DOI: https://doi.org/10.56712/latam.v5i6.3061

Modelo predictivo basado en algoritmos de *machine learning* para la estimación del peso de racimos de banano en una hacienda

Predictive model based on machine learning algorithms for estimating the weight of banana bunches on a farm

Pedro Santiago Muñoz Torres
juniorpsmt@gmail.com
Universidad Estatal de Milagro
Milagro – Ecuador

Artículo recibido: 13 de noviembre de 2024. Aceptado para publicación: 27 de noviembre de 2024.
Conflictos de Interés: Ninguno que declarar.

Resumen

La agricultura en el ámbito Bananero no se ha visto beneficiada fuertemente por los diferentes avances tecnológicos en materia de Ciencia de Datos y Machine Learning, estas técnicas ayudan 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 para predecir pesos de racimos, para posteriormente realizar un análisis que le permita a un Productor Bananero tomar las mejores decisiones para sus plantaciones y obtener una buena productividad. En primera instancia de este trabajo investigativo consistió en recopilar todos los datos necesarios para entrenamiento y validación de los modelos de predicción para la variable peso (peso del racimo). Los datos son preprocesados estadísticamente, se entrenan y se validan los modelos con el programa R Studio, en este caso se utilizó el algoritmo Random Forest, y el Xgboost que resultó ser más eficiente en la predicción de los pesos de racimos, aunque no hubo mucha diferencia entre estos dos algoritmos, se eligió este último como el mejor algoritmo de predicción por el resultados de sus indicadores de métricas de evaluación del error como: Error Absoluto Medio (MAE) para Random Forest fue de 13.2 y en el modelo Xgboost fue de 13.1 por lo consiguiente se tiene una mínima diferencia entre los dos modelos. Posteriormente se espera realizar muchas más combinaciones con diferentes variables en torno a la producción Bananera, para una mejor eficiencia de este modelo de predicción de peso de racimos de banano se necesitaría que las plantaciones se encuentren monitoreadas por sensores IOT y dependiendo su naturaleza obtener diferentes variables y obtener conclusiones relacionadas a una mayor productividad de las Haciendas Bananeras.

Palabras clave: productividad, machine learning, random forest, xgboost, regression, IOT

Abstract

Agriculture in the Banana field has not been strongly benefited by the different technological advances in Data Science and Machine Learning, these techniques help to understand the patterns and dynamics that your data follows, in this way, the main objective of This research is to use Machine Learning techniques to predict bunch weights, to subsequently carry out an analysis that allows a Banana Producer to make the best decisions for their plantations and obtain good productivity. In the first instance of this investigative work, it consisted of collecting all the necessary data for training and validation of the prediction models for the weight variable (bunch weight). The data are statistically

preprocessed, the models are trained and validated with the R Studio program, in this case the Random Forest algorithm was used, and the Xgboost, which turned out to be more efficient in predicting the cluster weights, although there was not much difference between these two algorithms, the latter was chosen as the best prediction algorithm due to the results of its error evaluation metric indicators such as: Mean Absolute Error (MAE) for Random Forest was 13.2 and in the model Xgboost was 13.1, therefore there is a minimal difference between the two models. Subsequently, it is expected to make many more combinations with different variables around Banana production. For better efficiency of this banana bunch weight prediction model, it would be necessary for the plantations to be monitored by IOT sensors and depending on their nature, obtain different variables. and obtain conclusions related to greater productivity of Banana Farms.

Keywords: productivity, machine learning, random forest, xgboost, regression, IOT

Cómo citar: Muñoz Torres, P. S. (2024). Modelo predictivo basado en algoritmos de machine learning para la estimación del peso de racimos de banano en una hacienda. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades 5 (6), 986 – 1015. https://doi.org/10.56712/latam.v5i6.3061*

I*NTRODUCCIÓN*

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 para 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 máquina 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.

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.

Pregunta general 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?

Preguntas específicas 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?

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.

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.

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.

DESARROLLO

Se ha realizado una búsqueda de distintos antecedentes nacionales e internacionales relacionados al tema de investigación presentado. Así como, la definición de los 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.

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 (Rezk et 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 múltiples desafíos, 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 de- nominando “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 nueva herramienta, 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.

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.

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 del análisis predictivo se pueden usar más dentro del análisis prescriptivo para impulsar acciones basadas en información predictiva (Cedric et al., 2022).

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).

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 de predicción 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.

METODOLOGÍA

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.

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.

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.

