PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

MODELO ALGORÍTMICO PARA LA CLASIFICACIÓN DE UNA HOJA DE PLANTA EN BASE A SUS CARACTERÍSTICAS DE FORMA Y TEXTURA

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

Susana Milagros Malca Bulnes

ASESOR: Dr. Héctor Andrés Melgar Sasieta.

Lima, abril del 2015

RESUMEN

A lo largo de los años, las plantas han sido consideradas parte vital e indispensable del ecosistema, ya que están presentes en todos los lugares donde vivimos y también donde no lo hacemos. Su estudio es realizado por la ciencia de la botánica, la cual se encargar del estudio de la diversidad y estructura de las mismas. La disminución y extinción de la variedad de las plantas es un tema serio, por lo cual ante el descubrimiento de nuevas especies, se propone una rápida identificación y clasificación a fin de poder monitorearlas, protegerlas y usarlas en el futuro.

El problema de la clasificación de hojas es una tarea que siempre ha estado presente en la labor diaria de los botánicos, debido al gran volumen de familias y clases que existen en el ecosistema y a las nuevas especies que van apareciendo. En las últimas décadas, se han desarrollado disciplinas que necesitan de esta tarea. Por ejemplo, en la realización de estudios de impacto ambiental y en el establecimiento de niveles de biodiversidad, es de gran importancia el inventariado de las especies encontradas.

Por este motivo, el presente proyecto de fin de carrera pretende obtener un modelo algorítmico mediante la comparación de cuatro modelos de clasificación de Minería de Datos, J48 Árbol de Decisión, Red Neuronal, K-Vecino más cercano y Naive Bayes o Red Bayesiana, los cuales fueron adaptados y evaluados para obtener valores de precisión. Estos valores son necesarios para realizar la comparación de los modelos mediante el método de Área bajo la curva ROC (AUC), resultando la Red Bayesiana como el modelo más apto para solucionar el problema de la Clasificación de Hojas.

Tabla de contenido

INDICE DE FIGURAS 5

INDICE DE TABLAS 6

CAPÍTULO 1 7

1.1 PROBLEMÁTICA 7

1.1.1 OBJETIVO GENERAL 9
1.1.2 OBJETIVOS ESPECÍFICOS 9
1.1.3 RESULTADOS ESPERADOS 9

1.2 HERRAMIENTAS, MÉTODOS, METODOLOGÍAS Y PROCEDIMIENTOS 10

1.2.1 HERRAMIENTAS 11
1.2.1.1 WEKA (WAIKATO ENVIRONMENT KNOWLEDGE ANALYSIS) 11
1.2.1.2 MICROSOFT WORD Y MICROSOFT OFFICE 12
1.2.1.3 NETBEANS IDE 13
1.2.2 MÉTODOS Y PROCEDIMIENTOS 13
1.2.2.1 ÁRBOLES DE DECISIÓN 13
1.2.2.2 REDES NEURONALES 14
1.2.2.3 REDES BAYESIANAS 15
1.2.2.4 K – VECINO MÁS CERCANO 16

1.3 ALCANCE 16

1.4 JUSTIFICACIÓN 17

CAPÍTULO 2 18

2.1 MARCO CONCEPTUAL 18

2.1.1 CONCEPTOS RELACIONADOS CON BOTÁNICA 18
2.1.1.1 BOTÁNICA 18
2.1.1.2 FORMA Y ESTRUCTURA DE LAS HOJAS 18
2.1.2 CONCEPTOS RELACIONADOS CON MINERÍA DE DATOS 19
2.1.2.1 DATOS, INFORMACIÓN Y CONOCIMIENTO (DATA, INFORMATION, KNOWLEDGE) 19
2.1.2.2 DESCUBRIMIENTO DE CONOCIMIENTO EN BASE DE DATOS (KDD) 20
2.1.2.3 MINERÍA DE DATOS (DATA MINING) 21
2.1.2.4 CLASIFICACIÓN (CLASSIFICATION) 22
2.1.2.5 APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING) 23
2.1.3 CONCLUSIÓN 23

2.2 ESTADO DEL ARTE 24

2.2.1 Introducción 24
2.2.2 Método usado en la revisión del estado del arte 24
2.2.2.1 Formulación de la pregunta 24
2.2.2.2 Selección de las fuentes 24
2.2.3 Investigaciones sobre el tema 25
2.2.3.1 Clasificación de hojas usando características de forma, color y textura 25
2.2.3.2 Clasificación de hojas usando las características de su estructura y una maquina vector de apoyo 25
2.2.3.3 Clasificación de plantas medicinales usando procesamiento de imágenes 26
2.2.4 Productos actuales presentes en el mercado 26
2.2.4.1 KEEL (Knowledge Extraction base on Evolutionary Learning) 26
2.2.5 Conclusiones sobre el estado del arte 28

CAPÍTULO 3 30

3.1 Introducción 30
3.2 Resultado Esperado 1 (RE1): Conjunto de datos con la estructura y formato adecuados para la adaptación y evaluación de los modelos de clasificación. El conjunto de datos incluye un conjunto de hojas de planta en donde cada una está conformada por siete atributos de forma, seis de textura y finalmente la clase. 30
3.2.1 Conjunto de datos 30
3.2.2 Clases dentro del conjunto de datos 31
3.2.3 Atributos 31
3.2.4 Estructura interna y extensión del archivo 32
3.3 Conclusión 34

CAPÍTULO 4 35

4.1 Introducción 35
4.2 Resultado Esperado 2 (RE2): Conjunto de clasificadores adaptados y evaluados 35
4.2.1 Opciones de prueba 35
4.2.2 Construcción y evaluación de los métodos de clasificación 36
4.2.3 Criterios de precisión 36
4.3 Resultado Esperado 3 (RE3): Resultado de los criterios de precisión por clasificador. 38
4.3.1 Árboles de decisión 38
4.3.2 Redes Neuronales 40
4.3.3 Redes Bayesianas 42
4.3.4 K-Vecino más cercano 44
4.4 Conclusión 46

CAPÍTULO 5 47

5.1 Introducción 47
5.2 Resultado Esperado 4 (RE4): Clasificador seleccionado 47
5.3 Conclusión 50

CAPÍTULO 6 51

6.1 Introducción 51
6.2 Resultado Esperado 5 (RE5): Prototipo funcional que muestre los resultados de clasificar una nueva instancia de hoja de planta. 51
6.2.1 Prototipo Funcional 51
6.2.2 Clasificación 53
6.2.3 Resultado 53
6.3 Conclusión 54

CAPÍTULO 7 55

7.1 INTRODUCCIÓN 55
7.2 CONCLUSIONES 55
7.3 RECOMENDACIONES Y TRABAJOS FUTUROS 56

REFERENCIAS BIBLIOGRÁFICAS 58

ÍNDICE DE FIGURAS

Figura 1: Interfaz de explorador de Weka 11
Figura 2: Interfaz del pre procesamiento de WEKA. 12
Figura 3: Representación de un árbol de decisión para determinar si un cliente es apropiado para comprar una computadora. 14
Figura 4: Ejemplo de un Red Bayesiana y sus componentes. (a) Grafo acíclico dirigido y (b) Tabla de probabilidad condicional para los valores de la variable Cáncer al pulmón (CP) mostrando todas las posibles combinaciones de sus nodos padres Historia Familia (HF) y Fumador (F) 15
Figura 5: Estructura física de una hoja de geranio. 19
Figura 6: Estructura física de una hoja de geranio 19
Figura 7: Minería de Datos como un paso del proceso de descubrimiento del conocimiento (KDD) 20
Figura 8: Taxonomía de Minería de Datos 21
Figura 9: Relación entre Aprendizaje Automático (ML), Minería de Datos (DM) y Descubrimiento de Conocimiento (KDD) 23
Figura 10: Formato Excel con instancias de hojas ya clasificadas. 31
Figura 11: Estructura del archivo con el conjunto de datos a utilizarse en el proyecto. 33
Figura 12: Resultado de abrir un archivo con extensión .arff 33
Figura 13: Detalle del Árbol de decisión adaptado. 38
Figura 14: Detalle de la adaptación de una Red Neuronal. 40
Figura 15: Detalle de la adaptación de una Red Bayesiana. 42
Figura 16: Detalle de la adaptación de K-Vecino más cercano. 44
Figura 17: Gráfico ROC con cinco clasificadores 48
Figura 18: Prototipo inicial. 51
Figura 19: Prototipo funcional de clasificación. 52
Figura 20: Prototipo funcional con el resultado de la clasificación. 53

ÍNDICE DE TABLAS

Tabla 1: Cuadro de las herramientas a utilizarse versus los resultados esperados..... 10
Tabla 2: Cadenas de búsqueda con sus resultados obtenidos respectivamente..... 25
Tabla 3: Resumen de las investigaciones. 27
Tabla 4: Resumen de los productos del mercado. 28
Tabla 5: Nombres científicos de las clases de hojas. 31
Tabla 6: Atributos que forman parte del conjunto de datos. 32
Tabla 7: Porcentaje de instancias correctamente clasificadas para determinar el número de pliegues a utilizar al momento de construir y evaluar los métodos de clasificación. 36
Tabla 8: Ejemplo de una matriz de confusión con dos clases A y B. 37
Tabla 9: Porcentaje de instancias correcta e incorrectamente clasificadas para el árbol de decisión J48. 39
Tabla 10: Detalles de precisión para el árbol de decisión. 39
Tabla 11: Matriz de confusión para el árbol de decisión 39
Tabla 12: Porcentaje de instancias correcta e incorrectamente clasificadas para una red neuronal 40
Tabla 13: Detalles de precisión para una red neuronal. 41
Tabla 14: Matriz de confusión para una red neuronal. 41
Tabla 15: Porcentaje de instancias correcta e incorrectamente clasificadas para una red bayesiana. 42
Tabla 16: Detalles de precisión para una red bayesiana. 43
Tabla 17: Matriz de confusión para una red bayesiana. 43
Tabla 18: Porcentaje de instancias correcta e incorrectamente clasificadas para el clasificador KNN. 44
Tabla 19: Detalles de precisión para un clasificador KNN. 45
Tabla 20: Matriz de confusión para un clasificador KNN. 45
Tabla 21: Valores de AUC para todos los métodos de clasificación descritos. 49
Tabla 22: Valores del AUC por clase para el modelo de Red Bayesiana. 50

CAPÍTULO 1

1.1 Problemática

A lo largo de los años, las plantas han sido consideradas parte indispensable y vital del ecosistema ya que existen en todas partes en donde habita el ser humano y también donde no lo hace [Wu, et al. 2007] [Gopal, et al. 2012]. Su estudio es realizado por la Ciencia de la Botánica, la cual se encarga del estudio de la diversidad y estructura de las mismas [Berg 2007]. La disminución y extinción de la variedad de las plantas es un tema serio; por lo cual ante el descubrimiento de nuevas especies, se sugiere una rápida identificación y clasificación para su monitoreo, protección y uso en el futuro [Gopal, et al. 2012]. Por lo tanto, si hablamos de diversidad de plantas estamos haciendo referencia a la diversidad de hojas que existen actualmente en nuestro ecosistema. Las hojas son el órgano más variable de la plantas y son características de la especie en que crecen. Por este motivo, muchas plantas pueden ser identificadas y clasificadas únicamente por sus hojas [Berg 2007].

El problema de la clasificación de hojas es una tarea que siempre ha estado presente en la labor diaria de los botánicos, debido al gran volumen de familias y clases que existen en el ecosistema y a las nuevas especies que van apareciendo. En las últimas décadas, se han desarrollado disciplinas que necesitan de esta tarea. Por ejemplo, en la realización de estudios de impacto ambiental y en el establecimiento de niveles de biodiversidad, es de gran importancia el inventariado de las especies encontradas [Clark, et al. 2012].

Por esta razón, el uso de un método automatizado de clasificación sería de gran ayuda por los siguientes motivos:

- A pesar de la reducción de la variedad de plantas, lo cual implica la reducción de la diversidad de hojas, existe un gran número de familias dentro de las cuales se pueden identificar diversas clases [Nilsback & Zisserman 2008].

  Bajo este factor, el uso de una herramienta automatizada contribuye a la reducción de los recursos de tiempo, dinero y mano profesional en el proceso de analizar cada clase dentro del volumen total de clases existentes.
- La necesidad de tener un especialista en botánica para el proceso de clasificación se reduciría, ya que la herramienta podría ser usada por personas capacitadas en su uso; lo cuales no tienen que ser necesariamente profesionales en el área.

Los métodos de clasificación examinan las características de un nuevo objeto y lo asignan a una clase predefinida dentro de un conjunto de clases [Berry and Linoff 1997]. Para esto, construyen modelos que son usados para predecir la tendencia futura de los datos. Esta consideración será vital al momento de clasificar una nueva instancia de hoja [Shazmeen, et al.]. Asimismo, es importante resaltar que el método de clasificación es un proceso de dos etapas. La primera se conoce como etapa de entrenamiento y consiste en construir un clasificador analizando o aprendiendo de un conjunto de datos de entrenamiento. La segunda etapa consiste en el uso del clasificador y la estimación de la precisión alcanzada en base al porcentaje de tuplas correctamente clasificadas [Han, et al. 2006].

Por otro lado, existen técnicas de procesamiento y análisis de datos que tienen como principal objetivo el descubrimiento de conocimiento sobre datos que están almacenados. Esta técnica es conocida como Minería de Datos, la cual es considerada como el proceso de exploración y análisis de datos con el fin de descubrir patrones y reglas significativas [Berry & Linoff 1997]. Asimismo, la Minería de Datos es considerada un paso esencial en el proceso de Descubrimiento de Conocimiento (KDD); el cual tiene como principal objetivo el descubrimiento de conocimiento útil dentro de un gran conjunto de datos [Fayyad, et al. 1996]. Cabe resaltar que Minería de datos es la etapa donde se aplican métodos inteligentes para la extracción de patrones [Han, et al. 2006]. Dentro de estos métodos inteligentes se encuentran los métodos de clasificación, los cuales son considerados una de las técnicas más comunes dentro de Minería de Datos.

En general, el uso de una herramienta de clasificación automática contribuiría a reducir los efectos que trae consigo el gran volumen de familias o clases existentes de plantas, tales como:

- Identificación poco acertada.
- Necesidad de profesionales con un alto grado de conocimiento y capacidad de observación.
- Alto uso de recursos como el tiempo y dinero.

El presente proyecto maneja un conjunto de datos que está formado por un conjunto de instancias de hojas de planta, las cuales están definidas por siete atributos numéricos de forma, seis atributos numéricos de textura y la clase como atributo nominal. Con el conjunto de datos como base, el objetivo del presente proyecto de fin de carrera es obtener un modelo de clasificación que será adaptado y evaluado tomando como datos

de entrada el conjunto de datos mencionado. Además, el modelo seleccionado se integrará en un prototipo funcional el cual que permitirá clasificar una nueva hoja de planta.

1.1.1 Objetivo General

Obtener un modelo algorítmico para la clasificación de una hoja de planta en base a sus características de forma y textura.

1.1.2 Objetivos Específicos

En base al objetivo general planteado anteriormente, se puede establecer los siguientes objetivos específicos.

- Objetivo Especifico 1 (OE1): Definir y estructurar el conjunto de instancias de hoja de planta que será utilizado para llevar a cabo las dos etapas del proceso de clasificación: adaptación del método de clasificación y evaluación del mismo.
- Objetivo Específico 2 (OE2): Realizar la etapa de adaptación y evaluación de cuatro diferentes métodos de clasificación.
- Objetivo Específico 3 (OE3): Realizar un análisis comparativo, en base a criterios de precisión, que justifique la elección del método que será empleado en el proyecto.
- Objetivo Específico 4 (OE4): Desarrollar un prototipo funcional que muestre el resultado de clasificar una nueva instancia de hoja haciendo uso de un modelo de clasificación.

1.1.3 Resultados Esperados

Objetivo Especifico 1 (OE1)

- (RE1) Conjunto de datos con la estructura y formato adecuados para la adaptación y evaluación de los modelos de clasificación. El conjunto de datos incluye un conjunto de hojas de planta en donde cada una está conformada por siete atributos de forma, seis de textura y finalmente la clase.

Objetivo Especifico 2 (OE2)

- (RE2) Conjunto de clasificadores adaptados y evaluados.
- (RE3) Resultado de los criterios de precisión por clasificador.

Objetivo Especifico 3 (OE3)

- (RE4) Clasificador seleccionado.

