UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA INDUSTRIAL Y COMERCIAL

Aplicación de técnicas de Machine Learning para identificar factores de predicción del estado de las cotizaciones en el sector de maquinaria ligera

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para obtener el título profesional de Ingeniero Industrial y Comercial

AUTORES

Leslie Alexandra Aliaga Urbina

Frank Junior Calle Burga

Carla Isabel Pacori Choquejahua

Noelia Yackelina Palma Bendezú

Hector Jhon Salinas Cevallos

ASESOR

Junior John Fabian Arteaga

ORCID Nº 0000-0001-9804-7795

Octubre, 2022

RESUMEN

En el presente estudio haremos uso de Machine learning, usando 4 técnicas en la categoría de aprendizaje supervisado, para la predicción del estado de las cotizaciones, buscando agilizar la toma de decisiones con respecto al tiempo y costos de importación, y evitar la pérdida de ventas.

Para la construcción del modelo predictivo se inició con la recopilación y limpieza de datos. Posteriormente, se utilizó el 80 % de datos recopilados para el entrenamiento de los modelos y 20% para la evaluación de las predicciones. Con la técnica k-NN se obtuvo un accuracy del 67.9% con un parámetro de k = 5; con la técnica Regresión logística, se obtuvo un 70.69% de accuracy; con la técnica SVM se obtuvo un 63.79% de accuracy y con la técnica Árbol de decisión se obtuvo un accuracy de 87.93%.

Se aplicó codificación y normalización como mejora a la base de datos y con ello, la técnica de Árbol de decisión obtuvo el valor más alto de accuracy - 88.79%. Se recomienda el empleo de técnicas adicionales de Aprendizaje Supervisado a fin de seleccionar la que mejor resultado obtenga en la predicción.

PALABRAS CLAVE: Machine learning, Aprendizaje Supervisado, Toma de decisiones, y Cotizaciones.

ABSTRACT

In this study we will make use of Machine learning, using 4 techniques in the category of supervised learning, for the prediction of the status of quotations, seeking to speed up decision-making with respect to the time and costs of importation, and to avoid the loss of sales.

The construction of the predictive model started with data collection and cleaning. Subsequently, 80% of the collected data was used to train the models and 20% for the evaluation of the predictions. With the k-NN technique, an accuracy of 67.9% was obtained with a parameter of k = 5; with the Logistic Regression technique, an accuracy of 70.69% was obtained; with the SVM technique, an accuracy of 63.79% was obtained; and with the Decision Tree technique, an accuracy of 87.93% was obtained.

Codification and normalization was applied as an improvement to the database and with this, the Decision Tree technique obtained the highest accuracy value - 88.79%. The use of additional Supervised Learning techniques is recommended in order to select the one with the best prediction result.

KEYWORDS: Machine learning, Supervised learning, Decision making and Quotations.

ÍNDICE DE CONTENIDO

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

INTRODUCCIÓN

La pandemia del COVID-19 obligó a varios sectores económicos a cambiar sus procesos dado que debían adaptarse a la nueva normalidad, haciendo uso de la transformación digital. Entre los sectores más afectados se encontraron aquellos cuyas operaciones dependían de las importaciones debido a que, durante la pandemia, el mercado internacional estuvo paralizado, ocasionando una escasez de productos y alza de precio de estos a nivel mundial. Por este motivo, varias empresas quebraron y otras decidieron cambiar de rubro o adaptar sus operaciones con apoyo de la tecnología.

Cabe resaltar que, en el año 2021 en el Perú, con el apoyo de la reactivación económica, diversos sectores han logrado ir incrementando su producción, teniendo un impacto positivo en sus proveedores y clientes, los cuales también se han enfocado en la mejora y transformación de sus procesos. Para una empresa importadora, por ejemplo, su demanda ha incrementado, sin embargo, debe tener en cuenta que sus productos dependen del momento y el costo al que importa del mercado internacional.

La empresa en estudio tiene un problema con respecto a la gestión de stock de sus productos (montacargas). Actualmente en la empresa, los tiempos de entrega a sus clientes son elevados (considerando el elevado tiempo de importación de 80 semanas), al igual que sus precios en comparación a la competencia, ambos influenciados por el momento en que se importa un montacarga y su disponibilidad. Todo lo mencionado está generando que las cotizaciones de la empresa sean rechazadas ya sea por tener plazos de entrega extensos, por la no disponibilidad de stock, o por los precios más elevados frente a la competencia, afectando así la satisfacción del cliente y ocasionando un impacto negativo en las ventas.

Por este motivo, en la presente investigación, como primer paso se identificarán los factores o variables relevantes en la decisión de aceptación o rechazo de una cotización por parte del cliente, analizadas a través de cuatro técnicas de Machine learning para poder predecir el estado de las cotizaciones de la empresa, buscando agilizar la toma de decisiones respecto a la adquisición de los montacargas, con el objetivo de ofrecer al cliente tiempos de entrega y precios más competitivos.

Asimismo, se busca que la identificación de factores relevantes y su análisis a través de Machine Learning para la predicción de las cotizaciones pueda ser ampliado a otras líneas de productos de la empresa, además de ser empleado o adaptado por otras empresas de rubros/sectores distintos o similares, beneficiando así su toma de decisiones.

El presente trabajo se divide en seis capítulos. Iniciamos el capítulo uno planteando la problemática de la empresa, continuando en el capítulo dos con los antecedentes de la investigación y sus bases teóricas. En el capítulo tres se describe el entorno empresarial y el modelo de negocio. En el capítulo cuatro se describe la metodología que se usará en la investigación y en el capítulo cinco se desarrolla la aplicación de las cuatro técnicas de Machine Learning planteadas. Finalmente, en el capítulo seis, se presentan las conclusiones y recomendaciones.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Descripción de la Realidad Problemática

Actualmente el sector de maquinaria ligera abastece a industrias como la construcción, minería, agricultura, hidrocarburos, entre otras, por lo que su contribución en el mercado peruano representa un papel relevante. Con respecto al crecimiento de la producción nacional. El Instituto Nacional de Estadística e Informática (marzo, 2022) indica:

"La producción nacional en el mes de enero 2022 registró un aumento de 2,86%, como resultado del incremento de los sectores: Alojamiento y Restaurantes, Minería e Hidrocarburos, Transporte y Almacenamiento, Agropecuario, Comercio, Telecomunicaciones, Servicios Prestados a Empresas y Electricidad Gas y Agua. Sin embargo, otros sectores productivos mostraron contracción como Financiero y Seguros, Manufactura, Pesca y Construcción. Cabe señalar que este resultado tiene como base de comparación enero 2021, mes que mostró restricciones adicionales que afectó la actividad productiva". (pág. 1).

CUADRO Nº 01
Evolución del Índice Mensual de la Producción Nacional: Enero 2022
(Año base 2007)

Sector	Ponderación 1/	Variación Porcentual	
		Enero 2022/2021	Feb 2021-Ene 2022/ Feb 2020-Ene 2021
Economía Total	**100,00**	**2,86**	**13,64**
DI-Otros Impuestos a los Productos	**8,29**	**5,55**	**20,26**
Total Industrias (Producción)	**91,71**	**2,61**	**13,08**
Agropecuario	5,97	4,96	4,06
Pesca	0,74	-30,27	-4,69
Minería e Hidrocarburos	14,36	4,53	8,70
Manufactura	16,52	-1,57	16,82
Electricidad, Gas y Agua	1,72	3,05	8,84
Construcción	5,10	-0,59	32,92
Comercio	10,18	2,34	18,04
Transporte, Almacenamiento, Correo y Mensajería	4,97	9,24	21,90
Alojamiento y Restaurantes	2,86	30,37	55,27
Telecomunicaciones y Otros Servicios de Información	2,66	3,50	7,22
Financiero y Seguros	3,22	-6,96	3,81
Servicios Prestados a Empresas	4,24	3,32	15,47
Administración Pública, Defensa y otros	4,29	3,84	4,10
Otros Servicios 2/	14,89	4,04	9,53

Tabla 01: "Evolución del Índice Mensual de la producción Nacional: Enero 2022"
Fuente : INEI

Como se observa en el gráfico, la reactivación económica sigue avanzando en nuestro país después de la pandemia del COVID-19; sin embargo, ocasionó grandes cambios y desafíos para los distintos sectores comerciales, ya que las restricciones de comercio internacional generaron que los productos de importación incrementen su precio al reducirse su disponibilidad.

El precio de las importaciones entre enero y marzo del 2022 se incrementó un 22.3%, mientras que el volumen se redujo en un 7.5 %, tal como se observa en el gráfico. Esto ocurrió principalmente por el alza de precios a nivel mundial debido a la pandemia, además de otros conflictos internacionales, alcanzando así el valor más alto en los últimos años. "(...) Además, como se ha mencionado, el valor de las compras de cada tipo de bienes se ha incrementado, mientras que el volumen de compras ha disminuido. En otras palabras, se está comprando más caro y menos cantidad de bienes." (Comex, 2022).

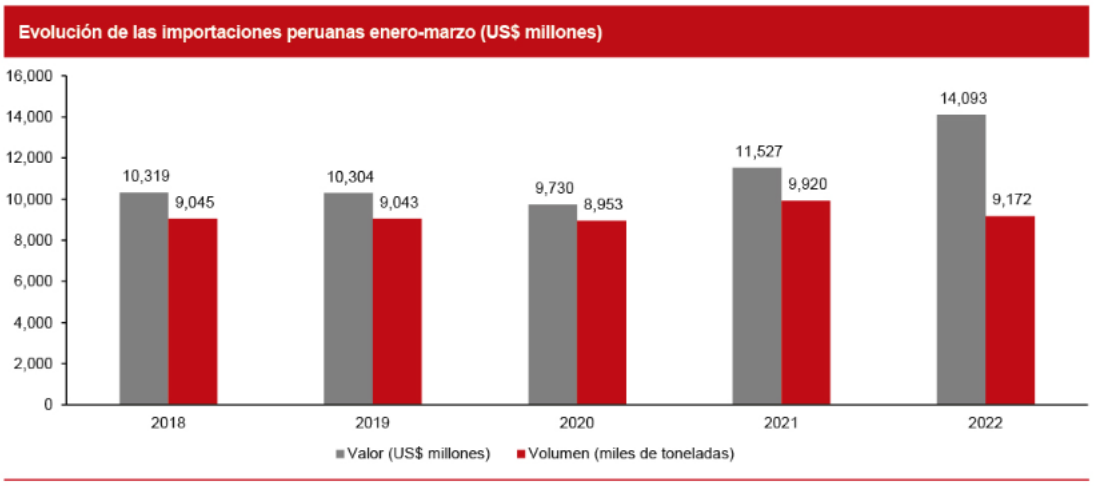


Gráfico 01: "Evolución de las importaciones peruanas enero - marzo 2022"
Fuente: Sunat. Elaborado por Comex Perú

Como consecuencia de la pandemia, los tiempos y costos de importación de los equipos se han elevado, lo que ha llevado a la empresa a tener desafíos para cumplir con las condiciones comerciales que se ofrecían anteriormente al cliente, lo que se traduce en pérdida de ventas para la compañía. Algunos clientes optan por comprar a la competencia porque esta les ofrece un mejor tiempo de entrega, ya sea porque disponen mayor stock de los productos solicitados o una ventaja competitiva en la estimación / predicción de la demanda.

Actualmente en la empresa, la planificación de compras de los equipos de importación para abastecer el inventario se basa en la información de las cotizaciones aprobadas y cantidades vendidas de los 2 años anteriores, y se realiza mediante una coordinación entre el área de logística y de producto. Al no ser precisa esta planificación, no se posee la debida cantidad de equipos en stock, por lo cual se inicia la importación recién cuando se realiza una venta. Esto conlleva, en primer lugar, a atender al cliente con tiempos de entrega elevados pues la importación se inicia con la venta y, en segundo lugar, a cobrar precios elevados al cliente pues el precio de los productos se encarece a medida que se tarde más en iniciar su importación, elevando los costos logísticos relacionados. Esto último como consecuencia de la pandemia y conflictos internacionales, tal como se comentó anteriormente.

Una mejora en la planificación permitiría a la empresa contar con el stock requerido para poder atender a sus actuales y nuevos clientes en el momento oportuno y con un precio más competitivo a fin de mejorar su participación en el mercado.

1.2 Justificación de la Investigación

1.2.1. Justificación Teórica

En la empresa se aplicará el tipo de aprendizaje supervisado, específicamente el enfoque de clasificación. La rama de Aprendizaje Supervisado se basa en que, dentro del conjunto de datos a analizar, se tiene el conocimiento existente o a priori de la variable que buscamos predecir a través del modelo. Entre sus dos principales técnicas se encuentran: Clasificación y Regresión. La técnica de clasificación, en base al análisis y entrenamiento previo de los datos de salida y entrada, permite distinguir y predecir a qué categoría pertenecerán los nuevos datos de estudio.

1.2.2. Justificación Práctica

El presente trabajo tiene como objetivo agilizar la toma decisiones con respecto a la compra de los equipos de importación, con el fin de aumentar las ventas al ofrecer al cliente mejores tiempos de entrega y precios. Se busca lograrlo mediante la aplicación de machine learning que nos ayudará a predecir cuándo una cotización es aprobada y en base a ello conocer la cantidad de equipos a comprar para cubrir la demanda, de esta manera la empresa será más competitiva en el mercado ya que podrá ofrecer al cliente mejores plazos de entrega y precios.

1.2.3. Justificación Metodológica

La investigación comenzará con un levantamiento de información sobre datos como cotizaciones pasadas de las cuales se obtendrán datos de tipo de producto, fecha, cantidad, monto, cliente, forma de pago, tiempo de entrega, condiciones comerciales, estado de cotizaciones (aprobado, rechazado), rubro de clientes, tipo de producto que solicita el cliente, vendedor, cantidad, entre otros.

Después de identificar el problema, se evaluará la técnica de Machine Learning que mejor asertividad tenga para solucionar nuestro problema principal, la cual podría ser cualquiera de las cuatro técnicas de aprendizaje supervisado propuestas.

1.3 Delimitación de la investigación

1.3.1 Espacial

El presente trabajo de investigación se realizará en la empresa con sede central en la ciudad de Lima, incluyendo también las operaciones de comercialización y alquiler de maquinaria ligera en todas sus sucursales.

Los datos para el presente trabajo se obtuvieron de la documentación y fuentes de información de la empresa, así como de libros, artículos, estudios, etc., relacionados a la metodología y herramienta a aplicar.

1.3.2 Temporal

El presente trabajo analizará la aplicación de Machine Learning del tipo Aprendizaje Supervisado, así como las actividades del área de Producto Aliados, específicamente de la línea de montacargas, de la empresa en el año 2022, además de data histórica de 3 años de antigüedad para su análisis.

1.3.3 Conceptual

Esta investigación se enfocará en el estudio, análisis y aplicación de la técnica de Aprendizaje Supervisado (Clasificación) de Machine Learning al proceso de compras del área de Productos Aliados (línea de montacargas) a través de la data histórica de sus cotizaciones y su información comercial; por este motivo, se recopilará información de SAP y se empleará el lenguaje de programación Python para la aplicación correcta de los algoritmos de predicción.

CAPÍTULO II: MARCO TEÓRICO

2.1 Antecedentes de la Investigación

Tesis Relacionadas:

Aguilar, C. (2021). Interpretable machine learning for promotional sales

Problema: De acuerdo con la tesis, se hace mención que las promociones tienen un alto nivel de impacto en la venta de los productos. En algunos países como Reino Unido pronostican el incremento de sus ventas al 200% gracias a las ofertas. La importancia de la estimación de las ventas por medio de las ofertas se debe a que afecta a toda la cadena de suministros. La tesis se enfocará en tres tópicos: la predicción de ventas promocionales, la predicción de ventas promocionales en escenarios cold-start (no hay histórico de ventas) y canibalización de ventas de productos causada por ofertas. Se tocarán esos temas ya que son los principales problemas recurrentes que tienen los analistas de ventas de la empresa.

Los principales **objetivos** de la tesis son:

- Creación, discusión y evaluación de la aplicabilidad de métodos interpretables a la predicción de ventas promocionales.
- Predicción cold-start en promociones y su solución a través de métodos interpretables.
- Canibalización de ventas desde una perspectiva causal.

Con respecto a la **metodología** que usaron en el trabajo de investigación se divide en distintos puntos: revisión extensiva de la literatura, preparación y validación de los datos, definición de métricas de interés, definición de cota baja de precisión empleando un método simple, definición cota alta de precisión empleando un método sofisticado encontrado en la búsqueda bibliográfica, diseñar modelos subrogados que permitan la evaluación de la interpretabilidad, investigar modelos que suplan las carencias de interpretabilidad de los métodos encontrados en la literatura, verificación mediante backtesting de la solución propuesta respecto de los modelos de alta y baja precisión. Finalmente, evaluación de los resultados, lo cual se aborda iterando los pasos anteriores hasta que la precisión se sitúe al menos entre las dos cotas.

Para el primer objetivo, se ha hecho uso de la técnica del algoritmo de los K vecinos más próximos (K-NN) con la finalidad de predecir las ventas por medio de las promociones. Los **resultados** de este método se basan en seleccionar variables que influyan directamente a las **ventas históricas**. Es importante resaltar que las promociones históricas son seleccionadas de acuerdo a las características. Con respecto a la interpretabilidad en el cálculo de promociones puede traducirse en beneficios para todos los involucrados en la cadena de suministros.

Con respecto al segundo objetivo se usó un método denominado Gradiente Boosting Decision Tree (GBDT). Con respecto a los resultados de este método generan predicciones cold start la cual es normal en productos nuevos. Dichos resultados en modelos subrogados detectan variables que influyen en la venta y es capaz de seleccionar las promociones semejantes a la actual.

Para el último objetivo, se ha tomado en cuenta el método Causal Impact. A comparación de los métodos mencionados anteriormente este **no requiere de data** histórica de los productos que pueden canibalizar. Tomar en consideración que todos los productos con ventas promocionales se pueden analizar. En base a la muestra determinada de la tesis (3067 productos en 13 departamentos de 11 supermercados), indican un total de 1965 episodios de canibalización que se traducen en un total de 719 271 unidades no vendidas debido a este fenómeno. Finalmente hacen mención que la canibalización promedio calculada para todos los productos es de un 31%.

Baldoceda Ramírez,A., Mamani Ccallohuari H. (2020), Perú. Modelo de aprendizaje supervisado para pronóstico de la deserción de estudiantes de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión - Lima. Universidad Peruana Unión.

El autor de esta tesis, señala que los principales motivos de la deserción universitaria son personales, académicos,socioeconómicos e institucionales, siendo el caso de la Universidad Peruana Unión. Por este motivo, esta tesis tiene como objetivo determinar el mejor modelo predictivo a usar para cada carrera de la Facultad de Ingeniería y Arquitectura.

Los datos utilizados en la tesis fueron, fueron definidas en base a sus edades, género, religión, nacionalidad, escuela, número de cursos desaprobados 1 vez, número de cursos desaprobados 2 veces, número de cursos desaprobados 3 veces, número de

cursos desaprobados 4 veces, número de cursos generales desaprobados, número de cursos específicos desaprobados, número de cursos de especialidad desaprobados, número de veces de traslado de carrera, número de cursos retirados, diferencia entre notas actuales y pasadas, cantidad de dinero que invertirá para finalizar la carrera y cantidad de dinero que lleva pagando en la carrera.

La metodología que el autor usó consta de 6 fases las cuales se explican a continuación: Comprensión del negocio, comprensión de los datos, preparación de los datos, diseño de los modelos predictivos, validación del modelo e implementación del modelo.

La primera fase consta de la comprensión del negocio en el que se determinó los objetivos por el cual se requiere desarrollar esta investigación. En la primera fase se comprende la necesidad de solucionar el problema para poder ayudar en la toma de decisiones que tendrá la Facultad.

En la segunda fase, se realizan las definiciones de las variables que influencian en la deserción de los alumnos de la Facultad, la cual son divididas en 3 categorías: personales, académicas y financieras.

En la tercera fase, se preparan los datos obtenidos para el diseño del modelo predictivo. Así mismo, en esta fase se pueden agregar o eliminar variables según corresponda. En caso algunas variables necesiten una estandarización y normalización, se les realiza para que puedan tener una mejor lectura de los datos.

Para la cuarta fase, se realizó el diseño del modelo predictivo utilizando el enfoque supervisado. Se validaron los datos y se separaron para el entrenamiento y testeo para seleccionar las técnicas del modelo más apropiado.

En la quinta fase, se valida el modelo utilizando la matriz de confusión para poder determinar el modelo más adecuado para cada carrera de la Facultad.

Las técnicas usadas por el autor para el desarrollo de la tesis fueron: K-NN, Naive Bayes, Decision Tree, Random Forest y XGBOOST.

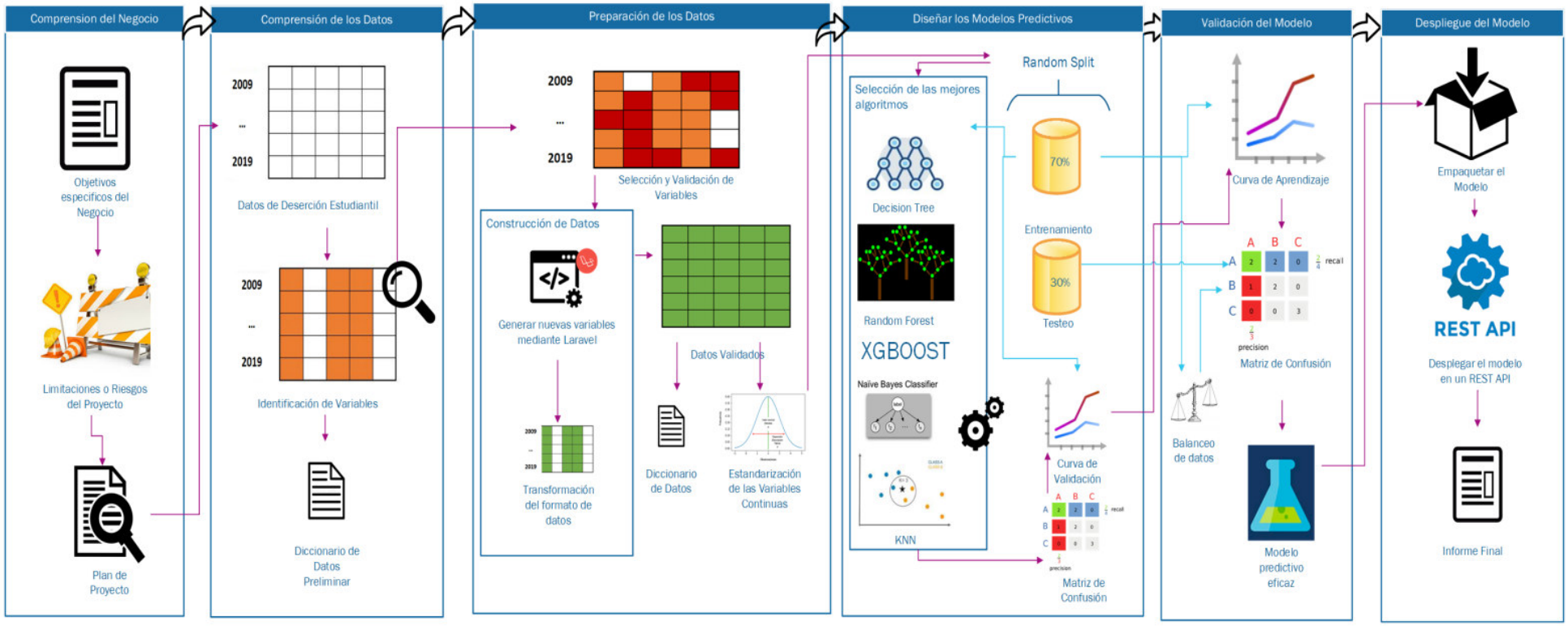


Gráfico 02: "Metodología propuesta en la tesis"

Fuente: Baldoceda Ramírez,A., Mamani Ccallohuari H. (2020)

Para el desarrollo del modelo se utilizaron 4 modelos predictivos las cuales son Naive de Bayes, XG BOOST, Árboles de Decisión (Decision Tree) y Bosques Aleatorios (Random Forest). A cada uno de estos modelos se le aplicó la técnica de Balanceo de Datos y poder equilibrar la cantidad de los estudiantes que continuarán el siguiente ciclo con la cantidad de los alumnos que no regresaron.

Para determinar el mejor modelo predictivo para cada carrera, se usó la técnica "Balanced Accuracy" y así encontrar el equilibrio entre las predicciones correctas e incorrectas.

Respecto a los resultados, se obtuvo que para la carrera de Ingeniería de Sistemas el mejor modelo es Random Forest, para Ingeniería Civil es Decisión Tree, para Ingeniería de Alimentos es KNN, para Ingeniería Ambiental es KNN y finalmente para Arquitectura es KNN.

Eiras-Franco, C. (2019), España. New Scalable machine learning methods: Beyond classification and regression. Universidade da Coruña.

El autor de esta tesis comenta que trabajar con grandes conjuntos de datos conlleva problemas logísticos, dado que el manejo y almacenamiento de grandes cantidades de datos se escapa de las capacidades de las tecnologías tradicionales. El aprendizaje a gran escala también limita la complejidad computacional y espacial de los algoritmos utilizados, siendo los algoritmos con coste lineal o menor los que mejor se prestan a este escenario, frente a alternativas que ofrecen mejores resultados pero a un coste computacional mayor. Este escenario también favorece los métodos con pocos o ningún hiper parámetro a configurar, dado el alto coste que tiene realizar muchas iteraciones de entrenamientos de prueba para ajustar dichos valores. Por último, el aprendizaje a gran escala muestra complicaciones específicas que dificultan el aprendizaje tales como la maldición de la dimensionalidad en el caso de conjuntos de datos con un gran número de variables.

El autor de esta tesis comenta que el problema de algunas empresas es manejar grandes cantidades de datos nos lleva a tener problemas logísticos, ya que almacenar una gran cantidad de datos en la actualidad no es controlable con las tecnologías tradicionales.

Existe, por tanto, una oportunidad en el estudio de algoritmos de aprendizaje máquina que puedan realizar aprendizaje a gran escala. En esta tesis se habla sobre la escalabilidad de los algoritmos de aprendizaje máquina y en ella explicaremos tanto modos de mejorar la escalabilidad de algoritmos existentes como nuevos desarrollos de algoritmos que tienen la escalabilidad como meta de diseño.

Los conjuntos de datos utilizados en la tesis se presentan en la siguiente tabla, donde se distinguen dos tipos. El primer tipo incluye conjuntos de datos de tamaño pequeño y mediano, y el segundo tipo, con un mayor número de muestras.

Dataset	# Features(N/C)	# Instances	Anomaly ratio
Arrhythmia (Arrhyth)	278 (271/7)	420	0.4357
German Credit (GC)	20 (7/13)	1000	0.3000
Abalone 1-8 (Ab. 1)	10 (7/3)	4177	0.3368
Abalone 9-11 (Ab. 9)	10 (7/3)	4177	0.3167
Abalone 11-29 (Ab. 11)	10 (7/3)	4177	0.3464
CoverType (CT)	12 (10/2)	286048	0.0096
KDD99 (full) (KDD)	41 (32/8)	4898431	0.8000
KDD99 (10%) (KDD10)	41 (32/8)	494021	0.8000
KDD99 (http) (KDDh)	40 (32/7)	623091	0.0065
KDD99 (smtp) (KDDs)	40 (32/7)	96554	0.0123
IDS	27 (8/19)	2071657	0.0333

Tabla 02: "Real datasets used for the comparative study"

Fuente: Eiras-Franco,C. (2019)

La metodología que usó el autor de la tesis se basa en explorar la data obtenida y verificar si hay datos los cuales son redundantes o nulos. Así mismo, también se detectó la presencia de datos con patrones que no se ajustan a la distribución del resto y por lo tanto necesitan una normalización.

En la tesis, se analizó el uso de cuatro estrategias para la obtención de algoritmos escalables nuevos o también para transformar costosos algoritmos ya existentes las cuales son seis.

Cabe resaltar que, para cada algoritmo se detallan los resultados experimentales que muestran su validez en comparación con los métodos previamente disponibles, como su capacidad de escalar a grandes conjuntos de datos.

