Análisis Exploratorio de Datos (EDA): TITANIC

Carrillo Hortua Juan Pablo

Mora Diaz Stephanie

Universidad Externado de Colombia UEC

Facultad departamento de matemáticas

Programa de Ciencia de datos

2025

Resumen

Este proyecto realiza un **Análisis Exploratorio de Datos (EDA)** sobre el conjunto de datos **"Titanic - Machine Learning from Disaster"**, disponible públicamente en Kaggle. El propósito es comprender los conceptos básicos de la librería orientada a manejo de datos **PANDAS**, y posteriormente comprender el naufragio del Titanic, empezando desde preguntas básicas hasta realizar cruces de datos y gráficos representativos de mortalidad segun edad, género, entre otros.

***Variables*:** Survival, Pclass, Sex, Age, Sibsp, Parch, Ticket, Fare, Cabin, Embarked

Tabla de Contenido

Introducción

En el presente proyecto se realiza un análisis exploratorio de los datos (EDA) con el conjunto de datos "**Titanic - Machine Learning from Disaster"** disponible en **Kaggle**, con el fin de comprender el naufragio del Titanic, con el fin de identificar patrones e información relevante sobre los pasajeros y su destino durante el naufragio.

Además, comprender los conceptos básicos de la librería orientada a manejo de datos **PANDAS**. Se busca transformar y visualizar la información para poder responder los diferentes interrogantes planteados, además de aplicar conceptos fundamentales de análisis estadístico y visualización de datos.

Objetivos

Objetivo General

Realizar un **Análisis Estadístico de los datos o Análisis Exploratorio de Datos (EDA)** de la base de datos expuesta en el encabezado, buscando correlaciones, Insights e información valiosa sobre los costes humanos de dicho naufragio.

Objetivos Específicos

Determinar la tasa de supervivencia por grupos etarios (niños, jóvenes, adultos y ancianos) para identificar la población con mayor índice de mortalidad. Además, se pretende analizar el género que presento la mayor mortalidad.

Hallar los diferentes grupos familiares de la base y ver cómo se distribuían por las cabinas (habitaciones).

Análisis exploratorio de los datos

Descripción de variables del Conjunto de Datos

Variable	*Definición*	*Key (Clave)*
Survival	Supervivencia	0 = No 1 = Sí
Pclass	Clase del billete/boleto	1 = 1ra 2 = 2da 3 = 3ra
Sex	Sexo	
Age	Edad en años	
Sibsp	# de hermanos / cónyuges a bordo del Titanic	
Parch	# de padres / niños a bordo del Titanic	
Ticket	Número del billete/boleto	
Fare	Tarifa del pasajero	
Cabin	Número de cabina	
Embarked	Puerto de Embarque	C = Cherbourg, Q = Queenstown, S = Southampton

Limpieza de datos

Antes de realizar el análisis exploratorio de los datos, se llevó a cabo el proceso de limpieza del conjunto de datos con el fin de garantizar un análisis correcto y representativo, a su vez previniendo errores futuros. En esta etapa se encontró variables como **Age, Cabin, Fare y Embarked** tenían valores faltantes; Esto se solucionó mediante el ingreso de valores representativos o eliminación de registros incompletos según su relevancia.

Cabe destacar el uso de herramientas estadísticas como medidas de tendencia central para escoger la mejor estrategia de imputación de datos, entre los principales criterios se encuentran pruebas de normalidad para datos continuos, como histogramas o la prueba shapiro para distribución normal.

Análisis estadístico básico:

Mediante la función **analizar** se realizó un contraste entre df_train y df_test, y se logró entender que las distribuciones de las variables presentaban una similitud marcada (cabe resaltar que la variable continua en ambos DataFrames no era normal, análisis se encuentra en la sección de limpieza, se basa en el test shapiro).

A su vez, cabe resaltar la similitud de las modas entre ambos sets de datos.

Unión de las dos bases de datos

si bien este apartado no presenta mayores insights es valioso mencionar que por la falta de la columna 'Survived' en **df_test** se optó por predecir está en base a un modelo de RandomForest importado desde sk-learn, dicho modelo se entrenó a partir del **df_train** y su rendimiento se evaluó, con una validación cruzada en base a ese mismo conjunto, dado los buenos resultados se utilizó el vector predicho por el modelo para rellenar la columna faltante en

el conjunto de train. Así, se evadió el problema de datos nulos al momento de concatenar ambos DataFrames.

Preguntas Univariadas:

Por medio de filtraciones de datos se logró encontrar respuestas a los interrogantes planteados inicialmente en el taller; entre los principales hallazgos se encuentra el hecho de que la edad promedio en la embarcación era de **29.88 años** el conteo de **muertos** ascendió a **809 personas**, mientras aquellos que lograron **sobrevivir** fueron tan solo **497** pasajeros, un saldo poco favorable de **supervivientes** siendo tan solo el **38.06%** de los tripulantes.

