UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS

CARRERA DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS COMPUTACIONALES

PROYECTO DE INVESTIGACIÓN

"CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS"

Proyecto de Titulación presentado previo a la obtención del Título de Ingeniería en Informática y Sistemas Computacionales.

Autor:

Fabian Rolando Cayambe Cajo

Tutor Académico:

Ing. MSc. Karla Susana Cantuña Flores

LATACUNGA – ECUADOR

2022

DECLARACIÓN DE AUTORÍA

Yo, Fabian Rolando Cayambe Cajo con C.I: 172318874-2, ser el autor del presente proyecto de Titulación: **"CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS"**, siendo la Ing. Mg. Karla Susana Cantuña Flores, tutor del presente trabajo, eximo expresamente a la Universidad Técnica de Cotopaxi y a sus representantes legales de posibles reclamos o acciones legales.

Además, certificamos que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de nuestra exclusiva responsabilidad.

Atentamente,

.....

Fabian Rolando Cayambe Cajo

CI: 172318874-2

AVAL DEL TUTOR DE PROYECTO DE TITULACIÓN

En calidad de Tutor del Trabajo de Investigación con el título:

"CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS", del estudiante: Fabian Rolando Cayambe Cajo de la Carrera de Ingeniería en Informática y Sistemas Computacionales, considero que dicho Informe Investigativo cumple con los requerimientos metodológicos y aportes científico-técnicos suficientes para ser sometidos a la evaluación del Tribunal de Validación de Proyecto que el Honorable Consejo Académico de la Facultad de Ciencias de la Ingeniería y Aplicadas de la Universidad Técnica de Cotopaxi designe, para su correspondiente estudio y calificación.

Latacunga, Marzo, 2022

.....

Ing. MSc. Karla Susana Cantuña Flores

C.C.: 0502305113

APROBACIÓN DEL TRIBUNAL DE TITULACIÓN

En calidad de Tribunal de Lectores, aprueban el presente Informe de Investigación de acuerdo a las disposiciones reglamentarias emitidas por la Universidad Técnica de Cotopaxi, y por la Facultad de **CIENCIAS DE LA INGENIERÍA Y APLICADAS**; por cuanto, el postulante: **FABIAN ROLANDO CAYAMBE CAJO**, con el título del proyecto de investigación: **"CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS",** ha considerado las recomendaciones emitidas oportunamente y reúne los méritos suficientes para ser sometido al acto de Sustentación del Proyecto.

Por lo antes expuesto, se autoriza realizar los empastados correspondientes, según la normativa institucional.

Latacunga, Marzo, 2022

.....

Lector 1 (Presidente)

Nombre: Ing. MSc. Medina Matute Víctor Hugo

CI: 050137395-5

.....

Lector 2

Nombre: Ing. MSc. Villa Quishpe Manuel

CI: 180338695-0

.....

Lector 3

Nombre: Ing. Mg Segundo Humberto Corrales

CI: 050240928-7

AGRADECIMIENTO

A mi querida madre por su esfuerzo constante

A los docentes de la Universidad Técnica de Cotopaxi por haberme brindado sus conocimientos y ser un apoyo más durante mi formación académica.

Fabian

DEDICATORIA

El presente proyecto de tesis lo dedico primero a Dios por ser el eje principal de mi vida.

A mis padres por motivarme a seguir adelante y apoyarme a superar cada obstáculo que se ha presentado a lo largo de mi vida y mi formación académica.

Fabian

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

UNIVERSIDAD TÉCNICA DE COTOPAXI

FACULTAD DE CIENCIAS DE LA INGENIERÍA Y APLICADAS

TÍTULO: "CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS"

Autores:

Fabian Rolando Cayambe Cajo

RESUMEN

En el presente proyecto de investigación, describe el desarrollo de una aplicación web basada en la comparación de dos técnicas de minería de datos tales como: regresión logística y SVM (máquina de vector de soporte). Para este caso de estudio se realizó la investigación de campo, donde se obtuvieron las imágenes para la creación de la base de datos con 353 registros. En el desarrollo de la aplicación web se recopila los datos como: área, perímetro, centroide y el tipo (monocotiledónea y dicotiledónea), estos datos son utilizados en el proceso de entrenamiento y aprendizaje de los dos algoritmos anteriormente mencionados; ya que es de utilidad para la clasificación automática. Para el desarrollo del prototipo se utilizó la segmentación de imágenes, operaciones morfológicas para el reconocimiento de la hoja y posteriormente se extrae los atributos de las mismas, dichos atributos son guardados en un cvs, el cual se utilizó dos modelos mediante las funciones model= LogisticRegression() y clf = SVC(kernel="rbf").fit(X_train, y_train), para obtener cómo resultado la clasificación de la planta esta puede ser (monocotiledónea y dicotiledónea) , finalmente nos da una precisión de validación de la clasificación en la planta monocotiledónea y dicotiledóneas con regresión logística un 97.75% y en SVM un 73.03%, lo que muestra que la técnica de minería de datos con menor error es de la regresión logística.

Palabras claves: regresión logística, SMV, minería de datos, monocotiledónea y dicotiledónea.

TECHNICAL UNIVERSITY OF COTOPAXI

FACULTY OF ENGINEERING SCIENCES AND APPLIED

THEME: "AUTOMATIC CLASSIFICATION OF MONOCOT AND DICOT PLANTS USING DATA MINING"

Authors:

Fabian Rolando Cayambe Cajo

ABSTRACT

This research project, it describes based on a web application development within the two data mining techniques comparison, such as: logistic regression and SVM (support vector machine). For this case study, it was performed the field research, where was got the images for the database creation with 353 records. Into web application development is collected data, such as: area, perimeter, centroid and type (monocotyledonous and dicotyledonous), these data are used in the two aforementioned algorithms training and learning process; since it is useful for automatic classification. For the prototype development was used the image segmentation, morphological operations for the leaf recognition and subsequently, it is extracted the same attributes, said attributes are saved in a cvs, what used two models, through the functions model = LogisticRegression () and clf = SVC(kernel="rbf").fit(X_train, y_train), to get as a result, the plant classification, this can be (monocotyledonous and dicotyledonous). At the end, it gives a classification validation accuracy in the monocotyledonous and dicotyledonous plant with logistic 97.75% regression and in 73.03%SVM, whose shows that the data mining technique with the least error is logistic regression.

Keywords: Logistic regression, SMV, data mining, monocot and dicot.

AVAL DE TRADUCCIÓN

UNIVERSIDAD TÉCNICA DE COTOPAXI

CENTRO DE IDIOMAS

AVAL DE TRADUCCIÓN

En calidad de Docente del Idioma Inglés del Centro de Idiomas de la Universidad Técnica de Cotopaxi; en forma legal **CERTIFICO** que:

La traducción del resumen al idioma Inglés del trabajo de titulación cuyo título versa: **"CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS"** presentado por: **Fabian Rolando Cayambe Cajo,** estudiante de la Carrera de: **Ingeniería en Informática y Sistemas Computacionales** perteneciente a la **Facultad de Ciencias de la Ingeniería y Aplicadas,** lo realizó bajo mi supervisión y cumple con una correcta estructura gramatical del Idioma.

Es todo cuanto puedo certificar en honor a la verdad y autorizo al peticionario hacer uso del presente aval para los fines académicos legales.

Latacunga, 25 marzo del 2022

Atentamente,

UNIVERSIDAD TECNICA DE COTOPAXI LATACUNGA - ECUADOR CENTRO DE IDIOMAS

Mg. Marco Paúl Beltrán Semblantes

DOCENTE CENTRO DE IDIOMAS-UTC
CI: 0502666514

1. INFORMACIÓN GENERAL

TÍTULO DEL PROYECTO: Clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos.

FECHA DE INICIO: 25 de octubre del 2021.

FECHA DE FINALIZACIÓN: 23 de febrero del 2022.

LUGAR DE EJECUCIÓN: Cotopaxi- Latacunga.

UNIDAD ACADÉMICA QUE AUSPICIA: Facultad en Ciencias de la Ingeniería y Aplicadas.

CARRERA QUE AUSPICIA: Carrera de Ingeniería en Informática y Sistemas Computacionales.

PROYECTO DE INVESTIGACIÓN VINCULADO: Análisis de imágenes en tiempo real de especies de malas hierbas ecuatorianas en cultivos de maíz sobre dispositivos móviles.

EQUIPO DE TRABAJO:

COORDINADOR:

Nombre: Ing. MSc. Karla Susana Cantuña Flores

Nacionalidad: Ecuatoriana

E-mail: karla.cantuna@utc.edu.ec

ESTUDIANTE:

Nombre: Fabian Rolando Cayambe Cajo

Nacionalidad: Ecuatoriano

Correo: fabian.cayambe8742@utc.edu.ec

Celular: 0968682233

ÁREA DEL CONOCIMIENTO: Tecnologías de la Información y Comunicación.

LÍNEA DE INVESTIGACIÓN: Tecnologías de la información y comunicación (TICS) y Diseño Gráfico.

SUB LÍNEA DE INVESTIGACIÓN DE LA CARRERA: Inteligencia artificial e inteligencia de negocios

2. INTRODUCCIÓN

2.1. EL PROBLEMA

2.1.1. Situación problemática

A lo largo del tiempo, las plantas han sido consideradas elementos necesarios y vitales del medio ambiente pues existen en todas partes en donde habita el ser humano [1]. La investigación es realizada por la ciencia de la botánica, la cual se encarga del estudio de la diversidad y estructura de las plantas. El decrecimiento y extinción de variedad de plantas es un asunto serio; por lo cual ante el hallazgo de nuevas especies, se sugiere una rápida identificación y clasificación para su monitoreo, protección y uso en el futuro [1].

Por lo tanto, si hablamos de variedad de plantas, actualmente existen en especies vegetales 8,7 millones, de las cuales 6,5 millones son terrestres y 2,2 millones acuáticas de plantas [2]. Por lo que la cantidad de plantas sería una dificultad para su clasificación. Toman en cuenta a las hojas que son el órgano más variable de la planta y son características de la especie en que crecen. Por este motivo, muchas plantas pueden ser identificadas y clasificadas únicamente por sus hojas [1].

El problema de la clasificación de las plantas es una misión que ha estado presente en las actividades de los botánicos e inclusive de los agricultores, debido a la gran cantidad de familias y clases que existen, sumando el aparecimiento de nuevas especies. En los últimos años, se han desarrollado disciplinas que necesitan de esta tarea. Por ejemplo, estudios de efecto ambiental, son de gran importancia en el inventariado de las especies encontradas. Por este motivo aparece la agricultura convencional.

De acuerdo a la Oficina de Información Científica y Tecnológica para el Congreso de la Unión: Los agricultores se basan en la "agricultura convencional" la cual considera las condiciones de terreno como homogéneas y aplica la misma cantidad de insumos, por ejemplo agua y fertilizantes, a toda la superficie de siembra [3]. Lo que provoca incremento de costos de inversión y aumenta los riesgos de contaminación ambiental.

Por los problemas presentados por la agricultura convencional aparece la "agricultura de precisión", donde es un sistema empleado para analizar y controlar la variación espacio-temporal del terreno y el cultivo [3]. La agricultura de precisión comprende los siguientes campos: los sistemas de posicionamiento global, sensores remotos, monitores de

rendimiento/aplicación y maquinaria inteligente. Es ahí donde ingresa el uso de las máquinas inteligentes a través de métodos inteligentes automáticos.

El uso de métodos automáticos de clasificación ha sido considerado una herramienta para facilitar cualquier actividad en el sector agrícola, la clasificación sería de gran ayuda, a pesar de limitar la variedad de plantas, lo cual implica la reducción de la diversidad de hojas [4]. Bajo este factor, el uso de métodos automáticos de clasificación contribuye a la reducción de los recursos de tiempo, dinero y mano profesional en el proceso de analizar cada clase dentro del volumen total de clases existentes.

Por otro lado, existen técnicas de procesamiento y análisis de datos que tienen como principal objetivo el descubrimiento de conocimiento sobre datos almacenados. Esta técnica es "Minería de Datos", la cual es considerada el proceso de exploración y análisis de datos con el fin de descubrir patrones y reglas significativas [5]. Cabe resaltar la "minería de datos", etapa donde se aplican métodos inteligentes para la extracción de patrones. El cual dentro de estos métodos inteligentes se encuentran los métodos de clasificación, los cuales son considerados una de las técnicas más comunes dentro de la minería de datos.

A nivel mundial en China, se desarrolló el estudio de "Técnicas de minería de datos para rasgos fenotípicos de plantas basados en imágenes identificación y clasificación" [6].Se basaron en la técnica de minería de datos como: análisis discriminante lineal, bosque aleatorio, máquina de vectores de soporte para la clasificación el estado de la planta (estrés/no estrés). La dificultad encontrada fue la demanda de control de enfermedades y numerosas tensiones para mantener los alimentos de calidad en todo el mundo y para reducir las enfermedades transmitidas por los alimentos originados en plantas infectadas.

En América del sur, otro estudio realizado fue una "Aplicación de técnicas de minería de datos para pronósticos del sector agrícola" en Chile [7]. Este estudio se basa en la aplicación de técnicas de minería de datos, encaminado en la identificación de patrones. Encontrar un modelo para el pronóstico de la producción y superficie sembrada de cultivos agrícolas, tales como la papa en la región. En Perú [1], se realizó el "Modelo algorítmico para la clasificación de una hoja de planta en base a sus características de forma y textura". Se encontró el problema de la clasificación de las plantas debido a la variedad de especies vegetales existentes. El propósito de la investigación fue obtener un modelo algorítmico mediante la comparación de cuatro modelos de clasificación de minería de datos.

A nivel nacional en el Ecuador la Escuela Superior Politécnica de Chimborazo, realizó el estudio: “Minería de datos para descubrir tendencias en la clasificación en plantas vegetales”. Se detallan de manera general los algoritmos utilizados, así como el porcentaje de validación para saber la eficiencia de la técnica.

La mayoría de estudios analizados anteriormente han desarrollado estudios con técnicas de minería de datos pero de manera general en plantas no tienen una especie definida. El trabajo de investigación presentado pretende establecer resultados de las técnicas de minería de datos mencionadas anteriormente, desarrollando un modelo de clasificación exclusivamente en el campo del reconocimiento de patrones y minería de datos para la clasificación de plantas monocotiledóneas y dicotiledóneas basadas en las diferencias de la estructura morfológica y características particulares de cada especie mediante el procesamiento digital.

2.1.2. Formulación del problema

¿El uso de técnicas de minería de datos permite la clasificación automática de plantas monocotiledóneas y dicotiledóneas?

2.2. OBJETO Y CAMPO DE ACCIÓN

2.2.1. Objeto de estudio

Técnicas de minería de datos usadas en la clasificación automática de plantas monocotiledóneas y dicotiledóneas

2.2.2. Campo de acción

Minería de datos en la agricultura.

2.3. BENEFICIARIOS

Tabla 1. Beneficiarios directos e indirectos

Beneficiarios Directos	Beneficiarios Indirectos
Estudiantes de la Universidad Técnica de Cotopaxi	Personas externas
Profesores de la Universidad Técnica de Cotopaxi	Sector agrícola

Elaborado por: Equipo de trabajo

2.4. JUSTIFICACIÓN

La agricultura es el eje principal para la economía, por lo que optimizar los procesos manuales a procesos automáticos ayuda a reducir costos a menor tiempo. El incremento y la mejora en la calidad de la producción agrícola es una tarea indispensable para satisfacer la demanda creciente de alimentos y garantizar al agricultor tener una cosecha eficiente [1]. Aproximadamente existen alrededor de 760 especies de gimnospermas y 250000 especies de angiospermas [2]. El agricultor no tiene la facilidad de clasificar las plantas sino de forma manual o conocida como "agricultura convencional". Sumado a esto el decrecimiento y extinción de especies de plantas angiospermas, es un tema serio, por lo cual, ante el descubrimiento de nuevas especies, se propone una rápida identificación y clasificación a fin de poder monitorear, proteger y usar.

El uso de tecnología es necesario para realizar los procesos automáticos en la agricultura. Es donde aparece la minería de datos, el cual permite realizar el proceso de clasificación automáticamente. Con la minería de datos se utiliza nueva tecnología muy poderosa con un gran potencial para ayudar al sector agrícola con la finalidad de enfocarse en la información más importante de sus bases de datos o almacenes de datos. Por ello la técnica más apta para la clasificación de plantas monocotiledóneas, es el método de la regresión logística, la cual permite los datos almacenados, que es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización [4].

La importancia de tener una clasificación automática es permitir la identificación de las características de plantas monocotiledóneas y dicotiledóneas, esto se lo realiza en base a la forma y textura de la hoja, lo cual ayuda al sector agrícola a enfocarse en la información más relevante de sus bases de datos o almacenes de datos. Por lo tanto, teniendo en cuenta lo analizado, se busca obtener un modelo de clasificación, el cual por medio de una aplicación web permite clasificar una nueva instancia de hoja reduciendo el consumo de recursos como: tiempo, dinero y mano de obra.

El proyecto de investigación, beneficiará a los actores indirectos quienes son los especialistas en botánica, brindando la facilidad de realizar clasificación de hojas y plantas de manera más rápida, reduciendo el tiempo invertido en una clasificación manual, así como los recursos de dinero y mano de obra requerida. Asimismo, especialistas ambientales se beneficiarán con la

investigación, dado que en muchas ocasiones necesitan la clasificación de hojas y plantas para realizar estudios de impacto ambiental.

Este proyecto de investigación tiene como objetivo brindar una herramienta tecnológica donde permita tener un adecuado control de las plantas así como la visualización de sus características que define a cada especie.

2.5. HIPÓTESIS

La regresión logística y la máquina de soporte vectorial permiten la clasificación automática de plantas monocotiledóneas y dicotiledóneas.

2.6. OBJETIVOS

2.6.1. Objetivo General

Clasificar plantas monocotiledóneas y dicotiledóneas usando técnicas de minería de datos como: la regresión logística y la máquina de soporte vectorial para el reconocimiento automático en imágenes digitales.

2.6.2. Objetivos Específicos

- Definir las bases teóricas acerca de minería de datos, técnicas de clasificación, reconocimiento de imágenes y plantas monocotiledóneas y dicotiledóneas para la redacción del marco teórico.
- Aplicar algoritmos de minería de datos como la regresión logística y máquina de soporte vectorial en la clasificación automática de plantas monocotiledóneas y dicotiledóneas.
- Desarrollar un prototipo de clasificación automática de plantas monocotiledóneas y dicotiledóneas.

