Estimación de la Calidad del Café a Partir de la Caracterización de los Defectos en el Grano Verde Utilizando Técnicas de Procesamiento de Imágenes

Trabajo de Grado
Modalidad: Trabajo de investigación

Jonathan Aldana Cristancho
100614010571
Luis Felipe López Pardo
100615011380

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Investigación de aplicaciones y servicios sobre internet.
Popayán - Octubre de 2022

Estimación de la Calidad del Café a Partir de la Caracterización de los Defectos en el Grano Verde Utilizando Técnicas de Procesamiento de Imágenes

Trabajo para optar por el título de Ingeniero Electrónico y de telecomunicaciones.

Trabajo de Grado
Modalidad: Trabajo de investigación

Jonathan Aldana Cristancho
100614010571
Luis Felipe López Pardo
100615011380

Director: PhD. Cristhian Nicolás Figueroa Martínez
Codirector: PhD. Juan Carlos Corrales
Asesor: MSc. Juan Fernando Casanova Olaya

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Departamento de Telemática
Línea de Investigación de aplicaciones y servicios sobre internet.
Popayán - Octubre de 2022

Tabla de contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

1. Introducción

La determinación de la calidad del café depende en gran medida del conocimiento que tienen los productores sobre la selección de las buenas cerezas de café, y de las mejores prácticas que les permita realizar un buen proceso de beneficio para obtener excelentes granos de café [1][2]. De este modo, permite que la taza de café producida pueda tener un aroma, una textura y un cuerpo adecuado [3][4]. No obstante, estos procesos requieren tiempo y conocimiento que en algunos casos los productores no disponen [5][6]. Esto ocasiona que en el momento de la venta a los intermediarios o cooperativas, el café no cumpla con los requisitos adecuados de calidad [7].

Por lo anterior, es necesario que los productores comprendan y apliquen buenas prácticas agrícolas para asegurar la calidad en los granos de café [8]. Para poder estimar la calidad del grano de café, un experto debe analizar y detectar visualmente los granos con defectos de una muestra según su criterio. Este proceso es subjetivo debido a su juicio y percepción, ya sea por la influencia de factores externos o cansancio por largas jornadas de trabajo del experto [9]. Por este motivo, en los estudios encontrados [10-19] realizan esta clasificación usando técnicas de procesamiento digital de imágenes y algoritmos de clasificación, con la limitante de que la mayoría, no realizan una estimación de la calidad a los granos de café clasificados.

En este sentido, cobra relevancia la clasificación de los granos de café haciendo empleo de imágenes y teniendo en cuenta el desarrollo de herramientas o procesos tecnológicos que faciliten la estimación de la calidad de una muestra de granos de café verde. Por consiguiente, el presente trabajo de grado se enfocó en el procesamiento de imágenes de granos de café con el fin de discriminar aquellos granos en buen estado de aquellos que poseían defectos. Para esto se realizó un sistema de clasificación automática de imágenes utilizando el marco de trabajo CRISP-DM para el análisis de datos y las fases de procesamiento digital de imágenes. En la primera fase de este marco, la adquisición de imágenes, obtuvimos un conjunto de imágenes de otro estudio al cual le adicionamos imágenes propias, que fueron tomadas en un entorno controlado para enriquecerlo. En la segunda fase, preprocesamiento, normalizamos las imágenes mediante el algoritmo White Patch, para mejorar la iluminación en las imágenes. En la tercera fase,

segmentación, implementamos el algoritmo Otsu a fin de extraer la imagen del grano del fondo de la fotografía. En la cuarta fase, extracción de características, seleccionamos las características más acordes a lo implementado, las cuales son color y textura. Y en la fase final, clasificación de patrones, entrenamos los modelos de clasificación SVM, KNN, Árboles de decisión y Bosque aleatorio haciendo uso del conjunto de datos construido. Posteriormente, elegimos el modelo de clasificación SVM por tener una precisión del 95,789% para identificar granos buenos y por tener una exhaustividad de 99.242% para los granos malos, evitando que estos fueran clasificados como buenos. Con este resultado, se logra estimar la calidad del café a partir de la identificación de los granos buenos presentes en una muestra de café.

Así, el presente trabajo de grado da respuesta a la siguiente pregunta de investigación:

¿Cómo estimar la calidad de los granos de café verde a partir de los defectos, mediante el uso de técnicas de procesamiento de imágenes y algoritmos de clasificación?

1.1. Objetivos

1.1.1. Objetivo general

- Estimar la calidad del grano de café verde a partir de sus defectos haciendo uso de técnicas de procesamiento de imágenes y algoritmos de clasificación

1.1.2. Objetivos específicos

- Caracterizar los defectos físicos del grano de café verde usando técnicas de procesamiento de imágenes
- Determinar el grado de calidad del grano de café verde a partir de la clasificación de los defectos caracterizados
- Evaluar los resultados obtenidos a través de la estimación de la calidad del grano de café verde

1.2. Contribuciones

- Se compartirá el conocimiento adquirido a través de un artículo en conjunto con lo desarrollado para que pueda ser replicado.

- Se compartirá el conjunto de imágenes propias construido para que pueda ser utilizado como base de otros estudios.
- Recomendaciones que permitirán la continuación del trabajo, relacionados con la mejora del conjunto de datos y el empleo de algoritmos más robustos.

1.3. Estructura del documento

El resto de este documento está dividido en los capítulos que se describen a continuación.

- Capítulo 2: **Estado actual del conocimiento**, presenta la información sobre los temas más relevantes del trabajo de grado. Estos temas incluyen el marco teórico, las brechas encontradas y los materiales y métodos empleados.
- Capítulo 3: **Metodología**, presenta el desarrollo de la propuesta de este trabajo de grado haciendo uso de WBS y CRISP-DM.
- Capítulo 4: **Evaluación y grado de calidad**, presenta la evaluación de los modelos y los resultados obtenidos en la estimación del grado de calidad.
- Capítulo 5: **Discusión**, presenta las conclusiones y trabajos futuros.

Capítulo 2

2. Estado actual del conocimiento

2.1. Marco teórico

Esta sección introduce temas asociados a la calidad del café, pasando por las fases del procesamiento digital de imágenes y finalizando con el entrenamiento y evaluación de los modelos de clasificación.

2.1.1. Café, defectos y calidad

Comúnmente la venta de café es realizada por medio de las cooperativas, que son el centro dende los caficultores llevan el café cosechado. Las cooperativas en dependencia de la calidad del producido y precio del mercado definen el valor a pagar a los caficultores. Por esta razón, los caficultores buscan mejorar los procesos de la producción del café con el fin de mejorar la calidad y obtener un mejor precio por su café.

En la actualidad existen buenas prácticas en el proceso productivo y en el beneficio del café. En el primer caso, los caficultores deben sembrar las mejores semillas, tener todos los cuidados adecuados y cosechar las cerezas maduras. En el otro, deben llevar controles en cada etapa de la postcosecha para asegurar que los granos de café que producen sean de alta calidad, obteniendo así la posibilidad de venderlos como café especial [20]. En la Figura 2.1 se observa el proceso de cosecha o recolección de los frutos del café.

Figura 2.1. Caficultor realizando su labor. Tomada de [15].

Normalmente, al interior de las cooperativas trabajan expertos en el control de la calidad del café, que son quienes verifican que los granos de café cumplan con los estándares adecuados. Este control se puede resumir en 6 pasos:

1. Determinar una muestra de los sacos de café.
2. Revisar que los granos de café no presenten olores desagradables.
3. Medir el porcentaje de humedad en los granos.
4. Comprobar el tamaño de los granos con mallas especiales.
5. Detectar los defectos en los granos de café visualmente.
6. Catación de la taza de café a partir de los granos tostados.

El análisis de estos factores es clave para determinar la calidad del café. El presente estudio está enfocado en el paso 5. El proceso de detección de defectos en los granos de café verde es complejo y requiere realizarse grano a grano, por lo cual es dispendioso para los agricultores, quienes muchas veces no cuentan con el tiempo y el conocimiento necesario para realizarlo [5][6]. En cuanto a los expertos, el proceso de detección de defectos puede ser muy subjetivo debido a su juicio y percepción, ya sea por la influencia de factores externos o cansancio por largas jornadas de trabajo [9]. Por esta razón, este trabajo de investigación tiene como objetivo plantar las bases de una herramienta de apoyo que ayude a los caficultores y expertos a estimar la calidad de los granos del café.

2.1.2. Procesamiento digital de imágenes

La detección de defectos en el grano de café es un proceso visual, de ahí que cobre importancia el uso de técnicas de procesamiento de imágenes. Estas técnicas pueden ser aplicadas a una imagen para resaltar características de interés y/o extraer información representativa de un objeto o zona de interés. Para esto implementan algoritmos que se aplican a una imagen para disminuir el ruido, mejorar el contraste, rotar, extraer regiones de interés, entre otras [22]. La Figura 2.2 describe las fases llevadas a cabo en el procesamiento digital de imágenes.

Figura 2.2. Fases resumidas del procesamiento digital de imágenes. Adaptada de [22]

Adquisición de imágenes

Esta fase consiste en la captura de imágenes a través de dispositivos ópticos encargados de transformar los objetos presentes en una escena a información digital o píxeles, que pueden ser almacenados en un ordenador.

Preprocesamiento

Esta fase analiza si las imágenes obtenidas en la etapa de adquisición requieren algún tipo de ajuste, como estandarizar los tamaños de las imágenes, y reparar algunos daños en la información como el ruido, el desenfoque, la mala iluminación, el movimiento, etc.

Segmentación

En esta fase son aplicadas diferentes técnicas para separar los elementos de interés del resto del contenido de una imagen. A nivel general, esta separación puede lograrse de 2 formas, una es identificando los cambios abruptos en la intensidad de los píxeles de la imagen, y la segunda es fijando un criterio para obtener regiones de píxeles con características similares.

Extracción de características

Esta fase tiene como objetivo describir sensaciones visuales que un ser humano podría tener al observar un objeto. En el procesamiento digital de imágenes puede traducirse como los métodos que describen cuantitativamente los atributos de los elementos que conforman una imagen.

Clasificación de patrones

Esta fase organiza o agrupa objetos con características en común. Existen algoritmos de clasificación permiten automatizar dicho proceso sin la intervención humana. Entre las técnicas más utilizadas se encuentran los algoritmos de aprendizaje supervisado como regresión y clasificación, no supervisado como *clustering* o agrupación, y aprendizaje por refuerzo [23].

2.1.3. Entrenamiento y evaluación

División entrenamiento/prueba o *Train-Test Split*

La división entrenamiento/prueba es una técnica para evaluar el rendimiento de un algoritmo de aprendizaje automático. Este puede ser empleado para problemas de

clasificación o regresión y puede ser empleado para cualquier algoritmo de aprendizaje supervisado.

El procedimiento consiste en tomar un conjunto de datos y dividirlo en dos subconjuntos. El primer subconjunto, conocido como “conjunto de datos de entrenamiento”, es usado para entrenar el modelo. El segundo subconjunto, es conocido como “conjunto de datos de prueba”, es utilizado con datos nunca vistos por el modelo, para luego realizar predicciones y comparar con los valores esperados.

- ➔ **Búsqueda de cuadrícula o *Grid search* [24]**

El rendimiento de un modelo depende significativamente de los valores de sus hiperparámetros. La búsqueda de cuadrícula es una herramienta para ajustar estos valores y encontrar un modelo óptimo para un algoritmo. Hay que tener en cuenta que, no hay forma de saber de antemano los mejores valores para los hiperparámetros. En el peor de los casos, debemos probar todos los posibles para conocer los valores óptimos. Por lo tanto, empleamos esta herramienta para automatizarlo.

El método consta de pasar una lista de valores predefinidos de hiperparámetros necesarios al clasificador. Este itera sobre la lista, entrenando el modelo y seleccionando los hiperparámetros con los cuales obtendría el mejor rendimiento.

Adicionalmente, este método usa *Cross validation* o validación cruzada [20], el cual es un método que permite validar el rendimiento de un modelo de aprendizaje automático al entrenarlo. El proceso consiste en realizar una división entrenamiento/prueba y usar los datos de entrenamiento para conformar subconjuntos llamados pliegues o *folds,* dejando un pliegue para prueba y el resto para entrenamiento. La Figura 2.3, ejemplifica una validación cruzada con 5 pliegues.

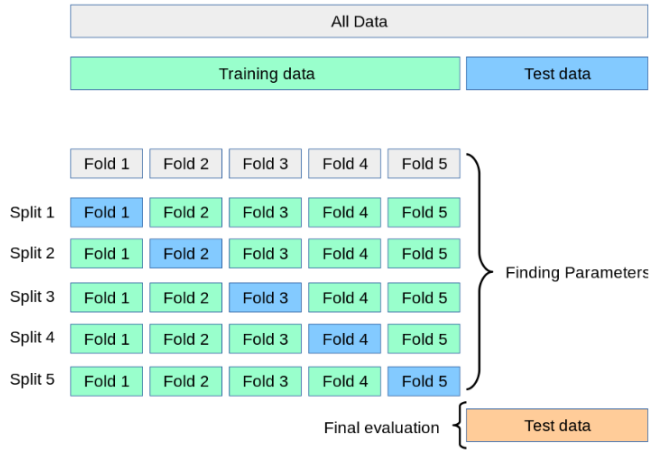


Figura 2.3. Ejemplo de *cross-validation* con 5 folds o pliegues. Tomada de [24].

Matriz de confusión [26]

La matriz de confusión es una herramienta utilizada para inspeccionar y evaluar visualmente el rendimiento de un algoritmo de clasificación. Calcular esta matriz permite tener una mejor idea de lo que el modelo de clasificación está haciendo bien y qué errores está cometiendo.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos positivos	Falsos negativos
	Negativos	Falsos positivos	Verdaderos negativos

Tabla 2.1. Matriz de confusión. Adaptado de [26].

La Tabla 2.1 describe los términos empleados en la matriz de confusión y cada uno se explica a continuación:

Verdaderos positivos (VP): Representan a las instancias que el modelo predice como positivas y que efectivamente son positivas. Por ejemplo, si la clase positiva son los granos buenos, esto representa cuando un grano bueno es clasificado como bueno.

Falsos positivos (FP): Son las instancias que el modelo predice como positivas cuando en realidad son negativas. Por ejemplo, si la clase negativa son los granos malos, entonces esto significa que un grano malo es clasificado como bueno.

Falsos negativos (FN): Ocurren cuando un modelo predice una observación como negativa, pero en realidad era positiva. Por ejemplo, cuando el grano bueno es clasificado como malo.

Verdaderos negativos (VN): Se presentan cuando un modelo predice una observación negativa y realmente lo es. Por ejemplo, el grano malo que es clasificado como malo.

Métricas de evaluación [27]

A partir de las matrices de confusión es posible calcular métricas que permitan evaluar que tan buenos son los modelos de clasificación. Entre todas las métricas existentes, las más comunes de calcular son la exactitud, la precisión, la exhaustividad y la puntuación F1.

Exactitud o *Accuracy*: Es una métrica que define como es la proporción de verdaderos positivos y verdaderos negativos con respecto a todas las observaciones. En otras palabras, la exactitud dice con qué frecuencia podemos esperar que el modelo predice correctamente un resultado del número total de veces que hizo predicciones. Esta es una buena métrica cuando el conjunto de datos está balanceado. Su fórmula es:

$$exactitud = \frac{VP + VN}{VP + FN + VN + FP}$$

Precisión o *Precision*: Mide la proporción de verdaderos positivos que son realmente correctos. La precisión puede considerarse como una medida de calidad y se emplea junto a la exhaustividad para compensar falsos positivos y falsos negativos. Si necesitamos minimizar los falsos positivos, elegimos un modelo con alta precisión. Su fórmula es:

$$precisión = \frac{VP}{VP + FP}$$

Exhaustividad o *Recall*: Representa la capacidad del modelo para predecir correctamente los aspectos positivos de los positivos reales. Es diferente de la precisión, ya que mide cuántas predicciones hechas por los modelos son realmente positivas de todas las predicciones positivas realizadas. En otras palabras, mide

qué tan bueno es nuestro modelo de aprendizaje automático para identificar todos los aspectos positivos reales de todos los positivos que existen dentro de un conjunto de datos. Si necesitamos minimizar los falsos negativos, elegiríamos un modelo con alta exhaustividad. Su fórmula es:

$$exhaustividad = \frac{VP}{VP + FN}$$

Puntuación F1 o *F1-score*: Representa la armónica entre la precisión y el *recall*. A menudo es utilizada como un valor único que proporciona información de alto nivel sobre la calidad del modelo. Esta es una medida útil en los escenarios en los que se intenta optimizar la precisión y/o la exhaustividad. Su fórmula es:

$$Puntuación\ F1 = 2 * \frac{precision\ *\ recall}{precision\ +\ recall} = \frac{VP}{VP\ +\ \frac{1}{2}(FP\ +\ FN\ )}$$

2.2. Revisión de literatura (SLR)

2.2.1. Revisión sistemática de literatura

El presente trabajo estudia el estado del arte sobre la detección de defectos en granos de café verde. Siguiendo la guía establecida por Kitchenham y Charters [28] para la construcción de una RSL. Esta busca obtener información de las técnicas y tecnologías empleadas para encontrar las brechas presentes en un campo de investigación. En la Figura 2.4 presentamos el protocolo seguido.

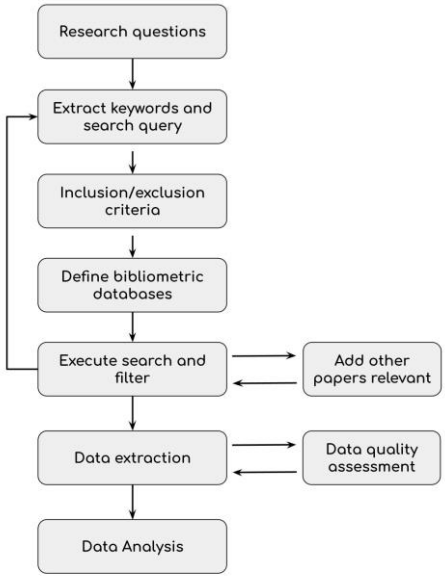


Figura 2.4. Diagrama de revisión sistemática de literatura. Creada a partir de [28].

Construcción de las preguntas de investigación

Iniciamos con la construcción de las preguntas de investigación para comprender y obtener información sobre como determinar la calidad física de los granos de café verde. Para lograr este objetivo, planteamos las siguientes preguntas:

- ➔ **RQ1** ¿Qué estudios determinan la calidad de los granos de café verde después de la clasificación?
- ➔ **RQ2** ¿Qué características de los granos de café verde se utilizan para la clasificación?
- ➔ **RQ3** ¿Hay disponibles conjuntos de datos de imágenes de granos de café verde?
- ➔ **RQ4** ¿Qué técnicas de procesamiento de imágenes se utilizan para la clasificación de los granos de café verde?
- ➔ **RQ5** ¿Cuáles son los algoritmos de clasificación más utilizados en la clasificación de los granos de café verde?
- ➔ **RQ6** ¿Qué brechas existen en los estudios en la clasificación de los granos de café verde?

Definición de las palabras clave y cadena de búsqueda

La búsqueda se realizó a partir de conceptos clave, omitiendo tecnologías para evitar limitaciones de los resultados. Se decidió colocar el término base seria “Green Coffee Beans” por qué es el tema principal de esta investigación y así acotar los estudios relacionados. La cadena de búsqueda definida es la siguiente:

green coffee bean AND (quality OR defect) AND (grading OR sort OR classif*)*

Criterios de inclusión y exclusión

Definimos un conjunto de criterios de inclusión y exclusión para determinar si un estudio es de calidad. Planteamos que los estudios deberían ser a partir del año 2000, ya que en este año salió la primera versión de *OpenC*V, herramienta ampliamente empleada para trabajar visión por computadora. En la Tabla 2.2 resumimos los criterios usados.

Criterio de inclusión
- Artículos de conferencias y revistas. - Artículos publicados entre los años 2000 y 2021. Primera - versión de OpenCV en 1999. - Artículos que clasifiquen al menos un módulo para la clasificación de granos de café verde. - Artículos que clasifican al menos un tipo de granos de café verde.
Criterio de exclusión
- Literatura gris. - Artículos que abordan las cerezas de café. - Artículos que abordan las enfermedades en los granos de café. - Artículos que no abordan procesamiento de imágenes o aprendizaje de máquina para clasificar granos de café verde. - Artículos que usen métodos invasivos o destructivos. - Fuentes secundarias - Informes breves

Tabla 2.2. Criterios de inclusión y exclusión. Fuente propia.

Selección de fuentes bibliográficas

Definimos las 4 fuentes bibliográficas de la Tabla 2.3, ya que son las más utilizadas para encontrar estudios relacionados con tecnología y diversos campos de aplicación.

Fuente bibliográfica	URL
Google Scholar	https://scholar.google.com
ISI Web of Science	http://www.isiknowledge.com
Science Direct	http://www.sciencedirect.com
Scopus	http://www.scopus.com

Tabla 2.3. Fuentes bibliográficas empleadas. Fuente propia.

Ejecución de búsqueda y filtrado

Realizamos la búsqueda de la cadena de búsqueda en cada una de las fuentes bibliográficas y encontramos los artículos listados en la Tabla 2.4.

Fuente bibliográfica	Número de artículos
Google Scholar	12
ISI Web of Science	24
Science Direct	0
Scopus	207
Total	**243**

Tabla 2.4. Resumen del número de artículos encontrados por fuente bibliográfica. Fuente propia.

Posteriormente, realizamos un filtraje a través de la lectura rápida del título, resumen y contenido en general de los 243 artículos. De los cuales encontramos los artículos duplicados, rechazados y aceptados según los criterios de inclusión y exclusión plantados. Los resumimos en la Tabla 2.5.

Clasificación	Número de artículos
Duplicados	11
Rechazados	216
Aceptados	16

Tabla 2.5. Resumen después del filtraje de los artículos encontrados. Fuente propia.

Para acotar aún más los estudios sin perder los de calidad, definimos una serie de criterios de calidad que enumeramos en la Tabla 2.6. Cinco preguntas donde es puntuada con los valores 1, 0,5 y 0, para representar las respuestas “Sí”, “Parcialmente” y “No”. El valor máximo que un artículo puede ser de 5 y el umbral

de selección lo hemos dejado en 2.9, los artículos que pasen el umbral serán seleccionados para esta investigación.

Pregunta	Puntaje
Q1. ¿Realiza una revisión comprensiva del estado actual del conocimiento o trabajos relacionados existentes?	Si / Parcialmente / No (1 / 0.5 / 0)
Q2. ¿Presenta un diagrama de los principales componentes arquitectónicos?	Si / Parcialmente / No (1 / 0.5 / 0)
Q3. ¿Muestra y describe los algoritmos o pseudocódigos utilizados o por lo menos indica claramente las técnicas utilizadas?	Si / Parcialmente / No (1 / 0.5 / 0)
Q4. ¿Analiza los principales hallazgos y resultados empíricos obtenidos?	Si / Parcialmente / No (1 / 0.5 / 0)
Q5. ¿Describe claramente la evaluación realizada de los resultados obtenidos?	Si / Parcialmente / No (1 / 0.5 / 0)

Tabla 2.6. Criterios de calidad. Fuente propia.

Después todo este proceso de filtraje obtuvimos un total de 10 artículos, los cuales utilizamos para esta investigación.

Extracción de datos

Después de haber seleccionado los artículos, son leídos con mayor detalle para extraer los siguientes metadatos y obtener información clave para plantear las brechas en el campo.

- **Re:** Referencia
- **AP:** Año de publicación
- **GC:** Grupo de clasificación
- **MG:** Muestras de granos de café
- **CU:** Características utilizadas
- **De:** Defectos
- **TP:** Técnicas de procesamiento digital de imágenes
- **AC:** Algoritmos de clasificación
- **Ca:** Calidad

La Tabla 2.7, muestra un resumen de la información principal de cada artículo.

Re	AP	GC	MG	CU	De	TP	AC	Ca
[10]	2020	-	2000	-	2	-	-	✗
[11]	2019	Binario	100	Color	2	- Filtro Gausiano - Otsu	- Relación de áreas	✗
[12]	2019	Binario	1103	Color Forma Textura	2	- Otsu	- Red Neuronal	✗
[13]	2020	Binario	72000	Color	2	- Detección de color - Aumento de datos	- Red Neuronal Convolucional	✗
[14]	2019	Binario	220	Color Forma	2	-	- k vecinos más cercanos	✓
[15]	2017	Multiclase	1584	Color Textura	12	- White patch - Otsu	- Máquina de vectores de soporte	✗
[16]	2016	Multiclase	340	Color Forma Textura	4	- Segmentación - Detección de bordes	- Red neuronal - Árboles de decisión - Clasificador Bayesiano - Máquina de vectores de soporte	✗
[17]	2017	Multiclase	1930	Textura	12	- White patch - Otsu	- Máquina de vectores de soporte	✗
[18]	2017	Multiclase	13000	Color	6	- Redimensión de imagen	- Red Neuronal Convolucional - Bosques aleatorios	✗
[19]	2020	Multiclase	635	Color Forma Textura	4	- Filtro Gausiano - Otsu	- Máquina de vectores de soporte - Red neuronal profunda	✗

Tabla 2.7. Resumen de los metadatos de los artículos relacionados. Fuente propia.

