Universidad de los Andes Colombia

Acreditación institucional de alta calidad 10 años Mineducación Resolución 582 del 9 de enero de 2015

Departamento de Ingeniería de Sistemas y Computación

Maestría en Ingeniería de Información

MINE 4101 – Ciencia de Datos Aplicada

Semestre: 2023-20

Accredited ABET Engineering Accreditation Commission

Autor(es):

- Daniela Chacón – 201910858
- Esteban Ortiz - 201913613
- Manuel Porras - 201913911
- Angie Rincón - 201114323

Fecha: Septiembre 17 de 2023

PROYECTO FINAL – PRIMERA ENTREGA

- **OBJETIVOS**
  - Proponer una solución a una problemática de una organización la cuál pueda ser abordada mediante la ciencia de datos y la elaboración de un producto de datos.
  - Realizar un entendimiento del negocio y de la problemática a solucionar.
  - Definir una primera propuesta de enfoque analítico a seguir, así como los elementos básicos del producto de datos a construir.
  - Recolectar los datos necesarios y hacer un análisis exploratorio de los mismos buscando validar su calidad y si son suficientes para la solución planteada.

- **ACTIVIDADES DEL SPRINT Y ENTREGABLES:**

En este sprint se realizará una iteración de la metodología ASUM-DM con énfasis en las fases de entendimiento del negocio, definición del enfoque analítico, recolección y entendimiento de los datos. Como entregable del sprint se debe incluir como mínimo el siguiente contenido:

1 [10%] DEFINICIÓN DE LA PROBLEMÁTICA Y ENTENDIMIENTO DEL NEGOCIO

Seleccionar la organización con la cual se trabajará así como la problemática a resolver o la oportunidad a aprovechar a través de la ciencia de datos y la construcción de un producto de datos. Documentar la información clave del negocio (estrategia, datos del sector, etc.) que sustenta la relevancia del problema o la oportunidad. Definir los objetivos del proyecto y métricas de negocio que se usarán para su validación

1.1 Descripción de la organización

El proyecto se llevará a cabo en colaboración OPAIN (Operadora Aeroportuaria Internacional S.A.). Empresa dedicada a la administración y gestión del Aeropuerto Internacional El Dorado. OPAIN opera bajo el marco de un contrato de concesión y se esfuerza por modernizar, expandir y mejorar constantemente las instalaciones y servicios del aeropuerto. Bajo este marco. Se planteó como objetivo definir un proyecto de ciencia de datos que esté estrechamente relacionado con la visión de OPAIN.

1.2 Problemática a Resolver

OPAIN está interesado en aprovechar al máximo el potencial comercial de los locales ubicados en el aeropuerto. Para lograr este objetivo es fundamental comprender en profundidad el comportamiento de dichos locales. Siendo así, es necesario desarrollar un mapeo completo del proceso de interacción entre pasajeros y los diferentes locales comerciales dentro del aeropuerto. Esto incluye patrones de compra, ubicaciones de interés, preferencias de productos, horarios de mayor afluencia, entre otros aspectos relevantes.

1.3 Objetivos del proyecto

- Analizar el comportamiento de los clientes: Comprender cómo los clientes se mueven dentro del aeropuerto, qué locales visitan, con qué frecuencia y qué compran.
- Identificar patrones de compra: Detectar patrones de compra recurrentes, preferencias de productos y servicios, así como oportunidades de mejora en la oferta comercial.
- Determinar si la ubicación de los locales comerciales dentro del aeropuerto afecta su desempeño y proponer recomendaciones de optimización.

1.4 Métricas de negocio

- Ventas netas totales: Cantidad total de ingresos generados en todas las tiendas o en una ubicación específica del aeropuerto durante un período de tiempo determinado (diarias, semanales, mensuales, etc.).
- Ventas promedio por transacción: Valor promedio de cada transacción o compra realizada en las tiendas. Esto puede ayudar a comprender cuánto gasta un cliente promedio en una visita a una marca/tienda específica.
- Ventas por hora: Análisis de las ventas por hora del día para identificar los momentos de mayor actividad comercial.
- Ventas por muelle: Análisis de las ventas por zona para identificar áreas de alto rendimiento y aquellas que podrían necesitar atención adicional.
- Tasa de conversión: Cálculo comprendido como el porcentaje de transacciones exitosas con respecto al número total de visitantes (para esto habría que hacer un cruce con la información de los vuelos que también proporciona OPAIN).
- Estacionalidad: Evaluación de las variaciones en las ventas a lo largo del año para identificar temporadas altas y bajas.

