eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Ingeniería de Computadores

Trabajo de Fin de Grado

Predicción de eventos Pump and Dump en datos de criptomonedas

Autor

Daniel Jiménez Cervera

2019

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Grado en Ingeniería Informática

Ingeniería de Computadores

Trabajo de Fin de Grado

Predicción de eventos Pump and Dump en datos de criptomonedas

Autor

Daniel Jiménez Cervera

Director

Borja Calvo Molinos

Resumen

En los últimos años, el mercado de las criptomonedas ha sufrido un crecimiento significativo. Se estima que actualmente, más de 300.000 millones de dolares se encuentran en formato digital. La falta de regulación en este tipo de mercados lo convierten en un objetivo perfecto para todo tipo de tácticas especulativas. Los avances en las aplicaciones de mensajería instantánea han permitido que enormes grupos de desconocidos se coordinen para especular utlizando un método conocido como Pump and Dump. Mediante este proyecto se pretende estudiar la viabilidad de crear un sistema de detección de eventos Pump and Dump en el mercado de las criptodivisas a través de técnicas de clasificación supervisada en series temporales. Asimismo, tras una descripción de los resultados, se proponen varias vías de mejora y líneas de investigación futuras con objetivos a largo plazo más ambiciosos.

Agradecimientos

Son muchas las personas que me han ayudado estos años de carrera que ya llegan a su final y es por ello que quiero dedicarles este apartado.

En primer lugar me gustaría dar las gracias a mis padres Piluka y Jose por estos cuatro años de apoyo incondicional y por haberme brindado la oportunidad de formarme hoy como Ingeniero Informático, no estaría aquí de no ser por vosotros. Os quiero.

A mis hermanos Imanol, Tania y Aritz por esa chispa de felicidad y por llenarme la vida de momentos alegres. A vosotros os debo mi actitud optimista y chistosa. A mi abuelo Rafael, por su entereza y por estar siempre a nuestro lado cuidando de nosotros.

A mis otros hermanos, Wiki, Simba, Nala y khaleesi por llenar nuestras vidas de amor sin nunca pedir nada a cambio.

A todos mis amigos y compañeros, Isma, Ander, Christian, Caja Laboral, Eduardo, Igor, Imanol, Marquitos, Juanistico, Badía, El Pesetas, Barrio, Plou, Gorgon, Obís, Alvaro, Alberto, Inés, Haritz, Irati y un sinfín más. Gracias por todos los buenos momentos que hemos pasado juntos.

A todos mis profesores y en especial a mi director Borja Calvo, que confió en mí y en el proyecto, aún habiéndole entregado un simple borrador de la idea. Si no llega a ser por el, hoy seguramente estaría presentando un proyecto muy diferente. Te la jugaste.

A las personas con las que he compartido la gran parte de estos cuatro años. Gracias Vesko, Sebas y Elsa por haber estado durante este curso siempre a mi lado, tanto disfrutando juntos de los buenos momentos, como apoyándome en los malos. Por mucho que ahora toque separarnos, estoy seguro de que nada cambiará entre nosotros.

Por último a mi parcero, desde que nos tropezamos en primero, se ha forjado una amistad increíble entre nosotros que ni siquiera la distancia o el tiempo ha podido desgastar lo más

mínimo. En ocasiones roza lo surrealista lo mucho que nos entendemos y compenetramos. Si es verdad lo que se dice de que todos tenemos un alma gemela, sin duda ese eres tú, gracias por todo Johan.

Finalmente me gustaría hacer una mención a una persona muy especiales.

A mi abuela Pilar, que falleció durante la realización de este proyecto. Una persona que se desvivía por sus seres queridos y llenaba de afecto nuestras vidas. Aunque no pueda estar hoy aquí viendo a su nieto graduarse, se que estaría muy orgullosa.

Índice general

Índice de figuras

1. CAPÍTULO

Introducción

Desde la aparición de la primera criptomoneda, el Bitcoin, en 2009, el mercado de las criptodivisas ha sufrido un crecimiento desmesurado. Se estima que actualmente existen en el mercado más de 2000 criptomonedas que mueven aproximadamente 300.000 millones de dolares en las distintas casas de cambio. Su influencia no pasa desapercibida para nadie. Los medios de comunicación se han hecho eco del tema, y a día de hoy es extraño encontrar a una persona que no haya escuchado o leído nada sobre estas nuevas divisas basadas en la tecnología blockchain.

El blockchain es un tipo de libro de cuentas distribuido para mantener un registro permanente y a prueba de manipulaciones de datos transaccionales. Funciona como una base de datos descentralizada que es administrada por ordenadores pertenecientes a una red de punto a punto, o peer to peer (P2P). Cada uno de los equipos de la red mantiene una copia del libro de registros para evitar fallos e inconsistencias y todas las copias se actualizan y validan simultáneamente. Sin duda, es una tecnología que seguirá creciendo y sobre la que se seguirá investigando en los próximos años.

Desde que surgió la segunda criptomoneda más famosa, el Ethereum[1], se investigó y se descubrió el enorme negocio de especulación que hay detrás. Grupos enormes de más de 100.000 personas comprando y vendiendo simultáneamente controlaban el precio de las monedas con total impunidad. En la literatura existen artículos de personas que ya proponían proyectos para aprovecharlo, pero la mayoría están enfocados a crear un cliente de

[1]El Ethereum, es una plataforma de programación, un lenguaje de programación, un protocolo y una moneda (Ether) creada para financiar el proyecto. Su creador, Vitalik Buterin utiliza esta tecnología con el fin de descentralizar programas y aplicaciones informáticas, para ser accesibles a todo el mundo.

Telegram con el que detectar las ordenes de compra y venta lo más rápido posible. Hasta donde sabemos, no existe ningún trabajo que plantee un sistema de detección basado en la evaluación del precio, motivo por el cual planteamos el presente proyecto.

1.1. Definiciones y conceptos

Este proyecto requiere de algunos conocimientos técnicos referentes a la economía y las criptodivisas. Estas son algunas de las definiciones que ayudarán a contextualizar el problema.

Mercados bursatiles. El mercado bursátil es el conjunto de todas aquellas instituciones, empresas e individuos que realizan transacciones de productos financieros en diferentes Bolsas alrededor del mundo.

Bolsa. La Bolsa o mercado de valores es el lugar donde se llevan a cabo todas las operaciones de compra y venta de valores, tanto las acciones como los bonos públicos o privados y otros activos financieros.

Símbolos Bursátiles. Un símbolo bursátil o ticker es un código alfanumérico que sirve para identificar de forma abreviada la empresa o divisa que cotizan en el mercado bursátil. Por ejemplo Apple - AAPL, Euro - EUR, Bitcoin - BTC.

Formato OHLC y Volumen. El valor de un activo o divisa normalmente se representa mediante el formato de velas japonesas OHLC, Open, High, Low, Close (Ver Figura 1.1). Una vela financiera representa un intervalo de tiempo para el cual se obtienen el valor de apertura (Open), el máximo valor alcanzado (High), el mínimo valor alcanzado (Low) y el valor de cierre (Close) para un ticker en concreto.

El espacio entre la apertura y el cierre se llama cuerpo de la vela y representa el cambio real sufrido en un intervalo concreto. A la parte, tanto por encima hasta valor de High, como por debajo hasta el valor de Low, se le denomina sombra superior e inferior respectivamente (Ver Figura 1.1).

Además de los datos anteriores, es muy común encontrarse el volumen de una acción o divisa, es decir la cantidad de unidades de esta que se ha movido (compras y ventas) en 24h.

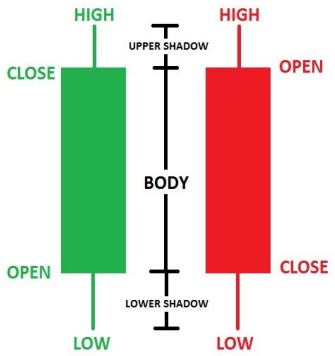

Figura 1.1: Ejemplo gráfico de velas japonesas y su significado.

Figura 1.2: Representación de una operación de compra a bajo precio y venta a alto

Para el objetivo de este proyecto, los datos de las velas se han recogido cada minuto, ya que se espera que la mayoría de eventos se produzcan en intervalos de tiempo pequeños. Para determinar el valor de una divisa en un fecha concreta, se utilizará únicamente el valor de cierre (Close).

Especulación bursatil. La especulación bursátil es el conjunto de operaciones financieras que tienen como fin obtener un beneficio económico aprovechando exclusivamente las fluctuación en el tiempo del mercado de valores, en lugar de tratar de sacar beneficio mediante intereses o dividendos. La Figura 1.2 muestra lo que sería una operación especulativa.

El beneficio de una operación especulativa es la diferencia entre el precio de venta y el de compra (sin tener en cuenta las comisiones).

Criptodivisas. Una criptodivisa o criptomoneda, es una moneda digital descentralizada

#	Nombre	Cap. de Mercado	Precio	Volumen (24h)	Acciones en circulación	Cambio (24h)	Precio (7 días)
1	Bitcoin	$172.165.430.966	$9.614,00	$11.220.747.440	17.907.787 BTC	0,17%	
2	Ethereum	$18.191.243.840	$169,14	$5.421.704.869	107.550.249 ETH	-0,21%	
3	XRP	$11.083.594.299	$0,257850	$808.514.398	42.984.656.144 XRP *	0,85%	
4	Bitcoin Cash	$5.000.393.101	$278,14	$1.149.690.424	17.977.675 BCH	-1,18%	
5	Litecoin	$4.024.911.903	$63,73	$2.183.953.117	63.153.412 LTC	-0,89%	
6	Tether	$4.024.742.308	$1,00	$13.127.898.107	4.008.269.411 USDT *	0,16%	
7	Binance Coin	$3.414.045.835	$21,95	$170.955.132	155.536.713 BNB *	-2,27%	
8	EOS	$2.987.154.140	$3,21	$942.546.623	929.773.648 EOS *	-0,76%	
9	Bitcoin SV	$2.292.693.728	$128,41	$280.022.573	17.854.986 BSV	-0,67%	
10	Stellar	$1.232.994.503	$0,062782	$98.021.351	19.639.376.193 XLM *	-0,31%	
11	Cardano	$1.164.043.250	$0,044897	$50.300.465	25.927.070.538 ADA	-1,20%	
12	Monero	$1.153.177.773	$67,11	$39.325.396	17.184.078 XMR	-0,95%	
13	UNUS SED LEO	$1.120.658.120	$1,12	$5.609.025	999.498.893 LEO *	-1,89%	

Figura 1.3: Ejemplo de de algunas de las criptomonedas de la página web CoinMarket

que emplea técnicas de cifrado para reglamentar la generación de unidades de moneda y verificar la transferencia de fondos. El Bitcoin es el ejemplo más común de criptomoneda, no obstante, no es la única. La Figura 1.3 muestra álgunas de las criptomonedas de la página web CoinMarket junto con su capitalización de mercado[2], el precio de una unidad, volumen, unidades en circulación, la variación del precio en 24h y una gráfica de su comportamiento en los últimos siete dias.

Desde su aparición en 2009 han surgido cientos de nuevas criptomonedas, algunas con un proyecto y una filosofía detrás, y otras únicamente con fines especulativos.

Casas de cambio. El intercambio (compra y venta) de criptomonedas, se realiza en casas de cambio o criptocurrency exchanges, plataformas digitales que permiten intercambiar criptomonedas por dinero de curso legal y/u otras criptomonedas o mercancías. Algunos ejemplos de casas de cambio son: Coinbase, Bitfinex, Poloniex, Bittrex. La Figura 1.4 muestra la página principal de la casa de cambios Bittrex.

[2]La capitalización de mercado es una medida de una empresa o su dimensión económica, y es igual al precio por acción en un momento dado multiplicado por el número de acciones en circulación de una empresa, e indica el patrimonio disponible para la compra y venta activa en bolsa.

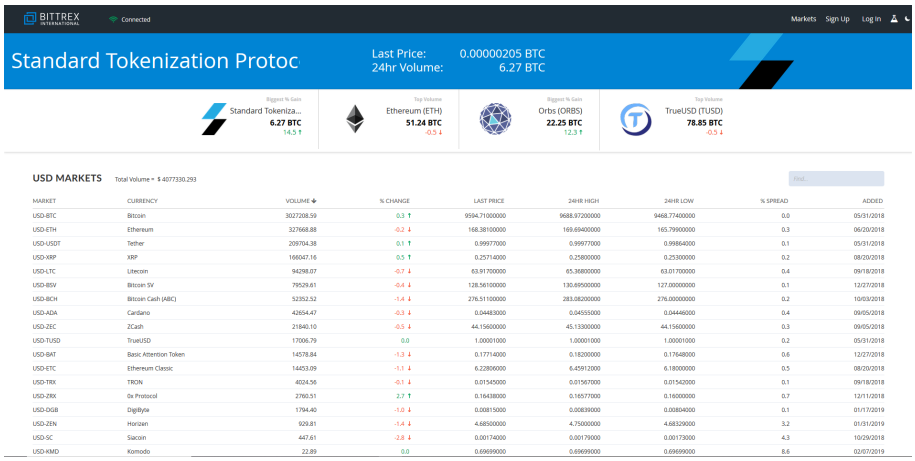

Figura 1.4: Página principal de Bittrex, uno de las casas de cambio más famosas

1.1.1. Indicadores financieros

Los indicadores financieros se centran en datos históricos, como el precio de cierre o volumen, y ayudan a analizar movimientos o tendencias a corto y largo plazo.

Dada la complejidad del estado financiero, es posible encontrar una gran cantidad de indicadores técnicos.

Para la extracción de las variables predictoras, utilizaremos únicamente dos indicadores.

No obstante en este proyecto utilizaremos únicamente dos muy sencillos.

Media Móvil

En estadística, una media móvil es un cálculo utilizado para analizar un conjunto de datos en modo de puntos para crear series de promedios. Así las medias móviles son una lista de números en la cual cada uno es el promedio de un subconjunto de los datos originales.

Constituyen una forma fácil y simple de suavizar la acción del precio ofreciendo una correlación suavizada entre la cotización de divisas y el transcurso del tiempo. El período de la media móvil es el intervalo de tiempo sobre el que se calcula la media de los precios, por ejemplo, un período de 5 hará que la media móvil se calcule sobre las últimas 5 velas.

Existen varios tipos de media móvil, pero la más simple y también más utilizada es la Media Móvil Simple, que es la media aritmética de los valores durante el periodo. Existen otras alternativas, como la Media Móvil Ponderada, que realiza una media ponderada dependiendo de la antigüedad de los datos, o la Media Móvil Exponencial, que es igual que

Figura 1.5: Ejemplo de 2 medias móviles, simple (roja) y exponencial (azul)

la Media Móvil Ponderada con un factor exponencial. La Figura 1.5 muestra la diferencia entre la media móvil simple y exponencial.

Volatilidad

La volatilidad es el término que mide la variabilidad de las trayectorias o fluctuaciones de los precios, de las rentabilidades de un activo financiero, de los tipos de interés y en general, de cualquier activo financiero en el mercado.

Una volatilidad alta, implica que el precio de un activo varia mucho y rápido.

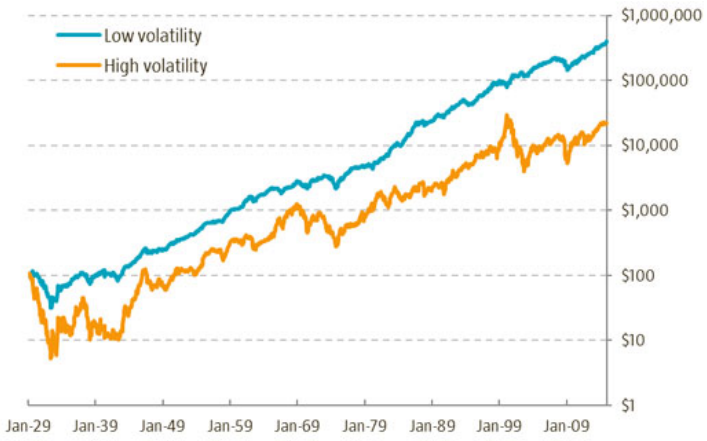

Figura 1.6: Ejemplo de volatilidad baja (azul) y alta (naranja).

Matemáticamente se expresa como:

$$\sigma = \frac{\sigma_{SD}}{\sqrt{P}} \tag{1.1}$$

Donde σ_{SD} es la desviación estándar y P es el periodo (normalmente en años) de retorno. La Figura 1.6 muestra gráficamente la diferencia entre alta y baja volatilidad.

1.2. Pump and Dump

El Pump and Dump es una práctica especulativa, que consiste en aumentar el precio de las acciones de una empresa, a través de recomendaciones basadas en declaraciones o noticias falsas, engañosas o muy exageradas con el fin de venderlas y obtener beneficio. Los organizadores de esta práctica, compran acciones a un precio determinado, para luego, mediante prácticas ilícitas, crear en ellos una expectativa de subidas. Esta expectación, atrae a nuevos inversores, aumentando la demanda, y en consecuencia, subiendo el valor de las acciones. Una vez que el valor haya alcanzado el límite fijado por los organizadores, estos venden todas las acciones en masa, generando una fuerte caída en el precio y obteniendo un gran beneficio, a costa de generar perdidas a los nuevos accionistas.

En los mercados convencionales, esta práctica es ilegal. No obstante, el mercado de las criptomonedas carece de regulación por lo que este tipo de prácticas son habituales.

