



BIG DATA, INTELIGENCIA ARTIFICIAL & MACHINE LEARNING

FULL STACK BOOTCAMP XI

PROYECTO FINAL DE BOOTCAMP

“Predicción del nivel de aceptación de un producto de una E-commerce (Flipkart) con implementación de Machine Learning”

AUTORES:

Dayana Franco, francosalcedod@gmail.com
Wilbert Vong, ww.pbsxerox@gmail.com
Marcos Salafranca, trope100@hotmail.com
Jefferson Erick Osorio, jefferson.osorio@pucp.pe

(Grupo: DataFriends)

Madrid, 26 de Octubre de 2023

Resumen

En el mundo del comercio electrónico, la rápida evolución del mercado y la diversidad de productos disponibles hacen que las decisiones relacionadas con el inventario, marketing y precios sean complejas. Las empresas invierten significativamente en la introducción de nuevos productos, pero no todos ellos son bien recibidos por los consumidores. Una elección errónea o un mal cálculo en la demanda puede resultar en pérdidas económicas, exceso de inventario y daño a la reputación de la marca.

Dentro de este escenario, Flipkart, como una prominente tienda e-commerce, enfrenta el desafío de anticipar cómo un producto será aceptado por su base de clientes. Tradicionalmente, las decisiones se han basado en tendencias pasadas y experiencia del equipo de marketing, pero estos métodos no siempre garantizan precisión. En el contexto global, donde la competencia es feroz y los márgenes de error son estrechos, es crucial tener herramientas más precisas y basadas en datos para prever la aceptación de un producto.

La falta de una herramienta predictiva puede llevar a una sobreinversión, pérdida de oportunidades, inadecuadas estrategias y desajuste en los inventarios.

Por lo tanto, en el presente trabajo, mediante el uso de machine learning y análisis de datos de compras de la base de datos de sus clientes se desarrolla un modelo que va a permitir medir el grado de aceptación de un producto dentro su catálogo permitiendo de esta manera que Flipkart pueda mejorar sus estrategias de marketing.

Tabla de contenido

1 Parte Introductoria.....	5
1.1Interés del estudio.....	5
2Parte General.....	5
2.1Fines y Objetivos.....	5
3Metodología y estado de la cuestión.....	6
3.1Marco tecnológico.....	6
3.1.1Herramientas a utilizar.....	6
3.1.2Hardware utilizado.....	7
3.2Metodología.....	7
3.3Fases Para Implementar.....	8
3.3.1Comprensión del Negocio.....	8
3.3.2Comprensión de los Datos.....	8
3.3.3Preparación de los datos.....	8
3.3.4Modelado.....	9
3.3.5Evaluación.....	9
3.4Estado del Arte.....	9
3.4.1Machine Learning.....	9
3.5Algoritmos utilizados.....	11
3.5.1Logistic Regression. Regresión Logística para Clasificación.....	11
3.5.2Árboles de decisión y Random Forest.....	12
3.5.3XGboost para clasificaciones (XGBClassifier).....	13
4Propuesta de Arquitectura.....	14
4.1Origen de Datos.....	14
4.2Almacenamiento de los Datos.....	15

4.3Limpieza, transformación y enriquecimiento de los Datos.....	15
4.4Persistencia:	15
4.5Modelado de Machine Learning.....	15
4.6Visualización de los Datos.....	15
Explorar un conjunto de datos para obtener respuestas a cuestiones previamente planteadas y a la vez elaborar nuevas preguntas.....	15
Analizar patrones, relaciones y valores atípicos existentes en el conjunto de los datos para identificar circunstancias no detectadas y extraer conclusiones.....	15
5ANÁLISIS DE RESULTADOS DE ML.....	16
5.1PRICING ML.....	16
5.2Classification ML:.....	16
6Desarrollo del departamento de big data.....	17
6.1Medios técnicos.....	17
6.2Medios humanos.....	18
7Visualización de los resultados.....	19
8Conclusiones.....	21

1 Parte Introductoria

1.1 Interés del estudio

Flipkart es uno de los gigantes del comercio electrónico en India, ofreciendo una amplia gama de productos desde electrónicos hasta moda y muebles. Fundada en 2007, la empresa ha experimentado un crecimiento explosivo, convirtiéndose en un referente en el mundo del e-commerce en Asia. Sin embargo, con el aumento de la competencia y la constante evolución del mercado, Flipkart enfrenta el desafío de mantenerse relevante y garantizar que su oferta de productos se alinee con las expectativas de los consumidores.

La principal problemática de Flipkart radica en predecir la aceptación de los nuevos productos que introduce en su plataforma. La elección incorrecta de un producto, basada en suposiciones y no en datos concretos, puede resultar en pérdidas económicas, acumulación de inventario no deseado y un posible daño a su imagen de marca. En este contexto, es esencial para Flipkart desarrollar herramientas y estrategias que le permitan anticipar las preferencias de sus clientes y adaptarse rápidamente a las demandas del mercado. La utilización de tecnologías avanzadas, como el machine learning, emerge como una solución potencial para abordar esta problemática.

2 Parte General

2.1 Fines y Objetivos

El objetivo se centra en emplear machine learning para analizar los datos de compras históricas de Flipkart, una tienda e-commerce, con el fin de predecir el grado de aceptación de un producto. Este indicador se basa en variables como ventas, opiniones de los clientes y devoluciones. La importancia de este estudio radica en su capacidad para ofrecer a las tiendas e-commerce una visión anticipada sobre el desempeño de un producto en el mercado. En un contexto global, donde el e-commerce juega un papel cada vez más dominante en el comercio, poder anticipar la respuesta de los consumidores es esencial. Esto permite a las empresas adaptar sus estrategias, optimizar el inventario y establecer precios adecuados, asegurando así una mayor eficiencia operativa y satisfacción del cliente.

