PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido

Tesis para optar el Título de Ingeniero Informático, que presenta el bachiller:

Luis Alejandro Pinto Valdiviezo

ASESOR: Dr. Héctor Andrés Melgar Sasieta.

Lima, Agosto de 2015

RESUMEN

En los últimos años la generación de información virtual ha aumentado considerablemente. Parte de esa información se encuentra almacenada en bases de datos de instituciones públicas y privadas. Sin embargo, no toda la información almacenada de forma electrónica tiene una estructura definida, tal es el caso de los documentos donde encontramos secuencias de palabras no estructuradas, los cuales según estudios representan el 80% de la información de las empresas.

La tarea de clasificar automáticamente documentos tiene como motivo principal brindar una herramienta de mejora en la gestión de la información, la cual es considerada como condición indispensable para el éxito de cualquiera empresa.

Ante esto, en el propósito del proyecto se propone la obtención de un modelo algorítmico para la clasificación automática de documentos de carácter judicial en lenguaje portugués según su contenido con el fin de automatizar las labores manuales involucradas en el proceso, y con ello disminuir los recursos implicados en la tarea de clasificación. La colección de documentos será brindada por una empresa en Brasil encargada de la clasificación manual de intimaciones a través de especialistas, llamados procuradores. Las intimaciones son documentos que son enviados desde los tribunales hacia las procuradurías durante un proceso de juicio.

CONTENIDO

Índice de ilustraciones

Índice de Tablas

CAPÍTULO 1

1 Problemática

En los últimos años la generación de información virtual ha aumentado considerablemente. Parte de esa información se encuentra almacenada en bases de datos de instituciones públicas y privadas, a las cuales se tiene acceso todos los días a través de internet o redes locales [Varguez-Moo, et al. 2014]. Sin embargo, no toda la información almacenada de forma electrónica tiene una estructura definida, tal es el caso de los documentos donde encontramos secuencias de palabras no estructuradas, los cuales según estudios representan el 80% de la información de las empresas [Tan 1999]. Para organizar dicha información es importante contar con métodos automatizados que permitan realizar una clasificación de una colección de documentos según su contenido lo más similar a como lo haría un especialista humano. De esta manera, el uso y gestión de los documentos se puede realizar de forma más ágil y eficiente.

Clasificación es el tema más común en el análisis de datos complejos, y es descrita como la tarea de asignar a un dato una categoría dentro de un conjunto predefinido de las mismas. Aplicado al dominio del problema de manejo y organización de documentos, es conocida como Categorización de Documentos y es definida como el proceso de encontrar la categoría correcta para cada documento dada una colección de categorías y una colección de documentos de texto [Feldman & Sanger 2007]. La clasificación de datos, es un proceso de dos etapas. En primer lugar, un clasificador es construido describiendo un conjunto de datos y conceptos predeterminados. Esta es la etapa de entrenamiento, donde un algoritmo de clasificación construye un modelo analizando o aprendiendo de un conjunto de datos de entrenamiento compuesto por tuplas de datos y su clase asociada. Por otro lado, en la segunda etapa el modelo es usado para la clasificación con el objetivo de evaluar su rendimiento. El rendimiento de un clasificador está dado por el porcentaje de tuplas correctamente clasificadas por el modelo [Han & Kamber 2006].

El problema de la categorización de documentos se inicia en la década de los 80's, donde se empleó un motor de conocimiento junto a un experto basado en un conjunto de reglas lógicas; sin embargo, debido a que se requería actualizar dichas reglas de manera manual, la técnica no alcanzo el éxito esperado. Es así como no fue hasta los 90's cuando se automatizó la tarea a través del Aprendizaje Automático, donde mediante un conjunto de datos el algoritmo es capaz de aprender un modelo de categorización. Entre las principales ventajas que ofrece este enfoque se encuentran el nivel de optimización de la tarea y la similitud de los resultados obtenidos en comparación con lo de un experto humano. Un ejemplo en el uso de este enfoque lo encontramos en los buscadores de *Google* [Varguez-Moo, et al. 2014].

La tarea de clasificar automáticamente documentos tiene como motivo principal brindar una herramienta de mejora en la gestión de la información, la cual es considerada como condición indispensable para el éxito de cualquiera empresa. Una adecuada gestión de la información, se basa en un tratamiento automático de la misma [García & Lucas 1987].

La tarea de encontrar una colección de documentos con un conjunto similar de características no sólo es compleja sino que además consume tiempo. En el caso que esta se realice de manera manual, significaría la necesidad que un experto humano leyera todo el documento, para que una vez comprendido en base a su experiencia y conocimiento le asigne una categoría bajo su propio criterio. Es así, como el costo y tiempo asociados a la categorización manual de documentos ha llevado a la investigación de técnicas que permitan la automatización de la tarea [Rüger & Gauch 2000].

Existen tecnologías de procesamiento y acceso a la información que se han dedicado al diseño de técnicas capaces de solucionar el problema de la clasificación automática de documentos. Estas tecnologías surgen como soporte para el descubrimiento del conocimiento sobre datos almacenados y son conocidas como Minería de Datos (*Data Mining)* y Minería de Textos (*Text Mining*). Mientras que la Minería de Datos es un paso en el proceso de *Knowledge Discovery in Data (KDD)*, que consiste en la aplicación de análisis de datos y algoritmos de descubrimiento que producen un conjunto particular de patrones o comportamiento de los datos [Fayyad, et al. 1996]; la Minería de Textos según Feldman y Ben-Dov (2005) es un descubrimiento automático de información previamente desconocida, cuyo proceso empieza con la extracción de hechos y eventos de fuentes de

texto no-estructurados con el propósito de formular hipótesis que son exploradas por métodos tradicionales de Minería de Datos y Análisis de Datos, tales como: Recuperación de Información, Análisis de Textos, Extracción de Información y Aprendizaje Automático [Feldman & Sanger 2007].

El presente proyecto tendrá como base de trabajo una colección de documentos de carácter judicial en lenguaje portugués, la cual será aplicada durante las etapas de construcción y evaluación del modelo de clasificación. La colección será brindada por una empresa en Brasil encargada de la clasificación manual de intimaciones a través de especialistas, llamados procuradores. Las intimaciones son documentos que son enviados desde los tribunales hacia las procuradurías durante un proceso de juicio.

Se tiene además como propósito obtener un modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido con el fin de automatizar las labores manuales involucradas en el proceso, y con ello disminuir los recursos implicados en la tarea de clasificación.

1.1. Objetivo general

Obtener un modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido.

1.2. Objetivos específicos

- Objetivo 1: Recopilar y estructurar el conjunto de documentos.
- Objetivo 2: Completar las fases de adaptación y evaluación de cuatro algoritmos de Aprendizaje Automático aplicables a tareas de clasificación.
- Objetivo 3: Realizar un análisis comparativo en base a indicadores de precisión que sustente la selección del modelo de clasificación más apto.
- Objetivo 4: Aplicar el modelo algorítmico en la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido.

1.3. Resultados esperados

- Objetivo 1: Recopilar y estructurar el conjunto de documentos.
  - ✓ RE 1: Conjunto de documentos estructurados según el formato aceptado por la herramienta Weka.
- Objetivo 2: Completar las fases de adaptación y evaluación de cuatro algoritmos de Aprendizaje Automático aplicables a tareas de clasificación.
  - ✓ RE 2: Conjunto de modelos de clasificación adaptados y evaluados.
  - ✓ RE 3: Conjunto de indicadores de precisión por modelo de clasificación evaluado.
- Objetivo 3: Realizar un análisis comparativo en base a indicadores de precisión que sustente la selección del modelo de clasificación más apto.
  - ✓ RE 4: Documentación del análisis de los modelos de clasificación en base a indicadores de precisión.
  - ✓ RE 5: Modelo de clasificación seleccionado.
- Objetivo 4: Aplicar el modelo algorítmico en la clasificación de documentos de carácter judicial en lenguaje portugués.
  - ✓ RE 6: Prototipo funcional que permita clasificar documentos de carácter judicial en lenguaje portugués según su contenido haciendo uso del modelo algorítmico de clasificación.

2 Herramientas y métodos

En el siguiente apartado se detallarán las diversas herramientas, métodos y procedimientos que serán utilizados para el correcto desarrollo de cada uno de los resultados esperados detallados previamente en este documento. A continuación, se presenta un cuadro introductorio de las herramientas a utilizarse versus los resultados esperados.

Tabla 1: Cuadro introductorio de las herramientas a utilizarse versus los resultados esperados.

Resultado esperado	Herramienta a usarse
• RE 1: Conjunto de documentos estructurados según el formato aceptado por la herramienta Weka.	**Editor de texto,** la herramienta servirá para estructurar el conjunto de documentos recopilados de una empresa en Brasil para su posterior uso en la construcción y evaluación del clasificador.
• RE 2: Conjunto de modelos de clasificación adaptados y evaluados. • RE 3: Conjunto de indicadores de precisión por modelo de clasificación evaluado.	**Weka**, herramienta empleada para realizar pruebas sobre los diversos modelos de clasificación existentes en el campo de la Minería de Datos en base a un conjunto de datos específico.
• RE 5: Modelo de clasificación seleccionado.	**Weka**, la herramienta permite guardar un modelo de clasificación previamente construido para su posterior uso en otros proyectos de desarrollo.
• RE 6: Prototipo funcional que permita clasificar nuevos documentos haciendo uso del modelo de clasificación seleccionado.	**Netbeans IDE**, el presente entorno de desarrollo será utilizado durante la etapa de desarrollo de la interfaz de usuario.

2.1 Herramientas

2.1.1 *Weka*

WEKA es una aplicación desarrollada en la Universidad de Waikato en Nueva Zelanda empleando el lenguaje de programación Java y distribuida como una aplicación de licencia libre. El aplicativo puede ser usado en cualquier plataforma, tales como Linux y Windows [Witten & Frank 2005].

Weka es una colección de algoritmos de aprendizaje automático para tareas de Minería de Datos. Desde el enfoque de la categorización de documento, el aprendizaje automático es un proceso inductivo que construye un modelo algorítmico de clasificación mediante el aprendizaje de propiedades de un conjunto de documentos previamente clasificados que cumplirán la función de ejemplos [Feldman & Sanger 2007].

Los algoritmos pueden ser aplicados directamente a un conjunto de datos o ser llamados desde el código fuente. Weka contiene herramientas para pre procesamiento de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización, además está estructurado para desarrollar nuevos esquemas de aprendizaje automático [Dan, et al. 2013].

Weka ofrece cuatro interfaces, a las que se puede acceder desde la pantalla de inicio. Cada una de estas interfaces, permite trabajar en un entorno diferente. Asimismo, Weka trabaja con datos provenientes de base de datos, archivos y datos que residen en servidores de Internet [Kirkby & Frank 2005]. Una de las características de la herramienta, es que trabaja con un archivo de formato arff (*Attribute- Relation- File- Format*) cuya estructura está conformada por tres secciones: cabecera, en donde se define el nombre de la relación; declaraciones de atributos, en donde se declaran los atributos que compondrán el archivo junto a su tipo; y sección de datos, en donde se declaran los datos que componen la relación separando entre comas los atributos y con saltos de línea las instancias [Corso & Alfaro]. En la ilustración 1 se muestra la pantalla inicial de la herramienta donde se pueden distinguir las cuatro interfaz que la componen.

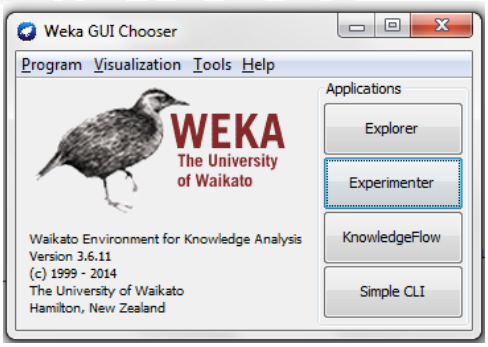


Ilustración 1: Pantalla de inicio de la herramienta Weka

Durante el desarrollo del proyecto se hará uso de la interfaz *Weka Explorer*, la cual permite llevar a cabo la ejecución de algoritmos de análisis sobre un solo archivo de datos. La pantalla principal de la interfaz se visualiza en la ilustración 2.

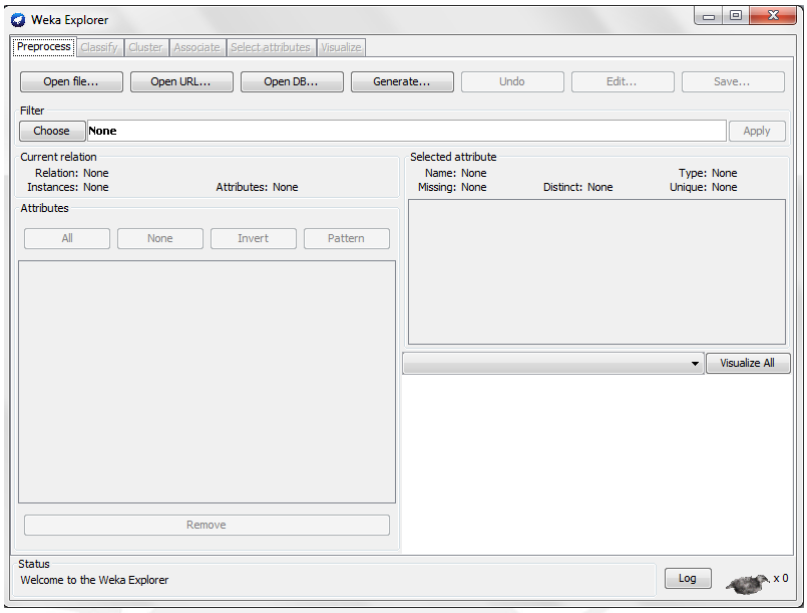


Ilustración 2: Pantalla principal de la interfaz *Weka Explorer*.