2.7. SISTEMA DE TAREAS

Tabla 2. Planificación de las actividades

OBJETIVOS ESPECÍFICOS	ACTIVIDADES	RESULTADO DE LAS ACTIVIDADES	DESCRIPCIÓN (TÉCNICAS E INSTRUMENTOS)
Definir las bases teóricas acerca de minería de datos, técnicas de clasificación y plantas monocotiledóneas y dicotiledóneas para la redacción del marco teórico.	Búsqueda de información en revistas científicas. Búsqueda de información en artículos científicos. Redacción del marco teórico.	Marco Teórico	**Técnica** Análisis documental **Instrumento** Ficha bibliográfica
Aplicar algoritmos de minería de datos como la regresión logística y máquina de soporte vectorial en la clasificación de plantas monocotiledóneas y dicotiledóneas	Elaboración de data set Aplicación de regresión logística. Aplicación de máquina de vectores de soporte.	Data set Modelos	**Técnica** Análisis documental **Instrumento** Ficha bibliográfica Cámara fotográfica
Desarrollar un prototipo de clasificación automática de plantas monocotiledóneas y dicotiledóneas.	Diseño del prototipo Integración de modelos obtenidos Aplicación de plan de pruebas	Prototipo de aplicación de web	Aplicación web-Metodología CRISP-DM

Elaborado por: Equipo de trabajo

3. FUNDAMENTACIÓN TEÓRICA

3.1. ANTECEDENTES

a. Reconocimiento de imágenes con técnicas de minería de datos.

Objetivo implementar una aplicación que permita el reconocimiento de imágenes con técnicas de minería de datos captando las fotografías del abecedario del lenguaje de señas a través de una cámara web [8].

La metodología aplicada es mixta, descriptiva, experimental, método inductivo-deductivo, incluyendo la metodología de trabajo CRISP-DM con 4 fases como: comprensión del negocio, preparación o procesamiento digital, entrenamiento, e identificación.

En la preparación se utilizó la cámara web y se vinculó con el programa MATLAB, capturando 24 imágenes por cada seña con fondo negro. Para identificar las características se efectuó dos procesos, en el primero se aplicó el algoritmo de binarización, y como siguiente paso el algoritmo histogramas de gradientes orientados.

b. Innovación en Minería de Datos para el Tratamiento de Imágenes: Agrupamiento K-media para Conjuntos de Datos de Forma Alargada y su Aplicación en la Agroindustria.

Por medio de este trabajo a través de minería de datos por medio de un método de agrupación K-media modificado basado en la teoría de conjunto junto con su aplicación en el ámbito de procesamiento de imágenes agroindustrial.[9]

K-media tradicional permite la agrupación de conjuntos en subconjuntos mediante la definición de centros según la fórmula de distancia. Cuando los datos se concentran en formas sin un sentido hiper-esférico, esta herramienta permite que el centro del conjunto, con un único punto, se convierta en un subconjunto de muchos puntos.

3.2. MARCO CONCEPTUAL

3.2.1. Agricultura de precisión

Según [3] define a la agricultura de precisión como: el manejo basado en el análisis y control con relación a la variabilidad espacio-temporal del terreno y del cultivo. Suministra distintas cantidades de insumos y toma en cuenta la variación en los componentes del suelo (como textura, acidez, humedad, topografía o relieve), en el desarrollo vegetal y en las condiciones entre temporadas de siembra.

3.2.1.1. Etapas de la agricultura de precisión

La agricultura de precisión se basa en las siguientes etapas detalladas en la **Figura 1**:

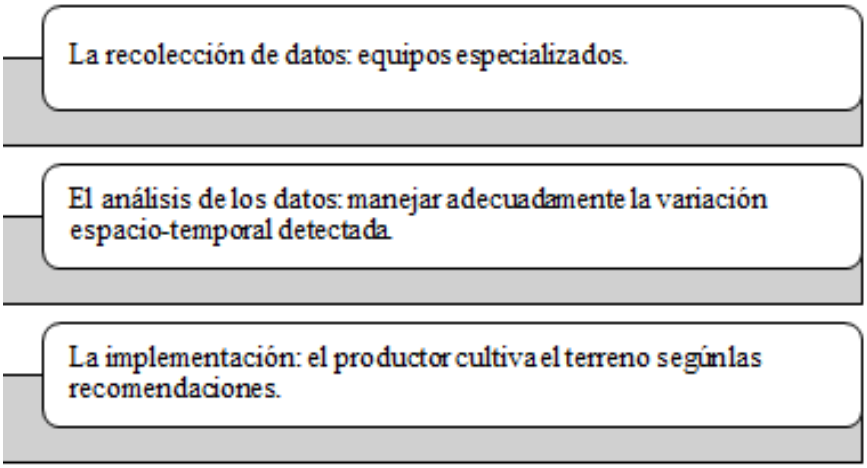


Figura 1.Etapas de la agricultura de precisión

3.2.1.2. Tecnologías asociadas con la agricultura de precisión

En la agricultura de precisión [3] se basa en las siguientes tecnologías detalladas a continuación:

a) **Sistemas de posicionamiento global:** este se emplea en monitores de rendimiento, banderilleros satelitales, pilotos automáticos o en equipos de aplicación variable.
b) **Sistemas de información geográfica:** en la agricultura de precisión los distintos receptores (v.g. sensores remotos), para tomar decisiones sobre el manejo de la variabilidad espacio-temporal.
c) **Sensores remotos:** este es emplean en la recolección de datos sobre la administración del agua de riego.
d) **Monitores de rendimiento y aplicación:** los monitores de rendimiento obtienen información sobre la cantidad del cultivo.
e) **Maquinaria inteligente:** sistemas de piloto automático.

3.2.2. Plantas

Es un organismo vegetal en el que se observa una diferenciación en tejidos tales como: raíces, tallo, hojas y flores, y las semillas donde permite la reproducción de la planta reproducción y un sistema circulatorio que permita el flujo de nutrientes a lo largo de todo su "cuerpo" [10]. Las plantas tienen dos características principales: a) Adaptarse a la vida terrestre b) obtención de la luz solar para vivir.

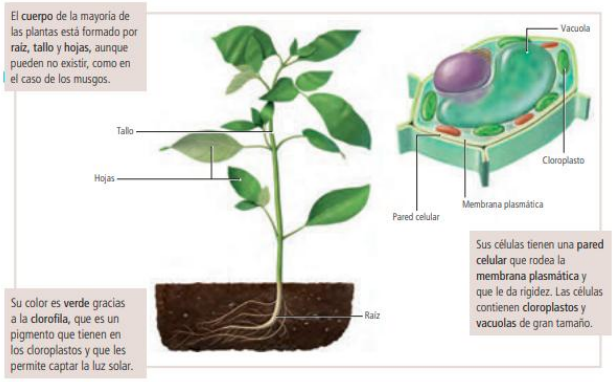


Figura 2.Esquema general de los órganos de una planta vascular

Presentan las siguientes características en la **Figura 3**:

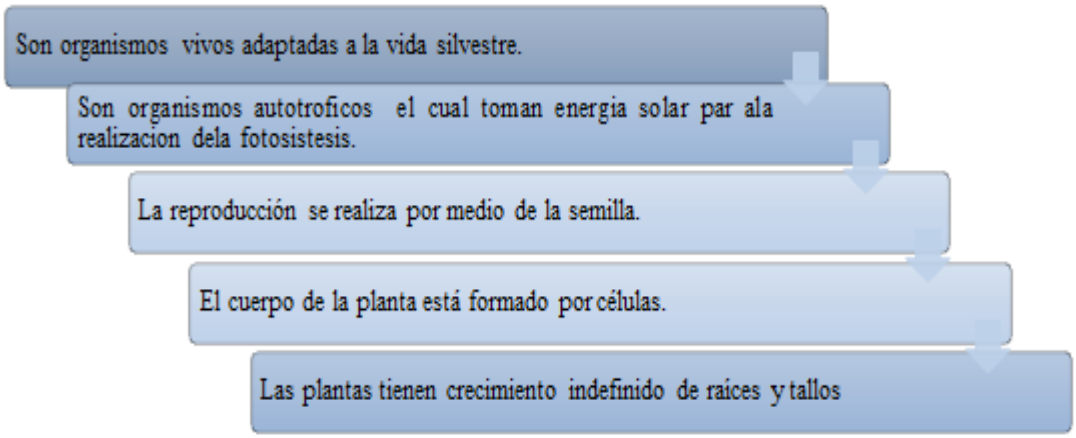


Figura 3. Características de las plantas vasculares

3.2.3. Clasificación de las plantas

3.2.3.1. Plantas con flores

Estos poseen flores, semillas y presentan vasos conductores. El cuerpo de las plantas con flores posee órganos vegetativos, que se encargan de la nutrición y el mantenimiento de la planta, y órganos reproductores, que llevan a cabo la función de reproducción. Se tiene dos grandes clasificaciones tales como: plantas gimnospermas y plantas angiospermas[11]. La gran mayoría de las plantas producen semillas y comprenden entre 760 especies gimnospermas y 250.000 especies angiospermas[11].

a. Plantas gimnospermas

Las gimnospermas [2] son arbustos o árboles que se extienden por el planeta, en especial, en las zonas más frías. Todas las plantas gimnospermas comparten estos rasgos:

- **Semilla:** no se encuentran encerradas en un fruto

- **Flores:** son unisexuales y pueden estar en diferentes ramas. Tienen una forma de cono.
- **Reproducción:** los granos de polen de las flores masculinas llegan, con el viento, los óvulos que están desnudos sobre las escamas del cono femenino y se produce la fecundación.

Figura 4. Reproducción de una planta gimnosperma - pino

b. Plantas angiospermas

Las angiospermas son el grupo de plantas más abundante y diversas que producen flores y frutos con aromas para atraer los polinizadores como dispensadores [11]. Los ejemplos de plantas angiospermas pueden ser árboles, arbustos y hierbas, y se pueden encontrar en todo tipo de ambientes, incluso en el desierto y en el mar. Un porcentaje mayoritario de plantas en el planeta son angiospermas.

Estas plantas tienen las siguientes características:

- **Semilla:** se encuentran encerradas en un fruto.
- **Flores:** son unisexuales y hermafroditas esto se presenta en los estambres y el pistilo de la misma flor.
- **Reproducción:** se realiza por medio de los óvulos que se encuentran protegidos en un ovario.

Figura 5. Reproducción de una planta angiosperma - manzana

3.2.4. Monocotiledóneas

Las plantas monocotiledóneas sus embriones tienen una única hoja o cotiledón, los nervios de sus hojas son paralelos y las envueltas florales suelen tener tres o seis partes. En la Tabla 3 se detalla cada parte de las plantas monocotiledóneas como: semilla, flor, hoja, tejido vascular, raíz, y finalmente el polen.

Tabla 3. Características de las plantas monocotiledóneas

Nombre	Característica
Semilla	Tiene el embrión con un cotiledón
Flor	Las partes de la flor son en múltiplo de tres.
Hoja	Las venas de las hojas son usualmente paralelas.
Tejido vascular	El tejido vascular en forma de haces vasculares dispersos.
Raíz	Sistema de raíces fibroso
Polen	El polen usualmente tiene 1 poro o surco

Elaborado por: Equipo de trabajo

3.2.5. Dicotiledóneas

Las plantas dicotiledóneas sus embriones poseen dos hojas o cotiledones, los nervios de las hojas son ramificados y las envueltas florales suelen tener cuatro o cinco partes. En la Tabla 4 se muestra el detalle de cada parte de las plantas dicotiledóneas como: semilla, flor, hoja, tejido vascular, raíz, y finalmente el polen.

Tabla 4. Características de las plantas dicotiledóneas

Nombre	Característica
Semilla	Tiene el embrión con dos cotiledones.
Flor	Las partes de la flor son en múltiplo de cuatro a cinco.
Hoja	Las venas de las hojas son usualmente ramificadas.
Tejido vascular	El tejido vascular en forma anillo en el tallo.
Raíz	Sistema de raíces pivotante.
Polen	El polen usualmente tiene 3 o más poro o surco

Elaborado por: Equipo de trabajo

3.2.6. Identificación de las plantas angiospermas

Las angiospermas son el grupo de plantas más diverso de la actualidad [12]. Se puede observar en todos los ecosistemas, conviviendo con otros grupos de plantas. Para la identificación de plantas angiospermas existen dos grandes clasificaciones como: monocotiledóneas y dicotiledóneas Mediante la forma de la hoja se puede realizar la identificación de las plantas angiospermas con las siguientes características. Hojas simples tienen un tipo ápice, tipo de margen, tipo de lámina, tipo de venación, tipo de base. Hojas compuestas a través de la morfología de los folios.

3.2.7. Aprendizaje automático

Se puede definir como un campo de la inteligencia artificial que se usa para el análisis y el descubrimiento del conocimiento en bases de datos [13]. El aprendizaje automático proporciona las técnicas a la minería de datos; usadas para extraer información de las bases de datos. Entre sus principales características [13] se mencionan:

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

La importancia de los métodos de minería de datos tienen su origen en el aprendizaje automático; sin embargo no se puede considerar que el aprendizaje automático es un subconjunto de minería de datos; ya que este también abarca otros campos [13].

3.2.8. Minería de datos

La minería de datos es utilizada en diferentes disciplinas para la búsqueda de patrones y modelos ocultos en las bases de datos [14].

> "La minería de datos es el proceso de descubrir correlaciones, patrones y tendencias significativas mediante la selección de grandes cantidades de datos almacenados en repositorios. La minería de datos emplea tecnologías de reconocimiento de patrones, así como técnicas estadísticas y matemáticas"[14].

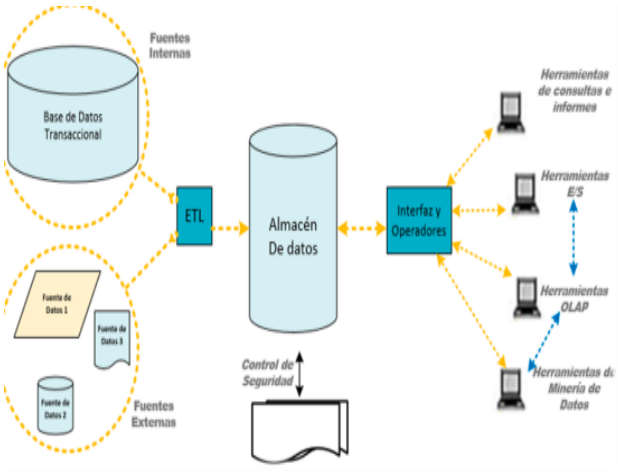


Figura 6. Proceso de minería de datos

La minería de datos de alguna forma facilita los procesos de clasificación de acuerdo a patrones que se requieran por medio de los datos almacenados, estos datos deben constar de un número significativo para la predicción, también seguido a los datos seleccionados se realiza el proceso de clasificación mediante estadísticas y matemáticas [15]. Los siguientes paradigmas de la minería de datos donde está la verificación y descubrimiento de nueva información **Figura 7.**

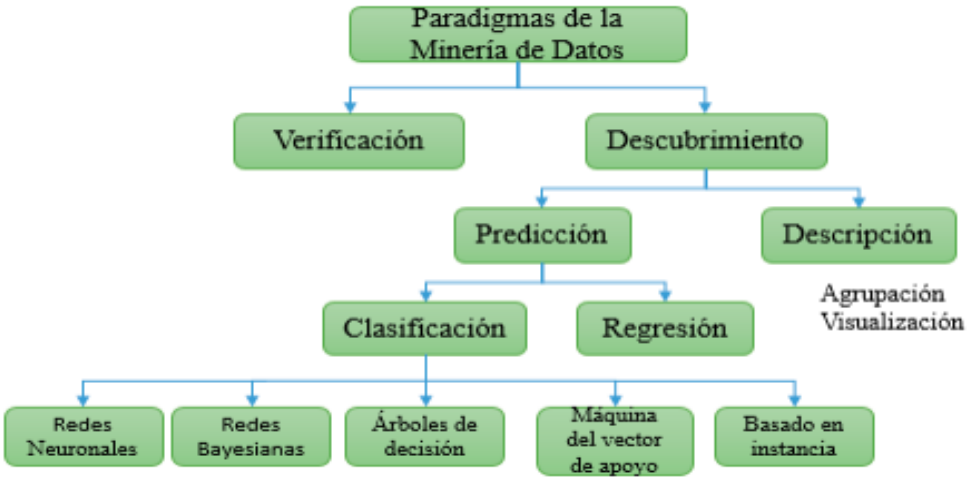


Figura 7.Paradigmas de la minería de datos [1]

3.2.8.1. Herramientas de minería de datos

Las herramientas de la minería de datos [8] ayudan a gestionar datos e identificar patrones, por ello se clasifican donde se detalla en la ***Figura 8*** con dos clasificaciones: a) método de descubrimiento y b) técnicas de verificación.

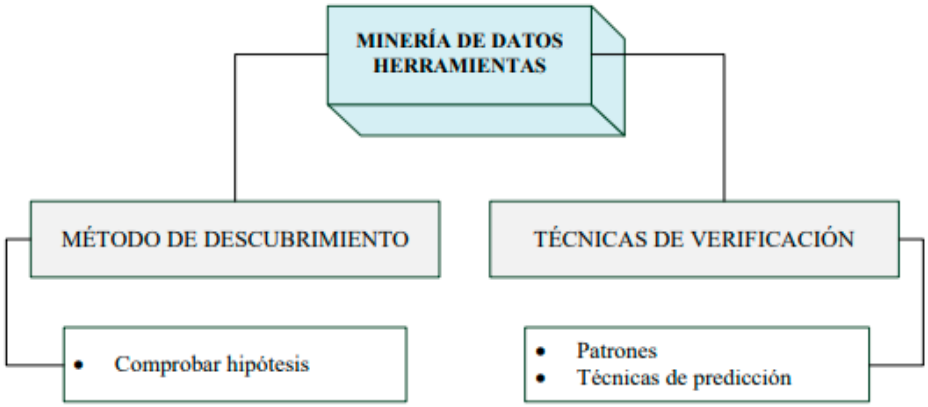


Figura 8.Herramientas de minería de datos [8]

Entre las herramientas más usadas tenemos [8]:

Tabla 5. Herramientas de minería de datos

Ítems	Herramientas	Características	Lenguaje de programación	Sistema operativo
1	**Orange**	Visualización de datos sin mucho conocimiento previo, análisis predilecto.	Núcleo del software: C++, ampliación y lenguaje de entrada: Python	Windows, macOS, Linux
2	**Weka**	Tiene muchos métodos de clasificación.	**Java**	
3	**RapidMiner**	Adaptación de los procesos y análisis de datos. Está basado en el análisis predictivo		
4	**KNIME**	Análisis predictivos		

Elaborado por: Equipo de trabajo

3.2.8.2. Aplicaciones de la minería de datos

Según [8], la minería de datos se aplica en diferentes áreas con el propósito de obtener patrones para la extracción de información.

