Proyecto Final
Programación II
Talle pandas titanic grupo 1
Juan Sebastian Vargas Brausin - Santiago Parra Mogollón

1.
```
dat = pd.read_csv("../data/raw/train.csv")
dat1 = pd.read_csv("../data/raw/test.csv")
```

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

En el archivo zip hay 2 *dataset* que serán analizados. El primero es *Train*, que contiene 891 registros; cada uno de estos corresponde a un pasajero y, asimismo, cada pasajero tiene a su vez 12 columnas que corresponden a las respectivas variables.

La primera variable corresponde al *PassengerID*, que es un identificador único del pasajero y está numerado del 1 al 891 sin ningún patrón asociado, dando siempre saltos de uno en uno. Luego está *Survived*, que, como su nombre indica, es la variable que describe si el pasajero sobrevivió o no; es una variable numérica binaria (0 = No sobrevivió, 1 = Sí sobrevivió).

Continuamos ahora con *Pclass*, que describe las 3 clases que tenía el Titanic, siendo la primera la más alta (identificada con el número 1), seguida por la segunda clase (identificada con el número 2) y, por último, la clase 3, que es la más baja (identificada con el número 3). Esta variable, a pesar de ser numérica, en la práctica se trata de una variable categórica.

Luego tenemos la variable *Name*, que sencillamente indica el nombre del pasajero y es una variable cualitativa. Posteriormente está la variable *Sex*, que corresponde al sexo biológico de los pasajeros; *Sex* es una variable cualitativa.

Continuamos ahora con *Age*, que indica la edad de los pasajeros; por el tipo de variable, corresponde claramente a una cuantitativa. Ahora se analizarán 2 variables que son muy parecidas: estas son *SibSp* y *Parch*. Ambas hacen referencia a la cantidad de familiares a bordo; la diferencia es que la primera corresponde al número de hermanos y/o cónyuge a bordo, y la segunda corresponde al número de hijos y/o padres a bordo. Dado que la naturaleza de las variables es la misma, ambas son del tipo cuantitativa.

Seguidamente está *Ticket*, que nos da el número de boleto del pasajero. En algunos casos esta variable es totalmente numérica, pero a veces también combina letras; por esto, *Ticket* termina siendo una variable cualitativa. De manera muy similar, *Cabin* nos da el código único (alfanumérico) de la cabina, y evidentemente también resulta siendo una cualitativa.

La variable *Fare* nos da el precio del pasaje en libras esterlinas; esta variable, en cambio, sí es cuantitativa.

Finalmente, la última variable de *Train* es *Embarked*, que nos dice el puerto en donde embarcó el pasajero y se clasificó de la siguiente manera: C = Cherbourg, Q = Queenstown, S = Southampton. Esta, al no tener ninguna característica numérica, termina siendo una variable cualitativa.

Test es un dataset prácticamente igual con las únicas diferencias de que tiene menos datos, con un total de 418. sin embargo la diferencia más notoria es que ya que el propósito original de este dataset es ser usado para probar un modelo de ML, específicamente con el propósito de predecir si el pasajero sobrevive o no, no cuenta con la variable survived. Por lo demás es dataset y sus variables se comportan prácticamente de la misma manera.
#Train :

	PassengerId	**Survived**	**Pclass**	**Name**	**Sex**	**Age**	**SibSp**	**Parch**	**Ticket**	**Fare**	**Cabin**	**Embarked**
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000	891.000000	891	891.000000	204	889
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	147	3
top	NaN	NaN	NaN	Dooley, Mr. Patrick	male	NaN	NaN	NaN	347082	NaN	G6	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	4	644
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008	0.381594	NaN	32.204208	NaN	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743	0.806057	NaN	49.693429	NaN	NaN

min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000	0.000000	NaN	7.910400	NaN	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000	0.000000	NaN	31.000000	NaN	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN	NaN

