Proyecto Final

Análisis de Sentimientos en Opiniones de Pasajeros de una Aerolínea con NLP y PySpark en un entorno Big Data

Andrea Buenaño **28/05/2024**

Introducción:

El objetivo de este proyecto es realizar un análisis de sentimientos sobre los comentarios de los pasajeros de una aerolínea utilizando técnicas de Procesamiento de Lenguaje Natural (NLP) en un entorno Big Data con Apache Spark y Python (PySpark). El análisis busca identificar opiniones positivas, negativas y neutrales en los comentarios, y correlacionar estos sentimientos con otras variables disponibles en el dataset, tales como la valoración general, el tipo de viajero y la clase de servicio (económica, business, etc.). Este estudio permitirá comprender cómo los diferentes aspectos de la experiencia de vuelo afectan la percepción y satisfacción del cliente.

Relevancia del Proyecto

El análisis de sentimientos es una herramienta poderosa para las aerolíneas, ya que les permite obtener una comprensión profunda de las opiniones y experiencias de sus clientes. Al correlacionar estos sentimientos con diferentes aspectos del servicio, la aerolínea puede identificar áreas de mejora y tomar decisiones informadas para mejorar la satisfacción del cliente y, en última instancia, su lealtad. En un entorno competitivo como el de las aerolíneas, la capacidad de interpretar y actuar sobre los comentarios de los pasajeros puede ofrecer una ventaja significativa.

Estructura de los Datos

El conjunto de datos, descargado de Kaggle, ha sido elegido por su detallada exploración de la experiencia de los pasajeros con Ryanair, una de las principales aerolíneas de bajo coste en Europa. Este conjunto reúne opiniones y calificaciones de los pasajeros, abarcando aspectos de los vuelos de Ryanair desde 2012 hasta 2024.

Este dataset ofrece una diversidad de opiniones y experiencias, permitiendo analizar en detalle las fortalezas y debilidades de los servicios ofrecidos por Ryanair. Con datos que abarcan más de una década, es posible identificar tendencias y cambios en la percepción de los pasajeros a lo largo del tiempo. Además, la información sobre tipos de viajeros y rutas voladas facilita la comparación entre diferentes segmentos y rutas, proporcionando una visión integral del desempeño de la aerolínea.

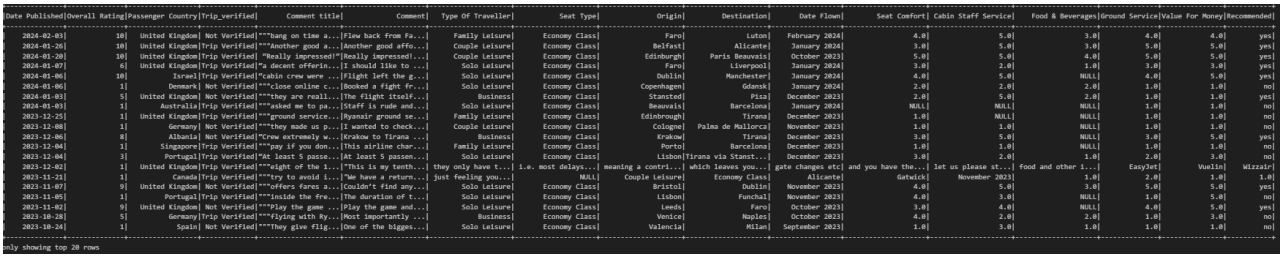
El conjunto de datos fue recopilado mediante técnicas de web scraping del sitio AirlineQuality.com, utilizando la biblioteca BeautifulSoup. Esta recolección de datos ha sido documentada y puesta a disposición en Kaggle, y se puede encontrar más detalles del proceso en este notebook de Kaggle.

El conjunto de datos tiene 21 columnas y 2249 registros, cada uno representando una revisión de un pasajero de la aerolinea:

1. **Unnamed: 0**: Un índice numérico que probablemente sea un remanente del proceso de importación de datos.
2. **Date Published**: Fecha en la que se publicó el comentario. Es crucial para analizar cambios en las percepciones a lo largo del tiempo.
3. **Overall Rating**: Calificación general del vuelo, en una escala numérica. Es un indicador clave de la satisfacción del pasajero.
4. **Passenger Country**: País de origen del pasajero, útil para identificar patrones geográficos en las opiniones.
5. **Trip_verified**: Indica si el viaje fue verificado, ayudando a determinar la autenticidad de las opiniones.
6. **Comment title**: Título del comentario del pasajero, que resume la opinión general.
7. **Comment**: Cuerpo del comentario proporcionado por el pasajero, ofreciendo detalles cualitativos sobre la experiencia de vuelo.
8. **Aircraft**: Tipo de aeronave en la que el pasajero voló, lo cual puede influir en la experiencia de vuelo.
9. **Type Of Traveller**: Tipo de viajero (negocios, ocio, familia), importante para segmentar las experiencias según el propósito del viaje.
10. **Seat Type**: Tipo de asiento ocupado por el pasajero, influyendo en la comodidad y satisfacción general.
11. **Origin**: Ciudad o aeropuerto de origen del vuelo. Junto con el destino, permite análisis sobre rutas específicas.
12. **Destination**: Ciudad o aeropuerto de destino del vuelo, útil para evaluar la satisfacción en diferentes rutas.
13. **Date Flown**: Fecha del vuelo. Permite correlacionar la experiencia con posibles eventos específicos o cambios en el servicio.
14. **Seat Comfort**: Calificación de la comodidad del asiento, en una escala numérica, crítica para evaluar este aspecto clave.
15. **Cabin Staff Service**: Calificación del servicio del personal de cabina, un indicador clave de la calidad del servicio al cliente.
16. **Food & Beverages**: Calificación de los alimentos y bebidas ofrecidos, relevante para vuelos largos.
17. **Ground Service**: Calificación del servicio en tierra, incluyendo el check-in, embarque y desembarque.
18. **Value For Money**: Calificación del valor recibido por el dinero pagado, una métrica importante para una aerolínea de bajo coste.

19. **Recommended**: Indica si el pasajero recomendaría la aerolina a otros, una medida directa de satisfacción y lealtad.
20. **Inflight Entertainment**: Calificación del entretenimiento a bordo, relevante aunque no todos los vuelos lo ofrezcan.
21. **Wifi & Connectivity**: Calificación del servicio de Wi-Fi y conectividad a bordo, importante para pasajeros que necesitan mantenerse conectados.

Estas variables permiten un análisis completo y detallado de la experiencia del pasajero, ofreciendo datos cuantitativos y cualitativos para evaluar la satisfacción del cliente y áreas de mejora.



Date Published	Overall Rating	Passenger Country	Trip_verified	Comment title	Comment	Type Of Traveller	Seat Type	Origin	Destination	Date Flown	Seat Comfort	Cabin Staff Service	Food & Beverages	Ground Service	Value For Money	Recommended
2024-02-03	10	United Kingdom	Not Verified	"bang on time a...	Flew back from Fa...	Family Leisure	Economy Class	Faro	Luton	February 2024	4.0	5.0	3.0	4.0	4.0	yes
2024-01-26	10	United Kingdom	Trip Verified	"Another good a...	Another good affo...	Couple Leisure	Economy Class	Belfast	Alicante	January 2024	3.0	5.0	3.0	5.0	5.0	yes
2024-01-20	10	United Kingdom	Trip Verified	"Really impressed!"	Really impressed!...	Couple Leisure	Economy Class	Edinburgh	Paris Beauvais	October 2023	5.0	5.0	4.0	5.0	5.0	yes
2024-01-07	6	United Kingdom	Trip Verified	"a decent offerin...	I should like to ...	Solo Leisure	Economy Class	Faro	Liverpool	January 2024	3.0	2.0	1.0	3.0	3.0	yes
2024-01-06	10	Israel	Trip Verified	"cabin crew were ...	Flight left the g...	Solo Leisure	Economy Class	Dublin	Manchester	January 2024	4.0	5.0	NULL	4.0	5.0	yes
2024-01-06	1	Denmark	Not Verified	"close online c...	Booked a flight fr...	Solo Leisure	Economy Class	Copenhagen	Gdansk	January 2024	2.0	2.0	2.0	1.0	1.0	no
2024-01-03	5	United Kingdom	Not Verified	"they are reall...	The flight itself...	Business	Economy Class	Stansted	Pisa	December 2023	2.0	5.0	2.0	1.0	1.0	yes
2024-01-03	1	Australia	Trip Verified	"asked me to pa...	Staff is rude and...	Solo Leisure	Economy Class	Beauvais	Barcelona	January 2024	NULL	NULL	NULL	1.0	1.0	no
2023-12-25	1	United Kingdom	Trip Verified	"ground service...	Ryanair ground se...	Family Leisure	Economy Class	Edinbrough	Tirana	December 2023	1.0	NULL	NULL	1.0	1.0	no
2023-12-08	1	Germany	Not Verified	"they made us p...	I wanted to check...	Couple Leisure	Economy Class	Cologne	Palma de Mallorca	November 2023	1.0	1.0	NULL	1.0	1.0	no
2023-12-06	8	Albania	Not Verified	"Crew extremely w...	Krakow to Tirana ...	Business	Economy Class	Krakow	Tirana	December 2023	3.0	5.0	NULL	3.0	5.0	yes
2023-12-04	1	Singapore	Trip Verified	"pay if you don...	This airline char...	Family Leisure	Economy Class	Porto	Barcelona	December 2023	1.0	1.0	NULL	1.0	1.0	no
2023-12-04	3	Portugal	Trip Verified	"At least 5 passe...	At least 5 passen...	Solo Leisure	Economy Class	Lisbon	Tirana via Stanst...	December 2023	3.0	2.0	1.0	2.0	3.0	no
2023-12-02	1	United Kingdom	Trip Verified	"sight of the 1...	This is my tenth...	they only have t...	i.e. most delays...	meaning a contri...	which leaves you...	gate changes etc	and you have the...	let us please st...	food and other i...	EasyJet	Vuelin	Wizzair
2023-11-21	1	Canada	Trip Verified	"try to avoid i...	We have a return...	just feeling you...	NULL	Couple Leisure	Economy Class	Alicante	Gatwick	November 2023	1.0	2.0	1.0	1.0
2023-11-07	9	United Kingdom	Not Verified	"offers fares a...	Couldn't find any...	Solo Leisure	Economy Class	Bristol	Dublin	November 2023	4.0	5.0	3.0	5.0	5.0	yes
2023-11-05	1	Portugal	Trip Verified	"inside the fra...	The duration of t...	Solo Leisure	Economy Class	Lisbon	Funchal	November 2023	4.0	3.0	NULL	1.0	5.0	no
2023-11-02	9	United Kingdom	Not Verified	"Play the game ...	Play the game and...	Solo Leisure	Economy Class	Leeds	Faro	October 2023	3.0	4.0	NULL	4.0	5.0	yes
2023-10-28	5	Germany	Trip Verified	"flying with Ry...	Most importantly ...	Business	Economy Class	Venice	Naples	October 2023	4.0	2.0	2.0	1.0	3.0	no
2023-10-24	1	Spain	Not Verified	"They give flig...	One of the bigges...	Solo Leisure	Economy Class	Valencia	Milan	September 2023	1.0	3.0	1.0	1.0	1.0	no

only showing top 20 rows

Objetivos del proyecto:

Objetivo Principal:

- Realizar un análisis de sentimientos: Analizar los comentarios de pasajeros para clasificarlos en sentimientos positivos, negativos y neutros.