Existen canales de Telegram[3] donde grupos de desconocidos se organiza para invertir si-

[3] Aplicación de mensajería instantánea desarrollada desde el año 2013 por los hermanos Nikolái y Pável

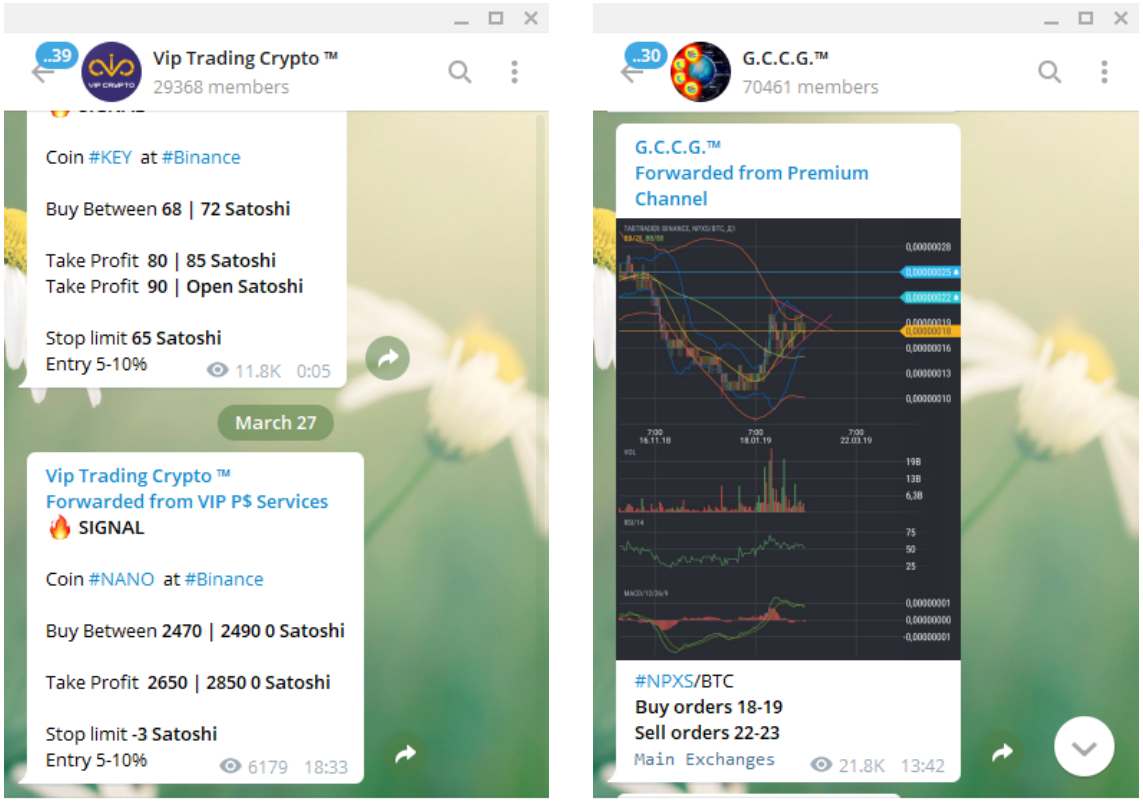

Figura 1.7: Grupos de difusión de Telegram con ordenes de compra y venta.

multaneamente sobre una criptomoneda determinada creando un efecto Pump and Dump [12] (ver Figura 1.7).

Estos grupos se organizan por jerarquias. En primer lugar, un grupo selecto de administradores deciden sobre que criptomoneda invertir. Esa información se pasa a grupos privados denominados “VIP” donde los usuarios disponen de un tiempo para invertir sobre las criptomonedas. Finalmente la información se difunde por canales públicos con un gran número de participantes, creando una subida artificial del precio (Pump).

Si la inversión de los miembros de los canales “VIP” es lo suficientemente grande, es capaz de generar una anomalía en la serie temporal. La Figura 1.8 muestra un ejemplo de este tipo de eventos.

Se observa que en los minutos antes del Pump, hay más movimiento de compra-venta de lo normal [11]. El objetivo del proyecto consiste en predecir los eventos Pump and Dump mediante la información que nos aportan esas pequeñas anomalías.

Dúrov.

Figura 1.8: Pump and Dump

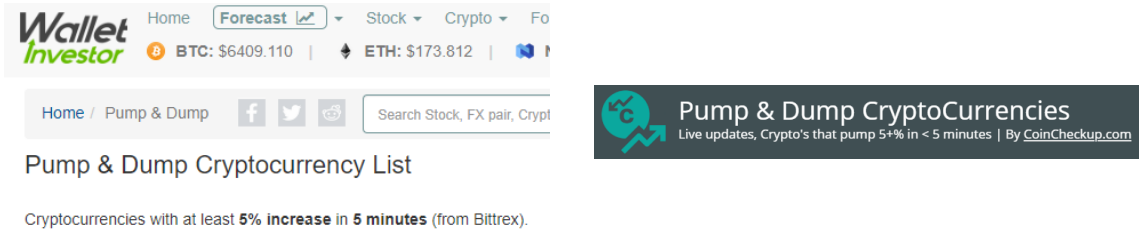

Figura 1.9: Wallet Investor y Pump & Dump CryptoCurrencies, páginas web dedicadas a detectar eventos Pump and Dump con el método del 5 %

1.2.1. Criterio de detección de eventos

Para la detección de eventos Pump and Dump existe un criterio muy común recogido en las páginas web dedicadas a ellos. Consiste en detectar un incremento del precio de cierre de un 5 % en un intervalo menor de 5 minutos [8, 3].

Dado que para el proyecto disponemos de una cantidad elevada de datos, utilizaremos el criterio de detectar subidas del 10 % en intervalos de 5 minutos para captar eventos más interesantes desde un punto de vista económico.

2. CAPÍTULO

Planificación y Gestión del Proyecto

2.1. Objetivos del proyecto

Como ya se ha comentado, el objetivo del proyecto consiste en crear un clasificador capaz de detectar eventos Pump and Dump a partir de los datos históricos.

Primeramente, se construirá una base de datos con los valores OHLC y volumen de algunas de las criptomonedas.

Se seleccionará un criterio para la detección de eventos Pump and Dump y se aplicará sobre nuestra base de datos.

A continuación, se someterán los datos a un primer procesado, donde se acondicionarán para el problema en cuestión.

Finalmente, se realizará una fase experimental, donde se crearán y evaluaran los modelos predictivos y se ejecutarán sobre la base de datos. Tras esto, se comentarán los resultados y se extraerán las conclusiones.

2.2. Tareas y estimación de tiempos

Hemos dividido el proyecto en las siguientes subtareas:

- Planificación y Gestión del proyecto: Comprende todas las tareas relacionadas con

Tareas	**Tiempo real**	**Tiempo estimado**
Planificación y gestión	11	13
Busqueda y lectura de información	17	20
Reuniones	12	12
Memoria	70	70
Captura y preprocesado de los datos	88	80
Experimentación	74	100
Visualización de los datos	27	10
Total	**299**	**300**

Tabla 2.1: Tabla con las tareas, el tiempo real de realización y el tiempo estimado

la planificación y seguimiento y control del proyecto, junto con las reuniones con el tutor.

- Búsqueda y lectura de información: Son todas aquellas tareas dedicadas a la formación sobre temas relacionados con el proyecto.
- Captura y preprocesado de datos: Son todas aquellas tareas relacionadas con la captura de datos, el preprocesado, acondicionamiento de los datos, calculo de derivadas, extracción de features y tratamiento de outliers.
- Diseño y ejecución de la experimentación: recopila todas las tareas relacionadas con la construcción de los clasificadores, optimización y evaluación.
- Visualización de los datos: Son las tareas relacionadas con la visualización de los datos, ya sea como gráficas, tablas y matrices de confusión.
- Memoria: Todas las tareas relacionadas con la redacción de la memoria.

2.3. Análisis de riesgos

Este proyecto, se trata de un trabajo completo que incluye todas las tareas, desde la captura de los datos, pasando por el preprocesado, hasta la construcción y evaluación de los algoritmos de clasificación. Asimismo, existen pocas publicaciones sobre temas relacionados. Por lo tanto, asumimos que no esta exento de riesgos.

Primero, la perdida de la información por problemas con la máquina de trabajo. Para evitar esto, se usarán servicios cloud (Drive para ficheros, BitBucket para el código, y Overleaf para la memoria) para guardar la información.

Por otra parte, necesitamos un conjunto de datos grande, para que contenga suficientes eventos Pump and Dump (ver sección 1.2) con los que entrenar nuestros algoritmos. se prevee capturar los datos directamente de las casas de cambio para generar nuestra propia base de datos. Existe el riesgo de que en ese tiempo no se den suficientes eventos Pump and Dump. En tal caso, se procurará a buscar bases de datos ya existentes.

Por ultimo, no hay mucha información sobre investigaciones en esta línea, por lo que desconocemos que tipo de resultados esperar y si resulta viable un proyecto de estas características. Cabe la posibilidad de que los resultados no sean satisfactorios. En tal caso se analizará la situación para evaluar la viabilidad de realizar mejoras que lleven a resultados positivos.

3. CAPÍTULO

Captura y preprocesado de los datos

Antes de comenzar a manipular los datos, es importante comprenderlos, especialmente en proyectos que requieren manejar información sobre dominios técnicos.

A menudo, cuando se trabajo con datos en crudo, por una parte, mucha de la información puede resultar repetida y/o irrelevante para nuestro problema. Por otra, la información puede que no este representada en un formato adecuado con el que trabajar.

El preprocesado de los datos engloba todas las operaciones o técnicas que se aplican para construir un conjunto de datos adecuado para un problema determinado. En otras palabras, mediante el preprocesado, pasamos de tener un conjunto de datos en crudo a tener un conjunto de datos preparado para los métodos a utilizar.

Asimismo, vamos a realizar un análisis previo de los datos con el objetivo de entender el formato, la estructura y la organización de los ficheros que tendremos que manipular. Se presentarán visualmente, las características principales de los datos para facilitar su comprensión.

Por último, comentaremos las operaciones aplicadas para el filtrado y acondicionado de los datos.

3.1. Captura de los datos

Una de las primeras tareas del proyecto fue la recogida de datos de la plataforma Bitfinex. Mediante un pequeño script se hicieron peticiones cada 3 segundos a su API pública

Figura 3.1: Ejemplo de la función ticker en python

pidiendo los valores OHLC y volumen para las criptomonedas Ethereum, Ripple y DashCoin.

Desafortunadamente, durante los 4 meses que estuvo en funcionamiento el script, no se detecto ni un solo evento Pump and Dump en esas monedas. De esta experiencia, se concluyó que hacía falta un volumen mucho más grande de datos y seleccionar criptomonedas menos conocidas, ya que estas tienden a ser el objetivo habitual de los especuladores.

Debido a la dificultad que suponía capturar datos útiles, se optó por buscar bases de datos públicamente disponibles. Encontramos una base de datos de la plataforma de exchange Bittrex con la información entre los años 2014 y 2017, cedida por Rami Kawach [1] para uso personal. Esta base de datos cuenta con los precios de apertura (open), cierre (close), máximo (high), mínimo (low) y volumen de 219 criptomonedas, con una frecuencia de muestreo de 1 minuto, al cual llamaremos tick. Para simplificar, se investigo sobre la frecuencia de eventos Pump and Dump en las criptomonedas y se selecionaron tres para el proyecto, las cuales, según la página Pump & Dump CryptoCurrencies[2], cuentan con un alto número de ellos. Estas monedas son Doge Coin, Cannabis Coin y Feather Coin

3.1.1. Ficheros de datos

Los datos están repartidos en 219 ficheros donde cada uno contiene los datos de los precios de una de las 219 criptomoneda respecto a otra, normalmente el Bitcoin.

[1] Rami Kawach, Co-Fundador y Chief Technology Officer de Bittrex.
[2] https://pumpdump.coincheckup.com/

```
# BTC-ETH Copyright Bittrex Inc. 2017

[TimeStamp], [Open], [Close], [High], [Low], [Volume], [BaseVolume]
8/14/2015 9:06:00 AM,0.00690000,0.00690000,0.00690000,0.00690000,1.34611713,0.00928820
8/14/2015 9:10:00 AM,0.00690000,0.00690000,0.00690000,0.00690000,0.37540654,0.00259030
8/14/2015 9:13:00 AM,0.00690000,0.00690000,0.00690000,0.00690000,0.18775374,0.00129550
8/14/2015 9:23:00 AM,0.00690000,0.00696000,0.00696000,0.00690000,0.09382613,0.00064758
8/14/2015 9:25:00 AM,0.00696000,0.00716000,0.00716000,0.00696000,4.98869811,0.03506968
8/14/2015 9:32:00 AM,0.00716000,0.00730000,0.00730000,0.00716000,2.59471613,0.01892697
8/14/2015 9:36:00 AM,0.00730000,0.00800000,0.00800000,0.00730000,3.49398315,0.02790592
8/14/2015 9:37:00 AM,0.00788684,0.00788684,0.00788684,0.00788684,0.70405177,0.00555274
8/14/2015 9:38:00 AM,0.00800000,0.08000000,0.08000000,0.00800000,0.10000000,0.00695606
8/14/2015 9:46:00 AM,0.00792292,0.07800000,0.07800000,0.00792292,0.50000000,0.01433105
8/14/2015 9:48:00 AM,0.00797500,0.00797500,0.00797500,0.00797500,0.17601294,0.00140370
8/14/2015 9:50:00 AM,0.01786500,0.01786500,0.01786500,0.01786500,2.42298565,0.04328662
8/14/2015 9:51:00 AM,0.00802553,0.00802553,0.00802553,0.00802553,0.08800648,0.00070629
8/14/2015 10:46:00 AM,0.01605106,0.01605106,0.01605106,0.01605106,0.36634234,0.00588018
8/14/2015 11:39:00 AM,0.01600000,0.01600000,0.01600000,0.01600000,1.80369760,0.02885916
8/14/2015 11:41:00 AM,0.01600000,0.01600000,0.01600000,0.01600000,0.05396974,0.00086351
8/14/2015 11:59:00 AM,0.01500000,0.01500000,0.01500000,0.01500000,19.63813163,0.29457197
8/14/2015 12:04:00 PM,0.01421000,0.01499999,0.01499999,0.01421000,5.20280695,0.07790501
8/14/2015 12:09:00 PM,0.01499999,0.01500000,0.01500000,0.01499999,77.92136250,1.16881969
8/14/2015 12:10:00 PM,0.01500000,0.01500000,0.01500000,0.01500000,218.92722355,3.28390835
8/14/2015 12:11:00 PM,0.01500000,0.01500000,0.01500000,0.01500000,3.07823047,0.04617345
8/14/2015 12:17:00 PM,0.01499999,0.01499999,0.01499999,0.01499999,1.25673126,0.01885095
8/14/2015 12:18:00 PM,0.01499990,0.01499990,0.01499990,0.01499990,0.22898634,0.00343477
8/14/2015 12:28:00 PM,0.01100005,0.01100005,0.01100005,0.01100005,4.98869811,0.05487592
8/14/2015 12:32:00 PM,0.01100003,0.01100003,0.01100003,0.01100003,4.82429683,0.05306740
8/14/2015 12:34:00 PM,0.01489999,0.01489999,0.01489999,0.01489999,0.16898197,0.00251782
8/14/2015 12:45:00 PM,0.01100006,0.01100002,0.01100006,0.01100002,21.96104832,0.24157227
8/14/2015 12:46:00 PM,0.01100005,0.01100001,0.01100005,0.01100001,7.76889020,0.08545789
8/14/2015 12:48:00 PM,0.01100005,0.01399998,0.01399998,0.01100005,1.79143807,0.02203685
8/14/2015 12:55:00 PM,0.01100001,0.01100000,0.01100001,0.01100000,4.34746827,0.04782214
8/14/2015 1:03:00 PM,0.01199996,0.01199996,0.01199996,0.01199996,0.09573673,0.00114883
```

Figura 3.2: Datos del fichero BTC-ETH.csv

La Figura 3.2 muestra un extracto del fichero BTC-ETH.csv que contiene el valor del Ethereum (ETH) respecto al Bitcoin (BTC).

Hay que tener en cuenta que no todas las criptomonedas tienen como referencia el Bitcoin. Teóricamente cualquier moneda puede expresarse en referencia a cualquier otra, no obstante, lo más común es encontrarlas expresadas respecto al Bitcoin o Ethereum, las dos criptomonedas más importantes en el mercado de las criptodivisas. Para evitar problemas de conversión, las monedas seleccionadas tienen todas como referencia el Bitcoin.

Como hemos visto, los ficheros csv contienen los datos de una criptomoneda en formato OHLC (ver Sección 1.1). Las filas repetidas están eliminadas para reducir el tamaño del fichero.

- TimeStamp: Fecha, fecha a la que se tomo el dato, con el siguiente formato de representacion. mm/dd/aaaa hh:mm:ss AM/PM. Por ejemplo 8/10/2014 7:19:00 AM.
- Open: Número real, precio de apertura de la criptomoneda para un tick.
- Close: Número real, precio de cierre de la criptomoneda para un tick.
- High: Número real, precio máximo alcanzado por la criptomoneda en un tick.
- Low: Numero real, precio mínimo alcanzado por la criptomoneda en un tick.
- Volume: Numero real, cantidad de transacciones de la moneda en las últimas 24h.

