UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL

INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

Técnicas de Machine Learning para determinar la producción de cultivos y personal requerido en las campañas de cosecha de la empresa Fundos Rejas SAC

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para:

Obtener el título profesional de Ingeniero Industrial y Comercial,

Obtener el título profesional de Ingeniero de Tecnologías de Información y Sistemas

AUTORES

Rafael Isaac Briceño Rodríguez
Marco Antonio Celedonio Rojas
Walter Javier Crisóstomo Fernández
Jose Luis Medrano Pelaez
Patricia Elizabeth Salas Castillo

ASESOR

Junior John Fabian Arteaga

ORCID N° 0000-0001-9804-7795

Octubre, 2022

RESUMEN

Diferentes empresas están utilizando técnicas de Machine Learning para analizar sus conjuntos de datos con la finalidad de encontrar comportamientos y patrones que les permitan crear modelos matemáticos predictivos, que a su vez pueden predecir diferentes variables de salida para determinar la producción y la cantidad de personal requerido para los cultivos de palta, arándano y mandarina. En el presente estudio, se utilizó una base de datos que comprende los años de campañas de cosecha (2019 a 2022). Para ello, la metodología CRISP-DM para obtener un mejor alineamiento en la etapa de desarrollo. Se utilizaron técnicas de aprendizaje supervisado entre ellas Regresión lineal Múltiple, Árbol de Regresión y Vectores de Soporte de Regresión, para medir el modelo que tiene mejor desempeño se utilizaron las métricas como el R2 y RMSE. Dentro de los resultados obtenidos, se obtuvo que, para determinar la producción del cultivo de palta, la mejor técnica fue la de Regresión Lineal Múltiple y para los cultivos de arándano y mandarina fue el Árbol de Regresión, por otro lado, para determinar la cantidad de trabajadores para el cultivo de palta el mejor modelo fue Árbol de Regresión y para los cultivos de mandarina y arándano fue el SVR.

Palabras Claves: Machine Learning, Regresión Lineal Múltiple, Árbol de regresión, Vectores de Soporte de Regresión, CRISP-DM.

ABSTRACT

Different companies are using Machine Learning techniques to analyze their datasets, to find behaviors and patterns that allows them to create predictive mathematical models, which can predict different output variables to determine the production & number of personnel required for avocado, blueberry & mandarin crops. In the present study, a database comprising the crop years (2019 to 2022) was used. For this, the CRISP-DM methodology was used to obtain a better alignment in the development stage. Supervised Learning Techniques, including Multiple Vector Regression, Regression Tree & Support Vector Regression, also, metrics such as R2 y RMSE were used to measure the best-performing model. Among the results obtained, it was accepted that, to determine the production of the avocado crop, the best technique was the Multiple Linear Regression and for the blueberry and mandarin crops it was the Regression Tree, on the other hand, to determine the number of workers for the avocado crop, the best model was Regression Tree and for the mandarin & blueberry crops were the SVR.

Key words: Machine Learning, Multiple Linear Regression, Regression Tree, Support Vector Regression, CRISP-DM.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

Según la FAO (2018) en los próximos años habrá una alta demanda de alimentos debido al incremento significativo de la población en la última década. Esto significa que el sector agrícola mundial tendrá un gran desafío para poder proveer de alimentos e insumos de calidad, con eficiencia, agilidad y en cantidades importantes. Las empresas agrícolas peruanas no serán ajenas a estos retos, las que además de satisfacer la demanda local y de los países a los que exportan, deben incluir tecnologías modernas en sus procesos para alcanzar competitividad. En este sentido y en la presente era de la digitalización, tienen la necesidad de realizar pronósticos de diversa índole y con un alto nivel de acierto, que ayuden a la correcta toma de decisiones en su campo de acción. Para este fin es necesario procesar gran volumen de datos relevantes para el sector, utilizando herramientas modernas y disruptivas.

En esta investigación planteamos el uso de tres algoritmos de machine learning en la rama de aprendizaje supervisado, como son regresión lineal múltiple, árboles de regresión y vectores de soporte de regresión, los cuales nos permitirán pronosticar la demanda de producción por tipo de cultivo (palta, mandarina y arándano), y el personal requerido para la labor de cosecha por tipo de cultivo de la empresa Fundos Rejas SAC.

Nuestra investigación se divide en seis capítulos, los cuales serán enfocados con criterios específicos.

En el capítulo I, presentamos el planteamiento del problema en el cual se detalla la realidad problemática actual desde una perspectiva genérica hacia una específica. Por otro lado, se realiza la descripción de la justificación teórica, práctica, metodológica, y la delimitación de la investigación.

En el capítulo II, se detalla el marco teórico con los antecedentes de la investigación, y las bases teóricas que nos permiten tener mayor conocimiento sobre las herramientas a utilizar.

En el capítulo III, se determina el entorno empresarial actual en el que se hace la descripción de la empresa, detallando el FODA cuantitativo, la cadena de suministros actual, el modelo de negocio actual (CANVAS) y el mapa de procesos actual.

En el capítulo IV, se detalla la metodología de la investigación y se hace uso de la metodología CRISP-DM, la cual bajo esta metodología permite seguir de manera adecuada proyectos de ciencia de datos, así como determina los lineamientos para la etapa siguiente, asimismo, se detalla el enfoque y el alcance de la investigación.

En el capítulo V, en el desarrollo de la solución, se presentan las actividades realizadas por cada algoritmo y se hace el análisis correspondiente con los indicadores determinados en

las bases teóricas, asimismo se realiza la comparativa para determinar el mejor algoritmo por cada modelo y cultivo.

Por último, en el capítulo VI, se detalla las conclusiones obtenidas en base a los resultados de nuestra investigación, asimismo detallamos las recomendaciones necesarias que podrían tomarse en cuenta para siguientes investigaciones o cuando el proyecto se encuentre en producción.

CAPITULO I: Planteamiento del problema

1.1. Descripción de la Realidad Problemática

Según la FAO (2018) la población en el mundo ha aumentado en la última década, lo que ha significado una alta demanda sustancial en la alimentación. Las Naciones Unidas proyecta que la población será de 9.7 billones para el 2050. Ello se refiere a la utilización directa de productos agrícolas para producir alimentos, y a la utilización de cultivos y otros materiales vegetales y animales para producir piensos para animales, que a su vez se utilizará para la producción de alimentos.

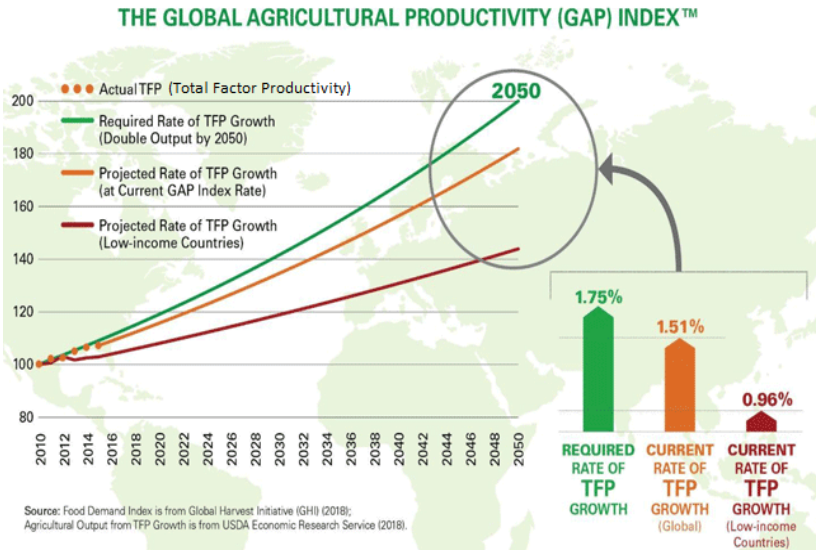


Figura N°1: Brecha de Productividad Agrícola Global

Fuente: *DTN Progressive Farmer (2018)*

A continuación, se detallan los principales destinos de las exportaciones peruanas:

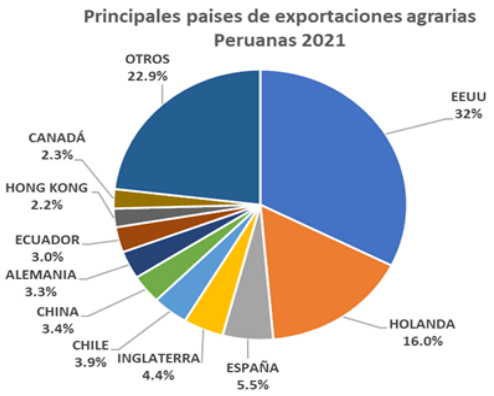


Figura N°2: Principales Destinos Agroexportaciones peruanas 2021

Fuente: *Sunat & Midagri (2022)*

El MIDAGRI (2022) en un comunicado de prensa señaló que, las agroexportaciones en el 2021 tuvieron un incremento del 18% en comparación del año anterior, a pesar de la pandemia por el COVID. La cantidad de exportaciones fue de US $9172 millones, dentro de estas los arándanos representaron US $1,203 millones, uva fresca US $1,200 millones, palta fresca US $1,083 millones y los demás cítricos US $164 millones, entre otros.

Figura N°3: Agroexportaciones peruanas de los últimos 10 años

Elaboración propia

Fuente: *MIDAGRI (2022)*

Uno de los principales retos en el sector agroindustrial es poder acertar con las estimaciones de cosecha. Conocer los volúmenes anticipadamente permite planear correctamente las operaciones agrícolas y comerciales. Dentro de las operaciones agrícolas nos permitirá planificar los presupuestos con un mayo ajuste, conocer la cantidad de personal a contratar para cada campaña. Por otro lado, en el ámbito comercial nos ayuda a conocer la cantidad a ofertar a nuestros clientes y evitar el incumplimiento de los pedidos

En el Fundo Rejas se realiza la contratación del personal en función a la proyección de la cosecha, por juicios empíricos y por datos históricos de años anteriores.

Actualmente no se puede determinar con exactitud la cantidad de personal requerido para las campañas de cosecha, debido a que no se puede asegurar la cantidad de fruta a cosechar por distintos factores climatológicos que afectan de manera aleatoria a los cultivos de palta, mandarina y arándano e imposibilitan su cosecha. Asimismo, la producción de frutas de cada cultivo presenta alta variabilidad lo que lleva a tener que asignar otras actividades al personal

que en un principio tenía destinada la labor de cosecha, lo que a su vez ocasiona que se deba balancear las horas pendientes de cosecha en fechas y horas posteriores.

Según Morales en el 2017, nos indica que el exceso de radiación puede influir en el acortamiento del periodo de maduración del arándano, lo que provocaría concentración de la cosecha y reduciendo la calidad, Por otro lado, a mayor nubosidad se incrementarían las enfermedades fúngicas que impactan la condición y el rendimiento de la fruta. Asimismo, las temperaturas óptimas para el desarrollo varían entre 18 -22 ° para las raíces y de 20-26°C para los brotes, hojas y frutas.

Según Grüter et al., (2022), El cambio climático en el Perú afectará las actuales áreas adecuadas para el cultivo de palta al Perú hacia el 2050, los principales criterios climáticos fueron las altas temperaturas, exceso de precipitaciones y estaciones secas prolongadas. Esto influirá negativamente en la producción de este cultivo.

Por lo descrito en los párrafos anteriores, se demuestra que los volúmenes de producción del arándano, palta y mandarina se ven influenciados directamente por variables climatológicas, entre ellos; radiación, nubosidad, temperaturas, etc. Asimismo, estos factores pueden incrementar las enfermedades y afectaciones por plagas que también influyen finalmente en la producción de las frutas.

1.2. Justificación de la Investigación

1.2.1. Teórica

La presente investigación se realiza con el propósito de aportar al conocimiento existente sobre el empleo del Machine Learning mediante aprendizaje supervisado, como herramienta de proyección acertada en la contratación de mano de obra agrícola, cuyos resultados podrán sistematizarse en una propuesta, para ser incorporado como conocimiento a las ciencias de Machine Learning.

En consecuencia, optimizar la asignación de personal para la labor de cosecha en los cultivos de arándano, palta y mandarina.

1.2.2. Práctica

Esta investigación desarrollará Machine Learning supervisado, a través de 2 modelos de regresión lineal, SVR y Árboles de Regresión, debido que existe la necesidad de la empresa

Fundo Rejas de poder conocer anticipadamente la cantidad de producción que dará su cosecha y la cantidad de trabajadores necesarios a contratar.

Para el primer modelo se considerarán variables climatológicas como entrada y la variable de cantidad de producción como salida.

Por otro lado, para el segundo modelo se considerarán variables climáticas, cantidad de producción como entrada y la variable de cantidad de trabajadores como salida.

1.2.3. Metodológica

Una vez demostremos mediante métodos científicos, la validez y confiabilidad de los modelos para la proyección de la mano de obra agrícola, podrán ser utilizados como modelo en otros trabajos de investigación y/o en otros negocios agrícolas.

1.3. Delimitación de la investigación

1.3.1. Espacial

La investigación se realizará en una empresa agroindustrial ubicada en el Perú. Para el desarrollo se utilizó información meteorológica, de producción y operatividad de fuerza laboral, la cual se origina dentro del contexto de la empresa ubicada en la provincia de Huaura.

1.3.2. Temporal

La investigación utilizará datos recopilados desde 2019 a 2022 de los cultivos de Palta, Mandarina y Arándanos, así como de los factores climáticos determinados por el microclima característico de la zona, que impacta en la producción de las plantas y la necesidad de obtener personal para cubrir esta demanda.

1.3.3. Conceptual

La presente investigación aplicará Machine Learning para desarrollar un modelo predictivo que nos permita identificar la cantidad de personas que se necesitan para los diferentes meses de cosecha que abarca la campaña de los cultivos de palta, mandarina y arándanos. Se utilizará la técnica de aprendizaje supervisado de regresión, en el cual se entrenará al modelo con variables de producción, meteorológicas y de operatividad de la fuerza laboral, con el fin de obtener el modelo predictivo.

CAPÍTULO II: Marco Teórico

2.1. Antecedentes de la investigación

Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia (2020)

La investigación se realiza en el país de Colombia y la principal problemática es incrementar el rendimiento del cultivo de cacao, por lo cual se aplica machine learning con la finalidad de determinar los factores que más influyen en el rendimiento. Las principales variables independientes son precipitaciones, radiación, temperatura, variedad del cultivo, nivel de fertilización, exposición al sol, lluvia y la humedad del suelo, estos datos fueron tomados desde el 2015 al 2017.

Las etapas de recolección de datos se enfocaron en las variables que se muestran en la siguiente tabla:

Variable name	Meaning	Type	Variable name	Meaning	Type
Cocoa_v	Cocoa variety	Cat[a]	P_accu_prev	Accumulated rainfall one month before harvest	Con[b]
Exp	Exposition	Cat[a]	T_avg	Temperature average on harvest month	Con[b]
F_level	Fertilization level	Cat[a]	T_avg _prev	Temperature average one month before harvest	Con[b]
EC_avg	Electrical conductivity on harvest month	Con[b]	Rad_accu	Accumulated photosynthetic active radiation (PAR)	Con[b]
Hum_avg	Soil humidity average on harvest month	Con[b]	Rad_accu_prev1	Accumulated photosynthetic active radiation (PAR) one month before harvest	Con[b]
P_accu	Accumulated rainfall on harvest month	Con[b]	Rad_accu_prev2	Accumulated photosynthetic active radiation (PAR) two months before harvest	Con[b]

a Categorical variable, b Continuous variable

Tabla N°1: Detalle de las variables de estudio

Fuente: *Lamos et al. (2020)*

Es así que estas variables sirvieron como entrada para desarrollar los múltiples algoritmos de machine learning (Regresión lineal Lasso, Support Vector Machines, Boosting y Random forest), luego se evalúa cuál modelo tuvo mejor desempeño a través de métricas como Root Mean Square Error, The Mean Absolute Error y coeficiente de determinación. En la tabla N° 2 se observa que el modelo Gradient boosting tuvo mejor desempeño respecto de los otros modelos, es así que obtuvo un MAE y RMSE más bajo y asimismo obtuvo un R2 Mayor.

Model	**MAE**	**RMSE**	**R^2 (%)**	**RI_{rmse} (%)**
LASSO	20.65	31.73	20.65	20.99
SVM	15.69	27.41	41.17	8.54
Random Forest	14.7	26.65	44.19	5.93
Gradient Boosting	12.94	25.07	49.29	-

Tabla N°2: Mediciones del desempeño de los modelos desarrollados

Fuente: *Lamos et al. (2020)*

Una vez determinado el mejor modelo, este se utiliza para identificar las variables más influyentes. A continuación, se presenta la interacción de las variables independientes más influyentes sobre la variable dependiente:

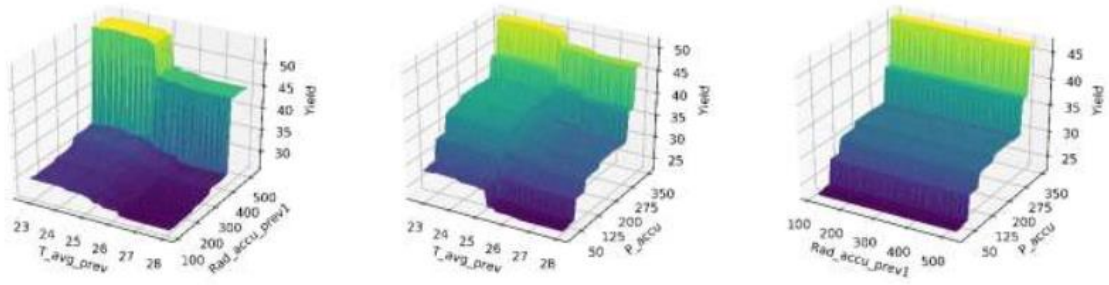


Figura N°4: Interacción de temperatura, radiación y humedad respecto del rendimiento del cultivo de cacao

Fuente: *Lamos et al (2020)*

Los resultados de la aplicación de este modelo de machine learning fue que la temperatura y radiación un mes antes de la cosecha, lluvias en el mes de la cosecha, humedad en el suelo son las principales variables que impactan en el rendimiento del cacao. Este estudio aporta a los agricultores conclusiones que les permiten poder tomar decisiones para el mejor manejo de sus cultivos lo que les generará mayores rentabilidades.

Aplicación del Machine Learning en Agricultura de Precisión (2020)

El artículo muestra herramientas para el procesamiento de información de los cultivos, las cuales permiten identificar patrones en la cosecha, lo que ayudará al agricultor a tomar decisiones basadas en los datos. En esta investigación primero se realiza una búsqueda del uso del término precision agriculture, analizando bases de datos de publicaciones de diferentes autores, periodos, entre otros. En este sentido, uno de los hallazgos es que en los últimos años la información con respecto a agricultura inteligente se incrementó, y también determinaron cuales son las temáticas principales mediante Vos viewer, para la construcción de redes de palabras clave.

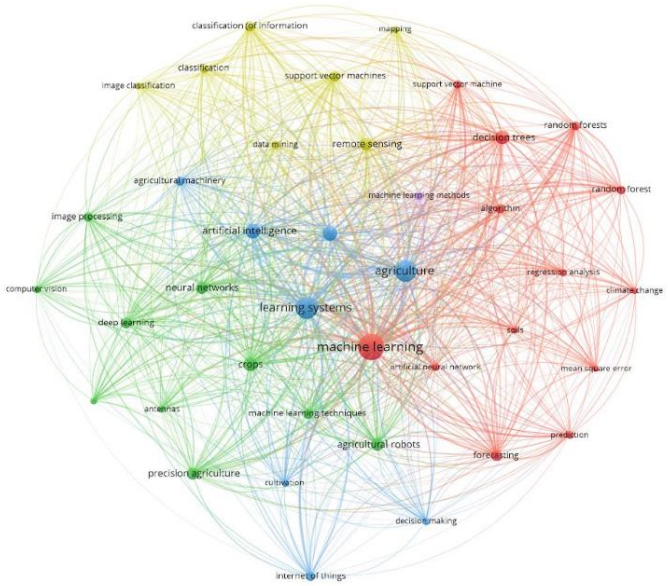


Figura N°5: Palabras clave sobre Smart Agriculture

Fuente: *Ramirez, C. (2020)*

En esta investigación se trata de predecir la variable categórica “daño a los cultivos”, la cual está dividida en cultivo saludable, daño por uso de pesticidas y otros daños. Para esto, primero hacen el tratamiento de la información de sus datos cuantitativos mediante promedios

debido a datos faltantes en su dataset; después realizan el análisis de los datos, la cual obtienen varios análisis, por ejemplo que los cultivos que son sanos disminuyen de manera proporcional si la cantidad de plagas aumenta; asimismo obtienen otro insight referente a que los cultivos se mantienen saludables si la plaga se encuentra dentro de un rango de 300-800 u/m2 sin incrementar el uso de agroquímicos, ya que estos afectan la calidad del cultivo.

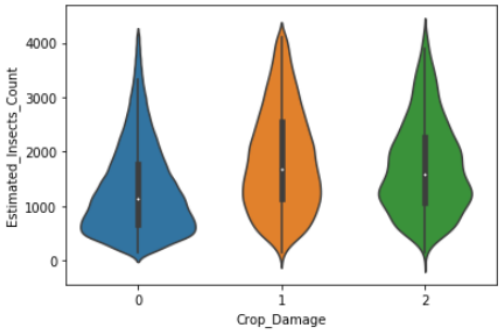


Figura N°6: Daño del cultivo

Fuente: *Ramirez, C. (2020)*

Como última fase en esta investigación, se hace uso de varios modelos como KNN, decision tree, SMV y Naive Bayes, para categorizar de manera correcta y de esta forma predecir el estado en el cual se encuentra el cultivo en base al uso de los pesticidas, asimismo se hace la comparación de cuál es el mejor modelo para sus datos, el cual el modelo de decision tree predice de mejor manera la variable dependiente.

Modelo	Accuracy train	Accuracy test
DecisionTreeClassifier	0.8403239150600447	0.8406372474169085
KNeighborsClassifier	0.8421540852909469	0.8410516982676306
SVM	0.8354601749943591	0.8410516982676306
Naive Bayes	0.8354601749943591	0.8354590096848997

Tabla N°3: Comparativa de modelos

Fuente: *Ramirez, C. (2020)*

En conclusión, esta investigación trata sobre un modelo de aprendizaje para proyectar el estado y las características de la cosecha considerando información sobre el uso de agroquímicos y otras variables a considerar del cultivo. Asimismo, la metodología de machine learning del estudio consta de cuatro pasos: proceso previo y análisis de la información

obtenida; discernimiento de los datos de testeo, training y validación; selección del modelo (clustering); y evaluación de los parámetros del modelo con indicadores y Se proponen 5 algoritmos para el modelo, por último, se realiza la medición de estos considerando el indicador el “accuracy score”.

Aplicación de nuevas tecnologías en el fortalecimiento de la cadena Agroindustrial (2020)

En esta investigación se describe el uso de nuevas tecnologías en distintos ámbitos como inteligencia artificial aplicada en la agricultura. Asimismo, se describen definiciones principales sobre el suelo, su empleo como pilar de los organismos de producción agroindustrial y aplicaciones de tecnologías innovadoras sobre sus propiedades. Por otro lado, se analiza el empleo de nuevas tecnologías en la selección del grano de cacao, la valoración de su calidad y mejora del producto en base a técnicas como el procesamiento de imágenes la cual ponen las imágenes puestas de manera matricial, para la segmentación por regiones binarios de 1 ó 0 de esta forma se realiza una detección de bordes y umbrales de la imagen del grano de cacao, para ello las imagenes deben estar en colores blanco y negro para evitar modificar las propiedades. También se expone el uso de redes neuronales, para generar modelos que sean capaces de determinar estándares de calidad a partir de las características que tiene el grano de cacao, para ello se determinan mediante dos estructuras, la primera es con el tamaño que tienen los granos, perímetro, ancho, RGB, longitud y color. La segunda en la fermentación, la cual contiene las variables de tonalidad, color e imágenes hiperespectrales la cual dieron resultados variados y esto se debió en mayor parte a la heterogeneidad en las formas de los granos.

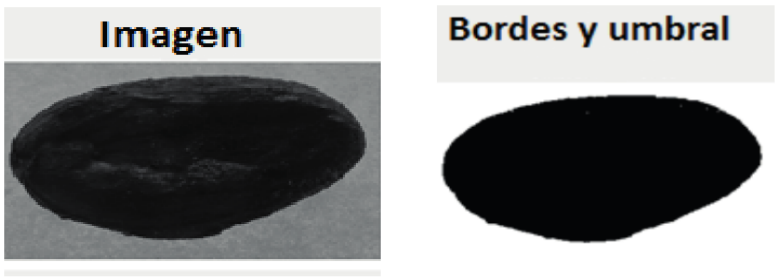


Figura N°7: Análisis de borde y umbral del grano de cacao

Fuente: *Machado, L. (2020)*

Finalmente se expone la relevancia de las redes neuronales en el reconocimiento del manto vegetal y la supervisión agrícola a través de imágenes satelitales, en la cual la herramienta fundamental es la teledetección espacial, que permite obtener imágenes en alta

resolución para obtener datos de las áreas de manera detallada en términos de composición y las características de la superficie, de esta forma se hace uso de las redes neuronales para la clasificación de las imágenes en aplicaciones como georreferenciación de puntos, la clasificación del suelo y la cobertura, otra metodología son las redes neuronales convolucionales, las cuales clasifican las imágenes 2D y 3D por último otra aplicación son las redes neuronales artificiales perceptrón multicapa, la cual las se determina mediante capas de entrada, que contienen imágenes multiespectrales la cual cada pixel es un número de neurona, las capas ocultas representan la clasificación con sesgos de ponderación en la separación de coberturas, todas estas técnicas son usada para hacer un adecuado procesamiento de imágenes, para su posterior análisis y clasificación.

IA en la gestión de personas (2020)

La investigación se centra en la relevancia y utilidad de la inteligencia artificial en los recursos humanos. Se fundamentan teóricamente en definiciones y estudios de diferentes expertos en Recursos Humanos que han aplicado la tecnología en sus respectivos campos de acción. Se logra comprobar que Inteligencia Artificial simplifica y automatiza las tareas rutinarias, analíticas y que constan de data, lo cual permite calcular de manera óptima la fuerza de trabajo real que se necesita en una operación.

Para ello, determinan áreas clave donde la inteligencia artificial genera impacto en la gestión de las personas, entre ellos se encuentra la eliminación de prejuicios en el proceso de reclutamiento, el cual permitiría suprimir el sesgo en la búsqueda de los perfiles y de esta forma aprovechar en mayor medida los talentos. Por otra parte, como apoyo en los sistemas de rendimiento para la evaluar el desempeño, de esta forma se mediría el rendimiento del personal, haciendo que la disertación sea mayormente objetiva, también en la colaboración de planificación de la sucesión analizando perfiles y detectando los candidatos con las coincidencias más similares para la función. Otro de los usos que se realiza es evitando los prejuicios en el aprendizaje, es decir, en el proceso de aprendizaje del talento otro de los usos es para tomar decisiones, la inteligencia artificial permite la elaboración de estrategias y nuevas prácticas para atraer talentos y volviendo más eficientes los procesos. Para ello, mencionan herramientas disponibles de IA como Cognizant, SAS, que permiten el análisis y gestión de los datos de RR.HH, de esta forma obtener la evaluación de talento, obteniendo como resultado los patrones de capacidad de liderazgo, talentos que vayan a salirse de la empresa, también herramientas especializadas para el proceso de selección como Textkernel, Watson Talent

Suite, que como toda herramienta permite trabajar con volúmenes de datos, rastreando información de los talentos en Linkedin y otras redes, permitiendo el filtrado de variedad de curriculums para agilizar el procesamiento de selección, evaluación y por último la contratación, de esta forma se obtendría el talento más apto.

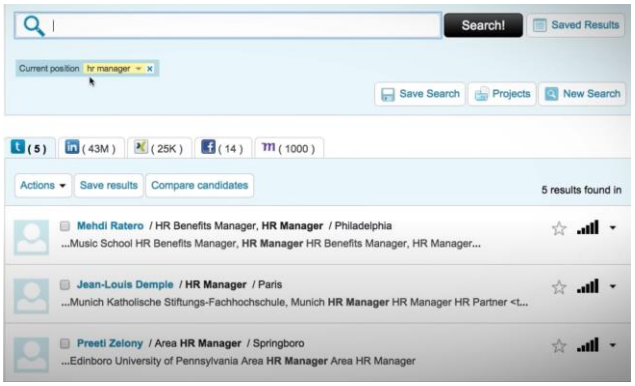


Figura N°8: TextKernel HR Suite

Fuente: *TextKernel(s.f.)*

De esta forma el estudio realiza entrevistas a varias empresas para obtener conclusiones del uso de la IA a varias empresas como Mercadolibre, YPF, etc. en las cuales obtuvieron el valor agregado para las operaciones de la empresa, así mismo obtuvieron como resultado que la incorporación de las tecnologías no reemplazaría al personal, sino que reforzaría o complementaria a los seleccionadores para una mejora en término de tiempo de selección de talentos y automatización.