Mediante un filtro se logró determinar que la tarifa promedio pagada por los pasajeros de primera clase fue de **87.556$,** además en el siguiente grafico se ve la relación entre tarifas promedio por clase de forma porcentual.

235 pasajeros viajaron exactamente con un familiar abordo y 778 solos.

El pasajero de mayor edad tuvo 80 años y para obtener el más joven se hizo un promedio del intervalo de edades entre 0 y 1 ya que las edades decimales representaban aproximaciones, y mediante este método se logró inferir que la menor edad era 0.8, es decir un poco menor de un año. se halló el número de pasajeros que zarparon de cada puerto:

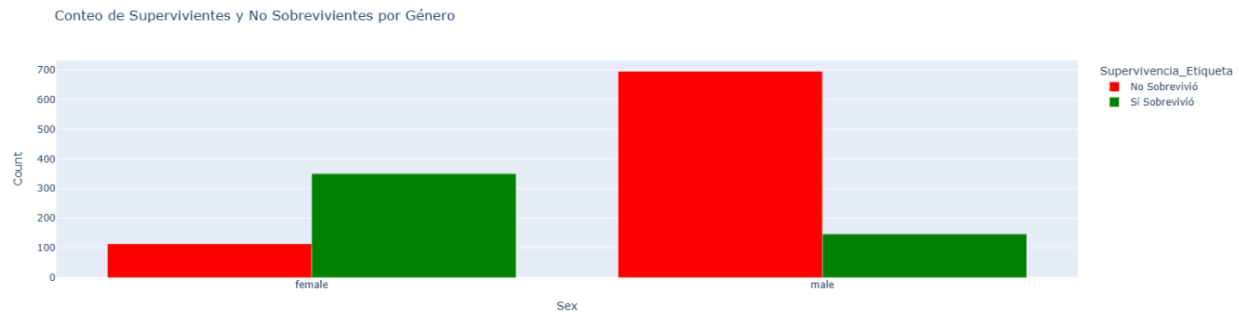
Embarked	Conteo
Cherbourg	270
Queenstown	123
Southampton	913

Cruces de variables

El siguiente apartado es realmente importante, pues en el radican algunos de los insights más interesantes y útiles del proyecto.

Iniciando por la supervivencia según el género del sujeto, en primera instancia si se observa el conteo bruto.

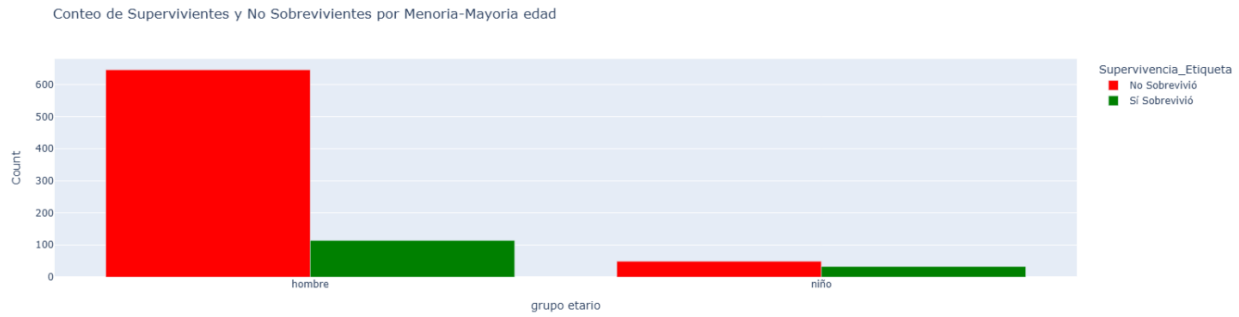
Sex	Survived	Count	Supervivencia_Etiqueta	Total_by_Sex	Percentage
female	0	114	No Sobrevivió	464	24.568965517241377
female	1	350	Sí Sobrevivió	464	75.43103448275862
male	0	695	No Sobrevivió	842	82.54156769596199
male	1	147	Sí Sobrevivió	842	17.458432304038006



