



**UNIVERSIDAD TÉCNICA DEL NORTE**  
**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**  
**CARRERA DE INGENIERÍA EN ELECTRÓNICA Y REDES DE**  
**COMUNICACIÓN**

**TRABAJO DE GRADO PREVIO A LA OBTENCIÓN DEL TÍTULO DE**  
**INGENIERA EN ELECTRÓNICA Y REDES DE COMUNICACIÓN**

**TEMA:**

**“ANÁLISIS DE DATOS BASADO EN TÉCNICAS DE BIG DATA Y DATA**  
**MINING PARA CULTIVOS DE HORTALIZAS EN EL INVERNADERO DE LA**  
**GRANJA LA PRADERA DE LA UNIVERSIDAD TÉCNICA DEL NORTE”**

**AUTORA:**

**KARINA LISETH PONCE GUEVARA**

**DIRECTOR:**

**MSc. EDGAR MAYA**

**IBARRA - ECUADOR**

**2017**



**UNIVERSIDAD TÉCNICA DEL NORTE**  
**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**BIBLIOTECA UNIVERSITARIA**

**AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA**  
**UNIVERSIDAD TÉCNICA DEL NORTE**

**1. IDENTIFICACIÓN DE LA OBRA**

La Universidad Técnica del Norte dentro del proyecto Repositorio Digital Institucional, determinó la necesidad de disponer de textos completos en formato digital con la finalidad de apoyar los procesos de investigación, docencia y extensión de la Universidad.

Por medio del presente documento dejo sentada mi voluntad de participar en este proyecto, para lo cual pongo a disposición la siguiente información:

<b>DATOS DEL CONTACTO</b>			
<b>CÉDULA DE IDENTIDAD:</b>		1004493209	
<b>APELLIDOS Y NOMBRES:</b>		Ponce Guevara Karina Liseth	
<b>DIRECCIÓN:</b>		Otavalo, Cdla.31 de Octubre	
<b>EMAIL:</b>		<a href="mailto:klponceg@utn.edu.ec">klponceg@utn.edu.ec</a>	
<b>TELÉFONO FIJO:</b>	062520420	<b>TELÉFONO MÓVIL</b>	0989577589
<b>DATOS DE LA OBRA</b>			
<b>TÍTULO:</b>		“Análisis de datos basado en técnicas de Big data y Data Mining para cultivos de hortalizas en el invernadero de la granja La Pradera de la Universidad Técnica del Norte”	
<b>AUTOR:</b>		Karina Liseth Ponce Guevara	
<b>FECHA:</b>		Abril de 2017	
<b>PROGRAMA:</b>		Pregrado	
<b>TÍTULO POR EL QUE OPTA:</b>		Ingeniería en Electrónica y Redes de Comunicación	
<b>ASESOR/DIRECTOR:</b>		MSc. Edgar Maya	

## **2. AUTORIZACIÓN DE USO A FAVOR DE LA UNIVERSIDAD**

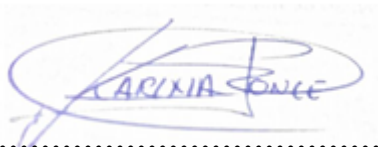
Yo, Karina Liseth Ponce Guevara con cédula de identidad Nro.100449320-9, en calidad de autor y titular de los derechos patrimoniales de la obra o trabajo de grado descrito anteriormente, hago entrega del ejemplar respectivo en formato digital y autorizo a la Universidad Técnica del Norte, la publicación de la obra en el Repositorio Digital Institucional y uso del archivo digital en la Biblioteca de la Universidad con fines académicos, para ampliar la disponibilidad del material y como apoyo a la educación, investigación y extensión; en concordancia con la Ley de Educación Superior Artículo 144.

## **3. CONSTANCIAS**

Yo, KARINA LISETH PONCE GUEVARA declaro bajo juramento que el trabajo aquí escrito es de mi autoría; y que este no ha sido previamente presentado para ningún grado o calificación profesional y que he consultado las referencias bibliográficas que se presentan en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondiente a este trabajo, a la Universidad Técnica del Norte, según lo establecido por las leyes de propiedad intelectual, reglamentos y normatividad vigente de la Universidad Técnica del Norte.

En la ciudad de Ibarra, abril de 2017.



Firma.....

**Karina Liseth Ponce Guevara**

**CI: 100449320-9**

**Ibarra, abril de 2017**



**UNIVERSIDAD TÉCNICA DEL NORTE**  
**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE GRADO A  
FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE**

Yo, KARINA LISETH PONCE GUEVARA, manifiesto mi voluntad de ceder a la Universidad Técnica del Norte los derechos patrimoniales consagrados en la Ley de Propiedad Intelectual del Ecuador artículos 4, 5 y 6, en calidad de autor del trabajo de grado con el tema: “ANÁLISIS DE DATOS BASADO EN TÉCNICAS DE BIG DATA Y DATA MINING PARA CULTIVOS DE HORTALIZAS EN EL INVERNADERO DE LA GRANJA LA PRADERA DE LA UNIVERSIDAD TÉCNICA DEL NORTE”, Que ha sido desarrollado con propósito de obtener el título de Ingeniero en Electrónica y Redes de Comunicación de la Universidad Técnica del Norte, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En mi condición de autor me reservo los derechos morales de la obra antes citada. En concordancia suscribo en el momento que hago entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Técnica del Norte.

Karina Liseth Ponce Guevara

100449320-9

Ibarra, abril 2017



**UNIVERSIDAD TÉCNICA DEL NORTE**  
**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**CERTIFICACIÓN**

MAGISTER EDGAR MAYA, DIRECTOR DEL PRESENTE TRABAJO DE TITULACIÓN  
CERTIFICA:

Que, el presente trabajo de Titulación “ANÁLISIS DE DATOS BASADO EN TÉCNICAS DE BIG DATA Y DATA MINING PARA CULTIVOS DE HORTALIZAS EN EL INVERNADERO DE LA GRANJA LA PRADERA DE LA UNIVERSIDAD TÉCNICA DEL NORTE” ha sido realizada en su totalidad por: KARINA LISETH PONCE GUEVARA bajo mi supervisión:

Es todo en cuanto puedo certificar en honor de la verdad.

.....

MSc. Edgar Maya

Director de Tesis



**UNIVERSIDAD TÉCNICA DEL NORTE**  
**FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS**

**DEDICATORIA**

*A mi madre Germania Guevara, quien ha sido el motor de mi vida y el ángel que siempre estuvo conmigo en los momentos de dificultad. Todos mis logros te los debo a ti, te amo.*

*A mis hermanos, Giss tu gran sentido del humor hace que cada día agradezca a Dios por tenerte en mi vida, Danny has sido como un padre, tus consejos han hecho de mi la mujer fuerte y decidida que poco a poco va cumpliendo sus metas, y a Germa que desde el cielo cuidas nuestros pasos, amarlos más que a mi vida.*

*A mis sobrinos Julissa y Rajnesh, son la luz que ilumina mi vida, con cada una de sus ocurrencias y travesuras me llenan de amor.*

*Karina*

## AGRADECIMIENTO

*Querido Dios, te agradezco por las bendiciones, por Tu misericordia y por mostrarme Tu camino, en Tus manos confío mi mañana.*

*Realizar este proyecto de titulación conllevó un gran esfuerzo, pero con la ayuda y guía correctos fue posible culminarlo a tiempo. Quiero agradecer al Ing. Edgar Maya que, siendo mi tutor en este proceso, compartió sus conocimientos y las pautas esenciales para lograrlo.*

*A Diego por ser un amigo incondicional que siempre busca la superación de sus estudiantes. Gracias por solventar todas mis inquietudes y ser el pilar fundamental para el desarrollo y conclusión de este proyecto.*

*De manera especial quiero agradecer al Ing. Juan Carlos Alvarado, quien compartió sus conocimientos acerca de Big Data y la minería de datos. Gracias por responder a mis dudas, sin tu apoyo esto no habría sido posible.*

*Estudiar esta carrera ha sido un reto, cada semestre exigía un poco más de esfuerzo. Agradezco también a mis docentes que impartieron sus conocimientos y me formaron como profesional, y al apoyo incondicional de mis amigos, gracias por estar siempre presentes, incluyendo los momentos más difíciles.*

*A la Asociación de Estudiantes CIERCOM, los tres hicimos que estos dos años al frente sean los mejores de nuestra vida universitaria. Como en todo lugar surgieron situaciones complicadas, pero siempre supimos solventarlas, esto sirvió para enriquecernos como personas y llenarnos de experiencias únicas. Gracias a ustedes y al apoyo incondicional de nuestros amigos, nuestra gestión fue exitosa.*

*Por último y no menos importante, quiero expresar mi gratitud a toda mi familia, mis tíos, primos, que con sus palabras de aliento hicieron que nunca desistiera, gracias por ser la familia bullying más adorable que pudo existir. A mi abue Lucilita te amo mi viejita linda, con tus frases, ocurrencias y el amor que nos brindas cada día nos haces felices, eres mi pilar y de toda la familia. Y a mi abuelito Emilio que, desde el cielo guía cada uno de mis pasos, muchas gracias y te extraño mucho.*

*Karina\**

## ÍNDICE DE CONTENIDOS

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE .....	ii
CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE GRADO A FAVOR DE LA UNIVERSIDAD TÉCNICA DEL NORTE.....	iv
CERTIFICACIÓN .....	v
DEDICATORIA.....	vi
AGRADECIMIENTO.....	vii
ÍNDICE DE CONTENIDOS .....	viii
ÍNDICE DE FIGURAS.....	xiii
ÍNDICE DE TABLAS .....	xvii
RESUMEN.....	xviii
ABSTRACT.....	xix
CAPÍTULO I.....	1
ANTECEDENTES.....	1
1.1. INTRODUCCIÓN .....	1
1.2. PROBLEMA.....	1
1.3. OBJETIVOS .....	2
1.3.1. Objetivo General.....	2
1.3.2. Objetivos Específicos.....	3
1.4. ALCANCE.....	3
1.5. JUSTIFICACIÓN .....	5



CAPÍTULO II .....	7
FUNDAMENTACIÓN TEÓRICA.....	7
2.1. INTRODUCCIÓN .....	7
2.2. CONTEXTO.....	7
2.2.1. Invernadero .....	8
2.2.2. Factores o variables a medir .....	9
2.3. BIG DATA.....	11
2.3.1. Herramientas para Big Data y minería de datos .....	14
2.4. ADQUISICIÓN DE DATOS Y BASES DE DATOS.....	21
2.4.1. Redes de sensores inalámbricos (WSN) .....	22
2.4.2. Bodega de datos .....	23
2.4.3. Base de datos.....	24
2.5. PROCESO KDD.....	24
2.5.1. Bodega de datos .....	26
2.5.2. Limpieza de datos (Data Cleaning).....	27
2.5.3. Minería de datos.....	28
2.5.4. Clasificación .....	30
CAPÍTULO III DISEÑO .....	37
3.1. INTRODUCCIÓN .....	37
3.2. DESCRIPCIÓN GENERAL.....	37
3.3. PROCESO KDD.....	38
3.3.1. Datos .....	39
3.3.2. Etapa de selección.....	41

3.3.3.	Etapa de pre-procesamiento/limpieza .....	42
3.3.4.	Etapa de transformación/reducción.....	43
3.3.5.	Etapa de minería de datos .....	44
3.3.6.	Etapa de interpretación y evaluación .....	45
3.4.	METODOLOGÍA DEL ANÁLISIS DE DATOS CRISP-DM .....	46
3.4.1.	Etapas de la metodología CRISP-DM .....	47
3.5.	DIAGRAMAS DE CASO DE USO .....	50
3.5.1.	Diagrama general de la interfaz de análisis de datos. ....	50
3.5.2.	Diagrama de caso de uso con predicción.....	52
3.6.	MÓDULOS DE LA INTERFAZ.....	53
3.6.1.	Módulo Datos.....	53
3.6.2.	Módulo Pre procesamiento .....	54
3.6.3.	Módulo Clasificación.....	54
3.6.4.	Módulo Visualización.....	55
3.6.5.	Módulo Predicción.....	55
3.6.6.	Módulo Interfaz .....	55
CAPÍTULO IV IMPLEMENTACIÓN .....		56
4.1.	INTRODUCCIÓN .....	56
4.2.	DESCRIPCIÓN DE LA HERRAMIENTA.....	56
4.3.	DESARROLLO DEL SOFTWARE.....	57
4.3.1.	Paquete utilidades .....	57
4.3.2.	Paquete filtros .....	58
4.3.3.	Paquete de algoritmos.....	59

4.3.4. Paquete gui.....	60
4.3.5. Interfaz de análisis de datos .....	61
4.3.6. Algoritmos y técnicas .....	64
4.4. PRUEBAS DE FUNCIONAMIENTO y RESULTADOS .....	71
4.4.1. Pruebas Usando el Data Set for Sustainability .....	71
4.4.2. Pruebas usando los datos reales del invernadero .....	82
4.4.3. Resultados .....	93
CAPÍTULO V .....	95
5.1. CONCLUSIONES .....	95
5.2. RECOMENDACIONES.....	98
5.3. GLOSARIO DE TÉRMINOS.....	99
REFERENCIAS .....	100
ANEXOS.....	105
ANEXO A.....	105
Estructura de los datos del Data Set for Sustainability .....	105
Estructura de los datos recolectados del invernadero.....	106
ANEXO B .....	107
EJEMPLO FUNCIONAMIENTO DEL ALGORITMO C4.5.....	107
ANEXO C: LIBRERÍAS USADAS .....	108
Java csv .....	108
Jama.....	108
Log4j .....	108
SLF4J .....	108

Swingx.....	108
Jcommon .....	109
JFreeChart .....	109
Prefuse.....	110
Substance.....	110
ANEXO D DIAGRAMAS DE PAQUETES Y CLASES .....	111

## ÍNDICE DE FIGURAS

Figura 1 Ejemplo de un invernadero de recubrimiento de plástico y cristal .....	8
Figura 2 Logo de Apache Hadoop .....	15
Figura 3 Uso de la interfaz web de Spark. ....	16
Figura 4 Ejemplo de Cluster jerárquico realizado en Orange.....	17
Figura 5 Uso del panel de clasificador en Weka.....	18
Figura 6 Navegador Anaconda, esta permite la selección de las herramientas para el desarrollo de software.....	19
Figura 7 Aplicación de clasificación de datos realizada en MatLab .....	20
Figura 8 Interfaz de clustering desarrollada en java. ....	21
Figura 9 Red de sensores inalámbricos WSN con el protocolo IEEE 802.15.4. ....	23
Figura 10 Proceso de KDD .....	25
Figura 11 Características de Data Mining .....	29
Figura 12 Diagrama explicativo de clasificación de datos de dos clases o especies .....	30
Figura 13 Algoritmo kNN con dos clases.....	32
Figura 14 Diagrama explicativo de la j-ésima red de un sistema neuronal de Q+M entradas .....	33
Figura 15 Etapas del proceso KDD .....	38
Figura 16 UMass Trace Repository .....	40
Figura 17 Registro para el acceso al repositorio.....	40
Figura 18 Archivo .csv del Data Set for Sustainability .....	41
Figura 19 Metodología CRISP-DM involucrando al proceso KDD.....	46
Figura 20 Riego en el invernadero de la granja La Pradera.....	47
Figura 21 Diagrama de la arquitectura de una herramienta débilmente acoplada. ....	49

Figura 22 Caso de uso del funcionamiento general de la interfaz .....	50
Figura 23 Caso de uso del funcionamiento de la interfaz con el modelo predictivo. ....	52
Figura 24 Funcionamiento de la interfaz en estructura de módulos .....	53
Figura 25 Paquetes y clases usados en la programación de la interfaz.....	57
Figura 26 Pantalla de inicio de la interfaz gráfica de usuario. ....	62
Figura 27 Texto mostrado por el botón informativo de la pantalla de inicio. ....	62
Figura 28 Descripción de los íconos formados por la pestaña Inicio .....	63
Figura 29 Descripción de los íconos formados en la pestaña Herramientas.....	64
Figura 30 Funcionamiento módulo datos de interfaz de análisis de datos. ....	65
Figura 31 Funcionamiento módulo selección en la interfaz de análisis de datos. ....	66
Figura 32 Funcionamiento módulo clasificación en la interfaz de análisis de datos. ....	67
Figura 33 Funcionamiento módulo visualización en la interfaz de análisis de datos. ....	68
Figura 34 Visualización del árbol de decisión en la interfaz de análisis de datos. ....	68
Figura 35 Visualización de las reglas de clasificación en la interfaz de análisis de datos. ....	69
Figura 36 Funcionamiento del módulo predicción en la interfaz de análisis de datos. ....	70
Figura 37 Visualización de la predicción de la variable target (Humedad del suelo) .....	70
Figura 38 Pantalla de inicio de la interfaz de análisis de datos en la prueba usando el Data Set for Sustainability.....	71
Figura 39 Elementos en el canvas de la interfaz de análisis de datos en la prueba usando el Data Set for Sustainability. ....	72
Figura 40 Carga de datos y visualización n en la prueba usando el Data Set for Sustainability. .	73
Figura 41 Datos visualizados en la prueba usando el Data Set for Sustainability. ....	73
Figura 42 Selección de la variable objetivo en la prueba usando el Data Set for Sustainability..	74

Figura 43 Conexión entre las etapas de selección y clasificación en la prueba usando el Data Set for Sustainability.....	75
Figura 44 Configuración de los parámetros del algoritmo c4.5 en la prueba usando el Data Set for Sustainability.....	75
Figura 45 Conexión entre la etapa de clasificación y visualización en la prueba usando el Data Set for Sustainability.....	76
Figura 46 Reglas de clasificación en la prueba usando el Data Set for Sustainability. ....	77
Figura 47 Visualización del árbol de decisión en la prueba usando el Data Set for Sustainability. ....	77
Figura 48 Íconos del proceso de predicción en la prueba usando el Data Set for Sustainability. ....	78
Figura 49 Carga del nuevo set de datos en la prueba usando el Data Set for Sustainability. ....	79
Figura 50 Ejecución de la etapa de predicción usando el Data Set for Sustainability.....	79
Figura 51 Ejecución de la etapa de predicción usando el Data Set for Sustainability.....	80
Figura 52 Resultados del proceso de predicción usando el Data Set for Sustainability. ....	81
Figura 53 Almacenamiento de los resultados usando el Data Set for Sustainability.....	81
Figura 54 Carga de datos obtenidos del invernadero, prueba con los datos reales.....	82
Figura 55 Datos obtenidos del invernadero, prueba con los datos reales. ....	83
Figura 56 Selección de la variable objetivo (humedad del suelo), prueba con los datos reales. ..	83
Figura 57 Selección de la variable objetivo (humedad del suelo), prueba con los datos reales. ..	84
Figura 58 Proceso de clasificación, prueba con los datos reales. ....	84
Figura 59 Proceso de clasificación, prueba con los datos reales. ....	85
Figura 60 Reglas de clasificación, prueba con los datos reales. ....	86
Figura 61 Árbol de decisión, prueba con los datos reales.....	87

Figura 62	Árbol de decisión, prueba con los datos reales.....	87
Figura 63	Árbol de decisión con los datos discretizados, prueba con los datos reales. ....	88
Figura 64	Íconos ubicados en el canvas para la predicción de datos, prueba con los datos reales. .....	89
Figura 65	Nuevo conjunto de datos cargado para la predicción, con la variable Humedad del Suelo, prueba con los datos reales. ....	90
Figura 66	Visualización de los resultados de predicción, con la variable Humedad del suelo, prueba con los datos reales. ....	91
Figura 67	Íconos ubicados en el canvas para la predicción de datos sin la variable Humedad del Suelo, prueba con los datos reales. ....	91
Figura 68	Carga del nuevo conjunto de datos sin la variable Humedad del suelo, prueba con los datos reales.....	92
Figura 69	Visualización de resultados de predicción de la variable Humedad del suelo, prueba con los datos reales. ....	92
Figura 70	Ejemplo del funcionamiento del Algoritmo C4.5 .....	107
Figura 71	Diagrama de paquetes de la interfaz.....	111
Figura 72	Diagrama de paquetes de GUI, interfaz gráfica de usuario.....	111
Figura 73	Diagrama de paquetes del paquete de utilidades.....	112
Figura 74	Diagrama de clases del algoritmo C4.5.....	113
Figura 75	Diagrama de clases del paquete de interfaz gráfica de usuario. ....	114



## ÍNDICE DE TABLAS

Tabla 1 Parámetros de Big Data .....	13
Tabla 2 Análisis de los factores que inciden en un cultivo.....	42
Tabla 3 Tipo de implementación de herramientas de minería de datos.....	48
Tabla 4 Usuarios de la interfaz .....	51
Tabla 5 Estructura de los datos almacenados en el invernadero.....	54
Tabla 6 Componentes del paquete utilidades.....	58
Tabla 7 Componentes del paquete Filtros.....	58
Tabla 8 Componentes del paquete algoritmos .....	59
Tabla 9 Componentes del paquete GUI.....	60
Tabla 10 Resultados de las pruebas con los diferentes archivos de datos .....	93

## RESUMEN

El presente proyecto es una propuesta del uso de técnicas de Big Data y Data Mining (minería de datos) aplicados a cultivos de hortalizas en el invernadero de la granja “La Pradera”, con el objetivo de analizar los factores que influyen en el crecimiento de los cultivos, y determinar un modelo predictivo de la humedad del suelo.

Dentro de un invernadero, las variables que inciden en el crecimiento de los cultivos son: Humedad relativa, humedad del suelo, temperatura ambiental, y niveles de iluminación y CO<sub>2</sub>. Estos parámetros son esenciales para la fotosíntesis, es decir, durante los procesos donde las plantas adquieren la mayoría de nutrientes, y por tanto, con un buen control de dichos parámetros las plantas podrían crecer más sanas y producir mejores frutos. El proceso de análisis de los factores en un contexto de minería de datos requiere diseñar un sistema de análisis y establecer una variable objetivo a ser predicha por el sistema. En este caso, con el fin de optimizar el gasto de recurso hídrico, se ha escogido como variable objetivo la humedad del suelo.

El sistema de análisis propuesto es desarrollado en una interfaz de usuario implementada en Java y NetBeans IDE 8.2, y consta principalmente de dos etapas: Una de ellas es la clasificación a través del algoritmo C4.5, el cual emplea un árbol de decisión basado en la entropía de los datos, y permite visualizar los resultados de manera gráfica. La segunda etapa principal es la predicción, en la cual, a partir de la clasificación obtenida en la etapa anterior, se predice la variable objetivo con base en un nuevo conjunto de datos. En otras palabras, la interfaz construye un modelo predictivo para determinar el comportamiento de la humedad de suelo.

## ABSTRACT

This work outlines the use of Big Data and Data Mining techniques on vegetable crops data from the greenhouse of the farm "The Pradera", which is aimed at analyzing the factors that influence the growth of the crops, and determine a predictive model of soil moisture.

Within a greenhouse, the variables that affect crop growth are: relative humidity, soil moisture, ambient temperature, and levels of illumination and CO<sub>2</sub>. These parameters are essential for photosynthesis, i.e. during processes where plants acquire the most nutrients, and therefore, if performing a good control on these parameters, plants might grow healthier and produce better fruits. The process of analysis of such factors in a data mining context requires designing an analysis system and establishing an objective variable to be predicted by the system. In this case, in order to optimize water resource expenditure, soil moisture has been chosen as the target variable.

The proposed analysis system is developed in a user interface implemented in Java and NetBeans IDE 8.2, and consists mainly of two stages. One of them is the classification through algorithm C4.5, which uses a decision tree based on the data entropy, and allows to visualize the results graphically. The second main stage is the prediction, in which, from the classification results obtained in the previous stage, the target variable is predicted from information of a new set of data. In other words, the interface builds a predictive model to determine the behavior of soil moisture.

# CAPÍTULO I

## ANTECEDENTES

### 1.1. INTRODUCCIÓN

Big Data es un concepto que ha tomado fuerza en los últimos años, debido principalmente a la gran cantidad de información que se genera en diversos contextos, entre ellos: Economía, educación, medio ambiente, redes de sensores y redes móviles. Por tanto, se puede decir que se ha convertido en un área multidisciplinaria. En efecto, existen diversas fuentes generando un flujo de información, el cual, a través de un procesado permite el análisis de diferentes variables (atributos o características) y provee elementos para realizar una toma de decisiones inteligente.

Particularmente, en la agricultura se tiene varios factores y variables que producen una cantidad inimaginable de información. A través de herramientas que usan internet de las cosas (IoT - *Internet of Things*) e internet de todo (IoE - *Internet of Everything*), es decir, haciendo uso de sistemas embebidos con redes de sensores inalámbricos, puede realizarse la adquisición de datos masivos y, adicionalmente, usando herramientas de análisis de datos puede lograrse una gestión agrícola sustentable que, en cierta medida, genere impactos favorables al medio ambiente debido a que una adecuada toma de decisiones permitiría optimizar los recursos naturales.

### 1.2. PROBLEMA

En la granja la pradera de la Universidad Técnica del Norte ubicada en el sector de Chaltura, se ha construido un invernadero para cultivos de diferentes hortalizas. Dicho sembrío ha reactivado su función gracias a profesores y estudiantes de la carrera de Ingeniería Agropecuaria.

Inicialmente, ellos han empezado por cultivar tomate riñón, no obstante, el plan es seguir con el sembrado de otras hortalizas. El riego que usa este invernadero, emplea una técnica irrigación por goteo, la cual consiste en un sistema de mangueras agujeradas que se extienden a lo largo del sembrío y, de esta manera, cada planta recibe agua suficiente; sin embargo, sigue siendo una técnica artesanal, manual y muy sujeta a fallos debido a la falta de monitoreo.

En el análisis del comportamiento de los cultivos se necesita tomar en cuenta variables que permitan al agropecuario evaluar el avance de los sembríos y la manera de proceder frente a alguna eventualidad. Particularmente, este invernadero no tiene implementado ninguna técnica de evaluación de los diversos factores que inciden en el desarrollo y crecimiento de las plantas; entre dichos factores se encuentra la temperatura ambiental, la humedad del suelo, la iluminación, e incluso el nivel de CO<sub>2</sub> (Castilla, 2007).

Por tanto, el estado de los cultivos se basa en la observación y no se tiene un almacenamiento de la información ordenada para su posterior análisis. Con el avance tecnológico y el desarrollo de áreas como Big Data y Data Mining, es decir, el análisis y exploración de grandes volúmenes de datos, se puede obtener los informes del sembrío en tiempo real. Además, a través del uso de la nube, dichos datos podrían ser analizados en un computador conectado a la red, obteniendo así una base de datos con un registro de los diferentes factores.

## **1.3. OBJETIVOS**

### **1.3.1. Objetivo General**

Diseñar una interfaz de análisis de datos basado en técnicas de Big Data y Data Mining para cultivos de hortalizas en el invernadero de la granja “La Pradera” de la Universidad Técnica del Norte”.

### **1.3.2. Objetivos Específicos**

- Investigar las normas, estándares y términos que utiliza Big Data y las técnicas de Data Mining para el análisis estadístico de datos, así como, la información sobre los factores que inciden en la agricultura de precisión, los cultivos en invernaderos y la relación entre estos.
- Buscar y analizar repositorios y bases de datos en Internet, las cuales permitan realizar el estudio las variables: humedad del suelo, humedad relativa, temperatura ambiental, niveles de iluminación y CO<sub>2</sub>.
- Diseñar una interfaz con técnicas de Big Data para cultivos en invernaderos que permita hacer predicciones del comportamiento de las variables como: humedad del suelo, humedad relativa, temperatura ambiental, niveles de iluminación y CO<sub>2</sub>.
- Realizar pruebas de funcionamiento de la plataforma para comprobar el correcto desarrollo del análisis de los datos con técnicas de Big Data.

## **1.4. ALCANCE**

Big Data es una metodología que implica el análisis de grandes cantidades de datos que se generan en ambientes donde las variables no pueden ser inspeccionadas directamente por los usuarios, es decir, no son inteligibles. Este estudio realiza un proceso basado en técnicas de minería de datos, las cuales aplicadas a la agricultura, se enfocan en el comportamiento de las variables o factores que inciden en el crecimiento y desarrollo de los cultivos. Por tanto, este proceso recibe la connotación de agricultura de precisión, y se realiza respetando las normas y estándares establecidos para el uso de la tecnología aplicada.

Para que el estudio propuesto recaiga dentro del área de Big Data, estrictamente dicha, debe considerarse un número suficientemente grande de muestras y un número de variables que no permitan un análisis visual directo (más de tres variables o no inteligibles para el ser humano). Una vez cumplida esta condición, el sistema permite establecer una relación las variables consideradas (temperatura ambiental, humedad del suelo, humedad relativa, iluminación y niveles de CO<sub>2</sub>) y lograr predicciones en el comportamiento del invernadero. Además, se estudia cómo éstas influyen en las otras para el buen desarrollo de las plantas, de esta manera, compararlas con las condiciones ideales.

En el invernadero, con el análisis de variables tales como temperatura ambiente, humedad de suelo, humedad relativa, iluminación y nivel de Co<sub>2</sub>, se busca analizar el comportamiento de éstas, y determinar cómo inciden en el desarrollo de los cultivos mediante un sistema de análisis y una interfaz de computador que permitan visualizar los resultados obtenidos. Cabe resaltar que, para evaluar el sistema, se necesita de una base de datos pre-establecida para realizar el análisis con técnicas de Big Data y minería de datos, razón por la cual, en este estudio, se usan bases de datos de repositorios de libre acceso en la internet, los cuales involucran las variables necesarias. Dichas bases de datos son seleccionadas bajo el criterio de que emulen el comportamiento de las variables reales de forma confiable y que los datos hayan sido adquiridos en condiciones de contexto similares al del invernadero real. Con las pruebas realizadas se puede observar el análisis, de forma que se pueda estudiar el comportamiento y la predicción de las variables en los cultivos del invernadero.

## 1.5. JUSTIFICACIÓN

El uso de Técnicas de Big Data y minería de datos en los procesos agrícolas es un tema relativamente nuevo y no ampliamente conocido en el país, ni a nivel internacional. Actualmente, muchos de los datos a nivel agrícola son tomados manual o artesanalmente, y no existe un almacenamiento ordenado, ni se explota dicha información para obtener diferentes resultados. Una buena alternativa para superar estos inconvenientes, es la implementación de una agricultura de precisión, la cual, con el uso de una base científica y documental, permita el aprovechamiento máximo de los recursos (El Comercio, 2015).

El Plan Nacional del Buen Vivir 2013-2017 plantea la Revolución del Conocimiento, proponiendo la innovación, la ciencia y la tecnología, como fundamentos para el cambio de la matriz productiva (Senplades, 2013). El presente proyecto es una aplicación de técnicas de Big Data y minería de datos en un invernadero de la granja “La Pradera”, donde se obtendrán datos masivamente para su posterior análisis estadístico, usando las variables más importantes que inciden el desarrollo de los cultivos, con el fin de desarrollar un enfoque de agricultura de precisión en dicho invernadero. Además, con este tipo de proyectos se aporta al cumplimiento de la misión de la Universidad Técnica del Norte, la cual busca la generación procesos de investigación, de transferencia de saberes, de conocimientos científicos, tecnológicos y de innovación (UTN, 2008).

Con la incursión tecnológica que hoy en día se promueve en casi todos los campos, el sector de la agricultura genera grandes cantidades de datos en las fases de pre y post producción, siendo necesaria la aplicación de técnicas que permitan analizar estos datos a gran escala y, de esta manera, generar una base de datos ordenada con la información de los diversos factores que afectan a los cultivos.



El invernadero en el que se va a plantear este proyecto, pertenece a la UTN y con la aplicación de herramientas tecnológicas como una WSN para la recolección de datos, se busca implementar una aplicación de técnicas de Big Data y minería de datos para agricultura de precisión.

## **CAPÍTULO II**

### **FUNDAMENTACIÓN TEÓRICA**

#### **2.1. INTRODUCCIÓN**

En este capítulo se trata los parámetros y aspectos necesarios para realizar el análisis de datos con técnicas de Big Data y minería de datos. A continuación, se describe cada uno de éstos y su aplicación en diversas áreas, siendo una de ellas la agricultura de precisión. El área Big Data no se refiere únicamente al aumento de la cantidad de datos, sino que también implica el análisis de los mismos para la toma de decisiones inteligente.

#### **2.2. CONTEXTO**

Con el avance tecnológico y científico, el sector agropecuario se ha convertido naturalmente en un área potencial de aplicación donde las técnicas de Big Data pueden incursionar. Es un tema que debe seguir en desarrollo, ya que generará sustentabilidad del modelo de negocio, a través del descubrimiento de patrones en bases de datos (Malvicinoa & Yoguelb, 2015).

El presente proyecto está enfocado al análisis de datos mediante el uso de técnicas de Big Data y de minería de datos en un invernadero de la granja “La Pradera” de la Universidad Técnica del Norte. El objetivo de este análisis es la predicción del comportamiento de los factores que inciden en el desarrollo de los cultivos y de esta manera influir en la toma de decisiones.

Las variables que se va a tomar en cuenta para el análisis de datos son: Humedad del suelo, humedad relativa, temperatura, y niveles de iluminación y CO<sub>2</sub>. Estos factores ambientales son los que intervienen en el crecimiento de las plantas, y de éstos depende el correcto desarrollo de los cultivos (Iglesias N. , 2009).

