

$$\sum_x \sum_y (I(x,y) * (x - \bar{x})^j * (y - \bar{y})^i)$$

$\bar{x} = \frac{m_{10}}{m_{00}}$

# SISTEMA DE IDENTIFICACIÓN AUTOMÁTICA DE ESPECIES VEGETALES DE AMARANTHACEAE MEDIANTE MORFOMETRÍA GEOMÉTRICA DE LA ARQUITECTURA FOLIAR

$$\sum_x \sum_y (I(x,y) * (x - \bar{x})^j * (y - \bar{y})^i)$$

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}}$$

$$mu_{ji} = \sum_x \sum_y (I(x,y) * (x - \bar{x})^j * (y - \bar{y})^i)$$

$$\bar{x} = \frac{m_{10}}{m_{00}}$$

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}}$$

JESUS EINER ZAPATA GRANADA  
CC 1097394172

CARLOS ANDRÉS FLÓREZ VILLARRAGA  
CC 1097398578

$$mu_{ji} = \sum_x \sum_y (I(x,y) * (x - \bar{x})^j * (y - \bar{y})^i)$$

$$\bar{x} = \frac{m_{10}}{m_{00}}$$

Trabajo de grado para obtener el título de Magister en Ingeniería de Sistemas y  
Computación.

$$nu_{ji} = \frac{mu_{ji}}{m_{00}^{(i+j)/2+1}}$$

$$nu_{ji} = \frac{mu_{ji}}{m_{00}^{(i+j)/2+1}}$$

$$\Delta(x,y) = \begin{cases} 1, & \text{if } \exists (x',y') \in N_8[(x,y)]: \\ & I_b(x',y') = 1 \\ 0, & \text{otherwise} \end{cases}$$

Universidad Tecnológica de Pereira.  
2018.



## INFORMACIÓN PROPONENTES

### ESTUDIANTES

**Nombre:** Jesús Einer Zapata Granada.

**Cédula:** 1097394172.

**Email:** [einerzg@gmail.com](mailto:einerzg@gmail.com)

**Celular:** 315 5733295

**Nombre:** Carlos Andrés Flórez Villarraga.

**Cédula:** 1097398578.

**Email:** [caflorezvi@gmail.com](mailto:caflorezvi@gmail.com)

**Celular:** 318 6459835

### DIRECTOR

**Nombre:** Julián Esteban Gutiérrez Posada, PhD.

**Cédula:** 14896480.

**Email:** [jugutier@uniquindio.edu.co](mailto:jugutier@uniquindio.edu.co).

**Celular:** 300 8386860

### ASESORES

**Nombre:** Hernán Felipe García Arias, PhD(c).

**Email:** [hernan.garcia@utp.edu.co](mailto:hernan.garcia@utp.edu.co).

**Nombre:** Carlos Alberto Agudelo, PhD.

**Email:** [cagudelo@uniquindio.edu.co](mailto:cagudelo@uniquindio.edu.co).



## TABLA DE CONTENIDOS

1. INTRODUCCIÓN .....	10
2. JUSTIFICACIÓN.....	17
3. IDENTIFICACIÓN DEL PROBLEMA.....	18
4. OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS: .....	19
4.1. Objetivo general.....	19
4.2. Objetivos específicos .....	19
5. METODOLOGÍA .....	20
6. CARACTERIZACIÓN DE LAS ESPECIES VEGETALES DE LA FAMILIA AMARANTHACEAE .....	22
6.1. Terminología de las partes básicas de una hoja .....	23
6.2. Características asociadas a la forma y contorno.....	24
6.2.1. Conexión de la hoja .....	24
6.2.2. Organización de la hoja.....	24
6.2.3. Tamaño laminar .....	25
6.2.4. Forma laminar .....	25
6.2.5. Simetría laminar .....	26
6.2.6. Ratio L:W laminar.....	26
6.2.7. Ángulo de la base .....	27
6.2.8. Ángulo del ápice.....	27
6.2.9. Forma de la base .....	28
6.2.10. Forma del ápice .....	28
6.2.11. Tipo de margen (dentado).....	29
6.2.12. Lobulación.....	29
6.3. Características asociadas al orden de las venas .....	30
6.3.1. Primer orden de venación .....	30
6.3.2. Segundo orden de venación.....	31
6.3.3. Tercer orden de venación .....	32
6.3.4. Areolación .....	33
6.4. Características y descripción de la familia Amaranthaceae .....	34
6.4.1. Achyranthes .....	34
6.4.2. Blutaparon.....	35
6.4.3. Celosia.....	35
6.4.4. Chamissoa .....	35
6.4.5. Cyathula.....	35

6.4.6. Gomphrena .....	35
6.4.7. Iresine .....	36
6.4.8. Amaranthus .....	36
6.4.9. Alternanthera .....	36
7. TÉCNICAS MORFOMÉTRICAS PARA LA EXTRACCIÓN DE CARACTERÍSTICAS DE LAS HOJAS .....	36
8. ENFOQUES MACHINE LEARNING APLICADOS AL ENTRENAMIENTO DE CARACTERÍSTICAS DE HOJAS.....	39
9. PROCEDIMIENTOS Y RECURSOS PARA LA CAPTURA DE IMÁGENES.....	41
9.1. Trabajos relacionados .....	41
9.2. Procedimiento y recursos para la captura digital de las imágenes de las hojas .....	42
10. BASES DE DATOS BASADAS EN HOJAS.....	46
10.1. Flavia .....	46
10.2. Leaf Data Set .....	47
10.3. Swedish .....	48
10.4. AmaranthaceaeUQ.....	49
11. TÉCNICAS DE EXTRACCIÓN DE CARACTERÍSTICAS SELECCIONADAS .....	52
11.1. Momentos de la imagen .....	55
11.2. Análisis de textura fractal .....	56
12. SELECCIÓN DE CARACTERÍSTICAS.....	57
12.1. Características que más aportan a la clasificación .....	58
13. SELECCIÓN DE ENFOQUE DE APRENDIZAJE DE MÁQUINA.....	59
13.1. Perceptrón Multicapa - MLP .....	60
13.1.1. Capas ocultas .....	60
13.1.2. Función de activación.....	60
13.1.3. Solver.....	61
13.1.4. Alpha.....	62
13.2. MAQUINA DE VECTOR DE SOPORTE - SVM.....	62
13.2.1. Kernel.....	62
13.2.2. C .....	63
13.2.3. Gamma .....	63
13.2.4. Grado.....	63
13.3. K VECINO MÁS CERCANO - K-NN.....	63
13.3.1. Número de vecinos .....	64
13.3.2. Pesos .....	64
13.3.3. Algorithm.....	64

13.4.	GRIDSEARCHCV .....	65
13.5.	PRUEBA DE CONCEPTO - SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS .....	65
13.6.	SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS PARA LA BASE DE DATOS AMARANTHACEAEUQ .....	68
13.7.	SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS PARA AMARANTHACEAUQ USANDO UNO CONTRA TODOS	71
13.7.1.	Configuraciones destacadas para MLP .....	71
13.7.2.	Configuraciones destacadas para SVM.....	72
13.7.3.	Configuraciones destacadas para K-NN.....	72
13.7.4.	Resultados .....	72
14.	ESQUEMA GENERAL DE LA SOLUCIÓN TECNOLÓGICA DESARROLLADA.....	73
15.	ARQUITECTURA DEL PROTOTIPO DE LA SOLUCIÓN TECNOLÓGICA .....	74
15.1.	Vista de escenarios .....	74
15.2.	Vista lógica.....	75
15.4.	Vista de despliegue .....	77
15.5.	Vista de implementación .....	77
16.	DESCRIPCIÓN DE LA APLICACIÓN MÓVIL.....	78
17.	PRUEBAS FUNCIONALES DE LA SOLUCIÓN TECNOLÓGICA.....	82
18.	RESUMEN DE RESULTADOS .....	89
	CONCLUSIONES Y TRABAJOS FUTUROS .....	90
	REFERENCIAS .....	93
	ANEXOS.....	96

## LISTA DE FIGURAS

Figura 1. Marco general del desarrollo del proyecto.....	13
Figura 2. Prueba de concepto para la selección de técnicas de extracción de características y enfoques <i>machine learning</i> . ....	14
Figura 3. Experimentación para la definición del proceso para la captura de fotografías de hojas de especies vegetales. ....	15
Figura 4. Adaptación de las características y los enfoques machine learning identificados en la prueba de concepto a la base de datos AmaranthaceaUQ.....	16
Figura 5. Proceso para la creación de la solución tecnológica. ....	17
Figura 6. Partes básicas de una hoja.....	23
Figura 7. Tipos de conexión hoja. ....	24
Figura 8. Tipos de organización de la hoja.....	25
Figura 9. Tipos de la forma laminar de la hoja.....	25
Figura 10. Simetría laminar de la hoja.....	26
Figura 11. Medidas laminares de la hoja.....	27
Figura 12. Diferentes ángulos de la base de la hoja.....	27
Figura 13. Diferentes ángulos del ápice de la hoja.....	28
Figura 14. Tipos de formas de la base de la hoja.....	28
Figura 15. Tipos de formas del ápice de la hoja.....	29
Figura 16. Tipos de márgenes de la hoja.....	29
Figura 17. Tipos de lóbulos de la hoja.....	30
Figura 18. Categorías de venas.....	30
Figura 19. Tipos de venas primarias de la hoja.....	31
Figura 20. Tipos de venas secundarias de la hoja.....	32
Figura 21. Tipos de venas terciarias de la hoja.....	33
Figura 22. Tipos de areolas de la hoja.....	34
Figura 23. Características foliares de la hoja de 03-Celosia argentea var. cristata.....	37
Figura 24. ERAF00.....	43
Figura 25. ERAF01.....	44
Figura 26. ERAF01 encendido.....	45
Figura 27. ERAF01 con una hoja.....	45
Figura 28. ERAF01 con un dispositivo móvil listo para capturar la foto.....	45
Figura 29. ERAF01 con un dispositivo móvil listo para capturar la foto.....	45
Figura 30. Fotografía de hoja madura de Amaranthus blitum.....	50
Figura 31. Fotografía de hoja inmadura de Amaranthus blitum.....	50
Figura 32. Matriz de covarianza aplicando las técnicas de extracción: Momentos de la imagen, Descriptores Elípticos de Fourier, Momentos polinómicos de Zernike, Momentos invariantes de Hu y Análisis de Textura Fractal.....	53
Figura 33. Matriz de covarianza aplicando las técnicas de extracción: Momentos de la imagen y Análisis de Textura Fractal.....	54
Figura 34. Imagen de la base de datos AmaranthaceaUQ segmentada.....	55
Figura 35. Matriz de confusión de la identificación todos contra todos usando SVM con la mejor configuración.....	70
Figura 36. Esquema general de la solución tecnológica desarrollada.....	73
Figura 37. Diagrama arquitectónico Cliente-Servidor.....	75
Figura 38. Diagrama arquitectónico de la aplicación móvil.....	76

Figura 39. Proceso para realizar la identificación taxonómica.....	77
Figura 40. Inicio .....	78
Figura 41. Menú desplegable.....	78
Figura 42. Clasificar inicio.....	79
Figura 43. Clasificar imágenes.....	79
Figura 44. Tarea en ejecución.....	80
Figura 45. Respuesta de la clasificación.....	80
Figura 46. Detalle del género.....	81
Figura 47. Lista de géneros.....	81
Figura 48. Descripción pasos para captura.....	82
Figura 49. Información acerca de la app.....	82
Figura 50. Resultados de la prueba todos contra todos iPhone 8 plus.....	85
Figura 51. Resultados de la prueba uno contra todos iPhone 8 plus.....	86
Figura 52. Resultados de la prueba todos contra todos iPhone 7.....	89
Figura 53. Resultados de la prueba uno contra todos iPhone 7.....	89

## LISTA DE TABLAS

Tabla 1.	Distribución en subfamilias y tribus de los géneros de Amaranthaceae presentes en Colombia [2].	10
Tabla 2.	Tamaños de la lámina.	25
Tabla 3.	Ejemplo de las muestras en la base de datos Flavia.	47
Tabla 4.	Ejemplo de las muestras en la base de datos Leaf Data Set.	48
Tabla 5.	Ejemplo de las muestras en la base de datos Swedish.	49
Tabla 6.	Fotos por género de la base de datos AmaranthaceaeUQ.	52
Tabla 7.	Resultados obtenidos con variada técnicas de extracción de características morfológicas.	54
Tabla 8.	Características extraídas por muestra de la base de datos AmaranthaceaeUQ.	59
Tabla 9.	Información base de la búsqueda de hiperparámetros para MLP con la base de datos Flavia.	66
Tabla 10.	Información base de la búsqueda de hiperparámetros para SVM con la base de datos Flavia.	67
Tabla 11.	Información base de la búsqueda de hiperparámetros para K-NN con la base de datos Flavia.	67
Tabla 12.	Información base de la búsqueda de hiperparámetros para MLP con la base de datos AmaranthaceaUQ.	68
Tabla 13.	Información base de la búsqueda de hiperparámetros para SVM con la base de datos AmaranthaceaUQ.	69
Tabla 14.	Información base de la búsqueda de hiperparámetros para K-NN con la base de datos AmaranthaceaUQ.	70
Tabla 15.	Resultados de entrenamiento uno contra todos con MLP.	72
Tabla 16.	Resultados de entrenamiento uno contra todos con SVM.	72
Tabla 17.	Representación del atributo de calidad de rendimiento.	75
Tabla 18.	Resultados prueba iPhone 8 plus.	85
Tabla 19.	Resultados prueba iPhone 7.	88
Tabla 20.	Tabla de resultados.	90

## 1. INTRODUCCIÓN

El Herbario de la Universidad del Quindío (HUQ) fundado hace más de 30 años y adscrito al Centro de Estudios e Investigaciones en Biodiversidad y Biotecnología (CIBUQ) del programa de Biología de la Universidad del Quindío se especializa en plantas de la región andina colombiana, debido a la gran cantidad de especies vegetales existentes, el personal del herbario estudia algunas familias de plantas específicas, una de ellas es la Amaranthaceae, la cual pertenece al orden Caryophyllales.

Según [1], la familia Amaranthaceae s. lat. (Incluyendo Chenopodiaceae) contiene aproximadamente 165 géneros y 2040 especies que se distribuyen por casi todo el planeta, exceptuando la región ártica, pero es más diversa en las regiones tropicales. En América tropical está representada por 20 géneros y 300 especies aproximadamente. En Colombia Amaranthaceae s. str. Presenta cerca de 14 géneros, 49 especies y siete variedades [2]. Los géneros mejor representados en Colombia son *Alternanthera*, *Amaranthus* y *Pfaffia*, los cuales se encuentran distribuidos principalmente en las regiones Andina y Caribe. El mayor número de especies de esta familia se encuentran entre 500 y 2500 m de altitud. En la Tabla 1 se destaca la división taxonómica de la familia Amaranthaceae en el territorio colombiano.

Subfamilias	Tribus	Géneros
Amaranthoideae (anteras biloculares)	Celosieae (hojas alternas y ovario con numerosos óvulos)	<i>Celosia</i> , <i>Pleuropetalum</i>
	Amarantheae (hojas alternas u opuestas y ovario con un óvulo)	<i>Amaranthus</i> , <i>Achyranthes</i> , <i>Chamissoa</i> , <i>Cyathula</i>
Gomphrenoideae (anteras uniloculares)	Gomphreneae (hojas opuestas y ovario con un óvulo)	<i>Alternanthera</i> , <i>Blutaparon</i> , <i>Froelichia</i> , <i>Gomphrena</i> , <i>Guilleminea</i> , <i>Hebanthe</i> , <i>Iresine</i> , <i>Pfaffia</i> .

**Tabla 1.** Distribución en subfamilias y tribus de los géneros de Amaranthaceae presentes en Colombia [2].

La taxonomía vegetal es la parte de la botánica que se ocupa de la clasificar y ordenar las plantas. Las plantas se ordenan en diferentes grupos o categorías taxonómicas de acuerdo con las afinidades que presentan. De esta manera se va formando un sistema jerárquico de categorías taxonómicas, debido a que cada categoría está subordinada o incluida en otra más amplia, y a su vez incluye otros

grupos de categoría inferior. La categoría taxonómica utilizada por la familia Amaranthaceae incluye lo siguiente: orden, familia, subfamilia, tribu, género, especie y variedad.

Amaranthaceae se compone generalmente de hierbas y arbustos, se destaca la Remolacha (*Beta vulgaris*), Acelga (*Beta vulgaris var. cicla*), Espinaca (*Spinacia oleracea*), las cuales son especies cultivadas para consumo, Escancel (*Aerva sanguinolenta*) que se usa como planta medicinal y las semillas de varias especies de *Chenopodium* y *Amaranthus* que son utilizadas en América para realizar harina.

Los herbarios se consideran como depósitos importantes y estructurados de conocimiento especializado en plantas, por ejemplo, el Herbario Kew en Londres tiene más de 7 millones de especímenes secados, que fueron recolectados de diferentes partes del mundo, algunas especies llevan más de 200 años en el herbario. Actualmente, y gracias a las nuevas tecnologías, estas colecciones de especies vegetales se están digitalizando para crear bases de datos gigantescas que incluyen las imágenes de cada planta y sus metadatos, lo cual mejorará el acceso a dicha información. El proceso de recolección y almacenamiento de especies vegetales es un trabajo minucioso que debe ser realizado por personas con amplios conocimientos en taxonomía vegetal. Por lo tanto, se resalta que todo esto ha llevado a un creciente interés en automatizar el proceso de identificación de especies vegetales y tareas relacionadas.

Los avances contemporáneos en la tecnología han dado lugar a nuevas herramientas que permiten la clasificación basada en enfoques alternativos e innovadores, por ejemplo: la transformada de *Fourier* espectroscopía infrarroja (FTIR) en hojas de *Camellia* para determinar si se pueden discriminar sobre la base de perfiles bioquímicos, o el uso de enfoques moleculares basados en información genética para la clasificación de especies de *Camellia* [3].

Los rasgos la arquitectura foliar de las hojas es comúnmente utilizado para identificar la taxonomía de las plantas, ya que sus características no son estacionarias, a diferencia de las flores y frutos, cabe destacar que por esta misma causa el color no es una característica útil para identificar una especie vegetal. Aunque la arquitectura foliar ha sido objeto de algunos estudios, la falta de definiciones estándar de las características foliares ha causado confusión al interpretar el valor de las clasificaciones resultantes [3]. La arquitectura foliar se refiere a la colocación y forma de varios elementos que constituyen la expresión externa de la estructura foliar, incluyendo la forma de las hojas, el tamaño, la configuración marginal, la posición de las glándulas y el patrón de venación [4].

El proceso taxonómico basado en la arquitectura foliar inicia ordenando las hojas en grupos con base su tipo de dentado, margen, patrón de venas primarias y secundarias, además la presencia y tipos de lóbulos. Estas características suelen ser usadas ya que son usualmente estables. Para las siguientes fases del proceso

taxonómico se usan las características restantes de las hojas [4].

La identificación de las especies vegetales se basa principalmente en una amplia gama de métodos morfométricos. La Morfometría se refiere al análisis cuantitativo de la forma, la cual se ha aplicado durante muchos años a las plantas, e incluso en estudios relacionados con la zoología, geología y medicina. Las plantas son extremadamente diversas en forma, tamaño y color. Un método que funciona muy bien en un grupo o familia de plantas (taxón) puede necesitar algunas características que no son relevantes en otro taxón. Por lo anterior, en [5] recomiendan centrarse en un grupo concreto de especies vegetales (en este proyecto *Amaranthaceae*).

La identificación automática de especies vegetales por medio de reconocimiento digital de imágenes necesita de un conjunto de características de las hojas que se deben extraer mediante técnicas de procesamiento de imágenes y visión artificial. El reconocimiento de imágenes es un proceso complejo ya que requiere una serie de pasos que sucesivamente transforman los datos icónicos a información que el computador puede reconocer [6]. En los últimos años ha habido un incremento muy importante en campos como: morfología de imágenes, Redes Neuronales Artificiales (ANN) o sistemas de análisis basados en conocimiento.

El reconocimiento de imágenes está directamente relacionado con la inteligencia artificial, la cual a través de ciertas técnicas permite reconocer ciertos elementos para determinar qué es lo que hay en la imagen. Los computadores no ven las imágenes de la misma manera que lo hacemos nosotros, para un computador una imagen no es más que una matriz de bits que definen su color y su intensidad, pero usando inteligencia artificial es posible entrenar un algoritmo para que reconozca patrones comunes de formas y colores con el fin de darle sentido a la imagen y extraer sus características de valor.

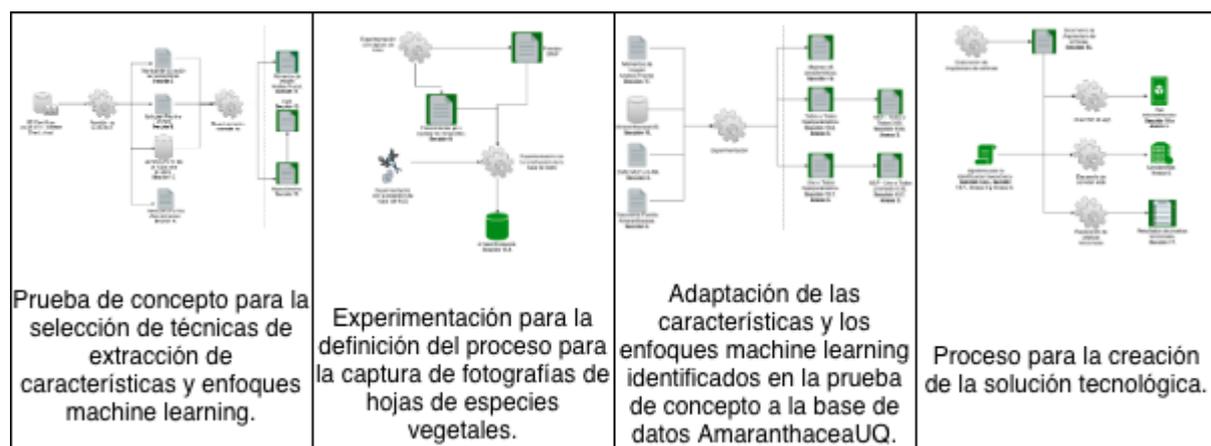
El procesamiento o tratamiento digital de imágenes consiste en procesos algorítmicos que transforman una imagen en otra en donde se resalta cierta información de interés, o se atenúa o elimina información irrelevante para la aplicación, este proceso se conoce como segmentación. Así, las tareas de procesamiento de imágenes comprenden la supresión de ruido, mejoramientos de contraste, eliminación de efectos no deseados en la captura como distorsiones por efectos ópticos o de movimiento, mapeos geométricos, transformaciones de color, etc [7].

Existen numerosas aplicaciones de reconocimiento de imágenes, de las cuales se pueden destacar entre otras las de reconocimiento de rostros y expresiones faciales, reconocimiento de firmas y de caracteres. El reconocimiento de rostros es ampliamente usado en cámaras digitales y redes sociales como Facebook, Snapchat y especialmente Google que usa reconocimiento de imágenes en sus mapas, Google Street View, Google Goggles, Google Photos, así como

reconocimiento de caracteres en Google Translate que junto con realidad aumentada permite traducir cualquier texto (de entre más de 100 idiomas) en tiempo real usando la cámara del dispositivo móvil. El reconocimiento de firmas se usa por ejemplo en sistemas biométricos para identificar personas. Este tipo de cosas se logra analizando un gran volumen de datos mediante algún algoritmo que se encarga de extraer características de las imágenes para encontrar similitudes entre ellas y que posteriormente llevan a reconocer objetos, personas, animales, plantas, textos, entre otras.

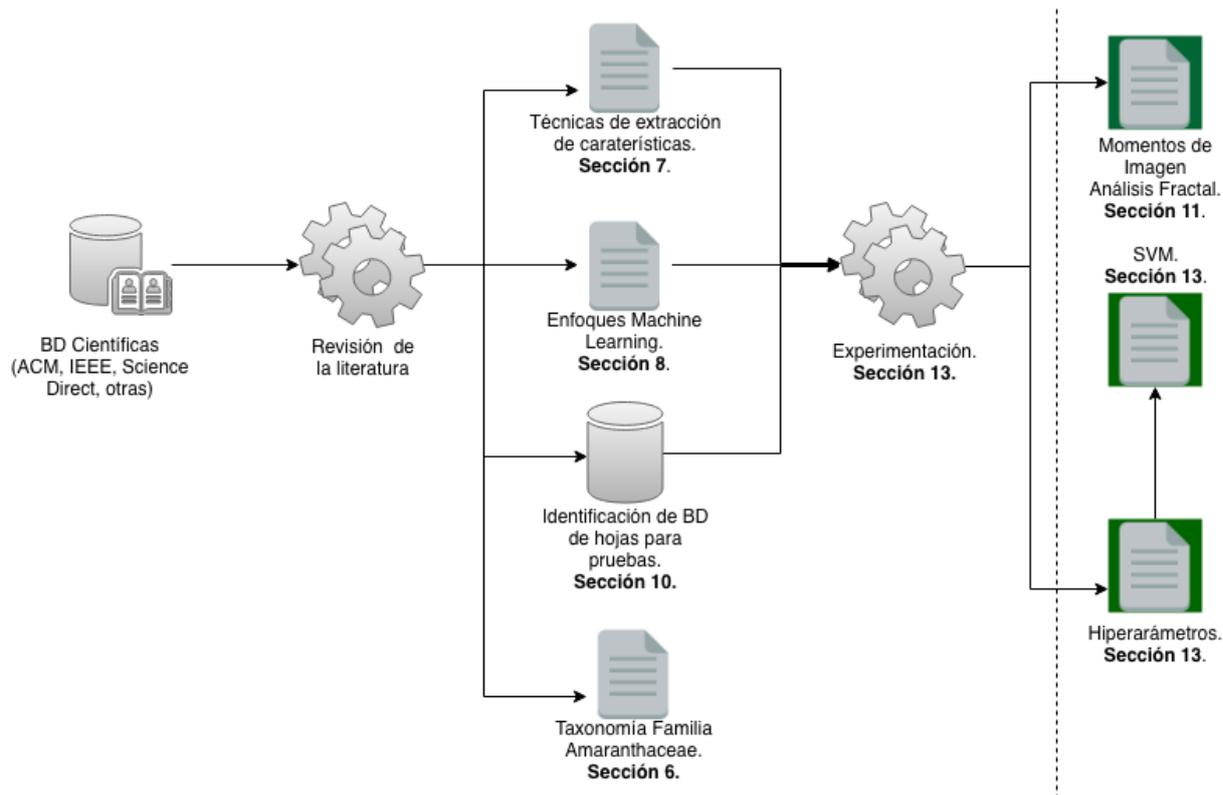
La identificación de plantas no es una tarea fácil ya que requiere conocimientos especializados en biología. Por lo anterior, en [8] plantearon la necesidad de construir un sistema que identifique plantas automáticamente, el cual recibe como insumo fotografías capturadas a través de la cámara de dispositivos móviles. Este sistema se encarga de extraer las características de la planta como la forma de su hoja, la flor, el fruto, tallo y la corteza, que permitan identificar la especie vegetal.

En este proyecto se ha construido un sistema basado en un algoritmo que permite facilitar la identificación taxonómica de especies vegetales de la familia Amaranthaceae, basado en información obtenida en el herbario HUQ. Este documento continúa con las secciones: dos, Justificación; tres, identificación del problema; cuatro, objetivos; y cinco, metodología. Las siguientes secciones del documento se dividen en cuatro fases principales, las cuales se visualizan en la siguiente figura.



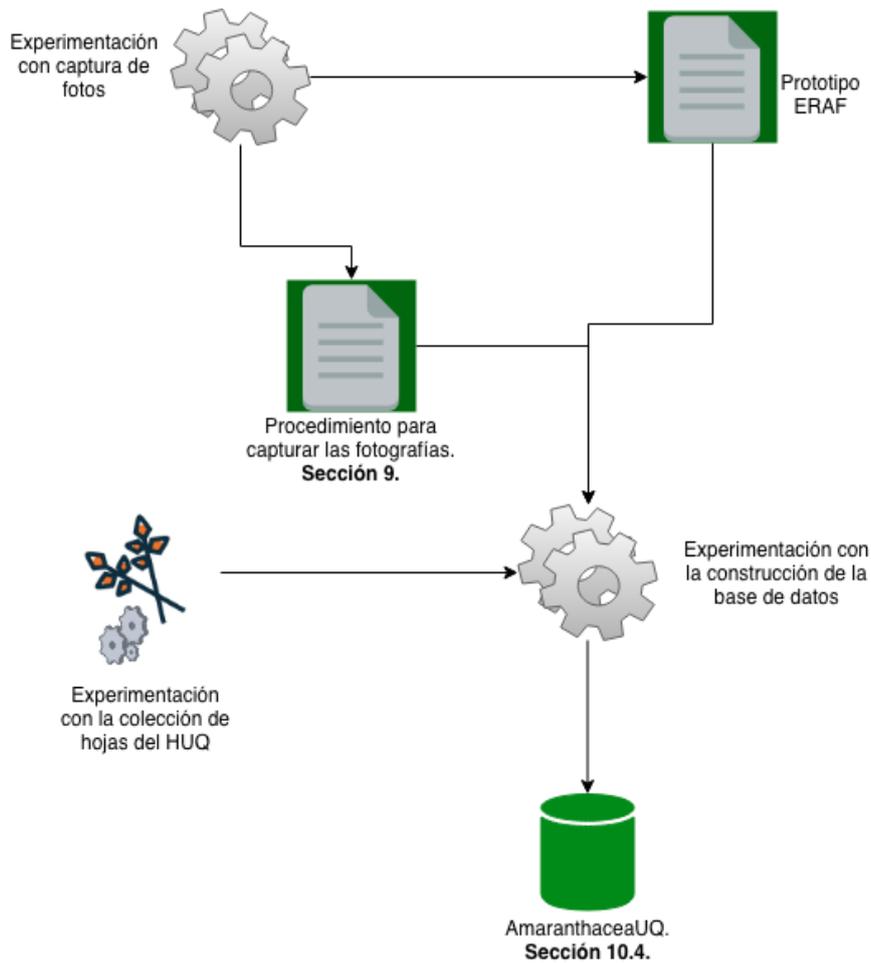
**Figura 1.** Marco general del desarrollo del proyecto.

