FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

“DESARROLLO DE UN MODELO PREDICTIVO PARA EL ANÁLISIS DE DATOS DEL SECTOR EXPORTADOR HORTOFRUTÍCOLA EN EL PERIODO 2016 -2020”

Tesis para optar el título profesional de:

Ingeniero de Sistemas Computacionales

Autores:
Bryam André Tapia Álvarez
Wernher D’alembert Chávez Sánchez

Asesor:
MBA Christiaan Michaell Romero Zegarra

Cajamarca - Perú

2020

DEDICATORIA

A mis familiares, por estar presentes y brindando
el apoyo moral a pesar de la distancia,
y a las personas que apoyaron a realizar
exitosamente este trabajo con
sus conocimientos.
Wernher D'alembert Chávez Sánchez

A mi padre, mi madre, por el apoyo incondicional,
y a mis hermanos que siempre están presentes.
Bryam André Tapia Álvarez

AGRADECIMIENTO

A mi abuela y hermanos por ser el apoyo
y fortaleza en momentos de dificultad,
por confiar y creer en mis expectativas
y por los principios morales que
inculcaron en mi persona.
Wernher D'alembert Chávez Sánchez

A las personas que nos apoyaron durante el
trancurso de este proyecto, a nuestro
asesor, y consultores que siempre
tuvieron la mejor disponibilidad ante
cualquier consulta.
Bryam André Tapia Álvarez

TABLA DE CONTENIDOS

DEDICATORIA *2*

AGRADECIMIENTO *3*

ÍNDICE DE TABLAS *7*

ÍNDICE DE FIGURAS *8*

RESUMEN *12*

CAPÍTULO I. INTRODUCCIÓN *13*

1.1. Realidad Problemática 13

1.2. Formulación del problema 32

1.3. Objetivos 33

1.3.1. Objetivo general 33

1.3.2. Objetivos específicos 33

1.4. Hipótesis 33

CAPÍTULO II. METODOLOGÍA *34*

2.1. Tipo de investigación 34

2.2. Población y muestra (Materiales, instrumentos y métodos) 35

2.2.1. Población 35

2.2.2. Muestra 35

2.3. Técnicas e instrumentos de recolección y análisis de datos 35

2.3.1. Descripción de la técnica para la recolección de datos 35

2.3.2. Confiabilidad y Validez de los instrumentos 36

2.4. Aspectos Éticos 36

2.5. Procedimiento 37

CAPÍTULO III. RESULTADOS 46

3.1. Resultados según Objetivos 46

IV. DISCUSIÓN Y CONCLUSIÓNES 74

4.1. DISCUSIÓN: 74

4.2. CONCLUSIONES: 77

REFERENCIAS 79

ANEXOS 85

Anexo Nº 1. Operacionalización de variables. 85

Anexo Nº 2. Población. 87

Anexo Nº 3. Fórmula para cálculo de muestra muestra. 87

Anexo Nº4. Fichas de Validación. 88

Anexo Nº5. Carta de autorización. 90

Anexo Nº6. Marco Conceptual. 91

Anexo Nº7. Análisis Preliminar. 97

Anexo Nº 8. Definiciones de algoritmos de minería de datos. 100

Anexo Nº 9. Selección de reportes. 101

Anexo Nº10. Reporte de descripción de datos. 103

Anexo Nº 11. Reporte Estándar Shipping Point 104

Anexo Nº 12. Reporte de calidad de los datos Nº1. 105

Anexo Nº 13. Reporte de selección, estructurado e integración de datos. 107

Anexo Nº 14. Reporte de calidad de datos Nº2. 111

Anexo Nº 15. Selección de Metricas. 114

Anexo Nº 16. Resultados obtenidos del modelado predictivo. 116

Anexo Nº 18. Informe comparativo de resultados. 125

Anexo Nº 19. Reporte de últimos 21 meses. 129

Anexo Nº 20. Estructura Organizacional Del Área Comercial. 132

ÍNDICE DE TABLAS

Tabla 1. Comparación de reportes según países. 46

Tabla 2. Top 3 de frutas que más se exportan. 47

Tabla 3. Relación de técnicas de minería de datos. 48

Tabla 4. Formato inicial de Reporte. 49

Tabla 5. Reporte Estándar de Precio. 52

Tabla 6. Formato final. 52

Tabla 7. Operacionalización de Variable Independiente 85

Tabla 8. Operacionalización de Variable Dependiente 86

Tabla 9. Número de reportes del 2016 al 2020 87

Tabla 10. Formato inicial de Reporte. 103

Tabla 11. Reporte Estándar de Precio. 107

Tabla 12. Empaquetado de Paltas. 108

Tabla 13. Reporte con conversión a Kg. 108

Tabla 14. Formato de reporte con datos principales. 109

Tabla 15. Formato final. 110

Tabla 16. Comparación Daily Only. 125

Tabla 17. Comparación Variety Price. 126

Tabla 18. Comparación Origin and Variety 127

ÍNDICE DE FIGURAS

Figura 1. Proceso de Selección de datos. 37

Figura 2. Resultados de reportes USDA. 38

Figura 3. Ejemplo de Reporte Original. 38

Figura 4. Arquitectura Data Mining. 39

Figura 5. Fase 1: Comprensión del negocio o problema. 39

Figura 6. Fase 2: Comprensión de los datos. 40

Figura 7. Fase 3: Preparación de los datos. 41

Figura 8. Fase 4: Modelado. 42

Figura 9. Fase 5: Evaluación. 43

Figura 10. Fase 6: Implementación. 44

Figura 11. Ciclo de Vida de CRISP-M. 45

Figura 12. Gráfico de lineas de Precio por Variedad. 50

Figura 13. Gráfico de lineas de Precio por Origen. 50

Figura 14. Top 10 de Algoritmos y métodos usados por Data Scientists. 54

Figura 15. Top 10 métodos de Data Science usados en 2018-2019. 54

Figura 16. Ejemplo de matriz de confusión. 55

Figura 17. Comparativa de sofware de data mining. 56

Figura 18. Modelado. 57

Figura 19. Evaluation Results Daily Price Only. 59

Figura 20. Evaluation Results Variety Price. 59

Figura 21. Postman de API desarrollada. 62

Figura 22. Gráfico de los últimos 4 años con respecto a precio. 63

Figura 23. Gráfico de los últimos 4 años con respecto a precio y variety. 64

Figura 24. Gráfico de los últimos 4 años con respecto a precio y origen. 64

Figura 25. Variedad Versión Móvil. 65

Figura 26. Origen Versión Móvil. 66

Figura 27. Diversos Datos Versión Móvil. 67

Figura 28. Pantalla Inicial con Predicciones. 68

Figura 29. Fórmula de Costo por Hora de Trabajo. 71

Figura 30. Costo de ejecución del Proyecto. 72

Figura 31. Fórmula de muestreo. 87

Figura 32. Países con Mayor Exportación 2019. 97

Figura 33. Formato Sunat. 98

Figura 34. Resultados de reportes USDA. 99

Figura 35. Resultados de reporte Shipping Point. 101

Figura 36. Resultados de reporte Terminal Market. 102

Figura 37. Resultados de reporte Movement. 102

Figura 38. Reporte Estándar Shipping Point USDA. 104

Figura 39. Linear Projection de un Reporte. 105

Figura 40. Linear Projection Extra Data. 106

Figura 41. Formato de Reporte Final más Return. 111

Figura 42. *Linear Projection Variety Price.* 112

Figura 43. Linear Projection Origin and Variety Price. 113

Figura 44. Matriz de Confusión y Métricas 114

Figura 45. *Daily Only Confusion Matrix SVM.* 116

Figura 46. *Daily Only Confusion Matrix Neural Network.* 116

Figura 47. Daily Only Confusion Matrix Naive Bayes. 117

Figura 48. Daily Only Confusion Matrix kNN. 117

Figura 49. Daily Only Confusion Matrix Logistic Regresion. 118

Figura 50. Daily Only Confusion Matrix Tree. 118

Figura 51. Variety Price Confusion Matrix kNN. 119

Figura 52. Variety Price Confusion Matrix Logistic Regression. 119

Figura 53. Variety Price Confusion Matrix Naive Bayes. 120

Figura 54. Variety Price Confusion Matrix Neural Network. 120

Figura 55. Variety Price Confusion Matrix SVM. 121

Figura 56. Variety Price Confusion Matrix Tree. 121

Figura 57. Variety and Origin Price Confusion Matrix kNN. 122

Figura 58. Variety and Origin Price Confusion Matrix Logistic Regression. 122

Figura 59. Variety and Origin Price Confusion Matrix Naive Bayes. 123

Figura 60. Variety and Origin Price Confusion Matrix Neural Network. 123

Figura 61. Variety and Origin Price Confusion Matrix SVM. 124

Figura 62. Variety and Origin Price Confusion Matrix Tree. 124

Figura 63. *Últimos 21 meses por Price* 129

Figura 64. Últimos 21 meses por Variety. 130

Figura 65. Últimos 21 meses por Origen. 130

Figura 66. Últimos 92 semanas solo Precio. 130

Figura 67. Últimos 92 semanas por Variety. 131

Figura 68. Últimos 92 semanas por Origen. 131

Figura 69. Estructura Organizacional Del Área Comercial. 132

RESUMEN

En el presente trabajo se busca el desarrollo de un modelo predictivo del sector hortofrutícola en el periodo de 2016 – 2020, tomando en cuenta datos de valor, identificando tanto las fuentes como metodologías óptimas para procesar dicha información. Se hizo uso de algoritmos que permiten la predicción de los precios del producto en estudio.

Se hizo uso de la metodología CRISP-DM, siguiendo paso a paso las fases de: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación .

Los resultados concluyeron que se puede mejorar el análisis de datos al desarrollar un modelo predictivo cubriendo las deficiencias en el manejo de información con respecto a la cantidad y calidad de los datos, reduciendo la cantidad de datos que se procesaban en un 50% y aumentando la calidad de los mismos, con un mejora en la toma de decisiones con datos concisos de predicción, con una precisión de 65.2% y eficacia de 66.4%. También se dio una disminución de un 39.08 % del costo anual en los recursos utilizados y un porcentaje de 180% menos tiempo invertido en el análisis.

Palabras clave: CRISP-DM, minería de datos, hortofrutícola

CAPÍTULO I. INTRODUCCIÓN

1.1. Realidad Problemática

La cantidad de información que existe hoy en día va en aumento. Según Agasys (2017) cada minuto que pasa, los 2.700 millones de personas con acceso a Internet que se calcula que hay actualmente en el mundo envían más de 200 millones de correos electrónicos, realizan 2 millones de consultas a Google, suben 48 horas de video a YouTube, escriben más de 100.000 mensajes en Twitter, publican casi 30.000 nuevos artículos en diversos sitios. Entonces si se genera tanta información, cómo es que se aprovecha la misma.

Se puede decir que desde el inicio de la computación se ha utilizado datos e información para el beneficio humano. Identificando la importancia del análisis de la información empezaron a surgir distintas técnicas para su procesamiento. Entre 1950 y 1969 se introduce la comercialización de la analítica mediante la generación del primer modelo de predicción meteorológica con el análisis de datos (Ariser, 2015). Es así como fue surgiendo el análisis predictivo y posteriormente la creación de modelos que se centraron en diversos sectores alrededor del mundo.

Aunque la analítica predictiva existe desde hace décadas, es una tecnología cuyo momento está en auge. Según Ariser (2015) cada vez más organizaciones recurren a la analítica predictiva para mejorar su base de operación y lograr una ventaja competitiva. Ejemplos de esto se pueden evidenciar en: el aumento y generalización del uso de la analítica a través de precios dinámicos en los billetes de avión; recomendaciones de compras, comprobación del tráfico, etc.; en el uso del

análisis para el día a día, por ejemplo: en educación, finanzas, sanidad, etc.; policía preventiva, que sea capaz de adelantarse mediante el análisis a un gran número de delitos; analítica anticipativa, que permita disminuir los accidentes, tanto domésticos como en transportes; cambio en las campañas de marketing como se conocen, ya que la publicidad será mucho más centrada en cada consumidor en particular.

Sin embargo, en Latinoamérica, algunos sectores aun no aprovechan el avance tecnológico del análisis predictivo. Siendo uno de estos, el sector hortofrutícola, y especialmente cuando se habla de la exportación de materia. En un artículo de La República (2019), Holger Matthey, experto de la FAO (Food and Agriculture Organization), aseguró en una conferencia en Roma que: "Latinoamérica será la principal región exportadora, por delante de Norteamérica y Europa, aunque necesita inversiones estratégicas para garantizar una producción sostenible". La expansión de las exportaciones de un país, por lo general, tiene efectos positivos en el crecimiento de la economía y en el de las empresas individuales (Cavusgil & Nevin, 1981). Sin embargo, a pesar de los numerosos beneficios que trae consigo la exportación, la mayoría de las empresas no exportan, a pesar de que se considera a la exportación como inevitable en los mercados mundiales cada vez más integrados (Milanzi, 2012).

Existen diversos factores por lo que las empresas no exportan, pero si se realiza un enfoque en uno de los principales, el correcto análisis de datos del mercado exportador termina destacando como uno de ellos. Por un lado, según Jha y Sinha (2013), la predicción de los precios es una de las cuestiones claves en el análisis sectorial. Sin embargo, tanto el nivel de producción como los precios de los sectores agro son características altamente variables con fuertes dependencias de eventualidades, ya que están sujetos a shocks climáticos y políticos, complejizando la

modelización de su comportamiento y, por consiguiente, la tarea de predecir o pronosticar su evolución futura.

En el estudio "Proyección de precios de exportación utilizando tipos de cambio: Caso peruano", realizado por Ferreyra y Vásquez (2012), se menciona que la economía peruana es vulnerable a las variaciones en los términos de intercambio debido a que es tomadora de precios en el mercado internacional de commodities (productos), principalmente de commodities mineros y agrícolas. La última crisis financiera global de 2008 mostró que los precios de estos productos estuvieron sujetos a cambios violentos, incluso más bruscos de los históricamente registrados y que los proyectados por la mayoría de los agentes del mercado.

En la empresa en la que se realiza el estudio, el análisis de datos es una herramienta utilizada por el área comercial de exportación. Gracias al análisis de datos existe una visión más completa al momento de tomar decisiones de exportación. Sin embargo, está lejos de darse de una manera confiable y que vaya acorde a la tecnología utilizada actualmente. En los últimos años se han identificado 3 factores concurrentes: el manejo de información es deficiente con respecto a la cantidad y calidad de los datos (cada vez es más difícil procesar gran cantidad de datos y la calidad de los mismos no es la mejor), la toma de decisiones se basa en datos subjetivos (estas dependen del criterio de quien realiza el análisis en lugar de valores concisos), y el aumento en el tiempo invertido y el costo de los recursos necesarios para obtener resultados en el análisis de datos (uso deficiente del personal).

Por lo previamente enunciado, el desarrollo de un modelo predictivo para un correcto análisis de datos en el sector exportador hortofrutícola se vuelve esencial. El presente proyecto de investigación busca establecer la creación de un modelo predictivo que satisfaga las necesidades que existen actualmente en la empresa utilizando las herramientas tecnológicas más relevantes de la actualidad. El modelo deberá ser capaz de utilizar información de calidad y optimizar los recursos utilizados, además de presentar una mejora con respecto al tiempo y la forma en que se da el análisis de datos. Teniendo en cuenta estos puntos, existen diversas investigaciones que aportaran al desarrollo del proyecto, estas son las siguientes:

En la investigación, Implementación de un modelo predictivo basado en data mining soportado por SAP Predictive Analitys en retails (Castro & Hernández, 2016), publicada por la Universidad de Ciencias Aplicadas en Lima, Perú, se decidió implementar un modelo predictivo con el objetivo de ayudar a disminuir pérdidas monetarias en la empresa retail prediciendo las ventas. Para el desarrollo del proyecto se realizó una investigación sobre la evolución de SAP Predictive Analytics, información relacionada a la implementación y configuración de la herramienta y casos de éxitos resaltantes de su implementación alrededor del mundo. Después de esto, se analizó la información consolidada para luego configurar e implementar el modelo predictivo en la empresa retail con información real de sus ventas, basándose en algoritmos de predicción que la herramienta brinda. Asimismo, se realizaron las validaciones correspondientes en base a una serie de indicadores. Los autores mencionan que las predicciones de venta juegan un rol importante en la eficiencia del proceso de abastecimiento de una empresa retail, ya que evita la falta o el sobre abastecimiento de stock, esto conlleva que la empresa evite tener pérdidas monetarias. Ellos, al contar con una enorme cantidad de información, pudieron lograr mejores

resultados. Gracias a que contaron con información de ventas de los dos últimos años, los algoritmos utilizados arrojaron resultados bastante favorables e incluso mucho más precisos a una estimación matemática. Al comparar los tres algoritmos usados se observó que el algoritmo con menor MAPE (Error Porcentual Absoluto Medio) y RMSE (Raíz del Error Cuadrático Medio) es el de Triple Exponential Smoothing con 23.6%, luego le sigue Linear Regression con 62.87%, finalmente sigue Monotone Multi Layer Perceptron con 80.91%. Al comparar el algoritmo Triple Exponential Smoothing frente a la predicción que se realizaba en el área de planeamiento de la empresa donde ser realizó el estudio, se identifica que se ha mejorado en la precisión de la proyección, reduciendo el Error Porcentual Absoluto Medio (MAPE) en un 28.49%. Finalmente se menciona, que la solución SAP Predictive Analytics, les ofreció un amplio alcance en cuanto a modelos predictivos gracias a su integración con el lenguaje estadístico R, proporcionando algoritmos de regresión, redes neuronales, series de tiempo entre otros.

La investigación Modelo Predictivo Machine Learning aplicado al análisis de datos climáticos capturados por una placa Sparkfun (Iribarren, 2016), publicada por la Universidad Pontificia Comillas en Madrid, España, señala el objetico es crear un modelo capaz de proporcionar una predicción precisa sobre posibles retrasos o cancelaciones de vuelos debidos a las condiciones climáticas. Para ello buscaron captar con exactitud las condiciones meteorológicas relevantes para la predicción; transmitir y procesar los datos obtenidos para alimentar con ellos el modelo, configurando correctamente las conexiones necesarias entre el dispositivo y la plataforma en la nube; crear un modelo predictivo en la nube que sea preciso y robusto, y que permita obtener una predicción correcta en la mayoría de los casos y presentar los datos almacenados y los resultados obtenidos de una forma visual y sencilla de comprender. El autor concluye que se logró

con éxito captar datos meteorológicos en tiempo real, cumpliendo así el primer objetivo del proyecto. La frecuencia de captación de datos y la precisión de las medidas fueron excelentes. Se logró con éxito configurar correctamente la interfaz en la nube y la conexión del dispositivo con dicho entorno. La transmisión de datos dispositivo-nube ha sido satisfactoria, y además se integró la consolidación en la nube de los datos recibidos. Se pudo crear un modelo predictivo en la nube, completando todas las etapas habituales en este tipo de proyectos (tratamiento de los datos, elección del algoritmo y despliegue del modelo). El autor menciona que el análisis realizado al seleccionar el algoritmo se dio evaluando 4. El primero fue Multiclass decision jungle: la parte del modelo que evaluaba este algoritmo falló, por lo que no lo tuvieron en cuenta como opción, el segundo fue Multiclass logistic regression: este presentó un 85,65% de precisión media, pero el acierto por clases en los vuelos cancelados y retrasados fue muy bajo, el tercero fue Multiclass neural network: este presentó un 85,66% de precisión media, pero el acierto por clases en los vuelos cancelados y retrasados siguió siendo bajo, finalmente usaron Multiclass decision forest: este presentó un 85,82% de precisión media, y fue el algoritmo más acertado a la hora de predecir retraso o cancelación, por lo que fue el algoritmo que eligieron. Se terminó comprobando la gran complejidad del modelo, y se determinaron posibles mejoras a implementar para aumentar la precisión de los resultados a la hora de realizar la predicción. Finalmente, menciona que queda patente la gran complejidad de este tipo de modelos predictivos, no sólo por la cantidad de opciones a configurar durante la creación del mismo, sino también por la necesidad de corrección de errores durante el proceso, los tiempos de ejecución del modelo y que a la vista de los resultados, las condiciones meteorológicas son sólo uno de los factores que pueden causar retrasos y

cancelaciones en los vuelos, y esta es la principal causa de que el porcentaje de acierto en estos casos sea menor.

La investigación Generación de modelos predictivos de satisfacción transaccional para un centro de atención a clientes (Gonzáles, 2012) publicada por el Instituto Tecnológico y de estudios superiores de Monterrey en Atizapán de Zaragoza, México, surgió de la necesidad de predecir con un alto grado de exactitud qué transacciones en un centro de atención telefónica a clientes tienen una tendencia a ser consideradas una mala experiencia para el cliente, contribuyendo a que se sienta insatisfecho, y describir qué características de la transacción determinan esa tendencia. El trabajo se centró en el uso de minería de datos para la construcción de clasificadores, usando la técnica de árboles de decisión con el algoritmo C4.5. Este algoritmo fue seleccionado debido a su amplio uso en la literatura y a que el resultado es fácilmente interpretable en comparación con modelos como Redes neuronales o Máquinas de soporte de vectores. Mencionan que los resultados obtenidos fueron positivos, ya que se logró una precisión de más del 78% en la predicción de casos con tendencia a ser satisfactorios contra casos con tendencia a ser insatisfactorios. Adicionalmente el clasificador generado tiene un tamaño adecuado para ser interpretado por el personal de atención y directivos del centro de llamadas, ayudándoles a tomar decisiones que conduzcan a mejoras sustanciales en el servicio que proveen. Sin embargo, aunque se construyó un modelo con una precisión que cumple el objetivo planteado, recomiendan que es importante que se realicen entrenamientos posteriores con una mayor cantidad de datos, lo que puede conducir a la creación de un modelo con aún mejores características.

En la investigación Modelo Predictivo Para Intensidades Sísmicas Superficiales en Chile (Bastías, 2016) publicada por la Universidad de Concepción en Concepción, Chile, menciona que

Chile, al estar emplazado cerca de la zona de subducción de las placas de Nazca y Sudamericana, se encuentra en un ambiente sísmico activo, por lo cual, está propenso a sufrir los efectos de terremotos de mediana y gran magnitud. Por esto, los terremotos representan una gran amenaza para ese país. Debido a lo expuesto, y con la intención de reducir los niveles de daño provocados por los terremotos en las estructuras, consideran que es necesario mejorar la predicción de los parámetros de intensidad sísmica que dominan el diseño y el comportamiento estructural, como lo son las aceleraciones máximas del suelo y espectrales. El autor realizó el desarrollo de un modelo predictivo de intensidades sísmicas que buscó cuantificar la intensidad sísmica en superficie (e.g. PGA o aceleraciones espectrales) y su incertidumbre. Modelando el fenómeno a través de variables explicativas tales como: la magnitud, la distancia desde la fuente sísmica al sitio de estudio, el mecanismo de falla, el efecto de sitio, entre otras. En este trabajo se desarrolló una robusta base de datos de registros sísmicos chilenos, distribuidos entre los años 1985 hasta el 2015, incluyendo los terremotos de Valparaíso (7.9Mw), Maule (8.8Mw), Iquique (8.1Mw), Illapel (8.3Mw), entre otros. Procesó todo esto a través de un esquema estandarizado por componente, con el fin de homologar el nivel de ruido entre los distintos registros sísmicos. Finalmente, ajustó los datos a través de una regresión de un modelo no lineal de efectos mixtos. Sus conclusiones fueron que el desarrollo de estos modelos predictivos se vuelve relevante para el análisis de la peligrosidad sísmica, que tiene por objetivo determinar las cargas sísmicas de diseño para ciertos proyectos de Ingeniería Civil (centrales nucleares, hidroeléctricas, estructuras mineras). Además, permite el desarrollo de mapas de peligro sísmico, los cuales son una herramienta para la planificación demográfica de grandes urbes, y la consecuente asignación de recursos a zonas probablemente más

riesgosas frente la acción de un terremoto. Todo esto con el fin de reducir el riesgo a las personas y a las estructuras.

La investigación Diseño e Implementación de un Sistema de Visión Artificial para Clasificación de al menos Tres Tipos de Frutas (Constante & Gordón, 2015) publicada por la Escuela Politécnica Nacional en Quito, Ecuador, menciona que buscaron usar técnicas de visión artificial aplicadas a la detección de características en frutas las cuales pueden ser destinadas a la industria alimenticia; para ello se utiliza un sistema de visión por computador basado en redes neuronales artificiales organizadas en una arquitectura profunda; el sistema fue entrenado mediante aprendizaje compensado por ruido, la finalidad de proyecto fue crear una fuerte relación entre la red neural artificial y el objeto (fruta), que permite reconocer características complejas de frutas seleccionadas: fresas, moras y uvillas; se consideraron condiciones cambiantes tanto en la iluminación, tamaño, así como en la orientación. El sistema fue probado en tiempo real con imágenes reales. Concluyen que fueron capaces de implementar de manera satisfactoria la herramienta de clasificación de frutas y que la implementación de nuevas tecnologías tales como la visión artificial por computador y redes neuronales son posible en los procesos agrícolas dentro de su país, con el fin de optimizar ciertos parámetros de producción como tiempo, espacio, calidad, higiene. Esto ha ayudado a crear una mejor competitividad dentro del campo agrícola pues este sector depende mucho de la calidad de los productos para satisfacer la necesidad de sus clientes.

La investigación App para móviles de detección de características hortofrutícolas mediante tratamiento de imágenes (Pizarro, 2017) publicada por la Universidad de Extremadura en Extremadura, España, menciona el desarrollo de una aplicación para teléfonos Android. Dicha aplicación permite fotografiar piezas de fruta y analizar sus características propias. El método para

dicho análisis consta de analizar el color de la fruta a través de un colorímetro, e insertar dichos valores en formulas matemáticas para obtener los resultados requeridos. Para esto, se necesita disponer de dicho dispositivo, y de las habilidades para manejarlo y utilizar los datos obtenidos. Para facilitar y simplificar todo ese trabajo, la aplicación desarrollada consiste en una interfaz que, tras tomar una fotografía de la fruta, indicará la información que se desea obtener. Las investigaciones sobre modos de color y regresiones polinomiales fueron vitales para comprobar si el proyecto se podría llevar a cabo, ya que la conversión de una fotografía a valores de un colorímetro se mostraba difícil de aproximar. La facilidad de uso fue una las metas del proyecto, que intentó acercar y acelerar el proceso de análisis de parámetros de una fruta. La posibilidad de conseguir dichos análisis sin necesidad de recolectar la fruta ni impedir su crecimiento la hacen una aplicación llamativa e interesante para el usuario final. Se concluye mencionando que fue un proyecto complejo, y con muchos frentes abiertos, como son la aplicación Android, el análisis fotográfico de frutas, y los valores generados por un colorímetro, pero al final todo es transformable y adaptable con un ligero margen de error. Las configuraciones en la toma de fotografías aumentaron la aproximación, pero limita su ejecución a días soleados con un dispositivo determinado. La variedad de dispositivos Android y sus distintas cámaras y enfoques dificultó la obtención de una única formula de conversión. Aun así, afirmaron que los resultados fueron satisfactorios.

Por otro lado, está la investigación de un Clasificador de imágenes de frutas basado en inteligencia artificial (Heras, 2017) publicada por la Universidad Católica de Cuenca en Cuenca, Ecuador, donde se menciona que la clasificación de imágenes es muy útil en la automatización de procesos en una empresa. Para realizar una tarea de clasificación de imágenes se requiere hacer la

extracción de características que identifiquen a cada tipo de imagen como, por ejemplo: color, forma, textura. Es por ello, que en su investigación se realizó la implementación de los algoritmos para la construcción de un clasificador de imágenes de frutas basado en la extracción de las características del color de las imágenes en determinadas regiones de interés. Para el desarrollo del clasificador de imágenes de frutas se utilizó la técnica de extracción del histograma a color en tres dimensiones y con la implementación de algoritmos de inteligencia artificial se efectuó la clasificación automática de imágenes. El conjunto de datos que utilizó consiste en: cuatro clases de frutas con el número variable de imágenes por cada clase, luego se preparan las imágenes seleccionando las regiones de interés mediante técnicas de enmascaramiento y se las divide en dos grupos de datos: los datos de entrenamiento y los datos de prueba. Luego de entrenado el clasificador, se realizaron pruebas de clasificación para evaluar la eficacia del clasificador de imágenes de frutas. El autor concluye que esta metodología de construcción e implementación del clasificador se puede usar en varias aplicaciones según las clases de imágenes de objetos a analizar en casos similares y automatizar procesos de clasificación y reconocimiento de objetos.

