ESCUELA SUPERIOR POLITÉCNICA AGROPECUARIA DE MANABÍ MANUEL FÉLIX LÓPEZ

CARRERA DE INFORMÁTICA

TRABAJO DE TITULACIÓN PREVIA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN INFORMÁTICA

MODALIDAD: SISTEMATIZACIÓN DE EXPERIENCIAS

TEMA:

MODELO DE INTELIGENCIA COMPUTACIONAL PARA LA DETERMINACIÓN DEL GRADO DE AFECTACIÓN DE LA ROYA EN CAFÉ ROBUSTA

AUTOR:

CESAR ISRAEL ANDRADE SÁNCHEZ

TUTOR:

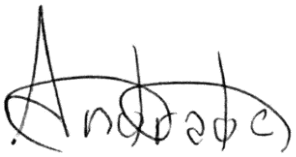
ING. JAVIER HERNÁN LÓPEZ ZAMBRANO MG. TI

CALCETA, OCTUBRE 2021

DERECHOS DE AUTORÍA

César Israel Andrade Sánchez, declara bajo juramento que el trabajo aquí descrito es de su autoría, que no ha sido previamente presentado para ningún grado o calificación profesional, y que ha consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cede los derechos de propiedad intelectual a la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López, según lo establecido por la Ley de Propiedad Intelectual y su reglamento.



.....

CÉSAR ISRAEL ANDRADE SÁNCHEZ

CERTIFICACIÓN DEL TUTOR

Javier Hernán López Zambrano certifica haber tutelado el trabajo de titulación **MODELO DE INTELIGENCIA COMPUTACIONAL PARA LA DETERMINACIÓN DEL GRADO DE AFECTACIÓN DE LA ROYA EN CAFÉ ROBUSTA**, que ha sido desarrollada por César Israel Andrade Sánchez, previa la obtención del título de Ingeniero en Informática, de acuerdo al **REGLAMENTO DE UNIDAD DE TITULACIÓN ESPECIAL DE PROGRAMAS DE GRADO** la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López.

Firmado digitalmente por JAVIER HERNAN LOPEZ ZAMBRANO

……

ING JAVIER HERNÁN LÓPEZ ZAMBRANO MG. TI

APROBACIÓN DEL TRIBUNAL

Los suscritos integrantes del tribunal correspondiente, declaramos que hemos **APROBADO** la Tesis titulada **MODELO DE INTELIGENCIA COMPUTACIONAL PARA LA DETERMINACIÓN DEL GRADO DE AFECTACIÓN DE LA ROYA EN CAFÉ ROBUSTA**, que ha sido propuesta, desarrollada y sustentada por César Israel Andrade Sánchez , previa la obtención del título de Ingeniero en Informática, de acuerdo al **REGLAMENTO DE UNIDAD DE TITULACIÓN ESPECIAL DE PROGRAMAS DE GRADO** de la Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López.

ANGEL ALBERTO VELEZ MERO
Firmado digitalmente.
avelez@espam.edu.ec

……

ING. ÁNGEL A. VÉLEZ MERO, MGTR

MIEMBRO

Firmado electrónicamente por:
ALFONSO TOMAS LOOR

……

ING. ALFONSO T. LOOR VERA, MGTR

MIEMBRO

Firmado electrónicamente por:
LUIS CRISTOBAL CEDENO VALAREZO

……

ING. LUIS C. CEDEÑO VALAREZO, MGTR

PRESIDENTE

AGRADECIMIENTO

En primer lugar, a la Escuela Superior Politécnica Agropecuaria de Manabí - Manuel Félix López, por abrirme las puertas para cumplir mi objetivo principal al culminar el bachillerato, el cual fue obtener un título de tercer nivel, a todas aquellas personas que se vieron involucradas de una u otra forma durante mi desarrollo profesional, sean estos docentes, técnico-docentes, colegas, amigos, compañeros, personal administrativo u operacional.

Al Ingeniero Luis Cedeño por saber recibir mi petición de participación en el proyecto de investigación que él lidera. A mi tutor de trabajo de titulación, el Ingeniero Javier López, por saber sobrellevar este trabajo dando siempre la mejor versión de sí mismo para que el proyecto siga en pie. A los Ingenieros Jorge Párraga y Víctor Pinargote por darme apoyo técnico cuando se los he solicitado estando siempre prestos a colaborar si está a su alcance.

DEDICATORIA

En primer lugar, a mi señora madre, ya que sin su apoyo no habría tenido las fuerzas necesarias para seguir en momentos complicados y de incertidumbre, a mi padre, el cual siempre estuvo allí cuando lo necesité, a todas las amistades que pude forjar durante mi paso por esta institución, con las que disfruté convivir durante todo este proceso.

CONTENIDO GENERAL

CONTENIDO DE CUADROS Y FIGURAS

TABLAS

FIGURAS

CONTENIDO DE GRÁFICOS

RESÚMEN

Esta investigación fue planteada con el propósito de determinar el grado de afectación de la roya en plantas de café robusta mediante la implementación de un modelo de inteligencia computacional (IC). Para ello: (1) se llevó a cabo una revisión bibliográfica para obtener información acerca de los modelos que se utilizan en investigaciones relacionadas al objeto de estudio, concluyendo que el Máquinas de vectores de soporte (SVM por sus siglas en inglés) es el más utilizado y presenta mejores resultados; (2) con el conjunto de datos proporcionados por la carrera de Agrícola, conformado por 12 características fenotípicas de ejemplares de café afectados por la roya, se construyeron varias versiones de conjuntos de datos utilizando 3 criterios de oversampling; (3) se implementó el modelo SVM, el cual se entrenó y testeó con cada una de las versiones de los conjuntos de datos creadas, (4) después, el uso de técnicas de validación de modelos de IC permitió realizar ajustes sobre el modelo hasta obtener resultados que reflejen un alto grado de precisión realizando inferencias. Este proceso permitió que el modelo definitivo obtuviese un rendimiento del 80% durante las pruebas y un 74% de precisión (en promedio) durante la validación, por lo que, se creó un modelo que registra un rendimiento relativamente alto que además puede ser usado en otras investigaciones afines a combatir la enfermedad de la roya en el café.

PALABRAS CLAVE

Inteligencia computacional, café robusto, roya, validación cruzada

ABSTRACT

This research was proposed with the purpose of determining the degree of affectation of rust in robusta coffee plants through the implementation of a computational intelligence (CI) model. To do this: (1) a bibliographic review was carried out to obtain information about the models that are used in research related to the object of study, concluding that the Support Vector Machines (SVM) is the most used and presents better results; (2) with the data set provided by Agrícola's career, made up of 12 phenotypic characteristics of coffee specimens affected by rust, several versions of data sets were constructed using 3 oversampling criteria; (3) the SVM model was implemented, which was trained and tested with each of the versions of the data sets created, (4) later, the use of IC model validation techniques They allow to make adjustments on the model until obtain results that reflect a high degree of precision by making inferences. This process that the definitive model obtained a performance of 80% during the tests and a 74% precision (on average) during the validation, therefore, a model was created that registers a relatively high performance that can also be used in other research related to combating coffee rust disease.

KEYWORDS

computational intelligence, robust coffee, rust, cross validation

CAPÍTULO I. ANTECEDENTES

1.1. DESCRIPCIÓN DE LA INSTITUCIÓN

La Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López (ESPAM MFL) fue creada en abril de 1999 para participar junto a otras instituciones en el auge y desarrollo de la provincia de Manabí y del país, a través de la enseñanza universitaria, la investigación científica y el emprendimiento. El contexto rural y socioeconómico manabita, con un alto potencial productivo, resultó determinante para la elección de las carreras, todas ellas vinculadas al agro y/o a la gestión de las empresas agroindustriales (ESPAM MFL, 2016).

Mediante la Coordinación General de investigación, se encuentra realizando trabajos mancomunados por el aporte de la universidad a la ciencia, contando con diversos grupos de investigación, entre los cuales está el de Sistemas Computacionales (SISCOM), cuya misión es desarrollar investigaciones computacionales con pertinencia, compromiso ético y social, en procura del mejoramiento del sector agro-productivo y de servicios, mientras que, la visión manifiesta que dicho grupo busca ser un referente en investigación y desarrollo relacionados a sistemas y tecnologías computacionales que contribuyan al progreso agro-productivo y de servicio a nivel local o nacional (ESPAM MFL, 2020).

SISCOM es dirigido actualmente por la Ing, Jessica J. Morales Carrillo, y cuenta con tres proyectos de investigación en ejecución; entre los cuales está: “Caracterización de la roya en cultivos de Café Robusta mediante técnicas avanzadas de inteligencia computacional”, cuyo objetivo general es: emplear inteligencia computacional para la caracterización fenotípica de la roya en cultivos de café robusta, con el fin de apoyar los mecanismos de control de la enfermedad y mejorar la productividad.

1.2. DESCRIPCIÓN DE LA INTERVENCIÓN

La agricultura ocupa una posición importante en el desarrollo económico de toda sociedad, y es una de las fuentes más importantes de ingresos para los seres humanos en muchos países (Jun-De *et al.*, 2020). Entre todos los productos naturales, el valor monetario del comercio del café sólo es superado por el del petróleo, este cultivo genera ingresos y brinda oportunidades laborales en muchos países tropicales (Daba *et a*l., 2019).

Aunque se conocen más de 100 especies de café, sólo dos especies, es decir, *Coffea arabica* (conocida como café arábica) y *Coffea canephora* (conocido como café robusta), son comercialmente cultivadas y consumidas (Kosaraju *et al.*, 2017).

En el Ecuador, el cultivo del café, constituye una de las principales actividades agrícolas, pues se encuentra entre los diez cultivos con mayor superficie, convirtiéndose así en una actividad de gran importancia económica y social (Guerrero, 2017) representando según BCE (Banco Central del Ecuador) un 5.8% del PIB hasta septiembre del 2019.

Al igual que cualquier cultivo, el cafeto es vulnerable a las plagas y las enfermedades. Un informe de 2012 de Fabienne Ribeyre, investigador en el centro de investigación agrícola de Francia CIRAD, afirma que “la mayoría de las enfermedades del café son causadas por hongos patógenos y, con menor frecuencia, por bacterias y virus” (Molina, 2019).

Una de las enfermedades que más se destaca por haber generado a lo largo de los años cuantiosas pérdidas en la producción de café sin duda es la roya, misma que según Chemura *et al*. (2018), es causada por el hongo Hemileia vastatrix y representa la mayor amenaza para la industria mundial del café.

El manejo integrado de cultivos define estrategias que abarcan varios conceptos a tener en cuenta en el desarrollo de éstos: la sustentabilidad del medio ambiente, el manejo de plagas y enfermedades, y el uso eficiente de los recursos con el objetivo de lograr el mejor rendimiento posible (Ehman *et al.,* 2018). Sin embargo, los agricultores enfrentan una serie de desafíos, incluidas diferentes enfermedades para las plantas (Jun-De *et al*., 2020).

Debido a esto existen múltiples investigaciones que tratan de abarcar este problema usando diferentes enfoques como es el machine learning, el cual, según BBVA (Banco Bilbao Vizcaya Argentaria, 2019), es una rama de la inteligencia artificial que permite que las máquinas aprendan y sean capaces de identificar patrones entre los datos para hacer predicciones.

Estas predicciones sirven para determinar qué sucederá a futuro, y una vez se tengan nuevos datos poder hacer nuevas inferencias. La idea principal es generar modelos computacionales que se puedan usar en cualquier momento y que faciliten esta tarea.

El presente trabajo plantea la creación de un modelo de predicción multiclase con el fin de determinar el grado de incidencia de roya definido en 5 niveles, partiendo de características fenotípicas, los cuales según López (2015), son las que “vemos” de las plantas y resultan de la interacción de la información genética (genotipo) y el ambiente donde estas se desarrollan. Además, este trabajo se encuentra ligado al proyecto de investigación “Caracterización de la Roya en Cultivos de Café Robusta mediante técnicas avanzadas de inteligencia computacional”, llevado a cabo por el grupo de investigación SISCOM, aportando a la continuidad de éste a través del cumplimiento de su segundo objetivo específico.

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Implementar un modelo de inteligencia computacional para identificar el grado de afectación de la roya en plantas de café robusta usando datos fenotípicos como entrada para la predicción.

