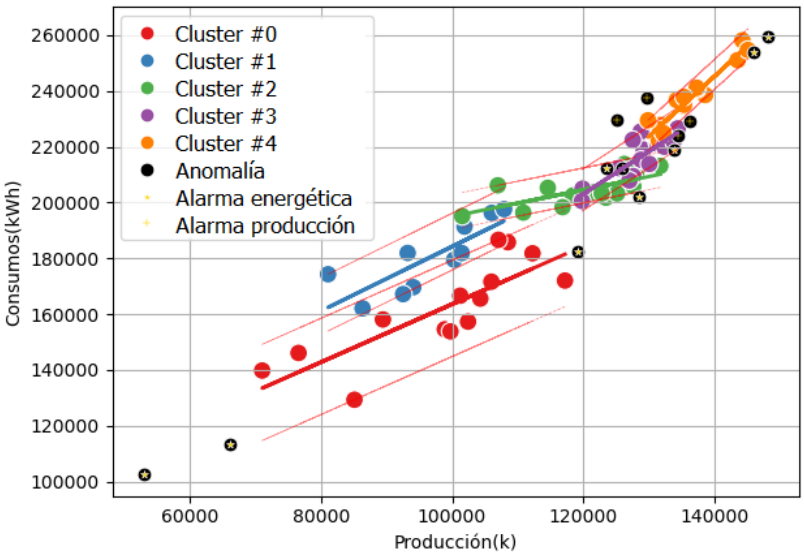
$$E_i = f(P_i) \tag{4.3}$$

En otras palabras, el objetivo es encontrar la función f que minimiza el error:

$$e_i = E_i - f(P_i) \tag{4.4}$$

Para ello, se realiza un ajuste por mínimos cuadrados con objeto de averiguar los puntos que maximizan el coeficiente de determinación en cada patrón entre ambas magnitudes. Aquellos puntos que no maximizan

Figura 4.4: Correlación del consumo energético agregado diario vs el nivel de producción en la planta industrial



el coeficiente, se les considera anomalías energéticas y se representan en color negro. Al coeficiente de determinación también se le denomina métrica R^2 o Chi2. Este valor estadístico cuyo propósito es validar hipótesis, permite cuantificar la calidad de la relación entre ambas variables. Oscila entre valores de 0 y 1. Cuanto más próximo a 1 se sitúe el coeficiente de determinación, más fiable será la relación entre las variables y viceversa. La definición matemática de R^2, para la regresión lineal simple en el caso que aplica, es el cuadrado del coeficiente de la correlación de Pearson.

Asimismo, cada patrón está representado por su recta de regresión (del mismo color que los puntos del patrón) que indica la línea que mejor ajusta el conjunto de puntos representado. En la literatura otros autores de dominio se refieren a ella como línea de eficiencia energética [270]. Esta recta sirve para representar la mejor relación energía vs producción.

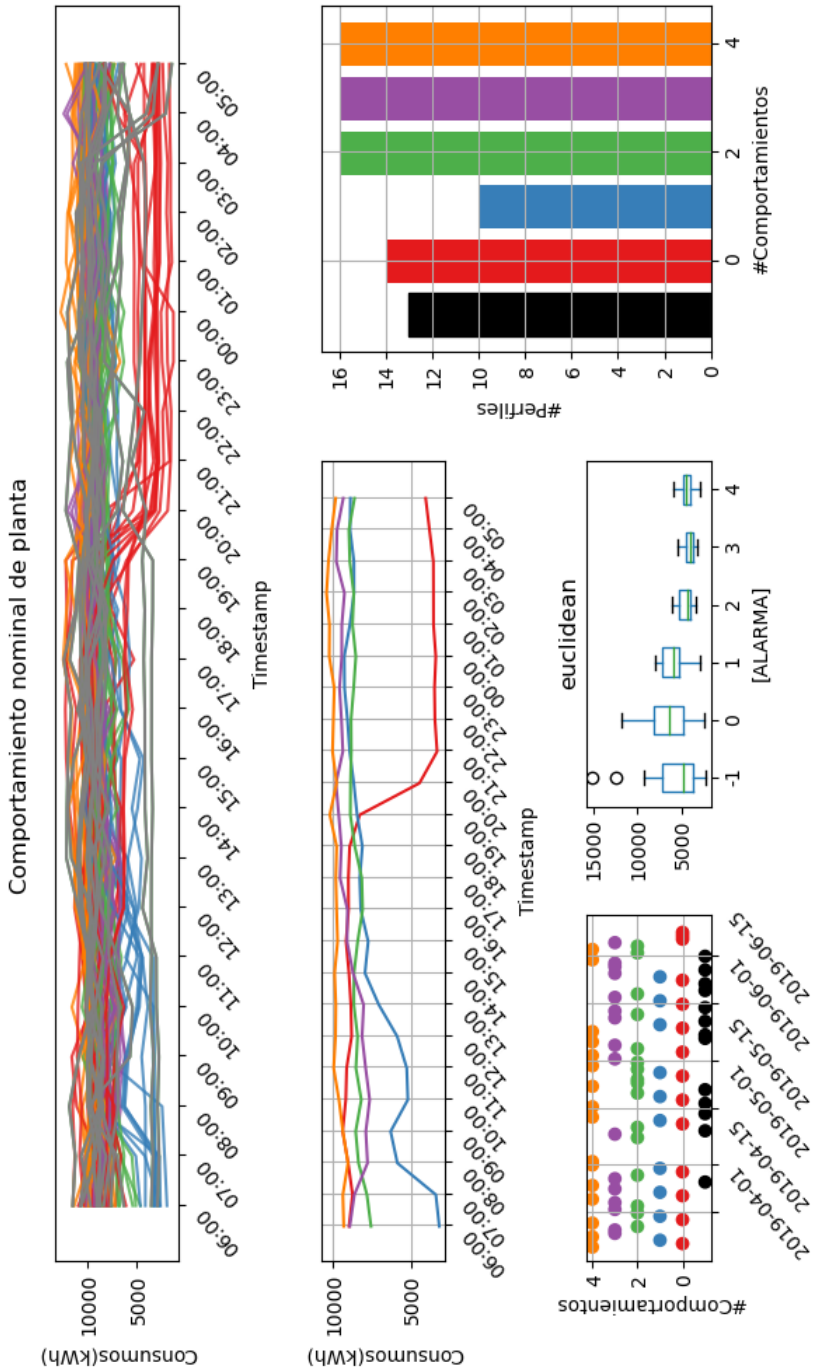
Posteriormente, en la etapa de validación, cualquier nuevo día puede ser representado y analizado de forma similar, evaluando así si un aumento del consumo de energía requerido para producir cierta cantidad de producción (k) es una ineficiencia energética o no. Al contrastar, por ejemplo, un nuevo día, se determina si ese día responde a la relación histórica (no empeora el Chi2) o si no lo hace (empeora el Chi2) en cuyo caso se puede disparar una alarma de producción. En otras palabras, si un aumento del consumo de energía se debe a un aumento de la producción, no debería activarse ninguna alarma. Por el contrario, si un aumento del consumo de energía no está asociado a un aumento de la producción, debería activarse una alarma. Por lo tanto, la producción es muy importante para evitar falsas alarmas. El objetivo es detectar consumos intensivos de energía, superiores a los esperados, no asociados a incrementos de producción.

4.2.3. Panel de control

El conocimiento adquirido en los dos apartados anteriores se resume en un Panel de Control. Panel que sintetiza las claves para que cualquier experto de dominio o gestor de eficiencia energética comprenda cuál es el funcionamiento energético de la planta de producción, sin necesidad de complejos conocimientos matemáticos.

La Figura 4.5 muestra, desde un punto de vista puramente energético, que una planta de producción no es más que un conjunto de curvas de carga de consumo de energía. El gestor de eficiencia energética gestiona el consumo global de la planta de producción como la agregación del conjunto de curvas de consumos individuales. Esta información, en sí misma, no aporta conocimiento sobre el modo de operar de la planta. Pero tras la aplicación de la metodología propuesta es capaz de comprender que tras las curvas individuales hay patrones de comportamiento que se repiten. Estos patrones están diferenciados en la gráfica por colores: patrón de modo-parada de planta (color rojo), patrón de modo-arranque de planta (color azul) y patrones de planta en funcionamiento (color naranja, morado y verde). La gráfica muestra, a través de un diagrama de puntos o scatter-plot, la representación espacio-temporal de cada uno de los patrones, permitiendo observar la periodicidad sistemática con la que aparecen algunos de ellos (el patrón rojo tiende a producirse los sábados y el patrón azul en domingo). El diagrama de barras de la figura muestra la distribución equitativa de los cinco comportamientos (colores rojo, azul, verde, morado y naranja) y de las curvas detectadas como ineficiencias energéticas (color negro).

Figura 4.5: Panel de control que define el funcionamiento de la planta en la fase de entrenamiento



4.2.4. Contrastación de un nuevos comportamientos. Generación de alarmas

Tras la fase de entrenamiento y conocido el número K de patrones que definen el comportamiento de planta y la relación consumo-producción esperada, se procede con nuevas curvas diarias para su evaluación.

Como **primer paso**, se determina a qué patrón pertenece la nueva curva (si es que pertenece a alguno). Esto se hace comparando la nueva curva con los K patrones. Esta comparación se realiza según la métrica de similitud definida o distancia euclídea. El grado de similitud/disimilitud de una nueva curva de consumo de energía con respecto a los patrones existentes puede evaluarse como:

$$x_j \epsilon C_k^t fork/argmin \left\| x_j - C_k^t \right\| \forall j = 1, ..., J, k = 1, ..., K \tag{4.5}$$

donde x_j es la j-ésima curva de consumo de energía y C_k^t es el k-ésimo patrón. La nueva curva de consumo de energía bajo evaluación se asocia al patrón del conjunto de K patrones que tiene la distancia euclídea mínima, siempre que dicha distancia mínima no sea mayor que la distancia de la curva más lejana dentro de dicho patrón. En otras palabras, una nueva curva de consumo energético x_j puede evaluarse con respecto a un patrón de consumo energético, esta evaluación puede realizarse utilizando formulaciones matemáticas de similitud/disimilitud, como las siguientes:

$$A_i^t = x_j I \left\| x_j - C_i^t \right\| \leq \left\| x_j - C_k^t \right\| \forall 1 \leq k \leq K \tag{4.6}$$

donde A_i^t es el i-ésimo grupo, x_j es la j-ésima curva de consumo energético, C_i^t es el patrón del i-ésimo grupo y K es el número total de patrones. Es decir, una nueva curva de consumo energético pertenece a un determinado patrón si la distancia euclídea al centroide de dicho patrón es menor que la distancia a cualquier otro centroide de cualquier otro patrón y siempre que dicha distancia mínima no sea mayor que la distancia de la curva más lejana dentro de dicho patrón.

En un primer escenario, **la nueva curva de consumo de energía se explica con un patrón ya existente**, es decir, la nueva curva de consumo de energía se ajusta matemáticamente a un patrón previamente definido en términos de la métrica de similitud. En este caso, la nueva curva de consumo energético se compara digitalmente con todas las curvas de consumo energético que pertenecen al comportamiento representado por el patrón al que pertenece. La nueva curva de consumo de energía se asocia al grupo de curvas representado por dicho patrón y el patrón de consumo de energía se actualiza incluyendo la nueva curva de consumo de energía. Cuando la nueva curva de consumo energético se compara digitalmente con todas las curvas de consumo energético que pertenecen al patrón, la nueva curva se compara, entre otras, con la curva de máximo consumo energético y con la curva de mínimo consumo energético dentro del grupo. La Figura 4.6 representa, por ejemplo, la curva de consumo de energía [2019-12-14] (línea descontinua negra) al cluster #0.

En un segundo escenario, **la nueva curva de consumo energético no se explica por un patrón** generado durante el horizonte temporal seleccionado. En otras palabras, la nueva curva de consumo de energía no se ajusta a un patrón previamente definido en términos de una métrica de similitud. En este caso, representa un consumo anómalo, en cuyo caso genera o dispara una alarma asociada al consumo de energía (**alarma energética**). Por ejemplo, la Figura 4.7 representa cómo la curva de consumo de energía [2019-07-14] identifica el cluster #1 como el más semejante el cluster, pero rebasa los límites de las curvas que pertenecen a dicho cluster. Alternativamente si esta circunstancia aparece de forma reiterada otros días, puede representar la aparición de un nuevo tipo de comportamiento de consumo, es decir, un nuevo patrón. En otras palabras, para cada nueva curva de consumo de energía, se comprueba si provoca una alarma energética o no. Esto se hace comparando la nueva curva de consumo de energía con los patrones K ya definidos, por ejemplo, durante la fase de entrenamiento.

Como **segundo paso**, la nueva curva de consumo energético debe evaluarse en términos de energía frente a los niveles de producción. Una vez que la nueva curva de consumo de energía ha sido sometida a la evaluación para determinar si pertenece o no a un patrón conocido, con su consiguiente generación o no de una alarma asociada al consumo de energía, la nueva curva de consumo de energía se compara con una relación esperada de energía frente a producción (obtenida durante la fase de entrenamiento, como se representa por ejemplo en la Figura 4.4). Se capturan los datos de la producción alcanzada durante el mismo periodo de tiempo de la nueva curva de consumo energético. El consumo de energía y la producción asociada durante el periodo de tiempo de la nueva curva pueden compararse con la relación energía vs. producción inferida de las curvas históricas. Si la relación energía vs. producción de la nueva curva de consumo energético no puede ser explicada por la variabilidad estadística del modelo inferido durante la fase de entrenamiento, se puede generar una **alarma de producción**. Es decir, si como resultado de la comparación digital se determina que hay ineficiencia energética, se dispara una alarma asociada a la producción. Si una nueva curva de consumo de energía, representada como un nuevo punto en el gráfico de la Figura 4.4, hace que la ya mencionada Chi2 empeore, entonces se puede deducir que dicha curva de consumo de energía no se corresponde con la relación producción-energía derivada. Por lo tanto, puede activarse una alarma de producción, véase el triángulo de bordes negros de la Figura 4.8 que debiendo ser asignado al cluster #1 en color azul se encuentra fuera de los límites que definen el patrón. En este caso se trata un perfil con consumo inferior al esperado en base a su nivel de producción. Véase que otros perfiles del mismo cluster #1, con semejantes niveles de producción (81000k), tienen consumos superiores (de 175000 kWh frente a 155000 kWh). Es decir, la curva estudiada no cumple la relación energía vs producción por subconsumo energético.

Finalmente, como **tercer paso**, la metodología contempla que los patrones se actualizan cada vez que se obtiene una nueva curva de consumo de energía. Volviendo a los dos escenarios ya comentados: en un primer

Figura 4.6: Nueva curva de consumo energético [2019-12-14] que SÍ se ajusta a patrón existente (cluster#0)

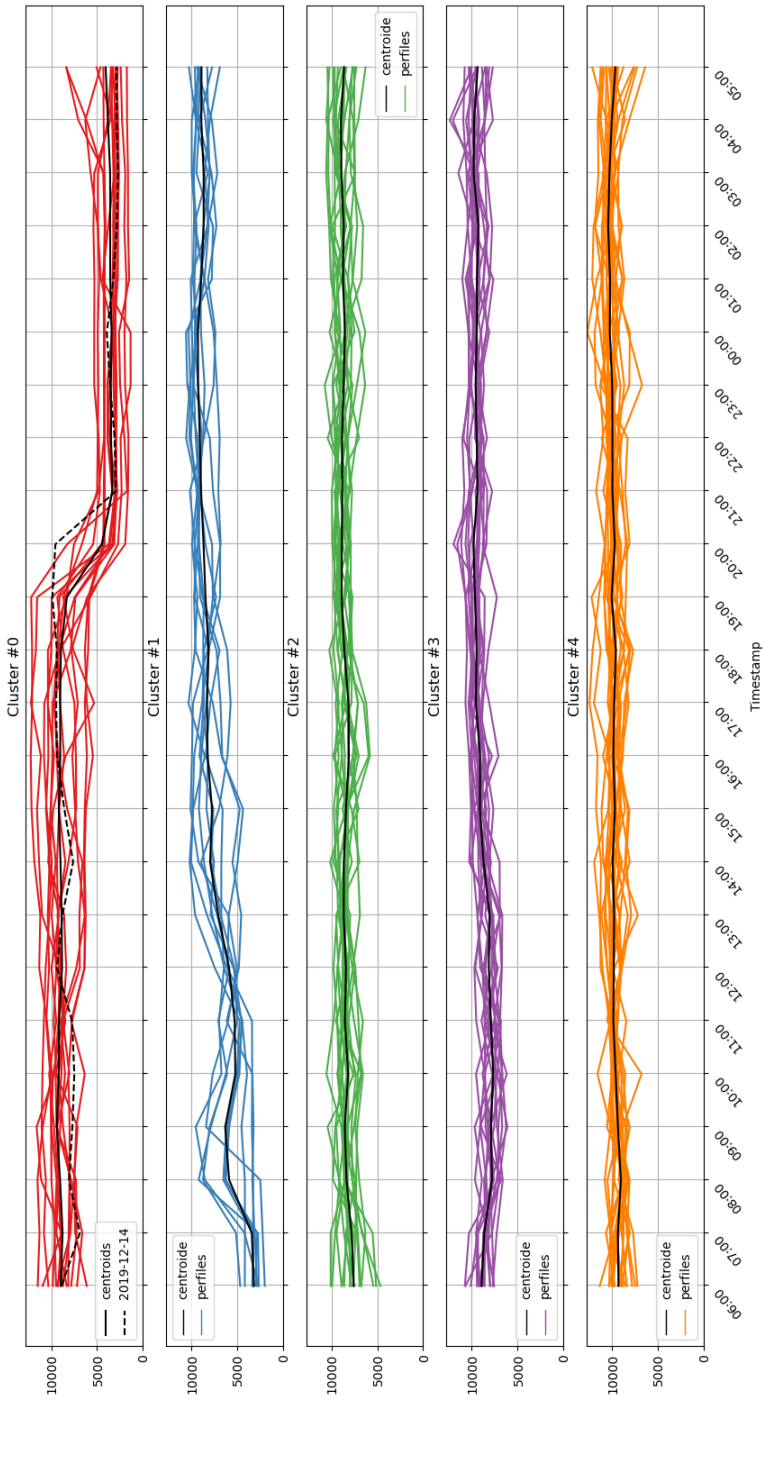


Figura 4.7: Nueva curva de consumo energético [2019-07-14] que NO se ajusta a patrón existente