Tabla 6.Aplicaciones de la minería de datos

Ítems	**Campo de aplicación**	**Características**
1	Medicina	Tener un valor científico o investigador que ayude a determinar causas de determinadas patologías o a identificar poblaciones de riesgo
2	Seguridad y detección de fraude	Reconocimiento facial, evasión fiscal, detección de acceso a redes restringidas.
3	Recuperación de datos no numéricos	Texto, web, imagen, audio, y voz.
4	Comercio y banca	Análisis de riesgo, pronóstico de ventas, segmentación de clientes.
5	Agricultura	Se realiza en el diseño de patrones de estratificación de cultivos en minería de datos para una agricultura de precisión. Generar estimaciones de superficie sembrada y cosecha de cultivos anuales esenciales en el territorio

Elaborado por: Equipo de trabajo

3.2.8.3. Etapas de la minería de datos

En [4] detalla las etapas de la minería de datos se encuentra la siguiente:

a) **Objetivo y recolección de datos:** elección de la información.
b) **Procesamiento y gestión de los datos**: requiere seleccionar la muestra representativa sobre la cual llevar a cabo el análisis.
c) **Selección del modelo:** crear un modelo o algoritmo (minería de datos) que nos brinde un posible resultado.

d) **Análisis y revisión de resultados:** tiene como objetivo analizar los resultados para tener una explicación lógica.

e) **Actualización del modelo:** tiene la característica de que las variables del modelo podrían volverse insignificantes.

3.2.8.1. Clasificación de minería de datos

La clasificación de datos es un proceso que consta de dos etapas. La técnica de clasificación examina las características de un nuevo objeto con el fin de asignarlo a una clase que esta predefinida dentro de un conjunto de clases [1]. La clasificación está representada por registros en la tabla de la base de datos o una fila, y el acto de clasificar consiste en agregar una nueva columna con un código de clase de algún tipo.

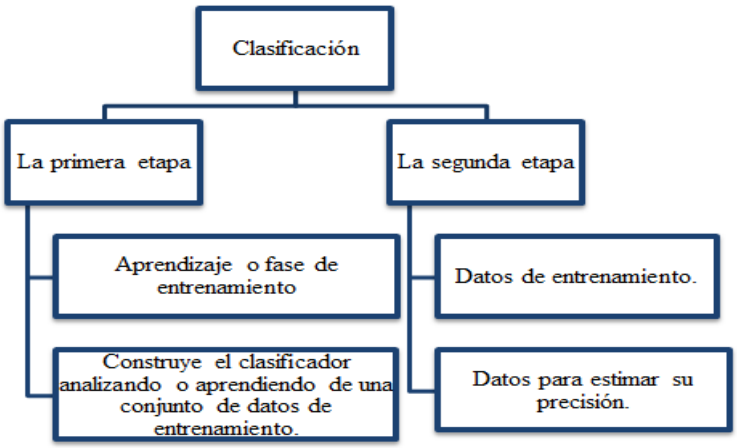


Figura 9. Clasificación de minería de datos

3.2.8.4. Técnicas de minería de datos

La minería de datos es un área que posee numerosas técnicas para realizar el conjunto de actividades que la comprenden, éstas se clasifican en dos grandes categorías [16]:

a. Predictivas: especifican el modelo para los datos con base en un conocimiento teórico previo, en el cual se busca realizar predicciones del valor de un atributo partiendo de las relaciones de datos conocidos. Entre los principales según [17] se detalla a continuación:

Series temporales: se parte de una serie de comportamientos históricos y se modelan series, tendencias, ciclos y estaciones para realizar predicciones.

Redes bayesianas: representan posibles sucesos o consecuencias mediante un grafo de probabilidades condicionales, el cual permite establecer relaciones causales y efectuar predicciones [18].

Árboles de decisión: presentan de forma visual las reglas de decisión, partiendo de datos históricos, su gran ventaja es la facilidad de interpretación.

Redes neuronales: imitan la estructura cerebral en relación con las neuronas y sus conexiones, buscando crear modelos artificiales para la solución de problemas mediante técnicas algorítmicas tradicionales [16].

Algoritmos genéticos: métodos adaptativos que pueden usarse para resolver problemas de búsqueda y optimización. Son funciones matemáticas o rutinas de software que toman como entrada los ejemplares y retorna como salida, cuáles de ellos deben producir descendencia para la nueva generación.

Regresión: técnica estadística que permite cuantificar la relación entre variables.

Análisis de la varianza y covarianza: medidas estadísticas, la primera mide la dispersión de los valores respecto a un valor central (media), la segunda es un indicador de la relación entre variables.

b. Descriptivas: no requieren de variables ni modelos previos de datos, intentan describir el comportamiento de los datos y las relaciones encontradas, los modelos se crean automáticamente partiendo del reconocimiento de patrones.

Clustering: técnica de agrupamiento de acuerdo con un criterio. Busca relación entre variables descriptivas y permite una descripción clara de un grupo de datos complejos.

Segmentación: clasificación de los datos en grupos específicos, incluye un aprendizaje no supervisado.

Asociación y dependencia: conocido como análisis exploratorio. La asociación se refiere a una alta frecuencia en la aparición de dos valores relacionados y la dependencia a que el valor de un atributo varía en relación con otros atributos.

Escalamiento multidimensional: Es muy importante en el reconocimiento de patrones, debido a que permite identificar la estructura de los datos y la extracción de información relevante del objeto de estudio.

3.2.9. Regresión logística

Este algoritmo de clasificación permite a través de un clasificador estimar la probabilidad de que un nuevo ejemplo pertenezca a una clase. Como para el presente problema se tuvo un total de ocho clases, se aplicó además una técnica denominada “one vs all” [19].

En el modelo logístico simple, consideramos el conjunto de datos X con variables respuesta binaria. Con el cual, para cada elemento xi de X la salida se denomina respuesta se puede ser yi = 1 o bien yi = 0. Entonces, el conjunto de entrenamiento X consistirá en "n" elementos del tipo (xi, yi), donde yi es el valor de la variable binaria de clasificación y xi es un vector de variables explicativas. Los elementos con salida yi = 1 se dice que pertenecen a la clase positiva, mientras que los que tengan yi = 0 pertenecen a la clase negativa. En este modelo el valor de la variable respuesta para un nuevo elemento, se obtiene en términos de probabilidad, empleando la función de distribución logística.

3.2.10-. Máquina de soporte vectorial

Una máquina de soporte vectorial [20] es un sistema de aprendizaje automático utilizado para resolver problemas de clasificación y regresión de manera eficiente, lo que le ha permitido posicionarse por encima de otras técnicas de clasificación.

La primera radica en que poseen una base matemática muy sólida y la segunda, en que se basan en el concepto de minimización del riesgo estructural. Lo que permite minimizar el riesgo de una clasificación errónea al introducir nuevas muestras y la tercera, en que dispone de potentes herramientas y algoritmos para hallar la solución de manera rápida y eficiente.

El funcionamiento de estas máquinas son capaces de clasificar muestras en dos posibles conjuntos: "positivos" y "negativos", que en el caso de detección de rostros corresponden a "rostros" y "no rostros" respectivamente.

3.2.11. Metodología CRISP-DM

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [21] detalla como: un ciclo de vida, similar a los de ingeniería de software, creada para proyectos de análisis de datos, con la diferencia que presenta un proceso más normalizado y racionalizado.

Es conocido como "un modelo de procesos jerárquicos, consistente en un conjunto de tareas descritas en 4 niveles de abstracción" [7], desde el general hasta lo específico contiene fases, tareas generales, tareas específicas e instancias de proceso.

La metodología CRISP-DM tiene una estructura de vida en seis fases **Figura 10**, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

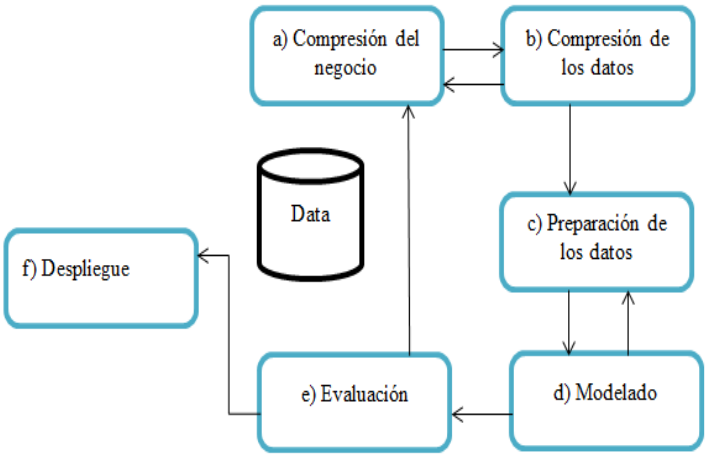


Figura 10. Metodología CRISP-DM

a. Comprensión del negocio

Entender los objetivos y requerimientos del proyecto sobre el negocio para posteriormente plasmarlo en la definición del problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

En donde se contextualiza la problemática y la interpretación que se quiere dar, trabajo a realizar en conjunto con el experto del negocio. Está basado en subfases como: establecimiento de los objetivos de negocio, evaluación de la situación, establecimiento de los objetivos de minería de datos y generación el plan del proyecto.

b. Comprensión de datos

Esta etapa se debe complementar con el análisis práctico de cada variable involucrada, modelo de datos, documentación asociada y la interpretación del experto del negocio.

Puede existir una constante retroalimentación con el entendimiento del negocio para una mejor comprensión general.

Esta etapa comprende de cuatro subtareas importantes detalladas a continuación: recopilación inicial de datos, descripción de datos, exploración de datos y verificación de calidad de datos.

c. Preparación de datos

En la preparación de datos están bajo ciertos criterios fundamentados se realiza la selección, limpieza y formateo de los datos. Con el objetivo de obtener resultados significativos en función de los datos seleccionados y del filtrado de las variables más representativas al problema.

Para esta fase de debe comprender las siguientes subtareas: selección de datos, limpieza de datos, construcción de datos y formateo de datos.

d. Modelado

Para la realización del modelado estará relacionado de acuerdo a la preparación de los datos por tanto interactúan sistemáticamente. Esta etapa consta de 4 subfases: selección de la técnica de modelado, diseño de la evaluación, construcción del modelo y evaluación del modelo.

e. Evaluación

En esta etapa se evalúa la aplicación del modelo de análisis adoptado. Para comprobar que el modelo se ajusta a los datos de entrada como para verificar que se ajusta a las necesidades establecidas en la primera fase; es decir que el modelo sirve para responder algunos requerimientos del negocio tales como: evaluación de resultados, revisar el proceso y establecimiento de los siguientes pasos de lista de posibles acciones.

f. Despliegue

Asegurar el mantenimiento de la aplicación del modelo de análisis y la difusión de los resultados, lo que en ciertas oportunidades puede implicar comenzar nuevamente el ciclo con el nuevo conocimiento aprendido en el ciclo anterior.

Tiene el propósito de tratar de explotar la potencialidad de los modelos integrándose en los procesos de toma de decisiones de la organización.

3.2.12. Procesamiento de imágenes digitales

El reconocimiento de imágenes se basa en un mecanismo para identificar de forma automática un objeto dentro de la imagen proyectada relacionados con formas, tamaños, colores, entre otros [8].

3.2.12.1. Etapas de procesamiento de imágenes digitales

Las fases del reconocimiento de imágenes se detallan en la **Figura 11.**

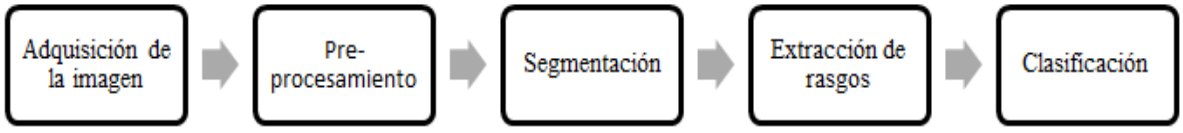


Figura 11. Etapas del reconocimiento de imágenes

a. Adquisición de la imagen

Esta fase es donde se toma en cuenta la calidad de la imagen mediante diversos mecanismos o sensores para digitalizar, procesar, almacenar y entregar del resultado. A través de esto se podrá obtener un mejor resultado para el reconocimiento de la imagen.

b. Pre-procesamiento

En esta fase busca un mecanismo (medios ópticos, electro ópticos o electrónicos) para mejorar el contraste, reducir errores para la siguiente fase.

c. Segmentación

Separación de objetos conforme al criterio de importancia de la imagen zonas disjuntas para diferenciar los distintos objetos y el fondo, la importancia es donde cada píxel de la imagen tendrá una etiqueta, la agrupación de puntos con la misma etiqueta y conectados espacialmente determinan la lista de objetos.

d. Extracción de rasgos

Se detalla de manera numérica los objetos separados según los atributos seleccionados.

e. Clasificación

Asignación de clase a los objetos en relación a los rasgos que se desea investigar.

3.2.12.2. Segmentación de imágenes

En [22] define la segmentación como: "una de las etapas más importantes en el tratamiento de imágenes". La segmentación se puede considerar como la forma de separar los objetos de interés en una imagen, del resto que se consideran no relevantes.

En todos los métodos de segmentación se trata de asignar cada píxel a un cierto grupo, llamado comúnmente "segmento". La imagen está compuesta por valores numéricos (uno o más valores de color para cada píxel).

La segmentación se basa en tres propiedades:

- Similitud: cada imagen tiene píxeles los cuales tienen valores parecidos para alguna propiedad.
- Discontinuidad: los objetos destacan del entorno y tiene por tanto unos bordes definidos.
- Conectividad: en la imagen los píxeles pertenecientes a la misma región tienen que ser contiguos.

a. Técnicas de segmentación de imágenes

Las técnicas utilizadas para la segmentación de imágenes se detalla en [22] el cual se basa en:

- **Frontera Global:** la obtención del umbral se da bajo diversos criterios, ya que puede ser manejado por el mismo usuario, donde se establece un punto de umbralización o se obtiene de forma automática.
- **Segmentación adaptativa:** esto se basa en un umbral para la obtención de aquellos píxeles necesarios para la obtención de alguna característica en común, ambos devuelven una imagen binarizada.
- **Segmentación por clustering:** el algoritmo trabaja con el fondo de la parte de la imagen para segmentación, por ello utilizará una componente o tres componentes.
- **Segmentación Otsu:** tiene el objetivo de buscar un umbral óptimo dado la imagen seleccionada.

3.2.12.3. Técnicas de extracción de características

Las imágenes son utilizadas en múltiples campos de aplicación [22]. La visión artificial utiliza métodos para procesar, analizar y comprender imágenes de una forma parecida a como lo hace el ser humano. Para ello, utiliza métodos de procesamiento que permiten extraer la información importante de la imagen. A través de las siguientes características de la imagen:

a. Altura y ancho

Mediante la altura y el ancho, son dos descriptores muy convenientes para objetos de poca complejidad.

b. Área

En la segmentación de imágenes, el área de un objeto está dada por el número de píxeles que representan al mismo, por lo tanto, el cálculo del área se realiza contando el número de píxeles.

Para hallar la segmentación por el área de la imagen está en un contorno codificado en código de cadena, el cálculo del área se realiza por un algoritmo que trabaja en forma similar a la integración numérica.

c. Perímetro

Para la segmentación de imágenes por medio de la característica de perímetro puede ser calculado a partir del código de cadena.

Para realizar este cálculo, es necesario contar la longitud del código y tomar en consideración que los pasos en direcciones diagonales deben ser multiplicados por un factor igual a raíz cuadrada de dos.

d. Rectangularidad

Una característica de la segmentación de imágenes es la rectangularidad que se presenta como la forma rectangular de la imagen.

e. Proyección

Las proyecciones son representadas por la forma de la región segmentada en los ejes x e y sobre el sistema de coordenadas cartesiano. Cada proyección está definida por un vector unidimensional.

f. Centro de gravedad o centroide

Para la segmentación de imágenes como característica se tomó el centroide. Conocido como el centro de gravedad también o también llamado centroide. Su posición debería ser fija en relación con la forma. Si una forma es representada por su función de región.

3.2.13. Operadores morfológicos

Las operaciones morfológicas simplifican imágenes y tienen como objetivo conservar las principales características de forma de los objetos. Un operador morfológico consiste en el pre-procesamiento de imágenes, el cual descarta, los ruidos y simplifica las formas, descarta la estructura de los objetos mediante la detección de los objetos y finalmente la descripción de los objetos como el área o perímetro [23].

3.2.13.1. Elemento estructural

Se define como un simple parámetro de forma para los filtros u operadores morfológicos.

3.2.13.2. Dilatación

Si algunos de los píxeles adyacentes al píxel probado pertenece al sujeto entonces estudio de píxeles perteneciente al sujeto [23].

3.2.13.3. Erosión

Si todos los pixeles vecinos al pixel de estudio pertenecen al objeto, entonces el pixel de estudio también pertenece al objeto. (Si alguno de los pixeles vecinos al pixel de estudio no pertenece al objeto entonces ese pixel de estudio tampoco pertenece a nuevo objeto) [23].

3.2.13.4. Apertura y clausura

En la apertura el objetivo es suavizar los contornos de una imagen y elimina pequeños salientes. Puede eliminar franjas o zonas de un objeto que sean "más estrechas" que el elemento estructural. Mientras que la clausura elimina pequeños huecos estos rellenándolo y por lo que une componentes conexas cercanas [23].

4. MATERIALES Y MÉTODOS

4.1. TIPOS DE INVESTIGACIÓN

4.1.1. Investigación bibliográfica

Mediante la investigación bibliográfica se realizó la recopilación de la información en fuentes bibliográficas primarias y secundarias de manera eficiente, contribuyendo con el sustento científico para el desarrollo del proyecto de investigación, en especial de la fundamentación teórica.