Test:
#

	PassengerId	**Pclass**	**Name**	**Sex**	**Age**	**SibSp**	**Parch**	**Ticket**	**Fare**	**Cabin**	**Embarked**
count	418.000000	418.000000	418	418	332.000000	418.000000	418.000000	418	417.000000	91	418
unique	NaN	NaN	418	2	NaN	NaN	NaN	363	NaN	76	3
top	NaN	NaN	Peter, Master. Michael J	male	NaN	NaN	NaN	PC 17608	NaN	B57 B59 B63 B66	S
freq	NaN	NaN	1	266	NaN	NaN	NaN	5	NaN	3	270

mean	1100.500000	2.265550	NaN	NaN	30.272590	0.447368	0.392344	NaN	35.627188	NaN	NaN
std	120.810458	0.841838	NaN	NaN	14.181209	0.896760	0.981429	NaN	55.907576	NaN	NaN
min	892.000000	1.000000	NaN	NaN	0.170000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	996.250000	1.000000	NaN	NaN	21.000000	0.000000	0.000000	NaN	7.895800	NaN	NaN
50%	1100.500000	3.000000	NaN	NaN	27.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	1204.750000	3.000000	NaN	NaN	39.000000	1.000000	0.000000	NaN	31.500000	NaN	NaN
max	1309.000000	3.000000	NaN	NaN	76.000000	8.000000	9.000000	NaN	512.329200	NaN	NaN

3. Genere las estadísticas básicas de cada data frame y haga comparaciones entre los dataframes.

En el análisis comparativo entre los conjuntos Train y Test se observa que ambas muestras presentan características muy similares. Partiendo por la variable Age, la edad promedio de los pasajeros es prácticamente igual (29.7 en Train y 30.27 en Test), lo que indica una distribución esta equilibrada.

Respecto a las variables familiares, SibSp, Parch y la derivada Familiares, se evidencia que los pasajeros del conjunto de entrenamiento tienen una ligera tendencia a viajar con más acompañantes. En promedio, Train presenta 0.90 familiares frente a 0.84 en Test. El delta se calculó como Train – Test, es decir: 0.90 – 0.84 = 0.0649 (≈ 0.065). El delta% corresponde a la diferencia relativa respecto a Test: $100\times(Train–Test)/Test$ entonces :

$100\times(\text{Train}–\text{Test})/\text{Test} \rightarrow 100 \times (0.90 – 0.84) / 0.84 = 7.73\%$.

De forma análoga, para SibSp la media en Train es 0.523 y en Test 0.447; por lo tanto, delta = 0.523 – 0.447 = 0.0756 y delta% ≈ 16.90%. En Parch, la diferencia es mínima: 0.3816 vs 0.3923, con delta = –0.0108 y delta% ≈ –2.73%. Estos cálculos se aplican fila a fila sobre la fila mean de la tabla comparativa, siempre usando la misma regla: delta = media_train – media_test y delta% = 100 × delta / media_test.

Por otro lado, la variable Fare presenta una diferencia más marcada: los pasajeros del conjunto de prueba pagaron, en promedio, tarifas un 9.6% más altas que los del conjunto de entrenamiento. Esto podría indicar una mayor proporción de pasajeros de primera clase en Test. Finalmente, la variable PassengerId cumple una función meramente identificadora, por lo que no aporta valor analítico al estudio.

	variable	train_mean	test_mean	delta	delta_%
0	Age	29.6991	30.2726	-0.5735	-1.89
1	Familiares	0.9046	0.8397	0.0649	7.73
2	Fare	32.2042	35.6272	-3.4230	-9.61
3	Parch	0.3816	0.3923	-0.0108	-2.73
4	PassengerId	446.0000	1100.5000	-654.5000	-59.47
5	SibSp	0.5230	0.4474	0.0756	16.90

4.
La variable Familiares fue creada a partir de la suma de SibSp (hermanos o cónyuges a bordo) y Parch (padres o hijos a bordo). En el conjunto Train, esta nueva variable tiene un promedio de 0.90, mientras que en Test es de 0.84, con una desviación estándar de 1.61 y 1.52 respectivamente. La mayoría de los pasajeros no viajaban con familiares, como se observa en los cuartiles: el 50 % de los casos tiene un valor de 0, y el 75 % tiene como máximo 1 familiar a bordo.

