Informe Final del Proyecto: Análisis de Datos del Titanic

Autor: William Gómez Marín

1. Introducción

El presente proyecto tiene como propósito analizar el conjunto de datos del Titanic mediante herramientas de análisis de datos en Python, principalmente utilizando la biblioteca *pandas*. A través del procesamiento, limpieza y exploración de la información contenida en los archivos *train.csv* y *test.csv*, se busca identificar los factores que influyeron en la supervivencia de los pasajeros.

El análisis incluye la descripción de las variables disponibles (como sexo, edad, clase social, tarifa y número de familiares), la generación de estadísticas básicas y comparativas entre distintos grupos, así como la creación de nuevas variables que permitan obtener una visión más completa del comportamiento de los datos. De esta forma, el proyecto no solo pretende comprender mejor los patrones de supervivencia en el desastre del Titanic, sino también aplicar conceptos fundamentales de análisis estadístico, manipulación de datos y visualización, fortaleciendo las habilidades prácticas en el uso de Python para la minería de datos.

2. Objetivo General del Proyecto:

El objetivo principal del proyecto fue analizar el conjunto de datos del Titanic disponible en la competencia de Kaggle "Titanic - Machine Learning from Disaster", con el propósito de identificar los factores que influyeron en la supervivencia de los pasajeros. El estudio tuvo un enfoque exploratorio y analítico, empleando la librería Pandas de Python para realizar la carga, limpieza, descripción y evaluación de las variables más relevantes.

A través de este análisis se buscó responder la pregunta central planteada:

¿Qué tipos de personas tenían más probabilidades de sobrevivir al hundimiento del Titanic?

De esta manera, el proyecto permitió aplicar técnicas de análisis de datos y fundamentos de aprendizaje automático en un contexto histórico y real, contribuyendo al desarrollo de habilidades prácticas en ciencia de datos.

3. Principales Hallazgos del Análisis con Pandas:

1. Durante el desarrollo del análisis se trabajó con tres archivos principales:

- **train.csv**: contiene información de 891 pasajeros con la variable de supervivencia ("Survived").

- **test.csv**: contiene 418 registros sin la variable de supervivencia, destinado a predicciones.
- **gender_submission.csv**: solo sirve como ejemplo del formato esperado para enviar resultados.

2. Los principales hallazgos fueron los siguientes:

- **Composición del conjunto de datos:** El dataset de entrenamiento incluye 891 pasajeros, mientras que el de prueba contiene 418. Ambos presentan una estructura similar en cuanto a variables, aunque con diferentes tamaños.
- **Valores faltantes:** Se detectaron valores ausentes principalmente en las columnas “Age” y “Cabin”, siendo esta última casi completamente vacía. Esto sugiere la necesidad de imputar datos o descartar la variable para análisis posteriores.
- **Distribución demográfica**: La edad promedio de los pasajeros es cercana a 30 años. La mayoría de los viajeros eran hombres, y el puerto de embarque predominante fue Southampton.
- **Variables socioeconómicas:** Se observó una mayor proporción de pasajeros en tercera clase, lo que refleja el perfil social del viaje. Las tarifas pagadas (Fare) mostraron alta variabilidad, indicando diferencias marcadas en el poder adquisitivo.
- **Relaciones con la supervivencia:**
- Las mujeres y los niños presentaron tasas de supervivencia considerablemente más altas.
- Los pasajeros de primera clase tuvieron mayores probabilidades de sobrevivir que los de clases inferiores.
- La edad mostró una tendencia: a menor edad, mayor posibilidad de supervivencia.
- Los hombres adultos en tercera clase fueron el grupo con menor tasa de supervivencia.

En conjunto, los resultados confirman la relevancia de las variables sexo, clase y edad como factores predictivos de la supervivencia, coherentes con el conocido principio histórico de *“mujeres y niños primero”*.

- Conclusiones del taller con pandas:

El análisis del Titanic permitió aplicar los fundamentos del análisis exploratorio de datos (EDA) para comprender las relaciones entre variables numéricas y categóricas, gestionar valores nulos y extraer conclusiones útiles sin recurrir a modelos predictivos complejos.

Entre los principales aprendizajes destacan:

- La importancia de inspeccionar la estructura de los datos antes de cualquier modelado.
- La necesidad de limpieza y tratamiento de valores faltantes para evitar sesgos en el análisis.
- La utilidad de las herramientas de Pandas para describir, agrupar y visualizar información de manera eficiente.
- La posibilidad de extraer conclusiones significativas a partir de patrones simples en los datos.