### 2.2.1. Invernadero

Un invernadero es una construcción de madera, hierro u otro material, que tradicionalmente está cubierto por cristales, aunque existen modelos básicos cubiertos por plástico. En general, su estructura está provista de calefacción y está iluminada artificialmente, y por tanto en su interior es factible cultivar diferentes hortalizas, flores u otras plantas fuera de su estación. Los materiales usados para recubrirlos, así como los sistemas de control de los factores ambientales son de gran variedad, en la Figura 1 se muestra la estructura de un invernadero recubierto de plástico. (Alpi & Tognoni, 1999).



Figura 1 Ejemplo de un invernadero de recubrimiento de plástico y cristal  
Fuente: (Asthor, 2016)

Desde hace varios años, los cultivos hortícolas se producen de manera anticipada y fuera de temporada dentro de estos invernaderos, gracias al uso de diversos sistemas protectores que permiten tener a los factores ambientales como la ventilación, humedad y temperatura interior en las condiciones ideales para el correcto desarrollo de las plantas (Barrios, 2004).

## **2.2.2. Factores o variables a medir**

### ***2.2.2.1. Humedad del suelo***

Este factor se refiere a la cantidad de agua por volumen de tierra que existe en el terreno de un cultivo. El buen manejo de la humedad del suelo permite mejorar la producción de las plantas. Su medida es gravimétrica, y se da entre 0.1 y 0.3 bares de presión. Se relaciona con la capacidad de las raíces de las plantas para realizar la absorción de nutrientes del suelo (Ibáñez, 2006).

La humedad en los cultivos en invernadero varía de acuerdo con los requerimientos. Es un factor que ayuda al desarrollo de las plantas, pero en exceso produce daños como enfermedades causadas por hongos y bacterias. El agua y la humedad suelen elevarse dentro de los invernaderos cuando éstos no se ventilan, lo que provoca evotranspiración, es decir, la pérdida de humedad del suelo y de las plantas.

### ***2.2.2.2. Humedad relativa***

La humedad relativa es uno de los factores medioambientales que influyen en los cultivos bajo invernadero, comportándose de manera única para cada tipo de planta, y representa la cantidad de agua contenida en el aire. En el interior, el aire es enriquecido por vapor de agua desde el suelo y por transpiración. Las plantas tienen que transpirar agua para poder transportar nutrientes y regular su crecimiento, este factor depende de la transpiración y de la temperatura que posea el invernadero. El porcentaje de humedad relativa en el cual las plantas tienen un correcto desarrollo es del 55% al 70% (Iglesias N. , 2009).

### ***2.2.2.3.Temperatura***

En el desarrollo de los cultivos, la temperatura es uno de los factores más importantes, es por esto que la posibilidad de crear condiciones climáticas se ha convertido en una de las ventajas primordiales de los invernaderos. Controlando este factor, se puede prevenir daños en los cultivos debido a las heladas o a las altas temperaturas, y una forma de realizarlo es a través de una cubierta doble, asegurándose de que en las noches quede completamente cerrada.

Para sus procesos de crecimiento y correcto desarrollo, las plantas necesitan de una temperatura adecuada, de no ser así, estos procesos se detienen. Cuando este factor desciende a cero grados o menos, las plantas pueden sufrir daños severos en sus tejidos, así como suele suceder cuando se encuentran al en el aire libre durante las heladas nocturnas. En general, el efecto favorable que produce el invernadero sobre el desarrollo de las raíces y del cultivo es mantener la adecuada temperatura tanto del aire como del suelo (Barrios, 2004).

El calor en un invernadero se pierde por las roturas o aberturas que existan en las cubiertas, a través de éstas se escapa el aire tibio y en su lugar entra el aire frío. Debido a que el aire caliente es más liviano y por esto sube a la parte alta del invernadero, el frío se mantiene en la parte baja por ser más pesado, produciendo daño los cultivos (Barrios, 2004).

La temperatura a la cual los cultivos realizan un buen desarrollo y crecimiento dentro de un invernadero está entre los 15°C y 25°C (Iglesias N. , 2009).

### ***2.2.2.4.Luminosidad***

Esencialmente toda la luz visible es capaz de promover la fotosíntesis, pero las regiones de 400 a 500 y de 600 a 700 nm son las más eficaces. La cantidad de iluminación necesaria para que los cultivos dentro de un invernadero sobrevivan está entre los 10000 a 40000 lux (Iglesias N. , 2009).

### **2.2.2.5.CO<sub>2</sub>**

Este gas carbónico es de suma importancia en el ciclo de vida de los cultivos, es un material indispensable para la fotosíntesis y la clorofila de las plantas. Combinado con agua y energía luminosa, el CO<sub>2</sub> se emplea para que las plantas puedan producir carbohidratos y oxígeno, además, este factor está presente en la actividad estomática. La concentración de CO<sub>2</sub> en la atmósfera debe estar entre los 100 ppm (0,2 g/m<sup>3</sup>) y los 2000 ppm (4g/m<sup>3</sup>), cuando los valores de este componente se encuentren en este rango el cultivo se desarrollará en condiciones ideales, en caso de superar estos valores la planta cerrará sus estomas provocando efectos perjudiciales. Una concentración óptima de CO<sub>2</sub> tendrá un efecto positivo en desarrollo de la planta y vigor en general, y en tamaño de fruto en particular (Marlow, 2011).

## **2.3. BIG DATA**

Big Data es un término que se refiere a la información que excede la capacidad de procesamiento de los sistemas convencionales de bases de datos, o también que dicha cantidad sea tan grande, constantemente cambiante, o simplemente que no encaje en la estructura de las bases de datos. Para obtener valores y resultados, se debe buscar alternativas viables para procesar la información, considerando el volumen, la velocidad y la variabilidad de los datos masivos. Dentro de dicha de información, se encuentra patrones e información oculta valiosos para la toma de decisiones. Actualmente, el hardware básico, las arquitecturas de cloud y el software libre implica casi que directamente el uso de Big Data para obtener resultados que no provee el hardware (Dumbill, Slocum, Croll, & Hill, 2012).

**Volumen:** La cantidad masiva y el alto crecimiento de datos en un sistema es un nuevo escenario que obliga a utilizar nuevas infraestructuras, más escalables y distribuidas (Raj & Chandra Deka, 2014).

**Velocidad:** Los datos son generados y adquiridos a velocidades cada vez más altas y además la transformación de éstos en información útil se debe realizar en un menor tiempo, requiriendo así un mayor nivel de procesamiento (Raj & Chandra Deka, 2014).

**Variedad:** Los datos son recolectados de varias fuentes, pero todo esto debe ser analizado en conjunto para generar resultados valorables. No se trabaja con el modelo tradicional soportado por datos estructurados (Bases de Datos relacionales), sino que se adentra a un nuevo escenario con una mayor cantidad de datos en formatos diferentes (bases de datos, HTML, XML, texto plano, imágenes, video, audio, código fuente, etc.).

Según (Tascón, 2013): “Tradicionalmente, los principales conceptos agrupados que han definido a Big Data han sido las denominadas ‘3 V’: Volumen, variabilidad y velocidad”. Esto se refiere a los parámetros que engloban a este concepto, es decir, a los grandes volúmenes de información que se mueven o analizan a altas velocidades, por ende, varían en su forma, estructura o composición; de la mano con estos conceptos debe tomarse en cuenta una cuarta ‘V’, denominada visualización, ya que el usuario puede interactuar más con los sistemas de análisis cuando los datos son mostrados de manera inteligible para interpretación del ser humano.

Algunos autores se refieren a la ‘4 V’ como el valor, siendo esta la información más relevante de todos los macrodatos que han sido adquiridos. A continuación, en la Tabla 1 se muestra una relación de los parámetros y descripciones de las características que definen a Big Data.

Tabla 1 *Parámetros de Big Data*

<b>PARÁMETRO</b>	<b>DESCRIPCIÓN</b>
<b>Volumen</b>	Grandes cantidades de datos adquiridos, y la infraestructura necesaria para almacenarlos.
<b>Velocidad</b>	Los datos se generan a velocidades impresionantes, y así exigiendo a los sistemas mayor procesamiento.
<b>Variabilidad</b>	Los formatos de los datos son diferentes, pues provienen de ambientes distintos, pero se deben analizar como un solo conjunto.
<b>Visualización</b>	Al ser una cantidad inmensa de datos y variables, se debe buscar técnicas que permitan su visualización y el entendimiento con la percepción humana.
<b>Valor</b>	Este factor mide la utilidad de los datos para la toma de decisiones.

Fuente: Propia.

Al referirse a Big Data, no existe un estándar que lo defina, de hecho, esta temática interseca con diversos temas relacionados con las técnicas y algoritmos de la inteligencia del negocio. Si es necesario el uso de la expresión estándar, se puede referir a las nuevas tecnologías que se han ido implementando para la recolección de información. Se puede mencionar a los estándares IEEE 802.11 utilizado para redes inalámbricas de área local o WLAN, e IEEE 802.15 usado en redes de área personal, en decir, estos pueden ser aplicados en las redes de sensores que recolecten los datos para el sistema.



### 2.3.1. Herramientas para Big Data y minería de datos

Actualmente existen varios *framework* que permiten realizar un análisis de datos con Big Data, o herramientas que facilitan el uso de las técnicas de minería de datos. Entre los frameworks más usados se encuentran los desarrollados por Apache: Hadoop y Spark, los cuales permiten el procesamiento de los datos a gran escala. Además, existen herramientas que permiten usar técnicas o algoritmos de Data Mining, tales como: Weka, Orange Canvas, Rapid Miner Studio; las cuales permiten al usuario seleccionar entre la variedad de técnicas para su respectivo análisis.

Para desarrollar herramientas personalizadas, es decir, enfocadas a una temática en especial, o la implementación de un software propio, se puede hacer uso de los entornos de desarrollo de programación, con esto se realiza aplicaciones más robustas, enfocadas al área de aplicación. A continuación, se describe los frameworks, herramientas y lenguajes de programación más usados en analítica de datos.

#### 2.3.1.1. Apache Hadoop

Apache Hadoop es un framework que permite el procesamiento distribuido de grandes conjuntos de datos usando simples modelos de programación. Está diseñado para pasar de simples servidores a miles de máquinas computacionales con almacenamiento propio. En lugar de depender de hardware para ofrecer alta disponibilidad, se ha diseñado para manejar fallas en la capa de aplicación (Hadoop, A, 2014).

HDFS, de sus siglas en inglés *Hadoop distributed file system*, es un sistema de archivos distribuido, y está compuesto de un único *NameNode* que es un servidor maestro. Administra el espacio de nombres del sistema de archivos y está vinculado a un *Datanode* para administrar el almacenamiento de datos. Esta estructura implica que todos los bloques de un archivo se pueden

guardar en varias máquinas. La información de grandes cantidades de datos (metadatos) consiste en particiones de archivos en bloques y la distribución de estos bloques en diferentes *Datanodes*. En la Figura 2, se muestra la interfaz web de Apache Hadoop (Hadoop, A, 2014).

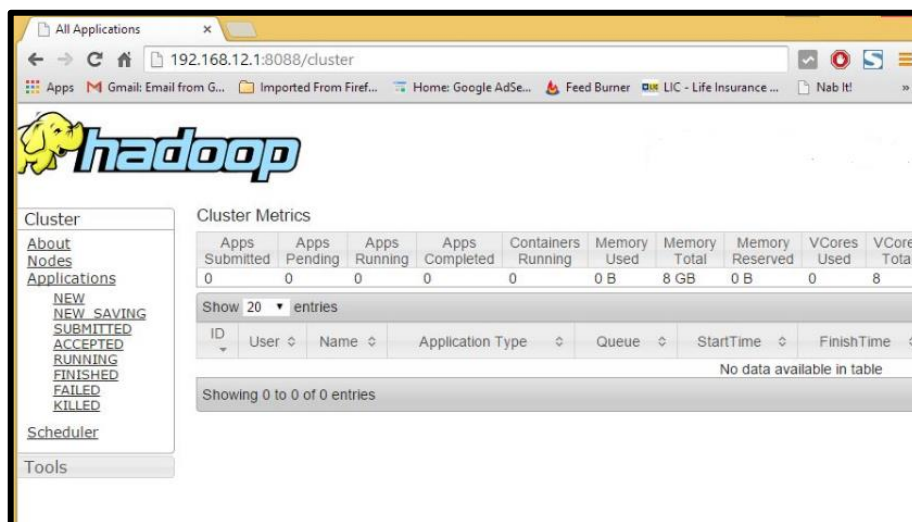


Figura 2 Logo de Apache Hadoop  
Fuente: (Hadoop, A, 2014)

### 2.3.1.2. Apache Spark

*Spark* es un *framework* considerado como un motor de procesamiento (denominado DAG) que se ha construido con base en la velocidad, facilidad de uso y análisis sofisticados. Permite la gestión de datos de diferente naturaleza como son los textos o los gráficos, además puede procesar a gran escala usando petabytes de datos en múltiples clusters con un mayor número de nodos. El acceso a la interfaz web del framework Apache Spark se muestra en la Figura 3. (Apache Spark, 2016).

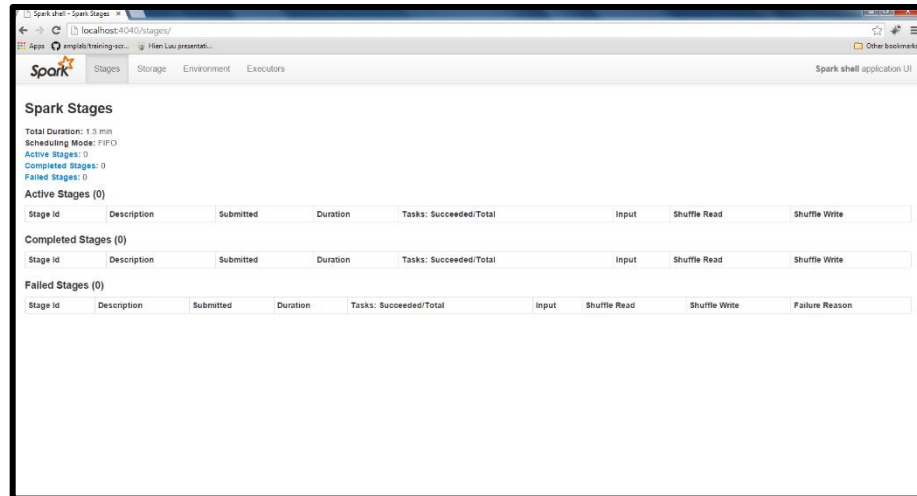


Figura 3 Uso de la interfaz web de Spark.

Fuente: (InfoQ, 2014)

### 2.3.1.3. Orange

Orange es una herramienta de Data Mining y Machine Learning. Su arquitectura multicapa soporta diferentes tipos de usuarios, desde principiantes hasta programadores. Este software se trata de visualización de datos que ayuda a descubrir patrones ocultos en estos, proporciona intuitivamente elementos del procedimiento de la analítica de datos y/o soporta la comunicación entre científicos de datos y expertos. Posee herramientas de visualización que incluyen gráficos de dispersión, cuadros e histogramas, y modelado de visualizaciones específicas como el dendograma. La Figura 4 muestra el framework de Orange con un ejemplo de cluster. ( University of Ljubljana, 2016).

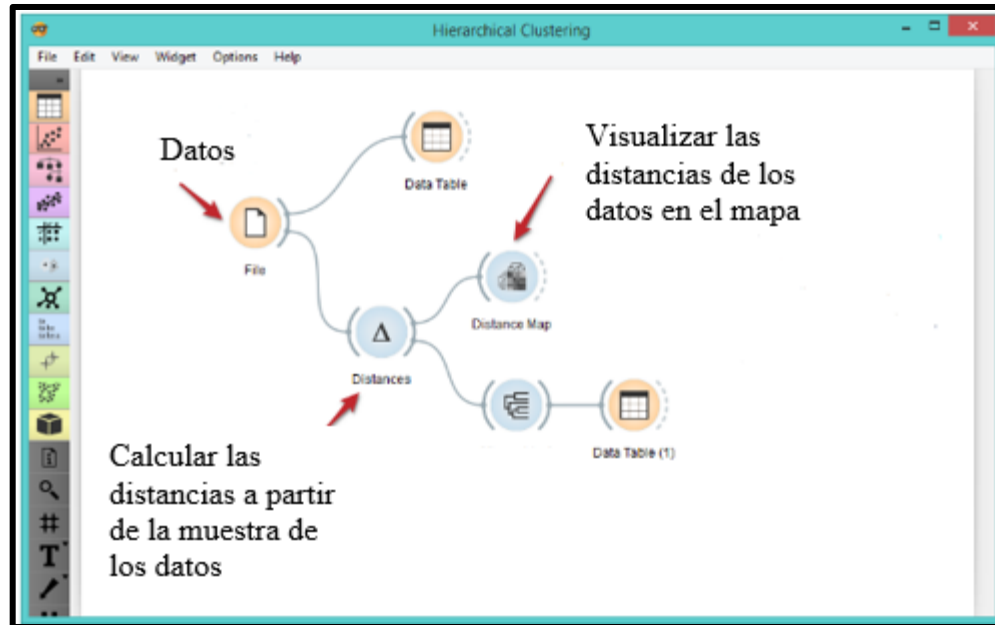


Figura 4 Ejemplo de Cluster jerárquico realizado en Orange  
Fuente: ( University of Ljubljana, 2016)

#### 2.3.1.4. Weka: Data Mining Software in Java

Weka es un software libre con una colección de algoritmos de Machine Learning para aplicaciones de minería de datos, Los algoritmos pueden ser directamente aplicados a un conjunto de datos (*data set*) o pueden ser llamados desde el propio código Java. Contiene herramientas para un procesamiento previo, clasificación, regresión, asociación y visualización. Es comúnmente usado para el desarrollo de nuevos esquemas de Machine Learning. Un esquema de clasificación realizado en weka se muestra en la Figura 5, en la cual se puede visualizar un árbol de decisión formado a partir de un set de entrenamiento. (Machine Learning Group at the University of Waikato, 2014).

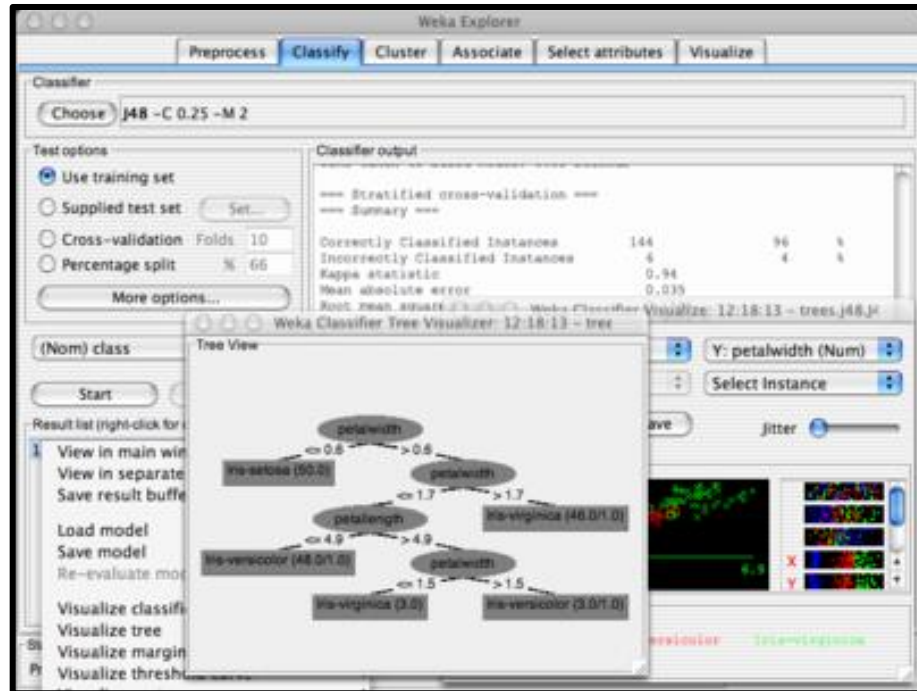


Figura 5 Uso del panel de clasificador en Weka  
Fuente: (The University of waikato, 2014)

### 2.3.1.5. Python

Python es una herramienta poderosa, flexible, de código abierto que posee numerosas librerías para la manipulación de datos y su análisis. Ha sido desarrollada la siguiente generación de herramientas, que permiten a Python ser uno de los softwares más sofisticados y poderosos para Big Data. Se han enfocado en ofrecer al usuario librerías para la estructuración, manipulación, análisis y visualización de los datos (Continun Analytics, 2016).

Python es un lenguaje de programación que con el uso de diversas herramientas permite el desarrollo de varias aplicaciones en el área de la minería de datos. Anaconda es una de las herramientas que utiliza Python, en la figura 6 se muestra la interfaz de anaconda para el desarrollo de aplicativos.

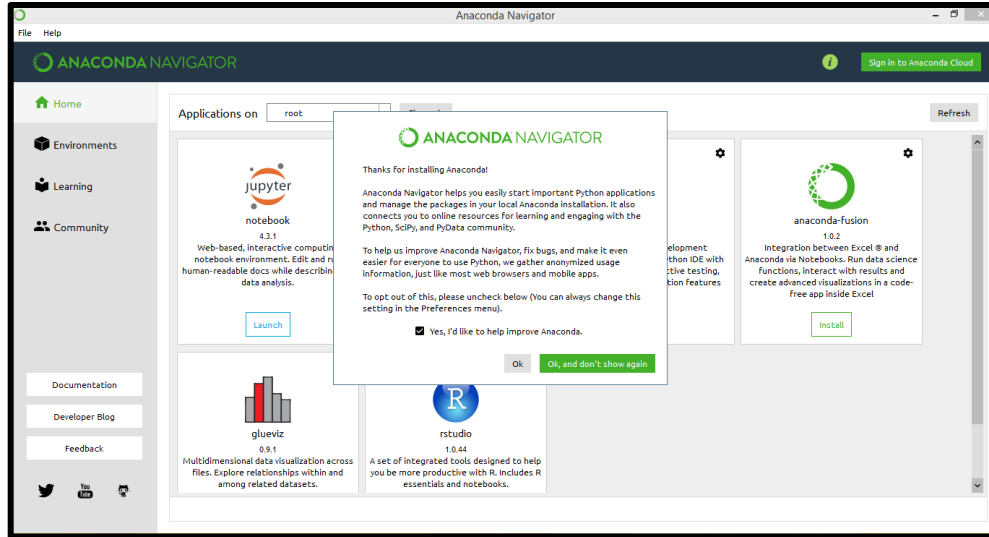


Figura 6 Navegador Anaconda, esta permite la selección de las herramientas para el desarrollo de software.

Fuente: (Python, 2015)

### 2.3.1.6. MATLAB

La plataforma MATLAB está optimizada para resolver problemas de ingeniería, el lenguaje basado en matrices es la manera más natural de expresar matemáticas computacionales. Los gráficos integrados facilitan la visualización y la obtención de información a partir de los datos. La Figura 7 muestra un aplicativo de clasificación de datos realizado en con las herramientas de MatLab para la minería de datos. Ingenieros y analistas trabajan con grandes cantidades de datos en varios formatos, usan técnicas de minería para encontrar patrones y construir modelos predictivos. MATLAB tiene acceso a funciones previamente diseñadas, varias herramientas y aplicaciones especializadas para clasificación, regresión y clustering (MathWorks, 2016).

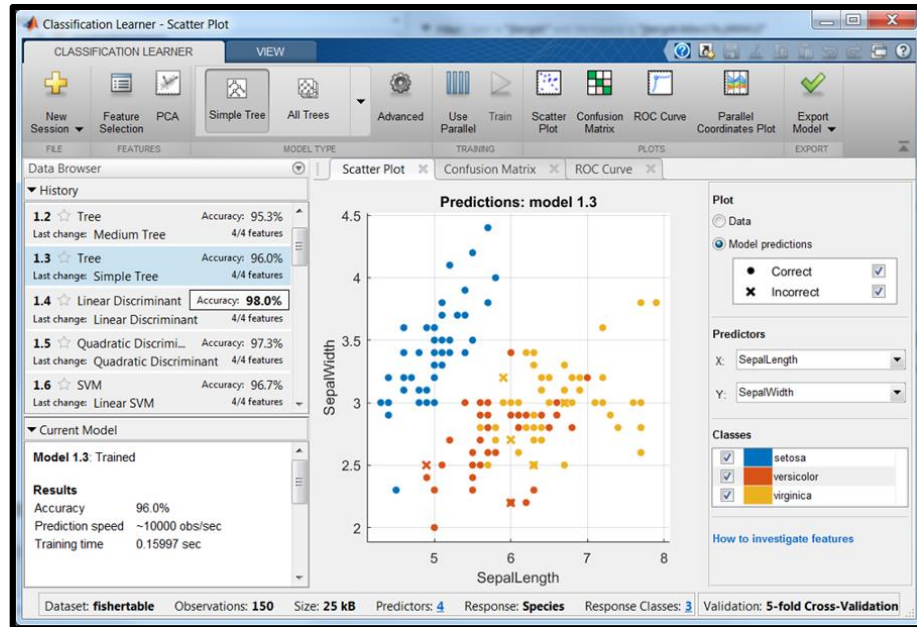


Figura 7 Aplicación de clasificación de datos realizada en MatLab  
Fuente: (MathWorks, 2016)

### 2.3.1.7. Java

Es un lenguaje de programación orientado a objetos, que es multiplataforma, pues el código es interpretado por una máquina virtual, propia del sistema (máquina virtual de java). Al ser robusto es utilizado para las diversas técnicas de minería de datos. Se puede acceder a diferentes librerías para clasificación, asociación y agrupamiento no supervisado (*clustering*), en la Figura 8 se aprecia una interfaz de clustering desarrollada en java. (Java, 2016).

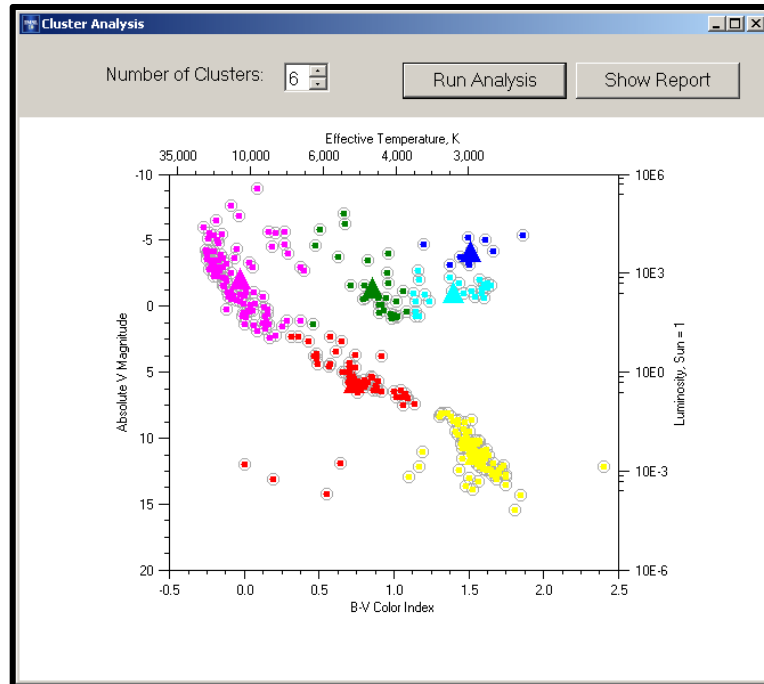


Figura 8 Interfaz de clustering desarrollada en java.  
Fuente: (RogueWave , 2017)

## 2.4. ADQUISICIÓN DE DATOS Y BASES DE DATOS

Como se ha mencionado anteriormente los datos pueden ser adquiridos de diferentes maneras, siendo una de ellas el uso de IoT (internet de las cosas), es decir, redes de sensores inalámbricos ubicados en las áreas de interés. Éstos capturan diferente tipo de información de todos los ambientes posibles. Con el nuevo concepto “smart cities” la recolección de datos se ha vuelto masiva, ya que existe una infinidad de sensores que pueden ser usados en muchas áreas, como el transporte, vigilancia, tráfico, turismo, factores ambientales, agricultura, etc.

Cloud Computing como un nuevo modelo de administración de la demanda de recursos computacionales, provee una plataforma flexible, la cual permite la recolección, análisis, procesamiento y visualización de datos masivos. El almacenamiento de Big Data es llevado a cabo por sistemas de archivos que son sumamente diferentes a los tradicionales y establecen la manera



en que el procesamiento de la información va a ser realizado, así como métodos estandarizados que facilitan el manejo de Big Data además de proveer una herramienta para los operadores en el cloud y así hacer a la plataforma más accesible.

Los avances tecnológicos a nivel de comunicaciones, computación y almacenamiento, han creado una gran cantidad de datos, adquiriendo información de valor para los negocios, la ciencia, el gobierno y la sociedad. Un gran ejemplo son las corporaciones desarrolladoras de software y hardware, tales como Google, Yahoo!, y Microsoft, las cuales han innovado completamente el negocio con la adquisición de información a través de la web (www o World Wide Web) y mostrándola a los usuarios en aplicaciones útiles. Estas compañías recolectan trillones de bytes de datos todos los días y añaden nuevos servicios continuamente, así como, imágenes satelitales, instrucciones de viaje, recuperación de imágenes, etc. Los beneficios sociales de estos servicios son incalculables ya que han transformado la manera en que las personas encuentran y hacen buen uso de la información diariamente (Bryant, Katz, & Lazowska, 2008).

#### **2.4.1. Redes de sensores inalámbricos (WSN)**

Una red de sensores inalámbricos está compuesta de sistemas embebidos individuales que son capaces de interactuar con su entorno a través de varios sensores, además de procesar y comunicar la información inalámbricamente con los nodos vecinos. Un diagrama de conexión de una WSN se muestra en la Figura 9 (Akyildiz & Vuran, 2010).

Los avances recientes en las comunicaciones inalámbricas, la electrónica digital y los dispositivos analógicos han desarrollado nodos de sensores que son de bajo costo y bajo consumo de energía para comunicarse sin inconvenientes a cortas distancias y que todos trabajen como un

solo grupo. Estos nodos aprovechan la fuerza de la colaboración para proveer una alta calidad de sensado en el tiempo y el espacio (LEWIS, 2006).

Existen varios tipos de aplicaciones de esta tecnología. Por ejemplo, en defensa, la detección de ataque nuclear, biológico y químico. En medio ambiente, el monitoreo de microclimas, detección de fuego, detección de inundaciones, agricultura. A través de estos nodos sensores se facilita la recolección de datos reales, usados hoy en día en busca de conectar todo lo posible, apegándose al concepto de Smart Cities (Libelium, 2010).

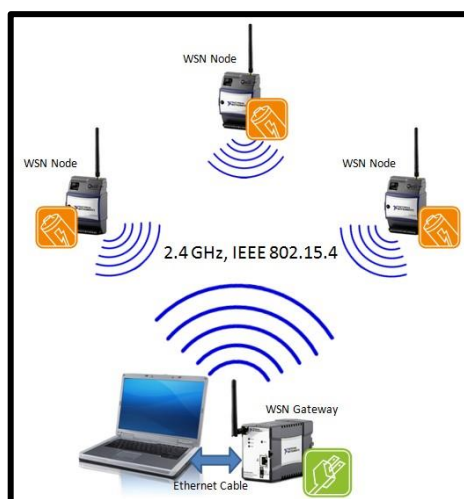


Figura 9 Red de sensores inalámbricos WSN con el protocolo IEEE 802.15.4.

Fuente: (Energy Monitoring, 2015)

#### 2.4.2. Bodega de datos

El término bodega de datos proviene del inglés *Data Warehouse*, y se refiere a los repositorios de bases de datos no relacionales existentes en la web o en los servidores de las entidades dedicadas a la adquisición de estos. Estas bodegas de datos ayudan al proceso KDD (Knowledge Discovery Data Bases) de dos maneras, en la limpieza y el acceso a los datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

### ***2.4.2.1. Open Data***

Este término es muchas veces confundido con Big Data, pero no se trata de lo mismo, ya que, Open Data es todo aquel conjunto de datos que se han distribuido libremente a través de la red de internet, es decir, la apertura de datos digitales. El objetivo de este concepto es tener información disponible para todo el mundo de manera libre, sin restricciones ni derechos de autor, y que pueden ser encontrados en la web, a través de repositorios (Ferrer-Sapena & Sánchez-Pérez, 2013).

### **2.4.3. Base de datos**

El almacenamiento masivo de los datos hace que las características de los estos varíen, tornándolos imperfectos, es decir, se encuentran sin una estructura previa. Las herramientas computacionales al no cumplir con las reglas Codd (propuestas por Edgar F. Codd, del modelo relacional para las bases de datos) se las conoce como No SQL o bases de datos no relacionales, siendo las más utilizadas para Big Data. Base de datos NoSQL se refiere a las diferentes tecnologías que proporcionan un enfoque alternativo para el almacenamiento de datos, ya que poseen una organización muy simple o sin estructura (Ferrer-Sapena & Sánchez-Pérez, 2013).

## **2.5. PROCESO KDD**

Este término se originó con investigaciones en el campo de la inteligencia artificial, este proceso involucra algunas etapas en el análisis de datos: Selección, procesamiento, transformación (en caso de ser necesaria), la realización de minería de datos (Data Mining) para extraer patrones y relaciones, y por último interpretación y evaluación de las estructuras descubiertas. En la figura 10 se muestra una ilustración del proceso KDD y sus etapas. (Hand, Mannila, & Smyth, 2001).