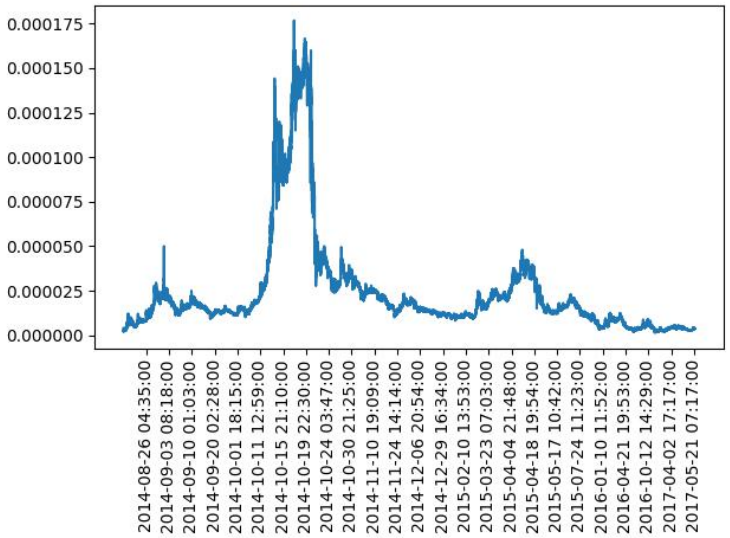

Figura 3.3: Representación gráfica de los valores de cierre de la criptomoneda BTC-CANN durante el periodo 2014-2017

Para realizar un análisis completo de la información disponible se utilizarían todos los indicadores. No obstante, dado que se trata de una primera aproximación, utilizaremos únicamente los indicadores de cierre y de volumen, los cuales contienen la mejor información. Notese que, en términos de la serie, el precio de apertura y de cierre son idénticos con un desfase de un minuto (ver Figura 3.2). Por otra parte, no es de esperar que los valores de High y Low estén muy alejados dado el corte intervalo de tiempo.

3.2. Visualización de datos

En esta sección vamos a presentar los datos de una manera gráfica, para analizar mejor sus características principales.

Si representamos los datos en una gráfica obtenemos una representación visual del comportamiento de la criptomoneda en un determinado intervalo de tiempo. A continuación se muestra en las Figuras 3.3, 3.4 y 3.5 el comportamiento global de las monedas.

3.2.1. Visualización de evento Pump and Dump

Como ya se ha indicado, los eventos Pump and Dump se caracterizan por fuertes subidas en intervalos de tiempo pequeños, por lo que, suelen ser fáciles de detectar visualmente.

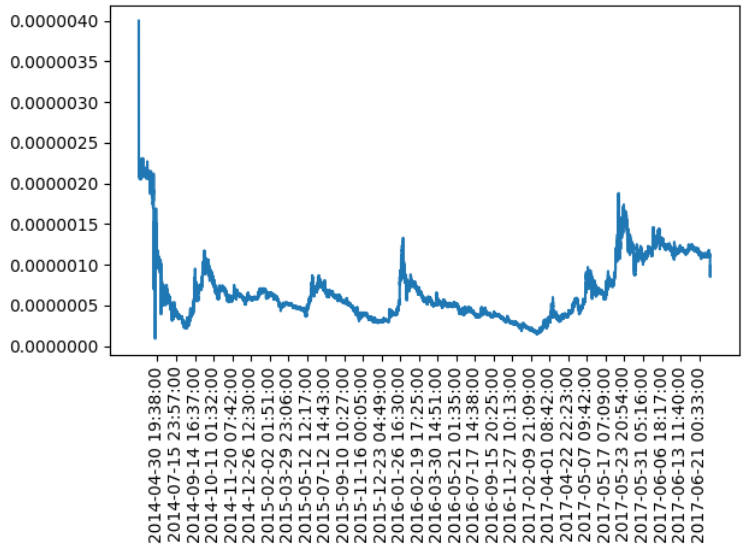

Figura 3.4: Representación gráfica de los valores de cierre de la criptomoneda BTC-DOGE durante los años 2014-2017

En las Figuras 3.6 y 3.7 se observa como se cumple el comportamiento previsto, el Pump, es decir, una rápida subida ($>10\,\%$) y escasos minutos después la caída o Dump.

En la Figura 3.6 se observa un caso en el que el crecimiento del precio de la moneda es muy elevado en torno al $100\,\%$s. Por otra parte, también se aprecia como minutos antes del evento Pump and Dump, el precio sufre unas fluctuaciones anormales, tal como comentábamos en la sección 1.2

3.3. Errores en el conjunto de datos

Dado que los datos son cedidos de la propia plataforma y que la extracción de datos no requiere de instrumentos físicos o de intervención humana, suponemos que no hay errores de ruido. Tampoco se ha encontrado ningún valor perdido ni se espera que haya atípicos (outliers).

3.4. Preprocesado de datos

El preprocesamiento de datos es un paso preliminar durante el proceso de minería de datos. Se trata de cualquier tipo de procesamiento que se realiza con los datos crudos, para transformarlos en otros más fáciles de interpretar por los algoritmos de aprendizaje.

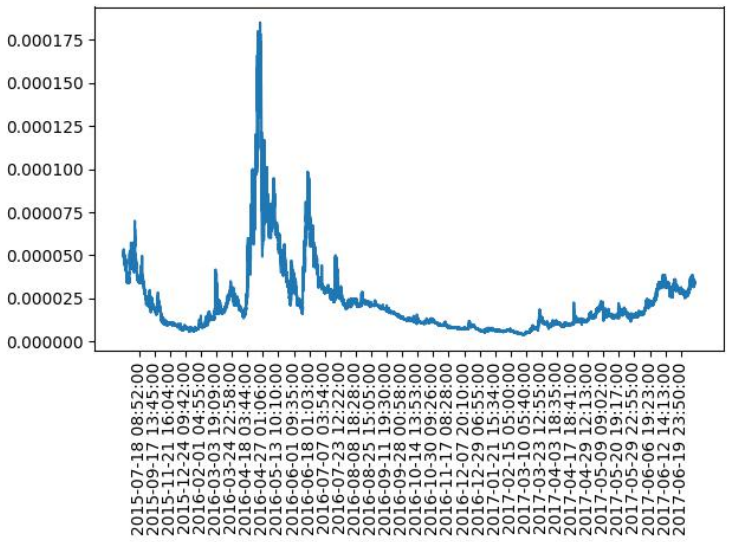

Figura 3.5: Representación gráfica de los valores de cierre de la criptomoneda BTC-FTC durante los años 2014-2017

El preprocesado de los datos es necesario para garantizar la aplicabilidad de las técnicas a utilizar, así como eliminar la información que no sea relevante para el problema reduciendo el volumen de los datos.

3.4.1. Estandarización y eliminación de datos

Lo primero que se aprecia de los datos en crudo es que tanto los nombres de las columnas como el formato de representación de la fecha no son los mas adecuados para trabajar con ellos. Asimismo, hay varias columnas que contienen información inútil para este proyecto.

En un primer filtrado cambiamos los nombres, el formato de representación de las fechas y eliminamos la información indeseada.

Estos son los cambios que se han realizado:

- Eliminamos las columnas de [Open], [High], [Low], [BaseVolume].
- Cambiamos los nombres de las columnas [Timestamp], [Close], [Volume] por Date, Close y Volume para facilitar la lectura y las labores de programación.
- Cambiamos la fecha al formato de representación estandar de Python, esto es, de mm/dd/yyyy HH:MM:SS AM/PM por yyyy-mm-dd HH:MM:SS.

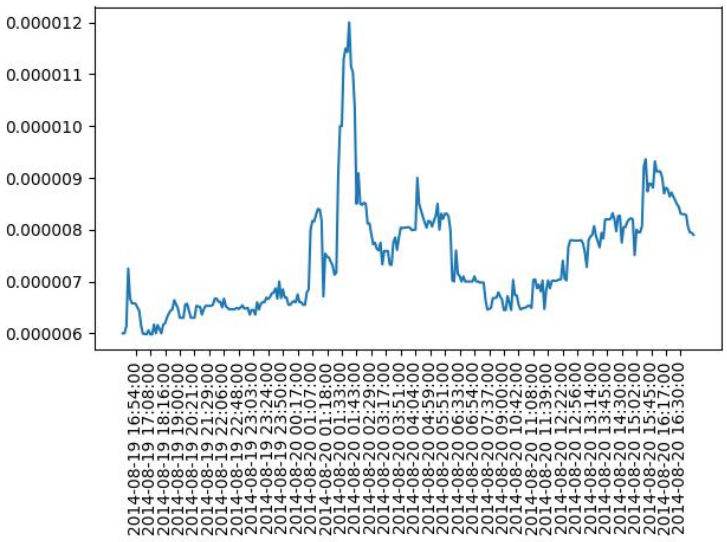

Figura 3.6: Ejemplo 1 de evento Pump and Dump detectado para la criptomoneda BTC-CANN

- Eliminamos las filas repetidas, es decir, en casos donde a lo largo de un periodo de tiempo la criptomoneda no ha sufrido ni un cambio (ni compra ni venta), la siguiente entrada es idéntica a la anterior por lo que no nos aporta ninguna información.

En la Figura 3.8 podemos ver la diferencia entre los datos antes y después del filtrado.

3.4.2. Datos auxiliares

Para el cálculo de ciertas variables predictoras necesitamos algunos datos auxiliares, los cuales únicamente nos sirven para operaciones concretas en la extracción de características. Estos son, la diferencia temporal en minutos entre un dato y el siguiente, necesario para el cálculo de la derivada (ver Sección 3.4.4) y el logaritmo neperiano del cociente entre un valor de cierre y el anterior, el cual denominamos variación de precio vp, necesario para el cálculo de la volatilidad.

$$vp(t) = ln(f(t)/f(t-h)) \tag{3.1}$$

Donde t es un instante de tiempo dado, h una cierta cantidad de tiempo y $f(t)$ es el valor de la serie temporal para el tiempo t.

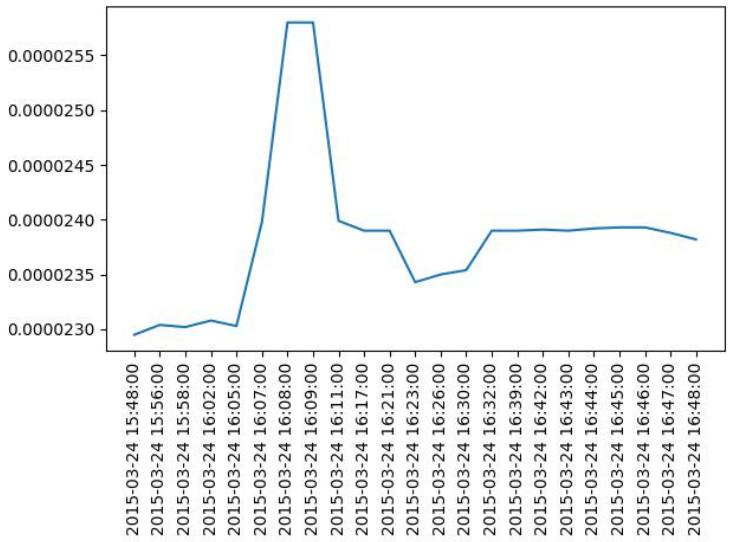

Figura 3.7: Ejemplo 2 de evento Pump and Dump detectado para la criptomoneda BTC-CANN

3.4.3. Detección de eventos

La siguiente tarea del proyecto consiste en detectar los eventos Pump and Dump para posteriormente clasificar las filas y entrenar algoritmos de clasificación supervisada.

Tal como comentábamos en la Sección 1.2.1, la tarea consiste en detectar subidas del 10 % en intervalos menores a 5 minutos. Mediante Python, creamos un programa que recorre el archivo y por cada fila, verifica si las 5 próximas filas están a menos de 5 minutos, si es así, comprueba la diferencia de los precios y si esta es mayor a un 10 % marca la fecha de la primera fila como el inicio de un evento Pump and Dump.

Debido a la simpleza del algoritmo, se detectan gran cantidad de eventos en intervalos muy cortos (ver Figura 3.9). Resulta lógico pensar que cuando se detectan varios Pumps en fechas muy cercanas, ambos forman parte de un único Pump de mayor duración. Para ello, a partir de un Pump, si se detecta otro en un intervalo menor a 6h, los uniremos en uno único con la hora de inicio del primero y hora de finalización del segundo (ver Figura 3.10).

Los resultados obtenidos los guardamos en un fichero csv con las siguientes columnas.

- PumpStarDate. Fecha, hora a la que se detecta el inicio del Pump.
- PumpEndDate. Fecha, hora a la que se detecta el final del Pump.
- PumpStartValue. Número real, precio al inicio del Pump.

[TimeStamp]	[Open]	[Close]	[High]	[Low]	[Volume]	[BaseVolume]
08/10/2014 07:19:00, AM	0.00000399	0.00000399	0.00000399	0.00000399	150	0.0005985
08/10/2014 07:23:00, AM	0.00000207	0.00000207	0.00000207	0.00000207	3085.782915	0.00638757
08/10/2014 07:34:00, AM	0.000003	0.000003	0.000003	0.000003	245.56854	0.0007367
08/10/2014 08:12:00, AM	0.00000346	0.00000346	0.00000346	0.00000346	250	0.000865
08/10/2014 08:14:00, AM	0.000003	0.000003	0.000003	0.000003	364.43146	0.00109328
08/10/2014 08:59:00, AM	0.00000227	0.0000022	0.00000227	0.0000022	630	0.00140544
08/10/2014 09:41:00, AM	0.0000027	0.0000027	0.0000027	0.0000027	485.3834526	0.00131053
08/10/2014 09:51:00, AM	0.0000028	0.0000028	0.0000028	0.0000028	1000	0.0028
08/10/2014 10:02:00, AM	0.00000227	0.00000227	0.00000227	0.00000227	501.1641077	0.00113764

Date	Close	Volume
2014-08-10 07:19:00	3.99E-06	150
2014-08-10 07:23:00	2.07E-06	3085.782915
2014-08-10 07:34:00	3.00E-06	245.56854
2014-08-10 08:12:00	3.46E-06	250
2014-08-10 08:14:00	3.00E-06	364.43146
2014-08-10 08:59:00	2.20E-06	630
2014-08-10 09:41:00	2.70E-06	485.3834526
2014-08-10 09:51:00	2.80E-06	1000
2014-08-10 10:02:00	2.27E-06	501.1641077

Figura 3.8: Tabla sin preprocesar (rojo) y tabla procesada (verde)

- PumpEndValue. Número real, precio al final del Pump.

En la Sección 3.4.4, veremos que estos datos se utilizan para el etiquetado.

3.4.4. Descripción de las variables predictoras

En esta sección, vamos a describir brevemente, las variables predictoras con las que trabajaremos. Recordamos que los eventos Pump and Dump (ver Sección 1.2) se caracterizan por un cambio brusco en los precios de una criptomoneda. Para detectar estos cambios, vamos a centrarnos en el estudio del valor de la moneda como tal, su tendencia o derivada y la variación de la tendencia o derivada segunda.

La **derivada** $f'(t)$ de una serie temporal es otra serie temporal donde cada punto se aproxima mediante la fórmula.

$$f'(t) = \frac{f(t+h) - f(t)}{h} \tag{3.2}$$

En caso h de que sea constante, puede expresarse mediante diferencias finitas progresivas.

$$\Delta_h f(t) = f(t+h) - f(t) \tag{3.3}$$

Donde t es un instante de tiempo, h una cierta diferencia de tiempo y $f(t)$ es el valor de la serie temporal para un tiempo t.

PumpStartDate	PumpEndDate	PumpStartValue	PumpEndValue
2014-08-10 12:24:00	2014-08-10 12:26:00	2.01e-06	2.42e-06
2014-08-10 12:26:00	2014-08-10 12:29:00	2.42e-06	2.73e-06
2014-08-10 13:34:00	2014-08-10 13:38:00	2.04e-06	2.33e-06
2014-08-10 14:14:00	2014-08-10 14:18:00	2.02e-06	2.32e-06
2014-08-10 14:25:00	2014-08-10 14:28:00	2.33e-06	2.72e-06
2014-08-10 14:28:00	2014-08-10 14:33:00	2.72e-06	3e-06
2014-08-10 14:29:00	2014-08-10 14:33:00	2.72e-06	3e-06
2014-08-10 14:31:00	2014-08-10 14:33:00	2.5e-06	3e-06
2014-08-10 14:32:00	2014-08-10 14:33:00	2.71e-06	3e-06
2014-08-10 14:44:00	2014-08-10 14:49:00	3.08e-06	3.5e-06
2014-08-10 14:47:00	2014-08-10 14:49:00	3.09e-06	3.5e-06
2014-08-10 14:48:00	2014-08-10 14:49:00	3.1e-06	3.5e-06
2014-08-10 14:51:00	2014-08-10 14:54:00	2.57e-06	3.3e-06
2014-08-10 14:59:00	2014-08-10 15:02:00	2.78e-06	3.2e-06
2014-08-10 15:04:00	2014-08-10 15:06:00	2.95e-06	3.3e-06
2014-08-10 15:39:00	2014-08-10 15:44:00	3.05e-06	3.44e-06
2014-08-10 15:40:00	2014-08-10 15:44:00	3.05e-06	3.44e-06
2014-08-10 15:42:00	2014-08-10 15:44:00	2.87e-06	3.44e-06
2014-08-10 17:01:00	2014-08-10 17:02:00	2.99e-06	3.3e-06
2014-08-10 20:04:00	2014-08-10 20:06:00	3.3e-06	3.8e-06
2014-08-10 20:05:00	2014-08-10 20:06:00	3.3e-06	3.8e-06
2014-08-12 07:47:00	2014-08-12 07:50:00	3e-06	3.49e-06

Figura 3.9: Ejemplo de fichero BTC-CANN-Pumps.csv

Con la Formula 3.2, calculamos los puntos de la derivada sustituyendo t por cada uno de los instantes de nuestro conjunto de datos y h, por la diferencia entre ese dato y el siguiente.