1.3.2. OBJETIVOS ESPECÍFICOS

- Determinar una técnica de aprendizaje automático acorde al fin de la investigación
- Construir un dataset con los datos históricos obtenidos a través de la recolección de información
- Desarrollar un modelo computacional usando la técnica seleccionada
- Validar la precisión del modelo

CAPÍTULO II. DESARROLLO METODOLÓGICO DE LA INTERVENCIÓN

Para llevar a cabo la ejecución del presente trabajo, se tomó como referencia la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos (Villena, 2016). Esta metodología, según Vallalta (2019) cuenta con las fases mostradas en la figura 2.1:

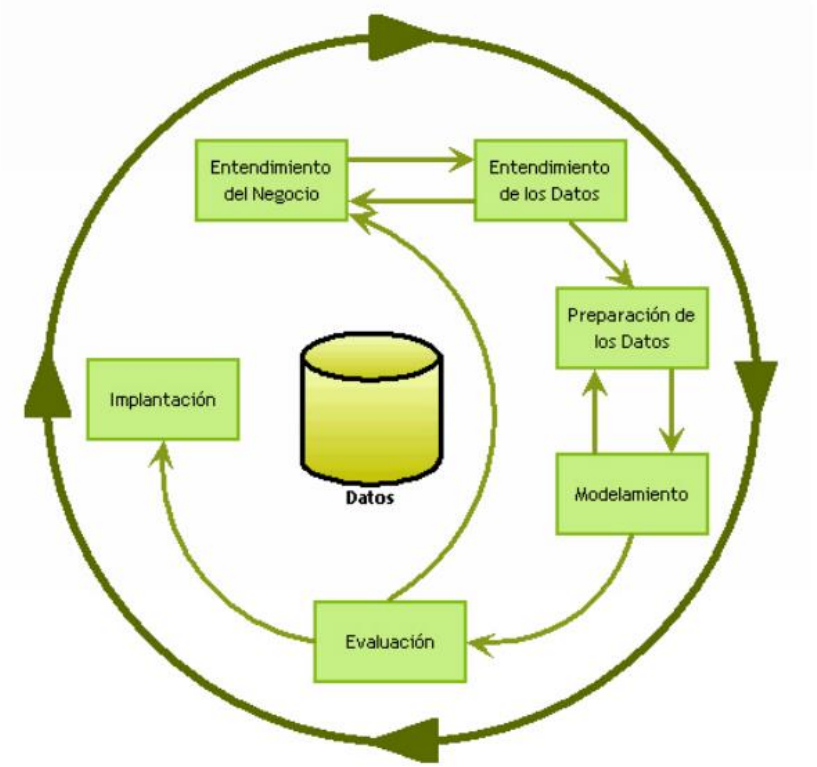


Figura 2.1 Fases de la metodología CRISP-DM
Fuente. Galán, 2015

Si bien esta metodología sirvió como base para el desarrollo de este trabajo, la misma sufrió algunas modificaciones con el fin de lograr un mejor acoplamiento a las

necesidades de la investigación, suprimiendo la fase despliegue, quedando como resultado las fases mostradas en la figura 2.2:

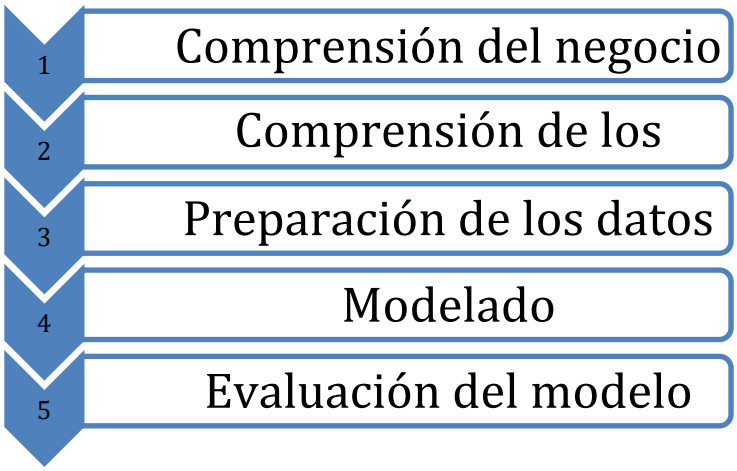


Figura 2.2 Fases de la investigación
Fuente. El autor, 2020

2.1. COMPRENSIÓN DEL NEGOCIO

Como parte de la comprensión del problema se llevó a cabo una revisión de carácter bibliográfico, todo esto con el fin de plantear un antecedente que pueda servir de base para plasmar de mejor forma una solución acorde a las necesidades de esta investigación.

Para este fin, se utilizó la metodología revisión sistemática, la cual según Moreno *et al.* (2018) permite definir resúmenes claros y estructurados de la información disponible orientada a responder una pregunta específica.

El análisis de artículos de carácter científico, trabajos de titulación de pregrado o grado y publicaciones de sitios web se llevó a cabo mediante el flujo de acciones mostrado en la figura 2.3.

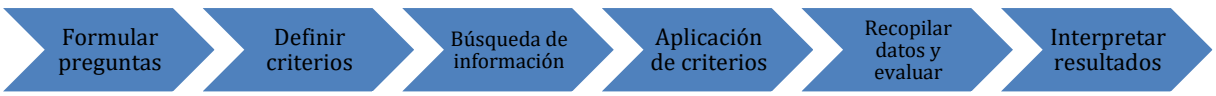


Figura 2.3 Etapas de la revisión sistemática
Fuente. El autor, 2020

2.1.1. FORMULAR LA PREGUNTA DE INVESTIGACIÓN

Toda investigación que aplique la revisión sistemática como metodología de búsqueda y selección de información, parte con la formulación de una pregunta, misma que permitirá definir el rumbo de dicha búsqueda.

Debido al creciente uso de la inteligencia artificial en el sector agro productivo, preguntas como ¿Existen investigaciones que documenten y sirvan de referencia para nuevos trabajos relacionados a esta área?, ¿De qué forma se pueden abarcar problemáticas relacionadas al agro utilizando estrategias de IC?, ¿Se puede considerar el uso de un modelo de IC basado en su rendimiento para resolver esta clase de problemas?

2.1.2. DEFINIR CRITERIOS DE ELEGIBILIDAD Y EXCLUSIÓN

Los trabajos o investigaciones a considerar deben estar inmersos bajo los siguientes criterios:

1. Utilizar técnicas de IA para resolver problemas relacionados con el agro;
2. Trabajos donde se haga alguna clase de caracterización vegetativa;
3. Trabajos de los últimos 5 años.

Mientras que, para los criterios de exclusión, se consideran:

A. Trabajos no relacionados a la problemática;
B. Investigaciones longevas;
C. Propuestas no relevantes o poco concluyentes.

2.1.3. BUSCAR INVESTIGACIONES

La búsqueda de información se realizó usando como fuente diversos repositorios digitales como ScienceDirect, Scielo, IEEExplorer e incluso la red de repositorios de acceso abierto del Ecuador; en donde se analizaron trabajos de muchas áreas, tratando siempre de que éstos aporten directamente a la investigación con información actualizada.

Información de cada una de las investigaciones encontradas se ubicó en un spreadsheet de Google, seleccionando datos como año, autor(es), título, entre otros. Una vez que la cantidad de trabajos sea convincente, se proseguirá con la aplicación de los criterios de selección.

2.1.4. APLICAR CRITERIOS DE SELECCIÓN

Para esta etapa, se realizó una lectura parcial de los documentos encontrados, en busca de puntos clave que aporten a la presente investigación, y que a su vez cumplan con los criterios planteados previamente.

2.1.5. RECOPILAR DATOS Y EVALUAR CRÍTICAMENTE

Una vez que los artículos fueron seleccionados, se realizó un análisis más detallado de los resultados obtenidos en cada investigación, ya que a través de esto se puede discernir qué modelo de IC es el más adecuado para resolver la problemática planteada, para lo cual, se analizaron los valores del rendimiento de los principales modelos utilizados en dichas investigaciones para establecer una comparativa entre los mismos.

2.1.6. INTERPRETAR LOS RESULTADOS

Con los resultados de la implementación de esta metodología, se pudo seleccionar un conjunto de investigaciones que cumplan con los criterios de selección, dichas investigaciones permitieron definir qué modelos son comúnmente utilizados para cubrir problemáticas relacionadas al agro, permitiendo de este modo determinar cuál sería ese modelo que logrará dar solución al objetivo de este trabajo.

2.2. COMPRENSIÓN DE LOS DATOS

El proceso de obtención de los datos se logró gracias a la coordinación entre las carreras de Agrícola y Computación, mismas que comparten un proyecto de investigación sobre la roya en plantas de café.

Este proceso tuvo lugar en la Ciudad de Investigación e Innovación y Desarrollo Agropecuario (CIIDEA) de la ESPAM MFL, en donde se analizaron especímenes de

café ubicados en esta unidad, realizando una recolección de información de 12 variables fenotípicas.

Con los datos obtenidos, un análisis de los datos es primordial, con el objetivo de según Huber *et al.* (2019) buscar nuevos patrones de frecuencia en los flujos de datos. Además, se establecieron reuniones con investigadores del área agro productiva con el fin de establecer el acercamiento conceptual adecuado para el uso de esta información.

2.3. PREPARACIÓN DE LOS DATOS

Una vez analizados los datos, se realizó la preparación del conjunto que servirá como entrada para el modelo a entrenar. Para esto, en primer lugar, se aplicaron técnicas de data cleansing, data transformation y oversampling, con la finalidad de eliminar campos vacíos, inconsistencias en la información, registros innecesarios, duplicados o desbalance de datos.

Después se preparó un conjunto de datos con las variables más relevantes, aplicando para ello una selección de características, misma que, según Shaikh (2018) es un proceso en el que se selecciona automática o manualmente las características que más contribuyen a su variable de predicción o salida.

2.4. MODELADO

Para esta fase, se utilizó el lenguaje de programación Python y sus múltiples librerías de inteligencia artificial, entre las cuales está Scikit-learn, misma que cuenta una considerable cantidad de algoritmos de IC implementados. Además, se definió un flujo de trabajo ágil, el cual abarca los cuatro pasos especificados en la figura 2.4:

1. Construir el modelo con Python, empleando las librerías de inteligencia artificial disponibles.
2. Entrenar el modelo usando el 80% del conjunto de datos.
3. Usar el otro 20% del conjunto de datos para probar el modelo entrenado.
4. Con los resultados de las pruebas preliminares hacer las modificaciones necesarias, incluir conjuntos de datos de validación para mejorar el rendimiento del modelo.

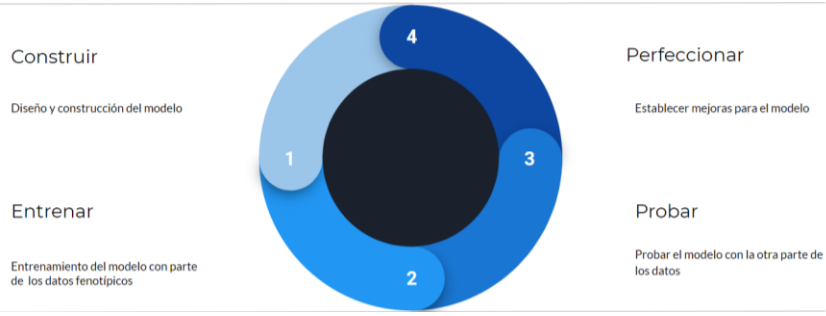


Figura 2.4 Etapas de desarrollo del modelo
Fuente. El autor, 2020

2.5. EVALUACIÓN DEL MODELO

Una vez entrenado el modelo, para la evaluación de su rendimiento se utilizaron varias métricas de calidad como son la precisión, la recuperación (recall) y la puntuación f1 (f1 score). Esto permitió realizar ajustes en su configuración del modelo y así obtener mejores resultados.

2.5.1. PRECISIÓN

La precisión es el número de verdaderos positivos (VP) dividido por el número de verdaderos positivos y falsos positivos (FP). Dicho de otra manera, es el número de predicciones positivas dividido por el número total de valores de clase positivos predichos. Una precisión baja también puede indicar una gran cantidad de falsos positivos (Brownlee, 2019).

$$Precisión = \frac{VP}{VP + FP}$$

2.5.2. RECUPERACIÓN (RECALL)

Según Brownlee (2019) la recuperación representa el número de verdaderos positivos dividido por el número de verdaderos positivos y el número de falsos negativos (FN). En otras palabras, es el número de predicciones positivas dividido por el número de valores de clase positivos en los datos de la prueba. Una recuperación baja indica muchos falsos negativos.

$$Recuperación = \frac{VP}{VP + FN}$$

2.5.3 PUNTUACIÓN F1 (F1 SCORE)

La puntuación F1, puntuación F o Medida F. transmite el equilibrio entre la precisión y la recuperación (Brownlee, 2019).

$$Puntuación_{F1} = 2 * \frac{Presición * Recuperación}{Presición + Recuperación}$$

Además de las métricas previamente mencionadas, el uso de técnicas de evaluación de modelos de IC permitirá probar múltiples configuraciones del modelo en simultáneo y obtener la que tuvo el mejor rendimiento. Adicionalmente, para la validación del modelo se contará con conjuntos de datos de validación, datos que el modelo desconoce. Esto permitirá conocer cómo se comporta el modelo realizando predicciones sobre nuevos conjuntos de datos.