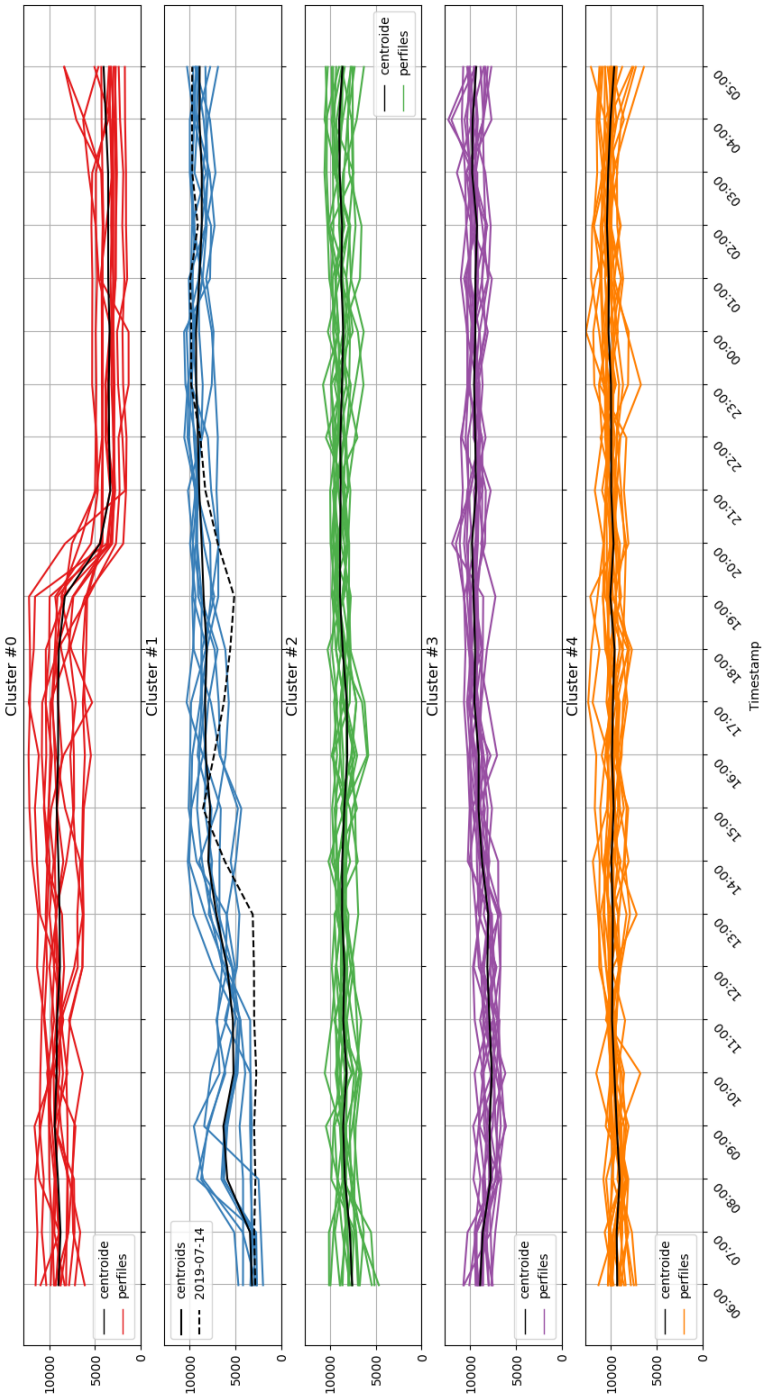
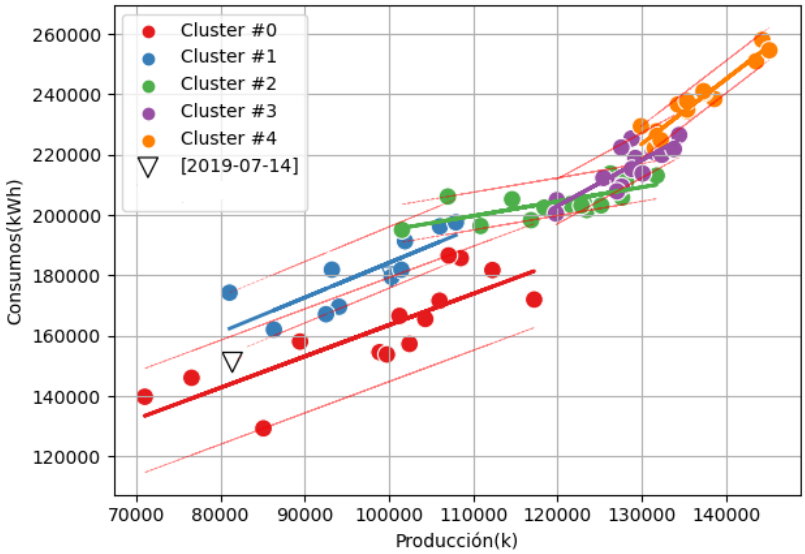


Figura 4.8: Correlación consumo vs producción para nueva curva [2019-07-14]. NO se ajusta a la relación producción-energía del patrón #1



escenario, en el que la nueva curva de consumo energético encaja matemáticamente en un patrón previamente definido, entonces la nueva curva de consumo energético se asocia al grupo de curvas que forman ese patrón y se recalcula el comportamiento característico (patrón). En un segundo escenario, en el que la nueva curva de consumo energético no se explica por un modelo energético discreto (patrón) ya generado: si se determina que la nueva curva representa un consumo anómalo, la nueva curva no debe contribuir a actualizar un patrón. Si, por el contrario, se determina que la nueva curva no representa un consumo anómalo, se puede determinar que representa la aparición de un nuevo tipo de comportamiento de consumo, es decir, un nuevo patrón. En todos los casos se le notifica al gestor de eficiencia energética de la aparición de una ineficiencia o de un nuevo comportamiento.

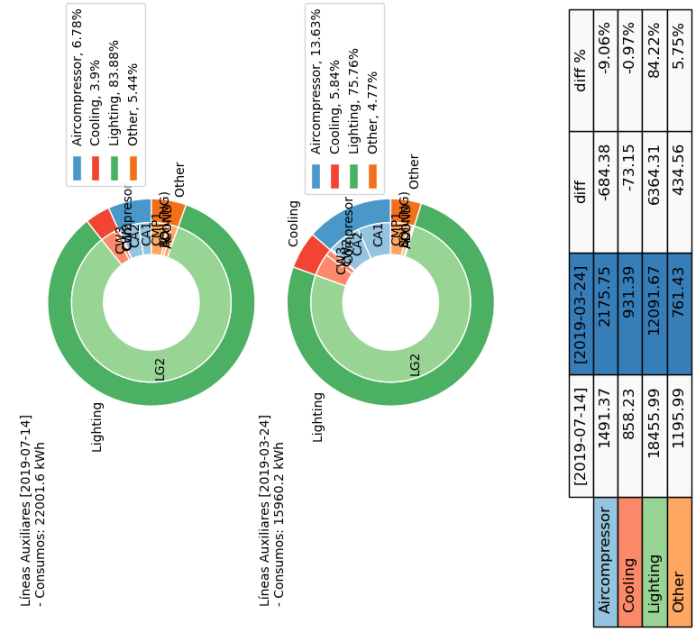
Además, cada cierto tiempo, bien definido por un parámetro configurable o bajo petición del gestor energético de planta, la generación de patrones de consumo energético debe actualizarse utilizando nuevas curvas de consumo energético que hayan sido medidas con posterioridad a la obtención de los patrones de consumo energético actuales en el horizonte temporal utilizado en la fase de entrenamiento. En otras palabras, los datos históricos utilizados, por ejemplo, para el entrenamiento, pueden ser actualizados. Así, se obtienen nuevos patrones de consumo de energía teniendo en cuenta las nuevas curvas.

4.2.5. Detección de ineficiencias energéticas. Identificación de causa raíz

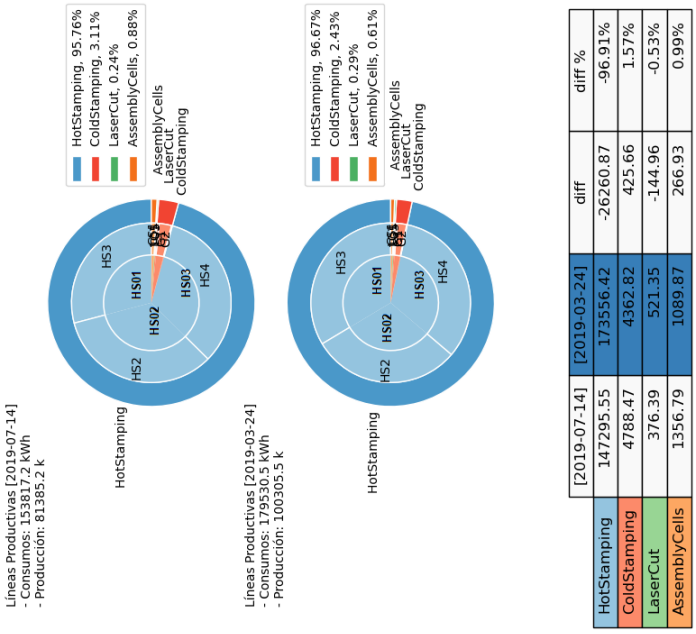
En el proceso del análisis de la causa raíz, se compara la curva de consumo de energía reconocida como anómala contra un curva o perfil de referencia. El perfil de referencia o curva del día más eficiente se define como aquella curva que minimiza la relación entre consumo energético y nivel de producción. Se trata de identificar las cargas energéticas concretas (de máquina, puesto, proceso o línea) de la nueva curva, qué han provocado que no pertenezca a ningún comportamiento conocido. Esto se deduce a través de la comparación de las mediciones de las cargas productivas entre la nueva curva de carga y las del perfil de referencia. Es decir, utilizando las mediciones de submedición para interpretar la diferencia de consumo de energía, se trata de identificar la causa raíz que ha provocado que la nueva curva de consumo energético de la planta industrial no pertenezca a ningún patrón conocido.

La herramienta provee de un nuevo panel de control que facilita la comparativa del consumo energético global entre una curva nueva y la curva de referencia, y que se muestra en la Figura 4.9. La figura indica el porcentaje de consumo energético que corresponde a cada carga de nivel inferior (carga de submedición) a través de diversos gráficos de tarta que comparan los consumos individuales de energía asociados a los niveles jerárquicos inferiores. Otro indicador matemático que se utiliza para identificar el consumo excesivo de energía - y que se representa en la tabla inferior de la figura - es la correlación entre los residuos (diferencia entre la curva asociada a la alarma energética y el perfil de referencia utilizado para la comparación) y procesos. A mayor correlación más sencillo es identificar la causa de la ineficiencia. Por ejemplo, en la gráfica la correlación del $99,39\,\%$ indica que la principal causa de diferencia en los procesos productivos se debe a los residuos del proceso de Estampación en caliente, y que una correlación del $84,22\,\%$ indica que la principal causa de diferencia en los procesos auxiliares se debe a los residuos de Iluminación. Es decir, los niveles de producción en la nueva curva son menores que los valores de referencia ($81385k$ frente a $100305k$) principalmente porque las máquinas relacionadas con el proceso de Estampación en caliente apenas estuvieron en funcionamiento y, por tanto, como muestra la figura apenas consumieron energía ($96\,\%$ menos de lo esperado). Asimismo, la Figura también muestra que el consumo en Iluminación es muy superior al esperado ($84\,\%$ de sobre consumo). Con esta información el técnico experto de dominio tiene información suficiente para deducir que ese día [2019-07-14] alguien se olvidó de apagar las luces de la planta de producción.

Figura 4.9: Desagregación de cargas productivas y auxiliares para dos periodos de tiempos identificados (el nuevo [2019-07-14] y el de referencia [2019-03-24])



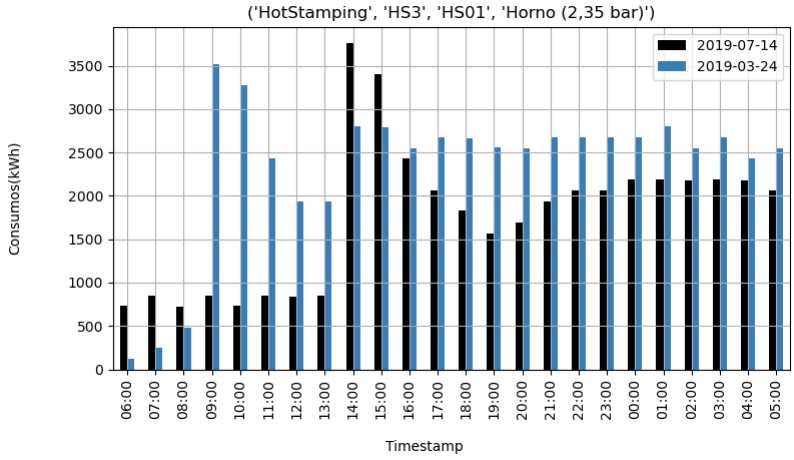
	[2019-07-14]	[2019-03-24]	diff	diff %
Aircompressor	1491.37	2175.75	-684.38	-9.06%
Cooling	858.23	931.39	-73.15	-0.97%
Lighting	18455.99	12091.67	6364.31	84.22%
Other	1195.99	761.43	434.56	5.75%



	[2019-07-14]	[2019-03-24]	diff	diff %
HotStamping	147295.55	173556.42	-26260.87	-96.91%
ColdStamping	4788.47	4362.82	425.66	1.57%
LaserCut	376.39	521.35	-144.96	-0.53%
AssemblyCells	1356.79	1089.87	266.93	0.99%

Este panel de control permite al gestor de eficiencia energética consultar los niveles de sub-medición de los consumos energéticos hasta llegar al nivel máquina. La Figura 4.10 muestra la comparación entre la nueva curva [2019-07-14] y el perfil de referencia [2019-03-24] para las cargas productivas del Horno HS3. La curva del perfil de referencia, en color azul, es la curva más eficiente del comportamiento al que debiera haber sido asignada la nueva curva (cluster #1), y como se observa en la Figura, con mayor consumo energético en HS3 (en el proceso de Estampación en caliente).

Figura 4.10: Comparación de la medición de cargas productivas de un Horno (en el proceso de Estampación en caliente) para la nueva curva ([2019-07-14]) y el perfil de referencia ([2019-03-24])



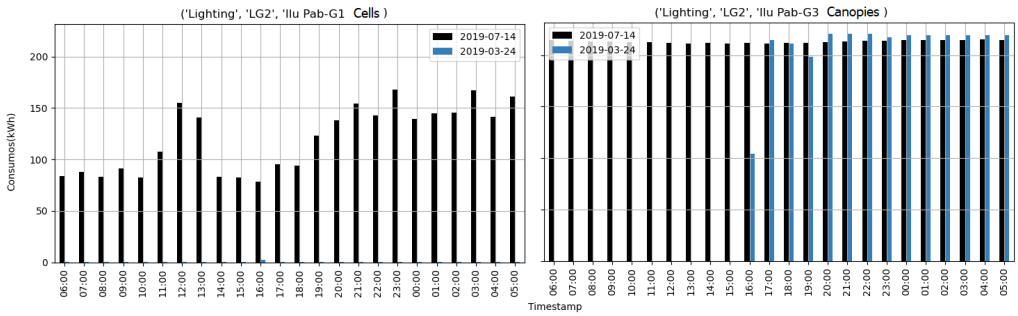
La Figura 4.11 muestra la comparación entre la nueva curva [2019-07-14] y el perfil de referencia [2019-03-24] para las cargas auxiliares de Iluminación en marquesinas y celdas. En ambos casos la nueva curva, en color negro, muestra un consumo energético muy superior al consumido por la curva de referencia, en color azul.

El panel de control demuestra que a través de los datos de submedición, es posible desglosar los porcentajes de consumo de energía atribuibles a los distintos componentes de un determinado nivel de producción (línea, proceso, puesto y máquina). Esto permite, identificar el elemento causante de la ineficiencia energética (por ejemplo, consumo excesivo en iluminación por el olvido de no apagar las luces).

4.2.6. Generación de informes

Para el gestor de eficiencia y los técnicos expertos es importante comprender cómo está funcionando la planta, además de comprender qué cargas energéticas y qué máquinas concretas causan las anomalías. Para ayudar en esta labor de comprensión, el sistema genera dos tipos de informes automáticos basándose en el nivel operativo y de responsabilidad:

Figura 4.11: Comparación de la medición de cargas auxiliares de Iluminación (Iluminación en marquesinas y celdas) para la nueva curva ([2019-07-14]) y el perfil de referencia ([2019-03-24])



- Un informe para el técnico experto de operaciones que necesita información diaria sobre el control de la energía para estimular acciones específicas de ahorro energético relacionadas directamente con el proceso de producción.
- Un informe para los gestores de eficiencia energética que necesitan información resumida con la que orientar el esfuerzo de gestión energética de la organización, con impacto a largo plazo en la eficiencia de la producción y con frecuencia mensual, véase la Figura 4.12.

