Consultoría a una tienda de materias primas

Axel Daniel Malváez Flores
Gabriel Zadquiel Peralta Rionda

Proyecto Final

Índice

Marco teórico de Calidad de Datos y Minería de Datos

La calidad de datos se refiere a la medida en que los datos son precisos, completos, consistentes, confiables y relevantes para su uso previsto. La falta de calidad de los datos puede tener graves consecuencias para las organizaciones, incluyendo la toma de decisiones inexactas, la pérdida de clientes, la disminución de la eficiencia y el aumento de los costos. Para mejorar la calidad de los datos, es necesario seguir un proceso sistemático que incluya las siguientes etapas: realizar un perfilado, identificación y detección de los problemas de calidad de datos, cuantificar y definir de los estándares de calidad de datos, implementación de controles de calidad de datos y gestión continua y monitoreo de la calidad de los datos. Este proceso debe involucrar a todas las partes interesadas y debe ser monitoreado y ajustado regularmente para asegurar que se cumplan los objetivos de calidad de datos.

Por otra parte tenemos la minería de datos que es el proceso de extracción de información útil y conocimientos ocultos en grandes conjuntos de datos utilizando ya sea técnicas estadísticas, de aprendizaje automático, de inteligencia artificial y de visualización de datos para descubrir patrones, relaciones y tendencias en los datos. La calidad de datos es una parte fundamental para que la minería de datos sea exitosa, ya que los resultados de la minería de datos serán tan buenos como la calidad de datos que tengamos. La selección y preparación de los datos deben incluir la evaluación de la calidad de los datos, la identificación y eliminación de valores atípicos y datos faltantes, y la normalización y estandarización de los datos. El modelado de datos implica la selección de las técnicas de minería de datos adecuadas, como clasificación, regresión, clustering o asociación, y la aplicación de estas técnicas a los datos para identificar patrones y relaciones. La evaluación y validación del modelo se lleva a cabo para determinar su precisión y su capacidad para predecir y generalizar a nuevos datos. Para garantizar el éxito de la minería de datos, es importante tener un equipo de expertos en minería de datos y conocimiento de dominio en el área en la que se están aplicando las técnicas. También es importante tener datos de alta calidad y un marco ético para garantizar la privacidad y seguridad de los datos. La minería de datos también puede ser utilizada para mejorar las ventas de una empresa, identificando patrones en la forma en que los clientes compran los productos y permitiendo a la empresa optimizar su estrategia de ventas. Esto puede incluir la segmentación del mercado, la personalización de ofertas y la identificación de oportunidades de venta cruzada y venta adicional.

En general, la minería de datos y la calidad de datos pueden mejorar la productividad de una empresa al proporcionar una mejor comprensión del segmento de mercado y de las necesidades de los clientes, así como al permitir la optimización de las estrategias de ventas. Es importante tener en cuenta la calidad de los datos en todo momento, desde la recolección hasta el análisis, para garantizar resultados precisos y confiables.

Objetivo del proyecto

Cada día la competitividad dentro del mundo empresarial es más grande, por ello las pequeñas, medianas y grandes empresas han tenido que adoptar nuevas estrategias para poder adaptarse a las tendencias del mercado actual. Para este proyecto, decidimos ayudar a un negocio de dulces y materias primas con nuestra experiencia y las herramientas necesarias en el ámbito de calidad de datos con la finalidad de que el negocio pueda aprovechar más sus recursos y obtener beneficio de sus datos como sus ventas.

Para poder orientar a este negocio, el dueño nos proporcionó los datos con los que él trabaja, algunos de los datos proporcionados fueron las ventas realizadas en el año 2022, el inventario de la tienda, entre otros. El propósito del proyecto será generar valor con los datos para generar más ingresos al igual que mejorar las decisiones que se toman en el negocio.

Problemáticas detectadas

Comenzamos realizando una investigación sobre el negocio, qué vende, cuánto vende, cómo lo hace, qué software utiliza, qué base de datos utiliza, dónde está alojada, entre otras preguntas inherentes al entendimiento del negocio. El propósito de la realización de esta investigación fue obtener un mejor entendimiento del negocio y durante este proceso se han identificado diversas problemáticas que están afectando su rendimiento. Entre ellas, se destaca la inexperiencia de los empleados en el negocio, lo que se traduce en un bajo rendimiento en sus tareas. Esto se debe, en gran medida, a que los empleados no poseen el conocimiento necesario para recomendar productos adecuados durante la venta.

Asimismo, se ha detectado un desconocimiento por parte de la empresa en relación a la cantidad de productos que deben ser abastecidos o desabastecidos en función del aumento o disminución de las ventas de ciertos productos en diferentes temporadas del año. Esto puede generar pérdidas significativas para el negocio, al tener un exceso o escasez de inventario. Además, se ha notado que algunos productos han dejado de venderse con la misma frecuencia que antes, por lo que se hace necesario determinar cuáles son los más convenientes para retirar del inventario y así optimizar los recursos disponibles.

Propuesta de solución.

Para solucionar las problemáticas del negocio detectadas, se emplearán distintas técnicas del área de ciencia de datos con los datos proporcionados por la empresa para que se pueda dar valor a los datos resolviendo estos problemas detectados.

Primordialmente es necesario utilizar las herramientas vistas en la clase de calidad y preprocesamiento de datos para revisar el estado actual de los datos. Para poder trabajar con los datos de manera adecuada es necesario realizar las siguientes actividades:

- Identificación de las fuentes relevantes: Para poder iniciar el proyecto, inicialmente es necesario identificar las fuentes relevantes que se nos proporcionan en los datos,

ya que los datos con los que contamos, aunque son proporcionados por una sola fuente de datos, en los datos se nos presenta una gran cantidad de tablas y en estas tenemos que identificar aquellas que son relevantes para resolver las problemáticas identificadas en el negocio.

- Extracción y perfilado de datos relevantes: Este punto se centra en extraer los datos relevantes de las fuentes originales y con ellos darles un diagnóstico en términos de completez, correctes, relevancia, falta de programación de reglas de negocio para asegurar integridad y consistencia en los datos.

- Limpieza de datos relevantes: Para la limpieza de los datos las actividades principales que se tienen que realizar son: Completar datos, es decir tenemos que rellenar los valores nulos o aquellos que los datos que estén mal capturados. Por otro lado, otra labor importante que se tiene que realizar es la deduplicación de los datos relevantes, este concepto refiere a que tenemos que identificar aquellos registros que representan a la misma información en el conjunto de datos.

- Preprocesamiento de datos: Esta parte del tratamiento de datos es crucial para la aplicación de los algoritmos, por ello al tener un entendimiento de los datos se debe escoger la mejor transformación de los datos para ver si la discretización, normalización, reducción de dimensión o estandarización es lo más adecuado para los datos.

Ya con el procesamiento de datos realizado, los datos ya son adecuados para aportar valor a la empresa, es por ello que con los datos procesados podemos proponer las siguientes soluciones a las problemáticas de la empresa:

1. Identificar los productos que se venden bien en conjunto para poder utilizar un algoritmo que nos ayude a recomendar dichos productos a los clientes sin la necesidad de experiencia del negocio por parte del vendedor. Esto facilitará la labor de venta de los vendedores ya que no necesitarán un conocimiento pleno de los productos.

2. Analizar las ventas semanales y mensuales para conocer la tendencia de venta de los productos a lo largo del tiempo, con la finalidad de poder estimar la fecha del desabastecimiento de ciertos productos o tener preparado todo el inventario para fechas de grandes ventas para no sufrir desabastecimiento de productos importantes.
3. Realizar un algoritmo de clusterización de productos para poder entender cuales son los grupos de productos que maneja la tienda, esta agrupación no tiene que centrarse en agrupar por tipo de producto, sino las características en venta que está teniendo este producto, esto nos ayudará a identificar rápidamente cuáles son los productos que tenemos que retirar del catálogo.