Machine learning and econometric applications for increasing profitability and efficiency: A case study on sustainable production and trade in agro-based industries (2021)

En el contexto actual, el crecimiento de la población mundial está dando lugar a que se necesite una mayor producción agrícola. Las empresas multinacionales no poseen suficientes tierras aptas para cultivo al igual que los países en vías de desarrollo, por lo que se ha convertido en una necesidad el enfocar las fuentes de inversión en el rubro agrícola para cumplir con la demanda de los bienes y servicios derivados del mismo. Asimismo, los cambios climáticos se han ido acentuando durante estos últimos años debido al efecto invernadero, lo cual según estudios se sugiere que ha disminuido el rendimiento de los cultivos entre un 1% a 2% por cada década durante este último siglo. Para obtener un ritmo creciente en la productividad agrícola

se debe tener en cuenta la disminución de las tierras, costes altos y el ritmo acelerado de aumento de la población mundial afectan este crecimiento, es por ello que se requieren innovaciones tecnológicas.

En la tesis de doctorado se analiza un caso del sector agropecuario, para el cual se diseña un método que combina algoritmos de aprendizaje supervisado y econometría con el fin de optimizar la gestión de recursos.

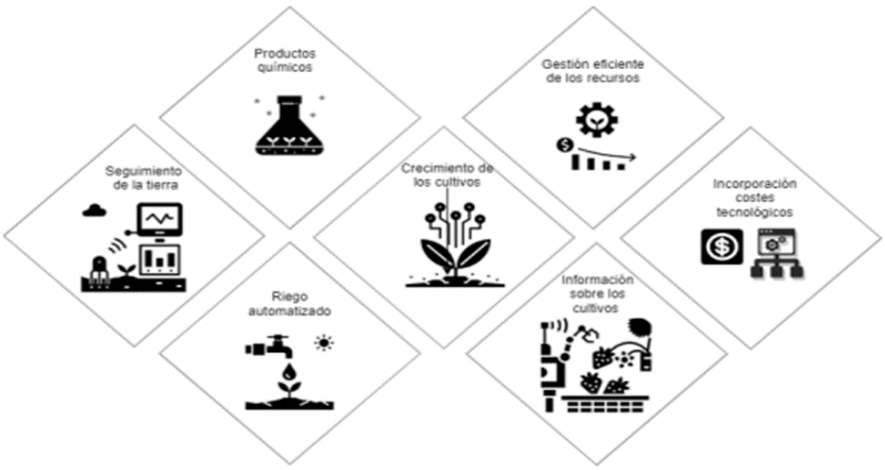


Figura N°9: Diagrama Agricultura de Precisión.

Fuente: *Perez-Pons M. (2021)*

Se realizaron 3 experimentos siendo el primero orientado a la medición de eficiencias de recursos empleando un método de programación lineal no paramétrico (Data envelopment analysis) para calcular fronteras de producción y sus costos; el segundo, en el cual se diseñó un sistema para proyectar oscilaciones de los precios en los mercado en productos agropecuarios y para tomar decisiones con parámetros medioambientales; y el tercero, en el que se diseñó una metodología de razonamiento para la recomendación de inversiones en empresas del sector agropecuario.

Con los resultados se ha demostrado que la utilización de nuevas tecnologías como el Edge Computing permite que las empresas reduzcan sus costes e incrementen su rentabilidad. Para ello se ha realizó un experimento en el que a partir del Data envelopment analysis se ha comparado como afecta la incorporación de los costes de enviar los datos al cloud utilizando Edge Computing.

Además, para mejorar el sistema de recomendación se ha implementado una mejora que permite identificar las métricas de evaluación más adecuadas en los casos de aprendizaje supervisado con clases no balanceadas

2.2. Bases teóricas

2.2.1. Inteligencia artificial

Rouhiainen (2018) define la inteligencia artificial como “la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano” (p. 17).

Benítez et al. (2013) nos define Inteligencia Artificial, una disciplina de la teoría de la informática la cual tiene como finalidad que sistemas artificiales puedan imitar las capacidades intelectuales de los humanos. Está capacidades humanas se hace referencia desde nuestros sentidos como visualizar, oir y van hasta el poder identificar patrones. Por estas razones los usos más comunes de la Inteligencia artificial son el tratamiento de datos y la identificación de sistemas.

Tenemos diferentes sectores que se han beneficiado de la inteligencia artificial, es así como Benítez et al. (2013) nos presenta los principales usos por sector:

Área	Aplicaciones
Medicina	Ayuda al diagnóstico Análisis de imágenes biomédicas Procesado de señales fisiológicas
Ingeniería	Organización de la producción Optimización de procesos Cálculo de estructuras Planificación y logística Diagnóstico de fallos Toma de decisiones
Economía	Análisis financiero y bursátil Análisis de riesgos Estimación de precios en productos derivados Minería de datos Marketing y fidelización de clientes
Biología	Análisis de estructuras biológicas Genética médica y molecular
Informática	Procesado de lenguaje natural Criptografía Teoría de juegos Lingüística computacional
Robótica y automática	Sistemas adaptativos de rehabilitación Interfaces cerebro-computadora Sistemas de visión artificial Sistemas de navegación automática
Física y matemáticas	Demostración automática de teoremas Análisis cualitativo sistemas no-lineales Caracterización de sistemas complejos

Tabla N°4: Principales aplicaciones de la IA

Fuente: *Benítez et al. (2013)*

2.2.2. Machine Learning

Kelleher et al (2015) define machine learning como un método que busca encontrar patrones en un conjunto de datos, con la finalidad de generar modelos predictivos que permitan clasificar o determinar futuros valores de salida para nuevos datos de entrada. Por otro lado, Vanderplas (2017) no dice que machine learning es conocida por ser una rama de la inteligencia artificial; sin embargo, también nos indica que un entorno de análisis de datos viene a ser la construcción de modelos matemáticos que ayuden la compresión de datos. Este aprendizaje inicia cuando se le entrega al modelo ciertos datos que ya han sido observados Una vez el modelo se ajuste a los datos históricos, recién se puede utilizar para predecir datos que recién serán evaluados.

De esta forma, según el libro Python Machine Learning (Raschka Sebastian & Vahid Mirjalili, 2018) clasifica en tres los tipos de aprendizaje automático, aprendizaje supervisado, no supervisado y reforzado.

2.2.2.1 Aprendizaje supervisado

Raschka (2015) nos indica que “el principal objetivo del aprendizaje supervisado es aprender un modelo a partir de datos de entrenamiento etiquetados que nos permite hacer predicciones sobre datos no vistos o futuros” (p. 3).

Asimismo, el Aprendizaje supervisado es definido por Kelleher et al (2015), como la creación de un modelo a partir de datos históricos que sirven para aprender automáticamente de la relación de un conjunto de datos descriptivos y conjunto de datos objetivos. Posteriormente, se realizan las predicciones con el modelo creado. Estos 2 pasos se muestran a continuación:

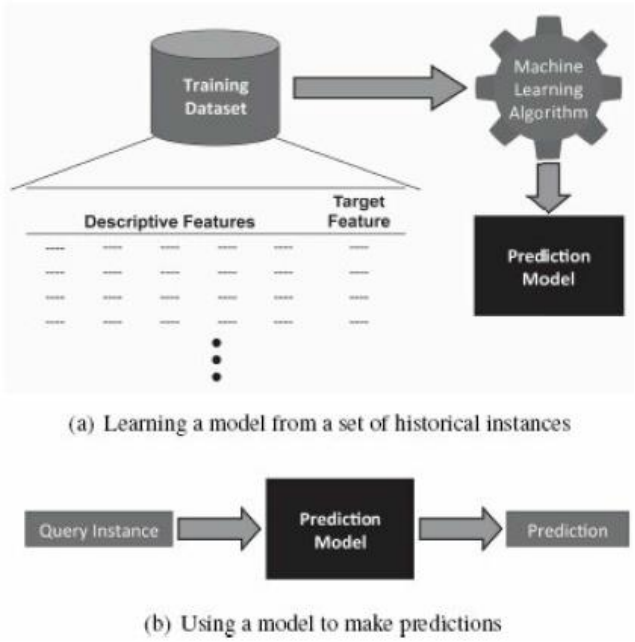


Figura N°10: Aprendizaje supervisado

Fuente: *Kelleher et al (2015)*

2.2.2.1.1. Regresión lineal

Regresión lineal se usa para predecir el valor de una variable con respecto al valor de otra. Es decir, la variable a predecir se denomina dependiente, mientras que las variables que se usan para predecir el valor de otra se denominan independientes. IBM (sf).

a) Regresión lineal simple

Según el libro de Introducción a la econometría (Jeffrey Wooldridge, 2015), a esta ecuación se le llama modelo de dos variable o modelo de regresión bivariada, parte de la premisa de explicar Y (variable dependiente) en función de X(variable independiente), o estudiar la variación de Y cuando varía X, para establecer un modelo que explique las premisas anteriores, se establece una ecuación que relacione Y con X:

$$Y = \beta_0 + \beta_1 X$$

Siendo β_0, es el parámetro de intercepto o constante, así como β_1, es el parámetro de la pendiente de la relación entre X e Y, asimismo según Joaquín A. (2016), indica que las estimaciones de estos valores desconocidos también se le denomina como coeficientes de regresión, debido a que se toman los valores que permitan minimizar la suma de cuadrados residuales.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \underline{x})(y_i - \underline{y})}{\sum_{i=1}^{n}(x_i - \underline{x})^2} = \frac{Sy}{Sx}R$$

$$\hat{\beta}_0 = \underline{y} - \hat{\beta}_1\underline{x}$$

El cual Sx y Sy, son las desviaciones estándar de cada variable, mientras que R es el coeficiente de correlación, por otro lado β_0 es el valor que se espera de Y cuando la variable X es cero

b) Regresión lineal múltiple

Según el libro de Introducción a la econometría (Jeffrey Wooldridge, 2015), permite que muchos factores observados afecten a la variable dependiente, puede expresarse como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

donde, β_0 es el intercepto, β_1 es el parámetro de pendiente asociado a la variable X_1, β_2 es el parámetro de pendiente asociado a la variable X_2, sucesivamente hasta las N variables independientes que tenga la variable a predecir.

2.2.2.1.2. Árboles de decisión

Carlos A. (2021), nos indica que es un modelo no paramétrico de aprendizaje supervisado, el algoritmo que se usan para armar los árboles se nombra como partición binaria recursiva, dado que se realizan sucesivas particiones de un subconjunto de datos. Asi mismo Lozano D. (2015) los árboles de decisión contienen nodos internos, que hacen referencia a los test sobre los valores de una de las propiedades, node de probabilidad, indica que debe ocurrir un evento aleatorio, los nodos hojas, son los valores que devolverá el árbol de decisión y las ramas son los posibles caminos que se tienen de acuerdo a la decisión tomada. Para ello los árboles de decisión tiene una estructura, que según IBM (s.f.), lo define de la siguiente manera:

- Nodo raíz, la cual es el primer nodo donde se realizará la división de la variable que tenga mayor relevancia
- Nodos Intermedios, son llamados también nodos de decisión la cual realizará evaluaciones para dividir el conjunto de datos, para crear subconjuntos
- Nodos hojas, se encuentran en la sección inferior del árbol, su función es representar los resultados posibles dentro del conjunto de datos

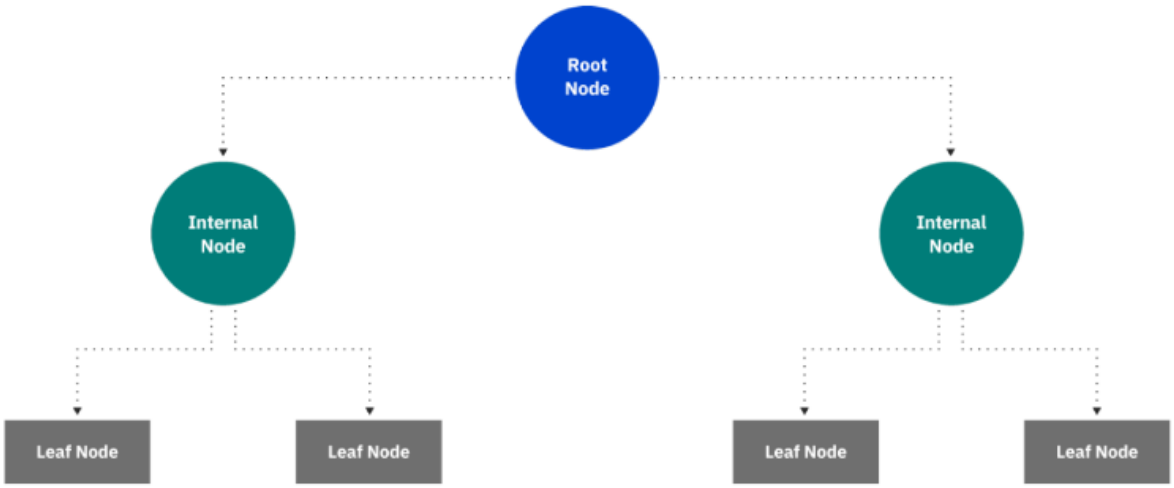


Figura N°11: Árbol de Decisión

Fuente: *IBM (s.f.)*

Algunos parámetros dentro del árbol de decisión, son los siguientes:

- splitter{"best", "random"}, default="best"

  La estrategia utilizada para elegir la división en cada nodo. Las estrategias soportadas son "best" para elegir la mejor división y "random" para elegir la mejor división aleatoria.

  max_depthint, default=None

- La profundidad máxima del árbol. Si es None, los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos muestras que min_samples_split.

2.2.2.1.3. Árbol de regresión

Según Joaquín A. (2020), este subtipo de árbol es usado cuando se quiere predecir variables continuas, el cual las observaciones se distribuyen mediante los nodos, creando la estructura de un árbol hasta alcanzar al nodo hoja.

- Entrenamiento del árbol: Según Joaquín A. (2020), este proceso se divide en dos etapas, la primera es la división sucesiva, la cual genera regiones de nodos terminales, y por último la predicción de las variables de respuesta. Para la construcción de un árbol se emplea el criterio de más frecuencia de esta forma identificar la suma de cuadrados de los residuos o RSS, el cual el principal objetivo es encontrar todas las "j" regiones que permitan minimizar el RSS

$$RSS = \sum_{j=1}^{J} \sum_{i \in Rj} (yi - \hat{y}R_j)^2$$

Donde la $\widehat{yR}_j$, es el promedio de las variables de la región.

- Predicción del árbol, para los árboles de regresión, el valor que se predice es la media de las respuestas que se encuentran en el mismo nodo

2.2.2.1.4. Random Forest

Random Forest es una combinación de árboles predictivos (clasificadores débiles); es decir, una modificación del Bagging, el cual funciona con una colección de árboles incorrelacionados que posteriormente son promediados (Hastie, Friedman y Tibshirani, 2001), los cuales se tienen para cada árbol, que dependen de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque. La generalización de error para los bosques se encuentra en un límite en cuanto el número de árboles en el bosque sea grande. El error de generalización de un bosque de árboles de clasificación depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos. El uso de una selección aleatoria de características para dividir cada nodo produce tasas de error que se comparan favorablemente al algoritmo AdaBoost (Freund y Schapire, 1996), pero son más robustos con respecto al ruido (clasificador fuerte). Estimaciones internas supervisan el error, la fuerza y la correlación. Además, estos se utilizan para mostrar la respuesta al aumento del número de características utilizadas en la división. Los cálculos internos se utilizan, asimismo, para medir la variable de importancia. Las ideas son también aplicables a la regresión.

El elemento común en todos estos procedimientos es que para el k-ésimo árbol se genera un vector aleatorio Θ_k, independiente de los últimos vectores aleatorios Θ_1, …, Θ_{k-1} pero con la misma distribución; y un árbol se desarrolla usando el conjunto de entrenamiento

y de Θ_k, lo que resulta en un clasificador donde h (x, Θ_k) es un vector de entrada. Como se ha señalado líneas arriba, el método Random Forest se basa en un conjunto de árboles de decisión, es decir, una muestra entra al árbol y es sometida a una serie de test binarios en cada nodo, llamados split, hasta llegar a una hoja en la que se encuentra la respuesta. Esta técnica puede ser utilizada para dividir un problema complejo en un conjunto de problemas simples. En la etapa de entrenamiento, el algoritmo intenta optimizar los parámetros de las funciones de split a partir de las muestras de entrenamiento.

$$\theta_k = argmax_{\theta j \in \tau j} I_j$$

Para ello se utiliza la siguiente función de ganancia de información:

$$I_j = H(j) - \sum_{\in I,2}^{i} \frac{|S_j^i|}{|S_j|} H(S_j^i)$$

Donde S representa el conjunto de muestras que hay en el nodo por dividir, y Si son los dos conjuntos que se crean de la escisión. La función mide la entropía del conjunto, y depende del tipo de problema que abordamos (Breiman, 2001)

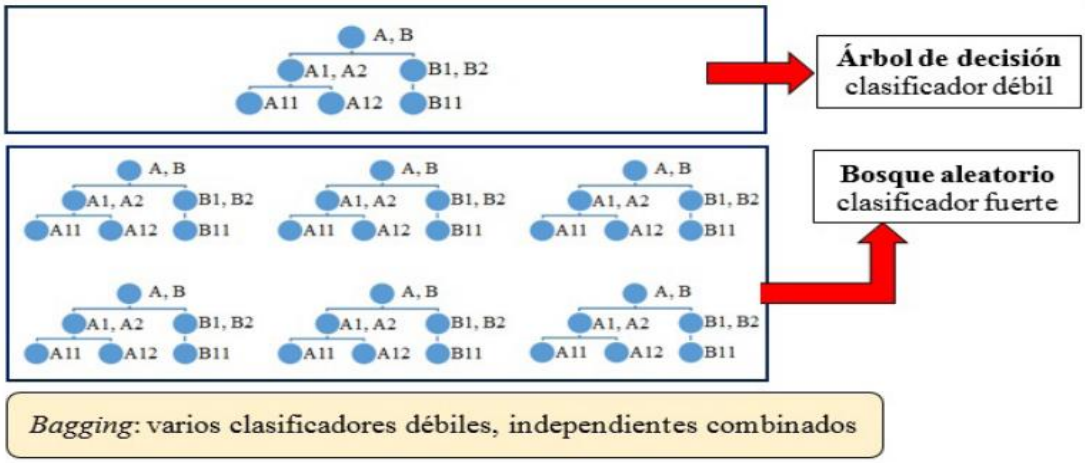


Figura N°12: Árbol de decisión vs Bosque aleatorio

Fuente: *Bosques aleatorios como Extensión de los Árboles de Clasificación con los Programas R y Python*

2.2.2.1.5. Gradient Boosting Machine (GBM)

Chambers y Dinsmore (2015), nos menciona que GBM es un tipo de algoritmo que se pueden derivar varios modelos individuales, como, por ejemplo, los árboles de decisión, de

manera individual. Siendo de modo que cada árbol de decisión genera resultados que se van agregando al resultado final (clasificador de ensamble). Con el fin de poder mejorar los errores y además, contar con una capacidad de predicción superior a los árboles anteriores. Por lo que una de las características de este modelo es que aprende de los errores de múltiples modelos a medida que se vaya generando.

James, Witten, Hastie y Tibshirani (2013), señalan que una de las principales fortalezas de este algoritmo reside en que se construye numerosos árboles de decisión, haciendo frente a las predicciones. Además, con ello se logra mediante la construcción de varios árboles. Sin embargo, señalan, que uno de los problemas es probar los otros métodos escoja un sólo árbol y esto conlleve a que haya poca capacidad predictiva.

Amat (2020) Por lo que se puede rescatar que algunas ventajas y desventajas de este modelo

Ventajas

- No se ven influenciadas por los outliers
- Son modelos que permiten identificar la exploración de datos, lo que hace que la búsqueda se haga de forma rápida.

Desventajas

- Son sensibles con los datos de entrenamiento, por lo que pueden llegar a prevalecer sobre los demás.
- Los valores que se encuentran fuera del rango, no son capaces de extrapolar.

2.2.2.1.6. Support vector machine

Harrington (2012), nos indica que las máquinas de vectores de soporte son un algoritmos de clasificación que basado en un entrenamiento de datos ya recolectados permite poder asignar nuevos datos a un grupo en específico, para esto se llevan a un plano los datos que se tienen y se debe trazar una línea que divide los grupos, a esta línea al no estar en un plano de solo 2 dimensiones normalmente se le asigna el nombre de hiperplano, en la figura 13 se trata de explicar que los conjuntos de datos del cuadrante A puede ser separado por una hiperplano como se muestra en los cuadrantes B, C Y D.

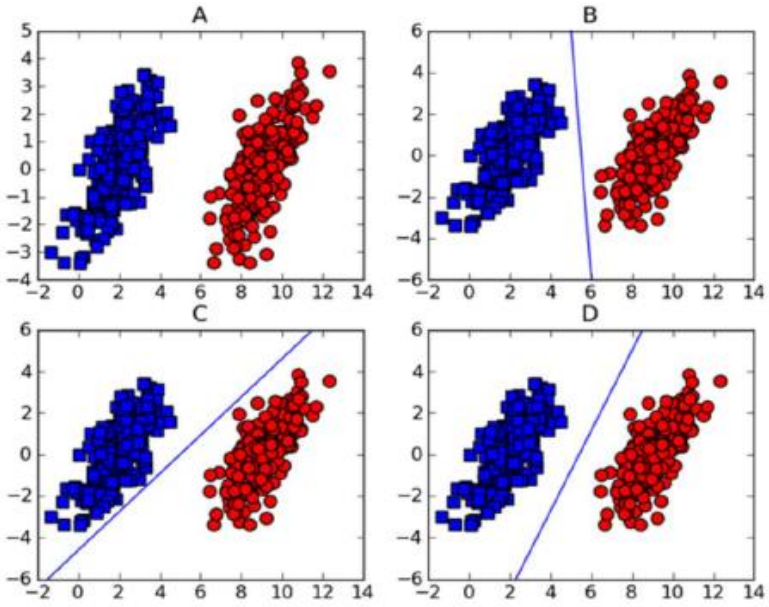


Figura N°13: División de datos por el Hiperplano

Fuente: *Harrington, (2012)*

Como se puede observar hay varias líneas que podrán dividir los grupos de datos, para determinar la mejor línea de separación se debe encontrar los puntos más cercanos al hiperplano y se de asegurar que esté lo más alejado de este, esta distancia se le conoce como margen y a los estos puntos más cercanos como vectores de soporte. Es así que el objetivo es maximizar el distanciamiento de los vectores de soporte con el hiperplano.

Support vector Regression, Según (Daniel R., 2021) el objetivo de esta técnica se basa en la definición de zonas en el hiperplano donde se ignora los errores, de esta forma aproximar el valor dentro de los márgenes, asimismo Bishop (2016) nos presenta la función que permite la predicción de nuevos valores de entrada:

$$y(x) = \sum_{n=l}^{N} (a_n - \hat{a}) k(x, x_n) \; + \; b$$

Para ello el SVR tiene varios conceptos entre los cuales se encuentran:

- Kernel, que permite la asignación de los puntos de conjuntos de datos que tengan menor dimensión a otra que tenga mayor, de esta forma trata de facilitar la averiguación de un hiperplano en un espacio que tenga mayor dimensión. (Daniel R., 2021)
    - Linear: El Kernel lineal, también conocido como margen suave (soft margin), tratará de encontrar un hiperplano que se asemeje a una línea recta, pero puede

tolerar una o más clasificaciones dispersas de datos. Aunque tolera algunos errores, el kernel lineal intenta encontrar una línea que maximice los márgenes y minimice la clasificación errónea. La cantidad de tolerancia de dispersión dada afecta en gran medida la precisión del hiperplano. En Sklearn, la tolerancia se llama C. Cuanto mayor sea el valor de C, menor será la tolerancia de clasificación errónea y más estrecho será el margen. (Chen, L., 2019)

$$K(x, z) = x.z + C$$

- Polynomial: Cuando los datos no se pueden separar con líneas rectas, se necesita un kernel polinomial. El kernel polinomial puede generar un límite de decisión no lineal, con una nueva función aplicando una combinación polinomial de funciones existentes. (Chen, L., 2019)

$$K(x, z) = (\gamma x.z + C)^d, d > 0$$

- Radial Basis Function(RBF): El kernel RBF se utiliza cuando los datos están realmente dispersos. Al entrenar un conjunto de datos con kernel RBF, hay dos parámetros a considerar, a saber, C y gamma. El parámetro C tiene como objetivo indicar cuánto error se debe evitar al clasificar los datos de entrenamiento, cuanto mayor sea el valor de C, menor será la clasificación errónea de los datos de entrenamiento. El parámetro gamma determina hasta qué punto es la influencia de una sola muestra de datos de entrenamiento. Esto significa que cuanto menor sea el valor gamma, mayor será la distancia de los puntos de datos que se van a calcular. (Pedregosa et al, 2011)

$$K(x, z) = exp(-\gamma||x - z||^2), \gamma > 0$$

- Hiperplano: es el subespacio plano y afin de dimensiones p - 1, por ejemplo cuando se tiene solo dos dimensiones, en este caso el hiperplano sería un subespacio de solo una dimensión, la cual vendría a ser una recta, en este caso se definiría bajo la siguiente fórmula, donde los pares de valores de X, si cumplen la igualdad serían puntos del hiperplano (Joaquín A., 2017)

$$0 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Líneas limítrofes, son las líneas que se dibujan en el hiperplano la cual contiene una distancia que se simboliza con ε, de esta forma se introduce la zona alrededor de la función de kernel.

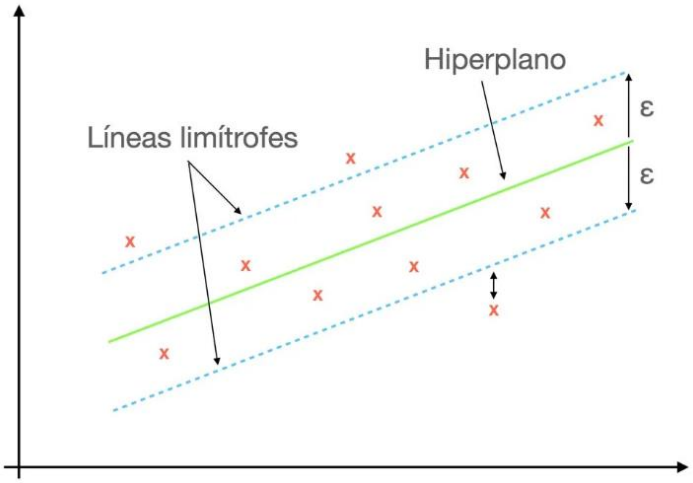


Figura N°14: SVR

Fuente: *Daniel R. (2021)*

- **Parámetro C:** Bishop (2006), para los modelos de SVM el parámetro C es similar a un coeficiente de regularización ya que se encarga de moderar la minimización de los errores que se puedan presentar en el entrenamiento de los datos y controlar la complejidad del modelo. Asimismo, Vanderplas (2017) nos indica que la constante C nos ayuda a controlar el margen, permitiendo que algunos datos se sitúan dentro o fuera de los márgenes para tener un mejor modelo, para un C más grande los márgenes serán más pequeño y para un C más pequeños el margen será más grande.

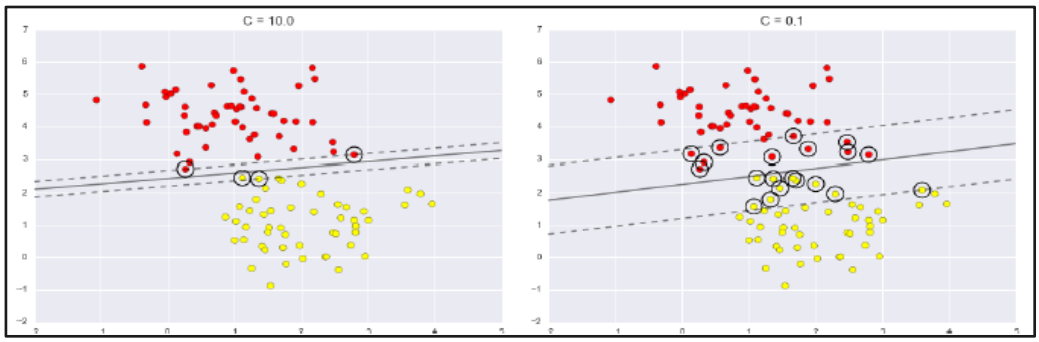


Figura N°15: Variación del margen al cambiar la constance C

Fuente: *Vanderplas (2017)*

- **Parámetro Gamma:** Vanderplas 2017, nos indica que gamma (**γ**) es un parámetro que controla el tamaño del Kernel radial basis function (RBF), Zaki & Wagner Meira (2014) nos indica que RBF viene representada por la siguiente ecuación:

$$K(x, y) = exp(-\frac{|\ x - y\ |^2}{2\sigma^2})$$

Donde:

x: *Son los valores de los datos ubicados en el espacio de un plano.*

y: *Son los valores de los datos ubicados en el espacio de un plano.*

σ: *es el nivel de dispersión similar a la desviación estándar en una función normal.*

Es así que el parámetro gamma es definido por:

$$gamma(\gamma) = \frac{1}{2\sigma^2}$$

2.2.2.1.7. Aprendizaje no supervisado

En el aprendizaje no supervisado no se conoce la variable objetivo como si lo hace el aprendizaje supervisado, aquí se le entrega a la máquina las variables de entrada y se espera que nos diga algo acerca de las variables, o que nos indique cuales son los grupos que puede generar a partir de las variables de entrada (Harrington, 2012).

