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.

Autor(es):

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

Fecha: Noviembre 6 de 2023

# PROYECTO FINAL – SEGUNDA ENTREGA

- **OBJETIVOS**
  - Finalizar la actividad de entendimiento de los datos aplicando diferentes técnicas pertinentes para la solución planteada.
  - Realizar la preparación y limpieza de datos requerida para la construcción del modelo de machine learning y/o dashboard planteado.
  - Construir una primera versión del producto de datos y realizar una primera evaluación de resultados.
- **ACTIVIDADES DEL SPRINT Y ENTREGABLES:**

## 1 [35%] PREPARACIÓN DE DATOS

### 1.1 Descripción del proceso

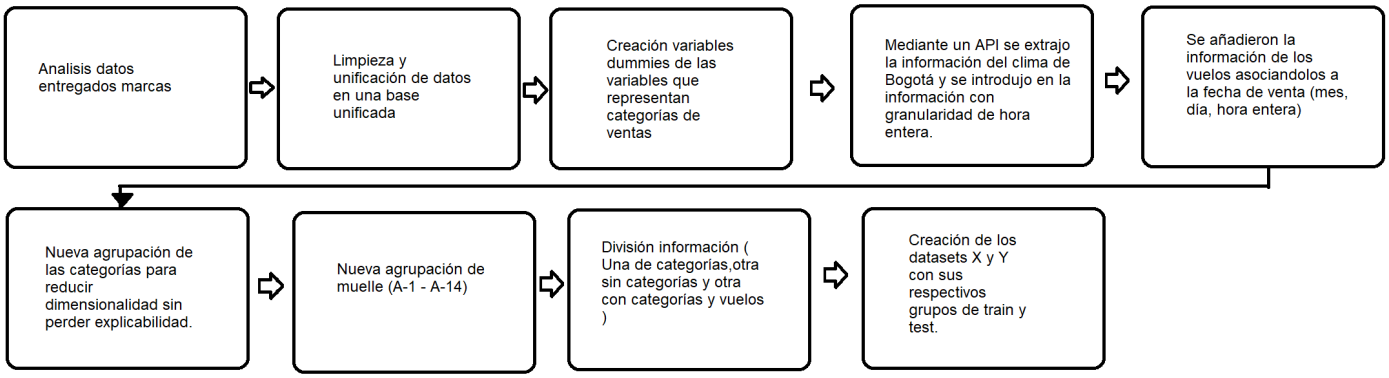



*Figura 1. Diagrama de bloques procesamiento de datos.*

**Análisis datos entregados de las marcas:** Se miró la estructura que tenían los datos de las 10 marcas que entregaron información y se identificó que había columnas que tenían nombres distintos pero sus datos significaban lo mismo por ejemplo Ubicación, muelle y lugar de venta. Se identificaron que había marcas que no tenían la misma granularidad y agrupación de sus ventas, es decir, había marcas que tenían registrada todas las ventas sin agrupaciones, otras que tenían las ventas agrupadas por minuto y la mayoría la tenían agrupada por hora entera. Además, había unas 4 marcas que tenían información de categorías solamente, las demás no tenían información al respecto. También, había marcas que tenían ciertas columnas que no tenían otras pero que se podían generar como, por ejemplo: hora_entera, mes_string, etc. Finalmente, al ver tanta disparidad de nombres de columnas se decidió cambiar todos los nombres de las columnas quitando las tildes, espacios y mayúsculas de las mismas.

**Limpieza y unificación de datos en una base unificada:** Se crea una base de datos con todos los datos de las 10 marcas procesados y organizados según lo dicho en el inciso de análisis de los datos entregados esto para tener un mayor control y una

facilidad de cara al procesamiento que se describirá en los siguientes pasos. Además, se eliminaron los datos con ventas negativas y escogimos una temporalidad de ventas de solo los años 2022 y 2023.

**Creaciones variables dummies:** Se identificó la necesidad de crear columnas binarias que representarán clases para las columnas de mes_string, categoría, tienda y marca. Esto con la finalidad de tener una explicabilidad de estas variables más clara cuando se cree la regresión respectiva.

**Extracción información clima por día y hora mediante API:** Mediante la API de la página: https://www.worldweatheronline.com/bogota-weather-history/cundinamarca/co.aspx se extrajo la información del clima de Bogotá sobre toda la línea de tiempo que nos interesa sobre los datos que es todo el año 2022 y los datos que tenemos del 2023. Este proceso se realizó con una granularidad igual que los datos de ventas (sobre hora_entera).

**Información de vuelos internacionales:** Teniendo en cuenta que los datos de ventas están sobre el muelle internacional se decidió relacionar esta información sobre los datos de vuelos internacionales que ocurrieron en la misma temporalidad de las ventas agrupándolos por categorías dummies en una granularidad mínima de hora_entera. Se tomaron en cuenta las variables del destino, el tipo de aerolínea, el día de la semana, y la aerolínea; variables que fueron convertidas a dummy para el procesamiento de los modelos.

**Nueva agrupación de categorías:** Se crearon nuevas categorías agrupando a las categorías mucho más específicos de la iteración anterior :

```
dataframe_unificado_final['viaje_ejecutivo_negocio'] =
dataframe_unificado_final['tabaco_alcohol'] = dataframe_
dataframe_unificado_final= dataframe_unificado_final.dr
dataframe_unificado_final['accesorios'] = dataframe_uni
dataframe_unificado_final['regalos_obsequios'] = datafr
dataframe_unificado_final['artesanias']= dataframe_unif
dataframe_unificado_final['calzado'] = dataframe_unific
dataframe_unificado_final['accesorios_de_mano']= datafr
dataframe_unificado_final['accesorios_belleza_cuidado']
dataframe_unificado_final['prendas_vestir'] =  datafram
dataframe_unificado_final['prendas_vestir_cabeza'] = da
dataframe_unificado_final['lujo'] = dataframe_unificado
dataframe_unificado_final['marroquineria'] = dataframe_
dataframe_unificado_final['cafe'] = dataframe_unificado
dataframe_unificado_final['tecnologia'] = dataframe_uni
dataframe_unificado_final['libros_revistas_papel']= dat
dataframe_unificado_final['cosas_hogar']= dataframe_uni
dataframe_unificado_final['comida_galguerias']= datafra
```

*Figura 2. Nuevas categorías.*

**Nueva agrupación de muelle:** Se crearon las siguientes agrupaciones de acuerdo al lugar espacial de las tiendas dentro del puente internacional con las siguientes denominaciones: A10-A12, A6-A8, A2-A4, A4, A5-A6 y A8-A10.