CAPÍTULO III. DESCRIPCIÓN DE LA EXPERIENCIA

3.1. COMPRENSIÓN DEL NEGOCIO

Mediante la revisión sistemática (proceso que se grafica en la figura 3.1) se pudo llevar a cabo una investigación con criterios de inclusión y exclusión bien definidos, evitando en toda medida seleccionar literatura que no represente un aporte considerable a este trabajo. La idea principal fue tener conocimiento de aquellas investigaciones en donde se apliquen técnicas de IA para resolver problemas relacionados con el agro.

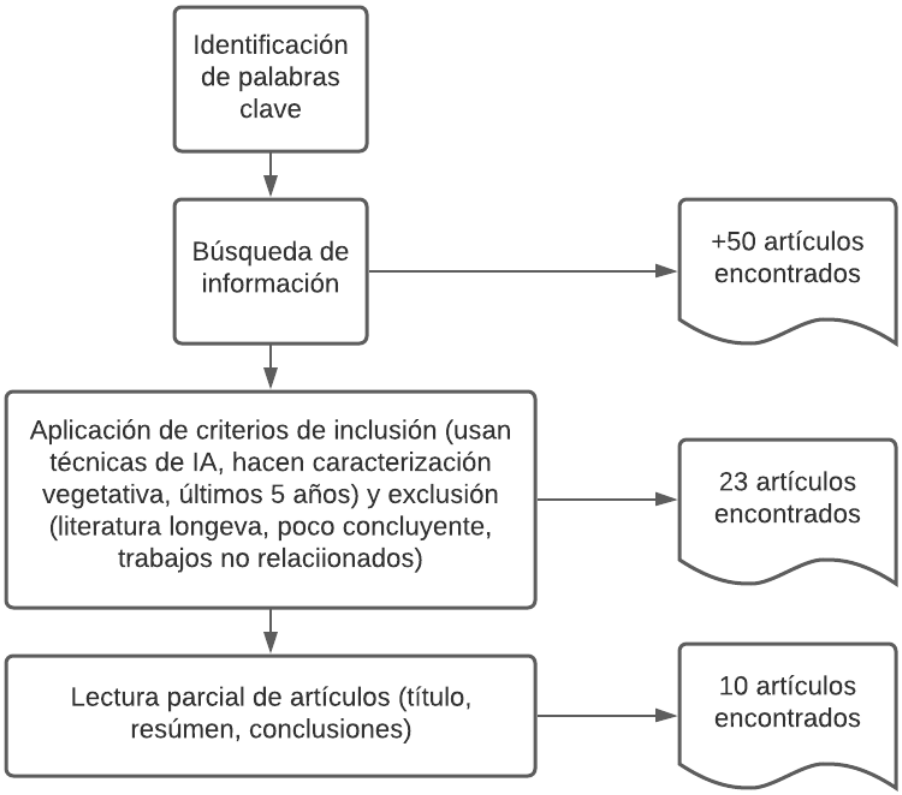


Figura 3.1 Aplicación de la revisión sistemática
Fuente. El autor, 2020

Se consideraron investigaciones de varias fuentes, como Science Direct, Scielo, IEEExplorer o repositorios de universidades, trabajos de diferentes grados profesionalizantes como tesis de pre y posgrado. La información de los artículos más relevantes fue ubicada en un documento de Google Sheets, en donde se extrajeron datos como el título del trabajo, el año, los autores, modelos de IC y técnicas de IA usadas, así como el link de acceso directo a cada investigación.

Cuando se inició la fase de revisión se logró identificar alrededor de 50 investigaciones presumiblemente relevantes, sin embargo, durante la aplicación de los criterios de inclusión y exclusión ese número se redujo a 23, estas investigaciones fueron sometidas a una lectura parcial (título, resumen, conclusiones), lo cual permitió "filtrar" un poco más el conjunto de literatura encontrada; al finalizar esta actividad, se obtuvo un compendio de 10 investigaciones, las cuales se detallan en la tabla 3.1.

Tabla 3.1 Fuentes seleccionadas después de la revisión

Año	Tema	Autor (es)	Técnica	Tipo	Modelos
2020	A self-adaptive classification method for plant disease detection using GMDH-Logistic model	Chen Jun-Dea Yin Huayi Zhang De-Fua	Machine Learning	Aprendizaje **Supervisado**	GMDH-Logistic
2020	Hybrid System for Detection and Classification of Plant Disease Using Qualitative Texture Features Analysis	Anjnaa Meenakshi Sooda Pradeep Kumar Singhb	Machine Learning	Aprendizaje Supervisado	Support Vector Machine
2019	Aplicación Móvil de Detección y Clasificación de "la Roya" en hojas de Café Robusta mediante Aprendizaje Automático	Cusme Zambrano, Kevin Daniel Loor Pinargote, Angélica María	Machine Learning	Aprendizaje no Supervisado	Convolutional Neural Networks
2018	Plant leaf disease identification using exponential spider monkey optimization	Sandeep Kumar Basudev Sharma Vivek Kumar Sharma Harish Sharma Jagdish Chand Bansal	Machine Learning	Aprendizaje Supervisado	Support Vector Machine
2018	Deep learning models for plant disease detection and diagnosis	Konstantinos Ferentinos	Machine Learning	Aprendizaje no Supervisado	Convolutional Neural Network
2018	Plant Disease Detection Using Machine Learning	Shima Ramesh Maniyath Vinod P V Niveditha M Pooja R Prasad Bhat N Shashank N Hebbar Ram	Machine Learning	Aprendizaje Supervisado	Support Vector Machine
2016	Detection and classification of diseases of Grape plant using opposite colour Local Binary Pattern feature and machine learning for automated Decision Support System	Harshal Waghmare Radha Kokare Yogesh Dandawate	Machine Learning	Aprendizaje Supervisado	Support Vector Machine

2016	Ethiopian Coffee Plant Diseases Recognition Based on Imaging and Machine Learning Techniques	Abrham Debasu Mengistu Dagnachew Melesew Alemayehu Seffi Gebeyehu Mengistu	Machine Learning	Aprendizaje Supervisado	Self Organizing Map
2016	SVM and ANN Based Classification of Plant Diseases Using Feature Reduction Technique	Jagadeesh D.Pujari Rajesh Yakkundimath Abdulmunaf. Syedhusain Byadgi	Machine Learning	Aprendizaje Supervisado	Support Vector Machine
2018	Machine learning prediction of coffee rust severity on leaves using spectroradiometer data	Abel Chemura Onisimo Mutanga Mbulisi Sibanda Pardon Chidoko	Machine Learning	Aprendizaje Supervisado	Radial Basis Function-Sparse Partial Least Squares

La tabla 3.1 muestra información como el año de publicación, título de la investigación, autores, técnica y tipo de inteligencia artificial usados, así como el modelo con el mejor rendimiento obtenido.

Como resultado de esta revisión se pudo notar que la técnica de inteligencia artificial más utilizada es el machine learning o aprendizaje automático, y que el aprendizaje supervisado es el tipo de aprendizaje aplicado con mayor frecuencia representando un 80% de uso como lo muestra el gráfico 3.1.

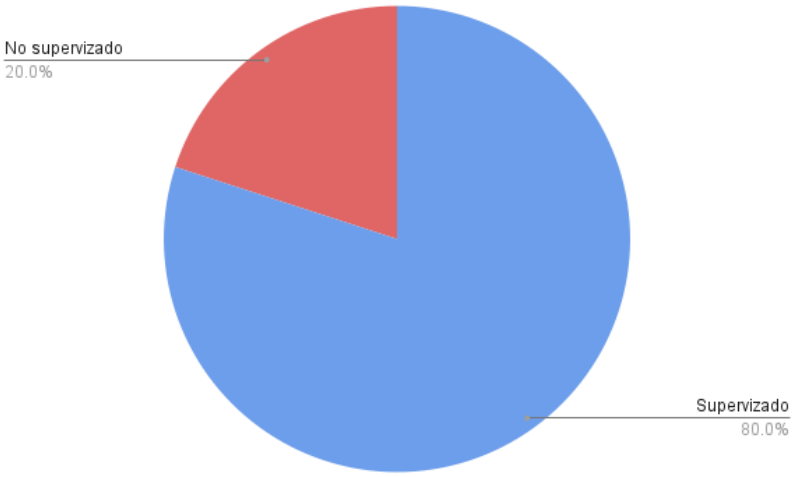


Gráfico 3.1 Porcentaje de selección de tipo de aprendizaje automático
Fuente. El autor, 2020

Los modelos de inteligencia artificial utilizados y la frecuencia con la que éstos fueron implementados se detallan en el gráfico 3.2, en donde además se muestra que el modelo más utilizado es Support Vector Machine, lo cual refleja que éste registra un rendimiento bastante satisfactorio abarcando problemáticas relacionadas al agro.

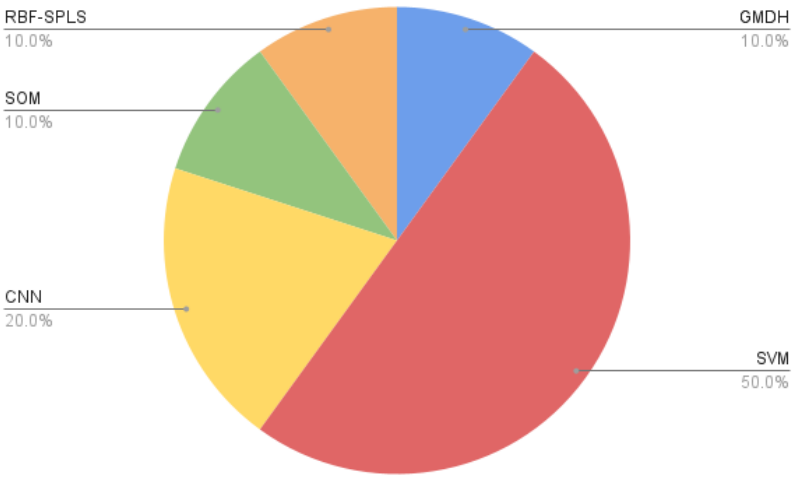


Gráfico 3.2 Porcentaje de uso de modelos de aprendizaje supervisado
Fuente. El autor, 2020

Si bien la gráfica 3.2 muestra el porcentaje de uso de los modelos de IC, el gráfico 3.3 muestra qué porcentaje de precisión obtuvo el mejor modelo de cada una de las investigaciones seleccionadas.

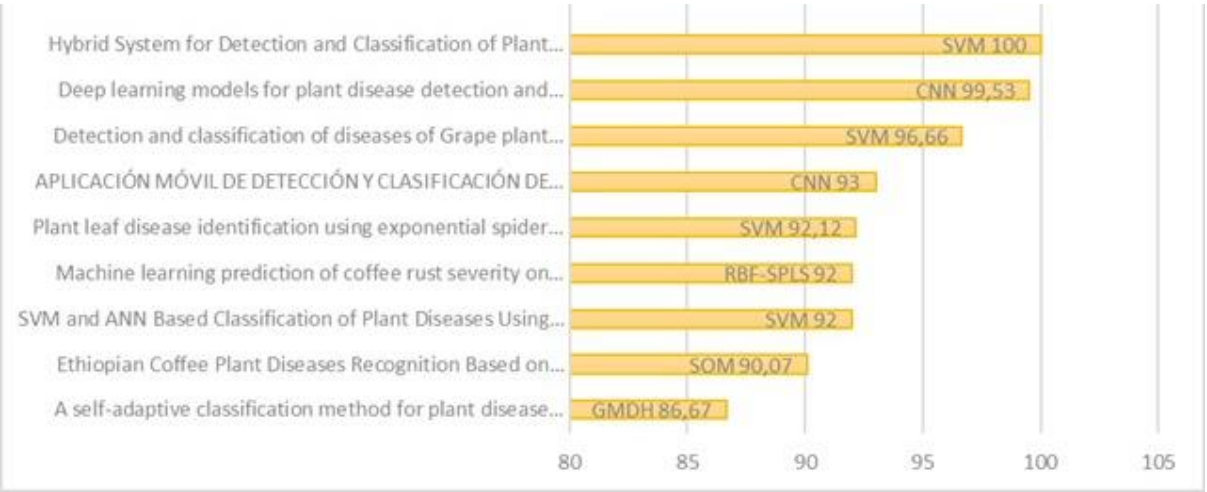


Gráfico 3.3 Representación del rendimiento de los algoritmos usados en la literatura consultada
Fuente. El autor, 2021

3.2. COMPRENSIÓN DE LOS DATOS

En esta fase se llevó a cabo la recolección y comprensión de los datos necesarios para el entrenamiento y prueba del modelo, dichos datos fueron obtenidos en CIIDEA en la ESPAM MFL. La recolección fue llevada a cabo por la carrera de Agrícola, misma que generó un conjunto de datos que contemplaba mediciones de 215 especímenes de café, agrupados en 17 “clones” o variedades.

