Máster Universitario en

Ingeniería **C**omputacional y **S**istemas **I**nteligentes

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Konputazio Zientziak eta Adimen Artifiziala Saila –
Departamento de Ciencias de la Computación e Inteligencia Artificial

K I S A I C S

Tesis de Máster

Lisset Alexandra Neyra Romero

CATEGORIZACIÓN AUTOMÁTICA DE RESPUESTAS APLICANDO ALGORITMOS DE CLASIFICACIÓN SUPERVISADA AL ANÁLISIS DE LAS CONTESTACIONES DE ESTUDIANTES A UNA SERIE DE PREGUNTAS TIPO TEST

Tutores

Basilio Sierra Araujo
Departamento de Ciencia de la Computación e Inteligencia Artificial
Facultad de Informática, UPV/EHU

Samanta Cueva Carrión
Departamento de Ciencias de la Computación y Electrónica
Universidad Técnica Particular de Loja
Ecuador

informatika fakultatea
facultad de informática

KZAA

Septiembre 2016

CATEGORIZACIÓN DE RESPUESTAS APLICANDO ALGORITMOS DE CLASIFICACIÓN SUPERVISADA AL ANÁLISIS DE LAS RESPUESTAS CONTESTADAS POR ESTUDIANTES A UNA SERIE DE PREGUNTAS TIPO TEST

RESUMEN

Durante los últimos años se ha evidenciado el creciente interés por el aprendizaje automático para la clasificación y categorización de documentos, textos, preguntas. Esto permite automatizar procesos que si se hicieran con la intervención del ser humano podrían tener un alto costo en tiempo, y abre las puertas para su implementación con sistemas incluyentes para estudiantes con discapacidades físicas.

En este artículo se describe un trabajo de investigación que utiliza técnicas de minería de datos para obtener clasificadores que permitan identificar automáticamente las respuestas correctas expresadas por los estudiantes y éstas son asociadas a una pregunta con distintas opciones que son parte del proceso de evaluación de los conocimientos adquiridos de los estudiantes durante su proceso formativo.

Atendiendo a estas consideraciones se utilizó un corpus con preguntas de diferentes categorías, donde cada pregunta tenía múltiples opciones factibles para ser seleccionadas; sin embargo a cada pregunta le pertenecía una sola respuesta correcta. Se transcribió las respuestas dadas por los estudiantes de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja teniendo un total de 12960 transcripciones de las respuestas verbales que se obtuvieron de los estudiantes en español. Los resultados obtenidos mediante diferentes algoritmos de clasificación son presentados, analizados y comparados.

Palabras clave: Clasificación supervisada, categorización de preguntas, Minería de datos, algoritmos

INTRODUCCIÓN

En la educación superior, la evaluación es un proceso principal para la formación de los estudiantes. En términos amplios, la evaluación educativa es un proceso a través del cual se valora el mérito de un objeto determinado en el campo de la educación, con el fin de tomar decisiones particulares (Méndez Martínez & Ruiz Méndez, 2015) .

La educación a distancia (EaD) es una forma de enseñar y aprender basada en "un diálogo didáctico mediado entre el profesor (institución) y el estudiante que, ubicado en espacio diferente al de aquél, aprende de forma independiente (cooperativa)". (García Aretio, 2002) (Universidad Técnica Particular de Loja, 2014)

La Universidad Técnica Particular de Loja dispone de dos modalidades de estudios: presencial y Abierta y a Distancia (**MAD**), siendo esta última con mayor impacto y acogida en el Ecuador. La Modalidad Abierta y a Distancia de la UTPL es un sistema de estudios superiores que pretende llegar a todos los rincones del país. Consiste en un diálogo didáctico entre el profesor y el estudiante el cual aprende de una forma independiente y colaborativa (Universidad Técnica Particular de Loja, 2012b). Esta modalidad de estudios se oferta en centros universitarios del Ecuador y en los centros internacionales de New York, Roma, Madrid y Bolivia.

En la Universidad Técnica Particular de Loja (UTPL), la evaluación es un proceso continuo y un elemento central en el proceso enseñanza-aprendizaje a través del cual el alumno recibe retroalimentación de su progreso. Del mismo modo permite valorar la eficacia de las técnicas didácticas empleadas, la capacidad científica y pedagógica del educador, la calidad de los materiales didácticos, entre otros (Universidad Técnica Particular de Loja, 2012a).

El modelo educativo de la Universidad Técnica Particular de Loja (UTPL) para la Modalidad Abierta y a Distancia contempla tres tipos de evaluación: autoevaluación, heteroevaluación y coevaluación. Según la UTPL la heteroevaluación es de carácter permanente, en ella el docente evalúa permanentemente al estudiante para conocer su trabajo, actuación, rendimiento y estilos de aprendizaje. Se realiza mediante las evaluaciones a distancia, las actividades en línea y la evaluación presencial (Universidad Técnica Particular de Loja, 2012a).

La evaluación presencial se realiza mediante un sistema móvil de preguntas randómicas que el estudiante debe responder, sin embargo para los estudiantes con discapacidades especiales esta actividad no puede ser llevada a cabo de forma individual por lo que deben ser acompañados y asistidos en la evaluación por un pariente, amigo o conocido.

La importancia de este trabajo radica en el principio de que "toda persona tiene derecho a aprender y a participar en los planes de educación y currículo normalizado" (Aitken, Pedejo Fairley, & Carlson, 2012). Al respecto en el Capítulo VII aprendizaje de personas con discapacidad del artículo 49 del reglamento de Régimen Académico del Ecuador (Consejo de Educación Superior, 2016), se da la siguiente disposición para las Instituciones de Educación Superior (IES). " En cada carrera o programa, las lES deberán garantizar a las personas con discapacidad ambientes de aprendizaje apropiados que permitan su acceso, permanencia y titulación dentro del proceso educativo, propiciando los resultados de aprendizaje definidos en la respectiva carrera o programa. Como parte de los recursos de aprendizaje, las lES deberán asegurar a las personas con discapacidad, la accesibilidad a sistemas y tecnologías de información y comunicación (TIC) adaptados a sus necesidades."

Por lo antes mencionado el presente trabajo abarca un primer alcance donde las respuestas de los alumnos a cada pregunta es transformada a texto para su posterior clasificación y a través de la aplicación de algoritmos de clasificación supervisada, se determina si la respuesta es correcta o no,

de esta forma se pretende automatizar el proceso de evaluación a través del sistema de evaluación en línea de la Universidad Técnica Particular de Loja.