Objetivos Especificos:

- Correlación de sentimientos y valoración global: Determinar la correlación entre los sentimientos de los comentarios y la valoración global del servicio.
- Relación con el tipo de viajero y clase de servicio: Analizar la relación entre los sentimientos expresados y el tipo de viajero, así como la clase de servicio utilizada.
- Identificación de patrones y tendencias: Identificar patrones y tendencias que ayuden a mejorar la calidad del servicio ofrecido por las aerolíneas.
- Desarrollo y aplicación de modelos de Machine Learning:
  - Random Forest: Implementar un modelo de Random Forest para manejar grandes volúmenes de datos y mejorar la precisión en la clasificación de sentimientos.
  - Logistic Regression: Utilizar Logistic Regression para ofrecer una solución simple y eficiente en la clasificación de sentimientos.
  - Naïve Bayes: Emplear el modelo de Naïve Bayes por su rapidez y eficiencia, especialmente adecuado para grandes conjuntos de datos y la clasificación de sentimientos.

Estos objetivos permiten no solo comprender mejor los comentarios de los pasajeros, sino también aprovechar técnicas avanzadas de machine learning para realizar predicciones y análisis más precisos, contribuyendo así a la mejora continua del servicio ofrecido por las aerolíneas.

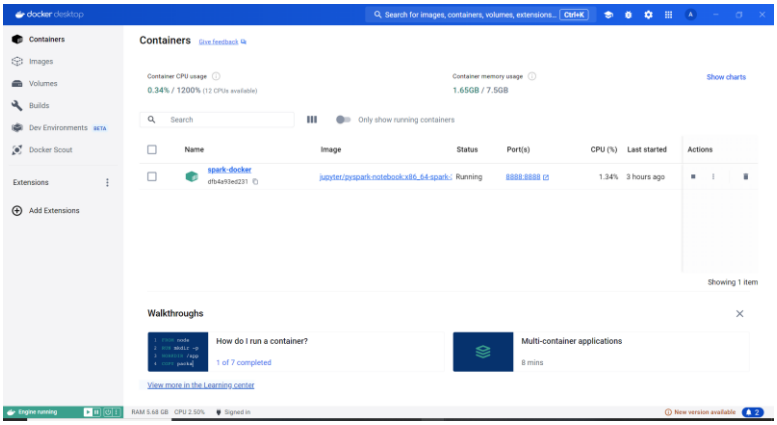
Configuración del Entorno Big Data

Para nuestro proyecto de análisis de sentimientos hemos decidido configurar un entorno Big Data usando un clúster de Spark dentro de un contenedor Docker. Este clúster está diseñado para trabajar con Jupyter Notebook y PySpar

Esta configuración es muy ventajosa debido a varias razones prácticas. Primero, Spark nos permite procesar grandes volúmenes de datos de manera distribuida, ideal para analizar los extensos comentarios de los pasajeros. Docker facilita escalar el clúster según sea necesario y asegura que cada componente opere de manera aislada, minimizando conflictos de dependencias.

Además, Jupyter Notebook con PySpark nos proporciona una plataforma interactiva para el análisis de datos, permitiendo escribir y ejecutar código y visualizar resultados en tiempo real. Esta configuración es intuitiva y ampliamente utilizada, lo que facilita la colaboración.

Finalmente, Docker asegura la portabilidad y reproducibilidad del entorno, garantizando que las configuraciones sean consistentes en todas las plataformas y eliminando problemas de compatibilidad. Esta combinación nos permite manejar datos de manera eficiente y mejorar nuestra capacidad de análisis y desarrollo colaborativo.



```
(base) C:\Users\acbon>Docker ps
CONTAINER ID   IMAGE                                      COMMAND                  CREATED      STATUS                  PORTS                                      NAMES
dfb4a93ed231   jupyter/pyspark-notebook:x86_64-spark-3.5.0   "tini -g -- start-no…"   5 days ago   Up 4 hours (healthy)   4040/tcp, 0.0.0.0:8888->8888/tcp   spark-docker

(base) C:\Users\acbon>
```

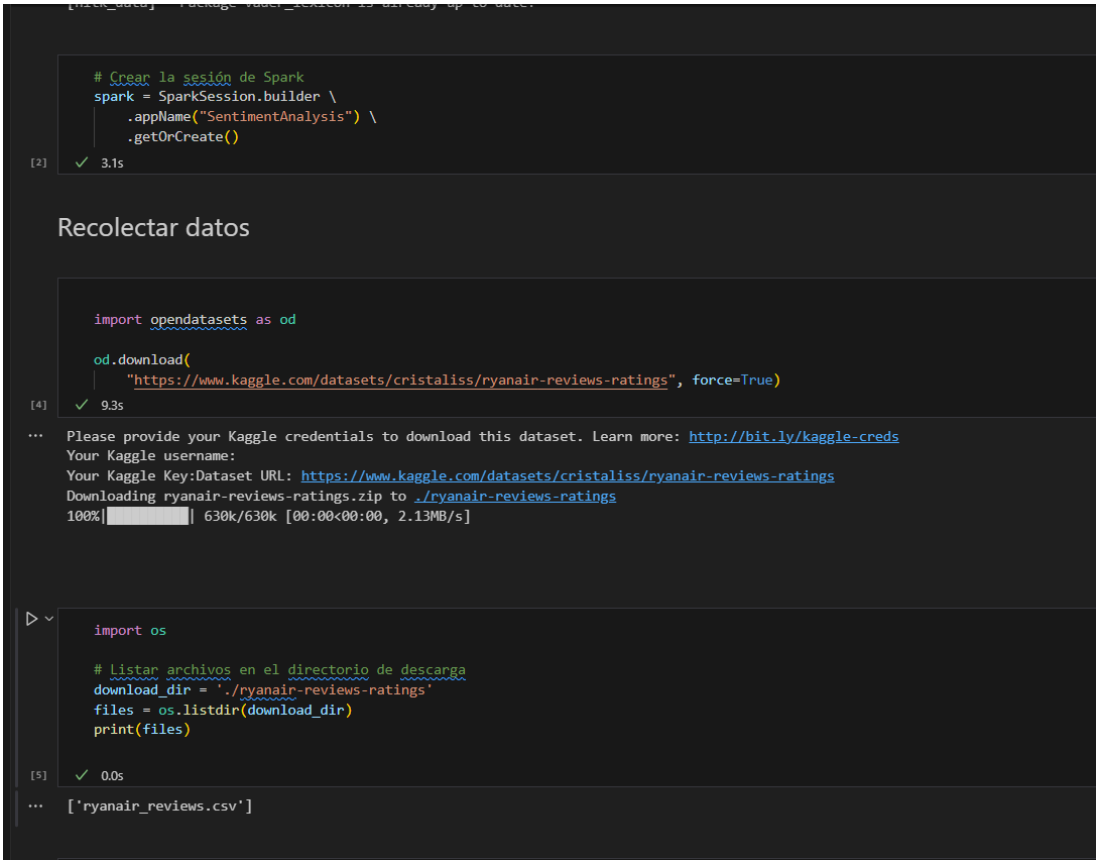
Preparación de Datos

Cargar el dataset

Después de importar todas las librerías necesarias, hemos creado una sesión de Spark, lo cual es fundamental para trabajar con grandes volúmenes de datos de manera distribuida. Este paso se realiza con el comando **SparkSession.builder** y se le asigna el nombre "SentimentAnalysis" para identificar nuestra tarea de análisis de sentimientos.

A continuación, procedimos a recolectar los datos necesarios para el análisis. Utilizamos la biblioteca **opendatasets** para descargar un conjunto de datos de comentarios de pasajeros de la aerolínea Ryanair directamente desde la plataforma Kaggle. Para ello, proporcionamos el enlace del dataset y ejecutamos el comando de descarga, que solicita las credenciales de Kaggle para acceder a los datos.

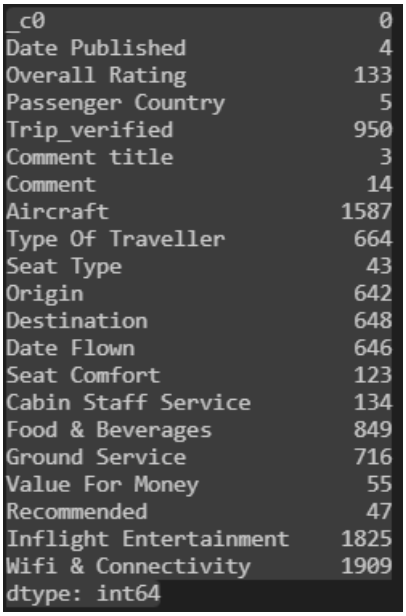
Una vez descargado el archivo zip que contiene los datos, lo descomprimimos y verificamos que el archivo específico, "ryanair_reviews.csv", esté presente en el directorio de descarga. Utilizamos la biblioteca **os** para listar los archivos en el directorio y confirmar que el archivo necesario está disponible para el análisis posterior. Este proceso asegura que los datos estén preparados y listos para ser utilizados en el análisis de sentimientos de los comentarios de los pasajeros



Limpieza y preprocesamiento de datos:

Manejo de valores nulos.

Para cumplir con los objetivos del proyecto, se implementaron diversas estrategias de limpieza de datos. Tras verificar los valores nulos y los datos repetidos, como se muestra en las imágenes adjuntas, se procedió de la siguiente manera:



En primer lugar, se eliminaron las columnas que no proporcionaban información relevante para el análisis de sentimientos y la correlación con la valoración global, el tipo de viajero y la clase de servicio. Esta acción simplificó el conjunto de datos y mejoró la eficiencia de los modelos de aprendizaje automático. Además, se transformaron las columnas a tipo numérico, lo que facilitó tanto los cálculos estadísticos como el entrenamiento de los modelos de machine learning.

En cuanto al manejo de valores nulos, se eliminaron las filas sin comentarios, ya que estos son fundamentales para el análisis de sentimientos y su ausencia no aportaba valor. Similarmente, se eliminaron las filas que carecían de fecha de publicación, dado que esta información es crucial para identificar tendencias a lo largo del tiempo. Para la imputación de valores nulos en la valoración general ('Overall Rating'), se utilizó la mediana, debido a que es menos sensible a los valores atípicos en comparación con la media, proporcionando así una imputación más robusta.

Por otro lado, los valores nulos en varias columnas se reemplazaron con valores específicos como 'Unknown' o 'No', ayudando a conservar las filas en el conjunto de datos mientras se mantenían datos significativos para el análisis. En las columnas de servicio, la imputación de valores nulos se realizó utilizando la media, lo que preservó la tendencia general de los datos sin introducir sesgos significativos. Adicionalmente, la columna 'Recommended' fue limpiada de espacios en blanco y sus valores fueron convertidos a formato numérico, facilitando el análisis y el modelado.

Estas estrategias de limpieza aseguraron que el conjunto de datos fuera de alta calidad, consistente y adecuado para el análisis de sentimientos y el desarrollo de modelos de machine learning, cumpliendo así con los objetivos del proyecto.

Normalización de texto y filtrado de valores anómalos

La normalización y el filtrado de valores anómalos en nuestro conjunto de datos fueron pasos esenciales para garantizar la calidad y la coherencia de los datos utilizados en nuestro análisis de sentimientos. Dada la naturaleza del dataset, se identificaron varios valores anómalos y caracteres que no aportaban información útil para los objetivos del proyecto.