Este taller reforzó la comprensión de las etapas iniciales del proceso de ciencia de datos: exploración, preprocesamiento y análisis descriptivo.

4. Conclusiones del Equipo:

El equipo concluye que este proyecto constituye un excelente punto de partida para adentrarse en el análisis de datos y en la aplicación práctica de Python.
A partir del trabajo se obtuvieron los siguientes aprendizajes:

- Se adquirió experiencia en el manejo de datasets reales con información incompleta y heterogénea.
- Se fortalecieron las habilidades en limpieza, filtrado y agrupamiento de datos con Pandas.
- Se comprendió el valor de las variables categóricas (sexo, clase, puerto de embarque) en la explicación de resultados.
- Se consolidó la capacidad de interpretar hallazgos de manera crítica y contextual.

En conclusión, el proyecto del Titanic demostró cómo el análisis de datos puede revelar patrones sociales y económicos históricos, al tiempo que refuerza competencias técnicas esenciales para la ciencia de datos aplicada.

Conclusiones por punto del taller de pandas (sin incluir código):

1. **Cargue cada conjunto de datos en un dataframe distinto. Ignore el archivo *gender_submission.csv*.**

Interpretación resultados: El archivo de entrenamiento (train.csv) contiene información de los pasajeros del Titanic junto con la columna de supervivencia, mostrando características como edad, sexo, clase, familiares a bordo y tarifa; lo impreso reflejó las primeras filas de estos datos, permitiendo ver quién sobrevivió y quién no. Por otro lado, el archivo de prueba (test.csv) incluye los mismos tipos de información de los pasajeros, pero sin la columna de supervivencia, y lo impreso mostró las primeras filas con las características disponibles para predecir la supervivencia. Finalmente, el archivo gender_submission.csv se ignoró, ya que solo sirve como ejemplo de formato de predicciones. En resumen, el entrenamiento tiene respuestas conocidas para aprender patrones, mientras que la prueba solo contiene los datos necesarios para hacer predicciones.

Entrenamiento:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Prueba:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

2. **Describa las variables que se involucran en los datasets.**

Interpretación resultados: Los datos del Titanic incluyen un conjunto de entrenamiento con 891 pasajeros y uno de prueba con 418. El archivo de entrenamiento contiene la variable Survived, mientras que el de prueba no. Ambos presentan valores faltantes, especialmente en Age y sobre todo en Cabin, que está casi vacía. Las estadísticas muestran que los pasajeros tenían una edad promedio cercana a 30 años, la mayoría viajaba en tercera clase, las tarifas son muy variables y la mayoría de las personas no viajaba con familiares. En las variables categóricas predomina el sexo masculino y el embarque desde Southampton. El archivo gender_submission.csv únicamente muestra el formato esperado para una predicción y no aporta información útil para el análisis. En conjunto, los datos revelan patrones claros por edad, clase, sexo y tarifa, aunque requieren limpieza debido a valores faltantes.

3. **Genere las estadísticas básicas de cada dataframe y haga comparaciones entre los dataframes.**

Interpretación resultados: El análisis de las estadísticas básicas de los datasets muestra que los dos conjuntos son comparables en términos de distribución de variables, aunque difieren en tamaño: el dataset de entrenamiento tiene 891 registros, mientras que el de prueba tiene 418. La edad promedio de los pasajeros es similar, alrededor de 29.7 años en entrenamiento y 30.3 en prueba, con valores extremos que van desde bebés hasta adultos mayores, lo que indica una población heterogénea. Las tarifas (Fare) también presentan medias comparables, aunque con alta dispersión debido a boletos de lujo, mientras que la mayoría de los pasajeros viaja en tercera clase (Pclass=3) y los promedios de familiares a bordo (SibSp y Parch) son bajos, mostrando que la mayoría de los pasajeros viajaba solo o con pocos acompañantes.