En la investigación Identificación del estado de madurez de las frutas con redes neuronales artificiales, una revisión (Figueredo & Ballesteros, 2016) publicada por la Universidad Tecnológica y Pedagógica de Colombia en Boyacá, Colombia, enfatiza en que la aplicación de las Redes Neuronales Artificiales (RNA) y la visión artificial tiene cada vez más acogida en la industria de productos alimenticios; estas técnicas priorizan la clasificación, el reconocimiento de patrones, la predicción de las cosechas y de los cambios físicos de sus productos. Debido a la capacidad de las redes neuronales para aprender patrones de un conjunto de datos no lineales y con presencia de ruido, en los estudios mencionados en este artículo se presentaron en su

investigación una amalgama de arquitecturas, topologías y algoritmos de entrenamiento de redes neuronales como alternativas a la clasificación manual, los cuales ofrecieron, en muchos casos, una solución eficiente y efectiva. Se identificó que, en cada investigación considerada, los autores se centraron en la importancia de los datos de entrada, así como en su previo tratamiento, en la arquitectura y topología de la red y en la selección de los algoritmos de entrenamiento, asegurando que estos ítems proporcionan mejor o peor clasificación, según el manejo dado. Dentro de los estudios descritos, la red neuronal más utilizada fue la Backpropagation (BPNN), seguida de la de Boltzman (BPN) y la red Probabilística (PNN).

En el artículo titulado "Portugal usa un sistema de inteligencia artificial para incrementar las exportaciones" (Tecno, 2019) se menciona que la inteligencia artificial (IA) protagoniza la escena tecnológica contemporánea. Actualmente las personas llevan esos sistemas en el bolsillo, por ejemplo, en los asistentes virtuales de los celulares. Además, de los laboratorios de investigación surgen innovaciones basadas en IA con soluciones sin dudas atractivas en muchos campos. Uno de estos campos es el comercio internacional, como en Portugal, donde están usando sistemas de IA para impulsar las exportaciones. AICEP (Agência para o Investimento e Comércio Externo de Portugal), una agencia pública portuguesa dedicada a las inversiones y exportaciones, apuesta a la inteligencia artificial para ayudar que las empresas del país incrementen sus exportaciones. Según señalan en el sitio web, esas tecnologías se postulan como un posible salvavidas en un contexto de desaceleración de la economía global. La plataforma denominada "Portugal Exporta" incluye aprendizaje automático y análisis de datos para, en base a ellos, ofrecer servicios personalizados a las compañías exportadoras. Los ejes son la generación de vínculos entre compañías e inversores, el ofrecimiento de información sobre potenciales socios y respecto a planes de expansión

internacional, que se personalizan para cada caso. El artículo cita a Luis Castro Henriques, presidente de AICEP, que comenta "Este sistema nos permitirá atraer a más empresas, atender mejor sus requisitos y ser más productivos. Ciertamente, traerá resultados importantes para el crecimiento de las exportaciones.". Se concluye mencionando que más de la mitad de las empresas portuguesas con capacidad para exportar lo hacen con regularidad: 23.000 de un total de 44.000. Con la aplicación de sistemas de IA, la agencia lusa quiere que las exportaciones representen el 50% del PBI en el país hacia 2025, lo que implicaría un crecimiento del 6% respecto a 2018 y del 20% en relación con las cifras de 2010.

Finalmente, en el artículo Agronometrics: Precios y volúmenes de frutas de todo el mundo en un solo lugar y en el momento que se requiera (Simfruit, 2014) se menciona una entrevista realizada al CEO de Agronometrics, Colin Fain, que comenta "Agronometrics es una plataforma de inteligencia de mercado que recolecta, estandariza y presenta, diariamente, precios y volúmenes de frutas de variados países del mundo, permitiendo conocer de manera fácil y rápida la realidad de un mercado y un producto en específico. La información permite comparar datos y conocer cifras históricas, por ejemplo, por producto, orígenes y mercado, entre otras estadísticas que sean de interés, en ese momento, por el cliente". Colin Fain menciona que la idea de crear Agronometrics surgió luego de que este trabajara con exportadores y al identificar la complejidad de información que se requería para tomar decisiones apropiadas, decidió crear esta plataforma. La información es actualizada diariamente, así como también enriquecida. Por ejemplo, constantemente va introduciendo nuevas frutas y países, y actualmente, está trabajando en incorporar cerca de veinte nuevas fuentes de información, lo que le permitirá entregar un mejor servicio.

Para comprender mejor la investigación, es importante conocer de manera más detallada ciertas definiciones. A continuación, se profundizará en lo que es un modelado predictivo, tipos de modelado, sus ventajas y limitaciones, la relación del modelo predictivo con el análisis de datos y cual es actualmente el estado del análisis de datos en el sector exportador.

"El modelado predictivo es un sistema que emplea datos y estadísticas para predecir resultados a partir de unos modelos de datos. Estos modelos se pueden utilizar para predicciones de todo tipo; desde resultados deportivos y audiencias televisivas hasta avances tecnológicos y ganancias empresariales" (Microstrategy, 2020).

El modelado predictivo se suele conocer también como análisis predictivos, analítica predictiva o aprendizaje automático, estos sinónimos suelen ser intercambiables. Sin embargo, el análisis predictivo se refiere casi siempre a las aplicaciones comerciales del modelado predictivo, mientras que este último se usa de manera más general o académica. En esta investigación, se usó principalmente el término "modelado predictivo", pero los términos modelado predictivo, análisis predictivo y analítica predictiva se pueden usar de manera intercambiable (Microstrategy, 2020).

El modelado predictivo es útil porque proporciona información precisa sobre cualquier pregunta y permite a los usuarios crear previsiones. Para mantener una ventaja competitiva es fundamental tener información detallada de los eventos y resultados futuros que desafíen las presuposiciones (Microstrategy, 2020).

Los profesionales del análisis suelen extraer datos de diversas fuentes para alimentar sus modelos predictivos. Por ejemplo, datos sobre transacciones, datos de servicio al cliente, datos de encuestas o sondeos, datos de marketing digital y publicidad, datos económicos, datos

demográficos, datos generados por máquinas (por ejemplo, datos telemétricos o datos de sensores), datos geográficos, datos de tráfico web, etc (Microstrategy, 2020).

Los líderes de análisis deben alinear las iniciativas de modelado predictivo con los objetivos estratégicos de la empresa o proyecto donde este será utilizado. Por ejemplo, un fabricante de chips informáticos podría establecer como prioridad estratégica producir chips con el mayor número de transistores del sector de aquí a 2025. Los profesionales del análisis podrían crear un modelo predictivo que pronosticará el número necesario de transistores por chip para convertirse en líderes. Para ello, se cargan en el modelo datos de producto, geográficos, ventas y otros datos relacionados con las tendencias. Como fuentes adicionales se incluyen datos sobre los chips con mayor densidad de transistores, la demanda comercial de capacidad de computación y las alianzas estratégicas entre fabricantes de chips y fabricantes de hardware. Una vez puesta en marcha la iniciativa, los analistas pueden realizar análisis retrospectivos para evaluar la precisión de los modelos predictivos y el éxito de dicha iniciativa. (Microstrategy, 2020)

Los analistas deben organizar los datos con el fin de alinearlos a un modelo. Así, es posible crear informáticamente previsiones y resultados de las pruebas de hipótesis. Las herramientas de inteligencia de negocios proporcionan información detallada en forma de paneles, visualizaciones e informes. Es necesario establecer un proceso que garantice una mejora continua. Aspectos importantes a tener en cuenta para la integración de modelos predictivos en la práctica son: Análisis de referencia, recopilación de datos, limpieza de datos, análisis (en general), evaluación de objetivos, creación de planes de acción basados en los análisis, ejecución de planes, optimización de procesos,etc. (Microstrategy, 2020)

Los tipos más comunes de modelos predictivos son los de: Regresión, que consiste en predecir una respuesta cuantificable. Este tipo de modelos abordan cuestiones como la cantidad de unidades de un producto vendidas, el precio de mercado o el retorno de la inversión. Y el de: Clasificación, estos modelos predictivos pronostican una respuesta categórica que responde a una cuestión abierta, como la probabilidad de que un consumidor se convierta en cliente, la existencia de intención de fraude en una transacción o la marca que resultará más demandada en el plazo de un año. (Logicalis, 2015)

En términos más específicos, algunos modelos predictivos conocidos son: mínimos cuadrados ordinarios, GLM (Modelos lineales generalizados), regresión logística, bosques aleatorios, árboles de decisión, redes neuronales , y otros. (Microstrategy, 2020)

Cada uno de éstos tiene un uso particular y responde a una pregunta específica o utiliza un determinado tipo de conjunto de datos. A pesar de las diferencias metodológicas y matemáticas entre los tipos de modelos, el objetivo general de todos ellos es similar: predecir resultados futuros o desconocidos basándose en datos pasados. (Microstrategy, 2020)

A grandes rasgos, el modelado predictivo reduce significativamente los costes de previsión de resultados empresariales, factores ambientales, inteligencia competitiva y condiciones del mercado. El modelado predictivo puede aportar valor de muchas formas, como: pronóstico de la demanda, planificación y análisis de pérdida de plantilla, predicción de factores externos, análisis de la competencia, mantenimiento de flotas o equipos, simulación de crédito u otros riesgos financieros y otros. (Microstrategy, 2020)

A pesar de sus numerosos y valiosos beneficios, es cierto que el modelado predictivo tiene sus limitaciones. A menos que se cumplan ciertas condiciones, el modelado predictivo no puede

alcanzar todo su potencial. De hecho, si no se dan estas condiciones, los modelos predictivos pueden no proporcionar ningún valor respecto a los antiguos métodos o conocimientos convencionales. Entre algunas limitaciones están: el etiquetado de datos, especialmente en aprendizaje automático, donde un ordenador construye un modelo predictivo, los datos se deben etiquetar y clasificar de manera adecuada. Este proceso puede ser impreciso, estar lleno de errores y convertirse en una actividad colosal. (Microstrategy, 2020)

Otra limitación sería la obtención de enormes conjuntos de datos para entrenamiento, para que los métodos estadísticos tengan éxito en la predicción de resultados, se debe cumplir con un principio básico: que la muestra sea suficientemente grande. Si un profesional del modelado predictivo no tiene suficientes datos para construir el modelo, éste será sin duda deficiente. Por supuesto, los conjuntos de datos relativamente pequeños tienden a mostrar más variaciones o, en otras palabras, más ruido. En la actualidad, la cantidad de registros necesarios para alcanzar un buen rendimiento oscila de miles a millones de datos. Además del tamaño, los datos utilizados deben ser representativos de la población. Si la muestra es lo suficientemente grande, los datos deben tener variedad de registros, incluidos los casos únicos o infrecuentes, para perfeccionar el modelo en la medida de lo posible. (Microstrategy, 2020)

Finalmente se puede considerar la generalización del aprendizaje, esta se refiere a la capacidad del modelo para aplicarse de un caso de uso a otro. A diferencia del ser humano, los modelos tienden a luchar con la generalización, también conocida como validez externa. En general, cuando un modelo se construye para un caso en particular, no debe usarse para otro diferente. Si bien se están desarrollando métodos como el aprendizaje por transferencia, un

enfoque que intenta remediar este problema, la generalización sigue siendo una limitación importante del modelado predictivo. (Microstrategy, 2020)

El análisis de datos es una técnica y por medio de ésta se inspeccionan, purifican y transforman datos, con la finalidad de destacar toda la información que sea de gran utilidad, a fin de poder elaborar conclusiones que sirvan de apoyo en la toma de decisiones. (Conceptodefinicion, 2019)

El análisis de datos se distingue de la extracción de datos por su alcance, su propósito y su enfoque sobre el análisis. Los extractores de datos clasifican inmensos conjuntos de datos usando software sofisticado para identificar patrones no descubiertos y establecer relaciones escondidas. El análisis de datos se centra en la inferencia, el proceso de derivar una conclusión basándose solamente en lo que conoce el investigador (Rouse, 2012).

Éste generalmente se divide en análisis exploratorio de datos (EDA), donde se descubren nuevas características en los datos, y en análisis confirmatorio de datos (CDA), donde se prueba si las hipótesis existentes son verdaderas o falsas. El análisis cuantitativo de datos (QDA) es usado en las ciencias sociales para sacar conclusiones de datos no numéricos, como palabras, fotografías o videos. En la tecnología de la información, el término tiene un significado especial en el contexto de las auditorias informáticas, cuando se examinan los sistemas, operaciones y controles de los sistemas de la información de una organización. El análisis de datos se usa para determinar si el sistema existente protege los datos efectivamente, opera eficientemente y cumple con las metas de la organización (Rouse, 2012).

Los modelos predictivos, para poder llevar a cabo su misión, requieren de predictores y de la observación de los conjuntos de datos. Es aquí donde entra a tallar el análisis. A mayor número

de predictores y mayor profundidad en su investigación, aumentará la complejidad del análisis. Aunque éste no es el reto más complicado. El verdadero desafío para los modelos predictivos es encontrar buenos subconjuntos de predictores o variables explicativas, es decir, hallar los que mayor utilidad aportan y los que mejor se ajustan a los datos. Al considerar los problemas de negocios, se utilizan los datos disponibles para predecir los datos que todavía no se tienen. Se trata de un proceso de extrapolar y predecir, con los riesgos que ello implica. Por eso, es importante tener en cuenta que los mejores modelos predictivos, los de mayor valor son los que aportan predicciones de más calidad (Logicalis, 2015).

Según un artículo de Emprendedores (2015), antes de dar el salto al exterior, toda empresa necesita conocer en profundidad el mercado que desea abordar. Para garantizar el éxito de su producto o servicio en un país concreto, la compañía debe disponer de información precisa y actualizada del mercado en el que actúa y valorar la inversión que está dispuesta a realizar para obtener la información que necesita. Existen fuentes primarias (que facilitan información adaptada a los objetivos que se persiguen, mediante entrevistas en profundidad con compradores potenciales) y secundarias (disponible en publicaciones y bases de datos públicas, aunque menos adaptadas a las necesidades de la empresa). En todo caso, los elementos que nunca deben faltar en una investigación de mercados exteriores son los siguientes:

En entorno internacional, donde las variables más significativas de los ámbitos económico (PIB, renta per cápita, tipos de interés,etc), comercial (estadísticas de comercio exterior), político (datos demográficos, geográficos, infraestructuras, riesgos y estabilidad política), cultural (idiomas, usos y costumbres, actitudes, preferencias) y legal (aranceles, licencias de importación

o exportación, impuestos, homologaciones y certificaciones, normas sanitarias, control de cambios, paquetes y marcas, etc) con las que la empresa va a operar (Emprendedores, 2015).

La demanda, que consiste en analizar, cuantitativa (análisis por subsectores, regiones, áreas geográficas, habitante y año, porcentaje y año) y cualitativamente (tipología del comprador, motivaciones de compra, hábitos y ritmos de consumo, preferencias de calidad, precio o segmentos), la demanda real y potencial de cada mercado (Emprendedores, 2015).

La competencia, pues en un mercado global y competitivo como el actual es preciso conocer la oferta de otras empresas del sector para descubrir posibles nichos de mercado. Para ello es necesario conocer estructura, situación y perspectivas de la industria local; principales fabricantes nacionales; volumen, origen y cuota de mercado de las importaciones; fabricantes extranjeros; segmentos de mercado cubiertos por la competencia y ranking de cuotas de mercado y zonas geográficas (Emprendedores, 2015).

Estructura de mercado, donde se debe tomar en cuenta los precios de la competencia, márgenes comerciales, costes de transporte, almacenamiento y distribución; canales de distribución, técnicas de promoción o cobertura de medios publicitarios (Emprendedores, 2015).

1.2. Formulación del problema

¿Cómo un modelo predictivo mejora el análisis de datos del sector exportador hortofrutícola en el periodo 2016-2020?

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un modelo predictivo que mejore el análisis de datos del sector exportador hortofrutícola en el periodo 2016-2020.

1.3.2. Objetivos específicos

- Realizar un análisis de las fuentes y reportes de datos de valor en el sector hortofrutícola, para el uso de la información en el desarrollo del modelo predictivo.
- Establecer un formato de reporte estándar tomando en cuenta los datos de mayor importancia y su calidad, para el correcto procesamiento de los datos.
- Identificar un algoritmo que permita realizar predicciones con los datos más relevantes, para determinar el aumento o disminución del precio de los productos del sector hortofrutícola con un buen nivel de precisión y eficacia.
- Implementar una solución tecnológica que permita el correcto procesamiento y visualización de los datos.

1.4. Hipótesis

El desarrollo de un modelo predictivo utilizando la metodología CRISP-DM mejora el análisis de datos del sector exportador hortofrutícola en el periodo 2016-2020.

CAPÍTULO II. METODOLOGÍA

2.1. Tipo de investigación

La investigación fue aplicada porque busca la generación de conocimiento con aplicación directa a los problemas de la sociedad o el sector (Lozada, 2014). En este caso el modelo predictivo está enfocado en el sector exportador. Esta se basa fundamentalmente en el uso de herramientas tecnológicas y avances metodológicos que permitan mejorar el problema presente.

Así mismo, el diseño de investigación fue cuasiexperimental. Los diseños cuasiexperimentales manipulan deliberadamente, al menos, una variable independiente para observar su efecto sobre una o más variables dependientes, sólo que difieren de los experimentos "puros" en el grado de seguridad que pueda tenerse sobre la equivalencia inicial. (Hernandez et al., 2014).

Lo que se busca establecer en esta investigación es si existe una mejora del análisis de datos del sector hortofrutícola al implementar el modelo predictivo (Ver Anexo Nº 1. Operacionalización de variables).

2.2. Población y muestra (Materiales, instrumentos y métodos)

2.2.1. Población

La población consiste en 21416 reportes de exportación del periodo 2016-2020 del sector hortofrutícola de la USDA (United States Department of Agriculture). (ver Anexo Nº 2)

2.2.2. Muestra

La muestra serán los reportes del producto palta dividido en valores diarios exceptuando sábado y domingo (no existe data en este periodo) del año 2019 (250 reportes).(ver Anexo Nº 3)

2.3. Técnicas e instrumentos de recolección y análisis de datos

2.3.1. Descripción de la técnica para la recolección de datos

La técnica utilizada para la recolección de datos fue la revisión documentaria, según Hurtado (2012) es una técnica en donde se recolecta información sobre un determinado tema, teniendo como fin proporcionar variables que se relacionan indirecta o directamente con el tema establecido, vinculando esta relaciones, posturas o etapas, en donde se observe el estado actual de conocimiento sobre ese fenómeno o problemática existente, en este caso se realizó la revisión del documentaria de producto palta (Avocado).

En la elección del instrumento según Castro (2010) se debe complementar el estudio a través de la utilización de fuentes primarias y secundarias como textos, manuales, folletos revistas,

Internet entre otros recursos. En este caso se utilizarán reportes de datos obtenidos, de fuentes validadas (USDA), de manera online (el modelo de reporte se puede ver en el Anexo Nº 11).

2.3.2. Confiabilidad y Validez de los instrumentos

El formato de reporte que se utilizará es el de Shipping Point, la confiabilidad del mismo fue determinado por un informe de selección de reportes (Anexo Nº 9). Se tuvo como prioridad la cantidad de datos presentes, la facilidad de interpretación, su accesibilidad, y consistencia como puntos principales para la elección.

La validación del instrumento fue realizada por un experto en el sector exportador y uno en desarrollo de modelos predictivos. Se buscó cumplir la mayoría de puntos del formato de validación y se tomaron en cuenta algunas observaciones brindadas por los expertos (Anexo Nº 4).

2.4. Aspectos Éticos

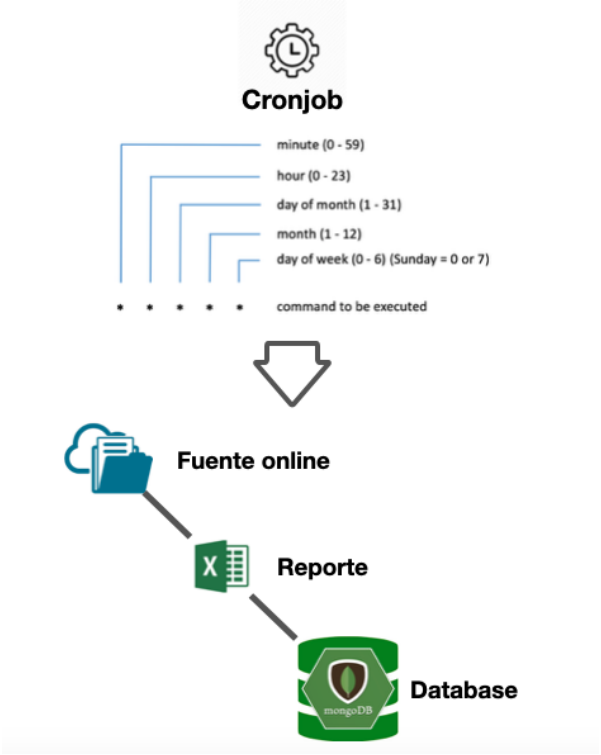
El presente estudio fue realizado por Bachilleres en Ingeniería de Sistemas Computacionales, que cuentan con la experiencia necesaria para realizar el análisis en el sector hortofrutícola. La investigación se basó en estudios de desarrollo de modelos predictivos en otros sectores. Se ha respetado las condiciones éticas de la recolección de datos, estos fueron sacados de los reportes públicos de la USDA (United States Department of Agriculture) e información de la empresa, para lo cual se contó con los permisos necesarios (Anexo Nº 5). Así mismo, se solicitó el permiso correspondiente a la empresa para poder mostrar los resultados más relevantes del proyecto.

2.5. Procedimiento

Para la recopilación de los datos, se siguieron los siguientes pasos:

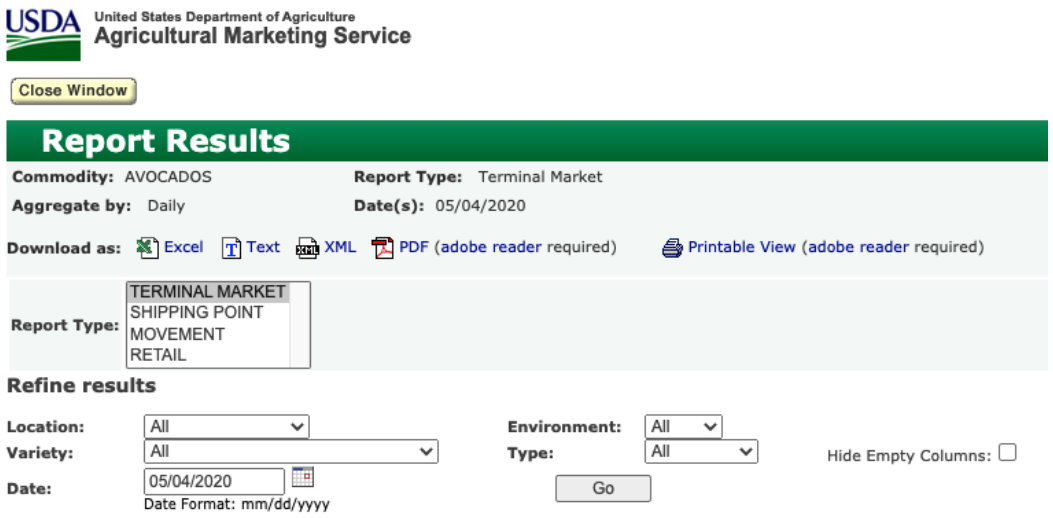
- Se utilizó un Cron Job (una tarea cronometrada) personalizado, que permita descargar los reportes de la fuente online de manera diaria, como se puede ver en la Figura 1:

Figura 1. *Proceso de Selección de datos.*



- Se tomó como fuente a la USDA, que nos permite tener acceso a diversos reportes oficiales en diversos formatos (en este estudio el formato elegido será en Excel) de manera gratuita diariamente, como se ve en la Figura 2:

Figura 2. *Resultados de reportes USDA.*



Nota: Tomado de https://www.marketnews.usda.gov (USDA, 2020).

- Luego que los reportes son descargados, son almacenados en una base de datos MongoDB, en el formato original brindado por la USDA (ver la Figura 3). En este reporte los datos más relevantes son: la fecha, la variedad, el origen y el precio del producto, representados por las columnas Date, Variety, Origin y Price respectivamente.

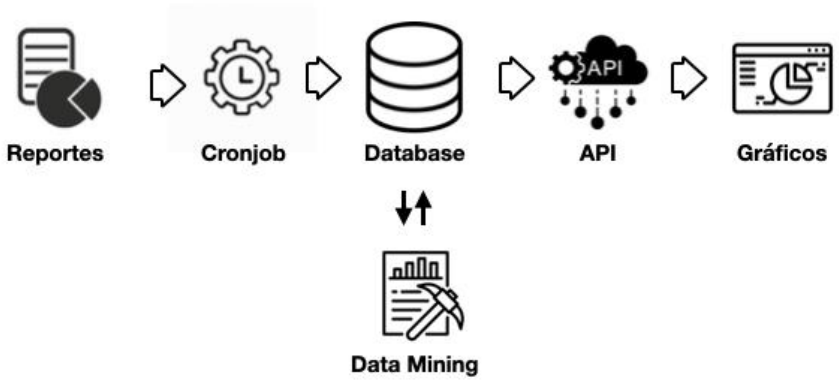
Figura 3. *Ejemplo de Reporte Original.*

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Commodity Name	City Name	Type	Package	Variety	Sub Variety	Grade	Date	Low Price	High Price	Mostly Low	Mostly High	Origin
2	AVOCADOS	ATLANTA		cartons 2 layer	HASS			5/04/20	51.5	53			CALIFORNIA
3	AVOCADOS	ATLANTA		cartons 2 layer	HASS			5/04/20	48	53.5	49	51.5	CALIFORNIA
4	AVOCADOS	ATLANTA		cartons 2 layer	HASS			5/04/20	45	51	48.5	48.5	MEXICO
5	AVOCADOS	ATLANTA		cartons 2 layer	HASS			5/04/20	44	51	45.5	45.5	MEXICO
6	AVOCADOS	ATLANTA		cartons 2 layer	VARIOUS GREENSKIN VARIETIES			5/04/20	27.5	27.5			FLORIDA
7	AVOCADOS	BALTIMORE		cartons 2 layer	HASS			5/04/20	47	49	47	48	MEXICO
8	AVOCADOS	BALTIMORE		cartons 2 layer	HASS			5/04/20	47	49	47	48	MEXICO
9	AVOCADOS	BALTIMORE		cartons 2 layer	HASS			5/04/20	46	46			MEXICO
10	AVOCADOS	BALTIMORE		cartons 2 layer	HASS			5/04/20	40	43			MEXICO
11	AVOCADOS	BALTIMORE		cartons 2 layer	HASS			5/04/20	44	48	44	45	MEXICO

Tomado de https://www.marketnews.usda.gov (USDA, 2020).