**División información:** Se divide la información entre dos datasets: Uno que contiene solo la información de ventas que tiene categorías y otro que tiene que solo la información de las ventas.

**Creación de los datasets X y Y:** Se divide la información entre dos datasets X y Y:

```
X=dataframe_unificado_final_solo_marca_categorias.drop(columns=['unidades','transacciones','ventas','muelle
','fecha_de_venta','mes_string','tienda','marca','descripcion','objeto1','objeto2','objeto3','objeto4','obj
eto5','tipo_de_tienda','ubicacion_esp','ventas_sin_iva','precio_bruto_total','impuesto_total','descuento_to
tal','tienda a','tienda b','personas_que_ingresan'])
```

```
y = dataframe_unificado_final_solo_marca_categorias['ventas']
```

Después se dejan los datos del año 2022 para los datasets del training y los del año 2023 para los datasets de prueba:

```
X_train = X[X['anio_entero']==2022]
X_test = X[X['anio_entero']==2023]
y_train = y[X['anio_entero']==2022]
y_test = y[X['anio_entero']==2023]
```

## 1.2 Antes y después de los datos

Datos unificados primera versión:

```
1  dataframe_unificado.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 246221 entries, 0 to 2032
Columns: 244 entries, fecha_de_venta to categoria 99
dtypes: datetime64[ns](1), float64(232), object(11)
memory usage: 460.2+ MB
```

*Figura 3. Primera versión datos unificados.*

Se tienen un total de 246221 filas con un total de 244 columnas. Con muchos problemas de valores nulos entre las columnas.

Datos unificados última versión:

```
1  dataframe_unificado_final.shape
```

```
(240436, 196)
```

*Figura 4. Última versión de datos unificados.*

Finalmente, se tienen un total de 240436 filas con 196 columnas.

# 2 [10%] ESTRATEGIA DE VALIDACIÓN Y SELECCIÓN DE MODELO

Para este apartado nos apoyamos en la librería de lazypredict que permite probar 42 modelos de regresión para ver cuál o cuáles son los que más se ajustan a nuestros datos e hicimos dos pruebas uno con el conjunto de datos que tiene la información de las categorías y otro con el que no.

```python
# Lista de modelos a excluir
modelos_a_excluir = ['SVR', 'NuSVR', 'GaussianProcessRegressor']
reg = LazyRegressor(verbose=0, ignore_warnings=False, custom_metric=None)
models, predictions = reg.fit(X_train, X_test, y_train, y_test)

print(models)
```

*Figura 5. Uso de la librería lazypredict y la clase LazyRegressor.*

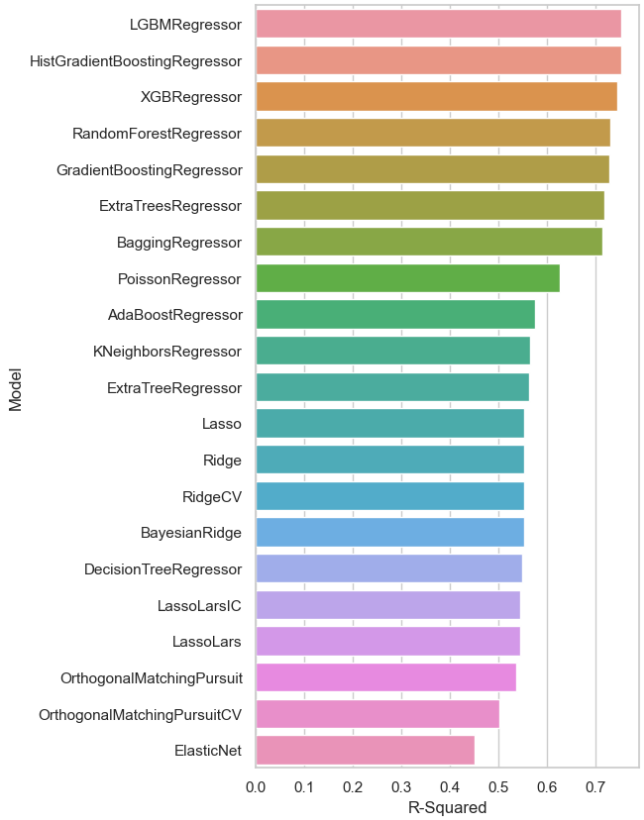



*Figura 6. Puntaje R2 modelos de regresión. (Con categorías).*

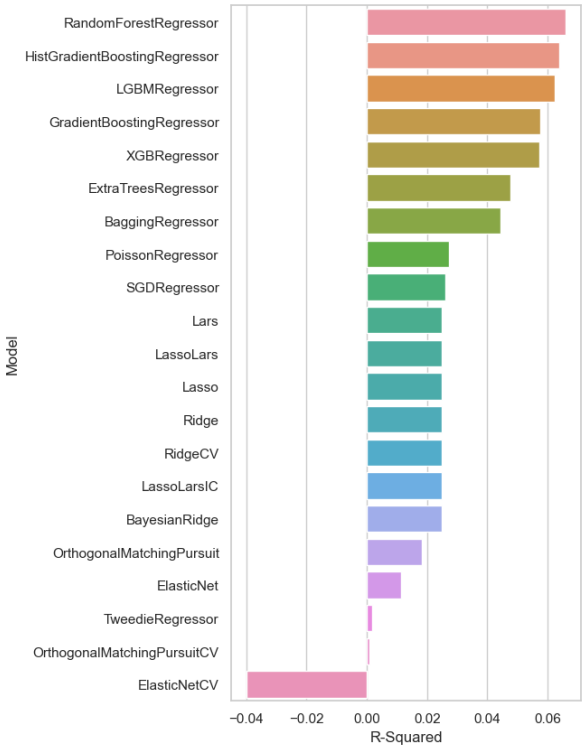



*Figura 7. Puntaje R2 de los modelos de regresión. (Sin categorías).*

De acuerdo a las figuras 6 y 7, se escogen los mejores 3 modelos para cada sub-dataset para desarrollar el siguiente punto de construcción de modelo.