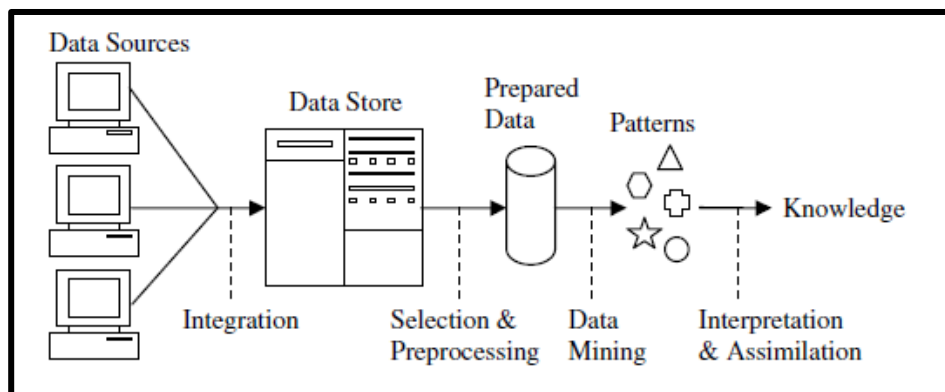


Figura 10 Proceso de KDD  
Fuente: (Bramer, 2007)

KDD ha sido definido como un proceso de extracción de información implícita, desconocida previamente, potencial y útil de los datos, este es el proceso central de Data Mining. Como se indica en la figura 10, los datos vienen probablemente de varias fuentes, se integran y son ubicados en alguna tienda común, parte de estos son pre procesados hacia un formato estándar. Los datos preparados son pasados a un algoritmo de data mining, el cual produce una salida en forma de reglas o algún otro tipo de patrón, esto es interpretado como un conocimiento útil y potencial (Bramer, 2007).

Según (Brachman & Anand, 1996), el proceso KDD envuelve numerosos pasos con varias decisiones hechas por el usuario, a continuación, se describe brevemente cada uno de estos:

- Desarrollar un entendimiento del dominio de la aplicación y del conocimiento previo, además de identificar la meta del proceso KDD desde el punto de vista del cliente.
- Crear un conjunto de datos, seleccionando o enfocándose en una subred de variables o muestras de datos en las que se va a realizar el descubrimiento de patrones.

- Limpieza de los datos y pre procesamiento, estas operaciones básicas incluyen remover el ruido, recolectar la información necesaria para modelarlo, decidir estrategias para manejar los datos perdidos y conocer los cambios que se han realizado.
- Reducción y proyección de los datos, encontrar características útiles para la meta o tarea. Con la reducción de dimensión o métodos de transformación, el número efectivo de variables a tomar en consideración puede ser reducido.
- Hacer encajar las metas del proceso KDD para un particular método de minería de datos, como: sumarización, clasificación, regresión, clustering, entre otros.
- Análisis exploratorio, modelamiento y selección de hipótesis; selección de métodos a ser usados en la búsqueda de patrones similares. En este proceso se decide el modelo y los parámetros apropiados, que encaje con un método particular de minería de datos.
- Minería de datos, búsqueda de patrones de interés en una forma particular representativa, incluye reglas de clasificación o árboles, regresión y clustering.
- Interpretar los patrones, posiblemente se regrese a los pasos anteriores para una iteración futura. Este paso puede también involucrar la visualización de los patrones extraídos.
- Actuar sobre el conocimiento descubierto: usando el conocimiento directamente, incorporándolo a otro sistema o futura iteración o simplemente documentarlo y entregarlo a las partes interesadas (Brachman & Anand, 1996).

### **2.5.1. Bodega de datos**

El término Bodega de datos se refiere al conjunto de arquitecturas, tecnologías que facilitan o apoyan procesos de KDD, estos conceptos surgieron en los años 70 y 80, debido a la necesidad de almacenar y analizar los datos que no poseen una estructura previa. En los procesos de Data

Warehousing normal, una de las técnicas de análisis de información más utilizadas es el procesamiento por lotes (Batch), donde los datos generados en ventanas de tiempo son divididos en segmentos de tamaño fijo y procesados a través de diferentes capas, cuyo resultado es almacenado en bodegas de datos para su uso futuro en análisis y/o visualización (Rahm & Hai Do, 2000).

Una bodega de datos requiere y provee un extensor soporte para la limpieza de los datos. Ésta se carga y refresca continuamente enormes cantidades de datos de diversas fuentes, que probablemente contienen datos contaminados (*dirty data*). Además, son usadas para la toma de decisiones, por lo tanto, la corrección de los datos erróneos es de vital importancia para evitar conclusiones equivocadas (Rahm & Hai Do, 2000).

### **2.5.2. Limpieza de datos (*Data Cleaning*)**

A este concepto también se lo conoce como *Data Cleansing* o *scrubbing*, y se refiere a la detección y eliminación de errores e inconsistencias de los datos, mejorando así la calidad de la información. Las múltiples fuentes necesitan ser integradas en un solo sistema de almacenamiento (denominado warehouses), que carga y continuamente actualiza grandes cantidades de información y por esto la probabilidad de datos repetidos o errados es alta (IEEE Computer Society, 2010).

Data Cleaning trata los problemas de los datos una vez estos ya hayan ocurrido. Las estrategias de prevención de errores suelen reducir algunos de los problemas, pero no eliminarlos por completo; pues estos pueden ocurrir en cualquier etapa del flujo de datos, es decir, mientras los extraen, transfirieren, editan, seleccionan, transforman y presentan (Van den Broek, Argeseanu, Eeckels, & Herbss, 2005).

No todo se trata de la infraestructura, el uso de técnicas de Big Data y Data Mining es un 80% la limpieza de los datos (Data Cleaning), es decir, el esfuerzo y el tiempo que se invierten en convertir la gran cantidad de información en algo valioso (Dumbill, Slocum, Croll, & Hill, 2012).

### **2.5.3. Minería de datos**

Minería de datos o *data mining* viene a ser el proceso de la extracción de información implícita y potencialmente útil de los datos. Como lo definen (Han, Kamber, & Pei, 2012) “Data mining es el proceso de descubrir patrones interesantes y conocimiento de una gran cantidad de datos”, donde la idea es construir programas computacionales que examinen cuidadosamente las bases de datos, en busca de aspectos similares o patrones para hacer predicciones exactas en los datos futuros. Muchos de los valores pueden ser banales y sin interés, mientras que otros falsos o inconsistentes, algunas partes serán ilegibles y otras se perderán en el proceso (Witten & Eibe, 2013).

La minería de datos es el proceso de análisis de conjuntos de datos que genera resultados que permiten encontrar relaciones y resumir los datos en formas novedosas que sean entendibles y útiles para el usuario. Dichos resultados son a menudo denominados modelos o patrones subyacentes de los datos en bruto. El proceso en sí de minería incluye ecuaciones, reglas, conformación de *clusters* (grupos o conglomerados), gráficos, estructuras de árbol y patrones similares en series de tiempos (Hand, Mannila, & Smyth, 2001).

Como una ciencia interdisciplinaria hace uso de: estadísticas, bases de datos, machine learning, reconocimientos de patrones, inteligencia artificial, y visualización. Una explicación gráfica de las características y técnicas que abarca la minería de datos, se aprecia en la Figura 11.

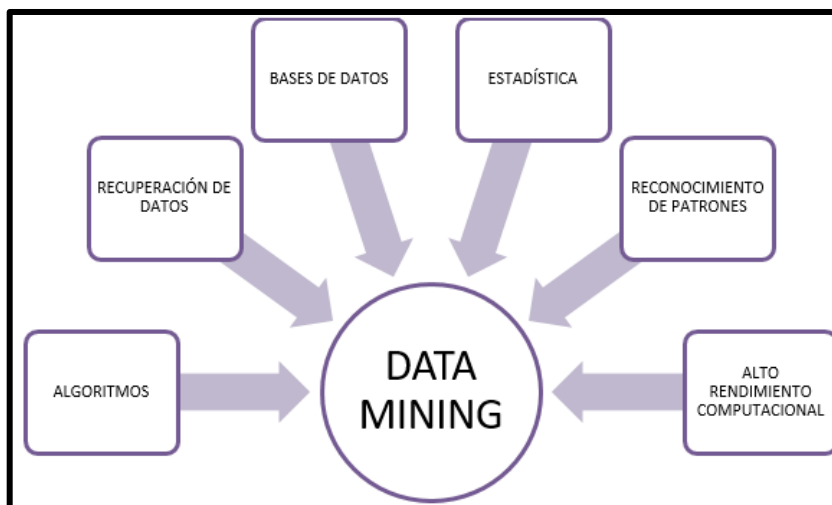


Figura 11 Características de *Data Mining*  
Fuente: Propia

### 2.5.3.1. Tareas de la minería de datos

**Análisis de datos exploratorio (EDA. Exploratory Data Analysis).** - La meta en esta etapa es simple, explorar los datos sin tener una idea clara de los que se está buscando.

**Modelo Descriptivo.** - La meta de este modelo es describir todo acerca de los datos (o del proceso para generarlos). Entre las diferentes descripciones están: la probabilidad general de la distribución de los datos (estimación de densidad), particionamiento del espacio (clustering y segmentación), y modelos de descriptivos de la relación entre variables (modelo de dependencia).

**Modelo Predictivo.** - Clasificación y Regresión, el objetivo es construir un modelo que permita predecir el valor de una variable a partir de los valores conocidos de otras variables. En la clasificación la variable a predecir es categórica, mientras que en la regresión es cuantitativa.

**Descubrimiento de patrones y reglas.** - Las aplicaciones con minería de datos contienen detección de patrones, otra tarea es encontrar combinaciones de ítems que ocurren frecuentemente en la transición de las bases de datos, esto se ha realizado con el uso de algoritmos y técnicas basados en las técnicas y reglas de asociación.



**Recuperación del contenido.** - En este punto el usuario tiene un patrón de interés y desea encontrar similares en el conjunto de datos (Hand, Mannila, & Smyth, 2001).

#### 2.5.4. Clasificación

La clasificación es una función de la minería de datos que asigna elementos de un conjunto de datos a categorías o clases. El objetivo de la clasificación es predecir con exactitud la clase objetivo para cada caso. Este proceso inicia con un conjunto de datos en el cual la clase es previamente conocida, por ende, es el aprendizaje de una función que mapea un elemento de los datos en una de las clases pre definidas (Oracle, 2016). La Figura 12 explica gráficamente el proceso de clasificación de un conjunto de datos en dos clases.

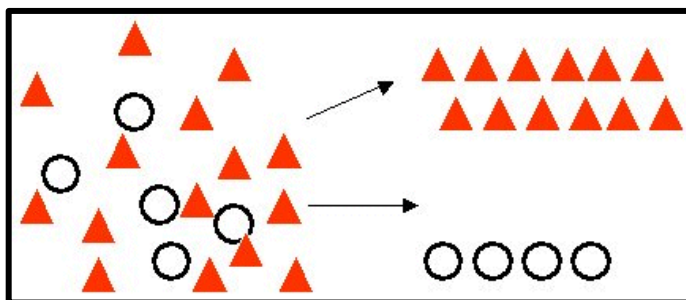


Figura 12 Diagrama explicativo de clasificación de datos de dos clases o especies  
Fuente: (Oracle, 2016)

##### 2.5.4.1. Clasificación supervisada

El proceso y las metas de las diferentes disciplinas de la minería de datos son básicamente lo mismo, pero los patrones que se buscan son diferentes. La clasificación supervisada es una subdisciplina que empieza con un conjunto de datos conformado por instancias que tienen atributos o características diferentes, por lo menos una de ellas es nominal, a lo que se llama clase. La meta es encontrar patrones que permitan predecir la clase de nuevas instancias y así poder clasificarlas.

- ***Máquinas de Vectores de Soporte (SVM)***

Esta técnica de clasificación supervisada ha sido desarrollada en orden inversa que las redes neuronales, las máquinas de vectores de soporte envuelven desde el sonido teórico a la implementación y experimentación, mientras que las redes neuronales siguen una ruta heurística, es decir, desde las aplicaciones y extensos experimentos hacia lo teórico. El desarrollo teórico detrás de esta técnica SVM no es ampliamente apreciado al inicio (Bramer, 2007).

- ***K – Nearest – Neighbor (K-*nn*)***

Este algoritmo basado en instancias es el más común en el aprendizaje para la clasificación de los datos. A pesar de sus deficiencias kNN aplica una variedad de clasificaciones de datos real. Estos datos consisten en algunas características como atributos continuos y nominales; para hacer frente a estos las funciones de distancia de kNN juegan un papel importante en la clasificación (Ishii, Hoki, Okada, & Bao, 2009).

Este método ha sido usado ampliamente en problemas de la clasificación, kNN está basado en una función de distancia que mide la diferencia o similitud entre dos instancias. La distancia Euclidiana estándar  $d(x_2, x_1)$ , como lo muestra la figura 13, es a menudo usada como la función de distancia kNN. Dada una instancia  $x$ , kNN asigna la etiqueta de clase más común de los  $x'$ s vecinos K más cercanos a  $x$  (Jiang, Zhang, & Cai, 2006).

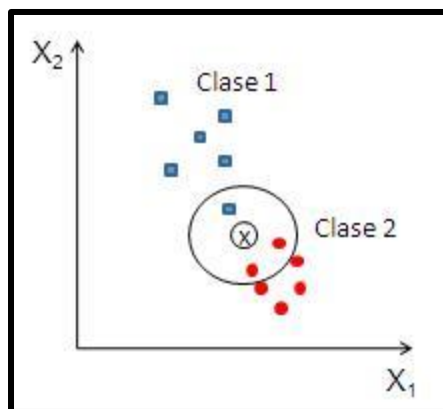


Figura 13 Algoritmo kNN con dos clases.  
Fuente: (EcuRed, 2011)

KNN es el ejemplo típico de un aprendizaje perezoso, el cual simplemente guarda sus datos entrenados y retarda su aprendizaje hasta la hora de clasificación. Aunque kNN ha sido usado como método de clasificación por décadas, existen problemas respecto a su exactitud. Motivados por estos, los investigadores han hecho un esfuerzo para mejorar la exactitud de KNN. Para mejorar el proceso se enfocó en el voto del vecino más cercano  $k$ , de acuerdo a su distancia a la instancia de prueba  $x$ , proporcionando una mayor ponderación a los vecinos más cercanos. Como resultado se tiene al clasificador  $K$  – Nearest – Neighbor con Distancia Ponderada ( $K$  – Nearest – Neighbor – with Distance Weighted, KNNDW) (Jiang, Zhang, & Cai, 2006).

#### - **Redes Neuronales**

Una Red Neuronal (NN, Neural Network) es usada para denotar modelos matemáticos de las funciones del cerebro, de ahí su nombre. Estas pretenden expresar las propiedades del procesamiento paralelo masivo y de la representación distribuida existente en el cerebro, a partir de la experiencia, las redes neuronales artificiales generan su conocimiento. Ha sido propuesto para los dos tipos de aprendizaje, tanto supervisado como no supervisado. Este método puede adaptarse a valores bastante indefinidos e incluso ausentes, pero son difíciles en el momento de

inspeccionar. Con el uso de una buena herramienta de visualización permite al usuario reconstruir el "razonamiento" de la red neuronal (Weber, 2000).

Al igual que el sistema nervioso humano el conocimiento se encuentra en los pesos de la conexión de las neuronas. Estos pesos varían de acuerdo al algoritmo de aprendizaje usado hasta tomar un valor constante y es ahí cuando la red ya ha aprendido. Dependiendo del tipo de clasificación usada el peso de las conexiones neuronales toma un valor, al ser supervisada ya existe información de ejemplo para el aprendizaje, mientras que en una clasificación no supervisada los pesos de la red fluctúan libremente hasta que se estabilizan, la Figura 14 muestra gráficamente este proceso. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida (Anegón, Herrero Solana, & Guerrero Bote, 1998).

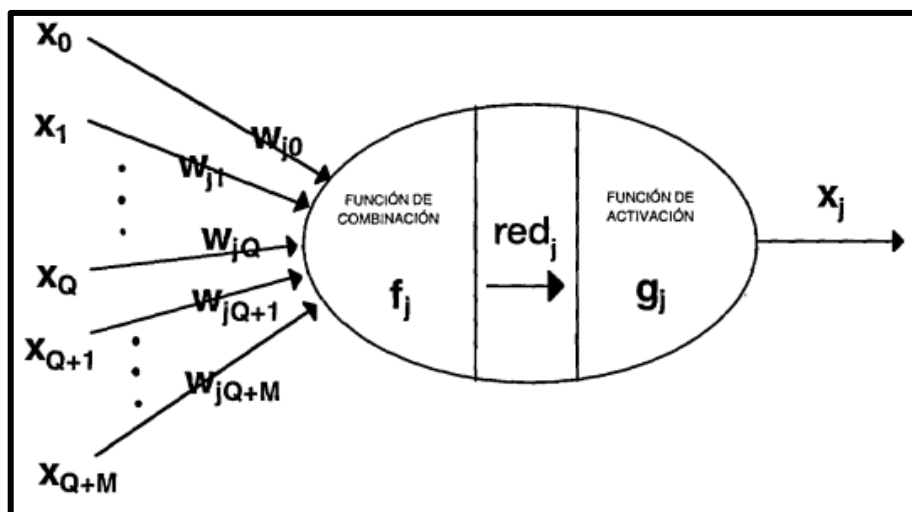


Figura 14 Diagrama explicativo de la j-ésima red de un sistema neuronal de Q+M entradas  
Fuente: (Anegón, Herrero Solana, & Guerrero Bote, 1998)

### - Árboles de Decisión

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy

similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema (Corso, 2009).

Es un modelo predictivo usado para identificar la estrategia que sea más probable para alcanzar la meta del análisis. El árbol de decisión es útil como una técnica de exploración. Cada nodo interno del árbol se divide en dos o más sub espacios de acuerdo a una función discreta de los atributos de entrada. Cada hoja es asignada a una clase representativa, la más cercana al objetivo, alternativamente la hoja guarda un vector, el cual indica la probabilidad de que los atributos tiendan a cierto valor (Rokach & Maimon, 2014).

#### - C 4.5

Es un algoritmo que permite trabajar con valores sensibles numéricos, es decir, con valores perdidos, podando para controlar el ruido. Clasifica los datos en forma de un árbol de decisión, a partir de un conjunto de datos que ya han sido clasificados previamente. Este es de aprendizaje es supervisado, desde el entrenamiento el conjunto de datos es etiquetado con clases.

Basado en la entropía de la información, la cual se refiere a un aprendizaje a través del desconocimiento. Se clasifican las instancias desde la raíz hacia las hojas, en cada nodo se selecciona una variable y un umbral (Rokach & Maimon, 2014).

#### ***2.5.4.2. Clasificación no supervisada***

En este tipo de clasificación no se conoce de antemano las clases de datos que se están entrenando, entre los diversos métodos se encuentran: clustering, K-means, e incluso pueden ser usadas las redes neuronales.

### - *Clustering*

Es un procedimiento de agrupación de una serie de ítems según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes (Corso, 2009).

Esta es una técnica de clasificación no supervisada, juega un papel muy importante en aplicaciones de minería de datos. Utiliza algoritmos matemáticos que se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases (Corso, 2009).

### - *K-Means*

El método k-means tiene como objetivo minimizar la suma de las distancias cuadradas entre todos los puntos y el centro del cluster. Es comúnmente usado para dividir automáticamente un conjunto de datos en  $k$  grupos. El algoritmo converge cuando no existe un cambio repentino en la asignación de instancias en el cluster, se inicia con valores aleatorios del conjunto de datos (Zha, He., Ding., Gu., & Simon, 2001).

Es uno de los métodos de clasificación no supervisada que resuelve el problema de agrupación. El proceso sigue un camino sencillo dado un conjunto de datos, a través de un cierto número de clusters, la idea es definir  $k$  centroides, uno por cada cluster. Estos deben ser ubicados de manera intuitiva. porque diferentes locaciones provocan diferentes resultados, entonces la mejor opción es escoger el lugar para ubicarlos lo más lejano posible de cada uno. El siguiente paso es tomar cada punto perteneciente a un conjunto de datos dado y asociarlo al centroide más cercano. Al terminar la agrupación se debe recalcular el número de centroides  $k$  y esta vez como baricentros de los

clusters resultado del paso anterior. Después, con estos nuevos centroides se debe realizar un nuevo vínculo entre los mismos puntos del conjunto de datos y el nuevo centroide, generando un lazo (loop). Como resultado de este lazo los centroides  $k$  han cambiado su ubicación paso a paso hasta que no ocurran más cambios (MacQueen, 1967).

## **CAPÍTULO III**

### **DISEÑO**

#### **3.1. INTRODUCCIÓN**

En este capítulo se mostrará una visión general del análisis de datos basado en técnicas de Big Data y Data Mining para los cultivos de hortalizas en el invernadero de la granja “La pradera”, se detallan las características y el modo de operación de los métodos a usar en el desarrollo de la interfaz.

#### **3.2. DESCRIPCIÓN GENERAL**

El presente proyecto tiene como objetivo realizar una interfaz de análisis de datos, esto involucra diferentes procesos, uno de ellos es la clasificación de los datos a través de una variable objetivo y el uso de un algoritmo de clasificación, se pretende proponer un modelo predictivo con los datos recolectados.

Los datos han sido obtenidos por medio de repositorios en línea que son confiables y poseen las variables necesarias, tales como: humedad del suelo, humedad relativa, temperatura ambiental, nivel de iluminación y Co<sub>2</sub>, las cuales interfieren en el buen desarrollo de los cultivos. El desarrollo de la solución posee varias etapas, entre estas: Selección de datos (búsqueda del archivo), pre procesamiento (selección de la variable objetivo), algoritmo de clasificación, y la predicción.

Para un mejor desarrollo de la propuesta, este capítulo está dividido en dos partes, la primera de estas enfocada a la analítica de datos con el proceso KDD (*Knowledge Discovery Databases*) con todas las fases que se debe seguir, y la segunda se refiere al desarrollo con la metodología CRIPS-DM, la cual incorporará todo el proceso en una sola solución, y así el usuario pueda observar el resultado final.



### 3.3. PROCESO KDD

El descubrimiento de conocimiento en bases de datos (Knowledge Discovery Databases) o comúnmente llamado KDD, ha sido definido por (Bramer, 2007) como un “proceso de extracción de información implícita, desconocida previamente, potencial y útil de los datos”. Para realizar el análisis de datos del presente proyecto se va a utilizar la estructura del proceso KDD, el cual posee varias etapas a seguir, tal como se muestran en la figura 15.

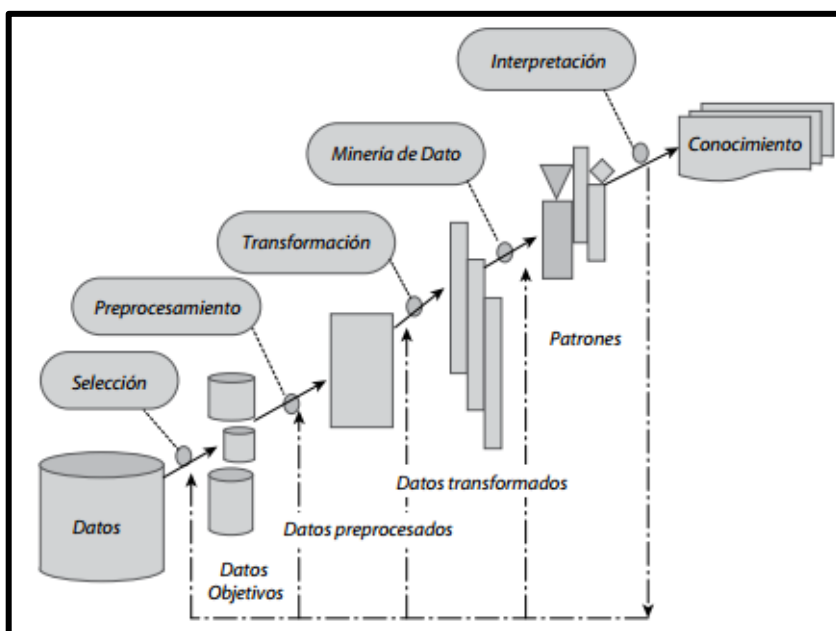


Figura 15 Etapas del proceso KDD  
Fuente: (Weber, 2000)

El Descubrimiento de conocimiento en bases de datos se trata de un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Como lo afirma (Hand, Mannila, & Smyth, 2001) “el objetivo es construir un modelo que permita predecir el valor de una variable a partir de los valores conocidos de otros datos. A continuación, se describe la metodología que este usa para el análisis de datos en todas sus etapas.

### 3.3.1. Datos

El primer paso en el proceso KDD es obtener los datos, para así poder aplicar en ellos las diferentes técnicas y algoritmos de este proceso. Para uso de este proyecto se realizó una búsqueda de repositorios y bases de datos, los cuales poseen información con características similares a las que se generarían en el ambiente del invernadero.

Se usó la base de datos *Environmental data (indoor and outdoor)* del repositorio *UMass Trace Repository*, el cual provee información recolectada por becarios de la *National Science Foundation*, o que ha sido donada a la fundación. Este repositorio permite descargar archivos con extensión .csv, es decir, separados por comas.

Particularmente, para este proyecto se utilizó el *Data Set for Sustainability*. El cual posee información con datos ambientales y variables similares a las que recolectarían en el invernadero, pese a tener más variables, esto no es un inconveniente ya que en la etapa de pre procesamiento se puede seleccionar las adecuadas. Cada archivo posee 13 variables y 300 muestras, siendo estos útiles para las pruebas de la interfaz.

Al *UMass Trace Repository* se puede acceder desde siguiente link <http://traces.cs.umass.edu/> , es de acceso gratuito, se necesita registrarse y de esta manera permiten descargar la información. Las figuras 16 y 17 indican el proceso a seguir para acceder a este repositorio, y en la Figura 18 se apreciar el set de datos descargado.

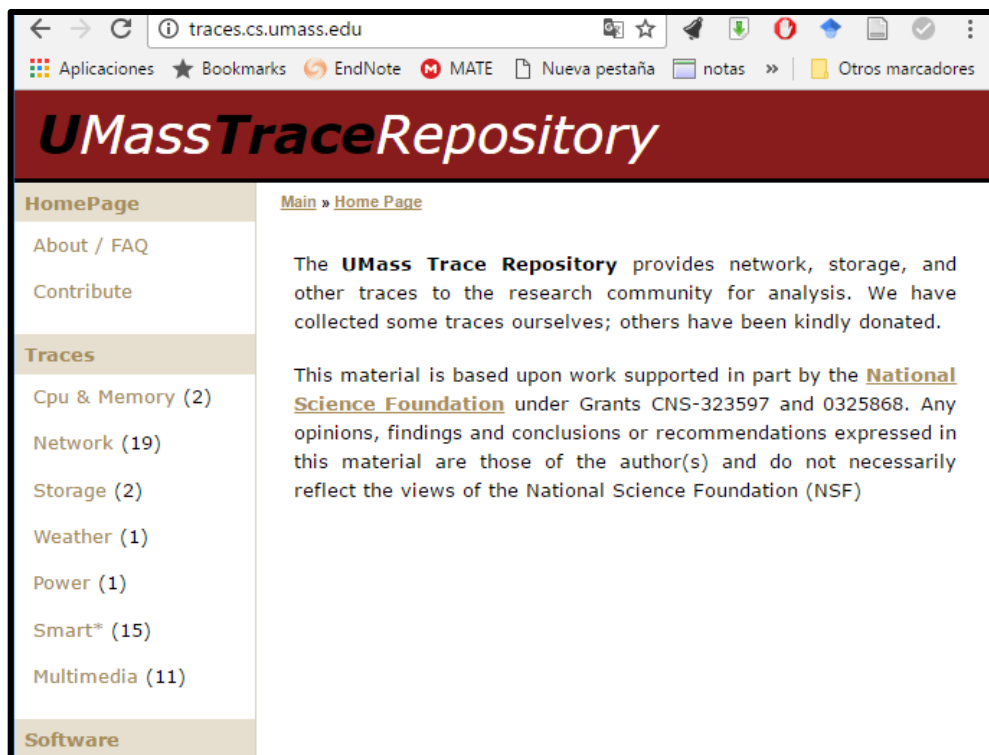


Figura 16 *UMass Trace Repository*  
Fuente: (UMassTraceRepository, 2013)

The screenshot displays the registration form for the Smart\* Repository. The form is part of the LASS website. It contains a disclaimer and several input fields for user information. The fields are filled with the following data: Name: karina.ponce, Affiliation: Universidad Técnica del Norte, Country: Ecuador, and Email (optional): kponceg@utn.edu.ec. A 'Submit' button is present at the bottom of the form. Below the form, there are two links for returning to download pages.

Figura 17 Registro para el acceso al repositorio  
Fuente: (LASS, 2013)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	TimestampUTC,insideTemp,outsideTemp,insideHumidity,outsideHumidity,windSpeed,windDirectionDegrees,windGust,windGustDirectionDegrees,rainRate,rain,windChill,heatindex													
2	1341048300,73.580009,59.269997,45.0	,83.699997,0.3	,105.0	,1.0	,293.0	,0.0	,0.0	,59.269997,59.269997						
3	1341048600,73.490005,59.161999,45.5	,83.800003,0.0	,0.0	,0.0	,0.0	,0.0	,59.161999,59.161999							
4	1341048900,73.45401,59.197998,45.700001,86.5	,0.1	,41.0	,2.0	,23.0	,0.0	,0.0	,59.197998,59.197998						
5	1341049200,73.400009,59.161999,46.0	,84.400002,0.0	,37.0	,1.0	,338.0	,0.0	,0.0	,59.161999,59.161999						
6	1341049500,73.400009,59.125996,46.0	,85.599998,0.3	,81.0	,2.0	,113.0	,0.0	,0.0	,59.125996,59.125996						
7	1341049800,73.400009,59.089996,46.0	,85.699997,0.0	,0.0	,0.0	,0.0	,0.0	,59.089996,59.089996							
8	1341050100,73.400009,59.18	,46.0	,86.0	,0.0	,0.0	,0.0	,59.18,59.18							
9	1341050400,73.274002,59.0	,46.0	,86.300003,0.0	,0.0	,0.0	,0.0	,59.0,59.0							
10	1341050700,73.219994,59.089996,46.0	,83.5	,0.0	,0.0	,0.0	,0.0	,59.089996,59.089996							
11	1341051000,73.219994,59.197998,46.0	,85.599998,0.0	,0.0	,0.0	,0.0	,0.0	,59.197998,59.197998							
12	1341051300,73.219994,59.359997,46.0	,87.0	,0.0	,0.0	,0.0	,0.0	,59.359997,59.359997							
13	1341051600,73.183998,59.539997,46.0	,86.300003,0.0	,0.0	,0.0	,0.0	,0.0	,59.539997,59.539997							
14	1341051900,73.039993,59.827995,46.0	,83.699997,0.0	,0.0	,0.0	,0.0	,0.0	,59.827995,59.827995							
15	1341052200,73.039993,60.062004,46.0	,83.599998,0.0	,0.0	,0.0	,0.0	,0.0	,60.062004,60.062004							
16	1341052500,73.039993,60.439995,46.0	,85.699997,0.0	,0.0	,0.0	,0.0	,0.0	,60.439995,60.439995							
17	1341052800,73.039993,60.637993,46.0	,84.099998,0.0	,0.0	,0.0	,0.0	,0.0	,60.637993,60.637993							
18	1341053100,73.039993,61.106007,46.0	,82.300003,0.0	,0.0	,0.0	,0.0	,0.0	,61.106007,61.106007							
19	1341053400,73.039993,61.43	,46.0	,82.900002,0.0	,0.0	,0.0	,0.0	,61.43,61.43							

Figura 18 Archivo .csv del *Data Set for Sustainability*

Fuente: Propio.

### 3.3.2. Etapa de selección

Una vez definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & AlvaradoPérez, 2016).

Como lo enuncian (Brachman & Anand, 1996), en los primeros pasos se debe enfocar en la meta del proceso desde el punto de vista del cliente. Siendo así, se ha evaluado las variables a medir en el contexto de mejorar el crecimiento de los cultivos. En un invernadero los factores que intervienen en un buen desarrollo de los sembríos son: Humedad del suelo, humedad relativa, temperatura, nivel de iluminación y CO<sub>2</sub>, las características de es estos parámetros se pueden apreciar en la Tabla 2. (Barrios, 2004).

Tabla 2 Análisis de los factores que inciden en un cultivo

<b>VARIABLE</b>	<b>IMPORTANCIA</b>
<b>Humedad del suelo y humedad relativa</b>	Factores que ayudan al desarrollo de las plantas, pero en exceso produce daños como enfermedades que son causadas por hongos y bacterias
<b>Temperatura</b>	Contralando este factor se pueden prevenir daños en los cultivos debido a las heladas o a las altas temperaturas.
<b>Iluminación</b>	Esencialmente toda la luz visible es capaz de promover la fotosíntesis, pero las regiones de 400 a 500 y de 600 a 700 nm son las más eficaces.
<b>CO<sub>2</sub></b>	Este gas carbónico es de suma importancia en el ciclo de vida de los cultivos, es un material indispensable para la fotosíntesis y la clorofila de las plantas. Combinado con agua y energía luminosa, el CO <sub>2</sub> se emplea en la fotosíntesis y mediante este proceso las plantas puedan producir carbohidratos y oxígeno.

Fuente: Adaptado de (Barrios, 2004).