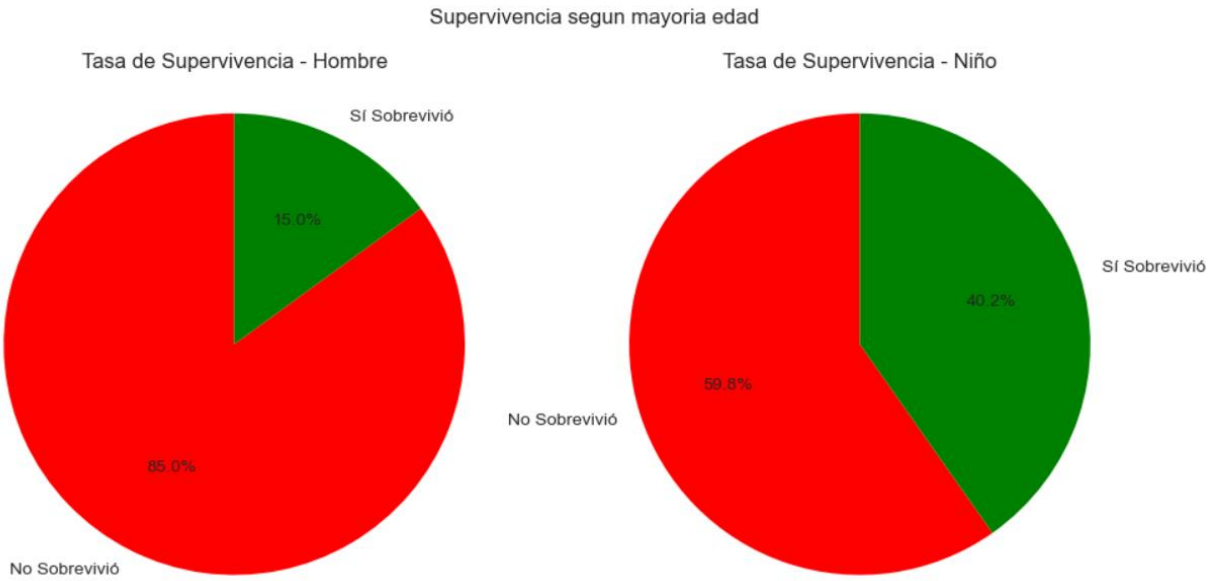
Se puede ver que en relación es más común que las mujeres sobrevivan. Sin embargo, al observar el gráfico según el porcentaje de supervivencia es donde realmente se aprecia la marcada diferencia que había entre la probabilidad de sobrevivir siendo mujer y siendo hombre, donde evidentemente las mujeres tenían más chances de sobrevivir al naufragio.

Si se observa las chances que tenía de sobrevivir niños o hombres adultos, en primera instancia con el conteo bruto podría parecer que era más probable que un hombre sobreviviera.

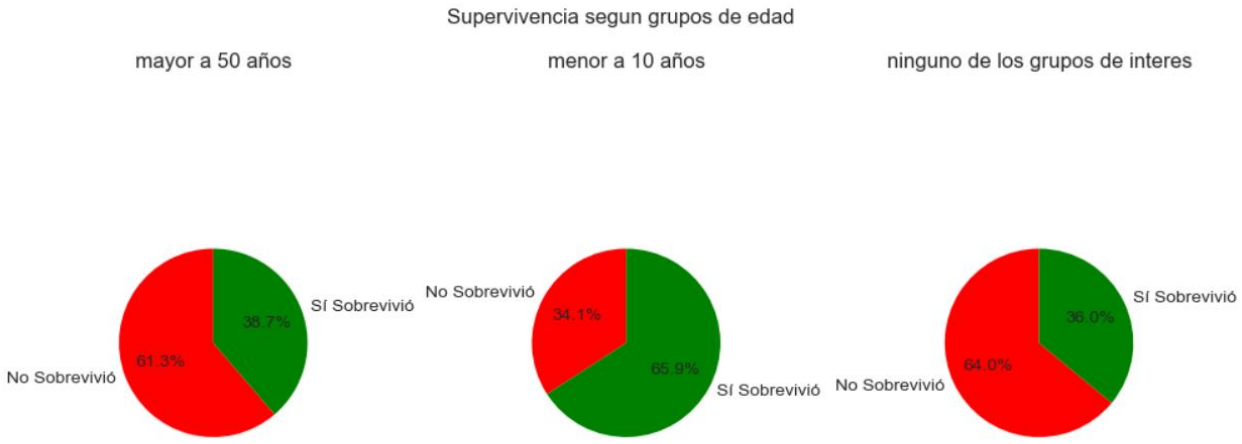
grupo etario	Survived	Count	Supervivencia_Etiqueta	Total por grupo etario	Percentage
hombre	0	646	No Sobrevivió	760	85.0
hombre	1	114	Sí Sobrevivió	760	15.0
niño	0	49	No Sobrevivió	82	59.756097560975604
niño	1	33	Sí Sobrevivió	82	40.243902439024396



