REPÚBLICA DEL ECUADOR

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO

PROYECTO DE INVESTIGACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE:

MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

TÍTULO DEL PROYECTO:

DISEÑO DE UN MODELO PREDICTIVO BASADO EN ALGORITMOS DE MACHINE LEARNING PARA LA ESTIMACIÓN DEL PESO DE RACIMOS DE BANANO, CASO DE ESTUDIO HACIENDA BANANERA EN ECUADOR.

TUTOR

ORTIZ MATA JHONNY DARWIN

AUTOR

MUÑOZ TORRES PEDRO SANTIAGO

MILAGRO, MARZO 2023

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO

Milagro, 27 de septiembre, 2022

CERTIFICACIÓN DE ACEPTACIÓN DEL TUTOR

En calidad de Tutor del Proyecto de Investigación, nombrado por el Comité Académico del Programa de Maestría en Tecnologías de la Información de la Universidad Estatal de Milagro.

CERTIFICO

Que he analizado el Proyecto de Investigación con el tema **DISEÑO DE UN MODELO PREDICTIVO BASADO EN ALGORITMOS DE MACHINE LEARNING PARA LA ESTIMACIÓN DEL PESO DE RACIMOS DE BANANO, CASO DE ESTUDIO HACIENDA BANANERA EN ECUADOR.**, elaborado por **PEDRO SANTIAGO MUÑOZ TORRES**, el mismo que reúne las condiciones y requisitos previos para ser defendido ante el tribunal examinador, para optar por el título de **MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN**.

Firmado electrónicamente por:
JHONNY
DARWIN ORTIZ

JHONNY DARWIN ORTIZ MATA
C.I: 0927159111

DECLARACIÓN DE AUTORÍA DE LA INVESTIGACIÓN

El autor de esta investigación declara ante el Comité Académico del Programa de Maestría en Tecnologías de Información de la Universidad Estatal de Milagro, que el trabajo presentado es de mi propia autoría, no contiene material escrito por otra persona, salvo el que esta referenciado debidamente en el texto; parte del presente documento o en su totalidad no ha sido aceptado para el otorgamiento de cualquier otro Título de una institución nacional o extranjera.

Milagro, a los 09 días del mes de Marzo de 2023

Firmado electrónicamente por:
PEDRO SANTIAGO
MUNOZ TORRES

MUÑOZ TORRES PEDRO SANTIAGO
C.I: 0926300088

UNEMI
UNIVERSIDAD ESTATAL DE MILAGRO

VICERRECTORADO DE INVESTIGACIÓN Y POSGRADO

CERTIFICACIÓN DE LA DEFENSA

El TRIBUNAL CALIFICADOR previo a la obtención del título de **MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN**, presentado por **ING. MUÑOZ TORRES PEDRO SANTIAGO**, otorga al presente proyecto de investigación denominado "DISEÑO DE UN MODELO PREDICTIVO BASADO EN ALGORITMOS DE MACHINE LEARNING PARA LA ESTIMACIÓN DEL PESO DE RACIMOS DE BANANO, CASO DE ESTUDIO HACIENDA BANANERA EN ECUADOR", las siguientes calificaciones:

TRABAJO DE TITULACION	58.33
DEFENSA ORAL	35.33
PROMEDIO	**93.67**
EQUIVALENTE	**Muy Bueno**

Firmado electrónicamente por:
KERLY VANESSA PALACIOS ZAMORA

PALACIOS ZAMORA KERLY VANESSA

PRESIDENTE (A) DEL TRIBUNAL

Firmado electrónicamente por:
CHRISTIAN ALBERTO BERMEO VALENCIA

BERMEO VALENCIA CRISTHIAN ALBERTO

MIEMBRO DEL TRIBUNAL

Firmado electrónicamente por:
OSCAR XAVIER BERMEO ALMEIDA

BERMEO ALMEIDA OSCAR JAVIER

SECRETARIO (A) DEL TRIBUNAL

DEDICATORIA

Este trabajo se lo dedico a mis Padres Isabel Magali Torres Torres y Pedro Amable Muñoz Flor por su esfuerzo y amor brindado apoyos incondicionales en el camino a cumplir mis metas. Porque a pesar de los buenos y malos momentos siempre me supieron apoyar sin importar las circunstancias, siempre pensé en que el mejor regalo que se le puede dar es ver a su hijo convertirse en un profesional competente.

También está dirigido a mi Abuelito Antonio Torres Lascano quien supo apoyarme siempre que lo necesitaba, y aunque mi Abuelo se fue de este mundo hace mucho, siento que cada logro va dedicado para él.

.

AGRADECIMIENTO

Agradezco a Dios, en primer lugar, por ser guía de mi vida y por bendecirla mediante este logro que pasando por momentos muy difíciles de pandemia jamás perdí la esperanza y la fe en el para poder pasar esos duros momentos y alcanzar mi formación profesional.

A mis Padres Isabel Torres y Pedro Muñoz, por el amor que me han brindado cada día de mi vida y ser mi mayor ejemplo de trabajo duro y perseverancia, por darme amor, apoyo, motivación y mostrarme lo importante que es la disciplina y firmeza en la vida siendo un pilar fundamental de mi vida.

A mis hermanos porque de una u otra forma me brindaron su respaldo y su aportación de cosas buenas en mi vida, pues mis logros también son suyos.

A mi Esposa Martha Pilco Paredes, por motivarme a cumplir mis metas, ser siempre mi apoyo incondicional y compartir mis anhelos.

UNEMI
UNIVERSIDAD ESTATAL DE MILAGRO

CESIÓN DE DERECHOS DE AUTOR

Doctor
ING. FABRICIO GUEVARA VIEJÓ, PhD

Presente

Mediante el presente documento, libre y voluntariamente procedo a hacer entrega de la Cesión de Derecho del Autor al Trabajo realizado como requisito previo para la obtención de mi Título de Cuarto Nivel, cuyo tema **DISEÑO DE UN MODELO PREDICTIVO BASADO EN ALGORITMOS DE MACHINE LEARNING PARA LA ESTIMACIÓN DEL PESO DE RACIMOS DE BANANO, CASO DE ESTUDIO HACIENDA BANANERA EN ECUADOR**., elaborado por ING. PEDRO SANTIAGO MUÑOZ TORRES y que corresponde al Vicerrectorado de Investigación y Posgrado.

Milagro, 09 de Marzo de 2023

Firmado electrónicamente por:
PEDRO SANTIAGO
MUNOZ TORRES

ING. MUÑOZ TORRES PEDRO SANTIAGO
C.I: 092630008-8

ÍNDICE

Índice de tablas **xi**

Índice de figuras **xii**

Resúmen **xiv**

Abstract **xvi**

INTRODUCCION **1**

1 Planteamiento del problema **3**

1.1 Planteamiento del problema....3

1.1.1 Pregunta Principal de Investigación....4

1.1.2 Preguntas Secundarias de Investigación....4

1.2 Objetivos....5

1.3 Alcance....6

1.3.1 Entre la Limitaciones mencionamos las Siguientes:....6

1.4 Estado del arte....7

2 Metodología **16**

2.1 Tipo y diseño de investigación....16

2.2 La población y la muestra....17

2.2.1 Características de la población....17

2.3 Técnicas de Observación e Instrumentos de Colecta de Datos.... 18

2.3.1 Entrevista....18

2.3.2 El análisis documental....18

2.4 Operacionalización de las Variables....19

2.5 Instrumentos....19

2.6 Procedimientos....21

2.6.1 Aprendizaje supervisado....21

2.6.2 Random Forest 22
2.6.3 Xgboost 23
2.6.4 Software R 24
2.6.5 Recopilación de los datos 25
2.6.5.1 Fase de preprocesamiento de los datos 25
2.6.5.2 Fase de entrenamiento 27
2.6.5.3 Fase de análisis, evaluación 27
2.6.6 Métricas de Evaluación 28
2.6.6.1 Error cuadrático medio (MSE) 28
2.6.6.2 Error cuadrático medio (RMSE) 29
2.6.6.3 Error absoluto medio (MAE) 29
2.6.6.4 Coeficiente de determinación R^2 29
2.6.6.5 Error porcentual absoluto medio (MAPE) 30

3 Propuesta de solución 32
3.1 Preparación y manejo de variables 32
3.2 Correlación 33
3.3 Análisis y validación de los datos 39
3.4 Modelamiento y Métricas 42
3.4.1 Random Forest 42
3.4.2 Xgboost 46
3.5 Dependencia de Variables 51

Conclusiones y trabajo futuro 53

Recomendaciones 54

Bibliografía general 55

Apéndice 58
Apéndice A: Instalación de R Studio 58
Apéndice B: Carga y Ajuste de datos R Studio 59
Apéndice C: Análisis y validación de datos 65
Apéndice D: Modelamiento y métricas 67

Apéndice E: Preguntas... 72

ÍNDICE DE TABLAS

Tabla 1.1 Principales conceptos usado en ML 13
Tabla 2.1 Matriz de operacionalización de variables 19
Tabla 2.2 Campo de los datos de parámetros de producción. 20
Tabla 2.3 Tipo de variable 26
Tabla 3.1 Variables de la Base de Datos consulta racimos.xlsx 31
Tabla 3.2 Correlación de Variable del dataframe ConsultaRacimos.csv 32
Tabla 3.3 Media Mediana IQR Cuantiles 38
Tabla 3.4 Media - Mediana- IQR- Cuantiles 40
Tabla 3.5 Variables escogidas 41
Tabla 3.6 Importancia de variables en el modelo 43
Tabla 3.7 Métricas RF-NTREE = 100 43
Tabla 3.8 Métricas RF-NTREE = 200 44
Tabla 3.9 Métricas para xgboost-nrounds=50 48
Tabla 3.10 Métricas para xgboost-nrounds=150 48
Tabla 3.11 Métricas para el modelo de random forest 48
Tabla 3.12 Métricas para el modelo XGBoost 48

ÍNDICE DE FIGURAS

Figuras 2.1 Jerarquía de Machine Learning 21
Figuras 2.2 Tipos de Algoritmos Supervisados 22
Figuras 2.3 Algoritmo Random Forest 23
Figuras 2.4 Realización del Modelo Predictivo 28

Figuras 3.1 Correlación peso – manos 33
Figuras 3.2 Correlación peso – fumigación 33
Figuras 3.3 Correlación peso – Herbicidas 34
Figuras 3.4 Correlación peso – deschive_f2 34
Figuras 3.5 Correlación peso – calibración superior 34
Figuras 3.6 Correlación peso – riego 35
Figuras 3.7 Correlación peso – selección 35
Figuras 3.8 Matriz de Correlación 36
Figuras 3.9 Diagrama de Caja (variable peso) 36
Figuras 3.10 Histograma (variable peso) 37
Figuras 3.11 Diagrama de Caja (variable peso) 39
Figuras 3.12 Histograma (variable peso) 40
Figuras 3.13 Modelo Random Forest /ntree =100 42
Figuras 3.14 Modelo Random Forest 42
Figuras 3.15 Modelo Random Forest /ntree =200 44
Figuras 3.16 Descripción del modelo 45
Figuras 3.17 Curva de error modelo XGBoost 50 nround 46
Figuras 3.18 Curva de error modelo XGBoost 150 nround 46
Figuras 3.19 Importancia de las variables 47
Figuras 3.20 Estimación vs Real-XGBoost 47
Figuras 3.21 DP-Fumigación 49

Figuras 3.22 DP-Fertilización 49
Figuras 3.23 DP-Deschive_f2 50
Figuras 3.24 DP-Edad 50
Figuras 25 Carga de Datos en RStudio 57
Figuras 26 Dataframe 57
Figuras 27 Summary 59
Figuras 28 str 59
Figuras 29 Valores Nulos - Resultados 61
Figuras 30 Correlación-Resultado 62

RESÚMEN

La agricultura no sólo proporciona alimentos y materias primas, es una actividad con mucho potencial y a su vez la columna vertebral del sistema económico del país, por tal razón, se debe dirigir esfuerzos para ir mejorando esta indispensable práctica. La agricultura en Ecuador no se ha visto beneficiada fuertemente por los diferentes avances tecnológicos en materia de Ciencia de Datos y Machine Learning, estas técnicasayudan a entender los patrones y las dinámicas que sigue sus datos, de esta forma, el objetivo principal de esta investigación es utilizar técnicas de Machine Learning parapredecir pesos de racimos, para posteriormente realizar un análisis que le permita a un Productor tomar las mejores decisiones para sus plantaciones.

En primera instancia de este trabajo investigativo consistió en recopilar todos los datos necesarios para entrenamiento y validación de los modelos de predicción, en la cual las variables de los parámetros de producción tales como peso, riego, fumigación, fertilización, deschive_f2, deschive_f3, # manos del racimo, edad del racimo. La Base de Datos con la que se trabajó, es mediante el Sistema de BI de producción que tiene implementado la Hacienda Bananera San Humberto, cabe resaltar que se tiene datos entre el año 2021 y 2022, por lo tanto, las conclusiones que se obtuvieron en el presente trabajo están respaldadas por datos validados y absolutos.

Ahora bien, con los datos organizados y preprocesados, se entrenan y se validan los modelos para esto se utilizó el programa RStudio y sus diferentes paquetes y librerías. En este caso se utilizó el algoritmo Random Forest, el cual recibe los parámetros internos para adecuar el modelo a los requerimientos del problema, además recibe los datos de entrada y la salida esperada. Después se realizó el modelo con el algoritmo Xgboost que resultó ser más eficiente en su ejecución, aunque no hubo un amplio margen o diferencia entre estos dos algoritmos, se eligió este último como el mejor algoritmo de predicción por sus indicadores de métricas de evaluación. Se usó una variable objetivo Dependiente: peso de racimos, para el primer caso el resultado

de la evaluación del modelo como el Error Absoluto Medio (MAE) para Random Forest fue de 13.2 y en el modelo Xgboost fue de 13.1, en la segunda métrica de evaluación del modelo de predicción fue el Error Porcentual Absoluto Medio (MAPE) para Random Forest fue del 21.39 y para el Modelo Xgboost fue 21.17 por lo consiguiente se tiene una mínima diferencia entre los dos modelos de predicción.

Mediante el análisis del Modelo se obtuvo las variables más influyentes para la predicción y son las siguientes: fumigación, fertilización, deschive_f2 y queda claro que en estas variables se debe poner más atención en la Hacienda Bananera para poder ir mejorando su productividad.

Posteriormente se espera realizar muchas más combinaciones con diferentes variables en torno a la producción Bananera y para una mejor eficiencia en la predicción de peso de racimos de banano se necesitaría que estas se encuentren monitoreadas por sensores dependiendo su naturaleza y así llegar a conclusiones que sean provechosos para el desarrollo de Haciendas Bananeras.

Palabras Claves: producción bananera, machine learning, random forest, xgboost, regression

ABSTRACT

Agriculture not only provides food and raw materials, it is an activity with a lot of potential and in turn the backbone of the country's economic system, for such for this reason, efforts should be made to improve this indispensable practice. The Agriculture in Ecuador has not been strongly benefited by the different advance's technology in Data Science and Machine Learning, these techniques help to understand the patterns and dynamics that your data follows, in this way, the objective The main objective of this research is to use Machine Learning techniques to predict weights of clusters, to later carry out an analysis that allows a Producer make the best decisions for your plantations.

In the first instance of this investigative work consisted of collecting all the data necessary for training and validation of the prediction models, in which the variables of production parameters such as weight, irrigation, fumigation, fertilization, deschive_f2,deschive_f3, # cluster hands, cluster age. The database with which we worked, is through the production BI System that has been implemented Hacienda Bananera SanHumberto, it should be noted that there is data between the year 2021 and 2022, therefore, the conclusions obtained in the present work are supported by validated data.

Now, with the data organized and pre-processed, the models are trained and validated. For this, the RStudio program and its different packages and libraries were used. In In this case, the Random Forest algorithm was used, which receives the internal parameters to adapt the model to the requirements of the problem, it also receives data from expected input and output. After the model was made with the Xgboost algorithm,it turned out be more efficient in its execution, although there was not a wide margin or difference between these two algorithms, the latter was chosen as the best

pre- diction algorithm for its evaluation metric indicators. A Dependent target variable was used: weight of bunches, for the first case the result of the evaluation of the model asthe Mean Absolute Error (MAE) for Random Forest was 13.2 and in the Xgboost modelwas 13.1, in the second evaluation metric of the prediction model it was the Error MeanAbsolute Percentage (MAPE) for Random Forest was 21.39 and for the Model Xgboost was 21.17 therefore there is a minimal difference between the two models of prediction.

Through the analysis of the Model, the most influential variables for the predictions were obtained. and they were: fumigation, fertilization, deschive_f2 and it is clear that these variables It should pay more attention to the Banana Farm in order to improve its productivity.

Later it is expected to make many more combinations with different variables in around the banana production and for a better efficiency in the prediction it would be necessary that these are monitored by sensors depending on their nature and thus reach to conclusions that are beneficial for the development of Banana Farms.

Keywords: banana production, machine learning, random forest, xgboost, regression

INTRODUCCIÓN

El Mundo Bananero le urge la necesidad de innovar sus procesos de producción para ser más eficientes y competitivos, con el buen uso de los recursos naturales y materiales se podrá disminuir costos y aumentar su productividad. Pero lograr esta eficiencia, se requiere de nuevas tecnologías de análisis de Información que permitirán integrarse y así se tomarán decisiones a corto y largo plazo. El banano es el principal producto agropecuario de exportación en el Ecuador, cuyos procesos se ha ido mejorando en todo ámbito, desde la siembra, cosecha y exportación como los controles fitosanitarios y de la fertilización, con el objetivo de aumentar los rendimientos de la fruta. (León Serrano et al., 2021)

Según el Informe de la Organización para la Administración y Agricultura de las Naciones Unidas (FAO, por sus siglas en inglés) la implementación de nuevas tecnologías en la agricultura podría contribuir a mejorar las condiciones de hambre y de pobreza extremas. Estas nuevas tecnologías como el aprendizaje de maquina a través de sus algoritmos de predicción en la cual va tomando fuerza y poco a poco estas diferentes opciones se tendría la implementación de agricultura inteligente. (Chlingaryan et al., 2018)

El Productor que use de estas avanzadas tecnologías está demandando algo más, que es la inteligencia integrada de estos nuevos sistemas en la cual estas tecnologías necesitan de una trama que las interconecte e irremediablemente la solución viene de la mano del análisis de datos y si no se aplica dichas instancias nos daremos cuenta que poco sirve almacenar gran cantidad de información si no sabemos qué hacer con ella. Las empresas lo saben y están poniendo mucho esfuerzo en implementar y desarrollar sistemas automatizados de análisis de datos con inteligencia Artificial/Machine Learning que permitan extraer información realmente útil para que el productor o Administrador, pueda tomar mejores decisiones

agrícolas y comerciales.