La primera parte se enfocó en la revisión de la literatura y la experimentación con la información identificada. En la Figura 2 se muestran los pasos realizados en esta fase, de color verde se pueden identificar los resultados obtenidos.



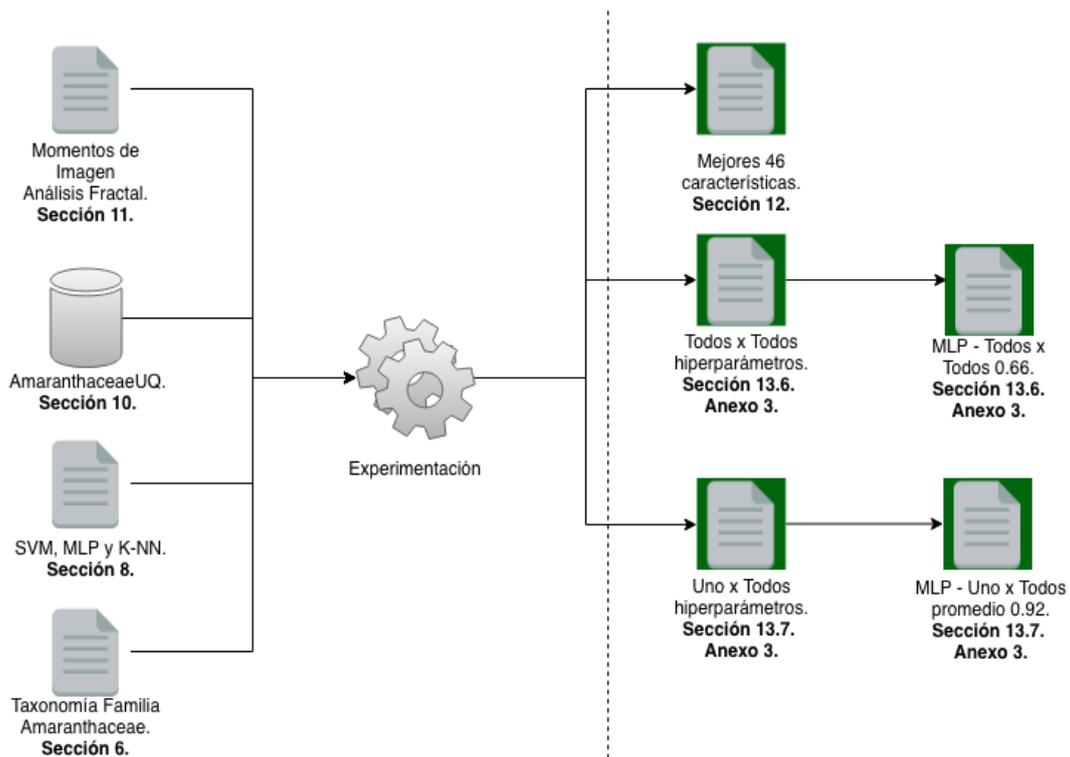
**Figura 2.** Prueba de concepto para la selección de técnicas de extracción de características y enfoques *machine learning*.

La fase dos, representada por la Figura 3, se centró en la definición de un proceso que permitiera la creación de una base de datos de la familia Amaranthaceae con todas las características necesarias para realizar el entrenamiento con los enfoques *machine learning* seleccionados.



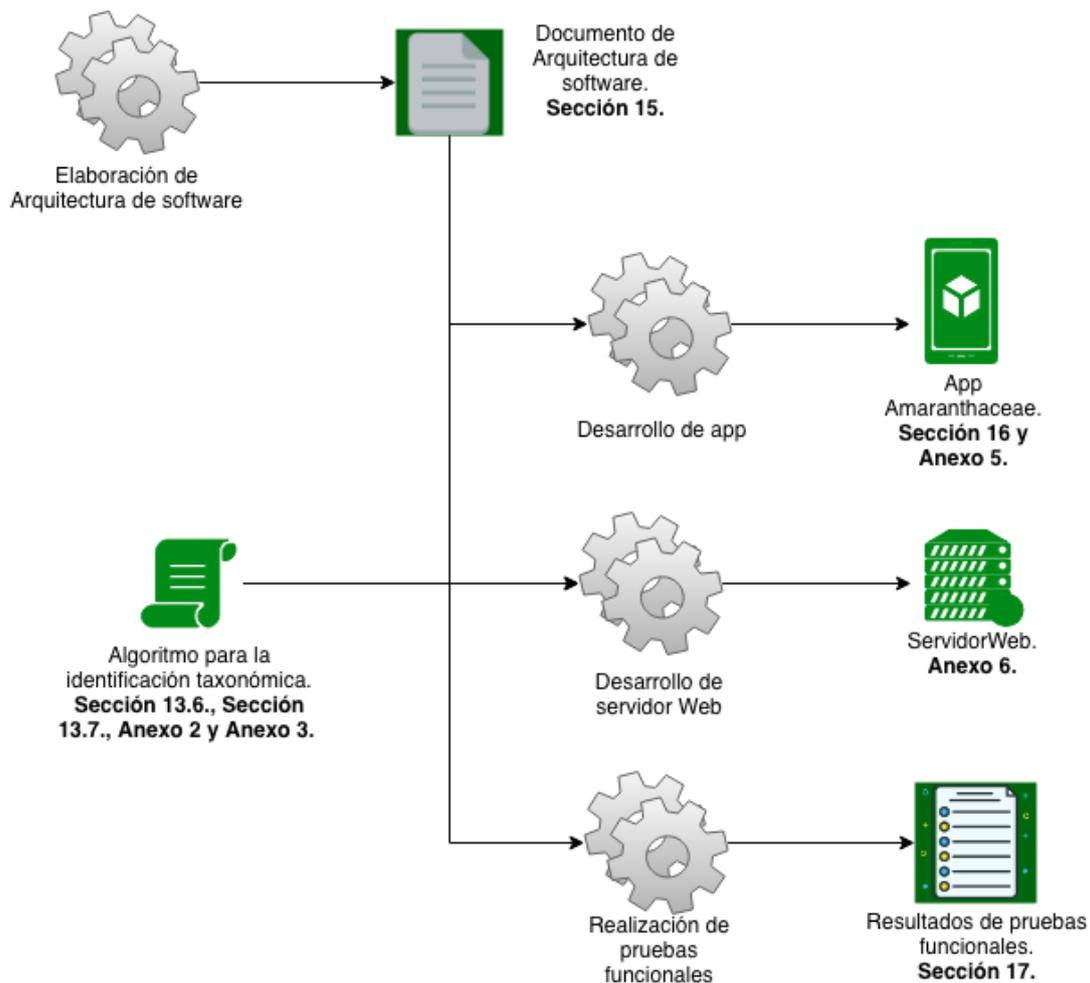
**Figura 3.** Experimentación para la definición del proceso para la captura de fotografías de hojas de especies vegetales.

La fase tres (Figura 4), muestra los pasos y los resultados obtenidos al experimentar con los productos de la prueba de concepto. Los resultados (resaltados en verde) definen el tipo de entrenamiento y la configuración que debe tener el algoritmo para realizar la clasificación de las especies Amaranthaceae usando la base de datos construida.



**Figura 4.** Adaptación de las características y los enfoques machine learning identificados en la prueba de concepto a la base de datos AmaranthaceaeUQ.

En la fase cuatro, representada con la Figura 5, se muestra la secuencias de cómo se desarrolló la solución tecnológica de cuáles fueron los resultados.



**Figura 5.** Proceso para la creación de la solución tecnológica.

Para finalizar se dan conclusiones y los trabajos futuros.

## 2. JUSTIFICACIÓN

Las especies vegetales familia Amaranthaceae se destacan por su variedad de usos. Según [2], el 54% de las especies que crecen en Colombia tienen importancia económica, debido a que son utilizadas principalmente como medicinales (82%), alimenticias (36%) y ornamentales (21%), aquí se puede resaltar la existencia de plantas que pueden ser medicinales, ornamentales o alimenticias a la vez. A nivel medicinal, estas plantas son utilizadas para combatir enfermedades cardíacas, cancerígenas, respiratorias e infecciosas. Sin embargo, falta información sobre la composición química, la actividad biológica y las dosis adecuadas para el consumo humano, entre otros aspectos, que respalden su uso con fines medicinales.

La importancia de la realización de la taxonomía de las plantas radica en identificar con precisión las especies o géneros vegetales por medio de nombres científicos únicos, además es de utilidad para las organizaciones e investigadores que buscan

plantas desde un punto de vista económico, como es el caso de aquellas que pueden ser utilizadas para desarrollar nuevos fármacos o alimentos. Por otra parte, sin una taxonomía adecuada es complicado determinar cuáles especies están amenazadas.

En [3] resaltan que el estudio de las formas de las hojas, sus pétalos y la planta entera es muy importante para la ciencia, ya que puede ayudar a medir la salud vegetal de las plantas e incluso se puede modelar el cambio climático, por lo tanto, el interés creciente del cuidado por la biodiversidad y la implementación de herramientas para la identificación automática de especies vegetales se combinan para hacer este tema oportuno. Aunque las características de las plantas mencionadas son aplicables en sistemas informáticos, el reconocimiento de especies vegetales basado en las hojas es el más recomendado [8].

Usar las hojas de las plantas para la identificación de especies vegetales, es una tarea que según [8] tiene tres problemáticas principales: la primera es que la postura arbitraria de la hoja influye mucho a la hora de identificarla, ya que si no se obtiene una buena imagen de la hoja o del foliolo (en el caso de las hojas compuestas) no se puede garantizar una buena clasificación; la segunda es la variación morfológica, dado que las formas de las hojas tienen una gran diferencia incluso entre su misma familia y clase, además cambia según la región a la que pertenece la especie vegetal; y la tercera es que las hojas que son compuestas son particularmente difíciles de reconocer, esto es debido a que la variación de una hoja compuesta no solo es causada por sus diferencias morfológicas, sino también por el cambio en el número de las hojas y la disposición de estas.

En la Universidad del Quindío y concretamente en el herbario HUUQ hay una gran variedad de procesos que tienen potencial de ser mejorados por medio de la tecnología como aplicaciones para dispositivos móviles, uno de ellos es la identificación y clasificación de especies vegetales. Por tal motivo la construcción de una aplicación móvil que ayude a identificar especies vegetales dentro de la taxonomía de la familia Amaranthaceae, será de gran utilidad para el herbario HUUQ porque agilizará el proceso de clasificación de dichas especies y evitará la pérdida de información valiosa en la recolección de las muestras. De igual manera los conceptos y componentes tecnológicos generados durante la construcción de la solución tecnológica servirán como base para realizar taxonomías de otras familias de plantas estudiadas en el herbario.

### **3. IDENTIFICACIÓN DEL PROBLEMA**

La identificación de la taxonomía de especies vegetales es una de las tareas más importantes y complejas que se realizan en el Herbario de la Universidad del

Quindío (HUQ), según lo especificado por el Ph.D. Carlos Alberto Agudelo Henao director de la Maestría en Biología Vegetal de la Universidad del Quindío y especialista en taxonomía, morfología y anatomía de plantas de la familia Amaranthaceae. La clasificación de especies vegetales es una tarea complicada debido a que se realiza manualmente e implica el análisis de numerosas características, la toma de múltiples medidas, la evaluación de la forma de los bordes de las hojas, el patrón de las venas en sus diferentes niveles, la textura de la hoja, las formas y características de la flor, el fruto y la semilla, a nivel microscópico, la forma y tamaño de sus células o del polen, el análisis de las características moleculares, genéticos, entre otros.

Las especies vegetales que recolectan los integrantes del HUQ deben ser analizadas y clasificadas a nivel de taxón, usos y demás propiedades. Esta es una tarea que debe ser realizada por expertos en el tema. Además, debido a la gran cantidad de plantas que existen, los biólogos se especializan en un número reducido de familias de plantas, por lo tanto, encontrar un biólogo experto en cierta especie puede ser una tarea difícil. Sin embargo, según Hernández en [9], las observaciones de características taxonómicas de las especies dependen de la apreciación particular del taxónomo, por lo tanto, la subjetividad humana puede también representar un problema, dando como resultado el sesgo de cualquiera de las características observadas. Por tal, la escasez mundial de taxónomos expertos aumenta aún más la demanda de herramientas de software que pueden reconocer y caracterizar las plantas a partir de imágenes de una forma más rápida y ágil [3].

El problema enunciado anteriormente lleva a realizar el siguiente cuestionamiento: ¿Es posible identificar automáticamente la taxonomía de una especie vegetal de la familia Amaranthaceae, por medio de una aplicación para dispositivos móviles basada en el reconocimiento digital de imágenes utilizando morfometría geométrica de la arquitectura foliar?

#### **4. OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS:**

##### **4.1. Objetivo general**

Desarrollar una aplicación para dispositivos móviles basada en el reconocimiento digital de imágenes y aprendizaje de máquina utilizando la arquitectura foliar para facilitar la identificación taxonómica de especies vegetales de la familia Amaranthaceae automáticamente.

##### **4.2. Objetivos específicos**

- Especificar un modelo formal de taxonomía para la familia Amaranthaceae utilizando características morfométricas de las hojas de cada una de las especies involucradas en el estudio.
- Desarrollar una metodología para la extracción de características morfológicas de las hojas por medio de reconocimiento digital de imágenes y aprendizaje de máquina, que faciliten la identificación automática de especies vegetales de la familia Amaranthaceae usando morfometría geométrica.
- Desarrollar una metodología para el reconocimiento taxonómico de especies vegetales dentro de la familia Amaranthaceae utilizando técnicas de aprendizaje de máquina.
- Desarrollar el prototipo de un sistema para dispositivos móviles de reconocimiento de patrones basado en técnicas de aprendizaje de máquina, para facilitar la identificación automática de especies vegetales de la familia Amaranthaceae.
- Validar la exactitud del reconocimiento obtenido usando el prototipo desarrollado para facilitar la identificación automática de especies vegetales de la familia Amaranthaceae.

## 5. METODOLOGÍA

Para el desarrollo de la tesis se aplicó una adaptación del método científico como un marco general que implica las siguientes etapas [13]:

1. **Observación:** Se realizaron charlas con los integrantes del herbario HUQ en las cuales se plantearon algunas problemáticas que se presentan al momento de identificar taxonómicamente las especies vegetales de la familia Amaranthaceae, y las alternativas que pueden dar soluciones a estos problemas usando la computación científica. Las secciones desarrolladas durante esta etapa fueron: Introducción, Justificación e Identificación del problema.
2. **Inducción:** Se realizó una revisión literaria detallada con el fin de establecer antecedentes de proyectos de investigación relacionados con la identificación de especies vegetales por medio de hojas, esto permitió generar las bases de conocimiento para la realización de experimentos enfocados en la solución de la pregunta de investigación. Las secciones desarrolladas durante esta etapa fueron: Caracterización de las especies vegetales de la familia amaranthaceae, Técnicas morfométricas para la extracción de características de las hojas, Enfoques machine learning aplicados al entrenamiento de

características de hojas, junto a Procedimientos y recursos para la captura de imágenes.

- 3. Hipótesis:** Sí es posible desarrollar una solución tecnológica que ayude a la identificación automática de la taxonomía de especies vegetales de la familia Amaranthaceae, por medio de reconocimiento digital de imágenes utilizando morfometría geométrica sobre las hojas, aplicando los descriptores elípticos de Fourier, análisis de textura fractal y momentos invariantes de Hu. El desarrollo de dicha solución tecnológica implica usar máquinas de vector de soporte para clasificar las especies vegetales. Finalmente, se debe desarrollar una aplicación para dispositivos móviles que capturen la fotografía de la hoja y que con la ayuda de librerías de visión computacional obtenga sus características, seguido a esto, mediante el enfoque de aprendizaje automático (máquinas de vector de soporte) se identificará taxonómicamente la especie vegetal.
- 4. Experimentación:** Se probaron las técnicas de extracción de características: Momentos invariantes de Hu, Momentos polinomios de Zernike, Descriptores Elípticos de Fourier, Momentos de las Imágenes y Análisis de textura fractal. La selección de las técnicas de extracción fue basada en un método de ensayo y error apoyado por la prueba de los clasificadores SVM, MLP (Perceptrón Multicapa) y K-NN (K- vecinos más cercanos). La base de datos usada para validar el algoritmo desarrollado fue Flavia. La secciones desarrolladas durante esta etapa fueron: Bases de datos de hojas, Técnicas de extracción de características seleccionadas, Selección de características y Selección de enfoque de aprendizaje de máquina.
- 5. Demostración:** se realizaron pruebas de precisión y exhaustividad para determinar cuáles eran los clasificadores y los hiperparámetros con los que se obtenían los mejores resultados. Se pudo determinar que el algoritmo tiene muy buenos resultados con bases de datos balanceadas y controladas. La secciones desarrolladas durante esta etapa fueron: Bases de datos de hojas, Técnicas de extracción de características seleccionadas, Selección de características y Selección de enfoque de aprendizaje de máquina.
- 6. Teoría científica:** En esta fase se realizará el informe de todas las actividades realizadas durante el desarrollo del Trabajo de Investigación, la documentación de los resultados obtenidos y sus respectivas conclusiones, además, de un paper enviado al congreso ICPRAM-2019.

## 6. CARACTERIZACIÓN DE LAS ESPECIES VEGETALES DE LA FAMILIA AMARANTHACEAE

El Herbario HUQ de la Universidad del Quindío asociado al Centro de Estudios e Investigaciones en Biodiversidad y Biotecnología (CIBUQ) cuenta con 4046 registros de plantas identificadas a nivel de especie (con un total de 563 especies únicas), 337 registros a nivel de géneros, 134 a nivel de familia y 27 identificados a otros niveles taxonómicos; el herbario tiene como objetivo almacenar especímenes vegetales, para estudiar la flora de la región Centro Andina Colombiana, con el fin de fomentar su conservación y uso sustentable, en el tiempo y en el espacio, como un patrimonio nacional.

Entre las familias estudiadas en el Herbario HUQ, existe una con un interés especial, la familia Amaranthaceae, la cual es la familia en la que se centra este trabajo. Las especies de la familia Amaranthaceae usadas en este trabajo son: *Achyranthes aspera* var. *aspera*, *Achyranthes aspera* var. *pubescens*, *Blutaparon vermiculare*, *Celosia argentea* var. *cristata*, *Celosia argentea* var. *argentea*, *Celosia virgata*, *Chamissoa altissima*, *Cyathula achyranthoides*, *Cyathula prostrata*, *Gomphrena serrata*, *Gomphrena globosa*, *Iresine angustifolia*, *Iresine diffusa*, *Iresine herbstii*, *Amaranthus blitum*, *Amaranthus caudatus*, *Amaranthus dubius*, *Amaranthus hybridus*, *Amaranthus spinosus*, *Amaranthus viridis*, *Alternanthera albotomentosa*, *Alternanthera brasiliana*, *Alternanthera lanceolata*, *Alternanthera pungens*, *Alternanthera sessilis*.

Aunque en comparación a nivel mundial hay una reducción considerable de las especies vegetales dentro de esta familia, la identificación de estas no es sencilla, los botánicos usan las características morfológicas de las plantas, como tallos, raíces, flores y hojas, para la identificación de las especies vegetales, siendo las hojas una de las más importantes sobre la identidad taxonómica de las plantas [5].

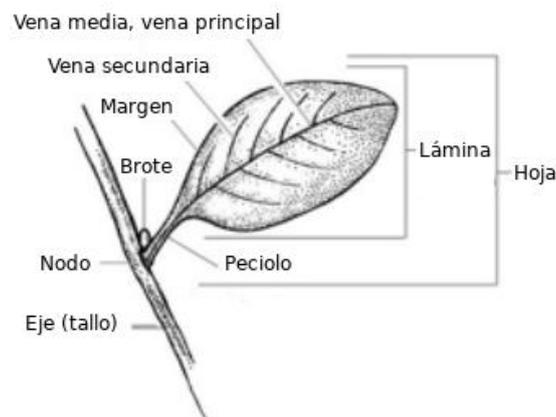
En [4] se indica que desde hace muchos años la identificación de plantas se ha basado en las características de sus órganos reproductivos, como son las flores y frutos, pero hay situaciones donde estos órganos no están disponibles, por ello, el trabajo de identificación y clasificación necesita de otras características, la hoja por ejemplo proporciona una gran cantidad de información por lo que se puede usar para reconocer plantas y asignarlas a taxones, además las características de las hojas no son estacionarias, haciéndolas aún más interesantes. En [4] definen la arquitectura de la hoja como un conjunto de características que la describen, debido a que las características propias de las hojas son muy relevantes y son de gran utilidad para la clasificación de especies.

A continuación, se describen aquellas características de las hojas que aportan valor para la identificación de especies vegetales, estas características serán más adelante llevadas a un modelo matemáticamente analizable para posteriormente desarrollar un algoritmo que permita extraer, de la fotografía de una especie vegetal,

las características representativas de cada género de la familia Amaranthaceae para su clasificación. Cabe destacar que las imágenes y la descripción de cada característica están definidas en [4].

### 6.1. Terminología de las partes básicas de una hoja

En la Figura 6 se ilustran las características más importantes de la hoja y se definen algunos términos básicos asociados a la arquitectura foliar, estas definiciones son extraídas de [4].



**Figura 6.** Partes básicas de una hoja

- **Ápice:** generalmente es la parte superior de la hoja, ocupa el 25% de la lámina.
- **Apical (distal):** que se acerca hacia el ápice.
- **Basal (proximal):** que se acerca hacia la base.
- **Base:** usualmente es el 25% más bajo de la lámina.
- **Cóncavo:** forma curvada hacia el centro de la lámina.
- **Convexo:** forma curvada que se aleja del centro de la lámina.
- **Curso de la vena:** el camino que sigue la vena de un punto a otro.
- **Eje mayor:** se denota como una línea, que se encuentra entre el ápice y la base de la hoja.
- **Eje menor:** zona de mayor ancho que se encuentra perpendicular al eje de mayor longitud.
- **Excentricidad:** es un escalar, es la relación de la distancia entre los focos de la elipse y la longitud de su eje mayor. Su valor varía entre 0 y 1. [11]
- **Lámina:** la parte plana y expandida de una hoja.
- **Margen (contorno o perímetro):** el borde de la lámina
- **Media vena:** vena principal de la hoja que se encuentra en toda la mitad de la lámina.
- **Nodo:** el lugar donde una hoja está (o estaba) unida al eje (tallo).
- **Pecíolo:** el tallo de la hoja
- **Vena Primaria:** la vena más ancha de la hoja. La vena primaria usualmente se originan en o justo arriba del pecíolo.

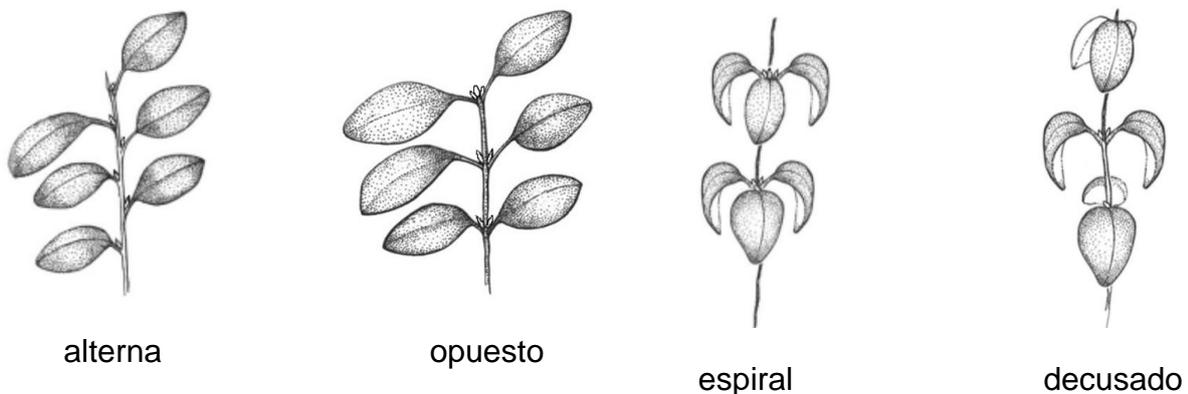
- Vena Secundaria: la siguiente clase de venas más estrechas después de la primaria.
- Vena Terciaria: la siguiente clase más estrecha de venas después de las secundarias.

## 6.2. Características asociadas a la forma y contorno

Las características asociadas a la forma o contorno de la hoja son lejos las que más se usan para clasificar especies vegetales, y es que la forma tiene quizás el poder más discriminativo en la clasificación, además que es fácil de extraer mediante procesamiento digital de imágenes y técnicas de visión computacional.

### 6.2.1. Conexión de la hoja

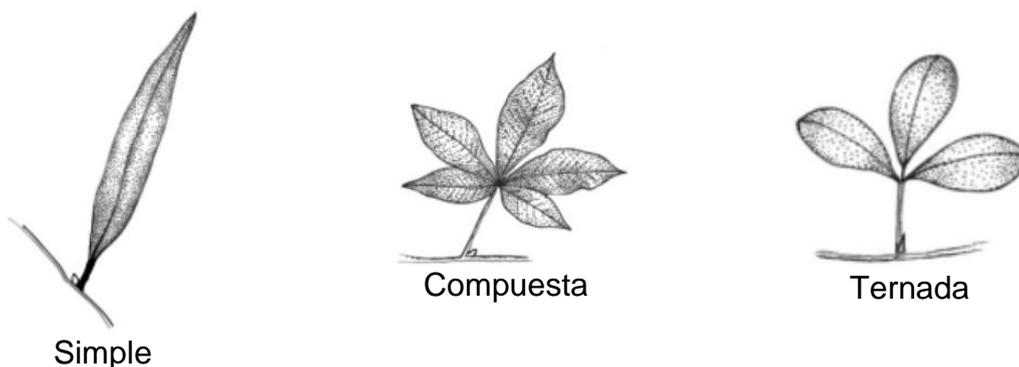
La hoja se puede conectar de diversas maneras: alterna (una hoja en cada nodo), opuesto (dos hojas en cada nodo), espiral (tres o más hojas en cada nodo), decusado (disposición en que un par de hojas opuestas crece en sentido perpendicular al par anterior, cada hoja está unida a 90 ° de las de arriba y abajo).



**Figura 7.** Tipos de conexión hoja.

### 6.2.2. Organización de la hoja

La hoja se puede componer de las siguientes maneras principalmente:



**Figura 8.** Tipos de organización de la hoja.

### 6.2.3. Tamaño laminar

El tamaño laminar se determina midiendo el área de la hoja. Se puede hacer una aproximación midiendo la longitud y el ancho de la hoja en milímetros y multiplicando la longitud x el ancho x 2/3.

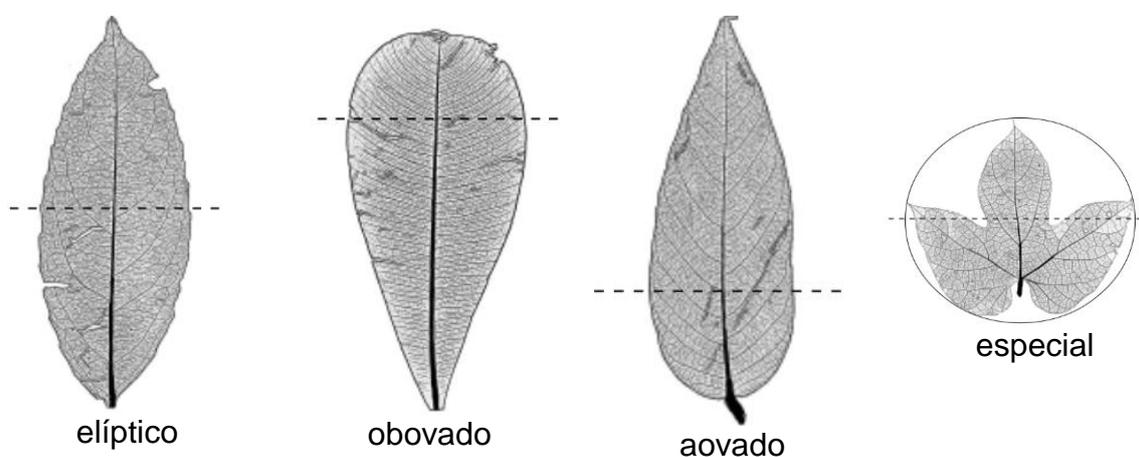
La siguiente tabla (Tabla 2) muestra los rangos de áreas para las diferentes clases de hojas:

Tipo de lámina	Área de hoja en mm <sup>2</sup> .
Leptofila	< 25
Nanofila	25 - 225
Microfila	225 - 2,025
Notofilia	2,025 - 4,500
Mesofilia	4,500 - 18,225
Macrofilia	18,225 - 164,025
Megafilia	>164,025

**Tabla 2.** Tamaños de la lámina.

### 6.2.4. Forma laminar

La forma más sencilla de describir la forma general de la lámina es ubicar el eje o, en algunos casos, la zona de mayor ancho que se encuentra perpendicular al eje de mayor longitud (eje largo):



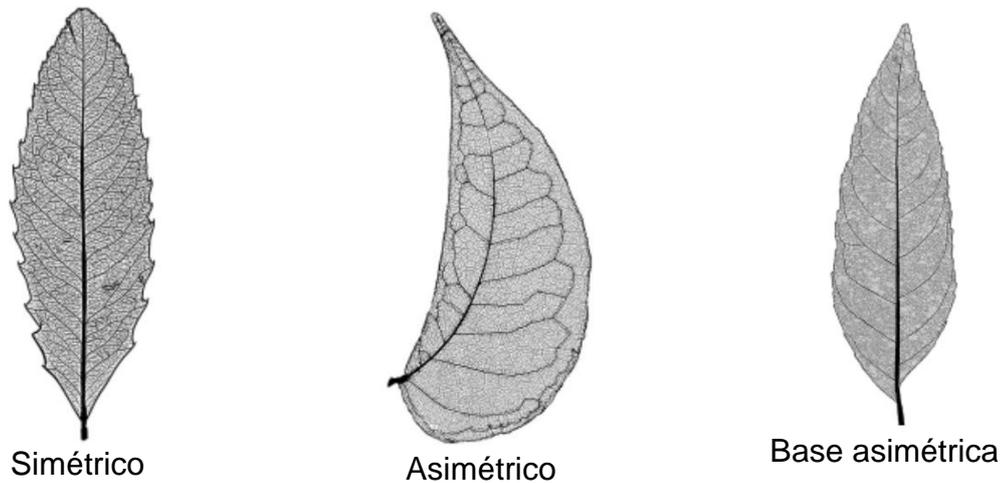
**Figura 9.** Tipos de la forma laminar de la hoja.

