



**UNIVERSIDAD MAYOR DE SAN SIMÓN**  
**FACULTAD DE CIENCIAS Y TECNOLOGÍA**



**DIRECCIÓN DE POSGRADO**

# **DIPLOMADO ESTADÍSTICA APLICADA A LA TOMA DE DECISIONES**

**TERCERA VERSIÓN**

## **ANÁLISIS EXPLORATORIO DE DATOS Y TENDENCIAS DE UNA COMERCIALIZADORA DE HERRAMIENTAS USANDO BUSINESS INTELLIGENCE**

**PROYECTO PRESENTADO PARA OBTENER EL GRADO DE LICENCIATURA EN  
INGENIERÍA DE SISTEMAS.**

**MODALIDAD DOBLE TITULACIÓN**

**POSTULANTE : CARLOS ALFREDO ORIHUELA BERRIOS**

**TUTOR : M.SC. ING. JORGE GERARDO ZAMBRANA VILAR**

**Cochabamba – Bolivia**

**2024**







# ANÁLISIS EXPLORATORIO DE DATOS Y TENDENCIAS DE UNA COMERCIALIZADORA DE HERRAMIENTAS USANDO BUSINESS INTELLIGENCE

Por

Carlos Alfredo Orihuela Berrios

El presente documento, Trabajo de Grado es presentado a la Dirección de Posgrado de la Facultad de Ciencias y Tecnología en cumplimiento parcial de los requisitos para la obtención del grado académico de Licenciatura (o sólo diplomado) en Ingeniería de Sistemas, modalidad Doble Titulación, habiendo cursado el Diplomado “Estadística Aplicada a la Toma de Decisiones” propuesta por el Centro de Estadística Aplicada (CESA) en su tercera versión.

ASESOR/TUTOR

M.Sc. Ing. JORGE GERARDO ZAMBRANA VILAR

COMITÉ DE EVALUACIÓN

Ing. M.Sc. Ronald Edgar Patiño Tito. (Presidente)

Ing. M.Sc. Guillen Salvador Roxana, (Coordinador)

Ing. M.Sc. Valentín Laime Zapata (Tribunal)

Ing. M.Sc. Jose Espinoza Orosco (Tribunal)



**DIRECCIÓN DE POSGRADO, FACULTAD DE CIENCIAS Y TECNOLOGIA**

Cochabamba, Bolivia

### **Aclaración**

**Este documento describe el trabajo realizado como parte del programa de estudios de Diplomado “Estadística Aplicada a la Toma de Decisiones” en el Centro de Estadística Aplicada CESA y la Dirección de Posgrado de la Facultad de Ciencias y Tecnología. Todos los puntos de vista y opiniones expresadas en el mismo son responsabilidad exclusiva del autor y no representan necesariamente las de la institución.**

---

## Resumen

---

En el contexto del proyecto "Análisis exploratorio de datos y tendencias de una comercializadora de herramientas usando Business Intelligence", se ha llevado a cabo un análisis exhaustivo del comportamiento del equipo del área de ventas de la empresa MASTER TOOLS. Basado en técnicas avanzadas de Business Intelligence y Machine Learning, tiene como objetivo optimizar el rendimiento y objetivos comerciales del área de ventas en dicha empresa.

La metodología empleada en el presente proyecto con un enfoque de diseño dimensional propuesto por Ralph Kimball para el desarrollo de la solución de Business Intelligence (BI). Adicionalmente, se implementaron modelos de Machine Learning, específicamente Regresión Logística y Árbol de Decisiones, los cuales alcanzaron una validación de 0.85 y 0.88, respectivamente, en la predicción del comportamiento de las ventas. El proceso metodológico abarcó desde la definición de requerimientos hasta el desarrollo de dashboards interactivos, facilitando la integración y el análisis efectivo de los datos.

La visualizaciones se proporcionó con la herramienta de Power BI identificando las tendencias, como ser la marca Lynus con mayor preferencia entre los clientes y la máquina de pintura sin aire lynus 1000w como su producto estrella, también con una participación superior de Santa Cruz (39,84%) y Cochabamba (32,84%) registraron los mayores ingresos consolidándose como las de mayor rendimiento comercial de todas las sucursales. Adicionalmente, se identificó un pico estacional en ventas durante el segundo trimestre (abril, mayo y junio).

En síntesis, el proyecto se enfocó en el análisis de modelos de Machine Learning para la predicción de tendencias clave con el objetivo de optimizar los ingresos por ventas. Se evaluaron modelos como Regresión Logística y Random Forest, utilizando variables relevantes como 'Precio de Venta', 'Marca' y 'Sucursal'. Los resultados fueron validados mediante métricas de rendimiento, incluyendo accuracy, F1-Score y la matriz de confusión, garantizando así una evaluación precisa del desempeño de los modelos

Esta ficha resumen proporciona una visión integral de los capítulos clave del proyecto, resaltando su estructura y los hallazgos más significativos.

### Palabras clave

Análisis de datos, Business Intelligence, Tendencias, Dashboard, Indicadores clave de rendimiento, toma de decisiones

*Dedicatoria a mis padres Francisco Oribuela y Elizabeth Berrios;*  
*Por ser el pilar fundamental en todo lo que soy, en toda mi educación, tanto*  
*académica, como de la vida, por su incondicional apoyo en todo el trayecto de*  
*mi carrera profesional.*

*A mis hermanos Jorge, Brandon y Dana;*  
*Por su cariño y compañía dentro del hogar, por todos los buenos momentos y*  
*las varias alegrías que compartimos.*



# Agradecimientos

---

*A Dios por toda la fuerza, compañía y haberme permitido llegar hasta este punto, por haberme dado salud para lograr mis objetivos.*

*A mis queridos docentes, por compartir sus conocimientos, orientación y experiencias;*

*Y a todos mis amigos que me ayudaron y apoyaron a lo largo de mi carrera universitaria.*

*¡Muchas Gracias!*

# Tabla de contenidos

---

## Contenido

1.	Introducción.....	1
1.1.	Antecedentes .....	1
1.2.	Justificación.....	2
1.2.1	Justificación Social.....	2
1.2.2	Justificación Tecnológica .....	2
1.2.3	Justificación Económica.....	3
1.3.	Planteamiento del problema .....	3
1.3.1	Formulación del problema.....	4
1.4.	Objetivo general .....	5
1.4.1	Objetivos específicos.....	5
2.	Marco teórico .....	6
2.1.	Recolección de Información.....	6
2.1.1	Ventajas y Desventajas de la recolección de información .....	6
2.1.2	Técnicas recolección de información .....	7
2.2.	Análisis exploratorio de Datos .....	7
2.2.1	Importancia del Análisis exploratorio de Datos en el Comercio .....	8
2.2.2	Herramienta para el análisis exploratorio de datos.....	8
2.3	Identificación de Tendencias.....	10
2.3.1	Indicadores de rendimiento y métricas de ventas.....	10
2.3.2	Beneficios para las ventas .....	10
2.4	Componentes de Business Intelligence .....	11
2.4.1	Fuente de Datos .....	11
2.4.2	Procesamiento ETL .....	12
2.4.3	Modelado de Datos.....	13
2.4.4	Visualización de datos.....	13
2.5	Metodología desarrollo del proyecto.....	14

2.6	Modelo Machine Learning .....	15
2.6.1	Tipos de Modelos de Machine Learning .....	15
2.6.2	Métricas de evaluación .....	16
2.7	Tableros de Control (Dashboard) .....	18
2.7.1	Herramientas para los Tableros de Control .....	18
2.7.2	Características de los tableros de Control .....	19
3	Marco metodológico.....	21
3.1	Área de estudio.....	21
3.2	Flujograma metodológico.....	22
3.3	Recopilar y organizar la información .....	24
3.3.1	Fuentes de Información Primaria .....	24
3.3.2	Fuentes de Información Secundaria .....	25
3.3.3	Entendimiento del negocio .....	26
3.3.4	Identificación de Requerimientos.....	26
3.4	Análisis exploratorio de datos .....	27
3.4.1	Limpieza de los datos .....	28
3.4.2	Análisis descriptivo .....	31
3.4.3	Selección de las variables.....	33
3.4.4	Detección de valores atípicos .....	34
3.5	Modelo Machine Learning .....	37
3.5.1	Preparación del Modelo.....	37
3.5.2	Entrenamiento del Modelo sobre Train .....	39
3.5.3	Validación de los modelos ML.....	40
3.6	Diseño tableros de control (Dashboard).....	41
3.6.1	Selección Herramienta BI.....	41
3.6.2	Obtención de los datos .....	43
3.6.3	Modelo Dimensional .....	43
3.6.4	Identificación de tendencias .....	<b>¡Error! Marcador no definido.</b>
3.7	Visualización indicadores y tendencias.....	<b>¡Error! Marcador no definido.</b>

4. Resultados y Discusión .....	46
4.1. Resultados Recopilación y organización de la información.....	46
4.1.1 Resultados de la encuesta .....	46
4.2. Resultados Análisis exploratorio de datos.....	51
4.2.1 Resultados Limpieza de datos .....	52
4.2.1 Resultados Análisis descriptivo.....	53
4.2.1 Resultados selección de variables .....	54
4.2.1 Resultados valores atípicos .....	55
4.2. Resultados Modelo Machine Learning.....	56
4.3. Resultados de tableros de control.....	59
4.5. Visualizaciones tendencias.....	61
4.6. Discusión de los resultados .....	65
5. Conclusiones .....	67
6. Recomendaciones.....	68
Referencias bibliográficas .....	69
Anexos.....	71
Anexo 1. Encuesta comprensión del negocio .....	71
Anexo 2. Planilla de datos extraídos sobre las ventas del negocio.....	73
Anexo 3. Valores Atípicos de las variables numéricas .....	76
Anexo 4. Resultados corrección de los Outsiders .....	77

## Lista de figuras

---

Figura 1-1 Árbol de Problemas .....	3
Figura 2-1 Esquema de Análisis Exploratorio de Datos .....	7
Figura 2-2 Evolución de proveedores de BI (2022) .....	18
Figura 2-3 DashBoards con Power BI .....	19
Figura 3-1 Mapa Bolivia – Sucursales de la tienda de Herramienta .....	21
Figura 3-2 Flujograma Metodológico .....	22
Figura 3-3 DataSet Área Ventas archivo Excel.....	25
Figura 3-4 Importar data set en entorno Google Colab.....	28
Figura 3-5 Importar data set en entorno Google Colab.....	28
Figura 3-6 Análisis de valores nulos .....	29
Figura 3-7 Corrección de valores nulos .....	29
Figura 3-8 Errores Tipográficos.....	30
Figura 3-9 Corrección errores tipográficos.....	31
Figura 3-10 Cambiar tipo de dato variable Cantidad y Fecha Venta.....	31
Figura 3-11 Variables Categóricas .....	32
Figura 3-12 Variables Categóricas .....	32
Figura 3-13 Comportamiento Variables Numéricas .....	33
Figura 3-14 Eliminación de variables irrelevantes.....	34
Figura 3-15 Matriz de correlación .....	34
Figura 3-16 Valores atipicos.....	35
Figura 3-17 Grafico de Cajas valores atipicos.....	35
Figura 3-18 Error en registro de la vaiable PrecioVenta .....	36
Figura 3-19 Corrección de valores Atipicos.....	36
Figura 3-20 Exportar datos para crear los tableros de control .....	37
Figura 3-21 Copia del dataset para los modelos.....	37
Figura 3-22 verificar percentiles variable Utilidad.....	37
Figura 3-23 modelado Random Forest.....	38

Figura 3-24 Modelado Regresión Logística .....	39
Figura 3-25 Entrenamiento Modelo Random Forest.....	39
Figura 3-26 Entrenamiento Modelo Regresión Logística.....	40
Figura 3-27 Evaluación Modelo random Forest.....	40
Figura 3-28 Evaluación Modelo regresión Logística.....	41
Figura 3-29 Importación datos en Power BI .....	43
Figura 3-30 Esquema de estrella Tabla de Hechos .....	44
Figura 3-31 Columnas del nuevo dataset de ventas .....	44
Figura 3-32 Visualizaciones con Power BI .....	45
Figura 4-1 Resultado procedimiento de ventas.....	46
Figura 4-2 Herramientas de apoyo en las ventas.....	47
Figura 4-3 Seguimiento a los clientes .....	47
Figura 4-4 Metas y objetivos en las ventas.....	48
Figura 4-5 Revisión desempeño equipo de ventas .....	48
Figura 4-6 Métricas para evaluar las ventas.....	49
Figura 4-7 Capacitación al equipo de ventas.....	49
Figura 4-8 Tipo capacitación para las ventas .....	50
Figura 4-9 Retos para el área de ventas .....	50
Figura 4-10 Sugerencias para el área de ventas .....	51
Figura 4-11 Registros de las ventas importadas .....	51
Figura 4-12 Limpieza de valores nulos .....	52
Figura 4-13 Cambiar tipo de dato variable Id y fecha .....	52
Figura 4-14 Descriptivos variables categóricas .....	53
Figura 4-15 Subcategoría de la variable Sucursal.....	53
Figura 4-16 Descriptivos variables numéricas .....	54
Figura 4-17 Variables con mayor relevancia .....	54
Figura 4-18 Datos relevantes para la empresa .....	55
Figura 4-19 Corrección error de registro en las ventas.....	56
Figura 4-20 Métrica matriz de confusión.....	56

Figura 4-21 Interpretación de los resultados.....	58
Figura 4-22 Grafico comparación de precisión modelo Random Forest.....	58
Figura 4-23 Datos en general .....	59
Figura 4-24 Marcas más y menos vendidas.....	60
Figura 4-25 Comportamiento de ventas por mes .....	60
Figura 4-26 Productos con mejor rendimiento en ventas .....	60
Figura 4-27 rendimiento porcentual de las sucursales.....	61
Figura 4-28 Productos con más y menos Ventas .....	62
Figura 4-29 Marca con mejor rendimiento .....	62
Figura 4-30 Comportamiento de ventas en las Sucursales .....	63
Figura 4-31 Mejores Vendedores según la sucursal.....	63
Figura 4-32 Fluctuación de las ventas por temporada.....	64
Figura 4-33 Clientes más rentables para la empresa.....	64
Figura 4-34 Script calcular la Precisión Modelo Random Forest.....	65
Figura 4-35 Script calcular la Precisión Modelo Regresión Logística.....	65
Figura 4-36 Script calcular la Precisión Modelo Regresión Lineal.....	66

# Lista de tablas

---

Tabla 2-1 Tabla comparativa Leguajes análisis de datos..... 9

Tabla 2-2 Componentes de proceso ETL ..... 12

Tabla 3-1 Ficha resumen Encuesta a vendedores ..... 25

Tabla 3-2 Requerimientos según los empleados ..... 27

Tabla 3-3 Tabla comparativa de los proveedores..... 42

Tabla 4-1 Comparación resultados métricas de evaluación ..... 57

Tabla 4-2 Tabla Comparativa Modelo de machine Learning..... 66



# 1. Introducción

El análisis de datos y las tendencias son elementos cruciales en la gestión de ventas, especialmente en el contexto de una tienda de herramientas. La historia del Business Intelligence (BI) se remonta a 1958, cuando Hans Peter Luhn, investigador de IBM, acuñó el término en su artículo "A Business Intelligence System", definiéndolo como la habilidad de aprender relaciones entre hechos para guiar acciones hacia metas deseadas. (IBM, 2010)

En la década de 1980, Howard Dresner popularizó el término y lo asoció con metodologías que mejoran la toma de decisiones empresariales utilizando sistemas basados en datos. (Kaizen, 2008)

Con el crecimiento exponencial de los datos en las últimas décadas, las organizaciones han comenzado a implementar soluciones de Business Intelligence (BI) con el objetivo de extraer valor significativo de la información disponible. El presente análisis examina cómo la aplicación de BI puede revolucionar la toma de decisiones en el ámbito de ventas, al ofrecer un enfoque fundamentado en datos que optimiza la eficiencia operativa y potencia la rentabilidad del negocio.

El presente proyecto consiste en implementar una solución Business Intelligence, utilizando la metodología de Ralph Kimball, para el proceso de toma de decisiones del área de ventas de la empresa comercial de herramientas "MASTER TOOLS", el cual tiene 4 sucursales en las ciudades de La Paz, Cochabamba, Santa Cruz y Oruro.

El área de ventas constituye uno de los pilares fundamentales de toda empresa u organización, al ser responsable de la generación de ingresos y la distribución eficiente de productos desde el inventario hacia los clientes. Este proceso no solo asegura la rentabilidad, sino que también impulsa la expansión hacia nuevos mercados estratégicos y la apertura de sucursales en ubicaciones clave. La realización de un análisis exhaustivo de las ventas resulta crucial para fundamentar decisiones estratégicas, como la identificación de los productos más rentables, determinar las ubicaciones geográficas con mayor potencial de retorno económico y reconocer a los clientes que aportan significativamente al volumen de ingresos.

## 1.1. Antecedentes

Los datos históricos, ya sean de ventas, inventarios, informes clínicos, deportivos, entre otros, han adquirido una creciente relevancia en diversas industrias de países desarrollados. A través de un análisis adecuado, utilizando herramientas como Business Intelligence y otros modelos de análisis de datos, las empresas han logrado interpretar esta información y extraer conclusiones valiosas que contribuyen significativamente a la toma de decisiones estratégicas y al fortalecimiento de su competitividad.

Las herramientas de Business Intelligence están pensadas para ayudar al personal de las empresas a dar sentido a todos los datos complejos con los que tienen que trabajar diariamente (GALIANA, 2022)

En el caso de Bolivia, son pocas las empresas o negocios que gestionan sus datos de manera eficiente. Esto es particularmente evidente en pequeñas y medianas empresas, donde la mayoría carece de software especializado para el procesamiento y uso adecuado de la información. En su lugar, continúan utilizando métodos tradicionales, como registros en hojas de papel o cuadernos. Esta situación se ve agravada por la falta de infraestructura tecnológica adecuada, como computadoras, tabletas o aplicaciones móviles, lo que limita significativamente su capacidad para modernizar y optimizar sus procesos de gestión de datos

En el caso de la empresa comercial de herramientas “MASTER TOOLS” usaban un libro de Excel y recién a partir del año 2022 comenzaron a integrar un software de ventas para manejar mejor las ventas.

Análisis e inteligencia de negocios (ABI) es un término general que incluye las aplicaciones, infraestructura y herramientas, y las mejores prácticas que permiten el acceso y el análisis de la información para mejorar y optimizar las decisiones y el rendimiento (GARTNER, 2021)

## **1.2. Justificación**

La presente investigación propone la implementación de Business Intelligence de empresa comercial de herramientas “MASTER TOOLS” con el propósito de proporcionar información actualizada, clara, precisa y accesible en todo momento, garantizando así una alta calidad en los datos..

Considerando que, tras la finalización de la emergencia sanitaria por la pandemia de COVID-19 en el año 2020, numerosos negocios experimentaron un significativo avance tecnológico en el ámbito de las ventas en línea mediante la implementación de sistemas digitales, estrategias de marketing y otras innovaciones, resulta indispensable que la tienda de herramientas adopte una solución de Business Intelligence (BI).

### **1.2.1 Justificación Social**

La adopción de BI en el área de ventas contribuye a una mejor experiencia del cliente al personalizar ofertas y mejorar la atención. Comprender las preferencias del consumidor permite a las tiendas adaptar su stock y promociones, resultando en una mayor satisfacción y lealtad del cliente. Asimismo también beneficiará a los vendedores, permitiéndoles cumplir con sus objetivos mensuales

### **1.2.2 Justificación Tecnológica**

Las herramientas de BI permiten integrar datos provenientes de diversas fuentes, facilitando un análisis más profundo, efectivo y de forma más rápida. Incluye el uso de algoritmos avanzados para identificar patrones y predecir comportamientos futuros, lo cual es esencial para enfrentar a la competencia de la empresa. También podrán acceder al estado actual de las herramientas más vendidas y a la sucursal con

mejores resultados en ventas. Esto les permitirá enfocar sus esfuerzos en promover las herramientas con mayor demanda durante un período determinado o en las regiones con mayor potencial comercial

### 1.2.3 Justificación Económica

La implementación de BI puede resultar en un aumento significativo en las ventas y la rentabilidad. Al optimizar los procesos y recursos basados en análisis precisos, las empresas pueden reducir costos y maximizar el margen de utilidad al enfocar sus esfuerzos en áreas con mayor potencial.

### 1.3. Planteamiento del problema

Se plantea investigar de qué manera la utilización de herramientas de análisis exploratorio de datos y modelos de aprendizaje automático puede optimizar los procesos en el área de ventas. Tal como se ilustra en la Figura 1-1, se presenta un árbol de problemas que permite identificar las causas y efectos asociados a los procesos en esta área.

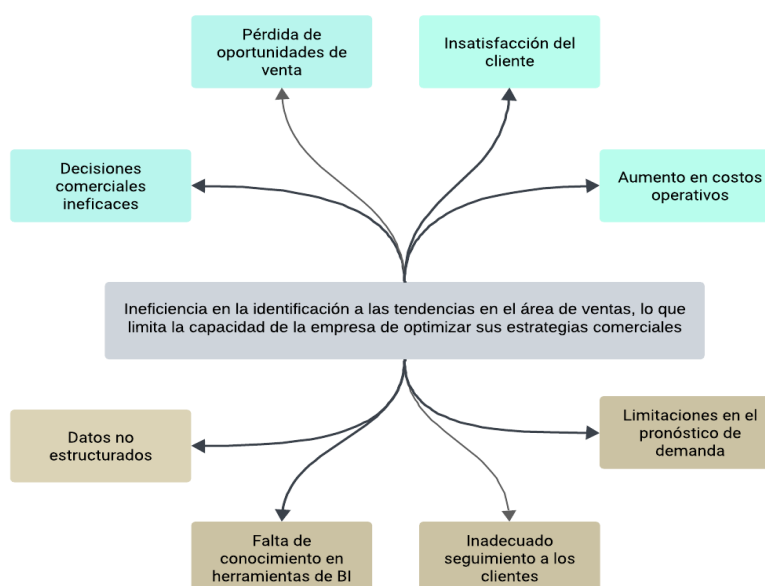


Figura 1-1 Árbol de Problemas

Fuente: Elaboración Propia, 2024.

#### Problema Central

Ineficiencia en la identificación a las tendencias en el área de ventas, lo que limita la capacidad de la empresa de optimizar sus estrategias comerciales.

#### Causas

- Datos no estructurados: La falta de un sistema para analizar los datos dificulta la obtención de información coherente sobre las tendencias de consumo.

- Falta de conocimiento en herramientas de BI: Los vendedores no están familiarizados con las herramientas de BI, lo que limita su capacidad para interpretar datos y tomar decisiones informadas.
- Inadecuado seguimiento a los clientes: Sin un análisis adecuado, los vendedores no logran realizar el seguimiento a los clientes potencialmente rentables, lo que lleva a bajas ventas de parte de los vendedores.
- Limitaciones en el pronóstico de demanda: La ausencia de técnicas avanzadas para predecir la demanda puede resultar en exceso o falta de inventario, afectando la satisfacción del cliente, las ventas e incrementos en los costos de almacenamiento.

### **Efectos**

- Decisiones comerciales ineficaces: La falta de información precisa puede llevar a decisiones que no responden a las necesidades reales del mercado, afectando las ventas.
- Pérdida de oportunidades de venta: La incapacidad para identificar tendencias emergentes puede resultar en la pérdida de oportunidades para capitalizar sobre nuevos productos o servicios.
- Insatisfacción del cliente: La falta de productos adecuados en el momento del cierre de las ventas puede generar frustración entre los clientes, afectando su lealtad hacia la tienda.
- Aumento en costos operativos: La ineficiencia en la gestión del inventario y la respuesta a las tendencias puede incrementar los costos operativos, disminuyendo así las utilidades del negocio.

Este árbol de problemas proporciona una visión clara sobre los desafíos que enfrenta la empresa comercial “MASTER TOOLS” al implementar BI en su área de ventas, así como las consecuencias que pueden derivarse si no se abordan adecuadamente.

### **1.3.1 Formulación del problema**

A pesar del auge de Business Intelligence (BI) en las empresas, muchos negocios de herramientas aún carecen de ese enfoque estructurado para analizar sus datos de ventas. Esta deficiencia limita la comprensión de las preferencias y comportamientos del consumidor, lo que puede derivar en decisiones que impacten negativamente tanto en el desempeño de las ventas como en la satisfacción del cliente. En este contexto, surge la pregunta central:

- ¿cómo puede la implementación del análisis de datos mediante Business Intelligence contribuir a la mejora de las decisiones estratégicas en el área de ventas?

Esta a su vez se descompone en varias interrogantes específicas:

- ¿Qué datos son más relevantes para analizar el comportamiento del consumidor?
- ¿Qué métricas deben ser monitoreadas para evaluar el impacto de las estrategias implementadas?

## **1.4. Objetivo general**

Realizar un análisis exploratorio de datos para determinar los patrones y tendencias de venta de herramientas aplicando técnicas de Business Intelligence, con el fin de optimizar la toma de decisiones estratégicas y mejorar el rendimiento de la comercializadora “MASTER TOOLS”.

### **1.4.1 Objetivos específicos**

- Recopilar y organizar la información histórica de las ventas de la comercializadora de herramientas MASTER TOOLS para su posterior análisis con técnicas de Business Intelligence.
- Realizar un análisis exploratorio de los datos para identificar los principales patrones y tendencias en el comportamiento de los consumidores.
- Entrenar y seleccionar el modelo de Machine Learning más óptimo con el objetivo de mejorar la gestión de ventas en la empresa.
- Desarrollar tableros de control que faciliten a los usuarios la visualización clara y concisa de los principales indicadores de desempeño (KPIs) y las tendencias identificadas en el ámbito de Business Intelligence.

## 2. Marco teórico

En este capítulo se aborda un conjunto de investigaciones, teorías y conceptos en que se basa el proyecto, estos temas servirán de fundamento para la propuesta de aplicación de business intelligence y Machine learning para la toma de decisiones enfocado al área de ventas de un negocio de comercialización de herramientas.

### 2.1. Recolección de Información

En la actualidad, prácticamente todas las actividades generan información, ya sea a través de textos, videos, comentarios, registros u otros medios. Por ello, el proceso de recopilación de información se convierte en un paso fundamental para iniciar cualquier investigación orientada a analizar los datos relacionados con los procesos bajo estudio. La recolección de información se define como la búsqueda, recopilación y medición de datos provenientes de múltiples fuentes, lo que permite obtener un panorama detallado sobre los procesos, servicios y productos de una empresa. Este enfoque facilita la evaluación de los resultados, lo que a su vez permite tomar decisiones más informadas y estratégicas. (Falcón Morales, 2023)

#### 2.1.1 Ventajas y Desventajas de la recolección de información

La principal ventaja es el conocimiento en sí, pues conocer es poder de alguna manera en la empresa o negocio, es saber lo que tus clientes opinan que es algo negativo o positivo en tu producto, servicio o proceso. (Falcón Morales, 2023)

Algunas ventajas más para destacar sobre la recolección de información.