Los resultados experimentales demuestran que, en una variedad de conjuntos de datos muy diversos en sus dimensiones, las características seleccionadas por ReliefFLSH difieren poco de las seleccionadas por el más costoso ReliefF, mientras que el tiempo de computación se reduce sensiblemente. Además, ReliefF-LSH ofrece mejores resultados que otros métodos de aproximar ReliefF, pudiendo, a diferencia de estos, enfrentarse a todos los tipos de datos que se pueden procesar con la versión exacta de ReliefF.

Barrueta, R y Castillo, E. (2018) Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos. Universidad Peruana de Ciencias Aplicadas.

El problema que abarca esta investigación es que el modelo actual de los bancos está enfocado en la venta de sus productos sin enfocarse en los clientes, además de trabajar con una base de datos con poca información y sin un modelo que permita el análisis y manejo de mayor cantidad de datos.

La metodología empleada en la investigación consta de tres etapas importantes para el diseño de este modelo: La carga y procesamiento de los datos a través del aplicativo web, la selección de variables, algoritmos y la ejecución del modelo.

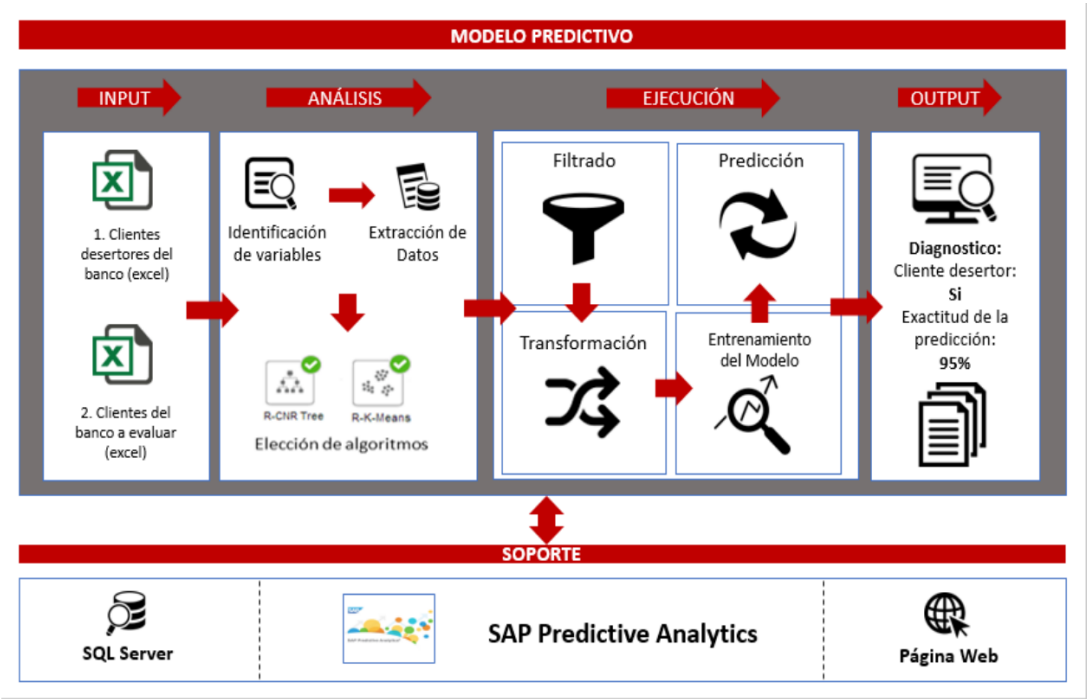


Gráfico 03: "Modelo de análisis predictivo de la tesis"

Fuente: Barrueta, R y Castillo, E. 2018

Como parte del desarrollo de esta tesis, el autor plantea usar las técnicas de agrupamiento y redes neuronales / Clasificación con el propósito de conseguir la más próxima cercanía en cuanto a la predicción de clientes desertores. El criterio a emplear para determinar la aplicabilidad del algoritmo se basará en los datos de input que se utilizarán en el modelo y la aplicabilidad de la funcionalidad de predicción de cada uno, mientras que la eficacia será medida en base a las predicciones correctas del conjunto de datos de prueba

El alcance de la investigación abarca todos los clientes con intención de abandonar las entidades bancarias. De esta manera se realiza la clasificación en dos grupos: desertores y no desertores.

El trabajo de investigación recopiló 20 000 clientes de los cuales se utilizó el 67 % para entrenamiento del modelo aplicado y el 33 % restante fue utilizado para la fase de validación. A continuación se muestra el cuadro de datos recopilados.

	Cantidad	Porcentaje
Desertores	1,764	9%
No Desertores	18,236	91%
Total	20,000	100%

Tabla 03: Clientes desertores y no desertores

Fuente: Barrueta, R y Castillo, E. 2018

El autor define sus variables dentro de cuatro grandes grupos los cuales son comportamiento, percepción, información demográfica y variables macroeconómicas.

Entre las variables de comportamiento se encuentran edad, fecha de nacimiento, situación laboral, estado civil, nivel educacional, etc.

El grupo de percepción abarca variables como encuesta de satisfacción y encuesta de calidad mientras que el grupo de información demográfica abarca variables como dirección, distrito, agencia.

Por otro lado se tienen las variables macroeconómicas entre las cuales figuran la utilidad neta, la clasificación de riesgo de un cliente, el sueldo neto del cliente, el ingreso conyugal, entre otras. A continuación se muestra la figura de clasificación de las variables en los cuatro grupos antes mencionados:

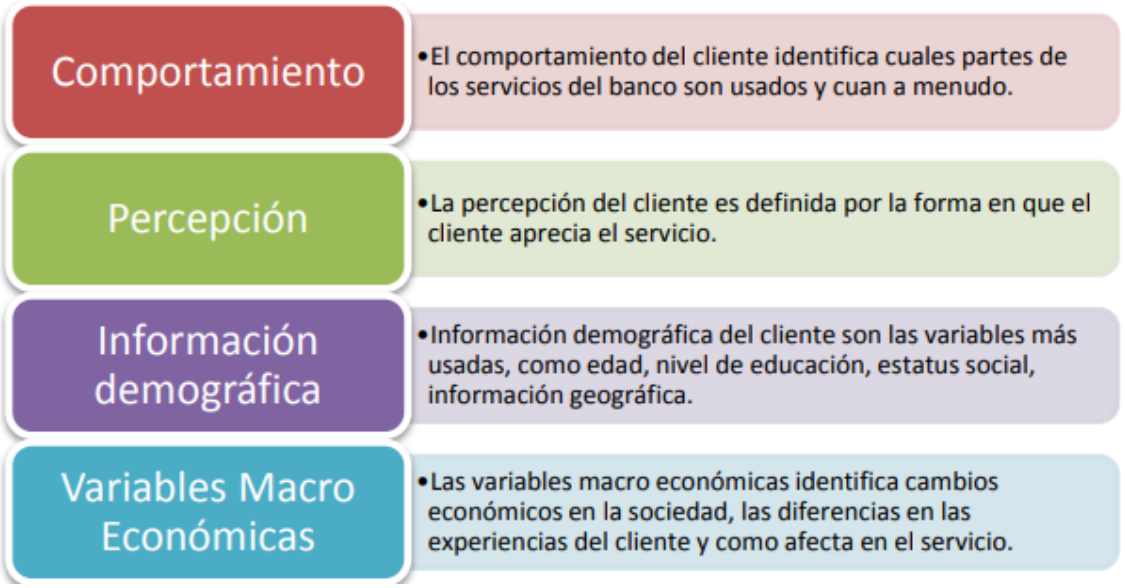


Gráfico 04: "Grupos de clasificación de variables"

Fuente: Barrueta, R y Castillo, E. 2018

La solución planteada se forman 4 grupos con un máximo de 100 iteraciones para de esta forma poder estabilizar el modelo, en estos grupos se representan los comportamientos que tiene un cliente desertor considerando las variables anteriormente mencionadas. Se utilizó el método R-K-Means con el que se tuvo una precisión de 93.20% y se identificó 8 falsos positivos que indican que el cliente no estaba desertando cuando no era así. También se utilizó el árbol R-CNR, donde se identificaron 10 falsos positivos y negativos, y la red neuronal R-NNet, donde se identificaron 14 falsos negativos y ningún falso positivo, con las que se obtuvo precisiones de 92.5% y 87.3%.

De la Fuente, A. (2022) Diseño de soluciones avanzadas basadas en Técnicas de Machine Learning para la toma de decisiones en Gestión de activos. Universidad de Sevilla.

El problema que abarca esta investigación es que en la actualidad la gestión de mantenimiento de los activos de una empresa es fundamental para no tener retrasos en la entrega de productos por fallos en las máquinas, ocasionando pérdidas significativas en las empresas. Por este motivo, la tesis aborda la gestión de activos y la ingeniería del mantenimiento, por otro lado, también desarrollar un proceso para la transformación del conocimiento extraído a partir de los datos generados por los activos, en herramientas técnicas de Machine Learning que al final serán comparadas con el objetivo de elegir la que mejor se ajuste.

La presente investigación toma como muestra a los equipos de bombeo sumergidos en tanques criogénicos. Estas bombas se ubican en plantas de regasificación, la cual están diseñadas para almacenar gas natural en estado líquido (a temperatura criogénica) y así puedan elevar la presión para el cambio de estado líquido a gaseoso, favoreciendo el transporte posterior por gasoductos a temperaturas mayores que facilitan su operación.

Para el análisis, se usaron los siguientes datos de las bombas: Flujo de las bombas, Presión en el tanque, Temperatura en el tanque, Temperatura en la impulsión, Nivel de líquido en el tanque, Densidad GNL, Horas desde último mantenimiento, Consumo de la bomba, entra otras.

La metodología empleada en la investigación consta de 9 etapas importantes para diseñar el modelo predictivo, las cuales son las siguientes: Selección del equipo, contexto operacional, selección de variables, selección del periodo temporal, selección

del intervalo de registro de datos, Procesado y validación de la información, normalización de la información, implementación en Vensim y obtención del modelo predictivo. A continuación se muestra el flujo de trabajo que se utilizó para la tesis.

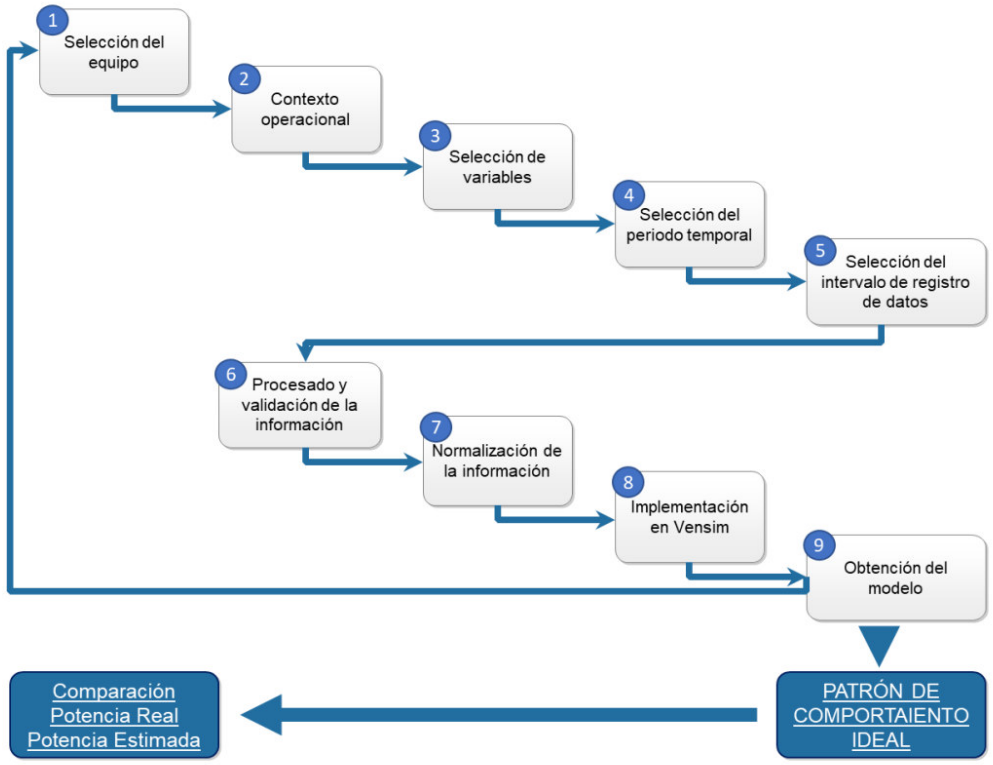


Gráfico 05: "Flujo de trabajo de la metodología a usar en la investigación"

Fuente: De la Fuente, A. (2022)

La selección del equipo se refiere a elegir el activo de la empresa a estudiar, tomando en cuenta la representatividad de este activo en base a las utilidades. También se toma en cuenta el histórico de datos de operación y los datos de mantenimiento de los activos.

El contexto operacional se refiere a los factores que influyen en los activos, como la localización, la altitud, parámetros operacionales y entre otros.

La selección de variables se refiere a elegir las variables que influyen en el mantenimiento de un activo, en este caso para las bombas.

La definición del periodo temporal, es para dar a conocer el periodo de tiempo la cual se realizará el estudio de la presente investigación.

La selección del intervalo de registro de datos, se refiere a la definición de la antigüedad de los datos que estarían entrando en el modelo predictivo, para tener una predicción cercana al tiempo real.

Después se define la metodología que seguirá el trabajo de investigación, dando a conocer las técnicas de machine learning que se usarán y también el lenguaje de programación. Seguidamente se realiza la normalización de datos, la cual se realiza para que el modelo predictivo no tenga distorsiones en las variables. Por último, se obtiene el modelo predictivo que se usará en la investigación.

Las técnicas que se usaron en la presente investigación son las siguientes: Redes neuronales, árboles de decisión, bosques aleatorios (Random Forest), máquinas de vectores de soporte, Deep learning, entre otros.

Los resultados de la investigación son una comparación de las técnicas de machine learning para predecir el mantenimiento de las máquinas (bombas), la cual se obtuvo como conclusión que la mejor técnica a usar, mediante la métrica del MSE, es la técnica de árboles de decisión con un 97% de accuracy. Esta técnica permitirá a la empresa una mejor Gestión de activos y un adecuado mantenimiento preventivo de sus activos.

Céspedes, A. (2017), Santiago, Chile. ***Construcción De Modelo De Forecast Para Estimación De Demanda En Una Empresa Multinacional De Retail*****, Universidad Técnica Federico Santa María.**

En la presente tesis, se comenta sobre una empresa multinacional de dulces en Chile llamada Mars, la cual cuenta con 72 000 colaboradores presente con más de 20 plantas en más de 78 países y con ventas netas de más de 33 mil millones de dólares. La empresa importa chocolates de Estados Unidos, México y Argentina, trabajando con marcas como Snickers, M&Ms, Milky Way, entre otras.

El problema que abarca la investigación se refiere a los tiempos elevados de importación desde el país de Estados Unidos, debido a que aproximadamente el tiempo de importación es de tres meses y la vida útil de los chocolates es doce meses, teniendo como resultado que la vida útil de los chocolates cuando llegan a Chile se reduce a nueve meses. Así mismo, un problema adicional es que la demanda de chocolates es irregular, teniendo periodos de sobre stock por cambios en el mercado, obligando a la empresa a tener un inventario de seguridad.

La empresa Mars plantea como objetivo mejorar la precisión del pronóstico de la demanda a un 70% como mínimo, debido a que actualmente llegan a un 55%. Por este motivo, la empresa para llegar a esa precisión utilizó el histórico de demandas de los años 2014 al 2016, teniendo como principales datos lo siguiente: fecha de compra, demanda del producto, cliente y tipo de producto.

La metodología utilizada para la presente investigación se detalla de la siguiente manera: Obtención de datos del equipo comercial, limpieza de datos, análisis de datos, comprensión de datos, preparación de datos, selección de técnica a utilizar, construcción de modelos, evaluación de modelos, plan de implementación y monitoreo del plan.

Así mismo, el autor usa la metodología SEMMA la cual hace referencia a Muestreo, Explore, Modificación, Modelado y Valoración. A continuación se presenta el siguiente gráfico como modelo:

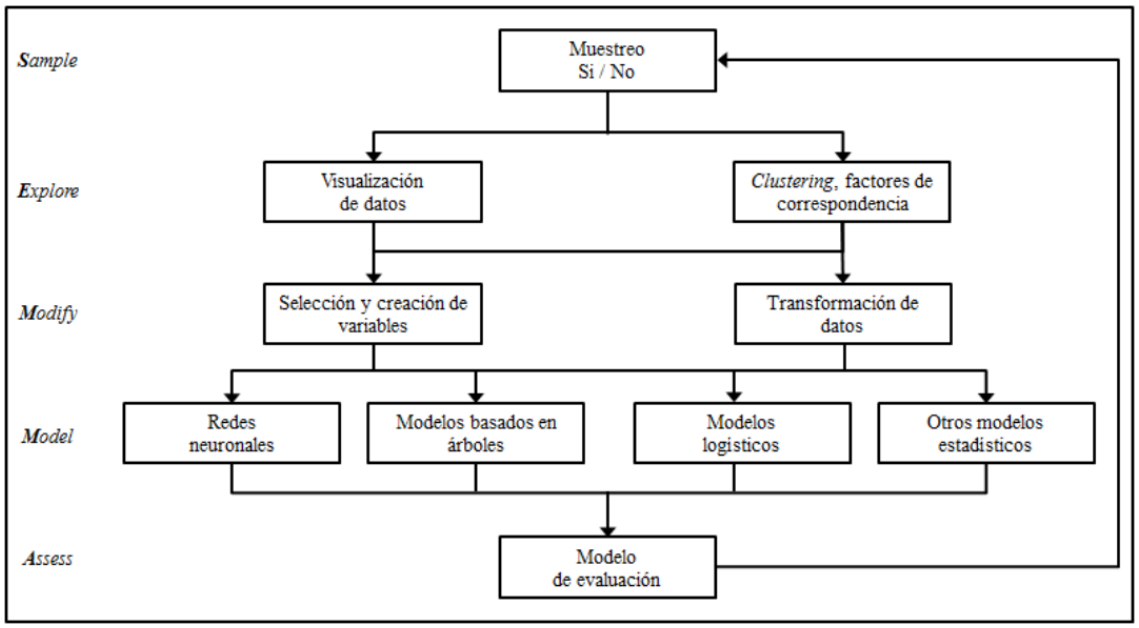


Gráfico 06: "Metodología SEMMA usada en la investigación"

Fuente: Céspedes, A. (2017)

Las técnicas usadas en la presente investigación fueron las Redes neuronales artificiales (RNA), debido a que es la técnica más usada en base a sus referencias. Así mismo, utilizaron el lenguaje de programación R, ya que es una herramienta del análisis estadístico y cuenta con diversas bibliotecas con métodos predictivos.

Los resultados de la investigación están relacionados a una mejor evaluación de pronósticos de demanda con el modelo de redes neuronales, ya que brinda una predicción a tiempo real y automatiza los procesos de predicción. Así mismo, también se comenta que deberán tener una mejor integración del conocimiento de la demanda

con los equipos de la empresa, debido a que no todos estaban capacitados para entender la información. Finalmente, realizando el análisis con el modelo de redes neuronales podrá tener una precisión del pronóstico de la demanda de por lo menos el 80%, sobrepasando 10% el objetivo planteado por la empresa.

2.2 Bases Teóricas

2.2.1 Inteligencia Artificial

Inteligencia Artificial o IA es un concepto complejo dado que incluso definir lo que es "inteligencia" aún continúa en debate. Por ejemplo, Russell & Norvig (2010) nos presenta 4 grupos de definiciones para la IA:

Sistemas que piensan como humanos	Sistemas que piensan racionalmente
«El nuevo y excitante esfuerzo de hacer que los computadores piensen… máquinas con mentes, en el más amplio sentido literal». (Haugeland, 1985) «[La automatización de] actividades que vinculamos con procesos de pensamiento humano, actividades como la toma de decisiones, resolución de problemas, aprendizaje…» (Bellman, 1978)	«El estudio de las facultades mentales mediante el uso de modelos computacionales». (Charniak y McDermott, 1985) «El estudio de los cálculos que hacen posible percibir, razonar y actuar». (Winston, 1992)
Sistemas que actúan como humanos	**Sistemas que actúan racionalmente**
«El arte de desarrollar máquinas con capacidad para realizar funciones que cuando son realizadas por personas requieren de inteligencia». (Kurzweil, 1990) «El estudio de cómo lograr que los computadores realicen tareas que, por el momento, los humanos hacen mejor». (Rich y Knight, 1991)	«La Inteligencia Computacional es el estudio del diseño de agentes inteligentes». (Poole *et al.*, 1998) «IA… está relacionada con conductas inteligentes en artefactos». (Nilsson, 1998)

Figura 1.1 Algunas definiciones de inteligencia artificial, organizadas en cuatro categorías.

Gráfico 07: "Algunas definiciones de inteligencia artificial"
Fuente: Russell & Norvig (2010)

Una definición más aterrizada es brindada por Rouhiainen (2018) quien sostiene que:

> (...) IA es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. Sin embargo, a diferencia de las personas, los dispositivos basados en IA no necesitan descansar y pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente

> menor en las máquinas que realizan las mismas tareas que sus contrapartes humanas. (Rouhiainen,2018,p. 14).

Otro ejemplo de lo que es IA es presentado por Theobald (2017), quien sostiene que:

> La inteligencia artificial, o IA, abarca la capacidad de las máquinas para realizar tareas inteligentes y cognitivas. Similar a la forma en que la Revolución Industrial dio origen a una era de máquinas que podían simular tareas físicas, la IA está impulsando el desarrollo de máquinas capaces de simular habilidades cognitivas. (Theobald,2017,p.12).

Es decir, Inteligencia Artificial se entiende como la simulación de la inteligencia y capacidad de razonamiento de los seres humanos a través de las máquinas, con el fin de llevar a cabo acciones determinadas o actuar de acuerdo a la situación. Por tal motivo, actualmente, la IA está siendo aplicada en diversos ámbitos o sectores. Tenemos, por ejemplo, el reconocimiento de imágenes tanto para sistemas de seguridad como para el sector salud, el mantenimiento predictivo que beneficia en gran medida al sector industrial, la identificación de rostros y recomendaciones de contenido para las redes sociales, etc. Rouhiainen (2018) sostiene que:

> La IA también será capaz de ofrecernos sugerencias y predicciones relacionadas con asuntos importantes de nuestra vida, lo que tendrá su impacto en áreas como la salud, el bienestar, la educación, el trabajo y las relaciones interpersonales. De la misma manera, cambiará la forma de hacer negocios al proporcionar ventajas competitivas a las empresas que busquen entender y aplicar estas herramientas de forma rápida y eficaz. (Rouhiainen,2018,p. 15).

Ahora bien, lo que sería Inteligencia Artificial se inicia aproximadamente en el año 1950. Russell & Norvig (2010) sostienen que:

> Hay un número de trabajos iniciales que se pueden caracterizar como de IA, pero fue Alan Turing quien articuló primero una visión de la IA en su artículo Computing Machinery and Intelligence, en 1950. Ahí, introdujo la prueba de Turing, el aprendizaje automático, los algoritmos genéricos y el aprendizaje por refuerzo. (Russell & Norvig,2010,p. 20).

Sin embargo, el término “Inteligencia Artificial” o IA es acuñado por John McCarthy unos años después, luego de llevar a cabo un taller en Darmouth con el fin de

acrecentar el interés de los investigadores por las teorías de autómatas, inteligencia y redes neuronales. “Quizá lo último que surgió del taller fue el consenso en adoptar el nuevo nombre propuesto por McCarthy para este campo: Inteligencia Artificial. Quizá «racionalidad computacional» hubiese sido más adecuado, pero «IA» se ha mantenido” (Russell & Norvig, 2010, p. 20).

Sin embargo, cabe resaltar que IA no lo es todo pues está incluida dentro de lo que es la Ciencia de los Datos, pero, a su vez, incluye otras ramas, de las que forma parte Machine Learning, tal como se observa a continuación:

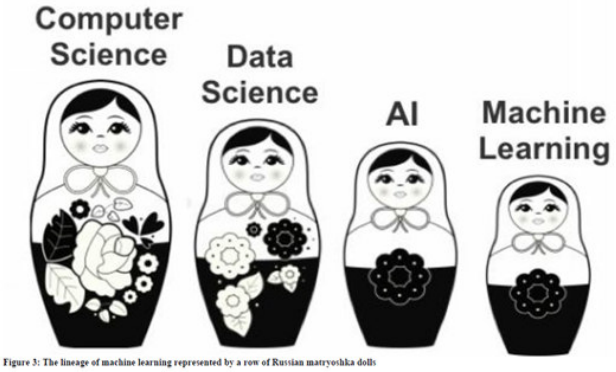


Figure 3: The lineage of machine learning represented by a row of Russian matryoshka dolls

Gráfico 08: “El linaje del aprendizaje automático representado por una fila de muñecas rusas matryoshka”
Fuente: Theobald, 2017, p.12

“IA contiene numerosos subcampos que son populares hoy en día. Estos subcampos incluyen la búsqueda y la planificación, el razonamiento y la representación del conocimiento, la percepción, el procesamiento del lenguaje natural (PNL) y, por supuesto, el aprendizaje automático” (Theobald, 2017, p.12).

2.2.2 Machine Learning

Cada vez más se producen grandes cantidades de datos desde distintas fuentes, por lo cual, saber usarlos (analizarlos) cobra gran importancia. “Las organizaciones modernas recopilan cantidades masivas de datos. Para que los datos sean valiosos para una organización, deben analizarse para extraer información que pueda usarse para tomar mejores decisiones” (Kelleher et. al, 2015, p. 36). Esto se ilustra a continuación:

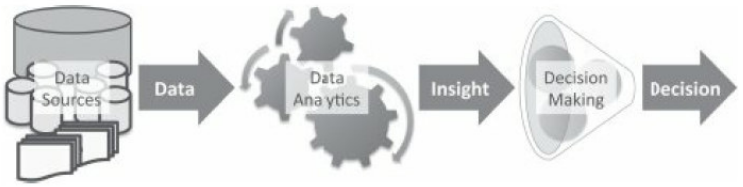


Gráfico 09: "Análisis de datos predictivos que pasan de los datos a la información y a la decisión"
Fuente: Kelleher et. al, 2015, p. 36

Pero, ¿cómo obtenemos este insight o información valiosa para tomar decisiones en el momento oportuno? Esto es logrado a través de Machine Learning: "El aprendizaje automático (machine learning) se define como un proceso automatizado que extrae patrones de los datos" (Kelleher et. al, 2015, p. 39). Es decir, esa información valiosa es obtenida gracias al análisis de las bases de datos. "Con machine learning podemos obtener información de un conjunto de datos; vamos a pedirle a la computadora que encuentre algún sentido a partir de los datos" (Harrington, 2012, p.3).

Como bien se mencionó, machine learning es un proceso, sin embargo, cabe resaltar que engloba algoritmos o técnicas dependiendo del resultado que se desee obtener y de los datos a analizar. "Machine Learning se basa en algoritmos para analizar grandes conjuntos de datos. (...) puede realizar análisis predictivos mucho más rápido que cualquier humano. Como resultado, puede ayudar a los humanos a trabajar de manera más eficiente" (Mueller & Massaron, 2021, p. 11). Así también, "Machine learning incorpora varios cientos de algoritmos basados en estadísticas y elegir el algoritmo o la combinación de algoritmos correctos para el trabajo es un desafío constante para cualquiera que trabaje en este campo" (Theobald, 2017, p.15).

Si bien Machine Learning abarca algoritmos específicos para el tipo de problema que se desee resolver, estos algoritmos son agrupados en 3 tipos de categorías, las cuales dependen de las características de los datos que poseemos. "(...) antes de examinar algoritmos específicos, es importante comprender las tres categorías generales de machine learning. Estas tres categorías son supervisadas, no supervisadas y de refuerzo" (Theobald, 2017, p. 15).