### 6.2.5. Simetría laminar

Simétrico: lámina aproximadamente de la misma forma en ambos lados de la vena media.

Asimétrico: tamaño o forma diferente de la lámina en cada lado de la vena media.

Base asimétrica: base de la lámina de forma marcadamente diferente en cada lado de la línea media.

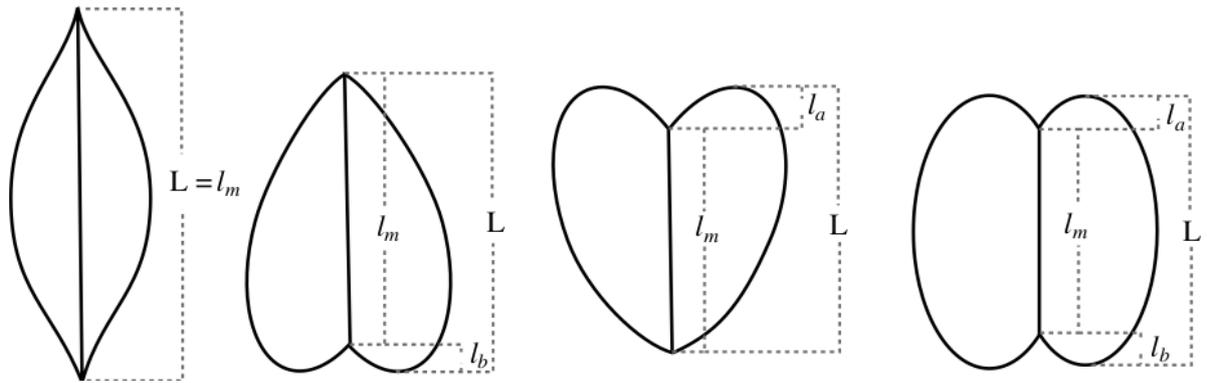


**Figura 10.** Simetría laminar de la hoja.

### 6.2.6. Ratio L:W laminar

Se obtiene al medir la longitud de la lámina (L) y dividir este número por el ancho de la lámina. A continuación todos los ratios que se pueden extraer de una lámina:

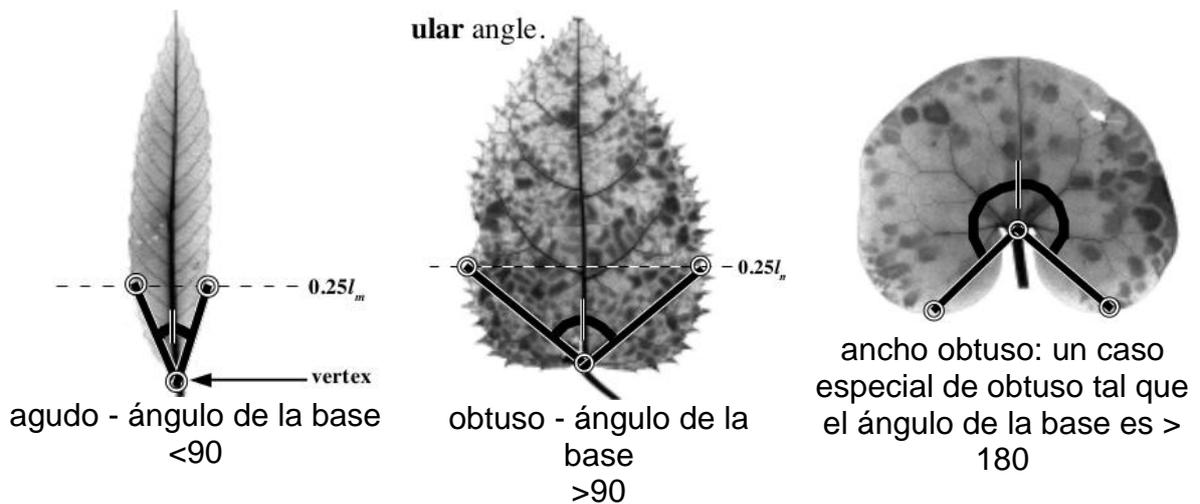
- Longitud de la vena media,  $lm$  = distancia desde la parte más proximal al punto más distal de la vena media.
- Longitud de extensión apical,  $la$  = distancia en una perpendicular desde el punto más distal de la vena media hasta la extensión más distal del tejido de la hoja. Puede ser igual a cero.
- Longitud de la extensión basal,  $lb$  = distancia en una perpendicular desde el punto más proximal de la vena media hasta la extensión más proximal del tejido de la hoja. Puede ser igual a cero.
- Longitud de la hoja,  $L = lm + la + lb$
- Mucronato: ápice que termina en un punto afilado que es la continuación de la vena media.



**Figura 11.** Medidas laminares de la hoja.

### 6.2.7. Ángulo de la base

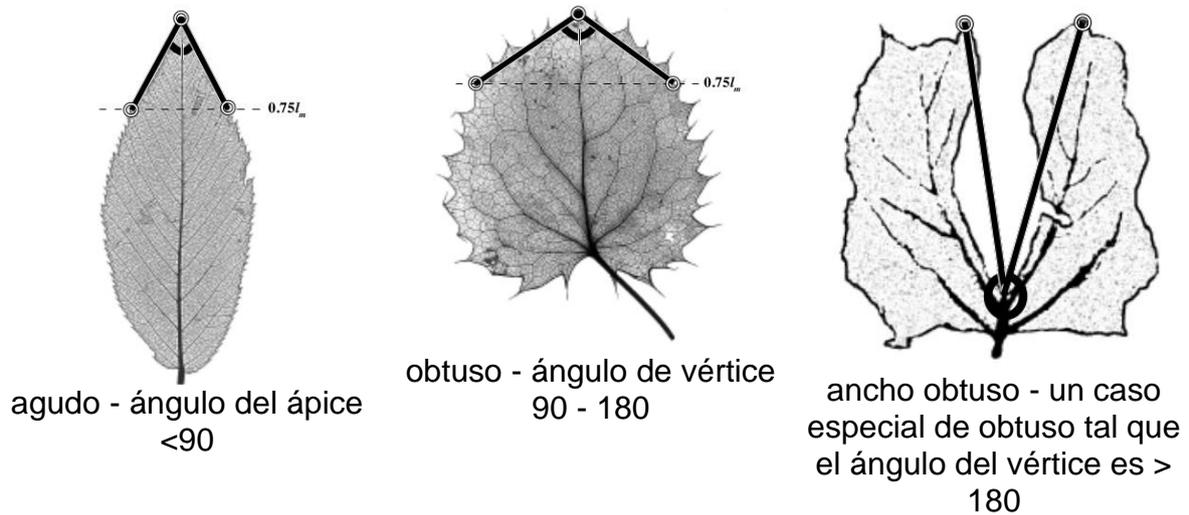
El vértice del ángulo de la base se encuentra en el centro del pecíolo en el punto donde el tejido más laminar basal toca el pecíolo. El ángulo de la base es el ángulo desde el vértice hasta los puntos donde una línea perpendicular a la vena media a  $0.25 l_m$  desde la base cruza el margen.



**Figura 12.** Diferentes ángulos de la base de la hoja.

### 6.2.8. Ángulo del ápice

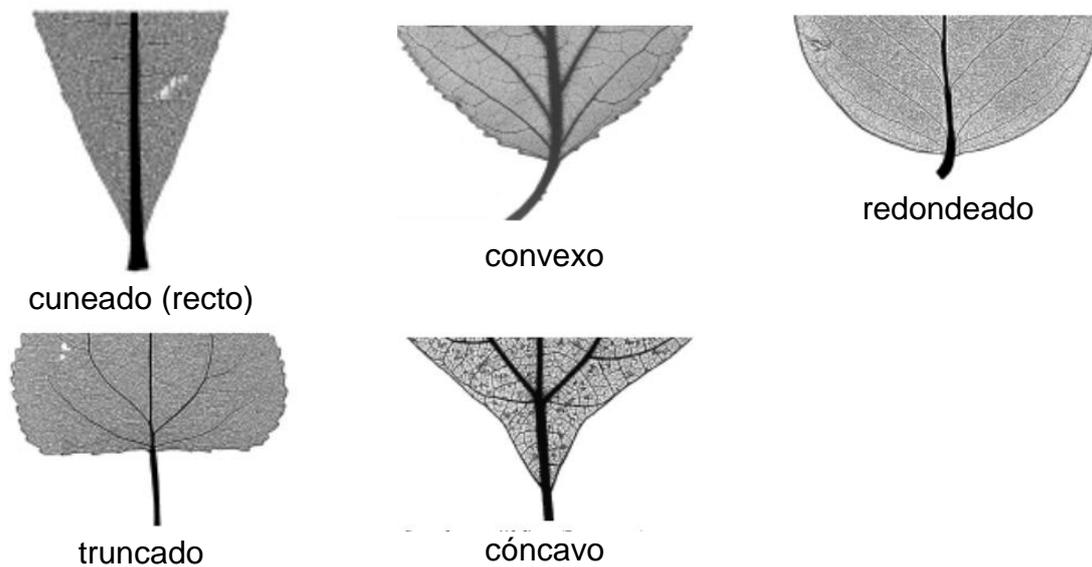
El ángulo de ápice es el ángulo desde la terminación apical de la vena media hasta el par de puntos donde se traza una línea perpendicular a la vena media a  $0,75 l_m$  desde el ápice de la lámina.



**Figura 13.** Diferentes ángulos del ápice de la hoja.

### 6.2.9. Forma de la base

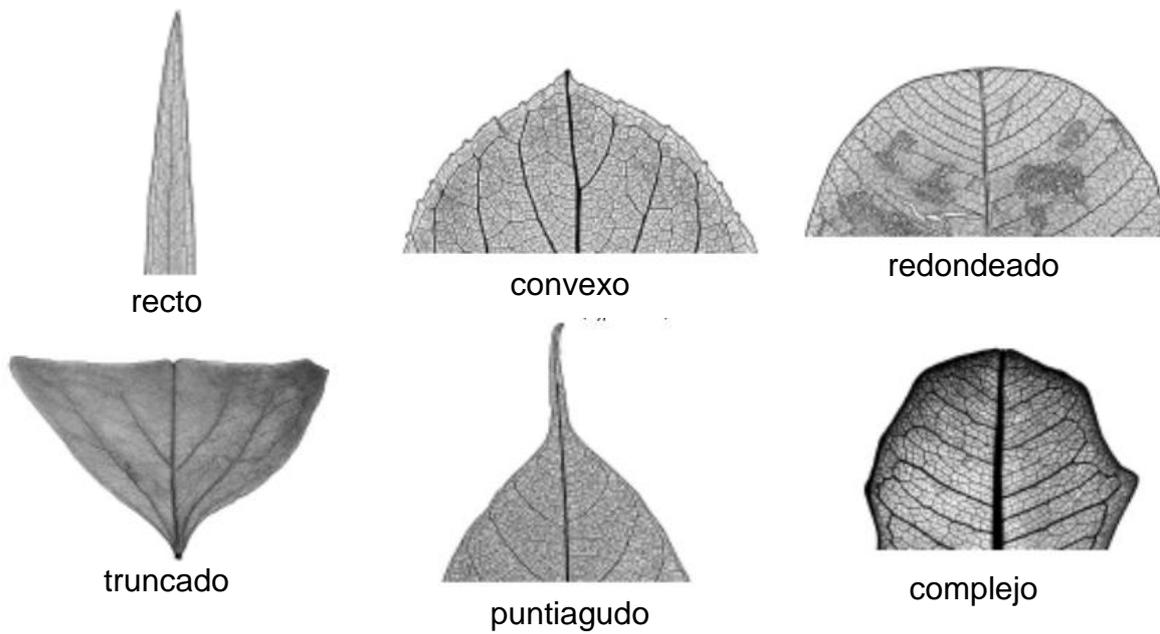
La forma de la base representa el 25% de la lámina y puede ser de diferentes tipos:



**Figura 14.** Tipos de formas de la base de la hoja.

### 6.2.10. Forma del ápice

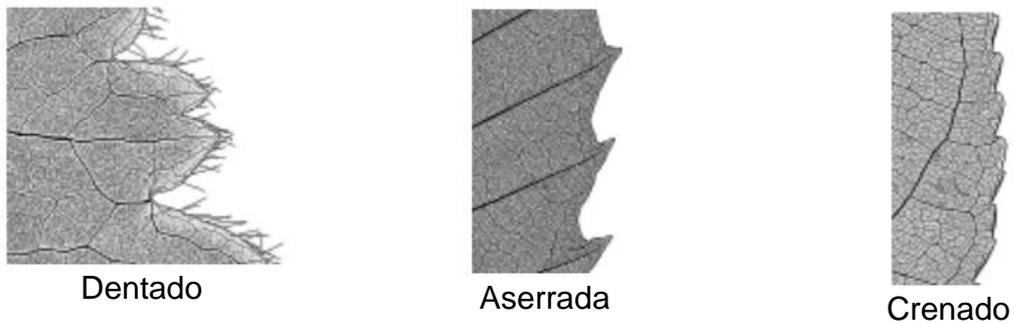
La forma del ápice representa el 25% de la lámina y puede ser de diferentes tipos:



**Figura 15.** Tipos de formas del ápice de la hoja.

#### 6.2.11. Tipo de margen (dentado)

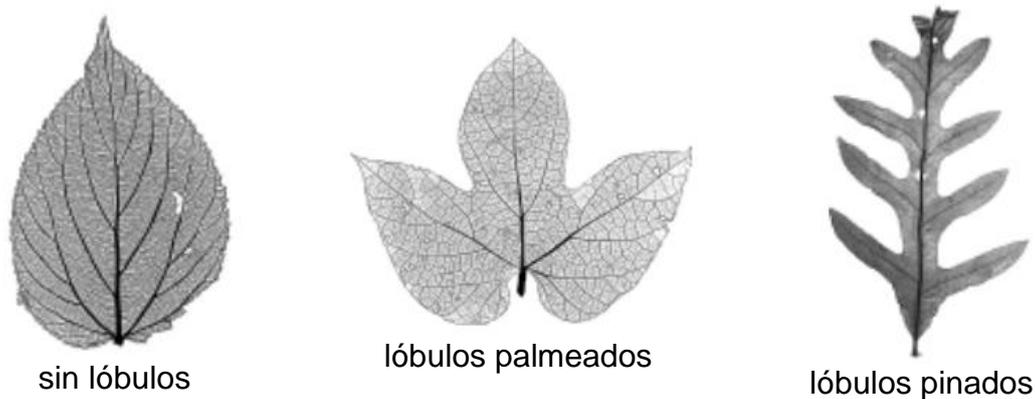
El dentado son proyecciones marginales con senos indentados de menos de 1/4 de la distancia hasta la vena central o eje largo de la hoja. Los dientes pueden ser dentados, aserrados o crenados.



**Figura 16.** Tipos de márgenes de la hoja.

#### 6.2.12. Lobulación

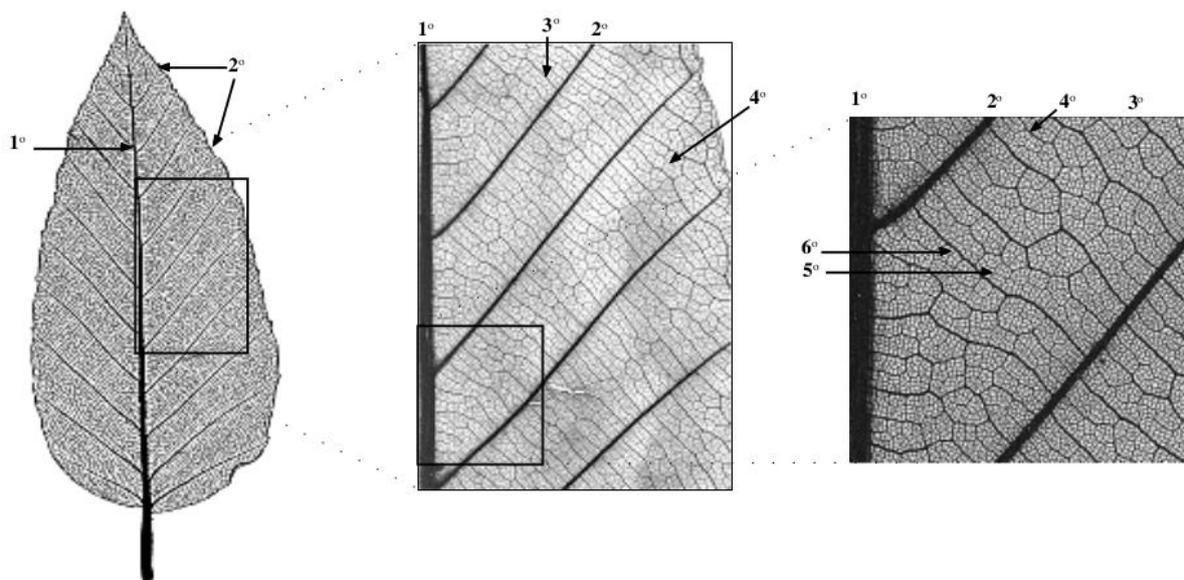
Los lóbulos son indentaciones marginales que alcanzan 1/4 o más de la distancia a la vena media, medida paralela al eje de simetría del lóbulo.



**Figura 17.** Tipos de lóbulos de la hoja.

### 6.3. Características asociadas al orden de las venas

El primer paso para describir el patrón de venación en una hoja es reconocer las categorías de venas que tienen anchos y cursos similares. La mayoría de las hojas tienen entre cuatro y siete órdenes de venación (Figura 18). Principalmente se recomienda reconocer los tres primeros órdenes de venas [4]. En general, las venas primarias y secundarias son las principales venas estructurales de la hoja, mientras que las venas terciarias son las venas que llenan el campo o lámina de la hoja. En la siguiente figura se muestra los diferentes órdenes o categorías de las venas en una hoja.



**Figura 18.** Categorías de venas.

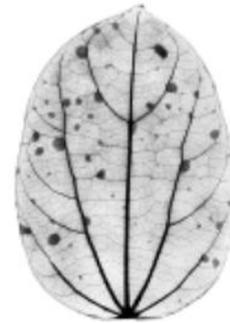
#### 6.3.1. Primer orden de venación

La vena primaria o de primer orden, es la vena más ancha en la hoja. La mayoría de las hojas tienen una vena primaria única y se llaman pinnadas. Si más de una vena se origina en o cerca de la base de la hoja, significa que la hoja tiene más de una

vena primaria. Después de reconocer la vena única más ancha de la hoja como primaria (generalmente la vena media), otras primarias se reconocen al ser el 74% o más del ancho de la primaria más ancha (en el punto de origen de la primaria más ancha).



Pinnado - con una sola vena primaria



actinodroma: tres o más venas primarias que divergen de un solo punto.



actinodroma

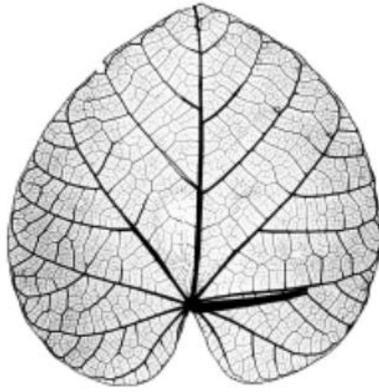


campylodromous: varias venas primarias o sus ramas, que se originan cerca de un solo punto y se dirigen en arcos fuertemente recurvados que convergen apicalmente

**Figura 19.** Tipos de venas primarias de la hoja.

### 6.3.2. Segundo orden de venación

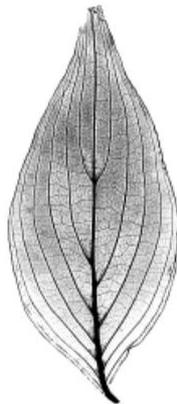
Las venas secundarias son análogas a las ramas principales de un árbol. Son el siguiente conjunto de venas en cuanto a anchura después de las venas primaria(s), también suelen afilarse a lo largo de su curso, y generalmente se extienden desde la base de la hoja o desde una vena primaria hacia el margen.



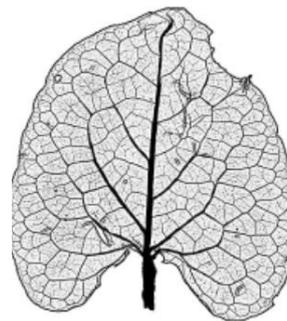
brochidódromo - venas secundarias unidas en una serie de arcos prominentes.



brochidódromo débil - venas secundarias unidas en una serie de arcos.



eucamptódromos



brochidódromo engalanada

**Figura 20.** Tipos de venas secundarias de la hoja.

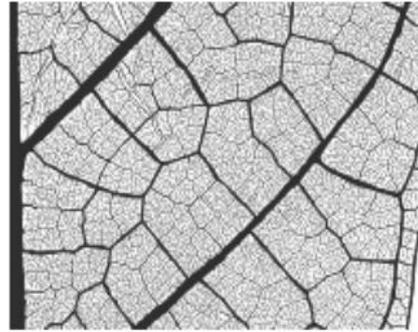
A las venas segundo orden también se le calcula otras medidas como ángulos de las venas, espacios entre las venas, número de venas basales (se aclara que el número de venas de 1, 2 y 3 orden se originan en o cerca de la base de la hoja, el cual es la parte superior del pecíolo), entre otros.

### 6.3.3. Tercer orden de venación

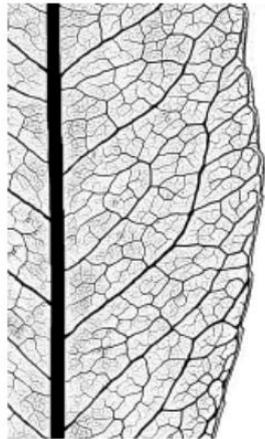
Las venas terciarias suelen ser considerablemente más estrechas que el conjunto secundario y tienen cursos que conectan las venas primarias y secundarias entre sí de manera similar en toda la hoja. Las terciarias suelen ser las venas más anchas que forman un "campo" más o menos organizado en la gran mayoría del área foliar.



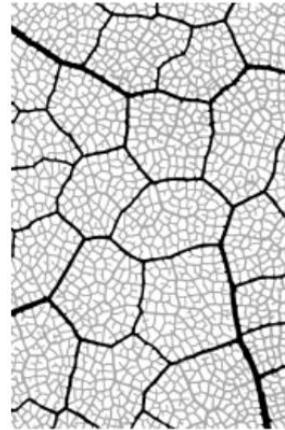
percurrente opuesto



percurrente alternativo



retículo aleatorio



reticulado poligonal regular

**Figura 21.** Tipos de venas terciarias de la hoja.

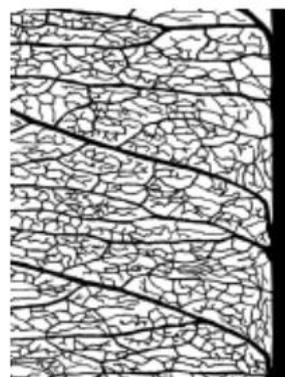
A las venas de tercer orden también se le calcula el curso de la vena (convexo, sinuoso, recto), ángulos, variabilidad de los ángulos,

#### 6.3.4. Areolación

Las areolas son las áreas más pequeñas del tejido de la hoja rodeadas de venas; en conjunto, forman un campo contiguo sobre la mayor parte del área de la lámina. Cualquier orden de venación puede formar uno o más lados de una areola.



carente - (raro)

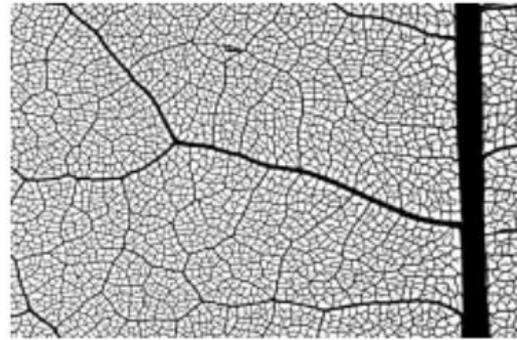


pobremente desarrolladas: areolas con muchas caras y de formas y tamaños



Areolas moderadamente desarrolladas de forma irregular, de tamaño más o menos variable

altamente irregulares.



bien desarrolladas - areolas de tamaño y forma relativamente consistentes.

**Figura 22.** Tipos de areolas de la hoja.

Otras características de valor para la clasificación de especies vegetales es la textura de la hoja, las células vegetales de la hoja presentes en la lámina, las cuales se pueden extraer a través de un corte en la misma.

#### **6.4. Características y descripción de la familia Amaranthaceae**

Según [2], las hojas de las plantas de Amaranthaceae son: pecíolo 0.2-1.8 cm de largo, estrigoso; lámina ovada, ampliamente obovada o elíptica, estrigosa, base cuneada, ápice agudo a mucronado, haz y envés estrigosos, venación primaria y secundaria estrigosa, axilas con aglomerados de tricomas. Como se puede ver, las características de las hojas de esta familia son bastante variadas y dificultan la tarea de identificación, por lo cual se puede tomar en cuenta un taxón más específico como el género y sus especies.

A continuación, se hace una descripción de los géneros Amaranthaceae presentes en el herbario HUQ, cabe destacar que toda la información de cada género es obtenida de [2].

##### **6.4.1. Achyranthes**

Generalmente son hierbas erectas, con tricomas pluricelulares, uniseriados. Hojas opuestas y pecioladas; lámina elíptica u obovada. El género comprende cerca de diez especies propias del trópico y subtrópico del Viejo Mundo, con unas pocas especies que alcanzan el Nuevo Mundo. En Colombia se conoce una sola especie de amplia distribución. Hierba erecta, 30-80 cm de altura.

#### **6.4.2. Blutaparon**

Se compone de hierbas heliófilas de tallo rizomatoso, con raíces adventicias. Hojas opuestas, sésiles, suculentas, bases concrecentes, con un anillo de tricomas en la cara adaxial. El género Blutaparon se distribuye en las costas de la zona tropical de América y África. El género consta de cuatro especies y siete variedades. En Colombia se encuentran dos especies en las costas Caribe y Pacífica que se distribuyen por las zonas costeras, dunas, médanos y playas marinas.

#### **6.4.3. Celosia**

Son hierbas de 0.5-2 m de altura. Hojas alternas. El género Celosia comprende cerca de 50 especies que crecen en las regiones tropicales y subtropicales de todo el mundo, pero está mejor representado en el continente africano. En Colombia existen dos especies distribuidas en zonas cálidas y húmedas de las regiones Andina y Caribe, entre el nivel del mar y 1300 m de altitud. Medicinalmente se utiliza para la inflamación de la garganta y la decocción de sus flores secas se usa para la disentería y como diurética.

#### **6.4.4. Chamissoa**

Se representa mediante hierbas erectas o trepadoras, tricomas ramificados o simples. Hojas alternas, pecioladas. El género Chamissoa se distribuye en el neotrópico. En Colombia se registran dos especies ampliamente distribuidas en bosques secundarios y zonas perturbadas, entre el nivel del mar y 2500 m de altitud.

#### **6.4.5. Cyathula**

Se compone de hierbas rastreras, erectas o subarborescentes. Hojas opuestas, pecioladas. El género es cosmopolita y comprende cerca de 25 especies, la mayor parte de las cuales están en el viejo mundo. En Colombia hay dos especies que se distribuyen por todas las regiones geográficas, entre el nivel del mar y 1900 m de altitud.

#### **6.4.6. Gomphrena**

Hierbas erectas a postradas, con tricomas pluricelulares, uniseriados o abietiformes. Hojas opuestas. El género está compuesto por aproximadamente 100 especies que se distribuyen principalmente en el neotrópico. Para Colombia se reconocen dos especies nativas y una cultivada, las que se distribuyen por todas las regiones geográficas, desde el nivel del mar hasta 2600 m de altitud, en áreas alteradas y sabanas.

Se cultiva en todo el mundo como planta ornamental. Por otra parte, algunos usos medicinales incluyen el uso de las hojas como antimaláricas y para la caída de los

dientes; toda la planta se utiliza como antipirética y diaforética; las hojas y las flores se usan como emolientes y para controlar la salivación sangrosa

#### **6.4.7. Iresine**

Hierbas erectas a bejucosas. Hojas opuestas, pecioladas. Algunos autores calculan que Iresine puede contener entre 40 o 70 especies, dado que algunas especies son muy variables morfológicamente y que se requiere una adecuada delimitación de las entidades específicas. Es un género con distribución pantropical. Para Colombia se registran tres especies distribuidas en todas las regiones geográficas, una de las cuales es cultivada, desde el nivel del mar hasta 3500 m de altitud.