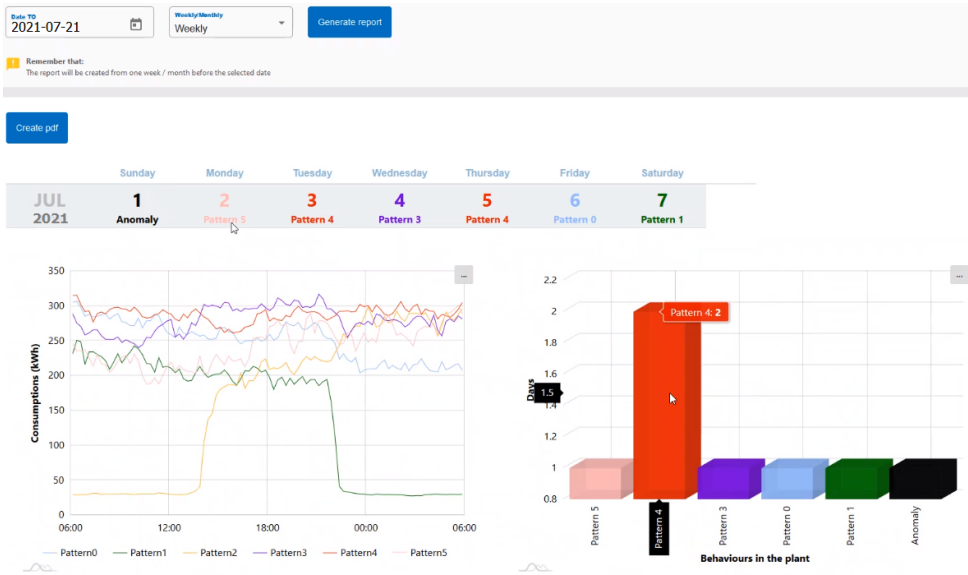
La función de estos informes es:

- Crear motivación para las acciones de ahorro energético.
- Informar regularmente sobre el rendimiento.

4.2.7. Cuantificación económica de los beneficios

Se ha realizado una serie de cálculos estimando la cuantificación en la reducción de consumo y la reducción de emisiones CO_2. Respecto a la cuantificación/estimación de la reducción de consumo, el consumo de energía eléctrica en una de las plantas en 2019 fue de $22,5GWh$ con un coste de 2,200,000N. Gracias a la instalación en planta de la metodología propuesta, se estima la reducción de la factura eléctrica en un 5 %, lo que supone en términos absolutos 150,000N/ao. Extrapolando estos datos al resto del grupo, la empresa tiene un consumo total de $493GWh$ con un coste de 40.000.000 €. Lo que significa 2 millones de euros al año en ahorro económico directo. Respecto a la cuantificación/estimación reducción emisiones CO_2 y gracias a la instalación de la metodología propuesta en una de las plantas se estima la reducción de las emisiones de CO_2 en $531kg$ al año. Extrapolando estos datos al resto del grupo, se estima que se podrían reducir las emisiones de CO_2 en 6.162 kg al año.

Figura 4.12: Interfaz para la creación de informes del funcionamiento de la planta de producción industrial



A fecha de febrero de 2022, esta metodología está implantada en la realidad de cinco plantas de producción de componentes de automoción pertenecientes al mismo grupo industrial.

4.3. Comparativa de experimentos

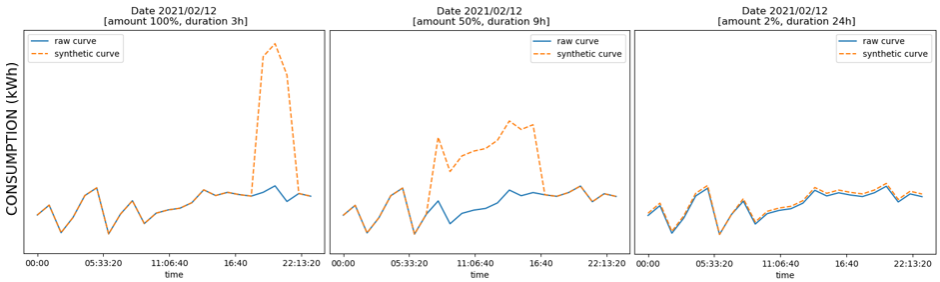
En esta sección se compara el rendimiento de la metodología propuesta, con algoritmos específicos para la detección de anomalías a nivel de agregación de planta industrial.

4.3.1. Descripción de los datos

La fase de entrenamiento utiliza 72 curvas de consumo energético capaces de caracterizar el comportamiento energético y la relación energía vs. producción de la planta industrial.

La etapa de evaluación utiliza 140 curvas de consumo energético: 18 % son nuevas curvas reales de la planta industrial y 82 % son curvas ficticias o pseudo-curvas creadas a partir de las 72 curvas de consumo energético de la fase de entrenamiento. El proceso de creación de anomalías ficticias permite simular los efectos en el consumo de energía observados por los gestores de la eficiencia energética en las anomalías ocurridas en el pasado, y produce un conjunto de datos de evaluación que refleja la diversidad de anomalías que la planta industrial puede encontrar en el futuro. La Figura 4.13 ejemplifica algunas muestras de curvas ficticias.

Figura 4.13: Ejemplos de diferentes anomalías generadas sintéticamente para la fase de evaluación



4.3.2. Métodos base de detección de anomalías

Existen técnicas específicas basadas en datos no supervisados y semi-supervisados especialmente diseñadas para la detección de anomalías. La gama de técnicas revisadas aquí abarca desde técnicas estadísticas bien conocidas y métodos clásicos de ML hasta nuevos enfoques basados en arquitecturas de aprendizaje profundo (DL). Se seleccionan varios métodos de detección de anomalías del estado del arte como líneas de base:

- Algoritmos de detección de anomalías basados en medidas de proximidad entre muestras vecinas cuyo cálculo del índice de anomalía se basa en el cálculo de las distancias y las densidades de los otros vecinos, como Local Outlier Factor (LOF) [271], Connectivity Outlier Factor (COF) [272], Clustering-Based Local Outlier Factor (CBLOF) [272], Histogram-Based Outlier Detection (HBOS) [273], Subspace Outlier Detection (SOD) [274], Local Outlier Correlation Integral (LOCI) [275], k Nearest Neighbors (kNN) [276] and Rotation-based Outlier Detection (ROD) [277].

- Algoritmos de detección de anomalías basados en conjuntos, como Isolation Forest (iSOFOREST) [278] un método de detección de anomalías basado en conjuntos con una complejidad de tiempo lineal y una gran precisión. iSOFOREST utiliza un árbol binario para segmentar los datos. La profundidad del punto de datos en el árbol binario refleja el grado de distancia de los datos.

- Algoritmos basados en métodos kernel y lineales, los datos se transforman de forma no lineal a un espacio de mayor dimensión sin perder las propiedades originales del método lineal, como One-Class Support Vector Machines (OSVM) [279], Principal Component Analysis (PCA) [280], Minimum Covariance Determinant (MCD) [281], Linear Model Deviation-base outlier Detection (LMDD) [282].

- Los algoritmos basados en métodos probabilísticos, dada una distribución de datos de entrada, basan el resultado devuelto en decisiones aleatorias, donde en promedio se obtiene una buena solución al problema,

como Angle-Based Outlier Detection (ABOD) [283], Stochastic Outlier Selection(SOS) [284].

- Algoritmos de aprendizaje profundo, como los autocodificadores (AE) que aprenden correlaciones complejas de los datos y utilizan esta comprensión para codificar puntos de datos de entrada de alta dimensión en una representación de baja dimensión, que contiene la información necesaria para reconstruir el punto de datos de entrada. Los valores atípicos se declaran cuando la calidad del punto de datos reconstruido se degrada gravemente [285]. Variational AutoEncoders (VAE) [286] también pretenden comprimir y reconstruir los datos de entrada. La VAE genera nuevas entradas falsas en el espacio latente mediante una distribución probabilística de los datos reales, mientras que los modelos basados en el codificador lo hacen de forma determinista a partir de un único valor.

- Redes Generativas Adversarias, o en ingles Generative Adversarial Networks (GAN [287]). Las GAN comprenden un sistema de dos redes neuronales (*generador* y *discriminador*) que compiten entre sí en una especie de juego de suma cero finita [288]. El generador tiene como objetivo producir muestras aleatorias que se aproximen a las del conjunto de entrenamiento, mientras que el discriminador aprende a distinguir entre las muestras generadas por la red generadora y las reales. Las dos redes se entrenan simultáneamente en una lucha adversaria. En [289] los autores proponen dos enfoques basados en el aprendizaje activo generativo adversario (Generative Adversarial Active Learning GAAL) para la generación adversaria de valores atípicos basados en el modelado de la distribución de las anomalías reales. Para evitar el problema de colapso de modo (mode collapse problem [290]) que se conoce en el generador de arquitecturas GAN nativas, los autores proponen el uso de un único generador, acuñando el algoritmo Single-Objective Generative Adversarial Active Learning (SO_GAAL), o el uso de múltiples generadores con el algoritmo Multiple-Objective Generative Adversarial Active Learning (MO_GAAL). Otro enfoque basado en GANs es el que los autores proponen en [291] con TAnoGAN, una arquitectura innovadora diseñada específicamente para la detección de valores atípicos en conjuntos de datos especialmente pequeños mediante el uso de GANs. Los autores proponen una red discriminadora compuesta por una capa LSTM y una red generadora basada en varias capas LSTM. Se han requerido algunas modificaciones adicionales para la adecuación a este problema y su inclusión en el benchmark de este estudio. Los autores en [291] prescriben un umbral de 5.000 instancias, por debajo del cual esta red puede experimentar una disminución de su rendimiento. En este estudio, la fábrica necesita al menos 3.360 instancias para definir su rendimiento nominal. Esta diferencia en el volumen de datos obliga a la red a trabajar en condiciones extremas. Para adaptar la red TAnoGAN a la casuística del estudio, ha sido necesario modificar no sólo el valor de sus hiperparámetros, sino también algunas características adicionales: (i) debido a la función hiperbólica tangencial, las variables de entrada se han escalado en el rango [-1,1], (ii) también se han redefinido las épocas, los pesos y

los umbrales de la red discriminadora para maximizar la precisión, (iii) se ha utilizado una ventana deslizante solapada en lugar de una ventana deslizante no solapada, (iv) se ha adoptado un método de inicialización de Xavier para la creación de instancias falsas en el espacio latente, (v) se utiliza un descenso de gradiente estocástico como solución al problema del desvanecimiento o explosión del gradiente [292] y (vi) se han modificado las tasas de aprendizaje para lograr la convergencia de la red. La Tabla 4.1 resume los hiperparámetros utilizados en la definición de la arquitectura TAnoGAN original, y los nuevos hiperparámetros elegidos en este estudio.

Tabla 4.1: Configuración de los hiperparámetros en el enfoque original de TAnoGAN y de la versión propuesta en este estudio

	TAnoGAN original	TAnoGAN propuesto en este trabajo
Épocas	20-30	350
Tamaño del minilote	32	32
Tamaño del conjunto de datos	1,000-22,000 muestras	$105 \cdot 24$ muestras
Tamaño de la instancia	Etiqueta por valor	Etiqueta por instancia (1 instancia = 24 valores)
Inicialización	normal	Xavier normal
Método de optimization	Adam	Stochastic Gradient Descent
Tasa de aprendizaje (LR)	Entrenamiento = 0.0002, Evaluación = 0.01	Entrenamiento = 0.008, Evaluación = 0.01
Tipo de ventana	Ventana deslizante no superpuesta	Overlapping sliding window
Capas Generadoras	3 capas LSTM (32, 64, 128 hidden units)	3 LSTM layers (32, 64, 128 unidades ocultas)
Capas Discriminadoras	1 capa LSTM (100 unidades ocultas)	1 capa LSTM (100 unidades ocultas)

Para la implementación de los experimentos se utiliza el toolbox de Python para detección de valores atípicos escalables (PyOD, [293]). Todos los algoritmos de detección de valores atípicos considerados se enumeran en la Tabla 4.2, junto con el rango de valores óptimos considerados por sus hiperparámetros. Con el fin de replicar este estudio, los hiperparámetros se nombran como en la Toolbox utilizada para la implementación [293].

4.3.3. Métricas de Evaluación

Cabe mencionar que no existen métricas y esquemas unificados para evaluar el rendimiento de los algoritmos de detección de anomalías [294]. Así, una comparación justa entre diferentes enfoques de detección de anomalías debe llevarse a cabo utilizando un conjunto de métricas estándar, y debe realizarse en las mismas condiciones, por ejemplo, utilizando el mismo conjunto de datos [294]. Se han utilizado seis métricas para evaluar la calidad de la solución:

- Tasa de falsos positivos (FPR), la FPR se calcula como la relación entre el número de muestras negativas erróneamente categorizadas como positivas (FP) y el número total de muestras negativas (FP + TN).

$$FPR = \frac{FP}{FP + TN}, \tag{4.7}$$

Tabla 4.2: Rangos de valores de hiper-parámetros explorados para los algoritmos de detección de anomalías en la prueba de referencia. $\mathbb{N}[a, b]$ representa todos los números naturales entre a y b; $\mathbb{R}[c, d, e]$ representa los números de valor real uniformes (cada e) entre c y d; finalmente, $\exp[f, g]$ representa el conjunto $\{10^f, 10^{f+1}, \ldots, 10^g\}$. El ajuste de los hiperparámetros no mostrados en la tabla no tuvo ningún impacto en la calidad de las soluciones, y se establecieron sus valores por defecto

Algoritmos	Hiper-parámetros
LOF	$n\_neighbors \in \mathbb{N}[1, 20]$, $p \in \mathbb{N}[1, 3]$
ISOFOREST	$n\_estimators \in \mathbb{N}[10, 100, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
OSVM	$gamma \in \exp[-3, 5]$
PCA	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
MCD	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
LMDD	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
COF	$n\_neighbors \in \mathbb{N}[1, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
CBLOF	$n\_clusters \in \mathbb{N}[2, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
HBOS	$n\_bins \in \mathbb{N}[1, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $alpha \in \mathbb{R}[0{,}05, 1, 0, 05]$
SOD	$n\_neighbors \in \mathbb{N}[1, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
ABOD	$n\_neighbors \in \mathbb{N}[1, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
LOCI	$k \in \mathbb{N}[2, 11]$, $alpha \in \mathbb{R}[0{,}05, 1, 0, 05]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
KNN	$n\_neighbors \in \mathbb{N}[1, 20]$, $contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $p \in [1, 2, 3]$, $method \in ['largest', 'mean', 'median']$
ROD	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$
SOS	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $perplexity \in \mathbb{N}[1, 9]$
AutoEncoder	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $l2\_regularizer \in \mathbb{R}[0{,}0, 1{,}0, 0{,}1]$, $dropout\_rate \in [0{,}0, 0{,}1, 0{,}2]$, $hidden\_neurons = [24, 12, 12, 24]$
VAE	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $gamma \in [0{,}5, 0{,}75, 1., 2.]$, $capacity \in [0{,}0, 0{,}1, 0{,}2, 0{,}4]$, $dropout\_rate \in [0{,}0, 0{,}1, 0{,}2]$, $encoder\_neurons = [24, 12]$, $decoder\_neurons = [12, 24]$
SO_GAAL	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $lr\_d \in \mathbb{R}[0{,}001, 0{,}1, 0{,}005]$, $lr\_g \in \mathbb{R}[0{,}001, 0{,}1, 0{,}005]$
MO_GAAL	$contamination \in \mathbb{R}[0{,}05, 0{,}5, 0{,}05]$, $k \in [3, 5, 7, 10]$, $lr\_d \in \mathbb{R}[0{,}001, 0{,}1, 0{,}005]$, $lr\_g \in \mathbb{R}[0{,}001, 0{,}1, 0{,}005]$

- Precision, mide la probabilidad de que una muestra clasificada como anomalía sea realmente una anomalía.

$$Precision = \frac{TP}{TP + FP}, \tag{4.8}$$

- Tasa de verdaderos positivos (TPR) o Recall, cuantifica la proporción de falsos positivos (TP) dentro de las muestras negativas (TP+FN). La precisión se centra más en la clase positiva que en la negativa, ya que en realidad mide la probabilidad de detección correcta de los valores positivos, mientras que la TPR y la recuperación (métrica ROC) miden

la capacidad de distinguir entre las clases.

$$Recall = TPR = \frac{TP}{TP + FN}, \tag{4.9}$$

- Balanced accuracy (bAcc), es la media de Recall/TPR y la tasa de verdaderos negativos (TN), y amplía el concepto de exactitud al caso de datos desequilibrados. Tiene en cuenta las muestras positivas y negativas, y no se ve influida por las diferencias en la representación de las distintas clases en los datos de prueba.

$$bAcc = \frac{Recall + Specificity}{2}, \text{where} \tag{4.10}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{4.11}$$

- Medidad F1, viene dada por la media armónica de la Precision y Recall, teniendo en cuenta ambas métricas para penalizar los valores extremos.

$$F1 = 2 * \frac{Precision \cdot Recall}{Precision + Recall}, \tag{4.12}$$

- Matthews Correlation Coefficient (MCC), tiene en cuenta los verdaderos negativos, los verdaderos positivos, los falsos negativos y los falsos positivos. La principal diferencia entre el MCC y la puntuación F1 es que el F1 ignora el recuento de verdaderos negativos, mientras que el MCC considera las cuatro entradas de la matriz de confusión. En consecuencia, MCC arroja valores altos si el clasificador lo hace bien en esas cuatro medidas (muestras positivas y negativas), constituyendo así una métrica más fiable en el caso de clases desequilibradas.

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)))}{\sqrt{TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{4.13}$$

4.3.4. Resultado del experimento

La Tabla 4.3 recoge los resultados de la comparación de rendimiento en términos de las métricas mencionadas, a saber, precisión, recall, FPR, F1, precisión equilibrada y coeficiente de correlación de Matthews. La configuración de los hiper-parámetros de cada modelo cuyos resultados se recogen en esta tabla corresponde a la que conduce al mayor valor de precisión equilibrada sobre el conjunto de datos de prueba. La información de esta tabla se complementa con los gráficos anidados en la Figura 4.14, que presenta gráficamente la matriz de confusión de los diferentes algoritmos, junto con la de la metodología anteriormente propuesta en la parte más a la derecha de la Figura.

