MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA CALIDAD DEL CAFÉ.

Javier Andrés Suárez Peña

20112196013

Universidad Distrital Francisco José De Caldas

Facultad de Ingeniería

Maestría en Ingeniería Industrial

Diciembre de 2019

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE LA CALIDAD DEL CAFÉ.

Proyecto de Grado para obtener el título de Magister en Ingeniería Industrial

Javier Andrés Suárez Peña

20112196013

Director

Ing. José Ignacio Rodríguez Molano PhD

Codirector

Dr. Ing. Hugo Fabián Lobatón García

Codirector

Mg. Ing. William Camilo Rodríguez Vásquez

Universidad Distrital Francisco José De Caldas

Facultad de Ingeniería

Maestría en Ingeniería Industrial

Diciembre de 2019

AGRADECIMIENTOS

A Dios en primer lugar por permitirme realizar este trabajo y bendecirme en todo momento. A mi esposa e hijas por su apoyo y comprensión durante el tiempo de este trabajo. A mi familia por acompañarme y apoyarme en todo momento. A mis directores de tesis por su acompañamiento, críticas constructivas y aportes teóricos que me permitieron completar este trabajo. A mi jefe laboral inmediato por la compresión y apoyo que facilitaron dedicar tiempo a este trabajo. A la empresa Almacafé que hace parte de la Federación Nacional de Cafeteros de Colombia y particularmente al jefe de la oficina de aseguramiento de calidad de café, Rodrigo Alarcón y a la analista de laboratorio Zulma Betancourt, quienes me apoyaron con la recopilación y análisis de muestras trabajadas en esta investigación y a todos aquellos quienes directa o indirectamente contribuyeron a la realización de este trabajo.

CONTENIDO

LISTADO DE TABLAS

LISTADO DE FIGURAS

RESUMEN

En este trabajo se desarrolla un modelo de aprendizaje automático (ML: Machine Learning) para la caracterización de la calidad del café en Colombia. Con el apoyo de la oficina de calidad de café de Colombia, Almacafé, se realiza un proceso de medición de varios atributos del café verde, en muestras provenientes de diferentes partes de Colombia, como datos de entrada del modelo, luego se realiza el proceso de tostado y preparación de bebida de café para realizar la caracterización de la calidad del café, denotada por un puntaje que se asigna a atributos del café como su aroma, su cuerpo, limpieza de taza entre otros (Datos de salida o etiquetas). Posteriormente, se entrenan los algoritmos de aprendizaje automático escogidos con el conjunto de datos construido, bajo en un enfoque de clasificación y otro de regresión. Se analizan las medidas de desempeño de estos y se hacen los ajustes requeridos. Finalmente, se realiza una validación cruzada del modelo.

PALABRAS CLAVE

Machine learning, calidad del café, SVM, redes neuronales, clasificación, regresión, catación.

INTRODUCCIÓN

El café es uno de los principales productos de exportación del país, y uno de los productos por los que se conoce a Colombia internacionalmente. Se cultiva en diferentes regiones a lo largo de todo el territorio y es reconocido mundialmente por su sabor y frescura, resultantes de climas y topografías propias de cada región (Buencafé, 2019). De acuerdo con la Federación Nacional de Cafeteros (FNC) de Colombia, organismo que representa y agrupa a los caficultores de todo el territorio, la cuidadosa selección de los granos de café durante sus etapas de cosecha, y postcosecha (despulpado, lavado, secado y trilla), aseguran una calidad única.

Dada su importancia para el país y sus procesos llenos de tareas manuales, se vuelve interés del investigador aplicar herramientas computacionales de análisis que permitan proponer mejoras, revelen relaciones o dependencias entre variables, posiblemente arraigadas dentro de esquemas de conocimiento tácito de los expertos, como lo es la actividad de calificación de la calidad del café, una tarea que se requiere de formación y entrenamiento para poder llegar a evaluar una muestra específica.

En el aprendizaje automático (ML: Machine Learning), se entrenan algoritmos para realizar tareas que son de alta complejidad y que podrían con otros métodos, llegar a requerir de una programación muy exhaustiva. En contraste, las técnicas de ML se limitan por la cantidad de datos o ejemplos de entrenamiento con los que se pueda alimentar un modelo en específico. Las aplicaciones del ML son tan variadas como campos del conocimiento existen, sin embargo, existen un grupo de usos comunes o de categorías donde se usan ciertas técnicas en específico y dentro de las cuales las 2 principales son clasificación y predicción (nombrada regresión por algunos autores). Esta investigación se emplearán técnicas de ML para la predicción de la calidad del café, contribuyendo al avance del sector, generando nuevo conocimiento, brindando herramientas para futuras investigaciones y aportando conclusiones que sirvan de base para proponer estrategias que permitan mejorar la competitividad del sector.

1. CAPITULO 1: EL PROBLEMA DE INVESTIGACIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

Colombia es el tercer productor de café en el mundo y el principal productor de café de la variedad arábico (Banco Mundial, 2002), reconocido por ser un café de calidad, que pasa por rigurosos procesos de selección y una calificación final que le permite su definición como café de exportación. Dicha calificación es realizada por catadores de café profesionales, quienes, según el método de certificación, clasifican o dan una puntuación a las diferentes características evaluadas.

Existen diferentes métodos para la calificación del café, dependiendo el país o la empresa tostadora que lidera un mercado en específico. Un ejemplo es el instituto de calidad del café de Estados Unidos, que cuenta con una base de datos donde se aprecia la calificación del café arábico de varios países, mediante una puntuación de 1 a 10 de diversos atributos que sumados dan una calificación final denominada puntuación de taza. Estas calificaciones son dadas por catadores profesionales entrenados para calificar los atributos de una taza de café líquido. De la calificación obtenido de un lote específico de café, la FNC define si el café se puede exportar o se queda para consumo interno. Es precisamente en este punto, donde se evidencia una oportunidad de aplicar técnicas de aprendizaje automático, que reduzcan el tiempo de calificación de un lote de café y reduzcan los costos de esta actividad.

Una primera aproximación a este tema se puede apreciar en el artículo titulado "*A computer vision system for coffee beans classification based on computational intelligence techniques*" (De Oliveira, Leme, Barbosa, & Rodarte, 2015), en el cual los autores desarrollan un modelo capaz de clasificar los granos de café verde mediante las diferencias en su color, usando primero redes neuronales para convertir los colores del espacio RGB al CIE L*a*b. Posteriormente usan un clasificador bayesiano, para clasificar las imágenes de café en cuatro colores blanco, verde, verde azuloso y verde caña. Este estudio se justifica en la importancia que tiene el color para determinación de la calidad de un café, asociada a un mayor valor de mercado.

Otro ejemplo, es un sistema de selección electrónico de café excelso basado en el color mediante procesamiento de imágenes (Ruge, Pinzon, & Moreno, 2012). En este, los autores usan un software para el reconocimiento de defectos de imágenes de granos de café individuales, mediante el conteo de los pixeles de una imagen asociadas a áreas defectuosas, este lo integran a una máquina capaz de separar granos defectuosos de granos buenos. Una de las limitantes de esta aproximación como la mencionan los autores, es que las imágenes requieren de unas condiciones controladas con buena iluminación, adicionalmente el método de segmentación de trabajo fijo y no ajustable empleados por ellos, deja de funcionar si las condiciones bajo las que se parametrizó cambian.

En contraste (Faridah, Parikesit, & Ferdiansjah, 2011), desarrollan un modelo para clasificar imágenes de grupos de grano de café, de acuerdo con el sistema de defectos nacional de Indonesia, el cual los divide en 7 categorías, la primera categoría tiene un número mínimo de defectos y la séptima tiene el mayor número de defectos. Para cada muestra toman 10 imágenes, las procesan para eliminar ruido y extraer las características con las que desean entrenar la red neuronal empleada. Estas características son textura y color. La textura la expresan en términos de la energía, la entropía, el contraste y la homogeneidad. El color la expresan como el valor medio de los colores rojo, verde y azul (Espacio de color RGB). Al final del entrenamiento los autores muestran que el modelo no logra predecir con precisión algunas de las categorías, concluyendo que posiblemente requieren de más desarrollo en la determinación de las características que realmente los representan.

En el campo de la agricultura se han desarrollado estudios mediante el uso de técnicas de aprendizaje automatizado, con el fin de predecir las mejores condiciones para los cultivos analizados. Algunos se centran en factores climáticos, otros en características y composición del suelo (Kouadio, y otros, 2018), el uso de agentes químicos que contrarrestan algún patógeno del cultivo o potencian su crecimiento y varios en el desarrollo de sistemas de reconocimiento visual para la clasificación automatizada de diversos frutos. (Cubero, Aleixos, Moltó, Gómez-Sanchis, & Blasco, 2011) , muestran en su trabajo los avances en reconocimiento visual para la automatización de la inspección y evaluación de calidad de varias frutas y vegetales. En el marco teórico se detalla un poco más los hallazgos de este trabajo.

De acuerdo con (Gomez, Paéz, Buitrago, & Ceballos, 2014). La academia debe jugar un papel fundamental en la capacitación de la población que trabaja en esta actividad, así como contribuir con programas e investigaciones que permitan a los caficultores adquirir conocimientos y aplicar nuevas tecnologías para incrementar la calidad del grano y su productividad, así como detectar necesidades de primera mano. Esta investigación busca caracterizar el café verde y generar un modelo que permita predecir la calidad del café líquido.

1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo predecir la calidad del café líquido mediante un modelo de aprendizaje automático (ML), que use como variables de entrada, características del grano de café verde en Colombia?

1.3 OBJETIVOS

1.3.1 Objetivo general

Definir un modelo de aprendizaje automático que permita, predecir la calidad del café líquido, partiendo de variables medidas en granos de café verde en Colombia.

1.3.2 Objetivos específicos

- Identificar, medir y recopilar datos o imágenes de variables descriptoras de los granos de café verde, para usarlos como datos de entrada del modelo.
- Realizar evaluación tradicional de café líquido, para obtener las calificaciones de las diferentes variables de café, las cuales se usarán como salidas o etiquetas del modelo.
- Entrenar, evaluar y ajustar los algoritmos definidos de aprendizaje automático con los datos recopilados.
- Validar la efectividad del modelo de aprendizaje automático implementándolo, con muestras de café no consideradas en el estudio.

1.4 JUSTIFICACIÓN

La actividad de calificación del café es una tarea realizada por catadores profesionales especializados en reconocer los diferentes atributos del café, sin embargo, para disminuir los errores humanos de las mediciones realizadas, se requiere de métodos, formularios y análisis estadísticos estandarizados, además de experiencia por parte de los catadores en la interpretación de los resultados (Salamanca, 2015), convirtiéndolo en una tarea compleja y costosa. Evidenciando la necesidad de encontrar métodos estadísticos basados en mediciones numéricas de las propiedades fisicoquímicas que puedan ser realizados por un rango más amplio de profesionales, con resultados reproducibles y con menos variación.

Dentro de la revisión del estado del arte, se aprecia el uso de algoritmos de ML para la clasificación de café en alguna de sus formas, según la etapa del proceso productivo en el que se encuentra, no obstante, no se ha encontrado quien emplee técnicas de ML para predecir las etiquetas o salidas de un estado posterior, mediante el análisis de las características de un estado anterior. Un ejemplo de esto sería predecir la calidad del café verde, por medio del análisis de las variables del proceso de beneficio de los granos, o a través del análisis de los atributos de los cerezos de café recolectados. Varios autores justifican sus enfoques en la importancia tácita o en el juicio experto de quienes, por experiencia, saben que la calidad del café final se reduce, cuando en alguna etapa de proceso de cosecha o postcosecha, se mezclan granos buenos y granos con defectos. Sin embargo, no se aprecia quien relacione un nivel específico de defectos con uno de calidad.

Debido a la forma en como está organizada la cadena productiva del café en Colombia, en donde para cada etapa y dependiendo la región, se pueden encontrar diferentes intermediarios quienes agrupan la producción de varios productores, es fácil que se pierda la trazabilidad de la cadena, generando como resultado que sea más fácil evaluar la calidad sobre un lote especifico, perteneciente a una región y a una variedad predominante en la región.

Las técnicas de aprendizaje automático permiten que una máquina sea capaz de aprender, reconocer patrones, características y vínculos entre un conjunto de datos, que muchas veces están dentro del dominio del conocimiento tácito de expertos o artesanos dedicados a una labor específica. Este aprendizaje se logra mediante el uso de datos de

entrenamiento, que relacionan unos datos de entrada con unos resultados de salida. En el mundo se vienen desarrollando diferentes investigaciones como las citadas en el marco teórico que abordan casos similares, en donde la complejidad o la especialidad de una tarea, representan un desafío muy grande de programación con las técnicas tradicionales, convirtiendo al ML en una herramienta que permite emular dichos conocimientos o técnicas y facilitar el uso de máquinas para la realización de tareas complejas.

Lo anterior hace que sea del interés del investigador aplicar estas técnicas de ML a una situación como la descrita respecto a la calidad del café, que permita evidenciar su alcance y las estrategias con las que debe definir y entrenar el modelo objeto de esta investigación.

1.5 HIPÓTESIS.

La calidad del café (líquido) se predecirá mediante un modelo de aprendizaje automático, tomando como entradas características de los granos de café verde y entregando como salida el puntaje estimado en los diferentes atributos definidos.

2. CAPITULO 2: MARCO DE REFERENCIA

2.1 MARCO TEÓRICO

2.1.1 Café de Colombia.

El café de Colombia es del tipo arábico, su arbusto pertenece a la familia de las Rubiáceas, y se considera originario de la parte central de etiopia, datándose a finales del primer milenio en la península arábica, hoy día Yemen (Wikipedia, 2019). Este tipo de arbusto crece mejor entre los 1200 y 2000 metros sobre el nivel del mar (msnm), su temperatura ideal esta entre los 15 y 24 grados centígrados, el clima donde se cultiva debe tener una precipitación entre 1200 a 2500 mililitros por año y el suelo debe ser de un pH bajo, no superior a 7 (Coffe IQ, 2019). En Colombia, por la diversidad de regiones se encuentran cultivos a niveles inferiores a los 1000 msnm y por encima de 2000 msnm (Coffee Quality Institute, 2019).

Los orígenes comerciales del café colombiano se remontan al año 1835, en donde principalmente se cultivaba en los “Santanderes” pertenecientes a la zona este del país y se realiza la primera exportación de 60kg de café (Buencafé, 2019). De acuerdo con el Departamento Nacional de Estadística (2019) (DANE, 2019) las exportaciones de café en Colombia para el 2017 fueron de 710.440 Toneladas métricas y 710.836 para el 2018, representando el 7% y 5% de las exportaciones totales para sus respectivos años, en la figura 1 se puede observar las exportaciones de café históricas para el periodo comprendido entre el año 2000 al 2018.

El café de Colombia se produce por más de 560.000 pequeños productores de café, quienes se encuentran agrupados en la Federación Nacional de Cafeteros (fundada en 1927) y siguen estándares para garantizar la calidad del café; la figura 2 muestra la distribución de los cultivos de café del país por departamentos en miles de hectáreas, entre el periodo del año 2002 al 2018. Los productores recolectan manualmente los granos de café maduros y adicionalmente realizan procesos de selección manual posteriores,

denominados “beneficios” para descartar cualquier grano defectuoso (Federación Nacional de Cafeteros, 2019).

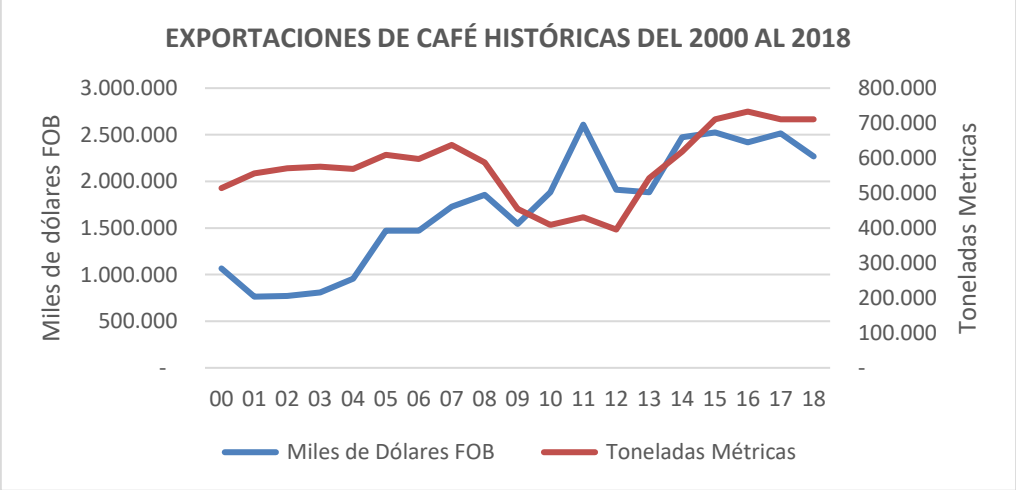


Figura1. Exportaciones de café históricas del 2000 al 2018. Fuente: (DANE, 2019)

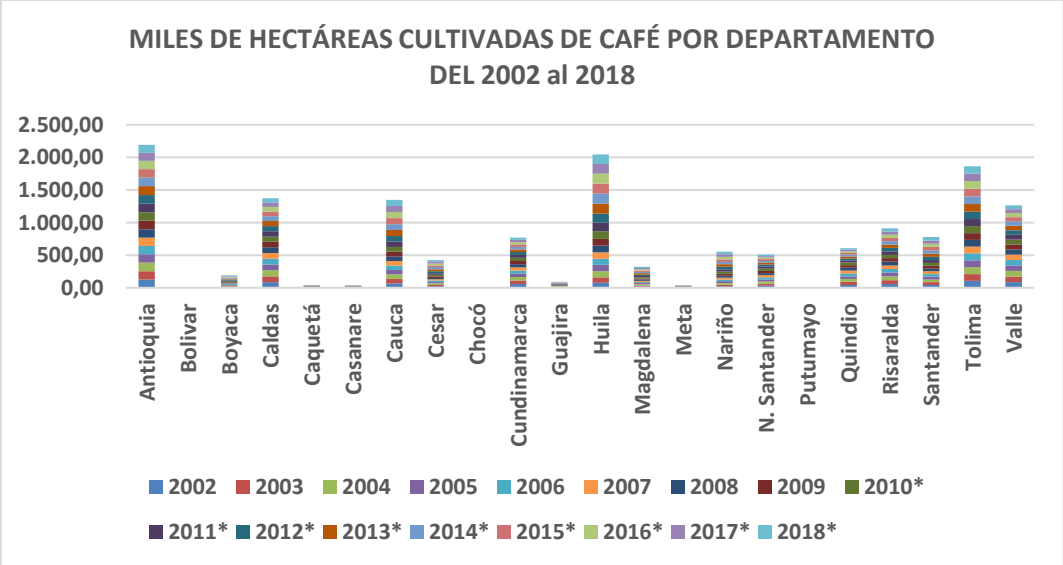


Figura 2. Miles de hectáreas cultivadas de café por departamento

En el mundo, el café es uno de los comodities de mayor importancia, el cual tiene un aporte considerable para la economía de países ubicados en el este africano, el sur de Asia, y en la región central del continente americano. Adicionalmente, el café se ha vuelto parte de la identidad de muchas culturas. Pendergast (citado por (Kushnir, 2016)) argumenta que el

café ha tendido una fuerte influencia en la cultura laboral americana, donde se consideraba que el café proporcionaba energía a los trabajadores. En Colombia, el café es parte de su identidad, y el café colombiano es sinónimo de calidad en el mundo entero. En la siguiente imagen (figura 3) se muestra un resumen del proceso productivo del café verde en Colombia.

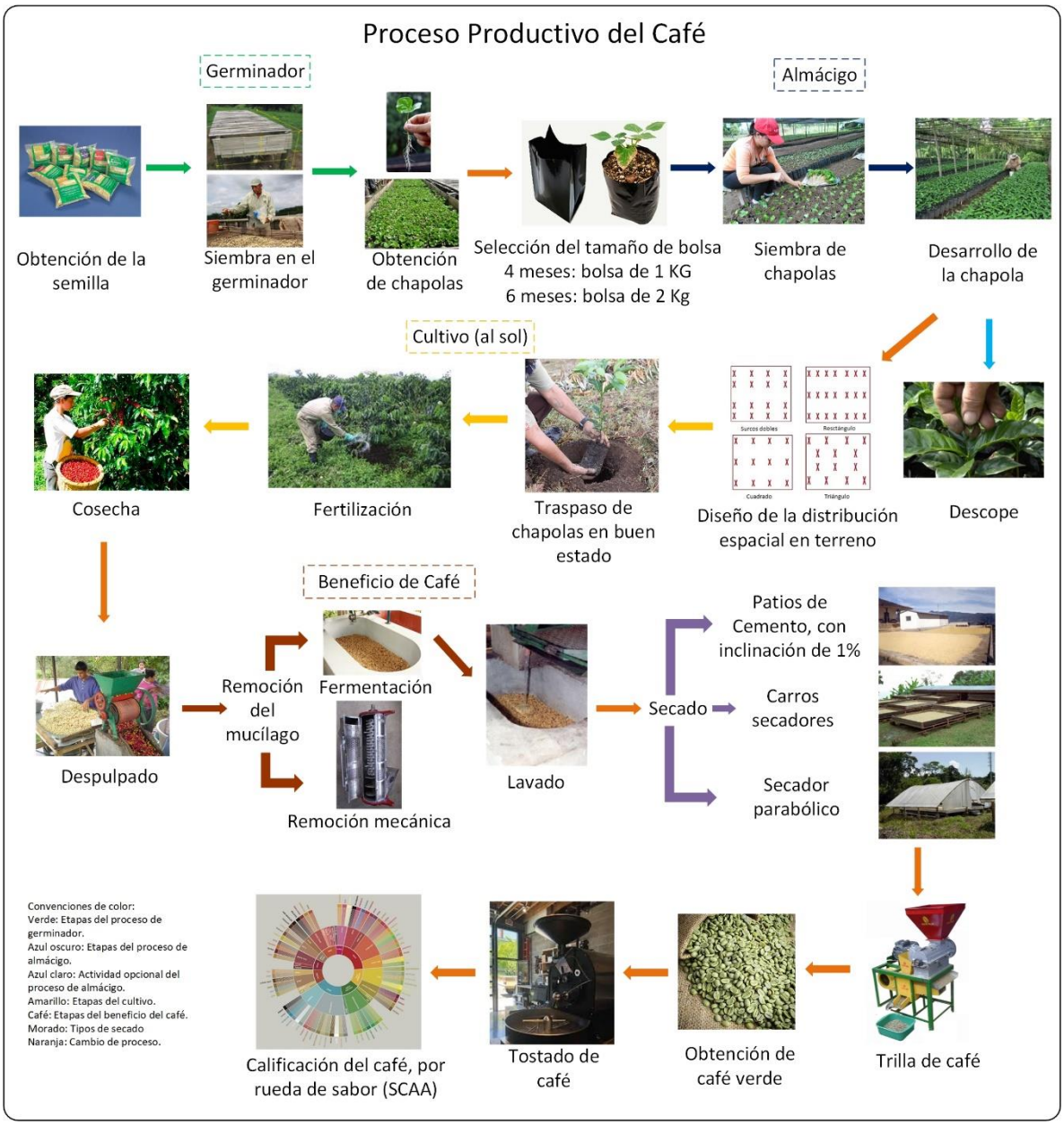


Figura 3. Proceso productivo del café verde en Colombia. Fuente: elaboración propia. Imágenes recolectadas de diversas fuentes ver anexo 2.

La productividad del café Colombiano, medida en cargas por hectárea es baja en comparación con países competidores (Espinal, Martínez, & Acevedo, 2005), esto se explica en parte, por el rigoroso proceso manual de selección del grano para cumplir con los estándares de calidad del mercado (Federación Nacional de Cafeteros, 2019), la falta de mano de obra capacitada y de recolectores (Banco Caja Social S.A., 2017) y el uso de árboles tradicionales que tienen poca resistencia al deterioro (USDA, United States Department of Agriculture Foreign Agricultural Service, 2016). Como parte de la solución a esta situación, desde el 2012 se implementó un programa de reemplazo de cultivos, por variedades con mayor resistencia, para finales del 2017 se habían completado 420.000 hectáreas, aproximadamente la mitad del área total (USDA 2016).

En cuanto al reemplazo de cultivos, surge otra necesidad y es la de cultivar variedades que por sus características tengan un mayor precio en el mercado, en los últimos años las fluctuaciones a la baja del precio del café internacional, con el cual se regula el precio de venta del café, ha afectado a muchos cultivadores en Colombia, quienes optan por reducir su inversión en tecnología, en fertilizantes que aumenten la productividad del cultivo y han incrementado su nivel de endeudamiento para pagar compromisos adquiridos. Para Botero (2019), el sector cafetero en Colombia debe implementar cambios profundos que le permitan hacer frente a las bajas en el precio del café, a los altos costos de fabricación de la variedad arábica y la creciente oferta del mercado impulsada por una mayor producción de países como Brasil, Vietnam e Indonesia.

A lo anterior, también se suma una tendencia del mercado convertidor de orientar sus compras hacia la variedad “robusta” que, aunque es un café más amargo que el arábico, sus costos de producción son menores, su productividad por hectárea es mayor y su precio es menor en comparación con el café arábico (Läderach, 2007). De hecho, algunos convertidores (tostadores) mezclan café colombiano con cafés de menor calidad para mejorar su rentabilidad, ofreciéndolos al mercado como café colombiano (Botero, 2019).

Las regiones tropicales montañosas poseen unas características específicas, las cuales, proporcionan el escenario ideal para la producción de ciertas variedades de café que son mejor pagas en el mercado. Generalmente en estas regiones se ubican en pequeñas y

medianas fincas productoras de café, las cuales pueden adaptarse con más facilidad a cambios en los cultivos, lo que les brinda una oportunidad para mejorar la rentabilidad de los cultivos de café (Läderach, 2007). No obstante, también presentan varios retos y es que su gran mayoría se encuentra ubicada en laderas que dificultan su recolección y la implementación de nuevas tecnologías.

2.1.2 Instituto de calidad del café (CQI)

El instituto de calidad del café (CQI por sus siglas en ingles), es uno de los principales referentes en el continente americano para la calificación de las distintas variedades de café comercializadas. De acuerdo con su página web, desde el 2002, a través del programa de ayuda estadounidense (USAID), trabajó con Colombia en un proyecto para el fomento de cultivos de café en reemplazo de cultivos ilegales. También ha logrado alianzas con la FNC, Catación Pública, Tecnicafé, Asoexport (ASOEXPORT, 2019) y el SENA, apoyando a Colombia con transferencia de buenas prácticas de procesamiento y entrenamiento de catadores para lograr la certificación Q. La base de datos referenciada para este trabajo fue recopilada de las páginas de revisión del CQI en enero de 2018 (Coffee Quality Institute, 2019), la cual se encuentra organizada y disponible en la página de GitHub (James L. , 2019). Teniendo en cuenta que en Colombia la variedad cultivada es la arábica, se realizó la revisión de la base de datos titulada “arabica_data_cleaned.cvs”, esta es una base de datos que contiene 1311 revisiones de café arábico, proveniente de distintos países, realizadas por catadores entrenados del CQI. A continuación, se relacionan las características presentes en la base de datos.

Mediciones de calidad: son los diferentes atributos medidos al café líquido. En una escala de 0 a 10 puntos, se califican los siguientes: Aroma, Sabor, Regusto, Acidez, Cuerpo, Balance, Uniformidad, Limpieza de la taza, Dulzura. Adicional se mide la Humedad y se realiza un conteo de Defectos.

Metadatos del grano: son atributos que dan información acerca de los granos de café, en la base de datos se tienen presentes el Método de procesamiento, el Color y la Especie.

Metadatos del lugar del cultivo: son atributos que dan información acerca del lugar de cultivo o la granja, se encuentra el Propietario, el País de origen, el Nombre de la granja, Número de lote, Molino, Compañía, Altitud, Región.

Si bien esta es una base de datos grande, las variables que están consignadas son informativas o de trazabilidad, lo que no permite llegar a predecir la calidad a partir de esta.

2.1.3 Aprendizaje automático aplicado al café

Alrededor de esta temática se han desarrollados varios trabajos, que emplean técnicas de ML para la clasificación de café en etapas del proceso de cultivo o postcosecha. A continuación, se relacionan algunos de los trabajos realizados.

En el trabajo titulado “Identificación y clasificación de frutos de café en tiempo real, a través de la medición de color” (Ramos, Sanz, & Oliveros, 2010), presentaron un método para clasificar el café partiendo de su color. En la figura 4 se muestra los estados del desarrollo del fruto del café. El sistema consistía en comparar el color de un fruto, como valor del espectro visible, versus datos guardados en el sistema ya etiquetados, logrando una precisión en la predicción entre el 87% y el 98%.

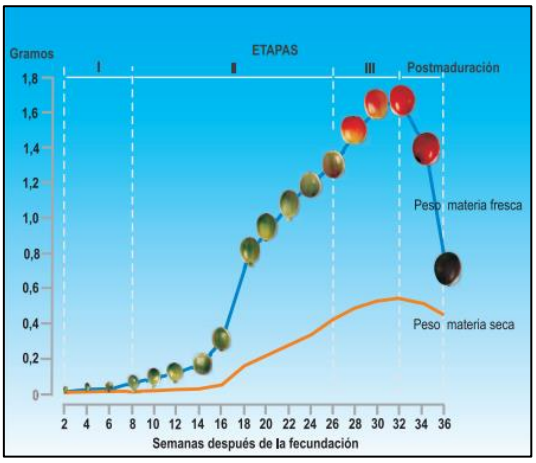


Figura 4. Fuente: (Ramos, Sanz, & Oliveros, 2010)

Un enfoque similar es abordado por (Carvajal, Aristizábal, Oliveros, & Mejía, 2006), quienes, usando un espectrofotómetro de esfera, determinaron cuantitativamente el color del fruto durante sus diferentes estados de desarrollo, que, aunque no emplean algoritmos de ML su trabajo relaciona varios aspectos y resultados que pueden ser empleados para la aplicación de clasificadores.