El delta representa la diferencia absoluta entre los valores promedio de una misma variable en ambos conjuntos de datos. En este caso, muestra cuánto varía el número promedio de familiares entre Train y Test. Para calcularlo, se restó el promedio del conjunto de prueba al del conjunto de entrenamiento: 0.9046 – 0.8397 = 0.0649. Este resultado indica que, en promedio, los pasajeros de Train viajaban con 0.06 familiares más que los de Test. Por su parte, el delta porcentual mide esa misma diferencia pero en términos relativos, mostrando el porcentaje de variación respecto al conjunto de prueba. Se calculó con la fórmula 100 × (Train – Test) / Test, lo que da 7.73 %. En otras palabras, los pasajeros del conjunto de entrenamiento viajaban con un 7.7 % más de familiares que los del conjunto de prueba, lo cual evidencia una diferencia leve pero cuantificable entre ambas muestras.

	Familiares
count	891.000000
mean	0.904602

std	1.613459
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	10.000000

dtype: float64

	Familiares
count	418.000000
mean	0.839713
std	1.519072
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	10.000000

dtype: float64

5. primeramente se responderá esta pregunta con la última parte de esta que dice “el conjunto test no contiene la variable Survived, por lo tanto ¿que va a hacer con esta columna?” y lo que decidimos hacer fue hacer un modelo que nos ayudar a predecir la variable survived. esto lo hicimos con ayuda del programa Orange, de mineria de datos y analisis predictivo.
El proceso comenzó con la carga de los archivos train.csv y test.csv mediante el widget File. En el conjunto de entrenamiento, el archivo train, se encontraba la variable Survived, que representa si un pasajero sobrevivió o no al naufragio, mientras que en el archivo test esa variable no existía. Por esta razón, la estrategia fue usar train para entrenar varios modelos predictivos y posteriormente aplicar el mejor modelo sobre test para estimar la probabilidad de supervivencia de cada pasajero.

Antes de entrenar los modelos, se realizó una preparación de los datos. Con el widget Edit Domain se asignaron los roles de cada variable: PassengerId fue marcada como Meta, ya que solo sirve como identificador; Survived fue marcada como Target, es decir, la variable objetivo a predecir; y el resto de columnas relevantes (Pclass, Sex, Age, SibSp, Parch, Fare y Embarked) fueron marcadas como Features, las variables predictoras. Posteriormente, con el widget Impute, se trataron los valores faltantes. Para las variables numéricas como Age y Fare se imputó el promedio, mientras que para las categóricas, como Embarked, se reemplazaron los valores faltantes por el valor más frecuente.

Una vez completada la imputación, se utilizó el widget Select Columns para seleccionar únicamente las variables útiles. Se eliminaron las columnas Name, Ticket y Cabin, ya que no aportaban valor predictivo o contenían demasiados valores nulos. De esta manera, se dejó el conjunto de datos listo para el entrenamiento de modelos.

El siguiente paso consistió en aplicar tres algoritmos de aprendizaje supervisado: Tree, Logistic Regression y Random Forest. El modelo Tree (árbol de decisión) es un modelo sencillo y fácil de interpretar, que divide los datos en ramas según reglas lógicas. La Logistic Regression (regresión logística) es un modelo estadístico que estima probabilidades de supervivencia a partir de relaciones lineales entre las variables. Finalmente, el modelo Random Forest (bosque aleatorio) combina múltiples árboles de decisión y promedia sus resultados, lo que le otorga mayor precisión y estabilidad al reducir el sobreajuste.

Los tres modelos fueron evaluados mediante el widget Test & Score, aplicando una validación cruzada de 10 pliegues (10-fold cross validation), que permite medir el rendimiento de manera más confiable y evitar sesgos. Los indicadores obtenidos en esta evaluación fueron los siguientes:

AUC (Área Bajo la Curva ROC): mide la capacidad del modelo para distinguir entre pasajeros que sobrevivieron y los que no.