El estudio de la comparativa entre experimentos revela que la *Precisión* es la métrica que presenta la menor variabilidad entre los algoritmos, con valores entre $0,86$ y $0,98$. Este resultado indica que la mayoría de

Tabla 4.3: Métricas obtenidas para varios algoritmos de detección de anomalías y el enfoque de la metodología propuesta: Precision, recall, FPR, balanced accuracy, F1-measure and Matthews Correlation Coefficient (MCC). Los mejores resultados para cada indicador se destacan en azul

Algoritmo	Precision	TPR/Recall	FPR	F1	bAcc	MCC
LOF	**0,986**	0,589	0,057	0,737	0,766	0,360
ISOFOREST	0,933	0,779	0,371	0,849	0,704	0,309
OSVM	0,928	0,784	0,400	0,850	0,692	0,293
PCA	0,933	0,784	0,371	0,852	0,706	0,314
MCD	0,984	0,532	0,057	0,691	0,738	0,322
LMDD	0,868	**1,000**	1,000	**0,930**	0,500	0,000
COF	0,887	0,511	0,429	0,648	0,541	0,056
CBLOF	0,926	0,649	0,343	0,763	0,653	0,212
HBOS	0,935	0,810	0,371	0,868	0,719	0,343
SOD	0,964	0,576	0,143	0,721	0,716	0,293
ABOD	0,965	0,719	0,171	0,824	0,774	0,387
LOCI	0,970	0,139	**0,029**	0,242	0,555	0,113
KNN	0,969	0,667	0,143	0,790	0,762	0,361
ROD	0,922	0,615	0,343	0,738	0,636	0,186
SOS	0,915	0,325	0,200	0,479	0,562	0,091
AutoEncoder	0,936	0,758	0,343	0,837	0,707	0,307
VAE	0,937	0,775	0,343	0,848	0,716	0,325
SO_GAAL	0,974	0,654	0,114	0,782	0,770	0,370
MO_GAAL	0,932	0,416	0,200	0,575	0,608	0,150
TAnoGAN	0,88	0,766	0,686	0,625	0,54	0,063
Metogología propuesta	0,947	0,82	0,045	0,879	**0,887**	**0,549**

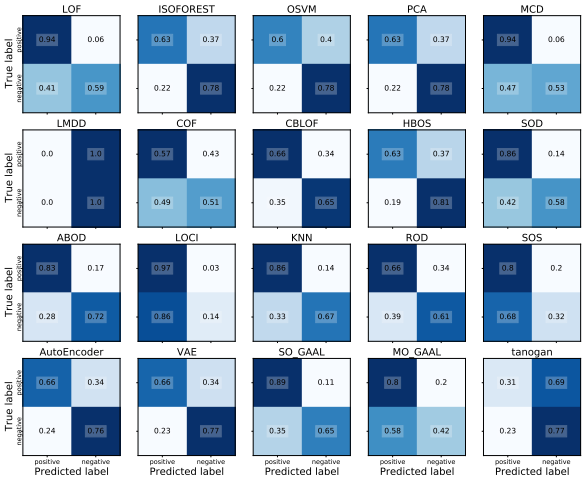
las anomalías detectadas son identificadas como tales por los algoritmos. Por el contrario, la métrica *Recall* sufre la mayor variabilidad, con valores que oscilan entre 0,13 y 1. En particular, el algoritmo LMDD (Linear Model Deviation-base outlier Detection) arroja valores de *Precisión* $0,86$, *Recall* $1,0$ y *F1* $0,93$, lo que podría llevar a la conclusión de que este enfoque de detección de anomalías funciona muy bien para la aplicación en cuestión. Sin embargo, el valor de la métrica MCC (Matthews Correlation Coefficient) es 0, lo que, junto con un valor de *FPR* de $1,0$, sugiere que este modelo funciona mal cuando predice muestras negativas. La matriz de confusión correspondiente a este modelo en la Figura 4.14.A. corrobora esta afirmación: LMDD siempre declara que se ha producido un valor atípico. Esto ejemplifica la necesidad de calcular y registrar medidas de rendimiento que sean sensibles a la presencia de desequilibrio de clases en el conjunto de datos de prueba.

El estudio de la métrica *FPR* indica que el algoritmo LOCI (Local Outlier Correlation Integral) alcanza un valor de $0,029$, algo más de la mitad del *FPR* respecto a la solución propuesta $0,045$. Visto de otro modo, LOCI detecta un $1,6\,\%$ más de falsos positivos que la metodología objeto de este capítulo. Sin embargo, la matriz de confusión de este algoritmo muestra que es muy conservador y no detecta la mayoría de los verdaderos positivos (valores atípicos). Además, LOF, MCD, SOD, ABOD, LOCI, SO_GAAL y KNN alcanzan valores de precisión elevados, entre $0,96$ y

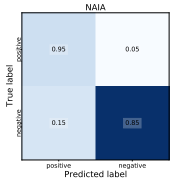
0, 98, mientras que el rendimiento de la metodología propuesta en términos de la misma puntuación es de 0, 94, que es ligeramente inferior, aunque está mejor equilibrado con respecto a su *Recall*. De hecho, la metodología propuesta supera al resto de las pruebas de referencia en cuanto a la métrica de recuerdo 0, 82, con la excepción de LMDD que, como se ha mencionado anteriormente, está extremadamente sesgado hacia una clase. Como resultado de sus buenos estadísticos de detección tanto de valores atípicos como de valores inliers, la mejor puntuación de precisión equilibrada la obtiene de nuevo la metodología propuesta, con un valor de 0, 88, seguido de ABOD, que se queda atrás con más de 10 % menos de valor de esta métrica 0, 77. A excepción de LMDD, con una puntuación *F1* de 0, 93, la metodología propuesta también domina el punto de referencia 0, 87. Por último, la puntuación *MCC* de la metodología propuesta es de 0, 54, por delante del segundo mejor (ABOD), con un valor de 0, 38. Este último resultado es notable, dado que se sabe que el *MCC* es un indicador fiable del rendimiento de un modelo en circunstancias con distribuciones de clase muy sesgadas.

Figura 4.14: Matrices de confusión correspondientes a los modelos considerados en el estudio comparativo (izquierda), y la de la metodología propuesta. La metodología propuesta logra un buen equilibrio de detección de ejemplos positivos y negativos, lo que se suma a la transparencia e interpretabilidad de sus principales etapas de procesamiento detalladas en Sección 4.2.

(a) Resto de algoritmos de la comparativa



(b) Metodología propuesta



Conforme a lo anterior, los resultados obtenidos por la metodología propuesta son consistentes en todas las métricas y superiores a los correspondientes a otras opciones de modelización del estado del arte. Esto subraya la importancia de tener en cuenta el conocimiento del dominio disponible y el público al que se dirige cuando se diseñan soluciones basadas

en la IA. Aunque esta aproximación en primera instancia podría indicar valores peores para las métricas, éstas están preconcebidas para datos con porcentajes similares de muestras para cada clase (es decir, para clases equilibradas). Este no es ciertamente el caso cuando se trata de la eficiencia energética en los procesos industriales, donde los eventos anómalos de consumo de energía son raros, pero, a menos que se detecten y se aborden adecuadamente, tienen un impacto dramático en la eficiencia de costes de la fábrica. Las métricas para las que la metodología propuesta muestra mejores resultados y que están menos influenciadas por los datos desequilibrados son *bAcc* y *MCC*. Es decir, entre todos los esquemas utilizados para la detección de anomalías, la metodología propuesta es el esquema que mejor equilibra entre los falsos positivos y los verdaderos positivos, como se expone claramente en su matriz de confusión de la Figura 4.14.B.

4.4. Conclusiones

En este capítulo se ha propuesto una novedosa metodología que identifica los comportamientos nominales de una Planta de Producción Industrial en base a su consumo energético y su correlación con la tasa de producción. La metodología de forma inherente, además de la modelización de los procesos industriales, ha permitido la detección de ineficiencias energéticas y la prospección de la causa raíz de estas. Las curvas de consumo o cargas y la jerarquía de monitorización energética de la planta permiten la descomposición de los consumos energéticos a diferentes niveles de desagregación: a nivel de máquina, proceso o línea productiva. La metodología, desde su conceptualización, ha perseguido la simplicidad y la comprensibilidad en su diseño, evitando modelos de caja negra que podrían obstaculizar la confianza de las partes interesadas relevantes en la planta (gestores de eficiencia energética y técnicos de dominio). En virtud de esta simplicidad (pero no menos eficientes funcionalidades basadas en la IA), la metodología ha implementado una herramienta totalmente interpretable capaz de interpretarse y entenderse asumiendo únicamente conocimientos básicos de regresión lineal, coeficiente de determinación y agrupación de comportamientos en base a distancias euclídeas) que son comunes a muchas disciplinas. Asimismo, la herramienta está provista de un panel de control con capacidades visuales interactivas para identificar aquellas cargas causantes de las anomalías identificadas, lo cual es imprescindible dada la enorme complejidad y heterogeneidad de los grandes procesos de producción. En consecuencia, un usuario sin conocimientos analíticos puede entender las decisiones generadas desde la herramienta, interactuar con ellas y, por tanto, entrar en el bucle de decisión sin mostrar ningún tipo de reticencia.

Más allá de los beneficios inherentes a la simplicidad y la interactividad, el estudio también ha analizado los resultados de la metodología sobre un conjunto de datos del mundo real capturados sobre una planta de fabricación industrial del sector de la automoción. Las comparaciones con un conjunto de otros 20 algoritmos de detección de valores atípicos

utilizados a menudo en estudios similares han sido concluyentes: la metodología no sólo se aleja de la naturaleza de caja negra de los métodos modernos para la detección de valores atípicos (por ejemplo, los modelos neuronales generativos), sino que también logra estadísticas de detección superiores. En concreto, los resultados cuantitativos de estos experimentos han revelado que la metodología sobresale en el aprendizaje de conjuntos de datos extremadamente desequilibrados, lo que enriquece aún más su interpretabilidad con una sólida solidez frente a los sesgos de aprendizaje resultantes del desequilibrio de clases. Esto informa con evidencia empírica de la principal lección aprendida en el presente estudio: la sofisticación de los modelos en los problemas industriales debe optarse por alternativas más simples, no necesariamente de peor rendimiento.

Este estudio que fue acometido con el objetivo de la optimización de líneas industriales, procesos, puestos y/o máquinas, a traves de la identificación de la causa raíz causante de las ineficiencias energéticas, ha conseguido un ahorro energético respecto a la situación original del 5 % del consumo total de la Planta de Producción.

Parte III

Observaciones finales

Capítulo 5

Aportaciones y Conclusiones

En este capítulo se presentan las principales aportaciones científicas y tecnológicas derivadas de esta Tesis, así como reconocimientos a la calidad de sus resultados. Se describen las lecciones aprendidas a nivel de investigación en este campo y se indican las líneas futuras de actuación que derivan de las conclusiones presentadas.

5.1. Aportaciones de la Tesis

Pese a que las evidencias nos indican que la Inteligencia Artificial está alcanzando niveles de desempeño nunca vistos, ¿por qué la Industria no adopta masivamente esta tecnología, aun habiendo ecosistemas digitales cada vez más maduros?

Esta Tesis ejemplifica la utilización de la Inteligencia Artificial aplicada a la realidad de tres casos de investigación científica industrial no-supervisados: la optimización en el equilibrado de cargas en redes eléctrica, el control y supervisión de procesos industriales mediante sensores virtuales, y la gestión de la eficiencia energética en plantas de producción industrial.

Mediante la investigación llevada a cabo en esta tesis, se demuestra, mediante contribuciones reales que es posible incorporar elementos de interpretabilidad en problemas industriales, a través de la codificación de leyes físicas, técnicas de inteligencia artificial y algoritmos estadísticos, conservando la precisión predictiva. Se demuestra que es posible democratizar la tecnología para garantizar su utilización y, por tanto, facilitar su adopción y proliferación efectiva para otros casos de uso industrial.

Capítulo 2. Contribuciones meta-heurísticas para la Optimización en el equilibrado de cargas en Redes de Baja Tensión
En el contexto de sistemas desequilibrados en la red de Baja Tensión, no siempre existe un equilibrio de cargas entre las tres fases de una Línea, CT o Caja General de Protección. Este trabajo de Tesis ha propuesto una metodología para el desarrollo de soluciones de optimización heurísticas y meta-heurísticas para establecer sistemas equilibrados, a nivel de cargas de fases, en la red eléctrica de Baja Tensión. La herramienta permite una mejor asignación de cargas a fases en base al nivel de congestión de la fase y de la complementariedad entre los consumidores. El estudio ha contemplado la definición de métricas específicas de dominio y técnicas de optimización metaheurísticas, alimentadas con información real de contadores telegestionados. Finalmente se ha concluido, que no hay una única solución ideal, y que dependiendo del tamaño y de la complejidad del espacio de búsqueda, la solución óptima puede ser un algoritmo de búsqueda estocástico tradicional (basado en conocimiento de dominio) para espacios de soluciones pequeños o algoritmos agnósticos de búsqueda no exacta (probabilístico meta-heurístico) para espacios de búsqueda de grandes dimensiones.

Capítulo 3. Contribuciones predictivas para el Control y Supervisión de procesos industriales mediante sensores virtuales
En el contexto de control y supervisión de procesos industriales no siempre todas las variables del proceso son fácilmente medibles en tiempo real. Este trabajo de Tesis ha propuesto un procedimiento para el desarrollo de soluciones predictivas de inferencia y modelado de sensores virtuales para control y supervisión en plantas industriales. La metodología, basándose en la virtualización de sensores mediante estrategias de aprendizaje

adaptativas, es capaz de inferir el valor en tiempo real de dichas variables críticas a partir de otras variables del proceso fáciles de medir. Se valida la generalización de la metodología, en tres casos de investigación industrial reales que, aunque dispares en concepto son afines en técnicas de preprocesamiento, técnicas de selección de variables relevantes, técnicas de reducción de la dimensionalidad y estrategias de aprendizaje adaptativo. Los tres casos de uso están relacionados con la inferencia de la temperatura del punto de inflamación flash en una planta Petroquímica para el proceso de la desulfuración del gasóleo de automoción, la inferencia de componentes químicos en una planta química para el proceso del cracking del etileno y la inferencia de emisiones contaminantes en Planta de Reciclaje para la propagación de partículas PM_{10} a áreas limítrofes. Los resultados cuantitativos de estos tres casos han revelado que el procedimiento propuesto sobresale en la inferencia de valores críticos sobre los métodos actuales, hecho que ha permitido disponer además de una herramienta de detección de anomalías basada en la diferencia sustancial entre ambos valores.

Capítulo 4. Contribuciones descriptivas en la Gestión de la Eficiencia Energética en plantas de Producción Industrial

En el contexto de plantas de producción basadas en sistemas ciberfísicos, las máquinas, procesos y/o líneas productivas y no-productivas han de intercambiar continuamente datos entre sí. Este trabajo de Tesis ha propuesto una metodología para el desarrollo de soluciones descriptivas para el modelado de la eficiencia energética en plantas de producción para la detección de ineficiencias y el descubrimiento de la causa raíz de estas. La metodología ha sido capaz de inferir los comportamientos de normalidad de la planta a partir de la jerarquía de monitorización energética y de las curvas de consumos, a través de sencillas técnicas descriptivas clásicas. Abordar este caso desde la doble perspectiva del conocimiento del dominio y del modelado basado en datos ha añadido potencialidad e interpretabilidad al resultado final. Asimismo, la capacidad de la metodología de indagar en la submedición de los consumos energéticos ha posibilitado el análisis de la causa raíz que origina tales ineficiencias energéticas. Los resultados de los experimentos de diferentes técnicas de detección de anomalías han revelado que la metodología propuesta sobresale en el aprendizaje de conjuntos de datos extremadamente desequilibrados y que la sofisticación de los modelos en los problemas industriales debe optarse por alternativas más simples y explicables.

Finalmente, este trabajo de Tesis ha demostrado que, ante la necesidad de la Industria de disponer de aplicaciones de Inteligencia Artificial aplicadas a casos reales, el verdadero esfuerzo consiste en la adecuación de las técnicas a la realidad del problema. Actualmente y con las limitaciones de cómputo actuales, es del todo innecesario continuar en la creación de nuevas técnicas analíticas, cuando la adecuación de las técnicas clásicas es suficiente para la resolución de problemas reales. Es una obligación de servicio y moral transmitir estas experiencias, que han de servir para mejorar la competitividad productiva del ecosistema industrial actual.

5.2. Diseminación de resultados

Con objeto de transmitir los resultados de la investigación, no solo al mundo académico sino al público general, la presente Tesis ha realizado tareas de diseminación científica en revista y congresos, así como tareas de diseminación tecnológica - propias de un proyecto de investigación industrial - como patentes, registro de software y actividades de divulgación. Fruto de este trabajo, ha recibido reconocimientos a la calidad del mismo.

5.2.1. Diseminación de investigación

- **Publicaciones de Revista:**
  - Izaskun Mendia, Javier Del Ser, Sergio Gil-López, **"A Novel Approach for the Detection of Anomalous Energy Consumption Patterns in Industrial Cyber-physicals Systems"** (17 págs.), en Revista Expert Systems 2022 (Q2 en la categoría Computer Science, Theory and Methods, JCR 2020 2.587).