En el libro Python Machine Learning definen aprendizaje no supervisado como:

El aprendizaje no supervisado nos ayuda a descubrir categorías que a simple vista no podemos diferenciar en los datos. “El objetivo del clustering es encontrar una agrupación natural en los datos de modo que los elementos en el mismo grupo son más similares entre sí que de los diferentes grupos” (Raschka, 2015, p.311).

2.2.3. Métricas de error estadístico

a) ERROR CUADRÁTICO MEDIO (MSE): Kelleher et al (2015) nos indica que, el error cuadrático medio no ayuda a determinar el rendimiento de los modelos de predicción en los cuales los objetivos a predecir son valores numéricos continuos. Asimismo, los valores de que puede tomar MSE van desde el 0 hasta el infinito, mientras más se acerque a 0 el modelo será más efectivo.

Definimos el error cuadrático medio a continuación:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Ytest(i) - Ypredic(i))^2$$

donde:

n= Es la cantidad de muestra

Ytest: son los resultados de los datos esperados

Ypred: son los resultados de la predicción del modelo

b) RAÍZ DEL ERROR CUADRÁTICO MEDIO (RMSE): La raíz del error cuadrático medio en términos sencillos es la raíz de MSE. La diferencia respecto de MSE nos la indica Kelleher et al (2015), los valores obtenidos con RMSE están en las mismas unidades que el variable objetivo y esto nos puede ayudar a comprender mejor el error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Ytest(i) - Ypredic(i))^2}$$

donde:

n= *Es la cantidad de muestra*

Ytest: son los resultados de los datos esperados

Ypredic: son los resultados de la predicción del modelo

c) R^2: El coeficiente R^2 es una métrica que nos ayuda a determinar el rendimiento que un modelo nos ayude a predecir futuros resultado, los resultados del coeficiente pueden ser negativos y van hasta el valor de 1, mientras más se acerque a 1 significa que nuestro modelo es más acertado en la predicción (Kelleher et al, 2015)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Ytest(i) - Ypredic(i))^2}{\sum_{i=1}^{n}(Ytest(i) - Yprom)^2}$$

donde:

n= Es la cantidad de muestra

Ytest: son los resultados de los datos esperados

Ypredic: son los resultados de la predicción del modelo

Yprom: Es el promedio de los resultados esperados

d) Coeficiente de Pearson: Según Morales (2011) el coeficiente de correlación de Pearson expresa la relación entre 2 variables continuas. El valor del coeficiente varía entre 0 y ±1, y en ese sentido, si el valor es igual a 0, significa que no existe relación; si el valor se encuentra entre el 0 y 1, la relación es positiva (a mayor X, mayor Y); y si el valor se encuentra entre el 0 y el -1, la relación es negativa (a menor X, mayor Y).

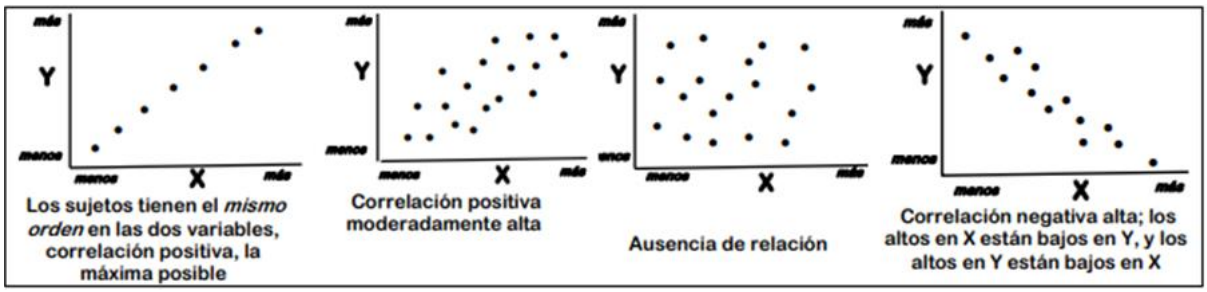


Figura N°16: Errores estadísticos

Fuente: *Morales (2011)*

2.2.4. Metodología CRISP-DM

IBM (2021), cross industry standard process for data mining es una metodología para proyectos de Data Science incluye seis fases que indican las dependencias entre fases, el cual incluye la compresión del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue.

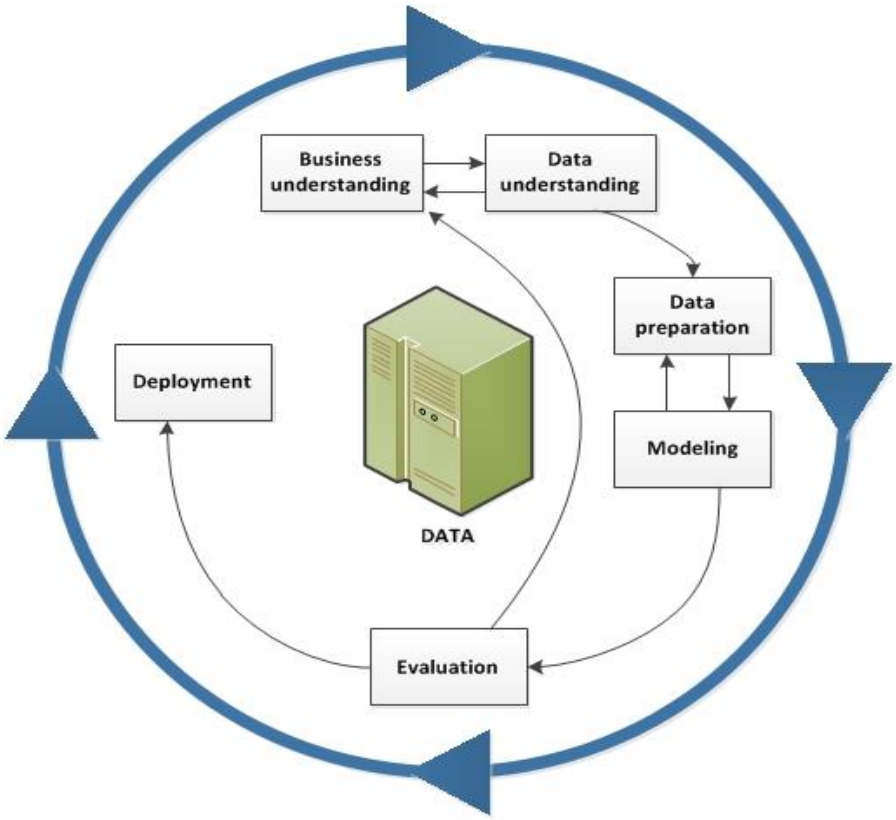


Figura N°17: Diagrama de proceso de CRISP-DM

Fuente: *IBM (2021)*

Compresión del negocio, según IBM (2021), esta primera fase es comprender el negocio, donde se establece los parámetros de los objetivos comerciales, que a su vez el proyecto debe estar alineado a dichos objetivos, así mismo se recopila información acerca de la situación comercial actual.

Entendimiento de los datos, en esta fase se accede a los datos y se exploran con la ayuda de tablas y gráficos que permitan tener un mejor entendimiento de las variables que vayan a usar, por otro lado, este paso es importante para evitar problemas durante la fase de preparación de datos.IBM (2021)

Preparación de los datos, según IBM (2021) es la fase más importante, dado que implica la fusión de conjuntos y registros de datos, se realiza las eliminaciones de valores en blanco o perdidos, o se aplican métodos como media, moda, etc, para el llenado de los valores vacíos, también se realiza la división en conjuntos de datos de prueba y entrenamiento

Modelado, según IBM (2021) en esta fase se escogen los tipos de modelos, para ello se tiene en consideración ciertos aspectos, se deben dividir el dataset en dos conjuntos de

entrenamiento y prueba, también se debe disponer de realizar estimaciones de correlación para hacer la debida elección de variables, de esta se tendrán resultados más fiables.

Evaluación, el proyecto debe cumplir con los criterios de rendimiento comercial, así mismo se determina si los modelos realizan una clasificación o predicción correcta, por otro lado, se evalúa el rendimiento de los mejores modelos. IBM (2021)

Despliegue, en esta última fase según IBM (2021), consiste en utilizar los conocimientos para implementar las mejoras en la organización, la cual incluye actividades como:

- Resumir los modelos para determinar cuáles se pueden integrar a los sistemas de base de datos
- En cada modelo crear una planificación para el despliegue e integración con los sistemas
- Cómo se controlará el despliegue
- Identificación de los problemas y realizar planes de contingencia.

2.2.5. Metodología KDD

KDD o Knowledge discovery in database, según ESAN (2018) esta técnica se basa en analizar patrones que responden a factores importantes, útiles y entendibles. Por otro lado, su finalidad es la interpretación de patrones, modelos y análisis de la información de la organización para tomar mejores decisiones.

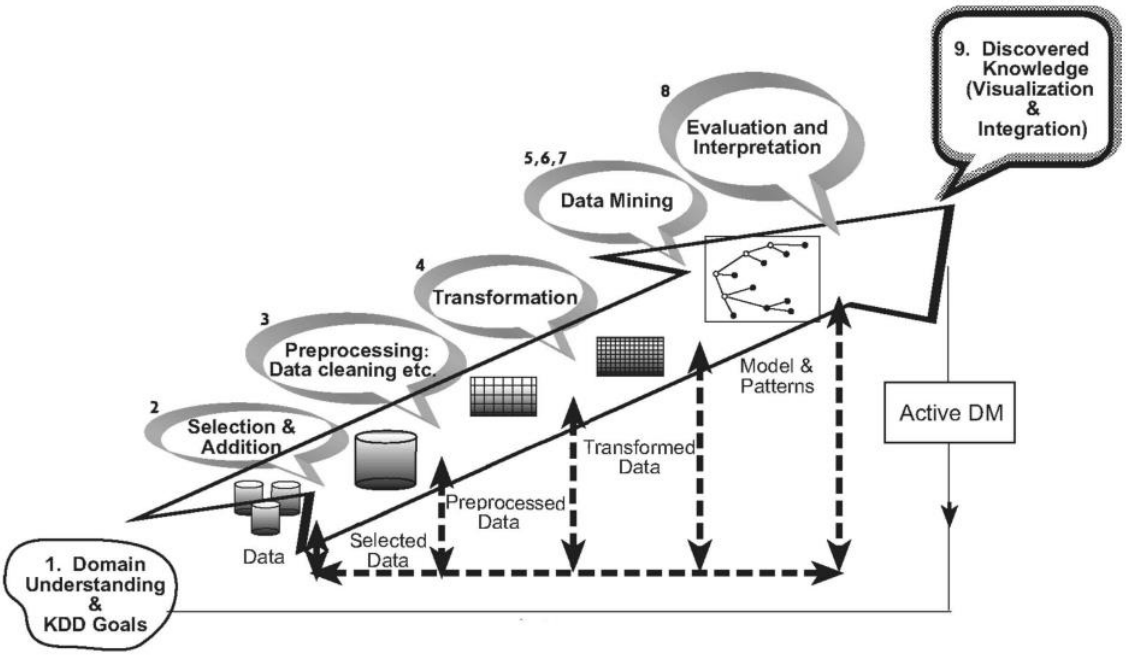


Figura N°18: Diagrama de proceso de KDD

Fuente: *Tel-Aviv University(sf)*

Esta metodología cuenta con las siguientes fases:

- **Comprensión del área de estudio**
- **Implementación de una dataset objetivo,** Según DataScience-PM (2021), sobre una base de datos, se determina los datos objetivos y se determinan las variables que se utilizarán para evaluar el descubrimiento del negocio
- **Limpieza y procesamiento de la información,** Según Tel-Aviv University(sf), en esta etapa se mejora la confiabilidad de los datos, incluyendo cómo la imputación de datos, manejo de valores faltantes y la eliminación de ruido o valores atípicos
- **Minería de datos,** Según DataScience-PM (2021) se centra en filtrar los datos transformados para buscar patrones de interés.
- **Interpretación y análisis de patrones encontrados, en** esta fase se evalúa e interpreta los patrones (fiabilidad), con respecto a los objetivos definidos. Tel-Aviv (sf)
- **Utilización del conocimiento obtenido para tomar decisiones.**

2.2.6. Comparación entre metodologías

- Scrum, Según DataScience-PM (2021) es un marco de referencia mayormente utilizado para proyectos de desarrollo de software uno de los desafíos para usar scrum para data science es el “timebox”, dado que los tiempos necesarios para implementar una solución de ciencia de datos es ambiguo, por otro lado los sprints generados en fechas limites, los equipos de trabajo pueden conducir las pruebas del modelamiento incompletas o realizar análisis apresurados,
- KDD, Según DataScience-PM (2021) es metodología no aborda las realidades de los proyectos de ciencia de datos actuales, como la configuración de la arquitectura de big data, así mismo el proceso iterativo de KDD permite obtener conocimiento y reincorporarse a las fases, mejorándolo de manera eficaz.
- CRISP-DM, esta metodología requiere mucha documentación, en todas las fases, por lo que puede llevar a ralentizar los entregables, así mismo la implementación de CRISP-DM puede adaptarse a los principios y prácticas de metodologías ágiles, dado que los conocimientos aprendidos, sirve como input para otras fases. Esta metodología incorpora una fase fundamental, que es la comprensión del negocio, ya que sirve para alinear el trabajo técnico a las necesidad u objetivos de la empresa. DataScience-PM (2021)

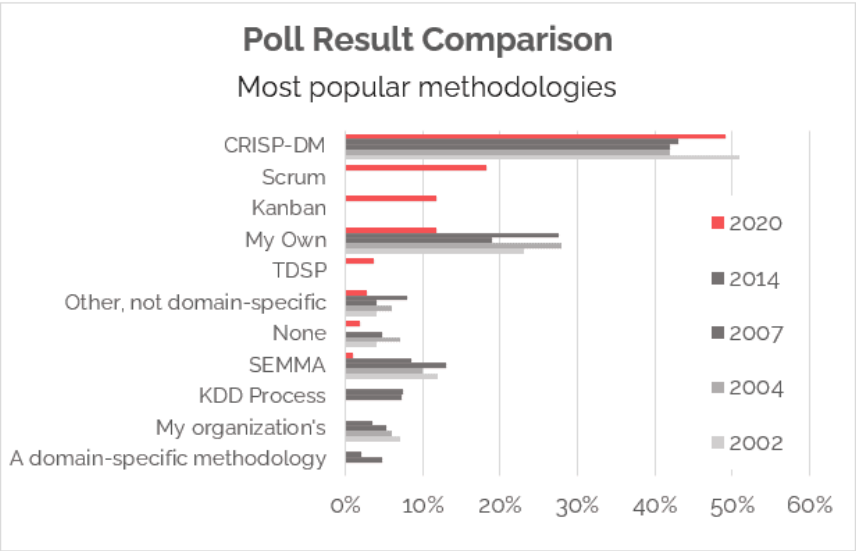


Figura N°19: Metodologías más populares de Data Science

Fuente: *datascience-pm (2022)*

CAPITULO III: Entorno empresarial

3.1 Descripción de la empresa

3.1.1 Reseña histórica y actividad económica

El Fundo Rejas S.A.C, cuenta con 22 años en el rubro agroindustrial, habiendo iniciado sus operaciones manejando hectáreas de paltas hass y una hectárea pequeña de tangelos. En el 2010 se realiza la implementación de 15 ha de mandarinas de variedad Tango y en el 2017 se optó por eliminar un lote de paltos para cambiarlo por la siembra de arándanos, dado el boom que se presentó por la alta demanda de mercado durante este y años posteriores, dándose las primeras cosechas en el 2019. Ubicada en la Irrigación Santa Rosa, Km 29 Carretera Río Seco, Sayán - Huaura, Lima; realiza la actividad de agroexportación contando con aproximadamente 56 hectáreas de cultivo distribuidos de la siguiente forma:

- Palta Hass 22 ha
- Mandarina Tango 22 ha
- Arándanos en Variedades Ventura, Biloxi, Springhigh y Kestrell 12 ha

El tipo de riego empleado es a goteo, proveniente de la irrigación Santa Rosa, Río Huaura, al contar con riego tecnificado las aplicaciones de fertilizantes se realizan por el método de fertirriego.

Figura N°20: Cultivo de Arándano

Fuente: *Elaboración Propia*

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

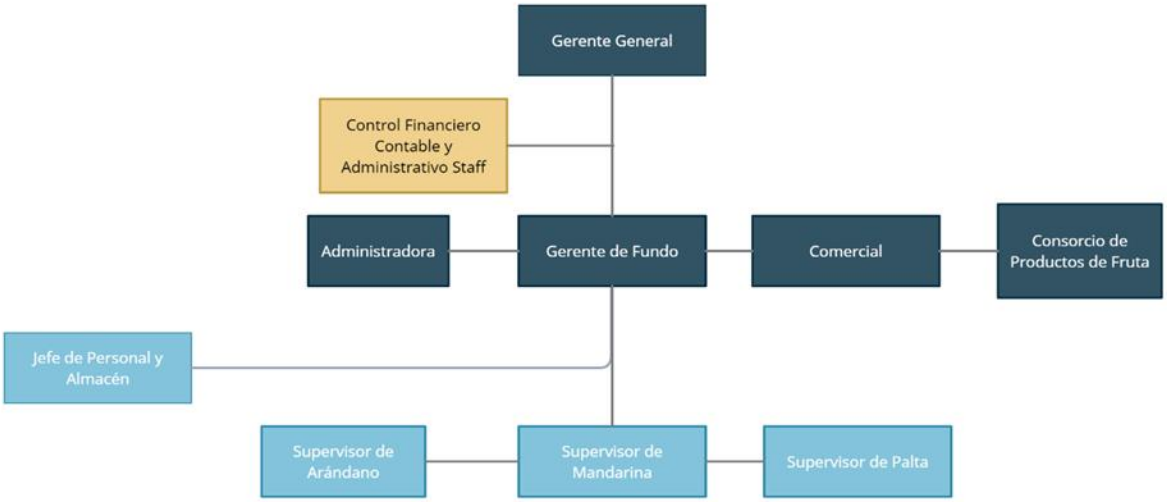


Figura N°21: Organigrama Fundo Rejas SAC

Fuente: *Elaboración Propia*

La figura muestra que la empresa cuenta con una Gerencia General que encabeza a la gerencia en fundo, ubicada en la operación, a diferencia de la parte contable y administrativa que se ubican en la sede corporativa en Lima. Asimismo, el área comercial interactúa con el Consorcio de Productores de Frutas y su personal ubicado en la planta procesadora. Finalmente, en la parte operativa, se tiene un jefe de personal y almacén, encargado de validar la distribución de horas de las actividades del personal de cosecha y operadores de tractores y supervisores de cada cultivo del fundo encargados de hacer cumplir las labores dentro su alcance de manera eficiente.

3.1.2.2 Cadena de suministros

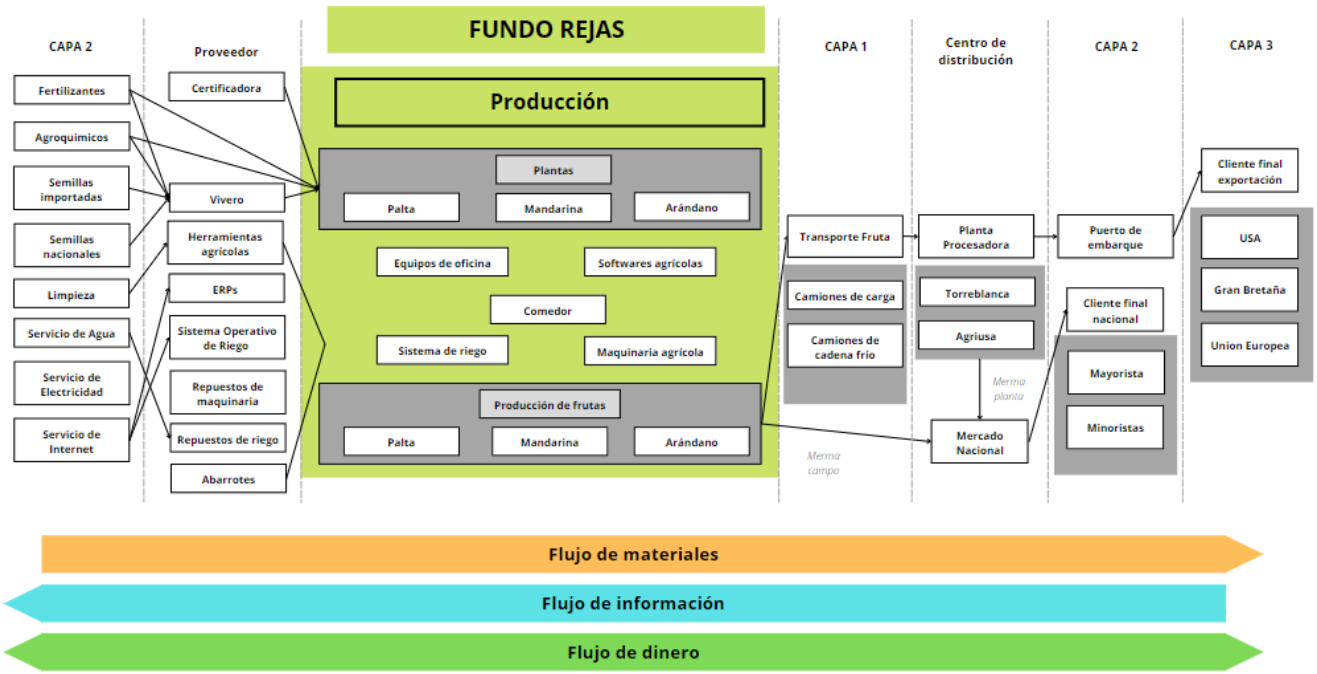


Figura N°22: Cadena de Suministros Fundos Rejas SAC

Fuente: *Elaboración Propia*

Fundo Rejas obtiene los plantones e injertos que reemplazan hectáreas de cultivo con plantas antiguas o se usan para hectáreas nuevas. Estos se consiguen mediante proveedores locales y extranjeros que poseen los permisos respectivos para la venta de los productos. Asimismo, se tiene proveedores locales para la adquisición de agroquímicos que son usados para evitar dañar la integridad de las plantas y sus frutos; y, fertilizantes importados obtenidos de empresas representantes de la matriz, los cuales se usan para nutrir a las plantas y lograr un rendimiento/planta óptima.

Respecto a la producción, en el caso de las mandarinas se exporta la variedad Tango, que tiene como fecha de inicio de campaña los meses de junio a octubre; con respecto a la campaña de paltas, se produce la variedad Hass y esta se da entre los meses de julio a setiembre y finalmente los arándanos son cosechados durante los meses de agosto a diciembre, y las variedades producidas son la Kestrell, Springhigh, Ventura y Biloxi. La producción de los tres cultivos varía dependiendo, principalmente, de factores meteorológicos, rendimiento de las plantas que no es uniforme en las zonas del lote de cultivo y precio de venta en el mercado no

es competitivo, lo cual ocasiona que se tenga que retrasar la cosecha de la fruta. Una vez cosechadas en campo son trasladadas a la zona de despacho mediante tractores de carga que a su vez contienen una carreta que permite aumentar su capacidad.

Una vez en la zona de despacho, los arándanos pasan por evaluaciones de calidad exhaustivas debido a ser un producto susceptible a daños, luego es pesado y embarcado en bulks dentro de un camión frigorífico que debe tener una temperatura de 18°C. En el caso de las paltas y mandarinas, estas también pasan un control de calidad, pero son enviados en camiones de carga con mayor capacidad de carga por el mayor volumen que ocupan, sin tener la necesidad de ser enviados por cadena en frío.

La planta procesadora recibe la carga del Fundo y brinda el servicio de pesaje, control de calidad y selección de la fruta en base al país destino. Este procesamiento tarda 2 días en llevarse a cabo lo cual lleva al almacenamiento de la fruta en frío para no perjudicar la integridad de esta. Finalmente, se colocan las mandarinas y paltas en cajas con sus respectivas categorías definidas por la calidad, peso y calibre y en el caso de los arándanos estos son enviados en clamshells o bulks para su venta al granel.

Respecto a la producción, Fundo Rejas tiene como giro de negocio la producción de frutas de los cultivos de paltos, plantas de mandarina y arándanos. En época de campañas es preciso contar la infraestructura óptima para operar que proviene de recursos físicos como equipos de oficina (laptops, tabletas, celulares), sistema de riego por goteo que servirá para la ejecución del plan de fertirriego; maquinaria agrícola, como tractores de carga de la fruta en jabas desde los lotes de cultivo a la zona de despacho, así como también para las aplicaciones de agroquímicos usando bombas de aplicación e insumos para la alimentación del personal operativo y administrativo en las instalaciones.

Por otro lado, se requiere intangibles como softwares agrícolas para la programación y ejecución de actividades como lo son la programación por turnos, asignación de lotes a irrigar, monitoreo de los caudales y flujo de agua. Asimismo, se cuenta con un ERP que permite realizar procesos administrativos de corte agrícola, como guías de remisión de cosecha, transformación de productos, stock de fertilizantes y agroquímicos, entre otros.

Finalmente se procede a realizar la cosecha de los frutos en base a los inicios de campaña estimados que no se dan en paralelo, lo cual requiere una planificación operativa detallada de las actividades y labores involucradas aparte de la principal que es la cosecha.

Proveedores:

En el caso de las plantas, es necesario contar con un vivero que tenga las plantas nuevas a sembrar para cambiar antiguas que ya no presenten un rendimiento óptimo o alguna enfermedad que pueda afectar la producción del lote. Se requiere contar con certificaciones como Global, Fsma, ETI, Tesco, entre otros para obtener las habilitaciones para exportar a los mercados extranjeros. Las herramientas agrícolas para la cosecha como tijeras de cosecha, telescópicas para alcanzar los frutos que se encuentran en la copa de los árboles, guantes, jabas, así como también los EPPs del caso. Repuestos de maquinaria y riego para reparaciones y mantenimiento de los tractores (discos, filtros de aire, etc.) y de igual forma repuestos para el sistema de riego (tuberías PVC, mangueras, entre otros). Luego, los insumos para alimentación que incrementa en los picos de cosecha en la campaña, el sistema operativo del proveedor Talgil de origen israelí que se usa para la programación del riego y finalmente el ERP usado para procesos administrativos y agrícolas.

Clientes:

Una vez hecho el traslado de las jabas con frutas a la zona de despacho, los arándanos pasan por evaluaciones de calidad exhaustivas debido a ser un producto susceptible a daños, luego es pesado y embarcado en bulks dentro de un camión frigorífico que debe tener una temperatura de 18°C. En el caso de las paltas y mandarinas, estas también pasan un control de calidad, pero son enviados en camiones de carga con mayor capacidad de carga por el mayor volumen que ocupan, sin tener la necesidad de ser enviados por cadena en frío, El servicio de transporte es tercerizado.

La planta procesadora recibe la carga del Fundo y brinda el servicio de pesaje, control de calidad y selección de la fruta en base al país destino. Este procesamiento tarda 2 días en llevarse a cabo lo cual lleva al almacenamiento de la fruta en frío para no perjudicar la integridad de esta. Finalmente, se colocan las mandarinas y paltas en cajas con sus respectivas categorías definidas por la calidad, peso y calibre y en el caso de los arándanos estos son enviados en clumshells o bulks para su venta al granel.

Asimismo, al terminar el proceso se produce merma tanto de planta como campo que tiene como destino el mercado nacional que se compone de mayoristas y minoristas.

La fruta que aprueba los controles de calidad es enviada al mercado extranjero en embarcaciones de cadena frío a 15°C que tienen un tiempo de llegada de 15 en promedio al mercado de USA y un mes a los mercados de Gran Bretaña y Unión Europea.

3.1.3 Datos generales estratégicos de la empresa

3.1.3.1 Visión, misión y valores o principios

Misión: Llegamos al mundo con alimentos naturales, saludables, y de alta calidad.

Visión: Ser reconocidos como una empresa saludable, en nuestros negocios, en nuestros productos y en nuestras relaciones con la gente y el entorno

Valores

- Pasión
- Compromiso
- Excelencia
- Respeto

Filosofía

Pasión por hacer las cosas bien

3.1.3.2 Objetivos estratégicos

- Maximizar ventas exportables de paltas en el mercado en 25% de la Unión Europea
- Incrementar la rentabilidad del cultivo de arándanos con la variedad Spring high en un 10% con respecto al periodo 2021
- Adquirir 15 hectáreas para cultivo de paltas Antillano para incrementar la producción en 25% del 2023 con respecto al periodo actual.
- Optimizar los gastos operativos del Fundo en maquinaria y mano de obra en 10% para el periodo 2022-2023.

