DESARROLLO DE MODELO PARA PREDICCIÓN DE VENTAS B2B EN EMPRESA DEL SECTOR AGROINDUSTRIAL

Presentado por:

Antonio Giacometto Cheij
Antonio Jose Fajardo Macías
Wilmer Castaño Mejía

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Daniel Enrique González Gómez
Director

Diego Luis Linares
Jurado

Mario Julian Mora
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

Camilo Rocha

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 06 de Julio de 2023

Facultad de Ingeniería y Ciencias

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, Julio 06 de 2023

Autores: Antonio Giacometto Cheij
Antonio Jose Fajardo Macías
Wilmer Castaño Mejía

Título del Trabajo de Grado: "DESARROLLO DE MODELO PARA PREDICCIÓN DE VENTAS B2B EN EMPRESA DEL SECTOR AGROINDUSTRIAL"

Director: Daniel Enrique González Gómez

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

Firma del Director del Trabajo de Grado

Santiago de Cali, 29 de Mayo de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad optar por el título de Magíster en Ingeniería, nos permitimos presentar a su consideración y solicitar la sustentación del Trabajo de Grado denominado Desarrollo de modelo para predicción de ventas B2B en empresa del sector agroindustrial, realizado por los estudiantes: Antonio Giacometto Cheij con C.C. 1140874520 y código 8972718, Antonio Fajardo Macías con C.C 94449334 y código 8971760 y Wilmer Castaño Mejía con C.C 1130617439 y código 0051135; pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección del profesor Daniel Enrique Gonzalez Gomez.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Antonio Giacometto.

Antonio Giacometto Cheij
C.C. 1140874520 de Barranquilla

Antonio Fajardo Macías
C.C. 94449334 de Cali

Wilmer Castaño Mejía
C.C. 1130617439 de Cali - Valle

Daniel Enrique Gonzalez Gomez
C.C. 16669372 de Cali -Valle

Barranquilla, 5 de Julio de 2023

Ing. Yoseff Ghisays
Gerente de Línea Estratégica Procesamiento y Acondicionamiento de Granos

Señores Pontificia Universidad Javeriana Cali,

Por medio de la presente autorizo el uso de los datos provenientes de las bases de datos del CRM utilizado por el Departamento Comercial de SuperBrix para el desarrollo del proyecto de grado de los estudiantes Antonio Fajardo, Wilmer Castaño y Antonio Giacometto cuyo objetivo es desarrollar una herramienta para la predicción del total de ventas e ingresos (Forecasting) basados en el histórico de ofertas de SuperBrix S.A.

Se solicita que la información sensible, tal como información de contacto de clientes y representantes de ventas de SuperBrix, sea anonimizada para evitar filtraciones de dicha información.

Atentamente,

Yoseff Ghisays

Yoseff Ghisays
Gerente de Línea Estratégica Procesamiento y Acondicionamiento de Granos

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TÍTULO: Desarrollo de modelo para predicción de ventas B2B en empresa del sector agroindustrial.

1. **ÁREA DE TRABAJO:** Sector Productivo - Agroindustrial
2. **TIPO DE PROYECTO (Aplicado, Innovación, Investigación):** Aplicado
3. **ESTUDIANTE(S):** Antonio Giacometto, Antonio Fajardo, Wilmer Castaño
4. **CORREO ELECTRÓNICO:** antonio.giacometto@javerianacali.edu.co, ajfajardom@javerianacali.edu.co, wcastano@javerianacali.edu.co
5. **DIRECCIÓN Y TELÉFONO:** Cra 57A No 13-106 (Cali) - 3016612340
6. **DIRECTOR:** Daniel Enrique González Gómez
7. **VINCULACIÓN DEL DIRECTOR:** Profesor Facultad Ingeniería y Ciencias - Javeriana Cali
8. **CORREO ELECTRÓNICO DEL DIRECTOR:** dgonzalez@javerianacali.edu.co
9. **GRUPO O EMPRESA QUE LO AVALA:** SuperBrix S.A.
10. **PALABRAS CLAVE:** Ventas, Predicción, Clientes, Forecasting, Tasa de conversión.
11. **ODS QUE APLICA EL PROYECTO (Agenda 2030)**: 8. Trabajo decente y crecimiento económico / 9. Industria, innovación e infraestructura / 12. Producción y consumo responsable / 13. Acción por el Clima
12. **FECHA DE INICIO:** Julio 04 de 2022
13. **RESUMEN:** *El trabajo presentado corresponde al desarrollo de un modelo de aprendizaje automático para predecir la probabilidad de conversión en venta de las cotizaciones recibidas en una empresa del sector agroindustrial colombiano. Los diferentes modelos probados, fueron entrenados utilizando un dataset consolidado con los datos históricos de ventas de la organización y algunas fuentes externas; La preparación de este dataset involucró diferentes etapas de limpieza, mejoramiento de datos y pre procesamiento, las cuales permitieron además de alimentar los modelos de predicción probados, desarrollar un modelo de agrupamiento que permitió identificar perfiles de clientes de acuerdo a algunas de sus principales características basados en la información de cotizaciones. Como complemento fue desarrollada una herramienta de visualización para el monitoreo y control de indicadores claves de desempeño dentro del área comercial.*

DESARROLLO DE MODELO PARA PREDICCIÓN DE VENTAS B2B EN EMPRESA DEL SECTOR AGROINDUSTRIAL

Antonio Giacometto Cheij
Antonio Jose Fajardo Macías
Wilmer Castaño Mejía

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
Daniel Enrique González Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO 29 DE 2023

TABLA DE CONTENIDO

Contenido

Pág.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El presente trabajo se desarrolla en una empresa colombiana dedicada al diseño, fabricación y comercialización de soluciones y bienes de capital en el sector agroindustrial, tiene dentro de su portafolio los siguientes productos y servicios: 1. Maquinaria para el acondicionamiento y transformación de granos y legumbres, 2. Servicios especializados de ingeniería y 3. Ejecución de proyectos llave en mano para la agroindustria. Debido a la naturaleza del negocio, la interacción de la empresa y sus clientes es categorizada dentro del tipo de ventas conocido como B2B[1]. El área comercial de la organización se encuentra interesada en utilizar la información recopilada durante los últimos 5 años desde la implementación del sistema de CRM[2] para generar valor agregado que permita mejorar la toma de decisiones estratégicas en el área.

En el presente documento, se detalla el desarrollo de un modelo de aprendizaje automático, basado en la información histórica recopilada en el CRM y otras variables externas identificadas a lo largo del proyecto. Como resultado se dota a la empresa con una herramienta útil en la predicción de la probabilidad de conversión en ventas de las solicitudes de cotización recibidas, lo cual apoya la toma de decisiones basadas en datos con respecto a la estrategia comercial y de mercadeo, así como de permitir a la organización estimar el volumen de ventas esperada en un periodo y organizar su producción en base a estos supuestos.

[1] B2B: Negocio a Negocio, del Inglés Business to Business.

[2] CRM: Gestión de la relación con el cliente, del Inglés Customer Relationship Management.

1. DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del problema

La compañía ha identificado un problema recurrente que consiste en la falta de enfoque de los recursos tanto del área de ventas como de producción hacia las oportunidades de proyectos con mayor nivel de certeza de convertirse en ventas reales y que permitan un mayor margen de contribución para la organización. Actualmente, la compañía carece de un método eficiente para estimar la probabilidad de conversión en ventas desde una etapa preliminar. Como resultado, el análisis de rentabilidad se lleva a cabo después de la ejecución del proyecto o de la fabricación. La ausencia de un proceso de predicción adecuado impide proyectar las ventas a lo largo de un período determinado, lo cual dificulta la posibilidad de adelantar la producción con el objetivo de ofrecer tiempos de entrega más competitivos, mejorar la eficiencia del abastecimiento, obtener ahorros en logística y afinar las estrategias de marketing y ventas, anticipándose así a las tendencias del mercado.

Conscientes sobre la necesidad urgente de contar con herramientas que permitan monitorear el embudo de ventas o "Pipeline"[3], logrando así tener una imagen en tiempo real del estado de las oportunidades y negocios en cada uno de los mercados; para lograrlo se ha identificado que es posible construir herramientas basadas en ciencias de datos para atender las siguientes necesidades:

[3] Pipeline: Conjunto de acciones realizadas para convertir una oportunidad de venta

- Mejorar el monitoreo de indicadores de desempeño en el área comercial.
- Establecer una metodología para predecir o estimar la probabilidad de obtener ventas futuras y la posibilidad de cierre de negocios.

- Definir y categorizar perfiles de clientes objetivo con base en la información histórica.

Actualmente la compañía ha tenido una alta solicitud de proyectos; pero debido a que los recursos de personal son limitados, no es posible atender con la misma prioridad cada una de las solicitudes. Para mejorar los márgenes de utilidad, a la vez que se optimiza el talento humano con que cuenta la compañía, se hace necesario identificar con antelación y agilidad las estrategias comerciales y el orden de atención a los potenciales proyectos con una alta confiabilidad; para apoyar este objetivo se construyó un modelo predictivo alimentado por información histórica de los clientes y de cotizaciones, que mediante herramientas de machine learning[4] permite determinar la probabilidad de conversión en venta de cada cotización ingresada al sistema.

1.2. Formulación del problema

- ¿Cuáles son las variables que inciden significativamente en la tasa de conversión[5] de los potenciales proyectos a realizarse?
- ¿Cuál es el perfil y las características de los clientes que presentan el mayor potencial de retorno y probabilidad de éxito para la empresa?
- ¿Qué variables deben optimizarse para garantizar el logro de la rentabilidad esperada para cada proyecto?
- ¿Qué información de los clientes es relevante de ser monitoreada para optimizar la toma de decisiones y planteamiento de la estrategia comercial?

---

[4] Machine Learning: Aprendizaje automático

[5] Tasa de Conversión:El número de cotizaciones que se convierten en ventas reales en un periodo determinado, Tasa de conversión = (Número de ventas / Número de cotizaciones) * 100

2. OBJETIVOS DEL PROYECTO

2.1. Objetivo general

Desarrollar una herramienta para la predicción de la conversión en ventas de las oportunidades potenciales mediante el empleo de técnicas de ciencias de datos, utilizando las bases de datos del CRM e históricos de ventas para una empresa del sector agroindustrial.

2.2. Objetivos específicos

- Desarrollar un modelo de aprendizaje automático adecuado a la empresa, que permita determinar la probabilidad de conversión en ventas de las oportunidades (potenciales negocios) que se presentan.
- Identificar los perfiles de clientes de mayor potencial de retorno y tasa de conversión para la organización.
- Diseñar una herramienta de visualización que facilite el monitoreo en tiempo real de indicadores claves de desempeño (KPI[6]), dentro del área comercial.

[6] KPI: Del inglés Key Performance Indicator.

3. MARCO TEÓRICO

3.1. Ventas B2B

Las ventas B2B (Business to Business) son intercambios comerciales que tienen lugar entre dos empresas [1]. Estas ventas se distinguen de las orientadas al consumidor (Business to Consumer) en varios aspectos, como un promedio de venta más alto, la participación de diferentes o múltiples tomadores de decisiones y una duración del ciclo de venta más prolongada debido a la complejidad, monto de la inversión y otras particularidades.

Es común identificar las siguientes etapas dentro de una venta B2B: Presentación, Demostración, Propuesta, Evaluación, Negociación, Cierre. Estas etapas, que pueden variar según la metodología de cada organización, se suelen agrupar y supervisar mediante una herramienta conocida como embudo de ventas o "Pipeline" [9],.

3.1.1. Pronóstico de conversión en ventas

Un pronóstico de conversión ventas es un elemento esencial en cualquier plan de negocios. Esta herramienta muy útil consiste en la predicción de ventas futuras basado en datos históricos, tendencias en la industria y en las características de las oportunidades presentes y su estado dentro del pipeline de ventas. La forma en la cual se realiza esta predicción y la información utilizada varía en cada organización dependiendo de su tamaño, antigüedad, industria, entre otros.

Entre los métodos o fuentes tradicionales utilizados en el pronóstico de ventas se encuentran: Método Histórico, Método de Evaluación de Oportunidades y Análisis Multivariado [2].

Método Histórico: Utiliza información de pasadas ventas de una organización para identificar tendencias y realizar predicciones con respecto a futuras ventas. Se puede dar mayor complejidad a este método al incluir datos del mercado para identificar tendencias dentro de él y buscar motivos de estos comportamientos. De acuerdo con este método cuantitativo, se espera que aquellos patrones identificados sean repetitivos a lo largo del tiempo, lo cual puede ser una debilidad en sí misma al asumir que se mantendrán las mismas condiciones de un tiempo pasado.

Evaluación de Oportunidades: La evaluación de oportunidades implica la recopilación de opiniones o perspectivas de los vendedores o líderes de cada negocio en cómo será el comportamiento de ventas en las próximas semanas, meses, o cualquier periodo en estudio. La fortaleza de este método se basa en las variables cualitativas ponderadas por cada participante con respecto a su mercado, clientes y de la dinámica dentro del pipeline de ventas. Sin embargo, no tiene una base cuantitativa y puede incluir cierto sesgo o "wishful thinking" que afecte la precisión y confiabilidad del método.

Análisis Multivariado: El método más sofisticado y completo para predicción de ventas podría ser el de análisis multivariado, donde se pueden combinar diferentes fuentes de información cuantitativa y cualitativa para alcanzar resultados más cercanos a la realidad. La empresa "Collective[i]", especializada en forecasting y servicios tecnológicos para empresas, define como "Pronóstico Prescriptivo" aquellas técnicas que gracias a inteligencia artificial y ciencia de datos utiliza información histórica de ventas, prospectos en el pipeline de la compañía, información de los mercados en tiempo real y demás información de diferentes fuentes privadas y públicas para entregar predicciones más precisas y confiables que permiten la toma de decisiones dentro de una organización [3].

3.1.2. KPIs de ventas

Los indicadores claves de desempeño son métricas que permiten administrar los procesos de una organización, estos son definidos por cada empresa de modo que resuman información útil que permita el control, identificación de oportunidades para alcanzar los objetivos y la toma de decisiones [4]. en cada etapa del embudo. Los KPIs pueden categorizarse en dos grupos; Indicadores rezagados (Lagging) los cuales miden los resultados, e indicadores proactivos

(Leading) que ayudan a predecir condiciones futuras [4].

Lagging KPIs: muestran los resultados generales de ventas. Por ejemplo, el volumen de ventas, volumen de ingresos o ticket promedio.

Leading KPIs: muestran los resultados de acciones específicas que permiten llegar a los resultados finales. Por ejemplo, el volumen de oportunidades en el embudo, la tasa de conversión promedio (prospecto/cliente), duración del ciclo de ventas, tasa de conversión MQL[7] a SQL[8]. Actividades agendadas, actividades completadas, actividades exitosas [10].

Uno de los indicadores centrales para el modelo a desarrollar es la tasa de conversión; este relaciona la cantidad de oportunidades generadas con las que en realidad se convierten en ventas.

3.1.3. Pipeline de ventas

El pipeline de ventas es una herramienta de gestión que permite organizar de forma visual y dinámica el flujo de ventas de una empresa. Esta herramienta refleja el movimiento de cada cliente potencial, desde su primer contacto con la marca hasta la etapa de postventa. Cada empresa tiene un proceso de ventas distinto, por lo cual el pipeline debe ser personalizado para cada realidad.

3.2. Preprocesamiento de datos

3.2.1. Técnicas de balanceo de clases

El balanceo de clases es una técnica importante para abordar los desequilibrios de clase en el preprocesamiento de los datos; Estas técnicas son un enfoque importante en el preprocesamiento. En el proyecto, los registros de las cotizaciones recibidas pueden presentar desequilibrios en la representatividad de las clases, lo que podría afectar la precisión de los modelos de predicción de ventas.

[7] MQL: Oportunidad de venta calificada por mercadeo, del inglés Marketing Qualified Lead.

[8] SQL: Oportunidad de venta calificada por ventas, del inglés Sales Qualified Lead.

Técnicas de sobre-muestreo: El uso de estas técnicas implica duplicar los registros de la clase minoritaria o generar muestras sintéticas de esta clase para aumentar su representatividad en el conjunto de datos. Esta técnica puede ayudar a mejorar la precisión de los modelos, pero también puede aumentar el riesgo de sobreajuste y una disminución de la velocidad de entrenamiento.

Técnicas de sub-muestreo: El uso de estas técnicas implica reducir el número de registros de la clase mayoritaria o seleccionar aleatoriamente registros de la clase minoritaria para obtener un equilibrio en la representatividad de las clases. Esta técnica puede ayudar a reducir el tiempo de entrenamiento y prevenir el sobreajuste, pero también puede disminuir la precisión del modelo.

Técnicas de reequilibrio de peso: Las técnicas de reequilibrio de peso implican asignar pesos diferentes a las clases durante el entrenamiento del modelo, para reflejar el desequilibrio de clase en el conjunto de datos. Esta técnica puede ayudar a mejorar la precisión de los modelos y prevenir el sobreajuste, pero requiere una comprensión cuidadosa de cómo se deben asignar los pesos para maximizar la efectividad.

3.2.1.1. Técnicas de imputación

Es importante identificar los valores faltantes en el conjunto de datos para entender su distribución y su impacto en el análisis. Se pueden utilizar herramientas de análisis de datos para identificar automáticamente los valores faltantes o visualizarlos gráficamente. Es importante comprender la razón detrás de los valores faltantes en los datos. Pueden ser causados por una variedad de factores, incluyendo errores de medición, registro o incluso la no disponibilidad de información. Conocer la causa de los valores faltantes es importante para elegir la técnica de imputación adecuada pues depende de la naturaleza de los datos y el objetivo del análisis. Por ejemplo, si la causa de los valores faltantes es un error de medición, puede ser apropiado utilizar la media o la mediana para imputarlos. Si la causa es una no disponibilidad de información, puede ser necesario utilizar una técnica más avanzada como la imputación por vecinos cercanos.

Imputación por promedio: en este método, se calcula el promedio de los valores presentes en una columna o variable y se reemplaza cualquier valor faltante con este promedio. Este método es sencillo y fácil de implementar, pero puede distorsionar los resultados si los valores faltantes están distribuidos de manera diferente a los valores presentes. Es útil cuando los datos faltantes son aleatorios y no tienen un patrón específico.

Imputación por mediana: en este método, se calcula la mediana de los valores presentes en una columna o variable y se reemplaza cualquier valor faltante con esta mediana. Este método es menos sensible a los valores extremos que el promedio y es útil para datos con distribuciones sesgadas o valores atípicos (outliers[9]).

Imputación por moda: Este método es útil cuando los datos son categóricos o discretos. Consiste en reemplazar los valores faltantes con la categoría o valor más frecuente en los datos no faltantes.

Imputación basada en modelos: en este método, se utiliza un modelo de aprendizaje automático para estimar los valores faltantes. Por ejemplo, se puede entrenar un modelo de regresión lineal para predecir una variable basada en otras variables presentes en el conjunto de datos. Este método es más preciso que los métodos simples, pero puede requerir más tiempo y recursos para entrenar y evaluar el modelo.

Imputación basada en la distribución: en este método, se asume una distribución para los datos y se generan valores aleatorios que siguen esa distribución para reemplazar los valores faltantes. Por ejemplo, se puede asumir una distribución normal y generar valores aleatorios siguiendo una

[9] Outlier: Estadísticamente se trata de una observación que es numéricamente distante del resto de observaciones

distribución normal con la misma media y desviación estándar que los valores presentes en el conjunto de datos. Este método es fácil de implementar y es útil cuando no se conocen las relaciones entre las variables.