El comercio electrónico ha transformado la forma en que hacemos negocios y ha elevado las expectativas de los consumidores en cuanto a variedad, precio y servicio. En este entorno dinámico, empresas como Flipkart deben estar un paso adelante, no solo satisfaciendo las necesidades actuales de los clientes, sino también anticipando futuras demandas. La aplicación de machine learning para predecir el grado de aceptación de productos, basándose en datos históricos de ventas, opiniones y devoluciones, representa una innovadora solución a este desafío. Al adoptar tales herramientas predictivas, las tiendas e-commerce pueden tomar decisiones más informadas y estratégicas, lo que se traduce en una ventaja competitiva en el mercado global y en una experiencia de compra optimizada para el cliente.

3 Metodología y estado de la cuestión


3.1 Marco tecnológico

3.1.1 Herramientas a utilizar

Las herramientas que se van a utilizar a lo largo del trabajo se resumen en la tabla 1:

Tabla 1- Herramientas empleadas

	<p>Como lenguaje de programación se han utilizado Python apoyándonos de las siguientes librerías:</p> <ul style="list-style-type: none">• Pandas: Provee series de datos y estructuras de Data Frame bidimensionales.• NumPy: Ofrece una estructura de datos universal que facilita el análisis e intercambio entre diversos algoritmos<ul style="list-style-type: none">• Matplotlib: Facilita la generación de gráficos.• Sklearn: Especializada en análisis predictivo, incluye clasificadores, algoritmos de agrupación y más. Se basa en NumPy, SciPy y Matplotlib.
	<p>Para el desarrollo y la ejecución del lenguaje Python se ha utilizado el Google Collab</p>
	<p>La arquitectura empleada para desarrollar el proyecto se basa en los servicios ofrecidos por Google Cloud Platform (GCP)</p>
 Cloud Storage	<p>Para la persistencia de los datos una vez procesados se ha utilizado un bucket de Google Cloud Storage</p>

	<p>Para la elaboración de los cuadros de mando y la visualización de los datos se ha utilizado Power Bi</p>
---	---

3.1.2 Hardware utilizado

Para el desarrollo de este trabajo se han utilizado ordenadores portátiles con diferentes características que pasamos a detallar:

- Portátil Aspire – E15

Disco solido de 1TB - Procesador Intel(R) Core™ i7-6500U 2.5GB, (4 CPUs), ~ 3.1 GHz.8 Gb
8 GB Ram DDR4

- PC sobremesa

Disco solido de 1TB - Intel core i7 6500 3.1ghzProcesador 16 GB Memoria Ram

- AMD Ryzen 9 5980HX

Radeon Graphics 3.30 Ghz 16 GB RAM

- Prestige 15 A10SC , Windows 10 Pro

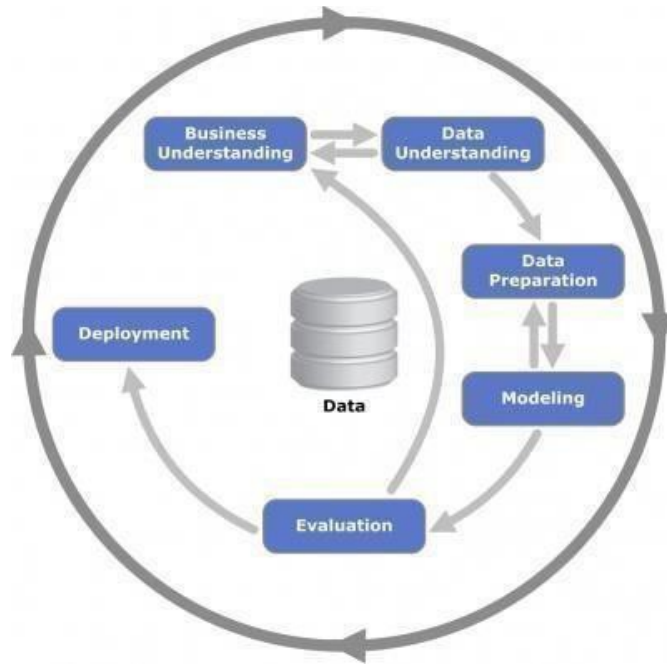
Intel(R) Core(TM) i7-10710U CPU @ 1.10GHz (12 CPUs), ~1.6GHz
16 Gb RAM

- DESKTOP-FNNCLS5

Procesador 11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 2.80 GHz
RAM instalada 20,0 GB

3.2 Metodología

El desarrollo de este proyecto se realiza en base a la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Dicha metodología consiste en 6 fases como se muestra a continuación:



Grafica 1: Fases del Proceso de Modelado metodología CRISP-DM

Estas fases interactúan entre ellas de manera iterativa durante todo el desarrollo del proyecto.

3.3 Fases Para Implementar

A continuación, se detallan los pasos a seguir para implementar la solución al problema planteado.

3.3.1 Comprensión del Negocio

Predecir el grado de aceptación de un producto en Flipkart utilizando machine learning. El criterio de éxito definido es desarrollar un modelo que pueda predecir con un alto grado de precisión el éxito de un producto basándose en ventas pasadas, opiniones y/o devolución, así como también el rating de cada producto.

3.3.2 Comprensión de los Datos

Se tomó como base de datos dos archivos csv de la e-commerce Flipkart la cual se encuentra en la plataforma Kaggle. Este csv contiene información histórica de ventas, reseñas de clientes y rating de los productos entre otros datos relevantes.

Además, se realiza una exploración inicial utilizando pandas y matplotlib para analizar estadísticas descriptivas, visualizar distribuciones y detectar posibles outliers.