- Finalmente, con esta recopilación de datos, se inicia el proceso de modelado de data mining, que se da como se muestra en la Figura 4:

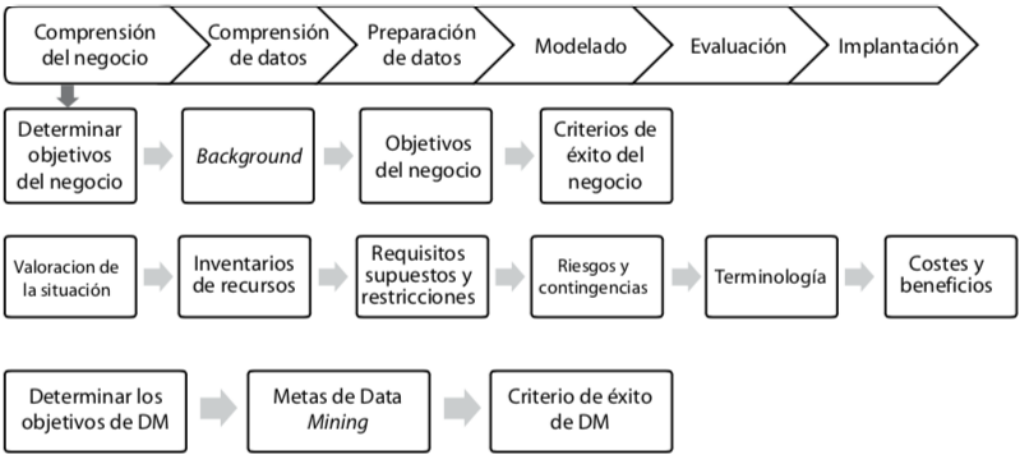
Figura 4. *Arquitectura Data Mining.*



Para el desarrollo del modelo predictivo se siguió la metodología CRISP-DM (el detalle metodológico es expandido en el marco conceptual presente en el Anexo Nº 6) y se trabajaron las siguientes fases:

- Fase 1: Comprensión del negocio o problema. Representada en la Figura 5.

Figura 5. *Fase 1: Comprensión del negocio o problema.*

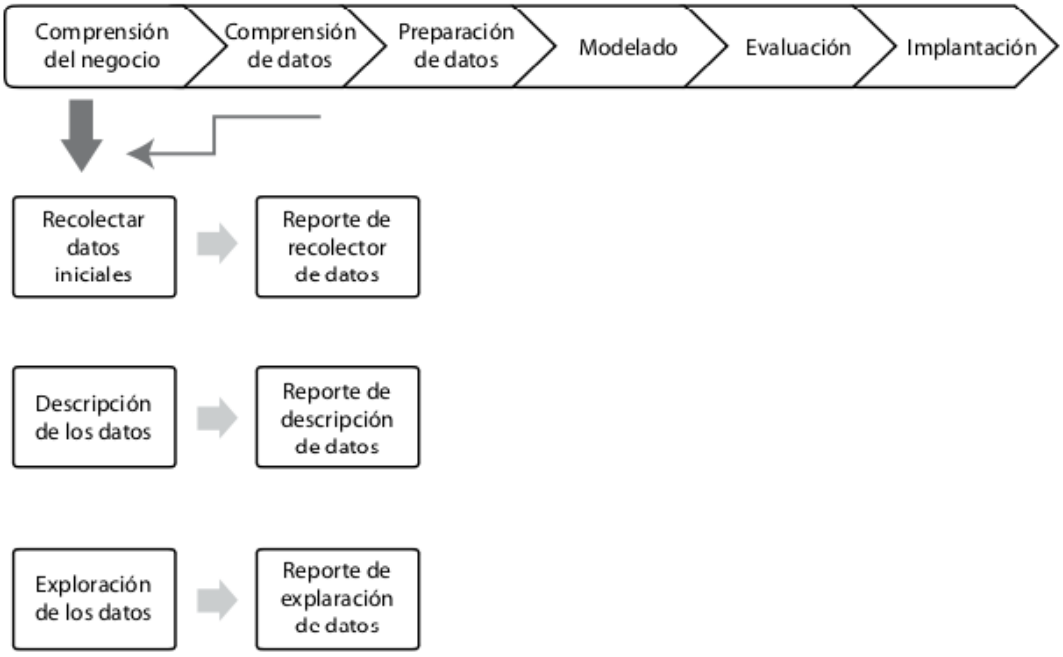


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, se realizó un análisis global y detallado del sector hortofrutícola, también una valoración de la información relevante y la forma en la que será utilizada.

- Fase 2: Comprensión de los datos. Representada en la Figura 6.

Figura 6. *Fase 2: Comprensión de los datos.*

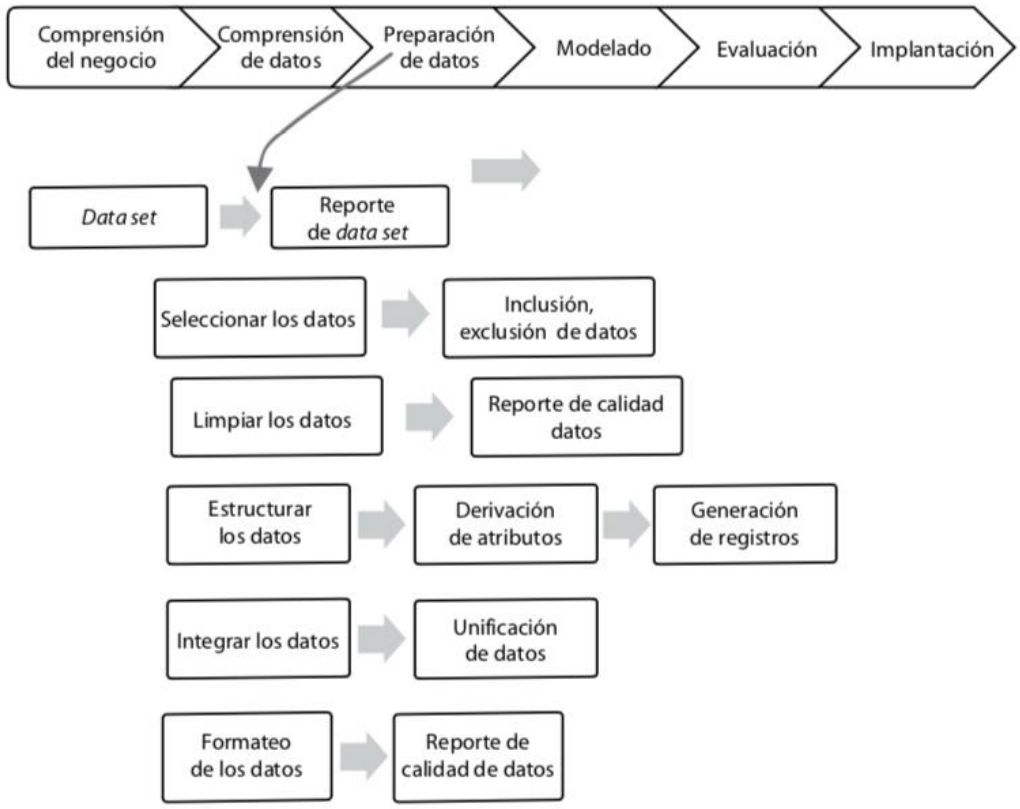


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, se recolectaron los datos iniciales, se produjo el entendimiento de los datos y el alcance de lo que se puede hacer con los mismos.

- Fase 3: Preparación de los datos. Representada en la Figura 7.

Figura 7. *Fase 3: Preparación de los datos.*

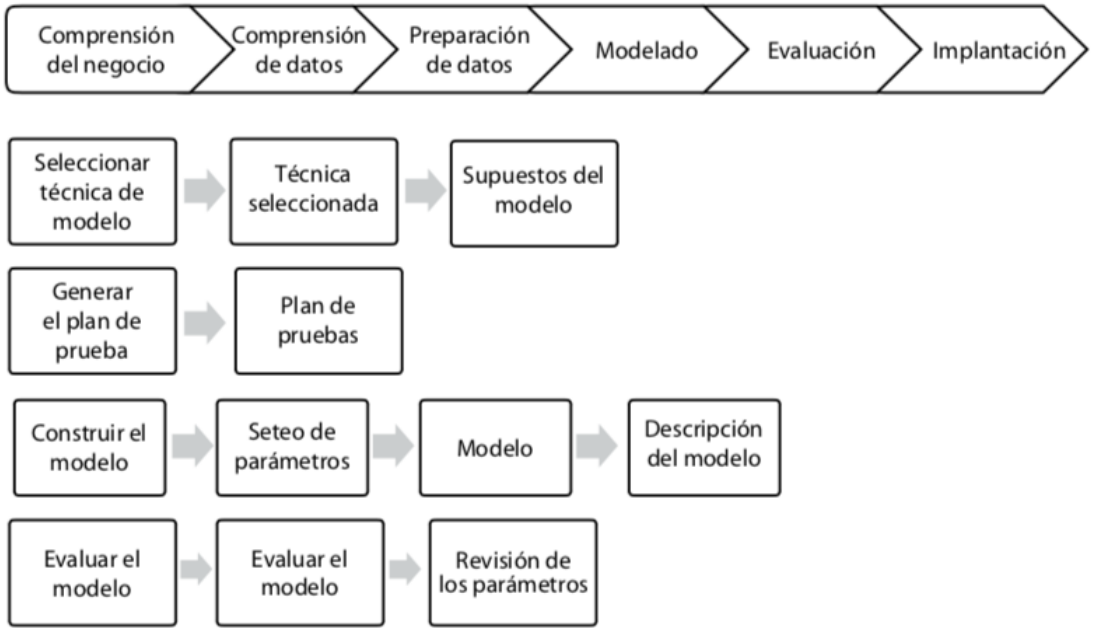


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, los datos pasaron por un proceso de limpieza, selección, estructurado e integración, finalmente se verificó que el resultado de todos los procesos derive en una mejora en la calidad de los datos.

- Fase 4: Modelado. Representada en la Figura 8.

Figura 8. *Fase 4: Modelado.*

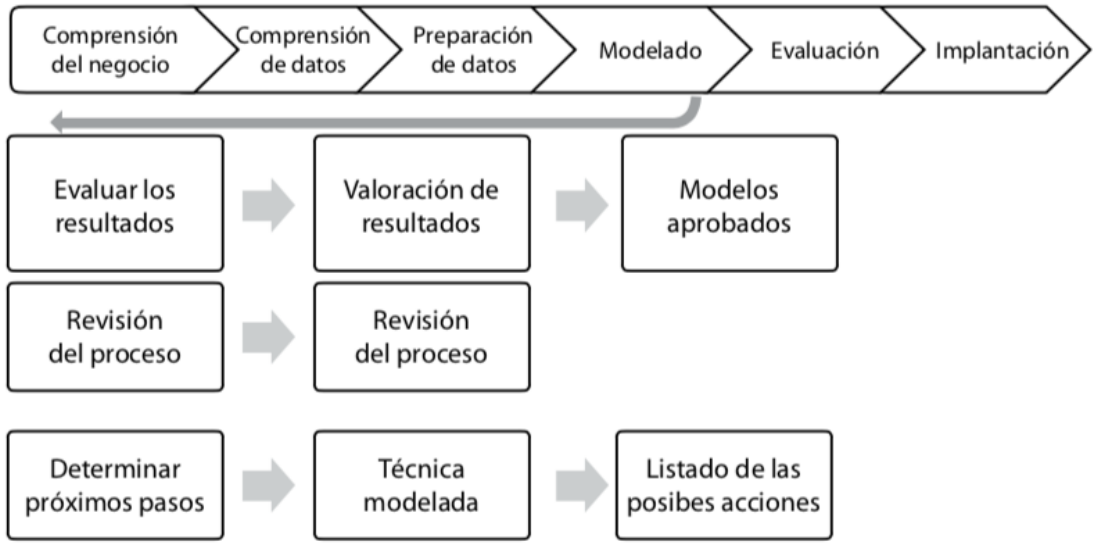


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, se seleccionaron algoritmos y un plan de cómo utilizarlos para construir el modelo, se construyó el modelo y se evaluó cada uno de los algoritmos elegidos.

- Fase 5: Evaluación. Representada en la Figura 9.

Figura 9. *Fase 5: Evaluación.*

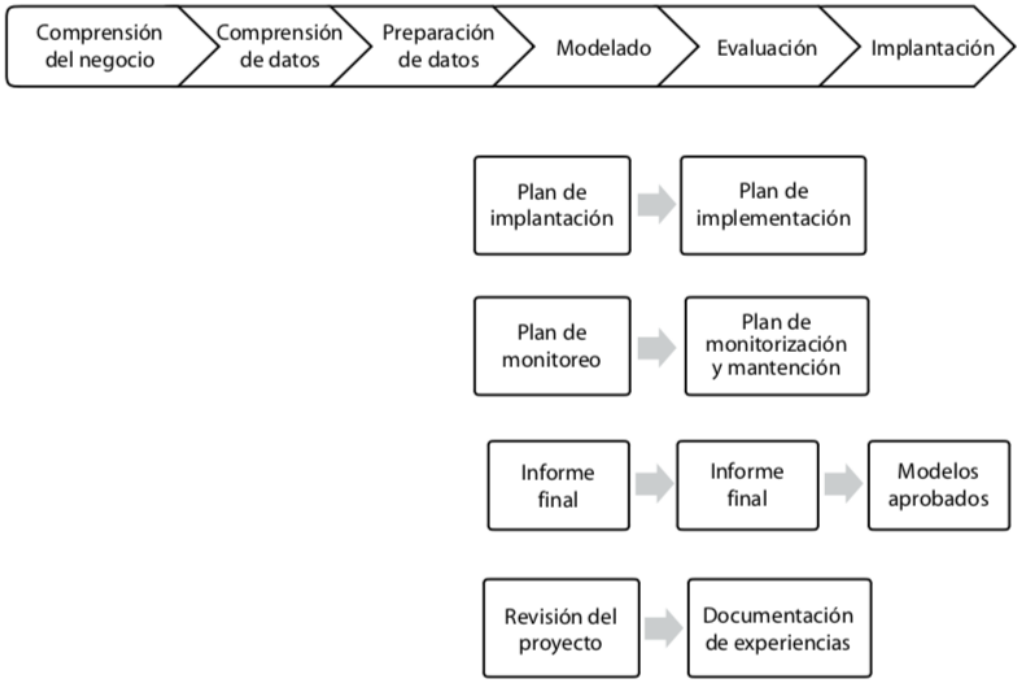


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, se realizó una comparación de acuerdo con el criterio de éxito de la investigación.

- Fase 6: Implementación. Representada en la Figura 10.

Figura 10. *Fase 6: Implementación.*

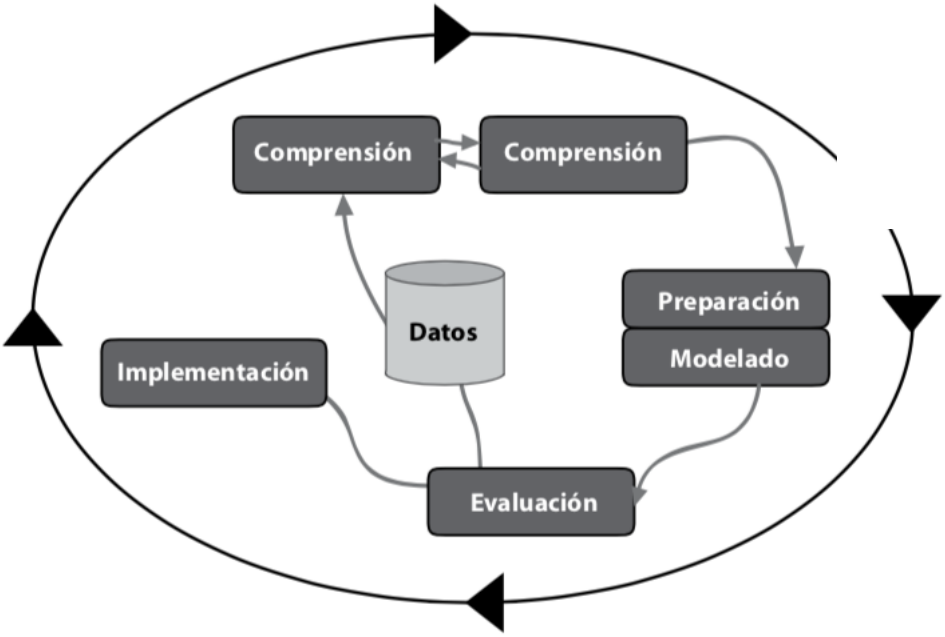


Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

En esta fase, se implementó el modelo acorde a la tecnología actual siguiendo los pasos planificados.

Estas fases ayudan a entender el proceso y proveen de un camino a seguir. De la fase 1 a la fase 5 el proceso es iterativo, hasta que sea decidido el pase a implementación como se puede ver el la Figura 11, que representa el ciclo de vida de la metodología .

Figura 11. *Ciclo de Vida de CRISP-M.*



Nota: Tomado de Crisp-dm 1.0 Step-by-Step Data Mining Guide (Chapman et al., 2000).

CAPÍTULO III. RESULTADOS

Para el desarrollo del proyecto se usaron las pautas brindadas por la metodología CRISP-DM.

3.1. Resultados según Objetivos

3.1.1. Objetivo específico 1:

Realizar un análisis de las fuentes y reportes de datos de valor en el sector hortofrutícola, para el uso de la información en el desarrollo del modelo predictivo.

Este objetivo se centró en las fases 1 y 2 de la metodología CRISP-DM. La fase 1 consiste en la comprensión del problema, estas se divide en los siguientes pasos:

- Análisis preliminar.
- Valoración Comparativa.
- Objetivos de minería de datos.

Se empezó trabajando en un análisis preliminar del sector hortofrutícola (Anexo Nº 7). El análisis se dio tomando en cuenta Latinoamérica, Europa, Asia y Estados Unidos. Basandose en dicho análisis, se realizó una valoración comparativa como se muestra en la Tabla 1:

Tabla 1. *Comparación de reportes según países.*

	SUNAT	LATIN AMERICAN FRUITS	CE	cirad	USDA
	Perú	Latinoamérica	China y Asia	Europa	USA

Cantidad de datos	√	√	√	√	√
Facilidad de interpretación	√	√			√
Accesibilidad	√				√
Procesamiento y filtrado				√	√

Tomando en cuenta la comparación, se identifica a la USDA como la opción que cumple con la mayoría de requisitos, es por esto que se usó como la fuente de información del modelo.

Los datos brindados por la USDA son variados con respecto a productos, según Workman (2020), en la Tabla 2 se tienen los 3 primeros productos que más se exportan.

Tabla 2. *Top 3 de frutas que más se exportan.*

1	Avocados	US$2.9 billion (17.5% of US-imported fruits)
2	Bananas Not Plantains	$2.5 billion (15.5%)
3	Fresh Grapes	$2.5 billion (15.5%)

En este caso, se elige a la palta (Avocado) como producto para realizar el modelo, por ocupar el primer puesto.

Por otro lado, la metodología CRISP-DM recomienda también evaluar los objetivos de la minería de datos, investigando sobre la misma. En este apartado se realizó un informe con las técnicas más usadas (Anexo Nº 8). Estas son representadas en la Tabla 3:

Tabla 3. *Relación de técnicas de minería de datos.*

Objetivo	Técnica de Minería de Datos
Asociación	Reglas de asociación
Clustering	Detección de cluster.
	Mapas autoorganizados.
Clasificación	Reglas de clasificación
	Árboles de decisión
	K-NN (K-Nearest Neighbor)
	Redes neuronales
Estimación	Árboles de decisión
	Modelos de regresión
	Redes neuronales
	Análisis de supervivencia
Predicción	Árboles de decisión
	Modelos de regresión
	Redes neuronales
	Series temporales
	Redes bayesianas
Descripción	Redes bayesianas
	Reglas de asociación
Explicación	Redes bayesianas

Nota: Tomado de Objetivos de un proyecto de data mining (Vallalta, 2020) .

Estas técnicas se tendrán en cuenta más adelante, cuando se defina cuales son las más convenientes para usar en el modelo.

Se continuó, aplicando la fase 2 de la metodología, la comprensión de los datos. Esta consiste en:

- Recolección datos iniciales.
- Descripción de los datos.
- Exploración de los datos.

Se fijó como objetivo analizar los datos específicos y determinar la forma en la que estos serán representados. Se inició, realizando una selección de los reportes brindados por la USDA, la selección de los mismos se dió a través de un informe de selección de reportes (Anexo Nº9), donde se terminó concluyendo que el reporte de Shipping point, era el más adecuado para el estudio.

El reporte de Shipping point cuenta con 12 campos utilizables, los cuales son descritos a más detalle en un reporte de descripción de los datos (Anexo Nº 10), a continuación en la Tabla 4 se muestra el formato inicial del reporte:

Tabla 4. *Formato inicial de Reporte.*

Commodity Name	City Name	Type	Package	Variety	Sub Variety	Grade	Date	Low Price	High Price	Mostly Low	Mostly High
AVOCADO	MEXICO CROSSINGS THROUGH TEXAS		cartons 2 layer	HASS			9/03/20	40.25	43.25	40.25	42.25

Al explorar estos datos, se lograron identificar un par de coincidencias:

Una de ellas es que la variedad Hass siempre tiene mayor precio que Green Skin, como se puede ver en la Figura 12. Y la otra coincidencia encontrada, es que el origen de la variedad Hass es siempre de México y el de Green Skin generalmente de Florida, como se puede ver comparando la Figura 12 y la Figura 13.

Figura 12. *Gráfico de lineas de Precio por Variedad.*

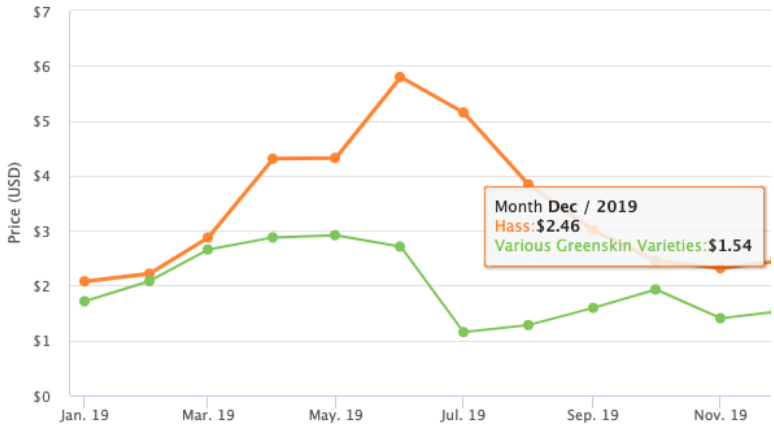
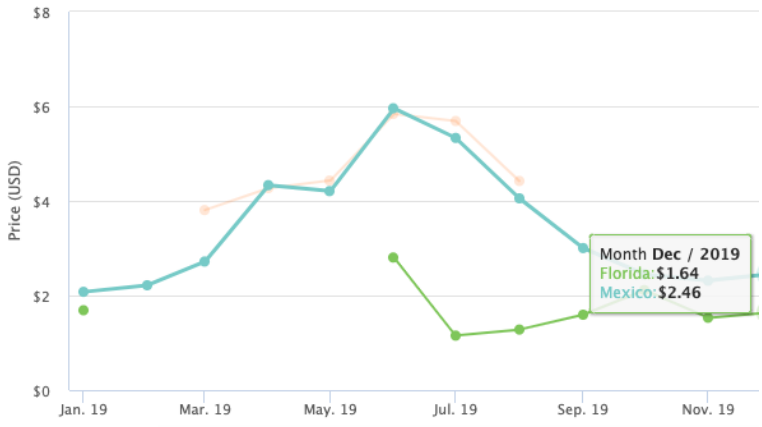


Figura 13*. Gráfico de lineas de Precio por Origen.*



Una vez finalizada la exploración de los datos, donde se encontraron coincidencias entre el Origen, Variedad y Precio, se pasa a la fase 3.

3.1.2. Objetivo específico 2:

Establecer un formato de reporte estándar tomando en cuenta los datos de mayor importancia y su calidad, para el correcto procesamiento de los datos.

Este objetivo se centró en la fase 3 de la metodología CRISP-DM: la preparación de los datos. Los pasos de la misma consisten en:

- Reporte a utilizar.
- Realizar una limpieza de datos.
- Selección, estructurado e integración de los datos.
- Reporte de calidad.

Se trabajó con los datos del reporte estándar de Shipping Point (ver Anexo Nº 11). En una primera instancia, para la limpieza de los datos, se realizó una análisis de la homogeneidad de los mismos, a través del gráfico Linear Projection (ver Anexo Nº 12) y se identificó que existe la presencia de datos confusos o faltantes, por lo que se planteó removerlos. Esto se dió en un proceso de selección, estructurado e integración de los datos, lo que permitió transformar el reporte original a un nuevo reporte con los datos más estables y de mayor valor (ver Anexo Nº 13). El formato inicial del precio (presente en la Tabla 5), se transformó en una versión simplificada del mismo, como se puede ver en la Tabla 6.

Tabla 5. *Reporte Estándar de Precio.*

Commodity Name	City Name	Package	Variety	Date	Low Price	High Price	Mostly Low	Mostly High
AVOCADO	MEXICO CROSSINGS THROUGH TEXAS	cartons 2 layer	HASS	9/03/20	40.25	43.25	40.25	42.25

Tabla 6. *Formato final.*

Origin	Variety	Date	Price Kg
MEXICO	HASS	9/03/20	3.58

Para finalizar, se realizó un segundo análisis de calidad de datos utilizando Linear Projection (Anexo Nº 14), donde se terminó identificando una mejora en la homogeneidad de los datos.

3.1.3. Objetivo específico 3:

Identificar un algoritmo que permita realizar predicciones relacionadas con los datos más relevantes, para determinar el aumento o disminución del precio de los productos del sector hortofrutícola con un buen nivel de precisión y eficacia.

Este objetivo empezó siguiendo la fase 4 de la metodología CRISP-DM , que es la de modelado. Esta fase consiste en:

- Selección de técnica.
- Plan de pruebas .
- Construcción de modelo .
- Evaluar el modelo .

Una vez hecha la investigación de los objetivos de minería en la fase 1, se empezó la fase 4 con una etapa de selección de técnicas para el modelo donde, de todo el rango de algoritmos existentes, se identificó cuales son lo que tienen más presencia en la actualidad. Piatetsky (2016), indica que el top 3 de los algoritmos más utilizados en el análisis de datos son Regression, Clustering y Decision/Rules, esto se puede evidenciar en la Figura 14. Estos algoritmos coinciden con lo que menciona Mayo (2019), con respecto a los más utilizados durante el periodo 2018–2019, como se puede ver en la Figura 15.

Figura 14. *Top 10 de Algoritmos y métodos usados por Data Scientists.*

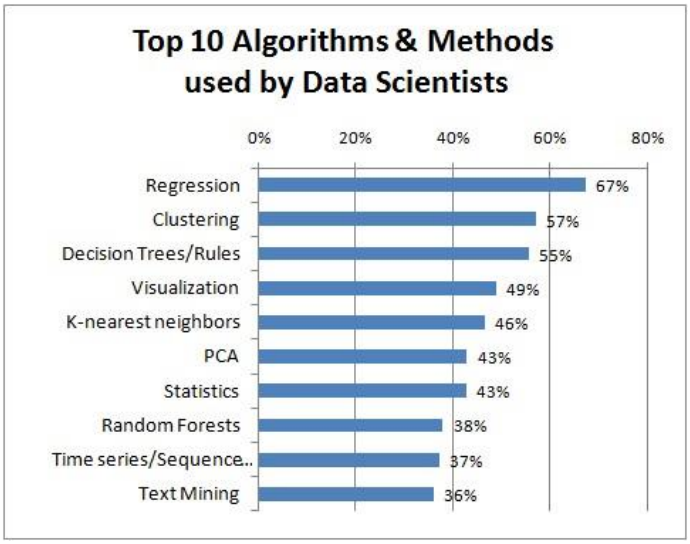
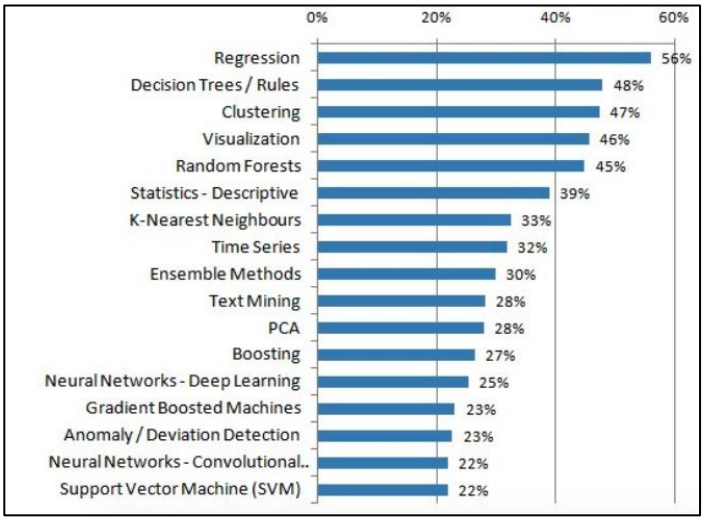


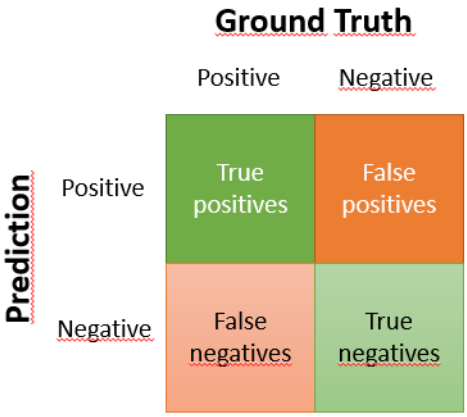
Figura 15. *Top 10 métodos de Data Science usados en 2018-2019.*



Tomando en cuenta esta información, se decidió utilizar los algoritmos: Logistic Regression, Decision Trees, kNN (estos presentes en el top). Así como los algoritmos: Neural Network, SVM y Naive Bayes, para aumentar la variedad de los resultados obtenidos. Finalmente, se utilizó Clustering para el agrupamiento de datos y su visualización en la implementación.