El Machine Learning, es una técnica que permite detectar patrones “a bajo nivel” en miles de datos individuales y los modelos predictivos es una de las potencias destacables, ya que facilitan la automatización de procesos, la toma de decisiones y el continuo aprendizaje basado en datos, en la cual estos sistemas esta diseñados para ir mejorando automáticamente con el tiempo y formar partes en las diferentes mejoras informáticas de la compañía.

Por lo tanto, para acelerar el rendimiento de los cultivos, se han propuesto diferentes técnicas de ML en todo el mundo y en el presente proyecto se muestra un resumen de estos en diferentes enfoques, específicamente son los algoritmos supervisados de Machine Learning de regresión, estos algoritmos toman en cuenta varios factores y que son de distinta naturaleza o variable para dar así los mejores resultados de predicción aplicados en datos de producción Bananera.

Finalmente, una de las grandes razones que impulsan la realización de este proyecto es el reto de incorporar estrategias tecnológicas al sector Bananero, que le permitirá no solamente a esta Hacienda Bananera sino a pequeños y grandes productores, el acceso a nuevas herramientas para la toma de decisiones y así puedan tener un crecimiento organizacional, económico y tecnológico estable y notable en dicho sector es decir que por medio del desarrollo de este tipo de proyectos se aporte el conocimiento del negocio basado en la analítica de datos y Machine Learning ayudando el crecimiento de las capacidades del sector y por ende en el país.

CAPITULO 1

1.1. Planteamiento del problema

El Sector Bananero y la agricultura se considera un pilar importante en muchos países por ser la principal fuente de empleo y en su mayoría la agricultura la realizan en forma tradicional, en la cual los agricultores y productores son reacios a utilizar tecnologías avanzadas mientras cultivan y cosechan, debido, a la falta de conocimiento, el alto costo o porque no son conscientes de las ventajas de estas nuevas tecnologías informáticas. La falta de conocimiento para el buen rendimiento agrícola, cosechas erróneas tienden a la pérdida y agrega un costo adicional. Según el famoso dicho “La información es poder”, llevar un registro de la información sobre los cultivos, como los parámetros de producción, labores agrícolas, el medio ambiente y el mercado puede ayudar a los agricultores a tomar mejores decisiones y aliviar los problemas relacionados con la agricultura.

Cabe mencionar que actualmente, poco a poco las Haciendas Bananeras en el Ecuador y especialmente la Hacienda San Humberto ha implementado Sistemas BI en diferentes áreas, como la información en línea de parámetros de producción de cada día de proceso que intervienen las siguientes variables de entrada: peso de los racimos, cantidad de manos por racimos, calibración de dedos, edad de cosecha de cada racimo, # de cuadrilla, # de lotes y por otro lado los pesos de las cajas procesadas, toda esta información se encuentra parametrizados con indicadores establecidos en un Dashboard.

Es muy importante resaltar que esta información es absoluta, es decir cada característica del racimo es ingresado al sistema y solamente el valor de la variable peso es calculada electrónicamente por la balanza de racimos y las demás variables

sí son ingresadas manualmente por experticia del operador.

En la Hacienda Bananera San Humberto se realizan proyecciones o estimaciones de producción semanal con todos los parámetros o indicadores antes mencionados mediante una plantilla de Excel y por ende no cuenta con tecnologías predictivas para la realización de la misma.

Con el análisis de la información y la aplicación de modelos predictivos de machine learning sobre los datos de parámetros de producción se mejoran estas proyecciones ya que un factor elemental en las estimaciones de producción la variable más importante es el peso del racimo.

1.1.1. Pregunta Principal de Investigación. ¿Cuál es el mejor modelo de Machine Learning para obtener una buena predicción de Pesos de Racimos según los Datos de Parámetros de Producción para mejorar la productividad de la Hacienda Bananera?

1.1.2. Preguntas Secundarias de Investigación. ¿Cuáles serán las variables de alta incidencia para los modelos de predicción de machine learning en torno a los parámetros de producción bananera?

¿Qué modelos de Machine Learning será el óptimo para obtener una buena predicción de pesos de racimos según los datos de parámetros de Producción Bananera?

¿Sera Suficiente los datos de los Parámetros de Producción de las Haciendas Bananeras para el análisis y la realización de proyecciones de producción aplicando los modelos de predicción con Machine Learning?

1.2.Objetivos

Objetivo General

Establecer un Modelo Predictivo basado en algoritmos de Machine Learning que permitan obtener proyecciones más exactas de peso de racimos de banano

Objetivos Específicos

Describir las variables relevantes que influyen en el peso de un Racimos de banano y construir una base de datos para el modelo predictivo.

Determinar los diferentes tipos de algoritmos de Machine Learning para el desarrollo del modelo predictivo de pesos de racimos.

Seleccionar 2 algoritmos de Machine Learning e implementar en el software R para comprobar y elegir el de mejor resultado.

1.3. Alcance

La finalidad de obtener un modelo predictivo de Machine Learning aplicado a los datos de parámetros de producción en la Hacienda Bananera San Humberto permite analizar e identificar el mejor algoritmo que mediante sus métricas de evaluación obtendremos los mejores resultados en las proyecciones de pesos de racimos y también ayuda a tener una visión más clara en la producción para semanas futuras dando lugar a posibles incrementos en la productividad de la Hacienda Bananera. Para aquello se obtiene la base de datos del sistema de BI de parámetros de producción que tiene la Hacienda y mediante el entorno del lenguaje de programación R se comienza a trabajar y verificar los datos, en el cual una vez que esta Data se encuentra funcional, se aplica los diferentes algoritmos de Machine Learning y se crea el modelo de predicción de peso de racimos.

Se establece un modelo escalable, que pueda ser fácilmente implementado en las diferentes Bananeras de la región que contengan implementados sistemas de BI de recolecta de datos en el área de producción.

1.3.1. Entre la Limitaciones mencionamos las Siguientes: El operador de la balanza de pesos de racimos, ingresa incorrectamente la información al Sistema BI de parámetros de producción.

Solamente la variable peso del racimo es calculada de forma electrónica, las demás variables son ingresadas manualmente.

El personal técnico y administrativo de la Hacienda no está capacitado para trabajar con herramientas tecnológicas predictivas.

Por razones atípicas en el día del proceso hay fallas técnicas tanto en el dispositivo de la Balanza o celular, en la cual no vamos a obtener datos reales de producción en el Sistema.

1.4.Estado del arte

Se ha realizado una búsqueda de distintos antecedentes nacionales e internacionales relacionados al tema de investigación presentado. Así como, la definición delos términos y palabras clave del proyecto de investigación.

González and Hernandez (2020), presentaron un sistema, que mediante la implementación de técnicas de algoritmos de machine learning contribuyen a un sistema de identificación de imágenes en tiempo real que daba lugar a la supervisión, identificación y clasificación de la calidad de productos. Esta investigación permitió implementar algoritmos de clasificación y enviar los resultados en tiempo on line. Las frutas utilizadas para este estudio fueron naranjas, banano, plátano y manzanas.

La mayoría de los agricultores toman estas decisiones basándose en sus creencias ancestrales, observaciones y propias experiencias. Sin embargo, adquirir experiencia lleva mucho tiempo y por lo general no es práctico observar cada actividad en una granja comercial o Hacienda. Para obtener más información sobre sus granjas, los agricultores confían cada vez más en los datos y del mismo modo recopilan y analizan la mayor cantidad de datos posible. Las granjas recopilan varios tipos de datos utilizando varios tipos de sensores (Degfie et al., 2019). Obviamente, la mayoría de los agricultores no son capaces de procesar los datos sin procesar por sí mismos y confían en las funciones disponibles en los sistemas de información de gestión agrícola que utilizan para administrar y procesar los datos.

En el pasado, los sistemas de información agrícola (SIA), solían ser simples sistemas de gestión de recursos agrícolas, pero hoy en día, algunos de estos sistemas son capaces de procesar datos de sensores detallados y proporcionar amplias funcionalidades de apoyo a la toma de decisiones (Cantero Díaz et al., 2019). Sin embargo, el potencial completo de los datos de varios sensores solo se puede utilizar cuando los SIA comienzan a incorporar algoritmos de aprendizaje automático para

respaldar o automatizar los procesos de toma de decisiones en el sector agrícola.

Según Mohd Shafri and Arenas París (2019), el aprendizaje automático (ML por sus siglas en inglés de Machine Learning) es un subcampo de la inteligencia artificial(IA) y utiliza algoritmos complejos para resolver problemas difíciles de resolver con enfoques tradicionales. Un modelo de predicción de ML se desarrolla, entrenando primero los algoritmos utilizados con un conjunto de datos de entrenamiento y luego validando el modelo con un conjunto de datos de validación separado.

Un conjunto de datos utilizado para ML consta de características (es decir, variables independientes) y el resultado correspondiente (es decir, una variable dependiente). Usando un conjunto de datos de entrenamiento, un algoritmo puede calcular parámetros óptimos para el algoritmo. El algoritmo junto con los parámetros, constituye un modelo de predicción. Se utiliza un modelo de predicción para predecir el resultado para un conjunto determinado de valores de las características utilizadas, lo que respalda la toma de decisiones sobre el resultado previsto (Pallares Cabrera, 2015).

Antes de que un modelo de predicción se utilice en la práctica para respaldar la toma de decisiones, debe validarse. Por lo tanto, una vez que se construye el modelo de predicción, se compara con un conjunto de datos de validación que contiene características y los resultados correspondientes que no se usaron en el entrenamiento del modelo para verificar qué tan bien funciona el modelo. En una situación ideal, el modelo proporciona un rendimiento similar al del conjunto de entrenamiento. Los modelos que funcionan bien en las pruebas se pueden utilizar en la práctica (Rezket al., 2021). Si bien el proceso de capacitación, prueba y uso de modelos ML es sencillo, la creación de un modelo de predicción altamente preciso presenta múltiplesdesafíos, como qué funciones usar, qué algoritmos elegir y cómo manejar grandes cantidades de datos.

Lo antedicho se evidencia en varios documentos, por ejemplo el estudio denominando “Implementación de un módulo de análisis estadístico y predictivo para agricultura utilizando big data y machine learning, integrado al sistema iotmach, en Ma- chala”.(Herrera-Díaz, 2016), tiene como principal objetivo, hacer una implementación de un módulo de análisis estadístico y predictivo para la agricultura, para ello utiliza el Big Data y Machine Learning integrado exclusivamente al IOTMACH, utilizando lenguaje R de programación, con el exclusivo propósito de poder contar con una nuevaherramienta, que permitirá la realización de predicciones, clasificaciones, segmenta-ción o agrupación de los diferentes datos que satisfagan necesidades o problemas que surgen dentro de un negocio.

La investigación de Slob et al. (2021) denominada: Aplicación de aprendizaje automático para mejorar la gestión de granjas lecheras: una revisión sistemática de la literatura; En los últimos años, varios investigadores y profesionales aplicaron algoritmos de aprendizaje automático en el contexto de las granjas lecheras y discutieron varias soluciones para predecir diversas variables de interés, la mayoría de las cuales estaban relacionadas con enfermedades incipientes. El objetivo de este artículo es identificar, evaluar y sintetizar los artículos que discuten la aplicación del aprendizaje automático en el contexto de gestión de granjas lecheras.

El aprendizaje automático o machine learning, implica problemas en los que se desconoce la relación de entrada y salida. El aprendizaje especifica la adquisición automática de descripciones estructurales. A diferencia de los métodos estadísticos tradicionales, el aprendizaje automático no hace suposiciones sobre la construcción exacta del modelo de datos, que describe los datos. Esta función es muy útil para describir comportamientos no lineales complejos, como la predicción del rendimiento de un cultivo (Balducci et al., 2018).

El aprendizaje automático es una parte de la inteligencia artificial empleada para construir un sistema inteligente. Al utilizar las muestras de entrenamiento, se pueden

identificar las muestras de prueba. La precisión del sistema se puede medir utilizando métricas como el error cuadrático medio, la precisión, el recuerdo, la especificidad de sensibilidad, etc. Además, el aprendizaje automático se puede emplear para abordar una variedad de aplicaciones, incluida la predicción del rendimiento de cultivos a través de Métodos de aprendizaje supervisado, no supervisado y por refuerzo (Chandraprabha and Dhanaraj, 2020).

Clasificación, agrupamiento, regresión, predicción son algunas de las técnicas involucradas para lograr el sistema inteligente. En este estudio, se considera la predicción y los métodos utilizados para la predicción, que se elaboran en la siguiente subsecciónes (Maduranga and Abeysekera, 2020)

Para el desarrollo de este proyecto es importante resaltar la importancia de tener herramientas para la toma de decisiones de manera preventiva, con el fin de mitigar posibles riesgos de pérdidas económicas para el agricultor, logrando tener un escenario para una buena productividad y mejorar sus procesos agrícolas. Adicionalmente es la oportunidad de demostrar que el agro ecuatoriano debe transformarse utilizando metodologías y todas herramientas disponibles de tecnologías basadas en aprendizaje de máquina.

Las organizaciones líderes adoptan una cultura basada en datos, realizando un cambio sutil pero significativo en los procesos de toma de decisiones, esta evolución está marcada por usuarios que mejoran los conjuntos de habilidades para que puedan integrar herramientas de análisis en la forma habitual de trabajar para descubrir información estratégica y los principales desafíos en el rendimiento de los cultivos pueden resolverse para mostrar el camino y obtener ganancias (Chandraprabha and Dhanaraj, 2020). Aquellas empresas que obtienen el mayor valor de la analítica aprenden cómo lograr el equilibrio preciso entre el uso de la analítica y los instintos gerenciales, así como también cómo administrar las reglas comerciales junto con la analítica.

Análisis predictivo apunta a diferentes técnicas como la minería de datos, estadística de modelización, aprendizaje automático y se basa en torno a un análisis de datos presentes e históricos para determinar secuencias de patrones y su realizaciónde predicciones sobre ciertas variables desconocidas. (Marqués Gozalbo, 2022).

Slob et al. (2021), establece una serie de procedimientos para adquirir diferentes decisiones como la inteligencia artificial, la minería de datos, el aprendizaje automático y las estadísticas, por ejemplo, la minería de datos implica el análisis de grandes conjuntos de datos que a partir de ellos se detecta diferentes patrones. Hay aplicativos que utilizan modelos estadísticos de datos para realizar estimaciones. Estas innovaciones son útiles para las empresas a administrar inventarios, desarrollar estrategias de marketing y pronosticar las ventas, también ayuda a las empresas a sobrevivir, especialmente aquellas en industrias altamente competitivas, como la atención médica y el comercio minorista.

El análisis predictivo a menudo se asocia con big data y ciencia de datos, las empresas de hoy poseen bases de datos transaccionales, archivos de registro de equipos, imágenes, videos, sensores u otras fuentes de datos. Para obtener información de estos datos, los científicos de datos utilizan algoritmos de aprendizaje profundo y aprendizaje automático para encontrar patrones y hacer predicciones sobre eventos futuros, estos incluyen regresión lineal y no lineal, redes neuronales, máquinas de vectores de soporte y árboles de decisión. Los aprendizajes obtenidos a través delanálisis predictivo se pueden usar más dentro del análisis prescriptivo para impulsar acciones basadas en información predictiva (Cedric et al., 2022).

El análisis predictivo es una técnica analítica importante utilizada por muchas organizaciones para determinar el riesgo, identificar tendencias comerciales futuras y reconocer cuándo se necesita mantenimiento. Utilizando datos históricos como fuente, los científicos de datos aplican varios análisis de regresión y técnicas de

aprendizaje automático para identificar patrones y tendencias contenidos en esos datos. El propósito principal del análisis predictivo es identificar, con un alto grado de probabilidad, lo que sucederá en el futuro. Esto separa el análisis predictivo del análisis descriptivo, así ayuda a los analistas a comprender lo que sucede, y del análisis prescriptivo, que utiliza software de optimización para determinar las mejores decisiones para hacer frente a las tendencias reveladas por el análisis predictivo (Balducci et al., 2018).

La fiabilidad de las predicciones depende en gran medida del modelo de análisis predictivo elegido y de la calidad de los datos utilizados, existen muchas formas diferentes de análisis predictivo, y es importante elegir una que se adapte al problema en cuestión. Aunque las organizaciones pueden tener acceso a una gran cantidad de datos, muchos no están estructurados y deben estar preparados para que puedan ser procesados. Esto incluye la limpieza de datos para eliminar información incorrecta y distorsionada, así como la organización de los datos en un formato adecuado. Las técnicas analíticas más utilizadas incluyen el análisis de regresión y el aprendizaje automático (Deepa et al., 2021).

Machine Learning (ML) es un campo de investigación que se centra formalmente en los sistemas de aprendizaje y la teoría, el rendimiento y las propiedades de los algoritmos. Es un campo altamente interdisciplinario basado en diferentes áreas como la inteligencia artificial, la teoría de la optimización, la teoría de la información, la estadística, la ciencia cognitiva, el control óptimo y muchas otras disciplinas científicas, de ingeniería y matemáticas. Debido a sus muchas aplicaciones, ML ha cubierto casi todos los dominios científicos, por lo que tiene un impacto significativo en la ciencia y la sociedad (Allouhi et al., 2021).