La variable Survived solo existe en el dataset de entrenamiento, con una media de 0.38, lo que indica que aproximadamente el 38 % de los pasajeros sobrevivió. En general, las estadísticas muestran que la distribución de características como edad, clase, familiares y tarifas es muy similar entre los datasets, asegurando que un modelo entrenado pueda aplicarse adecuadamente al dataset de prueba. La principal diferencia radica en la presencia de la variable objetivo y en la cantidad de registros.

```
=== Estadísticas del dataset de entrenamiento ===
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
=== Estadísticas del dataset de prueba ===
```

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

4. **Cree una nueva variable que se llame *Familiares* y que sea la suma de las variables *SibSp* y *Parch*. ¿Qué opina de esta nueva variable?**

Interpretación resultados: La variable Familiares, que suma SibSp y Parch, refleja el número total de familiares directos a bordo de cada pasajero. Los datos muestran que la mayoría de los pasajeros viajaba solo, ya que la mediana es 0 tanto en entrenamiento como en prueba. La media se sitúa alrededor de 0.9 en entrenamiento y 0.84 en prueba, indicando que pocos pasajeros tenían uno o más familiares acompañándolos. Sin embargo, existen casos con familias numerosas, alcanzando un máximo de 10 familiares. Esta variable es útil porque resume en un solo valor la información familiar, lo que facilita el análisis de cómo la presencia de familiares pudo haber influido en la supervivencia de los pasajeros.

5. **Junte verticalmente los dataframes *train* y *test* en un solo conjunto de datos. Antes de hacerlo, agregue una nueva columna llamada source que indique el origen de cada fila ("train" o "test").**
   Recuerde que el conjunto *test* no contiene la variable *Survived*, por lo tanto ¿qué va a hacer con esta columna?

Interpretación resultados: El conjunto de datos de prueba no incluye la variable Survived, que indica si un pasajero sobrevivió o no. Para poder unir verticalmente los datasets de entrenamiento y prueba en un solo dataframe, es necesario que ambos tengan las mismas columnas. Por ello, se agrega la columna Survived al dataset de prueba, pero con valores NaN. Esto permite diferenciar claramente los registros de entrenamiento, donde los valores son conocidos, de los de prueba, donde los valores deben ser predichos. Además, garantiza que la estructura del dataframe combinado sea consistente, facilitando análisis y procesamiento posteriores.

6. **Revise las estadísticas nuevamente y analice (discuta) si los conjuntos de datos *train.csv* y *test.csv* son representativos del conjunto de datos del ítem anterior.**

Interpretación resultados: El análisis muestra que los conjuntos de entrenamiento y prueba del Titanic son coherentes y muy similares entre sí. Las distribuciones de edad, tarifa y número de familiares, así como las proporciones de clase, sexo y puerto de embarque, prácticamente coinciden en ambos. Esto indica que provienen de la misma población y no presentan diferencias que generen sesgos. La única variable exclusiva del entrenamiento es Survived, como es normal en un problema supervisado. En conjunto, los datos son consistentes, equilibrados y adecuados para construir modelos predictivos confiables.

7. **¿Qué variables presentan datos faltantes? ¿Qué significado tienen esos datos faltantes?**

Interpretación resultados: El análisis muestra que los conjuntos de entrenamiento y prueba del Titanic son coherentes y muy similares entre sí. Las distribuciones de edad, tarifa y número de familiares, así como las proporciones de clase, sexo y puerto de embarque, prácticamente coinciden en ambos. Esto indica que provienen de la misma población y no presentan diferencias que generen sesgos. La única variable exclusiva del entrenamiento es Survived, como es normal en un problema supervisado. En conjunto, los datos son consistentes, equilibrados y adecuados para construir modelos predictivos confiables.

```
=== REPORTE DE DATOS FALTANTES: TRAIN ===
```

	Faltantes	% del total
Age	177	19.87
Cabin	687	77.10
Embarked	2	0.22

```
=== REPORTE DE DATOS FALTANTES: TEST ===
```

	Faltantes	% del total
Age	86	20.57
Fare	1	0.24
Cabin	327	78.23

8. **Responda las siguientes preguntas basándose en los datos:**

a) ¿Cuál es la edad promedio de los pasajeros del Titanic?
b) ¿Cuántos pasajeros sobrevivieron y cuántos murieron?
c) ¿Cuál es la tarifa promedio pagada por los pasajeros de primera clase?
d) ¿Cuántos pasajeros viajaron con un familiar a bordo?
e) ¿Cuál es la edad más joven y la más vieja de los pasajeros?
f) ¿Cuántos pasajeros viajaron desde cada puerto de embarque?
g) ¿Cuántos pasajeros viajaron solos y cuántos con familiares?