3.1.3.3 Evaluación interna y externa. FODA cuantitativo

Fortalezas	**Debilidades**
F1: Ingenieros agrónomos especializados F2: Empresa con certificaciones Global Gap, Grasp, ETI, FSMA. F3: Mejora en automatización de procesos administrativos y reportería. F4: Implementación de sistema de control presupuestario F5: Rendimiento óptimo de la planta F6: Uso de aplicativos para planificación de labores operativas	D1: Mala coordinación entre el área corporativa y operativa en temas de facturación D2: Mal aprovechamiento del software de riego D3: Falta de mantenimiento preventivo de maquinaria D4: Automatización de procesos agrícolas a mejorar D5: Mala cuantificación del personal requerido en las campañas de cosechas D6: Mal uso del software de clima, para la planificación de los cambios climáticos.
Oportunidades	**Amenazas**
O1: Avances tecnológicos en el sector. O2: Reducción de barreras de mercado y aranceles mediante TLCs y Alianza Pacifico O3: Incremento de la demanda de productos por mercados internacionales O4: Mayor demanda de productos orgánicos O5: Ampliación en áreas para cosecha. O6: Perú cuenta con una geografía apta para la siembra de diferentes cultivos	A1: Factores climatológicos A2: Inestabilidad política A3: Incidencia de plagas y/o enfermedades en los cultivos A4: Variabilidad de precios en insumos agrícolas A5: Incremento precio de insumos por el cierre de fronteras. A6: Tendencia a nuevas regulaciones en el sector.

Tabla N°5: FODA de Rejas SAC

Fuente: *Elaboración Propia*

		OPORTUNIDADES							AMENAZAS						
	FODA	O1	O2	O3	O4	O5	O6	Promedio	A1	A2	A3	A4	A5	A6	Promedio
FORTALEZAS	F1	7	4	3	9	9	9	6.8	8	3	9.5	8	8	8.5	7.5
	F2	5	7	9	8	4	6	6.5	8	1	10	1	6	8	5.7
	F3	8	2	2	4	7	6	4.8	1	2	3	6	1	1	2.3
	F4	7	4	8	3	5	3	5.0	1	7	6	8	8	2	5.3
	F5	7	6	9.5	7	8	9	7.8	1	1	6	8	2	8	4.3
	F6	8.5	5	7.5	8	9.5	7	7.6	1	1	8	2	4	5	3.5
	Promedio	7.1	4.7	6.5	6.5	7.1	6.7		3.3	2.5	7.1	5.5	4.8	5.4	
DEBILIDADES	D1	7	1	6	2	7	3	4.3	1	1	1	9	4	1	2.8
	D2	8	6.5	7	8.5	8	7	7.5	7	1	4	1	1	9	3.8
	D3	9	8	8	7	7.5	7	7.8	2	1	9	1	9	8	5.0
	D4	9.5	5	8	8	8.5	7	7.7	8	1	8	7	8	8	6.7
	D5	9.5	3	9.7	9	9.5	9.5	8.4	1	1	8	1	4	6	3.5
	D6	9.5	2	8	10	7	9	7.6	10	1	8	5	1	1	4.3
	Promedio	8.8	4.3	7.8	7.4	7.9	7.1		4.8	1.0	6.3	4.0	4.5	5.5	

Tabla N°6: FODA Cuantitativo

Fuente: *Elaboración Propia*

Se puede concluir que la fortaleza 5, rendimiento óptimo de la planta, tiene mayor relevancia en el aprovechamiento de las oportunidades, en la cual la que mayor puntuación tiene es en el incremento de la demanda de productos. Así mismo, la oportunidad 1, avances tecnológicos en el sector, es la más importante, ya que permite tener un mejor aprovechamiento con la implementación de sensores, drones, etc. sobre la información en los cultivos, programación de rutas de fumigación, o usar la información de los sensores y crear modelos predictivos que ayuden a tener una mejor productividad, reducir costos.

La debilidad 5, mala cuantificación del personal requerido en las campañas de cosechas es la principal debilidad que presenta esta empresa, así como el objetivo de la creación de este proyecto, debido a que Fundos no tiene la forma de determinar el personal óptimo para suplir con las labores de cosecha, lo cual tiene una relación directa con los planes estratégicos de la empresa. Por otro lado, la amenaza más importante que se obtuvo es A3, la incidencia de plagas y/o enfermedades en los cultivos, ya que al ser un factor externo la presencia de estos impacta de manera significativa mermando la productividad y calidad de los cultivos, lo que conlleva en posibles pérdidas para la empresa.

3.2 Modelo de negocio actual (CANVAS)

<table>
<tr>
<td rowspan="2">Socios Claves

1. Procitrus S.A.C
2. Prohass
3. Consorcio de productores de frutas
4. Coorporación Financiera de Inversiones
Proveedores</td>
<td>Actividades Claves

1. Cultivo/cosecha de palta
2. Cultivo/cosecha de mandarina
3. Cultivo/cosecha de arándano</td>
<td rowspan="2">Propuesta de valor

1. Productos de alta calidad para la exportación</td>
<td>Relación con Clientes

Correo electronico
Web de consulta: Sedex</td>
<td rowspan="2">Segmentos de Clientes

CPF (Europea, Gran Bretaña y Estados Unidos)
Comerciantes mayoristas</td>
</tr>
<tr>
<td>Recursos Claves

1. Fertilizantes
2. Mochila de aplicación de Agroquimicos
3. Maquinaria y equipos
4. Sistema de Riego
5. Herramientas de cosecha
6. Pozos Agua
7. Operarios de cosecha
8. Personal especialista en estándares de calidad en los cultivos
9. Ingenieros agrónomos especializados en manejo de cultivos de palta, mandarina y arándano</td>
<td>Canales

Transporte maritimo
Transporte terrestre (cadena de frio - almacén)
Cadena de distribución</td>
</tr>
<tr>
<td colspan="3">Estructura de Costes

1. Infraestructura
2. Insumos,(Fertilizantes, agroquimicos)
3. Maquinaria (Tractores) y equipos (bombas de aplicación)
4. Mantenimiento de maquinaria
5. Mantenimiento de campo (Poda)
6. Costos Logisticos (Flete, transporte)
7. Alquiler de maquinaria
8. Servicio de Polinización
9. Mano de Obra
10. Almacenamiento
11. Electricidad - Sistema de Riego</td>
<td colspan="2">Fuente de Ingresos

1. Exportación
2. Venta local
3. %Merma</td>
</tr>
</table>

Tabla N°7: Modelo CANVAS Fundos Rejas SAC

Fuente: *Elaboración Propia*

En el modelo CANVAS planteado, podemos identificar a nuestros aliados clave, como Procitrus SAC que brindan precios competitivos de insumos como agroquímicos, fertilizantes para mandarina, al igual que Prohass brindando insumos para el cultivo del palto. Por otra parte, las principales actividades dentro del fundo son la cosecha y cultivo del palto, mandarina y arándano. Además, nuestro principal cliente es el Consorcio de productores de fruta (CPF), que se encarga de la exportación a EE.UU. y varios países de Europa, por otro lado, otro principal cliete, son los comerciantes mayoristas.

3.3 Mapa de procesos actual

Los procesos que se logran identificar en la empresa están repartidos en tres grupos: procesos estratégicos, procesos operativos y procesos de soporte. Los procesos estratégicos consisten en procesos que dan como resultado la toma de decisiones y plantear objetivos en la organización en base a las proyecciones de cosecha de la parte operativa, lo cual va ligado al flujo de caja proyectado y el control de gastos permisibles para cada mes (siendo los gastos más fuerte del rubro en fertilizantes, agroquímicos, mano de obra y maquinaria); los procesos de operaciones, o procesos claves, constituyen en la razón de ser de la organización, siendo en rubro agrícola desde la parte del campo, la cosecha de la fruta (producción), control interno en el fundo de las frutas para mejorar la cantidad de fruta exportada y disminuir la cantidad de fruta vendida como descarte y mercado nacional, los cuales no son el Core Business de la empresa; con respecto a la Sanidad, está compuesta por evaluación de las plagas y enfermedades que atacan a los cultivos para determinar acciones que reduzcan el impacto en los mismos, antes de llegar al inicio de las campañas. Con respecto al Transporte, se refiere a los camiones de carga que llevan las frutas en jabas a la planta procesadora y también al transporte de personal al inicio y fin de la jornada laboral. Finalmente, los procesos de soporte ayudan a llevar a cabo estos objetivos de la organización.

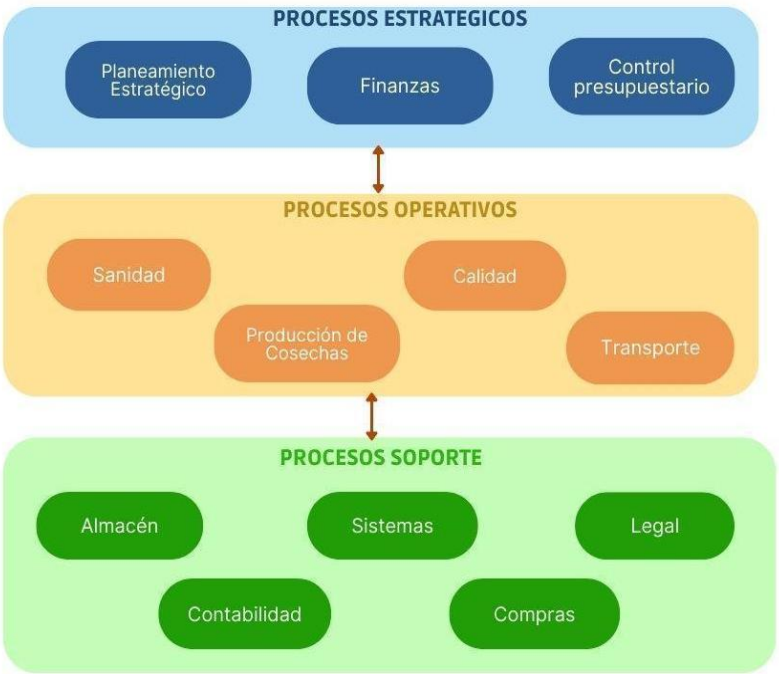


Figura N°23: Mapa de procesos

Fuente: *Elaboración Propia*

CAPITULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

En este capítulo daremos a conocer el enfoque, diseño y alcance de la investigación. Con ello, podremos diseñar la metodología a implementar, a fin de poder cumplir con nuestros objetivos de investigación.

4.1 Diseño de la Investigación

4.1.1 Enfoque de la investigación

Hernández & Mendoza (2018) nos indican que, "los estudios cuantitativos pretenden describir, explicar y predecir los fenómenos investigados, buscando regularidades y relaciones causales entre elementos (variables). Esto significa que la meta principal es la prueba de hipótesis y la formulación y demostración de teorías".

El siguiente trabajo tiene un enfoque cuantitativo, dado que el objetivo es predecir la producción por cultivos, además de la cantidad de personal requerido en las campañas de cosecha utilizando datos históricos cuantitativos.

4.1.2 Alcance de la investigación

Los alcances de la investigación pueden ser exploratorios, descriptivos, correlacionales y explicativos, afirman Hernández & Mendoza (2018), Asimismo nos indican que un alcance correlacional es "Investigaciones que pretenden asociar conceptos, fenómenos, hechos o variables. Miden las variables y su relación en términos estadísticos".

Por lo anterior descrito, la presente investigación tendrá un alcance correlacional, en el cual mediremos nuestras variables climatológicas relacionadas con nuestras variables de producción y tareo. Finalmente se pretende obtener 2 modelos de regresión que nos permita predecir la cantidad de producción de frutos y el número de personas a contratar según cada campaña de cosecha en el Fundo Rejas.

4.1.3 Diseño o tipo de la investigación

Una investigación de diseño experimental es definida por Pimienta & De La Orden (2017) como:

Este tipo de investigación consiste en la manipulación deliberada o recreación de alguna situación u objeto estudiado (variables) que no han sido comprobados con saber, con la finalidad de descubrir, analizar e identificar por qué, cómo, cuándo o para qué tienen lugar determinadas reacciones, cambios, modificaciones del objeto o situación; para finalmente, conocer y comprender las consecuencias o los posibles efectos causados por dicho objeto, situación o acontecimiento en particular (p. 10)

En este sentido, el presente trabajo de investigación tiene un diseño experimental, dado que las variables dependientes del primer modelo como los factores climatológicos se identificarán en diferentes momentos a lo largo de los años, con la finalidad de determinar cómo fluctúa nuestra variable dependiente producción por cultivo. Así mismo, en nuestro segundo modelo las variables independientes como los factores climatológicos, tareo y producción permiten determinar la cantidad de personal requerido, por lo que se buscará determinar el comportamiento de las variables para poder obtener 2 modelos de regresión.

4.1.4 Población y muestra

Esta investigación tuvo como población la producción por cultivo de palta, arándano y mandarina, así como la labor de cosecha por tipo de cultivo y las variables climatológicas, el intervalo de tiempo de los datos es desde abril de 2019 hasta Julio de 2022.

4.2 Metodología de implementación de la solución

Se conocen distintas metodologías para la implementación de la solución, entre ellas tenemos, CRISP DM, KDD y SCRUM.

En este trabajo se utilizará la metodología CRISP DM, debido a que es una de las metodologías más completas y de la que mejor se acomoda a la propuesta. Esta metodología consta de 6 etapas: Comprensión del negocio, Entendimiento de datos, Preparación de los datos, Modelado, Evaluación y Despliegue. Sin embargo, para este trabajo, sólo se utilizará las primeras 5 etapas, ya que se trata de un trabajo experimental.

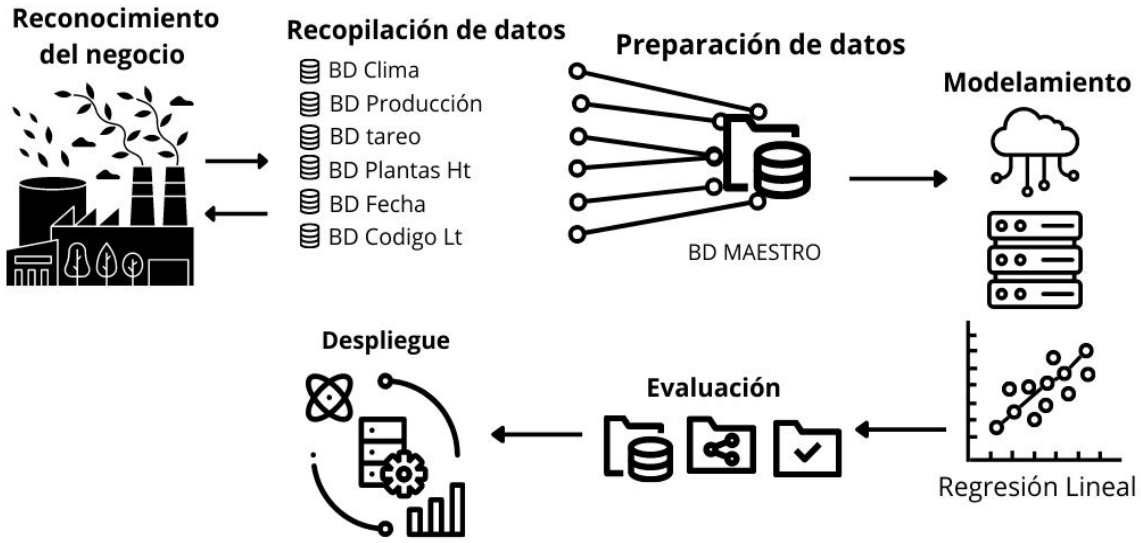


Figura N°24: Diseño de la implementación

Fuente: *Elaboración Propia*

1. Etapa 1 - Reconocimiento del negocio:

En esta etapa, entendemos el negocio en sí, como funciona y los objetivos estratégicos de la organización. También, se analiza la problemática de la empresa, a fin de poder solucionar con el proyecto.

2. Etapa 2 - Recopilación de la base datos:

En esta etapa, se recopiló toda la información brindada por la empresa. Sin embargo, para la realización de este proyecto, se utilizaron seis bases de datos. Debido a que esos datos nos sirven para la formulación de la metodología.

3. Etapa 3 - Preparación de datos:

En esta etapa, se realiza una limpieza de los datos, quiere decir, se detecta anomalías en la data. En este caso, se encontró que en las bases de datos, tiene data repetida, espacios vacíos. Por lo que se procedió a depurar la data sobrante, organizarla y fusionarla en una base de datos Maestra.

4. Etapa 4 - Modelado:

Una vez organizada la data según los parámetros establecidos, se pasa a la siguiente etapa que es el modelamiento de la data, en él se aplica machine learning utilizando para un modelo supervisado, debido a que se cuenta con los datos de entrada y de salida.

Etapa 5 - Evaluación de los resultados

Finalmente, en esta etapa, se pone a prueba el modelamiento de la data. Se valida el modelo empleado, donde se realizan varias pruebas, para poder comprobar el funcionamiento del modelo, y así poder pronosticar la cantidad de trabajadores a contratar para cada cultivo.

4.3 Metodología para la medición de resultados de la implementación

Para la evaluación y/o validación del modelo se utilizará la regresión lineal, SVR y Árboles de Regresión. Para este modelo se tiene que evaluar los errores del modelo. En él se encuentran RMSE y R^2.

Para la programación en el programa Python, primero se utilizará la librería *Sklearn* del módulo de *SimplelImpute para* poder reemplazar los valores perdidos para las categorías. Se pueden utilizar diferentes estrategias para modular los datos, entre ellas, la mediana, la media o una constante.

```
from sklearn.impute import SimpleImputer
```

Figura N°25: Sklearn.impute

Además, para la evaluación del método planteado, utilizaremos la regresión lineal de la librería *Sklearn*, el módulo *LinearRegression.*

```
from sklearn.linear_model import LinearRegression
```

Figura N°26: LinearRegression

Para algunas variables utilizaremos el método de *Pearson* para poder hallar la correlación entre ellas, y poder saber que tan relevante son para el modelo.

```
corr_mand=Base_Mandarina.corr(method="pearson")
```

Figura N°27: Correlación de Pearson

4.4 Cronograma de actividades y presupuesto

A continuación, en la tabla N° 8 se presenta el cronograma de actividades que se realizan en cada semana, desde junio a agosto del 2022, para poder llevar a cabo el proyecto.

Actividades	JUNIO		JULIO					AGOSTO			
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
INICIO											
Indentificación de la empresa											
Indentificación de la problemática											
PLANIFICACIÓN											
Recopilación de datos											
Reconocimiento de variables											
Desarrollo del marco de referencia											
Planteamiento de la metodología											
DESARROLLO											
Limpieza - pre procesamiento datos											
Modelamiento de datos											
Evaluación/Validación de Resultados											
CIERRE											
Conclusiones y Recomendaciones											
Presentación de la propuesta											
Sustentación											

Tabla N°8 Cronograma de Actividades

Fuente: *Elaboración propia*

Así mismo, se presenta el presupuesto en la tabla N°9, en él se muestran los recursos a utilizar a lo largo del proyecto.

Recursos	Cantidad	Monto	Sub Total
INICIO			
Equipo de trabajo: Laptop	5	S/ 200.00	S/1,000.00
Personal de trabajo	5	S/1,200.00	S/6,000.00
PLANIFICACIÓN - DESARROLLO - CIERRE			
Programas			
MS Office	5	S/ 150.00	S/ 750.00
Anaconda: Python	5	S/ -	S/ -
Meet	1	S/ -	S/ -
Servicios			
Electricidad	5	S/ 100.00	S/ 500.00
Internet	5	S/ 50.00	S/ 250.00
		Total	**S/8,500.00**

Tabla N°9 Presupuesto del proyecto

Fuente: *Elaboración propia*

CAPÍTULO V: Desarrollo

5.1 Propuesta Solución

El trabajo presentado tiene como finalidad desarrollar dos modelos predictivos. El primero logrará determinar la producción a proyectar en kg de cada cultivo cosechado dentro del Fundo Rejas en base las variables meteorológicas que predominan en la zona y variables de agrícolas como cantidad de plantas por lote, hectáreas y edad de las plantas. Asimismo, el segundo modelo permitirá predecir la cantidad de fuerza laboral para cada cultivo usando para este propósito variables climatológicas, de producción, agrícolas y de fuerza laboral. Dentro de los alcances presentados en este trabajo, se presenta hasta la evaluación de la propuesta mediante el uso de la herramienta de programación Python para predecir los modelos.

5.1.1 Planeamiento y descripción de Actividades

Para el planteamiento de la solución se utilizó la metodología CRISP-DM, siguiendo las etapas mostradas en la figura 24, donde se detallan los pasos a seguir.

5.1.1.1 Reconocimiento de la empresa

Para esta primera etapa, se mantuvo constante contacto con la empresa puesto que se contó con información de primera mano. Con ello, se buscó lograr una mejor comprensión de los procesos internos y datos históricos de los reportes extraídos del ERP de la empresa.

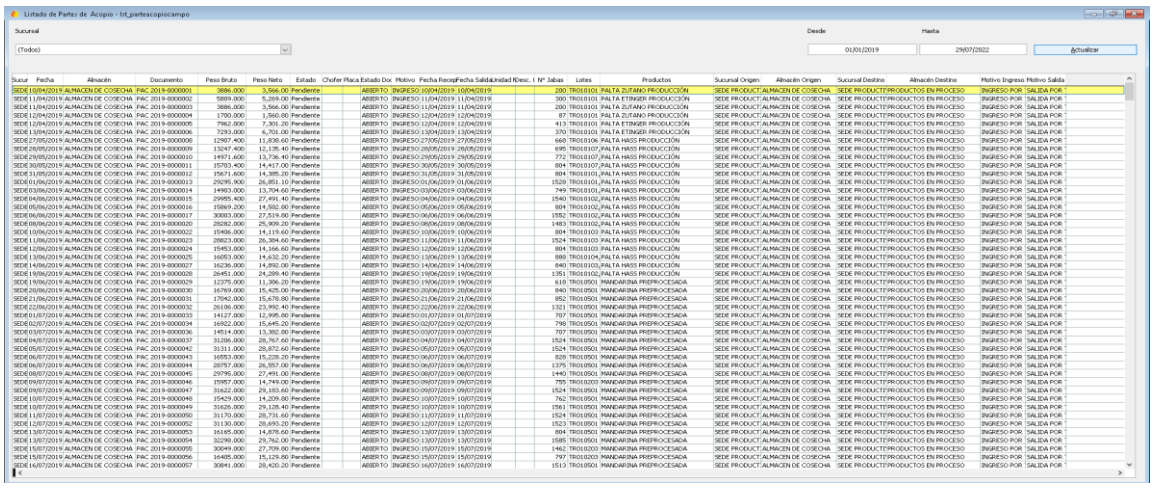


Figura N°28: Módulo de Cosecha - Parte de Acopio de Campo Fundo Rejas

Para ello se inició el proceso dando una explicación a grandes rasgos sobre el rubro del negocio, mostrando el ERP que se maneja en la empresa, a manera de instructivo, para los módulos de Cosecha, Recursos Humanos y Factores Climatológicos, estando este último ubicado en otro repositorio llamado Weatherlink.

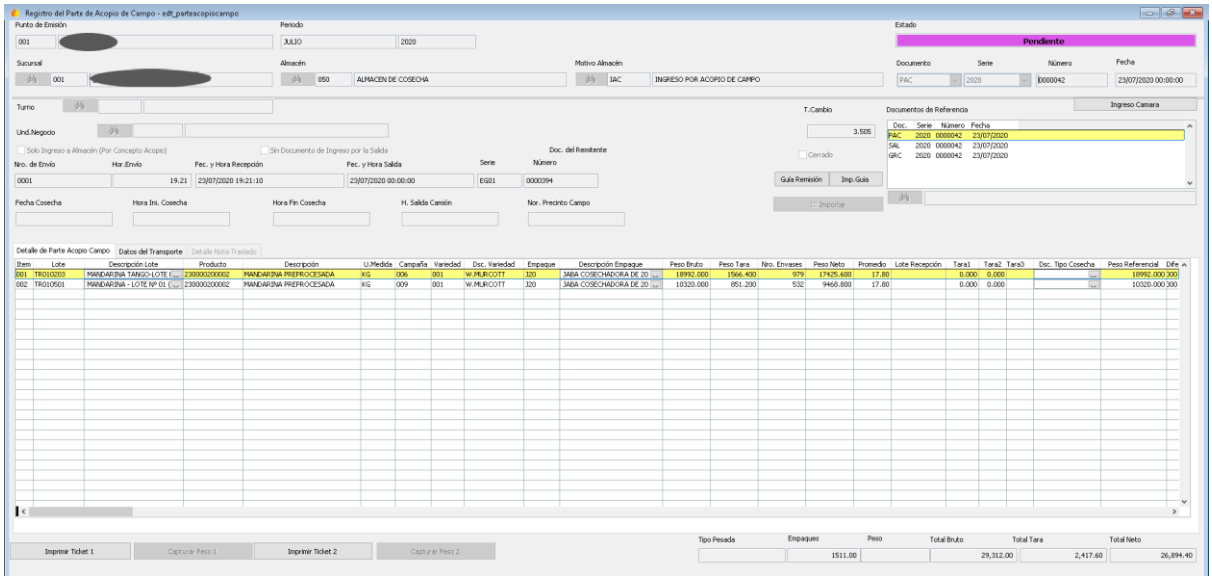


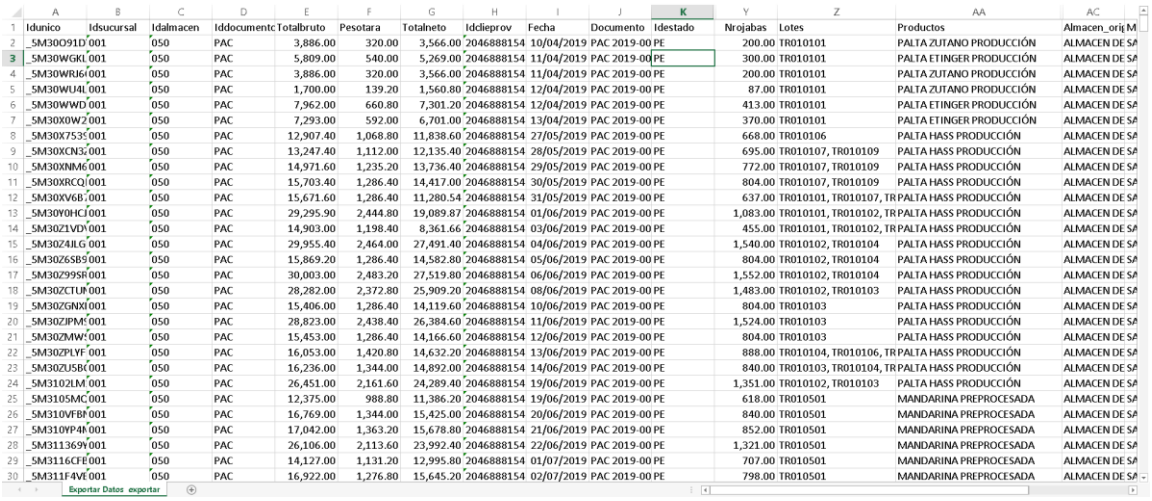
Figura N°29: Módulo de Cosecha - Registro de Parte de Acopio de Campo Fundo Rejas

5.1.1.2 Recopilación de Datos

En esta etapa, se recopiló información que nos brindó la empresa proveniente de su ERP. Con ellas, se estructuraron seis bases de datos (BD). Entre ellas, BD Clima, BD Producción, BD Tareo, BD Plantas Ht/Lt, BD Fecha y BD de Códigos de los Lotes de Cultivo.

- BD Clima: Tabla de datos de los factores meteorológicos, como temperatura promedio, velocidad del viento, índice UV, nivel de humedad, densidad del aire, etc. de los cultivos (palta, mandarina y arándanos)
- BD Producción: En esta tabla, se encuentran la cantidad de cosecha por día, desde 2019 a 2022.
- BD Tareo: Se encuentran los reportes de los trabajadores por actividad que realizan en el fundo, sea cosecha, limpieza, preparación, fumigación, etc.
- BD Planta Ht/Lote: En esta base de datos, se encuentra la cantidad de plantas que hay por hectáreas y las edades promedio de las plantas sembradas en las mismas. Esta información es importante, y que se requiere para tener mapeado el cultivo.

- BD Fecha: Se ubican las fechas agrupadas en semanas, en las cuales hubo actividad de cosecha, desde el año 2019 a 2022.
- BD Código Lote: Tabla que resume los códigos de consumidores o lotes de cultivos de palta, mandarina y arándanos, respectivamente.