Objetivo Especifico 4 (OE4):

- (RE5) Prototipo funcional que muestre los resultados de clasificar una nueva instancia de hoja de planta.

1.2 Herramientas, métodos, metodologías y procedimientos

En esta sección se presentará diferentes herramientas, métodos y procedimientos que serán usados para el desarrollo de los objetivos específicos descritos anteriormente y por consecuencia el desarrollo y cumplimiento del objetivo general descrito. Además, se detallará el alcance que tendrá el proyecto de fin de carrera considerando las limitaciones, obstáculos y riesgos que podrán presentarse en el futuro. A continuación en la tabla 1 se introduce las herramientas usadas versus los resultados esperados.

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

Resultados esperado	Herramientas a usarse
RE1: Conjunto de datos con la estructura y formato adecuados para la construcción y evaluación de los modelos de clasificación. El conjunto de datos incluye un conjunto de instancias en donde cada una está conformada por siete atributos de forma y seis de textura.	**Editor de texto:** es una herramienta que se usará para dar el formato adecuado al conjunto de datos.
RE2: Conjunto de clasificadores construidos y evaluados. **RE3:** Resultado de los criterios de precisión por clasificador. **RE4:** Clasificador seleccionado	**WEKA:** es una herramienta de prueba de métodos de clasificación y agrupación. Esta herramienta proporciona la implementación en JAVA de diferentes algoritmos de aprendizaje que se pueden aplicar a un conjunto de datos. Asimismo, incluye una variedad de herramientas para transformar conjunto de datos; los cuales pueden ser pre procesados en un esquema de aprendizaje.
RE5: Prototipo funcional que muestre los resultados de clasificar una nueva instancia de hoja de planta.	**Netbeans IDE:** entorno de programación que permite el desarrollo fácil y rápido de aplicaciones de escritorio, móviles y web con JAVA, HTML5, PHP, C, C++.

1.2.1 Herramientas

1.2.1.1 WEKA *(Waikato Environment Knowledge Analysis)*

WEKA es una de las herramientas más completas de evaluación y prueba de los métodos de Clasificación [Charalampopoulos & Anagnostopoulos 2011]; la cual está diseñada para probar de una manera flexible los métodos existentes en nuevos conjuntos de datos. Esta herramienta proporciona la implementación en JAVA de algoritmos de aprendizaje que se pueden aplicar fácilmente a un conjunto de datos. Asimismo, incluye una variedad de herramientas para transformar conjuntos de datos; los cuales pueden ser pre procesados para luego ser incorporados en un esquema de aprendizaje y finalmente poder analizar el rendimiento del clasificador resultante; todo esto sin la necesidad de escribir algún código de programa.

Este entorno incluye métodos para los principales problemas relacionados con Minería de Datos: regresión, clasificación, agrupación, extracción de reglas de asociación y selección de atributos. Dentro de las principales formas de usar WEKA tenemos:

- Aplicar un método de aprendizaje al conjunto de datos y analizar su salida para aprender más sobre los datos.
- Usar modelos aprendidos para realizar predicciones en nuevas instancias.
- Aplicar diferentes aprendices y comparar su rendimiento para elegir uno por medio de la predicción.

Un importante recurso que brinda WEKA es la documentación en línea; la cual ofrece una lista completa de los algoritmos disponibles desde su código fuente. Esto es posible ya que WEKA está creciendo continuamente y su documentación en línea siempre está al día [Witten & Frank 2005].

Figura 1: Interfaz de explorador de Weka [Witten & Frank 2005].

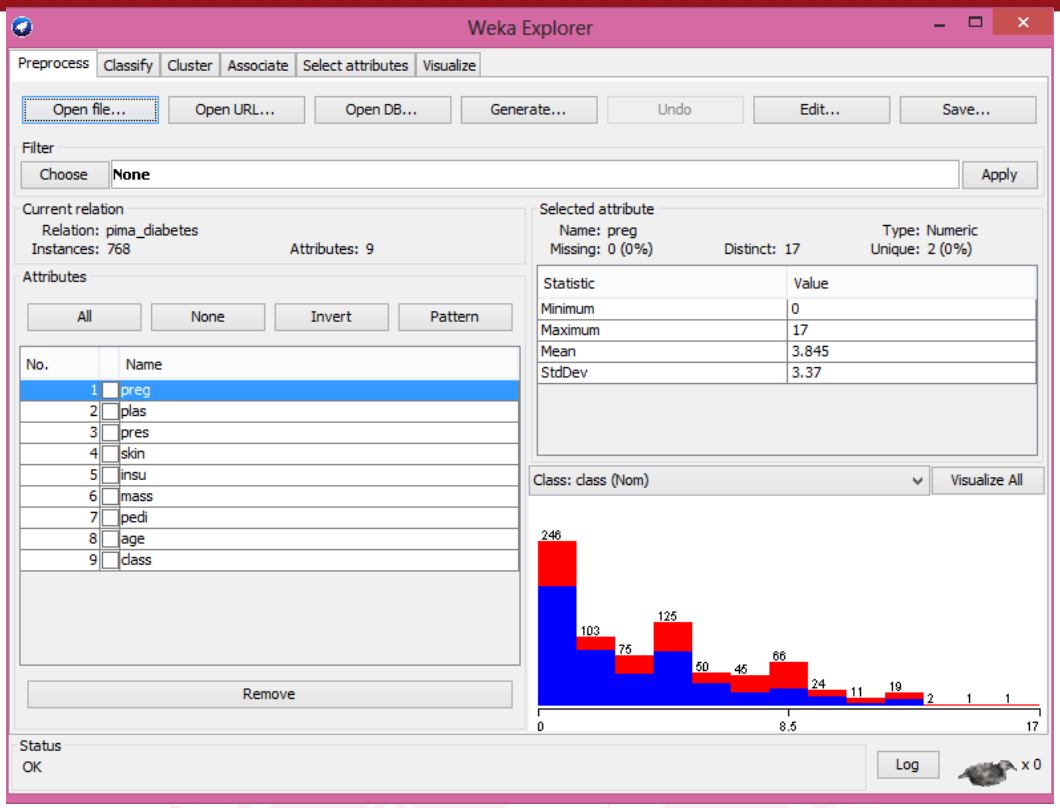


Figura 2: Interfaz del pre procesamiento de WEKA [Witten & Frank 2005].

Justificación: El uso de esta herramienta es vital para el desarrollo, cumplimiento del objetivo general y la finalización del proyecto de fin de carrera. El principal motivo se basa en que, **WEKA** es una herramienta que ofrece una serie de algoritmos desde su código fuente; los cuales son flexibles y se acomodan a algún objetivo específico. Esta característica hace posible la realización del **OE2** el cual, tiene como resultado esperado una serie de clasificadores adaptados y evaluados en base a un conjunto de datos. Además, considerando que WEKA tiene la capacidad de analizar el rendimiento de cada clasificador se podrá cumplir con el **OE3** y obtener el clasificador con mayor precisión y que se acomode más al objetivo del proyecto.

1.2.1.2 Microsoft Word y Microsoft Office

Microsoft Word es un software que está destinado al procesamiento de texto, mientras que Microsoft Excel es un software que es utilizado en la realización de tareas contables y financieras.

Justificación: El uso de estas herramientas es ser el soporte para las tareas de documentación incluidas dentro del proyecto de fin de carrera. Mediante la herramienta Microsoft Word se podrá detallar los resultados del análisis comparativo realizado en el

OE3. Por otro lado, Microsoft Excel esta motivadas principalmente en el uso de funciones y gráficos los cuales serán útiles al momento de realizar el análisis comparativo de los modelos de clasificación.

1.2.1.3 Netbeans IDE

Es un entorno de desarrollo integrado de aplicaciones de escritorio, móviles y web. Asimismo, es una herramienta para programadores pensada para escribir, compilar, depurar y ejecutar programas que pueden ser implementados en diferentes lenguajes tales como: JAVA, HTML5, PHP, C/C++.

Justificación: Esta herramienta permitirá la realización y el cumplimiento del **OE4**, el cual tiene como resultado esperado el desarrollo de un prototipo funcional que permita clasificar una nueva instancia de hoja. Después de haber seleccionado el clasificador con mayor precisión, **Netbeans** ofrece un entorno en el que se podrá hacer uso de una biblioteca propia del WEKA para la implementación de las principales funciones de clasificación.

1.2.2 Métodos y procedimientos

A continuación se listará una serie de métodos de clasificación que serán utilizados para la adaptación y evaluación de los modelos de clasificación utilizados en el proyecto. La herramienta WEKA incluye estos métodos por defecto permitiendo su uso para pruebas y evaluaciones.

1.2.2.1 Árboles de Decisión

Es un método inductivo que realiza divisiones sucesivas de un conjunto de datos, utilizando algún criterio de selección, manteniendo organizada su estructura con el fin de maximizar la distancia entre los grupos de datos generados en cada iteración [Gervilla, et al. 2008].

Esta técnica representa una serie de reglas que llevan hacia una clase de datos, los cuales se examinan para realizar predicciones futuras.
Los arboles de decisión poseen una estructura que está conformada por [Rokach 2008]:

- Nodos: nombres o identificadores de los atributos que caracterizan al conjunto de datos.
- Ramas: condiciones o variables de decisión que cumplen los objetos para poder separarse unos de otros.

- Hojas: conjuntos o grupos de datos resultantes de la división que realiza el algoritmo.

La Figura 3 muestra un ejemplo de árbol de decisión con sus componentes previamente explicados.

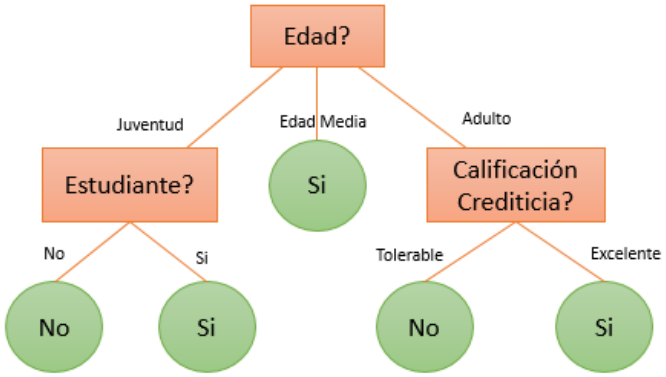


Figura 3: Representación de un árbol de decisión para determinar si un cliente es apropiado para comprar una computadora [Rokach 2008].

Los árboles de decisión realizan la clasificación de la siguiente manera, dada una tupla que tiene asociada una clase desconocida, los valores de sus atributos son probados en el árbol de decisión planteado, trazándose un camino desde el nodo raíz hasta el nodo hoja, el cual contiene la predicción de la clase asociada a dicha tupla.

El uso de un árbol de decisión es una técnica muy popular en Minería de Datos, debido a que para muchos investigadores ofrece transparencia y simplicidad [Rokach 2008]. Asimismo, la construcción de un árbol de decisión no requiere algún dominio de conocimiento o un conjunto de parámetros, por lo tanto es apropiado para el descubrimiento de conocimiento exploratorio.

Los arboles de decisión pueden manejar gran cantidad de datos y su representación es intuitiva y generalmente fácil de asimilar por los humanos. En general, los clasificadores árbol de decisión tienen buena precisión, pero su éxito podría depender de los datos con los que se cuente [Han, et al. 2006].
WEKA es una herramienta que soporta varios tipos de Árboles de decisión pero en el proyecto se usará el árbol de decisión J48.

1.2.2.2 Redes Neuronales

Redes neuronales es un paradigma que está basado en el desarrollo de estructuras matemáticas con la habilidad de aprender. Este método es el resultado de intentos

académicos por modelar el aprendizaje del sistema nervioso [Pujari 2001]. Las redes neuronales usan elementos de procesamiento conocidos como nodos, los cuales son análogos a las neuronas en el cerebro. Estos elementos son interconectados en una red que pueden identificar patrones en los datos [Pujari 2001].

Las ventajas de las redes neuronales incluyen su alta tolerancia a los datos ruidosos; es decir datos que no pueden ser entendidos e interpretados de manera correcta por las máquinas. Además, poseen la habilidad de clasificar patrones en datos que no han sido entrenados; por lo tanto las redes neuronales pueden ser usadas cuando se posee poco conocimiento de las relaciones entre los atributos y clases [Han, et al. 2006].

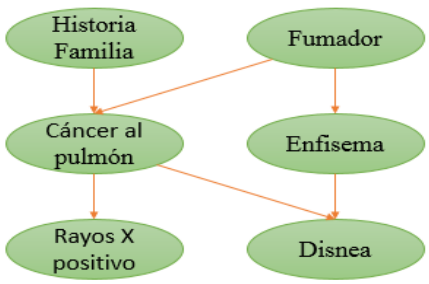
WEKA es una herramienta que soporta varios tipos de Redes Neuronales pero en el proyecto se usará la Red Neuronal multicapas (*Multilayer Perceptron*).

1.2.2.3 Redes Bayesianas

Los clasificadores Bayesianos son clasificadores estadísticos, los cuales especifican distribuciones conjuntas de probabilidad condicional. Asimismo, proporcionan un modelo gráfico de relaciones causales en las que el aprendizaje puede ser realizado.

Una red bayesiana está definida por dos componentes: un grafo acíclico dirigido y un conjunto de tablas de probabilidad condicional. Cada nodo en el grafo representa una variable aleatoria, las cuales pueden ser discretas o continuas. Asimismo, estas variables podrían corresponder a los atributos actuales de los datos o variables ocultas. Cada arco representa una dependencia probabilística. Si un arco es dibujado del nodo Y al nodo Z entonces Y es padre o predecesor de Z, y Z es un descendiente de Y [Han, et al. 2006].

(a) (b)



	HF, F	HF, ~F	~HF, F	~HF, ~F
CP	0.8	0.5	0.7	0.1
~CP	0.2	0.5	0.3	0.9

Figura 4: Ejemplo de un Red Bayesiana y sus componentes. (a) Grafo acíclico dirigido y (b) Tabla de probabilidad condicional para los valores de la variable Cáncer al pulmón

(CP) mostrando todas las posibles combinaciones de sus nodos padres Historia Familia (HF) y Fumador (F) [Han, et al. 2006].

1.2.2.4 K – Vecino más cercano

Es un método utilizado en el área de reconocimiento de patrones, el cual está basado en aprendizaje por analogía. Este método consiste en comparar una tupla de prueba dada con tuplas entrenadas que son similares a la tupla de prueba. Las tuplas entrenadas están descritas por n atributos y cada una de ellas representa un punto en un espacio de dimensión n; es decir, todas las tuplas entrenadas son almacenadas en un espacio de patrones.

Cuando se tiene una tupla desconocida, el clasificador busca en el espacio de patrones las k tuplas entrenadas más cercanas a la tupla desconocida. Estas k tuplas son los k vecinos más cercanos. La tupla desconocida es asignada a la clase más común entre sus k vecinos más cercanos.

La "cercanía" está definida en términos de distancia métrica, tal como la distancia euclidiana. Esta distancia entre dos puntos o tuplas propone, $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ y $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$ es:

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (\text{x1i} - x2i)^2}$$

1.3 Alcance

El presente proyecto de fin de carrera es del tipo Investigación Aplicada y tiene como objetivo obtener un modelo algorítmico que permita clasificar una hoja de planta. Para cumplir con el objetivo descrito se establece el siguiente alcance:

- La adaptación y evaluación del clasificador se realizará en base a un conjunto de datos que recoja información relacionada con la forma y textura de las hojas; es decir, la clasificación se realizará a través de métodos que procesen los atributos de forma y textura.
- Debido a la gran cantidad de clases o tipos de hojas existentes en nuestro ecosistema, la herramienta tendrá como objetivo clasificar una nueva instancia de hoja en una clase dentro de un conjunto de clases limitado, en este caso 17 clases distintas. La limitación del conjunto de clases asegurará el cumplimiento del proyecto en el tiempo estipulado para este.

1.4 Justificación

Según lo planteado en el presente proyecto, las plantas son parte vital e indispensable para nuestro ecosistema por lo cual, ante nuevos descubrimientos, se sugiere una rápida identificación y clasificación de las mismas. Esta acción tiene una gran contribución con el ecosistema, ya que permite realizar un adecuado monitoreo y protección para su futuro uso [Gopal, et al. 2012] .

Por lo tanto, teniendo en cuenta lo analizado, se busca obtener un modelo de clasificación, el cual por medio de un prototipo funcional permita clasificar una nueva instancia de hoja reduciendo el consumo de recursos como tiempo, dinero y mano de obra.