Imputación basada en KNN (K-Nearest Neighbors / K Vecinos más cercanos): en este método, se utiliza los K registros más cercanos a un registro con valores faltantes para estimar el valor faltante. Por ejemplo, se puede calcular la distancia Euclidiana entre un registro con valores faltantes y los K registros más cercanos y utilizar los valores de esos K registros para estimar el valor faltante. Este método es útil para datos con alta dimensionalidad y es menos sensible a los valores extremos que los métodos basados en el promedio o la mediana.

3.3. Tipos de modelos

Los modelos de Machine Learning son herramientas que utilizan algoritmos para aprender de los datos y hacer predicciones o tomar decisiones sin ser explícitamente programados para hacerlo. Estos modelos son una forma de inteligencia artificial que permite encontrar patrones en los datos y utilizarlos para realizar tareas específicas. Se pueden dividir en tres categorías principales: aprendizaje supervisado, no supervisado y por refuerzo, cada uno con diferentes enfoques y aplicaciones.

Para predecir qué tipo de cotizaciones tienen mayor probabilidad de ser convertidas en ventas reales, se pueden utilizar diferentes tipos de modelos de aprendizaje automático. En términos sencillos, el modelo "aprende" a partir de los datos históricos a identificar patrones y relaciones entre las variables que pueden influir en la probabilidad de conversión en ventas.

Al contar con un conjunto de datos históricos etiquetados (es decir, con información sobre qué cotizaciones se convirtieron en ventas y cuáles no), se puede entrenar el modelo para que aprenda a hacer predicciones en función de las variables relevantes identificadas

Una vez que el modelo está entrenado, se puede utilizar para hacer predicciones sobre la probabilidad de conversión en ventas para nuevas cotizaciones que ingresen al sistema. De esta forma, se tendrá una herramienta que permitirá tomar decisiones más informadas y enfocar mejor los recursos disponibles en el área comercial.

3.3.1. Técnicas supervisadas

Son un tipo de aprendizaje automático que se utiliza para modelar la relación entre variables de entrada y una variable objetivo. Para entrenar un modelo que pueda hacer predicciones precisas sobre esa variable, se utilizan datos etiquetados, es decir, que contienen valores conocidos para la variable objetivo. Se pueden encontrar técnicas de clasificación y regresión.

Modelos de clasificación: un modelo de clasificación es un tipo de modelo que se utiliza para predecir una variable categórica o discreta, como "sí" o "no", "perdedor" o "ganador". Es decir, se asigna una etiqueta o categoría a una observación basada en ciertos criterios. Por ejemplo, la idea sería asignar a cada solicitud de cotización una etiqueta binaria, por ejemplo, "convertido en venta" o "no convertido en venta".

- Modelos de regresión: un modelo de regresión es un tipo de modelo que se utiliza para predecir una variable continua, como por ejemplo el monto de una cotización o el porcentaje de probabilidad de conversión a venta. En este tipo de técnica, se busca establecer una relación entre una variable independiente y una variable dependiente y se utiliza para predecir el valor de la variable dependiente dado un valor específico de la variable independiente. Por ejemplo, estimar un valor continuo para la probabilidad de conversión comprendido entre 0 y 1

Para este tipo de modelos, se necesita un conjunto de datos que incluya información sobre las cotizaciones que se han convertido en ventas reales, como por ejemplo el monto de la cotización, la categoría de producto, el cliente, el vendedor, etc.

3.3.1.1. Regresión logística

La regresión logística es una técnica de modelado estadístico utilizada para analizar la relación entre una variable dependiente binaria y una o más variables independientes. Es ampliamente utilizada en problemas de clasificación y predicción en diversas áreas como la medicina, la economía, la biología, entre otras.

La regresión logística utiliza una función logística para estimar la probabilidad de que una observación pertenezca a una de las dos categorías posibles. La función logística es una curva S que mapea cualquier valor de entrada a un valor de salida entre 0 y 1, lo que permite interpretar la salida como una probabilidad.

Dado que la variable dependiente es binaria (venta o no venta), la regresión logística es una técnica útil para el tipo de problema de clasificación presentado en este proyecto. Para lo cual se utiliza un conjunto de datos que incluye información sobre las cotizaciones previas que se han convertido en ventas reales y las que no, junto con información sobre las variables independientes como el monto de la cotización, la categoría de producto, el cliente, el vendedor, entre otros. La regresión logística se utiliza para modelar la probabilidad de conversión a venta real en función de estas variables, lo que ayuda a identificar patrones y factores que influyen en la conversión de cotizaciones en ventas reales.

3.3.1.2. Árboles de decisión y bosques aleatorios

Los árboles de decisión son una técnica de aprendizaje automático que permite estructurar los datos en forma de árbol, donde cada nodo representa una decisión basada en una característica determinada y las hojas representan las etiquetas de clasificación o valores de salida. Esta técnica es particularmente útil para analizar datos no lineales o complejos, ya que permite analizar múltiples características a la vez y determinar cuáles son las más importantes para la predicción.

En este proyecto se pueden utilizar los árboles de decisión para determinar qué variables son más importantes para la conversión de una cotización a una venta, así como para clasificar las cotizaciones en diferentes categorías en función de sus características. Por ejemplo, es posible construir un árbol de decisión que tome en cuenta variables como la categoría del producto, el tamaño de la empresa, el histórico de compras del cliente y el precio de la cotización para predecir la probabilidad de conversión.

Por otro lado, el bosque aleatorio o random forest es una técnica de ensamblado basada en árboles de decisión que consta de una colección de árboles que trabajan juntos para mejorar la precisión. Cada árbol se construye a partir de un subconjunto aleatorio de las características y observaciones del conjunto de datos original, lo que permite reducir la varianza y mejorar la generalización del modelo.

Al construir múltiples árboles de decisión y combinar sus resultados, se puede reducir la probabilidad de sobreajuste y mejorar la precisión general del modelo. Además, el random forest también puede proporcionar información sobre la importancia relativa de diferentes variables en la predicción de la conversión, lo que podría ser útil para optimizar la estrategia de ventas.

3.3.1.3. Máquina de soporte vectorial

Es una técnica de clasificación que busca construir un hiperplano en un espacio de características con el objetivo de separar de manera óptima dos o más clases. La separación óptima se logra maximizando la distancia entre el hiperplano y los puntos de datos más cercanos a él, que se conocen como vectores de soporte.

Un modelo de SVM[10] es particularmente útil cuando se trata de problemas de clasificación no lineales, ya que permite transformar los datos de entrada en un espacio de características de mayor dimensión, donde la separación lineal es posible. Esto se logra mediante el uso de funciones de kernel, que permiten la transformación de los datos sin necesidad de calcular explícitamente las características de mayor dimensión.

[10] SVM: Máquinas de soporte vectorial, del inglés Support Vector Machines

En el proyecto se puede aplicar SVM para clasificar las cotizaciones según su probabilidad de conversión a venta. Se utilizan diferentes funciones de kernel para transformar las características de las cotizaciones en un espacio de características de mayor dimensión, SVM busca la mejor línea o hiperplano que pueda separar las diferentes cotizaciones en dos clases: aquellas con alta probabilidad de conversión y aquellas con baja probabilidad de conversión. Además, se podría ajustar la complejidad del modelo mediante la selección del parámetro de regularización adecuado, lo que permitiría evitar el sobreajuste y mejorar la generalización del modelo. Luego, utiliza esta línea o hiperplano para hacer predicciones sobre la probabilidad de conversión en ventas de nuevas cotizaciones que se presenten en el futuro.

3.3.1.4. Otras técnicas supervisadas

Redes Neuronales Artificiales (ANN: Artificial Neural Network): Las redes neuronales artificiales son una técnica de aprendizaje automático que busca imitar la estructura del cerebro humano. Están compuestas por capas de nodos o "neuronas" que procesan y transmiten información a través de conexiones ponderadas. Cada neurona recibe entradas de otras neuronas o del conjunto de datos y aplica una función de activación para generar una salida que se transmite a la siguiente capa. Esta técnica es particularmente útil para problemas de clasificación y regresión no lineales o complejos, donde los modelos lineales no son suficientes para capturar la relación entre las características y la variable objetivo. Además, las redes neuronales pueden aprender a identificar patrones y relaciones en los datos por sí mismas, lo que las hace adecuadas para problemas donde las características importantes no son conocidas a priori.

Al entrenar una red neuronal con datos históricos de consumo de servicios, se podría obtener un modelo capaz de predecir la demanda futura, patrones de consumo y la detección de anomalías en los datos, por parte de las empresas clientes.

Potenciación del gradiente (Gradient boosting): Es una técnica de aprendizaje automático supervisado que combina varios modelos más simples para crear un modelo más preciso y complejo. El enfoque consiste en entrenar los modelos secuencialmente, con el objetivo de corregir los errores cometidos por los modelos anteriores. El algoritmo utiliza la técnica de gradiente para optimizar la función de pérdida y ajustar los pesos asignados a cada modelo individual. El resultado final es un modelo en el que la suma de los modelos simples individuales brinda una precisión significativamente mejor que cualquiera de los modelos por separado.

La técnica de Gradient Boosting podría ser una buena opción para mejorar la precisión de los modelos de clasificación y predicción. Sin embargo, es importante tener en cuenta que esta técnica puede ser computacionalmente intensiva y requiere un ajuste adecuado de los parámetros para evitar el sobreajuste.

3.3.2. Técnicas no supervisadas

Las técnicas no supervisadas son una clase de técnicas de aprendizaje automático en las que no se proporciona una etiqueta o salida objetivo. A diferencia de las técnicas supervisadas, en las que se busca predecir una variable objetivo conocida, en las técnicas no supervisadas se busca descubrir patrones y relaciones ocultas en los datos sin tener una salida o etiqueta conocida.

Estas técnicas son particularmente útiles en el análisis exploratorio de datos, ya que pueden ayudar a identificar patrones y grupos de datos similares, lo que puede llevar a una mejor comprensión de los datos y posteriormente, a una mejor toma de decisiones.

Las técnicas de aprendizaje no supervisado se clasifican principalmente en las categorías de agrupamiento y reducción de dimensionalidad.

3.3.2.1. Agrupamiento (Clustering):

Mediante las técnicas de agrupamiento se dividen los datos en grupos basados en similitudes entre ellos. En el proyecto, se utilizó este tipo de algoritmos para identificar patrones, por ejemplo, agrupar a los clientes en diferentes categorías basadas en sus características comunes, como el tamaño de la empresa, la ubicación geográfica, el sector de la industria, entre otros. Una vez agrupados en diferentes categorías, se podría analizar cada grupo para entender mejor las necesidades y comportamientos de cada tipo de cliente. Esto permitiría personalizar las estrategias de marketing y ventas para cada grupo específico, lo que podría mejorar significativamente la efectividad de los esfuerzos comerciales.

- **K Medias (K-Means):** un algoritmo iterativo que trata de dividir los datos en K grupos, donde K es un número especificado por el usuario. Este algoritmo funciona calculando la media o centroide de cada grupo, y asignando cada objeto al grupo cuyo centroide esté más cerca. Este proceso se repite iterativamente hasta que el centroide de los grupos no cambie o se alcance un criterio de convergencia.
  Uno de los puntos clave del algoritmo k-means es la elección del número k de clusters, lo cual puede ser un problema si no se conoce de antemano. Otro aspecto importante es que el algoritmo puede verse afectado por la presencia de outliers o datos atípicos.
  k-means es un algoritmo fácil de implementar y eficiente en términos computacionales, y se utiliza comúnmente en una amplia variedad de aplicaciones, incluyendo la segmentación de clientes, la clasificación de texto y la reducción de dimensionalidad.

- **Agrupación Jerárquica (Hierarchical clustering):** un algoritmo que crea una estructura jerárquica de los datos y los agrupa en clusters. Hay dos tipos principales de agrupamiento jerárquico: aglomerativo y divisivo. En el aglomerativo, se comienzan con tantos grupos como observaciones y se fusionan gradualmente hasta formar un solo grupo o un número predeterminado de grupos. En el divisivo, se comienza con un solo grupo que contiene todas las observaciones y se divide gradualmente hasta alcanzar un número determinado de grupos. El resultado final es un dendrograma que muestra la relación entre los grupos y se puede usar para seleccionar el número óptimo de grupos a partir de los puntos de corte en la jerarquía.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** un algoritmo que identifica agrupaciones densas de datos y las asigna a clusters. El algoritmo funciona identificando puntos densamente poblados y agrupando aquellos que están dentro de un radio especificado (eps: epsilon) y que tienen un número mínimo de vecinos (minPts: Minimum points) en común. Todos los puntos que no cumplen estos criterios se marcan como ruido o outliers. DBSCAN es un algoritmo flexible que puede manejar clusters de diferentes formas y tamaños, y es especialmente útil para identificar estructuras complejas en datos multivariantes.

3.3.3. Procesamiento de texto

El procesamiento de lenguaje natural o NLP (Natural Language processing) se enfoca en la interacción entre la máquina y el lenguaje humano. Este emplea técnicas de procesamiento de lenguaje y aprendizaje automático, con el objetivo de permitir a las computadoras comprender, interpretar y generar lenguaje humano.

Podría ser aplicado en la tarea de clasificación automática de las cotizaciones y la extracción de información relevante de las mismas. Por ejemplo, se podría utilizar el procesamiento de lenguaje natural para identificar automáticamente el tipo de producto o servicio que se está cotizando, así como también la cantidad, el precio y otros términos clave que podrían influir en la decisión de compra. Esto permitiría automatizar el proceso de clasificación de las cotizaciones y ahorrar tiempo y recursos en la revisión manual de cada una de ellas. También se podría utilizar el procesamiento de lenguaje natural para analizar los comentarios y opiniones de los clientes en las cotizaciones y así obtener información valiosa para mejorar la calidad del servicio o producto ofrecido.

Las técnicas de NLP también pueden clasificarse entre supervisadas y no supervisadas.

Técnicas supervisadas de NLP: Implican el uso de etiquetas previas y conocidas para entrenar un modelo, con el objetivo de realizar tareas como la clasificación de texto o la extracción de información. Algunos ejemplos de técnicas supervisadas de NLP incluyen:

- Clasificación de texto: donde se entrena un modelo para clasificar texto en diferentes categorías, como spam, noticias, opiniones, etc.
- Reconocimiento de Entidades Nombradas (NER: Named Entity Recognition): donde se entrena un modelo para detectar y etiquetar entidades nombradas, como personas, lugares, organizaciones, etc.
- Análisis de sentimiento: donde se entrena un modelo para determinar la polaridad (positiva, negativa o neutral) de un texto.

Técnicas no supervisadas de NLP: No requieren etiquetas previas y se centran en la exploración y descubrimiento de patrones en los datos de texto sin un objetivo específico previamente definido. Algunos ejemplos de técnicas no supervisadas de NLP incluyen:

- Agrupación de texto: donde se agrupan documentos similares en categorías o clústeres.
- Modelos de tópico: donde se identifican temas o tópicos en un corpus de texto y se asignan a cada documento una distribución de probabilidad sobre estos temas.
- Reducción de dimensionalidad: donde se aplican técnicas como Latent Dirichlet Allocation (LDA) para reducir la dimensionalidad del espacio de características y visualizar los temas más relevantes en un corpus de texto.

Para el desarrollo del modelo es necesario realizar procesamiento del texto de descripción de cotizaciones con el propósito de categorizar los productos ofertados.

Teniendo en cuenta que el texto es una secuencia de caracteres y símbolos el primer paso para el procesamiento de lenguaje consiste en convertir las líneas de texto o párrafos en un listado de palabras con la cual se pueda trabajar; para este proceso se pueden considerar las siguientes etapas principales:

Tokenización: Consiste en convertir las secuencias de caracteres y palabras en una lista de palabras.

Filtrado: Dado que las palabras pueden ir acompañadas de signos o presentadas entre mayúsculas y minúsculas es necesario homogeneizar las palabras con fin de que no sean identificadas cada vez como una palabra diferente, para esto se pueden convertir todos los caracteres a minúsculas, eliminar espacios, saltos de línea y caracteres no alfabéticos, entre estos los puntos y las tildes, Algunos números pueden ser convertidos a palabras para no ser eliminados.

Existe una librería llamada unidecode la cual realiza la transliteración de un alfabeto al otro y es útil para este proyecto en el cual las descripciones se encuentran en español. Al utilizar esta función se pueden sustituir las vocales con tildes a vocales sin tildes.

Otra etapa del filtrado o limpieza del texto consiste en excluir aquellas palabras muy habituales en el idioma que no aportan significado dentro de la frase o párrafo, estas palabras son llamadas **Stopwords** y están definidas de acuerdo a cada idioma,

Considerando el número de palabras que pueden agregar o no valor, existen estrategias como poner límites máximos o mínimos a sus apariciones, esto considerando que una palabra que aparece en todas las descripciones puede no aportar información diferenciable, pero una palabra que solo aparece una vez entre muchas descripciones de igual manera podría no ser relevante.

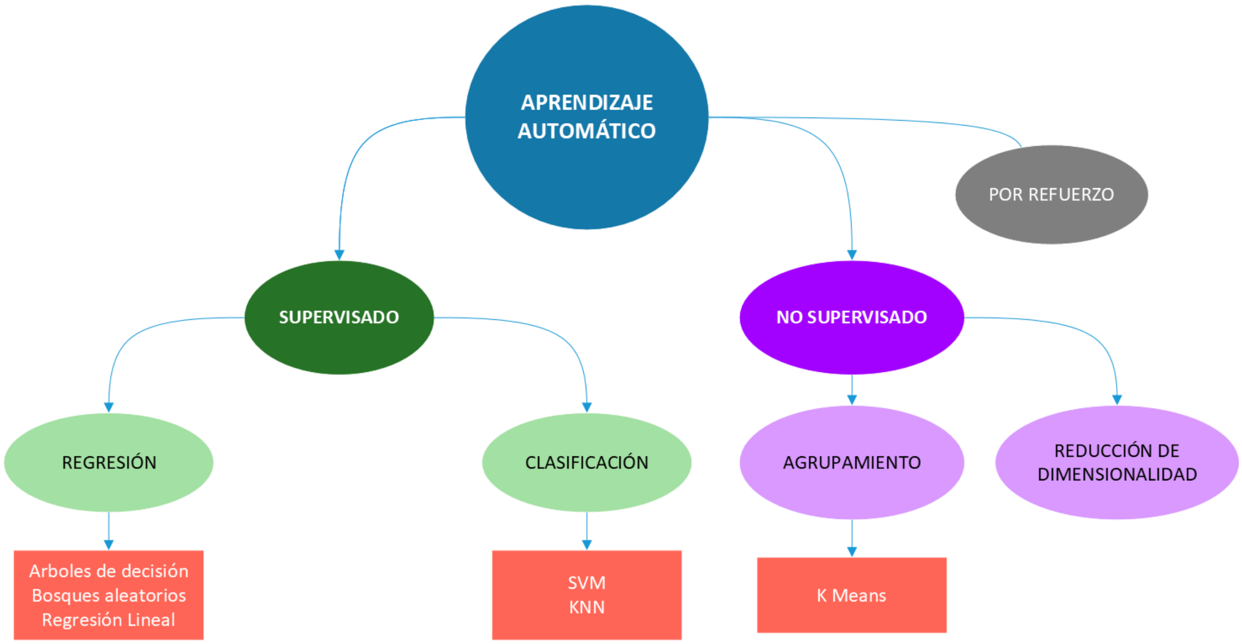
Una vez se cuenta con la lista de palabras filtradas es necesario transformar los textos a números que puedan ser procesados por los algoritmos de aprendizaje, para esto son utilizadas herramientas como la bolsa de palabras

Bolsa de Palabras (Bag of Words): este método permite obtener una representación numérica de los textos, básicamente se incluyen las palabras en una bolsa y por cada mensaje se construye un vector que contabiliza el número de veces que cada palabra está presente en el mensaje, si la bolsa contiene 100 palabras, se tendría una matriz de 100 columnas cada una asociada a una palabra y un número de filas correspondiente a cada descripción.