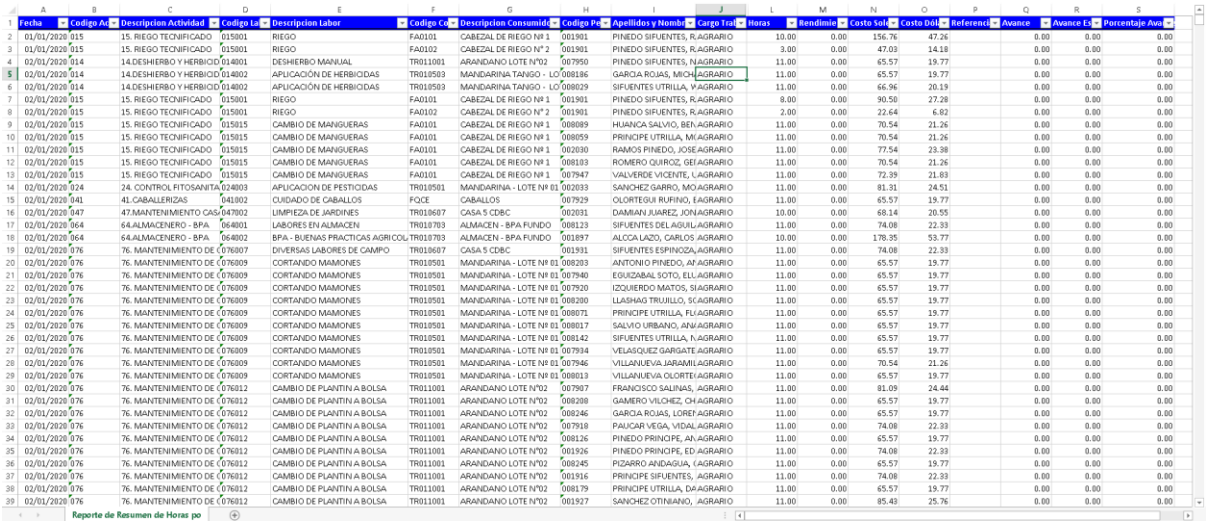


Idunico	Idsucursal	Idalmacen	Iddocumentc	Totalbruto	Pesotara	Totalneto	Iddieprov	Fecha	Documento	Idestado	Nrojabas	Lotes	Productos	Almacen_orig
_5M30O91D	001	050	PAC	3,886.00	320.00	3,566.00	2046888154	10/04/2019	PAC 2019-00	PE	200.00	TR010101	PALTA ZUTANO PRODUCCIÓN	ALMACEN DE SA
_5M30WGKL	001	050	PAC	5,809.00	540.00	5,269.00	2046888154	11/04/2019	PAC 2019-00	PE	300.00	TR010101	PALTA ETINGER PRODUCCIÓN	ALMACEN DE SA
_5M30WRJ6	001	050	PAC	3,886.00	320.00	3,566.00	2046888154	11/04/2019	PAC 2019-00	PE	200.00	TR010101	PALTA ZUTANO PRODUCCIÓN	ALMACEN DE SA
_5M30WU4L	001	050	PAC	1,700.00	139.20	1,560.80	2046888154	12/04/2019	PAC 2019-00	PE	87.00	TR010101	PALTA ZUTANO PRODUCCIÓN	ALMACEN DE SA
_5M30WWD	001	050	PAC	7,962.00	660.80	7,301.20	2046888154	12/04/2019	PAC 2019-00	PE	413.00	TR010101	PALTA ETINGER PRODUCCIÓN	ALMACEN DE SA
_5M30X0W2	001	050	PAC	7,293.00	592.00	6,701.00	2046888154	13/04/2019	PAC 2019-00	PE	370.00	TR010101	PALTA ETINGER PRODUCCIÓN	ALMACEN DE SA
_5M30X753	001	050	PAC	12,907.40	1,068.80	11,838.60	2046888154	27/05/2019	PAC 2019-00	PE	668.00	TR010106	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30XCN3	001	050	PAC	13,247.40	1,112.00	12,135.40	2046888154	28/05/2019	PAC 2019-00	PE	695.00	TR010107, TR010109	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30XNM6	001	050	PAC	14,971.60	1,235.20	13,736.40	2046888154	29/05/2019	PAC 2019-00	PE	772.00	TR010107, TR010109	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30XRCQ	001	050	PAC	15,703.40	1,286.40	14,417.00	2046888154	30/05/2019	PAC 2019-00	PE	804.00	TR010107, TR010109	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30XV6B	001	050	PAC	15,671.60	1,286.40	11,280.54	2046888154	31/05/2019	PAC 2019-00	PE	637.00	TR010101, TR010107, TR	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30Y0HC	001	050	PAC	29,295.90	2,444.80	19,089.87	2046888154	01/06/2019	PAC 2019-00	PE	1,083.00	TR010101, TR010102, TR	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30Z1VD	001	050	PAC	14,903.00	1,198.40	8,361.66	2046888154	03/06/2019	PAC 2019-00	PE	455.00	TR010101, TR010102, TR	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30Z4JLG	001	050	PAC	29,955.40	2,464.00	27,491.40	2046888154	04/06/2019	PAC 2019-00	PE	1,540.00	TR010102, TR010104	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30Z6SB	001	050	PAC	15,869.20	1,286.40	14,582.80	2046888154	05/06/2019	PAC 2019-00	PE	804.00	TR010102, TR010104	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30Z99S	001	050	PAC	30,003.00	2,483.20	27,519.80	2046888154	06/06/2019	PAC 2019-00	PE	1,552.00	TR010102, TR010104	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZCTU	001	050	PAC	28,282.00	2,372.80	25,909.20	2046888154	08/06/2019	PAC 2019-00	PE	1,483.00	TR010102, TR010103	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZGNX	001	050	PAC	15,406.00	1,286.40	14,119.60	2046888154	10/06/2019	PAC 2019-00	PE	804.00	TR010103	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZJPM	001	050	PAC	28,823.00	2,438.40	26,384.60	2046888154	11/06/2019	PAC 2019-00	PE	1,524.00	TR010103	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZMW5	001	050	PAC	15,453.00	1,286.40	14,166.60	2046888154	12/06/2019	PAC 2019-00	PE	804.00	TR010103	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZPLYF	001	050	PAC	16,053.00	1,420.80	14,632.20	2046888154	13/06/2019	PAC 2019-00	PE	888.00	TR010104, TR010106, TR	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M30ZU5B	001	050	PAC	16,236.00	1,344.00	14,892.00	2046888154	14/06/2019	PAC 2019-00	PE	840.00	TR010103, TR010104, TR	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M3102LM	001	050	PAC	26,451.00	2,161.60	24,289.40	2046888154	19/06/2019	PAC 2019-00	PE	1,351.00	TR010102, TR010103	PALTA HASS PRODUCCIÓN	ALMACEN DE SA
_5M3105MC	001	050	PAC	12,375.00	988.80	11,386.20	2046888154	19/06/2019	PAC 2019-00	PE	618.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA
_5M310VFB	001	050	PAC	16,769.00	1,344.00	15,425.00	2046888154	20/06/2019	PAC 2019-00	PE	840.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA
_5M310YP4	001	050	PAC	17,042.00	1,363.20	15,678.80	2046888154	21/06/2019	PAC 2019-00	PE	852.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA
_5M311369	001	050	PAC	26,106.00	2,113.60	23,992.40	2046888154	22/06/2019	PAC 2019-00	PE	1,321.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA
_5M3116CF	001	050	PAC	14,127.00	1,131.20	12,995.80	2046888154	01/07/2019	PAC 2019-00	PE	707.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA
_5M311F4V	001	050	PAC	16,922.00	1,276.80	15,645.20	2046888154	02/07/2019	PAC 2019-00	PE	798.00	TR010501	MANDARINA PREPROCESADA	ALMACEN DE SA

Exportar Datos exportar

Figura N°30: Reporte de Producción (2019 - 2022)

En el caso del reporte de producción, se identificaron datos relevantes como Kg netos de cosecha, números de jabas empleadas para la labor, la fecha en la cual se realizó el trabajo, el código de los lotes de los cultivos respectivos y el tipo de cultivo al cual corresponden.



Fecha	Codigo Ac	Descripcion Actividad	Codigo La	Descripcion Labor	Codigo Co	Descripcion Consumid	Codigo Pe	Apellidos y Nombr	Cargo Tral	Horas	Rendimie	Costo Sole	Costo Dól	Referenci	Avance	Avance Es	Porcentaje Ava
01/01/2020	015	15. RIEGO TECNIFICADO	015001	RIEGO	FA0101	CABEZAL DE RIEGO Nº 1	001901	PINEDO SIFUENTES, R	AGRARIO	10.00	0.00	156.76	47.26		0.00	0.00	0.00
01/01/2020	015	15. RIEGO TECNIFICADO	015001	RIEGO	FA0102	CABEZAL DE RIEGO Nº 2	001901	PINEDO SIFUENTES, R	AGRARIO	3.00	0.00	47.03	14.18		0.00	0.00	0.00
02/01/2020	014	14.DESHIERBO Y HERBICID	014001	DESHIERBO MANUAL	TR011001	ARANDANO LOTE N°02	007950	PINEDO SIFUENTES, N	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	014	14.DESHIERBO Y HERBICID	014002	APLICACIÓN DE HERBICIDAS	TR010503	MANDARINA TANGO - LO	008186	GARCIA ROJAS, MICH	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	014	14.DESHIERBO Y HERBICID	014002	APLICACIÓN DE HERBICIDAS	TR010503	MANDARINA TANGO - LO	008029	SIFUENTES UTRILLA, V	AGRARIO	11.00	0.00	66.96	20.19		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015001	RIEGO	FA0101	CABEZAL DE RIEGO Nº 1	001901	PINEDO SIFUENTES, R	AGRARIO	8.00	0.00	90.50	27.28		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015001	RIEGO	FA0102	CABEZAL DE RIEGO Nº 2	001901	PINEDO SIFUENTES, R	AGRARIO	2.00	0.00	22.64	6.82		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015015	CAMBIO DE MANGUERAS	FA0101	CABEZAL DE RIEGO Nº 1	008089	HUANCA SALVIO, BEN	AGRARIO	11.00	0.00	70.54	21.26		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015015	CAMBIO DE MANGUERAS	FA0101	CABEZAL DE RIEGO Nº 1	008059	PRINCIPE UTRILLA, M	AGRARIO	11.00	0.00	70.54	21.26		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015015	CAMBIO DE MANGUERAS	FA0101	CABEZAL DE RIEGO Nº 1	002030	RAMOS PINEDO, JOSE	AGRARIO	11.00	0.00	77.54	23.38		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015015	CAMBIO DE MANGUERAS	FA0101	CABEZAL DE RIEGO Nº 1	008103	ROMERO QUIROZ, GEI	AGRARIO	11.00	0.00	70.54	21.26		0.00	0.00	0.00
02/01/2020	015	15. RIEGO TECNIFICADO	015015	CAMBIO DE MANGUERAS	FA0101	CABEZAL DE RIEGO Nº 1	007947	VALVERDE VICENTE, L	AGRARIO	11.00	0.00	72.39	21.83		0.00	0.00	0.00
02/01/2020	024	24. CONTROL FITOSANITA	024003	APLICACION DE PESTICIDAS	TR010501	MANDARINA - LOTE Nº 01	002033	SANCHEZ GARRO, MO	AGRARIO	11.00	0.00	81.31	24.51		0.00	0.00	0.00
02/01/2020	041	41.CABALLERIZAS	041002	CUIDADO DE CABALLOS	FQCE	CABALLOS	007929	OLORTEGUI RUFINO, E	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	047	47.MANTENIMIENTO CAS	047002	LIMPIEZA DE JARDINES	TR010607	CASA 5 CDBC	002031	DAMIAN JUAREZ, JON	AGRARIO	10.00	0.00	68.14	20.55		0.00	0.00	0.00
02/01/2020	064	64.ALMACENERO - BPA	064001	LABORES EN ALMACEN	TR010703	ALMACEN - BPA FUNDO	008123	SIFUENTES DEL AGUIL	AGRARIO	11.00	0.00	74.08	22.33		0.00	0.00	0.00
02/01/2020	064	64.ALMACENERO - BPA	064002	BPA - BUENAS PRACTICAS AGRICOL	TR010703	ALMACEN - BPA FUNDO	001897	ALCCA LAZO, CARLOS	AGRARIO	10.00	0.00	178.35	53.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076007	DIVERSAS LABORES DE CAMPO	TR010607	CASA 5 CDBC	001931	SIFUENTES ESPINOZA,	AGRARIO	11.00	0.00	74.08	22.33		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008203	ANTONIO PINEDO, AF	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	007940	EGUIZABAL SOTO, ELI	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	007920	IZQUIERDO MATOS, SI	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008200	LLASHAG TRUJILLO, S	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008071	PRINCIPE UTRILLA, FL	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008017	SALVIO URBANO, ANA	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008142	SIFUENTES UTRILLA, N	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	007934	VELASQUEZ GARGATE	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	007946	VILLANUEVA JARAMI	AGRARIO	11.00	0.00	70.54	21.26		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076009	CORTANDO MAMONES	TR010501	MANDARINA - LOTE Nº 01	008013	VILLANUEVA OLORTE	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	007907	FRANCISCO SALINAS,	AGRARIO	11.00	0.00	81.09	24.44		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	008208	GAMERO VILCHEZ, CH	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	008246	GARCIA ROJAS, LORE	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	007918	PAUCAR VEGA, VIDAL	AGRARIO	11.00	0.00	74.08	22.33		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	008126	PINEDO PRINCIPE, AN	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	001926	PINEDO PRINCIPE, ED	AGRARIO	11.00	0.00	74.08	22.33		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	008245	PIZARRO ANDAGUA, J	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	001916	PRINCIPE SIFUENTES,	AGRARIO	11.00	0.00	74.08	22.33		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	008179	PRINCIPE UTRILLA, DA	AGRARIO	11.00	0.00	65.57	19.77		0.00	0.00	0.00
02/01/2020	076	76. MANTENIMIENTO DE C	076012	CAMBIO DE PLANTIN A BOLSA	TR011001	ARANDANO LOTE N°02	001927	SANCHEZ OTINIANO,	AGRARIO	11.00	0.00	85.43	25.76		0.00	0.00	0.00

Reporte de Resumen de Horas po

Figura N°31: Reporte de Tareo (2020)

En lo respecta el reporte de tareo, tenemos entre los datos más relevantes: horas trabajadas (las cuales son asignadas por cada supervisor dependiendo la labor) por persona, persona que desempeñó la labor, código de lote, descripción del consumidor o lote, para reconocer si es un lote de cultivo o no y costo soles por horas totales trabajadas por persona en

la respectiva labor. Cabe resaltar que, para ambos modelos predictivos, la labor de cosecha será la única tomada en cuenta dentro las labores y actividades del Fundo, ya que la problemática se sitúa en el contexto de campañas de cosecha, donde esta labor es prioritaria.

Por último, para las tablas de BD Planta Ht/Lote, BD Fecha, BD Código Lote, se recopilaron los datos en base al levantamiento de información hecho en campo con el apoyo de gerente agrónomo de la empresa, para clasificar los lotes en sus cultivos, las plantas por lotes, hectáreas de cosecha y semanas de campaña (Ver Tablas en Anexos)

5.1.1.3 Preparación de los Datos

Una vez que se realizó la estructuración de las bases de datos, se procedió a depurar los datos. Para ello se empleó el software de Power BI, para hacer la manipulación de la data y tener una base relacionada con respecto a las bases de datos de producción, planta Ht/Lote, etc.

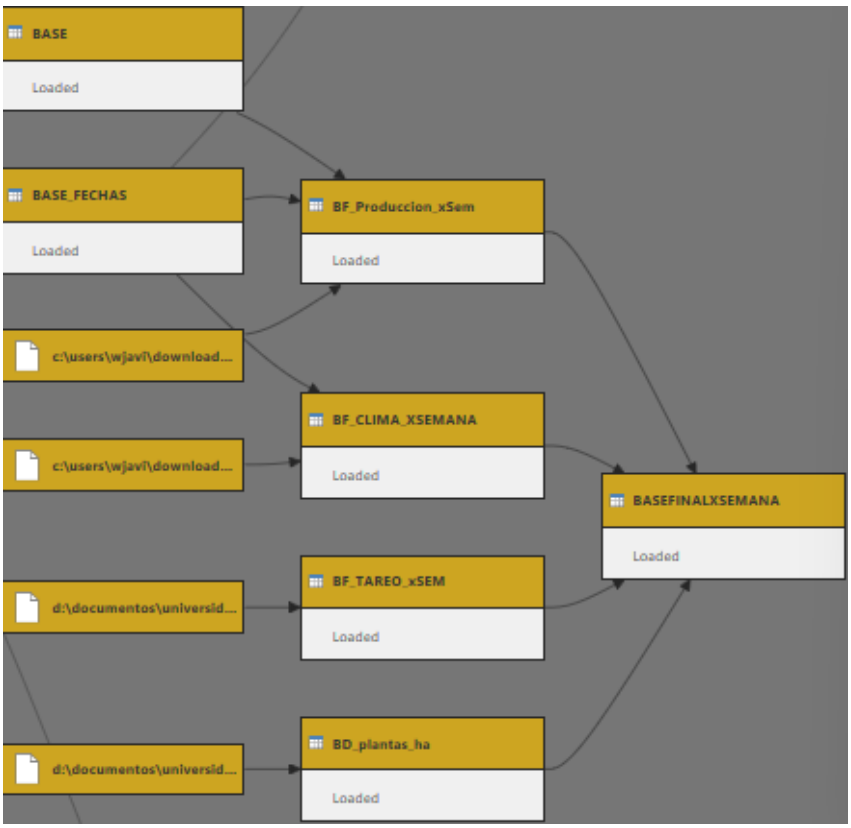


Figura N°32: Creación de la base final

Fuente: *Elaboración Propia*

Para ello, se identifican las variables a procesar, tanto las variables independientes, como dependientes. La nueva base de datos dará la unión de todas las BD presentados

anteriormente, para poder tener un mejor manejo de los datos, a esta BD se le llamó Maestro (BD_FINAL).

Variables	Grupo de Variable	Descripción	Unidad de medida
PERIODO	Identificador	Semanas de actividad de cosecha por año	-
DESCRIPCION	Datos de Cultivo	Referido al tipo de Cultivo del Fundo	Cultivo: Palta, Mandarinas y Arándanos
LT_Total	Datos de Cultivo	Lote de cultivo cosechados en periodo	Categorica (1: Un lote cosechado, 2: Dos lotes cosechados; 3: Tres lotes cosechados
AÑO	Datos de Cultivo	Año de actividad de cosecha	Periodo del 2019 - 2022
HorasTotales	Fuerza Laboral	Horas trabajadas efectivas de los cosechadores	Horas ejecutadas para cosecha
TotalNeto	Producción	Cantidad cosechada en KG de los cultivos	Kg por cultivo
Nrojabas	Producción	Representa la cantidad de jabas usadas para cosechar	Número de jabas usadas
Cantidad_Trabajadores	Fuerza Laboral	Total de trabajadores que cosecharon en el periodo respectivo	Cantidad de trabajadores
CostoTotal	Fuerza Laboral	Costo Total por horas trabajadas del personal de cosecha	Costo Mano de Obra
avg Temp Out	Clima	Promedio de temperatura externa en el ambiente	°C
avg Hi Temp	Clima	Promedio de temperatura máxima en el ambiente	°C
avg Low Temp	Clima	Promedio de temperatura mínima en el ambiente	°C
avg Out Hum	Clima	Promedio de humedad relativa en el ambiente	%
avg Dew Pt.	Clima	Promedio de punto de rocío (temperatura a que se debe enfriar el aire para que el vapor de agua se condense)	°C
avg Wind Speed	Clima	Promedio de velocidad del viento	metros/segundo, km/h
Wind Dir	Clima	Dirección del viento del periodo	Categorica, grados
avg Wind Run	Clima	Promedio de carrera del viento (monto de viento que transita durante la estación de un periodo)	m, km
avg Hi Speed	Clima	Promedio de alta velocidad del viento	metros/segundo, km/h
Hi Dir	Clima	Dirección del viento predominante del periodo	Categorica
avg Wind Chill	Clima	Promedio de sensación térmica (temperatura y velocidad del viento que calcula la pérdida de calor)	°C
avg Heat Index	Clima	Promedio de sensación térmica percibido en condiciones de calor del ambiente	°C
avg THW Index	Clima	Promedio de sensación térmica causado por el viento, temperatura en °C y humedad relativa	°C
avg Bar	Clima	Promedio de Presión Barométrica	Hpa (Hectopascal) o milibares
avg Rain	Clima	Promedio de cantidad de lluvia	mm
avg Rain Rate	Clima	Promedio de intensidad de lluvias	mm/hora
avg Solar Rad.	Clima	Promedio de cantidad de radiación solar	Watts por metro cuadrado
avg Solar Energy	Clima	Promedio de monto de energía solar	Langleys
avg Hi Solar Rad.	Clima	Promedio de radiación solar máxima	Langleys
avg UV Index	Clima	Promedio de indice de radiación ultravioleta	Índice ultravioleta
avg UV Dose	Clima	Promedio de intensidad de radiación ultravioleta	Meds (Dosis eritemal mínima)
avg Hi UV	Clima	Promedio de índice de radiación ultravioleta alta	Índice ultravioleta
avg Heat D-D	Clima	Promedio de energía necesaria diseñada para la calefacción	°C
avg Cool D-D	Clima	Promedio de energía necesaria diseñada para el enfriamiento	°C
avg In Temp	Clima	Promedio de temperatura interna	°C
avg In Hum	Clima	Promedio de humedad interna	°C
avg In Dew	Clima	Promedio de punto de rocío interno	°C
avg In Heat	Clima	Promedio de indice de calor en el interior	°C
avg In EMC	Clima	Promedio de compatibilidad electromagnética en el ambiente	-
avg In Air Density	Clima	Promedio de densidad del aire	kg/mL, g/mL
avg ET	Clima	Promedio de evapotranspiración (transferencia de agua a la atmósfera desde el suelo por evaporación)	mm
avg Wind Samp	Clima	Promedio de muestras de velocidad del viento	-
HA_TOTALES	Agrícola	Tamaño de lote de cultivo trabajado en el periodo respectivo	Hectáreas
PlantasTotales	Agrícola	Total de plantas cosechadas en el periodo respectivo	Cantidad de plantas
Edad_Prom	Agrícola	Edad promedio de las plantas cosechadas	Promedio de Edad de las Plantas

Tabla N°10: Descripción de variables

Fuente: *Elaboración propia*

La tabla muestra una total de 44 variables de tipo numéricas, descriptivas y categóricas de las cuales fueron usadas para los 2 modelos. Estas fueron elegidas mediante juicio de expertos en el rubro y una depuración realizada en Python, previa a la etapa de modelado de datos.

Modelo 1:

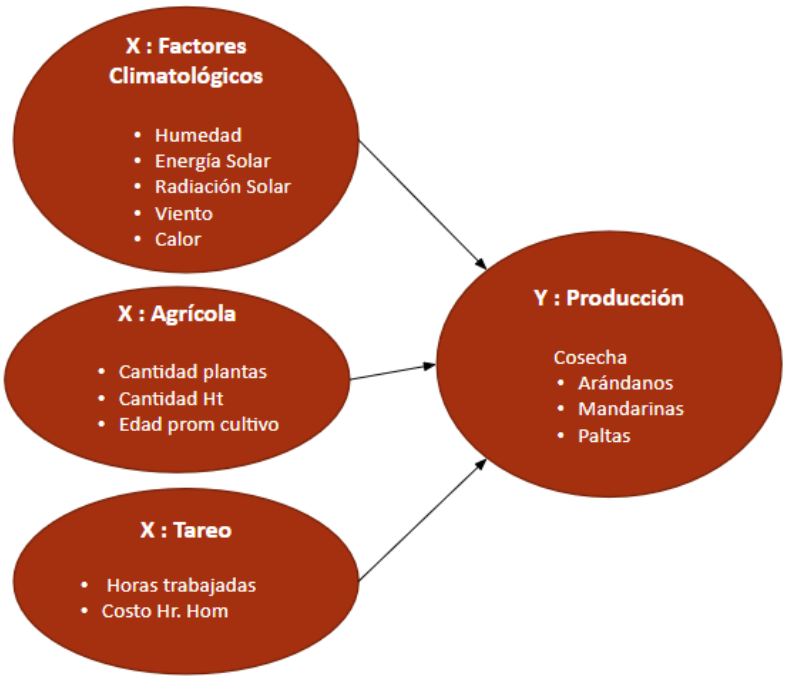


Figura N°33: Diagrama de Bayes para el Modelo 1

Fuente: *Elaboración Propia*

En este caso, se optó por promediar las edades de las plantas de cada lote de un mismo cultivo; para las hectáreas, se sumaron las de un mismo cultivo que fue trabajado por cada semana, siendo estos datos enlazados por los códigos de consumidores y semanas por año, para así evitar cruces de códigos entre cultivos y cruces de fechas con cultivos que también se cosecharon la misma semana, esto dada la operatividad que maneja la empresa para asignar tareas.

Modelo 2:

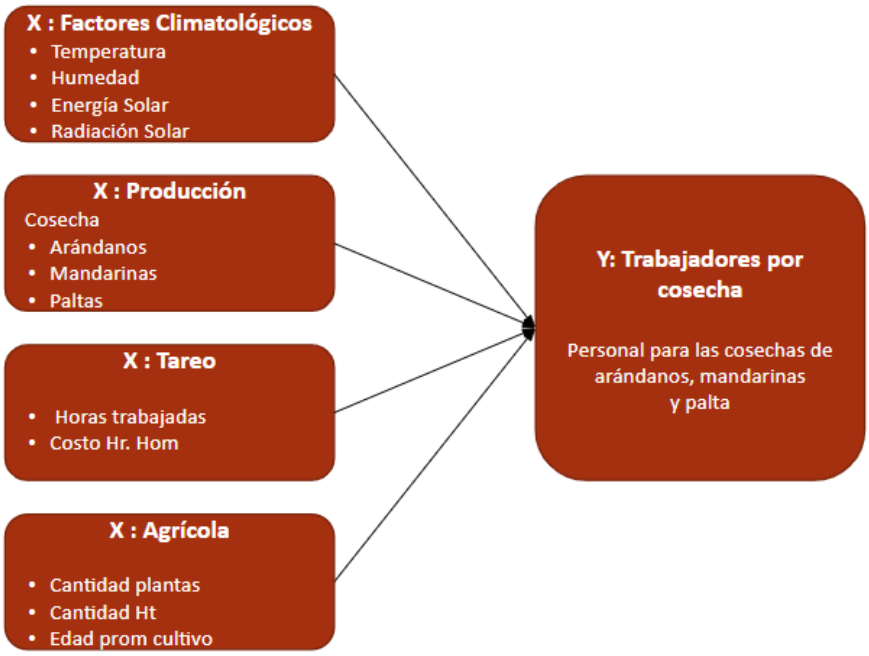


Figura N°34: Diagrama de Bayes para el Modelo 2

Fuente: *Elaboración Propia*

Para este modelo se unieron las tablas mediante las variables PERIODO, DESCRIPCION y LT_Total, con el fin de saber en qué semana del año se realizó asignó las tareas de cosecha, en qué cultivo se dio la labor y a que lotes se asignó el personal para dicha actividad, respectivamente.

En ambos casos, se detectaron errores de ingreso de datos, por tests de prueba del sistema, anulaciones de ingreso en producción, errores de lectura a usar el aplicativo de tareo, siendo estas filas eliminadas de los reportes que empleamos, por considerarse datos atípicos.

Limpieza o imputación de datos

Las bases de datos escogidas contaban con desperfectos, para la imputación de las variables se estará tomando en consideración que las variables deben tener como mínimo un 60% de registros completos, el cual se realizará mediante el siguiente código.

```
na_ratio = ((1-(data.isnull().sum() / len(data)))).sort_values(ascending = False)
na_ratio.to_csv("Porcentaje de Registros Completos.csv")
print(na_ratio)
```

Figura N°35: Cálculo de registros completos

Fuente: *Elaboración Propia*

Como se observa en la Tabla 1 del anexo se observa que todas las variables contienen más del 90% de información.