En la primera pestaña de Pre procesado, se permite al usuario cargar la fuente de datos sobre la cual se aplicarán las técnicas de Minería de Datos, tales como: clasificación, regresión, agrupamiento, entre otras. La fuente de datos puede ser una base de datos, un archivo Excel en formato .csv o un archivo con formato .arff que la herramienta es capaz de cargar. Además, en esta pestaña Weka permite aplicar una diversidad de filtros sobre los datos, permitiendo realizar transformaciones de todo tipo sobre ellos. Los filtros implementados por Weka son clasificados en filtros de atributos y de instancias. Los primeros, se pueden utilizar para transformar los datos, mientras que los segundos se aplican para eliminar registros o atributos según criterios previamente definidos por el usuario [Corso & Alfaro]. En la ilustración 3 se visualiza la pestaña de Pre procesado una vez cargado un archivo de datos en la herramienta.

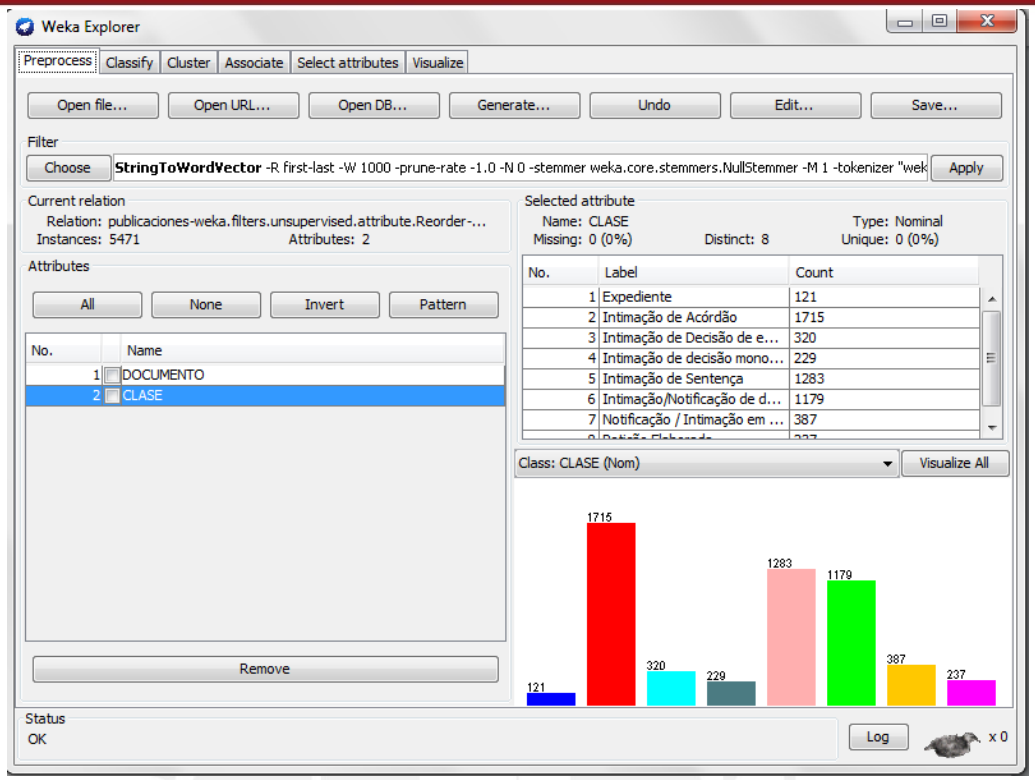


Ilustración 3: Pestaña de Pre Procesado de la interfaz *Weka Explorer*.

Asimismo, se hará uso de la pestaña de Clasificación, la cual permite entrenar y evaluar diversos algoritmos de aprendizaje automático para ejecutar tareas de clasificación o regresión. La herramienta permite configurar los parámetros de los algoritmos, así como definir la técnica de estimación de rendimiento del modelo [Kirkby & Frank 2005]. Finalmente, Weka permite visualizar la confiabilidad del modelo algorítmico construido aplicado sobre los datos de entrada, asimismo detalla la proporción de instancias bien y mal clasificadas; que permite al usuario dar una primera aproximación de qué tan bueno es el modelo [Corso & Alfaro]. En la ilustración 4 se puede visualizar la pestaña de Clasificación para un modelo construido en base al algoritmo de Redes Bayesiana y el conjunto de datos mostrado en la ilustración previa.

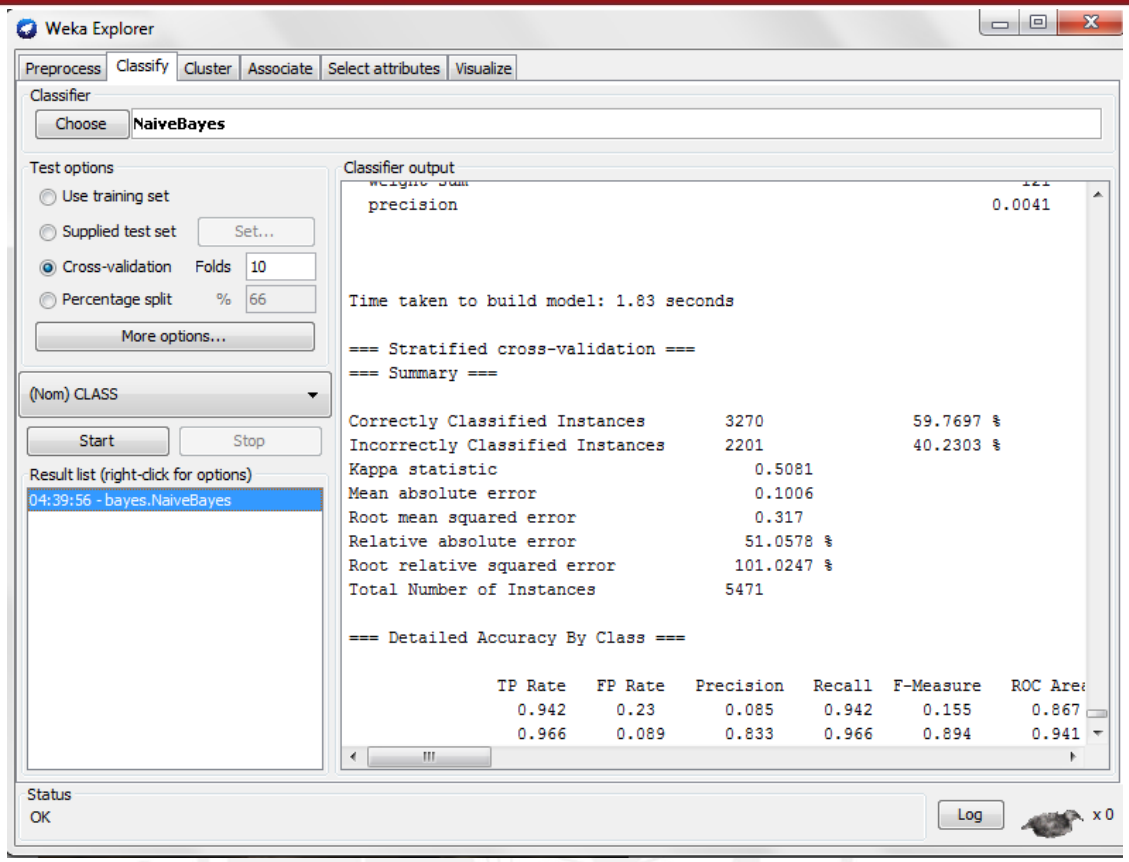


Ilustración 4: Pestaña de Clasificación de la interfaz *Weka Explorer*.

A continuación, se detallarán brevemente los cuatro algoritmos de Minería de Datos aplicables a tareas de clasificación que serán usados durante el proyecto, los cuales están incluidos en la librería Weka.

2.1.1.1 Red Bayesiana

Una red Bayesiana constituye un método basado en teoría probabilística, siendo un modelo de relaciones causales entre eventos.

La red Bayesiana consiste en un conjunto de nodos y arcos, los cuales juntos componen un grafo acíclico dirigido (DAG, por sus siglas en inglés). Los nodos representan variables aleatorias, donde todas tienen un conjunto finito de estados, por otro lado, los arcos indican la existencia de una conexión causal directa entre las variables enlazadas, y la fuerza de estas conexiones es expresada en términos de condiciones probabilísticas [Larranaga, et al. 1996]. A continuación, se presenta un ejemplo ilustrativo del modelo de

red bayesiana aplicado a la predicción de un paciente propenso a padecer cáncer de pulmón.

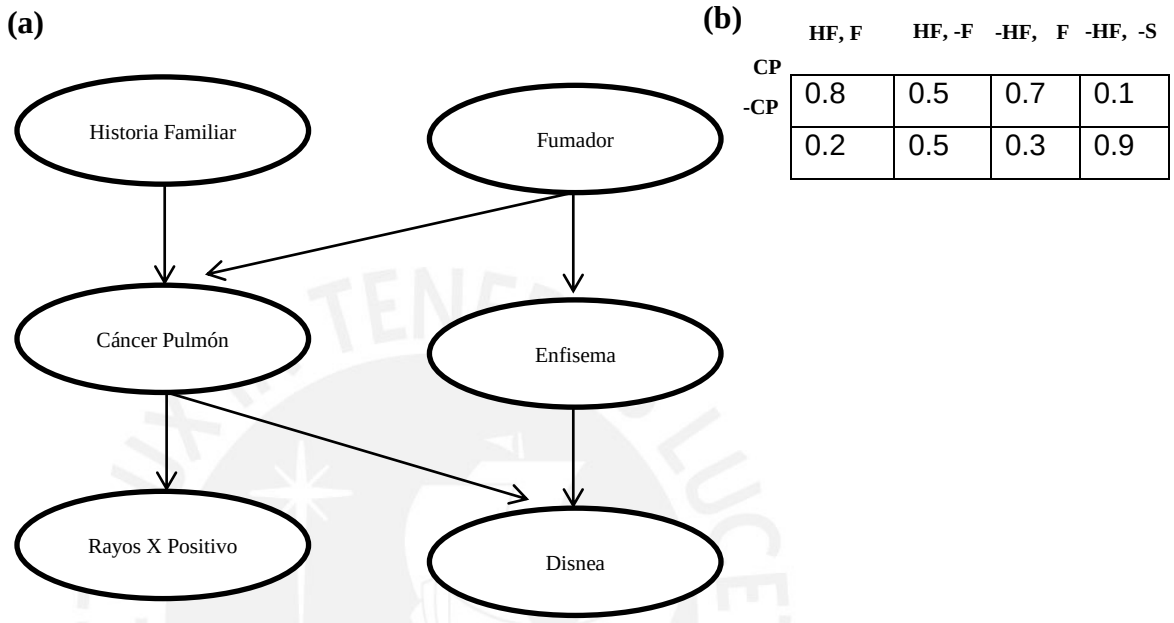


Ilustración 5: Ejemplo simple de una Red Bayesiana: (a) Modelo causal representado como un grafo dirigido acíclico. (b) Tabla de probabilidades condicionales para la variable Cáncer Pulmón [Han & Kamber 2006].

Una red Bayesiana está definida además por un segundo componente, un conjunto de tablas condicionales probabilísticas. Los arcos representan dependencias probabilísticas, si un arco es dibujado del nodo Y al nodo Z, entonces Y es padre o inmediato predecesor de Z, y Z es un descendente de Y, de esta manera cada variable es condicionalmente independiente de sus no descendientes [Han & Kamber 2006].

La construcción de una red Bayesiana consiste en dos sub problemas, el primero de ellos llamado aprendizaje de la estructura consiste en la búsqueda del grafo que mejor refleje todas las relaciones de dependencia entre las variables, y el aprendizaje de parámetros,

el cual consiste en determinar todas las condiciones probabilísticas en la red [Larranaga, et al. 1996].

2.1.1.2 Complemento de la Red Bayesiana

Es una solución heurística a algunos de los problemas que se presentan en los clasificadores basados en Redes Bayesianas, abordando tanto los problemas sistemáticos como los problemas que nacen debido a que los textos no son actualmente generados siguiendo un modelo multinomial. El modelo de Complemento de Red Bayesiana básicamente resuelve dos problemas sistemáticos, el primero en la asignación de pesos pobres para los límites de decisión cuando una clase contiene más instancias de entrenamiento que otra. Por otro lado, las Redes Bayesianas asumen que las características son independientes entre sí, como resultado inclusive cuando las palabras son dependientes unas con otras contribuyen individualmente al modelo, es así como se agregó un método de normalización de los pesos de clasificación con el propósito de evitar que aquellas clases con más dependencias dominen sobre el resto. Adicionalmente, a los problemas sistemáticos las Redes Bayesianas no modelan correctamente los textos, es por ello que se incluyó una simple transformación que coincida más de cerca con la distribución real de frecuencia de las palabras. El nuevo modelo de clasificación se acerca al rendimiento del *Support Vector Machine* además de ser más rápido y fácil de implementar [Rennie, et al. 2003].