Síntesis

Puede destacarse que, los artículos pueden agruparse en clasificación binaria, donde un grano puede ser bueno o malo sin importar el defecto; y multiclase, donde se determina el defecto presente en el grano a evaluar. En el grupo de clasificación binaria, en [14] proponen una manera para obtener el grado de calidad de un grano, pero agrupando los defectos como única clase, lo que lleva a la pérdida de información y no llevar un control de calidad adecuado, como lo mencionan [15][16].

En el grupo de clasificación multiclase, ninguno de los artículos sugiere una forma para estimar la calidad, pero en [17] los autores brindan información a los expertos sobre la clasificación realizada. Con base en lo anterior, nuestra propuesta aportará en la detección de distintos defectos para estimar el grado de calidad en los granos de café y a través de un umbral.

- ➔ **Clasificación binaria**

En el trabajo de E. R. Arboleda et al. [10], recomiendan un algoritmo para detectar el borde de los granos de café en imágenes, con el fin de obtener mejores resultados en el procesamiento de imágenes. El algoritmo consiste en tomar un píxel y analizar los píxeles adyacentes para detectar si hay cambios bruscos entre dichos valores, lo que se consideraría un borde. Adicionalmente, para probar que este método funciona, obtuvieron características morfológicas a partir de 2000 muestras de granos de café y compararon su algoritmo con los algoritmos más usados en la detección de bordes como Sobel, Prewitt y Roberts, obteniendo como resultado una mejora en cuanto al comportamiento de su algoritmo respecto a los algoritmos comparados por los autores.

En el trabajo de A. F. Sánchez et al. [11], realizan una clasificación binaria de 100 granos de café verde etiquetados, en una proporción 50:50, como granos buenos o malos, según el color y la forma. Llevan a cabo la adquisición de imágenes mediante un entorno controlado con buena iluminación, fondo uniforme y una cámara de buena resolución. A las imágenes obtenidas se les aplica un filtro Gaussiano para la reducción del ruido. Para obtener el área total del grano utilizan el canal H del espacio de color HSV y para determinar el área del defecto en el grano emplean el canal Cr del espacio de color YCbCr. La clasificación la efectúan por medio de una relación de áreas, entre el área defectuosa y el área total del grano.

En el trabajo de J. P. L. Pizzaia et al. [12], realizan una propuesta para clasificar granos sanos y defectuosos de café arábica con una red neuronal perceptrón multicapa (MLP). Esta consiste en adquirir imágenes de un conjunto de granos de café clasificados por expertos, realizar el proceso de segmentación y hacer la extracción de características, como el área, la redondez y el promedio de cada uno de los canales RGB. Para obtener los datos que la red neuronal necesita para la clasificación. Los autores obtuvieron un buen resultado, excepto por los granos con defecto de broca, que no fue posible detectar porque no había una característica que pudiera diferenciar los pequeños huecos en los granos.

El trabajo de N. F. Huang et al. [13] presenta la implementación de un sistema en tiempo real con una cámara web conectada a un servidor GNU/Linux para clasificar

granos de café sanos y defectuosos mediante una red neuronal convolucional (CNN). Como datos iniciales fotografiaron un total de 1000 granos buenos y 1000 granos malos, pero al ser insuficientes optaron por aumentar el número de datos con las técnicas de girar y voltear las imágenes originales, donde obtuvieron un nuevo conjunto de imágenes 36 veces mayor, con un total de 72000 granos de café verde. Cada imagen fue convertida a escala de grises para realizar la extracción del grano a fin de usarlas en el entrenamiento y pruebas de la red neuronal.

En el trabajo de M. García et al. [14], los autores realizan la clasificación de granos de café verde utilizando 4 características de forma y el algoritmo k-nearest neighbors (KNN) para agrupar los granos como buenos y malos, con el fin de obtener el porcentaje estimado de calidad de un grano seleccionado. Para esto emplearon 444 granos de café para la construcción de un conjunto de 220 imágenes, de las cuales 100 imágenes tenían un solo grano con la etiqueta del defecto correspondiente para entrenar el modelo y el restante, grupos de 4 y 10 granos para comprobar la predicción del modelo. A cada imagen obtenida, le aplicaron un filtro para reducir el ruido y mejorar la nitidez de la imagen. Además, la imagen fue segmentada dos veces, la primera para obtener el grano del fondo y la segunda para identificar las manchas en la superficie de los granos, usando los espacios de color HSV y LUV.

- ➔ **Clasificación multiclase**

En el artículo de J. Ramírez-Ticona et al. [15], proponen la clasificación de 12 clases de defectos más la clase de grano sano implementando una máquina de soporte vectoriales (SVM) como clasificador multiclase. A diferencia de otros estudios, los autores realizan la adquisición de imágenes con un smartphone, tomando las fotografías en diagonal para evitar la sombra del dispositivo. Para el procesamiento de imágenes utilizan *White patch*, un filtro en el dominio del espacio que permite contrarrestar las variaciones de iluminación. Posteriormente, realizaron la segmentación de los granos y la extracción de características mediante los histogramas de intensidad, donde obtuvieron de cada uno de los canales RGB 256 intensidades, representando así los granos de café como un vector de 768 intensidades que emplearon como características para el entrenamiento del modelo.

En el trabajo de J. C. Borrero Becerra & C. A. Diaz Molano [16], generaron una base de fotografías con los principales defectos de los granos de café seco, en compañía de un experto calificado, con el objetivo de implementar 4 clasificadores diferentes, perceptrón multicapa, máquina de soporte vectoriales, clasificador bayesiano y

árbol de decisión. Cada imagen fue segmentada con el método que los autores proponen, ya que las técnicas existentes para la detección de bordes no daban los resultados esperados. Los autores extraen 63 características teóricas y sugieren otras 90, entre color, textura y forma, para verificar si son útiles para la clasificación. Dándose cuenta de la gran cantidad de características, los autores redujeron a un total de 18 teóricas y 27 propuestas.

En el trabajo realizado por C. E. Portugal-Zambrano et al. [17], desarrollaron un sistema de clasificación automática de 12 defectos en granos de café verde para apoyar la labor del experto. Esto lo hicieron implementando un sistema modular compuesto por el prototipo hardware, donde crearon un ambiente controlado para la adquisición de imágenes. En el módulo de preprocesamiento aplicaron un algoritmo de constancia de color, el módulo de segmentación para obtener las máscaras de los granos, el módulo de clasificación donde hicieron uso de máquina de soporte vectoriales (SVM) y un módulo software para adquirir los datos extraídos por los expertos para generar un reporte completo del control de calidad.

En el trabajo de C. Pinto et al. [18], estudiaron la clasificación de granos de café verde mediante una red neuronal convolucional (CNN), entrenada con un total de 13000 imágenes de 6500 granos de café verde por ambas caras, las imágenes fueron recortadas a un tamaño de 256x256 píxeles y etiquetadas manualmente dependiendo de su defecto. Los autores utilizan 5 defectos, negro, agrio, flotador, roto, peaberry y el grano sano. Los datos fueron separados en tres grupos. Datos de entrenamiento empleados para el aprendizaje de la red neuronal. Datos de validación validan la precisión de la clasificación en la fase de aprendizaje. Y datos de prueba usados para evaluar el rendimiento en la capacidad de clasificación.

En el trabajo F. Santos et al. [19], proponen una comparación de tres algoritmos para la clasificación de 4 defectos en granos de café, estos son, bosques aleatorios, máquina de soporte vectoriales y red neuronal profunda. Los autores emplean una muestra de 635 granos a fin de realizar la captura de imágenes con un escáner Epson L210 y posteriormente, extraen 7 características de color y 8 características de forma para entrenar los algoritmos. Al final, determinan que los tres clasificadores muestran un comportamiento similar.

Análisis de información

RQ1 ¿Qué estudios determinan la calidad de los granos de café verde después de la clasificación?

- ➔ En la revisión realizada solo se encuentra un estudio de clasificación binaria que propone una forma de obtener la calidad. Estos autores proponen obtener la calidad a partir del resultado de clasificar un grano de café con el algoritmo de *k vecinos más cercanos (KNN).* Su propuesta consiste en contar el número de vecinos de granos buenos implicados en la clasificación y luego es dividido por el valor de número total de vecinos tenidos en cuenta para la clasificación. [14] Por ejemplo, si 10 es el número de vecinos tenidos en cuenta para clasificar un grano de café cualquiera y entre esos vecinos hay 6 que corresponden a la clase de granos buenos, el porcentaje de calidad será definido al dividir 6 entre 10.

RQ2 ¿Qué características de los granos de café verde son utilizados para la clasificación?

- ➔ A partir del resumen presentado en la Tabla 2.7 las características de color, textura y forma son las utilizadas para el proceso de calificación. Pero de todos los estudios, solo 3 trabajos utilizan todas esas características a la vez. [12][16][19]

RQ3 ¿Hay disponibles conjuntos de datos de imágenes de granos de café verde?

- ➔ Ninguno de los trabajos referencia algún enlace o menciona repositorios donde compartan su conjunto de imágenes o datos. Ni siquiera el trabajo de J. C. Borrero Becerra y C. A. Diaz Molano que describe el proceso realizado para construir un conjunto de imágenes. [16]

RQ4 ¿Qué técnicas de procesamiento de imágenes se utilizan para la clasificación de los granos de café verde?

- ➔ En el caso de la etapa de preprocesamiento de imágenes encontramos que en algunos estudios es utilizado el Filtro Gaussiano para eliminación de ruido [11][19], el aumento de datos a fin de ampliar la información de un conjunto de imágenes [13], la mejora de la iluminación en una imagen mediante *White patch* [15][17] y la redimensión de las imágenes para reducir el tiempo de entrenamiento de una red neuronal convolucional. [18]
- ➔ En la etapa de segmentación, la mayoría de los estudios hacen uso del algoritmo de Otsu por su facilidad en obtener las máscaras binarias que ayudan a separar los granos de café del fondo de la imagen. [11][12][15][17][19]

- ➔ Las técnicas de las etapas de extracción de características y clasificación de patrones se responden en **RQ2** y **RQ5** respectivamente.

RQ5 ¿Cuáles son los algoritmos de clasificación más utilizados en la clasificación de los granos de café verde?

- ➔ Con base en los datos extraídos, la máquina de vectores de soporte comienza la lista de los algoritmos más utilizados [15][17][16][19], luego le siguen diferentes tipos de redes neuronales [12][13][16][18][19] y por último, ya aparecen algoritmos que solo son usados en un estudio como los modelos de árboles de decisión [16], bosques aleatorios [18] y el clasificador bayesiano [16].

RQ6 ¿Qué brechas existen en los estudios en la clasificación de los granos de café verde?

- ➔ En la sección siguiente se resumen las brechas encontradas en los estudios.

2.2.2. Brechas

A través del análisis de los artículos relacionados encontramos las siguientes brechas a abordar en esta investigación.

- ➔ **Falta de un conjunto de imágenes o datos abierto:** No fue posible encontrar un conjunto de datos o imágenes abierto que podamos emplear como base para este estudio.
- ➔ **Replicabilidad del estudio:** Como no se encontraron datos disponibles, no fue posible replicar los estudios a pesar de que explicaban técnicas de procesamiento digital de imágenes y/o los parámetros configurados de los algoritmos en otros estudios.
- ➔ **Falta de evaluación:** No se encontró en los estudios una fase de evaluación adicional donde se emplee imágenes diferentes tomadas del entorno controlado usado del estudio.
- ➔ **Estimación de la calidad:** En la mayoría de los artículos no se encontró que plantearan una forma de estimar la calidad de una muestra de granos de café verde.

2.3. Materiales y métodos

El objetivo de esta sección es presentar en detalle el hardware y el software utilizado para dejar la posibilidad de replicación de este estudio.

2.3.1. Hardware

Esta sección presenta cada una de las herramientas hardware utilizadas con su descripción detallada y la razón por la cual fue seleccionada.

Computador utilizado en el estudio

En primer lugar, el computador seleccionado para llevar a cabo todas las etapas del procesamiento digital de imágenes es el portátil Acer Nitro 5. Esta elección es realizada porque sus características son de generaciones más recientes a las del portátil Lenovo G40-70, que era otro computador contemplado para desarrollar el estudio. A continuación, son descritas las características del computador seleccionado y en la Figura 2.5 es presentada una fotografía del equipo.

Figura 2.5. Computador seleccionado. Fuente propia.

Características:

- ➔ **Modelo:** AN515-52
- ➔ **Sistema operativo:** Windows 10 Home
- ➔ **CPU:** Intel Core i5 Quad-core 2.30GHz 8300H (8.ª Generación)
- ➔ **GPU:** NVIDIA GeForce GTX 1050 4 GB
- ➔ **RAM:** 16 GB DDR4 SDRAM
- ➔ **Almacenamiento:** 1T SSD

Cámara digital para la adquisición de imágenes

Como en la etapa de adquisición de imágenes es necesario tomar fotografías de granos de café, la selección de una cámara toma relevancia. Para realizar esta selección, primero realizamos pruebas donde unos modelos de clasificación son entrenados con el conjunto de imágenes que nos comparten de otro estudio y luego

validamos que tan bien clasificaban con fotografías tomadas con un Smartphone Xiaomi Redmi Note 8 Pro. Los resultados de esas pruebas fueron un fracaso, entonces como es presentado en la Figura 2.6, revisamos los metadatos de las imágenes del conjunto que nos compartieron para tener fotografías similares a las que nos compartieron. Así encontramos que es necesario utilizar una cámara digital, ya que permiten configurar más parámetros que la cámara de un smartphone.

Cámara	
Fabricante de cámara	Canon
Modelo de cámara	Canon EOS REBEL T3
Punto F	f/8
Tiempo de exposición	1/160 s
Velocidad ISO	ISO-100
Compensación de exposición	0 paso
Distancia focal	36 mm
Apertura máxima	
Modo de medición	Diseño
Distancia al objeto	
Modo de flash	Sin flash, obligatorio
Intensidad de flash	
Longitud focal de 35 mm	

Figura 2.6. Metadatos de una de las imágenes del conjunto compartido. Fuente propia.

Por lo anterior, realizamos una búsqueda y conseguimos las cámaras Sony α6500 y Canon EOS Rebel T3. Aunque esta última fue la misma referencia de la cámara que usaron en el otro estudio para tomar las fotos del conjunto de imágenes, en pruebas realizadas la clasificación de los modelos mencionados anteriormente seguía fallando. Por lo que elegimos la cámara Sony α6500 para tomar las fotografías que se integrarían con el conjunto de datos compartido a fin de tener más diversidad de fotografías. En la Figura 2.7 y en la Figura 2.8 es presentada la fotografía de la cámara y los metadatos de las imágenes tomadas respectivamente.

Figura 2.7. Cámara digital seleccionada. Fuente propia.

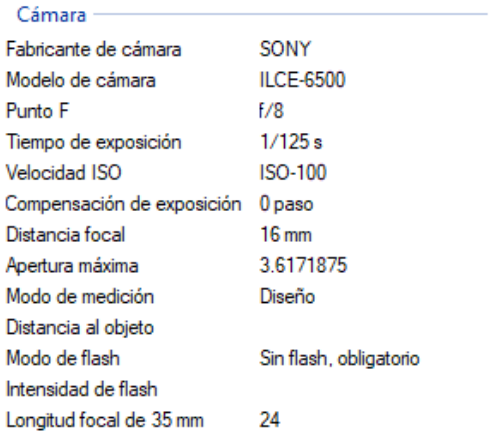


Figura 2.8. Metadatos de una de las imágenes del conjunto compartido. Fuente propia.

Entorno controlado

El último elemento hardware que es definido para usar en este estudio, es la construcción de un entorno controlado. Esta necesidad surge cuando se realizan pruebas al algoritmo de segmentación con fotografías tomadas por nosotros y este no logra separar bien los granos del fondo, como lo muestra la Figura 2.9.

Figura 2.9. Fallas de segmentación. Fuente propia.

Como estas fallas fueron ocasionadas por sombras de elementos externos, el entorno controlado ayuda a tener una iluminación más constante y a evitar dichas sombras. Para construirlo conseguimos una caja de cartón y la forramos con cartulina blanca por dentro. Después, como teníamos que seleccionar la fuente de luz que aportaría la iluminación constante, revisamos los estudios realizados en el estado del arte para conocer cómo se debía distribuir esa luz y encontramos que con una distribución circular se aseguraba una buena iluminación [14]. Por esta

razón, hacemos uso del aro de luz Phottix Nuada Ring 10 LED Go Kit, que nos asegura la iluminación constante en el entorno controlado. En la Figura 2.10 se observa el resultado final.

Figura 2.10. Entorno controlado ensamblado. Fuente propia.

2.3.2. Software

En esta sección se elige el lenguaje de programación para desarrollar las etapas del procesamiento digital de imágenes. Este lenguaje es Python, ya que por su amplia gama de librerías permite realizar todo lo que requiera cada etapa. Además, por ser también un lenguaje usado en el desarrollo aplicaciones web y de escritorio, permitiría integrar fácilmente todo este proceso en una interfaz. A continuación, son listadas las librerías con sus respectivas versiones utilizadas:

- **Entornos de ejecución:**
  - Versión del intérprete de Python - 3.8.13
  - Anaconda - 2.2.0
  - Jupyter Notebook - 6.4.12
- **Procesamiento de datos:**
  - OpenCV - 4.6.0
  - Numpy - 1.23.1
  - Pandas - 1.4.3
  - Scikit image - 0.19.3
  - Pillow - 9.2.0
  - Imageio - 2.21.0
- **Representación de gráficos:**
  - Matplotlib - 3.5.2

- Seaborn - 0.11.2

➔ **Aprendizaje automático:**

- Scikit learn - 1.1.1
- Joblib - 1.1.0
- Mahotas - 1.4.13

Capítulo 3

3. Metodología

3.1. WBS y CRISP-DM

Para el desarrollo del presente trabajo utilizamos la *Work Breakdown Structure* (*WBS*), el cual es una herramienta que permite detallar y estructurar las actividades que se van a desarrollar [29]. *WBS* define una estructura jerárquica por niveles en la cual el nivel inicial es la actividad más abstracta y los niveles subsiguientes equivalen a entregables simples para su ejecución. Además, empleamos *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) que es una metodología iterativa para describir el ciclo de vida de un proyecto orientado a la minería de datos. [31][33]

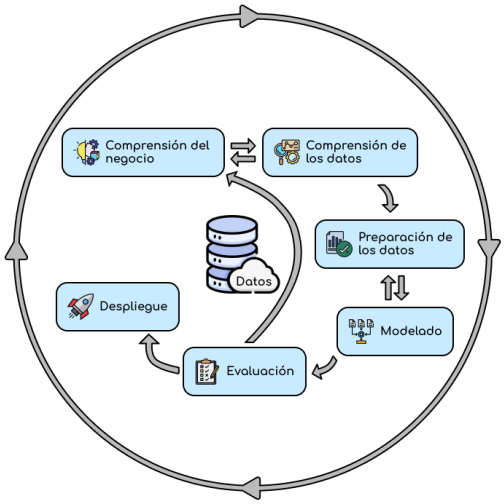


Figura 3.1. Diagrama de las fases de *CRISP-DM*. Imagen adaptada de [30].

Como se observa en la Figura 1, *CRISP-DM* consta de 6 fases:

- ➔ La 1° fase, **Compresión del negocio**, consiste en recopilar toda la información posible del tema a abordar y a partir de ese conocimiento, formular los objetivos o requerimientos del proyecto a desarrollar.

- ➔ La 2° fase, **Comprensión de los datos**, comprende la recolección, familiarización y descubrimiento de información con base en los conjuntos de datos relacionados con el proyecto.

- ➔ La 3° fase, **Preparación de los datos**, abarca las actividades de selección, limpieza y depuración de los datos más relevantes, para construir el conjunto de datos final.

- ➔ La 4° fase, **Modelado**, consiste en seleccionar las técnicas de modelado más acordes al proyecto, configurar los modelos con diversos parámetros y ejecutarlos para obtener los resultados.

- ➔ La 5° fase, **Evaluación**, es el análisis de los resultados para determinar si los modelos construidos cumplen con los objetivos formulados en la fase 1.

- ➔ La 6° fase, **Despliegue**, es la implementación de los modelos que superaron la fase 5 en un entorno productivo donde ya sean utilizados.

Las actividades de *WPS* por desarrollar, implementado *CRISP-DM*:

WP1. Generación de la base de conocimiento (Comprensión del negocio)

Este paquete tiene como objetivo la construcción del estado actual del conocimiento relacionado con la implementación de algoritmos para la clasificación de granos de café verde. El resultado de este paquete de trabajo es un documento con la construcción de la base del conocimiento.

- ➔ Revisión del estado actual del conocimiento
- ➔ Síntesis de la información
- ➔ Construcción de la base del conocimiento

WP2. Implementación de las etapas del procesamiento de imágenes (Comprensión y preparación de los datos)

En este paquete aplicamos las fases del procesamiento digital de imágenes en fotografías con granos de café verde. El resultado de este paquete es la construcción de un conjunto de datos con las características más relevantes de los defectos en los granos de café verde.

- ➔ Recopilación de imágenes de granos de café verde a través de un conjunto de datos existente o construcción de uno
- ➔ Extracción de características a partir de imágenes de granos de café verde
- ➔ Construcción de un conjunto de datos a partir de las características extraídas

WP3. Clasificación de granos de café verde según sus defectos (Modelado)

En este paquete realizamos el desarrollo y codificación de algoritmos para determinar los defectos en los granos de café verde. El resultado de este paquete es un algoritmo que clasifica granos de café según su defecto.

- ➔ Estudio de los algoritmos de clasificación
- ➔ Implementación de los algoritmos seleccionados
- ➔ Clasificación de los granos de café según su defecto
- ➔ Selección del algoritmo con el mejor resultado

WP4. Determinación de la calidad de los granos de café verde (Modelado)

En este paquete determinamos un método para obtener el grado de calidad de los granos de café verde. El resultado de este paquete es un indicador que estime la calidad del café clasificado.

- ➔ Definir la relación entre la calidad del café y los defectos del café
- ➔ Determinar un método para estimar la calidad del café clasificado

WP5. Pruebas y análisis de los resultados (Evaluación)

Este paquete tiene como finalidad evaluar la clasificación y estimación de la calidad de los granos de café verde. El resultado de este paquete es analizar el desempeño de la clasificación y determinación de la calidad.

- ➔ Evaluar los algoritmos de clasificación
- ➔ Evaluar la calidad de los granos según los resultados
- ➔ Análisis de resultados

WP6. Divulgación

Este paquete tiene como finalidad efectuar el conglomerado de cada una de las fases del trabajo de grado. El resultado de este paquete es la elaboración de un artículo y un documento final.

- ➔ Elaboración del artículo
- ➔ Elaboración del documento final

3.2. Implementación

Esta sección presenta el desarrollo de las actividades planteadas en el *WBS*, distribuidas en las fases definidas en *CRISP-DM*. Primero, la comprensión del negocio resume el proceso llevado a cabo para generar la base de conocimiento. Segundo, la comprensión de los datos muestra el proceso de recolección y familiarización del conjunto de imágenes. Tercero, la preparación de los datos para construir el conjunto de datos definitivo. Cuarto, el proceso de modelado para la selección y entrenamiento de los algoritmos más acordes al proyecto. La fase de evaluación se describe en el capítulo 4.

3.2.1. Comprensión del negocio

El desarrollo de esta fase comienza cuando somos motivados por la situación de amigos y conocidos donde su principal fuente de ingresos son los cultivos de café, por lo que empezamos a investigar en cuál de las etapas del proceso productivo podíamos realizar un aporte. De este modo, definimos como tema de trabajo la identificación de los defectos del café, realizamos una revisión en la literatura para conocer las brechas de los estudios relacionados, planteamos los objetivos de nuestro trabajo y construimos la base de conocimiento que nos permitió orientar este trabajo de grado.

3.2.2. Comprensión de los datos

Esta fase describe como realizamos la adquisición de imágenes descrita en la primera etapa del procesamiento digital de imágenes, esto consolidó el conjunto de imágenes usado en este trabajo de grado y la familiarización de ese conjunto de imágenes para identificar información relevante a tener en cuenta.

Recolección de imágenes

La recolección de imágenes comenzó cuando realizamos una búsqueda de un conjunto de datos o imágenes de granos de café verde para uso público, pero como no encontramos, decidimos listar los correos electrónicos de los autores de los trabajos que seleccionamos en la revisión sistemática de la literatura. Por lo tanto, enviamos un correo electrónico a cada uno de ellos con la intención de que nos brindaran información sobre los conjuntos de imágenes que utilizaron, logrando así conseguir el conjunto de imágenes usado en el estudio realizado por C. E. Portugal-Zambrano et al. [17]

Como necesitábamos comprobar si ese conjunto de imágenes era suficiente para nuestro trabajo, realizamos una serie de pruebas donde usamos ese conjunto de imágenes para entrenar unos algoritmos de clasificación y así validar su efectividad mediante unas fotografías de granos de café verde sanos que conseguimos. El resultado de las pruebas fue un fracaso, ya que los modelos no lograban identificar que las fotografías eran granos sanos; esto nos permitió comprender que había una limitación con la cantidad de información que teníamos para entrenar los modelos. Por este motivo, decidimos enriquecer con otras fotografías el conjunto de imágenes que nos habían compartido.