3.3.3 Preparación de los datos

En este punto la limpieza se realizó con Pandas para limpiar datos faltantes, eliminar duplicados y corregir errores. Además, las transformaciones considerando así mismo el uso de Pandas y Numpy para transformar datos categóricos en numéricos, normalizar datos y crear nuevas características en base a columnas existentes.

3.3.4 Modelado

En este punto se probaron diversos modelos. Entre los cuales el que dio mejores resultados fue Random Forest. Como toda metodología se dividió el conjunto de datos en entrenamiento y prueba (train y test). Posteriormente, se evaluó este modelo con datos nuevos para ver su comportamiento y, frente a esto, se tiene un alto grado de precisión.

3.3.5 Evaluación

La evaluación se hizo usando métricas como:

Despliegue

En esta etapa del proyecto, una vez definido el modelo, lo evaluamos revisando los pasos ejecutados comprobando que da cumplimiento a los objetivos planteados para la solución del problema y examinamos errores por si alguna cuestión importante del negocio no ha sido considerada suficientemente.

Explotación

En esta fase se implementa el modelo, se validan los resultados obtenidos generando, con toda la información obtenida, un informe a la entidad con las conclusiones y recomendaciones a desarrollar para la futura toma de decisiones.

3.4 Estado del Arte

3.4.1 Machine Learning

El Machine Learning o aprendizaje automático es una disciplina que combina ciencia y arte, y se dedica a la creación de sistemas que pueden aprender y mejorar su desempeño a partir de datos.

El Machine Learning es útil en:

- Problemas que requieren constante ajuste y actualización.
- Problemas complejos donde los métodos tradicionales no son efectivos.
- Situaciones que cambian y evolucionan con el tiempo.

Existen múltiples categorías y sistemas dentro del Machine Learning.

Una de las clasificaciones más comunes es:

- Por Supervisión:
 - Supervisado
 - No supervisado
 - Semisupervisado
 - Aprendizaje por refuerzo

Por Modalidad de Aprendizaje:

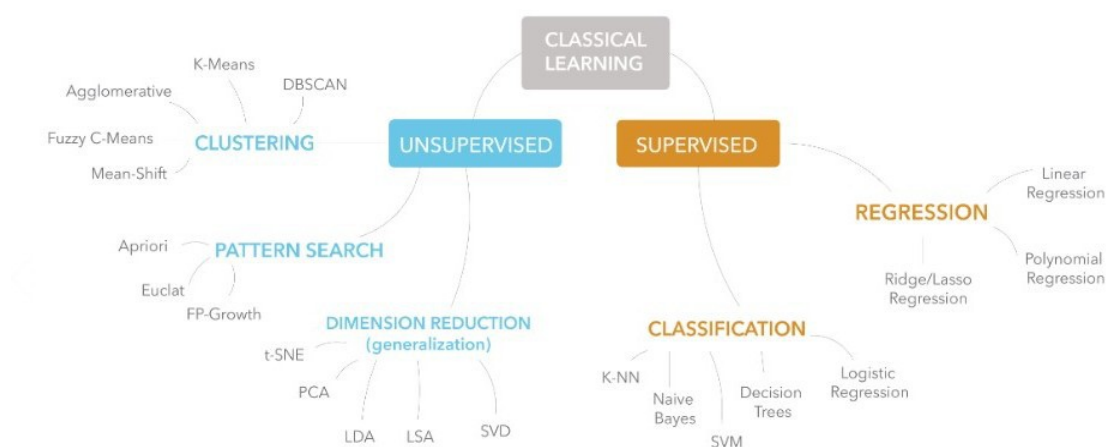
- Aprendizaje en línea
- Aprendizaje por lotes

Por Tipo de Enfoque:

- Basado en instancias
- Basado en modelos

El aprendizaje basado en instancias se enfoca en comparar nuevos datos con datos previamente conocidos, mientras que el aprendizaje basado en modelos busca detectar patrones en el conjunto de entrenamiento y construir un modelo predictivo

En la siguiente figura se muestra un resumen de los algoritmos más comunes de Machine Learning, divididos entre aprendizaje supervisado y no supervisado. A continuación, se detallan los algoritmos más populares en cada categoría.



Grafica 2 Mapa de aprendizaje automático

Supervisados:

- Clasificación:

Árboles de decisión: Estos algoritmos dividen el espacio de datos en subconjuntos basados en decisiones jerárquicas. Son intuitivos y fácilmente visualizables, y son empleados frecuentemente para problemas de clasificación y regresión

Naïve Bayes: Basado en el teorema de Bayes, este algoritmo asume independencia entre los predictores. A pesar de su simplicidad, es eficaz, especialmente en clasificación de textos

Support Vector Machines (SVM): Intenta encontrar un hiperplano que separe de forma óptima las clases en un espacio de características. Es potente para clasificación lineal y no lineal

Regresión Logística: A pesar de su nombre, se utiliza para problemas de calificación binaria. Estima la probabilidad de que una instancia pertenezca a una clase particular

- Regresión:

Regresión lineal: Intenta encontrar una línea recta que mejor se adapte a los datos. Es uno de los métodos más básicos y ampliamente utilizado para la predicción numérica

Regresión por mínimos cuadrados: Es un método que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos

Métodos “Ensemble” (Conjuntos de clasificadores): Estos métodos combinan las predicciones de varios modelos base (por ejemplo, árboles de decisión) para mejorar la robustez y precisión

No supervisados:

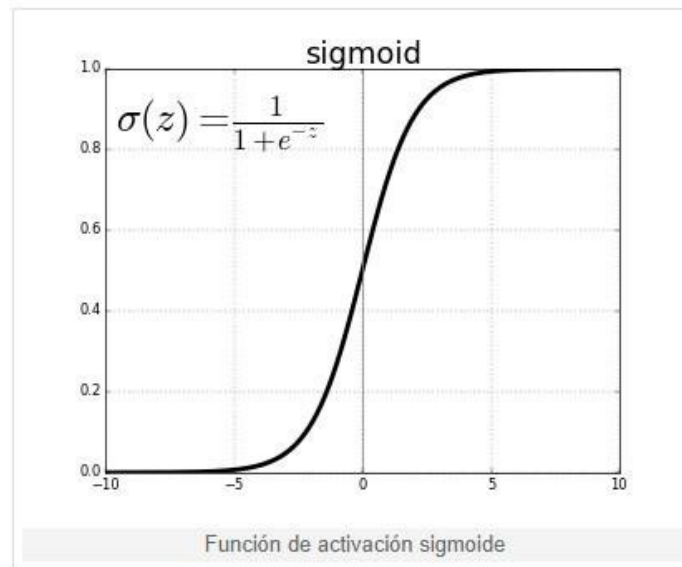
- Algoritmos de clustering: Estos algoritmos agrupan datos en conjunto basados en similitudes sin tener etiquetas predefinidas. Ejemplos incluyen K-means y DBSCAN.
- Análisis de componentes principales: Es una técnica de reducción de dimensionalidad que busca transformar variables correlacionadas en un conjunto de valores no correlacionados llamados componentes principales
- Descomposición en valores singulares (singular value decomposition): Es una técnica de factorización utilizada en reducción de dimensionalidad y en sistemas de recomendación
- Análisis de componentes principales (Independent Component Analysis): Similar al PCA, pero en lugar de encontrar componentes ortogonales con la mayor varianza, busca componentes que son estadísticamente independientes.

3.5 Algoritmos utilizados

3.5.1 Logistic Regression. Regresión Logística para Clasificación

La Regresión Logística es un método ampliamente utilizado en aprendizaje automático, específicamente para tareas de clasificación. Se puede conceptualizar como una red neuronal simplificada, equivalente a una única neurona. Su principal función es predecir la probabilidad de un resultado específico para una variable dependiente categórica. Esta variable, en muchos casos, es binaria y se representa con codificaciones como 1 y 0, sí y no, abierto y cerrado, entre otros.

La esencia de la Regresión Logística radica en su capacidad para transformar sus predicciones en un rango entre 0 y 1. Esto se logra mediante la función Sigmoide, que produce una curva en forma de 'S'. Esta curva permite que los valores resultantes se interpreten como probabilidades, facilitando la clasificación de las observaciones en una de las dos categorías.



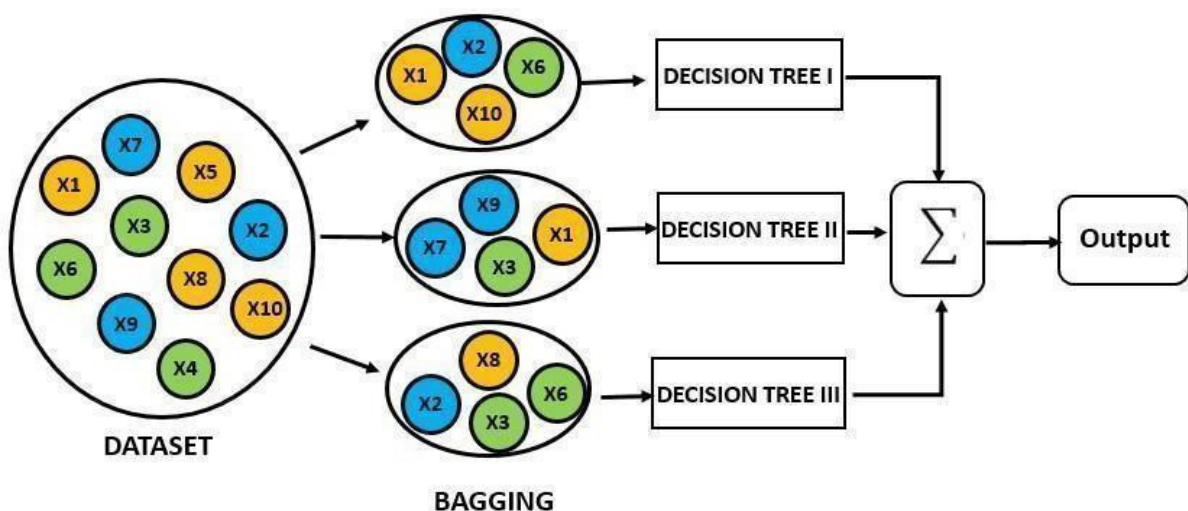
Grafica 3 Función sigmoide

Una gran ventaja de este algoritmo es que es muy eficiente y no requiere muchos recursos computacionales, pero por otro lado la regresión logística presenta el inconveniente de que requiere tamaños de muestra bastante grandes.

3.5.2 Árboles de decisión y Random Forest

En nuestra exploración de la ciencia de datos, seleccionamos el uso de la técnica de Random Forest, una extensión avanzada del tradicional árbol de decisión. El algoritmo Random Forest opera generando múltiples árboles de decisión durante la fase de entrenamiento y combina sus resultados para producir una predicción más precisa y robusta en la fase de prueba. Para nuestro análisis, adoptamos una división convencional de datos: 70% para entrenamiento y 30% para prueba.