El siguiente paso en la investigación, es realizar un plan de pruebas, para esto se buscó elegir la herramienta que permita realizar una comparación detallada de los resultados de cada algoritmo. Se terminó eligiendo la Matriz de Confusión, dado que esta herramienta, también llamada tabla de contingencia, muestra el desempeño de un algoritmo, describiendo cómo se distribuyen los valores reales y sus predicciones (Burrueco, 2020). La representación de la misma se puede ver en la Figura 16:

Figura 16. *Ejemplo de matriz de confusión.*



Como se puede ver en la figura, la fila Positive son aquellos elementos que han sido clasificados como positivos, y la fila Negative son los que han sido clasificados como negativos. Por otro lado, la columna Positive son los que, en realidad, son positivos, y la columna Negative los que, en realidad, son negativos (téngase en cuenta que los conceptos de "positivo" o "negativo" son completamente arbitrarios) (Burrueco, 2020).

De esta forma los resultados se interpretaran según lo que busca el estudio, en este caso es la predicción en el aumento o disminución de precios.

La evaluación de los resultados de la matriz de confusión se realizó gracias a métricas específicas, pero no todas son relevantes, esto varía según a que resultado se le da prioridad. En este punto, se realizó un análisis de selección de métricas (Anexo Nº 15). Donde se terminó elegiendo a las métricas de precisión y exactitud (accuracy) como las principales para este estudio.

Para la construcción de modelo, se buscó identificar que herramientas de data mining son las más utilizadas y cual de estas se adecua al propósito del estudio. Estas herramientas se pueden ver en la comparativa realizada por Ionos (2018) en la Figura 17:

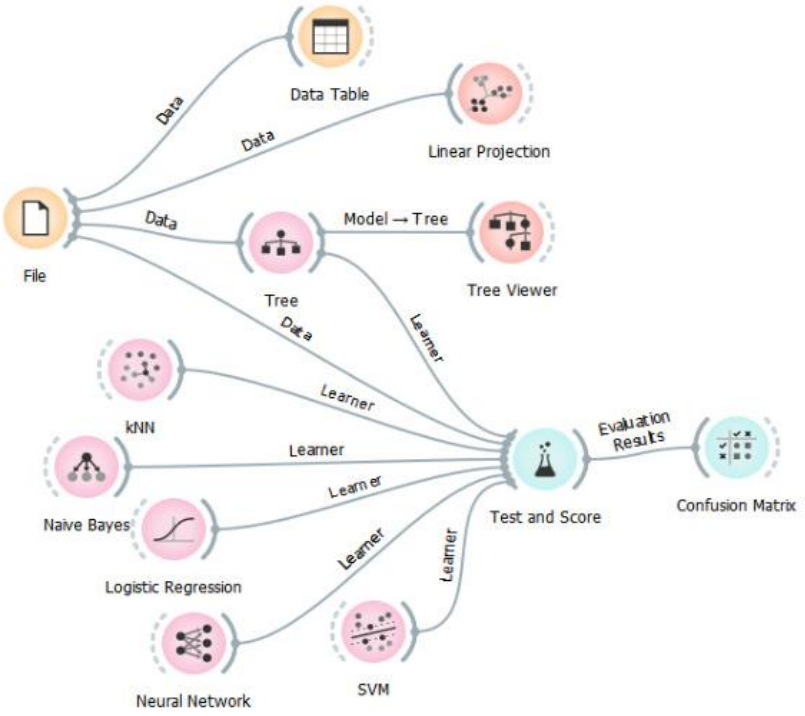
Figura 17. *Comparativa de sofware de data mining.*

	Características	Lenguaje de programación	Sistema operativo	Precio/Licencia
RapidMiner	Apto para todos los procesos. Destaca en el análisis predictivo	Java	Windows, macOS, Linux	Freeware, diferentes versiones de pago
WEKA	Muchos métodos de clasificación	Java	Windows, macOS, Linux	Software libre (GPL)
Orange	Crea una visualización de datos atractiva sin que se requieran muchos conocimientos previos para ello	Núcleo del software: C++, ampliación y lenguaje de entrada: Python	Windows, macOS, Linux	Software libre (GPL)
KNIME	Software de data mining de código abierto que ha democratizado el acceso a los análisis predictivos	Java	Windows, macOS, Linux	Software libre (GPL) (a partir de la versión 2.1)
SAS	Caro, pero potente para grandes empresas	Lenguaje SAS	Windows, macOS, Linux	Freeware limitado a instituciones públicas, el precio se establece tras solicitud, diferentes modelos disponibles

Se terminó eligiendo a Orange Data Mining, por que se tuvo como prioridad la visualización de datos y además brinda muchas funciones en una versión gratuita, lo que permitió testeos previos.

Utilizando la herramienta Orange se agregaron los algoritmos que se han seleccionado previamente en el proyecto y se procedió con el modelado. Donde se subió el archivo del reporte (File), y se verifica su contenido con un Data Table. Después se seleccionan los algoritmos a utilizar y se los agrega, estos algoritmos pasan por un proceso de aprendizaje, luego en Test and Score y Confusion Matrix se evaluan los resultados.Como se puede ver en la Figura 18.

Figura 18. *Modelado.*



Los resultados obtenidos fueron representados en un informe de resultados (Anexo Nº 16). Para la evaluación de los mismos se tomó en cuenta una fórmula para la eficacia, presente en un informe comparativo de resultados (Anexo Nº 17). Por otro lado, también se obtuvieron resultados con las métricas brindadas por la Confusion Matrix. Estas métricas son: la Exactitud, que consiste en el porcentaje de predicciones correctas frente al total, y la precisión, que se refiere a lo cerca que está el resultado de una predicción del valor verdadero. Haciendo una comparación de las métricas se identificó que el resultado de la formula de la eficacia es equivalente a el resultado de la exactitud (Confusion Matrix).

Siguiendo con los pasos, se realizó una evaluación más detallada basándose en la fase 5 (Fase de Evaluación), para que exista un mayor entendiento en como la variación de la cantidad de los datos afecta a los resultados que brinda el modelo. Cada una de las métricas seleccionadas se observaron tomando en cuenta 3 versiones de reporte :

- Daily Price Only, solo precio diario con dos columnas: fecha y precio (500 campos).
- DailyVariety, 3 columnas: fecha, variedad y precio (750 campos)
- DailyOriginVariety: fecha, variedad, origen y precio, 4 columnas (1000 campos).

  Los resultados obtenidos fueron:

- En la evaluación de resultados del reporte de Solo Precio Diario (Daily Price Only), el mayor valor obtenido en exactitud (CA) es de 0.652, este valor estuvo presente en 4 de los 6 algoritmos utilizados (SVM, Neural Network, Naive Bayes y Logistic Regression). Por otro lado, el mejor valor obtenido con respecto a precisión, fue de 0.555 en el de Classification Tree, como se puede ver en la Figura 19:

Figura 19. *Evaluation Results Daily Price Only.*

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.519	0.584	0.553	0.542	0.584
Tree	0.510	0.592	0.564	0.555	0.592
SVM	0.484	0.652	0.515	0.425	0.652
Neural Network	0.498	0.652	0.515	0.425	0.652
Naive Bayes	0.468	0.652	0.515	0.425	0.652
Logistic Regression	0.417	0.652	0.515	0.425	0.652

- En la evaluación de resultados del reporte de precio por Variedad (Daily Variety), el mayor valor obtenido en exactitud (CA) es de 0.644, este valor estuvo presente en el algoritmo de Neural Network. Por otro lado, el mejor valor obtenido en precisión fue de 0.618, también en Neural Network, como se puede ver en la Figura 20:

Figura 20. *Evaluation Results Variety Price.*

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.532	0.624	0.595	0.596	0.624
Tree	0.530	0.596	0.571	0.566	0.596
SVM	0.456	0.616	0.523	0.544	0.616
Neural Network	0.572	0.644	0.581	0.618	0.644
Naive Bayes	0.515	0.600	0.531	0.534	0.600
Logistic Regression	0.587	0.632	0.510	0.585	0.632

- En la evaluación de resultados del reporte de Variedad y Origen (Daily Origin Variety), el mayor valor obtenido en exactitud (CA) es de 0.664, este valor estuvo presente en el

algoritmo de Logistic Regression. Por otro lado, el mejor valor obtenido en precisión fue de 0.652, también en Logistic Regression, como se puede ver en la Figura x:

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.563	0.596	0.579	0.573	0.596
Tree	0.557	0.616	0.572	0.576	0.616
SVM	0.562	0.632	0.541	0.581	0.632
Neural Network	0.544	0.648	0.596	0.621	0.648
Naive Bayes	0.579	0.632	0.530	0.576	0.632
Logistic Regression	0.592	0.664	0.605	0.652	0.664

La conclusión de los resultados fue:

- El algoritmo más exacto y preciso terminó siendo el de Logistic Regression con 0.664 de exactitud y 0.652 de precisión.
- La eficacia de los algoritmos varió solo en 1 o 2 %. Donde verdaderamente se encuentra una diferencia, es en las predicciones de aumento o disminución, que terminan variando de distintas maneras dependiendo de cada algoritmo, esto se puede ver a más detalle el informe comparativo de resultados (Anexo Nº 17).
- Al aumentar la cantidad de datos empiezan a destacar diferentes algoritmos, por lo que se deduce que hay algoritmos que funcionan mejor con cierta cantidad de datos.
- La precision aumento en el reporte con más campos, lo cual es evidencia que a mayor cantidad de datos existe un aumento en la precisión.

3.1.4. Objetivo específico 4:

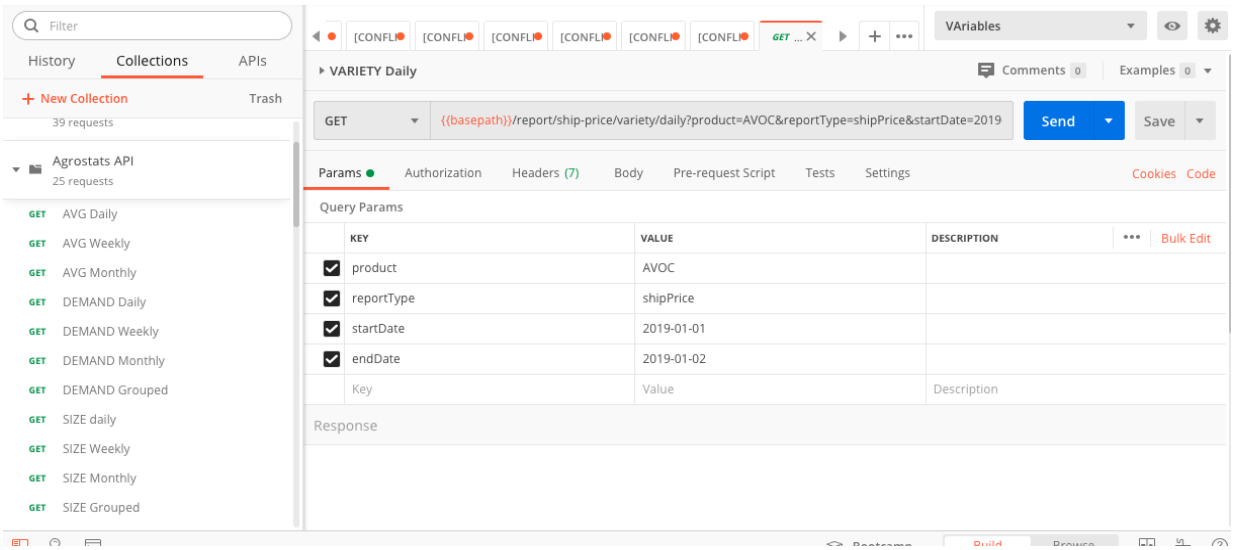
Implementar una solución tecnológica que permita el correcto procesamiento y visualización de los datos.

Este objetivo se centró en la fase final (fase 6) de la metodología CRISP-DM.

Para realizar la implementación del modelo se siguió el plan establecido por la empresa, con herramientas acorde a la actualidad y los recursos asignados al desarrollo del modelo. Esto se dio de la siguiente manera:

Una vez realizada la recolección de los datos, se identificó que herramientas actuales facilitan el acceso a la información y son adaptables, es por esto que se decide realizar la implementación de un API que permita estructurar los datos y al mismo tiempo facilite su uso, como se puede ver en la Figura 21.

Figura 21. *Postman de API desarrollada.*



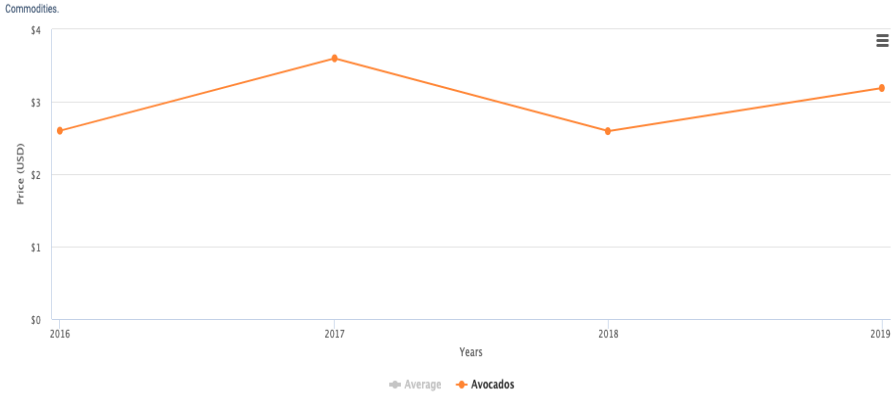
Una vez implementada el API, se decidió elegir un tipo de gráfico que se adapte a la visualización del agrupamiento requerido (este se dará utilizando Clustering), es aquí donde se eligió a los gráficos de lineas. Se utilizan gráficos de líneas para hacer un seguimiento de los cambios a lo largo de períodos de tiempo breves o extensos y para ayudar en análisis de datos predictivos. Si existen cambios pequeños y frecuentes en la serie, los gráficos de líneas son más eficaces que los gráficos de barras para visualizar el cambio a lo largo del tiempo. Los gráficos de líneas también resultan útiles para comparar los cambios a lo largo del mismo período de tiempo en varios grupos o categorías (Galvanize, 2020).

Para mostrar los gráficos de lineas se decidió utilizar Highcharts, una biblioteca gráfica basada en Javascript.

Una vez implementada la herramienta, se puede visualizar los datos de precio de paltas (Avocados) agrupados como se ve en la Figura 22, en esta, se muestra por ejemplo la tendencia

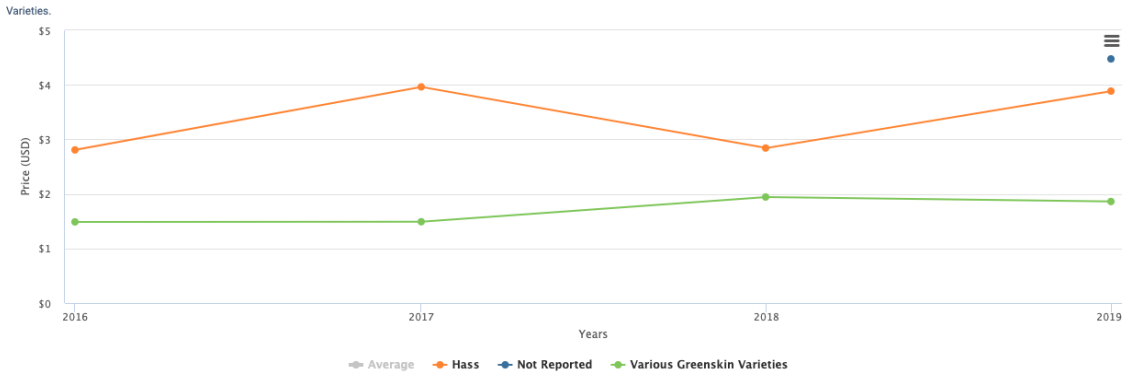
promedio del precio en los últimos 4 años. Se pueden visualizar que los puntos más altos están tanto en el 2017 como el 2019.

Figura 22. *Gráfico de los últimos 4 años con respecto a precio.*



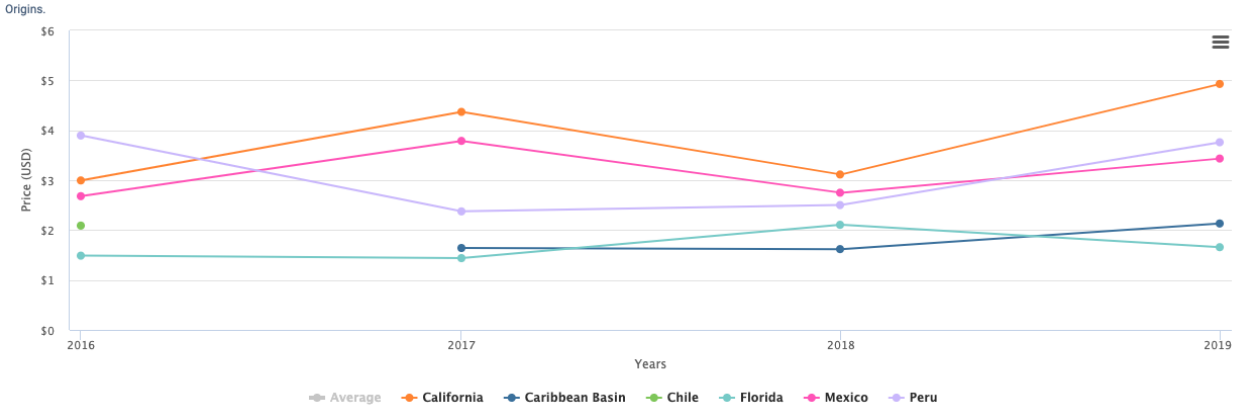
La herramienta permite ver otros valores en el gráfico, como se muestra en la Figura 23 . Donde se agregó Variety, para definir que tanto afecta el tener este campo en la visualización del precio. Se puede identificar que en variedad el precio de Hass es el que define el aumento en los años 2017 y 2019, y en el Various Green Skin ocurre lo contrario, dándose el aumento el 2018.

Figura 23. *Gráfico de los últimos 4 años con respecto a precio y variety.*



Finalmente, se muestra como se ven todos los origenes en un solo gráfico, como se puede identificar en la Figura 24, la mayor diferencia es que esta vez los datos se muestran con su lugar de origen creando una mayor amplitud en la visualización de los mismos.

Figura 24. *Gráfico de los últimos 4 años con respecto a precio y origen.*



Como se ha ido trabajando, se identificó que la cantidad de data a utilizar para que los análisis sean relevantes debe ser mínima de un año. Gracias a Highcharts se pueden evidenciar fluctuaciones más entendibles visualemente, esto se puede ver a detalle de manera ilustrativa en el reporte gráfico de los últimos meses (ver Anexo Nº 19).

También se terminó implementando un versión móvil, donde se agregó más variedad de gráficos por motivos de responsividad, esto se hizo con el uso de Android Studio. A continuación, se puede ver algunas capturas de la versión móvil, desde la Figura 25 a la Figura 27:

Figura 25. *Variedad Versión Móvil.*

***Figura 26**. Origen Versión Móvil.*

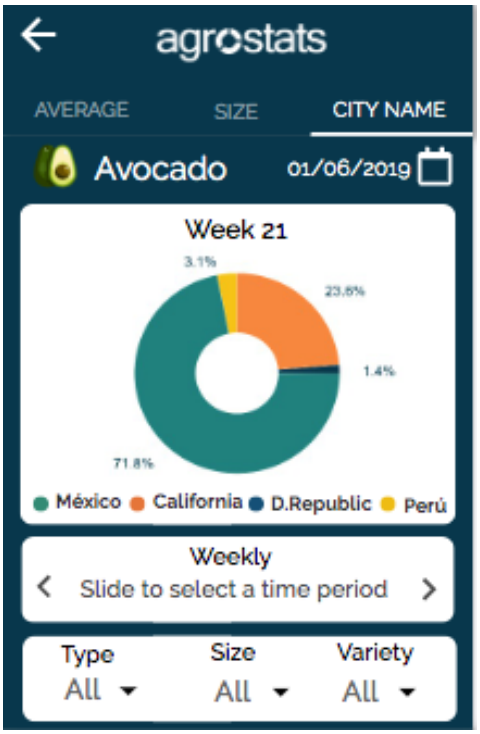


Figura 27. *Diversos Datos Versión Móvil.*

La visualización de la predicción son valores que se implementaron en la parte final. Como se puede ver en la Figura 28:

Figura 28. *Pantalla Inicial con Predicciones.*

3.1.5. Objetivo General:

Desarrollar un modelo predictivo que mejore el análisis de datos del sector exportador hortofrutícola en el periodo 2016-2020.

La mejora en el proceso de análisis de datos se realizó tomando en cuenta 3 principales deficiencias que presentó la empresa:

- El manejo de información es deficiente con respecto a la cantidad y calidad de los datos.

- La toma de decisiones se basa en datos subjetivos.
- El progresivo aumento de tiempo y recursos necesarios para obtener resultados del análisis de datos.

El desarrollo del modelo predictivo con la metodología CRISP-DM, permitió enfocarse en mejorar estos puntos. Esto se dio de la siguiente manera:

Con respecto a deficiencia en el manejo de la información a nivel de cantidad y calidad de los datos. Se comenzó realizando un análisis del estado actual de los datos, a través de un gráfico de Linear Projection en la Fase 2. Donde se pudo identificar que los datos diferian sustancialmente entre ellos. Por lo que, a través de un proceso de selección, estructurado e integración (Anexo 13), se creó un nuevo formato de reporte con los datos más completos. Este formato de reporte se utilizó, en el modelo predictivo, con datos del año 2019 (muestra elegida). La cantidad de datos, con respecto a a la versión original, varió de 2000 campos a 1000 (una reducción de 50%). Finalmente, teniendo datos de calidad, se decidió, por motivos comparativos, testear el modelo con 3 versiones de reporte:

- Daily Price Only, solo precio diario con dos columnas: fecha y precio (500 campos).
- DailyVariety, 3 columnas: fecha, variedad y precio (750 campos)
- DailyOriginVariety: fecha, variedad, origen y precio, 4 columnas (1000 campos).

Esto permite que exista un mayor entendiento en como la variación de la cantidad de los datos afecta a los resultados que brinda el modelo.

El análisis específico de los datos siempre se dio de una forma poco convencional, a pesar del valor que tienen en la toma final de decisiones, se basaban en suposiciones según el criterio de los analistas, sin ningún tipo de sustento o apoyo tecnológico, salvo la observación de los reportes en excel. El propósito del modelo fue cambiar esto, por lo que su relevancia fue representada principalmente por el porcentaje de precisión (precision) en la predicción y la exactitud(accuracy).

Analizando los resultados obtenidos de la Matriz de Confusión, se identificó que los valores de la Exactitud coinciden con el de la Eficacia. Esto se dá por que la definición de ambos en este contexto está orientada al mismo resultado, la exactitud es el porcentaje de predicciones correctas y la eficacia el resultado obtenido entre el resultado previsto. Los resultados terminaron brindando un porcentaje de eficacia de 66.4%, esto sustentado a más detalle en el informe comparativo de resultados (Anexo Nº 18) y 65.2% de precisión (Confusion Matrix), las conclusiones finales fueron:

- El algoritmo más eficaz y preciso terminó siendo el de Logistic Regression (con un porcentaje de 66.4% y 65.2% respectivamente).
- La eficacia de los algoritmos varió solo en 1 o 2 %. Donde verdaderamente se encuentra una diferencia, es en las predicciones de aumento o disminución, que terminan variando de distintas maneras dependiendo de cada algoritmo.
- La variación en la cantidad de campos utilizados influye en predecir el aumento más no la disminución, la demostración de esto se vio de manera más evidente en el algoritmo kNN.

- La precision aumento en el reporte con más campos, lo cual es evidencia que a mayor cantidad de datos existe un aumento en la precisión.

Finalmente, enfocándose en el tiempo y recursos necesarios para el análisis de datos. Se tuvo como datos preliminares que: los costos mensuales que invierte la empresa en el análisis de datos se pueden deducir tomando en cuenta las horas que invierte el personal del área comercial en este proceso, donde los asistentes de exportaciones (son 3 asistentes) tienen como parte de sus funciones realizar el análisis del sector exportador frutícola (la estructura organizacional se puede ver a más detalle en Anexo Nº 20).

Primero, se empezó evaluando el impacto económico del desarrollo del modelo, esto se hizo utilizando la fórmula presente en la Figura 29:

Figura 29. *Fórmula de Costo por Hora de Trabajo.*

$$\frac{\text{COM} + \text{SEM}}{160^*} = \text{CxH}$$

COM = Costos Operativos Mensuales
SEM = Sueldo Estimado Mensual
CxH = Costo por hora de trabajo
* 8 hrs diarias x 5 días a la semana x 4 semanas al mes = 160

Y se determinó el costo aproximado de ejecución del proyecto con la fórmula presente en la Figura 30:

Figura 30. *Costo de ejecución del Proyecto.*

$$CxH \times HT + GE = CEP$$

CxH = Costo por hora de trabajo
HT = Horas Trabajo (tiempo que toma desarrollar proyecto)
GE = Gastos Extra(gastos adicionales como hosting, sofware, etc)
CEP = Costo de ejecución del Proyecto

Se obtuvo que:

El costo de la ejecución del proyecto se calcula con el costo por hora de trabajo de los ingenieros de sistemas, que fué de S/. 26 de manera individual y S/. 52 en total (dado que fueron 2 desarrolladores) , las horas de trabajo planificadas para el desarrollo del proyecto fueron de 480, los gastos extra equivalen a S/. 3000 (mantenimiento, servidores y software). Esto da un valor de costo de ejecución del proyecto de S/. 27960.

Por otra lado, aplicando los cálculos para determinar el costo que se invierte actualmente, se tiene que:

Como costos operativos mensuales, para el área, se utiliza un promedio de S/. 3000 y S/.2000 como sueldo estimado mensual, pero este es multiplicado por cada analista (actualmente en la empresa son 3), esto da un valor de S/. 6000. Esto resulta en un costo de S/. 9000 soles mensuales. El costo por hora de trabajo de los 3 asistentes termina siendo de S/. 56.25.

Para hacer un cálculo más específico de que tiempo se invierte en el análisis de datos, se identificó que cada asistente demora aproximadamente 1 hora diaria buscando fuentes de información, esto se hace buscando información de manera online y reportes brindados por una empresa externa encargada de recopilar datos. También tardan aproximadamente 2 horas más

realizando un informe en Excel con los datos más relevantes a mostrar. Este proceso se da de manera diaria y mensualmente se debe realizar una recolección de datos históricos y análisis de estos, que toma aproximadamente de 8 horas. Esto da un total de 68 horas mensuales utilizadas en el análisis de datos. Si se multiplica esto por el costo por hora de S/. 56,25. Se obtiene un costo mensual de S/.3825, trimestral de S/. 11475 y anual de S/. 45900. Si comparamos la inversión del proyecto que es de 3 meses existe un costo mayor que el gasto que se realiza en la empresa. Pero considerando que esta implementación ya no generará más gasto (ya se consideraron los gastos extra), se podría decir que anualmente se logra un ahorro de S/. 17940. Que equivale a una reducción del 39,08% del costo anual.