#### **6.4.8. Amaranthus**

Hojas alternas, pecioladas; lámina romboide, elíptica u obovada, ápice obtuso, brevemente mucronado. Es un género con cerca de 60 especies que se distribuyen por todo el mundo, pero la mayor parte de ellas están concentradas en América. En Colombia se encuentran siete especies nativas y dos cultivadas, distribuidas por todas las regiones naturales, y desde el nivel del mar hasta 2700 m de altitud. En Colombia se cultiva como planta ornamental.

#### **6.4.9. Alternanthera**

Hierbas erectas, postradas o rastreras, glabras a tomentosas. Hojas opuestas, sésiles o pecioladas, tricomas simples. El género Alternanthera posee cerca de 100 especies, la mayor parte de las cuales se distribuyen en el neotrópico. En Colombia, Alternanthera está representado por doce especies nativas, las cuales crecen preferentemente en hábitats alterados. Las especies se distribuyen en todas las regiones naturales de Colombia, desde el nivel del mar hasta 3200 m de altitud, pero están mejor representadas por debajo de 1500 m.

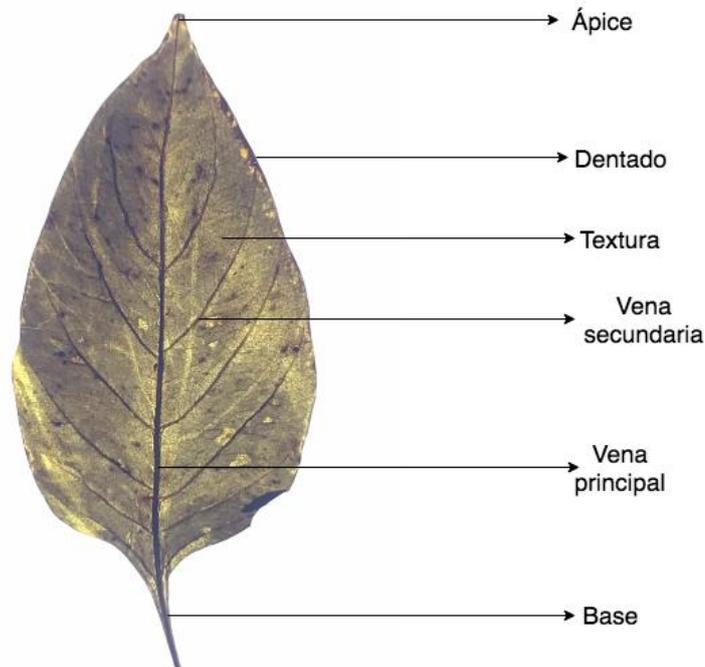
Dentro de Alternanthera hay gran variación morfológica a nivel intraespecífico, lo que ha dificultado su taxonomía y el reconocimiento de las diferentes entidades que lo componen.

## **7. TÉCNICAS MORFOMÉTRICAS PARA LA EXTRACCIÓN DE CARACTERÍSTICAS DE LAS HOJAS**

Uno de los retos más importantes al momento de identificar especies vegetales por medio de imágenes de las hojas, es la extracción de características. Las hojas cuentan con numerosas características morfológicas que permiten a los biólogos expertos identificar las especies vegetales, por lo cual está no es una tarea que cualquier persona pueda realizar [3].

En [4], especifican que las características de las hojas que se deben tomar en

cuenta para una identificación taxonómica basada en la arquitectura foliar son: el dentado, el patrón de la venas primarias y secundarias, orden de venación, rango de venación, los tipos de lóbulos, características de la lámina, características del peciolo, ángulo del ápice, forma de la base y textura de la hoja. Algunas de estas foliares se resaltan en la Figura 23.



**Figura 23.** Características foliares de la hoja de 03-Celosia argentea var. cristata.

En [11], se han centrado en la morfometría geométrica para extraer características las hojas, obteniendo datos como: excentricidad, factor de forma, relación de longitud del eje secundario con la longitud del eje principal, relación de longitud de casco convexa del perímetro, extensión, rectangularidad, relación del perímetro de la longitud del eje mayor, relación del área, relación de extensión y relación de regularidad. Este tipo de características morfométricas son bastante usadas en técnicas de visión artificial.

En el campo de la visión artificial se han usado diversas técnicas para la extracción de las características de las hojas, en donde regularmente se necesita de imágenes en escala grises o binarizadas (blanco y negro). Du *et al.* por ejemplo, desarrollaron una técnicas denominada *Ring Projection Wavelet Fractal Feature (RPWFF)* para la extracción de las características de las imágenes de las hojas. Esta técnica reduce la dimensionalidad de la imagen de dos a una por medio de un método llamado proyección de anillos. Usando el vector resultante después de aplicar dicho método se generan umbrales de diferentes niveles, y con base a ellos se calcula la dimensión fractal en cada una de los vectores. Con el anterior proceso lograron obtener información referente a la venación, margen y textura de las hojas. Para ampliar las características obtenidas usando RPWFF aplicaron Momentos

Invariantes de HU, obteniendo siete características que en la imagen principal binarizada son invariantes al zoom y a la rotación, estos momentos hacen referencias al contorno de la hoja [12].

Valliammal y Geethalakshmi, proponen una técnica que llaman Extracción de Fusión de Características (en inglés *Feature Fusion Extraction*) para la recuperación de imágenes de hojas [13]. Esta técnica extrae características basadas en la forma, color y textura, donde los cálculos morfométricos toman un valor importante. El color es tomado en cuenta dado que el objetivo del estudio de Valliammal y Geethalakshmi es la recuperación de imágenes de hojas de especies vegetales y no su identificación o clasificación. Otros estudios relacionados con el procesamiento de imágenes de las hojas en donde el color es una de las características más importante, es la detección de enfermedades de las hojas [14]. Las 13 características tomadas en cuenta en el trabajo descrito en [13] son: rectangularidad, esfericidad, relación de ejes, diámetro, complejidad de la hoja (con o sin folíolos), perímetro, energía, entropía, contraste, homogeneidad, media, oblicuidad y kurtosis. Para determinar cuáles de las 13 características aportaban más a la clasificación usaron algoritmos genéticos. Así mismo, en [15] se describe otro estudio basado en técnicas morfométricas, en donde usaron Descriptores Elípticos de Fourier (DEF) y momentos de las imágenes para extraer las características del contorno de las hojas. En este último estudio destacan el área de las hojas como una de las características más importantes al momento de la identificación o clasificación de las especies vegetales, lo que hace inferir que la base de datos usada debería ser controlada.

Dos de las técnicas que más se destacan en la extracción de características son DEF y momentos de las imágenes, una muestra de esto son los trabajos realizados en [12 - 15], otro ejemplo de esto, es el trabajo realizado por Liu y Kan en [16], en el cual usan Descriptores Elípticos de Fourier y Momentos Invariantes de Hu para la extracción de las características de hojas por medio de fotografías. Otras técnicas relacionadas a momentos de las imágenes como *Scale-invariant feature transform* (SIFT) han sido usadas con éxito en estudios enfocados en la identificación de especies vegetales [17], al igual que Análisis Elípticos de Fourier (AEF) en [18], donde resaltan que a pesar de que DEF se destaca ampliamente en la literatura, AEF fue la técnica que les permitió obtener mejores resultados.

Como se ha podido notar, todas las técnicas nombradas con anterioridad están asociadas a la extracción de características del contorno (forma, área, etc), dejando en un segundo plano características tan importantes como la venación y textura. En estudios como [19], han dado especial énfasis en estas dos propiedades por medio de un proceso de extracción basado en la generación de puntos en dos fases. La primera es generar los puntos que forma el margen de la hoja por medio de una relación espacial entre vecinos más cercanos. En la segunda fase aplican la técnica conocida como detector de Harris para obtener los puntos más relevantes de la

hoja, principalmente asociados con la textura y venación. Otra técnica asociada a la venación y textura que ha sido documentada como exitosa en [20, 21], es el Kernel Descriptor (KDES), el cual usa como base los puntos existentes en un Kernel formado por varios pixeles. KDES ha sido acompañado con métodos de agrupamientos como K-mean y generación de parches para obtener las características de interés con éxitos.

Aunque este trabajo está asociado al estudio de técnicas morfométricas para la extracción de características de las hojas, es importante tener en cuenta que existen otros enfoques que ha reportado casos de éxito, y tiene el mismo objetivo de ayudar a la identificación de plantas. Técnicas como generación de parches en [22], optimización de colonia de hormigas en [23] o Redes Convolucionales y Desconvolucionadas en [24], también se han documentado éxito, incluso en [24] reportan precisiones en la identificación de 0.99.

Para finalizar la referencia a las técnicas morfométricas de interés, se destaca la investigación realizada en [5], donde se centran en la revisión de algunas de las técnicas usadas para la extracción de características de hojas usando imágenes, algunas de esas son: señales y medidas lineales, características de la forma, ajuste de polígonos y dimensiones fractales, algoritmo de análisis de espacio de escala, un algoritmo de suavizado y detección de bordes (en inglés *smoothing and edge detection algorithm*) y dimensiones fractales usando múltiples capas. Las dimensiones fractales es una de las técnicas que más se pueden destacar, ya que permite representar matemáticamente detalles de las hojas asociados a la venación y textura de la hoja.

En toda literatura revisada para el desarrollo de este proyecto se han encontrado estudios totalmente enfocados en la identificación especies vegetales, sin tomar en cuenta la taxonomía de las mismas, algo que sí realizamos en este proyecto. Incluso en el concurso descrito en [22], destacan que ninguno de los participantes tomó en cuenta la taxonomía de las especies vegetales para realizar su identificación.

## **8. ENFOQUES MACHINE LEARNING APLICADOS AL ENTRENAMIENTO DE CARACTERÍSTICAS DE HOJAS**

Al extraer las características de las especies vegetales se debe buscar el enfoque de aprendizaje de máquina (clasificador) que brinde la precisión y fiabilidad más alta. Amplia variedad de trabajos han sido desarrollados para ayudar en la identificación de especies vegetales por medio de sus hojas, y con ellos el uso de diversos enfoques de aprendizaje de máquina.

Máquinas de vector de soporte (en inglés Support Vector Machines - SVM) es uno de los clasificadores con los que más se ha tenido resultados exitosos en este tipo de investigaciones. En [13, 20, 23, 25], aplicaron dicho clasificador obteniendo precisiones de 0.94, 0.98, 0.95 y 0.91 respectivamente. En otros casos como [21], se a usando SVM sin obtener precisiones tan altas, 0.715, sin embargo, hay que destacar que el entrenamiento fue realizado sobre la base de datos ImageClef (base de datos basada en fotografías con fondos complejos de hojas, frutos, flores y plantas), en la cual se suelen obtener precisiones bajas. La precisión más alta identificada usando SVM es la obtenida por Lee, Chan, Mayo y Remagnino en [24], donde se logra entrenar información generando un 0.995 de acierto, no obstante con los mismos datos Perceptrón Multicapa (en inglés Multilayer Perceptron - MLP) obtiene un precisión de 0.998. Otra solución alternativa con SVM fue basada en la generación de características usando un Modelo de Bolsa de Palabras (del inglés, *Bag of Words*) personalizado, obteniendo una precisión de 0.9237 [17]. En estas investigación, se centran directamente en la identificación de especies vegetales, sin embargo en trabajos como [26], el objetivo es la identificación taxonómica de las plantas con ayuda de SVM, la precisión obtenida a nivel de especies es de 0.7321 y en cuanto a género es 0.8488 (con una especie por género para evitar errores de clasificación con plantas parecidas en sus hojas).

Vecino más cercano (en inglés Nearest Neighbor - K-NN) es otro clasificador que suele usarse en la identificación de especies vegetales. La investigación realizada por Du, Zhai y Wang resalta una precisión de 0.96 obtenida con K-NN [12]. Por otra parte, en [19], usan Vecino más cercano para clasificar las especies vegetales de una base de datos propia llamada *Middle European Woody plants (MEW)*, obteniendo una presión de 0.8891. El mismo clasificar fue usando en [15] teniendo como resultado un acierto de 0.65.

Las redes neuronales también se destacan en los proyectos de investigación enfocados en la identificación de especies vegetales. Uluturk y Ugur utilizaron redes neuronales probabilísticas (PNN) para la identificación de plantas obteniendo una precisión de 0.92 [11]. En [16] usaron redes de creencia profunda (DBN) para el entrenamiento de 220 tipos de hojas, obteniendo una precisión de 0.939, hay que tener en cuenta que para usar DBN como clasificador se necesita de muchas instancias para el entrenamiento. Como se ha mencionado con anterioridad, en [24] se han obtenido resultados destacados en identificación de especies vegetales usando Redes Neuronales Convolucionales (CNN) para extraer datos crudos de las imágenes de las hojas de las plantas y Redes Descovolucionales (DN) para generar la intuición de las características extraídas. La precisión obtenida en este último proyecto de investigación fue de 0.998 usando MLP.

Aunque SVN, K-NN y diferentes tipos de redes neuronales (ANN) son los enfoques de aprendizaje de máquina que más sobresalen, hay opciones alternativas como: Mezcla de Gaussianas, el cual es usado en [27] obteniendo una precisión de

0.9237; Análisis Discriminante Lineal (del inglés, Linear Discriminant Analysis - LDA), en [18] con precisión de 0.843; y Vector de Fisher, en [22] con precisión de 0.607.

Todos los anteriores estudios nombrados tiene la característica común que realiza entrenamiento con aprendizaje supervisado. En [14] usan aprendizaje no supervisado por medio de K-Means para la clasificación de grupos de plantas por medio de las características de las hojas, y aunque no resaltan la precisión obtenida, destacan la investigación como caso de éxito.

Para seleccionar el clasificador adecuado para entrenar las características extraídas en este proyecto de investigación para ayudar a identificar taxonómicamente las especies de la familia Amaranthaceae, con base a fotografías de las especies del herbario HUQ, se tuvieron en cuenta las precisiones reportadas en algunos proyecto de investigación, las técnicas de extracción de características empleadas y las bases de datos usadas, por tal, la precisión no es el único que se consideró. Tomando en cuenta todo lo anterior, los enfoques de aprendizaje de máquina que se probaron en este proyecto fueron SVM, K-NN y ANN, dado su amplio uso en proyectos de investigación asociados y los destacados resultados reportados.

## **9. PROCEDIMIENTOS Y RECURSOS PARA LA CAPTURA DE IMÁGENES**

En esta sección se describe el procedimiento elegido para la captura digital de las imágenes de las especies vegetales, para esto, se mencionan algunos trabajos relacionados con el tema.

### **9.1. Trabajos relacionados**

En [28] se define una metodología para procesar imágenes de hojas, que va desde su adquisición o captura mediante una cámara digital, hasta el análisis y obtención del porcentaje de severidad de la enfermedad, pasando la imagen por diferentes filtros. Se resalta que la captura digital de la hoja debe cumplir ciertas condiciones de iluminación para evitar brillos o sombras, además se debe tener en cuenta la ubicación de la hoja y el fondo sobre el cual será fotografiada. En [25], por ejemplo, definieron que cada hoja recolectada debe ser la tercera hoja sana y madura de ramas viejas que haya estado completamente expuesta a la luz solar.

En la investigación descrita en [3], usaron la visión artificial y enfoques de aprendizaje de máquina para realizar la clasificación de especies vegetales del género *Camellia* con base en las características obtenidas de las hojas. Para resaltar el patrón de las venas de las hojas, han colocado las hojas en tubos de vidrio y les han añadido Hidróxido de Sodio al 10% (NaOH) en cantidad suficiente para cubrir su lámina, a 70-80oC, durante 3-4 horas. Dado que la textura de las

hojas puede diferir entre las especies, las hojas más gruesas fueron tratadas durante períodos de tiempo más largos.

Por otra parte, la base de datos ImageCLEF [22] está compuesta por imágenes de hojas, flores, frutos y tallos. Las imágenes son 58% fotografías y 42% escaneos. Sin contar que las imágenes no son capturadas bajo algún proceso definido y las aportan variada cantidad de personas. La totalidad de las imágenes en esta base de datos es 26 mil. Se resalta de igual manera que la digitalización de las hojas, flores, frutos y tallos tiene una buena aceptación si se hace por medio de capturas obtenidas por una cámara digital.

Finalmente, Global Plants es una colección de plantas creada por la colaboración de JSTOR y aproximadamente 300 herbarios. Esta colección contiene 2.48 millones de fotografías de amplia variedad de especies. Las muestras usadas para el ejercicio de recomendación taxonómica fue creada realizando un escaneo plano con un Epson Expression 10000XL. Las muestras fueron fotografiadas a una resolución de 5144 x 3599 píxeles y 300 dpi. Para cada especie, los especímenes fueron seleccionados para cubrir un rango de material de herbario típico, incluyendo hojas plegadas rotas, hojas dañadas por herbívoros y hojas superpuestas. [26]

Los anteriores estudios definen que la digitalización de las hojas se puede hacer por medio de una cámara o un escáner, en este trabajo se opta por la cámara y especialmente se recomienda una cámara de un dispositivo móvil dado que es un dispositivo de fácil acceso y permite agilizar la captura digital de la hoja de la especie vegetal y que se puede articular con una aplicación móvil que se encargue de realizar su clasificación.

## **9.2. Procedimiento y recursos para la captura digital de las imágenes de las hojas**

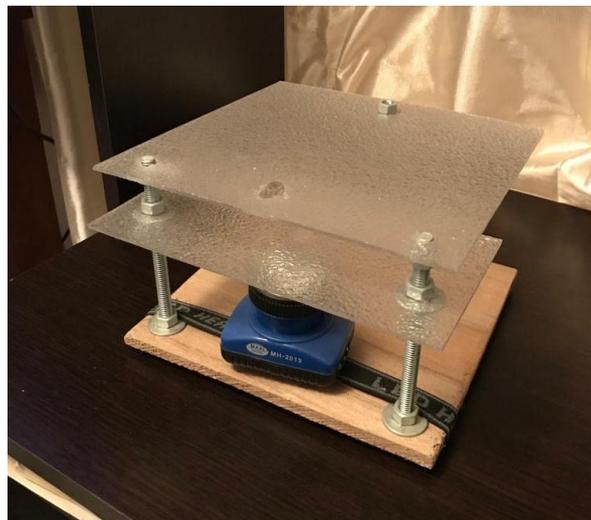
La captura digital de las fotografías de las hojas de las especies vegetales es parte fundamental en todo el proceso de clasificación de las mismas. Definir un procedimiento para la captura de las fotografías implica especificar los pasos que se deben seguir y los recursos necesarios para tal fin. Las fotografías a capturar corresponden a las especies vegetales de la familia Amaranthaceae del HUQ y aquellas que no hagan parte del herbario pero que requieren ser clasificadas dentro de la taxonomía de esta familia.

El punto de partida es la generación de una base de datos con fotografías de hojas de la familia Amaranthaceae disponibles en HUQ, esta base de datos será el cimiento para entrenar el algoritmo de clasificación, y posteriormente, el resultado de este entrenamiento dará los parámetros y la configuración adecuada para clasificar futuras especies mediante técnicas de *machine learning*.

Para la captura de la fotografía se debe tener en cuenta que lo ideal es ubicar la hoja de la especie vegetal sobre un fondo blanco e iluminado. El fondo blanco tiene

la finalidad de resaltar la características de las fotos y facilitar la segmentación de la foto, ya que si el fondo no es blanco se puede considerar como ruido en la fotografía, y esto puede afectar el contorno de la hoja y su forma. La iluminación del fondo es fundamental (todo el fondo debe tener la misma intensidad para evitar brillos y sombras) para resaltar las venas de la hoja, principalmente, las venas de segundo y tercer orden, así como la geometría fractal o la areolación de la hoja. Una buena captura fotográfica de la hoja garantiza una mejor clasificación.

Por lo enunciado con anterioridad se construyó un dispositivo de fácil uso, llamado *Eliminador de Ruido de la Arquitectura Foliar* (ERAF), que facilita la captura de las fotografías a las hojas. Inicialmente se construyó un dispositivo (ERAF00) sencillo (Figura 24), con una base de madera y una linterna ubicada sobre dicha base, encima se ubicaron dos acrílicos transparentes, uno bajo para ubicar la hoja y otro más alto para ubicar el dispositivo móvil, en dicho acrílico se hizo un hueco para no interrumpir ni bloquear el lente de la cámara. La idea es tomar la fotografía a la hoja iluminada por la linterna con el dispositivo móvil fijo. Con este dispositivo se tuvo algunos problemas, el más importante es que el acrílico no era totalmente transparente por lo que cuando se iluminaba la hoja con la linterna, en la hoja quedaban las sombras producidas por el diseño del acrílico, añadiendo ruido y dañando la venación de la hoja capturada en la fotografía.



**Figura 24.** ERAF00.

La segunda versión (ERAF01) del dispositivo (Figura 25) se diseñó de tal manera que estuviera compuesto por una base totalmente blanca con iluminación LED y un acrílico transparente en la parte superior, este acrílico tiene la finalidad de servir como base para ubicar la cámara del dispositivo móvil y que quede fijo, que el dispositivo móvil esté fijo garantiza que la foto quede bien enfocada y además, estandariza la distancia a la que se toma la fotografía para dar consistencia a las imágenes. El acrílico transparente está sobre cuatro tornillos que lo sostienen, cada tornillo tiene su propio perno, el perno permite ubicar el acrílico a la distancia deseada.



**Figura 25.** ERAF01.

Adicionalmente se cuenta con una regla que permite ajustar cada uno de los cuatro pernos para que el acrílico quede bien balanceado y a la misma distancia de la base en todos sus bordes. El objetivo del dispositivo es capturar una buena fotografía de la hoja de una manera sencilla, bajo un ambiente controlado.

El procedimiento para la captura de la imagen es el siguiente:

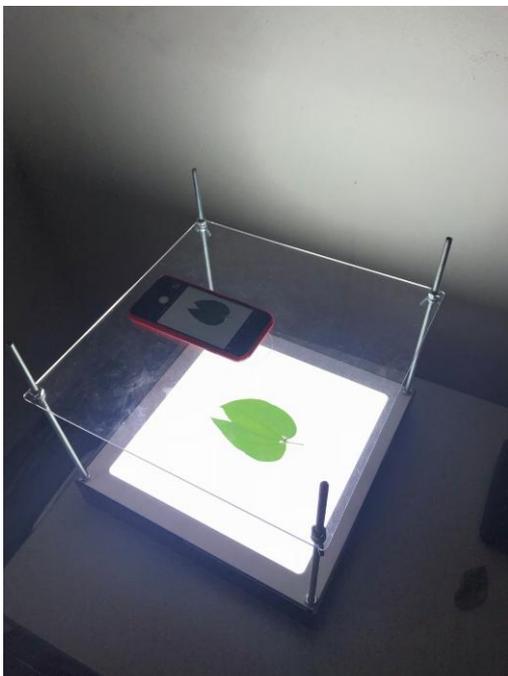
1. Se obtiene la tercera hoja sana y madura una rama vieja que estén completamente expuestas a la luz solar [25].
2. Ubicar la hoja verticalmente con el ápice en la parte alta sobre el centro de la base blanca luminosa de ERAF.
3. Ubicar el dispositivo móvil sobre el acrílico transparente en posición retrato. El acrílico tiene una marca donde se sugiere ubicar la cámara del dispositivo móvil.
4. Ubicar el acrílico a la distancia requerida, para la toma del contorno de la hoja se sugiere una distancia de 18 cm aunque puede variar según el tamaño de la hoja. Para la captura del zoom de la hoja se sugiere la misma distancia pero usando un zoom de al menos 3x.
5. Capturar la fotografía del contorno. La fotografía se debe capturar sin flash, y se recomienda tomar la foto en modo automático.
6. Capturar la fotografía del zoom. La fotografía se debe capturar sin flash, y se recomienda ajustar el nivel de apertura del lente de la cámara para capturar la luz adecuada que resalte con mayor nitidez las venas de las hojas.



**Figura 26.** ERAF01 encendido.



**Figura 27.** ERAF01 con una hoja.



**Figura 28.** ERAF01 con un dispositivo móvil listo para capturar la foto.



**Figura 29.** ERAF01 con un dispositivo móvil listo para capturar la foto.

Con el dispositivo antes mencionado se fotografiaron 619 especies vegetales de la familia Amaranthaceae del HUQ con diferentes dispositivos móviles (iPhone 8 plus, iPhone 7 y Google Nexus 5), a cada especie vegetal se le capturaron dos fotografías, una de la hoja completa para extraer características asociadas a su forma y contorno, y otra con zoom para extraer características asociadas a su

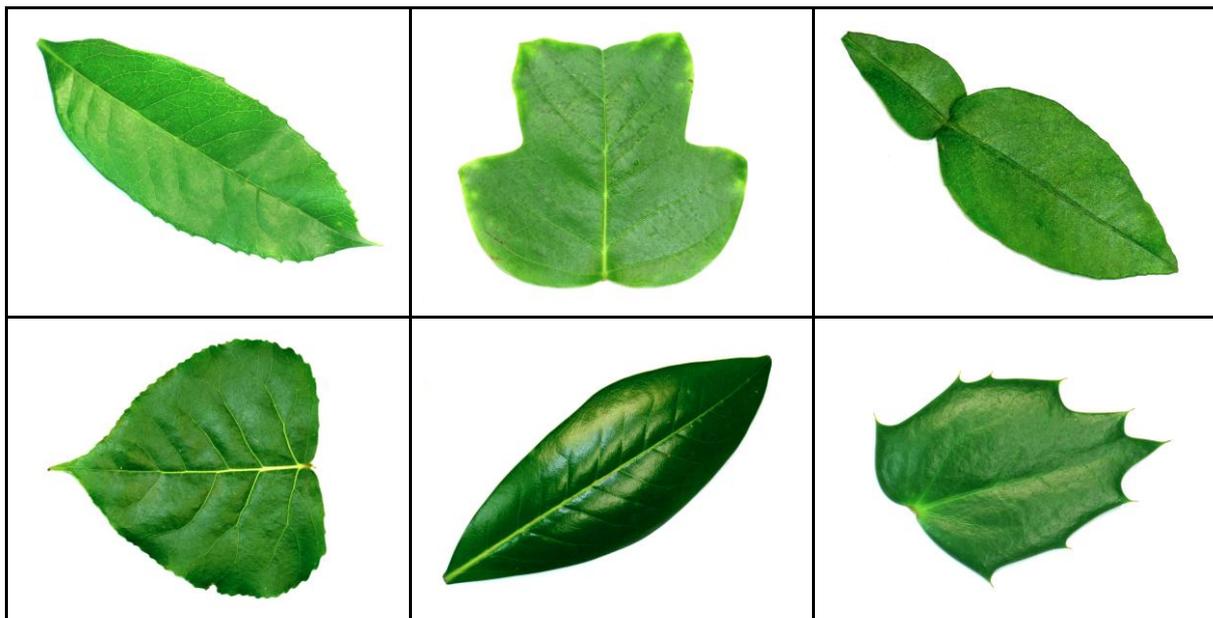
textura fractal y venación, debido a que, como se definió anteriormente, estas características son las que más aportan en una clasificación basada en las hojas.

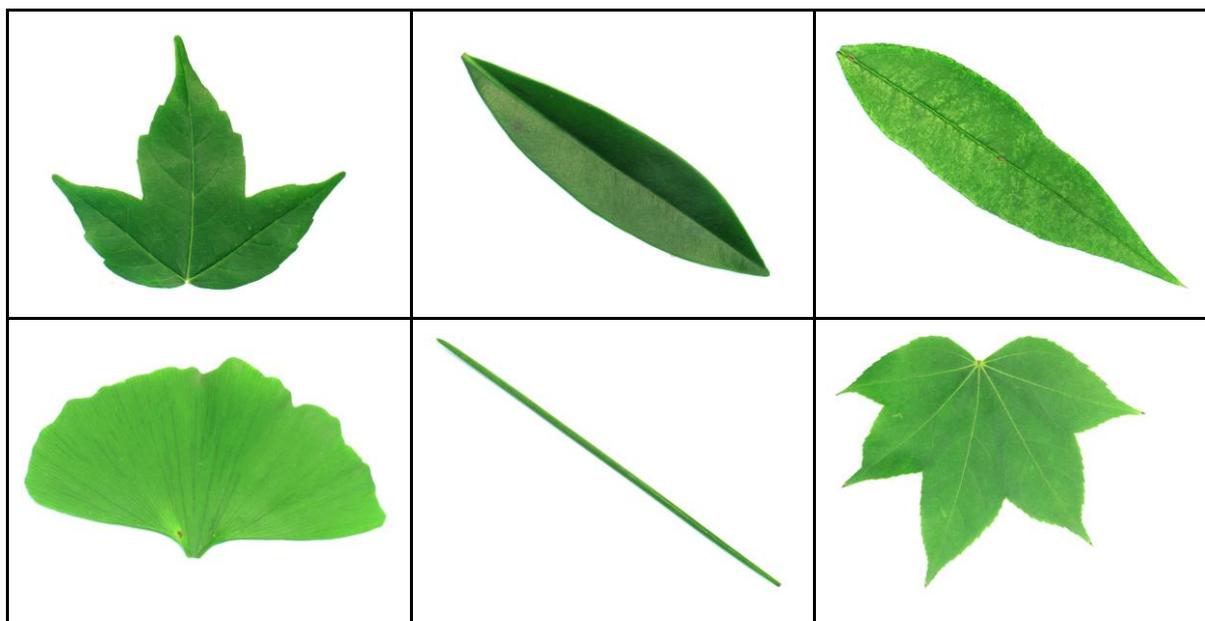
## 10. BASES DE DATOS BASADAS EN HOJAS

El desarrollo tecnológico ha permitido que muchos herbarios colaboren en la digitalización de especies vegetales, un ejemplo de estos es la base de datos descrita en [26], conocida como Global Plants. Esta base de datos es una colección de plantas creada por la colaboración de JSTOR y aproximadamente 300 herbarios. Esta colección contiene 2.48 millones de fotografías de amplia variedad de especies. En cuanto a bases de datos usadas para el estudio de la identificación de especies vegetales por medio de extracción de características a las hojas, se pueden destacar:

### 10.1. Flavia

Es una base de datos que contiene 1907 imágenes de hojas divididas en 33 especies de plantas. Cada especie tiene entre 50 a 77 imágenes con una resolución de 1200 x 1600. Las fotografías en esta base de datos son basadas en hojas simples sobre un fondo blanco. La Tabla 3 es ejemplo de las fotografías que conforman la base de datos Flavia.