En el campo de la visión artificial que es una de las ramas del ML (Sandoval & Prieto, 2007), realizan una caracterización del café cereza (clasificación del grado de maduración),

empleando un clasificador Bayesiano y una red neuronal, concluyendo que, aunque el primer método tuvo menos error, su tiempo computacional fue mayor, respecto a la red neuronal. Un aspecto que destaca en este trabajo es la extracción de un conjunto de 208 características por muestra de textura, color y forma, obtenidas de las imágenes tomadas, de las cuales a través de un análisis univariado y multivariado reducen a 9.

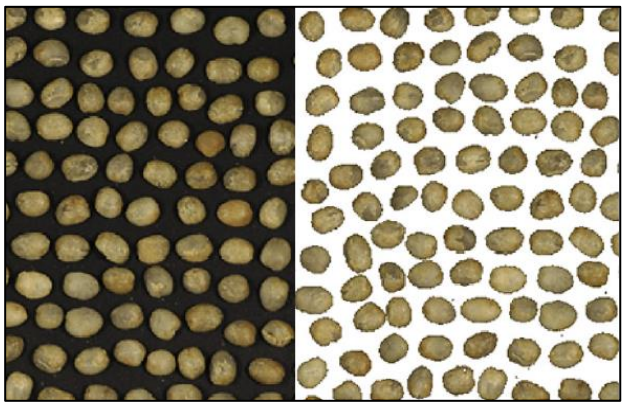


Figura 5. Imagen de café verde para identificar la calidad por su color. Fuente: (De Oliveira, Leme, Barbosa, & Rodarte, 2015)

También se encuentran los sistemas de reconocimiento visual, como el desarrollado por (De Oliveira, Leme, Barbosa, & Rodarte, 2015), en el cual clasificaron los granos de café verde, de acuerdo a su color (figura 5), siendo este, un determinante de la calidad del café, donde los autores destacan la mayor eficiencia computacional de una red neuronal versus un clasificador bayesiano. En el trabajo de (Ruge, Pinzon, & Moreno, 2012), se realizó un clasificador de café excelso mediante el procesamiento de imágenes de granos de café individuales, para lo cual usaron el dispositivo mostrado en la figura 6. La red se entrenó para reconocer un café con defectos (malo) de uno sin defectos, adicional se realizó una validación con profesionales expertos dedicados a realizar esta labor manualmente, para validar la precisión del modelo.

Otro ejemplo es el clasificador realizado por (Faridah, Parikesit, & Ferdiansjah, 2011), para muestras de café de varios granos, mediante un algoritmo de red neuronal, según los estándares de Indonesia. Para ello se entrenó la red con imágenes de varios grados como se muestra en

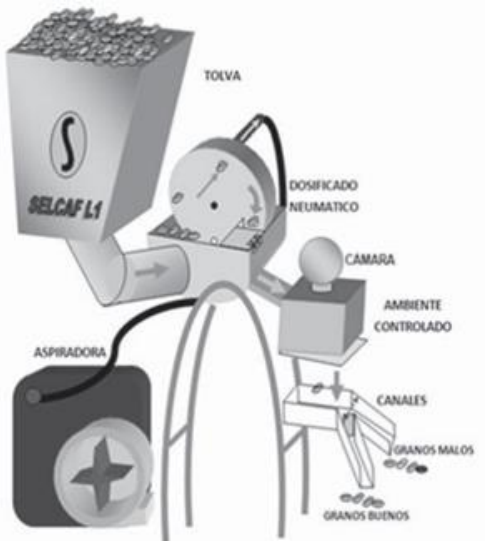


Figura 6. Sistema colector de café en la etapa excelso. Fuente: (Ruge, Pinzon, & Moreno, 2012)

la siguiente figura (7), capturadas mediante una cámara web y una fuente de iluminación dentro de una cámara.

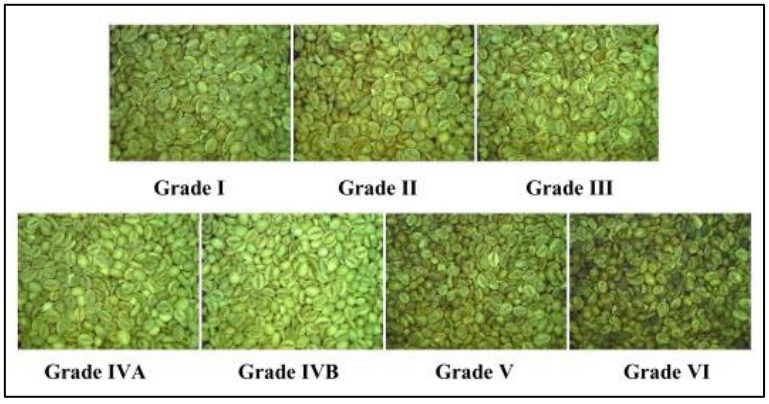


Figura 7. Grupos de clasificación definidos para la clasificación de granos de café verde Robusta. Fuente: (Faridah, Parikesit, & Ferdiansjah, 2011).

(Montes, 2003), realiza una segmentación de frutos de café para clasificarlos según su estado de maduración, para ello, emplea un algoritmo de conectividad, que emplea criterios de homogeneidad para agrupar pixeles, en donde inicialmente, se aplican herramientas como un filtro de mediana para reducir el tamaño de las imágenes y mejorar el uso de los recursos computacionales, un filtro de Sobel y Laplaciano Gausiano, para la detección de bordes, análisis de color mediante la transformada de coordenadas esféricas (SCT), y una técnica de crecimiento de regiones para segmentar de acuerdo al color y al contraste. En la siguiente imagen (figura 8) se aprecia, el proceso de separación de frutos empleado por (Montes, 2003).

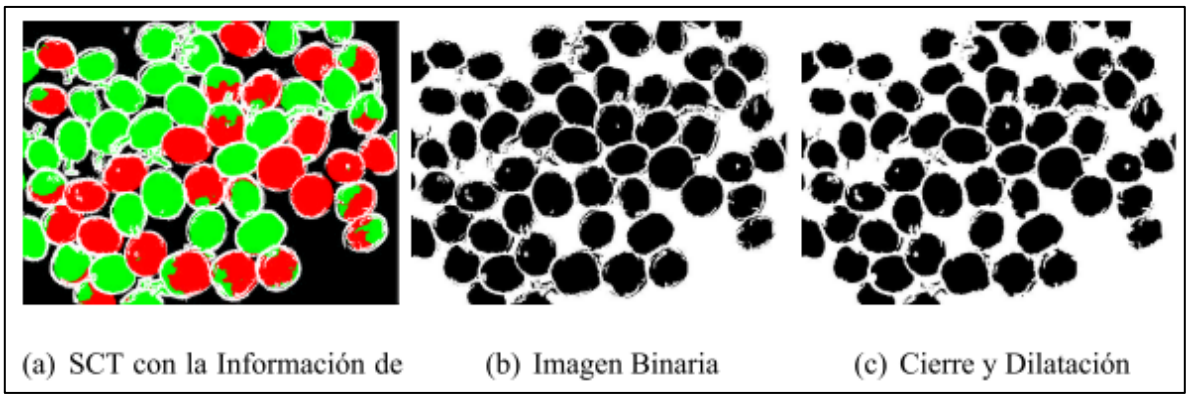


Figura 8. Proceso de separación de frutos en una imagen con granos de café en diferentes etapas de maduración, empleando la técnica SCT. Fuente: (Montes, 2003)

De los trabajos investigados en relación con la temática de este estudio, se aprecia que la mayoría de los autores han dirigido sus estudios hacia la clasificación de frutos o granos de café, por su importancia en la determinación de la calidad, construyendo sistemas, modelos o algoritmos capaces de identificar, segmentar o eliminar de una muestra los defectos o las categorías no deseadas, como lo pueden ser frutos inmaduros.

2.1.4 Revisión de trabajos de ML afines.

Ampliando el rango de aplicaciones a agrícolas diferentes al café, encontramos el trabajo de (Cubero, Aleixos, Moltó, Gómez-Sanchis, & Blasco, 2011) quienes realizan una recopilación de aplicaciones de ML para la clasificación orientada en la calidad de diferentes productos. A continuación, se muestra una imagen (tabla 1) tomada de Cubero, et al., (2011), con fragmentos de la tabla resumen que presentan los autores.

Tabla 1. Resumen de sistemas de inspección para frutas y vegetales. Fuente: (Cubero, et al., 2011)

Table 1 Summary of inspection systems for fruit and vegetables

Fruit	Imaging system	Focus on (colour coordinates and/or data processing methods in brackets)	References
Apple	CCD camera UVA fluorescent tubes, tungsten halogen lamp	Defects (ANN)	Ariana et al. (2006a)
Banana	Photographic camera, Diffuse fluorescent 45°	Texture (fractal Fourier descriptors)	Quevedo et al. (2008b)
Cucumber	Imaging spectrograph 950–1,350 nm, halogen lamps	Defects (PCA, band ratio, band difference)	Ariana et al. (2006b)
Grapefruit	Imaging spectrograph 450–930 nm, halogen lamps	Spectral information divergence	Qin et al. (2009)
Lemon	CCD camera UV lamps	Defects (thresholding)	Obenland and Neipp (2005)
Mandarin	Hyperspectral LCTF 460–1,020 nm, halogen lamps	Defects (SW, GALDA, CA, MI, CART, LDA)	Gómez-Sanchis et al. (2008a)
	Hyperspectral LCTF 460–1,020 nm, halogen lamps	Shape (digital elevation)	Gómez-Sanchis et al. (2008b)
Mango	Photographic camera, diffuse fluorescence	Colour (L*a*b*)	Kang et al. (2008)
Mushroom	Imaging spectrograph 450–950 nm	Colour (L), defects (PCA, LDA)	Gowen et al. (2009)
Olives	CCD camera	Colour (RGB, HSV), defects (ANOVA)	Riquelme et al. (2008)
Orange	Photographic camera, backlighting	Shape (Fourier descriptors, *k*-means)	Costa et al. (2009)
	Photographic camera, UV lamps 365 nm	Defects (thresholding)	Slaughter et al. (2008)
Peach	Multispectral camera, halogen lamps	Maturity, defects (clustering R/NIR)	Lleó et al. (2009)

El eje central de los trabajos de aplicación de ML o técnicas asociadas a esta rama del conocimiento era la de encontrar la forma de capturar características que permitieran, a partir de una imagen, entregar información a un algoritmo encargado de procesarla y dar un resultado. Dentro de los trabajos con más antigüedad, se aprecia, un mayor énfasis en el uso de técnicas para la extracción de datos, como color, textura, forma, separación de objeto del fondo y el uso de algoritmos condicionales o comparativos, para clasificar esas características, con ayuda de herramientas computacionales, para lograr realizar la clasificación. Dos ejemplos que contrastan los avances de este campo, se aprecian al revisar el trabajo de (Kaewapichai, Kaewtrakulpong, & Prateepasen, 2006) quienes propusieron un sistema de inspección en tiempo real para piñas, mediante un método denominado histograma de regiones segmentadas, que compara una imagen a inspeccionar, con una librería, construida, con información de la forma de la piña, separación del fondo e información del color y de la piel de la piña, versus el trabajo de (Dittakan, Theera-Ampornpunt, & Boodliam, 2018) quienes propusieron un sistema de inspección para clasificar piñas según su textura, extraída mediante un proceso denominado LBP (Local Binary Pattern), y lo evaluaron en 8 algoritmos diferentes como se muestra en la tabla 2, en donde el mejor resultado lo proporcionó una red neuronal.

Tabla 2. Resultados de 8 métodos de clasificación de piñas. Fuente: (Dittakan, Theera-Ampornpunt, & Boodliam, 2018)

Learning methods	AC	AUC	SN	SP	PR
Decision Tree (C4.5)	0.602	0.586	0.602	0.599	0.602
Binary Decision Tree	0.687	0.646	0.687	0.686	0.687
Random Forest	0.651	0.774	0.651	0.649	0.650
Nave Bayes	0.723	0.826	0.723	0.732	0.742
Bayesian Network	0.723	0.823	0.723	0.732	0.742
Logistic Regression	0.867	0.876	0.867	0.868	0.868
SMO	0.880	0.834	0.880	0.877	0.880
Neural Network	**0.940**	**0.979**	**0.940**	**0.940**	**0.940**

En el trabajo realizado por (Wang, y otros, 2015), titulado “Fruit Classification by Wavelet-Entropy and Feedforward Neural Network Trained by Fitness-Scaled Chaotic ABC and Biogeography-Based Optimization”, se proponen dos nuevos métodos de clasificación de frutas basados en aprendizaje automático. Dentro del desarrollo del trabajo se emplean herramientas de transformación y análisis de imagen tradicionales (ya que su estudio se

basa en las imágenes obtenidas de las frutas durante seis meses por medios fotográficos), frente a las herramientas nuevas de optimización y reducción de características, en ambos casos se usan la combinación de algoritmos de análisis de imagen, reducción de características y redes neuronales de aprendizaje avanzado; para validar estadísticamente los resultados, se usa la validación cruzada estratificada K-fold. A continuación, se muestra una imagen (figura 9) con parte de las frutas objeto de estudio del trabajo,

Figura 9: Frutas del estudio. Fuente: (Wang, y otros, 2015)

Con los métodos tradicionales de análisis de imagen, se necesita de muchas características para entrenar la red neuronal, lo que hace que se requiera una inversión computacional elevada, con los nuevos métodos de optimización es posible reducir el número de características para entrenar la red (mediante algoritmos de reducción), que combinados con algoritmos de clasificación permiten tener rendimientos competitivos contra los tradicionales.

Como resultados se obtiene que el uso de los nuevos optimizadores de procesamiento de imágenes para entrenamiento de redes neuronales tiene un rendimiento superior a los obtenidos con los métodos tradicionales. Como métodos tradicionales de análisis de imagen se usaron: combinación de histograma de color (CH), características basadas en la morfología (MP) y características de Unser (US); como redes neuronales de clasificación:

el genetic algorithm (GA), el particle swarm optimization (PSO) y el artificial bee colony (ABC). Por otro lado, como método nuevo de análisis de imagen se usa el wavelet entropy (WE); como redes neuronales de clasificación el fitness-scaled chaotic artificial bee colony (FSCABC) y el biogeography-based optimization (BBO); para ambos casos (método tradicional y nuevo) como reductor de características el principal component analysis (PCA). En la siguiente tabla (3) se muestran los resultados obtenidos para las diferentes combinaciones,

Tabla 3: Precisión de clasificación basada en diferentes algoritmos de entrenamiento. Fuente: (Wang, y otros, 2015), modificada por el autor

		# de características reducidas	**Exactitud (%)**
Algoritmos existentes	(CH+MP+US) +PCA+GA-FNN	14	84.8
	(CH+MP+US) +PCA+PSO-FNN	14	87.9
	(CH+MP+US) +PCA+ABC-FNN	14	85.4
	(CH+MP+US) +PCA+ kSVM	14	88.2
	(CH+MP+US) +PCA+FSCABC-FNN	14	89.1
Algoritmos propuestos	WE+PCA+FSCABC-FNN	12	89.5
	WE+PCA+BBO-FNN	12	89.5

Dada la complejidad que presenta la clasificación de frutas por medio de visión por computadora, por diversas propiedades características de cada tipo de fruta, se han realizado varios trabajos en torno a este tema, otro trabajo similar (al descrito anteriormente), es el desarrollado por (Zhang & Wu, 2012), el cual tiene por título “Classification of Fruits Using Computer Vision and a Multiclass Support Vector Machine” en donde también se buscó mejorar la clasificación de las frutas mediante aprendizaje automático, dentro de su experimento los autores propusieron un método de clasificación basado en el algoritmo de multiclase de kernel para support vector machine (kSVM), con el objetivo de lograr una clasificación precisa y rápida de las frutas. Los parámetros establecidos para entrenar la red neuronal fueron color, textura y forma, todos extraídos por

algoritmos que capturan esa información de las imágenes de entrada y luego las procesan para poder identificarlas.

Posterior al trabajo de extracción de la información, los autores usan el algoritmo de análisis de componentes principales (PCA) para hacer una reducción de las características obtenidas y que de esta manera la red pueda ser más eficiente. Para el entrenamiento de la red usaron tres tipos de SVM, los cuales fueron, Winner-Takes-All (WTA-SVM), Max-Wins-Voting (MWV-SVM), Directed Acyclic Graph (DAG-SVM) y tres clases de núcleos, el núcleo lineal (LIN), el núcleo dth Polynomial Homogeneous (HPOL) y el núcleo de Base Radial Gaussiana (GRB) , los cuales combinaron para determinar que combinación era más efectiva, los resultados obtenidos se muestran en la tabla 4, en donde se puede apreciar que la combinación más efectiva se encuentra entre el MWV-SVM con el núcleo GRB.

Tabla 4. Precisión de clasificación SVM. Resultados. Fuente: (Zhang & Wu, 2012)

	LIN	HPOL	GRB
WTA-SVM	48.1%	61.7%	55.4%
MWV-SVM	53.5%	75.6%	**88.2%**
DAG-SVM	53.5%	70.1%	84.0%

Siguiendo por el ámbito de las frutas, proyectado hacia otras ramas, se tiene como ejemplo el trabajo el realizado por (Cortez, Cerderia, Almeida, Matos, & Reis, 2009), denominado “Modeling wine preferences by data mining from physicochemical properties”, en el cual los investigadores propusieron un enfoque de minería de datos para predecir las preferencias de los consumidores del sabor del vino, tomando como conjunto de datos las pruebas realizadas al vino para certificación, específicamente a las del vino verde de Portugal, en sus variedades blanco y rojo. Para ello los autores aplicaron dos técnicas de regresión, bajo un proceso computacional eficiente que hace una selección simultanea de variables y modelo.

Las muestras para alimentar el algoritmo fueron obtenidas mediante la calificación de catadores profesionales, los cuales probaron a ciegas las muestras y las puntuaron en un rango de 0 a 10 en donde cero era muy malo y diez excelente, estas calificaciones fueron procesadas mediante una red neuronal perceptrón multicapa y una máquina de vectores de soporte (SVM), luego del entrenamiento de ambos sistemas los resultados (el

aprendizaje) fueron gestionados mediante el algoritmo análisis de sensibilidad, el cual permite descartar mediciones (resultados) irrelevantes dentro del proceso de aprendizaje de la red.

Los mejores resultados de aprendizaje para el estudio realizado se encontraron dentro del modelo SVM y para la clase blanco (la más común) del vino verde. De igual forma dentro del estudio los investigadores plantean que la superioridad de los resultados del método SVM sobre el perceptrón multicapa, gire en torno a las diferencias en las fases de entrenamiento, ya que el SVM garantiza un ajuste optimo, mientras que el perceptrón multicapa puede caer en un óptimo local.

Como se ha mencionado a lo largo de este trabajo cada vez son más campos los que emplean inteligencia artificial y su rama del aprendizaje automatizado; en la rama de la agricultura encontramos el trabajo titulado “Agricultural Crop Yield Prediction Using Artificial Neural Network Approach” desarrollado por (Dahikar & Rode, 2014), quienes intentaron desarrollar un modelo de predicción del rendimiento de varios cultivos utilizando un modelo de redes neuronales, con el fin de que dicha red aprendiera efectivamente la relación existente entre los factores climáticos y el rendimiento de los cultivos, datos que pudieran ser empleados posteriormente para estimar la producción a largo o corto plazo.

Por último fuera del ámbito agrícola tenemos el trabajo realizado por (Ahmed & Moustafa, 2016), titulado “House Price estimation from visual and textual features”, en el cual los investigadores propusieron una nueva metodología para estimar el precio de las viviendas, debido a que en los métodos empleados anteriormente solo se usaba información textual para estimar dicho precio (por medio de redes neuronales), ellos propusieron mezclar la información textual (datos de la vivienda) junto con imágenes de los predios (las imágenes fueron usadas para extraer sus características visuales), la cual iba a ser proporcionada a dos métodos de clasificación los SVM (Support Vector Machine) y a una red neuronal que empleaba el algoritmo Levenberg-Marquardt (LMA) como entrenamiento.

Las características extraídas de las imágenes de las viviendas alimentaron a una red neuronal multicapa, la cual estaba totalmente conectada, que estimaba el precio de las viviendas como única salida, (las viviendas empleadas también debian tener la información textual). Los datos visuales de las casas estuvieron compuestos por una serie de 4

imágenes, la parte frontal de la vivienda, un dormitorio, un baño y la cocina, como se muestra a continuación en la figura 10, mientras que los datos textuales fue información de la vivienda, tal como el número de habitaciones, de baños, área total del predio y el código postal para tener referencia de la ubicación de este.

Figura 10: Ejemplo de las imágenes tomadas de una de las viviendas para alimentar la red neuronal. Fuente: (Ahmed & Moustafa, 2016)

Para la extracción de características de la imagen los autores emplearon el algoritmo SURF (Speeded Up Robust Features), el cual extrae las partes cercanas alrededor de un punto de interés, una de las cosas que detectaron los investigadores fue que el algoritmo extrae "las partes más importantes" (Ahmed & Moustafa, 2016) de la imagen, en la figura 11 se puede apreciar un ejemplo de lo descrito.

Figura 11: Ejemplo de extracción de características visuales con SURF. Fuente: (Ahmed & Moustafa, 2016)

Al culminar su experimento (Ahmed & Moustafa, 2016), lograron concluir que la combinación de información textual y visual producía una mejor precisión de la estimación del precio de las viviendas en comparación con las estimaciones realizadas solo con los

datos textuales. Adicionalmente encontraron que los resultados entregados por la red neuronal fueron mejores que los entregados por la máquina de soporte de vectores (SVM), dado el mismo conjunto de datos.

2.2 MARCO CONCEPTUAL

2.2.1 Herramientas computacionales.

En esta sección se describe el lenguaje de programación, las aplicaciones de interfaz de programación, las librerías y el software en general utilizados en este trabajo.

Python: Python es un lenguaje de programación interpretado, de alto nivel, orientado a objetos, multiplataforma, de código abierto, caracterizado por ser de fácil aprendizaje (Hinojosa Gutiérrez, 2016). Fue desarrollado por el holandés Guido Van Rossum a finales de 1980, con el fin de intentar solucionar los problemas de conexión de la interfaz de usuario del bourn shell en el sistema operativo Amoeba, en el cual trabajaba (Chazallet, 2016).

Jupyter notebook: "Es una aplicación web de código abierto", nacida del proyecto IPython en el 2014, que permite utilizar diferentes lenguajes de programación, para elaborar documentos que albergan ejecución de código, texto narrativo, vídeo y ecuaciones. Es utilizado en análisis numérico, estadística y aprendizaje automático" (Jupyter, 2019).

Google Colab: "Colaboratory es un entorno gratuito de Jupyter Notebook que no requiere configuración y que se ejecuta completamente en la nube. Colaboratory permite escribir y ejecutar código, guardar y compartir análisis y tener acceso a recursos informáticos muy potentes, todo de forma gratuita desde el navegador" (Google Colab, 2019).

Keras: "Keras es una API (interfaz de programación de aplicaciones) de redes neuronales de alto nivel, escrita en Python y capaz de ejecutarse sobre TensorFlow, CNTK o Theano" (Keras, 2015). Permite el desarrollo de modelos de aprendizaje profundo, de una manera sencilla y legible (Torres, 2018).

TensorFlow: Es una plataforma de código abierto que permite expresar algoritmos de machine learning y una implementación para ejecutar los mismos. Tensorflow puede ejecutarse en diferentes sistemas operativos, con una variabilidad casi nula entre ellos. Posee un ambiente integral y flexible de herramientas, bibliotecas y recursos comunitarios, con los cuales se pueden efectuar una amplia variedad de algoritmos, incluidos los algoritmos para modelos de redes neuronales profundas, lo que ha permitido que se emplee

en diferentes campos del conocimiento, como la robótica, el reconocimiento de voz, entre otras (Abadi, y otros, 2015).

Numpy: “Numpy es una biblioteca de Python de código abierto, que permite trabajar matrices, debido a que contiene una larga lista de funciones matemáticas, dentro de las que se incluyen algunas de algebra lineal, transformación de Fourier y generación aleatoria de números” (Idris, 2011). “Fue escrito originalmente por Travis Oliphant como base de un entorno informático en Python (SciPy), posteriormente se disocio del módulo SciPy y fue lanzado de manera independiente en el 2006” (Chin & Dutta, 2016).

Seaborn: “Es una biblioteca de visualización de datos en Python basada en matplotlib, que proporciona una interfaz de alto nivel para graficar datos estadísticos” (Seaborn, 2012).

Matplotlib: Es una biblioteca de Python para e trazado 2D que genera gráficos, en entornos interactivos y que pueden ser exportados en diferentes formatos de impresión (Hunter, 2007).

Pandas: “Es una biblioteca de código abierto que proporciona estructuras de datos y herramientas de análisis de alto rendimiento y fáciles de usar para el lenguaje de programación Python” (Pandas, 2019). “El nombre Pandas se deriva de la palabra Panel Data, una ecometría de datos multidimensionales” (Tran, 2019).

Scikit-Learn: “Es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de mediana escala” (Pedregosa, y otros, 2011), además “proporciona algoritmos para tareas de aprendizaje automático que incluyen clasificación, regresión, reducción de dimensionalidad y agrupamiento” (Hackeling, 2014). En el 2007 se inició como un proyecto de Google Summer of Code de David Cournapeau; pero fue hasta el 2010 que Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort y Vincent Michel de INRIA, retomaron el proyecto y lo difundieron al público (Pedregosa, y otros, 2011).

2.2.2 Machine Learning (ML).

Para (Hackeling, 2014), el ML es el diseño y estudio de artefactos de programas computacionales que usan experiencia pasada para realizar decisiones futuras. Samuel (citado por (Hackeling, 2014)), dijo que el ML es el estudio que le da a los computadores la habilidad de aprender sin haber sido explícitamente programados.

Mitchell (citado por (Hackeling, 2014)), define el ML de la siguiente forma; un programa aprende de una experiencia E con respecto a una clase de tareas T y una medida de desempeño P, si su desempeño de la tarea en T, medida por P, mejora la experiencia E.

Dentro del ML encontramos 2 enfoques, el aprendizaje supervisado y el aprendizaje no supervisado. De acuerdo con (Hastie, Tibshirani, & Friedman, 2009), el aprendizaje supervisado consiste en predecir los valores de un conjunto de datos de salida, a partir de un conjunto de datos de entrada. Se le llama supervisado porque conforme el modelo predice las salidas para datos de prueba, se calcula el error entre lo que predijo el algoritmo y el valor real. El objetivo es minimizar el error, ajustando la función de densidad de probabilidad que relaciona las entradas con las salidas.

Por otra parte, en el aprendizaje no supervisado, solo se tienen conjuntos de datos de entrada, sin conocer su relación con variables de salida, por lo que no tiene como verificar fácilmente si el desempeño del modelo es el adecuado. En esta categoría, los modelos de ML realizan principalmente agrupaciones o clasificaciones de los datos, según las variables de entrada y su diferenciación. Son modelos con métricas no muy precisas y que en ocasiones dependen de ciertas heurísticas para validar sus resultados Hastie et al., (2009). Esta complejidad ha generado diversos métodos para trabajar estos modelos, no obstante, son de mucha utilidad para analizar relaciones o dependencias que muchas veces no son tan evidentes, como la detección de anomalías (González, 2018).

2.2.3 Modelos lineales.

Dentro de los modelos lineales, se encuentran los algoritmos de espacios medios (halfspaces), regresión lineal, y regresión logística. Muchos algoritmos de aprendizaje se

basan en predictores lineales, dado su eficiente habilidad de aprendizaje, su facilidad de interpretación y su capacidad para representar problemas de aprendizaje natural (Shalev-Shwartz & Ben-David, 2014). La regresión lineal permite modelar la relación entre unas variables de entrada (explicativas) con valores de salida reales. Dado un conjunto de datos *X* (ϵ R^d) y un conjunto de salidas reales dependientes *Y*, se busca establecer una función lineal h: $R^d \rightarrow R$ que aproxime la relación entre las variables de entrada y salida. La forma más básica es la regresión lineal simple, en donde existe una sola variable predictora x1 de Y, cuya relación lineal se escribe mediante la siguiente ecuación:

$$Y = \omega_0 + \omega_1 X$$

Los términos ω_0 y ω_1, son los coeficientes de regresión lineal, ω_0 y ω_1 son el punto de intercepto y la pendiente respectivamente. Para hallar estos coeficientes se requiere de una función de pérdida, que permita penalizar las diferencias entre las predicciones *h(x)* y los valores reales *Y*. Aunque existen varios métodos, el más común es usando el descenso de gradiente aplicado a la ecuación del error medio, descrito por la siguiente ecuación:

$$L_s(h) = \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$$

La función de pérdida también es conocida como función de costo. La forma de minimizar esta función es mediante la aplicación del descenso de gradiente. Esta técnica consiste primero en calcular la función de costo a partir de unos valores de ω_0 y ω_1 iniciales (aleatorios o iguales a cero), y aplicar la derivada parcial a nuestra función de costo para determinar ω_0 y ω_1, una vez se obtienen estos parámetros se calcula la función de costo nuevamente y así sucesivamente hasta que la función de costo converja, es decir alcance un óptimo local (Ng, 2018) . A continuación, se muestran las ecuaciones derivadas de la función de costo para hallar los parámetros del modelo o los coeficientes de la regresión lineal.

$$\omega_0 = \omega_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\omega(x_i) - y_i)$$

$$\omega_1 = \omega_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left((h_\omega(x_i) - y_i) x_i \right)$$

Donde m es el tamaño del conjunto de datos de entrenamiento, ω_0 y ω_1 son constantes que se actualizan con cada iteración y x_i, y_i son los valores del conjunto de datos de entrenamiento.