Cabe resaltar que las 2 categorías más ampliamente usadas son: Aprendizaje supervisado y Aprendizaje No Supervisado. A continuación se muestran estas 2 categorías con las técnicas que engloban:

Supervised learning tasks	
k-Nearest Neighbors	Linear
Naive Bayes	Locally weighted linear
Support vector machines	Ridge
Decision trees	Lasso
Unsupervised learning tasks	
k-Means	Expectation maximization
DBSCAN	Parzen window

Gráfico 10: "Algoritmos comunes utilizados para realizar tareas de clasificación, regresión, agrupamiento y estimación de densidad"
Fuente: Harrington, 2012, p. 10

Por lo tanto, dependiendo del problema que deseemos resolver y de los tipos de datos que poseemos, debemos escoger la técnica que mejor se ajuste a nuestras necesidades. "Con todos los diferentes algoritmos ¿cómo puedes elegir cuál usar? En primer lugar, debe considerar su objetivo. ¿Qué estás tratando de sacar de esto? ¿Qué datos tienes o puedes recopilar? Esas son las grandes preguntas" (Harrington, 2012, p.11). De esta manera, Harrington (2012) sostiene que:

> Si está tratando de predecir o pronosticar un valor objetivo, entonces debe considerar el aprendizaje supervisado. Si ha elegido el aprendizaje supervisado y su valor objetivo es un valor discreto, entonces desea buscar en la clasificación. Si el valor objetivo puede tomar varios valores, entonces debe considerar la regresión. Si no está tratando de predecir un valor objetivo, entonces debe considerar el aprendizaje no supervisado. ¿Está tratando de encajar sus datos en algunos grupos discretos? Si es así y eso es todo lo que necesita, debería considerar la agrupación en clústeres. ¿Necesita tener alguna estimación numérica de qué tan fuerte es el ajuste en cada grupo? Si responde que sí, entonces probablemente debería buscar un algoritmo de estimación de densidad. (Harrington, 2012,p. 11).

2.2.3 Aprendizaje Supervisado

El Aprendizaje Supervisado es una de las ramas más importantes de Machine Learning y se basa en el conocimiento previo o a priori de los datos, tanto de la variable que deseamos predecir, también llamada “Y”, como de las variables independientes que nos ayudan en dicha predicción, también llamadas “X”. Theobald (2017) sostiene que:

> El aprendizaje supervisado funciona alimentando los datos de muestra de la máquina con varias características (representadas como "X") y la salida del valor correcto de los datos (representada como "y"). El hecho de que los valores de salida y características sean conocidos califica el conjunto de datos como "etiquetado". Luego, el algoritmo descifra los patrones que existen en los datos y crea un modelo que puede reproducir las mismas reglas subyacentes con nuevos datos. (Theobald, 2017,p. 15).

Lo mencionado anteriormente se ilustra en el siguiente gráfico:

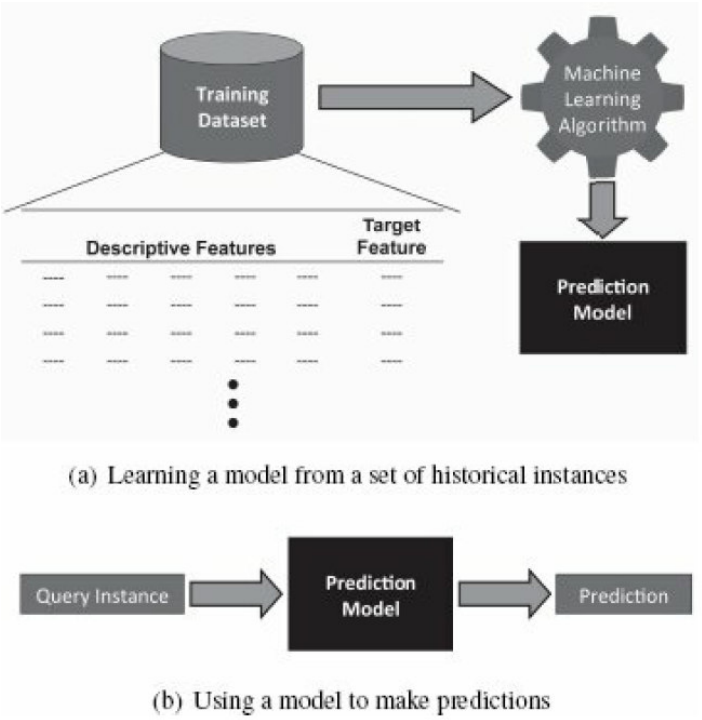


Gráfico 11: “Los dos pasos en el aprendizaje supervisado”

Fuente: Kelleher et. al, 2015, p. 39

De esta manera, Mueller & Massaron (2021) afirman que:

> El aprendizaje supervisado ocurre cuando un algoritmo aprende de datos de ejemplo y respuestas objetivo asociadas que pueden consistir en valores

> numéricos o etiquetas de cadena, como clases o etiquetas, para luego predecir la respuesta correcta cuando se le plantean nuevos ejemplos. El enfoque supervisado es de hecho similar al aprendizaje humano bajo la supervisión de un maestro. El maestro proporciona buenos ejemplos para que el estudiante los memorice, y el estudiante luego deriva reglas generales de estos ejemplos específicos. (Mueller & Massaron, 2021,p. 140).

Como se detalló en el apartado de Machine Learning, la categoría de aprendizaje supervisado agrupa distintas técnicas, las cuales, dependiendo de lo que se busque predecir, forman parte del enfoque que se desee emplear, de esta manera, Harrington (2012) sostiene que:

> En la clasificación, nuestro trabajo es predecir en qué clase debe caer una instancia de datos. Otra tarea en machine learning es la regresión. La regresión es la predicción de un valor numérico. (...) La clasificación y la regresión son ejemplos de aprendizaje supervisado. Este conjunto de problemas se conoce como supervisado porque le estamos diciendo al algoritmo qué predecir. (Harrington, 2012,p. 10).

2.2.4 Técnica/Algoritmo k-Vecinos más cercanos (k-NN)

El método k-NN, también conocido como algoritmo de k-Vecinos más cercanos, es una técnica o algoritmo de aprendizaje supervisado en el que se utilizan la proximidad de datos más cercanos para hacer clasificaciones definiendo experimentalmente la cantidad de grupos que se requiere para la clasificación (k). De esta manera, Mueller & Massaron (2021) sostienen que:

> “Ya sea que el problema sea adivinar un número o una clase, la idea detrás de la estrategia de aprendizaje del algoritmo K-Nearest Neighbors (KNN) es siempre la misma. El algoritmo encuentra las observaciones más similares a la que tiene que predecir y de las que deriva una buena intuición de la posible respuesta promediando los valores vecinos, o eligiendo la clase de respuesta más frecuente entre ellos. (Mueller & Massaron, 2021,p. 238).

Como se mencionó anteriormente, para asignar la clasificación a un nuevo dato, este debe compararse con los “k” datos, o vecinos, más cercanos, y se le asignará la categoría mayoritaria. De esta manera, Harrington (2012) afirma que:

> (...) Cuando recibimos un nuevo dato sin etiqueta, comparamos ese nuevo dato con los datos existentes. Luego tomamos los datos más similares (los vecinos más cercanos) y miramos sus etiquetas. Observamos los k datos más similares de nuestro conjunto de datos conocido; de ahí viene la k (k es un número entero y generalmente es menor que 20). Por último, tomamos un voto mayoritario de los "k" datos más similares, y la mayoría es la nueva clase que asignamos a los datos que se nos pidió que clasificáramos. (Harrington, 2012, p. 19).

Un ejemplo gráfico, empleando valores de k iguales a 3 y 7, es mostrado a continuación:

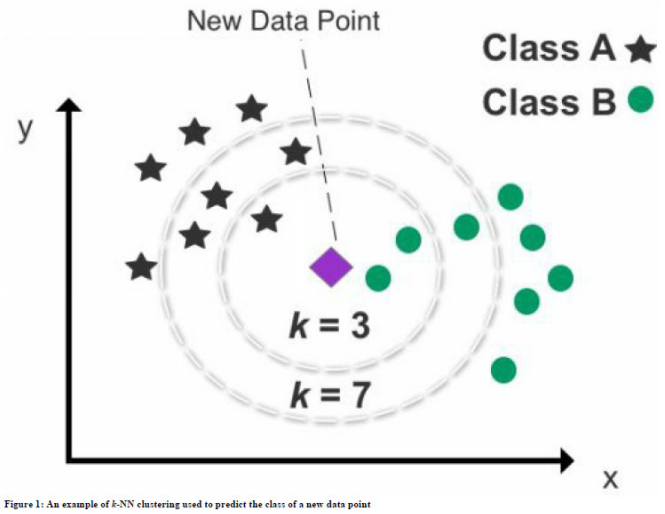


Figure 1: An example of *k*-NN clustering used to predict the class of a new data point

Gráfico 12: "Un ejemplo de agrupamiento k-NN utilizado para predecir la clase de un nuevo punto de datos"

Fuente: Theobald, 2017, p. 63

De esta manera y según el ejemplo, si asignamos un valor de k igual a 3, el nuevo dato pertenecería a la clase B (círculo), sin embargo, si empleamos un k igual a 7, el nuevo pertenece a la clase A (estrella). Cabe resaltar que el valor de k es definido por la persona que lleva a cabo el estudio o investigación, ajustando su valor dependiendo del nivel de confiabilidad obtenida. Muelle & Massaron (2021) sostienen que:

> El parámetro k es el que se puede modificar para hacer que un algoritmo KNN funcione bien en predicción y regresión. El valor k, un número entero, es el número de vecinos que el algoritmo debe considerar para encontrar una

> respuesta. Cuanto más pequeño sea el parámetro k, más se adaptará el algoritmo a los datos que está presentando, con el riesgo de sobreajuste pero ajustando muy bien los límites de separación complejos entre las clases. Cuanto mayor sea el parámetro k, más se abstrae de los altibajos de los datos reales, lo que deriva en curvas bien suavizadas entre clases de datos, pero lo hace a expensas de tener en cuenta ejemplos irrelevantes. (Muelle & Massaron, 2021,p. 238).

2.2.4 Técnica/Algoritmo de Regresión o Análisis de Regresión

Al igual que la técnica k-NN, el análisis de regresión también forma parte de la rama de Aprendizaje Supervisado. Esta técnica analiza la relación entre variables, bajo un proceso estadístico, buscando construir un "modelo" que permita estimar o predecir dicha relación. De acuerdo con Harrington (2012), respecto a la técnica de regresión nos indica que:

> El objetivo de la técnica de regresión es predecir un valor objetivo, escribiendo una ecuación tomando en cuenta las variables de entrada. Un ejemplo podría ser predecir los caballos de fuerza de un automóvil tomando en cuenta todas las variables que influyen directamente. (Harrington, 2012, p. 154)

Existen distintos tipos de análisis de regresión como, por ejemplo, regresión lineal, regresión logística, máquinas de vectores soporte (Support Vector Machine), entre otras. De esta manera, Harrington (2012) nos indica que:

> Cuando hablamos de regresión, a menudo nos referimos a la regresión lineal, pero existe la regresión lineal y la regresión logística. La regresión lineal significa que puedes sumar las variables de entrada multiplicadas por alguna constante para obtener la salida.

Con respecto a la **regresión lineal**, como su nombre lo dice, busca construir un modelo lineal, es decir, la gráfica de estimación que obtendremos será una línea recta. La regresión lineal busca construir una línea recta que pase por los datos graficados con la menor distancia posible a cada uno de ellos. Theobald (2017) sostiene que:

> La regresión lineal comprende una línea recta que divide los puntos de datos en un diagrama de dispersión. El objetivo de la regresión lineal es dividir

los datos de manera que se minimice la distancia entre la línea de regresión y todos los puntos de datos en el diagrama de dispersión. Esto significa que si dibujara una línea vertical desde la línea de regresión hasta cada punto de datos en el gráfico, la distancia total de cada punto equivaldría a la distancia más pequeña posible hasta la línea de regresión (Theobald, 2017, p. 48).

A continuación se muestra un gráfico sobre regresión lineal:

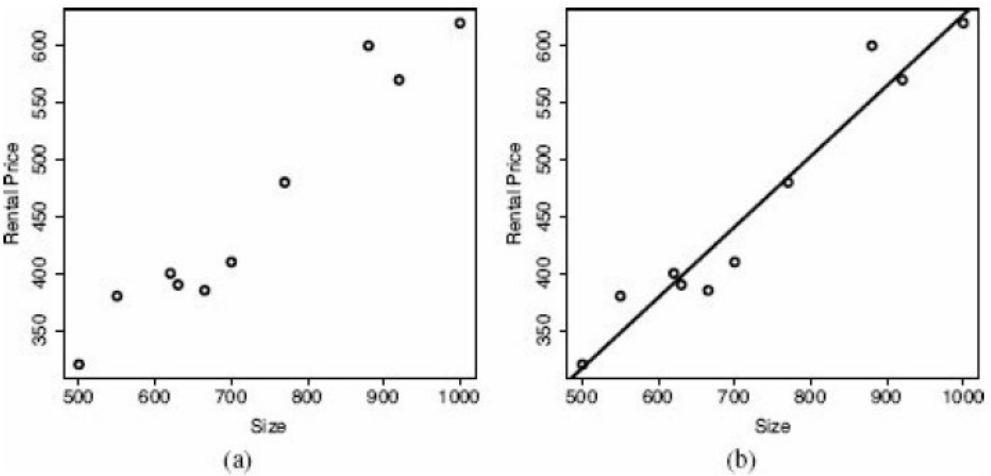


Gráfico 13: "(a) Diagrama de dispersión de las características Tamaño y Precio de alquiler; (b) Modelo lineal que relaciona el Precio de alquiler con el Tamaño."

Fuente: Kelleher, J., et. al, 2015, p. 355

Como se observa, se busca construir un gráfico lineal que pase por todos los puntos graficados, es decir, se construye una línea recta (función lineal) en base a la menor distancia de cada uno de los datos. Tal como se observa en este otro gráfico:

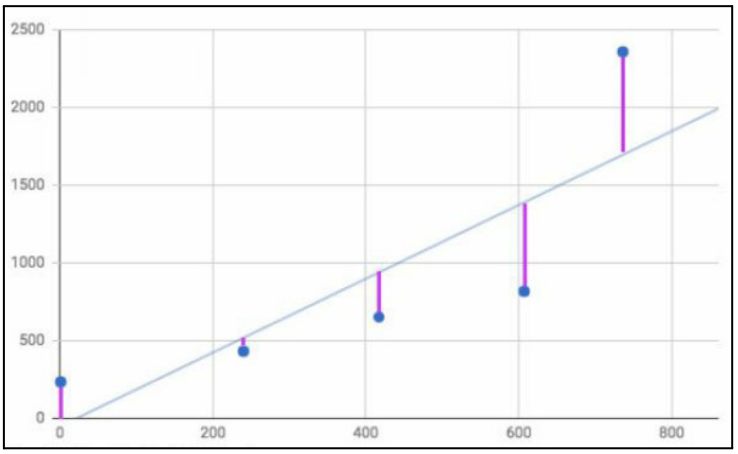


Gráfico 14: "Distancia de los puntos de datos al hiperplano"

Fuente: Theobald, 2017, p. 51

La exactitud de la predicción dependerá de la distancia de los puntos hacia la línea de regresión. “Cuanto más cerca estén los puntos de la línea de regresión, más precisa será la predicción final. Si hay un alto grado de desviación entre los puntos y la línea de regresión, la pendiente proporcionará predicciones menos precisas” (Theobald, 2017, p. 51).

Con respecto a la **regresión logística**, a diferencia de la regresión lineal que se usa principalmente en valores numéricos, esta se dirige más en el enfoque de clasificación. Theobald (2017) afirma que:

> Aunque la regresión logística comparte un parecido visual con la regresión lineal, técnicamente es una técnica de clasificación. Mientras que la regresión lineal aborda ecuaciones numéricas y forma predicciones numéricas para discernir relaciones entre variables, la regresión logística predice clases discretas (Theobald, 2017, p. 55).

Asimismo, Harrington (2012) sostiene que:

> Tal vez haya visto algunos puntos de datos y luego alguien ajustó una línea llamada la línea de mejor ajuste a estos puntos; eso es regresión. Lo que sucede en la regresión logística es que tenemos un montón de datos, y con los datos tratamos de construir una ecuación para hacer la clasificación por nosotros (Harrington, 2012, p. 83).

La ecuación empleada en la regresión logística es la ecuación o función sigmoidal. “La regresión logística adopta la función sigmoidal para analizar datos y predecir clases discretas en un conjunto de datos” (Theobald, 2017, p.55). A continuación se presenta la función sigmoidal:

$$h_{\mathbf{w}}(\mathbf{x}) = Logistic(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w} \cdot \mathbf{x}}}$$

Gráfico 15: Función Sigmoidal en la regresión logística

Fuente: Russell, S., & Norvig, P., 2010, p. 726

De esta manera, “(...) 1 dividido por 1 más *e* a la x negativa, x = valor numérico a transformar, e = constante de Euler, 2.718” (Theobald, 2017, p. 55). Asimismo, Harrington (2012) afirma que:

> Para el clasificador de regresión logística, tomaremos nuestras características y multiplicaremos cada una por un peso y luego las sumaremos. Este resultado se pondrá en el sigmoide, y obtendremos un número entre 0 y 1. Cualquier cosa por encima de 0,5 la clasificaremos como 1, y cualquier cosa por debajo de 0,5 la clasificaremos como 0. También puedes pensar en logística regresión como una estimación de probabilidad (Harrington, 2012, p. 85).

De manera gráfica, un ejemplo de diagrama de dispersión de la regresión logística es mostrado a continuación:

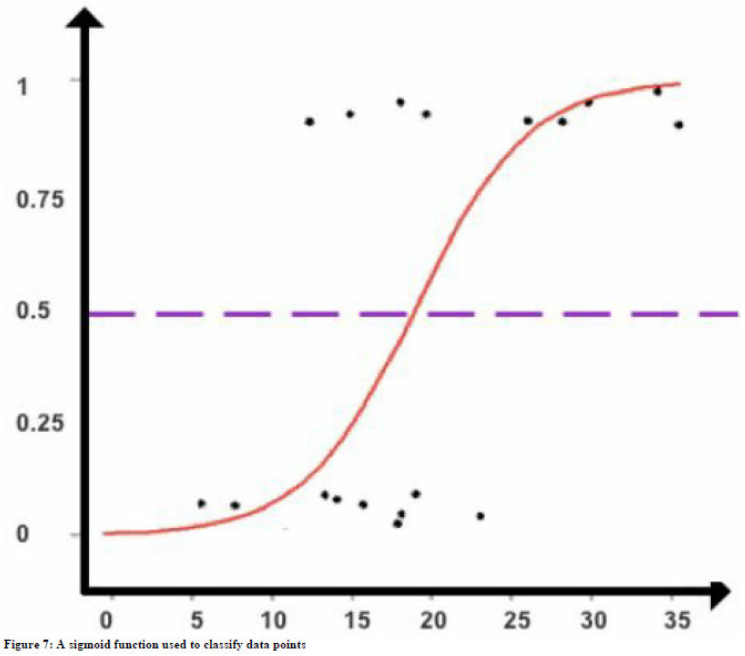


Gráfico 16: "Una función sigmoidal utilizada para clasificar puntos de datos"

Fuente: Theobald, 2017, p. 56

De acuerdo al gráfico anterior, la interpretación es brindada por Theobald (2017), quien sostiene que:

> Como se ve en la Figura 7, podemos crear un punto de corte en 0,5 para clasificar los puntos de datos en clases. Los puntos de datos que registran un valor por encima de 0,5 se clasifican como Clase A, y cualquier punto de datos por debajo de 0,5 se clasifica como Clase B. Los puntos de datos que registran un resultado de exactamente 0,5 no son clasificables, pero estos casos son raros debido al componente matemático de la función sigmoidea (Theobald, 2017, p. 56).

La regresión logística es usada ampliamente en diferentes sectores e industrias, tal como afirman Russell & Norvig (2010):

> (...) Estas ventajas tienden a trasladarse a las aplicaciones del mundo real y la regresión logística se ha convertido en una de las técnicas de clasificación más populares para problemas en medicina, marketing y análisis de encuestas, calificación crediticia, salud pública y otras aplicaciones (Russell, & Norvig, 2010, p. 727).

2.2.6 Técnica/Algoritmo Support Vector Machines

Máquinas de Vectores de Soporte o SVM (por sus siglas en inglés, Support Vector Machine) es un algoritmo que también se encuentra dentro de la categoría de Aprendizaje Supervisado, y es empleado para resolver problemas de clasificación. Esta técnica fue creada en la década de los 90 por el matemático Vladimir Vapnik. Mueller & Massaron (2021) afirman que:

> Las SVM, una especie de algoritmo que funciona tan bien como las redes neuronales en muchas ocasiones, son creación del matemático Vladimir Vapnik y algunos de sus colegas (Boser, Guyon y Cortes) que trabajaban en los laboratorios de AT&T en la década de 1990 (Mueller & Massaron, 2021, p. 309).

Cabe resaltar que SVM forma parte de la técnica de regresión, pero en un nivel más avanzado. Theobald (2017) sostiene que:

> Como categoría avanzada de regresión, la máquina de vectores de soporte (SVM) se parece a la regresión logística pero con condiciones más estrictas. Con ese fin, SVM es superior en el dibujo de líneas de límite de clasificación (Theobald, 2017, p. 57).

Para poder comprender la técnica de SVM debemos tener en claro 3 términos importantes: Hiperplano, márgen y vectores de soporte (support vectors). Al igual que vimos en la técnica de clasificación, el hiperplano puede entenderse como la línea que separa las categorías o los datos en un plano o gráfico. Habiendo separado gráficamente las categorías, el márgen viene a ser la distancia del punto más cercano de cada categoría hacia el hiperplano. Finalmente, los vectores de soporte son estos puntos más cercanos al hiperplano. “El hiperplano es nuestro límite de decisión. Lo que está a cada lado pertenece a una clase diferente” (Harrington, 2012, p. 103). Asimismo, “la

distancia desde el límite de decisión hasta la instancia de entrenamiento más cercana se conoce como margen" (Kelleher, et. al, 2015, p. 408). "Los puntos más cercanos al hiperplano de separación se conocen como vectores de soporte" (Harrington, 2012, p. 103).

A continuación se brinda un ejemplo gráfico sobre los 3 términos mencionados anteriormente:

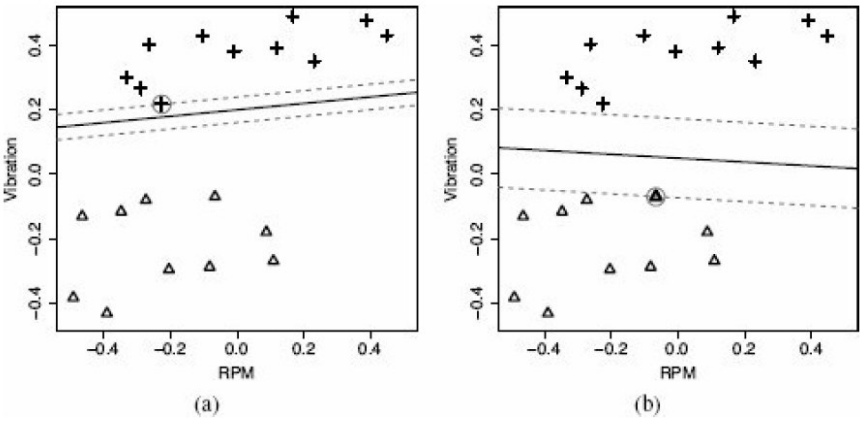


Gráfico 17: Gráficos de RPM (a) y Vibración (b) con límites de decisión y márgenes distintos

Fuente: Kelleher, et. al, 2015, p. 408

Por lo tanto, de acuerdo al gráfico observado, lo que se busca con SVM es determinar el margen máximo, a fin de poder distinguir o clasificar los elementos de la mejor manera, sin dejar de prestar atención a los vectores soporte. Kelleher, et. al (2015) afirma que:

> Entrenar una SVM implica buscar el límite de decisión, o separar el hiperplano, que conduce al margen máximo, ya que esto separará mejor los niveles de la entidad de destino (...) Las instancias en un conjunto de datos de entrenamiento que caen a lo largo de las extensiones de los márgenes y, por lo tanto, definen los márgenes, se conocen como vectores de soporte. Estas son las instancias más importantes en el conjunto de datos porque definen el límite de decisión (Kelleher, et. al, 2015, p. 408).

Cabe resaltar que encontrar la separación en un plano de dos dimensiones tiene sus limitaciones, por este motivo, los SVM realizan el análisis en un plano de mayor dimensionalidad con el fin de encontrar la manera de clasificar o determinar las clases

en los datos. De esta manera, los SVM emplean lo que se conoce como “funciones Kernel”. Russell & Norvig (2010) sostienen que:

> La función kernel se puede aplicar a pares de datos de entrada para evaluar productos escalares en algún espacio de características correspondiente. Entonces, podemos encontrar separadores lineales en el espacio de características de mayor dimensión con una función kernel K(xj , xk).. Por lo tanto, podemos aprender en el espacio de dimensiones superiores, pero solo calculamos las funciones del kernel en lugar de la lista completa de características para cada punto de datos (Russell & Norvig, 2010, p. 747).

Un término importante en SVM es también el truco Kernel o “Kernel trick”. Este “truco” es lo que permite llevar los datos a un plano de mayores dimensiones. Harrington (2012) afirma que:

> Una gran ventaja de la optimización SVM es que todas las operaciones se pueden escribir en términos de productos internos. Los productos internos son dos vectores multiplicados para producir un número escalar o único. Podemos reemplazar los productos internos con nuestras funciones del kernel sin hacer simplificaciones. Reemplazar el producto interno con un kernel se conoce como truco kernel o subestación kernel (Harrington, 2012, p. 119).

Este cambio de dimensión es mostrado a continuación:

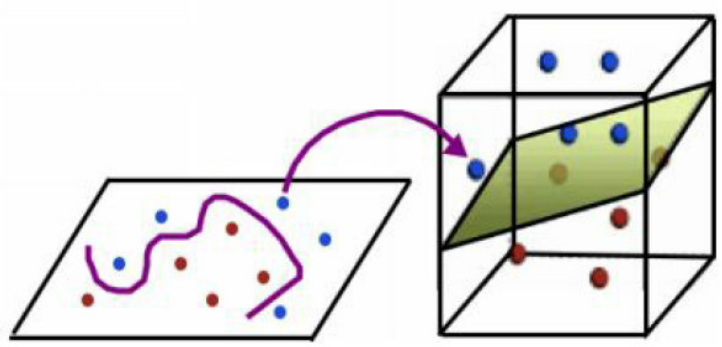


Gráfico 18: Transición de un espacio bidimensional a uno tridimensional
Fuente: (Theobald, 2017, p. 61)

Finalmente, la técnica SVM, al igual que las demás mencionadas, también es empleada en diferentes campos o sectores donde se requieren resolver problemas de clasificación o identificación. Mueller & Massaron (2021) sostienen que:

> Hoy en día, las SVM tienen un uso generalizado entre los científicos de datos, que las aplican a una increíble variedad de problemas, desde el diagnóstico médico hasta el reconocimiento de imágenes y la clasificación de textos (Mueller & Massaron, 2021, p. 310).

2.2.7 Técnica/Algoritmo de Árboles de Decisión

Los árboles de decisión es una de las técnicas más usadas para resolver problemas de clasificación, además de regresión, asimismo, se encuentra dentro de la rama de Aprendizaje Supervisado. Russell & Norvig (2010) afirman que:

> Un árbol de decisión representa una función que toma como entrada un vector de valores de atributo y devuelve una "decisión": un único valor de salida. Los valores de entrada y salida pueden ser discretos o continuos. (Russell & Norvig, 2010, p. 698).