4.1.2. Investigación de campo

Mediante la investigación de campo se va recopilar información sobre las imágenes de las plantas monocotiledóneas y dicotiledóneas en el lugar de viveros, áreas verdes, invernaderos entre otros, se tomara la foto para obtener la base de datos para la realización del proyecto de investigación. Se basa esta tesis con un método comprobado de recopilación (fotografías), y análisis de la información (creación de la base de datos)

4.1.3. Investigación descriptiva

Se utilizó la investigación descriptiva es detallar el objetivo de estudio (las técnicas de minería de datos usadas en la clasificación automática de plantas monocotiledóneas y dicotiledóneas) donde se detalla la minería de datos, etapas, clasificación por medio de gráficas o imágenes que represente el tema que se pueda tener una idea cabal sobre el estudio de la investigación, incluyendo sus características, sus elementos o propiedades, comportamientos y particularidades.

4.1.4. Investigación aplicada

Se utiliza la investigación aplicada se caracteriza por aplicar los conocimientos que surgen de la investigación pura para resolver el problema de la investigación que es el siguiente: ¿El uso de técnicas de minería de datos permite la clasificación automática de plantas monocotiledóneas y dicotiledóneas?, el cual será de carácter práctico para esto se realiza la verificación a través de la realización de una aplicación web, los beneficiarios sería, el sector agrícola y las personas externas (botánicos o especialistas agrónomos).

4.2. MÉTODO TEÓRICO

Con los métodos teóricos se relaciona al descubrir el objeto de investigación las relaciones esenciales y las cualidades fundamentales de técnicas de minería de datos y plantas monocotiledóneas y dicotiledóneas lo que permitió conceptualizar la realidad. Por tanto, en la investigación se utilizó el método analítico – sintético e hipotético – deductivo

4.2.1. Método analítico - sintético

Este método analítico- sintético tiene la utilidad en el proyecto de investigación con respecto al análisis lo que posibilita descomponer en lo más esencial en relación al objeto de estudio "Técnicas de minería de datos usadas en la clasificación automática de plantas monocotiledóneas y dicotiledóneas" que técnica es la más adecuada para la clasificación automática de las plantas y en el caso de la síntesis las generalizaciones que contribuyan a la solución del problema científico.

4.2.2. Método hipotético - deductivo

Este método se basa en verificar la existencia del problema ¿El uso de técnicas de minería de datos permite la clasificación automática de plantas monocotiledóneas y dicotiledóneas? de esta manera resolver con el desarrollo de la aplicación web donde permita la clasificación automática ya sea mediante la observación del fenómeno de estudio y luego es generar la hipótesis para explicar la existencia de dicho problema anteriormente mencionado y después proponer una solución al problema.

4.2.3. Método comparativo

Mediante este método se va refutar dos técnicas de minería de datos tales como: la regresión logística y la máquina de vectores de soporte, con el objeto de determinar cuál de ellas dispone de mayor rendimiento y eficiencia al momento de la clasificación automática de las plantas monocotiledóneas y dicotiledóneas.

4.3.1. Técnicas de investigación

4.3.1.1. Bibliografía

La técnica de investigación bibliográfica consiste en la revisión de material bibliográfico existente con respecto al tema: "Clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos", esta técnica es uno de los principales pasos para cualquier investigación e incluye la selección de fuentes de información.

4.3.2. Instrumentos de investigación

4.3.2.1. Ficha bibliográfica

Este instrumento de investigación, su función principal es servir como base (relacionado a las técnicas de minería de datos y clasificación automática) y sustento para anotar las fuentes que serán consultadas al momento de realizar un trabajo, estas pueden ser libros, guías, revistas, folletos, artículos científicos.

4.3.2.2. Cámara fotográfica

Este instrumento ayuda en el proyecto de investigación debido a que se encarga capturar y representar imágenes, aquí se toma las fotos de las hojas de las plantas monocotiledóneas y dicotiledóneas para hacer la base de datos que se va ocupar para posteriormente realizar la clasificación automática.

4.4. MÉTODO ESPECÍFICO

4.4.1. Metodología CRISP-DM

Para la utilización del proyecto se basó en la metodología CRISP-DM, porque se basa principalmente en la minería de datos por este motivo se optó para el desarrollo. Por lo cual se va desarrollar mediante 6 fases cada una con subtareas. En la primera fase se dará la comprensión del negocio con los objetivos y requerimiento del negocio, segunda fase se va dar la comprensión de los datos en el cual se va recopilar y familiarizar los datos este caso la base de datos de las fotos de las plantas angiospermas y gimnospermas, en la tercera fase la preparación de los datos una vista de datos para la minería de datos ,la cuarta fase el modelado consiste en elegir los parámetros ya un modelos para el desarrollo de la minería de datos, la quinta fase la evaluación consiste en comprobar si el modelo se ajusta los datos de entrada y finalmente el despliegue.

5. ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

Para el desarrollo del método específico se optó por la metodología CRISP-DM por motivo que se basa en el “Proceso estándar de la industria cruzada para la minería de datos” donde se el proyecto de investigación a realizarse y también como la mayoría de las investigaciones y proyectos del área de minería de datos adopta el estándar CRISP-DM para su desarrollo, en el caso del proyecto de investigación clasificación automática de plantas monocotiledóneas y dicotiledóneas mediante minería de datos, los proyectos de reconocimiento de patrones no existe un proceso metodológico

que describa de manera estructurada la planificación de un proyecto a partir de las etapas que un sistema de reconocimientos de patrones precisa.

5.1. METODOLOGÍA CRISP-DM

5.1.1. Comprensión del negocio

5.1.1.1. Objetivos comerciales

En la etapa de clasificación de las plantas se utiliza la agricultura convencional para este proceso realizado en el campo. En efecto, la clasificación supone la generación automática de clasificación a través de la minería de datos con la utilización de algoritmos que satisfagan la necesidad donde permita clasificar a dos tipos de plantas: monocotiledóneas y dicotiledóneas. Por lo cual se propone el objetivo de:

- Clasificar plantas monocotiledóneas y dicotiledóneas usando técnicas de minería de datos como la regresión logística y la máquina de soporte vectorial para el reconocimiento automático en imágenes digitales.

Se considera que la aproximación propuesta es de:

- La tasa de error de clasificación de plantas monocotiledóneas y dicotiledóneas en imágenes digitales es inferior a 0.02%.

5.1.1.2. Situación actual y objetivos

Para el desarrollo de la proyecto de investigación dispone de un total de 5 especies de plantas monocotiledóneas y 5 especies de plantas dicotiledóneas por lo que en cada especie se toma una cantidad de 10 fotografías, fueron tomadas en condiciones de fondo blanco para una visualización de la hoja, en un huerto de casa y viveros.

A través de técnicas de minería de datos y procesamiento de imágenes se pretende extraer las imágenes, los píxeles que identifican la capa vegetal y a continuación construir un clasificador que permita distinguir dos tipos de plantas monocotiledóneas y dicotiledóneas basándose características geométricas ancho y altura también el color de la hoja.

En la técnica de minería de datos el clasificador construido se integrará en una aplicación web que permita seleccionar la base de datos de la plantas monocotiledóneas y dicotiledóneas y a continuación muestre a qué clase de planta pertenece, así como el porcentaje de validación sobre el algoritmo utilizado para la clasificación.

5.1.1.3. Objetivos de minería de datos

Los objetivos comerciales enunciados anteriormente se detallan a continuación en relación a la minería de datos:

- ✓ Emplear las imágenes digitales mediante la utilización de una base de datos de plantas monocotiledóneas y dicotiledóneas para la extracción de zonas de interés.
- ✓ Generar un modelo de clasificación o reconocimiento de patrones permitiendo la clasificación automática de las plantas monocotiledóneas y dicotiledóneas.
- ✓ Integrar el clasificador en una aplicación web que permita obtener el porcentaje de validación del algoritmo de minería de datos y el tipo de la planta monocotiledónea o dicotiledónea

5.1.1.4. Plan del proyecto

Tabla 7. Plan de proyecto

ID	Nombre de la tarea	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
1	Comprensión del negocio													
2	Comprensión de datos													
3	Preparación de datos													
4	Segmentación													
5	Extracción de características													
6	Modelado-clasificación													
7	Evaluación													
8	Despliegue													

Elaborado por: Equipo de trabajo

En la Tabla 7 detalla el tiempo de desarrollo de cada etapa de la metodología CRISP-DM, el tiempo estimado para cumplir con cada una de las fases.

a. Valoración de herramientas y técnicas

Para generar el modelo de clasificación de plantas monocotiledóneas y dicotiledóneas en imágenes digitales se ha resuelto desarrollar una aplicación web con lenguaje de programación Python que cuenta con las siguientes características

- Es un lenguaje interpretado, no compilado, usa tipado dinámico.
- Es multiplataforma, lo cual es ventajoso para hacer ejecutable su código fuente entre varios sistema operativos.

- En Python, el formato del código es estructural.
- Tiene poderosas bibliotecas estadísticas y numéricas como Pandas, Numpy, Matplotlib, SciPy, scikit-learn.

5.1.2. Comprensión de los datos

5.1.2.1. Adquisición de datos

Para la comprensión de los datos se creó una base de datos con imágenes formato jpg con una resolución inicial de 1600* 1200 píxeles, las imágenes se tomaron un fondo blanco.

5.1.2.2. Segmentación

En la segmentación es donde la discriminación de la capa vegetal respecto al resto de elementos en la imagen que formarían el fondo. Para el funcionamiento es importante que esta fase sea adecuada en todo el sistema de clasificación que se está desarrollando. La imagen de entrada en modo RGB (verde) es convertida a una imagen binaria (blanco y negro), donde la capa vegetal es negra mientras que el fondo es blanco ejemplo se muestra en la **Figura12.**

(a) (b)

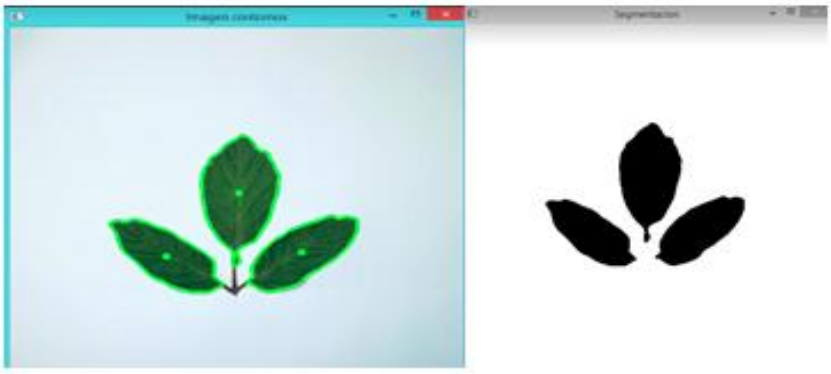


Figura 12. Ejemplo de segmentación de la capa vegetal.
(a) Imagen original en RGB, (b)

5.1.2.3. Descripción del proceso de segmentación

Para realizar la segmentación de la imagen digital se pasó la imagen de BGR a Gray y HSV (cambio a escala de gris) como se muestra en las siguientes líneas de código:

```
image_gray = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
image_gray = cv2.cvtColor(image_gray, cv2.COLOR_GRAY2BGR)
image_HSV = cv2.cvtColor(image, cv2.COLOR_BGR2HSV)
```

Luego se procedió a detectar el color verde mediante lo siguiente:

```
mask_green = cv2.inRange(image_HSV, low_green, tall_gren)
mask = cv2.add(mask_green, mask_green)
```

```
mask = cv2.medianBlur(mask, 7)
green_detected = cv2.bitwise_and(image, image, mask=mask)
image_gray2 = cv2.cvtColor(green_detected, cv2.COLOR_BGR2GRAY)
ret, th = cv2.threshold(image_gray2, 0.75,255,cv2.THRESH_BINARY_INV)
```

Una vez realizada la detección del color verde como se muestra en la **Figura 13.**

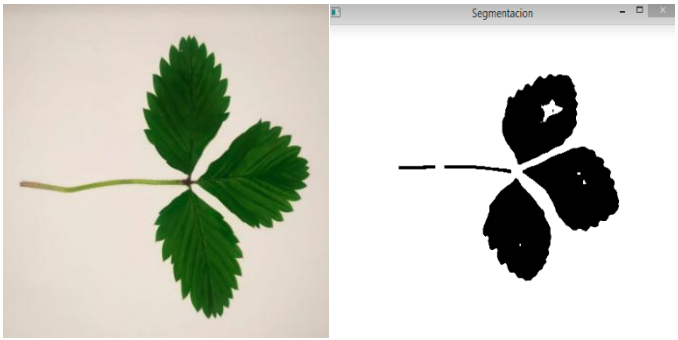


Figura 13. Detección del color verde- planta dicotiledónea

Se utilizó un operador morfológico que es la dilatación con esto los objetos se van hacer más visibles y se va eliminar el tallo para hacer uso sólo de la forma de la hoja como muestra en la **Figura 14.** La siguiente línea de código detalla cómo se utilizó el operador morfológico-dilatación:

```
img_dilate = cv2.dilate(th, kernel, iterations=1)
cv2.imshow("Dilatacion", img_dilate)
th_inv = cv2.bitwise_not(img_dilate)
```

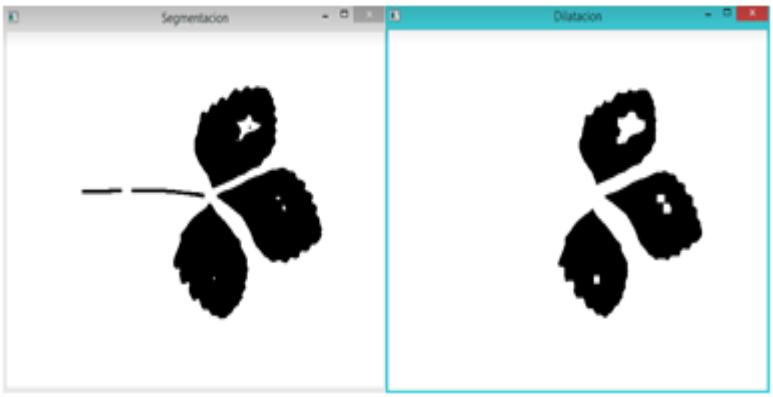


Figura 14. Segmentación con el operador morfológico - dilatación

Después la detección de contorno, donde se dibuja una línea del contorno de la hoja de la planta monocotiledónea y dicotiledónea. Para encontrar y dibujar el contorno de la hoja de la planta. Como se muestra en la **Figura 15.** Se empleó dos funciones: findContours, este permite encontrar el contorno y drawContours permite dibujar el contorno a través del siguiente línea de código:

```
contorns, jerarquy1 = cv2.findContours(th_inv,cv2.RETR_EXTERNAL,cv2.CHAIN_APPROX_SIMPLE)
```

cv2.drawContours (image, contorns,-1,(0,255,0),3)

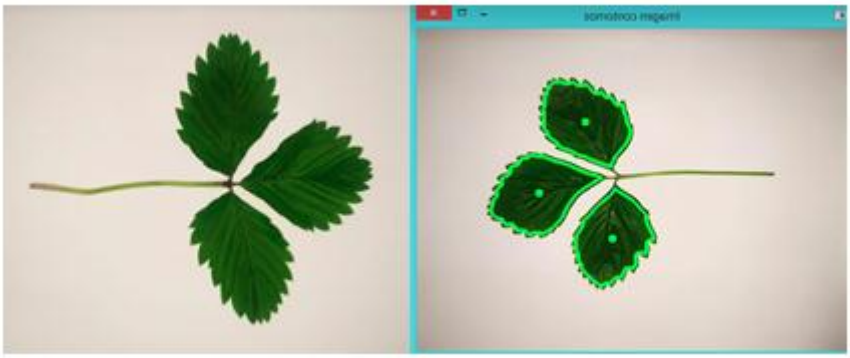


Figura 15. Segmentación por contorno de la hoja

5.1.2.4. Extracción de características

Para la extracción de características se basó en las siguientes características:

- Área. Número de píxeles en la región.
- Perímetro. Es la distancia entre cada par de píxeles adyacentes en el borde de la región.
- Centro de gravedad también o también llamada centroide. Su posición debería ser fija en relación con la forma.

El área del contorno puede calcularse usando la función cv2.contourArea (), o puede usar el momento (momento 0), M ['m00']. En el perímetro de contorno o también se llamado longitud de arco. Se puede calcular utilizando la función cv2.arcLength (). El segundo parámetro de esta función se puede usar para especificar si la forma del objeto está cerrada (Verdadero) o abierta (una curva). Para obtener el centro de gravedad se utilizó dos elementos: el método cv2.circle () el cual se usa para dibujar un círculo de la hoja de la planta y el center_coordinates donde da las coordenadas del centro del círculo. Estas coordenadas se representan como tuplas de dos valores, es decir (valor de la coordenada X , valor de la coordenada Y).

5.1.2.5. Aprendizaje supervisado- minería de datos

Para construir el conjunto de entrenamiento se requiere un grupo de pares de patrones que debe tener entrada – salida deseada. Por lo que se añade un nuevo atributo denominado "tipo" que representará (monocotiledónea o dicotiledónea) se asignará manualmente a los elementos del conjunto de entrenamiento después de un análisis visual de la región por parte de un experto. En la **Figura 16** se muestra un detalle del archivo generado en formato csv, para el conjunto de entrenamiento.

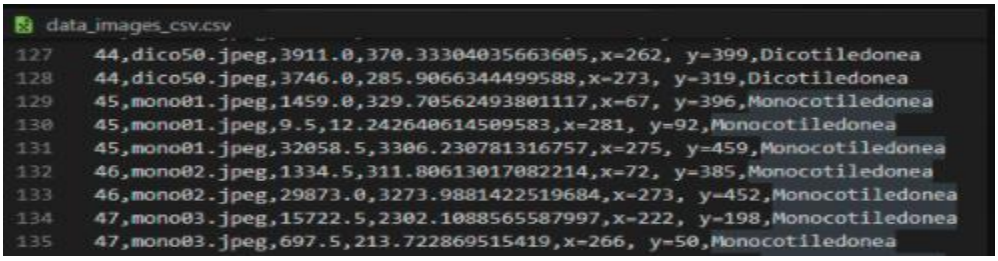


Figura 16.Base de datos de plantas monocotiledóneas y dicotiledóneas -Formato csv

5.1.2.6. Recopilación de datos

Los datos recolectados provienen de los vectores de características de cada una de las regiones capturados automáticamente mediante las siguientes funciones:

area = cv2.contourArea(cnt),

perimetro= cv2.arcLength(cnt,True)

cv2.circle(image,(cX,cY),5,(0,255,0),-1)

Esto fue implementado en Python y almacenados posteriormente en formato csv. Por otra parte, los cuatro atributos (área, perímetro, centroide y tipo) seleccionados son adecuados para la obtención de un clasificador. Así, las monocotiledóneas tienden a ser alargadas frente a las dicotiledóneas que presentan una superficie más redondeada. También es importante resaltar que la clasificación manual de las regiones para la construcción del conjunto de entrenamiento se realizó sobre el mismo tipo de imágenes que serán la entrada al clasificador.