Los pasos que llevamos a cabo para enriquecer el conjunto de imágenes comienzan con la recolección de muestras de granos de café, la cual consistió en comunicarnos con nuestros amigos y conocidos con fincas cafeteras para que nos compartieran muestras de sus granos de café. Cuando terminamos el proceso de recolección, hablamos con un catador de café que nos ayudó a separar los granos sanos de los defectuosos y a etiquetar cada uno de los defectos hallados en las muestras. Luego, utilizamos la cámara Sony α6500 y el entorno controlado a fin de realizar la etapa de adquisición de imágenes, tomando fotografías de las muestras organizadas por el catador. Por último, combinamos nuestras fotografías con las compartidas por los autores del otro estudio, consolidando así el conjunto de imágenes definitivo para este trabajo de grado. En la Figura 3.2 se puede observar una parte del conjunto de imágenes.

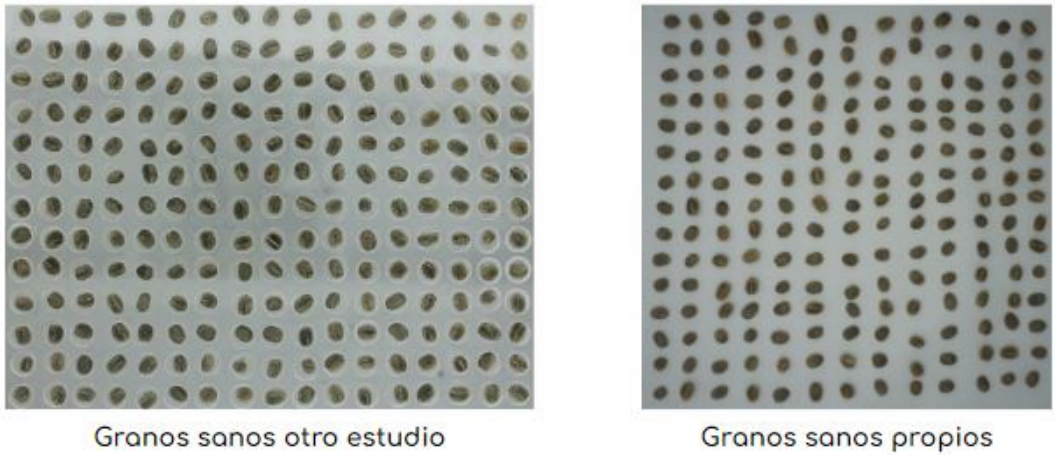


Figura 3.2. Granos sanos del conjunto de imágenes definitivo. Fuente propia.

Familiarización del conjunto de imágenes

Esta parte del trabajo fue muy importante para conocer más sobre el conjunto de imágenes. En primer lugar, revisamos cada una de las imágenes de las muestras

entre granos sanos y defectuosos para conocer más sobre sus diferencias visuales. Y en segundo lugar, agrupamos los granos sanos en la clase buenos y todos los granos defectuosos en la clase malos, donde posteriormente realizamos un conteo total de las muestras disponibles. En la Figura 3.3 se observa la agrupación realizada y las diferencias entre los granos sanos y los defectuosos. Por ejemplo, los granos con defecto partido, mordido, cortado, concha y brocado severo tienen diferencias de textura y morfología con los granos malos.

Figura 3.3. Muestras de granos sanos y defectuosos. Fuente propia.

A partir de la agrupación presentada anteriormente, el conjunto de imágenes usado en este estudio queda conformado con la información de 2654 granos de café verde. De los cuales, 2012 granos son del conjunto de imágenes compartido por C. E. Portugal-Zambrano et al. [17] y 642 granos son las imágenes tomadas por nosotros. En la Tabla 3.1 y la Tabla 3.2 se describe la cantidad de granos sanos y granos malos respectivamente.

Clase	**Granos sanos**	221
	Granos malos	1791
Total de granos		2012

Tabla 3.1. Distribución entre granos sanos y malos de las imágenes de C. E. Portugal-Zambrano et al. [17]. Fuente propia.

Clase	**Granos sanos**	255
	Granos malos	387
Total de granos		642

Tabla 3.2. Distribución entre granos sanos y malos de las imágenes tomadas por nosotros. Fuente propia.

3.2.3. Preparación de los datos

La preparación de los datos representa el proceso para construir el conjunto de datos final, esto fue llevado a cabo aplicando otras etapas del procesamiento digital de imágenes como el preprocesamiento, la segmentación y la extracción de los atributos o características del conjunto de imágenes. El resultado de este proceso es el conjunto de datos para entrenar los algoritmos en la etapa de modelado.

Preprocesamiento

Teniendo en cuenta los problemas de segmentación que tuvimos, en esta etapa implementamos el algoritmo de *White Patch* [36][37] para realizar un balance de blancos en cada una de las fotografías del conjunto de imágenes. De este modo, los granos de café se diferenciaban mejor del fondo de la imagen, como se muestra en la Figura 3.4.

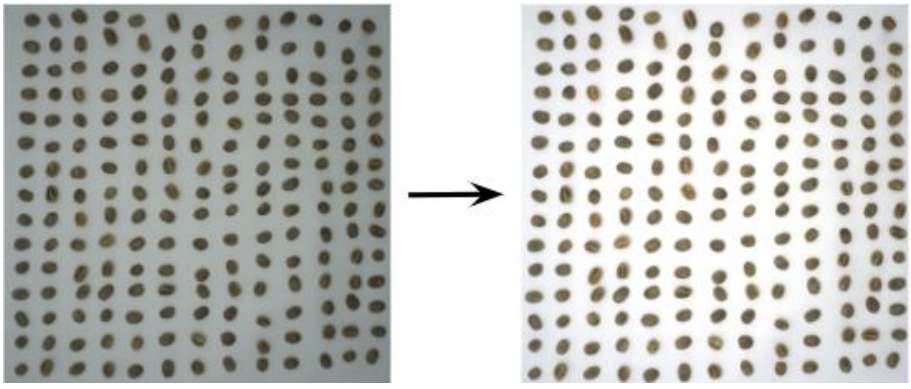


Figura 3.4. Granos buenos antes y después de White Patch. Fuente propia.

Segmentación

Para esta sección se tuvieron en cuenta 2 opciones para realizar esta etapa, la segmentación basada en umbrales mediante el algoritmo Otsu y la segmentación basada en agrupación o *clustering* mediante el algoritmo *k-means*. En el primer caso, a partir del histograma de una imagen en escala de grises, el algoritmo realiza el cálculo para identificar el valor intermedio óptimo que permita definir un umbral que separe al objeto de interés o grano de café del fondo de la imagen. Y en el segundo caso, partiendo de la información de cada uno de los píxeles en una imagen, el algoritmo *k-means* realiza un proceso iterativo donde forma k grupos de forma aleatoria para aplicar unos cálculos que perfeccionan dichos grupos hasta que estén estables o hasta que termine las iteraciones.

Como solo es necesario uno de esos algoritmos, decidimos hacer una serie de pruebas de segmentación con las fotografías del conjunto de imágenes. Los resultados arrojaron que ambos algoritmos tienen resultados similares, ver la Figura 3.5 y la Figura 3.6, pero en el caso *k-means* estos resultados no eran constantes y a veces segmentaba incorrectamente, como se observa en la Figura 3.7.

Figura 3.5. Grano segmentado con Otsu. Fuente propia.

Figura 3.6. Grano segmentado con *K-means.* Fuente propia.

Figura 3.7. Segmentación errónea con *K-means.* Fuente propia.

En definitiva, con las pruebas realizadas decidimos utilizar el algoritmo de segmentación por umbrales Otsu para separar los granos de café del fondo. A continuación, en la Figura 3.8 y la Figura 3.9 puede observar cómo fue el resultado de segmentación para los granos de la clase buenos.

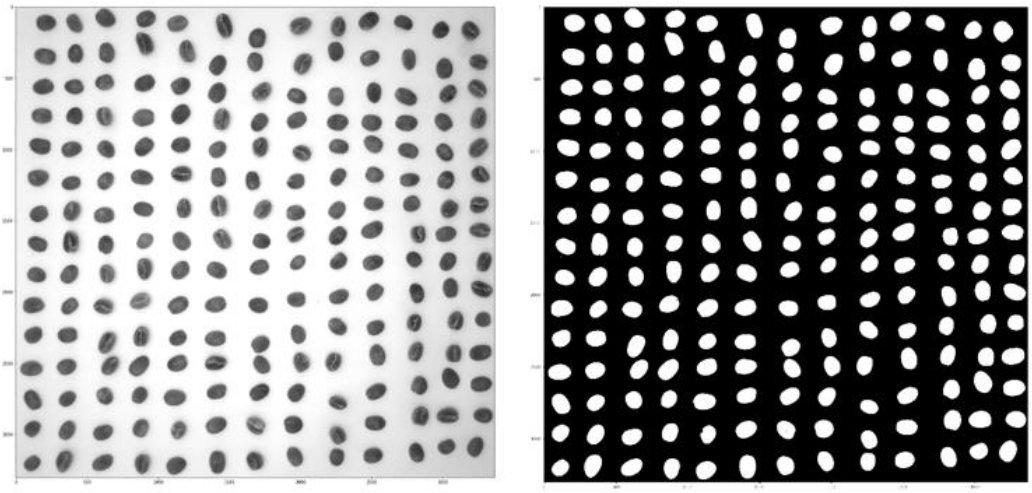


Figura 3.8. Imagen en escala de grises y máscara creada con Otsu para clase buenos. Fuente propia.

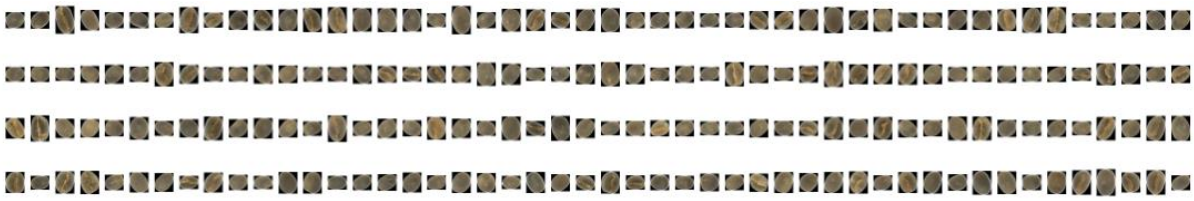


Figura 3.9. Granos segmentados. Fuente propia.

Extracción de características

A partir de la información extraída en la revisión sistemática de la literatura, identificamos que las características de color, textura y forma son las que normalmente se usan, pero no en todos los estudios emplean los tres tipos. Así que para conocer el comportamiento de todas las características halladas definimos 2 casos de prueba para realizar esa extracción de información. El primer caso consistió en tomar fotografías a 25 cm, porque era aproximadamente la altura del entorno controlado. Y en el segundo caso, las fotografías fueron tomadas a 51.5 cm de altura porque si se tomaban más alto, el lente no alcanzaba a enfocar los granos; y a una altura menor, el trípode no tenía la extensión suficiente para poderse sostener. El resultado en ambos casos fue una baja variación en las características de color y textura, pero una gran diferencia entre las características de forma, encontrando así que estas características dependen de la altura. En la Figura 3.10 y la Figura 3.11 se muestran esas variaciones respectivamente.

	area	eccentricity	extent	perimeter_crofton	solidity
0	24126	0.671993	0.731446	922.795921	0.918561
1	21302	0.526988	0.738883	720.942984	0.944447
2	19776	0.644582	0.733803	640.116333	0.940416
3	23280	0.439237	0.718607	937.757955	0.884936
4	20666	0.670262	0.779202	537.321653	0.980547

Figura 3.10. Resultado de las características de forma a 25 cm. Fuente propia.

	area	eccentricity	extent	perimeter_crofton	solidity
0	2410	0.480742	0.728978	179.638523	0.977688
1	2510	0.630091	0.748137	179.963846	0.982003
2	2689	0.734760	0.807023	190.190370	0.982822
3	2748	0.544827	0.810142	190.745730	0.978632
4	2707	0.593599	0.752153	191.991204	0.974091

Figura 3.11. Resultado de las características de forma a 51.1 cm. Fuente propia.

Por lo anterior, en este estudio se seleccionan las características de color y textura como los atributos extraídos de cada uno de los granos segmentados. Para el caso de las características de color, estas representan como se distribuye el color en una imagen y su extracción consistió en obtener los histogramas de los canales RGB y HSV de cada grano. Por ejemplo, los histogramas de color para un grano bueno se pueden observar en la Figura 3.12 y en la Figura 3.13 respectivamente.

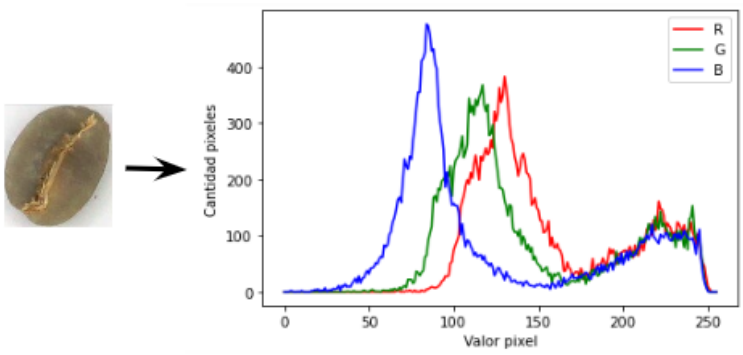


Figura 3.12. Histograma de color RGB para un grano bueno. Fuente propia.

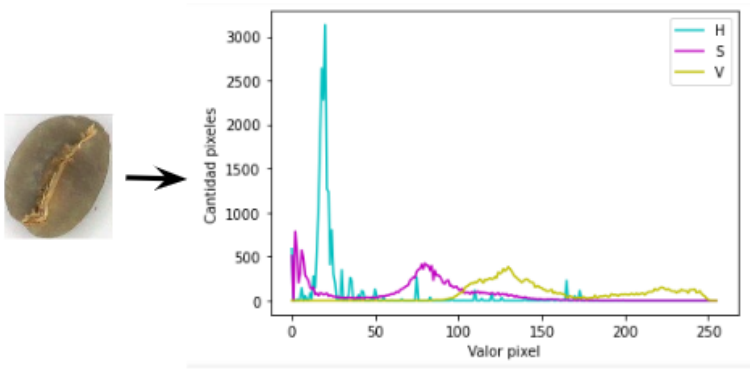


Figura 3.13. Histograma de color HSV para un grano bueno. Fuente propia.

En el caso de las características de textura, estas pueden representar si una superficie es suave o rugosa y su proceso de extracción comienza convirtiendo las imágenes de los granos segmentados a escala de grises. A partir de esa transformación, son construidas 4 matrices de co-ocurrencia de nivel de gris (GLCM) para obtener de cada matriz 13 de las 14 características de textura propuestas por Haralick, ya que esta última no es recomendable obtenerla por su inestabilidad computacional [39]. Por último, cada una de las características es promediada y recibe el cálculo del rango, conformando así un total de 26 características de textura.

Análisis de los datos

Con el fin de comprender las características extraídas anteriormente, realizamos un análisis en dicha información mediante la aplicación de estadística básica, gráficas de distribución y tablas de frecuencia. Para empezar, las medidas estadísticas generales son computadas, ya que estas medidas ayudan a identificar problemas o valores atípicos como el número total de granos (*count*), promedio (*mean*), desviación estándar (*std*), dato con el valor mínimo (*min*), cuartiles (*25%, 50%, 75%*) y dato con el valor máximo (*max*). En la figura 3.14 es resumida dicha información.

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_10	...	sum_squares_variance_range
count	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	2505.000000	...	2505.000000
mean	0.032335	0.025948	0.022754	0.040319	0.047505	0.050699	0.067864	0.091417	0.112176	0.122555	...	137.715238
std	0.305945	0.242551	0.257246	0.352381	0.399354	0.408848	0.508106	0.656762	0.825773	1.002269	...	57.168775
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	12.051858
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	102.922528
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	141.788076
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	175.523524
max	5.000000	6.000000	7.000000	6.000000	8.000000	9.000000	9.000000	13.000000	18.000000	25.000000	...	380.856384

Figura 3.14. Resumen de las medidas estadísticas por característica extraída. Fuente propia.

Como en el caso de las características de color son una gran cantidad y un análisis por grano con dichas características no proporcionaría demasiada información, en la Figura 3.15 y en la Figura 3.16 son comparados los histogramas promedio para RGB y HSV entre las dos clases de granos respectivamente. Esto permitió identificar que las distribuciones en los histogramas de RGB y HSV para granos buenos tiene una concentración de píxeles en intensidades más centrales, mientras que en los granos malos cada canal se distribuye en prácticamente todas las intensidades disponibles.

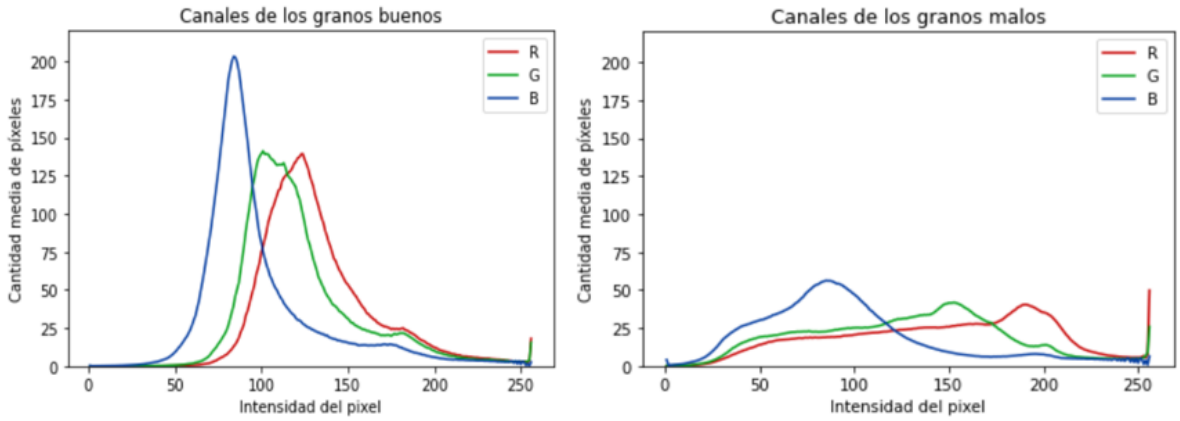


Figura 3.15. Histogramas RBG promedio para granos buenos y malos. Fuente propia.

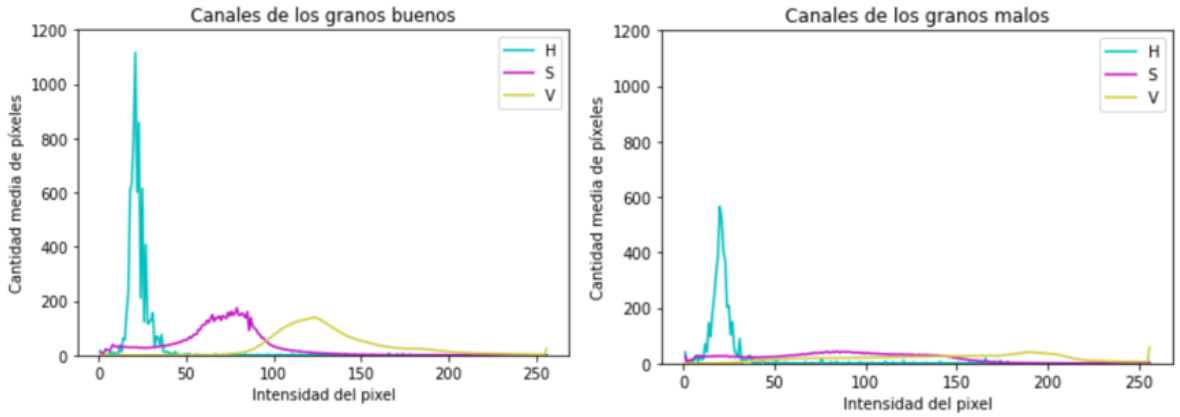


Figura 3.16. Histogramas RBG promedio para granos buenos y malos. Fuente propia.

Para el caso de las características de textura, los datos fueron resumidos en la Tabla 3.3 y graficados en histogramas como se muestra en la Figura 3.17 y en la Figura 3.18. Esto permitió identificar que la textura en los granos de café tiende a tomar la forma de una distribución normal y a tener algunas similitudes entre ambas clases.

Clase	Media				Rango			
Características de textura	**Bueno**		**Malo**		**Bueno**		**Malo**	
Valor	**Mínimo**	**Máximo**	**Mínimo**	**Máximo**	**Mínimo**	**Máximo**	**Mínimo**	**Máximo**
Segundo momento angular	0.0141	0.0846	0.0085	0.1895	0.0026	0.0135	0.0019	0.0327
Contraste	495.7403	3007.2885	426.0506	4310.1619	138.2185	1717.3960	182.0112	3285.4581
Correlación	0.6438	0.9207	0.5005	0.9584	0.0235	0.2074	0.0168	0.4287
Suma de cuadrados (Varianza)	2130.3013	7396.0869	1465.6707	9048.1826	28.1526	310.0138	12.0518	380.8563
Momento de diferencia inversa	0.1783	0.5533	0.1497	0.5456	0.0221	0.1048	0.0214	0.1286
Suma Promedio	146.3678	302.2601	69.1242	320.0437	0.7170	6.1179	0.0242	7.1802
Suma de varianza	7865.5637	27593.155	5100.9257	33760.125	252.1331	2631.8234	311.1111	4225.5504
Suma de entropía	5.6971	7.6300	5.3076	8.0172	0.0030	0.1846	0.0053	0.3106
Entropía	7.8339	10.8828	7.1182	11.218	0.1683	0.6129	0.1726	0.7624
Diferencia de varianza	0.000133	0.001175	0.000105	0.001085	0.000038	0.000340	0.000041	0.000352
Diferencia de entropía	2.8685	5.6967	2.6220	5.9290	0.3070	0.8917	0.2998	1.2929
Medida de información de correlación 1	(-0.5112)	(-0.293)	(-0.5275)	(-0.2591)	0.0167	0.0968	0.015	0.1066
Medida de información de correlación 2	0.9757	0.9985	0.9765	0.999	0.0002	0.0146	0.0001	0.0144

Tabla 3.3. Valores mínimos y máximos de las características de textura. Fuente propia.

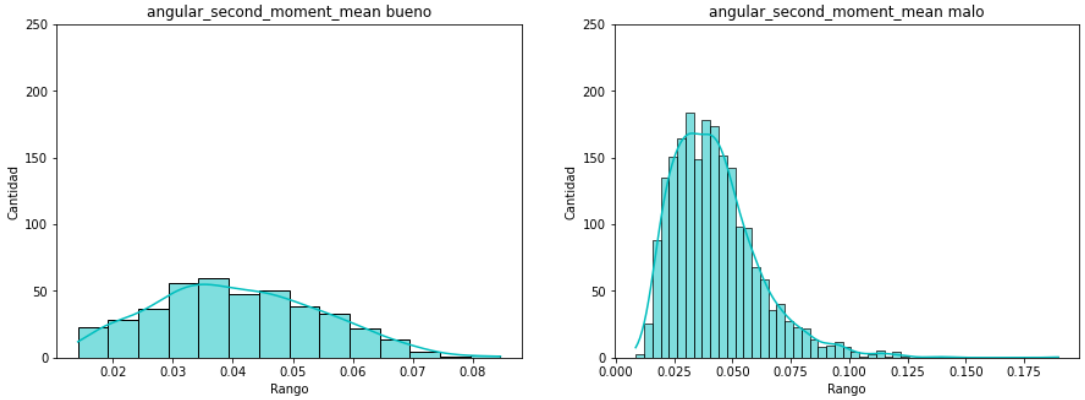


Figura 3.17. Distribución de la media del segundo momento angular entre granos buenos y malos. Fuente propia.