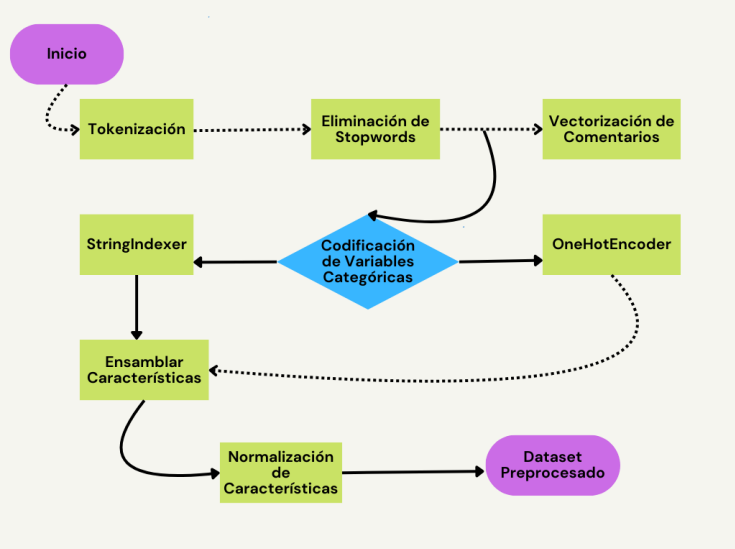
Para comenzar, se normalizó la columna 'Trip_verified' para unificar las distintas formas de indicar si un viaje estaba verificado o no. Se reemplazaron las variantes 'Trip Verified' y 'Verified Review' por 'Verified', y las variantes 'Not Verified', 'NotVerified', 'Unverified' y 'No' por 'Not Verified'. Esta normalización simplifica el análisis al reducir la variabilidad en la nomenclatura de esta columna.

Posteriormente, se filtraron los valores incorrectos en 'Type Of Traveller'. Solo se consideraron los tipos válidos de viajeros como 'Solo Leisure', 'Couple Leisure', 'Family Leisure', 'Business Traveller' y 'Solo Business'. De manera similar, se aseguraron los valores válidos en la columna 'Seat Type', manteniendo únicamente las categorías 'Economy Class', 'Premium Economy', 'Business Class' y 'First Class'. Este filtrado es crucial para evitar la distorsión de los resultados debido a entradas inválidas o mal categorizadas.

Además, para las columnas 'Origin' y 'Destination', se obtuvieron ubicaciones válidas y se filtraron las filas para asegurarse de que ambos campos contuvieran solo ubicaciones reconocidas y coherentes. También se eliminaron las filas con la fecha de vuelo 'Unknown Date', dado que estas no aportan información útil para el análisis temporal de los datos.

Finalmente, se corrigieron los valores en la columna 'Recommended', asegurando que cualquier valor mayor a 1 se ajustara a 1. Esto garantiza que los valores en esta columna sean binarios (1 o 0), lo cual es esencial para los modelos de machine learning que se utilizarán en el análisis.

Preprocesamiento de Texto y Codificación



Para realizar un análisis de sentimientos en comentarios de pasajeros de una aerolínea usando técnicas de Procesamiento de Lenguaje Natural (NLP) y modelos de Machine Learning, es esencial seguir un proceso adecuado de preprocesamiento de texto y codificación de datos. Este proceso asegura que los datos estén preparados de la mejor manera para ser utilizados por modelos como Random Forest, Logistic Regression y Naïve Bayes.

Comenzamos con la tokenización, que descompone los comentarios en palabras individuales. Luego, eliminamos las stopwords, palabras comunes que no aportan mucho al análisis y pueden añadir ruido a los modelos. Después, utilizamos un vectorizador de conteo (CountVectorizer) para transformar el texto en vectores de características basados en la frecuencia de las palabras, facilitando el procesamiento por los algoritmos de machine learning.

Además del texto, debemos codificar las variables categóricas como 'Passenger Country', 'Trip_verified', 'Type Of Traveller', 'Seat Type', 'Origin', y 'Destination'. Esto se hace en dos pasos: primero, se utilizan StringIndexer para convertir las categorías en índices numéricos, y luego, OneHotEncoder para transformar estos índices en vectores binarios. Este método permite que los modelos de machine learning manejen estas variables categóricas de manera efectiva.

Para unificar todas las características en un solo vector, usamos un ensamblador de vectores (VectorAssembler). Este paso combina las características textuales y categóricas en una única representación, facilitando su uso por los modelos de machine learning. Posteriormente, aplicamos una normalización de características con StandardScaler para asegurar que todas las características contribuyan de manera equitativa al modelo.

Finalmente, encapsulamos todas estas etapas en un pipeline, que ajusta y transforma los datos originales en un formato listo para el modelado. Esto asegura la reproducibilidad y facilita la actualización del flujo de trabajo.Este enfoque integral al preprocesamiento de texto y codificación de datos prepara adecuadamente el conjunto de datos para el análisis de sentimientos, permitiendo clasificar los comentarios en sentimientos positivos, negativos y neutros de manera precisa y eficiente.

Análisis de Sentimientos

Por medio del SentimentIntensityAnalyzer de VADER, se asignaron puntuaciones de sentimiento a cada comentario, clasificándolos en positivos, negativos o neutrales.

Los comentarios muestran una amplia gama de puntuaciones de sentimiento, reflejando tanto experiencias muy positivas como negativas. Por ejemplo, un comentario con una puntuación de 0.9543 fue clasificado como positivo, elogiando la puntualidad y el buen servicio del personal de cabina. En contraste, otro comentario con una puntuación de -0.8947 fue clasificado como negativo, mencionando cargos inesperados y problemas con el check-in.

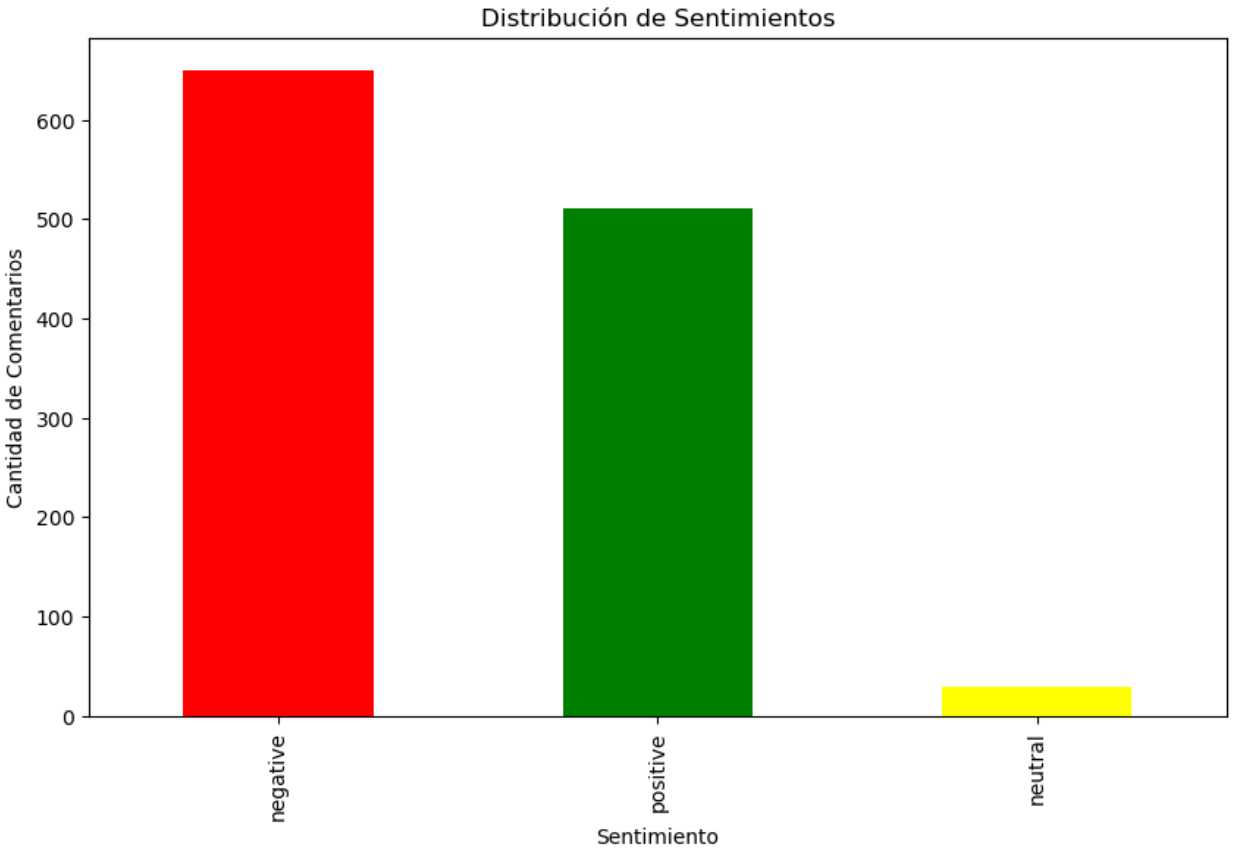
Clasificación de los comentarios:

Comentario	Puntuación de Sentimiento	Sentimiento
Flew back from Faro to London Luton Friday 2nd February. Ryanair in both directions was bang on time and smooth flights in both directions. We always sit in Front for more space and this was very comfortable for just under a 3 hour flight. The cabin crew were polite and efficient with nice sense of humour especially and engagement especially Ethan and his female colleague at the front section. For their human touch [unlike sometimes the stand offish BA crews] merit a 10/10 marking	0.9543	positive
Another good affordable flight with Ryanair. On time, pleasant staff at check-in and on board. We use Ryanair as our first choice on every flight we take.	0.7351	positive
Really impressed! You get what you pay for, this flight only cost £19.99. The seats were soft, and there was tons of legroom! (not in an emergency exit) The cabin was spotless. The colours were a little bright but it was good. Cheap and no frills! Highly recommend, flies almost everywhere.	0.8232	positive
I should like to review my flight from Faro to Liverpool with Ryanair. I booked the seat with Ryanair several months before my planned travel and found the website clear and easy to use. Ryanair sent quite frequent emails relating to the flight and attempted to sell me extras. The check-in procedure was followed by a bag-drop at Faro Airport. This was not efficient and the app they recommended for speedy bag-drop did not work. Setting that aside, I was able to check-in my bag within less than 30 minutes of having arrived at the airport. Boarding was supposedly via group but this did not materialise at all. The cabin crew had a reasonable presence and delivered a serious safety procedure. During the flight the crew worked hard and had several rounds of drinks for sale. I found the seat suitable for a flight just under 3 hours. The flight departed on-time and arrived on schedule too. Disembarkation was delayed in Liverpool while this was not the fault of Ryanair, there was no communication with the passengers as to why this was the case.	0.5605	positive