En la Figura 1 se presenta un diagrama reducido de las clasificaciones generales para los modelos de aprendizaje automático y se muestra en el nivel inferior solo los modelos involucrados en el desarrollo de este proyecto.

Figura 1

Diagrama de la clasificación de modelos de aprendizaje automático utilizados



3.4. Antecedentes

En febrero de 2022, se publica un trabajo en el cual se utilizan datos recopilados mediante solicitudes de ofertas (QRF), se aplican técnicas de Machine Learning y Procesamiento de Lenguaje Natural y se busca analizar información de entrada de cada solicitud para permitir una predicción de ventas futuras (forecasting) [5]. Según los resultados reportados por los autores (D. Rohaan, Et. al.), el método aplicado es capaz de identificar aproximadamente un 70% de casos de ventas reales con una precisión del 50%. Este caso-estudio, realizado con datos de una empresa comercializadora de repuestos, da una guía con respecto a la metodología, expectativas y alcance para la aplicación de la ciencia de datos y la inteligencia artificial en el análisis en ventas B2B.

Por otro lado, en la tesis de maestría publicada por K. Kasinathan (2021) [6], se plantea un caso estudio en el cual se analiza el histórico de ventas de una organización y con base a esta información se evalúan las oportunidades de negocios recibidas evaluándose de acuerdo con su probabilidad de cierre y permitiendo predecir las ventas en un periodo determinado. Este modelo predictivo fue acompañado de un módulo de visualización para permitir el seguimiento del pipeline en tiempo real. En el caso de este documento, nos permitió identificar aquellos desafíos a enfrentar durante el desarrollo del proyecto.

En la actualidad y en el contexto de la empresa, se han comenzado a implementar algunas herramientas de visualización y monitoreo de información referente al proceso comercial. Sin embargo, no existe una formalización de aquellos indicadores claves (KPI) dentro del área comercial y no se hace un análisis detallado de la información recopilada. Periódicamente se elabora un informe de gestión comercial donde se analiza información relacionada con: Total de negocios cerrados, Cumplimiento de metas de ventas y facturación y estado del pipeline con respecto al periodo en estudio. Este informe se organiza manualmente y no existe una herramienta para seguimiento en tiempo real, lo cual dificulta su consulta rápida y confiable.

4. DESARROLLO DEL MODELO PREDICTIVO

El procedimiento utilizado para el desarrollo del proyecto puede enmarcarse en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), comúnmente utilizada en proyectos de ciencia de datos. De acuerdo a sus principales etapas, el proyecto se desarrolló como se resumen en las siguientes actividades.

Comprensión del negocio: Entendiendo la necesidad en la empresa de predecir si una cotización recibida podría convertirse en una venta real; el desarrollo de este objetivo se centró en diseñar diferentes modelos de aprendizaje automático, cuyo rendimiento fue evaluado mediante diferentes métricas para de esta manera seleccionar el que se adecuaba de mejor manera al conjunto de datos de trabajo y las restricciones estratégicas del negocio.

Comprensión de los datos: Para desarrollar el modelo predictivo, se empleó la información histórica de la empresa como base de datos de trabajo, utilizando el CRM de la organización como fuente de datos primaria. Esta etapa del desarrollo se enfocó en realizar un análisis exhaustivo del CRM y su estructura de base de datos. Durante este proceso, se llevó a cabo una investigación minuciosa para comprender el funcionamiento del CRM, así como la naturaleza y el alcance de la información almacenada en él. Se examinaron diversos aspectos, como la cantidad de campos y registros disponibles, y se identificaron los patrones y las relaciones existentes entre los datos.

En el desarrollo del modelo, Se hizo hincapié en la calidad de los datos utilizados en el proceso de entrenamiento reconociendo su impacto en la precisión de las predicciones. Por esto se incluyó información clave, como datos de cotizaciones anteriores que abarcaban detalles sobre el cliente, el producto o servicio cotizado, el valor de la cotización y si se convirtió en una venta real.

Adicionalmente, se exploraron fuentes externas de datos que podrían enriquecer la información del CRM y proporcionar una perspectiva más amplia del entorno comercial. Entre los diversos factores que podrían influir en la probabilidad de conversión de una cotización en una venta real, se consideró el entorno económico, comprendiendo que sus condiciones podrían afectar la disponibilidad de recursos de los clientes y su disposición a realizar compras. Por lo tanto, estas fuentes externas incluyen información relevante como lo son: datos demográficos, tendencias del mercado e indicadores económicos.

Es importante destacar que el análisis de factores macroeconómicos es altamente sensible, ya que pueden producirse rupturas en las tendencias establecidas. Hechos coyunturales, como recesiones económicas, problemas de estabilidad política o falencias estructurales, pueden tener un impacto significativo en los resultados del modelo predictivo. Estos eventos imprevistos pueden alterar los patrones históricos de ventas y comportamiento del mercado, así como incertidumbre en el entorno empresarial, lo que afectaría la precisión de las predicciones realizadas por el modelo. Estos factores externos requieren una vigilancia constante y por lo tanto, resulta necesario establecer períodos de revisión, actualización y mejoramiento del modelo para adaptarse a los cambios en el panorama económico, cuando se presenten estas circunstancias. Esto implica considerar variables adicionales que no hayan sido incluidas dentro de este proyecto.

Una vez realizada la extracción de los datos necesarios del CRM que implicó la selección de las variables relevantes, se realizó un análisis exploratorio de los datos para comprender mejor su distribución, identificar patrones, detectar posibles inconsistencias o valores atípicos, y obtener ideas iniciales sobre las relaciones entre las variables.

Preparación de los datos: Una vez recopiladas y analizadas las variables necesarias para la construcción del modelo predictivo, se continuó con el procesamiento de los datos utilizando técnicas y algoritmos específicos. Este proceso implicó la definición de la variable a predecir, el filtrado de registros necesarios y la identificación y la corrección de posibles inconsistencias en las cuales fueron aplicadas técnicas de imputación y eliminación de registros para tratar los valores nulos, faltantes y atípicos.

Se realizó la transformación de los datos en un formato adecuado para el análisis; lo que incluyó la transformación de variables categóricas y normalización de atributos. Finalmente se realizó la división del conjunto de entrenamiento y pruebas y se aplicaron técnicas de balanceo de clases para abordar desequilibrios en la distribución de las clases objetivo. Estas etapas fueron fundamentales para establecer una base sólida y confiable sobre la cual construir el modelo y lograr pronósticos precisos y útiles para la organización.

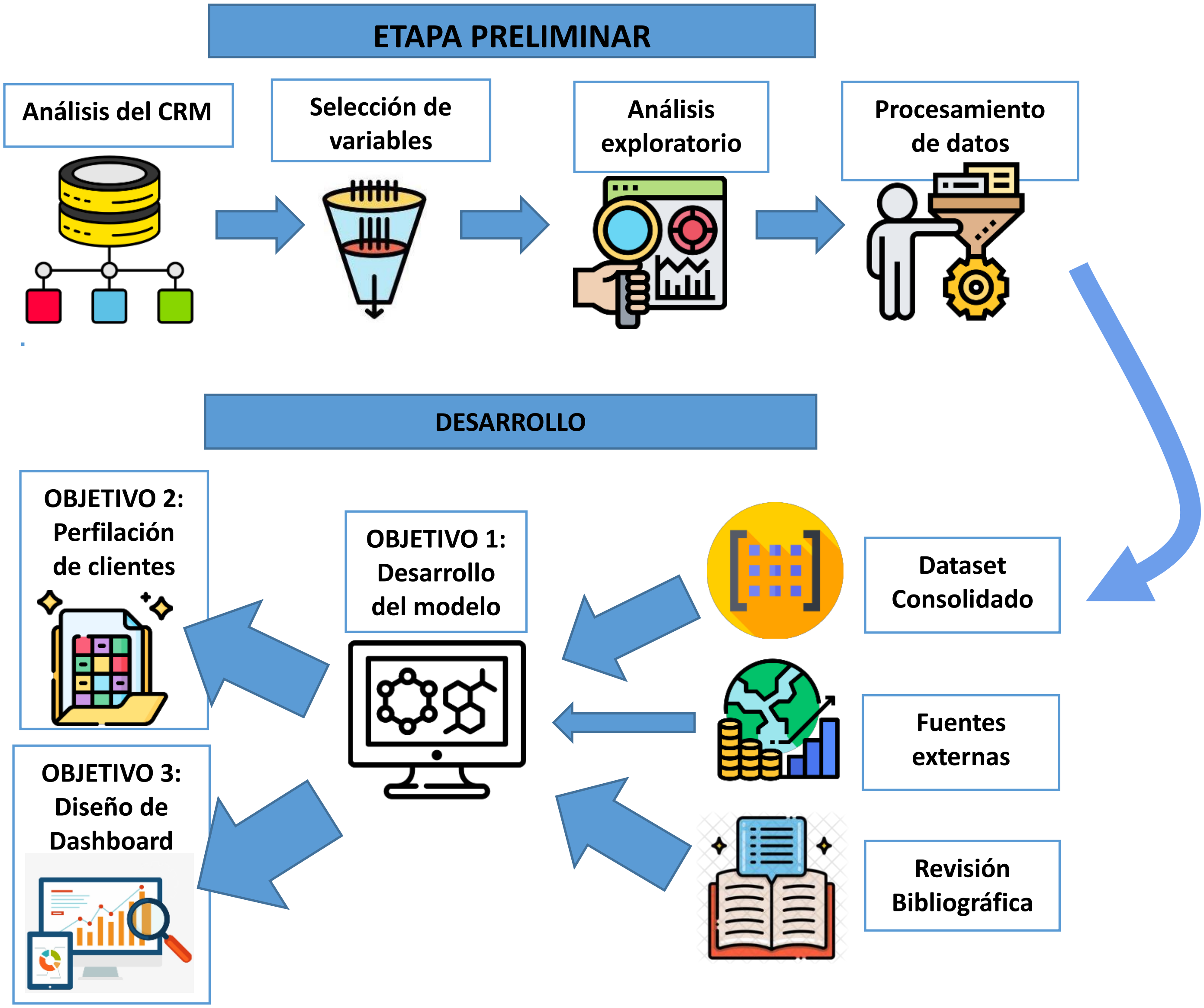
Modelado: Con una visión sobre las mejores prácticas y los enfoques utilizados en el desarrollo de modelos predictivos similares en el contexto empresarial, se procedió con el desarrollo de diferentes modelos de aprendizaje automático; Se desarrollaron modelos de regresión logística, árboles de decisión, bosque aleatorio y máquina de soporte vectorial. Cada modelo fue ajustado probando múltiples hiper parámetros.

Evaluación: En esta etapa, se evaluó el rendimiento de los modelos utilizando métricas y técnicas de validación adecuadas; se compararon los resultados obtenidos por diferentes modelos y entre estos se seleccionó el más adecuado al objetivo propuesto. Fue primordial tener en cuenta los criterios de evaluación establecidos según la estrategia del negocio, además de los resultados de las métricas y el rendimiento.

En la figura 2 se presenta un esquema general del diagrama de flujo seguido para el desarrollo del proyecto.

Figura 2

Diagrama de flujo de las etapas de desarrollo del proyecto



Cada una de las etapas y actividades desarrolladas se describen con más detalle a continuación.

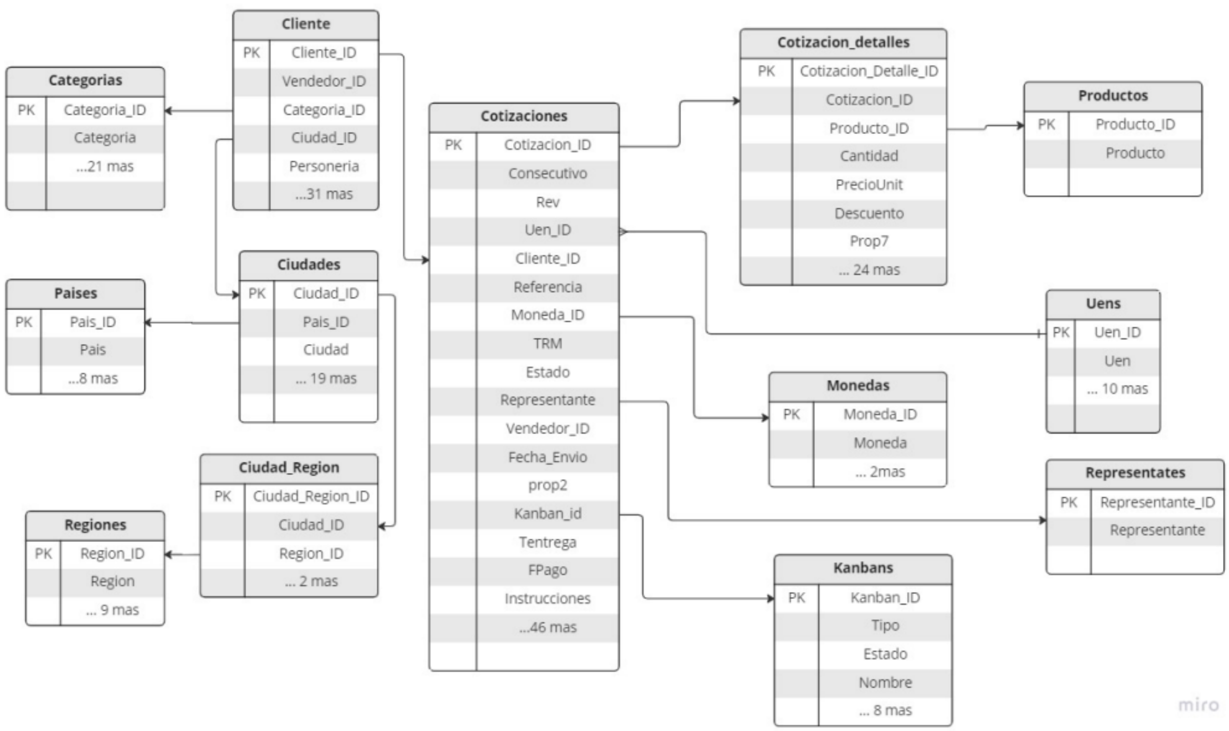
4.1. Análisis de CRM y estructura de base de datos

En primera instancia era necesario tener una visión clara de cómo se organizan y almacenan los datos e información comercial con que cuenta la organización; por esto la comprensión y análisis de la estructura del CRM desempeñó un papel fundamental en el proyecto. Al comprender el modelo de datos utilizado por el sistema, fue posible identificar las diferentes tablas y relaciones existentes. En particular, en este proyecto, se tuvo enfoque en las cotizaciones y los clientes como elementos clave. de esta manera se construyó un modelo entidad-relación de la base de datos para representar de manera visual y comprensible la estructura de estas entidades, sus atributos y sus interacciones.

El esquema simplificado del modelo entidad relación de la base de datos se aprecia en la Figura 3.

Figura 3

Modelo Entidad Relación base de datos de ventas



El modelo entidad-relación fue esencial para comprender la arquitectura de datos, garantizar una correcta interpretación de la información; a su vez que permitió realizar consultas y extracciones de datos de manera efectiva, asegurando que los datos utilizados en el desarrollo del proyecto fueran precisos y relevantes.

Considerando que cada tabla tiene varios atributos, se seleccionaron inicialmente los campos más relevantes buscando no sobredimensionar el dataset inicial; posteriormente en el análisis exploratorio fue posible definir si algunos atributos adicionales podrían excluirse o quizás reintegrarse.

4.2. Extracción de datos y preparación del dataset

Categoría de la cotización

Uno de los desafíos iniciales fue considerar para el modelo los productos presentes en cada cotización con fin de identificar aquellos productos más probables de ser comprados por los clientes de la compañía. Esto hubiera implicado que cada producto sería ubicado como un parámetro o columna adicional en el dataset, lo cual finalmente no resultaría viable técnicamente pues incrementaría la complejidad de manera significativa y los algoritmos podrían tomar demasiado tiempo de procesamiento. Para evitar este inconveniente fue necesario modificar la consulta de extracción de información del sistema.

Se consideró importante incluir dentro del dataframe el campo "categoría de producto". Esto representó un reto puesto que el dataframe original de cotización tiene un solo registro por cada cotización, incluyendo la categoría; esto implica que se podría repetir el número de línea cuando en una cotización existan productos de diferentes categorías.

Para no afectar el proceso de extracción de cotizaciones se decidió extraer las categorías por cotización en un archivo diferente. La estructura de este nuevo archivo entrega la cotización y la categoría a nivel de línea; esto significa que, si una cotización tiene varias categorías, cada nueva categoría se representa en una línea adicional en el archivo. Un ejemplo de lo mencionado se representa en la Tabla 1.

Tabla 1.

Agrupamiento de productos según Id Cotización y Categoría

Id Cotización	Id Categoría	Cantidad
1	A	11
1	C	3
1	E	3
2	B	6
2	D	1

Nota: *Para cada cotización (Id Cotización), se presenta la categoría (Id Categoría) asociada a cada producto presente en la cotización, Cantidad representa el número de productos presentes por cada categoría dentro de cada cotización.*

Una vez cargado el archivo en el notebook de Jupiter, se realizó la traspuesta con el propósito de organizar las categorías a nivel de columna; de esta manera se evitó el problema de duplicación del número de cotización en diferentes líneas.

Al realizar un análisis exploratorio de las categorías existentes, se encontró que algunas categorías podían recategorizarse dado que hacían referencia a otras ya existentes. con el fin de disminuir el número de variables además de evitar las posibles correlaciones entre categorías similares, se realizó una matriz de equivalencias y fue posible reducir el número de variables del dataset.

Una vez se realizó la unión con el dataframe original de cotizaciones, se obtuvieron 23 nuevas columnas, una por cada categoría y en caso que una cotización no contenga productos de alguna de ellas, el valor será de 0. Si tiene varios productos que pertenecen a la misma categoría, el sistema coloca la suma bajo la columna de la categoría correspondiente.

En la tabla 2 se muestra una representación del ejemplo anterior mostrado en la tabla 1 después del proceso de reorganización.

Tabla 2.

Reorganización de productos según Id cotización y categoría reorganizada

Id Cotización	...	A	B	C	D	E	...
1		11	0	3	0	3	
2		0	6	0	1	0	

Nota: *Se reorganizan las categorías de productos a nivel de columna, con una sola fila por cada cotización haciendo posible su integración al dataframe general*

Valor de la cotización

Dado que el valor de la cotización es calculado y no se almacena en la base de datos, Fue necesario probar diferentes formas de calcular el costo total de cada línea, verificando que los totales de las cotizaciones generados por la consulta correspondan con los totales de la cotización en el sistema.

Estado de la cotización

Cada cotización tiene un parámetro correspondiente al estado, es decir si fue ganada, modificada, pérdida, etc. En este sentido se excluyen de la consulta aquellas que no se tiene una clasificación concluyente, es decir algunas cotizaciones vigentes o en proceso de negociación, pero sin haberse convertido en venta o finalmente perdida.

Fecha de la cotización

Revisando el dataframe con PANDAS se encontró que una gran cantidad de registros tienen la fecha de cotización vacía, al ser un porcentaje alto se vería comprometida la calidad de los modelos generados. para corregir esta situación en acuerdo con la persona experta en el sistema de la empresa, se decidió corregir los registros vacíos, para ello se utiliza la fecha de creación del registro adicionando 15 días laborales. Este es el promedio de días que toma la construcción de una oferta de un proyecto.

Eliminación de variables

Entre algunas de las columnas que inicialmente se incluyeron, se decidió eliminar varios campos, por ejemplo, los identificadores como id de la cotización, fechas auxiliares como fecha de envió o fecha de pago, etc.