*2.1.1.3 kNN (*k-Nearest-Neighbor*)*

El método k-Nearest-Neighbor fue descrito por primera vez al comienzo de los años 50. Está basado en el aprendizaje por analogía es decir compara la tupla de prueba dada con aquellas tuplas de entrenamiento a las cuales se asemeja, cada una de estas tuplas esta descrita por n atributos, siendo cada tupla un punto en un espacio de n dimensiones. Es así como todas las tuplas de entrenamiento son almacenadas en un espacio de patrones de n dimensiones. De esta manera, para cada tupla desconocida dada al modelo, el kNN busca el espacio de k tuplas, siendo k un parámetro configurable desde la herramienta, más cercano a la tupla de prueba, a este conjunto de k tuplas se le denomina el vecindario k más cercano de una tupla desconocida. El modelo define la cercanía en términos de distancia métrica, tal como es la distancia Euclidiana, la cual se define siendo dos tuplas, del tipo X1 = (X11, X12,…, X1n) y X2 = (X21, X22,…, X2n) como:

$$dist(\boldsymbol{X_1}, \boldsymbol{X_2}) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Ilustración 6: Distancia Euclidiana para dos tuplas dadas [Han & Kamber 2006].

En otras palabras, para cada atributo numérico, se toma la diferencia entre los valores correspondiente de ambas tuplas, elevando al cuadrado la misma y acumulándola. La raíz cuadrada es tomada del total de la distancia acumulada [Han & Kamber 2006].

*2.1.1.4 SVM (*Support Vector Machine*)*

Es un método para la clasificación de datos lineales y no lineales. En pocas palabras, el *Support Vector Machine* es un algoritmo que trabaja de la siguiente manera. Este usa un mapeo no lineal para transformar el conjunto de entrenamiento en una dimensión más alta. Dentro de esta nueva dimensión, el SVM busca el hiperplano de separación óptima lineal, siendo este el límite de decisión que separa dos tuplas de una clase de otra. Es así, como con un apropiado mapeo no lineal a una dimensión más grande los datos de dos clases pueden estar siempre separados por un hiperplano. El Support Vector Machine encuentra el hiperplano usando vectores de soporte (esencialmente tuplas de entrenamiento) y márgenes, estos últimos definidos por los vectores de soporte [Han & Kamber 2006].

Sin embargo el tiempo de entrenamiento de inclusive el más rápido SVM puede ser extremadamente lento, por otro lado son de alta precisión debido a su capacidad de modelar complejos límites de decisión no lineales. El método puede ser aplicado tanto a labores de predicción como clasificación [Han & Kamber 2006].

La herramienta será usada para completar las fases de construcción y evaluación de los cuatro algoritmos de aprendizaje automático descritos anteriormente haciendo uso de nuestro conjunto de documentos previamente definido, asimismo se aprovecharan los resultados arrojados durante la etapa de evaluación para sustentar la elección del modelo más apto.

3 Alcance

El presente proyecto corresponde al campo de investigación aplicada y busca seleccionar un modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido. Durante el proyecto, se trabajará con una colección de documentos brindados por una empresa en Brasil encargada de clasificar manualmente los textos emitidos por los tribunales del estado.

El escenario ha sido escogido debido al constante incremento de los volúmenes de información almacenada en documentos virtuales que busca ser mejor gestionada por los usuarios automatizando las labores de clasificación disminuyendo de esta manera las horas hombre involucradas en el proceso.

Finalmente, debido a la diversidad de clases existentes dentro del ámbito judicial, el proyecto se basará en asignar a cada documento una clase dentro de un conjunto finito, en este caso se trabajara únicamente con ocho clases distintas, las cuales serán documentadas más adelante.

4 Justificación

De acuerdo a estudios el 80% de la información de las empresas se encuentra almacenada en documentos, los cuales corresponden a secuencias de palabras no estructuradas, de esta manera para realizar una eficiente y mejor gestión de la información se requiere contar con métodos automatizados que permitan realizar una clasificación de una colección de documentos lo más similar a como lo haría un especialista humano.

Por tal motivo, se busca seleccionar un modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido, de esta manera se trata de resolver una parte implicada en el problema de la gestión de información, la cual es considerada condición indispensable para el éxito de cualquier empresa.

La investigación es conveniente porque contribuirá en mejorar el manejo de documentos, además de disminuir costo y tiempo asociados a la categorización manual de los mismos, ya que para cumplir esta tarea se requiere contar con un experto humano que en base a su experiencia y conocimiento le asigne una clase específica a cada documento. Por otro lado, dada la naturaleza del problema, el proyecto servirá de base para el desarrollo de futuras herramientas de clasificación de documentos en otros contextos.

CAPÍTULO 2

1 Marco Conceptual

En el siguiente apartado se procederá a desarrollar todos los términos y conceptos claves inmersos en la clasificación automática de documentos dentro del ámbito de la Minería de Datos que permitan al lector familiarizarse con el tema del proyecto. Se encontrarán los conceptos básicos, complementarios y específicos relacionados al problema, los cuales han sido obtenidos de la revisión y consulta de fuentes documentales.

1.1 Conceptos de Minería de Datos y relacionados

1.1.1 Datos, Información y Conocimiento

Los datos son individualmente características, atributos o hechos sin ningún significado individualmente [Weiss & Davison 2010]. En este sentido, la información es la colección de hechos organizados y producidos, de tal manera que contengan un valor adicional al valor de los hechos individuales [Stair & Reynolds 2013].

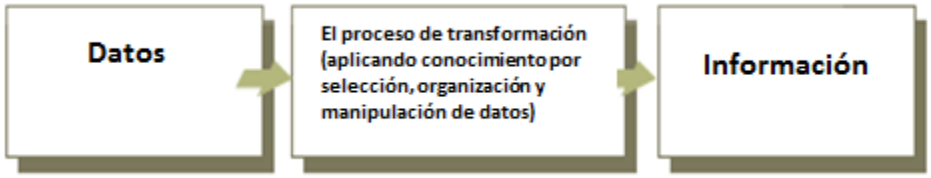


Ilustración 7: Proceso de transformación de datos en información [Stair & Reynolds 2013].

La conversión de los datos a información es un proceso, o un conjunto de tareas ejecutadas para alcanzar un resultado definido. El proceso de definir relaciones en los datos para crear información útil requiere conocimiento. De esta manera, se define al conocimiento como el entendimiento de un conjunto de información y la forma en que esta puede ser usada de manera útil para dar soporte a una tarea o tomar una decisión [Frakes & Baeza-Yates 1992].

1.1.2 Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery from Databases - KDD)

KDD es un proceso no trivial para identificar patrones de datos válidos, originales, potencialmente usables y entendibles en grandes bases de datos, este proceso busca hacer frente a un problema que la era de la información digital ha hecho realidad en la vida de todos nosotros: la sobrecarga de datos [Fayyad, et al. 1996].

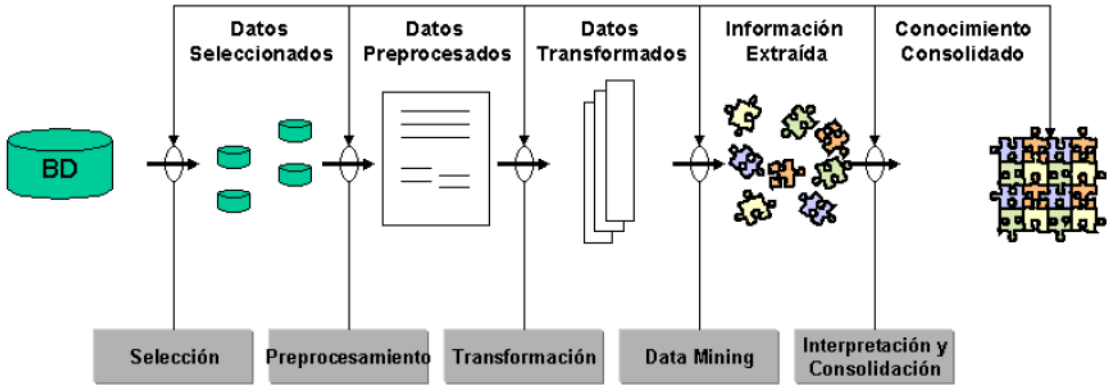


Ilustración 8: Proceso de Descubrimiento de Conocimiento [Maimon & Rokach 2005].

Los pasos definidos anteriormente sirven como guía para cualquier proyecto que esté relacionado con Descubrimiento de Conocimiento en Base de Datos, el cual a veces es confundido, y a la vez es relacionado a la Minería de Datos como un mismo proceso [Maimon & Rokach 2005].

1.1.3 Minería de Datos

La Minería de Datos es un paso en el proceso de *Knowledge Discovery in Data (KDD)*, que consiste en la aplicación de análisis de datos y algoritmos de descubrimiento que producen un conjunto particular de patrones o comportamiento de los datos [Fayyad, et al. 1996]. Muchas personas tratan la Minería de Datos como sinónimo para otro popular término usado, *Knowledge Discovery from Data*, o KDD. Alternativamente, otros aprecian la Minería de Datos como una etapa esencial en el proceso de descubrimiento de conocimiento [Han & Kamber 2006]. Hay muchos métodos de Minería de Datos usados para diversos propósitos y metas.

La taxonomía es llamada para ayudar en el entendimiento de la variedad de métodos, su interrelación y agrupamiento. Es útil distinguir entre dos tipos principales: orientados a verificación, el sistema verifica la hipótesis del usuario, y orientados a descubrimiento, donde el sistema encuentra nuevas reglas y patrones autónomos [Maimon & Rokach 2005].

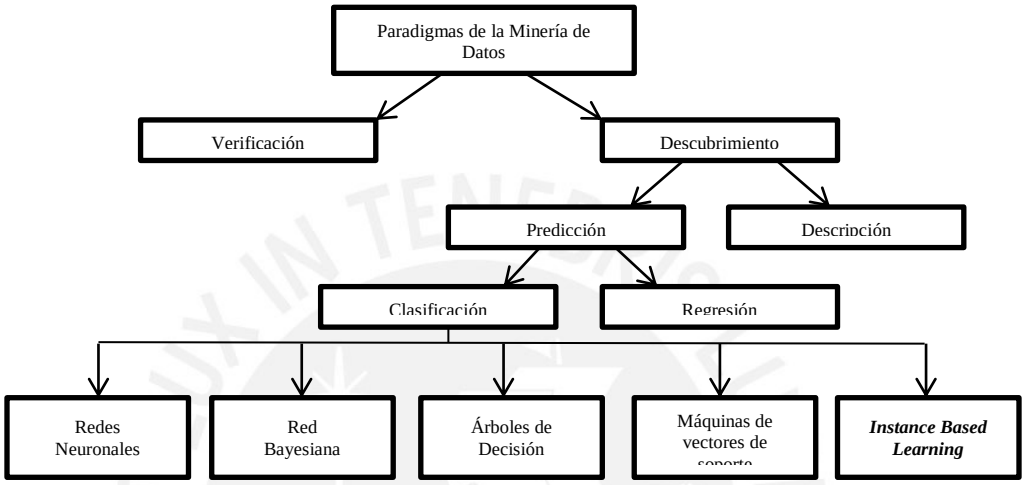


Ilustración 9: Taxonomía de la Minería de Datos [Maimon & Rokach 2005].

Los métodos de descubrimiento son aquellos que automáticamente identifican patrones en los datos. Estos se dividen en métodos de predicción y métodos de descripción.

Los métodos descriptivos buscan patrones en el conjunto de datos con el fin de presentárselos al usuario en una forma entendible para su posterior interpretación [Fayyad, et al. 1996]. [Maimon & Rokach 2005]En primer lugar, los métodos de predicción son aquellos que tienen el objetivo específico de permitirnos predecir el valor de ciertas características de un objeto en base a los valores observados en características de otros objetos [Hand, et al. 2001]. Por otro lado, los métodos de verificación trabajan en la evaluación de una hipótesis brindada por un recurso externo, estos están menos asociados a la Minería de Datos debido a que la mayoría de problemas en el área conciernen en descubrir una hipótesis, más que en probar una. Es además útil distinguir entre dos modelos principales de predicción: clasificación y regresión. Estos asignan al espacio de entrada dentro de un dominio de valor real. Por ejemplo, un regresor puede predecir la demanda para un cierto producto dadas sus características, mientras que un clasificador puede ser usado para clasificar un cliente hipotecario como bueno o malo [Maimon & Rokach 2005].

1.1.4 Clasificación o Categorización

Clasificación es una tarea encargada de mapear o clasificar un elemento del conjunto de datos dentro de un grupo de clases previamente definidas [Fayyad, et al. 1996]. Por otro lado, es definida como el proceso de dividir los datos en un número de grupos independientes o dependientes unos con otros y en donde cada grupo se comporta como una clase [Joshi & Nigam 2011].