5.1.2.7. Descripción de los datos

En cuanto a los datos de partida, se dispone de 4 atributos y 353 registros, de forma que la mayoría de los datos son de tipo numérico a excepción del atributo tipo que es categórico. En la Tabla 8 se muestra la descripción de cada atributo, el tipo de datos asociado y el valor mínimo y máximo obtenido de las 97 regiones.

Tabla 8. Tipo de datos asociado y rango de valores para cada atributo

Atributo	Formato	Valor mínimo	Valor máximo
Área	numérico	44	36710
Perímetro	numérico	29,3137085	5617,47085
Centroide	numérico	x=79 y=139	x=364 y=472
Tipo	categórico	Ninguno	Ninguno

Elaborado por: Equipo de trabajo

5.1.2.8. Exploración de datos

En esta etapa de exploración de datos se separan en dos grupos los atributos: variables cuantitativas o numéricas y variables cualitativas o categóricas con la finalidad de efectuar un análisis previo de los datos y asegurar la calidad de los mismos.

En las variables cuantitativas (área, perímetro y centroide) se consideró analizar los datos la media, el límite inferior y superior, la mediana, la desviación típica, el valor mínimo y máximo, y finalmente el rango con respecto al atributo de clase.

a. Variables cuantitativas

Los datos presentados en la Tabla 9, detalla la los atributos con respecto a la clase monocotiledóneas. Los datos muestran para monocotiledónea, el área promedio entre 611.5 y 9723.5 pixeles, en el perímetro 159.72286 y 1052.87719 unidades.

Tabla 9. Exploración de datos clase monocotiledóneas.

Atributo	Mínimo	Máximo	Mediana
Área	611.5	9723.5	5167.5
Perímetro	159.72286	1052.87719	606.300025

Elaborado por: Equipo de trabajo

Los datos presentados en la Tabla 10**,** detalla los mismos resultados del mismo análisis para el caso de dicotiledóneas. Se puede observar una variación en el área entre 2740.0 y 23599.0 pixeles, con respecto al perímetro 188.793936 y 644.121928 unidades.

Tabla 10. Exploración de datos clase dicotiledóneas.

Atributo	Mínimo	Máximo	Mediana
Área	2740.0	23599.0	13169.5
Perímetro	188.703936	644.121923	416,41293

Elaborado por: Equipo de trabajo

b. Variables cualitativas

Para la variable cualitativa, el tipo de datos se analizaron a través de la frecuencia simple como se aprecia en la Tabla 11. La variable cualitativa tipo contempla tres categorías: dicotiledóneas, monocotiledóneas con 49.85% y 50.15% respectivamente.

Tabla 11. Exploración de datos variable cualitativa

Tipo	Frecuencia	Porcentaje sobre el total
Dicotiledónea	176	49.85
Monocotiledónea	177	50.15

Elaborado por: Equipo de trabajo

5.1.2.9. Verificación de la calidad de datos

Para verificar la calidad de los datos se efectuó un análisis previo durante las etapas de descripción y exploración. En este marco se consideró solo el tipo de categoría con el nombre "monocotiledónea y "dicotiledónea".

Datos Perdidos: a tenor de lo expuesto, existe un 0% de datos perdidos tanto cuantitativos como cualitativos.

5.1.3. Preparación de los datos

5.1.3.1. Selección de datos

Para la selección de datos se va realizar la integración, recopilación, la selección de datos y atributos. Se detalla a continuación:

a. Selección de registros o filas

Para la selección de filas se eligieron 97 registros de los cuales son de tipo monocotiledónea y dicotiledónea. Para estos registros se validaron las imágenes según el tipo de la planta anteriormente mencionada.

b. Selección de atributos

Para la selección de atributos se eligió área, perímetro, centroide y tipo. Los atributos son de tipo numérico mientras que tipo es una carácter se monocotiledónea o dicotiledónea.

5.1.3.2. Limpieza de datos

En la limpieza de datos se verificó los valores perdidos, erróneos. Mediante la transformación de valores binarios. Por lo que se verificó que todas las imágenes deben ser correctas.

5.1.3.3. Construcción de nuevo datos

En la obtención del clasificador automático no se consideró la creación ni sustitución de atributos, porque los seleccionados cumplen con el objetivo de clasificar.

5.1.3.4. Formato de datos

En el formato de datos para la construcción del modelo de clasificación con técnicas de minería de datos, es un requisito que los datos estén clasificados en la etapa de entrenamiento.

5.1.4. Modelado

La finalidad de la etapa de modelado es extraer conocimiento a partir de los datos recogidos de la imagen, identificando patrones útiles y comprensibles.

5.1.4.1. Técnica de modelado

Se realizan dos comparaciones de técnicas de minería de datos como: regresión logística y máquina de vector de soporte, para verificar cual es el funcionamiento de los mismos.

a. Regresión logística

Para la creación de la regresión logística se crea x_train consiste en incluir todas sus variables independientes, estas se utilizarán para entrenar el modelo, este medio de observaciones de sus datos completos se utilizará para entrenar / ajustar el modelo y el resto se utilizará para probar el modelo.test_size = 30% y 70%.

El x_test que permite la parte restante de las variables independientes de los datos que no se utilizarán en la fase de entrenamiento y se utilizarán para hacer predicciones para probar la precisión del modelo.

El y_train consiste en la variable dependiente que necesita ser predicha por este modelo, esto incluye etiquetas de categoría contra sus variables independientes y finalmente el y_test que permite etiquetas de categoría para sus datos de prueba, estas etiquetas se utilizarán para probar la precisión entre las categorías reales y predichas como se muestra en la **Figura 17**.

```
X_train, X_test, y_train, y_test = train_test_split(
    x, y, test_size=0.3, random_state=0)
```

Figura 17. Entrenamiento y prueba del algoritmo

En el siguiente código **Figura 18,** se realiza una instancia del modelo. Aquí todos los parámetros no especificados se establecen en sus valores predeterminados.

El entrenamiento del modelo se realiza en una sola llamada de método llamada ajuste consiste en los dos primeros parámetros del método de ajuste especifican las características y el resultado del conjunto de datos de entrenamiento.

```
model = LogisticRegression()
model.fit(X_train, y_train)
```

Figura 18.Instancia del modelo

En la **Figura 19** detalla la predicción de la regresión logística mediante el área y perímetros.

```
data_new = {'Area': [area],
            'Perimetro': [perimetro]}

y_pred = model.predict(X_test)

df2 = pd.DataFrame(data_new, columns=["Area", "Perimetro"])
prediction = model.predict(df2)
print("Prediccion", prediction)

score = accuracy_score(y_test, y_pred)
print("Logistica %", score)
```

Figura 19.Prediccion de la regresión logística

b. Máquina de vectores de soporte

Se realiza el mismo procedimiento de la **Figura 20,** para el entrenamiento y las pruebas**.** Se utilizó la clf = svm.SVC (kernel='rbf'). Cada una de estas funciones tiene sus características, pros y contras y su ecuación, pero como no hay una forma sencilla de saber la función que mejor funciona, elegimos utilizar diferentes funciones y comparar los resultados. Utilicemos la función por omisión, RBF (Función Basada en Radio).

```
clf = SVC(kernel="rbf").fit(X_train, y_train)
print("SVM %", clf.score(X_test, y_test))
```

Figura 20. Predicción del modelo de máquina de vectores de soporte

5.1.4.2. Ejecución del modelo

Para la ejecución del modelo una vez implementado el modelo de regresión logística y de máquina de vector de soporte se realiza la ejecución del modelo para la predicción de clasificación automática en base a tipo de monocotiledónea y dicotiledónea.

a. **Regresión logística**

Mediante la siguiente línea de código **Figura 21** permite llamar al modelo entrenado para que se realice la predicción de regresión logística. Y la función permite guardar el modelo entrenado.

```
def category(self):
    model = joblib.load("./models/logistic.pkl")
    data_new = {'Area': [self.area],
                'Perimetro': [self.perimetro]}

    df2 = pd.DataFrame(data_new, columns=["Area", "Perimetro"])
    prediction = model.predict(df2)
    print(prediction)
    return str(prediction).replace("['","").replace("']","")
```

Figura 21. Categoría de regresión logística

Se observa en la **Figura 22** como se realiza la predicción de la regresión logística.

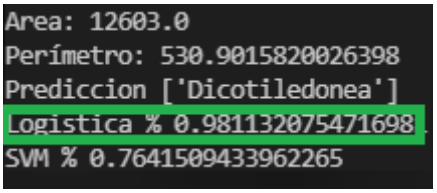


Figura 22. Predicción de la regresión logística

b. MSV

Mediante la siguiente línea de código **Figura 23** permite llamar al modelo entrenado para que se realice la predicción de regresión logística. Y la función permite guardar el modelo entrenado.

```
def score_svm(self, X_test, y_test):
    model_svm = joblib.load("./models/svm.pkl")
    score = model_svm.score(X_test, y_test)
    return score
```

Figura 23. Categoría de MSV

Se observa en la **Figura 24** como se realiza la predicción de la MSV

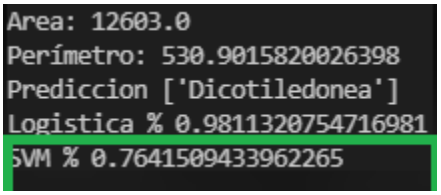


Figura 24. Predicción de la MSV

5.1.5. Evaluación

Para medir la calidad de las dos técnicas de minería de datos se empleó una técnica de validación como: la validación cruzada. Con la cantidad de 353 registros de plantas en la ***Figura 25***.

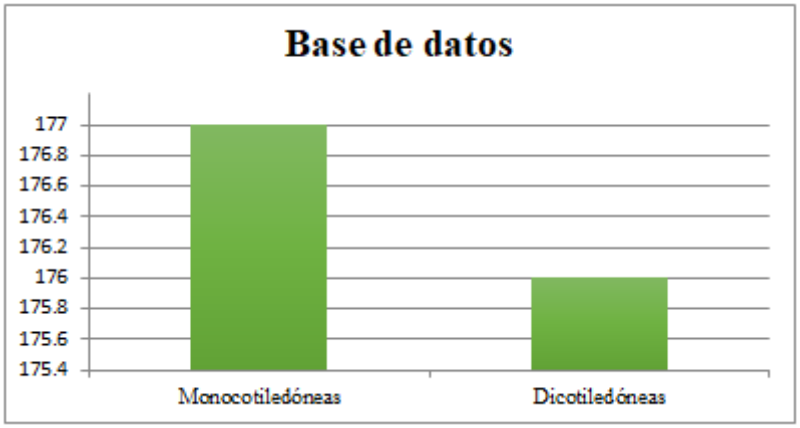


Figura 25. Estadística de base de datos

5.1.5.1. Validación cruzada con regresión logística

Se utilizó la técnica de la validación cruzada para el cual se separó en conjunto de datos inicial de 353 registros que se agruparon en 2: conjunto de entrenamiento (train) 70% y conjunto de pruebas 30% (test).

Tabla 12. Cantidad de datos para la validación de la regresión logística

Datos	Entrenamiento	Pruebas	Total
Cantidad de datos	248	105	353
Porcentaje	70%	30%	100%

Elaborado por: Equipo de trabajo

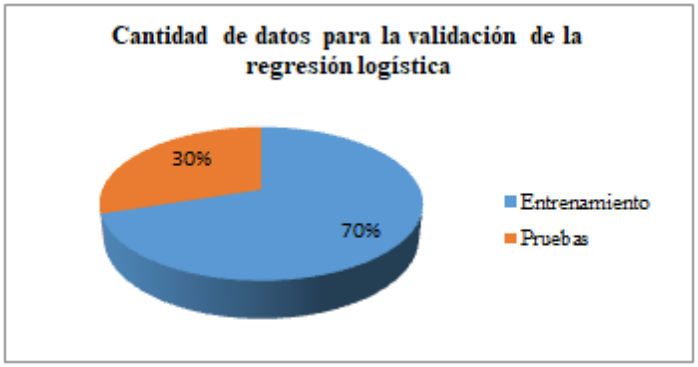


Figura 26. Cantidad de datos para la validación cruzada de regresión logística

a. Iteración 1

Tabla 13. Validación cruzada de la primera iteración en RL

Resultados Iteración N° 1:	
Steps/Epoch:	20
Accuracy:	84.8%
Loss:	15.2%

Elaborado por: Equipo de trabajo

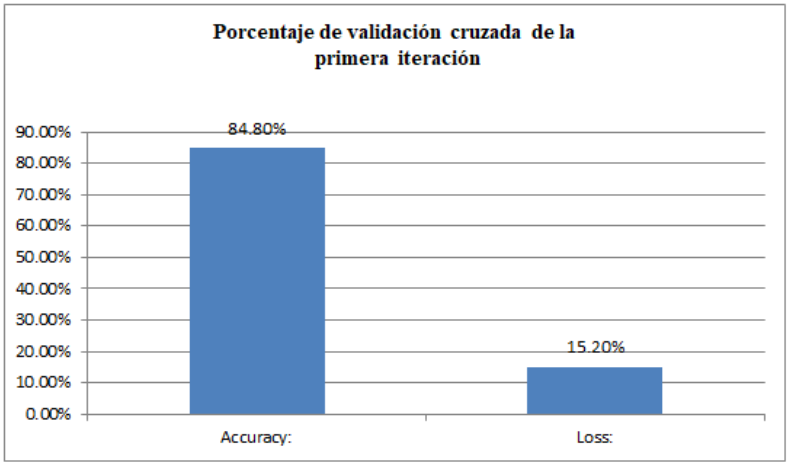


Figura 27. Porcentaje de validación cruzada de la primera iteración en RL

b. Iteración 2

Tabla 14. Validación cruzada de la segunda iteración en RL

Resultados Iteración N° 2:	
Steps/Epoch:	15
Accuracy:	88.8%
Loss:	11.2%

Elaborado por: Equipo de trabajo

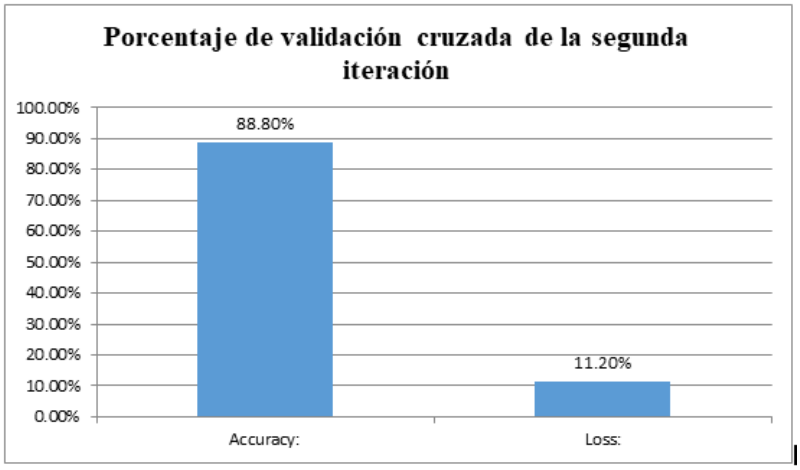


Figura 28. Porcentaje de validación cruzada de la segunda iteración en RL

c. Iteración 3

Tabla 15. Validación cruzada de la tercera iteración en RL

Resultados Iteración N° 3:	
Steps/Epoch:	10
Accuracy:	92.8%
Loss:	7.2%

Elaborado por: Equipo de trabajo

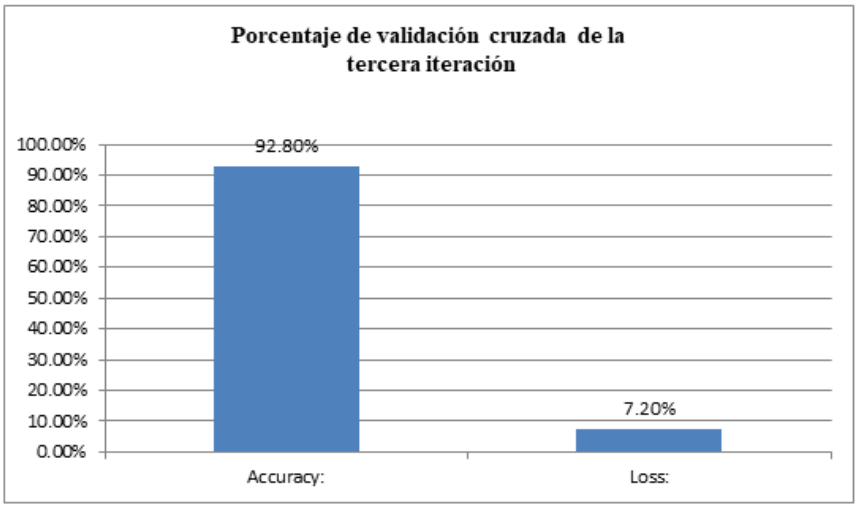


Figura 29. Porcentaje de validación cruzada de la tercera iteración en RL

d. Iteración 4

Tabla 16. Validación cruzada de la cuarta iteración en RL

Resultados Iteración N° 4:	
Steps/Epoch:	5
Accuracy:	97.1%
Loss:	2.9%

Elaborado por: Equipo de trabajo

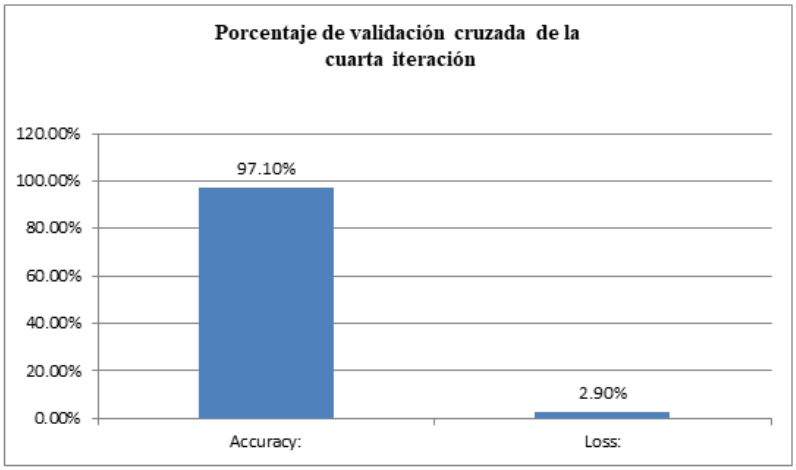


Figura 30. Porcentaje de validación cruzada de la cuarta iteración en RL

Una vez realizado la validación cruzada durante las 4 iteraciones nos dio un margen de error de 2.9% de la regresión logística y una precisión de 97.1%, el cual mide la precisión de eficiencia de los datos seleccionados en para la base de datos.