El tiempo estimado de respuesta del modelo es de 3 a 5 segundos, y se recopila datos de forma diaria, este tiempo de respuesta también incluye datos históricos de ser necesario. Aparte de esto, gracias a la implementación realizada en Highcharts es posible realizar análisis visuales de manera inmediata con la data que se desee seleccionar, sin que sea necesario el uso de Excel. El tiempo que tomaría obtener un reporte sería máximo de 1 minuto.

En resultados numéricos el tiempo invertido en el análisis de datos disminuye en un 180% con respecto al que se solía invertir. Por otro lado, el personal necesario para encargarse de consultar los reportes o la lectura de los gráficos es solo de una persona.

IV. DISCUSIÓN Y CONCLUSIÓNES

4.1. DISCUSIÓN:

En la realización de este proyecto se destacan dos limitaciones (dividir en prácticas metodológicas y etc). Como primer punto, se tuvo el acceso libre a los datos utilizados, que permitió que el desarrollo de la etapa planeación del proyecto y la aplicación de la metodología CRISP-DM, desde la fase 1 a la fase 5, se dé de manera exitosa. Sin embargo, por motivos que involucraron la pandemia del 2020, el acceso al desarrollo Backend y planes de implementación fueron limitados. A pesar de esto, se formó parte del proceso y los resultados obtenidos siguen siendo datos relevantes que permiten que se evidencie un avance en la realización de este tipo de modelos. Como segundo punto, dado que las dos personas a cargo de la investigación se encontraban en diferentes ciudades, existió una escasez de tiempo en la que se pudieran realizar avances, los horarios de trabajo tampoco apoyaron a esta situación, sin embargo, gracias a las herramientas virtuales y reuniones planificadas, se fue capaz de concluir el proyecto.

Tomando en cuenta la investigación Implementación de un modelo predictivo basado en data mining soportado por SAP Predictive Analitys en retails (Castro & Hernández, 2016), existe una coincidencia en que a más data existe mayor precisión, sin embargo hay una diferencia respecto a los algoritmos usados. Se tiene al algoritmo Triple Exponential Smoothing con 23.6 % de margen de error, lo cual equivale a un 76.4% de precisión. Comparadose con el estudio realizado existe un mayor valor de precisión dado que en este se obtuvo 65.2% . Sin embargo, solo

se muestra ese porcentaje en el algoritmo Triple Exponential, el segundo algoritmo que utilizaron, Linear Regression, les dio como resultado un 62.87% de porcentaje de error. Lo que evidencia que existe una diferencia abismal entre los resultados de esos algoritmos. Por otro lado, la similitud de los algoritmos utilizados es más cercana en la investigación Modelo Predictivo Machine Learning aplicado al análisis de datos climáticos capturados por una placa Sparkfun (Iribarren, 2016), donde se usó Logistic Regression (obtuvieron un porcentaje de precisión de 85.65%), Neural Network(obtuvieron un porcentaje de precisión de 85.66) y Multiclass decision forest (obtuvieron un porcentaje de precisión de 85.82). Se concluyó como el más eficaz al Multiclass decision forest (este no fue utilizado en el presente estudio). Aun así,en el segundo y tercer lugar ,Neural Network y Logistic Regression (ambos presente en los resultados del proyecto), el margen de diferencia de ambos fue solo de 0.1%. Un porcentaje parecido al del estudio. Por lo que se puede decir, que depende mucho que algoritmos son seleccionados para cada modelo, pues, la poca variedad de algoritmos puede llevar a que no todos sean eficaces.

La reducción de tiempo invertido y costos también estuvo presente en la investigación Diseño e Implementación de un Sistema de Visión Artificial para Clasificación de al menos Tres Tipos de Frutas (Constante & Gordón, 2015) , ellos lograron optimizar ciertos parámetros de producción como tiempo, espacio, calidad, higiene. Claro que esto más enfocado en el sector de campo. Estos ayudaron a crear una mejor competitividad dentro del campo agrícola.

Como implicancia teórica, el estudio se basó en definiciones de minería de datos, más que nada presente en los algoritmos. Estos también se usan en machine learning, inteligencia artificial y redes neuronales, que coinciden en un propósito: unificar datos. En el estudio realizado los

algoritmos fueron seleccionados basandose en un ranking de los más utilizados, por lo que se recomienda que se tenga en cuenta investigar siempre que algoritmos resaltan y se adecuan más al tipo de modelo que se quiere construir. Así se tendrá mayor diversidad en los resultados, y por ende, mejorará la valoración de los mismos.

Como implicancias metodológicas, se tiene que la metodología utilizada (CRISP-DM), está basada en el proceso de KDD (Knowledge Discovery in Databases), usado ampliamente en el conocimiento de datos. Esta metodología no es la única que se usa en proyectos de minería, pero es muy útil para comprender esta tecnología y extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características. Hay que tener en cuenta la capacidad de iteración de la metodología, que permite evaluar y mejorar el modelo para versiones futuras.

Como implicancias prácticas, se debe mencionar que el modelado fue mostrado de manera gráfica y evidencio el potencial de la predicción de los datos en el sector de exportación hortofrutícola,en este caso se tomó el precio como variable central y se lograron mejoras en el análisis. Este proceso es replicable a otras frutas (por ahora solo se utilizó en palta) y también pueden ser usados no solo en este sector, si no ser replicados en otros, como precios de otros productos o hasta servicios.

Finalmente, este estudio (como otros estudios en el sector), busca servir como un primer peldaño para expandir y ampliar el interés en el tema de modelos predictivos en el análisis de datos. Si bien los resultados se dieron basandose en cierta cantidad de datos, se recomienda ampliar la misma y analizar si traen mejores resultados, además se debe buscar utilizar siempre información histórica que permitirá realizar análisis más profundos y específicos.

4.2. CONCLUSIONES:

Se logró realizar el desarrollo de un modelo predictivo para la mejora del análisis de datos del sector hortofrutícola, durante el periodo 2016 – 2020, utilizando la metodología CRISP-DM. Esto se dio cubriendo las deficiencias existentes en la empresa, con un agregado de datos concisos en la predicción de precios, una disminución de 39.08% del costo anual en los recursos utilizados y un porcentaje de 180% menos tiempo invertido en el análisis de datos.

Se identificaron las fuentes y reportes de datos de valor en el sector hortofrutícola; se siguieron las pautas brindadas en la primeras fases de la metodología CRISP-DM. El análisis preliminar del sector hortofrutícola fue vital para tener un mejor entendimiento de cómo se usan los datos, la fuente elegida termino siendo la USDA, esta presentó una amplia gamma de datos que facilitó el manejo de los mismos.

Se estableció un formato de reporte final tomando en cuenta los datos de mayor importancia y su disponibilidad, se logró reducir en un 50% la cantidad de datos a procesar. Logrando un estándar que origina un avance significativo para procesar datos orientados al data mining del sector.

Se analizaron diversos algoritmos que permitieron realizar predicciones con respecto a el aumento o disminución del precio de los productos del sector hortofrutícola, no todos dieron como resultado una precisión alta, pero la variación entre ellos no fue mucha. El algoritmo más eficaz fue el de Logistic Regression con un porcentaje de 65.2% de precisión y eficacia de 66.4%.

Se implementó el modelo utilizando un API, la herramienta Highcharts y Android Studio. La visualización de los datos de manera gráfica es vital para el mejor entendimiento de estos. Aparte de esto, el poder realizar consultas variadas, que pueden ser hasta históricas, de manera fácil permitió que no sea necesario que sea un experto quien se encargue del análisis de datos.

REFERENCIAS

Agasys. (23 de noviembre de 2017). Importancia del análisis de información en las empresas. Obtenido de http://www.agasys.com.mx/importancia-del-analisis-de-informacion-en-las-empresas/

Ariser. (2015). Historia del Big Data – Del comienzo del análisis de datos a nuestros días. Obtenido de Big data para curiosos: https://bigdataparacuriosos.wordpress.com/historia-big-data/

B12. (03 de febrero de 2020). Qué es un modelo predictivo y cómo se aplica al negocio. Obtenido de https://agenciab12.pe/noticia/que-es-modelo-predictivo-como-aplica-negocio#:~:text=Un%20modelo%20predictivo%20es%20un,vez%2C%20detectar%20oportunidades%20de%20negocio.

Bastías, N. (2016). Modelo Predictivo Para Intensidades Sísmicas Superficiales en Chile. Concepción: Universidad de Concepción.

Burrueco, D. (2020). Matriz de confusión. Obtenido de Interactive Chaos: https://www.interactivechaos.com/manual/tutorial-de-machine-learning/matriz-de-confusion

Castro, A. P., & Hernández, J. P. (2016). Implementación de un modelo predictivo basado en data mining soportado por SAP Predictive Analitys en retails. Lima: Universidad de Ciencias Aplicadas.

Castro, L. (2010). Metodología de la Investigación. Caracas: Universidad Central de Venezuela.

Cavusgil, S., & Nevin, J. (1981). Internal Determinants of Export Marketing Behavior: An Empirical Investigation. Journal of Marketing Research.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0 Step-by-Step Data Mining Guide. Obtenido de https://www.the-modeling-agency.com/crisp-dm.pdf

Conceptodefinicion. (23 de Julio de 2019). Definición de análisis. Obtenido de https://conceptodefinicion.de/analisis-de-datos/

Constante, P., & Gordón, A. (2015). Diseño e Implementación de un Sistema de Visión Artificial para Clasificación de al menos Tres Tipos de Frutas. Quito: Escuela Politécnica Nacional.

Córdova, M., & Monsalve, C. (2006). Tipos de Investigación: Predictiva, proyectiva, interactiva, confirmatoria y evaluativa.

Emprendedores. (18 de octubre de 2015). Cómo hacer un estudio de mercado si vas a exportar. Obtenido de https://www.emprendedores.es/gestion/a22198/como-hacer-estudio-de-mercado-exportacion-comercio-exterior/

Ferreyra, J., & Vásquez, J. L. (2012). Proyección de precios de exportación utilizando tipos de cambio: Caso peruano. Lima: Banco de Reserva del Perú.

Figueredo, G., & Ballesteros, J. (2016). Identificación del estado de madurez de las frutas con redes neuronales artificiales, una revisión. Boyacá: Revista Ciencia y Agricultura.

Galvanize. (2020). Gráfico de Lineas. Obtenido de https://help.highbond.com/helpdocs/highbond/es/Content/visualizations/interpretations/charts/line_chart.html#:~:text=Los%20gr%C3%A1ficos%20de%20l%C3%ADneas%20muestran,en%20un%20gr%C3%A1fico%20de%20l%C3%ADneas.

Gonzáles, M. (2012). Generación de modelos predictivos de satisfacción transaccional para un centro de atención a clientes. Atizapán de Zaragoza: Instituto Tecnológico y de Estudios Superiores de Monterrey.

Heras, D. (2017). Clasificador de imágenes de frutas basado en inteligencia artificial. Cuenca: Universidad Católica de Cuenca.

Hernandez Sampieri, R., Fernandes Collado, C., & Baptista Lucio, M. d. (2014). Metodología de la Investigación Sexta Edición. Ciudad de México: McGrawHill Education.

Hurtado, J. (11 de junio de 2012). Capítulo 3: Marco Metodológico. Obtenido de http://virtual.urbe.edu: http://virtual.urbe.edu/tesispub/0092769/cap03.pdf

Ionos . (30 de enero de 2018). Software de data mining: realiza análisis de datos más efectivos. Obtenido de https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Iribarren, D. (2016). Modelo Predictivo Machine Learning aplicado al análisis de datos climáticos capturados por una placa Sparkfun. Madrid: Universidad Pontificia Comillas.

Jha, G. K., & Sinha, K. (2013). Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System. Agricultural Economics Research Review, 229-239.

La República. (8 de Julio de 2019). Exportaciones de alimentos de Latinoamérica se incrementará al 25 % en el 2028. Obtenido de https://larepublica.pe/economia/2019/07/08/exportaciones-de-alimentos-de-latinoamerica-se-incrementara-al-25-en-el-2028/

Logicalis. (03 de mayo de 2015). Modelos Predictivos: Reforzando el valor de una buena decisión. Obtenido de https://blog.es.logicalis.com/analytics/modelos-predictivos-reforzando-el-valor-de-una-buena-decision

Lozada, J. (2014). Investigación Aplicada: Definición, Propiedad Intelectual e Industria. Cienciamérica, 34-39.

Marín Castro, H. M. (2015). Minería de datos. Obtenido de https://www.tamps.cinvestav.mx/~hmarin/Mineria/EC2.pdf

Mayo, M. (Abril de 2019). Top data science machine learning methods 2018-2019. Obtenido de Kdnuggets: https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

Microstrategy. (2020). Predictive modeling the only guide you need. Obtenido de https://www.microstrategy.com/es/resources/introductory-guides/predictive-modeling-the-only-guide-you-need

Milanzi, M. (2012). The Impact of Barriers on Export Behavior of a Developing Country Firms: Evidence from Tanzania. International Journal of Business and Management Vol. 7.

OMC. (2019). Examen Estadístico del Comercio Mundial 2019. Ginebra, Suiza: secretaria de la OMC.

Piatetsky, G. (Septiembre de 2016). Top Algorithms and Methods Used by Data Scientists. Obtenido de KDnuggets: https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html

Pizarro, C. (2017). App para móviles de detección de características hortofrutícolas mediante tratamiento de imágenes. Extremadura: Universidad de Extremadura.

Question Pro. (2020). ¿Qué es el análisis de datos?. Obtenido de https://www.questionpro.com/es/analisis-de-datos.html

Recuero de los Santos, P. (9 de septiembre de 2020). Blog Think Big Empresas. Obtenido de Sitio Web de Telefónica: https://empresas.blogthinkbig.com/como-interpretar-la-matriz-de-confusion-ejemplo-practico/

Roman, V. (25 de abril de 2019). Algoritmos Naive Bayes: Fundamentos e Implementación. Obtenido de https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

Rouse, M. (noviembre de 2012). Análisis de Datos. Obtenido de https://searchdatacenter.techtarget.com/es/definicion/Analisis-de-Datos

Simfruit. (17 de diciembre de 2014). Agronometrics: Precios y volúmenes de frutas de todo el mundo en un sólo lugar y en el momento que se requiera. Obtenido de https://www.simfruit.cl/agronometrics-precios-y-volumenes-de-frutas-de-todo-el-mundo-en-un-solo-lugar-y-en-el-momento-que-se-requiera/

Spain, K. (5 de mayo de 2020). Las 11 técnicas más utilizadas en el modelado de análisis predictivos. Obtenido de https://keyrusspainblog.com/2020/05/05/las-11-tecnicas-mas-utilizadas-en-el-modelado-de-analisis-predictivos/

SUNAT. (2018). Manifiestos de exportación marítimo. Obtenido de http://www.aduanet.gob.pe/cl-ad-itconsmanifiesto/manifiestoITS01Alias?accion=cargarFrmConsultaManifiestoExportacion

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional. Bogotá, Colombia: Ediciones Universidad Cooperativa de Colombia.

USDA. (2020). Report Results. Obtenido de https://www.marketnews.usda.gov/mnp/fv-report-top-filters?locName=&commAbr=AVOC&commName=AVOCADOS&className=FRUITS&rowDisplayMax=25&startIndex=1&navClass=FRUITS&navType=byComm&repType=termPriceDaily&type=termPrice

Vallalta, J. F. (2020). Objetivos de un proyecto de data mining. Obtenido de Healthdataminer: https://healthdataminer.com/data-mining/objetivos-de-un-proyecto-de-data-mining/

Workman, D. (10 de agosto de 2020). Top Imported Fruits Most Loved by Americans. Obtenido de World's top exports: http://www.worldstopexports.com/top-imported-fruits-most-loved-by-americans/

ANEXOS

Anexo Nº 1. Operacionalización de variables.

Variable independiente (VI)

Tabla 7. *Operacionalización de Variable Independiente*

Variable	Definición	Dimensiones	Indicadores	Rango de Valores
Modelo Predictivo	Un modelo predictivo es un conjunto de procesos ejercidos a través de técnicas computacionales de análisis de datos que ayudan a inferir la probabilidad de que ocurran determinadas situaciones previas a su consecución y, a su vez, detectar oportunidades de negocio (B12, 2020).	Predicción	• Precisión.	• Porcentajes (0 a 100%) ó Rango de valores (0 a 1)

Variable dependiente (VD)

Tabla 8. *Operacionalización de Variable Dependiente*

Variable	Definición	Dimensiones	Indicadores	Rango de Valores
Análisis de Datos del Sector Exportador Hortofrutícola	El análisis de datos es un proceso que te permitirá conocer e interpretar información con la finalidad de identificar puntos de valor (Question Pro, 2020) .	• Calidad • Recursos • Tiempo	• Eficacia • Costo • Horas invertidas	• Porcentajes (0 a 100%) ó Rango de valores (0 a 1) • Nuevos Soles (S/.) • Horas.

Anexo Nº 2. Población.

Tabla 9. *Número de reportes del 2016 al 2020*

Año	Nº de reportes
2016	4997
2017	3494
2018	4213
2019	4752
2020	3960
Total	21416

Fuentes de información:

- Reportes Anuales USDA (USDA, 2020)

Anexo Nº 3. Fórmula para cálculo de muestra muestra.

La fórmula es:

Figura 31. *Fórmula de muestreo.*

$$n = \frac{N * Z_a^2 * p * q}{d^2 * (N-1) + Z_a^2 * p * q}$$

Donde:

N = $Total\ \ población$ (21416)
Z_a = $Confianza\ al$ 90%
p = $proporción\ esperada$ 5%
q = 1 − p ($En\ este\ caso$ 1 − 0.5 = 0.95) d = $precisión$ (5%)

n=Tamaño de muestra

n=266

Anexo Nº4. Fichas de Validación.

FICHA PARA VALIDACIÓN DEL INSTRUMENTO

I. REFERENCIA

- **1.1. Experto: Miguel Angel Luzquiños Riojas**
- **1.2. Especialidad: Exportación Internacional**
- **1.3. Cargo Actualidad: Analista de Comercial de Exportación**
- **1.4. Institución: Agrícola Pampa Baja S.A.C.**
- **1.5. Tipo de Instrumento: Revisión Documentaria**
- **1.6. Lugar y fecha:07/05/2020**

II. TABLA DE VALORACIÓN POR EVIDENCIAS

Nº	EVIDENCIAS	VALORACIÓN					
		5	4	3	2	1	0
1	Pertinencia de indicadores		X				
2	Formulado con lenguaje apropiado		X				
3	Adecuado para los sujetos de estudio	X					
4	Facilita la prueba de hipótesis		X				
5	Suficiencia para medir la variable		X				
6	Facilita la interpretación del instrumento		X				
7	Acorde al avance de la ciencia y tecnología	X					
8	Expresado en hechos perceptibles	X					
9	Tiene secuencia lógica	X					
10	Basado en aspectos teóricos	X					
	Total	**25**	**20**				

Coeficiente de valoración porcentual: c = 95%

III. OBSERVACIONES Y/O RECOMENDACIONES

La nomenclatura de city name se refiere a Origen, tratar de especificar

……

Firma del Experto

FICHA PARA VALIDACIÓN DEL INSTRUMENTO

I. REFERENCIA

- **1.1. Experto: Carlo André Alva Cohello**
- **1.2. Especialidad: Ingeniería de Software**
- **1.3. Cargo Actualidad: Developer Senior**
- **1.4. Institución: Wunderman Thopmson**
- **1.5. Tipo de Instrumento: Revisión Documentaria**
- **1.6. Lugar y fecha:15/05/2020**

II. TABLA DE VALORACIÓN POR EVIDENCIAS

Nº	EVIDENCIAS	VALORACIÓN					
		5	4	3	2	1	0
1	Pertinencia de indicadores		X				
2	Formulado con lenguaje apropiado		X				
3	Adecuado para los sujetos de estudio	X					
4	Facilita la prueba de hipótesis	X					
5	Suficiencia para medir la variable	X					
6	Facilita la interpretación del instrumento		X				
7	Acorde al avance de la ciencia y tecnología		X				
8	Expresado en hechos perceptibles	X					
9	Tiene secuencia lógica		X				
10	Basado en aspectos teóricos		X				
	Total	**20**	**24**				

Coeficiente de valoración porcentual: c = 88%

III. OBSERVACIONES Y/O RECOMENDACIONES

Tratar de eliminar valores que no aportan a la investigación.

.....

Firma del Experto

Anexo Nº5. Carta de autorización.

CARTA DE AUTORIZACIÓN DE USO DE INFORMACIÓN DE EMPRESA

UNIVERSIDAD PRIVADA DEL NORTE

Yo Jorge Alberto Luzquiños Rodriguez, identificado con DNI 17534978, en mi calidad de Gerente General de la empresa/institución Luzquiños Corp EIRL con R.U.C N° 20601819105, ubicada en la ciudad de Chiclayo.

OTORGO LA AUTORIZACIÓN,

A los señores: Chávez Sánchez Wernher D'alembert y Tapia Álvarez Bryam André, identificados con DNI Nº 71744016 y 47589836 egresados de la (X)Carrera profesional o ( ) Programa de Postgrados de Ingeniería de Sistemas Computacionales para que utilice la siguiente información de la empresa:

Reportes de exportación históricos de la empresa e información adicional de exportación.

con la finalidad de que pueda desarrollar su ( )Trabajo de Investigación, (X)Tesis o ( )Trabajo de suficiencia profesional para optar al grado de ( )Bachiller, ( )Maestro, ( )Doctor o ( )Título Profesional.

Recuerda que para el trámite deberás adjuntar también, el siguiente requisito según tipo de empresa:

- Vigencia de Poder. *(para el caso de empresas privadas).*
- ROF / MOF / Resolución de designación, u otro documento que evidencie que el firmante está facultado para autorizar el uso de la información de la organización. *(para el caso de empresas públicas)*
- Copia del DNI del Representante Legal o Representante del área para validar su firma en el formato.

Indicar si el Representante que autoriza la información de la empresa, solicita mantener el nombre o cualquier distintivo de la empresa en reserva, marcando con una "X" la opción seleccionada.
(X) Mantener en Reserva el nombre o cualquier distintivo de la empresa; o
( ) Mencionar el nombre de la empresa.

LUZQUINOS CORP. EIRL.
Jorge A. Luzquiños R.
GERENTE

Firma y sello del Representante Legal o Representante del área
DNI:17534978

El Egresado/Bachiller declara que los datos emitidos en esta carta y en el Trabajo de Investigación, en la Tesis son auténticos. En caso de comprobarse la falsedad de datos, el Egresado será sometido al inicio del procedimiento disciplinario correspondiente; asimismo, asumirá toda la responsabilidad ante posibles acciones legales que la empresa, otorgante de información, pueda ejecutar.

Firma del Egresado
DNI: 71744016

Firma de Egresado
DNI:47589836

Anexo Nº6. Marco Conceptual.

1. Introducción

El presente documento busca definir y ampliar el conocimiento sobre la metodología CRISP-DM, la cual fue la elegida para el desarrollo del proyecto.

2. Metodología CRISP DM.

El modelo provee una representación completa del ciclo de vida de un proyecto de minería de datos. El proceso es dinámico e iterativo, por lo que la ejecución de los procesos no es estricta y con frecuencia se puede pasar de uno a otro proceso, de atrás hacia delante y viceversa. Éstos dependen del resultado de cada fase o la planeación de la siguiente tarea por ejecutar (Timarán et al., 2016).

Las fases de esta metodología se adaptan a los objetivos del proyecto:

Fases de la metodología

Fase 1. Comprensión del negocio o problema. Comprende los requisitos y objetivos del proyecto desde una perspectiva empresarial o institucional para convertirlos en objetivos técnicos y en un plan de proyecto, para lo cual es necesario comprender de manera completa el problema por resolver (Timarán et al., 2016).

• Determinar los objetivos.

• Evaluar la situación actual.

• Determinar los objetivos de la minería de datos.

• Producir un plan de proyecto.

Fase 2. Comprensión de los datos. Corresponde a la recolección inicial de los datos para establecer un primer contacto con el problema; esta fase, junto con la fase 3 y la fase 4, demanda mayor esfuerzo y tiempo (Timarán et al., 2016).

Las principales tareas que se deben desarrollar en la fase de comprensión de los datos son:

- Recolectar datos iniciales.
- Describir los datos.
- Explorar los datos
- Verificar la calidad de los datos.

Fase 3. Preparación de los datos. Se usa para adaptarlos a la técnica de minería de datos, mediante la visualización de los datos y la búsqueda de relaciones entre las variables. Esta fase es la de modelado, ya que los datos requieren ser procesados de diferentes formas; por ende, las fases de preparación y modelado interactúan permanentemente (Timarán et al., 2016).

Los pasos que se consideran para la preparación de los datos son:

- Seleccionar los datos.
- Limpiar los datos.
- Estructurar los datos.
- Integrar los datos.

- Formatear los datos.

Fase 4. Modelado. Corresponde a la selección de un modelo adecuado y específico; para ello se usan técnicas que cumplan los siguientes criterios:

- Ser apropiada para el problema.
- Disponer de datos adecuados.
- Cumplir con los requisitos del problema.
- Técnica adecuada para obtener un modelo.
- Conocimiento pleno de la técnica.

Por ejemplo, si el problema es de clasificación, podemos elegir entre árboles de decisión, *k-nearest neighbour* o razonamiento basado en caos (Timarán et al., 2016).

- Generar plan de prueba.
- Construir el modelo.
- Evaluar el modelo.

En esta fase se busca el descubrimiento de patrones insospechados y de interés, se cuenta con técnicas que permiten crear dos tipos de modelos: predictivo y descriptivo, en este proyecto se realizaron algunas de las tareas/técnicas más importantes de modelado con la minería de datos:

Segmentación o clustering (descriptivo):

Esta técnica consiste en agrupar los datos, ayudando a construir particiones de dicho conjunto de datos, permitiendo con la segmentación tener subpoblaciones de la información tales

como por ejemplo la fecha de los reportes de las frutas, y/o si la demanda de dicha fruta es poca o moderada, ciudad, tamaño de la fruta, entre otros. A diferencia de la clasificación, no busca el agrupamiento de datos para una futura predicción (Timarán et al., 2016).

Clasificación (predictiva):

Permite obtener resultados a partir de un proceso, permitiendo encontrar propiedades comunes entre los objetos de una base de datos tales como la demanda, por ejemplo, si la fruta tiene o no una demanda. Haciendo uso datos anteriores y los más recientes, se puede hacer una clasificación para datos futuros de si una fruta tendrá demanda o no (Timarán et al., 2016).

Regresión:

A diferencia de la Clasificación, la regresión permite predecir valores. Consiste en aprender una función real que asigna a cada instancia un valor real, de manera que el objetivo es minimizar el error entre el valor predicho y el valor real (Marín Castro, 2015).

Cada una de las técnicas se pueden combinar para establecer un modelo más preciso, todo depende de los objetivos. En este caso se guio de la variedad de algoritmos que se aplican con esta técnicas. Estos serán:

Regresión Logística

Las regresiones logísticas son utilizadas para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictivas. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. Por ejemplo, puede utilizarse para predecir el riesgo crediticio (Spain, 2020).

Redes Neuronales

Imita las neuronas del cerebro humano ya que es capaz de modelar relaciones extremadamente complejas y suele utilizarse cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y los de salida (Spain, 2020).

Máquinas de Vectores de Soporte (SVM)

Son algoritmos de aprendizaje automático supervisado de cara a reconocer patrones, estando relacionados con problemas de clasificación o regresión (Spain, 2020).

K-Vecinos más Cercanos

Consiste en reconocer patrones para conocer la probabilidad de que un elemento pertenezca a una clase según su cercanía en el espacio a los elementos de esa clasificación (Spain, 2020).