Arquitectura de solución

El plan de trabajo que se realizará para el proyecto es aquel que se presenta en la Figura 1, la cual nos presenta la arquitectura solución, este es el plan de trabajo el cual se seguirá para realizar el procesamiento de datos y la realización de algoritmos.

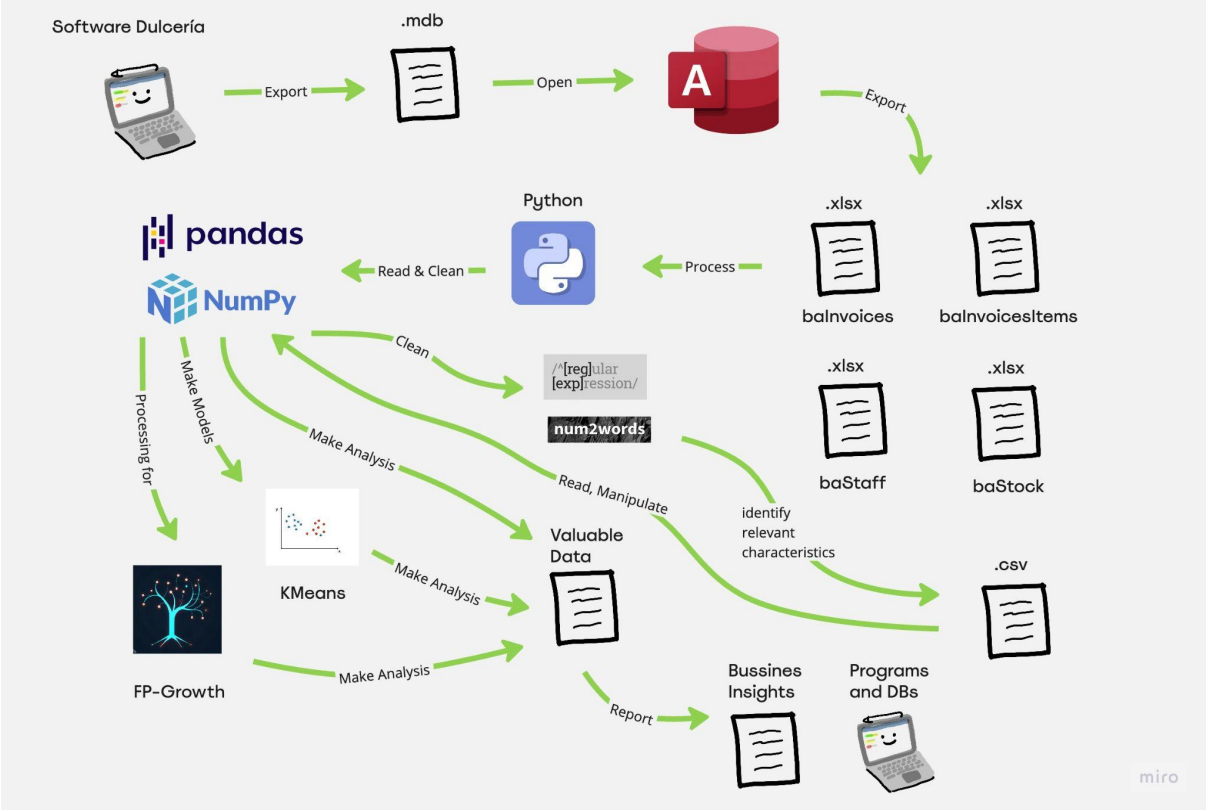


Figura 1 : Arquitectura solución de nuestro proyecto

Etapas de la arquitectura solución:

1. La primera etapa del proyecto nos presenta la recopilación de los datos, los datos son adquiridos del software con el que trabaja la dulcería. Estos archivos son adquiridos en un formato .mdb, los cuales se tiene que abrir en el software Access. Dentro de este software se hizo una elección de las principales tablas que contenía el conjunto de datos, ya que la base de datos proporcionada contenía una gran cantidad de información que no era necesaria para el análisis.

2. Ya con la identificación de las tablas más significativas para el análisis, las tablas que se consideran para la realización del proyecto fueron las tablas de baInvoices (Tabla que contiene la información de las ventas realizadas), baInvoicesItems (Tabla que contiene la información de los productos vendidos), baStaff (Tabla que contiene la información del personal activo en la tienda) y baStock (Tabla que contiene la

información del inventario de la tienda). Estos datos fueron exportados a formato .xlsx para facilitar el procesamiento de los datos.

3. Los datos en formato .xlsx se trabajaron en python, en este lenguaje de programación se trabajó con las librerías de numpy y pandas para realizar una exploración, procesamiento de limpieza y transformación de los datos.

4. El primer paso del procesamiento de datos fue la realización de un diagnóstico en términos de completez y correctitud de los datos. Ya con el diagnóstico de los datos realizado, la primera acción de limpieza que se realizó fue eliminar las columnas innecesarias que nos presentaban cada una de las tablas del conjunto de datos, ya que muchas de las columnas que tenían almacenadas estaban con una grán cantidad de valores nulos o algunas presentaban información irrelevante.

5. La siguiente tarea fue la limpieza y transformación de los datos, para realizar estas tareas fue la revisión de que los datos representarán la información correcta (los valores que no tenían la información correcta se investigó con el dueño los correctos y se modificó los valores) , también la revisión de valores duplicados. Ya con los datos revisados y modificados, estos se exportaron en archivos .csv para continuar con con la arquitectura de solución planteada.

6. Ya con los datos limpios, a algunos datos se les aplicó las transformaciones necesarias para la aplicación de los modelos.

    a. **K-Means**: Para la aplicación de este modelo se realizó una adquisición de las características principales y estas características se les aplicó una estandarización con Min-Max escaler y una reducción de dimensionalidad con PCA y con TSNE.
    b. **FP-Growth**: Para la aplicación de de este algoritmo se les hizo la transformación necesaria a los datos para que estos fueran admisibles en el algoritmo fp-growth.
    c. Se aplicó un procesamiento de los datos para que se pudieran extraer los patrones de datos de la empresa.

Esquemas sugeridos a todas y cada una de las fuentes de datos

Teniendo en cuenta las 4 tablas de la base de datos que tenemos, haremos una sugerencia de una posible esquematización de la base de datos justificando el nombre y la funcionalidad de cada una de las características de cada tabla y se evaluará la estructura, relaciones y campos recomendados. Además, se discutirán las implicaciones y posibles

mejoras en el esquema propuesto. Este análisis se basará en buenas prácticas de diseño de bases de datos y consideraciones específicas del dominio de datos.

El esquema bien diseñado debe ser fundamental para garantizar la eficiencia, integridad y por supuesto la calidad de los datos. A continuación tendremos una descripción detallada de las tablas incluidas en el esquema que propusimos:

- Staff

  Dentro de las 64 columnas con las que ésta tabla contaba inicialmente, únicamente decidimos quedarnos con las esenciales las cuáles son : **StaffID, Name** y **Function**. Esto debido a que a pesar de que son las únicas que nos interesaban, las demás columnas se encontraban totalmente vacías por lo que no se tuvo información perdida. Con esto logramos en otras tablas donde se encuentran las ventas ligarlas con el vendedor que logró dicha venta dado su **StaffID**. Con esto logramos realizar una analítica para ver el rendimiento de cada vendedor y así mismo el tener tantas columnas basura nuevamente hace que el sistema del dueño de la dulcería sea lento.