El problema de resolver la categorización de texto se remonta a la década de los 80's, donde lo que se usaba era un motor de conocimiento y un experto basado en un conjunto de reglas de la forma if X then Y; esta técnica sin embargo, no resultó exitosa debido a que el conjunto de reglas requería ser actualizada constantemente de forma manual, lo que era totalmente impráctico. No fue hasta los 90's cuando se comenzó a utilizar un enfoque de Aprendizaje Automático (Bishop, 2006) para la automatización de esta tarea, donde a través de un conjunto de datos el algoritmo es capaz de aprender un modelo que le permite categorizar los documentos. Entre algunas de las ventajas que ofrece este enfoque se encuentra su alto grado de automatización, puesto que el mismo algoritmo se encarga del entrenamiento del clasificador. Lo único indispensable es enviarle la cantidad de datos necesarios para que se construya dicho modelo clasificador. Además, los resultados que se obtienen en términos de exactitud son comparables con los de los expertos humanos. Es por estas razones que el uso del Aprendizaje Automático se ha ido extendiendo. Una muestra clara puede observarse al usar las diferentes aplicaciones de búsqueda que ofrece la compañía Google (Varguez Moo, Brito Loeza, & Uc Cetina, 2012).

1. ESTADO DEL ARTE

Esta sección tiene como objetivo presentar la base teórica sobre la que se fundamenta el desarrollo del trabajo: la clasificación de textos, aprendizaje automático, minería de datos y los principales algoritmos de clasificación.

1.1. Clasificación de textos

La clasificación automática de textos, también conocida como categorización de textos, es la tarea de asignar un documento dentro de un grupo de clases o categorías predefinidas (Sebastiani, Machine Learning in Automated Text Categorization, 2002).

En la clasificación automática de textos, es necesaria la presencia de un conjunto de categorías o clases $C = \{c_1, ..., c_{|C|}\}$ y un corpus inicial $D = \{d_1, ..., d_{|D|}\}$, el cual contiene una colección de documentos etiquetados con algunos de los valores del conjunto C. A través de un proceso inductivo, el clasificador aprende las características de cada una de las categorías del conjunto de entrenamiento $Dt = \{d_1, ..., d_{|Dt|}\}$. Por lo tanto, la clasificación de textos, puede ser formalizada como la tarea de aprender una función objetivo $F : Dt \rightarrow C$, llamado clasificador (Sebastiani, Text categorization, 2005) (Ramírez de la Rosa, 2010). El desempeño del clasificador se mide evaluando la función F en un conjunto de prueba $D_p = D - Dt$ (López Condori, 2014).

Se puede simplificar en dos etapas el proceso de clasificación de textos, la primera etapa abarca el entrenamiento de un conjunto de documentos pre-clasificados de cada clase o categoría, de esta forma se pretende conseguir que el clasificador generalice el modelo que ha aprendido con los

documentos precalificados. La segunda etapa se usa el modelo obtenido de la primera etapa para clasificar nuevos documentos.

1.2. Aprendizaje automático

Las técnicas de aprendizaje computacional se han utilizado frecuentemente para resolver problemas donde se manejan grandes cantidades de información y es necesario encontrar un patrón que permita determinar el comportamiento dicha información. El objetivo del aprendizaje computacional es desarrollar modelos que sean capaces de aprender de la experiencia previa de los eventos que se presentan, a partir de conjuntos de datos (Sierra Araujo, 2006). La finalidad de los modelos es extraer información implícita dentro de los datos para poder hacer predicciones y tomar decisiones sobre nuevos datos (Guzmán Cabrera, 2009).

Una definición comúnmente utilizada es la siguiente: "El aprendizaje automático es el estudio de algoritmos computacionales que van mejorando automáticamente su desempeño a través de la experiencia" (Mitchell, 1997). De manera más formal, esta definición se puede enunciar de la siguiente manera: Un programa de computadora aprende a partir de una experiencia E al realizar una tarea T, si su rendimiento al realizar T, medido con P, mejora gracias a la experiencia E (Guzmán Cabrera, 2009).

Los algoritmos de aprendizaje automático más utilizados son los algoritmos de clasificación supervisados, estos consisten en asignar a un objeto (persona, documento, fenómeno físico, etc) diversas categorías o clases previamente especificadas. (Ramírez de la Rosa, 2010)

Se puede deducir que con el aprendizaje automático se puede automatizar el tiempo del proceso de evaluación que tradicionalmente se ha venido realizando de forma manual, mejorando la operatividad y tiempo de respuesta hacia los estudiantes.

1.3. Minería de Datos

La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. (Pérez López & Santín González, 2008). En resumen, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos.

Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos. (Pérez López & Santín González, 2008) . Por lo tanto la minería de datos engloba un conjunto de técnicas enfocadas en la extracción de conocimiento que se encuentra alojado en los datos de forma implícita.

En este trabajo se empleará un proceso típico de minería de datos que consta de los pasos generales mostrados en la Figura 1.

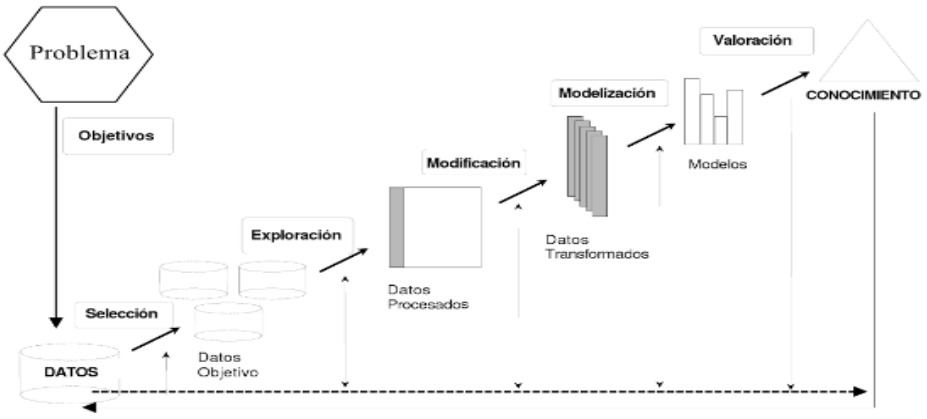


Figura 1. Proceso de minería de datos según SAS Institute *(Pérez López & Santín González, 2008)*

En la práctica, los modelos para extraer patrones pueden ser de dos tipos: predictivos y descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo o dependientes, usando otras variables o campos de la base de datos, llamadas variables independientes o predictivas. Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2007).