- Identificación de oportunidades facilitando el descubrimiento de nuevas oportunidades para el crecimiento de la empresa.
- Facilita el análisis del comportamiento y las preferencias de los clientes, lo que ayuda a personalizar estrategias y mejorar la experiencia del usuario.
- Almacenar la información de forma estructurada garantiza su disponibilidad y conservación para análisis posteriores más detallados y profundos.
- Ayuda a encontrar métodos efectivos para organizar y estructurar la información recopilada.

La principal desventaja es que las personas suelen pensar que “la recolección de información es magia” y no es así. Es un proceso de mejora continua, por lo tanto, no tiene fin. (Falcón Morales, 2023)

Otras desventajas a considerar en la recolección de información.

- Uso indebido de datos ya que existe el riesgo de que la información de los usuarios sea utilizada de manera malintencionada, lo que puede comprometer la privacidad y la confianza del cliente.
- Posibilidad de filtración de información sensible a la competencia lo que puede afectar negativamente la ventaja competitiva de la empresa.

### 2.1.2 Técnicas recolección de información

Las técnicas de recolección de información son métodos o herramientas utilizadas para recopilar datos de manera sistemática y organizada, con el fin de obtener información relevante para un análisis posterior. Estas técnicas son fundamentales en procesos de investigación, toma de decisiones y mejora continua, ya que permiten recabar datos precisos y confiables. (Clientify, 2024)

Se mencionan a continuación las principales técnicas de recolección de información:

- Entrevista: Consiste en una conversación directa entre el investigador y el entrevistado, donde se recopila datos cualitativos detallados.
- Encuesta: Consiste en realizar preguntas estructuradas o semiestructuradas a un grupo de personas para obtener información específica, esta puede ser en papel o de forma digital.
- Software CRM: Método más rápido al obtener la información de la empresa de los registros históricos como ser de las ventas, inventarios, estados financieros entre otros.

## 2.2. Análisis exploratorio de Datos

El análisis exploratorio de datos se emplea principalmente para identificar qué información puede ser revelada tras la recopilación de datos, transformando dicha información en un recurso valioso para la empresa. Este proceso proporciona una comprensión más profunda de las variables presentes en el conjunto de datos y de las relaciones que existen entre ellas. Además, el análisis exploratorio puede ser útil para determinar la idoneidad de las técnicas estadísticas que se aplicarán en el análisis posterior de los datos. Desarrolladas originalmente por el matemático estadounidense John Tukey en la década de 1970, las técnicas de análisis exploratorio de datos continúan siendo ampliamente utilizadas en el proceso de descubrimiento de datos en la actualidad. (IBM, 2024)

Como se muestra en la figura 2-1 el esquema general para un análisis exploratorio de datos con la entrada de datos sin procesar y la salida con datos limpios y de mayor valor.

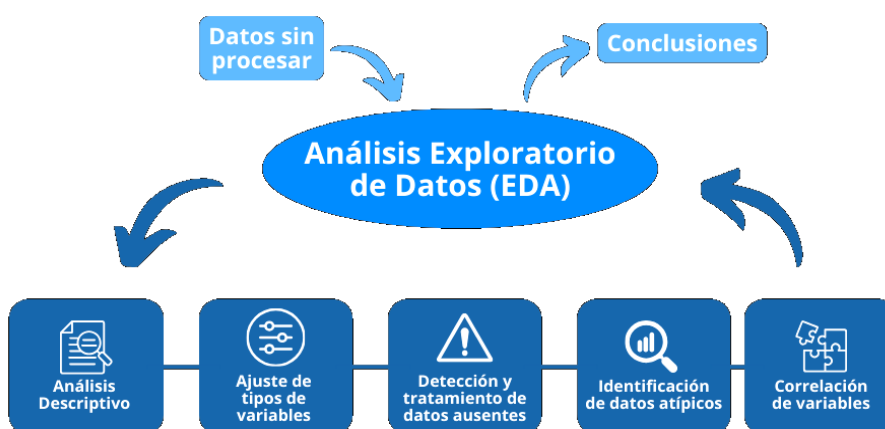


Figura 2-1 Esquema de Análisis Exploratorio de Datos

Fuente: (Gravitar, 2022)

### 2.2.1 Importancia del Análisis exploratorio de Datos en el Comercio

El análisis de datos constituye una herramienta fundamental que ha transformado el comercio minorista. A través de la recopilación y el análisis de grandes volúmenes de información, los minoristas son capaces de identificar patrones y tendencias que anteriormente resultaban imperceptibles. Esta capacidad de profundizar en los datos permite a las empresas obtener una comprensión más precisa de las necesidades y preferencias de sus clientes, lo que les facilita ofrecer experiencias de compra altamente personalizadas y ajustadas a sus expectativas. (Prometeusgs, 2024)

El análisis de datos permite a los minoristas personalizar sus estrategias publicitarias y promociones dirigidas a los clientes. Al contar con una comprensión exhaustiva de las preferencias y comportamientos de los consumidores, los minoristas pueden diseñar campañas publicitarias que resuenen de manera efectiva con su audiencia, maximizando así el retorno sobre la inversión. Por ejemplo, si un cliente ha mostrado interés en productos de una categoría específica, el minorista tiene la capacidad de enviarle ofertas y promociones relacionadas con dicha categoría. En un entorno comercial cada vez más competitivo, la personalización basada en el análisis de datos se convierte en un factor clave para fomentar la lealtad del cliente y optimizar el rendimiento comercial. (Prometeusgs, 2024)

### 2.2.2 Herramienta para el análisis exploratorio de datos

Algunas de las herramientas de ciencia de datos más comunes utilizadas para crear un análisis exploratorio de datos incluyen:

#### 2.2.2.1 Lenguaje de Análisis

- **Python:** un lenguaje de programación interpretado, orientado a objetos y con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con la tipificación dinámica y la vinculación dinámica, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de programación o aglutinante para conectar componentes existentes entre sí. Python y EDA pueden utilizarse conjuntamente para identificar los valores faltantes en un conjunto de datos, lo que es importante para poder decidir cómo tratar esos valores faltantes para el aprendizaje automático. (IBM, 2024)
- **R:** lenguaje de programación de código abierto y entorno de software gratuito para computación estadística y gráficos respaldado por R Foundation for Statistical Computing. El lenguaje R se utiliza ampliamente entre los estadísticos de la ciencia de datos para desarrollar observaciones estadísticas y análisis de datos. (IBM, 2024)
- 

En la Tabla 2-1 se presentan las características de ambos lenguajes de programación. Por un lado, R, un lenguaje ampliamente utilizado durante mucho tiempo para el análisis exploratorio de datos, y por otro, Python, que en los últimos años ha incrementado su popularidad entre los analistas de datos debido a su versatilidad y facilidad de uso.



Tabla 2-1 Tabla comparativa Leguajes análisis de datos

Características	R	Python
Tipo de lenguaje	Lenguaje especializado en estadística y análisis de datos	Lenguaje de programación de propósito general, versátil y fácil de aprender
Facilidad de aprendizaje	Más fácil para tareas estadísticas simples, pero complejo para funcionalidades avanzadas	Sintaxis intuitiva, ideal para principiantes y desarrollo en diversas áreas
librerías y paquetes	Colección de paquetes para análisis estadístico (ej. Ggplot2, dplyr)	Librerías especializadas para ciencia de datos (ej. Pandas, NumPy , Matplot)
Rendimiento	Generalmente más lento en ejecución que Python	Más rápido en ejecuciones generales y manipulación de datos
Integración con Power BI	Permite análisis estadísticos avanzados y visualizaciones personalizadas	Ideal para limpieza, transformación de datos y análisis predictivo
Uso de machine learning	Menos común, pero tiene capacidades a través de paquetes específicos (ej. Caret)	Muy utilizado en machine learning con bibliotecas como scikit-learn y tensorflow

Fuente: adaptado de <https://bauc3m.com/r-vs-python/>

Es importante destacar que ambos lenguajes presentan características sobresalientes; sin embargo para el desarrollo del proyecto, se considerarán las ventajas de Python este lenguaje se distingue por su fácil aprendizaje, la amplia variedad de bibliotecas disponibles para el análisis de datos y además por su rendimiento de ejecución, el cual supera al de R. Estas características fundamentan su elección como herramienta de análisis de datos para el presente proyecto.

#### 2.2.2.2 Google Colab

Es un servicio de Google que permite ejecutar código de Python en la nube. Asimismo, es una herramienta muy útil para entrenar modelos de inteligencia artificial y realizar ciencia de datos ya que permite utilizar la potencia de los servidores de Google de manera gratuita. También, el servicio de Google permite utilizar GPUs y TPUs para entrenar modelos de manera más rápida. (Colab.Research, 2022)

El entorno de Google Colab es una herramienta online o en la nube que permite escribir, ejecutar y código en Python directamente desde un navegador. Es especialmente popular para trabajos relacionados con

Machine Learning, Data Science, y análisis de datos, ya que proporciona una configuración rápida y sencilla sin necesidad de instalar software adicional. (Colab.Research, 2022)

Este ambiente de desarrollo en la nube se ofrece sin costo y cuenta con una interfaz intuitiva, similar a la de Jupyter Notebook. Su integración con Google Drive facilita el almacenamiento de archivos e información para su posterior uso en el análisis exploratorio de datos, la evaluación de variables relevantes, la creación de visualizaciones y el desarrollo de modelos de Machine Learning.

### 2.3 Identificación de Tendencias

Las tendencias de consumo son patrones observables en el comportamiento de compra de los consumidores a lo largo del tiempo. Estas tendencias pueden ser influenciadas por factores económicos, sociales y tecnológicos. La capacidad de una empresa para identificar y responder a estas tendencias puede determinar su éxito en el mercado. Un estudio sobre el uso de minería de datos para analizar tendencias en medios noticiosos destaca cómo el análisis sistemático puede revelar preferencias estables entre los consumidores (González, 2021)

Esto implica la necesidad de mantenerse atentos a las tendencias emergentes, escuchar activamente a los clientes, invertir en tecnología avanzada, promover la innovación interna y ofrecer experiencias personalizadas de alta calidad, la evaluación continua del impacto de las estrategias implementadas son elementos clave para lograr una adaptación exitosa en un entorno empresarial dinámico.

#### 2.3.1 Indicadores de rendimiento y métricas de ventas

En un entorno de negocios caracterizado por una creciente competitividad, la disponibilidad de Indicadores Clave de Rendimiento (KPIs) de ventas claras y efectivas resulta esencial para que las empresas puedan evaluar y optimizar su desempeño comercial. Los KPIs, o métricas clave de rendimiento, permiten a los equipos de ventas analizar su progreso e identificar áreas susceptibles de mejora, garantizando así que cada acción esté alineada con los objetivos estratégicos de la organización. Sin una medición adecuada, resulta complejo determinar la eficacia real de las estrategias implementadas, identificar los factores que limitan el crecimiento y determinar las áreas en las que se deben realizar ajustes para maximizar los resultados. (Torrez, 2024)

Los KPI ayudan a establecer objetivos de rendimiento, evaluar el progreso hacia esos objetivos y valorar la eficacia de las estrategias de ventas.

#### 2.3.2 Beneficios para las ventas

Estos indicadores de rendimiento ofrecen una visión clara y cuantificable del desempeño en las actividades de ventas, permitiendo a las empresas identificar áreas de oportunidad y aspectos que requieren fortalecerse mediante estrategias específicas. Entre los principales beneficios que aportan a la gestión de ventas se encuentran: (Torrez, 2024)

- **Medición del rendimiento:** estas métricas permiten a las empresas medir objetivamente el rendimiento de su equipo de ventas y la efectividad de sus estrategias comerciales. Al establecer

indicadores claros y específicos, las organizaciones pueden evaluar con precisión el desempeño de sus actividades de venta e identificar las áreas que necesitan mejorar.

- **Identificación de tendencias:** a través del monitoreo continuo de los KPIs de ventas a lo largo del tiempo, las empresas pueden identificar tendencias y patrones tanto en el comportamiento de sus clientes como en el rendimiento de sus productos. Esta información resulta invaluable para ajustar las estrategias comerciales y anticipar cambios en el mercado.
- **Optimización de recursos:** Al comprender qué actividades de ventas son más efectivas y qué áreas necesitan mejoras, las empresas pueden optimizar el uso de sus recursos y esfuerzos de ventas. Esto puede incluir asignar recursos adicionales a las estrategias que están generando resultados positivos y ajustar o eliminar aquellas que no están funcionando.
- **Toma de decisiones informada:** estos indicadores proporcionan datos objetivos que respaldan la toma de decisiones empresariales. Al fundamentar las decisiones en datos concretos, en lugar de suposiciones o intuiciones, las empresas pueden adoptar medidas más informadas y estratégicas para impulsar el crecimiento y alcanzar el éxito.
- **Fomento de la competitividad:** al establecer y monitorear los patrones de ventas, las empresas pueden mantenerse competitivas en su industria. (Torrez, 2024)

## 2.4 Componentes de Business Intelligence

“La inteligencia de negocios (BI) se refiere a técnicas informáticas utilizadas para detectar, desenterrar y analizar datos comerciales, como ingresos por ventas por productos y/o departamentos, o por costos e ingresos asociados. Las tecnologías de BI proporcionan vistas históricas, actuales y predictivas de las operaciones comerciales.” (Cebotarean, 2011)

La inteligencia de negocios tiene como objetivo extraer conocimiento para ayudar con las decisiones estratégicas de un negocio, para ello lo único necesario es la recopilación de datos estos son luego procesados con herramientas dedicados a inteligencia de negocios.

Business Intelligence (BI) se refiere a un conjunto de tecnologías, aplicaciones y prácticas que permiten la recopilación, integración, análisis y presentación de información empresarial. Su objetivo es apoyar la toma de decisiones informadas mediante la transformación de datos en información útil. A lo largo de las décadas, BI ha evolucionado desde simples sistemas de informes hasta complejas plataformas que utilizan análisis predictivo y minería de datos para descubrir patrones y tendencias en grandes volúmenes de datos. (Cruaños, 2020)

Este proceso permite obtener una comprensión más profunda de las operaciones de un negocio, identificar tendencias, patrones y tomar decisiones informadas basadas en evidencia.

### 2.4.1 Fuente de Datos

Parte fundamental para la inteligencia de negocios. Incluye la extracción y recolección de diversas fuentes como bases de datos, sistemas ERP (Planificación de recursos empresariales), documentos, archivos, redes sociales, entre otros.

Las fuentes de información externas son esenciales para enriquecer la información que tenemos de nuestros clientes. En algunos casos es interesante incorporar información referente, por ejemplo, área ventas, inventarios, finanzas, entre otros. Acceder a distintas bases de datos requiere distintas habilidades y los conocimientos. (Cano, 2007)

#### 2.4.2 Procesamiento ETL

Lo siguiente será procesar y transformar los datos en un formato que sea adecuado para su análisis. Lo cual implica su limpieza de datos, la normalización y la integración de datos de diferentes fuentes.

Los elementos, las operaciones realizadas, y los resultados esperados de un ETL para cada uno de los tres subprocesos: (Aguilar, 2017)

Como se muestra en la siguiente tabla 2-2 como los procesos de los componentes de ETL, detallando las entradas los procesos y salidas esperadas de cada componente.

Tabla 2-2 Componentes de proceso ETL

Componentes	Entrada Elementos	Proceso Operaciones	Salida Resultado
Extracción	Fuentes de datos (base de datos, hojas de cálculo, archivos)	Seleccionar	Datos sin procesar
Transformación	Datos sin procesar	Limpieza, crear métricas, transformación, y aplicación de funciones	Datos formateados, estructurados y resumidos
Carga	Datos formateados, estructurados y resumidos	Inserción	Datos formateados, estructurados y modelados (en data warehouse)

*Fuente:* Elaboración propia (2024)

Dichos procesos son la parte fundamental para el análisis de datos aplicando Business Intelligence y obtener los resultados para la toma de decisiones dentro la empresa.

- **Extracción:** Donde se extraen datos de diferentes fuentes los cuales pueden ser del tipo Estructurado como una base de datos relacional, Semiestructurados que se obtienen de documentos como Excel, pdf, textos entre otros, también hay los datos no estructurados que pueden venir de videos, páginas web, imágenes, mensajes de redes sociales. Es una etapa importante para luego analizar estos datos y comprender la naturaleza del negocio.

- También se debe analizar el volumen de los datos que se extraen si son demasiados podría ser que la solución deba ser tratada de forma diferente y con diferentes herramientas más avanzadas.
- **Transformación:** Implica realizar una manipulación mínima de los datos según las necesidades de negocio; estas transformaciones pueden ser las siguientes: eliminar datos que no sean relevantes, obtención de nuevos valores calculados, la unión o concatenación de dos o más columnas, separar una columna en otras según sea conveniente, identificación de columnas clave en las tablas para cumplir con los objetivos impuestos por el negocio.
- **Carga:** Consiste en categorizar a los datos en diferentes niveles, realizar el modelo de datos, donde se irán almacenando en memoria como los Datawarehouse, y dividirlos en datamart según las áreas del negocio.

### 2.4.3 Modelado de Datos

El modelado de datos es el proceso de analizar y definir todos los diferentes tipos de datos que su negocio recopila y produce, así como las relaciones entre esos bits de datos. Mediante el uso de texto, símbolos y diagramas, los conceptos de modelado de datos crean representaciones visuales de los datos que se capturan, almacenan y utilizan en su negocio. (IBM, [www.ibm.com](http://www.ibm.com), 2016)

El proceso iterativo consta de 4 pasos, y se realiza a partir de procesos que se priorizan tales como:

- **Seleccionar el proceso de negocio:** Se selecciona el área del negocio para realizar el modelado, según el requerimiento de gerencia para la toma de decisiones.
- **Definir el nivel de granularidad:** Depende de los requisitos que requiera el negocio, especificando un nivel alto de detalle.
- **Seleccionar las dimensiones:** Generalmente las tablas de dimensiones cuentan con atributos que permiten un mejor análisis acerca de una medida en una tabla de hechos. Para poder determinar las tablas de dimensiones, tomamos en cuenta sus atributos para ser candidatos para ser encabezados en los informes, tablas, gráficos entre otros.
- **Determinar medidas y tabla de hechos:** Se identifican medidas que se desean analizar, agrupando datos. La tabla de hechos cuenta con atributos adecuados a los requerimientos del negocio. En esta parte, la granularidad es el nivel de detalle que tiene cada registro en una tabla de hechos. (Yaipén, 2022)

### 2.4.4 Visualización de datos

Una vez que los datos han sido limpiados, se procede a la correcta carga de la información y al inicio de la elaboración de visualizaciones. Estas visualizaciones pueden representar los datos a través de una amplia variedad de gráficos dinámicos, lo que permite investigar tendencias, patrones, errores y características inherentes a los datos. Esto se puede lograr mediante el uso de diagramas de distribución, histogramas, gráficos de dispersión, entre otros. (Chavez Huapaya, 2018)

## 2.5 Metodología desarrollo del proyecto

Como se muestra en la tabla 2-2 se encuentra un cuadro comparativo de las características de las metodologías para desarrollar proyectos BI, donde se analizan las fases del proceso de las 3 metodologías como ser KIMBALL (Kimball, 2015), INMON desarrollado por (Inmon, 2005) y HEFESTO desarrollado por (Bernabeu, 2010)

Tabla 2-2 Cuadro Comparativo de metodologías

METODOLOGIAS	KIMBALL	INMON	HEFESTO
Descripción	Se centra en los procesos de la empresa, extrayendo la información relevante para modelar los cubos y generar el almacén de datos	Se fundamenta en el enfoque de un entorno corporativo analítico de apoyo a la toma de decisiones	Basado en las necesidades de información clave del negocio, identificando indicadores y estrategias de análisis.
Requerimientos	Determinación de las necesidades, características y usuarios de la organización	Determinación de las necesidades, características y usuarios de la organización	Determinación de las necesidades, características y usuarios de la organización
Enfoque e integración de datos	Formar data mart para cada área del negocio	Formar data warehouse global de toda la empresa	Formar data warehouse global de toda la empresa
Modelación	Análisis de requerimientos, granularidad, estructura, dimensiones, hechos, ETL, repositorio, cuadros de mando.	Análisis de requerimientos, granularidad, estructura, dimensiones, hechos, ETL, tablero de control	Análisis de requerimientos, granularidad, estructura, dimensiones, hechos, ETL. Sin tableros de control
Adaptabilidad	Crecimiento de forma evolutiva	Aplicación de toda la funcionalidad del proyecto	Aplicación de toda la funcionalidad del proyecto
Mantenimiento	Medio-Altas	Sencillas	Medianas
Costos	Bajo costo inicial y para fases posteriores. Implementa por áreas.	Alto costo inicial, pero menor para fases posteriores.	Alto costo inicial, pero menor para fases posteriores.

*Fuente:* adaptado de <https://gravitar.biz/datawarehouse/metodologias-data-warehouse/>

De la anterior tabla 2-2 se puede concluir que la metodología KIMBALL es la más compatible para el desarrollo del proyecto, por los requerimientos, su enfoque incremental y la visualización de los tableros de control.

## **2.6 Modelo Machine Learning**

El Machine Learning (ML) es una disciplina perteneciente a la informática, la ciencia de datos y la inteligencia artificial (IA), que permite a los sistemas aprender y mejorar automáticamente a partir de los datos, sin necesidad de intervención explícita en la programación.

A diferencia de los enfoques tradicionales que dependen de instrucciones programadas, los modelos de ML se fundamentan en algoritmos y técnicas estadísticas que ejecutan tareas mediante la identificación de patrones en los datos y la realización de inferencias. (Chrystal R, 2023)

Un ejemplo práctico de su aplicación se observa las páginas web de ventas, donde los algoritmos de machine learning influyen en las decisiones de compra de los consumidores al ofrecer recomendaciones personalizadas basadas en su historial de compras y preferencias.

### **2.6.1 Tipos de Modelos de Machine Learning**

Los algoritmos de machine learning se dividen 2 categorías principales: aprendizaje supervisado y aprendizaje no supervisado.

#### **2.6.1.1 Machine learning supervisado**

El machine learning supervisado es una categoría dentro del aprendizaje automático en la cual el modelo se entrena utilizando un conjunto de datos etiquetados, es decir, donde la variable objetivo o de resultado es conocida. Por ejemplo, si un equipo de científicos de datos desarrollara un modelo para predecir la ocurrencia de tornados, las variables de entrada podrían incluir datos como la fecha, la ubicación geográfica, la temperatura, los patrones de flujo del viento, entre otros, mientras que la variable de salida correspondería a la actividad real de tornados registrada en esos días específicos. (Chrystal R, 2023)

El aprendizaje supervisado se utiliza habitualmente para la evaluación de riesgos, el reconocimiento de imágenes, el análisis predictivo y la detección del fraude a continuación, se mencionan algunos algoritmos:

- Algoritmos de regresión: predicen valores de output identificando relaciones lineales entre valores reales o continuos (por ejemplo, temperatura, salario).
- Algoritmos de clasificación: predicen variables de output categóricas (por ejemplo, "basura" o "no basura") etiquetando piezas de datos de entrada, su objetivo es asignar una categoría a cada muestra en función de sus características.
- Clasificadores Naïve Bayes: permiten tareas de clasificación para grandes conjuntos de datos. También forman parte de una familia de algoritmos de aprendizaje generativo que modelan la distribución de entrada de una clase o categoría determinada.

- Redes neuronales: simulan el funcionamiento del cerebro humano, con un enorme número de nodos de procesamiento conectados que pueden facilitar procesos como la traducción de lenguaje natural, el reconocimiento de imágenes, el reconocimiento del habla y la creación de imágenes.
- Algoritmos de bosque aleatorio: Estos algoritmos predicen un valor o una categoría combinando los resultados de múltiples árboles de decisión. Este enfoque mejora la precisión y la robustez de las predicciones al reducir la posibilidad de sobreajuste.

### **2.6.1.2 Machine learning No supervisado**

Los algoritmos de aprendizaje no supervisado, como Apriori, los modelos de mezclas gaussianas (GMM) y el análisis de componentes principales (PCA), extraen conclusiones de conjuntos de datos no etiquetados, lo que facilita el análisis exploratorio y permite el reconocimiento de patrones, así como el modelado predictivo. El método de aprendizaje no supervisado más común es el análisis de conglomerados, que emplea algoritmos de agrupación para categorizar los puntos de datos en función de la similitud de sus valores, siendo útil en aplicaciones como la segmentación de clientes y la detección de anomalías. (Chrystal R, 2023)

Los algoritmos de asociación permiten a los científicos de datos identificar asociaciones dentro de grandes bases de datos, facilitando la visualización de los datos y la reducción de la dimensionalidad, algunos de estos algoritmos son:

- K-medias: Esta técnica asigna los puntos de datos a k grupos, donde los puntos más cercanos a un centroide determinado se agrupan en la misma categoría. El número k representa el número de grupos y se determina en función del tamaño y nivel de granularidad deseado. La agrupación K-medias es ampliamente utilizada para la segmentación del mercado, la agrupación de documentos, la segmentación de imágenes y la compresión de imágenes.
- Agrupación jerárquica: Describe un conjunto diverso de técnicas que incluyen tanto la agrupación aglomerativa como divisiva. En el caso aglomerativo, los puntos inicialmente se aíslan en grupos individuales que luego se fusionan iterativamente según su similitud hasta quedar reducidos a un solo clúster. Por otro lado, en el método divisivo, un único clúster inicialmente abarca todos los datos y luego se divide progresivamente basándose en las diferencias entre ellos.
- Agrupación probabilística: Ayuda a resolver problemas relacionados con la estimación de densidad o con tipos "suaves" de agrupamiento al clasificar los puntos según su probabilidad de pertenecer a una distribución específica.

### **2.6.2 Métricas de evaluación**

Estas métricas se utilizan para comparar diferentes modelos y seleccionar el que tenga un mejor rendimiento en base a las necesidades y objetivos del problema a resolver.