Disembarkation was not orderly and was poorly managed. Overall, a decent offering from Ryanair.		
Flight left the gate ahead of schedule, fare was really cheap and cabin crew were welcoming and friendly. Flight was so much cheaper than Aer Lingus and flying with Ryanair is a much better experience.	0.8402	positive
Booked a fight from Copenhagen to Poland though booking.com Somewhere in the email from booking.com it states that checkin must be done online from home. I figure I'll do it in the morning since I have plenty of time. it's low season and I don't have any checked in luggage. I live in Copenhagen not far from the airport. Morning comes and I try to check in online. No luck, apparently they close online checkin 3 hours before the flight. I figure I'll just check in at the airport at those self serve terminals. I arrive at the airport 2 hours before my flight and head for a terminal. Prompt says: kindly go to service desk. Head for service desk. They charge me 42 euro to check me in. The flight itself was 57 euro. They almost charged me the price of the flight just to check me in manually! And they forced me to check in manually by closing online and self-serve checkin. This is Ryanair policy. Apparently. Wizzair and booking are not without blame either, as they decide who they want to do business with.	-0.608	negative
Staff is rude and has no manners, let alone be professional. I had a backpack with which I flew from Vienna to Paris through Ryanair without any issues. But while flying from Paris to Barcelona, they asked me to pay for the backpack. They have kept a small rack to check size of your carryon baggage, which is way smaller than the one we have in flight and as a result, high chances you will always end up paying. Because of this there was a long queue and one couple even missed their flight and instead of being considerate, they smiled and asked them to purchase ticket for another flight.	0.6486	positive
I wanted to check in online a night before our flight, but my nationality was not listed in their websites and when we wanted to check in in person they asked for a fee, and that still was not a problem! They asked for a credit card to check for late check-in fee, we asked them to pay by our debit card or by cash, but they did not accept it. So we missed our flight, and they made us pay a No show fee on the application so that they give us another flight in the evening. The flight we got was about 30 Euro and the no show fee we paid was 100 Euro. When I told them in Germany we can pay by debit card or cash, they said we are an Irish airline it has nothing to do with Germany! Then why do you have your business here? The other thing is when I mentioned my nationality was not listed on your website they did not accept that and they inserted my nationality manually!	-0.8947	negative
This airline charges you for almost every thing. Every child needs to be seated with an adult. Got it, it is for Safety. But you must pay for the seat selection of the accompanying adult so that you can choose the seat next to the child. Only allowed a bag pack to be carried on for free, ay other cabin bag you must pay. You pay if you don’t check in online. It’s €55 to check in at the airport. That’s more than twice of my ticket price. No room for any negotiation. By the way, even if the flight is delayed and you have time to do online check in for free, too bad, just pay. They insisted that we have an email to inform us of this. Too bad if you ignored the email.	-0.8945	negative
Couldn’t find any reason to complain. Outbound flight on time, clean aircraft and Cabin Crew friendly and well presented. Great value for money. It was cheaper to fly to Dublin for the day on Ryanair rather than go to London from Bristol for the day by rail. Although the seats do not recline, I thought the leg room was fine. I have flown Ryanair several times over the years and have come to the conclusion that as long as you stick to the rules with regards checking in on line and adhere to luggage sizes, journeys are hassle free generally. Flight was late on the return leg but they emailed an apology before the dedicated check in time. That didn’t bother me personally. Their website is designed to capture more money related to seat allocation, earlier checkin, luggage upgrades, insurance, etc etc but they are optional and as long as you know what you want, there is no obligation to increase the price of your journey. I find their website and online check in very easy to use. I’m a pensioner and I’m grateful Ryanair offers fares at great value.	0.9824	positive
Play the game and you'll be fine. Got a return flight from Leeds Bradford to Faro for £101. I booked 3 months ahead, didn't reserve a seat, and took only a small bag. Had my boarding pass both on my app, and in paper format (just in case). Took my own food and drink on the plane, didn't buy any of theirs, and didn't buy any of their duty free etc.	0.6452	positive

It really does pay to do your research. At Leeds Bradford someone had lost their boarding pass (I think that cost them £60), while at Faro someone's cabin bag was too big. That had to go in the hold, and I expect that cost him. He said it was OK at Leeds, but maybe he got away with it there. My old bag would have been too big, but I checked Ryanair's website, and bought a new and conforming bag for £9.99. I'm 6 foot tall, and the leg room is tight, but just about OK for a 3 hour flight. The seats were uncomfortable though and needed more cushioning. The outward flight took off and arrived on time. The cabin crew were friendly and efficient. No hard sales pitch for food, drink, duty free. The announcements could have been clearer, but that was down to the Portuguese captain. The return flight was about an hour late taking off, due to delays experienced earlier in the day. This happens. Allow for it. They made up 20 minutes and the plane landed at Leeds Bradford 40 minutes behind schedule. Absolutely no complaints whatsoever and fantastic value for money. I'd reckon that the negative views on here are hugely in the minority. In some cases things have gone wrong, but in others part of the blame must be with the posters who haven't researched and prepared well.		

Análisis de la Distribución de los Sentimientos



El gráfico de distribución de sentimientos revela una clara predominancia de comentarios negativos entre los pasajeros de la aerolínea, seguidos por una cantidad significativa de comentarios positivos y una minoría de comentarios neutrales. Este análisis nos proporciona información valiosa sobre la percepción de los clientes y destaca áreas críticas para la mejora del servicio.

La barra roja, que representa los comentarios negativos, es la más alta, indicando que muchos pasajeros han tenido experiencias insatisfactorias. Estos comentarios son cruciales porque reflejan aspectos específicos del servicio que necesitan atención. Por ejemplo, problemas recurrentes mencionados en los comentarios negativos pueden incluir retrasos en los vuelos, mal servicio al cliente, cargos adicionales inesperados y problemas con el equipaje. Identificar estos puntos críticos permite a la aerolínea tomar medidas correctivas para mejorar la satisfacción del cliente.

La barra verde, correspondiente a los comentarios positivos, aunque menor que la de los negativos, sigue siendo significativa. Los comentarios positivos suelen destacar aspectos del servicio que funcionan bien, como la puntualidad de los vuelos, la amabilidad del personal y la buena relación calidad-precio. Estos aspectos positivos son fortalezas que la aerolínea puede seguir promoviendo para atraer y retener a los clientes.

La barra amarilla, que representa los comentarios neutrales, es la más baja. Los comentarios neutrales reflejan experiencias que no fueron buenas ni malas. Aunque representan una minoría, estos comentarios pueden proporcionar información sobre áreas donde la experiencia del cliente es inconsistente o donde se pueden realizar mejoras para convertir una experiencia neutral en una positiva.

La alta proporción de comentarios negativos sugiere la necesidad de un análisis detallado para identificar los problemas más comunes que enfrentan los pasajeros, como demoras, cargos adicionales y problemas con el servicio al cliente. Abordar estos problemas puede mejorar significativamente la percepción general del servicio.

Por otro lado, los comentarios positivos indican áreas donde la aerolínea ya está cumpliendo o superando las expectativas de los pasajeros. Es importante mantener y fortalecer estos aspectos, como la puntualidad y el buen trato del personal, para seguir generando experiencias satisfactorias.

Los comentarios neutrales ofrecen una oportunidad para convertir experiencias mediocres en experiencias positivas. Analizar estos comentarios puede ayudar a identificar áreas donde se pueden hacer ajustes menores para mejorar la satisfacción del cliente.

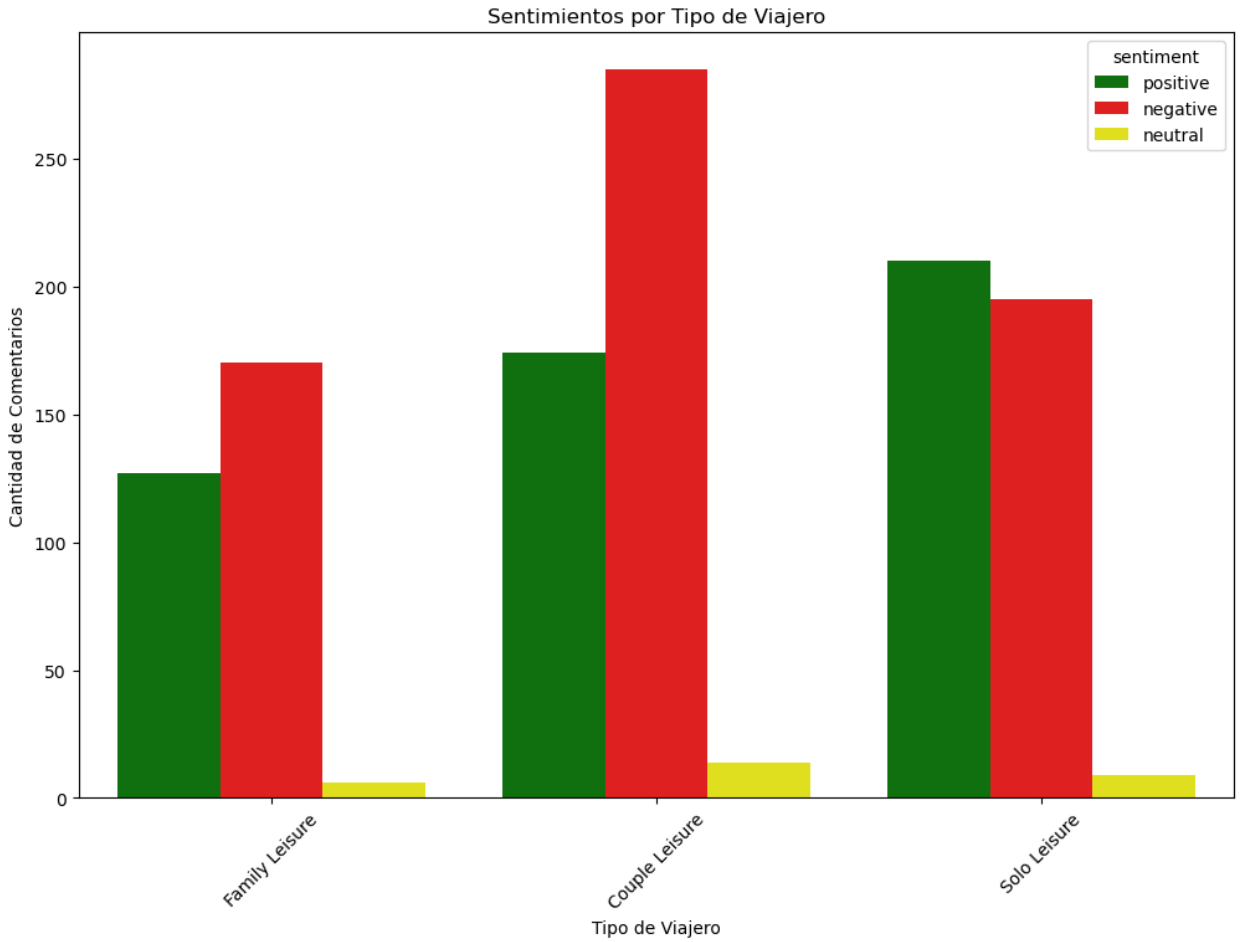
Además, mejorar la comunicación con los pasajeros sobre políticas y procedimientos puede reducir la frustración y los comentarios negativos relacionados con malentendidos o falta de información. Esto incluye claridad en los cargos adicionales y procesos de check-in.

Finalmente, la aerolínea debe implementar medidas correctivas basadas en los comentarios negativos, monitorear su efectividad y hacer ajustes según sea necesario. Esto puede incluir

capacitación adicional del personal, mejoras en la gestión del tiempo y revisiones de políticas para hacerlas más amigables para el cliente.

Correlación de Sentimientos con Otras Variables

Sentimientos por Tipo de Viajero



El gráfico de distribución de sentimientos por tipo de viajero muestra tendencias claras entre los diferentes grupos de pasajeros. Los comentarios de las familias que viajan por ocio presentan una mayor cantidad de opiniones negativas, con aproximadamente 190 comentarios negativos frente a 130 positivos, y muy pocos comentarios neutrales. Esto sugiere que las familias enfrentan problemas específicos que afectan su experiencia de viaje, posiblemente relacionados con la gestión de asientos familiares, servicios a bordo para niños o procesos de embarque.

En el caso de las parejas que viajan por ocio, la tendencia también es hacia comentarios negativos, aunque la diferencia con los positivos es menor en comparación con las familias. Hay alrededor de 250 comentarios negativos y 200 positivos, con pocos neutrales. Los problemas pueden ser

similares a los enfrentados por las familias, aunque probablemente con menos complicaciones relacionadas con los niños. La aerolínea debería considerar mejorar áreas como la comodidad de los asientos, la privacidad y el servicio al cliente para este grupo.

Los viajeros solitarios, por otro lado, muestran una proporción mayor de comentarios positivos, con aproximadamente 220 positivos y 180 negativos, y pocos neutrales. Esto sugiere que los servicios y la experiencia general son más adecuados para individuos que viajan solos. Sin embargo, la cantidad de comentarios negativos es considerable y no debe ignorarse.