- Stock

  Originalmente en esta tabla contábamos con 106 columnas, de las cuáles varias de ellas estaban totalmente vacías o fungían como banderas de **True** o **False** para el software del negocio. No obstante este tipo de columnas no eran de nuestro interés debido a que no contenían información valiosa y además existían algunas columnas que repetían en cada registro de la columna el mismo valor ya fuera booleano o numérico**.** Igualmente es necesario notar que existían columnas **Comments** que eran columnas que contenían texto sobre la cantidad de productos, sin embargo ya teníamos una columna que nos indicaba esta característica sobre los productos por lo que en este caso al ser información que requiere mayor procesamiento y cómputo. Por lo que de las 106 columnas, reducimos nuestra tabla con las características relevantes a tan solo 16 columnas lo cuál es bueno debido a que esta reducción nos ayuda al procesamiento de los datos en la parte del rendimiento computacional. Las columnas elegidas fueron : **StockID** (identificador del producto), **Family** (familia o marca a la que pertenece el prodcuto), **Item** (descripción del producto como su nombre), **CAJA_onehot** (indicador de si es caja), **BULTO_onehot** (indicador de si es BULTO)**, Costprice** (costo del producto), **QtyMinimum** (cantidad mínima que debemos de tener el producto en stock), **Barcode** (código de barras, SKU, identificador del producto), **PriceList** (Precio de lista del producto), **Package** (indicador de si viene en paquete), **QtyAvailable** (cantidad del producto actualmente disponible), **QtyAvailableForStore** (cantidad disponible para la tienda). Cada una de estas columnas tiene información importante para el negocio pues al ser una base de datos relacional podemos en cada ticket asociar cada producto con su StockID para no poner todas las características de dichos productos en las facturas. Esta tabla nos será de gran utilidad para realizar nuestro algoritmo de clustering para encontrar qué productos se asocian con cuáles y otros.

- Invoices

  En esta tabla originalmente contábamos con aproximadamente 133 columnas, lo cuál es algo considerable sin embargo, nuevamente teníamos columnas enteramente vacías, con el mismo valor repetido en cada registro, banderas, comentarios, etc. Al reducir la tabla, nos quedamos únicamente con 10 columnas las cuáles son : **InvoiceID** (identificador de la factura en cuestión), **InvoiceNumber** (número de factura), **Date Creation** (fecha de la creación de la factura), **Date Invoice** (fecha en la que se validó la factura en el sistema), **Sub Total** (valor numérico del precio pagado por el producto sin impuesto), **IVA** (valor numérico del precio pagado como impuesto), **Total** (precio final pagado), **Total letter** (precio final pagado escrito con letras), **Barcode** (código de barras de cada factura, va en el ticket). Esta es una tabla imprescindible para el análisis que queremos realizar debido a que contamos con cada uno de los tickets (sin los productos comprados) que se han generado en el negocio lo cuál nos ayudará a encontrar patrones de ventas en los consumidores.

- InvoicesItems

  Esta es una de las tablas más importantes para la analítica que realizaremos puesto que esta tabla contiene una descripción producto a producto de cada ticket generado en el negocio. Originalmente, contábamos con 33 columnas (un número relativamente pequeño), de las cuáles únicamente nos resultaron 10 relevantes. Estas fueron : **InvoiceItemID** (identificador en la tabla del registro), **InvoiceID** (identificador de la factura en la que el producto se encuentra), **Description** (descripción del producto, es la misma descripción que aparece en la tabla *Stock*), **StockID** (identificador del producto en la tabla *Stock*), **Price Unit** (precio unitario de nuestros productos), **Total_sin_iva** (precio total de nuestros productos antes de impuestos), **IVA** (impuesto del 16% aplicado al costo del producto), **Total** (precio final del costo del producto), **CostPrice** (precio del costo del producto al negocio), **Ganancia** (ganancia generada sobre cada producto, este es el margen que nos interesa).

El análisis del esquema sugerido para cada una de las tablas en nuestra base de datos nos provee de una estructura sólida para el almacenamiento de la información sin necesidad de generar información basura que solo provoca alentar el sistema debido al mayor cómputo que debe realizar.

Diagrama que presente las relaciones entre las diferentes fuentes de datos

El diagrama que presenta la relación entre los datos utilizados en el proyecto es el que se muestra en la Figura 2, este diagrama es un modelo relacional de los datos, el cual presenta 4 tablas y las relaciones que existen entre los datos. La tabla Invoices contiene la

información de las ventas realizadas en la dulcería (es decir los tickets), la tabla Invoices contiene la información de los productos vendidos en tickets, la tabla Staff contiene la información del personal de la tienda y la tabla Stock contiene la información de los productos de la tienda.

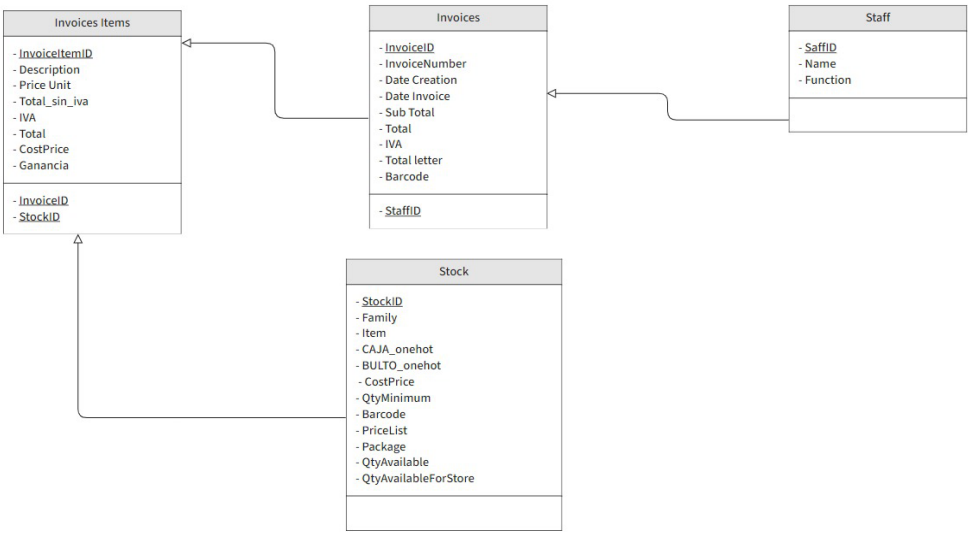


Figura 2 : DIagrama de relación entre los datos

Procedimiento realizado para la integración de cada fuente de datos

La integración de datos es el proceso de combinar y unificar datos de múltiples fuentes en una ubicación centralizada, esta acción es primordial para el correcto desarrollo de un proyecto de ciencia de datos, ya que esto permite acceder, consultar y analizar los datos de manera más eficiente. Es por ello que se consideró el siguiente procedimiento para la integración de los datos.

- *Definir los requisitos*: Los requisitos que necesitábamos para este proyecto eran obtener los datos de las ventas de la tienda sin importar el formato en que se presentarán y las distintas fuentes de datos en las que se encontraran, pues los objetivos del proyecto se plantearon con estos datos.

- *Identificar las fuentes de datos*: En nuestro proyecto en particular el negocio con el que estamos colaborando tiene 3 máquinas en las que se están generando las ventas en el negocio y existe otra máquina que contiene toda la información unificada del negocio.

- *Extracción de los datos del negocio*: La ventaja con la que contamos con este proyecto es que aunque la tienda cuenta con varios equipos de cómputo, toda la información que generan las computadoras sobre las ventas de la tienda se

encuentran unificadas en una misma base de datos. Por ello al momento de la extracción de datos se extrajo toda la información guardada en la base, esta información se exportó en formato .accdb el cual es el formato que genera el software Microsoft Access Database. Dado que contábamos con una única donde de datos, no fue necesario realizar un proceso de integración de distintas fuentes de datos.