Por ejemplo alguno de estas métricas que se usan dependiendo del modelo de machine learning son:



- Matriz de Confusión: Es una herramienta fundamental en problemas de clasificación que permite evaluar el rendimiento de un modelo de aprendizaje automático. Esta matriz presenta de manera detallada la cantidad de predicciones correctas e incorrectas realizadas por el modelo para cada clase, proporcionando una visión clara de su desempeño.
- Exactitud (Accuracy): Es una métrica de evaluación utilizada en problemas de clasificación que mide la proporción de predicciones correctas realizadas por el modelo en relación con el total de casos analizados. Representa la capacidad del modelo para clasificar de manera precisa tanto las instancias positivas como las negativas.
- Precisión (Precision): Es una métrica que evalúa la proporción de casos positivos identificados correctamente por el modelo en relación con el total de casos clasificados como positivos (incluyendo falsos positivos). Esta métrica es especialmente útil cuando el costo de los falsos positivos es alto.
- Sensibilidad (Recall): También conocida como exhaustividad, es una métrica que mide la proporción de casos positivos identificados correctamente por el modelo en relación con el total de casos positivos presentes en el conjunto de datos. Es relevante cuando es crucial minimizar los falsos negativos.
- F1 Score: Es una métrica que combina la precisión y la sensibilidad en un único valor, proporcionando un balance entre ambas. Es especialmente útil en escenarios donde se requiere equilibrar la identificación de casos positivos y la minimización de errores de clasificación, como en problemas de clasificación binaria.
- AUC-ROC: Es una métrica utilizada para evaluar la calidad de las predicciones en problemas de clasificación binaria. Representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos ( $1 - \text{especificidad}$ ) para distintos umbrales de decisión. Un valor cercano a 1 indica un excelente desempeño del modelo.
- Error Absoluto Medio (MAE): Es una métrica comúnmente empleada en problemas de regresión que mide la diferencia absoluta promedio entre las predicciones del modelo y los valores reales. Proporciona una medida directa del error promedio cometido por el modelo.
- Error Cuadrático Medio (MSE): Es una métrica utilizada en problemas de regresión que calcula el promedio de las diferencias al cuadrado entre las predicciones del modelo y los valores reales. Esta métrica penaliza más los errores grandes, lo que la hace sensible a outliers.
- Raíz del Error Cuadrático Medio (RMSE): Es una métrica derivada del MSE que representa la raíz cuadrada del promedio de las diferencias al cuadrado entre las predicciones y los valores reales. Al estar en la misma unidad que la variable objetivo, facilita la interpretación del error.
- Coeficiente de Determinación ( $R^2$ ): Es una métrica que evalúa la capacidad del modelo de regresión para explicar la variabilidad de la variable dependiente. Su valor oscila entre 0 y 1, donde 0 indica que el modelo no explica la variabilidad de los datos, y 1 indica que el modelo explica toda la variabilidad. Un valor cercano a 1 sugiere un ajuste óptimo del modelo a los datos (Databitai, 2023)

## 2.7 Tableros de Control (Dashboard)

La presentación de datos de manera clara y comprensible es fundamental para ello se usan los Tableros de Control (DashBoard en inglés) para ello se realizan con herramientas de visualización concretas, donde se crean los gráficos y tablas, facilitando la interpretación de los resultados del análisis para cualquier empleado del negocio.

### 2.7.1 Herramientas para los Tableros de Control

Existen múltiples industrias y empresas que operan en el sector del análisis de datos, cada una desarrollando sus propias herramientas para la visualización. Estas empresas ofrecen servicios de Business Intelligence a otras organizaciones, independientemente de su tamaño, ya sean pequeñas, medianas o grandes.

En el "Cuadrante Mágico" de Gartner de la Figura 2-2, se puede observar un panorama completo de las empresas que ofrecen estos servicios y sus respectivas herramientas tecnológicas. Cada año, estas empresas compiten entre sí para mejorar su posición en el mercado, los tres proveedores más destacados, junto con sus herramientas respectivas: Qlik (con su plataforma Qlik Sense), Microsoft (con Power BI) y Salesforce (que adquirió Tableau).



Figura 2-2 Evolución de proveedores de BI (2022)

Fuente: <https://gravitar.biz/wp-content/uploads/2022/05/gartner-2022.png>

Un resumen de cada cuadrante según Gartner tenemos por ejemplo los siguientes:

- **Líderes.** - Proveedores con un alto nivel de madurez, se desempeñan bien de acuerdo a la visión actual del mercado y están bien posicionados para el futuro.
- **Visionarios.** - Proveedores que entienden hacia dónde se dirige el mercado o tienen una visión para cambiar sus reglas, pero su capacidad de ejecución aún es limitada.
- **Contendientes.** - Proveedores se centran con éxito en un segmento pequeño, o están dispersos y no innovan por encima de la media ni superan a los demás.
- **Participantes eventuales.** - Proveedores con niveles bajos de puntuación se desempeñan bien hoy e incluso pueden dominar un gran segmento, pero no demuestran poseer una buena comprensión de hacia dónde se dirige el mercado. (Gartner, 2022)

### 2.7.2 Características de los tableros de Control

La presentación de datos de manera clara y comprensible es fundamental. Se realizan con herramientas de visualización concretas, donde se crean los gráficos y dashboards, facilitando la interpretación de los resultados del análisis para cualquier empleado del negocio. Debe representar las KPIs necesarias, realizar una buena limpieza de los datos. Además, se deben presentar las KPIs de forma que sean relevantes, con un contexto. También, es importante una visualización que sea fácil de interpretar, que nos permita tomar decisiones de forma correcta y rápida. (Rivera, 2018)

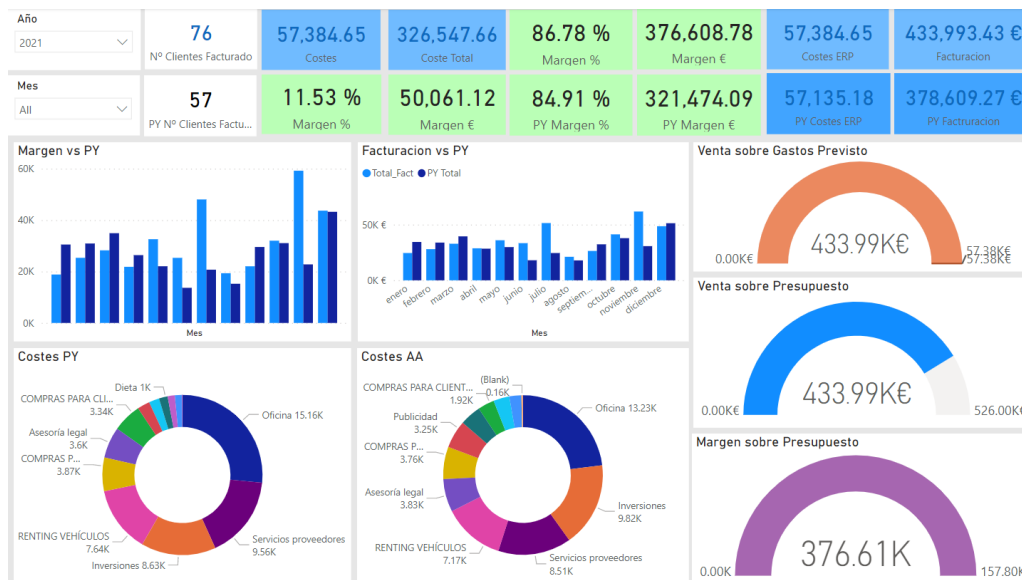


Figura 2-3 Dashboards con Power BI

Fuente: [https://evotic.es/wp-content/uploads/2022/10/dashboard\\_ventas.png](https://evotic.es/wp-content/uploads/2022/10/dashboard_ventas.png)

Como se ilustra en la anterior Figura 2-3, un tablero de control (Dashboard) está compuesto por una variedad de elementos, incluyendo métricas, indicadores de rendimiento, gráficos y tablas. La

configuración específica del tablero varía según las necesidades particulares del negocio y las áreas que requieren apoyo. El objetivo principal es proporcionar información oportuna que permita tomar decisiones eficientes basadas en el análisis detallado de los datos presentados.

Como se mencionaba los Dashboard deben personalizarse según las necesidades y requerimientos del negocio o el área designada. Para permitir a los usuarios explorar la información que requieran como ser por ejemplo:

- **Panel de Resumen:**
  - Muestra las ventas totales en un período específico.
  - Visualiza las ganancias generales después de los costos.
  - Número total de clientes atendidos
- **Análisis de Ventas:**
  - Gráficos o mapas que muestran las ventas desglosadas por regiones geográficas.
  - Análisis de ventas por segmentos de clientes (por ejemplo, nuevos clientes, clientes habituales).
  - Gráficos de líneas que muestran las tendencias de ventas a lo largo del tiempo.
- **Desempeño del Producto:**
  - Top de los productos más vendidos.
  - Rentabilidad de cada producto.
- **Análisis de Clientes:**
  - Información demográfica y datos relevantes sobre los clientes.
  - Porcentaje de clientes que regresan en un período específico.
- **Finanzas:**
  - Desglose de ingresos, costos y beneficios.
  - Visualización del flujo de efectivo y las proyecciones.
- **Marketing:**
  - Desempeño de campañas de marketing y retorno de inversión.
  - Efectividad de diferentes canales de marketing.
- **Cuadros de Mando Personalizados:**
  - Métricas KPI específicas para el éxito de tu empresa.
  - Configurar alertas para eventos importantes.

Finalmente, se deben generar informes detallados y realizar reuniones periódicas con el objetivo de respaldar la toma de decisiones estratégicas y operativas, asegurando que todos los usuarios clave de la empresa cuenten con la información necesaria para optimizar sus procesos y alcanzar los objetivos establecido

### 3 Marco metodológico

La metodología Ralph Kimball propone crear una matriz de negocio que contenga los elementos comunes que son utilizados por los data marts, dimensión, measures, etc., teniendo esta información, el usuario puede desarrollar soluciones que apoyen el análisis a través de los procesos de negocio para la venta cruzada. (Gravitar, 2022)

El análisis de los datos de un negocio se ha convertido en un punto fundamental para aumentar la competitividad, con ello se logra mejorar tiempos de respuesta en la presentación hacia el usuario final, alta confiabilidad, toma de decisiones oportuna y acertada.

#### 3.1 Área de estudio

En la empresa se tiene una amplia selección de herramientas de alta calidad, tanto eléctrica como manual, para todo tipo de proyectos de construcción y bricolaje. Ofrecen marcas reconocidas y asesoramiento experto para ayudarte a encontrar la herramienta adecuada para tus necesidades.

El área de estudio se centra en el análisis del área de Ventas de la empresa comercial de herramientas “MASTER TOOLS”, el cual tiene 4 sucursales en las ciudades de Cochabamba - Av. Beijing y Av. Topater, La Paz - Calle José María Serrano Zona San Sebastián, Oruro – Calle Lira esquina Soria Galvarro y Santa Cruz - Av. La Playa, a una cuadra del 6to. Anillo Zona Cambodromo, como se observa en la Figura 3-1 resaltando las sucursales de las 4 ciudades con color verde.

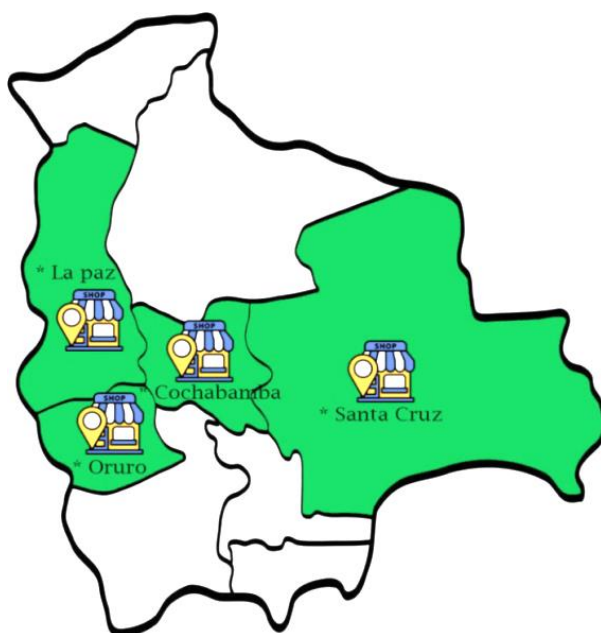


Figura 3-1 Mapa Bolivia – Sucursales de la tienda de Herramienta

Fuente: Elaboración Propia (2024)

### 3.2 Flujograma metodológico

El flujograma metodológico de trabajo se observa en la figura 3-2, el cual favorece como una guía estructurada y secuencial de las actividades desde la comprensión del negocio hasta la presentación de resultados, garantizando un enfoque efectivo de Business Intelligence y Machine Learning.

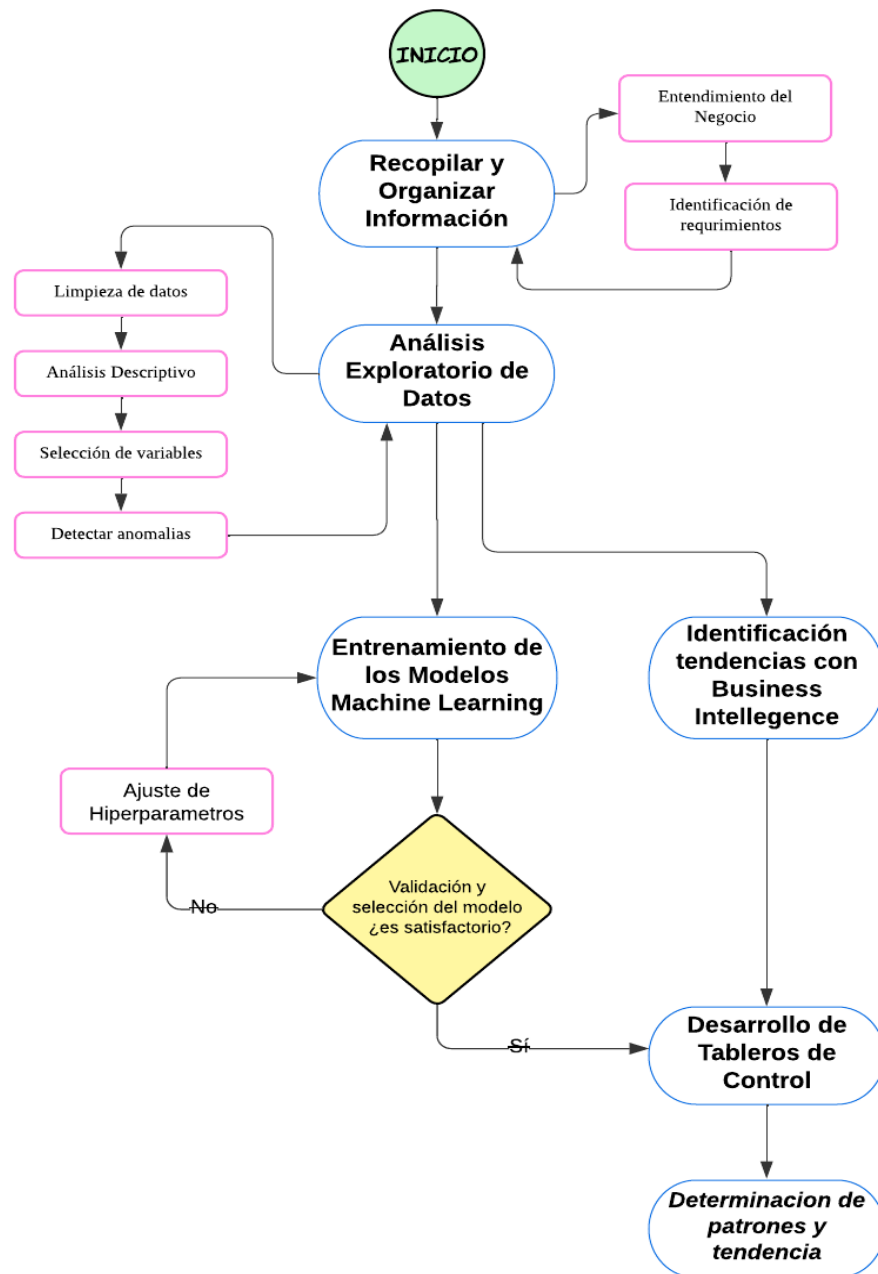


Figura 3-2 Flujograma Metodológico

Fuente: Elaboración Propia (2024)

- **Recopilar y organizar la información:** Se recopila información relevante de la empresa MASTER TOOLS con el propósito de obtener un conocimiento más profundo, ayudado con 2 subprocesos:
  - *Entendimiento del negocio:* conocer el funcionamiento, identificar oportunidades de mejora, detectar posibles deficiencias, entre otros.
  - *Identificación de requerimientos:* tiene como objetivo comprender las necesidades tanto de la empresa como de los vendedores, con el fin de optimizar el análisis y la estructuración de los datos recopilados, permitiendo así una toma de decisiones más eficiente y fundamentada.
- **Análisis Exploratorio de datos:** Contribuye a profundizar el entendimiento de los datos, mejorar la calidad de los datos recolectados, eliminando errores, datos inconsistentes para luego facilitar su análisis con diferentes procesos.
  - El proceso implica la eliminación de datos atípicos y la corrección de errores que puedan distorsionar el análisis y el modelo, así como el manejo de datos faltantes para garantizar la calidad y consistencia de la información.
  - Identificar las variables cualitativas y cuantitativas mediante la realización de análisis descriptivos.
  - Extraer las variables relevantes que sean de interés para la construcción de los modelos.
  - Aplicar cambios o transformaciones a los datos para que se ajusten adecuadamente a los requisitos del análisis
- **Entrenamiento de los Modelo Machine Learning:** Se lleva a cabo una selección meticulosa de modelos en función de los objetivos del proyecto y los datos preparados. Este proceso se estructura en tres subprocesos fundamentales:
  - Ajuste de Hiperparámetros: Es un conjunto de hiperparametros ideal para mejorar la eficiencia del modelo de machine learning con el uso de técnicas como el ‘param\_grid’.
  - Validación de los modelos: Se aplican métricas especializadas según el modelo utilizado, permitiendo evaluar su precisión y desempeño. Este análisis facilita la selección del modelo más adecuado para cumplir con los objetivos estratégicos de la empresa.
- **Identificación de las tendencias con Business Intelligence:** Proceso orientado a la detección de patrones y tendencias en el área de ventas mediante el uso de herramientas de Business Intelligence, mejorando su capacidad competitiva en el mercado y optimizando su rendimiento comercial.
- **Desarrollo Tableros de Control:** Selección de la herramienta tecnológica más adecuada para la construcción de tableros de control (dashboards), con el objetivo de representar de manera efectiva la información clave, preparación del modelo dimensional para organizar los datos y optimizar su análisis.

- **Determinación de patrones y tendencia:** Implementación de gráficos y tablas para representar de manera clara y estructurada los objetivos y requerimientos de la empresa, facilitando una interpretación precisa de la información. Se emplean los indicadores clave de rendimiento (KPIs) más relevantes, permitiendo un seguimiento eficiente y una representación visual de los patrones tendencias en el área de ventas, lo que contribuye a una toma de decisiones informada y estratégica.

### 3.3 Recopilar y organizar la información

Es un aspecto fundamental para clarificar el contexto del proyecto. Esta etapa incluye una evaluación del estado actual del manejo de datos y de las herramientas existentes, lo que permitirá obtener los requerimientos necesarios, su utilización y los pasos a seguir para completar el proyecto de manera eficiente y efectiva.

#### 3.3.1 Fuentes de Información Primaria

Con el propósito de comprender la problemática y la situación actual de la empresa comercial de herramientas MASTER TOOLS, se elaboró una encuesta con el objetivo es proponer soluciones pertinentes mediante la aplicación de Business Intelligence y el uso de las técnicas de Machine Learning más adecuadas.

La encuesta se puede observar en el Anexo 1, del cual se observó algunas debilidades, datos relevantes y la comprensión general de cómo opera la empresa MASTER TOOLS.

Para calcular el tamaño de la muestra de una población de 57 vendedores activos dentro la empresa MASTER TOOLS, se puede utilizar la fórmula general para poblaciones finitas:

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

Dónde:

n = Tamaño de la muestra

N= Tamaño de la población (57 vendedores en este caso)

Z= Valor de la distribución normal para el nivel de confianza deseado (1.96 para 95%, 1.645 para 90%, 2.576 para 99%)

p = Proporción esperada de la población

q = 1-p

e = Margen de error permitido (en este caso par aun nivel de confianza del 95% se usó e=5%)

Aplicando la formula y reemplazando los valores, para una población de 57 vendedores, con un nivel de confianza del 95% y un margen de error del 5%, el tamaño de muestra necesaria es de aproximadamente 49 personas para ser encuestadas.



Como se observa en la Tabla 3-1 la encuesta se realizó con un formulario online a los vendedores activos de la empresa, teniendo un total de 54 encuestados. La encuesta se puede observar en el Anexo 1.

Tabla 3-1 Ficha resumen Encuesta a vendedores

<b>Objetivo</b>	Identificar la problemática del área de ventas de un negocio de venta de herramientas, para la implementación de una solución BI
<b>Población</b>	Personal del área de ventas (57 personas)
<b>Muestra</b>	54 personas
<b>Herramienta</b>	Google Forms
<b>Fecha</b>	5 y 6 de octubre 2024
<b>Localización</b>	Ciudades de Santa cruz, Oruro, Cochabamba y La Paz

**Fuente:** Elaboración Propia (2024)

### 3.3.2 Fuentes de Información Secundaria

Se observa en la Figura 3-3 un archivo de Excel que son los datos históricos del área de ventas extraídos del sistema utilizado por la empresa en todas sus sucursales. Esta información abarca la totalidad del año 2022 y parte del año 2023, específicamente desde enero hasta junio. La planilla se puede observar por completo en el Anexo 2.

ID	Deposito	Vendedor	Cliente	FechaVer	FechaCien	Cantida	PrecioLoca	PrecioVen	PrecioFinal	ComisionUni	Comisi
1333	Santa Cruz		DAM/SM/UC7	2022-01-10	2022-01-11	1	1315	1650	1650	150	
695	Santa Cruz		2	2022-01-10	2022-01-11	1	837	950	950	80	
381	Santa Cruz		SA	2022-01-10	2022-01-11	1	75	125	125	30	
396	Santa Cruz		ced AC	2022-01-10	2022-01-11	1	1285,1	1400	1300	50	
130	Santa Cruz		ced AL	2022-01-10	2022-01-11	1	435	550	550	50	
1253	Santa Cruz		ced RA	2022-01-11	2022-01-12	1	74,5	107	96		
1085	Santa Cruz		ced RA	2022-01-11	2022-01-12	2	55,5	75	72		
562	Santa Cruz		ced CF	2022-01-11	2022-01-12	1	130	185	160		
687	Santa Cruz		ced SE	2022-01-11	2022-01-12	1	504,96	650	650	50	
1337	Santa Cruz		ced RA	2022-01-11	2022-01-12	1	535	770	770	50	
442	Santa Cruz		ced AL	2022-01-11	2022-01-12	1	1160	1500	1500	300	
114	Santa Cruz		VIN	2022-01-11	2022-01-12	1	443,5	550	550	50	
132	Santa Cruz		GL	2022-01-11	2022-01-12	1	1091	1300	1300	100	
396	Santa Cruz		ced CF	2022-01-11	2022-01-12	1	1285,1	1300	1300	50	
256	Santa Cruz		ced RA	2022-01-11	2022-01-12	1	186	280	280	40	
1328	Santa Cruz		ced RE	2022-01-11	2022-01-12	1	626,4	800	800	70	

Figura 3-3 DataSet Área Ventas archivo Excel

Fuente: Captura de Excel, Elaboración Propia (2024)

La información extraída del sistema de ventas de la Tienda de Herramientas se encuentra en un archivo Excel (VentasRango.xlsx) con 20883 filas y 18 columnas, las cuales se detalla a continuación:

1. "idProd": Identificador general de la herramienta.
2. "Descripción": Nombre y detalle específico de la herramienta
3. "ID": Identificador de la herramienta según el depósito de origen.
4. "Sucursal": Ubicación donde se encuentra la tienda o sucursal (categórica: "Oruro", "La Paz", "Cochabamba", "Santa Cruz")
5. "Vendedor": Nombre del vendedor encargado de la tienda o del vendedor freelancer.
6. "Cliente": Nombre del cliente
7. "FechaVenta": Fecha cuando se realizó la venta de la herramienta
8. "FechaCierre": Fecha cuando se realizó la transferencia de caja a la cuenta bancaria.
9. "Cantidad": Cantidad que se vendió de la misma herramienta (numérico entero).
10. "PrecioCompra": Precio de compra de la herramienta (numérico decimal).
11. "PrecioVenta": Precio de venta de la herramienta (numérico decimal).
12. "PrecioFinal": Precio Final de la venta de las herramientas después de gastos como (descuento o comisiones a los vendedores)
13. "comisiónUnit": monto percibido por la venta de la herramienta vendida por el vendedor (numérico decimal)
14. "comisionTotal": monto total para cancelar al vendedor por venta realizada (cantidad \* comisionUnit)
15. "Total": Monto total de la venta (cantidad \* PrecioFinal)
16. "Utilidad": Monto de utilidad por venta realizada (Total - comisionTotal)
17. "Marca": Nombre de la Marca de fabricación de la herramienta (categórica)
18. "Proveedor": Nombre de la empresa de quien se importan las herramientas (categórica)

### **3.3.3 Entendimiento del negocio**

La obtención de información del área de ventas es rápida, precisa y eficiente para el proceso de toma de decisiones, una vez que se implementen los análisis y herramientas necesarios. Los resultados al finalizar todo el proceso se mostrarán a los usuarios finales tanto los vendedores como los administradores de datos que ayudarán para identificar patrones y tendencias en el comportamiento del consumidor para la empresa MASTER TOOLS. Esto permitirá a la empresa anticipar cambios en la demanda y ajustar su oferta en consecuencia.

### **3.3.4 Identificación de Requerimientos**

Entender las necesidades de los clientes como también de los empleados dentro de un negocio es fundamental para ofrecer una mejor experiencia. Por tanto con el apoyo de los resultados de la encuesta realizada a los vendedores activos de la empresa del cual se obtuvo los siguientes requerimientos que se observan en la tabla 3- 2, para apoyar en la optimización y decisiones del área de las ventas.

Tabla 3-2 Requerimientos según los empleados

NRO	Requerimientos	Indicador
1	Identificar productos más y menos vendidos en general	Producto/cantidad
2	Identificar productos más y menos vendidos por sucursal	Producto/ Sucursal
3	Conocer las herramientas más vendidos según la marca	Producto/cantidad/marca
4	Conocer las herramientas menos vendidos según la marca	Producto/Marca
5	Conocer las Ventas Totales generadas anualmente	Ventas
6	Graficar la evolución de las herramientas según los meses.	Producto/fecha
7	Conocer las Utilidades generadas anualmente	Utilidades
8	Conocer el Total de Clientes y Total de Vendedores	Clientes/Vendedores
9	Conocer el margen porcentual de las Ventas Anualmente	Ventas/porcentaje
10	Identificar los mejores empleados	Empleados/ventas
11	Conocer las ventas totales por meses según la sucursal	Ventas/sucursal/ fecha
12	Calcular el porcentaje de ventas según la sucursal	Sucursal/porcentaje

Fuente: Elaboración propia (2024)

### 3.4 Análisis exploratorio de datos

Para facilitar el análisis exploratorio de datos, se optó por utilizar el entorno de trabajo de “Google Colab”. Como se observa en la Figura 3-4 Mediante el comando 'drive.mount', se estableció una conexión directa con Google Drive, especificando la ruta exacta del archivo que contiene el data set del área de ventas de la empresa, para luego cargarlo en un dataframe, el cual facilito su análisis usando el lenguaje de Python."