    DOI: https://onlinelibrary.wiley.com/doi/10.1111/exsy.12959

    Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 4 (*"Contribuciones descriptivas para la caracterización y gestión de la eficiencia energética en plantas industriales de producción"*), y describe la metodología que permite la caracterización del rendimiento nominal de una fábrica manufacturera del sector de la automoción.
  - Izaskun Mendia, Sergio Gil-López, Itziar Landa-Torres, Lucía Orbe Erik Maqueda **"Machine Learning based adaptive soft sensor for flash point inference in a refinery realtime process"** (11 págs.), en Revista Results and Engineering (Q2, SJR 2020 sin determinar). DOI: https://doi.org/10.1016/j.rineng.2022.100362

    Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 3 (*"Contribuciones predictivas para la mejora de la monitorización de plantas industriales con sensores virtuales"*), y describe el proceso de inferencia de la temperatura de flash en un proceso de refino en base a las relaciones no lineales con las variables de entrada del proceso.
- **Publicaciones de Conferencia:**
  - Mingozzi, E., Tanganelli, G., Vallati, C., Martínez, B., Mendia, I., Gonzalez-Rodriguez, M. (2016, June). **"Semantic-based context modeling for quality of service support in IoT platforms"**. In 2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM) (pp. 1-6).

    Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 2 (*"Contribuciones meta-heurísticas para la optimización de la distribución de energía en redes eléctricas"*), y describe cómo la gestión semántica del contexto puede ser utilizada para inferir conocimiento a través del razonamiento semántico.

- Mendia, I., Gil-López, S., Del Ser, J., Bordagaray, A. G., Prado, J. G., Vélez, M. (2017, February). **"Optimal phase swapping in low voltage distribution networks based on smart meter data and optimization heuristics"**. In International Conference on Harmony Search Algorithm (pp. 283-293). Springer, Singapore.

  Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 2 (*"Contribuciones meta-heurísticas para la optimización de la distribución de energía en redes eléctricas"*) e introduce una versión modificada del algoritmo HS como algoritmo de optimización para el intercambio de consumidores a fases en redes de distribución de BT.

- Mendia, I., Gil-Lopez, S., Del Ser, J., Grau, I., Lejarazu, A., Maqueda, E., Perea, E. (2020, November). **"An Intelligent Procedure for the Methodology of Energy Consumption in Industrial Environments"**. In International Conference on Intelligent Data Engineering and Automated Learning (pp. 92-103). Springer, Cham.

  Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 4 (*"Contribuciones descriptivas para la caracterización y gestión de la eficiencia energética en plantas industriales de producción"*), e introduce un caso practico de sistema industrial ciber-físico (ICPS) y su integración de funcionalidades basadas en la IA.

5.2.2. Diseminación de interés industrial

- **Patentes:**

  - **"Method, system and computer program product for evaluation of energy consumption in industrial environments"** Fecha de la patente Expedida el 8 julio 2018. Número y entidad emisora de la patente EU, EP3677976 (P204127ES, P204128FR, P204129DE).

    Esta patente está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 2 (*"Contribuciones meta heurísticas para la optimización de la distribución de energía en redes eléctricas"*) y que protege el proceso y método de optimización de la asignación de una pluralidad de contadores inteligentes a un Centro de Transformación inteligente. Tecnalia cree en que es una patente con impacto en el futuro impacto y ha sido recientemente renovada (el 21 de diciembre 2020).

  - **"Computer-implemented method for assessment of a power distribution grid"**. Fecha de la patente Expedida el 31 marzo 2022. Número y entidad emisora de la patente EU, EP22382305.5.

    Esta patente está basada en estudios realizados a partir de los resultados y aportaciones obtenidos en el Capítulo 2 (*"Contribuciones meta-heurísticas para la optimización de la distribución de*

energía en redes eléctricas"). La patente protege un novedoso método basado en datos para evaluar el estado de la red de distribución de energía.

- **"Assignation and connection of electricity customers to phases of a distribution feeder"**. Fecha de la patente Expedida el 26 junio 2019. Número y entidad emisora de la patente es, EP3502627A1.

  Esta patente está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 4 (*"Contribuciones descriptivas para la caracterización y gestión de la eficiencia energética en plantas industriales de producción"*) y que protege el proceso y método de un conjunto de herramientas genéricas para el diagnóstico de (in)eficiencias energéticas en plantas de producción industrial basado a la relación Consumo vs Producción. Tecnalia tiene copropiedad con la empresa que lo está explotando.

- **Registro de Software:**

  - **"SW Soft-Sensing"**. Fecha de registro de software Expedida el 2 junio 2018. Número de solicitud: 1-6269344511.

    Esta publicación está relacionada con los estudios, aportaciones y resultados obtenidos en el Capítulo 3 (*"Contribuciones predictivas para la mejora de la monitorización de plantas industriales con sensores virtuales"*)

- **Diseminación de divulgación industrial:**

  - Ponente del Congreso **"Basque Industry 4.0"**. Fecha de la potencia: junio 2018.

    Asistencia como ponente en el taller relativo a "Sensórica e Inteligencia Artificial" con la charla "Sensores virtuales: inferir lo indetectable y asegurar lo inestable", explicando las virtudes y forma de operar de los sensores virtuales que se introducen en el Capítulo 3: *"Contribuciones predictivas para la mejora de la monitorización de plantas industriales con sensores virtuales"*.

5.2.3. Reconocimientos a la calidad de los resultados

- **"Best use of Data Science for Industry 4.0"**. Fecha del reconocimiento: septiembre 2019. Entidad emisora del reconocimiento: European DatSci Awards 2019.

  Los premios DatSci, promovidos por BDVA (Big Data Value Association), reconocen cada año las mejores iniciativas europeas de Big Data, tratamiento de datos e inteligencia artificial y su aplicación en la industria y en la sociedad. En septiembre 2019, como reconocimiento a la calidad de los resultados obtenidos en el Capítulo 4 de la presente Tesis, se le otorga el premio en la categoría "Mejor uso de la Ciencia de Datos/Inteligencia Artificial para la Industria 4.0".

- **"Research and development of artificial intelligence applied to industrial plants"**. Fecha del reconocimiento: febrero 2020. Entidad emisora del reconocimiento: Advanced Factory Awards 2020.

  Reconocimiento a los resultados obtenidos en el Capítulo 4 de la presente Tesis por su labor como aplicación de investigación y desarrollo de Inteligencia Artificial aplicada en plantas industriales.

- **"Caso Práctico de Innobasque"**. Fecha del reconocimiento: diciembre 2021. Entidad emisora del reconocimiento: Innobasque 2021.

  Las aportaciones, estudios y resultados en el caso de uso para la inferencia de la temperatura de Flash del Capítulo 3 ha sido seleccionado por Innobasque como experiencia innovadora implantada y con resultados satisfactorios en el ámbito de la transformación digital aplicada. Incluido como referencia en el Banco de Casos Prácticos de Innobasque [295].

5.3. Lecciones aprendidas

Muchas!. Muchas son las lecciones aprendidas en el periodo de elaboración de estas Tesis. A modo resumido, las siguientes son las más determinantes:

- Trabajar con datos reales conlleva un esfuerzo extra principalmente en tareas de preprocesamiento. El dato hay que trabajarlo hasta que se esté en condiciones de aplicar inteligencia sobre él. Los tratamientos más habituales son: la ausencia de dato, datos con ruido, anomalías, identificación de procesos intermedios que no son el objeto de la modelización y, por tanto, es necesario identificarlos y eliminarlos. Reconocer cuál es la información qué realmente aporta valor y construir los modelos a partir de esa información de calidad, condiciona las soluciones que se proponen. La operativa es muy distinta a cuando se trabaja con datos sintéticos o en entornos controlados.

- Dado un mismo problema, puede haber un gran número de aproximaciones capaces de resolverlo. Resolver problemas complejos es un reto que requiere muchos esfuerzos y un profundo conocimiento de las estrategias y herramientas de resolución de problemas. A menudo la solución surge a medida que se exploran los datos. Es recomendable recurrir al pensamiento paralelo en la ideacion de soluciones creativas.

- Siempre hay espacio para mejorar. La continua búsqueda de la excelencia en soluciones cuantificables por métricas, es siempre posible a través de nuevas técnicas a cada cuál más prometedora que la anterior. Es un proceso iterativo agotador. Por ello, es importante aprender a celebrar los logros, y aprender a reconocer la excelencia en la simplicidad de las técnicas y en su comprensión.

- Para saber comprender la importancia de los datos hay que comprender su dominio de aplicación. Un buen profesional del dato debiera evitar interminables procesos iterativos de prueba y error en la búsqueda de la mejor precisión del modelo.
- Proponer enfoques incrementales en complejidad en la resolución de problemas.
- Transmitir que las soluciones basadas en Ciencia de Datos no tienen como objetivo suplir el conocimiento adquirido a través años de experiencia del humano. Sino ayudar a transformar e intentar dotar de más capacidades a las metodologías clásicas. Estas soluciones se centran en transformar el conocimiento de dominio adquirido por el trabajador a través de la experiencia, y convertirlo en (i) parámetros de entrada a los esquemas numéricos, (ii) en funciones de coste específicas que definan de una manera biunívoca y precisa el espacio solución o (iii) incluir la compatibilidad de los sistemas numéricos basados en Ciencia de Datos con las leyes de la física que rigen el comportamiento de dichos sistemas, máquinas o procesos es clave a la hora de conferir confianza en los resultados provistos.

5.4. Futuras Líneas de Investigación

Considerando distintos aspectos de las áreas de conocimiento relacionadas con las contribuciones propuestas en estos estudios, se han identificado diferentes líneas de investigación para futuros trabajos. En esta sección se presentan algunas de las más destacadas.

- Investigar en el concepto de IA consistente con los principios Físicos (Physics-Aware IA). Se trata de utilizar técnicas de Inteligencia Artificial, que tienen en cuenta las leyes físicas, para dotar a los modelos de ciertos matices del proceso de forma indirecta (a traves de determinadas características, de la identificación del espacio de estados o de la definición de la función de coste). Esta área de estudio tiene como objetivo incorporar restricciones físicas -como la ley conservación de la masa, la ley de conservación de la energía o la ley de conservación del momento- a las arquitecturas de aprendizaje automático profundo.
- Investigar en el concepto de Aprendizaje Automático Automatizado (ML automatizado o AutoML). Se trata del proceso de automatizar ciertas acciones mediante aprendizaje automático, sin necesidad de tener conocimientos de IA para la identificación de los modelos más idóneos. Esta democratización de las tecnicas se puede aplicar a las labores de pre-procesamiento de los datos, de selección del algoritmo idóneo, optimización de hiper-parámetros, selección de métricas e incluso, visualizaciones e implementación de interfaces de usuario.
- Investigar e integrar las soluciones que se han presentado en esta tesis en ecosistemas digitales que den soporte al ciclo de vida del dato.

Con presencia en la ingesta, preparación, entrenamiento y entrega del dato. Lo que da pie a entrenar modelos con más rapidez, escalado y productificación en variedad de entornos on-premise o en cualquier servicio gestionado en Cloud, de forma eficiente. Ya existen plataformas de analítica de colaboración que conforman estos ecosistemas a través de herramientas como Azure Synapse o Databricks.

- Y finalmente, investigar e integrar soluciones en la incipiente computación cuántica. A corto plazo, será posible desarrollar aplicaciones basadas en problemas de optimización, pero a medio plazo la computación cuántica será capaz de proporcionar el salto cualitativo y cuantitativo que la inteligencia artificial necesita para abordar problemas más complejos.

Bibliografía

[1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld y M. Hoffmann, «Industry 4.0,» *Business & information systems engineering*, vol. 6, n.º 4, págs. 239-242, 2014.

[2] B. Dafflon, N. Moalla e Y. Ouzrout, «The challenges, approaches, and used techniques of CPS for manufacturing in Industry 4.0: a literature review,» *The International Journal of Advanced Manufacturing Technology*, págs. 1-18, 2021.

[3] M. Hermann, T. Pentek y B. Otto, «Design principles for industrie 4.0 scenarios,» en *2016 49th Hawaii international conference on system sciences (HICSS)*, IEEE, 2016, págs. 3928-3937.

[4] R. Duray, «Mass customization origins: mass or custom manufacturing?» *International Journal of Operations & Production Management*, 2002.

[5] A. G. Frank, G. H. Mendes, N. F. Ayala y A. Ghezzi, «Servitization and Industry 4.0 convergence in the digital transformation of product firms: A business model innovation perspective,» *Technological Forecasting and Social Change*, vol. 141, págs. 341-351, 2019.

[6] H. von Scheel, «2nd wave of Industry 4.0 by Henrik von Scheel 2019,» mayo de 2019.

[7] H. Von Scheel, *Putting the Industry 4.0 into Practice*. Morgan Kaufmann, 2021.

[8] A. Ali, W. Hamouda y M. Uysal, «Next generation M2M cellular networks: challenges and practical considerations,» *IEEE Communications Magazine*, vol. 53, n.º 9, págs. 18-24, 2015.

[9] A. M. Turing, «Computing machinery and intelligence,» en *Parsing the turing test*, Springer, 2009, págs. 23-65.

[10] K. Daniel, *Thinking, fast and slow*, 2017.

[11] B. Lepri, N. Oliver, E. Letouze, A. Pentland y P. Vinck, «Fair, transparent, and accountable algorithmic decision-making processes,» *Philosophy & Technology*, vol. 31, n.º 4, págs. 611-627, 2018.

[12] S. Wu, «Advanced control for large scale non-conventional power systems,» 2012.

[13] Baringa, «The future role of network operators: The emerging active DSO model,» 2016. dirección: https://www.baringa.com/getmedia/9174062a-ecc8-4032-9129-04b5573e44f8/The-future-role-of-network-operators-the-emerging-active-DSO-model/ (visitado 04-04-2022).

[14] H. King, «Synthesis and Characterization of Efficient and Economical Surfactants for Use in Enhanced Oil Recovery and Oilfield Corrosion Inhibition Applications,» 2021.

[15] G. de Trabajo de centro de transformación inteligente, «Visión FUTURED hacia 2050,» *FutuRed*, 2020.

[16] L. Fink, «Systems engineering challenges emerge as electric energy network increases in complexity,» *Prof. Eng.(Wash., DC);(United States)*, vol. 47, n.º 12, 1976.

[17] CNMC. «Los contadores inteligentes integrados en el sistema de telegestión en 2018.» (), dirección: https://www.cnmc.es/los-contadores-inteligentes-integrados-en-el-sistema-de-telegestion-alcanzaron-el-98-en-2018-376837 (visitado 04-04-2022).

[18] H. Farhangi, «The path of the smart grid,» *IEEE power and energy magazine*, vol. 8, n.º 1, págs. 18-28, 2009.

[19] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet y D. Irwin, «Private memoirs of a smart meter,» en *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, 2010, págs. 61-66.

[20] C. Beckel, L. Sadamori, T. Staake y S. Santini, «Revealing household characteristics from smart meter data,» *Energy*, vol. 78, págs. 397-410, 2014.

[21] H. Youn y H. J. Jin, «The effects of progressive pricing on household electricity use,» *Journal of Policy Modeling*, vol. 38, n.º 6, págs. 1078-1088, 2016.

[22] G. de Trabajo de centro de transformación inteligente, «Centro de Transformación Inteligente: Fundamentos para distribución de funcionalidades,» *FutuRed*, 2021.

[23] S.-y. Chen, S.-f. Song, L.-x. Li y J. Shen, «Survey on smart grid technology,» *Power system technology*, vol. 33, n.º 8, págs. 1-7, 2009.

[24] R. M. Ciric, A. P. Feltrin y L. F. Ochoa, «Power flow in four-wire distribution networks-general approach,» *IEEE Transactions on Power Systems*, vol. 18, n.º 4, págs. 1283-1290, 2003.

[25] E. M. de Industria y Energía, «Reglamento electrotécnico para baja tensión e ITC,» Publicaciones de la Administración General del Estado, 2021. dirección: https://www.boe.es/legislacion/codigos/abrir_pdf.php?fich=326_Reglamento_electrotecnico_para_baja_tension_e_ITC.pdf (visitado 04-04-2022).

[26] «MT 2.11.03 (Ed.08) - Proyecto tipo para centro de transformación en edificio de otros usos.» (visitado 04-04-2022).

[27] K. Wang, S. Skiena y T. G. Robertazzi, «Phase balancing algorithms,» *Electric Power Systems Research*, vol. 96, págs. 218-224, 2013.

[28] N. Gupta, A. Swarnkar y K. Niazi, «A novel method for simultaneous phase balancing and mitigation of neutral current harmonics in secondary distribution systems,» *International Journal of Electrical Power & Energy Systems*, vol. 55, págs. 645-656, 2014.

[29] L. Fortuna, S. Graziani, A. Rizzo y M. G. Xibilia, *Soft sensors for monitoring and control of industrial processes*. Springer Science & Business Media, 2007.

[30] P. Kadlec, B. Gabrys y S. Strandt, «Data-driven soft sensors in the process industry,» *Computers & chemical engineering*, vol. 33, n.º 4, págs. 795-814, 2009.

[31] J. M. Pinto, M. Joly y L. F. L. Moro, «Planning and scheduling models for refinery operations,» *Computers & Chemical Engineering*, vol. 24, n.º 9-10, págs. 2259-2276, 2000.

[32] M. Joly, D. Odloak, M. Y. Miyake, B. C. Menezes y J. D. Kelly, «Refinery production scheduling toward Industry 4.0,» *Frontiers of Engineering Management*, vol. 5, n.º 2, págs. 202-213, 2018.

[33] H. Albazzaz y X. Z. Wang, «Historical data analysis based on plots of independent and parallel coordinates and statistical control limits,» *Journal of Process Control*, vol. 16, n.º 2, págs. 103-114, 2006.

[34] A. C. Pereira y F. Romero, «A review of the meanings and the implications of the Industry 4.0 concept,» *Procedia Manufacturing*, vol. 13, págs. 1206-1214, 2017.

[35] J. Lee, H.-A. Kao y S. Yang, «Service innovation and smart analytics for industry 4.0 and big data environment,» *Procedia Cirp*, vol. 16, págs. 3-8, 2014.

[36] A. Jindal, M. Gerndt, M. Bauch y H. Haddouti, «Scalable Infrastructure and Workflow for Anomaly Detection in an Automotive Industry,» en *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, IEEE, 2020, págs. 1-6.