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

Técnicas de observación e instrumentos de recolección 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.

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.

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).

Operacionalización de las Variables

Tabla 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

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

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

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 se debe realizar al inicio y luego aplicarse al conjunto de entrenamiento y al de test o 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 les 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.

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.

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)

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).

Error cuadrático medio (MSE). Es un indicador de medición más simple para la evaluación del modelo de regresión 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.

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

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

Coeficiente de determinación R2. 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 entre 0 y 1. Cuanto más cerca de 1 su valor, mayor será el ajuste del modelo a la variable que estamos intentando predecir.

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.

RESULTADOS Y DISCUSIÓN

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 contiene información 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.

Tabla 3

Variables

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 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.

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

Tabla 4

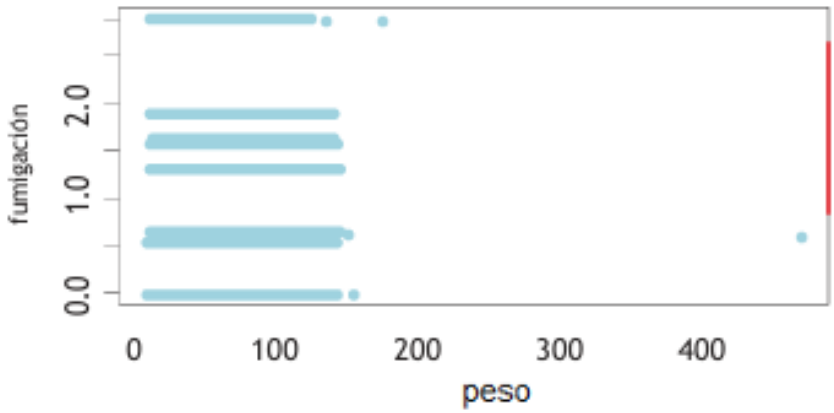
Variables

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.

Gráfico 1

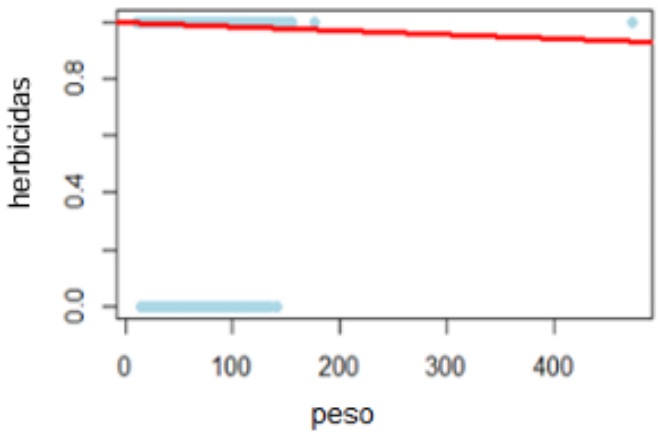
Correlación peso – fumigación



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.

Gráfico 2

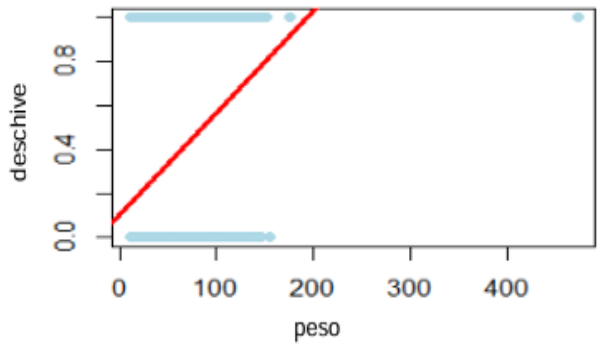
Correlación peso – Herbicidas



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.