Después del proceso de integración de datos el proyecto estaba listo para empezar, ya que después de esto ya se contaban con todas las fuentes de datos necesarias para la elaboración del proyecto.

Procedimiento realizado para el perfilado de datos

El perfilado de datos es el proceso de analizar y comprender los datos de una fuente específica, este es un procedimiento indispensable para un proyecto de ciencia de datos ya que ya que este procedimiento nos dará conocimiento de la calidad, la estructura, el contenido y las características principales de los datos.

El procedimiento realizado para realizar el perfilado de fue el siguiente:

- **Exploración inicial de los datos**: Con la base de datos del negocio se realizó una primera exploración de las tablas que podrían servir para el proyecto, de esta primera etapa se seleccionaron las tablas según su completez ya que existían muchas tablas en la base las cuales no contenían ningún tipo de información. Después de realizar este procedimiento de selección nos quedamos al final con alrededor de 50 tablas distintas.

- **Identificación de fuentes importantes**: Al conseguir estas tablas, se realizó otra selección de los datos pues necesitábamos la información más importante de las ventas del negocio y muchas de las tablas extraídas contenían información redundante o información innecesaria para el análisis ya que mucha de esta información era información que innecesaria que genera el software de la tienda. Después de realizar una selección minuciosa ya con el entendimiento del negocio y con los objetivos planteados para el proyecto se concluyó que las únicas tablas necesarias para el entendimiento de las ventas del negocio eran las que se muestran en la Figura 2.

- **Análisis de calidad de datos**: Ya con las fuentes de datos seleccionadas a estas se les realizó un análisis de calidad en el cual identificó los valores faltantes, inconsistencias, errores en el formato de datos.

- **Entendimiento del esquema de datos**: Era importante entender todos los esquemas y estructuras que nos presentaban los datos, unas columnas eran fácilmente entendibles a lo que estaban representando, pero otras columnas no era sencillo, por ello es que se preguntó al dueño de la tienda sobre estas columnas. Existían columnas en las que sí se nos dio información y otras en las que no.

Después del procedimiento de perfilado de datos teníamos un mejor entendimiento sobre la información con la que contábamos, pero este proceso se siguió aplicando a lo largo del proyecto ya que el perfilado es un proceso iterativo y continuo ya que se siguieron descubriendo distintas características de los datos a lo largo del proyecto.

Procedimiento realizado para la limpieza de datos

El procedimiento de limpieza de datos es un paso clave ya que al realizar una buena limpieza de datos mejora la confiabilidad de estos lo que conlleva a un mejor entendimiento de los datos y evita decisiones erróneas a la hora de la toma de decisiones con estos.

El procedimiento que se realizó en el proyecto de limpieza de datos es el siguiente:

- **Eliminación de columnas innecesarias**: Inicialmente las tablas que se presentan en la Figura 2 contaban con muchas más columnas que las que se presentan en el esquema, pero muchas de estas columnas presentaban información irrelevante las cuales generaba de manera automática el software de la tienda, por lo que se eliminaron estas columnas y se conservaron las que sí presentaban información de utilidad.

- **Identificación de campos duplicados:** Se exploró los datos para la identificación de campos duplicados, en las ventas se realizó con cuidado esta operación ya que era importante identificar si existían ventas duplicadas. Después de realizar la búsqueda de duplicados en los tickets se observó que no existen duplicados. Otro proceso de deduplicación importante fue con los productos, esta fue una tarea complicada ya que muchos productos son sumamente parecidos ya que en la tienda se vende el mismo producto pero en diferentes presentaciones, este fue un factor a tomar en cuenta en la deduplicación de estos registros, tomando en cuenta esto se concluyó que que tampoco existían productos duplicados.

- **Identificación de la correctez de los datos**: Durante la exploración de los productos se identificaron productos con precio negativo, por lo que se contactó al dueño de la tienda para corregir y colocar los precios reales de los productos.

- **Estandarización de los datos**: Algunas columnas de los datos no presentaban el mismo formato en todos los registros, por ejemplo en las fechas que se presentaban en los datos, por lo que se estandarizó este tipo de datos, así como otros que se identificaron que no estaban estandarizados.

- **Transformaciones de datos**: Se realizaron las siguientes modificaciones en los datos:
    - Se agregaron en los tickets el precio total sin iva, el iva del producto y la ganancia que había generado la venta.
    - En los productos se agregó la presentación con lo que se estaban contando los productos, porque algunos productos se venden por bulto y otros por caja,

por lo que se agregó una columna a la base de datos para tener una identificación más rápida de este tipo de productos.

Después de la limpieza de datos, los datos tenían mayor confiabilidad y presentaban de manera más precisa la información. Ya con la limpieza de datos realizada, los datos estaban listos para realizar los distintos algoritmos que se tenían pensados para resolución de los conflictos de la tienda y también para tener un mayor entendimiento de los patrones de venta que presentaba la tienda.

Procedimiento realizado para el análisis de los datos

Se llevaron a cabo tres tipos de análisis con los datos recopilados. En primer lugar, se realizó un agrupamiento de los productos en función de sus características de precio, ganancia y nivel de ventas. Este análisis de agrupamiento nos brinda una mejor comprensión de los diferentes productos que se ofrecen en la dulcería.

En segundo lugar, se llevó a cabo un análisis para identificar qué productos tienen un buen desempeño en ventas cuando se venden juntos. Esto nos permite entender las relaciones y las oportunidades de venta cruzada entre los productos de la dulcería.

Por último, se realizó un análisis para identificar patrones de ventas en la tienda. Este análisis nos ayuda a descubrir tendencias y comportamientos recurrentes en las ventas, lo cual puede ser valioso para la planificación de inventario, promociones y estrategias de ventas.

Con estos tres análisis, se obtiene una visión más completa y detallada de los productos y las ventas en la dulcería, lo que proporciona información valiosa para la toma de decisiones y la optimización de los resultados.

Algoritmo de agrupación: K-means

Para llevar a cabo la agrupación de productos, se procedió a identificar las principales características que se utilizarían en el proceso. Estas características consideradas fueron el número de ventas, el costo del producto, el número mínimo de unidades requeridas en la tienda y la ganancia que cada producto aporta.Estas características se seleccionaron debido a la amplia variedad de productos presentes en la tienda, donde no era relevante agruparlos por tipo, sino más bien explorar si existían agrupaciones basadas en sus características de venta.

Para la clasificación, se utilizó el algoritmo K-means. Dado que este algoritmo requiere un número inicial de grupos (clusters), se aplicó el método del codo para determinar el número óptimo de clusters para los datos en cuestión. Según los resultados de este método, la mejor agrupación se logró con 6 grupos. Además, se realizó una transformación de los datos mediante el escalamiento, una técnica importante para mejorar la eficiencia del algoritmo y garantizar una agrupación precisa.

Una vez aplicado el algoritmo de agrupación, fue necesario utilizar los algoritmos PCA y t-SNE para visualizar los resultados. Dado que se estaban clasificando los productos en función de 4 características, la visualización directa no era posible. Por lo tanto, se ejecutaron estos algoritmos para obtener una representación visual adecuada de los datos agrupados.