La investigación, beneficiará a sus actores directos como son los especialistas en botánica, brindándoles la facilidad de realizar el inventariado de hojas y plantas de manera más rápida, reduciendo el tiempo invertido en una clasificación manual, así como los recursos de dinero y mano de obra requeridos. Asimismo, especialistas ambientales se beneficiaran con la investigación, dado que en muchas ocasiones necesitan del inventariado de hojas y plantas para realizar estudios de impacto ambiental.

En conclusión, teniendo en cuenta que el estudio de la botánica y los elementos pertenecientes a ella tienen un alcance muy grande y además dado la cantidad de métodos de clasificación pertenecientes al contexto de Minería de Datos, el proyecto servirá como base para el desarrollo de futuras herramientas de clasificación automática de hojas, brindando mayor precisión en sus resultados.

CAPÍTULO 2

2.1 Marco conceptual

En la presente sección se definirá los términos y conceptos que serán mencionados en la problemática. Por lo tanto el objetivo principal de este marco conceptual es desarrollar de manera concreta todos los conceptos y términos que serán de aporte para entender de manera clara la problemática presentada que es la clasificación de hojas en torno a la Minería de Datos.

2.1.1 Conceptos relacionados con Botánica

2.1.1.1 Botánica

Es el estudio científico de las plantas [Mauseth 2012]. Este estudio abarca el origen, diversidad, estructura y procesos internos de las plantas así como su relación con otros organismos y con el medio ambiente [Berg 2007].

El alcance de la botánica es extenso, algunos biólogos de plantas estudian como el clima afecta a las plantas mientras que otros examinan su composición [Berg 2007].

2.1.1.2 Forma y estructura de las hojas

Las hojas son el órgano más variable de las plantas por lo cual, especialistas en biología de plantas han tenido que desarrollar un conjunto entero de terminologías que permitan describir sus formas, márgenes, patrones venosos y la forma en que están unidas al tallo [Berg 2007]. Además, las hojas son consideradas parte indispensable de toda planta ya que le proveen protección (escamas, espinas), soporte, almacenamiento e inclusive obtención de nitrógeno (atrapar y digerir insectos) [Mauseth 2012].

Como cada hoja es característica de la especie en que crece, muchas plantas pueden ser identificadas únicamente por sus hojas las cuales pueden ser redondas, alargadas, cilíndricas, delgadas y estrechas, en forma de abanico o en forma de corazón [Berg 2007]. A continuación la Figura 5 muestra la estructura física de una hoja.

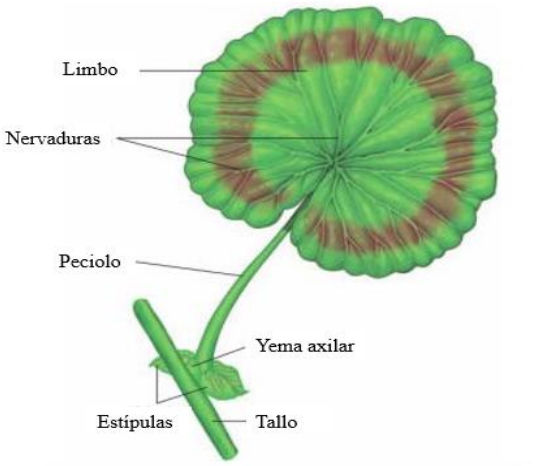


Figura 5: Estructura física de una hoja de geranio [Berg 2007].

2.1.2 Conceptos relacionados con Minería de Datos

2.1.2.1 Datos, Información y Conocimiento *(Data, Information, Knowledge)*

Los datos son hechos o características individuales que están en crudo; es decir no presentan información relevante [Weiss & Davison 2010]. Cuando estos hechos están organizados de una manera significativa se convierten en información; por lo tanto se puede definir información como un conjunto de hechos organizados y procesados de tal manera que tengan un valor adicional más allá del valor que tienen cuando son hechos individuales.

Asimismo, el proceso de definir relaciones entre los datos para crear información que sea útil requiere de conocimiento; por lo tanto el conocimiento se puede interpretar como la comprensión de un conjunto de información y las formas en que la información puede convertirse en útil para ayudar en tareas específicas o para la toma de decisiones [Stair & Reynolds 2008].

Figura 6: Estructura física de una hoja de geranio [Berg 2007].

2.1.2.2 Descubrimiento de Conocimiento en Base de Datos *(KDD)*

KDD es el proceso organizado de identificar patrones nuevos, válidos, útiles y comprensibles de un conjunto de datos grande y complejo [Maimon & Rokach 2005]. Asimismo, Fayyad propone KDD como el proceso de descubrir conocimiento útil de los datos [Fayyad, et al. 1996].

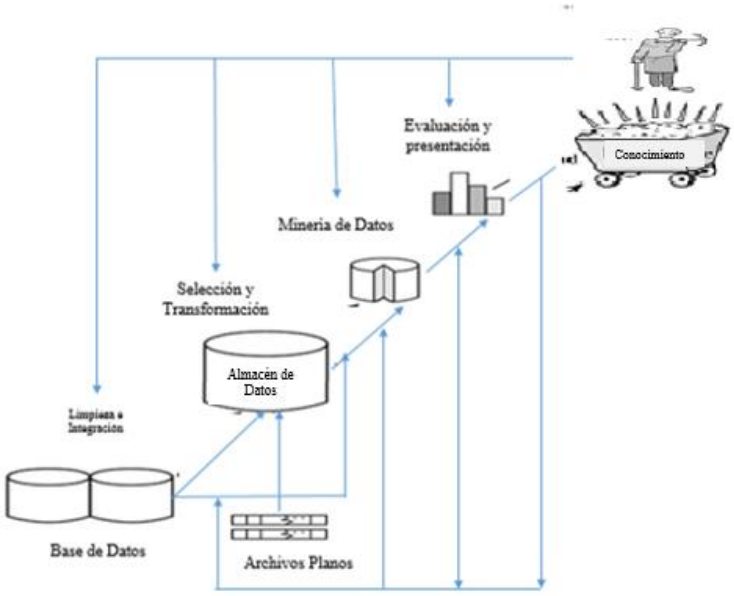


Figura 7: Minería de Datos como un paso del proceso de descubrimiento del conocimiento (KDD) [Han, et al. 2006].

Minería de Datos es un paso esencial del Proceso de Descubrimiento de Conocimiento [Han, et al. 2006]. El proceso del Descubrimiento del Conocimiento está representado en la Figura 7 y consiste de una secuencia iterativa de los siguientes pasos:

- Limpieza de datos: En esta etapa la fiabilidad de los datos es mejorada. Esta limpieza incluye manejar valores perdidos y la eliminación de ruidos o valores atípicos [Maimon & Rokach 2005].
- Integración de datos: Esta etapa incluye la combinación de datos provenientes de múltiples fuentes [Han, et al. 2006].
- Selección de Datos: Esta etapa consiste en seleccionar y recuperar datos relevantes de la Base de Datos [Han, et al. 2006].
- Transformación de Datos: En esta etapa los datos son transformados o consolidados en formas apropiadas para la Minería de Datos [Han, et al. 2006].
- Minería de Datos: Esta etapa es un proceso esencial donde métodos inteligentes son aplicados para extraer patrones de datos [Han, et al. 2006].

- Evaluación: En esta etapa se evalúa e interpretan los patrones extraídos (reglas, confiabilidad). En esta etapa el conocimiento descubierto es documentado para su uso posterior [Maimon & Rokach 2005].
- Presentación del Conocimiento: En esta etapa, técnicas de visualización y representación del conocimiento son usadas para presentar el conocimiento extraído al usuario [Han, et al. 2006].

2.1.2.3 Minería de Datos *(Data Mining)*

Minería de Datos es la exploración y análisis de una gran cantidad de datos con el fin de descubrir patrones y reglas significativas [Berry & Linoff 1997]. Este proceso incluye el uso de algoritmos para hallar patrones en los datos y finalmente generar conocimiento [Weiss & Davison 2010]. Hoy en día, existen muchos métodos que son usados para diferentes propósitos y objetivos. La Figura 8 presenta la taxonomía de Minería de Datos mostrando la interrelación y agrupamiento de los métodos.

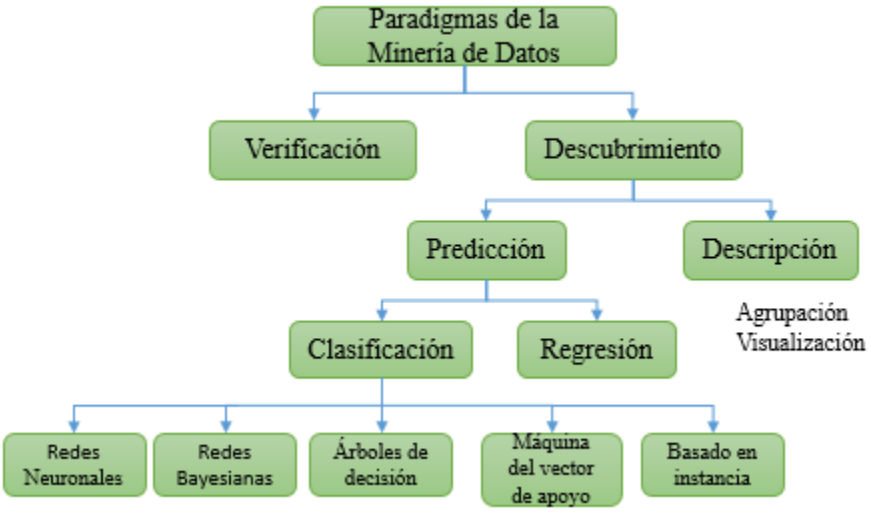


Figura 8: Taxonomía de Minería de Datos [Maimon & Rokach 2005].

Es útil distinguir dos métodos importantes en Minería de Datos: Método orientado a la verificación y Método Orientado al descubrimiento [Maimon & Rokach 2005]. Los métodos de Verificación se ocupan de evaluar una hipótesis propuesta por una fuente externa. Estos métodos incluyen estadística tradicional, pruebas de hipótesis y análisis de varianza por lo que son métodos poco asociados con problemas de Minería de Datos; ya que no están relacionados con el descubrimiento de una hipótesis y por el contrario evalúan y prueban una hipótesis conocida [Maimon & Rokach 2005].

Por otro lado, los métodos de Descubrimiento identifican patrones en la data automáticamente. Estos métodos se pueden diferenciar en métodos descriptivos y métodos predictivos [Maimon & Rokach 2005]. Los métodos descriptivos están orientados a la interpretación de los datos. [Maimon & Rokach 2005]. Esto se realiza con el fin de encontrar patrones en los datos que sean interpretados por los humanos [Fayyad, et al. 1996]. Por el contrario los métodos predictivos usan algunas variables o campos en la base de datos para predecir futuros valores de otras variables de interés [Fayyad, et al. 1996] [Maimon & Rokach 2005].

2.1.2.4 Clasificación *(Classification)*

Es una de las técnicas más comunes de Minería de Datos [Berry & Linoff 1997]; estas técnicas construyen modelos que son usados para predecir la tendencia futura de los datos [Shazmeen, et al.].

La clasificación de datos es un proceso que consta de dos etapas. La primera, conocida como etapa de aprendizaje o fase de entrenamiento, consiste en la construcción de un clasificador donde un algoritmo de clasificación construye el clasificador analizando o aprendiendo de una conjunto de datos de entrenamiento. La segunda etapa consiste en evaluar el modelo construido sobre un conjunto de datos para estimar su precisión; es decir el porcentaje de tuplas correctamente clasificadas por el clasificador [Han, et al. 2006].

Asimismo, la técnica de Clasificación examina las características de un nuevo objeto con el fin de asignarlo a una clase que esta predefinida dentro de un conjunto de clases. El objeto a ser clasificado esta generalmente representado por registros en la tabla de la Base de Datos o una fila , y el acto de clasificar consisten en agregar una nueva columna con un código de clase de algún tipo [Berry & Linoff 1997].

Las principales técnicas usadas en Clasificación son [Shazmeen, et al.]:

- Arboles de decisión
- Redes neuronales artificiales
- K-Vecino más cercano
- Naive-Bayes.
- Máquina de vectores de apoyo

2.1.2.5 Aprendizaje Automático *(Machine Learning)*

El Aprendizaje Automático es un campo de la Inteligencia Artificial que es usado para el análisis de datos y el de

scubrimiento del conocimiento en Base de Datos [Kononenko & Kukar 2007]. Según Ian Witten y Eibe Frank el Aprendizaje Automático proporciona una base técnica de Minería de Datos; la cual es usada para extraer información de una Base de Datos. La información extraída es expresada en una forma comprensible y puede ser usada para diferentes propósitos [Witten & Frank 2005].

Entre sus principales enfoques de investigación se encuentra [Carbonell, et al. 1983]:

- El análisis y desarrollo de sistemas de aprendizaje para mejorar el rendimiento en un predeterminado conjunto de tareas.
- La exploración teórica del espacio de posibles métodos de aprendizaje y algoritmos independientes del dominio de aplicación.
- La investigación y el equipo de simulación del proceso de aprendizaje humano.

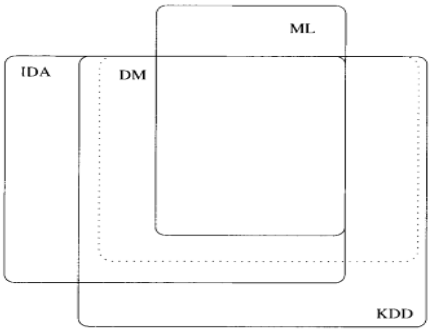


Figura 9: Relación entre Aprendizaje Automático (ML), Minería de Datos (DM) y Descubrimiento de Conocimiento (KDD) [Kononenko & Kukar 2007].

La mayoría de los métodos de Minería de Datos tienen su origen en Aprendizaje Automático; sin embargo no se puede considerar que el Aprendizaje Automático es un subconjunto de Minería de Datos; ya que este también abarca otros campos [Kononenko & Kukar 2007].

2.1.3 Conclusión

Después de haber desarrollado el marco conceptual, podemos afirmar que los conceptos definidos previamente han sido abarcados por muchos autores, lo cual refleja los diferentes puntos de vista que tienen en torno a estos.

Por este motivo, se concluye que para poder tener una visión más clara sobre la problemática que está inmersa en la clasificación de hojas se requiere tener conocimiento de los términos expuestos previamente.

2.2 Estado del arte

2.2.1 Introducción

Hoy en día la Minería de Datos es un tema que presenta muchas aplicaciones en diferentes ámbitos del mundo entero, tales como Medicina, Marketing, Botánica, Economía y algunos otros. Por esta razón, en la presente sección se dará a conocer las investigaciones, productos y proyectos que están relacionados con la clasificación de hojas en torno a la Minería de Datos. El principal objetivo planteado en esta revisión del estado del arte es brindar al lector una noción del estado actual del tema del presente proyecto de fin de carrera, mediante la presentación de los avances más actuales sobre Minería de Datos y Clasificación de hojas los cuales, incluyendo productos, investigaciones y proyectos relacionados con el tema.

2.2.2 Método usado en la revisión del estado del arte

El método usado en la revisión del estado del arte es la Revisión Sistemática, la cual se define como un medio para evaluar e interpretar toda información e investigación disponible correspondiente a una pregunta en particular, tema de área o fenómeno de interés. En general, las revisiones sistemáticas tienen como objetivo presentar una evaluación de un tema de investigación mediante el uso de metodologías rigurosas, confiables y auditables [Kitchenham 2004]. Este método se llevó a cabo en diferentes Bases de Datos como: IEEE Explorer y SCOPUS

2.2.2.1 Formulación de la pregunta

Para realizar la Revisión Sistemática se formuló la siguiente pregunta: ¿Qué herramientas ase han utilizado para resolver problemas de Minería de Datos y más específicamente Clasificación de hojas? Además, se formuló la pregunta ¿Qué investigaciones sobre Minería de Datos y Clasificación de hojas se han realizado recientemente?

Para responder estas preguntas se identificaron las siguientes palabras claves como: “Data Mining”, “Classification”, “Botany” and “Plants”.

2.2.2.2 Selección de las fuentes

Por medio de la agrupación de las palabras claves mencionadas anteriormente y con el uso de operadores lógicos se llevó a cabo la Revisión Sistemática en la cual se usó principalmente el operador lógico “AND”. Además, el criterio de exclusión empleado en la revisión se enfoca principalmente en el año de publicación de la investigación y en el tipo de documento.