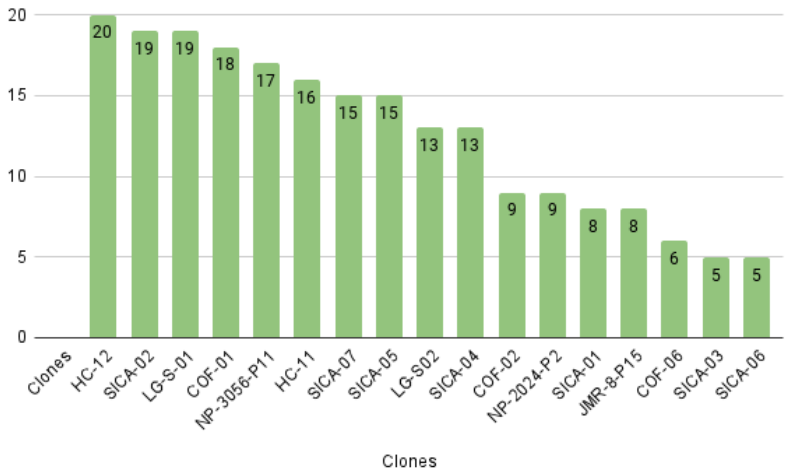


Gráfico 3.4 Porcentaje de plantas por clon
Fuente. El autor, 2021

El gráfico 3.4 muestra la distribución del número de plantas por clon que conforman este banco de información, de las cuales se tabularon una cantidad de 12 variables fenotípicas que describen cómo se comporta una planta bajo ciertas condiciones de cualquier índole.

Tabla 3.2 Variables del conjunto de datos inicial

Variable	Tipo de variable	Unidad de medida	Descripción
Clones	Cualitativa	Ninguna	Especifica cuál es la variedad (clon) de una planta
Código	Cualitativa	Ninguna	Refleja un código único para cada planta
Plantas Vivas	Cualitativa	Ninguna	Representa el número de plantas vivas por clon
Altura de planta	Cuantitativa	cm	Define la longitud de la planta desde el suelo hasta la copa
Número de ramas	Cuantitativa	N°	Especifica el número de ramas de la planta
Producción	Cualitativa	Si/No	Determina si una planta está produciendo o no
Si está en producción	Cualitativa	0,1,2,3,4,5	Establece una escala del 0 al 5 el nivel de producción de una planta, en donde 0 representa que la planta no produce y 5 el nivel más alto de producción
Presencia de plagas	Cualitativa	Si/No	Representa (de forma global) la presencia o no de plagas en una planta
Si hay plagas	Cualitativa	0,1,2,3,4	Define en una escala del 0 al 4 el nivel de afectación de plagas de una planta, en donde 0 representa la no presencia de plagas y 4 el nivel más alto de afectación
Presencia de roya	Cualitativa	Si/No	Determina si una planta se encuentra infectada por roya o no
Si hay roya	Cualitativa	1,2,3,4,5	Define en una escala del 0 al 4 el nivel de afectación de roya de una planta, en donde 0 representa la no presencia de roya y 4 el nivel más alto de afectación
Si hay otras enfermedades	Cualitativa	Si/No	Determina si una planta se encuentra infectada por otras enfermedades o no
Incidencia de enfermedades distintas de roya	Cualitativa	0,1,2	Define en una escala del 0 al 4 el nivel de afectación de otras enfermedades de una planta, en donde 0 representa la no presencia de otras

			enfermedades y 2 el nivel más alto de afectación
Producción (gramos/planta)	Cuantitativa	gramos	Mide el peso total de todos los frutos producidos por una planta durante toda su vida
Peso de 50 frutos	Cuantitativa	gramos	Mide el peso de los 50 últimos frutos producidos por una planta

La tabla 3.2 sintetiza las variables consideradas durante la recolección de información, variables fenotípicas y variables complementarias, así como la unidad de medida de cada una de ellas y el tipo de variable (cualitativa o cuantitativa).

Una vez realizada la caracterización de los atributos/variables, fue necesario la aplicación de métodos que permitan conocer con mayor profundidad el comportamiento de ciertas características. Para llevar a cabo esta tarea, el Exploratory Data Analysis (EDA) es un enfoque ampliamente utilizado, permitiendo obtener información del conjunto de datos mediante gráficos estadísticos facilitando así la toma de decisiones.

Durante la realización del EDA uno de los descubrimientos más notorios fue encontrar un considerable desbalance de los datos con respecto a la variable objetivo (Nivel_roya) tal y como se muestra en el gráfico 3.5, motivo por el cual se aplicaron técnicas de oversampling con el fin de reducir el impacto negativo que este desbalance pueda ocasionar al rendimiento del modelo.

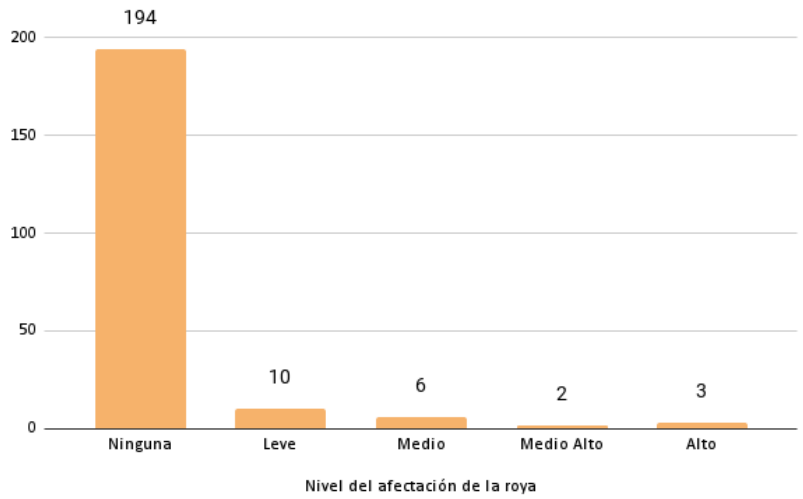


Gráfico 3.5 Distribución inicial de la variable de estudio Nivel Roya
Fuente. El autor, 2021

La figura 3.2 muestra cómo luce una hoja de café infectada por roya dependiendo del nivel de afectación.

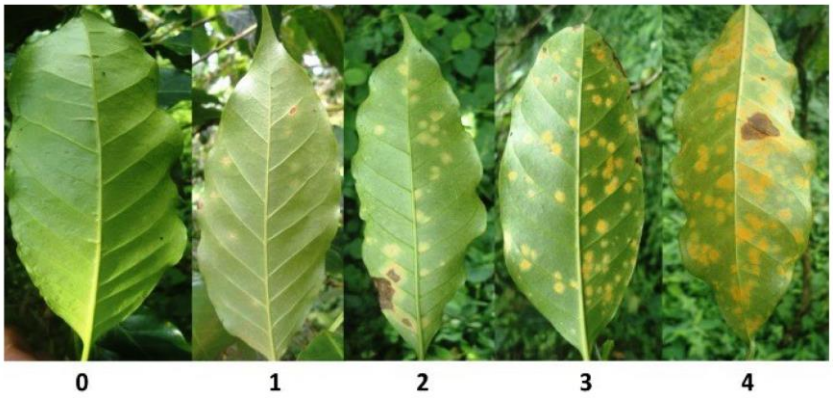


Figura 3.2 Así luce una hoja de café dependiendo del grado de afectación de la roya
Fuente. De Melo & Astorga, 2020

Continuando con el EDA, si se analizan los clones de café con respecto a la variable de estudio, tal y como se muestra en el gráfico 3.6: clones como LG-S02, COF-01 o SICA-07 son quienes presentan un mayor número de especímenes infectados por roya, lo cual puede sugerir que estos clones son más propensos a sucumbir ante esta enfermedad.

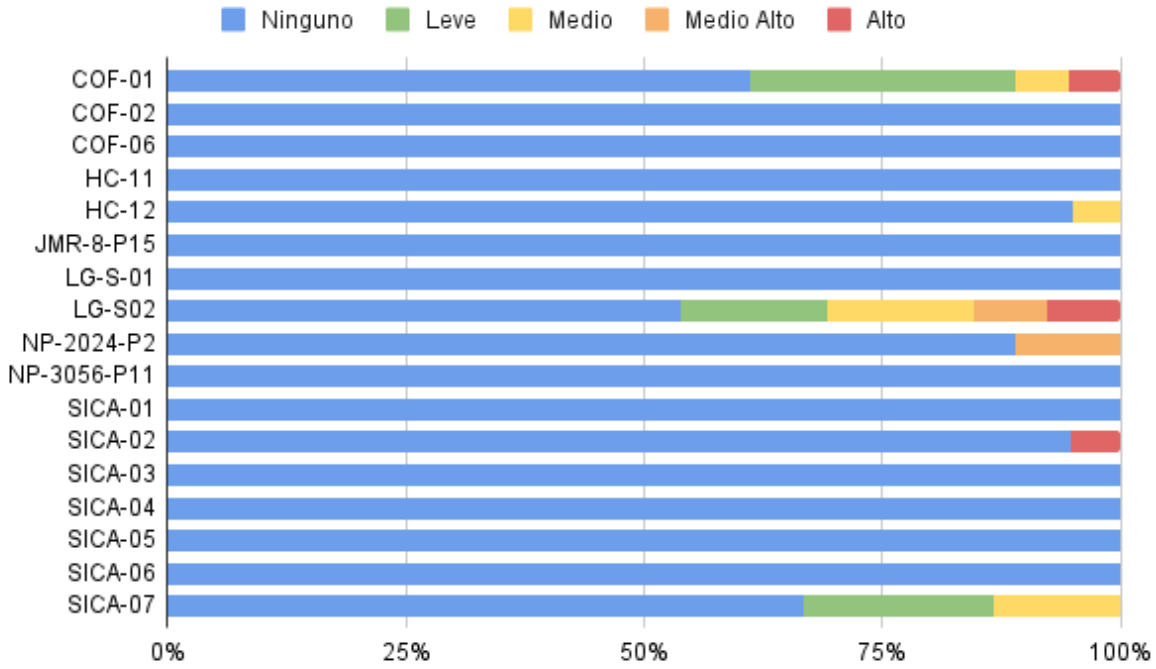


Gráfico 3.6 Distribución del nivel de roya con respecto a los clones analizados
Fuente. El autor, 2021

Por otro lado, si se analizan los niveles de producción de todos los clones, la gráfica 3.7 muestra que, clones como el HC-11 o los clones HC-12, LG-S01, LG-S02 y SICA-07 presentan un rendimiento productivo bastante elevado a pesar de que varios de ellos cuentan con un número importante de especímenes infectados por roya.

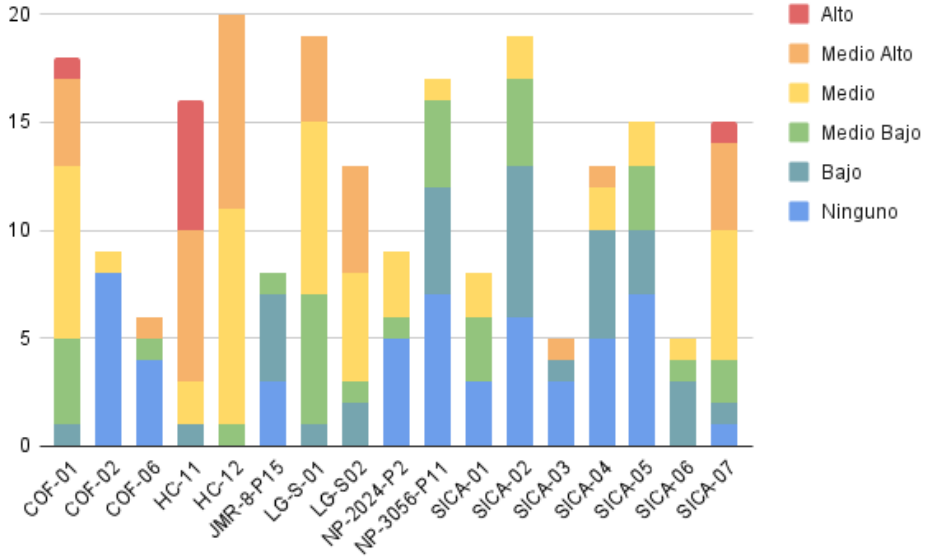


Gráfico 3.7 Distribución del nivel de producción con respecto a los clones analizados
Fuente. El autor, 2021

3.4. PREPARACIÓN DE LOS DATOS

Para la ejecución de esta etapa, se realizaron diversas actividades con el fin de que los datos tuvieran el formato más adecuado para el modelo. Actividades como el renombramiento de las columnas, transformación e imputación de datos o reducción de características fueron llevadas a cabo.