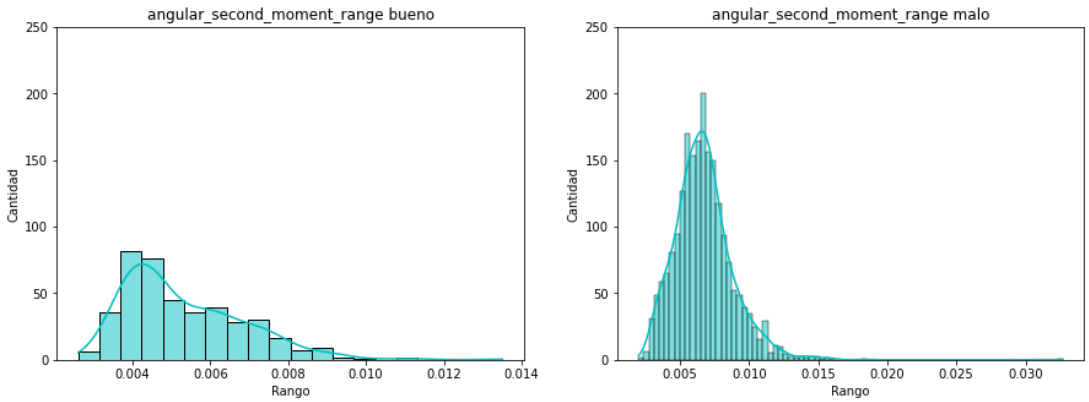


Figura 3.18. Distribución del rango del segundo momento angular entre granos buenos y malos. Fuente propia.

Adicionalmente, para comprender que representan las características de textura en los granos de café, en la Figura 3.19 y en la Figura 3.20 se muestra el gráfico de la diferencia de varianza para un grano bueno y malo respectivamente. Esto permitió comprender que en este caso la rugosidad es representada por las zonas más cercanas al negro y las secciones lisas por tonalidades blancas o grises claras.

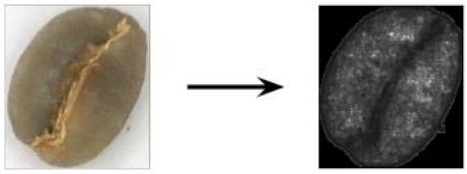


Figura 3.19. Diferencia de varianza para un grano bueno. Fuente propia.

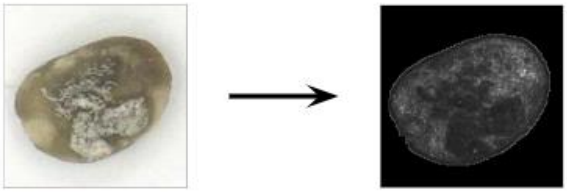


Figura 3.20. Diferencia de varianza para un grano malo. Fuente propia.

Limpieza de datos

En esta sección se realiza la identificación y eliminación de las características que no aportan información, como las que tienen el mismo valor en todas las muestras sin importar la clase, las características que están correlacionadas, entre otras. Aunque este proceso se realizó en un principio, fue descartado porque las características extraídas representan la información de contenidas en las imágenes y eliminarlas implicaría perder detalles importantes. Por ejemplo, la limpieza en las

características de color ocasionó dificultad para diferenciar el histograma RGB entre un grano bueno y un grano malo, ver Figura 3.21 y Figura 3.22 respectivamente.

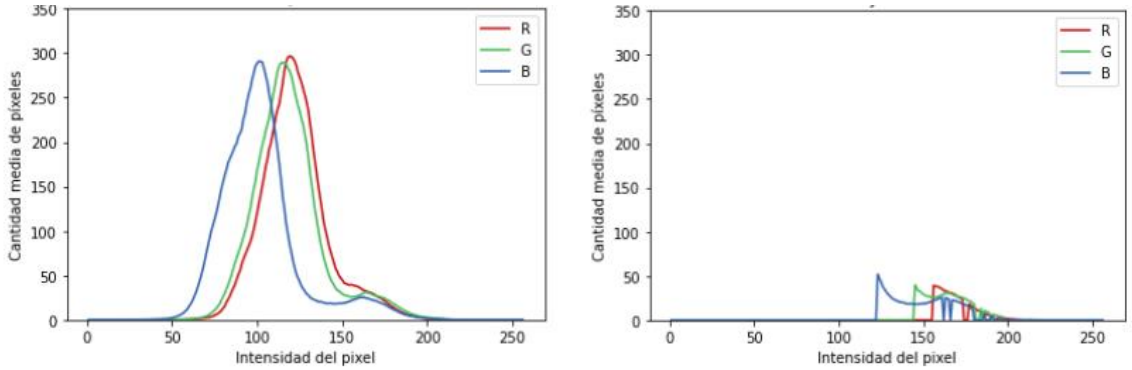


Figura 3.21. Histograma RGB para un grano sano antes y después de la limpieza. Fuente propia.

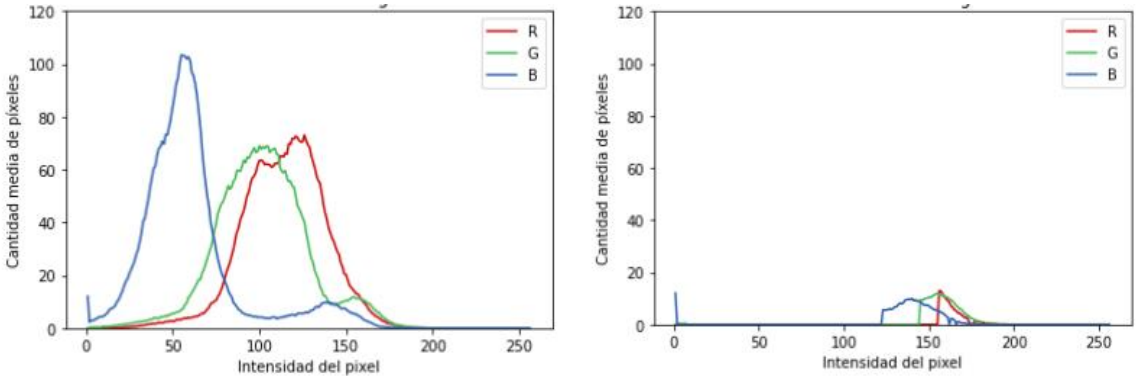


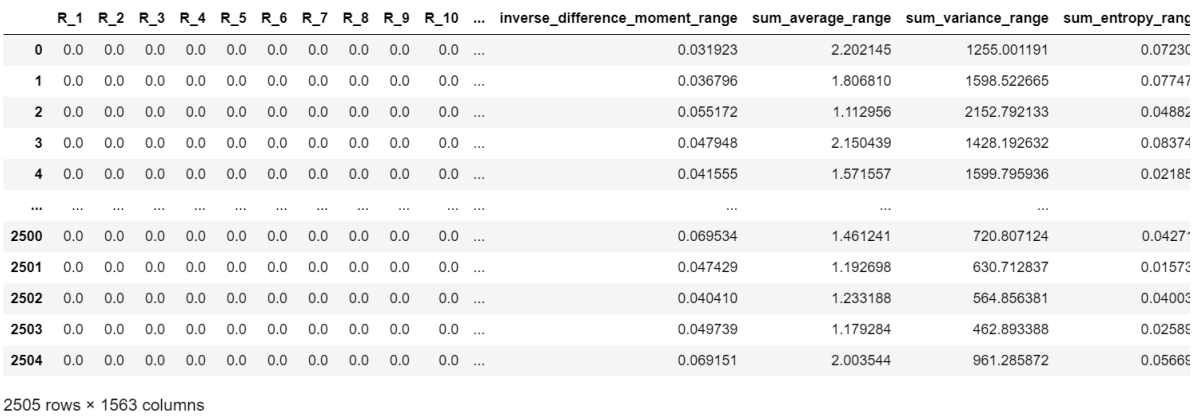
Figura 3.22. Histograma RGB para un grano malo antes y después de la limpieza. Fuente propia.

Estructurar, integrar y formateo de los datos

En resumen, esta parte incluye la generación de nuevos atributos a partir de atributos ya existentes, añadir nuevos registros, transformar algunos valores de atributos existentes e integrar las variaciones hechas. De los cuales no se efectúa ninguna modificación en esta parte, por lo que el conjunto de datos queda compuesto por un total de 2505 granos de café verde, 1562 características por grano y la etiqueta para identificar si es grano bueno o malo. La Tabla 3.4 presenta la distribución de las características y la Figura 3.23 presenta el conjunto de datos resultante.

Distribución de las características extraídas	
Canal R	256
Canal G	256
Canal B	256
Total de características para el espacio de color RGB	**768**
Canal H	256
Canal S	256
Canal V	256
Total de características para el espacio de color HSV	**768**
Media por cada característica de textura	13
Rango por cada característica de textura	13
Total de características para textura	**26**

Tabla 3.4. Distribución de las características que conforman el conjunto de datos. Fuente propia.



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_10	...	inverse_difference_moment_range	sum_average_range	sum_variance_range	sum_entropy_rang
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.031923	2.202145	1255.001191	0.07230
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.036796	1.806810	1598.522665	0.07747
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.055172	1.112956	2152.792133	0.04882
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.047948	2.150439	1428.192632	0.08374
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.041555	1.571557	1599.795936	0.02185
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2500	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.069534	1.461241	720.807124	0.0427
2501	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.047429	1.192698	630.712837	0.01573
2502	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.040410	1.233188	564.856381	0.04003
2503	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.049739	1.179284	462.893388	0.02588
2504	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.069151	2.003544	961.285872	0.05669

2505 rows × 1563 columns

Figura 3.23. Conjunto de datos resultante. Fuente propia.

3.2.4. Modelado

Como el objetivo de esta fase es la de elegir y configurar las técnicas de modelado más acordes al proyecto, se decide realizar este proceso en 3 etapas. En la primera es realizado el proceso para seleccionar los algoritmos de clasificación. En la segunda es efectuada la búsqueda de los hiperparámetros que permiten configurar dichos algoritmos. Y en la última etapa, se exportan los modelos definitivos a fin de utilizarlos posteriormente en la fase de evaluación.

Seleccionar técnicas de modelado

Con la información extraída en la revisión sistemática de la literatura, identificamos que los modelos o algoritmos más utilizados son la máquina de vectores de soportes o *Support Vector Vachines (SVM),* k vecinos más cercanos o *k-nearest neighbors (KNN), árboles de decisión o Decision Trees,* Bosque aleatorio o *Random Forest* y la red neuronal convolucional o *Convolutional Neural Network (CNN)*. De este modo, decidimos implementar esos algoritmos para realizar pruebas que nos permitieran conocer su desempeño.

En los resultados obtenidos pudimos notar que la *CNN* no tenía unas buenas métricas de evaluación para identificar los granos buenos. Esto es debido a la poca cantidad de imágenes de granos buenos que disponíamos en el conjunto de imágenes definitivo, ya que la *CNN* necesita la mayor cantidad de imágenes para su entrenamiento. Mientras que, en el caso de los otros algoritmos, como el conjunto de datos reunía una buena información por las características extraídas, tuvieron mejores resultados que la CNN. Por esta razón, en el presente trabajo la *CNN* es descartada y son elegidos los otros 4 algoritmos para realizar la calcificación de los granos de café verde.

Búsqueda de los hiperparámetros óptimos

Después de seleccionar los algoritmos de clasificación a usar, entrenamos cada modelo con el conjunto de datos preparado anteriormente. Este proceso de entrenamiento requiere configurar los hiperparámetros que permiten determinar las características que tendrá el modelo. Como los hiperparámetros pueden tener un amplio rango de valores, definimos de forma empírica un rango donde notamos que los modelos daban métricas de evaluación buenas. Partiendo de lo anterior, utilizamos la búsqueda de cuadrícula con *cross-validation* para encontrar los hiperparámetros que conformarían el modelo óptimo.

SVM

Los hiperparámetros para el modelo SVM, *gamma* y *C*, [1.e+00, 1.e+01, 1.e+02, 1.e+03, 1.e+04, 1.e+05] y [1.e-10, 1.e-09, 1.e-08, 1.e-07, 1.e-06, 1.e-05, 1.e-04, 1.e-03, 1.e-02, 1.e-01, 1.e+00], respectivamente. El hiperpárametro *kernel*, solo se usa como “rbf”, ya que tiene algunas ventajas, una de ella es su flexibilidad, este puede ir desde un clasificador lineal a uno muy complejo.

```
In [18]: param_grid = {
             'kernel': ['rbf'],
             'C': np.logspace(0, 5, 6),
             'gamma': np.logspace(-10, 0, 11),
         }

         grid = GridSearchCV(
             estimator = SVC(random_state=42),
             param_grid = param_grid,
             scoring = "accuracy",
             refit = True,
             verbose = 1,
             n_jobs = -1,
             cv = 10
         )

         grid.fit(X_train, y_train)
         print("_____")
         print("Mejores hiperparametros encontrados para SVM: ")
         print(grid.best_estimator_)

Fitting 10 folds for each of 66 candidates, totalling 660 fits
_____
Mejores hiperparametros encontrados para SVM:
SVC(C=1000.0, gamma=1e-08, random_state=42)
```

Figura 3.24. Encontrando los mejores hiperparámetros para el clasificador SVM. Fuente propia.

Como es mostrado en la Figura 3.24, son ejecutados 10 grupos para *cross-validation* de 66 candidatos posibles con los hiperparámetros configurados, para un total de 660 entrenamientos. El resultado del proceso realizado arroja que los hiperparámetros óptimos seleccionados son los siguientes:

- kernel: rbf
- C: 1000
- gamma: 1e-08

KNN

Los hiperparámetros para el modelo KNN, *n_neighbors, weights* y *p*, varían entre y [3, 5, 7, 9, 11, 13, 15, 17, 19], [“uniform”, “distance”] y [1, 2], respectivamente.

```
In [25]: param_grid = {
             'n_neighbors': [ x for x in range(3,21) if x % 2 != 0],
             'weights': ['uniform', 'distance'],
             'p': [1, 2]
         }

         grid = GridSearchCV(
             estimator = KNeighborsClassifier(),
             param_grid = param_grid,
             scoring = "accuracy",
             refit = True,
             verbose = 1,
             n_jobs = -1,
             cv = 10
         )

         grid.fit(X_train, y_train)
         print("_____________________________________________________________")
         print("Mejores hiperparametros enontrados para KNN: ")
         print(grid.best_estimator_)

Fitting 10 folds for each of 36 candidates, totalling 360 fits
_____________________________________________________________
Mejores hiperparametros enontrados para KNN:
KNeighborsClassifier(n_neighbors=9, p=1)
```

Figura 3.25. Encontrando los mejores hiperparámetros para el clasificador KNN. Fuente propia.

Como es presentado en la Figura 3.25, son realizados 10 grupos para *cross-validation* de 36 candidatos posibles con los hiperparámetros configurados, para un total de 360 entrenamientos. Después del proceso de entrenamiento, los hiperparámetros óptimos seleccionados son:

- ***n_neighbors:*** 9
- ***weights:*** uniform
- **p:** 1

Decision trees

Los hiperparámetros para el modelo *decision trees, criterion*, *max_depth* y *min_sample_split,* varían entre [“*gini*”, “*entropy*”], una lista entre 1 y 20 y [10, 12, 14, 16, 18], respectivamente.

```
In [6]: param_grid = {
            'criterion': ['gini', 'entropy'],
            'max_depth': list(range(1, 21)),
            'min_samples_split': list(range(10, 20, 2))
        }

        grid = GridSearchCV(
            estimator = DecisionTreeClassifier(random_state=42),
            param_grid = param_grid,
            scoring = "accuracy",
            refit = True,
            verbose = 1,
            n_jobs = -1,
            cv = 10
        )

        grid.fit(X_train, y_train)
        print("_____________________________________________________")
        print("Mejores hiperparametros encontrados para decision trees: ")
        print(grid.best_estimator_)

Fitting 10 folds for each of 200 candidates, totalling 2000 fits
_____________________________________________________
Mejores hiperparametros encontrados para decision trees:
DecisionTreeClassifier(max_depth=8, min_samples_split=10, random_state=42)
```

Figura 3.26. Encontrando los mejores hiperparámetros para el clasificador *decision trees*. Fuente propia.

Como se observa en la Figura 3.26, son definidos 10 grupos para *cross-validation* de 200 candidatos posibles con los hiperparámetros configurados, para un total de 2000 entrenamientos. Los hiperparámetros óptimos resultantes son:

- criterion: gini
- max_depth: 8
- min_sample_split*:* 10

Random forest

Los hiperparámetros para el modelo *random forest, criterion, max_depth y min_sample_split*, varían entre [“*gini*”, “*entropy*”], [5, 6, 7, 8, 9, 10] y [40, 45, 50, 55, 60, 65, 70, 75], respectivamente.

```
In [12]: param_grid = {
             'criterion'   : ['gini', 'entropy'],
             'min_samples_split': list(range(40, 80, 5)),
             'max_depth': list(range(5, 11)),
         }

         grid = GridSearchCV(
             estimator = RandomForestClassifier(random_state=42),
             param_grid = param_grid,
             scoring = "accuracy",
             refit = True,
             verbose = 1,
             n_jobs = -1,
             cv = 10
         )

         grid.fit(X_train, y_train)
         print("_____")
         print("Mejores hiperparametros encontrados para random forest: ")
         print(grid.best_estimator_)

         Fitting 10 folds for each of 96 candidates, totalling 960 fits
         _____
         Mejores hiperparametros encontrados para random forest:
         RandomForestClassifier(criterion='entropy', max_depth=7, min_samples_split=40,
                                random_state=42)
```

Figura 3.27. Encontrando los mejores hiperparámetros para el clasificador *random forest*. Fuente propia.

Como es expuesto en la Figura 3.27, son conformados 10 grupos para *cross-validation* de 96 candidatos posibles con los hiperparámetros configurados, dando un total de 960 entrenamientos. Los hiperparámetros óptimos seleccionados para este modelo son:

- criterion: entropy
- max_depth: 7
- min_sample_split*:* 40

Exportación de los algoritmos definitivos

Como se observa en la Figura 3.28, los mejores modelos son entrenados con los valores óptimos encontrados y son exportados para usarlos en la fase de evaluación realizada en el siguiente capítulo.

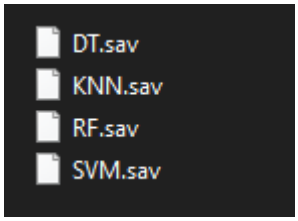


Figura 3.28. Modelos entrenados listos para la fase de evaluación. Fuente propia.

Capítulo 4

4. Evaluación y grado de calidad

4.1. Evaluación de los algoritmos

El objetivo de esta actividad es interpretar los modelos de acuerdo con las medidas de evaluación estudiadas para seleccionar el mejor modelo a fin de utilizarlo en la estimación del grado de calidad. Para esto es conformado el conjunto de datos de prueba a partir del 25% del conjunto de datos sin la etiqueta o clase que identifica los granos. De este modo, son presentadas las matrices de confusión y las métricas calculadas para los cuatro modelos elegidos.

KNN

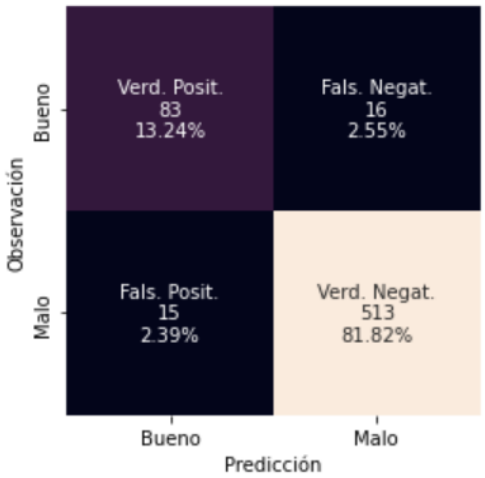


Figura 4.1. Matriz de confusión del modelo *KNN*. Fuente propia.

A partir de las predicciones realizadas por el algoritmo de clasificación *KNN*, la matriz de confusión es construida como lo muestra la Figura 4.1. Y con esta información son calculadas las métricas de evaluación descritas en la Figura 4.2.

	precision	recall	f1-score	support
Bueno	0.84694	0.83838	0.84264	99
Malo	0.96975	0.97159	0.97067	528
accuracy			0.95056	627
macro avg	0.90835	0.90499	0.90666	627
weighted avg	0.95036	0.95056	0.95046	627

Figura 4.2. Medidas de evaluación del modelo *KNN*. Fuente propia.

SVM

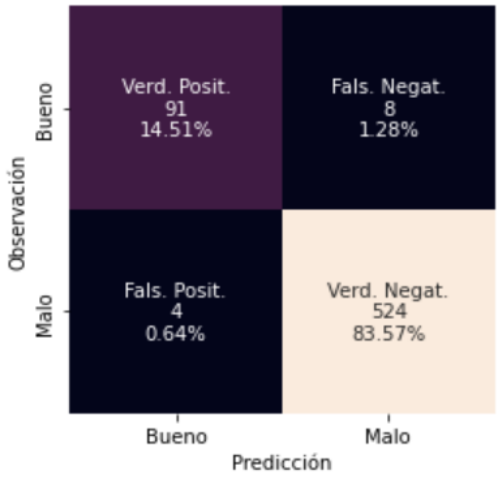


Figura 4.3. Matriz de confusión del modelo *SVM*. Fuente propia.

Con las predicciones realizadas por el algoritmo de clasificación *SVM*, la matriz de confusión es conformada como lo muestra la Figura 4.3. Y con esta información son calculadas las métricas de evaluación referidas en la Figura 4.4.

	precision	recall	f1-score	support
Bueno	0.95789	0.91919	0.93814	99
Malo	0.98496	0.99242	0.98868	528
accuracy			0.98086	627
macro avg	0.97143	0.95581	0.96341	627
weighted avg	0.98069	0.98086	0.98070	627

Figura 4.4. Medidas de evaluación del modelo *SVM*. Fuente propia.

Decision trees (DT)

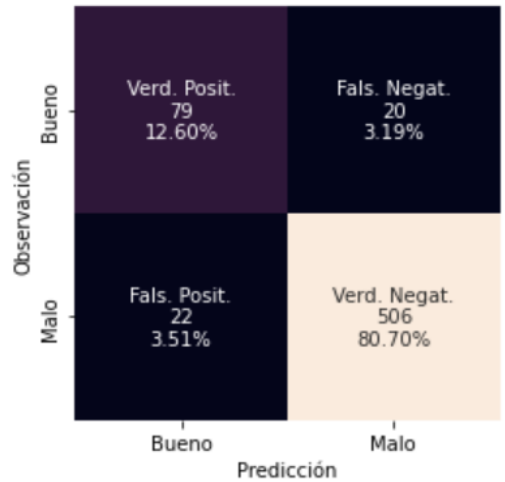


Figura 4.5. Matriz de confusión del modelo *DT*. Fuente propia.

Igual que en los casos anteriores, la matriz de confusión para el modelo de clasificación *DT* es construido como lo muestra la Figura 4.5 y son calculadas las métricas de evaluación de la Figura 4.6.

	precision	recall	f1-score	support
Bueno	0.78218	0.79798	0.79000	99
Malo	0.96198	0.95833	0.96015	528
accuracy			0.93301	627
macro avg	0.87208	0.87816	0.87508	627
weighted avg	0.93359	0.93301	0.93329	627

Figura 4.6. Medidas de evaluación del modelo *DT*. Fuente propia.

Random forest (RF)

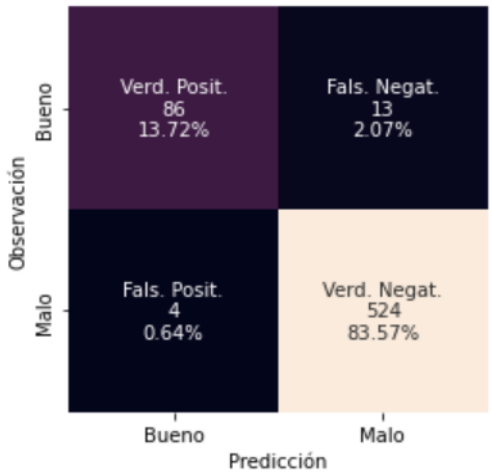


Figura 4.7. Matriz de confusión del modelo *RF*. Fuente propia.

Por último, siguiendo un proceso similar a lo anterior, la matriz de confusión y las métricas de evaluación para el modelo de clasificación *RF* es presentado en la Figura 4.7 y en la Figura 4.8 respectivamente.

	precision	recall	f1-score	support
Bueno	0.95556	0.86869	0.91005	99
Malo	0.97579	0.99242	0.98404	528
accuracy			0.97289	627
macro avg	0.96567	0.93056	0.94705	627
weighted avg	0.97260	0.97289	0.97236	627

Figura 4.8. Medidas de evaluación del modelo *RF*. Fuente propia.

En la Tabla 4.1, resumimos los resultados obtenidos para comparar las métricas de cada modelo:

Modelo	*KNN*		*SVM*		*DT*		*RF*	
Clase	**Bueno**	**Malo**	**Bueno**	**Malo**	**Bueno**	**Malo**	**Bueno**	**Malo**
Exactitud	95.056%		98.086%		93.301%		97.289%	
Precisión	84.694%	96.975%	95.789%	98.496%	78.218%	96.198%	95.556%	97.579%
Exhaustividad	83.838%	97.159%	91.191%	99.242%	79.798%	95.833%	86.869%	99.242%
F1 Score	84.264%	97.067%	93.814%	98.868%	79.000%	96.015%	91.005%	98.404%

Tabla 4.1. Resumen de las medidas de evaluación de cada modelo. Fuente propia.