Variables Meteorológicas

Para ello primero se cambiará las variables categóricas “Wind Dir” y “Hi Dir”, a númericas

```
direccion_={'W':1,'SSW':2,'SSE':3,'NNW':4,'S':5,'WSW':6,'SE':7,'WNW':8}
```

```
data['Wind Dir']=data['Wind Dir'].map(direccion_)
data['Hi Dir']=data['Hi Dir'].map(direccion_)
```

Figura N°36: Cambio de variables categóricas a numéricas

Fuente: *Elaboración Propia*

El cual obtendremos el siguiente resultado

	Wind Dir	Hi Dir
0	1.0	1.0
1	1.0	1.0
2	1.0	1.0
3	1.0	1.0
4	1.0	1.0
...	...	...
220	NaN	NaN
221	NaN	NaN
222	7.0	7.0
223	7.0	NaN
224	7.0	NaN

225 rows × 2 columns

Tabla N°11: Descripción de la variable Wind Dir y Hi Dir

Fuente: *Elaboración Propia*

Una vez cambiando los datos categóricos a numéricos, realizaremos la imputación de las variables NaN, para que no haya imprecisión al realizar los modelos. Para los valores categóricos usaremos la estrategia del más frecuente o moda:

```
from sklearn.impute import SimpleImputer
import numpy as np
```

```
imp= SimpleImputer(missing_values=np.nan,strategy='most_frequent')
```

```
imp=imp.fit(data[['Wind Dir']])

```

```
data[['Wind Dir']] = imp.transform(data[['Wind Dir']])

```

```
imp= SimpleImputer(missing_values=np.nan,strategy='most_frequent')
imp=imp.fit(data[['Hi Dir']])
data[['Hi Dir']] = imp.transform(data[['Hi Dir']])

```

Figura N°37: Imputación de las variables Wind Dir y Hi Dir

Fuente: *Elaboración Propia*

Para comprobar los resultados que no haya ninguna variable con valores NaN, en estas dos categorías, para ello se validará con el siguiente código:

```
data[['Wind Dir','Hi Dir']].isnull().sum()
```

```
Wind Dir    0
Hi Dir      0
dtype: int64
```

Figura N°38: Comprobación de vacíos en variables imputadas

Fuente: *Elaboración Propia*

Para los valores numéricos se maneja los datos faltantes con la media, de la siguiente manera:

```
impNum= SimpleImputer(missing_values=np.nan,strategy='mean')
impNum=impNum.fit(data[['avg Temp Out','avg Hi Temp','avg Low Temp','a
```

```
data[['avg Temp Out','avg Hi Temp','avg Low Temp','avg Out Hum','avg D
```

Figura N°39: Imputación de variables numéricas

Fuente: *Elaboración Propia*

Validamos que no se tiene variables con valores NaN, en las variables numéricas imputadas:

```
data.select_dtypes(include= np.number).isnull().sum().sort_values(ascending = False)
```

```
LT_Total                0
avg In Temp             0
avg Rain Rate           0
avg Solar Rad.          0
avg Solar Energy        0
avg Hi Solar Rad.       0
avg UV Index            0
avg UV Dose             0
avg Hi  UV              0
avg Heat D-D            0
avg Cool D-D            0
avg In Hum              0
avg Bar                 0
avg In Dew              0
avg In Heat             0
avg In EMC              0
avg In Air Density      0
avg ET                  0
```

Figura N°40: Comprobación de nulos en las variables numéricas

Fuente: *Elaboración Propia*

Variables de Datos de Cultivo

Debido a que cada cultivo presenta distintos requerimientos para su cosecha, fue necesario diferenciarlos mediante una variable categórica, donde el tipo de cultivo que se desea hallar, tanto la producción y el personal requerido por campaña, dividirá la base de datos en 3 datasets, con el siguiente código:

```
Palto=data[data['DESCRIPCION']=='PALTO']
Arandano=data[data['DESCRIPCION']=='ARANDANO']
Mandarina=data[data['DESCRIPCION']=='MANDARINA']
```

Figura N°41: Creación de dataset

Fuente: *Elaboración Propia*

Se eliminó las columnas que son innecesarias para realizar el modelamiento de las variables:

```
Base_Palto=Palto.drop(['PERIODO','DESCRIPCION','AÑO','TotalBruto','Pesotara','avg Wind Tx','avg Arc. Int.'],axis=1)
Base_Mandarina=Mandarina.drop(['PERIODO','DESCRIPCION','AÑO','TotalBruto','Pesotara','avg Wind Tx','avg Arc. Int.'],axis=1)
Base_Arandano=Arandano.drop(['PERIODO','DESCRIPCION','AÑO','TotalBruto','Pesotara','avg Wind Tx','avg Arc. Int.'],axis=1)
Base_Palto
```

Figura N°42: Eliminación de las variables no requeridas

Fuente: *Elaboración Propia*

Avg Arc Int - variables metereológicas por cada tiempo

Avg Wind Tx - radio frecuencia en tiempo

5.1.1.4 Modelamiento de datos

Con los dataset “Base Palta”, “Base Mandarina” y” Base Arándano”, se procedió a realizar el modelamiento para los pronósticos de la cantidad de producción en kg y la cantidad de trabajadores para la labor de cosecha.

Para estos casos, se tomaron para los dos modelos el 80% de datos para el entrenamiento y el 20 % restante como data de prueba de funcionamiento del modelo.

MODELO 1: Pronóstico de la cantidad de producción por cultivos.

a) Pronóstico de la cantidad de producción de Palto.

Para nuestro primer dataset “Base Palta”, definiremos las variables dependiente e independiente de la siguiente forma

```
Prod_Palto=Base_Palto['TotalNeto']
VarInd_Palto=Base_Palto.drop(['TotalNeto'],axis=1)
```

Figura N°43: Variable independiente de Palto

Fuente: *Elaboración Propia*

De esta forma, se realizará un split del dataset en 80% Train y 20% para probar el modelo:

```
import sklearn
from sklearn.model_selection import train_test_split
x_train_ppalto, x_test_ppalto, y_train_ppalto,y_test_ppalto=train_test_split(VarInd_Palto,Prod_Palto,test_size=0.2,random_st
```

Figura N°44: Split en conjunto de entrenamiento y prueba

Fuente: *Elaboración Propia*

i) Aplicación del algoritmo Regresión lineal por producción de Palto

Para este algoritmo, se usó la función *LinearRegression* de la librería de *sklearn*, de la cual se usó el método fit, que ajusta el modelo entre los puntos de datos dispersos. Asimismo, se le estará dando como parámetro las variables independientes y dependientes del cultivo de palto.

```
from sklearn.linear_model import LinearRegression
LinRegPalto=LinearRegression()
LinRegPalto.fit(x_train_ppalto,y_train_ppalto)
```

Figura N°45: Parámetro de variables de producción palta

Fuente: *Elaboración Propia*

Con nuestra variable “LinRegPalto”, que fue entrenado con el 80% de la información, se pasará a predecir la variable dependiente para nuestro dato de Test, que nos mostrará el siguiente resultado, el cual se procederá a realizar la comparación con las variables

```
1 y_predict_palto=LinRegPalto.predict(x_test_ppalto)
2 y_predict_palto
```

```
array([43157.38644745, 13042.15898209, 13927.92128755, 14428.8580459 ,
       14261.95161973, 13650.69180364, 14022.01727629, 12302.62753606,
       15001.715591  , 28754.48749426])
```

Figura N°46: Entrenamiento de la variable producción palta

Fuente: *Elaboración Propia*

Como el objetivo de una regresión lineal, es encontrar las pendientes de cada variable independiente y el coeficiente o valor de intersección, para la variable producción de palto sería lo siguiente:

In [193]:
```
print('Valor de la pendiente son:')
print(LinRegPalto.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegPalto.intercept_)
```

```
Valor de la pendiente son:
[ 8.95304055e+01  1.79359530e+01 -1.50123599e+02  5.16883505e-01
  2.62678607e+00 -9.15474181e-02 -8.03353956e+01  4.60285981e-01
 -5.06617621e+02  1.68261707e+03 -4.74646342e+03 -1.99471820e+03
  6.00702947e+03 -9.80241569e+03 -4.64601101e+02 -1.18143497e+03
  9.50160315e+03 -2.34389987e+02 -1.27796075e+04  1.03428920e+04
 -9.84941142e+02 -2.22748480e+02 -7.10764342e+01 -6.26724005e+03
  2.04515498e+03 -3.92765007e+01 -1.94892904e+03 -4.04684413e+03
 -1.93319901e+02 -8.67330629e+03 -4.17298824e+01 -3.72031439e+01
 -1.19357582e+04  5.36497685e+03 -1.99402994e+04  2.86815735e+04
 -1.67009462e+03  2.89132481e+01 -4.40789875e+02  2.88177495e+03
 -4.36936235e+03]
Valor de la interseccion o coeficiente "b" es:
81928.15746261823
```

Figura N°47: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Con respecto al coeficiente de determinación múltiple y el error cuadrático medio, nos indica lo siguiente.

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_ppalto,y_predict_palto))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_ppalto,y_predict_palto))
print("RMSE es:",sqrt(mean_squared_error(y_test_ppalto,y_predict_palto)))

```

```
Coefienciente de determinacion
R^2 es: 0.8496017227738417
Error cuadratico medio
MSE es: 15795735.85100837
RMSE es: 3974.384965124588
```

Figura N°48: R^2 y RMSE de Producción Palto

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) por producción de Palto

Para este algoritmo, se usó la función SVR de la librería de sklearn, de la cual se usó el método fit, que permite ajustar el modelo entre los datos dispersos. Asimismo, se está creando arrays para el parámetro de Kernel, los cuales se probará “linear”, “rbf”, “poly”, de la misma manera para los parámetros.

```
K_array=['linear','rbf','poly']
Prod_cols = ['Kernel','C','R^2','RMSE']
resultados_ProdPalto_array=[]
```

Figura N°49: Uso de array para Kernel Producción Palta

Fuente: *Elaboración Propia*

Para este algoritmo, se realizará bucle con la función “for” declarando una variable “k”,”C_”,”GM”, que tendrá como entrada el array creado anteriormente, de esta forma obtendremos los resultados por cada Kernel, C, gamma y coeficiente de determinación.

```
for K in K_array:
            SVRPalto=SVR(kernel=K, C=1)
            SVRPalto.fit(x_train_ppalto,y_train_ppalto)
            y_SVRLin_predict_palto=SVRPalto.predict(x_test_ppalto)
            corr_pp=r2_score(y_test_ppalto,y_SVRLin_predict_palto)
            rmse_pp=sqrt(mean_squared_error(y_test_ppalto,y_SVRLin_predict_palto))
            resultados_ProdPalto_array.append([K,1,corr_pp,rmse_pp])
resultados_ProdPalto_df = pd.DataFrame(resultados_ProdPalto_array,columns=Prod_cols)
resultados_ProdPalto_df =resultados_ProdPalto_df.sort_values(by='R^2',ascending=False)
resultados_ProdPalto_df
```

Figura N°50: Modelo de SVR Producción Palta

Fuente: *Elaboración Propia*

Para visualizar los resultados que obtenemos de cada Kernel, lo almacenaremos en un dataframe, el cual nos permitirá ordenarlo de manera descendente por el coeficiente de determinación. Como resultado para la producción del palto obtenemos lo siguiente:

	Kernel	C	R^2	RMSE
0	linear	1	0.997231	539.260968
1	rbf	1	-0.055373	10528.137634
2	poly	1	-0.056225	10532.382567

Tabla N°12: Coeficiente de determinación Producción Palta

Fuente: *Elaboración Propia*

iii) Aplicación del algoritmo Árboles de Decisión: Regresión por producción de Palto

Para este algoritmo determinaremos mediante un array, valores aleatorios para determinar el "Max Depth", asimismo almacenaremos en un dataframe los resultados por cada "r2_score"

```
resultados_DT_palto_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
```

Figura N°51: Uso de array para Max Depth Producción Palta

Fuente: *Elaboración Propia*

Para este algoritmo definiremos con un bucle "for" se usará el array declarado anteriormente, de esta forma la función "DecisionTreeRegressor" tendra los parámetros random state de valor 0 por defecto, lo cual nos permitirá que el r2_score no salga de manera aleatoria por cada ejecución, el parámetro splitter se usará la estrategia"best", para obtener la mejor división en cada nodo.

```
for K in Max_Depth_array:
            dtP_Palto=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtP_Palto.fit(x_train_ppalto,y_train_ppalto)
            y_DT_pred_palto=dtP_Palto.predict(x_test_ppalto)
            corr_DT_P=r2_score(y_test_ppalto,y_DT_pred_palto)
            rmse_DT_P=sqrt(mean_squared_error(y_test_ppalto,y_DT_pred_palto))
            resultados_DT_palto_array.append([K,corr_DT_P,rmse_DT_P])
resultados_DT_palto_df = pd.DataFrame(resultados_DT_palto_array,columns=DT_cols)
resultados_DT_palto_df =resultados_DT_palto_df.sort_values(by='R^2',ascending=False)
resultados_DT_palto_df
```

Figura N°52: Modelo de Árbol de regresión Producción Palta

Fuente: *Elaboración Propia*

Por último, creamos un dataframe el cual nos permitirá obtener el valor de “r2_score” ordenado de manera descendente:

	Max Depth	R^2	RMSE
5	10.0	0.838650	4116.550302
1	3.0	0.828640	4242.318418
4	8.0	0.780314	4803.415460
0	NaN	0.780213	4804.510010
6	12.0	0.780213	4804.510010
7	15.0	0.780213	4804.510010
8	25.0	0.780213	4804.510010
9	30.0	0.780213	4804.510010
10	45.0	0.780213	4804.510010
11	50.0	0.780213	4804.510010
12	100.0	0.780213	4804.510010
13	1000.0	0.780213	4804.510010
3	6.0	0.779724	4809.853490
2	4.0	0.743036	5194.994150

Figura N°53: R^2 y RMSE de Producción Palto

Fuente: *Elaboración Propia*

b) Pronóstico de la cantidad de producción de Mandarina.

Para el dataset de Mandarina, se definieron las variables independientes y dependiente.

```
Prod_Mandarina=Base_Mandarina['TotalNeto']
VarInd_Mandarina=Base_Mandarina.drop(['TotalNeto'],axis=1)
print(Prod_Mandarina.head())
print(VarInd_Mandarina.head())
```

Figura N°54: Variable independiente de Mandarina

Fuente: *Elaboración Propia*

De esta forma, se realizará un split del dataset en 80% Train y 20% para probar el modelo.

```
x_train_pmand, x_test_pmand, y_train_pmand,y_test_pmand=train_test_split(Var
```

Figura N°55: Split en conjunto de entrenamiento y prueba

Fuente: *Elaboración Propia*

i) Aplicación del algoritmo de Regresión Lineal de producción de Mandarina

En base a la función *LinearRegression*, se obtendrá la variable Y predecida, la que nos permitirá evaluarlo con la variable de Y Test

```
LinRegMand=LinearRegression()
LinRegMand.fit(x_train_pmand,y_train_pmand)
y_predict_mand=LinRegMand.predict(x_test_pmand)
y_predict_palto
```

```
array([43157.38644745, 13042.15898209, 13927.92128755, 14428.8580459 ,
       14261.95161973, 13650.69180364, 14022.01727629, 12302.62753606,
       15001.715591  , 28754.48749426])
```

Figura N°56: Entrenamiento de la variable producción Mandarina

Fuente: *Elaboración Propia*

Por otro lado, para la función lineal múltiple obtendremos las pendientes de nuestro modelo con la funcion “coef_” y el intercepto con “intercept_”

```
print('Valor de la pendiente son:')
print(LinRegMand.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegMand.intercept_)
```

```
Valor de la pendiente son:
[ 4.06372222e+04  1.87852395e+01 -1.32484497e+03  2.97001375e-01
 -3.82828613e+02 -1.79048692e+00  1.87712887e+04  1.00929117e+01
 -7.18057699e+06  3.78299230e+06 -1.78806628e+06 -4.14585581e+04
  2.49087462e+05 -3.23210877e+05 -2.70723498e+03  3.48268491e+06
  3.72713377e+04 -1.82222605e+02  3.00061692e+06  3.42704777e+06
 -3.22281315e+06  3.28608735e+05 -3.64862095e+05  3.78758936e+03
 -6.81225352e+05  9.65313846e+07 -8.82102028e+03 -4.09476968e+06
  9.97998424e+07  5.79341165e+05 -5.07888923e+08  5.01061682e+08
  4.91805682e+05  4.45405737e+05 -1.72478653e+06 -1.79872778e+05
 -4.95012555e+05 -2.67930557e+08 -6.15282669e+07  1.63905042e+04
 -2.25299985e+04]
Valor de la interseccion o coeficiente "b" es:
14905642.071692357
```

Figura N°57: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Para el cálculo de coeficiente de determinación usaremos la función “r2_score”, la cual proviene de la librería sklearn, tendrá como parámetros la variable de test de producción de mandarina y la variable de predicción de la producción de mandarina

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_pmand,y_predict_mand))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_pmand,y_predict_mand))
print("RMSE es:",sqrt(mean_squared_error(y_test_pmand,y_predict_mand)))
```

```
Coefienciente de determinacion
R^2 es: 0.7251029713954724
Error cuadratico medio
MSE es: 512348644.2233295
RMSE es: 22635.119708614962
```

Figura N°58: R^2 y RMSE de Producción Mandarina

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) de producción de Mandarina

Para este algoritmo, se usó la función SVR de la librería de sklearn, de la cual se usó el metodo fit, que permite ajustar el modelo entre los datos dispersos. Asimismo, se está creando arrays para el parametro de Kernel, los cuales se probará “linear”, “rbf”, “poly”.

```
K_array=['linear','rbf','poly']
resultados_ProdMand_array=[]
Prod_cols = ['Kernel','C','R^2','RMSE']
```

Figura N°59: Uso de array para Kernel Producción Mandarina

Fuente: *Elaboración Propia*

Para este algoritmo, se realizará bucle con la función “for” declarando una variable “k”, que tendrá como entrada el array creado anteriormente,

```
for K in K_array:
            SVRMand=SVR(kernel=K, C=1)
            SVRMand.fit(x_train_pmand,y_train_pmand)
            y_SVR_predict_Mand=SVRMand.predict(x_test_pmand)
            corr_mand=r2_score(y_test_pmand,y_SVR_predict_Mand)
            rmse_pm=sqrt(mean_squared_error(y_test_pmand,y_SVR_predict_Mand))
            resultados_ProdMand_array.append([K,1,corr_mand,rmse_pm])
resultados_ProdMand_df = pd.DataFrame(resultados_ProdMand_array,columns=Prod_cols)
resultados_ProdMand_df =resultados_ProdMand_df.sort_values(by='R^2',ascending=False)
resultados_ProdMand_df
```

Figura N°60: Modelo de SVR Mandarina

Fuente: *Elaboración Propia*

Visualizamos el resultado por cada Kernel, el cual mediante un dataframe nos permitirá ordenarlo de manera descendente por el coeficiente de determinación. Como resultado para la producción del palto obtenemos lo siguiente:

	Kernel	C	R^2	RMSE
0	linear	1	0.998100	1881.900295
2	poly	1	-0.060706	44462.647407
1	rbf	1	-0.064172	44535.219303

Tabla N°13: Coeficiente de determinación y RMSE Mandarina

Fuente: Elaboración Propia

iii) Aplicación del algoritmo Árboles de Decisión: Regresión de producción de Mandarina

Para este algoritmo determinaremos mediante un array, valores aleatorios para determinar el “Max Depth”

```
resultados_DT_Mand_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
```

Figura N°61: Uso de array para Max Depth Producción de Mandarina

Fuente: *Elaboración Propia*

la función “DecisionTreeRegressor” tendra los parámetros random state de valor 0 por defecto, lo cual nos permitirá que el r2_score no salga de manera aleatoria por cada ejecución, el parámetro splitter se usará la estrategia“best”, para obtener la mejor división en cada nodo.

```
for K in Max_Depth_array:
            dtP_Mand=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtP_Mand.fit(x_train_pmand,y_train_pmand)
            y_DT_pred_Mand=dtP_Mand.predict(x_test_pmand)
            corr_DT_M=r2_score(y_test_pmand,y_DT_pred_Mand)
            rmse_DT_M=sqrt(mean_squared_error(y_test_pmand,y_DT_pred_Mand))
            resultados_DT_Mand_array.append([K,corr_DT_M,rmse_DT_M])
resultados_DT_Mand_df = pd.DataFrame(resultados_DT_Mand_array,columns=DT_cols)
resultados_DT_Mand_df =resultados_DT_Mand_df.sort_values(by='R^2',ascending=False)
resultados_DT_Mand_df
```

Figura N°62: Modelo de Árbol de Regresión Mandarina

Fuente: Elaboración Propia

Determinaremos el coeficiente de determinación, para determinar la mejor opción del modelo.

	Max Depth	R^2	RMSE
5	10.0	0.838650	4116.550302
1	3.0	0.828640	4242.318418
4	8.0	0.780314	4803.415460
0	NaN	0.780213	4804.510010
6	12.0	0.780213	4804.510010
7	15.0	0.780213	4804.510010
8	25.0	0.780213	4804.510010
9	30.0	0.780213	4804.510010
10	45.0	0.780213	4804.510010
11	50.0	0.780213	4804.510010
12	100.0	0.780213	4804.510010
13	1000.0	0.780213	4804.510010
3	6.0	0.779724	4809.853490
2	4.0	0.743036	5194.994150

Figura N°63: Coeficiente de determinación y RMSE Mandarina

***Fuente:** Elaboración Propia*

c) Pronóstico de la cantidad de producción de arándano

i) Aplicación del algoritmo Regresión lineal por producción de arándano

Una vez declarado las variables dependientes e independientes, se procede a realizar los conjuntos de datos de prueba y test, dado que esto permitirá evaluar al modelo si está realizando una correcta predicción de la variable producción. Por otro lado, para calcular la fórmula de regresión lineal múltiple de la cantidad de producción de arándano, se determina mediante la pendiente y el coeficiente de intersección, para ello se realiza lo siguiente:

```
In [205]:
print('Valor de la pendiente son:')
print(LinRegArand.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegArand.intercept_)
```

```
Valor de la pendiente son:
[ 1.14812689e+02  3.09650730e+00  6.88767214e+02 -8.60801311e-02
  1.04475871e+01 -1.43885403e-01 -4.91144417e+02  8.97440909e-01
 -5.87604694e+04 -1.69052152e+04 -9.11800019e+04  2.30240677e+02
  2.75384339e+02 -1.86035442e+04  2.57441411e+01  2.13723016e+05
  5.75452434e+02 -1.96346113e+01  1.63444651e+05  1.54667529e+05
 -1.61213628e+05  8.67972420e+02 -9.42377483e-08 -4.00882012e+04
 -4.48250116e+03  6.29038121e+05 -6.27671232e+00  9.78319547e+04
 -2.78802072e+06  4.22833214e+03 -2.86758083e+06  3.49110207e+06
 -2.16827092e+04 -5.57823397e+03  1.40872422e+04  3.92747402e+03
  7.45487387e+03 -7.10823578e+05 -8.13024641e+05  4.72344372e+02
  0.00000000e+00 -7.12715013e+02  0.00000000e+00]
Valor de la interseccion o coeficiente "b" es:
580659.6740105003
```

Figura N°64: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Para determinar si el modelo realiza una correcta predicción de la variable producción de arándanos, se usará las funciones “r2_score” y “mean_squeared_error” que tendrá como parámetros las variables de producción proyectada mediante el modelo y la producción real.

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_parand,y_predict_Arand))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_parand,y_predict_Arand))
print("RMSE es:",sqrt(mean_squared_error(y_test_parand,y_predict_Arand)))
```

```
Coefienciente de determinacion
R^2 es: 0.9446274863737195
Error cuadratico medio
MSE es: 363154.5453539861
RMSE es: 602.6230541175687
```

Figura N°65: R^2 y RMSE de Producción Arándano

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) producción de arándano

Para este algoritmo, se está creando arrays para el parámetro de Kernel, los cuales se probará “linear”, “rbf”, “poly”.

```
K_array=['linear','rbf','poly']
resultados_ProdArand_array=[]
Prod_cols = ['Kernel','C','R^2','RMSE']
for K in K_array:
```

Figura N°66: Uso de array para Kernel Arándano

Fuente: *Elaboración Propia*

Para este algoritmo, se realizará bucle con la función “for” declarando una variable “k”, que tendrá como entrada el array creado anteriormente

```
for K in K_array:
            SVRArand=SVR(kernel=K, C=1)
            SVRArand.fit(x_train_parand,y_train_parand)
            y_SVR_predict_Arand=SVRArand.predict(x_test_parand)
            corr_arand=r2_score(y_test_parand,y_SVR_predict_Arand)
            rmse_pa=sqrt(mean_squared_error(y_test_parand,y_SVR_predict_Arand))
            resultados_ProdArand_array.append([K,1,corr_arand,rmse_pa])
resultados_ProdArand_df = pd.DataFrame(resultados_ProdArand_array,columns=Prod_cols)
resultados_ProdArand_df =resultados_ProdArand_df.sort_values(by='R^2',ascending=False)
resultados_ProdArand_df
```

Figura N°67: Modelo de SVR Arándano

Fuente: *Elaboración Propia*

El coeficiente de determinación, ordenando de manera descendente por cada Kernel, obtenemos lo siguiente:

	Kernel	C	R^2	RMSE
0	linear	1	0.837764	1031.508489
1	rbf	1	-0.284704	2902.686242
2	poly	1	-0.285150	2903.190716

Tabla N°14: Coeficiente de determinación y RMSE producción de Arándano

Fuente: Elaboración Propia

iii) Aplicación del algoritmo Árboles de Decisión: Regresión producción de arándano

Para este algoritmo determinaremos mediante un array, valores aleatorios para determinar el "Max Depth". Por otro lado, la función "DecisionTreeRegressor" tendra los parámetros random state de valor 0 por defecto, lo cual nos permitirá que el r2_score no salga de manera aleatoria por cada ejecución, el parámetro splitter se usará la estrategia"best", para obtener la mejor división en cada nodo.

```
resultados_DT_Arand_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
for K in Max_Depth_array:
            dtP_Arand=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtP_Arand.fit(x_train_parand,y_train_parand)
            y_DT_pred_Arand=dtP_Arand.predict(x_test_parand)
            corr_DT_A=r2_score(y_test_parand,y_DT_pred_Arand)
            rmse_DT_A=sqrt(mean_squared_error(y_test_parand,y_DT_pred_Arand))
            resultados_DT_Arand_array.append([K,corr_DT_A,rmse_DT_A])
resultados_DT_Arand_df = pd.DataFrame(resultados_DT_Arand_array,columns=DT_cols)
resultados_DT_Arand_df =resultados_DT_Arand_df.sort_values(by='R^2',ascending=False)
resultados_DT_Arand_df
```

Figura N°68: Modelo de Árbol de Regresión de Producción Arándano

Fuente: Elaboración Propia

Ordenaremos el coeficiente de determinación de manera descendente, por cada valor ingresado en el parámetro "Max_Depth"

	Max Depth	R^2	RMSE
2	4.0	0.964826	480.296790
1	3.0	0.941149	621.262791
3	6.0	0.929827	678.395374
4	8.0	0.929670	679.153012
0	NaN	0.927070	691.596142
6	12.0	0.927070	691.596142
7	15.0	0.927070	691.596142
8	25.0	0.927070	691.596142
9	30.0	0.927070	691.596142
10	45.0	0.927070	691.596142
11	50.0	0.927070	691.596142
12	100.0	0.927070	691.596142
13	1000.0	0.927070	691.596142
5	10.0	0.926010	696.600885

Figura N°69: Coeficiente de determinación y RMSE de producción de Arándano

Fuente: *Elaboración Propia*

MODELO 2: Pronóstico de la cantidad de trabajadores por cosecha.

Para calcular la predicción de los trabajadores para la labor de cosecha por cultivo, se determinará la dependencia entre la variable cantidad de trabajadores y todas las variables independientes del dataset, para ello se usará el coeficiente de correlación de pearson.

a) Pronóstico de la cantidad de trabajadores por cosecha de Palto

Para determinar el coeficiente de correlación de pearson usaremos la siguiente función y determinar qué variables tienen mayor relación con Cantidad de trabajadores

```
1 corr_Palto=Base_Palto.corr(method="pearson")
2 corr_Palto['Cantidad_Trabajadores'].sort_values(ascending = False)
3
```

Figura N°70: Cálculo del coeficiente de correlación de Pearson

Fuente: *Elaboración Propia*

los cual se escogen las variables “HorasTotales”,”CostoTotal”,”HA_Totales” y “TotalNeto", este último a pesar de tener una correlación baja, es una variable importante.

Cantidad_Trabajadores	1.000000
HorasTotales	0.675751
CostoTotal	0.604015
HA_TOTALES	0.379864
avg Solar Energy	0.312362
avg Solar Rad.	0.312302
avg Hi Solar Rad.	0.311645
LT_Total	0.308786
PlantasTotales	0.297581
avg UV Dose	0.284783
avg ET	0.284724
avg UV Index	0.284501
avg Hi UV	0.283580
avg Wind Samp	0.265389
avg ISS Recept	0.244042
avg Bar	0.190470
Edad_Prom	0.149202
avg Rain Rate	0.123893
avg Rain	0.123893

Tabla N°15: Coeficiente de correlación de Pearson

Fuente: Elaboración Propia

Para nuestros dataset de plata, se define la variable dependiente como la cantidad de trabajadores e independiente en base a la selección que se realizó en el paso previo.