La transformación de datos fue necesaria para cambiar el valor de ciertas características cualitativas y reemplazarlos por valores numéricos, ya que el modelo seleccionado (SVM) únicamente acepta como entrada este tipo de datos.

Por otro lado, para la selección de características, se utilizó el valor estadístico de correlación de las variables (valores reflejados en la gráfica 3.8), en donde se eliminaron variables fuertemente correlacionadas, siempre y cuando el valor de correlación supere el 60%.

Gráfico 3.8 Representación del porcentaje de correlación de las variables que conforman el conjunto de datos
Fuente. El autor, 2021

Además del uso del porcentaje de correlación para eliminar ciertas variables, se descartaron aquellas que poseían valores de carácter booleano (0, 1). Dichas columnas fueron “Plagas”, “Roya” y “Otras_enfermedades”, ya que el conjunto de datos original cuenta con otras variables que también son numéricas pero, con una escala no binaria que complementan a las variables antes indicadas, por ejemplo la columna “Plaga” define si una planta tiene plagas o no (1, 0), mientras que la columna “Nivel_plaga” determina el nivel de afectación de una plaga, representada en una escala de 0 a 4, donde 0 corresponde a la no presencia de plagas y 4 el nivel más alto de afectación concluyendo que estas variables de tipo booleano no aportan mayor información al conjunto de datos.

Como se mencionó en la etapa anterior, la variable de estudio Nivel_roya presenta un desbalance considerable, lo cual puede causar overfitting del modelo o predicciones erróneas a la hora de probarlo, motivo por el cual se aplicaron varios criterios de oversampling.

Uno de estos fue mediante la implementación de ADASYN, el cual según Aditsania et al. (2017) es un algoritmo que usa la distribución de densidad β como referencia para determinar el número de datos sintéticos generados a partir de datos menores.

Otro de los criterios aplicados fue uno en donde se desarrolló un método Custom (es decir, creado por el autor), el cual genera registros sintéticos a partir de la media y de la desviación estándar de los valores de ciertas características del conjunto de datos. Para lograr esto, se establece un límite superior (media + desviación estándar) y un límite inferior (media - desviación estándar) como lo muestra la gráfica 3.9.

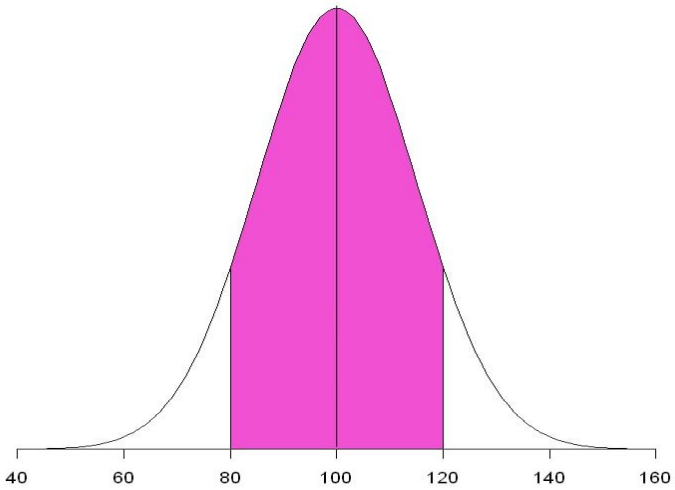


Gráfico 3.9 Ejemplificación del rango de generación de registros sintéticos para el método Custom
Fuente. El autor, 2021

Y, por último, el criterio en donde se combina el método ADASYN con el método Custom para generar un tercer conjunto de datos, dando como resultado 3 conjuntos de datos finales, los cuales fueron usados en las fases posteriores. La distribución de la variable Nivel_roya una vez aplicados todos los criterios se muestra en la tabla 3.3

Tabla 3.3 Porcentaje de registros por clase después de aplicar los criterios de oversampling

Criterios	Leve	Medio	Medio Alto	Alto
ADASYN	24.6%	24.9%	25.2%	25.3%
Custom	25%	25%	25%	25%
Híbrido	25%	25%	25%	25%

3.5. MODELADO

Para la construcción del modelo se tomaron como base los resultados obtenidos de la revisión sistemática realizada en la primera etapa DD de la investigación, en la cual

se demuestra que el modelo SVM es aquel que presenta un mejor rendimiento siendo utilizado en investigaciones aplicadas a la agricultura.

Para la implementación del modelo se siguió el flujo de acciones especificado en la figura 3.3, en donde se establece un ciclo de mejora progresiva basado en el rendimiento del modelo.

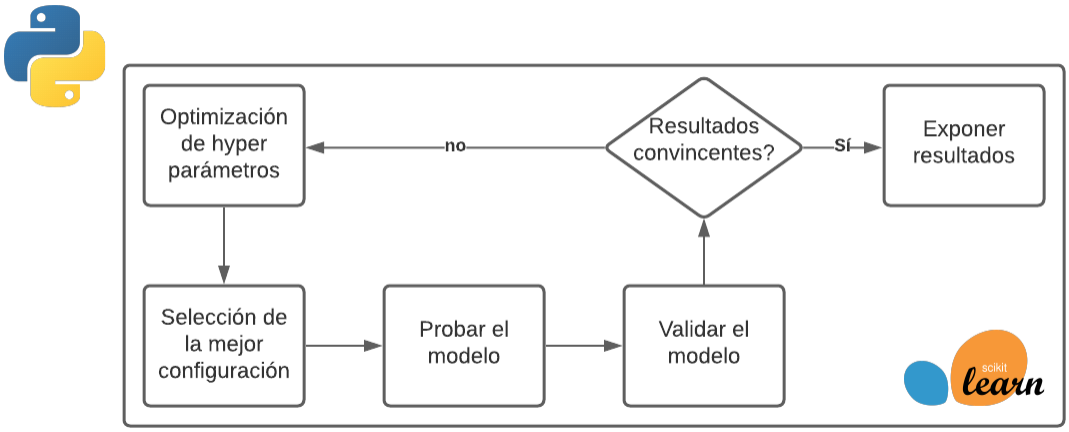


Figura 3.3 Flujo de trabajo durante la fase de modelado
Fuente. El autor, 2021

Durante el proceso de preparación de los datos se usaron 3 criterios de oversampling: ADASYN, Custom, e híbrido que combina los dos criterios ya mencionados; por tal motivo se generaron 3 conjuntos de datos, mismos que fueron usados para crear tres versiones del modelo.

Para ello, en primer lugar, se cargó la data con la que se iba a entrenar/probar el modelo, usando Pandas se hace referencia a la ubicación del conjunto de datos a utilizar, tal y como muestra la figura 3.4.

```
import pandas as pd

balanced_data = pd.read_csv('./data/balanced_data_custom.csv')
```

Figura 3.4 Carga de los datos
Fuente. El autor, 2021

Para el entrenamiento/prueba de cada versión del modelo se dividieron los datos en subconjuntos comprendidos por el 80% de datos para el entrenamiento y 20% para prueba. Para la separación del conjunto de datos se utilizó Scikit Learn mediante la sentencia mostrada en la figura 3.5.

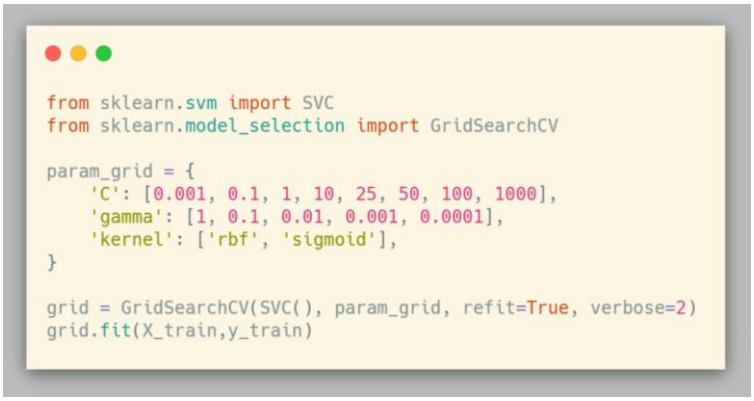
```
from sklearn.model_selection import train_test_split

X = balanced_data.drop('Nivel_roya', axis=1)
y = balanced_data['Nivel_roya']

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, random_state=0)
```

Figura 3.5 Proceso de separación de los datos
Fuente. El autor, 2021

Para la optimización hiper paramétrica del modelo se utilizó GridSearchCV, el cual es un método que a través de la búsqueda de múltiples combinaciones de parámetros de un modelo (partiendo de un conjunto de valores iniciales), obtiene la mejor configuración de éste en función de su rendimiento. Cabe resaltar que la selección es realizada por el propio método y, al final solo queda realizar predicciones sobre los datos de prueba para evaluar el rendimiento del nuevo modelo.



```
from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV

param_grid = {
    'C': [0.001, 0.1, 1, 10, 25, 50, 100, 1000],
    'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
    'kernel': ['rbf', 'sigmoid'],
}

grid = GridSearchCV(SVC(), param_grid, refit=True, verbose=2)
grid.fit(X_train,y_train)
```

Figura 3.6 Uso de GridSearchCV
Fuente. El autor, 2021

La figura 3.6 muestra de forma sencilla la implementación de GridSearchCV para realizar el model tunning del SVM, en donde en primer lugar, se define un diccionario que contiene los posibles valores que tomará el modelo y que serán usados por este método. Acto seguido, se crea una “rejilla” de posibles configuraciones del modelo (usando los valores iniciales) y se realiza un entrenamiento con cada una de éstas, obteniendo así la mejor.

```
from sklearn.metrics import classification_report

grid_predictions = grid.predict(X_test)
print(classification_report(y_test,grid_predictions))
```

Figura 3.7 Obtención de predicciones y del reporte de clasificación
Fuente. El autor, 2021

Una vez obtenida la mejor configuración para el modelo, sólo queda probarlo con el subconjunto de pruebas representado por el 20% de los datos originales, en donde, además, tal y como lo grafica la figura 3.7, se imprime el reporte de clasificación, el cual permitirá visualizar cómo se comportó el modelo realizando predicciones, reporte que se muestra en la figura 3.8.

	precision	recall	f1-score	support
1.0	0.72	0.80	0.76	49
2.0	0.71	0.50	0.59	60
3.0	0.98	1.00	0.99	64
4.0	0.75	0.88	0.81	67
accuracy			0.80	240
macro avg	0.79	0.79	0.79	240
weighted avg	0.80	0.80	0.79	240

Figura 3.8 Reporte de clasificación
Fuente. El autor, 2021

Para la fase de validación del modelo se utilizaron los conjuntos de datos de validación creados durante la etapa de preparación de estos, graficando también el reporte de clasificación para visualizar cómo se comporta el modelo realizando predicciones sobre datos que desconoce y en base a esto realizar ajustes para mejorar su rendimiento, este procedimiento se analizará con mayor detalle en la siguiente sección.

3.6. EVALUACIÓN DEL MODELO

Como ya se mencionó en la etapa anterior, el uso de GridSearchCV fue necesario para evaluar diversas configuraciones hiper paramétricas del modelo (model tuning) y de esta forma obtener la mejor configuración para éste.

Para llevar a cabo la validación de las versiones del modelo entrenado, se utilizaron los conjuntos de datos que no se usaron para entrenar una versión específica. Por ejemplo, si se entrena una versión del modelo con el conjunto de datos generado por ADASYN, los dos conjuntos de datos restantes generados a partir de los otros dos criterios (Custom e Híbrido) fueron usados para “validar” el rendimiento de esta versión del modelo.

Tabla 3.4 Resultados de experimentos iniciales

Método de Oversampling	Precisión	Conjunto de validación 1	Conjunto de validación 2

Adasyn	100%	36%	44%
Custom	**80%**	**65%**	**67%**
Adasyn + Custom	100%	56%	57%

La tabla 3.4 muestra de forma sintetizada los resultados en términos de precisión obtenidos por cada una de las versiones del modelo en fase de pruebas, además, se incluyen los resultados de validación obtenidos para cada una de estas versiones.