### 3.3.3. Etapa de pre-procesamiento/limpieza

Esta etapa como dice su nombre es previa, donde se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para

su reemplazo (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & AlvaradoPérez, 2016).

Limpieza de datos viene del inglés Data Cleaning, trata los problemas de los datos una vez estos ya hayan ocurrido. Las estrategias de prevención de errores suelen reducir algunos de los problemas, pero no eliminarlos por completo; pues estos pueden ocurrir en cualquier etapa del flujo de datos, es decir, mientras los extraen, transfirieren, editan, seleccionan, transforman y presentan (Van den Broek, Argeseanu, Eeckels, & Herbss, 2005).

Existen maneras de realizar la limpieza de los datos: eliminar datos que faltan, suavizar el efecto del ruido, eliminar datos fuera de rango y corregir inconsistencias. En este caso se va a utilizar la herramienta de “selección”, la cual selecciona la variable target u objetivo, siendo esta aquella que depende de las demás y se la utiliza para el modelo predictivo.

En la implementación de la interfaz, el usuario será capaz de seleccionar la variable objetivo que el crea conveniente, así como, seleccionar los datos con los que desea trabajar, pues los archivos que contienen la información poseen las 5 variables que intervienen en el desarrollo de los cultivos en invernaderos, siendo esta etapa la de intuición del usuario donde podrá escoger los datos con los que desee trabajar. Particularmente, para este proyecto se usará a la humedad el suelo como variable objetivo, pues a futuro se la podría controlar por medio de un sistema de riego automatizado.

#### **3.3.4. Etapa de transformación/reducción**

Una forma natural de visualizar los datos es a través de un diagrama de dispersión de 2 ó 3 dimensiones, como se muestra en la figura 8, teniendo en cuenta que el sistema de percepción humano trabaja adecuadamente en baja dimensión, esto supone que el conjunto inicial de datos

debe representarse en un espacio de dimensión menor que la original. Este proceso se conoce como reducción de dimensión (RD) y es una etapa importante dentro de los sistemas reconocimiento de patrones y visualización de datos puesto a que está orientada a representar los datos en una dimensión menor en donde el desempeño tanto perceptual (por parte del humano) como el costo computacional mejoren, para obtener una visualización realística y más inteligible para el usuario.

Los métodos de reducción de dimensión (RD) se usan para representar los datos de manera que sean visualmente entendibles a los usuarios, pueden simplificar una tabla de una base de datos horizontal o verticalmente. Entre los métodos clásicos de RD, se encuentra el análisis de componentes principales -principal component analysis (PCA) y classical multidimensional scaling (CMDS), los cuales se basan en criterios de conservación de la varianza y la distancia, respectivamente.

Para el presente proyecto no es necesario aplicar una técnica de reducción de dimensión, pues se está manejando de cinco variables que influyen directamente en el crecimiento de los cultivos, por lo tanto, en este proceso se van a utilizar todos los datos recolectados sin necesidad de reducir su dimensión, pues se está enfocando a la minería de datos y el entendimiento de este proceso por parte del usuario, sin embargo, era necesario resaltar el funcionamiento de este proceso, debido a que forma parte del proceso KDD.

### **3.3.5. Etapa de minería de datos**

La minería de datos es la búsqueda y descubrimiento de patrones, mediante el uso de técnicas de asociación, clustering y clasificación. Con el objetivo de crear modelos predictivos o descriptivos. Para el desarrollo del análisis del presente proyecto se busca implementar un modelo predictivo. Por esto, la metodología dentro del proceso KDD permite escoger un algoritmo de

minería de datos incluyendo la selección de los métodos por aplicar en la búsqueda de patrones importantes, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos a utilizar (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & AlvaradoPérez, 2016).

#### ***3.3.5.1. Técnica de clasificación (Clasificación no supervisada)***

Uno de los métodos más usados ha sido el árbol de decisión, para este caso se ocupará el método C4.5 que es uno de los más completos. Éste es un algoritmo de clasificación supervisado, ya que el usuario interactúa seleccionando una variable objetivo (target), en base a la cual se construyen las reglas de clasificación.

En esta sección el usuario debe seleccionar el porcentaje de datos que se va a utilizar para el entrenamiento del sistema, y con este saber si el modelo aplicado trabaja de manera correcta. Al utilizar el algoritmo c4.5 que viene a ser un árbol de decisión se debe escoger el porcentaje de poda (pruning), es decir, que tan frondoso se desea que este árbol sea, y de esta manera obtener los resultados esperados.

#### **3.3.6. Etapa de interpretación y evaluación**

Esta etapa es técnicamente donde todo el proceso se retroalimenta, ya que, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. En este paso se puede visualizar los datos extraídos y se verifica que la limpieza del conjunto de datos sea correcta y no posea valores irrelevantes.

En esta sección se puede implementar el modelo predictivo, el cual necesita de parámetros que sean enviados desde procesos anteriores. Un nuevo conjunto de datos, el cual no ha sido tratado



aún, donde no se ha seleccionado la variable target previamente, además de las reglas de clasificación generadas por el algoritmo C4.5. Como se ha mencionado con anterioridad la variable a escoger es la humedad del suelo, por lo tanto, esta será la que se va a calcular, entonces, el modelo predictivo lo que hará es mostrar los futuros valores de este parámetro.

### 3.4. METODOLOGÍA DEL ANÁLISIS DE DATOS CRISP-DM

CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining, esta metodología incluye un modelo estructurado en seis fases, algunas de estas fases son bidireccionales, lo que significa que permitirán revisar parcial o totalmente las fases anteriores, en la Figura 19 se puede apreciar las etapas de esta metodología y donde interviene el proceso KDD.

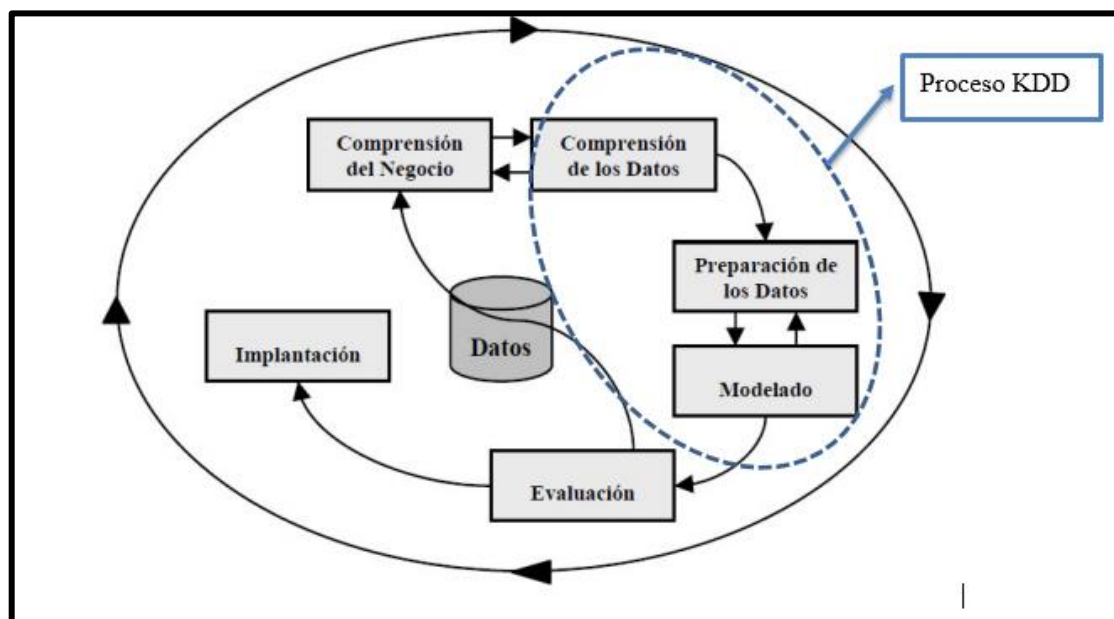


Figura 19 Metodología CRISP-DM involucrando al proceso KDD  
 Fuente: Adaptada de (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, & AlvaradoPérez, 2016)

### 3.4.1. Etapas de la metodología CRISP-DM

#### 3.4.1.1. *Comprensión del negocio*

En esta etapa se describe los objetivos y requerimientos desde una perspectiva no técnica. Un invernadero está enfocado a la crianza de diferentes plantas fuera de temporada, es decir, crea condiciones ambientales adecuadas para el desarrollo de los cultivos, la figura 20 muestra una imagen del invernadero de la granja “La Pradera”. Los factores clave en la crianza de hortalizas u otro tipo de planta son los niveles de: humedad del suelo (que varía dependiendo del riego), humedad relativa, temperatura, iluminación y CO<sub>2</sub>. En este caso de estudio se busca optimizar los recursos, mediante el análisis del comportamiento de estas variables.



Figura 20 Riego en el invernadero de la granja La Pradera  
Fuente: Propia.

Las siguientes tres etapas: Comprensión del negocio (4.4.1.2), preparación de los datos (4.4.1.3) y modelado (4.4.1.5), involucran al proceso KDD, en estas se encuentran implícitas actividades o procesos que se detallan en la sección 4.3.

### ***3.4.1.2.Comprensión de los datos***

Familiarizarse con los datos teniendo presente los objetivos del negocio.

### ***3.4.1.3.Preparación de los datos***

Obtener la vista minable o el conjunto de datos para analizar.

### ***3.4.1.4.Modelado.***

Aplicar las técnicas de minería a los datos.

### ***3.4.1.5.Evaluación***

De los modelos de las fases anteriores para determinar si son útiles a las necesidades del negocio.

### ***3.4.1.6.Despliegue***

Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización.

Al implementar soluciones de minería de datos existen tres maneras de hacerlo, como se muestra en la Tabla3.

Tabla 3 Tipo de implementación de herramientas de minería de datos

<b>Herramienta</b>	<b>Característica</b>
<b>Débilmente acoplada</b>	Las técnicas y algoritmos se encuentran fuera del Sistema Gestor de Base de Datos (SGBD), implementando la solución a través de una interfaz. (R, 2012)

<b>Medianamente Acoplada</b>	Ciertas funciones y tareas forman parte del SGBD. (R, 2012)
<b>Fuertemente acoplada</b>	Todas las funciones, tareas, algoritmos se encuentran en el SGBD, con operaciones primitivas.

Fuente: Propia

En la Figura 21 se muestra la arquitectura de una herramienta débilmente acoplada, en esta la interfaz gráfica de usuario y los algoritmos de minería de datos se encuentran desarrollados de manera que sean fáciles de usar, y que permita realizar el modelo predictivo de la variable objetivo (humedad del suelo) con base en los valores de las otras variables ambientales que inciden en el desarrollo de los cultivos del invernadero. El sistema gestor de base de datos es totalmente independiente, por lo tanto, el usuario es quien está encargado de seleccionar el archivo con la información necesaria para el análisis, de manera que los datos sean cargados fácilmente desde un panel de selección.

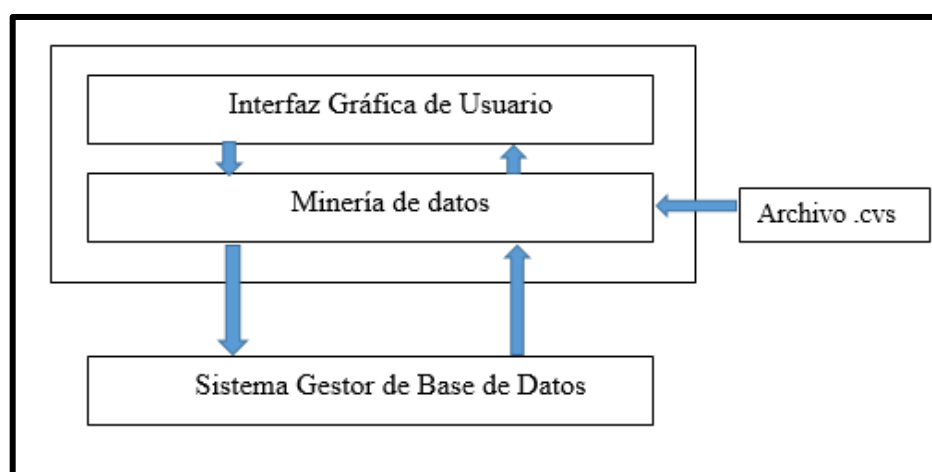


Figura 21 Diagrama de la arquitectura de una herramienta débilmente acoplada.  
Fuente: Propia.

### 3.5. DIAGRAMAS DE CASO DE USO

Se han propuesto dos casos de uso debido a que, la interfaz genera dos procesos: Clasificación y predicción. En el primero se podrá observar el árbol de decisión que genera el algoritmo C4.5 y las reglas de clasificación que han sido calculados por el mismo. Con esta información y adicionando un conjunto de datos nuevos, el modelo predictivo podrá ser ejecutado.

#### 3.5.1. Diagrama general de la interfaz de análisis de datos.

En la interfaz desarrollada, existe un caso de uso general, donde se muestra el proceso que un usuario debería tomar para tener la visualización del árbol de decisión y las reglas de clasificación, que vienen a ser el resultado del análisis de los datos.

El usuario tiene el control del funcionamiento de la interfaz, desde que esta inicia (Figura 22). Existen tres tipos de usuario: el encargado del invernadero (horticultor), un analista de datos, o un usuario externo que quiera acceder a la interfaz, y el administrado.

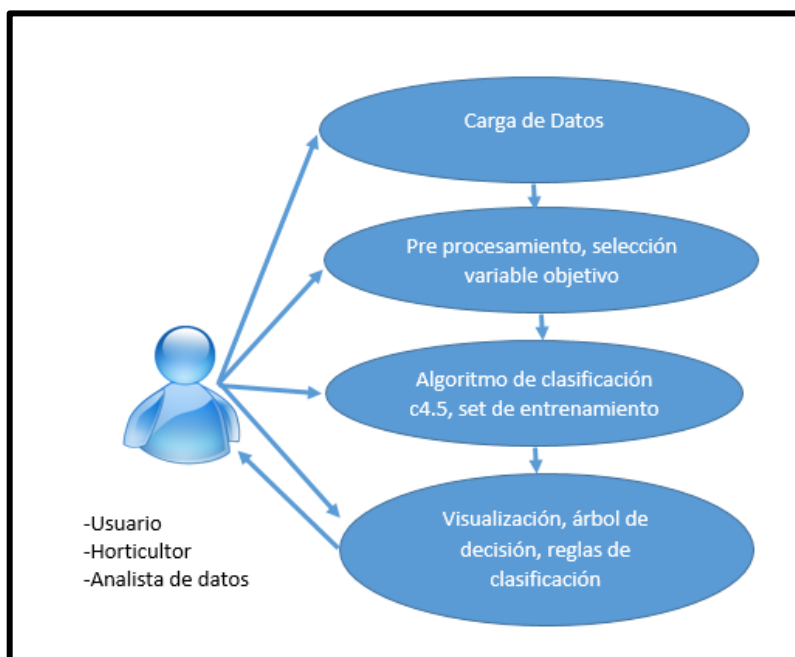


Figura 22 Caso de uso del funcionamiento general de la interfaz  
Fuente: Propia.

A continuación, en la tabla 4 se describe cada uno de estos y su función.

Tabla 4 Usuarios de la interfaz

Tipo de usuario	Uso de la interfaz
Encargado del invernadero	Puede acceder a la interfaz, y mediante recomendaciones usar los diferentes módulos de los cuales está conformada.
Analista de datos	Podrá modificar los valores en los algoritmos a su conveniencia, su conocimiento le permitirá escoger el porcentaje del set de entrenamiento que le parezca más útil.
Administrador	Es quien ha creado el sistema, así como el analista por su conocimiento será capaz de modificar como le parezca conveniente los parámetros solicitados en cada proceso.

Fuente: Propia.

Por lo tanto, el proceso que debe tomar el usuario para el funcionamiento de este aplicativo debe iniciar con la carga del archivo que contiene la información. Seguido de esto debe entrar al pre procesamiento de los datos, donde se debe escoger a la variable objetivo (humedad del suelo). Después, pasa por el algoritmo de clasificación c4.5, aquí se debe configurar los parámetros necesarios como el set de entrenamiento y el porcentaje de poda para la generación del árbol de decisión. Por último, se ejecuta la visualización donde el árbol las reglas de clasificación pueden ser observados.

### 3.5.2. Diagrama de caso de uso con predicción

Para generar el modelo predictivo, el usuario debe realizar nuevas configuraciones. En este caso de uso se debe partir desde el proceso anterior, para generar un modelo predictivo se necesita de un nuevo conjunto de datos que no haya sido entrenado y en este se va a predecir el comportamiento de la variable objetivo (humedad del suelo), entonces, el usuario debe proporcionar dos parámetros: conectar las reglas de clasificación generadas en el proceso anterior (clasificación C4.5) y el nuevo set de información, con base en estos parámetros se realiza un algoritmo de regresión para predecir valores futuros de la variable objetivo, el proceso a seguir se aprecia en la Figura 23.

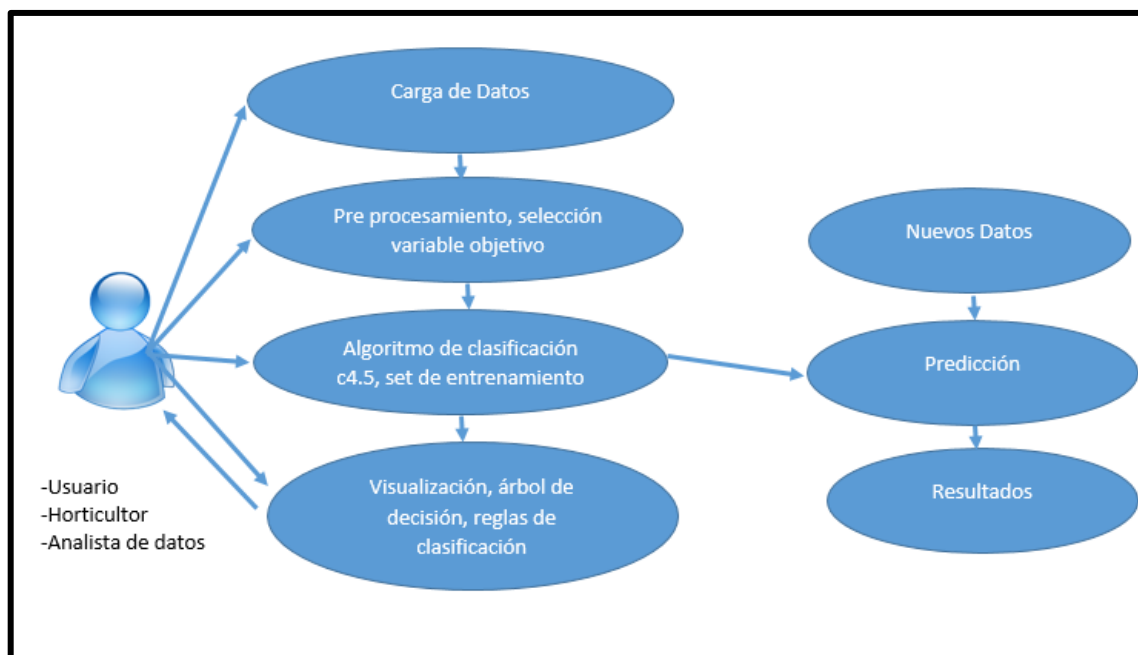


Figura 23 Caso de uso del funcionamiento de la interfaz con el modelo predictivo.

Fuente: Propia.

### 3.6. MÓDULOS DE LA INTERFAZ

A continuación, se describe el funcionamiento de la interfaz de manera modular, explicando cada etapa y las funciones del usuario. En la Figura 24 se describe cada uno de los módulos que conforman la interfaz.

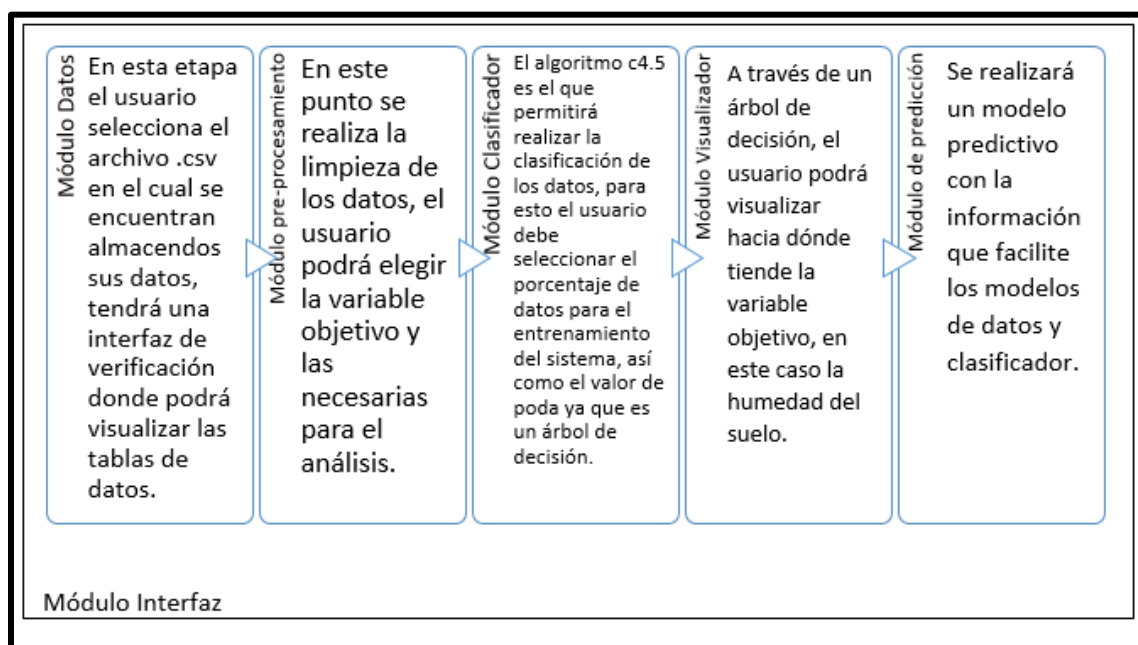


Figura 24 Funcionamiento de la interfaz en estructura de módulos  
Fuente: Propia.

#### 3.6.1. Módulo Datos

Este módulo posee una sub interfaz a la cual se accede para cargar los datos, la información debe estar almacenada en un archivo de extensión csv, es decir separado por comas “,”. De esta manera el sistema reconocerá la separación entre atributos y podrá manejarlos de una manera adecuada. La tabla 5 muestra la estructura de los datos a manejar, en esta se puede observar el tipo de variables a usar, ya que este contiene un encabezado que indica a qué tipo de información se ha almacenado; en el anexo A se muestra un set de datos completo.



Tabla 5 Estructura de los datos almacenados en el invernadero

Temperatura	Humedad relativa	Luz	CO2	Humedad de suelo
27	28	2082	19	650
27	27	2583	19	781
26	29	2571	18	762
26	30	2640	18	763
26	30	560	19	762

Fuente: Propia.

### 3.6.2. Módulo Pre procesamiento

A este proceso lo denominan también limpieza de los datos (Data Cleansing o Data Scrubbing), aquí se preparan los datos para el usuario pueda aplicar el algoritmo de clasificación que genere el modelo deseado. En esta etapa se utiliza la herramienta de selección, donde el usuario interviene al escoger los datos con los que desea trabajar, a través de un panel podrá seleccionar las columnas a su conveniencia. Al usar la herramienta de selección, se permite indicar la variable objetivo, que será usada en los siguientes procesos o algoritmos. Particularmente, se debe seleccionar a la variable humedad del suelo como como objetivo, ya que, en la búsqueda de optimizar recursos, ésta permitirá realizar un análisis para el uso adecuado del riego de agua.

### 3.6.3. Módulo Clasificación

Este módulo recibe los datos previamente preparados y con la variable objetivo (target) ya seleccionada en el módulo pre procesamiento. Se utiliza el algoritmo C4.5 para realizar la clasificación de los datos, siendo un árbol de decisión basado en la entropía, es decir, que busca el conocimiento con base en el desgaste del sistema; el funcionamiento del C4.5 se muestra en el anexo B. El usuario deberá seleccionar el porcentaje de poda para el árbol de clasificación, así

como el porcentaje del set de entrenamiento. Usualmente se trabaja con el 80% de los datos para generar las reglas de clasificación.

#### **3.6.4. Módulo Visualización**

Este recibe las reglas de clasificación que son generadas por el proceso anterior, de manera que grafica la tendencia de la variable objetivo a través de un árbol, donde la ramificación va a depender de los parámetros configurados previamente en el módulo de clasificación.

#### **3.6.5. Módulo Predicción**

Este módulo genera una nueva solución, es decir, con la información que es proporcionada por el módulo de clasificación y con un nuevo conjunto de datos, se genera un modelo predictivo. La predicción se realiza a través de la regresión, es decir, que las reglas obtenidas del proceso anterior (clasificación) son usadas para calcular el valor futuro de la variable objetivo (humedad del suelo). Este cálculo se basa en los valores que tienen las variables independientes (humedad relativa, temperatura ambiental, iluminación y CO<sub>2</sub>) y usando las reglas de clasificación se obtiene los valores futuros de la variable objetivo (humedad del suelo).

#### **3.6.6. Módulo Interfaz**

El módulo interfaz es usado para integrar todos los módulos los anteriormente mencionados en una sola solución, es decir, a los módulos: Datos, pre procesamiento, clasificación, visualización y predicción, permitiendo que trabajen unos con otros para lograr el funcionamiento del sistema.

## **CAPÍTULO IV IMPLEMENTACIÓN**

### **4.1. INTRODUCCIÓN**

A continuación, se describe todos los parámetros y procesos que se usan en el desarrollo de la solución. Este proyecto es una interfaz de analítica de datos en agricultura, enfocado a los factores ambientales que inciden el crecimiento de cultivos en invernaderos, es una aplicación débilmente acoplada, es decir, que es fácil de usar e intuitivo para el usuario, con un bajo costo computacional.

### **4.2. DESCRIPCIÓN DE LA HERRAMIENTA**

Esta interfaz está diseñada para el análisis de datos en cultivos de invernaderos, ésta será intuitiva y sencilla de usar para el horticultor. Con un proceso de arrastrar y soltar los íconos, el usuario podrá seleccionar las técnicas necesarias, conectarlas entre sí y configurar los parámetros necesarios para el funcionamiento del aplicativo. Siendo así, se la puede definir como una herramienta débilmente acoplada, pues ha sido implementada de manera independiente del Sistema Gestor de Base de Datos.

En este desarrollo se usó el software Net Beans IDE 8.2 que maneja el lenguaje de programación java. La interfaz se encuentra conformada por 6 módulos: Datos, pre procesamiento, clasificador, visualizador, predicción y la interfaz gráfica de usuario; cada uno de estos posee una *JFrame form* y un *java class*, los cuales poseen los métodos necesarios para el funcionamiento de los módulos, todos estos son llamados en el *JFrame MyCanvas* en el cual se unifican todas las etapas antes mencionadas.

### 4.3. DESARROLLO DEL SOFTWARE

Este software de análisis de datos en agricultura, enfocado especialmente a los factores ambientales de un invernadero, comprende 5 etapas las cuales son: Datos, pre procesamiento, minería de datos, visualización y predicción; estos módulos se detallan en la Figura 24, en la cual se indica las funciones del usuario en cada uno de estos pasos.

Este software ha sido desarrollado con el uso de varios paquetes y clases, las cuales se describen a continuación en la Figura 25 para un mejor entendimiento, los diagramas de paquetes y de clases se muestran en el Anexo D, donde se puede apreciar como se encuentra estructurado el software.

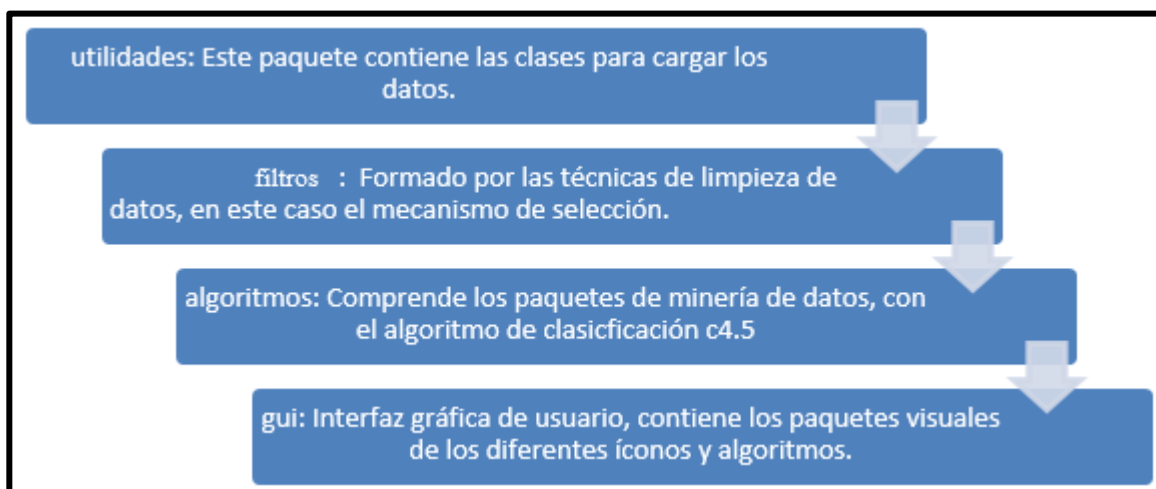


Figura 25 Paquetes y clases usados en la programación de la interfaz  
Fuente: propia.

#### 4.3.1. Paquete utilidades

Este paquete se compone de clases como AsoReg, filemanager, e itemset, las cuales permiten generar las reglas de asociación, manejar los archivos .csv y la generación de los nodos intermedios en los nodos del árbol de decisión, expuestas en la Tabla 6.

Tabla 6 Componentes del paquete utilidades

Clase	Descripción
<b>AsoReg</b>	Genera las reglas de asociación para los algoritmos, a partir de la variable de confianza.
<b>fileManager</b>	Encargada del funcionamiento de los archivos .csv que se cargan al sistema.
<b>itemset</b>	Clases encargadas de la generación de nodos intermedios en los árboles decisión, y las ramificaciones de este.

Fuente: Propia.

#### 4.3.2. Paquete filtros

Este ha sido utilizado para realizar la limpieza de los datos a través de la selección de los atributos necesarios para el análisis. El usuario debe seleccionar la variable objetivo, la cual servirá de guía para realizar la clasificación en los pasos posteriores, la Tabla 7 muestra las clases que usa este paquete para funcionar.

Tabla 7 Componentes del paquete Filtros

Clase	Descripción
<b>seleccion</b>	Posee los métodos para trabajar en la nueva tabla a crearse, como son: <i>nuevaTabla</i> , <i>getRowCount</i> , <i>getColCount</i> , <i>getValue</i> , <i>setValue</i> .
<b>abrirSel</b>	Es el <i>JFrame</i> que permite visualizar la tabla con los tipos de variables originales (se toman de la primera fila del conjunto de datos original, que comúnmente posee los nombres de las variables), y seleccionar las columnas con las que el usuario desea trabajar, así como la variable objetivo, denominada como <i>target</i> .

<b>verSel</b>	Este es un JFrame que muestra tanto el conjunto de datos original, como la nueva tabla generada con la información seleccionada por el usuario.
---------------	---

Fuente: Propia.

### 4.3.3. Paquete de algoritmos

Este contiene las clases que permiten el uso de los diferentes algoritmos de clasificación, en esta solución se utiliza el algoritmo c4.5, el cual permite realizar un modelo basado en la entropía de los datos y así generar conocimiento. Se realiza un conteo de los valores que ha tomado la variable objetivo, la cual ya ha sido seleccionado en el proceso anterior, mediante esta se realiza un cálculo de la entropía de las demás variables, buscando el valor que genere más ganancia, ya que este se convertirá en un nodo dentro del árbol que genera este algoritmo. El proceso se detiene cuando ya no existan valores que clasificar.

En este paquete se ubican las clases que permiten los cálculos de las reglas de clasificación del árbol c4.5, la Tabla 8 describe a cada una de estas.

Tabla 8 Componentes del paquete algoritmos

<b>Clase</b>	<b>Descripción</b>
<b>Atributo</b>	Es una clase que posee los métodos para calcular la entropía del sistema.
<b>C45TreeModel</b>	Clase que contiene métodos para la generación de los nodos del árbol de decisión.

<b>C45TreeGUI</b>	Métodos de características gráficas del árbol generado, para expandir o aumentar el gráfico del árbol.
-------------------	--

Fuente: Propia

#### 4.3.4. Paquete gui

La interfaz gráfica de usuario está diseñada de tal manera que sea amigable con el usuario, con la opción de arrastrar y soltar los íconos y que el usuario pueda configurar los parámetros necesarios en los algoritmos. Este paquete contiene las clases de los diferentes íconos, los cuales permiten al usuario desempeñar las actividades de un analista de datos de una manera fácil y entendible, la Tabla 9 indica las características y funciones de cada clase que compone el paquete GUI.