A continuación se muestran dos ejemplos de árbol de regresión (clasificación y regresión):

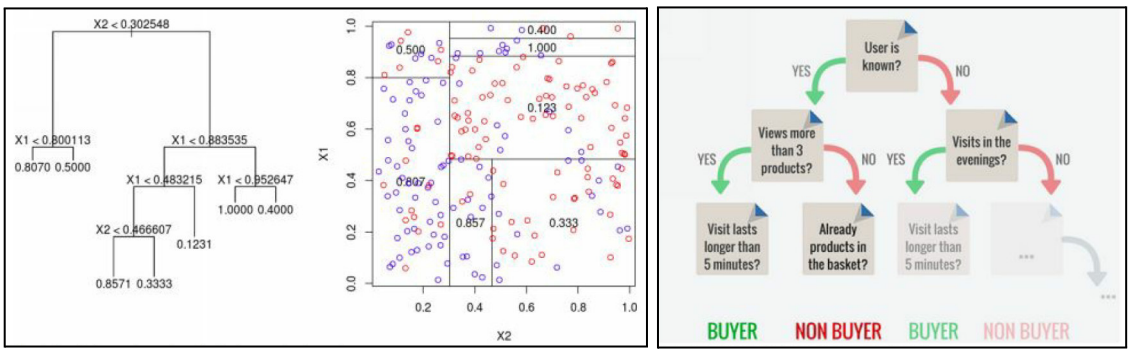


Gráfico 19: "Árbol de regresión (izquierda) y árbol de clasificación (derecha)"

Fuente: Theobald, 2017, p. 89

Como se observa, los árboles de decisión son un conjunto de nodos o pasos secuenciales, dependiendo del camino en base a las respuestas o valores que va tomando una variable, hasta llegar al valor o clase final. De esta manera, Mueller & Massaron (2021) sostiene que:

> Utilizando una muestra de observaciones como punto de partida, el algoritmo vuelve a trazar las reglas que generaron las clases de salida (o los

> valores numéricos al trabajar con un problema de regresión) dividiendo la matriz de entrada en particiones cada vez más pequeñas hasta que el proceso activa una regla para detener. En un contexto de aprendizaje automático, dicho razonamiento inverso se logra aplicando una búsqueda entre todas las formas posibles de dividir el entrenamiento en muestra y decidir usar la división que maximiza las mediciones estadísticas en las particiones resultantes (Mueller & Massaron, 2021, p. 181).

Por lo tanto, el árbol de decisión está compuesto de nodos, que determinan el camino que se seguirá dependiendo de las condiciones a cumplirse. Kelleher, et. al (2015) afirma que:

> Al igual que con todas las representaciones de árboles, un árbol de decisión consta de un nodo raíz (o nodo de inicio), nodos interiores y nodos de hoja (o nodos de terminación) que están conectados por ramas. Cada nodo que no es hoja (raíz e interior) del árbol especifica una prueba que se llevará a cabo en una característica descriptiva. El número de niveles posibles que puede tomar una característica descriptiva determina el número de ramas descendentes desde un nodo que no es una hoja. Cada uno de los nodos hoja especifica un nivel predicho de la característica de destino (Kelleher, et. al, 2015, p. 161).

De esta manera, la técnica busca ir clasificando/separando los datos después de cada nodo, hasta el punto de poder separar los datos en los grupos a los que pertenecen, en la medida de lo posible. Harrington (2012) afirma que:

> Para crear un árbol de decisiones, se debe tomar una primera decisión sobre el conjunto de datos para dictar qué función se utiliza para dividir los datos. Para determinar esto, prueba cada función y mide qué división te dará los mejores resultados. Después de eso, dividirá el conjunto de datos en subconjuntos. Los subconjuntos luego atravesarán las ramas del primer nodo de decisión. Si los datos en las ramas son de la misma clase, entonces los clasificó correctamente y no necesita continuar dividiéndolos. Si los datos no son los mismos, se debe repetir el proceso de división en este subconjunto. La decisión sobre cómo dividir este subconjunto se realiza de

la misma manera que el conjunto de datos original y repite este proceso hasta que haya clasificado todos los datos (Harrington, 2012, p. 39).

A continuación se muestra un ejemplo de árbol de decisión aplicado a la un caso de la vida real, siendo este la decisión de esperar por una mesa o no en un restaurante:

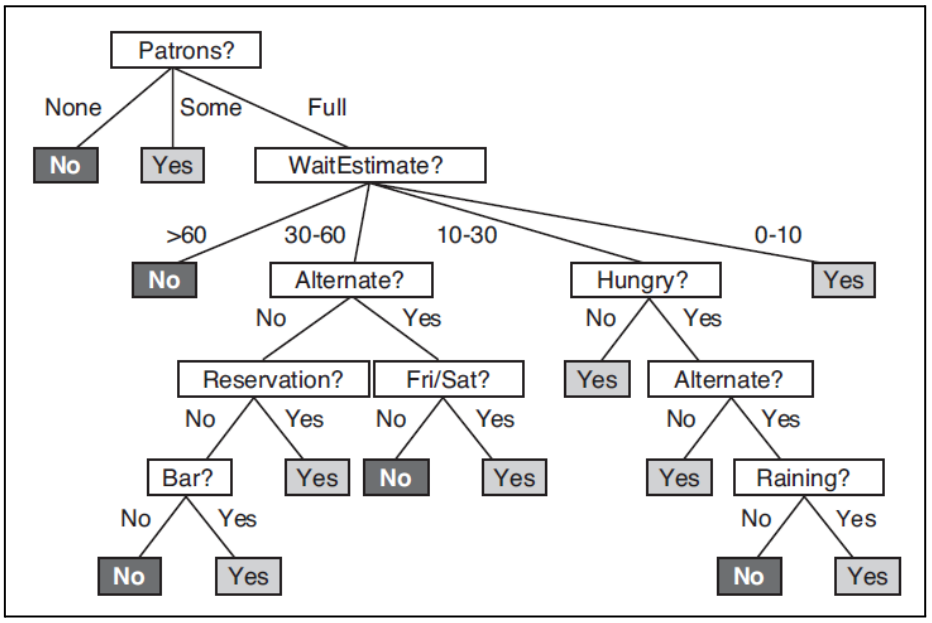


Gráfico 20: Un árbol de decisión para decidir si esperar por una mesa.

Fuente: Russell & Norvig, 2010, p. 699

Se debe tener en cuenta las reglas de parada o de detención, las cuales permitirán detener la ramificación del árbol de decisión, de esta manera, Mueller & Massaron (2021) sostienen que:

> Las reglas de parada son límites a la expansión de un árbol. Estas reglas funcionan al considerar tres aspectos de una partición: el tamaño de la partición inicial, el tamaño de la partición resultante y la ganancia de información que se puede lograr con la división. Las reglas de parada son importantes porque los algoritmos de árboles de decisión aproximan una gran cantidad de funciones; sin embargo, el ruido y los errores de datos pueden influir fácilmente en este algoritmo. En consecuencia, dependiendo de la muestra, la inestabilidad y la varianza de las estimaciones resultantes afectan las predicciones del árbol de decisión (Mueller & Massaron, 2021, p. 182).

Finalmente, el algoritmo de árbol de decisión se aplica en distintos casos de la vida real. “Los ejemplos de la vida real incluye elegir al beneficiario de una beca, evaluar a un solicitante de un préstamo hipotecario, predecir las ventas de comercio electrónico o seleccionar al solicitante de empleo adecuado” (Theobald, 2017, p. 90).

2.2.7 Técnica/Algoritmo de Redes Neuronales

Seleccionar un algoritmo o técnica de aprendizaje ayuda a obtener resultados de predicción una vez introducido un modelo con base de datos históricos, de esta manera el modelo producirá una respuesta cercana a la correcta. “En las Redes Neuronales Artificiales (ANN) se debe considerar que la unidad análoga a la neurona biológica es el *process element* (PE). Este elemento cuenta con varias entradas y las mezcla mediante una suma básica. La suma de las entradas es cambiada por medio una función de transferencia y el resultado de salida de esta se pasa directo a la salida del elemento procesador” (Basogsin, X., s.f., pág. 3)

En el siguiente gráfico se puede visualizar un elemento procesador de una red neuronal artificial

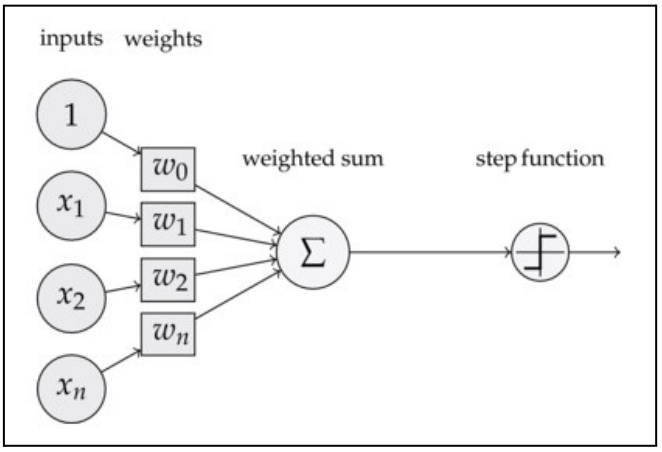


Gráfico 21: “Diagrama de una Neurona Artificial”

Fuente: Basogsin, X., s.f., pág. 3

Se debe considerar que una red neuronal abarca un conjunto de unidades de elementos PE interconectados. Por otro lado, el ANN tiene un interés por la forma de cómo se conectan los elementos procesadores. La red típica consiste en una serie de capas de conexiones entre capas adyacentes consecutivas.

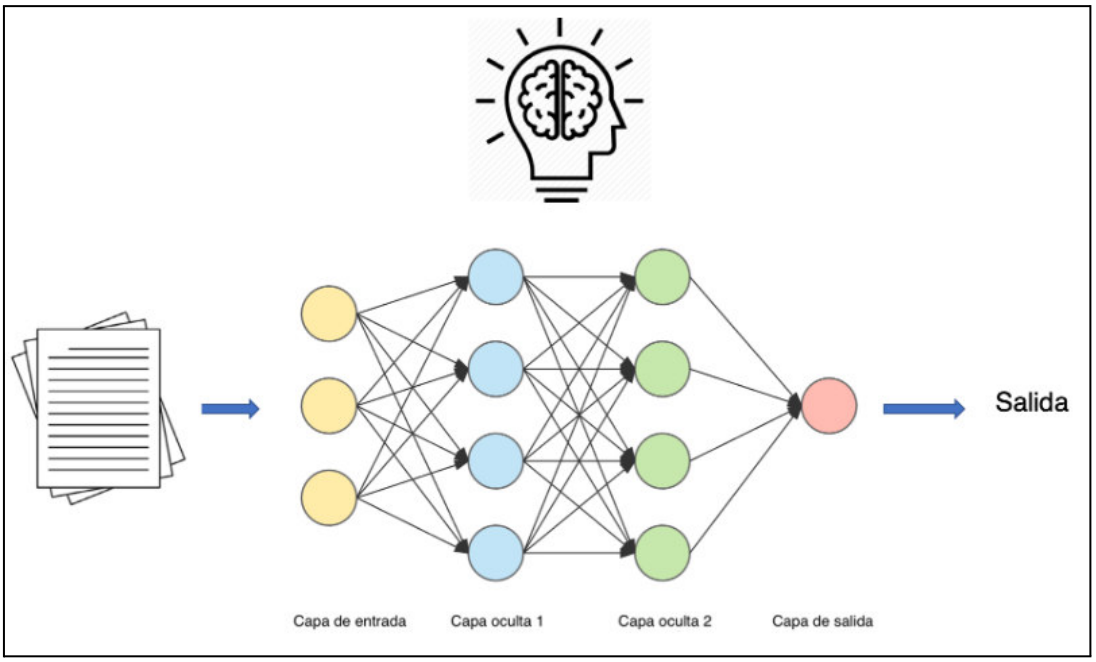


Gráfico 22: "Arquitectura de una Red Neuronal Simple"

Fuente: Basogsin, X., s.f., pág. 3

Las redes neuronales artificiales son una muestra del proceso evolutivo y de sofisticación propio del ser humano. Partiendo de una idea simple, donde años después está tecnología ha tenido un crecimiento vertiginoso por ende nos han permitido solucionar problemas hasta el momento imposibles. Adicionalmente, el sistema es capaz de aprender por sí mismo en lugar de ser programado, por lo que la clasifican como una de las grandes ventajas que tiene.

Las redes neuronales artificiales cuentan con las siguientes características principales similares a las del cerebro:

- Cuenta con un **aprendizaje adaptativo** en la cual hace mención que las redes neuronales aprenden mediante una base a un entrenamiento donde no es necesario hacer uso de un modelo previo.
- Cuenta con una **autoorganización,** donde una red neuronal crea su organización mediante un entrenamiento y de esta manera las partes de la red se encargan de identificar los patrones.
- Presenta **tolerancia a fallos,** es decir que cuentan con la capacidad de poder de eliminar el ruido de los datos.

- Brinda una **operación en tiempo real,** al tener las redes neuronales entrenadas se efectúa el reconocimiento y la clasificación de los datos en tiempo real.
- Rápida i**nserción en la tecnología** existente, debido a la naturaleza matricial de las operaciones para el entrenamiento se puede crear chips que se encarguen de agilizar dichas tareas.

Como podemos visualizar en la siguiente gráfica la red neuronal se componen por neuronas interconectadas y estructuradas con tres capas, en el cual el último puede tener variaciones. Tomar en consideración que los datos ingresan mediante la capa de entrada la cual pasan por la capa oculta para que finalmente salgan por la capa de salida (cuenta con varias capas).

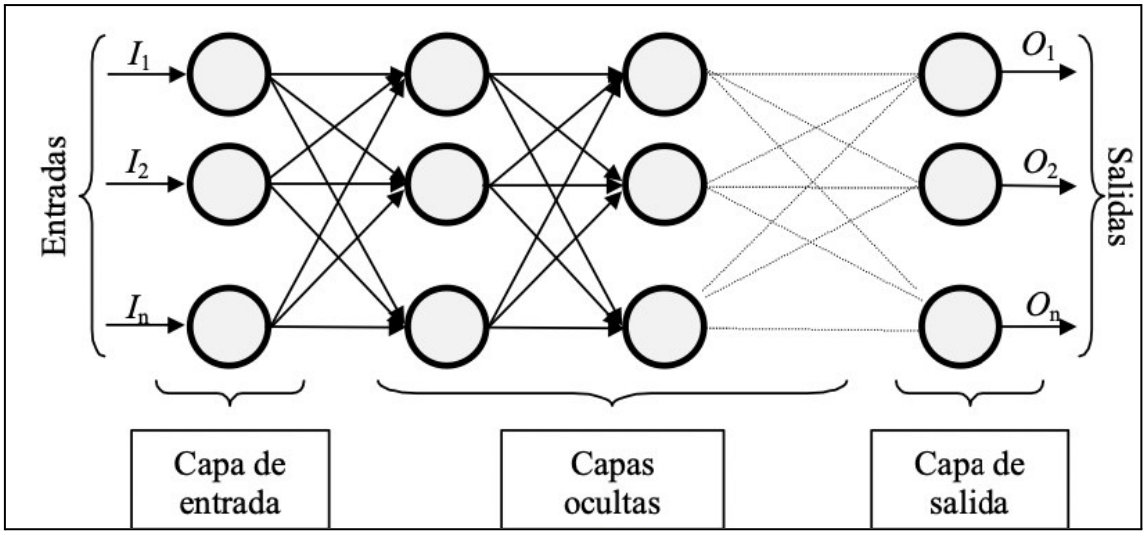


Gráfico 23: "Elementos básicos de la red neuronal"
Fuente: Matich, D. (2001)

Se debe tener en cuenta los mecanismos de aprendizaje, las cuales permitirán entender con mayor profundidad sobre las redes neuronales, de esta manera, Matich, D. (2001) sostienen que:

> Se ha visto que los datos de entrada se procesan a través de la red neuronal con el propósito de lograr una salida. También se dijo que las redes neuronales extraen generalizaciones desde un conjunto determinado de ejemplos anteriores de tales problemas de decisión. Una red neuronal debe aprender a calcular la salida correcta para cada constelación (arreglo o vector) de entrada en el

conjunto de ejemplos. Este proceso de aprendizaje se denomina: *proceso de entrenamiento o acondicionamiento*. El conjunto de datos o conjunto de ejemplos sobre el cual este proceso se basa es, por ende, llamado: *conjunto de datos de entrenamiento*.

2.2.5 Metodología para la aplicación de Machine Learning

Como en toda ciencia que emplea datos, se requiere un proceso o serie de pasos para poder llevar a cabo una correcta aplicación de las técnicas con el fin de obtener el resultado o predicción correcta. De esta manera, Harrington (2012) sostiene que:

> Nuestro enfoque para comprender y desarrollar una aplicación utilizando machine learning seguirá un procedimiento similar a este:
>
> 1. Recolectar datos
> 2. Preparar los datos de entrada: Una vez que tenga los datos, debe asegurarse de que estén en un formato utilizable.
> 3. Analizar los datos de entrada: (...) Asegúrese de que los pasos 1 y 2 realmente funcionen y que no tenga varios valores vacíos.
> 4. Si está trabajando con un sistema de producción y sabe cómo deberían verse los datos, o confía en su fuente, puede omitir este paso. Este paso requiere la participación humana.
> 5. Entrenar el algoritmo: Usted alimenta el algoritmo con datos limpios de los primeros dos pasos y extrae conocimiento o información. En el caso del aprendizaje no supervisado, no hay un paso de entrenamiento porque no tienes un valor objetivo.
> 6. Evaluar el algoritmo: Cuando esté evaluando un algoritmo, lo probará para ver qué tan bien funciona. En el caso del aprendizaje supervisado, tiene algunos valores conocidos que puede usar para evaluar el algoritmo. En el aprendizaje no supervisado, es posible que deba usar otras métricas para evaluar el éxito.
> 7. Usar el algoritmo: Aquí haces un programa real para hacer alguna tarea, y una vez más ves si todos los pasos anteriores funcionaron como esperabas. (Harrington, 2012, p. 13).

De esta manera, en primer lugar se deben obtener los datos en caso no los poseas. Se pueden obtener a través de sensores, cámaras, base de datos de tu empresa,

etc. Acto seguido, se deben preparar, es decir, corregir errores en los datos (limpieza) o entender las razones en caso sean datos atípicos, asimismo, puede requerir otras actividades como adaptar los datos al formato requerido. Como tercer paso se tiene en análisis, lo cual abarca la selección y aplicación de la técnica a usar. Posteriormente se presentarán los resultados obtenidos en el análisis para poder realizar los ajustes necesarios en los pasos anteriores, de ser requeridos, o continuar con el último paso. Finalmente, si todo está conforme, el modelo puede ser aplicado en la empresa o con el fin requerido.

2.2.6 Métricas

Una vez construído el modelo y aplicado a los datos que se poseen, debe evaluarse el resultado de dicha aplicación, es decir, emplear una métrica acorde para determinar si el resultado es aceptado o requiere mayor evaluación. Respecto a las métricas de evaluación de modelos, Reina & Solís (2018) sostienen que:

> Todas las metodologías necesitan una función de evaluación métrica para estimar cuantitativamente la capacidad de generalización del modelo, es decir, su desempeño sobre la distribución completa de posibles datos (y no solo sobre el conjunto usado para aprender) para comparar diversas opciones en la fase ajuste y validación. La más simple: tasa de aciertos en la predicción (accuracy) pero hay muchas otras y no necesariamente tiene que ser un solo número. La elección de una medida concreta debe ser guiada por el objetivo final. (Reina & Solís , 2018, p. 103).

Asimismo, mediante la siguiente tabla se indica de manera resumida las métricas de evaluación de modelos:

Métrica	Definición	Expresión
Tasa de Aciertos	Predicciones correctas sobre el total de predicciones.	$\frac{TP+TN}{TP+TN+FP+FN}$
Tasa de Errores	Predicciones incorrectas sobre el total de predicciones.	$\frac{FP+FN}{TP+TN+FP+FN}$
Accuracy (Exactitud)	Número total de predicciones correctas del modelo, es decir, proporción de aciertos en la clasificación	$\frac{TP+TN}{TP+TN+FP+FN}$
Precisión	Qué porcentaje de clientes de los que se contacten en la campaña de marketing del banco, estarán interesados en adquirir el depósito a plazo. (Calidad del modelo en la tarea de clasificación). Qué tan preciso ha sido el modelo para detectar la predicción sobre los clientes que SI tomarán el depósito a plazo**Precisión (P) alta significa pocos falsos positivos**	$\frac{TP}{TP+FP}$
Recall (Sensibilidad, Exhaustividad)	Qué porcentaje de los clientes que están interesados en tomar el depósito, es capaz el modelo de identificar.	$\frac{TP}{TP+FN}$
Tasa FP (Falsos Positivos) o Tasa de Error Tipo I	Proporción de negativos, clasificados como positivos.	$\frac{FP}{TP+FN}$
Tasa FN (Falsos Negativos) o Tasa de Error Tipo II	Número de elementos identificados erróneamente como negativos del total de verdaderos positivos. Número de elementos identificados como negativos de manera equivocada.	$\frac{FN}{TP+FN}$
Medida F1 (F measure o F1 Score)	Resume la precisión y el recall con el fin de comparar el rendimiento entre varias soluciones.	$F1 = 2 \cdot \frac{P \cdot R}{P+R}$
Especificidad	Mide qué tan exacta es la asignación a la clase negativa. Número de ítems correctamente identificados como negativos. Es lo opuesto al recall. En este caso, que el cliente no se haya suscrito a un depósito a plazo (NO).	$\frac{1-FPR}{\frac{TN}{FP+TN}}$

Tabla 04: "Métricas de evaluación de modelos"

Fuente: Lena & García, 2021, p. 95

Como se observa en la tabla, la métrica Accuracy (exactitud) mide o permite evaluar las predicciones correctas, las cuales, en un modelo con categoría de clasificación, permite evaluar qué tan bien se han acertado las predicciones o clasificaciones.

Matriz de confusión (Confusion Matrix)

Una herramienta que permite representar gráficamente los resultados de las técnicas de aprendizaje supervisado enfocadas en la clasificación es la matriz de confusión. "Existe una herramienta comúnmente utilizada en el aprendizaje automático

que le brinda una mejor vista de los errores de clasificación llamada matriz de confusión" (Harrington, 2012, p. 143).

Esta matriz de confusión muestra las predicciones en una tabla de doble entrada de la correcta categoría de las clases vs. la predicción de la categoría llevada a cabo por el modelo. Mueller & Massaron (2021) sostienen que:

> (...) una matriz de confusión, que es una herramienta muy útil en las clasificaciones para comparar sus predicciones con las etiquetas correctas (la llamada verdad fundamental - ground truth). En una matriz de confusión, lee una tabla donde las filas representan las clases correctas y las columnas las predichas (Mueller & Massaron, 2021, p. 235).

De esta manera, se tendrían 4 resultados para un problema de predicción con una función objetivo binaria, es decir, donde se buscan predecir 2 clases (A o B, 0 o 1, sí o no, etc.). Kelleher, et. al (2015) afirman que:

> (...) solo hay cuatro resultados cuando el modelo hace una predicción:
>
> - Verdadero positivo (TP): una instancia en el conjunto de prueba que tenía un valor de característica de destino positivo y que se predijo que tendría un valor de característica de destino positivo
> - Verdadero negativo (TN): una instancia en el conjunto de prueba que tenía un valor de característica de destino negativo y que se predijo que tendría un valor de característica de destino negativo
> - Falso positivo (FP): una instancia en el conjunto de prueba que tenía un valor de característica de destino negativo pero que se predijo que tendría un valor de característica de destino positivo
> - Falso negativo (FN): una instancia en el conjunto de prueba que tenía un valor de característica de destino positivo pero que se predijo que tendría un valor de característica de destino negativo (Kelleher, et. al, 2015, p. 431).

A continuación se presenta un ejemplo gráfico de una matriz de confusión:

		Prediction	
		positive	negative
Target	positive	*TP*	*FN*
	negative	*FP*	*TN*

Gráfico 24: La estructura de una matriz de confusión

Fuente: Kelleher, et. al, 2015, p. 432

2.2.7 Cotización

Respecto a las cotizaciones, la OCDE (2017) indica:

> El precio de cotización de los productos básicos normalmente refleja un acuerdo entre compradores y vendedores independientes del sector, en relación con el precio y la cantidad de un tipo específico de producto básico, negociando según unas condiciones concretas y en un momento dado.(...) Por tanto, dependiendo de los hechos y circunstancias de cada caso, los precios de cotización podrán servir de referencia para determinar el precio de las operaciones con productos básicos entre empresas asociadas. Los contribuyentes y las administraciones tributarias deben aplicar el precio de cotización seleccionado de forma coherente. (Directrices de la OCDE aplicables en materia de precios de transferencia a empresas multinacionales y administraciones tributarias, 2017, p.116).

Así mismo, la OCDE (2017) indica:

> Cuando existan diferencias entre las condiciones de las operaciones vinculadas y las condiciones de las operaciones no vinculadas o las que determinan el precio de cotización del producto básico, que afecten significativamente al precio de las operaciones con productos básicos examinadas, será necesario practicar ajustes razonables precisos para garantizar que las características económicas relevantes de las operaciones son comparables. (Directrices de la OCDE aplicables en materia de precios de transferencia a empresas multinacionales y administraciones tributarias, 2017,p. 117).

2.2.8 Montacarga

En la presente investigación se hará el estudio del estado de cotizaciones sobre las montacargas, las cuales son parte de máquinas ligeras. Las montacargas son un tipo de producto que vende y alquila la empresa. Respecto a las montacargas, Santiago et. al (2007) indican que:

> Las montacargas son para transporte exclusivo de materiales, accionado por cabrestante, por ser este el tipo de aparato más frecuente. Sin embargo, gran parte de las características o dispositivos que citaremos son perfectamente aplicables a montacargas-elevadores movidos por cables o de cremallera, efectuando las debidas correcciones u observaciones complementarias. (Espeso Santiago, Fernández Zapico, Espeso Expósito, & Ferández Muñiz, 2007, p.168)

Así mismo, "Un montacargas tiene unos elementos estructurales característicos, que son base de apoyo, estructura flotante y sistema de deslizamiento" (Santiago et. al, 2007, p.168).

2.2.9 Variables Independientes

Las variables dependientes para el siguiente proyecto a considerar son los siguientes:

- Vendedor (código de vendedor)
- Sucursales (código de oficina)
- Tiempo de entrega del producto
- Precio del producto
- Toneladas del producto
- Cliente (código de cliente)
- Mercado o rubro del cliente (código de mercado)
- Mes de la cotización
- Año de la cotización
- Condición de pago de la cotización

2.2.10 Variable Dependiente

La variable dependiente del trabajo de investigación es el estado de cotizaciones. Tomar en consideración que, previamente al pre-análisis y limpieza de datos,

actualmente existen tres estados: aprobado (se ha concretado la venta), en proceso (aún se encuentra en negociación) y rechazado (el cliente desistió de la compra).

CAPÍTULO III: ENTORNO EMPRESARIAL

3.1 Descripción de la empresa

La empresa nace de la fusión de dos empresas líderes en el mercado (Unimaq y Rentando Cat Rental Store) del grupo Ferreycorp, con el objetivo de especializarse en la venta y alquiler de equipos ligeros. La empresa actualmente cuenta con sedes en Lima, Arequipa, Cajamarca, Piura, Trujillo, Huancayo, Ilo, Chiclayo y Cusco.

En el año 2009, obtuvo la representación oficial de la línea de Construcción General de Caterpillar, anteriormente comercializada por Ferreyros. Años más tarde, otras líneas de máquinas y equipos como Blend y Rival le encomendaron su representación.

En el año 2015, recibió el premio al mejor distribuidor de equipos ligeros Caterpillar a nivel mundial gracias a su desempeño comercial. Asimismo, fue galardonada por Mitsubishi Cat como el mejor distribuidor de Latinoamérica.

En el año 2017, lanzó la nueva plataforma de venta de repuestos PARTS.CAT.COM como un nuevo canal de ventas enfocado en el cliente a fin de agilizar sus procesos.

La empresa, en la actualidad, está presente en los mercados de Construcción, Minería, Energía, Logística, Industria, Puertos marítimos y Saneamiento, dedicándose a la comercialización y alquiler de maquinaria ligera brindando, además, un soporte de post-venta, como lo son el soporte técnico y opciones de reparación de las maquinarias.