Aplicado al dominio de manejo de documentos, la tarea es conocida como Categorización de Textos (*TC, text categorization*), consiste en el proceso de dada un conjunto de categorías (asunto, temas) y una colección de documentos encontrar el correcto tema o temas para cada uno de los documentos [Feldman & Sanger 2007]. También es definido como la asignación de categorias a nuevos documentos basado en el aprendizaje ganado por los sistemas de clasificación en su etapa de entrenamiento [Antonie & Zaiane 2002].

Como muchas de las tareas de Inteligencia Artificial, hay dos enfoques principales inmerso en la categorización de texto: El primero es la Ingeniería del conocimiento (*Knowledge Engineering*) basado en que el conocimiento de los expertos acerca de las categorias se codifica directamente en el sistema mediante reglas. El otro es el Aprendizaje Autómatico (*ML, Machine Learning*) en el que un proceso inductivo contruye un clasificador mediante el aprendizaje de un conjunto de ejemplos [Feldman & Sanger 2007].

1.1.5 Minería de Textos (Text Mining)

La Minería de Textos es un descubrimiento automático de información previamente desconocida, cuyo proceso empieza con la extracción de hechos y eventos de fuentes de texto con el propósito de formular hipótesis que son exploradas por métodos tradiciones de Minería de Datos y Análisis de Datos [Feldman & Sanger 2007].

Su objetivo es derivar el conocimiento implícito que se esconde en textos no estructurados, para ello se divide el proceso en cuatro fases: recuperación de información, extracción de información, descubrimiento del conocimiento y generación de hipótesis [Zhu, et al. 2013].

Estableciendo una analogía con la Minería de Datos, la Minería de Textos busca extraer información útil de un conjunto de datos a través de la identificación y exploración de

patrones. En el caso de Minería de Datos, sin embargo, el conjunto de datos son una colección de documentos, y los patrones son encontrados no en registros formalizados de una base de datos sino de documentes no estructurados.

La Minería de Datos asume que los datos están almacenados en una estructura formalizada, la labor de pre procesamiento se centra únicamente en dos tareas: depuración y normalización de los datos. Por el otro lado, para la Minería de Textos, dicha tarea se centra en la identificación y extracción de características representativas en el lenguaje natural de los documentos [Feldman & Sanger 2007].

1.1.6 Colección de documentos (Document Collection)

La colección de documentos es uno de los elementos principales en la Minería de Textos. Se define como cualquier agrupación de documentos basados en texto, los cuales pueden ser de dos tipos: estático, cuando la colección inicial de documentos permanece sin cambios, o dinámico, aplicado a la colección de documentos caracterizado por la inclusión de nuevos o actualizados documentos alrededor del tiempo.

Un ejemplo típico de colección de documentos como entrada inicial para la Minería de Datos es *PubMed*, la cual es considerada la colección en línea más comprensiva de investigaciones biomédicas publicadas en el idioma inglés [Feldman & Sanger 2007].

1.1.7 Recuperación de la Información (Information Retrieval)

Es la rama de la ciencia de la computación que se encarga del estudio de los procesos de búsqueda eficiente de información relativa a un tema en particular en grandes volúmenes de documentación. La recuperación de información se diferencia de la recuperación de datos, en que la primera trata con textos en lenguaje natural, los cuales en la mayoría de casos no se encuentra estructurado y en consecuencia semánticamente ambiguo; en cambio la recuperación de datos trabaja con datos estructurados y semántica definida [Baeza-Yates & Ribeiro-Neto 1999].

1.1.8 Extracción de Información (Information extraction - IE)

Es un proceso que requiere como entrada documentos no revisados y produce una estructura tabulada como salida y tiene como objetivo la formulación de un conjunto de reglas efectivas que permitan el reconocimiento de la información relevante para el usuario [Jung, et al. 2005].

La extracción de información se diferencia de la recuperación de la información en que los sistemas de recuperación de información encuentran textos relevantes y los presenta al usuario, mientras que las aplicaciones de extracción de información analizan los textos y brindan al usuario únicamente la información que a este le interesa [Cunningham 2005].

1.1.9 Reconocimiento de Patrones

El reconocimiento de patrones es una disciplina que trata con el problema del desarrollo de algoritmos y metodologías que sean capaces de implementar computacionalmente muchas de las tareas de reconocimiento que los seres humanos normalmente realizamos. La motivación recae en realizar dichas tareas de manera más precisa, o rápida, e inclusive más económica que los humanos. En este camino, su meta es idear formas de automatización para ciertos procesos de toma de decisiones que desembocan en la clasificación y reconocimiento [Pal & Mitra 2004].

1.2 Conclusión

La categorización automática de documentos es un problema que abarca una gran cantidad de conceptos de diversas índoles y ha venido siendo tratada desde hace muchos años por diversos especialistas dentro de las áreas de Minería de Datos y Minería de Textos, cuyos planteamientos en diversas fuentes documentales servirán de base para la elaboración del proyecto planteado.

Es importante para el lector comprender y familiarizarse con los conceptos previamente expuestos para una mejor comprensión del problema debido a que serán usados con frecuencia a lo largo del documento.

2 Estado del arte

En el siguiente apartado está sustentado en la revisión de literatura para el problema de clasificación automática de documentos brindando información relacionada a proyectos, investigaciones y productos que han sido realizados en los últimos años.

El objetivo principal es de brindar un enfoque global de las diversas investigaciones que se han realizado hasta el momento en relación al tema a tratar en el presente proyecto de fin de carrera, con el propósito de brindar una noción al lector del estado actual del tema,

desde sus limitaciones hasta sus alcances, así como resumir el conocimiento y conclusiones obtenidas hasta el momento por los investigadores.

2.1 Método usado en la revisión

El método usado para la revisión del estado del arte es la revisión sistemática, la cual es usada para recopilar y evaluar la evidencia disponible relacionada a un tema en particular. A diferencia del proceso usual de revisión de literatura, la revisión sistemática, como su nombre lo denota, es desarrollada de una manera formal y sistemática. De esta manera, el proceso de conducción de la investigación de este tipo sigue un conjunto bien definido y estricto de pasos metodológicos, según el desarrollo de un protocolo, el cual es construido alrededor de una regla central, la cual representa el núcleo de la investigación y es expresado usando conceptos y términos específicos que debe direccionarnos hacia información relacionada a una pregunta específica, pre-definida y estructurada [Biolchini, et al. 2005].

La revisión sistemática se llevó a cabo el 21 de Abril del 2014 en el buscador IEEE Explore.

2.1.1 Formulación de la pregunta

Para la revisión sistemática se formuló la siguiente pregunta: ¿Qué investigaciones acerca de clasificación automática de documentos en el campo de la Minería de Datos se han realizado recientemente?, para ello se identificaron las siguientes palabras claves: "*document classification*", "*document categorization*", "*document*", "*data mining*" y "*text mining*".

2.1.2 Selección de las fuentes

A partir del agrupamiento de las palabras claves, que se mencionaron en el punto anterior y haciendo uso de los operadores lógicos, se llevó a cabo la revisión con la siguiente cadena: ("*document classification*" *OR* "*document categorization*") *AND* ("*data mining*" OR "*text mining*")

Entre los detalles importantes de la búsqueda, el criterio de exclusión empleado se ha enfocado en la fecha de publicación, de esta forma, únicamente se consideraron aquellos documentos cuyo año de publicación va desde el 2010 hasta la actualidad. Por otro lado,

como criterio de inclusión se tomó en cuenta la presencia de las palabras claves previamente mencionadas dentro del resumen y título.

A partir de los criterios mencionados de exclusión e inclusión y aplicando la cadena de búsqueda en IEEE Explore se encontraron 56 resultados, de los cuales 20 han ayudado a responder la pregunta planteada en un inicio.

2.2 Productos similares

En este apartado se procederá a brindar una referencia general acerca de productos desarrollados con un propósito similar al del presente proyecto de fin de carrera.

2.2.1 ADaM (*Algorithm Development and Mining*)

ADaM es una herramienta para la Minería de Datos diseñada para el uso con datos científicos, de esta manera ha sido usado en varios proyectos de investigación, con el objetivo de explotar un gran conjunto variado de datos científicos usando diferentes metodologías de procesamiento. ADaM provee un conjunto de herramientas por cada uno de los procesos básicos de Minería de Datos, tales como: clasificación, agrupamiento, reglas de asociación y pre procesamiento [Rushing, et al. 2005].

2.2.2 KEEL *(Knowledge Extraction based on Evoluationary Learning)*

KEEL (*Knowledge Extraction based on Evoluationary Learning*) es un *open source* implementado en lenguaje Java el cual permite al usuario evaluar el comportamiento del aprendizaje evolutivo y *Soft Computing* basado en técnicas aplicadas para diversos problemas en la rama de la Minería de Datos, tales como: regresión, clasificación, agrupamiento y patrones de explotación [Alcalá-Fdez, et al. 2009].

KEEL ha sido diseñado con dos propósitos: investigación y educativo, y ofrece el siguiente conjunto de ventajas: reduce el trabajo de programación al incluir una librería con algoritmos de aprendizaje evolutivo basados en diversos paradigmas simplificando la integración de los algoritmos de aprendizaje evolutivo con las diversas técnicas de pre-procesamiento, extiende el rango de posibles usuarios, investigadores con menor conocimiento en computación evolutiva, a través del *framework*, podrán ser capaces de aplicar satisfactoriamente los algoritmos a sus problemas y cualquier usuario puede usar KEEL en su máquina independientemente del sistema operativo [Alcalá-Fdez, et al. 2009].

2.3 Investigaciones acerca del tema

2.3.1 Algoritmo de clasificación de textos kNN (*K* nearest neighbor) basado en aprendizaje ansioso

En esta investigación se introduce un algoritmo mejorado de clasificación de textos basado en el aprendizaje ansioso (*eager learning*) con el algoritmo básico de clasificación kNN (*K nearest neighbor*) con el objetivo de superar una de las deficiencias de este último: el aprendizaje perezoso (*lazy learning*), el cual consiste en almacenar una muestra de entrenamiento dada únicamente cuando la muestra de prueba llega, se comienza a categorizar en base a las similitudes con las muestras de entrenamiento. De esta manera, se emplea menos tiempo en el entrenamiento, pero más tiempo en la predicción, sin embargo, la inclusión de un aprendizaje ansioso (*eager learning*) permite dada la muestra de entrenamiento construir un modelo de categorización antes de recibir las muestras de pruebas disminuyendo el tiempo requerido en las predicciones.

Los resultados experimentales realizados comprueban que el algoritmo introducido decremento el costo computacional y mejora la eficiencia del kNN tradicional, siendo un método aplicable, factible y efectivo [Dong, et al. 2012].

2.3.2 Sistema de agente múltiple para la clasificación de documentos

En este proyecto se introduce el concepto de sistema de agente múltiple (*Multi-Agent system – MAS*) aplicado a la clasificación de documentos, MAS es un sistema basado en múltiples agentes de inteligencia que interactúan entre sí y con el entorno para mejorar sus capacidades de aprendizaje. Es así como puede ser usado para resolver problemas que pueden resultar difíciles o imposibles de resolver para un único agente.

El sistema propuesto usa un clasificador *Naive Bayes,* el cual calcula la posibilidad máxima y mínima de que un documento pertenezca a cierta clase, el sistema trabaja en un ambiente de distribución donde cada agente es autónomo en la clasificación de textos y puede compartir su información con otros agentes cuando es necesario.

El sistema de agente múltiple trabaja de manera más rápida en comparación al enfoque jerárquico, además permite distribuir el procesamiento de los documentos de manera paralela, en base a los resultados obtenidos en la investigación se puede afirmar que el nuevo enfoque tiene un rendimiento aceptable en términos de tiempo de ejecución y precisión [Ahmad, et al. 2012].

2.3.3 Mejora en el método de clasificación *Support Vector Machine* usando la función de distancia Euclidiana para la categorización de documentos

La siguiente investigación presenta un nuevo marco para la clasificación de documentos de texto basado en el Support Vector Machine (SVM) y la función de distancia Euclidiana, la primera es usada en la fase de entrenamiento para identificar el conjunto de vectores asociados para cada categoría, mientras que la segunda en la fase de clasificación para calcular las distancias entre el elemento de prueba y los vectores de categorías cuyo valor definirá la decisión a tomar por el clasificador, el mencionado nuevo marco se denomina Euclidean-SVM [Lee, et al. 2012].

2.4 Conclusiones sobre el estado del arte

En los puntos anteriores se ha brindado información relacionada a proyectos, investigaciones y productos que han sido realizados en los últimos años por diversos investigadores del área. En cuanto a las herramientas existentes, se pueden encontrar productos de similares características a las de la herramienta Weka, la cual será usada como base para nuestro proyecto, ambos productos le dan la libertad al usuario de probar los diversos modelos de clasificación disponibles sobre un conjunto de datos definido por uno mismo. Asimismo, debido a que un modelo de aprendizaje se construye en base a un conjunto de datos de entrenamiento definido no se han encontrado en el mercado herramientas iguales a las que el presente proyecto plantea, sin embargo si existen productos capaces de resolver la categorización de documentos pero aplicados a otros escenarios, tal como: clasificación de reseñas de películas, clasificación de publicaciones en diarios, entre otros.