Para seleccionar el mejor modelo, tuvimos en cuenta las siguientes consideraciones:

1. No podemos utilizar la exactitud como medida de evaluación, ya que el conjunto de datos no contiene las clases “Bueno” y “Malo” balanceadas.
2. Buscamos un equilibrio entre las 2 clases, desde el punto de vista de la clase “Bueno” esperamos una precisión alta, porque deseamos disminuir la posibilidad de falsos positivos; y desde el punto de vista de la clase “Malo” esperamos una exhaustividad alta, puesto que deseamos disminuir la posibilidad de tener falsos negativos. Lo anterior podemos traducir como, un modelo que identifique bien los granos buenos y cuando clasifique los granos malos, se reduzca la posibilidad de que los clasifique como buenos.
3. F1 score, es una medida primordial para la selección. Este valor debe ser alto para ambas clases.

El modelo con las mejores métricas de evaluación es SVM. Puesto que la precisión de la clase “Bueno” es del 97.789% y la exhaustividad de la clase “Malo” es del 99.242%, superior al de los demás clasificadores. Por este motivo es seleccionado como el mejor clasificador para ser utilizado en las pruebas de evaluación.

4.2. Pruebas de evaluación

Después de haber seleccionado el modelo, realizamos las pruebas de evaluación con imágenes propias de granos previamente clasificados por un experto. Se tomaron 149 granos de café en imágenes y las distribuimos en 11 pruebas. En cada prueba reutilizamos los granos sanos como es mostrado en la Figura 4.9.

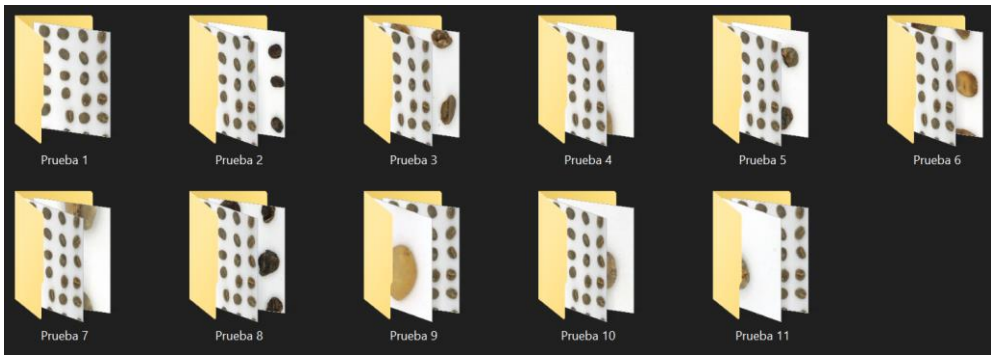


Figura 4.9. Imágenes de evaluación para obtener grado de calidad. Fuente propia.

Los grupos quedan constituidos de la siguiente manera:

- **Prueba 1:** 60 sanos.
- **Prueba 2:** 60 sanos y 9 cerezas secas.
- **Prueba 3:** 60 sanos y 9 conchas.
- **Prueba 4:** 60 sanos, 3 dañados por hongo.
- **Prueba 5:** 60 sanos y 10 brocados severos.
- **Prueba 6:** 60 sanos y 12 vinagres.
- **Prueba 7:** 60 sanos y 6 brocados leves.
- **Prueba 8:** 60 sanos y 20 negros.
- **Prueba 9:** 60 sanos y 10 pergaminos.
- **Prueba 10:** 60 sanos y 4 inmaduros.
- **Prueba 11:** 60 sanos y 6 partidos/mordidos/cortados.

Aunque habíamos pensado en solo usar el mejor modelo para que clasificara las pruebas propuestas, decidimos realizar la clasificación con todos los algoritmos y obtuvimos los siguientes resultados:

Modelo / Clase / Prueba	Original		*DT*		*KNN*		*RF*		*SVM*	
	Bueno	**Malo**	**Bueno**	**Malo**	**Bueno**	**Malo**	**Bueno**	**Malo**	**Bueno**	**Malo**
1	60	0	54	6	60	0	60	0	60	0
2	60	9	54	15	59	10	60	9	60	9
3	60	9	54	15	62	7	62	7	61	8
4	60	3	54	9	60	3	60	3	60	3
5	60	10	54	16	59	11	60	10	60	10

6	60	12	54	18	60	12	61	11	60	12
7	60	6	54	11	63	3	62	4	62	4
8	60	20	54	26	60	20	60	20	60	20
9	60	10	54	16	60	10	60	10	60	10
10	60	4	54	9	61	3	61	3	61	3
11	60	6	54	12	60	6	60	6	60	6

Tabla 4.2. Resumen de clasificación de los modelos construidos. Fuente propia.

Como se observa en la Tabla 4.2, el color azul representa la muestra de los granos verdaderos en cada prueba. En color rojo, están las pruebas donde los clasificadores se equivocaron. En color verde, son resaltadas las pruebas donde los clasificadores identifican correctamente. Esto permitió validar de nuevo que *SVM* es el mejor modelo. Y en el caso de los demás algoritmos, *DT* no clasifica correctamente los granos buenos, *KNN* se equivoca en dos pruebas más que *SVM* y *RF* se equivoca en una prueba más que *SVM*.

4.3. Estimar el grado de calidad

Para realizar la estimación del grado de calidad a partir de la identificación de los granos buenos y malos en las pruebas realizadas, proponemos la siguiente relación para obtener el grado de calidad:

$$Grado\ de\ calidad\ [\%] = \frac{Numero\ de\ granos\ buenos\ clasificados}{Numero\ total\ de\ granos\ evaluados}\ x\ 100$$

Adicionalmente, establecemos un error porcentual para determinar cuál es el error que tiene *SVM* al clasificar los granos de evaluación.

$$Error\ de\ clasificación[\%] = \frac{|Numero\ de\ granos\ buenos\ clasificados\ -\ Numero\ total\ de\ granos\ buenos|}{Numero\ total\ de\ granos\ buenos}\ x\ 100$$

A partir del resultado de la clasificación realizada por *SVM* presentado en la Tabla 4.2, el grado de calidad y el error de clasificación es calculado en cada una de las pruebas. La Tabla 4.3 describe los resultados obtenidos para cada caso.

Nombre de la prueba	Grado de calidad	Error de clasificación
Prueba 1	100,000%	0,000%
Prueba 2	86,957%	0,000%
Prueba 3	88,406%	1,667%
Prueba 4	95,238%	0,000%
Prueba 5	85,714%	0,000%
Prueba 6	83,333%	0,000%
Prueba 7	93,939%	3,333%
Prueba 8	75,000%	0,000%
Prueba 9	85,714%	0,000%
Prueba 10	95,313%	1,667%
Prueba 11	90,909%	0,000%

Tabla 4.3. Resumen de la estimación del grado de calidad. Fuente propia.

4.4. Análisis de resultados

Con base en lo anterior, encontramos que la prueba 1 nos sirvió de control para saber que el modelo cataloga correctamente los granos sanos en la clase “Bueno”. Esto asegura que en las siguientes pruebas si llega a existir un error porcentual, este será debido a granos malos clasificados como buenos.

Las pruebas 2, 4, 5, 6, 8, 9 y 11, son las que obtienen una clasificación correcta porque los defectos de cereza seca, dañados por hongo, brocados severos, vinagres, negros, pergaminos y partidos/mordidos/cortados, tienen en común que son fáciles de identificar por su color.

Y en el caso de las pruebas fallidas 3, 7 y 10, conformadas por los defectos concha, brocado leve e inmaduro. El modelo confundió algunos granos porque esos defectos tienen secciones verdes en su superficie y/o tenían el defecto en la otra cara del grano. A continuación, mostramos los granos que fueron confundidos por el modelo.

Figura 4.10. Clasificación fallida de granos con defecto concha. Fuente propia.

En la Figura 4.10, el grano con defecto concha señalado en el recuadro rojo, no fue detectado porque desde el punto de vista donde fue tomada la fotografía, parece que es un grano sano por su color verde en casi toda la totalidad del grano.

Figura 4.11. Clasificación fallida de granos con brocado leve. Fuente propia.

Para el caso de la prueba 7, los granos encerrados en los recuadros rojos de la Figura 4.11 no fueron detectados porque casi en su totalidad son de color verde y el punto de negro de la broca era muy pequeño o se ubicaba en la otra cara del grano.

Figura 4.12. Clasificación fallida de granos con defecto inmaduro. Fuente propia.

Por último, la prueba 10 falló porque el grano destacado en el recuadro rojo de la Figura 4.12 no tenía la película plateada que normalmente está bien adherida en su superficie. Además, por el ángulo de la fotografía no es posible distinguir el borde afilado que tiene este grano y sería la otra forma para identificarlo.

Capítulo 5

5. Discusión

5.1. Conclusiones

Este trabajo estuvo enfocado en responder la pregunta de investigación ¿Cómo estimar la calidad de los granos de café verde a partir de los defectos, mediante el uso de técnicas de procesamiento de imágenes y algoritmos de clasificación? Para esto, el presente trabajo presentó la implementación de un módulo para obtener el grado de calidad a través de la detección de los defectos físicos en los granos de café verde. Siguiendo el marco de trabajo de CRISP-DM, fue posible implementar las fases del procesamiento digital de imágenes, obteniendo los resultados esperados.

En la adquisición de imágenes, conseguimos el conjunto de imágenes utilizado por el estudio de C. E. Portugal-Zambrano et al. [17] y fue enriquecido con fotos nuestras tomadas en un entorno controlado. En el preprocesamiento, normalizamos las imágenes del conjunto de datos usando el filtro White-Patch. Para la fase de segmentación, utilizamos el algoritmo de segmentación por umbralización denominado Otsu. Por último, para la clasificación entrenamos cuatro algoritmos de aprendizaje automático junto a Grid-Search para encontrar los mejores hiperparámetros y seleccionar el mejor modelo.

A continuación, se presentan las conclusiones obtenidas en el desarrollo del presente trabajo de grado:

- En las pruebas que realizamos, encontramos que para la construcción de un conjunto de imágenes la altura de la cámara y su resolución pueden afectar especialmente a las características de forma. Y la limpieza de las características extraídas de una imagen puede ser contraproducente, ya que, aunque esto es importante cuando existen datos redundantes e independientes entre sí, las características eliminadas en una imagen puede ser la información de otra.
- A partir de las pruebas donde entrenamos los modelos con el conjunto de imágenes del estudio de C. E. Portugal-Zambrano et al. [17] y evaluamos con fotografías propias, descubrimos que la información contenida en las imágenes de entrenamiento no era suficiente para que los modelos clasifiquen cualquier imagen con granos de café verde. Por lo que, para

contrarrestar esta limitante, adicionamos nuestras fotografías al conjunto de imágenes. Esto nos permitió obtener mejores resultados en los modelos y determinar que para un mejor desempeño de los modelos de clasificación, es necesario un conjunto de imágenes que disponga de diferentes tipos de iluminación, alturas, grados de inclinación, fotografías de diversas cámaras, y entre otros factores.

- Por último, logramos estimar la calidad del grano de café verde de forma satisfactoria, ya que el clasificador SVM demostró tener métricas de evaluación superiores a los otros modelos y una gran capacidad para diferenciar los granos sanos de los defectuosos. Esto pone de manifiesto la efectividad de nuestro sistema como una base para el control de calidad para el caficultor, experto y la industria.

5.2. Trabajos futuros

Para continuar el desarrollo de este trabajo proponemos los siguientes trabajos futuros:

- Crear un repositorio de imágenes de granos de café verde categorizado con sus defectos correspondientes, en diferentes ángulos, alturas, iluminación y mayor variedad de granos.
- Implementar algoritmos más robustos dentro del campo del aprendizaje no supervisado, reforzado o aprendizaje profundo.
- Construir un módulo de retroalimentación más preciso y accesible para los caficultores y expertos.

Bibliografía

[1] G. I. Puerta, «Calidad física del café de varias regiones de Colombia según altitud suelos y buenas prácticas de beneficio», ISSN: 0120-0275, 2016. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/676/1/arc067%2801%297-40.pdf

[2] G. I. Puerta, «Buenas prácticas para la prevención de los defectos de la calidad del café: Fermento reposado fenólico y mohoso», Centro Nacional de Investigaciones de Café (Cenicafé), ISSN: 0120-0178, 2015. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/675/1/avt0461.pdf

[3] G. I. Puerta, «Los catadores de café», Centro Nacional de Investigaciones de Café (Cenicafé), ISSN: 0120-0178, 2013. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/367/1/avt0381.pdf

[4] G. I. Puerta, «Buenas prácticas agrícolas para el café», Centro Nacional de Investigaciones de Café (Cenicafé), ISBN: 0120-0178, 2013. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/359/1/avt0349.pdf

[5] C. Arias, D. Yolima, y L. M. Caro Gutiérrez, «Modelo para la gestión de información y del conocimiento en la cadena productiva del café, enfocado al eslabón de comercialización del municipio de Moniquirá-Boyacá», doi: 10.15332/tg.pre.2020.000.

[6] D. R. Gómez Restrepo, «Condiciones laborales en el sector cafetero: la jornada, el salario y el descanso: El caso del municipio de Andes, Antioquia», Universidad EAFIT, 2013. [En línea]. Disponible en: https://repository.eafit.edu.co/bitstream/handle/10784/7242/CONDICIONES%20LABORALES%20EN%20EL%20SECTOR%20CAFETERO.pdf?seque

[7] C. A. Serna, J. F. Trejos, y G. Cruz, «Estudio económico de sistemas de producción cafeteros certificados y no certificados en dos regiones de Colombia», 2010, [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/294/1/arc061%2803%29222-240.pdf

[8] A. García y D. Sandoval, «Posibles acciones en favor de los pequeños productores cafeteros en medio de la crisis actual del sector», JJ Echavarría P Esguerra McAllister CF Robayo Misión Estud. Para Compet. Caficultura En

Colomb., 2013, [En línea]. Disponible en: https://www.urosario.edu.co/Mision-Cafetera/Archivos/Evolucion-Pequenos-Cafeteros-Arturo-Garcia.pdf

[9] A. M. Feria-Morales, «Examining the case of green coffee to illustrate the limitations of grading systems/expert tasters in sensory evaluation for quality control», Food Qual. Prefer., vol. 13, n.o 6, pp. 355-367, 2002, doi: 10.1016/S0950-3293(02)00028-9.

[10] E. R. Arboleda, A. C. Fajardo, y R. P. Medina, «Green coffee beans feature extractor using image processing», TELKOMNIKA Telecommun. Comput. Electron. Control, vol. 18, n.o 4, pp. 2027-2034, 2020, doi: 10.12928/TELKOMNIKA.v18i4.13968.

[11] A. F. Sánchez-Aguiar, A. M. Ceballos-Arroyo, y A. Espinosa-Bedoya, «Toward the recognition of non-defective coffee beans by means of digital image processing», en 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), 2019, pp. 1-5. doi: 10.1109/STSIVA.2019.8730267.

[12] J. P. L. Pizzaia, I. R. Salcides, G. M. de Almeida, R. Contarato, y R. de Almeida, «Arabica coffee samples classification using a Multilayer Perceptron neural network», en 2018 13th IEEE International Conference on Industry Applications (INDUSCON), 2018, pp. 80-84. doi: 10.1109/INDUSCON.2018.8627271.

[13] N.-F. Huang, D.-L. Chou, y C.-A. Lee, «Real-time classification of green coffee beans by using a convolutional neural network», en 2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC), 2019, pp. 107-111. doi: 10.1109/ICISPC.2019.8935644.

[14] M. García, J. E. Candelo-Becerra, y F. E. Hoyos, «Quality and defect inspection of green coffee beans using a computer vision system», Appl. Sci., vol. 9, n.o 19, p. 4195, 2019, doi: 10.3390/app9194195.

[15] J. Ramirez-Ticona, J. C. Gutiérrez-Cáceres, y C. E. Portugal-Zambrano, «Cell-phone based model for the automatic classification of coffee beans defects using white patch», en 2016 XLII Latin American Computing Conference (CLEI), 2016, pp. 1-6. doi: 10.1109/CLEI.2016.7833335.

[16] J. C. Borrero Becerra y C. A. Diaz Molano, «Elaboración de base de datos de fotografías de granos de café seco con diferentes defectos físicos, caracterizados con métodos estándar de PDI y clasificación.».

[17] C. E. Portugal-Zambrano, J. C. Gutiérrez-Cáceres, J. Ramirez-Ticona, y C. A. Beltran-Castañón, «Computer vision grading system for physical quality evaluation of green coffee beans», en 2016 XLII Latin American Computing Conference (CLEI), 2016, pp. 1-11. doi: 10.1109/CLEI.2016.7833383.

[18] C. Pinto, J. Furukawa, H. Fukai, y S. Tamura, «Classification of Green coffee bean images basec on defect types using convolutional neural network (CNN)», en 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017, pp. 1-5. doi: 10.1109/ICAICTA.2017.8090980.

[19] F. F. L. dos Santos, J. T. F. Rosas, R. N. Martins, G. de M. Araújo, L. de A. Viana, y J. de P. Gonçalves, «Quality assessment of coffee beans through computer vision and machine learning algorithms», 2020, doi: 10.25186/.v15i.1752.

[20] C. Velásquez Agudelo y M. Trávez Velásquez, «Café especial, una alternativa para el sector cafetero en Colombia», Universidad EAFIT, 2019. [En línea]. Disponible en: https://docs.google.com/viewerng/viewer?url=https://repository.eafit.edu.co/bitstream/handle/10784/15236/Mateo_Travez_Camilo_Velasquez_2019.pdf?sequence%3D2&isAllowed=y

[21] M. Barbee, «Fotogalería en flicker de Mareen Barbee». https://www.flickr.com/photos/27781737@N05/with/40480901872/

[22] R. C. Gonzalez y R. E. Woods, Digital Image Processing, ISSN: 978-0-13-335672-4. Pearson, 2018.

[23] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, y M. DATA, «Practical machine learning tools and techniques», en Data Mining, 2005, vol. 2, n.o 4. doi: 10.1016/C2009-0-19715-5.

[24] «3.2. Tuning the hyper-parameters of an estimator», scikit-learn. https://scikit-learn/stable/modules/grid_search.html (accedido 24 de octubre de 2022).

[25] «3.1. Cross-validation: evaluating estimator performance», scikit-learn. https://scikit-learn/stable/modules/cross_validation.html (accedido 23 de octubre de 2022).

[26] S. V. Stehman, «Selecting and interpreting measures of thematic classification accuracy», Remote Sens. Environ., vol. 62, n.o 1, pp. 77-89, 1997, doi: 10.1016/S0034-4257(97)00083-7.

[27] «3.3. Metrics and scoring: quantifying the quality of predictions», scikit-learn. https://scikit-learn/stable/modules/model_evaluation.html (accedido 23 de octubre de 2022).

[28] B. Kitchenham, «Procedures for performing systematic reviews», Keele UK Keele Univ. ISSN1353-7776, vol. 33, n.o 2004, pp. 1-26, 2004.

[29] A. Guide, «Project management body of knowledge (pmbok® guide)», en Project Management Institute, ISSN: 978-1933890517, 2001, vol. 11, pp. 7-8. [En línea]. Disponible en: http://lms.aambc.edu.et:8080/xmlui/bitstream/handle/123456789/160/PROJECT%20MANAGEMENT%20BODY%20OF%20KNOWLEDGE%20(PMBOK%20GUIDE)%20(%20PDFDrive.com%20).pdf?sequence=1

[30] P. Chapman et al., «CRISP-DM 1.0: Step-by-step data mining guide», SPSS Inc, vol. 9, n.o 13, pp. 1-73, 2000.

[31] «IBM Documentation», 8 de marzo de 2021. https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide (accedido 23 de octubre de 2022).

[32] E. H. Land, «The retinex theory of color vision», Sci. Am., vol. 237, n.o 6, pp. 108-129, 1977, doi: 10.1007/978-1-4419-8071-7_260.

[33] J. Cepeda-Negrete y R. E. Sanchez-Yanez, «Experiments on the white patch retinex in RGB and CIELAB color spaces», Acta Univ., vol. 22, pp. 21-26, 2012, doi: 10.15174/au.2012.337.

[34] M. V. Boland, «Quantitative description and automated classification of cellular protein localization patterns in fluorescence microscope images of mammalian cells», Carnegie Mellon University, 1999.

[35] International Coffee Organization, «Explanatory Note for the Coffee Production Report», International Coffee Organization (ICO), 2021. [En línea]. Disponible en: https://www.ico.org/prices/po-production.pdf

[36] M. L. Q. Rizzuto y M. Rosales, «El mercado mundial del café: tendencias recientes, estructura y estrategias de competitividad», Visión Gerenc. ISSN 1317-8822, n.o 2, pp. 291-307, 2014.

[37] M. N. Clifford, Coffee: botany, biochemistry and production of beans and beverage, ISSN: 1-4615-6657-6. Springer Science & Business Media, 2012.

[38] «¿Café 100% Arábica? ¿Blend? ¿80% Arábica-20% Robusta?», Kenós Café. https://kenoscafe.cl/blogs/aprende-con-kenos/cafe-100-arabica-blend-80-arabica-20-robusta (accedido 23 de octubre de 2022).

[39] B. Heredia, «Guía Técnica para el Cultivo del Café», Instituto del Café de Costa Rica (ICAFE), ISSN: 978-9977-55-041-4, 2011. [En línea]. Disponible en: http://www.icafe.cr/wp-content/uploads/cicafe/documentos/GUIA-TECNICA-V10.pdf

[40] Asociación nacional del café, «Manual Técnico para la Producción de Café Robusta», Asociación nacional del café (Anacafé), 2016. [En línea]. Disponible en: https://www.anacafe.org/uploads/file/283f6fd107ef4ce38af855880c47c49d/Manual-Cafe-Robusta.pdf

[41] Federación nacional de cafeteros, «Guia ambiental para el sector cafetero», Federación nacional de cafeteros (FNC), ISSN 2248-8731, 2012. [En línea]. Disponible en: https://redjusticiaambientalcolombia.files.wordpress.com/2012/09/guia-ambiental-para-el-subsector-cafetero.pdf

[42] J. Arcila, F. F. Farfán, A. M. Moreno, L. F. Salazar, y E. Hincapié, Sistemas de producción de café en Colombia, ISSN: 958-98193-0-3. 2007. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/720/1/Sistemas%20producci%c3%b3n%20caf%c3%a9%20Colombia.pdf

[43] G. I. Puerta, «Sistema de aseguramiento de la calidad y la inocuidad del café en la finca», Centro Nacional de Investigaciones de Café (Cenicafé), ISSN: 0120-0178, 2013. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/415/1/avt0351.pdf

[44] P. C. R. para el Desarrollo, «Guía técnica para el beneficiado de café protegido bajo una indicación geográfica ó denominación de origen», Denominaciones Origen Café ISSN 978-92-9248-267-1, 2010, [En línea].

Disponible en:
https://repositorio.iica.int/bitstream/handle/11324/14124/BVE21011258e.pdf

[45] Federación nacional de cafeteros, «Cartilla 20 Beneficio del café: Despulpado, remoción de mucilago y lavado», Federación nacional de cafeteros (FNC), ISSN 2248-8731. [En línea]. Disponible en: https://caldas.federaciondecafeteros.org/app/uploads/sites/11/2020/07/Cartilla_20-Beneficio-del-caf%C3%A9-I.-Despulpado-remoci%C3%B3n-de-mucilago-y-lavado..pdf

[46] G. I. PUERTA, «Influencia del proceso de beneficio en la calidad del café», ISSN: 0120-0275, 1999. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/58/1/arc050%2801%29078-088.pdf

[47] J. Pabón y V. Osorio, «Factores e indicadores de la calidad física, sensorial y química del café», ISSN: 958-8490-39-1., Cenicafé, 2019. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/4227/1/Cap07.pdf

[48] A. PEÑUELA, C. OLIVEROS, y J. SANZ, «Remoción del mucílago de café a través de fermentación natural», ISSN: 0120-0275, 2014. [En línea]. Disponible en: https://biblioteca.cenicafe.org/bitstream/10778/494/1/arc061%2802%29159-173.pdf

[49] C. de Colombia, Fermentación del Café. 2010. [Photo]. Disponible en: https://www.flickr.com/photos/100porcientocafedecolombia/8552584870/

[50] «Proceso de beneficiado - AECafé», 10 de enero de 2020.https://www.asociacioncafe.com/proceso-de-beneficiado-cafe/

[51] S. C. Association, «El café Arábica lavado Guía de defectos del café verde», Recuperado Httpsbootcoffee Comwp-Contentuploads201909SCAThe-Arab.-Green-Coffee-Defect-Guid. Pdf, 2019, [En línea]. Disponible en: https://bootcoffee.com/wp-content/uploads/2019/09/SCA_The-Arabica-Green-Coffee-Defect-Guide_Spanish_updated.pdf

[52] «Procesamiento Digital de Imágenes - Facultad de Ciencias Exactas ...», documentop.com. https://documentop.com/procesamiento-digital-de-imagenes-facultad-de-ciencias-exactas-_59fd9c191723dded73187406.html (accedido 23 de octubre de 2022).

[53] E. L. van den Broek, T. Kok, T. E. Schouten, y L. G. Vuurpijl, «Human-centered content-based image retrieval», en Human Vision and Electronic Imaging XIII, 2008, vol. 6806, pp. 575-586. doi: 10.1117/12.767190.