## Información del area de Ventas – importación de los datos

```
[ ] # Lectura de los datos desde el drive de google
from google.colab import drive
drive.mount('/gdrive', force_remount=True)

ruta = "/gdrive/MyDrive/dataset/VentasRango.xlsx"
df_data = pd.read_excel(ruta)
```

Mounted at /gdrive

Figura 3-4 Importar data set en entorno Google Colab

Fuente: Elaboración Propia (2024)

En la Figura 3-5 se observa mediante el comando info() que los datos fueron cargado exitosamente y se tiene en total 20883 registros de las ventas de la empresa y 18 columnas que son las variables para analizar .

```
#informacion de los datos en general
df_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20883 entries, 0 to 20882
Data columns (total 18 columns):
#   Column          Non-Null Count  Dtype
---  -
0   IdProd           20883 non-null  object
1   Descripcion      20883 non-null  object
2   ID               20883 non-null  int64
3   Sucursal         20883 non-null  object
4   Vendedor         20883 non-null  object
5   Cliente          20883 non-null  object
6   FechaVenta       20883 non-null  object
7   FechaCierre      20781 non-null  object
8   Cantidad          20839 non-null  float64
9   PrecioCompra     20883 non-null  float64
10  PrecioVenta       20882 non-null  float64
11  PrecioFinal       20870 non-null  float64
12  ComisionUnit       9617 non-null   float64
13  ComisionTotal     9617 non-null   float64
14  Total             20838 non-null  float64
15  Utilidad          20883 non-null  float64
16  Marca             20883 non-null  object
17  Proveedor         20883 non-null  object
dtypes: float64(8), int64(1), object(9)
memory usage: 2.9+ MB
```

Figura 3-5 Importar data set en entorno Google Colab

Fuente: Elaboración Propia (2024)

### 3.4.1 Limpieza de los datos

Uno los pasos para la limpieza es analizar los valores nulos o faltantes que pueden causar problemas cuando se use el modelo de machine learning. Por ejemplo en la Figura 3-6. Se observó ve que las variables 'ComisiónUnit' y 'ComisiónTotal' tienen 11266 valores nulos, también se tiene algunos valores nulos en las variables de Fecha Cierre, Total, Cantidad, Precio Final.

#### ANÁLISIS DE NULOS

```
[5] df_data.isna().sum().sort_values(ascending = False)
```

	0
ComisionTotal	11266
ComisionUnit	11266
FechaCierre	102
Total	45
Cantidad	44
PrecioFinal	13
PrecioVenta	1
IdProd	0
Marca	0

Figura 3-6 Análisis de valores nulos

Fuente: Elaboración Propia (2024)

Como se observa en la figura 3-7 dado que los campos 'ComisiónUnit' y 'ComisiónTotal' presentaban una alta proporción de valores nulos, se optó por modificarlo por el valor '0' en este caso. Esta decisión se fundamenta con la lógica de que, si no se registra una comisión, se asume que la venta fue asignada directamente a la sucursal y no generó una comisión para el vendedor. Asimismo, se procedió a eliminar los registros con valores nulos en el campo 'Venta', dado que la ausencia de un valor en este campo indica que la transacción no se completó de manera exitosa.

```
# Modificar valores nulos de ComisiónTotal y ComisiónUnit por el valor numerico de "0"
df_data["ComisionUnit"] = df_data["ComisionUnit"].fillna(0)
df_data["ComisionTotal"] = df_data["ComisionTotal"].fillna(0)

# Eliminar filas con valores nulos en la columna 'Total' ya que la venta no es valida
df_data = df_data.dropna(subset=['Total'])
```

Figura 3-7 Corrección de valores nulos

Fuente: Elaboración Propia (2024)

En la Figura 3-8 se ha detectado la presencia de inconsistencias en las variables sucursales y proveedores. Se observan múltiples registros para una misma sucursal o proveedor, pero con variaciones en la escritura de sus nombres (por ejemplo, 'Sta cruz' y 'Santa Cruz' para la misma sucursal) también en la variable 'Proveedor' (por ejemplo, 'Roghur S.A.' y 'Roghur', Forza S.R.L y Forza SRL). Estas situaciones generan ambigüedades en el análisis, ya que el sistema interpreta estas variaciones como sucursales y proveedores diferentes, por tanto se debe corregir todas las que presentes el mismo error tipográfico.

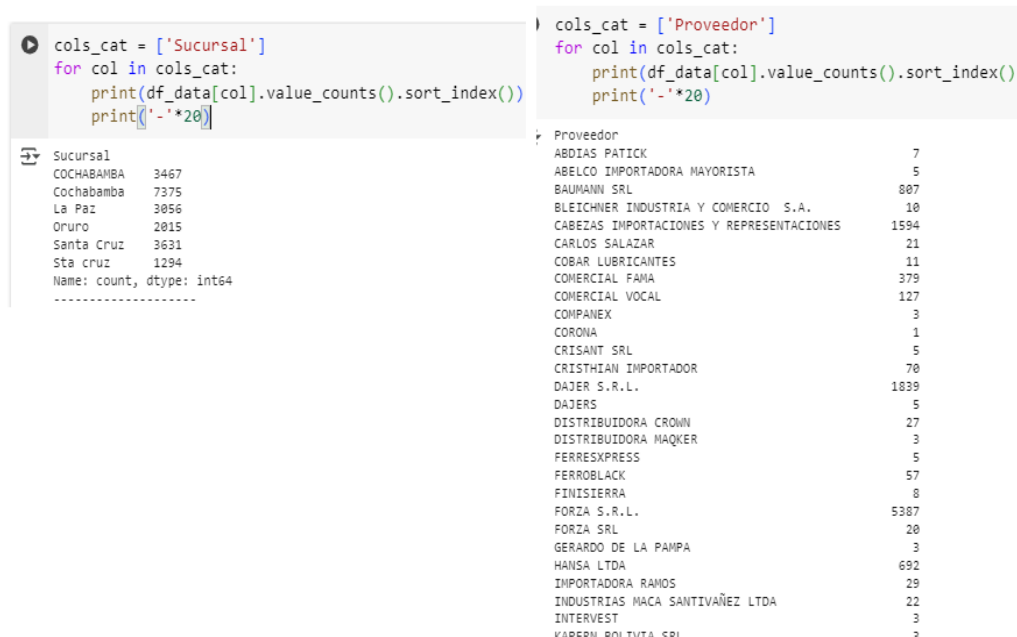


Figura 3-8 Errores Tipográficos

Fuente: Elaboración Propia (2024)

Por esta razón se procedió a corregir la escritura de los nombres de las sucursales y los proveedores. Se cambió a mayúscula todos los nombres sucursales y se reemplazó 'Sta Cruz' por 'SANTA CRUZ', también se corrigieron los nombres de los proveedores, como 'Forza S.R.L.' por 'Forza SRL', asegurando así que cada entidad fuera representada de manera única en el conjunto de datos con el código de la Figura 3-9.

```
[ ] #Poner todo en mayuscula
df_data['Sucursal'] = df_data['Sucursal'].str.upper()

# Reemplazar 'STA CRUZ' por 'SANTA CRUZ' en la columna 'Sucursal'
df_data['Sucursal'] = df_data['Sucursal'].replace('STA CRUZ', 'SANTA CRUZ')

[ ] #Corregir los nombres de los proveedores
df_data['Proveedor'] = df_data['Proveedor'].str.replace('VADIKO S.R.L.', 'VADIKO', regex=False)
df_data['Proveedor'] = df_data['Proveedor'].str.replace('ROGHUR S.A.', 'ROGHUR', regex=False)
df_data['Proveedor'] = df_data['Proveedor'].str.replace('SALCEDO IMPORTACIONES S.R.L.', 'SALCEDO IMPORTACIONES S.R.L.', regex=False)
df_data['Proveedor'] = df_data['Proveedor'].str.replace('DAJER S.R.L.', 'DAJERS', regex=False)
df_data['Proveedor'] = df_data['Proveedor'].str.replace('FORZA S.R.L.', 'FORZA SRL', regex=False)
df_data['Proveedor'] = df_data['Proveedor'].str.replace('MUEBLETEKA IMPORTACIONES Y REPRESENTACIONES', 'MUEBLETEK', regex=False)
```

Figura 3-9 Corrección errores tipográficos

Fuente: Elaboración Propia (2024)

En la Figura 3-10 se realizó la conversión de tipo de datos de las variables 'Cantidad' y 'FechaVenta', con el objetivo de garantizar la coherencia y la precisión en el análisis. Se transformó la variable 'Cantidad' de tipo float a tipo entero (int), y la variable 'FechaVenta' de tipo objeto a tipo fecha (datetime). Estas modificaciones fueron necesarias para asegurar la correcta interpretación de los datos para su análisis posterior.

```
#Cambiar tipo de dato de decimal a entero
df_data['Cantidad'] = df_data['Cantidad'].astype(int)

#Corregir el formato de fecha a datetime
df_data['FechaVenta'] = pd.to_datetime(df_data['FechaVenta'], format="%Y-%m-%d")
```

Figura 3-10 Cambiar tipo de dato variable Cantidad y Fecha Venta

Fuente: Elaboración Propia (2024)

### 3.4.2 Análisis descriptivo

En la Figura 3-11 se verificó el comportamiento de las variables categóricas de los cuales se evidenció que la empresa cuenta con 4 sucursales, tiene registrado 3390 herramientas vendidas, 121 vendedores y 8284 clientes.

```

0] # Subniveles de cada variable categórica
cols_cat = ['Descripcion', 'Sucursal', 'Vendedor', 'Cliente', 'Marca',
            'Proveedor']

for col in cols_cat:
    print(f'Columna {col}: {df_data[col].nunique()} subniveles')

Columna Descripcion: 3390 subniveles
Columna Sucursal: 4 subniveles
Columna Vendedor: 115 subniveles
Columna Cliente: 8284 subniveles
Columna Marca: 120 subniveles
Columna Proveedor: 44 subniveles

```

Figura 3-11 Variables Categóricas

Fuente: Elaboración Propia (2024)

En la Figura 3-12 también se observó que la sucursal con más registro de ventas está en la ciudad de Cochabamba con 10842, la marca más vendida es la Truper con 4158 unidades y el producto con más cantidad vendida es mascara de soldadura de la marca Lynus. Este resultado preliminar podría cambiar si se toma en cuenta el rendimiento de la utilidad.

```

#Como se comportan de variables categoricas
df_data.describe(include=['O'])

```

	IdProd	Descripcion	Sucursal	Vendedor	Cliente	FechaCierre	Marca	Proveedor
count	20838	20838	20838	20838	20838	20736	20838	20838
unique	3298	3390	4	115	8284	453	120	44
top	00012196.4	MASCARA DE SOLDA AUTO.C/ CONTROLADOR LYNUS MSL...	COCHABAMBA	Tienda Cochabamba	CLIENTE	2022-09-21	TRUPER	FORZA SRL
freq	411	401	10842	9644	790	1499	4158	5407

Figura 3-12 Variables Categóricas


Fuente: Elaboración Propia (2024)

En la Figura 3-13 se ve el comportamiento de las Variables Numéricas, los cuales denota valores negativos en las variables 'ComisionUnit' y 'ComisionTotal', los cuales se podrían eliminar o modificar si los valores corresponden a un error humano y si estos no afectarían con el modelo de Machine Learning que se vaya a utilizar.



```
[ ] def estadisticos_cont(num):
    #Calculamos describe
    estadisticos = num.describe().T
    #Añadimos la mediana
    estadisticos['median'] = num.median()
    #Reordenamos para que la mediana esté al lado de la media
    estadisticos = estadisticos.iloc[:, [0,1,8,2,3,4,5,6,7]]
    #Lo devolvemos
    return(estadisticos)
```

```
▶ estadisticos_cont(df_data.select_dtypes('number'))
```



	count	mean	median	std	min	25%	50%	75%	max
Cantidad	20838.0	1.353254	1.00	2.066784	0.50	1.00	1.00	1.0000	100.0
PrecioLocal	20838.0	380.458355	99.08	691.990847	1.00	23.25	99.08	436.6000	12043.0
PrecioVenta	20838.0	487.602683	135.00	925.495531	1.04	32.00	135.00	560.0000	22440.0
PrecioFinal	20838.0	484.188019	130.00	862.646920	2.00	32.00	130.00	560.0000	14500.0
ComisionUnit	20838.0	11.984362	0.00	318.786116	-20200.00	0.00	0.00	20.7500	4681.0
ComisionTotal	20838.0	10.316749	0.00	629.900834	-80800.00	0.00	0.00	25.0000	4681.0
Total	20838.0	531.496194	166.50	996.225934	2.00	36.00	166.50	630.0000	38500.0
Utilidad	20838.0	134.375190	41.60	813.174771	-15067.20	10.00	41.60	114.2375	88039.0

Figura 3-13 Comportamiento Variables Numéricas

Fuente: Elaboración Propia (2024)

### 3.4.3 Selección de las variables

Las variables que no se tomara en cuenta para el posterior análisis para continuar con el proyecto son:

- ID, por son ser relevantes para el análisis ya que el sistema otorga una numeración de los registros automáticamente.
- IdProd, por tener menos relevancia a la hora de identificar a los productos, por lo que se tomó en cuenta mejor la variable de 'Descripción'
- FechaCierre, no otorgan ninguna relevancia saber qué fecha se depositó el registro de la caja, al no tener un periodo definido lo cual podría ser al día siguiente, a la semana o dentro de 1 mes.

Como se observa en la Figura 3-14 con el comando 'drop', se eliminara las 3 variables o las columnas que no son relevantes, así también poder tener mayor espacio en la memoria de la computadora para las variables que si son necesarios para el análisis.

```
df_data = df_data.drop(columns=['ID', 'FechaCierre', 'IdProd'])
```

Figura 3-14 Eliminación de variables irrelevantes

Fuente: Elaboración Propia (2024)

En la Figura 3-15 con la matriz de correlación se observó el grado de relación que tienen las variables numéricas, de las cuales tanto el PrecioVenta y Utilidad están fuertemente relacionadas con un valor de 0,55 lo cual se tomara en cuenta para el análisis del modelo de Machine Learning

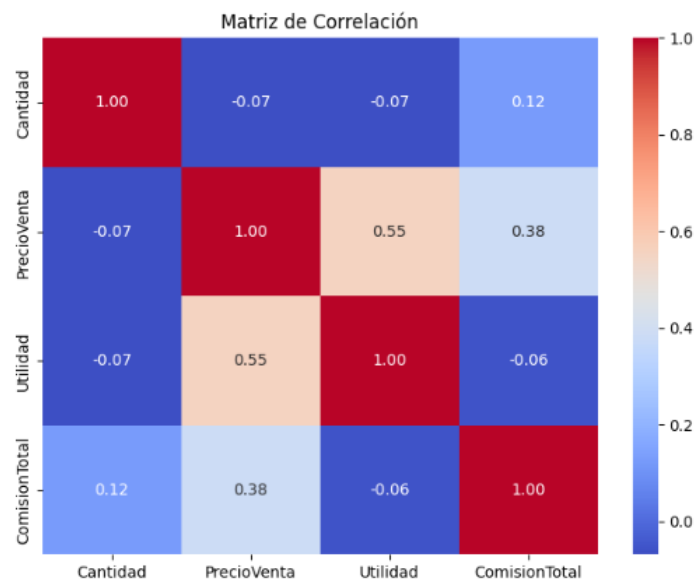


Figura 3-15 Matriz de correlación

Fuente: Elaboración Propia (2024)

#### 3.4.4 Detección de valores atípicos

En la Figura 3-16, se llevó a cabo una verificación de la presencia de valores atípicos en las variables numéricas mediante el código correspondiente. La existencia de estos valores extremos podría comprometer la precisión necesaria de los modelos de Machine Learning.

```
# Generar gráficas individuales pues las variables numéricas
# están en rangos diferentes
cols_num = ['Cantidad', 'PrecioCompra', 'PrecioVenta', 'PrecioFinal', 'ComisionUnit',
            'ComisionTotal', 'Total', 'Utilidad']

fig, ax = plt.subplots(nrows=8, ncols=1, figsize=(8,30))
fig.subplots_adjust(hspace=0.5)

for i, col in enumerate(cols_num):
    sns.boxplot(x=col, data=df_data, ax=ax[i])
    ax[i].set_title(col)
```

Figura 3-16 Valores atipicos

Fuente: Elaboración Propia (2024)

En la Figura 3-17, se presentan diagramas de caja que revelo la presencia de valores atípicos algo peculiares en dos variables, la variable 'ComisionTotal' muestra valores de -20000 y -80,000, mientras que la variable 'Utilidad' presenta valores de 20,000 y 80,000. Esta discrepancia sugiere la posibilidad de un error de escritura. Los diagramas de las demás variables pueden consultarse en el Anexo 3.

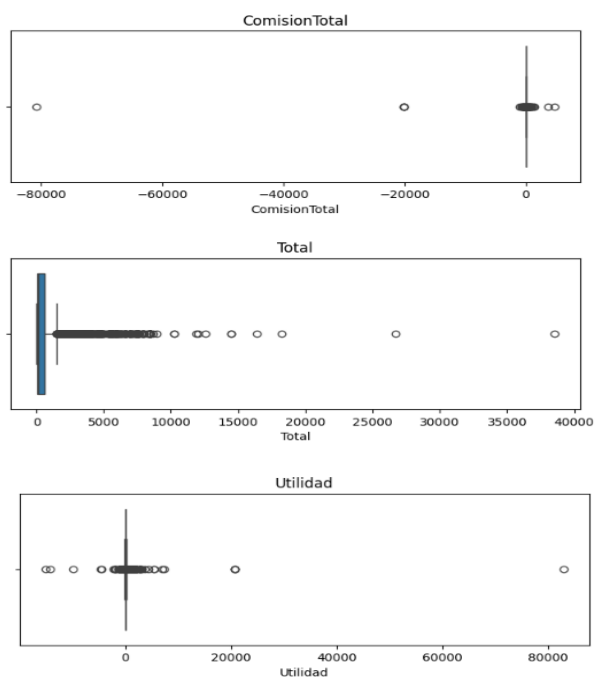


Figura 3-17 Grafico de Cajas valores atipicos

Fuente: Elaboración Propia (2024)

En la figura 3-18 se observó que efectivamente los outsiders son errores de registro en la columna 'precioVenta' en lugar de digitar 2240 se puso 22440 aumentando el valor y afectando el cálculo de las columnas de ComisionTotal y de la Utilidad.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Sucursal	Vendedor	Cliente	FechaVenta	Cantidad	PrecioCompra	PrecioVenta	PrecioFinal	ComisionUnit	ComisionTotal	Total	Utilidad	Marca	Proveedor	
3924	SANTA CRUZ	mamier	Oferton Santa Cruz	2022-04-13 00:00:00	1	1721	22440	2240	-20200	-20200	2240	20719	ELECTROP	FORZA SRL	
3925	SANTA CRUZ	mamier	Oferton La Paz	2022-04-13 00:00:00	1	1721	22440	2240	-20200	-20200	2240	20719	ELECTROP	FORZA SRL	
3926	SANTA CRUZ	mamier	Oferton Cochabamba	2022-04-13 00:00:00	1	1721	22440	2240	-20200	-20200	2240	20719	ELECTROP	FORZA SRL	
4122	SANTA CRUZ	mamier	David Garcia	2022-04-18 00:00:00	1	1721	22440	2240	-20200	-20200	2240	20719	ELECTROP	FORZA SRL	
4294	SANTA CRUZ	mamier	Martin Mier	2022-04-21 00:00:00	4	1721	22440	2240	-20200	-80800	8960	82876	ELECTROP	FORZA SRL	

Figura 3-18 Error en registro de la variable PrecioVenta

Fuente: Elaboración Propia (2024)

En la figura 3-19 se procedió a eliminar los demás valores atípicos de las variables, usando como método el IQR, el cual define la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), para la disminución de los valores atípicos dejando en total 12355 registros y 15 variables.

```
# Eliminación de outliers
df = df_data

#función para eliminar outliers
def remove_outliers_iqr(dataframe, columns):
    for column in columns:
        Q1 = dataframe[column].quantile(0.25)
        Q3 = dataframe[column].quantile(0.75)
        IQR = Q3 - Q1

        # Calcular límites inferior y superior
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR

        # Filtrar los outliers
        dataframe = dataframe[(dataframe[column] >= lower_bound) & (dataframe[column] <= upper_bound)]

    return dataframe

# Especificar las columnas a limpiar
columnas_a_limpiar = ['Cantidad', 'Total', 'Utilidad', 'PrecioVenta', 'ComisionTotal']

# Limpiar el DataFrame
df_sin_outliers = remove_outliers_iqr(df, columnas_a_limpiar)

print(f"Nueva forma del DataFrame: {df_sin_outliers.shape}")
```

Nueva forma del DataFrame: (12355, 15)

Figura 3-19 Corrección de valores Atípicos

Fuente: Elaboración Propia (2024)

Una vez finalizado el proceso, se dispone de un conjunto de datos más limpio, el cual se exportará para su posterior utilización con la herramienta de visualización. Este nuevo archivo será denominado "VentasRangoLimpio.xlsx", tal como se ilustra en la Figura 3-20.

```
[ ] ruta = "/gdrive/MyDrive/dataset/VentasRangoLimpio.xlsx"
data.to_excel(ruta, index=False)
```

Figura 3-20 Exportar datos para crear los tableros de control

Fuente: Elaboración Propia (2024)

### 3.5 Entrenamiento de Modelos Machine Learning

Se implementaron técnicas de Machine Learning utilizando dos modelos: de regresión logística y de Random Forest. El primero, al ser un modelo de regresión, predice una probabilidad que posteriormente se traduce en una clasificación, de manera similar al segundo modelo. En el contexto de nuestro proyecto, el objetivo es predecir si una venta se clasificará como alta, media o baja. Esto permitirá optimizar las ventas realizadas y determinar qué características deben ser consideradas y cuáles deben ser excluidas en el área de las ventas de la empresa de herramientas.

#### 3.5.1 Preparación datos para los Modelos

Para preparar los datos de modelo antes se realizara un copia de respaldo que sera usada solo para los modelos de Machine learning, como se observa en la Figura 3-21, con el nombre de 'df\_ml' esto para evitar conflictos con los datos que serán utilizados en la visualizaciones.

```
[56] #Copia del dataset para las tecnicas de Machine Leraning
df_ml = df_sin_outliers.copy()
```

Figura 3-21 Copia del dataset para los modelos

Fuente: Elaboración Propia (2024)

En la Figura 3-22 se observó los percentiles de la variable 'Utilidad', para tener conocimiento del comportamiento de su distribución.

```
percentiles = df_ml['Utilidad'].quantile([0.25, 0.5, 0.75])
print(percentiles)
```

0.25	6.6
0.50	15.5
0.75	49.4

Name: Utilidad, dtype: float64

Figura 3-22 verificar percentiles variable Utilidad

Fuente: Elaboración Propia (2024)

### 3.5.2 Modelamiento con Random Forest

En la figura 3-23 se observa el proceso de preparación del modelo usando RandomForest, donde se procedió a normalizar las variables Marca con la función Categorical para tener mayor presión y también se convierte las variables categóricas de Marca, Sucursal usando la función Dummy esto par estandarizar los datos en el modelo y que se pueda facilitar la interpretación con los algoritmos del modelo. Para el modelo de Random Forest se utilizó 3 variables marca, Sucursal y Precio Venta, y como Variable Objetivo esta 'Nivel\_ventas' la cual se cree en el nuevo dataset de df\_ml con las categorías de Baja. Media y Alta.

```
# usando Modelo RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier

# Convertir 'Marca' a variable categórica
df_ml['Marca'] = pd.Categorical(df_ml['Marca'])

# Preparar los datos
# Dividir datos en entrenamiento y prueba
X = df_ml[['Marca', 'Sucursal', 'PrecioVenta']]
y = df_ml['Nivel_ventas']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Convertir la columna 'Marca' a variables dummy (codificación one-hot)
# Crear variables dummy para la columna 'Marca' en X_train
dummies_train = pd.get_dummies(X_train['Marca'], drop_first=True)
X_train = pd.concat([X_train.drop('Marca', axis=1), dummies_train], axis=1)

# Crear variables dummy para la columna 'Sucursal' en X_train
dummies_train = pd.get_dummies(X_train['Sucursal'], drop_first=True)
X_train = pd.concat([X_train.drop('Sucursal', axis=1), dummies_train], axis=1)

# Crear variables dummy para la columna 'Marca' en X_test
dummies_test = pd.get_dummies(X_test['Marca'], drop_first=True)
X_test = pd.concat([X_test.drop('Marca', axis=1), dummies_test], axis=1)

# Crear variables dummy para la columna 'Sucursal' en X_test
dummies_test = pd.get_dummies(X_test['Sucursal'], drop_first=True)
X_test = pd.concat([X_test.drop('Sucursal', axis=1), dummies_test], axis=1)

# Asegurar que X_train y X_test tengan las mismas columnas
missing_cols = set(X_train.columns) - set(X_test.columns)
for col in missing_cols:
    X_test[col] = 0
X_test = X_test[X_train.columns]
```

Figura 3-23 modelado Random Forest

Fuente: Elaboración Propia (2024)

En la Figura 3-24 se realizó el entrenamiento el modelo con el conjunto de entrenamiento llamado TRAIN los cuales son el 80% del data set, donde también se ajustan los parámetros del modelo, para luego para a la validación de la precisiones.

```
# Entrenar el modelo de Random Forest
rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train, y_train)
```

Figura 3-24 Entrenamiento Modelo Random Forest

Fuente: Elaboración Propia (2024)

### 3.5.3 Modelamiento con Regresión Logística

Para el siguiente modelo de Machine Learning se usó la de regresión Logística, con la diferencia que se tomó solo las variables 'Marca' y 'Precio venta', para observar su funcionamiento y rendimiento como se observa en la Figura 3-25 donde igualmente se tomó como variable objetivo 'Nivel\_ventas'

```
# Convertir marca a variable categórica
df_ml['Marca'] = pd.Categorical(df_ml['Marca'])

#Preparar datos

# Dividir datos en entrenamiento y prueba
X = df_ml[['Marca', 'PrecioVenta']]
y = df_ml['Nivel_ventas']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Convertir marca a variable numérica
# Crear variables dummy para la columna 'Marca' sin asignarla directamente a 'X_train['Marca']'
dummies = pd.get_dummies(X_train['Marca'], drop_first=True)

# Concatenar las variables dummy con el DataFrame original y eliminar la columna original 'Marca'
X_train = pd.concat([X_train.drop('Marca', axis=1), dummies], axis=1)

# Generar variables dummy para la columna 'Marca' en X_test
dummies_test = pd.get_dummies(X_test['Marca'], drop_first=True)

# Concatenar las variables dummy con el DataFrame original X_test y eliminar la columna 'Marca'
X_test = pd.concat([X_test.drop('Marca', axis=1), dummies_test], axis=1)
```

Figura 3-25 Modelado Regresión Logística

Fuente: Elaboración Propia (2024)

En la Figura 3-26 se realizó el entrenamiento el modelo regresión logística, usando los datos de entrenamiento X\_train y Y\_train.

```
#Entrenar modelo

# Entrenar modelo de regresión logística
model = LogisticRegression()
model.fit(X_train, y_train)
```

Figura 3-24 Entrenamiento Modelo Regresión Logística

Fuente: Elaboración Propia (2024)

### 3.5.4 Métricas de validación Modelo Random Forest

Para la validación del modelo Random Forest, se emplearon las métricas pertinentes conforme a lo establecido. Tal como se ilustra en la Figura 3-27, se utilizaron las métricas de Exactitud (accuracy), F1-Score y la matriz de confusión.

```
# Evaluar el modelo
# Predecir resultados
y_pred = rf_model.predict(X_test)

# Reporte de clasificación
print("Reporte de clasificación:")
print(classification_report(y_test, y_pred))

# Calcular precisión
print("==== Calcular precisión ====")
accuracy = accuracy_score(y_test, y_pred)
print(f'accuracy: {accuracy:.2f}\n')

f1 = f1_score(y_test, y_pred, average='weighted') # 'weighted' para manejar desbalance
print(f'F1-Score: {f1:.2f}\n')

# Matriz de confusión
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index=model.classes_, columns=model.classes_)
print("Matriz de Confusión:\n", cm_df)
```

Figura 3-25 Evaluación Modelo random Forest

Fuente: Elaboración Propia (2024)

### 3.5.5 Métricas de Validación Modelo Regresión Logística

Para el modelo de regresión logística, se emplearon las mismas métricas de Exactitud (accuracy), F1-Score y matriz de confusión que se utilizaron en el anterior modelo de Random Forest, como se puede observar en



la Figura 3-28. Esto se realizó con el propósito de comparar los resultados y determinar cuál de las dos opciones sería la más adecuada para su aplicación en la empresa MASTER TOOLS.

```
#Evaluar modelo

# Predecir resultados
y_pred = model.predict(X_test)

# Evaluar precisión
accuracy = accuracy_score(y_test, y_pred)
print(f'Precisión: {accuracy:.2f}\n')

f1 = f1_score(y_test, y_pred, average='weighted')
print(f'F1-Score: {f1:.2f}\n')

# Evaluar clasificación
print(classification_report(y_test, y_pred))

# Evaluar Matriz de confusión
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index=model.classes_, columns=model.classes_)
print("Matriz de Confusión:\n", cm_df)
```

Figura 3-26 Evaluación Modelo regresión Logística

Fuente: Elaboración Propia (2024)

### 3.6 Desarrollo de tableros de control (Dashboard)

La selección de la herramienta tecnológica para llevar a cabo el análisis y la visualización de datos fue fundamental para asegurar la eficiencia y efectividad del proceso. Se consideró la información registrada en las encuestas, así como la necesidad de contar con un conjunto de datos (dataset de ventas) limpio y de alta calidad. Con esta base, se desarrollaron los tableros de control requeridos por la empresa, que permiten la visualización de los indicadores de rendimiento y las tendencias, facilitando así la comprensión del comportamiento de los productos. Esto, a su vez, beneficia la toma de decisiones y la implementación de estrategias de manera anticipada.