Se requiere un esfuerzo de procesamiento para convertir los datos sin procesar en datos valiosos. Este esfuerzo generalmente incluye: (a) limpieza de datos para eliminar elementos inconsistentes o faltantes y ruido, (b) integración de datos para

reunir datos de muchas fuentes, y (c) transformación de datos para normalizar y diferenciar datos (Deepa et al., 2021).

En este caso, se hace una revisión de las definiciones de ML basado en los modelos agrícolas que se observaron en el estado del arte.

Tabla 1.1

Principales conceptos usado en ML

Siglas	**Descripción inglés**	**Descripción español**
DL	Deep learning	Aprendizaje profundo
ANN	Artificial neural networks	Redes neuronales artificiales
SVM	Support vector machines	Máquinas de vectores de soporte
DT	Decision trees	Árboles de decisión
NN	Neural networks	Redes neuronales
RF	Random forest	Bosque aleatorio
CNN	Convolutional neural networks	Redes neuronales convolucionales
RNN	Recurrent neural networks	Redes neuronales recurrentes
RBN	Restricted Boltzmann machine	Máquina de Boltzmann restringida
DBN	Deep belief network	Red de creencias profundas
SNIC	Simple non-iterative clustering	Agrupación simple no iterativa
SLIC	Simple linear iterative clustering	Agrupación iterativa lineal simple
KC	K-means clustering	Agrupamiento de K-medias
BC	Bagged clustering	Agrupación en bolsas
RPT	Recursive partition trees	Árboles de partición

Continúa en la siguiente página

Tabla 1.1 – Continúa en la siguiente página

Siglas	Descripción inglés	Descripción español
		recursivos
BDT	Booster decision trees	Árboles de decisión de refuerzo
BCT	Bootstrap classification trees	Árboles de clasificación Bootstrap
SB	Stochastic boosting	Impulso estocástico
LR	Logistic regression	Regresión logística
AR	Autoregression	Autorregresión
ARIMA	Autoregressive integrated moving average	Media móvil integrada autorregresiva
VAR	Vector autoregression	Autorregresión vectorial
KNN	K-nearest neighbors	K-vecinos más cercanos
GLM	Generalized linear model	Modelo lineal generalizado
GBM	Gradient-boosting machine	Máquina de aumento de gradiente

En este trabajo investigativo se aplica dos modelos predictivos supervisados de regresión en la cual en base a su eficiencia se elegirá el mejor, tenemos el modelo depredicción por Randon Forest y el modelo de predicción XGBoost.

Random forest o Bosques Aleatorios es un algoritmo de Machine Learning muy utilizado entre los científicos de datos, presenta un sinnúmero de ventajas en comparación con otros algoritmos de predicción. Este algoritmo es muy popular por su capacidad de combinar los resultados de sus diferentes árboles para obtener un resultado final más confiable, por ejemplo, se tiene la predicción del rendimiento de los cultivos en la cual se usó 3 algoritmos y se crearon los modelos. En el primer intento, el que sobresalió fue un modelo de red neuronal, con un Error Cuadrático Medio de

0.0081, después se tiene el modelo Random Forest con un Error Cuadrático Medio de 0.0004, y por último el modelo de Arboles de decisión con una métrica de 0.0168, se observa una buena métrica en la medición de errores de tipo de regresión, donde su potencial predictivo de Random forest fue del 95 % (Arteaga et al., 2020).

Xgboost es una técnica de machine learning que se basa en árboles de decisión, es el más usado en la actualidad por su velocidad y el rendimiento, tiene un dual de resolución de modelos tanto lineal como de aprendizaje de árboles entonces, lo que lo hace rápido es su capacidad para realizar cálculos paralelos en una sola máquina, el uso del algoritmo Xgboost, utilizando los indicadores de evaluación para seleccionar aquel modelo que permita obtener mejores pronósticos para tener una mejor pre visualización de los datos y en base a esto, tomar las mejores decisiones (Villafuerte Chacnama, 2021). En consecuencia Swami et al. (2020), presentan un paper, en el cual comparan el poder predictivo de los modelos Xgboost, Long Short Term Memory (LSTM) aplicados a una serie de ventas mensuales extraída de la plataforma Kaggle, basándose en el Error cuadrático medio (RMSE), se encontró que el modelo Xgboost brindaba mejores resultados que el modelo LSTM.

CAPITULO 2

2.1.Tipo y diseño de investigación

La investigación predictiva como trabajo principal se ocupa de pronosticar resultados, consecuencias, costos, es decir la dirección futura de los eventos investigados. Este tipo de investigación trata de anteponerse al análisis de anomalías, políticas u otras entidades existentes para predecir algo que no se ha intentado, probado o propuesto Pereyra (2020).

La investigación está basada en un enfoque cuantitativo, en función de identificar los factores o parámetros de producción que rigen el proceso bananero. Meshram et al. (2021), la tecnología Blockchain, la computación en la nube, Internet de las cosas (IoT), el aprendizaje automático (Machine Learning) y el aprendizaje profundo (Deep Learning), son las últimas tendencias emergentes en el campo de la informática. Ya se ha utilizado en diferentes dominios como la sanidad, el cybercrimen, la bioquímica, la robótica, la metrología, la banca, la medicina, la alimentación, etc., para resolver los complejos problemas de los investigadores.

Según Ortega (2018), en el enfoque cuantitativo para comprobar una hipótesis se requiere de la recolección de datos que permite mediante la medición numérica y el análisis estadístico, determinar patrones de comportamiento y examinar teorías; así mismo afirman que si en una investigación no se efectúa la manipulación intencional de variables y solo se limitan a observar los fenómenos en su contexto original para analizarlos pertenecen a un tipo de investigación no experimental de enfoque cuantitativo, ya que busca comparar diferentes metodologías de predicción mediante un criterio empírico, el cual, se construye con información cuantificable, siguiendo un orden metodológico riguroso y secuencial, tomando en cuenta información

correspondiente a investigaciones previas como punto de partida referencial, y generando un modelo que describa el comportamiento de la población en estudio.

En consecuencia, por tener las características expuestas anteriormente, la presente investigación es no experimental de enfoque cuantitativo. (Ortega, 2018) indican que un diseño de investigación transversal se caracteriza porque el proceso de recolección de los datos se lleva a cabo en un único momento, en un determinado tiempo; así mismo tiene como propósito la descripción de representar y analizar su influencia e interacción en un determinado momento, por lo tanto:

El diseño correspondió al transversal retrospectivo porque se trabajó con datos históricos de parámetros de producción de la Hacienda Bananera San Humberto de la Provincia del Guayas cantón Duran Zona 8, registrados marzo del 2021 hasta junio del 2022.

2.2.La población y la muestra

La población seleccionada es la Hacienda Bananera “San Humberto” ubicada en el Cantón Durán Zona 8 Ecuador, en la cual se obtuvo del sistema de BI (Inteligencia de Negocios) que permitirá recolectar una gran cantidad de datos de parámetros de Producción con un total 498526 registros en el rango de fechas que abarca desde marzo 2021 hasta junio 2022.

2.2.1. Características de la población. Es una hacienda bananera que queda ubicada a 11 kilómetros de la ciudad de Durán en la zona 8 de Ecuador, su capacidad de producción es 2650 cajas por hectáreas anual en la cual tiene contrato fijo con empresa exportadoras de Banano, cuenta con 204.1 hectáreas de Producción Establecida, 6.2 Has como R1 y 11.11 Has como R0 que totaliza 221.41 hectáreas de Producción, con una densidad de 1450 plantas por Has. La hacienda tiene 22 lotes y cada lote tiene aproximadamente 10 hectáreas.

2.2.2. Dispone de una Fuerza Laboral de casi 200 colaboradores que se encuentra dividido en áreas administrativas, agrícolas y de empacadora.

2.3.Técnicas de Observación e Instrumentos de Colecta de Datos

Bernal Pablo (2018) indica que, para evitar ambigüedades o sesgos en la recolección de datos, los instrumentos deben ser revisados y avalados, cuyo requisito será, además de la validez sostenida del estudio, la confiabilidad del instrumento elaborado por el investigador. Los criterios éticos de la investigación se fundamentan en la explicación del carácter interpretativo del investigador y la necesidad de dar sentido a las expresiones de los sujetos a partir de la calidad de las expresiones de los hechos. De esta forma, el análisis de los hallazgos puede apoyarse en los planteamientos de procesos específicos que pueden reforzar la validez y confiabilidad de los estudios cuantitativos.

La técnica que se utilizará para la obtención de la información para nuestra investigación, fue la entrevista y el análisis documental. Se utilizó las métricas del modelo propuesto y se obtuvo el indicador de Evaluación del Error necesario para medir el éxito de la investigación.

2.3.1. Entrevista. La entrevista fue realizada al Gerente General de Producción y el Analista de Producción. En esta etapa de la investigación, a las personas encargadas sobre la producción de la Hacienda Bananera se les propuso una serie de preguntas referente a buenas labores agrícolas y productividad bananera tanto para el Gerente del área y el Analista de producción, tal como se presenta en el anexo de la página 71 de este trabajo de investigación.

2.3.2. El análisis documental. Es el instrumento por el cual se obtiene los datos primarios a través del personal entrevistado, se obtuvo la información solicitada desde que la Hacienda implemento su Sistema de BI para llevar indicadores de los parámetros de producción. Las fuentes fueron una base de datos en hoja de cálculo (Excel).

2.4.Operacionalización de las Variables

Tabla 2.1
Matriz de operacionalización de variables

Alcance	Segmentación	Dimensiones	Indicadores	Técnicas	Instrumentos
Modelos de Aprendizaje automáticos basado en técnicas supervisadas de regresión	Modelo Presupuesto (Variables Independientes)	Sistema BI	Cantidad de resgitro de parámetros de Producción contenida en el Sistema BI Calibración # Manos Edad Deschive Deschante Selección Deshoje Fertilización Riego Fumigación	Entrevistas Análisis Documental	Cuestionarios Grabadora de Audio y Video Base de Datos en Formato de Excel
	Predicción de peso de Racimos –(Variable Dependiente)	Métricas	(MSE-RMSE-MAPE-COEFICIENTE DE DETERMINACIÓN)	Técnicas de Random Forest- Tecnicas de Extreme Gradient Boosting	Lenguaje R

2.5.Instrumentos

En este caso considerando los instrumentos validados y del Sistema de BI que tiene la Hacienda Bananera San Humberto, la recolección de datos se realizaron desde el primer día que se implementó este sistema en la cual obtenemos todos los datos de producción.

Para la elaboración de los modelos de utilizó el software R versión 4.2.0 (2022-04-22 ucrt), contiene herramientas específicas e inflexibles ya que dispone de una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, clasificación, agrupamiento y graficas que permiten incluir todos los procesos para el análisis requerido en esta investigación, desde importar las tablas excel que en este caso fueron obtenidas del Sistema de BI de Producción hasta

validar los modelos con herramientas como la matriz de confusión y el análisis de curvas ROC (receiver operating characteristic curve) que viene a ser un método estadístico para precisar la exactitud de los modelo incluyendo conceptos de sensibilidad y especificidad que permite evaluar y discriminar entre datos correctamente predichos e incorrectamente predichos, como también validar los modelo de regresión con técnicas de evaluaciones de errores como el Error Medio Absoluto (MAE), Error Cuadrático Medio (MSE), Raiz cuadrada del error cuadrático medio (RMSE), Error Porcentual Absoluto Medio (MAPE).

Los campos para cada registro obtenidos de los instrumentos anteriores se detallan en la siguiente tabla:

Tabla 2.2
Campo de los datos de parámetros de producción.

Nro	Campo	Tipo de dato registrado	Descripción del campo	Medición	Ref.
1	Fecha	Fecha	Fecha de Registro	22/5/2022	Fecha
2	Calibración	Numérico	Grosor de dedos de lo racimos	44,5	Grados
3	Nro de manos	Numérico	Cantidad de (manos) gajos en el racimo	9	Cantidad
4	Edad	Numérico	Edad del racimo	11,8	Semanas
5	Peso	Numérico	Peso del racimo	77,5	Libras
6	Palanca	Numérico	Cuadrilla de corte	6	Cantidad
7	Lote	Numérico	Ubicación de la hacienda del racimo cortado	2	Ubicación
8	Deschive	Alfanumérico	Labor realizada en los racimos	V/F	TRUE/FALSE
9	Deschante	Numérico	Labor realizada en la planta	1/0	TRUE/FALSE
10	Selección	Numérico	Labor realizada en la planta	1/0	TRUE/FALSE
11	Deshoje	Numérico	Labor realizada en la planta	1/0	TRUE/FALSE
12	Fertilización	Numérico	Labor realizada en la planta	1/0	TRUE/FALSE
13	Riego	Numérico	Labor realizada en la plantación	1/0	TRUE/FALSE
14	Fumigación	Numérico	Labor realizada en la plantación	1/0	TRUE/FALSE

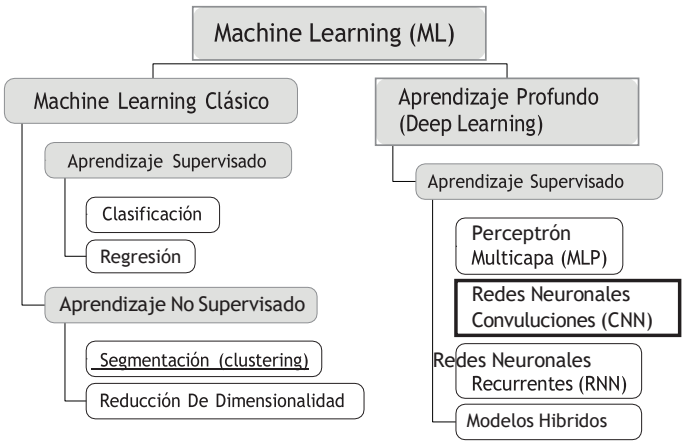
2.6.Procedimientos

Lopez Briega (2015) dice que el proceso de aprendizaje automático se descompone en varios pasos, entre los que se incluyen la recopilación de los datos, el preprocesamiento de dichos datos, el entrenamiento - prueba de los modelos y la evolución de los mismos.

Aprendizaje automático (machine Learning) es el estudio Informático de algoritmos que mejoran el rendimiento en función de la experiencia o entrenamiento, consiste en programar un equipo para optimizar un criterio de desempeño utilizando datos de experiencia pasada o ejemplos si no existe experiencia humana.

Como indica VanderPlas (2016), en su trabajo investigativo Machine Learning, comprende la elaboración modelos mediante algoritmos y así analizar la estructura de los datos, el "aprendizaje" de estos modelo sucede cuando se va cambiando poco a poco los parámetros de los algoritmos y los modelos se van adaptando a las necesidades y con ello se evalúa y se pueden realizar todo tipo de predicciones.

Figura 2.1. Jerarquía de Machine Learning

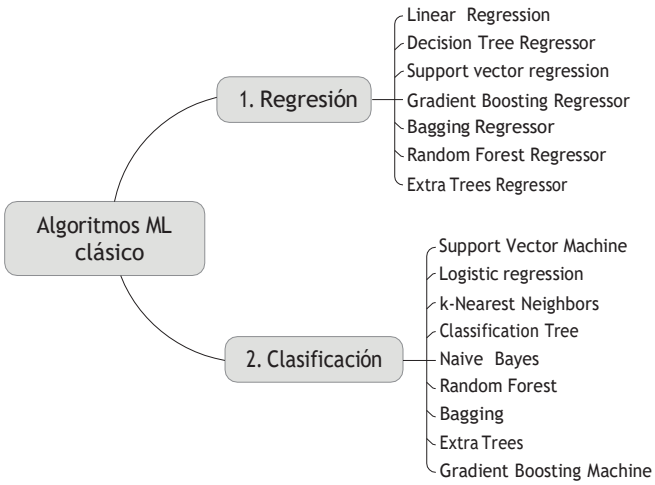


2.6.1. Aprendizaje supervisado. En el aprendizaje supervisado las variables predictoras y las variables objetivas están estipuladas y sirven para construir los modelos con la predisposición de estimar situaciones que se obtengan

posteriormente, este tipo de métodos se consideran supervisados por que el modelo se construye con variables conocidas, es decir, el sistema “aprende” de las variables conocidas con el objetivo de estimar resultados.

Los algoritmos de aprendizaje supervisado se categorizan mediante la diferenciación con respecto al tipo (cuantitativo o cualitativo) de la variable de salida involucrada en el problema, la regresión se utiliza cuando el resultado es cuantitativo y la clasificación cuando el resultado es cualitativo como proponer un modelo de predicción de aprendizaje supervisado que se ajuste a los requerimientos de clasificación deuna empresa (Crisóstomo Fernández et al., 2021). En la siguiente Figura se muestra los diferentes tipos de algoritmos de aprendizaje supervisados.

Figura 2.2. Tipos de Algoritmos Supervisados



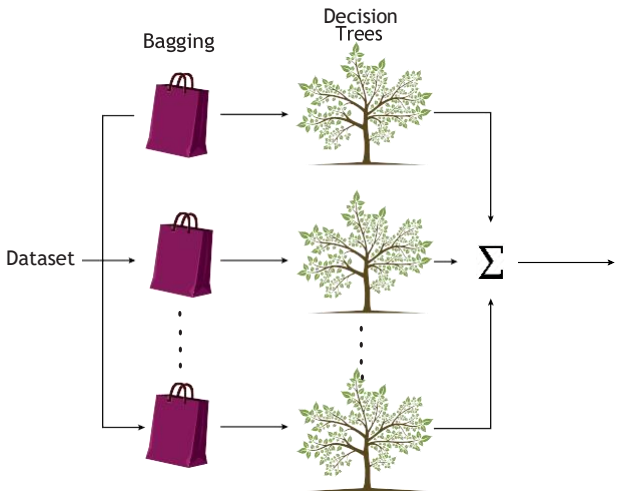
Se investigó los diferentes algoritmos de aprendizaje supervisado de regresión y según su eficiencia, se seleccionaron dos algoritmos el Xgboost y Random Forest que mediante sus métricas de rendimientos se elegirá el mejor modelo para la predicción de pesos de racimos de banano para una buena productividad bananera y mejoras en los procesos agrícolas.