2.2.4 Algoritmo árbol de decisión.

Según (Palma Méndez & Marín Morales, 2008) “los árboles de decisión son una representación gráfica de un procedimiento para clasificar o evaluar un concepto”. Dichos arboles están constituidos por nodos de decisión, los cuales se despliegan en ramas para cada una de las alternativas (Barber, 2012).

Dentro de los datos utilizables en los árboles de decisión es posible tener datos binarios (como por ejemplo atributos donde las respuestas a las preguntas sean verdadero / falso o si / no) o datos que sean propensos a categorización pero no sean de naturaleza binaria. Para los datos binarios los ensayos son sencillos si se le asignan valores de 0 y 1 a las características, por el contrario, cuando los datos no son binarios, las pruebas pueden establecerse formando subconjuntos excluyentes y exhaustivos entre sí, “si los atributos son numéricos las evaluaciones pueden contener, pruebas de intervalos” (Nilsson, 1998).

Una de las dificultades en este método es el establecimiento de las pruebas y del orden de las mimas (Nilsson, 1998), que conlleva a un alto índice de incertidumbre, debido a que las pruebas pudiesen clasificar los atributos con poca frecuencia, para disminuir dicha incertidumbre se emplea una técnica denominada entropía (Hackeling, 2014).

La entropía según (Palma Méndez & Marín Morales, 2008) es una medida generalmente usada para determinar el grado de incertidumbre de una variable y está dada por la siguiente ecuación:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i)$$

"En la cual n es el número de resultados y $P(x_i)$ es la probabilidad del resultado i. Los valores comunes para b son 2, *e* y 10, debido a que el logaritmo de un número menor a 0 sería negativo, toda la suma se niega para devolver el valor positivo" (Hackeling, 2014).

2.2.5 Algoritmo vecinos cercanos (*K-Neighbours*).

Los algoritmos de vecinos cercanos memorizan el conjunto de entrenamiento (características), para luego predecir la etiqueta (resultado) de un nuevo ítem, basado en los atributos de sus vecinos cercanos en el conjunto de entrenamiento. Esto es posible debido a que el método se fundamenta "en el supuesto de que las características empleadas para describir los puntos de dominio son relevantes para sus etiquetas" (Shalev-Shwartz & Ben-David, 2014) haciendo que los puntos cercanos a estos posean las mismas etiquetas (Shalev-Shwartz & Ben-David, 2014).

Para hallar una etiqueta el método usa métricas de distancia al punto dominio, las más usadas para ello son la euclidiana y la mahalanobis al cuadrado. La distancia Euclidiana es la distancia entre dos variables (x_{11}, x_{12}, …, x_{1n}) y (x_{21}, x_{22}, …, x_{2n}), teniendo como fórmula (Nilsson, 1998):

$$D = \sqrt{\sum_{j=1}^{n} \left(x_{1j} - x_{2j}\right)^2}$$

Esta distancia es apropiada cuando los elementos del conjunto se miden de la misma forma, es decir en la misma característica y dimensión, debido a que modificar la escala de una dimensión podría cambiar el conjunto de elementos cercanos (Russell & Norvig, 2004). Una solución a este inconveniente es homogenizar la escala para cada dimensión, midiendo "la desviación estándar de cada característica sobre el conjunto de datos y expresar los valores de la característica como múltiplo de la desviación de esa característica" (Russell & Norvig, 2004), para ello sirve la distancia mahalanobis, cuya ecuación se muestra a continuación (Cuadras, 1989),

$$M^2(i,j) = (x_i - x_j)' \sum^{-1} \left(x_i - x_j\right)$$

Donde x_i e x_j representan los vectores de distancia entre dos variables *i, j* pertenecientes a una población Ω "caracterizada por *p* variables aleatorias, siendo $\mu = (\mu_1, \ldots, \mu_p)'$ el vector de medias y ∑ la matriz de covarianzas no singular" (Cuadras, 1989).

Igualmente, "la distancia entre un individuo *i* y la población es," (Cuadras, 1989)

$$M^2(i, \Omega) = (x_i - \mu)' \sum^{-1} (x_i - \mu)$$

"Y la distancia entre dos poblaciones es," (Cuadras, 1989)

$$M^2(\Omega_1, \Omega_2) = (\mu_1 - \mu_2)' \sum^{-1} (\mu_1 - \mu_2)$$

2.2.6 Algoritmo Support Vector Machine (SVM)

Las máquinas de vectores de soporte (SMV) son algoritmos de aprendizaje supervisado (MathWorks, s.f.), "desarrollados por Vapnik y Cortés (1995) y su equipo AT&T" (Sánchez, 2015) usados para la clasificación binaria o regresión, en problemas de reconocimiento de patrones. Dicha clasificación se cimienta en el fundamento de separación optima entre clases, de tal manera que, si las clases son susceptibles de separación, el resultado se elige para segregar las clases tanto como sea posible (Figueira, y otros, 2009).

Las SVM construyen un hiperplano en un espacio de dimensionalidad muy alta (que puede llegar a ser infinita) que separe las clases que se tienen en un conjunto de datos. Una buena disociación entre las clases generará una categorización correcta de la nueva muestra, en otros términos, se requiere hallar la máxima separación a los puntos más cercanos al hiperplano generado (Gala, 2013).

Debido a que los SVM son en un método conocido de ML e implementado en varios campos del conocimiento, han surgido bibliotecas y herramientas para facilitar su uso, una de ellas es LIBSVM (biblioteca de SVM), usada dentro de Scikit-Learn.

La LIBSVM permite varias formulaciones SVM para clasificación, regresión y estimación de distribución, algunas de las formulaciones empleadas se nombran a continuación (Chang & Lin, 2013):

Clasificación de vectores de soporte C: para esta clasificación se tiene dos vectores de entrenamiento X_i que pertenece al conjunto R^n, donde $i = 1, \ldots, l$, en dos clases, y un vector indicador y que pertenece al conjunto R^l tal que y_i pertenece a {1,-1}, C-SVM resuelve el problema de optimización primaria, según Boser et al.,1992; Cortes & Vapnik, 1995; citados por (Chang & Lin, 2013), de la siguiente manera,

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

Sujeto a,

$$y_i(w^T \emptyset(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, l$$

Donde $\emptyset(x_i)$ asigna x_i a un espacio de dimensiones superiores y C > 0 es el parámetro de regularización. En vista de la posible alta dimensionalidad de la variable W, se resuelve el problema dual con la siguiente ecuación (Chang & Lin, 2013),

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

Sujeto a,

$$y^T \alpha = 0,$$

$$0 \leq \alpha_i \leq C \qquad i = 1, \ldots, l$$

Donde $e = [1, \ldots, 1]^T$ es el vector de cada uno de estos, Q es una matriz semidefinida positiva l por l.

Luego de la aplicación de ecuación anterior, que soluciona el problema de dual se aplica la ecuación de satisfacción optima (w) (Chang & Lin, 2013),

$$w = \sum_{i=1}^{l} y_i \, \alpha_i \emptyset(x_i)$$

Y finalmente la función de decisión (Chang & Lin, 2013),

$$sgn(w^T \emptyset(x) + b) = sgn\left(\sum_{i=1}^{l} y_i \, \alpha_i K(x_i, x) + b\right)$$

Soporte de Regresión de Vectores (ϵ- SVR):

Según Vapnik (1998) citado por (Chang & Lin, 2013), la forma estándar de la regresión de vectores de soporte es la siguiente formula,

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^*$$

Sujeto a,

$$w^T \Phi(x_i) + b - z_i \leq \epsilon + \xi_i,$$
$$z_i - w^T \Phi(x_i) - b \leq \epsilon + \xi_i^*,$$
$$\xi_i \xi_i^* \geq 0, i = 1, \ldots, l$$

Esta formulación se ha realizado teniendo en cuenta que se considera un conjunto de entrenamiento $\{(x_1, z_1), \ldots, (x_l, z_l)\}$, donde x_i pertenece al conjunto R^n y es un vector de características y z_i que pertenece al conjunto R^1 es la salida del modelo. Esto pude generar un problema dual, el cual se resuelve así (Chang & Lin, 2013),

$$\min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \epsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i \, (\alpha_i - \alpha_i^*)$$

Sujeto a,

$$e^T (\alpha - \alpha^*) = 0,$$
$$0 \leq \alpha_i \alpha_i^* \leq C, i = 1, \ldots, l,$$

Donde $Q_{ij} = K(x_i, x_j) \equiv \Phi(x_i)^T \Phi(x_j)$. Posterior a la solución del problema dual, la función aproximada es (Chang & Lin, 2013),

$$\sum_{i=1}^{l} (-\alpha_i + \alpha_i^*) \, K(x_i, x) + b$$

2.2.7 Algoritmo regresión logística.

La regresión logística es un algoritmo que sirve para la clasificación de datas de información, basado en la probabilidad de que una variable corresponda a una clase; “las probabilidades deben ser continuas en el conjunto y delimitadas entre (0,1)” (Bonaccorso, 2017), de esta manera se establece un umbral de discriminación, en donde si la predicción (respuesta) de una muestra es superior a dicho umbral, la clase será positiva, de lo contrario será negativa (Hackeling, 2014). El nombre logístico proviene de usar la función sigmoide (o logística), la cual se expone a continuación (Bonaccorso, 2017):

$$F(t) = \frac{1}{1 + e^{-t}}$$

Donde *t*, es equivalente a la conjunción lineal de variables explicativas, por lo cual la anterior ecuación queda (Hackeling, 2014):

$$F(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_x)}}$$

En este modelo también es utilizada la función inversa a la función logística la cual es denominada función logic, en donde se vincula la función logística con una combinación lineal de las variables explicativas, quedando finalmente, la siguiente ecuación (Hackeling, 2014):

$$g(x) = \frac{F_{(x)}}{1 - F_{(x)}} = \beta_0 + \beta_x$$

2.2.8 Red Neuronal artificial.

En ML uno de los modelos más usados es el de redes neuronales (NN por Neural Network en inglés), una red neuronal artificial, es un modelo de computación conformado por un número de unidades computacionales básicas, llamados neuronas que están conectados entre sí, formando una red de comunicación la cual permite realizar computaciones complejas (Shalev-Shwartz & Ben-David, 2014). La forma básica de una red neuronal se muestra en la siguiente imagen (figura 12):

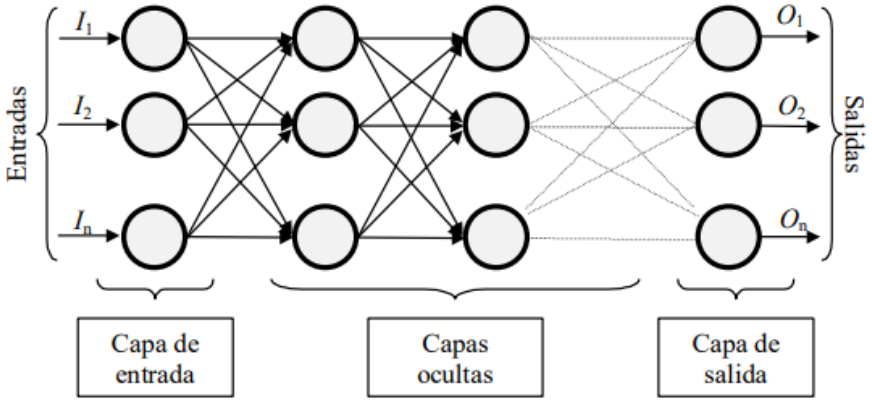


Figura 12. Esquema de una red neuronal. Fuente: (Matich, 2001)

En donde cada nodo es una neurona y los arcos representan las conexiones entre ellas. Las neuronas están agrupadas en capas. Una capa de entrada que recibe los datos de entrada del modelo, los cuales pasan a través del siguiente grupo, denominado capas ocultas. En este grupo, se pueden encontrar tantas capas como la complejidad del problema o de los datos a entrenar lo requiera. Finalmente, una capa de salida, donde se obtendrán los valores a las funciones resultantes del procesamiento de la red.

La manera en que las redes adquieren su conocimiento está contenida en los pesos de conexión que posee cada neurona (nodo) dentro de la red, asumiendo sus valores (pesos de conexión) gracias a la fase de entrenamiento (Gallo, 2015).

La manera en cómo se interconectan las neuronas permite distinguir dos tipos de arquitectura (Gallo, 2015):

- La arquitectura de retroalimentación, en la cual existen conexiones entre neuronas de la misma capa o de la anterior;
- La arquitectura de avance (según Hornik, Stinchcombe y White, 1989 citados por (Gallo, 2015), sin conexiones de retroalimentación, es decir, las señales van solo a las neuronas de la capa siguiente.

Las redes neuronales pueden ser configuradas de múltiples maneras (las posibilidades llegan a ser infinitas), por ello la elección de la configuración óptima para los datos o

problema manejados, debe ser principalmente una función objetivo de la aplicación. Existen dos arquitecturas principales, las cuales se describen a continuación:

Perceptrón: Es según Rosenblatt (1958) Minsky y Papert (1969) citados por (Gallo, 2015), la red más simple, la cual está conformada por una única neurona, con *n* entradas y una sola salida. El algoritmo empleado por esta red analiza los patrones de entrada y, ponderando las variables a través de las sinapsis (conexiones) decide la salida que está asociada con la configuración. Este tipo de arquitectura se ve limitado a resolver problemas de soluciones linealmente separables.

Perceptrón multicapa (MLP): Las redes con una capa de entrada, capas intermedias (que pueden ir de una hasta las que se requieran) y una capa de salida son denominadas perceptrón multicapa (MLP - por su nombre en inglés multi layer perceptrón) según Hornik, Stinchcombe y White (1998) citados por (Gallo, 2015). Son de tipo de alimentación avanzada (feed forward) que emplean el algoritmo de aprendizaje de retro propagación (en su gran mayoría), este aprendizaje consiste en que la red calcula los pesos entre capas a partir de valores aleatorios, hace cambios graduales y progresivos, posterior a una evaluación de errores de las salidas, hasta que los resultados converjan a una aproximación de error aceptable (Gallo, 2015).

Las redes neuronales tienen ciertas ventajas como lo indica (Matich, 2001),

- Aprendizaje Adaptativo. Capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.
- Autoorganización. Una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.
- Tolerancia a fallos. La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.
- Operación en tiempo real. Los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.

- Fácil inserción dentro de la tecnología existente. Se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilitará la integración modular en los sistemas existentes
- Estas ventajas, hacen que las NN se apliquen en diversos enfoques y sean pilar del reconocimiento visual.

2.2.9 Red neuronal convolucional

El concepto de redes neuronales convolucionales (CNN) fue acuñado por (LeCun, Bottou, Bengio, & Haffner, 1998), igualmente son conocidas también como LeNet, debido a su inventor. Sirven para hacer la clasificación de imágenes y consisten en un conjunto de neuronas interconectadas que atribuyen las características de entrada a las salidas, por medio de aplicar filtros, los cuales se encargan de extraer partes de las imágenes de entrada (LeCun, Bottou, Bengio, & Haffner, 1998).

Las CNN además de poseer las características de las redes neuronales tradicionales, también disponen de propiedades específicas, algunas de ellas son, en primer lugar que son profundas, tienen un numero típico de capas que varía entre 10 a 30, pero este puede extenderse hasta 1000 o más en algunos casos, en segundo lugar, las neuronas están conectadas de manera que múltiples neuronas comparten pesos, lo cual admite que la red realice convoluciones (o coincidencia de plantilla) de la imagen de entrada con los filtros (definidos por los pesos) dentro de la CNN (Mazurowski, Buda, Saha, & Bashir, 2018) por último, puede hacer que las neuronas no están conectadas todas entre sí, sino con un grupo pequeño local, que además comparte los mismos pesos que la capa anterior (esto se denomina perceptrón multicapa). Igualmente son más robustas con relación a las variaciones de los datos de entrada, como la traslación o rotación de la imagen (Simonyan & Zisserman, 2015).

En las primeras versiones de las CNN, estas tomaban la imagen de entrada, aplicaban una convolución y luego un submuestreo o reducción dimensional, para finalmente nutrir a una red conectada totalmente, la cual se encargaba de hacer la clasificación (figura 13) (Singla, Yuan, & Ebrahimi, 2016).

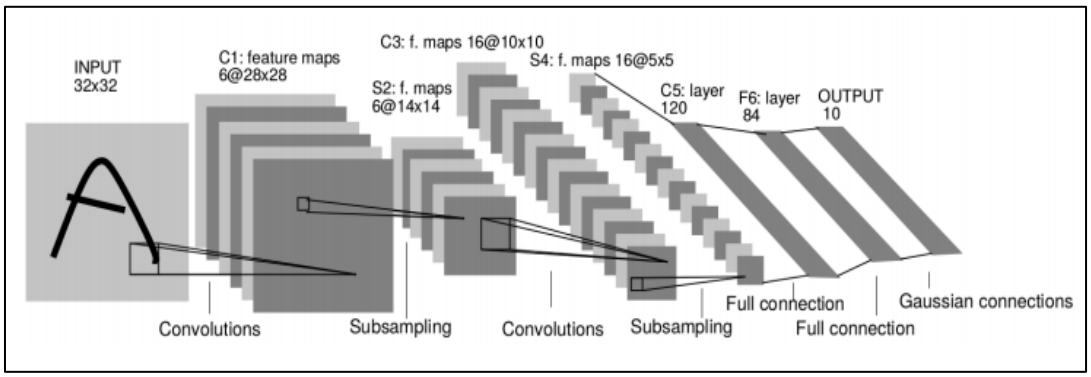


Figura 13: Arquitectura de la red LeNet-5. Fuente: (LeCun, Bottou, Bengio, & Haffner, 1998)

Posteriormente se propusieron mejoras y variantes con mayor profundidad de capas , las cuales se usan actualmente como la red VGG16 (figura 14), propuesta por (Simonyan & Zisserman, 2015), la cual se sirve de una gran capacidad computacional para adentrarse en las capas y lograr una mayor sustracción de características, reconocida por ser la ganadora del desafío ImageNet (ILSVRC) (Simonyan & Zisserman, 2015).

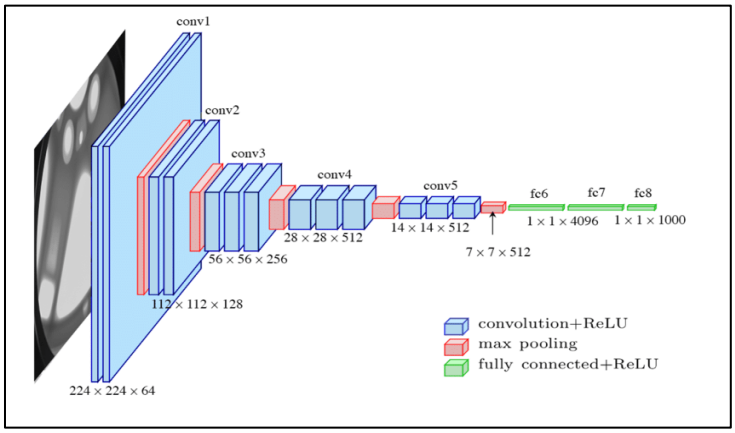


Figura 14: Red convolucional VGG16. Fuente: (Ferguson, Ak, Lee, & Law, 2017)

Otra de las redes reconocidas dentro del concurso nombrado anteriormente es la red Inception, la cual aplica varios tipos de filtros a la vez y los enlaza en uno solo (ver figura 15, para visualizar el funcionamiento), esta manera de trabajo posibilita la reducción de la carga computacional y al mismo tiempo manejar un alto volumen de información (Szegedy, Ioffe, Vanhoucke, & Alemi, 2016).

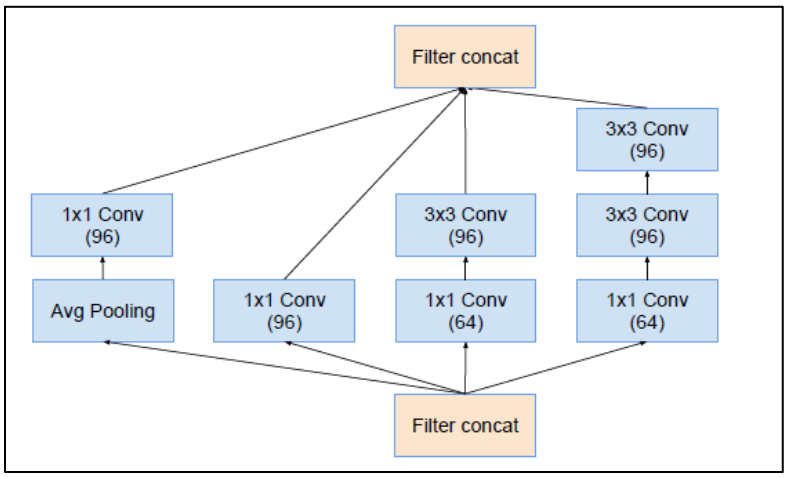


Figura 15: Esquema del bloque A de la red InceptionV4. Fuente: (Szegedy, Ioffe, Vanhoucke, & Alemi, 2016)

2.2.10 Validación cruzada

Según (Aguilar, Torres, & Martín, 2019) "la validación cruzada es un método para estimar el rendimiento de algoritmos ML con la menor varianza posible cuando se utiliza un único conjunto de entrenamiento". Este algoritmo divide al conjunto de datos en k grupos, a los cuales se les denomina *fold*, y son aproximadamente del mismo tamaño. El primer fold se trabaja como un conjunto de comprobación, y el modelo se entrena con los *k-1* folds restantes; dicho entrenamiento se efectúa haciendo que cada uno de los folds pase a ser un grupo de testeo cada vez (ver figura 16 para esquema del proceso). De modo que al final se tienen k diferentes medidas del error de prueba (MSE_k) del algoritmo; que se puede concluir a partir de la media de las k medidas del rendimiento (James, Witten, Hastie, & Tibshirani, 2013) (Aguilar, Torres, & Martín, 2019).

La media está dada por la siguiente ecuación (James, Witten, Hastie, & Tibshirani, 2013):

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

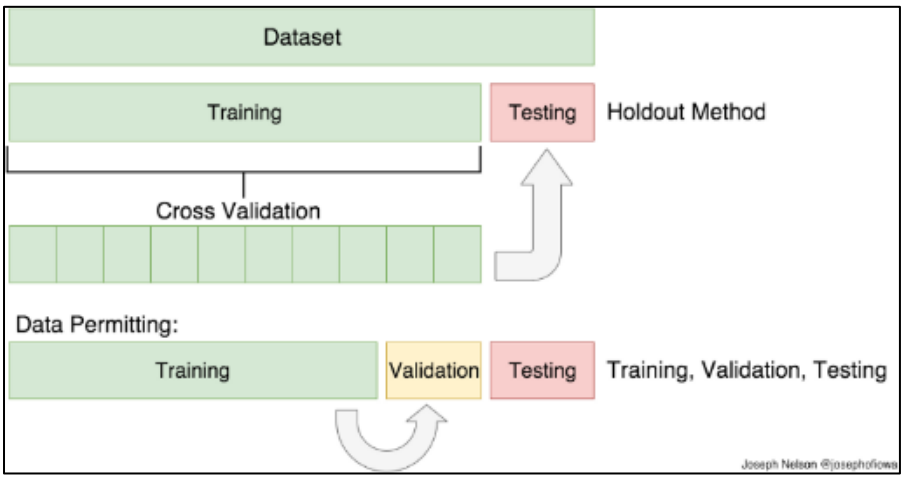


Figura 16: Proceso del método de validación cruzada. Fuente: (Shah, 2019)

2.2.11 Medidas de desempeño.

Para realizar la validación del desempeño de los modelos de Machine Learning, se emplean diferentes medidas según el tipo de problema que se esté analizando. En la tabla 5, se muestra la tabla de confusión, la cual permite entender las diferentes medidas partiendo del cruce entre la condición real y la condición predicha. Dentro de las más empleadas se encuentran Accuracy, Precision, Recall, F1-score, para los modelos de clasificación y R2, Mean Squared Error y Mean Absolute Percentage Error, para los modelos regresivos (Powers, 2007).

Tabla 5. Tabla de confusión (o matriz de errores) con las diferentes métricas disponibles para validar los algoritmos de Machine Learning. Fuente: (Wikipedia, 2019)

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

A continuación, se expone una breve descripción de las medidas de desempeño anteriormente nombradas:

Accuracy: es el porcentaje de tuplas (listas ordenada de elementos) del conjunto de prueba que fueron correctamente clasificadas por el modelo (UNLU, Universidad Nacional de Luján , 2015)**.**

Precision: es la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa y está dada por tp / (tp + fp) donde tp es el número de positivos verdaderos y fp el número de falsos positivos, siendo como resultados 1 el mejor valor y 0 el peor (Pedregosa, y otros, 2011).

Recall: es la relación tp / (tp + fn) donde tp es el número de verdaderos positivos y fn el número de falsos negativos. Entonces puede decirse que es la capacidad del clasificador para encontrar todas las muestras positivas (Pedregosa, y otros, 2011).

F1-score: puede ser interpretado como un promedio ponderado de precision y el recall, donde un puntaje F1 alcanza su mejor valor en 1 y el peor puntaje en 0. “En el caso de clases y etiquetas múltiples, esté es el promedio del puntaje F1 de cada clase con una ponderación que depende del parámetro promedio” (Pedregosa, y otros, 2011).

R^2: es la función de puntuación de regresión o coeficiente de determinación, en el cual la mejor puntuación posible es 1.0 y dicha puntuación puede ser negativa (debido a que el modelo puede ser arbitrariamente peor). Un modelo constante que siempre predice el valor esperado de y, sin tener en cuenta las características de entrada, obtendría una puntuación R^2 de 0.0 (Pedregosa, y otros, 2011).

Mean Squared Error: el MSE “mide el error cuadrado promedio de las predicciones. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el objetivo y luego promedia esos valores” (Drakos, 2018).

Mean Absolute Percentage: El MAE es un puntaje lineal, en el cual se calcula el error como un promedio de diferencias absolutas entre los valores objetivo y las predicciones, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio (Drakos, 2018).

2.2.12 Conceptos relacionados con el café.

Esta parte de este marco conceptual es tomado de la página web de la FNC (Federación Nacional de Cafeteros, 2019), en este se encuentran muchos de los términos usados en el desarrollo de este trabajo en referencia al café.

Acidez: Sabor primario agradable y positivo que se percibe, con mayor o menor intensidad, en los cafés de la especie arábica. Esta característica desaparece con un tueste acentuado. Junto con el sabor, aroma y cuerpo es uno de los parámetros principales usados por los catadores profesionales en la valoración sensorial del café.

Amargo: Sabor primario del café proporcionado por la cafeína y otras sustancias. Es agradable dentro de un determinado límite. Usualmente causado por exceso de tueste.

Aromas: Los aromas del café se despliegan únicamente bajo el efecto del calor, durante la torrefacción. Un fenómeno complejo de mutaciones entre proteínas, ácidos e hidratos de carbono provoca la emanación de muchísimos aromas (se pueden contar hasta un millar de compuestos aromáticos diferentes). Estas notas aromáticas pertenecen a familias de aromas muy diferentes, tales como las etc. Estas notas son volátiles y frágiles. Es conveniente que después de la empireumáticas, florales, maderosas, especiadas, afrutadas y otras balsámicas, torrefacción se proceda muy rápidamente a envasar el café, preferentemente en envases metálicos (sistema que permite una mejor conservación de los aromas).

Beneficio húmedo: Se realiza mediante la utilización de agua. Comprende el despulpado, desmucilaginado o la fermentación y el lavado y el secado. Por esta vía se obtienen los llamados cafés lavados, finos o suaves.

Broca del café: Este insecto del grano del café es cilíndrico, de color negro y de una longitud de 1.50 mm aproximadamente. Las hembras ponen los huevos en el interior de los propios frutos en donde se desenvolverán. Las posturas son de 12 a 33 huevos, que necesitan entre 3 y 14 días para su eclosión, según sea la temperatura y la humedad. Las

larvas, que son blancas, necesitan entre 28 y 50 días para trocarse en ninfas. Durante ese período se alimentan de preferencia de la semilla verde y tierna del café.

Café almendra / café verde en almendra / Endospermo / café trillado: A partir de la cual se produce el café tostado y molido. Es básicamente el café pergamino sin su parte externa.

Café Cereza: Es el fruto del arbusto de café –cafeto-, que como tal se recoge en las fincas, en las épocas de cosecha, en las zonas cafeteras y luego se somete a un proceso de adecuación para que pueda ser comercializado (beneficio húmedo), el cual se realiza en la misma finca cafetera.

Café Despulpado: café procesado por la vía húmeda o sea al que se le ha quitado la pulpa para entrar al proceso de fermentación antes de ser lavado.

Café Excelso: Es una calidad del café almendra / Es aquella almendra producto de un esmerado esfuerzo de selección (muy bien clasificada) en todos sus atributos físicos y sensoriales. Esta selección se realiza en las trilladoras de café.

Café Pasilla: Las pasillas son los granos de café que presentan defectos, como brocados, vinagres, negros, partidos, astillados.

Café Pergamino Seco: Es el producto del beneficio del grano el cual se obtiene después de quitarle la cáscara y el mucílago, lavarlo y secarlo hasta una humedad del 12%. / Nombre del café que comercializa el caficultor al interior del país. El contenido de agua está entre 10-12%.

Café Reposado: El café reposado o café reposo o café viejo o envejecido, es un defecto del café, da un sabor muy amargo y un aroma muy fuerte que evoca el nombre reposo o viejo, Sucede cuando los granos se almacenan durante mucho tiempo o se ha guardan en condiciones inadecuadas con alta humedad del grano, por encima del 12%, y en sitios muy calientes por encima de 25 °C.