# 3 [20%] CONSTRUCCIÓN DEL MODELO

Para cada modelo, se han ajustado los hiperparámetros para maximizar la métrica R ^ 2 a través de una validación cruzada con 5 folds para los modelos (excepto en Random Forest Regressor para reducir la complejidad).

## 3.1 Modelos con categorías:

- **LightGBM Regressor:**
  - **Hiperparámetros del Grid Search:**

```
param_grid = {
    'num_leaves': [31, 50, 70],
    'reg_alpha': [0.1, 0.5],
    'min_data_in_leaf': [30, 50, 100],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [-1, 5, 10]
}

lgbm = lgb.LGBMRegressor()

grid_search = GridSearchCV(estimator=lgbm, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)
```

- **Mejores hiperparámetros encontrados:**
  - learning_rate: 0.05
  - max_depth: -1 (sin límite)
  - min_data_in_leaf: 100
  - num_leaves: 70
  - reg_alpha: 0.5

- **Histogram-based Gradient Boosting Regressor**

  - **Hiperparámetros del Grid Search:**

```
param_grid = {
    'max_iter': [100, 200],
    'max_depth': [3, 5, 10],
    'min_samples_leaf': [20, 40],
    'learning_rate': [0.01, 0.1, 0.2],
    'l2_regularization': [0.0, 0.1]
}

hist_gbr = HistGradientBoostingRegressor()

grid_search = GridSearchCV(estimator=hist_gbr, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)
```

  - **Mejores hiperparámetros encontrados:**
    - learning_rate: 0.05
    - max_depth: -1 (sin límite)
    - min_data_in_leaf: 100
    - num_leaves: 70
    - reg_alpha: 0.5

- **XGBoost Regressor**

  - **Hiperparámetros del Grid Search:**

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 6, 9],
    'colsample_bytree': [0.5, 0.7],
    'subsample': [0.6, 0.8, 1.0],
    'gamma': [0, 0.1, 0.2]
}

# Inicializar el modelo XGBRegressor
xgb_regressor = XGBRegressor(objective='reg:squarederror')

# Configurar GridSearchCV
grid_search = GridSearchCV(estimator=xgb_regressor, param_grid=param_grid,
                           scoring='r2', cv=5, verbose=1, n_jobs=-1)
```

- **Mejores hiperparámetros encontrados:**
  - **colsample_bytree: 0.7**
  - **gamma: 0**
  - **learning_rate: 0.01**
  - **max_depth: 9**
  - **n_estimators: 100**
  - **subsample: 0.8**

## 3.2 Modelos sin categorías:

- **RandomForest Regressor**

  - **Hiperparámetros del Grid Search:**

```
param_grid = {
    'n_estimators': [100, 200],  # Menos árboles
    'max_depth': [10, 20, None],  # Profundidad limitada y una opción sin límite
    'min_samples_split': [2, 5],  # Menos opciones para la cantidad mínima de muestras para dividir
    'min_samples_leaf': [1, 2],   # Menos opciones para la cantidad mínima de muestras en una hoja
    'max_features': ['auto', 'sqrt'],  # Opciones para el número de características a considerar al buscar la mejor división
}

# Inicializar el modelo RandomForestRegressor
rf_regressor = RandomForestRegressor(random_state=0)

# Configurar GridSearchCV
grid_search = GridSearchCV(estimator=rf_regressor, param_grid=param_grid,
                           scoring='r2', cv=3, verbose=1, n_jobs=-1)
```

  - **Mejores hiperparámetros encontrados:**
    - max_depth: 5
    - min_samples_leaf: 1
    - n_estimators: 200
    - min_samples_split: 2
    - max_features: 'auto'

- **Histogram-based Gradient Boosting Regressor**

  - **Hiperparámetros del Grid Search:**

```
# Definir una cuadrícula de parámetros más pequeña y menos exhaustiva
param_grid = {
    'max_iter': [100, 200],
    'max_depth': [3, 5, 10],
    'min_samples_leaf': [20, 40],
    'learning_rate': [0.01, 0.1, 0.2],
    'l2_regularization': [0.0, 0.1]
}

hist_gbr = HistGradientBoostingRegressor()

grid_search = GridSearchCV(estimator=hist_gbr, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)
```

- **Mejores hiperparámetros encontrados:**
  - l2_regularization: 0.1
  - learning_rate: 0.01
  - max_depth: 3
  - max_iter: 100
  - min_samples_leaf: 40

- **LightGBM Regressor**

  - **Hiperparámetros del Grid Search:**

```
param_grid = {
    'num_leaves': [31, 50, 70],
    'reg_alpha': [0.1, 0.5],
    'min_data_in_leaf': [30, 50, 100],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [-1, 5, 10]
}

lgbm = lgb.LGBMRegressor()

grid_search = GridSearchCV(estimator=lgbm, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)
```

  - **Mejores hiperparámetros encontrados:**
    - learning_rate: 0.01
    - max_depth: 5
    - min_data_in_leaf: 30
    - num_leaves: 50
    - reg_alpha: 0.5

## 3.3 Modelo categoría con información de vuelos