Al igual que la derivada $f'(t)$, la **derivada segunda** $f''(t)$ de una serie temporal es otra serie temporal donde cada punto se estima mediante la fórmula.

$$f''(t) = \frac{f'(t+h) - f'(t)}{h} \tag{3.4}$$

Habiendo calculado previamente la derivada primera, calculamos los puntos de la derivada segunda sustituyendo t por cada uno de los instantes de nuestro conjunto de datos y h, por la diferencia entre ese dato y el siguiente.

La **volatilidad** es una medida de la intensidad de los cambios aleatorios o impredecibles en la rentabilidad o en el precio de un título.

PumpStartDate	PumpEndDate	PumpStartValue	PumpEndValue
2014-08-10 12:24:00	2014-08-10 20:06:00	3.3e-06	3.8e-06
2014-08-10 20:04:00	2014-08-12 07:50:00	3e-06	3.49e-06
2014-08-12 07:47:00	2014-08-13 02:37:00	2.81e-06	3.19e-06
2014-08-13 02:36:00	2014-08-13 12:34:00	2.8e-06	3.25e-06
2014-08-13 12:30:00	2014-08-13 22:38:00	5.15e-06	6e-06
2014-08-13 22:37:00	2014-08-14 06:01:00	5.40e-06	5.99e-06
2014-08-14 05:57:00	2014-08-14 12:08:00	1e-05	1.125e-05
2014-08-14 12:03:00	2014-08-14 21:38:00	8.01e-06	9.33e-06
2014-08-14 21:34:00	2014-08-15 03:44:00	7.03e-06	8.19e-06
2014-08-15 03:42:00	2014-08-15 11:45:00	7.9e-06	8.83e-06
2014-08-15 11:40:00	2014-08-15 18:47:00	7.75e-06	8.60e-06
2014-08-15 18:42:00	2014-08-16 03:07:00	6.23e-06	6.93e-06
2014-08-16 03:02:00	2014-08-16 11:53:00	6.3e-06	7.00e-06
2014-08-16 11:48:00	2014-08-17 13:39:00	4.52e-06	5.54e-06
2014-08-17 13:34:00	2014-08-19 07:00:00	5.01e-06	5.74e-06
2014-08-19 06:55:00	2014-08-19 16:27:00	6.14e-06	7.25e-06
2014-08-19 16:26:00	2014-08-20 01:06:00	6.85e-06	7.99e-06
2014-08-20 01:05:00	2014-08-20 12:33:00	7.02e-06	7.79e-06
2014-08-20 12:30:00	2014-08-20 18:49:00	7.75e-06	9.07e-06

Figura 3.10: Ejemplo de fichero BTC-CANN-Pumps-Filtered.csv tras realizar el filtrado

Para el cálculo de la volatilidad, dado que ya hemos calculado previamente la variación de precio (vp) o rentabilidad, aplicamos la siguiente fórmula. Sean C_t y C_{t-1} los indicadores de cierre para los periodos t y $t-1$ respectivamente, la rentabilidad se obtiene como

$$vp = ln(\frac{C_t}{C_{t-1}})$$

Una vez que tenemos calculadas las rentabilidades, calculamos su media

$$vp_{avg} = \sum_{i=1}^{n} \frac{vp_i}{n}$$

Finalmente, para el hallar la volatilidad, calculamos la desviación estándar de las rentabilidades

$$volatility = \sqrt{\frac{\sum_{i=1}^{n}(vp_i - vp_{avg})^2}{n-1}} \tag{3.5}$$

A continuación, vamos a definir y explicar como se han calculado las variables predictoras que utilizaremos en la creación de los modelos de clasificación.

Recordamos que nuestra hipótesis (ver sección 1.2) es que antes de un evento Pump and Dump, debído a la difusión por los grupos “VIP”, el precio sufre ligeras anomalías en su comportamiento. No podemos saber el tiempo que pasa desde que el mensaje se difunde a los grupos “VIP” hasta que llega a los grupos convencionales, por lo tanto, utilizamos ventanas de distintos tamaños para resumir la información aportada por la serie.

La estrategia consiste en, dado un instante de tiempo, calcular la media de los datos de cierre y derivadas y por otro lado la desviación estándar de la volatilidad, para cada uno de los tamaños de las ventanas.

Los tamaños de las ventanas que utilizamos son 1 dia, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos.

Con todo lo anterior, calculamos las siguientes características.

- Date: Fecha sobre la que se calcularan las caracteristicas. Necesaria para el etiquetado.
- Close1d, Close3h, Close1h, Close30m, Close15m, Close5m: Media del valor de cierre 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.
- dClose1d, dClose3h, dClose1h, dClose30m, dClose15m, dClose5m: Media del valor de la primera derivada 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.
- d2Close1d, d2Close3h, d2Close1h, d2Close30m, d2Close15m, d2Close5m: Media del valor de la segunda derivada 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.
- Volume1d, Volume3h, Volume1h, Volume30m, Volume15m, Volume5m: Media del valor del volumen 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos anteriores de la fecha Date.
- dVolume1d, dVolume3h, dVolume1h, dVolume30m, dVolume15m, dVolume5m: Media del valor del volumen 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.
- d2Volume1d, d2Volume3h, d2Volume1h, d2Volume30m, d2Volume15m, d2Volume5m: Media del valor del volumen 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.

- Volatility1d, Volatility3h, Volatility1h, Volatility30m, Volatility15m, Volatility5m: Media del valor de la volatilidad 1 día, 3 horas, 1 hora, 30 minutos, 15 minutos y 5 minutos antes de la fecha Date.
- Class: Variable nominal que indica la clase. Los posibles valores son 'P' - Pump y 'NP' - No Pump.

Todas las características que correspondan a la media de un periodo de tiempo determinado, están divididos por la media de 1 semana atrás, para estandarizar los valores y que sean comparables entre ellos. Esto se hace para compensar la evolución a largo plazo del precio y volumen de la moneda.

3.4.5. Balanceo de las clases

Uno de los problemas a los que nos enfrentamos es la escasez de eventos Pump and Dump en comparación con el conjunto total de datos.

Estas son las cifras si contamos las filas repetidas.

- BTC-CANN: 1512000 entradas, 247 eventos detectados. Proporción NP/P, 6121:1
- BTC-DOGE: 1766820 entradas, 49 eventos detectados. Proporción NP/P, 36058:1
- BTC-FTC: 1762500 entradas, 206 eventos detectados. Proporción NP/P, 8556:1

Si quisiésemos simular un entorno realista se debería cumplir esa proporción. No obstante, es inviable para este proyecto. En lugar de eso, para que sea factible el aprendizaje de los modelos con el número de datos que disponemos, la proporción entre los eventos NP/P se de establecerá en 10:1 aproximadamente, haciendo un submuestreo de la clase NP.

El siguiente paso es generar las instancias, para ello como hemos comentado, por cada instancia de la clase P, generamos aproximadamente 10 instancias de la clase NP. Las instancias NP se generan de forma aleatoria, pero han de cumplir tres condiciones:

- No puede haber 2 instancias iguales.
- Una instancia NP no puede estar en el intervalo entre el inicio de un evento Pump y una semana antes.

- Una instancia NP no puede estar en el intervalo entre el inicio de un evento Pump y un día despues.

En caso de no cumplir alguna de las tres, la instancia se descarta y se comprueba la siguiente.

Tras este proceso, obtenemos los siguientes tres datasets:

- BTC-CANN: 2000 instancias. 247 de clase P y 1753 de clase NP.
- BTC-DOGE: 554 instancias. 49 de clase P y 505 de clase NP.
- BTC-FTC: 1859 instancias. 205 de clase P y 1654 de clase NP.

En algunos casos no se conserva exactamente la proporción 10:1 debido al descarte de instancias por no cumplir alguna de las tres condiciones.

3.4.6. Outliers o valores atípicos

Un outlier, es una observación anormal y extrema en una muestra estadística o serie temporal de datos, que puede afectar potencialmente a la estimación de los parámetros del mismo.

Es decir, un outlier sería una observación dentro de una muestra o una serie temporal de datos que no es consistente con el resto.

Primeramente, asumimos que los datos cedidos por la plataforma Bittrex son buenos, es decir reflejan la realidad sin errores.

No obstante, al extraer las características, nos encontramos con valores para la primera y segunda derivada con tendencia a $\pm\infty$. A primera vista, los valores anómalos dentro de un rango controlado, es posible que aporten información sobre la existencia o no de un evento, ya que estos, se caracterizan por fuertes subidas en intervalos de tiempo pequeños. Sin embargo cuando nos encontramos con valores desmesurados, resulta lógico pensar que puede tratarse de un error numérico, ya que operamos con valores muy próximos a 0 en el denominador al llevar a cabo la normalización.

Para detectar y eliminar los outliers, planteamos dos métodos.

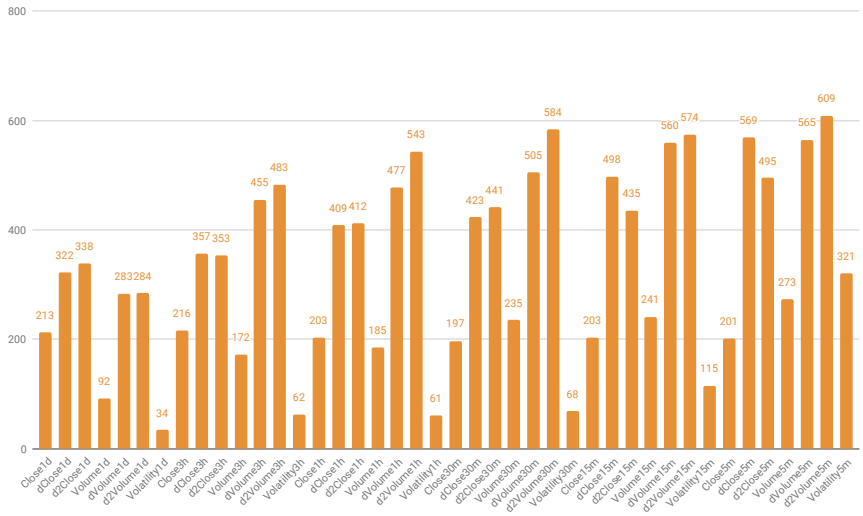

Figura 3.11: Resultados del método IQR

Rango Intercuartílico, IQR

Este método clasificará como outlier cualquier valor que se encuentre fuera de un intervalo de valores definido en base al rango intercuartilíco de los datos (IQR). El rango intercuartílico se define como $IRQ = Q_3 - Q_1$, siendo Q_3 y Q_1 el tercer y primer cuartil, respectivamente.

Calculamos los outliers aplicando la regla del rango intercuartílico, la cual considera un valor atípico a cualquiera que no este incluido en el rango
$[(Q_1 - 1{,}5 \cdot IQR), (Q_3 + 1{,}5 \cdot IQR)]$.

Es decir

$$x\ is\ outlier \longleftrightarrow x \notin [(Q_1 - 1{,}5 \cdot IQR), (Q_3 + 1{,}5 \cdot IQR)]$$

La Figura 3.11 muestra los resultados de la ejecución.

Dado que las criptomonedas sobre las que estamos ejecutando este método son significativamente volatiles, mediante este método obtenemos una cantidad demasiado grande de valores atípicos como para tratarlos. Concluimos que este método no es el más adecuado para su identificación.

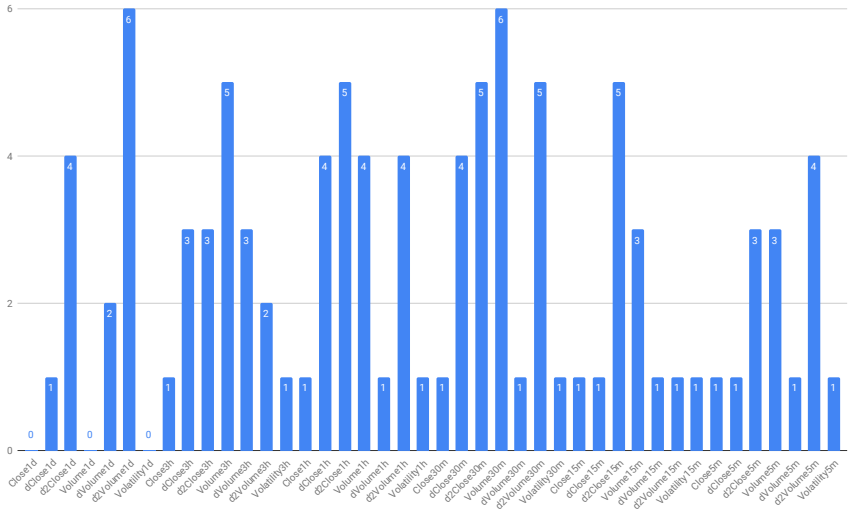

Figura 3.12: Resultados del método Z-Score

Método Z-Score

El Z-Score, es una forma de describir un punto del conjunto de datos en términos de su relación con la media y la desviación estándar.

El objetivo del Z-Score es eliminar los efectos de la ubicación y la escala de los datos, lo que permite comparar directamente diferentes conjuntos de datos. La intuición detrás del método de Z-Score de detección de valores atípicos es que, una vez que hemos centrado y reescalado los datos, cualquiera que se encuentre demasiado lejos de cero (el umbral estándar suele ser un Z-Score de 3 o -3) debe considerarse outlier.

En nuestro caso, dado que lo que más nos interesa son los datos atípicos, es decir los que se generan cuando ocurre un evento Pump and Dump (fuerte variación de los precios), utilizaremos un valor límite de Z muy grande (>8) para que únicamente detecte valores desmesurados debidos a la estandarización.

La Figura 3.12 muestra los resultados de la ejecución del método Z-Score.

En ambos casos (ver Figuras 3.11 y 3.12) se observa que a intervalos de tiempo menores aumenta el número de outliers. Ocurre lo mismo con la derivada y segunda derivada que tienen mayor número de outliers. Esto es debido a lo comentado en la Sección 3.4.6 sobre la normalización de los valores. Estos resultados son coherentes con lo esperado y dado que no son demasiados los datos detectados, comparándolos con la entrada, simplemente se eliminaran del conjunto de datos.

4. CAPÍTULO

Creación y Evaluación de los modelos

En este capítulo vamos a exponer y desarrollar el trabajo realizado en cuanto al análisis de los datos, describiendo brevemente los modelos utilizados y los criterios de selección, validación, optimización y las métricas utilizadas.

Recordamos que el objetivo del proyecto es predecir los eventos Pump and Dump. Para ello, usaremos los datos creados para entrenar modelos y evaluaremos su rendimiento. Este capítulo comienza con algunas ideas generales sobre los métodos a utilizar. A continuación se describe el marco experimental y finalmente se presentan y discuten los resultados.

4.1. Modelos predictivos

Comenzamos presentando los modelos predictivos que utilizaremos para realizar el análisis de los datos. Dado que no hay referencias sobre que tipo de modelo será el más adecuado para este problema, la experimentación se llevará a cabo usando una selección de algoritmos clásicos. Estos son: K-NN [4], Naive Bayes [9], Regresión Logística [7], Random Forest [2] y Redes Neuronales [6].

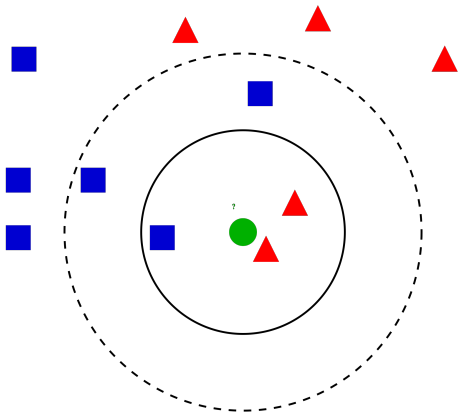

Figura 4.1: Representación esquemática del algoritmo K-NN

4.1.1. K Vecinos más cercanos

El método de los k vecinos más cercanos, en inglés k-nearest neighbours o K-NN es un método de clasificación supervisada que pertenece a la familia de algoritmos denominados vagos o lazy. El aprendizaje vago es un método de aprendizaje en el que la creación de un modelo se pospone hasta que se desea una predicción, al contrario que otros aprendizajes, donde el sistema intenta generalizar los datos de entrenamiento. En este caso, no se genera ningún modelo a partir de los datos de entrada, simplemente se comparan entre ellos.

Las reglas de clasificación por vecindad, están basadas en la búsqueda en un conjunto de prototipos de los k prototipos más cercanos al patrón a clasificar (ver Figura 4.1).

Se debe especificar una métrica para poder medir la proximidad entre instancias. Suele utilizarse por razones computacionales la distancia Euclídea, no obstante, existen otras distancias ampliamente utilizadas como la norma 1 o la norma ∞.