En la Figura 2, se muestran los principales métodos de minería de datos:

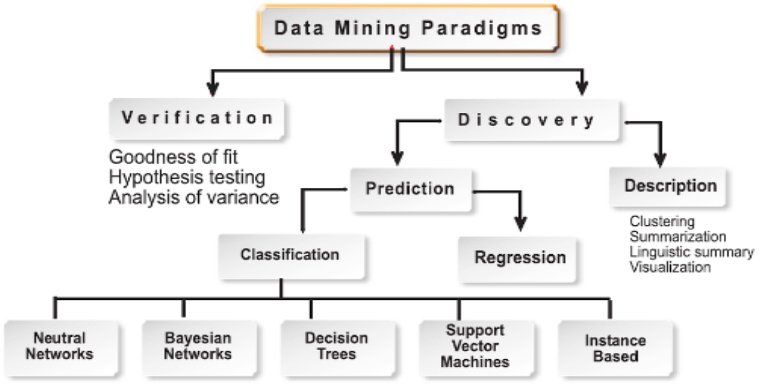


Figura 2. Taxonomy of data mining methods *(Rokach & Maimon, 2014)*

Entre las técnicas más usadas están: redes neuronales, arboles de decisión y agrupamiento o clustering, los cuales se describen más adelante detalladamente.

1.4. Clasificación Supervisada.

La clasificación supervisada se podría definir como la asignación de una o varias categorías predefinidas sobre un grupo de respuestas (instancias) a clasificar. En general se puede dividir en las siguientes fases:

Representación: el conjunto de respuestas que componen los datos a clasificar deben ser transformadas a texto de acuerdo al sistema de clasificación a utilizar.

Clasificación: El proceso de clasificación se lo puede realizar de dos formas, que se detallan a continuación:

- **Entrenamiento**: permite obtener la descripción de las categorías con ayuda de la colección de respuestas previamente clasificadas. Lo que se denomina "modelo".
- **Test**: una vez obtenido el modelo, se podrá predecir las categorías de las respuestas por clasificar.

Evaluación: Una vez realizada la clasificación se procede con el análisis de las clasificaciones obtenidas, para poder evaluar la calidad de la clasificación, para lo cual se procede de acuerdo al esquema de la Figura 3.

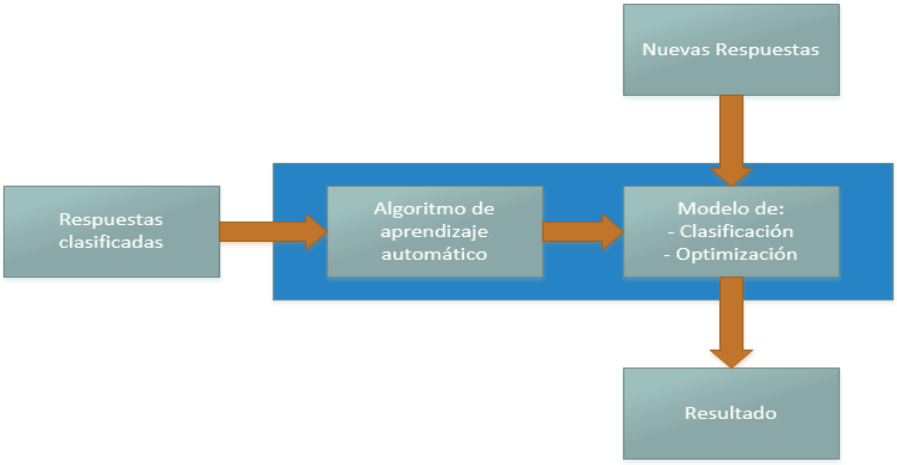


Figura 3. Esquema del proceso de aprendizaje automático

1.5. Algoritmos de clasificación

1.5.1. K- vecinos más cercanos

De los métodos de clasificación supervisados más utilizados por su simplicidad y bajo costo computacional son los llamados basados en instancias, que utilizan criterios de vecindad. La idea fundamental detrás de los métodos basados en instancias, en especial del método llamado K-Vecinos más cercanos (o kNN), es que las muestras pertenecientes a una misma clase, probablemente, se encontraran cercanas en un espacio de representación común. (Sierra Araujo, 2006).

La asignación de clase de un documento **d** se hace considerando los **k** documentos, en el conjunto de entrenamiento, más cercanos a **d**; la clase de la mayoría de esos **k** documentos es la clase asignada al documento **d**. (Ramírez de la Rosa, 2010)
Un vecino más cercano es un documento que se define en términos de la distancia Euclidiana como sigue:

$$distancia\left(d_i, d_j\right) = \sqrt{\sum_{r=1}^{M}(t_r^i - t_r^j)^2} \qquad (1)$$

Fuente: (Ramírez de la Rosa, 2010)

Donde $\mathbf{d_i}$ y $\mathbf{d_j}$ son los vectores de los documentos a comparar, **M** es el total de atributos en el diccionario, $\mathbf{t_r^i}$ es el r-ésimo atributo del documento $\mathbf{d_i}$ y $\mathbf{t_r^i}$ es el r-ésimo atributo del documento $\mathbf{d_j}$. (Ramírez de la Rosa, 2010)

1.5.2. Naive Bayes

El aprendizaje Bayesiano se basa en la asunción de que las cantidades de interés están gobernadas por distribuciones de probabilidades y que las decisiones óptimas se pueden hacer razonando sobre estas probabilidades y sobre los datos observados (Buill Vilches, 2014).

El clasificador Naive Bayes es un clasificador probabilístico que se basa en aplicar el teorema de Bayes, que asume la independencia dado el valor de la clase, es por eso que se le llama clasificador bayesiano ingenuo (naive). (Martis Cáceres, 2012)

Como se explica en (Tomás Diaz, 2010), desde el punto de vista de la clasificación de textos se puede decir que se asume la independencia de las palabras, es decir, la probabilidad condicional de una palabra dada una clase se asume que es independiente de la probabilidad condicional de otras palabras dada esa clase.