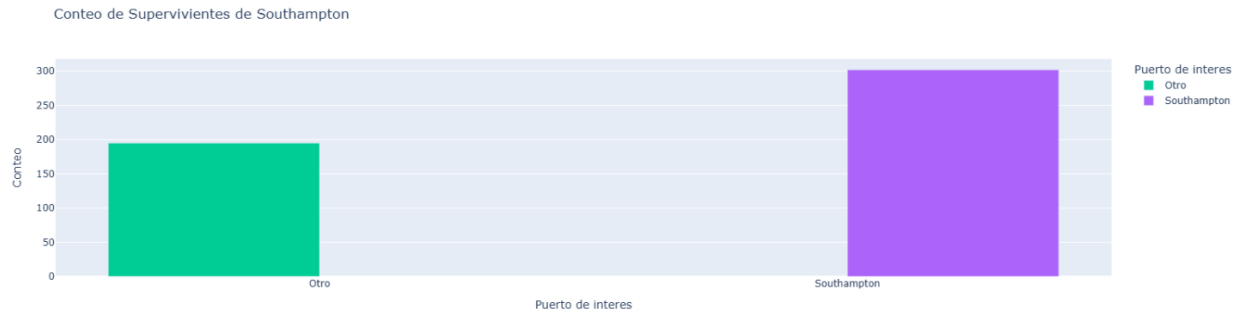
Sin embargo, los porcentajes nos indican que realmente los niños tenían muchas más posibilidades de sobrevivir, solo que el conteo bruto daba dicha ilusión a raiz de la menor frecuencia de niños.



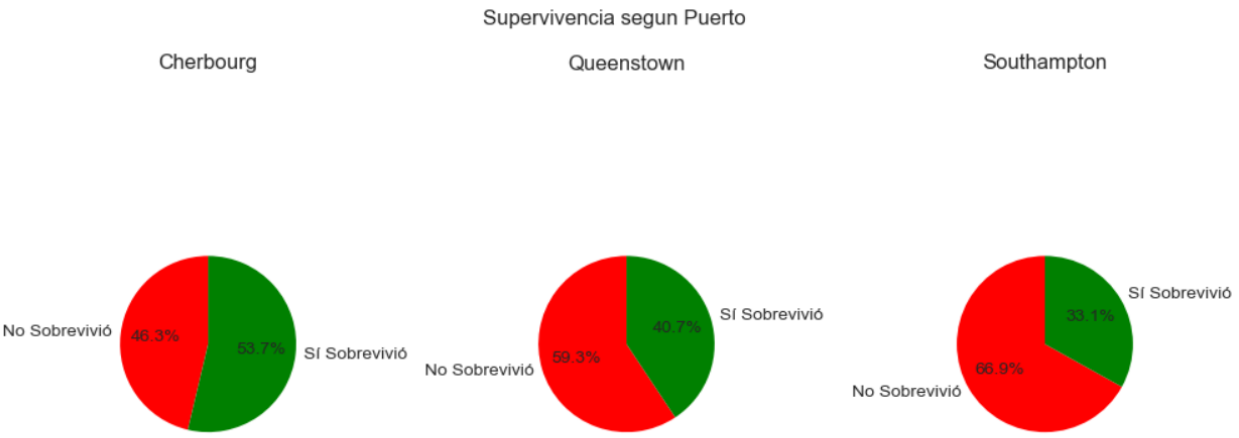
De igual manera ocurre con la supervivencia de pasajeros con edad mayor a 50 años o menor a 10 años, donde el conteo bruto hace pensar que los demás grupos tienen mejor chance de sobrevivir. No obstante, los porcentajes nos muestran que los menores a 10 años tenían posibilidades mayores de sobrevivir.



En cuanto a la supervivencia por puertos, desde el cruce inicial es evidente que la mayor parte de supervivientes partieron de Southampton, lo que hace pensar que tenían más probabilidad de sobrevivir. Aun así, tenga en cuenta que la mayoría de los pasajeros zarparon de dicho lugar.

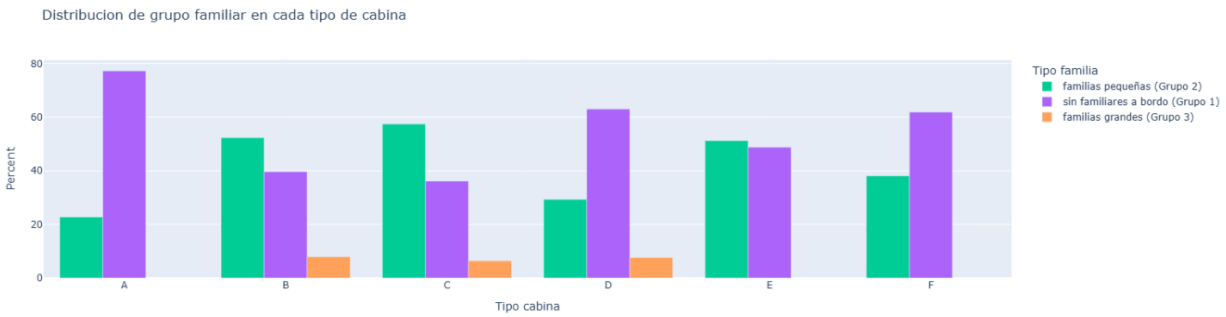


No obstante, revisando las probabilidades individuales de supervivencia por cada puerto, realmente las personas que zarparon de Southampton, tenían una probabilidad baja de sobrevivir.