Las predicciones se realizan basándose en los ejemplos mas parecidos al que hay que predecir. El coste del aprendizaje es nulo, todo el coste pasa al cálculo de la predicción.

4.1.2. Naive Bayes

Los modelos de Naive Bayes, son una clase especial de clasificadores basados en redes Bayesianas, que son modelos probabilísticos que factorizan distribuciones de probabilidad conjunta, teniendo en cuenta las independencias condicionales entre variables. A la

hora de usar las Redes Bayesianas, como clasificadores, se hace uso de la regla de Bayes:

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})} \tag{4.1}$$

donde X son las variables predictoras y C la clase.

El modelo de Naive Bayes proporciona una manera fácil de construir funciones predictoras con un buen comportamiento debido a su simplicidad. La fórmula para calcular la probabilidad de la clase C condicionada por las variables predictoras $X_1, ..., X_n$ mediante el algoritmo de Naive Bayes es el siguiente.

$$P(C = c, X_1 = x_1, ..., X_n = x_n) \propto P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c) \tag{4.2}$$

Los parámetros del modelo son las probabilidades $P(C = 0)$ y $P(X_i = x_i)$, que se estiman de los datos.

4.1.3. Random Forest

Los árboles de clasificación se basan en una estructura en forma de árbol, donde las ramas representan conjuntos de decisiones, las cuales generan reglas para la clasificación de un conjunto de datos en subgrupos de datos. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Las ramificaciones se generan de forma recursiva hasta que se cumple un cierto criterio de parada. Es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía.

El Random Forest es un método de clasificación basado en los árboles de decisión. Es una modificación sustancial de bagging [1] que construye una larga colección de árboles no correlacionados y luego los promedia. La Figura 4.2 muestra un ejemplo conceptual del método de Random Forest.

4.1.4. Regresión Logística

La regresión logística puede considerarse un caso especial del análisis de regresión en donde la variable dependiente es dicotómica.

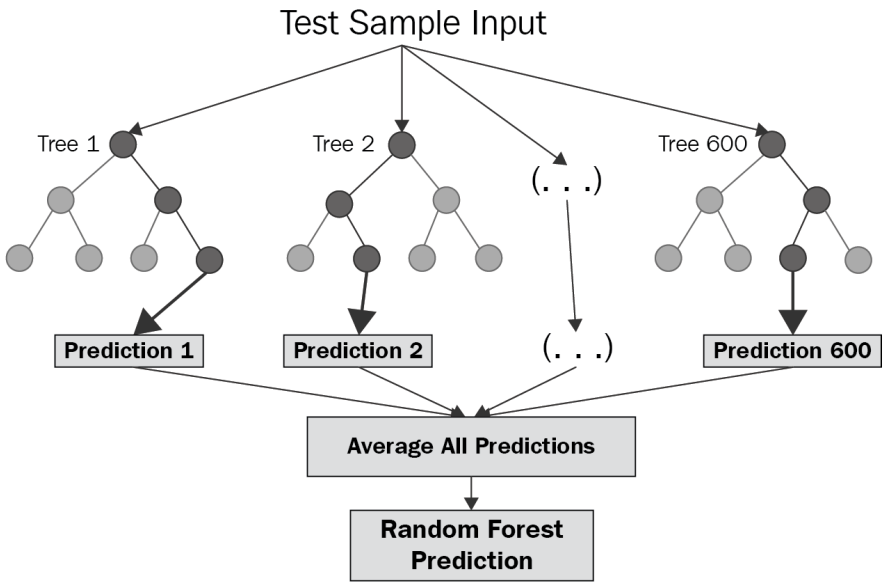

Figura 4.2: Ejemplo gráfico del método Random Forest

El objetivo de la regresión logística es encontrar el mejor modelo que describa la relación entre variables dicotómicas y el conjunto de variables predictoras. La regresión logística, genera los coeficientes de la formula para calcular una transformación logarítmica de la probabilidad de presencia de la característica de interés.

Suponiendo que hay únicamente dos clases, la regresión logistica reemplaza la variable objetivo original

$$Pr[1|a_1, a_2, ..., a_k]$$

que no se puede aproximar con precisión usando una función lineal, por

$$log(Pr[1|a_1, a_2, ..., a_k])/(1 - Pr[1|a_1, a_2, ..., a_k])$$

Los valores resultantes ya no están limitados al intervalo de 0 a 1 sino que puede estar entre $-\infty$ y $+\infty$.

La variable transformada se aproxima usando una función lineal como las generadas por la regresión lineal. El modelo resultante es el siguiente.

$$Pr[1|a_1, a_2, ..., a_k] = 1/(1 + exp(-w_0 - w_1 a_1 - ... - w_k a_k))$$

Igual que en la regresión lineal, se deben encontrar los pesos que se ajusten al conjunto de entrenamiento. La regresión lineal, mide la bondad del ajuste mediante el estimador error cuadrático medio. En la regresión logística, en su lugar se utiliza la función de verosimilitud del modelo. Esta viene dada por

$$\sum_{i=1}^{n}(i-x^{(i)})log(1-Pr[1|a_1^{(1)},...,a_n^{(n)}])+x^{(i)}log(Pr[1|a_1^{(1)},...,a_n^{(n)}])$$

donde $x^{(i)}$ es 0 o 1.

Los pesos w_i se deben elegir para maximizar la función de verosimilitud. Existen varios métodos para resolver este problema de maximización. Uno simple es resolviendo iterativamente una secuencia de problemas de regresión de mínimos cuadrados, hasta que la probabilidad logarítmica converge a un máximo, lo que generalmente ocurre en unas pocas iteraciones.

4.1.5. Redes Neuronales

Una red neuronal es un paradigma de aprendizaje y procesamiento automático inspirado en el funcionamiento del sistema nervioso humano.

Una red neuronal está compuesta por un conjunto de neuronas interconectadas entre sí mediante enlaces. Cada neurona toma como entradas las salidas de las neuronas de las capas antecesoras, cada una de esas entradas se multiplica por un peso, se agregan los resultados parciales y mediante una función de activación se calcula la salida. Esta salida es a su vez la entrada de la neurona a la que precede.

La unión de todas estas neuronas interconectadas es lo que compone una red neuronal artificial.

Una red neuronal normalmente esta formada por 3 capas.

- Capa de entrada: Constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.
- Capas ocultas: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.
- Capa de salida: Neuronas cuyos valores de salida se corresponden con las salidas de

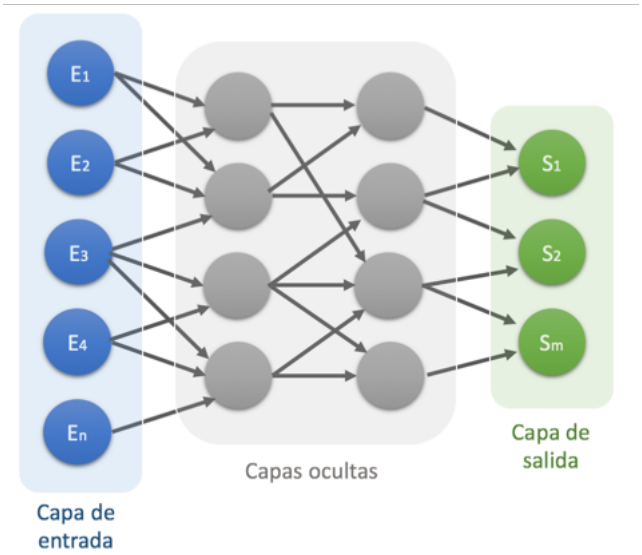

Figura 4.3: Ejemplo de una red neuronal artificial.

toda la red. En el caso de clasificación, la capa de salida contiene el valor asignado a la clase.

El número de capas y neuronas de la red se determina por el usuario. Su aprendizaje, consiste en ajustar los pesos para minimizar una cierta función de error en la capa de salida.

4.2. Métodos de evaluación de clasificadores

Para determinar que algoritmos se comportan mejor es fundamental llevar a cabo una evaluación correcta de su rendimiento. Esto implica dos aspectos básicos, las métricas y la forma de estimarlas. A continuación se presentarán las métricas y estimadores que se han utilizado en el proyecto.

Hemos seleccionado tres métodos de validación para simular dos escenarios concretos. Por una parte, entrenaremos el modelo y lo testearemos con el mismo conjunto de datos mediante los métodos de resustitución y validación cruzada, cuyo comparación nos dará una idea del sobreajuste. Por otra parte, para simular condiciones reales, separamos el conjunto de entrenamiento y de testeo. Entrenaremos el modelo con los datos correspondientes al intervalo 2014-2016 y lo testearemos con los datos de 2017.

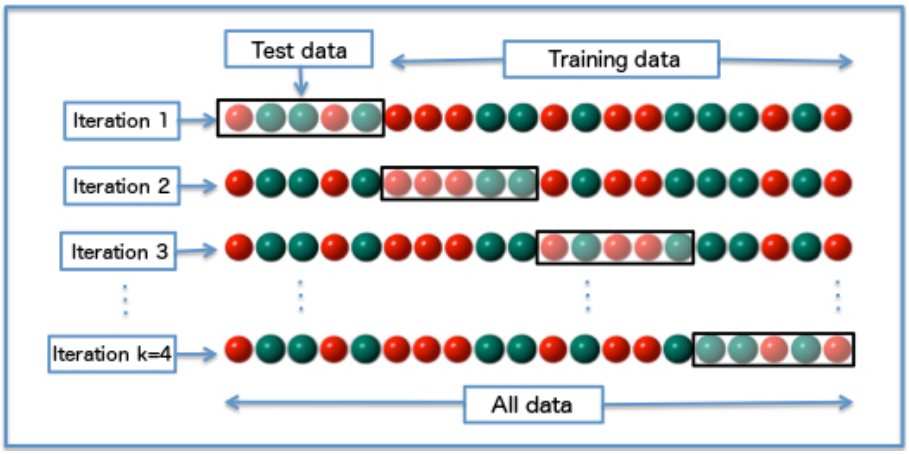

Figura 4.4: Ejemplo de validación cruzada k-CV

4.2.1. Resustitución

El método de resustitución consiste en entrenar el modelo con unos datos y estimar las métricas en esos mismos datos. Es un método no honesto que da lugar a una estimación optimista, pero un buen reflejo de la capacidad de ajuste del modelo. Es decir, un mal resultado en este estimador puede ser indicador de que el ajuste del modelo o los datos es subóptimo.

4.2.2. Validación cruzada

La validación cruzada o cross validation, es una técnica utilizada para la evaluación honesta de resultados de análisis estadísticos que garantiza, la independencia entre el subconjunto de datos de entrenamiento y el de prueba.

La independencia es garantizada debido a la forma en que se construyen cada uno de los subconjuntos de test y entrenamiento, impidiendo el solapamiento de datos.

El conjunto se divide en k subconjuntos, donde uno de ellos sera el conjunto de test y los k-1 restantes de entrenamiento. Esta operación se realiza k veces, cambiando el conjunto de test (y consecuentemente los de entrenamiento) en cada una de las iteraciones. La Figura 4.4 muestra un esquema del proceso.

La estimación de la métrica es la media aritmética de las estimaciones en cada una de las k iteraciones.

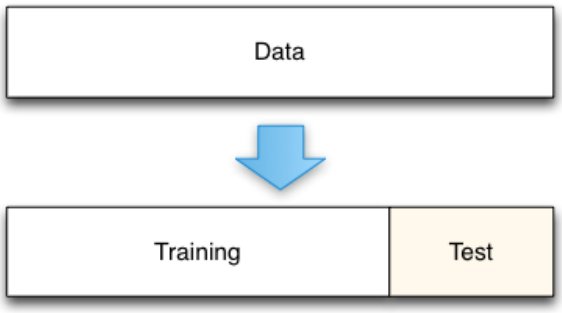

Figura 4.5: Representación visual de la técnica hold-out

4.2.3. Hold-out temporal

El hold-out es una técnica utilizada para la evaluación de resultados de análisis estadísticos. Consiste en dividir el conjunto de datos en 2 subconjuntos, el conjunto de entrenamiento y el conjunto de pruebas. El conjunto de entrenamientos, es el utilizado para entrenar el modelo, mientras que el conjunto de pruebas será el conjunto sobre el que se estimará la métrica. La Figura 4.5 muestra una representación visual de esta técnica.

El objetivo de un proyecto de estas características, consistiría en detectar eventos Pump and Dump en tiempo real a partir de datos históricos del pasado. En nuestro caso, para simular esto, dividimos el conjunto de datos en dos subconjuntos, el conjunto de entrenamiento con los datos de 2014-2016 y el conjunto de pruebas, con datos del año 2017.

De esta forma, obtenemos una estimación del comportamiento de nuestro algoritmo en un escenario real (en el cual tendríamos solo 3 años de datos).

4.3. Métricas de evaluación

A la hora de determinar el rendimiento de un clasificador se pueden utilizar diferentes métricas, siendo las más habituales la probabilidad de error de clasificación o su complementaria, el accuracy. Estas (y otras medidas) pueden ser engañosas cuando la variable clase no está balanceada. Esto es facil de ver si tomamos un problema donde el 90 % de las instancias son negativas, un clasificador que siempre asigne esa clase tendrá una tasa de acierto del 90 %.

Para evitar este problema se puede usar otras métricas como la medida F. Aunque estas métricas solventan este problema, su estimación esta ligada a un umbral de decisión pa-

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Figura 4.6: Representación gráfica de una matriz de confusión.

ra el clasificador, por lo que en ciertas situaciones no son apropiadas. En este proyecto utilizaremos el área bajo la curva ROC (AUC).

4.3.1. Matriz de confusión

En el campo de la inteligencia artificial, una matriz de confusión es una herramienta que permite la visualización de los resultados obtenidos en la ejecución de un algoritmo de clasificación supervisada.

La matriz contiene información de los aciertos y fallos de las predicciones en forma de matriz. En el caso de un problema binario, la tabla contiene estos valores (ver Figura 4.6).

- Verdaderos positivos (VP). Número de elementos clasificadas como clase positiva cuya clase real es positiva.
- Falsos positivos (FP). Número de elementos clasificados como clase positiva cuya clase real es negativa.
- Falsos negativos (FN). Número de elementos clasificados como clase negativa cuya clase real es positiva.
- Verdaderos negativos (VN). Número de elementos clasificados como clase negativa cuya clase real es negativa.

4.3.2. Accuracy

El accuracy o exactitud de un algoritmo de clasificación es el cociente entre el número de verdaderos positivos más verdaderos negativos y el número total de muestras.

$$Accuracy = \frac{VP + VN}{VP + FP + FN + VN} \tag{4.3}$$

4.3.3. Área bajo la curva ROC

Una curva ROC (Receiver Operating Characteristic) es una representación gráfica de la sensibilidad, frente a uno menos la especificidad, para un clasificador binario, según se varía el umbral de decisión.

La sensibilidad es el cociente entre los verdaderos positivos y verdaderos positivos más falsos negativos, es decir, el total de positivos reales. La medida nos da una idea de la capacidad del clasificador para identificar los positivos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

La especificidad es el cociente entre los verdaderos negativos y verdaderos negativos más falsos positivos, es decir, el total de negativos. Nos da una idea de la capacidad del clasificador para detectar los negativos.

$$Especificidad = \frac{VN}{VN + FP}$$

Para cada umbral de decisión, tendremos un par de valores de sensibilidad y uno menos la especificidad. El conjunto de todos estos pares constituyen la curva ROC. La Figura 4.7 muestra un ejemplo de curva ROC.

El área bajo esta curva, AUC, permite resumir en un número la bondad del clasificador. Cuando un clasificador clasifica aleatoriamente como positivo o negativo, su curva ROC se acerca a la diagonal y el área asociada es próxima a 0.5. Cuando un clasificador clasifica perfectamente las instancias, su AUC es de 1.0.

4.4. Selección de características

No tenemos una idea clara de que atributos pueden funcionar, pero parece razonable pensar que algunos de los propuestos no tengan valor predictivo. Por ello, se han utilizado

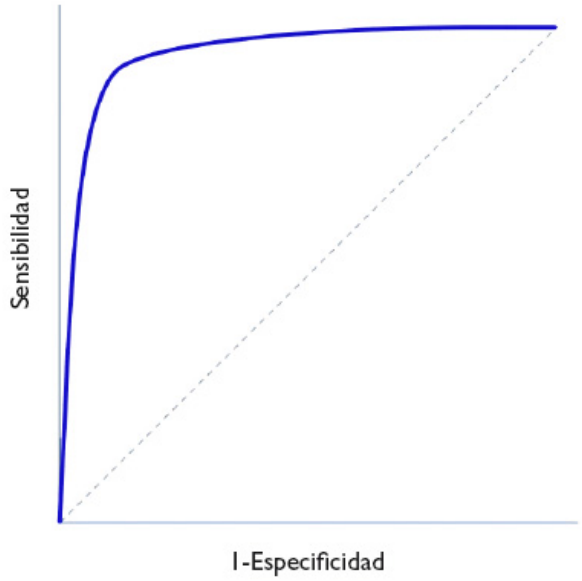

Figura 4.7: Ejemplo de una curva ROC

técnicas de selección de atributos para eliminar aquellas innecesarias para la predicción. Los modelos se han creado tanto con el conjunto original como con el reducido

4.4.1. Recursive feature elimination RFE

El algoritmo usado para la selección de características es Recursive Feature Elimination (RFE) [5]. Es un algoritmo que elimina de forma iterativa las características mas débiles para, crear un ranking ordenado de mayor a menor importancia. Comienza utilizando todas las características para la creación del modelo. En cada iteración calcula la característica más débil y la elimina. Este proceso se repite hasta llegar al número de características deseado, normalmente una, para que realice un ranking de todas ellas. Una vez finalizado, añadimos las características de una en una en el orden del ranking y volvemos a entrenar los modelos para calcular el número de características óptimo.