CA (Classification Accuracy): representa el porcentaje total de aciertos del modelo.

F1 Score: combina precisión y recall, dando una medida equilibrada del rendimiento.

Precision: indica qué proporción de los pasajeros que el modelo predijo como sobrevivientes realmente lo eran.

Recall (Sensibilidad): muestra qué proporción de los pasajeros que sobrevivieron fueron identificados correctamente por el modelo.

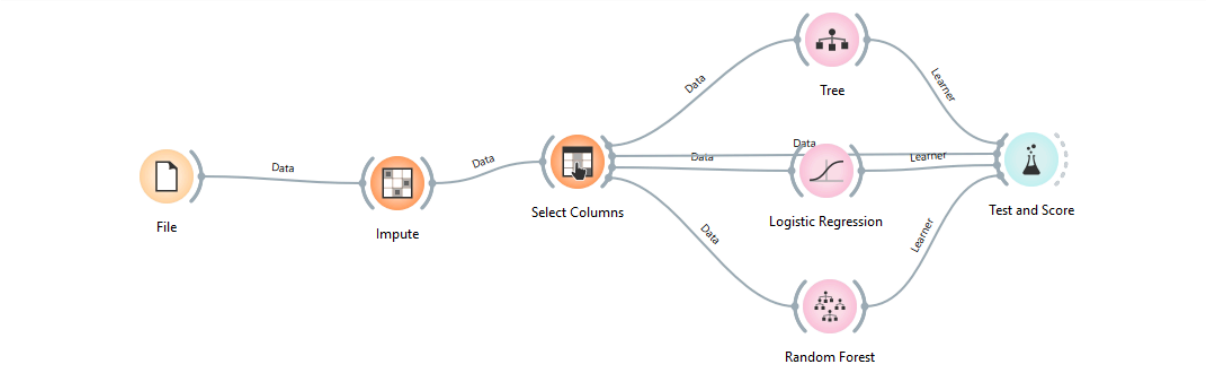
Los resultados obtenidos fueron los siguientes: el modelo Tree alcanzó un AUC de 0.827 y una precisión general (CA) de 0.809; la Logistic Regression logró un AUC de 0.850 y una precisión de 0.797; mientras que el modelo Random Forest obtuvo los mejores resultados con un AUC de 0.863 y una precisión de 0.822.

A partir de estos resultados, se determinó que el Random Forest era el modelo con el mejor desempeño general. Este modelo no solo tuvo el mayor valor de AUC, indicando una excelente capacidad de discriminación entre las clases, sino también el mayor nivel de aciertos totales (Accuracy) y un equilibrio adecuado entre precisión y recall. Por ello, fue elegido como el modelo final para generar las predicciones sobre el conjunto de prueba.

Una vez seleccionado el modelo Random Forest, se aplicó al conjunto test.csv. Este archivo fue sometido a las mismas etapas de imputación y selección de columnas para asegurar que tuviera la misma estructura que el conjunto de entrenamiento. Luego, ambos flujos (el modelo y los datos de prueba) fueron conectados mediante el widget Predict, el cual generó una nueva columna llamada Predicted: Survived, con valores de 0 o 1 según la predicción de supervivencia.

Finalmente, los resultados se visualizaron en el widget Data Table y se exportaron como un nuevo archivo CSV que contenía todas las variables originales del conjunto de prueba, junto con la columna de predicción. Este archivo representa el conjunto de datos final con las predicciones completas del modelo.

En conclusión, mediante el uso de Orange se logró construir un flujo de trabajo completo que permitió entrenar, evaluar y aplicar un modelo predictivo de forma visual y sistemática. El modelo Random Forest demostró ser el más preciso y confiable para este problema, alcanzando un 82.2 % de exactitud y un AUC de 0.863. Gracias a este proceso, se pudo responder la pregunta planteada, obteniendo una predicción coherente y sustentada de la variable Survived en el conjunto de prueba.