5.1.5.1. Validación cruzada con SVM

Se utilizó la técnica de la validación cruzada para el cual se separó en conjunto de datos inicial de 353 registros que se agruparon en 2: conjunto de entrenamiento (train) 70% y conjunto de pruebas 30% (test).

Tabla 17. Cantidad de datos para la validación de la SVM

Datos	Entrenamiento	Pruebas	Total
Cantidad de datos	247	105	352
Porcentaje	70%	30%	100%

Elaborado por: Equipo de trabajo

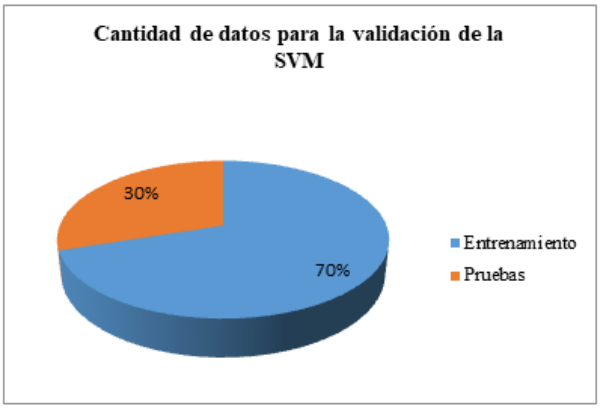


Figura 31. Cantidad de datos para la validación de la SVM

a. Iteración 1

Tabla 18. Validación cruzada de la primera iteración en SVM

Resultados Iteración N° 1:	
Steps/Epoch:	20
Accuracy:	59.26%
Loss:	17.56%

Elaborado por: Equipo de trabajo

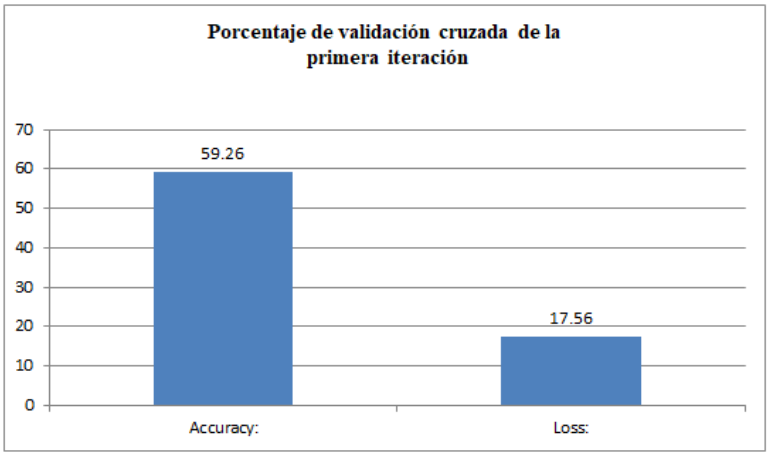


Figura 32. Porcentaje de validación cruzada de la primera iteración en SVM

b. Iteración 2

Tabla 19. Validación cruzada de la segunda iteración en SVM

Resultados Iteración N° 2:	
Steps/Epoch:	15
Accuracy:	64.74%
Loss:	20.96%

Elaborado por: Equipo de trabajo

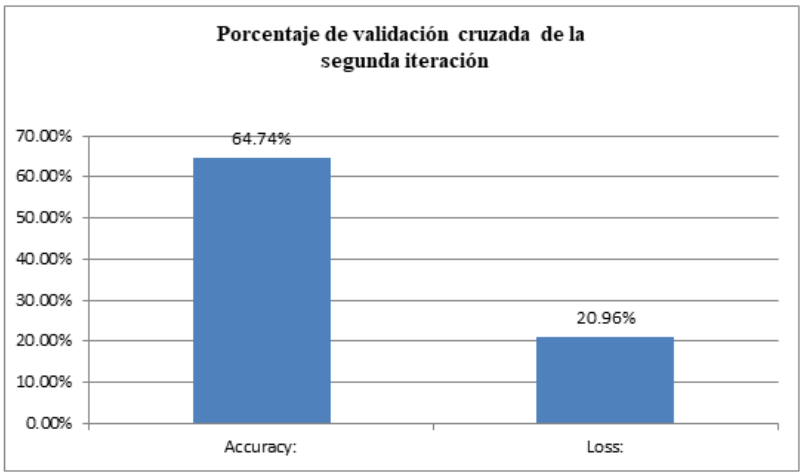


Figura 33. Porcentaje de validación cruzada de la segunda iteración en SVM

c. Iteración 3

Tabla 20. Validación cruzada de la tercera iteración en SVM

Resultados Iteración N° 3:	
Steps/Epoch:	10
Accuracy:	68.15%
Loss:	23.22%

Elaborado por: Equipo de trabajo

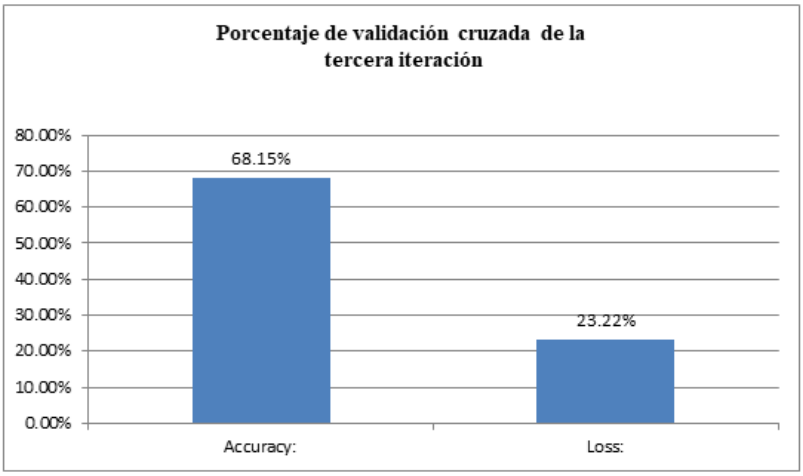


Figura 34. Porcentaje de validación cruzada de la tercera iteración en SVM

d. Iteración 4

Tabla 21. Validación cruzada de la cuarta iteración en SVM

Resultados Iteración N° 4:	
Steps/Epoch:	5
Accuracy:	74.51%
Loss:	27.76%

Elaborado por: Equipo de trabajo

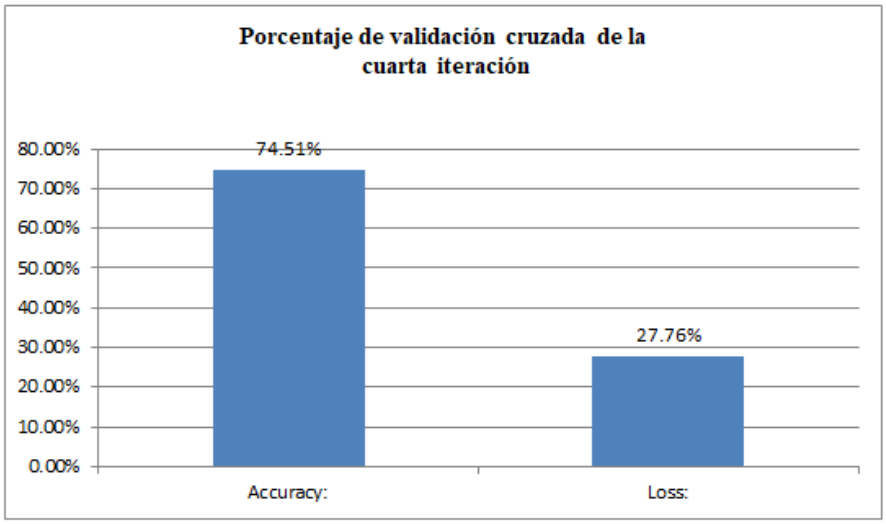


Figura 35. Porcentaje de validación cruzada de la cuarta iteración en SVM

Una vez realizado la validación cruzada durante las 4 iteraciones nos dio un margen de error de 27.76% de la regresión logística y una precisión de 74.51%, el cual mide la precisión de eficiencia de los datos seleccionados en para la base de datos.

5.1.6. Distribución

Una vez evaluado los clasificadores con técnicas de minería de datos (regresión logística y máquina de vectores de soporte). El objetivo de esta etapa es integrar en una aplicación web. La aplicación desarrollada se denomina "Prototipo de investigación" y la información que se obtiene a partir de ella puede ser usada para la elaboración de un inventario de plantas.

La aplicación, que se desarrolló en Python, emplea a un conjunto de imágenes de partida para efectuar las pruebas necesarias. Las salidas que tienes son:

- Identificación de dicotiledóneas
- Identificación de monocotiledóneas

Para la identificación se emplea el porcentaje de área de cada una de las clases expresadas en píxeles. El porcentaje solo es de la capa vegetal, la cual clasifica de manera automática en monocotiledóneas y dicotiledóneas. En su salida se pintan de color verde el borde de la hoja para poder clasificar.

En la **Figura 36**, se aprecia la imagen seleccionada, el cual se pinta de color verde el borde de la hoja, indica el tipo que pertenece este caso es monocotiledónea y finalmente da el margen de porcentaje en los dos tipos de técnicas de minería de datos: la primera regresión logística con 97.75% y la segunda con 73.03%, lo que clasifica de manera automática las plantas.

Figura 36. Clasificación automática de plantas monocotiledóneas y dicotiledóneas

5.1.6.1. Requerimientos técnicos de la aplicación

a. Hardware

- Cámara fotográfica
- Procesador: Intel Core I5
- Disco duro: 250 GB
- Memoria RAM: 8 GB

b. Software

- Sistema operativo: Ubuntu 20.04 lts
- Base de datos: Posgresql 12
- Lenguaje de Programación: Python versión 3.7, java script css3 html5
- Framework: Django 3.2 , flask y doker.
- Plataforma del Servicio: aws (Amazon Web Service) s2

5.1.6.2. Funcionalidades

La aplicación web "Prototipo de investigación" ofrece funcionalidades básicas que permiten la carga e identificación de las plantas monocotiledóneas y dicotiledóneas.

Estas funcionalidades son:

- Iniciar sesión
- Carga de imagen digital en la aplicación web.
- Clasificación automática de la planta.
- Salir de la aplicación web.

5.1.6.3. Uso de la herramienta

a. Ejecución de la aplicación

En la ejecución de la aplicación es necesario comenzar desde Python por el comando python manage.py Cuando te lo pida, escribe tu nombre de usuario (en minúscula, sin espacios), email y contraseña, una vez ingresado nos da como resultado la página principal de la aplicación **http://127.0.0.1:8000/** como se muestra en la **Figura 37.**

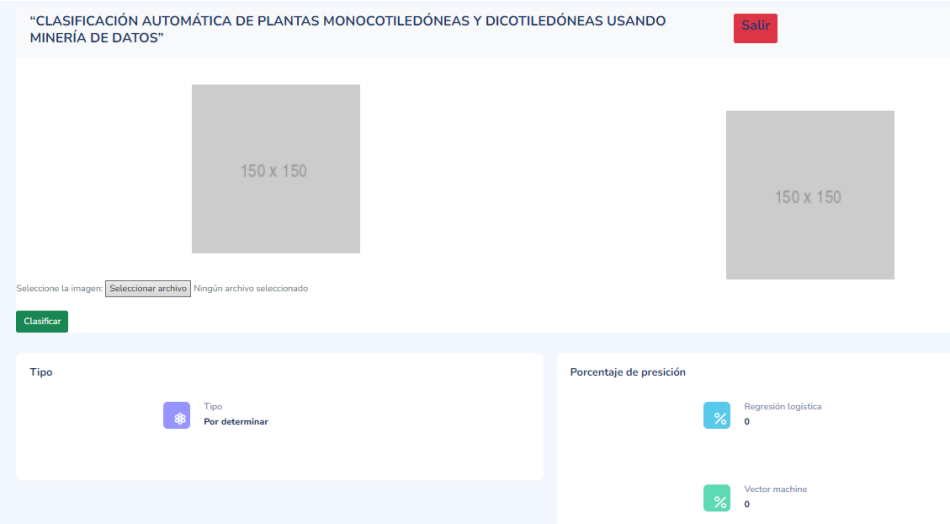


Figura 37. Ventana principal de la aplicación web

A continuación se describe de forma ordenada cada una de las opciones disponibles de la aplicación web.

b. Ventana inicio de sesión

En la ventana principal de la aplicación móvil en la **Figura 38** del inicio de sesión se debe colocar el usuario y la contraseña creada.

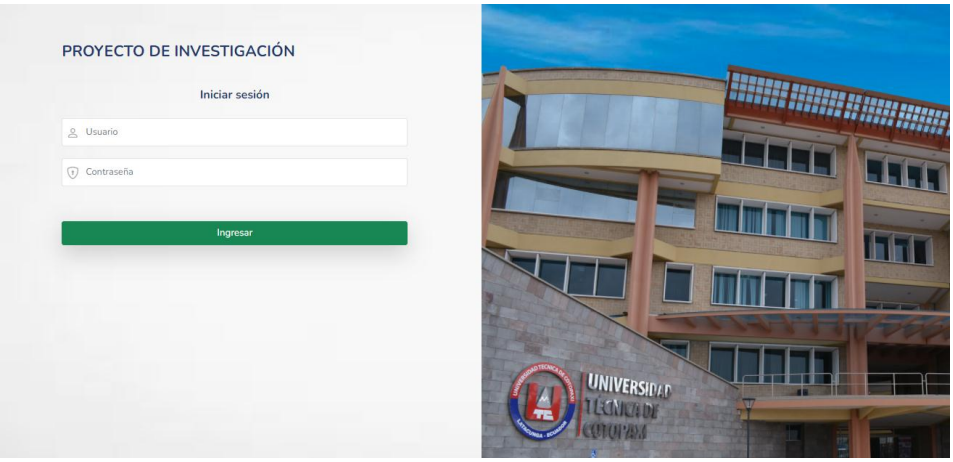


Figura 38. Ventana de inicio de sesión

c. Cargar una imagen

Para el efectuar la carga de una imagen en la aplicación web se debe seguir los siguientes pasos. Elegir el botón de "seleccionar carga", el cual dará lugar a la aparición de una ventana de exploración para buscar la imagen como se muestra en la **Figura 39.**

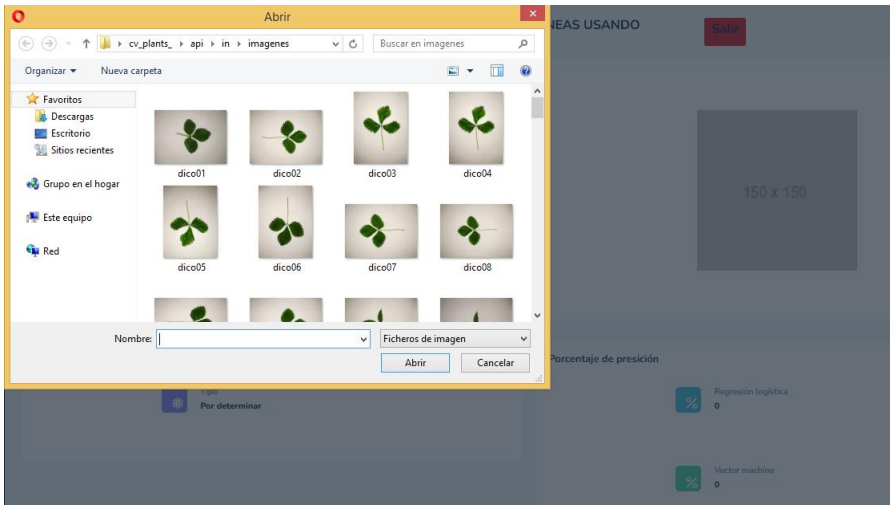


Figura 39.Ventana de exploración para abrir una imagen

Seleccionar la imagen de tipo jpg en los directorios y dar click en el botón abrir. Aparecerá el nombre de la imagen seleccionada **Figura 40.**

Figura 40. Imagen cargada

d. Clasificar de forma automática de plantas monocotiledóneas y dicotiledóneas

Para clasificar las plantas monocotiledóneas y dicotiledóneas mediante el borde de la hoja, una vez cargada la imagen, aparecerá en la página principal, en la cual se debe seleccionar el botón de "clasificar". A continuación aparece la imagen de la planta con el borde, el tipo de planta que pertenece, el porcentaje del algoritmo de regresión logística y finalmente el porcentaje del algoritmo de vector machine en plantas monocotiledóneas **Figura 41** y en plantas dicotiledóneas **Figura 42**