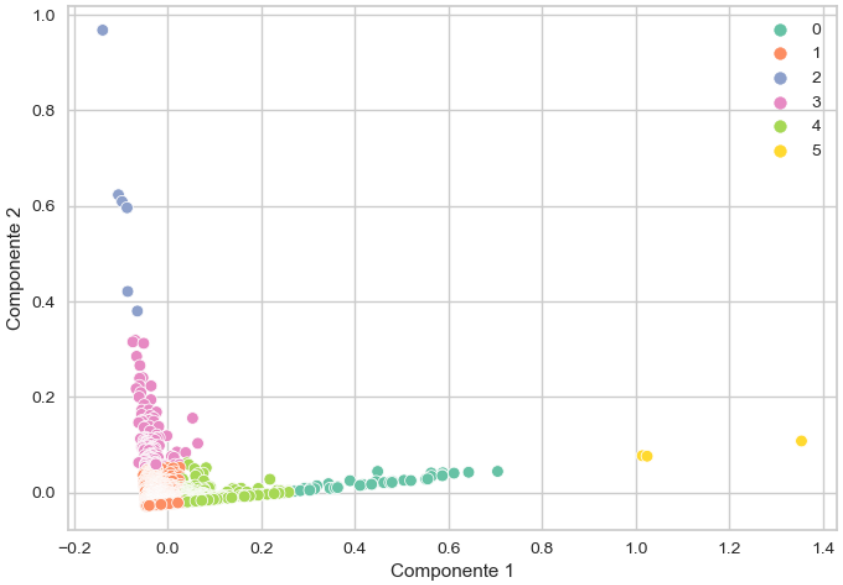


Figura 3 : Aplicación del algoritmo PCA

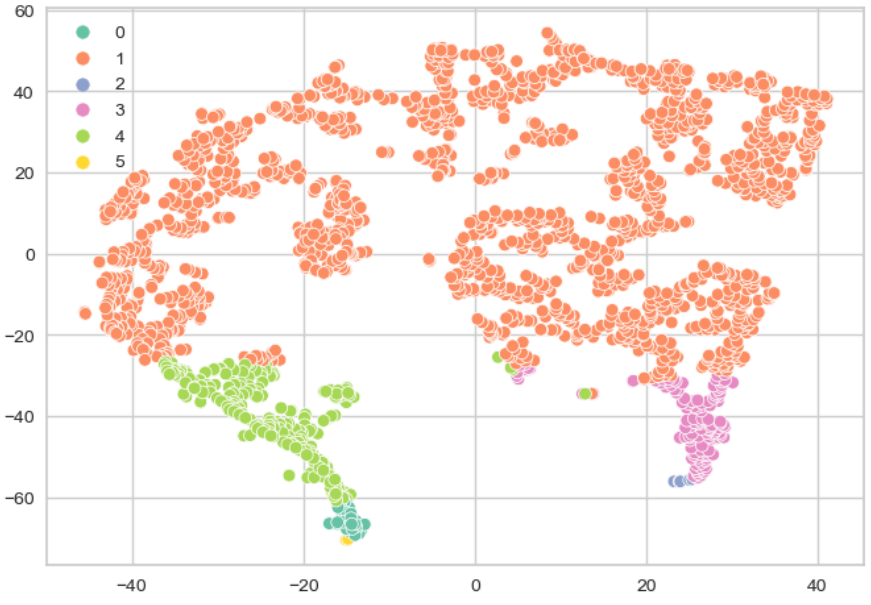


Figura 4 : Aplicación del algoritmo t-SNE

Las dos aplicaciones de los algoritmos son fundamentales para comprender la similitud y diferencia entre los distintos grupos. Por ejemplo, al observar la Figura 3, podemos apreciar la lejanía que tiene al grupo 5 y el grupo 2 del resto, lo cual nos indica que se están presentando características atípicas en comparación con el resto de los productos, por ello

es importante tener en cuenta estos productos debido a su comportamiento distintivo. Por otro lado, notamos que los grupos 0, 1, 3 y 4 muestran una mayor proximidad entre sí, lo que sugiere que comparten características similares. Sin embargo, la clasificación diferente de estos productos nos indica que cada uno tiene características significativas propias.

En cuanto a la Figura 4, se puede apreciar que el grupo más grande es el grupo 1. Es importante destacar que la agrupación mostrada en esta figura difiere de la Figura 3, pero esto no significa que se estén generando grupos completamente diferentes. Más bien, representa una forma alternativa de visualizar los datos. La elección de esta visualización se realiza debido a que los datos en la Figura 3 están muy densamente agrupados, mientras que en la Figura 4 se presenta una agrupación más uniforme de los productos, lo cual facilita una comprensión más clara de las agrupaciones generadas por el algoritmo de agrupación.

Después de visualizar la agrupación de los datos es importante darle una interpretación a los resultados que estamos obteniendo, a continuación se dará la interpretación de los grupos:

Grupo 0

Este grupo tiene en total 36 productos, este nos agrupa a los productos cuyas ventas son bajas, pero su precio es alto. Indagando un poco más en los productos de este grupo, nos damos cuenta que son productos agrupados en CAJA o BULTO. Esto nos da una idea de por qué sus ventas anuales de dichos productos son pocas y los precios altos, esto pues sabemos que es una tienda minorista.Todo esto se ve reflejado en la siguiente tabla que describe la información descriptiva del grupo.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia
mean	4.666667	0.694444	0.166667	825.063889	190.572222
std	5.286100	0.467177	0.377964	210.491861	53.818709
min	1.000000	0.000000	0.000000	510.000000	64.400000
max	27.000000	1.000000	1.000000	1391.600000	299.000000

Grupo 1

En este grupo tenemos 1637 productos, en este grupo contamos con aquellos productos que tienen ventas bajos y precios bajos. Esto nos indica que son productos que no son muy demandados por los clientes. Por consecuencia, el dueño de la tienda debería de considerar la posibilidad de retirar una parte de estos productos de su inventario o tener poco inventario de los mismos debido a que no son muy rentables y tienen poco movimiento.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia

mean	15.725718	0.001222	0.0	53.197936	15.141222
std	16.607615	0.034943	0.0	27.019856	7.528334
min	1.000000	0.000000	0.0	5.000000	0.500000
max	77.000000	1.000000	0.0	144.000000	50.000000

Grupo 2

En este grupo tenemos 6 productos, donde estos productos tienen ventas anuales muy altas, pero su precio es bajo. Esto nos da una idea de que son productos que se venden en gran cantidad, pero que no son muy caros. Al indagar un poco más en los productos de este grupo, nos damos cuenta que son productos de materias primas que son de uso cotidiano como cucharas y platos de plástico. No obstante también contamos con totis que son un producto de comida que se vende en gran cantidad y es barato.
Al ser un grupo con muchas ventas anuales ahora entendemos porque en la Figura 3 estaban separados del resto de productos. En la siguiente tabla se puede apreciar la información descriptiva del grupo.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia
mean	635.333333	0.0	0.0	18.841667	7.491667
std	211.066498	0.0	0.0	11.443357	3.978746
min	412.000000	0.0	0.0	9.600000	3.400000
max	1009.000000	0.0	0.0	38.800000	13.000000

Grupo 3

Este es un grupo tiene en total 126 productos, podemos notar que tenemos en esta agrupación aquellos productos que tienen ventas medias. Además, sus precios son bajos. Esto nos indica que son productos medianamente demandados por los clientes y que no son muy caros. En la siguiente tabla se puede apreciar la información descriptiva del grupo.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia
mean	138.063492	0.007937	0.0	45.052087	12.997988
std	61.946654	0.089087	0.0	29.299917	9.286240
min	64.000000	0.000000	0.0	0.068000	0.029000
max	351.000000	1.000000	0.0	164.060000	61.100000