[54] K. Erdoğan y N. Yılmaz, «Shifting Colors to Overcome not Realizing Objects Problem due to Color Vision Deficiency», dic. 2014. doi: 10.15224/978-1-63248-034-7-27.

[55] M. Sonka, V. Hlavac, y R. Boyle, Image processing, analysis, and machine vision, ISSN: 1-285-98144-8. Cengage Learning, 2014.

[56] Pixelci, «Sala de estar con poca iluminación en una casa de planta abierta con un altillo», iStock. https://www.istockphoto.com/es/foto/sala-de-estar-con-poca-iluminaci%C3%B3n-en-una-casa-de-planta-abierta-con-un-altillo-gm1265084899-370700242

[57] N. Otsu, «A threshold selection method from gray-level histograms», IEEE Trans. Syst. Man Cybern. ISSN 0018-9472, vol. 9, n.o 1, pp. 62-66, 1979, doi: 10.1109/TSMC.1979.4310076.

[58] M. Hall-Beyer, «GLCM texture: A tutorial v. 3.0 March 2017», 2017, doi: 10.13140/RG.2.2.12424.21767.

[59] R. M. Haralick, K. Shanmugam, y I. H. Dinstein, «Textural features for image classification», IEEE Trans. Syst. Man Cybern., n.o 6, pp. 610-621, 1973, doi: 10.1109/TSMC.1973.4309314.

[60] M.-C. Desseroit et al., «Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort», J. Nucl. Med., vol. 58, n.o 3, pp. 406-411, 2017, doi: 10.2967/jnumed.116.180919.

[61] W. Burger y M. J. Burge, «Regions in Binary Images», Digital Image Processing: An Algorithmic Introduction Using Java. Springer London, London, pp. 209-252, 2016. [En línea]. Disponible en: https://doi.org/10.1007/978-1-4471-6684-9_10

[62] G. James, D. Witten, T. Hastie, y R. Tibshirani, An introduction to statistical learning, ISBN: 978-1-4614-7138-7., vol. 112. Springer, 2013.

Anexos

A. Teoría adicional

A.1. Generalidades del café

El café es uno de los *commodities* o materias primas más comercializadas en el mundo, siendo Brasil, Vietnam y Colombia los principales productores [35]. Desde 1970 el consumo de café ha ido en aumento año tras año, gracias al mejoramiento en la calidad de los granos de café y a la demanda en Europa de cafés especiales [36]. Por esa razón, los productores deben aplicar buenas prácticas y llevar un control de calidad en su café, ya que deben evitar la aparición de defectos. Estos defectos son características que pueden formarse en cualquiera de las etapas del proceso productivo del café, sobre todo en la postcosecha, y se ven reflejados en las cualidades organolépticas o sensoriales del café que comprenden el aroma, la acidez, el amargor, el cuerpo y el sabor de la bebida [2].

Aunque existen más de 10 especies de café, el café robusta y el café arábigo son las dos especies más importantes a nivel comercial por cubrir el total del mercado mundial. El café robusta se caracteriza por cultivarse en alturas desde 0 a los 700 metros sobre el nivel del mar. Es más resistente a enfermedades como la roya, tiene un sabor amargo, un olor más vegetal e intenso, es utilizado en cafés de menor calidad, su producción y su rendimiento es mayor al del café arábigo por su grano de café más pequeño, y sus costos de producción son más bajos. Por otra parte, el café arábigo es cultivado en alturas desde los 1000 a 2000 metros sobre el nivel del mar. Es más susceptible a enfermedades, tiene casi la mitad de cafeína que un grano de café robusta, es un café de alta calidad y dependiendo de los cuidados que reciba, el café puede adquirir olores y sabores únicos como frutales o cítricos sin perder él toque ácido que representa su sabor [37]. La Figura A.1 ilustra la diferencia física entre el grano de café robusta y el grano de café arábigo.

Figura A.1. Comparación de granos de tipo arábica contra robusta. Tomada de [38].

Independientemente de la especie, el cultivo de café se debe sembrar en zonas donde la temperatura promedio es de 22 grados centígrados, las precipitaciones llegan a una media de 2500 milímetros anuales, la humedad relativa llega hasta un 85% [39], la ubicación del sembrado debe estar protegida de los vientos y los suelos deben ser fértiles, planos o con una leve inclinación [40].

Cuando hay un terreno adecuado y ya es clara la especie a cultivar. Los caficultores deben cumplir con 3 fases para el sembrado de café. Fase de germinación y trasplante, consiste en ubicar las semillas en un sustrato con arena de río para un correcto desarrollo de las raíces. Fase del mucilago, consiste en el desarrollo de las plántulas y posteriormente seleccionar las que serán sembradas definitivamente. Fase de sembrado, consiste en limpiar el terreno, realizar el trazado y marcado para distribuir el espacio, y sembrar las plántulas [41].

Es importante llevar buenas prácticas en todas las fases, como hacer constantemente la remoción de malezas, llevar un control fitosanitario para evitar la aparición de plagas o enfermedades, llevar prácticas de conservación de suelos, entre otros; de este modo, asegurar una alta calidad en la cereza producida por los cafetos sembrados [42].

Para el desarrollo de la cereza del café, es necesario que el cafeto se encuentre en la fase reproductiva. Esta fase comienza con la aparición de inflorescencias en la zona donde las hojas unen con las ramas, que crecerán hasta convertirse en flores; y finaliza cuando las flores florecen y son fecundadas mediante polinización cruzada o autofecundación, como es el caso de la especie arábiga [37]. Con las flores fecundadas, inicia el desarrollo de la cereza del café que comprende 4 etapas que duran 45 semanas en la especie Robusta [40] y 32 semanas en la especie Arábiga.

- ➔ **1° etapa:** Crecimiento de la flor hasta convertirse en un fruto muy pequeño
- ➔ **2° etapa:** El fruto crece hasta adquirir su tamaño final y la semilla mantiene una consistencia gelatinosa.
- ➔ **3° etapa:** La semilla completa su formación al adquirir una consistencia sólida.
- ➔ **4° etapa:** La cereza ya está totalmente formada para dar comienzo el proceso de maduración.

Cuando la cereza de café está madura, su composición está distribuida en 4 capas, donde la primera es el exocarpio o epidermis, que es la capa externa de color rojizo que contiene a la pulpa de la cereza; la segunda es llamada mesocarpio, el cual contiene los azúcares y mucílagos del fruto; la tercera capa es el endocarpio o

pergamino, que es una membrana de color amarillo pálido; y por último el endospermo, que es el grano comúnmente conocido como café verde [42]. Durante las etapas donde se desarrolla la cereza, el cafeto debe estar bien hidratado para evitar la mayoría de los defectos que afectan directamente a la formación del grano y por ende, su calidad.

El beneficio del café

El beneficio es conocido como el conjunto de procesos que someten a las cerezas o bayas del café a transformarlas en pergamino seco. Este proceso consta de separar las diferentes capas que recubren a las semillas o granos de café de manera eficiente, sin afectar su calidad o rendimiento [42]. Cabe aclarar que como en todo proceso de preparación y almacenamiento de alimentos, deben llevar un control sanitario adecuado y buenas prácticas para mantener la inocuidad y la calidad del café que va a ser consumido por el usuario final [41][43].

El beneficio puede ocurrir de dos formas [44]:

- ➔ **Beneficio seco:** método más antiguo y amigable con el medio ambiente, en este método extienden las cerezas de café en el suelo, a la luz del sol durante 3 semanas, moviéndose constantemente para que el fermentado y secado se realice uniformemente.

- ➔ **Beneficio húmedo:** método donde usan máquinas especializadas para retirar la pulpa de los granos, después son llevados a una pila de fermento donde es retirado el mucílago y posteriormente lavado para obtener el grano que será secado y almacenado para su venta o exportación.

Entraremos en detalle en el beneficio húmedo, ya que es el más empleado actualmente a la hora de procesar el café en fincas cafeteras [45][46][47].

- ➔ **Cosecha:** Cuando las bayas del cafeto o cerezas del café ya se encuentran maduras, son recolectadas manualmente como lo hace la persona de la Figura A.2.

Figura A.2. Recolección de cerezas de café en temporada de cosecha. Tomado de [21].

- **Despulpado:** Consiste en retirar la pulpa de la cereza madura, mediante maquinaria especial, como lo realiza en la Figura A.3, donde está efectúa presión sobre el fruto para obtener los dos granos con mucílago. Este paso debe iniciar antes de que los frutos cumplan 6 horas de cosechados, ya que es muy posible que aparezcan defectos y por ende afecte la calidad.

Figura A.3. Despulpado de las cerezas de café haciendo uso de una despulpadora. Tomado de [21].

- **Remoción del mucílago:** Consiste en retirar el mucílago del grano. Para lograr removerlo, utilizan dos métodos. La fermentación, como en la Figura A.4, donde los granos son llevados a un tanque lleno de agua durante 12 a 16 horas, donde el mucílago se remueve lentamente por el proceso de fermentación [48]. La otra opción son las máquinas especializadas. En esta

etapa es donde los defectos que más afectan la calidad del café aparecen, ya que los tiempos deben ser exactos para no dejar el grano con mucílago en sus paredes o llegar a un sobre fermento.

Figura A.4. Remoción del mucílago empleando tanques de fermentación. Tomado de [49].

- **Lavado:** Este paso buscan lavar y retirar el mucílago fermentado de los granos de café con agua limpia, como en la Figura A.5, puesto que, si no es retirado, los compuestos fermentados podrían ocasionar la aparición de defectos. Hay varios métodos que son empleados para lavar los residuos, pero lo más importante es buscar consumir la menor cantidad de agua posible.

Figura A.5. Lavando granos de café con residuos de mucílago con agua limpia. Tomado de [50].

- ***Secado:*** Este proceso es llevado a cabo para reducir el contenido de humedad del grano al 12%, porque en un porcentaje mayor, podrían dañar el grano por proliferación de hongos. Los granos son dejados al sol, en patios

o sobre marquesinas, como lo hace la persona de la Figura A.6, o también emplean secadores eléctricos o con combustibles en los cuales reducen el contenido del agua rápidamente.

Figura A.6. Secando granos de café bajo el sol. Tomada de [21]

El resultado del proceso de beneficio es el café pergamino o también conocido como café oro, que será empaquetado y almacenado, conservando las buenas prácticas para evitar la aparición de defectos. La temperatura de almacenamiento no debe ser superior a los 20 °C y una humedad relativa del 65%.

La calidad del café

Un producto alimenticio es considerado de calidad cuando tiene un conjunto de cualidades valoradas o aceptadas por los consumidores como buenas prácticas de manufactura, cualidades nutricionales, entre otros. En el caso del café, la calidad depende de unas buenas cualidades organolépticas que derivan de los granos de café verde sano, como se observan de la Figura A.7. Ya que en investigaciones donde simularon malas prácticas para obtener granos defectuosos, encontraron que las bebidas hechas con estos granos adquirían aromas y sabores desagradables, e incluso llegaban a ser nada bebible. Por este motivo, es necesario llevar un control de calidad y aplicar buenas prácticas agrícolas y de manufactura en todo el proceso productivo del café, especialmente en postcosecha, donde aumenta la posibilidad de que los granos adquieran algún tipo de defecto [2].

Figura A.7. Granos de café verde sanos o normales. Fuente propia.

Como existen diversos tipos de defectos en el grano de café, estos son clasificados en 2 grupos. El primer grupo contiene a los defectos que dañan todas las cualidades organolépticas del café y en el segundo grupo, se encuentran solo los que afectan una o algunas de las cualidades organolépticas con menor intensidad que el primer grupo [51]. A continuación, detallamos cada uno de los defectos y la causa dentro del proceso productivo del café:

- **Negro Total o Parcial:** Este defecto es reconocido por su color negro o marrón y por su forma marchita, como en la Figura A.8. La causa de este defecto puede ser por la falta de agua durante el desarrollo del fruto, fermentaciones prolongadas, recolecta de cerezas sobremaduradas o grano mal secado.

Figura A.8. Granos con el defecto negro total y parcial. Tomada de [51].

- **Agrio Total o Parcial:** Este defecto es reconocido por su color marrón, rojizo o amarillento, como en la Figura A.9. Cuando el grano es raspado, tiende a tener un olor avinagrado. La causa de este defecto es retrasos entre la

recolección y despulpado, fermentaciones prolongadas, limpieza deficiente de los tanques de fermentación o almacenamiento de café húmedo.

Figura A.9. Granos con el defecto agrio total y parcial. Tomada de [51].

- **Cereza seca:** Este defecto es reconocido porque la pulpa seca generalmente cubre parte o todo el pergamino, como en la Figura A.10. La causa de este defecto es el resultado de un deficiente proceso de despulpado, trilla y de selección o falta de mantenimiento o mal ajuste de la maquinaria.

Figura A.10. Granos con el defecto cereza seca. Tomada de [51].

- **Daño por hongos o Cardenillo:** Este defecto es reconocido porque el grano es atacado por hongos, lo cual lleva que sea recubierto de polvillo amarillo o amarillo rojizo, como en la Figura A.11. La causa de este defecto es por fermentaciones prolongadas, interrupciones en el proceso de lavado o almacenamiento con humedad.

Figura A.11. Granos con el defecto daño por hongos o cardenillo. Tomada de [51].

- **Materia Extraña:** Este defecto incluye todo objeto no originario del café encontrado en el café verde, tal como piedras, palos, clavos, etc. Como en la Figura A.12. La causa de este defecto es la falta de cuidado en la recolección, limpieza de los patios de secado, etc.

Figura A.12. Granos con el defecto de materia extraña. Tomada de [51].

- **Brocado severo o leve:** Este defecto generalmente presenta pequeños agujeros en la superficie del grano, como en la Figura A.13. También puede verse en la superficie del fruto a la hora de recolectar. La causa de este defecto son ataques de insectos como el gorgojo y la broca.

Figura A.13. Granos con el defecto brocado severo y leve. Tomada de [51].

- **Pergamino:** Este defecto es reconocido porque el grano verde está cubierto parcial o totalmente por el pergamino, como en la Figura A.14. La causa de este defecto es el desajuste de la máquina trilladora.

Figura A.14. Granos con el defecto pergamino. Tomada de [51].

- **Flotador:** Este defecto es reconocido porque son extremadamente blancos y decolorados, como en la Figura A.15. La causa de este defecto es el mal secado del café o malas condiciones de almacenamiento.

Figura A.15. Granos con el defecto flotador. Tomada de [51].

- **Inmaduro y/o paloteado:** Este defecto es reconocido por tener una superficie rugosa y de tamaño pequeño, como en la Figura A.16. La causa de este defecto es la recolección de frutos verdes, cultivo en zonas marginales bajas, falta de fertilización de los cafetales o cultivar en lotes paloteados por diversas causas.

Figura A.16. Granos con el defecto inmaduro o paloteado. Tomada de [51].

- **Averanado o arrugado:** Este defecto es reconocido por ser granos pequeños de baja densidad, malformados y con superficie arrugada, como

en la Figura A.17. La causa de este defecto es la falta de agua o sequía durante el desarrollo del grano. Las plantas en mal estado de salud o indebidamente fertilizadas.

Figura A.17. Granos con el defecto averanado. Tomada de [51].

- **Conchas:** Este defecto es reconocido por las malformaciones que consisten en dos partes, que por fricción o golpes generalmente están separados, como en la Figura A.18. La parte externa tiene la forma de una concha de mar y la parte interna tiene forma cónica o cilíndrica. La causa de este defecto es debido a factores genéticos del árbol.

Figura A.18. Granos con el defecto concha. Tomada de [51].

- **Partido/Mordido/Cortado:** Este defecto es reconocido porque la parte mordida o cortada generalmente presentan una coloración rojiza oscura, como en la Figura A.19, debido a una oxidación del área cortada. Está cortada puede ser inicio de actividad bacteriana, fermentaciones y formación de hongos. La causa de este defecto se puede producir durante los procesos de despulpado y trilla, por mal ajuste o falta de calibración de los equipos, causan excesiva fricción o presión al grano.

Figura A.19. Granos con el defecto partido, mordido o cortado. Tomada de [51].

- **Cáscara o Pulpa Seca:** Este defecto es reconocido por la cáscara o pulpa, como en la Figura A.20, que son fragmentos secos de cereza de color rojo oscuro. La causa de este defecto aparece cuando el producto no ha sido limpiado correctamente o una mala calibración de la máquina despulpadora, puede resultar en pedazos de pulpa seca.

Figura A.20. Granos con el defecto cáscara o pulpa seca. Tomada de [51].

En la Tabla A.1, resumimos cada uno de los defectos, en los grupos y sus causas correspondientes.

Grupos	Nombre del defecto	Causa
Grupo 1	Negro	+Falta de agua durante el desarrollo del fruto. +Fermentaciones prolongadas. +Recolecta de cerezas sobremaduradas o granos mal secados.
	Agrio	+Retrasos entre la recolección y despulpado. +Fermentaciones prolongadas. +Limpieza deficiente de los tanques de fermentación. +Almacenamiento de café húmedo.
	Cereza Seca	+Deficiente proceso de despulpado, trilla y de selección. +Falta de mantenimiento o mal ajuste de la maquinaria.

	Daño por Hongos	+Fermentaciones prolongadas. +Interrupciones en el proceso de lavado. +Almacenamiento en húmedo.
	Materia Extraña	+Falta de cuidado en la recolección. +Limpieza de los patios de secado.
	Brocado Severo	+Ataques de insectos como el gorgojo y la broca.
Grupo 2	Negro Parcial	+Falta de agua durante el desarrollo del fruto. +Fermentaciones prolongadas. +Recolecta de cerezas sobre maduras o mal secado del grano.
	Agrio Parcial	+Retrasos entre la recolección y despulpado. +Fermentaciones prolongadas. +Limpieza deficiente de los tanques de fermentación. +Almacenamiento de café húmedo.
	Pergamino	+Desajuste de la máquina trilladora
	Flotador	+Mal secado del café. +Malas condiciones de almacenamiento.
	Inmaduro	+Recolección de frutos verdes. +Cultivo en zonas marginales bajas. +Falta de fertilización de los cafetales. +Cultivar en lotes paloteados.
	Averanado o Arrugado	+Falta de agua o sequía durante el desarrollo del grano. +Las plantas en mal estado de salud. +Indebidamente fertilizadas.
	Conchas	+Factores genéticos del árbol.
	Partido/Mordido/Cortado	+Mal ajuste o falta de calibración de los equipos de despulpado y trilla.
	Cáscara o Pulpa Seca	+El producto no ha sido limpiado correctamente. +Una mala calibración de la máquina despulpadora
	Brocado Leve	+Ataques de insectos como el gorgojo y la broca.

Tabla A.1. Defectos en los granos de café. Construida a partir de [51].

A.2. Procesamiento digital de imágenes

El procesamiento de imágenes es un conjunto de técnicas aplicadas a una imagen para resaltar características de interés y/o extraer información representativa de un objeto o zona de interés, en otras palabras, es la implementación de algoritmos que se aplican a una imagen para disminuir el ruido, mejorar el contraste, rotar, extraer regiones de interés, entre otras [22].

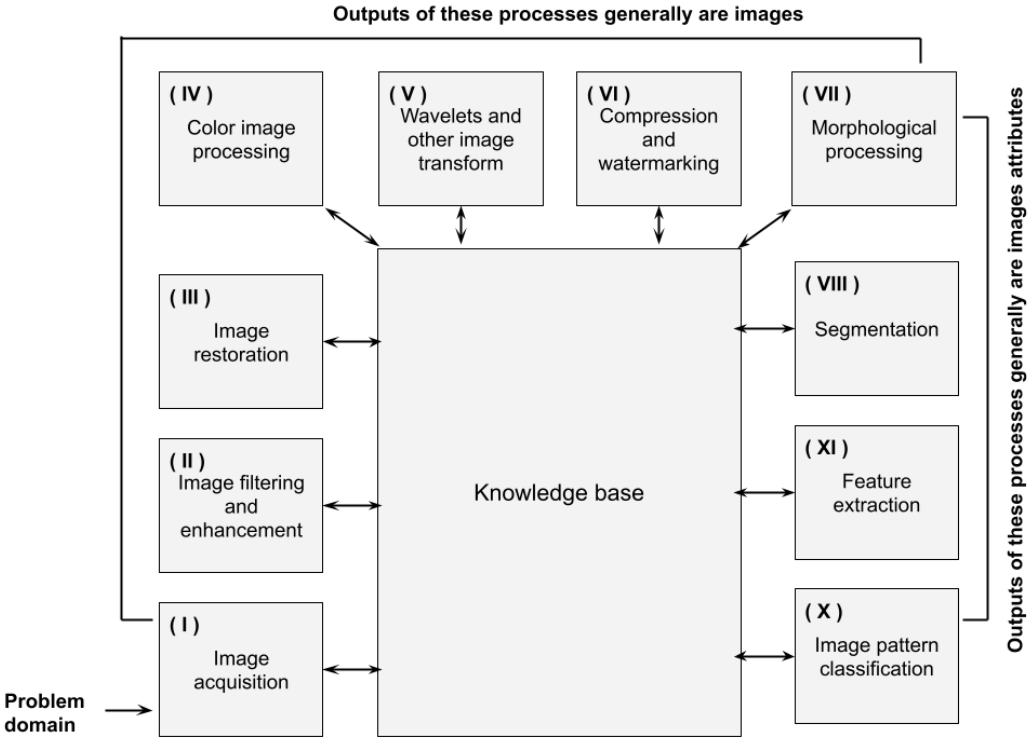


Figura A.21. Fases en el procesamiento digital de imágenes. Imagen adaptada de [227

En la Figura A.21, están contemplados los procesos fundamentales que desarrolla el procesamiento de imágenes. Estos son [17]:

I. **Adquisición**, es la obtención de imágenes o videos en forma digital.

II. **Filtrado y mejora**, es la mejora subjetiva, que es realizado a una imagen mediante filtros o modificaciones.

III. **Restauración**, es la mejora objetiva a una imagen, según modelos matemáticos.

IV. **Procesamiento del color**, es el procesamiento y análisis del dominio de color de la imagen.

V. **Wavelet y transformación**, es el procesamiento de la resolución de una imagen mediante transformada de Fourier.

VI. **Compresión**, son técnicas para reducir el almacenamiento que ocupa una imagen.

VII. **Procesamiento morfológico**, son las técnicas para extraer formas o componentes que son encontrados en una imagen.

VIII. **Segmentación**, es la extracción, según umbrales, para dividir la imagen en regiones.

IX. **Extracción de características**, es la detección y descripción de los atributos de una imagen.

X. **Clasificación de patrones**, predicción de una etiqueta según su descripción.

Para resumir este proceso en pasos simples, se ha llegado al siguiente diagrama, como se muestra en la Figura A.22:

Figura A.22. Fases resumidas del procesamiento digital de imágenes. Adaptada de [22].

Adquisición de imágenes y composición de una imagen

Una imagen es una representación visual bidimensional de píxeles, donde su valor equivale a la información de su tono o luminosidad. Una imagen digital puede ser representada como un arreglo o una matriz de M x N. En una imagen con profundidad de 1 bit tiene solo dos valores de píxeles posibles: 1 (blanco) y 0 (negro). En una imagen en escala de grises, son representadas en 8 bits, o sea que tiene 2^8 o 256 valores, donde el 0, es el valor más bajo y representa el color negro y donde es 255, es el valor más alto y representa el color blanco, como se observa en la Figura A.23.

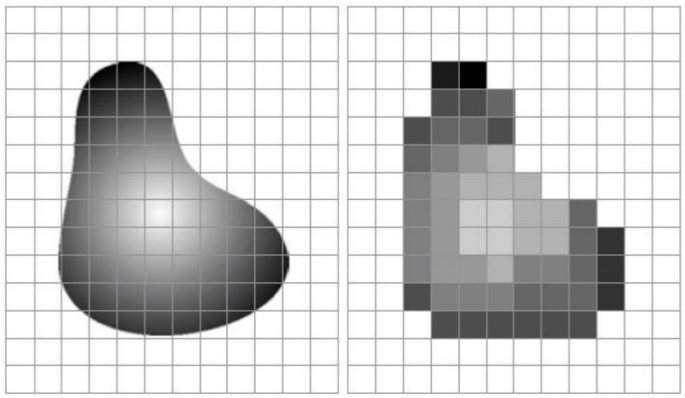


Figura A.23. Muestreo en píxeles de una imagen en escala de grises. Imagen tomada de [52].

Histograma de color

El histograma es una de las formas de entender una imagen a través de un gráfico. En él, es posible obtener información sobre el contraste, brillo, intensidad, etc. El histograma es un gráfico que puede ser representado a través de la distribución de intensidad de los píxeles de la imagen, en otras palabras, es la representación de la frecuencia de una intensidad o píxeles de la imagen. En el eje X, encontramos el rango del valor del píxel, que puede variar de 0 a 255, y el eje Y, la cantidad correspondiente en píxeles. En la Figura A.24, observamos una imagen en escala de grises y su histograma.