**Tabla 3.** Ejemplo de las muestras en la base de datos Flavia.

Las fotografías de Flavia fueron realizadas a hojas maduras sin sombras y una iluminación que permite resaltar en amplio detalle las características las hojas. Para más información de esta base de datos ver el siguiente link: <http://flavia.sourceforge.net/>.

## 10.2. Leaf Data Set

Es una base de datos que contiene 340 imágenes de hojas divididas en 40 especies de plantas. Cada especie tiene entre 5 a 16 imágenes con una resolución de 720 x 960. Las fotografías en esta base de datos son basadas en hojas simples y compuestas sobre un fondo de color rosa. La Tabla 4 es ejemplo de las fotografías que conforman esta base de datos.



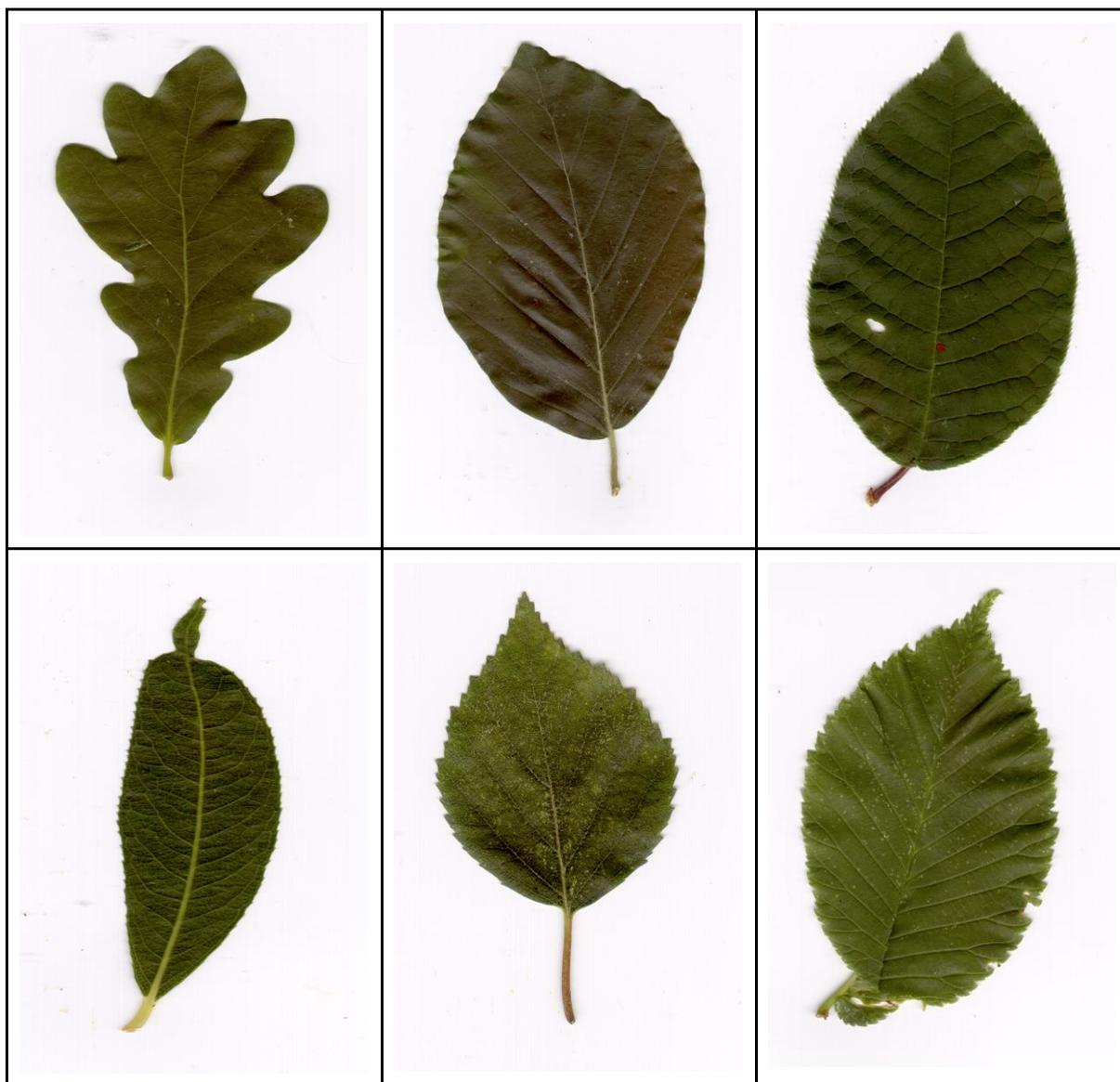


**Tabla 4.** Ejemplo de las muestras en la base de datos Leaf Data Set.

Las fotografías de *Leaf Data Set* fueron realizadas a hojas en un entorno controlado con un ligero nivel de sombreado, lo que permite inferir que faltó un poco más de iluminación. Esta base de datos puede encontrarse en el siguiente link: <https://archive.ics.uci.edu/ml/datasets/leaf>.

### 10.3. Swedish

Es una base de datos que contiene 1125 imágenes de hojas divididas en 15 especies de plantas. Cada especie tiene 75 imágenes con diferentes resoluciones, no superiores a 1200 x 1600. Las fotografías en esta base de datos son basadas en hojas simples y compuestas sobre un fondo blanco. La Tabla 5 es ejemplo de las fotografías que conforman esta base de datos.



**Tabla 5.** Ejemplo de las muestras en la base de datos Swedish.

Según las características de las imágenes se puede decir que las fotografías de Swedish fueron realizadas en un entorno controlado con iluminación que permitió la eliminación de las sobras. Esta base de datos puede encontrarse en el siguiente link: <http://www.cvl.isy.liu.se/en/research/datasets/swedish-leaf/>.

#### **10.4. AmaranthaceaeUQ**

Para extraer la información necesaria para la identificación taxonómica de especies vegetales de la familia Amaranthaceae, fue necesario crea una base de datos de la cual se obtuviera las características con las que se pueda realizar la clasificación. Esta base de datos fue llamada AmaranthaceaeUQ.

AmaranthaceaeUQ es una base de datos construida con base a fotografías de las hojas de 25 especies vegetales (todas con hojas simples) de la familia

Amaranthaceae existentes en el herbario HU de la Universidad del Quindío. Las especies que conforman de esta base de datos son: *Achyranthes aspera* var. *aspera*, *Achyranthes aspera* var. *pubescens*, *Blutaparon vermiculare*, *Celosia argentea* var. *crinata*, *Celosia argentea* var. *argentea*, *Celosia virgata*, *Chamissoa altissima*, *Cyathula achyranthoides*, *Cyathula prostrata*, *Gomphrena serrata*, *Gomphrena globosa*, *Iresine angustifolia*, *Iresine diffusa*, *Iresine herbstii*, *Amaranthus blitum*, *Amaranthus caudatus*, *Amaranthus dubius*, *Amaranthus hybridus*, *Amaranthus spinosus*, *Amaranthus viridis*, *Alternanthera albotomentosa*, *Alternanthera brasiliana*, *Alternanthera lanceolata*, *Alternanthera pungens* y *Alternanthera sessilis*. Todas las anteriores hacen parte de los siguiente nueve géneros: *Achyranthes* (Achy), *Blutaparon* (Blut), *Celosia* (Celos), *Chamissoa* (Cham), *Cyathula* (Cyat), *Gomphrena* (Gomp), *Iresine* (Ires), *Amaranthus* (Amar) y *Alternanthera* (Alte). Para más detalle ver Anexo 1.

Una de las problemáticas encontradas al construir la base de datos *AmaranthaceaeUQ*, es que las hojas fotográficas no fueron seleccionadas de forma controlada, por tal, no su pudo cumplir con el paso 1, el proceso propuesto en la sección 9.2, un ejemplo de eso es que un misma especie hay fotos de hojas maduras y sanas, y al mismo tiempo fotos de hojas inmaduras y con daños. Esta situación no es la deseada, dado que como se dice en [25], la selección de las hojas debe ser basada en un proceso controlado realizado por biólogos especializados, como por ejemplo, obtener la tercera hoja sana y madura de ramas viejas que hayan estado completamente expuestas a la luz solar. Un ejemplo de las diferencias que pueden existir entre hojas de las misma especie se detalla en las Figuras 30 y 31, donde se ven dos hojas de la especie vegetal *Amaranthus blitum*, que a pesar de ser de la misma especie, se diferencian ampliamente por su tamaño.



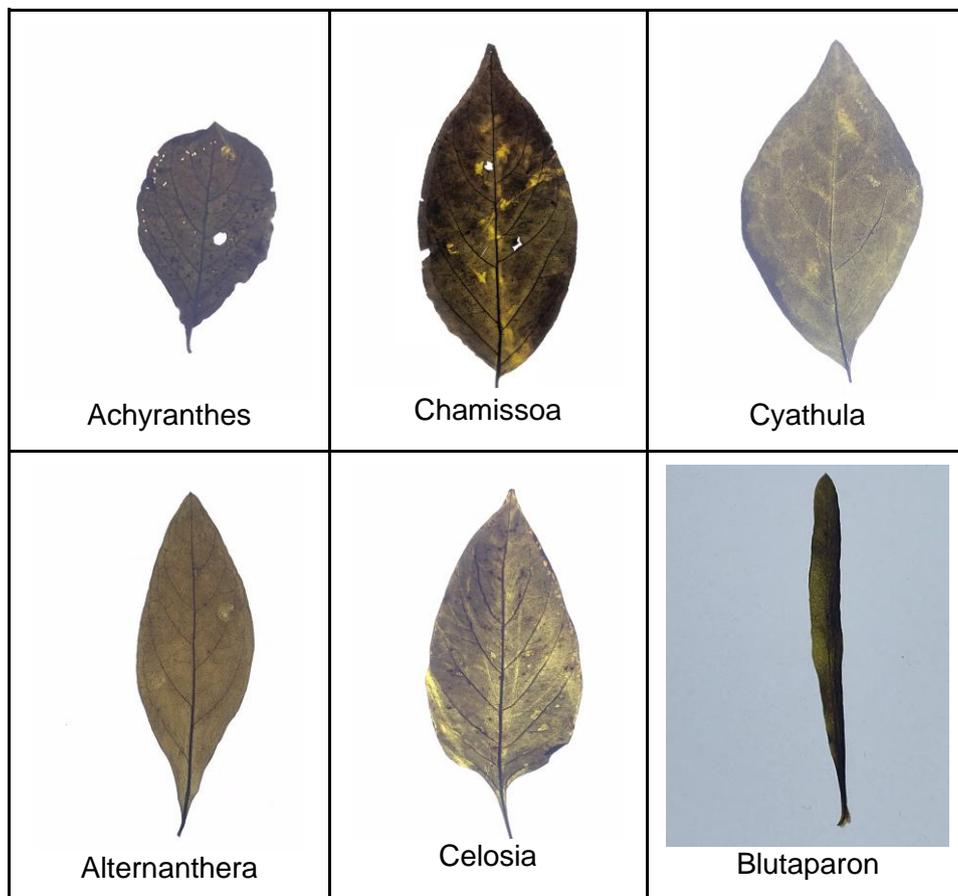
**Figura 30.** Fotografía de hoja madura de *Amaranthus blitum*.

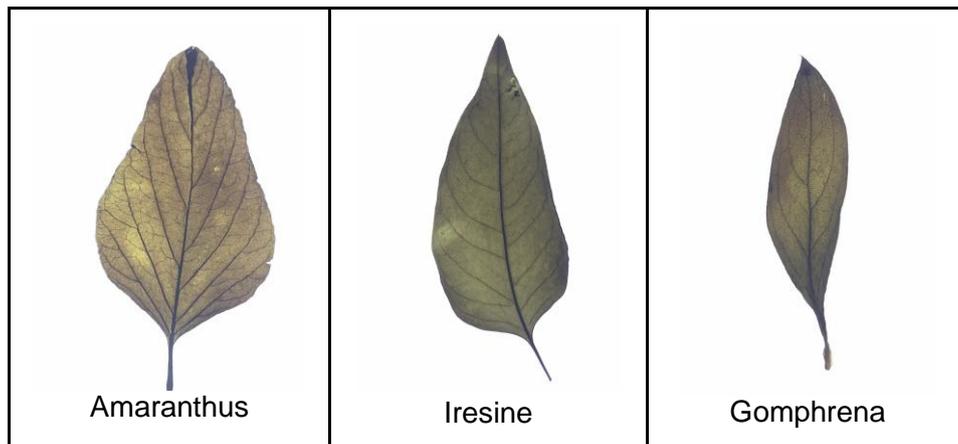


**Figura 31.** Fotografía de hoja inmadura de *Amaranthus blitum*.

Pese a los problemas comentados con anterioridad, se buscó que la captura de las fotos se realizará de forma controlada. Se garantizó que las fotografías de las hojas fueran tomadas en un fondo blanco e iluminado que eliminará la existencia de posibles sombras, además que todas las fotografías fueron realizadas a una altura de 18 cm. Este control se realizó gracias al diseño y construcción del dispositivo llamado *Eliminador de Ruido de la Arquitectura Foliar* versión 001 (ERAF001), el cual se destaca en la sección 9.2. El fotografiar las hojas desde la misma altura permitió tener una referencias de profundidad unificada, no obstante, también causó que en algunas de las hojas inmaduras no se notara con claridad las características como venación o dentado, por tal, para obtener dichas características en la fase de extracción, se decidió tomar una segunda fotografía con zoom, y allí destacar información complementaria para la posterior identificación taxonómica.

Como resultado se obtuvo una base de datos desbalanceada basada en hojas, que cuenta con nueve géneros, 25 tipos de especies vegetales, 206 especies vegetales y por cada una de estas seis fotografías, dos (con zoom y sin zoom) por cada uno de los tres dispositivos nombrados con anterioridad. En total 231 fotografías. En la siguiente tabla (Tabla 6) se destacan algunas de las fotografías de hojas de especies vegetales que hacen parte de *AmaranthaceaeUQ*.



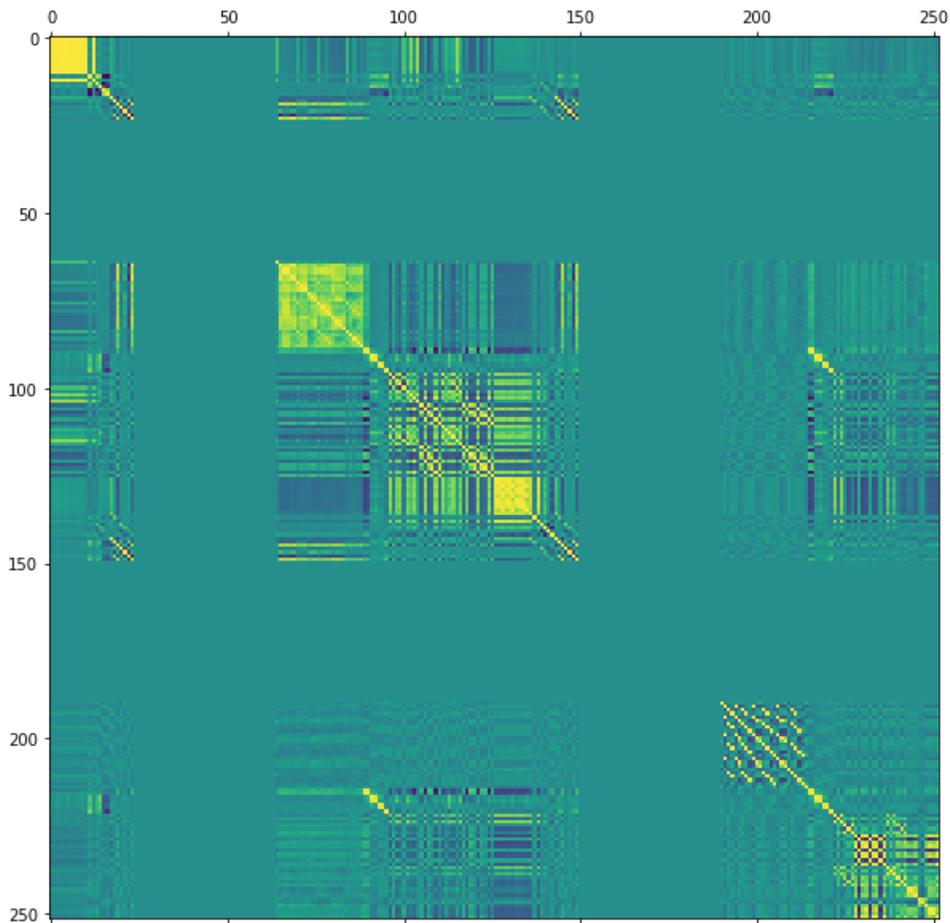


**Tabla 6.** Fotos por género de la base de datos AmaranthaceaeUQ.

Los porcentajes de muestra por género son los siguientes: Achyranthes 6.9%, Blutaparon 4.3%, Celosia 9.5%, Chamissoa 3.5%, Cyathula 9.1%, Gomphrena 8.2%, Iresine 12.1%, Amaranthus 24.7% y Alternanthera 21.6%. Estos porcentajes enmarcan el claro desbalance que tiene AmaranthaceaeUQ.

## 11. TÉCNICAS DE EXTRACCIÓN DE CARACTERÍSTICAS SELECCIONADAS

Las técnicas usadas en la experimentación para la extracción de características fueron Momentos de la imagen, Descriptores Elípticos de Fourier, Momentos polinómicos de Zernike, Momentos invariantes de Hu y Análisis de Textura Fractal. Estas técnicas fueron seleccionadas dado que fueron identificadas en casos de éxito y generan las medidas morfométricas planteadas desde los objetivos. Para generar una intuición de cuáles eran las técnicas que menos aportaban a la clasificación se generó la siguiente matriz de covarianza.



**Figura 32.** Matriz de covarianza aplicando las técnicas de extracción: Momentos de la imagen, Descriptores Elípticos de Fourier, Momentos polinómicos de Zernike, Momentos invariantes de Hu y Análisis de Textura Fractal.

Para el análisis de la figura anterior, se debe tomar en cuenta que las primeras 122 características son extraídas de las imágenes de contorno y las siguientes 122 son de la imagen con zoom. Los colores en la Figura 32 reflejan el grado de variación conjunta entre las características asociadas, entre más vivo el color mayor variación (amarillo alto, verde poco, morado sin variación conjunta).

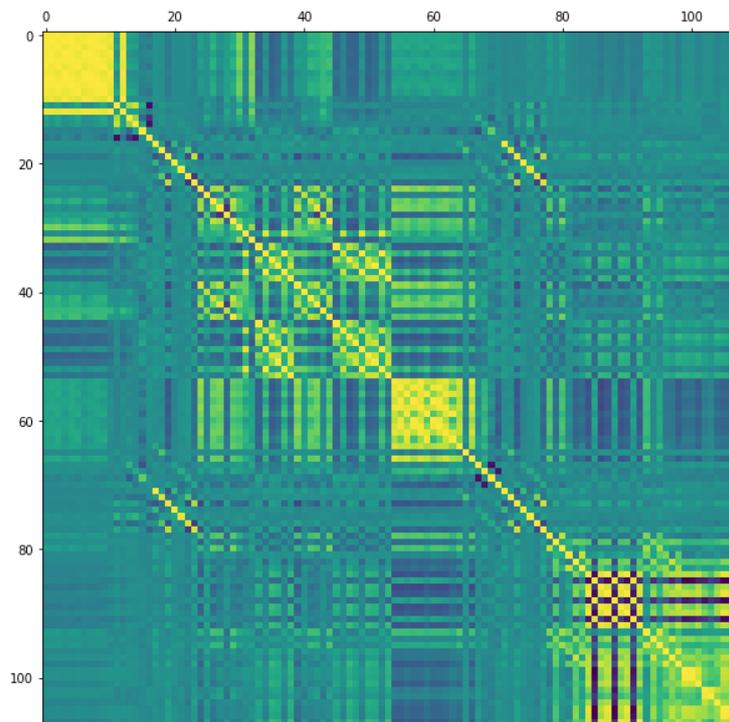
Las primeras 24 características asociadas a los Momentos de la Imagen claramente tienen un alto nivel de variación conjunta para las dos imágenes. Las siguientes 40 características asociadas a los Descriptores Elípticos de Fourier se visualizan en verde, por lo que no deberían de aportar mucha a la clasificación de las plantas. Las siguientes 25 características asociadas a los momentos polinómicos de Zernike muestra una destacada variación conjunta para la primera imagen, pero poco para la segunda. Las siguientes siete características asociadas a los momentos invariantes de HU no parecen aportar mucho a la clasificación. Para finalizar, Análisis de textura fractal en ambas imágenes muestran un alto grado de variación conjunta. Para validar la intuición, se realizaron pruebas de validación cruzada

usando SVM y la base de datos Flavia, con lo que se obtuvieron los resultados mostrados en la Tabla siguiente.

Momentos de la Imagen	Análisis de Textura Fractal	Momentos Polinomicos de Zernike	Momentos Invariantes de Hu	Descriptores Elípticos de Fourier	Precisión
Sí	Sí	Sí	Sí	Sí	0.91
Sí	Sí	Sí	Sí	No	0.91
Sí	Sí	Sí	No	No	0.93
Sí	Sí	No	No	No	0.93

**Tabla 7.** Resultados obtenidos con variada técnicas de extracción de características morfométricas.

La Tabla 7 muestra que la intuición que generó la Figura 32 era correcta y las técnicas de extracción de características que aportan más a la clasificación son: Momento de la Imagen y Análisis de Textura Fractal. La Figura 33 muestra la matriz de covarianza solo con las dos técnicas seleccionadas.

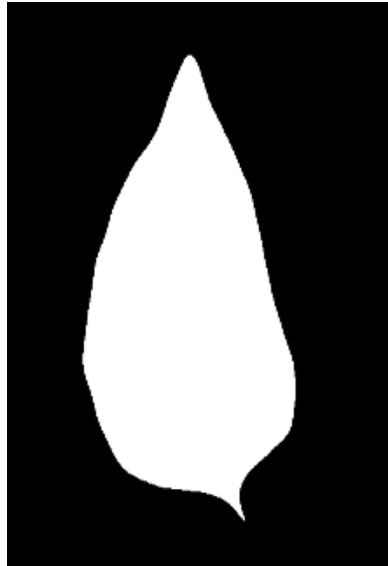


**Figura 33.** Matriz de covarianza aplicando las técnicas de extracción: Momentos de la imagen y Análisis de Textura Fractal.

A continuación se describen las técnicas seleccionadas para la extracción de características morfométricas: Momentos de las imágenes y Análisis de textura fractal.

### 11.1. Momentos de la imagen

Los momentos de la imagen descritos en [30], fueron usados para obtener información del contorno de las hojas. Esta técnica se usó en imágenes de hojas segmentadas (blanco y negro), un ejemplo de una imagen después de ser segmentada se puede ver en la Figura 34.



**Figura 34.** Imagen de la base de datos AmaranthaceaUQ segmentada.

Los momentos de las imágenes se componen de los momentos crudos, centrales y centrales normalizados. Los 10 momentos crudos  $m_{ji}$  son calculados por medio de la Ecuación 1, con esta ecuación se genera una comparación espacial de los puntos generados con base a las dos dimensiones de la imagen  $I$  en el orden  $(j + i)$ .

$$m_{ji} = \sum_x \sum_y (I(x, y) * x^j * y^i) \quad (1)$$

Donde  $j, i = 0, 1, 2, 3$  y  $(x, y)$  la posición de cada pixel. Por otro lado, los siete momentos centrales  $mu_{ji}$  están dados por la Ecuación 2, en donde se realizan comparaciones espaciales a las que se les resta el centro de masa de cada dimensión, esto causa que las relaciones espaciales sean relativamente menores en comparación con los generados con la ecuación 1.

$$mu_{ji} = \sum_x \sum_y (I(x, y) * (x - \bar{x})^j * (y - \bar{y})^i) \quad (2)$$

Los centros de masa  $\bar{x}$  y  $\bar{y}$  están dados por las ecuaciones 3 y 4 respectivamente.

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad (3)$$

$$\bar{y} = \frac{m_{01}}{m_{00}} \quad (4)$$

En cuanto a los siete momentos centrales normalizados  $nu_{ji}$  están representados por la Ecuación 5. Estos momentos son generados usando la Ecuación 2 y son la base para generar los momentos invariantes de Hu.

$$nu_{ji} = \frac{mu_{ji}}{m_{00}^{(i+j)/2+1}} \quad (5)$$

## 11.2. Análisis de textura fractal

Esta técnica fue usada para extraer las características asociadas a la venación y textura. El algoritmo usado fue *Segmentation-based Fractal Texture Analysis* (SFTA), el cual es propuesto en [29] por Costa *et al.* Este algoritmo usa la dimensión de Hausdorff que se presenta con la Ecuación 6.

$$D_0 = \lim_{\varepsilon \rightarrow 0} \frac{\log N(\varepsilon)}{\log \varepsilon^{-1}} \quad (6)$$

Teniendo como entrada la imagen  $I$  para la Ecuación 6, una aproximación  $D$  para la dimensión  $D_0$  puede ser obtenida por medio del algoritmo de conteo de cajas. Donde  $N(\varepsilon)$  es el conteo de cajas en una dimensión euclidiana y  $\varepsilon$  es el tamaño de las cajas.

SFTA se divide en dos partes, la primera es el uso de la técnica *Two-Threshold Binary Decomposition* (TTBD) y la segunda la aplicación del algoritmo de extracción de características. TTBD es una técnica propuesta en [29], basada en dos pasos. El primer paso de TTBD es calcular un conjunto  $T$  de umbrales que se genera usando el algoritmo de descomposición de la pila binaria (en inglés *Binary Stack*).  $T$  Se obtiene seleccionando valores de nivel de grises igualmente espaciados. Para el algoritmo de descomposición de la pila binaria se usó la imagen  $I$  en escala de grises, que posteriormente está limitada por la Ecuación 6 y por un umbral  $t \in \{1, \dots, n_l\}$  generado por la imagen binaria correspondiente, donde  $I(x, y)$  es el valor en escala de grises o la intensidad del píxel en la posición  $(x, y)$ .

$$I_b(x, y) = \begin{cases} 1, & \text{if } I(x, y) \geq t \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

El siguiente paso de TTBD recibe una imagen  $I$  y devuelve imágenes binarias basada en la selección de pares de umbrales de  $T$  y la aplicación de una segmentación de dos umbrales como se muestra en la siguiente ecuación.

$$I_b(x, y) = \begin{cases} 1, & \text{if } t_l < I(x, y) \leq t_u \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Donde  $t_l$  y  $t_u$  denotan respectivamente el umbral bajo y alto. Para representar la dimensión fractal, se generó una imagen de borde denotada por  $\Delta(x, y)$ , que se calcula con base a cada una de las imágenes binarias generadas con el TTBD y la siguiente ecuación.

$$\Delta(x, y) = \begin{cases} 1 & \text{if } \exists(x', y') \in N_8[(x, y)] : \\ & I_b(x', y') = 0 \wedge \\ & I_b(x, y) = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Donde  $N_8[(x, y)]$  son los 8 píxeles que están conectados a  $(x, y)$ .  $\Delta(x, y)$  es 1 si el valor de  $(x, y)$  es 1 y 0 si no tiene algún pixel vecino igual a 1. Finalmente la dimensión fractal es calculada usando  $\Delta(x, y)$  en la Ecuación 6. SFTA no solo da como resultado dimensiones fractales, también devuelve el nivel de gris promedio y el tamaño de cada imagen binaria. Para este proyecto SFTA fue usado con 5 umbrales, los cuales generaron 30 características por imagen analizada.

## 12. SELECCIÓN DE CARACTERÍSTICAS

Para la reducción de dimensionalidad y selección de las características que más aportan a la clasificación se usó Eliminación Recursiva de Características (RFE). RFE es un algoritmo de selección de características propuesto en [30], que usa los pesos generados por un clasificador SVM lineal como base para la selección de las mejores características. El uso del kernel lineal en SVM hace que RFE tenga buenos resultados en características separables linealmente. El algoritmo RFE se describe a continuación.

### Entrada

- Vector de características  $X_0 = [x_1, x_2, x_3, \dots, x_l]^T$  y las clases etiquetadas  $y = [y_1, y_2, y_3, \dots, y_l]$ . Siendo  $l$  el número total de características.
- $m$  = Número de características a ser seleccionadas.

### Salida

- $X$  con  $m$  características.

### Inicio RFE

1. Inicializar  $X$  con todas las características.  $X = X_0$ .
2. Entrenar la SVM.  $\alpha = SVM - train(X, y)$ .

3. Calcular los pesos de los vectores con base a  $\alpha$ .  $w = \sum_k \alpha_k y_k x_k$ .
4. Calcular los criterios de clasificación.  $c_i = (w_i)^2$ .
5. Buscar y eliminar las características con el menor criterio de clasificación de  $X$ .
6. Repetir el proceso de 2 a 5 hasta que el número de características igual a  $m$ .

### 12.1. Características que más aportan a la clasificación

Tal como se ha descrito con anterioridad, cada ejemplo de la base de datos *AmaranthaceaeUQ* consta de dos imágenes. Por imagen se extraen 54 características dando como total 108 por ejemplo. Después de una selección por medio de ensayo y error se definió un RFE con 46 características. Las 46 características seleccionadas se pueden visualizar en la Tabla 8.