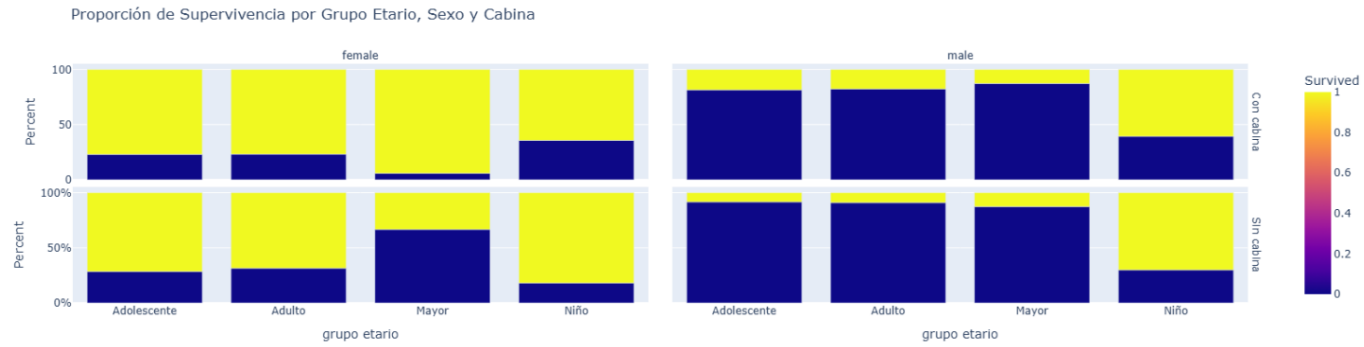


Por último, fue realmente claro que los pasajeros de primera clase tenían más chances de supervivencia.

Para obtener la distribución de tipos de grupo familiar, se decidió separar las cabinas según sus letras de inicio, ya que sin esta clasificación no se lograba apreciar una distribución clara, posteriormente se logró observar que la mayor concentración de personas está en las **cabinas D**, con una clara predominancia de personas sin familiares abordo, asimismo el tipo de cabina con menos gente son las de tipo **A**. También observe que es poco común ver familias numerosas, independientemente de la cabina.



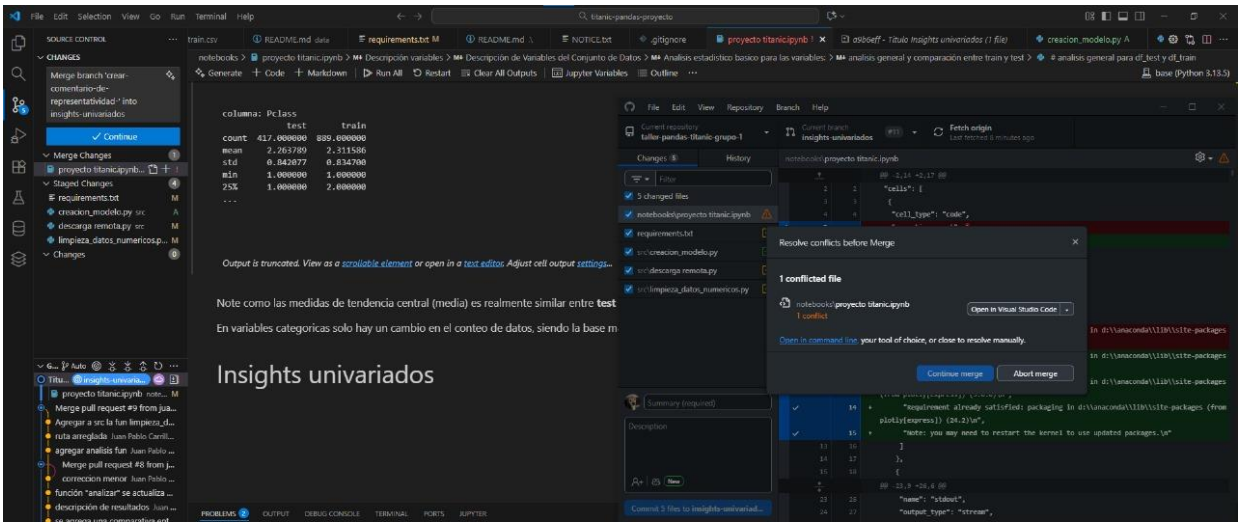
Por último, se cruza según edad, genero, posesión de cabina y supervivencia, y en el grafico obtenido se logra apreciar como el grupo con la mayor probabilidad de sobrevivir es aquel que corresponde a mujeres de la tercera edad con cabina. En general las mujeres tienen más chance de sobrevivir que los hombres. No hay mayor diferencia de supervivencia entre gente con y sin cabina.