Interpretación resultados: Los datos del Titanic muestran que los pasajeros tenían en promedio 29.7 años, con edades entre 0.42 y 80 años. De los 891 pasajeros, 549 murieron y 342 sobrevivieron, evidenciando una alta mortalidad. Los pasajeros de primera clase pagaron en promedio 84.15 por su tarifa, reflejando su mayor estatus económico. En cuanto a la composición familiar, 354 viajaban con algún familiar, mientras que 537 iban solos. La mayoría embarcó en Southampton (644), seguido de Cherbourg (168) y Queenstown (77). En conjunto, estos datos permiten entender mejor el perfil de los pasajeros y las condiciones del viaje.

9. **Responda las siguientes preguntas y luego verifique con los datos (utilice groupby, agg o tablas de contingencia para justificar sus respuestas y, cuando se pida proporcionalmente, calcule tasas en lugar de conteos):**

a) ¿Cree usted que sobrevivieron más mujeres que hombres? ¿Y proporcionalmente?
b) ¿Cree usted que sobrevivieron más niños que hombres adultos? ¿Y proporcionalmente?
c) ¿Cree usted que pasajeros con edad mayor a 50 años o menor a 10 años sobrevivieron más que pasajeros de cualquier otro grupo etario? ¿Y proporcionalmente?
d) ¿Cree usted que la mayor cantidad de pasajeros sobrevivientes partieron del puerto de Southampton o de los otros puertos? ¿Y proporcionalmente?
e) ¿Cree usted que pasajeros con tiquetes de primera clase sobrevivieron más que pasajeros con otros tipos de tiquetes? ¿Y proporcionalmente?

Interpretación resultados: El análisis de supervivencia del Titanic revela patrones claros influenciados por sexo, edad, puerto de embarque y clase socioeconómica. Las mujeres presentan una gran ventaja, con una supervivencia del 74.2% frente al 18.9% de los hombres. Los niños también muestran una supervivencia proporcionalmente alta (61.3%), especialmente los menores de 10 años, que alcanzan el 59.3%, aunque son un grupo pequeño. Sobre los puertos, Southampton aporta más sobrevivientes por cantidad de pasajeros, pero Cherbourg posee la mayor tasa proporcional (55.4%). La clase social también marca una diferencia importante: los pasajeros de primera clase tienen la tasa más alta (62.9%), seguidos por segunda (47.3%) y tercera clase (24.2%). En conjunto, estos resultados evidencian que factores sociales como sexo, edad y clase influyeron decisivamente en las probabilidades de supervivencia.

10. **Usando la nueva variable llamada *Familiares* (suma de *SibSp* y *Parch*), clasifique a los pasajeros en tres grupos familiares:**

- Grupo 1: sin familiares a bordo (*Familiares* = 0)
- Grupo 2: familias pequeñas (*Familiares* entre 1 y 3)
- Grupo 3: familias grandes (*Familiares* ≥ 4)

Luego, muestre cómo era la distribución de cabinas (*Cabin*) en cada grupo familiar.

Interpretación resultados: La mayoría de los pasajeros no tiene cabina registrada en los tres grupos familiares, siendo especialmente alto en el Grupo 1 (sin familiares), que concentra la mayor cantidad de valores faltantes. En el Grupo 2

(familias pequeñas) aparece una mayor variedad de cabinas identificadas, lo que sugiere mayor presencia en clases superiores. El Grupo 3 (familias grandes) también presenta principalmente cabinas sin registrar, aunque algunas cabinas compartidas como C23 C25 C27 se repiten. En general, los datos muestran que los registros de cabina son escasos y más comunes en grupos familiares pequeños y con mejor posición socioeconómica.

11. **¿Cuál es la proporción de personas del mismo sexo, por cada grupo etario, que tenían y no tenían cabina, que sobrevivieron y no? Finalice con una adecuada conclusión (análisis).**
    Antes de responder, defina una nueva variable llamada *grupo etario* a partir de la variable *Age*, usando las tres categorías:

- Niños (menores de 10 años)
- Adultos (de 18 a 49 años)
- Mayores (50 años o más)