```
regression = LinearRegression()
regression.fit(X_train, y_train)
```

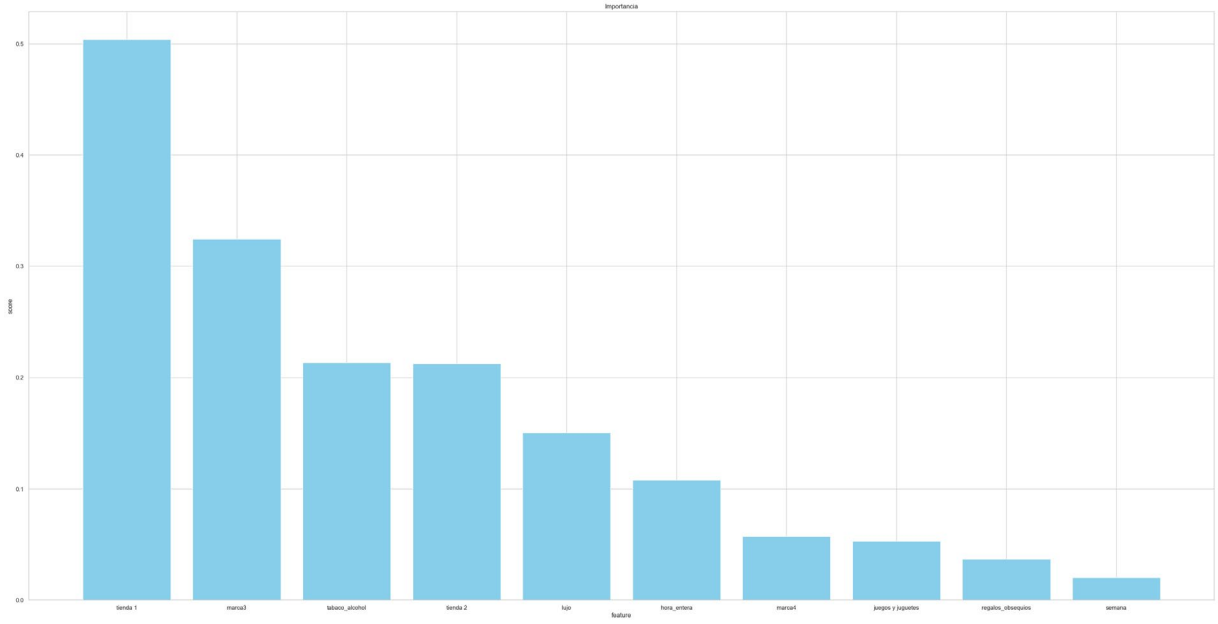
# 4 [20%] EVALUACIÓN DEL MODELO

## 4.1 Modelos con categorias

| Modelo | MAE Train | MAE Test | RMSE Train | RMSE Test | R2 Train | R2 Test |
|---|---|---|---|---|---|---|
| LGBM | 7,862,636.43 | 11,218,845.19 | 15,646,572.29 | 20,700,288.21 | 0.84 | 0.75 |
| HISTGBR | 8,111,847.00 | 11,220,512.78 | 16,018,202.92 | 20,371,537.14 | 0.83 | **0.76** |

| | | | | | | |
|---|---|---|---|---|---|---|
| XGBR | 10,261,177.53 | 15,207,824.47 | 23,283,554.16 | 29,066,946.27 | 0.65 | 0.51 |

Importancia de los atributos del mejor modelo:




## 4.2 Modelos sin categorias

| Modelo | MAE Train | MAE Test | RMSE Train | RMSE Test | R2 Train | R2 Test |
|---|---|---|---|---|---|---|
| RFRegressor | 14,924,398.46 | 16,056,810.96 | 28,082,341.46 | 31,236,064.37 | 0.093 | **0.041** |
| HISTGBR | 15,497,472.35 | 16,409,970.83 | 28,641,938.71 | 31,416,677.50 | 0.057 | 0.030 |
| LGBM | 15,330,679.43 | 16,202,557.48 | 28,468,636.89 | 31,341,046.67 | 0.068 | 0.034 |

## 4.3 Modelo con categorías e información de vuelos

| Modelo | MAE Train | MAE Test | RMSE Train | RMSE Test | R2 Train | R2 Test |
|---|---|---|---|---|---|---|
| Regresión lineal | 17227486.6 | 16,056,810.96 | 28,082,341.46 | 31,236,064.37 | 0.093 | **0.041** |

## 4.4 Evaluación de los resultados

En los modelos con categorías:

- El Histogram-based Gradient Boosting Regressor (HISTGBR) obtiene el mejor rendimiento en el conjunto de prueba con un R2 de 0.76, lo cual indica que el modelo puede explicar aproximadamente el 75.7% de la varianza en los datos de prueba. Aunque tiene un error absoluto medio (MAE) y un error cuadrático medio (RMSE) ligeramente más altos en la fase de prueba en comparación con el LGBM, su mayor R2 en la fase de prueba sugiere una mejor generalización. De la misma manera sus atributos más importantes son: tienda_1, marca3, tabaco_alcohol, tienda_2 y lujo.
- El LightGBM (LGBM) tiene un buen rendimiento en el conjunto de entrenamiento, pero muestra una caída en la puntuación R2 cuando se evalúa en el conjunto de prueba. Esto puede ser indicativo de un ligero sobreajuste al conjunto de entrenamiento.
- El XGBoost Regressor (XGBR) tiene la puntuación más baja de R2 en el conjunto de prueba, lo que sugiere que tiene un poder predictivo inferior en comparación con los otros modelos con categorías.

En los modelos sin categorías:

- Todos los modelos tienen un rendimiento muy bajo, con puntuaciones R2 que indican que los modelos no están capturando bien la variabilidad de los datos. El Random Forest Regressor tiene el R2 más alto en el conjunto de prueba de los tres, pero sigue siendo muy bajo (0.041), lo que indica que prácticamente no tiene capacidad predictiva.

# 5 [15%] CONCLUSIONES

## 5.1 Dificultades Encaradas:

Durante el desarrollo del proyecto, nos enfrentamos a varias dificultades, entre las cuales las más destacadas fueron:

- Overfitting en Modelos con Categorías: Se observó que algunos modelos, como el LightGBM, mostraban un alto rendimiento en los datos de entrenamiento, pero una disminución significativa en los datos de prueba, lo que sugiere una especialización excesiva en los datos de entrenamiento y una generalización insuficiente.
- Bajo Rendimiento en Modelos sin Categorías: Los modelos sin categorías mostraron una capacidad predictiva extremadamente baja, indicada por puntuaciones R2 cercanas a cero, lo que implica que no fueron capaces de capturar la variabilidad de los datos y realizar predicciones fiables.

## 5.2 Estrategias de Mitigación:

Para abordar estas dificultades, proponemos las siguientes estrategias:

- Regularización y Ajuste de Hiperparámetros: Implementaremos técnicas de regularización y optimización de hiperparámetros para mejorar la generalización de los modelos.
- Análisis de Sesgo: Realizaremos un análisis para detectar y mitigar cualquier sesgo presente en los datos que pueda estar afectando el rendimiento del modelo.

## 5.3 Condiciones de los Datos para Mejorar Resultados:

Se considera que los datos podrían mejorar en los siguientes aspectos:

- Reducción de Sesgo: Una selección y procesamiento de datos que busque activamente reducir sesgos puede resultar en modelos más justos y precisos.

## 5.4 Evaluación del Mejor Modelo:

El modelo más prometedor identificado hasta la fecha es el Histogram-based Gradient Boosting Regressor (HISTGBR) con categorías, el cual ha mostrado una capacidad considerable para explicar la varianza en los datos de prueba (R2 de 0.757).

Autor(es):

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

Fecha: Diciembre 3 de 2023

# PROYECTO FINAL – TERCERA ENTREGA

- **OBJETIVOS**
  - Finalizar las etapas de modelado y evaluación, teniendo en cuenta la retroalimentación brindada durante la sustentación de la segunda entrega.
  - Construir el producto de datos con los diferentes componentes establecidos durante la actividad de ideación.
  - Presentar los resultados del análisis y el producto de datos a los stakeholders de la organización y obtener retroalimentación respecto a los aspectos positivos logrados y elementos a mejorar.
- **ACTIVIDADES DEL SPRINT Y ENTREGABLES:**

## 1 [35%] CONSTRUCCIÓN DEL PRODUCTO DE DATOS:

### 1.1 Generación datos finales para la creación de modelo:

```
df_final.shape

(114279, 302)
```

*Figura 1. Tamaño DataFrame final información unificada.*

Este es el tamaño final de nuestro dataframe unificado el cual contiene la información del clima, vuelos, marcas, tiendas, categorías, ventas, etc.

(El proceso más al detalle se puede encontrar en el notebook Proyecto_OPAINV_Alistamiento_Datos_Entrega3.ipynb en la carpeta de Notebooks entrega final de nuestro repositorio)

## 1.2 Se escoge el mejor modelo con los mejores resultados obtenidos:

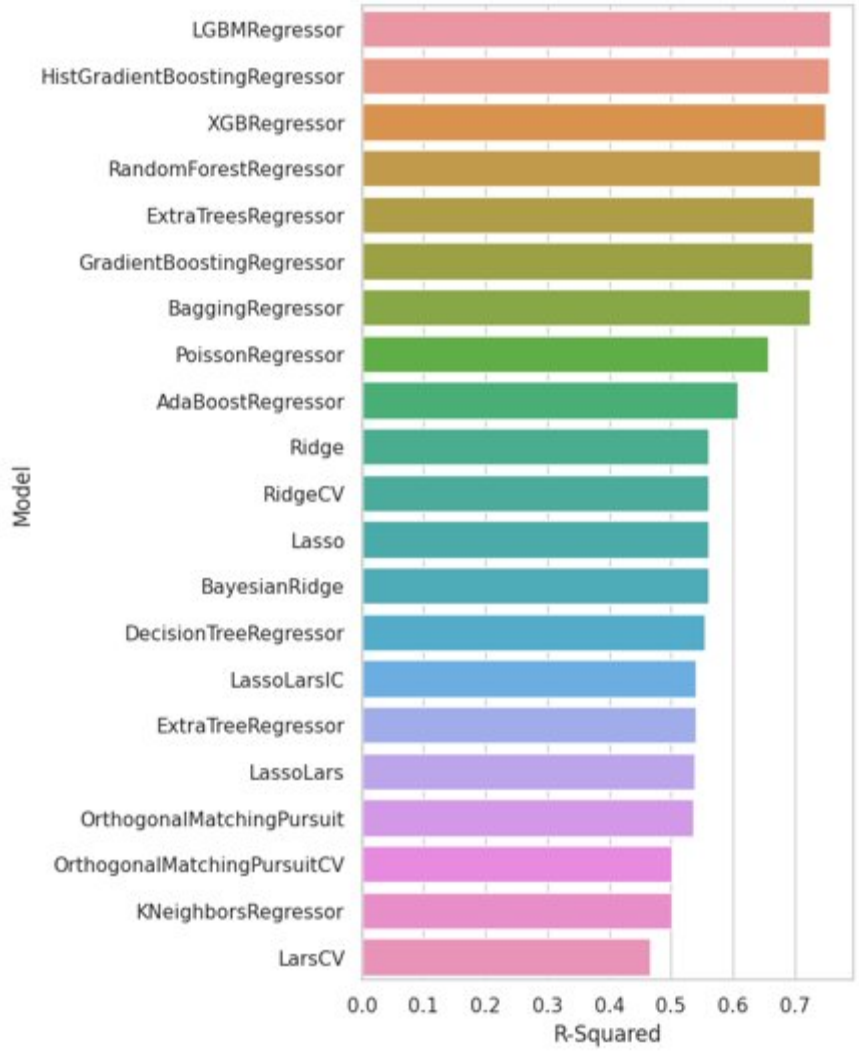



*Figura 2. Pruebas diferentes modelos de regresión.*

A partir de estos resultados se escoge el modelo LGBMRegressor para realizarle la respectiva búsqueda de hiperparámetros mostrados en la Figura 3.