#### 3.6.1 Selección Herramienta BI

En el contexto de los proveedores líderes en visualización de datos, se elaboró un cuadro comparativo que resume las características más destacadas de las tres empresas con mayor rendimiento, como se muestra en la Tabla 3-3. Esta comparación permitió evaluar sus características y las funcionalidades de cada proveedor, facilitando así la selección de la herramienta para la Visualizaciones teniendo en cuenta su adaptación a las necesidades específicas del proyecto y la empresa.

Tabla 3-3 Tabla comparativa de los proveedores

Características	POWER BI	TABLEAU	QLIK
Costo	Ofrece una versión gratuita, el plan mínimo es de 10\$ por usuario/mes	Ofrece un periodo de prueba gratis, el plan mínimo es de 15\$ por usuario/mes	Ofrece un periodo de prueba gratis, el plan mínimo es de 20\$ por usuario/mes
Capacidad almacenamiento	Solo 10 GB en la nube, por un costo mínimo.	100 GB, pero con un costo medio-alto.	50Gb, con un costo medio-bajo.
Interfaz de Usuario	Intuitiva y familiar, especialmente para usuarios de Microsoft.	Altamente visual, fácil de usar.	Interfaz sencilla con arrastrar y soltar.
Comunidad y Soporte	Gran comunidad de usuarios y amplia documentación, foros y video tutoriales.	Comunidad activa, recursos en línea y soporte de Tableau, foros y video tutoriales.	Comunidad activa y recursos de soporte, incluyendo foros y documentación extensa
Seguridad	Seguridad sólida, con los datos de sus clientes.	Herramientas de seguridad avanzadas.	Enfoque en seguridad y control de acceso.
Integración con Otros Sistemas	Integración estrecha con productos de Microsoft.	Integración con diversas aplicaciones y sistemas.	Integración con múltiples sistemas y plataformas.
Conectividad de Datos	Conexión a numerosas fuentes de datos.	Amplia variedad de fuentes de datos.	Soporte para diversas fuentes de datos.
Escalabilidad	Escalable y adecuado para empresas de todos los tamaños.	Escalable para grandes empresas	Escalabilidad para grandes conjuntos de datos.
	Fuente: <a href="https://powerbi.microsoft.com/es-es/">https://powerbi.microsoft.com/es-es/</a>	Fuente: <a href="https://www.tableau.com/es-es">https://www.tableau.com/es-es</a>	Fuente: <a href="https://www.qlik.com/es-es/">https://www.qlik.com/es-es/</a>

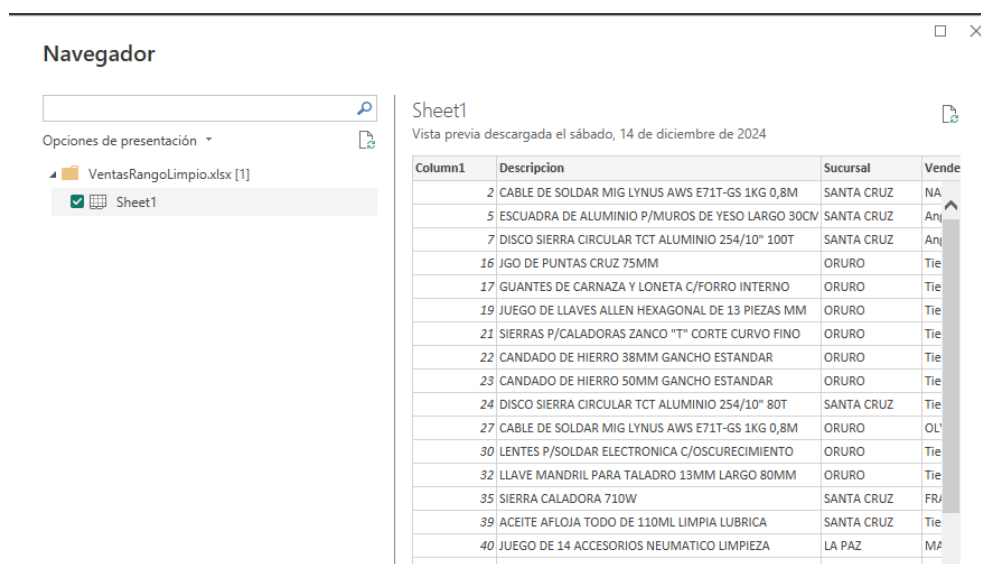
Fuente: Adaptado de las plataformas oficiales (2024)

Por lo tanto, se optó por utilizar Power BI debido a su bajo costo y a la disponibilidad de una versión gratuita. Esta herramienta destaca por su facilidad de uso, ofreciendo una interfaz intuitiva que permite

a los usuarios interactuar con ella sin complicaciones. Además, Power BI proporciona capacidades de análisis avanzado al permitir la integración de lenguajes como Python y DAX.

### 3.6.2 Obtención de los datos

En este proceso, se realizó la importación del nuevo archivo que contiene el dataset de ventas de la empresa, esta información ya está tratada, limpiada y sin ningún tipo de errores, listo para comenzar su organización de los datos y registros que se observan en la Figura 3-29 para su posterior modelado dimensional el cual contara con una tabla de hechos y las demás tablas de dimensiones con sus respectivas relaciones.



Column1	Descripción	Sucursal	Vendedor
2	CABLE DE SOLDAR MIG LYNUS AWS E71T-GS 1KG 0,8M	SANTA CRUZ	NA
5	ESCUADRA DE ALUMINIO P/MUROS DE YESO LARGO 30CM	SANTA CRUZ	Anj
7	DISCO SIERRA CIRCULAR TCT ALUMINIO 254/10" 100T	SANTA CRUZ	Anj
16	JGO DE PUNTAS CRUZ 75MM	ORURO	Tie
17	GUANTES DE CARNAZA Y LONETA C/FORRO INTERNO	ORURO	Tie
19	JUEGO DE LLAVES ALLEN HEXAGONAL DE 13 PIEZAS MM	ORURO	Tie
21	SIERRAS P/CALADORAS ZANCO "T" CORTE CURVO FINO	ORURO	Tie
22	CANDADO DE HIERRO 38MM GANCHO ESTANDAR	ORURO	Tie
23	CANDADO DE HIERRO 50MM GANCHO ESTANDAR	ORURO	Tie
24	DISCO SIERRA CIRCULAR TCT ALUMINIO 254/10" 80T	SANTA CRUZ	Tie
27	CABLE DE SOLDAR MIG LYNUS AWS E71T-GS 1KG 0,8M	ORURO	OL'
30	LENTES P/SOLDAR ELECTRONICA C/OSCURECIMIENTO	ORURO	Tie
32	LLAVE MANDRIL PARA TALADRO 13MM LARGO 80MM	ORURO	Tie
35	SIERRA CALADORA 710W	SANTA CRUZ	FR/
39	ACEITE AFLOJA TODO DE 110ML LIMPIA LUBRICA	SANTA CRUZ	Tie
40	JUEGO DE 14 ACCESORIOS NEUMATICO LIMPIEZA	LA PAZ	MA

Figura 3-27 Importación datos en Power BI

Fuente: Elaboración Propia (2024)

### 3.6.3 Modelo Dimensional

Como se muestra en la figura 3-30 se diseñó un esquema en estrella, teniendo una tabla de hechos donde se organizó toda la información de las ventas generadas, de las cuales se creó cinco tablas de dimensiones y una tabla de hechos (Fact\_Ventas) que se sitúa en el centro conectadas mediante la relación de muchos a uno con las demás tablas.

- Dim\_Cliente: información de los clientes.
- Dim\_Producto: información de las herramientas que ofrece la empresa
- Dim\_Sucursal: información de las ubicaciones de la sucursales
- Dim\_Vendedor: información del equipo del área de ventas
- Dim\_Calendario: información de fechas en que se registró las ventas
- Fact\_Ventas: datos que relación con las demás tablas, adema tienen los montos de cada venta.

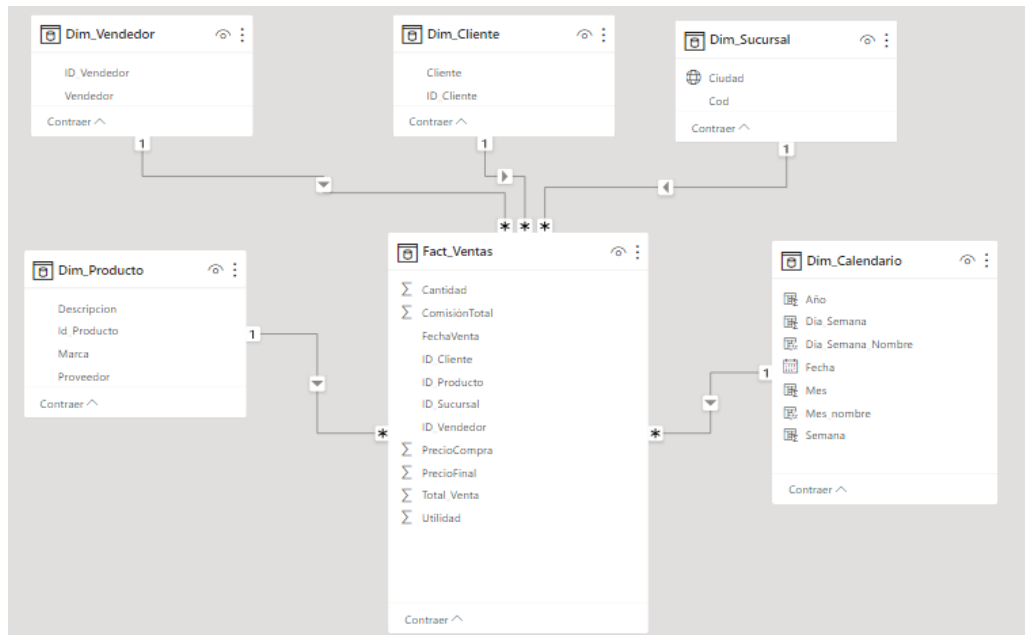


Figura 3-28 Esquema de estrella Tabla de Hechos

Fuente: Elaboración propia en Power BI (2024)

### 3.7 Determinación de patrones y tendencias.

Teniendo actualmente solo 15 columnas o variables los cuales se observa en la Figura 3-31 y por su lado se tienen 12355 registros de las ventas, para obtener las visualizaciones que requiere la empresa.

```
column_names = df_sin_outliers.columns.tolist()
print(column_names)

['Descripcion', 'Sucursal', 'Vendedor', 'Cliente', 'FechaVenta', 'Cantidad', 'PrecioCompra', 'PrecioVenta',
 'PrecioFinal', 'ComisionUnit', 'ComisionTotal', 'Total', 'Utilidad', 'Marca', 'Proveedor']
```

Figura 3-29 Columnas del nuevo dataset de ventas

Fuente: Elaboración propia en Power BI (2024)

Con los datos mencionados anteriormente se identificó algunas de las tendencias más importantes que ayudarían a la empresa a tomar mejores decisiones para la optimización en el área de las ventas como ser:

- Tendencias por Producto: conocer que productos son los más y los menos vendidos, analizando su comportamiento según las ventas generadas.
- Rendimiento por Marca: Saber de la marca con mejor aceptación de los consumidores con mayores ventas.
- Comportamiento de Sucursal: Conocer cuál de las 4 sucursales tiene un mejor comportamiento según los meses.

- Rendimiento de Vendedores: Analizar que vendedores tienen mejor rendimiento y su contribución a las ventas totales en la empresa.
- Tendencias Temporales: Evaluar la fluctuación de las ventas fluctúan a lo largo del año, identificando picos en festividades o eventos especiales.
- Comportamiento del Cliente: Identificar y clasificar que clientes más rentables para la empresa según la sucursal.

Estas tendencias pueden ser visualizadas a través de dashboards interactivos y reportes, facilitando la comprensión de datos complejos y apoyando la toma de decisiones informadas en la estrategia de ventas y marketing.

Con lo que se presenta de forma visual e interactiva para los usuarios que lo necesiten como es el caso de la empresa MASTER TOOLS. Como se ve en el ejemplo de la Figura 3-32 que va representando los datos obtenidos con tarjetas con la cantidad de vendedores y clientes, gráficos con información de los montos, mapas de las sucursales, barras y líneas del comportamiento de las ventas, según sea necesario o eficiente. Los cuales también responden a los requerimientos del negocio y tendencias usando la herramienta de Power BI.

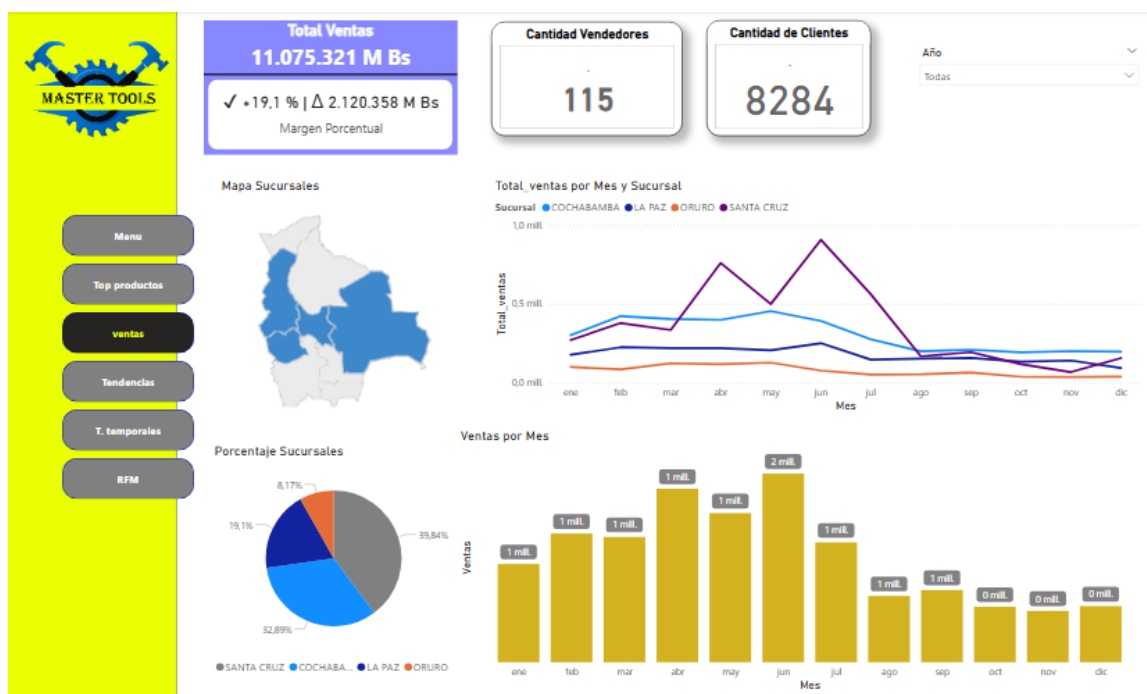


Figura 3-30 Visualizaciones con Power BI

Fuente: Elaboración Propia Power BI (2024)

## 4. Resultados y Discusión

En este capítulo se presentan exhaustivamente los resultados obtenidos de los productos de la empresa MASTER TOOLS, derivados de los procedimientos llevados a cabo en el capítulo anterior. Se detallan los resultados de las evaluaciones, de los modelos, así como también de los requerimientos establecidos desde la perspectiva de Business Intelligence, en relación con el conjunto de datos de las ventas registradas durante los años 2022 y parte de 2023.

### 4.1. Resultados Recopilación y organización de la información

Este proceso permitió una mejor percepción del funcionamiento y del desempeño de la empresa “MASTER TOOLS”, así también se logró determinar algunas tendencias, patrones, oportunidades y también observar algunas debilidades presentes los cuales se debe hacer.

#### 4.1.1 Resultados de la encuesta

En la Figura 4-1 Se observa que según los resultados de la primera pregunta del cuestionario, se tiene que el proceso de ventas dentro de la empresa tiene ciertas características de organización pero que le falta mejorar o automatizar ciertos procesos. Con una votación de 59,3% donde los vendedores denotaron una falta de organización para mejorar las ventas.

1. ¿Cómo describiría el proceso de ventas actual en la empresa?  
54 respuestas



Figura 4-1 Resultado procedimiento de ventas.

Fuente: Elaboración Propia (2024)

En la Figura 4-2 Se observa como resultado de la encuesta que la empresa y los vendedores usan la herramienta EXCEL como apoyo en las ventas con un 55,6% y también casi la mitad de los vendedores solo usa cuadernos físicos para llevar un registro, lo cual es algo preocupante ya que no sería muy eficiente, ya que en la mayoría de los casos no se podría realizar cálculos automáticos o generar algún análisis de datos con la información que almacenan ya que no estaría organizada ni uniforme, tendrían información incompleta.

## 2. ¿Qué herramientas o sistemas utiliza el equipo de ventas para gestionar clientes y ventas?

54 respuestas

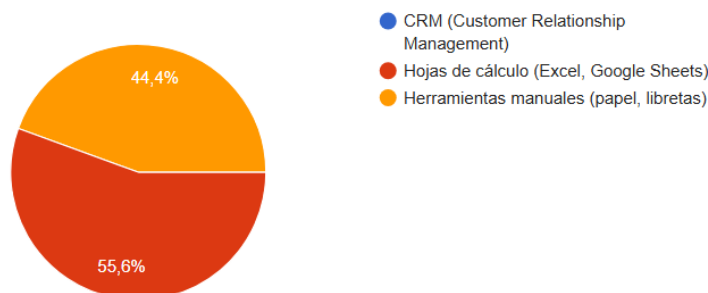


Figura 4-2 Herramientas de apoyo en las ventas

Fuente: Elaboración Propia (2024)

En la figura 4-3 se observa que los vendedores tienen un deficiente manejo para el seguimiento de los clientes, ya que toda la información lo manejan por medio de whatsapp o por sus celulares con un porcentaje del 70,4%, esto haría que los clientes potenciales no sean evaluados con la importancia necesaria, para que otros vendedores o el gerente pueda dar seguimiento y fidelización a dichos clientes.

## 3. ¿Cómo se realiza el seguimiento a los clientes potenciales y existentes?

54 respuestas

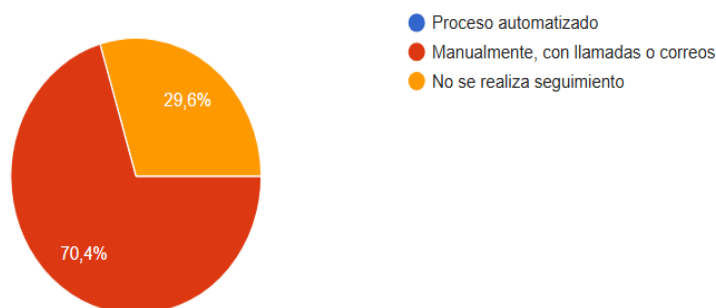


Figura 4-3 Seguimiento a los clientes

Fuente: Elaboración Propia (2024)

En la figura 4-4 se observa otra debilidad en la empresa, al no existir un análisis de las ventas y no tener objetivos o metas claras, podría llevar a cabo que el gerente realice promoción o enfoque la fuerza de los vendedores a la comercialización de productos que no sean rentables, perdiendo así el liderazgo de alguna marca o su producto estrella contra su competencia.

#### 4. ¿Cómo se establecen las metas de ventas en la empresa?

54 respuestas

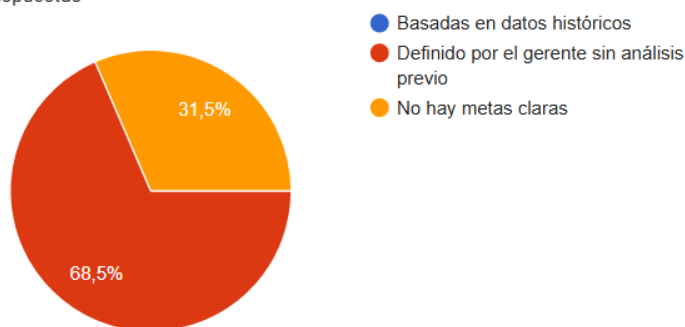


Figura 4-4 Metas y objetivos en las ventas

Fuente: Elaboración Propia (2024)

En la Figura 4-5 se observa que a pesar que tiene una revisión mensual de cómo van los resultados de las ventas, aún existe un 38,9% que no logra asistir a dichas reuniones o evaluaciones, Por lo tanto se recomendaría tener alguna herramienta que permita ver en tiempo real los resultados de las ventas o realizar la reuniones de forma online.

#### 5. ¿Con qué frecuencia se revisa el desempeño del equipo de ventas?

54 respuestas

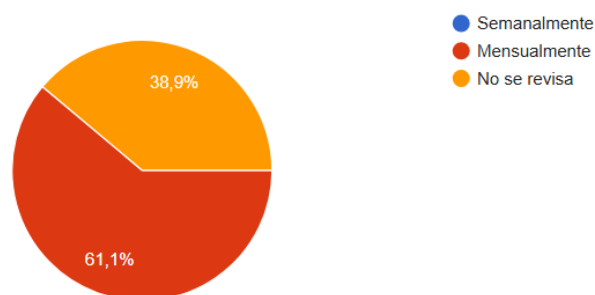


Figura 4-5 Revisión desempeño equipo de ventas

Fuente: Elaboración Propia (2024)

Se observa que la Figura 4-6 que un total de 44 empleados toma como métrica de venta solo el total de las ganancias generadas, para ello faltaría mejorar sobre la satisfacción del cliente para tener la fidelización del mismo o ganar más clientes a la competencia. Otorgando a los mejores clientes de cada mes un paquete promocional o algún incentivo.



**6. ¿Qué métricas se utilizan para medir el éxito de las ventas? (Seleccione todas las que apliquen)**

54 respuestas

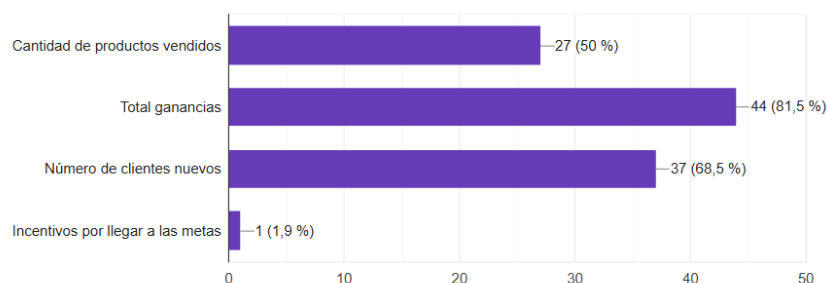


Figura 4-6 Métricas para evaluar las ventas

Fuente: Elaboración Propia (2024)

Se observa en la Figura 4-7 una deficiente capacitación al equipo de ventas, sobre todo al nuevo personal que ingresa a veces sin conocimientos sobre maquinas eléctricas sus características y las funcionalidad para diferentes áreas de trabajo (carpintería, plomería, albañilería, entre otros). Esto podría generar una mala experiencia con los clientes sobre todo con algún cliente frecuente lo que haría el abandono total con la empresa.

**7. ¿Recibe el equipo de ventas capacitación regular?**

54 respuestas

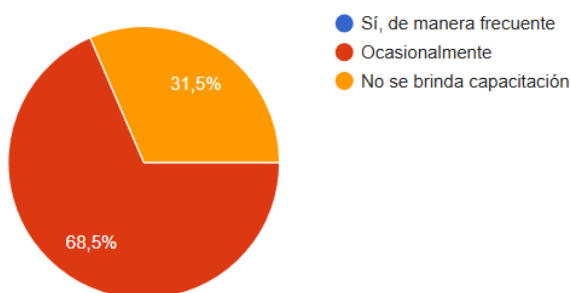


Figura 4-7 Capacitación al equipo de ventas

Fuente: Elaboración Propia (2024)

Se observa en la figura 4-8 que los empleados con un 48,1% solicitaron una capacitación sobre las técnicas de ventas para mejorar su rendimiento y atraer más clientes potenciales, también con un 35,2% sobre el manejo de herramientas tecnológicas para tener un mejor control sobre sus ventas y sus clientes.

### 8. ¿Qué tipo de capacitación considera más útil para el equipo de ventas?

54 respuestas

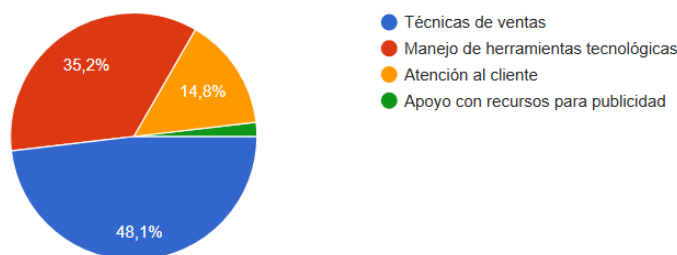


Figura 4-8 Tipo capacitación para las ventas

Fuente: Elaboración Propia (2024)

Se observa en la Figura 4-9 que uno de los retos que tienen los vendedores es que no tienen la información precisa ya sea de las herramientas más demandadas, las herramientas que tengan alguna promoción que ofrece su proveedor. Otra de las falencias es sobre la capacitación para el manejo y comprensión de las máquinas eléctricas (tipos de taladros, sierras, cortadoras de mesa, entre otros), esto hace que el vendedor no tenga las respuestas necesarias para poder cerrar una venta con algún cliente que es experto en el uso de alguna herramienta para algún área en específica (carpintería, plomería, etc).