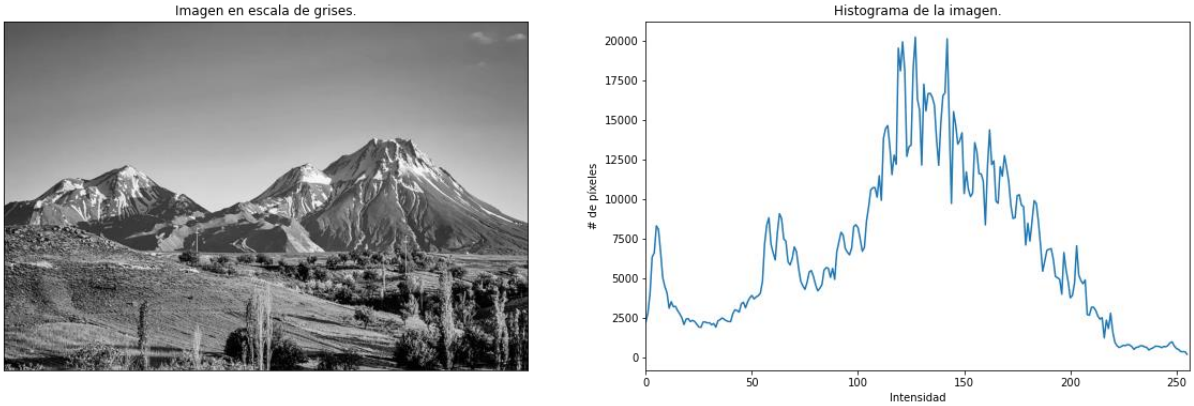


Figura A.24. Histograma de una imagen en escala de grises. Fuente propia.

Color y espacios de color

El color es una de las características más usadas en el procesamiento de imágenes, ya que este permite identificar rasgos específicos de objetos en una escena. El ojo humano capta las ondas electromagnéticas que se reflejan en el objeto y envía señales al cerebro, que luego se convierten en información del objeto observado. En imágenes digitales el proceso es muy similar, el sensor dentro de los dispositivos

ópticos convierte la luz en información o matrices de píxeles, para luego ser almacenada en memoria.

Para una mejor comprensión, son definidos algunos conceptos:

- ➔ **Tono o matiz:** Es el estado puro del color, sin mezclar el color blanco o negro.
- ➔ **Brillo o Luminosidad:** Es la cantidad de luz presente en el color.
- ➔ **Saturación:** Es la cantidad del color presente en la imagen.

En el procesamiento digital de imágenes, son utilizados varios espacios de color para representar una misma imagen, entre ellos esta RGB, CMYK, HSV, etc. Donde es posible crear, representar o visualizar cualquier color. El ser humano puede distinguir los colores dependiendo de la variación del tono, brillo y saturación [22].

- ➔ **Espacio de color RGB**

En el espacio RGB, cada color aparece en sus componentes espectrales *Red*, *Green* y *Blue*. Este modelo está basado en un sistema de coordenadas cartesianas. La representación de los colores de interés es representada en un cubo, como son mostrados en la Figura A.25. Donde los valores principales RGB están en tres esquinas; los colores secundarios *Cyan*, *Magenta* y *Yellow* están en otras tres esquinas; el negro está en el origen; el blanco está en la esquina más alejada del origen y la escala de grises, está extendida del negro al blanco, en la diagonal del cubo.

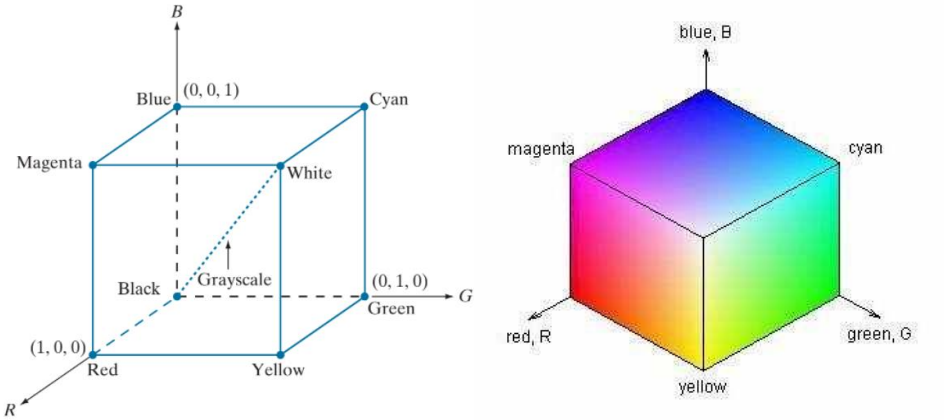


Figura A.25. Espacio de color RGB. Tomadas de [17] y [58]

Los diferentes colores en este modelo son puntos, donde es posible colocar sobre o dentro del cubo y están definidos por vectores que están extendidos desde el punto de origen. En la Figura A.25, Los valores R, G y B están normalizados y representados en el rango [0, 1]. Pero normalmente, estos valores se encuentran en el rango [0, 255], por ejemplo, el blanco estaría en el punto [255, 255, 255] en el cubo. Las imágenes representadas en este espacio de color están compuestas de tres imágenes base, una para cada color primario. Cuando las imágenes son renderizadas en un monitor o pantalla RGB, estas tres imágenes son combinadas en una, para producir una imagen de color compuesta.

➔ Espacio de color HSV

En el espacio de color HSV, como es nombrado anteriormente, está basado en el modo de percepción del ser humano. Este es caracterizado por el color en función de tono (*Hue*), saturación (*Saturation*) y brillo (*intensity*). En este espacio, es posible representar el tono o color puro de un color, sin depender de la intensidad lumínica. La conversión del canal RGB a HSV, se consigue a través de las siguientes fórmulas [22]:

$$H = \{ \frac{\theta}{360 - \theta} \quad \frac{si\ B \leq G}{si\ B > G} \text{ donde } \theta = cos^{-1}\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-G)+(R-B)}}\}$$

$$S = 1 - \frac{3}{(R+G+B)}[min(R,G,B)]$$

$$V = max(R,G,B)$$

Así como el espacio de color RGB está representado como un cubo, el espacio HSV está representado como un cono, como es mostrado en la Figura A.26.

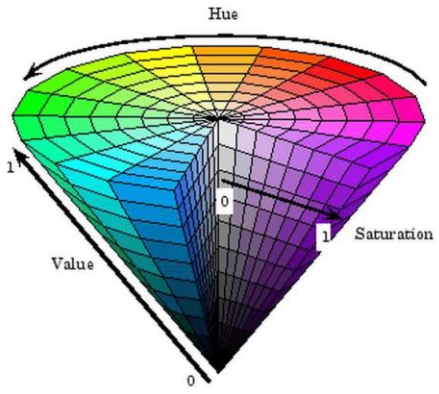


Figura A.26. Espacio de color HSV. Tomada de [59].

En el cono, el color lo define el tono (H) el cual se representa en grados (rojo es 0° o 360°, verde es 120° y azul es 240°). La saturación (S), está representado en porcentaje (sin color es 0% y color puro es 100%) e indica la distancia al centro del cono y la intensidad. El valor (V), está representado en porcentaje (negro es 0% y brillo total del color es 100%) e indica la altura del cono.

Preprocesamiento

El preprocesamiento consiste en revisar si las imágenes obtenidas en la etapa de adquisición requieren algún tipo de ajuste, como estandarizar los tamaños de las imágenes, y/o reparar algunos daños en la información como el ruido, el desenfoque, la mala iluminación, el movimiento, etc. El preprocesamiento de imágenes ofrece una variedad de operaciones matemáticas que se aplican en una imagen para corregir el brillo, resaltar la información mediante umbrales, corregir la distorsión, rotarla, escalarla, utilizar diversos filtros para disminuir o corregir el ruido y resaltar los bordes del objeto de interés [60].

➔ **Normalización de color usando el algoritmo *White Patch***

Es uno de los algoritmos pertenecientes a la teoría de constancia de color [32][33]. Este algoritmo consta de escalar las intensidades de los píxeles de cada canal de acuerdo con un valor determinado. Este valor puede ser el máximo, la media, mediana o percentil de la distribución de los valores de los píxeles de la imagen. La elección de este valor varía dependiendo del contexto y las condiciones de iluminación de la toma. La ecuación que describe esta normalización es:

$$fnorm_i(x,y) = \frac{f_i(x,y)}{I_i}$$

Donde,

- $fnorm_i(x,y)$, es el valor de intensidad de cada píxel de la imagen.
- $f_i(x,y)$, es el valor de cada pixel de la imagen
- I_i, es de intensidad al cual se normalizará la imagen.

Si requerimos hacer una normalización más robusta, es posible usar cada canal que componen la imagen. La ecuación quedaría de la siguiente manera y trabajamos en el canal RGB:

$$fnorm_i(x,y) = (\frac{R_i(x,y)}{IR_i}, \frac{G_i(x,y)}{IG_i}, \frac{B_i(x,y)}{IB_i})$$

Donde,

- $R_i(x,y)$, $G_i(x,y)$ y $B_i(x,y)$, son los valores de los píxeles de cada canal de una imagen en RGB
- IR_i, IG_i y IB_i, son los valores de intensidad normalizados de cada canal de la imagen RGB.

A continuación, mostramos la transformación de unas imágenes con diferentes valores de intensidad para realizar la normalización.

En la Figura A.27, observamos el cambio en la iluminación de la habitación. Para la selección de las intensidades de normalización utilizamos el 99% de los percentiles de cada canal. Estos valores son, 173, 166, 167, respectivamente para R, G y B.

Figura A.27. Comparación de una imagen de una habitación con poca luz (Extraída de [61]) y misma imagen con mayor iluminación empleando *White Patch* (Fuente propia).

En la Figura A.28, empleamos el 25% de los percentiles de cada canal. Estos valores son, 143, 138 y 135, respectivamente para R, G y B.

Figura A.28. Comparación de una imagen de unas moneda y misma imagen con mayor iluminación usando *White Patch*. Fuente propia.

Segmentación

La segmentación es una de las etapas más importante. Es la fase donde son aplicadas diferentes técnicas para separar los elementos de interés del resto del contenido de una imagen. A nivel general, esta separación puede lograrse de 2 formas, una es identificando los cambios abruptos en la intensidad de los píxeles de la imagen, y la segunda es fijando un criterio para obtener regiones de píxeles con características similares. Entre los métodos más conocidos para segmentar se encuentra la segmentación basada en bordes, la segmentación por umbrales y la segmentación por agrupación o clustering [22].

- **Segmentación basada en bordes**

Consiste en identificar la ubicación de los diferentes elementos que componen una imagen. Como se muestra en la Figura A.29, Esto es logrado con imágenes en escala de grises y detectando las diferencias o discontinuidades en su textura, contraste, nivel de gris, etc. El objetivo de este tipo de segmentación es la de realizar una segmentación intermedia que resalta los cambios abruptos de intensidad o bordes de los elementos de una imagen para complementar la segmentación basada en regiones u otro tipo de segmentación que permita obtener los elementos de interés en una imagen final totalmente segmentada. Entre los métodos para detectar bordes podemos encontrar Sobel, Prewitt, Roberts, entre otros [55].

Figura A.29. Ejemplo de segmentación basada en bordes. Fuente propia.

- **Segmentación basada en umbrales**

Consiste en obtener el histograma de una imagen en escala de grises para analizar la distribución de intensidad de la imagen y hallar una intensidad intermedia $\boldsymbol{k}$ que permita separar el objeto o elemento de interés del fondo de la imagen, como lo expone la Figura A.30. En este tipo de segmentación algunas veces es complicado encontrar un $\boldsymbol{k}$ óptimo manualmente, por lo que el método de Otsu es ampliamente utilizado para encontrar dicho parámetro automáticamente [17]. El método consiste

en representar una imagen en $\boldsymbol{L}$ niveles de gris $[1,\cdots,L]$ donde el nivel intermedio k tomará valores en el rango $[1,\cdots,L-1]$ para encontrar cuál de los valores maximiza la varianza entre clases. Por ende, identificar ese k óptimo [62].

El proceso para realizar dicha operación es iterar los diferentes valores de k y en cada caso crear las clases.

$$C_0 = [1,\cdots,k] \text{ y } C_1 = [k+1,\cdots,L]$$

Que representan las intensidades del fondo y del elemento de interés, contar el número de píxeles n_i por cada nivel de gris i y sumar el total de píxeles.

$$N = n_1 + \cdots + n_i$$

Calcular las probabilidades de cada clase.

$$\omega_0 = Pr(C_0) = \sum_{i=1}^{k} \frac{n_i}{N} \text{ y } \omega_0 = Pr(C_1) = \sum_{i=k+1}^{L} \frac{n_i}{N}$$

Calcular las medias de cada clase.

$$v_0 = \sum_{i=1}^{k} i \cdot \frac{n_i}{N \cdot \omega_o} \text{ y } v_1 = \sum_{i=k+1}^{L} i \cdot \frac{n_i}{N \cdot \omega_1}$$

Y por último, obtener el valor de la varianza entre clases.

$$\sigma_B{}^2 = \omega_0 \omega_1 (v_1 - v_0)^2$$

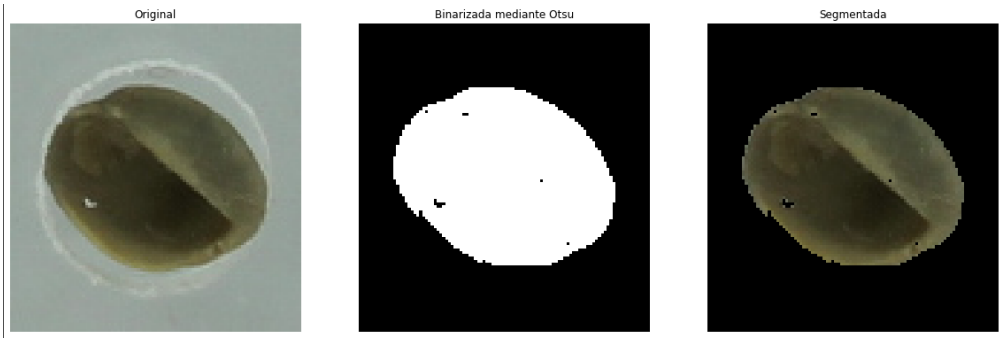


Figura A.30. Ejemplo de segmentación basada en umbrales. Fuente propia.

- **Segmentación basada en agrupación o clustering**

Consiste en buscar grupos de datos que estén relacionados entre sí, en el caso de una imagen consiste en asociar los píxeles que tienen más similitud a partir de variables como el color y la intensidad. El algoritmo clásico para realizar este tipo de agrupación es k-means, el cual consiste en dividir un conjunto de elementos en un número k de clústeres o grupos. Para formar estos grupos, el algoritmo efectúa

un proceso iterativo que comienza eligiendo centroides al azar para cada uno de los grupos, asocia los elementos o puntos más cercanos a dichos centroides, calcula nuevos centroides y asigna los puntos a los nuevos centroides, así continúa hasta que los datos estén estabilizados o terminen el número de iteraciones; cuando ese proceso termina, los grupos o clústeres quedan conformados [17] y se puede extraer el objeto de interés como en la Figura A.31.

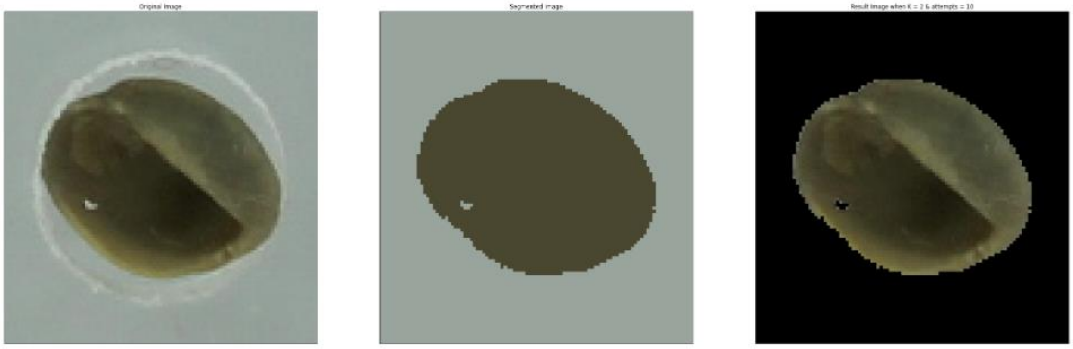


Figura A.31. Ejemplo de segmentación basada en agrupación o clustering. Fuente propia.

Extracción de características

La extracción de características tiene como objetivo describir sensaciones tridimensionales que un ser humano podría tener al observar un objeto. En computador, puede traducirse como los métodos para describir cuantitativamente características sobre los objetos presentes en la imagen. Algunos grupos de estudio comprenden diferentes métodos o algoritmos, estos son:

- ➔ **Extracción de características de color**

Consiste en la extracción de características utilizando los componentes de los canales de color, siendo posible distinguir entre regiones similares u obtener información de la iluminación en la imagen. Estas características pueden extraerse a partir del cálculo del histograma de canal correspondiente. Por ejemplo, si implementamos los canales RGB y HSV, obtendríamos un total de 1536 características, ya que cada canal tiene 256 intensidades.

- ➔ **Extracción de características de textura**

Consiste en la extracción de las características que representan la distribución espacial de los niveles de grises, de este modo pueden evaluar si la superficie de la imagen representa rugosidad o suavidad. Para obtener estas características sugieren emplear las matrices de co-ocurrencia de nivel de gris (GLCM), ya que este método permite obtener 4 matrices que describen la relación de intensidad

entre un píxel de referencia y sus vecinos a 0º (distribución horizontal), 45º (distribución diagonal), 90º (distribución vertical) y 135º (distribución diagonal) [58]. Posteriormente, a cada una de las matrices construidas les calculan las 14 características de textura propuestas por Haralick, luego promedian y obtienen el rango de cada característica para formar las 28 características que serían usadas en un clasificador [59]. Antes de presentar las fórmulas para calcular las características de textura, es necesario tener en cuenta las siguientes notaciones:

- N_g Representa el total de niveles de gris en la imagen.
- $N = N_g - 1$ Como la matriz de co-ocurrencia tendría un tamaño de N_g filas por N_g columnas, el valor de N representa el nivel de gris máximo de la matriz al empezar desde 0.
- $V(i,j)$ Representa el valor en las coordenadas (i,j) de la matriz de co-ocurrencia, donde i es la fila que representa el píxel de referencia y j la columna que representa el píxel vecino.
- $R = \sum_{i,j=0}^{N} V(i,j)$ Es la suma de todos los valores de la matriz de co-ocurrencia.
- $p(i,j) = \frac{V(i,j)}{R}$ Representa la probabilidad en la posición (i,j) de la matriz de co-ocurrencia normalizada.
- $p_x(i) = \sum_{j=0}^{N} p(i,j)$ Entrada i-ésima de la matriz de probabilidad marginal obtenida al sumar los valores de las filas $p(i,j)$.
- $p_y(j) = \sum_{i=0}^{N} p(i,j)$ Entrada j-ésima de la matriz de probabilidad marginal obtenida al sumar los valores de las columnas $p(i,j)$.
- $\upsilon_x = \sum_{i=0}^{N} i \cdot p_x(i)$, $\upsilon_y = \sum_{j=0}^{N} j \cdot p_y(j)$ Medias de las matrices de probabilidad marginal.
- $\sigma_x = \sqrt{\sum_{i=0}^{N}(i-\upsilon_x)^2 p_x(i)}$,$\sigma_y = \sqrt{\sum_{j=0}^{N}(j-\upsilon_y)^2 p_y(j)}$ Desviaciones estándar de las matrices de probabilidad marginal.
- $p_{x+y}(k) = \sum_{i,j=0}^{N} p(i,j);\ k = i + j, k \in [0,\ldots,2N]$ Distribución de probabilidad de la suma de 2 niveles de gris.

- $p_{x-y}(k) = \sum_{i,j=0}^{N} p(i,j);\ k = |i-j|, k \in [0,\ldots,N]$ Distribución de probabilidad de la diferencia de 2 niveles de gris.
- $v_{x-y}(k) = \sum_{k=0}^{N} k \cdot p_{x-y}(k)$ Media de la distribución de probabilidad de la diferencia de 2 niveles de gris.
- $v_{x+y}(k) = \sum_{k=0}^{N} k \cdot p_{x+y}(k)$ Media de la distribución de probabilidad de la suma de 2 niveles de gris.
- $HX = -\sum_{i=0}^{N} p_x(i) \cdot log_2(p_x(i))$ $HY = -\sum_{j=0}^{N} p_y(j) \cdot log_2(p_y(j))$ Entropías marginales.
- $HXY = -\sum_{i,j=0}^{N} p(i,j) \cdot log_2(p(i,j))$ Entropía conjunta
- $HXY1 = -\sum_{i,j=0}^{N} p(i,j) \cdot log_2(p_x(i) \cdot p_y(j))$
- $HXY2 = -\sum_{i,j=0}^{N} p_x(i) \cdot p_y(j) \cdot log_2(p_x(i) \cdot p_y(j))$
- $Q(i,j) = \sum_{k=0}^{N} \frac{p(i,k) \cdot p(j,k)}{p_x(i) \cdot p_y(k)}$

A continuación, listamos las características de textura propuestas:

1. **Segundo momento angular**

$$f_1 = \sum_{i,j=0}^{N} (p(i,j))^2$$

2. **Contraste**

$$f_2 = \sum_{k=0}^{N} k^2 \cdot p_{x-y}(k)$$

3. **Correlación**

$$f_3 = \frac{\sum_{i,j=0}^{N} (i \cdot j) \cdot p(i,j) - v_x v_y}{\sigma_x \sigma_y}$$

4. **Suma de cuadrados (Varianza)**

$$f_4 = \sum_{i,j=0}^{N} (i - \upsilon)^2 p(i,j)$$

Como Haralick no es claro de cómo obtener υ [64], realizamos un análisis a la ecuación y encontramos lo siguiente:

La sección $(i - \upsilon)^2$, al no depender de j, la ecuación puede reescribirse como,

$$f_4 = \sum_{i=0}^{N} (i - \upsilon)^2 \sum_{j=0}^{N} p(i,j).$$

Identificamos que la sección

$$\sum_{j=0}^{N} p(i,j)$$

es la definición de $p_x(i)$, de este modo reescribiendo de nuevo la ecuación quedaría como

$$f_4 = \sum_{i=0}^{N} (i - \upsilon)^2 p_x(i).$$

Ahora, teniendo en cuenta que la varianza de una variable aleatoria discreta X está dada por la ecuación [43]

$$V[X] = E[(X - E[X])^2] = \sum_{x \epsilon R_x} (x - E[X])^2 p_x(x)$$

Asociamos cada parte de esta definición con la ecuación descrita anteriormente y encontramos que $\upsilon = E[X]$, de este modo, en nuestro estudio la ecuación para calcular esta característica estará dada por la siguiente ecuación:

$$f_4 = \sum_{i=0}^{N} (i - \upsilon_x)^2 p_x(i)$$

5. **Momento de diferencia inversa**

$$f_5 = \sum_{i,j=0}^{N} \frac{p(i,j)}{1 + (i - j)^2}$$

6. **Suma Promedio**

$$f_6 = \sum_{k=0}^{2N} k \cdot p_{x+y}(k)$$

7. **Suma de varianza**

$$f_7 = \sum_{k=0}^{2N} (k - f_6)^2 \cdot p_{x+y}(k)$$

Por la definición de varianza identificamos que Haralick utiliza la característica f_8 en lugar de utilizar la $f_6$que sería la correcta [60].

8. **Suma de entropía**

$$f_8 = -\sum_{k=0}^{2N} p_{x+y}(k) \cdot log_2(p_{x+y}(k))$$

9. **Entropía**

$$f_9 = HXY$$

10. **Diferencia de varianza**

$$f_{10} = \sum_{k=0}^{N} (k - v_{x-y}(k))^2 p_{x-y}(k)$$

11. **Diferencia de entropía**

$$f_{11} = -\sum_{k=0}^{N} p_{x-y}(k) \cdot log_2(p_{x-y}(k))$$

12. **Medida de información de correlación 1:**

$$f_{12} = \frac{HXY - HXY1}{max(HX, HY)}$$

13. **Medida de información de correlación 2:**

$$f_{13} = \sqrt{1 - e^{-2 \cdot (HXY2 - HXY)}}$$

14. **Máximo coeficiente de correlación:**

$$f_{14} = \sqrt{2^{do}\ mayor\ valor\ propio\ de\ Q}$$

➔ **Extracción de características morfológicas**

Consiste en la extracción de características analizando la estructura o morfología del objeto presente en la imagen [61]. Esto ayuda a determinar aspectos físicos de la estructura interna del objeto. En la imagen, se usan los píxeles para calcular las siguientes características:

1. **Área:** Número total de píxeles del objeto en la imagen.
2. **Área bbox:** Número total de píxeles del cuadro delimitador.
3. **Área convexa:** Número total de píxeles de la imagen convexa o el polígono más pequeño que encierra el objeto en la imagen.
4. **Longitud eje mayor:** Longitud en píxeles del eje mayor de la elipse según el objeto en la imagen.
5. **Longitud eje menor:** Longitud en píxeles del eje menor de la elipse según el objeto en la imagen.
6. **Excentricidad:** Excentricidad de la elipse según el objeto en la imagen. La excentricidad es la relación entre la distancia focal. El valor está comprendido entre el intervalo [0, 1]. Cuando es 0, la elipse se convierte en un círculo.
7. **Área de diámetro equivalente:** Diámetro de un círculo con la misma área que la región.
8. ***Extent*:** Proporción de píxeles en la región (área) a píxeles en el cuadro delimitador (bbox).
9. **Perímetro de crofton:** Perímetro aproximado en píxeles. Calculado por medio de la fórmula de Crofton.
10. **Solidez:** Proporción de píxeles en la región (área) a píxeles en el área convexa.