En cuanto a las investigaciones sobre el problema matriz, la clasificación de documentos, se han encontrado diversos avances durante los últimos años relacionados a la aplicación de modelos de clasificación sobre un conjunto estándar de prueba, lo cual demuestra la diversidad de posibles modelos aplicables al escenario. Es por ello, que no se podría definir en base a la literatura que modelo de clasificación es el más adecuado para nuestro problema, sin embargo nos da la seguridad de que el problema en cuestión se puede resolver según lo planteado en puntos anteriores.

CAPÍTULO 3: Selección del conjunto de documentos

En el presente capitulo profundizará sobre el primer objetivo específico de nuestro proyecto de fin de carrera detallando los resultados obtenidos para este.

El presente objetivo tiene como propósito estructurar el conjunto de documentos que será usado para la construcción y evaluación del modelo de clasificación bajo el formato aceptado por la herramienta Weka, la cual será usada para completar ambas fases mencionadas. Es así como este objetivo se ha dividido en dos etapas, la primera de ellas corresponde a la recopilación y posterior estructuración de la información extraída desde la base de datos de la empresa cliente, mientras que el segundo abarca el pre-procesamiento de la misma mediante el uso de las funciones de manejo de cadenas disponibles en Weka obteniendo un conjunto de datos aplicable a los modelos de clasificación.

1 Recopilación y estructuración del conjunto de datos

El presente proyecto de fin de carrera será aplicado a un conjunto de documentos de carácter legal en idioma portugués, los cuales han sido brindados por una empresa en Brasil encargada de la clasificación manual de los mismos. Actualmente, la empresa cuenta con un gran número de instancias ya clasificadas por especialistas en el área, las cuales han sido recopiladas, sintetizadas y estructuradas en el siguiente documento en formato Excel, donde la columna A almacena la clase asignada al documento mientras que la columna B contiene la cadena de palabras que conforman el texto. Es importante mencionar que el formato aceptado por la herramienta Weka difiere de la hoja de cálculo de Excel, y será detallado más adelante.

A	B
CLASE	DOCUMENTO
Expediente	Processo 053.09.011884 0 Mandado de Seguranca Maria Jose Barreto da Silva Dirigente Regional de Ensino da Diretoria de Ensino Sul 3 Vistos. Ante as infc
Expediente	Processo 053.08.609268 8 Procedimento Ordinario (em geral) MARCOS ALBERTO DE LIMA Fazenda Publica do Estado de Sao Paulo Indiquem e justifiquem a
Expediente	Processo 053.09.006141 4 Procedimento Ordinario (em geral) Ieda Maria Vieira e outros Instituto de Previdencia do Estado de Sao Paulo IPESP nota de car
Expediente	Processo 053.09.014776 9 Procedimento Ordinario (em geral) Celso Ramos da Silva INSTITUTO DE PREVIDENCIA DO ESTADO DE SAO PAULO IPESP Fls.56
Expediente	Processo 053.09.010113 0 Procedimento Ordinario (em geral) Nair Gomes de Assis e outros Fazenda do Estado de Sao Paulo FESP Vistos Recebo a(s) apela
Expediente	Processo 053.09.017054 0 Procedimento Ordinario (em geral) Jandira Ogane Simao e outros Fazenda do Estado de Sao Paulo Certifico e dou fe que nos ter
Expediente	Processo 053.09.015531 1 Procedimento Ordinario (em geral) Adavenisa Gomes de Holanda Adao e outros Fazenda do Estado de Sao Paulo FESP (Ao petici
Expediente	920.295.5/1 SAO PAULO FAZ PUBLICA REL. DES. URBANO RUIZ AGTE(S): JOSE VICENTE ROSA E JOSE VIEIRA DE CAMPOS E NEYDE ZORZI CARCHEDI E PAULIN
Expediente	EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N 0534/2009Processo 053.09.010826 7 Procedimento Ordinario (em geral) Silas Pereira da Costa e outros
Expediente	EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N 0538/2009Processo 053.09.016074 9 Declaratoria (em geral) Zeferino Batista Camargo Fazenda Publica
Expediente	EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N 0546/2009Processo 053.08.615908 1 Procedimento Ordinario (em geral) Ademar Pinheiro Sotta e outros
Expediente	JULGAMENTOS SEÇÃO DE PROCESSAMENTO DA 5ª CÂMARA DE DIREITO PÚBLICO SALA 203 SESSÃO ORDINÁRIA DA 5ª CÂMARA DE DIREITO PÚBLICO. REALIZAI
Expediente	EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N 0790/2009Processo 053.09.012991 4 Procedimento Ordinario (em geral) Francisco Pereira Sobrinho Faz
Expediente	EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N 0237/2009Processo 053.09.009646 3 Declaratoria (em geral) Marcelo Antonio Scatena Franco Fazenda

Ilustración 10: Conjunto de datos en hojas de Excel

1.1 Clases y atributos dentro del conjunto de datos

El presente proyecto está destinado a trabajar en base a una colección de documentos en portugués, los cuales están clasificados sobre 8 clases, las cuales son presentadas en la tabla 2, por otro lado únicamente se contará con dos atributos que representarán al elemento: la cadena de texto correspondiente al documento y la clase asignada para la cadena mencionada.

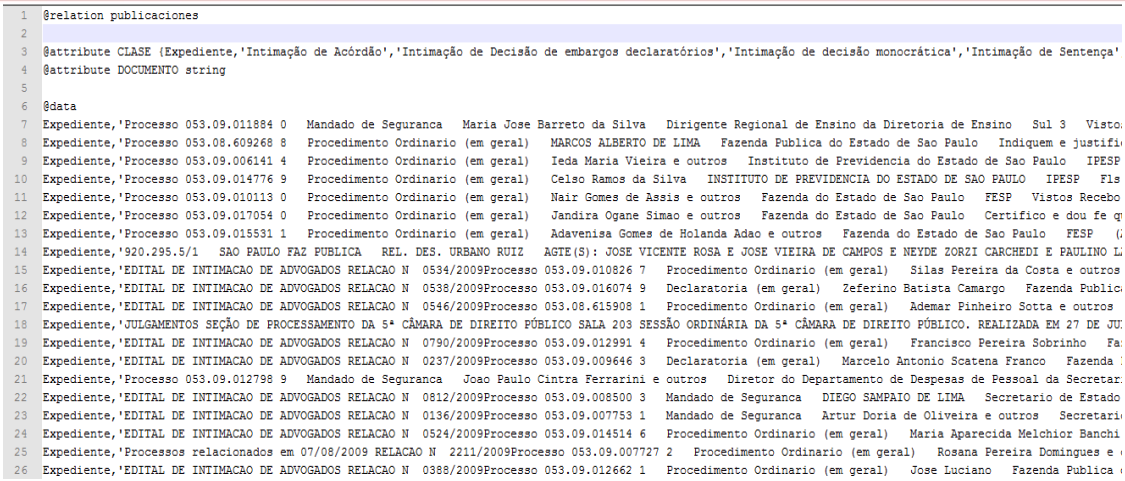
Tabla 2: Clases de trabajo

Clase	Nombre	Número de instancias
1	Expediente	121
2	Intimação de Acórdão	1715
3	Intimação de Decisão de embargos declaratórios	320
4	Intimação de decisão monocrática	229
5	Intimação de Sentença	1283
6	Intimação/Notificação de despacho	1179
7	Notificação / Intimação em Mandado de Segurança	387
8	Petição Elaborada	237

1.2 Estructura del archivo

El archivo a utilizar debe estar estructurado en base al formato aceptado por la herramienta Weka; la cual será empleada para la construcción y evaluación de los modelos de clasificación. A continuación se presenta la estructura del archivo aceptado por la herramienta, el cual debe ser de extensión seguir la extensión arff.

- **@Relation** <nombre-relación> (línea 1), todo archivo .arff debe comenzar con esta primera declaración.
- **@Attribute** <nombre-atributo> <tipoDato> (línea 3 hasta línea 4), en esta sección se incluye una línea por cada atributo que se vaya a incluir en el conjunto de datos, indicando su nombre y el tipo de dato.
- **@Data** (a partir de la línea 6), en esta sección se incluyen los datos, cada columna es separada por una coma, el número de columnas debe coincidir con el número de atributos previamente definidos.



```
@relation publicaciones

@attribute CLASE {Expediente,'Intimação de Acórdão','Intimação de Decisão de embargos declaratórios','Intimação de decisão monocrática','Intimação de Sentença'
@attribute DOCUMENTO string

@data
Expediente,'Processo 053.09.011884 0   Mandado de Seguranca   Maria Jose Barreto da Silva   Dirigente Regional de Ensino da Diretoria de Ensino   Sul 3   Visto
Expediente,'Processo 053.08.609268 8   Procedimento Ordinario (em geral)   MARCOS ALBERTO DE LIMA   Fazenda Publica do Estado de Sao Paulo   Indiquem e justifi
Expediente,'Processo 053.09.006141 4   Procedimento Ordinario (em geral)   Ieda Maria Vieira e outros   Instituto de Previdencia do Estado de Sao Paulo   IPESP
Expediente,'Processo 053.09.014776 9   Procedimento Ordinario (em geral)   Celso Ramos da Silva   INSTITUTO DE PREVIDENCIA DO ESTADO DE SAO PAULO   IPESP   Fls
Expediente,'Processo 053.09.010113 0   Procedimento Ordinario (em geral)   Nair Gomes de Assis e outros   Fazenda do Estado de Sao Paulo   FESP   Vistos Recebo
Expediente,'Processo 053.09.017054 0   Procedimento Ordinario (em geral)   Jandira Ogane Simao e outros   Fazenda do Estado de Sao Paulo   Certifico e dou fe q
Expediente,'Processo 053.09.015531 1   Procedimento Ordinario (em geral)   Adavenisa Gomes de Holanda Adao e outros   Fazenda do Estado de Sao Paulo   FESP   (
Expediente,'920.295.5/1   SAO PAULO FAZ PUBLICA   REL. DES. URBANO RUIZ   AGTE(S): JOSE VICENTE ROSA E JOSE VIEIRA DE CAMPOS E NEYDE ZORZI CARCHEDI E PAULINO L
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0534/2009Processo 053.09.010826 7   Procedimento Ordinario (em geral)   Silas Pereira da Costa e outros
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0538/2009Processo 053.09.016074 9   Declaratoria (em geral)   Zeferino Batista Camargo   Fazenda Public
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0546/2009Processo 053.08.615908 1   Procedimento Ordinario (em geral)   Ademar Pinheiro Sotta e outros
Expediente,'JULGAMENTOS SEÇÃO DE PROCESSAMENTO DA 5ª CÂMARA DE DIREITO PÚBLICO SALA 203 SESSÃO ORDINÁRIA DA 5ª CÂMARA DE DIREITO PÚBLICO. REALIZADA EM 27 DE JU
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0790/2009Processo 053.09.012991 4   Procedimento Ordinario (em geral)   Francisco Pereira Sobrinho   Fa
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0237/2009Processo 053.09.009646 3   Declaratoria (em geral)   Marcelo Antonio Scatena Franco   Fazenda
Expediente,'Processo 053.09.012798 9   Mandado de Seguranca   Joao Paulo Cintra Ferrarini e outros   Diretor do Departamento de Despesas de Pessoal da Secretar
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0812/2009Processo 053.09.008500 3   Mandado de Seguranca   DIEGO SAMPAIO DE LIMA   Secretario de Estado
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0136/2009Processo 053.09.007753 1   Mandado de Seguranca   Artur Doria de Oliveira e outros   Secretari
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0524/2009Processo 053.09.014514 6   Procedimento Ordinario (em geral)   Maria Aparecida Melchior Banchi
Expediente,'Processos relacionados em 07/08/2009 RELACAO N  2211/2009Processo 053.09.007727 2   Procedimento Ordinario (em geral)   Rosana Pereira Domingues e
Expediente,'EDITAL DE INTIMACAO DE ADVOGADOS RELACAO N  0388/2009Processo 053.09.012662 1   Procedimento Ordinario (em geral)   Jose Luciano   Fazenda Publica
```

Ilustración 11: Conjunto de datos estructurado sin pre-procesar

2 Pre-procesamiento del conjunto de datos

El objetivo de esta segunda etapa es la obtención de un conjunto de datos aplicable a los modelos de clasificación disponibles en la herramienta de trabajo Weka, dado que dichos modelos únicamente admiten atributos de tipo numérico es necesario procesar mediante funciones de manejo de cadenas el archivo previamente mostrado. La función que ha sido empleada en este caso es *StringToWordVector*, la cual se encuentra disponible en el Weka Explorer y se encarga de convertir un atributo de cadena en un conjunto de atributos que represente la información de ocurrencia de las palabras dentro del texto, es importante detallar que la dimensión del diccionario de palabras, la técnica de estimación de pesos por palabras, y otros atributos correspondientes a la función son definidos como parámetros.

2.1 Técnica de estimación de peso por palabra

La técnica de estimación de peso por palabra a ser usada para el pre-procesamiento del texto es TF-IDF. Esencialmente, trabaja determinando la frecuencia relativa de la palabra en un documento específico comparado a la proporción inversa de dicha palabra en la colección total de documentos [Ramos 2003]. La principal razón por la cual se ha decidido aplicar esta técnica es debido a que permite controlar los pesos de aquellas palabras que generalmente son más comunes que otras.