```
Q_Trabj_Palto=Base_Palto['Cantidad_Trabajadores']
Ind_QT_Palto=Base_Palto[['HorasTotales','CostoTotal','HA_TOTALES','TotalNeto']]
```

Figura N°71: Variable independiente de Palto

Fuente: *Elaboración Propia*

En base a los dataframes, emplearemos el método de 80 y 20 para separar la data en prueba y entrenamiento

```
x_train_Q_palto, x_test_Q_palto, y_train_Q_palto,y_test_Q_palto,=train_test_split(Ind_QT_Palto,Q_Trabj_Palto,test_size=0.2,
```

```
LinRegQPalto=LinearRegression()
LinRegQPalto.fit(x_train_Q_palto,y_train_Q_palto)
```

Figura N°72: Split en conjunto de entrenamiento y prueba

Fuente: *Elaboración Propia*

i) Aplicación del algoritmo Regresión lineal por cosecha de Palto

Para la función lineal múltiple, tendríamos las siguientes pendientes y coeficiente de intersección

```
print('Valor de la pendiente son:')
print(LinRegQPalto.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegQPalto.intercept_)
```

```
Valor de la pendiente son:
[ 4.70656453e-02 -2.73896464e-03  9.13078538e-01 -4.32688003e-04]
Valor de la interseccion o coeficiente "b" es:
25.781634834170674
```

Figura N°73: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Con la variable LinRegQpalto, se determina el coeficiente de determinación y el error cuadrático medio

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_Q_palto,y_predict_QPalto))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_Q_palto,y_predict_QPalto))
print("RMSE es:",sqrt(mean_squared_error(y_test_Q_palto,y_predict_QPalto)))
```

```
Coefienciente de determinacion
R^2 es: 0.2268436261779595
Error cuadratico medio
MSE es: 241.53405118200544
RMSE es: 15.541365808126564
```

Figura N°74: R2 y RMSE de Trabajadores por Cosecha de Palto

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) por cosecha de Palto

Para el modelado de trabajadores para el cultivo de palto, se hará uso solo del Kernel “Linear”, debido al costo computacional que se emplea al usar los parámetros de C y Gamma, de este mismo modo, se realizará 3 array para el Kernel, C y Gamma.

```
K_array=['linear']
resultados_QPalto_array=[]
QT_cols = ['Kernel','C','Gamma','R^2','RMSE']
C_array=[0.001,0.01,0.1]
G_array=[0.001,0.01,0.1]
```

Figura N°75: Uso de array para Kernel Trabajadores Cultivo de Palta

Fuente: *Elaboración Propia*

Se hace uso de la función SVR, en un bucle "for" para las variables "K", "C_", "G", lo que nos permitirá obtener 9 combinaciones entre C y gamma. Las cuales se almacenarán en un dataframe ordenados de manera descendente por R2 score.

```
for K in K_array:
    for C_ in C_array:
        for G in G_array:
            SVRQPalto=SVR(kernel=K, C=C_,gamma=G)
            SVRQPalto.fit(x_train_Q_palto,y_train_Q_palto)
            y_SVRLin_predict_Qpalto=SVRQPalto.predict(x_test_Q_palto)
            corr_Qp=r2_score(y_test_Q_palto,y_SVRLin_predict_Qpalto)
            rmse_qp=sqrt(mean_squared_error(y_test_Q_palto,y_SVRLin_predict_Qpalto))
            resultados_QPalto_array.append([K,C_,G,corr_Qp,rmse_qp])
resultados_QPalto_df = pd.DataFrame(resultados_QPalto_array,columns=QT_cols)
resultados_QPalto_df =resultados_QPalto_df.sort_values(by='R^2',ascending=False)
resultados_QPalto_df
```

Figura N°76: Modelo de SVR Trabajadores Cultivo de Palta

Fuente: *Elaboración Propia*

El siguiente dataframe, nos muestra el coeficiente de determinación por el Kernel "linear".

	Kernel	C	Gamma	R^2	RMSE
6	linear	0.100	0.001	0.515705	12.300150
7	linear	0.100	0.010	0.515705	12.300150
8	linear	0.100	0.100	0.515705	12.300150
3	linear	0.010	0.001	0.230001	15.509597
4	linear	0.010	0.010	0.230001	15.509597
5	linear	0.010	0.100	0.230001	15.509597
0	linear	0.001	0.001	0.212403	15.685829
1	linear	0.001	0.010	0.212403	15.685829
2	linear	0.001	0.100	0.212403	15.685829

Tabla N°16: Coeficiente de determinación Trabajadores Cultivo de Palta

Fuente: Elaboración Propia

iii) Aplicación del algoritmo Árboles de Decisión: Regresión por cosecha de Palto

Se empleará la función "DecisionTreeRegressor", la cual usaremos los parámetros random state de valor 0 por defecto, lo cual nos permitirá que el r2_score no salga de manera aleatoria por cada ejecución, el parámetro splitter se usará la estrategia"best", para obtener la mejor división en cada nodo.

```
resultados_DT_Qpalto_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
for K in Max_Depth_array:
            dtQ_Palto=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtQ_Palto.fit(x_train_Q_palto,y_train_Q_palto)
            y_DT_pred_Qpalto=dtQ_Palto.predict(x_test_Q_palto)
            corr_DT_QP=r2_score(y_test_Q_palto,y_DT_pred_Qpalto)
            rmse_DT_QP=sqrt(mean_squared_error(y_test_Q_palto,y_DT_pred_Qpalto))
            resultados_DT_Qpalto_array.append([K,corr_DT_QP,rmse_DT_QP])
resultados_DT_Qpalto_df = pd.DataFrame(resultados_DT_Qpalto_array,columns=DT_cols)
resultados_DT_Qpalto_df =resultados_DT_Qpalto_df.sort_values(by='R^2',ascending=False)
resultados_DT_Qpalto_df
```

Figura N°77: Modelo de Árbol de Regresión Trabajadores Cultivo de Palta

Fuente: Elaboración Propia

En base al resultado obtenido por cada "Max_depth", se ordenará el r2_score:

	Max Depth	R^2	RMSE
1	3.0	0.557698	11.754794
2	4.0	-0.254058	19.793122
3	6.0	-0.315123	20.269298
0	NaN	-0.376761	20.738852
4	8.0	-0.376761	20.738852
5	10.0	-0.376761	20.738852
6	12.0	-0.376761	20.738852
7	15.0	-0.376761	20.738852
8	25.0	-0.376761	20.738852
9	30.0	-0.376761	20.738852
10	45.0	-0.376761	20.738852
11	50.0	-0.376761	20.738852
12	100.0	-0.376761	20.738852
13	1000.0	-0.376761	20.738852

Tabla N°17: Coeficiente de determinación y RMSE Trabajadores de Cultivo de Palta

Fuente: Elaboración Propia

b) Pronóstico de la cantidad de trabajadores por cosecha de Mandarina

Se usará el coeficiente de correlación de pearson, para determinar las variables independientes que tengan mayor relación con la cantidad de trabajadores para la labor de cosecha del cultivo de mandarina.

```
corr_mand=Base_Mandarina.corr(method="pearson")
corr_mand['Cantidad_Trabajadores'].sort_values(ascending = False)
```

Figura N°78: Cálculo del coeficiente de correlación de Pearson

Fuente: Elaboración Propia

i) Aplicación del algoritmo Regresión lineal

Las variables por usar serán las que tengan una correlación dentro del rango de 0.4 a 1 y -0.4 a -1.

Cantidad_Trabajadores	1.000000
HorasTotales	0.861118
CostoTotal	0.826280
avg Out Hum	0.585547
avg Heat D-D	0.477727
HA_TOTALES	0.469130
avg In Hum	0.430901
avg In EMC	0.429860
LT_Total	0.425668
avg In Air Density	0.364267
Edad_Prom	0.299271
Nrojabas	0.254118
TotalNeto	0.227122

Tabla N°18: Coeficiente de correlación de Pearson

Fuente: Elaboración Propia

Determinaremos la pendiente y coeficiente de intersección

```
print('Valor de la pendiente son:')
print(LinRegQMand.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegQMand.intercept_)
```

```
Valor de la pendiente son:
[ 2.02932230e-02 -1.73392081e-04 -1.13946144e+01  1.86948233e+02
  6.59629018e-01  3.63669641e-01  8.35631910e+00  9.45911291e+00
 -1.50987214e-04 -3.51453712e+03  3.71451454e+03  3.35810103e+03
 -3.45459571e+02 -2.44694959e+03 -7.67697228e+02  1.37163310e+03
 -1.14410027e+02 -5.46323411e+04  3.99863086e+02 -5.51020641e+00
 -4.83813781e+04 -5.53715844e+00  2.19795135e+02 -2.37237753e+04
  5.82876140e+02]
Valor de la interseccion o coeficiente "b" es:
941.0230056210306
```

Figura N°79: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Se calcula el coeficiente de determinación y el error cuadrático medio.

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_Q_Mand,y_predict_QMand))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_Q_Mand,y_predict_QMand))
print("RMSE es:",sqrt(mean_squared_error(y_test_Q_Mand,y_predict_QMand)))
```

```
Coefienciente de determinacion
R^2 es: 0.6290499449307232
Error cuadratico medio
MSE es: 101.66491891916498
RMSE es: 10.082902306338438
```

Figura N°80: R^2 y RMSE de Producción Mandarina

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) por cosecha de Mandarina

Para el modelado de trabajadores para el cultivo de mandarina, se hará uso solo del Kernel “Linear”, debido al costo computacional que se emplea al usar los parámetros de C y Gamma, de este mismo modo, se realizará 3 array para el Kernel, C y Gamma.

```
K_array=['linear']
resultados_QTMand_array=[]
QT_cols = ['Kernel','C','Gamma','R^2','RMSE']
C_array=[0.001,0.01,0.1]
G_array=[0.001,0.01,0.1]
```

Figura N°81: Uso de array para Kernel Trabajadores Cultivo de Mandarina

Fuente: *Elaboración Propia*

Del mismo modo que para la mano de obra de la cosecha de palto, se estará usando bucles "for", para obtener los 9 resultados del modelo.

```
for K in K_array:
    for C_ in C_array:
        for G in G_array:
            SVRQMand=SVR(kernel=K, C=C_,gamma=G)
            SVRQMand.fit(x_train_Q_Mand,y_train_Q_Mand)
            y_SVR_predict_QMand=SVRQMand.predict(x_test_Q_Mand)
            corr_Qm=r2_score(y_test_Q_Mand,y_SVR_predict_QMand)
            rmse_qm=sqrt(mean_squared_error(y_test_Q_Mand,y_SVR_predict_QMand))
            resultados_QTMand_array.append([K,C_,G,corr_Qm,rmse_qm])
resultados_QTMand_df = pd.DataFrame(resultados_QTMand_array,columns=QT_cols)
resultados_QTMand_df =resultados_QTMand_df.sort_values(by='R^2',ascending=False)
resultados_QTMand_df
```

Figura N°82: Modelo de SVR Trabajadores Cultivo de Mandarina

Fuente: *Elaboración Propia*

Los resultados obtenidos están ordenados de manera descendente, para escoger el mejor resultado del modelo de cantidad de trabajadores para el cultivo de mandarina.

	Kernel	C	Gamma	R^2	RMSE
3	linear	0.010	0.001	0.821648	6.991450
4	linear	0.010	0.010	0.821648	6.991450
5	linear	0.010	0.100	0.821648	6.991450
0	linear	0.001	0.001	0.789545	7.594646
1	linear	0.001	0.010	0.789545	7.594646
2	linear	0.001	0.100	0.789545	7.594646
6	linear	0.100	0.001	-4.608225	39.204920
7	linear	0.100	0.010	-4.608225	39.204920
8	linear	0.100	0.100	-4.608225	39.204920

Tabla N°19: Coeficiente de determinación Trabajadores Cultivo de Mandarina

Fuente: Elaboración Propia

iii) Aplicación del algoritmo Árboles de Decisión: Regresión por cosecha de Mandarina

Se empleará la función "DecisionTreeRegressor", con los parámetros random state de valor 0 y el parámetro splitter con la estrategia"best", para obtener la mejor división en cada nodo.

```
resultados_DT_QMand_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
for K in Max_Depth_array:
            dtQ_Mand=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtQ_Mand.fit(x_train_Q_Mand,y_train_Q_Mand)
            y_DT_pred_QMand=dtQ_Mand.predict(x_test_Q_Mand)
            corr_DT_QM=r2_score(y_test_Q_Mand,y_DT_pred_QMand)
            rmse_DT_QM=sqrt(mean_squared_error(y_test_Q_Mand,y_DT_pred_QMand))
            resultados_DT_QMand_array.append([K,corr_DT_QM,rmse_DT_QM])
resultados_DT_QMand_df = pd.DataFrame(resultados_DT_QMand_array,columns=DT_cols)
resultados_DT_QMand_df =resultados_DT_QMand_df.sort_values(by='R^2',ascending=False)
resultados_DT_QMand_df
```

Figura N°83: Modelo de Árbol de Regresión Trabajadores Cultivo de Mandarina

Fuente: *Elaboración Propia*

Por cada "Max_depth" declarado en el array, se obtendrá un "r2_score", distinto lo que permitirá decidir cuál es la mejor opción.

	Max Depth	R^2	RMSE
0	NaN	0.765959	8.008924
4	8.0	0.765959	8.008924
5	10.0	0.765959	8.008924
6	12.0	0.765959	8.008924
7	15.0	0.765959	8.008924
8	25.0	0.765959	8.008924
9	30.0	0.765959	8.008924
10	45.0	0.765959	8.008924
11	50.0	0.765959	8.008924
12	100.0	0.765959	8.008924
13	1000.0	0.765959	8.008924
2	4.0	0.762882	8.061392
3	6.0	0.240278	14.429632
1	3.0	-0.158905	17.821810

Tabla N°20: Coeficiente de determinación y RMSE Trabajadores Cultivo de Mandarina

Fuente: Elaboración Propia

c) Pronóstico de la cantidad de trabajadores por cosecha de Arándano

i) Aplicación del algoritmo Regresión lineal por cosecha de Arándano

Una vez determinado el coeficiente de correlación de Pearson, y declarado las variables independientes con mejor correlación, se calcula los valores de la pendiente e intersección, de esta forma tener la fórmula de regresión lineal múltiple.

```
print('Valor de la pendiente son:')
print(LinRegQArand.coef_)
print('Valor de la interseccion o coeficiente "b" es:')
print(LinRegQArand.intercept_)
```

```
Valor de la pendiente son:
[ 1.16156460e+00  6.70910064e-04  6.96938919e+00  4.85640748e-03
 -2.79103528e-02  6.92827317e+03 -1.21532575e+03 -6.03132133e+02
  4.58766891e+00 -9.99379866e+02  1.19132552e+04 -7.00255342e+00
 -5.07055476e+03 -4.95656796e+03  4.91975649e+03 -3.11123177e+01
  4.17838921e+03  6.76983737e-01  3.57508880e+03 -1.00262522e+05
  2.87352662e+00  1.44268129e+03  8.36311900e+00 -3.76096563e+01
  1.56421802e+04]
Valor de la interseccion o coeficiente "b" es:
-482.09056693772715
```

Figura N°84: Pendiente y Coeficiente

Fuente: *Elaboración Propia*

Con las variables de predicción de trabajadores y reales del dataset de arándanos, se calcula el coeficiente de determinación y el error cuadrático medio.

```
from sklearn.metrics import mean_squared_error, r2_score
print("Coefienciente de determinacion")
print("R^2 es:",r2_score(y_test_Q_Arand,y_predict_QArand))

print("Error cuadratico medio")
print("MSE es:",mean_squared_error(y_test_Q_Arand,y_predict_QArand))
print("RMSE es:",sqrt(mean_squared_error(y_test_Q_Arand,y_predict_QArand)))
```

```
Coefienciente de determinacion
R^2 es: 0.8522738308253484
Error cuadratico medio
MSE es: 271.4184504581471
RMSE es: 16.474782258292432
```

Figura N°85: R2 y RMSE de Producción Arándano

Fuente: *Elaboración Propia*

ii) Aplicación del algoritmo Vector Soporte Regresión (SVR) por cosecha de Arándano

Se hace uso solo del Kernel “linear”, por el alto costo computacional que emplea los kernel “rbf” y “poly”, del mismo modo para escoger los mejores parámetros entre C y gamma, se emplea un array de valores numéricos que se usarán como valores de entrada para la función SVR.

```
resultados_QTArand_array=[]
K_array=['linear']
QT_cols = ['Kernel','C','Gamma','R^2','RMSE']
C_array=[0.001,0.01,0.1]
G_array=[0.001,0.01,0.1]
```

Figura N°86: Uso de array para Kernel Trabajadores Cultivo de Arándano

Fuente: *Elaboración Propia*

Para el uso de la función SVR, se emplea el bucle “for”, que nos permitirá agilizar la obtención de los resultados, donde pondremos las variables creadas anteriormente como parámetros de entrada. Por otra parte, emplearemos el método “fit” que ajusta el modelo entre los puntos de datos dispersos y usaremos la función “r2_score”, para obtener el coeficiente de determinación por cada combinación entre C y Gamma.

```
for K in K_array:
    for C_ in C_array:
        for G in G_array:
            SVRQArand=SVR(kernel=K, C=C_,gamma=G)
            SVRQArand.fit(x_train_Q_Arand,y_train_Q_Arand)
            y_SVR_predict_QArand=SVRQArand.predict(x_test_Q_Arand)
            corr_QA=r2_score(y_test_Q_Arand,y_SVR_predict_QArand)
            rmse_qm=sqrt(mean_squared_error(y_test_Q_Arand,y_SVR_predict_QArand))
            resultados_QTArand_array.append([K,C_,G,corr_QA,rmse_qm])
resultados_QTArand_df = pd.DataFrame(resultados_QTArand_array,columns=QT_cols)
resultados_QTArand_df =resultados_QTArand_df.sort_values(by='R^2',ascending=False)
resultados_QTArand_df
```

Figura N°87: Modelo de SVR Trabajadores Cultivo de Arándano

Fuente: *Elaboración Propia*

En base a los resultados obtenidos, lo dejaremos en un dataframe ordenado de manera descendente por el coeficiente de determinación.

	Kernel	C	Gamma	R^2	RMSE
6	linear	0.100	0.001	0.886292	14.453916
7	linear	0.100	0.010	0.886292	14.453916
8	linear	0.100	0.100	0.886292	14.453916
3	linear	0.010	0.001	0.848801	16.667292
4	linear	0.010	0.010	0.848801	16.667292
5	linear	0.010	0.100	0.848801	16.667292
0	linear	0.001	0.001	0.828581	17.746804
1	linear	0.001	0.010	0.828581	17.746804
2	linear	0.001	0.100	0.828581	17.746804

Tabla N°21: Coeficiente de determinación y RMSE Trabajadores Cultivo de Arándano

Fuente: Elaboración Propia

iii) Aplicación del algoritmo Árboles de Decisión: Regresión por cosecha de Arándano

Se empleará la función “DecisionTreeRegressor”, con los parámetros random state de valor 0, para no alterar el resultado del “r2_score” por cada ejecución, asimismo el parámetro splitter con la estrategia“best”, para obtener la mejor división en cada nodo y el parámetro “max_depth” mediante un array**.**