Café Semitostado (Café Desnaturalizado): Café verde en almendra sometido a un proceso térmico cuyo objetivo es el control sanitario y fitosanitario que garanticen la esterilización del producto y la diferenciación de otras materias primas, pues la almendra sometida a este proceso cambia de aspecto (color). Desde el punto de vista normativo se

entiende como aquel con un porcentaje de pérdida de peso no mayor del 10%. durante el proceso térmico. Al igual que el café verde en almendra, no puede consumirse directamente.

Café Tostado: Café obtenido por torrefacción del café verde en almendra, mediante tratamientos térmicos que producen cambios físicos y químicos fundamentales en la estructura y composición de la materia prima, brindándole oscurecimiento en su color y desarrollando las características de sabor de un café tostado.

Café Trillado: Es el café al que se ha sometido al proceso de quitarle sea la cascara cuando se trata de café sin lavar o el pergamino cuando se trata de café lavado.

Café Verde: A este estado del grano de café también se le conoce en ciertos países como “café oro”, “semilla” o “almendra”. Se obtiene después de la remoción del pergamino como resultado del proceso de trilla. El café verde es el insumo básico para la elaboración del café tostado, el soluble y los extractos de café, y es la forma más común en la que es exportado a otros países.

Calidad del café: Clasificación de los cafés de acuerdo con la altitud, variedad botánica, tipo de beneficiado, densidad, tamaño del grano, calidad de taza, color, imperfecciones del grano y la presencia de materia extraña. Cada país establece sus propios estándares de clasificación de la calidad. El Café de Colombia es reconocido mundialmente por su buena calidad, por lo cual se vende a un mayor precio. Esta calidad depende de los cuidados y prácticas que siguen los caficultores, recolectores, procesadores, comercializadores, tostadores y consumidores, en los diferentes procesos, a través de las etapas de la cadena productiva del café

Calidad física: El fruto de café de buena calidad, sano y maduro; el grano pergamino tiene apariencia homogénea, olor fresco característico a café, color amarillo claro y homogénea y sana, olor fresco, color verde azulado, humedad entre el 10% y el 12%; su tamaño varía según la variedad y se mide en mallas de 12/64 a 18/64 de pulgada.

Caramelo (aroma): Este descriptor de aroma presenta reminiscencias con el olor y el sabor que se producen cuando se carameliza el azúcar sin llegar a quemarlo. El catador deberá tener cuidado de no confundirlo con el aroma a chamuscado.

Catación de Café: Es el método usado para medir el aroma, el sabor y la canidad del café. Mediante las evaluaciones sensoriales se pueden identificar los defectos presentes en la bebida de café, conocer la intensidad de una característica sensorial como la acidez y el dulzor, reconocer y calificar el sabor y el aroma, y de igual forma, medir la calidad global del producto.

Cosecha Principal: En Colombia es el más importante de los dos periodos de recolección realizados en el curso de un año.

Cosecha Secundaria: Recolección de café de menor importancia, llamada "traviesa" o "mitaca", efectuada comúnmente 6 meses después de la cosecha principal.

Cup Testing: Expresión que significa prueba de la taza. Es decir, juzgar los méritos de un café por medio de tostarlo y hacer una taza de café negro para determinar si tiene cuerpo y si es fuerte, rico, ácido o suave; si es vinoso, neutral, agrio, si tiene el sabor viejo, a tierra, madera, si es agrio, etc. El experto retiene un poquito de café en la boca, solamente durante el tiempo necesario para saborearlo bien en toda su fuerza, después de lo cual se deposita el café en una especie de escupidera.

Coups of Coffee to the Pound: Expresión que significa el número de tazas de café sacadas de una libra. El promedio es de 40 tazas

Dulzura: Se trata de un descriptor básico del sabor que deriva de las soluciones de sacarosa o frutosa, las cuales se asocian normalmente con descriptores de aromas dulces, como el de la fruta, el chocolate o el caramelo. Se utiliza comúnmente para describir cafés que están libres de olores rancios.

Fisiología y Catación de Café: La clasificación del café por cualidades exige gran práctica, aunque nos llegue ser una ciencia. La educación del paladar es el punto principal. Se sabe que todo órgano o sistema orgánico se perfecciona por el uso metódico. El punto esencial es conocer exactamente la definición básica de cada descripción y fijarla, ya que cada una de ellas influirá en el precio del producto. Un buen clasificador debe tener gran agudeza de los sentidos. La vista, el gusto y el olfato son los más importantes para juzgar la consistencia del grano. El tacto sirve para estimar su consistencia. Aún el oído se utiliza para apreciar el estado de secación por el ruido que producen los granos al caer unos sobre otros. La

degustación, o sea el acto de probar, requiere utilizar del sentido del gusto, el cual debe ser el más desarrollado para poder diferenciar las diversas sanciones.

Floral (aroma): Este sabor se asemeja a la fragancia de las flores. Presenta un ligero aroma a ciertos tipos de flores, entre las que podemos incluir madreselva, jazmín, diente de león y ortiga. Está presente principalmente cuando se percibe un intenso aroma a fruta o verdura, pero raramente aparece en una intensidad excesivamente alta.

Frutoso/Cítrico (aroma): Este aroma recuerda al olor y sabor de la fruta. Este atributo está íntimamente relacionado con el aroma natural de las bayas. La alta acidez de algunos cafés se relaciona con los cítricos. El catador no debe utilizar este atributo para describir el aroma de la fruta verde o excesivamente madura.

Frutos secos (aroma): Hay que tener en cuenta que este aroma se parece al aroma o sabor de los frutos secos frescos (distintos de los rancios) y no al sabor amargo de las almendras.

Quemado/Ahumado (aroma): El olor se parece al del humo que se produce cuando se quema madera. El descriptor se utiliza a menudo para indicar el grado de torrefacción que se encuentra habitualmente en los cafés torrefactos o tostados en horno.

2.3 MARCO ESPACIAL

Esta investigación se enfocó en el café verde de Colombia, recolectado de diferentes regiones, y recopilado por Almacafé, una empresa de la Federación Nacional de Cafeteros que presta el servicio de calificación del café.

3. CAPITULO 3: METODOLOGÍA

3.1 DISEÑO METODOLÓGICO.

Para este proyecto de investigación aplicada, se implementa la metodología de fases por cada objetivo específico, a continuación, se detallan las actividades, herramientas necesarias y resultados esperados en cada actividad.

Objetivo 1: Identificar, medir y recopilar datos o imágenes de variables descriptoras de los granos de café verde, para usarlos como datos de entrada del modelo.		
Actividades	**Herramientas**	**Resultados**
Identificar variables descriptoras de los granos de café.	Revisión documental.	Sección con descripción de los atributos y variables a medir.
Toma de datos y medición de características definidas.	Instrumentos de medición para las variables seleccionadas.	Matriz con resultados de las mediciones de las muestras seleccionadas.
Toma de imágenes.	Cámara fotográfica	Conjunto de imágenes de granos de café por cada muestra a analizar.
Realizar análisis estadístico de los datos recopilados	Software Python.	Estadísticos descriptivos, gráficos de dispersión, correlación entre variables, histogramas.

Objetivo 2: Realizar evaluación tradicional de café líquido, para obtener las calificaciones de las diferentes variables de café, las cuales se usarán como salidas o etiquetas del modelo.		
Actividades	**Herramientas**	**Resultados**
Revisión de métodos disponibles de calificación de la calidad del café.	Revisión documental	Capitulo descriptivo con los diferentes métodos encontrados, detalle de los criterios empleados y selección del método a usar en la investigación.
Calificación de muestras de café.	Catadores profesionales de Almacafé	Calificación de la calidad del café de las muestras bajo estudio, según el método seleccionado.
Realizar análisis estadístico de los datos recopilados	Software Python.	Estadísticos descriptivos, gráficos de dispersión, correlación entre variables, histogramas.

Objetivo 3: Entrenar, evaluar y ajustar los algoritmos definidos de aprendizaje automático con los datos recopilados.		
Actividades	**Herramientas**	**Resultados**
Construcción de base de datos.	Software Python	Base de datos con las variables medidas a cada muestra y los resultados de calidad evaluados.
Preparación de los datos para ingresar al modelo.	Software Python	Estadísticos descriptivos, gráficos de dispersión, correlación entre variables, histogramas. Datos normalizados. División de los datos entre conjunto de entrenamiento y prueba

Preparar imágenes tomadas.	Software TensorFlow	Base de datos de imágenes del café verde etiquetas listas para entrenar el modelo.
Elección y entrenamiento de los algoritmos para la predicción de la calidad	Software Python	Modelo entrenado, resultados de indicadores de desempeño del modelo.
Entrenamiento de red neuronal para el reconocimiento de imágenes.	Software Python, API Keras, TensorFlow	Modelo entrenado, resultados de indicadores de desempeño del modelo.
Ajuste de parámetros del modelo entrenado.	Software Python, API Keras, TensorFlow	Modelo ajustado con mejoras en sus indicadores de desempeño.

Objetivo 4: Validar la efectividad del modelo de aprendizaje automático implementándolo, con muestras de café no consideradas en el estudio.		
Actividades	**Herramientas**	**Resultados**
Calificación de muestras nuevas de café real.	Catadores profesionales de Almacafé	Calificación de la calidad del café de las muestras bajo estudio, según el método seleccionado.
Realizar validación cruzada	Librería de sci-kit learn para validación cruzada	Descripción de las medidas de desempeño obtenidas, de los modelos con mejor desempeño.
Validación de resultados y conclusiones.	Microsoft Word	Reporte final.

3.2 RECOLECCIÓN DE DATOS PARA EL MODELO DE CALIDAD DE CAFÉ.

Los datos recolectados corresponden a muestras de diferentes partes del país de granos de café almendra, analizadas conforme llegaron a las instalaciones del laboratorio de Almacafé (Almacafé, 2019) ubicado en la ciudad de Bogotá durante los meses de mayo, junio y julio de 2019. A continuación, se relacionan las variables medidas, el equipo y el método de medición:

Tabla 6. Resumen de variables de entrada medidas. Fuente: elaboración propia.

Variable	**Unidad de medida**	**Equipo de medición**	**Método de medición**	**Tamaño de la muestra**
Altitud	metros	Google Earth	Ubicación aproximada según la ubicación del lugar proveniente de la muestra.	N/A.
Humedad	%	MEDIDOR DE HUMEDAD KAPPA	AL-CC-EC-P-0003 Café pergamino – Café Excelso: Determinación de humedad (kappa)	400g
Distribución granulométrica	% de retención por malla	ZARANDA MECANICA	AL-CC-EC-P-0008 Café Verde: Distribución Granulométrica.	210gr aprox.
Color	Cielab	Espectro fotómetro	Medición directa	
Fotografía	Pixeles	Cámara Samsung	Medición directa o de captura	3264x2448 pixeles
Calidad del café	Escala ordinal	Catador profesional	AL-CC-EC-I-0002 ANÁLISIS SENSORIAL PREPARACIÓN DE LA PRUEBA	5 tazas de café

Altitud. Corresponde a la altitud del lugar del cultivo medido en metros sobre el nivel del mar.

Humedad: Se mide el % de humedad presente en el grano, de acuerdo con la resolución 02 de 2016 expedida por el Comité Nacional de Cafeteros no debe estar por encima del 12.5%, medida en equipos calibrados según método ISO 6673. Para este trabajo se midió la humedad en granos de café es su estado “pergamino”.

Peso inicial pergamino. Se abrevia como “PESO.INI.PERG”, y corresponde al peso del grano de café en su forma pergamino.

Peso almendra. Se abrevia como “PESO.ALMEN”, y corresponde al peso del grano de café en su forma almendra, es decir después de trillar el pergamino.

Merma: es el porcentaje resultante de dividir la diferencia entre el peso del pergamino y la almendra, dividido el peso del pergamino.

Distribución granulométrica: para determinar el tamaño se emplean mallas o tamices de diferentes tamaños. Se considera que un café es de calidad superior si (entre otros requisitos) su tamaño está por arriba de la malla 15/64 de pulgada. De acuerdo con la experiencia de Almacafé, un grano pequeño, puede ser un grano que no se desarrolló completamente, causando que sus atributos no sean óptimos. En la siguiente imagen (tabla 7) se aprecia un ejemplo de los resultados de estas mediciones. El fondo corresponde al último nivel o la base del conjunto de tamices.

Tabla 7. Ejemplo de resultados del tamaño del grano en porcentaje. Fuente: Almacafé.

Distribución de tamaño de grano (%)							
%Malla 18/64"	%Malla 17/64"	%Malla 16/64"	%Malla 15/64"	%Malla 14/64"	%Malla 13/64"	%Malla 12/64"	% fondo
29,09	33,81	24,78	8,25	3,51	0,00	0,54	0,02
30,17	32,00	24,40	9,40	2,99	0,00	0,97	0,02
36,77	35,52	17,81	6,63	2,63	0,00	0,67	0,05
25,51	31,88	24,98	11,31	5,06	0,00	1,22	0,09

Figura 17. Ejemplos de granos de café con defecto de o de color 3 y 4. Fuente: Almacafé.

Granos de color grupo1: en esta variable se registra el número de granos identificados manualmente en una muestra de aproximadamente 200 gramos de café almendra, con los defectos de color pertenecientes al grupo 1 de la resolución 02 de 2016 emitida por el Comité Nacional de Cafeteros. En la base de datos se relaciona como GRANOS.G1. Este primer grupo hace relación a colores del café asociados a baja calidad como se muestra en la figura 17. En la variable PESO.GRANOS.G1 se coloca el peso en gramos de los granos identificados como defectuosos del grupo1.

Granos grupo2: es la cantidad de granos defectuosos pertenecientes al grupo 2 de la resolución 02 de 2016 emitida por el Comité Nacional de Cafeteros. Estos defectos hacen referencias a irregularidades en la superficie o geometría de los granos. En la base de datos se identifica como GRANOS.G2. el número y el peso como PESO.GRANOS.G2

Grano brocado punto: hace referencia al número de granos identificados manualmente con presencia de brocado en forma de pequeños orificios. Se coloca el número en la base de datos como GRANO.BROC.P y el peso como PESO.GRANO.B.P.

Total grano defectuoso: se suma el conteo de los defectos grupo 1 grupo 2 y brocado en la columna GRANOS.DEF y peso en la columna PESO.DEF.

Porcentaje grano defectuoso: Relacionada como PORCENTAJE.DEF, es el resultado de dividir PESO.DEF entre PESO.ALMEN.

Fotografía: Se tomaron 2 fotografías a una muestra de 200gr de café en almendra y a la misma muestra después de tostado con una cámara de un dispositivo Samsung SM-P585M una resolución de 3264 x 2448 pixeles, sobre un fondo negro. La foto corresponde a la muestra de granos resultantes después de la identificación y exclusión de granos defectuosos realizados por un analista de Almacafé. En la tabla 8, se muestran los detalles de cada imagen y de la cámara.

Tabla 8. Características de las fotos de granos de café adquiridas

Cámara		Imagen	
Fabricante de cámara	samsung	Id. de imagen	J08LLJA00AM J08LLL...
Modelo de cámara	SM-P585M	Dimensiones	3264 x 2448
Punto F	f/1.9	Ancho	3264 píxeles
Tiempo de exposición	1/30 s	Alto	2448 píxeles
Velocidad ISO	ISO-125	Resolución horizontal	72 ppp
Compensación de exposición	0 paso	Resolución vertical	72 ppp
Distancia focal	3 mm	Profundidad en bits	24
Apertura máxima	1.85	Compresión	
Modo de medición	Promedio central pond...	Unidad de resolución	2
Distancia al objeto		Representación del color	sRGB
Modo de flash	Sin flash	Bits comprimidos/píxel	
Intensidad de flash			
Longitud focal de 35 mm	27		

Como se mencionó en la sección 2.1.3, se han realizado varios trabajos con Machine learning aplicado el café, mediante clasificación de los granos de café verde, en este trabajo se incluye un paso adicional y es la inclusión de la foto de café tostado, así como la verificación con la catación. En la figura 18, se muestra una imagen de una de las muestras de almendra verde y su correspondiente imagen después del proceso de tostado

Figura 18. Izquierda, muestra 1 imagen de grano almendra verde seleccionado. Derecha, imagen del grano tostado de la muestra 1.

Color: Este se mide con un espectrofotómetro en la escala CIELAB y CIEXYZ definido por la La Commission Internationale de lÉclairage (CIE), según (Tobijaszewska, Mills, & Jøns, 2018), los espectrofotómetros miden y recopilan todo el espectro de cada longitud de onda y mediante un algoritmo lo transforman en valores que asemejan la percepción humana. Para este trabajo se midió el color de los granos de café verde y de los granos tostados, para cada uno de estos dos granos se registra en la base de datos las coordenadas de espacio de color L*, a*, b*, X, Y y Z., relacionadas en la base de datos como ALM.VERD.L* y CAFÉ.TOS.L*, modificando la letra final por cada característica.

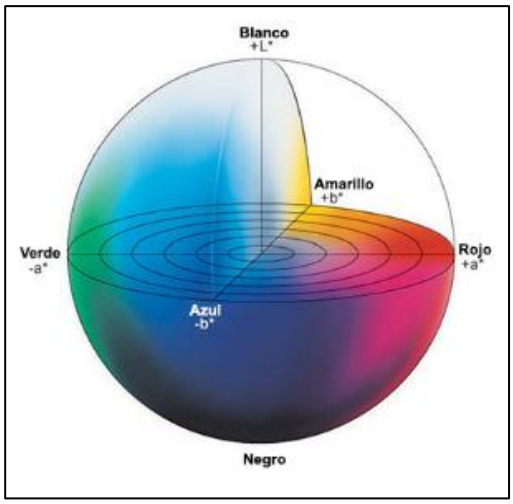


Figura 19. Espacio de color CIELAB. L alterna entre el negro (0) y luminosidad (100). a* alterna entre rojo (positivo) y verde (negativo). b* alterna entre amarillo (positivo) y azul (negativo). Fuente: http://sensing.konicaminolta.com.mx/2014/09/entendiendo-el-espacio-de-color-cie-lab/*

Adicional a estos datos se cuentan con metadatos como el departamento y el municipio.

Evaluación sensorial: se miden 10 descriptores de calidad de café en una escala ordinal de 1 a 10, estos son aroma, sabor, sabor residual, Acidez, cuerpo, balance, uniformidad, limpieza de taza, dulzura y puntaje general de taza. Para realizar esta medición se realiza el tostado de los granos de café verde de cada muestra, bajo un método estandarizado y con condiciones controladas de Almacafé, posteriormente se muele el café tostado y se preparan 5 tazas de café, la cuales son evaluadas por catadores profesionales.

En la tabla 9, se muestra un ejemplo de los resultados realizados a muestras de ejemplo. Una precisión importante respecto a los catadores es que ellos llegan al lugar de la evaluación, sin ningún conocimiento de los resultados de los análisis físicos realizados a los granos de café, o de trazabilidad del cultivo, permitiéndoles no tener ningún sesgo en el momento de realizar las calificaciones a las respectivas tazas de café.

Tabla 9. Ejemplo de resultados de atributos de calidad del café medidos por catadores profesionales. Fuente: Elaboración propia.

N° Muestra	Fragancia/ Aroma	Sabor	Sabor Residual	Acidez	Cuerpo	Balance	Uniformidad	Taza Limpia	Dulzor	Overall
1	6	5,5	5	5,5	5	5	5,5	5,5	5,5	5
2	5,5	5	5	5	4,5	5	4,5	4,5	5	5
3	5,5	5	5,5	5,5	5	5	5	5	5,5	5
4	6	5	5,5	5,5	5	5,5	5,5	5,5	5,5	5,5
5	5	5	5	5	5	4,5	5	5	5,5	5
6	5	5	4,5	5	4,5	4,5	4,5	4,5	4,5	4,5
7	4	4	4	4,5	3,5	4	4	3,5	4	4
8	5,5	4	3,5	4	3,5	3,5	4	3	3	3,5

Adicional al puntaje asignado a cada descriptor de la calidad del café, los catadores, relacionan, descripciones para el aroma, el sabor, el sabor residual, y dan una descripción cualitativa general de la muestra, como se muestra en la tabla 10. En total se completaron 56 análisis de muestras, con las variables anteriormente descritas.

Tabla 10. Ejemplo de los atributos cualitativos dados por los catadores a atributos de calidad del café. Fuente: elaboración propia.

Código de identificación	Descriptor 1 Fragancia/Aroma	Descriptor 2 Fragancia/Aroma	Descriptor 1 Sabor	Descriptor 2 Sabor	descriptor1,s.residual	Descriptor 2 S.Residual	Notas Generales
1	dulce	chocolate	astringente	dulce	astringente		dulce, limpio, astringente
2	floral	dulce	dulce	limpio	dulce		limpio, dulce
3	herbal	canela	astringente	herbal	astringente	herbal	astringente, amargo, herbal
4	panela	inmaduro	suave		floral	suave	suave, floral
5	caramelo		inmaduro		inmaduro		inmaduro, astringente

Totas las mediciones realizadas por Almacafé, esta soportadas en normas y procedimientos internos aprobados por la FNC y que se han construido de la experiencia y de estándares internacionales de clientes en el exterior. En la siguiente tabla (11), se relacionan las normas que rigen las mediciones realizadas de las muestras empleadas para esta investigación.

Tabla 11. Normas que soportan los métodos de medición realizados por Almacafé.

DESCRIPCIÓN	FECHA EXPEDICIÓN
RESOLUCIÓN 02/2016. POR LA CUAL SE UNIFICAN Y ACTUALIZAN LAS NORMAS DE CALIDAD DEL CAFÉ VERDE EN ALMENDRA PARA EXPORTACIÓN	2016-04-25
ISO 8586:2012 SENSORY ANALYSIS -- GENERAL GUIDELINES FOR THE SELECTION, TRAINING AND MONITORING OF SELECTED ASSESSORS AND EXPERT SENSORY ASSESSORS. Corrected version (en): 2014-06	2014-06
ISO 6668: 2008 GREEN COFFEE - PREPARATION OF SAMPLES FOR USE IN SENSORY ANALYSIS	2008-05
NTC 5248:2013 “CAFÉ VERDE. ANÁLISIS DEL TAMAÑO. TAMIZADO MANUAL”.	2013-08
NTC 2442:2004 CAFE TOSTADO EN GRANO Y/O MOLIDO. DETERMINACION DEL GRADO DE TOSTION.	2004-02

La base de datos construida para este trabajo reunió procedimientos o mediciones adicionales a los realizados habitualmente por Almacafé, como lo fue la medición de color, las fotos de los granos y la evaluación sensorial ampliada por parte de los catadores, el objetivo fue sacar la mayor cantidad de variables por muestra, para encontrar cuales podrían aportar más a la definición de las diferentes métricas y permitir al modelo entrenado lograr una mejor precisión. Otra nota adicional, es que, aunque la escala de resultados posibles va de 0 a 10 (con incrementos de 0.5), las muestras medidas solo arrojaron resultados entre el rango de 2 a 7.

3.3 ANÁLISIS Y PREPARACIÓN DE LOS DATOS RECOLECTADOS.

En esta sección se realiza el análisis de los datos recolectados por Almacafé, descritos en la sección 3.2. Dado que se tenían hojas de cálculo en Microsoft Excel para la recolección de los datos, no fue necesario realizar una limpieza exhaustiva de los datos por lo que se

procede a continuación a describir los pasos realizados para el análisis de los datos de Almacafé.

El primer paso consistió en agrupar las 2 bases de datos en una sola. Como se ejemplifica en la figura 20, los datos de entrada estaban en 2 hojas de cálculo y los de salida en otra. Adicional, se renombran las variables para facilitar su interpretación, dejarlas en una sola fila y/o reducir su tamaño, facilitando así la visualización en el cuaderno de trabajo en Jupyter.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Q	R	S	T	U	V	Y
1	Código interno de muestra	FECHA DE ANÁLISIS	Material	humedad en pergamino	Peso inicial Pergamino (g)	peso almendra (g)	merma (%)	Distribución de tamaño de grano (%)								Granos de color (grupo I). Negro,vinagre,ambar,reposo		Otros defectos (Grupo 2)		Gr. Brocado de punto		Total
2								%Malla 18/64"	%Malla 17/64"	%Malla 16/64"	%Malla 15/64"	%Malla 14/64"	%Malla 13/64"	%Malla 12/64"	% fondo	Cantidad de granos	peso (g)	Cantidad de granos	peso (g)	Cantidad de granos	peso (g)	Cantidad de granos
3	001	2019-08-23	pergamino	11,32	250,38	207,88	16,97	29,09	33,81	24,78	8,25	3,51	0,00	0,54	0,02	5	0,71	33	4,72	0	0,00	38
4	002	2019-09-13	pergamino	8,74	250,42	206,51	17,53	30,17	32,00	24,40	9,40	2,99	0,00	0,97	0,02	1	0,26	8	1,25	1	0,16	10
5	003	2019-09-13	pergamino	10,57	250,20	208,94	16,49	36,77	35,52	17,81	6,63	2,63	0,00	0,67	0,05	5	0,61	22	3,23	4	0,69	31
6	004	2019-09-13	pergamino	10,64	250,14	207,22	17,16	25,51	31,88	24,98	11,31	5,06	0,00	1,22	0,09	1	0,11	12	1,51	8	1,12	21
7	005	2019-09-13	pergamino	9,82	250,58	205,76	17,89	20,87	28,19	27,94	13,69	6,84	0,00	2,39	0,05	0	0,00	39	4,39	11	1,67	50
8	006	2019-09-13	pergamino	9,96	250,29	210,03	16,09	46,98	28,59	14,44	7,41	1,82	0,00	0,74	0,02	4	0,58	20	2,61	1	0,07	25
9	007	2019-09-13	pergamino	9,76	250,07	210,77	15,72	30,12	34,17	22,36	9,19	2,98	0,00	0,63	0,08	2	0,27	14	1,78	6	0,89	22
10	008	2019-09-13	pergamino	10,44	250,16	208,96	16,47	24,21	34,22	23,15	11,88	5,44	0,00	1,09	0,00	2	0,20	26	3,90	15	2,25	43
11	009	2019-08-15	pergamino	9,08	250,18	205,81	17,74	26,47	31,71	23,88	10,23	5,32	0,00	2,08	0,26	1	0,10	35	4,26	8	1,29	44
12	010	2019-08-15	pergamino	10,03	250,45	206,86	17,40	33,81	37,93	19,61	6,21	2,16	0,00	0,24	0,00	0	0,00	17	2,52	7	1,00	24

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Código de identificación	Fecha cata	Fragancial/Aroma	Descriptor 1 Fragancial/Aroma	Descriptor 2 Fragancial/Aroma	Sabor	Descriptor 1 Sabor	Descriptor 2 Sabor	Sabor Residual	descriptor1,s.residual	Descriptor 2 S.Residual	Acidez	Cuerpo	Balance	Uniformidad	Taza Limpi	Dulzor	Overall
2	1	2019-09-06	6,5	dulce	chocolate	5	astringente	dulce	4,5	astringente		5,5	4,5	5	5,5	5	5,5	5
3	2	2019-09-06	6	floral	dulce	6	dulce	limpio	6	dulce		6	5,5	5,5	5	5,5	5,5	5,5
4	3	2019-09-06	5,5	herbal	canela	4,5	astringente	herbal	4	astringente	herbal	5	4	4,5	4	4,5	5	4
5	4	2019-09-06	5,5	panela	inmaduro	6	suave		6	floral	suave	6	6	6	6,5	6	5,5	6
6	5	2019-09-06	5,5	caramelo		5	inmaduro		5	inmaduro		4,5	4,5	5	5	5	5	4

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	MUESTRA	DEPARTAMENTO	MUNICIPIO	ALTITUD	HUMEDAD	PESO.INI.PERG	PESO.ALMEN	MERMA	%M 18/64"	%M 17/64"	%M 16/64"	%M 15/64"	%M 14/64"	%M 13/64"	%M 12/64"
2	1	CUNDINAMARCA	VILLETA	820	11,32	250,38	207,88	16,97	29,09	33,81	24,78	8,25	3,51	0,00	0,54
3	2	CUNDINAMARCA	FALTANTE	3000	8,74	250,42	206,51	17,53	30,17	32,00	24,40	9,40	2,99	0,00	0,97
4	3	CUNDINAMARCA	GUADUAS	990	10,57	250,2	208,94	16,49	36,77	35,52	17,81	6,63	2,63	0,00	0,67
5	4	CUNDINAMARCA	GUADUAS	990	10,64	250,14	207,22	17,16	25,51	31,88	24,98	11,31	5,06	0,00	1,22

Figura 20. En la parte superior se muestra un ejemplo en baja definición de la base de datos de las variables de entrada. En la mitad, la base de datos de las salidas y en la parte inferior, la unificación. Fuente: elaboración propia.

Si bien Python tiene incorporada librerías para leer directamente desde las hojas de datos de Excel, la base de datos se convierte a formato separado por comas (csv), por preferencia propia. A partir de este punto todo se trabaja desde un cuaderno web en Jupyter notebook o Google Colaboratory (Colab).

```
pd.set_option('display.max_columns', None)
df = pd.read_csv("AlmacafeDataBase.csv", sep=";", decimal=",")
display(df.head(3))
print("Dataset dimensions : ", df.shape)
```

\FE.TOS.a*	CAFE.TOS.b*	CAFE.TOS.X	CAFE.TOS.Y	CAFE.TOS.Z	Aroma	Sabor	Sabor.Residual	Acidez	Cuerpo	Balance
8.10	13.43	5.40	5.00	2.94	6.5	5.0	4.5	5.5	4.5	5.0
7.30	11.79	4.81	4.50	2.80	6.0	6.0	6.0	6.0	5.5	5.5
8.07	12.93	5.27	4.88	2.92	5.5	4.5	4.0	5.0	4.0	4.5

Figura 21. imagen de cargue de datos en un cuaderno web (notebook) de Jupyter y visualización de una parte de este.