2.6.2. Random Forest. Los bosques aleatorios es un conjunto de Árboles de decisión no entrenados con diferentes muestras de datos con reemplazo

y a esto se llama bootstrap. En la estructura del algoritmo (RF) se tiene los árboles de decisión en la cual aprende al dividir los datos desde el inicio hacia adelante de acuerdo con la métrica de división en cada nodo de decisión. Como mencionamos anteriormente, la agregación Bootstrap reduce las diferentes variaciones del modelo (sobreajustes) y arroja como resultado el aprendizaje de la muestra o conjuntos de datos diferentes.

Los procedimientos de agrupación de conjuntos de datos utilizan diversos modelos de aprendizaje para tener mejores resultados precisos: debido a un bosque indiferente, el algoritmo realiza que el bosque total de árboles de elección continua no similares comience a emerger y así obtener la mejor correlacione entre ellos. En un bosque irregular se debe reducir el problema de relación al que se hace referencia anteriormente, principalmente su alcance es relacionar árboles y podarlos mediante la ejecución de una medida de parada para las divisiones de nodo (Eulogio, 2017)

Figura 2.3. Algoritmo Random Forest



Con las características principales del random forest como el bagging toma muestras de los datos de entrenamiento para generar múltiples arboles decisión, se realiza el re-muestreo y a su vez reduce la varianza de las predicciones.

2.6.3. Xgboost. Es una técnica de aprendizaje automático basada en arboles de decisión en la cual sus bases vienen del Gradient Boosting para obtener eficiencia en su velocidad de entrenamiento y el rendimiento de los modelos. Unas de sus destacables características es que utiliza todos los núcleos de la CPU durante el entrenamiento, optimiza los recursos de memoria y computación distribuida, lo que permite manejar grandes conjuntos de datos. Sus notables cualidades como en ser eficiente, flexible, portátil, optimizado, además de que resuelve muchas problemáticas de ciencia de datos.

La eficiencia de Xgboost surge porque se utiliza la potenciación de gradientes, esta técnica realiza el modelo predictor a partir del empaquetamiento de modelos débiles ose de grupos de árboles de decisión y los muestra de forma general a través de una función de pérdida común. El algoritmo de Xgboost que se utiliza en este proyecto, permite definir algunos hiperpárametros que determinan el comportamiento del mismo en la cuales tenemos los siguientes:

Target: encuentra el valor a predecir, en este caso un número.

MaxDepth: Indica la profundidad máxima de los árboles de decisión mientras más profundos son los árboles, más probabilidad de sobre ajustar los datos.

2.6.4. Software R. En base a lo indicado anteriormente y para el desarrollo de esta investigación utilizaremos las ventajas del Software R, el cual se encarga del análisis de los datos sin procesar y convertirlos en nuevo conocimiento, R es un lenguaje de programación de código abierto que viene otro lenguaje llamado S, cuenta con más de 15000 librerías en la cual abarca casi en su totalidad las diferentes áreas de estudio como el análisis datos, áreas financieras, bayesiana. Unasde sus características más conocidas son sus diferentes herramientas estadísticas que posee para el análisis de datos, permitiendo a los usuarios crear sus propias funciones, también posee capacidades graficas muy destacables (Jiménez, 2019).

Para complementar el entorno para la programación en R es RStudio y tiene un panel de control muy bien distribuida, brinda los mejores paquetes y librerías y para la ciencia de datos. (Wickham and Grolemund, 2017).

En la siguiente parte se mencionan algunas de las librerías que se aplicaron para el desarrollo de los modelos predictivos:

RandomForest: dispone en sus parámetros como el tree y rpart: permiten controlar la asignación de árboles.

Xgboost: Significa (Extra Gradient Boosting), en la cual unos de sus parámetros fundamentales es nrounds que ayuda ingresar la cantidad de cuantas iteraciones realizara el algoritmo al realizar el modelo. Dplyr: Uno de los paquetes principales de tidyverse en el lenguaje de programación R, dplyr es principalmente un conjunto de funciones diseñadas que contiene una colección de funciones para realizar operaciones de manipulación de datos comunes como: filtrar por fila, seleccionar columnas específicas, reordenar filas, añadir nuevas filas y agregar datos.

Caret: Kuhn et al. (2016) menciona que este paquete contiene diferentes funciones que nos ayuda en la realización de los diferentes casos de clasificación y regresión.

2.6.5. Recopilación de los datos. Esta fase inicial para el desarrollo de los modelos predictivos se realizó con el apoyo de profesionales encargados del BI, en este caso fueron el nexo para el acceso a los datos de registros de parámetros producción desde marzo 2021 hasta junio 2022.

2.6.5.1. Fase de preprocesamiento de los datos. Este es el inicio del análisis exploratorio de los datos, a pesar que los datos obtenidos de registros de parámetros de producción, se pudo identificar que previo al desarrollo de los modelos predictivos se requiere verificar los atributos o características de las diferentes

variables en la cual se muestra la siguiente tabla:

Tabla 2.3
Tipo de variable

Nro	Campo	Tipo de dato	Descripción del campo	Variable	Tipo
1	Fecha	Numérico	Fecha de Registro	Independiente	Cuantitativa
2	Calibración	Numérico	Grosor de dedos de lo racimos	Independiente	Discreta
3	Nro de manos	Numérico	Cantidad de (manos) gajos en el racimo	Independiente	Discreta
4	Edad	Numérico	Edad del racimo	Independiente	Discreta
5	Peso	Numérico	Peso del racimo	Dependiente	Continua
6	Palanca	Numérico	Cuadrilla de corte	Independiente	Discreta
7	Lote	Numérico	Ubicación de la hacienda del racimo cortado	Independiente	Discreta
8	Deschive	Numérico	Labor realizada en los racimos	Independiente	Binaria
9	Deschante	Numérico	Labor realizada en la planta	Independiente	Binaria
10	Selección	Numérico	Labor realizada en la planta	Independiente	Binaria
11	Deshoje	Numérico	Labor realizada en la planta	Independiente	Binaria
12	Fertilización	Numérico	Labor realizada en la planta	Independiente	Binaria
13	Riego	Numérico	Labor realizada en la plantación	Independiente	Binaria
14	Fumigación	Numérico	Labor realizada en la plantación	Independiente	Binaria

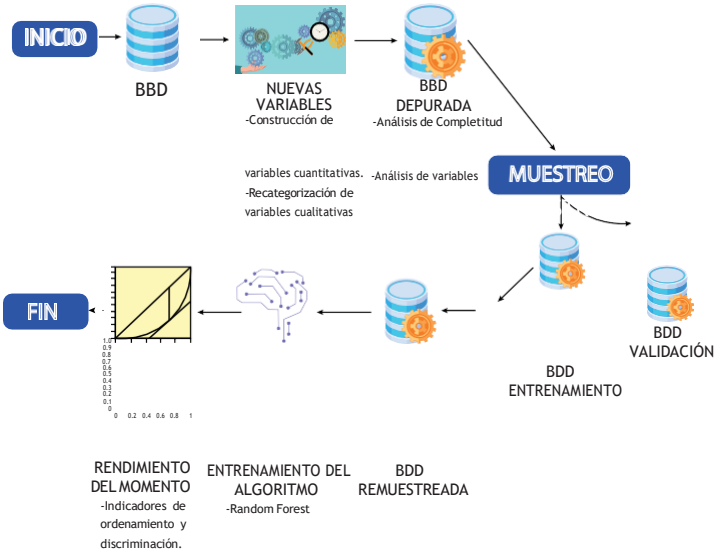
El preprocesamiento de datos son las transformaciones de los datos realizados con la finalidad de que puedan admitir los algoritmos de machine learning y para que también puedan mejorar sus resultados. El preprocesamiento de datos sedebe realizar al inicio y luego aplicarse al conjunto de entrenamiento y al de testo prueba. Esto es muy importante para no cambiar la condición de que ninguna información procedente de las observaciones de test puede participar o influir en el ajuste del modelo. Algunos pasos de preprocesamiento que más suelen aplicarse son: Imputación de valores ausentes Estandarización de las variables numéricas Binarización de las variables cualitativas.

Kotu and Deshpande (2015) expresan que la visualización exploratoria, es el proceso de mostrar datos en coordenadas visuales y que le permiten a los usuarios encontrar patrones y relaciones en los datos y de esta manera comprender el comportamiento de un conjunto de datos de tamaño considerable; es decir, en forma similar que las estadísticas descriptivas, son integrales en las etapas de preprocesamiento y posprocesamiento.

2.6.5.2. Fase de entrenamiento. En la siguiente Fase se debe calificar de una manera acertada el error, y es por esta razón se necesita tener un conjunto separado, de las que se conozca la variable objetivo, pero que el modelo no haya reconocido, es decir, que no hayan participado en su ajuste. Con esta finalidad, se dividen los datos, en un conjunto de entrenamiento y un conjunto de test o prueba. El tamaño aconsejado de las divisiones depende en gran medida de la base de datos datos disponibles y de la seguridad que se necesite en la estimación del error, 70 %-30 % suele dar buenos resultados.

2.6.5.3. Fase de análisis, evaluación. En este estudio para el desarrollo de cada uno de los modelos de predicción se utilizó el software R que es una aplicación que no sólo ha permitido el desarrollo de cada uno de los modelos de predicción, sino también fue muy útil para realizar el análisis estadístico descriptivo del comportamiento de los atributos más importantes. Se encontró algunas técnicas para calificar los resultados de un algoritmo en la de predicción de datos, entre ellos son las pruebas de bondad y ajuste, y es entonces al revisar los modelos se recomienda que al realizar las pruebas se tome como referencia resultados históricos. (Dadas et al., 2019)

Figura 2.4. Realización del Modelo Predictivo



2.6.6. Métricas de Evaluación. El objetivo de un modelo de aprendizaje automático, es aprender, desde un conjunto de datos, patrones que permitan generalizar la predicción a datos nunca antes vistos. Para evaluar un modelo se divide el conjunto original en partes, el set de entrenamiento, set de pruebas. El set de entrenamiento es usado para “construir” el modelo (encontrar sus parámetros), el set de validación se usa para evaluar el modelo entrenado con el set de entrenamiento, mientras se ajustan los parámetros del modelo en entrenamiento, y por último el set de prueba se usa para ver qué tan bien lo hizo el modelo, con los hyperparámetros ajustados, sobre datos no mencionados. Con la medición del rendimiento del modelo predictivo y más aún si es un modelo de regresión su medición se basa en el error, estos ayudan a tomar una decisión qué tan eficiente es el modelo prediciendo con nuevos datos o variables. (Developers, 2021).

2.6.6.1. Error cuadrático medio (MSE). Es un indicador de medición más simple para la evaluación del modelo de regresión y su ecuación es la siguiente:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_1 - \hat{y}_i)^2 \qquad (2.1)$$

donde y1 es el resultado real esperado y yˆi es la predicción del modelo.

En su cálculo esta ecuación predice para cada punto estimado es la diferencia cuadrada entre la predicción y la variable objetivo y por último calcula el promedio de dichos valores. Si es resultado o la cantidad es muy grande el modelo creado es malo, tampoco el resultado nunca da negativo y si en algún momento es cero concluyéramos que el modelo es perfecto.

2.6.6.2. Error cuadrático medio (RMSE). Esta métrica de precisión RMSE se obtiene solo con la raíz cuadrada de MSE, se introduce para hacer que la escala de los errores sea igual a la escala de los objetivos, su ecuación es la siguiente:

$$\text{RMSE} = \frac{1}{m}\sum_{i=1}^{m}(y_1 - \hat{y}_i)^2 \qquad (2.2)$$

2.6.6.3. Error absoluto medio (MAE). Una vez teniendo los resultados anteriores de la métrica se encuentra el Mean Absolute Error (MAE), esta métrica obtiene el promedio de los errores de cada predicción entre lo predicho y lo real. Es decir, esta métrica es muy importante porque permite tener un mayor control sobre el error promedio que tiene el algoritmo sobre lo real y lo estimado. El Error medio Absoluto trabaja de manera lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio, y su ecuación es la siguiente:

$$\text{MAE} = \frac{1}{m}\sqrt{\sum_{i=1}^{m} |c_i - \hat{c}_1|} \tag{2.3}$$

2.6.6.4. Coeficiente de determinación R^2***.*** Como parte Final el coeficiente de determinación es una métrica muy conocida en los problemas de regresión.Es importante destacar que el resultado del coeficiente de determinación oscila entre0 y 1. Cuanto más cerca de 1 su valor, mayor será el ajuste del modelo a la variable que estamos intentando predecir.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (c_i - \hat{c}_1)^2}{\sum_{i=1}^{m} (c_i - \overline{c})^2} \tag{2.4}$$

en donde $\bar{c}$ se define como:

$$\overline{c} = \frac{1}{m}\sum_{i=1}^{m} c_i \tag{2.5}$$

2.6.6.5. Error porcentual absoluto medio (MAPE). La métrica de Machine Learning que se usa para evaluar el desempeño de los modelos es el porcentaje medio del error absoluto MAPE, que es una relación media entre el error absoluto y el valor absoluto y permite dar una idea del tamaño de los errores en comparación con los valores. El MAPE, mide la precisión como un porcentaje. Esta es la medida más común para pronosticar el error ya que las unidades de la variable se escalan a unidades porcentuales y facilita la comprensión.

La ecuación que representa al MAPE es la siguiente:

$$\text{MAPE} = \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \tag{2.6}$$

Donde:

x_i son las observaciones actuales de las series de tiempo.

x_1 son las series de tiempo estimadas o pronosticadas.

N es el número de punto de datos no faltantes.

3. Propuesta de solución

3.1.Preparación y manejo de variables

Tanto R (R Project) y R-Studio, son softwares Open Source y gratuitos. Para poder utilizar R-Studio, previamente ha sido instalado R (Ver Apéndice A de la página 58) y de acuerdo el avance del proyecto se irá instalando los respectivos paquetes y librerías para la realización del mismo.

Los datos están recogidos en ficheros Excel con formato .xlsx llamado “consulta racimos.xlsx”. Esta Base de datos se la obtiene mediante el sistema de BI de producción que tiene implementado la Hacienda Bananera San Humberto y contieneinformación sobre los parámetros de producción obtenidas en este rango de fechas desde el 24/04/2021 hasta 28/06/2022. La información corresponde a distintos aspectos relacionados con la producción Bananera como lo muestra en la Tabla 3.1

Tabla 3.1
Variables de la Base de Datos consulta racimos.xlsx

Variables	
fecha	año
manos	deschive F/2
calibracion_superior	deschive F/3
calibracioninferior	deshoje
longitud_dedos	deschante
palanca	selección
lote	fertilización
peso	riego
edad	herbicidas
mes	fumigación

En la Tabla 3.1 se observa las diferentes variables a utilizar en la cual mediante el programa R Studio se carga la base de datos (Ver Apéndice B de la página 59) y se comienza a realizar la exploración de datos, mediante los comandos *summary* y *str* muestra un resumen general sobre las variables del Dataframe, indicando de qué tipo de variable pertenece y un resumen estadístico como la media, Valor máximo, mínimo, percentiles, como resultados se obtuvo que la variable dependiente(peso) y la variable manos está ingresada en la base datos como variable de tipo categórica o cualitativa , entonces se procedió a cambiar el tipo de datos a cuantitativas o numérica (Ver Apéndice B de la página 59). Otro paso a seguir es la verificación de datos nulos en las respectivas variable del Dataframe mediante el comando is.na, pero mostró error en las variables deschive F/2 y deschive F/3 por el espacio que forman el nombre de la respectiva variable y por ende se cambió el nombre de las variables con el comando *repage*y quedaron así: deschive_f/2, deschive_f/3 una vez rectificado el error se procede a la verificación de la existencia de valores nulos en los variables del Dataframe (Ver Apéndice B de la página 59), y como resultado no se obtuvo valores nulo en cada una de las variables del Dataframe.

3.2.Correlación

Se verificó si existían correlación de datos mediante el comando *cor* en la cual consiste en analizar la relación de al menos dos variables y como resultados se muestran en el Apéndice B de la página 59 y como resultados se muestra en la siguiente

Tabla 3.2
Correlación de Variable del dataframe ConsultaRacimos.csv

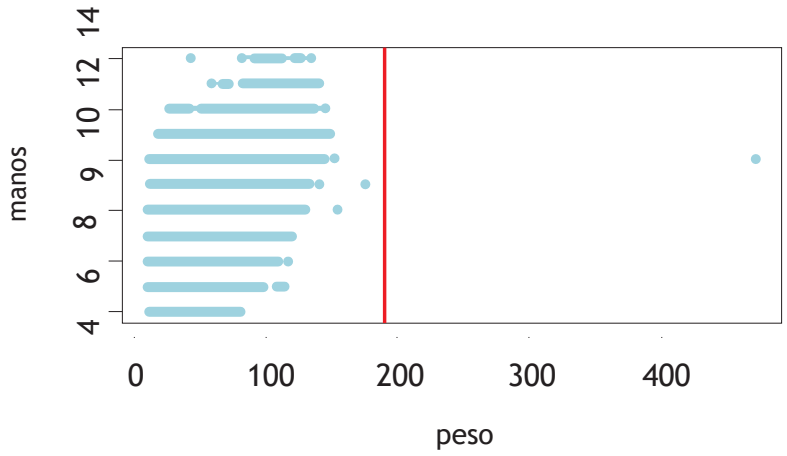
Variables	Coef. Correlación (%)
manos	74,5
calibracion_superior	19,5
deschive_f2	16,2
fumigacion	8,1
palanca	3
fertilizacion	0,04
herbicidas	-2,6
deschante	-2,6
seleccion	-2,6
edad	-4
riego	-7,85
deschive_f3	-16,2

Se tiene un criterio al momento de analizar los resultados del coeficiente de correlación utilizando los siguientes rangos:

Si es 0 y 0,10: correlación cero Si es 0,10 y 0,29: correlación baja Si es 0,30 y 0,50: correlación estable Si es 0,50 y 1,00: correlación alta

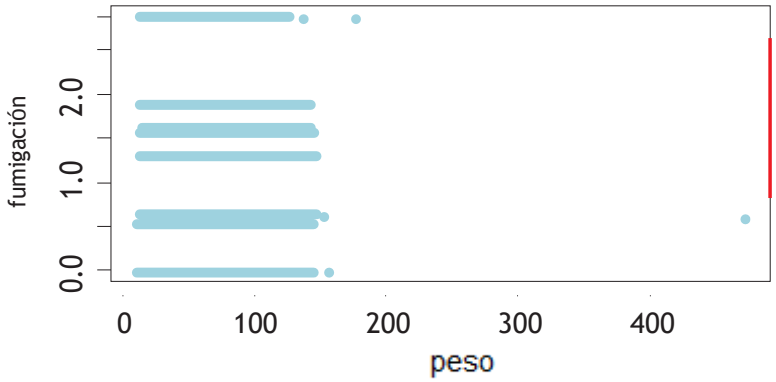
Mediante la gráfica de correlación de la variable dependiente(peso) con las variables independientes, se visualiza la relación entre las variables y cuando la recta se inclina hacia la derecha la correlación es positiva, pero cuando se inclina hacia la izquierda es negativa como se muestran en las siguientes imágenes de Correlaciones.