Gráfico 3

Correlación peso – deschive_f2



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

Gráfico 4

Correlación peso – calibración superior

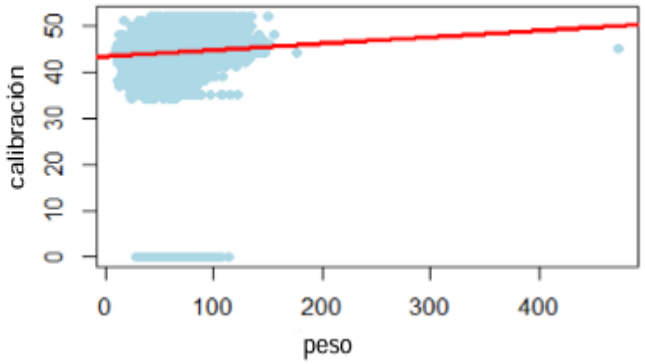
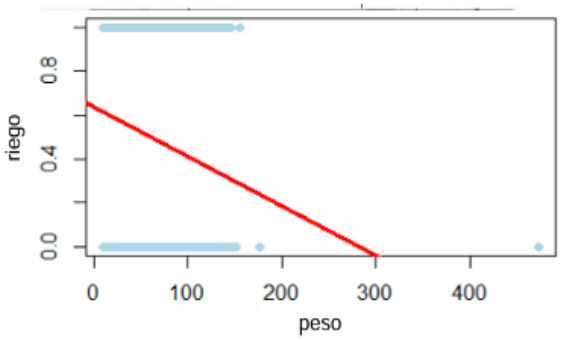


Gráfico 5

Correlación peso – riego

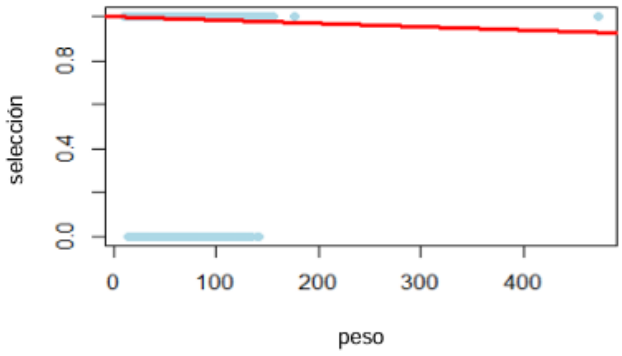


Se observa una correlación negativa a medida que el peso del racimo aumenta el riego disminuye o viceversa.

Gráfico 6

Correlación peso – selección

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



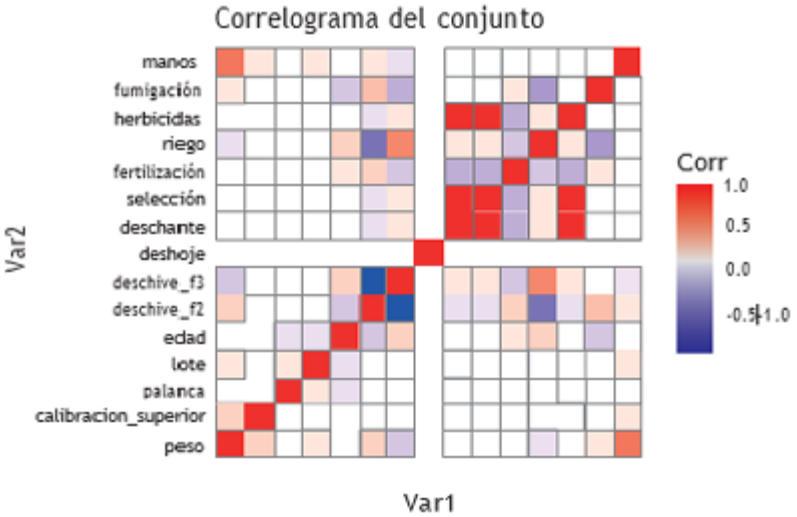
o viceversa.

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 el gráfico.

Gráfico 7

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 lineal entre 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

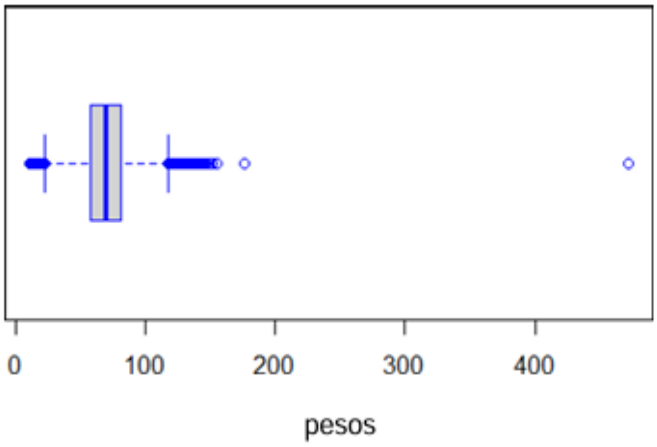
grado de relación lineal entre cada par de elementos o variables. Esta prueba permite cuantificar la magnitud de la correlación entre dos variables y ayuda a predecir valores. Si estas variables tuvieran una correlación perfecta se podría inferir el valor de la variable y conociendo el valor de x. Debido 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)

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:

Gráfico 8

Diagramas 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.