Eliminación de registros vacíos

Dentro de este proceso se identificaron algunos registros donde el precio o la referencia estaban vacíos. Se procedió a eliminar estos registros particularmente debidos a las siguientes razones: 1. Cotizaciones de prueba o 2. Garantías con valor 0.

4.3. Fuentes de datos externas

Producto Interno Bruto (PIB)

La compañía tiene un alto porcentaje de ventas internacionales, como variable de análisis para los modelos se decidió incluir el crecimiento en producto interno bruto del país y el año al momento de generar la cotización.

La fuente de la información seleccionada fue el fondo monetario internacional[11], que tiene disponible un archivo consolidado con información histórica y también proyecciones futuras para todos los países. El archivo se descargó en formato excel y se procesó con herramientas de Python.

Para poder incluir el porcentaje del PIB se inició depurando el archivo recibido desde el Fondo monetario internacional, eliminando múltiples columnas relacionadas al PIB de años comprendidos entre 1980 y 2010. Esto dadas las características de los registros en el CRM que datan a partir de esta fecha.

[11] Dataset de crecimiento economico por paises del Fondo Monetario Internacional: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/OEMDC/ADVEC/WEOWORLD

El archivo del FMI, contiene el año a nivel de columna, pero para hacer el cruce con las cotizaciones este debe ir a nivel de fila junto al país. para solucionarlo, se procedió a realizar la traspuesta.

Al momento de unir el archivo que contiene la información del PIB con el archivo de cotizaciones se encontró que hay registros de países que han tenido cotizaciones pero que su valor no fue actualizado, la razón es que los nombres de los países no coinciden, para solucionarlo se decide crear un archivo intermedio donde se asocian los nombres de los países en el sistema de la empresa. con el nombre del país como lo maneja el FMI.

Luego de este paso, el PIB del país que genera la cotización en el año de la misma se asocia en una columna adicional.

Tasa representativa del mercado (TRM)

Considerando que los clientes se encuentran en diferentes países, algunas cotizaciones contienen los valores en pesos colombianos y otras en dólares; para homogeneizar las magnitudes se realiza el cambio de las cotizaciones en pesos a dólares usando la serie histórica de la tasa representativa del mercado (TRM) obtenida de la página web del banco de la república[12]. Esta serie se vincula a nuestro dataset para obtener la TRM en la fecha de envío de la cotización y de esta manera realizar la conversión a Dólares.

4.4. Análisis exploratorio

La etapa de análisis exploratorio fué crucial en el proyecto al facilitar una comprensión más completa de los patrones, tendencias y características presentes en el conjunto de datos. Después

[12] Página web Banco de la Republica de Colombia: https://www.banrep.gov.co/es/estadisticas/trm

de haber pasado por la etapa de preparación del dataset, en la cual se realizaron diversas acciones como la eliminación de variables y registros redundantes, además de la limpieza y depuración de datos inconsistentes y faltantes; se obtuvo un dataframe con 46 columnas y 3519 registros.

El análisis exploratorio brindó la capacidad de revelar información valiosa y detectar posibles anomalías o patrones inesperados en los datos, lo cual permitió obtener una visión general de la calidad y la integridad del dataframe antes de adentrarse en la implementación de modelos. Se evaluaron relaciones entre las variables, así como posibles correlaciones significativas. También se logró identificar la presencia de valores atípicos que debieron ser ajustados.

En la tabla 3 se detallan las columnas seleccionadas y su descripción.

Tabla 3

Descripción de las columnas utilizadas en el dataframe

Columnas	Tipo	Descripción
rev	Continua	Número de Revisiones de la Oferta
uen	Categórica	Unidad de Negocio o Sector correspondiente. Ejemplo: Molinería de Arroz
cliente_id	Categórica	Código de Identificación del cliente dentro de la BBDD
client_cat	Categórica	Clasificación del cliente por tamaño de organización. Ejemplo: Empresa Familiar
referencia	Categórica	Descripción breve del objeto de la cotización
moneda	Categórica	Moneda en la cual se realiza la cotización
TRM	Continua	Tasa de cambio de presupuestación al momento de la cotización
estado	Categórica	Variable predictora, indica el estado final de la cotización. Ganada/Perdida/Descartada
prop2	Categórica	Tipo de Oferta (Proyecto, Servicio, Maquina Suelta, Repuestos)
tentrega	Categórica	Tiempo de entrega del proyecto en días calendario. Hace ref. a ID en BBDD
ciudad	Categórica	Ciudad de ubicación del cliente
pais	Categórica	País de ubicación del cliente

Columnas	Tipo	Descripción
precio_venta	Continua	Precio Total de Cotización en USD
Envio_year	Categórica	Año de envío de la cotización
Envio_month	Categórica	Mes de envío de la cotización
12 al 12927	Categórica	Indica si la oferta incluye elementos de una categoría de productos (Nombre hace referencia a ID de cada categoría de productos en BBDD)
gpd	Continua	PIB del país del cliente en el año en que se envía la cotización

Nota: *Se presenta la descripción de cada una de las columnas presentes en el dataset final utilizado para el desarrollo de los modelos*

A continuación se presentan resumen de algunas características encontradas para las variables categóricas y continuas.

En la tabla 4 se presenta para cada variable categórica el valor de la moda acompañado del porcentaje de registros que tienen este valor sobre el total.

Tabla 4

Medidas de tendencia central para las variables Categóricas

Columnas	Moda	Porcentaje de registros para el valor de la Moda
uen	Acondicionamiento o centros de acopio	49,3
cliente_id	3791	5,9
client_cat	Corporación	50,0
referencia	Equipos Agroindustriales	4,3
moneda	USD	68,0
estado	X	55,7
prop2	Proyectos	44,9
tentrega	91	27,6

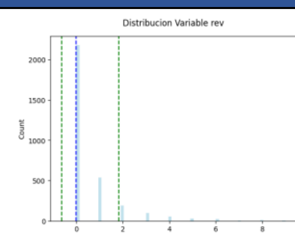
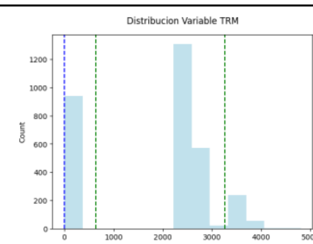
ciudad	Barranquilla	10,4
pais	Colombia	32,0
Envio_year	2019	25,4
Envio_month	11	10,2
12 al 12927	[1-12]	[0,1 - 21,1]

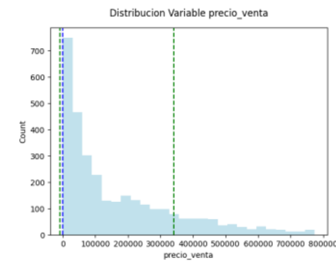
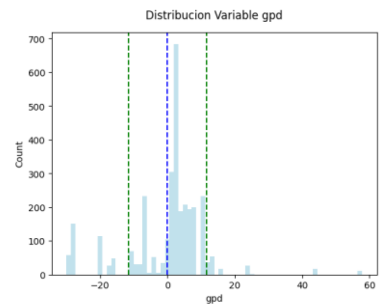
Dado que en este punto el dataset no ha sido ajustado de acuerdo a los hallazgos sobre los datos; Los detalles adicionales sobre las variables relevantes en el proceso, se presentan en el numeral 4.6.4 en el cual se describe el conjunto de datos final una vez culminado el proceso de preparación de los datos, limpieza e imputación de registros.

En la tabla 5 se presentan medidas estadísticas para las variables numéricas.

Tabla 5

Medidas estadísticas para las variables Continuas

Columna	Media	Desv. Est.	min	max	Media - Desviacion
rev	0.598	1.225	0	9	
TRM	1953	1288	1	4806	

Columna	Media	Desv. Est.	min	max	Media - Desviacion
precio_venta	166654	1.00E+06	1	773522	
gpd	0.05	11.58	-30	57	

4.5. Tratamiento de datos faltantes y valores atípicos

El tratamiento de datos faltantes, valores nulos y valores atípicos en el conjunto de datos resultó fundamental para el desarrollo de los modelos. Estos defectos podrían afectar significativamente los resultados y la calidad de las predicciones; pues al limitar la cantidad de información disponible para el modelo, se dificulta la interpretación precisa de los patrones y relaciones en los datos.

Los valores inusuales en el conjunto de datos pueden ser resultado de errores en la entrada de datos, pero también pueden representar casos excepcionales que son legítimos. Así que, si no se tratan adecuadamente, los valores atípicos pueden generar predicciones erróneas o sesgadas. Para evitar estos problemas y garantizar la integridad de los resultados, se aplicaron técnicas de imputación. Estas técnicas permiten estimar y reemplazar los valores faltantes o nulos utilizando métodos estadísticos o basados en modelos. Al imputar estos valores faltantes de manera apropiada, se preserva la coherencia y la representatividad de los datos, lo que mejora la precisión del modelo y los resultados obtenidos.

4.5.1. Valores nulos o faltantes

Al realizar el análisis exploratorio se encontraron valores nulos o faltantes en diferentes columnas y fueron utilizadas diferentes opciones de tratamiento desde eliminación, imputación simple e imputación basada en modelos para las siguientes columnas:

prop2 (Tipo de oferta)

El 29% de los registros tenían valor nulo en la variable prop2. Teniendo en cuenta la recomendación de expertos de la empresa y la importancia que tiene identificar si una oferta corresponde a un proyecto, servicio o máquina individual se decidió aplicar una técnica de imputación para incluir esta variable en el modelo sin eliminar los registros nulos y/o correr el modelo sin una transformación de este campo.

La técnica utilizada consistió en utilizar el paquete **rapidfuzz**[13] de Python para la comparación de las cadenas de texto contenidas en la columna descripción. Este paquete proporciona el algoritmo **Extract0ne** que permite encontrar la mejor coincidencia única entre una cadena de consulta y un conjunto de opciones. para calcular un puntaje de similitud entre la cadena de consulta y cada una de las opciones en el conjunto se utilizó la distancia de Jaro–Winkler como métrica**,** la cual tiene en cuenta la cantidad de caracteres en común y su posición relativa, dando más peso a determinados caracteres según su posición en la cadena**.**

Se realizó con el siguiente orden:

- Se dividió el dataframe entre registros con prop2 nulas (data_revisar) y aquellas con información valida (data_to_match).
- Para cada registro en data_to_match se buscó el primer registro con el mayor porcentaje de similitud en data_revisar y se guarda su index como nueva columna en el dataframe.
- Se ejecuta un ciclo for para buscar utilizando cada index registrado en el punto 2 el campo

[13] Rapidfuzz: biblioteca rápida de coincidencia de cadenas para Python y C++, de: https://pypi.org/project/rapidfuzz/

prop2 en data_to_match y se sobreescribe en data_revisar.

Como resultado de esta transformación fue posible completar todos los registros de la variable prop2, quedando claramente categorizada entre Proyectos, Equipos agroindustriales, Repuestos y Servicios.

Categorías de productos

Se encontraron registros nulos en cada una de las columnas de categorías, esto se puede explicar porque las cotizaciones con alguna columna de categoría nula, no contaban con productos que pertenecían a esa categoría.

Las columnas de categoría con valores nulos se corrigen asignado el valor cero, puesto que estas cotizaciones no contenían productos en ellas

En el caso de **tentrega**, se decide eliminar los 37 registros vacíos ya que no se encuentra una forma de recuperarlos.

4.5.2. Valores atípicos

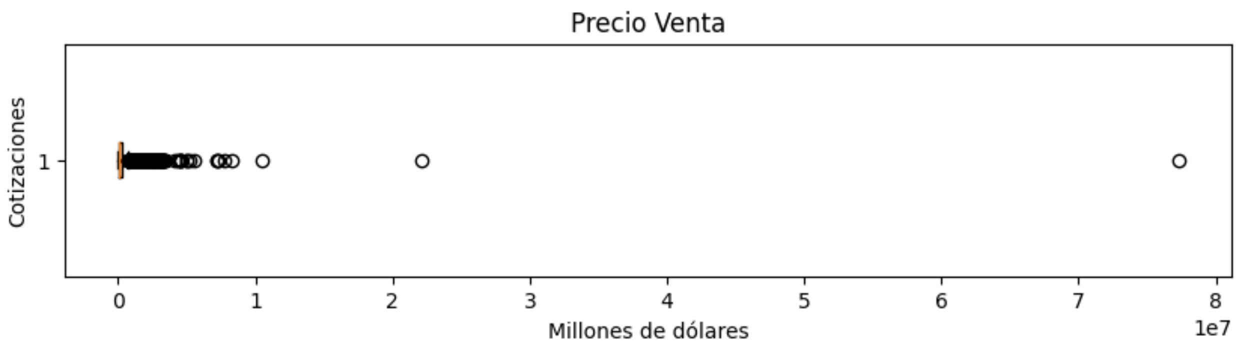
El dataframe cuenta con las siguientes columnas numéricas, excluyendo las columnas de categorías:

- rev
- cliente_id (A pesar de ser numérica, es una variable categórica, ya que representa el código de cada cliente en la BBDD)
- TRM
- precio_venta

Se utiliza un gráfico de cajas y bigotes (Figura 4) para visualizar posibles valores atípicos en la columna de precio_venta observando que hay una considerable cantidad de registros con precios de venta fuera del rango intercuartílico.

Figura 4.

Diagrama de cajas para la variable "precio venta"



Nota: *En el gráfico de cajas es posible identificar visualmente la cantidad de registros con valores atípicos*

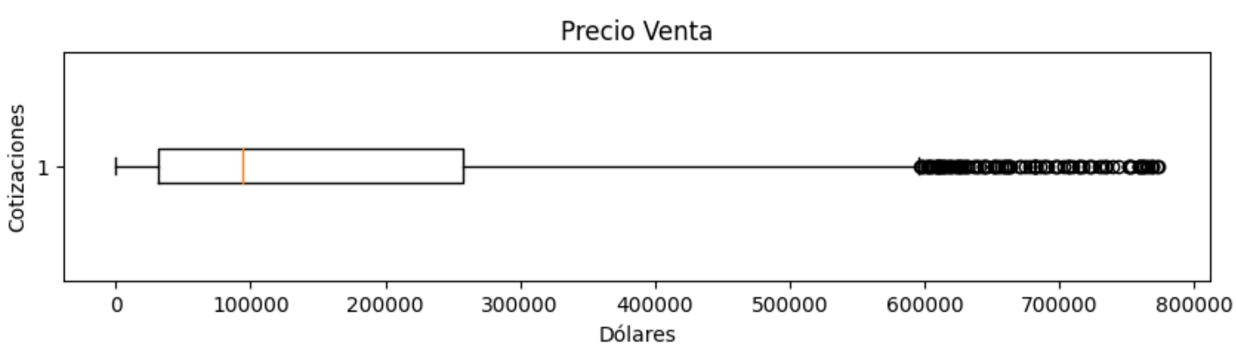
Para mitigar los efectos negativos de estos datos atípicos se eliminan aquellos registros que estén por fuera de los dos intervalos definidos como:

Límite Inferior: Q1 - 1.5 * IQR
Límite Superior: Q3 + 1.5 * IQR

Una vez eliminados los registros con valores atípicos, ya no se cuenta con valores fuera del rango intercuartílico, como se puede observar en el nuevo diagrama presentado en la figura 5:

Figura 5

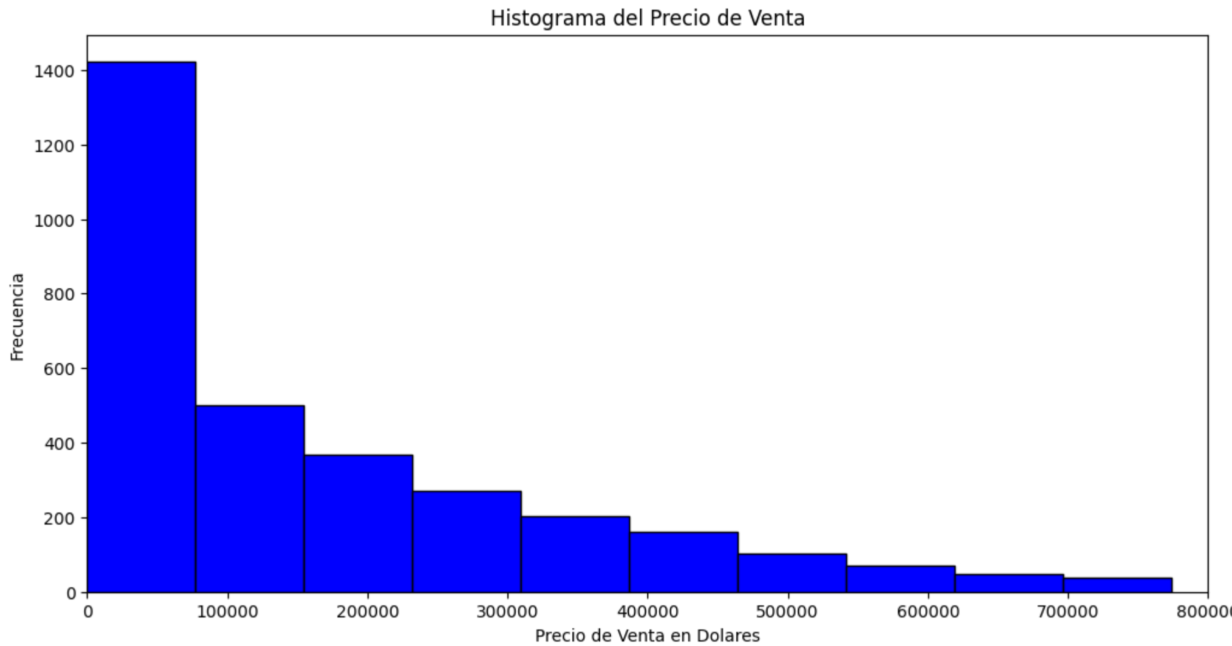
Diagrama de caja para la variable precio venta ajustada



En el histograma presentado en la figura 6 se visualiza la distribución de valores para esta variable.

Figura 6

Histograma de frecuencia para la variable precio de venta ajustada



4.6. Preparación de datos y pre procesamiento

4.6.1. Definición de la variable objetivo:

Con base en su relevancia y utilidad para los objetivos del proyecto, se definió el estado como la variable objetivo o variable de respuesta del modelo**.** El estado es una variable binaria que indica si una cotización se convierte en una venta real o no, esto permite explorar y comprender las características, variables y acciones están asociadas con el éxito o el fracaso de las cotizaciones. se representa con los siguientes valores:

1: Representa que una cotización es ganada, es decir que efectivamente se convirtió en una venta real

0: La cotización no fué ganada, es decir que no se convirtió en venta.

Para simplificar y estandarizar la variable objetivo, se Realizó la recategorización de la variable Estado; En este proceso se convirtió el estado de las cotizaciones ganadas (denominado como Estado “C”) en la clase 1, mientras que los demás estados se clasificaron como 0, indicando que no fueron ganadas.

4.6.2. Transformación de las variables categóricas en variables numéricas

La transformación de variables categóricas en variables numéricas permite aplicar técnicas estadísticas y algoritmos de modelado que requieren datos numéricos, como regresión, árboles de decisión, de esta manera se aprovechó el potencial de la información contenida en estas variables en la construcción de los modelos de análisis y predicción,
En el dataframe se cuenta con las siguientes variables categóricas:

- uen
- client_cat
- moneda
- prop2
- tentrega
- ciudad
- pais

Para realizar esta transformación se utilizó el método fit_transform() de la instancia LabelEncoder de la librería sklearn.preprocessing. Este método analiza las categorías presentes en cada variable categórica y les asigna un valor numérico único a cada una de ellas. Luego, reemplaza los valores originales en la columna con los valores numéricos asignados.