Definiendo más formalmente, sea {1…K} el conjunto de clases posibles y {xi,1, … , xi,m} el conjunto de valores de las características del ejemplo xi el algoritmo Naive Bayes selecciona la clase que maximiza $P(k|x_{i,1},\ldots,x_{i,m})$: (Martis Cáceres, 2012)

$$\arg\max_x \; P(k|x_{i,1},\ldots,x_{i,m}) \approx \arg\max_x \; P(k)\prod_j P\left(x_{i,j}|k\right) \qquad (2)$$

Fuente: (Martis Cáceres, 2012)

Las probabilidades P(k) y P ($x_{i,j}$,K), se estiman a partir del corpus de aprendizaje a través de las frecuencias relativas.
Es conocido que el algoritmo de Naive Bayes tiene una limitante para trabajar en espacios de gran dimensionalidad, en otros términos no puede trabajar con un gran número de características de aprendizaje.

1.5.3. Árboles de decisión

Son árboles cuyos nodos están etiquetados por términos, las ramas salientes están etiquetadas por los pesos de estas y las hojas corresponden a las categorías. De esta forma se recorre el árbol de arriba abajo para cada uno de los documentos, hasta llegar a una hoja y asignar una categoría.

En la Figura 4, se puede observar la estructura típica de un árbol de decisión donde cada rama representa una decisión.

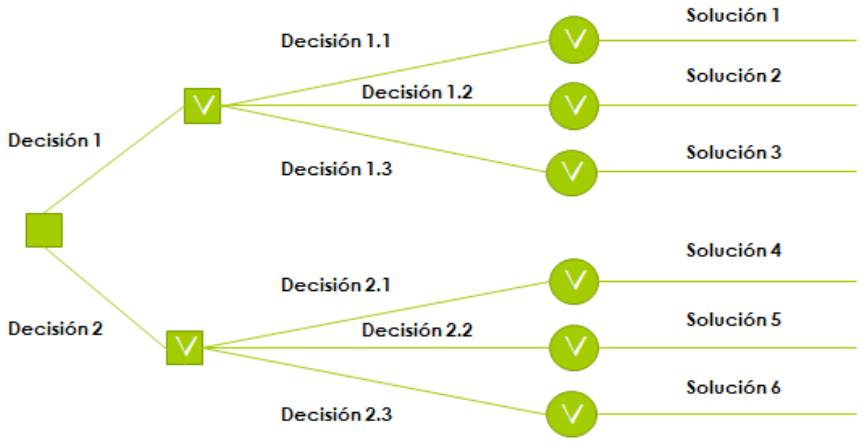


Figura 4. Árbol de decisión *(Vergara García, 2014)*

1.5.4. Redes neuronales artificiales

Las redes neuronales han mostrado resultados exitosos con cálculos más sencillos para categorización de documentos.

El concepto de Red Neuronal Artificial está inspirado en las Redes Neuronales Biológicas. Una Red Neuronal Biológica es un dispositivo no lineal altamente paralelo, caracterizado por su robustez y su tolerancia a fallos. Sus principales características son las siguientes: aprendizaje mediante adaptación de sus pesos sinápticos a los cambios en el entorno, manejo de imprecisión, ruido e información probabilística y generalización a partir de ejemplos (Jimenez, Paz -Arias, & Larco, 2015). En la Figura 5, se puede observar una red totalmente conectada.

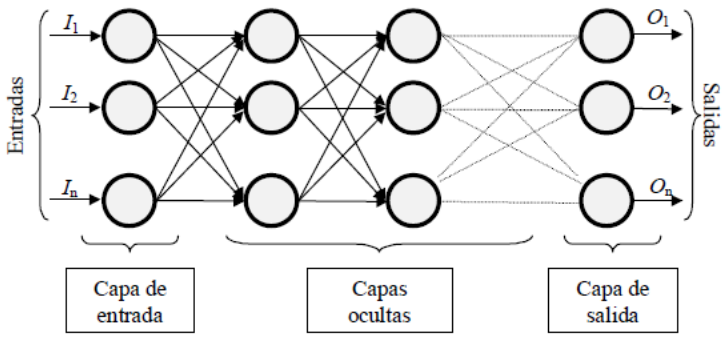


Figura 5. Ejemplo de una red totalmente conectada *(Jimenez, Paz -Arias, & Larco, 2015)*

Las redes neuronales tienen características que permiten llevar a cabo una categorización de documentos eficiente y de alta calidad, en los últimos años ha sido notoria la implementación de redes neuronales con aprendizaje competitivo para la categorización de documentos. El aprendizaje competitivo se basa en el criterio de: "*el ganador* se queda con todo" (winner-takes-all).

1.5.5. Máquina de soporte vectorial (SVM)

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase (Buill Vilches, 2014).

Se lo puede definir de la siguiente forma, dado un conjunto de datos de entrenamiento $\{\boldsymbol{x_i}, \boldsymbol{y_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{n}} \in \boldsymbol{R^m x} \{\pm \mathbf{1}\}$, se desea es encontrar el hiperplano óptimo que divida las dos clases de datos. El correspondiente hiperplano puede ser definido como: (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

$$\boldsymbol{w^T x + b = 0}, \qquad (3)$$

Fuente: (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

Donde x es un vector de datos, $\boldsymbol{w^T}$ es el vector de parámetros de nuestro modelo y b es un término independiente que ofrece mayor libertad al momento de encontrar el hiperplano óptimo para clasificar los datos. (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

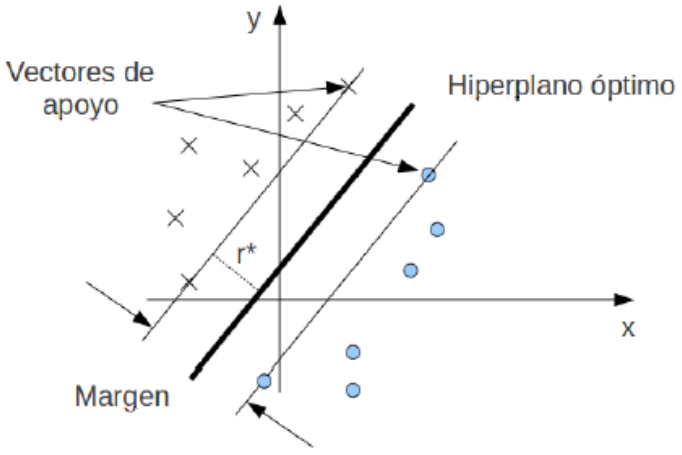


Figura 6. Hiperplano para un caso linealmente separable *(Varguez Moo, Brito Loeza, & Uc Cetina, 2012)*