3.1.1 Reseña histórica y actividad económica

A continuación se brindan los datos de la empresa:

- **RUC:** Anónimo
- **Razón Social:** Anónimo
- **Nombre Comercial:** Anónimo
- **Tipo Empresa:** Sociedad Anónima
- **Fecha Inicio Actividades:** 12/12/1956
- **Actividad Comercial:** Venta al por mayor de otros tipos de maquinaria y equipos
- **Dirección Legal:** Anónimo
- **Departamento / Ciudad / Distrito:** Lima / Lima / Ate

3.1.2 Descripción de la organización

3.1.2.1 Organigrama

A continuación, se muestra el organigrama funcional:

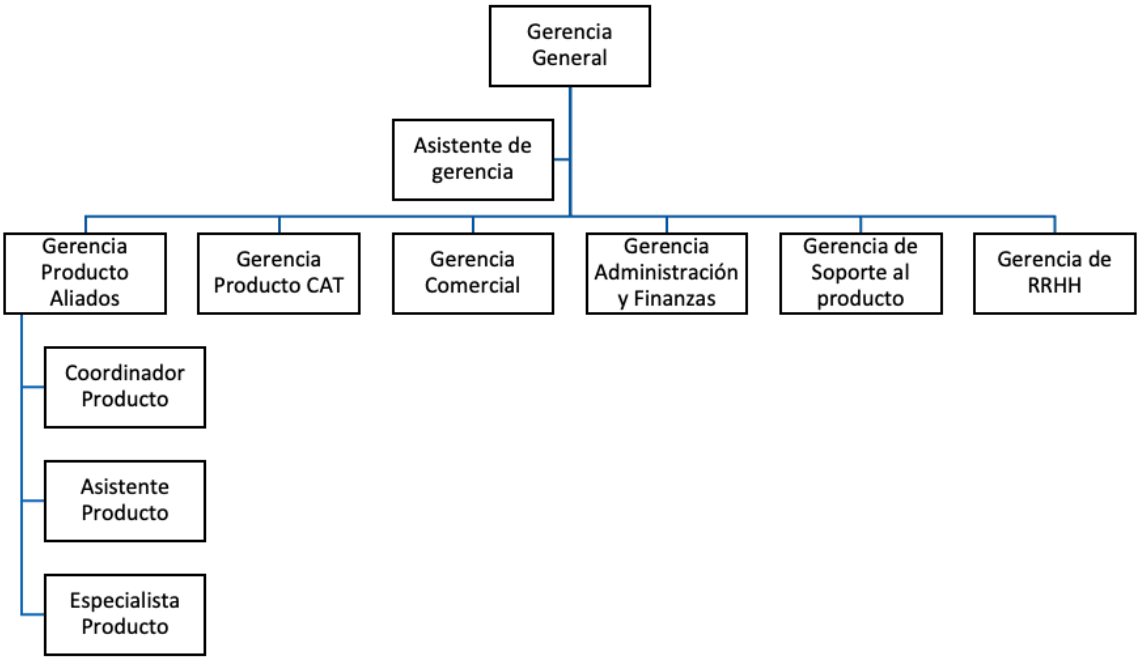


Gráfico 25: "Organigrama de la empresa"
Fuente: La empresa

Tal como se observa en la imagen anterior, la empresa cuenta con distintas gerencias las cuales están en constante comunicación y coordinación para cumplir los objetivos principales.

En la presente investigación nos enfocamos en la línea de montacargas, la cual forma parte de la Gerencia de Producto. Esta área está conformada por: Coordinador de producto, Asistente de Producto y el Especialista de Producto.

- Coordinador de producto: Colaborador encargado de la coordinación con la fábrica para la realización de capacitaciones, compras y actualizaciones de costos.
- Asistente producto: Colaborador encargado de las actividades administrativas del área.
- Especialista de producto: Colaborador encargado de brindar soporte técnico al vendedor para establecer las especificaciones del montacarga de acuerdo a la necesidad del cliente.

3.1.2.2 Cadena de suministros

A continuación se presenta la Cadena de Suministros:

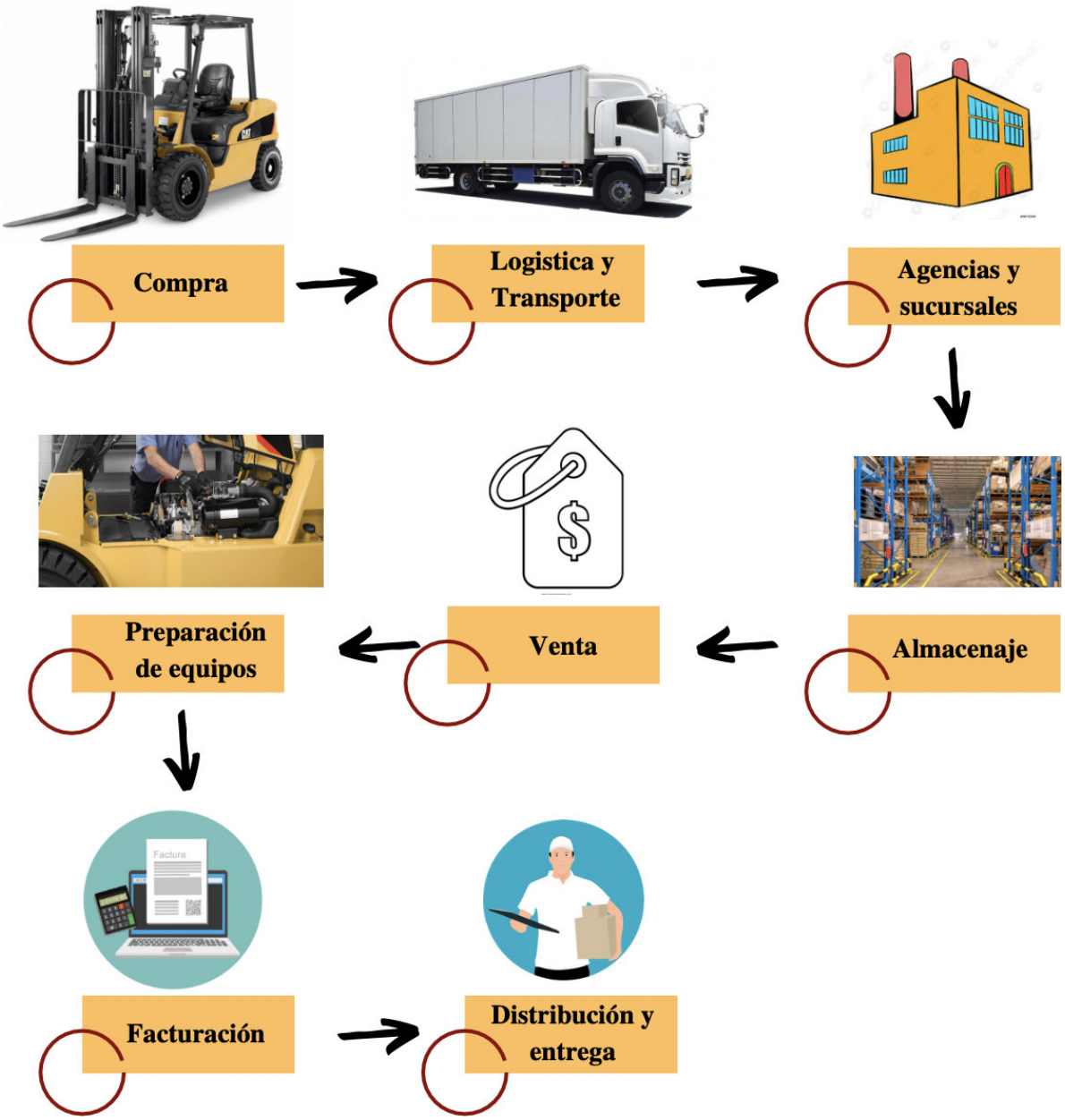


Gráfico 26: "Cadena de suministros"
Fuente: Elaboración propia

3.1.3 Datos generales estratégicos de la empresa

La empresa presenta una misión, visión y valores marcados desde sus inicios, los cuales se detallan en las siguientes líneas, así como también los objetivos estratégicos.

3.1.3.1 Visión, misión y valores o principios

Misión:

Satisfacer las necesidades de nuestros clientes mediante soluciones integrales en equipos ligeros a través de un amplio portafolio de marcas, productos e infraestructura a nivel nacional; así como un servicio eficiente, ágil y de calidad con personal altamente motivado y calificado.

Visión:

Ser la mejor opción en soluciones integrales de equipos ligeros en un solo lugar y líderes en cada línea que representamos.

Valores:

- Innovación
- Integridad
- Compromiso
- Vocación de servicio
- Dinamismo

3.1.3.2 Objetivos estratégicos

Se tienen los siguientes objetivos estratégicos:

- OE1: Adecuar el negocio y la empresa para garantizar la rentabilidad adecuada
- OE2: Óptima satisfacción del cliente gracias a la búsqueda de la excelencia operacional
- OE3: Cubrir adecuada y eficientemente el mercado
- OE4: Uso eficiente de los activos

3.1.3.3 Evaluación interna y externa. FODA cuantitativo

- **Fortalezas**:

  F1: Distribuidores exclusivos de la marca Caterpillar.

  F2: Se cuenta con técnicos certificados por la institución IPAF.

  F3: Incremento de proyectos de mejora relacionados a la aplicación y uso de tecnología.

F4: Se obtuvo por tercer año consecutivo el premio al mejor distribuidor de equipos ligeros Caterpillar.

F5: Se cuenta con las certificaciones ISO 9001:2015, ESR, ABE y OEA.

- **Oportunidades**:

  O1: Aumento de la producción en ciertos sectores a nivel nacional debido a la reactivación económica.

  O2: Tendencia nacional e internacional hacia la automatización de procesos.

  O3: Posibilidad de incrementar la participación de mercado al aumentar la producción de los sectores de nuestros clientes.

- **Debilidades**:

  D1: Demora en el inicio del proceso de compras (importación)

  D2: Precios elevados con respecto a la competencia

  D3: Ausencia de procesos automatizados

- **Amenazas**:

  A1: Incremento de empresas Chinas en el mercado peruano con precios y tiempos de importación más competitivos

  A2: La competencia cuenta con planes de marketing más efectivos y desarrollados

  A3: Incremento del costo de los componentes y materiales eléctricos para fabricación de motores, además del costo de importación

  A4: Retraso en las importaciones por altos niveles de congestión logística

Factores internos determinantes de éxito	Peso	Calificación	Peso Ponderado
Fortalezas			
Distribuidores exclusivos de la marca Caterpillar	14%	4	0,56
Se cuenta con técnicos certificados por la institución IPAF	5%	3	0,15
Incremento de proyectos de mejora relacionados a la aplicación y uso de tecnología	14%	4	0,56
Se obtuvo por tercer año consecutivo el premio al mejor distribuidor de equipos ligeros Caterpillar	8%	3	0,24
Se cuenta con las certificaciones ISO 9001:2015, ESR, ABE y OEA	8%	3	0,24
Debilidades			
Demora en el inicio del proceso de compras (importación)	18%	1	0,18
Precios elevados con respecto a la competencia	18%	1	0,18
Ausencia de procesos automatizados	15%	2	0,30
Total	**1,00**		**2,41**

Factores externos determinantes de éxito	Peso	Calificación	Peso Ponderado
Oportunidades			
Aumento de la producción en ciertos sectores a nivel nacional debido a la reactivación económica	16%	4	0,64
Tendencia nacional e internacional hacia la automatización de procesos	16%	3	0,48
Posibilidad de incrementar la participación de mercado al aumentar la producción de los sectores de nuestros clientes	23%	4	0,92
Amenazas			
Incremento de empresas Chinas en el mercado peruano con precios y tiempos de importación más competitivos	8%	2	0,16
La competencia cuenta con planes de marketing más efectivos y desarrollados	5%	2	0,10
Incremento del costo de los componentes y materiales eléctricos para fabricación de motores, además del costo de importación	18%	1	0,18
Retraso en las importaciones por altos niveles de congestión logística	14%	1	0,14
Total	**1,00**		**2,62**

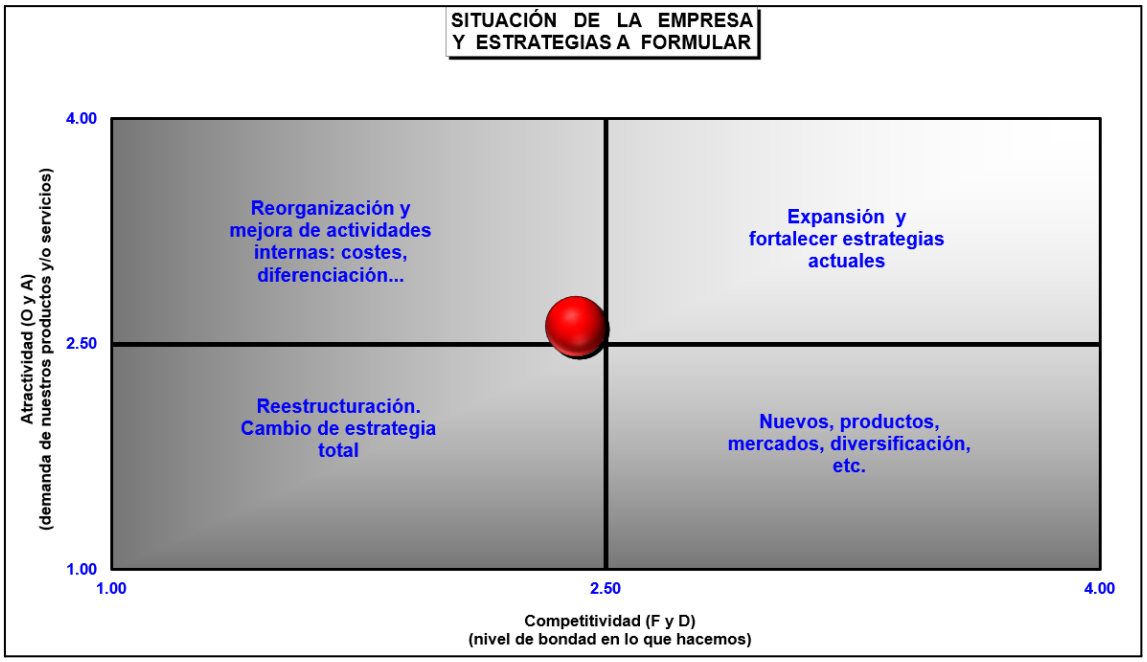


Gráfico 27: "Análisis de la situación interna y externa"
Fuente: Elaboración propia

Como se observa en el gráfico y de acuerdo a la situación actual de la empresa en base al análisis FODA cuantitativo, esta se encuentra en el primer cuadrante, por lo tanto, la estrategia a realizar debe estar enfocada hacia la reorganización y mejora de actividades internas.

3.2 Modelo de negocio actual (CANVAS)

A continuación se presenta el CANVAS de la empresa:

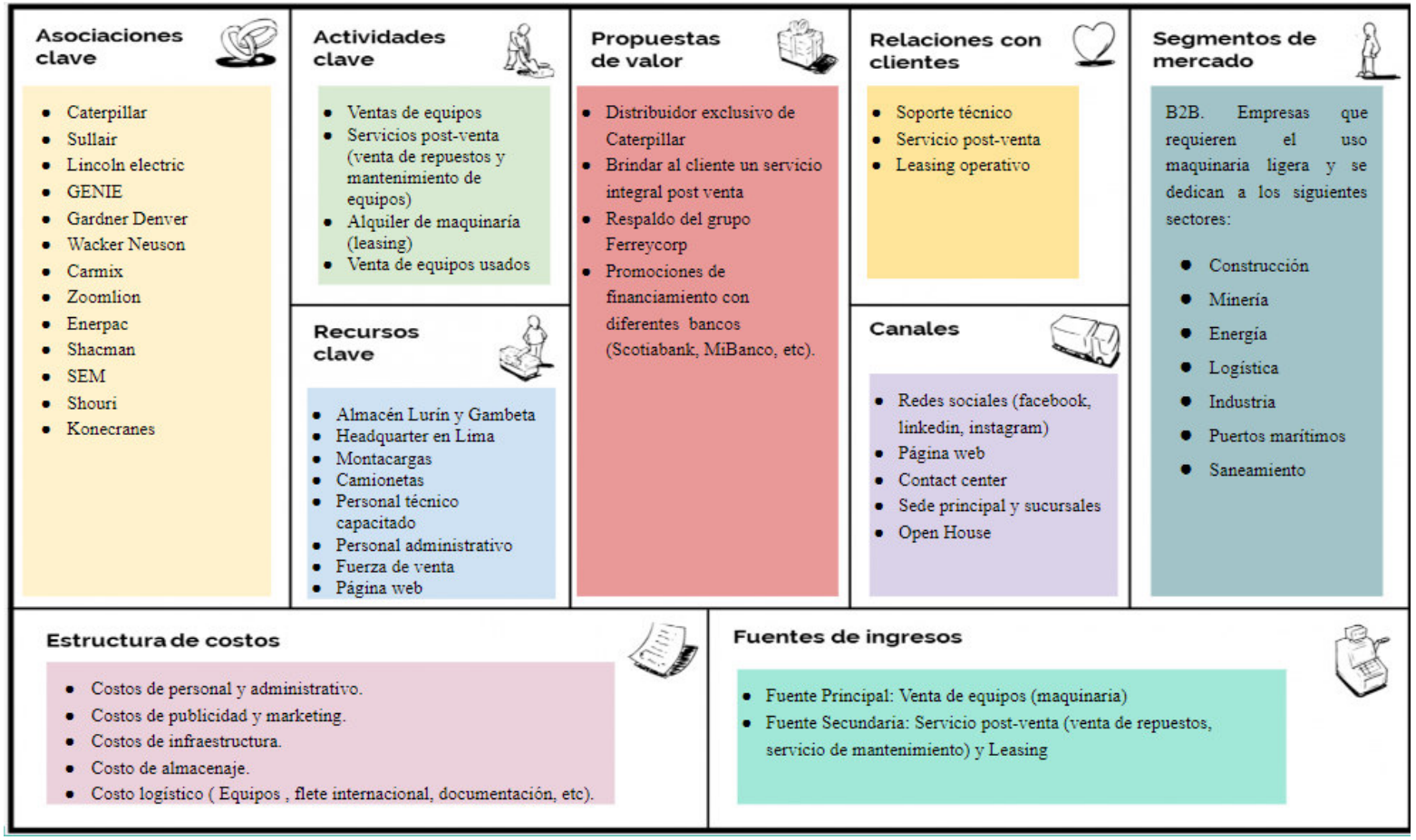


Gráfico 28: "Modelo de negocio de la empresa"
Fuente: Elaboración propia

3.3 Mapa de Procesos actual

A continuación se muestra el Mapa de Procesos actual de “La empresa”:

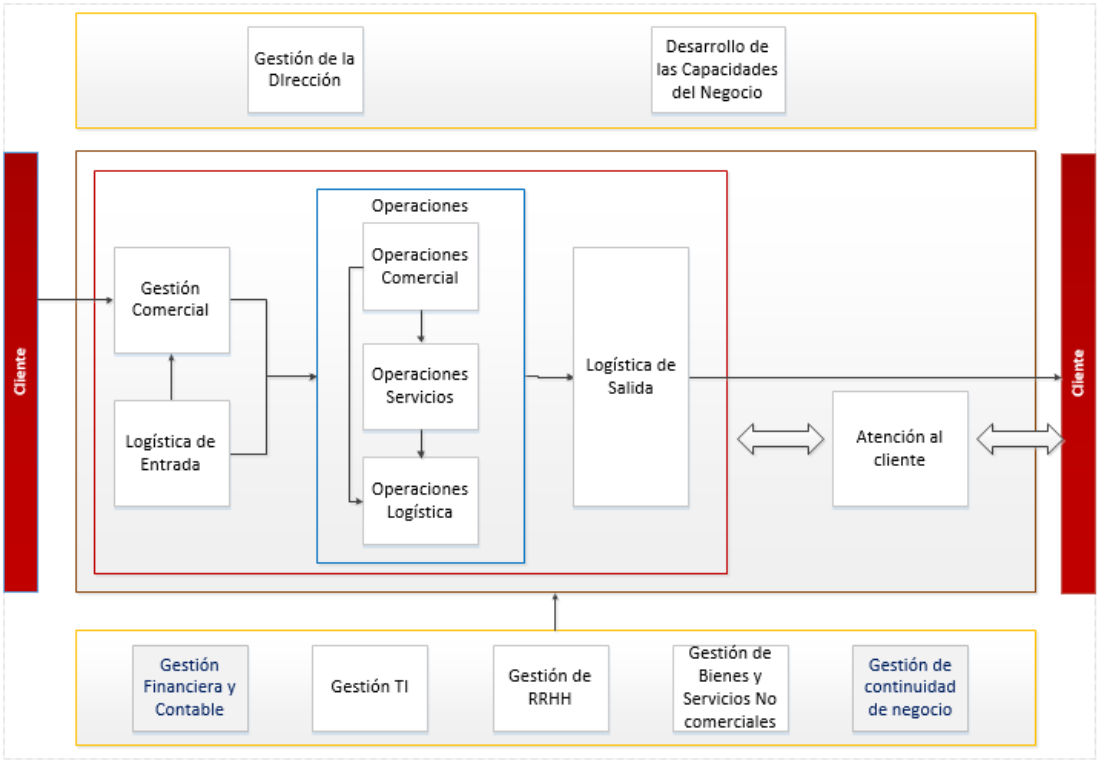


Gráfico 29: “Mapa de procesos de la empresa”
Fuente: La empresa

A continuación, se brinda una breve descripción de los procesos:

- Procesos Estratégicos
    - Gestión de la Dirección: Velar por el adecuado liderazgo a los colaboradores para gestionar correctamente los procesos del negocio.
    - Desarrollo de las Capacidades del Negocio: Es el área encargada de velar por las correctas aplicaciones de estrategias para aumentar la ventas y brindar el enfoque del negocio en las áreas de la empresa. Para el crecimiento del negocio tienen la posibilidad de agregar nuevas líneas aliadas que agreguen valor a la empresa.
- Procesos Core
    - Gestión Comercial: Se encarga del manejo de las ventas y promociones de los equipos acorde al mercado. Además, es la encargada de velar por el cumplimiento de los objetivos de ventas mensuales y satisfacción del cliente, la cual se mide con un indicador NLS.

- Logística de Entrada: Se encarga de realizar las compras de los equipos para el abastecimiento del stock acorde a la demanda del mercado. Adicionalmente, se hacen responsable de las compras indirectas a las ventas de los equipos para brindar un servicio personalizado a comparación de la competencia.
- Operaciones:
  - Operaciones Comerciales: Hace referencia a las actividades diarias de la fuerza comercial quienes hacen posible que las ventas sean concretadas con un asesoramiento técnico. Entre los objetivos principales se tiene los siguientes: brindar una solución integral, incrementar los márgenes y mantener el liderazgo en el mercado.
  - Operaciones Servicios: Es la que se encarga de la parte intangible del negocio como los servicios post-venta. Adicionalmente, el área se encarga de brindar entregas técnicas a los clientes en la cual se da conformidad de lo acordado con el cliente y, de está manera, activar la garantía del equipo. Esta garantía consiste en cubrir los gastos de reparación y/o reposición de repuestos si es que presenta problemas de fábrica. Finalmente, se realizan mantenimientos a los equipos cada cierto tiempo cuando se requiera.
  - Operaciones Logística: Se encarga de todas las actividades relacionadas a compras, distribución y almacenaje.
- Logística de Salida: Se encarga de realizar las gestiones para la distribución de los pedidos hacia los puntos de operaciones de los clientes así como el abastecimiento adecuado a las sucursales de la empresa para contar con stock a nivel nacional.
- Atención al cliente: Velar por la entrega de productos cumpliendo con los estándares de calidad a sus clientes, además de brindar un correcto soporte en el mantenimiento de las maquinarias. Es importante mencionar que el tiempo de entrega de los pedidos de los clientes es un tema crítico a la hora de evaluar los indicadores de satisfacción.

- Procesos de Soporte:
    - Gestión financiera y contable: Velar por la rentabilidad de la empresa en base a su giro del negocio.
    - Gestión de TI: Velar por el correcto funcionamiento de los sistemas de Tecnología e Información de la empresa, así como apoyar en proyectos tecnológicos. Adicionalmente, con la colaboración del área de procesos, cuentan con proyectos de mejora continua de la herramienta principal de la empresa (SAP logon).
    - Gestión de RRHH: Velar por la obtención, desarrollo y mantenimiento del recurso humano de la empresa.
    - Gestión de Bienes y Servicios No Comerciales: Velar por la disponibilidad de insumos no comerciales que son requeridos por los demás procesos de la empresa.
    - Gestión de Continuidad del Negocio: Velar por el flujo ininterrumpido del negocio (Core).

CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN

4.1 Diseño de la investigación

4.1.1 Enfoque de la investigación

El enfoque de la presente investigación es cuantitativo debido a que se propone un conjunto de modelos predictivos que hacen uso de técnicas estadísticas de Machine Learning. Cabe resaltar que todos estos modelos podrán predecir el estado de las cotizaciones de la empresa y también podrán medir el porcentaje de confiabilidad a través de técnicas estadísticas.

4.1.2 Alcance de la investigación

El alcance de la investigación es correlacional dado que analiza la relación entre dos o más variables. En este caso, en base a la correlación entre la variable dependiente con las independientes, se busca predecir el estado de una cotización (aprobada o rechazada) a través de las variables independientes definidas.

4.1.3 Diseño o tipo de investigación

El diseño de la investigación es experimental dado que se trabajan / analizan las variables tanto dependientes como independientes con el fin de predecir el resultado de la variable dependiente, haciendo uso del tipo de aprendizaje supervisado (4 técnicas) de machine learning.

4.1.4 Población y muestra

En la siguiente tabla detallamos la población y muestra de la presente investigación:

Población:	Base de datos de las cotizaciones realizadas por el área de ventas en la empresa.
Muestra:	576 registros de las cotizaciones realizadas por el área de ventas de la empresa desde el año 2020 hasta junio del 2022 del tipo de producto montacargas.