Después de ejecutado este paso las columnas contienen valores numéricos en lugar de las categorías originales, lo cual permite que la información contenida en estas variables se utilice en el modelo de aprendizaje automático

4.6.3. Normalización de los atributos predictores

Se realizó la estandarización de los atributos para evitar que las características con valores notablemente grandes o pequeños dominen la contribución al resultado final del modelo, lo cual podría generar un sesgo indeseado. Además, otro beneficio de la estandarización es que puede ayudar a mejorar la precisión y la velocidad de convergencia de algunos algoritmos de aprendizaje automático.

Para lograrlo, se utilizó la función scale() del módulo preprocessing de la biblioteca Scikit-learn. Esta función realiza una estandarización de los atributos del dataset, transformándolos para que sigan una distribución normal con una media de cero y una desviación estándar de uno. De esta manera se ajustaron los datos a una escala común que facilita su comparación y análisis.

4.6.4. Descripción del conjunto de datos final

El conjunto de datos con la información depurada está compuesto por 45 columnas (variables) y 3136 filas (registros). Con respecto a la variable resultado (estado de cotización), el conjunto de datos se distribuye como se muestra a continuación en la tabla 6:

Tabla 6

Distribución registros según las clases de la variable objetivo

Estado de Oferta	Número de Registros	Porcentaje Sobre Total
0 (Perdida o Descartada)	2583	82.36%
1 (Ganada)	553	17.64%

Este desbalance en la clase objetivo fue revisado como se describe en el **apartado 4.6.6**, donde se aplicó la técnica de oversampling con el objetivo de mejorar el rendimiento de los modelos a entrenar.

Con respecto a la variable "prop2", la cual hace referencia al tipo de Oferta (Proyecto, Servicio, Maquina Suelta, Repuestos) tenemos una mayor cantidad de registros correspondientes a proyectos, mientras que "Servicios" es el tipo con menor número de ofertas dentro del dataset luego del preprocesamiento. La distribución de registros se muestra en la tabla 7.

Tabla 7

Distribución de registros para la variable "prop2" (Tipo de Oferta)

Tipo de Oferta (prop2)	Número de Registros	Porcentaje Sobre Total
Proyectos	1812	57.8%
Máquina Suelta	1051	33.5%
Repuestos	153	4.9%
Servicios	120	3.8%

Analizando la variable **"moneda"**, como se puede apreciar en la tabla 8, se observa que ⅔ de las ofertas fueron realizadas en dólares lo cual puede indicar dos puntos interesantes:

1. Posible mayor número de ofertas para clientes en el exterior con respecto a nacionales.
2. Posible mayor número de ofertas de equipos y proyectos con componentes extranjeros.

Tabla 8

Distribución de registros según moneda

Moneda	Número de Registros	Porcentaje Sobre Total
USD	2075	66.2%
COP	940	30.0%
EUR	121	4.8%

Al revisar la variable **“pais”**, se encuentra que solo el 33.7% de los registros corresponden a clientes en Colombia. Lo cual corrobora lo expuesto anteriormente.

Por último, el conjunto de datos cuenta con 30 columnas correspondientes a las categorías de los productos incluidas en cada registro (oferta). Es decir, para cada ítem dentro de una oferta que haga parte de una categoría, se le suma “1” a la columna de la categoría correspondiente.

Ejemplo: La oferta con ID 1200 incluye cuatro equipos de la categoría “Secadoras” (id 112), por tanto, el valor de la columna correspondiente al id 112 es igual a 4.

4.6.5. División del conjunto de datos

Para garantizar una evaluación confiable del rendimiento del modelo, es fundamental dividir el conjunto de datos en conjuntos de entrenamiento y prueba; esto permite evaluar el desempeño del modelo en datos no vistos previamente, lo cual sería un indicador de cómo se comportaría el modelo cuando se aplique a nuevos datos.

Para este propósito se utilizó la función train_test_split[14] del módulo model_selection de Scikit-learn, En este caso se optó por dividir el conjunto de datos en conjuntos de entrenamiento y

[14] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

prueba para el modelo de acuerdo a la siguiente proporción:
70% para el conjunto de entrenamiento y el 30% para el conjunto de prueba. Esto son 2195 registros de entrenamiento y 940 registros para pruebas
Como resultado se obtuvo cuatro elementos:

X_train: Un subconjunto aleatorio de las variables predictoras para el conjunto de entrenamiento.
X_test: Un subconjunto aleatorio de las variables predictoras para el conjunto de prueba.
y_train: Un subconjunto aleatorio de las variables objetivo para el conjunto de entrenamiento.
y_test: Un subconjunto aleatorio de las variables objetivo para el conjunto de prueba.

4.6.6. Balanceo de clases

Se identificó que la clase objetivo tiene un desbalance significativo, donde se encuentran 2583 cotizaciones No Ganadas y solo 553 cotizaciones ganadas. Este desequilibrio puede afectar el rendimiento y la precisión del modelo, ya que puede sesgar su capacidad para predecir correctamente ambas clases.

Para abordar este desbalance en el dataset, se utilizó el método de **oversampling**. Utilizando la librería **imblearn**[15] se aplica una técnica de sobre muestreo aleatorio en el conjunto de datos de entrenamiento, de manera que la clase minoritaria (aquella que tiene menos muestras) sea aumentada hasta que alcance el 90% de la cantidad de la clase mayoritaria.

La salida de esta función son dos conjuntos de datos sobre muestreados: x_train_res y y_train_res
En la tabla 9 se resume el impacto de la técnica de oversampling sobre el conjunto de datos para entrenamiento “x_train”

[15] **imblearn**: Biblioteca de python utilizada para abordar el problema de desbalance de clases. https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

Tabla 9

Relación de registros antes y después de aplicar oversampling

Conjunto de Datos de Entrenamiento Original		
Estado de Oferta	**Número de Registros**	**Porcentaje Sobre Total**
0 (Perdida o Descartada)	1823	83%
1 (Ganada)	372	17%
Conjunto de Datos de Entrenamiento Ajustado (90%)		
0 (Perdida o Descartada)	1823	52.64%
1 (Ganada)	1640	47.36%

Se puede observar en la anterior tabla que el conjunto de datos de entrenamiento, previo a el balanceo en la clase objetivo, mantiene una relación entre ofertas ganadas y no ganadas muy similar a la del dataset en general.

4.7. Modelo 1 - Regresión logística

Inicialmente se decidió entrenar un modelo de regresión logística usando los datos de entrenamiento previamente balanceados (X_train_res y y_train_res), este modelo fue implementado a través de la clase LogisticRegression[16] de la librería sklearn.linear_model. Se definieron los siguientes parámetros:

El parámetro "solver" se estableció en "lbfgs" (Limited memory Broyden Fletcher Goldfarb Shanno), que es un método de optimización utilizado para estimar los parámetros del modelo. El solver "lbfgs" Es eficiente en términos de tiempo de cálculo y requiere menos memoria en

[16] **LogisticRegression**: Algoritmo de regresion logistica implementado en la biblioteca sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

comparación con otros algoritmos de optimización, una ventaja es que no requiere que se especifiquen tasas de aprendizaje o pasos de tamaño fijo, lo que simplifica su uso y reduce la necesidad de ajustar hiperparámetros adicionales. Además, puede manejar conjuntos de datos grandes y converge rápidamente hacia una solución óptima.

El parámetro "max_iter" indica el número máximo de iteraciones permitidas para que el algoritmo de optimización converja hacia una solución. Si el número es bajo puede que el algoritmo no alcance la convergencia adecuada, mientras que si es muy alto se estaría utilizando más tiempo de entrenamiento que el necesario. En este caso, se estableció en 500 para garantizar que el algoritmo tenga suficientes iteraciones para converger.

El parámetro "tol" establece la tolerancia para el criterio de parada del algoritmo de optimización. Si el cambio en el valor de la función objetivo es menor que "tol", se considera que el algoritmo ha convergido. Si se establece un valor muy bajo, el algoritmo puede requerir más iteraciones para converger, lo que puede aumentar el tiempo de entrenamiento del modelo. Por otro lado, si se establece un valor demasiado alto, el algoritmo puede converger prematuramente y no alcanzar la precisión deseada. En conjuntos de datos grandes, un valor de tolerancia más alto puede ser aceptable ya que las diferencias pequeñas pueden no ser significativas en términos prácticos. Sin embargo, en conjuntos de datos más pequeños o problemas donde se requiere una mayor precisión, es recomendable utilizar un valor de tolerancia más bajo, En este caso, se estableció en 0.01, lo que indica que el algoritmo se detendrá si la mejora en la función objetivo es menor que ese valor.

Una vez creada la instancia de LogisticRegression considerando los anteriores parámetros, se ajustó el modelo a los datos de entrenamiento representados por las variables predictoras X_train_res y la variable objetivo y_train_res. Posteriormente se realizó la predicción de la variable objetivo (y) para los datos de prueba (X_test) utilizando el modelo de regresión logística previamente entrenado. Los resultados del desempeño del modelo se muestran a continuación.

Este modelo inicial arrojó un accuracy de 0.789 y un recall de 0.808; se obtuvo la matriz de confusión[17] [[TN FP] [FN TP]] mostrada en la tabla 10:

Tabla 10

Matriz de confusión: Modelo de regresión logística

Matriz de confusión: Modelo 1		Predicción			
		0		1	
Estado Real	0	TN	615	FP	169
	1	FN	30	TP	127

TN (True Negatives): *Número de casos negativos que fueron correctamente clasificados como negativos.*

FP (False Positives): *Número de casos negativos que fueron incorrectamente clasificados como positivos.*

FN (False Negatives): *Número de casos positivos que fueron incorrectamente clasificados como negativos.*

TP (True Positives): *Número de casos positivos que fueron correctamente clasificados como positivos.*

En términos de la precisión del modelo, se tiene que la tasa de verdaderos positivos (TPR) = (127/(127+30)) es del 80,88% , Esto indica que el modelo clasifica correctamente el 80.88% de los casos positivos en relación con el total de casos positivos.

La tasa de verdaderos negativos (TNR) = (615/(615+169)) es del 78,45%, siendo este porcentaje

[17] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html

de casos negativos que el modelo clasifica correctamente en relación con el total de casos negativos.

La tasa de falsos positivos (FPR) = (169/(615+169)) es del 21,55% , mientras que la tasa de falsos negativos (FNR) = (30/(30+127)) es del 19,12% . Estos porcentajes representan respectivamente la cantidad de casos que incorrectamente se clasifican como positivos siendo en realidad negativos y los que incorrectamente se clasifican como negativos siendo realmente positivos.

4.8. Modelo 2 - Árboles de decisión

Buscando mejorar el desempeño del modelo se creó un segundo modelo utilizando árboles de decisión[18]. El funcionamiento se basa en seguir el camino desde la raíz del árbol hasta una hoja, buscando maximizar la pureza en cada rama, respondiendo las preguntas en cada nodo de acuerdo con los valores de las características de entrada. Finalmente, se asigna la etiqueta de clasificación de la hoja alcanzada como respuesta. Algunos de los parámetros que se plantearon para el modelo fueron los siguientes:

"criterion": Especifica la función de impureza utilizada para evaluar la calidad de una división en el árbol de decisión. En este caso, se estableció en "gini"como medida de impureza; este mide la probabilidad de que un elemento seleccionado aleatoriamente sea clasificado incorrectamente en un nodo dado.

max_depth: Este parámetro establece la profundidad máxima del árbol de decisión. En este caso, no se impusó un límite máximo en la profundidad del árbol, de modo que los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las muestras se hayan asignado a una sola clase. Esto puede conducir a un modelo más complejo y con mayor capacidad de ajuste, aunque también puede aumentar el riesgo de sobreajuste.

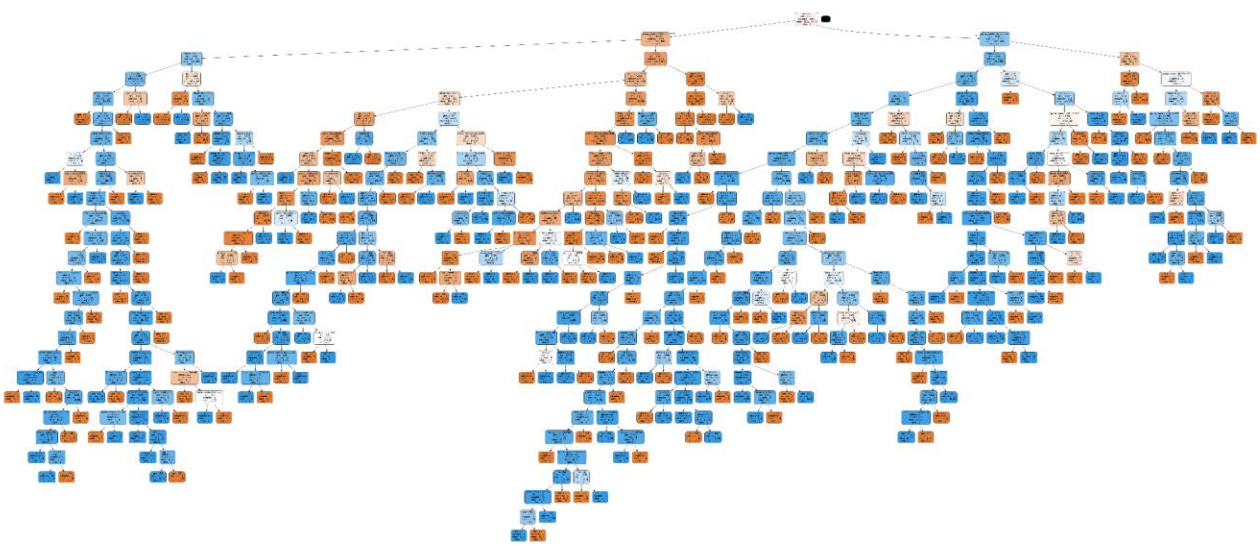
[18] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Después de probar diferentes parámetros los resultados de precisión más altos se encontraron al no limitar la profundidad, para este caso utilizando criterio de Gini la profundidad resultante del árbol fue de 26.

Una vez definido el modelo, se realizó el entrenamiento con el mismo conjunto de datos y se realizaron las predicciones sobre los datos de prueba. Como resultado se obtuvo una mejora en el accuracy pero también un decremento en el recall comparado con la regresión logística

Figura 7

Visualización modelo de árbol de decisión



La matriz de confusión obtenida con el árbol de decisión se presenta en la tabla 11.

Tabla 11.

Matriz de confusión: Modelo de Árbol de decisión

<table>
<tr><th colspan="2" rowspan="2">Matriz de confusión: Modelo 2</th><th colspan="4">Predicción</th></tr>
<tr><th colspan="2">0</th><th colspan="2">1</th></tr>
<tr><td rowspan="2">Estado Real</td><td>0</td><td>TN</td><td>690</td><td>FP</td><td>70</td></tr>
<tr><td>1</td><td>FN</td><td>96</td><td>TP</td><td>85</td></tr>
</table>

Muestra una mayor precisión en la predicción de las cotizaciones que se convertirán en ventas, ya que la diagonal principal (690 y 85) representa el número de predicciones correctas para las cotizaciones que no se convierten en ventas y las que sí se convierten en ventas, respectivamente. Por otro lado, la matriz de confusión obtenida con la regresión logística muestra una mayor cantidad de falsos positivos (169) y menos verdaderos positivos (127) que la matriz de confusión obtenida con el árbol de decisión, lo que indica que la regresión logística puede estar sobreestimando la probabilidad de conversión en ventas

4.9. Modelo 3 - Bosque aleatório

Entendiendo el modelo de bosque aleatorio o random forest[19], el cual utiliza múltiples árboles de decisión para mejorar los resultados se procede a crear este modelo. Este tipo de modelo utiliza una combinación de muestras de entrenamiento y características aleatorias para construir varios árboles independientes y luego combina sus predicciones para obtener un resultado final.

[19] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Entre los parámetros de este modelo se comparten varios utilizados en el modelo de árbol de decisión, un parámetro adicional para este es el siguiente:

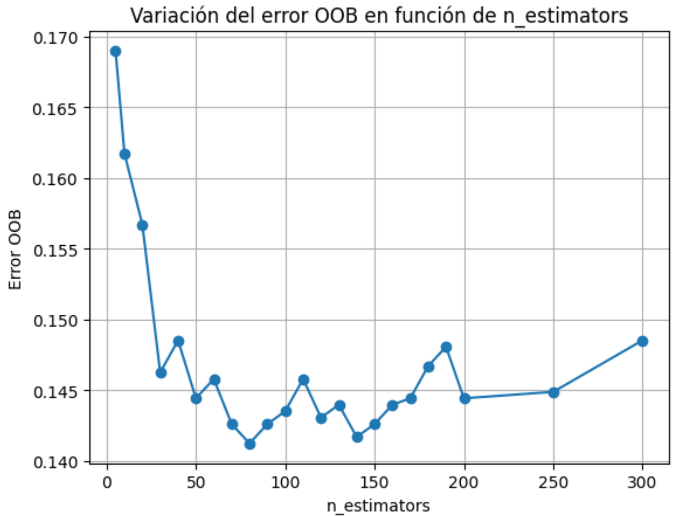
n_estimators: Indica el número de árboles que se utilizarán en el bosque. El aumento en el número de estimadores generalmente mejora el rendimiento del modelo, ya que se consideran más árboles para realizar las predicciones. Sin embargo, un número demasiado alto puede llevar a un aumento en el tiempo de entrenamiento y la complejidad del modelo sin necesariamente mejorar significativamente su rendimiento.

Al probar diferente número de árboles se encontró que la métrica de evaluación accuracy converge rápidamente con bajo número de árboles, y las mejoras dejan de ser significativas alrededor de 40.

Al tomar como criterio el método Out-of-Bag" (OOB) donde las muestras que no se incluyeron en la construcción de cada árbol (muestreo OOB) se utilizan para evaluar el rendimiento del modelo, se encontró que el error alcanza valores mínimos alrededor de 80 árboles y después comienza a variar entorno a valores bajos sin una convergencia clara, pero después de 140 árboles la tendencia es creciente.

Figura 8

Variación del error OOB en función de n_estimators



Nota: *Grafica de del error OOB en calculado en función del número árboles (n_estimators) permite observar el punto donde el error alcanza su mínimo y se estabiliza*

Después de entrenar el modelo y predecir sobre los datos de prueba, se obtuvieron los resultados para la matriz de confusión mostrados en la tabla 12.

Tabla 12

Matriz de confusión: Modelo de Bosque Aleatorio

<table>
<tr><th colspan="2" rowspan="2">Matriz de confusión: Modelo 3</th><th colspan="4">Predicción</th></tr>
<tr><th colspan="2">0</th><th colspan="2">1</th></tr>
<tr><td rowspan="2">Estado Real</td><td>0</td><td>TN</td><td>706</td><td>FP</td><td>54</td></tr>
<tr><td>1</td><td>FN</td><td>97</td><td>TP</td><td>84</td></tr>
</table>

En comparación con las dos matrices de confusión anteriores, se puede decir que el modelo de Random Forest tiene una métrica de accuracy más alta tanto frente al modelo de regresión logística como el árbol de decisión. Esto se puede observar en el hecho de que la diagonal principal (los verdaderos positivos y verdaderos negativos) es más grande en la matriz de confusión del modelo de Random Forest en comparación con las otras dos matrices. Además, el número de falsos positivos y falsos negativos también es menor. En general, esto indica que el modelo de Random Forest está teniendo un mejor rendimiento que los otros dos modelos en términos de clasificar correctamente las cotizaciones en general, considerando tanto las ganadas como las perdidas.

4.10. Modelo 4 - Regresión logística modificado

Utilizando los resultados del modelo de árboles de decisión, en el cual uno para la clasificación entregada uno de los criterios significativos recae sobre el valor de la oferta, se procedió a crear una nueva variable cuyo objetivo es separar las ventas en dos grupos, uno con las ventas de alto valor y otro con las de menor valor, buscando mejorar los resultados.
Se crea un nuevo campo en el dataframe identificando las cotizaciones mayores a US$ 140.000.