Los puntos que aparecen en la Figura 6 que se encuentran sobre las líneas no punteadas se les conoce como vectores de apoyo. Cuando los datos de prueba no son linealmente separables se puede adoptar dos técnicas para resolverlo: con optimización "margen suave" y a través de kernel. (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

Nos enfocaremos en el uso del kernel para ello se debe obtener el clasificador óptimo de la siguiente forma:

$$\boldsymbol{f(x) = \sum_i^n \alpha_i^* y_i K(x_i, x) + b,} \qquad (4)$$

Fuente: (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

Donde α es el multiplicador óptimo de Lagrange y $\boldsymbol{K(x_i, x)}$es una función kernel. Los kernels comúnmente usados son: (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

- Polinomial $\boldsymbol{K(x_i, x) = (x_i^T . x + c)^d}$
- Función de base radial $\boldsymbol{K(x_i, x) = \exp(-\gamma \|x_i - x\|^2), \gamma > 0}$
- Sigmoidal $\boldsymbol{K(x_i, x) = \tanh(x_i^T . x + c)}$ (Varguez Moo, Brito Loeza, & Uc Cetina, 2012)

Es importante mencionar que la elección del kernel adecuado para el algoritmo de Máquina de Soporte Vectorial permite entrenar correctamente el clasificador y esto se evidencia en los resultados de la clasificación.

2. METODOLOGÍA

Para el desarrollo de este trabajo de investigación se utilizó como base el proceso de minería de datos detallado en la Figura 1, donde se identifican 4 fases. Este proceso fue aplicado al caso de estudio realizado en la Universidad Técnica Particular de Loja y la implementación se la va a realizar en el sistema de evaluaciones.

2.1 Fase 1: Selección de datos

El objetivo de esta etapa es seleccionar las fuentes de datos para el proceso de minería de datos, como fuente interna de la Universidad Técnica Particular de Loja, se seleccionaron evaluaciones de los bancos de preguntas de la Dirección de Tecnologías de la Educación (DTE) referentes a dos componentes educativos, que tienen la mayor cantidad de estudiantes. Los componentes académicos que se utilizarán son: Metodología de Estudio y Realidad Nacional y Ambiental, estos componentes son comunes en todas las titulaciones de la universidad, en cada periodo académico cursan estos componentes académicos 7.000 estudiantes aproximadamente.

Posterior a ello se procedió a realizar la transcripción de las respuestas de 20 estudiantes referente a 648 preguntas, obteniendo un total de 12960 respuestas tipo texto; por cada pregunta existe una única respuesta correcta y puede ser de tipo dicotómica o muchas opciones.

En la Figura 7, se observa una pregunta del componente educativo denominado Metodología de Estudio donde las opciones de respuestas son verdadero o falso, a estas se las considera de tipo dicotómica. La respuesta correcta es FALSO.

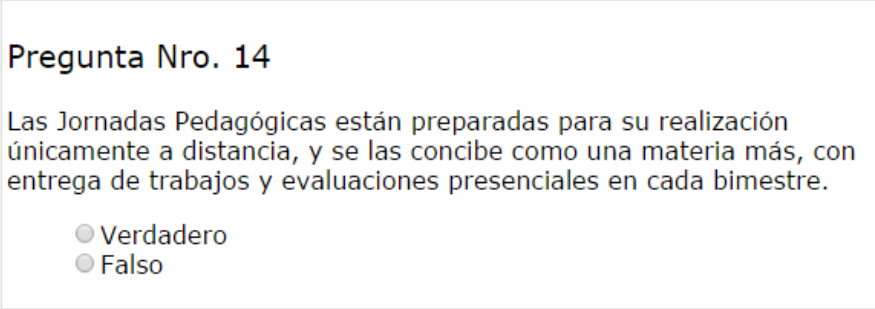


Figura 7. Pregunta dicotómica de Metodología de Estudio *(Universidad Técnica Particular de Loja, 2013)*

En la Figura 8, se representa una pregunta del componente educativo denominado Realidad Nacional y Ambiental, la cual tiene múltiples opciones de respuesta. La respuesta correcta es: Rechazan a los partidos tradicionales.

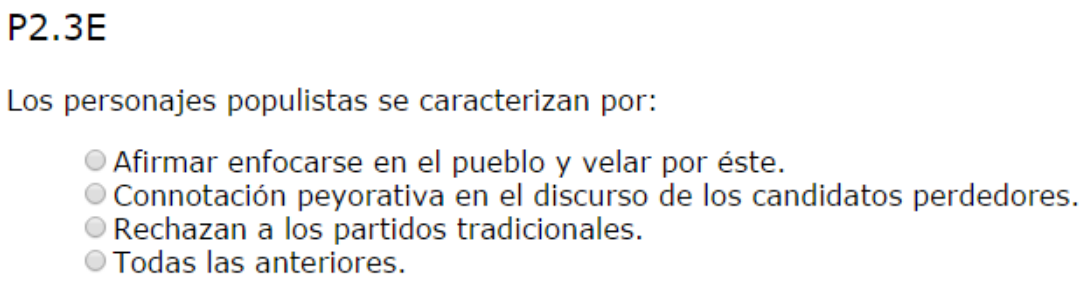


Figura 8. Pregunta con muchas opciones de Realidad Nacional y Ambiental *(Universidad Técnica Particular de Loja, 2013)*

1.1. Fase 2: Preprocesamiento de datos

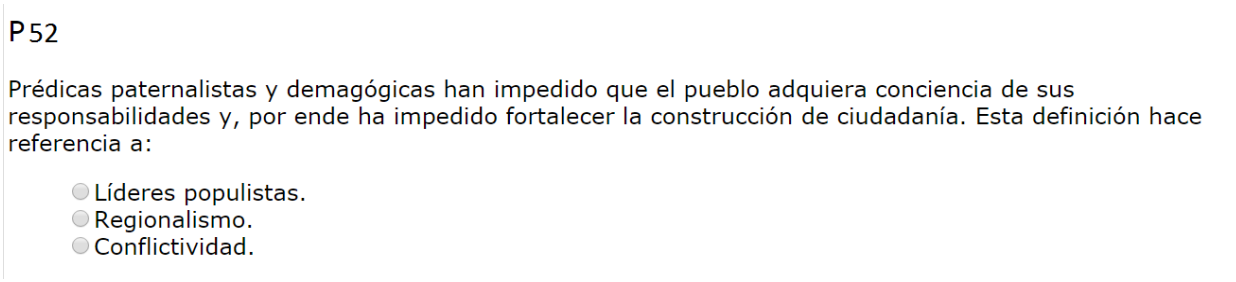
El objetivo de esta etapa es obtener los datos limpios, datos sin valores nulos o anómalos que permitan obtener patrones de calidad. Se reemplazó todos los caracteres extraños (letra ñ, tildes, doble espacios), que no admitía la herramienta que se utilizó para el proceso de entrenamiento WEKA (Waikato Environment for Knowledge Analysis)[1]. Este proceso se lo realizó de forma masiva y automática con la ayuda de comandos específicos para esta tarea.