Cada árbol se construye utilizando un conjunto de observaciones seleccionadas aleatoriamente del conjunto de entrenamiento a través del método de bootstrap. Al finalizar, todos los árboles individuales se "ensamblan" para formar un modelo cohesivo. En el caso de predicciones numéricas, el valor promedio de todos los árboles se toma como resultado, mientras que, para las categóricas, se adopta un enfoque de "mayoría de votos". A continuación, una imagen del método



Los motivos por los que hemos elegido este método son los siguientes:

- Versatilidad: Es ampliamente aplicado en tareas de clasificación y predicción, alineándose con el propósito central de nuestro estudio
- Eficiencia: Es un modelo que se entrena con relativa facilidad y rapidez
- Flexibilidad: Al ser no paramétrico, no impone restricciones estrictas sobre la distribución de los datos
- Diversidad de los datos: Puede manejar tanto datos numéricos como categóricos sin complicaciones
- Preparación Simplificada: Requiere menos preprocesamiento y limpieza de datos en comparación con otros métodos
- Relevancia Práctica: Es una herramienta comúnmente empleada en el ámbito del e-commerce para anticipar comportamientos clave, como si un cliente realizara una compra

3.5.3 XGboost para clasificaciones (XGBClassifier)

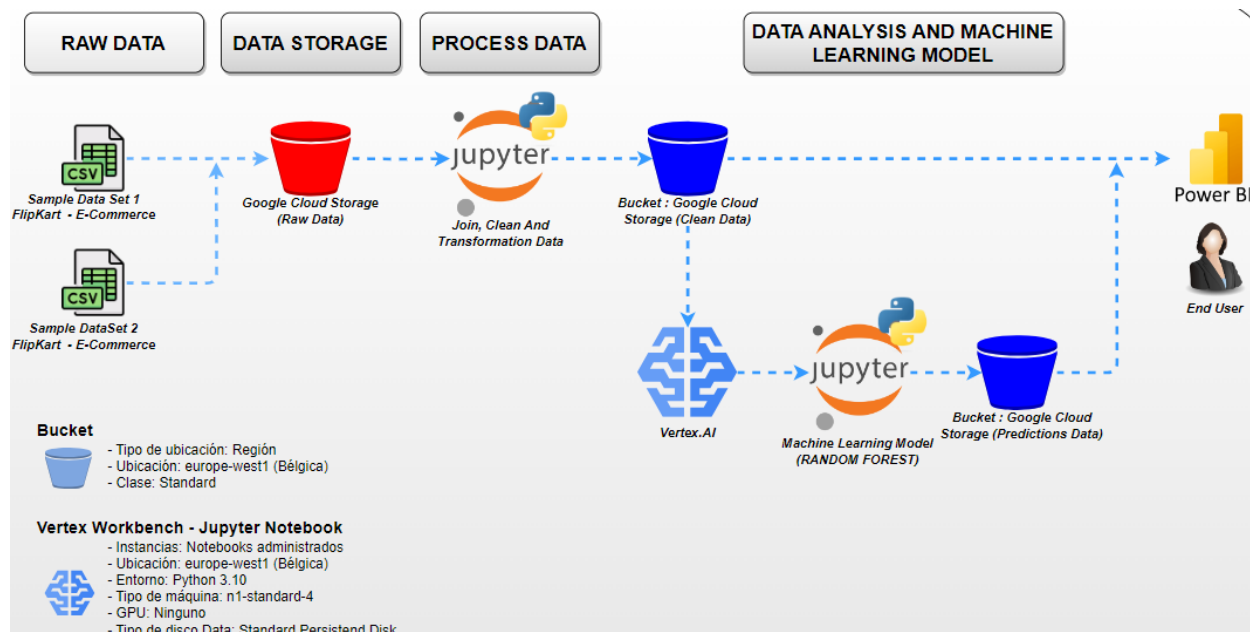
El XGBClassifier es una interfaz de programación de aplicaciones (API) diseñada específicamente para tareas de clasificación, y se fundamenta en la potencia del algoritmo XGBoost, o Extreme Gradient Boosting. Este algoritmo ha ganado notoriedad en el mundo del aprendizaje automático por su rendimiento excepcional en diversas competencias y aplicaciones del mundo real.

XGBoost es una evolución de los métodos de "Tree Boosting". Fue concebido y popularizado gracias al proyecto "XGBoost: A Scalable Tree Boosting System", liderado por Tianqi Chen y Carlos Guestrin. En su innovador trabajo, no sólo adaptaron el concepto de "Tree Boosting", sino que también implementaron técnicas de optimización para mejorar la eficiencia en términos de escalabilidad y rendimiento. Entre las optimizaciones destacadas, lograron una mejor utilización de la memoria caché y la incorporación de procesamiento en paralelo, lo que permite que XGBoost maneje grandes conjuntos de datos con una fracción del recurso requerido por otros algoritmos.

4 Propuesta de Arquitectura

Para el análisis de datos, es fundamental adoptar un esquema de trabajo bien definido y una estructura de información coherente. Con el respaldo de diversas herramientas de software y la arquitectura propuesta basada en los servicios de Google Cloud Platform (GCP), logramos una comprensión profunda del proceso que atraviesan los datos, desde su adquisición inicial hasta su visualización final.

En la siguiente figura se observa la arquitectura propuesta:



Grafica 5 Diagrama de Arquitectura

Los datos serán procesados en base a la arquitectura propuesta y seguirán los siguientes procesos.

- Origen de Datos
- Almacenamiento de los datos
- Limpieza, transformación y enriquecimiento de los datos
- Persistencia: almacenamiento de los datos tratados
- Modelado de Machine Learning
- Visualización de los datos

4.1 Origen de Datos

El primer paso fue identificar y obtener un conjunto de datos 'Reales' de flipkart. Se recurrió a Kaggle, plataforma conocida por hospedar vastos conjuntos de datos de diversos dominios. Este conjunto de datos proporcionó una visión integral de las transacciones, preferencias de los clientes y otros aspectos cruciales relacionados al e-commerce

4.2 Almacenamiento de los Datos

Una vez adquiridos los datos, era esencial asegurar un almacenamiento adecuado. Se optó por Google Cloud Storage, una solución en la nube que no solo garantiza seguridad, si no también escalabilidad y accesibilidad inmediata a los datos cuando se requiera.