A continuación la tabla 2 muestra un resumen con los resultados de la revisión sistemática detallando los criterios de inclusión y exclusión utilizados.

Tabla 2: Cadenas de búsqueda con sus resultados obtenidos respectivamente.

Cadena de Búsqueda	Base de Datos	Criterios de exclusión		N° Resultados
		Año	Tipo de Documento	
"Classification" AND "Data Mining" AND "Botany"	IEEE Xplore	2010 - 2014	Todos	7
"Classification" AND "Plants" AND "Botany"	IEEE Xplore	2010 - 2014	Conferencias y publicaciones Diarios y revistas Artículos de acceso	66
"Plants" AND "Classification" AND "Data Mining"	IEEE Xplore	2008 - 2014	Conferencias y publicaciones Diarios y revistas Artículos de acceso	37
"Classification" AND "Data Mining" AND "Botany"	SCOPUS	2008- 2014	Todos	5

2.2.3 Investigaciones sobre el tema

En esta sección se presentarán algunas investigaciones que intentan resolver el problema de clasificación de hojas o un entorno similar.

2.2.3.1 Clasificación de hojas usando características de forma, color y textura

Hasta la actualidad, muchos investigadores han propuesto métodos para identificar y clasificar plantas. Comúnmente, estos métodos no capturaban color ya que, el color no era reconocido como una característica importante. En esta investigación se incorpora el uso de características de forma, color y textura para la aplicación de un método llamado red neuronal o red neuronal probabilística que permitirá la clasificación de hojas. Descriptores de Fourier, relación de ancho y alto ***(aspect ratio)***, redondez ***(roundness ratio)*** y dispersión son usados como características de forma mientras que momentos de color que consisten en desviación estándar y asimetría representan características de color.

El resultado experimental muestra que el método de clasificación tiene una precisión del 93.75% cuando es probado en un conjunto de datos que contiene 32 especies de hojas [Kadir, et al. 2013].

2.2.3.2 Clasificación de hojas usando las características de su estructura y una maquina vector de apoyo

Esta investigación propone el método SVM (Maquina vector de apoyo) para realizar la clasificación automática de hojas. Este método está basado en la teoría del aprendizaje

automático que intenta mapear la presente información al límite óptimo en función del espacio buscando la distancia mínima entre dos clases. El método propuesto fue implementado por Microsoft Visual C++ y la librería SVM. Las imágenes usadas para la prueba varían en tipo, color, tamaño, estructura, textura, etc.

Finalmente, se puede concluir que la eficiencia del método decrementará conforme se incremente las características de las hojas [Watcharabutsarakham, et al. 2012].

2.2.3.3 Clasificación de plantas medicinales usando procesamiento de imágenes

Ante la diversidad de plantas existentes y considerándolas como parte indispensable en nuestros ecosistema, la investigación se centra en brindar a los profesionales en botánica una herramienta capaz de identificar a las plantas de manera rápida y haciendo uso de la información disponible; para ello se propone la implementación de un sistema para la identificación automática de plantas medicinales de una base de datos de hojas a través de procesamiento de imágenes.

La investigación sustenta que la eficiencia de este método es de un 92% y que para futuras investigaciones se podrá mejorar la eficiencia con la elección de un algoritmo que consuma menos recursos computacionales mediante una optimización de la lógica [Gopal, et al. 2012].

2.2.4 Productos actuales presentes en el mercado

En esta sección se presentarán productos actuales que sirven de apoyo al tema del presente proyecto de fin de carrera; el cual está relacionado principalmente con el concepto de Minería de Datos y las técnicas aplicadas a este.

2.2.4.1 KEEL *(Knowledge Extraction base on Evolutionary Learning)*

KEEL es un software no comercial en JAVA que le permite al usuario evaluar el comportamiento del aprendizaje evolutivo para las diferentes técnicas que son aplicadas a los problemas relacionados con la Minería de Datos: regresión, clasificación, agrupación y extracción de patrones [Fernández, et al. 2009]. Esta herramienta ofrece muchas ventajas como [Alcalá-Fdez, et al. 2008]:

- Reduce el trabajo de programación, ya que incluye una librería con algoritmos de aprendizaje evolutivo basado en diferentes paradigmas. Asimismo, simplifica la integración de los algoritmos de aprendizaje evolutivos con diferentes técnicas de pre procesamiento.
- Extiende el rango de posibles usuarios, ya que las librerías que están incluidas en el software son fáciles de usar; lo cual reduce el nivel de conocimiento

requerido; por lo que los investigadores con poco conocimiento pueden ser capaces de aplicar los algoritmos en sus problemas.

- El software puede ser instalado en cualquier máquina que contenga JAVA; es decir es independiente del sistema operativo.

2.2.4.2 *ADaM* (Algortihm Development and Mining)

ADaM es una herramienta usada en Minería de datos; la cual está diseñada para usar con datos científicos. Esta herramienta está siendo usada para explorar una amplia variedad de conjunto de datos científicos en diferentes plataformas y usando diferentes técnicas y metodologías. ADaM proporciona un conjunto de herramientas para cada proceso básico de Minería de Datos, incluyendo clasificación, agrupación, extracción de reglas de asociación y pre procesamiento. Además, ADaM incluye herramientas para extraer características de las imágenes y convertir esos datos a la forma de vector de patrones.

Los componentes están diseñados para ser independientes de la plataforma y las versiones están disponibles para MS Windows y Linux [Rushing, et al. 2005].

A continuación se mostrará dos tablas que resumen las investigaciones y productos de mercado presentados anteriormente.

Tabla 3: Resumen de las investigaciones.

Investigación	Método usado	Descripción
Clasificación de hojas usando características de forma, color y textura	Redes neuronales	• Uso de una red neuronal que emplea características de forma, color y textura para la clasificación de hojas. • El método de clasificación tiene una precisión de 93.75%.
Clasificación de hojas usando las características de su estructura y una maquina vector de apoyo	SVM (Máquina de vector de apoyo)	• El método SVM intenta mapear la información presente al límite óptimo en función del espacio buscando la distancia mínima entre dos clases. • La eficiencia del método decrementará conforme se incremente el número de características.

Clasificación de plantas medicinales usando procesamiento de imágenes	Procesamiento de imágenes	• Se propone la implementación de un sistema basado en procesamiento de imágenes para la identificación automática de las plantas medicinales. • La eficiencia del método es del 92%.

Tabla 4: Resumen de los productos del mercado.

Productos en el mercado	Descripción
WEKA	• Está diseñada para probar de manera flexible los métodos de clasificación existentes en nuevos conjuntos de datos. • Proporciona la implementación en JAVA de algoritmos de aprendizaje. • WEKA brinda documentación en línea, la cual ofrece una lista completa de los algoritmos disponibles desde su código fuente.
KEEL	• Software no comercial en JAVA que es empleado en la evaluación el comportamiento del aprendizaje evolutivo de las diferentes técnicas de Minería de datos.
ADaM	• Explora una amplia variedad de conjuntos de datos científicos en diferentes plataformas y usando diferentes técnicas y metodologías. • Incluye herramientas para extraer características de las imágenes y convertir esos datos a la forma de vector de patrones.

2.2.5 Conclusiones sobre el estado del arte

Luego de haber realizado la revisión del estado del arte, se puede apreciar lo que nos ofrece el mercado en cuanto a soluciones para el problema de clasificación en el contexto de Minería de Datos. Las soluciones son variadas en cuanto a diseño y complejidad. Por este motivo, en base a todo lo investigado se puede establecer un punto de partida para la realización del proyecto tomando como base alguna solución ya planteada.

En cuanto a los productos que existen en la actualidad, se puede apreciar que son beneficiosos y útiles pero que no están hechos a medida de un problema en específico; lo cual les da mayor flexibilidad; ya que se pueden adecuar a cualquier problema relacionado con Minería de Datos.

CAPÍTULO 3

Objetivo Específico 1 (OE1): Definir y estructurar el conjunto de instancias de hoja de planta que será utilizado para llevar a cabo las dos etapas del proceso de clasificación: adaptación del método de clasificación y evaluación del mismo.

3.1 Introducción

Después de revisar y entender toda la información disponible sobre Minería de Datos y métodos de clasificación, se puede plantear el punto de partida para la realización del proyecto de fin de carrera, clasificación automática de hojas. En este capítulo se planteará el primer objetivo específico que consiste en definir y estructurar el conjunto de datos que se usará en la adaptación y evaluación del clasificador de hojas. Este objetivo presenta un solo resultado esperado. Este único resultado propone un conjunto de datos el cual tiene que poseer un adecuado formato y estructura. Por este motivo, en lo que resta del capítulo se detallará el formato, estructura, clases y atributos del conjunto de datos.

3.2 Resultado Esperado 1 (RE1): Conjunto de datos con la estructura y formato adecuados para la adaptación y evaluación de los modelos de clasificación. El conjunto de datos incluye un conjunto de hojas de planta en donde cada una está conformada por siete atributos de forma, seis de textura y finalmente la clase.

3.2.1 Conjunto de datos

El proyecto de fin de carrera se aplicará sobre un conjunto de datos que recopila información sobre la forma y textura que presentan ciertas clases de hojas. Este archivo contiene un conjunto de instancias de hojas ya clasificadas con el fin de cumplir con las dos etapas del proceso de clasificación, adaptación y evaluación. Cabe resaltar que la información relacionada con cada instancia de hoja es el resultado de aplicar técnicas de procesamiento de imágenes reales de hojas [Silva, et al. 2013]. Cabe mencionar que el proyecto no se enfocará en realizar el procesamiento de las imágenes para obtener el valor de los atributos. El proyecto utilizará la información de las hojas ya procesadas.

A continuación se presenta un conjunto de instancias de hojas en formato Excel. La columna A representa la clase de la hoja, las columnas B hasta H representan los atributos de forma y las columnas I hasta N representan los atributos de textura de las hojas.

Es importante resaltar que las instancias se almacenan en un archivo Excel con el objetivo de que puedan ser administradas de una manera adecuada y ordenada. Sin embargo, este formato no es aceptado por la herramienta WEKA. El formato adecuado así como la estructura que incluye las clases y atributos se detallarán más adelante.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Clase	Excentricidad	Relación de aspecto	Estiramiento	Solidez	Convexidad	Factor isoperimétrico	de penetración máxima	Intensidad promedio	Contraste promedio	Suavidad	Tercer momento	Uniformidad	Entropía
2	Quercus suber	0.72694	1.4742	0.32396	0.98535	1	0.83592	0.0046566	0.04779	0.12795	0.016108	0.0052323	0.00027477	1.1756
3	Quercus suber	0.74173	1.5257	0.36116	0.98152	0.99825	0.79867	0.0052423	0.02416	0.090476	0.0081195	0.002708	0.0000748	0.69659
4	Quercus robur	0.82268	1.6657	0.59909	0.84493	0.85614	0.32056	0.048803	0.079282	0.18571	0.033338	0.013523	0.00027502	1.5377
5	Quercus robur	0.82135	1.6035	0.6025	0.84981	0.87544	0.31304	0.056079	0.069239	0.16319	0.02594	0.0092918	0.00032134	1.4686
6	Nerium oleander	0.98502	6.0596	0.83612	0.97628	0.98772	0.31546	0.02322	0.03683	0.10661	0.011238	0.0032908	0.00022206	1.0185
7	Nerium oleander	0.98487	6.0229	0.83628	0.96522	1	0.29509	0.024667	0.0303	0.097095	0.0093394	0.0027833	0.00018341	0.85025
8	Betula pubescens	0.51577	1.2902	0.32454	0.9208	0.99474	0.62017	0.023879	0.047501	0.12228	0.014732	0.0042129	0.00056945	1.1232
9	Betula pubescens	0.59837	1.3539	0.35704	0.93647	0.99825	0.64448	0.019818	0.045326	0.13137	0.016966	0.0059817	0.0003532	0.93682
10	Betula pubescens	0.55112	1.2821	0.39683	0.88375	0.99649	0.49492	0.030576	0.062015	0.14621	0.02093	0.0063685	0.00067747	1.2659
11	Tilia tomentosa	0.44941	1.1374	0.38458	0.90736	0.97368	0.54595	0.047809	0.11385	0.18608	0.033467	0.0082395	0.0012682	2.2422
12	Tilia tomentosa	0.46727	1.1876	0.36748	0.91542	0.97719	0.57995	0.051335	0.13764	0.18732	0.033899	0.0061636	0.0029358	2.5551
13	Tilia tomentosa	0.55362	1.2438	0.39517	0.91092	0.97193	0.51113	0.046323	0.083154	0.17465	0.0296	0.0094574	0.00074607	1.5807
14	Acer palmatum	0.38501	1.0656	0.63042	0.51223	0.59123	0.13705	0.12292	0.033373	0.098907	0.0096879	0.0027872	0.00021426	1.0015
15	Acer palmatum	0.26758	1.1316	0.60128	0.54301	0.77368	0.20311	0.15354	0.017945	0.07145	0.0050792	0.0014652	0.0000725	0.63273
16	Acer palmatum	0.24465	1.047	0.60511	0.56524	0.79474	0.21788	0.12522	0.037595	0.127	0.015874	0.006587	0.00010798	0.8331

Figura 10: Formato Excel con instancias de hojas ya clasificadas.

3.2.2 Clases dentro del conjunto de datos

Para el proyecto de fin de carrera se procesará una cantidad limitada de hojas, en este caso 17 clases las cuales se detallan a continuación.

Tabla 5: Nombres científicos de las clases de hojas.

Clase	Nombre científico	Clase	Nombre científico
1	Quercus Suber	**10**	Primula Vulgaris
2	Quercus Robur	**11**	Boungainvillea
3	Nerium Oleander	**12**	Euonymus Japonicus
4	Betula Pubescens	**13**	Magnolia Soulangeana
5	Tilia Tomentosa	**14**	Buxus Sempervirens
6	Acer Palmatum	**15**	Urtica Dioica
7	Celtis	**16**	Podocarpus
8	Corylus Avellana	**17**	Acca Selloviana
9	Castanea Sativa		

La Tabla 5 muestra las clases que serán parte del conjunto de datos a utilizarse en el proyecto y estarán acompañadas de un conjunto limitado de atributos.

3.2.3 Atributos

El conjunto de datos cuenta con 14 atributos incluyendo la clase. Este conjunto se divide en atributos de forma y textura de la hoja. Todos los valores de los atributos son reales y fueron el resultado del procesamiento de ciertas imágenes de hojas, como se explicó

anteriormente. A continuación la Tabla 6 detalla los atributos que formarán parte de la estructura del conjunto de datos, la cual se detallará en el siguiente apartado.

Tabla 6: Atributos que forman parte del conjunto de datos.