1.2. Fase 3: transformación de datos

El objetivo de esta fase es transformar la fuente de datos en un conjunto listo para aplicar las diferentes técnicas de minería de datos. Para facilitar el proceso de entrenamiento se adecuaron todas las respuestas por preguntas al formato ARFF (Atribute Relation File Format) requerido por

[1] http://www.cs.waikato.ac.nz/ml/weka/

WEKA (Waikato Environment for Knowledge Analysis)[2] y se definió la clase con valores *YES* o *NO*, donde a cada respuesta correcta transcrita de los estudiantes se le fijó el valor *YES* y las incorrectas el valor *NO*. En la Figura 9, se muestra la pregunta 52 cuya respuesta correcta es: Conflictividad.


P 52

Prédicas paternalistas y demagógicas han impedido que el pueblo adquiera conciencia de sus responsabilidades y, por ende ha impedido fortalecer la construcción de ciudadanía. Esta definición hace referencia a:

- Líderes populistas.
- Regionalismo.
- Conflictividad.

Figura 9. Pregunta 52

En la figura 10, se puede ver el resultado del proceso de transformación de la pregunta 52 con sus respectivas respuestas.

```
@relation Pregunta52

@attribute text string
@attribute class {YES, NO}

@data
"los Lideres populistas",NO
"seria los Lideres populistas",NO
"creo que son Lideres populistas",NO
"me parece que son los Lideres populistas",NO
"supongo que son Lideres populistas",NO
"imagino que son Lideres populistas",NO
"sin lugar a dudas son los Lideres populistas",NO
"el Regionalismo",NO
"Imagino el Regionalismo",NO
"me parece que el Regionalismo",NO
"podria ser el Regionalismo",NO
"imagino que es el Regionalismo",NO
"la Conflictividad",YES
"seria la Conflictividad",YES
"Imagino que la Conflictividad",YES
"considero que la Conflictividad",YES
"supongo la Conflictividad",YES
"podria ser la Conflictividad",YES
"definitivamente la Conflictividad",YES
"sin lugar a dudas la Conflictividad",YES
```

Figura 10. Resultado del proceso de transformación de la pregunta 52

Posterior a ello se aplicó el filtro *StringToWordVector* que convierte los atributos de tipo String en un conjunto de atributos representando la ocurrencia de las palabras del texto.

[2] http://www.cs.waikato.ac.nz/ml/weka/

1.3. Fase 4: Experimentación

En esta etapa se aplicó una serie de algoritmos por cada pregunta donde el objetivo es alcanzar el 100% en la clasificación correcta de cada respuesta.

EVALUACIÓN Y RESULTADOS

Una vez generado el clasificador es necesario medir su capacidad de predicción sobre las nuevas instancias, para ello es necesario contar con un conjunto de ejemplos (diferente al conjunto con el cual se generó el clasificador), llamado conjunto de evaluación. Cuando no se cuenta con un conjunto de evaluación independiente del conjunto de entrenamiento se realiza una validación cruzada (k-fold cross validation), donde se divide el conjunto de entrenamiento en k subconjuntos de forma aleatoria, luego se entrena un clasificador utilizando k − 1 subconjuntos y el restante se utiliza para evaluar los resultados. Este procedimiento se realiza k veces, al final se obtiene un promedio que resume los resultados obtenidos. Existen dos métodos para obtener una evaluación general del desempeño del clasificador, macro promedio y micro-promedio. Para ambos métodos es necesario obtener cuatro valores por cada clase: (Ramírez de la Rosa, 2010).

- a_j - el número de documentos asignados correctamente a la clase j.
- b_j - el número de documentos asignados incorrectamente a la clase j.
- c_j - el número de documentos rechazados incorrectamente de la clase j.
- d_j - el número de documentos rechazados correctamente de la clase j.

A partir de estos cuatro valores se definen las medidas de desempeño recuerdo, precisión y medida-F. Utilizando macro promedio, estas medidas son calculadas como se muestra en las Ecuaciones 5 a 7, donde K es el número de clases diferentes.

$$\boldsymbol{macro\ recuerdo} = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \frac{\boldsymbol{a_j}}{\boldsymbol{a_j + c_j}} \qquad (5)$$

Fuente: *(Ramírez de la Rosa, 2010)*

$$\boldsymbol{macro\ precisición} = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \frac{\boldsymbol{a_j}}{\boldsymbol{a_j + b_j}} \qquad (6)$$

Fuente: *(Ramírez de la Rosa, 2010)*

$$\boldsymbol{macro\ medida\ F} = \frac{\mathbf{2} * \boldsymbol{marco\ recuerdo} * \boldsymbol{macro\ precisión}}{\boldsymbol{macro\ recuerdo} + \boldsymbol{macro\ precisión}} \qquad (7)$$

Fuente: *(Ramírez de la Rosa, 2010)*

Para calcular el micro promedio se considera la colección completa, las Ecuaciones 8 a 10 muestran cómo se obtienen utilizando los valores a_j , b_j , c_j y d_j descritos anteriormente.

$$\boldsymbol{micro\ recuerdo} = \frac{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{a}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} (\boldsymbol{a_j + c_j})} \qquad (8)$$

Fuente: *(Ramírez de la Rosa, 2010)*

$$\boldsymbol{micro\ precisión} = \frac{\sum_{j=1}^{K} \boldsymbol{a}}{\sum_{j=1}^{K} (\boldsymbol{a_j} + \boldsymbol{b_j})} \qquad (9)$$

Fuente: *(Ramírez de la Rosa, 2010)*

$$\boldsymbol{micro\ medidaF} = \frac{\boldsymbol{2 * micro\ recuerdo * micro\ precisión}}{\boldsymbol{micro\ recuerdo + micro\ precisión}} \qquad (10)$$

Fuente: (Ramírez de la Rosa, 2010)

Resultados

En la Tabla 1, se muestran los porcentajes obtenidos en cada pregunta por cada algoritmo ejecutado, para el caso de las preguntas dicotómicas se consideró que al tener la misma opción de respuesta para todas las preguntas se englobó en una única ejecución tanto para cuando la respuesta correcta es *VERDADERO* y también para cuando lo es *FALSO.*

Por otra parte se puede observar en los resultados de la Tabla 1, que en la mayoría de las preguntas se obtuvo un 100% de clasificación de las respuestas con más de un algoritmo, sin embargo existen algunas preguntas cuyas respuestas no pudieron ser clasificadas al 100% con ningún algoritmo ejecutado.