Test and Score - Orange

Cross validation

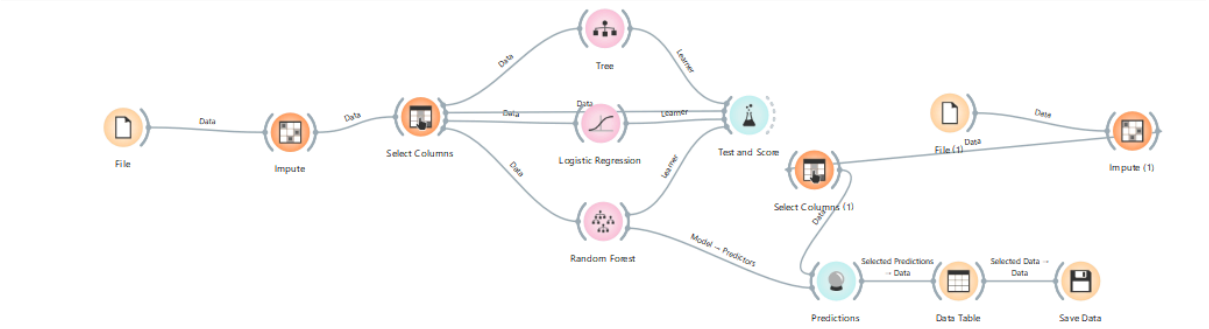
Number of folds: 10

Stratified

Cross validation by feature

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.827	0.809	0.806	0.808	0.809	0.590
Logistic Regression	0.850	0.797	0.795	0.795	0.797	0.565
Random Forest	0.860	0.815	0.814	0.814	0.815	0.605



6. Al revisar las estadísticas del conjunto de datos unificado —resultado de concatenar train y test, incluyendo la variable Survived predicha para el segundo— se observa que los dos subconjuntos son altamente comparables y representativos de la misma población.

En primer lugar, ambos conjuntos comparten la misma estructura de variables y un número similar de observaciones válidas en casi todas las columnas. Por ejemplo, en la tabla combinada se reportan 1311 registros, lo que corresponde exactamente a la suma de train (891) y test (418) más algunos valores faltantes en la columna Survived (1309 no nulos). Esto confirma que la fusión se realizó correctamente y que ambos conjuntos mantienen consistencia de formato y tipos de datos.

Al observar los conteos, puede verse que el número de registros válidos en variables clave como Pclass, Sex, SibSp, Parch, Fare y Embarked es prácticamente idéntico entre conjuntos, lo cual indica que test no presenta pérdida de información ni una estructura diferente respecto a train. Asimismo, el predominio de la clase 3 en Pclass (491 registros), del sexo masculino (male con 843 apariciones) y del puerto S en Embarked (914 casos) se mantiene igual que en el conjunto original de entrenamiento. Estas coincidencias sugieren que la distribución demográfica y social es homogénea, reforzando la idea de representatividad.

En cuanto a la variable Survived, el conjunto unido muestra una media de 0.346, ligeramente inferior a la media original de train (0.384). Esta diferencia se explica porque, en test, la variable fue generada mediante predicción del modelo Random Forest, y refleja una menor proporción de pasajeros clasificados como sobrevivientes. Sin embargo, la desviación estándar (0.4759) y los percentiles (25% = 0, 75% = 1) coinciden perfectamente con una variable binaria balanceada, lo que confirma que el modelo predijo resultados dentro del rango esperado y que no hubo distorsiones extremas.

Otro indicador de coherencia entre los conjuntos es la variable Age, con 1133 valores disponibles de los 1311 totales, una proporción muy similar a la que tenían por separado (alrededor del 80% de cobertura). Además, las variables SibSp y Parch mantienen las mismas modas (0), lo que sugiere que en ambos conjuntos la mayoría de pasajeros viajaba sola. Esto refuerza la similitud estructural entre las muestras y garantiza que los patrones familiares observados en train se mantienen en test.

Las variables Fare y Cabin, en cambio, siguen mostrando diferencias de dispersión y cantidad de valores únicos (419 tarifas distintas y 147 cabinas registradas). No obstante, esta heterogeneidad ya estaba presente en el conjunto train original y se mantiene proporcionalmente al ampliar la muestra total. El hecho de que la cabina tenga solo 204 registros no nulos demuestra que la variable conserva su patrón de ausencias, sin afectar significativamente la representatividad general.