Gráfico 9

Histograma (variable peso)

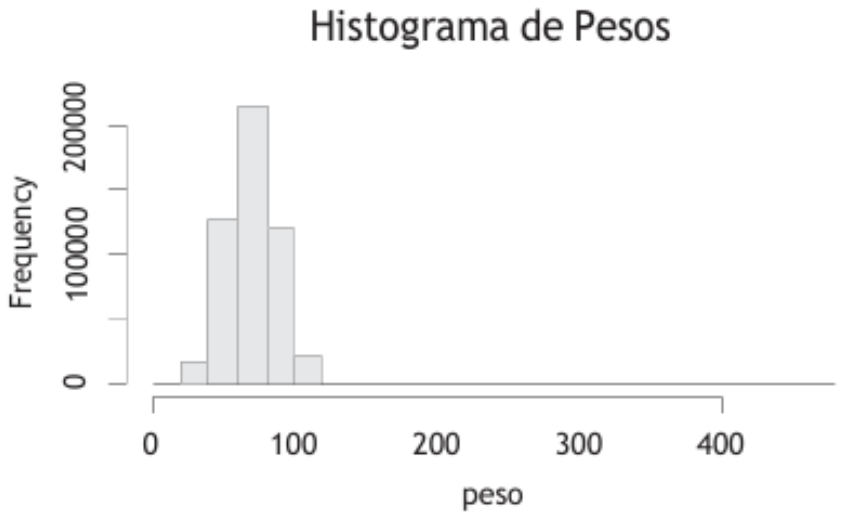


Tabla 5

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 el gráfico 9 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:

Gráfico 10

Diagrama de Caja (variable peso)

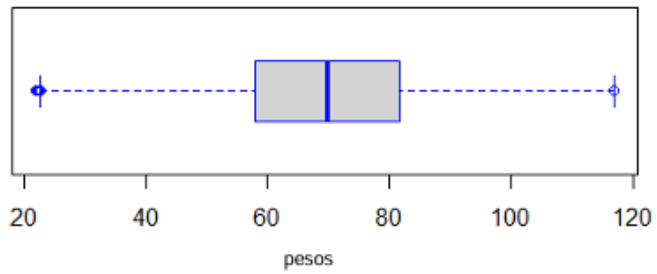


Gráfico 11

Histograma (variable peso)

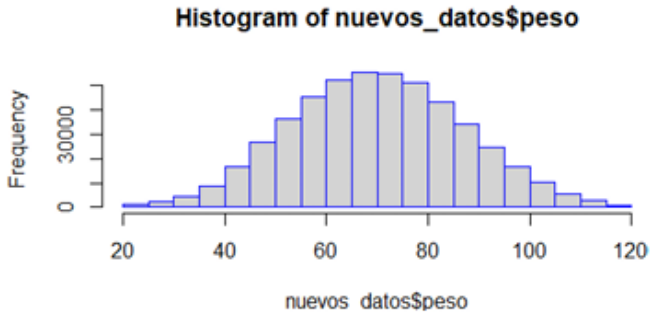


Tabla 6

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 tabla 6 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).

Modelamiento y Métricas

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 7.

Tabla 7

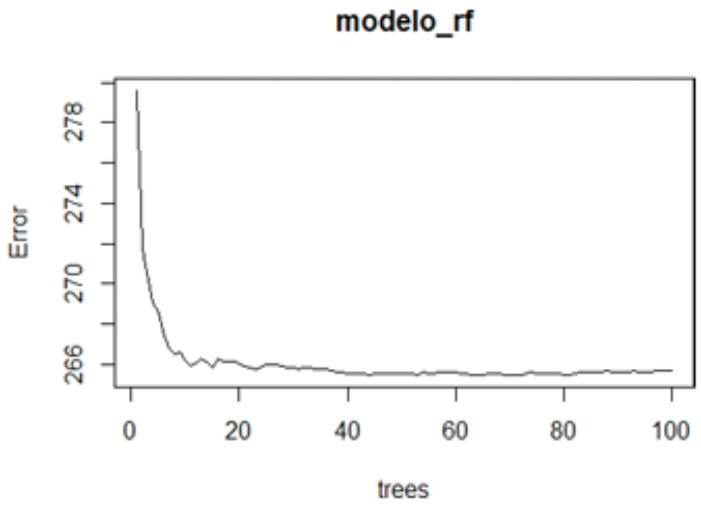
Variables escogidas

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 en el gráfico 12.