Uno de los primeros pasos para trabajar con Python es instalar las librerías y programas necesarios, en el marco conceptual se describen los principales para el desarrollo de esta investigación y se pueden visualizar al inicio del código del cuaderno web en el anexo 3. Lo anterior aplica si se está empleando el procesador y la RAM del ordenador usado, es decir desde un entorno local. Para el caso de un entorno alojado como es Colab, muchas librerías están instaladas por defecto y en caso de que sea una muy específica, Colab la identifica y facilita el método para incluirla rápidamente o sugiere alguna alternativa. Por defecto, se trabaja desde un entorno local, no obstante, para algunas pruebas que requieran más capacidad, se empleará un cuaderno web de Colab. Este último es muy similar al de Jupyter, con otra ventaja adicional y es que permite acceder a entonos de ejecución alojados, permitiendo el uso gratuito de hasta 25 GB de RAM, y procesadores GPU de gran capacidad que reducen la velocidad de computación. En ejercicios comparativos realizados se encontró una mejora de hasta 3 veces en la ejecución de los modelos más complejos o con bucles muy grandes. No obstante, la información alojada en Colab se elimina cada 12 horas (Google Colab, 2019) y se requiere una conexión a internet permanente para poder trabajar los archivos. En el entorno local la versión de Python usada es la 3.6.8, esta no es la última versión disponible durante el desarrollo de esta investigación, sin embargo, era la última versión compatible con Tensorflow, librería clave para la ejecución de modelos de redes neuronales como el perceptrón multicapa. La versión de Tensorflow instalada es la 1.14.0 y la de Keras es la 2.2.4, por efectos prácticos se coloca en el anexo 1, la lista de las librerías instaladas con su respectiva versión. Estas librerías y programas se instalan desde la ventana de comandos de Windows. Para acceder a esta ventana se puede colocar el

buscador de Windows, “cmd” u oprimir las teclas Windows+R y escribir cmd. Una nota particular, es que la librería GRAPHVIZ, con la que se grafican los árboles de decisión, requiere de dos acciones de instalación en Windows. Por un lado, la instalación pip (pip graphviz) y por otro, descargar una carpeta de archivos, de la página de esta librería. Una vez se descomprime, se le debe indicar la ruta desde el notebook como se muestra en la figura 22. Se sugiere no copiar esta ruta, dado que se movió de la carpeta descargas y al guardar los archivos se creó una carpeta adicional, se recomienda descomprimir desde la carpeta Descargas y copiar la ruta a esta para facilidad.

```
import os
os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/graphviz-2.38/release/bin'
import graphviz
```

Figura 22. Ejemplo de llamado de la librería graphviz para generar una imagen de algoritmo árbol de decisión. Fuente: elaboración propia.

Seguidamente se realiza un preprocesamiento de los datos, como el realizado en la sección 3.1, se calculan los estadísticos descriptivos (figura 23), se verifica el número de datos (figura 24), se realiza conversión de las variables con cadenas (o tipo objeto) a número, y se procede a realizar una visualización de las distribuciones de estos (figura 25).

	ALTITUD	HUMEDAD	PESO.INI.PERG	PESO.ALMEN	MERMA	%M 18/64"	%M 17/64"	%M 16/64"
count	56.000000	56.000000	56.000000	56.000000	56.000000	56.000000	56.000000	56.000000
mean	1628.464286	10.552857	250.437679	206.194821	17.666071	29.866607	30.087857	23.315000
std	555.439775	0.849658	0.289118	2.477325	0.989256	10.999099	3.269189	6.085673
min	820.000000	8.740000	250.030000	200.070000	15.120000	6.740000	22.170000	9.440000
25%	1230.000000	10.030000	250.195000	204.477500	17.135000	24.360000	28.140000	19.395000
50%	1556.500000	10.640000	250.415000	206.005000	17.735000	29.285000	29.710000	22.995000
75%	1740.000000	11.050000	250.602500	207.740000	18.342500	37.060000	32.262500	25.022500
max	3341.000000	12.880000	251.280000	212.620000	20.070000	60.130000	37.930000	38.680000

Figura 23. Ejemplo de estadísticos descriptivos básicos empleando la función df.describe(). Fuente: elaboración propia.

El resumen de los descriptivos permite validar los rangos de cada dato, conocer su, media, su desviación y sus percentiles.

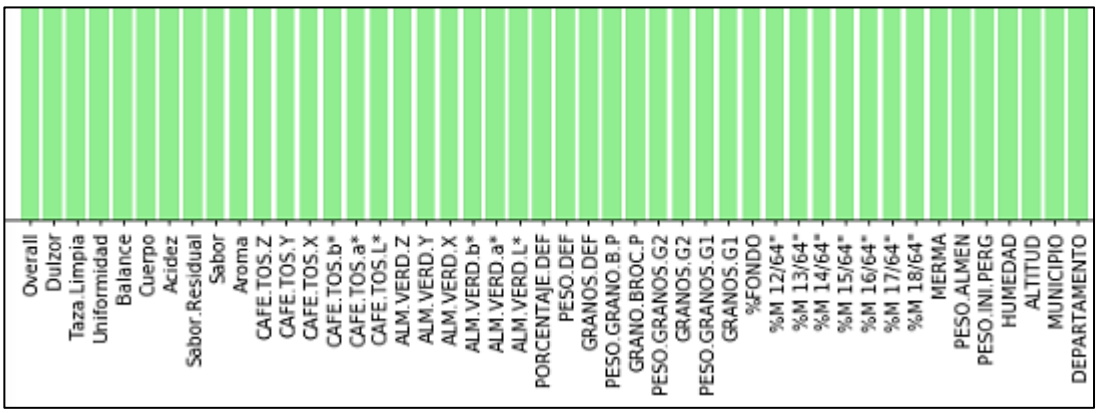


Figura 24. Gráfico de validación de conteo de datos, se aprecia que todas las variables de la base datos tienen las mismas cantidades. Fuente: Elaboración propia usando la librería Matplotlib.

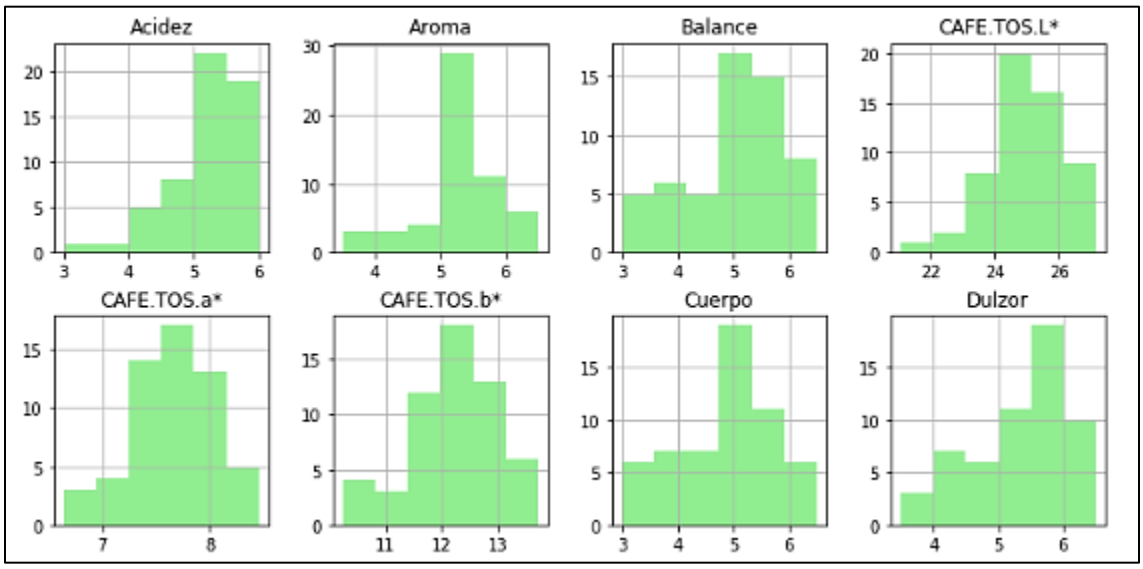


Figura 25. Visualización de los histogramas de las variables presente en el conjunto de datos analizado de Almacafé. Se muestran solo unos ejemplos por la cantidad de variables. Fuente: elaboración propia.

De los histogramas de cada variable se evidencia la dispersión de algunos datos, siendo de interés particular la dispersión de las variables de salida ("Aroma", "Sabor", "Sabor.Residual", "Acidez", "Cuerpo", "Balance", "Uniformidad", "Taza.Limpia", "Dulzor", "Overall"). La importancia de esta dispersión radica en que, si se tienen muchos datos agrupados en algún punto específico de todo el rango, los algoritmos pueden tender a encontrar un optimó local y predecir lo que se denomina la clase mayoritaria, debido a que

esta primera aproximación le dará una baja pérdida en la función de costo para la mayoría de los casos. Tomando de ejemplo la variable Acidez, la mayoría de los datos están dentro del rango de puntaje entre 5 y 6. Esto se debe tener en cuenta para realizar el análisis e interpretación de los resultados. El siguiente diagrama de cajas y bigotes (figura 26) nos confirman la dispersión de los datos, la cual puede sesgar a los algoritmos al darles más importancia a las variables con mayor rango.

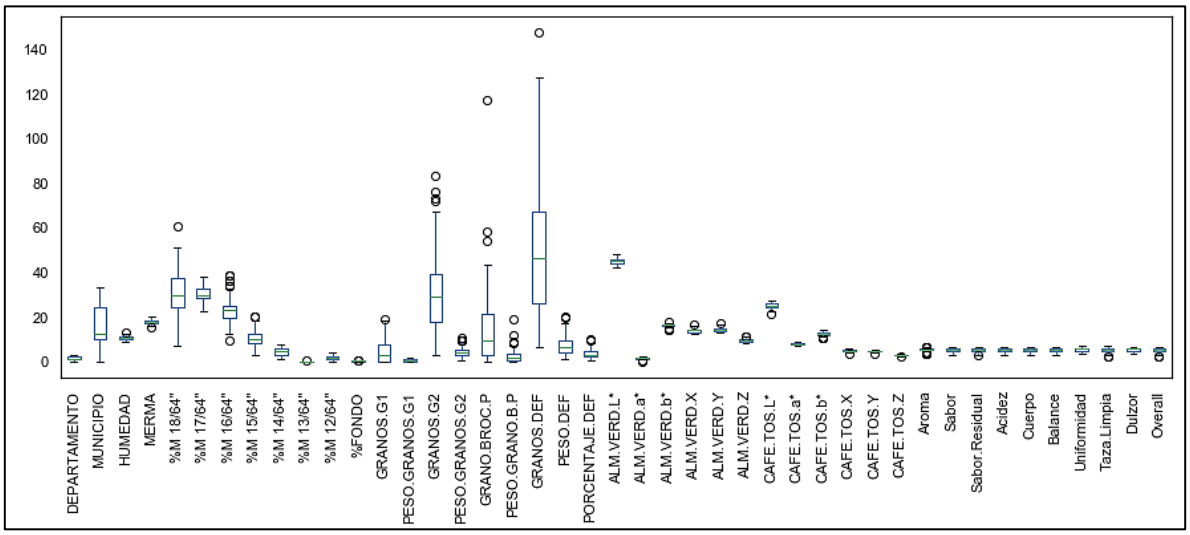


Figura 26. Diagrama de cajas y bigotes de las variables presentes en la base de datos construida

Prosiguiendo con el análisis de los datos, se realizan algunos gráficos adicionales para interpretar y conocerlos aún más, como los mostrados en las figuras 27 a 29.

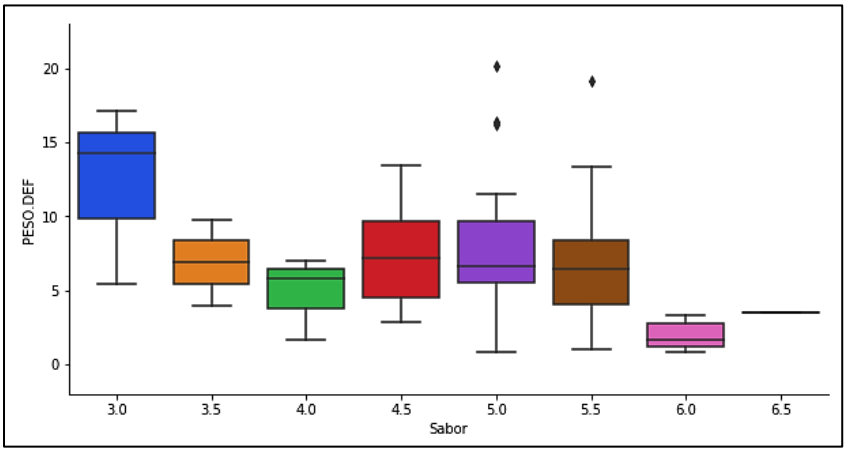


Figura 27. Diagrama de cajas y bigotes entre la salida "Sabor" y la entrada "Peso defectos"

En la figura 27, se aprecia una tendencia inversa leve entre la cantidad de defectos y el puntaje obtenido en la variable de salida “Sabor”. Aunque se podría relacionar una disminución del puntaje, con la aparición de más defectos, se aprecian 2 casos particulares que obtuvieron mediciones de 5 y 5.5, aún con un alto número de defectos. En la figura 28 se aprecia en el gráfico de la izquierda la relación entre “Overall” y la variable de entrada color b* en el grano verde. Un valor de b* más positivo, es decir más en dirección al espectro amarillo, marca una disminución en el puntaje “Overall”. El de la derecha es otra forma de representar la relación mostrada en la figura 27.

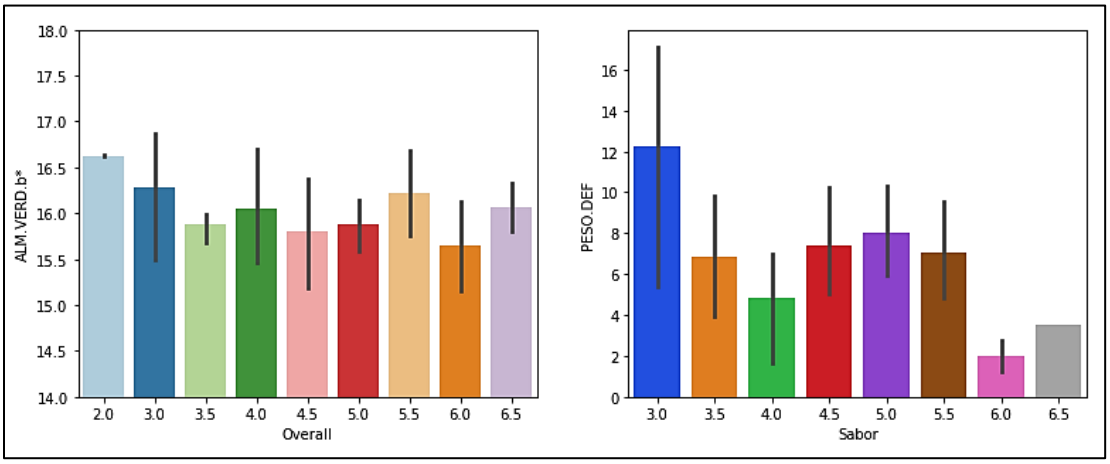


Figura 28. Gráficos de barras. A la izquierda la relación entre la variable “Overall” y el color b en la almendra verde. En la derecha, se muestra la relación entre el “Sabor” y el peso de los granos identificado como defectuosos. Fuente: elaboración propia.*

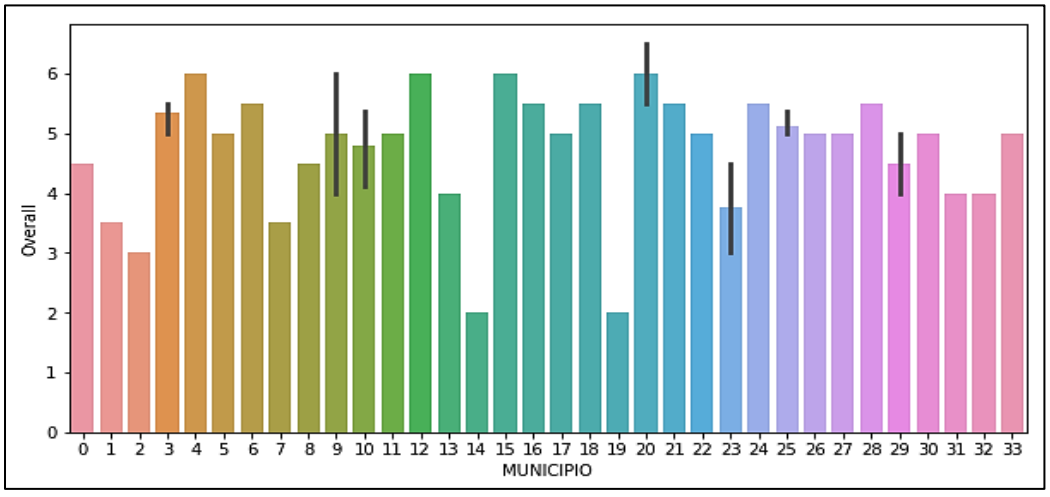


Figura 29. Promedio de puntaje salida “Overall” por Municipio. Fuente: elaboración propia.

La figura 29, se muestra el promedio del puntaje por municipio para la variable “Overall”, siendo Pescador (14) y Piendamo (19) los municipios con el menor puntaje. Es importante, entender que por el número reducido de muestras, no se puede generalizar la calidad de todo el municipio. Para este investigación, esté actua más como un referente de origen de la finca donde esta ubicado el cultivo del que provino la muestra. Si bien estos gráficos permiten la comprensión del comportamiento de los datos en detalle, los sigueintes gráficos de correlación, muestran de una forma visual todo el panorama. Teniendo en cuenta el número de variables, solo se muestran algunos ejemplos.

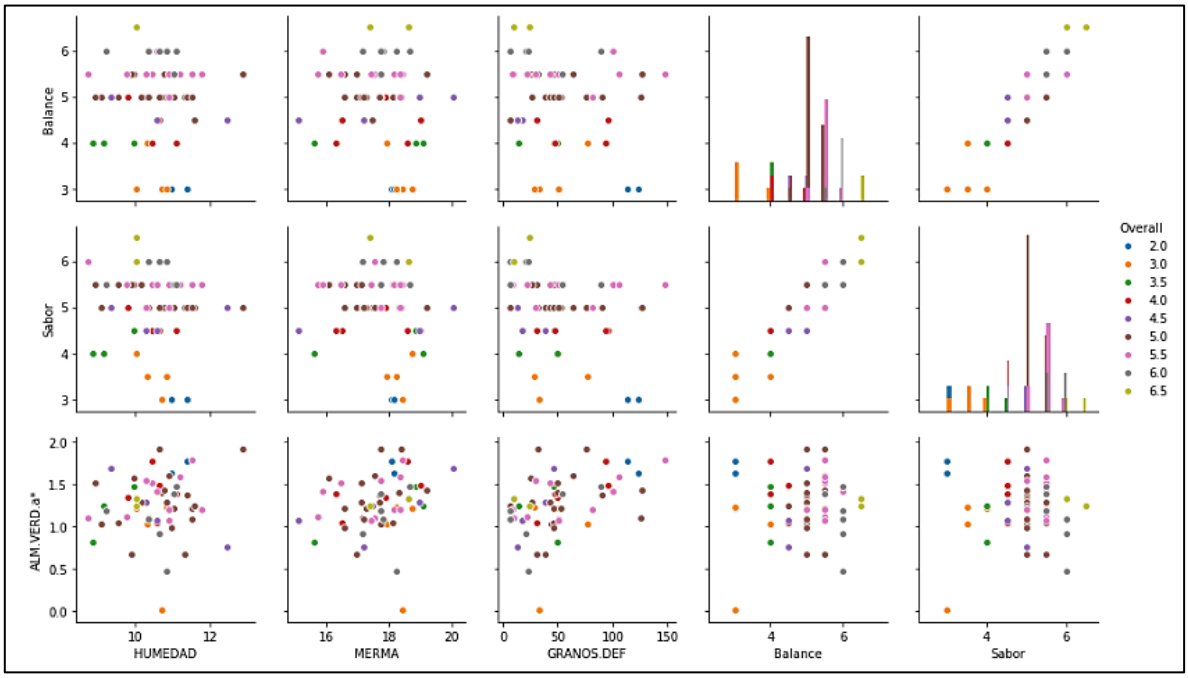


Figura 30. Gráficos de correlación entre algunas variables del conjunto de datos de Almacafé. Fuente: elaboración propia.

En la imagen anterior (figura 30), se muestra las correlaciones entre cada una de las variables del modelo, estos gráficos dan una primera apreciación visual de las relaciones entre las variables. Otra forma de visualizarlas en mediante un mapa de calor. En la figura 31, se aprecia en la parte izquieda una vista general del mapa de calor y a la derecha la correlación entre las varialbes, granos defectuosos totales y el color a* de los granos verdes. Un mayor a* significa que el color del grano tiene más componente del espectro rojo en su color, lo cual es identificado por los expertos de calidad como un defecto de color que afecta la calidad de la bebida del café.

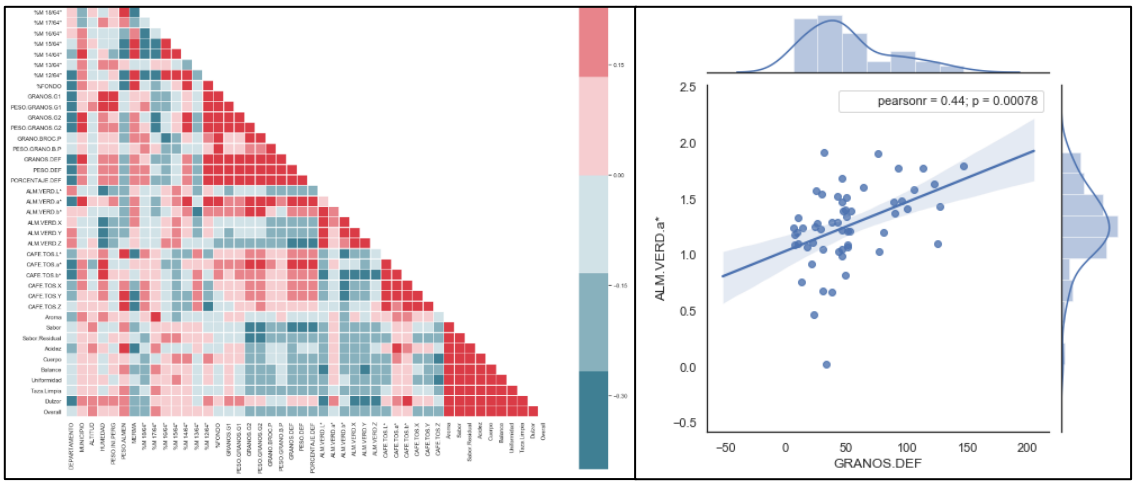


Figura 31. A la izquierda se muestra una vista general del mapa de calor de correlación entre las variables y a la derecha la correlación de las variables Granos defectuosos y el color b de los granos.*

En la figura 31 se muestra el mapa de calor de correlación entre las variables, elaborado mediante la función df.corr() la cual devuelve una matriz con los valores de la correlación Pearson entre las variables del conjunto de datos. Los colores son configurables, así como los rangos, para este caso como las correlaciones eran medias o bajas, es decir cercanas a cero, se colocó el rango en -0.4 a 0.4 (lo normal es de -1 a 1) y se les dio a las correlaciones positivas el color rojo y a las negativas el color azul (esta regla de color solo aplica para el mapa). En la figura 32, se aprecia el detalle del mapa que relaciona las variables de salida (eje Y) con las de entrada (eje X).

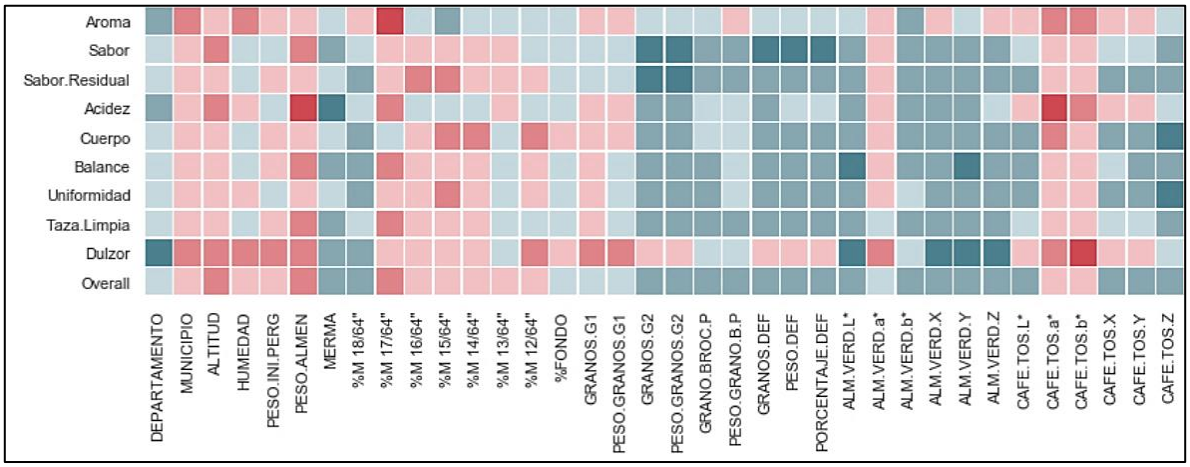


Figura 32. Detalle del mapa de correlación de las variables de salida versus la de entrada. Fuente: elaboración propia.

Por la intensidad de los tonos se aprecia en la imagen anterior que la variable Aroma tiene una correlación medida (según la escala definida en el mapa) con el % en la malla 17/64", el Sabor, presenta una correlación media inversa versus el número de defectos. El Dulzor esta inversamente correlacionado con las mediciones de color y medianamente correlacionado, con variables como la altitud, la humedad, el peso del grano y los granos defectuosos de grupo 1. Aunque en general las correlaciones sean medianas o bajas, se infiere, que el resultado de los catadores guarda una relación con las variables analizadas, por la oficina de calidad de Café. Adicional, se aprecia que el grupo de variables de entrada con más relación versus las de salida, son las de grupo de la escala de color CIELAB.

Retomando las diferencias de rangos entre las variables, esto se corrige realizando una normalización de los datos, es decir, dejarlos todos en la misma escala. La técnica más usada es el escalamiento entre el máximo y mínimo, el cual está dado por la siguiente ecuación, válida para normalizar entre 0 y 1 (Jayalakshmi & Santhakumaran, 2011):

$$X' = \frac{X - v_{min}}{v_{max} - v_{min}}$$

Para este trabajo se usa la función que viene integrada con la librería scikt.learn, en la figura 33, se muestra el fragmento del código con el que se realiza la transformación. Para esto, se crea un nuevo marco de datos (df: Data Frame), llamado df3 en donde se guardan los valores normalizados del marco de datos original df. Para facilitar el entendimiento de esta transformación se muestran nuevamente los gráficos de histograma y de cajas y bigotes.

```
df3 = df.copy()
x = df.to_numpy(copy=True)
min_max_scaler=preprocessing.MinMaxScaler()
x_scaled=min_max_scaler.fit_transform(x)
df2=pd.DataFrame(x_scaled)
for column in list(df2):
    df3[df3.columns[int(column)]]=df2[column].tolist()
```

Figura 33. Código con el cual se normaliza el conjunto de datos Fuente: elaboración propia.

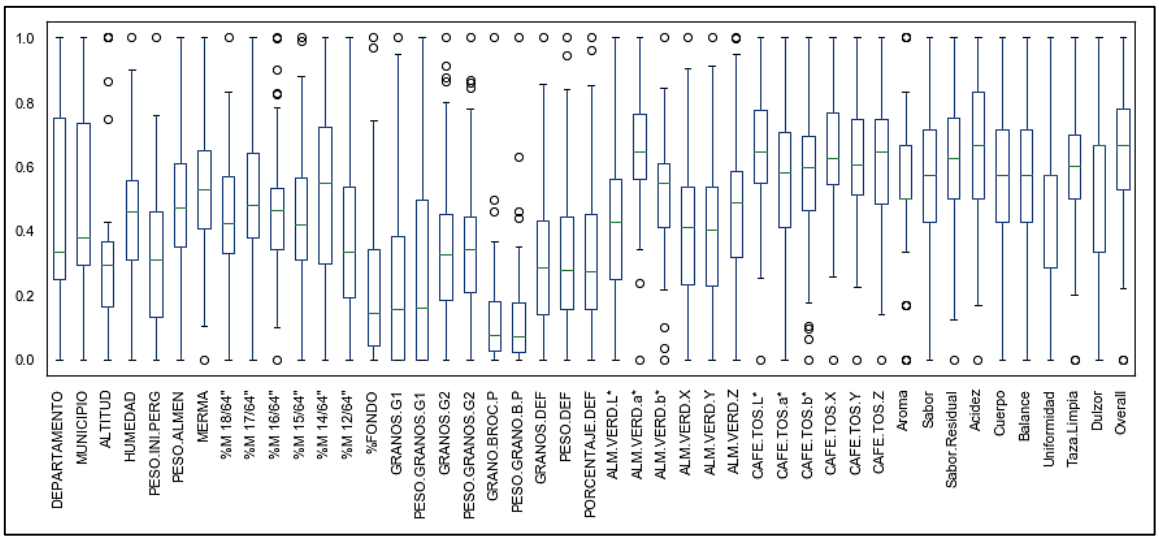


Figura 34. Ejemplo de histograma de variables normalizadas. Fuente: elaboración propia.

Como se aprecia en la figura anterior (34), todas las variables se encuentran en 0 y 1. Esto permite a los algoritmos encontrar las relaciones y determinar que combinaciones de estas, permiten predecir las variables de salida. En la figura 35, se muestran las variables de la figura 33, después de la normalización.