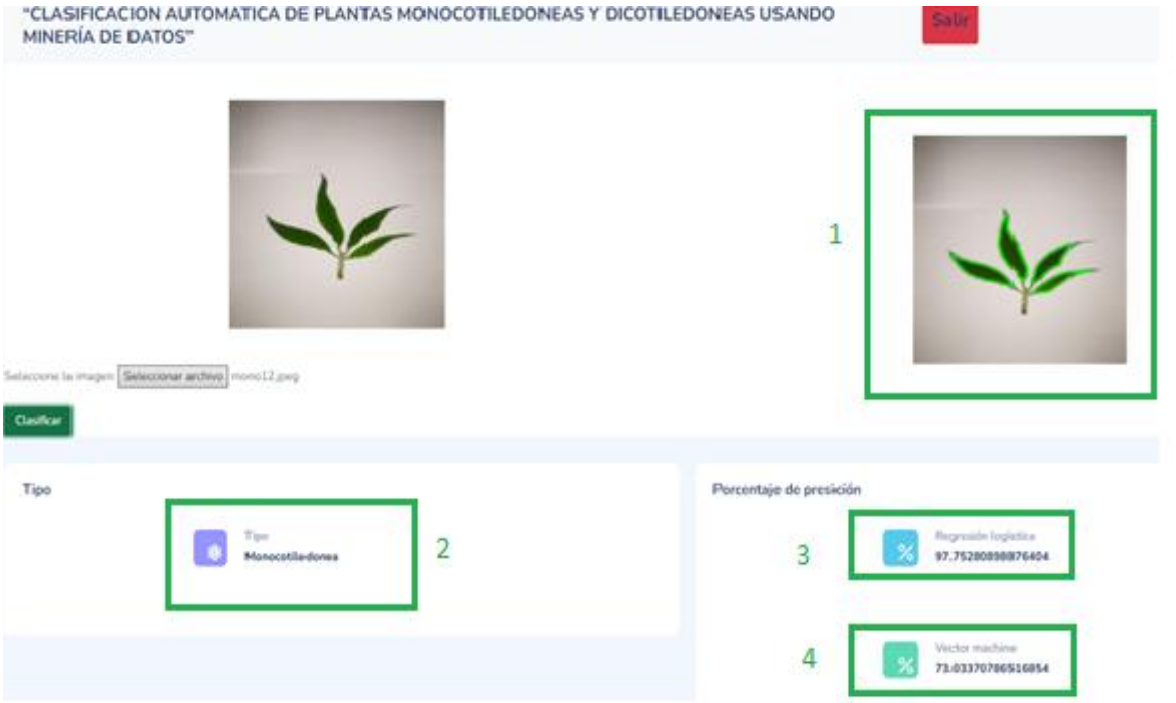


Figura 41. Clasificación de la planta monocotiledónea

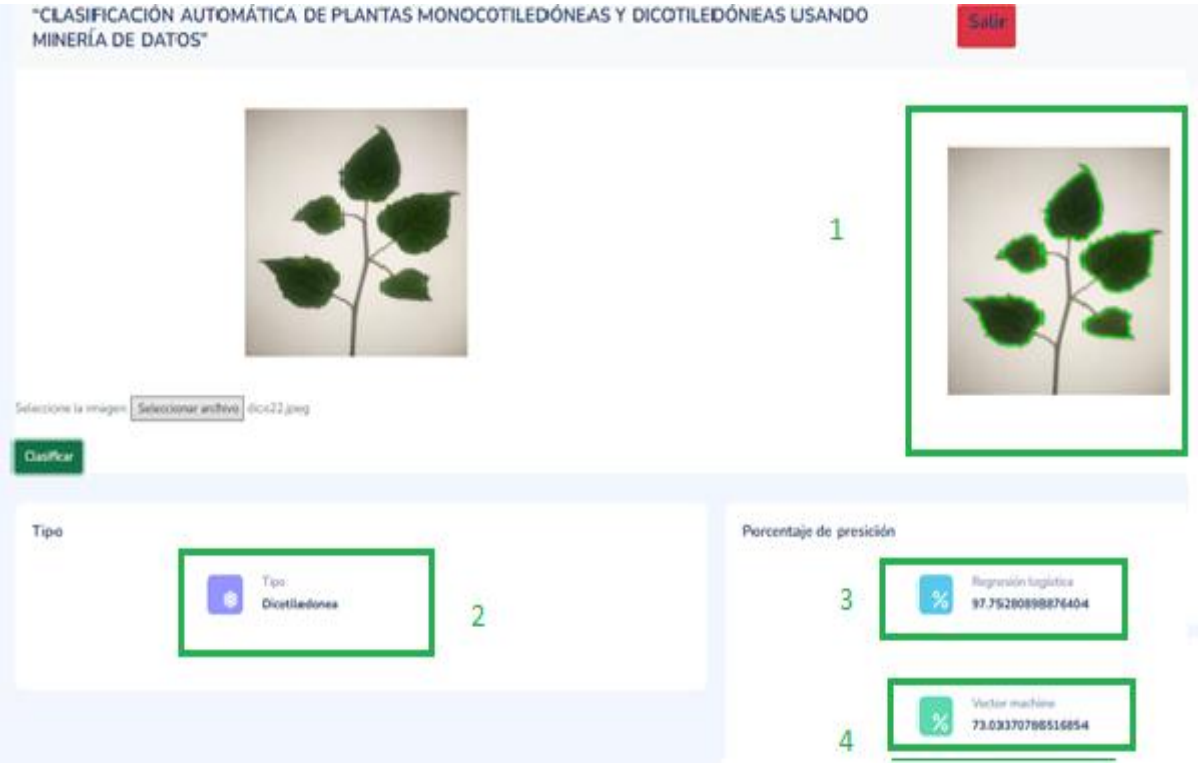


Figura 42.Clasifiacion de la planta dicotiledónea

e. Salir de la aplicación móvil

Esta opción permite la salida del sistema, para lo cual se debe dar click en el botón "salir" como se muestra en la **Figura 43.**

Figura 43. Salida de la aplicación web

5.1.6.4. Funciones principales

En la **Tabla 14** se enumeran las funciones principales desarrolladas a lo largo del proyecto de investigación.

Tabla 14. Librerías utilizadas para el desarrollo de la aplicación web

Nombre	Tarea
Cv2	La imagen se trata ahora como una matriz con valores de filas y columnas almacenados en img
Numpy	Cálculo numérico y análisis de datos, especialmente para un gran volumen de datos.
Pandas	Define nuevas estructuras de datos basadas en los arrays de la librería NumPy. Permite leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL. Permite acceder a los datos mediante índices o nombres para filas y columnas. Ofrece métodos para reordenar, dividir y combinar conjuntos de datos.
Imutils	Realizar una serie de funciones de conveniencia para hacer que las funciones básicas de procesamiento de imágenes como la traslación, rotación, cambio de tamaño, esqueletización, visualización de imágenes
Joblib	Facilita la escritura de código paralelo legible y depurarlo rápidamente. Persistencia comprimida rápida: un reemplazo para que pickle funcione de manera eficiente en objetos Python que contienen datos grandes

Elaborado por: Equipo de trabajo

Tabla 15.Funciones principales en el desarrollo de la aplicación móvil

Nombre	Tarea
model = LogisticRegression()	Es el módulo utilizado para implementar la regresión logística.
model.fit(X_train, y_train)	Consiste en X_train es toda la instancia con atributos, y_train es la etiqueta de cada instancia.
prediction = model.predict(df2)	Modulo que se utiliza para la predicción de la regresión logística y máquina de vectores.
score = accuracy_score(y_test, y_pred)	Para probar la salida predicha del valor objetivo "y_pred" con el y_test, utilizando accuracy_score (y_test, y_pred),
clf = SVC(kernel="rbf").fit(X_train, y_train)	Es el módulo utilizado para implementar el algoritmo de máquina de vector de soporte.

Elaborado por: Equipo de trabajo

5.2. COMPROBACIÓN DE LA HIPÓTESIS

Para la comprobación de la hipótesis se utilizó el método de validación cruzada de esta manera se evaluó los resultados de un análisis estadístico y garantizar que los datos sean utilizados correctamente (datos de entrenamiento y prueba). Por lo que en la regresión logística dio un margen de error de 2.9% y una precisión de 97.1%, mientras en la SVM dio un margen de error de 27.76% y una precisión de 74.51%, el cual mide la precisión de eficiencia de los datos seleccionados en para la base de datos.

5.3. ANÁLISIS DE IMPACTO

5.3.1. Impacto social

El presente proyecto de investigación tiene un impacto social, debido al beneficio que da la aplicación web usando dos tipos de algoritmos de minería de datos como: regresión logística y máquina de vectores de soporte. Mediante la clasificación automática permite determinar en un tiempo mínimo a qué clase de planta pertenece sea monocotiledónea o dicotiledónea lo que aporta a la sociedad con futuras investigaciones sobre minería de datos.

5.3.2. Impacto tecnológico

El desarrollo de una aplicación web para la clasificación automática de plantas monocotiledóneas y dicotiledóneas, permitió aprovechar nuevas tecnologías innovadoras como: hardware y software, con la ayuda de las técnicas de minería de datos, para ello también el uso de nuevas tecnologías en la agricultura ayudando a los agricultores cambiar la forma de clasificar manual a la forma automática, con la finalidad de obtener resultados de forma rápida y precisa.

5.3.3. Impacto económico

En el impacto económico se incluye la inversión en el aspecto económico, por lo cual en el proyecto de investigación tendrá un costo de software de la aplicación móvil. Para ello se utilizó el modelo de COCOMO, para la estimación de costes software.

5.3.3.1. Costo del software

a. Estimación de cantidad de instrucciones

Para esta fase de estimación se eligió la cantidad de número de líneas de código que tiene la aplicación web. En la ecuación 1 significa L=cantidad de líneas de código, FD E/S Flujo de Entrada + Flujo de Salida de la aplicación web. En la ecuación 2 significa ML=Miles de código fuente que tendrá el sistema.

$$L = 200 * FD\ E/S \tag{1}$$

$$M\,L = \frac{L}{1000} \tag{2}$$

b. Estimación de esfuerzo

Esta fase consiste en el proceso de predecir el esfuerzo requerido para desarrollar o mantener un sistema de software. Se eligió el modo intermedio dependió del problema y se aplica personas no experimentadas. En la ecuación 3 significa el modo de desarrollo semiencajado=3 y el ML=miles de línea de código del sistema.

$$\mathrm{ESF} = 3 * \mathrm{ML}^{1.12} \tag{3}$$

c. Estimación de tiempo de desarrollo

Esta fase detalla el tiempo estimado que se ocupó para el desarrollo del proyecto de investigación de la aplicación web de clasificación automática de las plantas. En la ecuación 4 significa 2.5 y 0.35= modo de semi encajado de la estimación del software.

$$\mathrm{TDES} = 2.5 * \mathrm{ESF}^{0.35} \tag{4}$$

d. Estimación del personal necesario

En esta etapa se determina el personal que se requiere para el desarrollo del software. Se utilizó la ecuación 5 que significa ESF=estimación de esfuerzo y TDES= estimación de tiempo de desarrollo.

$$\mathrm{CP} = \frac{\mathrm{ESF}}{\mathrm{TDES}} \tag{5}$$

e. Estimación de productividad

Para la estimación de la productividad se basa en el tiempo que deben desarrollarse la actividad del proyecto de investigación. En la ecuación 6 significa L=líneas del código y ESF= estimación del esfuerzo.

$$P = \frac{L}{ESF} \quad (6)$$

f. Estimación de costo

Finalmente la estimación del costo del proyecto de investigación. En la ecuación 7 significa ESF= estimación del esfuerzo y CHM= sueldo del personal del proyecto.

$$EC = ESF * CHM \quad (7)$$

Para sacar el costo del proyecto primero se define el número de cantidad de instrucciones que tiene la aplicación web se basó en 200 líneas de código. Con esta información se realiza el proceso del costo del software.

Tabla 16. Modelo COCOMO – Costo de la aplicación web

Estimación del costo de desarrollo de la aplicación web	
Estimación de cantidad de instrucciones (1) y (2)	L=400 ML=0.4
Estimación de esfuerzo (3)	ESF=1.07≈1 persona
Estimación de tiempo de desarrollo (4)	TDES=2.55≈3 meses
Estimación del personal necesario (5)	CP=0.69≈1
Estimación de productividad (6)	P=374 instrucciones/personas_mes
Estimación de coste (7)	EC=$857

Elaborado por: Equipo de trabajo

En la Tabla 16 detalla el costo del desarrollo de la aplicación web en base al esfuerzo se requiere de una persona, en el tiempo de desarrollo se toma 3 meses para el proyecto de investigación, el personal necesario se requiere de una sola persona, y nos da un costo de $857 de la aplicación web.

5.3.3.2. Costo del alojamiento de la aplicación web

Para el alojamiento de la aplicación web se detalla en la Tabla 17.

Tabla 17. Costo de alojamiento de la aplicación web

Ítems	Descripción	Cantidad por mes	Costo unitario
1	Servidor Ubuntu 20.04-Instancia en amazon web service s2 (api)	1 mes	30$
2	Base de datos	1 mes	15$
		TOTAL	45$

Elaborado por: Equipo de trabajo

5.3.3.3. Costo total

Para la realización del costo total se obtiene del costo del desarrollo del software más el costo del alojamiento de la aplicación web como muestra en la Tabla 18 dio un total de $902 para la realización de la aplicación web.

Tabla 18. Costo total de la aplicación web

Ítems	Descripción	Costo unitario
1	Costo del desarrollo del software	$857
2	Costo del alojamiento de la aplicación web	$45
	TOTAL	$902

Elaborado por: Equipo de trabajo

6. CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

- La minería de datos permitió descubrir patrones en grandes cantidades con la finalidad de seleccionar la base de datos, entre las técnicas de clasificación están las descriptivas dando un valor a un atributo partiendo de las relaciones de datos conocidos mientras las predictivas se crean automáticamente partiendo del reconocimiento de patrones, con el reconocimiento de imágenes permite identificar de forma automática un objeto dentro de la imagen y finalmente las características en plantas monocotiledóneas y dicotiledóneas; las venas de las hojas son usualmente paralelas y ramificadas.
- La clasificación se basa en el tipo y forma de la hoja en las plantas monocotiledóneas y dicotiledóneas. Por ello, se ha construido un conjunto de datos de entrenamiento para el uso de algoritmos de regresión logística y máquinas de soporte de vectores, basadas en el reconocimiento visual de regiones con el proceso de seleccionar hojas claramente identificables.
- Para el desarrollo del prototipo de clasificación se recopilo un conjunto de imágenes digitales donde se extrajo las características (área, perímetro y centroide), mediante un proceso de segmentación esta información será almacenada en una base de datos , que será utilizada para la clasificación. Finalmente se utilizó dos funciones como: model=LogisticRegression() y clf = svm.SVC (kernel='rbf'), lo que permite la clasificación automática el cual se visualiza en la aplicación web.

6.2. Recomendaciones

- Es importante la recopilación de información sobre técnicas de minería de datos, las características para la segmentación de las imágenes, aspectos de las plantas monocotiledóneas y dicotiledóneas (hoja, flor, etc.). Estos datos deben ser investigados en repositorios, artículos académicos entre otros, para tener mayor eficiencia permitiendo tener una información verídica.
- Es fundamental conocer el funcionamiento de los algoritmos de regresión logística y máquina de vectores de soporte, con esto permita generar el modelo clasificador para la implementación en la aplicación.
- En el desarrollo del prototipo es indispensable realizar una validación (cross validation) sobre el porcentaje de precisión con la finalidad de obtener una correcta clasificación de las plantas.

7. BIBLIOGRAFÍA

[1] S. M. Malca Bulnes, "Modelo algorítmico para a clasificación ce una hoja de planta en base a sus características de forma y textura," *Propues. Pucp*, p. 61, 2018, [Online]. Available: http://tesis.pucp.edu.pe/repositorio/bitstream/handle/123456789/6053/MALCA_SUSANA_MODELO_ALGORITMICO_CLASIFICACION.pdf?sequence=1&isAllowed=y.

[2] S. M. Gómez and D. S. Sierra, "Pinos ponderosa," *Gimnospermas*, pp. 1–22, 2018, [Online]. Available: https://www.uv.mx/personal/tcarmona/files/2010/08/Cocucci-et-al-1994.pdf.

[3] C. S. Catarina, "Agricultura de Precisión," *INCYTU*, vol. 52, no. 55, 2018.

[4] H. B. D. Lucas and X. Lisbeth, "La minería de datos y algunas de sus aplicaciones contextual," *Grup. Editor. "Ediciones Futur.*, vol. 13, no. 11, pp. 17–25, 2020.

[5] I. G. Leiva, P. Díaz, J. Vicente, and R. Muñoz, "Técnicas y usos en la clasificación automática de imágenes," pp. 1–14, 2019, [Online]. Available: http://www.dspace.uce.edu.ec/handle/25000/19364.

[6] M. Rahaman, A. Asif, and M. Chen, "Data-mining Techniques for Image- based Plant Phenotypic Traits Identification and Classification," pp. 1–11, 2019, doi: 10.1038/s41598-019-55609-6.

[7] T. Zamora, "Aplicación de técnicas de minería de datos para pronósticos del sector agrícola," 2018.

[8] R. Danilo, "Reconocimiento de imágenes con técnicas de minería de datos.," 2019.

[9] G. A. L. Trung T. Plam, "Innovación en Minería de Datos para el Tratamiento de Imágenes: Agrupamiento K-media para Conjuntos de Datos de Forma Alargada y su Aplicación en la Agroindustria," *Scielo-Información tecnológica*, vol. vol.30, no. 0718–0764, p. 6, 2019, [Online]. Available: http://dx.doi.org/10.4067/S0718-07642019000200135.

[10] ANAYA, "El reino Plantas," *Univ. del pacifico*, vol. 1, pp. 172–196, 2017, [Online]. Available: file:///C:/Users/Master/Downloads/BIOLOGIA_GEOLOGIA_1_ESO_U08_ReinoPlantas.pdf.

[11] UPR-Mayagüez, "Angiospermas: Plantas vasculares con flores y frutos," *Biol. Org. Veg.*, pp. 1–11, 2017.

[12] D. Aguilar Sandí, "Notas para la identificación de familias de plantas con flores (angiospermas)," *Rev. Biol. Trop.*, p. 2019, 2019.

[13] S. M. Malca Bulnes, "Pontificia Universidad Católica Del Perú Facultad De Ciencias E Ingeniería Modelo Algorítmico Para La Clasificación De Una Hoja De Planta En Base a Sus Características De Forma Y Textura," *Propues. Pucp*, p. 61, 2018, [Online]. Available: http://tesis.pucp.edu.pe/repositorio/bitstream/handle/123456789/6053/MALCA_SUSANA_MODELO_ALGORITMICO_CLASIFICACION.pdf?sequence=1&isAllowed=y.

[14] J. I. Salazar Torres and E. Girón Cardenas, "Análisis y aplicación de algoritmos de minería de datos," *Perspectivas*, vol. 1, no. 21, pp. 71–88, 2021, [Online]. Available: https://revistas.uniminuto.edu/index.php/Pers/issu e/view/195.

[15] B. Kavitha and M. Nagarani, “Advanced Agriculture System Using Predictive,” vol. 5, no. 7, pp. 1280–1284, 2018.