4.3 Limpieza, transformación y enriquecimiento de los Datos

Antes del análisis, es crucial asegurar que los datos sean coherentes, completos y estén en el formato adecuado. Para ello, se utilizó librerías de Python, específicamente Pandas y Numpy para limpiar, transformar y enriquecer el conjunto

de datos. Estas herramientas permitieron eliminar valores atípicos, tratar datos faltantes y transformar variables para optimizar su uso en el modelado posterior

4.4 Persistencia:

Posterior a la limpieza, los datos enriquecidos necesitan ser almacenados de nuevo para garantizar su integridad. Se recurrió nuevamente a Google Cloud Storage asegurando que las versiones más refinadas del conjunto de datos estuvieran listas y disponibles para el análisis y modelado posterior

4.5 Modelado de Machine Learning

Con los datos preparados, el siguiente paso fue implementar un modelo que pudiera predecir con precisión el éxito de los productos. Se eligió el algoritmo **Random Forest** por su robustez y capacidad para manejar grandes conjuntos de datos. A través de múltiples árboles de decisión, este algoritmo proporciona predicciones precisas y tiene la ventaja de reducir el riesgo de sobreajuste

4.6 Visualización de los Datos

El objetivo principal de la visualización es la transformación de un conjunto de datos en un resultado visual para comunicar información de forma clara, precisa y eficiente. Para ello, es esencial seleccionar el tipo de gráfico adecuado para cada conjunto de datos teniendo en cuenta la clase de análisis que se quiere proyectar.

La visualización de datos tiende a alcanzar tres objetivos:

- 5** Explorar un conjunto de datos para obtener respuestas a cuestiones previamente planteadas y a la vez elaborar nuevas preguntas.
- 6** Analizar patrones, relaciones y valores atípicos existentes en el conjunto de los datos para identificar circunstancias no detectadas y extraer conclusiones

Explicar el significado de los datos analizados permitiendo llegar la información de forma adecuada y que ayude a la toma de decisiones según las conclusiones obtenidas

Para el desarrollo de esta parte del TFB se ha estudiado el uso de herramientas BI para agregar valor a los datos procesados a través del fácil acceso y visualización de la información, así como la toma de decisiones.

Se empleo Power Bi, una herramienta de visualización líder en la industria, para representar gráficamente los datos y resultados. Esto permitió una interpretación más intuitiva y ofreció insights valiosos para los stakeholders de flipkart.

7 ANÁLISIS DE RESULTADOS DE ML

7.1 PRICING ML

En un principio se suponía que brand y sub_category eran las características más importantes al momento de determinar el precio de un producto.

Luego de aplicar el modelo de ML, se descubrió que la suposición era correcta, pero no se tomaba en cuenta las características de seller ni product_details; que se descubrió, eran importantes para determinar el precio de los productos.

Las métricas seleccionadas han sido correctas porque al visualizar la Matriz de Correlación, se pudo visualizar características que compartían lazos, como por ejemplo seller y brand.

La arquitectura ha sido la correcta porque los datos son del tipo batch, el tamaño de los datasets no es de gran tamaño, la funcionalidad de los requerimientos no requería gran escalabilidad y el acceso a los datos para análisis no es masiva.

Los mejores modelos:

- Regresión Lineal: R^2 (train): 86.8% / R^2 (test): 86.7%
- Árbol de decisión: Train Accuracy: 94.5% Test Accuracy: 91.7%
- Random Forest: Train Accuracy: 79.2% Test Accuracy: 77.9%

Se decidió utilizar el modelo Random Forest por mejor generalización en la predicción de precios y porque Random Forest es más robusto a valores atípicos que pudieran estar aplicando el modelo a un dataset diferente.

7.2 Classification ML:

En un principio se suponía que brand, category, sub_category y title eran las características más importantes al momento de determinar la aceptación de los productos.

Luego de aplicar el modelo de ML, se descubrió que la suposición era correcta, pero no se tomaba en cuenta la característica de seller para determinar la aceptación de los productos.

Las métricas seleccionadas han sido correctas porque al visualizar la Matriz de Correlación, se pudo visualizar características que compartían lazos, como por ejemplo seller y brand.

La arquitectura ha sido la correcta porque los datos son del tipo batch, el tamaño de los datasets no es de gran tamaño, la funcionalidad de los requerimientos no requería gran escalabilidad y el acceso a los datos para análisis no es masiva.

Los mejores modelos:

- XGBoost: Accuracy (train): 82% / Accuracy (test): 78%
- Random Forest: Train Accuracy: 90% / Test Accuracy: 83%

Se decidió utilizar el modelo Random Forest porque se encontró un problema con el rango de etiquetado (1, 2, 3, 4, 5), que forzaba que las etiquetas fuera 0, 1, 2, 3, 4. Por lo que, si no hay registros de alguna etiqueta en particular, el modelo da

error. El rendimiento con XGBoost generaliza mejor, pero por esa particularidad con las etiquetas, se escogió en primera instancia, Random Forest. Es más flexible en ese aspecto.

8 Desarrollo del departamento de big data

8.1 Medios técnicos

En caso de necesitar una arquitectura más potente a utilizar para el departamento se han considerado varias alternativas entre los principales proveedores Cloud, Amazon, Google y Microsoft.