Tabla 9 Componentes del paquete GUI

<b>Clase</b>	<b>Descripción</b>
<b>MiCanvas</b>	Frame principal, contiene un JPanel en el que se instancian los JLabels como íconos y así acceder a las propiedades de los mismos. Posee métodos de serialización (setGrapho y getGrapho), que permite guardar y abrir los proyectos realizados.
<b>SeleccPanel</b>	Esta es la clase que contiene los métodos para seleccionar las ventanas del <i>ScrollPane</i> , que permite mostrar el menú inicio y herramientas en orden.
<b>Contenedor</b>	Con JSplitPane, permite dividir al entorno gráfico en dos secciones, con dos <i>ScrollPane</i> uno dedicado a mostrar las herramientas,

	ubicado a la derecha, y el espacio restante es donde se encuentra el canvas para arrastrar y soltar los íconos.
<b>mipnlHerramientas</b>	Panel que posee los íconos: Datos, Selección, Clasificación, Visualización y Predicción.
<b>mipnlInicio</b>	Panel de Inicio que contiene las opciones para abrir, crear o guardar un proyecto nuevo.

Fuente: Propia.

#### 4.3.5. Interfaz de análisis de datos

El presente proyecto se ha implementado de tal forma que sea de fácil uso para el usuario, a continuación, se muestra cada una de las partes de las que está conformado y cómo estas funcionan. La interfaz se encuentra conformado por dos partes que se muestran en forma de panel de selección, posee un panel de selección con dos pestañas, la primera ha sido denominada inicio posee las opciones para abrir, crear o guardar un proyecto, además de tener un botón de ayuda. Las figuras 26, 27, 28 y 29 muestran la interfaz desarrollada, indicando las partes que esta posee.



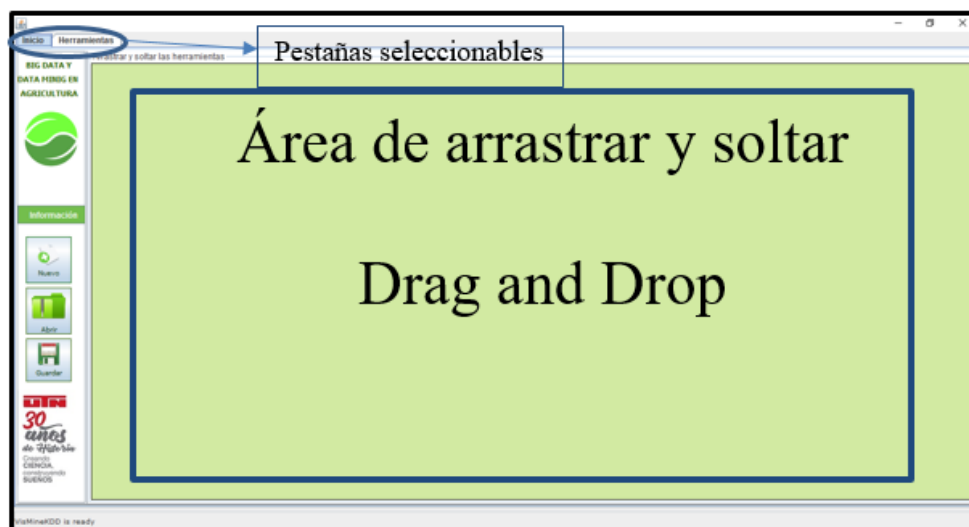


Figura 26 Pantalla de inicio de la interfaz gráfica de usuario.  
Fuente: Propia.

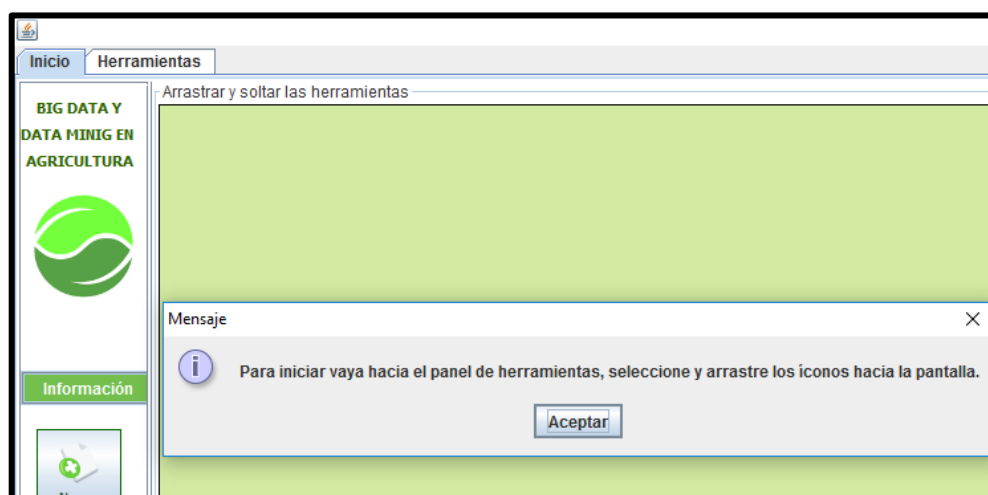


Figura 27 Texto mostrado por el botón informativo de la pantalla de inicio.  
Fuente: Propia

Esta pestaña de inicio posee seis partes, un área informativa que muestra el motivo del sistema, a continuación, un botón informativo que muestra el siguiente paso a seguir, los tres botones son para empezar un nuevo proyecto, abrir uno ya realizado o guardar lo que se ha venido trabajando, por último, se tiene el logo representativo de la institución.



Figura 28 Descripción de los íconos formados por la pestaña Inicio  
Fuente: Propia.

La segunda pestaña llamada herramientas posee todos los procesos que se usan en el análisis: la selección del archivo de datos, el pre procesamiento con la elección de la variable objetivo, el algoritmo de clasificación de datos c4.5, y la visualización que se realiza a través de un árbol de decisión, además de una opción de predicción. En el centro se encuentra el canvas, es decir, el área donde el usuario podrá configurar los parámetros de cada uno de los procesos y conectarlos entre sí para su funcionamiento.

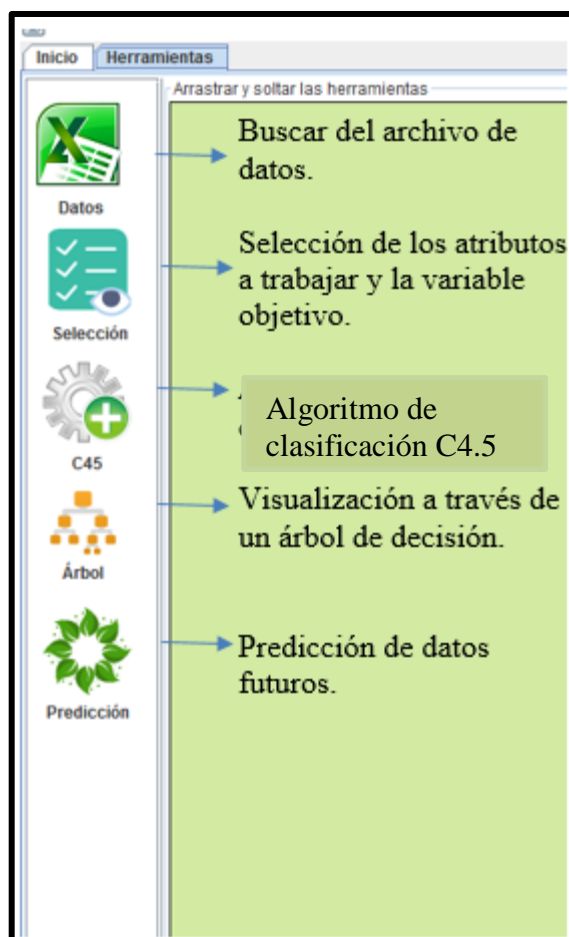


Figura 29 Descripción de los íconos formados en la pestaña Herramientas.

Fuente: Propia.

#### 4.3.6. Algoritmos y técnicas

Las secciones expuestas a continuación, son los módulos que conforman la interfaz de análisis de datos en invernaderos.

##### 4.3.6.1. Datos

En esta sección el usuario puede seleccionar un archivo de extensión .csv (valores separados por comas) alojado en su computador, dependiendo del carácter por el que se encuentren separados

los datos sea una “,” o “;”. Este proceso permite visualizar la información en una tabla, y de esta manera se pueda cerciorar que el archivo es el correcto. Se utilizó la librería *JavaCSV*, la cual permite la lectura y escritura de archivos de este tipo. En la figura 30 aprecia la carga del archivo .csv para el análisis de datos.

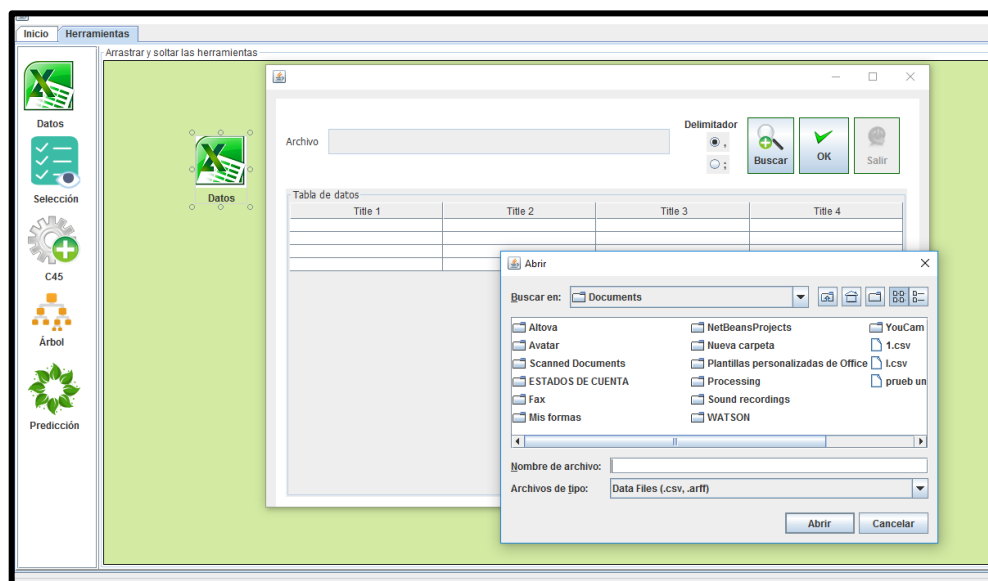


Figura 30 Funcionamiento módulo datos de interfaz de análisis de datos.

Fuente: Propia.

#### 4.3.6.2. Selección

Como su nombre lo indica, este módulo permite al usuario seleccionar la variable objetivo o target (humedad del suelo), así como también las variables con las que se desea trabajar. Este proceso almacena los datos seleccionados en una nueva tabla y la muestra al usuario, también etiqueta a la variable objetivo para que el siguiente módulo la tome en cuenta. La Figura 31 muestra el funcionamiento del módulo selección.

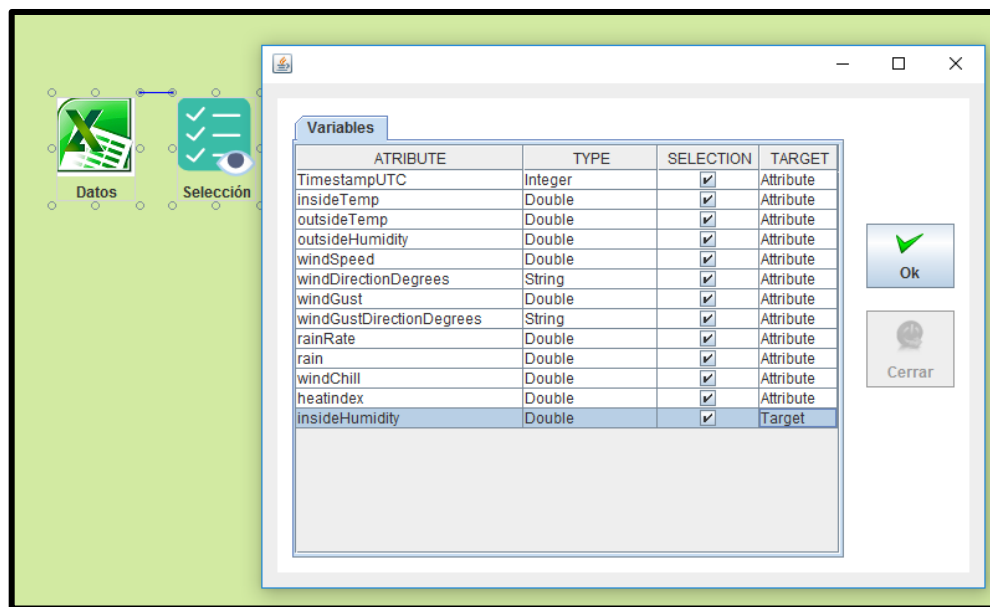


Figura 31 Funcionamiento módulo selección en la interfaz de análisis de datos.

Fuente: Propia.

#### 4.3.6.3. Clasificación

En este módulo se desarrolló el algoritmo de clasificación, el cual forma las reglas para el árbol de decisión C4.5, para esto el usuario debe dar parámetros al sistema para su correcto funcionamiento, esto dependerá del punto de vista del analista. El set de entrenamiento es el conjunto de datos con el que se va a construir las reglas de clasificación, las filas por nodo, son el número de filas de datos que algoritmo analizará por cada nodo que ser forme, y el porcentaje de límite se refiere a cuan frondoso se quiere visualizar el árbol, este se puede ir cambiando de acuerdo al nivel de entendimiento del usuario. En la Figura 32 se aprecia al módulo clasificación funcionando.

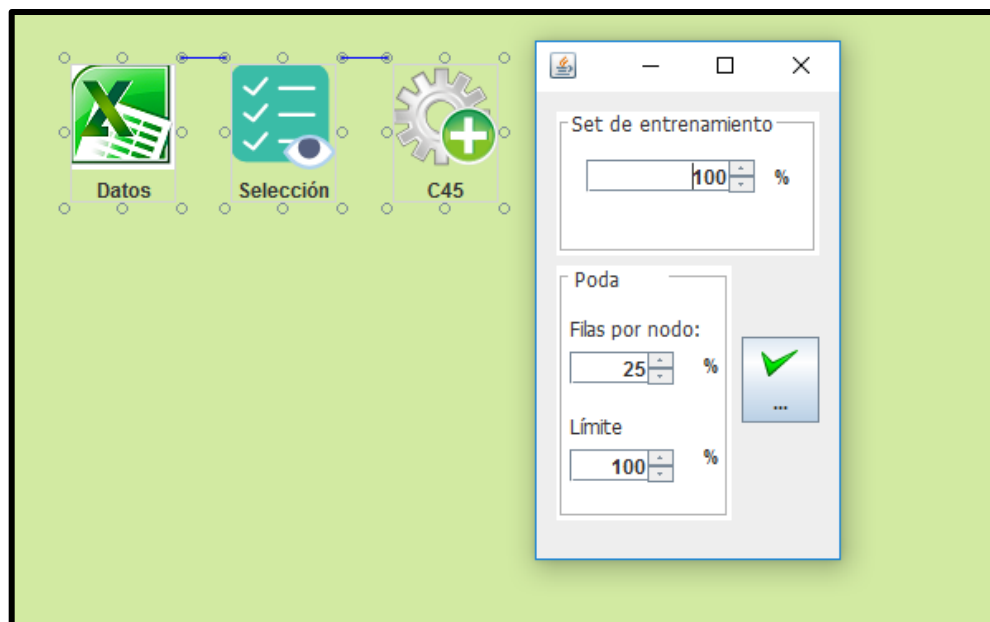


Figura 32 Funcionamiento módulo clasificación en la interfaz de análisis de datos.  
Fuente: Propia.

#### 4.3.6.4. Visualización

Para visualizar los resultados del algoritmo de clasificación se ha utilizado el WekaTreeView, tomado de la herramienta weka. Con base en estos datos y las reglas que se generaron en el algoritmo C4.5, el árbol se gráfica, sin necesidad de pedir otro parámetro adicional. La Figura 33 muestra un esquema de conexión para llegar a este módulo, en la Figura 34 se puede observar un árbol generado por la herramienta, el cual ha resultado particularmente muy frondoso debido a la naturaleza numérica de los datos y la cantidad de los mismos. Las reglas que han sido utilizadas en este proceso de clasificación, se muestran la pestaña Rules, la Figura 35 indica aquellas que han sido particularmente calculadas para este proceso y el porcentaje de confianza que genera.



Confidence Tree : 100%

Weka Tree Rules

Rules set

#	Rules	Class	Confidence
1	TimestampUTC=1341134700	insideHumidity=48.299999 [1/1]	100%
2	TimestampUTC=1341135000	insideHumidity=48.0 [1/1]	100%
3	TimestampUTC=1341135300	insideHumidity=48.200001 [1/1]	100%
4	TimestampUTC=1341135600	insideHumidity=49.0 [1/1]	100%
5	TimestampUTC=1341135900	insideHumidity=48.299999 [1/1]	100%
6	TimestampUTC=1341136200	insideHumidity=48.900002 [1/1]	100%
7	TimestampUTC=1341136500	insideHumidity=48.200001 [1/1]	100%
8	TimestampUTC=1341136800	insideHumidity=49.0 [1/1]	100%
9	TimestampUTC=1341137100	insideHumidity=49.0 [1/1]	100%
10	TimestampUTC=1341137400	insideHumidity=49.0 [1/1]	100%
11	TimestampUTC=1341137700	insideHumidity=48.599998 [1/1]	100%
12	TimestampUTC=1341138000	insideHumidity=48.799999 [1/1]	100%
13	TimestampUTC=1341138300	insideHumidity=48.900002 [1/1]	100%
14	TimestampUTC=1341138600	insideHumidity=48.799999 [1/1]	100%
15	TimestampUTC=1341138900	insideHumidity=48.900002 [1/1]	100%
16	TimestampUTC=1341139200	insideHumidity=48.900002 [1/1]	100%
17	TimestampUTC=1341139500	insideHumidity=48.799999 [1/1]	100%
18	TimestampUTC=1341139800	insideHumidity=48.799999 [1/1]	100%
19	TimestampUTC=1341140100	insideHumidity=48.900002 [1/1]	100%
20	TimestampUTC=1341140400	insideHumidity=48.599998 [1/1]	100%
21	TimestampUTC=1341140700	insideHumidity=48.900002 [1/1]	100%

Save Report

Figura 35 Visualización de las reglas de clasificación en la interfaz de análisis de datos.

Fuente: Propia.

#### 4.3.6.5. Predicción

El algoritmo de predicción trabaja con base en las reglas formadas por el módulo clasificación C4.5. El set de datos con el que se realizaron las etapas anteriores funciona ahora como el conjunto de información de entrenamiento, ahora para realizar predicción se necesita de un archivo con datos nuevos, el cual desconoce la variable objetivo (humedad del suelo). Por lo tanto, los parámetros que se deben enviar a esta herramienta, mediante la conexión entre estas (esto se realiza manualmente en la interfaz, conectándolos en un punto, el proceso se muestra en las figuras 36 y 37) son la clasificación weka realizada y el nuevo archivo. Con la información ya mencionada, se realiza el proceso de predicción, y dando como resultado los valores que tomará la variable objetivo.



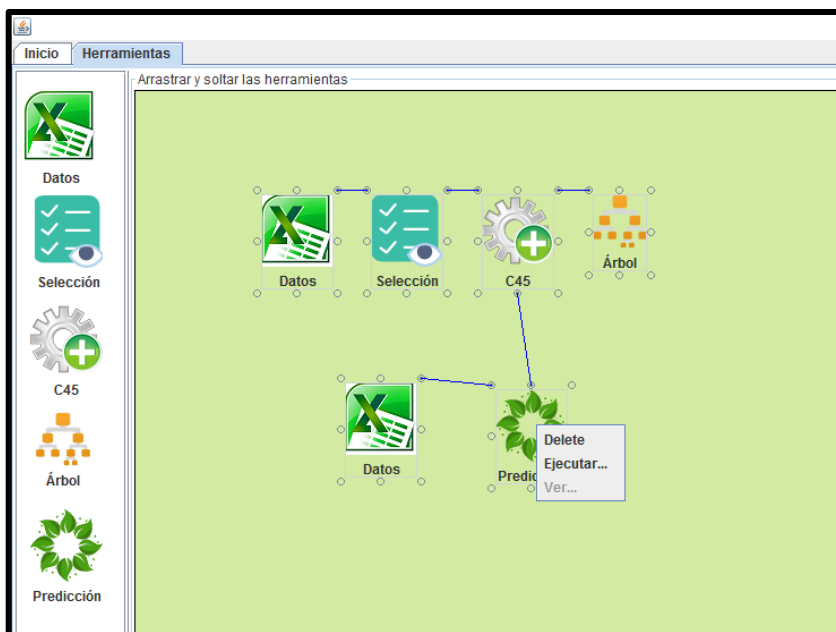


Figura 36 Funcionamiento del módulo predicción en la interfaz de análisis de datos.  
Fuente: Propia.

The screenshot shows a data table with columns for 'Variables', 'Input Data', and 'Prediction'. The 'Prediction' column shows the predicted values for the target variable 'insideHumidity'. The table contains 20 rows of data, with the first row being the header and the subsequent rows representing individual data points. The predicted values are consistently 0.0, indicating that the model is predicting zero for the target variable.

Time...	inside...	outsid...	outsid...	windS...	wind...	wind...	wind...	rainR...	rain	windC...	heatin...	inside...	inside...
1341...	72.5	61.286	89.3	0		0		0	0	61.286	61.286	48.3	48.29...
1341...	72.5	61.178	88.8	0		0		0	0	61.178	61.178	48	48.0
1341...	72.5	61.322	88.8	0		0		0	0	61.322	61.322	48.2	48.20...
1341...	72.5	61.034	92	0		0		0	0	61.034	61.034	49	49.0
1341...	72.5	61.052	90.6	0		0		0	0	61.052	61.052	48.3	48.29...
1341...	72.5	60.8	92.7	0.39	0	1.315	0	0	0	60.8	60.8	48.9	48.90...
1341...	72.5	61.034	91.1	0.29	0	1.90	0	0	0	61.034	61.034	48.2	48.20...
1341...	72.5	61.034	91.7	0		0		0	0	61.034	61.034	49	49.0
1341...	72.5	60.836	91.1	0		0		0	0	60.836	60.836	49	49.0
1341...	72.5	60.98	90	0		0		0	0	60.98	60.98	49	49.0
1341...	72.5	61.034	91.5	0		0		0	0	61.034	61.034	48.6	48.59...
1341...	72.41	61.25	91.8	0.38	0	2.315	0	0	0	61.25	61.25	48.8	48.79...
1341...	72.32	61.304	93.3	0.19	0	2.338	0	0	0	61.304	61.304	48.9	48.90...
1341...	72.32	61.664	93.6	0.9	0	3.135	0	0	0	61.664	61.664	48.8	48.79...
1341...	72.32	61.934	94.2	0.67	0	3.338	0	0	0	61.934	61.934	48.9	48.90...
1341...	72.32	62.204	91.7	0		0		0	0	62.204	62.204	48.9	48.90...
1341...	72.32	62.456	91	0		0		0	0	62.456	62.456	48.8	48.79...
1341...	72.302	62.762	90.5	0		0		0	0	62.762	62.762	48.8	48.79...
1341...	72.212	63.086	91	0		0		0	0	63.086	63.086	48.9	48.90...
1341...	72.14	63.392	90.9	0.23	0	1.0	0	0	0	63.392	63.392	48.6	48.59...
1341...	72.14	63.878	91.2	0.15	0	0.45	0	0	0	63.878	63.878	48.4	48.40...
1341...	72.032	64.418	89.6	0		0		0	0	64.418	64.418	48	48.0
1341...	71.96	64.85	88.6	0		0		0	0	64.85	64.85	48	48.0
1341...	71.978	65.354	87.6	0		0		0	0	65.354	65.354	48	48.0
1341...	72.14	65.804	87.4	0		0		0	0	65.804	65.804	48	48.0

Figura 37 Visualización de la predicción de la variable target (Humedad del suelo)  
Fuente: Propia.

## 4.4. PRUEBAS DE FUNCIONAMIENTO y RESULTADOS

### 4.4.1. Pruebas Usando el Data Set for Sustainability

#### 4.4.1.1. Caso 1: Clasificación

Para evaluar los resultados de este proyecto, en primera instancia se usó la base datos *Data Set for Sustainability* del repositorio *UMass Trace Repository*, explicado en la sección 3.3.1. A continuación, se mostrará cada uno los procesos y algoritmos de la interfaz.

La pantalla de inicio de la interfaz, en la cual se puede crear un nuevo proyecto, abrir uno en el que ya se haya trabajado o guardar el progreso que se tiene. El botón Información da un mensaje al usuario indicando lo que debe hacer para iniciar su análisis de datos, como se muestra en la Figura 38.

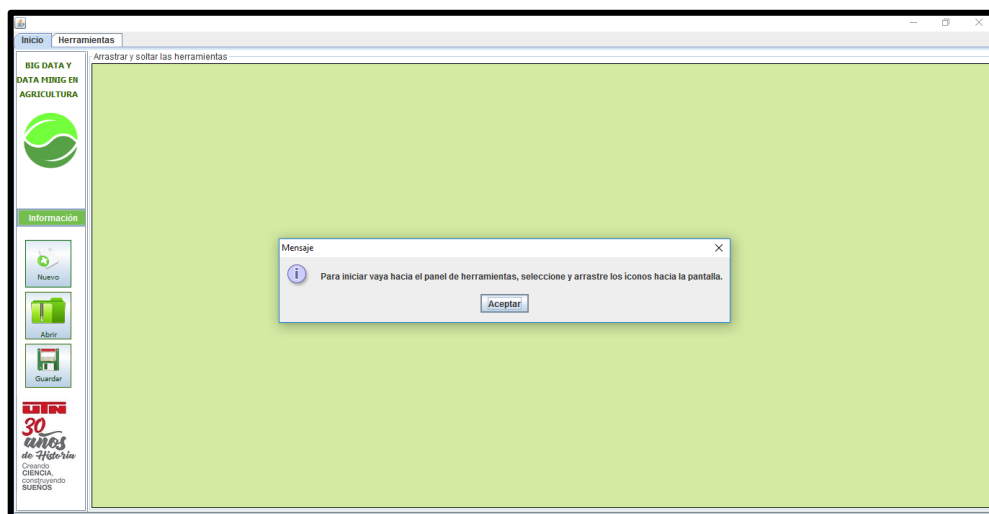


Figura 38 Pantalla de inicio de la interfaz de análisis de datos en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Se procede al panel herramientas, en este se encuentran las etapas en orden para ser aplicadas, esta interfaz funciona con el proceso de arrastrar y soltar (*Drag and Drop*), siendo fácil de usar para el usuario e interactivo. En la Figura 39 se muestra los elementos en el *canvas*, que viene a ser el área de trabajo, y el primer paso que es seleccionar los datos a analizar.

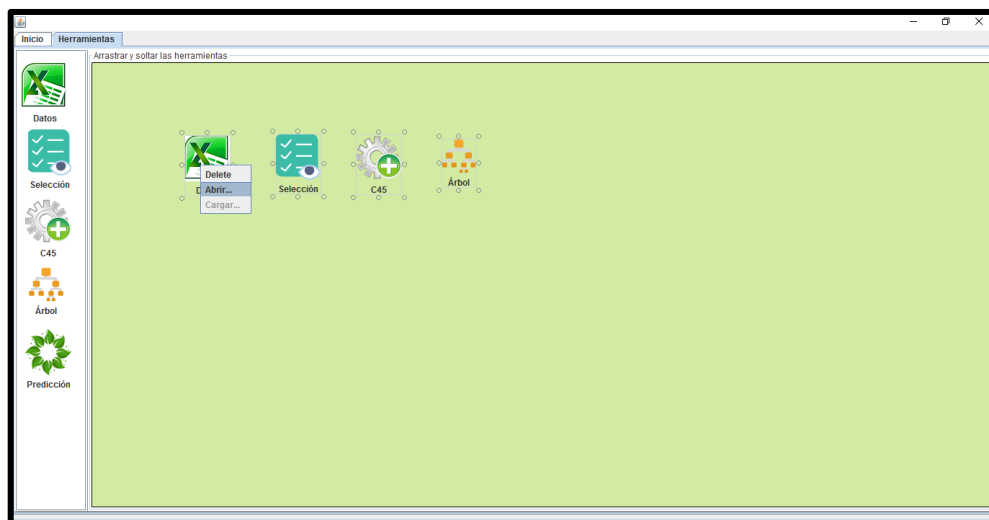


Figura 39 Elementos en el canvas de la interfaz de análisis de datos en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Para seleccionar el archivo de extensión .csv, se abre un *frame*, el cual posee las opciones para buscar la información, además de permitir seleccionar el caracter con el que se encuentran separados los datos. La Figura 40 muestra el *frame* donde se van a cargar los datos, y en la Figura 43 su visualización en una tabla.

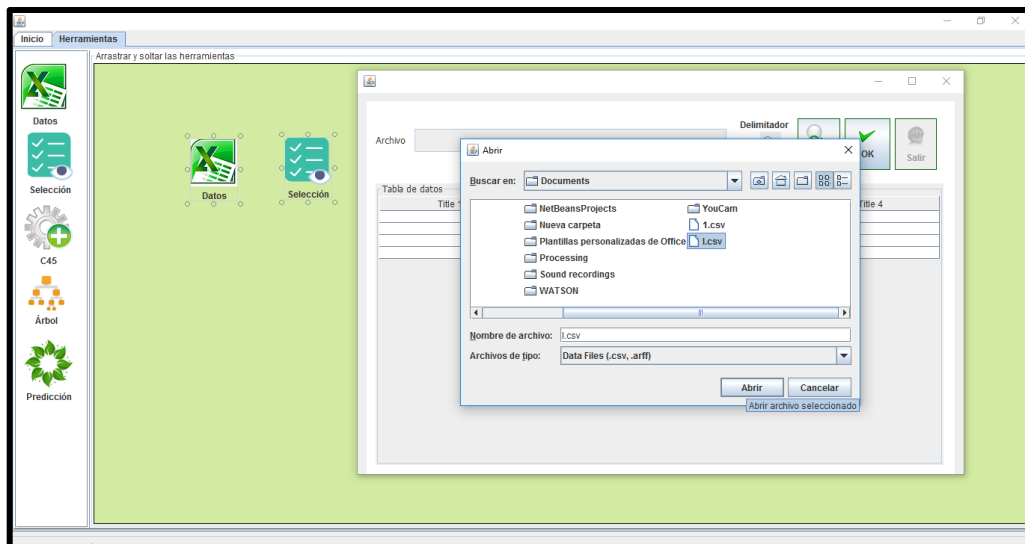


Figura 40 Carga de datos y visualización n en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Timestamp...	insideTemp	outsideTemp	outsideHu...	windSpeed	windDirecti...	windGust	windGustDi...	rainRate
1341134700	72,5	61,286	89,3	0		0		0
1341135000	72,5	61,178	88,8	0		0		0
1341135300	72,5	61,322	88,8	0		0		0
1341135600	72,5	61,034	92	0		0		0
1341135900	72,5	61,052	90,6	0		0		0
1341136200	72,5	60,8	92,7	0,39,0		1,315,0		0
1341136500	72,5	61,034	91,1	0,290,0		1,90,0		0
1341136800	72,5	61,034	91,7	0		0		0
1341137100	72,5	60,836	91,1	0		0		0
1341137400	72,5	60,98	90	0		0		0
1341137700	72,5	61,034	91,5	0		0		0
1341138000	72,41	61,25	91,8	0,384,0		2,315,0		0
1341138300	72,32	61,304	93,3	0,196,0		2,338,0		0
1341138600	72,32	61,664	93,6	0,9110,0		3,135,0		0
1341138900	72,32	61,934	94,2	0,670,0		3,338,0		0
1341139200	72,32	62,204	91,7	0		0		0
1341139500	72,32	62,456	91	0		0		0
1341139800	72,302	62,762	90,5	0		0		0

Figura 41 Datos visualizados en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Para pasar la información de un proceso a otro, los íconos se interconectan a través del puntero sostenido de un punto a otro, como se muestra en la Figura 42. La etapa de pre procesamiento se usa a través de la herramienta selección, donde se escoge la variable objetivo (humedad del suelo) y los atributos con los que se desea trabajar. Particularmente para este caso se ha trabajado con todas las variables del conjunto de datos, incluyendo el tiempo de toma de las muestras.

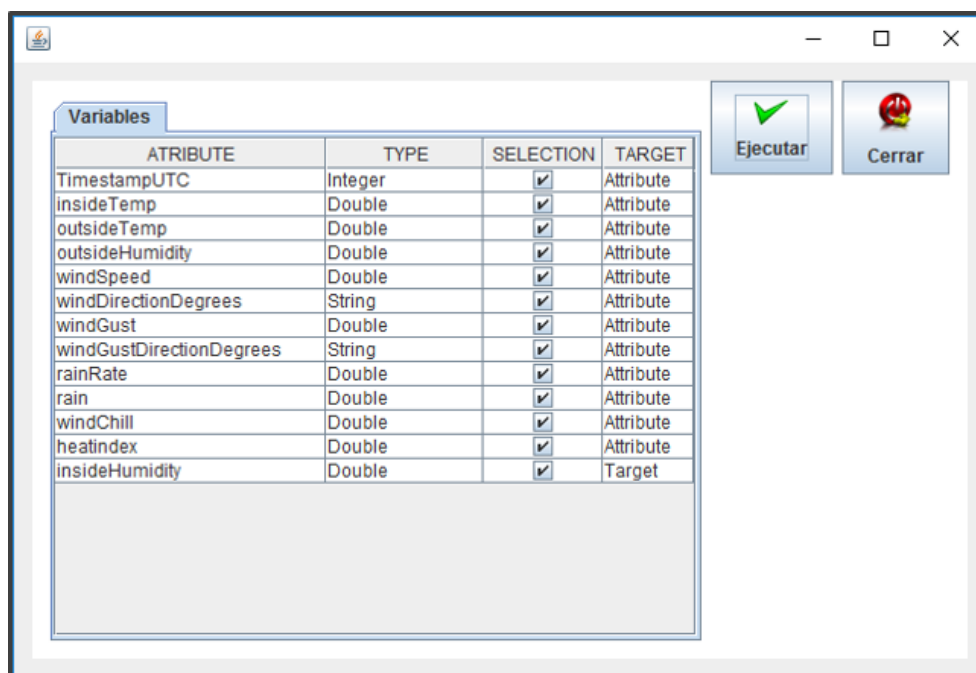


Figura 42 Selección de la variable objetivo en la prueba usando el Data Set for Sustainability.  
Fuente: Propia.