```
resultados_DT_QArand_array=[]
Max_Depth_array=[None,3,4,6,8,10,12,15,25,30,45,50,100,1000]
DT_cols = ['Max Depth','R^2','RMSE']
for K in Max_Depth_array:
            dtQ_Arand=DecisionTreeRegressor(random_state=0,max_depth=K,splitter="best")
            dtQ_Arand.fit(x_train_Q_Arand,y_train_Q_Arand)
            y_DT_pred_QArand=dtQ_Arand.predict(x_test_Q_Arand)
            corr_DT_QA=r2_score(y_test_Q_Arand,y_DT_pred_QArand)
            rmse_DT_QA=sqrt(mean_squared_error(y_test_Q_Arand,y_DT_pred_QArand))
            resultados_DT_QArand_array.append([K,corr_DT_QA,rmse_DT_QA])
resultados_DT_QArand_df = pd.DataFrame(resultados_DT_QArand_array,columns=DT_cols)
resultados_DT_QArand_df =resultados_DT_QArand_df.sort_values(by='R^2',ascending=False)
resultados_DT_QArand_df
```

Figura N°88: Modelo de Árbol de Regresión Trabajadores Cultivo de Arándano

Fuente: *Elaboración Propia*

Acorde al dataframe obtenido del resultado, la columna Max depth se repetirá las veces que se declaró en el array, de esta forma evaluar cada "r2_score" y decidir cuál es la mejor opción.

	Max Depth	R^2	RMSE
4	8.0	0.849173	16.646793
0	NaN	0.848568	16.680146
6	12.0	0.848568	16.680146
7	15.0	0.848568	16.680146
8	25.0	0.848568	16.680146
9	30.0	0.848568	16.680146
10	45.0	0.848568	16.680146
11	50.0	0.848568	16.680146
12	100.0	0.848568	16.680146
13	1000.0	0.848568	16.680146
2	4.0	0.847719	16.726817
3	6.0	0.835917	17.362940
5	10.0	0.832284	17.554105
1	3.0	0.790181	19.634222

Tabla N°22: Coeficiente de determinación y RMSE Trabajadores Cultivo de Arándano

Fuente: Elaboración Propia

5.2 Medición de la solución

Evaluación de resultados

Para medir el error estadístico dentro de la solución propuesta se usaron el R^2 (Coeficiente de Determinación) y el RMSE (Error cuadrático medio), los cuales muestran cómo los puntos se ajustan a la línea evalúan y que tan dispersos son los datos que se usaron en el modelo con respecto a la línea de regresión hallada.

5.2.1 Análisis de Indicadores cuantitativo y/o cualitativo

Al haber aplicado los algoritmos de Regresión Lineal Múltiple, SVR y de Árboles de Regresión, obtuvimos los siguientes resultados para cada caso:

Pronóstico 1: Cantidad producida por cultivos

Palta

Modelo 1: Producción Palta	R^2	RMSE
Regresión Lineal Múltiple	0.84	3’974.384965 kg
SVR	0.99	539.26 kg
Árbol de Regresión	0.84	4’116.55 kg

Tabla N°23: Resumen de métricas estadísticas para Producción en Kg de Palta

Fuente: Elaboración Propia

Mandarina

Modelo 1: Producción Mandarina	R^2	RMSE
Regresión Lineal Múltiple	0.72	22’635.12 kg
SVR	0.99	1’881.90 kg
Árbol de Regresión	0.83	4’116.55 kg

Tabla N°24: Resumen de métricas estadísticas para Producción en Kg de Mandarina

Fuente: Elaboración Propia

Arándano

Modelo 1: Producción Arándano	R^2	RMSE
Regresión Lineal Múltiple	0.94	602.62 kg
SVR	0.84	1031.51 kg
Árbol de Regresión	0.96	480.29 kg

Tabla N°25: Resumen de métricas estadísticas para Producción en Kg de Arándano

Fuente: Elaboración Propia

R^2

Con los resultados obtenidos podemos deducir que los modelos para los tres cultivos son factibles (no superan el 1) de usar para predecir valores futuros, partiendo de la base de que las campañas de cosecha pueden cruzarse entre sí por lo que es necesario que los 3 tengan que ser factibles. Asimismo, el modelo de Árbol de Decisión presenta valores de R^2 realistas con respecto a los modelos SVR y de Regresión Lineal Múltiple para los cultivos de mandarina y arándano, pero no siendo el caso de la palta, en el cual la Regresión Lineal Múltiple presenta un **R^2** de 0.84 pero de menor dispersión (RMSE) respecto al Árbol de Regresión.

RMSE

Podemos afirmar que los resultados del modelo del Árbol de Decisión son los adecuados ya que, si bien no presentan la menor dispersión, como la que muestra los resultados del SVR, no se opta por elegir este modelo dado su alto grado de ajuste a los datos, lo que la convierte en un modelo muy optimista. Sin embargo, la dispersión es menor para el caso de las paltas en el modelo de Regresión Lineal Múltiple, pero se valora los resultados de los cultivos en conjunto.

Pronóstico 2: Cantidad de mano de obra por cultivo

Palta

Modelo 2: Cantidad de Trabajadores Palta	R^2	RMSE
Regresión Lineal Múltiple	0.22	15.54
SVR	0.51	12.30
Árbol de Regresión	0.55	11.75

Tabla N°26: Tabla comparativa R2 y RMSE Palta

Fuente: Elaboración Propia

Mandarina

Modelo 2: Cantidad de Trabajadores Mandarina	R^2	RMSE
Regresión Lineal Múltiple	0.62	10.08
SVR	0.82	6.99
Árbol de Regresión	0.77	8

Tabla N°27: Tabla comparativa R2 y RMSE Mandarina

Fuente: Elaboración Propia

Arándano

Modelo 2: Cantidad de Trabajadores Arándano	R^2	RMSE
Regresión Lineal Múltiple	0.85	16.47
SVR	0.88	14.45
Árbol de Regresión	0.85	16.65

Tabla N°28: Tabla comparativa R2 y RMSE Arándano

Fuente: Elaboración Propia

R^2

Con los resultados obtenidos podemos afirmar que los modelos para los tres cultivos son factibles y entre estos el que mejor indicador presenta es el modelo SVR para el caso las mandarinas y arándanos más no para las paltas, en la cual el R^2 tiene un valor de 0.55 contra el 0.51 del SVR.

RMSE

Podemos afirmar que los resultados del modelo SVR presentan menor dispersión para los para los de mandarinas y arándanos. En el caso del cultivo de palta, los RMSE de los modelos de SVR y Árbol de Regresión presentan una mínima diferencia de un 4.47%, por tanto se opta por emplear el modelo SVR como mejor opción para el pronóstico de la cantidad de personal para cosecha de los 3 cultivos.

5.2.2 Simulación de solución. Aplicación de Software.

Pronóstico 1

Producción de Palta

Para la simulación de la producción del cultivo de palta, usaremos el 20% de las variables de prueba de nuestra variable dependiente con el resultado que nos predice la producción del modelo de regresión lineal con nuestras variables independientes de la prueba, de esta forma podemos observar la diferencia que existe entre los valores y estimados de la producción de palta en los 3 modelos, pudiendo demostrar que las filas de los resultados del modelo SVR tienen una dispersión muy baja a comparación de los otros 2 modelos.

	Prod Palto Real	Prod Palto Pred	Difference
180	42896.0	43157.386447	-261.386447
120	13221.4	13042.158982	179.241018
188	14163.0	13927.921288	235.078712
185	14214.4	14428.858046	-214.458046
191	13940.8	14261.951620	-321.151620
117	13935.6	13650.691804	284.908196
87	14892.0	14022.017276	869.982724
158	609.0	12302.627536	-11693.627536
91	14632.2	15001.715591	-369.515591
90	24289.4	28754.487494	-4465.087494

Tabla N°29: Regresión Lineal Múltiple: Test vs Predicción Producción de Palta

Fuente: Elaboración Propia

	Prod Palta Real	Prod Palta Real Pred	Difference
180	42896.0	43139.464488	-243.464488
120	13221.4	13499.056283	-277.656283
188	14163.0	14396.057203	-233.057203
185	14214.4	14174.502597	39.897403
191	13940.8	14191.895045	-251.095045
117	13935.6	13417.063110	518.536890
87	14892.0	14767.050397	124.949603
158	609.0	-534.174101	74.825899
91	14632.2	15636.643936	-1004.443936
90	24289.4	24060.375518	229.024482

Tabla N°30: SVR: Test vs Predicción Producción de Palta

Fuente: Elaboración Propia

	Prod Palta Real	Prod Palta Pred	Difference
180	42896.0	54670.8	-11774.8
120	13221.4	13739.2	-517.8
188	14163.0	14465.6	-302.6
185	14214.4	14465.6	-251.2
191	13940.8	14033.8	-93.0
117	13935.6	13739.2	196.4
87	14892.0	14033.8	858.2
158	609.0	5769.0	-5160.0
91	14632.2	14033.8	598.4
90	24289.4	25909.2	-1619.8

Tabla N°31: Árbol de regresión: Test vs Predicción Producción de Palta

Fuente: Elaboración Propia

Producción de Mandarina

De esta forma, podemos observar la diferencia que existe entre los valores estimados y reales de producción, en este caso se evidencia la alta dispersión del algoritmo de Regresión Lineal Múltiple, la baja del SVR y la moderada del Árbol de Regresión.

	Prod Mandarina Real	Prod Mandarina Pred	Difference
102	28327.20	-15469.730765	43796.930765
72	104869.00	101407.599073	3461.400927
164	31751.90	26998.446572	4753.453428
73	14540.40	16594.775026	-2054.375026
77	67584.60	100339.153984	-32754.553984
138	36348.00	28575.983831	7772.016169
101	159601.80	115605.579971	43996.220029
100	30513.00	-5067.777250	35580.777250
153	85146.20	95770.515963	-10624.315963
115	21080.32	24298.399549	-3218.079549
145	41677.00	42714.605190	-1037.605190
151	17191.00	9362.825632	7828.174368
98	109400.20	83077.772527	26322.427473
130	9670.00	8079.584540	1590.415460

Tabla N°32: Regresión Lineal Múltiple: Test vs Predicción Producción de Mandarina

Fuente: Elaboración Propia

	Prod Mandarina Real	Prod Mandarina Real Pred	Difference
102	28327.20	28631.483140	-304.283140
72	104869.00	106342.215777	-1473.215777
164	31751.90	31348.628447	403.271553
73	14540.40	15697.586474	-1157.186474
77	67584.60	70469.771782	-2885.171782
138	36348.00	37514.786619	-1166.786619
101	159601.80	162122.588532	-2520.788532
100	30513.00	30410.330915	102.669085
153	85146.20	85286.680527	-140.480527
115	21080.32	26044.231295	-4963.911295
145	41677.00	42651.763127	-974.763127
151	17191.00	16324.067598	866.932402
98	109400.20	111217.994038	-1817.794038
130	9670.00	9353.397075	316.602925

Tabla N°33: SVR: Test vs Predicción Producción de Mandarina

Fuente: Elaboración Propia

	Prod Mandarina Real	Prod Mandarina Pred	Difference
102	28327.20	38302.523529	-9975.323529
72	104869.00	149291.100000	-44422.100000
164	31751.90	38302.523529	-6550.623529
73	14540.40	17718.523000	-3178.123000
77	67584.60	73542.120000	-5957.520000
138	36348.00	38302.523529	-1954.523529
101	159601.80	149291.100000	10310.700000
100	30513.00	38302.523529	-7789.523529
153	85146.20	73542.120000	11604.080000
115	21080.32	38302.523529	-17222.203529
145	41677.00	38302.523529	3374.476471
151	17191.00	17718.523000	-527.523000
98	109400.20	149291.100000	-39890.900000
130	9670.00	17718.523000	-8048.523000

Tabla N°34: Árbol de regresión: Test vs Predicción Producción de Mandarina

Fuente: Elaboración Propia

Producción de Arándano

Se observan en las filas 52, 66 y 216 se observan las diferencias entre cada algoritmo, presentándose los valores más bajos en el Árbol de Regresión, siendo estos 12.23, 48.76 y 37.03, respectivamente, lo que de igual forma hace que se opte por esta alternativa dado que la dispersión en los 3 casos no es tan distinta.

	Prod Arandano Real	Prod Arandano Pred	Difference
196	1650.035	545.883775	1104.151225
63	8898.896	8558.643665	340.252335
214	163.800	-782.726891	946.526891
52	692.652	487.459878	205.192122
26	5325.416	5041.954821	283.461179
209	919.786	967.484432	-47.698432
37	4425.240	4273.113572	152.126428
193	913.214	679.561272	233.652728
19	1690.620	3053.301349	-1362.681349
47	7994.778	8556.197247	-561.419247
14	431.130	544.722184	-113.592184
3	3928.470	3854.314538	74.155462
161	1793.224	1755.875447	37.348553
2	4330.320	4783.913732	-453.593732
163	2384.188	2526.238821	-142.050821
48	7504.507	7562.647296	-58.140296
66	1366.310	663.901841	702.408159
39	4843.760	3493.705773	1350.054227
211	1047.718	1365.698532	-317.980532
216	717.450	42.723649	674.726351
1	3565.390	3597.451540	-32.061540
43	5333.750	4822.392964	511.357036

Tabla N°35: Regresión Lineal Múltiple: Test vs Predicción Producción Arándano

Fuente: Elaboración Propia

	Prod Arandano Real	Prod Arandano Real Pred	Difference
196	1650.035	-434.848901	1215.186099
63	8898.896	9262.393248	-363.497248
214	163.800	-253.900964	-90.100964
52	692.652	247.747155	444.904845
26	5325.416	4985.567592	339.848408
209	919.786	-763.758567	156.027433
37	4425.240	4212.733430	212.506570
193	913.214	-2085.233482	-1172.019482
19	1690.620	1795.854494	-105.234494
47	7994.778	8743.473268	-748.695268
14	431.130	145.550328	285.579672
3	3928.470	4410.791738	-482.321738
161	1793.224	1127.952290	665.271710
2	4330.320	4515.704996	-185.384996
163	2384.188	2104.556198	279.631802
48	7504.507	8095.084942	-590.577942
66	1366.310	1373.042588	-6.732588
39	4843.760	4198.086511	645.673489
211	1047.718	-246.398424	801.319576
216	717.450	421.169945	296.280055
1	3565.390	2151.495045	1413.894955
43	5333.750	4557.240684	776.509316

Tabla N°36: SVR: Test vs Predicción Producción de Arándano

Fuente: Elaboración Propia

	Prod Arándano Real	Prod Arándano Pred	Difference
196	1650.035	1317.544400	332.490600
63	8898.896	7999.315000	899.581000
214	163.800	680.415400	-516.615400
52	692.652	680.415400	12.236600
26	5325.416	5162.464889	162.951111
209	919.786	680.415400	239.370600
37	4425.240	5162.464889	-737.224889
193	913.214	680.415400	232.798600
19	1690.620	2142.093000	-451.473000
47	7994.778	7999.315000	-4.537000
14	431.130	680.415400	-249.285400
3	3928.470	4069.829000	-141.359000
161	1793.224	1317.544400	475.679600
2	4330.320	4069.829000	260.491000
163	2384.188	2678.133000	-293.945000
48	7504.507	7999.315000	-494.808000
66	1366.310	1317.544400	48.765600
39	4843.760	5162.464889	-318.704889
211	1047.718	680.415400	367.302600
216	717.450	680.415400	37.034600
1	3565.390	2142.093000	1423.297000
43	5333.750	5162.464889	171.285111

Tabla N°37 - Árbol de regresión: Test vs Predicción Producción de Arándano

Fuente: Elaboración Propia

Pronóstico 2:

Cantidad de Personal Palta

Para la simulación de la cantidad de personas necesarias para el cultivo de palta, usamos el 20% de nuestros datos para el test, aquí comparamos nuestro Ytest con los datos que ha predicho nuestro modelo. Donde se puede observar unos varios valores atípicos que puede darse por peticiones del cliente que pide que se coseche antes de lo programado o después de lo programado, por los cual puede variar las cosechas y hacen que se requiera más o menos personal, pero esto no siempre se da así, sino que se sigue un programa. En este caso la menor dispersión la presenta el Árbol de Regresión.

	Q Trabajadores Palto Real	Q Trabajadores Palto Pred	Difference
180	47	21	26
120	60	66	-6
188	63	39	24
185	38	30	8
191	55	34	21
117	70	63	7
87	51	57	-6
158	6	27	-21
91	40	44	-4
90	30	39	-9

Tabla N°38- Regresión Lineal Múltiple: Test vs Predicción de cantidad de trabajadores Palta

Fuente: Elaboración Propia

	Mano de Obra Palto Real	Mano de Obra Palto Pred	Difference
180	47	26	21
120	60	71	-11
188	63	45	18
185	38	31	7
191	55	41	14
117	70	67	3
87	51	56	-5
158	6	24	-18
91	40	43	-3
90	30	35	-5

Tabla N°39 - SVR: Test vs Predicción de cantidad de trabajadores Palta

Fuente: Elaboración Propia

	Q Trabajadores Palta Real	Q Trabajadores Palta Pred	Difference
180	47	42	5
120	60	51	9
188	63	51	12
185	38	31	7
191	55	51	4
117	70	51	19
87	51	51	0
158	6	18	-12
91	40	51	-11
90	30	51	-21

Tabla N°40 - Árbol de regresión: Test vs Predicción de cantidad de trabajadores Palta

Fuente: Elaboración Propia

Cantidad de Personal Mandarina

Para la simulación de la cantidad de personas necesarias para el cultivo de mandarina, usamos el 20% de nuestros datos para el test, aquí comparamos nuestro Ytest con los datos que ha predicho nuestro modelo. Donde en varias filas la variación es mínima, sin embargo, también encontramos algunos valores atípicos.

	Q Trabajadores Mandarina Real	Q Trabajadores Mandarina Pred	Difference
102	4	1	3
72	14	16	-2
164	35	24	11
73	18	19	-1
77	38	30	8
138	47	54	-7
101	18	-3	21
100	15	4	11
153	51	49	2
115	35	54	-19
145	13	-1	14
151	3	7	-4
98	1	-1	2
130	1	-1	2

Tabla N°41 - Regresión Lineal Múltiple: Test vs Predicción de cantidad de trabajadores Mandarina

Fuente: Elaboración Propia

	Mano de Obra Mandarina Real	Mano de Obra Mandarina Pred	Difference
102	4	8	-4
72	14	19	-5
164	35	44	-9
73	18	15	3
77	38	21	17
138	47	46	1
101	18	8	10
100	15	13	2
153	51	45	6
115	35	44	-9
145	13	12	1
151	3	4	-1
98	1	8	-7
130	1	3	-2

Tabla N°42 - SVR: Test vs Predicción de cantidad de trabajadores Mandarina

Fuente: Elaboración Propia

	Q Trabajadores Mandarina Real	Q Trabajadores Mandarina Pred	Difference
102	4	4	0
72	14	19	-5
164	35	29	6
73	18	11	7
77	38	39	-1
138	47	50	-3
101	18	19	-1
100	15	19	-4
153	51	36	15
115	35	57	-22
145	13	13	0
151	3	7	-4
98	1	1	0
130	1	7	-6

Tabla N°43 - Árbol de regresión: Test vs Predicción de cantidad de trabajadores Mandarina

Fuente: Elaboración Propia

Cantidad de Personal Arándano

Con respecto a la cantidad de personal para arándanos, se sabe que el requerimiento para su cosecha es de 10 personas por hectárea, alcanzando en picos de campaña un aproximado de 200 personas para atender los 2 lotes existentes, por lo cual la predicción de la muestra 193 que indica un total de 200 de la Regresión Lineal Múltiple hace referencia a este escenario, pero sin embargo se presenta una estimación con menor diferencia empleando el SVR, dada las hectáreas y el número de plantas en número más cercano al que debe llegar la estimación es de 140 en un escenario ideal.

	Q Trabajadores Arandano Real	Q Trabajadores Arandano Pred	Difference
196	91	136	-45
63	81	73	8
214	5	36	-31
52	54	49	5
26	36	33	3
209	72	87	-15
37	29	18	11
193	170	200	-30
19	12	16	-4
47	90	89	1
14	16	22	-6
3	13	13	0
161	35	33	2
2	14	29	-15
163	37	32	5
48	96	107	-11
66	11	-4	15
39	31	34	-3
211	113	101	12
216	16	39	-23
1	14	24	-10
43	101	94	7

Tabla N°44 - Regresión Lineal Múltiple: Test vs Predicción de cantidad de trabajadores Arándano

Fuente: Elaboración Propia

	Mano de Obra Arandano Real	Mano de Obra Arandano Pred	Difference
196	91	132	-41
63	81	60	21
214	5	29	-24
52	54	48	6
26	36	32	4
209	72	69	3
37	29	30	-1
193	170	186	-16
19	12	14	-2
47	90	94	-4
14	16	14	2
3	13	14	-1
161	35	33	2
2	14	17	-3
163	37	24	13
48	96	106	-10
66	11	15	-4
39	31	39	-8
211	113	82	31
216	16	29	-13
1	14	16	-2
43	101	102	-1

Tabla N°45 - SVR: Test vs Predicción de cantidad de trabajadores Arándano

Fuente: Elaboración Propia

	Q Trabajadores Arándano Real	Q Trabajadores Arándano Pred	Difference
196	91	145	-54
63	81	52	29
214	5	6	-1
52	54	55	-1
26	36	32	4
209	72	85	-13
37	29	45	-16
193	170	170	0
19	12	11	1
47	90	88	2
14	16	14	2
3	13	32	-19
161	35	33	2
2	14	32	-18
163	37	45	-8
48	96	88	8
66	11	11	0
39	31	32	-1
211	113	85	28
216	16	6	10
1	14	11	3
43	101	91	10

Tabla N°46 - Árbol de regresión: Test vs Predicción de cantidad de trabajadores Arándano

Fuente: Elaboración Propia

CAPÍTULO VI: Conclusiones y Recomendaciones

6.1 Conclusiones

El sector agroindustrial, presenta una significativa demanda en cuanto a la alimentación de la población. Asimismo, los cambios que se presentan son bastantes inciertos, como el cambio climático, una guerra, hasta una pandemia. Por lo que, cada vez más presenta nuevos retos en cuanto a la estimación y la cantidad de cosecha a producir, ya que se desarrolla en un contexto más difícil y voluble. Es por ello, que obtener los volúmenes anticipadamente, significa poder planear correctamente las operaciones agrícolas y comerciales. Por consiguiente, permitirá conocer los costos que representan estas operaciones.

En el presente trabajo al ser realizado por la metodología CRISP-DM, nos permitió obtener un mejor estructura al obtener el reconocimiento del negocio por parte del usuario experto, el cual nos explicó las bases del negocio, así como la explicación de las bases de datos no relacionadas que empleamos en este proyecto, la cual mediante la preparación de los datos se realizaron tareas de normalización de las diferentes fuentes de información, para crear una relación entre estas y tener una base de datos general que contiene la información general y dividirlo en dataset por cultivo (Palta, mandarina y arándano). Asimismo, se realizó las imputaciones correspondientes de los datos vacíos, de esta forma los data sets por cultivo tendrían la información completa.

Para el modelamiento, en base a las investigaciones realizadas, se concluyó que se usarían 3 algoritmos por cada modelo, producción y cantidad de trabajadores, que se iba a predecir, las cuales son regresión lineal múltiple, support vector regression y árboles de regresión, estas dos últimas se usaron parámetros específicos para que tengan los valores óptimos, asimismo con estos 18 resultados se evaluó con dos medidas de errores estadísticos (R^2 y RMSE), que permitió discernir el mejor algoritmo que se adecua a los dos modelos por palta, mandarina y arándano

Posteriormente se obtuvieron los resultados de los errores estadísticos de cada modelo para realizar la comparación entre las mismas, concluyendo que el algoritmo de Árbol de Regresión es el más adecuado para estimar la cantidad producida en kg para todos los cultivos, a pesar de mostrar el algoritmo SVR valores de R2 muy cercanos al 1 a

diferencia del Árbol de Regresión (0.84, 0.83 y 0.96), pero esto se debe al alto grado de ajuste de la data lo que hace esta alternativa sea muy optimista.

En el caso del segundo modelo, se concluye que el algoritmo SVR es el más óptimo dado que las R2 para el caso de las mandarinas y arándanos mas no para la palta que tienen 0.55 a diferencia del 0.51. De igual manera, RMSE para la mandarina y arándano presenta menor dispersión que con los otros algoritmos, sin embargo, es importante mencionar que se obtiene del algoritmo SVR un RMSE no tan diferenciado con respecto al Árbol de Regresión siendo de 12.30 y 11.75, respectivamente.

La finalidad de este trabajo es brindar opciones de herramientas que puedan usarse para una mejor planificación operativa tomando en cuenta otras variables que en la práctica no se utilizan para una más acertada proyección de cosecha, que impacta en el requerimiento de personal para las campañas y variaciones en las proyecciones presupuestales de los gastos del negocio. Podemos concluir que los patrones o tendencias hallados en cada modelo se ajustan a los valores de la BD, esto debido a un buen resultado de los métodos empleados.

6.2 Recomendaciones

Como principal recomendación, se sugiere utilizar mayor cantidad de datos para obtener un grado alto de confiabilidad en las pruebas de los modelos propuestos. Asimismo, es importante realizar un correcto procesamiento de datos, con el fin de eliminar los valores erróneos o atípicos. En este sentido, se sugiere realizar el cálculo de la correlación entre las variables dependientes e independientes (Coeficiente de Pearson) e indicar si dichas correlaciones son significativas. Por otro lado, también se deben determinar los datos atípicos dentro de las variables, de esta forma podremos reducir algún efecto de desproporción en los resultados de los modelos y tener un resultado más acertado.

También se propone el uso de otras técnicas de Machine Learning como Random Forest y Gradient Boosting Machine. Esto permitirá realizar una mayor comparación de los resultados de cada modelo y obtener el mejor modelo que permita predecir las variables de producción por cultivo y cantidad de trabajadores para la labor de cosecha por cultivo. Asimismo, se recomienda utilizar otras métricas de error estadístico como MSE Y R^2 ajustado que nos ayude a determinar mejor la precisión de los modelos evitando los sobre ajustes por la mayor cantidad de variables.

En el presente trabajo se han investigado los tipos de kernel más comunes (lineal, polinomial, rbf). Se propone profundizar en nuevas tipologías como el sigmoid, hyperbolic tangent, exponencial, entre otros. De esta manera, se podrían identificar otras funciones que nos permitan trabajar con un mejor modelamiento. Asimismo, debido al alto costo computacional que se hace uso de los kernels no lineales, dado que las muestras de entrenamiento que son grandes (en nuestro caso 80%) el algoritmo crea una matriz de NxN de distancias entre los puntos dichos puntos de entrenamiento, lo que conlleva a un uso excesivo de memoria RAM, por lo que se recomienda hacer uso de instancias en la nube que permite acelerar el resultado.

El estudio se limitó hasta la fase de evaluación según la metodología CRISP-DM. donde determinamos los modelos que tienen mejor resultado acorde por cada modelo y cultivó, la última etapa que es la de despliegue no se llegó a realizar, dado que este estudio es un estudio inicial, y esta etapa requiere actividades de planificación, monitoreo y mantenimiento.

Bibliografía

Arana, C. (2021). *Modelos de Aprendizaje Automático Mediante Árboles de Decisión*. Serie Documentos de Trabajo, (778), 1+. https://link.gale.com/apps/doc/A687634783/IFME?u=uesan&sid=bookmark-IFME&xid=40e4fb81

Benítez, R., Escudero S., Kanaan & Masip, D. (2013). Inteligencia Artificial Avanzada. Editorial UOC.

Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5. https://doi.org/10.1023/a:1010933404324

Clayton, C. (2018). Group warns food supply won't meet future demand without productivity gains. DTN Progressive Farmer. https://www.dtnpf.com/agriculture/web/ag/news/world-policy/article/2018/10/17/group-warns-food-supply-meet-future-2

Chambers, M., y Dinsmore, T. W. (2015). *Advanced analytics methodologies: Driving business value with analytics. NJ: Pearson Education*

ESAN. (2018). Minería de datos: ¿en qué consiste el knowledge discovery in databases? 27/07/2022, de ESAN Sitio web: https://www.esan.edu.pe/conexion-esan/mineria-de-datos-en-que-consiste-el-knowledge-discovery-in-databases#:~:text=El%20KDD%20es%20un%20proceso,recursos%20%C3%BAtiles%20para%20una%20compa%C3%B1%C3%ADa.

FAO, 2018. *The future of food and agriculture – Alternative pathways to 2050*. Rome. 224 pp. Licencia: CC BY-NC-SA 3.0 IGO.

Grüter, R., Trachsel, T., Laube, P., Jaisli, I. (2022) *Expected global suitability of coffee, cashew and avocado due to climate change*. PLoS ONE 17(1): e0261976. https://doi.org/10.1371/journal.pone.0261976

Hernández i, R. and Mendoza, C., 2018. *Metodología de la investigación*. 1st ed. MÉXICO: McGRAW-HILL INTERAMERICANA EDITORES.

IBM. (SF). Regresión lineal. 27/07/2022, de IBM Sitio web: https://www.ibm.com/pe-es/analytics/learn/linear-regression

IBM. (sf). Guía de CRISP-DM de IBM SPSS Modeler. 27/07/2022, de ibm Sitio web: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

Kelleher, J., Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (1.ª ed.). INGLATERRA: The MIT Press.

Lamos, H., Puentes, D. & Zarate, D. (2020). C*omparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia*. Revista Facultad de Ingeniería, 29(54), Artículo e10853. https://doi.org/10.19053/01211129.v29.n54.2020.10853

L. Chen. "Support Vector Machine – Simply Explained", towardsdatascience.com. [Online]. Available: towardsdatascience.com/support-vectormachine-simply-explained-fee28eba5496. [Accessed: 5-Jun-2021].

Lozano, D. (2015). *Modelos predictivos del churn – abandono de clientes – para operadores de telecomunicaciones.*

Machado, L., Rodríguez, L., Orduz S. A., S*aavedra, D., Murcia, V., Vásquez, A.M., Coronado, H.H., & Méndez, D.A. (2020). Aplicación de nuevas tecnologías en el fortalecimiento de la cadena agroindustrial*. Revista SENNOVA, vol1, pp.101. 2022, julio 5, De SENA Base de datos.

MIDAGRI (2022) Las *agroexportaciones suman nuevo récord y superaron los US$ 9,000 millones en ventas el 2021.* [Comunicado de prensa] https://www.gob.pe/institucion/midagri/noticias/583476-las-agroexportaciones-suman-nuevo-record-y-superaron-los-us-9-000-millones-en-ventas-el-2021

Morales, C. (2017) *Manual de manejo agronómico del arándano* [en línea]. Villa Alegre, Chile: Boletín INIA - Instituto de Investigaciones Agropecuarias. no. 371. Disponible en: https://hdl.handle.net/20.500.14001/6673 (Consultado: 22 julio 2022).

Oded Maimon. (sf). *INTRODUCTION TO KNOWLEDGE DISCOVERY IN DATABASES*. 27/07/2022, de Tel-Aviv University Sitio web: https://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf

Perez, M.. (2017, diciembre). *Machine learning and econometric applications for increasing profitability and efficiency: A case study on sustainable production and trade in agro-based industries*. Tesis Doctoral, Vol1, pp. 144. 2022, julio 22, De USAL Base de datos

Pimienta, J. & De La Orden, A., 2017. *Metodología de la investigación*. Tercera ed. México: Pearson education. ISBN 978-607-32-3933-2

Prestifilippo, C. (2020). *IA en la gestión de las personas. Escuela de Administración y Negocios,* vol.1, pp.60. 2022, julio 26, De Universidad de San Andrés Base de datos.

¿Qué es un árbol de decisión? | IBM. (s. f.). Recuperado 11 de octubre de 2022, de https://www.ibm.com/es-es/topics/decision-trees

Ramirez, C.A. (2020, diciembre 3). *Aplicación del Machine Learning en Agricultura de Precisión*. Revista Cintex, Vol25(2), pp. 14-27. 2022, julio 25, De CEASOFT Base de datos.

Raschka, S., & Miirjalili, V. (2019). *Python Machine Learning* (2.ª ed.). MARCOMBO S.A.

Raschka, S. (2015). *Python Machine Learning: Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics*. Packt Publishing.

Rodrigo, J. A. (s. f.-a). Arboles de decision, Random Forest, Gradient Boosting y C5.0. Recuperado 11 de octubre de 2022, de https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting

Rodrigo, J. A. (s. f.-b). Correlación lineal y Regresión lineal simple. Recuperado 11 de octubre de 2022, de https://www.cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

Rodrigo, J. A. (s. f.). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Recuperado 11 de octubre de 2022, de

https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Rodríguez, D. (2021b, noviembre 6). Regresión de Vectores de Soporte (SVR, Support Vector Regression). Analytics Lane. Recuperado 11 de octubre de 2022, de https://www.analyticslane.com/2021/12/17/regresion-de-vectores-de-soporte-svr-support-vector-regression/

Rouhiainen, L. (2018). I*nteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro*. España: Alienta Editorial.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

VANDERPLAS, J., 2017. *Python data science handbook: essential tools for working with data*. New York: O'Reilly Media. ISBN 978-1-491-91205-8.

Wooldridge, J. M. (2015). *Introducción a la econometría*. Cengage Learning. https://www.ebooks7-24.com:443/?il=1305

Zaki, M. J., & Wagner Meira. (2014). Data mining and analysis : fundamental concepts and algorithms. Cambridge University Press.

Anexos

1. **Tabla de % registros completos**

Variable	% Registro Completo
PERIODO	100%
PlantasTotales	100%
DESCRIPCION	100%
HorasTotales	100%
Edad_Prom	100%
CostoTotal	100%
Cantidad_Trabajadores	100%
AÑO	100%
HA_TOTALES	100%
TotalNeto	100%
Pesotara	100%
TotalBruto	100%
Nrojabas	100%
LT_Total	100%
avg In Dew	99%
avg UV Dose	99%
avg Hi UV	99%

avg Heat D-D	**99%**
avg Cool D-D	**99%**
avg In Temp	**99%**
avg In Hum	**99%**
avg In EMC	**99%**
avg In Heat	**99%**
avg Hi Solar Rad.	**99%**
avg In Air Density	**99%**
avg ET	**99%**
avg Wind Samp	**99%**
avg Wind Tx	**99%**
avg ISS Recept	**99%**
avg Arc. Int.	**99%**
avg UV Index	**99%**
avg Rain Rate	**99%**
avg Solar Energy	**99%**
Wind Dir	**99%**
avg Temp Out	**99%**
avg Hi Temp	**99%**

avg Low Temp	**99%**
avg Out Hum	**99%**
avg Dew Pt.	**99%**
avg Wind Speed	**99%**
avg Wind Run	**99%**
avg Solar Rad.	**99%**
avg Hi Speed	**99%**
Hi Dir	**99%**
avg Wind Chill	**99%**
avg THW Index	**99%**
avg Bar	**99%**
avg Rain	**99%**
avg Heat Index	**99%**

2. Tabla de Tareo de personal de cosecha

PERIODO	Codigo Con	DESCRIPCION	Hora	Costo	Cantidad_Trabajado
W322019	TR011001	ARANDANO	289	S/ 1,615.90	15
W332019	TR011001	ARANDANO	206	S/ 1,142.47	16
W342019	TR011001	ARANDANO	365	S/ 2,335.56	14
W352019	TR011001	ARANDANO	343	S/ 2,223.16	12
W362019	TR011001	ARANDANO	513	S/ 3,181.38	13
W372019	TR011001	ARANDANO	571	S/ 3,265.48	13
W382019	TR011001	ARANDANO	117	S/ 668.33	13
W392019	TR011001	ARANDANO	416	S/ 2,495.79	12
W402019	TR011001	ARANDANO	397	S/ 2,625.58	11
W412019	TR011001	ARANDANO	500	S/ 3,119.55	17
W422019	TR011001	ARANDANO	509	S/ 2,910.68	14
W432019	TR011001	ARANDANO	691	S/ 4,367.93	14
W442019	TR011001	ARANDANO	651	S/ 4,461.03	13
W452019	TR011001	ARANDANO	482	S/ 3,328.24	13
W462019	TR011001	ARANDANO	526	S/ 2,946.91	11
W472019	TR011001	ARANDANO	498	S/ 2,945.76	11
W482019	TR011001	ARANDANO	410	S/ 2,467.30	11
W492019	TR011001	ARANDANO	410	S/ 2,425.57	11
W502019	TR011001	ARANDANO	451	S/ 2,704.12	11
W512019	TR011001	ARANDANO	332	S/ 1,973.00	12
W522019	TR011001	ARANDANO	346	S/ 2,520.06	14
W532019	TR011001	ARANDANO	195	S/ 1,456.28	11
W022020	TR011001	ARANDANO	210	S/ 3,125.22	12
W312020	TR011001	ARANDANO	1023	S/ 7,059.37	37
W322020	TR011001	ARANDANO	1198	S/ 7,099.08	28
W332020	TR011001	ARANDANO	1455	S/ 8,805.54	29
W342020	TR011001	ARANDANO	2039	S/ 11,984.63	42
W352020	TR011001	ARANDANO	1687	S/ 10,689.50	36
W362020	TR011001	ARANDANO	1629	S/ 10,820.22	35
W372020	TR011001	ARANDANO	1538	S/ 9,110.30	32

3. Tabla de cantidad de plantas, hectáreas y edades planta

Año	Cod Consumi	# de plan	HA	Eda
2019	TR010503	676	1.5	1
2019	TR010106	1200	2.5	17
2019	TR010404	1068	3.1	17
2019	TR010102	1491	4.7	17
2019	TR010103	2123	6.6	17
2019	TR010109	1170	3	17
2019	TR010104	777	2.8	17
2019	TR010501	10053	21.4	6
2019	TR010110	560	0.7	17
2019	TR011001	45000	6	2
2019	TR010203	2579	5	5
2019	TR011002	0	6	0
2019	TR010107	485	1	17
2020	TR010110	560	0.7	18
2020	TR010107	401	1	18
2020	TR010103	2061	6.6	18
2020	TR010102	1468	4.7	18
2020	TR011001	45000	6	3
2020	TR011002	60000	6	1
2020	TR010104	703	2.8	18
2020	TR010106	1108	2.5	18
2020	TR010503	676	1.5	2
2020	TR010404	1068	3.1	18
2020	TR010203	2579	5	6
2020	TR010501	10053	21.4	7
2020	TR010109	1138	3	18
2021	TR010404	1004	3.1	19
2021	TR010503	676	1.5	3
2021	TR011001	45000	6	4
2021	TR010107	401	1	19
2021	TR010203	2579	5	7
2021	TR010104	703	2.8	19
2021	TR010103	2061	6.6	19
2021	TR010501	10053	21.4	8
2021	TR010110	560	0.7	19
2021	TR010102	1468	4.7	19
2021	TR010109	1138	3	19
2021	TR011002	60000	6	2
2021	TR010106	1108	2.5	19
2022	TR010102	1972	4.7	1
2022	TR010110	605	0.7	20
2022	TR011001	43842	6	5
2022	TR010503	676	1.5	4
2022	TR010106	1248	2.5	20
2022	TR010501	10053	21.4	9
2022	TR010109	1154	3	20
2022	TR010203	2579	5	8
2022	TR010104	1150	2.8	1
2022	TR010404	1635	3.1	20
2022	TR010103	2756	6.6	1
2022	TR010107	500	1	20
2022	TR011002	56557	6	3

4. Tabla de códigos por lote

CODPRD	DESCRIPCION	IND_EXISTE
TR0101	PALTO	1
TR010101	PALTO	1
TR010102	PALTO	1
TR010103	PALTO	1
TR010104	PALTO	1
TR010106	PALTO	1
TR010107	PALTO	1
TR010109	PALTO	1
TR010110	PALTO	1
TR010111	PALTO	1
TR010203	MANDARINA	1
TR010404	PALTO	1
TR010501	MANDARINA	1
TR010503	MANDARINA	1
TR010708	CAMINOS	1
TR011001	ARANDANO	1
TR011002	ARANDANO	1