2.2 Aplicando StringToWordVector al conjunto de datos

La aplicación de la función sobre el conjunto de datos devolverá un nuevo archivo plano estructurado, que será aplicable directamente a los modelos de clasificación de textos incluidos en la herramienta. La *StringToWordVector* convierte cada palabra incluida en los documentos, en un nuevo atributo de tipo numérico cuyo valor se define según la fórmula de transformación del vector definida en los parámetros de la función. A continuación, se muestra la estructura y datos del archivo procesado, el cual sigue el formato detallado en el punto anterior.

```
@relation publicaciones

@attribute às numeric
@attribute § numeric
@attribute xvi numeric
@attribute xiv numeric
@attribute www numeric
@attribute walvis numeric
@attribute wagner numeric
@attribute voto numeric
@attribute volume numeric
@attribute vistos numeric
@attribute vista numeric
@attribute vinte numeric
@attribute vigencia numeric
@attribute vieira numeric
@attribute vida numeric
@attribute vicente numeric
@attribute vi numeric
@attribute vez numeric
@attribute verbas numeric
@attribute adicional numeric
@attribute adicionais numeric
@attribute adiados numeric
@attribute acrescimos numeric
@attribute acrescidas numeric
@attribute acordo numeric
@attribute acordaon numeric
@attribute acordao numeric
@attribute acoes numeric
@attribute acima numeric
@attribute acesso numeric
@attribute acerca numeric
@attribute acao numeric
@attribute CLASS {Expediente,'Intimação de Acórdão','Intimação de Decisão de embargos declaratórios',

@data
{9 1.615964,93 2.594484,99 1.605671,126 2.908091,186 2.979557,242 1.677028,256 2.829469,345 0.67978,3
{91 2.576081,136 0.117144,227 1.550001,231 4.483232,242 1.849074,244 1.409763,263 2.399196,304 0.1121
{13 4.065172,136 0.120103,170 3.181635,242 1.895783,244 1.445374,253 3.953095,263 2.4598,290 2.682235
{91 1.995084,99 1.371111,136 0.090724,147 3.211045,212 3.734006,242 1.432043,244 1.091812,253 2.98610
{9 1.27564,42 1.758737,72 3.475351,79 2.937465,133 2.571376,134 0.708358,136 0.132929,201 2.379623,21
{1 3.101926,55 1.941063,91 2.394732,130 4.586838,136 0.108897,137 2.122608,175 1.385013,242 1.718904,
{9 1.431834,81 2.089075,97 3.235611,121 1.938079,134 1.590183,136 0.094138,139 3.892027,146 2.487245,
{13 3.930921,15 4.404538,86 2.669311,94 3.952682,136 0.184072,142 1.327598,145 3.798079,171 4.582644,
```

.

Ilustración 12: Conjunto de datos estructurado pre-procesado

3 Conclusión

La definición del conjunto de datos constituye el punto de partida para la elaboración total del proyecto, es importante detallar que los resultados obtenidos a partir de este punto en adelante están totalmente ligados a los datos definidos en este capítulo, debido a que los modelos predictivos se construyen y definen sus reglas en base a ellos. De esta manera, se recalca la relevancia de contar con información confiable y elaborada por expertos humanos.

CAPÍTULO 4: Adaptación y evaluación de los modelos de clasificación

En el presente capitulo profundizará sobre el segundo objetivo específico del proyecto, el cual tiene como propósito completar las etapas de construcción y evaluación de los cuatro algoritmos de clasificación, documentados previamente para el lector, para extraer de estas los indicadores de precisión que sustente la posterior selección del modelo algorítmico más apto.

1 Técnica de estimación de rendimiento de modelos predictivos

La técnica de estimación de rendimiento a ser usada para el presente proyecto es la validación cruzada, la cual es una técnica que divide el conjunto de datos de manera aleatoria en k partes exclusivas de aproximadamente el mismo tamaño. Es decir, siendo nuestro conjunto de datos D, la validación cruzada divide este en k partes de la siguiente manera: D1, D2, D3,…, Dk. El modelo es construido y evaluado k veces, siendo construido usando D/Di instancias y evaluado en Di instancias. La precisión del modelo es estimada como el total de instancias correctamente clasificadas sobre el total de instancias en el conjunto de datos [Kohavi 1995]. La razón principal por la cual se ha decido usar esta técnica de estimación, es que es capaz de entrenar al modelo con todo el conjunto de datos, y a la vez evaluarlo indirectamente con la totalidad del mismo.

En el proyecto se ha definido a la variable k con un valor igual a 25, dado que para este valor se ha obtenido un mayor porcentaje de instancias clasificadas correctamente, es importante recalcar que la calibración se ha realizado con la totalidad del conjunto de datos, es decir 5471 instancias. A continuación, se muestra una tabla resumen con los resultados de las pruebas de calibración realizadas para la selección del valor más apto de k.

Tabla 3: Cuadro resumen de los resultados para las pruebas de selección del valor de k para la técnica de validación cruzada.

Valor de la variable k	Red Bayesiana (Instancias clasificadas correctamente)	CNB (Instancias clasificadas correctamente)	SVM (Instancias clasificadas correctamente)	kNN (Instancias clasificadas correctamente)
k = 5	60.080%	72.308%	84.280%	79.144%
k = 10	59.769%	72.144%	84.883%	79.638%
k = 15	59.897%	72.089%	84.993%	79.930%
k = 20	59.842%	72.180%	85.103%	79.894%
k = 25	59.824%	72.180%	85.267%	79.839%
k = 30	59.550%	72.198%	85.158%	79.820%

2 Adaptación y evaluación de los modelos de clasificación

La clasificación de datos, es un proceso de dos etapas. En primer lugar, un clasificador es construido describiendo un conjunto de datos y conceptos predeterminados. Esta es la etapa de entrenamiento, donde un algoritmo de clasificación construye un modelo analizando o aprendiendo de un conjunto de datos de entrenamiento compuesto por tuplas de datos y su clase asociada, es así como también puede ser visto como el aprendizaje de una función, $y = f(X)$, que pueda predecir la clase asociada y dada una tupla X. Por otro lado, en la segunda etapa el modelo es usado para la clasificación con el objetivo de evaluar su rendimiento. El rendimiento de un clasificador está dado por el porcentaje de tuplas correctamente clasificadas por el modelo [Han & Kamber 2006].

En los puntos siguientes se procederá a brindar al lector el conjunto de resultados obtenidos durante la etapa de evaluación de los cuatro modelos construidos en base a nuestro conjunto de datos aplicado sobre los algoritmos de clasificación descritos previamente en el documento, este conjunto se divide de la siguiente manera: número total de instancias empleadas, porcentaje de instancias correctamente clasificadas, porcentaje de instancias incorrectamente clasificadas, detalle de precisión por clase y matriz de confusión. A continuación, se brindará al lector una breve definición de estos dos últimos elementos.

2.1 Detalle de precisión

La herramienta proporciona por cada modelo evaluado un detalle de precisión, el cual consiste en una tabla que agrupa los siguientes criterios.

- Tasa de verdaderos positivos, es decir el total de instancias correctamente clasificadas dada una clase.
- Tasa de falsos positivos, es decir el total de instancias incorrectamente clasificadas dada una clase.
- Precisión, proporción de instancias pertenecientes a una clase dividida por el total de instancias clasificadas en dicha clase.
- Exhaustividad, es el equivalente a los verdaderos positivos.
- *F-Measure*, medición calculada a partir de la precisión y exhaustividad.
- Área de ROC, es uno de los valores de salida más importantes brindados por la herramienta Weka, un clasificador optimo deberá tener un área de ROC cercana al valor de 1.

Cabe mencionar que en el capítulo siguiente se procederá a analizar el valor del área bajo la curva de ROC para sustentar la elección del modelo más apto, el cual será aplicado en el proyecto.

2.2 Matriz de confusión

Una matriz de confusión contiene información acerca de las predicciones hechas por un sistema de clasificación. El rendimiento de estos sistemas es comúnmente evaluado usando los datos en la matriz [Kohavi & Provost 1998]. A continuación, se muestra una matriz de confusión para un clasificador de dos clases.

Tabla 4: Matriz de confusión

Clase	Negativo	Positivo
Negativo	A	B
Positivo	C	D

Las entradas en la matriz contienen el siguiente significado según el contexto de clases detallado.

- A es el número de predicciones correctas para la clase negativo.
- B es el número de predicciones incorrectas para la clase positivo.
- C es el número de predicciones incorrectas para la clase negativo.
- D es el número de instancias correctas para la clase positivo.

2.3 Conjunto de resultados de rendimiento por modelo de clasificación

En este apartado se procederá a mostrar el conjunto de resultados obtenidos durante la etapa de evaluación de cada uno de los cuatro modelos de clasificación construidos, es importante recalcar que el análisis de los valores obtenidos será realizado más adelante.

2.3.1 *Red Bayesiana*

A continuación, se presentan los resultados de rendimiento obtenidos por el modelo construido a partir del algoritmo de Red Bayesiana en base a nuestro conjunto de datos ya previamente documentado.

- Número de instancias: 5471.
- Porcentaje de instancias correctamente clasificadas: 59.824%
- Porcentaje de instancias incorrectamente clasificadas: 40.175%

Tabla 5: Tabla de detalle de precisión para el modelo de Red Bayesiana

TP Rate	FP Rate	Precisión	Exhaustividad	F-Measure	AUC	Clase
0.934	0.227	0.085	0.934	0.156	0.868	Expediente
0.965	0.089	0.832	0.965	0.893	0.941	Intimação de Acórdão
0.431	0.013	0.68	0.431	0.528	0.79	Intimação de Decisão de embargos declaratórios
0.485	0.025	0.463	0.485	0.473	0.78	Intimação de decisão monocrática
0.812	0.045	0.848	0.812	0.83	0.931	Intimação de Sentença
0.014	0.003	0.552	0.014	0.026	0.731	Intimação/Notificação de despacho
0.488	0.043	0.464	0.488	0.476	0.821	Notificação / Intimação em Mandado de Segurança
0.038	0.007	0.188	0.038	0.063	0.655	Petição Elaborada

Tabla 6: Matriz de confusión para el modelo de Red Bayesiana

A	B	C	D	E	F	G	H	Clase
113	2	0	2	3	0	1	0	A = Expediente
16	1655	5	23	9	5	1	1	B = Intimação de Acórdão
73	55	138	10	38	2	3	1	C = Intimação de Decisão de embargos declaratórios
61	6	3	111	39	3	3	3	D = Intimação de decisão monocrática
84	0	43	64	1042	1	18	31	E = Intimação de Sentença
634	259	10	18	51	16	188	3	F = Intimação/Notificação de despacho
163	12	3	4	14	2	189	0	G = Notificação / Intimação em Mandado de Segurança
181	1	1	8	33	0	4	9	H = Petição Elaborada

2.3.2 *kNN (k-Nearest-Neighbor)*

A continuación, se presentan los resultados de rendimiento obtenidos por el modelo construido a partir del algoritmo kNN en base a nuestro conjunto de datos ya previamente documentado.

- Número de instancias: 5471.
- Porcentaje de instancias correctamente clasificadas: 79.839%
- Porcentaje de instancias incorrectamente clasificadas: 20.160%

Tabla 7: Tabla de detalle de precisión para el modelo kNN

TP Rate	FP Rate	Precisión	Exhaustividad	F-Measure	AUC	Clase
0.694	0.009	0.641	0.694	0.667	0.846	Expediente
0.897	0.043	0.904	0.897	0.9	0.926	Intimação de Acórdão
0.541	0.017	0.665	0.541	0.597	0.756	Intimação de Decisão de embargos declaratórios
0.568	0.013	0.653	0.568	0.607	0.779	Intimação de decisão monocrática
0.913	0.032	0.898	0.913	0.906	0.938	Intimação de Sentença
0.763	0.075	0.737	0.763	0.749	0.845	Intimação/Notificação de despacho
0.643	0.031	0.615	0.643	0.629	0.809	Notificação / Intimação em Mandado de Segurança
0.519	0.024	0.492	0.519	0.505	0.747	Petição Elaborada

Tabla 8: Matriz de confusión para el modelo kNN

A	B	C	D	E	F	G	H	Clase
84	2	1	2	1	7	4	20	A = Expediente
2	1538	38	19	5	100	8	5	B = Intimação de Acórdão
3	43	173	2	36	29	16	18	C = Intimação de Decisão de embargos declaratórios
3	13	2	130	23	26	15	17	D = Intimação de decisão monocrática
1	11	18	12	1172	40	12	17	E = Intimação de Sentença
13	80	20	18	31	899	86	32	F = Intimação/Notificação de despacho
11	7	3	5	10	84	249	18	G = Notificação / Intimação em Mandado de Segurança
14	7	5	11	27	35	15	123	H = Petição Elaborada

2.3.3 *CNB (Complemento de Red Bayesiana)*

A continuación, se presentan los resultados de rendimiento obtenidos por el modelo construido a partir del algoritmo CNB en base a nuestro conjunto de datos ya previamente documentado.