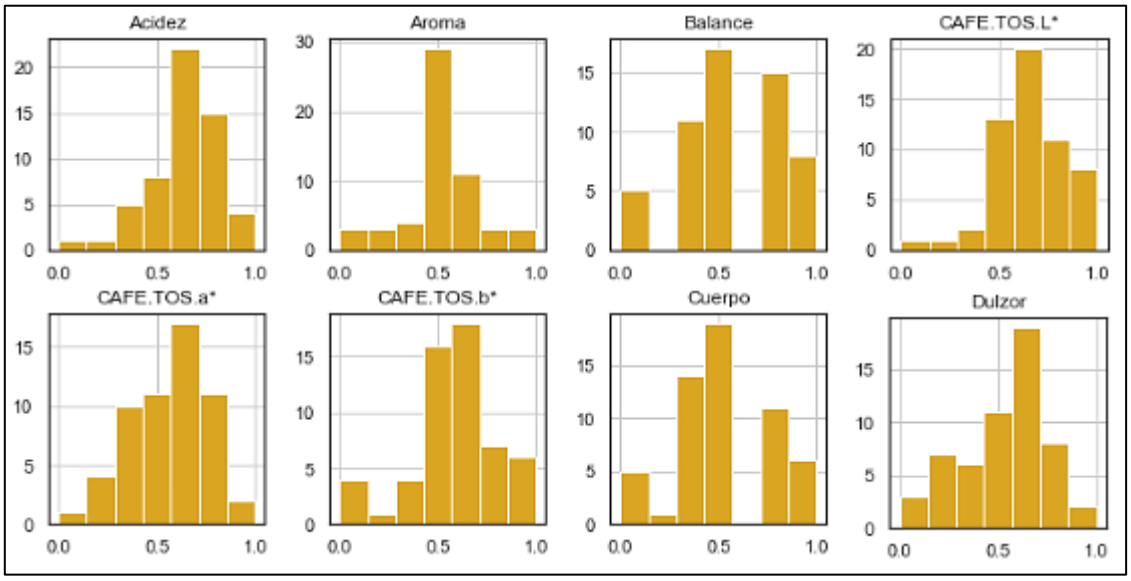


Figura 35. Histograma de variables normalizadas entre 0 y 1. Fuente: elaboración propia.

3.4 ENTRENAMIENTO DE ALGORITMOS PARA LA CALIDAD DE CAFÉ.

Una vez completado el análisis de las variables del conjunto de datos de Almacafé, se procede a implementar los algoritmos definidos. Como se indicó en el marco teórico, los modelos de ML se pueden trabajar bajo enfoque de clasificación o de regresión. Como el objetivo es poder predecir la calidad del café, que en este caso está reportada en 10 variables o descriptores, con una escala finita, se puede optar por proponer grupos o categorías para clasificarlos, bien sea por variable, o por el grupo de estas (ejemplo creando una nueva que las agrupe). Para el caso de regresión, se emplean algoritmos que predecirán ya sea individual o colectivamente los valores de las variables de salida. En esta sección se muestran los diversos enfoques que se trabajaron para cada algoritmo propuesto, producto de la revisión bibliográfica realizada y de su relación con el tipo de datos trabajados. En la siguiente tabla (12) se muestra un resumen de los algoritmos trabajados.

Tabla 12. Resumen de los algoritmos trabajados. Fuente: elaboración propia.

PROPUESTO / SUGERIDO POR:	NOMBRE ALGORTIMO	ENFOQUE	PARÁMETROS
(Palma Méndez & Marín Morales, 2008); **(Barber, 2012)**; **(Nilsson, 1998)**; **(Hackeling, 2014)**	Árbol de decisión	Clasificación	K: profundidad del árbol
(Shalev-Shwartz & Ben-David, 2014); **(Nilsson, 1998)**; **(Russell & Norvig, 2004)**; **(Cuadras, 1989)**	Vecinos cercanos (KNN)	Clasificación	K: Número de vecinos
(Sánchez, 2015); **(Figueira, y otros, 2009)**; **(Gala, 2013)**; **(Chang & Lin, 2013)**	SVM	Regresión	
(Gallo, 2015); **(Matich, 2001)**; **(Shalev-Shwartz & Ben-David, 2014)**	Red neuronal	Clasificación	Ver tabla 7, sección 3.4.4
(Gallo, 2015); **(Matich, 2001)**; **(Shalev-Shwartz & Ben-David, 2014)**	Red neuronal	Regresión	Ver tabla 8, sección 3.4.6
(Simonyan & Zisserman, 2015)	CNN-VGG16	Regresión	Ver tabla 10, sección 3.4.8

3.4.1 Árbol de decisión (análisis variables de salida).

Este algoritmo se ejecuta con la librería scikit-learn, incluyendo un bucle del hiperparámetro k, correspondiente a la profundidad del árbol. En un primer enfoque se aplica a cada una de las variables de salida, porque este algoritmo indica las variables de entrada que más están aportando para la respectiva predicción. Posteriormente se crea una etiqueta de 2 categorías calidad alta y calidad baja, definida a partir de los puntajes de las 10 salidas. El primer paso para la ejecución de estos algoritmos es asignar los datos de entrada una variable (X) y los de salida a otra (y), seguidamente se realiza la división de los datos entre entrenamiento y prueba con la función “train_test_split”, como se muestra en la figura 36.

```
y=df1[b].values
X=df1.drop(["Aroma","Sabor","Sabor.Residual","Acidez","Cuerpo","Balance","Uniformidad",
#print(X)
print(np.shape(X))
```

```
X_trn, X_tst, Y_trn, Y_tst = train_test_split(X, y, test_size=0.2,random_state=1)
print(np.shape(X_trn),np.shape(X_tst))
```

```
(44, 36) (12, 36)
```

Figura 36. Asignación de datos de entrada y de salida a las variables X, y respectivamente.

Para este conjunto de datos, se emplea una división 80% para los datos de entrenamiento y 20% para los datos de validación. Al imprimir las dimensiones de las variables X_trn, X_tst, se aprecia que los datos de entrenamiento tienen 44 registros con 36 variables de entrada, y los de validación 12. Para el caso de la variable de salida y, se inicia por solo asignar los valores de la variable “Sabor”, la cual contiene 8 puntajes diferentes y para efectos de este algoritmo se tratan como valores categóricos.

El siguiente paso, es realizar el entrenamiento del modelo con los datos de entrenamiento (X_trn), mediante la función DecisionTreeClassifier(max_depth=k), la cual, va ajustando la precisión en la predicción de los datos de entrenamiento y validación (X_tst). Adicionalmente se crea un bucle para ir aumentando la profundidad del árbol de decisión. En la figura 37 se aprecia el resultado obtenido por el algoritmo, tratando de clasificar el puntaje de la variable “Sabor”. Se aprecia que, sobre una profundidad de 8, el árbol logra predecir el 100% de los datos de entrenamiento. La mejor precisión para los datos de validación se alcanza hasta la octava profundidad, no obstante, es aún muy baja para los

datos de validación. Preliminarmente se evidencia, que el árbol de decisión no es un buen modelo para predecir la variable “Sabor”, partiendo de los datos de entrada. Lo anterior se explica en parte porque, en esta primera aproximación, los puntajes de la variable sabor se colocaron como categorías independientes, lo que demanda una gran cantidad de datos para lograr un algoritmo que logre diferenciarlas todas. Lo importante, de este primer análisis es el reconocimiento de que variables de entrada influyen más en la predicción de cada descriptor de calidad.

	k	acc_Tree_trn	acc_Tree_tst
0	1	0.409091	0.333333
1	2	0.568182	0.333333
2	3	0.727273	0.250000
3	4	0.840909	0.333333
4	5	0.909091	0.333333
5	6	0.954545	0.333333
6	7	1.000000	0.333333
7	8	1.000000	0.416667
8	9	1.000000	0.250000
9	10	1.000000	0.416667

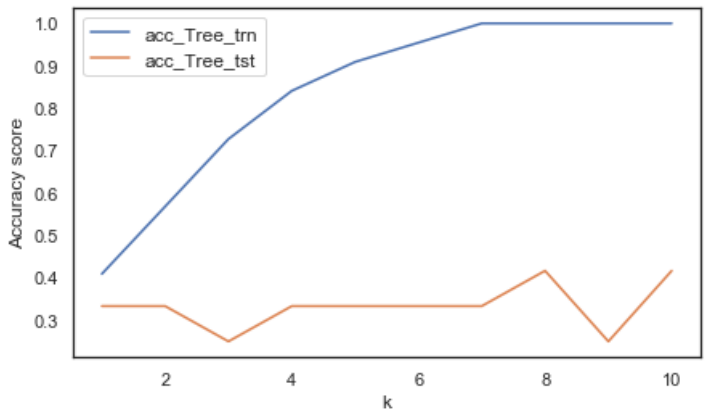


Figura 37. Gráficos de curvas de precisión para los datos de entrenamiento (acc_Tree_trn) y los datos de validación (acc_Tree_tst). Fuente: elaboración propia.

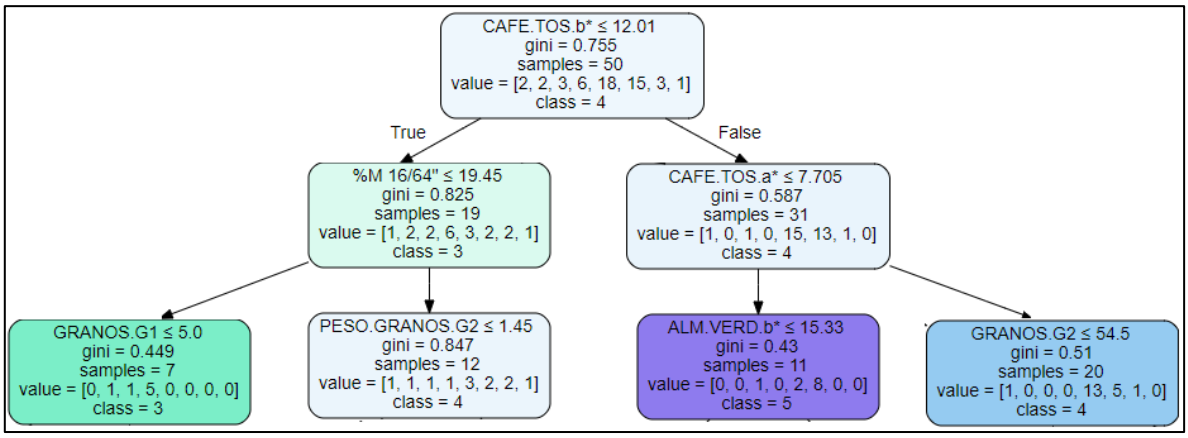


Figura 38. Diagrama de árbol de decisión para 3 profundidades al predecir la variable “Sabor”

En la figura 38, se aprecia un ejemplo del árbol creado usando como etiqueta de salida el “Sabor”. En esta, se aprecia que las 3 variables de entrada que más permiten generalizar

el modelo son el color b* y a* del café tostado y el % de granos retenidos en la malla 16/64". Bajo este enfoque se completa la tabla, en la cual se muestra para cada variable de salida, cual es la profundidad alcanzada que permite el valor de precisión de validación más alto, cuáles son las variables que más aportan y cuál es el resultado de la precisión de validación.

Tabla 13. Resumen de entrenamiento de árbol de decisión para predecir cada descriptor de calidad.

DESCRIPTOR DE CALIDAD	K	VARIABLES CON MAYOR APORTE	PRECISIÓN DE VALIDACIÓN
Aroma	6	%M 14/64", %M 15/64", PESO.GRANOS.G2	0.33
Sabor	8	CAFE.TOS.b*, CAFÉ.TOS.a*, %M 16/64"	0.416
Sabor.Residual	4	PESO.ALMEN, CAFE.TOS.b*, ALM.VERD.a*	0.416
Acidez	7	CAFE.TOS.b*, %M 16/64", MERMA	0.5
Cuerpo	3	MERMA, %M 15/64", ALM.VERD.b*	0.25
Balance	3	CAFÉ.TOS.a*, GRANOS.G2, GRANOS.DEF	0.33
Uniformidad	5	GRANOS.G1, PESO.ALMEN, %M 18/64"	0.33
Taza.Limpia	4	PESO.GRANOS.G2, %M 17/64", ALM.VERD.Z	0.33
Dulzor	5	CAFÉ.TOS.a*, ALM.VERD.b*, PESO.GRANOS.G1	0.416
Overall	7	CAFÉ.TOS.a*, ALM.VERD.a*, PESO.INI.PERG	0.33

De la tabla anterior (13), se puede apreciar, variables representantes de los 3 grandes grupos de datos de entrada, Malla, Color, Defectos, aportando a la determinación de las variables de salida, siendo las de color las más recurrentes y que refuerza la relación indicada por los autores citados, respecto al color de los granos, versus la calidad del café.

3.4.2 Árbol de decisión (clasificación).

Esta parte, se desarrolla creando una etiqueta para dividir la calidad del café obtenida, como alta (1) o baja (0). Iniciando con la definición una variable binaria o etiqueta, a partir de un condicional que asigna el valor, de acuerdo con el resultado de la suma aritmética de los 10 descriptores de calidad. En la figura 39, se muestra la instrucción creada para la creación de la etiqueta, se usa un valor por encima de la media, debido a que los granos por encima de este grupo son de más interés por su calidad.

```
for i in range(len(df5)):
    df5.loc[i,"label"] = 1 if df5.loc[i,"Suma"] >51.5 else 0
```

Figura 39. Bucle para crear una etiqueta binaria, para separar la calidad del café

Como resultado de la separación realizada, obtenemos 28 datos por cada categoría, en la tabla 14, se aprecia los rangos de los descriptores obtenidos. Teniendo en cuenta que la separación se realiza por el puntaje total, se puede ver algunas intersecciones en los rangos, no obstante, el entrenamiento permitirá analizar que tanto se pueden generalizar los datos a partir de estas categorías.

Tabla 14. Categorías creadas de puntaje para la clasificación

Categoría	**Café puntaje Alto (1)**		**Café puntaje Bajo (0)**	
Datos	28		28	
Descriptor	**Valor máx.**	**Valor mín.**	**Valor máx.**	**Valor mín.**
Aroma	6,5	4,5	6	3,5
Sabor	6,5	5	5,5	3
Sabor.Residual	6,5	4,5	5,5	2,5
Acidez	6	5	5,5	3
Cuerpo	6,5	4,5	6	3
Balance	6,5	5	5	3
Uniformidad	7	5	5,5	3,5
Taza.Limpia	7	5	5,5	2
Dulzor	6,5	5	5,5	3,5
Overall	6,5	5	5	2

Siguiendo el proceso descrito en el anterior numeral, se separan los datos y se entrena el algoritmo, obteniendo los resultados mostrados en la figura 40. En contraste con los anteriores, se aprecia que la precisión en validación logró aumentar en un 75% para los datos de validación, con una profundidad de 6. Por lo que se procede a graficar la matriz de confusión y a generar el reporte de clasificación.

	k	acc_Tree_trn	acc_Tree_tst
0	1	0.704545	0.583333
1	2	0.795455	0.666667
2	3	0.886364	0.583333
3	4	0.977273	0.583333
4	5	1.000000	0.500000
5	6	1.000000	0.750000
6	7	1.000000	0.583333
7	8	1.000000	0.666667
8	9	1.000000	0.666667
9	10	1.000000	0.500000

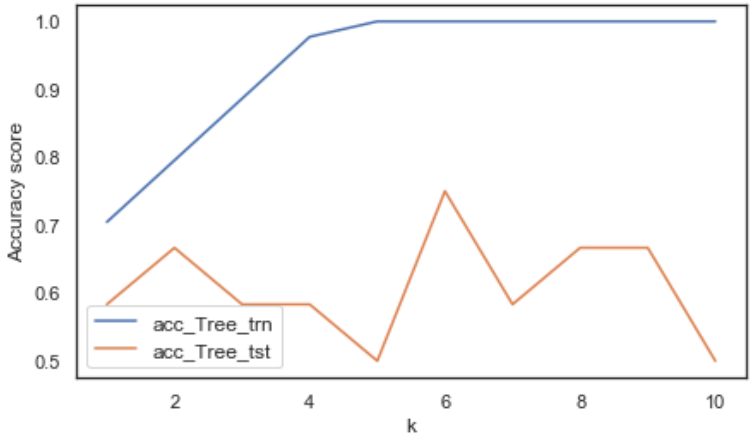
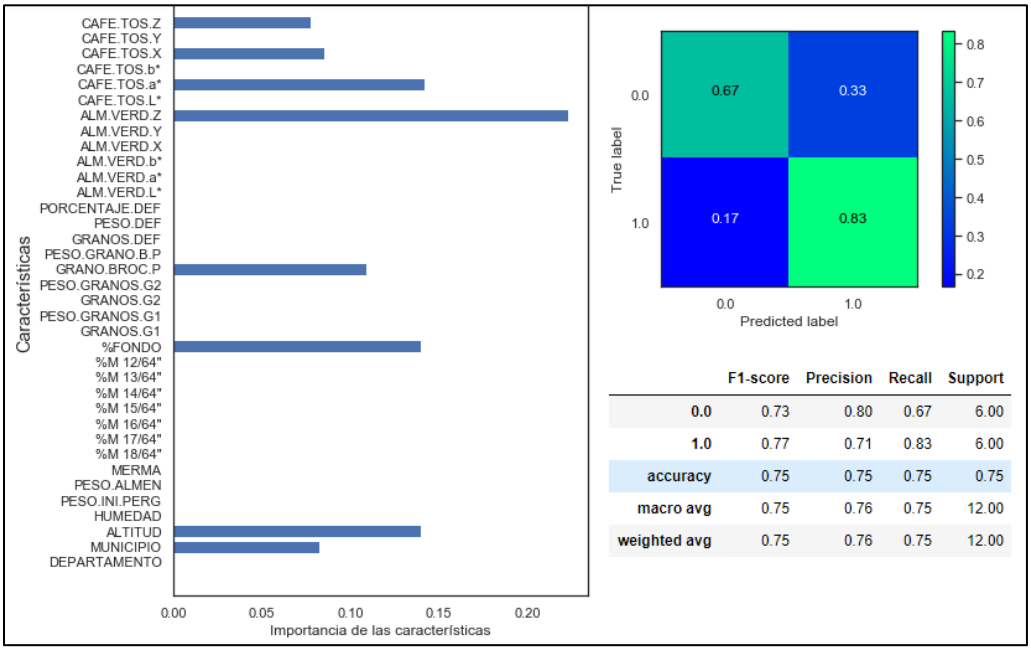


Figura 40. Resultados árbol de decisión, predicción de calidad de café global. Fuente: propia.



	F1-score	Precision	Recall	Support
0.0	0.73	0.80	0.67	6.00
1.0	0.77	0.71	0.83	6.00
accuracy	0.75	0.75	0.75	0.75
macro avg	0.75	0.76	0.75	12.00
weighted avg	0.75	0.76	0.75	12.00

Figura 41. Izquierda, importancia de las características en el resultado obtenido. Superior derecha, Matiz de confusión. Inferior derecha, reporte de clasificación. Fuente: elaboración propia.

De la figura 41, se aprecia que, para la predicción global, las variables con más importancia en la clasificación fueron el color Z de la almendra, la altitud y el color a* del café tostado. También se aprecia en la matriz de confusión que el árbol tuvo un mejor desempeño (83%) prediciendo el café con puntaje de calidad alto versus el bajo (67%). Para esto se escogió un conjunto de datos de validación aleatorio que tuviera datos balanceados de ambas categorías. La precisión más alta de validación alcanzada con este algoritmo para este set de datos fue del 75%. En la figura 42 se muestra el árbol resultante.

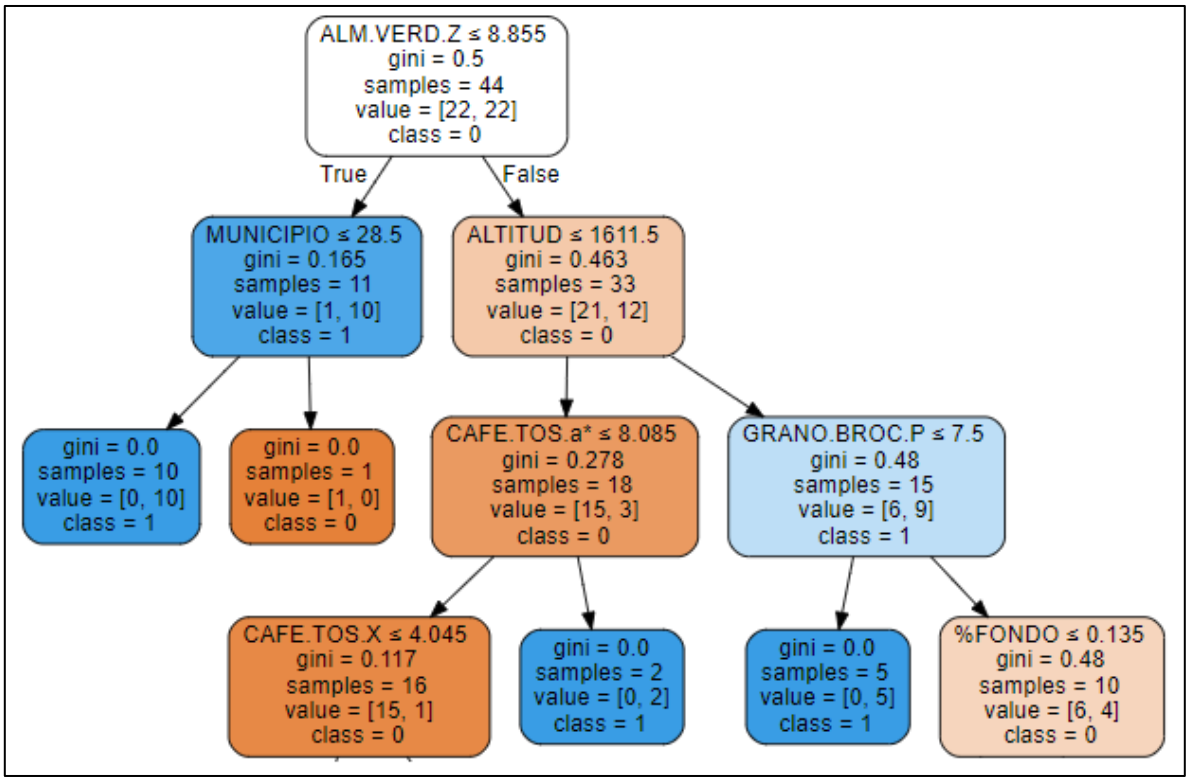


Figura 42. Árbol de decisión resultante. Se muestran solo los primeros niveles por visualización. Fuente: elaboración propia.

3.4.3 KNN, Vecinos cercarnos (clasificación).

Para este algoritmo se sigue el mismo enfoque del algoritmo anterior, se dividen los datos en entrenamiento y validación (80/20), y se entrena el algoritmo con la función KNeighborsClassifier(n_neighbors=k), en donde k, es el número de vecinos que el algoritmo

emplea para realizar la clasificación. En la figura 43, se aprecia que un mayor número de vecinos no influye positivamente en la capacidad del algoritmo para predecir este set de datos.

	k	Acc_KNN_trn	Acc_KNN_tst
0	1	1.000000	0.833333
1	2	0.750000	0.750000
2	3	0.727273	0.666667
3	4	0.681818	0.666667
4	5	0.750000	0.666667
5	6	0.659091	0.666667
6	7	0.681818	0.750000
7	8	0.636364	0.750000
8	9	0.659091	0.750000
9	10	0.636364	0.583333

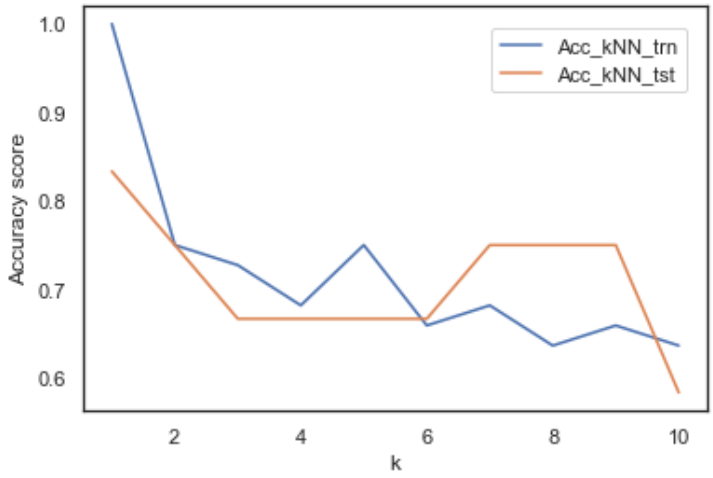
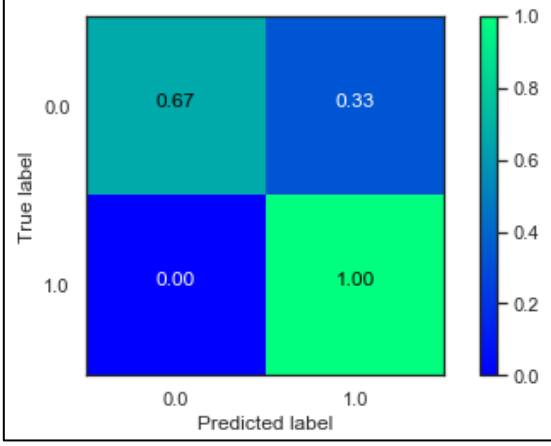


Figura 43. Resultados de entrenamiento con el algoritmo de vecinos cercanos, aumentando el número de vecinos. Acc_KNN_tst, corresponde a la precisión de la predicción en los datos de validación. Fuente: elaboración propia.



	F1-score	Precision	Recall	Support
0.0	0.80	1.00	0.67	6.00
1.0	0.86	0.75	1.00	6.00
accuracy	0.83	0.83	0.83	0.83
macro avg	0.83	0.88	0.83	12.00
weighted avg	0.83	0.88	0.83	12.00

Figura 44. Izquierda, matriz de confusión para el algoritmo KNN. Derecha, reporte de clasificación.

Respecto al reporte de clasificación, si se toma un solo vecino, se aprecia una precisión (*accuracy*) del 83% en los datos de validación. Tomando de referencia el F1-score, el algoritmo KNN, logra redecir los datos de alto puntaje de calidad en el 100% y solo logra predecir el 67% de los datos con baja calidad (figura 44). Partiendo de que en el algoritmo

anterior se presentó la misma dificultad, se podría pensar que se requiriere más detalle o quizás otra variable que tenga una relación más fuerte con los casos de baja calidad de café.

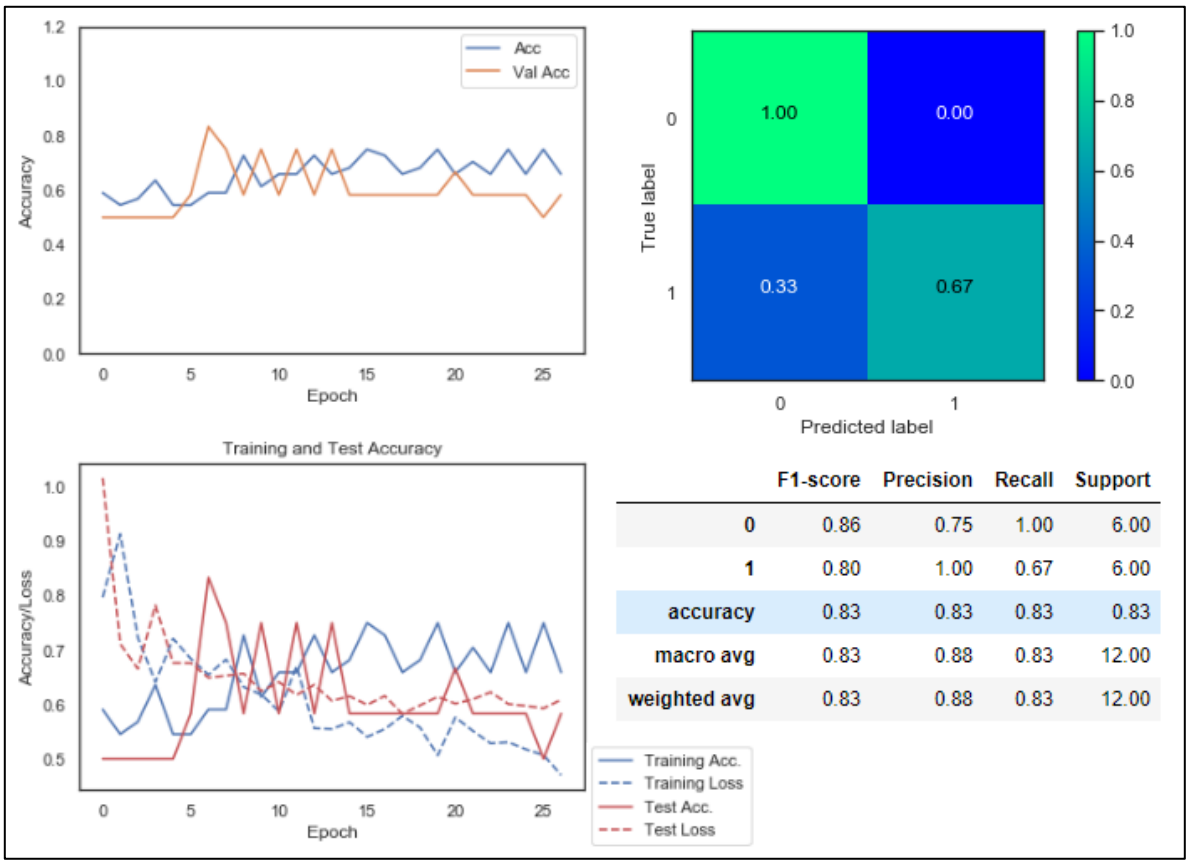
3.4.4 Red Neuronal (clasificación).