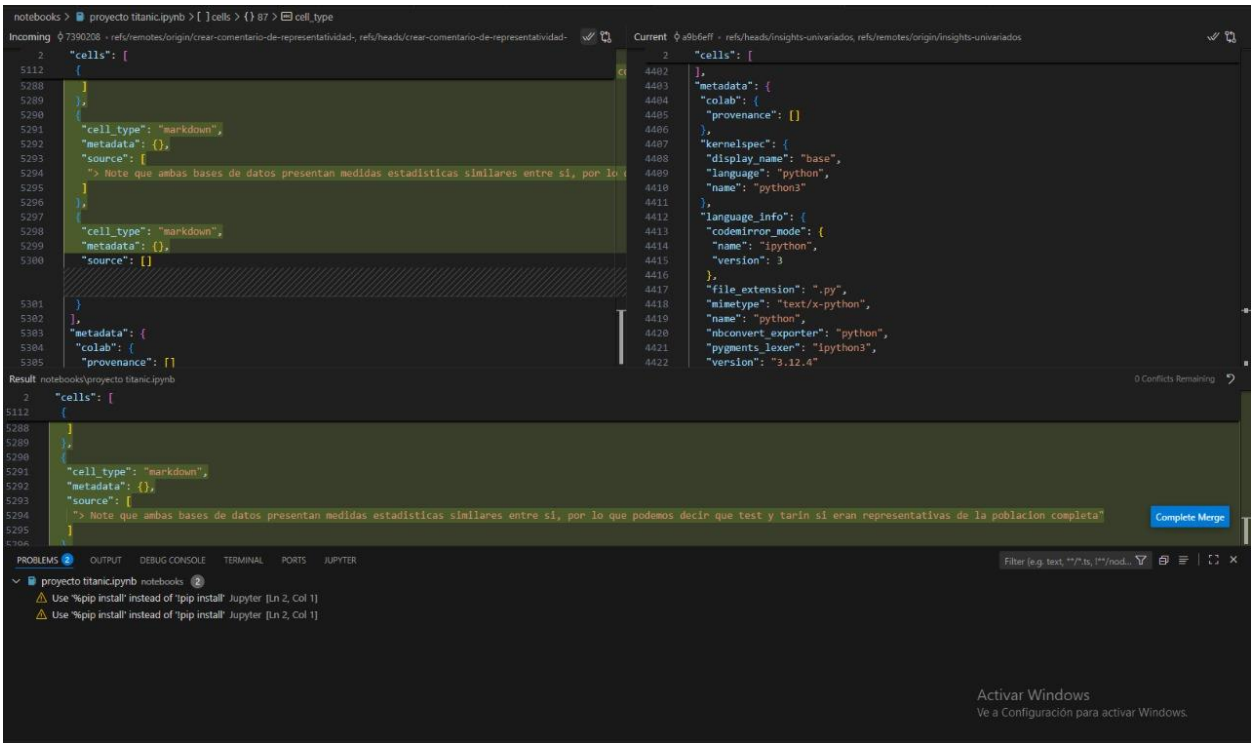
Descripción de resolución de conflictos

Durante el trabajo colaborativo se presentaron una serie de conflictos orgánicos a la hora de realizar el trabajo paralelo en ramas, en total se solucionaron alrededor de 4 conflictos y principalmente se usaron dos métodos de resolución de conflictos.

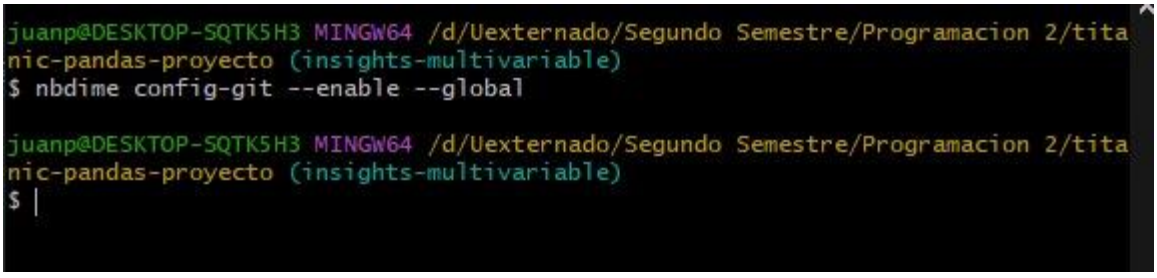
Un conflicto por destacar ocurrió cuando los dos miembros del equipo realizaron dos secciones del proyecto en paralelo desde el mismo punto de partida en **main**, más específicamente esto ocurre en la creación de las secciones de **insights univariado** e **insights multivariado**. Una vez finalizadas estas dos secciones, se crearon los pull request respectivos a cada rama, luego de esto se originó el merge de la rama univariada. No obstante, al momento de intentar mezclar la otra rama (multivariable), como era de esperar surgió un conflicto realmente significativo.