[16] R. Camana, “Potenciales Aplicaciones de la Minería de Datos en Ecuador,” *Rev. Tecnológica ESPOL-RTE*, vol. 29, no. 1, pp. 170–183, 2016.

[17] A. F. Gil-Torres, A. L. Monroy-García, and J. S. González-Sanabria, “Minería de datos espacial en la agricultura en Latinoamérica - Una aproximación conceptual,” *Pensam. y Acción*, no. 28, pp. 19–33, 2019, doi: 10.19053/01201190.n28.2020.10976.

[18] N. Yusuf and T. Rohmah, “ALGORITMOS DE CLASIFICACION LINEAL PARA LA IDENTIFICACIÓN DE ZONAS CEREBRALES,” *PENGARUH Pengguna. PASTA LABU KUNING (Cucurbita Moschata) UNTUK SUBSTITUSI TEPUNG TERIGU DENGAN PENAMBAHAN TEPUNG ANGKAK DALAM PEMBUATAN MIE KERING*, pp. 274–282, 2020.

[19] C. Barrionuevo, J. S. Ierache, and I. I. Sattolo, “Reconocimiento de emociones a través de expresiones faciales con el empleo de aprendizaje supervisado aplicando regresión logística,” *XXVI Congr. Argentino Ciencias la Comput.*, pp. 491–500, 2020, [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/114089.

[20] Y. Sisco and F. Moreno, “Minería de datos para la detección de rostros mediante el uso de Maquinas de Soporte Vectorial, Redes Neuronales Artificiales y Redes Neuronales Convolucionales,” *Researchgate.Net*, no. March 2018, 2018, doi: 10.13140/RG.2.2.17256.78082.

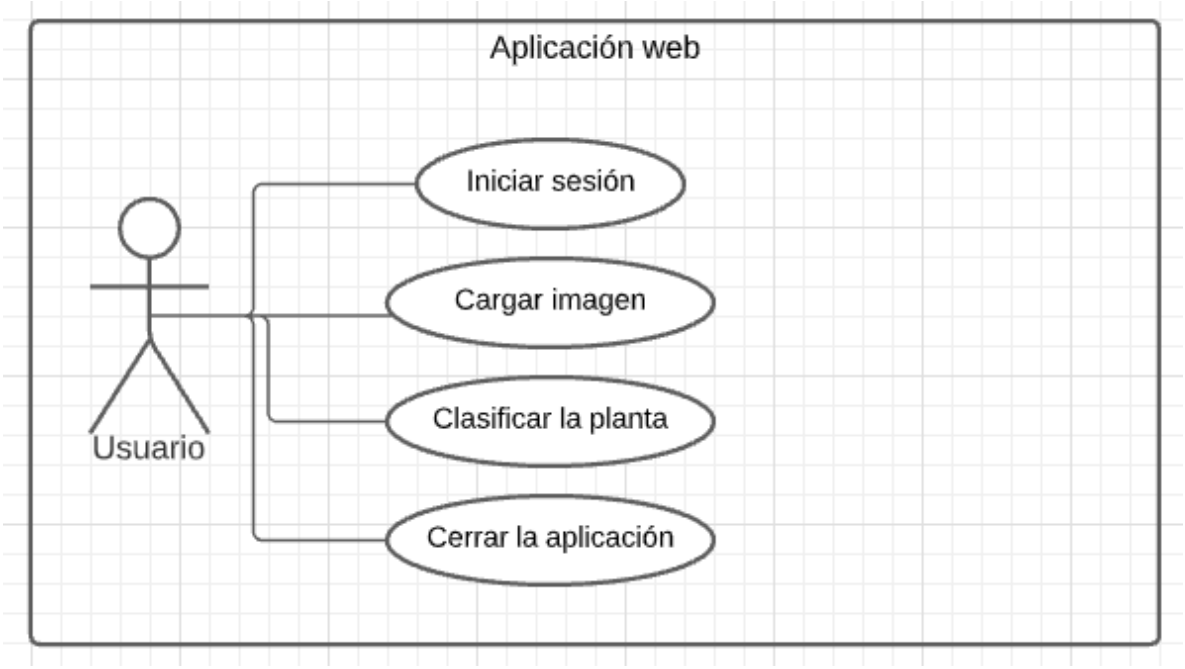
[21] B. Intelligence, “Tecnologías de Inteligencia de Negocios y Minería de datos para el análisis de la producción y comercialización de cacao,” *Revista Espacios*, vol. 39 (N^{o} 32), p. 6, 2018.

[22] A. Delgado, “Identificación automática de hojas utilizando un clasificador bayesiano,” Universidad Autónoma del Estado de México Centro Universitario UAEM Texcoco, 2018.

[23] J. V. Lozano, *Operadores morfológicos*, vol. 3. 2018, p. 53.

8. ANEXOS

Anexo A. Caso de uso general de la aplicación web



Anexo B. Caso de uso a detalle

Tabla 19. Caso de uso a detalle- Inicio de sesión

Iniciar sesión	
Código	CU001
Descripción	Este caso de uso permite al usuario el inicio de sesión en la aplicación web.
Actores	Administrador, usuario
Precondición	El usuario debe estar registrado, para ingresar a la aplicación web.
Flujo Principal "Iniciar sesión"	
1. El administrador ingresa a la interfaz de inicio de sesión 2. El usuario se autentica (nombre de usuario y clave) 3. La aplicación web muestra la página principal.	
Flujo Alterno *En caso de que usuario y clave no coincida con los registro de la BD:* 2A. La aplicación web notifica un mensaje "Por favor, introduzca un nombre de usuario y clave correctos. Observe que ambos campos pueden ser sensibles a mayúsculas"	
Post-Condición: Inicio de sesión exitoso.	

Elaborado por: El investigar

Tabla 22. Caso de uso a detalle-Cargar imagen

Cargar imagen	
Código	CU002
Descripción	Este caso de uso permite al usuario seleccionar la imagen sea monocotiledónea o dicotiledónea.
Actores	Administrador, usuario
Precondición	El usuario debe estar registrado, para ingresar a la aplicación web.
Flujo Principal "Cargar imagen"	
1. El usuario ingresa a la página principal. 2. El usuario elige el botón "seleccionar archivo" 3. La aplicación web muestra una ventana de exploración de los archivos. 4. El usuario selecciona la imagen y elige abrir. 5. La aplicación web muestra la imagen y en tipo da por mensaje "Por determinar".	
Post-Condición: Se carga la imagen exitosamente.	

Elaborado por: El investigar

Tabla 20. Caso de uso a detalle-Clasificar la panta

Clasificar la planta	
Código	CU003
Descripción	Este caso de uso permite clasificar la planta de manera automática sea de tipo monocotiledónea y dicotiledónea.
Actores	Administrador y usuario.
Precondición	El usuario tenía que seleccionar una imagen para la clasificación.
Flujo Principal "Clasificar la planta"	

1. El usuario debe estar en la página principal. 2. El usuario selecciono una imagen. 3. El usuario selecciona el botón “clasificar” 4. La aplicación web muestra la imagen, el tipo de planta y el porcentaje de precisión.
Flujo Alterno *En caso de que la imagen no pase el modelo de clasificación* **2A.** La aplicación web muestra un mensaje “Por determinar”.
Post-Condición: Se realiza la clasificación automática de manera exitosa.

Elaborado por: El investigar

Tabla 21. Caso de uso a detalle-Cerrar sesión

Cerrar la aplicación	
Código	CU004
Descripción	Este caso de uso permite cerrar la aplicación web.
Actores	Usuario
Precondición	El administrador debe estar registrado.
Flujo Principal “Editar usuario”	
1. El usuario está en la página principal. 2. El usuario selecciona el botón “cerrar” 3. La aplicación web muestra la interfaz de inicio de sesión.	
Post-Condición: Se cierra la aplicación web de manera exitosa.	

Elaborado por: El investigar

Anexo C. Proceso de regresión logística y MSV

```
def params(self, x, y):
    X_train, X_test, y_train, y_test = train_test_split(x,y, test_size=0.25, random_state=0)
    model = LogisticRegression()
    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)

    return X_train, X_test, y_train, y_test, y_pred

def category(self):
    model = joblib.load("./models/logistic.pkl")
    data_new = {'Area': [self.area],
                'Perimetro': [self.perimetro]}

    df2 = pd.DataFrame(data_new, columns=["Area", "Perimetro"])
    prediction = model.predict(df2)
    print(prediction)
    return str(prediction).replace("['","").replace("']","")


def score_logistic(self, y_test, y_pred):
    score = accuracy_score(y_test, y_pred)
    return score


def score_svm(self, X_test, y_test):
    model_svm = joblib.load("./models/svm.pkl")
    score = model_svm.score(X_test, y_test)
    return score
```

1. Esta función permite guardar el modelo entrenado (Maquina de soporte vectorial)

   X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.25, random_state=0)
   clf = SVC(kernel="rbf").fit(X_train, y_train)
   score = clf.score(X_test, y_test)
   utils = Utils()
   utils.model_export_svm(clf, score)

2. Esta función muestra como resultado si la imagen es monocotiledónea o dicotiledónea.

   def category(self):
   model = joblib.load("./models/logistic.pkl")
   data_new = {'Area': [self.area],
   'Perimetro': [self.perimetro]}
   df2 = pd.DataFrame(data_new, columns=["Area", "Perimetro"])
   prediction = model.predict(df2)
   print(prediction)
   return str(prediction).replace("['","").replace("']","")

3. En la siguiente funcion obtenemos el porcentaje del modelo entrenado regresión logística

```
def score_logistic(self, y_test, y_pred):
score = accuracy_score(y_test, y_pred)
return score
```

4. En la siguiente función obtenemos el porcentaje del modelo entrenado máquina de soporte vectorial

```
def score_svm(self, X_test, y_test):
model_svm = joblib.load("./models/svm.pkl")
score = model_svm.score(X_test, y_test)
return score
```

Anexo D. Juicio de expertos

Título: Clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos.

Tipo: Básico

Diseño: Experimental, transversal, correlacionar

Método: Hipotético-deductivo

a. Cuestionario de clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos.

El presente documento es anónimo y su aplicación será de utilidad para mi investigación, por ello pido su colaboración: Marque con una "X" la respuesta que considere acertada con su punto de vista, según las siguientes alternativas:

Tabla 23. Significado de los índices de juicio de expertos

TD	DA	NAD	ED	TDA	NIVELES Y RANGOS
TOTALEMENTE DE ACUERDO (5)	DE ACUERDO (4)	NI DE ACUERDO NI DESACUERDO (3)	EN DESACUERDO (2)	TOTALMENTE EN DESACUERDO (1)	Alto [40-34[Medio [35-29[Bajo [20-10[

Elaborado por: Equipo de trabajo

Tabla 24. Cuestionario para la evaluación de juicio de expertos

Nº	ÍTEMS	ÍNDICES				
		TD	DA	NAD	ED	TDA
1	¿Considera que es conveniente automatizar los procesos que se realiza en la agricultura?					
2	¿Considera que es importante conocer la clasificación de plantas monocotiledóneas y dicotiledóneas?					
3	Es importante para la clasificación la forma de la hoja					
4	¿Considera que es beneficioso aplicar técnicas de minería de datos en la agricultura					
5	¿Considera usted que es importante almacenar la información extraída de cada planta en una base de datos?					
6	La aplicación facilita el trabajo de clasificación de plantas monocotiledóneas y dicotiledóneas al agricultor					
7	¿La aplicación permite el reconocimiento de plantas monocotiledóneas y dicotiledóneas?					
8	¿Considera que es necesaria una capacitación para el uso de la aplicación?					

Elaborado por: Equipo de trabajo

a. Certificado de la clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos.

- Adecuación: La adecuación se válida para establecer si el requerimiento ha sido utilizado de manera satisfactoria, de igual modo se evalúa su facilidad de uso, precisión y el orden lógico con el cual han sido utilizado.

- Pertinencia: El criterio de pertinencia tiene la función de verificar si el requerimiento evaluado realmente es de utilidad para el ámbito de la ingeniería agrónoma y su valor a la lógica de negocio del prototipo.

A continuación, se presentan los promedios de las puntuaciones asignadas por los expertos que han colaborado en la validación del sistema:

Nº	ÍTEMS	ÍNDICES				
		TD	DA	NAD	ED	TDA
1	¿Considera que es conveniente automatizar los procesos que se realiza en la agricultura?	X				
2	¿Considera que es importante conocer la clasificación de plantas monocotiledóneas y dicotiledóneas?		X			
3	Es importante para la clasificación la forma de la hoja	X				
4	¿Considera que es beneficioso aplicar técnicas de minería de datos en la agricultura	X				
5	¿Considera usted que es importante almacenar la información extraída de cada planta en una base de datos?	X				
6	La aplicación facilita el trabajo de clasificación de plantas monocotiledóneas y dicotiledóneas al agricultor		X			
7	¿La aplicación permite el reconocimiento de plantas monocotiledóneas y dicotiledóneas?		X			
8	¿Considera que es necesaria una capacitación para el uso de la aplicación?	X				

Elaborado por: Equipo de trabajo

Opinión de aplicabilidad:

Aplicable [X] Aplicable después de corregir [ ] No aplicable [ ]

Apellidos y nombres del juez evaluador: Taipicaña Comasanta Carmen Viviana

Especialidad del evaluado: Ingeniera Agrónoma

SENESCYT: 1020-2019- 2115547

Mediante la evaluación de juicio de experto, llegó a la conclusión que el prototipo web cumple con las condiciones requeridas para realizar la clasificación automática de plantas monocotiledóneas y dicotiledóneas mediante regresión logística y VSM, que la interfaz es amigable con el usuario y que los datos obtenidos del porcentaje de precisión es el correcto. El experto llega a una calificación del 37 que tiene un equivalente de un nivel alto por lo que es aceptable el prototipo. También afirma que el prototipo web tiene las características necesarias para la clasificación automática.

b. Certificado de la clasificación automática de plantas monocotiledóneas y dicotiledóneas usando minería de datos.

- Adecuación: La adecuación se válida para establecer si el requerimiento ha sido utilizado de manera satisfactoria, de igual modo se evalúa su facilidad de uso, precisión y el orden lógico con el cual han sido utilizado.

- Pertinencia: El criterio de pertinencia tiene la función de verificar si el requerimiento evaluado realmente es de utilidad para el ámbito de la ingeniería agrónoma y su valor a la lógica de negocio del prototipo.

A continuación, se presentan los promedios de las puntuaciones asignadas por los expertos que han colaborado en la validación del sistema:

Nº	ÍTEMS	ÍNDICES				
		TD	DA	NAD	ED	TDA
1	¿Considera que es conveniente automatizar los procesos que se realiza en la agricultura?	X				
2	¿Considera que es importante conocer la clasificación de plantas monocotiledóneas y dicotiledóneas?	X				
3	Es importante para la clasificación la forma de la hoja	X				
4	¿Considera que es beneficioso aplicar técnicas de minería de datos en la agricultura	X				
5	¿Considera usted que es importante almacenar la información extraída de cada planta en una base de datos?	X				
6	La aplicación facilita el trabajo de clasificación de plantas monocotiledóneas y dicotiledóneas al agricultor	X				
7	¿La aplicación permite el reconocimiento de plantas monocotiledóneas y dicotiledóneas?		X			
8	¿Considera que es necesaria una capacitación para el uso de la aplicación?	X				

Elaborado por: Equipo de trabajo

Opinión de aplicabilidad:

Aplicable [X] Aplicable después de corregir [ ] No aplicable []

Apellidos y nombres del juez evaluador: Hidalgo Fernández Mónica del Pilar

Especialidad del evaluado: Ingeniera Agrónoma

SENESCYT: 1005-15-86070442

Mediante la evaluación de juicio de experto, llegó a la conclusión que el prototipo web cumple con las condiciones requeridas para realizar la clasificación automática de plantas monocotiledóneas y dicotiledóneas mediante regresión logística y VSM, que la interfaz es amigable con el usuario y que los datos obtenidos del porcentaje de precisión es el correcto. El experto llega a una calificación del 39 que tiene un equivalente de un nivel alto por lo que es aceptable el prototipo. También afirma que el prototipo web tiene las características necesarias para la clasificación automática.

AVAL DE JUICIO DE EXPERTOS

Yo, **HIDALGO FERNANDEZ MONICA DEL PILAR** con cédula de identidad N° **1712413655** en calidad de Ingeniera Agrónoma y Magister en Floricultura con número de registro de la **SENESCYT:** 1005-15-86070442; **CERTIFICO** que el proyecto realizado es funcional y cumple con todas las características requeridas para la realización de la clasificación, el tema del proyecto de investigación es:" **CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS",** por el estudiante **Cayambe Cajo Fabian Rolando** con C.I.: **172318874-2** de la Carrera de Ingeniería en Informática y Sistemas Computacionales de la Universidad Técnica de Cotopaxi, siendo la **Ing. Mtr. Karla Susana Cantuña Flores** tutor del presente trabajo.

Es todo cuanto puedo certificar en honor a la verdad y autorizo hacer uso del presente certificado de la manera ética que estimaren conveniente.

Quito, Marzo del 2022

Atentamente,

.....

Ing. MSc. HIDALGO FERNANDEZ MONICA DEL PILAR

C.I.: 1712413655

AVAL DE JUICIO DE EXPERTOS

Yo, **TAIPICAÑA COMASANTA CARMEN VIVIANA** con cédula de identidad **N° 050399535-9** en calidad de Ingeniera Agrónoma con número de registro de la **SENESCYT:** 1020-2019-2115547; **CERTIFICO** que el proyecto realizado es funcional y cumple con todas las características requeridas para la realización de la clasificación, el tema del proyecto de investigación es:" **CLASIFICACIÓN AUTOMÁTICA DE PLANTAS MONOCOTILEDÓNEAS Y DICOTILEDÓNEAS USANDO MINERÍA DE DATOS",** por el estudiante **Cayambe Cajo Fabian Rolando** con C.I.: **172318874-2** de la Carrera de Ingeniería en Informática y Sistemas Computacionales de la Universidad Técnica de Cotopaxi, siendo la **Ing. MSc. Karla Susana Cantuña Flores** tutor del presente trabajo.

Es todo cuanto puedo certificar en honor a la verdad y autorizo hacer uso del presente certificado de la manera ética que estimaren conveniente.

Quito, Marzo del 2022

Atentamente,

.....

Ing. TAIPICAÑA COMASANTA
CARMEN VIVIANA
C.I.: 050399535-9