Tabla 05: "Población y muestra de la investigación"

Fuente: Elaboración propia

4.2 Metodología de implementación de la solución

Teniendo como guía lo mencionado en las bases teóricas, se presenta la metodología de implementación de la solución a aplicar en el presente trabajo:

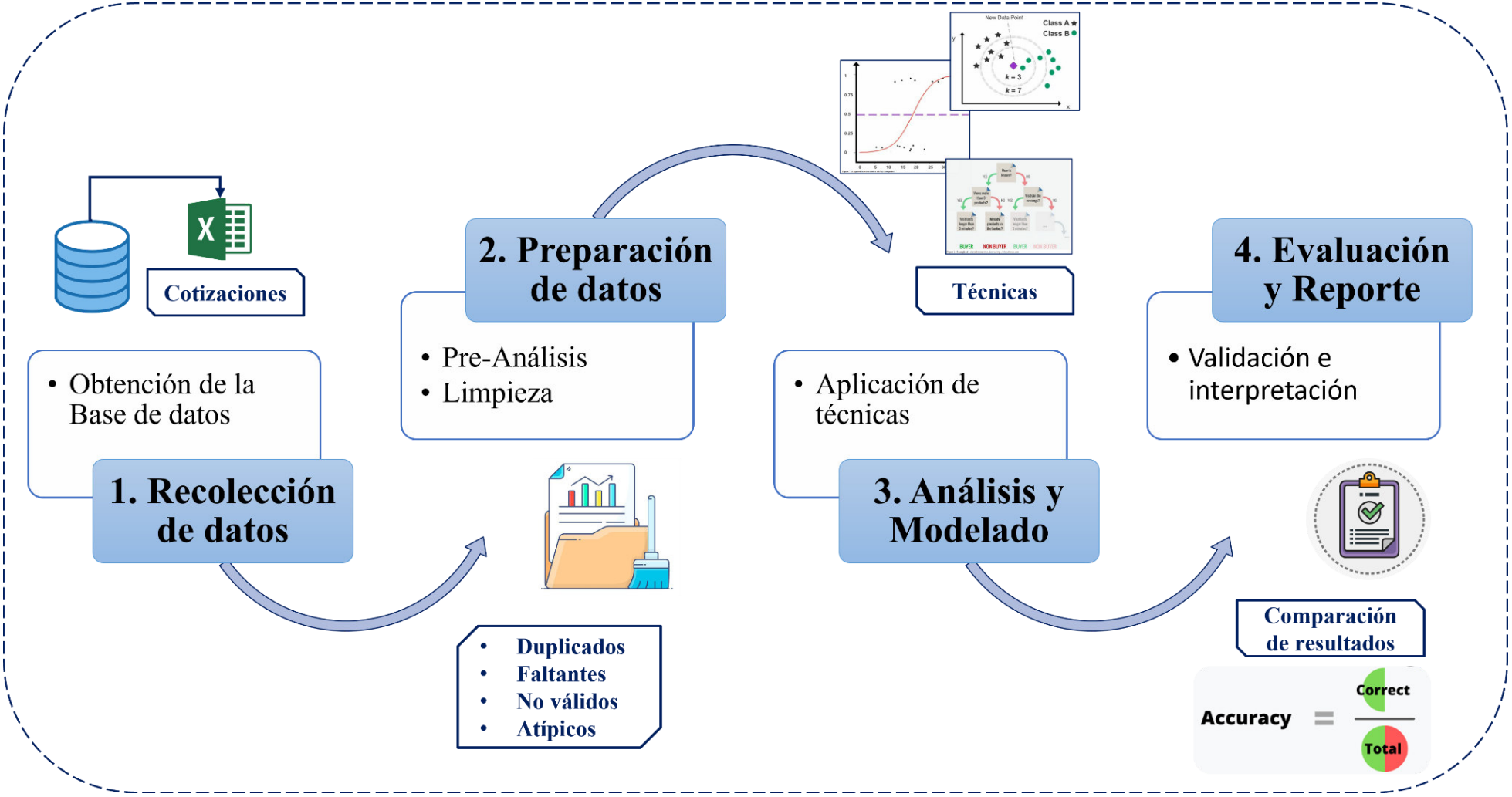


Gráfico 30: "Etapas de la Metodología de Implementación de la solución"

Fuente: Elaboración propia

Como se observa en el gráfico, la metodología está dividida en 4 pasos, siendo estos:

1. **Recolección de datos**: Involucra la obtención de los datos que emplearemos para llevar a cabo el presente proyecto de investigación. En el presente trabajo, los datos contienen información de las cotizaciones de la empresa desde el año 2020 hasta junio del 2022. Como variables X o independientes se poseen: Monto del equipo (MontoPosicion), peso en toneladas del equipo (Toneladas), tiempo de entrega del equipo (TiempoEntrega), código del vendedor (IDVendedor), código de oficina (CodOficina), código del cliente (CodCliente), código del mercado o rubro del cliente (CodMercado), mes de la cotización (MONTH), año de la cotización (YEAR), y condición de pago (Pago). Finalmente, como variable Y o dependiente se posee: Estado de la cotización (EstadoY).
2. **Preparación de datos**: En esta etapa llevaremos a cabo un análisis preliminar de los datos a fin de comprender su naturaleza y poder así llevar a cabo una correcta limpieza de datos. En esta limpieza buscaremos eliminar los problemas de calidad que pudieran estar presentes en los datos como, por ejemplo: Registros duplicados o valores repetidos, valores faltantes o ausentes, datos no válidos o nulos, y datos atípicos (outliers). En esta última situación, de presentarse el caso, evaluaremos si el dato es realmente atípico con el fin de llevar a cabo un análisis posterior.
3. **Análisis y modelado**: Abarca la selección de la categoría y las técnicas a aplicar en el presente trabajo de investigación, además de la construcción de los modelos en el software de interfaz Jupyter mediante el lenguaje Python. De acuerdo con los tipos de datos que se poseen para el presente trabajo, la categoría es Aprendizaje Supervisado dado que se busca predecir el resultado (aprobado o rechazado) de las cotizaciones, los cuales son datos conocidos, asimismo, se usarán 4 técnicas de las cuales se evaluará su nivel de accuracy o acierto en las predicciones. En esta etapa, en el software de interfaz Jupyter separaremos los datos en de *X_train*, *X_test*, *Y_train*, y *Y_test*. Cabe resaltar que, con respecto a la codificación en Python, se hará uso de las librerías *pandas* y *sklearn*. Asimismo, se hará uso de las técnicas: k-NN con apoyo de su herramienta *KNeighborsClassifier*, Regresión Logística con *LogisticRegression*, Support Vector Machine con *SVM*, y Árbol de decisión con *DecisionTreeClassifier*.

4. **Evaluación y Reporte**: En esta última etapa se evaluarán los resultados obtenidos en la predicción que llevaron a cabo los modelos. La métrica que se va a emplear para evaluar dicho resultado será el Accuracy dado que los datos están balanceados, es decir, ambos posibles resultados contienen la misma cantidad de datos para asegurar un correcto aprendizaje del modelo. Posteriormente, se mostrará el porcentaje total de cotizaciones correctamente clasificadas.

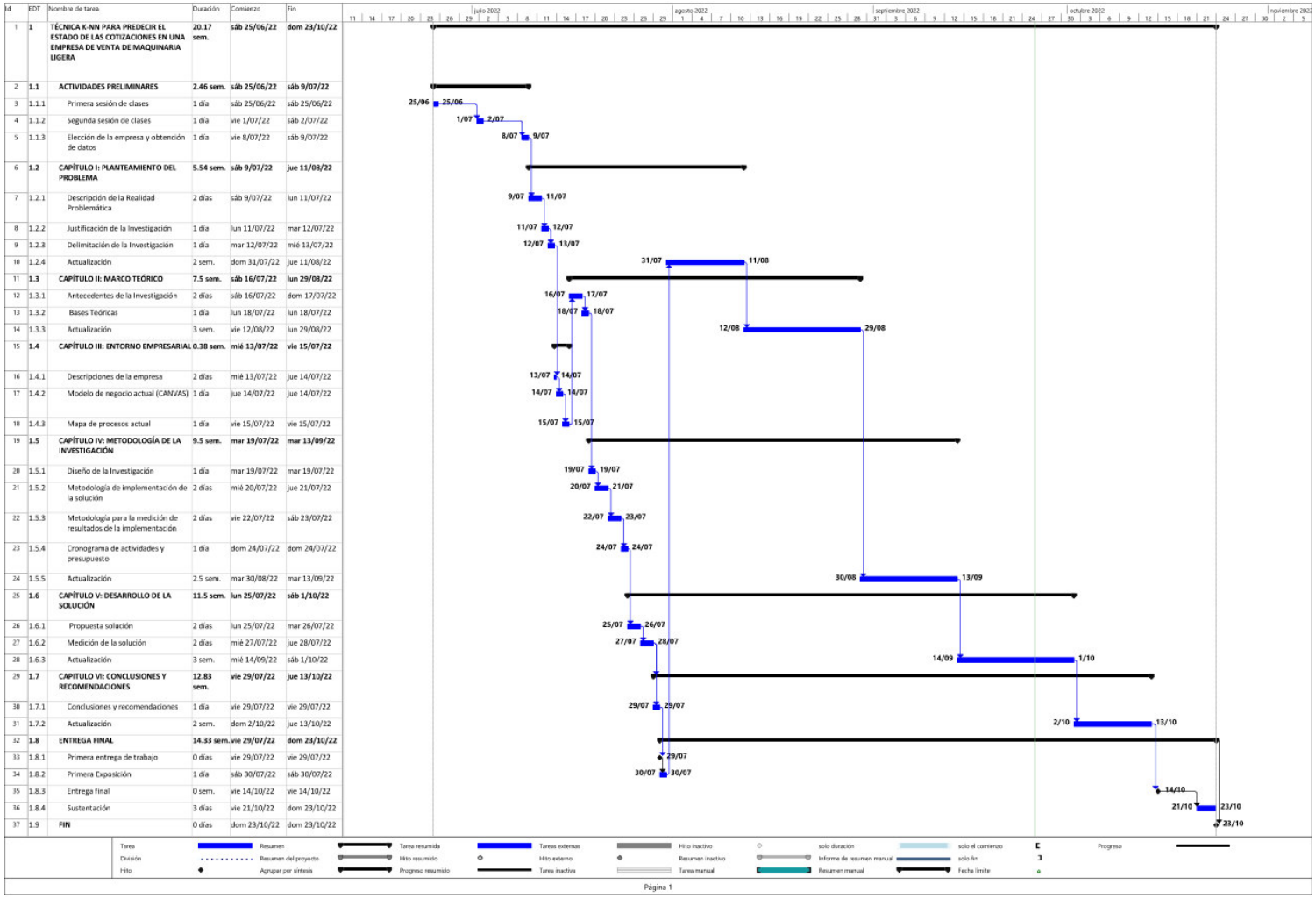
   Finalmente, si el resultado obtenido por el *accuracy* puede ser mejorado, se tomarán las acciones necesarias como la codificación y normalización de la base de datos, evaluando el nuevo valor de accuracy obtenido por los modelos. En caso sea favorable este último valor de accuracy, será tomado en cuenta, caso contrario, sería desestimado. En base a ello, se realizará una comparativa para evaluar el mejor accuracy de las diferentes técnicas usadas.

4.3 Metodología para la medición de resultados de la implementación

Para la evaluación del resultado de la aplicación del modelo planteado en el presente trabajo de investigación se emplea la métrica *Accuracy*, dado que, como se mencionó anteriormente, los datos o clases están balanceados.

Para poder aplicar esta métrica en la codificación mediante el lenguaje Python, haremos uso de la librería *sklearn*, específicamente *sklearn.metrics* para poder aplicar la herramienta *accuracy_score*. Esto nos permitirá evaluar el nivel de precisión en las predicciones llevadas a cabo por el modelo, es decir, con qué nivel de acierto se han logrado predecir correctamente el resultado de las cotizaciones.

4.4 Cronograma de actividades y presupuesto



Id	EDT	Nombre de tarea	Duración	Comienzo	Fin
1	1	**TÉCNICA K-NN PARA PREDECIR EL ESTADO DE LAS COTIZACIONES EN UNA EMPRESA DE VENTA DE MAQUINARIA LIGERA**	**20.17 sem.**	**sáb 25/06/22**	**dom 23/10/22**
2	1.1	**ACTIVIDADES PRELIMINARES**	**2.46 sem.**	**sáb 25/06/22**	**sáb 9/07/22**
3	1.1.1	Primera sesión de clases	1 día	sáb 25/06/22	sáb 25/06/22
4	1.1.2	Segunda sesión de clases	1 día	vie 1/07/22	sáb 2/07/22
5	1.1.3	Elección de la empresa y obtención de datos	1 día	vie 8/07/22	sáb 9/07/22
6	1.2	**CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA**	**5.54 sem.**	**sáb 9/07/22**	**jue 11/08/22**
7	1.2.1	Descripción de la Realidad Problemática	2 días	sáb 9/07/22	lun 11/07/22
8	1.2.2	Justificación de la Investigación	1 día	lun 11/07/22	mar 12/07/22
9	1.2.3	Delimitación de la Investigación	1 día	mar 12/07/22	mié 13/07/22
10	1.2.4	Actualización	2 sem.	dom 31/07/22	jue 11/08/22
11	1.3	**CAPÍTULO II: MARCO TEÓRICO**	**7.5 sem.**	**sáb 16/07/22**	**lun 29/08/22**
12	1.3.1	Antecedentes de la Investigación	2 días	sáb 16/07/22	dom 17/07/22
13	1.3.2	Bases Teóricas	1 día	lun 18/07/22	lun 18/07/22
14	1.3.3	Actualización	3 sem.	vie 12/08/22	lun 29/08/22
15	1.4	**CAPÍTULO III: ENTORNO EMPRESARIAL**	**0.38 sem.**	**mié 13/07/22**	**vie 15/07/22**
16	1.4.1	Descripciones de la empresa	2 días	mié 13/07/22	jue 14/07/22
17	1.4.2	Modelo de negocio actual (CANVAS)	1 día	jue 14/07/22	jue 14/07/22
18	1.4.3	Mapa de procesos actual	1 día	vie 15/07/22	vie 15/07/22
19	1.5	**CAPÍTULO IV: METODOLOGÍA DE LA INVESTIGACIÓN**	**9.5 sem.**	**mar 19/07/22**	**mar 13/09/22**
20	1.5.1	Diseño de la Investigación	1 día	mar 19/07/22	mar 19/07/22
21	1.5.2	Metodología de implementación de la solución	2 días	mié 20/07/22	jue 21/07/22
22	1.5.3	Metodología para la medición de resultados de la implementación	2 días	vie 22/07/22	sáb 23/07/22
23	1.5.4	Cronograma de actividades y presupuesto	1 día	dom 24/07/22	dom 24/07/22
24	1.5.5	Actualización	2.5 sem.	mar 30/08/22	mar 13/09/22
25	1.6	**CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN**	**11.5 sem.**	**lun 25/07/22**	**sáb 1/10/22**
26	1.6.1	Propuesta solución	2 días	lun 25/07/22	mar 26/07/22
27	1.6.2	Medición de la solución	2 días	mié 27/07/22	jue 28/07/22
28	1.6.3	Actualización	3 sem.	mié 14/09/22	sáb 1/10/22
29	1.7	**CAPITULO VI: CONCLUSIONES Y RECOMENDACIONES**	**12.83 sem.**	**vie 29/07/22**	**jue 13/10/22**
30	1.7.1	Conclusiones y recomendaciones	1 día	vie 29/07/22	vie 29/07/22
31	1.7.2	Actualización	2 sem.	dom 2/10/22	jue 13/10/22
32	1.8	**ENTREGA FINAL**	**14.33 sem.**	**vie 29/07/22**	**dom 23/10/22**
33	1.8.1	Primera entrega de trabajo	0 días	vie 29/07/22	vie 29/07/22
34	1.8.2	Primera Exposición	1 día	sáb 30/07/22	sáb 30/07/22
35	1.8.3	Entrega final	0 sem.	vie 14/10/22	vie 14/10/22
36	1.8.4	Sustentación	3 días	vie 21/10/22	dom 23/10/22
37	1.9	**FIN**	0 días	dom 23/10/22	dom 23/10/22

Gráfico 31: "Cronograma de la investigación"
Fuente: Elaboración propia

Presupuesto:

En la tabla 06 se muestra el presupuesto referencial de la presente investigación. Se ha tomado en cuenta las máquinas y equipos, servicios básicos y la mano de obra de los cinco integrantes del equipo.

Item	Cantidad	Importe por unidad (soles)	Semanas	Total (soles)
Maquinaria y equipos				
Laptops	5	S/.2.000,00		S/.10.000,00
Mouse	5	S/.50,00		S/.250,00
Sub total				**S/.10.250,00**
Mano de obra				
Bachilleres	5	S/.500,00	12	S/.30.000,00
Sub total				**S/.30.000,00**
Servicios Básicos				
Luz	5	S/.50,00	12	S/.3.000,00
Internet	5	S/.30,00	12	S/.1.800,00
Sub total				**S/.4.800,00**
Total				**S/.45.050,00**

Tabla 06 : "Presupuesto referencial de la investigación"

Fuente: Elaboración propia

CAPÍTULO V: DESARROLLO DE LA SOLUCIÓN

5.1 Propuesta solución

5.1.1 Planteamiento y descripción de actividades

En el presente trabajo, mediante la aplicación de Machine Learning, se busca predecir el resultado de las cotizaciones (aprobación o rechazo) con el fin de agilizar la toma de decisiones con respecto al tiempo y costos de importación a fin de evitar la pérdida de ventas. De acuerdo con la metodología establecida en el gráfico 30, se establecieron los siguientes pasos:

1. **Recolección de datos**: Este primer paso hace referencia a definir los datos que se recopilarán para el análisis de estos.
2. **Preparación de datos**: Luego de la obtención de los datos, se realizará un análisis preliminar. Este paso es esencial ya que el objetivo es la limpieza de los datos y obtener la base datos final, es decir, se detectarán y eliminarán los datos ausentes, erróneos o atípicos que puedan afectar la fase de aprendizaje del modelo.
3. **Análisis y modelado**: Con los datos finales obtenidos del paso previo, se aplicarán las técnicas a utilizar.
4. **Evaluación y reporte**: En este último paso se analizará e interpretará el resultado obtenido para determinar la exactitud de cada modelo. Si el resultado obtenido por el *accuracy* puede ser mejorado, se tomarán las acciones necesarias como la codificación y normalización de la base de datos. Se van a comparar los resultados obtenidos por cada modelo y se seleccionará aquel que lleve a cabo la mejor predicción de la variable dependiente.

5.1.2 Desarrollo de actividades. Aplicación de herramientas de solución

1. **Recolección de datos**

   Base de datos

   La base de datos para el presente trabajo fue obtenida de la empresa en estudio. Se cuenta con la información de las cotizaciones desde el año 2020 hasta junio del 2022. Esta base de datos posee 576 registros separados equitativamente en cotizaciones aceptadas y rechazadas, asimismo, se encuentra en formato xlsx, es decir, en el programa Excel.

Como variables independientes (X), ordenadas en columnas dentro del Excel, se encuentran las siguientes etiquetas: Monto cotizado, peso en toneladas del montacargas cotizado, tiempo de entrega estimado, código del vendedor, código de la oficina, código del cliente, código del mercado o rubro del cliente, mes de la cotización, año de la cotización, y condición de pago (contado o crédito). Como variable dependiente (Y) se encuentra la siguiente etiqueta: Estado de la cotización. Por lo tanto, se definieron 11 variables (10 independientes y 1 dependiente) a ser utilizadas en el presente trabajo. A continuación se brinda mayor detalle:

Nro	Variable Independiente	Descripción
1	MontoPosicion	Precio del montacargas cotizado
2	Toneladas	Tonelaje del montacargas cotizado
3	TiempoEntrega	Tiempo estimado de llegada de la importación (en meses)
4	IDVendedor	Código interno del vendedor
5	CodOficina	Código de oficina (Se asocia a la sucursal)
6	CodCliente	Código de cliente
7	CodMercado	Código del mercado (Construcción, transporte, agricultura, etc)
8	MONTH	Mes de inicio de la cotización
9	YEAR	Año de inicio de la cotización
10	Pago	Condición de pago (al contado, pago a 15 días, 30 días, hasta 90 días)
Nro	**Variable Dependiente**	**Descripción**
1	EstadoY	Estado de la cotización (Aprobado o Rechazado)

Tabla 07: "Descripción de variables"

Fuente: Elaboración propia

Asimismo, se detalla una leyenda para las variables de Pago, Código de mercado, Código de cliente, código de oficina, y código o ID de los vendedores:

- **Pago:**

Condición de pago	Pago
Contado	0
Crédito 7 días	7
Crédito 15 días	15
Crédito 30 días	30
Crédito 45 días	45
Crédito 60 días	60
Crédito 90 días	90

Tabla 08: "Condición de pago"

Fuente: Elaboración propia

- **CodMercado:**

MERCADO	CodMercado
CONSTRUCCIÓN	1
TRANSPORTE	2
INDUSTRIA	3
COMERCIO Y SERVICIOS	4
AGRICULTURA	5
MARINO	6
MINERÍA	7
PROVEEDORES PARA EQUIPOS	8
CANTERAS Y AGREGADOS	9
EMPRESAS FINANCIERAS	10
HIDROCARBURO Y ENERGÍA	11
GOBIERNO	12
AFILIADAS	13
GRAN MINERÍA	14
FORESTAL	15
PARTICULAR	16

Tabla 09: "Mercados"

Fuente: Elaboración propia

- **CodCliente** (se brinda un pequeño ejemplo de la codificación asignada. Cabe resaltar que el nombre de los clientes se mantendrá de manera anónima)**:**

cod_cliente	Cliente
1028275	Cliente 1
1371541	Cliente 2
1160903	Cliente 3
1382939	Cliente 4
1387439	Cliente 5
1048210	Cliente 6
1053884	Cliente 7

Tabla 10: "Código de clientes"

Fuente: Elaboración propia

- **CodOficina:**

CodOficina	Descripción Oficina de venta
2301	Lima
2318	Lambayeque
2320	Trujillo
2321	Arequipa
2323	Huaraz
2324	Juliaca
2331	Piura
2332	Cajamarca
2333	Ayacucho
2348	Cusco
2349	Huancayo
2350	Cerro de Pasco
2352	Ilo

Tabla 11: "Descripción de oficina de ventas"

Fuente: Elaboración propia

- **IDVendedor**

NOMBRE DEL VENDEDOR	IDVendedor
SERGIO ENRIQUE ESPINOZA MEDINA	81753
DANIEL ANDRE HUAMANTINCO MARMA	81754
JIMMY OSCAR BURGOS GUZMAN	81755
MARCO ANTONIO CARDOSO ORTIZ	81756
OSCAR ALBERTO YAGA GUZMAN	81757
HENRY ARMANDO PACHAS AVALOS	81758
JORGE EUGENIO SAMPEN MURRUGARR	81759
WALTER JUAN MANUEL ARTEAGA VAS	81760
CARLOS ALONSO VALDIVIA TORRES	81761
OMAR PICON TARRILLO	81762
DAVID ALEJANDRO CAYO ROMERO	81763
CARLOS LUIS CANEZ ALDECOA	81764
KEVIN ANTONIO GONZALES RUIZ	81765
JULIAN PATRICIO GARCIA HUACCAY	81766
GROWER ORLANDO USURIAGA DE LA	81767
IVAN HIPAHUANCA TAPIA	81768
BENGGI ROBERT VIZCARRA GUILLEN	81769
MARCO TITO SILVA FARFAN	81770
REYNALDO SOTELO IZAGUIRRE LOPE	81771
MAYRA ALEJANDRA CHUMAN CRUZ	81772
CESAR MANUEL LINCH PEÑA	81773
ELAR MAURICIO VALDIVIA ESCALAN	81774
ALBERT ALEXANDER AGREDA LI	81775
EDWIN RAFAEL QUILICHE HUAMAN	81776
CHRISTIAN EDUARDO YEPEZ LAVILL	81777
JEAN CARLOS RODRIGUEZ LAURA	81778
JUAN PABLO BELTRAN TUPAC YUPAN	81779
JHONATAN KENYO RUIZ JUMPA	81780
CESAR GUSTAVO CHIAN CHANG	81781
GILBERTH ANDRES CRUZADO VILLAR	81782
ELMER WILFREDO ZACARIAS ROJAS	81783
JUAN CARLOS FRANCO VELASQUEZ	81784
DANIEL HUAMAN NECIOSUP	81785
EMILIANO EDGARDO FLORES CORDOV	81786
RAMON EDUARDO ARROYO VIGIL	81787
LAURA EUGENIA ROJAS VERGARA	81788
MARCO ANTONIO SEGURA PERLECHE	81789
MARIO MARCO MARTIN BERGHUSEN S	81790
JOSE CARLOS LLANOS CASTAÑEDA	81791
EMPERATRIZ ALEJANDRA YARE LEON	81792
JEISON GABRIEL PACHECO BELLIDO	81793
ROGER REYNERIO ESPINAL DIAZ	81794
CARLOS ENRIQUE ALCANTARA PARED	81795

Tabla 12: "ID del vendedor"

Fuente: Elaboración propia

2. Preparación de datos

a. Análisis preliminar

El primer aspecto a entender de los datos es que el precio o monto cotizado guarda relación con las toneladas del montacargas, es decir, el monto cotizado de un montacargas será mayor si el tonelaje de este es mayor. Esto nos permitirá abordar de mejor manera la limpieza que se realice con estas variables, pudiendo entender los valores atípicos, corrigiendo los valores no válidos, y completando los valores faltantes.

El tiempo de entrega estimado del montacargas cotizado no necesariamente aumentará al incrementarse el tonelaje o monto cotizado.

Cada vendedor tiene asignado un código de 5 dígitos. Asimismo cada cotización tiene asignado un solo vendedor, el cual puede repetirse. Lo mismo sucede con las variables de código de oficina (4 dígitos), código de cliente (7 dígitos) y código de mercado (1 dígito).

La variable de meses (Month) solo podrá tener números de 1 a 12, mientras que año (Year) solo tendrá los años de 2020, 2021 y 2022. Esto permitirá corregir o completar los datos en el apartado de limpieza.

En cuanto a la variable Pago (condición de pago), esta solo puede tomar valores numéricos como: 0, 7, 15, 30, 45, 60 y 90.

Con respecto a la variable dependiente o Y, esta solo podrá tener las categorías de: Aceptado o Rechazado, es decir, no deben tener valores numéricos ni otra categoría. Por tal motivo, se desestimaron las cotizaciones con categoría En Proceso dado que no forman parte del presente estudio.

Finalmente, en el caso de que se posean valores nulos, no válidos, faltantes o duplicados y no se pudiera esclarecer su origen, se desestimará ese registro a fin de no influenciar en el resultado.

b. Limpieza

En esta actividad se procedió a retirar los datos nulos o con campos vacíos como, por ejemplo, los datos donde no se había colocado nombre del vendedor ya que en el proceso de aprobación de cotizaciones estos datos no son considerados por no tener los campos completos y se debe ingresar nuevamente la cotización con los datos correctos, lo que significa que se tendrían registros duplicados.

Así mismo, para poder analizar las cotizaciones aprobadas o rechazadas, se retiraron las cotizaciones que se encontraban en proceso y los datos que no aportan información relevante para el modelo de machine learning.

A continuación se brinda un ejemplo encontrado en la base de datos con respecto a datos nulos o no válidos:

RRVV COMPL.	SUCURSAL RRVV	REGION RRVV
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A
#N/A	#N/A	#N/A

Tabla 13: "Datos nulos"

Fuente: Empresa

3. **Análisis y modelado**

En este paso se procedió a realizar la construcción de los modelos aplicando las técnicas k-NN, Regresión logística, SVM y Árbol de decisión, las cuales forman parte del tipo de aprendizaje supervisado. Para ello se utilizó la plataforma Anaconda, el interfaz de Jupyter Notebook y el lenguaje de programación Python.

4. **Evaluación y reporte**

Finalmente, se evaluarán los resultados obtenidos de los modelos a través de la métrica de *accuracy* para conocer el porcentaje de acierto de la predicción, y se elegirá la técnica que logre un mayor nivel de *accuracy*. Como se mencionó anteriormente, si el resultado obtenido por el *accuracy* puede ser mejorado, se tomarán las acciones necesarias como la codificación y normalización de la base de datos.

5.2 Medición de la solución

5.2.1 Análisis de indicadores cuantitativos y/o cualitativos

Para la medición de la solución del modelado propuesto se utilizó la métrica de accuracy que nos indicará la exactitud de la predicción. “Esta métrica indica el número total de predicciones correctas del modelado, es decir, la proporción de aciertos en la clasificación” (Lena & García, 2021,pág. 03).

Asimismo, como refuerzo para la métrica de accuracy, se empleó la matriz de confusión a fin de poder identificar el número de cotizaciones correctamente predichas (sean aprobadas o rechazadas) por cada modelo, y así poder visualizar el mapa de calor correspondiente.

Cabe resaltar que también se llevaron a cabo las acciones de codificación y normalización con el fin de que los modelos puedan analizar de mejor manera los datos, sin embargo, solo en el caso de que se logre mejorar el accuracy obtenido, se tomará en cuenta.