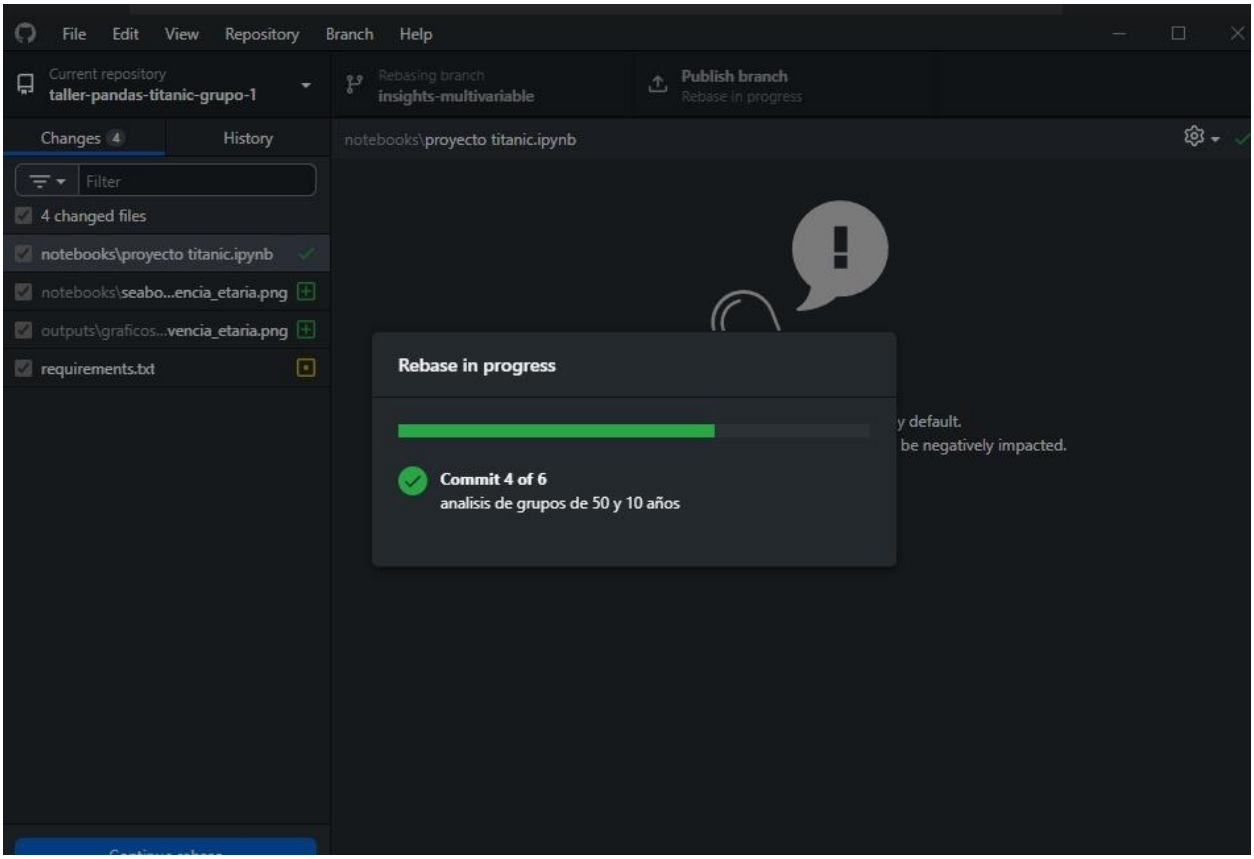
Dado que la sección multivariable estaba pensada para ir después de la sección univariada, se decide hacer una resolución de conflictos de tipo **rebase** la cual por practicidad se realizo de manera local desde el equipo del encargado de la rama multivariada.