N° Atributo	Atributos de forma	N° Atributo	Atributos de textura
1	Excentricidad *(Eccentricity)*	**8**	Intensidad promedio *(Average Intensity)*
2	Relación de aspecto *(Aspect Ratio)*	**9**	Contraste promedio *(Average Contrast)*
3	Alargamiento *(Elongation)*	**10**	Suavidad *(Smoothness)*
4	Solidez *(Solidity)*	**11**	Tercer momento *(Third Moment)*
5	Convexidad *(Stochastic Convexity)*	**12**	Uniformidad *(Uniformity)*
6	Factor isoperimétrico *(Isoperimetric Factor)*	**13**	Entropía *(Entropy)*
7	Profundidad de penetración máxima *(Maximal Indentation Depth)*		

3.2.4 Estructura interna y extensión del archivo

Teniendo en cuenta las clases y atributos definidos en los puntos anteriores, es importante resaltar que el archivo a utilizar tiene que tener una estructura interna apropiada; es decir una estructura que pueda ser utilizada por la herramienta WEKA para adaptar y evaluar el clasificador. Asimismo, a pesar que WEKA es una herramienta flexible en cuanto al conjunto de datos; es decir, tiene la capacidad de procesar diferentes tipos de datos, la extensión del archivo debe ser la correcta, en este caso .arff. La estructura del archivo es la siguiente:

- **@RELATION** <nombre-relación> (línea 2), todo archivo .arff debe comenzar con esta primera declaración.
- **@ATTRIBUTE** <nombre-atributo> <tipoDeDato> (línea 4 hasta línea 17), en esta sección se incluye una línea por cada atributo (columna) que se vaya a incluir en el conjunto de datos, indicando su nombre y el tipo de dato. Los tipos de datos aceptados por WEKA son NUMERIC (numérico), STRING (cadena), DATE (fecha).
- **@DATA** (a partir de la línea 21), en esta sección se incluyen los datos propiamente dichos, cada columna es separada por una coma y todas las filas

deben tener el mismo número de columnas. Este número debe coincidir con número de declaraciones (@ ATTRIBUTE) realizadas en la segunda sección.

hojasDataset.txt

```
@RELATION hojas

@ATTRIBUTE Eccentricity REAL
@ATTRIBUTE AspectRatio  REAL
@ATTRIBUTE Elongation   REAL
@ATTRIBUTE Solidity REAL
@ATTRIBUTE StochasticConvex REAL
@ATTRIBUTE IsoperimetricFact    REAL
@ATTRIBUTE MaximalIndentationDepth  REAL
@ATTRIBUTE AverageIntensity REAL
@ATTRIBUTE AverageContrast  REAL
@ATTRIBUTE Smoothness   REAL
@ATTRIBUTE ThirdMoment  REAL
@ATTRIBUTE Uniformity   REAL
@ATTRIBUTE Entropy  REAL
@ATTRIBUTE class    {QuercusSuber, QuercusRobur, NeriumOleander, BetulaPubescens, TiliaTomentosa, AcerPalmatum, CeltisSp, CorylusAvellana, CastaneaSativa, Primul

@DATA

0.72694 ,   1.4742  ,   0.32396 ,   0.98535 ,   1   ,   0.83592 ,   0.0046566 ,   0.04779 ,   0.12795 ,   0.016108  ,   0.0052323 ,   0.00027477 ,   1.175
0.74173 ,   1.5257  ,   0.36116 ,   0.98152 ,   0.99825 ,   0.79867 ,   0.0052423 ,   0.02416 ,   0.090476  ,   0.0081195 ,   0.002708  ,   0.0000748 ,
0.76722 ,   1.5725  ,   0.38998 ,   0.97755 ,   1   ,   0.80812 ,   0.0074573 ,   0.011897  ,   0.057445  ,   0.0032891 ,   0.00092068 ,   0.0000379 ,
0.73797 ,   1.4597  ,   0.35376 ,   0.97566 ,   1   ,   0.81697 ,   0.0068768 ,   0.01595 ,   0.065491  ,   0.0042707 ,   0.0011544 ,   0.0000663 ,   0
0.82301 ,   1.7707  ,   0.44462 ,   0.97698 ,   1   ,   0.75493 ,   0.007428  ,   0.0079379 ,   0.045339  ,   0.0020514 ,   0.00055986 ,   0.0000235 ,
0.72997 ,   1.4892  ,   0.34284 ,   0.98755 ,   1   ,   0.84482 ,   0.0049451 ,   0.010487  ,   0.058528  ,   0.0034138 ,   0.0011248 ,   0.0000248 ,
0.82063 ,   1.7529  ,   0.44458 ,   0.97964 ,   0.99649 ,   0.7677  ,   0.0059279 ,   0.018375  ,   0.080587  ,   0.0064523 ,   0.0022713 ,   0.0000415
0.77982 ,   1.6215  ,   0.39222 ,   0.98512 ,   0.99825 ,   0.80816 ,   0.0050987 ,   0.024875  ,   0.089686  ,   0.0079794 ,   0.0024664 ,   0.0001467
0.83089 ,   1.8199  ,   0.45693 ,   0.9824  ,   1   ,   0.77106 ,   0.0060055 ,   0.0072447 ,   0.040616  ,   0.0016469 ,   0.00038812 ,   0.0000329 ,
0.90631 ,   2.3906  ,   0.58336 ,   0.97683 ,   0.99825 ,   0.66419 ,   0.0084019 ,   0.0070096 ,   0.042347  ,   0.0017901 ,   0.00045889 ,   0.0000283
0.7459  ,   1.4927  ,   0.34116 ,   0.98296 ,   1   ,   0.83088 ,   0.0055665 ,   0.0057679 ,   0.036511  ,   0.0013313 ,   0.00030872 ,   0.0000318 ,
```

Figura 11: Estructura del archivo con el conjunto de datos a utilizarse en el proyecto.

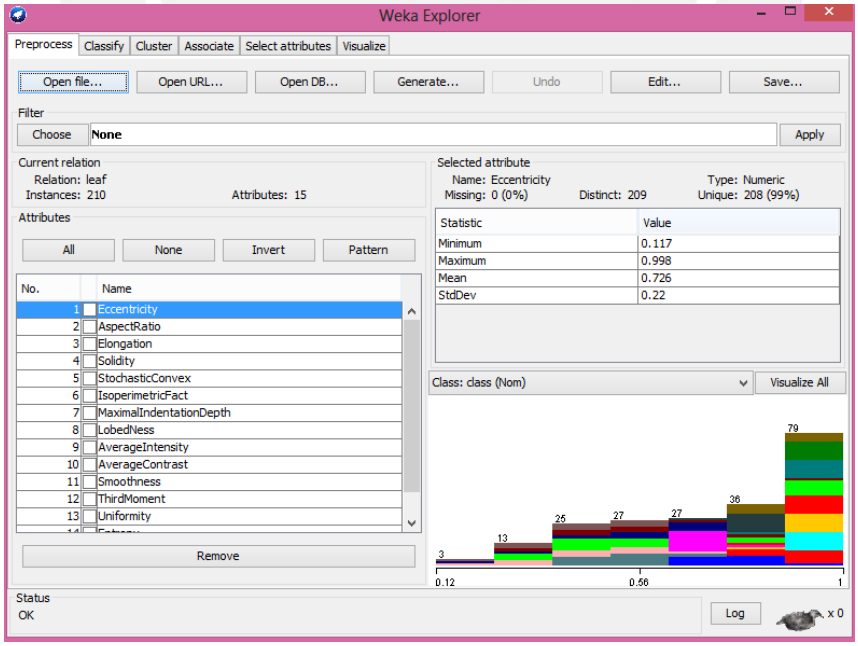


Figura 12: Resultado de abrir un archivo con extensión .arff

Es importante resaltar que la Figura 12 muestra algunas características del conjunto de datos a utilizarse en la clasificación. Número de instancias, número de atributos y algunos valores relacionados con los atributos.

3.3 Conclusión

Después de definir la estructura completa del archivo que contiene el conjunto de datos, se concluye que es de vital importancia contar con un conjunto de datos bien estructurado, ya que la eficiencia y precisión del clasificador dependerá de la calidad de los datos de entrada. Además, es importante señalar que para la realización de los próximos objetivos específicos se deberá contar con la estructura del archivo de conjunto de datos totalmente definida.

Objetivo Específico 2 (OE2): Realizar la etapa de adaptación y evaluación de cuatro diferentes métodos de clasificación.

4.1 Introducción

Como se explicó anteriormente, la minería de datos es la etapa en donde se aplican métodos inteligentes para la extracción de patrones. Por ese motivo, después de estructurar adecuadamente el conjunto de datos a utilizar a lo largo del proyecto, el capítulo 4 detalla el proceso de construcción y evaluación de los 4 métodos descritos en el apartado de Métodos y Herramientas del presente documento junto a las opciones de prueba que fueron utilizadas para completar el objetivo.

4.2 Resultado Esperado 2 (RE2): Conjunto de clasificadores adaptados y evaluados

4.2.1 Opciones de prueba

WEKA es una herramienta multiuso que soluciona problemas de Minería de Datos; es decir es útil para problemas de clasificación, agrupación y regresión. Asimismo, para cada categoría de problemas existe un número importante de parámetros. Dentro de la categoría de clasificación encontramos opciones de prueba, las cuales tienen el objetivo de delimitar el número de instancias que se utilizarán para la construcción y evaluación del clasificador seleccionado.

Para el presente proyecto de fin de carrera se utilizará la validación Cruzada *(Cross - validation)* como modo de prueba. Este método utiliza un número k de pliegues o dobleces *(folds)*, en donde el conjunto de datos (D) es dividido aleatoriamente en k subconjuntos exclusivos (D_1, D_2, D_3,…, D_K) del mismo tamaño aproximadamente. El método de clasificación es entrenado y evaluado k veces. Cada vez $t \in \{1, 2,\ldots, k\}$, este es entrenado en D/D_t y evaluado en D_t [Kohavi, 1995].

El número de pliegues o dobleces (folds) tomados en cuenta para el presente proyecto de fin de carrera es 15 (k=15), el número se determinó después de realizar la calibración con cuatro métodos, en donde los resultados fueron los siguientes.

Tabla 7: Porcentaje de instancias correctamente clasificadas para determinar el número de pliegues a utilizar al momento de construir y evaluar los métodos de clasificación.

Método	% Instancias correctamente clasificadas (k=10)	% Instancias correctamente clasificadas (k=15)	% Instancias correctamente clasificadas (k=20)
Árbol de decisión	74.3 %	76.7%	75.2%
K-Vecino más cercano	77.1%	78.6%	78.0%
Red Neuronal	80.95%	81.50%	80.95%
Red Bayesiana	79.52%	81.9%	80.5%

La principal razón del uso del método de Validación Cruzada (Cross-Validation), es que trabaja con divisiones en los datos de entrada, que son brindados en un archivo con formato .arff, lo que permite que el método seleccionado aprenda de un conjunto de datos y evalúe su precisión en otro conjunto de datos en un proceso iterativo. Este criterio de prueba es muy importante, ya que el prototipo funcional desarrollado tiene como función principal clasificar una nueva instancia.

4.2.2 Construcción y evaluación de los métodos de clasificación

Después de determinar el modo con el cual se adaptarán y evaluarán los diferentes métodos de clasificación se procede a realizar la adaptación y evaluación de los diferentes modelos para luego ser guardados para su uso posterior.

El proceso de construcción y evaluación se realiza simultáneamente; es decir se realiza en un solo paso. Es importante considerar que para cada uno de los métodos existen una cantidad de variaciones. Por cuestiones de documentación, en el proyecto se elegirán los métodos básicos.

4.2.3 Criterios de precisión

Después de la construcción del modelo de clasificación seleccionado, la herramienta WEKA proporciona una serie de criterios que son utilizados para determinar el método más adecuado para un conjunto de datos específico. Los criterios de precisión son los siguientes:

- **Instancias correctamente clasificadas:** porcentaje del total de instancias del conjunto de datos de entrada y se calcula teniendo el número total de instancias que han sido clasificadas correctamente.

  totalInstancias: número total de instancias.
  instClasCorrect= número total de instancias clasificadas correctamente.
  porcentajeCorrect= Porcentaje de instancias correctamente clasificadas.

porcentajeCorrect (%) = instClasCorrect / totalInstancias

- **Instancias incorrectamente clasificadas:** porcentaje del total de instancias del conjunto de datos de entrada y se calcula teniendo el número total de instancias que han sido clasificadas incorrectamente.

  totalInstancias: número total de instancias.
  instClasIncorrect: número total de instancias clasificadas incorrectamente.
  porcentajeIncorrect: Porcentaje de instancias correctamente clasificadas.

  porcentajeIncorrect (%) = instClasIncorrect / totalInstancias

- **Detalles de precisión por clase:** es una tabla que agrupa diferentes criterios por cada clase analizada. Los criterios son los siguientes:

  - ✓ **Tasa TP:** tasa de verdaderos positivos.
  - ✓ **Tasa FP**: tasa de falsos positivos.
  - ✓ **Precisión:** proporción de instancias que son verdaderamente de una clase dividida por el total de instancias clasificadas como esa clase.
  - ✓ **Recall:** proporción de instancias clasificadas como una clase dada dividido por el total de instancias de esa clase.
  - ✓ **F-mesure:** una medida que combina el valor de precisión y recall y se calcula 2 * precisión * recall / (precisión + recall).
  - ✓ **Área ROC:** es uno de los valores más importantes que proporciona la herramienta WEKA. Un clasificador óptimo tendrá sus valores de área ROC cercanos a 1 o en mejor caso igual a 1.

- **Matriz de confusión:** es una matriz que detalla el número de predicciones por clase en donde las columnas representan el número de predicciones por clase y cada fila la clase [Kohavi & Provost, 1998]. Por ejemplo:
  - ✓ A, B : clases
  - ✓ 100 instancias en total (50 instancias de la clase A y 50 de la clase B)

  Tabla 8: Ejemplo de una matriz de confusión con dos clases A y B.

a	**b**	
48	2	**a=A**
5	45	**b=B**

Como se puede apreciar en la Tabla 8, de las 50 instancias de clase A, 48 instancias han sido clasificadas correctamente; es decir como clase A, mientras que 2 instancias han sido clasificadas como clase B. De manera similar, de las 50 instancias de clase B, 45 han sido clasificadas correctamente; es decir como clase B y 5 instancias han sido clasificadas como clase A. Es importante resaltar que la matriz de confusión puede ser tratada como la base para determinar los valores de los otros criterios mencionados anteriormente.

4.3 Resultado Esperado 3 (RE3): Resultado de los criterios de precisión por clasificador.

Los siguientes resultados de los criterios de precisión se obtuvieron después de adaptar cada modelo usando la herramienta WEKA. Cada resultado contiene todos los criterios de precisión descritos anteriormente y fueron proporcionados por la herramienta WEKA. Estos criterios son importantes, ya que sirven de base para seleccionar el método más apto para realizar la clasificación de hojas.

4.3.1 Árboles de decisión

WEKA ofrece una serie de variaciones con respecto al método árbol de decisión. La variación usada para el proyecto de fin de carrera es el J48.

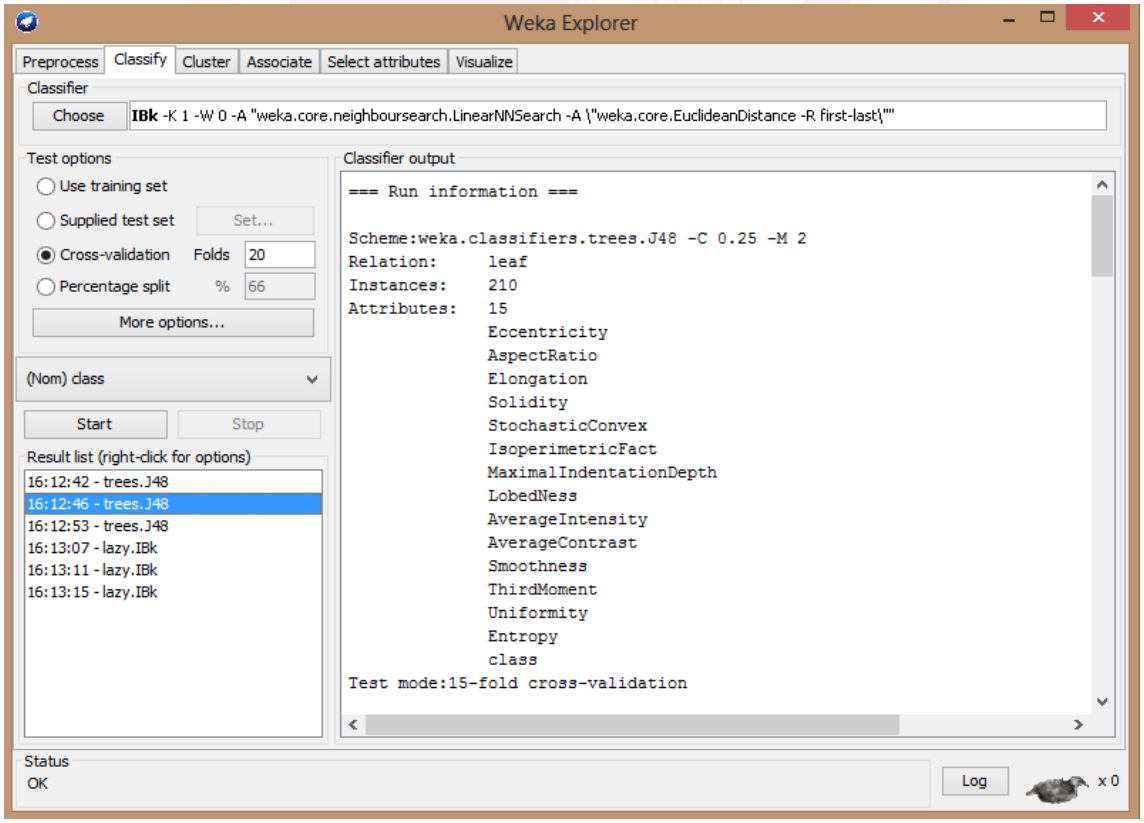


Figura 13: Detalle del Árbol de decisión adaptado.