4.5. Diseño Experimental

En esta sección primeramente vamos a hacer un resumen de toda la información disponible. Recordamos que disponemos de una base de datos con los valores de tres cripto-

monedas BTC-CANN, BTC-DOGE y BTC-FTC, así como el archivo ALL-COINS que contiene los datos de las tres entremezclados.

Como se ha comentado en la sección anterior, los algoritmos que utilizaremos para la creación de los modelos son: K-NN, Naive Bayes, Random Forest, Regresión Logistica y Redes Neuronales. Todas ellas con los parámetros por defecto de la librería de Python Sklearn. El código fuente del proyecto se encuentra alojado en el repositorio de Bitbucket

La selección de características se realizará mediante el algoritmo comentado en la Sección 4.4.1, Recursive Feature Elimination. Recordamos que mediante este proyecto, solo se pretende estudiar la viabilidad del problema en cuestión, por lo tanto, reduciremos el problema de la selección a una única ejecución del algoritmo de RFE utilizando el modelo de Random Forest, ya que este es el que ha obtenido los mejores resultados en la clasificación. El subconjunto de características obtenido aplicando este método son las siguientes 13:

- Close1d
- Close3h
- Close1h
- Volatility1h
- Close30m
- dVolume30m
- Close15m
- dVolume15m
- Close5m
- dClose5m
- d2Close5m
- dVolume5m
- d2Volume5m

Para evaluar los resultados, usaremos tres métodos con los que podremos ver el comportamiento de los algoritmos para tres escenarios diferentes. En primer lugar, la resustitución, nos dará una estimación de como es el sobreajuste de los algoritmos. Como método honesto de evaluación de resultados, utilizaremos la validación cruzada 10-CV. Finalmente, mediante el método hold-out, evaluaremos los clasificadores utilizando los datos de 2014 a 2016 como entrenamiento y los de 2017 como test, para ver si hay independencia del tiempo a la hora de clasificar. Esto nos da una estimación de como se comportaría el clasificador en una situación real en la que entrena con datos del pasado, para predecir los eventos futuros.

Las métricas que se utilizará para evaluar el rendimiento de los algoritmos será principal-

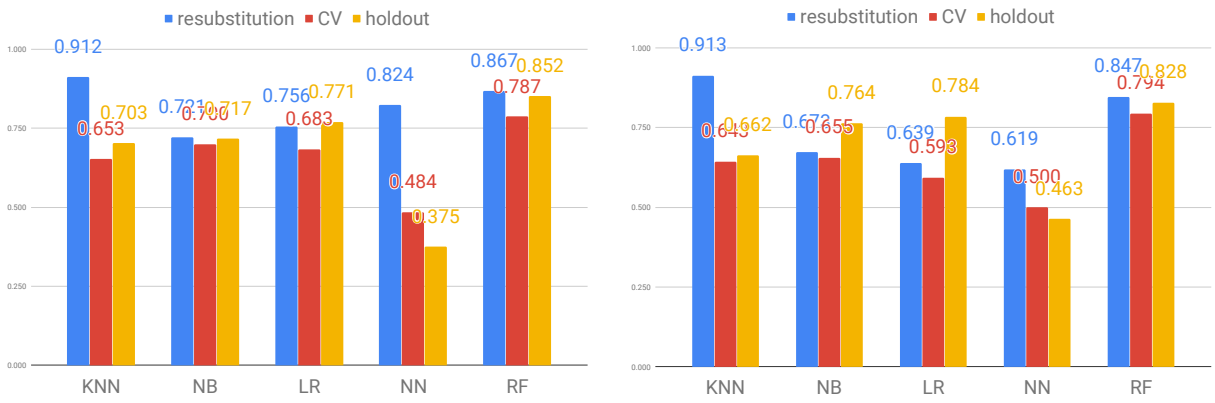

Figura 4.8: Barplot de los resultados de la ejecución para la criptomoneda BTC-CANN. A la izquierda los resultados sin selección de características, a la derecha con selección de características

mente el AUC, ya que como comentábamos en la Sección 4.3, debido al desbalanceo de las clases el accuracy es más confuso de interpretar.

4.6. Resultados y discusión

Los resultados de AUC obtenido para las criptomonedas BTC-CANN, BTC-DOGE, BTC-FTC y ALL-COINS se muestran en las Figuras 4.8, 4.9, 4.10 y 4.11 respectivamente. Cada figura cuenta con dos gráficos. El de la izquierda muestra los resultados obtenidos con todas las variables, mientras que el de la derecha muestra los resultados utilizando únicamente las 13 variables seleccionadas. En todos los casos se compara, para cada clasificador, el valor de AUC obtenido con resustitución, validación cruzada y hold-out. Los resultados en términos de accuracy pueden consultarse en los anexos de este documento.

En primer lugar, cabe recordar que este proyecto únicamente pretendía estudiar la viabilidad del problema. A la vista de los resultados obtenidos, podemos señalar que son cuanto menos prometedores. Se observa que la detección de estos eventos es factible, con unos resultados razonablemente buenos, lo que indica que es un proyecto viable y abre nuevas líneas de investigación. A continuación se da una descripción mas detallada de los resultados para cada clasificador y métodos.

Como cabía esperar, el **K-NN** se sobreajusta demasiado a los datos (buenos resultado en resustitución, mientras que en la validación cruzada apenas consigue diferenciarse de un clasificador aleatorio). En general, con las **redes neuronales** ocurre lo mismo, se sobreajustan demasiado (ver Figura 4.9).

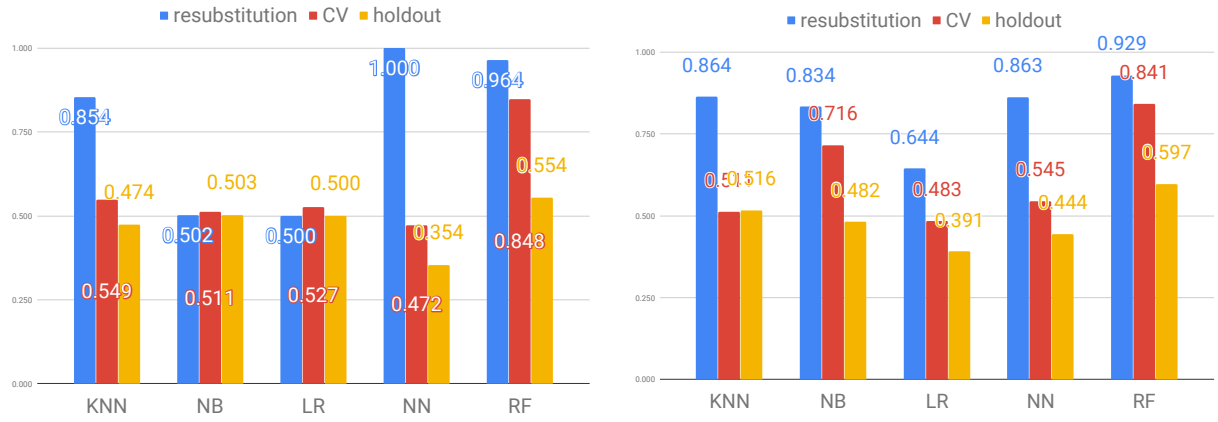

Figura 4.9: Barplot de los resultados de la ejecución para la criptomoneda BTC-DOGE. a la izquierda los resultados sin selección de características, a la derecha con selección de características

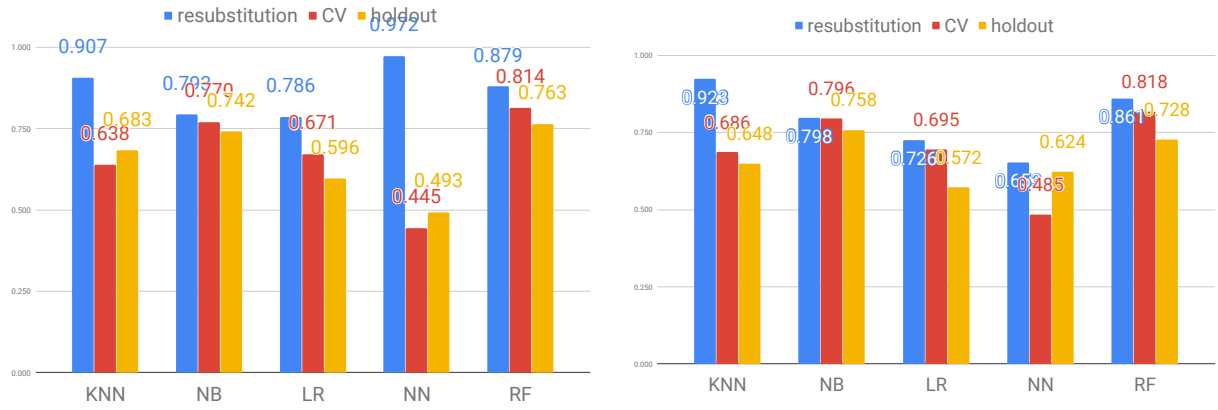

Figura 4.10: Barplot de los resultados de la ejecución para la criptomoneda BTC-FTC. A la izquierda los resultados sin selección de características, a la derecha con selección de características

Naive Bayes y **Regresión Logística** no muestran indicios de sobreajuste. No obstante, los pobres resultados obtenidos sugieren que tal vez su ajuste sea suboptimo. De todos los clasificadores, **Random Forest** es el que mejores resultados obtiene, tanto en la validación cruzada como en el hold-out.

Los resultados mediante el método de resustitución para la criptomoneda BTC-DOGE son especialmente buenos (ver Figura 4.9), esto puede deberse a que es la criptomoneda de la que menos datos disponemos (≈ 600) lo que facilita el sobreajuste. En cambio, para los conjuntos de datos mas grandes, el sobreajuste tiende a ser menor.

La selección de variables en la mayoría de los casos mejora ligeramente la clasificación, no obstante la diferencia es mínima e incluso en algunos empeora.

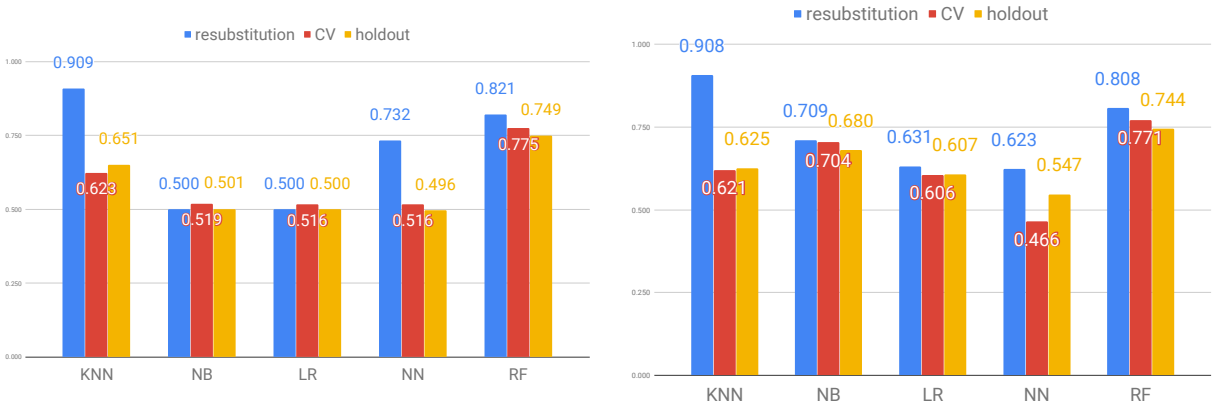

Figura 4.11: Barplot de los resultados de la ejecución para el conjunto de todas las criptomoneda ALL-COINS. A la izquierda sin selección de características, a la derecha con selección de características

Otro hecho destacable, es que los resultados cuando se combinan todas las monedas son similares a los de las monedas individuales. Esto sugiere que los patrones que permiten la clasificación son similares en las tres monedas. Con el objetivo de explorar esta idea en más profundidad hemos realizado otra ejecución utilizando como conjunto de entrenamiento los datos de las monedas BTC-CANN y BTC-FTC, y como test BTC-DOGE.

Como puede observarse en la Figura 4.12, los resultados han mejorado ligeramente (respecto a los de la moneda BTC-DOGE sola), lo que indica que por una parte al aumentar el conjunto de entrenamiento puede mejorar la predicción, y por otra, que se pueden combinar las monedas en el entrenamiento.

Globalmente, Random Forest es el mejor clasificador con unos resultados relativamente buenos, llegando a alcanzar un AUC de 0.852 para el caso de hold-Out (Figura 4.8). Era de esperar también que la mayoría de los clasificadores obtendrían peores resultados en el hold-Out. Especial mención a los casos de la criptomoneda BTC-DOGE (Figura 4.9).

Finalmente, indicar que como puede verse en la Figura 4.9, los resultados de hold-out para la criptomoneda BTC-DOGE son particularmente malos en comparación con la validación cruzada (ver Figura 4.9). Esto puede deberse a que el comportamiento de la moneda cambia significativamente a partir de 2016 tal y como puede verse en la Figura 4.13.

Se observa que desde 2014 hasta 2016 la tendencia es mayormente bajista, con algunos picos de subidas. No obstante, a partir de 2016, la tendencia cambia y comienza a ser alcista. Dado que para el hold-out, entrenamos el clasificador con los datos de 2014-2016 y los comprobamos con los de 2017, es razonable pensar que si entre ambos conjuntos las

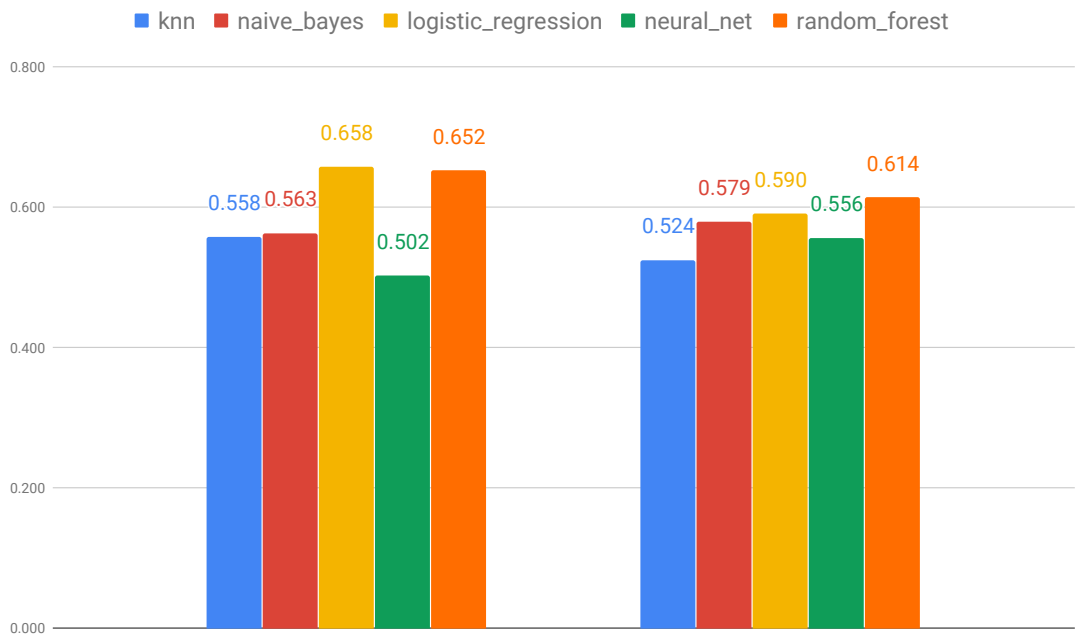

Figura 4.12: AUC para la moneda BTC-DOGE entrenada por los conjuntos de entrenamiento de BTC-CANN y BTC-FTC. A la izquierda con selección de características y a la derecha sin selección de características. Ambas evaluadas por el método hold-out.

tendencias son muy diferentes, los resultados de la predición no serán demasiado acertados.

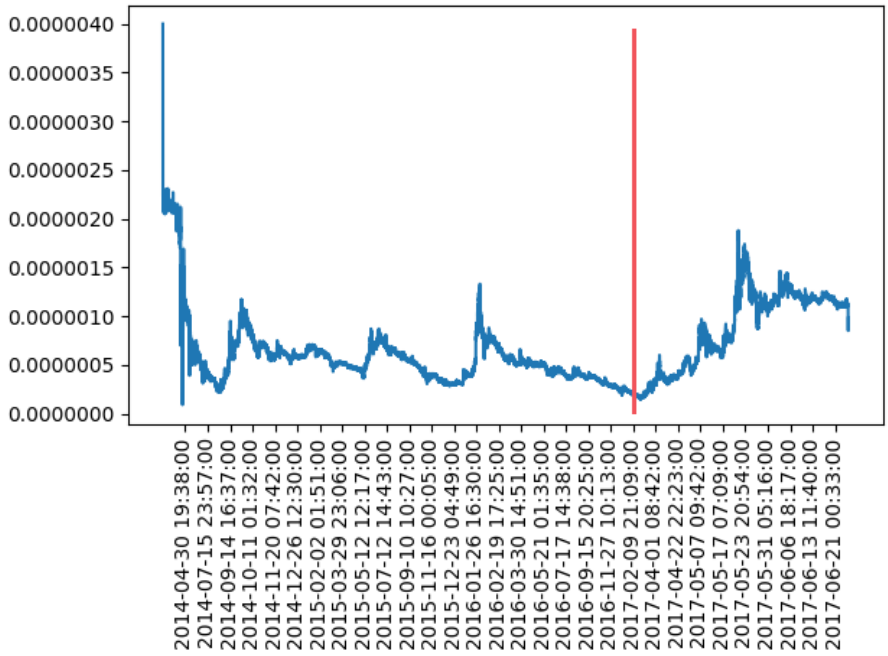

Figura 4.13: Evolución del precio de la criptomoneda BTC-DOGE a lo largo del tiempo (2014-2017)

5. CAPÍTULO

Conclusiónes y trabajo futuro

A la vista de los resultados y teniendo en cuenta los recursos de los que disponíamos podemos concluir que son unos resultados prometedores.