Naive Bayes

Son una clase especial de algoritmos de clasificación de Aprendizaje Automático, o Machine Learning, en ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica (Roman, 2019).

Árboles de Decisión

Son modelos de clasificación muy utilizados que tratan de encontrar la variable que permita dividir el dataset en grupos lógicos que son más diferentes entre sí. Cada árbol se va descomponiendo en distintas ramas y hojas que representan cada clasificación en función de las condiciones que se van seleccionando hasta llegar a la resolución del problema. Estos modelos son de gran ayuda a la hora de determinar las decisiones a lo largo de un proceso como por ejemplo el funnel de compra (Spain, 2020).

Fase 5. Evaluación. Evalúa el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema; para ello se emplean múltiples herramientas para la interpretación de los resultados, entre ellas matrices de confusión, que es una tabla que indica cuántas clasificaciones se han hecho para cada tipo. La diagonal de la tabla representa las clasificaciones correctas (Timarán et al., 2016).

Si es válido lo anterior, se procede a la explotación del modelo, que es el mantenimiento de la aplicación y la posible difusión de los resultados. (Timarán et al., 2016).

Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio; la retroalimentación generada por la monitorización y mantenimiento puede indicar si el modelo está siendo utilizado apropiadamente (Timarán et al., 2016).

Fase 6. Implementación. Es aquí donde el conocimiento obtenido se transforma en acciones dentro del proceso de negocio, ya sea observando el modelo y resultados, o aplicándolo a múltiples grupos de datos o como parte del proceso. Las tareas que se efectúan son: planear la implementación, monitorizar y mantener, informe final y revisar el proyecto (Timarán et al., 2016).

- Planear la implementación.
- Monitorizar y mantener.
- Informe final.
- Revisar el proyecto.

Anexo Nº7. Análisis Preliminar.

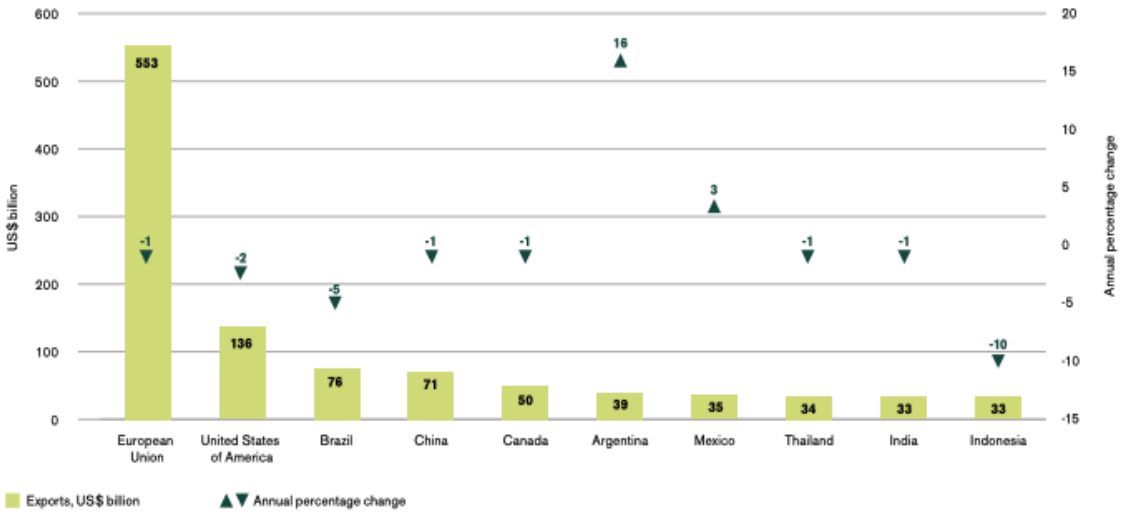
1. Introducción

El presente documento busca analizar los factores cuantitativos del estado en el que se encuentra el sector hortofrutícula.

2. Análisis del sector a nivel global.

Según los datos de la Organización Mundial del Comerio (2019), los países que tuvieron mayor exportación son los mostrados en la Figura 32.

Figura 32. *Países con Mayor Exportación 2019.*



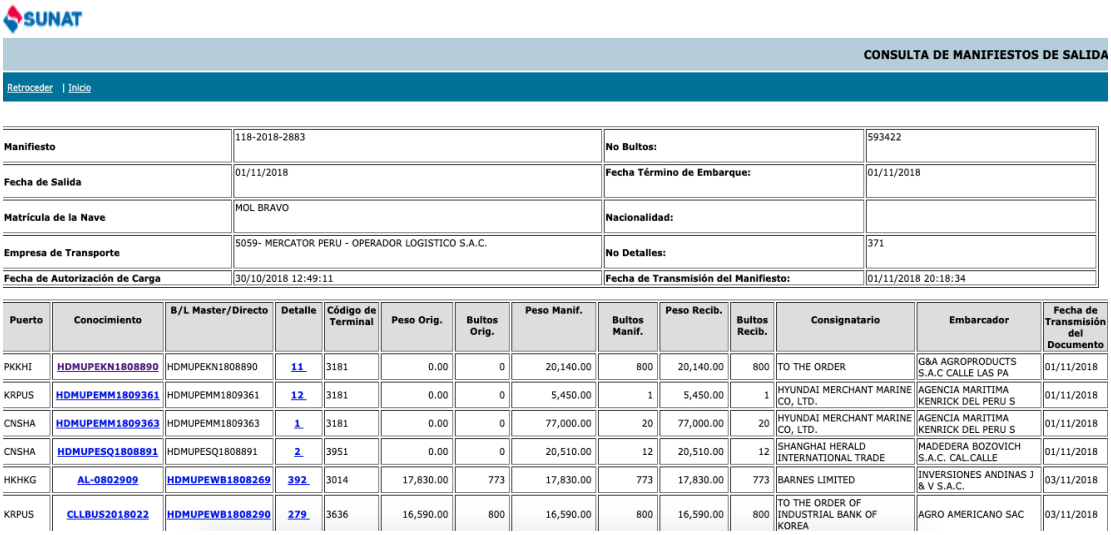
Nota:. Tomado de Exámen estadístico del Comercio Mundial 2019 (OMC, 2019).

3. Análisis del sector en Latinoamérica.

Existe aún un gran espacio diferencial con la recopilación de datos que se da en los países Latinoamericanos con respecto a los de otros continentes. Si bien países como Brasil, Argentina y México cuentan una gran presencia en el rubro exportador, los registros de exportación de los

mismos carecen de puntos importantes para su procesamiento como el poco ordenamiento de los datos y falta de los mismos. Por otro lado, en la mayoría de estos países cuentan con leyes que evitan el fácil acceso a los datos dado que ellos tienen estos procesos regulados. En el caso de Perú, las fuentes de información son aún dispersas siendo la más confiable la encontrada en la Superintendencia Nacional de Aduanas y de Administración (SUNAT), mostrados en la Figura 33; sin embargo estos datos no están para nada procesados, encontrándose data por manifiesto según el día pero sin ningún tipo de capacidad de filtrado que puedan ser usados a niveles estadísticos.

Figura 33. *Formato Sunat.*



SUNAT

CONSULTA DE MANIFIESTOS DE SALIDA

Retroceder | Inicio

Manifiesto	118-2018-2883	**No Bultos:**	593422
Fecha de Salida	01/11/2018	**Fecha Término de Embarque:**	01/11/2018
Matrícula de la Nave	MOL BRAVO	**Nacionalidad:**	
Empresa de Transporte	5059- MERCATOR PERU - OPERADOR LOGISTICO S.A.C.	**No Detalles:**	371
Fecha de Autorización de Carga	30/10/2018 12:49:11	**Fecha de Transmisión del Manifiesto:**	01/11/2018 20:18:34

Puerto	Conocimiento	B/L Master/Directo	Detalle	Código de Terminal	Peso Orig.	Bultos Orig.	Peso Manif.	Bultos Manif.	Peso Recib.	Bultos Recib.	Consignatario	Embarcador	Fecha de Transmisión del Documento
PKKHI	HDMUPEKN1808890	HDMUPEKN1808890	11	3181	0.00	0	20,140.00	800	20,140.00	800	TO THE ORDER	G&A AGROPRODUCTS S.A.C CALLE LAS PA	01/11/2018
KRPUS	HDMUPEMM1809361	HDMUPEMM1809361	12	3181	0.00	0	5,450.00	1	5,450.00	1	HYUNDAI MERCHANT MARINE CO, LTD.	AGENCIA MARITIMA KENRICK DEL PERU S	01/11/2018
CNSHA	HDMUPEMM1809363	HDMUPEMM1809363	1	3181	0.00	0	77,000.00	20	77,000.00	20	HYUNDAI MERCHANT MARINE CO, LTD.	AGENCIA MARITIMA KENRICK DEL PERU S	01/11/2018
CNSHA	HDMUPESQ1808891	HDMUPESQ1808891	2	3951	0.00	0	20,510.00	12	20,510.00	12	SHANGHAI HERALD INTERNATIONAL TRADE	MADEDERA BOZOVICH S.A.C. CAL.CALLE	01/11/2018
HKHKG	AL-0802909	HDMUPEWB1808269	392	3014	17,830.00	773	17,830.00	773	17,830.00	773	BARNES LIMITED	INVERSIONES ANDINAS J & V S.A.C.	03/11/2018
KRPUS	CLLBUS2018022	HDMUPEWB1808290	279	3636	16,590.00	800	16,590.00	800	16,590.00	800	TO THE ORDER OF INDUSTRIAL BANK OF KOREA	AGRO AMERICANO SAC	03/11/2018

Nota: Tomado de http://www.aduanet.gob.pe/cl-ad-itconsmanifiesto/ (SUNAT, 2018).

4. Análisis del sector en Europa, Asia y Estados Unidos.

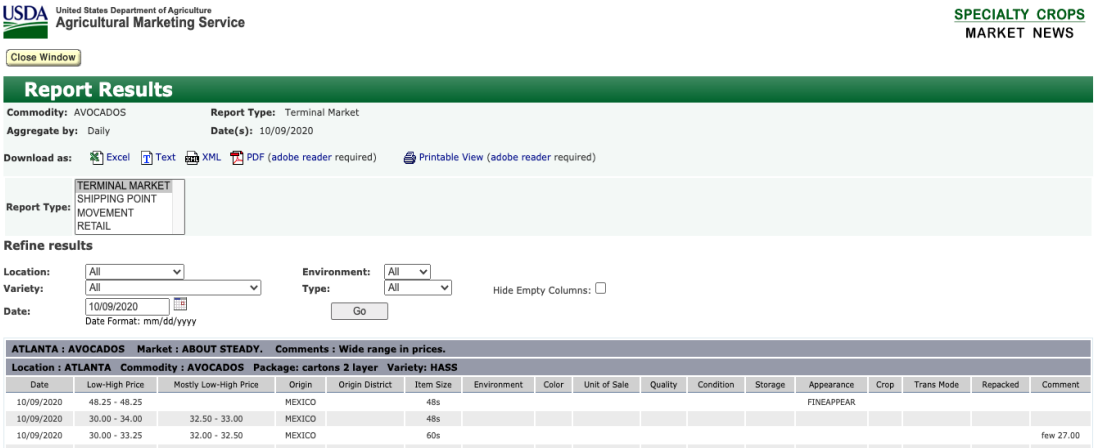
En el mercado europeo, las fuentes llegan a ser un poco más variadas, las más relevantes son las de países como Francia, Inglaterra y España. En este apartado se encuentra un enfoque orientado a datos que siguen formatos diferentes a nivel de nomenclatura, unidades de medida y

claro está una barrera de lenguaje en los mismos, es decir no existe uniformidad entre ellos, haciendo más complicada su recopilación.

China y los países asiáticos se convierten en una imposibilidad principalmente por barreras de idiomas, sin duda contiene datos de mucho valor, pero, por el momento, no es posible juntar los recursos para el acceso de los mismos.

Finalmente se tiene a Estados Unidos, la principal fuente de datos de este país proviene de la United States Department of Agriculture (USDA), los datos que brinda esta institución cumplen con la mayoría de requisitos buscados en el proyecto: gran cantidad de información, son de libre uso y cuentan con datos uniformes y ordenados, tales como muestra en la Figura 34.

Figura 34. *Resultados de reportes USDA.*



Nota: Tomado de https://www.marketnews.usda.gov/ (USDA, 2020).

Anexo Nº 8. Definiciones de algoritmos de minería de datos.

1. Introducción

En el siguiente documento se definen de manera genérica el alcance de las técnicas de minería de datos.

2. Objetivos data mining

Asociación: determinar conjunto de objetos que se dan frecuentemente juntos en el contexto del problema.

Clustering: segmentar una población de objetos heterogénea en un número de grupos más homogéneos o clústeres.

Clasificación: consiste en examinar las características de un objeto nuevo y asignarlo a un conjunto finito predefinido de clases.

Estimación: dado un conjunto de datos de entrada el modelo estima el valor de una magnitud continua desconocida.

Predicción: dado un conjunto de datos de entrada el modelo estima el valor de una magnitud futura.

Descripción: describir las características de un conjunto de objetos en forma de asociaciones significativas o causales entre diferentes variables.

Explicación: determinar las razones de un determinado comportamiento.

Anexo Nº 9. Selección de reportes.

1. Introducción

En el presente documento se busca determinar el tipo de reporte que será usado en la investigación, la USDA brinda tres reportes principales para la palta: SHIPPING POINT, TERMINAL MARKET y MOVEMENT. Se busca determinar cual de los 3 es el adecuado para la investigación.

2. Selección de reportes

SHIPPING POINT:

Muestra los datos generales del envío de un producto como el nombre, origen, variedad, tipo, tamaño, empaquetado, precio, etc. (ver Figura 35) Siendo su principal diferenciador con respecto a los otros el precio de llegada a puerto.

Figura 35. *Resultados de reporte Shipping Point.*

MEXICO CROSSINGS THROUGH TEXAS : AVOCADOS Demand : 32-36S GOOD, 40-48S MODERATE, OTHERS LIGHT. Market : ABOUT STEADY. Basis of Sale : Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis Supply : 32-36S VERY LIGHT, 40S FAIRLY LIGHT. Comments :Extra services included.

Location : MEXICO CROSSINGS THROUGH TEXAS Commodity : AVOCADOS Package: cartons 2 layer Variety: HASS Reporting City: FRESNO,CA

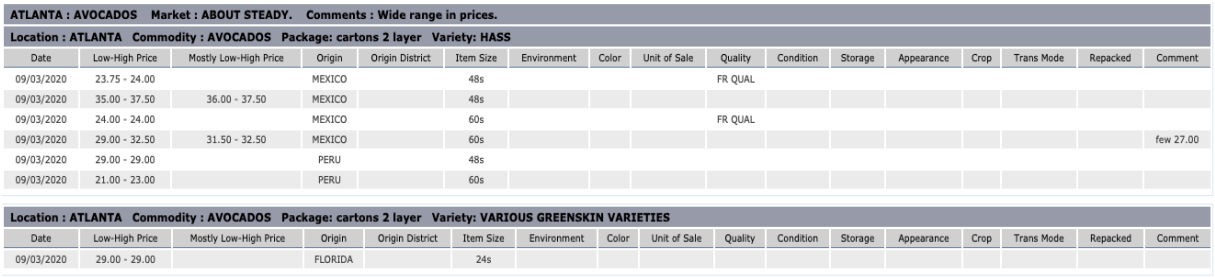
Date	Low-High Price	Mostly Low-High Price	Season	Item Size	Color	Environment	Unit of Sale	Quality	Condition	Storage	Appearance	Import/Export	Comment
09/03/2020	40.25 - 43.25	40.25 - 42.25	2020	32s									
09/03/2020	40.25 - 43.25	40.25 - 42.25	2020	36s									few 38.25 occasional lower
09/03/2020	34.25 - 38.25	36.25 - 37.25	2020	40s									occasional lower
09/03/2020	27.25 - 32.25	29.25 - 31.25	2020	48s									occasional lower
09/03/2020	22.25 - 26.25	24.25 - 25.25	2020	60s									occasional lower
09/03/2020	19.25 - 23.25	20.25 - 21.25	2020	70s									occasional higher
09/03/2020	17.25 - 21.25	18.25 - 20.25	2020	84s									

Nota: Tomado de https://www.marketnews.usda.gov (USDA, 2020).

TERMINAL MARKET:

Terminal Market muestra datos como nombre, variedad, tipo, (ver Figura 36) pero la diferencia es que sus precios son del mercado terminal es decir los destinos finales, siendo estos variados, puesto que se fijan acorde a factores menos confiables.

Figura 36. *Resultados de reporte Terminal Market.*



ATLANTA : AVOCADOS Market : ABOUT STEADY. Comments : Wide range in prices.

Location : ATLANTA Commodity : AVOCADOS Package: cartons 2 layer Variety: HASS

Date	Low-High Price	Mostly Low-High Price	Origin	Origin District	Item Size	Environment	Color	Unit of Sale	Quality	Condition	Storage	Appearance	Crop	Trans Mode	Repacked	Comment
09/03/2020	23.75 - 24.00		MEXICO		48s				FR QUAL							
09/03/2020	35.00 - 37.50	36.00 - 37.50	MEXICO		48s											
09/03/2020	24.00 - 24.00		MEXICO		60s				FR QUAL							
09/03/2020	29.00 - 32.50	31.50 - 32.50	MEXICO		60s											few 27.00
09/03/2020	29.00 - 29.00		PERU		48s											
09/03/2020	21.00 - 23.00		PERU		60s											

Location : ATLANTA Commodity : AVOCADOS Package: cartons 2 layer Variety: VARIOUS GREENSKIN VARIETIES

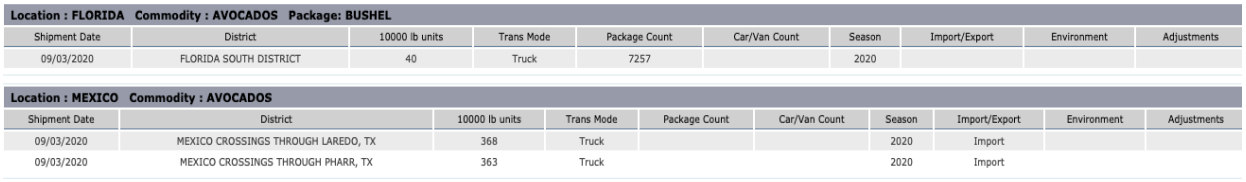
Date	Low-High Price	Mostly Low-High Price	Origin	Origin District	Item Size	Environment	Color	Unit of Sale	Quality	Condition	Storage	Appearance	Crop	Trans Mode	Repacked	Comment
09/03/2020	29.00 - 29.00		FLORIDA		24s											

Nota: Tomado de https://www.marketnews.usda.gov (USDA, 2020).

MOVEMENT:

Movement tiene como atributo principal la cantidad de producto en kg que ha sido exportado en diversas fechas, como se ve en la Figura 37.

Figura 37. *Resultados de reporte Movement.*



Location : FLORIDA Commodity : AVOCADOS Package: BUSHEL

Shipment Date	District	10000 lb units	Trans Mode	Package Count	Car/Van Count	Season	Import/Export	Environment	Adjustments
09/03/2020	FLORIDA SOUTH DISTRICT	40	Truck	7257		2020			

Location : MEXICO Commodity : AVOCADOS

Shipment Date	District	10000 lb units	Trans Mode	Package Count	Car/Van Count	Season	Import/Export	Environment	Adjustments
09/03/2020	MEXICO CROSSINGS THROUGH LAREDO, TX	368	Truck			2020	Import		
09/03/2020	MEXICO CROSSINGS THROUGH PHARR, TX	363	Truck			2020	Import		

Nota: Tomado de https://www.marketnews.usda.gov (USDA, 2020).

Concluyendo, tomando en cuenta que es el que contiene datos más variados y un precio más preciso, se elegirá para el estudio el formato de SHIPPING POINT.

Anexo Nº10. Reporte de descripción de datos.

Los datos adquiridos en un reporte estándar de Shipping point son los siguientes:

- Commodity: Producto
- City Name: Origen del producto.
- Type: Tipo.
- Package: Empaquetado.
- Variety: Variedad.
- Sub Variety: Sub Variedad.
- Grade: Grado.
- Date: Fecha.
- Low Price: Precio más bajo del producto.
- High Price: Precio más alto del producto.
- Mostly Low: Precio mayormente bajo.
- Mostly: Precio mayormente alto.

En la tabla 10, se puede identificar el formatio inicial del reporte con los datos respectivos.

Tabla 10. *Formato inicial de Reporte.*

Commodity Name	City Name	Type	Package	Variety	Sub Variety	Grade	Date	Low Price	High Price	Mostly Low	Mostly High
AVOCADO	MEXICO CROSSINGS THROUGH TEXAS		cartons 2 layer	HASS			9/03/20	40.25	43.25	40.25	42.25

Anexo Nº 11. Reporte Estándar Shipping Point.

Figura 38. *Reporte Estándar Shipping Point USDA.*

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Variety	Date	Low Price	High Price	Mostly Low	Mostly High	Season	Item Size	Supply Tone	Demand Tone	Basis of Sale	Market Tone	Price Comment	Comments	Rpt City
2	HASS	10/02/20	27.25	32.25	28.25	30.25	2020	32s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional higher and lower	Extra services included.	FRESNO, CA
3	HASS	10/02/20	27.25	32.25	28.25	30.25	2020	36s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional higher and lower	Extra services included.	FRESNO, CA
4	HASS	10/02/20	20.25	25.25	22.25	24.25	2020	40s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional higher	Extra services included.	FRESNO, CA
5	HASS	10/02/20	20.25	24.25	22.25	23.25	2020	48s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional lower	Extra services included.	FRESNO, CA
6	HASS	10/02/20	21.25	26.25	22.25	23.25	2020	60s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.		Extra services included.	FRESNO, CA
7	HASS	10/02/20	21.5	25.25	22.25	23.25	2020	70s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional lower	Extra services included.	FRESNO, CA
8	HASS	10/02/20	18.25	23.25	20.25	21.25	2020	84s		LIGHT, OTHERS MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S SLIGHTLY HIGHER, 32-36S LOWER, OTHERS ABOUT STEADY.	occasional higher	Extra services included.	FRESNO, CA
9	HASS	10/02/20	32.25	35.25	32.25	34.25	2020	32s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.		Extra services included. Wide range in prices.	FRESNO, CA
10	HASS	10/02/20	32.25	35.25	32.25	34.25	2020	36s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.		Extra services included. Wide range in prices.	FRESNO, CA
11	HASS	10/02/20	32.25	36.25	32.25	34.25	2020	40s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.		Extra services included. Wide range in prices.	FRESNO, CA
12	HASS	10/02/20	30.25	35.25	31.25	32.25	2020	48s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.		Extra services included. Wide range in prices.	FRESNO, CA
13	HASS	10/02/20	25.25	29.25	26.25	28.25	2020	60s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.	occasional higher	Extra services included. Wide range in prices.	FRESNO, CA
14	HASS	10/02/20	23.25	26.25	24.25	26.25	2020	70s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.		Extra services included. Wide range in prices.	FRESNO, CA
15	HASS	10/02/20	20.25	22.25			2020	84s	FAIRLY LIGHT.	MODERATE.	Sales F.O.B. Shipping Point and/or Delivered Sales, Shipping Point Basis	60S HIGHER, OTHERS ABOUT STEADY.	occasional higher	Extra services included. Wide range in prices.	FRESNO, CA

report - 2020-10-11T034701.333

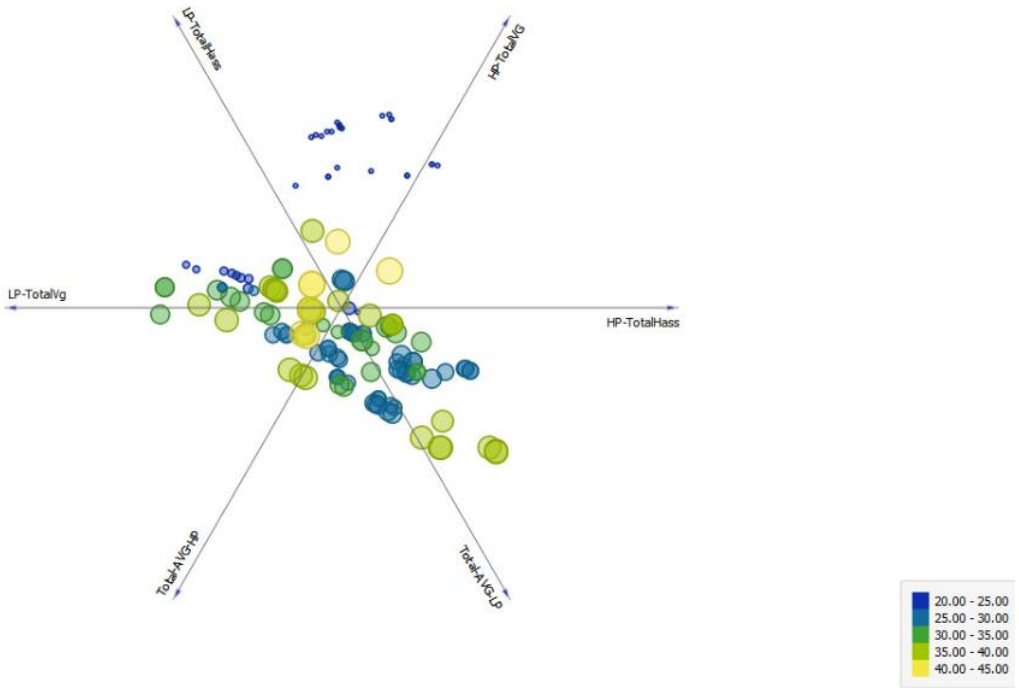
Nota: Tomado de https://www.marketnews.usda.gov (USDA, 2020).

Anexo Nº 12. Reporte de calidad de los datos Nº1.

1. Análisis de calidad de datos a través de Linear Projection

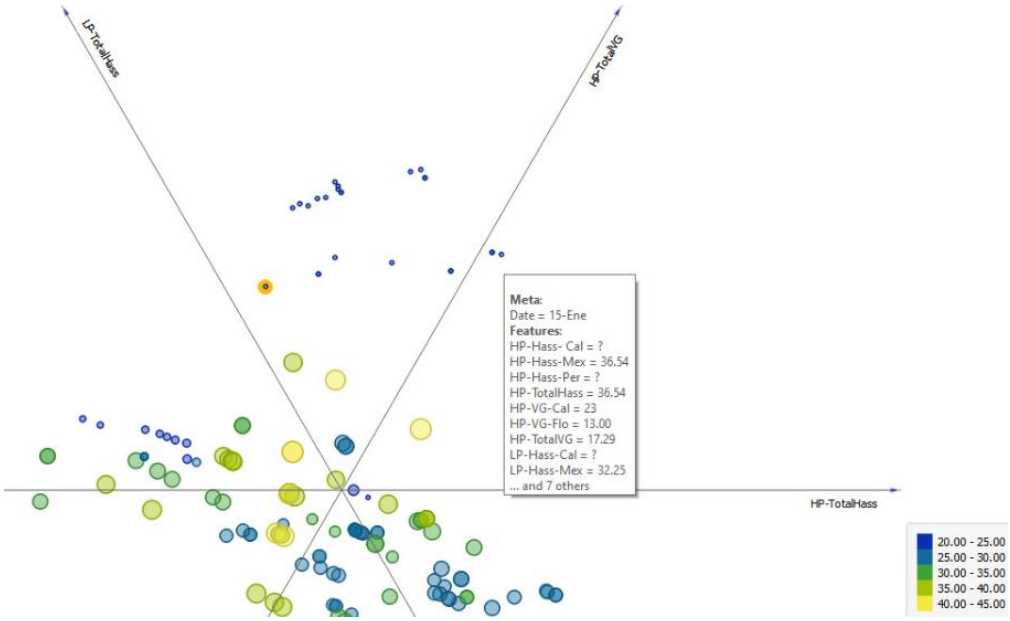
Se identificó que existe presencia de datos confusos o faltantes, como se puede ver en la data representada en la Figura 39.