Luego de haber conectado la etapa de selección con la de clasificación, se procede a configurar los parámetros para el algoritmo C4.5. Como se indica en las figuras 43 y 44, se debe seleccionar el set de entrenamiento, las filas por nodo y el porcentaje de poda.

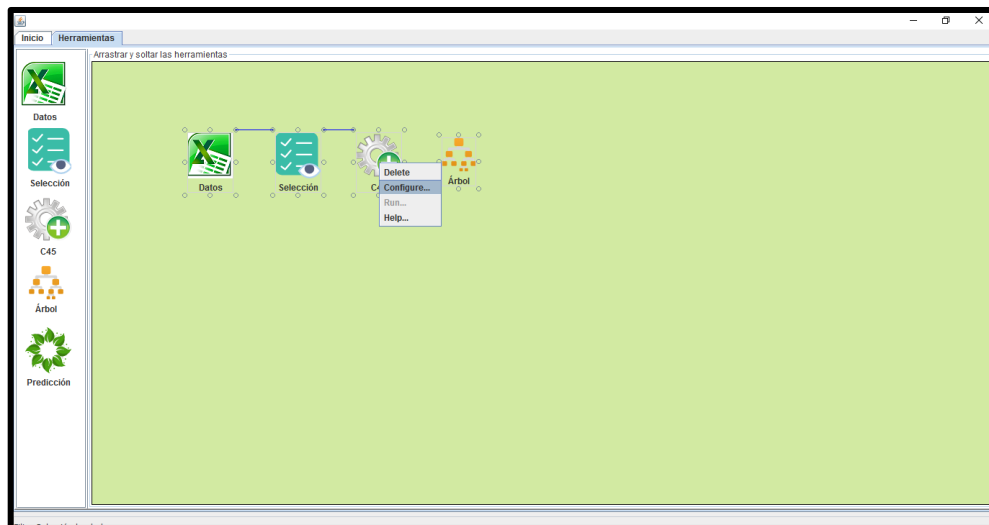


Figura 43 Conexión entre las etapas de selección y clasificación en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

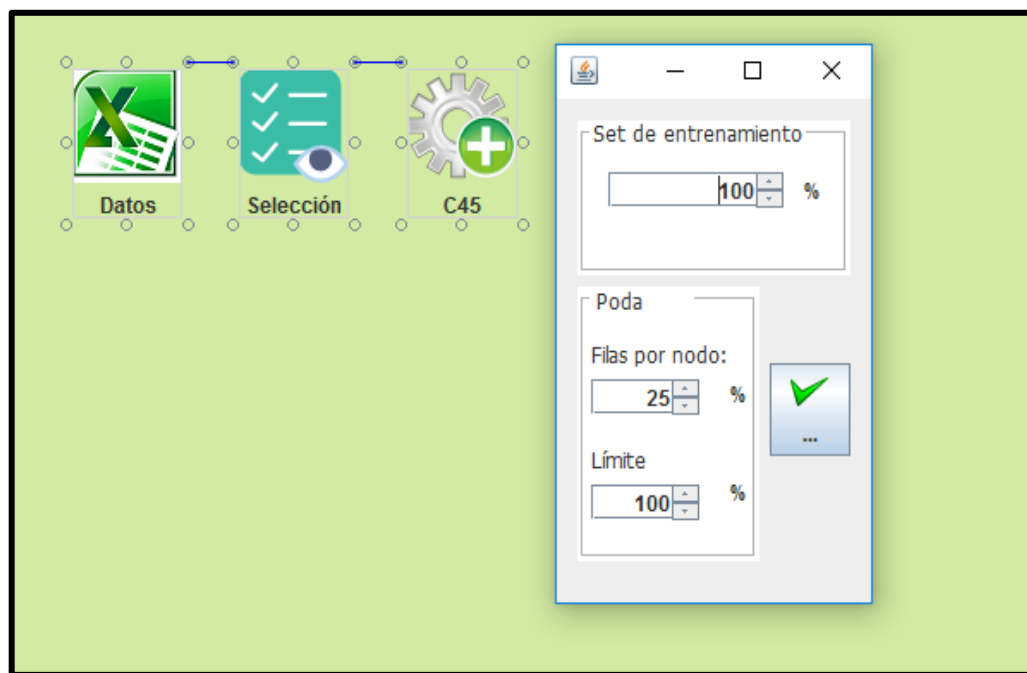


Figura 44 Configuración de los parámetros del algoritmo c4.5 en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

La última etapa en este caso de uso es la visualización, ésta se realiza mediante el gráfico del árbol de decisión. Para este se conectan los dos últimos objetos y se ejecuta al árbol. Las figuras 45, 46 y 47 permiten apreciar la conexión de los íconos y el árbol que se ha generado. Particularmente, este gráfico ha sido muy frondoso, es decir, que se han procesado muchas reglas de clasificación, dando como resultado que la variable objetivo (humedad del suelo), que para este caso es la humedad interna, depende de la variable tiempo.

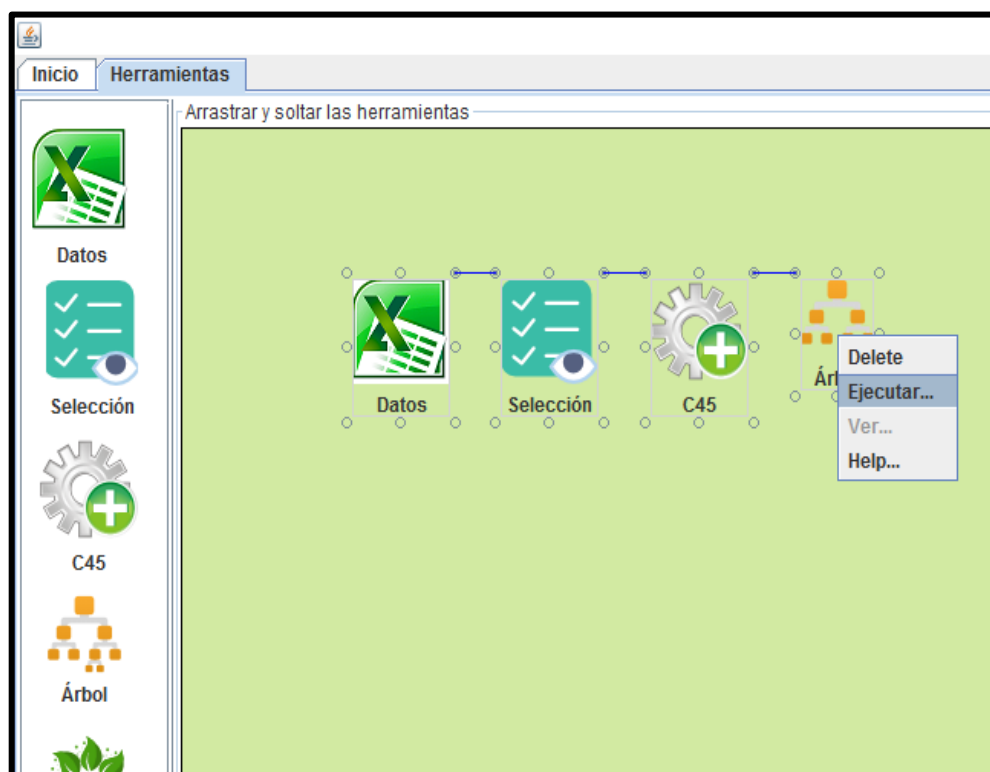


Figura 45 Conexión entre la etapa de clasificación y visualización en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Porcentaje de confianza : 100%

Visualización Rules

Set de reglas

#	Rules	Class	Confidence
1	TimestampUTC=1341134700	insideHumidity=48.299999 [1/1]	100%
2	TimestampUTC=1341135000	insideHumidity=48.0 [1/1]	100%
3	TimestampUTC=1341135300	insideHumidity=48.200001 [1/1]	100%
4	TimestampUTC=1341135600	insideHumidity=49.0 [1/1]	100%
5	TimestampUTC=1341135900	insideHumidity=48.299999 [1/1]	100%
6	TimestampUTC=1341136200	insideHumidity=48.900002 [1/1]	100%
7	TimestampUTC=1341136500	insideHumidity=48.200001 [1/1]	100%
8	TimestampUTC=1341136800	insideHumidity=49.0 [1/1]	100%
9	TimestampUTC=1341137100	insideHumidity=49.0 [1/1]	100%
10	TimestampUTC=1341137400	insideHumidity=49.0 [1/1]	100%
11	TimestampUTC=1341137700	insideHumidity=48.599998 [1/1]	100%
12	TimestampUTC=1341138000	insideHumidity=48.799999 [1/1]	100%
13	TimestampUTC=1341138300	insideHumidity=48.900002 [1/1]	100%
14	TimestampUTC=1341138600	insideHumidity=48.799999 [1/1]	100%
15	TimestampUTC=1341138900	insideHumidity=48.900002 [1/1]	100%
16	TimestampUTC=1341139200	insideHumidity=48.900002 [1/1]	100%
17	TimestampUTC=1341139500	insideHumidity=48.799999 [1/1]	100%
18	TimestampUTC=1341139800	insideHumidity=48.799999 [1/1]	100%
19	TimestampUTC=1341140100	insideHumidity=48.900002 [1/1]	100%
20	TimestampUTC=1341140400	insideHumidity=48.599998 [1/1]	100%

Guardar Reporte

Figura 46 Reglas de clasificación en la prueba usando el Data Set for Sustainability.  
Fuente: Propia.

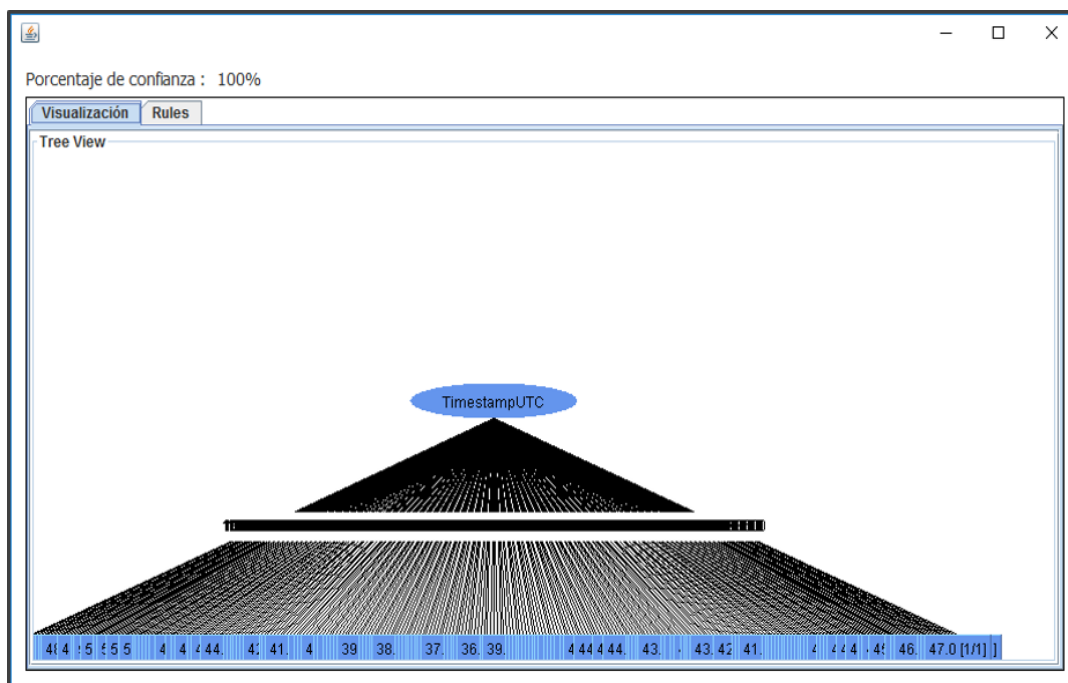


Figura 47 Visualización del árbol de decisión en la prueba usando el Data Set for Sustainability.  
Fuente: Propia.



#### 4.4.1.2. Caso 2: Predicción

La etapa de predicción parte del caso 1, para ésta se necesita un nuevo conjunto de datos y las reglas de clasificación generados anteriormente. En la Figura 48 se muestra los íconos necesarios para este proceso. Se debe cargar un nuevo conjunto de datos, donde no es necesario aplicar la regla de pre procesamiento, pues la variable objetivo ha sido seleccionada en el caso 1. La Figura 49 indica la carga del set de datos.

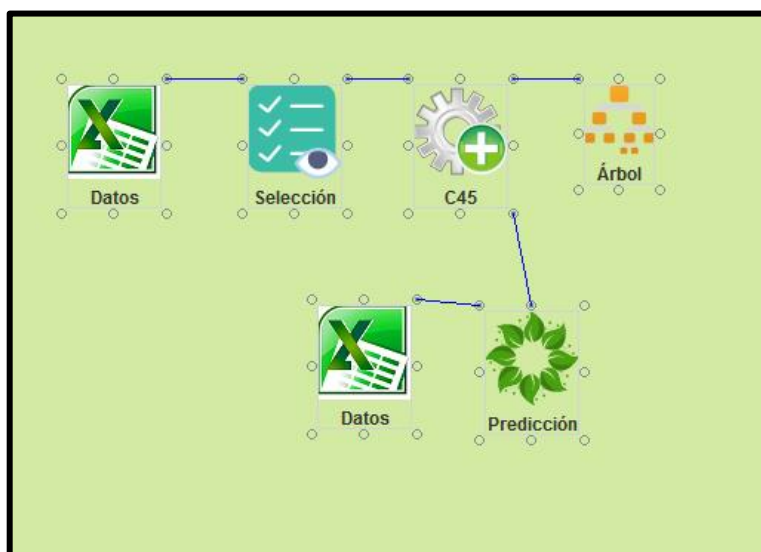
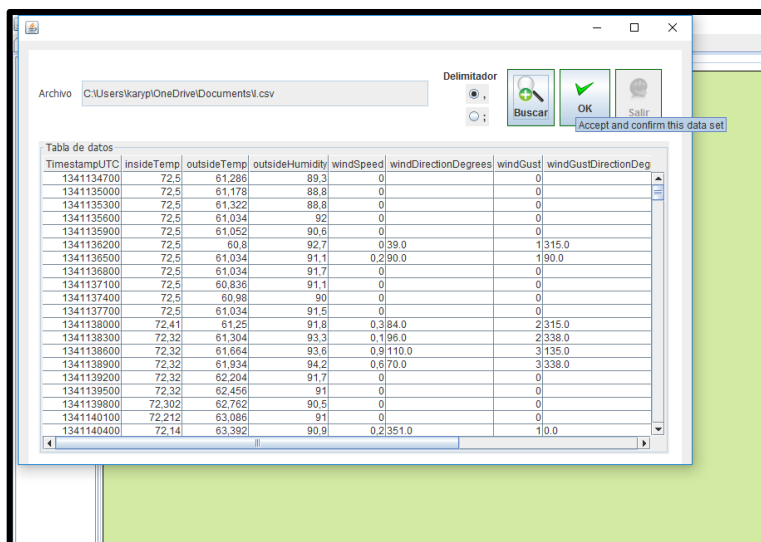


Figura 48 Íconos del proceso de predicción en la prueba usando el Data Set for Sustainability.  
Fuente: Propia.



TimestampUTC	insideTemp	outsideTemp	outsideHumidity	windSpeed	windDirectionDegrees	windGust	windGustDirectionDeg
1341134700	72.5	61.286	89.3	0		0	
1341135000	72.5	61.178	88.8	0		0	
1341135300	72.5	61.322	88.8	0		0	
1341135600	72.5	61.034	92	0		0	
1341135900	72.5	61.052	90.6	0		0	
1341136200	72.5	60.8	92.7	0.39.0		1.315.0	
1341136500	72.5	61.034	91.1	0.290.0		190.0	
1341136800	72.5	61.034	91.7	0		0	
1341137100	72.5	60.836	91.1	0		0	
1341137400	72.5	60.98	90	0		0	
1341137700	72.5	61.034	91.5	0		0	
1341138000	72.41	61.25	91.8	0.384.0		2.315.0	
1341138300	72.32	61.304	93.3	0.196.0		2.338.0	
1341138600	72.32	61.664	93.6	0.9110.0		3.135.0	
1341138900	72.32	61.934	94.2	0.670.0		3.338.0	
1341139200	72.32	62.204	91.7	0		0	
1341139500	72.32	62.456	91	0		0	
1341139800	72.302	62.762	90.5	0		0	
1341140100	72.212	63.086	91	0		0	
1341140400	72.14	63.392	90.9	0.2351.0		110.0	

Figura 49 Carga del nuevo set de datos en la prueba usando el Data Set for Sustainability.

Fuente: Propia.

Para que la etapa de predicción trabaje se conectan los íconos de datos (en este caso un conjunto nuevo) y las reglas generadas por c4.5. Y los pasos son dar clic a ejecutar y ver para que se genere el proceso, seguido de la opción ver y así generara el reporte con la predicción, como se aprecia en las figuras 50 y 51.

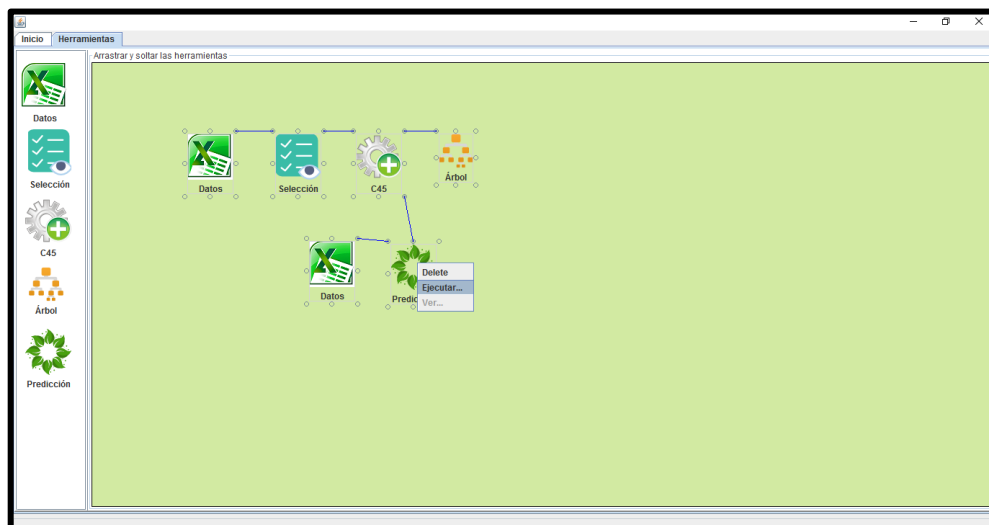


Figura 50 Ejecución de la etapa de predicción usando el Data Set for Sustainability.

Fuente: Propia.

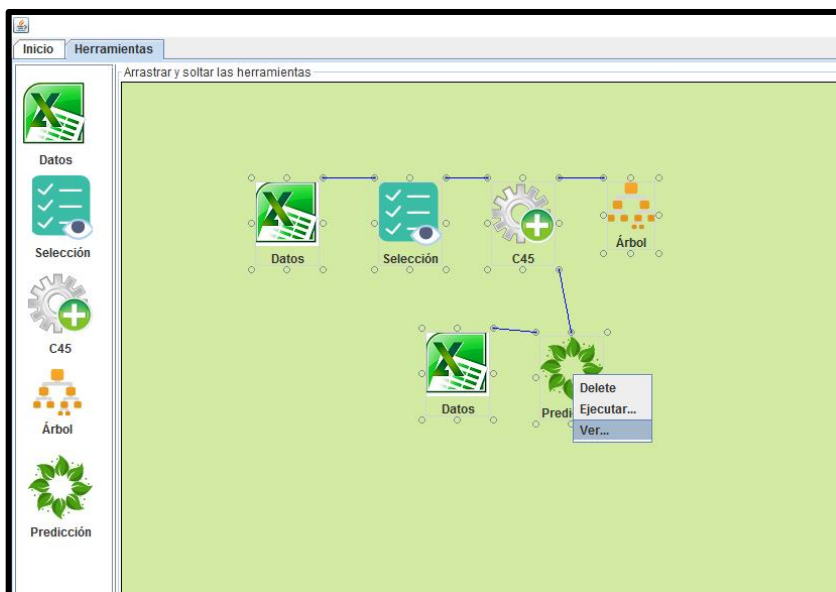


Figura 51 Ejecución de la etapa de predicción usando el Data Set for Sustainability.  
Fuente: Propia.

En la Figura 52 se puede observar los resultados de la etapa de predicción, el usuario puede acceder a los datos de entrada, variables usadas y la predicción realizada, mostrada en la última columna de este conjunto de datos. Con esta prueba se puede visualizar que la predicción con estas variables es notablemente acertada, es decir, tiene un porcentaje alto de confiabilidad.

La variable humedad del suelo que se ha predicho es muy cercana al valor real del conjunto de datos, la diferencia entre estos es de centésimas. Por lo tanto, se puede decir que los algoritmos de clasificación y predicción implementados si se ajustan a los datos utilizados. En la Figura 53, se muestra como guardar este reporte en un archivo nuevo de extensión .csv.

Time	insid...	outsid...	wind...	wind...	wind...	wind...	rain...	rain	wind...	heati...	insid...	insid...
1341...	72,5	61,2...	89,3	0		0		0	61,2...	61,2...	48,3	48,2...
1341...	72,5	61,1...	88,8	0		0		0	61,1...	61,1...	48	48,0
1341...	72,5	61,3...	88,8	0		0		0	61,3...	61,3...	48,2	48,2...
1341...	72,5	61,0...	92	0		0		0	61,0...	61,0...	49	49,0
1341...	72,5	61,0...	90,6	0		0		0	61,0...	61,0...	48,3	48,2...
1341...	72,5	60,8	92,7	0	39,0		1315,0	0	60,8	60,8	48,9	48,9...
1341...	72,5	61,0...	91,1	0,2	90,0		190,0	0	61,0...	61,0...	48,2	48,2...
1341...	72,5	61,0...	91,7	0		0		0	61,0...	61,0...	49	49,0
1341...	72,5	60,8...	91,1	0		0		0	60,8...	60,8...	49	49,0
1341...	72,5	60,98	90	0		0		0	60,98	60,98	49	49,0
1341...	72,5	61,0...	91,5	0		0		0	61,0...	61,0...	48,6	48,5...
1341...	72,41	61,25	91,8	0,3	84,0		2315,0	0	61,25	61,25	48,8	48,7...
1341...	72,32	61,3...	93,3	0,1	96,0		2338,0	0	61,3...	61,3...	48,9	48,9...
1341...	72,32	61,6...	93,6	0,9	110,0		3135,0	0	61,6...	61,6...	48,8	48,7...
1341...	72,32	61,9...	94,2	0,6	70,0		3338,0	0	61,9...	61,9...	48,9	48,9...
1341...	72,32	62,2...	91,7	0		0		0	62,2...	62,2...	48,9	48,9...
1341...	72,32	62,4...	91	0		0		0	62,4...	62,4...	48,8	48,7...
1341...	72,3...	62,7...	90,5	0		0		0	62,7...	62,7...	48,8	48,7...
1341...	72,2...	63,0...	91	0		0		0	63,0...	63,0...	48,9	48,9...
1341...	72,14	63,3...	90,9	0,2	351,0		10,0	0	63,3...	63,3...	48,6	48,5...
1341...	72,14	63,8...	91,2	0,1	54,0		045,0	0	63,8...	63,8...	48,4	48,4...
1341...	72,0...	64,4...	89,6	0		0		0	64,4...	64,4...	48	48,0
1341...	71,96	64,85	88,6	0		0		0	64,85	64,85	48	48,0
1341...	71,9...	65,3...	87,6	0		0		0	65,3...	65,3...	48	48,0
1341...	72,14	65,8...	87,4	0		0		0	65,8...	65,8...	48	48,0

Figura 52 Resultados del proceso de predicción usando el Data Set for Sustainability.

Fuente: Propia.

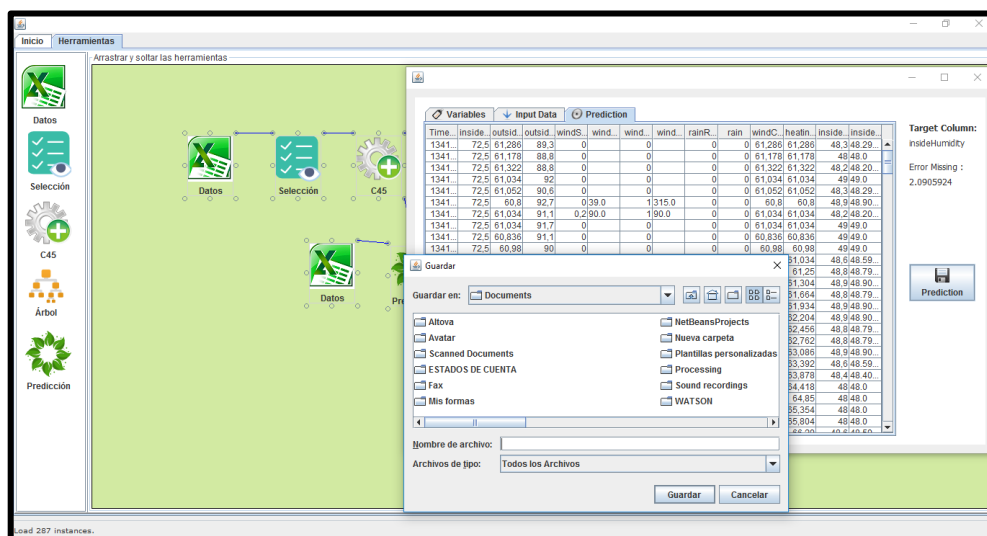


Figura 53 Almacenamiento de los resultados usando el Data Set for Sustainability.

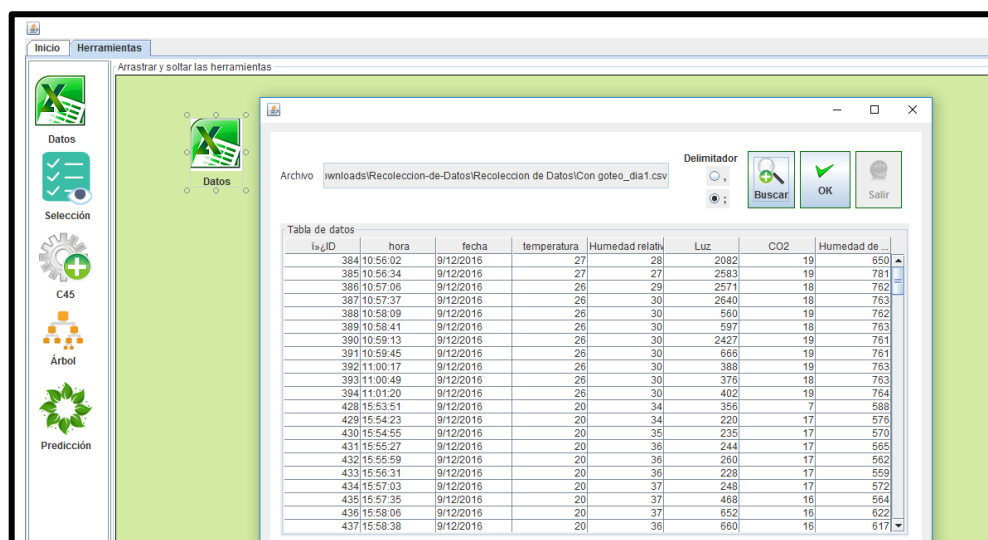
Fuente: Propia.

#### 4.4.2. Pruebas usando los datos reales del invernadero

Con los datos obtenidos a través de un WSN en el invernadero de la granja la pradera, se realizaron pruebas para comprobar que las técnicas usadas funcionan y se acoplan a los datos reales, usando los dos algoritmos: Clasificación y predicción.

##### 4.4.2.1. Caso 1: Clasificación

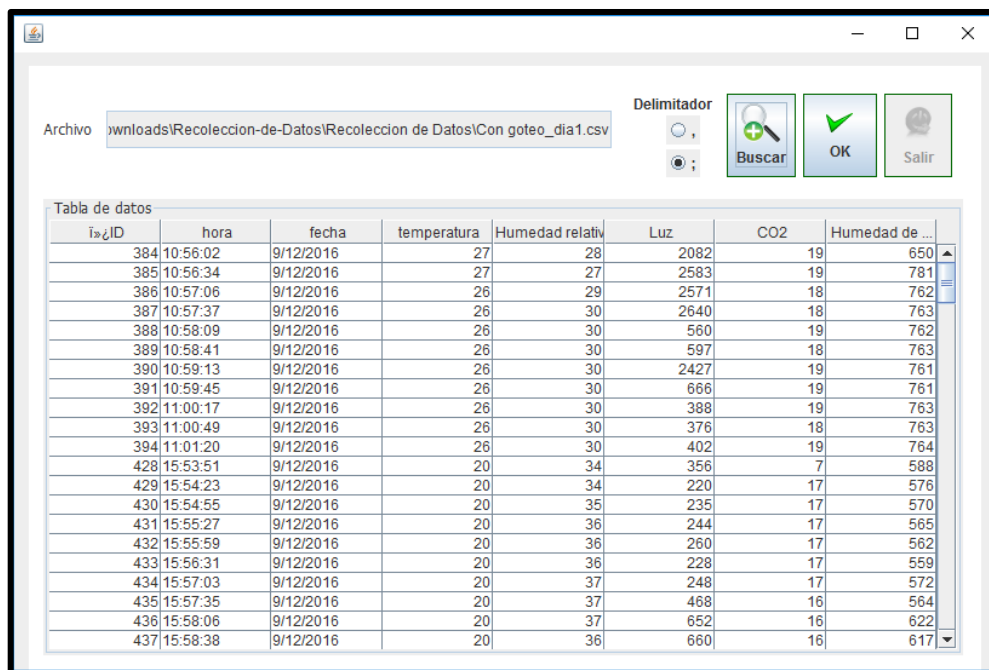
Al igual que en las pruebas anteriores, se ha cargado el archivo .csv que contiene los datos a analizar, en este caso el conjunto de datos posee las 5 variables necesarias, además de variables de control como: ID, hora, fecha, como se puede apreciar en las figuras 54 y 55. Estas pueden ser eliminadas en el proceso de selección, y se trabaja solo con las variables ambientales del invernadero.



Id	hora	fecha	temperatura	Humedad relati	Luz	CO2	Humedad de
384	10:56:02	9/12/2016	27	28	2082	19	650
385	10:56:34	9/12/2016	27	27	2583	19	781
386	10:57:06	9/12/2016	26	29	2571	18	762
387	10:57:37	9/12/2016	26	30	2640	18	763
388	10:58:09	9/12/2016	26	30	560	19	762
389	10:58:41	9/12/2016	26	30	597	18	763
390	10:59:13	9/12/2016	26	30	2427	19	761
391	10:59:45	9/12/2016	26	30	666	19	761
392	11:00:17	9/12/2016	26	30	388	19	763
393	11:00:49	9/12/2016	26	30	376	18	763
394	11:01:20	9/12/2016	26	30	402	19	764
428	15:53:51	9/12/2016	20	34	356	7	588
428	15:54:23	9/12/2016	20	34	220	17	576
430	15:54:55	9/12/2016	20	35	235	17	570
431	15:55:27	9/12/2016	20	36	244	17	565
432	15:55:59	9/12/2016	20	36	260	17	562
433	15:56:31	9/12/2016	20	36	228	17	569
434	15:57:03	9/12/2016	20	37	248	17	572
435	15:57:35	9/12/2016	20	37	468	16	564
436	15:58:06	9/12/2016	20	37	652	16	622
437	15:58:38	9/12/2016	20	36	660	16	617

Figura 54 Carga de datos obtenidos del invernadero, prueba con los datos reales.

Fuente: Propia.



Archivo: \\nloads\Recoleccion-de-Datos\Recoleccion de Datos\Con goteo\_dia1.csv

Delimitador: ;

Tabla de datos

ÍtemID	hora	fecha	temperatura	Humedad relativ	Luz	CO2	Humedad de ...
384	10:56:02	9/12/2016	27	28	2082	19	650
385	10:56:34	9/12/2016	27	27	2583	19	781
386	10:57:06	9/12/2016	26	29	2571	18	762
387	10:57:37	9/12/2016	26	30	2640	18	763
388	10:58:09	9/12/2016	26	30	560	19	762
389	10:58:41	9/12/2016	26	30	597	18	763
390	10:59:13	9/12/2016	26	30	2427	19	761
391	10:59:45	9/12/2016	26	30	666	19	761
392	11:00:17	9/12/2016	26	30	388	19	763
393	11:00:49	9/12/2016	26	30	376	18	763
394	11:01:20	9/12/2016	26	30	402	19	764
428	15:53:51	9/12/2016	20	34	356	7	588
429	15:54:23	9/12/2016	20	34	220	17	576
430	15:54:55	9/12/2016	20	35	235	17	570
431	15:55:27	9/12/2016	20	36	244	17	565
432	15:55:59	9/12/2016	20	36	260	17	562
433	15:56:31	9/12/2016	20	36	228	17	559
434	15:57:03	9/12/2016	20	37	248	17	572
435	15:57:35	9/12/2016	20	37	468	16	564
436	15:58:06	9/12/2016	20	37	652	16	622
437	15:58:38	9/12/2016	20	36	660	16	617

Figura 55 Datos obtenidos del invernadero, prueba con los datos reales.  
Fuente: Propia.