Si bien las versiones del modelo creadas a partir de los conjuntos de datos a los que se les aplicó ADASYN (directamente o combinado con el método Custom) presentan un porcentaje de precisión en fase de prueba más elevado con respecto a la versión del modelo creada a partir del conjunto de datos generado con el método Custom, si se analiza el rendimiento de las 3 versiones del modelo, se puede apreciar que la versión creada a partir del conjunto de datos generado con el método Custom refleja un porcentaje de precisión considerablemente mayor realizando predicciones sobre los conjuntos de datos de validación.

Es por esta razón que, se crearon 2 conjuntos de datos más utilizando el método Custom, pero incrementando el número de registros por clase a 250 y 300 para cada uno, dando como resultado, que a mayor número de registros generados por este método el rendimiento del modelo en la fase de validación crece considerablemente como se refleja en la tabla 3.5.

Tabla 3.5 Resultados de experimentos finales

Método de Oversampling	Adasyn	Custom	Adasyn + Custom	Custom 250	Custom 300
Adasyn	100%	36%	36%	33%	36%
Custom	60%	76%	62%	80%	77%
Adasyn + Custom	56%	54%	100%	56%	54%
Custom 250	63%	80%	70%	80%	78%
Custom 300	**63%**	**83%**	**68%**	**81%**	**80%**

En donde la versión del modelo creado a partir de usar el conjunto de datos “Custom 300” registra un porcentaje de precisión del 80% en la fase de prueba, mientras que en la fase de validación se refleja (en promedio) un 74% de precisión, por lo que, si

se analiza el rendimiento individual de esta versión del modelo se tienen los siguientes resultados:

Tabla 3.6 Matriz de confusión de la mejor versión del modelo

	Leve	**Medio**	**Medio Alto**	**Alto**
Leve	39	5	1	4
Medio	14	30	0	16
Medio Alto	0	0	54	0
Alto	1	7	0	59

La tabla 3.6 muestra la matriz de confusión del modelo entrenado con el conjunto Custom 300, en la cual se evidencia un alto rendimiento del modelo realizando predicciones. Esta matriz muestra el número de registros por cada clase (horizontalmente) versus el número de registros que el modelo clasificó como perteneciente a una clase (verticalmente).

Por ejemplo: el número de registros pertenecientes a la clase 1 (nivel de afectación 1 o leve) es 49, de los cuales, el modelo clasificó satisfactoriamente 39 como perteneciente a esta clase, 5 registros con grado de afectación 2 (medio), 1 registro con nivel medio alto y 4 registros con la clase 4, es decir, el nivel de afectación más alto. Para la clase 2 (que cuenta con 60 registros), 14 se clasificaron como de la clase 1, 30 se clasificaron correctamente como de la clase 2, 0 para la clase 3 y 16 para la clase 4, y así para las demás clases.

Tabla 3.7 Reporte de clasificación de la mejor versión del modelo

	Precision	Recall	F1-score	Support
1	72%	80%	76%	49
2	71%	50%	59%	60
3	98%	100%	99%	64
4	75%	88%	81%	67
Accuracy	80%	80%	80%	240
Macro avg	79%	79%	79%	240
Weighted avg	80%	80%	79%	240

Mientras que, la tabla 3.7 presenta el reporte de clasificación de rendimiento del modelo Custom 300, la precisión que tiene el modelo para detectar un registro como perteneciente a una clase en concreto, el “recall” que refleja cuán bien el modelo diferencia unas clases de otras y el “f1 score” que muestra un balance entre las dos métricas previamente mencionadas.

La fila “Accuracy” muestra el comportamiento global del modelo para las tres métricas especificadas anteriormente, en donde se puede apreciar que en las 3, el modelo obtuvo un 80% de exactitud, la cual se contrasta perfectamente con el alto rendimiento mostrado durante la fase de validación del mismo.

CAPÍTULO IV. CONCLUSIONES Y RECOMENDACIONES

4.1. CONCLUSIONES

- Con los datos obtenidos de la revisión sistemática, se pudo identificar que el modelo SVM es uno de los modelos con un rendimiento relativamente alto resolviendo problemas relacionados al agro.
- El conjunto de datos utilizado, se construyó con características fenotípicas de 215 plantas de café ubicadas en CIIDEA, el mismo presentaba un desbalance notable con respecto a la variable objetivo Nivel roya, debido a que el 90% de las plantas no presentaban incidencia de esta enfermedad, por lo que se presume que estas variedades de café son resistentes a la roya o no hay mucha incidencia de esta enfermedad en la zona.
- De los 3 métodos de oversampling utilizados en la construcción de los conjuntos de datos, se pudo determinar que aquellos generados a partir del método Custom permiten entrenar un modelo con un rendimiento global elevado en comparación a aquellas versiones del modelo que fueron creadas usando los conjuntos de datos balanceados con los métodos ADASYN o Híbrido.
- De las 5 versiones del modelo evaluado, la versión entrenada con el conjunto de datos Custom 300 obtuvo un 80% de precisión sobre el subconjunto de prueba, y un 74% de precisión (en promedio) durante la fase de validación, siendo esta la versión del modelo con el más alto rendimiento encontrado.
- Las versiones del modelo entrenado con conjuntos de datos balanceados con el método ADASYN presentan un rendimiento muy bueno en la fase de prueba, lo cual se contrasta con un rendimiento relativamente bajo durante el proceso de validación, mismo que podría ser causado por un overfitting, debido a que los registros generados de manera artificial con este método serían muy parecidos entre sí.

4.2. RECOMENDACIONES

- Al momento de la creación del conjunto de datos es necesaria la inclusión de profesionales relacionados al área agro productiva, ya que esto puede jugar un papel muy importante al momento de realizar una selección de

atributos/variables fenotípicas con mayor relevancia para el modelo, procurando así que éste se acople a todos los posibles escenarios sintomatológicos.

- Durante el transcurso de las sesiones mantenidas con profesionales relacionados a la agricultura, se pudo notar la ausencia de ciertas características dentro del conjunto de datos, la inclusión de variables como el número de ramas de una planta de café o características de las hojas fueron recomendaciones brindadas por parte de los expertos.
- Un correcto análisis de los datos es uno de los puntos más importantes a tener en cuenta cuando se trabaja con modelos de IC, puesto que un ligero cambio en los datos puede ocasionar que los resultados del modelo cambian radicalmente, por lo cual, se recomienda que antes de toda implementación de un modelo se realice un análisis exploratorio de datos adecuado.
- Si bien en esta investigación se utilizó un modelo de IC en concreto basados en los resultados de una revisión sistemática, es recomendable probar un enfoque multi modelo, ya que esto puede ayudar a conocer qué otros modelos (además del utilizado en este trabajo) presentan un rendimiento igual o superior al obtenido.

BIBLIOGRAFÍA

Aditsania, A., Adiwijaya, & Saonard, A. L. (2017). Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm. Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017, 2018-Janua, 533–536. https://doi.org/10.1109/ICSITech.2017.8257170

BBVA (Banco Bilbao Vizcaya Argentaria), 2019. 'Machine learning': ¿qué es y cómo funciona? En línea). Consultado el 16 de diciembre de 2019. Formato HTML. Disponible en https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/

Bce (Banco Nacional Del Ecuador), 2019. La Economía Ecuatoriana Creció Un 0.3% En El Segundo Semestre Del 2019. (En Línea). Consultado El 16 De diciembre De 2019. Formato Html. Disponible En Https://Www.Bce.Fin.Ec/Index.Php/Boletines-de-prensa-archivo/Item/1206-la-econom%C3%Ada-ecuatoriana-creci%C3%B3-03-en-el-segundo-trimestre-de-2019

Brownlee, J. (2019). Classification Accuracy is Not Enough: More Performance Measures You Can Use. https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

Chemura, A., Mutanga, O., Sibanda, M., Chidoko, P. (2018). Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. *Tropical Plant Pathology*, *43*(2), 117–127. https://doi.org/10.1007/s40858-017-0187-8

Daba, G., Helsen, K., Berecha, G., Lievens, B., Debela, A., & Honnay, O. (2019). Seasonal and altitudinal differences in coffee leaf rust epidemics on coffee berry disease-resistant varieties in Southwest Ethiopia. *Tropical Plant Pathology*, *44*(3), 244–250. https://doi.org/10.1007/s40858-018-0271-8

De Melo Virginio Filho, E.; Astorga, C. Prevención y Control de la Roya del Café: Manual de Buenas Prácticas para Técnicos y Facilitadores, 1st ed.; CATIE: Turrialba, Costa Rica, 2015; p. 67.

Ehman, M., Surraco, G., Eckert, K., Garrán, S., Hochmaier, V., Taie, A. (2018). Aplicación de minería de datos sobre un repositorio de variables fito fenológicas de cultivos cıtricos para la extracción de conocimientos. http://sedici.unlp.edu.ar/bitstream/handle/10915/71509/Documento_completo.pdf?sequence=1

ESPAM MFL (Escuela Superior Politécnica Agropecuaria de Manabí Manuel Félix López) 2016. Modelo Educativo. (En línea). Consultado el 16 de diciembre de 2019. Formato PDF. Disponible en http://espam.edu.ec/recursos/sitio/espam/ModeloEducativo2016.pdf

ESPAM MFL. (2020). *COORDINACIÓN GENERAL DE INVESTIGACIÓN*. http://www.espam.edu.ec/web/unidades/investigacion.aspx

Galán, V. 2015. Aplicación De La Metodología Crisp-dm A Un Proyecto De Minería De Datos En El Entorno Universitario. En Línea. Consultado el 10 de agosto de 2020. Formato Pdf. Disponible en https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Guerrero, M. (2017). Rendimientos de Café grano seco en el Ecuador 2017, 13.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Jun-De, C., Huayi, Y., & De-Fu, Z. (2020). A self-adaptive classification method for plant disease detection using GMDH-Logistic model. Sustainable Computing: Informatics and Systems, 100415. https://doi.org/10.1016/j.suscom.2020.100415

Kosaraju, B., Sannasi, S., Mishra, M. K., Subramani, D., & Bychappa, M. (2017). Assessment of genetic diversity of coffee leaf rust pathogen Hemileia vastatrix using SRAP markers. *Journal of Phytopathology*, *165*(7–8), 486–493.

López, H. M. (2015). Fenotipado de plantas - MasScience. https://www.masscience.com/2015/11/30/fenotipado-de-plantas/

Molina A. 2019. Guía de plagas y enfermedades comunes del café. (En línea). Consultado el 16 de diciembre de 2019. Formato HTML. Disponible en https://www.perfectdailygrind.com/2019/01/guia-de-plagas-y-enfermedades-comunes-del-cafe/

Moreno, B., Muñoz, M., Cuellar, J., Domancic, S., & Villanueva, J. (2018). Revisiones Sistemáticas: definición y nociones básicas. *Revista Clínica de Periodoncia, Implantología y Rehabilitación Oral*, *11*(3), 184–186. https://doi.org/10.4067/s0719-01072018000300184

Shaikh, R. (2018). Feature Selection Techniques in Machine Learning with Python | by Raheel Shaikh | Towards Data Science.

https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Vallalta, J. (2019). *CRISP-DM: una metodología para minería de datos en salud - health data miner.com*. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Villena, J. (2016). *CRISP-DM: La metodología para poner orden en los proyectos - Sngular*. https://www.sngular.com/es/data-science-crisp-dm-metodologia/