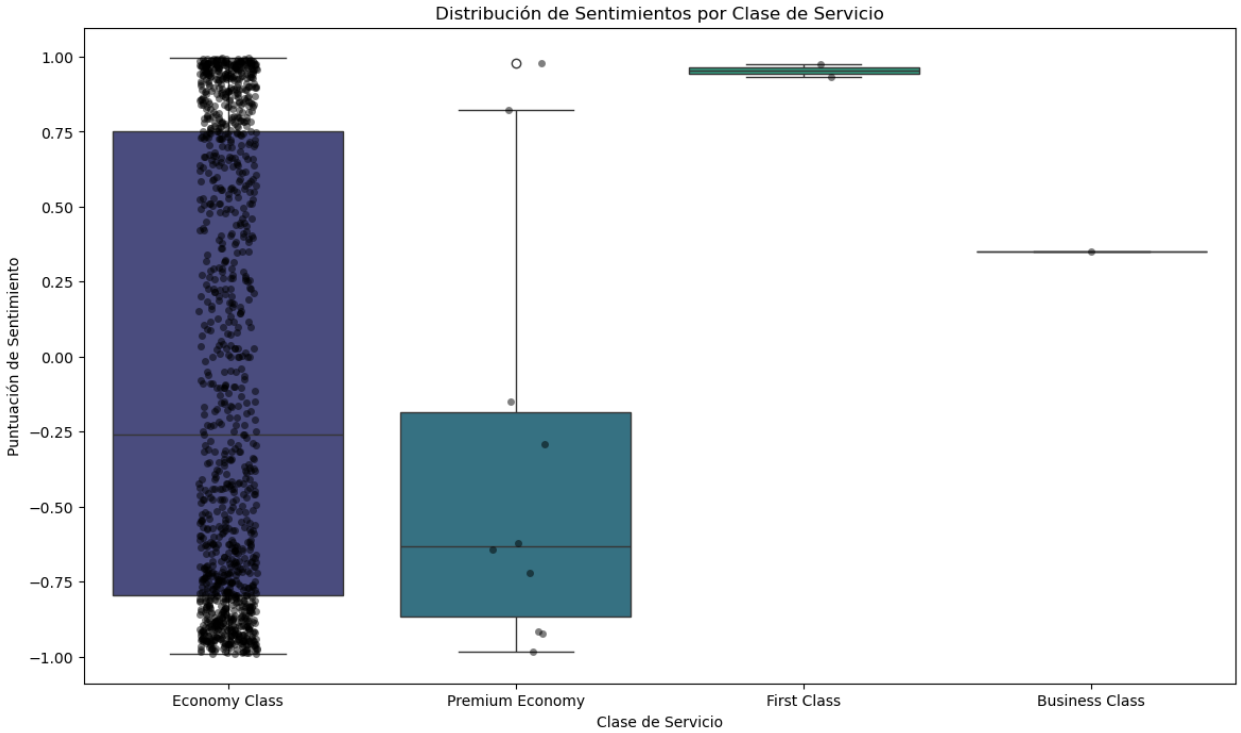
Para mejorar la satisfacción de los clientes, la aerolínea puede tomar medidas específicas para cada grupo de viajeros. Para las familias, se podrían mejorar la asignación de asientos juntos, el entretenimiento a bordo para niños y el proceso de embarque. Para las parejas, garantizar asientos juntos y ofrecer opciones de privacidad adicionales podría mejorar su experiencia. Para los viajeros solitarios, es importante continuar reforzando los aspectos positivos y abordar cualquier problema persistente.

La aerolínea debe realizar un análisis detallado de los comentarios negativos para identificar problemas recurrentes y específicos de cada grupo de viajeros. Esto permitirá implementar soluciones dirigidas y mejorar la satisfacción general. Además, mejorar la comunicación sobre políticas y servicios específicos para cada grupo puede reducir la insatisfacción. Por ejemplo, informando claramente sobre los procesos de embarque para familias y parejas, y destacando los beneficios para los viajeros solitarios.

Los comentarios negativos reflejan en gran medida problemas con el servicio al cliente. Capacitar al personal para manejar mejor las necesidades y expectativas de diferentes tipos de viajeros podría mejorar significativamente la percepción del servicio.

En definitiva, el gráfico proporciona una comprensión clara de cómo diferentes grupos de viajeros perciben los servicios de la aerolínea. Al enfocarse en las áreas problemáticas específicas para cada grupo, la aerolínea puede mejorar significativamente la satisfacción del cliente. Al mismo tiempo, reforzar las áreas positivas ayudará a mantener y atraer a más pasajeros. Implementar estas estrategias basadas en los insights obtenidos del análisis de sentimientos permitirá a la aerolínea ofrecer una experiencia de viaje más personalizada y satisfactoria.

Sentimientos por Clase de Servicio



El gráfico de caja y bigotes (boxplot) muestra cómo las puntuaciones de sentimiento varían según la clase de servicio: Economy, Premium Economy, First Class y Business Class. Esta visualización es útil para comprender cómo cada clase influye en la percepción de los pasajeros.

En Economy Class, los comentarios tienen una amplia dispersión en las puntuaciones de sentimiento, que van desde muy negativos hasta muy positivos. El rango intercuartil (IQR) es amplio, indicando una variabilidad considerable en las experiencias de los pasajeros. La alta densidad de puntos alrededor del valor cero sugiere que muchos pasajeros tienen sentimientos mixtos o neutros sobre esta clase de servicio.

Para Premium Economy, los comentarios muestran una tendencia más negativa en comparación con Economy Class. Aunque el IQR también es amplio, la mediana está por debajo de cero, indicando una tendencia general hacia sentimientos más negativos. La presencia de varios outliers, tanto positivos como negativos, sugiere experiencias muy variadas y extremas en esta clase.

En First Class, hay pocos datos disponibles, lo que podría indicar una menor cantidad de comentarios o menos pasajeros en esta clase. Los comentarios disponibles son relativamente positivos, con una puntuación de sentimiento centrada alrededor de la mediana positiva. La falta de dispersión significativa sugiere una experiencia más consistente y generalmente satisfactoria.

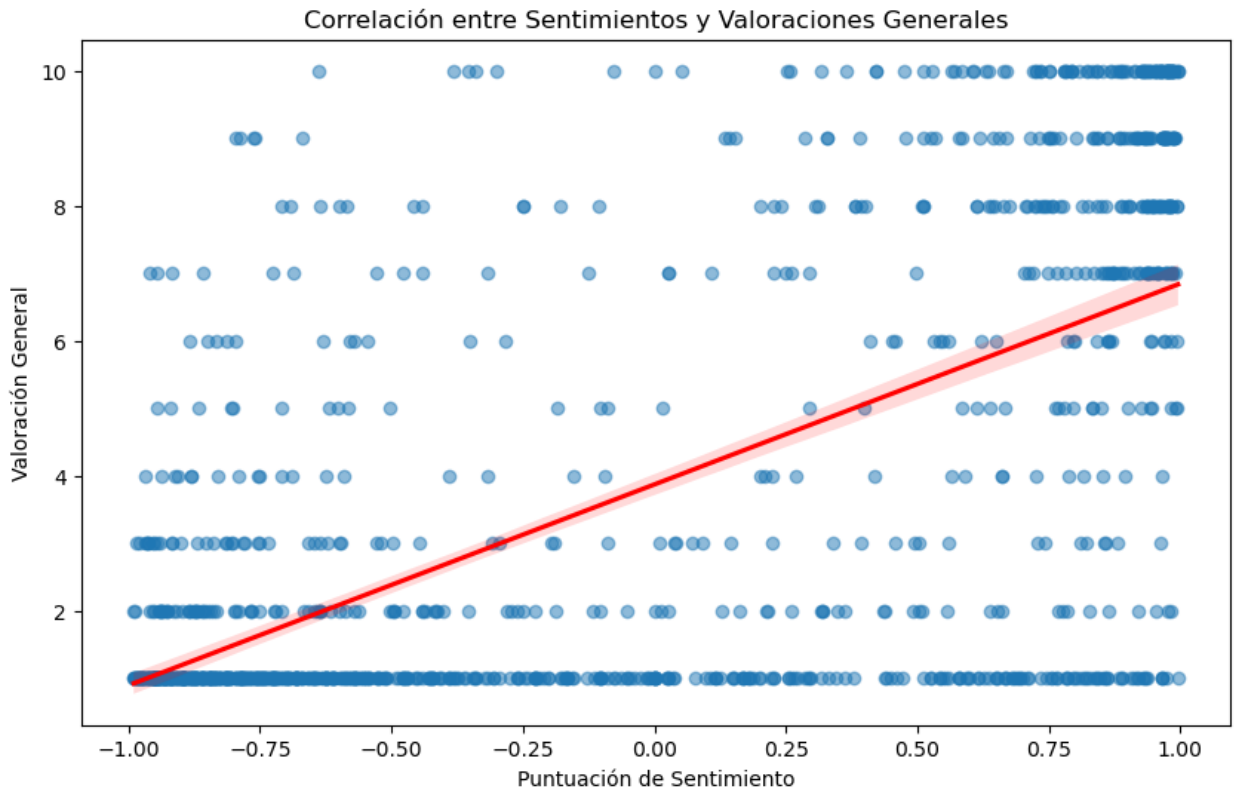
Por su parte, Business Class muestra una distribución de puntuaciones de sentimiento bastante positiva. Aunque existe una ligera dispersión, en general, los comentarios indican una alta satisfacción entre los pasajeros de Business Class.

La amplia dispersión de los comentarios en Economy Class sugiere que las experiencias son muy variadas. La aerolínea debería investigar más a fondo para identificar y abordar los factores que causan insatisfacción. Implementar medidas para estandarizar y mejorar la consistencia del servicio en esta clase podría reducir la variabilidad y aumentar la satisfacción del cliente.

En Premium Economy, la tendencia hacia comentarios más negativos indica problemas específicos que necesitan ser abordados. Podrían ser necesarios ajustes en la oferta de servicios o en la relación calidad-precio. Analizar los outliers podría proporcionar información valiosa sobre problemas recurrentes o excepcionales que afectan significativamente la percepción del servicio.

First Class y Business Class muestran una puntuación de sentimiento positiva y consistente, sugiriendo que los servicios ofrecidos están cumpliendo o superando las expectativas de los pasajeros. Es importante continuar ofreciendo un servicio de alta calidad en estas clases y considerar la implementación de algunos de estos estándares en clases más económicas para mejorar la percepción general.

Correlación entre Sentimientos y Valoraciones Generales



El gráfico de dispersión revela la correlación entre la puntuación de sentimientos y las valoraciones generales de los pasajeros sobre la aerolínea. Este análisis nos permite identificar la relación entre las emociones de los pasajeros y las calificaciones que otorgan al servicio.

La tendencia lineal positiva del gráfico indica que, a medida que aumenta la puntuación de sentimiento (se vuelve más positiva), la valoración general también tiende a aumentar. Esta correlación sugiere que los sentimientos positivos se asocian con valoraciones altas, mientras que los negativos se correlacionan con valoraciones bajas. Aunque los puntos en el gráfico están dispersos a lo largo del eje de la puntuación de sentimientos, la tendencia general es clara. Se observa una mayor concentración de puntos en los extremos, indicando que los pasajeros con sentimientos muy positivos o negativos tienden a dar valoraciones más extremas.

Existen algunos outliers que no siguen la tendencia general, como valoraciones bajas con sentimientos positivos y viceversa. Esto podría deberse a experiencias específicas o a factores externos no relacionados directamente con el servicio de la aerolínea. Dado que los sentimientos positivos se correlacionan fuertemente con valoraciones altas, es crucial para la aerolínea fomentar experiencias que generen estos sentimientos. Esto podría incluir mejoras en el servicio al cliente, puntualidad, comodidad de los asientos y calidad de los servicios a bordo.

Los sentimientos negativos están claramente asociados con valoraciones bajas. La aerolínea debe abordar las causas subyacentes de estos sentimientos negativos para mejorar la percepción general del servicio. Esto puede implicar la revisión de políticas que causan insatisfacción, como cargos adicionales inesperados o procesos de check-in confusos. Los outliers en el gráfico ofrecen una oportunidad para entender mejor las excepciones a la tendencia general. Analizar estos casos puede proporcionar información valiosa sobre áreas específicas donde el servicio puede ser inconsistente.