2 [10%] IDEACIÓN

Identificar los potenciales usuarios del producto de datos, los procesos que desempeñan actualmente relacionados con la problemática a resolver y sus dolores. Establecer los requerimientos del producto de datos a construir, los componentes que tendrá desde el punto de vista tecnológico así como un mockup del mismo.

2.1 Potenciales usuarios del producto de datos

De acuerdo a las necesidades que fueron comunicadas por parte de OPAIN proponemos dos tipos de clientes

- Gerentes de Operaciones de OPAIN: La eficiencia operativa y la optimización de distribución de tiendas y servicios dentro del aeropuerto son prioridad para el equipo de OPAIN. El producto de datos a desarrollar puede ayudar a tomar y justificar decisiones importantes en este ámbito.
- Propietarios de locales comerciales: Las marcas en cuestión buscan información relevante que les permita mejorar sus operaciones y aumentar las ventas.

2.2 Procesos y dolores actuales

- Adquisición y unificación de datos: Datos provenientes de diversas tiendas llegan en diferentes formatos y estructuras, lo que dificulta su integración y posterior análisis. La falta de comunicación y estandarización de los datos entrantes requiere un esfuerzo manual adicional. Un mal manejo de esta unificación manual puede llegar generar errores en los insights derivados.
- Contexto de eventos externos: La falta de una visión integrada que combine datos de ventas y comportamiento de los pasajeros con eventos externos, como retrasos de vuelos, condiciones meteorológicas o eventos especiales en la ciudad.

- Falta de contexto sobre los compradores: Existe una carencia de información demográfica, de destino, entre otros factores, que impide hacer un análisis de segmentación más refinado de los compradores. Sin información detallada sobre quiénes son los compradores, es difícil desarrollar estrategias de segmentación que podrían ser útiles tanto para OPAIN como para las marcas.

2.3 Requerimientos del producto de datos

2.3.1 Funcionales

- Unificación de datos: El sistema debe ser capaz de recibir datos de diferentes tiendas en múltiples formatos y estructuras, y unificarlos en una base de datos centralizada.
- Permitir el análisis en tiempo real de métricas como ventas totales, ventas por transacción, ventas por hora y ventas por muelle.
- Incorporación de datos externos: Capacidad para integrar información sobre eventos externos como retrasos de vuelos, condiciones meteorológicas, y eventos especiales en la ciudad.

2.3.2 No funcionales

- Rendimiento: El sistema debe ser rápido y eficiente en el procesamiento y análisis de los datos.

2.4 Componentes tecnológicos

- Base de datos centralizada: Utilizar una base de datos como PostgreSQL para almacenar los datos unificados.
- ETL (Extract, Transform, Load): Implementar un proceso ETL para extraer, transformar y cargar los datos de las diferentes tiendas al sistema centralizado.
- Interfaz de Usuario: Crear un dashboard usando una herramienta como Power BI para visualizar las métricas y análisis.
- APIS para datos externos: Usar APIs para integrar datos externos como condiciones meteorológicas, eventos en la ciudad, etc.

2.5 Mockup del producto

- Página principal (Panel de métricas): Muestra un resumen de las métricas clave como ventas totales, ventas por transacción, ventas por hora, y ventas por muelle.
- Pestaña de unificación y análisis de datos: Subsección que permite a los usuarios filtrar y visualizar los datos provenientes de distintas tiendas.
- Pestaña de eventos externos e impacto en ventas: Línea de tiempo o calendario que muestra eventos como retrasos de vuelos, condiciones meteorológicas y eventos especiales en la ciudad. Esta línea de tiempo será acompañada de gráficos que muestren cómo estos eventos afectan las ventas.
- Funcionalidades adicionales: Usuarios y Roles: Opciones para gestionar el acceso según el tipo de usuario (Gerente de Operaciones de OPAIN, Propietarios de locales comerciales).