<b>IMAGEN SIN ZOOM (PARA RESALTAR EL CONTORNO)</b>	
<b>Momentos de la imagen</b>	m00, m10, m01, m20, m02, m30, m03, mu11, mu02, nu20 y nu30.
<b>De SFTA con 5 umbrales</b>	
<b>Umbral 1</b>	promedio de escala de grises
<b>Umbral 2</b>	promedio de escala de grises conteo de cajas
<b>Umbral 3</b>	dimensión fractal promedio de escala de grises conteo de cajas
<b>Umbral 4</b>	dimensión fractal conteo de cajas
<b>Umbral 5</b>	promedio de escala de grises conteo de cajas
<b>Entre umbral 1 y 2</b>	dimensión fractal promedio de escala de grises conteo de cajas
<b>Entre umbral 2 y 3</b>	dimensión fractal promedio escala de grises conteo de cajas
<b>Entre el umbral 3 y 4</b>	dimensión fractal

	promedio de escala de grises conteo de cajas
<b>Entre el umbral 4 y 5</b>	dimensión fractal promedio de escala de grises conteo de cajas
<b>Entre el umbral 5 y el número 1</b>	promedio de escala de grises
<b>IMAGEN CON ZOOM (PARA RESALTAR LA VENACIÓN Y TEXTURA)</b>	
<b>Momentos de la imagen</b>	m01, m02, m21, mu20, mu11, mu12, nu20, nu02, nu30 y nu03.
<b>Umbral 3</b>	Dimensión fractal.
<b>Umbral 4</b>	Dimensión fractal.

**Tabla 8.** Características extraídas por muestra de la base de datos AmaranthaceaeUQ.

Se debe tener en cuenta que los valores de los umbrales se encuentran entre 0 y 1, además de que para cada imagen se calculan los diferentes umbrales, por tal los valores son diferentes por imagen.

### 13. SELECCIÓN DE ENFOQUE DE APRENDIZAJE DE MÁQUINA

Tal como se ha especificado en la sección de selección de enfoque de aprendizaje de máquina, los tres clasificadores seleccionados como opción para entrenar las características extraídas de las fotos de las hojas son: SVM, K-NN y ANN. Debido a que se usará el lenguaje de programación Python para la creación del algoritmo que ayuda a la identificación taxonómica de especies vegetales de Amaranthaceae (para ver la justificación del por qué usar este lenguaje de programación ver documento de arquitectura - Anexo 4), se usarán los enfoques de aprendizaje de máquina que provee la librería Scikit-Learn [31].

Los tres clasificadores seleccionados tienen hiperparámetros que permite ajustar el modelo a los datos. Con el objetivo de seleccionar el clasificador con que se obtenga la mayor precisión y fiabilidad, se realizaron entrenamientos con variados hiperparámetros y así determinar cuál es la configuración que mejor se ajustan a los datos por cada clasificador.

### 13.1. Perceptrón Multicapa - MLP

El perceptrón multicapa es un algoritmo de aprendizaje supervisado basado en la función  $f(.) = R^m \rightarrow R^o$  para entrenar un conjunto de datos, donde  $m$  es el número de dimensiones de la capa de entrada y  $o$  las dimensiones de la capa de salida. Para el uso de este algoritmo se debe proporcionar un conjunto de características  $X = \{x_1, x_2, x_3, \dots, x_m\}$  y una clasificación  $y$ . Uno de los rasgos que más se destaca en MPL es la existencia de capas internas, las cuales son llamadas capas ocultas.

El algoritmo MPL de Scikit-Learn recibe en la capa inicial (capa de entrada) las características de entrenamiento, en las capas ocultas se transforma la información y en la capa final se reciben los valores generados en la capa oculta y los transforma en los datos de salida. La complejidad de una red neuronal con retropropagación está dada por  $O(n * m * h^k * o * i)$ , donde  $n$  es el número de ejemplos de entrenamiento,  $m$  el número de características,  $k$  la cantidad de capas ocultas cada una con  $h$  neuronas,  $o$  las neuronas de la capa de salida e  $i$  el número de iteraciones.

Aunque MLP tiene un funcionamiento general definido, los hiperparámetros de éste permiten modificar su comportamiento para ajustarse a los datos. Los hiperparámetros que fueron objeto de estudio son: número de capas ocultas, función de activación, solver y alfa.

#### 13.1.1. Capas ocultas

Cada neurona en la capa oculta transforma los valores de la capa anterior en una suma lineal ponderada  $w_1x_1 + w_2x_2 + \dots + w_mx_m$  aplicada en una función de activación. Por tal, entre más capas ocultas hay más operaciones se realizan, incrementado el tiempo de ejecución del algoritmo. Para determinar la cantidad de capas ocultas a utilizar en MPL se usa la propiedad *hidden\_layer\_sizes*.

#### 13.1.2. Función de activación

Las funciones de activación disponibles en MLP de Scikit-Learn son: *identity*, *logistic*, *tanh* y *relu*. Estas funciones determinan la forma en que se desean separar las muestras en un espacio multidimensional.

La función *identity* está dada por la función identidad,  $f(x) = x$ , claramente muestra que el valor que entre es el mismo que sale, por lo cual es totalmente dependiente la suma de los valores transformados desde las neuronas de la capa anterior. La función *logistic* está dada por  $g(z) = 1/(1 + e^{-z})$  en el caso de realizar un entrenamiento biclase, o por la función softmax,  $softmax(z)_i = exp(z_i) / (\sum_{l=1}^k exp(z_l))$  en el caso que se desean entrenar múltiples clases. Para el caso biclase el umbral utilizado es de 0.5. En el caso de la función *tanh* hace referencia a  $g(z) = (e^z - e^{-z}) / (e^z + e^{-z})$ . Por otra parte, la función de activación *relu* se

representa  $f(x) = \max(0, x)$ , representado un crecimiento lineal al igual de la función identidad. Para definir la función de activación que se desea usar en MLP se usa la propiedad *activation*.

### 13.1.3. Solver

El solver es la forma como se quieren ajustar los pesos que multiplican las  $x_i$  de entrada a cada neurona. Las posibles opciones de solver en MPL son: *adam*, *sgd* y *lbfgs*. El solver por defecto es *adam* que según la documentación de Scikit-Learn se adapta bien a conjuntos de datos relativamente grandes (con miles de muestras de entrenamiento) en términos de tiempo de entrenamiento y puntaje de validación. De la misma forma recomiendan *lbfgs* para conjuntos de datos pequeños, dado que puede converger más rápido y funcionar mejor.

**Gradiente descendente estocástico (sgd):** es un método iterativo para optimizar una función objetivo diferenciable. Se denomina estocástico porque las muestras se seleccionan aleatoriamente. El gradiente descendente estocástico está dado por la siguiente ecuación:

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} := \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \eta \begin{bmatrix} \frac{\partial}{\partial w_1} (w_1 + w_2 x_i - y_i)^2 \\ \frac{\partial}{\partial w_2} (w_1 + w_2 x_i - y_i)^2 \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \eta \begin{bmatrix} 2(w_1 + w_2 x_i - y_i) \\ 2x_i(w_1 + w_2 x_i - y_i) \end{bmatrix}.$$

**Limited-memory BFGS (lbfgs):** es un algoritmo de optimización que utiliza una cantidad limitada de memoria de la computadora. Su objetivo es minimizar  $f(x)$ , donde  $f$  es una función diferenciable. L-BFGS usa una estimación de la matriz inversa de Hessiana para dirigir su búsqueda a través del espacio variable, pero donde BFGS almacena una aproximación densa  $n * n$  al inverso de la Hessiana (siendo  $n$  el número de variables en el problema), L-BFGS almacena solo unos pocos vectores que representan la aproximación implícitamente. *lbfgs* está dada por la siguiente ecuación.

$$H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top.$$

**Adam:** es un algoritmo para la optimización basada en el gradiente de primer orden de las funciones objetivas estocásticas, basado en estimaciones adaptativas de momentos de orden inferior. El método es computacionalmente eficiente, tiene pocos requisitos de memoria, es invariante para el reajuste diagonal de los gradientes, y es adecuado para problemas que son grandes en términos de datos y / o parámetros. Adam está dado por la función:

$$v_t = (1 - \beta_2) \sum_{i=1}^t \beta_2^{t-i} \cdot g_i^2$$

Para definir el solver a usar en MLP se usa la propiedad *solver*.

#### 13.1.4. Alpha

*Alpha* es el parámetro para ayudar a evitar el sobreentrenamiento en MLP, para esto *alpha* penaliza los pesos que tienen grandes magnitudes. El aumento de *alpha* puede corregir una gran varianza (un signo de sobreajuste) mediante el fomento de pesos más pequeños, lo que resulta en un diagrama de límites de decisión que aparece con curvaturas menores. Del mismo modo, la disminución de *alpha* puede corregir un sesgo alto (un signo de falta de ajuste) al fomentar pesos más grandes, lo que puede resultar en un límite de decisión más complicado.

### 13.2. MAQUINA DE VECTOR DE SOPORTE - SVM

SVM es un conjunto de algoritmos de aprendizaje de máquina supervisados utilizado para clasificación, regresión y detección de valores atípicos. Una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita). Scikit-Learn entre sus librerías ofrece tres métodos para usar SVM: SVC (C-Support Vector Classification), NuSVC (Nu-Support Vector Classification) y LinearSVC (Linear-Support Vector Classification). Para usar cualquiera de estos métodos se necesita de dos matrices como datos de entrada, la matriz  $X$  de características y el vector de etiquetas. Por otra parte la complejidad de SVM está dada por  $O(n_{características} * n_{ejemplos}^3)$ , la cual es claramente más baja que la de MPL. El método seleccionado para probar los hiperparámetros en este proyecto fue SVC, dado que el LinearSVC solo permite realizar entrenamiento basado en un kernel lineal y NuSVC es igual que SVC solo que tiene un hiperparámetro adicional para controlar la cantidad de los vectores de soporte (hiperparámetro que no se desea usar). Los hiperparámetros tomados en cuenta para determinar la mejor configuración en el entrenamiento son: *kernel*, *C*, *gamma* y *degree* (usado solo para kernel polinomial).

#### 13.2.1. Kernel

El kernel determina la función con base a la cual se realiza la separación de las clases que se desean clasificar en un espacio multidimensional. El algoritmo SVC permite usar cualquiera de las siguientes funciones: *linear*, *poly*, *rbf* y *sigmoid*.

El kernel *linear* está dado por la función  $k(x_1, x_2) = x_1 * x_2$ , donde  $x_1$  representa las características y  $x_2$  la etiqueta. El kernel *poly* usa la función  $k(x_1, x_2) = (\gamma * x_1 * x_2 + r)^d * d$ , donde  $d$  es el grado,  $r$  es especificada por medio del hiperparámetro *coef0* (por defecto es cero) y  $\gamma > 0$ . La función para el kernel *rbf*  $k(x_1, x_2) = \exp(-\gamma || x_1 * x_2 ||^2)$ . Para finalizar, la función la *sigmoid*, está dada por  $k(x_1, x_2) = \tanh(\gamma * x_1 * x_2 + r)$ .

### 13.2.2. C

Con  $C$  se determina que tanto se desea clasificar cada instancia correctamente. Una  $C$  baja permite que la superficie de decisión sea lisa, mientras que una  $C$  alta tiene como objetivo identificar correctamente la mayor cantidad de ejemplos de entrenamiento, dando libertad al modelo para seleccionar más muestras como vectores de soporte. Un  $C$  grande hace referencia a más vectores de soporte, y a más vectores de soporte más memoria.

### 13.2.3. Gamma

Intuitivamente, el parámetro gamma define hasta dónde llega la influencia de un único ejemplo de entrenamiento, valores bajos significan lejos y valores altos significan cercanos. El comportamiento del modelo es sensible al parámetro gamma. Si gamma es demasiado grande, el radio del área de influencia de los vectores de soporte solo incluye el vector de soporte en sí mismo y ninguna cantidad de regularización con  $C$  podrá evitar el sobreajuste. Aunque *gamma* y  $C$  se suele usar con kernel *rbf*, estos hiperparámetros también pueden generar un importante influencia en los otros *kernel*.

### 13.2.4. Grado

Este hiperparámetro solo es usado con el kernel polinomial, es usado para buscar un ajuste a los datos con base al grado. Para cambiar el grado al kernel se usa la propiedad *degree*.

## 13.3. K VECINO MÁS CERCANO - K-NN

Es un método de clasificación y regresión que permite el aprendizaje supervisado y no supervisado. K-NN estima el valor de la función de densidad de probabilidad o directamente la probabilidad posterior de que un elemento  $x$  pertenezca a una clase  $C_j$  a partir de la información proporcionada por el conjunto de prototipos.

El principio de K-NN en Scikit-Learn es encontrar el número predefinido de muestras de entrenamiento más cercanas en distancia al punto a clasificar, y predecir la etiqueta a partir de estos. El número de muestras puede ser una constante definida por el usuario (aprendizaje del vecino más cercano) o variar en función de la densidad local de puntos (aprendizaje vecino basado en el radio). La distancia puede, en general, ser cualquier medida métrica: la distancia euclidiana estándar es la opción más común.

Los elementos de entrada de este clasificador son los mismo que los dos anteriores, una matriz de características  $X = \{ x_1, x_2, x_3, \dots, x_m \}$  y un vector de etiquetas  $y$ . Los hiperparámetros probados para este clasificador fueron: *n\_neighbors*, *weights* y *algorithm*.

### 13.3.1. Número de vecinos

Es el número de vecinos con base al cual se realizará la estimación de la probabilidad (valor de la función de densidad de probabilidad). El número de vecinos se puede definir en K-NN usando la propiedad *n\_neighbors*.

### 13.3.2. Pesos

Este hiperparámetro se usa para ponderar la contribución de cada vecino, en este algoritmo de Scikit-Learn se ofrecen dos opciones preestablecidas. La primera es *uniform*, esta opción permite que cada punto del vecindario local contribuya de manera uniforme a la clasificación de un punto de consulta. La segunda es *distance*, donde los vecinos más cercanos al punto de consulta tendrán una mayor influencia que los vecinos que están más lejos. La estrategia de peso se puede determinar usando la propiedad *weights*.

### 13.3.3. Algorithm

Los algoritmos que provee Scikit-Learn para el uso en de K-NN son: *auto*, *ball\_tree*, *kd\_tree* y *brute*.

**Fuerza bruta (*brute*):** La implementación de búsqueda del vecino más cercano implica el cálculo de la fuerza bruta de distancias entre todos los pares de puntos en el conjunto de datos: para  $N$  muestras en  $D$  dimensiones, este enfoque tiene una complejidad de  $O(DN^2)$ . La fuerza bruta suele ser efectiva en muestras de datos pequeños, no es recomendable al contar con cientos de miles de datos o más.

**Árbol K-D (*kd\_tree*):** Para abordar las ineficiencias computacionales del enfoque de la fuerza bruta, se ha inventado una variedad de estructuras de datos basadas en árboles. La idea básica de este algoritmo es que si el punto  $A$  está muy lejos del punto  $B$  y el punto  $B$  está muy cerca del punto  $C$ , entonces se sabe que los puntos  $A$  y  $C$  son distantes, sin tener que calcular explícitamente su distancia. De esta forma, el costo computacional de la búsqueda de un vecino más cercano se puede reducir a  $O(D * N * \log N)$ . Donde  $N$  es el número de muestras y  $D$  las dimensiones.

El árbol KD es una estructura de árbol binario que divide recursivamente el espacio de parámetros a lo largo de los ejes de datos, dividiéndolo en regiones ortotrópicas anidadas en las que se archivan los puntos de datos. La construcción de un árbol KD es rápida debido a que el particionamiento se realiza solo a lo largo de los ejes de datos, no es necesario calcular distancias *D-dimensionales*. Una vez construido, el vecino más cercano de un punto de consulta se puede determinar con solo cálculos de distancia con complejidad de  $O(\log N)$ . Aunque el enfoque de árbol KD es muy rápido para búsquedas de vecinos de baja dimensión ( $D < 20$ ), se vuelve ineficiente a medida que  $D$  se vuelve muy grande.

**Árbol de bolas (*ball\_tree*):** Para abordar las ineficiencias de los árboles KD en dimensiones superiores, ha sido desarrollada la estructura de datos del árbol de bolas. Cuando los árboles KD dividen datos a lo largo de ejes cartesianos, los árboles de bolas dividen los datos en una serie de hiperesferas de anidación. Esto hace que la construcción de árboles sea más costosa que la del árbol KD, pero da como resultado una estructura de datos que puede ser muy eficiente en datos altamente estructurados, incluso en dimensiones muy altas.

Un árbol de bolas recursivamente divide los datos en nodos definidos por un centroide  $C$  y un radio  $r$ , de modo que cada punto del nodo se encuentra dentro de la hiperesfera definida por  $r$  y  $C$ . El número de puntos candidatos para una búsqueda de vecinos se reduce mediante el uso de la desigualdad del triángulo:

$$|x + y| \leq |x| + |y|$$

Con esta configuración, un solo cálculo de distancia entre un punto de prueba y el centroide es suficiente para determinar un límite inferior y superior en la distancia a todos los puntos dentro del nodo.

**Auto (*auto*):** Intenta decidir el algoritmo más apropiado en función de los valores pasados al método de ajuste.

Para definir el algoritmo a usar en K-NN se usa la propiedad *algorithm*.

#### 13.4. GRIDSEARCHCV

GridSearchCV del paquete *model\_selection* de Scikit-Learn realiza una búsqueda exhaustiva sobre valores de parámetros especificados para un estimador. Los parámetros del estimador utilizado para aplicar estos métodos se optimizan mediante una búsqueda de grillas con validación cruzada en una cuadrícula de parámetros. Los hiperparámetros usados en la búsqueda fueron: *estimator*, clasificador que se desea usar (en este caso SVM, MLP y K-NN); *param\_grid*, parámetros con los que se desea entrenar el clasificador; *cv*, número de subconjuntos (iteraciones) que se usarán en la validación cruzada (*K-fold cross-validation*), en el entrenamiento se usaron 5; y *scoring*, define lo que se desea evaluar, en este caso precisión y exhaustividad.

#### 13.5. PRUEBA DE CONCEPTO - SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS

Para validar que las características extraídas que se especifican en la sección 12 generan la información suficiente para clasificar especies vegetales, se decidió usar la base de datos Flavia (base de datos descrita en la sección 10) como prueba de

concepto, y se realizó la selección de los mejores hiperparámetros usando GridSearchCV de Scikit-Learn. Para la separación de los datos de entrenamiento con los de prueba se hizo uso de *train\_test\_split* del paquete *model\_selection* de Scikit-Learn. A continuación, en las Tablas 9, 10 y 11 se describen los resultados obtenidos con los clasificadores MLP, SVM y K-NN.

<b>Clasificador:</b>	MLP
<b>Hiperparámetros:</b>	[{'activation': ['identity', 'logistic', 'tanh', 'relu'], 'hidden_layer_sizes': [(25,),(50,),(12,25,12)], 'max_iter': [8000], 'solver': ['lbfgs', 'sgd', 'adam'], 'alpha': np.logspace(-9, 3, 13)}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 1907 instancias se usó el 70% de los datos para el entrenamiento y el 30% restante para pruebas.
<b>Búsqueda enfocada en la precisión:</b>	Hiperparámetros resultantes: {'activation': 'tanh', 'alpha': 1e-07, 'hidden_layer_sizes': (50,), 'max_iter': 6000, 'solver': 'adam'}  Resultados: Precisión 0.94 y exhaustividad 0.94.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'activation': 'tanh', 'alpha': 1e-09, 'hidden_layer_sizes': (50,), 'max_iter': 6000, 'solver': 'adam'}  Resultados: Precisión 0.95 y exhaustividad 0.95.

**Tabla 9.** Información base de la búsqueda de hiperparámetros para MLP con la base de datos Flavia.

<b>Clasificador:</b>	SVM
<b>Hiperparámetros:</b>	[{'kernel': ['rbf', 'sigmoid'], 'gamma': np.logspace(-4, 3, 8), 'C': np.logspace(-2, 7, 10)}, {'kernel': ['linear'], 'C': np.logspace(-2, 7, 10)}, {'kernel': ['poly'], 'C': np.logspace(-2, 7, 10), 'degree': [1, 2, 3], 'gamma': np.logspace(-4, 3, 8)}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 1907 instancias se usó el 70% de los datos para el entrenamiento y el 30% restante para pruebas.
<b>Búsqueda enfocada</b>	Hiperparámetros resultantes:

<b>en el precisión:</b>	{'C': 100.0, 'gamma': 0.01, 'kernel': 'rbf'}  Resultados: Precisión 0.95 y exhaustividad 0.95.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'C': 100.0, 'gamma': 0.01, 'kernel': 'rbf'}  Resultados: Precisión 0.95 y exhaustividad 0.95.

**Tabla 10.** Información base de la búsqueda de hiperparámetros para SVM con la base de datos Flavia.

<b>Clasificador:</b>	K-NN
<b>Hiperparámetros:</b>	[{'n_neighbors':[1,2,3,4,5,6,7,8,9,10],'weights':['uniform','distance'],'algorithm':['auto','ball_tree','kd_tree','brute']}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 1907 instancias se usó el 70% de los datos para el entrenamiento y el 30% restante para pruebas.
<b>Búsqueda enfocada en el precisión:</b>	Hiperparámetros resultantes: {'algorithm': 'auto', 'n_neighbors': 4, 'weights': 'distance'}  Resultados: Precisión 0.92 y exhaustividad 0.91.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'algorithm': 'auto', 'n_neighbors': 1, 'weights': 'uniform'}  Resultados: Precisión 0.91 y exhaustividad 0.91.

**Tabla 11.** Información base de la búsqueda de hiperparámetros para K-NN con la base de datos Flavia.

Tal como lo muestran en las Tablas 9, 10 y 11 los resultados obtenidos en la identificación de las especies vegetales de la base de datos Flavia, fueron satisfactorios para todos los clasificadores, sin embargo, cabe destacar el resultado obtenido con SVM (ver Tabla 10) que generó en precisión y exhaustividad de 0.95. Por lo anterior, se puede considerar que el modelo generado para la identificación de especies vegetales con base a fotografías de las hojas es válido. Para ver más detalle de la prueba de concepto del modelo ver Anexo 2.

### 13.6. SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS PARA LA BASE DE DATOS AMARANTHACEAEUQ

Debido a las pocas instancias que hay en algunas de las especies vegetales en la base de datos AmaranthaceaeUQ (para agosto de 2018), se decidió realizar el entrenamiento a nivel de género. AmaranthaceaeUQ cuenta con nueve géneros, cabe decir que la base de datos a nivel de género sigue estando desbalanceada.

Para la selección del mejor clasificador con sus respectivos hiperparámetros para la base de datos generada en este proyecto, se realizó el mismo proceso descrito en la sección de la prueba de concepto. Las características de la base de datos AmaranthaceaeUQ son descritas en la Sección 10. En las Tablas 12, 13 y 14 se detallan las pruebas realizadas con los clasificadores MLP, SVM y K-NN para la base de datos AmaranthaceaeUQ.

<b>Clasificador:</b>	<b>MLP</b>
<b>Hiperparámetros:</b>	[{'activation': ['identity','logistic','tanh','relu'], 'hidden_layer_sizes':[(25,),(50,),(12,25,12)],'max_iter': [8000], 'solver':['lbfgs', 'sgd', 'adam'],'alpha': np.logspace(- 9, 3, 13)}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 231 instancias se usó el 60% de los datos para el entrenamiento y el 40% restante para pruebas.
<b>Búsqueda enfocada en la precisión:</b>	Hiperparámetros resultantes: {'activation': 'relu', 'alpha': 1e-08, 'hidden_layer_sizes': (50,), 'max_iter': 6000, 'solver': 'adam'}  Resultados: Precisión 0.65 y exhaustividad 0.63.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'activation': 'tanh', 'alpha': 1e-08, 'hidden_layer_sizes': (25,), 'max_iter': 6000, 'solver': 'adam'}  Resultados: Precisión 0.66 y exhaustividad 0.65.

**Tabla 12.** Información base de la búsqueda de hiperparámetros para MLP con la base de datos AmaranthaceaeUQ.

<b>Clasificador:</b>	SVM
<b>Hiperparámetros:</b>	[{'kernel': ['rbf','sigmoid'], 'gamma': np.logspace(-4, 3, 8),'C': np.logspace(-2, 7, 10)}, {'kernel': ['linear'], 'C': np.logspace(-2, 7, 10)},{'kernel': ['poly'], 'C': np.logspace(-2, 7, 10),'degree':[1, 2, 3], 'gamma': np.logspace(-4, 3, 8)}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 231 instancias se usó el 60% de los datos para el entrenamiento y el 40% restante para pruebas.
<b>Búsqueda enfocada en el precisión:</b>	Hiperparámetros resultantes: {'C': 100.0, 'gamma': 0.001, 'kernel': 'sigmoid'}  Resultados: Precisión 0.66 y exhaustividad 0.62.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'C': 100000.0, 'gamma': 0.0001, 'kernel': 'rbf'}  Resultados: Precisión 0.65 y exhaustividad 0.67.

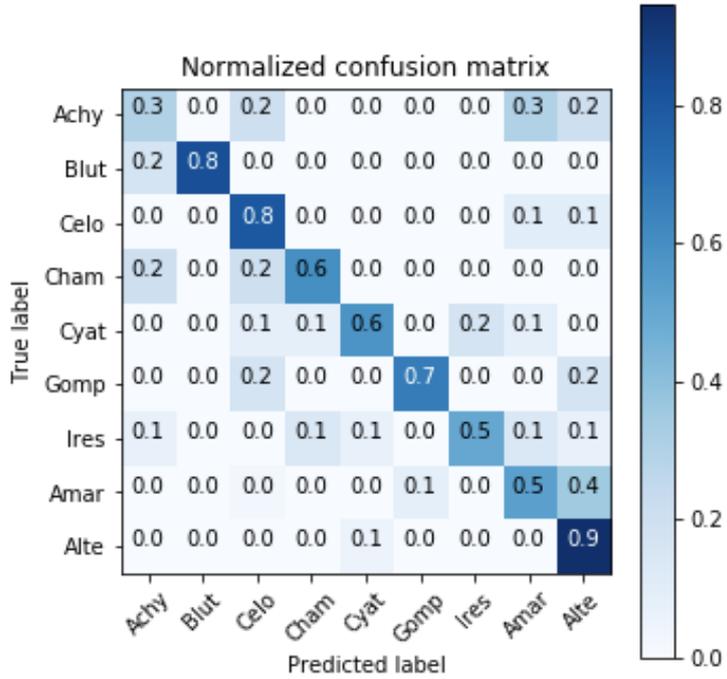
**Tabla 13.** Información base de la búsqueda de hiperparámetros para SVM con la base de datos AmaranthaceaUQ.

<b>Clasificador:</b>	K-NN
<b>Hiperparámetros:</b>	[{'n_neighbors':[1,2,3,4,5,6,7,8,9,10],'weights':['uniform','distance'],'algorithm':['auto', 'ball_tree', 'kd_tree', 'brute']}]
<b>Grupo de entrenamiento y pruebas:</b>	De las 231 instancias se usó el 60% de los datos para el entrenamiento y el 40% restante para pruebas.
<b>Búsqueda enfocada en el precisión:</b>	Hiperparámetros resultantes: {'algorithm': 'auto', 'n_neighbors': 5, 'weights': 'distance'}  Resultados: Precisión 0.51 y exhaustividad 0.57.
<b>Búsqueda enfocada en la exhaustividad:</b>	Hiperparámetros resultantes: {'algorithm': 'auto', 'n_neighbors': 5, 'weights': 'uniform'}  Resultados:

	Precisión 0.45 y exhaustividad 0.43.
--	--------------------------------------

**Tabla 14.** Información base de la búsqueda de hiperparámetros para K-NN con la base de datos AmaranthaceaUQ.

Como se ha podido observar en las Tablas 12, 13 y 14 los resultados obtenidos para la identificación de géneros en la base de datos AmaranthaceaeUQ son en general bajos. Los mejores resultados fueron obtenidos con SVM al igual que sucedió con el entrenamiento realizado para la base de datos Flavia. La configuración {'C': 100000.0, 'gamma': 0.0001, 'kernel': 'rbf'} para SVM alcanza un 0.65 de precisión y un 0.67 de exhaustividad, configuración parecida a la que genera los mejores resultados con Flavia. En la siguiente figura (Figura 35) se muestra la matriz de confusión en donde se pueden ver los géneros que se confunden en la clasificación.



**Figura 35.** Matriz de confusión de la identificación todos contra todos usando SVM con la mejor configuración.

El género Achyranthes en el algoritmo desarrollado tiene importantes problemas de clasificación, y se confunde con facilidad con los géneros: Celosia, Amaranthus y Alternanthera. Por otra parte, se puede destacar que el género Alternanthera tiene una importante probabilidad de acierto. La causa principal de los anteriores resultados es que la base de datos generada durante este proyecto cuenta con características que dificultan la clasificación, a diferencia de la base de datos Flavia. Por tal se ha tomado la decisión de realizar entrenamiento de uno contra todos con las configuraciones más destacadas al realizar todos contra todos.

### 13.7. SELECCIÓN DEL CLASIFICADOR CON SUS RESPECTIVOS HIPERPARÁMETROS PARA AMARANTHACEAUQ USANDO UNO CONTRA TODOS

El uno contra todos se enfoca en entrenar cada una de las clases contra todas las otras a la vez, por tal en este caso se obtiene nueve resultados (uno por género), una por clase. Dado a que cada entrenamiento genera su propia configuración, no se puede encontrar una configuración (clasificador con sus hiperparámetros) ideal para todas las clases a la vez, por tal, se usan las configuraciones que arrojaron los mejores resultados en el entrenamiento todos contra todos, y se seleccionará la que genere el mejor promedio de resultados. Para más detalle de las mejores configuraciones ver Anexo 3.