La variabilidad en las puntuaciones de sentimiento y valoraciones generales sugiere que la experiencia del pasajero puede ser inconsistente. La aerolínea debe trabajar para estandarizar y mejorar la consistencia del servicio, asegurando que todos los pasajeros tengan una experiencia positiva. Es importante que la aerolínea continúe monitoreando las puntuaciones de sentimiento y las valoraciones generales a lo largo del tiempo. Esto permitirá identificar rápidamente cualquier cambio en la percepción del cliente y responder de manera proactiva.

El gráfico de correlación entre sentimientos y valoraciones generales proporciona una visión clara de la relación entre cómo se sienten los pasajeros y cómo valoran su experiencia. Los sentimientos positivos están fuertemente asociados con valoraciones altas, lo que destaca la importancia de crear experiencias de viaje que generen estos sentimientos. Al abordar los sentimientos negativos y mejorar la consistencia del servicio, la aerolínea puede aumentar significativamente la satisfacción del cliente y mejorar su reputación en el mercado.

Análisis de Varianza (ANOVA) y Matriz de Correlación

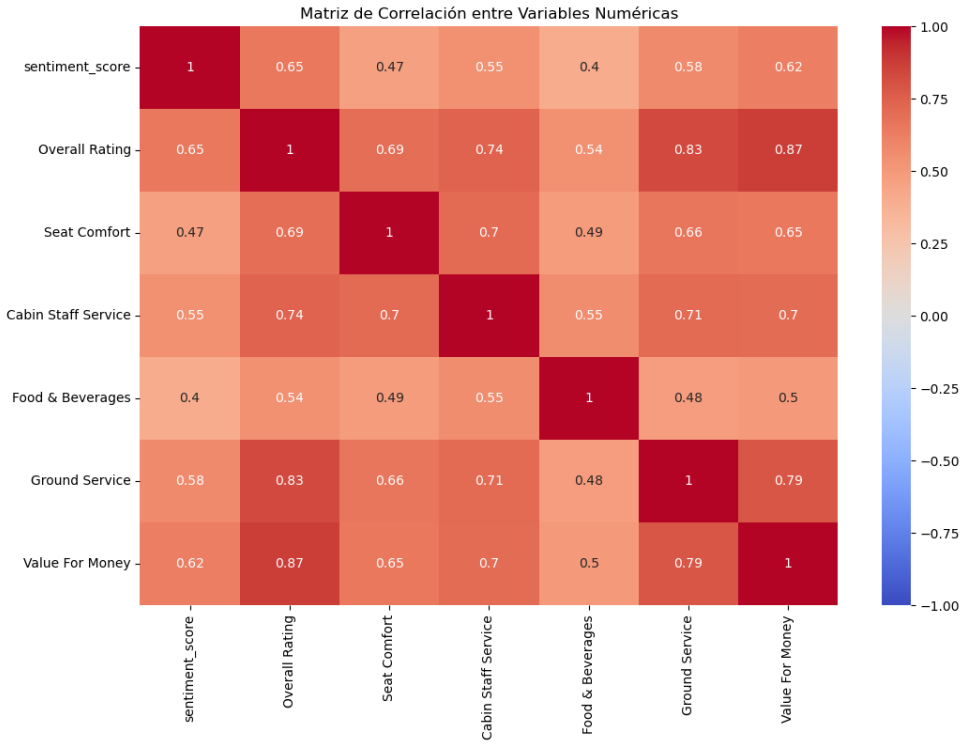
```python
# Asegurarse de que la columna 'sentiment_score' está en formato numérico
dataset = dataset.withColumn("sentiment_score", col("sentiment_score").cast("float"))

# Calcular la correlación entre el puntaje de sentimientos y las variables numéricas
numeric_columns = ['Overall Rating', 'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Value For Money']
for column in numeric_columns:
    correlation = dataset.stat.corr('sentiment_score', column)
    print(f"Correlation between sentiment_score and {column}: {correlation}")

# Ejecutar ANOVA para cada variable categórica
anova_results = {}
for column in ["Overall Rating", "Type Of Traveller", "Seat Type"]:
    anova_data = dataset.select(column, "sentiment_score")
    pandas_df = anova_data.toPandas()
    groups = [pandas_df[pandas_df[column] == val]["sentiment_score"].dropna() for val in pandas_df[column].unique()]
    anova_results[column] = stats.f_oneway(*groups)

# Mostrar los resultados del ANOVA
for column, result in anova_results.items():
    print(f"ANOVA result for {column}: F={result.statistic}, p-value={result.pvalue}")
```

```
Correlation between sentiment_score and Overall Rating: 0.6537098206488229
Correlation between sentiment_score and Seat Comfort: 0.46641779263793187
Correlation between sentiment_score and Cabin Staff Service: 0.5465088615909955
Correlation between sentiment_score and Food & Beverages: 0.40095227246767523
Correlation between sentiment_score and Ground Service: 0.5768188262585976
Correlation between sentiment_score and Value For Money: 0.6228901673332747
ANOVA result for Overall Rating: F=100.46377886633262, p-value=4.087288180229986e-139
ANOVA result for Type Of Traveller: F=7.639939958406461, p-value=0.000504882594799101
ANOVA result for Seat Type: F=2.122828392725131, p-value=0.09559650766716829
```



El análisis de los resultados obtenidos a través de ANOVA, correlaciones y la matriz de correlación nos proporciona una visión clara y detallada sobre cómo los diferentes aspectos del servicio de la aerolínea están relacionados con los sentimientos de los pasajeros.

Inicialmente, el análisis de correlaciones reveló información valiosa sobre las interrelaciones entre las diversas dimensiones del servicio y los sentimientos de los pasajeros. Se observó que la puntuación de sentimiento presenta una alta correlación con la valoración general (0.6537), lo que indica que los pasajeros con sentimientos positivos tienden a otorgar valoraciones generales más altas. También se identificó una fuerte correlación con la relación calidad-precio (0.6229) y el servicio en tierra (0.5768), sugiriendo que estos factores son críticos para una percepción positiva del servicio.

Los resultados del ANOVA respaldan y complementan estos hallazgos. La prueba de ANOVA para la valoración general mostró diferencias significativas en las puntuaciones de sentimiento, confirmando la fuerte relación entre estos dos factores. Además, se observó que el tipo de viajero presenta diferencias significativas, lo que sugiere que distintos grupos de pasajeros tienen experiencias y percepciones variadas. No obstante, el tipo de asiento no mostró diferencias significativas, lo que indica que la comodidad del asiento, aunque importante, no es un diferenciador clave en los sentimientos de los pasajeros.

La matriz de correlación proporciona una visualización detallada de las relaciones entre las variables numéricas, reafirmando varios de los hallazgos anteriores. La valoración general muestra fuertes correlaciones con la relación calidad-precio (0.87) y el servicio en tierra (0.83), sugiriendo que estos factores son fundamentales para la percepción general del servicio. Asimismo, la correlación con la puntuación de sentimiento es alta (0.65), consistente con los resultados del análisis de correlación inicial.

En cuanto a la comodidad del asiento, se observó una correlación significativa con la valoración general (0.69) y el servicio del personal de cabina (0.70), lo que indica que estos elementos están interrelacionados en la percepción del pasajero. El servicio del personal de cabina también presenta una alta correlación con la valoración general (0.74) y el servicio en tierra (0.71), subrayando la importancia del servicio al cliente en todos los aspectos del viaje. Por otro lado, la relación calidad-precio muestra una alta correlación con la valoración general (0.87) y el servicio en tierra (0.79), reforzando la idea de que la percepción del valor recibido es crítica para la satisfacción general.

En conclusión, los análisis combinados de correlación y ANOVA, junto con la matriz de correlación, proporcionan una comprensión robusta de los factores que influyen en los sentimientos de los pasajeros y sus valoraciones generales del servicio. La valoración general, la relación calidad-precio y el servicio en tierra emergen como factores clave que impactan significativamente los sentimientos de los pasajeros.

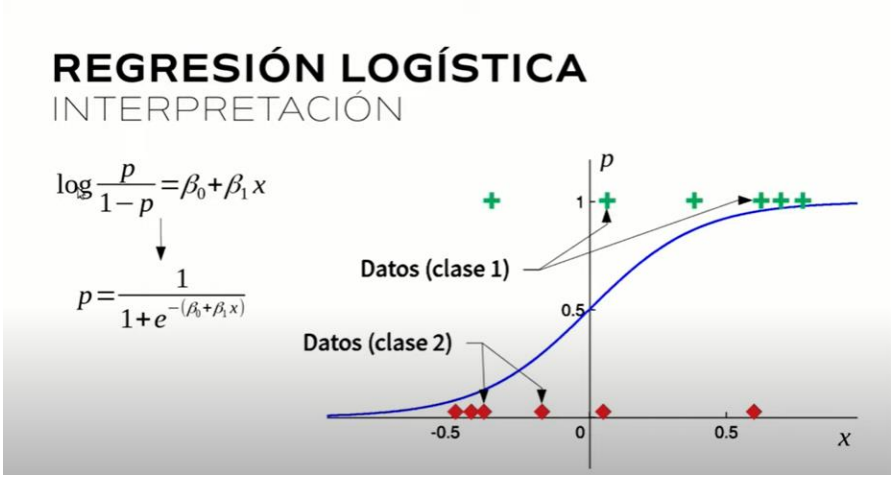
Para la aerolínea, estos hallazgos subrayan la importancia de mejorar estos aspectos críticos del servicio. Específicamente, se recomienda mejorar el servicio al cliente, asegurando que el personal

de cabina y de tierra esté bien capacitado y pueda ofrecer un servicio excelente y consistente. También es crucial optimizar la relación calidad-precio, revisando y ajustando las políticas de precios y servicios para asegurar que los pasajeros perciban que están recibiendo un buen valor por su dinero. Además, se debe estandarizar la comodidad y el servicio, implementando estándares consistentes de comodidad del asiento y servicio en todas las clases para mejorar la experiencia general del pasajero.

Algoritmos de Aprendizaje Automático

En este estudio, se han implementado tres algoritmos de clasificación con el objetivo de evaluar la precisión en la identificación de los sentimientos expresados en los comentarios de los pasajeros. Los algoritmos seleccionados para este análisis son la Regresión Logística, el Naive Bayes y los Bosques Aleatorios. A continuación, se ofrece una descripción general de cada clasificador, los resultados obtenidos y la justificación de su utilización en nuestro proyecto de análisis de sentimientos.

Regresión Logística (Logistic Regression)



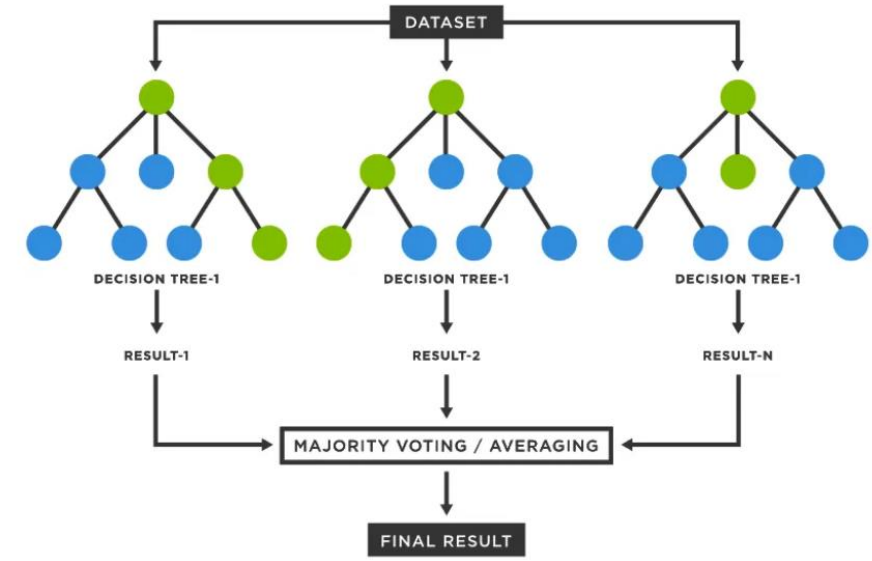
La Regresión Logística es un método de clasificación lineal que modela la probabilidad de que una observación pertenezca a una clase binaria, basándose en una combinación lineal de características independientes. Este clasificador es apreciado por su simplicidad y eficiencia, además de proporcionar probabilidades interpretables, lo que facilita la comprensión y explicación de los resultados obtenidos.