- **Instancias correcta e incorrectamente clasificadas**

Tabla 9: Porcentaje de instancias correcta e incorrectamente clasificadas para el árbol de decisión J48.

Método	% Instancias correctamente clasificadas	% Instancias incorrectamente clasificadas
J 48	76.7 %	23.3%

- **Detalles de precisión por clase**

Tabla 10: Detalles de precisión para el árbol de decisión.

Tasa TP	Tasa FP	Precisión	Recall	F-Medida	Area ROC	Clase
0.667	0.015	0.727	0.667	0.696	0.897	QuercusSuber
0.833	0.01	0.833	0.833	0.833	0.915	QuercusRobur
0.818	0.005	0.9	0.818	0.857	0.908	NeriumOleander
0.714	0.015	0.769	0.714	0.741	0.95	BetulaPubescens
0.846	0.01	0.846	0.846	0.846	0.957	TiliaTomentosa
1	0	1	1	1	1	AcerPalmatum
0.75	0.025	0.643	0.75	0.692	0.859	CeltisSp
0.769	0.015	0.769	0.769	0.769	0.874	CorylusAvellana
0.583	0.02	0.636	0.583	0.609	0.77	CastaneaSativa
0.417	0.015	0.625	0.417	0.5	0.698	PrimulaVulgaris
0.692	0.02	0.692	0.692	0.692	0.873	BougainvilleaSp
0.667	0.03	0.571	0.667	0.615	0.85	EuonymusJaponicus
0.75	0.02	0.692	0.75	0.72	0.906	MagnoliaSoulangeana
0.917	0.005	0.917	0.917	0.917	0.955	BuxusSempervirens
0.75	0.015	0.75	0.75	0.75	0.901	UrticaDioica
1	0.01	0.846	1	0.917	0.995	PodocarpusSp
0.818	0.015	0.75	0.818	0.783	0.946	AccaSelloviana

- **Matriz de confusión**

Tabla 11: Matriz de confusión para el árbol de decisión

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q				Clase clasificada como:
8	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	2	|	a	=	QuercusSuber
0	10	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	|	b	=	QuercusRobur
0	0	9	0	0	0	0	0	0	1	0	0	0	0	0	1	0	|	c	=	NeriumOleander
0	0	0	10	2	0	0	0	0	0	0	0	0	0	2	0	0	|	d	=	BetulaPubescens
0	0	0	0	11	0	0	0	0	0	2	0	0	0	0	0	0	|	e	=	TiliaTomentosa
0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	|	f	=	AcerPalmatum
0	1	0	0	0	0	9	0	0	1	0	0	1	0	0	0	0	|	g	=	CeltisSp
2	0	0	0	0	0	0	10	1	0	0	0	0	0	0	0	0	|	h	=	CorylusAvellana
0	0	0	0	0	0	1	0	7	0	0	0	3	0	0	0	1	|	i	=	CastaneaSativa
0	0	1	0	0	0	2	0	0	5	0	2	0	1	0	1	0	|	j	=	PrimulaVulgaris
0	0	0	1	0	0	0	0	0	0	9	2	0	0	1	0	0	|	k	=	BougainvilleaSp
0	0	0	0	0	0	0	1	0	1	2	8	0	0	0	0	0	|	l	=	EuonymusJaponicus
0	0	0	0	0	0	0	0	3	0	0	0	9	0	0	0	0	|	m	=	MagnoliaSoulangeana
0	0	0	0	0	0	0	0	0	0	0	1	0	11	0	0	0	|	n	=	BuxusSempervirens
0	1	0	2	0	0	0	0	0	0	0	0	0	0	9	0	0	|	o	=	UrticaDioica
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	|	p	=	PodocarpusSp
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	9	|	q	=	AccaSelloviana

4.3.2 Redes Neuronales

Al igual que para el árbol de decisión WEKA ofrece una serie de métodos de redes neuronales. El utilizado en el proyecto será MultilayerPerceptrón.

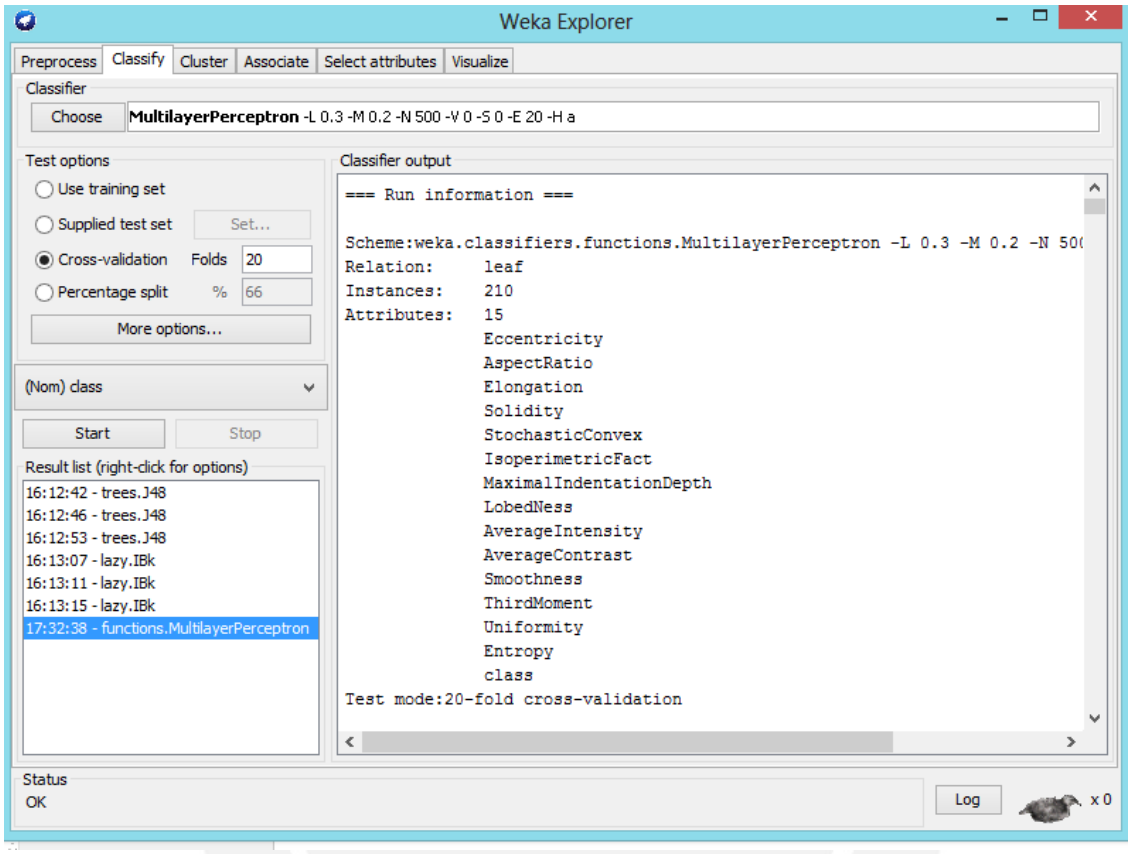


Figura 14: Detalle de la adaptación de una Red Neuronal.

- **Instancias correcta e incorrectamente clasificadas**

Tabla 12: Porcentaje de instancias correcta e incorrectamente clasificadas para una red neuronal.

Método	% Instancias correctamente clasificadas	% Instancias incorrectamente clasificadas
Red neuronal	80.95%	19.05%

- **Detalles de precisión por clase**

Tabla 13: Detalles de precisión para una red neuronal.

Tasa TP	Tasa FP	Precisión	Recall	F-Medida	Area ROC	Clase
0.75	0.25	0.643	0.75	0.692	0.973	QuercusSuber
1	0	1	1	1	1	QuercusRobur
1	0	1	1	1	1	NeriumOleander
0.929	0.02	0.765	0.929	0.839	0.993	BetulaPubescens
0.615	0.005	0.889	0.615	0.727	0.984	TiliaTomentosa
1	0	1	1	1	1	AcerPalmatum
0.833	0.015	0.769	0.833	0.8	0.959	CeltisSp
0.846	0.01	0.846	0.846	0.846	0.966	CorylusAvellana
0.75	0.03	0.6	0.75	0.667	0.964	CastaneaSativa
0.833	0.01	0.833	0.833	0.833	0.995	PrimulaVulgaris
0.769	0.025	0.667	0.769	0.714	0.981	BougainvilleaSp
0.333	0.02	0.5	0.333	0.4	0.871	EuonymusJaponicus
0.5	0.005	0.857	0.5	0.632	0.993	MagnoliaSoulangeana
0.917	0.005	0.917	0.917	0.917	0.998	BuxusSempervirens
0.917	0.01	0.846	0.917	0.88	0.989	UrticaDioica
1	0.005	0.917	1	0.957	1	PodocarpusSp
0.727	0.015	0.727	0.727	0.727	0.972	AccaSelloviana

- **Matriz de confusión**

Tabla 14: Matriz de confusión para una red neuronal.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q				Clase clasificada como:
9	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	|	a	=	QuercusSuber
0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|	b	=	QuercusRobur
0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|	c	=	NeriumOleander
0	0	0	13	1	0	0	0	0	0	0	0	0	0	0	0	0	|	d	=	BetulaPubescens
0	0	0	3	8	0	0	0	0	0	0	1	0	0	1	0	0	|	e	=	TiliaTomentosa
0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	|	f	=	AcerPalmatum
0	0	0	0	0	0	10	0	1	1	0	0	0	0	0	0	0	|	g	=	CeltisSp
2	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	|	h	=	CorylusAvellana
0	0	0	0	0	0	1	0	9	0	0	0	1	0	0	0	1	|	i	=	CastaneaSativa
0	0	0	0	0	0	0	0	0	10	0	0	0	1	0	1	0	|	j	=	PrimulaVulgaris
1	0	0	0	0	0	0	0	0	0	10	1	0	0	1	0	0	|	k	=	BougainvilleaSp
0	0	0	0	0	0	1	1	0	1	4	4	0	0	0	0	1	|	l	=	EuonymusJaponicus
0	0	0	0	0	0	1	0	5	0	0	0	6	0	0	0	0	|	m	=	MagnoliaSoulangeana
0	0	0	0	0	0	0	0	0	0	1	0	0	11	0	0	0	|	n	=	BuxusSempervirens
0	0	0	1	0	0	0	0	0	0	0	0	0	0	11	0	0	|	o	=	UrticaDioica
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	|	p	=	PodocarpusSp
2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	8	|	q	=	AccaSelloviana

4.3.3 Redes Bayesianas

De manera similar a los métodos anteriores, la herramienta WEKA ofrece una serie de variaciones respecto al método Bayes. El método usado para el proyecto de fin de carrera es el Naive Bayes.

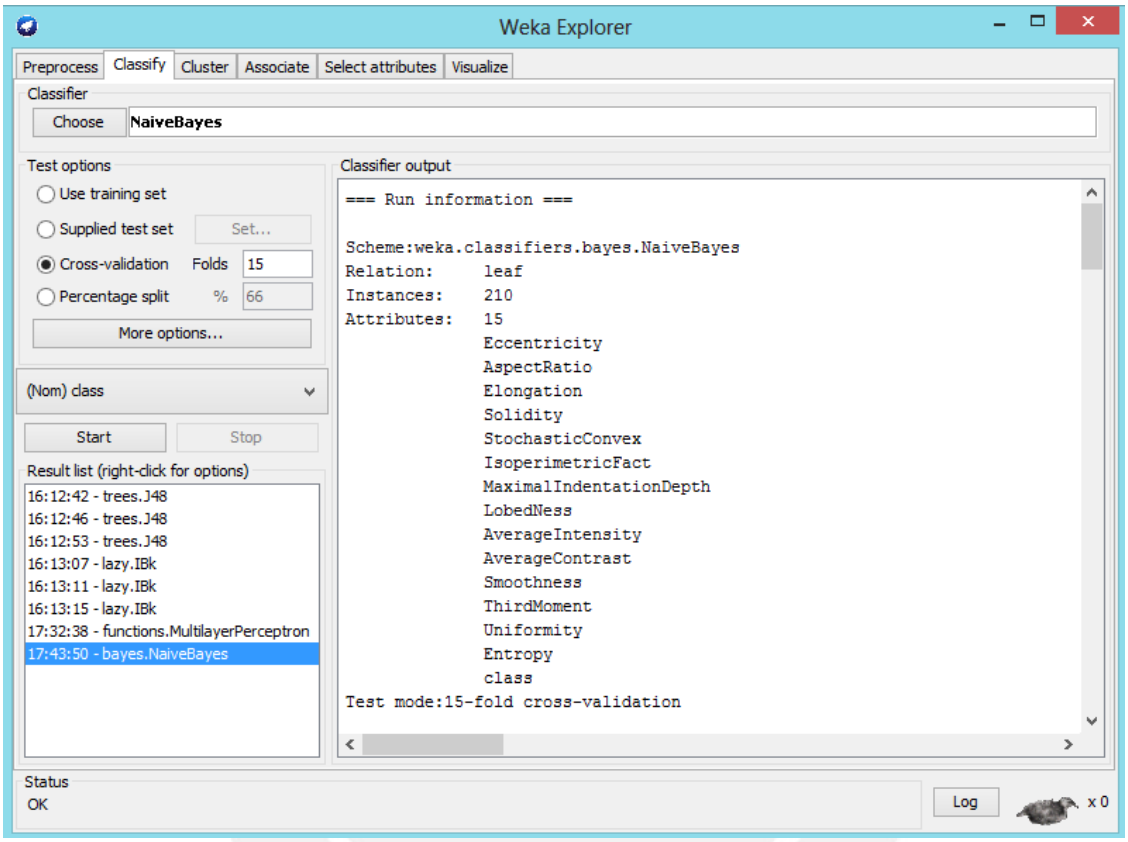


Figura 15: Detalle de la adaptación de una Red Bayesiana.

- **Instancias correcta e incorrectamente clasificadas**

Tabla 15: Porcentaje de instancias correcta e incorrectamente clasificadas para una red bayesiana.

Método	% Instancias correctamente clasificadas	% Instancias incorrectamente clasificadas
Red Bayesiana	80.5%	19.5%

- **Detalles de precisión por clase**

Tabla 16: Detalles de precisión para una red bayesiana.

Tasa TP	Tasa FP	Precisión	Recall	F-Medida	Area ROC	Clase
0.833	0.01	0.833	0.833	0.833	0.974	Quercus Suber
0.917	0	1	0.917	0.957	0.985	QuercusRobur
1	0	1	1	1	1	NeriumOleander
0.786	0.02	0.733	0.786	0.759	0.987	BetulaPubescens
0.846	0.005	0.917	0.846	0.88	0.998	TiliaTomentosa
1	0	1	1	1	1	AcerPalmatum
0.75	0.04	0.529	0.75	0.621	0.967	CeltisSp
0.846	0.015	0.786	0.846	0.815	0.995	CorylusAvellana
0.5	0.03	0.5	0.5	0.5	0.897	CastaneaSativa
0.5	0.02	0.6	0.5	0.545	0.976	PrimulaVulgaris
0.846	0.02	0.733	0.846	0.786	0.98	BougainvilleaSp
0.667	0.005	0.889	0.667	0.762	0.964	EuonymusJaponicus
0.5	0.02	0.6	0.5	0.545	0.975	MagnoliaSoulangeana
0.917	0	1	0.917	0.957	0.999	BuxusSempervirens
0.917	0.005	0.917	0.917	0.917	0.999	UrticaDioica
1	0	1	1	1	1	PodocarpusSp
0.818	0.015	0.75	0.818	0.783	0.991	AccaSelloviana

- **Matriz de confusión**

Tabla 17: Matriz de confusión para una red bayesiana.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q				Clase clasificada como:
10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	|	a	=	QuercusSuber
0	11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	|	b	=	QuercusRobur
0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|	c	=	NeriumOleander
0	0	0	11	1	0	0	0	0	0	1	0	0	0	1	0	0	|	d	=	BetulaPubescens
0	0	0	2	11	0	0	0	0	0	0	0	0	0	0	0	0	|	e	=	TiliaTomentosa
0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	|	f	=	AcerPalmatum
0	0	0	0	0	0	9	0	0	2	0	0	1	0	0	0	0	|	g	=	CeltisSp
1	0	0	1	0	0	0	11	0	0	0	0	0	0	0	0	0	|	h	=	CorylusAvellana
0	0	0	0	0	0	1	0	6	1	0	0	3	0	0	0	1	|	i	=	CastaneaSativa
0	0	0	0	0	0	5	0	1	6	0	0	0	0	0	0	0	|	j	=	PrimulaVulgaris
0	0	0	1	0	0	0	0	0	0	11	1	0	0	0	0	0	|	k	=	BougainvilleaSp
0	0	0	0	0	0	0	2	1	0	0	8	0	0	0	0	1	|	l	=	EuonymusJaponicus
0	0	0	0	0	0	2	0	4	0	0	0	6	0	0	0	0	|	m	=	MagnoliaSoulangeana
0	0	0	0	0	0	0	0	0	0	1	0	0	11	0	0	0	|	n	=	BuxusSempervirens
0	0	0	0	0	0	0	0	0	0	1	0	0	0	11	0	0	|	o	=	UrticaDioica
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	|	p	=	PodocarpusSp
1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9	|	q	=	AccaSelloviana

4.3.4 K-Vecino más cercano

Este método pertenece al grupo de clasificadores lazy y está representado en la herramienta WEKA como IBk.