3 [10%] RESPONSIBLE

Identificar las posibles implicaciones éticas, de privacidad, confidencialidad, transparencia, aspectos regulatorios, entre otros, a considerar con el uso de datos y técnicas de IA en el contexto particular de la problemática abordada.

- Es fundamental que los insights obtenidos no sean sesgados y favorezcan injustamente a ciertas tiendas o marcas respecto a otras, especialmente si se van a tomar decisiones operativas.
- Debe quedar claro para las marcas que se recopilan, procesan y utilizan sus datos. Toda parte interesada debe tener un entendimiento claro de estos procesos.
- Se realizó un acuerdo de confidencialidad entre todos los integrantes del equipo, junto con el docente de la materia, y el representante legal de la SOCIEDAD CONCESIONARIA OPERADORA AEROPORTUARIA INTERNACIONAL S.A. – OPAIN

S.A, en el que se prohíbe la reproducción, divulgación de los datos entregados; de igual manera se garantiza que la información entregada será utilizada únicamente con fines académicos para este curso.

- Los datos entregados por la SOCIEDAD CONCESIONARIA OPERADORA AEROPORTUARIA INTERNACIONAL S.A. – OPAIN S.A se encuentran anonimizados, es decir, los nombres de las marcas contenidas dentro del conjunto de datos cuentan con un título genérico y una descripción general de la prestación de sus productos y/o servicios.

4 [15%] Enfoque analítico

Definir las hipótesis o preguntas de negocio que guiarán el proceso de experimentación. Proponer las técnicas estadísticas, de visualización de datos y/o de machine learning que se aplicarán para dar respuesta a dichas preguntas. Plantear las métricas que se utilizarán para evaluar la calidad del modelo.

4.1 Preguntas de negocio

- ¿La ubicación de la tienda X dentro del aeropuerto tiene un impacto significativo en sus ventas netas?
- ¿Existen patrones temporales (diarios, semanales, mensuales) en las ventas totales de ciertos locales comerciales?
- Eventos específicos (feriados, conciertos, partidos de futbol, ¿etc.) afectan significativamente las ventas?
- ¿Hay alguna correlación entre las condiciones climáticas y/o los retrasos en vuelos con la actividad comercial del aeropuerto?

4.2 Técnicas estadísticas de visualización de datos y/o machine learning

En este contexto empresarial, la aplicación de técnicas de aprendizaje supervisado, como la construcción de un modelo de regresión, representa una herramienta invaluable para anticipar y comprender mejor las dinámicas de ventas. Este modelo, alimentado con datos históricos que abarcan diversas condiciones como la fecha, ubicación, tipo de tienda, marca y más, tiene el potencial de revelar patrones y relaciones complejas en los datos. Esto no solo ayudaría a pronosticar con mayor precisión si se producirá un cambio sustancial en las ventas, sino que también proporcionaría información valiosa para la toma de decisiones estratégicas, como la optimización de inventario, la gestión de precios y la planificación de campañas de marketing, mejorando así la eficiencia y la rentabilidad del negocio.

Con el fin de evaluar el modelo generado, se utilizarán métricas asociadas a la tarea de regresión, tales como MAE, MSE y RMSE. Consideramos que este último es el más adecuado preliminarmente para tomar como fuente de entrenamiento del modelo.

5 [10%] Recolección de datos

Describir las fuentes de datos a utilizar en función de su estructura y utilidad para la solución del problema o aprovechamiento de la oportunidad identificada.

Para efectos de este proyecto, se trabajará con los datos entregados por los stakeholders, es decir OPAIN S.A. Contamos con varias fuentes de datos entregadas y trabajaremos sobre las siguientes:

5.1 Descripción de marcas

Archivo de texto en formato .docx que provee información acerca de las marcas (de forma anonimizada). En caso de que la marca cuente con más de una tienda dentro del aeropuerto, se encontrará una descripción de los productos y/o servicios ofrecidos exclusivamente allí, que pueden ser diferentes a los de otras tiendas pertenecientes a la misma marca. Esta información es relevante para efectos del proyecto ya que podremos diferenciar las marcas con respecto a su descripción y los servicios que ofrecen o productos que venden. Así pues, luego del tratamiento de datos contaremos con los siguientes campos