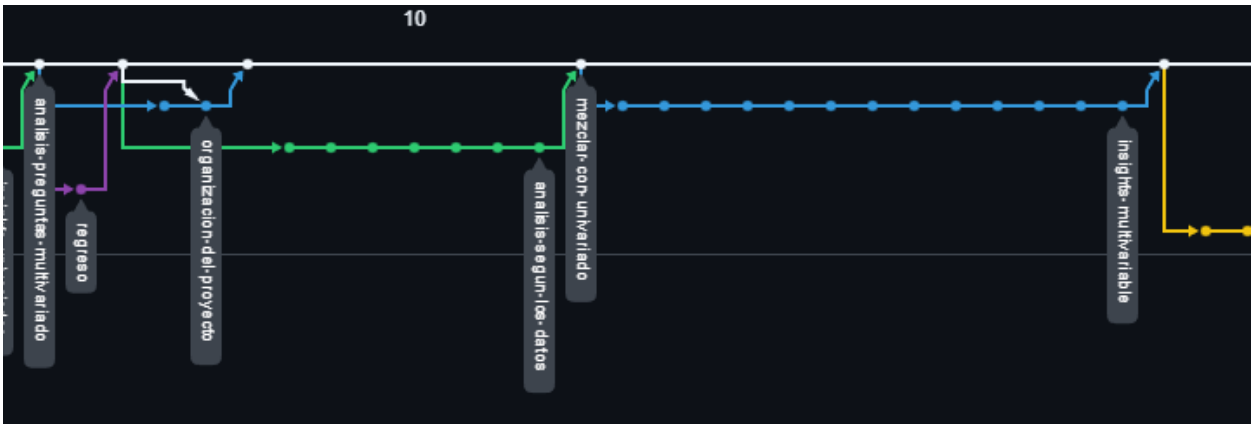


Sin embargo, observe que por la estructura interna del archivo **.ipynb** la resolución de conflictos clásica con edición de texto, aunque fuera por rebase era complicada por la interpretación de **.json**, por lo cual se procedió a utilizar comandos desde visual studio code, se utilizó el comando Git: Rebase Branch para insertar todos los commits de la rama deseada después de lo que ya estaba mezclado en main. Sin embargo, este método fallo debido a la falta de un paquete especifico de resolución de conflictos por un archivo **.ipynb**, por lo cual se hizo la descarga de nbdime y su configuración global.



Una vez instalada esta herramienta el rebase se llevo a cabo de manera automática y sin ningún percance mayor.





Observe como la rama de color verde (insights univariados) parte de un punto en main y se guia de esta la rama de insights multivariados, luego mire como todos los commits de insights multivariados están después de los commits de insights univariados gracias al método rebase.

El resto de los conflictos fueron resueltos de manera típica, en el caso de archivos .py desde el propio git hub aceptando o rechazando cambios y en el caso de archivos .ipynb desde nbdime, cabe resaltar que se prefirió resolver la mayoría de los conflictos desde una rama auxiliar que emula main para después hacer el merge oficial.

Conclusiones

Después del análisis estadístico realizado, se puede inferir que el grupo con mayor probabilidad de supervivencia corresponden a las niñas de primera clase que partieron desde Cherbourg, seguido de las mujeres de tercera edad bajo las mismas condiciones. Estos resultados muestran como el género, la edad y la clase fueron factores determinantes en las probabilidades de supervivencia durante el naufragio. Por el contrario, los hombres adultos de tercera clase representaron el grupo con menos probabilidad de sobrevivir.

Estos hallazgos concuerdan con los patrones observados en los cruces de variables, donde pudimos evidenciar que las mujeres tenían tasas de supervivencia significativamente superiores a la de los hombres. Además, los niños presentaban unas mejores posibilidades de sobrevivir que los adultos, aunque había menor cantidad de niños abordo. Otro hallazgo significativo fue que la clase tuvo un papel crucial, mostrando que los pasajeros de primera clase tienen ventaja en comparación con los pasajeros de segunda y tercera clase.

Referencias y Anexos

Kaggle (2012). *Titanic - Machine Learning from Disaster*. https://www.kaggle.com/competitions/titanic/overview

Python Software Foundation. (2025). *Python 3.14.0 documentation*. https://docs.python.org/3/

Pandas community. (2025). *pandas documentation (versión 2.3.3).* https://pandas.pydata.org/docs/

Pandas community. (s. f.). *pandas cheat sheet*. https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

Plotly Technologies Inc. (2025). *Plotly Python graphing library* (versión 6.4.0). https://plotly.com/python/

scikit-learn developers. (2025). *scikit-learn: Machine learning in Python 1.7.2 (stable documentation)*. https://scikit-learn.org/stable/

NumPy Developers. (2025). *NumPy documentation* (versión 2.x). https://numpy.org/doc/

Seaborn. (2024). *Statistical data visualization* (v 0.13.2). https://seaborn.pydata.org/ (seaborn.pydata.org)

Matplotlib Development Team. (2025). *Matplotlib 3.10.7 documentation*. https://matplotlib.org/stable/index.html