Grupo 4

En este grupo contamos con 214 productos, este nos agrupa a aquellos productos que tienen ventas anuales un poco bajas y que coinciden con ser algunas cajas o bultos o productos normales. Sin embargo, sus precios están alrededor de 198 pesos. Esto nos indica que son productos que no son muy demandados por los clientes y que tienen un precio más alto que el promedio de productos. En la siguiente tabla se puede apreciar la información descriptiva del grupo.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia
mean	10.948598	0.126168	0.028037	198.539953	55.629206
std	13.693224	0.332817	0.165467	94.779279	24.818800
min	1.000000	0.000000	0.000000	0.055900	26.016000
max	83.000000	1.000000	1.000000	533.000000	175.000000

Grupo 5

Finalmente tenemos al grupo 5 en el cual solo pertenecen 3 productos, estos productos tienen ventas anuales demasiado bajas entre 5 y 7 unidades anuales. No obstante los precios de estos productos son muy altos y por ende las ganancias que se obtienen de estos son igualmente altas, al ser precios tan altos podemos ver la razón por la que se encuentran tan separados en la Figura 3. En la siguiente tabla se puede apreciar la información descriptiva del grupo.

	no_ventas	CAJA_onehot	BULTO_onehot	CostPrice	Ganancia
mean	6.0	0.666667	0.333333	1868.000000	467.000000
std	1.0	0.577350	0.577350	304.945897	76.236474
min	5.0	0.000000	0.000000	1684.000000	421.000000
max	7.0	1.000000	1.000000	2220.000000	555.000000

En resumen, podemos observar en la Figura 3 que la Componente 1 representa el precio de los productos, mientras que la Componente 2 refleja el número de ventas principalmente. Esto nos permite deducir que los grupos con mayores costos, como los grupos 0 y 5, tienden a ubicarse más hacia la derecha en el gráfico. Por otro lado, los grupos con mayores ventas, como los grupos 2 y 3, se sitúan en la parte superior, ya que son los grupos que presentan un alto volumen de ventas.

Es evidente que estos grupos mencionados anteriormente son la principal fuente de ingresos de la tienda, ya que representan productos que se venden de manera exitosa o tienen un precio más elevado. Sin embargo, los grupos 1 y 4 muestran un desempeño inferior en términos de ventas y ganancias. Por lo tanto, sería recomendable realizar una revisión cuidadosa de estos productos para determinar cuáles sería conveniente retirar del inventario.

Con el fin de facilitar esta revisión, sería beneficioso proporcionar al dueño una lista de los ID de los productos pertenecientes a estos grupos. De esta manera, el dueño podrá tomar las medidas adecuadas con respecto a estos productos y tomar decisiones informadas sobre su gestión en el inventario.

En conclusión, la Figura 3 nos proporciona información valiosa sobre la relación entre el precio y las ventas de los productos, identificando los grupos más rentables y los que requieren una atención especial. La revisión de los productos menos exitosos puede llevarse a cabo de manera efectiva al proporcionar una lista de los ID correspondientes, brindando al dueño una herramienta práctica para la toma de decisiones.

Recomendador de productos: Algoritmo FP-growth

El algoritmo FP-growth (Frequent Pattern growth) es una técnica de minería de datos utilizada para la extracción de patrones frecuentes en conjuntos de datos transaccionales o basados en transacciones (tickets de venta). Se utiliza principalmente para el análisis de asociación, que consiste en descubrir relaciones y conexiones entre diferentes elementos de un conjunto de datos, es decir, se encarga de descubrir qué productos se compran juntos con mayor frecuencia, lo que permite a los comerciantes tomar decisiones sobre estrategias de marketing.

Viendo la utilidad de este algoritmo para encontrar que productos se venden bien en conjunto se decidió aplicar este algoritmo al conjunto de dato de la dulcería para poder mejorar la labor de venta del personal para que ellos sepan que productos se venderán bien en conjunto.

Para empezar a trabajar con el algoritmo empezamos realizando una modificación a los productos que se encuentran en InvoicesItems, ya que al querer realizar recomendaciones lo más importante no es la presentación en la que se encuentra el producto sino el producto en sí, por ejemplo en la base de datos tenemos registrados productos como ‘ferrero rocher 24pz 300gr’ y ‘ferrero rocher 8pz 100gr’, nosotros sabemos que es el mismo producto, pero este viene en distinta presentación, por lo que realizamos un procesamiento de datos para transformar los productos como este a que solo contuviera ‘ferrero rocher’, esto para ver con que otros productos se venden bien este chocolate por ejemplo.

Después de realizar el procesamiento de datos y la preparación de los datos para el algoritmo, este fue aplicado con un soporte de 0.00024 y una confianza del 50%, estos valores se escogieron ya que al contar con mucha variedad de productos la tienda y una gran cantidad de tickets generados al año con valores más grandes para los parámetros anteriores

encontrábamos muy pocas reglas de asociación, con estos parámetros encontramos 300 reglas de asociación, podemos ver algunas en la Figura 5:

	antecedents	consequents
0	(reyma plato termico bio)	(classy cuchara pastelera bio)
1	(dart tapa normal)	(dart blanco)
2	(goplas tapa plana termoformada pet)	(goplas vaso cristal)
3	(dart tapa)	(dart)
4	(reyma plato pastelero bio)	(classy cuchara pastelera bio)

Figura 5 : Algunas reglas de asociación

Las reglas de asociación encontradas se visualizan en la Figura 5, esta nos indica que se han encontrado buena relación de ventas entre los productos de plato termico y cuchara pastelera, la tapa del vaso y el vaso, plato pastelero y cuchara pastelera. Estas como otras reglas el algoritmo encontró, estas reglas tienen sentido ya que si alguien compra un vaso dart (de unicel) probablemente quiera comprar la tapa también, por lo que sería buena recomendarselo al cliente cuando este realizando la compra.

Una forma en la que podría ser empleado este algoritmo es la siguiente:
Supongamos que un cliente llega y quiere comparar el producto reyma plato térmico el vendedor debería recomendar el producto classy cuchara pastelera bio, ya que estos productos se venden bien en conjunto y la venta de la cuchara se realizará con un 50% de probabilidad de ser efectuada.

Patrones de venta de la tienda

La identificación de patrones de ventas desempeña un papel fundamental en la toma de decisiones empresariales. Al comprender y analizar los patrones de ventas, las empresas pueden realizar una planificación más efectiva de sus actividades y estrategias. Estos patrones ofrecen información valiosa que puede ser utilizada en diversas áreas de la empresa, incluyendo la planificación de promociones, la gestión de los días de descanso del personal, la asignación de recursos internos entre otros.

Por ejemplo, al identificar los patrones de ventas estacionales, una empresa puede ajustar su adquisición y oferta de productos de acuerdo con la demanda esperada. Esto puede ayudar a evitar la escasez o el exceso de inventario, maximizando así la eficiencia y reduciendo los costos asociados.

Además, los patrones de ventas también pueden proporcionar información sobre las preferencias y comportamientos de los clientes. Al comprender qué productos o servicios se venden en conjunto o cuáles son más populares en ciertos períodos, una empresa puede personalizar su estrategia de marketing y desarrollar promociones específicas para

maximizar las ventas. Otro aspecto importante de la identificación de patrones de ventas es la capacidad de anticiparse a tendencias y cambios en el mercado. Al analizar datos históricos y detectar patrones emergentes, una empresa puede adaptarse rápidamente a nuevas oportunidades o desafíos, lo que puede marcar la diferencia en un entorno empresarial competitivo.

Por el momento la empresa con la que estamos trabajando solo cuenta con datos históricos del año 2022, así que haremos un análisis del comportamiento de las ventas del negocio en este año.

Las Figuras 6 y 7 nos brindan una visión detallada del comportamiento de las ventas de la dulcería a lo largo del año 2022. Estas gráficas revelan patrones interesantes y contrastantes, ya que las ventas no se comportan de la misma manera en todos los meses, especialmente en Abril, Octubre y Diciembre, que se destacan como los mejores meses en términos de venta para la tienda.