ANEXOS

ANEXO 1. TABLA DE REVISIÓN BIBLIOGRÁFICA

Año	Tema	Autoress
2019	Evaluación De La Incidencia De Problemas Fitosanitarios Y Productividad Del Café Robusta (Coffea Canephora Pierre) En Cinco Densidades De Siembra En El Cantón Caluma, Provincia Bolívar.	Marco Antonio López
2019	Evaluación Morfo – Agronómica Del Híbrido De Café Robusta (Coffea Canephora Pierre) En Cinco Densidades Poblacionales Mediante La Poda De Agobio, En La Granja El Triunfo Cantón Caluma	Clariza Maribel García Lombeida
2017	Revealing The Diversity Of Introduced Coffea Canephora Germplasm In Ecuador: Towards A National Strategy To Improve Robusta	Fabien De Bellis, Thierry Leroy, Luis Plaza, Hilton Guerrero, Cristian Subia, Darío Calderón, Fabián Fernández, Iván Garzón, Diana Lopez, Danilo Vera
2016	Caracterización Física y Organoléptica de Árboles Cabezas de Clon de Genotipos de Café Robusta de Alta Productividad para la Provincia de Los Ríos	Dennise Xiomara Suarez Prieto
2018	Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties	Louis Kouadio, Ravinesh C. Deo, Vivekananda Byrareddy, Jan F. Adamowski, Shahbaz Mushtaq, Van Phuong Nguyen
2020	Few-Shot Learning approach for plant disease classification using images taken in the field	David Argüesoab, Artzai Picona, Unai Irustab, Alfonso Medelaa, Miguel G San-Emeterioa, Arantza Bereciartuaa, Aitor Alvarez-Gilaa
2019	A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning	Hafiz Tayyab Raufa, Basharat Ali Saleemb, M. Ikram Ullah Lalia, Muhammad Attique Khanc, Muhammad Sharifd, Syed Ahmad Chan Bukharie
2015	CARACTERIZACIÓN FENOTÍPICA DEL GERMOPLASMA DE Coffea canephora Pierre BASE PARA SU MEJORAMIENTO EN ECUADOR	Luis Fernando Plaza Avellán, Rey Gaston Loor Solórzano, Hilton Ecuador Guerrero Castillo, Luis Alberto Duicela Guambi
2019	RoCoLe: A robusta coffee leaf images dataset for evaluation of machine learning based methods in plant diseases recognition	Jorge Parraga Alava, Kevin Cusme, Angélica Loor, Esneider Santandera
2018	Machine Learning in Agriculture: A Review	Konstantinos G Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, Dionysis Bochtis
2018	A Review of Neural Networks in Plant Disease Detection using Hyperspectral Data	Kamlesh Golhani, Siva K Balasundram, Ganesan Vadamalai, Biswajeet Pradhan
2016	Identifying Multiple Plant Diseases Using Digital Image Processing	Jayme Garcia Arnal Barbedo, Luciano Koenigkan, Thiago Teixeira Santos
2017	Identification of plant disease infection using soft-computing: Application to modern botany	Ehsan Kiani, Tofik Mamedov
2020	A Guideline for Building Large Coffee Rust Samples Applying Machine Learning Methods	Jhonn Pablo Rodríguez, Edwar Javier Girón, David Camilo Corrales, Juan Carlos Corrales
2020	Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases	Lucas Eduardo de Oliveira Aparecido, Glauco de Souza Rolim, Jose Reinaldo da Silva Cabral De Moraes, Cicero Teixeira Silva Costa, Paulo Sergio de Souza
2015	An Empirical Multi-classifier for Coffee Rust Detection in Colombian Crops	David Camilo Corrales, Apolinar Figueroa, Agapito Ledezma, Juan Carlos Corrales
2018	Towards an Alert System for Coffee Diseases and Pests in a Smart Farming Approach Based on Semi-supervised Learning and Graph Similarity	Emmanuel Lasso, Juan Carlos Corrales
2017	Separability of coffee leaf rust infection levels with machine learning	Abel Chemura, Onisimo Mutanga,

	methods at Sentinel-2 MSI spectral resolutions	Timothy Dube
2016	Two-Level Classifier Ensembles for Coffee Rust Estimation in Colombian Crops	David Camilo Corrales, Apolinar Figueroa Casas, Agapito Ismael Ledezma Espino, Juan Carlos Corrales

De la tabla anterior sólo se consideraron los trabajos que cumplieron los siguientes criterios: utilizar técnicas de IC para resolver problemas relacionados con el agro, trabajos donde se haga alguna clase de caracterización vegetativa y trabajos de los últimos 5 años. Mientras que, para los criterios de exclusión, se consideran trabajos no relacionados a la problemática, investigaciones longevas o propuestas no relevantes o poco concluyentes.

ANEXO 2. DATOS FENOTÍPICOS PROPORCIONADOS POR LA CARRERA DE AGRÍCOLA

Clones	Código	Plantas vivas	Altura de planta	Numero de ramas	Producción	Si esta en produccion	Presencia de plagas	Si hay plagas	Presencia de roya	Si hay roya	Si hay otras enfermedades	Incidencia de enfermedades distintas de roya	Produccion (gramos/planta)	Peso de 50 frutos
	SICA-01_08	8	108	18	NO	0	SI	1	NO	0	NO	0	0	NO
	SICA-01_10	10	128	23	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-01_11	1	130	26	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-01_12	2	163	31	SI	2	SI	3	NO	0	NO	0	0	VERDE
	SICA-01_13	3	120	26	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-01_15	5	155	30	SI	3	SI	2	NO	0	NO	0	52	52
	SICA-01_18	8	140	35	SI	2	SI	2	NO	0	SI	2	0	VERDE
	SICA-01_19	9	123	27	SI	3	SI	2	NO	0	NO	0	0	VERDE
SICA-02	SICA-02_01	1	108	19	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-02_02	2	109	29	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-02_03	3	98	26	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-02_05	5	110	26	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-02_06	6	108	15	NO	0	SI	1	NO	0	NO	0	0	NO

	SICA-02_07	7	112	27	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-02_08	8	110	31	NO	0	SI	1	NO	0	NO	0	0	NO
	SICA-02_09	9	89	10	NO	0	SI	1	NO	0	NO	0	0	NO
	SICA-02_10	10	128	40	SI	2	SI	2	NO	4	NO	0	0	VERDE
	SICA-02_11	1	120	32	SI	3	SI	2	NO	0	NO	0	0	VERDE
	SICA-02_12	2	130	32	SI	1	SI	4	NO	0	NO	0	0	VERDE
	SICA-02_13	3	145	40	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-02_14	4	120	34	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-02_15	5	135	36	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-02_16	6	136	28	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-02_17	7	110	31	SI	1	SI	2	NO	0	SI	1	0	VERDE
	SICA-02_18	8	115	35	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-02_19	9	146	32	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-02_20	10	170	28	SI	3	SI	2	NO	0	SI	1	0	VERDE
SICA-03	SICA-03_01	1	107	34	SI	4	SI	2	NO	0	NO	0	292	47
	SICA-03_02	2	140	26	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-03_04	4	135	30	NO	0	SI	3	NO	0	NO	0	0	NO

	SICA-03_05	5	92	32	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-03_06	6	124	29	SI	1	SI	2	NO	0	NO	0	0	VERDE
SICA-04	SICA-04_01	1	80	18	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-04_03	3	119	23	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-04_04	4	143	36	SI	3	SI	3	NO	0	NO	0	0	VERDE
	SICA-04_05	5	101	13	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-04_06	6	138	30	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-04_11	1	175	40	SI	4	SI	2	NO	0	SI	1	0	VERDE
	SICA-04_12	2	169	36	SI	3	SI	3	NO	0	SI	1	0	VERDE
	SICA-04_13	3	95	23	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-04_14	4	88	22	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-04_15	5	108	32	SI	1	SI	3	NO	0	NO	0	0	VERDE
	SICA-04_16	6	117	27	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-04_17	7	130	35	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-04_20	10	84	28	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-05_02	2	102	23	SI	1	SI	3	NO	0	SI	1	0	VERDE
	SICA-05_03	3	128	23	SI	1	SI	2	NO	0	SI	1	0	VERDE

	SICA-05_04	4	84	14	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-05_05	5	89	18	SI	1	SI	3	NO	0	SI	1	0	VERDE
	SICA-05_06	6	105	37	SI	2	SI	2	NO	0	NO	0	137	38
	SICA-05_07	7	157	42	SI	2	SI	2	NO	0	NO	0	44	44
	SICA-05_08	8	165	34	SI	3	SI	3	NO	0	SI	1	0	VERDE
	SICA-05_09	9	128	33	SI	2	SI	3	NO	0	SI	1	102	36
	SICA-05_10	10	118	28	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-05_12	2	142	49	SI	3	SI	4	NO	0	NO	0	0	VERDE
	SICA-05_13	3	120	32	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-05_14	4	111	22	NO	0	SI	3	NO	0	NO	0	0	NO
	SICA-05_17	7	134	28	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-05_19	9	120	28	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-05_20	10	114	32	NO	0	SI	2	NO	0	NO	0	0	NO
SICA-06	SICA-06_01	1	104	35	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-06_03	3	138	17	SI	1	SI	3	NO	0	SI	1	0	VERDE
	SICA-06_05	5	155	37	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-06_06	6	97	18	SI	1	SI	2	NO	0	NO	0	0	VERDE

	SICA-06_07	7	133	32	SI	3	SI	3	NO	0	NO	0	0	VERDE
SICA-07	SICA-07_01	1	133	24	SI	3	SI	3	NO	0	SI	1	256	41
	SICA-07_02	2	130	22	SI	4	SI	2	NO	0	NO	0	138	49
	SICA-07_03	3	128	19	SI	4	SI	3	NO	0	NO	0	431	44
	SICA-07_04	4	165	27	SI	3	SI	2	NO	0	NO	0	927	49
	SICA-07_05	5	148	27	SI	3	SI	3	NO	0	SI	1	115	55
	SICA-07_06	6	157	35	SI	3	SI	3	SI	1	NO	0	114	50
	SICA-07_07	7	133	22	SI	2	SI	2	SI	1	NO	0	150	41
	SICA-07_08	8	129	30	SI	4	SI	2	SI	2	NO	0	756	51
	SICA-07_09	9	135	26	SI	5	SI	3	SI	1	SI	1	734	46
	SICA-07_10	10	158	36	SI	4	SI	3	SI	2	SI	2	422	41
	SICA-07_11	1	128	34	SI	3	SI	2	NO	0	SI	1	0	VERDE
	SICA-07_12	2	116	22	SI	1	SI	2	NO	0	NO	0	0	VERDE
	SICA-07_13	3	134	21	NO	0	SI	2	NO	0	NO	0	0	NO
	SICA-07_18	8	140	30	SI	2	SI	2	NO	0	NO	0	0	VERDE
	SICA-07_20	10	145	44	SI	3	SI	3	NO	0	NO	0	0	VERDE
	NP-2024-P2_03	3	63	20	NO	0	SI	2	NO	0	NO	0	0	NO

	NP-2024-P2_04	4	107	30	SI	3	SI	3	NO	0	SI	2	0	VERDE
	NP-2024-P2_05	5	118	38	SI	3	SI	2	NO	0	NO	0	0	VERDE
	NP-2024-P2_07	7	123	33	NO	0	SI	3	NO	0	NO	0	0	NO
	NP-2024-P2_08	8	104	26	NO	0	SI	3	NO	0	NO	0	0	NO
	NP-2024-P2_09	9	108	18	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-2024-P2_10	10	140	24	SI	2	SI	4	NO	0	NO	0	0	VERDE
	NP-2024-P2_11	1	113	20	SI	3	SI	3	SI	3	NO	0	0	VERDE
	NP-2024-P2_12	2	95	19	NO	0	SI	2	NO	0	NO	0	0	NO
NP-3056-P11	NP-3056-P11_01	1	96	21	SI	1	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_02	2	85	22	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-3056-P11_03	3	130	30	SI	2	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_04	4	116	38	SI	2	SI	2	NO	0	NO	0	0	VERDE

	NP-3056-P11_05	5	75	9	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-3056-P11_06	6	124	31	SI	3	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_07	7	133	35	SI	2	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_08	8	94	24	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-3056-P11_09	9	100	26	SI	1	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_10	10	95	24	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-3056-P11_11	1	115	28	SI	2	SI	2	NO	0	NO	0	0	VERDE
	NP-3056-P11_12	2	120	27	SI	1	SI	3	NO	0	NO	0	42	42
	NP-3056-P11_13	3	72	12	NO	0	SI	3	NO	0	NO	0	0	NO
	NP-3056-P11_14	4	116	28	SI	1	SI	4	NO	0	NO	0	0	VERDE
	NP-3056-P11_15	5	126	25	NO	0	SI	2	NO	0	NO	0	0	NO
	NP-3056-P11_17	7	110	23	NO	0	SI	3	NO	0	NO	0	0	NO