Naive Bayes

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

El algoritmo Naive Bayes se basa en el teorema de Bayes y asume la independencia de las características. Este enfoque calcula la probabilidad de que un ejemplo pertenezca a una clase determinada, considerando la presencia de ciertas características. Entre sus ventajas, destaca su rapidez y eficiencia, así como su buen desempeño con conjuntos de datos grandes y características independientes, lo que lo convierte en una opción robusta para el análisis de sentimientos.

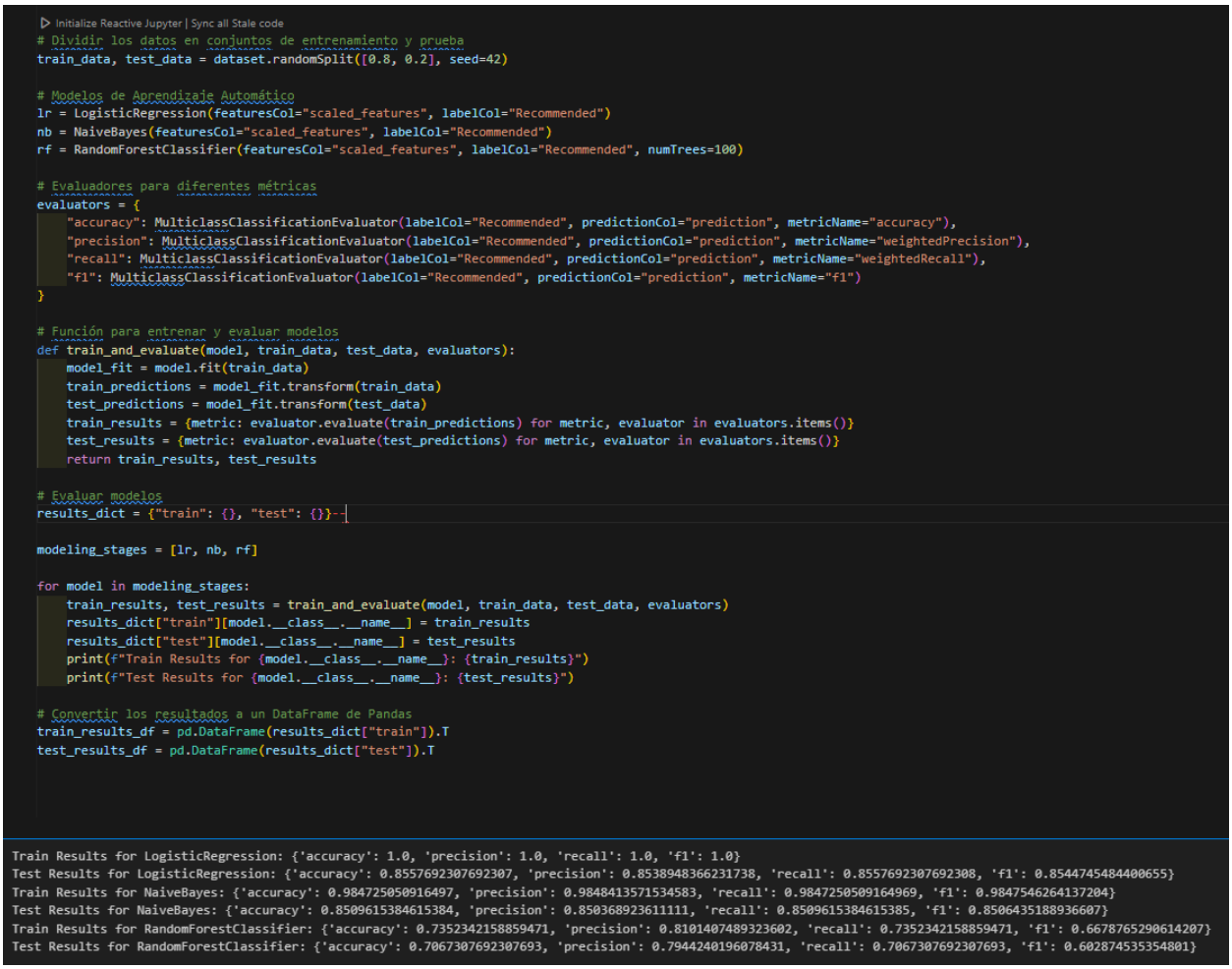
Bosques Aleatorios (Random Forest)



El método de Bosques Aleatorios utiliza múltiples árboles de decisión para mejorar la precisión del modelo y evitar el sobreajuste. Cada árbol en el bosque vota, y la clase con la mayoría de los votos se considera la predicción final del modelo. Este enfoque es robusto frente al sobreajuste y

maneja eficazmente grandes conjuntos de datos y características con correlaciones, proporcionando resultados consistentes y precisos en el análisis de sentimientos.

Implementación de Modelos de Clasificación



```
Initialize Reactive Jupyter | Sync all Stale code
# Dividir los datos en conjuntos de entrenamiento y prueba
train_data, test_data = dataset.randomSplit([0.8, 0.2], seed=42)

# Modelos de Aprendizaje Automático
lr = LogisticRegression(featuresCol="scaled_features", labelCol="Recommended")
nb = NaiveBayes(featuresCol="scaled_features", labelCol="Recommended")
rf = RandomForestClassifier(featuresCol="scaled_features", labelCol="Recommended", numTrees=100)

# Evaluadores para diferentes métricas
evaluators = {
    "accuracy": MulticlassClassificationEvaluator(labelCol="Recommended", predictionCol="prediction", metricName="accuracy"),
    "precision": MulticlassClassificationEvaluator(labelCol="Recommended", predictionCol="prediction", metricName="weightedPrecision"),
    "recall": MulticlassClassificationEvaluator(labelCol="Recommended", predictionCol="prediction", metricName="weightedRecall"),
    "f1": MulticlassClassificationEvaluator(labelCol="Recommended", predictionCol="prediction", metricName="f1")
}

# Función para entrenar y evaluar modelos
def train_and_evaluate(model, train_data, test_data, evaluators):
    model_fit = model.fit(train_data)
    train_predictions = model_fit.transform(train_data)
    test_predictions = model_fit.transform(test_data)
    train_results = {metric: evaluator.evaluate(train_predictions) for metric, evaluator in evaluators.items()}
    test_results = {metric: evaluator.evaluate(test_predictions) for metric, evaluator in evaluators.items()}
    return train_results, test_results

# Evaluar modelos
results_dict = {"train": {}, "test": {}}

modeling_stages = [lr, nb, rf]

for model in modeling_stages:
    train_results, test_results = train_and_evaluate(model, train_data, test_data, evaluators)
    results_dict["train"][model.__class__.__name__] = train_results
    results_dict["test"][model.__class__.__name__] = test_results
    print(f"Train Results for {model.__class__.__name__}: {train_results}")
    print(f"Test Results for {model.__class__.__name__}: {test_results}")

# Convertir los resultados a un DataFrame de Pandas
train_results_df = pd.DataFrame(results_dict["train"]).T
test_results_df = pd.DataFrame(results_dict["test"]).T
```

```
Train Results for LogisticRegression: {'accuracy': 1.0, 'precision': 1.0, 'recall': 1.0, 'f1': 1.0}
Test Results for LogisticRegression: {'accuracy': 0.8557692307692307, 'precision': 0.8538948366231738, 'recall': 0.8557692307692308, 'f1': 0.8544745484400655}
Train Results for NaiveBayes: {'accuracy': 0.984725050916497, 'precision': 0.9848413571534583, 'recall': 0.984725050916497, 'f1': 0.9847546264137204}
Test Results for NaiveBayes: {'accuracy': 0.8509615384615384, 'precision': 0.850368923611111, 'recall': 0.8509615384615385, 'f1': 0.8506435188936607}
Train Results for RandomForestClassifier: {'accuracy': 0.7352342158859471, 'precision': 0.8101407489323602, 'recall': 0.7352342158859471, 'f1': 0.6678765290614207}
Test Results for RandomForestClassifier: {'accuracy': 0.7067307692307693, 'precision': 0.7944240196078431, 'recall': 0.7067307692307693, 'f1': 0.602874535354801}
```

Para comenzar, dividimos los datos en conjuntos de entrenamiento y prueba utilizando un método de división aleatoria. Esto garantiza que el 80% de los datos se utilicen para entrenar los modelos y el 20% restante para evaluarlos. Esta distribución equitativa asegura una evaluación objetiva del rendimiento de los modelos.

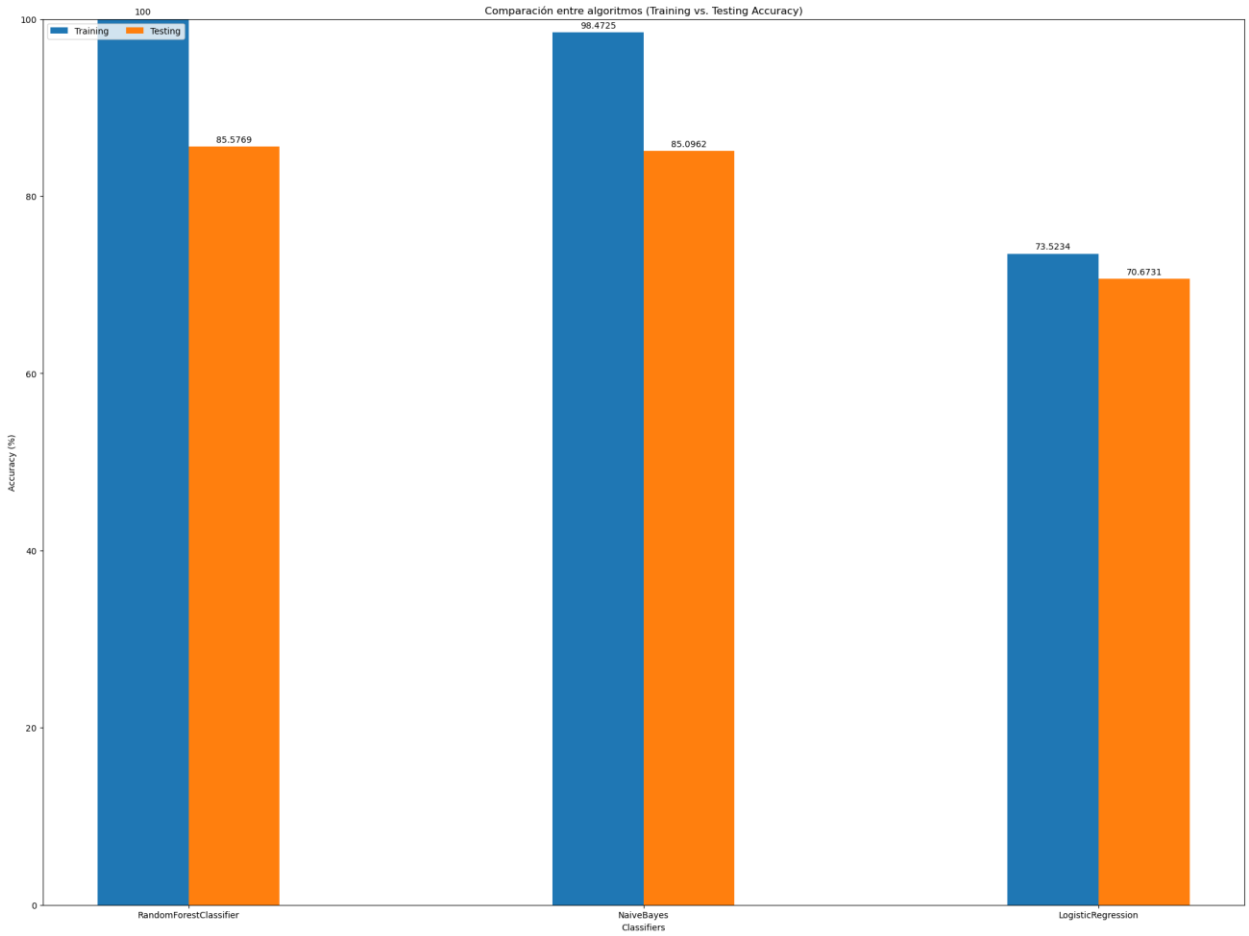
Para el análisis de sentimientos, hemos configurado tres modelos de clasificación diferentes: Regresión Logística, Naive Bayes y Bosques Aleatorios (Random Forest). Cada uno de estos modelos tiene sus propias ventajas y se adapta bien a diferentes tipos de datos y contextos. Hemos configurado varios evaluadores para medir el rendimiento de los modelos. Estos evaluadores miden métricas clave como la precisión (accuracy), la precisión ponderada (weighted precision), la exhaustividad ponderada (weighted recall) y la puntuación F1 ponderada (weighted F1 score).