5.2.2 Simulación de solución

A continuación se mostrarán los pasos que se siguieron para llevar a cabo la simulación de la solución aplicando las 4 técnicas de aprendizaje supervisado:

1. **Ingreso a la interfaz de Jupyter**

Luego de obtenida la data final, se procedió a realizar la construcción del modelado bajo el tipo de aprendizaje supervisado. Para ello se usó el lenguaje de programación Python, usando la interfaz de Jupyter en la plataforma de Anaconda.
A continuación se brinda una imagen de la interfaz y de la plataforma:

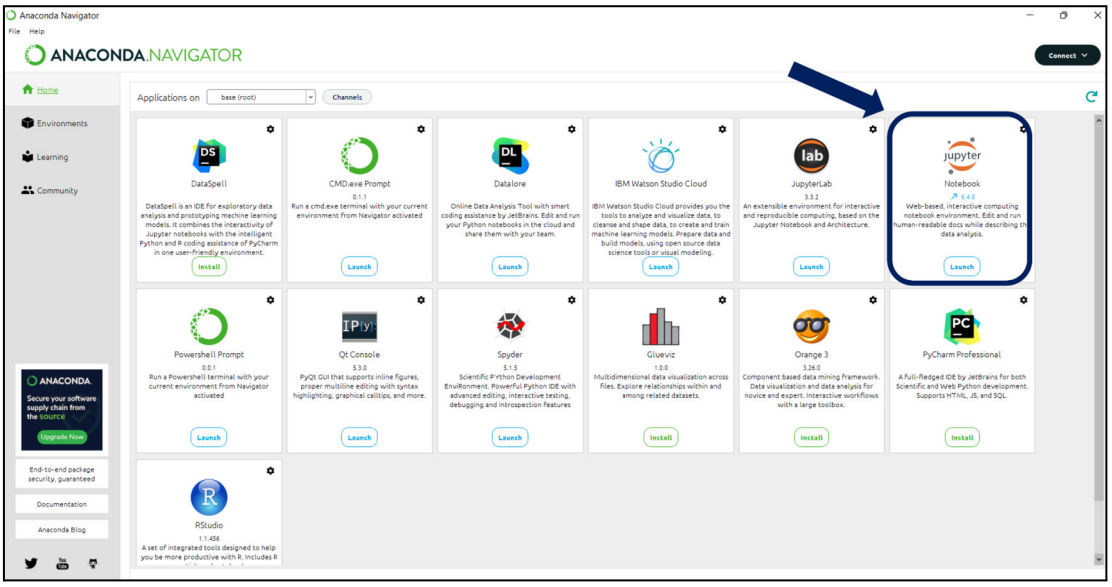


Gráfico 32: "Plataforma de Anaconda"

Fuente: Anaconda Navigator

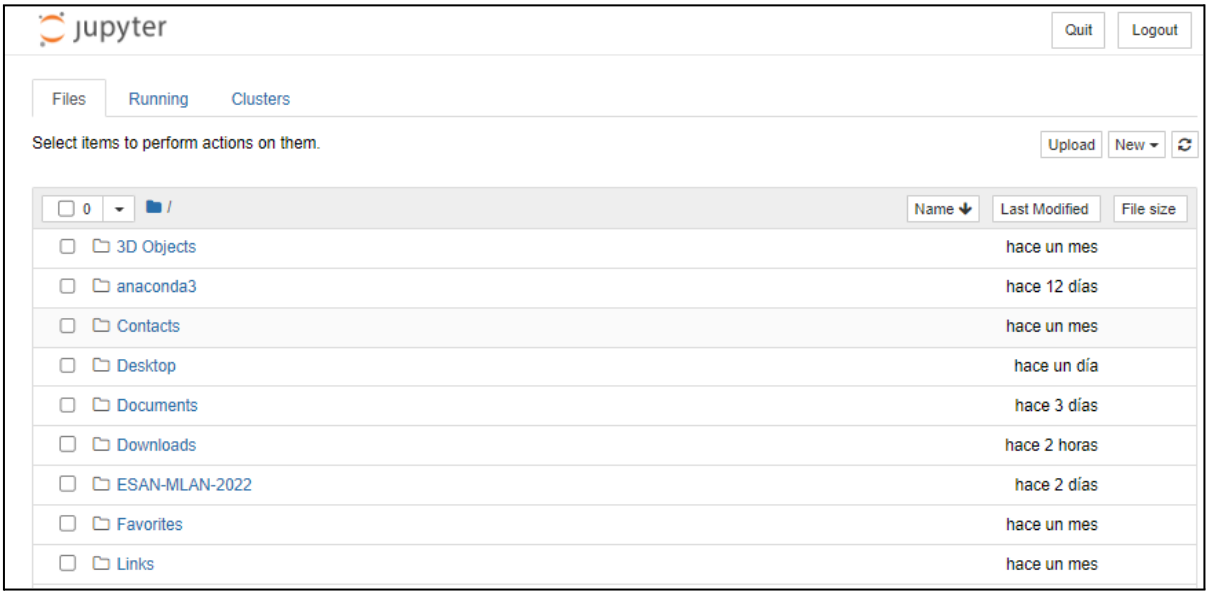


Gráfico 33: "Interfaz de Jupyter"

Fuente: Jupyter

2. Importación de los datos

Para importar los datos para cada uno de los 4 modelos, haremos uso de la librería *pandas*, tal como se observa a continuación:

```
1 import pandas as pd
```

Gráfico 34: "Empleo de la librería *pandas*"

Fuente: Elaboración propia

3. Lectura del archivo

Una vez importada la librería *pandas*, se procede a subir el archivo de Excel con los registros de las cotizaciones. Asimismo, se guardará el archivo con el nombre de "datos" (variable). Esto se muestra a continuación:

```
1  datos = pd.read_excel("COTIZACIONES2.xlsx")
```

Gráfico 35: "Lectura del archivo de Excel"

Fuente: Elaboración propia

A fin de corroborar que los datos del archivo se han adjuntado correctamente, se realiza el comando datos.head() para evidenciar los primeros registros:

```
1  datos.head()
```

	MontoPosicion	Toneladas	TiempoEntrega	IDVendedor	CodOficina	CodCliente	CodMercado	MONTH	YEAR	Pago	EstadoY
0	82000.0	5.0	89	81756	2301	1061190	8	1	2022	0	ACEPTADO
1	135000.0	7.0	69	81760	2301	1008321	9	1	2022	30	ACEPTADO
2	135000.0	7.0	69	81754	2349	1046927	7	1	2022	30	ACEPTADO
3	45000.0	3.0	46	81760	2301	1054270	3	1	2022	30	ACEPTADO
4	45000.0	3.0	46	81759	2324	1350252	1	1	2022	0	ACEPTADO

Gráfico 36: *"datos.head()* para evidenciar los primeros registros"

Fuente: Elaboración propia

4. Pasos a seguir para la programación

Paso 1: Separación de variables en X - Y

Con el archivo listo para ser trabajado, se llevará a cabo la separación de las variables en X e Y. Recordar que el archivo está guardado con el nombre (variable) de "datos". Esto se muestra a continuación:

```
1  X = datos[["MontoPosicion","Toneladas","TiempoEntrega","IDVendedor","CodOficina","CodCliente","CodMercado","MONTH","YEAR","Pago"]]
2  Y = datos[["EstadoY"]]
```

Gráfico 37: "Separación de variables X - Y"

Fuente: Elaboración propia

Paso 2: Separación de datos en train y test

Habiendo identificado correctamente las variables en X e Y, se separarán los datos en train y test, para esto, como primer paso, se debe importar la librería *sklearn*, luego, dentro de esta librería, se utilizará la técnica de selección de modelos, es decir, *sklearn.model_selection* y, finalmente, se empleará la

herramienta *train_test_split* para separar los datos en 4 grupos de manera aleatoria: X_train, X_test, Y_train, y Y_test. Cabe resaltar que, del conjunto de registros, el 80% será tomado para entrenamiento y, el 20% restante, para la evaluación (predicción), tal como se muestra a continuación:

```
1  import sklearn

1  from sklearn.model_selection import train_test_split
```

Gráfico 38: "Uso de la librería *sklearn* y de la técnica *train_test_split*"

Fuente: Elaboración propia

```
1  X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.20, random_state=10)

1  X_train.shape, X_test.shape

((460, 10), (116, 10))
```

Gráfico 39: "Separación de datos en train y test (80% y 20%)"

Fuente: Elaboración propia

El 80% del total de datos (576) es 460.8, debido a que se trata de registros, este valor fue tomado como 460. De la misma manera, el 20% restante abarca 116 registros de cotizaciones. Para el presente caso, se desestimará el parámetro *random_state* dado que es solo es relevante para la reproducibilidad de experimentos.

Paso 3: Elección del modelo

Con los datos ya separados, se elige el modelo a usar, para ello se hace uso de la librería *sklearn.* Para la técnica k-NN se usó la herramienta *KNeighborsClassifier*, para la técnica Regresión logística se usó la herramienta *LogisticRegression*, para la técnica SVM se usó la herramienta *svm* y para la técnica árbol de decisión se usó la herramienta *DecisionTreeClassifier*. Cada herramienta se guardó con el nombre "ml" como variable.

<u>Técnica k-NN:</u>

Se iniciará con el parámetro de k = 5, tal como se observa:

```
from sklearn.neighbors import KNeighborsClassifier

ml = KNeighborsClassifier(n_neighbors=5) # K=5
```

Gráfico 40: "Elección del modelo y del parámetro k"

Fuente: Elaboración propia

Técnica Regresión logística:

```
from sklearn.linear_model import LogisticRegression

ml = LogisticRegression(random_state=10)
```

Gráfico 41: "Elección del modelo de regresión logística"

Fuente: Elaboración propia

Técnica SVM:

- **Kernel Linear**

```
from sklearn import svm

ml = svm.SVC(kernel='linear')
```

Gráfico 42: "Elección del modelo SVM con kernel lineal"

Fuente: Elaboración propia

- **Kernel Poly**

```
from sklearn import svm

ml = svm.SVC(kernel='poly')
```

Gráfico 43: "Elección del modelo SVM con kernel Poly"

Fuente: Elaboración propia

- **Kernel rbf**

```
from sklearn import svm

ml = svm.SVC(kernel='rbf')
```

Gráfico 44: "Elección del modelo SVM con kernel rbf"

Fuente: Elaboración propia

Técnica Árbol de decisión:

```
from sklearn.tree import DecisionTreeClassifier

ml = DecisionTreeClassifier()
```

Gráfico 45: "Elección del modelo de árbol de decisión"

Fuente: Elaboración propia

Paso 3: Entrenamiento del modelo (X_train, Y_train, *fit*)

En este paso llevaremos a cabo el entrenamiento, es decir, el modelo analizará el 80% de los datos de las variables independientes (X_train), así como sus respuestas (categoría de la cotización o Y_train). Para cada técnica haremos uso de la herramienta *fit*.

Técnica k-NN:

```
ml.fit(X_train,y_train)
```

```
KNeighborsClassifier()
```

Gráfico 46: "Entrenamiento del modelo k-NN"

Fuente: Elaboración propia

Técnica Regresión logística:

```
ml.fit(X_train, y_train)
```

```
LogisticRegression(random_state=10)
```

Gráfico 47: "Entrenamiento del modelo de regresión logística"

Fuente: Elaboración propia

Técnica SVM:

- **Kernel Lineal**

```
ml.fit(X_train, y_train)
```

```
SVC(kernel='linear')
```

Gráfico 48: "Entrenamiento del modelo SVM con kernel lineal"

Fuente: Elaboración propia

- **Kernel Poly**

```
ml.fit(X_train, y_train)
```

```
SVC(kernel='poly')
```

Gráfico 49: "Entrenamiento del modelo SVM con kernel Poly"

Fuente: Elaboración propia

- **Kernel rbf**

```
ml.fit(X_train, y_train)
```

```
SVC()
```

Gráfico 50: "Entrenamiento del modelo SVM con kernel rbf"

Fuente: Elaboración propia

Técnica Árbol de decisión:

```
ml= ml.fit(X_train,y_train)
```

Gráfico 51: "Entrenamiento del modelo de Árbol de decisión"

Fuente: Elaboración propia

Paso 4: Predicciones llevadas a cabo por el modelo (X_test, *res*)

Una vez entrenado el modelo, llevaremos a cabo las predicciones, para esto haremos uso de los datos X_test y de la herramienta *predict*. Asimismo, asignaremos la variable "res" a las respuestas o predicciones del modelo.

Técnica k-NN:

```
res = ml.predict(X_test)
```

Gráfico 52: "Predicciones del modelo (X_test, res) k-NN"

Fuente: Elaboración propia

Técnica Regresión logística:

```
res = ml.predict(X_test)
```

Gráfico 53: "Predicciones del modelo (X_test, res) de regresión logística"

Fuente: Elaboración propia

Técnica SVM:

- **Kernel linear, Poly y rbf**

```
res = ml.predict(X_test)
```

Gráfico 54: "Predicciones del modelo (X_test, res) SVM con kernel linear, kernel poly y kernel rbf"

Fuente: Elaboración propia

Técnica Árbol de decisión:

```
res = ml.predict(X_test)
```

Gráfico 55: "Predicciones del modelo (X_test, res) de árbol de decisión"

Fuente: Elaboración propia

Paso 5: Evaluación de las predicciones (Y_test, *accuracy*)

Para la evaluación de las predicciones llevadas a cabo por el modelo se utilizará el conjunto de datos Y_test y la métrica accuracy para corroborar el nivel de precisión del modelo, asimismo, se hará uso de la matriz de confusión y su correspondiente mapa de calor.

Técnica k-NN:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.6551724137931034
```

Gráfico 56: "Accuracy de la técnica k-NN"

Fuente: Elaboración propia

Como se observa, el resultado obtenido es del 0.655, es decir, el modelo ha acertado en el 66% de las predicciones. Adicionalmente, se comparan los resultados del modelo al variar el parámetro de k. Para esto se hace uso de las herramientas *matplotlib* y *pyplot*.

```
from matplotlib import pyplot
```

```
pyplot.plot(range(1,50,2),lista_resultados)
```

```
[<matplotlib.lines.Line2D at 0x24fb96180a0>]
```

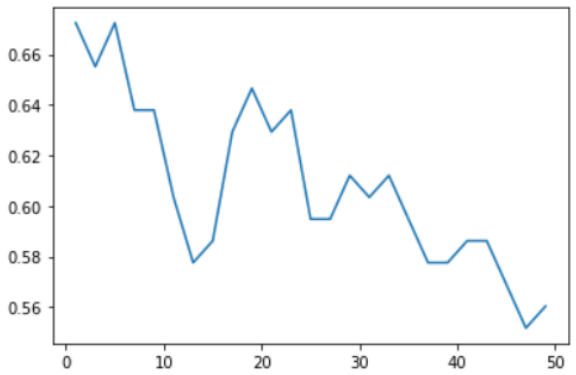


Gráfico 57: "Variación de accuracy frente a cambios en el parámetro k"

Fuente: Elaboración propia

```
lista_resultados = []

for k in range (1, 50, 2):
    ml = KNeighborsClassifier(n_neighbors=k)
    ml.fit(X_train, y_train.values.ravel())
    res=ml.predict(X_test)
    print('Para k = ',k, ' el accuracy es: ',accuracy_score(y_test, res))
```

```
Para k =  1  el accuracy es:  0.6724137931034483
Para k =  3  el accuracy es:  0.6551724137931034
Para k =  5  el accuracy es:  0.6724137931034483
Para k =  7  el accuracy es:  0.6379310344827587
Para k =  9  el accuracy es:  0.6379310344827587
Para k =  11  el accuracy es:  0.603448275862069
Para k =  13  el accuracy es:  0.5775862068965517
Para k =  15  el accuracy es:  0.5862068965517241
Para k =  17  el accuracy es:  0.6293103448275862
Para k =  19  el accuracy es:  0.646551724137931
Para k =  21  el accuracy es:  0.6293103448275862
Para k =  23  el accuracy es:  0.6379310344827587
Para k =  25  el accuracy es:  0.5948275862068966
Para k =  27  el accuracy es:  0.5948275862068966
Para k =  29  el accuracy es:  0.6120689655172413
Para k =  31  el accuracy es:  0.603448275862069
Para k =  33  el accuracy es:  0.6120689655172413
Para k =  35  el accuracy es:  0.5948275862068966
Para k =  37  el accuracy es:  0.5775862068965517
Para k =  39  el accuracy es:  0.5775862068965517
Para k =  41  el accuracy es:  0.5862068965517241
Para k =  43  el accuracy es:  0.5862068965517241
Para k =  45  el accuracy es:  0.5689655172413793
Para k =  47  el accuracy es:  0.5517241379310345
Para k =  49  el accuracy es:  0.5603448275862069
```

Gráfico 58: "Lista de resultados de accuracy frente a cambios en el parámetro k"

Fuente: Elaboración propia

El máximo valor de accuracy que se puede obtener con el modelo al variar el parámetro de k es de 0.6724 con k tomando el valor de 5, es decir, un 67% de aciertos en la predicción aproximadamente. De esta manera, ajustando el valor de k, el accuracy también puede ser corroborado en la matriz de confusión.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.6724137931034483

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[42, 16],
       [22, 36]], dtype=int64)
```

Gráfico 59: "Matriz de confusión k-NN"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 42 y 36
    - Valor 42: Son las predicciones correctas realizadas por el modelo, verdaderos positivos.
    - Valor 36: Son las predicciones correctas realizadas por el modelo, verdaderos negativos.
- Diagonal de predicciones incorrectas: Valores 16 y 22
    - Valor 16: Son las predicciones incorrectas realizadas por el modelo, falsos negativos.
    - Valor 22: Son las predicciones incorrectas realizadas por el modelo, falsos positivos.

Otra manera de representar gráficamente la matriz de confusión es a través de un mapa de calor, el cual es mostrado a continuación para la técnica k-NN:

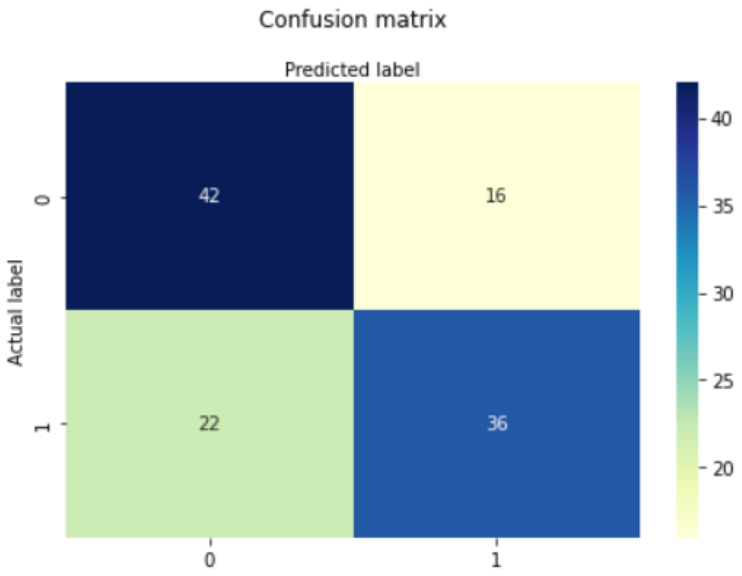


Gráfico 60: "Mapa de calor k-NN"

Fuente: Elaboración propia

De manera similar que en la matriz de confusión, en el mapa de calor se tiene la diagonal de predicciones correctas, es decir, aquellos cuadrantes donde la categoría que se ha predicho ha acertado con la categoría real que posee el dato o valor, en este caso, esta diagonal de predicciones correctas es representada por los valores 42 y 36, de la misma forma, la diagonal de predicciones incorrectas, es decir, aquellas categorías que fueron predichas de manera errónea dado que no acertaron con la categoría real que posee el dato o valor, es representada por los valores 22 y 16.

En el mapa de calor, el rango de colores va desde más claros (valores más bajos, crema) hacia colores más oscuros (valores más altos, azul oscuro/marino). En este caso, el valor más alto es 42, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 36, obteniendo el color azul. Estos dos valores son los más altos de los 4 de la matriz de confusión, además de ser los cuadrantes o diagonales con las predicciones correctas, representando así un resultado favorable del modelo en cuanto a las predicciones, habiendo predicho 78 (42 + 36) datos de manera correcta de un total de 116. Por consiguiente, los 2 valores más bajos son los de la diagonal de predicciones incorrectas, siendo estos 22 (color verde) y 16 (color crema).

Técnica Regresión logística:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.7068965517241379
```

Gráfico 61: "Accuracy de técnica Regresión Logística"

Fuente: Elaboración propia

Como se puede apreciar, el Accuracy Score obtenido es de 0.7, es decir, el modelo ha acertado el 70% de las predicciones. Adicionalmente, se elaborará la matriz de confusión y el gráfico de calor para visualizar de diferente manera el accuracy obtenido anteriormente.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.7068965517241379

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[44, 14],
       [20, 38]], dtype=int64)
```

Gráfico 62: "Matriz de confusión Regresión Logística"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 44 y 38
    - Valor 44: Son las predicciones correctas realizadas por el modelo, verdaderos positivos.
    - Valor 38: Son las predicciones correctas realizadas por el modelo, verdaderos negativos.
- Diagonal de predicciones incorrectas: Valores 14 y 20
    - Valor 14: Son las predicciones incorrectas realizadas por el modelo, falsos negativos.
    - Valor 20: Son las predicciones incorrectas realizadas por el modelo, falsos positivos.

A continuación se muestra el mapa de calor para la matriz de confusión de la técnica regresión logística:

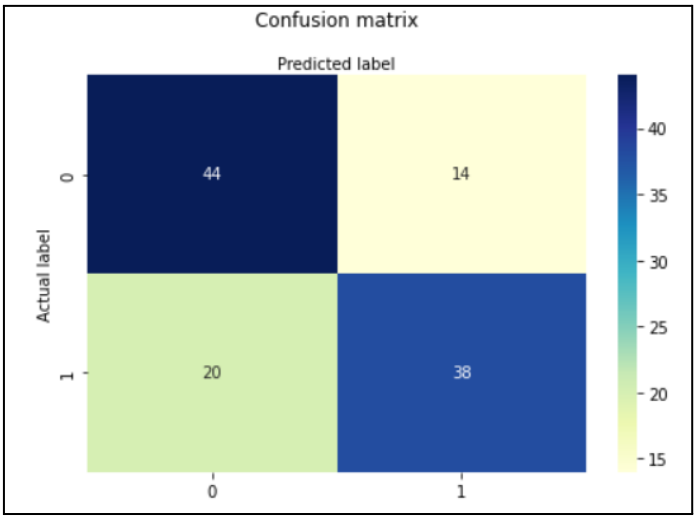


Gráfico 63: "Mapa de calor Regresión Logística"

Fuente: Elaboración propia

En este caso, el valor más alto es 44, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 38, obteniendo el color azul. Estos dos valores son los más altos de los 4 de la matriz de confusión, además de ser los cuadrantes o diagonales con las predicciones correctas, representando así un resultado favorable del modelo en cuanto a las predicciones, habiendo predicho 82 (44+38) datos de manera correcta de un total de 116. Por consiguiente, los 2 valores más bajos son los de la diagonal de predicciones incorrectas, siendo estos 20 (color verde) y 14 (color crema).

Técnica SVM - Linear

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.6379310344827587
```

Gráfico 64: "Accuracy de técnica SVM - Linear"

Fuente: Elaboración propia

Como se puede apreciar, el Accuracy Score obtenido es de 0.64, es decir, el modelo ha acertado el 64% de las predicciones. Adicionalmente, se elaborará la matriz de confusión y el gráfico de calor para visualizar de diferente manera el accuracy obtenido anteriormente.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.6379310344827587

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[24, 34],
       [ 8, 50]], dtype=int64)
```

Gráfico 65: "Matriz de confusión SVM - Linear"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 24 y 50
    - Valor 24: Son las predicciones correctas realizadas por el modelo, verdaderos positivos.
    - Valor 50: Son las predicciones correctas realizadas por el modelo, verdaderos negativos.
- Diagonal de predicciones incorrectas: Valores 8 y 34
    - Valor 8: Son las predicciones incorrectas realizadas por el modelo, falsos negativos.
    - Valor 34: Son las predicciones incorrectas realizadas por el modelo, falsos positivos.

A continuación se muestra el mapa de calor para la matriz de confusión de la técnica SVM - Linear:

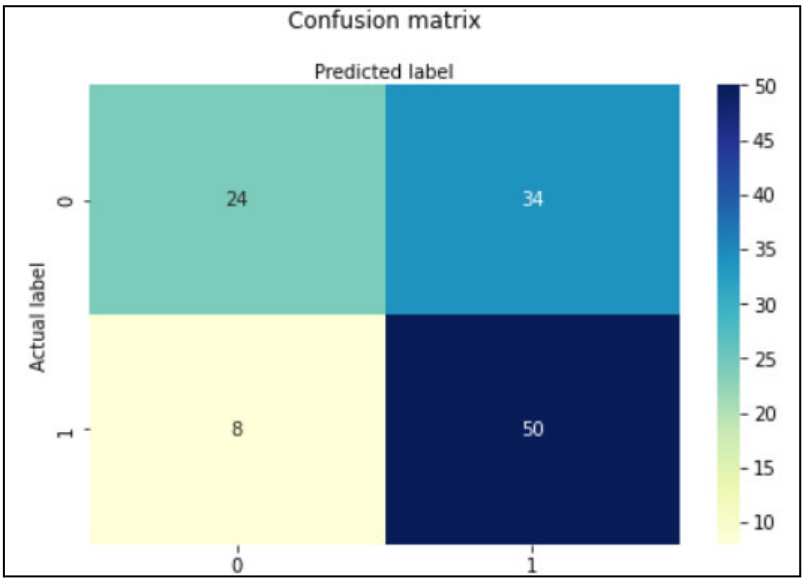


Gráfico 66: "Mapa de calor SVM - Linear"

Fuente: Elaboración propia

Como se observa, el valor más alto es 50, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 34, obteniendo el color azul. En este caso, los dos valores más altos corresponden tanto a un cuadrante de la diagonal de predicciones correctas como a un cuadrante de predicciones incorrectas. El siguiente valor es 24, obteniendo un color verde y, finalmente, el valor de 8 con un color crema. De esta manera, el presente modelo ha predicho correctamente solo 74 (50+24) valores de un total de 116.

Técnica SVM - Poly

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.5775862068965517
```

Gráfico 67: "Accuracy de técnica SVM - Poly"

Fuente: Elaboración propia

Como se puede apreciar, el Accuracy Score obtenido es de 0.57, es decir, el modelo ha acertado el 57% de las predicciones. Adicionalmente, se elaborará la matriz de confusión y el gráfico de calor para visualizar de diferente manera el accuracy obtenido anteriormente.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.5775862068965517

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[47, 11],
       [38, 20]], dtype=int64)
```

Gráfico 68: "Matriz de confusión SVM - Poly"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 47 y 20
    - Valor 47: Son las predicciones correctas realizadas por el modelo, verdaderos positivos.
    - Valor 20: Son las predicciones correctas realizadas por el modelo, verdaderos negativos.
- Diagonal de predicciones incorrectas: Valores 11 y 38
    - Valor 11: Son las predicciones incorrectas realizadas por el modelo, falsos negativos.
    - Valor 38: Son las predicciones incorrectas realizadas por el modelo, falsos positivos.

A continuación se muestra el mapa de calor para la matriz de confusión de la técnica SVM - Poly:

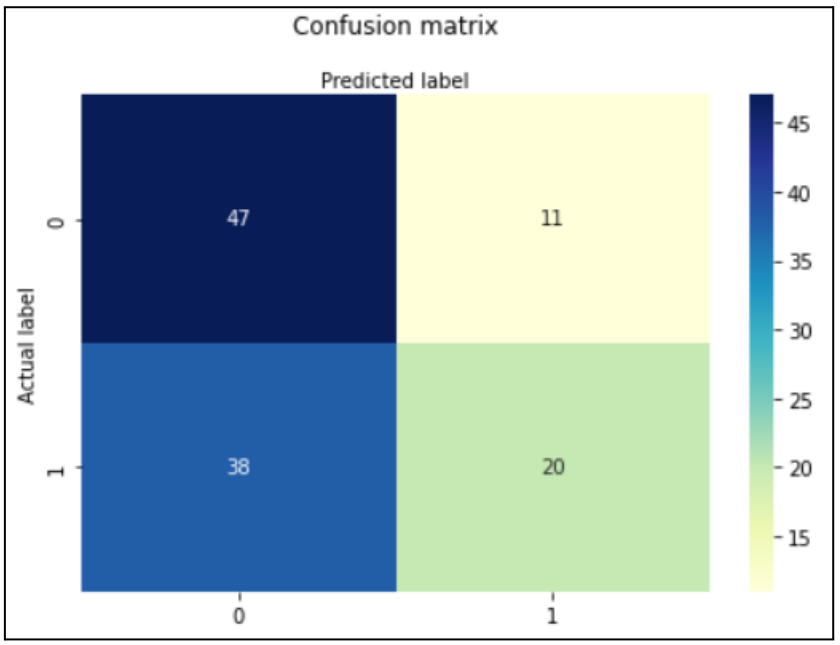


Gráfico 69: "Mapa de calor SVM - Poly"

Fuente: Elaboración propia

Como se observa, el valor más alto es 47, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 38, obteniendo el color azul. En este caso, los dos valores más altos corresponden tanto a un cuadrante de la diagonal de predicciones correctas como a un cuadrante de predicciones incorrectas. El siguiente valor es 20, obteniendo el color verde y, finalmente, el valor de 11 con un color crema. De esta manera, el presente modelo ha predicho correctamente solo 67 (47+20) valores de un total de 116.