Figura 3.1. Correlación peso – manos



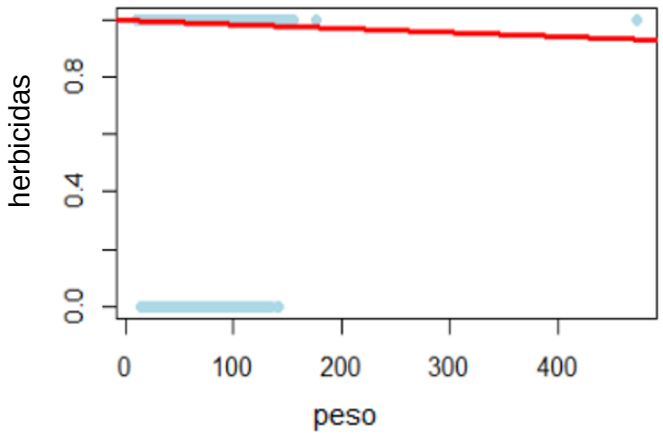
En esta imagen de Correlación de la variable manos (y) con la variable principal peso (x) se observa que ha mayor número de manos puede mejorar el peso del racimo, pero no necesariamente suele pasar aquello ya que existe la correlación, pero los datos se encuentran lejos de curva.

Figura 3.2. Correlación peso – fumigación



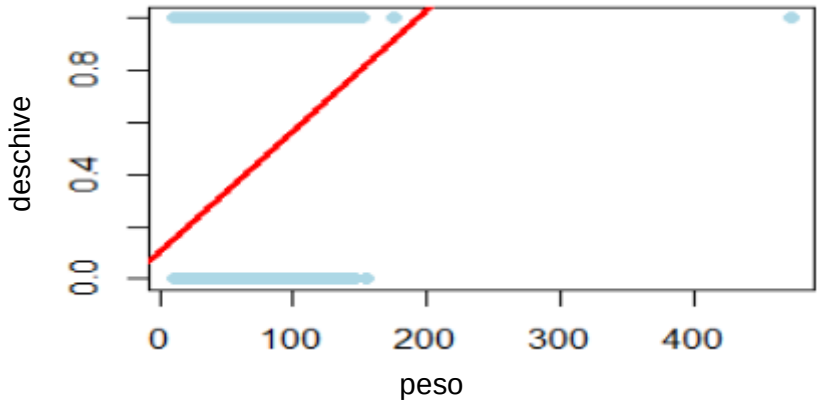
En esta imagen de correlación de la variable fumigación (y) con la variable principal peso (x) se deduce que de forma muy general se aplica y en su consecuencia también se obtiene racimos con buen peso, pero no hay correlación directa por que los datos están muy lejos de la curva.

Figura 3.3. Correlación peso – Herbicidas



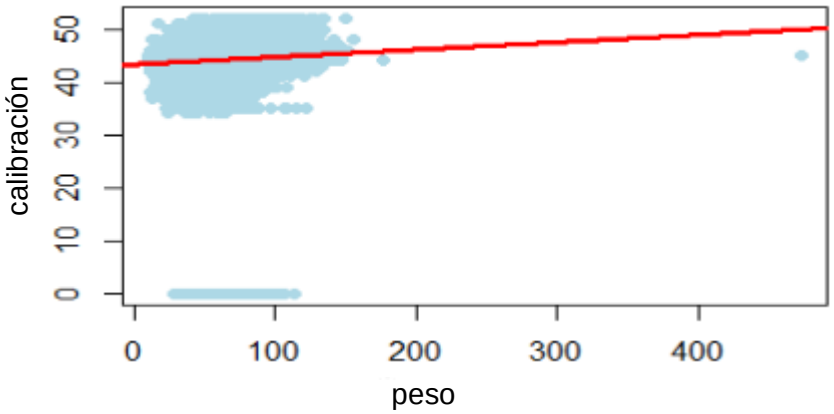
A diferencia de las correlaciones anteriores si existe correlación de la variable herbicida (y) con la variable principal peso (x) pero es negativa o inversa

Figura 3.4. Correlación peso – deschive_f2



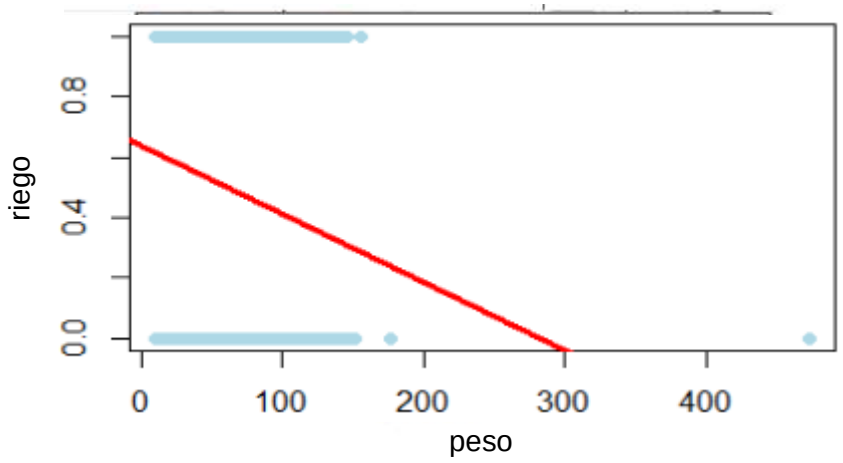
Aquí se observa una correlación positiva, aunque no muy fuerte por que los datos se alejan de la curva creciente.

Figura 3.5. Correlación peso – calibración superior



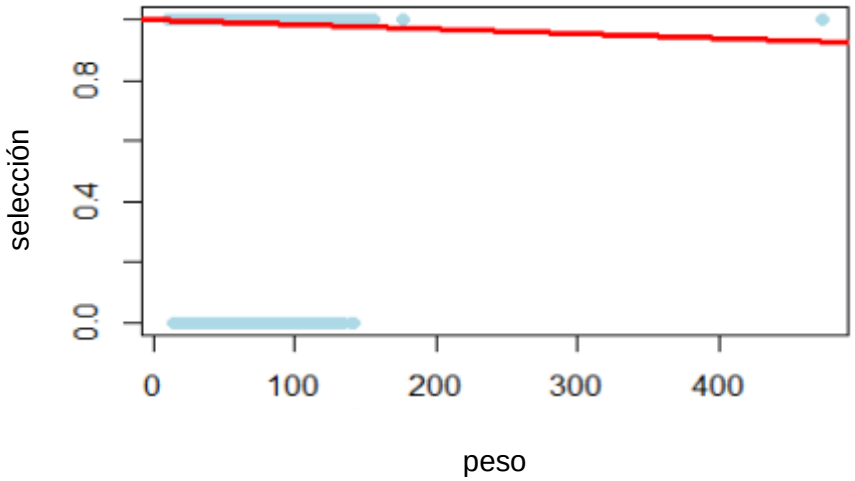
En esta imagen se observa una muy buena correlación a medida que la curva va creciendo la variable peso de racimos tienden a aumentar.

Figura 3.6. Correlación peso – riego



Se observa una correlación negativa a medida que el peso del racimo aumenta el riego disminuye o viceversa.

Figura 3.7. Correlación peso – selección

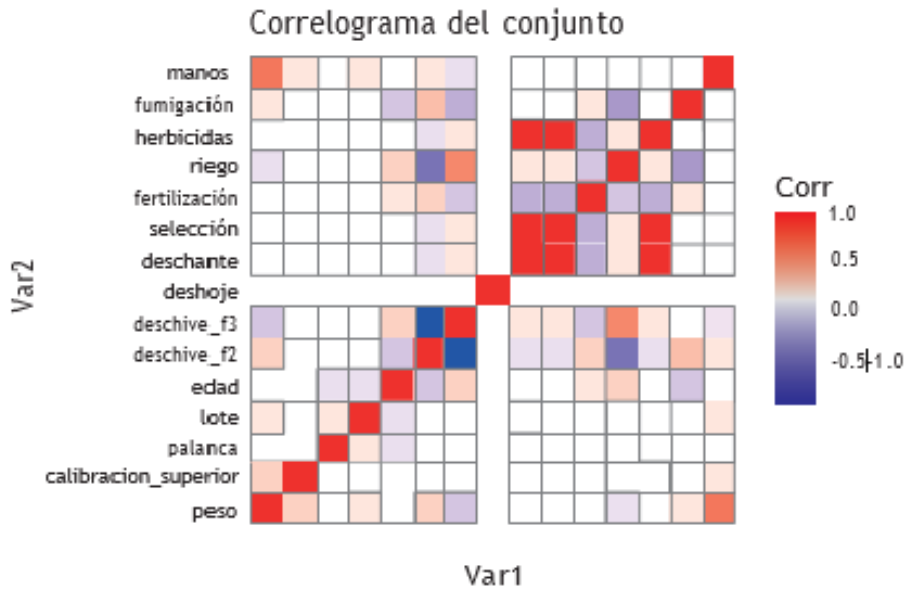


Se observa una correlación negativa a medida que el peso del racimo aumenta las labores de selección o viceversa.

Continuando con este ejercicio de correlación de variables en esta parte se tiene la Matriz de correlaciones donde se indica la relación entre variables, para así

detectar la importancia con la variable dependiente como se muestra en la figura 3.8.

Figura 3.8. Matriz de Correlación



Se observa que las variables manos, calibración superior, lote, son las que tienen un mayor porcentaje o las que mayor correlación positiva tienen por qué se acercan a 1.

La matriz de correlación (Ver Apéndice B de la página 59 - matriz de correlaciones) explica cómo se encuentran relacionadas cada una de las variables con otra variable.Los resultados de la correlación de las variables se pueden ubicar entre -1 y +1. Si estos elementos suben o bajan al mismo tiempo, el resultado de la correlación es positivo. Si un elemento sube y el otro baja o viceversa, entonces la correlación es negativa. De igual forma, valores cerca a cero indica que no existe una relación linealentre las variables, por lo tanto, es una matriz simétrica con unos en la diagonal (ya que la correlación entre una variable y ella misma es perfecta), en donde cada posición denota el coeficiente de correlación lineal de Pearson, que mide el grado de relación lineal entre cada par de elementos o variables. Esta prueba permite cuantificar la magnitud de la correlación entre dos variables y ayuda a predecir valores. Si estas variables tuvieran una correlación perfecta se podría inferir el valor de la variable y conociendo el valor de x. Debido a estas ventajas, la correlación es una de las pruebas más usadas en todo ámbito, ya que además de medir la dirección

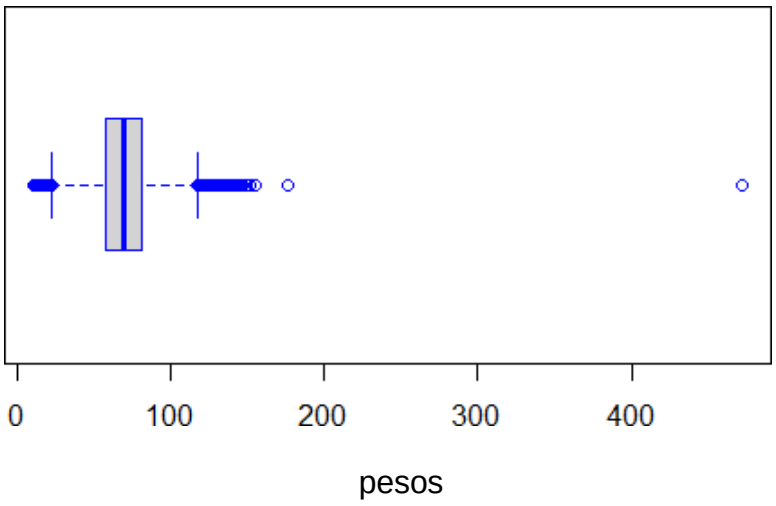
y magnitud de la asociación de dos variables, es uno de los fundamentos de los modelos de predicción, como los modelos de regresión lineal, random forest. (Ivonne Roy.2020)

3.3.Análisis y validación de los datos

En esta segunda parte se verifica los datos atípicos que tiene la variable dependiente (peso) en la cual, mediante la realización de diferentes funciones en R se detecta diferentes anomalías como las siguientes:

Diagrama de cajas

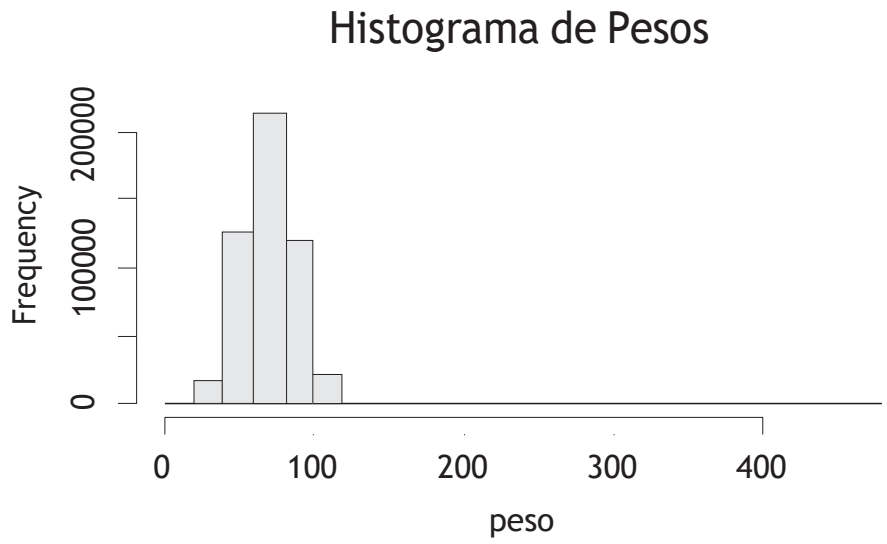
Figura 3.9. Diagrama de Caja (variable peso)



En el diagrama de Cajas se observa que los valores que se escapan a los «bigotes» se consideran valores extremos porque «pareciera» que no forman parte del grupo. Estos datos atípicos se pueden considerar mala práctica in situ por parte del operador al manipular e ingresar incorrectamente los datos al sistema BI de producción.

Histograma

Figura 3.10. Histograma (variable peso)



Media – Mediana – IQR - Cuantiles

Tabla 3.3
Media Mediana IQR Cuantiles

			QUANTILES				
MEDIANA	**MEDIA**	**IQR**	**0 %**	**25 %**	**50 %**	**75 %**	**100 %**
69,687	69,95	23,716	11,02	57,99	69,78	81,8	472

En la Figura 3.10 las barras en su distribución de datos indica que hay grupos de pesos en rango que mayores a 10, y menores 120 Libras aproximadamente, en la Tabla 3.3 se detallan datos como: la mediana de la variable peso es 69.68, la media 69.65 y el rango intercuartílico IQR se obtuvo 23.7, esto representa la diferencia del tercer cuartil con el primero y se entiende que en esta parte se encuentra el 50 % del total de los datos y por último se tiene la clasificación por quintiles de la variable peso que muestra los percentiles con su respectivo rango de valores.

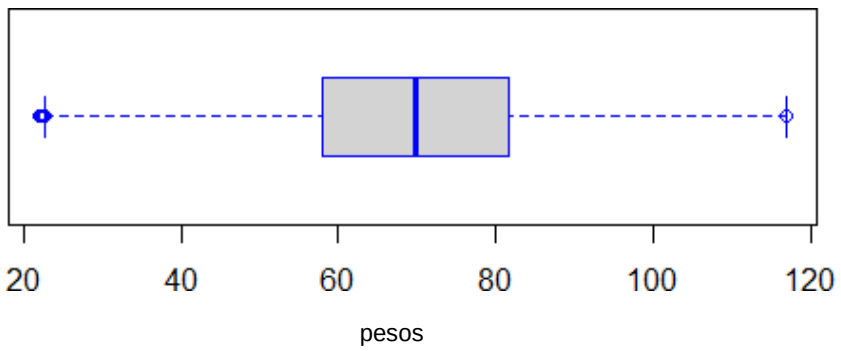
Una vez realizado el análisis del comportamiento de la variable dependiente (peso) urge tener mejor estructurado y consolidado los datos de esta variable, mediante comandos en R Studio se divide los datos de la variable peso en un mejor rango de valores en la cual solo se escogerá información de pesos < 117.28 libras y pesos > 22.41 libras.

Se obtuvo este rango de valores por que se trabajó con el IQR de la variable peso y arrojó 23.71 libras esto significa la diferencia del cuartil 3 menos el cuartil 1, para tener un tope estándar de peso de racimos y según lo indicado anteriormente se multiplica al IQR por 1.5 y se suma por el cuartil 3 (81.708).Como resultado se tiene un peso máximo de 117.28 libras, así mismo se realiza el ejercicio para peso mínimo de racimos, pero con el único cambio que se suma con el cuartil 1 y el peso mínimo aceptable es 22.41 libras.