A.3. Algoritmos de clasificación

La clasificación es aquella acción que permite organizar o agrupar objetos con características en común; por consiguiente, los algoritmos de clasificación permiten automatizar dicho proceso sin la intervención humana. Estas técnicas agrupan los algoritmos de aprendizaje supervisado, donde encontramos métodos de regresión y clasificación; no supervisado, donde encontramos el *clustering*; y por último, el aprendizaje por reforzamiento [23].

Gracias a la amplia gama de posibilidades que ofrecen. Estos algoritmos son cada vez más usados en diversas áreas del conocimiento. Algunos ejemplos de ello son, el análisis de atributos de una planta para identificar alguna enfermedad presente. La revisión de negociaciones laborales para determinar si un contrato es aceptable o inaceptable. O el análisis para identificar el comportamiento de compra de clientes potenciales [23].

A continuación, detallamos algunos algoritmos de aprendizaje supervisado más usados para la tarea de clasificación:

- **Máquinas de vectores de soporte o *Support vector machine* (*SVM*) [62]**

Las máquinas de vectores de soporte, es una técnica de aprendizaje automático que busca encontrar la mejor separación posible entre dos grupos de datos. A esta separación la conocemos como, hiperplano, que es un subespacio de $n-1$ dimensiones. Por ejemplo, un espacio de dos dimensiones tiene un hiperplano de 1 dimensión, es decir, una recta. Normalmente, los problemas de aprendizaje automático tienen muchas dimensiones. Así que en vez de encontrar una recta, el *SVM* encuentra el hiperplano que maximiza el margen de separación entre clases.

La ecuación o definición matemática de un hiperplano es la siguiente:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n = 0$$

Cuando el punto

$$X = (x_1, x_2, \ldots, x_n)$$

Termina a un lado o al otro del hiperplano, *SVM* lo clasifica dependiendo si satisface la ecuación de la recta. Para saber en qué lado del hiperplano está clasificado, solo es cuestión de calcular el signo resultante.

Ahora bien, para encontrar el hiperplano que separa ambos grupos de datos, como en la Figura A.32, buscamos que la distancia entre el hiperplano y las observaciones sean la más grande posible. A esto le conocemos como, margen máximo. Aunque esta técnica tiene sentido, no es posible aplicarla, ya que habría infinitas formas de separar ambos grupos. Para resolver este problema, utilizamos algoritmos de optimización para encontrar el hiperplano óptimo de separación.

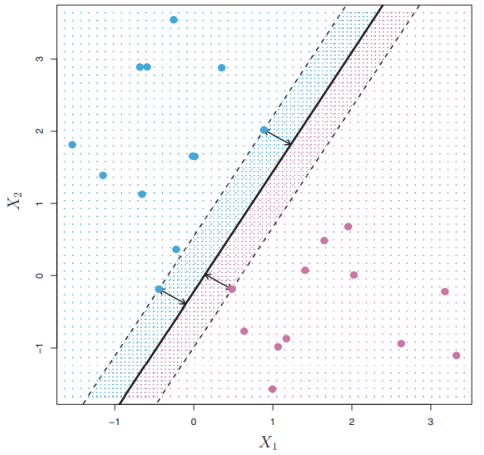


Figura A.32. Vectores de soporte en un clasificador SVM. Tomada de [62].

En la mayoría de los casos, los datos no pueden separarse linealmente de manera perfecta, por lo que no existe un hiperplano que logre separarlos. Una manera de resolver este problema es añadiendo un hiperpárametro C, que controla el número o severidad de las violaciones que sobrepasan el margen de separación, en otras palabras, permite que cometan unos pocos errores, como observamos en la Figura A.33. Si $C \approx \infty$, no permitiría ninguna violación del margen, y cuando $C \approx 0$, se penalizará los errores y más observaciones pueden estar en el lado incorrecto, tanto margen como del hiperplano.

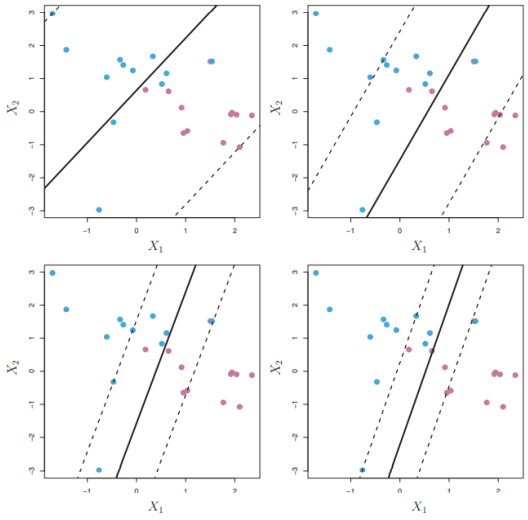


Figura A.33. Ajuste del clasificador SVM. Tomada de [62].

Hasta el momento, únicamente hablamos del *SVM* lineal. Si el límite de separación de los datos no es lineal, la capacidad de clasificación decae drásticamente. Para resolver este problema de la no linealidad, utilizamos la técnica de aumentar la dimensionalidad, o también conocida como, truco del kernel, como observamos en

la Figura A.34. Este busca transformar un espacio de datos a otro para encontrar la solución óptima que logre acomodarse a los datos.

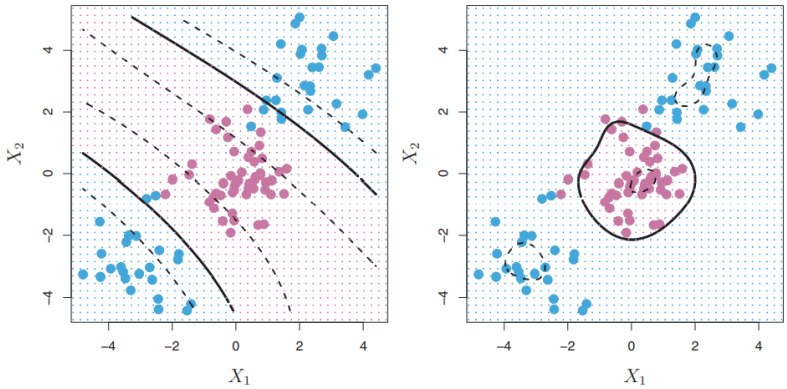


Figura A.34. De izquierda a derecha, clasificador SVM con kernel polinómico y clasificador SVM con kernel gaussiano. Tomado de [62].

Existen varios, pero los más utilizados son:

Kernel lineal: Se comporta como un *SVM* lineal. Su espacio de transformación es el siguiente:

$$K(x, x') = x \cdot x'$$

Kernel polinómico: Con $d = 1$ y $C = 0$, tendrá un comportamiento similar al *SVM* lineal. Con $d > 1$, tiende a generar límites de decisión no lineales a medida que aumenta. Su espacio de transformación es el siguiente:

$$K(x, x') = (x \cdot x' + c)^d$$

Kernel Gaussiano o *Gaussian Kernel* (*RBF*): El valor de γ controla el comportamiento del kernel, cuando el valor es pequeño tiende comportarse como un SVM lineal, a medida que aumenta su valor, la flexibilidad del modelo lo hace. Su espacio de transformación es el siguiente:

$$K(x, x') = e^{(-\gamma||x-x'||)^2}$$

- **K-Vecinos más cercanos o *K-Nearest Neighbors* (KNN) [62]**

K-Vecinos más cercanos, es la técnica más simple de los algoritmos de clasificación, calcula la distancia o proximidad de un nuevo punto de datos entre los demás, como observamos en la Figura A.35. En pocas palabras, busca determinar la clase a la que pertenece el nuevo punto dependiendo del número K vecinos

similarmente cercanos. El cálculo de la distancia puede ser de cualquier tipo, por ejemplo, euclidiana, Manhattan, etc.

Para seleccionar el K óptimo, toca ejecutar el algoritmo KNN múltiples veces con diferentes valores de K y elegir el K que reduce el número de errores, manteniendo la capacidad del algoritmo para hacer predicciones con precisión cuando pasamos datos que no ha visto antes.

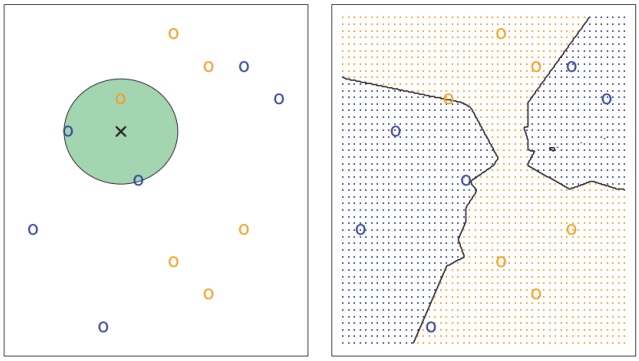


Figura A.35. De izquierda a derecha, observación evaluada por el clasificador KNN con k = 3 y límite de decisión construido. Tomado de [62].

Algunos puntos para considerar son:

1. A medida que el valor de K disminuye a 1, las predicciones del clasificador se vuelven menos estables. Por ejemplo, cuando K = 1, solo consideraría un vecino u observación, para determinar el dato a clasificar.

2. Inversamente, a medida que aumentamos el valor de K, las predicciones se vuelven más estables debido al gran número de observaciones. Por lo tanto, es más probable que hagan predicciones más precisas. Eventualmente, comenzará a presenciarse un número creciente de errores, ya que un número demasiado grande reducirá la capacidad de predicción.

3. Generalmente, usamos valores de K como un número impar, por el hecho de que en algunos casos se necesita un desempate.

- **Árboles de decisión o *Decision Trees* (DT) [23] [62]**

El árbol de decisión es un algoritmo que pertenece a la familia de algoritmos de aprendizaje supervisado, a diferencia de otros algoritmos de aprendizaje, este puede ser utilizado para resolver problemas de regresión y clasificación. La meta de usar un árbol de decisión es crear un modelo entrenado que pueda ser usado para

predecir la clase o valor de la variable objetivo por reglas simples de aprendizaje inferidas por datos previos.

En los árboles de decisión, para predecir una etiqueta de clase para un registro, comenzamos desde el nodo raíz del árbol. Los datos se comparan con el atributo raíz y dependiendo del resultado, siguen en la rama correspondiente a ese valor y continúa con el siguiente nodo, hasta encontrar el resultado. Para entender lo anterior, supongamos que desea salir a jugar un día cualquiera, pero su decisión depende del clima, la temperatura, la humedad y el viento que esté haciendo, por lo que decide esperar 10 días para anotar el comportamiento de dichos factores durante ese tiempo. Al finalizar tiene los datos presentados en la Tabla A.2.

Día	Clima	Temperatura	Humedad	Viento	¿Salgo a jugar?
1	Soleado	Caluroso	Alta	Débil	No
2	Nublado	Caluroso	Alta	Débil	Sí
3	Soleado	Templado	Normal	Fuerte	Sí
4	Nublado	Templado	Alta	Fuerte	Sí
5	Lluvioso	Templado	Alta	Fuerte	No
6	Lluvioso	Frío	Normal	Fuerte	No
7	Lluvioso	Templado	Alta	Débil	Sí
8	Soleado	Caluroso	Alta	Fuerte	No
9	Nublado	Caluroso	Normal	Débil	Sí
10	Lluvioso	Templado	Alta	Fuerte	No

Tabla A.2. Datos anotados del clima de 10 días para construcción del árbol decisión. Adaptado de [23].

Hasta este momento podríamos usar la tabla para decidir si salir o no, pero si en el onceavo día observa un patrón que no está en la tabla, sería difícil decidir. De este modo, un árbol de decisión sería una buena forma de representar estos, ya que este tiene en cuenta todas las rutas posibles que lo pueden conducir a la decisión final. Así que para construir el árbol de decisión de este ejemplo debemos seguir los siguientes pasos:

1. Elegir el criterio de decisión para los nodos del árbol de decisión. Puede elegir entre los dos criterios más conocidos, la ganancia de entropía o el índice Gini.

2. Obtener el nodo raíz. Esto consiste en construir un árbol que agrupe los estados de cada clase con la decisión resultante y calcular su respectivo valor del criterio elegido. Los árboles construidos para el ejemplo y los valores obtenidos para ambos criterios son mostrados en la Figura A.36.

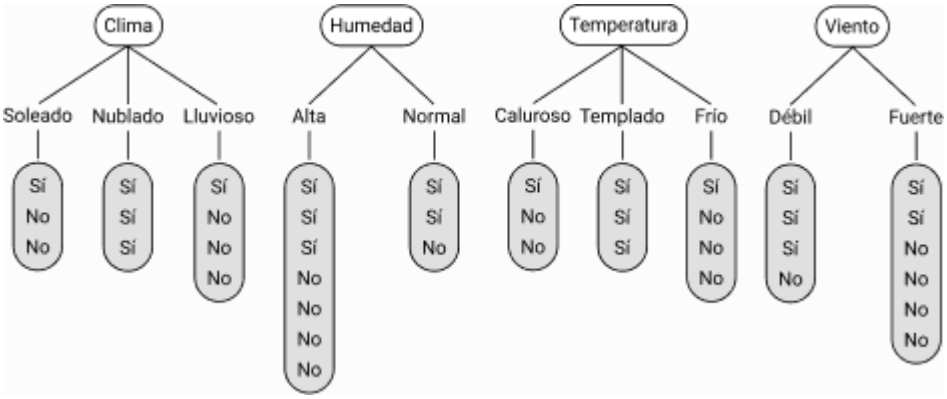


Figura A.36. Adquisición del nodo raíz. Adaptado de [23].

3. Valores para ganancia de entropía. Para obtener estos valores primero calculamos la entropía de cada estado de la clase mediante la ecuación

$$E(S) = -\sum_{i=1} p_i \cdot log_2(p_i)$$

donde S es el estado a analizar y p_i es la probabilidad de la decisión i-ésima del estado S. Segundo, obtenemos la entropía de clase mediante la ecuación

$$E(C) = \sum_{s=1}^{s} p_s \cdot E(S)$$

donde p_s es la probabilidad del estado y $E(S)$ la entropía del estado.

Tercero, usando la misma ecuación calculamos la entropía general y aplicamos la ecuación de ganancia $Gain(S) = E(General) - E(S)$. Por último, el nodo principal será aquella ganancia que tenga el valor más alto. En el caso del ejemplo obtuvimos los siguientes valores de entropía y ganancia:

- ➔ $E(General[5\ Sí, 5\ No]) = -\frac{5}{10} log_2(\frac{5}{10}) - \frac{5}{10} log_2(\frac{5}{10}) = 1$
- ➔ $E(clima_{soleado}[1\ Sí, 2\ No]) = -\frac{1}{3} \cdot log_2(\frac{1}{3}) - \frac{2}{3} \cdot log_2(\frac{2}{3}) = 0.9183$
- ➔ $E(clima_{nublado}[3, 0]) = -\frac{3}{3} \cdot log_2(\frac{3}{3}) - 0 = 0$

- ➔ $E(clima_{lluvioso}[1,3]) = -\frac{1}{4} \cdot log_2(\frac{1}{4}) - \frac{3}{4} \cdot log_2(\frac{3}{4}) = 0.8115$
- ➔ $E(clima[3,3,4]) = \frac{3}{10} \cdot 0.9183 + 0 + \frac{4}{10} \cdot 0.8115 = 0.6$
- ➔ $E(humedad_{alta}[3,4]) = 0.9852$
- ➔ $E(humedad_{normal}[2,1]) = 0.9183$
- ➔ $E(humedad[7,3]) = \frac{7}{10} \cdot 0.9852 + \frac{3}{10} \cdot 0.9183 = 0.9652$
- ➔ $E(temperatura_{caluroso}[2,2]) = 1$
- ➔ $E(temperatura_{templado}[3,2]) = 0.8844$
- ➔ $E(temperatura_{frío}[1,0]) = 0$
- ➔ $E(temperatura[4,5,1]) = 0.8422$
- ➔ $E(viento_{débil}[3,1]) = 0.7013$
- ➔ $E(viento_{fuerte}[2,4]) = 0.9183$
- ➔ $E(humedad[4,6]) = 0.8315$
- ➔ $Gain(clima) = E(General) - E(clima) = 1 - 0.6 = 0.4$
- ➔ $Gain(humedad) = 1 - 0.9652 = 0.0348$
- ➔ $Gain(temperatura) = 1 - 0.8422 = 0.1578$
- ➔ $Gain(humedad) = 1 - 0.8315 = 0.1685$

Por lo que seleccionamos la clase clima como nodo principal, ya que tuvo la mayor ganancia de entropía, por ende, tiene más información.

4. Valores para índice Gini. En el caso de Gini seguimos un proceso similar a entropía, calculamos dicho índice por cada árbol, pero el nodo raíz será el que tenga el índice Gini más pequeño. La fórmula para calcular el índice Gini por cada estado de clase está dada por

$$Gini\ Index(S) = \sum_{i=1} p_i(1 - p_i)$$

donde p_i es la probabilidad de la decisión i-ésima del estado S; y la fórmula para obtener el índice Gini por clase está dado por la ecuación

$$Gini\ index(C) = \sum_{s=1}^{S} p_s \cdot Gini\ Index(S)$$

donde p_s es la probabilidad del estado S. En el caso del ejemplo obtuvimos los siguientes índices Gini:

- ➔ $Gini\ Index(clima_{soleado}[1\ Sí, 2\ No]) = \frac{1}{3}(1 - \frac{1}{3}) + \frac{2}{3}(1 - \frac{2}{3}) = 0.4444$-
- ➔ $Gini\ Index(clima_{nublado}[3,0]) = \frac{3}{3}(1 - \frac{3}{3}) + 0 = 0$
- ➔ $Gini\ Index(clima_{lluvioso}[1,3]) = 0.375$
- ➔ $Gini\ Index(clima[3,3,4]) = 0.4444 \cdot \frac{3}{10} + 0 \cdot \frac{3}{10} + 0.375 \cdot \frac{4}{10} = 0.2833$
- ➔ $Gini\ Index(humedad_{alta}[3,4]) = 0.4898$
- ➔ $Gini\ Index(humedad_{normal}[2,1]) = 0.4444$
- ➔ $Gini\ Index(humedad[7,3]) = 0.4762$
- ➔ $Gini\ Index(temperatura_{caluroso}[2,2]) = 0.5$
- ➔ $Gini\ Index(temperatura_{templado}[3,2]) = 0.48$
- ➔ $Gini\ Index(temperatura_{frío}[1,0]) = 0$
- ➔ $Gini\ Index(temperatura[4,5,1]) = 0{,}44$
- ➔ $Gini\ Index(viento_{débil}[3,1]) = 0.48$
- ➔ $Gini\ Index(viento_{fuerte}[2,4]) = 0.4444$
- ➔ $Gini\ Index(viento[4,6]) = 0{,}4166$

Con todos los datos obtenidos, de nuevo el clima será el nodo principal al tener el valor más bajo del índice Gini.

5. Obtener los nodos hijos. Con el nodo raíz definido, lo que sigue es usar de nuevo las ecuaciones mencionadas en el anterior punto y obtener los nodos hijos para cada uno de los estados de la clase del nodo raíz. Continuando con el ejemplo, presentamos los datos para cada estado o rama de la clase Clima, en las Tablas A.3, A.4 y A.5.

Datos para soleado				
Día	**Temperatura**	**Humedad**	**Viento**	**¿Salgo a jugar?**
1	Caluroso	Alta	Débil	No
3	Templado	Normal	Fuerte	Sí
8	Caluroso	Alta	Fuerte	No

Tabla A.3. Datos del clima para días soleados. Adaptado de [23].

Datos para nublado				
Día	**Temperatura**	**Humedad**	**Viento**	**¿Salgo a jugar?**
2	Caluroso	Alta	Débil	Sí
3	Templado	Alta	Fuerte	Sí
9	Caluroso	Normal	Débil	Sí

Tabla A.4. Datos del clima para días nublado. Adaptado de [23].

Datos para lluvioso				
Día	**Temperatura**	**Humedad**	**Viento**	**¿Salgo a jugar?**
5	Templado	Alta	Fuerte	No
6	Frío	Normal	Fuerte	No
7	Templado	Alta	Débil	Sí
10	Templado	Alta	Fuerte	No

Tabla A.5. Datos del clima para días lluviosos. Adaptado de [23].

Con los datos presentados se omite el cálculo para temperatura, ya que en ningún caso contempla todas las posibilidades; en la rama “nublado” también omitiría el cálculo de todas las clases restantes, porque independientemente del estado siempre sale a jugar; y para los demás casos calculamos la humedad y el viento para ambos criterios. Al realizar los cálculos usando las ecuaciones de los criterios mencionados anteriormente, nos da como resultado el árbol de decisión presentado en la Figura A.37.

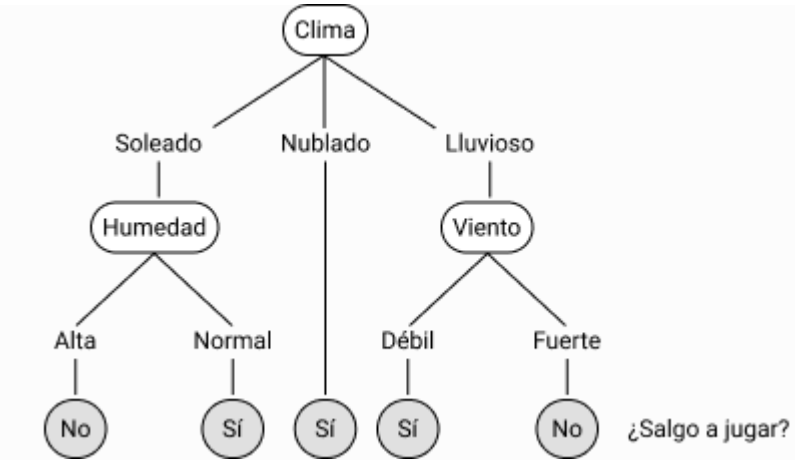


Figura A.37. Árbol de decisión construido. Adaptado de [23].

- **Bosque aleatorio o *Random Forest* (RF) [62]**

Como su nombre lo indica, el bosque aleatorio consiste en entrenar múltiples árboles de decisión individuales que clasifican en conjunto. Un bosque aleatorio puede estar conformado por m árboles de decisión, los cuales pueden ser entrenados mediante bootstrapping que consiste en entrenar cada árbol de decisión con un conjunto de características seleccionadas aleatoriamente del conjunto de datos, dando como resultado árboles diferentes entre sí. El funcionamiento del algoritmo consiste en que cada árbol de decisión dentro del bosque predice una clase y luego suman los resultados, la clase con mayor valor obtenido de la suma, es la clase que clasifica el bosque aleatorio. Por ejemplo, supóngase que tiene un bosque aleatorio conformado por 9 árboles de decisión y desea clasificar si un elemento es de la clase 1 o la clase 0, si 7 árboles clasifican al elemento en la clase 1 y los 2 restantes en la clase 0, el resultado de la clasificación es que el elemento es de la clase 1. A continuación mostramos una representación gráfica del ejemplo en la Figura A.38, los árboles los hemos coloreado de diferentes colores para representar el bootstrapping.

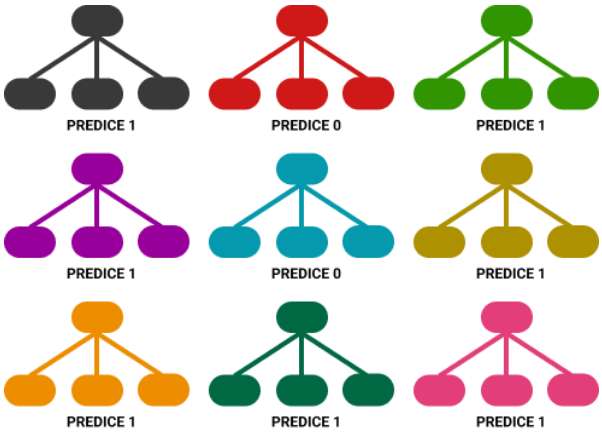


Figura A.38. Ejemplo de un bosque aleatorio. Fuente propia.

Algo importante para que el modelo de bosque aleatorio funcione correctamente y tenga una buena precisión, es que los árboles de decisión que lo conformen deben tener una muy baja correlación, ya que esto asegura que los árboles protejan entre sí de su error individual y no afecte el resultado de los demás.