[37] IPRI. «Índice de Precios Industriales 2021.» (), dirección: `https://www.ine.es/daco/daco42/daco423/ipri0821.pdf` (visitado 04-04-2022).

[38] M. Pacce, I. Sanchez-Garcia y M. Suarez-Varela, «Recent Developments in Spanish Retail Electricity Prices: The Role Played by the Cost of CO2 Emission Allowances and Higher Gas Prices (El papel del coste de los derechos de emision de CO2 y del encarecimiento del gas en la evolucion reciente de los precios minoristas de la electricidad en España),» 2021.

[39] M. d. l. L. Matea Rosa, F. Martinez Casares y S. Vazquez Martinez, «El coste de la electricidad para las empresas españolas,» *Boletin economico/Banco de España [Articulos], n. 1, 2021*, 2021.

[40] S. Gil-López. «Las 5Ws de la Ciencia De Datos: qué, quién, dónde, cuándo y por qué.» (), dirección: http://blogs.tecnalia.com/inspiring-blog/2021/04/08/las-5ws-la-ciencia-datos-quien-donde-cuando/ (visitado 04-04-2022).

[41] «ISO 50001:2018. Sistemas de gestión de la energía. Requisitos con orientación para su uso,» 19 de dic. de 2018. dirección: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060594 (visitado 28-09-2021).

[42] I. Mendia, S. Gil-Lopez, J. Del Ser y col., «An Intelligent Procedure for the Methodology of Energy Consumption in Industrial Environments,» en *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2020, págs. 92-103.

[43] B. Lepri, J. Staiano, D. Sangokoya, E. Letouze y N. Oliver, «The tyranny of data? the bright and dark sides of data-driven decision-making for social good,» en *Transparent data mining for big and small data*, Springer, 2017, págs. 3-24.

[44] P Czop, G Kost, D Sławik y G Wszołek, «Formulation and identification of first-principle data-driven models,» *Journal of Achievements in materials and manufacturing Engineering*, vol. 44, n.º 2, págs. 179-186, 2011.

[45] A. J. Trappey, C. V. Trappey, U. H. Govindarajan, J. J. Sun y A. C. Chuang, «A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing,» *IEEE Access*, vol. 4, págs. 7356-7382, 2016.

[46] B. Eiteneuer, N. Hranisavljevic y O. Niggemann, «Dimensionality reduction and anomaly detection for cpps data using autoencoder,» en *2019 IEEE International Conference on Industrial Technology (ICIT)*, IEEE, 2019, págs. 1286-1292.

[47] N. Hranisavljevic, O. Niggemann y A. Maier, «A novel anomaly detection algorithm for hybrid production systems based on deep learning and timed automata,» *arXiv preprint arXiv:2010.15415*, 2020.

[48] O. E. Oluyisola, S. Bhalla, F. Sgarbossa y J. O. Strandhagen, «Designing and developing smart production planning and control systems in the industry 4.0 era: a methodology and case study,» *Journal of Intelligent Manufacturing*, págs. 1-22, 2021.

[49] R. R. Lam, L. Horesh, H. Avron y K. E. Willcox, «Should you derive, or let the data drive? an optimization framework for hybrid first-principles data-driven modeling,» *arXiv preprint arXiv:1711.04374*, 2017.

[50] F. T. Chan y H. K. Chan, «A comprehensive survey and future trend of simulation study on FMS scheduling,» *Journal of Intelligent Manufacturing*, vol. 15, n.º 1, págs. 87-102, 2004.

[51] L. Pehrsson, A. H. Ng y D. Stockton, «Industrial cost modelling and multi-objective optimisation for decision support in production systems development,» *Computers & Industrial Engineering*, vol. 66, n.º 4, págs. 1036-1048, 2013.

[52] A. Rasheed, O. San y T. Kvamsdal, «Digital twin: Values, challenges and enablers from a modeling perspective,» *Ieee Access*, vol. 8, págs. 21 980-22 012, 2020.

[53] O. Auchet, P. Riedinger, O. Malasse y C. Iung, «First-principles simplified modelling of glass furnaces combustion chambers,» *Control Engineering Practice*, vol. 16, n.º 12, págs. 1443-1456, 2008.

[54] F. Shrouf, J. Ordieres y G. Miragliotta, «Smart factories in Industry 4.0: A review of the concept and of energy management approached in production based on the Internet of Things paradigm,» en *2014 IEEE international conference on industrial engineering and engineering management*, IEEE, 2014, págs. 697-701.

[55] F. Almada-Lobo, «The Industry 4.0 revolution and the future of Manufacturing Execution Systems (MES),» *Journal of innovation management*, vol. 3, n.º 4, págs. 16-21, 2015.

[56] E. A. Lee, «The past, present and future of cyber-physical systems: A focus on models,» *Sensors*, vol. 15, n.º 3, págs. 4837-4869, 2015.

[57] N. Nazemzadeh, A. A. Malanca, R. F. Nielsen, K. V. Gernaey, M. P. Andersson y S. S. Mansouri, «Integration of first-principle models and machine learning in a modeling framework: An application to flocculation,» *Chemical Engineering Science*, vol. 245, pág. 116 864, 2021.

[58] Gartner. «Gartner Says 70 % of Organizations Will Shift Their Focus From Big to Small and Wide Data By 2025.» (), dirección: https://www.gartner.com/en/newsroom/press-releases/2021-05-19-gartner-says-70-percent-of-organizations-will-shift-their-focus-from-big-to-small-and-wide-data-by-2025 (visitado 04-04-2022).

[59] G. Marcus, «Deep learning: A critical appraisal,» *arXiv preprint arXiv:1801.00631*, 2018.

[60] I. Pan, L. Mason y O. Matar, «Data-Centric Engineering: integrating simulation, machine learning and statistics. Challenges and Opportunities,» *arXiv preprint arXiv:2111.06223*, 2021.

[61] V. Venkatasubramanian, «The promise of artificial intelligence in chemical engineering: Is it here, finally,» *AIChE J*, vol. 65, n.º 2, págs. 466-478, 2019.

[62] D. Jayaratne, D. De Silva, D. Alahakoon y X. Yu, «Continuous detection of concept drift in industrial cyber-physical systems using closed loop incremental machine learning,» *Discover Artificial Intelligence*, vol. 1, n.º 1, págs. 1-13, 2021.

[63] G. Schirner, D. Erdogmus, K. Chowdhury y T. Padir, «The future of human-in-the-loop cyber-physical systems,» *Computer*, vol. 46, n.º 1, págs. 36-45, 2013.

[64] C. Molnar, G. König, J. Herbinger y col., «General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models,» *arXiv preprint arXiv:2007.04131*, 2020.

[65] L. Monostori, «Cyber-physical production systems: Roots, expectations and R&D challenges,» *Procedia Cirp*, vol. 17, págs. 9-13, 2014.

[66] Gartner. «Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage.» (), dirección: https://www.gartner.com/en/newsroom/press-releases/2016-08-16-gartners-2016-hype-cycle-for-emerging-technologies-identifies-three-key-trends-that-organizations-must-track-to-gain-competitive-advantage (visitado 04-04-2022).

[67] Gartner1. «The 4 Trends That Prevail on the Gartner Hype Cycle for AI, 2021.» (), dirección: https://www.gartner.com/en/articles/the-4-trends-that-prevail-on-the-gartner-hype-cycle-for-ai-2021/ (visitado 04-04-2022).

[68] ——, «Why Big Data Science and Data Analytics Projects Fail.» (), dirección: https://www.datascience-pm.com/project-failures/ (visitado 04-04-2022).

[69] VentureBeat. «Why do 87 % of data science projects never make it into production?» (), dirección: https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/ (visitado 04-04-2022).

[70] Gartner. «Our Top Data and Analytics Predicts for 2019.» (), dirección: https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/ (visitado 04-04-2022).

[71] F. Chollet, *Deep learning with Python*. Simon y Schuster, 2021.

[72] IDC. «Forecasts Improved Growth for Global AI Market in 2021.» (), dirección: https://www.idc.com/getdoc.jsp?containerId=prUS47482321 (visitado 04-04-2022).

[73] S. Ransbotham, D. Kiron, P. Gerbert y M. Reeves, «Reshaping business with artificial intelligence: Closing the gap between ambition and action,» *MIT Sloan Management Review*, vol. 59, n.º 1, 2017.

[74] Y. Thieullent Baerd, «The AI-powered enterprise: Unlocking the potential of AI at scale,» *Capgemini Research*, 2020. dirección: https://www.capgemini.com/research/the-ai-powered-enterprise/.

[75] W. Kong, «Investigation of Phase Imbalance Characteristics and Phase Balancing in Low Voltage Distribution Networks,» Tesis doct., University of Bath, 2021.

[76] W. Kong, K. Ma y Q. Wu, «Three-phase power imbalance decomposition into systematic imbalance and random imbalance,» *IEEE Transactions on Power Systems*, vol. 33, n.º 3, págs. 3001-3012, 2017.

[77] S. Beharrysingh, «Phase unbalance on low-voltage electricity networks and its mitigation using static balancers,» Tesis doct., Loughborough University, 2014.

[78] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa y T. Gozel, «Representative residential LV feeders: A case study for the North West of England,» *IEEE Transactions on Power Systems*, vol. 31, n.º 1, págs. 348-360, 2015.

[79] J. N. Fidalgo, C. Moreira y R. Cavalheiro, «Impact of load unbalance on low voltage network losses,» en *2019 IEEE Milan PowerTech*, IEEE, 2019, págs. 1-5.

[80] I. Berganza, A. Sendin y J. Arriola, «PRIME: Powerline intelligent metering evolution,» en *CIRED Seminar 2008: SmartGrids for Distribution*, IET, 2008, págs. 1-3.

[81] Mecfi. «Desequilibrios en el sistema trifásico.» (), dirección: `https://mecfi.es/equilibrio-trifasico` (visitado 04-04-2022).

[82] J. Hansell, H. F. Bernheim, Y. Liao, M. R. Martin y A. L. Abendschein. «System and method for inferring schematic and topological properties of an electrical distribution grid.» US Patent 10,554,257. (2020).

[83] M. Shakeel, S. A. Jaffar, M. F. Ali y S. Zaidi, «LV three phase automatic load balancing system,» en *4th International Conference on Energy, Environment and Sustainable Development*, 2016.

[84] S. Yan, S.-C. Tan, C.-K. Lee, B. Chaudhuri y S. R. Hui, «Electric springs for reducing power imbalance in three-phase power systems,» *IEEE Transactions on Power Electronics*, vol. 30, n.º 7, págs. 3601-3609, 2014.

[85] S. Soltani, M. Rashidinejad y A. Abdollahi, «Stochastic multiobjective distribution systems phase balancing considering distributed energy resources,» *IEEE Systems Journal*, vol. 12, n.º 3, págs. 2866-2877, 2017.

[86] F. R. Quintela, J. G. Arevalo y N. R. Melchor, «Desequilibrio y perdidas en las instalaciones electricas,» *Montajes e instalaciones*, 2000.

[87] M. W. Siti, D. V. Nicolae, A. A. Jimoh y A. Ukil, «Reconfiguration and load balancing in the LV and MV distribution networks for optimal performance,» *IEEE transactions on power delivery*, vol. 22, n.º 4, págs. 2534-2540, 2007.

[88] T.-H. Chen y J.-T. Cherng, «Optimal phase arrangement of distribution transformers connected to a primary feeder for system unbalance improvement and loss reduction using a genetic algorithm,» en *Proceedings of the 21st International Conference on Power Industry Computer Applications. Connecting Utilities. PICA 99. To the Millennium and Beyond (Cat. No. 99CH36351)*, IEEE, 1999, págs. 145-151.

[89] C.-H. Lin, C.-S. Chen, H.-J. Chuang y C.-Y. Ho, «Heuristic rule-based phase balancing of distribution systems by considering customer load patterns,» *IEEE Transactions on Power Systems*, vol. 20, n.º 2, págs. 709-716, 2005.

[90] C. Fei y R. Wang, «Using phase swapping to solve load phase balancing by ADSCHNN in LV distribution network,» *Int. J. Control Autom*, vol. 7, n.º 7, págs. 1-14, 2014.

[91] G. Vulasala, S. Sirigiri y R. Thiruveedula, «Feeder reconfiguration for loss reduction in unbalanced distribution system using genetic algorithm,» *International Journal of Electrical and Electronics Engineering*, vol. 3, n.º 12, págs. 754-762, 2009.

[92] C.-T. Su y C.-S. Lee, «Feeder reconfiguration and capacitor setting for loss reduction of distribution systems,» *Electric power systems research*, vol. 58, n.º 2, págs. 97-102, 2001.

[93] D. S. Hochba, «Approximation algorithms for NP-hard problems,» *ACM Sigact News*, vol. 28, n.º 2, págs. 40-52, 1997.

[94] J. Zhu, G. Bilbro y M.-Y. Chow, «Phase balancing using simulated annealing,» *IEEE Transactions on Power Systems*, vol. 14, n.º 4, págs. 1508-1513, 1999.

[95] C.-H. Lin, C.-S. Chen, H.-J. Chuang, M.-Y. Huang y C.-W. Huang, «An Expert System for Three-Phase Balancing of Distribution Feeders,» *IEEE Transactions on Power Systems*, vol. 23, págs. 1488-1496, 2008.

[96] E Dolatdar, S Soleymani y B Mozafari, «A new distribution network reconfiguration approach using a tree model,» *World Academy of Science, Engineering and Technology*, vol. 58, n.º 34, pág. 1186, 2009.

[97] O. Homaee, A. Najafi, M. Dehghanian, M. Attar y H. Falaghi, «A practical approach for distribution network load balancing by optimal re-phasing of single phase customers using discrete genetic algorithm,» *International Transactions on Electrical Energy Systems*, vol. 29, n.º 5, e2834, 2019.

[98] B. Cortes-Caicedo, L. S. Avellaneda-Gomez, O. D. Montoya, L. Alvarado-Barrios y H. R. Chamorro, «Application of the Vortex Search Algorithm to the Phase-Balancing Problem in Distribution Systems,» *Energies*, vol. 14, n.º 5, pág. 1282, 2021.

[99] A. G. Ruiz, M. G. Echeverri y R. A. Gallego, «Balance de fases usando colonia de hormigas,» *Lenguaje*, vol. 7, n.º 2, págs. 43-52, 2005.

[100] C. Peng, L. Xu, X. Gong, H. Sun y L. Pan, «Molecular evolution based dynamic reconfiguration of distribution networks with DGs considering three-phase balance and switching times,» *IEEE Transactions on Industrial Informatics*, vol. 15, n.º 4, págs. 1866-1876, 2018.

[101] S. Soltani, M Rashidinejad y A Abdollahi, «Dynamic phase balancing in the smart distribution networks,» *International Journal of Electrical Power & Energy Systems*, vol. 93, págs. 374-383, 2017.

[102] C. Bary, «Coincidence-factor reationships of electric-service-load characteristics,» *Transactions of the American Institute of Electrical Engineers*, vol. 64, n.º 9, págs. 623-629, 1945.

[103] A. Grandjean, J. Adnot y G. Binet, «A review and an analysis of the residential electric load curve models,» *Renewable and Sustainable energy reviews*, vol. 16, n.º 9, págs. 6539-6565, 2012.

[104] R. Hamilton, «Synthetic or equivalent load curves,» *Transactions of the American Institute of Electrical Engineers*, vol. 61, n.º 6, págs. 369-381, 1942.

[105] R. Hamilton y A. Grandjean, «The summation of load curves,» *Electrical Engineering*, vol. 63, n.º 10, págs. 729-735, 1944.

[106] S. Katoch, S. S. Chauhan y V. Kumar, «A review on genetic algorithm: past, present, and future,» *Multimedia Tools and Applications*, págs. 1-36, 2020.

[107] V. Chvatal, «A greedy heuristic for the set-covering problem,» *Mathematics of operations research*, vol. 4, n.º 3, págs. 233-235, 1979.

[108] J. Bang-Jensen, G. Gutin y A. Yeo, «When the greedy algorithm fails,» *Discrete optimization*, vol. 1, n.º 2, págs. 121-127, 2004.

[109] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[110] Z. W. Geem, *Music-inspired harmony search algorithm: theory and applications*. Springer, 2009, vol. 191.

[111] X.-S. Yang, «Harmony search as a metaheuristic algorithm,» en *Music-inspired harmony search algorithm*, Springer, 2009, págs. 1-14.

[112] S. Gil-Lopez, J. Del Ser, I. Landa, L. Garcia-Padrones, S. Salcedo-Sanz y J. A. Portilla-Figueras, «On the application of a novel grouping harmony search algorithm to the switch location problem,» en *International Conference on Mobile Lightweight Wireless Systems*, Springer, 2010, págs. 662-672.

[113] J. Del Ser, M. Matinmikko, S. Gil-Lopez y M. Mustonen, «Centralized and distributed spectrum channel assignment in cognitive wireless networks: a harmony search approach,» *Applied Soft Computing*, vol. 12, n.º 2, págs. 921-930, 2012.

[114] A. Friedman, *What Is The Crest Factor And Why Is It Used?* 2008.

[115] L. Xiaoming, S. Weihua, Y. Xianggen, L. Shiqi y Z. Lianmei, «The statistical algorithm of simultaneity coefficient based on real-time data of typical consumers from power measurement system,» en *Proceedings. International Conference on Power System Technology*, IEEE, vol. 4, 2002, págs. 2247-2250.

[116] T Chala. «Perdidas en distribucion de energia electrica.» (2012).

[117] J. A. Suarez, G. F. Di Mauro, D. O. Anaut y C. Aguero, «Parametros que afectan la corriente de neutro en presencia de armonicos,» *Informacion tecnologica*, vol. 21, n.º 1, págs. 77-89, 2010.