- Número de instancias: 5471.
- Porcentaje de instancias correctamente clasificadas: 72.180%
- Porcentaje de instancias incorrectamente clasificadas: 27.819%

Tabla 9: Tabla de detalle de precisión para el modelo CNB

TP Rate	FP Rate	Precisión	Exhaustividad	F-Measure	AUC	Clase
0.512	0	0.969	0.512	0.67	0.756	Expediente
0.941	0.091	0.825	0.941	0.879	0.925	Intimação de Acórdão
0.15	0.002	0.828	0.15	0.254	0.574	Intimação de Decisão de embargos declaratórios
0.052	0	0.857	0.052	0.099	0.526	Intimação de decisão monocrática
0.984	0.155	0.66	0.984	0.79	0.915	Intimação de Sentença
0.656	0.107	0.627	0.656	0.641	0.774	Intimação/Notificação de despacho
0.452	0.011	0.751	0.452	0.565	0.72	Notificação / Intimação em Mandado de Segurança
0.013	0	1	0.013	0.025	0.506	Petição Elaborada

Tabla 10: Matriz de confusión para el modelo CNB.

A	B	C	D	E	F	G	H	Clase
62	4	0	0	16	34	5	0	A = Expediente
0	1613	0	2	29	70	1	0	B = Intimação de Acórdão
2	58	48	0	198	13	1	0	C = Intimação de Decisão de embargos declaratórios
0	49	0	12	134	33	1	0	D = Intimação de decisão monocrática
0	4	3	0	1263	10	3	0	E = Intimação de Sentença
0	209	3	0	154	773	40	0	F = Intimação/Notificação de despacho
0	12	2	0	62	136	175	0	G = Notificação / Intimação em Mandado de Segurança
0	5	2	0	57	163	7	3	H = Petição Elaborada

2.3.4 *SVM (Support Vector Machine)*

A continuación, se presentan los resultados de rendimiento obtenidos por el modelo construido a partir del algoritmo de SVM en base a nuestro conjunto de datos ya previamente documentado.

- Número de instancias: 5471.
- Porcentaje de instancias correctamente clasificadas: 85.267%
- Porcentaje de instancias incorrectamente clasificadas: 14.732%

Tabla 11: Tabla de detalle de precisión para el modelo SVM

TP Rate	FP Rate	Precisión	Exhaustividad	F-Measure	AUC	Clase
0.669	0.001	0.953	0.669	0.786	0.834	Expediente
0.942	0.038	0.92	0.942	0.931	0.952	Intimação de Acórdão
0.722	0.012	0.783	0.722	0.751	0.855	Intimação de Decisão de embargos declaratórios
0.419	0.004	0.835	0.419	0.558	0.708	Intimação de decisão monocrática
0.962	0.041	0.878	0.962	0.918	0.96	Intimação de Sentença
0.902	0.078	0.761	0.902	0.826	0.912	Intimação/Notificação de despacho
0.54	0.007	0.853	0.54	0.661	0.766	Notificação / Intimação em Mandado de Segurança
0.57	0.007	0.78	0.57	0.659	0.781	Petição Elaborada

Tabla 12: Matriz de confusión para el modelo SVM

A	B	C	D	E	F	G	H	Clase
81	1	1	0	4	19	9	6	A = Expediente
1	1616	5	15	12	66	0	0	B = Intimação de Acórdão
2	51	231	0	24	9	1	2	C = Intimação de Decisão de embargos declaratórios
1	1	5	96	54	59	5	8	D = Intimação de decisão monocrática
0	0	34	1	1234	9	5	0	E = Intimação de Sentença
0	75	10	2	23	1063	5	1	F = Intimação/Notificação de despacho
0	10	4	0	14	129	209	21	G = Notificação / Intimação em Mandado de Segurança
0	3	5	1	40	42	11	135	H = Petição Elaborada

3 Conclusión

En el presente capítulo se completaron las etapas dentro del proceso de clasificación, construcción y evaluación. Ambas etapas fueron completadas haciendo uso de la herramienta para tareas de Minería de Datos, Weka asimismo, los algoritmos descritos previamente fueron construidos en base al conjunto de datos documentado en el capítulo previo. Es importante recalcar la importancia de esta sección, dado que en base a los resultados obtenidos durante la etapa de evaluación de los modelos construidos se sustentará la elección del modelo más apto.

CAPITULO 5: Selección del modelo de clasificación de documentos de carácter judicial en lenguaje portugués según su contenido

En el presente capitulo profundizará sobre el tercer objetivo específico del proyecto, el cual tiene como propósito sustentar la elección del modelo más apto haciendo uso del análisis del valor del área bajo la curva de ROC.

En el presente proyecto se hará uso de gráficas de ROC como medio análisis de rendimiento de clasificadores debido a que son capaces de proveer una medida más enriquecida de rendimiento que medidas escalares, tales como: costo de error, tasa de error y precisión debido a que las gráficas de ROC separan el rendimiento del modelo de las clases sesgadas y costos de error [Fawcett 2004].

1 Gráficas de ROC

Una gráfica de ROC es una técnica para visualizar, organizar y seleccionar clasificadores basados en su rendimiento, las cuales son representadas en un espacio de dos dimensiones, donde el valor del *FP rate* es colocado en el eje X, mientras que el valor de *TP rate* es colocado en el eje Y. Es así como un gráfico de ROC representa las ventajas y desventajas relativas entre los beneficios (verdaderos positivos) y costos (falsos positivos) de un modelo [Fawcett 2004].

Un clasificador discreto es aquel que únicamente asigna una clase para cada tupla. Cada clasificador discreto produce un par (*fp rate*, *tp rate*) correspondiente a un único punto en el espacio de ROC. Informalmente, un punto en el espacio de ROC es mejor que otro si este se ubica más al noroeste, es decir mayor tp rate, menor fp rate o ambos. Aquellos clasificadores que aparecen en el lado izquierdo de la gráfica de ROC son considerados conservadores, dado que realizan clasificaciones positivas únicamente cuando se encuentra con evidencia fuerte, mientras que aquellos que aparecen en el lado derecho de la gráfica son considerados liberales, dado que realizan clasificaciones positivas con débil evidencia entonces consiguen clasificar casi todos los valores positivos correctamente, sin embargo frecuentemente tienen tiene una alta tasa de falsos positivos. A continuación, se muestra una gráfica de ROC para cinco clasificadores discretos.

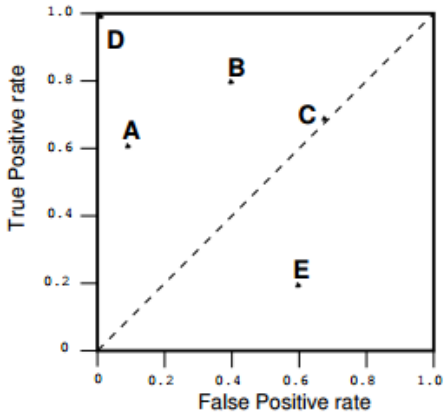


Ilustración 13: Ejemplo de gráfica de ROC para cinco clasificadores discretos [Fawcett 2004].

1.1 Curva de ROC

Las curvas de ROC son una herramienta visual para comparar modelos de clasificación. A fin de graficar una curva de ROC para un modelo de clasificación dado, es necesario que este devuelva un valor probabilístico por cada tupla del conjunto de datos de prueba, es así como es necesario ordenar de manera decreciente las tuplas de prueba de acuerdo a su valor probabilístico, de tal manera que la tupla más propensa a pertenecer al valor positivo aparezca en la cima de la lista. El eje vertical de una curva de ROC representa la tasa de verdaderos positivos, mientras que el eje horizontal representa la tasa de falsos positivos [Han & Kamber 2006].

1.2 Gráficas de ROC multi clases

Al trabajar con más de dos clases la situación se vuelve un poco más compleja si se desea manejar la totalidad del espacio de ROC, es así que con *n* clases, la matriz de confusión se convierte en una matriz *n x n* donde la mayor diagonal contiene las *n* clasificaciones correctas, mientras que las demás entradas contienen los $n^2 - n$ posibles errores. A diferencia de manejar compensaciones entre los verdaderos y falsos positivos, se tienen *n* beneficios y $n^2 - n$ errores. Uno de los métodos básicos aplicados para el manejo de n clases es producir *n* diferentes gráficas de ROC, una por cada clase del conjunto. A este método se le denomina formulación de referencia de clase, en otras palabras si C es el conjunto total de clases, la gráfica de ROC *i* muestra el rendimiento de

clasificación usando la clase c_i como la clase positiva, y todas las demás clases como la clase negativa [Han & Kamber 2006].

1.3 Área bajo la curva de ROC (AUC)

Una curva de ROC es una representación en dos dimensiones del rendimiento de un clasificador. Es así, que para comparar clasificadores debemos reducir la representación ROC a un único valor escalar que represente el rendimiento. Un método común es calcular el área bajo la curva de ROC. Debido a que la región es una porción del área de una unidad de cuadrado, este valor siempre estará entre 0 y 1. Sin embargo, debido a que una predicción aleatoria produce una diagonal entre los puntos (0,0) y (1,1) cuyo valor de área es 0.5, ningún clasificador real debería contar con un AUC menor a 0.5.

El área bajo la curva de ROC tiene una propiedad estadística importante, dado que es equivalente a la probabilidad de que un clasificador pueda ubicarse en un valor positivo aleatorio antes que en uno negativo [Fawcett 2004].

2 Análisis de resultados de área bajo la curva de ROC

En esta sección se procederá a sustentar la elección del modelo de clasificación en base a los valores de AUC obtenidos en el capítulo anterior. A continuación, se presenta una tabla resumen con los valores del área bajo la curva de ROC obtenidos por cada uno de los modelos evaluados en el capítulo previo. A partir de los resultados descritos en la 13, se procede a la elección del *Support Vector Machine* como el modelo de clasificación más apto, debido a que cuenta con el valor de área bajo la curva de ROC más próximo a 1.

Tabla 13: Tabla de resultado por área bajo la curva de ROC por modelo de clasificación

AUC (kNN)	AUC (SVM)	AUC (CNB)	AUC (Red Bayesiana)	Clase
0.846	0.834	0.756	0.868	Expediente
0.926	0.952	0.925	0.941	Intimação de Acórdão
0.756	0.855	0.574	0.79	Intimação de Decisão de embargos declaratórios
0.779	0.708	0.526	0.78	Intimação de decisão monocrática
0.938	0.96	0.915	0.931	Intimação de Sentença
0.845	0.912	0.774	0.731	Intimação/Notificação de despacho
0.809	0.766	0.72	0.821	Notificação / Intimação em Mandado de Segurança
0.747	0.781	0.506	0.655	Petição Elaborada
0.831	**0.846**	0.712	0.815	**Promedio AUC**

3 Conclusiones

En el presente capítulo se realizó la elección del modelo más apto, el candidato seleccionado es el *Support Vector Machine*. La elección se basó en el análisis del valor de AUC, cuyo valor refleja la probabilidad de que un modelo de clasificación ubique una instancia cualquiera en un valor positivo aleatorio antes que en uno negativo.

CAPITULO 6: Aplicación del modelo algorítmico en la clasificación de documentos de carácter judicial en lenguaje portugués

En el presente capítulo se procederá en aplicar el modelo algorítmico en la clasificación de documentos de carácter judicial según su contenido haciendo uso de una interfaz de usuario, la cual funcionará como una caja negra recibiendo el texto del documento y retornará al usuario una categoría dentro del conjunto previamente definido en los capítulos anteriores.

1 Prototipo funcional de interfaz de usuario

En este apartado se procederá a explicar el funcionamiento del prototipo de interfaz de usuario desarrollada para el proyecto, la cual tiene como finalidad brindarle al usuario un entorno que soporte la clasificación de documentos de carácter judicial en lenguaje portugués. La clasificación será realizada haciendo uso del modelo algorítmico seleccionado en el capítulo previo, y funcionará como una caja negra desde la perspectiva del usuario. De esta manera, el usuario deberá ingresar un archivo en formato Excel de dos columnas, texto y categoría, donde la columna categoría estará inicialmente vacía, mientras que la interfaz se encargará de procesar cada uno de los textos incluidos en el archivo de entrada y le asignara la clase correspondiente. A continuación, se muestra una ilustración que corresponde a la vista previa de la interfaz de usuario desarrollada.