Como era de esperar, las estimaciones por validación cruzada obtienen mejores resultados ya que no se ven afectados por los cambios en la tendencia general. El caso mas interesante (hold-out) aun habiendo tenido los peores resultados (0.852 de AUC en el mejor de los casos), siguen siendo prometedor y sin duda mejorables con más dedicación y recursos.

El mercado de las criptodivisas, seguirá siendo durante un tiempo un mercado muy inestable debido a la falta de regulación, por lo que muy probablemente no faltarán investigaciones por esta línea.

A lo largo del proyecto hemos experimentado algunos contratiempos. Se ha invertido tiempo innecesario en la recogida y preprocesado de datos. Por una parte, era muy complicado conseguir un volumen adecuado de estos en tiempo real, debido a las limitaciones de las propias APIs (cantidad máxima de peticiones por minuto) y la duración del proyecto.

Un proyecto de estas características requiere de más tiempo y recursos para tener una cantidad y calidad de datos adecuada. En nuestro caso, gracias a que encontramos una base de datos grande el proyecto pudo llevarse a cabo, pero una parte importante de este debería ser su captura y acondicionamiento.

5.1. Propuestas de mejora y trabajo futuro

Basándonos en el trabajo y los análisis desarrollados a lo largo del documento, en este capítulo vamos a recopilar diferentes propuestas que consideramos de interés, así como futuras líneas de investigación.

5.1.1. Aprendizaje de datos en streaming

En este proyecto se ha utilizado un fichero de datos histórico, pero lo ideal sería recogerlos en tiempo real para ir aumentando la base de datos dinámicamente. Esto permitiría estar al día de las ultimas novedades del mercado y actualizar los modelos consecuentemente. Esta forma de operar podría solventar el problema del cambio de tendencia de las monedas.

Para ello proponemos utilizar la plataforma de intercambio Poloniex, por la facilidad que ofrece su API para obtener los datos de un gran número de criptomonedas en tiempo real. No obstante, también convendría realizar esta misma extracción para las distintas casas de cambio (Bitfinex, Binance, Coinbase...).

5.1.2. Mejora de los algoritmos

Dado que este proyecto es una primera aproximación al problema, los algoritmos que hemos utilizado tanto para la selección de variables, como para la construcción del modelo no están optimizados. Otra línea de mejora sería comprobar el funcionamiento de otros algoritmos de clasificación, así como optimizarlos adecuadamente.

5.1.3. Extracción de características

Para este proyecto hemos extraído una serie de características muy sencillas, sin embargo en el ámbito de las finanzas se utilizan indicadores muy específicos como los soportes y

las resistencias[1], Bandas de Bollinger[2], Indicador de Aroon[3], etc.

Recomendamos el cálculo y análisis de estas nuevas características más especificas para mercados bursátiles.

5.1.4. Mejoras de implementación

Otro aspecto a mejorar es la implementación del código añadiendo funciones de compra y venta automática de criptomonedas mediante la API de la plataforma. Por otra parte, para este proyecto la parte de software se ha hecho mediante scripts de ficheros. Proponemos para añadir más robustez al código utilizar un paradigma de programación orientado a objetos.

5.1.5. Grupos de Telegram

Los grupos de Telegram comentados en la Sección (1.2), proporcionan una información esencial para este problema. Cualquier línea de investigación futura debería trabajar con ellos. Nuestra propuesta es crear un cliente de Telegram, donde se recogen en tiempo real todos los mensajes de los grupos de especulación. En el momento de detectar un mensaje con orden de compra/venta ejecutarla automáticamente para adelantar al máximo número de compradores posibles [10].

5.1.6. Análisis de sentimientos

En el mercado de valores intervienen factores psicológicos y sociales, un análisis de las principales redes sociales sobre la opinión general de una criptomoneda concreta, o la opinión de expertos de renombre, podría dar como resultado un mayor poder predictivo.

[1]Un soporte es un nivel de precio por debajo del actual, se espera que la fuerza de compra supere a la de venta, por lo que un impulso bajista se verá frenado y por lo tanto el precio repuntará. Normalmente, un soporte corresponde a un mínimo alcanzado anteriormente.

Una resistencia es el concepto opuesto a un soporte. Es un precio por encima del actual, la fuerza de venta superará a la de compra, poniendo fin al impulso alcista, y por lo tanto el precio retrocederá. Las resistencias se identifican comúnmente en una gráfica como máximos anteriores alcanzados por la cotización

[2]Las bandas de Bollinger son dos curvas que envuelven el gráfico de precios. Se calcula a partir de una media móvil sobre el precio de cierre a la que envuelven dos bandas que se obtienen de añadir y sustraer al valor de la media K desviaciones estándar

[3]El indicador Aroon es un sistema que determina si el activo objeto de análisis está en tendencia o no y como de fuerte es esta tendencia.

Anexos

A. ANEXO

Resultados completos de la ejecución

coin	metric	validation	Selection	dummy	knn	naive_bayes	logistic_regressi	neural_net	random_forest
BTC-CANN	auc	resubstitution	FALSE	0,5	0,9123635483	0,7205767296	0,7563056518	0,8733933162	0,867426077
BTC-CANN	auc	resubstitution	TRUE	0,5	0,9131210958	0,6727829316	0,6392241968	0,6460126299	0,8468760478
BTC-CANN	auc	holdout	FALSE	0,5	0,7030612245	0,7165178571	0,7705994898	0,5433035714	0,8519770408
BTC-CANN	auc	holdout	TRUE	0,5	0,662372449	0,7638392857	0,7839923469	0,4753826531	0,828252551
BTC-CANN	auc	CV	FALSE	0,5	0,6527291347	0,7000161684	0,6828432589	0,5301323605	0,7868110402
BTC-CANN	auc	CV	TRUE	0,5	0,6433478278	0,6554511009	0,5925520895	0,5262572965	0,7938765607
BTC-CANN	accuracy	resubstitution	FALSE	0,8825865003	0,8927963698	0,1803743619	0,8933635848	0,9018718094	0,8854225752
BTC-CANN	accuracy	resubstitution	TRUE	0,8825865003	0,8973340896	0,7231990925	0,8876914351	0,867271696	0,889960295
BTC-CANN	accuracy	holdout	FALSE	0,8305084746	0,8262711864	0,2288135593	0,8601694915	0,8050847458	0,8305084746
BTC-CANN	accuracy	holdout	TRUE	0,8305084746	0,8347457627	0,7923728814	0,8516949153	0,8389830508	0,8601694915
BTC-CANN	accuracy	CV	FALSE	0,8825960269	0,8706928792	0,2660785274	0,8797163035	0,8428447428	0,8825960269
BTC-CANN	accuracy	CV	TRUE	0,8825960269	0,8723813376	0,6157974173	0,8814435028	0,8632708379	0,8837259704
BTC-DOGE	auc	resubstitution	FALSE	0,5	0,8541212019	0,501552795	0,5	0,9942084942	0,9644116166
BTC-DOGE	auc	resubstitution	TRUE	0,5	0,8644451905	0,8337250294	0,6443679705	0,7816854121	0,9291589726
BTC-DOGE	auc	holdout	FALSE	0,5	0,4744990893	0,5027322404	0,5	0,4817850638	0,5537340619
BTC-DOGE	auc	holdout	TRUE	0,5	0,5163934426	0,4817850638	0,3907103825	0,4676684882	0,5974499089
BTC-DOGE	auc	CV	FALSE	0,5	0,548922822	0,5114583333	0,5270833333	0,4771306818	0,8483664773
BTC-DOGE	auc	CV	TRUE	0,5	0,5113636364	0,7160511364	0,4831912879	0,5104876894	0,841110322
BTC-DOGE	accuracy	resubstitution	FALSE	0,8969359331	0,9052924791	0,1058495822	0,8969359331	0,9665738162	0,9108635097
BTC-DOGE	accuracy	resubstitution	TRUE	0,8969359331	0,8997214485	0,860724234	0,8997214485	0,9192200557	0,9136490251
BTC-DOGE	accuracy	holdout	FALSE	0,9384615385	0,9282051282	0,06666666667	0,9384615385	0,8102564103	0,9384615385
BTC-DOGE	accuracy	holdout	TRUE	0,9384615385	0,9333333333	0,7794871795	0,9179487179	0,8615384615	0,9282051282
BTC-DOGE	accuracy	CV	FALSE	0,8971085371	0,8914736165	0,1257486057	0,8913942514	0,8210767911	0,8971085371
BTC-DOGE	accuracy	CV	TRUE	0,8971085371	0,8913942514	0,858043758	0,8913148863	0,8606713857	0,8915529816
BTC-FTC	auc	resubstitution	FALSE	0,5	0,9071674354	0,7933456104	0,7855682644	0,8304921523	0,8792230538
BTC-FTC	auc	resubstitution	TRUE	0,5	0,9234711405	0,7981606943	0,7258893826	0,529383275	0,8607596006
BTC-FTC	auc	holdout	FALSE	0,5	0,6829389739	0,7415949649	0,5960404716	0,4886671447	0,7634839069
BTC-FTC	auc	holdout	TRUE	0,5	0,6484424793	0,7584847036	0,5719008923	0,6749123646	0,7279517208
BTC-FTC	auc	CV	FALSE	0,5	0,6382199484	0,7695891554	0,6710553485	0,467278664	0,8135471976
BTC-FTC	auc	CV	TRUE	0,5	0,6864263119	0,7961934288	0,6952951793	0,4896565945	0,8178617451
BTC-FTC	accuracy	resubstitution	FALSE	0,8781055901	0,8897515528	0,873447205	0,900621118	0,9588509317	0,8967391304
BTC-FTC	accuracy	resubstitution	TRUE	0,8781055901	0,902173913	0,8819875776	0,8967391304	0,8781055901	0,8944099379
BTC-FTC	accuracy	holdout	FALSE	0,9159369527	0,9054290718	0,8931698774	0,9089316988	0,8318739054	0,9036777583
BTC-FTC	accuracy	holdout	TRUE	0,9159369527	0,8984238179	0,8441330998	0,9124343257	0,877408056	0,9176882662
BTC-FTC	accuracy	CV	FALSE	0,878117453	0,8680094663	0,8680097458	0,8843309295	0,8245677736	0,8897757342
BTC-FTC	accuracy	CV	TRUE	0,878117453	0,87191553	0,8835619782	0,8913199724	0,8494527057	0,8928765094
ALL-COINS	auc	resubstitution	FALSE	0,5	0,9093745364	0,5001661682	0,5	0,8002405916	0,8206038576
ALL-COINS	auc	resubstitution	TRUE	0,5	0,9081206091	0,7093374904	0,6306185351	0,6553912659	0,8078888853
ALL-COINS	auc	holdout	FALSE	0,5	0,6508536585	0,5005543237	0,5	0,4854822616	0,7487416851
ALL-COINS	auc	holdout	TRUE	0,5	0,6250332594	0,680304878	0,6065022173	0,5232427938	0,7443847007
ALL-COINS	auc	CV	FALSE	0,5	0,6234762587	0,5189493355	0,5163039867	0,5328360776	0,7748290748
ALL-COINS	auc	CV	TRUE	0,5	0,6206202668	0,7040715123	0,6055936175	0,5853350978	0,7713878373
ALL-COINS	accuracy	resubstitution	FALSE	0,8824046921	0,8929618768	0,117888563	0,8824046921	0,8574780059	0,891202346
ALL-COINS	accuracy	resubstitution	TRUE	0,8824046921	0,8941348974	0,853372434	0,8882697947	0,864516129	0,8891495601
ALL-COINS	accuracy	holdout	FALSE	0,9001996008	0,878243513	0,1007984032	0,9001996008	0,83133732bio53	0,9081836327
ALL-COINS	accuracy	holdout	TRUE	0,9001996008	0,8852295409	0,8483033932	0,9061876248	0,8433133733	0,9081836327
ALL-COINS	accuracy	CV	FALSE	0,8824053478	0,8768412276	0,1917882006	0,8803525619	0,8328425925	0,8873915627
ALL-COINS	accuracy	CV	TRUE	0,8824053478	0,8750834168	0,8481064969	0,885927007	0,8568989909	0,8876813879