Se probó la incidencia de esta categorización en el rendimiento de un nuevo modelo de regresión logística sin embargo los resultados obtenidos para las métricas no resultaron significativamente mejores a las obtenidas en el modelo de regresión logística inicial. Aumentó cerca de 1% el Accuracy pero se disminuyó el recall. los resultados de la matriz de confusión se muestran en la tabla 13.

Tabla 13

Matriz de confusión: Modelo de Regresión Logística modificada

Matriz de confusión: Modelo 4		Predicción			
		0		1	
Estado Real	0	TN	615	FP	157
	1	FN	37	TP	132

Como se observa en la matriz de confusión la configuración resultante es bastante similar a la obtenida en el modelo 1.

4.11. Modelo 5 - Máquina de soporte vectorial

Se realizó la prueba de un modelo SVM[20] (máquina de soporte vectorial). Este tipo de modelo es eficiente en problemas de alta dimensionalidad, sin embargo suele tener muy buen comportamiento en conjunto de datos pequeños si se consideran algunos aspectos como la selección del Kernel adecuado si los datos no son linealmente separables; la versatilidad de los kernels permite adaptarse a diferentes tipos de datos.

SVM puede ser un modelo complejo precisamente al usar kernels no lineales, puede ser sensible al sobreajuste y además requiere mayores recursos computacionales. Entre los parámetros establecidos para este modelo se tuvo en cuenta:

C: Es el parametro de regularización, Este controla la penalización de los errores de clasificación. Un valor grande de C indica que se permite un mayor margen de error en la clasificación, En

[20] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

cambio, un valor menor de C enfatiza una clasificación más precisa, pero a costa de un posible sobreajuste. Se utilizó un valor de 100 considerando una penalización moderada de los errores.

Kernel: El kernel en SVM es una función que transforma los datos de entrada en un espacio de mayor dimensión, donde es más probable que los datos sean linealmente separables.
Se optó por utilizar el kernel radial puesto que es adecuado para problemas que no son linealmente separables. por esto se ajustó el parámetro como “rbf” (Radial basis function).

Los resultados para las métricas de interés obtenidas para este modelo fueron mínimamente superiores en accuracy al comparar con modelo de regresión, pero resultaron inferiores en cuanto a recall. Estos resultados se pueden apreciar al revisar la matriz de confusión mostrada en la tabla 14, donde el número de falsos negativos mediante el modelo de SVM se mantiene encima del resultante en el modelo 1.

Tabla 14

Matriz de confusión: Modelo de Máquina de soporte vectorial

<table>
<tr><th colspan="2" rowspan="2">Matriz de confusión: Modelo 5</th><th colspan="4">Predicción</th></tr>
<tr><th colspan="2">0</th><th colspan="2">1</th></tr>
<tr><td rowspan="2">Estado Real</td><td>0</td><td>TN</td><td>632</td><td>FP</td><td>145</td></tr>
<tr><td>1</td><td>FN</td><td>44</td><td>TP</td><td>120</td></tr>
</table>

4.12. Comparación de resultados y selección del modelo

Dado el contexto en el que se aplicaron los anteriores modelos, donde la intención es enfocar mejor los recursos comerciales, es prioritariamente importante identificar con certeza aquellas cotizaciones con baja probabilidad de conversión en ventas, esto se traduce a nivel estratégico en la oportunidad de priorizar los esfuerzos de venta otras oportunidades de venta distintas a las clasificadas con baja probabilidad de conversión; Además se estaría dispuesto a tolerar algunos falsos positivos, es decir casos clasificados con probabilidad de conversión en venta pero que finalmente no se concretan. Llevando lo comentado anteriormente a términos de las métricas de los modelos, se consideraron importantes las siguientes:

El recall (sensibilidad o exhaustividad): es la proporción de verdaderos positivos que son correctamente identificados como tal. Se calcula de la siguiente manera:

Recall = TP / (TP + FN)

Donde TP es el número de verdaderos positivos (true positives) y FN es el número de falsos negativos (false negatives).

En el caso de la regresión logística, el cálculo del recall se realizó de la siguiente manera:

recall = 127 / (127 + 30) = 0.809

Es aceptable que el modelo tenga un recall alto y un accuracy más bajo, ya que esto significaría que se están identificando correctamente a los no compradores, es decir, dado que el recall indicaría la proporción de cotizaciones convertidas en ventas que el modelo logró identificar correctamente como positivas. Un alto valor de recall significa que el modelo está identificando la mayoría de las cotizaciones que se convierten en ventas.y aunque algunos casos clasificados como ventas pueden no serlo, la identificación temprana de aquellos que no se convertirán en ventas permitiría enfocar los esfuerzos del departamento comercial en otros clientes potenciales que tengan una mayor probabilidad de conversión.

Si el objetivo principal es minimizar los falsos negativos (es decir, minimizar el número de cotizaciones que se convierten en ventas pero que el modelo predijo que no lo harían), entonces el criterio de selección adecuado es el recal [11]l.

La especificidad: se refiere a la capacidad del modelo para identificar correctamente a aquellos clientes que no están interesados en comprar o cotizaciones que no se convierten en ventas. Es decir, se trata de la proporción de casos negativos reales que el modelo identifica correctamente como negativos en relación con el total de casos negativos reales. Se calcula de la siguiente manera:

Especificidad = TN / (TN + FP),

donde TN son los verdaderos negativos y FP son los falsos positivos.

En el caso de la regresión logística, el cálculo de Especificidad se calcula se realizó de la siguiente manera:

Especificidad = 615 / (615 + 169) = 0.784

La especificidad mide la capacidad del modelo para evitar falsos positivos y reducir el número de oportunidades perdidas. Por tanto proporciona respuesta al objetivo de maximizar la capacidad del modelo para identificar las cotizaciones que no se convertirán en ventas, es decir, maximizar los verdaderos negativos.

En la tabla 15 se muestran los resultados consolidados para la última corrida de los modelos.

Tabla 15

Resultados consolidados de las métricas obtenidas para los modelos creados

Modelo	Accuracy	Precision	Recall	F1
Regresión Logística	0.767	0.406	0.806	0.540
Árboles de Decisión	0.817	0.528	0.464	0.494
Random Forest	0.845	0.638	0.458	0.533
Regresión Logística Ajustado	0.787	0.402	0.767	0.528
Máquina de Soporte Vectorial	0.806	0.454	0.732	0.545

De acuerdo con los resultados finales obtenidos se observa que en general los modelos de Regresión logística presentan un nivel de accuracy un poco menor a los demás, sin embargo un comportamiento significativamente superior en términos de Recall, lo cual sumado a la simplicidad del modelo lo hace mucho más práctico para los propósitos del problema planteado.

Adicionalmente este modelo permite estimar los efectos marginales producidos por cambios en las variables independientes y de esta manera determinar la incidencia de los factores.

4.13. Ajuste de Hiper Parámetros

Los hiper parámetros son los parámetros intrínsecos de un modelo de Machine Learning. Estos tienen un impacto en el rendimiento y la precisión de cada modelo. En el caso práctico de este proyecto, luego de la evaluación de los modelos comparados en el **apartado 4.12** se realizó una búsqueda manual de aquellos hiper parámetros que permitieran alcanzar alguna mejora en el rendimiento del modelo de aprendizaje supervisado.

A continuación, se presenta la tabla 16 donde se consolidan los resultados para cada uno de los modelos con los hiper parámetros ajustados buscando obtener el mayor Recall.

Tabla 16

Resultados de las métricas obtenidas después del nuevo ajuste de hiperparametros

Modelo	Accuracy	Precision	Recall	F1
Regresión Logística	0.787	0.458	0.816	0.587
Árboles de Decisión	0.82	0.508	0.651	0.571
Random Forest	0.849	0.644	0.481	0.551
Regresión Logística Ajustado	0.765	0.427	0.793	0.556
Máquina de Soporte Vectorial	0.799	0.478	0.761	0.586

Los resultados presentados en la nueva corrida usando las configuraciones finales de hiper parámetros para cada modelo (tabla 16), son muy cercanos a los resultados presentados en la tabla 15. Esto debido a que ya había sido incorporada durante la creación de los modelos la selección de parámetros de acuerdo a las características del dataset; sin embargo, se puede observar unos resultados consistentes entre ambas ejecuciones de acuerdo a las métricas utilizadas. En este caso el modelo de regresión logística continuó mostrándose como el más adecuado de acuerdo a la estrategia y necesidad de la empresa, en cuanto a Recall continuó siendo el modelo con el valor más alto en 0,82. También se pudo observar un incremento en las demás métricas, como por ejemplo en F1 score donde se pasó de 0,54 a 0,58.

5. CATEGORIZACIÓN DE CLIENTES

Para enfocar de la mejor manera los esfuerzos del área de mercadeo y ventas de la organización, es de inmensa utilidad conocer y segmentar los clientes de acuerdo a diferentes características que inclusive pueden encontrarse escondidas en diferentes patrones dentro de los datos del CRM. Esto permite una mayor personalización de las estrategias comerciales, ofrecer un mejor servicio y de esta manera mejorar la satisfacción y rentabilidad.

En el proceso de identificar los perfiles con mayor potencial de tasa de conversión y retorno para la organización, se buscó organizar en grupos los clientes con características similares; para poder desarrollarlo se aprovechó el conocimiento del sistema CRM de la compañía y las etapas iniciales de análisis desarrolladas al trabajar en el primer objetivo de este proyecto como lo son la extracción de datos, tratamiento del dataset, análisis exploratorio y preprocesamiento. Pero con este nuevo propósito fue necesario, enfocarlo en la información de los clientes para lo cual se adiciono o filtró la información de trabajo considerando aquellas cotizaciones que en efecto habían sido convertidas en ventas reales

5.1. Preparación del dataset

Para este objetivo se utilizó el dataset procesado anteriormente, lo cual brindó valiosas ventajas permitiendo capitalizar el tiempo y los recursos invertidos en las etapas anteriores, aprovechar las lecciones aprendidas y la experiencia acumulada y se evitó repetir tareas de extracción y limpieza de datos. Además, al trabajar con un dataset previamente preparado, se aseguró de contar con información de calidad y coherencia, lo que garantiza resultados más confiables y precisos en los análisis y modelos.

5.1.1. Definición de los campos a utilizar

El proceso de segmentación requiere que la información entregada esté directamente relacionada con los clientes. El data set extraído contenía los datos de clientes y cotizaciones requeridos, sin embargo la manera en que se debía alimentar el proceso requería una organización diferente de los datos; por tanto, para tal efecto se acordó en acuerdo con expertos de la empresa que debía contener la información que se describe en la tabla 17:

Tabla 17

Descripción de las columnas utilizadas en la categorización de clientes

Columnas	Tipo	Descripción
rev	Continua	Número de Revisiones de la Oferta
cotizacion_id	Continua	Número único de cada cotización
cliente_id	Categórica	Codigo de Identificacion del cliente dentro de la BBDD
client_cat	Categórica	Clasificación del cliente por tamaño de organización. Ejemplo: Empresa Familiar
estado	Categórica	indica el estado final de la cotización. Ganada/Perdida/Descartada
precio_venta	Continua	Precio Total de Cotización en USD
Envio_year	Categórica	Año de envío de la cotización

Luego de la limpieza y depuración de los datos se cuenta con un dataframe con 7 columnas.

5.1.2. Agrupamiento de los datos

La información obtenida en el dataframe original se encontraba organizada a nivel de número de cotización, lo que significa que cada registro corresponde a una cotización diferente. Esta información es valiosa, sin embargo el agrupamiento por cotización no es acorde para el proceso de categorización de clientes. Este proceso requiere que la información esté agrupada a nivel de cliente. Para lograrlo, se agrupó la información a nivel de cliente en python.

Al realizar un agrupamiento, se debe decidir qué operación de agregación aplicar a las demás columnas. Generalmente las funciones de agrupamiento suman el valor de una columna o toman el valor más alto o más bajo. En la tabla 18 se presenta la operación de agregación aplicada a cada columna y la información contenida

Tabla 18

Descripción de agregación aplicada para cada columna del dataframe

Columnas	Tipo Agregación	Descripción
client_cat	Máximo	Al ser categórica tomará el valor máximo
rev	Media	Toma el valor correspondiente a la media
tot_cotizaciones	Conteo	Número de cotizaciones por cliente
tot_cotizaciones_gan	Conteo	Número de cotizaciones ganadas por cliente
amnt_cotizaciones	Media	Calcula la media del valor de todas las cotizaciones por cliente
amnt_cotizaciones_gan	Media	Calcula la media del valor de todas las cotizaciones por cliente
envio_year	Mínimo	Corresponde al año de la primera cotización

5.2. Modelo K-Medias

En la identificación de los perfiles de los clientes, se utilizó un modelo de agrupamiento o Clustering. Este tipo de modelos permite identificar y agrupar a los clientes acorde con sus similitudes. El número de grupos o clusters se define con anterioridad a la ejecución del algoritmo, utilizando varias técnicas entre ellas la técnica del codo. Cabe resaltar que el proceso de clustering es un proceso no supervisado, lo que significa que no es necesario generar un conjunto de datos de entrenamiento y prueba.

Para realizar la función de agrupamiento no supervisado se decide utilizar el algoritmo de k medias[21] (k means) [7],[8],[13], el cual es ampliamente reconocido y utilizado en el campo del análisis de datos y la minería de patrones debido a su simplicidad y eficacia. Mediante este algoritmo se busca encontrar una configuración óptima de centroides que minimice las distancias dentro de cada grupo y maximice las distancias entre grupos.

Como punto de partida se requiere especificar el número de grupos deseados como entrada. En la etapa inicial, el algoritmo selecciona de forma aleatoria las coordenadas de los centroides para cada grupo. A continuación, calcula la distancia entre cada muestra y cada centroide, y asigna los puntos al centroide con la distancia más corta. Este proceso se repite varias veces, ajustando el valor de los centroides en cada iteración al calcular la media de las distancias de los puntos asignados a cada centroide. Este ciclo continúa hasta que los centroides dejen de cambiar o se alcance el número máximo de iteraciones establecido.

5.2.1. Estimación de hiperparametros

La estimación de hiperparámetros es un paso crucial en el modelado de k-means, ya que permite determinar los valores óptimos de los parámetros que afectan el rendimiento y la calidad de los resultados. Dos hiperparámetros clave a considerar son el número de grupos (k) y el estado aleatorio (random state).

Estimación del valor óptimo para el hiper parámetro K

La selección del número de grupos (k) es fundamental, ya que determina la cantidad de clusters en los que se dividirán los datos. Para una mejor operación del algoritmo de k medias, es necesario calcular antes el valor más adecuado del hiper parámetro K, para ello se utilizó el método del

[21] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

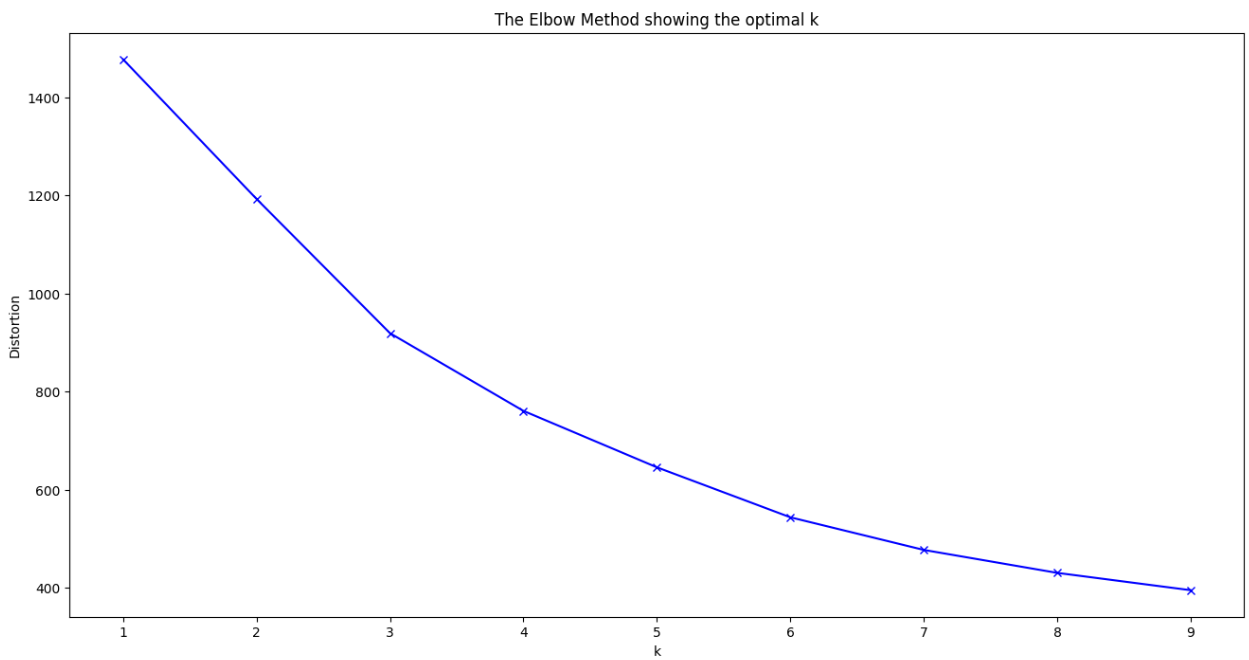
codo (elbow) .

Este método consiste en plasmar en una gráfica la distorsión del modelo y el valor de K en los ejes Y y X respectivamente. La distorsión representa la suma de las distancias al cuadrado entre cada punto de datos y el centroide de su cluster correspondiente.

La selección del mejor valor de K, corresponde a la identificación del punto a partir del cual el cambio en la distorsión se reduce significativamente, lo cual sugiere que ya no se mejora sustancialmente la calidad del modelo si se continúa aumentando el número de clusters K.

Figura 9

Gráfica del codo para el modelo K-medias



Nota: *En la gráfica se tiene la distorsión en función del número de clusters, lo cual permite visualmente identificar el valor óptimo del parámetro k*

Para el caso de categorización de clientes y basándose en la gráfica del codo, se puede identificar que el valor correcto de K es 3. lo que significa que el algoritmo de K medias agrupa a los clientes en 3 grupos.

Definición del valor del hiper parámetro random state

El estado aleatorio (random state) es importante para garantizar la reproducibilidad de los resultados; el hiper parámetro random state es el valor de inicialización de los centroides. Como se desea que el algoritmo K medias genere resultados consistentes en las diferentes ejecuciones del modelo se decide asignar un valor fijo al hiper parámetro random state. Para el proceso de categorización de clientes se utilizó el valor random state: 42.

Método de inicialización

El parámetro 'init' en el algoritmo K-Means se relaciona con la inicialización de los centroides. Una de las opciones disponibles es 'k-means++', la cual se ha demostrado que proporciona una mejor inicialización en comparación con otras alternativas. 'k-means++' selecciona de manera inteligente los centroides iniciales distribuyéndolos de forma equilibrada en función de la distancia de las muestras. Esta estrategia reduce la probabilidad de una inicialización deficiente que pueda llevar a soluciones subóptimas o a una convergencia lenta

Una vez definidos los hiperparametros se desarrolló un modelo de K medias ajustado al dataset de clientes, utilizando un valor fijo de semilla (Random_state = 42), una inicialización mejorada (init = k-means++) y buscando clasificar los clientes en 3 grupos (n_clusters = 3)

5.3. Resultados

Después de ajustar el modelo, se tomó como referencia el valor medio obtenido para cada uno de los segmentos de clientes. los resultados se consolidan en la tabla 19.