El siguiente algoritmo entrenado bajo el enfoque de clasificación de la calidad del café, es una red neuronal. La red neuronal tiene más parámetros y opciones para ajustarla. La red que se presenta a continuación es resultante de una revisión con diferentes tasas de aprendizajes, épocas, optimizadores y tamaño de nodos y capas. En la siguiente tabla (15) se resumen los parámetros de la red empleada.

Tabla 15. Parámetros de la red neuronal

PARÁMETRO	VALOR
Tamaño de los datos de entrada	36
Número de capas	3
Configuración de las capas	Capa de entrada con 36 nodos – capa densa de 18 nodos – Capa Dropout 50% - Capa densa con 2 nodos de salida y función de activación “Softmax” (Dunne & A., 1997)
Función de activación, capa densa	ReLu (Agarap, 2019)
Optimizador	Adam (Kingma & Ba, 2015)
Tasa de aprendizaje	1e-2
Reducción de la tasa de aprendizaje	50% cada 4 épocas sin mejora en la precisión de validación (val_acc)
Función de perdida	"categorical_crossentropy"
Parada temprana	Paciencia de 20 épocas según precisión de validación, máximo 100 épocas.
Tamaño del lote	1
Épocas resultantes	26

Esta red neuronal se realizó utilizando la API Keras, la cual integra a Tensorflow. En la figura 45 se aprecian los resultados de esta red. La precisión obtenida fue del 83%. Esta red fue mejor clasificando las muestras de bajo puntaje, acertando en el 100% de los casos como se muestra en la matriz de confusión. No obstante, esa mejora en la categoría 0 le implicó solo un 67% de aciertos en la clase 1



	F1-score	Precision	Recall	Support
0	0.86	0.75	1.00	6.00
1	0.80	1.00	0.67	6.00
accuracy	0.83	0.83	0.83	0.83
macro avg	0.83	0.88	0.83	12.00
weighted avg	0.83	0.88	0.83	12.00

Figura 45. Resultados Red Neuronal. Izquierda arriba, curva de precisión de entrenamiento y validación. Izquierda abajo, curvas de pérdida versus entrenamiento. Derecha arriba, matriz de confusión. Derecha abajo, reporte de clasificación. Fuente: elaboración propia.

En la siguiente sección se analiza el problema de la calidad del café como un problema de regresión, en este caso se busca predecir el valor los descriptores a partir de la regresión, no obstante, teniendo encueta las conclusiones de la sección 3.4.1. Se procede aplicarlos sobre la suma de los puntajes.

3.4.5 SVM, Support Vector Machine (regresión)

El algoritmo SVM para regresión se implementa usando la librería de sci-kit learn llamada SVR (Support Vector Regression) basada en libsvm (Chang & Lin, 2013). En la figura 46 se presentan los parámetros usados en la respectiva función.

```
clf= SVR(C=1.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.2, gamma='scale',
    kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)
clf.fit(X_trn, Y_trn)
```

Figura 46. Parámetros del algoritmo SVR empleado. Fuente: elaboración propia.

Usando la técnica de validación cruzada se probaron diferentes opciones para los parámetros gamma y Kernel, con lo cual se obtuvo que los mejores resultados se lograron con las opciones mostradas. En la figura 47, se muestran los resultados de la predicción realizada para los datos de validación con el algoritmo entrenado y se grafican para facilidad de comparación. A diferencia del enfoque de clasificación, en el enfoque de regresión las métricas cambian, para el caso de este algoritmo el puntaje por defecto es el R^2, el cual fue del -0.3338, un buen resultado es cuando este valor es muy cercano a cero. Adicional a esta métrica se incluyó el error medio absoluto, para compararlo versus la red neuronal. El error medio absoluto fue de 19.35%.

	Predicción	Real	Diferencia	% Diferencia
0	0.537193	0.516129	0.021064	4.081193
1	0.510324	0.693548	-0.183224	26.418401
2	0.490136	0.629032	-0.138897	22.081009
3	0.475654	0.612903	-0.137249	22.393334
4	0.594271	0.048387	0.545884	1128.159975
5	0.371587	0.338710	0.032877	9.706516
6	0.467970	0.354839	0.113131	31.882378
7	0.475730	0.967742	-0.492012	50.841258
8	0.681771	0.693548	-0.011778	1.698173
9	0.638280	0.709677	-0.071398	10.060590
10	0.583704	0.725806	-0.142102	19.578557
11	0.379393	0.612903	-0.233511	38.099091

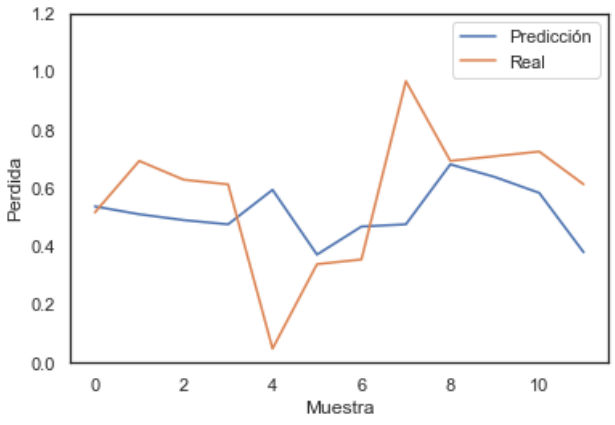


Figura 47. Predicción realizada con el algoritmo de SVM para los datos de validación versus los datos reales. Fuente: elaboración propia.

3.4.6 Red Neuronal (regresión).

Continuando con el enfoque de regresión, se propone evaluar el conjunto de datos con una red neuronal, construida para este fin. En la tabla 16, se presentan los parámetros empleados, tanto en la red, como el modelo entrenado.

Tabla 16. Parámetros de la red neuronal de regresión

PARÁMETRO	VALOR
Tamaño de los datos de entrada	36
Número de capas	6
Configuración de las capas	Capa de entrada con 36 nodos – capa densa de 36 nodos – Capa Dropout 25% - Capa densa con 36 nodos – Capa Dropout 25% - Capa densa con 9 nodos y una capa de salida densa con un nodo función de activación “Linear”
Función de activación, capas densas	ReLu
Optimizador	Adam
Tasa de aprendizaje	1e-2
Reducción de la tasa de aprendizaje	50% cada 4 épocas sin mejora en la perdida de validación (val_loss)
Función de perdida	"mean_absolute_percentage_error "
Parada temprana	Paciencia de 20 épocas según perdida de validación, máximo 100 épocas.
Tamaño del lote	4
Épocas resultantes	44

El resultado de este modelo es un error absoluto promedio del 19.22% con una desviación del 14.25%. Si bien, este no es un resultado que represente fielmente los datos de entrada, permite una aproximación al problema, con un error inferior al indicado en trabajos similares como el reportado por (Ahmed & Moustafa, 2016). En la figura 48, se aprecia la curva de

aprendizaje, en este caso reduciendo la pérdida del error medio absoluto, de acuerdo con la función de perdida definida.

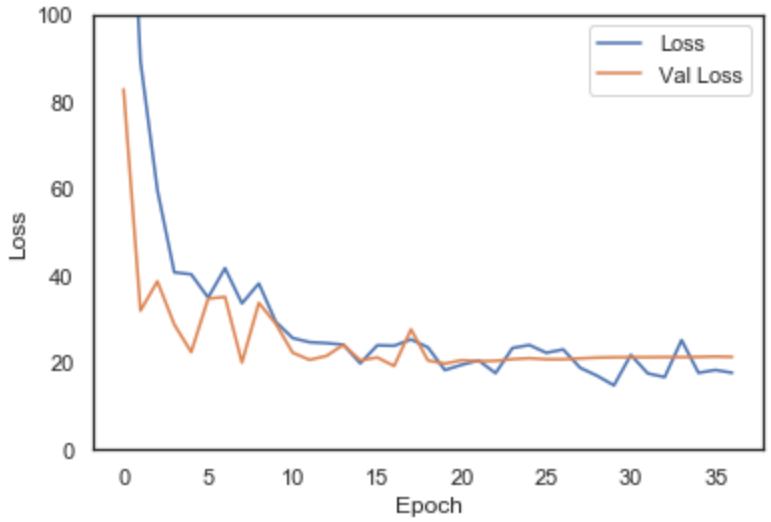


Figura 48. Gráfico de perdida de la red neuronal entrenada para regresión

En la figura 49, se muestran los resultados de las predicciones para los datos de validación y se muestra una gráfica para facilidad de comparación. En donde la línea naranja de nota los datos reales y la línea azul los datos predichos por la red.

	Predicción	Real	Diferencia	% Diferencia
0	34.23	34.00	0.23	0.68
1	36.56	34.50	2.06	5.96
2	33.12	63.50	-30.38	47.84
3	40.93	54.00	-13.07	24.20
4	29.99	49.50	-19.51	39.41
5	41.84	33.50	8.34	24.90
6	33.76	39.00	-5.24	13.43
7	41.80	51.50	-9.70	18.84
8	39.51	52.50	-12.99	24.75
9	52.62	54.00	-1.38	2.56
10	40.61	32.50	8.11	24.95
11	33.52	32.50	1.02	3.14

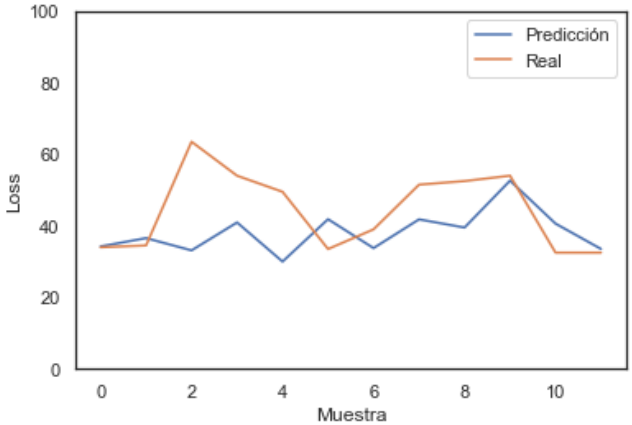


Figura 49. Resultados de predicción de la red neuronal de los datos de validación para el caso de regresión. Fuente: elaboración propia.

3.4.7 Red Neuronal (regresión múltiple).

En esta sección se trabaja el problema desde un enfoque de regresión múltiple, es decir, lograr predecir los 10 atributos de calidad, a partir de los datos de entrada. Para esto se modifica la red de la sección anterior, dejando 10 nodos en la capa de salida, uno por cada descriptor de calidad. Adicional se modifica la variable “y” que es la variable donde se cargan las etiquetas o datos de salida, para que reciba los 10 atributos por cada muestra. La división de los datos se mantiene en 80/20 y los parámetros de la red se muestran en la tabla 17.

Tabla 17. Parámetros de la red neuronal bajo el enfoque de regresión múltiple. Fuente: Elaboración propia

PARÁMETRO	**VALOR**
Tamaño de los datos de entrada	36
Número de capas	6
Configuración de las capas	Capa de entrada con 36 nodos – capa densa de 36 nodos – Capa Dropout 25% - Capa densa con 36 nodos – Capa Dropout 25% - Capa densa con 9 nodos y una capa de salida densa con 10 nodos función de activación “Linear”
Función de activación, capas densas	ReLu
Optimizador	Adam
Tasa de aprendizaje	1e-2
Reducción de la tasa de aprendizaje	Se adiciona un decay de 1e-2/200
Función de perdida	"mean_absolute_percentage_error "
Parada temprana	Ninguna
Tamaño del lote	4
Épocas resultantes	200

Una vez completado el entrenamiento se obtuvieron los resultados de la figura 50, con un error absoluto medio de 12.55% y una desviación del 5.1%. En términos generales se aprecia que la red trata de generalizar los resultados hacia la media de ellos, no obstante, se pueden apreciar diferencias en los datos individuales tanto en las variables de una misma muestra, como en las muestras.

	Aroma	Sabor	Sabor.Residual	Acidez	Cuerpo	Balance	Uniformidad	Taza.Limpia	Dulzor	Overall
0	4.169149	4.081212	3.904573	4.129179	3.863088	3.958924	4.205576	3.759331	4.053699	3.738016
1	4.186828	4.102150	3.927104	4.148110	3.883823	3.983422	4.228165	3.781557	4.075748	3.761476
2	4.153579	4.062771	3.884729	4.112505	3.844825	3.937347	4.185680	3.739755	4.034280	3.717353
3	4.166920	4.078571	3.901731	4.126791	3.860473	3.955834	4.202726	3.756527	4.050918	3.735056
4	4.092703	3.990672	3.807145	4.047317	3.773424	3.852988	4.107894	3.663220	3.958355	3.636569
5	4.077327	3.972460	3.787548	4.030851	3.755389	3.831680	4.088247	3.643888	3.939178	3.616164
6	4.157124	4.066969	3.889247	4.116301	3.848983	3.942259	4.190209	3.744212	4.038701	3.722057
7	4.253943	4.181639	4.012641	4.219980	3.962543	4.076427	4.313923	3.865937	4.159454	3.850540
8	4.227818	4.150697	3.979344	4.192003	3.931900	4.040223	4.280540	3.833091	4.126870	3.815870
9	4.124254	4.028039	3.847355	4.081102	3.810430	3.896709	4.148209	3.702887	3.997705	3.678437
10	4.116493	4.018847	3.837464	4.072792	3.801327	3.885954	4.138292	3.693129	3.988026	3.668138
11	4.249141	4.175951	4.006520	4.214838	3.956910	4.069772	4.307787	3.859900	4.153465	3.844167

Figura 50. Resultados de las predicciones del modelo de regresión múltiple. Fuente: elaboración propia.

	Aroma	Sabor	Sabor.Residual	Acidez	Cuerpo	Balance	Uniformidad	Taza.Limpia	Dulzor	Overall
0	5.5	5.0	5.0	4.5	4.5	5.0	5.0	5.0	5.0	4.0
1	5.5	5.5	5.0	5.5	5.0	5.5	5.5	5.5	5.5	5.5
2	5.5	5.0	5.5	5.0	5.0	5.0	5.5	5.0	5.5	5.0
3	5.0	5.5	5.0	5.0	5.0	5.0	5.5	5.0	5.5	5.0
4	3.5	3.5	3.0	3.5	3.0	4.0	3.5	3.0	4.0	3.0
5	5.0	4.5	4.5	4.5	4.0	4.5	4.0	4.0	4.0	4.0
6	5.0	4.5	4.0	4.5	4.5	4.0	4.5	4.0	4.5	4.0
7	6.5	6.0	6.5	5.5	6.0	6.5	7.0	6.5	5.5	6.5
8	5.0	5.5	5.0	5.5	5.0	5.5	5.5	5.5	6.0	5.5
9	5.0	5.5	5.0	5.0	5.5	5.5	5.5	5.5	6.0	6.0
10	5.5	5.5	5.5	6.0	5.0	5.5	5.5	5.5	5.5	5.5
11	5.0	5.0	5.0	5.0	6.0	5.0	5.5	5.0	5.0	5.0

Figura 51. Mediciones de calidad reales para los datos de validación. Fuente: elaboración propia.

3.4.8 Red Convolucional VGG16 (regresión múltiple).

En esta parte se emplea una red pre entrenada conocida como VGG16, tomada de (Keras, 2015), se modifican las capas de salida para emplearla como un algoritmo de regresión, en la figura 52, se muestran las imágenes de entrada para la red de fotos de granos de almendra verde y su respectiva foto por muestra después de tostados. En la figura 53, se muestra los resultados de la curva de entrenamiento. En la tabla 18 se muestran los parámetros empleados para esta red.

Tabla 18. Parámetros de la red CNN-VGG16

PARÁMETRO	VALOR
Tamaño de los datos de entrada	224 x 224 x 3
Número de capas	20
Configuración de las capas	VGG16 como en (Simonyan & Zisserman, 2015)+capa plana de 25088 nodos + capa densa de 128 nodos + capa dropout 50% y una capa de salida de 10 nodos con activación “linear”
Función de activación, capa densa	ReLu (Agarap, 2019)
Optimizador	Adam (Kingma & Ba, 2015)
Tasa de aprendizaje	1e-2
Reducción de la tasa de aprendizaje	50% cada 4 épocas sin mejora en la perdida de validación (val_loss)
Función de perdida	"mean_absolute_percentage_error"
Parada temprana	Paciencia de 15 épocas según precisión de validación, máximo 100 épocas.
Tamaño del lote	1
Épocas resultantes	39

Con esta red y el entrenamiento de imágenes se obtuvo un porcentaje de error medio absoluto de 17.56%, no es por lo pronto el mejor de los desempeños, sin embargo, se

resalta, que esta red obtuvo un mejor resultado que la red neuronal entrenada solo con datos textuales para el enfoque de regresión simple de la sección 3.4.6.

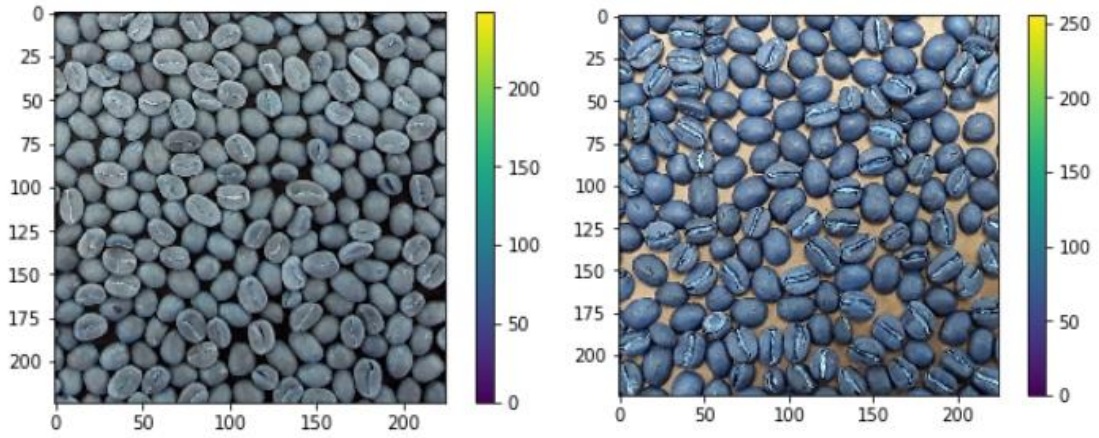


Figura 52. Imágenes convertidas a 224 x224 de café verde y café tostado, con la librería OPEN CV. Fuente: elaboración propia.

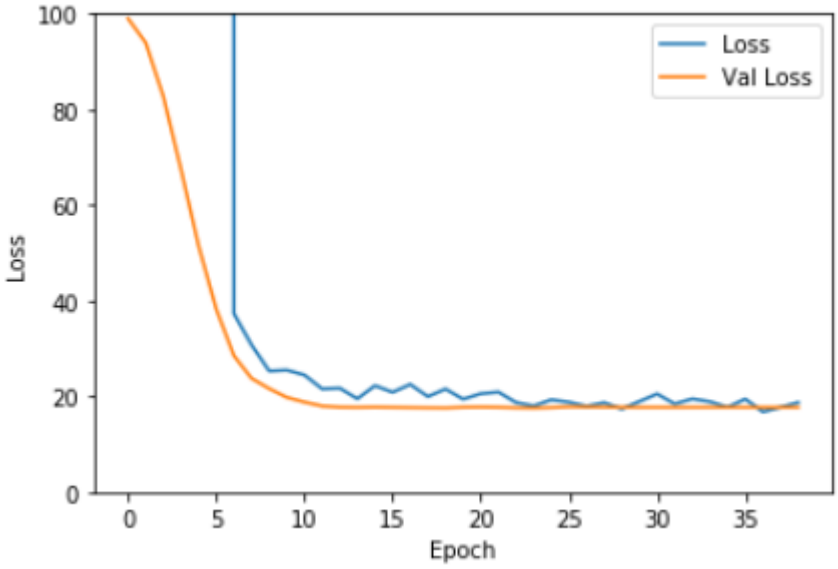


Figura 53. Curva de entrenamiento y validación de la CNN-VGG16 (Regresión múltiple). Fuente: elaboración Propia

	Aroma	Sabor	Sabor.Residual	Acidez	Cuerpo	Balance	Uniformidad	Taza.Limpia	Dulzor	Overall
0	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
1	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
2	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
3	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
4	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
5	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
6	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
7	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
8	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
9	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
10	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
11	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
12	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
13	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
14	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
15	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
16	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266
17	5.004977	4.814236	4.678197	4.847045	4.592027	4.856383	4.990659	4.692161	4.941185	4.621266

Figura 54. Fragmento de resultados de las predicciones de la CNN-VGG16 (Regresión múltiple). Fuente: Elaboración propia.

3.5 VALIDACIÓN DEL MODELO DESARROLLADO.

En esta sección se realiza la validación del modelo mediante la técnica de validación cruzada. Teniendo en cuenta que el mejor modelo en desempeño fue la red neuronal con regresión múltiple, se procede a realizar la validación sobre este modelo. La importancia del uso de esta técnica (Zhang & Yang, 2015), radica en que un modelo puede dar diferentes resultados según la separación de los datos realizados, como se mencionó en la sección 2.2.10, esta técnica permite ver la estabilidad de un modelo o la variación que puede presentarse al modificar los datos entrenamiento y validación. Para esta validación se procede a realizar una división en 10 grupos, empleando la función Kfold() de scikit-learn (Hackeling, 2014), y posteriormente, se realizan 10 variantes del modelo, usando un conjunto de validación diferente en cada corrida. Esta técnica es particularmente útil cuando se tienen pocos datos. Debido a que estamos evaluando 10 modelos diferentes, se presentan los resultados de cada modelo, así como el promedio resultante. En la figura 55 se muestra el código empleado con el resultado promedio obtenido. En la figura 56, se muestra el resultado de cada modelo.

```
#Se coloca una semilla para reproducibilidad
seed = 7
np.random.seed(seed)
estimator = KerasRegressor(build_fn=CNN5, epochs=200, batch_size=4, verbose=0)
kfold = KFold(n_splits=10, random_state=seed)
results = cross_val_score(estimator, X, y, cv=kfold)
print("Results: %.2f (%.2f) MAPE" % (results.mean(), results.std()))
```

```
Results: -14.02 (4.43) MAPE
```

Figura 55. Código para la realización de la validación cruzada de la red neuronal aplicada a regresión múltiple. Fuente: elaboración propia.

```
results
```

```
array([ -9.91677745,  -8.54442056, -18.54195182, -14.03436406,
       -16.71234576, -16.85513369, -14.37326279, -18.47187614,
        -5.06056237, -17.68474884])
```

Figura 56. Resultados de validación cruzada para cada uno de los cruces realizado.

De los resultados obtenidos en la validación cruzada, se aprecia, que el mejor modelo tuvo un porcentaje de error medio absoluto del 5.06%, posiblemente este fue un modelo donde los puntajes de las muestras seleccionadas para validación eran muy cercanos lo que le facilito al modelo realizar un mejor ajuste. En contraste, el modelo con error más alto fue el tercero con 18.54%, lo que se puede atribuir a muestras con puntajes opuestos en la escala, es decir bajos versus altos. No obstante, al realizar el promedio de los resultados, se aprecia un desempeño satisfactorio. Una nota importante es que al realizar la validación con 10 particiones (folds), estamos dejando para entrenamiento en cada modelo el 10% de los datos. Teniendo en cuenta que en el desarrollo de los modelos se usó una división de 80/20, se precede a realizar una validación con 5 particiones, mostrada en la figura 57.

```
estimator = KerasRegressor(build_fn=CNN5, epochs=200, batch_size=4, verbose=0)
kfold = KFold(n_splits=5, random_state=seed)
results = cross_val_score(estimator, X, y, cv=kfold)
print("Results: %.2f (%.2f) MAPE" % (results.mean(), results.std()))
```

```
Results: -14.61 (3.71) MAPE
```

```
results
```

```
array([ -9.51163006, -17.44620031, -12.55009174, -20.02535508,
       -13.52416767])
```

Figura 57. Código y resultado de validación cruzada con 5 particiones. Fuente: elaboración propia.

Aplicando solo 5 particiones, es decir dejando un 20% de los datos para validación, se aprecia que el promedio aumenta (14.61) y la desviación disminuye (3.71), este resultado da una mejor representación de la capacidad del modelo.

Nota: la precisión en inglés es “*accuracy*”, la palabra “*precision*”, en inglés es para denotar los valores positivos predictivos (*Positive Predictive Values*).

4. CAPITULO 4: RESULTADOS

4.1 RESUMEN DE RESULTADOS DE CADA ALGORITMO.

En las siguientes tablas se muestran los resultados de cada algoritmo empleado con sus respectivas métricas. En la tabla 19 se aprecian los resultados de los algoritmos de clasificación empleados. Los datos disponibles no permitieron realizar una separación de las variables en más de 2 categorías. Al intentar realizarlo la predicción no superaba el 50%, por ello se optó por crear una etiqueta en la que se pudiera clasificar el café como alto o bajo puntaje, según los datos disponibles. De igual manera en esta misma tabla (19) también se puede apreciar que tanto el algoritmo de vecinos cercanos con un vecino, así como la red neuronal, lograron una precisión del 83%, con la diferencia que para el algoritmo de vecinos cercanos el mayor éxito de predicción estuvo en las muestras de puntaje alto, mientras que para la red estuvo en las muestras de puntaje bajo.

Tabla 19. Resumen de resultados de algoritmos de clasificación empleados. Fuente: elaboración propia.

NOMBRE ALGORTIMO	ENFOQUE	PARÁMETROS	PRECISIÓN (ACCURACY)	F1 (clase 0)	F1 (clase 1)
Árbol de decisión	Clasificación	K = 6	0.75	0.67	0.83
Vecinos cercanos (KNN)	Clasificación	K=1	0.83	0.67	1.0
Red Neuronal	Clasificación	Ver tabla sección 3.4.4	0.83	1.0	0.67

Este tipo de enfoque es de interés para la oficina de calidad de Café, porque permite evidenciar, por un lado, que estas herramientas logran identificar patrones, a partir de las mediciones realizadas, y les permite una primera discriminación de las muestras. Adicionalmente, se aprecia una oportunidad para aumentar esta base de datos y lograr

tener una herramienta de verificación y porque no en un futuro emplearlo como calificador de línea.

Respecto al enfoque de regresión, y de acuerdo con lo mencionado en la sección 2.2.11, las métricas empleadas para este enfoque son diferentes de la de clasificación y se escoge el porcentaje de error medio absoluto (MAPE), porque permite entender de una manera sencilla la desviación de las predicciones del modelo versus los datos reales. En los primeros 2 modelos se empleó el enfoque de regresión para predecir el puntaje global de los descriptores de café, es decir un solo resultado. En el tercer modelo se emplea un modelo de regresión múltiple, siendo este modelo que mejor desempeño presentó. Aunque el resultado es bajo para este tipo de enfoque, igual que en el caso de clasificación, se ve la oportunidad de mejorar este desempeño con la inclusión de más datos.

Tabla 20. Resumen de los resultados de algoritmos de regresión empleados.

NOMBRE ALGORTIMO	ENFOQUE	PARÁMETROS	MEDIA % ERROR (MAPE)	DESV % ERROR (MAPE)
SVM	Regresión 1 variable	Gamma = “scale”, Kernel = “rbf”	19.35	12.3
Red neuronal	Regresión 1 variable	Ver tabla sección 3.4.6	19.22	14.25
Red neuronal	Regresión múltiple	Ver tabla sección 3.4.7	14.61	3.71
CNN-VGG16	Regresión múltiple	Ver tabla sección 3.4.8	17.56	9.5

Revisando los resultados de regresión con el jefe de la oficina de calidad de Almacafé, ellos consideran que un error de 0.5 unidades por encima o por debajo en la predicción es aceptable para estos descriptores, no obstante, un puntaje de 3.5, reportado como 4 puede, implicar autorizar el uso de dicho café para mercados donde podría ser rechazado, por ello, igual se ve la oportunidad de colectar aún más datos y de apoyar el modelo con algunas reglas adicionales para descartar valores dudosos, ejemplo un puntaje de 3.8, sería preferible reportarlo como 3.5 y no como 4.

4.2 DISCUSIÓN DE RESULTADOS.

- (de Oliveira, Leme, Barbosa, & Rodarte, 2015) midieron las coordenadas del espacio CIELAB en granos de café verde para clasificarlos con la ayuda de una red neuronal. De manera similar, en este trabajo se empleó un espectrofotómetro para obtener las coordenadas de color CIELAB, y se usaron estas mediciones como entrada para los algoritmos entrenados.
- (Ferreira, Pereira, Delbem, Oliveira, & Mattoso, 2007), utilizaron una lengua electrónica (ET) para la predicción de la calidad del café, entrenaron una red neuronal con las mediciones obtenidas con la ET y los puntajes de calidad asignados por los catadores profesionales, para clasificar en el grupos definidos En este trabajo también se utilizaron los puntajes de calidad asignados por los catadores profesionales, con la diferencia de que las variables o medidas de entrada no se realizaron en la bebida de café, sino en los granos, que se asemejan al filtro y la evaluación de calidad realizada por Almacafé.
- (Ruge, Pinzon, & Moreno, 2012) desarrollaron una máquina capaz de seleccionar granos de café defectuosos de los buenos, a partir de la implementación del método de segmentación de umbral multinivel, aplicado a imágenes de granos buenos y defectuosos, que permitieron determinar los rangos de cada color del espacio RGB para las categorías definidas. Por el contrario, en este trabajo se utilizó el espacio CIELAB, donde los valores de cada componente del espacio de color, junto con las otras variables, contribuyen con un peso a cada resultado de catación obtenido.

4.3 LÍMITACIONES.

Para este trabajo se realizaron mediciones únicamente a muestras de café de regiones específicas, recolectadas durante el 2019, permitiendo realizar una definición y evaluación inicial de modelo. No obstante, su desempeño con muestras recolectadas en diferentes condiciones puede ser bajo, caso en el cual se recomienda un reentrenamiento. Por otra parte, los valores de las variables de salida están dados por el promedio de la calificación de diferentes expertos, que aunque generalmente tienen una baja variación, esto genera un ruido en la variable de salida.