Realizado todo esto, se crea un nuevo Dataframe llamado *nuevos_datos*, solo con el nuevo rango de valores de la variable peso, y para corroborar lo realizado se ejecuta las pruebas de análisis de datos y verificación de los respectivos cambios con las funciones de R ya conocidas:

Diagrama de cajas

Figura 3.11. Diagrama de Caja (variable peso)



Histograma

Figura 3.12. Histograma (variable peso)

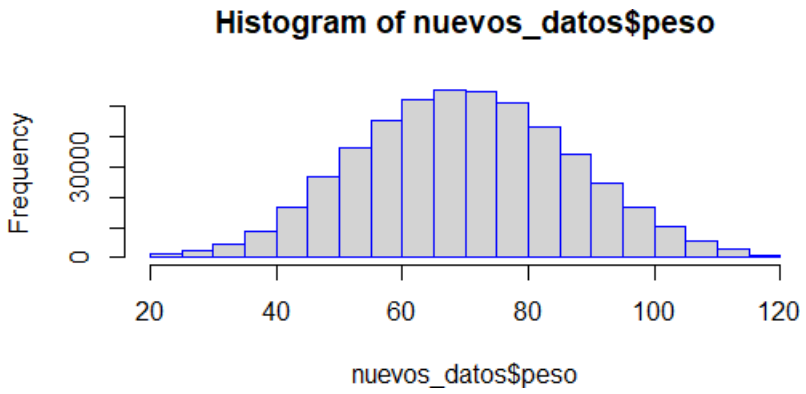


Tabla 3.4
Media - Mediana- IQR- Cuantiles

			QUANTILES				
MEDIANA	**MEDIA**	**IQR**	**0 %**	**25 %**	**50 %**	**75 %**	**100 %**
69,76	69,91	23,54	22	58,05	69,76	81,59	116,9

Como se observa en la nueva Figura 3.11 diagrama de cajas ya no contiene valores atípico o datos en los extremos y en el nuevo histograma 3.12, se obtuvo una buena distribución de sus barras indicando variaciones continuas aceptables, y por supuesto algo más que nos ayuda son las medidas estadísticas y se puede corroborar que el ultimo percentil es un valor adecuado (116.9 libras).

Recordar que toda la realización practica de esta sección de **Análisis y validación de Datos** se encuentra en el Apéndice B de la página 65.

3.4.Modelamiento y Métricas

3.4.1. Random Forest. Mediante la matriz de correlación se escoge las variables independientes para realizar el modelo predictivo para la variable dependiente pesos de racimos, una vez que se tiene bien delimitados las variables se crea un nuevo dataframe solo con las variables que se va a realizar los modelos, como lo muestra la siguiente Tabla 3.5.

Tabla 3.5
Variables escogidas

Variables
Edad
deschive F/2
deschante
selección
fertilización
riego
herbicidas
fumigación
peso

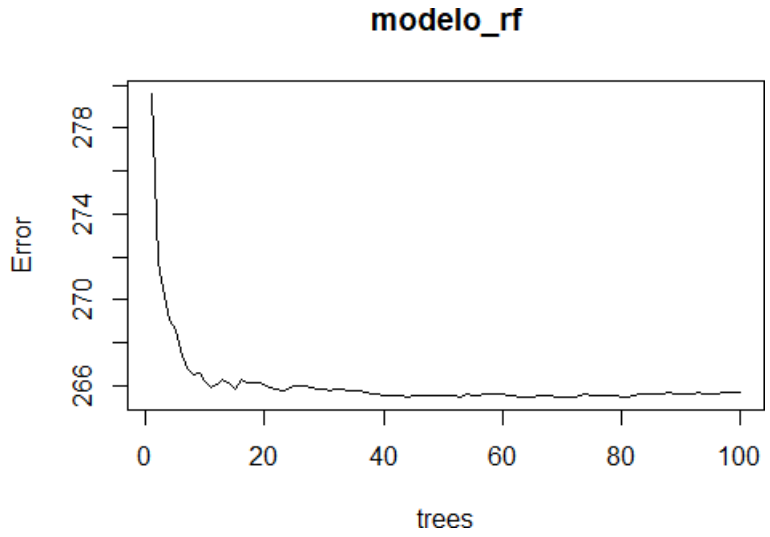
En el nuevo dataframe llamado (*data_modelamiento*) se divide en 2 conjuntos de datos (Ver anexo de la página 67), el primero será de entrenamiento (train) con 70 % aproximadamente de datos y el otro conjunto de datos para prueba (test) con el 30 % .En el conjunto de datos para prueba (test) se la divide en dos partes: un dataframe solo de variables predictoras (independientes) x_test y otra data solo la variable dependiente peso (y_test) y en el conjunto de datos de entrenamiento (*train*)se divide y_train solo con la variable dependiente.

Se crea el modelo predictivo con la librería *randomForest* y en sus parámetros en la opción *ntree* se empieza a crear modelo Random Forest (RF) con 100 árboles después se le aumentara hasta lograr una buena métrica del algoritmo RF, ejemplo:

```
library(randomForest)
modelo_rf = randomForest( peso ≃.,
data = train,
ntree =100
)
```

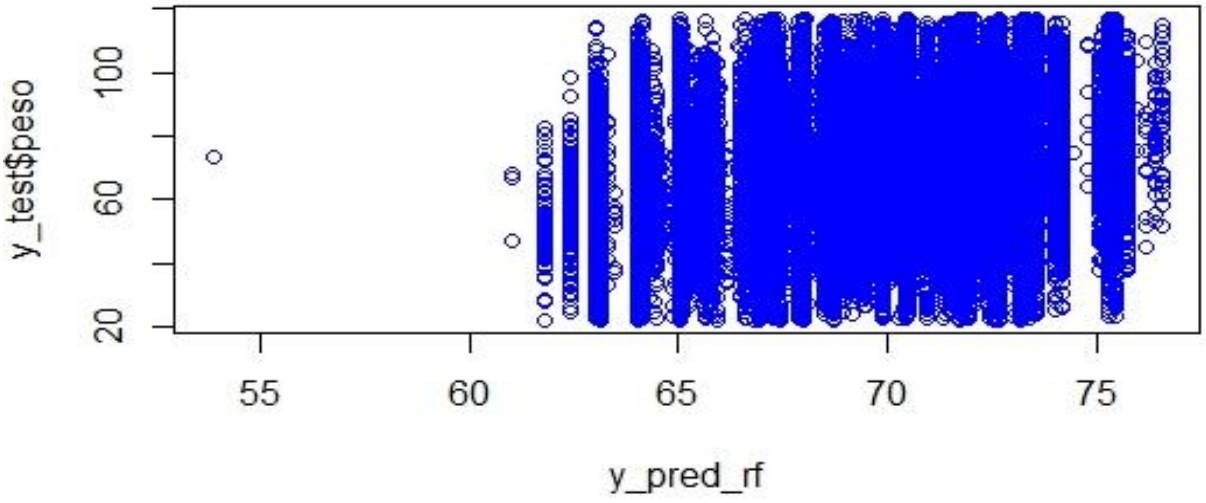
En esta primera instancia se tiene el primer modelo de RF como lo muestra la Figura 3.13

Figura 3.13. Modelo Random Forest /ntree =100



Una vez obtenido el primer modelo de RF con 100 árboles se observa que el error va disminuyendo, con la función *predict* se realiza la predicción de pesos de racimos incrustándole en sus parámetros el modelo de RF modelo_rf y el conjunto de datos x_test(variables independientes) , como resultado se obtiene la predicción de pesos de racimos en la cual lo comparamos con el conjunto de datos y_test(variable peso) como se observa en la siguiente en la Figura 3.14 en la cual los puntos se superponen y esto quiere decir que la predicción van a la par con los datos reales.

Figura 3.14. Modelo Random Forest



Mediante la función *importance* se obtiene la importancia de las variables que

ha contribuido en el modelo realizado que se detalla en la siguiente Tabla 3.6, entre cada vez mayor sea el resultado la variable es más importante en la creación del modelo.

Tabla 3.6
Importancia de variables en el modelo

Variables	IncNodePurity
deschive_f2	1.506.948,38
fertilización	1.358.940,77
fumigación	1.291.532,81
edad	483.435,13
riego	333.820,04
herbicidas	11.443,59
selección	8.917,99
deschante	8.775,63

Para validar y medir este modelo RF de regresión hay diferentes métricas como las que se muestran en la siguiente tabla 3.7:

Tabla 3.7
Métricas RF-NTREE = 100

Métricas para RF-ntree=100	
MAE	13,21
MSE	266,85
MAPE	21,4
RMSE	16,33
R2	0,059

Se Observa en el parámetro de Ntree = 100 árboles el Error Absoluto Medio (MAE) se encuentra elevado, entonces se procede a realizar las pruebas aumentando el número de árboles al algoritmo RF a Ntree = 200 y se obtiene lo siguientes resultados:

Figura 3.15. Modelo Random Forest /ntree =200

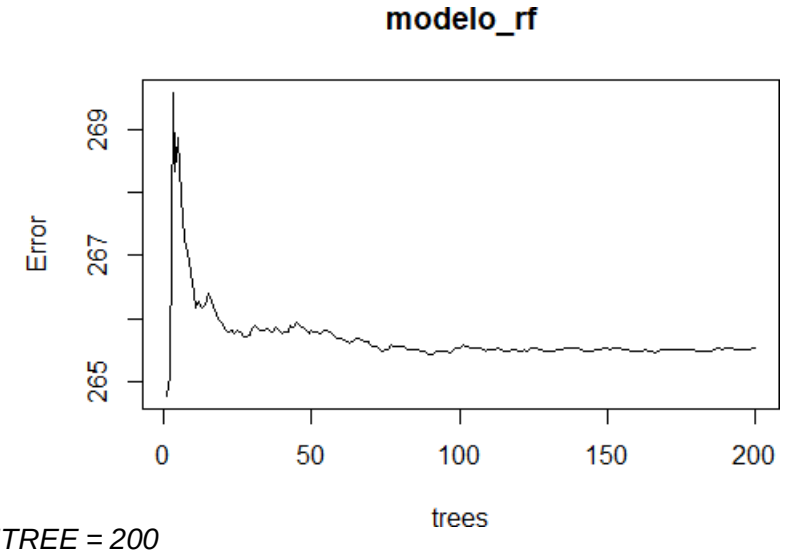


Tabla 3.8
Métricas RF-NTREE = 200

Métricas para RF-ntree=200	
MAE	13,2
MSE	266,77
MAPE	21,39
RMSE	16,33
R2	0,06

Se Observa que no hay mucha variación en la medición de errores del algoritmo y en este caso se elige el modelo que se entrenó con Ntree = 200 árboles.

Una de las métricas más importante para este caso de estudio, es la métrica del Error Absoluto Medio que indica que el error en el modelo predictivos es de 13.2 libras, en la predicción de peso de racimos, cabe recalcar que no es necesariamente que todas las métricas deben ajustar a unos buenos indicadores pues esto dependerá mucho del caso de estudio, la realización de este modelo RF se puede verificar en el Anexo de la página 81.

3.4.2. Xgboost. Para realizar este modelo predicción se instala los paquetes y librerías *library(xgboost)* seguido a esto el dataFrame de entrenamiento se la convierte en una matriz con la función *data.matrix* como se tiene en el Anexo de la página 81, con la función de *xgboost* y en sus parámetros *nround* como parte inicial se le ingresa un cantidad de 50, este número son iteraciones que se realizarán

antes de detener el proceso de ajuste. Un mayor número de iteraciones generalmente devuelve mejores resultados de predicción, como el siguiente ejemplo del algoritmo de Xgboost.

modelo_xgb = xgboost(data = data.matrix(train[,-9]),

eval_metric = rmse",

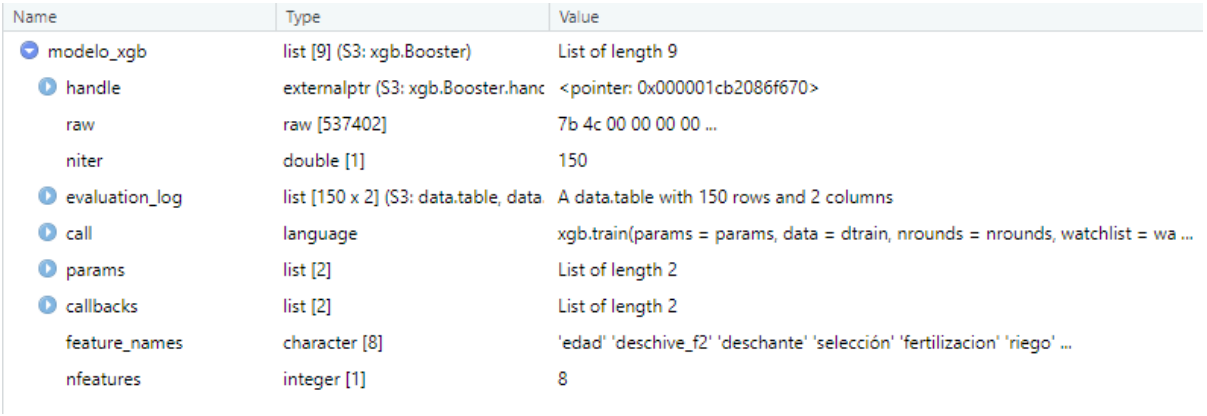
label = y_train$peso,

reg_lambda = 0.5,

nrounds = 50)

En este algoritmo Xgboost tiene algunos parámetros, pero los más importantes son los que mencionamos en el código, una vez realizado el modelo verificamos los resultados del mismo mediante su descripción con la función *print(modelo_xgb)* como se muestra en la siguiente figura 3.16

Figura 3.16. Descripción del modelo



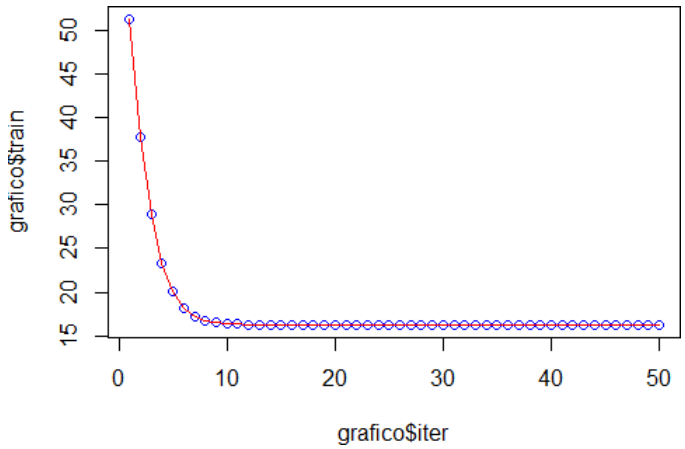
Name	Type	Value
modelo_xgb	list [9] (S3: xgb.Booster)	List of length 9
handle	externalptr (S3: xgb.Booster.hanc	<pointer: 0x000001cb2086f670>
raw	raw [537402]	7b 4c 00 00 00 00 ...
niter	double [1]	150
evaluation_log	list [150 x 2] (S3: data.table, data	A data.table with 150 rows and 2 columns
call	language	xgb.train(params = params, data = dtrain, nrounds = nrounds, watchlist = wa ...
params	list [2]	List of length 2
callbacks	list [2]	List of length 2
feature_names	character [8]	'edad' 'deschive_f2' 'deschante' 'selección' 'fertilizacion' 'riego' ...
nfeatures	integer [1]	8

En los respectivos detalles se muestra los diferentes parámetros con lo que se ejecutó el algoritmo como el siguiente ***eval_metric = rmse"***, que indica que las métricas de evaluación van a ser de tipo de errores, otro parámetro es *reg_lambda = 0.5* este sirve para guiar al algoritmo, al objetivo y reducir el error, entre mayor sea el numero menos error, pero no es recomendable que sea mayor a 1 y el *nrounds = 50*, como se mencionó anteriormente son la cantidad de secuencias o iteraciones que realiza el algoritmo para obtener un mejor resultado, también se muestra la cantidad de variables que tiene la matriz e incluso desde ya tenemos los resultados de la métricas

del algoritmo para su evaluación como lo es el RMSE.

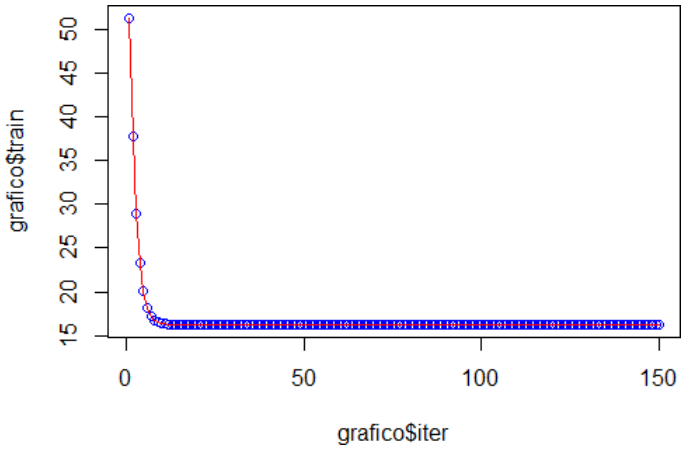
En la siguiente Figura 3.17 es fundamental hacer notar como desciende el error a medida que se desarrollan las 50 iteraciones

Figura 3.17. Curva de error modelo XGBoost 50 nround



En la siguiente Figura 3.18 se subió el número de iteraciones *nround* a 150 y se tiene el mismo panorama y el error llego a RMSE = 16.23.