Figura 39. *Linear Projection de un Reporte.*



Por otro lado en la Figura 40, se puede ver como existen datos que se alejan tanto en cantidad como en proporción de la data más homogénea.

Figura 40. *Linear Projection Extra Data.*



Se concluye que en ambas figuras existe la oportunidad de identificar y obviar los datos que no aportan de manera objetiva al proceso de modelado.

Anexo Nº 13. Reporte de selección, estructurado e integración de datos.

1. Introducción

En el presente documento se busca procesar los datos para poder establecer un nuevo formato de reporte, los procesos de selección, estructurado e integración han sido unificados para facilitar el entendimiento del la fase de preparación de datos.

2. Selección ,estructurado e integración de datos

Un reporte de palta cuenta con una gran cantidad de data, que puede ser interpretada de diversas maneras, pero el uso al que se le dio en este modelo está orientado al precio. En este punto existen 2 cosas a tomar en cuenta. El primero es que el objetivo es determinar el precio pero en Kg, el reporte lo da basado en Lb, pués este es el estándar en Estados Unidos. Por otro lado, se busca identificar en la variedad de precios, cual de estos tiene más valor con respecto a la información del reporte. A continuación se muestra el proceso.

Un reporte estándar de precio esta expresado en la Tabla 11, de la siguiente manera:

Tabla 11. *Reporte Estándar de Precio.*

Commodity Name	City Name	Package	Variety	Date	Low Price	High Price	Mostly Low	Mostly High
AVOCADO	MEXICO CROSSINGS THROUGH TEXAS	cartons 2 layer	HASS	9/03/20	40.25	43.25	40.25	42.25

Como se explicó, se debe cambiar el precio que está establecido en Libras a Kilogramos. Esto se realiza basándose en el empaquetado de la palta, éste se da como se muestra en la Tabla 12:

***Tabla 12**. Empaquetado de Paltas.*

	
Cartons 2 layer	Cartons 1 layer
Apróx. 11,34 lb	Apróx. 5,66 lb

Según estos valores estableceremos la siguiente fórmula:

Kg Price (Cartons 2 layer) = [(Low Price + High Price)/2]/11.34]

Kg Price (Cartons 1 layer) = [(Low Price + High Price)/2]/5.66]

Se ha definido el estándar para precio en Kg, éste es en dólares pero por motivos de procesado de datos numéricos se obvió el símbolo (ver tabla 13). El siguiente paso es realizar una reducción de la data basada en la misma:

***Tabla 13.** Reporte con conversión a Kg.*

Commodity Name	City Name	Package	Variety	Package Weight	Date	Low Price	High Price	Mostly Low	Mostly High	Price Kg
AVOCADOS	MEXICO CROSSINGS THROUGH TEXAS	cartons 2 layer	HASS	11.66	9/03/20	40.25	43.25	40.25	42.25	3.5806175

Los datos de precios varían en cantidad dependiendo de muchos factores, no todos pueden ser utilizados objetivamente. Se tiene dato como Mostly Low y Mostly High que si bien son de gran valor presentan una deficiencia al momento de estar presente en la mayoría de reportes. Package cumple con la función principal de brindar el precio en Kg, aparte de eso el dato es meramente representativo, al igual que Commodity Name. Tomando en cuenta estos puntos, se identificaron como datos de valor a los presentes en la tabla 14:

Tabla 14. *Formato de reporte con datos principales.*

City Name	Variety	Date	Price Kg
MEXICO CROSSINGS THROUGH TEXAS	HASS	9/03/20	3.5806175

Por motivos visuales, el precio fue redondeado a dos decimales, y para que exista mayor entendimiento, se cambió la nomenclatura de City Name a Origin().

Finalmente, para la predicción de datos, se tendra en cuenta un campo llamado Return que utilizará la fórmula de:

If((PrecioMañana-PrecioHoy)>0) {Aumento = True} else {Aumento=False}

Este campo no estará presente en el formato final, pero se agregará al procesar el reporte, con el objetivo de tener una columna con datos validados y referenciales para nuestra predicción.

Una vez pasado todo este proceso, se tiene toda la información necesaria para recrear un formato de reporte final, el cual es el de la Tabla 15.

Tabla 15. *Formato final.*

Origin	Variety	Date	Price Kg
MEXICO	HASS	9/03/20	3.58

Anexo Nº 14. Reporte de calidad de datos Nº2.

1. Análisis de calidad de datos a través de Linear Projection

Se termino definiendo el formato final de reporte más la columna Return,como se muestra en la Figura 41.

Figura 41. *Formato de Reporte Final más Return.*

	A	B	C	D	E	F	G	H	I	J
1	Date	H-Mex-Price	H-Per-Price	H-Cal-Price	TotalHass	VG-Car-Price	VG-Flor-Price	TOTALVG	TotalAvgPrice	Return
2	1-Ene	1.92			1.92	1.44	1.68	1.56	1.71	1
3	2-Ene	1.92			1.92	1.44	1.69	1.55	1.71	1
4	3-Ene	2.02			2.02	1.44	1.69	1.55	1.76	1
5	6-Ene	2.09			2.09	1.70	1.69	1.70	1.88	1
6	7-Ene	2.09			2.09	1.78	1.69	1.74	1.90	0
7	8-Ene	2.09			2.09	1.78	1.69	1.74	1.90	1
8	9-Ene	2.12			2.12	1.78		1.78	2.00	0
9	10-Ene	2.12			2.12	1.78		1.78	2.00	1
10	13-Ene	2.15			2.15	1.74		1.74	2.00	0
11	14-Ene	2.15			2.15	1.74		1.74	2.00	0
12	15-Ene	2.15			2.15	1.74		1.74	2.00	1
13	16-Ene	2.15			2.15	1.78		1.78	2.02	0
14	17-Ene	2.15			2.15	1.78		1.78	2.02	0
15	21-Ene	2.06			2.06	1.78		1.78	1.96	0
16	22-Ene	2.01			2.01	1.78		1.78	1.92	0
17	23-Ene	1.95			1.95	1.78		1.78	1.89	0
18	24-Ene	1.90			1.90	1.78		1.78	1.86	1
19	27-Ene	2.13			2.13	1.82		1.82	2.04	0
20	28-Ene	2.12			2.12	1.82		1.82	2.04	0
21	29-Ene	2.12			2.12	1.82		1.82	2.04	0
22	30-Ene	2.12			2.12	1.82		1.82	2.04	0
23	31-Ene	2.12			2.12	1.82		1.82	2.04	1
24	3-Feb	2.13			2.13	1.95		1.95	2.07	0
25	4-Feb	2.13			2.13	1.95		1.95	2.07	0

Nota : Donde: Date = Fecha, H-Mex-Price=Precio Hass México, H-Cal-Price=Precio Hass California, TotalHass=Precio Total Hass, VG-Car-Price=Precio Various Green Skin Caribean , VG-Flor-Price=Precio Various Green Skin Florida, TotalVG= Total Various Green Skin, TotalAvgPrice=Precio Promedio Total, Return=Aumento o disminución.

Se utiliza Linear Projection, dado que permite comprobar la diferencia entre calidad y cantidad de datos. Tanto en la figura 42 y figura 43 se puede identificar una mejora en la

homogeneidad de los datos, todo estos resultados se dan después del proceso de reestructurado de los datos.

Figura 42. *Linear Projection Variety Price.*

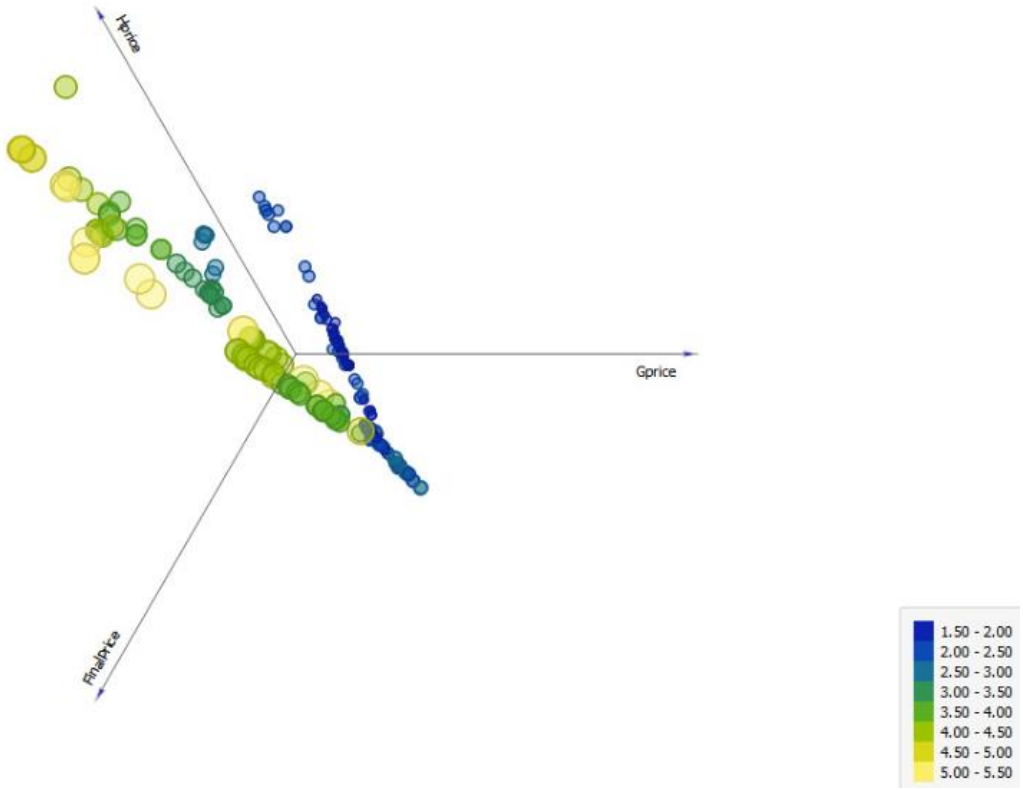
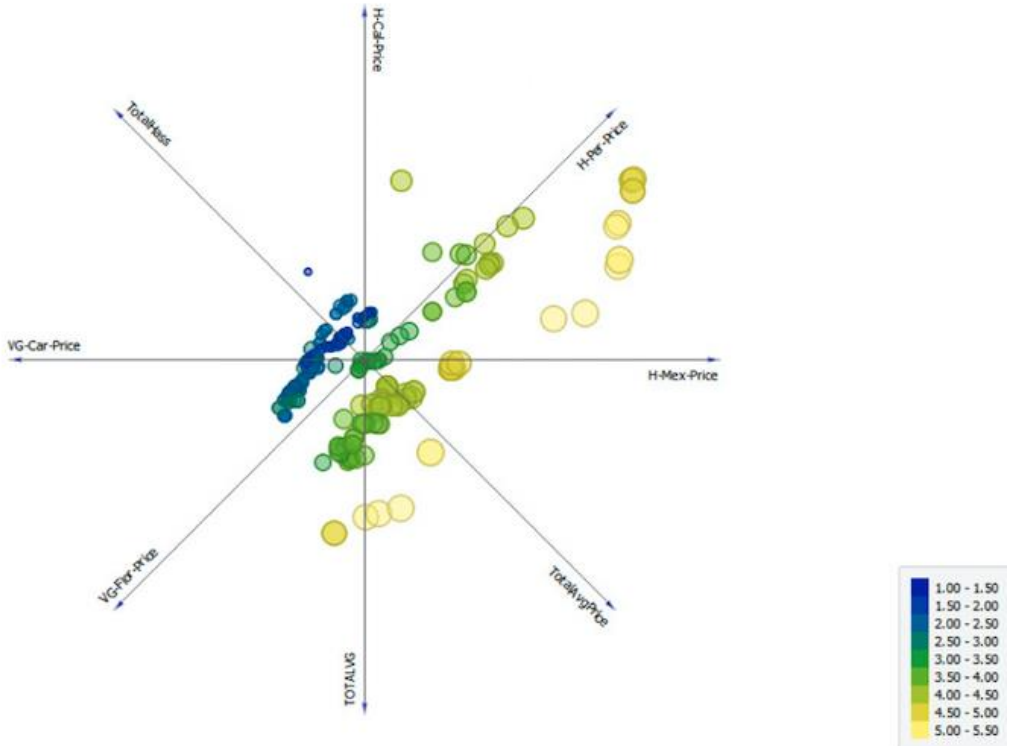


Figura 43. *Linear Projection Origin and Variety Price.*



Anexo Nº 15. Selección de Metricas.

1. Matriz de Confusión

Una matriz de confusión cuenta con sus respectivas métricas, estas se pueden ver en la Figura 44:

Figura 44. *Matriz de Confusión y Métricas*

<table>
<tr><td colspan="2" rowspan="2">Matriz de confusión</td><td colspan="2">Estimado por el modelo</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Negativo (N)</td><td>Positivo (P)</td></tr>
<tr><td rowspan="2">Real</td><td>Negativo</td><td>a: (TN)</td><td>b: (FP)</td><td></td><td></td></tr>
<tr><td>Positivo</td><td>c: (FN)</td><td>d: (TP)</td><td>Precisión
("precision")
Porcentaje predicciones positivas correctas:</td><td>d/(b+d)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Sensibilidad, exhaustividad
("Recall")
Porcentaje casos positivos detectados</td><td>Especifidad
(Specifity)
Porcentaje casos negativos detectados</td><td colspan="2">Exactitud
("accuracy")
Porcentaje de predicciones correctas
(No sirve en datasets poco equilibrados)</td></tr>
<tr><td>d/(d+c)</td><td>a/(a+b)</td><td colspan="2">(a+d)/(a+b+c+d)</td></tr>
</table>

Según Recuero de los Santos (2020), la métricas más importantes son:

La exactitud (accuracy) , que representa el porcentaje de predicciones correctas frente al total. Por tanto, es el cociente entre los casos bien clasificados por el modelo (verdaderos positivos y verdaderos negativos, es decir, los valores en la diagonal de la matriz de confusión), y la suma de todos los casos.

Sin embargo, cuando un conjunto de datos es poco equilibrado, no es una métrica útil. Por ejemplo, si se intenta predecir una enfermedad rara, y el algoritmo clasifica a todos los individuos como sanos, podría ser muy preciso (incluso un 99%), pero también, totalmente inútil.

Por ello, en estos casos se suele recurrir a otras métricas, como la sensibilidad (o recall), que representa la habilidad del modelo de detectar los casos relevantes.

La precisión, (precision) se refiere a lo cerca que está el resultado de una predicción del valor verdadero. Por tanto, es el cociente entre los casos positivos bien clasificados por el modelo y el total de predicciones positivas.

La sensibilidad o exhaustividad (recall) representa la tasa de verdaderos positivos (True Positive Rate) ó TP. Es la proporción entre los casos positivos bien clasificados por el modelo, respecto al total de positivos.

La especificidad, por su parte, es la tasa de verdaderos negativos, ("true negative rate")o TN. Es la proporción entre los casos negativos bien clasificados por el modelo, respecto al total de negativos.

Para concluir, la conveniencia de usar una métrica u otra como medida, dependerá de cada caso en particular. En este caso en concreto, el coste asociado a cada error de clasificación del algoritmo no es tan relevante como analizar el porcentaje de precisión de las mismos, dado que es una primera interacción con los algoritmos. Por lo que se decidio priorizar la métricas de exactitud y precisión en el estudio.

Anexo Nº 16. Resultados obtenidos del modelado predictivo.

Se obtuvieron los resultados solamente diarios (Daily Only) del año 2019 utilizando Matrices de Confusión, estos se muestran desde la Figura 45 hasta la 50.

Figura 45. *Daily Only Confusion Matrix SVM.*

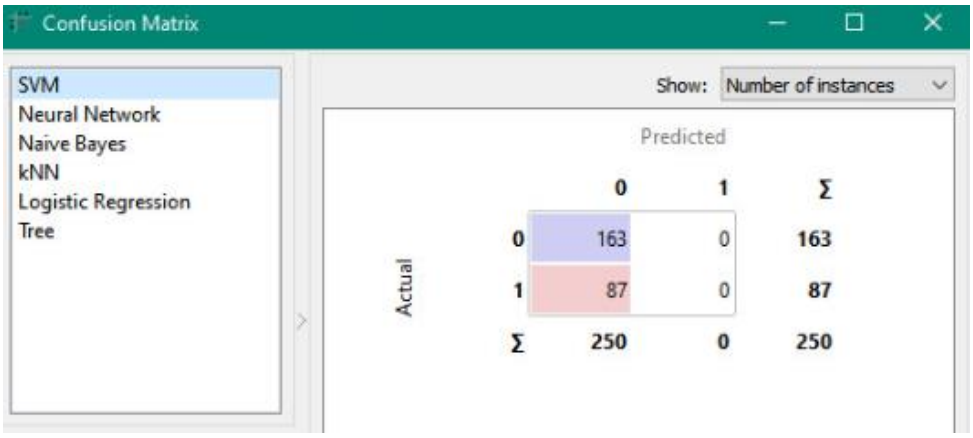


Figura 46. *Daily Only Confusion Matrix Neural Network.*

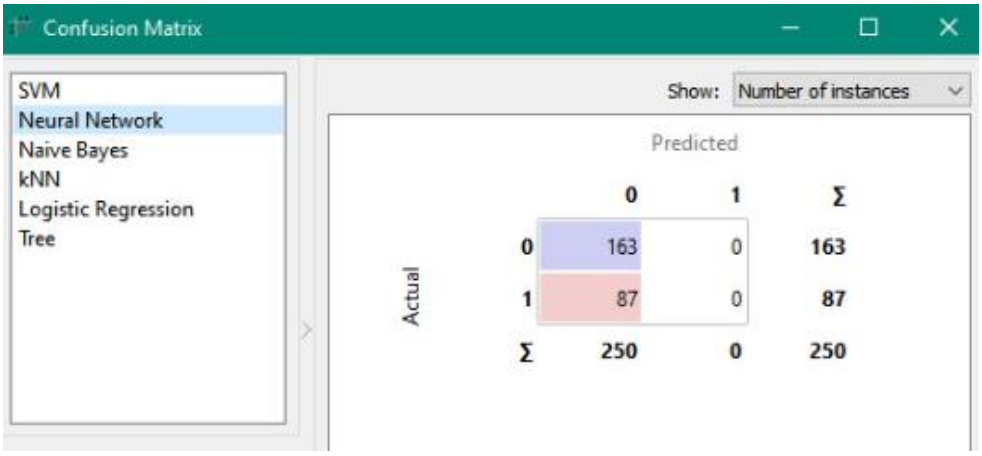


Figura 47. *Daily Only Confusion Matrix Naive Bayes.*

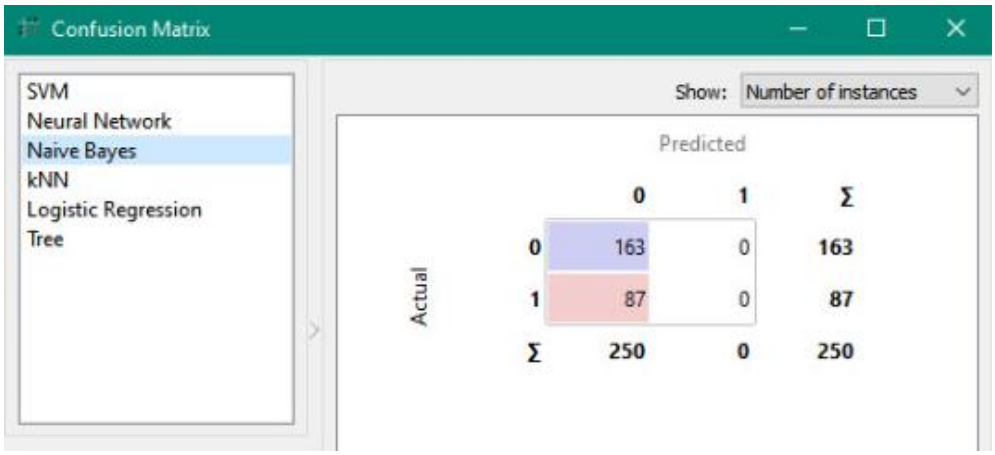


Figura 48. *Daily Only Confusion Matrix kNN.*

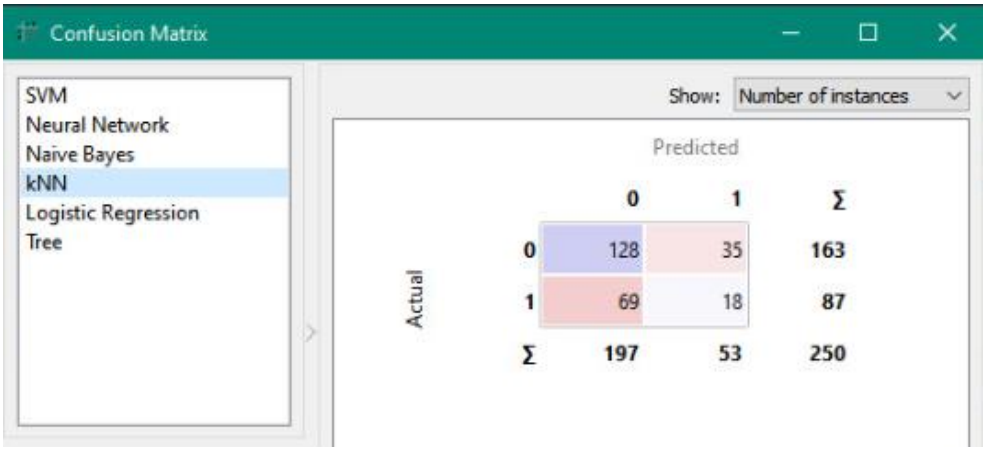


Figura 49. *Daily Only Confusion Matrix Logistic Regresion.*

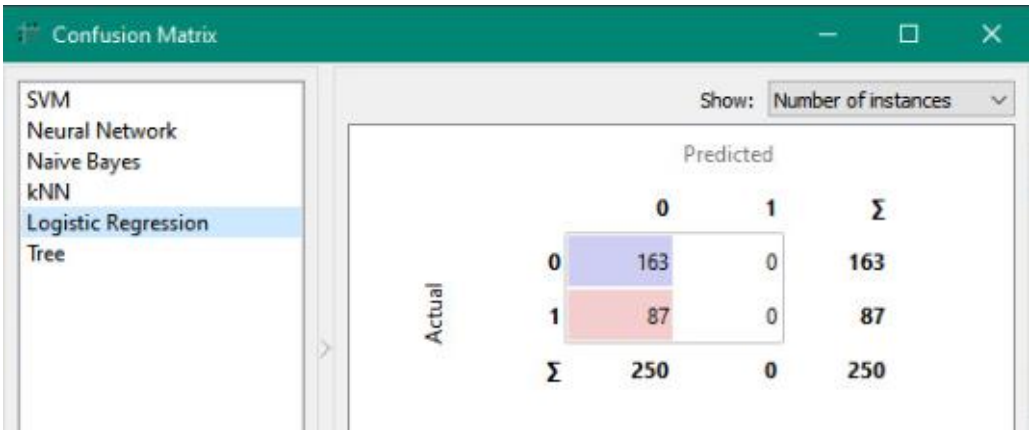
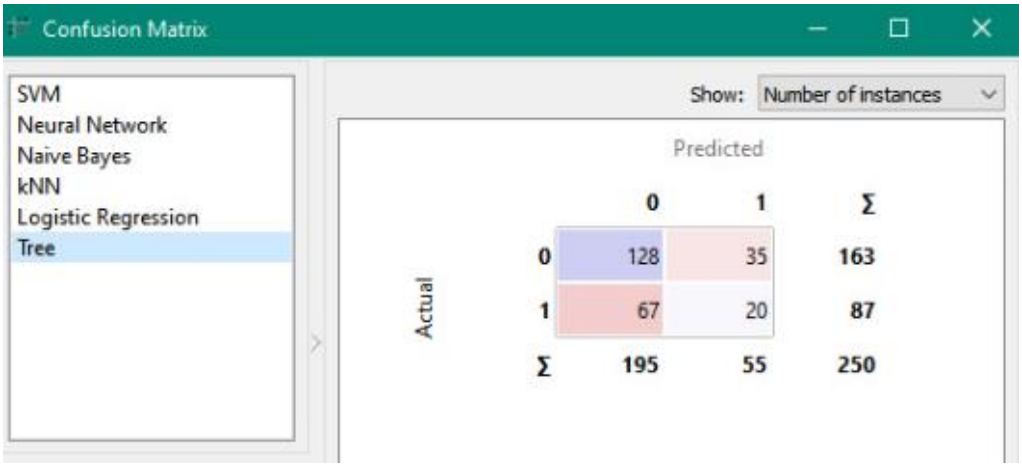


Figura 50*. Daily Only Confusion Matrix Tree.*



Se obtuvieron los resultados de variedad por precio (Variety Price) del año 2019 utilizando Matrices de Confusión, estos se muestran desde la Figura 51 hasta la 56.