### 9. ¿Cuáles son los principales retos que enfrenta el área de ventas? (Seleccione todas las que apliquen)

54 respuestas

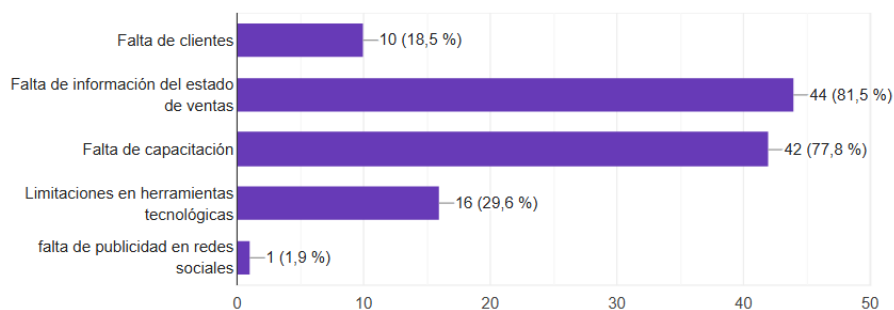


Figura 4-9 Retos para el área de ventas

Fuente: Elaboración Propia (2024)

En la Figura 4-10 se observa las sugerencias para aumentar las ventas están Implementar herramientas tecnológicas con el fin de automatizar la información para todo el equipo de ventas, también de optimizar el seguimiento y analizar a los posibles clientes frecuentes, por otro lado también hay gran aceptación para las capacitaciones para mejorar el conocimiento del campo de las herramientas eléctricas

10. ¿Qué sugerencias tiene para mejorar en el área de ventas? (Seleccione todas las que apliquen)

54 respuestas

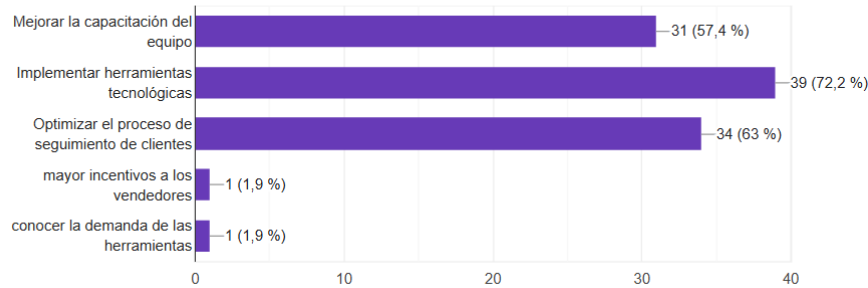


Figura 4-10 Sugerencias para el área de ventas

Fuente: Elaboración Propia (2024)

Como resultado de la encuesta tenemos que la mayor parte de los vendedores carece de información actualizada sobre los productos a ofertar, no cuentan con una herramienta que apoye un análisis sobre el estado de las ventas en general, la mayoría indica que es necesario incrementar las ventas con algún tipo de plan de marketing enfocados en atraer nuevos clientes con el apoyo de publicaciones en redes sociales y dar importancia a capacitaciones regulares sobre técnicas de Ventas para todos los vendedores.

4.2. Resultados Análisis exploratorio de datos

Como resultados de los análisis exploratorios de datos, como se observa en la Figura 4-11 se importó correctamente los datos de la ventas históricas del periodos 2022 y 2023, teniendo en total de 20883 registros o filas y 18 variables o columnas.

IDProd	Descripcion	ID	Sucursal	Vendedor	Cliente	FechaVenta	FechaCierre	Cantidad	PrecioCompra	PrecioVenta	PrecioFinal	ComisionUnit	ComisionTotal	Total	Utilidad	Marca	Proveedor
0	133001	ASPIRADORA 1400W BTA 133001	1333	Santa Cruz		2022-01-10	2022-01-11	1.0	1315.00	1650.0	1650.0	150.0	150.0	1650.0	185.00	BTA	VADIKO S.R.
1	4007325	COMPRESOR EINHELL DE 24LITROS	695	Santa Cruz		2022-01-10	2022-01-11	1.0	637.00	950.0	950.0	80.0	80.0	950.0	33.00	EINHELL	BAUMANN SF
2	00013245.3	CABLE DE SOLDAR MIG LYNUS 4HRS E71T-GS 1KG 0.8MM	381	Santa Cruz		2022-01-10	2022-01-11	1.0	75.00	125.0	125.0	30.0	30.0	125.0	20.00	LYNUS	FORZA S.R.
3	FTS18001	SIERRA DE MESA 1800W	396	Santa Cruz		2022-01-10	2022-01-11	1.0	1285.10	1400.0	1300.0	50.0	50.0	1300.0	-35.10	LYNUS	FORZA S.R.
4	MO.5221	ESCALERA MULTIFUNCION 4X3	130	Santa Cruz		2022-01-10	2022-01-11	1.0	435.00	550.0	550.0	50.0	50.0	550.0	65.00	TRAMONTINA	CABEZAS IMPORTACIONES REPRESENTACIONE
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20878	FI-44477	CARRITO DE CARGA PROFESIONAL DE 300 KG CON ROD...	22894	Santa Cruz		2023-07-08	2023-07-10	1.0	665.26	899.0	899.0	NaN	NaN	899.0	233.74	PIERO	SARUMADI S.R.
20879	HKTHP11801	JUEGO DE HERRAMIENTAS 108 PIEZAS CON TALADRO 680W	13730	COCHABAMBA		2023-07-08	NaN	1.0	555.00	750.0	700.0	NaN	NaN	700.0	145.00	INGCO	DAJER S.R.
20880	MCD1211155	DISCO DE CORTE METAL 4 1/2" GALLETA CAJA DE 25...	7083	COCHABAMBA		2023-07-08	NaN	2.0	4.07	5.5	6.0	NaN	NaN	12.0	3.06	INGCO	DAJER S.R.
20881	MS001	SIERRA DE INGLETE TELESCOPICA 1800W MS001	1557	Oruro		2023-07-08	NaN	1.0	1221.00	1650.0	1650.0	NaN	NaN	1650.0	429.00	MAKUTE	MIOMETAL S.R.
20882	VPK3708	BOMBA DE AGUA 370W	23801	Oruro		2023-07-08	NaN	1.0	234.30	340.0	340.0	NaN	NaN	340.0	105.70	INGCO	DAJER S.R.

Figura 4-11 Registros de las ventas importadas

Fuente: Elaboración Propia (2024)

### 4.2.1 Resultados Limpieza de datos

En la limpieza de los datos comenzando con los valores nulos o registro vacíos, se logró corregir todas las variables para que no haya problemas con registros incompletos, como se muestra en la figura 4-12 donde todas las variables indican que tienen cero valores vacíos.

	0
Descripcion	0
Sucursal	0
Vendedor	0
Cliente	0
FechaVenta	0
Cantidad	0
PrecioCompra	0
PrecioVenta	0
PrecioFinal	0
ComisionUnit	0
ComisionTotal	0
Total	0
Utilidad	0
Marca	0

Figura 4-12 Limpieza de valores nulos

Fuente: Elaboración Propia (2024)

Como se observa en la Figura 4-13 se modificó el tipo de datos de las variables Cantidad y Fecha Venta para que no se tenga conflicto a la hora de interpretar o realizar algún cálculo y se tenga mayor precisión en los resultados.

```
<class 'pandas.core.frame.DataFrame'>
Index: 12355 entries, 2 to 20882
Data columns (total 15 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Descripcion     12355 non-null  object
1   Sucursal        12355 non-null  object
2   Vendedor        12355 non-null  object
3   Cliente         12355 non-null  object
4   FechaVenta      12355 non-null  datetime64[ns]
5   Cantidad         12355 non-null  int64
6   PrecioCompra    12355 non-null  float64
7   PrecioVenta     12355 non-null  float64
8   PrecioFinal     12355 non-null  float64
9   ComisionUnit    12355 non-null  float64
10  ComisionTotal   12355 non-null  float64
11  Total           12355 non-null  float64
12  Utilidad        12355 non-null  float64
13  Marca           12355 non-null  object
14  Proveedor       12355 non-null  object
dtypes: datetime64[ns](1), float64(7), int64(1), object(6)
memory usage: 1.5+ MB
```

Figura 4-13 Cambiar tipo de dato variable Id y fecha

Fuente: Elaboración Propia (2024)

### 4.2.1 Resultados Análisis descriptivo

En la figura 4-14 se presentan las variables categóricas, destacando que el número de vendedores que trabajaron durante los períodos de 2022 y 2023 asciende a 121. Además, se observa la existencia de cuatro sucursales. La herramienta más vendida corresponde al proveedor 'Forza SRL', con una frecuencia de 5,407 unidades. La marca preferida por los clientes es 'Truper', mientras que la herramienta con el mayor volumen de ventas es la máscara de soldadura auto c/controlador Lynus.

```
#Como se comportan de variables categoricas
df_data.describe(include=['O'])
```

	Descripcion	Sucursal	Vendedor	Cliente	Marca	Proveedor
count	20838	20838	20838	20838	20838	20838
unique	3390	4	115	8284	120	44
top	MASCARA DE SOLDAAUTO.C/ CONTROLADOR LYNUS MSL...	COCHABAMBA	Tienda Cochabamba	CLIENTE	TRUPER	FORZA SRL
freq	401	10842	9644	790	4158	5407

Figura 4-14 Descriptivos variables categóricas

Fuente: Elaboración Propia (2024)

Como se visualiza en la Figura 4-15, dentro de las subcategorías de la variable "Sucursal", se identifican las cuatro ciudades: Santa Cruz, Oruro, La Paz y Cochabamba. De estas, Cochabamba destaca como la sucursal con la mayor afluencia de ventas, acumulando más de 10,000 registros

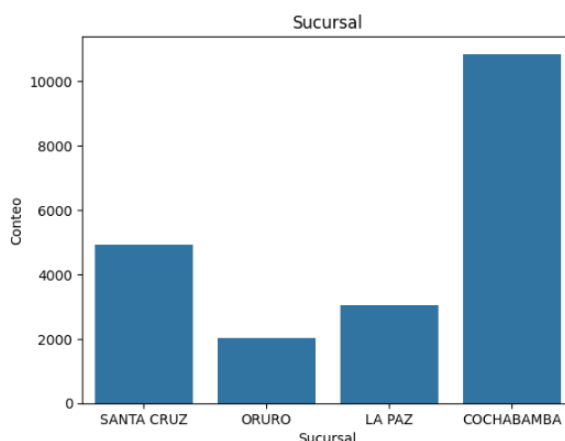


Figura 4-15 Subcategoría de la variable Sucursal

Fuente: Elaboración Propia (2024)

En la figura 4-16 se presentan los resultados correspondientes a las variables numéricas. Se destaca que, en promedio, los clientes adquieren una unidad de herramienta por transacción, mientras que el

promedio de venta es de 130 Bs. Asimismo, se observa un valor mínimo en la variable "Utilidad" de -15,067.2, lo cual sugiere un error en el registro durante el proceso de venta. De manera similar, se han encontrado valores de "Comisión" inferiores a -750 Bs. También se observa que los cuartiles 25 y 50 de la comisiones son de cero, lo que sugiere que la mayor parte de la ventas son otorgadas a la Sucursal y no así a los vendedores. Por consiguiente, se procedió a la eliminación de estos valores atípicos para garantizar la coherencia y precisión de los resultados en los análisis posteriores.

```
estadisticos_cont(df_data.select_dtypes('number'))
```

	count	mean	median	std	min	25%	50%	75%	max
<b>Cantidad</b>	20838.0	1.353297	1.00	2.067014	1.00	1.00	1.00	1.0	100.0
<b>PrecioCompra</b>	20838.0	380.458355	99.08	691.990847	1.00	23.25	99.08	436.6	12043.0
<b>PrecioVenta</b>	20838.0	482.755769	135.00	861.172376	1.04	32.00	135.00	560.0	14500.0
<b>PrecioFinal</b>	20838.0	484.188019	130.00	862.646920	2.00	32.00	130.00	560.0	14500.0
<b>ComisionUnit</b>	20838.0	16.900860	0.00	59.551017	-750.00	0.00	0.00	21.0	4681.0
<b>ComisionTotal</b>	20838.0	18.170190	0.00	70.048407	-750.00	0.00	0.00	25.0	4681.0
<b>Total</b>	20838.0	531.496194	166.50	996.225934	2.00	36.00	166.50	630.0	38500.0
<b>Utilidad</b>	20838.0	93.900944	37.25	277.979480	-15067.20	9.50	37.25	102.0	7294.0

Figura 4-16 Descriptivos variables numéricas

Fuente: Elaboración Propia (2024)

#### 4.2.1 Resultados selección de variables

Como resultado del proceso de selección de variables relevantes para el proyecto, se redujo el conjunto original de 18 columnas a 15, las cuales se consideran de mayor importancia para continuar con los análisis, lo cuales se observan en la Figura 4-17.

	Descripcion	Sucursal	Vendedor	Cliente	FechaVenta	Cantidad	PrecioCompra	PrecioVenta	PrecioFinal	ComisionUnit	ComisionTotal	Total	Utilidad	Marca	Proveedor
2	CABLE DE SOLDAR MIG LYNUS AWS E71T-GS 1KG 0,8M	SANTA CRUZ	NATALIA ROMERO	SANTIAGO	2022-01-10	1	75.00	125.0	125.0	30.0	30.0	125.0	20.00	LYNUS	FORZA SRL
5	ESCUADRA DE ALUMINIO P/MUROS DE YESO LARGO 30CM	SANTA CRUZ	Angie Johanna Carmona Salcedo	RAUL FLORES	2022-01-11	1	74.50	107.0	96.0	0.0	0.0	96.0	21.50	TRUPER	SALCEDO IMPORTACIONES S.R.L.
7	DISCO SIERRA CIRCULAR TCT ALUMINIO 254/10" 100T	SANTA CRUZ	Angie Johanna Carmona Salcedo	CRISTOBAL VARGAS	2022-01-11	1	130.00	185.0	160.0	0.0	0.0	160.0	30.00	TOLSEN	MIOMETAL S.R.L.
16	JGO DE PUNTAS CRUZ 75MM	ORURO	Tienda Oruro	antonio	2022-01-12	1	12.66	19.0	20.0	0.0	0.0	20.0	7.34	TRUPER	SALCEDO IMPORTACIONES S.R.L.
17	GUANTES DE CARNAZA Y LONETA C/FORRO INTERNO	ORURO	Tienda Oruro	ALFREDO CORRALES	2022-01-12	1	23.00	35.0	30.0	0.0	0.0	30.0	7.00	TRUPER	SALCEDO IMPORTACIONES S.R.L.

12355 rows x 15 columns

Figura 4-17 Variables con mayor relevancia

Fuente: Elaboración Propia (2024)

### 4.2.1 Resultados valores atípicos

En la figura 4-18 se presentan los diagramas de cajas, excluyendo los valores atípicos de mayor relevancia que podrían influir de manera negativa y errónea en el análisis del proyecto, especialmente en las variables "Comisión Total", "Total" y "Utilidad", las cuales son fundamentales para la empresa. Los restantes diagramas de cajas pueden consultarse en el Anexo 4. Los resultados obtenidos en los diagramas de cajas reflejan valores más confiables, aunque se identificaron algunos outliers de menor relevancia, los cuales podrían representar pérdidas mínimas para la empresa, pero no se consideran inusuales en el contexto de los negocios.

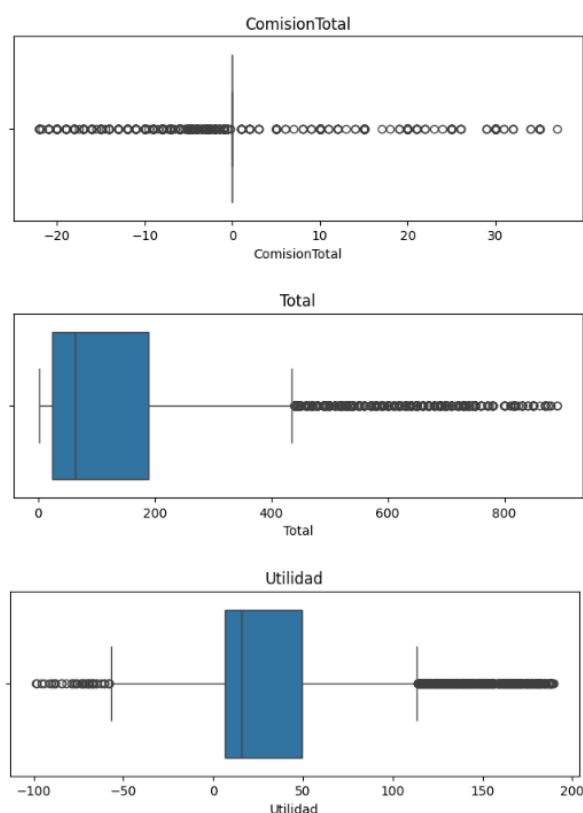


Figura 4-18 Datos relevantes para la empresa

Fuente: Elaboración Propia (2024)

En la figura 4-19 se ilustra la corrección de un error en el registro de datos, específicamente en la columna "PrecioVenta", donde se modificó el valor erróneo de 22,440 a 2,240, que representa el precio real de la herramienta en las cinco filas correspondientes. Es posible que existan otros registros con errores similares; sin embargo, estos no tendrían una relevancia o influencia significativa como el caso mencionado, el cual generaba un desbalance de 20,000 Bolivianos en las utilidades de la empresa.

	I	J	K	L	M	N	O	P	Q	
	Cantidad	PrecioCompra	PrecioVenta	PrecioFinal	ComisionUnit	ComisionTotal	Total	Utilidad	Marca	Proveedor
	1	1721	2240	2240	200	200	2240	319	ELECTROPLAST	FORZA S.R.L.
	1	1721	2240	2240	200	200	2240	319	ELECTROPLAST	FORZA S.R.L.
	1	1721	2240	2240	200	200	2240	319	ELECTROPLAST	FORZA S.R.L.
	1	1721	2240	2240	200	200	2240	319	ELECTROPLAST	FORZA S.R.L.
	4	1721	2240	2240	200	800	8960	1276	ELECTROPLAST	FORZA S.R.L.

Figura 4-19 Corrección error de registro en las ventas

Fuente: Elaboración Propia (2024)

### 4.3. Resultados entrenamiento Modelos Machine Learning

Se realizó el entrenamiento de modelos Regresión logística y RandomForest con el fin de predecir si una venta tiene un nivel Alto, Medio o Bajo, lo cual ayudaría con las toma de decisiones de ofertar ciertos productos o marcas. Las variables que se tomaron en cuenta para el entrenamiento fueron Marca, Sucursal y Precio venta.

#### 4.3.1. Resultados Modelo Random Forest

Como se observa en la figura 4-20 en la matriz de confusión presenta pocos errores y tanto en su accuracy como en F1-score tiene un valor de 0,88. Por lo que el modelo tiene un buen desempeño en la clasificación de los datos.

```

===== Calcular precisión =====
accuracy: 0.88

F1-Score: 0.88

Matriz de Confusión:
      alto  bajo  medio
alto    773    6    86
bajo    10   509    95
medio   45    51   896

```

Figura 4-20 Métricas modelo Random Forest

Fuente: Elaboración Propia (2024)

#### 4.3.2. Resultados Modelo Regresión Logística

Como se observa en la figura 4-21 los resultados de validación con las métricas de accuracy y F-score son 0.85 en ambos casos, también en la matriz de confusión se presentan pocos errores.



```
===== Calcular precisión =====  
Precisión: 0.85  
  
F1-Score: 0.85  
  
Matriz de Confusión:  
      alto  bajo  medio  
alto  715    0    150  
bajo   5    517    92  
medio  26    110   856
```

Figura 4-21 Métricas modelo Regresión Logística

Fuente: Elaboración Propia (2024)

4.3.3. Resultado Selección del Modelo

En la tabla 4-1 se pueden observar la comparación de las métricas de evaluación correspondientes a cada modelo. Por lo que se seleccionó el modelo de Random Forest que demostró un rendimiento superior al de regresión logística, alcanzando un valor de 0.88. Para mejorar el desempeño podríamos por ejemplo, incluir datos como la edad de los clientes el cual proporcionaría información valiosa sobre los grupos que generan mayores ingresos. Esta información permitiría desarrollar estrategias comerciales más efectivas, adaptadas a las características específicas de la clientela más rentable según su rango de edad.

Tabla 4-1 Comparación resultados métricas de evaluación

	Métrica Accuracy	Métrica F1-Score
Regresión logística	0.85	0.85
Random Forest	0.88	0.88

Fuente: Elaboración Propia (2024)

En la Figura 4-16, usando el modelo de Random Forest se aprecia la identificación de las características más influyentes en el éxito de las ventas. En particular, el 'precio de venta' destaca con una importancia relativa de 0.86, lo que sugiere una marcada preferencia de los clientes por opciones económicas o promocionales. Adicionalmente, aunque con una importancia menor, la marca 'Truper' exhibe una relevancia de 0.015 en las preferencias de los clientes, seguida por la marca 'Lynus' con una importancia

de 0.011. Por último, se observa que la sucursal ubicada en la ciudad de 'Santa Cruz' presenta una influencia superior en comparación con las demás sucursales.

Importancia de características:		
	Característica	Importancia
0	PrecioVenta	0.860666
88	TRUPER	0.015078
51	LYNUS	0.011672
97	SANTA CRUZ	0.010965
84	TOLSEN	0.010307
..	...	...
86	TOYAKI	0.000000
16	CORTAG	0.000000
76	ROS.ALEXIS	0.000000
39	HOTECHE	0.000000
72	PROVEEDOR NACIONAL	0.000000

[98 rows x 2 columns]

Figura 4-21 Interpretación de los resultados

Fuente: Elaboración Propia (2024)

En la figura 4-22 se muestra la gráfica de la relación entre el número de árboles en el modelo Random Forest y la precisión del modelo en los conjuntos de entrenamiento y prueba. Donde se entiende de La línea azul la precisión del modelo en el conjunto de entrenamiento mantiene muy alta y se estabiliza cerca del 90% a medida que aumenta el número de árboles, esto sugiere que el modelo está aprendiendo bien los patrones en los datos de entrenamiento. Por otro lado la línea naranja denota que el modelo se mantiene alrededor del 87-88% de precisión lo cual muestra un buen desempeño que podría mejorarse ajustando los parámetros del modelo.

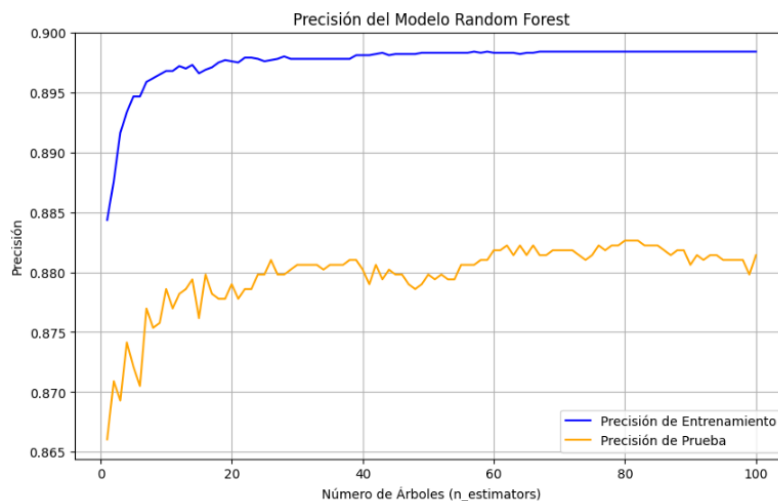


Figura 4-22 Gráfico comparación de precisión modelo Random Forest

Fuente: Elaboración Propia (2024)

#### 4.4. Resultados del desarrollo de tableros de control

La creación de los tableros de control (dashboards) permitió la visualización de los datos presentados en la Figura 4-23, donde se destacan aspectos clave como el registro de 115 vendedores durante los periodos 2022 y 2023, así como la participación de 8,284 clientes que realizaron al menos una compra. Asimismo, se evidencia que la empresa MASTER TOOLS alcanzó un ingreso total de 11,075,321 Bs en ambas gestiones, con un margen de utilidad del 19.1%.



Figura 4-23 Datos en general

Fuente: Elaboración Propia (2024)

Se generó también la visualización de la figuras 4-24 donde se muestra los 15 mejores marcas y además cuenta con botones interactivos para modificar el grafico y visualice los 15 marcas menos vendidas, con el fin de incrementar las estrategias de ventas en las mejores marcas como son: Lynus, ingco y Truper.



Figura 4-24 Marcas más y menos vendidas

Fuente: Elaboración propia Power Bi (2024)

La Figura 4-25 presenta un gráfico de barras que describe el comportamiento estacional de las ventas, destacando un incremento significativo durante los meses de abril, mayo y junio, periodo que corresponde al segundo trimestre del ciclo anual. Este gráfico incluye funciones de filtrado interactivo, permitiendo segmentar los datos según variables como el año o la sucursal, lo que facilita un análisis comparativo y contextualizado de las tendencias.

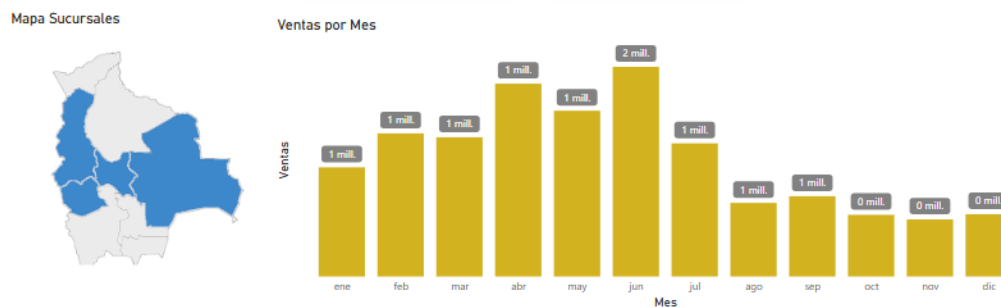


Figura 4-25 Comportamiento de ventas por mes

Fuente: Elaboración propia Power Bi (2024)

La Figura 4-26 presenta una tabla que resume los datos más relevantes correspondientes a los productos con el mejor rendimiento en ventas. Adicionalmente, la tabla incorpora una funcionalidad de ventana emergente, la cual permite visualizar el comportamiento individual de cada herramienta seleccionada a lo largo de los meses. Esta característica facilita un análisis detallado y comparativo del desempeño de los productos y fluctuaciones en la demanda.