Estas métricas nos proporcionan una visión completa de cómo se desempeñan los modelos en diferentes aspectos.

Definimos una función que se encarga de entrenar cada modelo y evaluar su rendimiento tanto en los conjuntos de datos de entrenamiento como en los de prueba. Esta función es crucial para asegurarnos de que los modelos están bien ajustados y que sus predicciones son precisas y fiables. Entrenamos y evaluamos cada uno de los modelos, almacenando los resultados en un diccionario para su posterior análisis. Este enfoque sistemático nos permite comparar fácilmente el rendimiento de los diferentes modelos y seleccionar el más adecuado para nuestro propósito.

Finalmente, convertimos los resultados en un DataFrame de Pandas para facilitar su análisis. Este formato nos permite manipular y visualizar los datos de manera más efectiva, proporcionando una base sólida para nuestras conclusiones. Este enfoque integral nos permite evaluar de manera sistemática y eficiente el rendimiento de diferentes modelos de clasificación. A través de este proceso, aseguramos una comprensión profunda y objetiva de las capacidades y limitaciones de cada modelo en el análisis de sentimientos de las opiniones de los pasajeros. Esta metodología no solo optimiza el proceso de análisis, sino que también mejora la calidad de las decisiones basadas en los datos obtenidos.

Resultados obtenidos del Modelado



El análisis de los resultados obtenidos de los modelos de clasificación y el gráfico adjunto nos permite determinar cuál es el mejor modelo para desarrollar nuestro proyecto de análisis de sentimientos en opiniones de pasajeros de una aerolínea, utilizando NLP y PySpark en un entorno Big Data.

Los tres modelos de clasificación configurados son Regresión Logística, Naive Bayes y Random Forest. Cada uno de estos modelos se ajustó utilizando las características escaladas y la etiqueta "Recommended" como variables de entrada y salida, respectivamente. Posteriormente, se evaluaron los modelos utilizando varias métricas, incluyendo precisión (accuracy), precisión ponderada, recall ponderado y F1 ponderado, para obtener una visión completa de su rendimiento.

El gráfico muestra las precisiones de entrenamiento y prueba para cada modelo. La Regresión Logística alcanzó una precisión del 100% en el conjunto de entrenamiento, lo que sugiere un posible sobreajuste, ya que su precisión disminuyó al 85.58% en el conjunto de prueba. Aunque el rendimiento en el conjunto de prueba es bueno, el sobreajuste indica que el modelo puede no generalizar bien en nuevos datos. En cuanto a otras métricas, la precisión, recall y F1 son todas del 85%, lo que demuestra consistencia en su rendimiento.

Naive Bayes mostró un rendimiento consistente entre los conjuntos de entrenamiento y prueba, con precisiones del 98.47% y 85.10%, respectivamente. Esta consistencia sugiere que el modelo generaliza mejor y es más robusto frente a nuevos datos. Además, Naive Bayes es conocido por su eficiencia y rapidez, lo que lo hace adecuado para manejar grandes volúmenes de datos en un entorno Big Data. Las otras métricas, incluyendo precisión, recall y F1, son todas cercanas al 85%, lo que refuerza la fiabilidad del modelo.

Random Forest presentó el rendimiento más bajo entre los tres modelos, con una precisión del 73.52% en el conjunto de entrenamiento y 70.67% en el conjunto de prueba. Aunque Random Forest es robusto y maneja bien las interacciones complejas entre características, en este caso no superó a los otros dos modelos en términos de precisión. Sus métricas de precisión, recall y F1 son también inferiores, lo que sugiere que este modelo no es el más adecuado para nuestro proyecto.

Dado el análisis de los resultados y el rendimiento mostrado en el gráfico, Naive Bayes se destaca como el mejor modelo para desarrollar en nuestro proyecto. Su capacidad para manejar eficientemente grandes conjuntos de datos y características independientes, junto con su rendimiento consistente en los conjuntos de entrenamiento y prueba, lo hacen ideal para el análisis de sentimientos en un entorno Big Data. Implementar Naive Bayes permitirá automatizar y mejorar la precisión del análisis de sentimientos, proporcionando insights valiosos que pueden guiar las estrategias de mejora del servicio de la aerolínea y aumentar la satisfacción del cliente.

Recomendaciones y Conclusiones Generales del Proyecto

Recomendaciones

Mejora del Servicio al Cliente

Para mejorar la satisfacción del cliente, es esencial enfocarse en la capacitación del personal y en la comunicación clara de políticas y cargos adicionales. La correlación positiva entre el servicio del personal de cabina y los sentimientos positivos indica que la formación continua del personal en atención al cliente y resolución de problemas puede tener un impacto significativo. Además, la aerolínea debe asegurarse de que todas las políticas y posibles costos adicionales se comuniquen de manera clara durante el proceso de reserva y check-in.

Mejora de la Comodidad y Servicios a Bordo

Dado que la comodidad del asiento tiene una correlación significativa con la satisfacción general, se recomienda mejorar la ergonomía y el espacio de los asientos, especialmente en clases económicas. Aunque la correlación entre la calidad de los alimentos y bebidas y los sentimientos

es moderada, mejorar estos aspectos también puede contribuir positivamente a la experiencia del pasajero.

Optimización de la Relación Calidad-Precio

La relación calidad-precio es un factor clave en la satisfacción del cliente. La aerolínea debe revisar sus políticas de precios para asegurarse de que los pasajeros perciban un buen valor por su dinero. Implementar ofertas y promociones estratégicas puede mejorar la percepción de valor y atraer a más clientes.

Monitoreo y Mejora Continua

Es crucial implementar un sistema de retroalimentación continuo que permita a los pasajeros dar sus opiniones de manera fácil y rápida. Esto ayudará a identificar problemas en tiempo real y permitirá a la aerolínea responder proactivamente. Además, los modelos de clasificación deben ser actualizados y refinados regularmente con nuevos datos para mantener su precisión y relevancia.

Próximos Pasos

Ampliación del Análisis

Es necesario realizar un análisis más detallado segmentando a los clientes por variables adicionales como edad, frecuencia de vuelo y destinos preferidos para obtener insights más específicos. Además, es importante evaluar si existen variaciones en los sentimientos de los pasajeros dependiendo de la temporada del año o eventos específicos.

Integración con Sistemas Operativos

Integrar los modelos de clasificación en el sistema de gestión de la aerolínea para automatizar la clasificación de comentarios y generar informes de satisfacción en tiempo real es un paso crucial. También, desarrollar un sistema de alertas que notifique al equipo de atención al cliente sobre problemas recurrentes o aumentos en los comentarios negativos permitirá una intervención rápida.

Expansión de la Investigación

Expandir el análisis de sentimientos para incluir datos de redes sociales y otras plataformas de revisión en línea ofrecerá una visión más completa de la percepción de la aerolínea. Además,

realizar un análisis comparativo con otras aerolíneas ayudará a identificar áreas de mejora y oportunidades de diferenciación.

Mejoras en la Implementación de Modelos de Clasificación:

Optimización del Modelo Naive Bayes:

El modelo Naive Bayes ha demostrado ser el más eficaz en nuestro proyecto. Para mejorar aún más su implementación, se recomienda:

- **Selección de Características:** Realizar una selección y extracción de características más exhaustiva para identificar las variables que tienen el mayor impacto en los sentimientos de los pasajeros.
- **Tuning de Hiperparámetros:** Ajustar los hiperparámetros del modelo Naive Bayes para optimizar su rendimiento. Esto puede incluir la exploración de diferentes técnicas de suavizado.
- **Validación Cruzada:** Implementar técnicas de validación cruzada para asegurar que el modelo generaliza bien a diferentes subconjuntos de datos y minimizar el riesgo de sobreajuste.

Mejoras en el Preprocesamiento de Datos

- **Limpieza de Datos Continua:** Asegurar que los datos se limpien y preprocesen de manera continua para mantener la calidad de los datos alta. Esto incluye la eliminación de outliers y la imputación de valores faltantes.
- **Vectorización Avanzada:** Explorar técnicas avanzadas de vectorización de texto, como TF-IDF y embeddings de palabras, para capturar mejor el contexto y el significado de los comentarios de los pasajeros.

Integración y Monitoreo

- **Integración en Sistemas de Producción:** Integrar el modelo Naive Bayes en los sistemas operativos de la aerolínea para automatizar el análisis de sentimientos en tiempo real.
- **Monitoreo en Tiempo Real:** Establecer un sistema de monitoreo en tiempo real para rastrear el rendimiento del modelo y ajustar rápidamente en caso de cambios en los datos o el comportamiento de los clientes.

Conclusiones Generales

El proyecto de análisis de sentimientos en opiniones de pasajeros de una aerolínea utilizando técnicas de NLP y PySpark ha proporcionado insights valiosos sobre la percepción del servicio

por parte de los pasajeros. Los modelos de clasificación, particularmente Naive Bayes, demostraron ser herramientas eficaces para automatizar y mejorar la precisión del análisis de sentimientos.

Los hallazgos clave del proyecto subrayan la importancia de la comodidad del asiento, la calidad del servicio al cliente y la percepción de la relación calidad-precio en la satisfacción general de los pasajeros. La aerolínea puede utilizar estos insights para implementar mejoras específicas en sus operaciones y estrategias de servicio al cliente.

En resumen, este proyecto no solo ha demostrado el valor del análisis de sentimientos en la comprensión de la percepción del cliente, sino que también ha proporcionado un marco práctico para la implementación de mejoras continuas. Con una implementación adecuada de las recomendaciones y una adaptación proactiva a las necesidades y expectativas de los pasajeros, la aerolínea puede mejorar significativamente la satisfacción del cliente y su competitividad en el mercado.

Referencias:

https://projectgurukul.org/

https://www.kaggle.com/

https://medium.com/@dishantkharkar9/about-random-forest-algorithms-62163357db25

https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/

https://www.youtube.com/watch?app=desktop&v=SeM4Rtoa4EU