Atributo	Descripción	Tipo de Dato
Marca	Nombre de la marca (anonimizado)	String
Descripción	Descripción de los productos y/o servicios que provee la marca	String
Tienda	Nombre de la tienda asociada a la marca, en caso de que esta cuente con varias ubicaciones dentro del aeropuerto	String
Descripción Tienda	Descripción de los productos y/o servicios ofrecidos en dicha tienda	String

5.2 Mapa de zonas del programa

Cada marca posee uno o varios mapas del aeropuerto, que cuentan con las puertas de embarque cercanas a estas, el número de tienda y la ubicación de esta dentro de la zona delimitada del mapa. Aunque la información de estas imágenes no será utilizada en el procesamiento de datos, si será de gran ayuda para entender como están ubicadas las tiendas con respecto a las puertas de embarque

5.3 Información bases de datos comercial

Esta es la fuente de datos principal para el desarrollo del proyecto; contamos con dos archivos en formato .xlsx con información de ventas. Cada hoja contiene la información discriminada por marca, y cuenta con columnas que, aunque son similares entre ellas, tienen etiquetas diferentes o pueden tener registros faltantes.

Esta información es importante ya que es crucial averiguar el flujo de ventas en distintos rangos de tiempo, días, ubicación etc.

El diccionario de datos es el siguiente:

Atributo	Descripción	Tipo de Dato
Fecha de Venta/ Fecha	Fecha en la cual las transacciones fueron realizadas	String (mm/dd/yyy)
Centro de Consumo/ Tienda/ Cód Tienda	En caso tal que la marca cuente con varias tiendas, se indica el id de esta. De lo contrario, está registrado el nombre de la marca	String
Ventas Netas/ Ventas/ Venta/ Ventas sin IVA	Valor total de las transacciones realizadas en la hora específica.	Float
Zona/ Muelle/ Ubicación	Área del aeropuerto donde se encuentra ubicada la tienda	String
Hora/ Hora Entera/ Hora de venta	Los registros están discriminados por hora. Muestra la hora donde dieron lugar las transacciones indicadas. Existen registros donde esta hora también incluye los minutos exactos de la transacción	Int
Personas que ingresan	Número de personas que ingresan a la tienda en la hora indicada	Int
Transacciones/ Tickets	Número de transacciones realizadas en la hora indicada	Int
Unidades/ Piezas	Número de productos y/o servicios vendidos.	Int
Categoría	Categoría del producto vendido	String
Subcategoría	Subcategoría del producto vendido	String
Descripción del producto	Producto vendido	String
Semana/ # Semana	Semana del año en que las transacciones fueron ejecutadas	Int
Año	Año en el cual fueron ejecutadas las transacciones	Int

Mes	Mes del año en que las transacciones fueron ejecutadas	Int
Día	Día del año en que las transacciones fueron ejecutadas	Int
Mes (2) / Mes	Mes del año en que las transacciones fueron ejecutadas	String

6 [35%] ENTENDIMIENTO DE LOS DATOS

Generar un reporte de análisis exploratorio y calidad de los datos recolectados. Deben ser evidentes las diferentes técnicas de análisis univariado/multivariado y gráficas/no gráficas utilizadas, así como el análisis de calidad desde sus diferentes dimensiones.

Con lo expuesto en el anterior punto, se plantea en este apartado poder unificar toda la información de las 10 marcas. Por lo tanto, decidimos generar una nomenclatura unificada para ciertas columnas que representan lo mismo y que estás estén en minúscula y sin espacios.

Las principales columnas del dataset que se agruparon de acuerdo a lo planteado en el anterior punto serían: fecha_de_venta, tienda, ventas, muelle

Columnas
fecha_de_venta
tienda
ventas
muelle
hora_entera
transacciones
semana
anio_entero
mes_entero
dia_entero
mes_string
marca
unidades

Nota: Es de aclarar que hay otras columnas que también se agregaron como ventas_sin_iva, precio_bruto_total, impuesto_total y descuento_total.