Como se ha mencionado anteriormente, en el proceso de selección la variable a escoger es la humedad del suelo, es decir, la variable objetivo, esto se puede apreciar en las figuras 56 y 57.

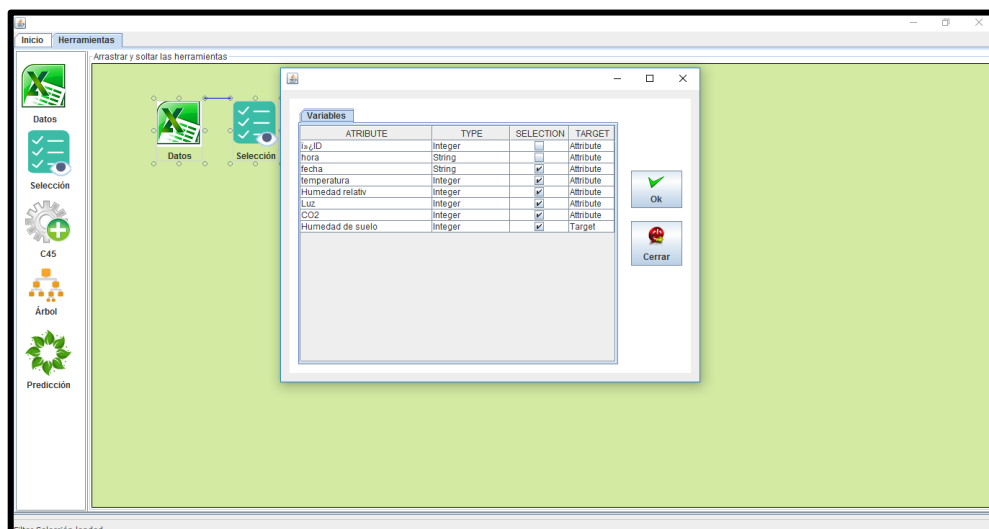


Figura 56 Selección de la variable objetivo (humedad del suelo), prueba con los datos reales.  
Fuente: Propia.

Las variables de identificación son omitidas en este proceso, se seleccionan aquellas que son de valor para el análisis.

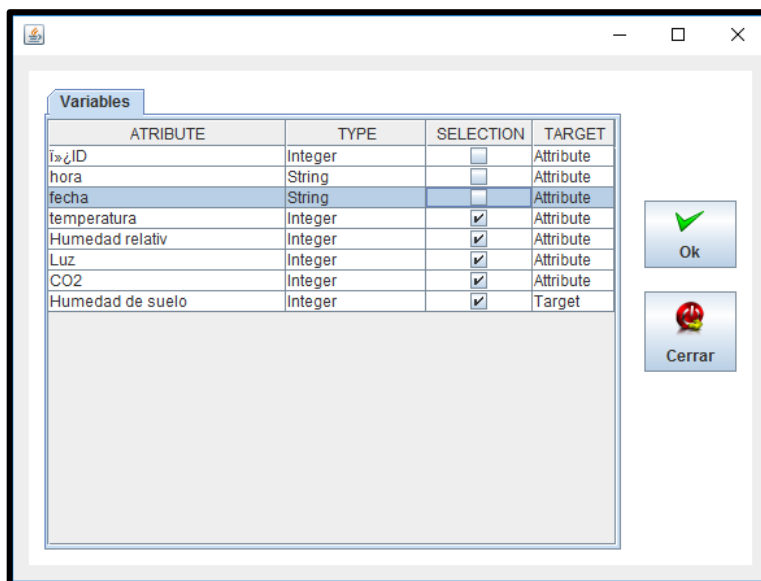


Figura 57 Selección de la variable objetivo (humedad del suelo), prueba con los datos reales.  
Fuente: Propia.

A continuación, en el proceso de clasificación se trabaja con los valores por defecto, es decir, los que ya vienen asignados. En las figuras 58 y 59 se aprecia la configuración de estos parámetros.

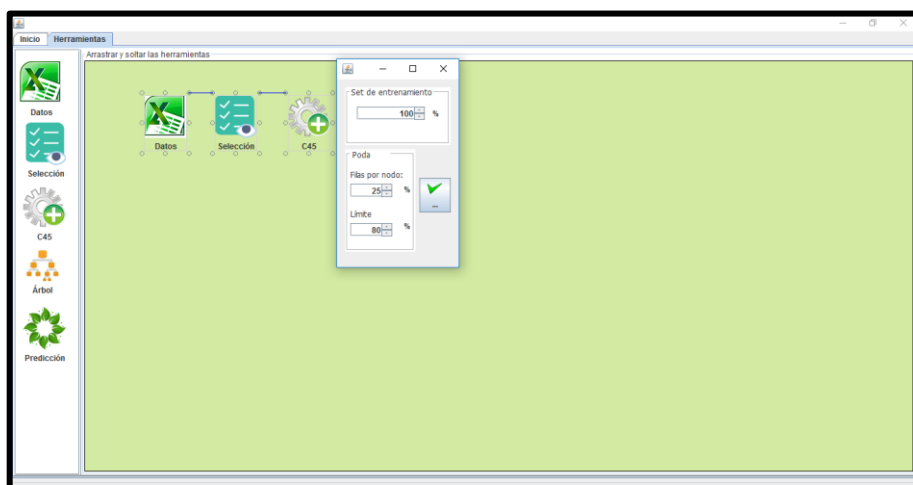


Figura 58 Proceso de clasificación, prueba con los datos reales.  
Fuente: Propia.

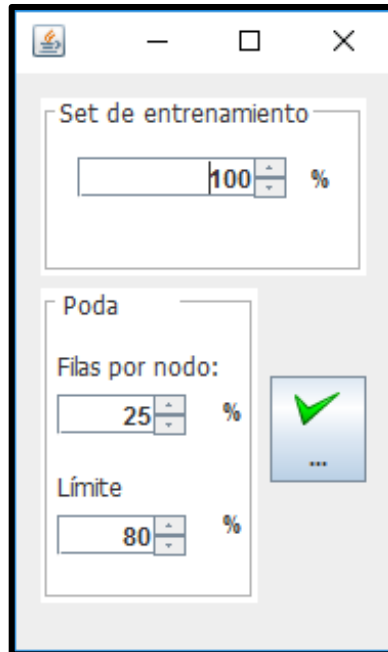


Figura 59 Proceso de clasificación, prueba con los datos reales.  
Fuente: Propia.

Por último, se tiene a la visualización, dónde se observan las reglas de clasificación y el árbol de decisión generados. En la Figura 60 se muestra las reglas que han sido generadas, donde se puede observar que la humedad del suelo se encuentra estrechamente relacionada con la cantidad de iluminación, y cuando este valor es 0, interviene la humedad relativa. En la figura 61 se aprecia el árbol de decisión que se ha generado, donde se puede visualizar que la variable objetivo humedad del suelo depende directamente del comportamiento de la variable luz (cantidad de iluminación).



Confidence Tree : 77,957%

Weka Tree Rules

Rules set

#	Rules	Class	Confidence
28	Luz=356	Humedad de suelo=588 [1/1]	100%
29	Luz=220	Humedad de suelo=576 [1/1]	100%
30	Luz=235	Humedad de suelo=570 [1/1]	100%
31	Luz=244	Humedad de suelo=565 [1/1]	100%
32	Luz=260	Humedad de suelo=562 [1/1]	100%
33	Luz=228	Humedad de suelo=559 [1/1]	100%
34	Luz=248	Humedad de suelo=572 [1/1]	100%
35	Luz=468	Humedad de suelo=564 [1/1]	100%
36	Luz=652	Humedad de suelo=622 [1/1]	100%
37	Luz=660	Humedad de suelo=617 [1/1]	100%
38	Luz=673	Humedad de suelo=611 [1/1]	100%
39	Luz=1	Humedad de suelo=566 [1/1]	100%
40	Luz=0 and Humedad relativ=58	Humedad de suelo=481 [1/1]	100%
41	Luz=2058	Humedad de suelo=587 [1/1]	100%
42	Luz=2133	Humedad de suelo=584 [1/1]	100%
43	Luz=2188	Humedad de suelo=580 [1/1]	100%
44	Luz=2399	Humedad de suelo=577 [1/1]	100%
45	Luz=2588	Humedad de suelo=575 [1/1]	100%
46	Luz=2711	Humedad de suelo=574 [1/1]	100%
47	Luz=2892	Humedad de suelo=571 [1/1]	100%
48	Luz=4093	Humedad de suelo=566 [1/1]	100%
49	Luz=864	Humedad de suelo=560 [1/1]	100%
50	Luz=1400	Humedad de suelo=562 [1/1]	100%
51	Luz=2266	Humedad de suelo=561 [1/1]	100%
52	Luz=2103	Humedad de suelo=561 [1/1]	100%
53	Luz=1330	Humedad de suelo=558 [1/1]	100%
54	Luz=1470	Humedad de suelo=559 [1/1]	100%
55	Luz=1249	Humedad de suelo=560 [1/1]	100%
56	Luz=911	Humedad de suelo=606 [1/1]	100%
57	Luz=729	Humedad de suelo=607 [1/1]	100%
58	Luz=725	Humedad de suelo=605 [1/1]	100%
59	Luz=3207	Humedad de suelo=525 [1/1]	100%
60	Luz=3436	Humedad de suelo=515 [1/1]	100%

Save Report

Figura 60 Reglas de clasificación, prueba con los datos reales.

Fuente: Propia.

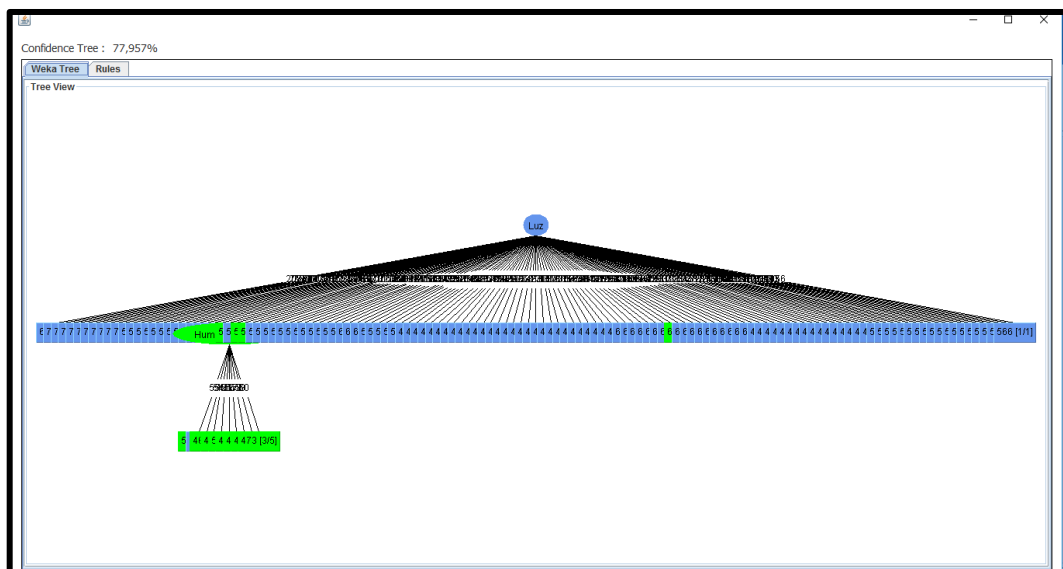


Figura 61 Árbol de decisión, prueba con los datos reales.  
Fuente: Propia.

En el árbol de la Figura 62, se puede observar las ramificaciones y la variable luz, es la que mayor peso tiene, es por esto que es el nodo central. De color verde se ve que otro nodo importante viene a ser la humedad relativa, cuando la luz no posee valoración es la variable humedad relativa la que infiere en el comportamiento de la humedad del suelo.

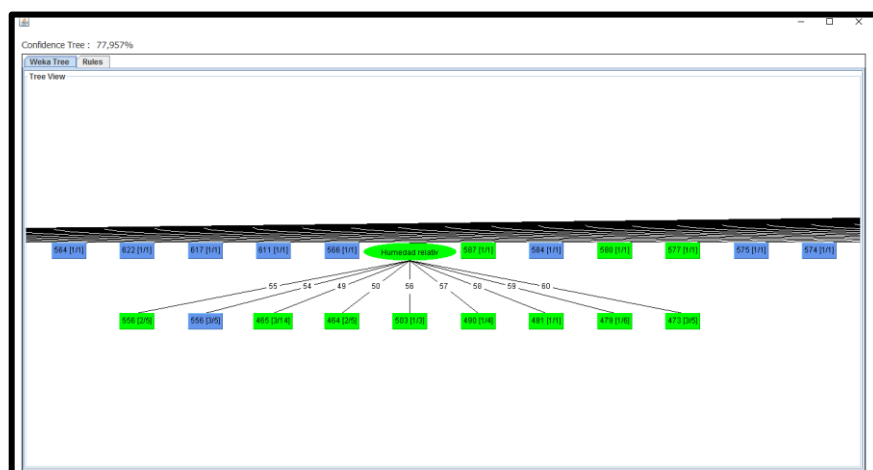


Figura 62 Árbol de decisión, prueba con los datos reales.  
Fuente: Propia.

Para obtener una mejor visualización del árbol de decisión es recomendable discretizar los valores, de manera que se pueda observar hacia donde tiende la variable objetivo, particularmente para este caso es la humedad del suelo. Al realizar este procedimiento la confianza de la clasificación disminuye en baja proporción, dando reglas confiables para la predicción. Esto se puede apreciar en la Figura 63.

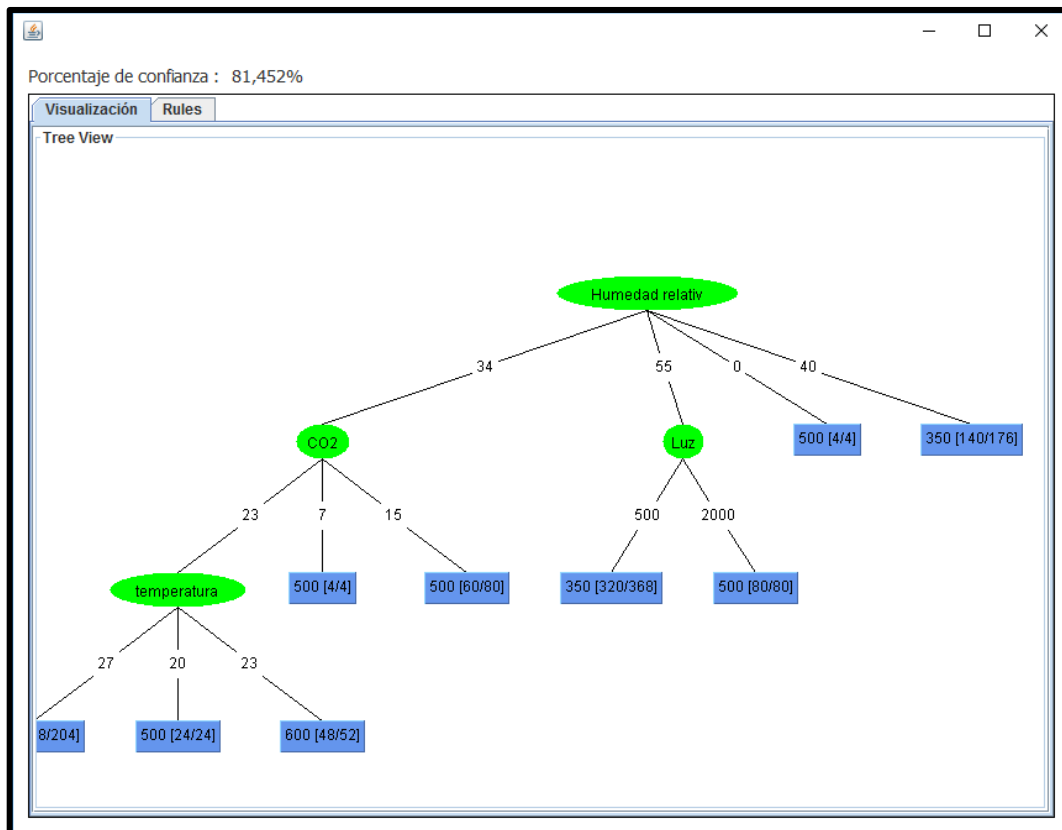


Figura 63 Árbol de decisión con los datos discretizados, prueba con los datos reales.  
Fuente: Propia.

#### 4.4.2.2. Caso 2: Predicción

Con los resultados obtenidos anteriormente se realiza la predicción de la variable objetivo que particularmente para este caso de estudio es la humedad del suelo. Para este proceso se debe cargar el nuevo conjunto de datos en el que se quiere predecir dicha variable. La figura 64 se muestra el

esquema de conexión para realizar este proceso, donde se puede visualizar que para obtener los resultados del algoritmo de predicción se necesita las reglas de clasificación, generadas por el algoritmo C4.5 y del nuevo conjunto de datos.

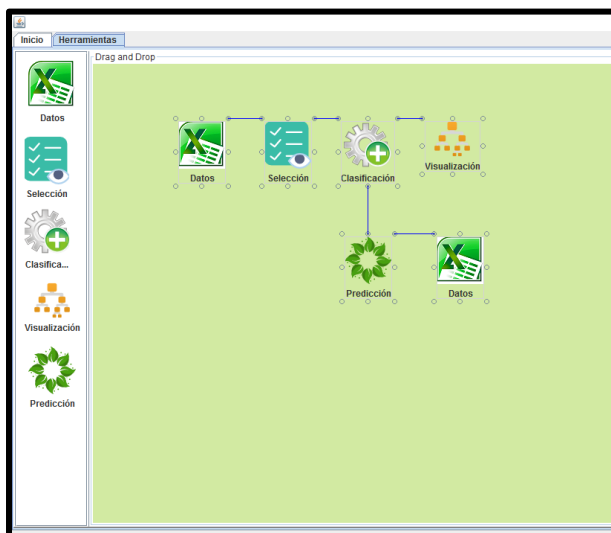


Figura 64 Íconos ubicados en el canvas para la predicción de datos, prueba con los datos reales.

Fuente: Propia.

En primera instancia se realizó pruebas con archivos que contienen las 5 variables recolectadas: Temperatura, humedad relativa, luz, CO<sub>2</sub>, y humedad del suelo. En la figura 65 se muestra los datos que se han cargado para la predicción de la variable objetivo (humedad del suelo), en este archivo consta la variable mencionada anteriormente, y de esta manera se puede verificar que el modelo predictivo funciona.

Data File  
C:\Users\karyp\OneDrive\Documents\Datooooos\cont 1.1.csv

Delimiter:  
 ,  
 ;

Play

Exit

temperatura	Humedad relativ	Luz	CO2	Humedad de s...
27	28	2082	19	650
27	27	2583	19	781
26	29	2571	18	762
26	30	2640	18	763
26	30	560	19	762
26	30	597	18	763
26	30	2427	19	761
26	30	666	19	761
26	30	388	19	763
26	30	376	18	763
26	30	402	19	764
20	34	356	7	588
20	34	220	17	576
20	35	235	17	570
20	36	244	17	565
20	36	260	17	562
20	36	228	17	559
20	37	248	17	572
20	37	468	16	564

Figura 65 Nuevo conjunto de datos cargado para la predicción, con la variable Humedad del Suelo, prueba con los datos reales.

Fuente: Propia.

El algoritmo de predicción se realiza en base a las reglas de clasificación calculadas por el sistema y con el nuevo conjunto de datos. La Figura 66 muestra la predicción del sistema donde se puede constatar que los valores de la variable Humedad del suelo son bastante cercanos, demostrando así que el modelo predictivo si se ajusta a los datos utilizados.

Cuando el nivel de iluminación (luz) se encuentra en valores relativamente bajos o nulos, las reglas de clasificación no poseen un nivel de confianza alto, por ende, no se calcula la predicción de manera adecuada, dando como resultado datos erróneos. Pero estos casos son pocos, debido a que la variable iluminación tiende a valores bajos cuando llega la noche, durante el día esta da valores que sirven para la creación de reglas de clasificación con alto nivel de confianza.

temperatura	Humedad relativ	Luz	CO2	Humedad de suelo	Humedad de suelo
27	26	2082	19	650	650
27	27	2533	19	781	781
28	29	2571	18	762	762
28	30	2640	18	763	763
26	30	560	19	762	762
26	30	597	18	763	763
26	30	2427	19	761	761
26	30	666	19	761	761
26	30	388	19	763	763
26	30	376	18	763	763
26	30	402	19	764	764
20	34	356	7	588	588
20	34	220	17	576	576
20	35	235	17	570	570
20	36	244	17	565	565
20	36	260	17	562	562
20	36	228	17	559	559
20	37	248	17	572	572
20	37	468	16	564	564
20	37	652	16	622	622
20	36	660	16	617	617
20	37	673	16	611	611
18	55	1	18	566	566
19	55	0	18	561	566
19	55	0	18	559	566
19	55	0	18	557	566
19	55	0	18	556	566
19	55	0	18	556	566
19	54	0	18	556	566
19	54	0	18	556	566
19	54	0	18	556	566

Figura 66 Visualización de los resultados de predicción, con la variable Humedad del suelo, prueba con los datos reales.

Fuente: Propia.

La interfaz desarrollada puede predecir la variable objetivo, es por esto que se ha realizado pruebas con archivos que poseen 4 variables, es decir, se omite a la humedad del suelo para que se realice la predicción. Las Figuras 67, 68, y 69 muestran el proceso para la predicción de la humedad del suelo, sin necesidad de que dicha variable se encuentre en el archivo del nuevo conjunto de datos.

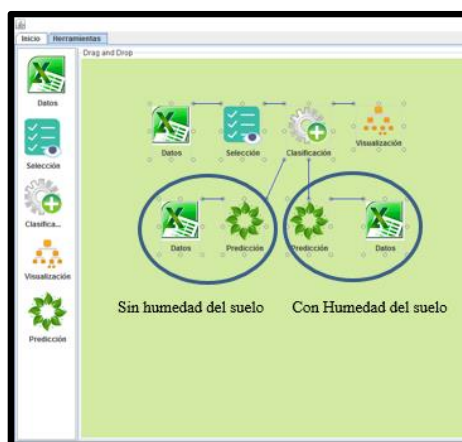


Figura 67 Íconos ubicados en el canvas para la predicción de datos sin la variable Humedad del Suelo, prueba con los datos reales.

Fuente: Propia.

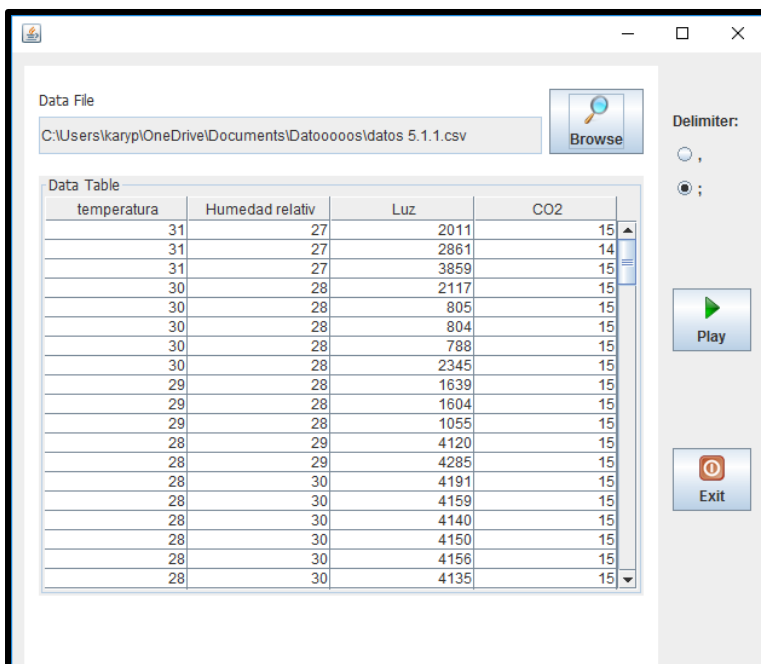


Figura 68 Carga del nuevo conjunto de datos sin la variable Humedad del suelo, prueba con los datos reales.

Fuente: Propia.

temperatura	Humedad relativ	Luz	CO2	Humedad de suelo
31	27	2011	15	15,459
31	27	2861	14	14,459
31	27	3859	15	15,459
30	28	2117	15	15,459
30	28	805	15	15,647
30	28	804	15	15,459
30	28	788	15	15,459
30	28	2345	15	15,459
29	28	1639	15	15,459
29	28	1604	15	15,459
29	28	1055	15	15,459
28	29	4120	15	15,459
28	29	4285	15	15,459
28	30	4191	15	15,459
28	30	4159	15	15,459
28	30	4140	15	15,459
28	30	4150	15	15,459
28	30	4156	15	15,459
28	30	4135	15	15,459
28	30	4125	15	15,459
28	31	4097	15	15,459
29	30	4063	15	15,459
29	31	4043	15	15,459
29	30	4018	15	15,459
29	30	4009	15	15,459
29	30	3977	15	15,459
29	30	3969	15	15,459
29	30	3926	15	15,459
29	30	3884	15	15,459
29	30	3898	15	15,459
29	30	3884	15	15,459

Figura 69 Visualización de resultados de predicción de la variable Humedad del suelo, prueba con los datos reales.

Fuente: Propia.

#### 4.4.3. Resultados

Se realizaron varias pruebas de funcionamiento de la interfaz de análisis de datos para invernaderos. De esta manera, se comprobó que las técnicas y algoritmos utilizados si se ajustan a los datos de interés. La Tabla 10 contiene los resultados obtenidos de las pruebas realizadas.

Tabla 10 Resultados de las pruebas con los diferentes archivos de datos

<b>Prueba</b>	<b>Resultados</b>
<b>Data Set for Sustainability Clasificación</b>	En este caso, se observó que las reglas de clasificación apuntaban a la variable TimeStamp como la de mayor peso, de esta dependía directamente la humedad.
<b>Data Set for Sustainability Predicción</b>	Al realizar el algoritmo de predicción, se comprobó que el modelo si funciona, pues la variable Humedad del Suelo fue predicha con un alto grado de confianza.
<b>Datos reales Clasificación</b>	Al poseer datos reales el algoritmo determinó que la humedad del suelo depende directamente de la variable iluminación, cuando esta no posee valor, la segunda con mayor peso es la humedad relativa. Por ende, en base a estos dos factores se forman las reglas de clasificación



---

<b>Datos reales Predicción</b>	La variable Humedad del Suelo fue predicha con un alto grado de confianza, ya que los datos obtenidos por el proceso de predicción se acercan a los reales, con alta precisión.
--------------------------------	---

---

Fuente: Propia.

## CAPÍTULO V

### 5.1. CONCLUSIONES

- Después de realizar una investigación acerca de los factores que inciden en el crecimiento de los cultivos dentro de un invernadero, se determinó que los factores más importantes son: Humedad del suelo, humedad relativa, temperatura ambiental, nivel de iluminación y CO<sub>2</sub>, dado que éstos influyen directamente en el proceso de fotosíntesis de las plantas y, estableciendo valores correctos, logran una mejor captación de nutrientes y obtienen mejores frutos.
- Las herramientas y técnicas de *Big Data* y, específicamente, de *data mining* son fundamentales para realizar procesos de analítica de datos. Si bien estas dos áreas surgieron como soporte para la toma de decisiones en economía y negocios, hoy en día son de uso transversal en diversos escenarios y se enfocan al descubrimiento de patrones dentro de una montaña de datos. A través de una revisión de la documentación empleada para esta investigación, se pudo encontrar que estas herramientas y técnicas de analítica de datos no se rigen, particularmente, por un estándar, sino que representan un conjunto de algoritmos que permiten realizar modelos descriptivos sobre un conjunto de datos con el fin de clasificar y/o predecir información.
- En las primeras etapas del desarrollo de este proyecto, se realizó una búsqueda de bases de datos que contengan diversas mediciones representando variables capturadas en un invernadero real. En este sentido, se determinó usar la base de datos *Environmental data (indoor and outdoor)* del repositorio *UMass Trace Repository*, debido a que se ajustó adecuadamente a los requerimientos del proyecto, es decir, que contiene un conjunto de muestras suficiente y posee las variables con los factores más importantes.

- Como resultado significativo de este proyecto, se encontró que el diseño de un software de *data mining* con una interfaz de uso intuitivo en un *framework* de *drag and drop* es una alternativa adecuada para procesar variables de invernadero. Se comprobó que puede realizarse el procesamiento de los datos conformando de forma secuencial las etapas del proceso de descubrimiento de conocimiento en base de datos (KDD) a través de la unión de objetos que representan módulos de programación. Específicamente, el entorno de desarrollo NetBean IDE 8.2, que trabaja con el lenguaje de programación java, comprobó ser un software adecuado y de precisión para el desarrollo de las técnicas y algoritmos de la analítica de datos, y que también permitió implementar una interfaz amigable con el usuario.
- Después del desarrollo de un software para el análisis de los parámetros ambientales que inciden en el crecimiento de cultivos en invernaderos y con el fin de comprobar su correcto funcionamiento, se realizó diversas pruebas con diferentes fuentes de información, siendo una de ellas la encontrada en el *UMass Trace Repository*, dando como resultado que la humedad interior se puede predecir con base en las demás variables con una precisión, significativamente buena. Particularmente, se logró determinar que dicha variable se encuentra ligada a la variable *TimeStampUTM*, la cual se refiere al tiempo en que ha sido tomada la muestra. Las pruebas comprobaron la usabilidad y confiabilidad del software.
- Para comprobar el funcionamiento del sistema con datos reales, se realizó pruebas con información obtenida del invernadero de la granja “La pradera”, a través de una red de sensores inalámbricos instalada en el lugar. Al igual que en los experimentos con los datos de prueba, se encontró nuevamente que la humedad del suelo depende del tiempo de toma de la muestra. Adicionalmente, usando la herramienta de selección se determinó las cinco

variables más importantes y, en ese caso, se obtuvo como resultado que el nivel de iluminación (denominado luz) es el factor más importante del cual depende la humedad del suelo. Además, con base en este factor se calcularon las reglas de clasificación, de forma que cuando éste tiene un valor de 0, la variable que considera el sistema es la humedad relativa, y se comprobó experimentalmente que así se genera mayor conocimiento a través de la exploración de datos.

- Las áreas de *Big Data* y *Data Mining* son relativamente emergentes y se encuentran en constante desarrollo, y, particularmente, su aplicación en el sector agrícola es un tema amplio y diverso que busca, entre otros aspectos, optimizar los recursos. En efecto, la agricultura de precisión está enfocada al uso de herramientas tecnológicas para hacer eficientes en el uso y administración de recursos. Dicho esto, las técnicas de minería de datos representan una buena alternativa para explorar la información de las variables relacionadas con agricultura y soportar la toma de decisiones inteligentes.

## 5.2. RECOMENDACIONES

- Al iniciar un proceso de analítica de datos, usando las técnicas, algoritmos y herramientas de *Big Data* y *Data Mining*, se debe investigar el funcionamiento de éstas y así poder tener un conocimiento amplio de la temática. Al realizar la búsqueda de las variables que interfieren en el crecimiento y correcto desarrollo de los cultivos en invernadero, se debe definir de una manera adecuada aquellas que aporten significativamente información para el respectivo análisis.
- Es esencial realizar una investigación adecuada de las técnicas y herramientas, más no, la búsqueda de estándares a los que se rijan *Big Data* o *Data Mining*. Este proyecto se encuentra enfocado al análisis de datos en cultivos de invernadero, pues al ser un desarrollo en software, no se ha implementado aún un estándar que aplicar al realizar un proyecto de este tipo.
- Los datos son de suma importancia para desarrollar este proyecto, es por esto que es aconsejable buscar repositorios con bases de datos, que posean características similares a la que se obtendrían directamente en el área de aplicación. Esto se debe realizar, pues en las fases de la implementación existen pruebas para ir verificando que el modelo de predicción de datos si se ajusta a la información usada.
- Para la implementación de un proyecto con analítica de datos, las pruebas de funcionamiento es uno de los pasos esenciales para comprobar el correcto funcionamiento del sistema. Es por esto, que se deben realizar varias pruebas y de esta manera, ir verificando que los algoritmos desarrollados tienen un funcionamiento adecuado.

### 5.3. GLOSARIO DE TÉRMINOS

**AGRICULTURA DE PRECISIÓN:** Se define como la gestión agrícola mediante el uso de herramientas tecnológicas con el objetivo de optimizar recursos.

**BIG DATA:** Llamado también datos masivos, conjuntos de datos tan grandes que no es posible procesarlos con los sistemas convencionales de bases de datos.

**MINERÍA DE DATOS:** Conjunto de técnicas usadas para el análisis de conjuntos de datos y encontrar relaciones, patrones que sean entendibles y útiles para el usuario.