N° de pregunta	**Algoritmos**						
	J48	**IBK**	**MLP**	**Naive Bayes**	**Random Forest**	**SMO**	**Bayes Net**
Preguntas (Respuesta = VERDADERO)	87,23%	82,98%	85,11%	78,72%	82,98%	87,23%	78,72%
Preguntas (Respuesta = FALSO)	87,23%	80,85%	87,23%	78,72%	80,85%	87,23%	78,23%
Pregunta 16	81,82%	100,00%	100,00%	100,00%	100,00%	100,00%	81,82%
Pregunta 17	90,00%	100,00%	100,00%	90,00%	100,00%	100,00%	90,00%
Pregunta 18	100,00%	100,00%	100,00%	100,00%	90,00%	100,00%	100,00%
Pregunta 19	90,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 20	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 21	90,00%	90,00%	100,00%	100,00%	100,00%	100,00%	90,00%
Pregunta 22	100,00%	95,24%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 23	95,24%	90,48%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 24	95,45%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 25	95,24%	100,00%	100,00%	95,24%	95,24%	100,00%	95,24%
Pregunta 26	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Pregunta 27	95,24%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 28	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 29	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 30	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 31	100,00%	100,00%	100,00%	100,00%	95,00%	100,00%	100,00%
Pregunta 32	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 33	100,00%	90,00%	100,00%	100,00%	95,00%	100,00%	100,00%
Pregunta 34	80,00%	95,00%	95,00%	100,00%	100,00%	95,00%	80,00%
Pregunta 35	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 51	100,00%	95,24%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 52	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 53	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 54	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 55	95,24%	95,24%	95,24%	100,00%	95,24%	95,24%	76,19%
Pregunta 56	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 57	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 58	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 59	95,00%	100,00%	100,00%	100,00%	95,00%	100,00%	95,00%
Pregunta 60	100,00%	89,47%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 61	100,00%	100,00%	100,00%	100,00%	95,24%	100,00%	100,00%
Pregunta 62	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 63	95,00%	100,00%	100,00%	100,00%	100,00%	100,00%	95,00%
Pregunta 64	100,00%	66,67%	100,00%	90,48%	100,00%	100,00%	100,00%
Pregunta 65	90,48%	90,48%	90,48%	90,48%	90,48%	90,48%	90,48%
Pregunta 66	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 67	100,00%	80,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 68	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 69	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 70	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 111	100,00%	80,00%	100,00%	95,00%	95,00%	100,00%	100,00%
Pregunta 112	100,00%	85,00%	100,00%	95,00%	100,00%	100,00%	100,00%
Pregunta 113	95,00%	95,00%	95,00%	95,00%	90,00%	95,00%	95,00%
Pregunta 114	100,00%	85,00%	100,00%	95,00%	100,00%	100,00%	100,00%
Pregunta 115	100,00%	95,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 116	90,00%	80,00%	100,00%	85,00%	90,00%	85,00%	85,00%
Pregunta 117	85,00%	75,00%	85,00%	75,00%	85,00%	85,00%	90,00%
Pregunta 118	85,00%	55,00%	85,00%	85,00%	55,00%	80,00%	85,00%
Pregunta 119	45,00%	65,00%	75,00%	55,00%	70,00%	75,00%	65,00%
Pregunta 120	70,00%	50,00%	80,00%	80,00%	85,00%	70,00%	80,00%
Pregunta 121	45,00%	60,00%	70,00%	50,00%	65,00%	70,00%	65,00%

Pregunta 122	100,00%	30,00%	100,00%	75,00%	95,00%	100,00%	70,00%
Pregunta 123	50,00%	75,00%	70,00%	80,00%	70,00%	75,00%	70,00%
Pregunta 164	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 165	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 166	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 167	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 168	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 169	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 170	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 171	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 172	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 173	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 174	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 175	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 216	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 257	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%
Pregunta 258	90,00%	100,00%	100,00%	100,00%	100,00%	100,00%	90,00%
Pregunta 259	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 1. Resultados en la ejecución de los algoritmos

En la Figura 11, se demuestra que el algoritmo que obtuvo un mayor porcentaje global de aciertos en el proceso de clasificación son las redes neuronales (MLP), seguidamente se encuentra el algoritmo de Máquina de Soporte Vectorial (SOM) obteniendo también buenos resultados; por otra parte el algoritmo que obtuvo un menor porcentaje global de aciertos en comparación con los otros algoritmos ejecutados es K- vecinos más cercanos (IBK).

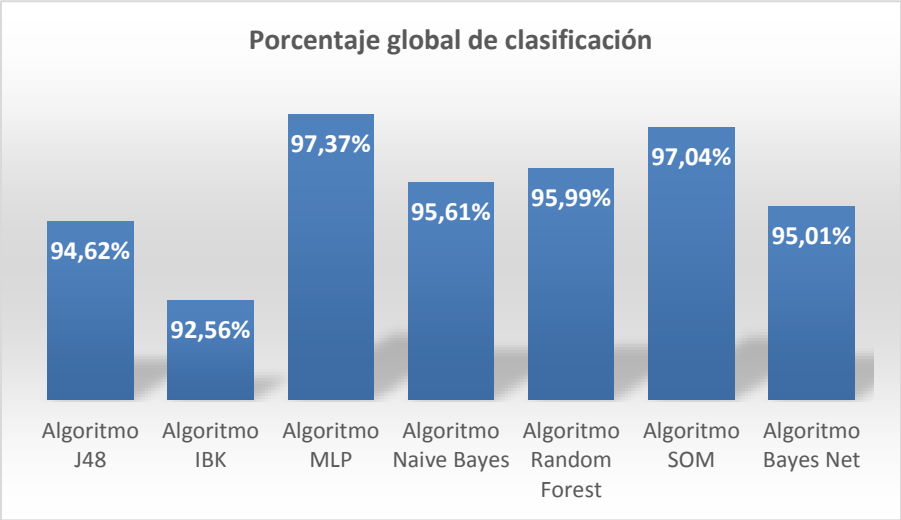


Figura 11. Promedio global de los porcentajes de los algoritmos.