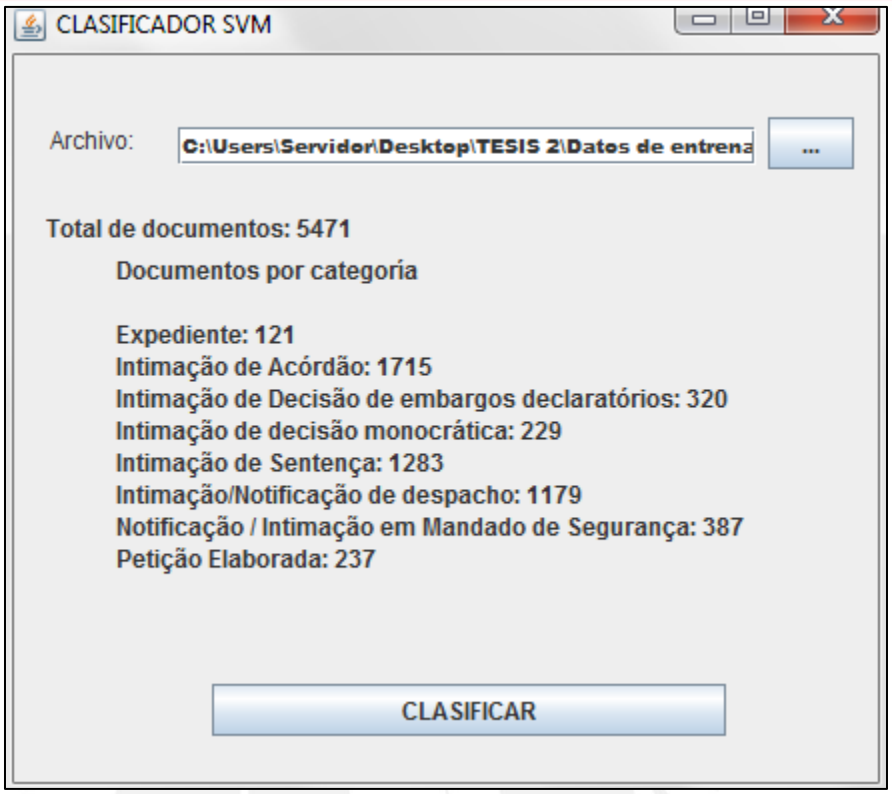


Ilustración 14: Prototipo funcional de interfaz de usuario para la clasificación de documentos

2 Clasificación del documento

En este apartado se profundizará sobre las etapas involucradas en la clasificación de cada documento incluido dentro del documento de entrada en la interfaz de usuario presentada en el punto anterior. El proceso de clasificación se divide en dos etapas, la primera consiste en el procesamiento interno del texto, mientras que la segunda aplica el modelo algorítmico de clasificación sobre el texto procesado para asignarle una categoría dentro del conjunto de clases previamente definido en los capítulos anteriores.

El procesamiento interno del documento tiene como objetivo la obtención de un vector de palabras, la creación de este se realizará mediante el uso de funciones de manejo de cadenas de caracteres. El vector estará conformado por cada una de las palabras dentro del documento y se le asignará asimismo un valor numérico a cada una de ellas, el cual

representa cuan relevante es la palabra para el documento. La técnica de estimación para el cálculo de dicho valor es TF-IDF. Esencialmente, trabaja determinando la frecuencia relativa de la palabra en un documento específico comparado a la proporción inversa de dicha palabra en la colección total de documentos [Ramos 2003]. La principal razón por la cual se ha decidido aplicar esta técnica es debido a que permite controlar los pesos de aquellas palabras que generalmente son más comunes que otras. Finalmente, la importancia de esta etapa radica en que se encarga de la construir la variable de entrada para el modelo algorítmico de clasificación. A continuación, se presenta un esquema gráfico del proceso interno de clasificación.

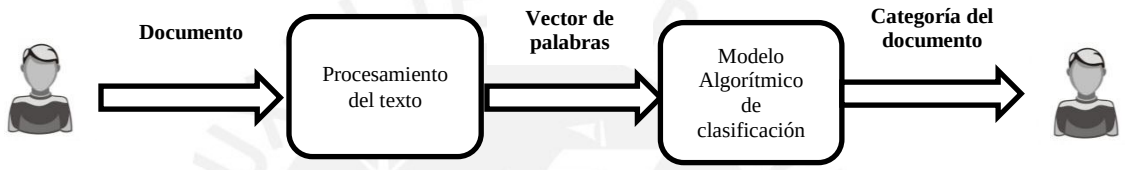


Ilustración 15: Esquema del proceso de clasificación de documento

3 Conclusiones

En el presente capítulo se ha detallado el proceso involucrado dentro de la categorización de un documento de carácter judicial en lenguaje portugués haciendo uso de un prototipo funcional de interfaz de usuario que integra el modelo algorítmico de clasificación construido en los capítulos previos. Es importante recalcar que la interfaz funciona desde la perspectiva del usuario como una caja negra, la cual únicamente requiere del texto del documento para ser este clasificado.

CAPITULO 7: Conclusiones, recomendaciones y trabajos futuros

En este capítulo se presentan las conclusiones obtenidas durante las etapas de desarrollo del proyecto, así como las recomendaciones y trabajos futuros que se consideran pertinentes.

1 Conclusiones

En esta sección se presentan las conclusiones del presente proyecto de fin de carrera. Las conclusiones se han realizado luego de haber completado todos los objetivos trazados para la obtención de un modelo algorítmico para la clasificación de documentos de carácter judicial en lenguaje portugués según su contenido.

En primer lugar, dado que el enfoque de aprendizaje automático es un proceso inductivo en el cual se construye un modelo de clasificación mediante el aprendizaje de propiedades de un conjunto de ejemplos, se requiere contar con una colección de documentos clasificados para el desarrollo del proyecto, dado que esta colección definirá el comportamiento del modelo algorítmico. De esta manera, el proyecto tendrá como base de trabajo una colección de documentos brindada por una empresa en Brasil encargada de la clasificación manual de intimaciones a través de especialistas, llamados procuradores. Las intimaciones son documentos que son enviados desde los tribunales hacia las procuradurías durante un proceso de juicio, y se clasifican en ocho categorías principales.

Dada la variedad de algoritmos de Minería de Datos aplicables al problema de categorización de elementos, se procedió a seleccionar un grupo de cuatro candidatos: Redes Bayesianas, Complemento de Red Bayesiana, Vecino más cercano (kNN) y *Support Vector Machine* (SVM). La selección de estos tiene sustentó en la revisión de la literatura durante la etapa de recolección de información para el desarrollo del Estado de Arte, en donde diversos autores les asignan una alta probabilidad de éxito al ser aplicados a problemas de categorización de textos. Posteriormente, a través del uso de la herramienta Weka se construyó para cada algoritmo un modelo de clasificación basado en la colección de documentos recopilados de la empresa cliente en Brasil.

Los cuatro modelos algorítmicos se llevaron a comparación de modo que se pueda demostrar cuál de ellos es el más apto para solucionar el problema de categorización de documentos. Para ello se elaboró un análisis de rendimiento basado en curvas de ROC, las cuales representan una herramienta visual para comparar modelos de clasificación. El análisis realizado demostró que el modelo algorítmico construido en base al *Support Vector Machine* (SVM) es el mejor de los cuatro en cuanto a rendimiento, obteniendo alrededor de un 85% de precisión.

También es pertinente mencionar que el modelo algorítmico propuesto en el presente proyecto tiene como principal limitante que trabaja únicamente sobre una base de ocho categorías distintas, propiedad que es adquirida por el modelo a partir del conjunto de documentos que fue usado durante la etapa de construcción.

Finalmente, en base a todo lo mencionado anteriormente, se concluye que el modelo algorítmico propuesto en el presente proyecto soluciona el problema de categorización de documentos de carácter judicial en lenguaje portugués según su contenido.

2 Recomendaciones y trabajos futuros

En esta sección se presentan las recomendaciones y trabajos futuros con el propósito de fomentar nuevas investigaciones con relación al problema de categorización de documentos.

Dado que el modelo algorítmico presentado en el proyecto ha sido construido usando una colección de documento base, difícilmente podría ser aplicado en otros escenarios o empresas, además la información viene incrementándose considerablemente en cortos periodos de tiempo, es así como el modelo quedaría obsoleto y desactualizado en un futuro no muy lejano. Es así, como se propone a forma de trabajo futuro, la implementación de un sistema de información de categorización de documentos flexible. La flexibilidad estará determinada por la capacidad del sistema de actualizar el modelo de categorización de documentos periódicamente conforme se incrementen los volúmenes de información de la base de datos donde se alojan los textos del usuario.

Inclusive, se puede ofrecer al usuario una interfaz de configuración de parámetros del algoritmo que será usado para la construcción del modelo, con el propósito de ir mejorando constantemente el porcentaje de rendimiento del modelo de categorización.

En cuanto al ámbito de análisis de modelos de categorización, en el proyecto se empleó las curvas de ROC como herramienta visual para la comparación de rendimiento, sin embargo, ante la gran variedad de métodos de análisis existentes en el área de la Minería de Datos se recomienda a manera de trabajo futuro usar alguno de estos con el propósito de sustentar la elección del *Support Vector Machine* (SVM) como algoritmo más apto desde otra perspectiva.

Por otro lado, se pueden realizar adaptaciones al problema únicamente variando el conjunto de documentos que han servido de entrenamiento para el modelo algorítmico, abriendo la posibilidad de nuevas investigaciones y soluciones.

Además, enfocándose en el ámbito de la investigación, se propone comparar el modelo algorítmico de categorización construido a partir del *Support Vector Machine* (SVM) con otros algoritmos de categorización no incluidos en el presente proyecto con el propósito de determinar qué modelo algorítmico es el más apto dentro del escenario planteado en el proyecto.

REFERENCIAS BIBLIOGRÁFICAS

R. Ahmad, S. Ali, and D. H. Kim, "A Multi-Agent system for documents classification," in *Open Source Systems and Technologies (ICOSST), 2012 International Conference on*, 2012, pp. 28-32.

J. Alcalá-Fdez, L. Sánchez, S. García, M. J. del Jesús, S. Ventura, J. Garrell*, et al.*, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing,* vol. 13, pp. 307-318, 2009.

M.-L. Antonie and O. R. Zaiane, "Text document categorization by term association," in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, 2002, pp. 19-26.

R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval* vol. 463: ACM press New York, 1999.

J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos, "Systematic review in software engineering," *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES,* vol. 679, p. 45, 2005.

C. L. Corso and S. L. Alfaro, "Alternativa de herramienta libre para la implementación de aprendizaje automático," *línea] disponible en http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/Alternativa_de_herramienta_para_Mineria_Datos_CNEISI_2009. pdf*.

H. Cunningham, "Information extraction, automatic," *Encyclopedia of language and linguistics,* pp. 665-677, 2005.

L. Dan, L. Lihua, and Z. Zhaoxin, "Research of Text Categorization on WEKA," in *Proceedings of the 2013 Third International Conference on Intelligent System Design and Engineering Applications*, 2013, pp. 1129-1131.

T. Dong, W. Cheng, and W. Shang, "The Research of kNN Text Categorization Algorithm

Based on Eager Learning," in *Industrial Control and Electronics Engineering (ICICEE), 2012 International Conference on*, 2012, pp. 1120-1123.

T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Machine learning,* vol. 31, pp. 1-38, 2004.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM,* vol. 39, pp. 27-34, 1996.

R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*: Cambridge University Press, 2007.

W. B. Frakes and R. Baeza-Yates, "Information retrieval: data structures and algorithms," 1992.

A. García and R. Lucas, *Documentación automatizada en los medios informativos*, 1987.

J. Han and M. Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*: Morgan kaufmann, 2006.

D. J. Hand, H. Mannila, and P. Smyth, *Principles of data mining*: MIT press, 2001.

S. Joshi and B. Nigam, "Categorizing the Document Using Multi Class Classification in Data Mining," in *Computational Intelligence and Communication Networks (CICN), 2011 International Conference on*, 2011, pp. 251-255.

H. Jung, E. Yi, D. Kim, and G. G. Lee, "Information extraction with automatic knowledge expansion," *Information processing & management,* vol. 41, pp. 217-242, 2005.

R. Kirkby and E. Frank, "WEKA Explorer User Guide for Version 3-4-5," ed, 2005.

R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, 1995, pp. 1137-1145.

R. Kohavi and F. Provost, "Confusion matrix," *Machine learning,* vol. 30, pp. 271-274, 1998.

P. Larranaga, C. M. Kuijpers, R. H. Murga, and Y. Yurramendi, "Learning Bayesian network structures by searching for the best ordering with genetic algorithms," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* vol. 26, pp. 487-493, 1996.

L. H. Lee, C. H. Wan, R. Rajkumar, and D. Isa, "An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization," *Applied Intelligence,* vol. 37, pp. 80-99, 2012.

O. Z. Maimon and L. Rokach, *Data mining and knowledge discovery handbook* vol. 1: Springer, 2005.

S. K. Pal and P. Mitra, *Pattern recognition algorithms for data mining*: CRC press, 2004.

J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naïve bayes text classifiers," in *ICML*, 2003, pp. 616-623.

S. M. Rüger and S. E. Gauch, *Feature reduction for document clustering and classification*: Citeseer, 2000.

J. Rushing, R. Ramachandran, U. Nair, S. Graves, R. Welch, and H. Lin, "ADaM: a data mining toolkit for scientists and engineers," *Computers & Geosciences,* vol. 31, pp. 607 618, 2005.

R. Stair and G. Reynolds, *Fundamentals of information systems*: Cengage Learning, 2013.

A.-H. Tan, "Text mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999, pp.

65-70.

M. Varguez-Moo, V. Uc-Cetina, and C. Brito-Loeza, "Clasificación de documentos usando Máquinas de Vectores de Apoyo," *Abstraction and Application Magazine,* vol. 6, 2014.

G. M. Weiss and B. D. Davison, "Data Mining," in *TO APPEAR IN THE HANDBOOK OF TECHNOLOGY MANAGEMENT, H. BIDGOLI (ED.)*, 2010.

I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.

F. Zhu, P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai*, et al.*, "Biomedical text mining and its applications in cancer research," *Journal of biomedical informatics,* vol. 46, pp. 200-211, 2013.