Con el objetivo ya claro, se analizaron los datos de cada marca y se realizaron los siguientes cambios sobre los mismos para poder realizar la unificación:

- Se quitaron todas las ventas que tenía el valor de 0 en todas las marcas, porque no aportaban ningún significado al análisis.
- Se unificaron en lo máximo posible todas las columnas de todas las marcas para que tuvieran el mismo nombre y una buena nomenclatura según las buenas prácticas de programación (sin tildes, minúsculas y sin espacios).
- Se eliminaron las columnas como destino, aerolínea de la marca 5, ya que en este análisis nos interesa saber las ventas de las tiendas más no conocer aún la incidencia de los vuelos como tal.
- Se eliminaron las columnas Departamento y Factura de la marca 10, ya que no aportaban ninguna información valiosa al análisis y era la única marca que tenía estas columnas, a excepción de la marca 4 que tenía num_factura, pero esta también se eliminó.

- Se realizaron agrupaciones (group by) por año, mes, día y hora conjuntamente de las marcas 3,4,5,6 y 10, para que todos los registros representarán la misma granularidad de ventas. Es decir, en el dataset final generado cada fila representa la venta en pesos colombianos de una tienda en un punto de venta, año, mes, día y hora entera especifica.
- Se calcularon nuevas columnas que ciertas marcas no tenían y que se podían generar a partir de la fecha, como: semana, día_entero, mes_entero, mes_string, hora_entera, anio_entero, etc.
- Sobre las columnas de categoría y/o descripción debido a que las marcas que tocaba transformar para unificar son la 3,4,5,6 y la 10, estas estaban agrupadas a las ventas por año, mes, día, hora y categoría. Por lo anterior, se tuvo que generar columnas dummies para cada categoría, ya que así era más fácil llevar el conteo sobre el tipo de productos y así no se perdiera esta información que es valiosa para este análisis.

Decisiones de modificación sobre la tabla unificada de datos:

- Se reemplazó con 0 todos los valores nulos de las columnas que se generaron dummies de las categorías.
- Se decidió solo tener las temporalidades del año del 2022 y la del año 2023, porque los datos hacia atrás son muy pocos y no van a ser tan significantes en los análisis.
- De momento, se decidió mantener las categorías definidas por las marcas, pero estas varían entre ellas, por lo cual son muchas categorías diferentes y esto puede hacer un poco engorroso el análisis de estas.

7 [10%] Primeras conclusiones, insights y acciones próximas a ser ejecutadas

- Se podría concertar y definir unas ciertas categorías más generalizadas de los productos para facilitar el análisis de los mismos datos.
- Definir con la empresa CAOBA si es posible conseguir los nombres de las categorías de la marca 10, ya que estas son muy difusas y solo están enumeradas, a la vez que la fila de Departamento de esta misma marca.
- Entender si los valores de ventas negativas son datos posibles dentro del contexto del negocio.
- En manera de aprendizaje creemos que pudo haber una mejor planeación entre el análisis de la calidad de datos con la parte del entendimiento, ya que se pudo planear un poco mejor la creación de gráficas y las conclusiones sacadas a partir de estas.
- Los años 2022 y 2023 son comparables hasta cierto punto, ya que del 2023 por cuestiones de temporalidad no tendremos todos los datos disponibles
- Las marcas que representan el mayor porcentaje de ventas (aprox. 76%) corresponde a las marcas 3, 8, 6 y 5.
- Los años con mayores ventas registradas son el 2022 y 2023.
- En cuanto a la evolución de las ventas en el año (semanas o meses) con los análisis actuales no es posible concluir que haya un mejor periodo que otro. Sin embargo, se tiene con la información de 2022, que las ventas son mejores en el segundo semestre. Igualmente, que las ventas de 2023 han sido mejores que las del 2022.
- A nivel de categorías de venta, la marca 4 con diferencia es la que más vende dentro del aeropuerto. Alcanza valores que sextuplican la categoría más vendida de la segunda marca con más ventas. Se resalta que no se tiene el nivel de granularidad necesario para incluir en el análisis a la marca 3.
- Las categorías con el mayor porcentaje de ventas son “alimentos y bebidas retail” y “golosinas” con alrededor de 600.000 unidades. Otras categorías con un número de ventas importantes son “tecnología”, “souvenirs”, “perfumería” y “vicios” con valores entre 30.000 y 40.000 unidades vendidas.