```
param_grid = {
    'num_leaves': [31, 50, 70],
    'reg_alpha': [0.1, 0.5],
    'min_data_in_leaf': [30, 50, 100],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [-1, 5, 10]
}

lgbm = lgb.LGBMRegressor()

grid_search = GridSearchCV(estimator=lgbm, param_grid=param_grid, cv=5, scoring='r2', verbose=1, n_jobs=-1)

grid_search.fit(X_train, y_train)
```

*Figura 3. Búsqueda de hiperparámetros.*

Finalmente, mediante tres subcojuntos de datos se obtuvieron los resultados mostrados en la figura 4.

| Modelo | MAE Train | MAE Test | RMSE Train | RMSE Test | R2 Train | R2 Test |
|---|---|---|---|---|---|---|
| **Todas las variables** | **7808426.93** | **11216316.69** | **15335847.97** | **20567723.23** | **0.84** | **0.75** |
| Sin variables de tienda, categorías no especificadas ni marca | 8896343.74 | 12128998.44 | 17769446.63 | 22354600.02 | 0.79 | 0.7 |
| Sin variables de tienda, categorías no especificadas, marcas y aerolíneas | 8907448.70 | 12131249.39 | 17786849.75 | 22361158.30 | 0.79 | 0.7 |

*Figura 4. Métricas resultados pruebas.*

Por lo anterior, se decidió escoger el modelo con todas las variables posibles, ya que tiene el mejor comportamiento sobre los datos de prueba.

## 1.3 Generación de predicciones sobre los datos de prueba:

En este apartado, describimos que a partir del archivo .pkl que se llama *modelo_completo_Entrega3.joblib* el cuál se encuentra en la carpeta llamada *Modelos última entrega* se generaron las predicciones de las ventas del año 2023 y se contrastaron estas predicciones con las ventas reales mediante un tablero de control.

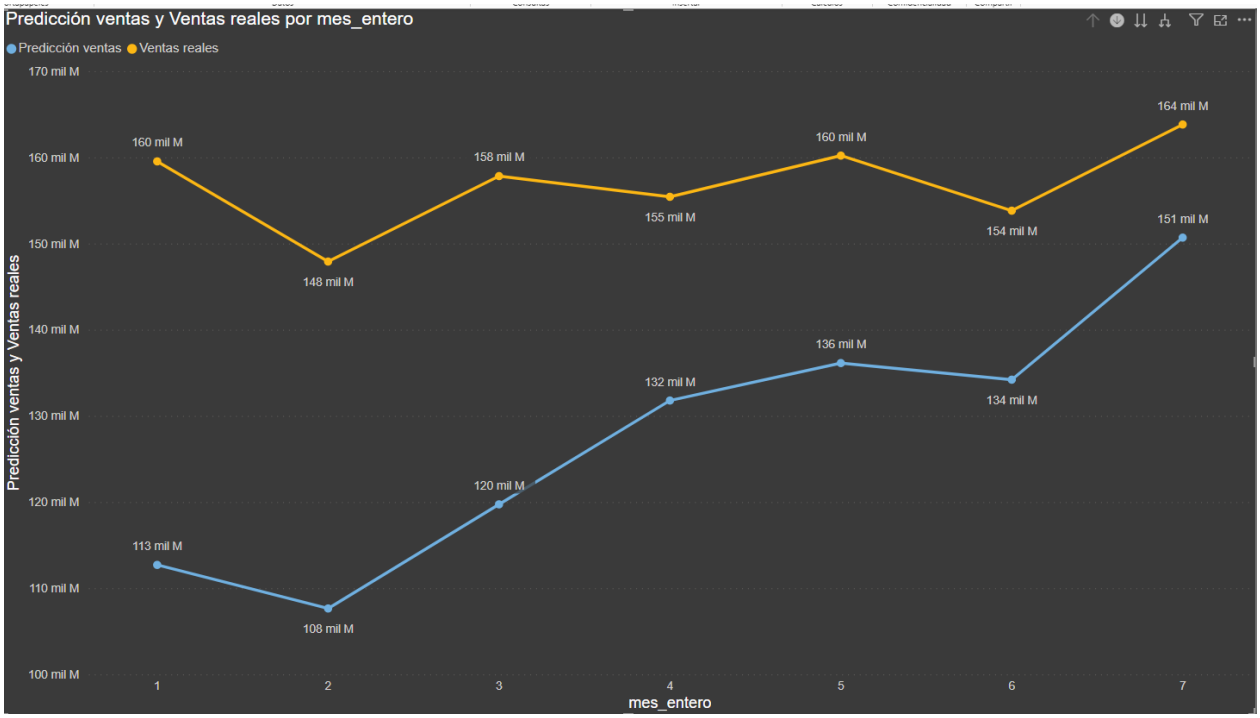



*Figura 5. Tablero de control de realidad vs predicción. (Por mes)*

Además, este gráfico permite ver la granularidad más baja que sería la de horas como lo muestra la figura 6.

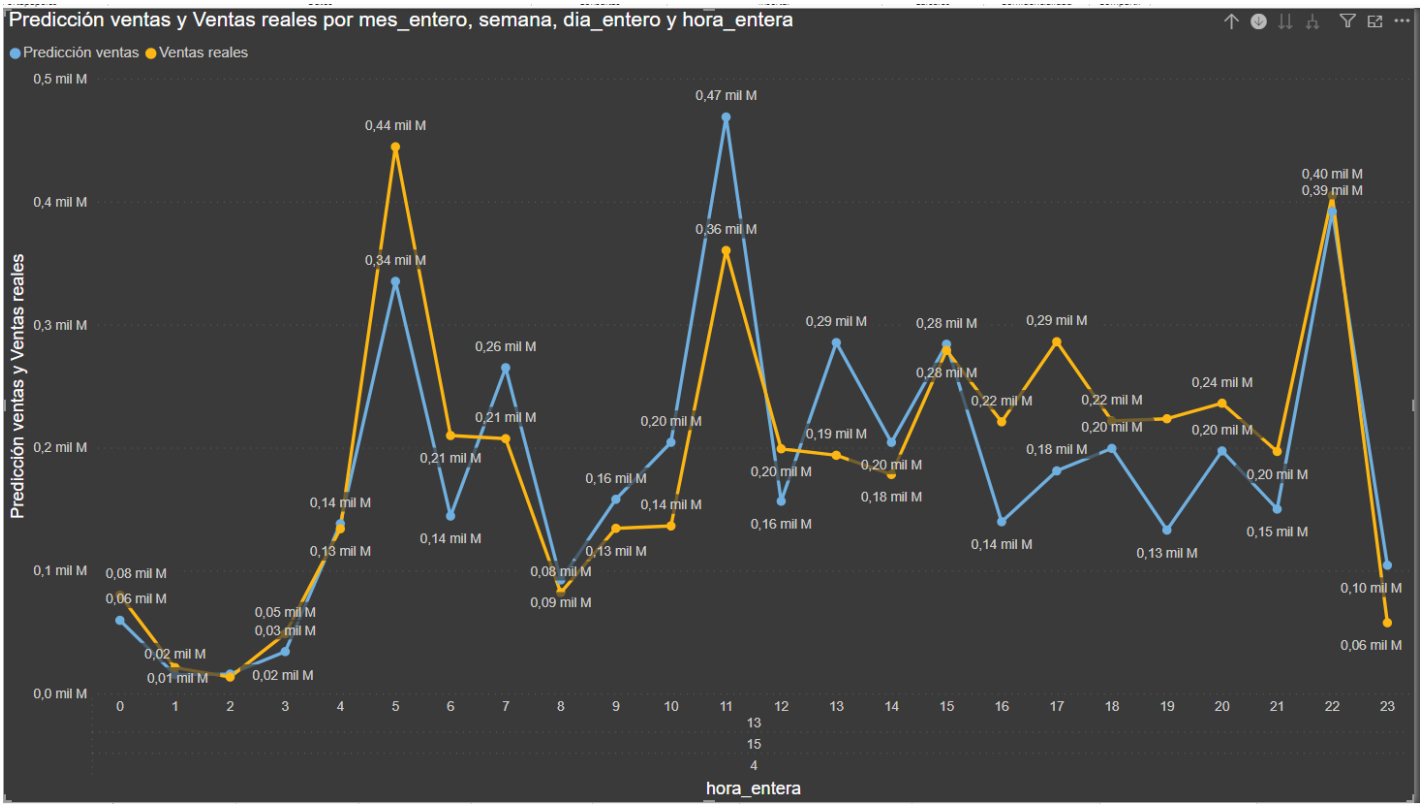



*Figura 6. Tablero de control realidad vs predicción. (Por hora entera).*

## 1.4 Creación de tableros de control:

Se diseñó un tablero de control unificado en Power BI como un producto de datos, que permite visualizar los comportamientos de las ventas en diversos aspectos, como las unidades vendidas de ciertas categorías y las ventas promedio según la agrupación de temperaturas, entre otros.

**Link al archivo compartido en la nube:**

https://app.powerbi.com/reportEmbed?reportId=ac2e0b60-2eab-4269-81ee-6416d55ecb99&autoAuth=true&ctid=fabd047c-ff48-492a-8bbb-8f98b9fb9cca

**Archivo original del tablero de control (.pbix):**

https://drive.google.com/file/d/1kiPRSB6vfhazjDihq1ZgPlCevhNnexV1/view?usp=sharing

**Nota:** Es de aclarar que para poder ver el informe en línea se tiene que pedir permiso primero, ya que, las políticas de la DSIT (Departamento de la seguridad y la información) no permite generar un link para que cualquier usuario lo pueda ver, si no que toca compartirle directamente a cada usuario el power BI. Por otro lado, se dejó enlazado también el archivo .pbix para que se pueda modificar, se pueda utilizar para mayor facilidad si es el caso o se pueda subir al power bi de su respectiva organización u espacio de trabajo propio.

# 2 [20%] RETROALIMENTACIÓN POR PARTE DE LA ORGANIZACIÓN:

## 2.1 Cronograma establecido por el cliente:

El cliente en este caso especifico OPAIN y CAOBA, nos entregaron y estuvimos de acuerdo con el siguiente cronograma con fecha de actualización al 15 de octubre del año 2023:

A continuación, se presentan las actividades que se realizaron hasta la fecha indicada:

| Fecha | Actividad |
|---|---|
| 11/08/2023 | Primera reunión Caoba, profesor de la universidad los Andes y Opain para revisar alcance del proyecto |
| 25/08/2023 | Primer acercamiento entre Opain y grupo 1 de los estudiantes de la universidad de los Andes |
| 29/08/2023 | Formalización de aceptación por parte de Caoba |
| 01/09/2023 | Primer acercamiento entre Opain y grupo 2 de los estudiantes de la universidad de los Andes |