Figura 51. *Variety Price Confusion Matrix kNN.*

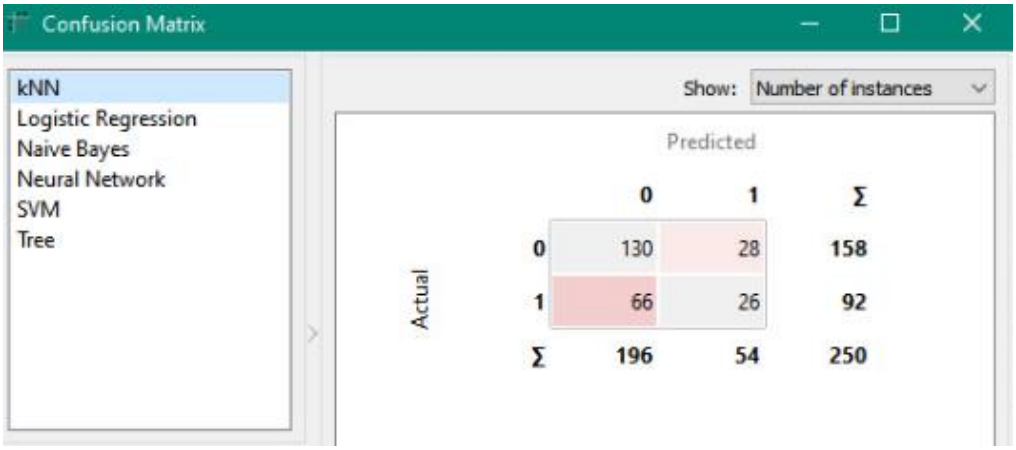


Figura 52. *Variety Price Confusion Matrix Logistic Regression.*

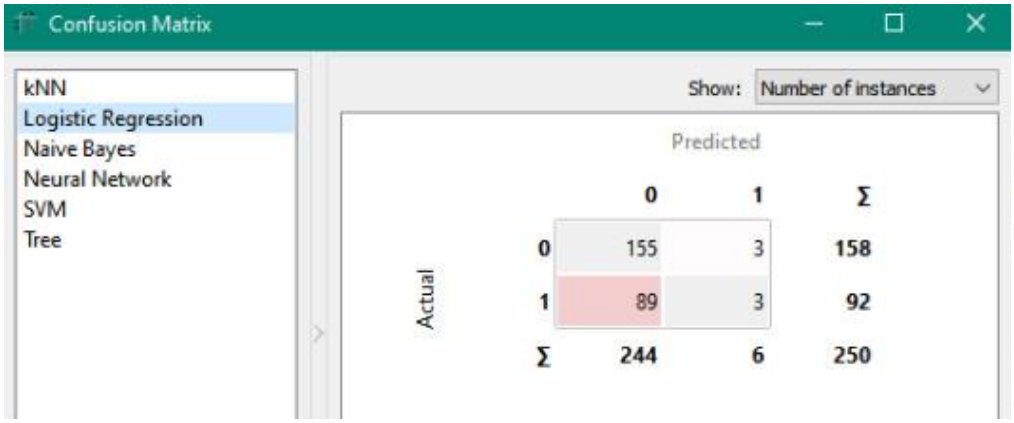


Figura 53. *Variety Price Confusion Matrix Naive Bayes.*

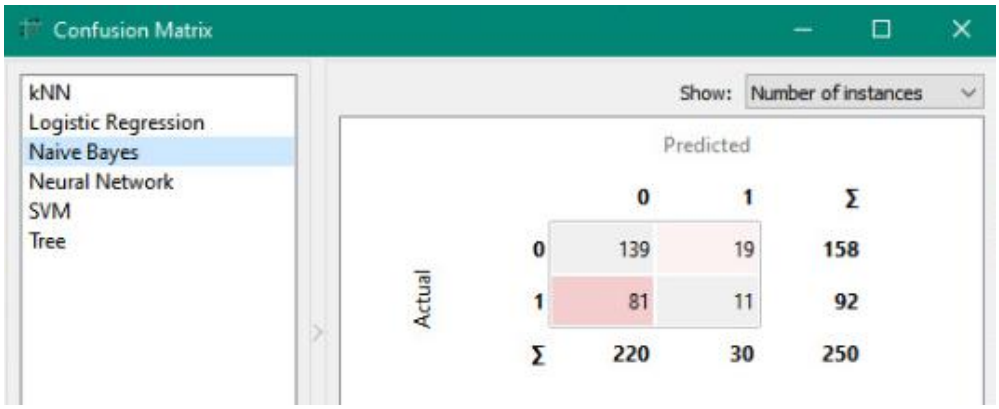


Figura 54. *Variety Price Confusion Matrix Neural Network.*

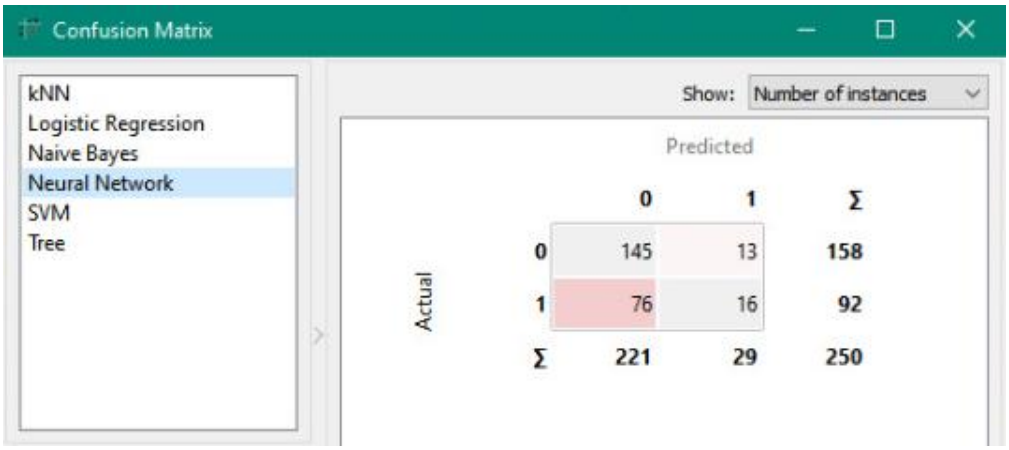


Figura 55. *Variety Price Confusion Matrix SVM.*

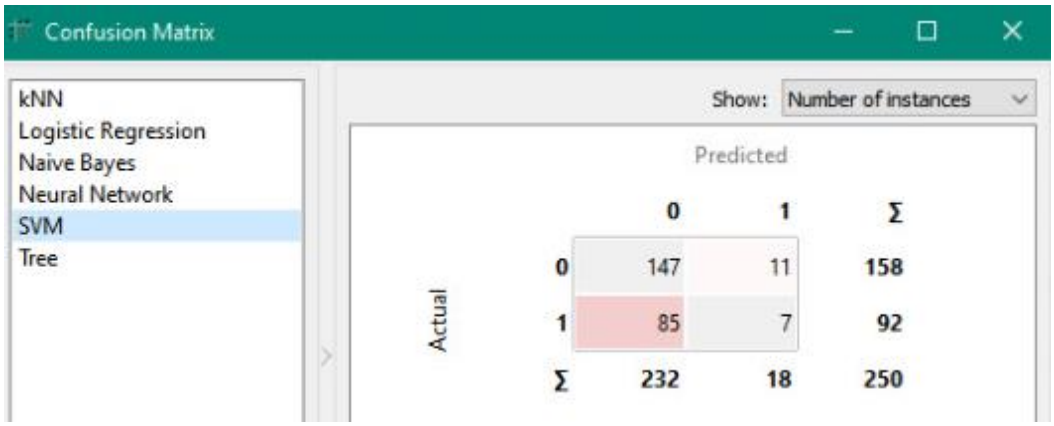
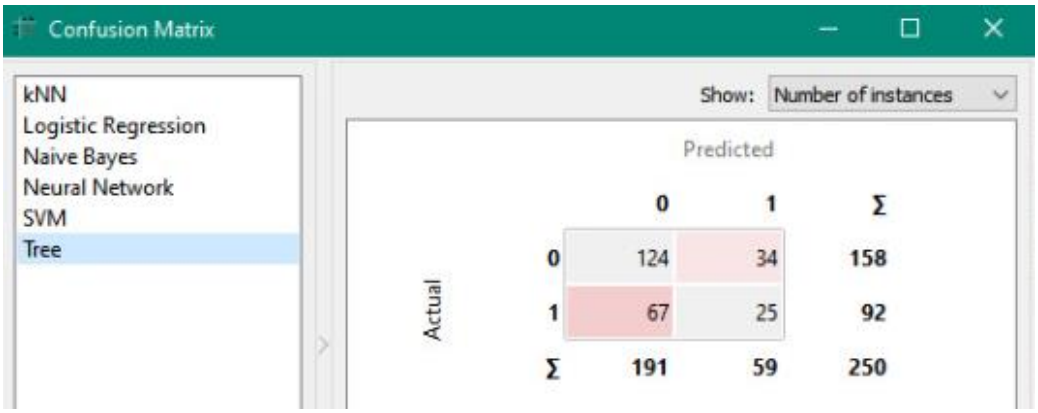


Figura 56. *Variety Price Confusion Matrix Tree.*



Se obtuvieron los resultados solamente diarios de variedad y precio (Daily Variety and Origin Price) del año 2019 utilizando Matrices de Confusión, estos se muestran desde la Figura 57 hasta la 62.

Figura 57. *Variety and Origin Price Confusion Matrix kNN.*

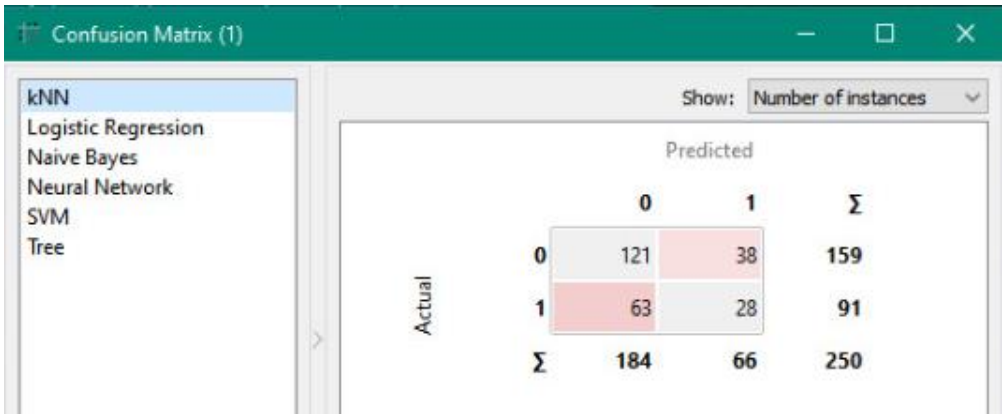


Figura 58. *Variety and Origin Price Confusion Matrix Logistic Regression.*

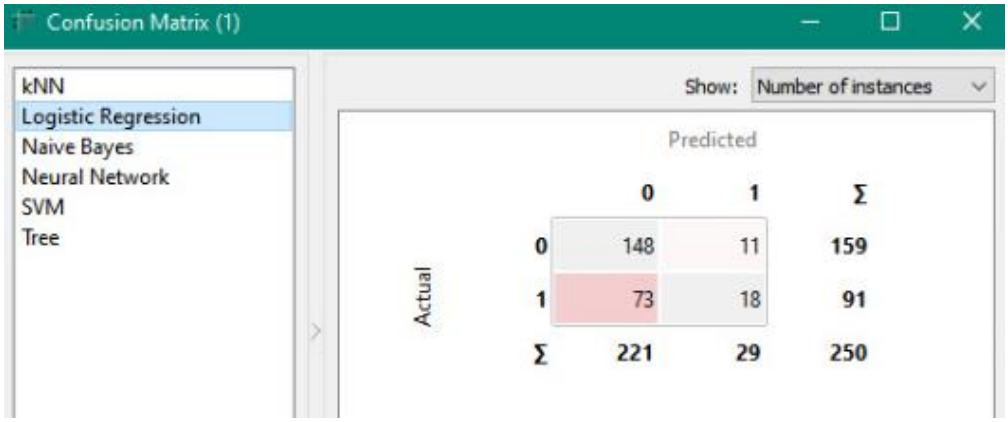


Figura 59. *Variety and Origin Price Confusion Matrix Naive Bayes.*

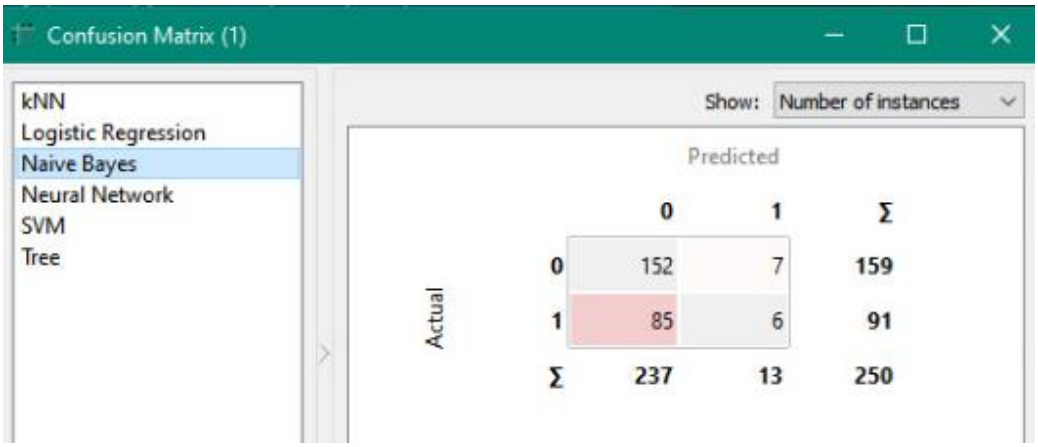


Figura 60. *Variety and Origin Price Confusion Matrix Neural Network.*

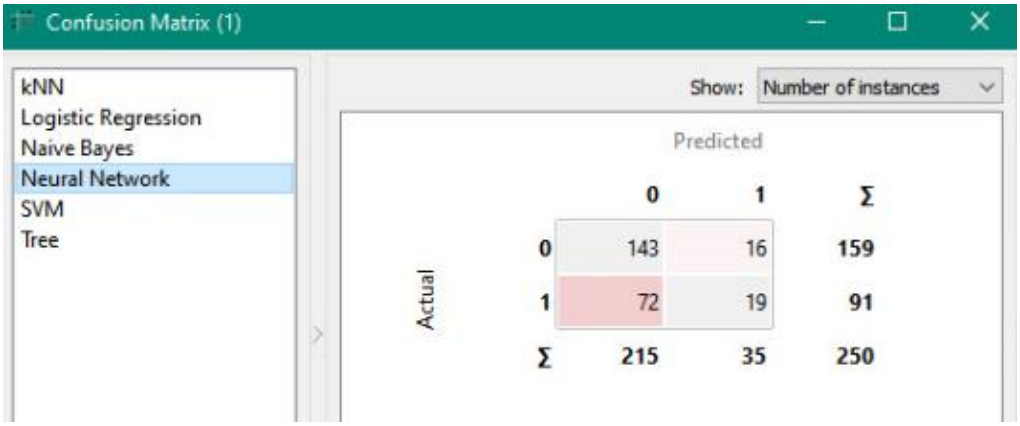


***Figura 61**. Variety and Origin Price Confusion Matrix SVM.*

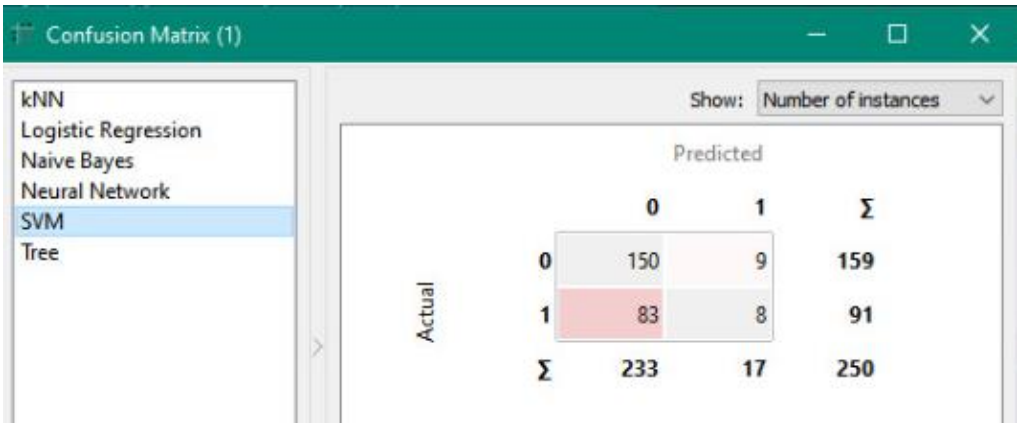
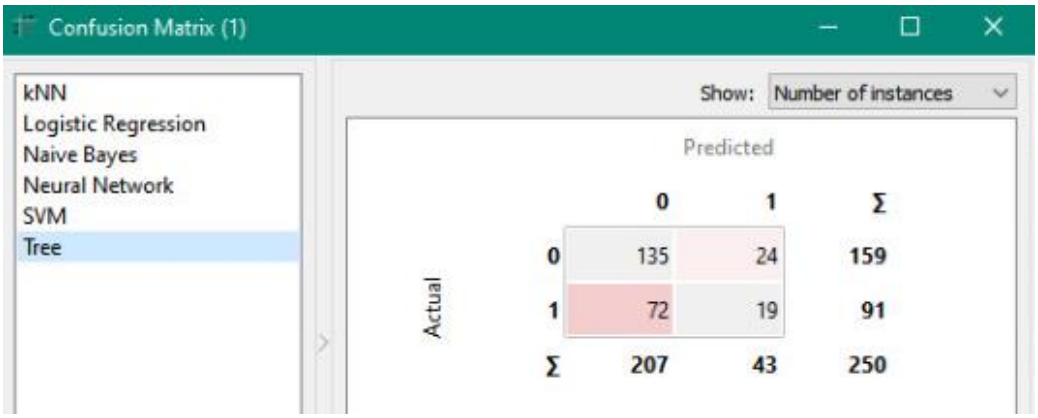


***Figura 62.** Variety and Origin Price Confusion Matrix Tree.*



Anexo Nº 18. Informe comparativo de resultados.

1. Introducción:

En el presente informe se agrupan los resultados obtenidos para cada algoritmos y se busca evaluar cual de ellos presento una mayor eficacia. Esto se dará a través de la fórmula:

Eficacia =(Resultado Obtenido * 100)/(Resultado Previsto)

Resultado obtenido = True positive+ True Negative

Resultado previsto = Cantidad de datos totales

Las filas en los reportes son de 250 cada una de ellas con diversos campos, estos son: Daily Price Only (500 campos), DailyVariety (750 campos) y DailyOriginVariety (1000 campos). Con esta variación en los campos, se busca determinar si estos influyen en el aumento o disminución.

2. Resultados con mayor eficacia:

Tabla 16. *Comparación Daily Only.*

	True Negative (0 0)	False Positive (0 1)	False Negative (1 0)	True Positive (1 1)
SVM	163	0	87	0
Neural Network	163	0	87	0
Naive Bayes	163	0	87	0

kNN	128	35	69	18
Logistic Regression	163	0	87	0
Tree	128	35	67	20

Analizando la Tabla 16, se identifica que solo existen valores de True Positive (Aumento) en los algoritmos de kNN y Tree. El True Negative (Disminución) está presente en todos, siendo el de mayor valor el de 163 presente en SVM, Neural Network, Naive Bayes y Logistic Regression. Estos 4 algoritmos terminaron siendo los más eficaces pero con resultados de poco valor. Esto puede ser debido a la cantidad de campos presentes en el reporte o características propias de los algoritmos.

$$\%Eficacia = 163*100/250 = 65,2\%$$

Tabla 17. *Comparación Variety Price.*

	True Negative (0 0)	False Positive (0 1)	False Negative (1 0)	True Positive (1 1)
SVM	147	11	85	7
Neural Network	145	13	76	16
Naive Bayes	139	19	81	11
kNN	130	28	66	26
Logistic Regression	155	3	89	3

Tree	124	34	67	25

En la Tabla 17, el algoritmo kNN presente un mayor número de valores True Positive, sin embargo, el de Neural Network terminó siendo el que tiene la eficacia más alta.

%Eficacia = 161*100/250 = 64,4%

***Tabla 18**. Comparación Origin and Variety*

	True Negative (0 0)	False Positive (0 1)	False Negative (1 0)	True Positive (1 1)
SVM	150	9	83	8
Neural Network	143	16	72	19
Naive Bayes	152	7	85	6
kNN	121	38	63	28
Logistic Regression	148	11	73	18
Tree	135	24	72	19

En la Tabla 18, de nuevo el algoritmo kNN presento un aumento en True Positives, y Logistic Regression continuó siendo el que presenta mayor valor de eficacia.

%Eficacia = 158*100/250 = 66,4%

3. Conclusiones:

Se puede concluir que:

- El algoritmo más eficaz terminó siendo el de Logistic Regression con un porcentaje de 66.4%.
- La eficacia de los algoritmos varió solo en 1 o 2 %. Donde verdaderamente se encuentra una diferencia, es en las predicciones de aumento o disminución, que terminan variando de distintas maneras dependiendo de cada algoritmo.
- La variación en la cantidad de campos utilizados influye en predecir el aumento más no la disminución, la demostración de esto se vio de manera más evidente en el algoritmo kNN.
- La precision aumento en el reporte con más campos, lo cual es evidencia que a mayor cantidad de datos existe un aumento en la precisión.

Anexo Nº 19. Reporte de últimos 21 meses.

En el presente reporte se representa de manera gráfica la fluctuación de los últimos meses por precio, origen y variedad del producto Palta (Avocado). El propósito de este reporte es de manera ilustrativa, para visualizar como se ven los datos con la herramienta Highcharts.

Lo parámetros finales están representados en: la Figura 30, últimos 21 meses por precio; en la Figura 31, últimos 21 meses por variedad; en la Figura 32, últimos 21 meses por origen; en la Figura 33, últimas 92 semanas por precio; en la Figura 34, últimas 92 semanas por variedad; y finalmente, Figura 35, últimas 92 semanas por origen.

Figura 63. *Últimos 21 meses por Price.*

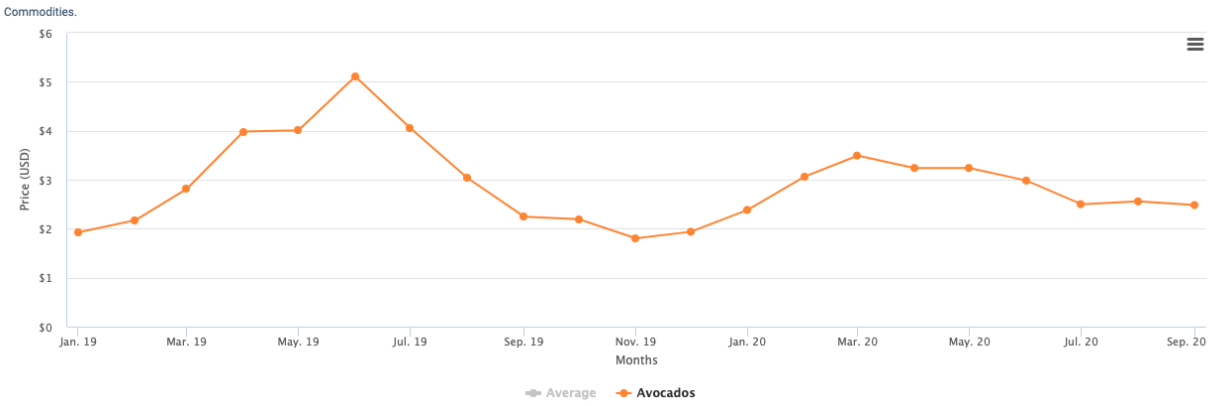


Figura 64. *Últimos 21 meses por Variety.*

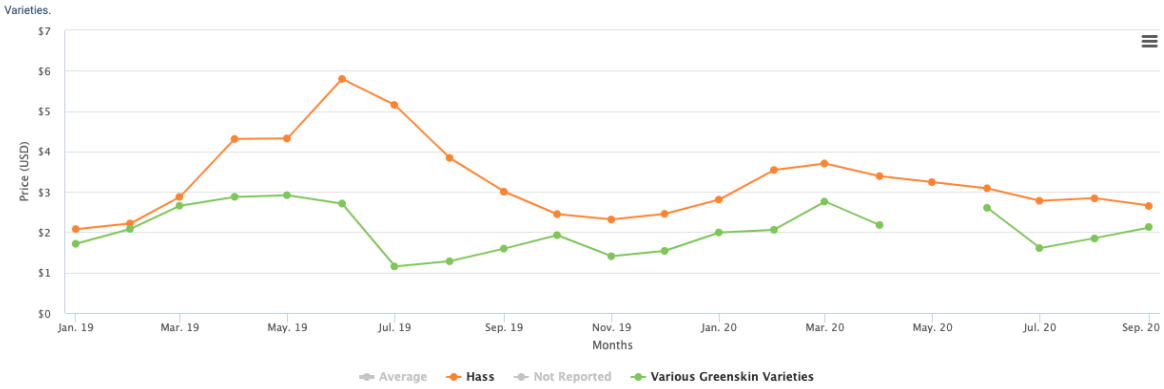


Figura 65. *Últimos 21 meses por Origen.*

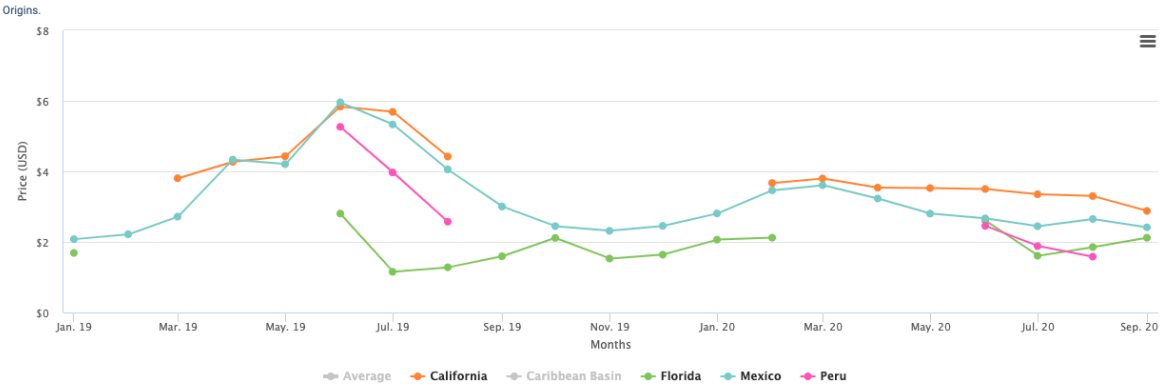


Figura 66. *Últimos 92 semanas solo Precio.*

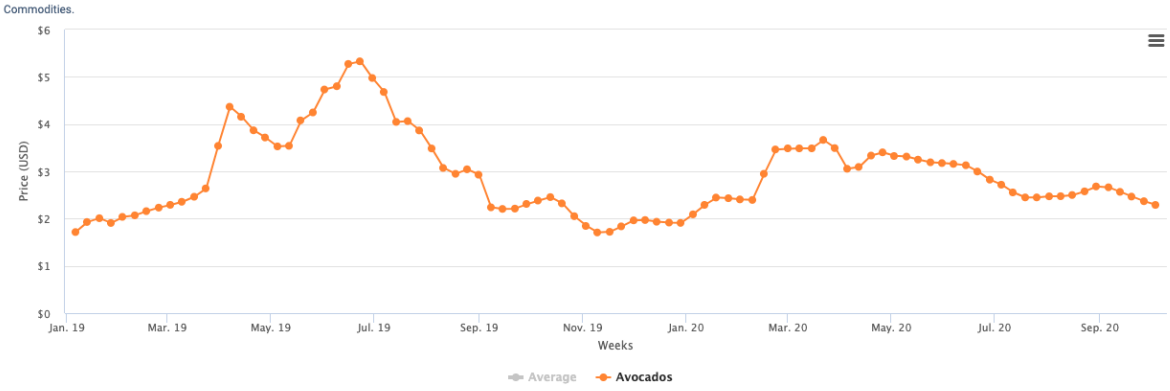


Figura 67. *Últimos 92 semanas por Variety.*

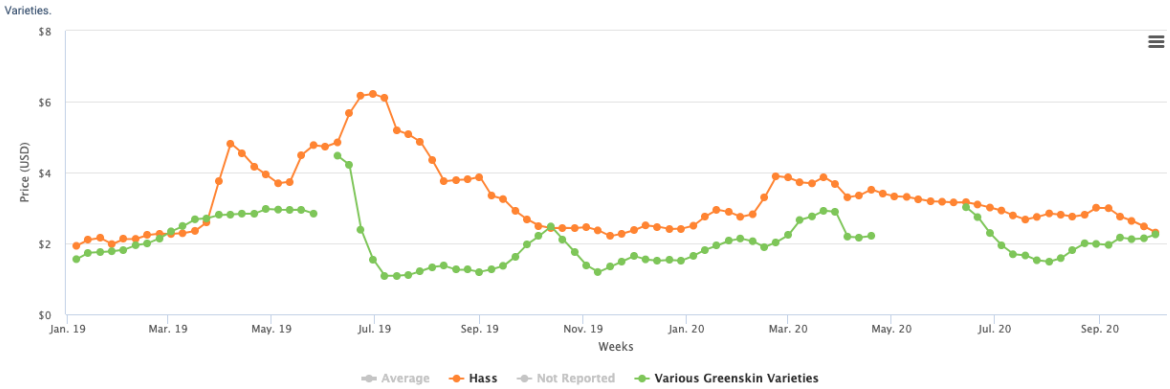
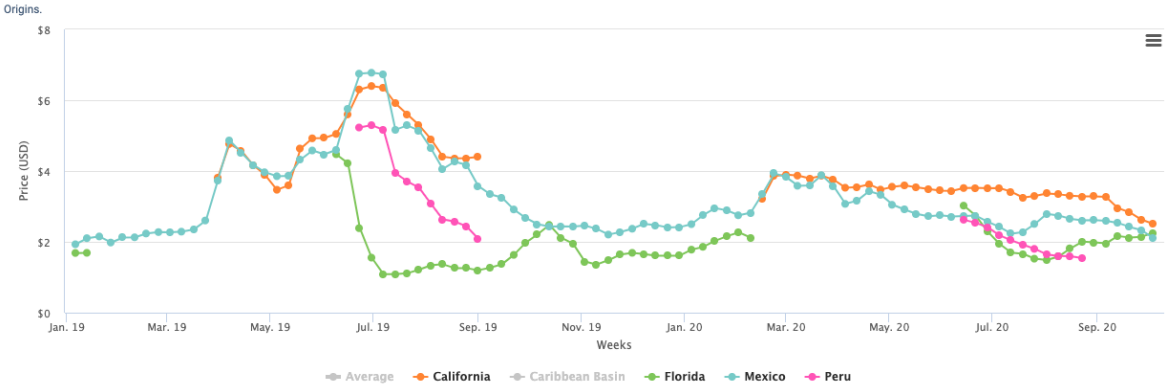


Figura 68. *Últimos 92 semanas por Origen.*



Anexo Nº 20. Estructura Organizacional Del Área Comercial.

Figura 69. *Estructura Organizacional Del Área Comercial.*