#### 13.7.1. Configuraciones destacadas para MLP

Según los resultados obtenidos en el entrenamiento todos contra todos las siguientes configuraciones son las que generar mejores resultados en precisión o exhaustividad.

- {'activation': 'relu', 'alpha': 1e-08, 'hidden\_layer\_sizes': (50,), 'max\_iter': 6000, 'solver': 'adam'}
- {'activation': 'tanh', 'alpha': 1e-08, 'hidden\_layer\_sizes': (25,), 'max\_iter': 6000, 'solver': 'adam'}
- {'activation': 'logistic', 'alpha': 1e-05, 'hidden\_layer\_sizes': (50,), 'max\_iter': 6000, 'solver': 'adam'}
- {'activation': 'relu', 'alpha': 0.1, 'hidden\_layer\_sizes': (50,), 'max\_iter': 6000, 'solver': 'adam'}

Agregando parámetros al azar se han obtenido buenos resultados con la siguiente configuración:

- {'activation': 'relu', 'alpha': 1.0, 'hidden\_layer\_sizes': (25,), 'max\_iter': 6000, 'solver': 'lbfgs'}

En la siguiente tabla se pueden ver los resultados obtenidos.

accuracy promedio	activación	alphan	hidden_layer_sizes	max_iter	solver
0.925	relu	1.0	(25,)	6000	lbfgs
0.922	relu	0.1	(50,)	6000	adam
0.922	relu	1e-08	(50,)	6000	adam
0.919	logistic	1e-05	(50,)	6000	adam

0.906	tanh	1e-08	(25,)	6000	adam
-------	------	-------	-------	------	------

**Tabla 15.** Resultados de entrenamiento uno contra todos con MLP.

### 13.7.2. Configuraciones destacadas para SVM

Según los resultados obtenidos en el entrenamiento todos contra todos las siguientes configuraciones son las que generar mejores resultados en precisión o exhaustividad.

- {'C': 100.0, 'gamma': 0.001, 'kernel': 'sigmoid'}
- {'C': 100000.0, 'gamma': 0.0001, 'kernel': 'rbf'}
- 

Agregando parámetros al azar se han obtenido buenos resultados con la siguiente configuración:

- {'C': 100.0, 'gamma': 0.01, 'kernel': 'rbf'}

En la siguiente tabla se pueden ver los resultados obtenidos.

accuracy promedio	C	gamma	kernel
0.918	100	0.01	rbf
0.916	100	0.001	sigmoid
0.89	100000	0.0001	rbf

**Tabla 16.** Resultados de entrenamiento uno contra todos con SVM.

### 13.7.3. Configuraciones destacadas para K-NN

Debido a que los resultados obtenidos con K-NN han sido claramente por debajo de los de MLP y SVM, no se tomarán en cuenta ninguna de las configuraciones con este clasificador.

### 13.7.4. Resultados

Los mejores resultado recopilados en el entrenamiento uno contra todos fueron aplicados generando resultado destacados, sin embargo, la mejor solución fue construida usando hiperparametrós al azar. Por tal se puede inferir que la separación que se debe realizar en el entrenamiento de todos contra todos, difiere bastante al disminuir la cantidad de clases a clasificar.

La configuración seleccionada para el entrenamiento definitivo de uno contra todos es usando MLP con:

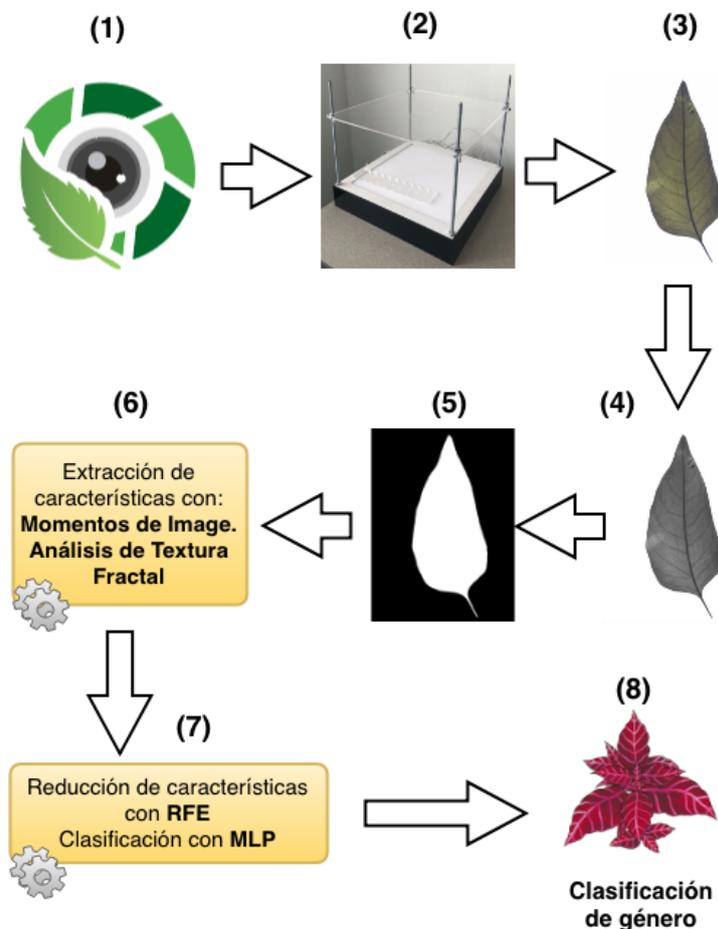
```

{ 'activation': 'relu',
  'alpha': 1.0,
  'hidden_layer_sizes': (25,),
  'max_iter': 6000,
  'solver': 'lbfgs'
}

```

#### 14. ESQUEMA GENERAL DE LA SOLUCIÓN TECNOLÓGICA DESARROLLADA

El proceso realizado para la identificación taxonómica de especies vegetales de la familia Amaranthaceae fue basado en los ocho pasos principales descritos en la Figura 36.



**Figura 36.** Esquema general de la solución tecnológica desarrollada.

Los pasos 1 y 2 están asociados en ubicar el dispositivo móvil en el ERAF01. Paso 3, usando la aplicación móvil desarrollada se toman las fotografías (o se cargan desde la galería) de contorno y zoom de la hoja, teniendo en cuenta el proceso

definido en la Sección 9.2. y se envían estas al servidor para iniciar la identificación taxonómica.

El paso 4 describe la transformación de las imágenes a escala de grises, para posteriormente segmentarlas en el paso 5. En el paso 6 se usan los momentos de la imagen y SFTA para la extracción de características. Para aplicar la primera técnica fue necesario tener una imagen binarizada para resaltar el contorno de la hoja. Para SFTA se usa inicialmente una imagen en escala de grises. Estas dos técnicas son aplicada en cada una de las imágenes, con momentos de la imagen se obtiene 24 características y con SFTA 30, en total por ejemplo se generan 108 características.

En el séptimo paso se seleccionan las 46 (de las 108) características que más aportan a la clasificación, para esto se usa el algoritmo RFE, seguido se realiza el entrenamiento y clasificación uno contra todos usando MLP. Para finalizar, en el paso ocho se da como resultado el género identificado con su respectiva taxonomía.

## 15. ARQUITECTURA DEL PROTOTIPO DE LA SOLUCIÓN TECNOLÓGICA

El documento de arquitectura de software está basado en el Modelo de Vistas de Arquitectura 4+1 propuesto Kruchten [32]. Con la arquitectura de software se busca proporcionar una base conceptual y tecnológica para los integrantes del equipo de desarrollo y stakeholders del prototipo de aplicación para dispositivos móviles para la identificación taxonómica de especies vegetales de la familia de plantas Amaranthaceae.

### 15.1. Vista de escenarios

En esta vista se agregaron los escenarios de calidad que representa los atributos de calidad que se deben satisfacer. A continuación se muestra en la Tabla 17 un ejemplo de los escenarios de calidad definidos.

<b>Rendimiento</b>	
Se desea que el prototipo desarrollado funcione de forma fluida en todas las funcionalidades que no necesitan internet. Todas las operaciones que realice el prototipo deben ser ejecutadas en un intervalo de tiempo aproximado de dos segundo, las tareas costosas deben de ejecutar en un hilo secundario.	
Elemento	Descripción
<b>Origen del estímulo</b>	Fuentes Externas
<b>Estímulo</b>	Rápida respuesta de la app en funcionalidades que no necesitan de internet.
<b>Ambiente</b>	Condiciones normales.

<b>Componentes</b>	Toda la app.
<b>Respuesta</b>	La app cargará la mayor cantidad de recursos en tiempo de compilación. Las navegación no generan más de tres actividades en la pila. Se evitarán procesos largos en segundo plano.
<b>Medida de la Respuesta</b>	La app funciona de forma fluida, cada una de las funcionalidades se ejecuta en un tiempo inferior o igual a dos segundos.

En los momentos en que se necesite conexión a internet para alguna funcionalidad, la app mostrará mensajes que le indiquen al usuario si necesita esperar o si necesita conexión a internet. En todo caso el usuario estará enterado del porqué del funcionamiento de la app.

**Tabla 17.** Representación del atributo de calidad de rendimiento.

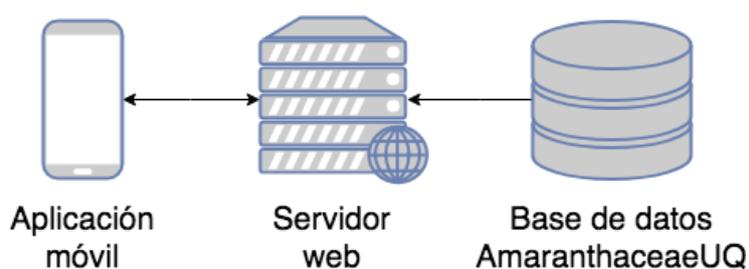
En total los atributos de calidad que deben satisfacer por la solución tecnológica planteada fueron:

- Rendimiento.
- Mantenibilidad.
- Disponibilidad.
- Portabilidad
- Integridad conceptual.

Para más detalle revise el Anexo 4.

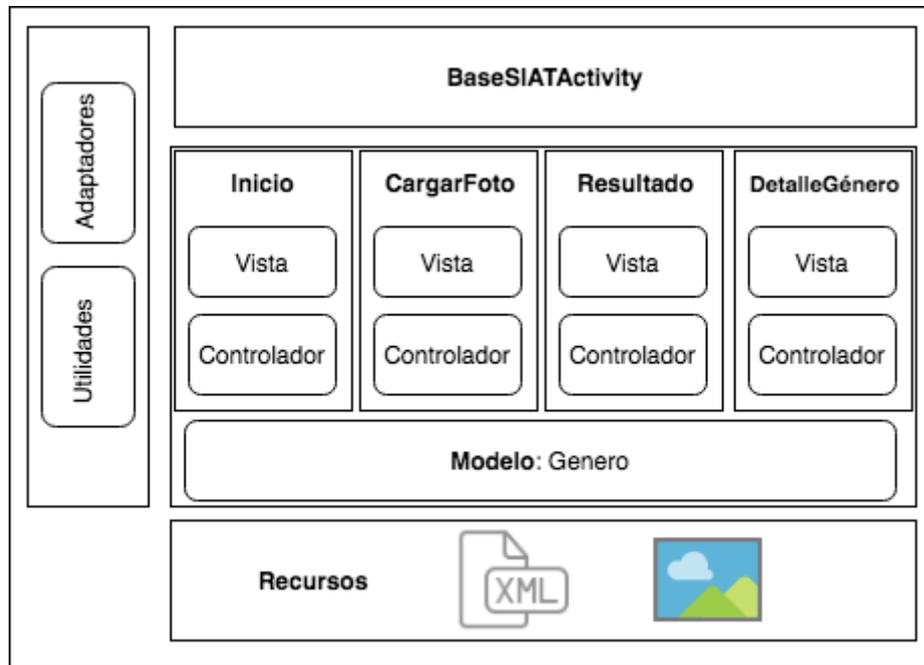
## 15.2. Vista lógica

En la vista lógica se definen la combinación de estilos y patrones arquitectónicos con los que se busca satisfacer los atributos de calidad identificados. El estilo arquitectónico cliente servidor se usa para la división de responsabilidades. Tal como se muestra en la Figura 37 la solución tecnológica definida está compuesta de una aplicación móvil y de un servidor web que se conecta a la base de datos AmaranthaceaeUQ.



**Figura 37.** Diagrama arquitectónico Cliente-Servidor.

La aplicación móvil está desarrollada con base a la combinación de estilos arquitectónico por capas y componentes, cada componente internamente es desarrollado usando el patrón Modelo Vista Controlador. La Figura 38 representa la distribución arquitectónica general de la aplicación móvil.

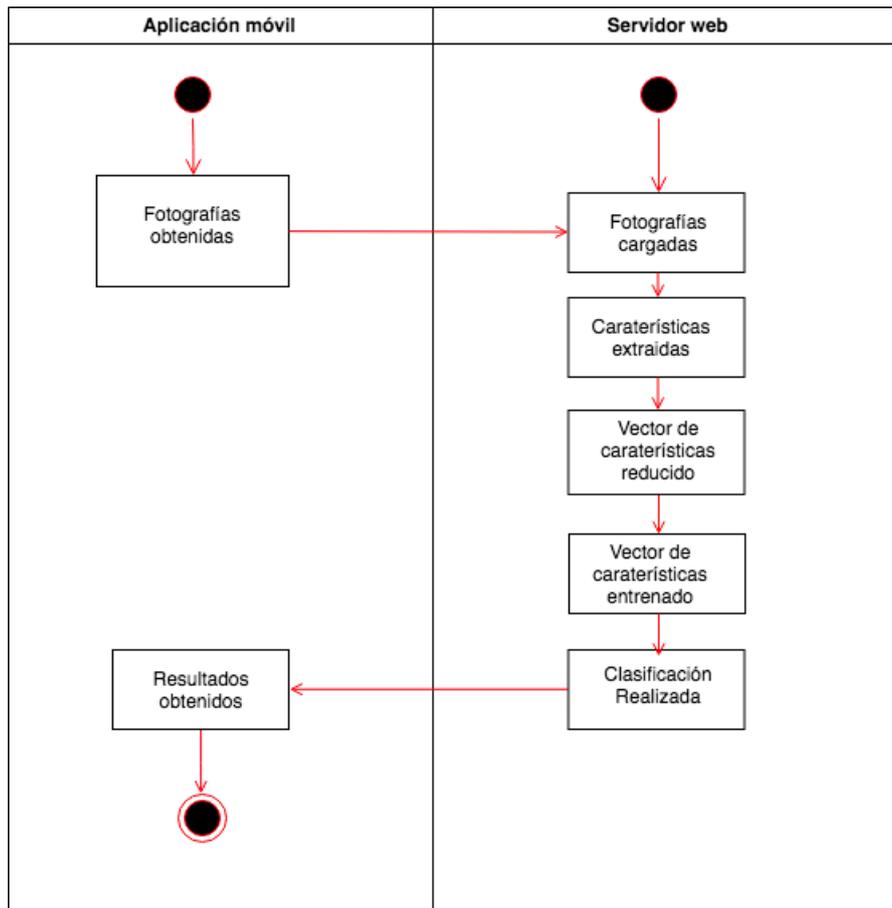


**Figura 38. Diagrama arquitectónico de la aplicación móvil.**

Para más detalle de la vista lógica revise el Anexo 4.

### 15.3. Vista de proceso

La solución tecnológica planteada tiene un proceso general y principal, el cual se centra en la identificación taxonómica de especies vegetales de la familia Amaranthaceae. Este proceso es representado en la Figura 39.



**Figura 39. Proceso para realizar la identificación taxonómica.**

Para más detalle de la vista de proceso revise el Anexo 4.

#### 15.4. Vista de despliegue

El servidor web está en ejecución sobre una máquina que cuenta con un procesador Intel Xeon E5-265 v3 a 2.30 GHz con 1 core de CPU, 2 GB de memoria RAM, el servidor está gestionado por Debian 3.16.43 como sistema operativo. El servidor está físicamente en la Universidad del Quindío y pertenece al Grupo de investigación en redes, información y distribución (GRID).

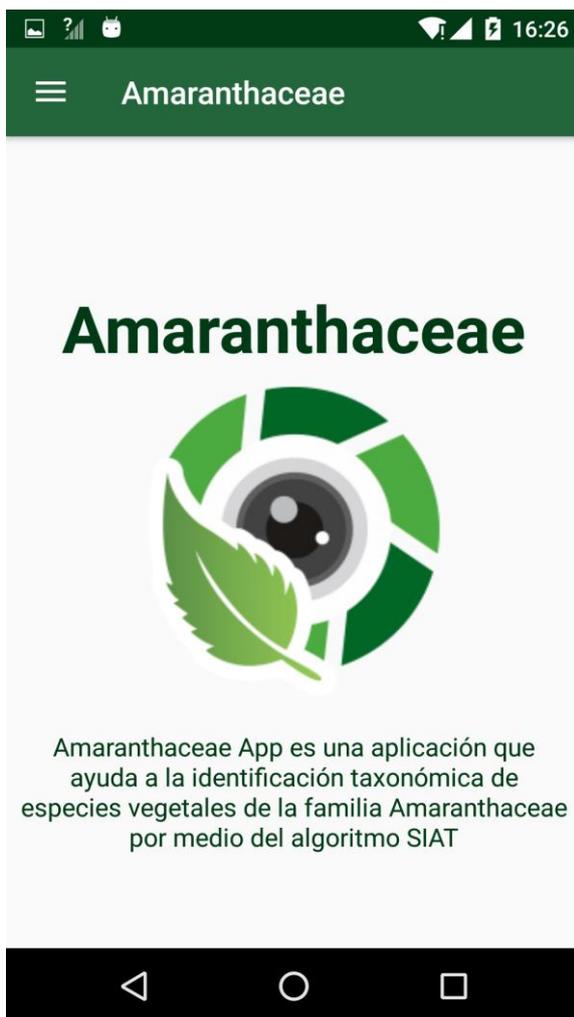
#### 15.5. Vista de implementación

En la vista de implementación se definió de manera específica los elementos que conforma la solución tecnológica planteada. Para más detalle de la vista de implementación revise el Anexo 4.

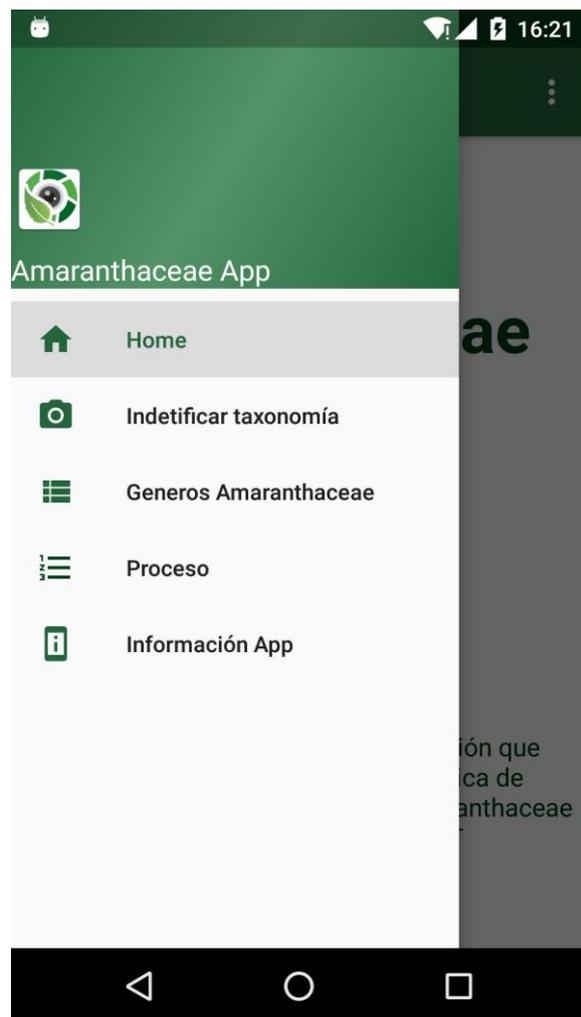
## 16. DESCRIPCIÓN DE LA APLICACIÓN MÓVIL

La aplicación móvil desarrollada está disponible para cualquier teléfono cuyo sistema operativo sea Android, y cumple con el diseño material design propuesto por Google. A continuación se muestran algunas capturas de pantalla de la aplicación.

La Figura 40 muestra la pantalla de inicio de la aplicación, en esta pantalla hay una descripción corta, el logo y hay un menú lateral. En la Figura 41 se puede observar el menú lateral que cuenta con cinco opciones: Inicio, Identificar taxonomía, Géneros Amaranthaceae, Proceso e Información de la App.



**Figura 40.** Inicio



**Figura 41.** Menú desplegable

La funcionalidad más importante de la aplicación móvil es la de identificar la taxonomía, su pantalla inicial es la que se muestra en la Figura 42, la cual cuenta con dos botones, uno para tomar la foto directamente desde la aplicación y otro para subir una foto existente en la galería de imágenes.



Figura 42. Clasificar inicio.

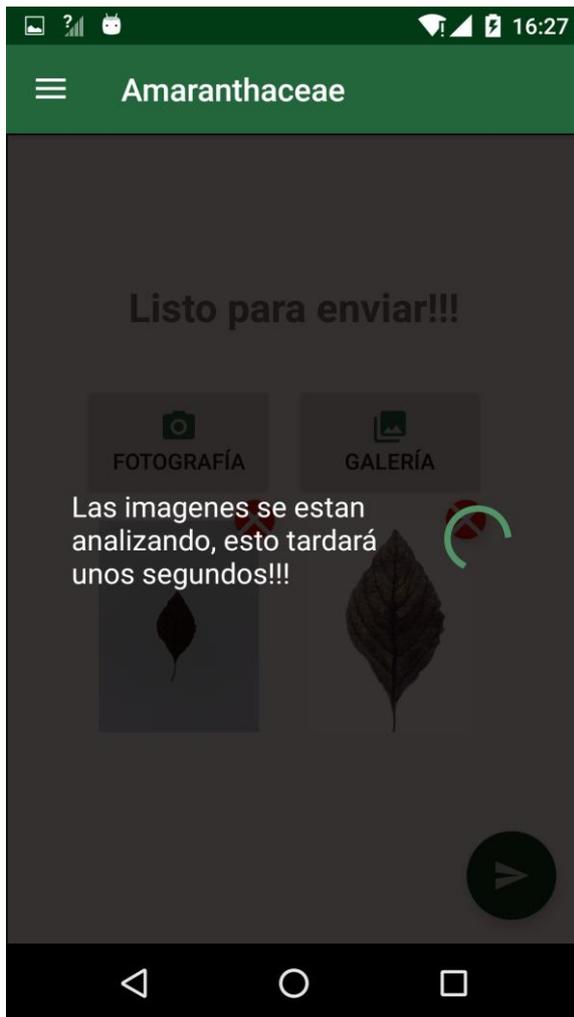


Figura 43. Clasificar imágenes.

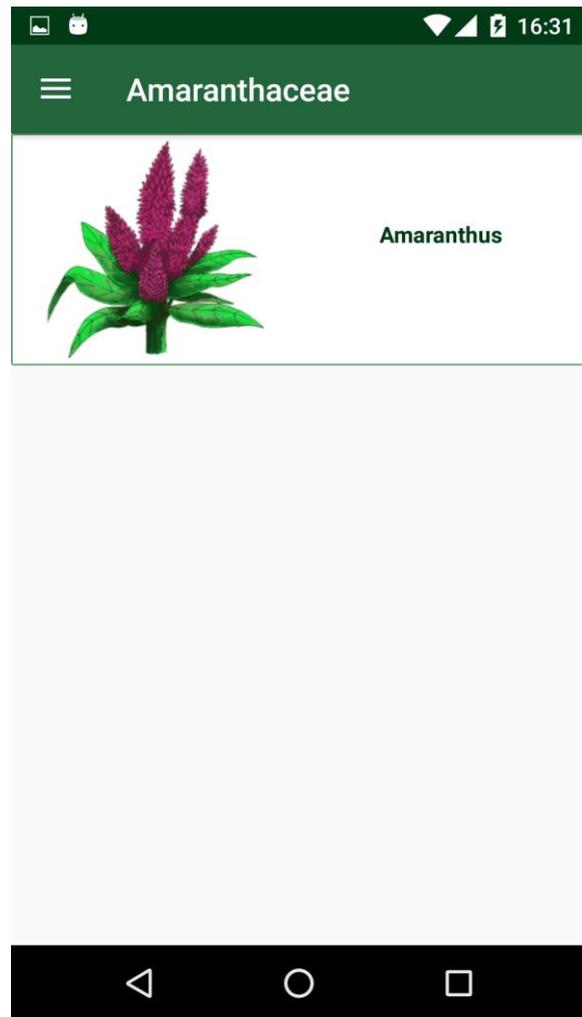
Sea cual sea la elección, se deben seleccionar dos fotografías (Figura 43) una para el contorno y otra para el zoom. Ambas son obligatorias.

Una vez seleccionadas las fotografías se pueden enviar mediante el botón correspondiente (botón flotante), en ese momento se lanza una tarea en segundo plano la cual se encarga de mandar las imágenes al servidor web mediante una petición HTTP, al ser en segundo plano se muestra un mensaje al usuario (Figura 44) donde se le indica que las imágenes se están analizando.

En la Figura 45 se muestra la respuesta devuelta por el servidor donde se indica el género al que pertenecen las fotografías de las hojas enviadas.



**Figura 44.** Tarea en ejecución



**Figura 45.** Respuesta de la clasificación.

Una vez obtenida la respuesta de la clasificación se puede ver el detalle e información adicional sobre el género devuelto (Figura 46), el cual brinda datos como subfamilias, tribus, especies, descripción, estado del género en Colombia, etc.

Tal como se muestra en la Figura 41 la siguiente funcionalidad consiste en listar todos los géneros de la familia Amaranthaceae contemplados este trabajo (en total son 9), cada ítem de la lista (Figura 47) es un género y al presionarlo muestra su detalle o descripción así como se muestra en la Figura 46.

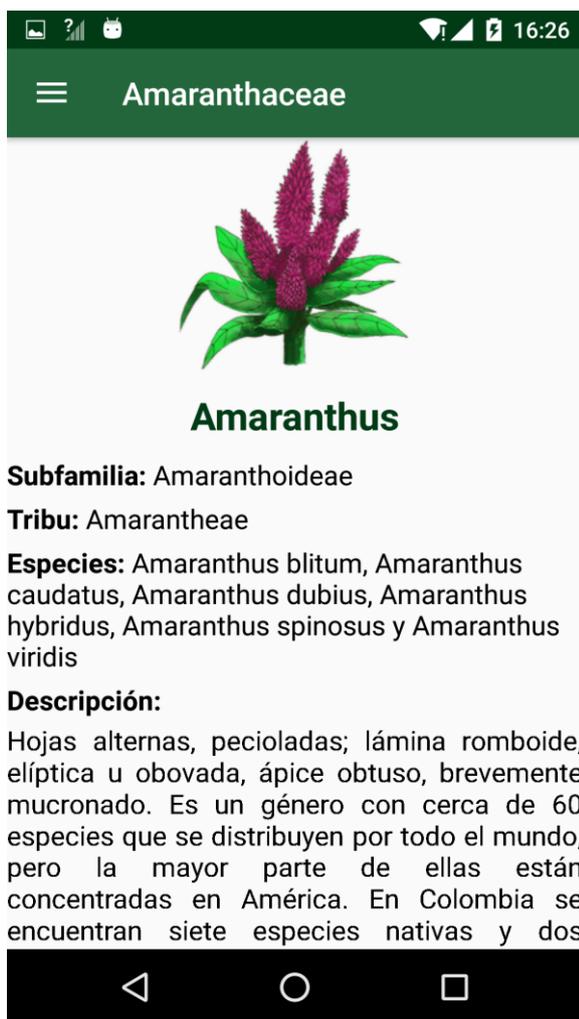


Figura 46. Detalle del género.