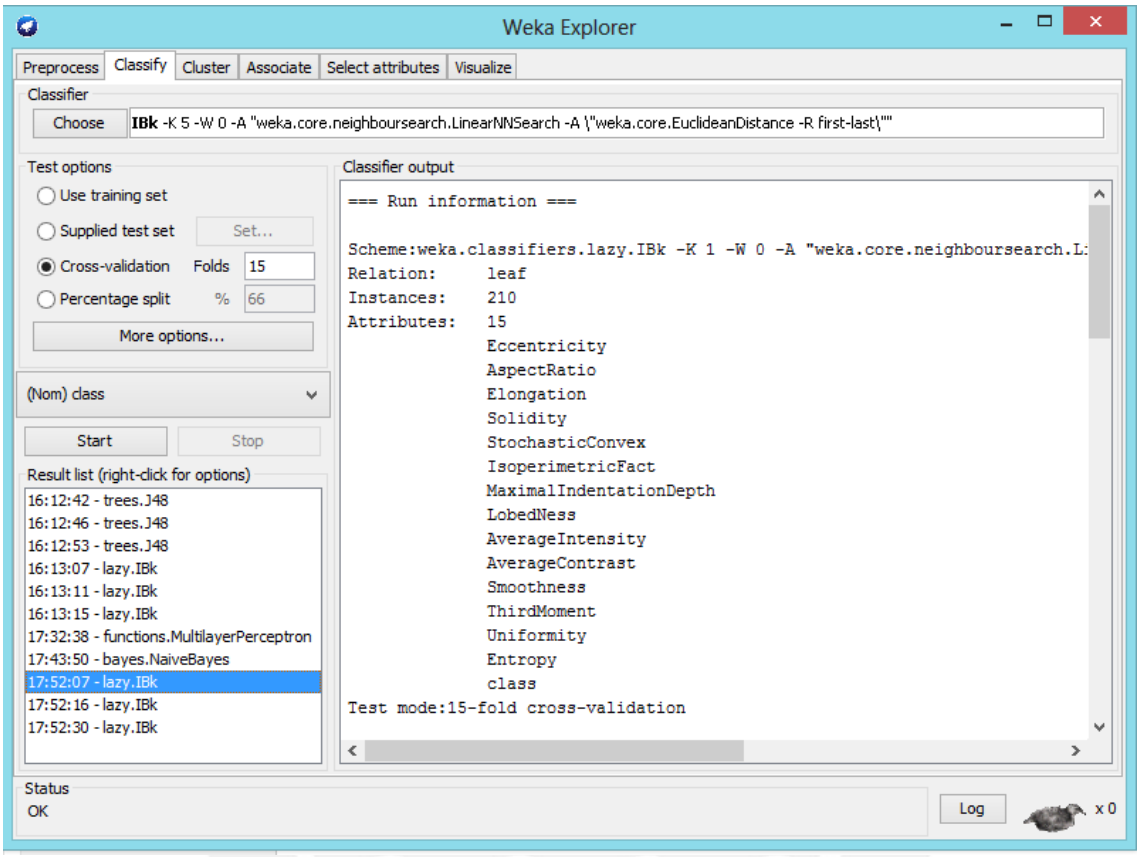


Figura 16: Detalle de la adaptación de K-Vecino más cercano.

- **Instancias correcta e incorrectamente clasificadas**

Tabla 18: Porcentaje de instancias correcta e incorrectamente clasificadas para el clasificador KNN.

Método	% Instancias correctamente clasificadas	% Instancias incorrectamente clasificadas
KNN (k=1)	78.57%	21.43%

- **Detalles de precisión por clase**

Tabla 19: Detalles de precisión para un clasificador KNN.

Tasa TP	Tasa FP	Precisión	Recall	F-Medida	Area ROC	Clase
0.75	0.015	0.75	0.75	0	0.867	QuercusSuber
1	0	1	1	1	1	QuercusRobur
1	0	1	1	1	1	NeriumOleander
0.786	0.026	0.688	0.786	0.733	0.88	BetulaPubescens
0.615	0.015	0.727	0.615	0.667	0.8	TiliaTomentosa
1	0	1	1	1	1	AcerPalmatum
0.667	0.02	0.667	0.667	0.667	0.823	CeltisSp
0.846	0.005	0.917	0.846	0.88	0.921	CorylusAvellana
0.583	0.01	0.778	0.583	0.667	0.787	CastaneaSativa
0.75	0.015	0.75	0.75	0.75	0.867	PrimulaVulgaris
0.769	0.025	0.667	0.769	0.714	0.872	BougainvilleaSp
0.583	0.02	0.636	0.583	0.609	0.782	EuonymusJaponicus
0.667	0.025	0.615	0.667	0.64	0.821	MagnoliaSoulangeana
0.917	0.005	0.917	0.917	0.917	0.956	BuxusSempervirens
0.667	0.02	0.667	0.667	0.667	0.823	UrticaDioica
1	0	1	1	1	1	PodocarpusSp
0.727	0.025	0.615	0.727	0.667	0.851	AccaSelloviana

- **Matriz de confusión**

Tabla 20: Matriz de confusión para un clasificador KNN.

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q				Clase clasificada como:
	9	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	|	a	=	QuercusSuber
0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|	b	=	QuercusRobur
0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|	c	=	NeriumOleander
0	0	0	11	1	0	0	0	0	1	0	0	0	0	1	0	0	|	d	=	BetulaPubescens
0	0	0	3	8	0	0	0	0	0	0	0	0	0	2	0	0	|	e	=	TiliaTomentosa
0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	|	f	=	AcerPalmatum
0	0	0	0	0	0	8	0	0	0	0	0	3	0	0	0	1	|	g	=	CeltisSp
0	0	0	0	0	0	0	11	0	0	1	1	0	0	0	0	0	|	h	=	CorylusAvellana
0	0	0	0	0	0	1	0	7	1	0	0	2	0	0	0	1	|	i	=	CastaneaSativa
0	0	0	0	0	0	1	0	0	9	0	1	0	0	0	0	1	|	j	=	PrimulaVulgaris
0	0	0	0	0	0	0	0	0	0	10	2	0	0	1	0	0	|	k	=	BougainvilleaSp
1	0	0	0	0	0	0	0	0	0	3	7	0	0	0	0	1	|	l	=	EuonymusJaponicus
0	0	0	0	0	0	2	0	2	0	0	0	8	0	0	0	0	|	m	=	MagnoliaSoulangeana
0	0	0	0	0	0	0	0	0	0	1	0	0	11	0	0	0	|	n	=	BuxusSempervirens
0	0	0	2	2	0	0	0	0	0	0	0	0	0	8	0	0	|	o	=	UrticaDioica
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	|	p	=	PodocarpusSp
2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	8	|	q	=	AccaSelloviana

4.4 Conclusión

Después de adaptar, evaluar y obtener el resultado de los diferentes criterios de precisión para cada método seleccionado, se concluye que todos los métodos tienen un porcentaje de precisión adecuado, en este caso mayor al 60%. Sin embargo, en primera instancia, las redes bayesianas son las que proporcionan una precisión de 81.9% sobrepasando al resto de métodos construidos. Por otro lado, es importante concluir que los resultados de los criterios de precisión son parte importante del proyecto, ya que la elección formal del método depende del análisis comparativo en base a los resultados obtenidos.

CAPÍTULO 5

Objetivo Específico 3 (OE3): Realizar un análisis comparativo, en base a criterios de precisión, que justifique la elección del método que será empleado en el proyecto.

5.1 Introducción

Después de obtener los resultados de los criterios de precisión para los diferentes métodos a emplearse, se procede a utilizar estos resultados como entrada para determinar el método más adecuado para la clasificación de hojas mediante un análisis que compara los resultados obtenidos. El presente capítulo detalla el análisis comparativo realizado para determinar el método más adecuado para el desarrollo de la herramienta de clasificación automática de hojas. El método usado es AUC (Área bajo la curva), el cual maneja gráficos ROC, los cuales proporcionan una medida más enriquecida de rendimiento que medidas escalares como: tasa de error, tasa de precisión, costo error [Fawcett 2004].

5.2 Resultado Esperado 4 (RE4): Clasificador seleccionado

En este apartado se detalla el método usado para determinar el método de clasificación que se adecue más al problema planteado y el método seleccionado propiamente dicho. El método a usarse será Área Bajo la Curva ROC (AUC). Antes de entender el funcionamiento del análisis, es importante entender el concepto relacionado a un gráfico ROC, curva de ROC.

- **Gráficos ROC *(Reciever Operating Characteristics)***

  Un gráfico ROC es un técnica usada para visualizar, organizar y seleccionar clasificadores basado en su rendimiento. Estos gráficos han sido utilizados durante mucho tiempo en la teoría de detección de señales para representar el equilibrio entre las tasas de éxito y las tasas de alarma de los clasificadores. Además de ser un gráfico de rendimiento útil, tienen propiedades que los hacen especialmente útiles para dominios con una distribución sesgada de clases.

  Los gráficos ROC son gráficos bidimensionales en donde el tasa TP es trazada en el eje Y y la tasa FP es trazada en el eje X. De esa manera, un gráfico ROC representa las ventajas y desventajas relacionadas con los beneficios (verdaderos positivos) y costo (falsos positivos). La figura 20 muestra un gráfico ROC con 5 clasificadores (A, B, C, D, E).

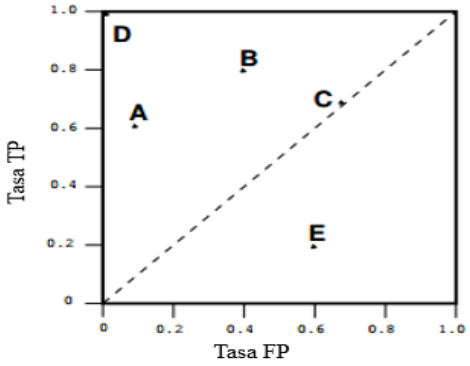


Figura 17: Gráfico ROC con cinco clasificadores [Fawcett 2004].

Es importante resaltar que los clasificadores representados en el gráfico ROC son conocidos como clasificadores discretos; es decir produce un par (tasa FP, tasa TP) correspondiente a un único punto en el espacio ROC. Adicionalmente, un punto en el espacio ROC es mejor que otro si este se ubica más al noroeste (tasa TP alto, tasa FP bajo). Los clasificadores que aparecen al lado izquierdo del espacio de ROC, cerca del eje X son considerados conservadores; es decir hacen clasificaciones positivas únicamente con evidencia fuerte. Los clasificadores que se encuentran en la parte alta del lado derecho del gráfico ROC son considerados liberales, ya que hacen clasificaciones positivas con evidencia débil; es decir clasifican casi todos los positivos correctamente pero obtienen una tasa FP alta [Fawcett 2004].

- **Gráficos de ROC Multiclases**

Cuando el número de clases dentro del conjunto de datos es mayor que dos, la situación se vuelve mucho más compleja si se intenta manejar todo el espacio de ROC. Con n clases, la matriz de confusión se convierte en una matriz n x n que contiene las n clasificaciones correctas y n^2 – n posibles errores. En los gráficos ROC multiclases no se maneja compensaciones entre la tasa TP y tasa FP, se maneja n beneficios y n^2-n errores. Un método para manejar n clases es producir n diferentes gráficos ROC; es decir uno por cada clase. Específicamente, si C es el conjunto de datos con todas las clases, el gráfico ROC i traza el rendimiento de la clasificación haciendo uso de la clase ci como la clase positiva y el resto de clases son consideradas clases negativas [Fawcett, 2004].

- **Curva de ROC**

Las curvas de ROC son una herramienta visual que es utilizada para comparar modelos de clasificación. Una curva de ROC muestra la compensación entre la tasa tp o sensibilidad (porcentaje de tuplas correctamente clasificadas) y la tasa fp

(proporción de tuplas negativas que son incorrectamente identificadas como positivas) para un modelo en específico [Han, et al. 2006].

- **AUC (Área bajo la curva)**

Una curva de ROC es una representación bidimensional del rendimiento de un clasificador. Para comparar el rendimiento de los clasificadores se desea reducirlo a un valor escalar que represente el rendimiento esperado. Un método común para obtener este valor escalar es el AUC o Área bajo la curva. El AUC es una porción del área de una unidad cuadrada, su valor estará en el rango de 0 y 1.0.

Por otro lado, debido a que una predicción aleatoria produce una línea diagonal entre (0,0) y (1,1), el cual tiene un área de 0.5, se recomienda que un clasificador no tenga un AUC menor a 0.5 [Fawcett, 2004].

- **Resultados de las AUC por cada método de clasificación**

A continuación se muestra la tabla con los valores de la tasa TP, tasa FP y AUC**.** Estos valores determinarán la elección del método de clasificación más adecuado.

Tabla 21: Valores de AUC para todos los métodos de clasificación descritos.

Método de clasificación	Tasa TP	Tasa FP	AUC
J48	0.767	0.014	0.899
Red Neuronal	0.819	0.011	0.977
Red Bayesiana	0.805	0.012	0.982
KNN	0.786	0.013	0.886

- **Resultados de las AUC por cada modelo de clasificación**

A continuación se muestra una tabla con los resultados de área bajo la curva ROC por cada clase para el modelo Red Bayesiana.

Tabla 22: Valores del AUC por clase para el modelo de Red Bayesiana.

Clase	AUC
QuercusSuber	0.974
QuercusRobur	0.985
NeriumOleander	1
BetulaPubescens	0.987
TiliaTomentosa	0.998
AcerPalmatum	1
CeltisSp	0.967
CorylusAvellana	0.995
CastaneaSativa	0.897
PrimulaVulgaris	0.976
BougainvilleaSp	0.98
EuonymusJaponicus	0.964
MagnoliaSoulangeana	0.975
BuxusSempervirens	0.999
UrticaDioica	0.999
PodocarpusSp	1
AccaSelloviana	0.991
Promedio	0.982

La elección de este método de análisis comparativo se basa en que los gráficos ROC son muy útiles para visualizar y evaluar clasificadores. Ellos proporcionan una medida más enriquecida del rendimiento de la clasificación que una medida escalar como precisión, tasa de error y costo de error. Estos gráficos separan el rendimiento del clasificador de los costos de error y clases sesgadas.

5.3 Conclusión

Tomando en cuenta la tabla anterior, se concluye que el método de clasificación más adecuado para realizar la clasificación automática de hojas es la red Bayesiana, ya que presenta el AUC más cercano a 1, en este caso 0.982.

Objetivo Específico 4 (OE4): Desarrollar un prototipo funcional que muestre el resultado de clasificar una nueva instancia de hoja haciendo uso de un modelo de clasificación.

6.1 Introducción

La técnica de Minería de Datos tiene como principal objetivo explorar y analizar una gran cantidad de datos con el fin de descubrir patrones y reglas significativas. Este proceso incluye el uso de algoritmos para hallar patrones en los datos y finalmente generar conocimiento [Berry & Linoff 1997] [Weiss and Davison 2010]. Estas características hacen posible la construcción de modelos de clasificación que permitan clasificar una nueva instancia de hoja. En base a esta idea, el presente capítulo presenta el funcionamiento del prototipo funcional desarrollado para el presente proyecto de fin de carrera, el cual permite clasificar una nueva instancia de hoja haciendo uso del modelo de clasificación seleccionado en el capítulo previo.

6.2 Resultado Esperado 5 (RE5): Prototipo funcional que muestre los resultados de clasificar una nueva instancia de hoja de planta.