Técnica SVM - rbf

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.603448275862069
```

Gráfico 70: "Accuracy de técnica SVM - rbf"

Fuente: Elaboración propia

Como se puede apreciar, el Accuracy Score obtenido es de 0.60, es decir, el modelo ha acertado el 60% de las predicciones. Adicionalmente, se elaborará

la matriz de confusión y el gráfico de calor para visualizar de diferente manera el accuracy obtenido anteriormente.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.603448275862069

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[53,  5],
       [41, 17]], dtype=int64)
```

Gráfico 71: "Matriz de confusión - rbf"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 53 y 17
    - Valor 53: Son las predicciones correctas realizadas por el modelo, verdaderos positivos
    - Valor 17: Son las predicciones correctas realizadas por el modelo, verdaderos negativos
- Diagonal de predicciones incorrectas: Valores 5 y 41
    - Valor 5: Son las predicciones incorrectas realizadas por el modelo, falsos negativos
    - Valor 41: Son las predicciones incorrectas realizadas por el modelo, falsos positivos

A continuación se muestra el mapa de calor para la matriz de confusión de la técnica SVM - rbf:

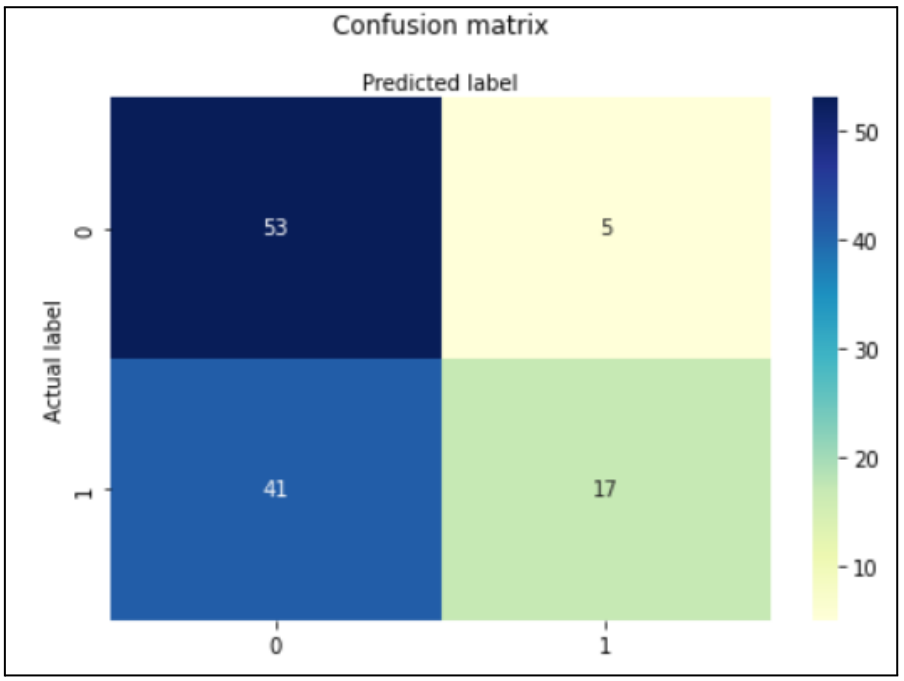


Gráfico 72: "Mapa de calor SVM - rbf "

Fuente: Elaboración propia

Como se observa, el valor más alto es 53, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 41, obteniendo el color azul. En este caso, los dos valores más altos corresponden tanto a un cuadrante de la diagonal de predicciones correctas como a un cuadrante de predicciones incorrectas. El siguiente valor es 17, obteniendo el color verde y, finalmente, el valor de 5 con un color crema. De esta manera, el presente modelo ha predicho correctamente solo 70 (53+17) valores de un total de 116.

Técnica árbol de decisiones

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.8793103448275862
```

Gráfico 73: "Accuracy de técnica Árbol de decisión"

Fuente: Elaboración propia

Como se puede apreciar, el Accuracy Score obtenido es de 0.83, es decir, el modelo ha acertado el 83% de las predicciones. Adicionalmente, se elaborará

la matriz de confusión y el gráfico de calor para visualizar de diferente manera el accuracy obtenido anteriormente.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))

Accuracy: 0.8793103448275862

ml_matrix = metrics.confusion_matrix(y_test, res)

ml_matrix

array([[49,  9],
       [ 5, 53]], dtype=int64)
```

Gráfico 74: "Matriz de confusión Árbol de decisión"

Fuente: Elaboración propia

Con la matriz de confusión se puede evidenciar que se obtienen 4 valores, de esta manera:

- Diagonal de predicciones correctas: Valores 49 y 53
    - Valor 49: Son las predicciones correctas realizadas por el modelo, verdaderos positivos
    - Valor 53: Son las predicciones correctas realizadas por el modelo, verdaderos negativos
- Diagonal de predicciones incorrectas: Valores 9 y 5
    - Valor 9: Son las predicciones incorrectas realizadas por el modelo, falsos negativos
    - Valor 5: Son las predicciones incorrectas realizadas por el modelo, falsos positivos

A continuación se muestra el mapa de calor para la matriz de confusión de la técnica Árbol de decisiones:

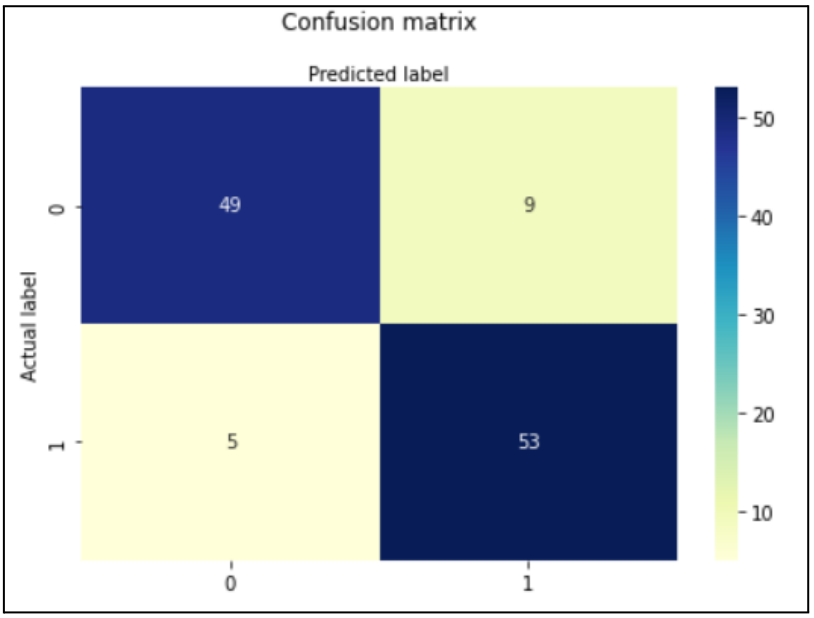


Gráfico 75: "Mapa de calor Árbol de decisión"

Fuente: Elaboración propia

En este caso, el valor más alto es 53, por lo cual ha obtenido un color azul oscuro o marino. El siguiente valor en orden descendente es 49, obteniendo el color azul. Estos dos valores son los más altos de los 4 de la matriz de confusión, además de ser los cuadrantes o diagonales con las predicciones correctas, representando así un resultado favorable del modelo en cuanto a las predicciones, habiendo predicho 102 (49+53) datos de manera correcta de un total de 116. Por consiguiente, los 2 valores más bajos son los de la diagonal de predicciones incorrectas, siendo estos 9 (color verde) y 5 (color crema).

Paso 6: Mejoras de la base de datos

Como medidas de mejora al modelo, se llevarán a cabo las acciones de codificación y normalización de la base de datos.

Una vez obtenidos los porcentajes de accuracy de las cuatro técnicas, se realizará la codificación y normalización. Cabe resaltar que, este paso se repetirá para las cuatro técnicas dado que el código/programación es idéntico para estas. Posterior a ello, se evaluará si el nuevo accuracy score aumentó, disminuyó o se mantuvo con la mejora en la base de datos.

Codificación de las columnas IDVendedor, CodOficina, CodCliente

Para llevar a cabo la codificación de las columnas IDVendedor, CodOficina y CodCliente haremos uso del conjunto de herramientas *sklearn.preprocessing*, específicamente de las técnicas *StandardScaler* y *LabelEncoder*, como se muestra a continuación:

```
1 from sklearn.preprocessing import StandardScaler,LabelEncoder
```

Gráfico 76: "Importando *sklearn.preprocessing*"

Fuente: Elaboración propia

Una vez importadas las técnicas, llevaremos a cabo la codificación de las 3 variables, como se muestra:

```
from sklearn.preprocessing import StandardScaler,LabelEncoder

codificadorIDV = LabelEncoder()
X['IDVendedor'] = codificadorIDV.fit_transform(X['IDVendedor'])

codificadorCOF = LabelEncoder()
X['CodOficina'] = codificadorCOF.fit_transform(X['CodOficina'])

codificadorCOC = LabelEncoder()
X['CodCliente'] = codificadorCOC.fit_transform(X['CodCliente'])
```

Gráfico 77: "Codificación de variables"

Fuente: Elaboración propia

A continuación se muestra la nueva base de datos previo a la normalización. Como se observa, las columnas de IDVendedor, CodOficina y CodCliente han cambiado sus valores con el fin de mejorar la predicción del modelo:

```
1 X
```

	MontoPosicion	Toneladas	TiempoEntrega	IDVendedor	CodOficina	CodCliente	CodMercado	MONTH	YEAR	Pago
0	82000.0	5.0	89	3	0	160	8	1	2022	0
1	135000.0	7.0	69	7	0	14	9	1	2022	30
2	135000.0	7.0	69	1	9	99	7	1	2022	30
3	45000.0	3.0	46	7	0	141	3	1	2022	30
4	45000.0	3.0	46	6	5	296	1	1	2022	0
...	...	...	...	...	...	...	...	...	...	...
571	31000.0	2.5	46	1	0	228	4	12	2020	0
572	178000.0	10.0	59	7	0	206	4	12	2020	0
573	190000.0	10.0	59	7	0	221	3	12	2020	0
574	70000.0	5.0	89	7	0	127	9	12	2020	30
575	39000.0	3.0	46	7	0	127	9	12	2020	30

Gráfico 78: "Base de datos codificada"

Fuente: Elaboración propia

Normalización de la base de datos

Una vez codificadas las columnas de IDVendedor, CodOficina y CodCliente, se identificó que los datos eran dispersos, por lo tanto, se procedió a normalizar la base de datos con el objetivo de mejorar la predicción del modelo. Para ello se hará uso del normalizador, es decir, la técnica *StandardScaler*, la cual consiste en restar a cada dato el valor de la media y dividirlo entre la desviación estándar, como se muestra a continuación:

```
normalizador = StandardScaler()
X = normalizador.fit_transform(X)
```

```
X
```

```
array([[ 0.5151514 ,  0.45701941,  2.05929074, ..., -1.82012571,
         1.76512125, -0.77919269],
       [ 1.76247496,  1.36008348,  0.86206493, ..., -1.82012571,
         1.76512125,  0.70941424],
       [ 1.76247496,  1.36008348,  0.86206493, ..., -1.82012571,
         1.76512125,  0.70941424],
       ...,
       [ 3.05686733,  2.71467959,  0.26345203, ...,  1.11430597,
        -0.94610499, -0.77919269],
       [ 0.23273852,  0.45701941,  2.05929074, ...,  1.11430597,
        -0.94610499,  0.70941424],
       [-0.49682809, -0.44604467, -0.51474474, ...,  1.11430597,
        -0.94610499,  0.70941424]])
```

Gráfico 79: "Normalizando los datos"

Fuente: Elaboración propia

<u>Técnica k-NN</u>

Nuevos valores del parámetro k con las mejoras aplicadas

Ahora que ya se cuenta con la base de datos normalizada, se vuelve a ejecutar el modelo con un "k" igual a 5, obteniéndose un *accuracy* mayor, es decir, se logró incrementar el *accuracy* de 67% a 86%, como se observa a continuación:

```
from sklearn.metrics import accuracy_score
```

```
accuracy_score(y_test, res)
```

```
0.8620689655172413
```

Gráfico 80: "Nuevo accuracy de la técnica k-NN"

Fuente: Elaboración propia

A modo de brindar mayor detalle, se brinda la lista de los nuevos posibles *accuracy* para los valores de "k" del 1 al 49:

```
lista_resultados = []

for k in range (1, 50, 2):
    ml = KNeighborsClassifier(n_neighbors=k)
    ml.fit(X_train, y_train.values.ravel())
    res=ml.predict(X_test)
    print('Para k = ',k, ' el accuracy es: ',accuracy_score(y_test, res))
```

```
Para k =  1  el accuracy es:  0.8620689655172413
Para k =  3  el accuracy es:  0.8620689655172413
Para k =  5  el accuracy es:  0.8620689655172413
Para k =  7  el accuracy es:  0.853448275862069
Para k =  9  el accuracy es:  0.8448275862068966
Para k =  11  el accuracy es:  0.8448275862068966
Para k =  13  el accuracy es:  0.8275862068965517
Para k =  15  el accuracy es:  0.8189655172413793
Para k =  17  el accuracy es:  0.8017241379310345
Para k =  19  el accuracy es:  0.8189655172413793
Para k =  21  el accuracy es:  0.8189655172413793
Para k =  23  el accuracy es:  0.8103448275862069
Para k =  25  el accuracy es:  0.8103448275862069
Para k =  27  el accuracy es:  0.8017241379310345
Para k =  29  el accuracy es:  0.8017241379310345
Para k =  31  el accuracy es:  0.7931034482758621
Para k =  33  el accuracy es:  0.8017241379310345
Para k =  35  el accuracy es:  0.7931034482758621
Para k =  37  el accuracy es:  0.7931034482758621
Para k =  39  el accuracy es:  0.7758620689655172
Para k =  41  el accuracy es:  0.7844827586206896
Para k =  43  el accuracy es:  0.7844827586206896
Para k =  45  el accuracy es:  0.7758620689655172
Para k =  47  el accuracy es:  0.7758620689655172
Para k =  49  el accuracy es:  0.7758620689655172
```

Gráfico 81: "Nuevos valores de k"

Fuente: Elaboración propia

A continuación se muestra la tabla anterior en un gráfico:

```
from matplotlib import pyplot

pyplot.plot(range(1,50,2),lista_resultados)
```

```
[<matplotlib.lines.Line2D at 0x24fbeca91f0>]
```

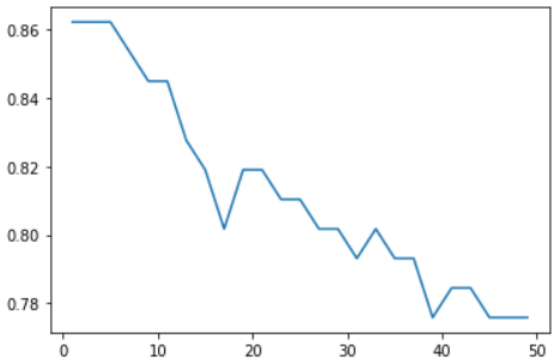


Gráfico 82: "Gráfica de nuevos valores de k"

Fuente: Elaboración propia

Como se puede observar, para un “k” igual o menor a 5, se obtiene el mayor accuracy score de 86% (0.86). Por lo tanto, con la mejora planteada se logró incrementar en casi un 20% el nivel de precisión del modelo.

Técnica Regresión Logística

Para la técnica de Regresión Logística se evidenció que la codificación y normalización no tuvo efecto alguno en el accuracy de esta técnica, manteniéndose en 70.6%, como se muestra en la siguiente imagen:

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, res))
Accuracy: 0.7068965517241379
```

Gráfico 83: “Nuevo accuracy de la técnica Regresión Logística”

Fuente: Elaboración propia

Técnica SVM - linear

Para la técnica de SVM con kernel Linear se evidenció que la codificación y normalización no tuvo efecto alguno en el accuracy de esta técnica, manteniéndose el mismo en un 63.79%, como se muestra en la siguiente imagen:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)
0.6379310344827587
```

Gráfico 84: “Nuevo accuracy de la técnica SVM - linear”

Fuente: Elaboración propia

Técnica SVM - poly

Para la técnica de SVM con kernel Poly se evidenció que la codificación y normalización no tuvo efecto alguno en el accuracy de esta técnica, manteniéndose el mismo en un 57.76%, como se muestra en la siguiente imagen:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.5775862068965517
```

Gráfico 85: "Nuevo accuracy de la técnica SVM - poly"

Fuente: Elaboración propia

Técnica SVM - rbf

Para la técnica de SVM con kernel rbf se evidenció que la codificación y normalización no tuvo efecto alguno en el accuracy de esta técnica, manteniéndose el mismo en un 60.34%, como se muestra en la siguiente imagen:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.6034482758620690
```

Gráfico 86: "Nuevo accuracy de la técnica SVM - rbf"

Fuente: Elaboración propia

Técnica Árbol de decisión

Para la técnica de Árbol de decisión se evidenció que la codificación y normalización sí tuvo un efecto positivo en el accuracy de esta técnica, incrementándose de un 87.93% a un 88.79%, como se muestra en la siguiente imagen:

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test, res)

0.8879310344827587
```

Gráfico 87: "Nuevo accuracy de la técnica Árbol de decisión"

Fuente: Elaboración propia

5: Comparación final de las técnicas empleadas:

Una vez obtenidos los resultados de predicción de las 4 técnicas empleadas, aplicando la codificación y normalización correspondientes, se obtuvo que la mejor técnica de predicción para nuestra base de datos de las cotizaciones es Árbol de Decisión (Decision Tree) con un *accuracy* de 88.79% aproximadamente, seguida por la técnica k-NN con un acierto en las predicciones del 86.21%, continuando con la técnica de Regresión logística, la cual obtuvo un acierto de 70.69%, finalmente, la técnica que tuvo un menor porcentaje de aciertos en las predicciones fue la técnica SVM (Support Vector Machine), aplicando 3 kernel distintos (linear, poly y rbf), siendo sus *accuracy* 63.79%, 57.76% y 60.34%, respectivamente. Estos resultados se observan en la siguiente tabla:

Descripción	k-NN (k = 5)	Regresión logística	SVM			Árbol de decisión
			Linear	Poly	RBF	
Sin normalización ni codificación	67.94%	70.69%	63.79%	57.76%	60.34%	**87.93%**
Con normalización y codificación	86.21%	70.69%	63.79%	57.76%	60.34%	**88.79%**

Tabla 14: Comparación de resultados de las técnicas empleadas

Fuente: Elaboración propia

CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones

En el presente trabajo se ha logrado el cumplimiento de la construcción de 4 modelos predictivos bajo el tipo de aprendizaje supervisado mediante las siguientes técnicas: k-NN, Regresión logística, SVM y Árbol de decisión. Todos los modelos tuvieron como objetivo predecir el estado (aprobación o rechazo) de las cotizaciones de la empresa, a fin de agilizar la toma de decisiones con respecto a los montacargas y así evitar la pérdida de ventas por demora de productos.

De las cuatro técnicas de machine learning aplicadas, el modelo con mayor porcentaje de acierto en las predicciones es el Árbol de decisión. Esta técnica se encuentra apta para su uso por la empresa dado que se ha logrado alcanzar una confiabilidad del 88.79% en la predicción de los estados de las cotizaciones (accuracy del 88.79%). Esto se debió principalmente a la aplicación de mejoras como codificación y normalización de la base de datos (variables X) dado que los valores/datos de cada categoría se encontraban muy dispersos. Los resultados de las demás técnicas fueron: k-NN con un acierto en las predicciones del 86.21%, Regresión logística con un acierto de 70.69%, y la técnica SVM (Support Vector Machine), aplicando 3 kernel distintos (linear, poly y rbf), obteniendo los aciertos en las predicciones de 63.79%, 57.76% y 60.34%, respectivamente.

La metodología de 4 pasos empleada para el presente trabajo de investigación ha permitido la correcta aplicación de 4 técnicas de modelo predictivo. Cabe resaltar que el último paso de la metodología incluye la aplicación de posibles mejoras como la codificación y normalización, evaluando si este cambio genera efectos positivos en la predicción de los modelos.

Con respecto a la metodología mencionada previamente, como primer paso se procedió a recopilar la data de cotizaciones pasadas entre los años 2020 y 2022 recopilando más de 500 cotizaciones para, posteriormente, proceder a realizar la limpieza de los mismos retirando registros que poseían valores nulos o datos que no aportaban a la aplicación del modelo planteado, definiendo de esta forma las variables X y la variable Y del modelo. Posteriormente se procedió a la aplicación de las técnicas tomando el 80 % de datos para entrenamiento y reservando el 20 % para la validación.

En cuanto al análisis de datos, limpieza de datos y aplicación de los pasos estudiados del enfoque clasificación de machine learning, se obtuvo inicialmente diferentes porcentajes de acierto para los cuatro modelos, teniendo como conclusión que es confiable, pero no totalmente seguro debido a que hay un margen de error considerado debido a que involucra miles de dólares.

Se procedió a realizar la codificación y normalización de datos para las cuatro técnicas debido a que los datos estaban muy dispersos, es decir, se tenía valores desde siete cifras hasta valores de una sola cifra; para ello se restó la media a cada valor y se dividió entre su desviación estándar, obteniendo datos más comprensibles para el algoritmo. Posterior a ello, se ejecutó nuevamente el código en las cuatro técnicas y la que obtuvo un mayor incremento en el porcentaje de confiabilidad fue la técnica de k-NN ya que, con la aplicación de la mejora en la base de datos su *accuracy* aumentó de 68% a 86% obteniendo un incremento del 18 %. Sin embargo, la técnica de machine learning con un mayor *accuracy* fue el Árbol de decisiones puesto que obtuvo un 89% de confiabilidad, a pesar de solo obtener un incremento del 1% luego de la mejora en la base de datos. Es por ello que al realizar la comparativa de las técnicas, se eligió Árbol de decisión (Decision Tree) como la técnica más óptima por su alto nivel de confiabilidad sobre las demás.

6.2 Recomendaciones

Dado que aún existe un porcentaje de error del 11% (nivel de confiabilidad del 89%) en la técnica con un mejor *accuracy*, se recomienda que las decisiones se tomen en conjunto con un experto en el tema así como también considerando factores externos a las variables X utilizadas durante la programación del algoritmo que puedan generar una desviación en la tendencia de ventas o en los cambios de decisiones de los clientes.

En caso la empresa tome en cuenta las predicciones llevadas a cabo por el modelo y ocurra el hecho de que se decida importar un montacargas pero finalmente la cotización sea rechazada, la maquinaria importada quedaría en stock a la espera de otro pedido, pudiéndose comercializar a un precio más competitivo debido a que el precio de compra fue menor por el momento en el que fue adquirido.

Se recomienda ampliar el alcance de la aplicación de machine learning en la empresa dado que existen oportunidades adicionales de predicción, como, por ejemplo, la estimación del tiempo de entrega de los montacargas y la estimación del volumen de ventas de la línea de montacargas de la empresa. Esto permitiría fortalecer la ventaja

competitiva de la empresa en el mercado al permitirle conocer información con anticipación y, de esta manera, tomar las decisiones adecuadas.

Asimismo, se recomienda ampliar el alcance en cuanto a técnicas de Machine Learning a utilizar como, por ejemplo, Redes neuronales, permitiendo una comparación adicional en los resultados (accuracy o precisión en las predicciones), beneficiando a la empresa y a los encargados de la toma de decisiones con respecto a la importación de los montacargas.

Finalmente, es recomendable continuar recopilando datos importantes que tengan relevancia en la decisión de compra de montacargas por parte de los clientes. Esto permitirá poder enriquecer la base de datos y analizar mayor cantidad de información relevante de la mano de las técnicas o modelos de predicción de Machine Learning, buscando alcanzar un mejor porcentaje de confiabilidad. De esta forma, se mejoraría en gran medida la toma de decisión anticipada generando una ventaja competitiva en la empresa, debido a que los tiempos de entrega serían menores y habría un incremento en los márgenes de ganancia al momento de adquirir una maquinaría antes que el proveedor (la fábrica) incremente el precio cada 5 o 6 meses en promedio.

REFERENCIAS BIBLIOGRÁFICAS

- Aguilar, C. (2021). Interpretable machine learning for promotional sales. (Tesis para optar el título profesional). Universidad Rey Juan Carlos
- Baldoceda Ramírez,A., Mamani Ccallohuari H. (2020). Modelo de aprendizaje supervisado para pronóstico de la deserción de estudiantes de la Facultad de Ingeniería y Arquitectura de la Universidad Peruana Unión - Lima (Tesis para obtener el Título Profesional de Ingeniero de Sistemas). Universidad Peruana Unión.
- Barrueta, R y Castillo, E. (2018). Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos. (Tesis para optar el título profesional de Ingeniero de Sistemas de Información). Universidad Peruana de Ciencias Aplicadas.
- Caja, O. (2020), en Valencia, realizó la investigación titulada: *Librería Python para el aprendizaje y la implementación de redes neuronales*. Recuperado de https://riunet.upv.es/bitstream/handle/10251/152226/Caja%20-%20Librer%C3%ADa%20Python%20para%20el%20aprendizaje%20y%20la%20implementación%20de%20redes%20neuronales.pdf?sequence=1
- Carmona, A.d.l. (2022). Diseño de soluciones avanzadas basadas en técnicas de machine learning para la toma de decisiones en gestión de activos. (Tesis Doctoral Inédita). Universidad de Sevilla, Sevilla.
- Céspedes, A. (2017), en Santiago, Chile, realizó la investigación titulada: *Construcción De Modelo De Forecast Para Estimación De Demanda En Una Empresa Multinacional De Retail*, Memoria De Titulación Para Optar Al Título De Ingeniero Civil Informático. Recuperado de https://repositorio.usm.cl/bitstream/handle/11673/41250/3560902038636UTFSM.pdf?sequence=1&isAllowed=y
- ComexPerú (abril, 2022). ¿Cómo fue el desempeño de las Importaciones en el Primer Trimestre de 2022?. *Semanario 1117 - Comercio Exterior*. Recuperado de https://www.comexperu.org.pe/articulo/como-fue-el-desempeno-de-las-importaciones-en-el-primer-trimestre-de-2022#:~:text=De%20acuerdo%20con%20cifras%20de,fue%20de%20US%24%2011%2C527%20millones.
- Eiras-Franco, C. (2019). *New scalable machine learning methods: Beyond classification and regression* (Tesis de Ph.D). University of A Coruña.

- Espeso, J., Fernández, A., Espeso, M. & Espeso, B. (2007). *Seguridad en el trabajo: Manual para la Formación del Especialista.* España: Lex Nova S.A.
- Harrington, P. (2012). *Machine Learning in Action*. Nueva York: Manning Publications Co.
- INEI (marzo, 2022). Producción Nacional. *Informe Técnico*. Recuperado de https://www.inei.gob.pe/media/principales_indicadores/03-informe-tecnico-produccion-nacional-ene-2022.pdf
- Kelleher, J., et. al (2015). *Fundamentals of Machine Learning for Predictive data analytics: Algorithms, worked examples, and case studies*. Massachusetts: The Massachusetts Institute of Technology Press.
- Lena, G. & García, M (2021). *Avances en Educación, TIC e innovación: Aportaciones para la mejora empresarial y social.* Madrid
- Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Recuperado de https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf.
- Mueller, J., & Massaron, L. (2021). *Machine Learning for dummies*. (2da. ed.) Nueva Jersey: John Wiley & Sons, Inc.
- OCDE (2017). *Directrices de la OCDE aplicables en materia de precios de transferencia a empresas multinacionales y administraciones tributarias.* (2017). OCDE.
- Ramos, L. (marzo, 2022). Producción nacional crece 2,86%, pero manufactura y construcción caen en enero. *Infomercado*. Recuperado de https://infomercado.pe/produccion-nacional-crecio-286-pero-manufactura-y-construccion-caen-en-enero-160322-lr/
- Rouhiainen, L. (2018). *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro*. Barcelona: Editorial Planeta.
- Ruiz, J. & Solís. D (2018). *Complementos de aprendizaje automático. Ampliación de Inteligencia artificial, 2018 - 2019*. Recuperado de: https://www.cs.us.es/cursos/aia-2018/temas/tema_Complementos_Aprendizaje_Autom%C3%A1tico.pdf
- Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. (3ra. ed.) New Jersey: Pearson Education.

- Theobald, O. (2017). *Machine Learning for Absolute Beginners*. (2da. ed.) London: Scatterplot Press.