Gráfico 12

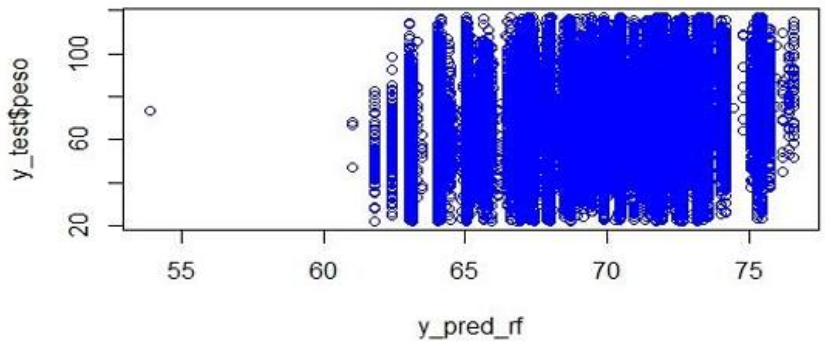
Modelo random forest/ntree=100



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 gráfico en la cual los puntos se superponen y esto quiere decir que la predicción van a la par con los datos reales.

Gráfico 13

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 8

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

Tabla 9

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:

Gráfico 14

Modelo random forest/ntree=200

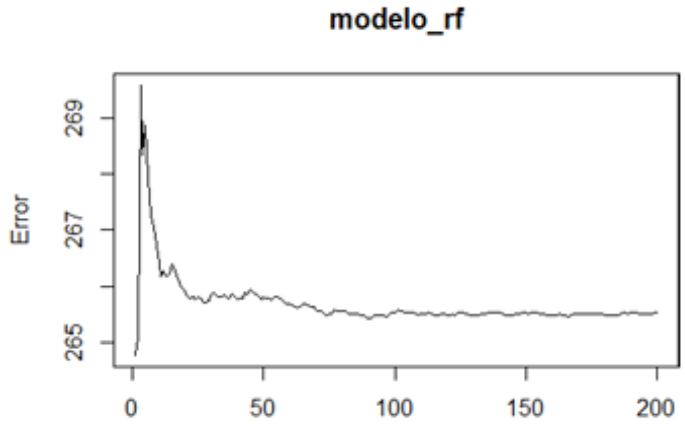


Tabla 10

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

Xgboost

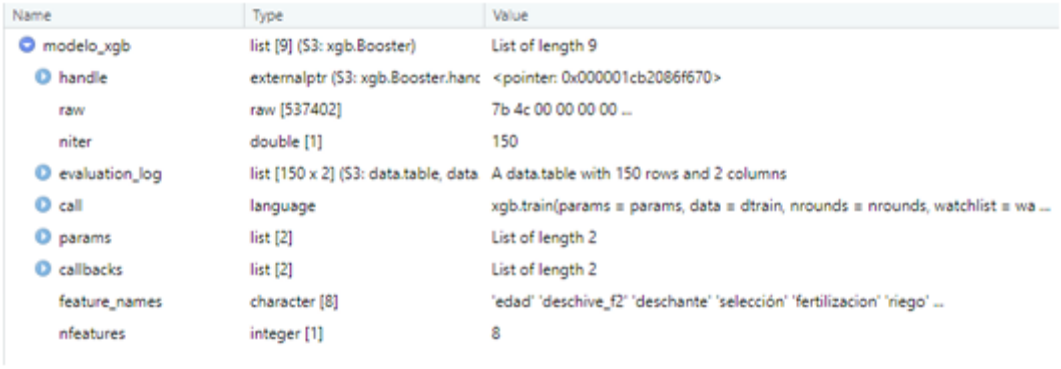
Para realizar este modelo predicción se instala los paquetes y librerías library(xgboost) seguido a esto el dataFrame de entrenamiento se la convierte en una matriz con la función data.matrix como se tiene en el Anexo de la página 81, con la función de xgboost y en sus parámetros nround como parte inicial 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.