| 17/09/2023 | Los estudiantes hicieron la primera entrega que tenían en el cronograma del curso al profesor |
| 29/09/2023 | Presentación a Opain de la primera entrega |
| 06/10/2023 | Opain hace entrega a estudiantes del cronograma propuesto por la Dirección Proyectos Aeroportuarios e Innovación Tecnológica |

***Programa de actividades***

Se presentan las entregas que deben realizar los estudiantes a Opain de acuerdo con las entregas que deben hacer a la universidad de los Andes:

| Fecha de Entrega | Meta | Entregable |
|---|---|---|
| 20/10/2023 | Comprensión de datos y elección de modelo. | • Verificar significancia del modelo seleccionado.<br>• Presentación de hallazgos o tendencias encontradas en el análisis exploratorio de datos.<br>• Código abierto comentado en formato ipynb.<br>• Sustentar que modelo se usó y ¿por qué? |

| | | |
|---|---|---|
| 24/11/2023 | Prototipo inicial del proyecto | • Construcción del prototipo inicial donde se detallen los procesos de desarrollo (así sean de overview o experimentales) y tengan un grado de significancia optimo, así como una relevancia constante.<br>• Presentación de hallazgos y análisis encontrados del desarrollo del prototipo inicial.<br>• Código abierto comentado en formato ipynb. |
| 15/12/2023 | Modelo finalizado | • El modelo debe ser capaz de sugerir tácticas para aumentar las ventas de los establecimientos ubicados en el Aeropuerto Internacional El Dorado, basándose en información histórica comercial y de operaciones.<br>• Presentación final a alto nivel donde se detalle el funcionamiento del modelo.<br>• Código abierto con comentado en formato ipynb. |

## 2.2 Observaciones generadas a partir de cada entrega:

A continuación, se presenta un resumen de los comentarios proporcionados por el cliente durante las reuniones realizadas casi semanalmente desde el inicio del semestre. Es importante señalar que no contamos con observaciones específicas escritas por ellos, ya que estas fueron recomendaciones expresadas durante las reuniones, las cuales también han sido grabadas en nuestro espacio en Microsoft Teams.

| Entrega | Observaciones hechas por el cliente: |
|---|---|
| Entrega 1 Universidad | En esta primera entrega, se presentaron al cliente las primeras visualizaciones de las construcciones de varios modelos, incluyendo un total de 43 modelos de regresiones. El objetivo era determinar cuál de ellos se ajustaba mejor a los datos. Se proporcionó una visión general de todos los datos disponibles, abarcando todas las marcas disponibles. A raíz de esta presentación, el cliente nos aconsejó evitar posibles problemas derivados de la creación de tantas columnas dummy en los modelos. Además, se enfatizó la falta de información con respecto a una marca que denomina sus categorías como "categoría 1", "categoría 2", etc. |

| | |
|---|---|
| Entrega 2 Universidad | En esta segunda entrega, se presentó al cliente un modelo de regresión seleccionado a partir de la búsqueda de hiperparámetros realizada en tres modelos distintos. El HistGradientBoostingRegressor fue identificado como el mejor modelo, demostrando un rendimiento equitativo tanto en los datos de entrenamiento como en los de prueba. Además, se llevaron a cabo dos pruebas: una considerando solo las categorías con información de las marcas que tenían información de las categorías y otra exclusivamente con información de ventas.<br><br>Es importante destacar que en esta entrega se incorporó información climática con la misma granularidad horaria que las ventas. A raíz de esta adición, el cliente expresó su agrado, ya que la inclusión de datos climáticos proporciona mayor explicabilidad al modelo. Esto resulta especialmente relevante para obtener conclusiones sobre las ventas, dado que la principal importancia de este modelo no radica tanto en su capacidad predictiva, sino en su habilidad para explicar de manera efectiva las ventas ocurridas en el año 2022. |