Tras un estudio económico se ha determinado que la mejor opción es la mostrada en la siguiente tabla

Tabla 2 ESTUDIO ECONOMICO

Estudio económico

Google Storage	Cloud	1 Procesador	4GB	110 €/mes
Google Colab		2 Procesadores	16GB	80 €/mes
Vertex Ai.		2 Tb		50 €/mes
			Total Mes	540 €/mes

La opción de tener un sistema Cloud es altamente valorable y recomendable en este caso. El coste anual no es excesivo y en caso de que el proyecto tenga necesidades de mayor potencia de procesamiento o almacenamiento la necesidad es fácilmente cubierta

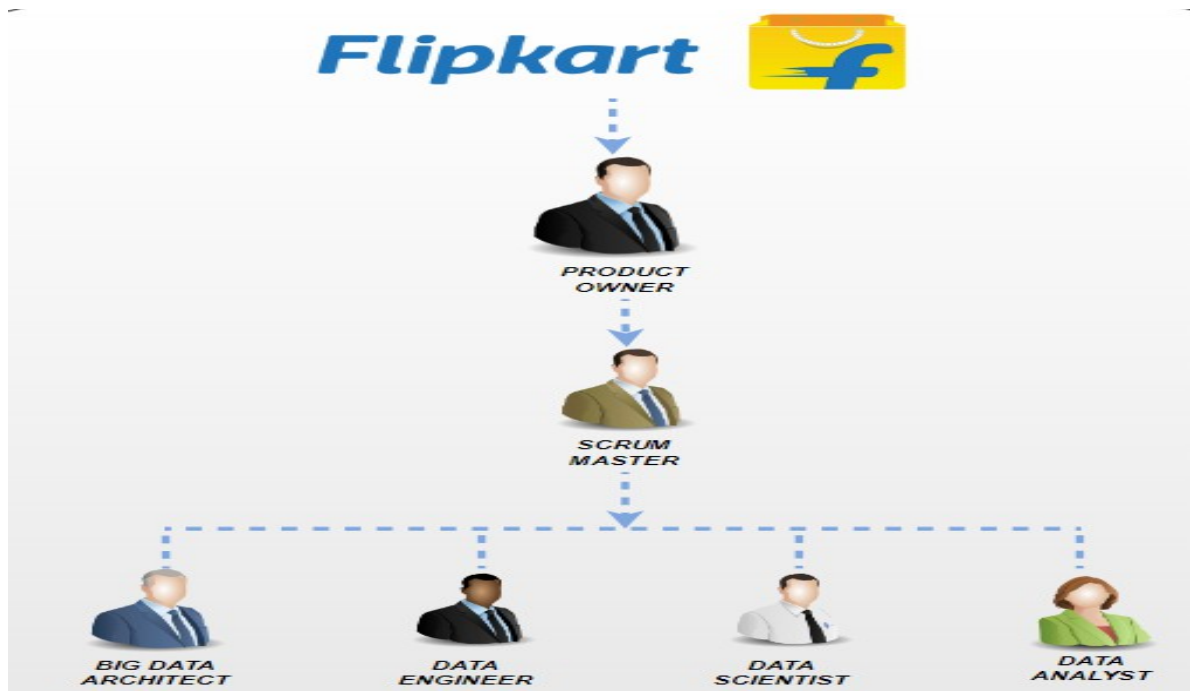
8.2 Medios humanos

El equipo para desarrollar el presente proyecto de Big Data considero los siguientes recursos humanos además de adoptar una metodología Agile:

- Big Data Architect: Diseñar y planificar la arquitectura de datos a gran escala
- Funciones en el proyecto:
- Determinar la arquitectura óptima y las tecnologías a utilizar, considerando la naturaleza de los datos de flipkart y las necesidades del proyecto

- Asegurar que la infraestructura sea escalable, resiliente y pueda manejar el volumen de datos previstos
- Colaborar estrechamente con el Data Engineer para implementar soluciones de almacenamiento y procesamiento
- Data Engineer: Construir y mantener la infraestructura de datos
Funciones en el proyecto:
 - Implementar la arquitectura diseñada
 - Preparar y preprocesar los datos, asegurando su disponibilidad y formato adecuado para el análisis
 - Administrar y optimizar consultas y procesos para garantizar un rendimiento eficiente
- Data Scientist: Desarrollar modelos de aprendizaje automático
Funciones en el proyecto:
 - Analizar el conjunto de datos para identificar patrones y tendencias
 - Diseñar, implementar y evaluar los modelos predictivos
 - Optimizar modelos en función de métricas específicas y feedback del Data Analyst
- Data Analyst: Analizar y presentar insights
Funciones en el proyecto:
 - Utilizar herramientas de visualización para representar gráficamente los datos y resultados
 - Producir informes y dashboards que informen decisiones estratégicas
- Product Owner: Definir la visión del producto y asegurarse de que el equipo entregue valor al negocio
Funciones en el proyecto:
 - Priorizar las características y tareas basadas en el valor y el ROI
 - Actuar como enlace entre los stakeholders y el equipo de desarrollo
 - Asegurar que los requisitos sean claros y que el equipo tenga todo lo necesario para implementar soluciones efectivas
- Scrum Master: Facilitar el proceso Scrum y asegurar que el equipo siga las prácticas ágiles
 - Funciones en el proyecto Flipkart:
 - Coordinar las ceremonias Scrum, como las daily stand-ups y las retrospectives
 - Eliminar obstáculos que impidan el progreso del equipo
 - Asegurarse de que el equipo mantenga un flujo de trabajo constante y eficiente

Estos roles asumidos, en conjunto, permitieron el desarrollo del proyecto y la finalización de este trabajando de manera colaborativa siguiendo prácticas ágiles para llevar el proyecto desde la concepción inicial hasta la entrega de soluciones basadas en datos para Flipkart. A continuación, un diagrama de los roles asumidos:



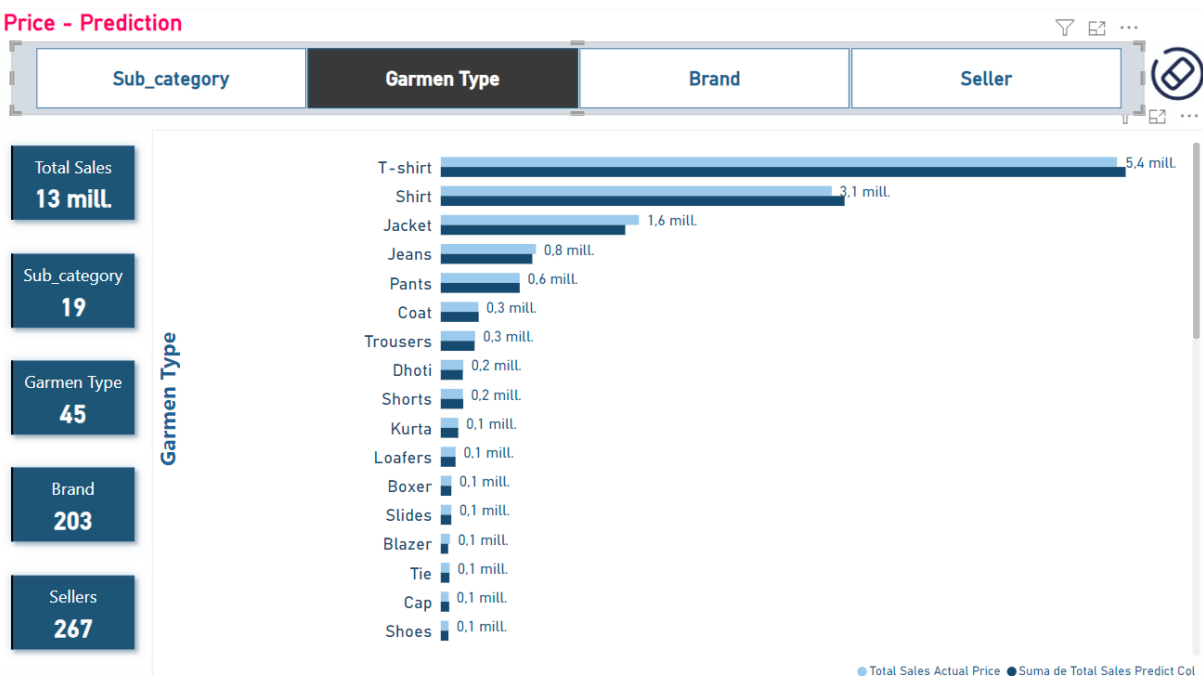
9 Visualización de los resultados

Una finalidad de los datos una vez procesados, es el análisis mediante los algoritmos propuestos en los puntos anteriores. Y la otra fundamental, es que sirvan para la toma de decisiones en el día a día del equipo directivo del centro.

Para ello se ha elaborado un cuadro de mando utilizando la aplicación Power BI:

1. Se establecieron parámetros para evaluar : Sub_Category, Garmen Type, Brand y Seller:

Price - Prediction



2. Se muestran las “Brands” y “Sellers” que reportan las mayores ventas

Best Sellers and Brands



Se guía al cliente para establecer las mejores opciones, con la predicción de precios y el promedio para la aceptación de Brand y Seller.

The Choice



Coat	Dhoti	Jacket	Jeans	Kurta	Pants	Shirt	Shorts	Trousers	T-shirt
------	-------	--------	-------	-------	-------	-------	--------	----------	---------

Brand	Predicted Price
ABC ANY BUDY CLE	1434.50
allan pet	1732.67
ALQI	1353.67
ARBO	1194.00
FOREVER YOU	1712.00
Lafant	1960.00
LDHSA	2386.27
Lucky Bi	1983.00
Mah	1712.55
Mo	1385.00
Purple Sta	2351.75
SATDEVANGIK HADIBHAND	1015.83
Solid Styl	1946.00
Total	1787.93

Brand	Predicted Rating
ABC ANY BUDY CLE	4.00
allan pet	4.14
ALQI	5.00
FOREVER YOU	3.60
Jagdish Garmen	4.00
Lafant	4.00
LDHSA	3.97
Mah	4.06
Mo	3.00
Purple Sta	3.14
SATDEVANGIKHA DIBHAND	4.33
SORA	5.00
Szto	2.00
Total	4.04

Seller	Predicted Rating
VOXATI	4.00
UTTRAKHAND CLOTHHOUSE	3.97
UNKNOWN	2.00
SORANG	5.00
SHREYASHFASHION	2.00
SH APPAREL ONLINE	3.60
Satdevangi Khadi Bhandar	4.33
Purplestate	3.14
Mahir Apparels	4.31
KrishnamEnterpri saes	5.00
KRISHNAM	3.00
Total	4.04

10 Conclusiones

- I. Teniendo en cuenta lo aprendido, se volvería a tomar en cuenta el punto de vista del cliente sobre los datos que tiene, como primer paso; y luego indagar más con herramientas como la Matriz de Correlación para buscar más patrones en los datos. Así se logra entender mejor la información que se cuenta. Lo que se pudiera hacer diferente o adicional, sería hacer más granular los datos del cliente. Porque en columnas como product_details y description, se pueden crear nuevas columnas para lograr más detalles de cada uno de los productos.
- II. De lo que se pudo obtener del dataset, es que los precios de los productos se basan principalmente de seller, brand, product_details, title, category, sub_category.
- III. Se concluye que el conocimiento experto del cliente es la primera información que se debe de consultar, y que con Data Mining es posible conseguir encontrar más información, que posiblemente no sea vista o tomada con la importancia que se debería.
- IV. De lo que se pudo obtener del dataset, es que seller es una característica muy importante para determinar el grado de aceptación de los productos.
- V. Se concluye que el conocimiento experto del cliente es la primera información que se debe de consultar, y que con Data Mining es posible conseguir encontrar más información, que posiblemente no sea vista o tomada con la importancia que se deber.