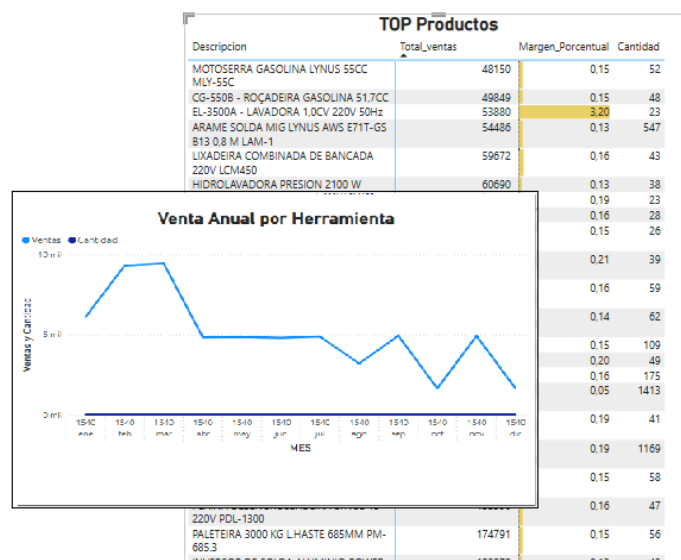


Figura 4-26 Productos con mejor rendimiento en ventas

Fuente: Elaboración Propia (2024)

La Figura 4-27 ilustra la distribución porcentual de los ingresos generados por cada sucursal, evidenciando a la de Santa Cruz presenta la mayor contribución a las ventas totales, con un 39,84%, seguida por la sucursal de Cochabamba, que aporta un 32,84%. Esto sugiere la necesidad de reforzar las estrategias de marketing en las sucursales de Oruro y La Paz, con el objetivo de optimizar su rendimiento comercial y equilibrar la distribución de ingresos.

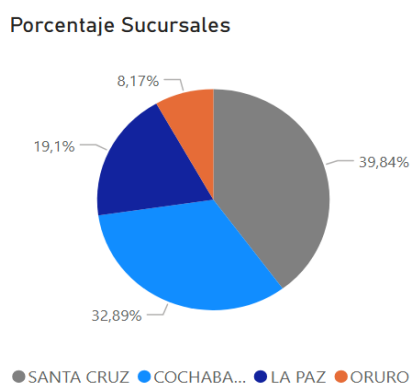


Figura 4-27 rendimiento porcentual de las sucursales

Fuente: Elaboración Propia (2024)

#### 4.5. Resultados determinación de patrones y tendencias

En la figura 4-28 se observó los productos con mayor y menor rentabilidad, como ser los más 3 más rentables son: Lpa-1000 máquina de pintura sin aire lynus 1000w 220v y lpa-650 máquina de pintura airless lynus 650w 220v y emphiladeria manual 2000 kg el-2000 con ventas de 342070, 254222 y 238946 respectivamente. Mientras los productos menos vendidos son motoserpa gasolina lynus 55cc mly-55c, cg-550b , roçadeira gasolina 51,7cc y el-3500a - lavadora 1,0cv 220v 50hz productos que no superaron los 60000 bs de ingresos.

Productos con mejores ventas		Productos con pocas ventas	
Descripcion	Total_ventas	Descripcion	Total_ventas
LPA-1000 MÁQUINA DE PINTURA SIN AIRE LYNUS 1000W 220V	342070	MOTOSERRA GASOLINA LYNUS 55CC MLY-55C	48150
LPA-650 MAQUINA DE PINTURA AIRLESS LYNUS 650W 220V	254222	CG-550B - ROÇADEIRA GASOLINA 51,7CC	49849
EMPHILADERIA MANUAL 2000 KG EL-2000	238946	EL-3500A - LAVADORA 1,0CV 220V 50Hz	53880
TC-TS 2025/2 U, SIERRA DE BANCO	228410	ARAME SOLDA MIG LYNUS AWS E71T-G5 B13 0,8 M LAM-1	54486
EMPILHADEIRA MANUAL 1T X 1.6M - 550 X 900MM EM-1000	224687	LIXADEIRA COMBINADA DE BANCADA 220V LCM450	59672
POWER INVERSORA DE SOLDA LYNUS MIG BIVOLT 140A LIM-140	216172	HIDROLAVADORA PRESION 2100 W	60690
POWER INVERSOR DE SOLDA LYNUS MIG 200A 220V LIS-220i	188519	HIDROLAVADORA A PRESION 2500W	65170
PLAINA DESENGROSSADEIRA LYNUS 8" 220V PDL-800	185250	PM - 520 PALETEIRA 2000KG - L. ASTE 520	74460
INVERSOR DE SOLDA ALUMINIO POWER 220V LIS-250AL	183870	LIT-516P INVERSOR DE SOLDA LYNUS MMA+TIG+CORTE 220V	79260
		GP-1500ABI - PICADOR C/MOTOR MON 1,5CV 220V/50HZ	79430
		LPLA - 750 LIJADORA DE PARED CON LUZ LED Y ASPIRACION	80440

Figura 4-28 Productos con más y menos Ventas

Fuente: Elaboración Propia (2024)

El análisis estratégico reveló que la sucursal en la ciudad de Santa Cruz tenía un rendimiento notablemente superior al promedio en cierta temporada más claro en el segundo trimestre del año, lo que sugiere que hubo buenas prácticas de marketing e incluso mayor personal en el área de ventas para esa temporada como se Observa en la figura 4-29, se identificaron también que a partir del cuarto trimestre más o menos al inicio de agosto las ventas decayeron no solo en santa cruz sino en la todas las sucursales debido al poco personal que hubo para esa temporadas.

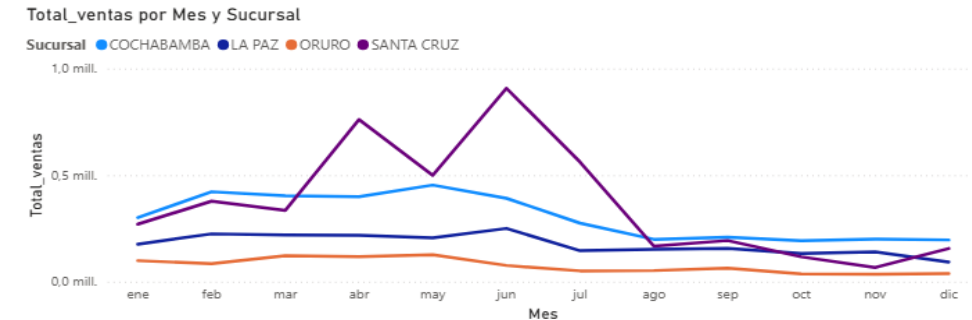


Figura 4-29 Marca con mejor rendimiento

Fuente: Elaboración Propia (2024)

Una evaluación de la Figura 4-30 revela una clara tendencia así la marca 'Lynus' el cual sobresale ampliamente en comparación con las demás marcas. Teniendo una diferencia sustancial de aproximadamente tres millones de Bs. con respecto a la segunda marca más vendida, 'Ingco'. Este hallazgo concuerda con el análisis previo que identifica la máquina de pintura sin aire de 1000w de la marca Lynus como el producto de mayor rentabilidad.



Figura 4-30 Comportamiento de ventas en las Sucursales

Fuente: Elaboración Propia (2024)

Se observa en la figura 4-31 el comportamiento del rendimiento de los mejores vendedores los cuales generan más ingresos para la empresa, teniendo a los 4 mejores vendedores de cada sucursal. En los que se destaca a Olymar Yelitza Andrade y Pablo Paz Peña los vendedores que registran ventas en 2 sucursales, representando un gran trabajo comercial.

Sucursal	Suma de Total
<b>SANTA CRUZ</b>	
mamier	1674960
Tienda Santa Cruz	1224318
Angie Johanna Carmona Salcedo	693429
Pablo Paz peña	99514
<b>COCHABAMBA</b>	
Tienda Cochabamba	2511723
Karen Angelica Navia Gutierrez	179640
Nury Fatima Mier Lopez	156600
NIA MONTECINOS	132684
<b>LA PAZ</b>	
Tienda La Paz	977049
OLYMAR YELITZA ANDRADE CABEZA	260409
Pablo Paz peña	141201
Karen Angelica Navia Gutierrez	76693
<b>ORURO</b>	
Tienda Oruro	550954
JHOSELYN VALLEJOS CANQUI	104760
OLYMAR YELITZA ANDRADE CABEZA	28644
TANIA MERCEDES MONTIEL AURAN	22890

Figura 4-31 Mejores Vendedores según la sucursal

Fuente: Elaboración Propia (2024)

Según la figura 4-32 donde se tiene las gráficas según el mes, los trimestres y el anual, se observó un incremento considerable de las ventas en el periodo del segundo trimestre del año en los meses comprendidos de abril mayo y junio.

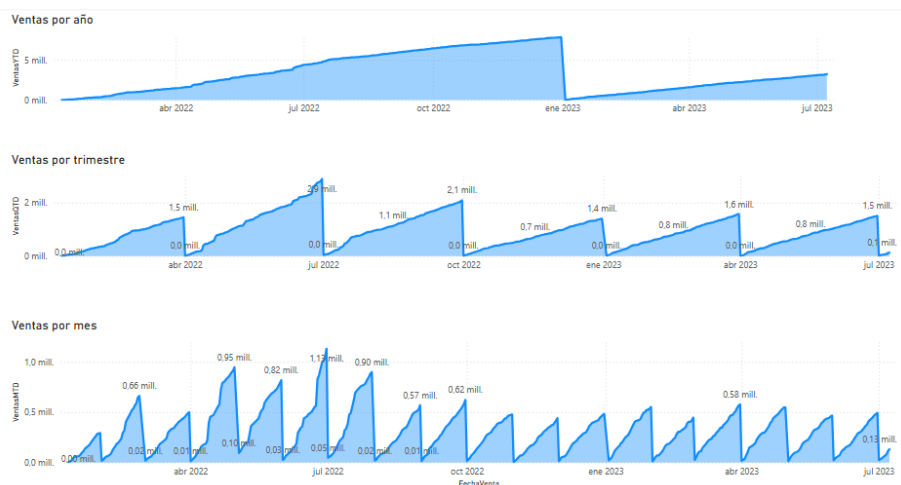


Figura 4-32 Fluctuación de las ventas por temporada

Fuente: Elaboración Propia (2024)

Se observa en la Figura 4-33 el comportamiento de los clientes según la sucursal, teniendo a los 4 mejores clientes de cada sucursal, por ejemplo en la Ciudad de La Paz se encuentra el cliente Freddy Huanca y en Cochabamba se tiene a la empresa de Flota Bolívar, esto apoyaría con el seguimientos a los clientes con mayor ingreso tratando de generar fidelización de dicho clientes, realizando visitas o llamadas ofreciendo las novedades de herramientas o algún paquete promocional.

Sucursal	Suma de Total
<b>COCHABAMBA</b>	
ALEJANDRO CLAURE	54900
CLIENTE	60033
FLOTA BOLIVAR	26697
MARTIN MIER	32074
<b>LA PAZ</b>	
FREDDY HUANCA	13230
JUAN TITIRICO	15600
JULIO CESAR	12258
ROBERTO CHAMBI	12196
<b>ORURO</b>	
GABINO CORO MAMANI	26760
JUAN ENRIQUE RAFAEL MARTINEZ	7500
VELASCO HUANCA FANOR	7930
XIMENA ZEBALLOS	7659
<b>SANTA CRUZ</b>	
MARTIN MIER	167839
Oferton Cochabamba	275776
Oferton La Paz	229272
Oferton Santa Cruz	303424

Figura 4-33 Clientes más rentables para la empresa

Fuente: Elaboración Propia (2024)



#### 4.6. Discusión de los resultados

En el estudio titulado "COMPARACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DESCUBRIR INFORMACIÓN RELEVANTE DE VENTAS DE UNA MYPE COMERCIAL" (Miñano Sánchez, 2022), se llevaron a cabo diversos análisis utilizando diferentes modelos de machine learning. Uno de los enfoques predictivos empleados fue el algoritmo de Regresión Logística, mediante el cual se realizaron mediciones basadas en indicadores previamente establecidos. El conjunto de datos utilizado constó de 1,113 registros, de los cuales el 80% se destinó al entrenamiento y el 20% restante se reservó para pruebas (testing). Cabe destacar que los datos fueron sometidos a un proceso de normalización previo para garantizar su adecuada preparación.

En la figura 4-34 y 4-35 se presenta la evaluación de presión de los modelos de Random Forest y Regresión Logística utilizados con las métricas correspondientes del proyecto presente.

```
# Entrenar el modelo de Random Forest
rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train, y_train)

accuracy = accuracy_score(y_test, y_pred)
print(f'accuracy: {accuracy:.2f}\n')

f1 = f1_score(y_test, y_pred, average='weighted')
print(f'F1-Score: {f1:.2f}\n')
```

Figura 4-34 Script calcular la Precisión Modelo Random Forest

Fuente: Elaboración Propia (2024)

```
# Entrenar modelo de regresión logística
model = LogisticRegression()
model.fit(X_train, y_train)

accuracy = accuracy_score(y_test, y_pred)
print(f'Precisión: {accuracy:.2f}\n')
```

Figura 4-35 Script calcular la Precisión Modelo Regresión Logística

Fuente: Elaboración Propia (2024)

En la Figura 4-36 se presenta la métrica de Accuracy, utilizada para calcular el nivel de precisión del modelo del proyecto del autor Miñano Sanchez, cuyo resultado fue del 0.993%. Esta métrica, ampliamente reconocida en los campos del aprendizaje automático y la estadística, se emplea para evaluar el desempeño de los modelos de Machine Learning.

```
# muestra la precision en el entrenamiento
accuracyTrain = model_7.score(x_train,y_train)
accuracyTest = model_7.score(x_test,y_test)
print('accuracyTrain: ', accuracyTrain , " accuracyTest: ", accuracyTest)
```

Figura 4-36 Script calcular la Precisión Modelo Regresión Lineal

Fuente: (Miñano Sanchez, 2022)

Como se mencionaba el análisis arrojó un nivel de precisión (P) del 0.993% al emplear un conjunto de datos de 1,113 registros. En contraste, para el presente proyecto se trabajó con un dataset significativamente mayor, compuesto por 12,355 registros. En este contexto, se implementaron dos modelos predictivos: Regresión Logística y Árbol de Decisiones. Entre ambos, el modelo de Árbol de Decisiones demostró un mejor desempeño, alcanzando una precisión del 88%, como se detalla en la Tabla 4.2.

Tabla 4-2 Tabla Comparativa Modelo de machine Learning

Proyecto	Matriz de confusión	Precisión	Modelo																
COMPARACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA DESCUBRIR INFORMACIÓN RELEVANTE DE VENTAS DE UNA MYPE COMERCIAL	<div>Matriz de Confusión de la Clasificación</div> <table><tr><td></td><td>BOLETA (NO)</td><td>FACTURA (SI)</td></tr><tr><td>BOLETA (NO)</td><td>386</td><td>0</td></tr><tr><td>FACTURA (SI)</td><td>1</td><td>726</td></tr></table> <div>Regresión Lineal</div>		BOLETA (NO)	FACTURA (SI)	BOLETA (NO)	386	0	FACTURA (SI)	1	726	0.993%	Regresión Logística							
	BOLETA (NO)	FACTURA (SI)																	
BOLETA (NO)	386	0																	
FACTURA (SI)	1	726																	
ANÁLISIS EXPLORATORIO DE DATOS Y TENDENCIAS DE UNA COMERCIALIZADORA DE HERRAMIENTAS USANDO BUSINESS INTELLIGENCE	<div>Matriz de Confusión:</div> <table><tr><td></td><td>alto</td><td>bajo</td><td>medio</td></tr><tr><td>alto</td><td>619</td><td>3</td><td>347</td></tr><tr><td>bajo</td><td>9</td><td>864</td><td>184</td></tr><tr><td>medio</td><td>127</td><td>380</td><td>1635</td></tr></table>		alto	bajo	medio	alto	619	3	347	bajo	9	864	184	medio	127	380	1635	0.85%	Regresión Logística
	alto	bajo	medio																
alto	619	3	347																
bajo	9	864	184																
medio	127	380	1635																
	<div>Matriz de Confusión:</div> <table><tr><td></td><td>alto</td><td>bajo</td><td>medio</td></tr><tr><td>alto</td><td>788</td><td>5</td><td>176</td></tr><tr><td>bajo</td><td>5</td><td>862</td><td>190</td></tr><tr><td>medio</td><td>149</td><td>126</td><td>1867</td></tr></table>		alto	bajo	medio	alto	788	5	176	bajo	5	862	190	medio	149	126	1867	0.88%	Árbol de Decisiones
	alto	bajo	medio																
alto	788	5	176																
bajo	5	862	190																
medio	149	126	1867																

Fuente: Elaboración Propia (2024)

## 5. Conclusiones

Los datos históricos recolectados comprenden las gestiones 2022 y 2023 donde el dataset cuenta con 20883 transacciones de ventas y 18 variables los cuales se puede observar en el Anexo 2. Cabe recalcar que el dataset presentaba algunos errores de cálculos en sus fórmulas los cuales informaron al gerente de la empresa MASTER TOOLS, también se realizó una encuesta a los vendedores activos con el fin de recolectar información relevante para una mejor comprensión del funcionamiento del negocio de cual de obtuvo que la empresa opera de manera tradicional, obteniendo información mediante los cálculos de libros en Excel y que carece de un seguimiento oportuno de sus clientes potenciales.

El análisis exploratorio de datos proporcionó una visión integral de las transacciones de la empresa MASTER TOOLS. Se identificaron valores nulos y errores tipográficos donde había registro de Proveedores similares pero escritos de forma distinta como el caso de 'Forza S.R.L.' y 'Forza SRL' los cuales se corrigieron en la limpieza de datos, se identificó una amplia variedad de marcas de las herramientas que ofrece la empresa teniendo 120 en total, una fuerte correlación entre el precio de venta y la utilidad con un 0,55 y una media de  $\mu=93$  Bs, por transacción registrada con la desviación estándar de  $\sigma=277$  en indicando la amplia dispersión.

Al aplicar modelos de Machine Learning, como la regresión logística y Random Forest, se obtuvieron valores de accuracy de 0.85 y 0.88, respectivamente, así como un F1-Score de 0.85 y 0.88 teniendo un buen rendimiento. Estos modelos no solo permiten predecir el rendimiento de las ventas, sino que también identifican las características más influyentes en su éxito. Entre los factores determinantes se destaca el precio, que influye significativamente en la concreción de una venta, y la marca, siendo 'Truper' y 'Lynus' las más demandadas. Asimismo, la ubicación geográfica del cliente juega un papel relevante, destacándose Santa Cruz como la región con mejor desempeño en ventas.

La creación de tableros de control (dashboards) en Power BI facilitó la visualización de indicadores clave de rendimiento (KPIs) y tendencias significativas en las transacciones comerciales. Los análisis revelaron una predominancia de la marca LYNUS en las preferencias de los clientes, con el producto "LPA-1000 Máquina de Pintura Sin Aire Lynus 1000W" como el más vendido. Geográficamente, las sucursales de Santa Cruz (39,84%) y Cochabamba (32,84%) registraron los mayores ingresos durante las gestiones 2022 y 2023, consolidándose como las de mayor rendimiento comercial. Adicionalmente, se identificó un pico estacional en ventas durante el segundo trimestre (abril, mayo y junio), periodo que concentró las transacciones más altas del ciclo anual. Estos hallazgos, respaldados por visualizaciones interactivas, permitieron correlacionar estrategias promocionales con incrementos en la demanda, particularmente en herramientas eléctricas.

## 6. Recomendaciones

A partir de los resultados obtenidos en el análisis de datos y la implementación de Business Intelligence en la empresa comercial "MASTER TOOLS", se presentan las siguientes recomendaciones para optimizar aún más la gestión y mejorar la experiencia del cliente:

Se recomienda establecer un sistema robusto para la recolección y análisis de datos históricos, que incluya no solo datos de ventas, sino también información sobre devoluciones, descuentos y datos de los clientes como su número de celular. Para poder mejorar el análisis y contactarse con el cliente para indicar novedades en los productos.

Es crucial implementar un programa de capacitación regular para los empleados, enfocándose en el uso efectivo de herramientas de Business Intelligence, así como en técnicas de ventas y atención al cliente. Esto garantizará que el equipo esté preparado para aprovechar al máximo las herramientas disponibles y mejorar la experiencia del cliente.

Utilizar los insights obtenidos del análisis de datos para optimizar la gestión del inventario. Esto incluye identificar productos de alta rotación y ajustar los niveles de stock para minimizar costos y asegurar la disponibilidad de productos populares.

Basándose en los hallazgos analíticos, es fundamental desarrollar estrategias personalizadas que mejoren la experiencia del cliente. Esto incluye promociones específicas basadas en el comportamiento del consumidor y un enfoque proactivo en la atención al cliente para resolver inquietudes antes de que se conviertan en problemas.

Finalmente, es crucial realizar análisis periódicos para trazar estrategias según la evolución del mercado y el desempeño de vendedores. Estableciendo reuniones regulares de revisión de datos y desempeño, utilizando indicadores clave de rendimiento que asegure a la empresa "MASTER TOOLS" se mantenga competitiva en un mercado dinámico.

Estas recomendaciones están diseñadas para fortalecer la capacidad analítica y estratégica de "MASTER TOOLS", garantizando una respuesta ágil ante las tendencias del mercado y mejorando la satisfacción del cliente a largo plazo.

## Referencias bibliográficas

- Aguilar, L. Arquitectura de una herramienta empresarial de toma de decisiones para la gestión del departamento bienestar Universitario de la Universidad tecnica del norte. (*Tesis Maestria*). Universidad Tecnica del Norte, Ibarra.
- Bernabeu, R. (19 de Julio de 2010). *www.businessintelligence.info*. Recuperado el 15 de Noviembre de 2024, de *www.businessintelligence.info*: <https://www.businessintelligence.info/resources/assets/hefesto-v2.pdf>
- Cano, J. (2007). Business intelligence: Competir con información. En J. Cano, *Business intelligence: Competir con información* (pág. 545). Madrid: Banesto: Fundación Cultural. Obtenido de <https://www.scribd.com/document/162811141/Business-Intelligence> Competir-Con- Informacion
- Cebotarean, E. (15 de febrero de 2011). *www.scientificpapers.org*. Recuperado el 16 de Noviembre de 2024, de *www.scientificpapers.org*: <https://www.scientificpapers.org/economics/business-intelligence/>
- Chavez Huapaya, S. M. (Enero de 2018). <https://repositorio.autonoma.edu.pe/>. Recuperado el 18 de Diciembre de 2024, de <https://hdl.handle.net/20.500.13067/435>
- Chrystal R, C. (20 de Diciembre de 2023). Recuperado el 20 de Enero de 2025, de *ibm.com*: <https://www.ibm.com/es-es/think/topics/machine-learning-types>
- Clientify. (25 de Junio de 2024). *Clientify.com*. Recuperado el 28 de Enero de 2025, de <https://clientify.com/blog/marketing/recoleccion-de-datos-metodos-tecnicas-e-instrumentos>
- Colab.Research. (16 de Septiembre de 2022). Recuperado el 07 de Diciembre de 2024, de *colab.research*: <https://colab.research.google.com/>
- Cruaños, H. (22 de 11 de 2020). Recuperado el 05 de Diciembre de 2024, de *www.hiberus.com*: <https://www.hiberus.com/crecemos-contigo/business-intelligence/>
- Databitai. (17 de Abril de 2023). Recuperado el 28 de Enero de 2025, de <https://databitai.com/>: <https://databitai.com/machine-learning/metricas-de-evaluacion-en-machine-learning/>
- Dataprix. (06 de Mayo de 2009). <https://www.dataprix.com/>. Recuperado el 23 de Octubre de 2024, de <https://www.dataprix.com/>: <https://www.dataprix.com/es/data-warehousing-y-metodologia-hefesto/34-datawarehouse-manager>
- Falcón Morales, L. E. (21 de Marzo de 2023). *blog.maestriasydiplomados.tec*. Obtenido de <https://blog.maestriasydiplomados.tec.mx/recoleccion-de-datos-que-es-ventajas-y-consejos-para-usarlos>