[118] T.-H. Chen, «Analysis of multi-grounded four-wire distribution systems considering the neutral grounding,» *IEEE Transactions on Power Delivery*, vol. 16, págs. 710-717, 2001.

[119] P. Kadlec, R. Grbic y B. Gabrys, «Review of adaptation mechanisms for data-driven soft sensors,» *Computers & chemical engineering*, vol. 35, n.º 1, págs. 1-24, 2011.

[120] M. Kano y M. Ogawa, «The state of the art in chemical process control in Japan: Good practice and questionnaire survey,» *Journal of Process Control*, vol. 20, n.º 9, págs. 969-982, 2010.

[121] S. Kim, M. Kano, S. Hasebe, A. Takinami y T. Seki, «Long-term industrial applications of inferential control based on just-in-time soft-sensors: Economical impact and challenges,» *Industrial & Engineering Chemistry Research*, vol. 52, n.º 35, págs. 12 346-12 356, 2013.

[122] S. J. Qin, «Recursive PLS algorithms for adaptive data modeling,» *Computers & Chemical Engineering*, vol. 22, n.º 4-5, págs. 503-514, 1998.

[123] J. Tang, W. Yu, T. Chai y L. Zhao, «On-line principal component analysis with application to process modeling,» *Neurocomputing*, vol. 82, págs. 167-178, 2012.

[124] H. Kaneko y K. Funatsu, «Ensemble locally weighted partial least squares as a just-in-time modeling method,» *AIChE Journal*, vol. 62, n.º 3, págs. 717-725, 2016.

[125] J. Yu, K. Chen, J. Mori y M. M. Rashid, «A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction,» *Energy*, vol. 61, págs. 673-686, 2013.

[126] L. Yao y Z. Ge, «Moving window adaptive soft sensor for state shifting process based on weighted supervised latent factor analysis,» *Control Engineering Practice*, vol. 61, págs. 72-80, 2017.

[127] W. Shao, S. Chen y C. J. Harris, «Adaptive soft sensor development for multi-output industrial processes based on selective ensemble learning,» *IEEE Access*, vol. 6, págs. 55 628-55 642, 2018.

[128] S. Suradhaniwar, S. Kar, S. S. Durbha y A. Jagarlapudi, «Time Series Forecasting of Univariate Agrometeorological Data: A Comparative Performance Evaluation via One-Step and Multi-Step Ahead Forecasting Strategies,» *Sensors*, vol. 21, n.º 7, pág. 2430, 2021.

[129] C. A. Ratanamahatana, J. Lin, D. Gunopulos, E. Keogh, M. Vlachos y G. Das, «Mining Time Series Data,» en *Data Mining and Knowledge Discovery Handbook*, O. Maimon y L. Rokach, eds. Boston, MA: Springer US, 2010, págs. 1049-1077, ISBN: 978-0-387-09823-4. DOI: 10.1007/978-0-387-09823-4_56. dirección: https://doi.org/10.1007/978-0-387-09823-4_56.

[130] Y. Du, Y. Liang, J. Jiang, R. J. Berry y Y Ozaki, «Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares,» *Analytica chimica acta*, vol. 501, n.º 2, págs. 183-191, 2004.

[131] H. Kaneko y K. Funatsu, «Smoothing-combined soft sensors for noise reduction and improvement of predictive ability,» *Industrial & Engineering Chemistry Research*, vol. 54, n.º 50, págs. 12 630-12 638, 2015.

[132] C. Kneale y S. D. Brown, «Small moving window calibration models for soft sensing processes with limited history,» *Chemometrics and Intelligent Laboratory Systems*, vol. 183, págs. 36-46, 2018.

[133] W. Ni, S. K. Tan, W. J. Ng y S. D. Brown, «Moving-window GPR for nonlinear dynamic system modeling with dual updating and dual preprocessing,» *Industrial & engineering chemistry research*, vol. 51, n.º 18, págs. 6416-6428, 2012.

[134] D. Markudova, E. Baralis, L. Cagliero y col., «Heterogeneous Industrial Vehicle Usage Predictions: A Real Case.,» en *EDBT/ICDT Workshops*, 2019, págs. 3-8.

[135] Z. Ge y Z. Song, «A comparative study of just-in-time-learning based methods for online soft sensor modeling,» *Chemometrics and Intelligent Laboratory Systems*, vol. 104, n.º 2, págs. 306-317, 2010.

[136] X. Yuan, Z. Ge, B. Huang, Z. Song e Y. Wang, «Semisupervised JITL framework for nonlinear industrial soft sensing based on locally semisupervised weighted PCR,» *IEEE Transactions on Industrial Informatics*, vol. 13, n.º 2, págs. 532-541, 2016.

[137] X. Yuan, B. Huang, Z. Ge y Z. Song, «Double locally weighted principal component regression for soft sensor with sample selection under supervised latent structure,» *Chemometrics and Intelligent Laboratory Systems*, vol. 153, págs. 116-125, 2016.

[138] Y. Liu y J. Chen, «Integrated soft sensor using just-in-time support vector regression and probabilistic analysis for quality prediction of multi-grade processes,» *Journal of Process control*, vol. 23, n.º 6, págs. 793-804, 2013.

[139] A. Urhan y B. Alakent, «Integrating adaptive moving window and just-in-time learning paradigms for soft-sensor design,» *Neurocomputing*, vol. 392, págs. 23-37, 2020.

[140] K. Hazama y M. Kano, «Covariance-based locally weighted partial least squares for high-performance adaptive modeling,» *Chemometrics and Intelligent Laboratory Systems*, vol. 146, págs. 55-62, 2015.

[141] S. Kim, R. Okajima, M. Kano y S. Hasebe, «Development of soft-sensor using locally weighted PLS with adaptive similarity measure,» *Chemometrics and Intelligent Laboratory Systems*, vol. 124, págs. 43-49, 2013.

[142] K. Fujiwara, M. Kano, S. Hasebe y A. Takinami, «Soft-sensor development using correlation-based just-in-time modeling,» *AIChE Journal*, vol. 55, n.º 7, págs. 1754-1765, 2009.

[143] Y. Liu, C. Yang, K. Liu, B. Chen e Y. Yao, «Domain adaptation transfer learning soft sensor for product quality prediction,» *Chemometrics and Intelligent Laboratory Systems*, vol. 192, pág. 103 813, 2019.

[144] T. G. Dietterich, «Ensemble methods in machine learning,» en *International workshop on multiple classifier systems*, Springer, 2000, págs. 1-15.

[145] D. Opitz y R. Maclin, «Popular ensemble methods: An empirical study,» *Journal of artificial intelligence research*, vol. 11, págs. 69-98, 1999.

[146] G. Brown, «Diversity in neural network ensembles,» Tesis doct., Citeseer, 2004.

[147] L. Breiman, «Bagging predictors,» *Machine learning*, vol. 24, n.º 2, págs. 123-140, 1996.

[148] Y. Freund, R. Schapire y N. Abe, «A short introduction to boosting,» *Journal-Japanese Society For Artificial Intelligence*, vol. 14, n.º 771-780, pág. 1612, 1999.

[149] T. Chen y C. Guestrin, «Xgboost: A scalable tree boosting system,» en *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, págs. 785-794.

[150] J. Cahill. «Applying Advanced Process Control to Distillation Processes.» (2009), dirección: https://www.emersonautomationexperts.com/2009/services-consulting-training/applying_advanc_1/.

[151] J. Riggs. «Distillation Column Control.» (2015), dirección: https://controlguru.com/distillation-introduction-to-control/.

[152] Y.-D. Hsiao, J.-L. Kang y D. S.-H. Wong, «Development of Robust and Physically Interpretable Soft Sensor for Industrial Distillation Column Using Transfer Learning with Small Datasets,» *Processes*, vol. 9, n.º 4, pág. 667, 2021.

[153] Z. Ma, S. Chen, Y. Yin, Y. Zhou, X. Lu y T. Lin, «Study on measurement model of lignin content in pulp after alkaline extraction,» *Journal of Environmental Engineering and Landscape Management*, vol. 29, n.º 2, págs. 94-100, 2021.

[154] X. Li, X. Yi, Z. Liu y col., «Application of novel hybrid deep leaning model for cleaner production in a paper industrial wastewater treatment system,» *Journal of Cleaner Production*, vol. 294, pág. 126 343, 2021.

[155] H. Zhang, C. Yang, X. Shi y H. Liu, «Effluent quality prediction in papermaking wastewater treatment processes using dynamic Bayesian networks,» *Journal of Cleaner Production*, vol. 282, pág. 125 396, 2021.

[156] D. C. de Souza, L. Cabrita, C. F. Galinha, T. J. Rato y M. S. Reis, «A Spectral AutoML Approach for Industrial Soft Sensor Development: Validation in an Oil Refinery Plant,» *Computers & Chemical Engineering*, pág. 107 324, 2021.

[157] I. Niño-Adan, I. Landa-Torres, D. Manjarres y E. Portillo, «Soft-sensor design for vacuum distillation bottom product penetration classification,» *Applied Soft Computing*, vol. 102, pág. 107 072, 2021.

[158] J. Foschi, A. Turolla y M. Antonelli, «Soft sensor predictor of E. coli concentration based on conventional monitoring parameters for wastewater disinfection control,» *Water Research*, vol. 191, pág. 116 806, 2021.

[159] B. S. Pattnaik, A. S. Pattanayak, S. K. Udgata y A. K. Panda, «Machine learning based soft sensor model for BOD estimation using intelligence at edge,» *Complex & Intelligent Systems*, vol. 7, n.º 2, págs. 961-976, 2021.

[160] W. Wang, C. Yang, J. Han, W. Li e Y. Li, «A soft sensor modeling method with dynamic time-delay estimation and its application in wastewater treatment plant,» *Biochemical Engineering Journal*, pág. 108 048, 2021.

[161] C. Mei, Y. Ding, X. Chen, Y. Chen y J. Cai, «Soft Sensor Modelling based on Just-in-Time Learning and Bagging-PLS for Fermentation Processes,» *Chemical Engineering Transactions*, vol. 70, págs. 1435-1440, 2018.

[162] Y. Liu, T. Chen y J. Chen, «Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions,» *Industrial & Engineering Chemistry Research*, vol. 54, n.º 18, págs. 5037-5047, 2015.

[163] H. Huang, X. Peng, C. Jiang, Z. Li y W. Zhong, «Variable-Scale Probabilistic Just-in-Time Learning for Soft Sensor Development with Missing Data,» *Industrial & Engineering Chemistry Research*, vol. 59, n.º 11, págs. 5010-5021, 2020.

[164] J. Wang, K. Qiu, Y. Guo, R. Wang y X. Zhou, «Soft sensor development based on improved just-in-time learning and relevant vector machine for batch processes,» *The Canadian Journal of Chemical Engineering*, vol. 99, n.º 1, págs. 334-344, 2021.

[165] B. Alakent, «Soft sensor design using transductive moving window learner,» *Computers & Chemical Engineering*, vol. 140, pág. 106 941, 2020.

[166] D. G. Stork, R. O. Duda, P. E. Hart y D Stork, «Pattern classification,» *A Wiley-Interscience Publication*, 2001.

[167] X. Yuan, J. Zhou, Y. Wang y C. Yang, «Multi-similarity measurement driven ensemble just-in-time learning for soft sensing of industrial processes,» *Journal of Chemometrics*, vol. 32, n.º 9, e3040, 2018.

[168] F. Yu, L. Cao, W. Li, F. Yang y C. Shang, «Feature based causality analysis and its applications in soft sensor modeling,» *IFAC-PapersOnLine*, vol. 53, n.º 2, págs. 138-143, 2020.

[169] L. Wiskott y T. J. Sejnowski, «Slow feature analysis: Unsupervised learning of invariances,» *Neural computation*, vol. 14, n.º 4, págs. 715-770, 2002.

[170] S. Sharma y S. S. Tambe, «Soft-sensor development for biochemical systems using genetic programming,» *Biochemical engineering journal*, vol. 85, págs. 89-100, 2014.

[171] X. Yuan, L. Ye, L. Bao, Z. Ge y Z. Song, «Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA,» *Chemometrics and Intelligent Laboratory Systems*, vol. 147, págs. 167-175, 2015.

[172] H. Li, T. Chai y H. Yue, «Soft sensor of technical indices based on KPCA-ELM and application for flotation process,» *Ciesc Journal*, vol. 63, n.º 9, págs. 2892-2898, 2012.

[173] D. Li, Z. Li y K. Sun, «Development of a novel soft sensor with long short-term memory network and normalized mutual information feature selection,» *Mathematical Problems in Engineering*, vol. 2020, 2020.

[174] F. Guo, B. Wei y B. Huang, «A just-in-time modeling approach for multimode soft sensor based on Gaussian mixture variational autoencoder,» *Computers & Chemical Engineering*, vol. 146, pág. 107 230, 2021.

[175] P. Domingos, «A few useful things to know about machine learning,» *Communications of the ACM*, vol. 55, n.º 10, págs. 78-87, 2012.

[176] D. H. Wolpert y W. G. Macready, «No free lunch theorems for optimization,» *IEEE transactions on evolutionary computation*, vol. 1, n.º 1, págs. 67-82, 1997.

[177] X. Zhu, K. U. Rehman, B. Wang y M. Shahzad, «Modern soft-sensing modeling methods for fermentation processes,» *Sensors*, vol. 20, n.º 6, pág. 1771, 2020.

[178] F. A. Souza, R. Araujo y J. Mendes, «Review of soft sensor methods for regression applications,» *Chemometrics and Intelligent Laboratory Systems*, vol. 152, págs. 69-79, 2016.

[179] B. Lin, B. Recke, J. K. Knudsen y S. B. Jørgensen, «A systematic approach for soft sensor development,» *Computers & chemical engineering*, vol. 31, n.º 5-6, págs. 419-425, 2007.

[180] D. Wang, J. Liu y R. Srinivasan, «Data-driven soft sensor approach for quality prediction in a refining process,» *IEEE Transactions on Industrial Informatics*, vol. 6, n.º 1, págs. 11-17, 2009.

[181] I. F. Ilyas y X. Chu, *Data cleaning*. ACM, 2019.

[182] K. Morad, B. R. Young y W. Y. Svrcek, «Rectification of plant measurements using a statistical framework,» *Computers & chemical engineering*, vol. 29, n.º 5, págs. 919-940, 2005.

[183] S. Xu, B. Lu, M. Baldea y col., «Data cleaning in the process industries,» *Reviews in Chemical Engineering*, vol. 31, n.º 5, págs. 453-490, 2015.

[184] R. K. Pearson, «Outliers in process modeling and identification,» *IEEE Transactions on control systems technology*, vol. 10, n.º 1, págs. 55-63, 2002.

[185] H. Martens y T. Naes, *Multivariate calibration*. John Wiley & Sons, 1992.

[186] T Jayalakshmi y A Santhakumaran, «Statistical normalization and back propagation for classification,» *International Journal of Computer Theory and Engineering*, vol. 3, n.º 1, págs. 1793-8201, 2011.

[187] A. Savitzky y M. J. Golay, «Smoothing and differentiation of data by simplified least squares procedures.,» *Analytical chemistry*, vol. 36, n.º 8, págs. 1627-1639, 1964.

[188] M. Budka, M. Eastwood, B. Gabrys y col., «From sensor readings to predictions: On the process of developing practical soft sensors,» en *International Symposium on Intelligent Data Analysis*, Springer, 2014, págs. 49-60.

[189] R. Bellman, «Dynamic programming,» *Science*, vol. 153, n.º 3731, págs. 34-37, 1966.

[190] R. E. Bellman, *Adaptive control processes: a guided tour*. Princeton university press, 2015.

[191] D. Mladenić, «Feature selection for dimensionality reduction,» en *International Statistical and Optimization Perspectives Workshop Subspace, Latent Structure and Feature Selection*, Springer, 2005, págs. 84-102.

[192] W. Zhong y J. Yu, «MIMO soft sensors for estimating product quality with on-line correction,» *Chemical Engineering Research and Design*, vol. 78, n.º 4, págs. 612-620, 2000.

[193] A. Altmann, L. Tolosi, O. Sander y T. Lengauer, «Permutation importance: a corrected feature importance measure,» *Bioinformatics*, vol. 26, n.º 10, págs. 1340-1347, 2010.

[194] A. Fisher, C. Rudin y F. Dominici, «All models are wrong but many are useful: variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance,» *arXiv preprint arXiv:1801.01489*, págs. 237-246, 2018.

[195] F. Curreri, S. Graziani y M. G. Xibilia, «Input selection methods for data-driven Soft sensors design: Application to an industrial process,» *Information Sciences*, vol. 537, págs. 1-17, 2020.

[196] I. Žliobaitė, «Learning under concept drift: an overview,» *arXiv preprint arXiv:1010.4784*, 2010.

[197] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy y A. Bouchachia, «A survey on concept drift adaptation,» *ACM computing surveys (CSUR)*, vol. 46, n.º 4, págs. 1-37, 2014.

[198] B. Schifferer, C. Deotte y E. Oldridge, «Tutorial: Feature Engineering for Recommender Systems,» en *Fourteenth ACM Conference on Recommender Systems*, 2020, págs. 754-755.

[199] J. Pan, V. Pham, M. Dorairaj, H. Chen y J.-Y. Lee, «Adversarial Validation Approach to Concept Drift Problem in User Targeting Automation Systems at Uber,» *arXiv e-prints*, arXiv-2004, 2020.

[200] H. Hotelling, «Analysis of a complex of statistical variables into principal components.,» *Journal of educational psychology*, vol. 24, n.º 6, pág. 417, 1933.

[201] I. T. Jolliffe, «Principal components in regression analysis,» en *Principal component analysis*, Springer, 1986, págs. 129-155.

[202] D. Engel, L. Huttenberger y B. Hamann, «A survey of dimension reduction methods for high-dimensional data analysis and visualization,» en *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[203] J. H. Cheong. «Four ways to quantify synchrony between time series data.» (4 de abr. de 2022), dirección: https://towardsdatascience.com/four-ways-to-quantify-synchrony-between-time-series-data-b99136c4a9c9.