Tabla 19

Media de cada uno de los campos por segmento de clientes

Segmento	Revisión	Total Cotizaciones	Total Cotizaciones Ganadas	Valor Cotizaciones USD	Valor Cotizaciones Ganadas USD	Fecha Primera Cotización
0	0.76	7.96	3.18	$92,936.83	$18,326.57	2018.2
1	2.31	3.11	2.31	$165,921.54	$99,101.98	2020.6
2	0.63	206.00	92.00	$27,506.02	$5,674.39	2017.0

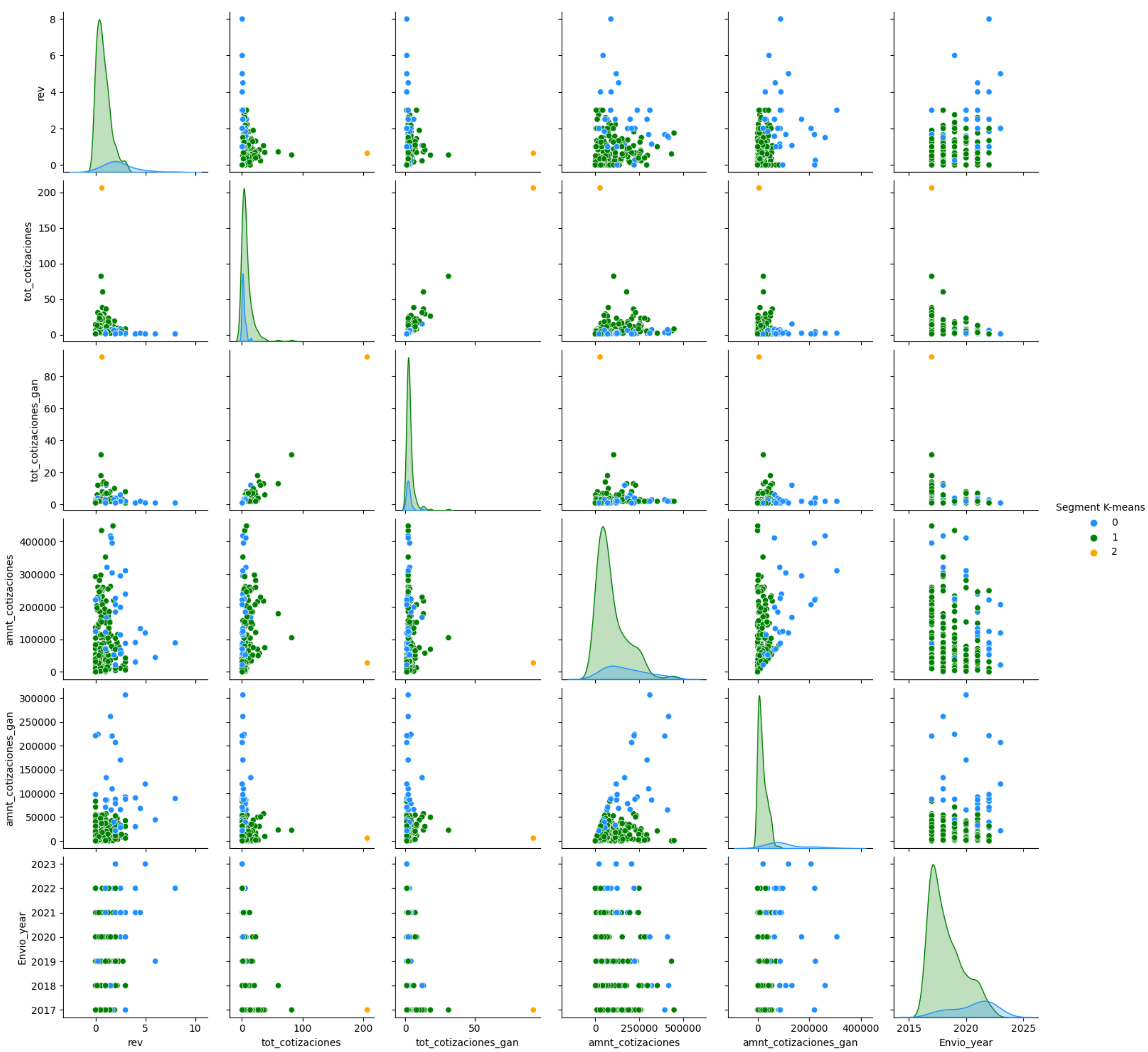
De los resultados fue posible identificar 3 grupos de clientes, los cuales se pueden identificar de la siguiente forma:

1) El primer grupo está compuesto por clientes más antiguos, puesto que el promedio del año de su primera cotizacion es 2018, estos clientes la tasa de conversión es del 40% es decir, de 10 cotizaciones se convierten 4 en promedio. El valor promedio de las cotizaciones y las compras es inferior a los demás grupos, tratándose mayoritariamente de venta de repuestos.
2) El segundo grupo corresponde a clientes donde su primera cotización es más reciente y donde su valor promedio de cotizaciones y compras es mucho más alto, En este grupo se pueden encontrar clientes principalmente enfocados en la compra de plantas o proyectos de producción compuestas por varios productos. Por esta razón el promedio de revisiones en las cotizaciones es significativamente superior a los demás grupos. La tasa de conversión para este grupo de clientes es cercana al 80%.
3) El tercer grupo lo conforman clientes con un alto número de cotizaciones, este número al ser tan alto llevó al equipo de trabajo a investigar y se encontró que todas las cotizaciones dentro de este grupo corresponden a un único cliente, este cliente es un socio comercial de la compañía. El valor de estas cotizaciones es significativamente inferior a los demás grupos, además presenta el menor número de revisiones. Al analizar la fecha de cotizaciones que resultan ser las más antiguas se encuentra una relación comercial consolidada desde 2017 y con una tasa de conversión de 45%.

Considerando que al trabajar con múltiples variables resulta complejo realizar una visualización de la salida del modelo, se utilizó la función pairplot para crear gráficos de pares de variables que se pueden apreciar en la figura 10.

Figura 10

Relación entre las variables utilizadas en el modelo k medias.



En el gráfico es posible identificar por cada pareja de variables, la salida del modelo de K medias que se encargó de generar la segmentación de los clientes de la compañía. En cada uno se presentan con diferentes colores los segmentos o grupos generados y su relación entre dos variables.

De esta observación se puede concluir que el monto de cotizaciones ganadas (Amt_cotizaciones_gan) es un factor determinante puesto que en cada uno de los gráficos donde se relaciona esta variable es posible visualizar una frontera de separación más marcada entre los diferentes segmentos de clientes.

6. DESARROLLO DE HERRAMIENTA DE VISUALIZACIÓN

6.1. Recopilación de requisitos

En conversación con las directivas del área comercial de la empresa se identificó la necesidad de reemplazar los reportes de "Pipeline" y ventas por dashboards que presenten la información en tiempo real [12]. Como expectativas mínimas, se deben incluir los siguientes elementos:

Embudo de Ventas - "Pipeline":

- Presentar la cantidad de ofertas en negociación.
- Dividir las ofertas en negociación por el estado de avance en el pipeline.
- Representar gráficamente la participación de cada vendedor en el pipeline.
- Presentar tablas resumen con ofertas en estados de negociación más avanzados.

Informe de Ventas Anual

- Total de ofertas negociadas (vendidas).
- Representar gráficamente las ofertas negociadas por mes.
- Hacer seguimiento al objetivo anual de ventas (USD 12.000.000).
- Hacer seguimiento a la tasa de conversión (KPI: 20%).
- Representar gráficamente la participación de cada vendedor en el total de negociados.
- Identificar en qué países hubo ventas durante el año en estudio.
- Analizar cómo se distribuyen las ventas por línea de negocio.

6.2. Fuentes de datos y recursos utilizados

Los datos utilizados en el dashboard fueron extraídos en su totalidad de la BBDD del CRM de la empresa. Sin embargo, no fueron empleadas todas las variables usadas durante el entrenamiento de los modelos de aprendizaje supervisado y no supervisado descritos anteriormente.
Para el desarrollo de las consultas y desarrollo de visualización se utilizaron los recursos disponibles actualmente por la empresa.

MySQL:
Se realizaron consultas mediante MySQL en la base de datos del CRM (Proveedor externo de la empresa) utilizando unas credenciales de sólo visualización autorizadas por la empresa. Los resultados de estas consultas fueron exportados a archivos CSV.

Power BI / Power Query:
Los archivos CSV fueron importados en el software de análisis de datos Power BI para construir las visualizaciones acordes a los requisitos de la compañía. Mediante Power Query, se automatizan tareas de limpieza y transformación de datos que faciliten el desarrollo del dashboard.

6.3. Desarrollo del dashboard

6.3.1. Extracción de tablas de la Base de datos:

- Tabla clientes (dim_1):

Tabla 20

Definicion dimension Clientes

Columnas	Tipo	Descripción
id	Categórica	ID de cliente en BBDD
Nombre Cliente	Categórica	Razon Social del Cliente

- Tabla users (dim_2):

Tabla 21

Definición dimension usuarios

Columnas	Tipo	Descripción
id	Categórica	ID de usuario en BBDD
name	Categórica	Nombre del usuario en la BBDD (Nombre vendedor)
cargo	Categórica	Cargo del usuario en la organización

- Tabla cotizaciones (facts)

Tabla 22

Definición tabla de hechos

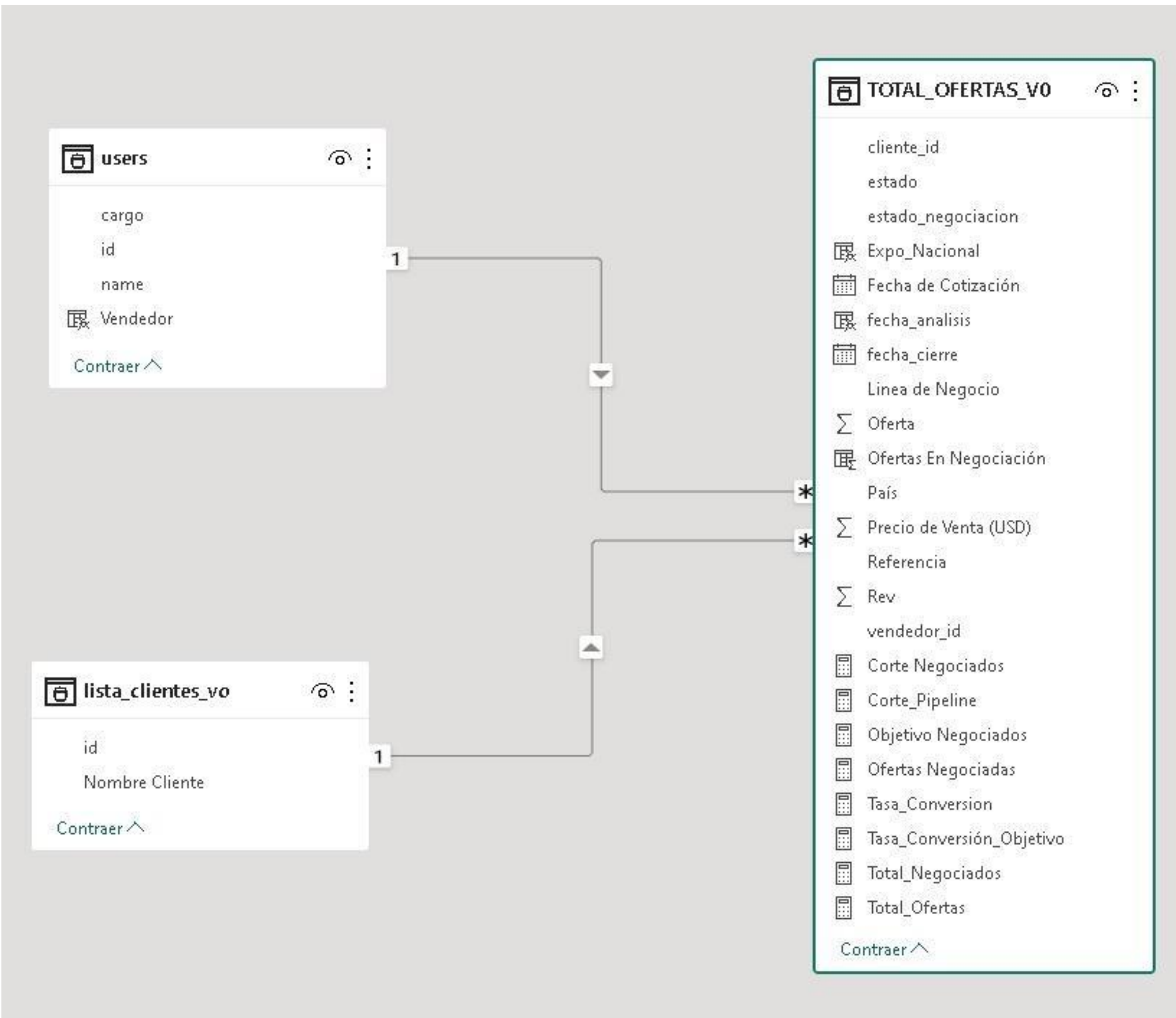
Columnas	Tipo	Descripción
consecutivo	Categórica	Número de consecutivo de oferta
Rev	Categórica	Número de Revisión de la Oferta
cliente_id	Categórica	Código de Identificación del cliente dentro de la BBDD
referencia	Categórica	Descripción breve del objeto de la cotización
fecha_envio	Categórica	Fecha de envío de oferta al cliente
fecha_cierre	Categórica	Fecha de cierre de la venta
estado	Categórica	Estado de la oferta (Ganada, Perdida, En Negociación, Descartada)
estado_negociacion	Categórica	Estado de la oferta en pipeline (Avance en negociación)
pais	Categórica	País destino de la venta
precio_venta	Continua	Precio de venta de la oferta (USD)
cliente_id	Categórica	ID de cliente en tabla lista_clientes_v0
vendedor_id	Categórica	ID de vendedor en tabla users

6.3.2. Creación de medidas y columnas nuevas

- Ofertas negociadas: Número total de ofertas vendidas en el periodo de tiempo a analizar.
- Total_Ofertas: Número total de ofertas generadas en el periodo de tiempo a analizar.
- Tasa_Conversion: Ofertas negociadas entre Total_Ofertas.
- Total_Negociados: Total de ventas en USD durante el periodo de tiempo a analizar.

Figura 11

Modelo de la bodega de datos



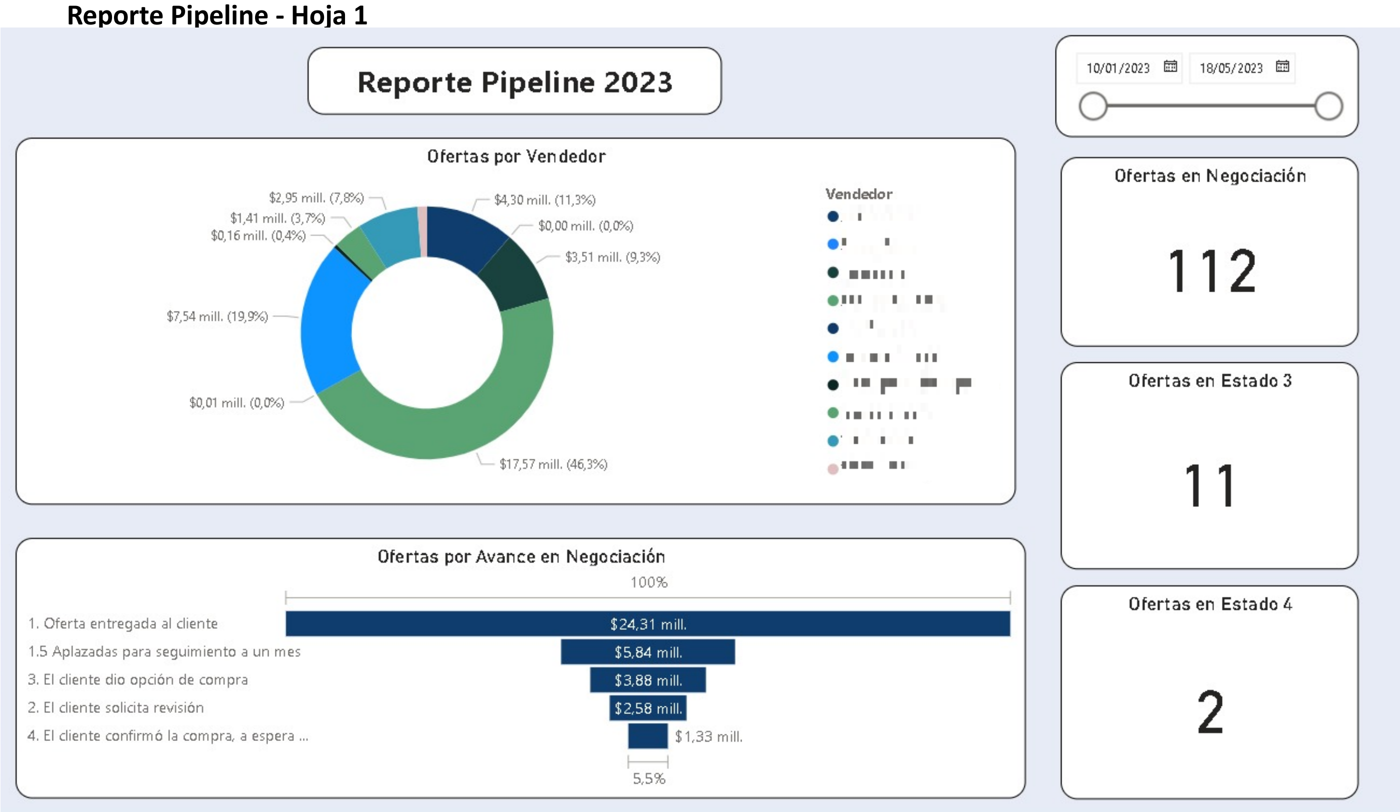
Nota: Esquema de relaciones entre tablas de dashboard

6.4. Dashboard finalizado:

Reporte pipeline

Figura 12

Reporte Pipeline - Hoja 1



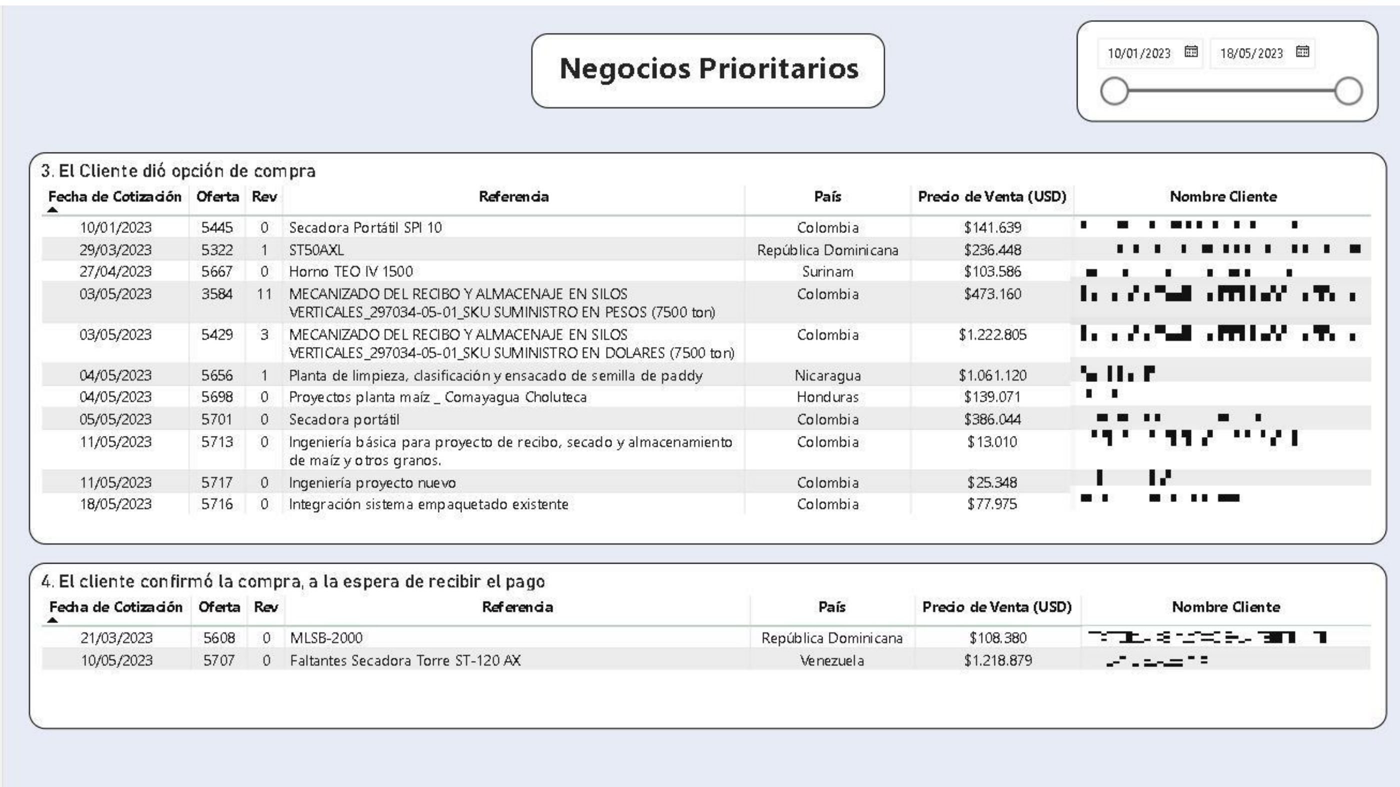
La primera página del dashboard presenta la información general del Pipeline, destacando los siguientes puntos:

1. Gráfico de anillos que presenta la suma total de ofertas (USD) agrupado por cada vendedor.
2. Gráfico de embudo que muestra la suma total de ofertas (USD) agrupado por estado de negociación.
3. Tarjeta que indica el número total de ofertas en negociación.
4. Tarjeta que indica el número total de ofertas en estado 3.