B. ANEXO

Resultados accuracy

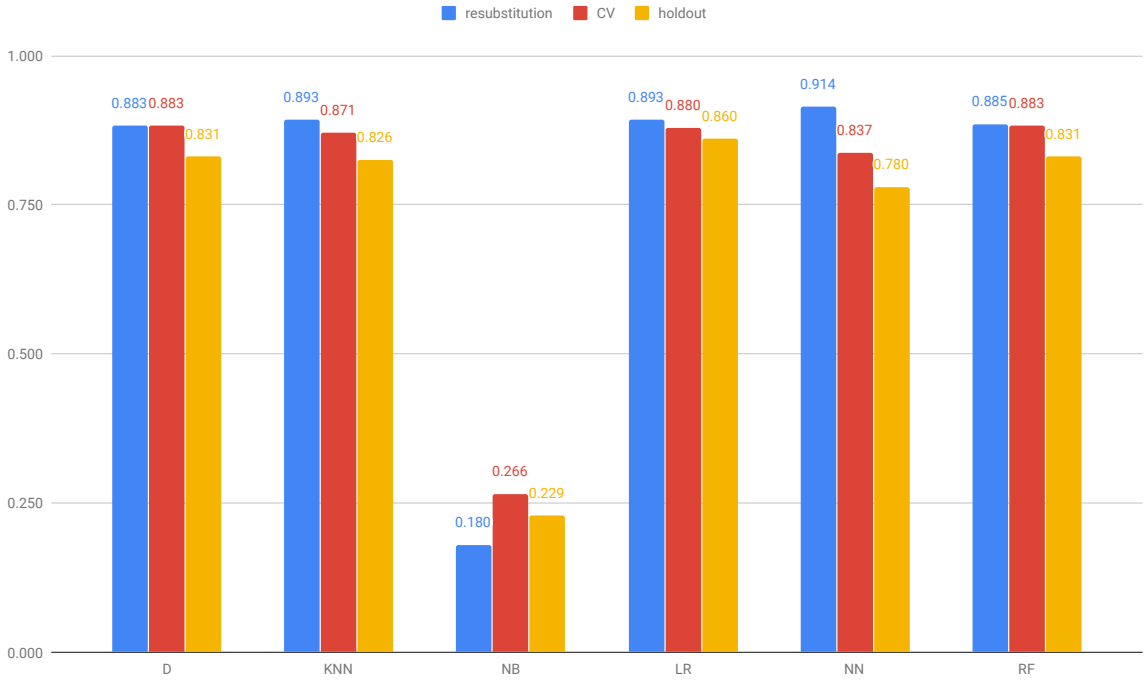

Figura B.1: Resultados BTC-CANN sin selección de características

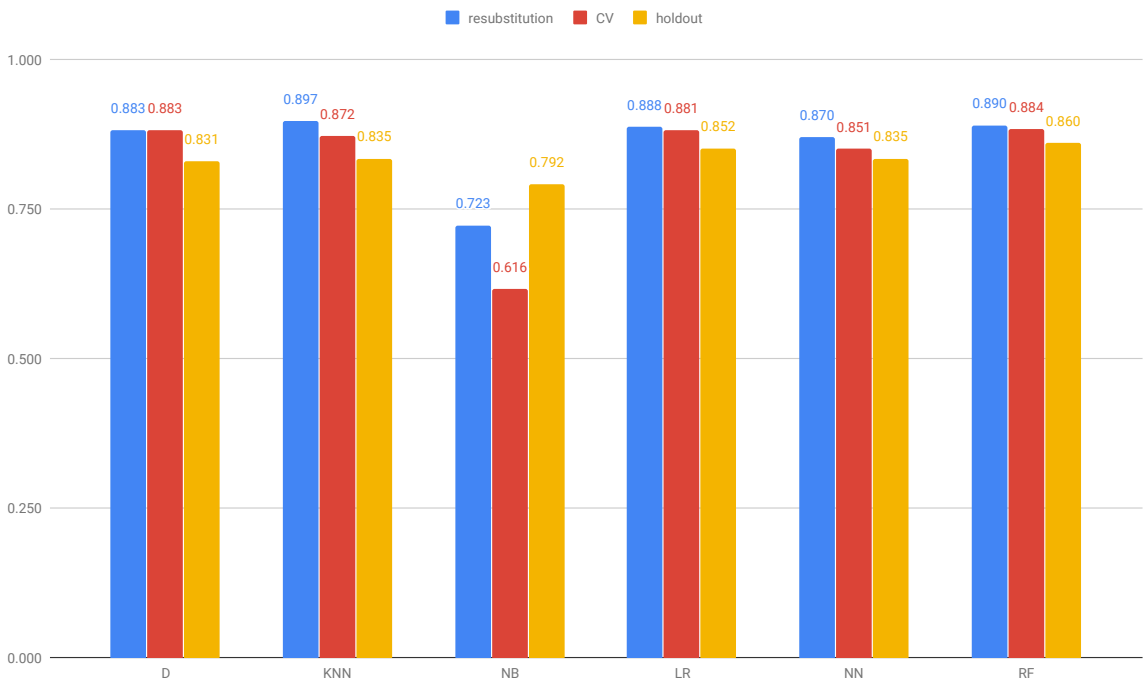

Figura B.2: Resultados BTC-CANN con selección de características

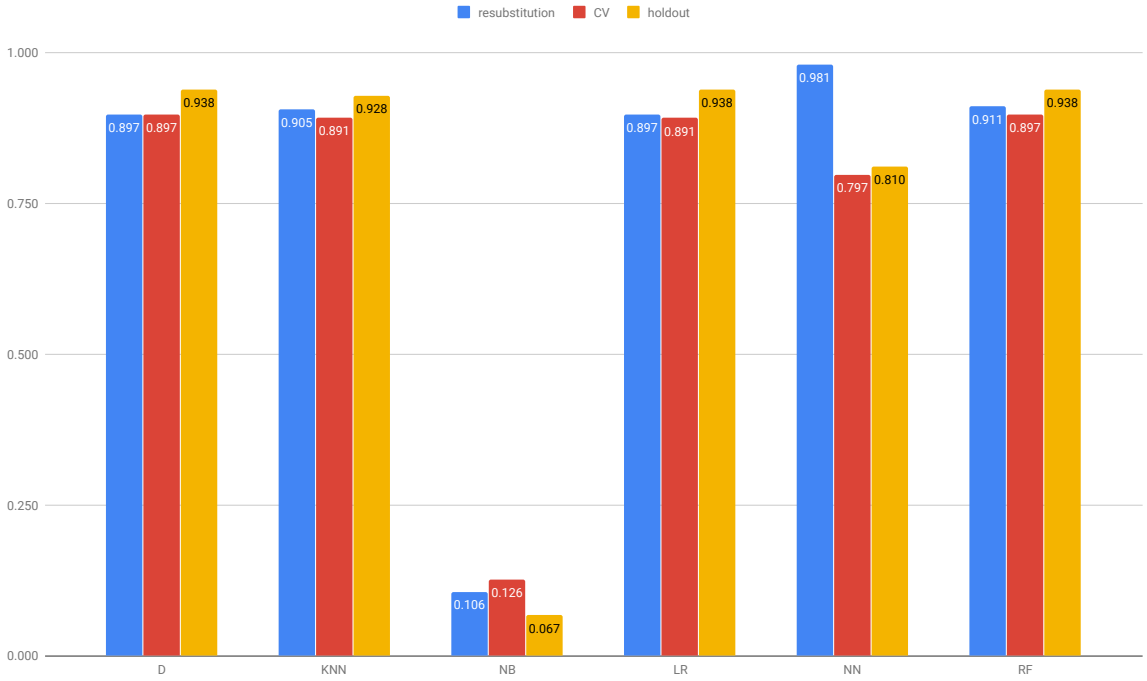

Figura B.3: Resultados BTC-DOGE sin selección de características

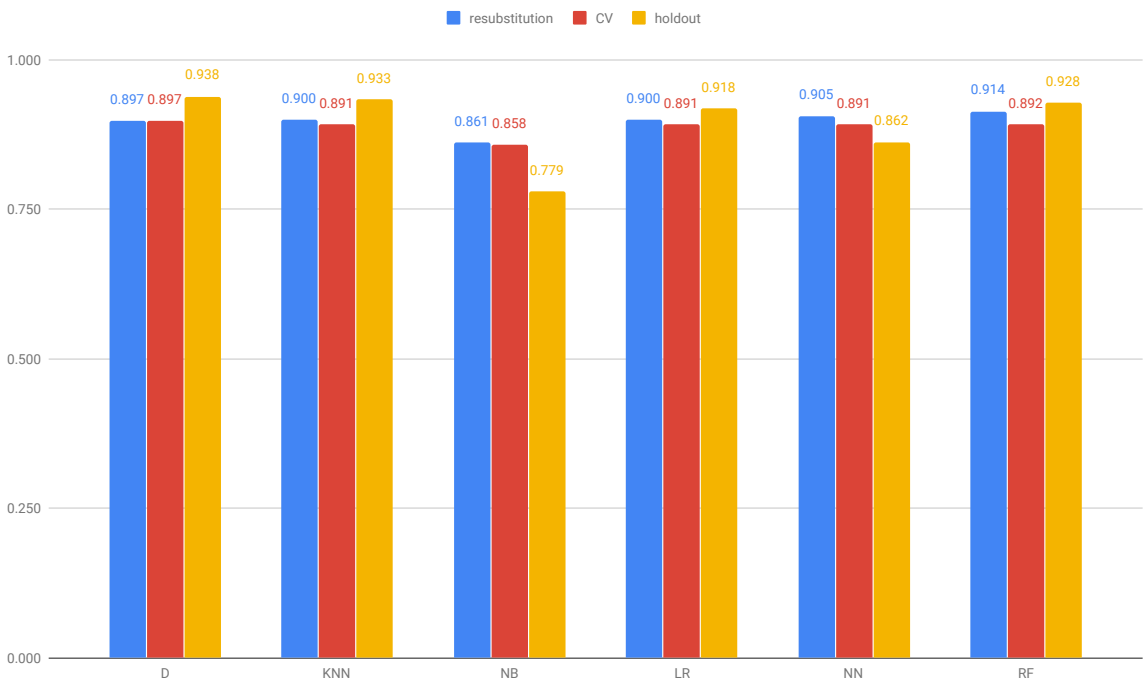

Figura B.4: Resultados BTC-DOGE con selección de características

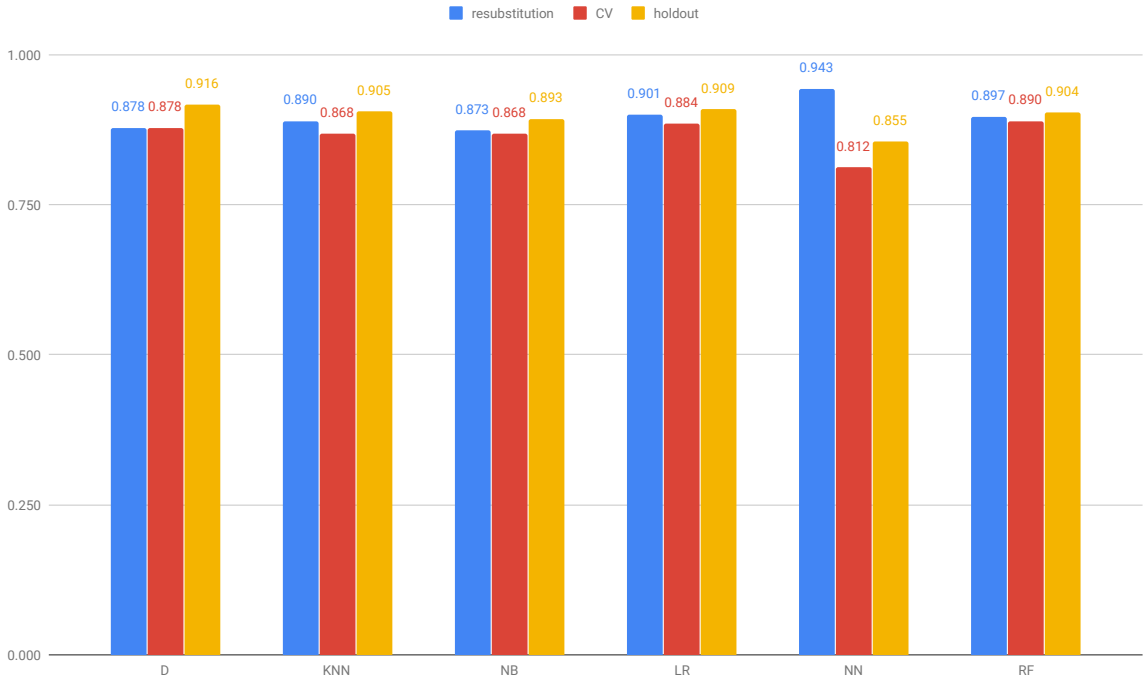

Figura B.5: Resultados BTC-FTC sin selección de características

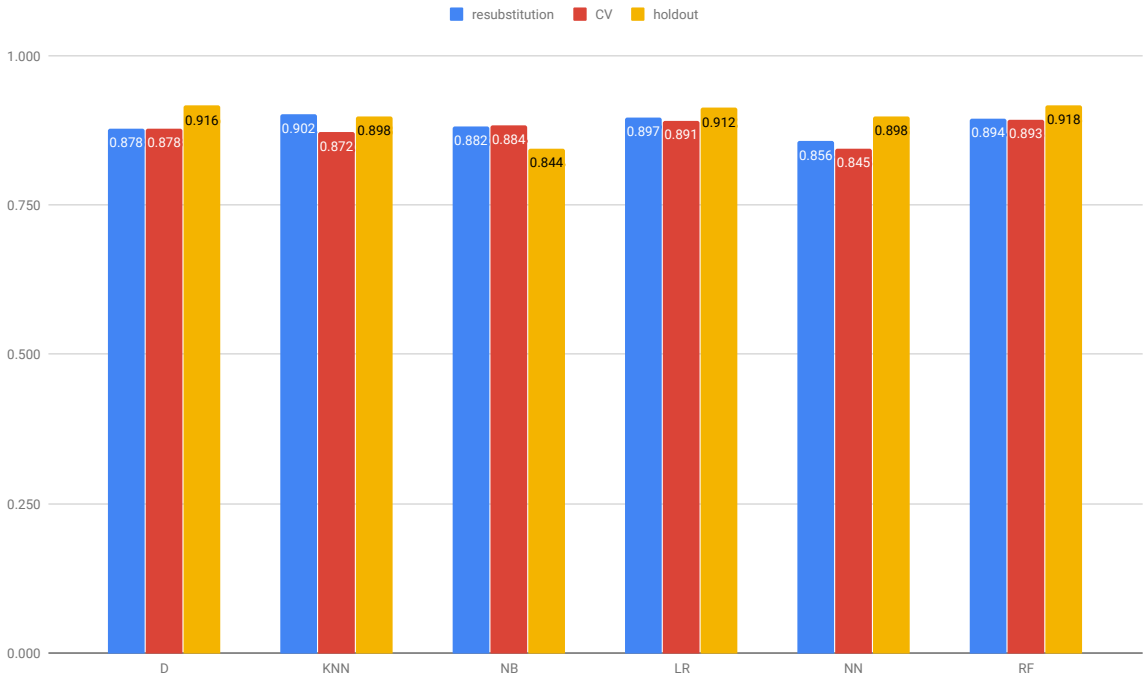

Figura B.6: Resultados BTC-FTC con selección de características

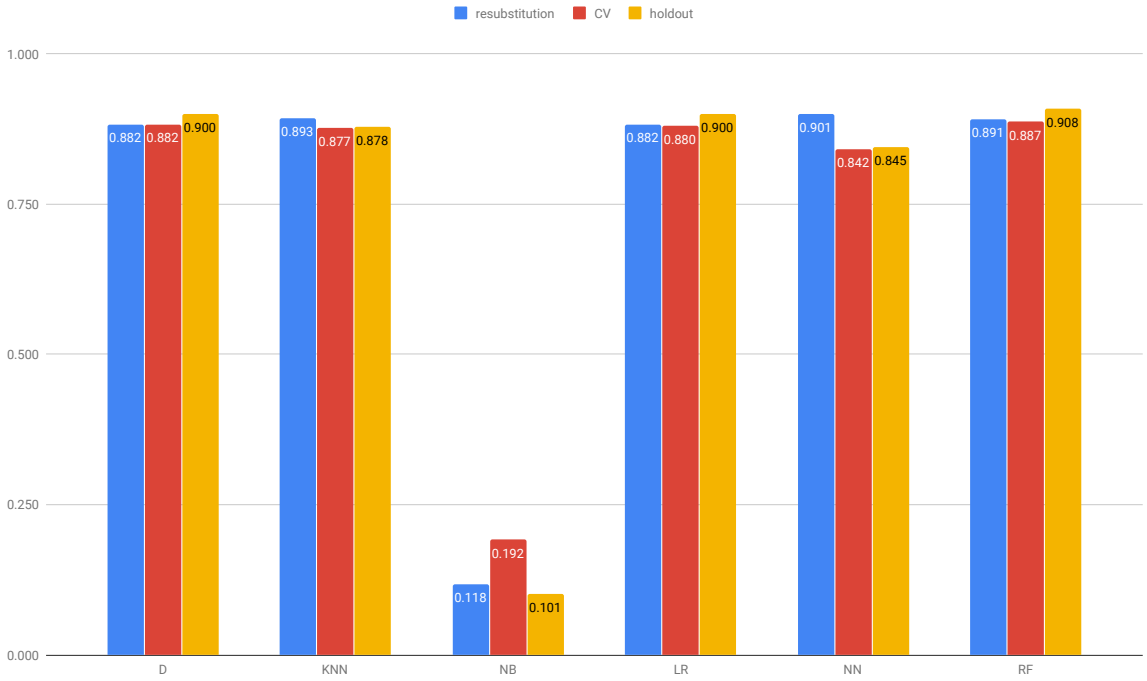

Figura B.7: Resultados ALL-COINS sin selección de características

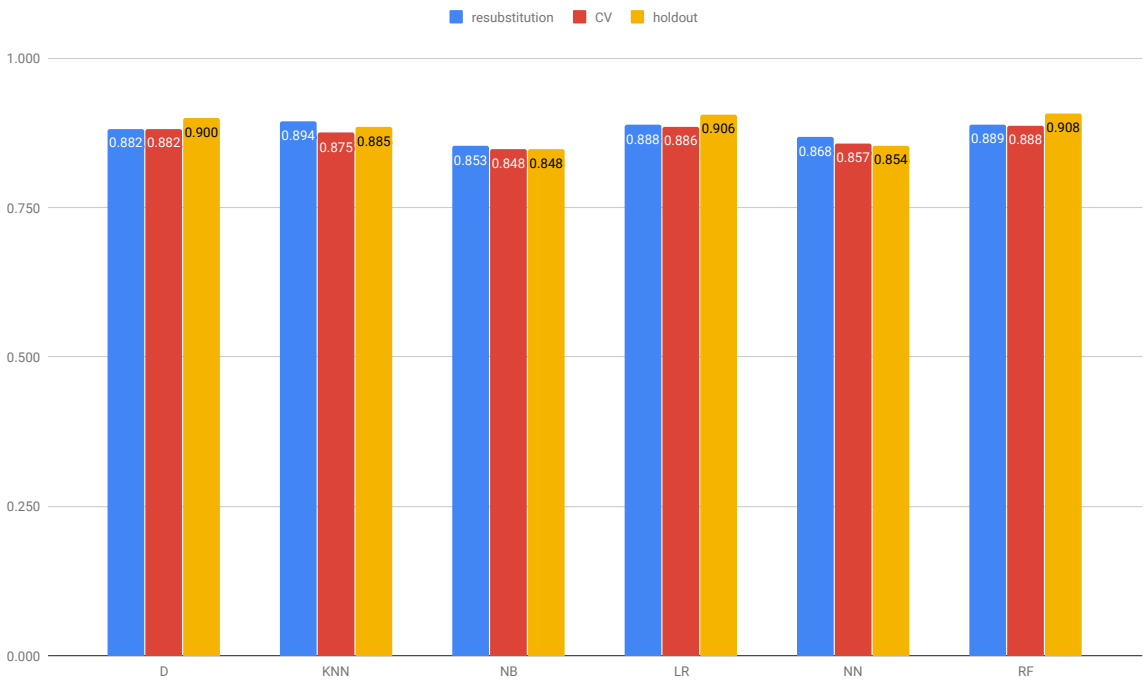

Figura B.8: Resultados ALL-COINS con selección de características

Bibliografía

[1] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[2] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] Markus K Brunnermeier, Joseph Abadi, Antonio Fatás, Beatrice Weder di Mauro, Ousmène Jacques Mandeng, Piroska Nagy-Mohacsi, Neil Gandal, JT Hamrick, Tyler Moore, and Tali Oberman. The economics of cryptocurrency pump and dump schemes.

[4] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[5] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[6] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):993–1001, 1990.

[7] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[8] Tao Li, Donghwa Shin, and Baolian Wang. Cryptocurrency pump-and-dump schemes. *Available at SSRN*, 2018.

[9] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[10] Mehrnoosh Mirtaheri, Sami Abu-El-Haija, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Identifying and analyzing cryptocurrency manipulations in social media. *arXiv preprint arXiv:1902.03110*, 2019.

[11] Andreas Isnes Nilsen. Limelight: Real-time detection of pump-and-dump events on cryptocurrency exchanges using deep learning. Master's thesis, UiT Norges arktiske universitet, 2019.

[12] Jiahua Xu and Benjamin Livshits. The anatomy of a cryptocurrency pump-and-dump scheme. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1609–1625, 2019.