Figura 1

Descripción del modelo



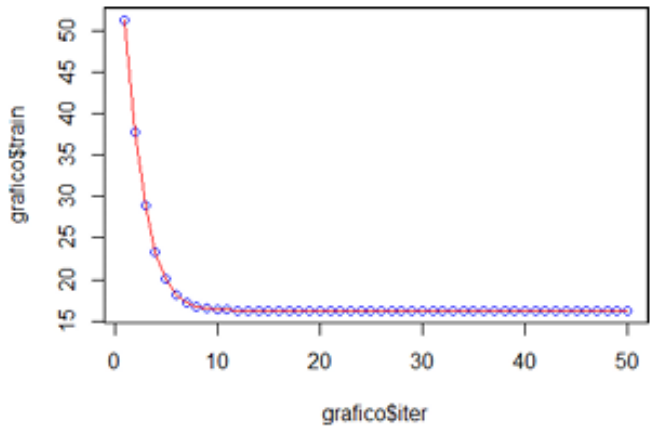
Name	Type	Value
modelo_xgb	list [9] (S3: xgb.Booster)	List of length 9
handle	externalptr (S3: xgb.Booster.hanc	<pointer: 0x000001cb2086f670>
raw	raw [537402]	7b 4c 00 00 00 00 ...
niter	double [1]	150
evaluation_log	list [150 x 2] (S3: data.table, data	A data.table with 150 rows and 2 columns
call	language	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 el siguiente gráfico es fundamental hacer notar como desciende el error a medida que se desarrollan las 50 iteraciones

Gráfico 15

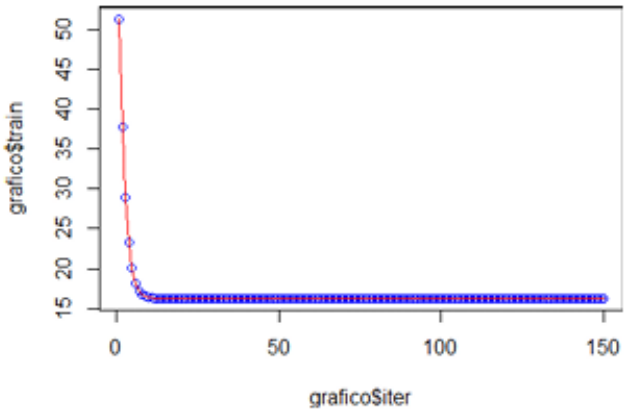
Curva de error modelo XGBoost 50 nround



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

Gráfico 16

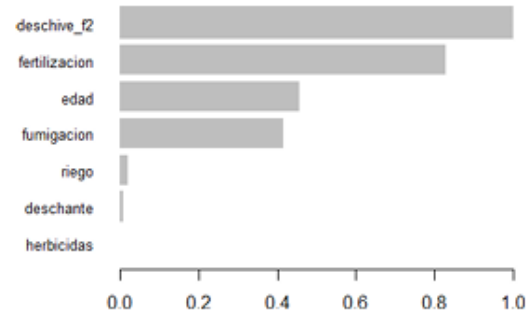
Curva de error modelo XGBoost 150 nround



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 el gráfico y las variables importantes fueron deschive_f2, fertilización, edad, fumigación

Gráfico 17

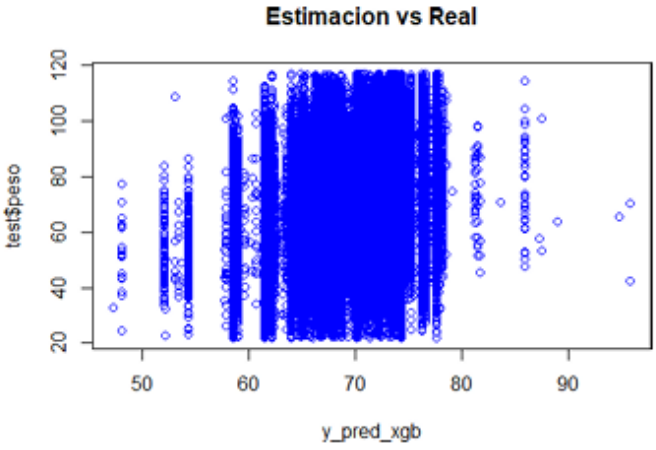
Importancia de las variables



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 del gráfico. Se observan datos superpuestos y a su vez datos un poco más abiertos a los datos reales.

Gráfico 18

Estimación vs Real-XXGBoost



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 11 y 12.

Tabla 11

Métricas para xgpoost-nrounds=50

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

Tabla 12

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 13 y 14.

Tabla 13

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 14

Métricas para el modelo XGBoost

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

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.

CONCLUSIÓN

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 predecir con 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.

REFERENCIAS

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., Asen- cio 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/mo- dules/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 Pi- neda, 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 learning to 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.