Interpretación resultados: El análisis de la proporción de supervivencia en función del sexo, el grupo etario y la disponibilidad de cabina muestra una clara interacción entre estos factores. Las mujeres presentan, en general, las mayores tasas de supervivencia, especialmente las adultas y mayores que tenían cabina registrada, donde las probabilidades superan ampliamente el 85% e incluso alcanzan el 96.9% en el caso de las mujeres adultas con cabina. Incluso sin cabina, las mujeres mantienen tasas relativamente altas, lo que refleja la prioridad que se les otorgó durante la evacuación. En contraste, los hombres adultos y mayores muestran las tasas más bajas, particularmente aquellos sin cabina: los adultos sin cabina apenas alcanzan un 12.6% de supervivencia, y los mayores, un 6.7%. Sin embargo, los niños, tanto hombres como mujeres, presentan proporciones elevadas de supervivencia, independientemente de la disponibilidad de cabina, lo que indica que este grupo también fue priorizado en el rescate. En conjunto, los datos evidencian que las probabilidades de sobrevivir dependían fuertemente del sexo, la edad y el acceso a cabina (como indicador de clase social), favoreciendo consistentemente a mujeres, niños y pasajeros con mejores condiciones socioeconómicas.

12. **Investigue, realice y analice gráficos asociados al conjunto de datos resultante del ítem 6 que considere necesarios. Agregue una interpretación de cada gráfico.**
    Ayuda: https://www.data-to-viz.com/

Interpretación resultados: El análisis de los grupos familiares revela diferencias marcadas en las condiciones de viaje y en la supervivencia. El Grupo 1, compuesto por pasajeros que viajaban solos, es el más numeroso y también el que muestra menor acceso a cabinas, lo que se refleja en una tasa de supervivencia relativamente baja. En contraste, el Grupo 2, formado por familias pequeñas, presenta mejores condiciones: una proporción más alta de pasajeros con cabina y, consecuentemente, la mayor tasa de supervivencia entre los tres grupos. Por último, el Grupo 3, correspondiente a familias grandes, es el más vulnerable, con muy poco acceso a cabinas y la supervivencia más baja, especialmente entre quienes no tenían cabina asignada. En general, los resultados indican que tanto el tamaño de la familia como el acceso a cabina influyeron significativamente en las probabilidades de sobrevivir al hundimiento.

13. **¿Quiénes tenían más probabilidades de sobrevivir?**
    Cheat sheet: https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

Interpretación resultados: El análisis del conjunto de datos del Titanic revela que solo el 38.38% de los pasajeros sobrevivió, lo que demuestra la magnitud de la tragedia. Al examinar los resultados por género, se observa una diferencia muy marcada: las mujeres presentaron una tasa de supervivencia del 74.2%, mientras que los hombres apenas alcanzaron un 18.9%, lo que confirma la aplicación de la regla “mujeres y niños primero” durante la evacuación. La clase social también tuvo un papel determinante: los pasajeros de primera clase tuvieron una probabilidad de supervivencia del 63%, los de segunda clase del 47% y los de tercera clase solo del 24%, reflejando las desigualdades sociales y el acceso privilegiado a los botes salvavidas. Al combinar ambos factores, las mujeres de primera y segunda clase fueron el grupo con mayores posibilidades de vivir (96.8% y 92.1%), mientras que los hombres de tercera clase fueron los más afectados, con solo un 13.5% de supervivencia. Además, los sobrevivientes eran ligeramente más jóvenes (edad promedio de 28.3 años) que los que fallecieron (30.6 años), lo que sugiere una leve ventaja para los pasajeros más jóvenes. En conclusión, las mujeres jóvenes de clases altas fueron quienes tuvieron más probabilidades de sobrevivir, mientras que los hombres adultos de tercera clase fueron el grupo con menor posibilidad de hacerlo, evidenciando que el género, la clase social y la edad influyeron decisivamente en las oportunidades de supervivencia a bordo del Titanic.

```
 Porcentaje total de supervivientes: 38.38%

Probabilidad de supervivencia por género:

Sex
female    74.203822
male      18.890815
Name: Survived, dtype: float64


Probabilidad de supervivencia por clase:

Pclass
1    62.962963
2    47.282609
3    24.236253
Name: Survived, dtype: float64
```

```
Probabilidad de supervivencia por género y clase:

Sex     Pclass
female  1         96.808511
        2         92.105263
        3         50.000000
male    1         36.885246
        2         15.740741
        3         13.544669
Name: Survived, dtype: float64


Promedio de edad (0 = no sobrevivió, 1 = sobrevivió):

Survived
0    30.626179
1    28.343690
```