Conclusiones:

- Se ha identificado que las redes neuronales y las Máquinas de Soporte Vectorial han demostrado ser los clasificadores más idóneos para este problema de clasificación, pudiéndose implementar en la integración con un sistema de evaluación en línea.
- El clasificador de acuerdo a los resultados obtenidos en el presente trabajo que ha obtenido el menor porcentaje global de clasificación correcta es el K- vecinos más cercanos (IBK), por lo tanto lo excluye para ser considerado en la fase de implementación en el sistema de evaluación en línea.
- En las preguntas que no se ha alcanzado un 100% en la clasificación de las respuestas con los algoritmos propuestos, se ha identificado que se debe a la existencia de respuestas cuya clasificación debe ser enfocada a la similitud semántica textual que existe entre las respuestas dadas. Donde existe el reto de encontrar el grado de similitud entre un párrafo y una sentencia, una sentencia y una frase, una frase y una palabra y una palabra y un sentido.

Trabajos Futuros:

- Implementación en el sistema de evaluaciones en línea que incluya un módulo de reconocimiento de voz para que sean transcritas las respuestas dadas por los estudiantes de forma hablada.
- Determinar el algoritmo que obtenga el 100% en la clasificación de las respuestas para las que no se ha obtenido una clasificación correcta total.
- Ampliar el proceso de experimentación con otros componentes educativos cuyo nivel de dificultad en el proceso de clasificación es mayor, ya que se requiere interpretar semánticamente las respuestas emitidas.
- Implementación del proceso de aprendizaje automático en el sistema de evaluaciones en línea de tal forma que permita clasificar correctamente las respuestas de tipo texto obtenidas por los estudiantes que han sido trascritas y que brinde de forma inmediata la calificación de la evaluación aplicada, de esta forma se beneficiará directamente a los estudiantes con discapacidad física y visual.

Referencias

Aitken, J. E., Pedejo Fairley, J., & Carlson, J. K. (2012). Communication Technology for Students in Special Education and Gifted Programs. In *Communication Technology for Students in Special Education and Gifted Programs* (pp. 105-116).

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Buill Vilches, J. (2014, Junio 19). Clasificación automática de textos y explotación BI. Barcelona, España.

Consejo de Educación Superior. (2016, Marzo 22). *CES.* Retrieved from http://www.ces.gob.ec/gaceta-oficial/download/file?fid=231.3707

García Aretio, L. (2002). *La educación a distancia: de la teoría a la práctica.* Ariel S.A.

Guzmán Cabrera, R. (2009, Noviembre). Categorización semi-supervisada de documentos usando la web como corpus. Valencia, España.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2007). *Introducción a la Minería de Datos.* Madrid: Pearson.

Jimenez, D., Paz -Arias, H., & Larco, A. (2015). Desarrollo de un sistema inteligente para la clasificación de documentos ya digitalizados aplicando redes neuronales supervisadas. *Revista Tecnológica ESPOL – RTE*, 8-23.

López Condori, R. E. (2014, Diciembre). Método de Clasificación Automática de Textos basado en Palabras Claves utilizando Información Semántica: Aplicación a Historias Clínicas. Perú.

Martis Cáceres, M. A. (2012, Enero). CLASIFICACIÓN AUTOMÁTICA DE LA INTENCIÓN DEL USUARIO EN MENSAJES DE TWITTER.

Méndez Martínez, J., & Ruiz Méndez, R. (2015). *Evaluación del aprendizaje y tecnologías de información y comunicación (TIC): De la precensialidad a la educación a distancia.* Retrieved from http://revalue.mx/revista/index.php/revalue/issue/current

Mitchell, T. (1997). *Machine learning.* McGraw-Hill.

Pérez López, C., & Santín González, D. (2008). *Minería de Datos: técnicas y herramientas.* Madrid: España.

Ramírez de la Rosa, A. G. (2010, Noviembre). Clasificacion de textos utilizando información inherente al conjunto a clasificar. Tonantzintla, Puebla, México.

Rokach, L., & Maimon, O. (2014). *Data Mining with Decision Trees: Theory and Applications.* World Scientific.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys 34*, 1-47.

Sebastiani, F. (2005). Text categorization. *In Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, 109-129.

Sierra Araujo, B. (2006). *Aprendizaje Automático: Conceptos básicos y avanzados.* Madrid: PRENTICE-HALL.

Tomás Diaz, D. (2010). Sistemas de clasificación de preguntas basados en corpus para la búsqueda de respuestas.

Universidad Técnica Particular de Loja. (2012a). *Evaluación*. Obtenido de http://distancia.utpl.edu.ec/modalidad-abierta/sistema-evaluacion

Universidad Técnica Particular de Loja. (2012b). *Modalidad Abierta y a Distancia*. Retrieved from http://distancia.utpl.edu.ec/modalidad-abierta/descripcion

Universidad Técnica Particular de Loja. (2013, Agosto). Banco de Preguntas de Metodología de Estudio. Loja, Loja, Ecuador.

Universidad Técnica Particular de Loja. (2013, Agosto). Banco de Preguntas de Realidad Nacional y Ambiental. Loja, Loja, Ecuador.

Universidad Técnica Particular de Loja. (2014, Julio 18). *Guía General MAD.* Retrieved from http://www.utpl.edu.ec/sites/default/files/pregrado/guia-general-MAD.pdf

Varguez Moo, M., Brito Loeza, C., & Uc Cetina, V. (2012). Clasificación de documentos usando Máquinas de Vectores de Apoyo. *Abstraction and Application Magazine 6*, 40-51.

Vergara García, P. M. (2014, Mayo). SELECCIÓN Y EVALUACIÓN DE ALGORITMOS PARA CLASIFICACIÓN DE DOCUMENTOS.