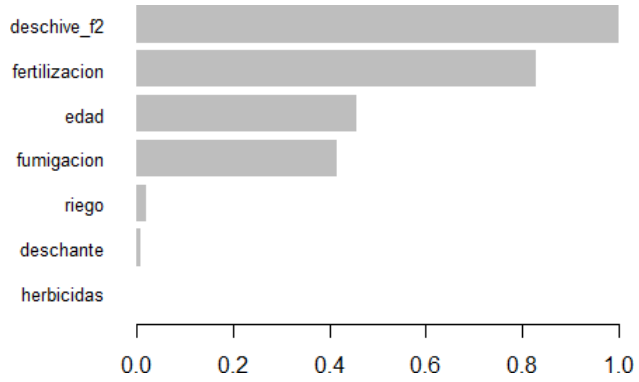
Figura 3.18. Curva de error modelo XGBoost 150 nround



Mediante la ejecución de una función que permite o detalla las variables mejores puntuadas que aportaron en la realización del modelo, como lo muestra en

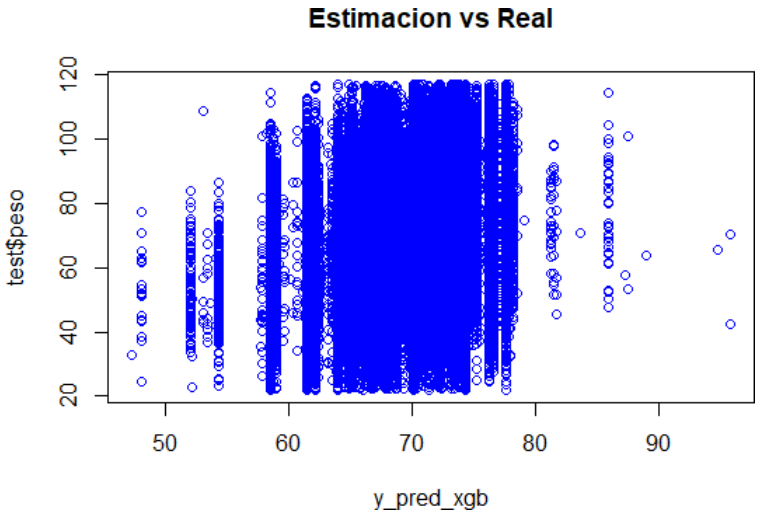
la siguiente Figura 3.19 y las variables importantes fueron deschive_f2, fertilización, edad, fumigación

Figura 3.19. Importancia de las variables



Una vez que se ha realizado el modelo con el Algoritmo de Xgboost viene la predicción de pesos en la cual, mediante la función *predict* en sus argumentos, se ingresa los dataset del modelo y los de prueba, como resultado se obtiene la predicción de pesos de racimos, dicha información se compara con los datos de prueba y se obtiene un gráfico como el de la Figura 3.20. Se observan datos superpuestos y a su vez datos un poco más abiertos a los datos reales.

Figura 3.20. Estimación vs Real-XGBoost



Como todas métricas de evaluación Xgboost no es la excepción y son las métricas de errores como los siguientes algoritmos tiene MSE-RMSE-MAE–MAPE, se calcula para 50 - 150 iteraciones como lo muestran las tablas 3.9 y 3.10:

Tabla 3.9
Metricas para xgboost-nrounds=50

Metricas xgboost - nrounds=25	
MAE	13,13
MSE	264,51
MAPE	21,21
RMSE	16,23

Tabla 3.10
Métricas para xgboost-nrounds=150

Metricas para XGboost-nrounds=150	
MAE	13,1
MSE	262,92
MAPE	21,17
RMSE	16,2

Se observa que no se tiene mucha variación tanto para 50 iteraciones que 150 iteraciones, para este caso la métrica MAE indica que el error Absoluto medio de la predicción de peso de racimos es de 13.1 libras.

Una vez realizado estos dos modelos de predicción tanto de RF y Xgboost y según el análisis y resultados de sus métricas, el algoritmo que ayudaría a dar mejor resultado de predicción es el Xgboost, aunque no tiene una amplia diferencia con los resultados del algoritmo RF como se muestra en las siguientes tablas 3.11 y 3.12:

Tabla 3.11
Métricas para el modelo de random forest

Métricas para RF-ntree=200	
MAE	13,2
MSE	266,77
MAPE	21,39
RMSE	16,33

Tabla 3.12
Métricas para el modelo XGBoost

Métricas para XGboost-nrounds=150	
MAE	13,1
MSE	262,92
MAPE	21,17
RMSE	16,23

3.5.Dependencia de Variables

La dependencia parcial ayuda a comprender el efecto marginal de una característica (o un subconjunto de la misma) en el resultado previsto. En esencia, permite comprender cómo cambia la variable de respuesta a medida que se cambia el valor de una característica (variables independientes), teniendo en cuenta el efecto promedio de todas las demás características del modelo y una vez definido el modelo que resulto con mejores métricas es el Xgboost se realizó los grafico de dependencia parcial donde se muestra como varia el peso de los racimos (Variable Objetivo) cuando camba los valores de las variables más importantes del modelo predictivo (Xgboost) o variables independientes.

Figura 3.21. DP-Fumigación

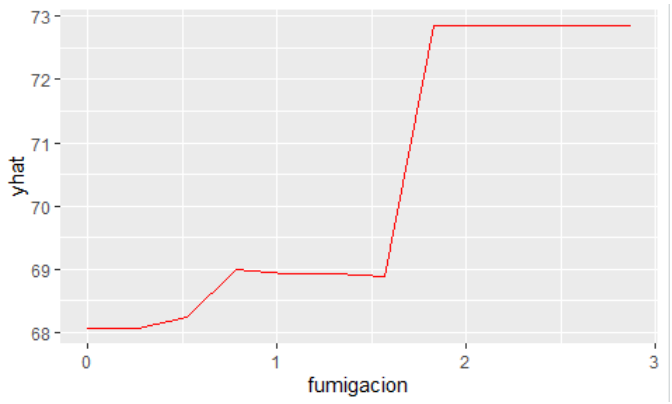


Figura 3.22. DP-Fertilización

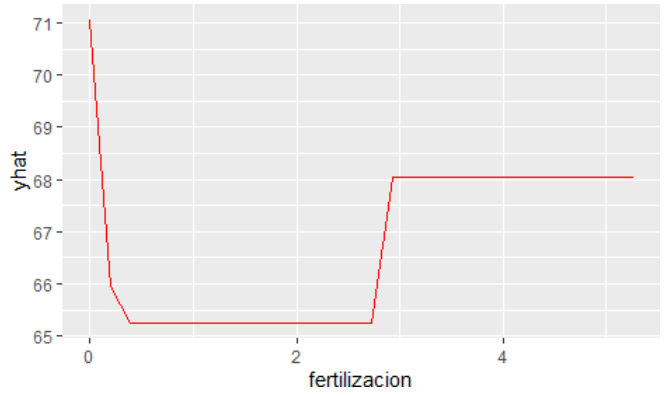


Figura 3.23. DP-Deschive_f2

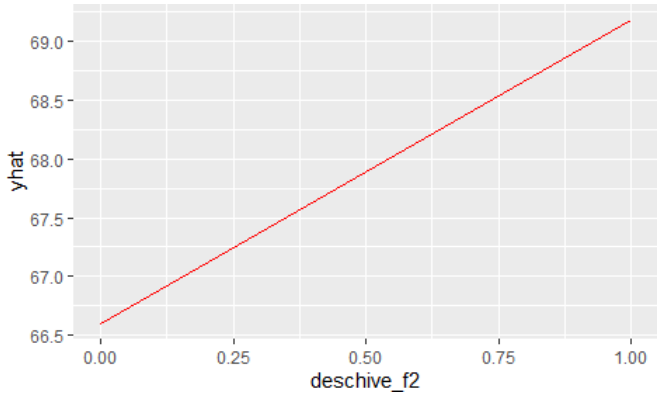
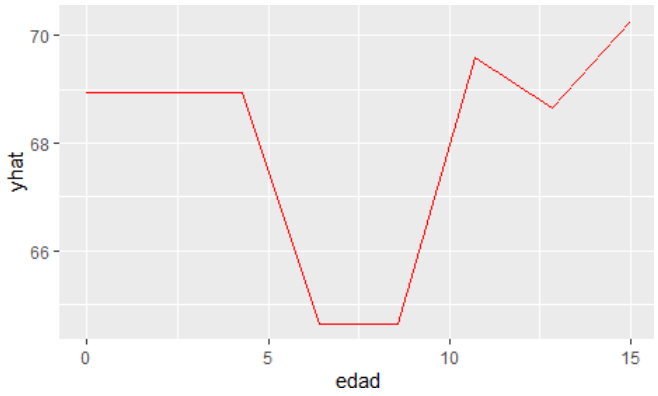


Figura 3.24. DP-Edad



CONCLUSIONES Y TRABAJO FUTURO

Al termino de este trabajo investigativo se pudo obtener modelos predictivos basados en machine learning, que fueran capaces de realizar buenas predicciones de pesos de racimos. También es importante analizar la afectación al modelo cada variable utilizada para entrenar, en términos generales se logró demostrar que los algoritmos robustos como Random Forest Regression y Xgboost pueden predecircon un muy buen desempeño, siempre que se tengan los datos y variables necesarias para entrenar los modelos.

Una vez desarrollados los distintos modelos, el mejor algoritmo predictivo fue Xgboost que a diferencia del Random Forest Regression en sus métricas de evaluación de errores para una buena predicción fue un poco más bajo, como en la métrica MAE (error absoluto medio) para Xgboost fue de 13.1 y para RF fue de 13.2 (Tabla 3.11 y Tabla 3.12) como se puede observar la diferencia fue por centésimas en la cual se escoge al algoritmo Xgboost como el mejor y según mis investigaciones realizadas este algoritmo al momento de su ejecución procesa la información de manera muy diferente y más eficiente que el algoritmo de Random Forest Regression y a su vez también ha permitido definir las variables influyentes de alto impacto para la predicción de pesos de racimos como lo son: variable de fertilización, fumigación(sigatoka), deschive_f2, en la cual se debe poner mucho énfasis en la realización de las mismas.

RECOMENDACIONES

Con los resultados obtenidos en esta investigación se pueden llegar a modelos predictivos que incorporen un mayor conjunto de variables de diferente naturaleza relacionada al mundo bananero, en la cual sus mediciones deberían realizarse por sensores que estén inmerso a la producción bananera con el objetivo de analizar esta misma problemática u otros problemas similares y así se tendría mejores modelos predictivos de pesos de racimos para ayudar en el desarrollo productivo de la Hacienda Bananera.

BIBLIOGRAFIA GENERAL

Allouhi, A., Choab, N., Hamrani, A., and Saadeddine, S. (2021). Machine learning algorithms to assess the thermal behavior of a moroccan agriculture greenhouse. *Cleaner Engineering and Technology*, 5:100346.

Arteaga, J. J. G., Zambrano, J. J. Z., Cevallos, R. A., and Romero, W. D. Z. (2020). Predicción del rendimiento de cultivos agrícolas usando aprendizaje automático. *Revista Arbitrada Interdisciplinaria Koinonía*, 5(2):144–160.

Balducci, F., Impedovo, D., and Pirlo, G. (2018). Machine learning applications on agricultural datasets for smart farm enhancement. *Machines*, 6(3):38.

Bernal Pablo, P. (2018). *La Investigación en Ciencias Sociales: Técnicas de recolección de la información*.

Cantero Díaz, A., Goire Castilla, M. M., and Quintana Cassulo, Y. (2019). Sistema para la gestión y análisis de datos de una red de sensores inalámbricos basado en un almacén de datos. *Revista Cubana de Ciencias Informáticas*, 13(3):76–90.

Cedric, L. S., Adoni, W. Y. H., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M., and Kimpolo, C. L. M. (2022). Crops yield prediction based on machine learning models: Case of west african countries. *Smart Agricultural Technology*, page 100049.

Chandraprabha, M. and Dhanaraj, R. K. (2020). Machine learning based pedantic analysis of predictive algorithms in crop yield management. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1340–1345.

Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture*, 151:61–69.

Crisóstomo Fernández, F. L., Lajo Aurazo, A. S., Hernández Quiroz, G. V., Asencio Diaz, L. d. l. A. M., and Chiang Cornejo, R. H. (2021). Técnicas de machine learning para la clasificación automática de clientes en una empresa de seguros.

Dadas, S., Protasiewicz, J., and Pedrycz, W. (2019). A deep learning model with data enrichment for intent detection and slot filling. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 3012–3018. IEEE.

Deepa, S., Alli, A., Gokila, S., et al. (2021). Machine learning regression model for material synthesis prices prediction in agriculture. *Materials Today: Proceedings*.

Degfie, T. A., Mamo, T. T., and Mekonnen, Y. S. (2019). Optimized biodiesel production from waste cooking oil (wco) using calcium oxide (cao) nano-catalyst. *Scientific reports*, 9(1):1–8.

Developers, S.-L. (2021). Metrics and scoring: Quantifying the quality of predictions. *User Guide,[entre 2007 e 2019]. Disponível em: https://scikit-learn. org/stable/modules/model_evaluation. html. Acesso em*, 26.

Eulogio, R. (2017). Introduction to random forests. *Oracle+ DataScience. com*.

González, C. A. G. and Hernandez, V. (2020). Clasificador de productos agrícolas para control de calidad basado en machine learning e industria 4.0. *Revista Perspectivas*, 2(2):21–28.

Herrera-Díaz, C. (2016). Implementación de un módulo de análisis estadístico y predictivo para agricultura utilizando bigdata y machine learning, integrado al sistema iotmach.[implementation of a statistical and predictive analysis module for agriculture using bigdata and machine learning, integrated to the iotmach system]. *Trabajo de titulación. Carrera de ingeniería de sistemas. Universidad Técnica de Machala. Recuperado de https*, (9).

Jiménez, J. U. (2019). Introducción a r y rstudio.

Kotu, V. and Deshpande, B. (2015). Data mining process. *Predictive analytics and data mining*, 1:17–36.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., and Engelhardt, A. (2016). Caret: classification and regression training package. *R package version*, pages 6–0.

León Serrano, L. A., Arcaya Sisalima, M. F., Barbotó Velásquez, N. A., and Bermeo Pineda, Y. L. (2021). Ecuador: Análisis comparativo de las exportaciones de banano orgánico y convencional e incidencia en la balanza comercial, 2018.

Lopez Briega, R. (2015). Machine learning con python.

Maduranga, M. and Abeysekera, R. (2020). Machine learning applications in iot based agriculture and smart farming: A review. *Int. J. Eng. Appl. Sci. Technol*, 4(12):24–27.

Marqués Gozalbo, M. Á. (2022). Modelos predictivos de producción agroindustrial con machine learning a partir de fuentes de información pública.

Meshram, V., Patil, K., Meshram, V., Hanchate, D., and Ramkteke, S. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, 1:100010.

Mohd Shafri, H. Z. M. S. and Arenas París, C. (2019). Artificial intelligence (ai) in oil palm remote sensing applications.

Ortega, A. O. (2018). Enfoques de investigación. *Métodos para el diseño urbano–Arquitectónico*.

Pallares Cabrera, F. (2015). Desarrollo de un modelo basado en machine learning para la predicción de la demanda de habitaciones y ocupacion en el sector hotelero.

Pereyra, L. E. (2020). *Metodología de la investigación*.

Rezk, N. G., Hemdan, E. E.-D., Attia, A.-F., El-Sayed, A., and El-Rashidy, M. A. (2021). An efficient iot based smart farming system using machine learning algorithms. *Multimedia Tools and Applications*, 80(1):773–797.

Slob, N., Catal, C., and Kassahun, A. (2021). Application of machine learningto improve dairy farm management: A systematic literature review. *Preventive Veterinary Medicine*, 187:105237.

Swami, D., Shah, A. D., and Ray, S. K. (2020). Predicting future sales of retail products using machine learning. *arXiv preprint arXiv:2008.07779*.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*.

Villafuerte Chacnama, F. F. (2021). Análisis comparativo de modelos de pronóstico arima y xgboost aplicados a las series mensuales de ventas en una empresa certificadora.

Wickham, H. and Grolemund, G. (2017). R for data science: Import. *Tidy, transform, visualize, and model data*, 1.

APENDICE

Apéndice: Instalación de R Studio

Para instalar R es necesario visitar la página web ubicada en el Uniform Resource Locator (URL) o dirección web www.r-project.org (González, 2017), y posteriormente bajar e instalar el paquete. En el margen izquierdo la página contiene una liga con la leyenda, “Download, Packages and CRAN”.