[204] T. Helleseth, «Some results about the cross-correlation function between two maximal linear sequences,» *Discrete Mathematics*, vol. 16, n.º 3, págs. 209-232, 1976.

[205] G. Ditzler, M. Roveri, C. Alippi y R. Polikar, «Learning in nonstationary environments: A survey,» *IEEE Computational Intelligence Magazine*, vol. 10, n.º 4, págs. 12-25, 2015.

[206] L. I. Kuncheva e I. Zliobaite, «On the window size for classification in changing environments,» *Intelligent Data Analysis*, vol. 13, n.º 6, págs. 861-872, 2009.

[207] A. J. Smola y B. Schölkopf, «A tutorial on support vector regression,» *Statistics and computing*, vol. 14, n.º 3, págs. 199-222, 2004.

[208] «Real Decreto 61/2006, de 31 de enero, por el que se determinan las especificaciones de gasolinas, gasóleos, fuelóleos y gases licuados del petróleo y se regula el uso de determinados biocarburantes,» 17 de feb. de 2006. dirección: https://https://www.boe.es/buscar/pdf/2006/BOE-A-2006-2779-consolidado.pdf (visitado 28-09-2021).

[209] «Combustibles para automoción. Combustibles para motor diésel (gasóleo). Requisitos y métodos de ensayo,» 4 de oct. de 2017. dirección: https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma/?Tipo=N&c=N0058993 (visitado 28-09-2021).

[210] H.-J. Liaw, Y.-H. Lee, C.-L. Tang, H.-H. Hsu y J.-H. Liu, «A mathematical model for predicting the flash point of binary solutions,» *Journal of Loss Prevention in the Process Industries*, vol. 15, n.º 6, págs. 429-438, 2002.

[211] R. W. Prugh, «Estimation of flash point temperature,» *Journal of Chemical Education*, vol. 50, n.º 2, A85, 1973.

[212] A Alibakhshi, H Mirshahvalad y S Alibakhshi, «Prediction of flash points of pure organic compounds: Evaluation of the DIPPR database,» *Process Safety and Environmental Protection*, vol. 105, págs. 127-133, 2017.

[213] A. International, *Standard Test Methods for Flash Point by Pensky-Martens Closed Cup Tester*. ASTM International, 2015.

[214] F. Gharagheizi, «A new molecular-based model for prediction of enthalpy of sublimation of pure components,» *Thermochimica acta*, vol. 469, n.º 1-2, págs. 8-11, 2008.

[215] L. Y. Phoon, A. A. Mustaffa, H. Hashim y R. Mat, «A review of flash point prediction models for flammable liquid mixtures,» *Industrial & Engineering Chemistry Research*, vol. 53, págs. 12 553-12 565, 2014.

[216] S. M. Santos, D. C. Nascimento, M. C. Costa, A. M. Neto y L. V. Fregolente, «Flash point prediction: Reviewing empirical models for hydrocarbons, petroleum fraction, biodiesel, and blends,» *Fuel*, vol. 263, pág. 116 375, 2020.

[217] T. N. G. Borhani, M. Saniedanesh, M. Bagheri y J. S. Lim, «QSPR prediction of the hydroxyl radical rate constant of water contaminants,» *Water research*, vol. 98, págs. 344-353, 2016.

[218] Z. Jiao, H. U. Escobar-Hernandez, T. Parker y Q. Wang, «Review of recent developments of quantitative structure-property relationship models on fire and explosion-related properties,» *Process Safety and Environmental Protection*, vol. 129, págs. 280-290, 2019.

[219] Z. Jiao, P. Hu, H. Xu y Q. Wang, «Machine learning and deep learning in chemical health and safety: A systematic review of techniques and applications,» *ACS Chemical Health & Safety*, vol. 27, n.º 6, págs. 316-334, 2020.

[220] Q. Sun, L. Jiang, M. Li y J. Sun, «Assessment on thermal hazards of reactive chemicals in industry: state of the art and perspectives,» *Progress in Energy and Combustion Science*, vol. 78, pág. 100 832, 2020.

[221] W. F. McClure, «Near-infrared spectroscopy the giant is running strong,» *Analytical chemistry*, vol. 66, n.º 1, 42A-53A, 1994.

[222] S. Hu y col., «A general framework for incorporating molecular modelling into overall refinery optimisation,» *Applied Thermal Engineering*, vol. 21, n.º 13-14, págs. 1331-1348, 2001.

[223] L Alves, S. M. Paixão, R Pacheco, A. F. Ferreira y C. M. Silva, «Biodesulphurization of fossil fuels: energy, emissions and cost analysis,» *RSC Advances*, vol. 5, n.º 43, págs. 34 047-34 057, 2015.

[224] R. J. Hyndman y G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.

[225] I. Mendia, S. Gil-López, I. Landa-Torres, L. Orbe y E. Maqueda, «Machine learning based adaptive soft sensor for flash point inference in a refinery realtime process,» *Results in Engineering*, pág. 100 362, 2022.

[226] U. R. Chaudhuri, *Fundamentals of petroleum and petrochemical engineering*. Crc Press Boca Raton, 2011.

[227] D. L. Trimm, «Catalysts for the control of coking during steam reforming,» *Catalysis today*, vol. 49, n.º 1-3, págs. 3-10, 1999.

[228] I. Amghizar, L. A. Vandewalle, K. M. Van Geem y G. B. Marin, «New trends in olefin production,» *Engineering*, vol. 3, n.º 2, págs. 171-178, 2017.

[229] A. R. Moss, J.-P. Jouany y J. Newbold, «Methane production by ruminants: its contribution to global warming,» en *Annales de zootechnie*, EDP Sciences, vol. 49, 2000, págs. 231-253.

[230] M. Maione, D. Fowler, P. S. Monks y col., «Air quality and climate change: Designing new win-win policies for Europe,» *Environmental Science & Policy*, vol. 65, págs. 48-57, 2016.

[231] J. Fenger, «Urban air quality,» *Atmospheric environment*, vol. 33, n.º 29, págs. 4877-4900, 1999.

[232] D. Balram, K.-Y. Lian y N. Sebastian, «Air quality warning system based on a localized PM2. 5 soft sensor using a novel approach of Bayesian regularized neural network via forward feature selection,» *Ecotoxicology and environmental safety*, vol. 182, pág. 109 386, 2019.

[233] H. Liu y C. Yoo, «A robust localized soft sensor for particulate matter modeling in Seoul metro systems,» *Journal of hazardous materials*, vol. 305, págs. 209-218, 2016.

[234] E. E. Directive, «Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32,» *Official Journal, L*, vol. 315, págs. 1-56, 2012.

[235] W. Wolf, «Cyber-physical systems,» *Computer*, vol. 42, n.º 03, págs. 88-89, 2009.

[236] K. Gillingham, R. G. Newell y K. Palmer, «Energy efficiency economics and policy,» *Annual Review of Resource Economics*, vol. 1, n.º 1, págs. 597-620, 2009.

[237] A. Harry y E. Beddingfield, «Transforming the market through energy management information systems,» 2016.

[238] K. Schwab, *La cuarta revolucion industrial*. Debate, 2016.

[239] M. Sangorski y A. Wierzbic, «Using computer software for energy saving determination in complex business processes–a case study of KGHM Polska Miedź SA,» *Informatyka Ekonomiczna*, n.º 4 (58), págs. 127-141, 2020.

[240] F. Milojkovic, F. Zuniga, A. Zandi, K. Posern, E. Uen y col., «The Quantification and Reporting of Negawatt-Hours with Flexible Energy Conservation Measure Verification Software (ECM-Tool),» *Open Journal of Energy Efficiency*, vol. 8, n.º 04, pág. 179, 2019.

[241] I. Rakhmonov y N. Kurbonov, «Analysis of automated software for monitoring energy consumption and efficiency of industrial enterprises,» en *E3S Web of Conferences*, EDP Sciences, vol. 216, 2020, pág. 01 178.

[242] V. Hodge y J. Austin, «A survey of outlier detection methodologies,» *Artificial intelligence review*, vol. 22, n.º 2, págs. 85-126, 2004.

[243] R. Chalapathy y S. Chawla, «Deep learning for anomaly detection: A survey,» *arXiv preprint arXiv:1901.03407*, 2019.

[244] S. Agrawal y J. Agrawal, «Survey on anomaly detection using data mining techniques,» *Procedia Computer Science*, vol. 60, págs. 8-13, 2015.

[245] A. Becue, I. Praça y J. Gama, «Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities,» *Artificial Intelligence Review*, vol. 54, n.º 5, págs. 3849-3886, 2021.

[246] V. Chandola, A. Banerjee y V. Kumar, «Anomaly detection: A survey,» *ACM computing surveys (CSUR)*, vol. 41, n.º 3, págs. 1-58, 2009.

[247] A. F. Bueno, M. Godinho Filho y A. G. Frank, «Smart production planning and control in the Industry 4.0 context: A systematic literature review,» *Computers & Industrial Engineering*, pág. 106 774, 2020.

[248] C. M. Salgado, C. Azevedo, H. Proenca y S. M. Vieira, «Noise versus outliers,» *Secondary Analysis of Electronic Health Records*, págs. 163-183, 2016.

[249] H. Rashid, P. Singh, V. Stankovic y L. Stankovic, «Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour?» *Applied energy*, vol. 238, págs. 796-805, 2019.

[250] H. Rashid, V. Stankovic, L. Stankovic y P. Singh, «Evaluation of non-intrusive load monitoring algorithms for appliance-level anomaly detection,» en *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, págs. 8325-8329.

[251] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali y A. Amira, «Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives,» *Applied Energy*, vol. 287, pág. 116 601, 2021.

[252] P. Calvo-Bascones, M. A. Sanz-Bobi y T. M. Welte, «Anomaly detection method based on the deep knowledge behind behavior patterns in industrial components. Application to a hydropower plant,» *Computers in Industry*, vol. 125, pág. 103 376, 2021.

[253] B. Eiteneuer y O. Niggemann, «Lstm for model-based anomaly detection in cyber-physical systems,» *arXiv preprint arXiv:2010.15680*, 2020.

[254] C. Feng y P. Tian, «Time Series Anomaly Detection for Cyber-physical Systems via Neural System Identification and Bayesian Filtering,» *arXiv preprint arXiv:2106.07992*, 2021.

[255] B. Zhao, D. Qin, D. Gao y L. Xu, «Energy saving diagnosis model of petrochemical plant based on intelligent curvelet support vector machine,» *Soft Computing*, págs. 1-11, 2021.

[256] M. Carletti, C. Masiero, A. Beghi y G. A. Susto, «Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis,» en *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, IEEE, 2019, págs. 21-26.

[257] K. Amarasinghe, K. Kenney y M. Manic, «Toward explainable deep neural network based anomaly detection,» en *2018 11th International Conference on Human System Interaction (HSI)*, IEEE, 2018, págs. 311-317.

[258] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova y C. Zhong, «Interpretable machine learning: Fundamental principles and 10 grand challenges,» *arXiv preprint arXiv:2103.11251*, 2021.

[259] C. Rudin, «Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,» *Nature Machine Intelligence*, vol. 1, n.º 5, págs. 206-215, 2019.

[260] Z. C. Lipton, «The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery.,» *Queue*, vol. 16, n.º 3, págs. 31-57, 2018.

[261] Y. Himeur, A. Alsalemi, F. Bensaali y A. Amira, «A novel approach for detecting anomalous energy consumption based on micro-moments and deep neural networks,» *Cognitive Computation*, vol. 12, n.º 6, págs. 1381-1401, 2020.

[262] G. Schuh, T. Gartzen, T. Rodenhauser y A. Marks, «Promoting work-based learning through industry 4.0,» *Procedia Cirp*, vol. 32, págs. 82-87, 2015.

[263] O. Penas, R. Plateaux, S. Patalano y M. Hammadi, «Multi-scale approach from mechatronic to Cyber-Physical Systems for the design of manufacturing systems,» *Computers in Industry*, vol. 86, págs. 52-69, 2017.

[264] L. D. Xu, E. L. Xu y L. Li, «Industry 4.0: state of the art and future trends,» *International Journal of Production Research*, vol. 56, n.º 8, págs. 2941-2962, 2018.

[265] A. Napoleone, M. Macchi y A. Pozzetti, «A review on the characteristics of cyber-physical systems for the future smart factories,» *Journal of manufacturing systems*, vol. 54, págs. 305-335, 2020.

[266] O. Niggemann y V. Lohweg, «On the diagnosis of cyber-physical production systems,» en *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[267] V. Satopaa, J. Albrecht, D. Irwin y B. Raghavan, «Finding a kneedle in a haystack: Detecting knee points in system behavior,» en *2011 31st international conference on distributed computing systems workshops*, 2011, págs. 166-171.

[268] G. Hamerly y C. Elkan, «Learning the k in k-means,» *Advances in neural information processing systems*, vol. 16, págs. 281-288, 2004.

[269] X. Wan, W. Wang, J. Liu y T. Tong, «Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range,» *BMC medical research methodology*, vol. 14, n.º 1, págs. 1-13, 2014.

[270] T. Energy. «Accounting: Facility Energy Use.» (), dirección: `https://what-when-how.com/energy-engineering/accounting-facility-energy-use/` (visitado 04-04-2022).

[271] M. M. Breunig, H.-P. Kriegel, R. T. Ng y J. Sander, «LOF: identifying density-based local outliers,» en *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, págs. 93-104.

[272] J. Tang, Z. Chen, A. W.-C. Fu y D. W. Cheung, «Enhancing effectiveness of outlier detections for low density patterns,» en *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2002, págs. 535-548.

[273] M. Goldstein y A. Dengel, «Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm,» *KI-2012: Poster and Demo Track*, págs. 59-63, 2012.

[274] H.-P. Kriegel, P. Kroger, E. Schubert y A. Zimek, «Outlier detection in axis-parallel subspaces of high dimensional data,» en *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2009, págs. 831-838.

[275] S. Papadimitriou, H. Kitagawa, P. B. Gibbons y C. Faloutsos, «LOCI: Fast outlier detection using the local correlation integral,» en *IEEE International Conference on Data Engineering*, IEEE, 2003, págs. 315-326.

[276] S. Ramaswamy, R. Rastogi y K. Shim, «Efficient algorithms for mining outliers from large data sets,» en *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, págs. 427-438.

[277] Y. Almardeny, N. Boujnah y F. Cleary, «A Novel Outlier Detection Method for Multivariate Data,» *IEEE Transactions on Knowledge and Data Engineering, in press*, 2020.

[278] F. T. Liu, K. M. Ting y Z.-H. Zhou, «Isolation forest,» en *IEEE International Conference on Data Mining*, IEEE, 2008, págs. 413-422.

[279] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola y R. C. Williamson, «Estimating the support of a high-dimensional distribution,» *Neural computation*, vol. 13, n.º 7, págs. 1443-1471, 2001.

[280] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn y L. Chang, «A novel anomaly detection scheme based on principal component classifier,» en *IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the IEEE International Conference on Data Mining*, IEEE, 2003, 172–-179.

[281] P. J. Rousseeuw y K. V. Driessen, «A fast algorithm for the minimum covariance determinant estimator,» *Technometrics*, vol. 41, n.º 3, págs. 212-223, 1999.

[282] A. Arning, R. Agrawal y P. Raghavan, «A linear method for deviation detection in large databases,» en *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 1141, 1996, págs. 164-169.

[283] H.-P. Kriegel, M. Schubert y A. Zimek, «Angle-based outlier detection in high-dimensional data,» en *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008, págs. 444-452.

[284] J. Janssens, F. Huszar, E. Postma y H. van den Herik, «Stochastic outlier selection,» *Tilburg centre for Creative Computing, technical report 2012-001*, 2012.

[285] C. C. Aggarwal, «Outlier analysis,» en *Data mining*, Springer, 2015, págs. 237-263.

[286] D. P. Kingma y M. Welling, «Auto-encoding variational bayes,» *arXiv preprint arXiv:1312.6114*, 2013.

[287] I. Goodfellow, J. Pouget-Abadie, M. Mirza y col., «Generative adversarial nets,» *Advances in Neural Information Processing Systems*, vol. 27, págs. 2672-2680, 2014.

[288] G. J. Szekely y M. L. Rizzo, «The uncertainty principle of game theory,» *The American Mathematical Monthly*, vol. 114, n.º 8, págs. 688-702, 2007.

[289] Y. Liu, Z. Li, C. Zhou y col., «Generative adversarial active learning for unsupervised outlier detection,» *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, n.º 8, págs. 1517-1528, 2019.

[290] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann y C. Sutton, «Veegan: Reducing mode collapse in gans using implicit variational learning,» en *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, págs. 3310-3320.

[291] M. A. Bashar y R. Nayak, «TAnoGAN: Time Series Anomaly Detection with Generative Adversarial Networks,» en *IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2020, págs. 1778-1785.

[292] S. Hochreiter, «The vanishing gradient problem during learning recurrent neural nets and problem solutions,» *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, n.º 02, págs. 107-116, 1998.

[293] Y. Zhao, Z. Nasrullah y Z. Li, «PyOD: A Python Toolbox for Scalable Outlier Detection,» *Journal of Machine Learning Research*, vol. 20, págs. 1-7, 2019.

[294] M. Gaur, S. Makonin, I. V. Bajić y A. Majumdar, «Performance evaluation of techniques for identifying abnormal energy consumption in buildings,» *IEEE Access*, vol. 7, págs. 62 721-62 733, 2019.

[295] «Un paso más en la transformación digital de Petronor: la experiencia del Soft Sensor Flash,» Innobasque. (), dirección: `https://mapa.innobasque.eus/casos-practicos/petroleos-del-norte-sa-petronor_un-paso-mas-en-la-transformacion-digital-de-petronor-la-experiencia-del-soft-sensor-flash` (visitado 04-04-2022).