CONCLUSIONES

- Esta investigación muestra que la calidad de la taza de café puede predecirse como alta o baja a partir del análisis de las características de los granos de café verde.
- El modelo de red neuronal definido es una alternativa potencial a la calificación experta de café en taza, reduciendo el costo de la evaluación, ahorrando tiempo en análisis y preparación de café en taza y proporcionando una herramienta adicional para catadores de café y productores de café.
- Los resultados del modelo estuvieron limitados por la cantidad de datos disponibles, no obstante, el modelo permite realizar una precalificación de los granos de café, lo que permitiría un mejor aprovechamiento de recursos valiosos, al dejar por fuera muestras con baja calidad.
- La precisión en la predicción fue del 83% para el enfoque de clasificación, tanto con el algoritmo de vecinos cercanos, como con la red neuronal artificial.
- Bajo el enfoque de regresión, el menor error entre las predicciones y los datos reales se obtuvo con una red neuronal, con un error absoluto medio de 14,61%
- Las variables de entrada, relacionadas con el color del grano en almendra verde y grano tostado, tienen un peso importante siendo fundamentales para la generalización del modelo y lograr mejores resultados.
- La devolución de llamada de detención temprana fue determinante para detener el entrenamiento de la red neuronal cuando el modelo no estaba mejorando el puntaje de precisión de validación.
- El aumento en la precisión de los resultados obtenidos, demanda de un mayor número de datos de entrenamiento, así como de variedad en los rangos de las variables.
- El aumento de la precisión para la predicción de algunos descriptores de calidad como cuerpo, balance, uniformidad podrían requerir de datos de entrada adicionales.
- El conjunto de datos permitió realizar una definición de la calidad generalizada (alta/baja calidad) respecto a las categorías definidas y la forma de agrupación de las variables de salida.

- Se sugiere la implementación de un piloto del modelo en Almacafé, para irlo ajustando y mejorando conforme se puedan ingresar más muestras a este.
- El enfoque de regresión múltiple con una red neuronal permitió obtener unos resultados muy aproximados a los reales con un error de validación de 14.61% lo que está dentro del rango de tolerancia de la catación realizada (+/-0.5).
- La división de los datos o separación no es tan sencilla, debido a que pueden estar diferenciados en algunas variables, pero en otras pueden tener las mismas puntuaciones.
- De los algoritmos trabajados, se identifica que el árbol de decisión no solo permite realizar predicciones, sino también constituye una herramienta de análisis de datos.
- Para el caso del algoritmo SVR, los datos deben estar normalizados.
- Respecto a la normalización de los datos para el entrenamiento de redes neuronales aplicado a regresión múltiple, no se encontró una diferencia significativa entre el uso de los datos normalizados y sin normalizar.
- Las variables de entrada, relacionadas con el color del grano en almendra verde y grano tostado, tienen un peso importante como se describió en la sección 3.4.1., siendo fundamentales para la generalización del modelo y lograr mejores resultados.

PRINCIPALES APORTES

- Un modelo entrenado para la determinación de la calidad del café líquido a partir de mediciones realizados a los granos de café verde, de acuerdo con la plataforma de 10 atributos empleada.
- Una tesis de investigación, con información de implementación, recomendaciones y pasos tomados para su realización. Incluyendo parámetros de los modelos.
- Un código guía de los análisis, algoritmos y librerías usadas para futuras implementaciones.
- Un artículo científico (en proceso), de aplicación de herramientas computacionales en aplicaciones no tradicionales, específicamente, la determinación de la calidad del café a partir de variables textuales y de imágenes.
- La apertura de una nueva tecnología al sector del café, iniciando por la oficina de calidad de la Federación Nacional de Cafeteros.

RECOMENDACIONES Y FUTUROS TRABAJOS

El trabajo realizado, constituye una primera aproximación a la determinación de la calidad del café mediante herramientas computacionales y establece un modelo para implementar en aplicaciones donde la calidad se determine mediante operaciones de catación o panel sensorial, realizado por humanos. A continuación, se relacionan algunas recomendaciones para trabajos o enfoques similares, así como algunas propuestas de trabajos futuros tanto en el campo de aplicación específico como en ámbitos más generales

- En el caso de la aplicación específica del café, el avance de las tecnologías permite automatizar varias tareas, a lo largo de su proceso productivo, a continuación, se nombran algunas de ellas.
    - Selección automática de frutos, mediante sistemas de reconocimientos de objetos, capaces de separarlos por su grado de maduración.
    - Selección automática de defectos en granos de café, para separar el café de alta calidad o excelso del café estándar o pasilla.
    - Clasificación de calidad de café a partir de imágenes. Aunque este punto se tocó en este trabajo, se precia un potencial enorme por los resultados obtenidos de seguir profundizando en esta dirección.
    - Uso de narices electrónicas y redes neuronales para la determinación de la calidad del café.
- El método de validación cruzada no solo permite realizar la validación de un modelo, también es muy útil para realizar ajuste fino de parámetros de los algoritmos empleados.
- Para el caso de entrenamiento con imágenes, existen redes ya pre entrenadas como la VGG16 o la inceptionV3 (Keras, 2015), con una gran de capacidad de reconocimiento de características en muchas aplicaciones, por lo que se recomienda usar estas redes pre entrenadas para problemas a partir de imágenes.
- Una tendencia general, como las mencionadas en el marco teóricos son las de clasificación y evaluación de diferentes frutos, mediante imágenes, dado que cada fruto es tan especifico y en Colombia tenemos variedad de cultivos, muchas de estas técnicas se pueden implementar y adaptar a los cultivos o frutos de interés.

- Uso de aplicaciones móviles para la detección del estado de madurez de frutas y vegetales. Una vez se realiza la implementación de Machine Learning y se obtinen un modelo entrenado, se pueden usar los pesos del modelo para la predicción de las variables a las que fue entrenado. Esto permite, por ejemplo, la creación de aplicaciones, que, desde el celular, les permitan a los consumidores, saber el estado de una fruta o un vegetal, una herramienta importante para consumidores o compradores no expertos.

Referencias

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (9 de Noviembre de 2015). *TensorFlow: large-scale machine learningon heterogeneous distributed systems.* Obtenido de https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf

Agarap, A. F. (07 de Feb de 2019). *Deep Learning using Rectified Linear Units (ReLU).* Obtenido de Cornell University: arXiv:1803.08375v2

Aguilar, R., Torres, J., & Martín, C. A. (2019). Aprendizaje automático en la identificación de sistemas. Un caso de estudio en la predicción de la generación eléctrica de un parque eólico. *Revista Iberoamericana de Automática e Información Industrial, 15*, 114-127. Obtenido de https://www.researchgate.net/publication/327721029_Aprendizaje_Automatico_en_la_Identificacion_de_Sistemas_Un_caso_de_estudio_en_la_generacion_de_un_parque_eolico

Ahmed, E. H., & Moustafa, M. N. (2016). House price estimation from visual and textual features. *Proceedings of the 8th International Joint Conference on Computational Intelligence. 3*, págs. 62-68. SciTePress. Obtenido de Computer Science and Engineering Department, The American University in Cairo, Road 90, New Cairo, Cairo, Egypt: https://arxiv.org/pdf/1609.08399.pdf

Almacafé, O. d. (1 de Sep de 2019). Analisis de muestra de café de diferentes regiones. Bogotá.

ASOEXPORT, A. N. (19 de 08 de 2019). *Cultivo de café*. Obtenido de ASOEXPORT: https://asoexport.org/

Banco Caja Social S.A. (2017). *Oportunidades y riegos: café.* Obtenido de Informes sectoriales: https://www.bancocajasocial.com/sites/default/files/page/file/cafe_-_primer_semestre_de_2017.pdf

Banco Mundial. (2002). *Estudio del sector cafetero en Colombia.* Obtenido de Café de Colombia: http://www.cafedecolombia.com/docs/ensayos182002/resumenejecutivobancomundial.pdf

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

Bonaccorso, G. (2017). *Machine learning algorithms.* Birmingham: Packt Publishing.

Botero, M. (04 de 04 de 2019). *Café: año tras año se repite la misma película*. Obtenido de Revista Dinero: https://www.dinero.com/opinion/columnistas/articulo/cafe-ano-tras-ano-se-repite-la-misma-pelicula-por-mauricio-botero/269124

Buencafé. (04 de 04 de 2019). *Buencafé liofilizado de Colombia*. Obtenido de https://www.buencafe.com/cafe-de-colombia/

Carvajal, J. J., Aristizábal, I. D., Oliveros, C. E., & Mejía, J. W. (2006). Colorimetría del fruto de café (Coffea arabica L.) durante su desarrollo y maduración. *International Journal of Developmental and Educational Psychology*(1), 37-48. Obtenido de http://www.redalyc.org/articulo.oa?id=349832311003

Chang, C.-C., & Lin, C.-J. (4 de Marzo de 2013). *LIBSVM: A library for support vector machines.* Obtenido de https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf

Chazallet, S. (2016). *Python 3. Los fundamentos del lenguaje.* Barcelona, España: Ediciones ENI.

Chin, L., & Dutta, T. (2016). *NumPy essentials.* Packt Publishing Ltd.

Coffe IQ. (15 de 03 de 2019). *Café arabicá, características*. Obtenido de http://www.coffeeiq.co/cafe-arabica-caracteristicas/

Coffee Quality Institute. (15 de 03 de 2019). *Coffee Quality Institute (CQI) database*. Obtenido de https://database.coffeeinstitute.org/

Cortez, P., Cerderia, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*, 547-553.

Cuadras, C. (1989). Distancias estadísticas. *Estadistica Española, 30*(119), 295-378. Obtenido de http://halweb.uc3m.es/esp/Personal/personas/dpena/publications/castellano/1989EE_cuadras_coment.pdf

Cubero, S., Aleixos, N., Moltó, E., Gómez-Sanchis, J., & Blasco, J. (2011). Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food and Bioprocess Technology*, 487-504. doi:10.1007/s11947-010-0411-8

Dahikar, S., & Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. *International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering*, 683-686.

DANE. (29 de Marzo de 2019). *DANE*. Obtenido de https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-internacional/exportaciones

De Oliveira, E. M., Leme, D. S., Barbosa, B. H., & Rodarte, M. P. (2015). A computer vision system for coffee beans classification based on computational intelligence techniques. *Journal of Food Engineering, 171*, 22-27. doi:10.1016/j.jfoodeng.2015.10.009

Dittakan, K., Theera-Ampornpunt, N., & Boodliam, P. (2018). Non-destructive grading of pattavia pineapple using texture analysis. *The 21st International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 144-149.

Drakos, G. (26 de Agosto de 2018). *How to select the right evaluation metric for machine learning models: part 1 regression metrics*. Obtenido de Towards Data Science: https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

Dunne, R. A., & A., C. N. (1997). *On The Pairing Of The Softmax Activation And Cross-Entropy Penalty Functions And The Derivation Of The Softmax Activation Function.* Obtenido de Proccedings of the 8th Australian Conference on Neural Networks, Melbourne: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.6403&rep=rep1&type=pdf

Espinal, C., Martínez, H., & Acevedo, X. (03 de 2005). *La cadena del café en Colombia. Una mirada global de su estructura y dinámica 1991-2005.* Bogotá: Ministerio de agricultura y desarrollo rural. Obtenido de Observatorio Agrocadenas Colombia: http://bibliotecadigital.agronet.gov.co/bitstream/11348/6111/1/200511215113_caracterizacion_cafe.pdf

Faridah, F., Parikesit, G. O., & Ferdiansjah, F. (2015 de 2011). Coffee bean grade determination based on image parameter. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 9*(3), 547-554. doi:10.12928/telkomnika.v9i3.747

Federación Nacional de Cafeteros. (20 de 03 de 2019). *El café de Colombia*. Obtenido de https://www.federaciondecafeteros.org/particulares/es/nuestro_cafe/el_cafe_de_colombia/

Ferguson, M., Ak, R., Lee, Y.-T. T., & Law, K. (2017). Automatic localization of casting defects with convolutional neural networks. *2017 IEEE International Conference on Big Data (Big Data)* (págs. 1726-1735). Boston : IEEE.

Ferreira, E., Pereira, R., Delbem, A., Oliveira, O., & Mattoso, L. (2007). Random subspace method for analysing coffee with electronic tongue. *Issue 21, 43*, 1138-1140.

Figueira, R., De Alcântara, E., Kampel, M., Stech, J., Leão de Moraes, E., & Garcia, L. (2009). O algoritmo Support Vector Machines (SVM): avaliação da separação ótima de classes em imagens CCD-CBERS-2. *Anais XIV simpósio brasileiro de sensoriamiento remoto*, (págs. 2079-2089). Natal.

Gala, Y. (25 de Septiembre de 2013). *Algoritmos SVM para problemas sobre big data.* Obtenido de https://repositorio.uam.es/bitstream/handle/10486/14108/66152_Yvonne_Gala_Garcia.pdf?sequence=1

Gallo, C. (2015). Artificial Neural Networks Tutorial . En M. Khosrow-Pour, *Encyclopedia of Information Science and Technology, Third Edition* (págs. 179-189). IGI Global.

Gomez, E., Paéz, C., Buitrago, L., & Ceballos, O. (2014). *Cadena productiva del café: oportunidades de inclusión productiva en el Quindío.* Armenia: Programa de las Naciones Unidas para el Desarrollo.

González, L. (10 de Agosto de 2018). *Todo sobre aprendizaje no supervisado en machine learning*. Obtenido de Ligdi González. Aprende todo sobre inteligencia artificial: http://ligdigonzalez.com/todo-sobre-aprendizaje-no-supervisado-en-machine-learning/

Google Colab. (2019). *Te damos la bienvenida a Colaboratory*. Obtenido de https://colab.research.google.com/notebooks/welcome.ipynb#scrollTo=xitplqMNk_Hc

Hackeling, G. (2014). *Mastering machine learning with scikit-learn.* Birmingham, UK: Packt Publishing Ltd.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning data mining, inference, and prediction (12th printing).* Springer. doi:10.1007/978-0-387-84858-7

Hinojosa Gutiérrez, Á. P. (2016). *Python paso a paso .* Madrid, España: RA-MA.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science \& Engineering, 9*(3), 90 - 95. Obtenido de https://matplotlib.org/

Idris, I. (2011). *NumPy 1. 5 beginner's guide.* Packt Publishing Ltd.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with aplications in R.* Nueva York: Springer.

James, L. (15 de 01 de 2019). *GitHub*. Obtenido de https://github.com/jldbc/coffee-quality-database

Jayalakshmi, T., & Santhakumaran, A. (Febrero de 2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering, 3*(1), 1793-8201. Obtenido de https://pdfs.semanticscholar.org/4cbd/5fcd6081cfcd16e9111f1bcc17b6e283d439.pdf

Joy, P., & Rashida, T. (2016). *Harvesting and post-harvest handling of pineapple.* Obtenido de ResearchGate: https://www.researchgate.net/publication/308171377_Harvesting_and_post-harvest_handling_of_pineapple

Jupyter. (22 de Agosto de 2019). *Jupyter*. Obtenido de https://jupyter.org/index.html

Kaewapichai, W., Kaewtrakulpong, P., & Prateepasen, A. (15 de Octubre de 2006). A real-time automatic inspection system for pattavia pineapples. *Key engineering materials, 321-323*, 1186-1191. doi:10.4028/www.scientific.net/KEM.321-323.1186

Keras. (2015). *Keras: la biblioteca de Python deep learning*. Obtenido de https://keras.io/

Kingma, D., & Ba, J. (2015). *Adam: A method for stochastic optimization.* Obtenido de arXiv:1412.6980v9 [cs.LG] 30 Jan 2017

Kouadio, L., Deo, R., Byrareddy, V., Adamowski, J., Mushtaq, S., & Nguyen, V. (2018). Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Computers and Electronics in Agriculture*(115), 324-338. Obtenido de https://doi.org/10.1016/j.compag.2018.10.014

Krizhevsky, A. (2014). *One weird trick for parallelizing convolutional neural networks.* Obtenido de CoRR, abs/1404.5997.

Kushnir, R. (12 de Febrero de 2016). Coffee-drinking in constructing finnish-americanness in three finnish-american migrant short-story collections. Vaasa, Finlandia.

Läderach, P. R. (26 de Noviembre de 2007). *Management of intrinsic quality characteristics for high-value specialty coffees of heterogeneous hillside landscapes.* Obtenido de Universität Bonn: http://hss.ulb.uni-bonn.de/2007/1290/1290.pdf

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning Aapplied to document recognition. *Proceedings of the IEEE*.

MathWorks. (s.f.). *Algortimos de Machine Learning para clasificación (SVM)*. Recuperado el 15 de Septiembre de 2019, de https://la.mathworks.com/discovery/support-vector-machine.html

Matich, D. J. (Marzo de 2001). *Redes neuronales: conceptos básicos y aplicaciones.* Obtenido de Universidad Tecnológica Nacional – Facultad Regional Rosario: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

Mazurowski, M., Buda, M., Saha, A., & Bashir, M. (2018). Deep learning in radiology: an overview of the concepts and a survey of the state of the art. *Cornell University. ArXiv*. Obtenido de https://arxiv.org/abs/1802.08717

Montes, N. (13 de 06 de 2003). *Segmentación de imágenes del café en el proceso de beneficio.* Obtenido de bdigital.unal.edu.co: http://www.bdigital.unal.edu.co/1219/

Ng, A. (16 de Octubre de 2018). *Lectures from the Machine Learning course.* Obtenido de COURSERA: https://www.coursera.org/learn/machine-learning/resources/JXWWS

Nilsson, N. J. (1998). *Introduction to machine learning.* Stanford University.

Palma Méndez, J., & Marín Morales, R. (2008). *Inteligencia artificial. Métodos, técnicas y aplicaciones.* Madrid: McGRAW-HILL/Interamericada de España, S. A. U.

Pandas. (2019). *Biblioteca de análisis de datos de Python*. Obtenido de https://pandas.pydata.org/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit - aprender: machine learning en Python. *Diario de Machine Learning Research*, 2825 - 2830.

Powers, D. M. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.* School of Informatics and Engineering. Adelaide • Australia: Flinders University.

Ramos, P., Sanz, J., & Oliveros, C. (2010). Identificación y clasificación de frutos de café en tiempo real, a través de la medición de color. *CENICAFE*, 61(4):315-326.

Ruge, I. A., Pinzon, A. S., & Moreno, D. E. (2012). Sistema de selección electrónico de café excelso basado en el color mediante procesamiento de imagenes. *Tecnura, 16*(34), 84-93. doi:http://dx.doi.org/10.14483/udistrital.jour.tecnura.2012.4.a06

Russell, S., & Norvig, P. (2004). *Inteligencia artificial. Un enfoque moderno.* Madrid: Pearson Educación, S.A.

Salamanca, C. (2015). *Métodos estadísticos para evaluar la calidad del café.* Tesis Doctoral, Universitat de Girona. Obtenido de http://www.tdx.cat/bitstream/handle/10803/327037/tcasr1de1.pdf?sequence=2

Sánchez, N. (2015). Máquinas de soporte vectorial y redes neuronales atificiales en la predicción del moviemiento USD/COP spot intradiario. *ODEON*, 113-172. Obtenido de https://revistas.uexternado.edu.co/index.php/odeon/article/view/4414/5004

Sandoval, Z., & Prieto, F. (2007). Caracterización de café cereza empleando técnicas de visión artificial. *Revista Facacultad Nacional de Agronomía Medellín, 60*(2), 4105-4127. Obtenido de https://revistas.unal.edu.co/index.php/refame/article/view/24461

Seaborn. (2012). *Seaborn: visualización de datos estadísticos*. Obtenido de https://seaborn.pydata.org/

Shah, I. (29 de Enero de 2019). *What is cross validation in machine learning?* Obtenido de Quora: https://www.quora.com/What-is-cross-validation-in-machine-learning

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: from theory to algorithms.* Cambridge University Press. doi:10.1017/CBO9781107298019

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image rocgnition.* Obtenido de The International Conference on Learning Representations (ICLR): arXiv:1409.1556v6

Singla, A., Yuan, L., & Ebrahimi, T. (16 de Octubre de 2016). *Food/non-food image classification and food categorization using pre-trained GoogLeNet model.* (ACM, Ed.) doi:http://dx.doi.org/10.1145/2986035.2986039

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (23 de Agosto de 2016). *Inception-V4, Inception-ResNet and the impact of residual connections on learning.* Obtenido de https://arxiv.org/pdf/1602.07261.pdf

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. (11 de Dec de 2015). *Rethinking the inception architecture for computer vision.* Obtenido de arXiv:1512.00567v3 [cs.CV] 11 Dec 2015

Tobijaszewska, B., Mills, R., & Jøns, J. (Junio de 2018). *Foss analytics.* Obtenido de https://www.fossanalytics.com/-/media/files/documents/papers/meat-segment/using-spectrometry-for-simultaneous-measurement_es.pdf

Torres, J. (2018). *Deep learning. Introducción práctica con Keras. Primera parte.* Barcelona: Lulu Press, Inc.

Tran, T. X. (2019). *Python pandas: the complete reference: guide to Python pandas.* Thanh Tran.

UNLU, Universidad Nacional de Luján . (12 de Noviembre de 2015). Criteios de selección de modelos. Bases de datos masivas. *Diapositvas clase magistral*. Luján, Argentina.

USDA, United States Department of Agriculture Foreign Agricultural Service. (2016). *Coffee: world markets and trade - 2016/17 Forecast overview.* Obtenido de https://apps.fas.usda.gov/psdonline/circulars/coffee.pdf

Wang, H., & Raj, B. (2015). *A survey: time travel in deep learning space: an introduction to deep learning models and how deep learning models evolved from the initial ideas.* Obtenido de Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213: https://arxiv.org/pdf/1510.04781.pdf

Wang, S., Zhang, Y., Ji, G., Yang, J., Wu, J., & Wei, L. (2015). Fruit classification by wavelet-entropy and feedforward neural network trained by fitness-scaled chaotic ABC and biogeography-based optimization. *Entropy, 17*(8), 5711-5728.

Wikipedia, C. d. (07 de 01 de 2019). *Wikipedia*. Obtenido de https://es.wikipedia.org/wiki/Coffea_arabica

Zhang, Y., & Wu, L. (2012). Classification of fruits using computer vision and a multiclass support vector machine. *Sensors, 12*(9), 12489-12505. Obtenido de https://www.mdpi.com/1424-8220/12/9/12489

Zhang, Y., & Yang, Y. (2015). *Cross-Validation for Selecting a Model Selection Procedure.* Obtenido de University of Minnesota: http://users.stat.umn.edu/~yangx374/papers/ACV_v30.pdf

Anexos

Anexo 1: Librerías compatibles con Python instaladas.

Package	Version
absl-py	0.7.1
astor	0.8.0
backcall	0.1.0
colorama	0.4.1
cycler	0.10.0
decorator	4.4.0
gast	0.2.2
google-pasta	0.1.7
graphviz	0.11
grpcio	1.21.1
h5py	2.9.0
ipykernel	5.1.1
ipython	7.6.0
ipython-genutils	0.2.0
jedi	0.14.0
Jinja2	2.10.1
joblib	0.13.2
jupyter-client	5.2.4
jupyter-core	4.5.0
Keras	2.2.4
Keras-Applications	1.0.8
Keras-Preprocessing	1.1.0
kiwisolver	1.1.0
Markdown	3.1.1
MarkupSafe	1.1.1
matplotlib	3.1.0
numpy	1.16.4
object-detection	0.1
opencv-contrib-python	4.1.0.25
pandas	0.24.2
parso	0.5.0

Package	Version
pickleshare	0.7.5
Pillow	6.0.0
pip	19.1.1
prompt-toolkit	2.0.9
protobuf	3.8.0
pydot	1.4.1
Pygments	2.4.2
pyparsing	2.4.0
python-dateutil	2.8.0
pytz	2019.1
PyYAML	5.1.1
pyzmq	18.0.2
scikit-learn	0.21.2
scipy	1.3.0
seaborn	0.9.0
setuptools	41.0.1
Shapely	1.6.4.post2
six	1.12.0
slim	0.1
tensorboard	1.14.0
tensorflow	1.14.0
tensorflow-estimator	1.14.0
termcolor	1.1.0
tornado	6.0.3
traitlets	4.3.2
wcwidth	0.1.7
Werkzeug	0.15.4
wheel	0.33.4
wrapt	1.11.2
xlrd	1.2.0

Anexo 2: Fuentes, imagen del proceso productivo del café verde (Figura 3)

- Abagensa. (2016). Bolsas para almácigos. Obtenido de http://www.abagensa.com/
- Castro Toro, Á. M., Rivillas Osorio, C. A., Serna Giraldo, C. A., & Mejía Mejía, C. G. (febrero de 2008). Avances Técnicos Cenicafé - 368. Germinadores de Café. Construcción, manejo de Rhizoctonia solani y costos. Obtenido de https://www.cenicafe.org/es/publications/avt0368.pdf
- Cenicafé. (2015). Cultivemos café / Almácigo. Obtenido de https://www.cenicafe.org/es/index.php/cultivemos_cafe/almacigo
- Cenicafé. (15 de diciembre de 2015). Cultivemos Café/ Germinadores. Obtenido de https://www.cenicafe.org/es/index.php/cultivemos_cafe/germinador
- Cenicafé. (30 de marzo de 2016). Cultivemos café. Obtenido de Beneficio: https://www.cenicafe.org/es/index.php/cultivemos_cafe/beneficio
- Contreras, D. (18 de agosto de 2018). Sin Fronteras. Obtenido de Procesamiento del Café: Entendiendo el Café Despulpado Natural. Café Despulpado Natural: Una Explicación: https://www.sinfronteraspangoa.com/2018/08/18/procesamiento-del-cafe-entendiendo-el-cafe-despulpado-natural/
- Dagett, Z. (mayo de 2017). Perfect Daily Grind. Obtenido de ¿Cuánto le Cuesta a un Caficultor Sembrar una Parcela Básica?: https://www.perfectdailygrind.com/2017/05/cuanto-le-cuesta-un-caficultor-sembrar-una-parcela-basica/
- El Diario. (25 de enero de 2015). Disminuye producción de café en norte de La Paz. Obtenido de https://www.eldiario.net/noticias/2015/2015_01/nt150125/nacional.php?n=47&-disminuye-produccion-de-cafe-en-norte-de-la-paz
- Federación Nacional de Cafeteros. (2013). Cartilla. Fertilización un excelente negocio. Obtenido de https://www.federaciondecafeteros.org/pergamino-fnc/CartillaFertilizacinUnExcelenteNegocio.pdf
- Federación Nacional de Cafeteros. (16 de mayo de 2018). Tostadora de Café. Obtenido de https://es-la.facebook.com/100porcientocafedecolombia/photos/sab%C3%ADas-que-al-

tostar-caf%C3%A9-la-primera-fase-es-la-de-secado-en-esta-fase-los-grano/10156450160772188/

- Federación Nacional de Cafeteros; Cenicafé. (2004). Cartilla Cafetera Cap. 21. Obtenido de Beneficio del café. 2. Secado del café pergamino.: https://es.scribd.com/doc/103217884/Cartilla-Cafetera-21-Beneficio-Del-Cafe-2
- Federación Nacional de Cafeteros; Cenicafé. (2004). Cartilla Cafetera Cap. 20. Obtenido de Beneficio del café. 1. Despulpado, remoción del mucílago y lavado: https://es.scribd.com/doc/103614650/Cartilla-Cafetera-20-Beneficio-Del-Cafe-1
- Frutos del País Meléndez. (2015). Café verde en granos. Obtenido de https://frutosdelpaismelendez.cl/producto/cafe-verde-en-granos-200-gramos/
- Gaitán, Á., Villegas, C., Rivillas, C., Hinapié, É., & Arcilla, J. (febrero de 2011). Avances Técnicos Cenicafé. Almácigos de café: Calidad fitosanitaria, manejo y siembra en el campo. Obtenido de https://www.cenicafe.org/es/documents/AVT0404.pdf
- Maquinarias Mavimar. (septiembre de 2010). Trilladora de Café. Obtenido de http://maquinariasmavimar.blogspot.com/2010/09/trilladora-de-cafe.html
- Pabón, J., Snaz, J., & Oliveros, E. (octubre de 2009). Avances Técnicos Cenicafé - 388. Obtenido de Manejo del café desmucilaginado mecánicamente: https://www.cenicafe.org/es/publications/avt0388.pdf
- Ramírez, V. (2013). Establecimiento de cafetales al sol. En F. N. Cafeteros, & Cenicafé, Manual del cafetero colombiano. Investigación y tecnología para la sostenibilidad de la caficultura. LEGIS.
- Rivillas Osorio, C. A., & Gaitán León, Á. (2013). Germinadores de café. En F. N. Cenicafé, Manual del cafetero colombiano. Investigación y tecnología para la sostenibilidad de la caficultura.
- SCAA. (18 de junio de 2016). Café Kinetic. Obtenido de Rueda de sabores, SCAA Flavor Wheel: http://www.cafekinetic.net/nueva-rueda-sabores/scaa_flavorwheel-01-18-15/

Anexo 3: Código Python empleado.

El código que se muestra en las siguientes páginas, constituye un soporte del trabajo realizado, no esta diseñado para explicar o detallar la metodología. Debido a las diferentes variantes implementadas durante el trabajo, algunas celdas solo pueden mostrar una variante, un ejemplo de esto es una gráfica de los datos de la variable de salida “Sabor”, con este mismo código se modificaba el campo “Sabor” por el siguiente (ejemplo “Aroma”) y se ejecutaba de nuevo. Este archivo fue un soporte y una herramienta de trabajo más que una guía o el trabajo en sí. Desde la siguiente página se muestra el código, como se genera al exportarlo para impresión, debido a esto algunas gráficas pueden aparecer cortadas.