Una vez que se le da click al link, se dirige a otra página con los principales servidores o mirrows de R en el mundo. Desde luego, hay que escoger aquí a un servidor que esté lo suficientemente cerca a la ubicación geográfica del ordenador, Acorde a lo anterior, se elige en enlace correspondiente al servidor, a través página web de la misma dirección URL y se da click en el enlace ”The Comprehensive R Archive Network” (González, 2017), en la sección titulada Download and Install R se encuentra tres enlaces según el Sistema Operativo del Ordenador

Instalación para Sistema Operativo Windows

En un sistema Windows se le debe de dar doble click al archivo .exe, en modo de administrador. El archivo al tiempo de este escrito se llama como sigue:

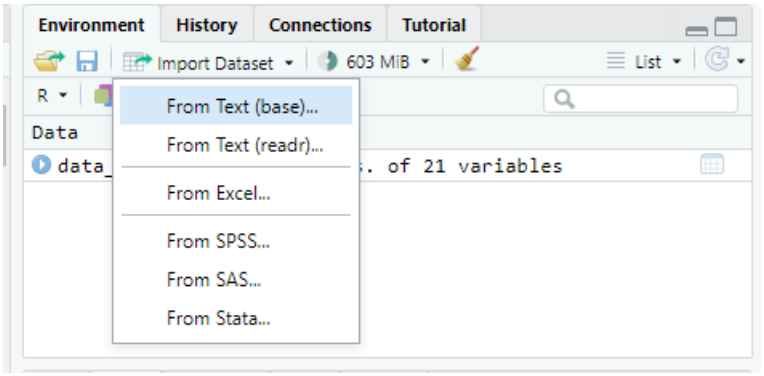
R-4-2-0-win.exe

Luego, se le da doble click, después de lo cual se siguen las instrucciones y después de pasar por varias opciones de configuración, el programa R queda instalado

Apéndice B: Carga y Ajuste de datos R Studio

Para realizar la carga de datos se realiza por consola de comandos o con el asistente para importar datos, en el proyecto se utiliza el asistente tal como sigue:

Figura 25. Carga de Datos en RStudio



Fuente.- R Studio

O por comandos como el siguiente:

Cargar datos

library(readxl)

library(tidyverse)

library(readr) data_tesis_pedro ← read_excel("F:\\Modulos Maestria\\Tesis\\Datos de Tesis\\tesis 2022-20220323T002110Z-001\\tesis 2022\\consulta racimos(2).xlsx")

Visualizacion del Dataframe

view(data_tesis_pedro)

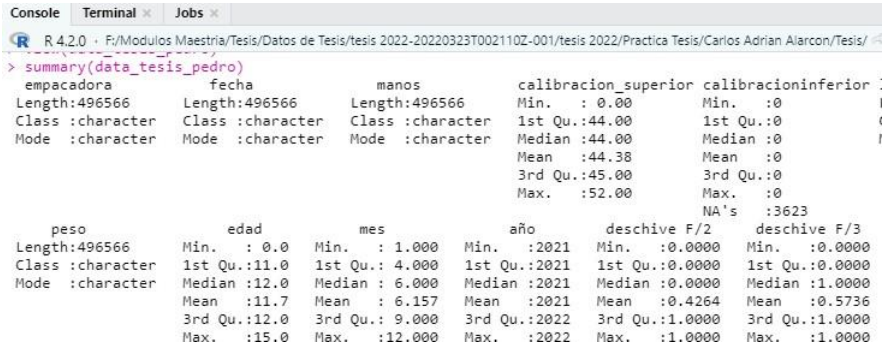
Figura 26. Dataframe

data_tesis_pedro | codigo_pedro(xgboost).R | codigo_pedroultimo.R

	edad	mes	año	deschive F/2	deschive F/3	deshoje	deschante	selección	fertilizacion
0	12	3	2021	1	0	1	0	0	0.26448
0	14	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	12	3	2021	1	0	1	0	0	0.26448
0	14	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448
0	13	3	2021	1	0	1	0	0	0.26448

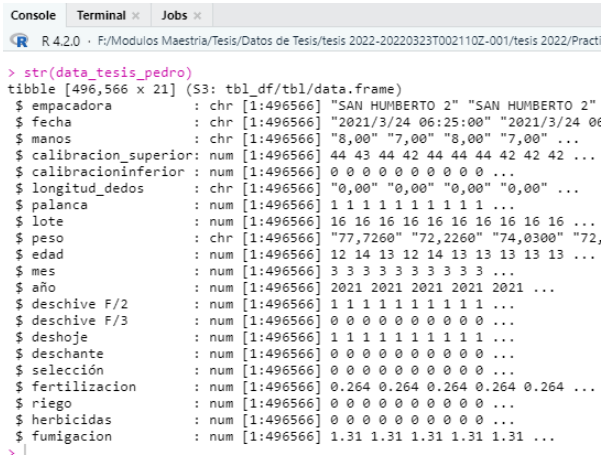
Fuente.- R Studio

Figura 27. Summary



Fuente.- R Studio

Figura 28. str



Fuente.- R Studio

Modificaciones en Variables

Cambiar el punto por la coma y transformar a numérico

data_tesis_pedro$peso← as.numeric(gsub(",", ".",data_tesis_pedro$peso))

data_tesis_pedro$manos<-as.numeric(gsub(",", .",data_tesis_pedro$manos))

Cambiar el nombre de una variable

data_tesis_pedro = data_tesis_pedro %>% rename(deschive_f2 =' deschive # F/2' , deschive_f3 = 'deschive #

Variables importantes

```
data_tesis_pedro = data_tesis_pedro %> % select( manos,
                                                 palanca,
                                                 calibracion_superior,
                                                 deschive_f3,
                                                 deschive_f2,
                                                 edad,
                                                 lote,
                                                 deshoje,
                                                 deschante,
                                                 selección,
                                                 fertilizacion,riego,
                                          herbicidas,fumigacion, peso )
```

Valores nulos

######## Revisión de datos en blanco

table(is.na(data_tesis_pedro$peso))

table(is.na(data_tesis_pedro$manos))

table(is.na(data_tesis_pedro$calibracion_superior))

table(is.na(data_tesis_pedro$palanca))

table(is.na(data_tesis_pedro$peso))

table(is.na(data_tesis_pedro$edad))

table(is.na(data_tesis_pedro$deschive_f2))

table(is.na(data_tesis_pedro$deschive_f3))

table(is.na(data_tesis_pedro$deshoje))

table(is.na(data_tesis_pedro$deschante))

table(is.na(data_tesis_pedro$selección))

table(is.na(data_tesis_pedro$fertilizacion))

table(is.na(data_tesis_pedro$riego))

table(is.na(data_tesis_pedro$herbicidas))

table(is.na(data_tesis_pedro$fumigacion))

Valores Nulos - Resultados

Figura 29. Valores Nulos - Resultados

```
> table(is.na(data_tesis_pedro$deschive_f2))

 FALSE
496566
> table(is.na(data_tesis_pedro$deschive_f3))

 FALSE
496566
> table(is.na(data_tesis_pedro$deshoje))

 FALSE
496566
> table(is.na(data_tesis_pedro$deschante))

 FALSE
496566
> table(is.na(data_tesis_pedro$selección))

 FALSE
496566
> table(is.na(data_tesis_pedro$fertilizacion))

 FALSE
496566
```

Correlación

cor(data_tesis_pedro$peso, data_tesis_pedro$manos)

cor(data_tesis_pedro$peso, data_tesis_pedro$calibracion_superior)

cor(data_tesis_pedro$peso, data_tesis_pedro$calibracioninferior)

cor(data_tesis_pedro$peso, data_tesis_pedro$palanca)

cor(data_tesis_pedro$peso, data_tesis_pedro$lote)

cor(data_tesis_pedro$peso, data_tesis_pedro$peso)

cor(data_tesis_pedro$peso, data_tesis_pedro$edad)

cor(data_tesis_pedro$peso, data_tesis_pedro$deschive_f2)

cor(data_tesis_pedro$peso, data_tesis_pedro$deschive_f3)

cor(data_tesis_pedro$peso, data_tesis_pedro$deshoje)

cor(data_tesis_pedro$peso, data_tesis_pedro$deschante)

cor(data_tesis_pedro$peso, data_tesis_pedro$selección)

cor(data_tesis_pedro$peso, data_tesis_pedro$fertilizacion)

cor(data_tesis_pedro$peso, data_tesis_pedro$riego)

cor(data_tesis_pedro$peso, data_tesis_pedro$herbicidas)

cor(data_tesis_pedro$peso, data_tesis_pedro$fumigacion)

cor(data_tesis_pedro$peso, data_tesis_pedro$lote)

Correlacion-Resultado

Figura 30. Correlacion-Resultado

```
> cor(data_tesis_pedro$peso, data_tesis_pedro$manos)
[1] 0.7454231
> cor(data_tesis_pedro$peso, data_tesis_pedro$calibracion_superior)
[1] 0.1958994
> cor(data_tesis_pedro$peso, data_tesis_pedro$palanca)
[1] 0.03035807
> cor(data_tesis_pedro$peso, data_tesis_pedro$lote)
[1] 0.1190231
> cor(data_tesis_pedro$peso, data_tesis_pedro$peso)
[1] 1
> cor(data_tesis_pedro$peso, data_tesis_pedro$edad)
[1] -0.04005629
> cor(data_tesis_pedro$peso, data_tesis_pedro$deschive_f2)
[1] 0.1626649
> cor(data_tesis_pedro$peso, data_tesis_pedro$deschive_f3)
[1] -0.1626649
> cor(data_tesis_pedro$peso, data_tesis_pedro$deschante)
[1] -0.02621184
> cor(data_tesis_pedro$peso, data_tesis_pedro$selección)
[1] -0.02621184
> cor(data_tesis_pedro$peso, data_tesis_pedro$fertilizacion)
[1] 0.004281288
```

correlación entre peso y manos

plot(data_tesis_pedro$peso, data_tesis_pedro$manos, pch = 19, col = ”light-blue”)

abline(lm(data_tesis_pedro$manos~ data_tesis_pedro$peso), col = red”, lwd = 3)

text(paste(Çorrelación:”, round(cor(data_tesis_pedro$peso, data_tesis_pedro$manos), 2)), x = 25, y = 95)

correlación entre peso y fumigación

plot(data_tesis_pedro$peso, data_tesis_pedro$fumigacion, pch = 19, col = ”lightblue”)

abline(lm(data_tesis_pedro$fumigacion~ data_tesis_pedro$peso), col = red",
lwd = 3)

text(paste(Çorrelación:", round(cor(data_tesis_pedro$peso, data_tesis_pedro$fumigacion),
2)), x = 25, y = 95)

Correlación entre peso y herbicidad

plot(data_tesi_pedro$peso, data_tesis_pedro$herbicidas, pch = 19, col = "light-
blue")

abline(lm(data_tesis_pedro$herbicidas~ data_tesis_pedro$peso), col = red",
lwd = 3)

text(paste(Çorrelación:", round(cor(data_tesis_pedro$peso, data_tesis_pedro$herbicidas),
2)), x = 25, y = 95)

correlación entre peso y deschive_f2

plot(data_tesis_pedro$peso, data_tesis_pedro$deschive_f2, pch = 19, col =
"lightblue")

abline(lm(data_tesis_pedro$deschive_f2~ data_tesis_pedro$peso), col = red",
lwd = 3)

text(paste(Çorrelación:", round(cor(data_tesis_pedro$peso, data_tesis_pedro$deschive_f2),
2)), x = 25, y = 95)

correlación entre peso y riego

plot(data_tesis_pedro$peso, data_tesis_pedro$riego, pch = 19, col = "lightblue")

abline(lm(data_tesis_pedro$riego~ data_tesis_pedro$peso), col = red", lwd = 3)

text(paste(Çorrelación:", round(cor(data_tesis_pedro$peso, data_tesis_pedro$riego),
2)), x = 25, y = 95)

correlación entre peso y seleccion

plot(data_tesis_pedro$peso, data_tesis_pedro$selección, pch = 19, col = "light-
blue")

abline(lm(data_tesis_pedro$selección~ data_tesis_pedro$peso), col = red", lwd

= 3)

text(paste(Çorrelación:", round(cor(data_tesis_pedro$peso, data_tesis_pedro$selección), 2)), x = 25, y = 95)

Matriz correlación

library(ggcorrplot)

corr ← round(cor(data_tesis_pedro,method = "kendall"), 1)

corr

ggcorrplot(corr) +

ggtitle(Çorrelograma del conjunto ") +

theme_minimal()

ggcorrplot(corr, method = 'circle') +

ggtitle(Çorrelograma del conjunto ") +

theme_minimal() ggcorrplot(corr, method = 'circle', type = 'lower') +

ggtitle(Çorrelograma del conjunto ") +

theme_minimal()

ggcorrplot(corr, method = 'circle', type = 'lower', lab = TRUE) +

ggtitle(Çorrelograma del conjunto ") +

theme_minimal() +

theme(legend.position="none")

Apéndice C: Análisis y validación de datos

Diagrama de cajas

boxplot(data_tesis_pedro$peso, border = c("blue"),horizontal = TRUE)

boxplot(data_tesis_pedro$peso ~ data_tesis_pedro$manos, horizontal = TRUE,border = c("blue", "green"))

Histograma

hist(data_tesis_pedro$peso, main = ”Histograma de Pesos”)

IQR – media- mediana-quantiles

mean(data_tesis_pedro$peso)

median(data_tesis_pedro$peso)

IQR(data_tesis_pedro$peso)

quantile(data_tesis_pedro$peso)

Parametrización de la Variable Peso

$81.708 + 1.5 * IQR(data\_tesis\_pedro\$peso)$

$81.708 + 1.5 * IQR(data\_tesis\_pedro\$peso)$

nuevos_datos = data_tesis_pedro[data_tesis_pedro$peso >= 22 &

data_tesis_pedro$peso <= 117,]

#verificacion del comprtamiento de la variable peso

boxplot(nuevos_datos$peso, border = c(”blue”),horizontal = TRUE)

hist(nuevos_datos$peso,border = c(”blue”))

mean(nuevos_datos$peso)

median(nuevos_datos$peso)

IQR(nuevos_datos$peso)

quantile(nuevos_datos$peso)quantile

Apéndice D: Modelamiento y métricas

Nuevo dataframe con variables escogidas

```
data_modelamiento = nuevos_datos %> % select( edad
                                              deschive_f2,
                                              deschante,
                                              selección,
                                              fertilización,
                                              riego,
                                              herbicidas,
                                              fumigación,
                                              peso )
```

División del Dataframe

```
ind = sample(2,nrow(data_modelamiento),replace = TRUE,prob = c(0.7,0.3))
train = data_modelamiento[ind ==1,]
test = data_modelamiento[ind==2,]
y_train = train[,9]
x_test = test[,-9] #variables independientes menos peso
view(x_test)
y_test = test[,10] # variable dependiente solo peso
view(y_test)
```

Creación del Modelo Random Forest RF

```
 library(randomForest)
modelo_rf = randomForest( peso ~.,
                          data=train
                          ntree=100
                          )
```

Resultado del modelo

plot(modelo_fr)

Predicción

y_pred_rf = predict(modelo_rf,x_test)

Comparación real vs estimados de pesos

plot(y_pred_rf, y_test$peso,main = .Estimacion vs Real", col = "blue")

Importancia de variables

importance(modelo_rf)

Métricas para el modelo RF

mse = mean((y_pred_rf - y_test$peso)2)

mse

rmse = sqrt(mean((y_pred_rf - y_test$peso)2))

print(rmse)

mae = mean(abs(y_pred_rf - y_test$peso))

mae

mape = mean(abs((y_test$peso - y_pred_rf)/y_test$peso))∗100

mape

r2 = 1 - sum((y_test$peso - y_pred_rf)2)/sum((y_test$peso - mean(y_test$peso))2)

r2

Creación del Modelo Xgboost

XGBoost

library(xgboost)

library(caret)

library(car)

```
modelo_xgb = xgboost(data = data.matrix(train[,-9]),
                     eval_metric = rmse",
                     label = y_train$peso,
                     reg_lambda = 0.5,
                     nrounds = 150 )
```

Características del modelo

```
print(modelo_xgb)
```

Gráfico de evaluación del modelo

```
grafico=data.frame(modelo_xgb$evaluation_log)
plot(grafico$iter, grafico$train, col = 'blue')
lines(grafico$iter, grafico$train, col = 'red')
```

Importancia de variables

```
xgb.importance(colnames(data.matrix(train[,-9])), model = modelo_xgb) impor-
tance=xgb.importance(feature_names = colnames(data.matrix(train[,-9])),model=
modelo_xgb)
xgb.plot.importance(importance_matrix = importance,rel_to_first = T)
```

Predicción

```
y_pred_xgb = predict(modelo_xgb,data.matrix(test[,-9]))
```

Comparación real vs estimados de pesos

```
plot( y_pred_xgb,test$peso,main = .Estimacion vs Real",
col = "blue")
```

Métricas para el modelo xgboost

rmse_xgb = sqrt(mean((y_pred_xgb-y_test$peso)2))

rmse_xgb

mse_xgb = mean((y_pred_xgb - y_test$peso)2)

mse_xgb

mae_xgb = mean(abs(y_pred_xgb - y_test$peso))

mae_xgb

mape_xgb = mean(abs((y_test$peso-y_pred_xgb)/y_test$peso))∗100

mape_xgb

Dependencia parcial para el modelo xgboost

library(pdp)

library(ggplot2)

library(magrittr)

library(ggfortify)

a=partial(modelo_xgb, pred.var = ”fumigacion”, train = data.matrix(test[,-9])) autoplot(a,contour = TRUE, colour = red”)

b=partial(modelo_xgb, pred.var = ”fertilizacion”, train = data.matrix(test[,-9])) autoplot(b,contour = TRUE, colour = red”)

c=partial(modelo_xgb, pred.var = ”deschive_f2”, train = data.matrix(test[,-9])) autoplot(c,contour = TRUE, colour = red”)

c=partial(modelo_xgb, pred.var = ."ad", train = data.matrix(test[,-9])) autoplot(c,contour = TRUE, colour = "red")

Apéndice E: Preguntas

Entrevista al Gerente de Producción

1. ¿El peso de racimos es un indicador de alto impacto en el sector bananero?

   El peso del racimo es un parámetro de producción bananero de gran aporte a la Hacienda Bananera pues sin ningún recurso para su mejora continua la Hacienda Bananera no tendría una buena productividad.

2. ¿Según su experiencia en este campo cuales son los factores que inciden en el peso del racimo?

   Los factores que inciden en un buen peso del racimo son la fertilización, cambios bruscos de temperaturas, control de plagas o fitosanitario, labores a la planta.

3. ¿Con su experiencia demostrada ¿Qué variables inciden en el peso del racimo?

   Son muchas las variables que inciden en el peso de los racimos pueden ser variables de labores manuales de proceso agrícola o indicadores atmosféricos.

4. ¿En qué forma ayudaría la predicción de pesos de racimos en la Hacienda Bananera?

   Nos ayudaría a identificar donde estaría el problema de baja producción y por ende donde mejorar, también tendríamos buenas estimaciones de producción a largo plazo una mejor visión a futuro.

5. ¿El sistema de BI implementado en la Hacienda durante este tiempo han sido datos confiables?

   NO hemos tenido ningún problema, pues no ha servido mucho para la toma de decisiones para el mejoramiento en diferentes áreas específicas de producción.

Entrevista al Analista de Producción

1. ¿Cómo se obtiene los Datos de Parámetros de Producción?

   Los datos de Parámetros de Producción se Obtienen mediante el sistema implementados de BI en la cual recoleta información de datos en los días de Producción.

2. ¿Desde qué tiempo tiene implementado la Hacienda Bananera el Sistema de BI de recoleta de Datos?

   La Hacienda Bananera tiene implementado el Sistema desde Abril/2021 hasta la fecha

3. ¿Desde la implementación del BI de producción ha surgido algún problema de gran envergadura?

   El sistema funciona correctamente y no hemos tenido ningún problema técnico ni logístico en el sistema

4. ¿Cuántos registros almacena el Sistema de BI por día de proceso o producción?

   El Sistema almacena aproximadamente de 2000 a 3000 registros diarios dependiendo la cantidad de cajas a procesar