Es importante destacar que el mes de Octubre sobresale como el mes que generó mayores ganancias para el negocio. Este patrón de ventas resulta intuitivamente claro debido a que, al ser una dulcería, se esperaría un aumento significativo en las ventas durante el período de celebración del Día de los Muertos. Sin embargo, es importante señalar que la información disponible para el mes de diciembre es incompleta, ya que solo se dispone de datos hasta el día 19. Por lo tanto, no se cuenta con una imagen completa de las ventas de diciembre, lo cual limita nuestro análisis.

Por otra parte, se observa que el mes de Abril registra el mayor número de transacciones, aunque en comparación con Octubre, genera menos ganancias. Esto indica que hay una mayor afluencia de clientes en abril, pero estos realizan compras de menor cuantía. Con base en este patrón, sería recomendable considerar la implementación de promociones durante este mes, para aprovechar la gran afluencia de clientes y estimular la compra de más productos.

En contraste, el mes de Enero presenta las ventas más bajas del año, así como una menor afluencia de clientes. Por lo tanto, sería beneficioso implementar promociones atractivas durante este período, con el objetivo de atraer a más personas a la tienda y aumentar las ventas.

Además, se observa que los otros meses del año muestran ventas bastante similares, lo cual es un indicador positivo para la tienda, ya que no existen meses significativamente bajos en términos de ventas, excepto Enero.

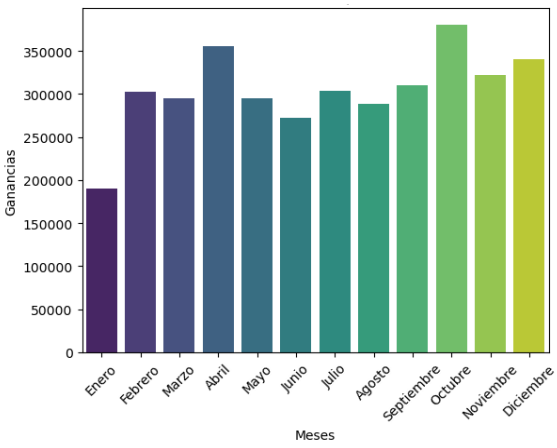


Figura 6: Total de ganancia por mes

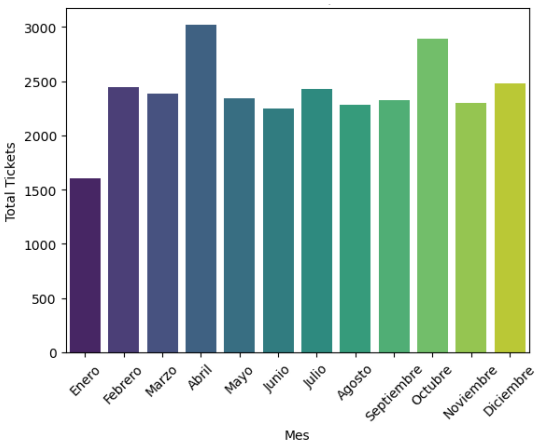


Figura 7: Total de ventas por mes

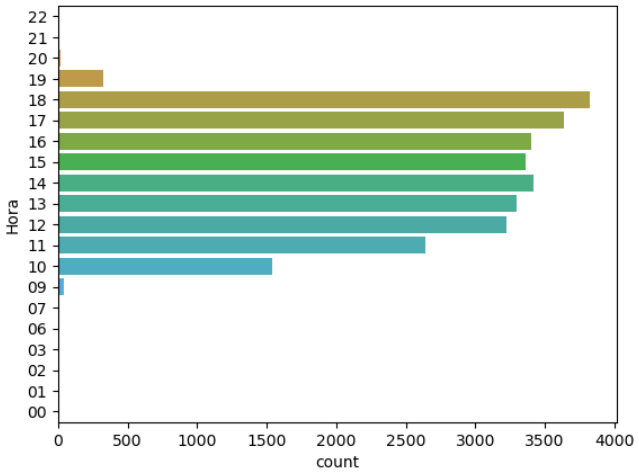
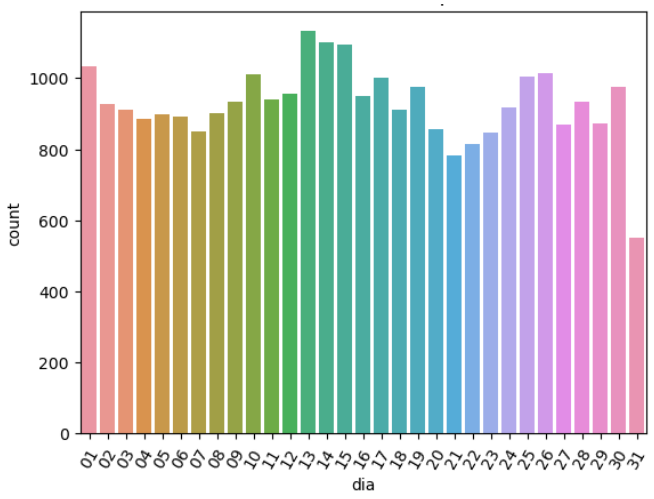


Figura 8: Transacciones en el año por hora

La Figura 8 presenta los patrones de ventas por hora de la tienda a lo largo del año 2022. Este análisis nos permite identificar las horas en las que se realizaron las ventas y observar tendencias significativas. Destaca que el horario pico de la tienda ocurre a las 18:00 horas, momento en el que se registra la mayor cantidad de transacciones. Además, se observa que las ventas van aumentando gradualmente a medida que avanza el día, pero experimentan una disminución abrupta a las 19:00 horas, que coincide con el cierre de la tienda.

Esta tendencia positiva de aumento de clientes durante el día se debe a que las personas suelen salir de sus trabajos en ese horario y aprovechan la oportunidad para realizar compras en la dulcería. Basándonos en estos patrones de venta, se puede concluir que el horario laboral actual de la tienda es adecuado, pero podría mejorarse dadas las condiciones de venta. No sería conveniente abrir la tienda antes de las 10:00 horas, al contrario, sería provechoso abrir el negocio más tarde, ya que durante estas primeras horas la tienda registra pocas ventas. Sin embargo, teniendo en cuenta la tendencia de ventas,

sería beneficioso extender el horario de cierre hasta las 20:00 horas, ya que aún podrían



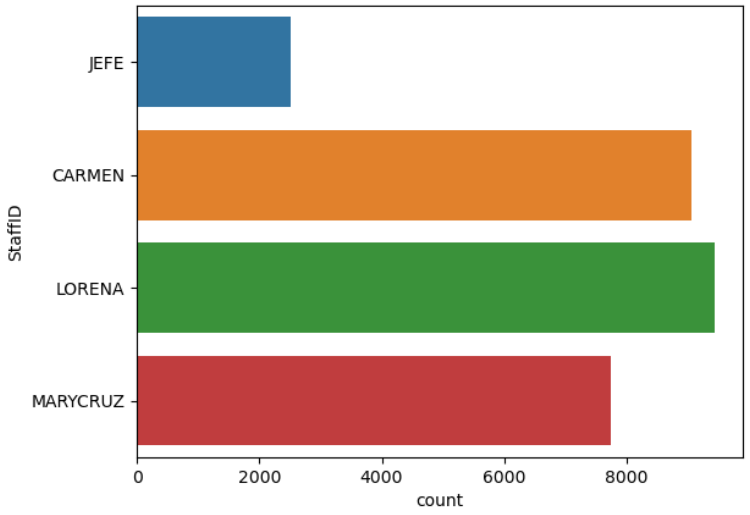
atender a algunos clientes y aprovechar oportunidades de venta adicionales.