5. Tarjeta que indica el número total de ofertas en estado 4.
6. Se incluye un segmentador de datos que permite filtrar las visualizaciones por fecha de envío de oferta.

Figura 13

Reporte Pipeline - Hoja 2



Fecha de Cotización	Oferta	Rev	Referencia	País	Precio de Venta (USD)	Nombre Cliente
10/01/2023	5445	0	Secadora Portátil SPI 10	Colombia	$141.639	[redacted]
29/03/2023	5322	1	ST50AXL	República Dominicana	$236.448	[redacted]
27/04/2023	5667	0	Horno TEO IV 1500	Surinam	$103.586	[redacted]
03/05/2023	3584	11	MECANIZADO DEL RECIBO Y ALMACENAJE EN SILOS VERTICALES_297034-05-01_SKU SUMINISTRO EN PESOS (7500 ton)	Colombia	$473.160	[redacted]
03/05/2023	5429	3	MECANIZADO DEL RECIBO Y ALMACENAJE EN SILOS VERTICALES_297034-05-01_SKU SUMINISTRO EN DOLARES (7500 ton)	Colombia	$1.222.805	[redacted]
04/05/2023	5656	1	Planta de limpieza, clasificación y ensacado de semilla de paddy	Nicaragua	$1.061.120	[redacted]
04/05/2023	5698	0	Proyectos planta maíz _ Comayagua Choluteca	Honduras	$139.071	[redacted]
05/05/2023	5701	0	Secadora portátil	Colombia	$386.044	[redacted]
11/05/2023	5713	0	Ingeniería básica para proyecto de recibo, secado y almacenamiento de maíz y otros granos.	Colombia	$13.010	[redacted]
11/05/2023	5717	0	Ingeniería proyecto nuevo	Colombia	$25.348	[redacted]
18/05/2023	5716	0	Integración sistema empaquetado existente	Colombia	$77.975	[redacted]

Fecha de Cotización	Oferta	Rev	Referencia	País	Precio de Venta (USD)	Nombre Cliente
21/03/2023	5608	0	MLSB-2000	República Dominicana	$108.380	[redacted]
10/05/2023	5707	0	Faltantes Secadora Torre ST-120 AX	Venezuela	$1.218.879	[redacted]

La segunda página del dashboard presenta una mirada más a detalle a aquellas ofertas consideradas como prioritarias, ya que se encuentran en los estados de negociación más avanzados, los elementos que la componen son:

1. Tabla indicando las ofertas en estado de negociación 3 (El cliente dió opción de compra).
2. Tabla indicando las ofertas en estado de negociación 4 (El cliente confirmó la compra, a la espera de recibir el pago).

3. Se incluyó un segmentador de datos que permite filtrar las visualizaciones por fecha de envío de oferta.

Figura 14

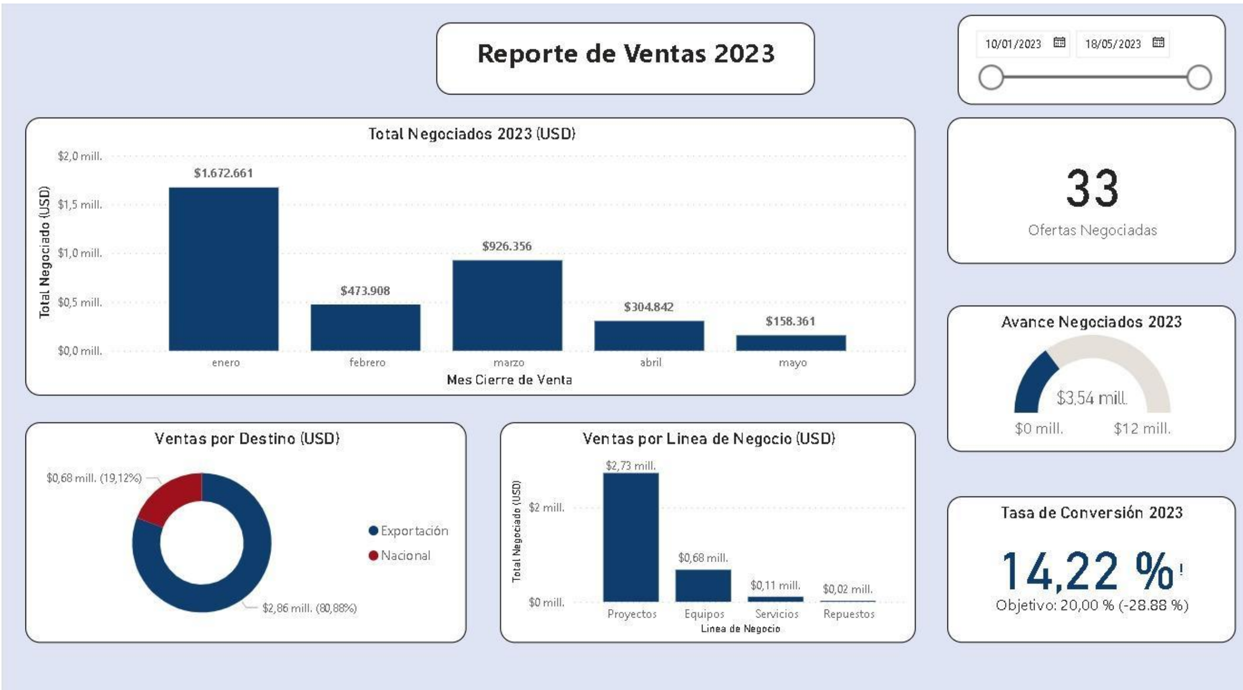
Reporte Pipeline - Hoja 3

La tercera página del dashboard, y última con información del pipeline, incluye una gráfica tipo mapa interactiva, que permite observar la ubicación de las ofertas en negociación. Donde el tamaño de la burbuja corresponde a la suma total de ofertas (USD).

Reporte Ventas

Figura 15

Reporte Ventas - Hoja 1



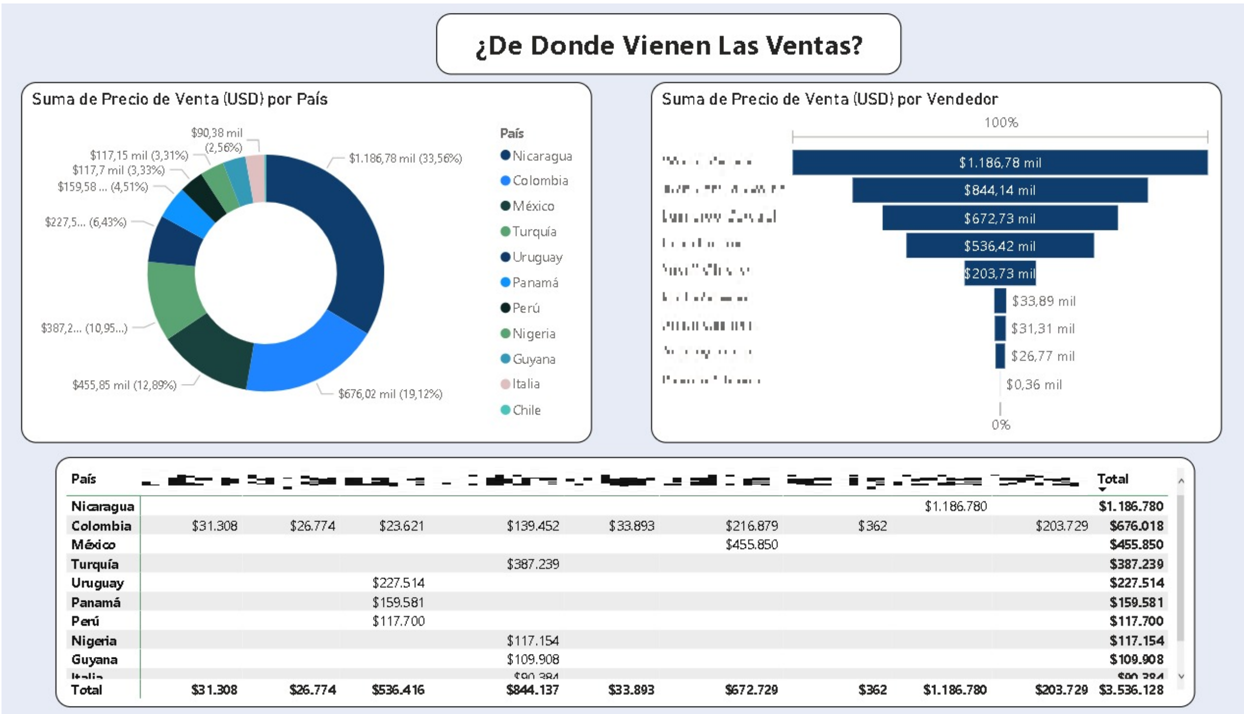
La cuarta hoja del pipeline muestra la primera parte del reporte de ventas, donde se presentan los resultados (resumen de ventas) en el año 2023. Se incluyen los siguientes elementos en esta parte del dashboard:

1. Gráfico de barras que muestra el total de ventas por mes (USD).
2. Gráfico de anillos que muestra el total de ventas (USD) categorizando las ventas en nacionales y exportaciones.
3. Gráfico de barras que muestra el total de ventas (USD) por línea de negocio.
4. Tarjeta que muestra el número de ofertas negociadas.
5. Medidor que muestra el avance del total de ventas anuales con respecto al objetivo anual.
6. Tasa de conversión (KPI) medido a lo largo del año 2023.

7. Se incluye un segmentador de datos que permite filtrar las visualizaciones por fecha de cierre de venta.

Figura 16

Reporte Ventas - Hoja 2



Por último, se incluye una página que tiene como objetivo presentar el origen de las ventas (país origen del cliente y vendedor responsable). Se incluyen los siguientes elementos en esta parte del dashboard:

1. Gráfica de anillos que muestra el total de ventas por país.
2. Gráfico de embudo que muestra la suma total de ofertas (USD) agrupado por vendedor responsable.
3. Matriz que resume el total de ventas (USD) por país (columnas) y vendedor (filas).

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

El modelo con mejores resultados permite pronosticar el estado final de una oferta con una exactitud (accuracy) del 78%. Con respecto al Recall (Tasa de Verdaderos Positivos), este modelo alcanza un valor de 80.6%. Teniendo en cuenta la importancia que es para el área de ventas identificar aquellas ofertas con alta probabilidad de venta (cierre) estas se consideran nuestra clase de interés.

Se evidenció que la variable ***"rev"*** (cantidad de revisiones de la oferta) tienen un impacto en la probabilidad de cierre de una venta. Esto fue validado a través de los coeficientes del modelo de regresión logística y la visualización del árbol de decisión. Este hallazgo corresponde con la realidad del negocio, donde se priorizan las revisiones de oferta ya que no son actividades con alta necesidad de recurso. Sin embargo, se deben tomar medidas para que en la base de datos el registro de cada oferta cerrada mantenga el número de la revisión vendida y evitar que se hagan revisiones posteriores a esta venta que alteren este indicador.

La variable de fuente externa **"GDP",** asociada al crecimiento económico del país de la cotización en dicho año, no agregó suficiente valor al modelo y no cumplió significativamente con el propósito inicial de complementar la base de datos de ventas con un indicador económico relacionado al país correspondiente a cada oferta (registro). Esta variable a pesar de mostrar una relación positiva con la probabilidad de conversión en ventas, como se podía intuir, tuvo el menor impacto dentro del modelo de regresión logística.

La segmentación de los perfiles de clientes giró en torno a dos principales variables: precio de venta y volumen de ofertas. Inicialmente se esperaba obtener una visión más detallada de estos clusters de clientes, sin embargo, la poca información que realmente era útil dentro de la BBDD no permitió desarrollar con más profundidad este objetivo. Por ejemplo, los datos como tamaño del cliente (número de empleados, ingresos) y sector del cliente, no habían sido correctamente incluidos en el CRM para permitir un análisis al detalle.

La implementación del modelo desarrollado para determinar la probabilidad de conversión en ventas de las oportunidades presentadas, ayudará a la compañía a enfocar los limitados recursos del área comercial hacia las oportunidades con mayor potencial. Considerando que no existía actualmente en la empresa, un criterio basado en datos para tomar este tipo de decisiones; en adelante será posible confrontar los resultados obtenidos mediante el modelo y medir los cambios en los indicadores de eficiencia del área comercial. La priorización de recursos permitiría incrementar las metas de conversión actuales al tomar decisiones de mejora en los tiempos de respuesta y la atención al cliente, lo cual posibilitará un aumento del volumen de ventas.

Con el desarrollo del dashboard para visualización de pipeline y reporte de ventas se dota a los directivos del área comercial con una herramienta que les permite hacer seguimiento en tiempo real de las ventas y de aquellos indicadores que consideren claves en su negocio. Por otro lado, la actualización en vivo de estos datos disminuye el tiempo que hasta el momento invierte el personal del área comercial en la construcción de informes periódicamente o bajo pedido de los stakeholders. El cálculo en tiempo real de la tasa de conversión es un paso importante en el monitoreo y evaluación permanente del proceso comercial, el cual se deberá seguir complementando con indicadores claves de desempeño como longitud del ciclo de ventas e incluir dentro del análisis las oportunidades o “leads” que se identifican periódicamente.

7.2. Trabajos futuros

1. Se harán esfuerzos para tener conexión directa con la BBDD y hacer consultas SQL directamente desde el Jupyter Notebook para automatizar el proceso de extracción de datos. Esta conexión directa también es necesaria para permitir la actualización en tiempo real del dashboard al permitir la conexión con Power BI. Se recomienda al departamento comercial de la empresa gestionar un acceso autorizado a la base de datos con los permisos requeridos con su proveedor externo.

2. Se buscará implementar el modelo más preciso en un web service para que reciba la información de una nueva cotización y que este retorne la probabilidad de conversión de venta.

3. Se debe implementar el webservice mencionado en el punto # 2 en el dashboard en desarrollo para predecir el porcentaje de cierre de ofertas en negociación. Esto permitirá tener con mayor certidumbre un valor esperado de ventas y se podrían tomar decisiones estratégicas de qué negocios deben ser prioritarios.

4. Se debe implementar el modelo con el CRM para predecir el porcentaje de cierre desde las ofertas en construcción. También se debe plantear la necesidad de incluir estos resultados en el dashboard (Reporte Pipeline) para tener una proyección de ventas teniendo en cuenta las ofertas en los distintos estados de negociación.

5. Se recomienda generar implementar buenas prácticas de manejo y almacenamiento de la información dentro del CRM resaltando el valor agregado y potencial que tiene su análisis y uso para modelos predictivos y de clasificación.

6. Se deberá hacer pruebas del modelo utilizando las ofertas generadas desde la última extracción de datos (Febrero 2023).

8. REFERENCIAS BIBLIOGRÁFICAS

[1] T. Grapsas, «Ventas B2B: ¿cómo funcionan y qué procesos tienen?,» rockcontent, 22 Diciembre 2018. [En línea]. Available: https://rockcontent.com/es/blog/ventas-b2b/. [Último acceso: 4 Julio 2022].

[2] DemandJump, «B2B Sales Forecasting Methods,» 24 Agosto 2020. [En línea]. Available: https://www.demandjump.com/blog/b2b-sales-forecasting-methods. [Último acceso: 4 Julio 2022].

[3] Collective[i], «B2B sales forecasting,» [En línea]. Available: https://collectivei.com/b2b-sales-forecasting/#:~:text=Sales%20forecasting%20is%20the%20process,market%20research%20and%20other%20insights. [Último acceso: 4 Agosto 2022].

[4] Pipedrive, «Tres tipos de KPIs de ventas y tableros para mejorar tus ingresos,» [En línea]. Available: https://www.pipedrive.com/es/blog/kpis-ventas. [Último acceso: 4 Julio 2022].

[5] D. Rohaan, E. Topan y C. Groothuis-Oudshoorn, «Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider,» *Expert Systems with Applications,* vol. 188, 2022.

[6] K. M. Kasinathan, «Infer - A Full Scale B2B Sales Predictor,» 17 Agosto 2021. [En línea]. Available: https://staff.fnwi.uva.nl/a.s.z.belloum/MSctheses/MSchesis_Muthiah-kasinathan.pdf. [Último acceso: 4 Julio 2022].

[7] Lloyd, Stuart P. (1957). "Least square quantization in PCM". *Bell Telephone Laboratories Paper*. Published in journal much later: Lloyd, Stuart P. (1982). "Least squares quantization in PCM" (PDF). *IEEE Transactions on Information Theory*. **28** (2): 129–137. CiteSeerX 10.1.1.131.1338. doi:10.1109/TIT.1982.1056489. S2CID 10833328. Retrieved 2009-04-15.

[8] Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications". *Biometrics*. **21** (3): 768–769. JSTOR 2528559.

[9] Cooper, H.B., Ewing, M.T. and Mishra, S. (2022) Text-mining 10-K (Annual) reports: A guide for B2B marketing research, Industrial Marketing Management. Elsevier. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0019850122002334 (Accessed: October 21, 2022).

[10] Marvasti, N.B. et al. (2021) Is this company a lead customer? estimating stages of B2B buying journey, Industrial Marketing Management. Elsevier. Available at:

https://www.sciencedirect.com/science/article/pii/S001985012100105X (Accessed: October 21, 2022).

[11] Rohaan, D., Topan, E. and Groothuis-Oudshoorn, C.G.M. (2021) Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider, Expert Systems with Applications. Pergamon. Available at: https://www.sciencedirect.com/science/article/pii/S0957417421012793 (Accessed: October 21, 2022).

[12] Rusthollkarhu, S. et al. (2022) Managing B2B customer journeys in digital era: Four management activities with artificial intelligence-empowered tools, Industrial Marketing Management. Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S0019850122000888 (Accessed: October 21, 2022).

[13] Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 135–139. Available at: https://ieeexplore.ieee.org/abstract/document/8769171