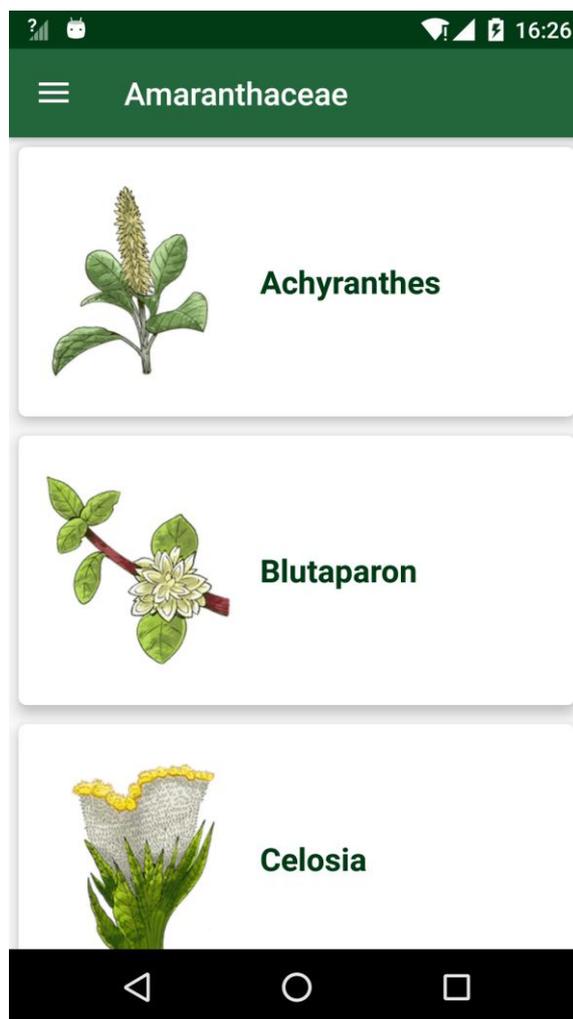
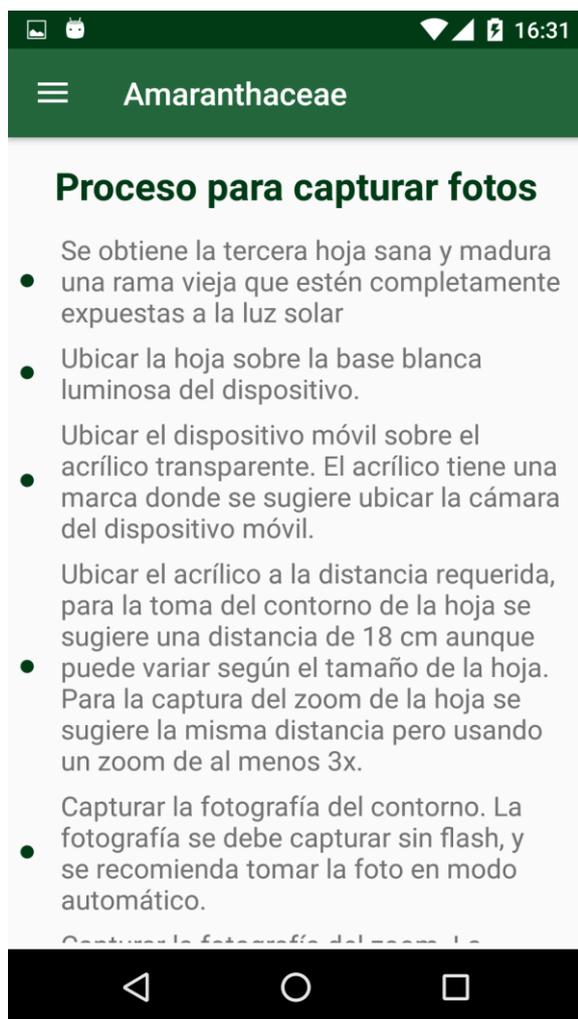
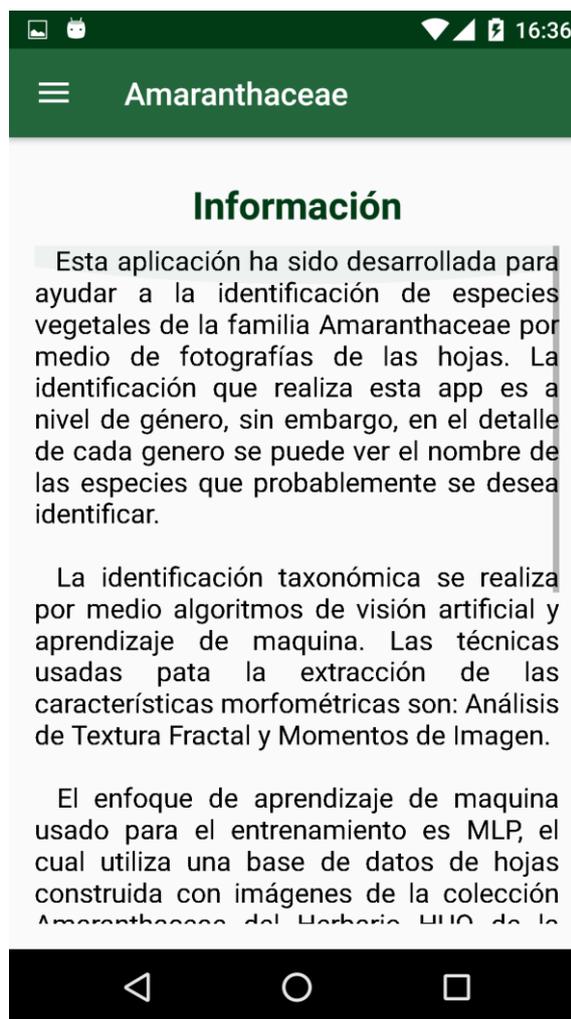


Figura 47. Lista de géneros.

Las dos funcionalidades restantes de la aplicación son: una guía donde se muestra el proceso que se debe seguir para capturar las fotografías de las hojas de las especies vegetales (Figura 48) y la información general de la aplicación y el contexto en el que está desarrollada, así como las técnicas y algoritmos usados para la clasificación de las fotografías (Figura 49).



**Figura 48.** Descripción pasos para captura.



**Figura 49.** Información acerca de la app.

## 17. PRUEBAS FUNCIONALES DE LA SOLUCIÓN TECNOLÓGICA

Las pruebas funcionales de la solución tecnológica giran en torno a comprobar la tasa de acierto en la clasificación de especies vegetales y su correcta identificación taxonómica. La clasificación está soportada por un algoritmo anteriormente definido, por cada instancia de la especie vegetal se hicieron dos pruebas, usando la técnica todos contra todos y uno contra todos.

Para la siguiente prueba se seleccionaron de manera aleatoria 40 fotografías de hojas de especies vegetales de la familia Amaranthaceae capturadas con un teléfono iPhone 8 plus, y se realizó su clasificación taxonómica a nivel de género, estos fueron los resultados teniendo en cuenta que a cada género se le asignaron los siguientes números: Achyrantes=0, Blutaparon=1, Celosia=2, Chamissoa=3, Cyathula=4, Gomphrena=5, Iresine=6, Amaranthus=7, Alternanthera=8.

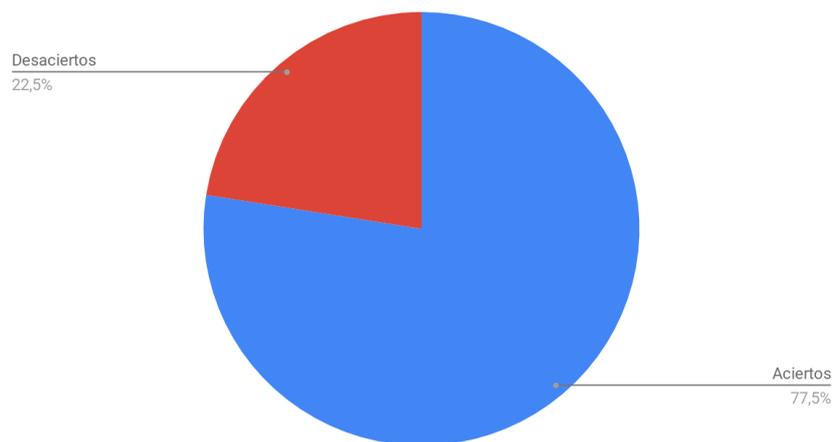
Especie vegetal	Clasificación		Resultado esperado	Validación	
	Todos contra todos (TxT)	Uno contra todos (UxT)		TxT	UxT
10a.png	5	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Gomphrena serrata	ok	x
3a.png	5	[1, 1, 1, 1, 1, 4, 1, 1, 1]	Gomphrena serrata	ok	ok
9a.png	5	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Gomphrena serrata	ok	x
2a.png	5	[1, 1, 1, 1, 1, 4, 1, 1, 1]	Gomphrena serrata	ok	ok
1a.png	5	[1, 1, 1, 1, 1, 4, 1, 1, 1]	Gomphrena serrata	ok	ok
1a.png	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus caudatus	ok	x
5a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus caudatus	x	x
7a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus caudatus	ok	ok
3a.png	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus caudatus	ok	x
4a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus caudatus	x	x
4a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Achyranthes aspera var. pubescens	x	x
1a.png	0	[4, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	ok
2a.png	2	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	x	x
6a.png	0	[4, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	ok
5a.png	2	[4, 1, 4, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	x	-

4a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus hybridus	x	x
11a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus hybridus	ok	ok
5a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus hybridus	ok	ok
1a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus hybridus	ok	ok
2a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus hybridus	ok	ok
11a.png	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus spinosus	ok	ok
6a.png	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus spinosus	ok	x
8a.png	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus spinosus	ok	x
7a.png	7	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Amaranthus spinosus	ok	x
2a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus spinosus	x	x
11a.png	7	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera albotomentosa	x	ok
6a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera albotomentosa	ok	ok
15a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera albotomentosa	ok	ok
3a.png	5	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Alternanthera albotomentosa	x	x
7a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera albotomentosa	ok	ok
3a.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
7a.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
4a.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok

8a.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
10a.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
4a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Alternanthera lanceolata	ok	x
7a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera lanceolata	ok	ok
3a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Alternanthera lanceolata	ok	x
2a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera lanceolata	ok	ok
8a.png	8	[1, 1, 1, 1, 1, 1, 1, 1, 4]	Alternanthera lanceolata	ok	ok

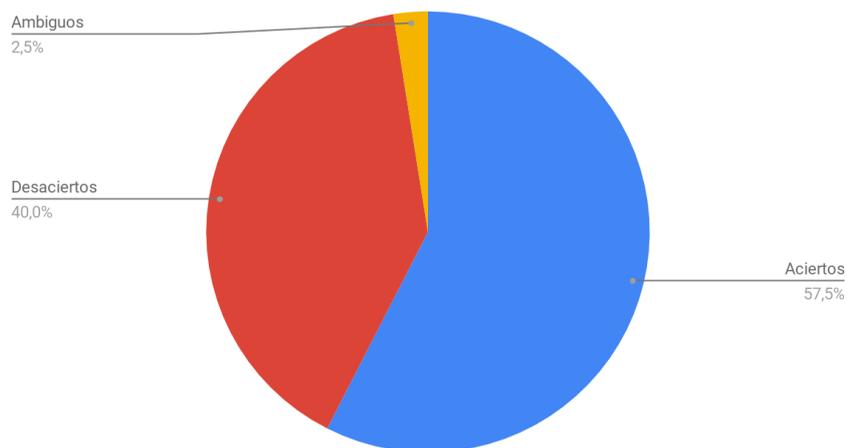
**Tabla 18.** Resultados prueba iPhone 8 plus.

En conclusión, se obtuvo: Todos contra todos: 31 aciertos, 9 desaciertos (Figura 50).



**Figura 50.** Resultados de la prueba todos contra todos iPhone 8 plus.

Uno contra todos: 23 aciertos, 16 desaciertos y 1 ambiguas (aunque uno de los valores obtenidos corresponde al género correcto), así como se muestra en la Figura 51.



**Figura 51.** Resultados de la prueba uno contra todos iPhone 8 plus.

La siguiente prueba funcional consiste en clasificar 40 fotografías de hojas de especies vegetales capturadas mediante un dispositivo móvil diferente al que se usó para el entrenamiento del algoritmo, específicamente un iPhone 7, las imágenes fueron elegidas al azar, a continuación los resultados:

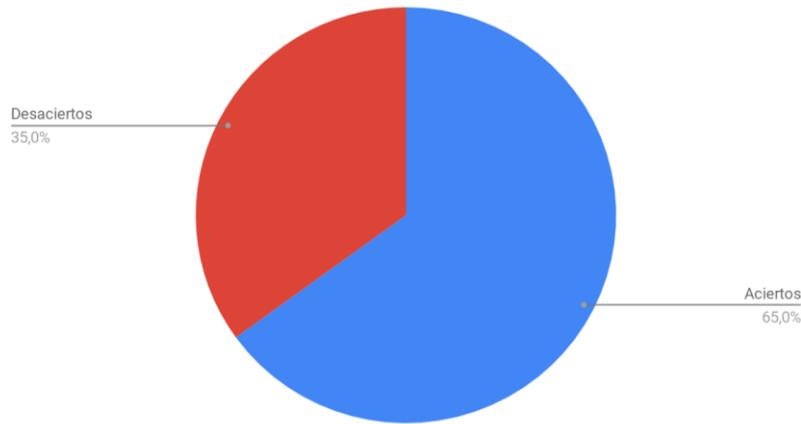
Especie vegetal	Clasificación		Resultado esperado	Validación	
	Todos contra todos (TxT)	Uno contra todos (UxT)		TxT	UxT
4b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia virgata	ok	ok
3b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia virgata	ok	ok
2b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia virgata	ok	ok
6b.jpg	2	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Celosia virgata	ok	x
5b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia virgata	ok	ok
5b.jpg	1	[1, 4, 1, 1, 1, 1, 1, 1, 1]	Blutaparon vermiculare	ok	x
3b.jpg	1	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Blutaparon vermiculare	ok	x
7b.jpg	6	[1, 1, 4, 1, 1, 1, 4, 1, 1]	Blutaparon vermiculare	x	-

4b.jpg	2	[4, 4, 4, 1, 1, 1, 1, 1, 1]	Blutaparon vermiculare	x	-
9b.jpg	2	[1, 4, 4, 1, 1, 1, 1, 1, 1]	Blutaparon vermiculare	x	-
2b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia argentea var. argentea	ok	ok
4b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia argentea var. argentea	ok	ok
5b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia argentea var. argentea	ok	ok
3b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Celosia argentea var. argentea	ok	ok
1b.jpg	8	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Celosia argentea var. argentea	x	x
9b.jpg	4	[1, 1, 4, 1, 4, 1, 1, 1, 1]	Cyathula prostrata	ok	-
2b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Cyathula prostrata	x	x
5b.jpg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Cyathula prostrata	x	x
11b.jpg	4	[1, 1, 4, 1, 4, 1, 1, 1, 1]	Cyathula prostrata	ok	-
8b.jpg	4	[1, 1, 1, 1, 4, 1, 1, 1, 1]	Cyathula prostrata	ok	ok
10b.jpeg	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus viridis	ok	x
1b.jpeg	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Amaranthus viridis	x	x
7b.jpeg	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus viridis	ok	x
2b.jpeg	2	[1, 1, 4, 4, 1, 1, 1, 1, 1]	Amaranthus viridis	x	-
4b.jpeg	2	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Amaranthus viridis	x	x
2b.jpg	4	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Iresine herbstii	x	x

1b.jpg	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
9b.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
3b.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
5b.png	6	[1, 1, 1, 1, 1, 1, 4, 1, 1]	Iresine herbstii	ok	ok
4b.jpg	0	[4, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	ok
5b.jpg	0	[4, 1, 4, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	-
7b.jpg	0	[4, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	ok
1b.jpg	0	[4, 1, 4, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	ok	-
3b.jpg	7	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Achyranthes aspera var. pubescens	x	x
8b.jpeg	7	[1, 1, 4, 1, 1, 1, 1, 1, 1]	Amaranthus blitum	ok	x
3b.jpeg	7	[1, 1, 1, 1, 1, 1, 1, 4, 1]	Amaranthus blitum	ok	x
7b.jpeg	2	[4, 1, 4, 1, 1, 1, 1, 1, 1]	Amaranthus blitum	x	-
2b.jpeg	4	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus blitum	x	x
4b.jpeg	0	[1, 1, 1, 1, 1, 1, 1, 1, 1]	Amaranthus blitum	x	x

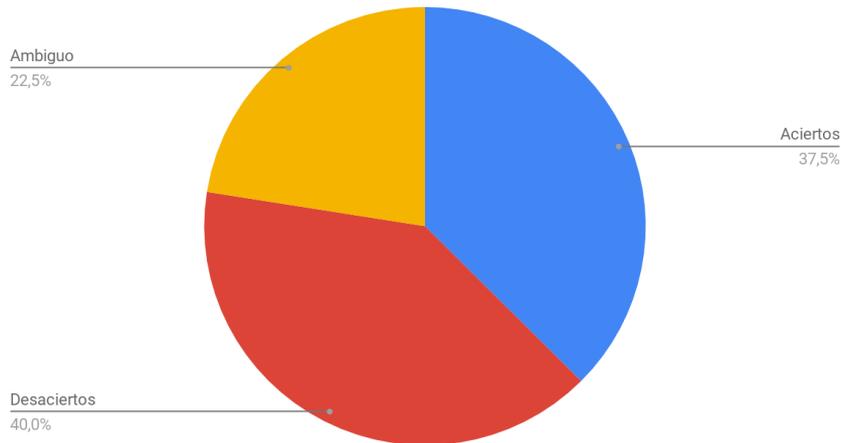
**Tabla 19.** Resultados prueba iPhone 7.

En conclusión, se obtuvo: Todos contra todos: 26 aciertos, 14 desaciertos. (Figura 52)



**Figura 52.** Resultados de la prueba todos contra todos iPhone 7.

Uno contra todos: 15 aciertos, 16 desaciertos y 9 ambiguas. (Figura 53)



**Figura 53.** Resultados de la prueba uno contra todos iPhone 7.

## 18. RESUMEN DE RESULTADOS

En la siguiente tabla se indican los resultados obtenidos por cada uno de los objetivos específicos planteados.

Objetivo	Resultados
Especificar un modelo formal de taxonomía para la familia Amaranthaceae utilizando características morfométricas de las hojas de cada una de las especies involucradas en el estudio.	Sección 6. Sección 12.
Desarrollar una metodología para la extracción de características morfológicas de las hojas por medio de	Sección 7. Sección 8. Sección 9.

reconocimiento digital de imágenes y aprendizaje de máquina, que faciliten la identificación automática de especies vegetales de la familia Amaranthaceae usando morfometría geométrica.	Sección 10. Sección 11. Anexo 1. Base de datos AmaranthaceaeUQ. Anexo 2. Prueba de concepto para validar el modelo generado para la clasificación de especies vegetales, Base de datos de prueba Flavia.
Desarrollar una metodología para el reconocimiento taxonómico de especies vegetales dentro de la familia Amaranthaceae utilizando técnicas de aprendizaje de máquina.	Sección 12. Sección 13. Sección 14. Anexo 3. Selección de clasificador con sus respectivos hiperparámetros.
Desarrollar el prototipo de un sistema para dispositivos móviles de reconocimiento de patrones basado en técnicas de aprendizaje de máquina, para facilitar la identificación automática de especies vegetales de la familia Amaranthaceae.	Sección 15. Sección 16. Anexo 4. Documento de Arquitectura de software SIAT. Anexo 5. Código fuente aplicación móvil. Anexo 6. Código fuente servidor.
Validar la exactitud del reconocimiento obtenido usando el prototipo desarrollado para facilitar la identificación automática de especies vegetales de la familia Amaranthaceae.	Sección 17. Anexo 2. Prueba de concepto para validar el modelo generado para la clasificación de especies vegetales, Base de datos de prueba Flavia. Anexo 3. Selección de clasificador con sus respectivos hiperparámetros.

**Tabla 20.** Tabla de resultados.

## CONCLUSIONES Y TRABAJOS FUTUROS

Aunque la precisión con la base de datos creada en este proyecto fue de 0.66, gracias al entrenamiento uno contra todos esta solución tecnológica es una herramienta que ayuda y facilita (gracias a la facilidad que brinda la aplicación móvil) destacadamente a la identificación taxonómica de géneros de Amaranthaceae, sin embargo, con un mejora en la base de datos AmaranthaceaeUQ, que permita tener imágenes de entrenamiento similares a las de la Flavia y sean capturadas siguiendo el proceso definido en este proyecto, se puede garantizar una identificación taxonómica de las especies de plantas de la familia Amaranthaceae con una precisión y exhaustividad cercana al 0.95, tal como se muestra en el Anexo 2. Esta solución tecnológica permitirá que los interesados no tengan que ser taxónomos expertos en la familia para realizar una clasificación adecuada y confiable.

En el procedimiento que se plantea en este proyecto para la captura de las fotografías de las especies vegetales, se define que la hoja debe estar en una postura específica sobre un fondo blanco (se debe usar ERAF o algún dispositivo similar), y para su correcta clasificación la fotografía se debe capturar en modo retrato donde la base de la hoja esté en la parte baja y su ápice en la parte alta de la imagen. Esto satisface la primera de las tres problemáticas que se plantea en la sección de Justificación.

Los estudios enfocados en la identificación de especies de plantas mediante técnicas de reconocimiento de patrones deben basarse en taxones, ya que, como se indica en [25], hay plantas de diferentes taxones que pueden tener patrones bastante similares entre sí, esto causa que ocurran errores en el reconocimiento de especies. Además es importante que los estudios de identificación de especies de plantas se realicen teniendo en cuenta la taxonomía de estas, ya que de esta forma se proporciona información más completa y valiosa. También es recomendable que si el entrenamiento se realiza a nivel de especie se tengan en cuenta las variedades de plantas por separado. Esto satisface la segunda de las tres problemáticas que se plantea en la sección de Justificación.

Las especies vegetales consideradas en este trabajo no son compuestas, todas son hojas simples, aunque de ser compuestas se deberían tratar igual, primero segmentarlas, extraer sus características y luego clasificarla mediante el algoritmo desarrollado. Esto satisface la tercera de las tres problemáticas que se plantea en la sección de Justificación.

Las hojas utilizadas para la creación de las bases de datos con el objetivo de identificar especies de plantas deben ser recolectadas por biólogos expertos o taxónomos en un ambiente controlado, un ejemplo de esto es la selección de hojas hechas en [25], donde obtienen la tercera hoja sana y madura de ramas viejas completamente expuestas a la luz solar.

De acuerdo con la base de datos que se usa, las características que más contribuyen a la clasificación variarán, por lo tanto, el algoritmo de identificación de especies vegetales desarrollado en este proyecto puede generar un modelo que permita reconocer especies por taxón.

Nuestras principales contribuciones en este proyecto son: la identificación de especies vegetales mediante el análisis de su taxonomía fractal, para modelar morfométricamente la venación y la textura de las hojas; utilizar RFE para la selección de características morfométricas de hojas, en función de los criterios generados por un SVM; una solución tecnológica que ayuda a la identificación taxonómica de especies vegetales de la familia Amaranthaceae; y principalmente, la creación de un proceso con el cual se establece la forma de cómo deben ser capturadas las fotos que se desean clasificar (o incluso se deseen agregar a AmaranthaceaeUQ) para su posterior identificación.

## TRABAJOS FUTUROS

Mejora la base de datos AmaranthaceaeUQ. Inicialmente se debería enfocar en la mejora de las imágenes asociadas a las especies como *Achyranthes*, dado que como se observa en la Figura 35, este género tiene problemas de clasificación. Cada especie debería tener entre 60 y 70 fotografías (muestras) capturadas siguiendo el proceso definido en esta la Sección 9.2. Además se debería ampliar el número de especies soportadas, esto hace que también se tenga que aumentar el número de muestras por clase.

Desarrollar una página web que permita gestionar la base de datos AmaranthaceaeUQ y realizar identificación taxonómica sin la necesidad de una aplicación móvil.

Desarrollar una aplicación móvil que permita el soporte para los usuarios del sistema operativo iOS.

Experimentar con nuevos algoritmos de extracción de características morfométricas que no hayan sido ensayados. Además de tener una base de datos que permita, usar aprendizaje profundo para la extracción de información cruda de las imágenes, tal como se muestra en investigaciones como [24].

Realizar experimentos usando identificación taxonómica con variadas familias de plantas.

## REFERENCIAS

- [1] M. J. & B. J. W. Christenhusz, «The number of known plants species in the world and its annual increase. *Phytotaxa*,» vol. 261, nº 3, 2016.
- [2] C. Agudelo Henao, «'Amaranthaceae'. 1st ed. Bogotá: Instituto de ciencias naturales - Museo de historia natural - Facultad de ciencias - Universidad nacional.,» 2008.
- [3] Lu H, Jiang W, Ghiassi M, Lee S, Nitin M (2012) Classification of *Camellia* (Theaceae) Species Using Leaf Architecture Variations and Pattern Recognition Techniques. *PLoS ONE* 7(1): e29704. <https://doi.org/10.1371/journal.pone.0029704>
- [4] Leaf Architecture Working Group, «Manual of Leaf Architecture: Morphological Description and Categorization of Dicotyledonous and Net-veined Monocotyledonous Angiosperms,» 1999.
- [5] Cope, JS, Remagnino, P, Corney, D, Clark, JY and Wilkin, P (2012) Plant species identification using digital morphometrics: A review *Expert Systems with Applications*, 39 (8). pp. 7562-7573.
- [6] D. Mendieta Domínguez, «Reconocimiento de Objetos Bidimensionales en Imágenes mediante la Transformada de Distancia utilizando Matlab,» Universidad de las Américas Puebla, 2003.
- [7] J. Alvarado Moya, «Procesamiento y Análisis de Imágenes Digitales,» Instituto Tecnológico de Costa Rica, 2012.
- [8] C. C. S. S. C. W. & C. L. Zhao, «Plant identification using leaf shapes—A pattern counting approach. *Pattern Recognition*,» vol. 48, nº 10, pp. 3203-3215, 2015.
- [9] A. Hernández Serna y L. F. Jiménez Segura, «Automatic identification of species with neural networks,» *Peerj*, 2014.
- [10] R. F. C. C. a. B. L. P. Hernández Sampieri, «Metodología de la investigación (4a. ed.). 1st ed. Distrito Federal: McGraw-Hill Interamericana.,» 2006.
- [11] C. Uluturk and A. Ugur, "Recognition of leaves based on morphological features derived from two half-regions," in *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 2012, pp. 1-4.
- [12] J.-x. Du, C.-M. Zhai, and Q.-P. Wang, "Recognition of plant leaf image based on fractal dimension features," *Neurocomputing*, vol. 116, pp. 150-156, 9/20/ 2013.
- [13] N. Valliammal and S. N. Geethalakshmi, "Efficient feature fusion, selection and classification technique for Plant Leaf Image Retrieval system," presented at the

Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology, Coimbatore UNK, India, 2012.

[14] S. Prasad, S. K. Peddoju, and D. Ghosh, "Unsupervised resolution independent based natural plant leaf disease segmentation approach for mobile devices," presented at the Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop, New Delhi, India, 2013.

[15] P. Novotný and T. Suk, "Leaf recognition of woody species in Central Europe," *Biosystems Engineering*, vol. 115, no. 4, pp. 444-452, 2013/08/01/ 2013.

[16] N. Liu and J. Kan, "Improved deep belief networks and multi-feature fusion for leaf identification," *Neurocomputing*, vol. 216, pp. 460-467, 12/5/ 2016.

[17] J. K. Hsiao, L. W. Kang, C. L. Chang, and C. Y. Lin, "Comparative study of leaf image recognition with a novel learning-based approach," in 2014 Science and Information Conference, 2014, pp. 389-393.

[18] P. F. Barros,. "Development of a System for Automatic Plant Species Recognition." in 2013.

[19] S. Mouine, I. Yahiaoui, and A. Verroust-Blondet, "Advanced shape context for plant species identification using leaf image retrieval," presented at the Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, Hong Kong, China, 2012.

[20] L. Thi-Lan, T. Duc-Tuan, and H. Van-Nam, "Fully automatic leaf-based plant identification, application for Vietnamese medicinal plant search," presented at the Proceedings of the Fifth Symposium on Information and Communication Technology, Hanoi, Viet Nam, 2014.

[21] L. Thi-Lan, D. Nam-Duong, N. Van-Toi, V. Hai, H. Van-Nam, and N. T. Thanh-Nhan, "Complex Background Leaf-based Plant Identification Method Based on Interactive Segmentation and Kernel Descriptor," presented at the Proceedings of the 2nd International Workshop on Environmental Multimedia Retrieval, Shanghai, China, 2015.

[22] Herv et al., "The imageCLEF plant identification task 2013," presented at the Proceedings of the 2nd ACM international workshop on Multimedia analysis for ecological data, Barcelona, Spain, 2013.

[23] M. Ali Jan Ghasab, S. Khamis, F. Mohammad, and H. Jahani Fariman, "Feature decision-making ant colony optimization system for an automated recognition of

plant species," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2361-2370, 4/1/2015.

[24] S. Lee, C. Chan, S. Mayo, and P. Remagnino, "How deep learning extracts and learns leaf features for plant classification," *Pattern Recognition*, vol. 71, pp. 1-13, 11// 2017.

[25] Jiang, Wu & Barshan, Billur & Mantri, Nitin & Tao, Zhengming & Lu, Hongfei. (2013). Classification of *Camellia* species from 3 sections using leaf anatomical data with back-propagation neural networks and support vector machines. *Turkish Journal of Botany*. 37. 1093-1103. 10.3906/bot-1210-21.

[26] J. Unger, D. Merhof and S. Renner, "Computer vision applied to herbarium specimens of German trees: testing the future utility of the millions of herbarium specimen images for automated identification", *BMC Evolutionary Biology*, vol. 16, no. 1, 2016.

[27] H. Laga, S. Kurtsek, A. Srivastava, and S. J. Miklavcic, "Landmark-free statistical analysis of the shape of plant leaves," *Journal of Theoretical Biology*, vol. 363, pp. 41-52, 2014/12/21/ 2014.

[28] O. Guzmán P., E. Gomez G. y C. Rivillas Osorio. (2003). Utilización del procesamiento de imágenes para determinar la severidad de la mancha de hierro, en hojas de café, CENICAFE Federación Nacional de Cafeteros de Colombia.

[29] Costa, A. F., Humpire-Mamani, G., & Traina, A. J. M. (2012, August). An efficient algorithm for fractal analysis of textures. In *Graphics, Patterns and Images (SIBGRAPI)*, 2012 25th SIBGRAPI Conference on (pp. 39-46). IEEE.

[30] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422.

[31] scikit-learn (2018, octubre 11) 1. Supervised learning. [Online]. Available: [http://scikit-learn.org/stable/supervised\\_learning.html](http://scikit-learn.org/stable/supervised_learning.html)

[32] Kruchten, P. B. (1995). The 4+ 1 view model of architecture. *IEEE software*, 12(6), 42-50.

## **ANEXOS**

- Anexo 1. Base de datos AmaranthaceaeUQ.
- Anexo 2. Prueba de concepto para validar el modelo generado para la clasificación de especies vegetales. Base de datos de prueba Flavia.
- Anexo 3. Selección de clasificador con sus respectivos hiperparámetros.
- Anexo 4. Documento de Arquitectura de software SIAT.
- Anexo 5. Código fuente de la aplicación móvil.
- Anexo 6. Código fuente del servidor.