	NP-3056-P11_19	9	87	19	SI	1	SI	3	NO	0	NO	0	0	VERDE
HC-11	HC-11_01	1	135	24	SI	3	SI	2	NO	0	SI	2	0	VERDE
	HC-11_02	2	165	30	SI	4	SI	2	NO	0	NO	0	0	VERDE
	HC-11_03	3	140	24	SI	4	SI	2	NO	0	NO	0	45	45
	HC-11_04	4	175	28	SI	1	SI	3	NO	0	NO	0	0	VERDE
	HC-11_05	5	142	33	SI	5	SI	2	NO	0	NO	0	0	VERDE
	HC-11_06	6	160	40	SI	5	SI	2	NO	0	NO	0	279	59
	HC-11_07	7	156	30	SI	5	SI	3	NO	0	NO	0	133	59
	HC-11_08	8	175	27	SI	5	SI	3	NO	0	NO	0	0	VERDE
	HC-11_09	9	156	49	SI	5	SI	3	NO	0	NO	0	0	VERDE
	HC-11_11	1	120	24	SI	4	SI	3	NO	0	NO	0	0	VERDE
	HC-11_12	2	145	28	SI	5	SI	2	NO	0	NO	0	0	VERDE
	HC-11_13	3	140	28	SI	4	SI	2	NO	0	NO	0	0	VERDE
	HC-11_14	4	126	28	SI	4	NO	0	NO	0	NO	0	0	VERDE
	HC-11_16	6	126	28	SI	4	SI	2	NO	0	NO	0	0	VERDE
	HC-11_18	8	130	28	SI	4	SI	2	NO	0	NO	0	0	VERDE

	HC-11_20	10	143	26	SI	3	SI	2	NO	0	NO	0	0	VERDE
HC-12	HC-12_01	1	105	20	SI	2	SI	2	NO	0	NO	0	0	VERDE
	HC-12_02	2	130	25	SI	3	SI	2	NO	0	NO	0	0	VERDE
	HC-12_03	3	135	23	SI	3	SI	2	NO	0	SI	2	382	51
	HC-12_04	4	130	21	SI	3	SI	3	NO	0	SI	1	541	49
	HC-12_05	5	175	28	SI	3	SI	2	SI	2	SI	1	153	57
	HC-12_06	6	150	22	SI	4	SI	2	NO	0	SI	1	183	51
	HC-12_07	7	170	45	SI	3	SI	3	NO	0	NO	0	0	VERDE
	HC-12_08	8	178	30	SI	3	SI	3	NO	0	NO	0	0	VERDE
	HC-12_09	9	148	24	SI	4	SI	2	NO	0	SI	1	421	34
	HC-12_10	10	172	36	SI	4	SI	2	NO	0	NO	0	0	VERDE
	HC-12_11	1	140	28	SI	4	SI	2	NO	0	NO	0	249	32
	HC-12_12	2	158	33	SI	4	SI	3	NO	0	NO	0	393	44
	HC-12_13	3	158	30	SI	4	SI	2	NO	0	NO	0	0	VERDE
	HC-12_14	4	165	31	SI	3	SI	3	NO	0	NO	0	124	41
	HC-12_15	5	165	41	SI	4	SI	2	NO	0	NO	0	230	50
	HC-12_16	6	160	34	SI	4	SI	3	NO	0	NO	0	323	44

	HC-12_17	7	160	24	SI	3	SI	2	NO	0	NO	0	0	VERDE
	HC-12_18	8	145	39	SI	3	SI	2	NO	0	NO	0	0	VERDE
	HC-12_19	9	135	26	SI	3	SI	2	NO	0	NO	0	0	VERDE
	HC-12_20	10	150	28	SI	4	SI	3	NO	0	NO	0	0	VERDE
LG-S-01	LG-S-01_01	1	120	23	SI	2	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_02	2	107	26	SI	2	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_03	3	122	32	SI	3	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_04	4	138	30	SI	2	SI	3	NO	0	SI	2	0	VERDE
	LG-S-01_05	5	158	36	SI	2	SI	3	NO	0	SI	1	0	VERDE
	LG-S-01_06	6	158	37	SI	3	SI	2	NO	0	NO	0	135	52
	LG-S-01_07	7	114	28	SI	2	SI	3	NO	0	NO	0	0	VERDE
	LG-S-01_08	8	135	27	SI	4	SI	3	NO	0	NO	0	0	VERDE
	LG-S-01_10	10	120	26	SI	2	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_11	1	108	25	SI	1	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_12	2	140	40	SI	4	SI	3	NO	0	NO	0	0	VERDE
	LG-S-01_13	3	143	39	SI	4	SI	3	NO	0	NO	0	0	VERDE

	LG-S-01_14	4	160	41	SI	3	SI	3	NO	0	NO	0	0	VERDE
	LG-S-01_15	5	138	26	SI	3	SI	2	NO	0	SI	1	0	VERDE
	LG-S-01_16	6	120	28	SI	3	SI	2	NO	0	NO	0	0	VERDE
	LG-S-01_17	7	156	28	SI	3	SI	3	NO	0	NO	0	109	56
	LG-S-01_18	8	118	27	SI	4	SI	3	NO	0	NO	0	283	54
	LG-S-01_19	9	156	27	SI	3	SI	2	NO	0	NO	0	59	59
	LG-S-01_20	10	173	28	SI	3	SI	3	NO	0	NO	0	0	VERDE
	LG-S02_04	4	142	34	SI	3	SI	2	NO	0	NO	0	367	48
	LG-S02_05	5	157	24	SI	1	SI	2	SI	3	NO	0	90	57
	LG-S02_06	6	130	30	SI	1	SI	2	SI	2	NO	0	0	VERDE
	LG-S02_07	7	138	26	SI	4	SI	3	NO	0	SI	2	658	68
	LG-S02_09	9	118	24	SI	4	SI	2	NO	0	NO	0	413	49
	LG-S02_11	1	150	35	SI	4	SI	3	NO	0	NO	0	0	VERDE
	LG-S02_12	2	145	29	SI	4	SI	3	NO	0	NO	0	908	54
	LG-S02_13	3	145	28	SI	4	SI	4	SI	1	SI	1	410	57
	LG-S02_15	5	145	29	SI	3	SI	2	SI	1	NO	0	655	42
	LG-S02_16	6	165	28	SI	3	SI	2	SI	4	SI	1	541	50
	LG-S02_17	7	140	34	SI	3	SI	2	SI	2	SI	1	392	48

	LG-S02_18	8	110	20	SI	2	SI	2	NO	0	NO	0	125	52
	LG-S02_19	9	110	26	SI	3	SI	3	NO	0	NO	0	388	47
	COF-01_02	2	135	23	SI	3	SI	2	SI	2	NO	0	364	63
	COF-01_03	3	132	28	SI	4	SI	2	NO	0	SI	1	0	VERDE
	COF-01_04	4	126	26	SI	3	SI	3	NO	0	NO	0	483	49
	COF-01_05	5	110	27	SI	4	SI	2	SI	1	SI	2	0	VERDE
	COF-01_06	6	112	30	SI	3	SI	3	SI	4	NO	0	266	52
	COF-01_07	7	130	24	SI	4	SI	3	NO	0	NO	0	738	51
	COF-01_08	8	124	20	SI	3	SI	3	NO	0	NO	0	277	42
	COF-01_09	9	121	36	SI	3	SI	3	SI	1	SI	1	0	VERDE
	COF-01_10	10	108	27	SI	3	SI	2	NO	0	NO	0	0	VERDE
	COF-01_11	1	155	19	SI	2	SI	3	NO	0	NO	0	0	VERDE
	COF-01_12	2	140	32	SI	2	SI	3	NO	0	NO	0	0	VERDE
	COF-01_13	3	176	28	SI	4	SI	4	NO	0	NO	0	0	VERDE
	COF-01_14	4	171	43	SI	2	SI	2	NO	0	SI	1	45	45
	COF-01_15	5	118	20	SI	2	SI	1	NO	0	SI	1	38	38
	COF-01_16	6	145	31	SI	5	SI	4	SI	1	SI	2	1765	38

	COF-01_17	7	170	25	SI	1	SI	2	SI	1	SI	2	0	VERDE
	COF-01_18	8	150	24	SI	3	SI	2	NO	0	SI	2	63	36
	COF-01_20	10	137	17	SI	3	SI	2	SI	1	SI	2	82	51
COF-02	COF-02_01	1	71	12	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-02_02	2	74	13	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-02_03	3	94	15	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-02_04	4	90	22	NO	0	SI	3	NO	0	NO	0	0	NO
	COF-02_05	5	80	20	NO	0	SI	3	NO	0	NO	0	0	NO
	COF-02_08	8	85	18	NO	0	SI	3	NO	0	NO	0	0	NO
	COF-02_11	1	120	21	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-02_12	2	95	28	SI	3	SI	2	NO	0	SI	1	0	VERDE
	COF-02_13	3	80	13	NO	0	SI	3	NO	0	NO	0	0	NO
	COF-06_06	6	87	18	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-06_07	7	92	20	SI	2	SI	3	NO	0	NO	0	0	VERDE
	COF-06_09	9	84	22	NO	0	SI	3	NO	0	NO	0	0	NO
	COF-06_10	10	82	24	NO	0	SI	2	NO	0	NO	0	0	NO
	COF-06_12	2	104	36	SI	4	SI	2	NO	0	NO	0	0	VERDE
	COF-06_18	8	96	32	NO	0	SI	2	NO	0	NO	0	0	VERDE

JMR-8-P15	JMR-8-P15_01	1	95	29	SI	1	SI	2	NO	0	NO	0	0	VERDE
	JMR-8-P15_02	2	116	29	NO	0	SI	2	NO	0	NO	0	0	NO
	JMR-8-P15_03	3	136	24	NO	0	SI	2	NO	0	NO	0	0	NO
	JMR-8-P15_04	4	120	36	SI	1	SI	2	NO	0	NO	0	0	VERDE
	JMR-8-P15_05	5	144	33	SI	1	SI	3	NO	0	NO	0	0	VERDE
	JMR-8-P15_07	7	105	34	SI	2	SI	4	NO	0	NO	0	0	VERDE
	JMR-8-P15_09	9	135	26	NO	0	SI	2	NO	0	NO	0	0	NO
	JMR-8-P15_10	10	105	20	SI	1	SI	2	NO	0	NO	0	0	VERDE

ANEXO 3. MÉTODOS DE OVERSAMPLING UTILIZADOS

ANEXO 3-A. APLICACIÓN DEL CRITERIO ADASYN

```python
from imblearn.over_sampling import ADASYN

def apply_adasyn_oversampling(data):
  X = data.drop('Nivel_roya', axis=1)
  y = data['Nivel_roya']
  ada = ADASYN(random_state=0, n_neighbors=1)
  x_ada, y_ada = ada.fit_resample(X, y)
  result = pd.DataFrame(x_ada, columns=X.columns)
  result['Nivel_roya'] = pd.Series(y_ada)
  print(result['Nivel_roya'].value_counts())
  return result

balanced_data = apply_adasyn_oversampling(data)
```

ANEXO 3-B. DEFINICIÓN DEL MÉTODO CUSTOM

```python
import math
import random
import numpy as np
import pandas as pd


def get_upsampled_class(df, rows_per_class, class_id, col_indexes, class_col_name='Nivel_roya'):
    df = df[df[class_col_name] == class_id][df.columns]
    means = np.mean(df)
    df['Produccion_gramos'].replace(0, means['Produccion_gramos'], inplace=True)
    means = np.mean(df)
    std = np.std(df)
    response = pd.DataFrame()
    for x in range(0, rows_per_class):
        for n in range(0, df.shape[1]):
            if n in col_indexes:
                response.at[x, n] = math.floor(random.uniform(means[n]-std[n], means[n]+ std[n]))
            else:
                response.at[x, n] = random.sample(df.iloc[:,n].values.tolist(), 1)[0]
    response.columns = df.columns
    return response
```

ANEXO 3-C. APLICACIÓN DEL CRITERIO CUSTOM

```python
def apply_custom_oversampling(data, col_indexes=[], new_rows_per_class=[]):
  data = data[data['Nivel_roya'] != 0]
  class1 = get_upsampled_class(data, new_rows_per_class[0], 1, col_indexes)
  class2 = get_upsampled_class(data, new_rows_per_class[1], 2, col_indexes)
  class3 = get_upsampled_class(data, new_rows_per_class[2], 3, col_indexes)
  class4 = get_upsampled_class(data, new_rows_per_class[3], 4, col_indexes)
  balanced_data = pd.concat([data, class1, class2, class3, class4])
  print(balanced_data['Nivel_roya'].value_counts())
  return balanced_data

balanced_data = apply_custom_oversampling(data)
```

ANEXO 3-D. APLICACIÓN DEL MÉTODO HÍBRIDO (ADASYN+CUSTOM)

```python
def apply_adasyn_plus_custom_oversampling(data, col_indexes=[], new_rows_per_class=[]):
  data = data[data['Nivel_roya'] != 0]
  ada = apply_adasyn_oversampling(data)
  cus = apply_custom_oversampling(ada, col_indexes, new_rows_per_class)
  balanced_data = cus
  print(balanced_data['Nivel_roya'].value_counts())
  return balanced_data

balanced_data = apply_adasyn_plus_custom_oversampling(data)
```