Figura 9: Transacciones por días del mes

La figura 9 presenta los patrones de venta de los días del mes, en este es importante destacar que se generan mejores ventas los días cercanos a la quincena de cada mes, los días 31 de cada mes hay una disminución de ventas ya que no todos los meses tienen este día, pero en general las ventas se compartan de manera bastante uniforme.

Figura 11: Ventas realizadas por el personal

La Figura 11 nos brinda información sobre el desempeño de ventas del personal de la tienda. Observamos que el jefe de la tienda presenta el menor volumen de ventas, lo cual puede atribuirse a sus responsabilidades adicionales, como encargarse de la reposición de mercancía o administrar la caja. Por otro lado, destaca que Maricruz, una de las empleadas, tuvo el rendimiento más bajo a lo largo del año. Sería importante analizar las razones detrás de este desempeño inferior en comparación con sus compañeras. Es posible que Maricruz requiere más capacitación o que sea una empleada reciente en la dulcería, lo cual podría explicar esta situación. En cualquier caso, sería beneficioso brindarle el apoyo y los recursos necesarios para mejorar su desempeño y contribuir al éxito general de la tienda.



En cuanto a la Figura 12, nos proporciona una visión de los patrones de venta a lo largo de la semana. Notamos que a medida que avanzan los días, las ventas en la tienda van en aumento, alcanzando su punto máximo los sábados, que se posicionan como el mejor día en términos de ventas. Por otro lado, durante los días de lunes a jueves, las ventas se mantienen de manera bastante similar y estable. Esta información es valiosa para la planificación y la toma de decisiones en la tienda, ya que permite identificar los días de mayor afluencia y demanda, lo cual puede influir en la asignación de personal, la gestión de inventario y la implementación de estrategias promocionales.

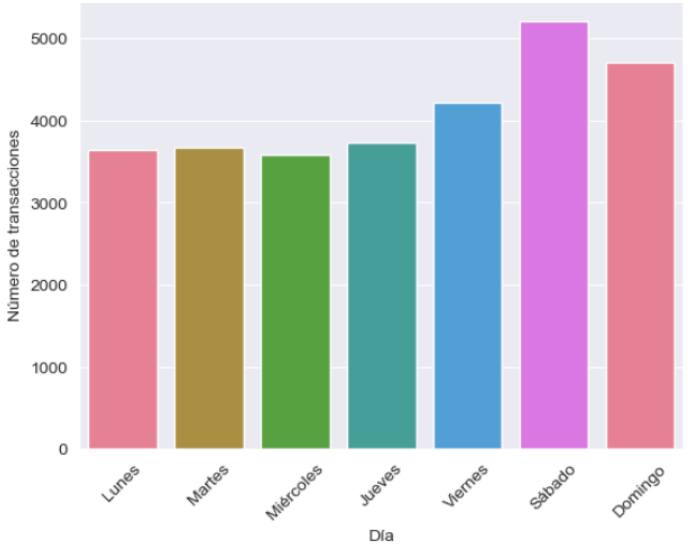


Figura 12: Transacciones por los días de la semana

Conclusiones

Este fue un proyecto basado en mejorar y generar conciencia sobre el impacto que tiene la calidad de datos en un negocio, en este caso de dulces y materias primas el cuál ha sido exitoso en la generación de información valiosa para la toma de decisiones. Utilizando modelos de clústering de Machine Learning, pudimos agrupar e identificar productos con una alta y baja rentabilidad. Esto nos permitirá optimizar el inventario y reducir los costos asociados a los productos que son de baja demanda. Pues obtuvimos que el menor margen de ganancia se da con respecto a ventas de BULTOS o CAJAS pues anualmente contamos con 168 ventas las cuáles nos generan un 24.77% de margen de utilidades. Esto nos hace pensar que debemos cambiar la estrategia de venta con respecto a este tipo de productos ya que estamos vendiendo muy poco y generando un margen muy bajo. Sin embargo, obtuvimos un segundo grupo el cuál cuenta con aproximadamente 1600 productos y el cual es el grupo con mayores ventas contando con 25743 ventas anuales, no obstante el margen de utilidad es del 28.26% lo cuál necesitaríamos hacer un análisis más profundo sobre este grupo para identificar aquellos productos que no se venden para quitarlos del inventario. Ahora con respecto al grupo 2, tenemos que este es el grupo con un mayor margen de ganancia con el 41.05% y con un total de 3812 ventas anuales lo que nos dice que este tipo de productos debe de mantenerse en el inventario. Posteriormente el grupo 3 cuenta con 106 productos los cuales tienen unas ventas anuales totales de 17396, sin embargo son

productos de bajo costo dejándonos 28.99% de margen de utilidad. Ahora para los últimos grupos el 4 tiene un total de 2343 ventas anuales con un margen de utilidad del 30.17%, sin embargo son productos que nos dejan poca ganancia por lo que debemos de hacer un análisis más profundo e identificar aquellos productos que quitaremos del inventario. Por último el grupo 5 cuenta con tan solo 3 productos el cuál tuvo en total unas 18 ventas anuales las cuáles generaron un 25% de margen de utilidad, esto nos dice que al ser productos CAJA, deberíamos de considerar el vender a mayoreo, pues no es el fuerte del negocio.

Por otro lado, el análisis de las horas más concurrentes dentro del negocio nos proporciona una visión más clara de los momentos del día con mayor afluencia de clientes, lo que nos permite ajustar el personal y los recursos en función de la demanda real. Del mismo modo, el análisis de los meses y días tanto más como menos rentables nos permite identificar patrones estacionales y realizar estrategias de promoción y marketing más efectivas. Dentro de este análisis fue bastante notorio que **Abril**, **Octubre** y **Diciembre** fueron los 3 meses con más ventas generadas con un total de ventas de $355,511.91, $380,521.70 y $340,585.23. Derivado de esto, podemos concluir que la temporada de primavera y otoño, representada por los meses de abril y octubre respectivamente, son periodos de alta demanda y generación de ingresos para nuestro negocio. Estos datos nos permiten planificar adecuadamente nuestras estrategias de abastecimiento, marketing y promociones, aprovechando al máximo estas épocas de mayor afluencia de clientes. Asimismo, el mes de diciembre destaca como un mes importante debido a la temporada navideña, donde se observa un incremento significativo en las ventas. De igual manera la implementación del recomendador utilizando el algoritmo FP-Growth nos brindará la capacidad de asociar productos basados en las ventas y las preferencias de nuestros clientes. Esto mejorará la experiencia del cliente al ofrecer sugerencias más personalizadas y con una alta probabilidad de compra del producto recomendado. Dado que aún no está en funcionamiento no podemos cuantificar el impacto en el negocio en ventas, no obstante una predicción que podemos dar es que esperaríamos que el mes de Diciembre del próximo año al realizar las ventas con el tendremos $880 pesos de ganancia neta más en este mes.

Finalmente, propusimos un nuevo esquema para la base de datos, que permitirá un mejor almacenamiento y acceso a la información necesaria para el proyecto. La implementación del esquema es un paso importante para evaluar el impacto de las mejoras realizadas y realizar ajustes adicionales en el futuro. En resumen, el proyecto de calidad de datos ha generado información valiosa que nos ayudará a tomar decisiones más informadas y optimizar la eficiencia y rentabilidad del negocio de dulces y materias primas. La implementación y evaluación continua del proyecto nos permitirá seguir mejorando y adaptándonos a las necesidades cambiantes del mercado, comprometiéndonos a realizar esta evaluación y monitoreo.