**KDD:** Acrónimo en inglés de Knowledge Discovery Data Bases, es el descubrimiento de conocimiento en bases de datos, es decir, es un proceso para encontrar patrones entre los datos que sean de utilidad para el usuario.

**OPEN DATA:** Es todo conjunto de datos que ha sido distribuido libremente a través de la red, sin restricciones de uso, ni derechos de autor.

**LIMPIEZA DE DATOS:** Conjunto de procesos que permiten preparar los datos antes de ingresar a los algoritmos, es decir, eliminar datos erróneos o corruptos, llenar los faltantes, o seleccionar con los que se desee trabajar, con el fin de que se encuentren óptimos para aplicar las diversas técnicas de analítica de datos.

**BODEGAS DE DATOS:** Data Warehouse, es un conjunto de bases de datos no relacionales alojadas en servidores que permiten almacenar cantidades enormes de información, independientemente de su origen.

**C4.5:** Es un algoritmo de clasificación que genera un árbol de decisión basado en la entropía del sistema, es decir, en el desgaste del mismo.

## REFERENCIAS

- University of Ljubljana. (2016). *Orange*. Retrieved from <http://orange.biolab.si/>
- Akyildiz, I. F., & Vuran, M. C. (2010). *Wireless sensor networks (Vol. 4)*. . John Wiley & Sons.
- Alpi, A., & Tognoni, F. (1999). *Cultivo en invernadero*. Madrid: Ediciones Mundi-Prensa.
- Anegón, F. d., Herrero Solana, V., & Guerrero Bote, V. (1998). La aplicación de Redes Neuronales Artificiales (RNA): a la recuperación de la información.
- Apache Spark. (2016). *Apache spark*. Retrieved from <http://spark.apache.org/>
- Asthor. (2016). *Asthor*. Retrieved from [http://asthor.com/?page\\_id=183](http://asthor.com/?page_id=183)
- Barrios, O. (2004). *Construcción de un Invernadero*. Santiago.
- Brachman, R., & Anand, T. (1996). The Process of Knowledge Discovery in Databases: A Human-Centered Approach. *Advances in Knowledge Discovery and Data Mining*, 35-58.
- Bramer, M. (2007). *Principles of Data Mining*. Springer.
- Bryant, R., Katz, R., & Lazowska, E. (2008). Big-Data Computing: Creating revolutionary. *Data Science Association*.
- Castilla, N. (2007). *Nicolás Castilla, Nicolás Castilla Prados*. Madrid: Mundi-Prensa.
- Continuum Analytics. (2016). *Anaconda*. Retrieved from <https://www.continuum.io/why-python>
- Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. *Córdoba: Universidad Tecnológica Nacional*.
- Dumbill, E. H., Slocum, M., Croll, A., & Hill, C. (2012). *Big Data Now-2012 Edition*. Sebastopol, CA: O'Reilly Media.
- EcuRed. (2011, Junio 15). *EcuRed*. Retrieved from [https://www.ecured.cu/Archivo:K\\_NN.JPG](https://www.ecured.cu/Archivo:K_NN.JPG)
- El Comercio. (2015, Mayo 30). BIG DATA, ¿EL FUTURO AGRÍCOLA? *El Comercio*.

- Energy Monitoring. (2015, Noviembre 7). *Energy Monitoring WSN*. Retrieved from <https://energymonitoringwsn.wordpress.com/>
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Data Bases. *AI Magazine*.
- Ferrer-Sapena, A., & Sánchez-Pérez, E. (2013). Open data, big data: ¿hacia dónde nos dirigimos? *ThinkEPI*, 150-156.
- Hadoop, A. (2014). *Wellcome to Apache Hadoop*. Retrieved from <http://hadoop.apache.org>
- Han, J., Kamber, M., & Pei, J. (2012). *DATA MINING Concepts and Techniques*. Morgab Kaufmann Publishers.
- Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining*. Massachusetts: MIT.
- Ibáñez, J. J. (2006, Junio). *Madrid*. Retrieved from <http://www.madrimasd.org/blogs/universo/2006/06/26/33002>
- IEEE Computer Society. (2010). Data Engineering. *IEEE Computer Society*.
- Iglesias, N. (2009). <http://inta.gob.ar/>. Retrieved from [http://inta.gob.ar/sites/default/files/script-tmp-inta\\_produccion-de-hortalizas-bajo-cubierta\\_2006.pdf](http://inta.gob.ar/sites/default/files/script-tmp-inta_produccion-de-hortalizas-bajo-cubierta_2006.pdf)
- Iglesias, N. (2009). *INTA*. Retrieved from [http://inta.gob.ar/sites/default/files/script-tmp-inta\\_produccion-de-hortalizas-bajo-cubierta\\_2006.pdf](http://inta.gob.ar/sites/default/files/script-tmp-inta_produccion-de-hortalizas-bajo-cubierta_2006.pdf)
- InfoQ. (2014). *InfoQ*. Retrieved from [https://cdn.infoq.com/statics\\_s2\\_20170314-0434/resource/articles/apache-spark-introduction/en/resources/4.png](https://cdn.infoq.com/statics_s2_20170314-0434/resource/articles/apache-spark-introduction/en/resources/4.png)
- Ishii, N., Hoki, Y., Okada, Y., & Bao, Y. (2009, Septiembre 29). Nearest Neighbor Classification by Relearning. *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, 42-49.
- Java. (2014). *Java.net*. Retrieved from <https://swingx.java.net/>



- Java. (2016). *Java*. Retrieved from <https://www.java.com/es/>
- JFree. (2014). *JFree*. Retrieved from <http://www.jfree.org/jcommon/>
- JFreeChart. (2014). *JFreeChart*. Retrieved from <http://www.jfree.org/jfreechart/>
- Jiang, L., Zhang, H., & Cai, Z. (2006). Dynamic K - Nearest - Neighbor Naive Bayes with Attribute Weighted. *Fuzzy Systems and Knowledge Discovery*, 365-372.
- LASS. (2013, Diciembre). *Laboratory for Advanced Software Systems*. Retrieved from <http://lass.cs.umass.edu/projects/smart/downloadform.php?t=homeA-environmental>
- LEWIS, F. L. (2006). *Wireless sensor networks. Smart environments: technologies, protocols, and applications*. Springer.
- Libelium. (2010). *Libelium*. Retrieved from Libelium Environment: <http://www.libelium.com/applications/Environment>
- Machine Learning Group at the University of Waikato. (2014). *Weka*. Retrieved from <http://www.cs.waikato.ac.nz/ml/weka/>
- MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.
- Malvicinoa, F., & Yoguelb, G. (2015). *Big Data: Avances Recientes a Nivel Internacional y Perspectivas para el Desarrollo Local*. Buenos Aires: Centro Interdisciplinario de Estudios en Ciencia Tecnología e Innovación.
- Marlow, D. (2011, Marzo). *Hortalizas*. Retrieved from <http://www.hortalizas.com/horticultura-prottegida/invernadero/aporte-de-co2-en-un-invernadero/>
- MathWorks. (2016). *MATLAB*. Retrieved from <https://www.mathworks.com/products/matlab/>
- NIST. (2012, Noviembre). *NIST*. Retrieved from <http://math.nist.gov/javanumerics/jama/>

- Oracle. (2016). *ORACLE*. Retrieved from [https://docs.oracle.com/cd/B28359\\_01/datamine.111/b28129/classify.htm#DMCON004](https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004)
- Powel, V. (2014). *Explained Visually*. Retrieved from Principal Component Analysis: <http://setosa.io/ev/principal-component-analysis/>
- Prefuse. (2012). *Prefuse*. Retrieved from [prefuse.org](http://prefuse.org)
- Python. (2015). *Python*. Retrieved from <https://www.python.org/>
- R, T. (2012). Arquitecturas de integración del proceso de descubrimiento de conocimiento con sistemas de gestión de bases de datos: un estado del arte. *Revista de Ingeniería y Competitividad*.
- Rahm, E., & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *Data Engineering*, 3-13.
- Raj, P., & Chandra Deka, G. (2014). *Handbook of Research on*. IGI Global.
- RogueWave . (2017). *RogueWave Software*. Retrieved from [//www.roguewave.com/getmedia/0c450a24-469d-4a96-bd1c-8b8820ffc7e8/vni\\_clusteranalysis](http://www.roguewave.com/getmedia/0c450a24-469d-4a96-bd1c-8b8820ffc7e8/vni_clusteranalysis)
- Rokach, L., & Maimon, O. (2014). *Data Mining with Decision Trees: Theory and Applications*. Singapore: World Scientific.
- Senplades. (2013). *Plan Nacional del Buen Vivir*. Quito: Senplaes.
- SLF4J. (2017). *SLF4J*. Retrieved from <https://www.slf4j.org/>
- Source Forge. (2017). *Source Forge*. Retrieved from <https://sourceforge.net/projects/javacsv/>
- Tascón, M. (2013). Introducción: Big Data- Pasado, presente y futuro . *Telos, Cuadernos de Comunicación e Innovación*, 49-51.

- The University of Waikato. (2014). *The University of Waikato*. Retrieved from [http://www.cs.waikato.ac.nz/~ml/weka/gui\\_explorer.html](http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html)
- Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A., & AlvaradoPérez. (2016). El proceso de descubrimiento de conocimiento en bases de datos. . In *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional+* (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.
- UMassTraceRepository*. (2013, Diciembre 3). Retrieved from <http://traces.cs.umass.edu/>
- Van den Broek, J., Argeseanu, S., Eeckels, R., & Herbss, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLOS MEDICINE*.
- Weber, R. (2000). Data Mining en la Empresa y en las Finanzas. *REVISTA INGENIERÍA DE SISTEMAS*, 61-78.
- Witten, I. H., & Eibe, F. (2013). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. MORGAN KAUFHANN PUBLISHERS.
- Zha, H., H. X., D. C., G. M., & Simon, H. (2001). Spectral relaxation for k-means clustering. *Advances in neural information processing systems*, 1057-1064.

## ANEXOS

## ANEXO A

## Estructura de los datos del Data Set for Sustainability

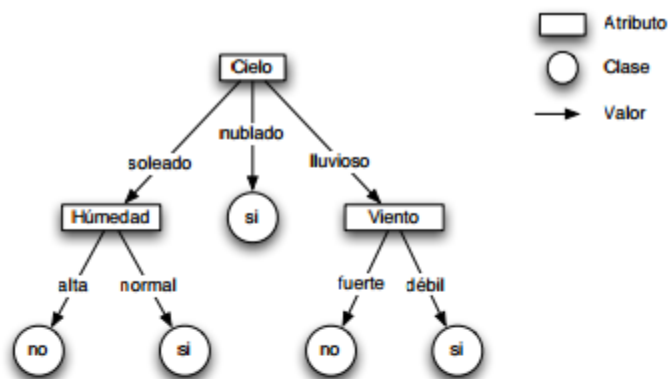
Timestamp	inside Temp	outsideTemp	Inside Humidit	Outside Humidit	Wind Speed	Wind Gust	Wind Chill	heatindex
1341912300	71.78 0.006	50.233.997	41.0	87.099. 998	0.0	0.0	50.233.997	50.233.997
1341912600	71.72 5.998	50.359.997	41.299	87.099. 998	0.0	0.0	50.359.997	50.359.997
1341912900	71.59 9.991	50.306	42.0	86.0	0.0	0.0	50.306	50.306
1341913200	71.59 9.991	50.107.998	42.0	84.400. 002	0.0	0.0	50.107.998	50.107.998
1341913500	71.59 9.991	49.964.001	42.0	84.199. 997	0.0	0.0	49.964.001	49.964.001
1341913800	71.45 5.986	49.820.004	42.0	85.800. 003	0.0	0.0	49.820.004	49.820.004
1341914100	71.41 9.991	49.802.002	42.0	86.0	0.0	0.0	49.802.002	49.802.002
1341914400	71.41 9.991	49.820.004	42.0	86.0	0.0	0.0	49.820.004	49.820.004
1341914700	71.41 9.991	49.91	42.0	86.699. 997	0.0	0.0	49.91	49.91
1341915000	71.41 9.991	5.019.799	42.0	87.099. 998	0.0	0.0	5.019.799	5.019.799
1341915300	71.41 9.991	50.558.002	42.0	86.699. 997	0.0	0.0	50.558.002	50.558.002
1341915600	71.41 9.991	50.990.009	42.0	88.5	0.0	0.0	50.990.009	50.990.009
1341915900	71.41 9.991	51.314.003	42.0	89.0	0.0	0.0	51.314.003	51.314.003
1341916200	71.41 9.991	51.655.998	42.0	89.099. 998	0.0	0.0	51.655.998	51.655.998
1341916500	71.59 9.991	51.835.999	42.0	88.800. 003	0.0	0.0	51.835.999	51.835.999
1341916800	71.59 9.991	52.159.996	42.0	88.699. 997	0.0	0.0	52.159.996	52.159.996
1341917100	71.59 9.991	52.25	42.0	88.699. 997	0.0	0.0	52.25	52.25
1341917400	71.59 9.991	52.340.004	42.0	89.199. 997	0.0	0.0	52.340.004	52.340.004
1341917700	71.59 9.991	52.483.997	42.0	88.400. 002	0.0	0.0	52.483.997	52.483.997
1341918000	71.59 9.991	52.772.003	42.299. 999	88.099. 998	0.0	0.0	52.772.003	52.772.003
1341918300	71.59 9.991	53.006.001	43.0	88.0	0.0	0.0	53.006.001	53.006.001
1341918600	71.76 2.001	53.275.993	42.099. 998	88.300. 003	0.0	0.0	53.275.993	53.275.993

### Estructura de los datos recolectados del invernadero

ID	hora	fecha	temperatura	Humedad relativa	Luz	CO2	Humedad de suelo
384	10:56:02	9/12/2016	27	28	2082	19	650
385	10:56:34	9/12/2016	27	27	2583	19	781
386	10:57:06	9/12/2016	26	29	2571	18	762
387	10:57:37	9/12/2016	26	30	2640	18	763
388	10:58:09	9/12/2016	26	30	560	19	762
389	10:58:41	9/12/2016	26	30	597	18	763
390	10:59:13	9/12/2016	26	30	2427	19	761
391	10:59:45	9/12/2016	26	30	666	19	761
392	11:00:17	9/12/2016	26	30	388	19	763
393	11:00:49	9/12/2016	26	30	376	18	763
394	11:01:20	9/12/2016	26	30	402	19	764
428	15:53:51	9/12/2016	20	34	356	7	588
429	15:54:23	9/12/2016	20	34	220	17	576
430	15:54:55	9/12/2016	20	35	235	17	570
431	15:55:27	9/12/2016	20	36	244	17	565
432	15:55:59	9/12/2016	20	36	260	17	562
433	15:56:31	9/12/2016	20	36	228	17	559
434	15:57:03	9/12/2016	20	37	248	17	572
435	15:57:35	9/12/2016	20	37	468	16	564
436	15:58:06	9/12/2016	20	37	652	16	622
437	15:58:38	9/12/2016	20	36	660	16	617
438	15:59:10	9/12/2016	20	37	673	16	611
476	18:23:04	9/12/2016	18	55	1	18	566
477	18:23:36	9/12/2016	19	55	0	18	561
478	18:24:08	9/12/2016	19	55	0	18	559
479	18:24:40	9/12/2016	19	55	0	18	557
480	18:25:12	9/12/2016	19	55	0	18	556
481	18:25:44	9/12/2016	19	55	0	18	556
482	18:26:16	9/12/2016	19	54	0	18	556
483	18:26:47	9/12/2016	19	54	0	18	556
484	18:27:19	9/12/2016	19	54	0	18	556
485	18:27:51	9/12/2016	19	54	0	19	557

## ANEXO B

## EJEMPLO FUNCIONAMIENTO DEL ALGORITMO C4.5



**Figura 10.1** Un ejemplo de árbol de decisión para el concepto “buen día para jugar tenis”. Los nodos representan un atributo a ser verificado por el clasificador. Las ramas son los posibles valores para el atributo en cuestión. Los textos en círculos, representan las clases consideradas, i.e., los valores posibles del atributo objetivo.

Como el atributo *Cielo*, tiene el valor *soleado* en el caso, éste es filtrado hacia abajo del árbol por la rama de la izquierda. Como el atributo *Humedad*, tiene el valor *alta*, el ejemplo es filtrado nuevamente por rama de la izquierda, lo cual nos lleva a la hoja que indica la clasificación del caso: *Buen día para jugar tenis = no*. El Algoritmo 2, define computacionalmente esta idea.

---

**Algoritmo 2** El algoritmo clasifica, para árboles de decisión
 

---

```

1: function CLASIFICA(Ej, Arbol)
Require: Ej: un ejemplo a clasificar, Arbol: un árbol de decisión
Ensure: Clase: la clase del ejemplo
2: Clase ← tomaValor(raíz(Arbol), Ej);
3: if hoja(raíz(Arbol)) then
4:   return Clase
5: else
6:   clasifica(Ej, subArbol(Arbol, Clase));
7: end if
8: end function
  
```

---

Figura 70 Ejemplo del funcionamiento del Algoritmo C4.5

Fuente: (Timarán-Pereira, Hernández-Arteaga, Caicedo-Zambrano, Hidalgo-Troya, &

AlvaradoPérez, 2016)

## **ANEXO C: LIBRERÍAS USADAS**

### **Java csv**

Java CSV es una librería Java de código abierto que permite leer y escribir archivos y archivos planos delimitados por cierto tipo de carácter. Se puede manejar todo tipo de archivo CSV. Contiene dos clases: `CsvReader` que permite la lectura de los archivos y `CsvWriter` que permite la modificación de los mismos. (Source Forge, 2017)

### **Jama**

JAMA es un paquete básico de álgebra lineal para Java. Proporciona clases a nivel de usuario para construir y manipular matrices densas y reales. Se pretende proporcionar suficiente funcionalidad para problemas de rutina, empaquetados de una manera que sea natural y comprensible para los no expertos. (NIST, 2012)

### **Log4j**

Es una librería de java que permite mostrar mensajes de información, comúnmente conocida como un log.

### **SLF4J**

sirve como una simple fachada o abstracción para varios marcos de registro (por ejemplo, `java.util.logging`, `logback`, `log4j`) que permite al usuario final conectar el marco de registro deseado en el momento de la implementación. (SLF4J, 2017)

### **Swingx**

Contiene extensiones del kit de herramientas GUI de Swing, que incluye componentes nuevos y mejorados que proporcionan la funcionalidad comúnmente requerida por las aplicaciones de cliente enriquecido. Con esta extensión se puede realizar:

Clasificación, filtrado, resaltado de tablas, árboles y listas

Buscar / buscar

Autocompletado

Marco de inicio de sesión / autenticación

Componente TreeTable

Componente de panel plegable

Componente del selector de fecha

Componente de punta del día. (Java, 2014)

### **Jcommon**

JCommon es una librería de java usada por JFreeChart, entre otros proyectos. Ésta contiene clases que soportan:

Configuración y dependencia de código.

Un framwork general de registro.

Utilidades de texto.

Clases de interfaces de usuarios para desplegar información acerca de las aplicaciones.

Un panel de selección de fechas.

Utilidades de serialización. (JFree, 2014)

### **JFreeChart**

Es una librería de gráficos que facilita a los desarrolladores mostrar gráficos de alta calidad en sus aplicaciones. Entre sus funciones están:

Una API consistente y bien documentada, que soporta una amplia gama de tipos de gráficos;

Un diseño flexible que es fácil de extender y apunta tanto a las aplicaciones del lado del servidor como del lado del cliente;



Soporte para muchos tipos de salida, incluyendo componentes Swing y JavaFX, archivos de imagen (incluyendo PNG y JPEG) y formatos de archivo de gráficos vectoriales (incluyendo PDF, EPS y SVG). (JFreeChart, 2014)

### **Prefuse**

Prefijo de herramientas de visualización. Éste es un conjunto de herramientas de software para crear visualizaciones de datos interactivas.

Prefuse soporta una amplia gama de características para el modelado, visualización e interacción de datos. Proporciona estructuras de datos optimizadas para tablas, gráficos y árboles, una gran cantidad de técnicas de diseño y codificación visual y soporte para animaciones, consultas dinámicas, búsqueda integrada y conectividad de bases de datos. (Prefuse, 2012)

### **Substance**

Esta librería te permite cambiar el diseño de las aplicaciones desarrolladas en Java.

## ANEXO D DIAGRAMAS DE PAQUETES Y CLASES

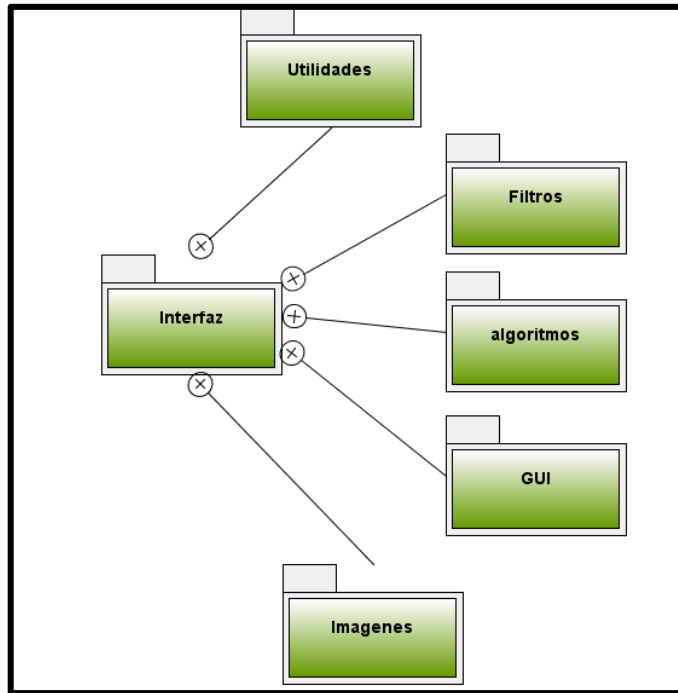


Figura 71 Diagrama de paquetes de la interfaz  
Fuente: Propia.

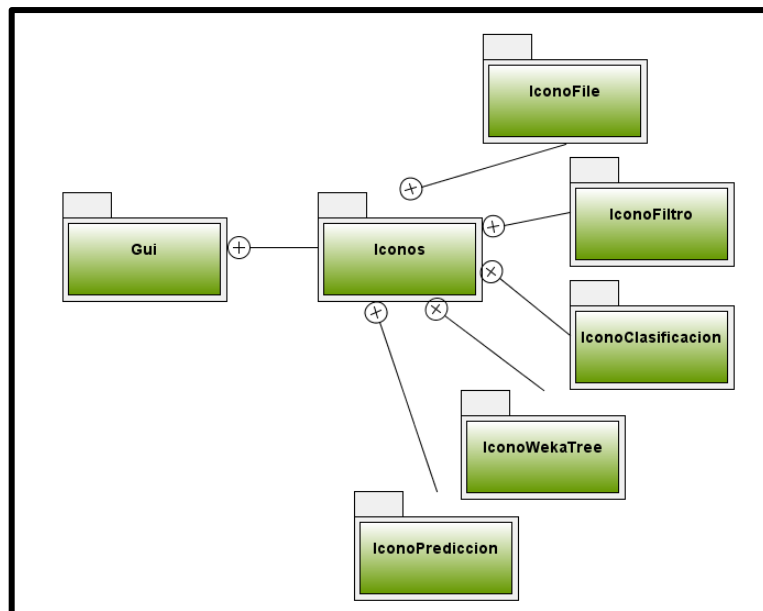


Figura 72 Diagrama de paquetes de GUI, interfaz gráfica de usuario.  
Fuente: Propia.

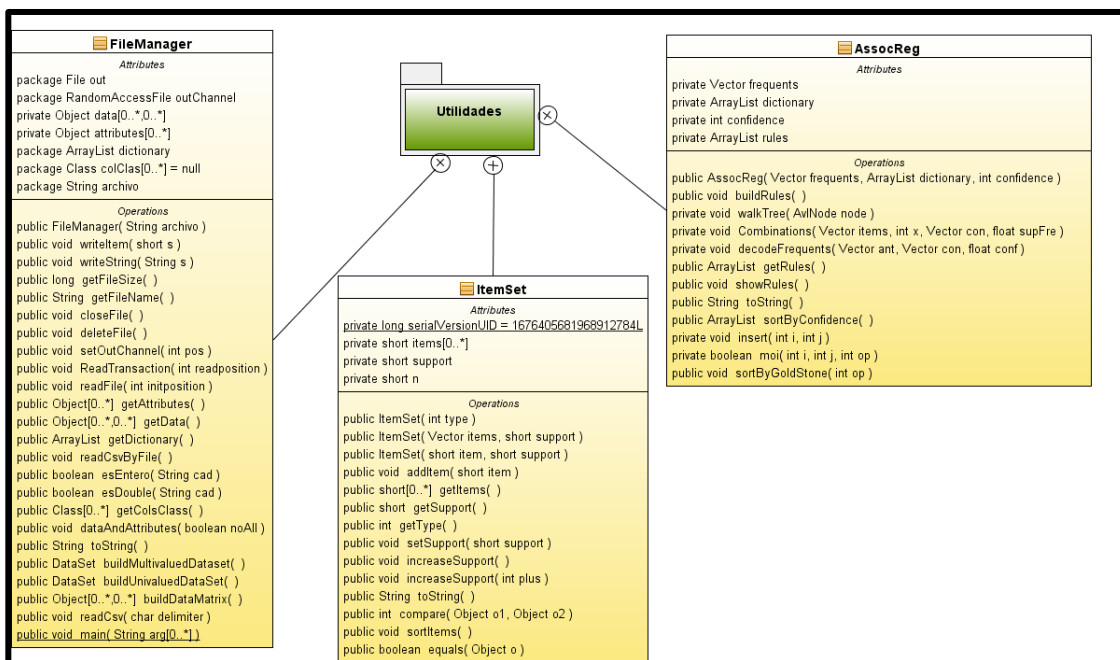


Figura 73 Diagrama de paquetes del paquete de utilidades.  
Fuente: Propia.

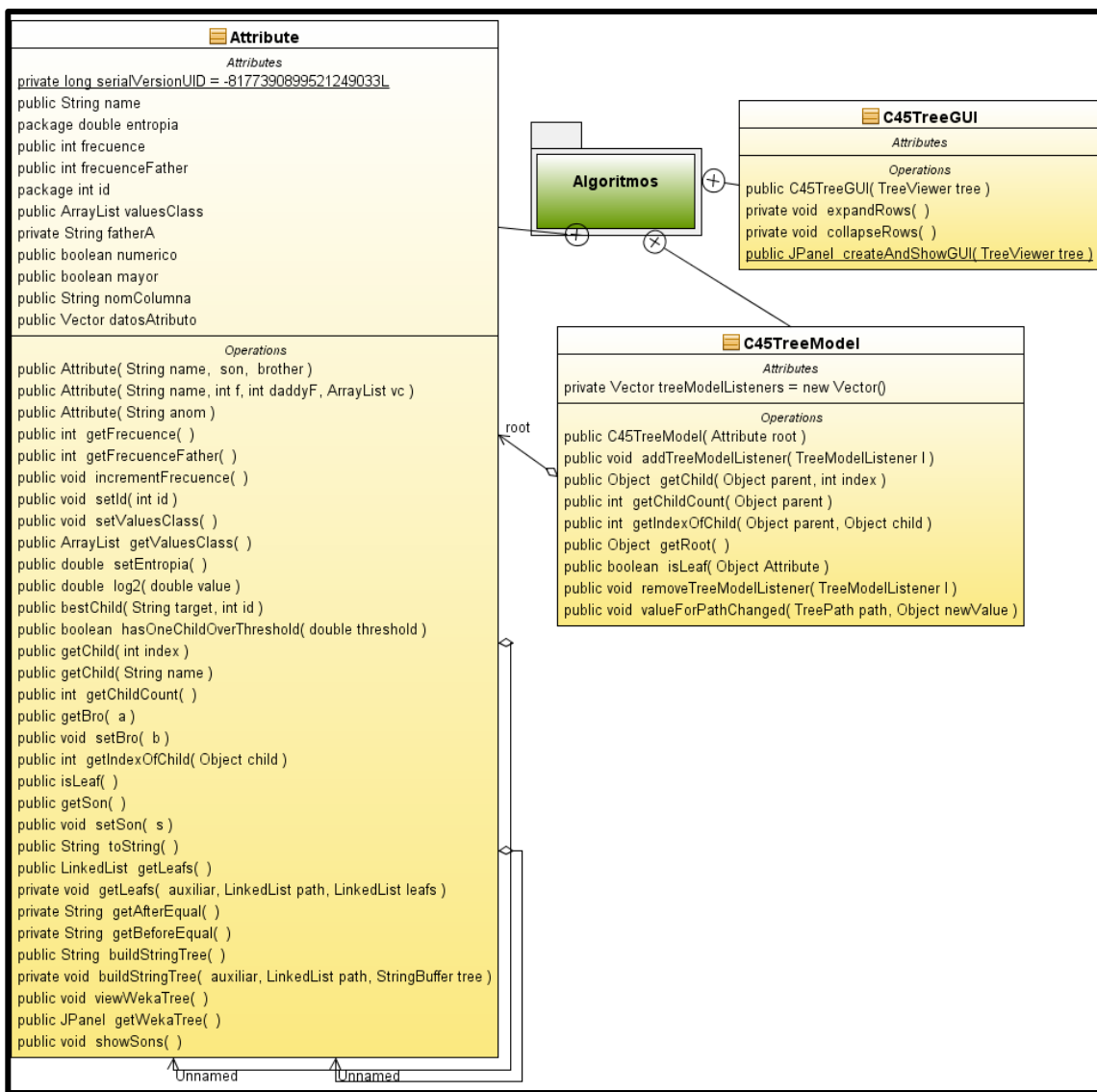


Figura 74 Diagrama de clases del algoritmo C4.5.

Fuente: Propia.

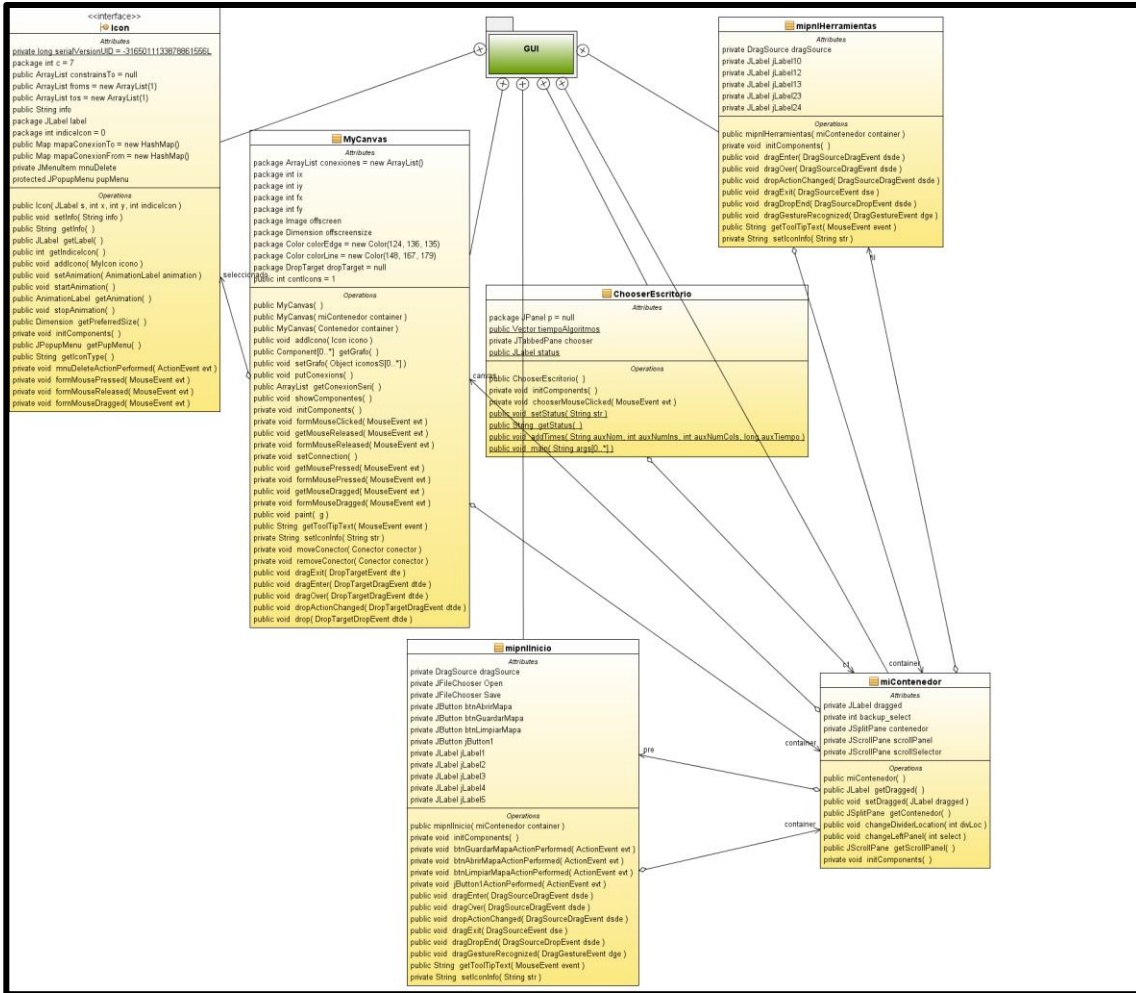


Figura 75 Diagrama de clases del paquete de interfaz gráfica de usuario.  
Fuente: Propia.