| Entrega 3 Universidad | En esta entrega final, se presentó al cliente la primera versión del tablero de control, que permite analizar diversos comportamientos de las ventas basándonos en los datos procesados y en las predicciones generadas por el modelo. Es importante señalar que en esta última entrega se incorporó un modelo con nuevas variables relacionadas con vuelos, con el objetivo de mejorar la explicabilidad del modelo.<br><br>A partir de esta adición, al cliente le complació que fuéramos más allá de simplemente proporcionar un modelo. Se destacó la importancia de presentar de manera gráfica la información, permitiendo tanto a ellos como a nosotros extraer conclusiones sobre las ventas específicas de cada marca o tienda. |

## 3 [35%] PRESENTACIÓN FINAL:

Realizamos la presentación con unas ciertas mejoras respecto a lo mostrado en la sustentación final en el siguiente link:
https://drive.google.com/file/d/1hK-8R1WPm76uhIev_Lz2Z5wJARRGkpoa/view?usp=sharing

**Nota:** Dejamos los archivos subidos en Google drive para evitar restricciones por temas de organización que impone la universidad.

## 4 [15%] CONCLUSIONES Y TRABAJO FUTURO:

- Las predicciones generadas por el modelo desarrollado resultan ser de gran utilidad para Opain, ya que permiten optimizar la cantidad de ventas al basarse en las tendencias de compra de los viajeros en el Aeropuerto El Dorado.
- El modelo presentado puede desempeñar un papel crucial como sistema de toma de decisiones para Opain, proporcionando seguridad al momento de reubicar tiendas, ajustar productos y/o establecer nuevas alianzas con compañías.
- Nuestro modelo final tiene la capacidad de informar sobre los factores que inciden en el comportamiento de las ventas. Permite indagar sobre qué variables y en qué medida influyen en las predicciones respectivas.
- El producto de datos entregado superó las expectativas del cliente, quienes esperaban únicamente un modelo funcional en términos de predicciones.
- Aunque los años 2022 y 2023 son comparables hasta cierto punto, debido a limitaciones temporales, no tendremos todos los datos disponibles para el año 2023.
- Las marcas que representan aproximadamente el 76% de las ventas corresponden a las marcas 3 y 6, siendo las tiendas Duty Free las que generan las mayores ventas.
- Las categorías con mayor influencia en las ventas incluyen tabaco y alcohol, juegos y juguetes, productos de lujo, regalos y obsequios, así como accesorios de mano, belleza y cuidado.
- Las zonas/ubicaciones en el muelle internacional que más influyen positivamente son A6-A8 y A10-A12, mientras que la zona A2-A4 muestra una relación inversa con las ventas.
- Los vuelos en horas tempranas, especialmente entre las 0-3 de la mañana, presentan una relación negativa con las ventas. Además, se observan picos de venta entre las 5-8, 10-15 y 19-22 horas.
- Los días de la semana con mayores ventas son el domingo, lunes y sábado.
- De cara al futuro, sería interesante incorporar aún más variables externas al propio aeropuerto. Al explicar el comportamiento de los pasajeros dentro del aeropuerto con estos datos adicionales, sería crucial para generar conclusiones más acertadas y desarrollar modelos de predicción de ventas más efectivos.
- La falta de información en relación con las categorías no clasificadas generó un problema en la construcción del modelo. Aunque es posible explicarlas, carecen de contexto, a diferencia de las marcas que proporcionaron descripciones detalladas de sus productos.
- Los datos se limitaron a las ventas en el muelle internacional. Sería interesante realizar también un análisis de estas ventas en vuelos nacionales, dado que se comprende que estos constituyen la mayor parte de los vuelos que atraviesan el aeropuerto.