- GALIANA, P. (9 de septiembre de 2022). *iebschool.com*. Recuperado el 12 de Octubre de 2024, de iebschool.com: <https://www.iebschool.com/blog/herramientas-business-intelligence-digital-business/>
- GARTNER. (2021). *gartner.com*. Recuperado el 12 de octubre de 2024, de gartner.com: <https://www.gartner.com/en/information-technology/glossary/business-intelligence-bi>
- Gartner. (2022). *www.gartner.es*. Recuperado el 20 de Noviembre de 2024, de *www.gartner.es*: <https://www.gartner.es/es/metodologias/magic-quadrant>
- González, R. &. (2021). Uso de minería de datos para analizar las tendencias de consumo noticioso en un diario en línea local. *Revista Internacional de Comunicación*.
- Gravitar. (23 de Noviembre de 2022). *gravitar.biz*. Recuperado el 16 de Noviembre de 2024, de *gravitar.biz*: <https://gravitar.biz/datawarehouse/metodologias-data-warehouse/>
- IBM. (marzo de 2010). Recuperado el 03 de Diciembre de 2024, de *www.ibm.com*: <https://www.ibm.com/mx-es/topics/business-intelligence>
- IBM. (2016). *www.ibm.com*. Recuperado el 18 de Noviembre de 2024, de *www.ibm.com*: <https://www.ibm.com/es-es/topics/data-modeling>
- IBM. (28 de Junio de 2024). Obtenido de <https://www.ibm.com/mx-es/topics/exploratory-data-analysis>
- Inmon, B. (2005). *es.scribd.com*. Recuperado el 20 de Noviembre de 2024, de *es.scribd.com*: <https://es.scribd.com/document/265678703/Paradigma-Bill-Inmon>
- Kaizen. (20 de agosto de 2008). Recuperado el 03 de Diciembre de 2024, de *www.kaizen.com*: <https://kaizen.com/es/insights-es/business-intelligence-ventaja-estrategica/>
- Kimball. (2015). *kimballgroup*. Recuperado el 20 de Noviembre de 2023, de *kimballgroup*: <https://www.kimballgroup.com/>
- Miñano Sanchez, C. J. (2022). *repositorio.uss.edu.pe*. Recuperado el 13 de 12 de 2024, de <https://repositorio.uss.edu.pe/handle/20.500.12802/10077>
- Noriega, R. V. (2015). Evolución de la inteligencia de negocios. *Núm. 57 (12)*, 299-308.
- Prometeusgs. (19 de Julio de 2024). Recuperado el 06 de Diciembre de 2024, de *prometeusgs.com*: <https://prometeusgs.com/como-el-analisis-de-datos-esta-transformando-el-comercio-minorista/>
- Rivera, R. F. Aplicación de Business intelligence en una pequeña empresa usando power BI. (*Tesis Grado en Ingenieria*). Universidad de valladolid Escuela de Ingenierias, Valladolid.
- Torrez, D. (6 de Noviembre de 2024). Recuperado el 24 de Enero de 2025, de *hubspot*: <https://blog.hubspot.es/sales/metricas-ventas>
- Yaipén, E. F. Business Intelligence para mejorarla toma de decisiones en la gerencia general. (*Tesis Ingenieria Sistemas e Informatica*). Universidad Nacional del Santa, Nuevo Chimbote.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Fecha	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10					
2024-10-05	Algo organizado	Herramientas: Manualmente	Definido por el Mensual	me	Cantidad d	Ocasional	me	Técnicas de v	Falta de clientes; Fa	Mejorar la capacitación del equipo; Implementar herramientas tecn					
2024-10-05	Poco estructura	Herramientas: Manualmente	Definido por el No se revisa	Cantidad d	No se brinda	Atención al cl	Falta de informació	Implementar herramientas tecnológicas; Optimizar el proceso de s							
2024-10-05	Algo organizado	Hojas de cálculo Manualmente	Definido por el No se revisa	Total ganar	Ocasional	me	Manejo de he	Falta de clientes; Fa	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Poco estructura	Herramientas: Manualmente	No hay metas	C Mensual	me	Total ganar	Ocasional	me	Atención al cl	Falta de capacitación	Implementar herramientas tecnológicas; Optimizar el proceso de s				
2024-10-05	Poco estructura	Hojas de cálculo No se realiza	s No hay metas	« No se revisa	Total ganar	No se brinda	Atención al cl	Falta de clientes; Fa	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Algo organizado	Herramientas: No se realiza	s Definido por el Mensual	me	Total ganar	No se brinda	Atención al cl	Falta de clientes; Li	Mejorar la capacitación del equipo; Implementar herramientas tecn						
2024-10-05	Algo organizado	Hojas de cálculo No se realiza	s Definido por el Mensual	me	Total ganar	Ocasional	me	Atención al cl	Falta de informació	Implementar herramientas tecnológicas; Optimizar el proceso de s					
2024-10-05	Algo organizado	Herramientas: No se realiza	s Definido por el Mensual	me	Total ganar	No se brinda	Atención al cl	Falta de informació	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Algo organizado	Herramientas: No se realiza	s No hay metas	« No se revisa	Cantidad d	No se brinda	Técnicas de v	Limitaciones en her	Optimizar el proceso de seguimiento de clientes						
2024-10-05	Algo organizado	Hojas de cálculo Manualmente	Definido por el Mensual	me	Total ganar	Ocasional	me	Manejo de he	Falta de capacitación	Mejorar la capacitación del equipo; Optimizar el proceso de seguim					
2024-10-05	Algo organizado	Hojas de cálculo Manualmente	Definido por el Mensual	me	Cantidad d	Ocasional	me	Manejo de he	Falta de informació	Mejorar la capacitación del equipo; Optimizar el proceso de seguim					
2024-10-05	Algo organizado	Herramientas: Manualmente	No hay metas	C Mensual	me	Total ganar	No se brinda	Atención al cl	Falta de informació	Mejorar la capacitación del equipo; Implementar herramientas tecn					
2024-10-05	Algo organizado	Hojas de cálculo No se realiza	s No hay metas	« Mensual	me	Total ganar	No se brinda	Atención al cl	Falta de capacitación	Mejorar la capacitación del equipo; Implementar herramientas tecn					
2024-10-05	Algo organizado	Hojas de cálculo Manualmente	Definido por el No se revisa	Total ganar	Ocasional	me	Manejo de he	Falta de informació	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Algo organizado	Herramientas: No se realiza	s Definido por el Mensual	me	Cantidad d	Ocasional	me	Atención al cl	Falta de clientes; Fa	Mejorar la capacitación del equipo; Optimizar el proceso de seguim					
2024-10-05	Algo organizado	Hojas de cálculo Manualmente	Definido por el Mensual	me	Total ganar	No se brinda	Atención al cl	Falta de informació	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Algo organizado	Herramientas: Manualmente	Definido por el Mensual	me	Total ganar	Ocasional	me	Técnicas de v	Falta de informació	Mejorar la capacitación del equipo; Implementar herramientas tecn					
2024-10-05	Poco estructura	Herramientas: No se realiza	s No hay metas	« No se revisa	Total ganar	No se brinda	Atención al cl	Falta de capacitación	Implementar herramientas tecnológicas; Optimizar el proceso de s						
2024-10-05	Poco estructura	Herramientas: No se realiza	s No hay metas	« No se revisa	Total ganar	No se brinda	Técnicas de v	Falta de capacitación	Mejorar la capacitación del equipo; Optimizar el proceso de seguim						
2024-10-05	Algo organizado	Herramientas: No se realiza	s No hay metas	« No se revisa	Total ganar	No se brinda	Atención al cl	Falta de clientes; Li	Mejorar la capacitación del equipo; Optimizar el proceso de seguim						
2024-10-06	Poco estructura	Hojas de cálculo No se realiza	s No hay metas	« No se revisa	Total ganar	Ocasional	me	Manejo de he	Falta de capacitación	Implementar herramientas tecnológicas					
2024-10-06	Algo organizado	Hojas de cálculo Manualmente	Definido por el Mensual	me	Cantidad d	Ocasional	me	Manejo de he	Falta de clientes; Fa	Mejorar la capacitación del equipo; Implementar herramientas tecn					

## Anexo 2. Encuesta comprensión del negocio

**Objetivo:** Comprender el funcionamiento del área de ventas, identificar fortalezas, debilidades y oportunidades de mejora.

1. **¿Cómo describiría el proceso de ventas actual en la empresa?**
  - Muy estructurado y eficiente
  - Algo organizado, pero con áreas de mejora
  - Poco estructurado y desorganizado
2. **¿Qué herramientas o sistemas utiliza el equipo de ventas para gestionar clientes y ventas?**
  - CRM (Customer Relationship Management)
  - Hojas de cálculo (Excel, Google Sheets)
  - Herramientas manuales (papel, libretas)
  - Otro: \_\_\_\_\_

3. **¿Cómo se realiza el seguimiento a los clientes potenciales y existentes?**
- ☐ Proceso automatizado
  - ☐ Manualmente, con llamadas o correos
  - ☐ No se realiza seguimiento
  - ☐ Otro: \_\_\_\_\_
4. **¿Cómo se establecen las metas de ventas en la empresa?**
- ☐ Basadas en datos históricos
  - ☐ Definido por el gerente sin análisis previo
  - ☐ No hay metas claras
5. **¿Con qué frecuencia se revisa el desempeño del equipo de ventas?**
- ☐ Semanalmente
  - ☐ Mensualmente
  - ☐ No se revisa
6. **¿Qué métricas se utilizan para medir el éxito de las ventas?** (Seleccione todas las que apliquen)
- ☐ Cantidad de productos vendidos
  - ☐ Total ganancias
  - ☐ Número de clientes nuevos
  - ☐ Otro: \_\_\_\_\_
7. **¿Recibe el equipo de ventas capacitación regular?**
- ☐ Sí, de manera frecuente
  - ☐ Ocasionalmente
  - ☐ No se brinda capacitación
8. **¿Qué tipo de capacitación considera más útil para el equipo de ventas?**
- ☐ Técnicas de ventas
  - ☐ Manejo de herramientas tecnológicas
  - ☐ Atención al cliente
  - ☐ Otro: \_\_\_\_\_
9. **¿Cuáles son los principales retos que enfrenta el área de ventas?** (Seleccione todas las que apliquen)



- Falta de clientes
- Falta de información del estado de ventas
- Falta de capacitación
- Limitaciones en herramientas tecnológicas
- Otro: \_\_\_\_\_

**10. ¿Qué sugerencias tiene para mejorar en el área de ventas?** (Seleccione todas las que apliquen)

- Mejorar la capacitación del equipo
- Implementar herramientas tecnológicas
- Optimizar el proceso de seguimiento de clientes
- Otro: \_\_\_\_\_

Fuente:

[https://docs.google.com/forms/d/e/1FAIpQLSdj\\_lPcj4mPD1vzxFKToEwgWrQ1ThOk4l\\_cDCmJL\\_OYPnEatKQ/viewform](https://docs.google.com/forms/d/e/1FAIpQLSdj_lPcj4mPD1vzxFKToEwgWrQ1ThOk4l_cDCmJL_OYPnEatKQ/viewform)

**Anexo 2. Planilla de datos extraídos sobre las ventas del negocio**

Portapapeles	Fuente	Alineación	Número	Estilos	Celdas	Modificar	
B39	f	LENTES P/SOLDAR ELECTRONICA C/OSCRECIMIENTO					
B		C	D	E	F	G	H
1	Descripcion	ID	Depósito	Vendedor	Cliente	Fecha	Vent
2	CABLE DE SOLDAR MIG LYNUS AWS E71T-GS 1/8" 0.8M	381	Santa Cruz	NATALIA ROMERO	SANTIAGO	2022-01-10	2022-01-10
3	COMPRESOR ENHELL DE 24LITROS	685	Santa Cruz	Francisco mariscal torres	ROLANDO CONDORI	2022-01-10	2022-01-10
4	ESCALERA MULTIFUNCION 4X3	130	Santa Cruz	Angie Johanna Camona Salced	ALEXANDER	2022-01-10	2022-01-10
5	ASPIRADORA 1400W BTA 133001	1333	Santa Cruz	Karen Angelica Navia Gutierrez	DANY SANCHEZ	2022-01-10	2022-01-10
6	SIERRA DE MESA 1800W	396	Santa Cruz	Angie Johanna Camona Salced	ADALDO TORREZ	2022-01-10	2022-01-10
7	ARRANCADOR AUXILIAR 350AMP	1328	Santa Cruz	Angie Johanna Camona Salced	RENE CORTEZ	2022-01-11	2022-01-11
8	HIDROLAVADORA 1200W MINI	114	Santa Cruz	PABLO SOLORZANO	VIVIAN GUARDIA	2022-01-11	2022-01-11
9	CC-PO 1100/2E PULIDORA AUTOMOVIL	687	Santa Cruz	Angie Johanna Camona Salced	SERGIO LOROÑO	2022-01-11	2022-01-11
10	SIERRA DE MESA 1800W	396	Santa Cruz	Angie Johanna Camona Salced	CRISTOBAL VARGAS	2022-01-11	2022-01-11
11	DISCO SIERRA CIRCULAR TCT ALUMINIO 254/10" 100T	562	Santa Cruz	Angie Johanna Camona Salced	CRISTOBAL VARGAS	2022-01-11	2022-01-11
12	LJADORA COMBINADA COMBINADA 6"	442	Santa Cruz	Angie Johanna Camona Salced	ALBERTO CESPEDES	2022-01-11	2022-01-11
13	CORTADORA ELECTRICA 1300W CE35M2	132	Santa Cruz	YUGUMAR MENDOZA	GUSTAVO GUZMAN	2022-01-11	2022-01-11
14	PISTOLA DE PINTAR TC SY 700	1337	Santa Cruz	Angie Johanna Camona Salced	RAUL FLORES	2022-01-11	2022-01-11
15	PRENSA DE ALUMINIO ESQUINERA 3 PROFESIONAL	1085	Santa Cruz	Angie Johanna Camona Salced	RAUL FLORES	2022-01-11	2022-01-11
16	JUEGO DE BROCAS MIXTAS 67 PZA P/MET. CON. MAD + PUNTAS	256	Santa Cruz	Angie Johanna Camona Salced	RAUL FLORES	2022-01-11	2022-01-11
17	ESCUADRA DE ALUMINIO P/MUROS DE YESO LARGO 30CM	1253	Santa Cruz	Angie Johanna Camona Salced	RAUL FLORES	2022-01-11	2022-01-11
18	JGO DE PUNTAS CRUZ 75MM	2043	Oruro	Tienda Oruro	antonio	2022-01-12	2022-01-12
19	TRAJE DE TRABAJO ALASKA TXL	1553	Oruro	Tienda Oruro	jose marca	2022-01-12	2022-01-12
20	CANDADO DE HIERRO 50MM GANCHO ESTANDAR	1890	Oruro	Tienda Oruro	jose marca	2022-01-12	2022-01-12
21	CANDADO DE HIERRO 38MM GANCHO ESTANDAR	1888	Oruro	Tienda Oruro	jose marca	2022-01-12	2022-01-12
22	SIERRAS PICALADORAS ZANCO "T" CORTE CURVO FINO	1921	Oruro	Tienda Oruro	calet villagas	2022-01-12	2022-01-12
23	LLAVE MANDRIL PARA TALADRO 13MM LARGO 80MM	1417	Oruro	Tienda Oruro	fernando	2022-01-12	2022-01-12
24	DISCO SIERRA CIRCULAR TCT ALUMINIO 254/10" 80T	538	Santa Cruz	Tienda Santa Cruz	YUNIOR	2022-01-12	2022-01-12
25	PIRÓMETRO MEDIDOR TEMPER-MLS BOSCH	1282	Santa Cruz	PABLO SOLORZANO	WILFREDO RIVERO	2022-01-12	2022-01-12
26	SIERRA CALADORA 710V	574	Santa Cruz	Francisco mariscal torres	RONALD VARGAS	2022-01-12	2022-01-12
27	MASCARA SOLDAR AUTO C/CONTRALADOR LYNUS	391	Santa Cruz	YUGUMAR MENDOZA	ALVARO MAMANI	2022-01-12	2022-01-12
28	MANGUERA SUCCION DE 2 PULGADAS 25M	2086	Santa Cruz	Nury Fatima Mer Lopez	ANDRES VILLARROEL	2022-01-12	2022-01-12
29	MOTOBOMBA A GASOLINA 2X2 CALDAL 40M3/H 6.5HP	2085	Santa Cruz	Nury Fatima Mer Lopez	ANDRES VILLARROEL	2022-01-12	2022-01-12
30	GATO HIDRAULICO DE PATIN DE 3 TON TRUPER	1217	Santa Cruz	YUGUMAR MENDOZA	LUIS GUTIERREZ	2022-01-12	2022-01-12
31	AMOLADORA 150MM 1550W	603	Santa Cruz	YUGUMAR MENDOZA	ELICO	2022-01-12	2022-01-12
32	CABLE DE SOLDAR MIG LYNUS AWS E71T-GS 1/8" 0.8M	1566	Oruro	OLYMAR YELITZA ANDRADE CA	LUIS ALBERTO GOMEZ	2022-01-12	2022-01-12
33	ESCALERA DE ALUMINIO CON BANDEJA 6 ESCALONES	1980	Oruro	Tienda Oruro	PASCUAL FLORES	2022-01-12	2022-01-12
34	JUEGO DE LLAVES ALLEN HEXAGONAL DE 13 PIEZAS MM	1840	Oruro	Tienda Oruro	CLIENTE	2022-01-12	2022-01-12

Ubicación: CD: 01\_BaseDeDatos/VentasRango.xls

### Anexo 3. Código Modelo Random Forest

```

from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, f1_score

# Convertir 'Marca' a variable categórica
df_ml['Marca'] = pd.Categorical(df_ml['Marca'])

# Preparar los datos
# Dividir datos en entrenamiento y prueba
X = df_ml[['Marca', 'Sucursal', 'PrecioVenta']]
y = df_ml['Nivel_ventas']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Convertir la columna 'Marca' a variables dummy (codificación one-hot)
# Crear variables dummy para la columna 'Marca' en X_train
dummies_train = pd.get_dummies(X_train['Marca'], drop_first=True)
X_train = pd.concat([X_train.drop('Marca', axis=1), dummies_train], axis=1)

# Crear variables dummy para la columna 'Sucursal' en X_train
dummies_train = pd.get_dummies(X_train['Sucursal'], drop_first=True)
X_train = pd.concat([X_train.drop('Sucursal', axis=1), dummies_train], axis=1)

# Crear variables dummy para la columna 'Marca' en X_test
dummies_test = pd.get_dummies(X_test['Marca'], drop_first=True)
X_test = pd.concat([X_test.drop('Marca', axis=1), dummies_test], axis=1)

# Crear variables dummy para la columna 'Sucursal' en X_test
dummies_test = pd.get_dummies(X_test['Sucursal'], drop_first=True)
X_test = pd.concat([X_test.drop('Sucursal', axis=1), dummies_test], axis=1)

# Asegurar que X_train y X_test tengan las mismas columnas
missing_cols = set(X_train.columns) - set(X_test.columns)
for col in missing_cols:
    X_test[col] = 0
X_test = X_test[X_train.columns]

# Entrenar el modelo de Random Forest
rf_model = RandomForestClassifier(random_state=42, n_estimators=100)
rf_model.fit(X_train, y_train)

# Evaluar el modelo
y_pred = rf_model.predict(X_test)

# Reporte de clasificación
print("Reporte de clasificación:")
print(classification_report(y_test, y_pred))

# Calcular precisión
print("==== Calcular precisión ====")
accuracy = accuracy_score(y_test, y_pred)
print(f'accuracy: {accuracy:.2f}\n')

```

Ubicación: CD: 02\_Notebook/Modelo\_Random\_Forest.ipynb

## Anexo 4. Código Modelo Regresión Logística

```
#Usando modelo LogisticRegression
from sklearn.linear_model import LogisticRegression

#Preparar datos

# Dividir datos en entrenamiento y prueba
X = df_ml[['Marca', 'PrecioVenta']]
y = df_ml['Nivel_ventas']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Convertir marca a variable numérica
# Crear variables dummy para la columna 'Marca' sin asignarla directamente a 'X_train['Marca']'
dummies = pd.get_dummies(X_train['Marca'], drop_first=True)

# Concatenar las variables dummy con el DataFrame original y eliminar la columna original 'Marca'
X_train = pd.concat([X_train.drop('Marca', axis=1), dummies], axis=1)

# Generar variables dummy para la columna 'Marca' en X_test
dummies_test = pd.get_dummies(X_test['Marca'], drop_first=True)

# Concatenar las variables dummy con el DataFrame original X_test y eliminar la columna 'Marca'
X_test = pd.concat([X_test.drop('Marca', axis=1), dummies_test], axis=1)

#Entrenar modelo

# Entrenar modelo de regresión logística
model = LogisticRegression()
model.fit(X_train, y_train)

#Evaluar modelo

# Predecir resultados
y_pred = model.predict(X_test)

# Evaluar precisión
accuracy = accuracy_score(y_test, y_pred)
print(f'Precisión: {accuracy:.2f}\n')

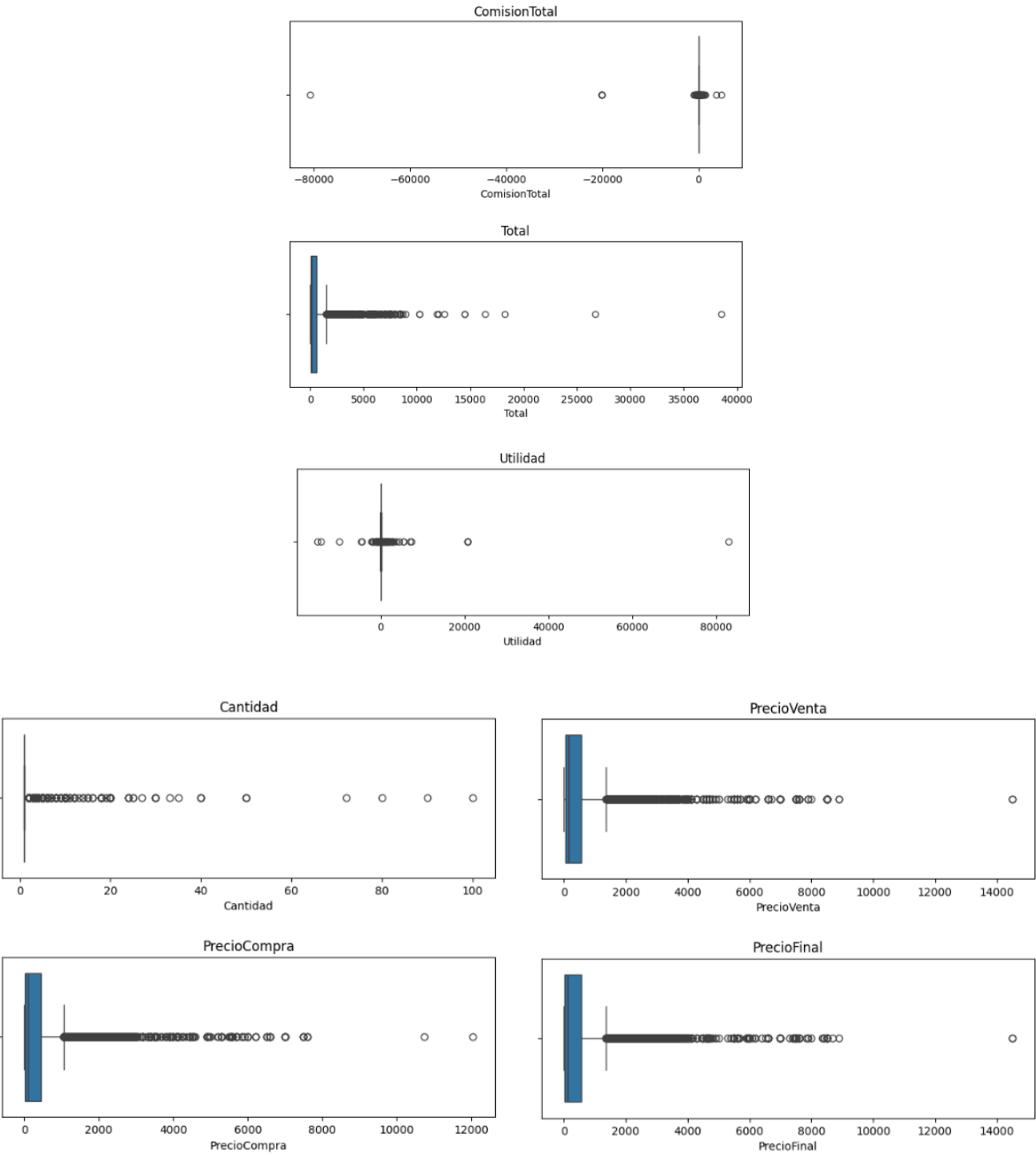
f1 = f1_score(y_test, y_pred, average='weighted')
print(f'F1-Score: {f1:.2f}\n')

# Evaluar clasificación
print(classification_report(y_test, y_pred))

# Evaluar Matriz de confusión
cm = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(cm, index=model.classes_, columns=model.classes_)
print("Matriz de Confusión:\n", cm_df)
```

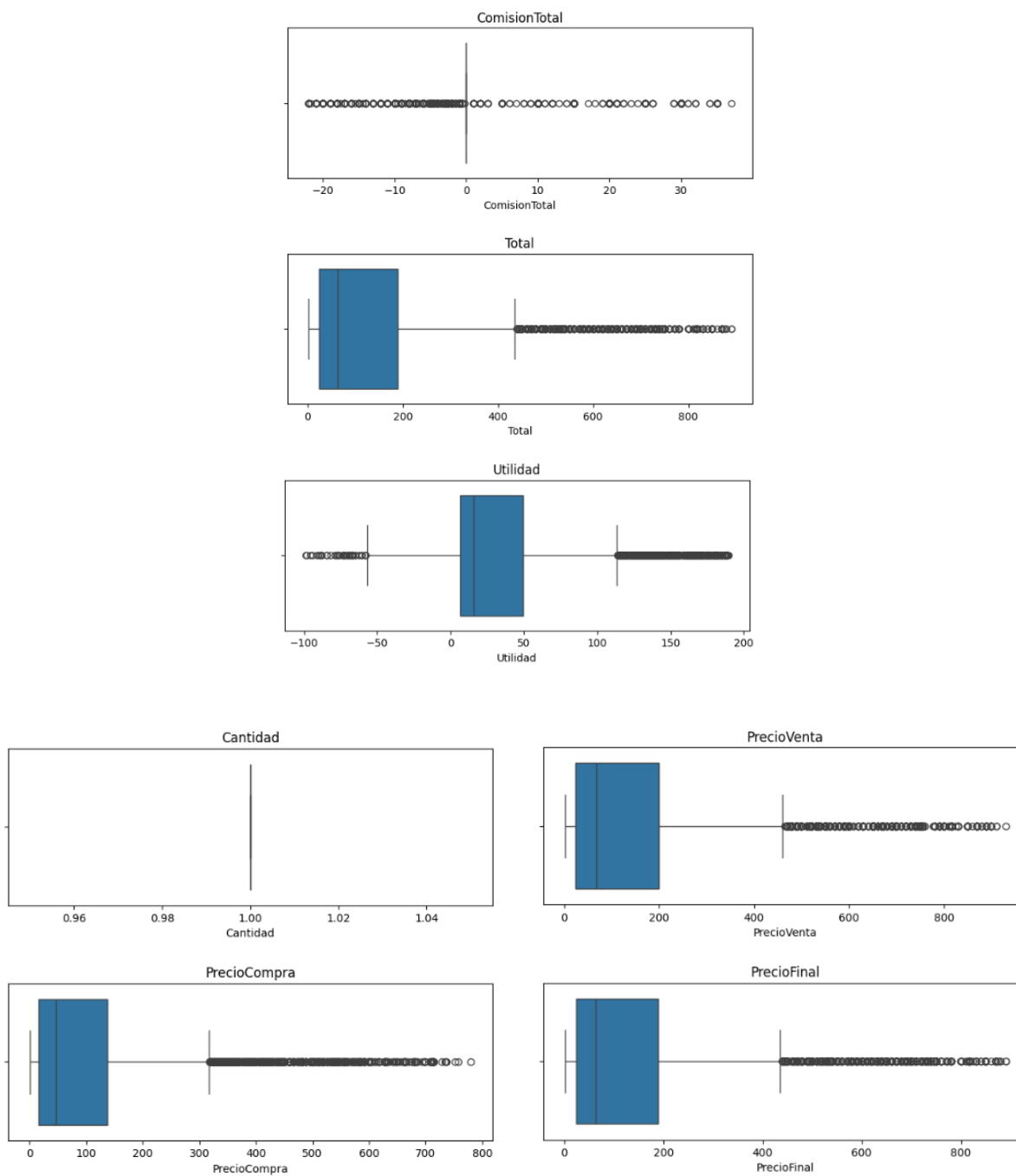
Ubicación: CD: 02\_Notebook/Modelo\_Regresión\_Logistica.ipynb

Anexo 4. Valores Atípicos de las variables numéricas



Ubicación: CD: 02\_Notebook/Proyecto\_Final\_BI.ipynb

#### Anexo 4. Resultados corrección de los Outsiders



Ubicación: CD: 02\_Notebook/Proyecto\_Final\_BI.ipynb

## Anexo PRINCIPAL: CD

El anexo contenido en el CD proporciona una compilación detallada de datos complementarios, gráficos explicativos y documentos adicionales que respaldan y enriquecen el contenido principal del proyecto.

Estos mismos datos se encuentran almacenados en el repositorio GitHub con la siguiente dirección:

[https://github.com/xKarlozx/Proyecto\\_Final\\_Diplomado.git](https://github.com/xKarlozx/Proyecto_Final_Diplomado.git)



Descripción del contenido:

### 01\_BaseDeDatos

- Formulario\_encuesta.csv
- VentasRango.xlsx

### 02\_Notebook

- Modelo\_Random\_Forest.ipynb
- Modelo\_Regresion\_Logistica.ipynb
- Proyecto\_Final\_BI.ipynb

### 03\_VisualizacionesPowerBI

- MapaBolivia3.svg
- ProyectoMasterTools.pbix
- VentasRangoLimpio.xlsx
- logo.png

### 04\_Documentación

- Perfil proyecto 2024.docx