6.2.1 Prototipo Funcional

En el presente apartado se explica el funcionamiento del prototipo desarrollado para cumplir con la tarea de clasificación de hojas. El prototipo presenta una estructura simple, fácil e intuitiva con la finalidad de que el usuario pueda clasificar una nueva hoja de planta de una manera rápida y sin problema alguno de entendimiento. A continuación se muestran los dos prototipos, el explorador y la vista principal de clasificación.

Figura 18: Prototipo inicial.

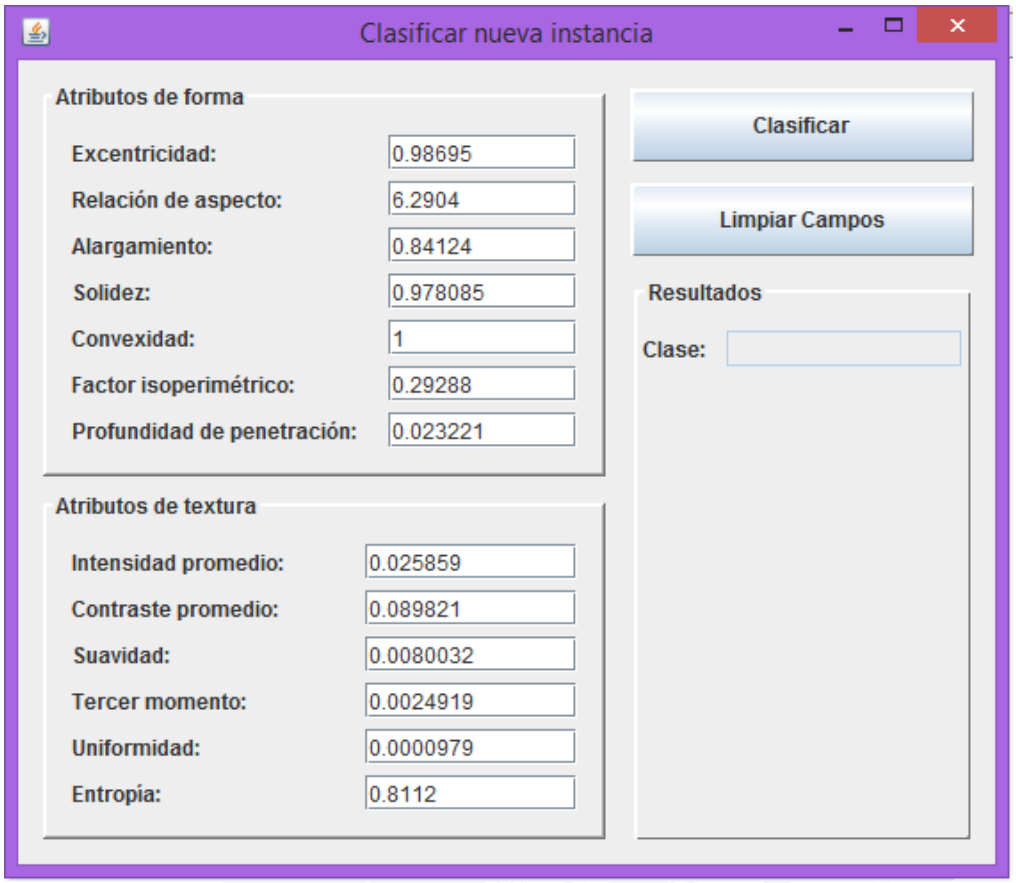


Figura 19: Prototipo funcional de clasificación.

En base a la Figura 19 se debe tener las siguientes consideraciones:

- Los campos están agrupados en atributos de forma, atributos de textura y resultados para mayor entendimiento del usuario.
- Los campos relacionados con los atributos de forma y textura se encontrarán inicialmente vacíos.
- El campo relacionado con el nombre de la clase se encontrará vacío inicialmente.
- Una vez ingresado los datos, se le permitirá al usuario limpiar los campos en caso haya ingresado algunos campos con valores incorrectos.

El funcionamiento del prototipo es el siguiente:

- El usuario debe ingresar obligatoriamente todos los campos mostrados en el prototipo.
- El usuario debe presionar el botón “Clasificar”.
- Se muestra al usuario el resultado de la clasificación.
- El usuario puede presionar el botón “Limpiar Campos” antes o después de realizar la clasificación.

6.2.2 Clasificación

El funcionamiento del prototipo tiene dos perspectivas, la externa en donde el usuario ingresa la información relacionada con la nueva instancia de hoja sin conocer cómo se llevará a cabo la clasificación y la segunda perspectiva que es la interna en donde se lleva a cabo la clasificación de la nueva instancia haciendo uso del modelo de clasificación que fue construido, evaluado y seleccionado como se explicó en los capítulos previos. El modelo de clasificación toma como parámetros los datos ingresados inicialmente por el usuario y le asigna una clase a la nueva instancia.

6.2.3 Resultado

Luego de conocer el funcionamiento interno y externo del prototipo, el resultado que se muestra finalmente es la clase a la que pertenece esa nueva instancia ingresada por el usuario, la cual se predijo haciendo uso de métodos inteligentes de Minería de datos, en este caso métodos de clasificación. A continuación se presenta una vista del prototipo funcional luego de haberse realizado la clasificación.

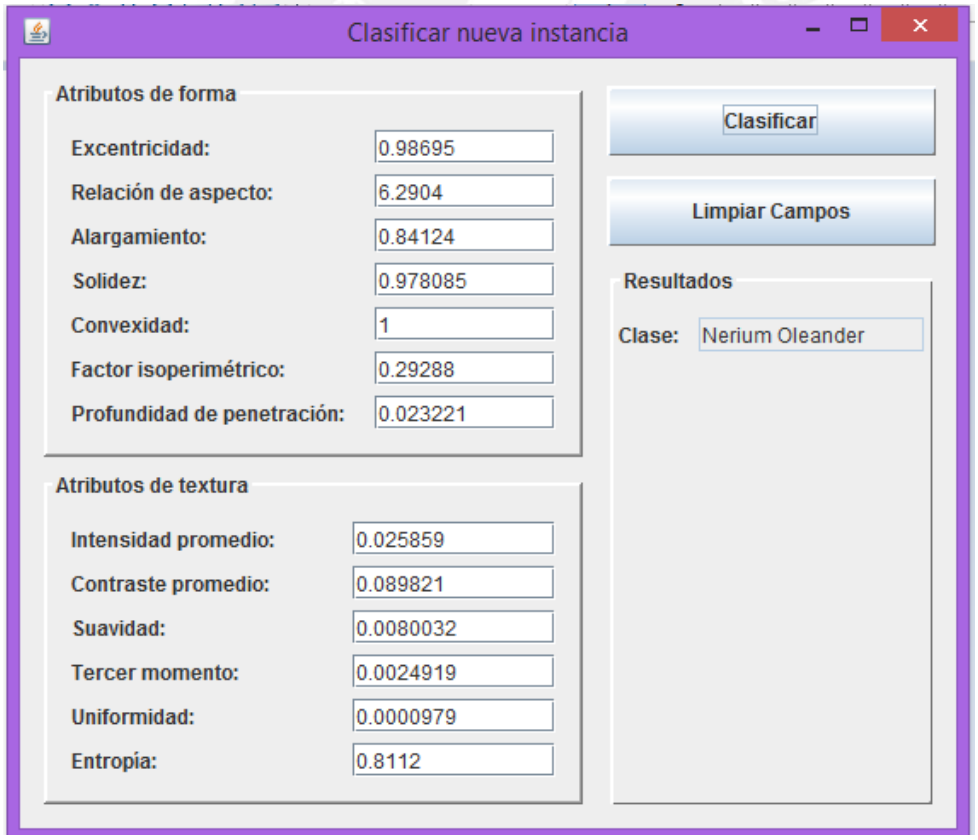


Figura 20: Prototipo funcional con el resultado de la clasificación.

6.3 Conclusión

En el presente capítulo se ha detallado el proceso de clasificar una nueva instancia de hoja haciendo uso de un prototipo funcional que integra un modelo de clasificación adaptado y evaluado en los capítulos previos a este. Por lo tanto, se concluye que desde la perspectiva del usuario final, el prototipo es simple, fácil e intuitivo, ya que solo requiere que el usuario ingrese los datos relacionados con la forma y textura de la hoja sin la necesidad de entender el modelo de clasificación que es usado internamente.

CAPÍTULO 7

7.1 Introducción

En el presente capítulo se detalla las conclusiones obtenidas después de desarrollar cada una de las etapas del proyecto. Además, se presenta algunas recomendaciones y trabajos futuros que pueden ayudar al desarrollo de una investigación como la realizada en el presente proyecto.

7.2 Conclusiones

En el presente apartado se presentan las conclusiones del presente proyecto de fin de carrera. Estas conclusiones se obtuvieron luego de completar los cuatro objetivos trazados para obtener un modelo de clasificación que permita clasificar nuevas instancias de hojas en base a sus atributos de forma y textura.

- Dado que el proceso de Minería de datos incluye el uso de algoritmos para hallar patrones en los datos y finalmente generar conocimiento, para el presente proyecto de fin de carrera se requiere un conjunto de datos que incluyan instancias de hojas clasificadas, debido a que el conjunto de datos determinará el comportamiento del modelo de clasificación. De esta forma, el proyecto tiene como punto de partida el uso de un conjunto de datos que incluye 210 instancias de hojas clasificadas. Cada instancia de hoja presenta 7 atributos de forma, 6 atributos de textura y la clase. Asimismo, es importante señalar que el proyecto trabaja únicamente con 17 clases de hojas debido al tipo de proyecto y tiempo estipulado para este. El conjunto de datos debe estar debidamente estructura, ya que de este depende la eficiencia y precisión de los modelos de clasificación.

- Teniendo en cuenta que dentro del proceso de Minería de datos existen una diversidad de métodos de clasificación, agrupación y regresión, el análisis desarrollado en el presente proyecto se limita a trabajar con 4 métodos de clasificación: Árboles de decisión, Redes neuronales, Redes bayesianas y k-Vecino más cercano. Estos cuatro métodos se seleccionaron teniendo como base investigaciones realizadas en temas similares al presente proyecto y de las cuales se obtuvieron resultados positivos. Por otro lado, con el conjunto de datos debidamente estructurado, se adaptan y evalúan los modelos de clasificación obteniéndose de estos una serie de criterios de precisión que son utilizados para evaluar su rendimiento.

- Luego de la construcción de los diferentes modelos de clasificación, se llevó a cabo un análisis comparativo en base a los resultados de los criterios de precisión de los modelos de clasificación seleccionado. El método de comparación fue AUC (Área bajo la curva), el cual hace uso de gráficos ROC para comparar el rendimiento de los modelos. En base a este análisis se demostró que la red bayesiana es el método más adecuado para solucionar el problema de la clasificación de hojas.

- Después de obtener el modelo de clasificación más adecuado para solucionar el problema de clasificación de hojas, el presente proyecto le provee al usuario un prototipo funciona como medio que le permita clasificar de manera fácil, simple e intuitiva una nueva hoja de planta sin la necesidad de involucrarse directamente con el modelo de clasificación usado.

- Finalmente, se concluye que después de completar los diferentes objetivos planteados para el proyecto, el problema de clasificación de hojas puede ser solucionado, con una alta precisión, haciendo uso del modelo de clasificación, el cual se obtuvo después de la adaptación y evaluación de una red bayesiana.

7.3 Recomendaciones y trabajos futuros

En este apartado se presenta las recomendaciones y trabajos futuros con la finalidad de proponer mejoras o nuevas investigación para solucionar el problema de clasificación de hojas.

- En primer lugar, se propone la implementación de un software que sea flexible; es decir que permita actualizar el modelo de clasificación de hojas cada cierto tiempo. La tarea de actualizar el modelo de clasificación dependerá de la actualización del conjunto de datos de entrada, el cual puede incrementarse en cuanto a cantidad de muestras o cantidad de clases debido a que nuevas especies de plantas van apareciendo con el paso del tiempo.

- Por otro lado, se propone un módulo en el que el usuario pueda configurar los parámetros para la construcción del modelo de clasificación con el fin de obtener mejores resultados para diferentes conjuntos de datos. Uno de los parámetros que podría incluirse en el módulo es el k de la Validación Cruzada.

- Dado que el conjunto de datos contiene instancias de hojas con atributos numéricos; es decir que se obtuvieron después de realizar procesamiento de imágenes de hojas, se propone un sistema que realice el procesamiento de cada imagen de hoja permitiendo extraer directamente los atributos y de esa manera construir el modelo de clasificación.

- Finalmente, dado que Minería de datos es una técnica que alberga una gran cantidad de métodos de clasificación, se propone analizar algunos métodos que no estén incluidos en el proyecto y que puedan solucionar el problema de clasificación de hojas. Los posibles métodos a utilizar son SVM (Support Vector Machine), variaciones del árbol de decisión, redes neuronales o redes bayesianas.

REFERENCIAS BIBLIOGRÁFICAS

Libros y artículos

L. Berg, *Introductory botany: plants, people, and the environment*: Cengage Learning, 2007.

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

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

J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.

I. Kononenko and M. Kukar, *Machine learning and data mining*: Elsevier, 2007.

L. Rokach, *Data mining with decision trees: theory and applications* vol. 69: World scientific, 2008.

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

S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang, and Q.-L. Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, 2007, pp. 11-16.

A. Gopal, S. Prudhveeswar Reddy, and V. Gayatri, "Classification of selected medicinal plants leaf using image processing," in *Machine Vision and Image Processing (MVIP), 2012 International Conference on*, 2012, pp. 5-8.

J. Y. Clark, D. P. Corney, and H. L. Tang, "Automated plant identification using artificial neural networks," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*, 2012, pp. 343-348.

M. E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," in *Computer Vision, Graphics & Image Processing, 2008. ICVGIP '08. Sixth Indian Conference on*, 2008, pp. 722-729.

S. F. Shazmeen, M. M. A. Baig, and M. R. Pawar, "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis."

M. J. Berry and G. Linoff, *Data mining techniques: for marketing, sales, and customer support*: John Wiley & Sons, Inc., 1997.

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.

I. Charalampopoulos and I. Anagnostopoulos, "A comparable study employing WEKA clustering/classification algorithms for web page classification," in *Informatics (PCI), 2011 15th Panhellenic Conference on*, 2011, pp. 235-239.

G. E. Gervilla, L. R. Jiménez, M. J. Montaño, A. A. Sesé, B. B. Cajal, and P. A. Palmer, "The methodology of Data Mining. An application to alcohol consumption in teenagers," *Adicciones,* vol. 21, pp. 65-80, 2008.

A. K. Pujari, *Data mining techniques*: Universities press, 2001.

J. D. Mauseth, *Botany: an introduction to plant biology*: Jones & Bartlett Publishers, 2012.

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

R. Stair and G. Reynolds, *Fundamentals of information systems*: Cengage Learning, 2008.

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.

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine,* vol. 17, p. 37, 1996.

J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," in *Machine learning*, ed: Springer, 1983, pp. 3-23.

B. Kitchenham, "Procedures for performing systematic reviews," *Keele, UK, Keele University,* vol. 33, p. 2004, 2004.

A. Kadir, L. E. Nugroho, A. Susanto, and P. I. Santosa, "Leaf classification using shape, color, and texture features," *arXiv preprint arXiv:1401.4447,* 2013.

S. Watcharabutsarakham, W. Sinthupinyo, and K. Kiratiratanapruk, "Leaf classification using structure features and Support Vector Machines," in *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in*, 2012, pp. 697-700.

A. Fernández, J. Luengo, J. Derrac, J. Alcalá-Fdez, and F. Herrera, "Implementation and integration of algorithms into the KEEL data-mining software tool," in *Intelligent Data Engineering and Automated Learning-IDEAL 2009*, ed: Springer, 2009, pp. 562-569.

J. Alcalá-Fdez, S. Garcia, F. J. Berlanga, A. Fernández, L. Sánchez, M. del Jesus*, et al.*, "KEEL: A data mining software tool integrating genetic fuzzy systems," in *Genetic and Evolving Systems, 2008. GEFS 2008. 3rd International Workshop on*, 2008, pp. 83-88.

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.

P. F. Silva, A. R. Marçal, and R. M. A. da Silva, "Evaluation of Features for Leaf Discrimination," in *Image Analysis and Recognition*, ed: Springer, 2013, pp. 197-204.