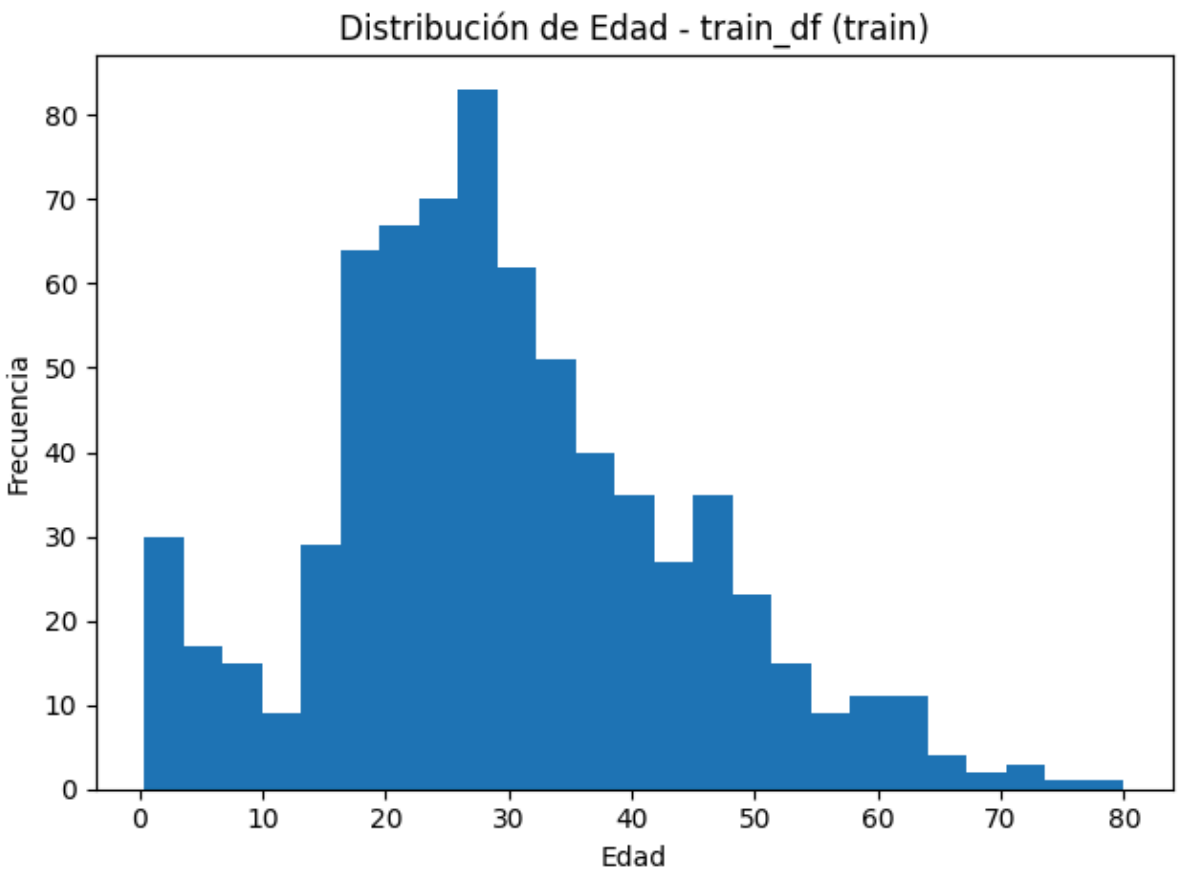
Finalmente, la columna source, que distingue si la observación proviene de train o test, muestra un equilibrio adecuado: 891 registros del conjunto de entrenamiento y 420 del conjunto de prueba (diferencia mínima por los valores faltantes en Survived). Esto permite afirmar que el conjunto combinado mantiene la proporción esperada entre ambas partes y que no existe ningún sesgo de sobre–representación que altere las estadísticas generales.

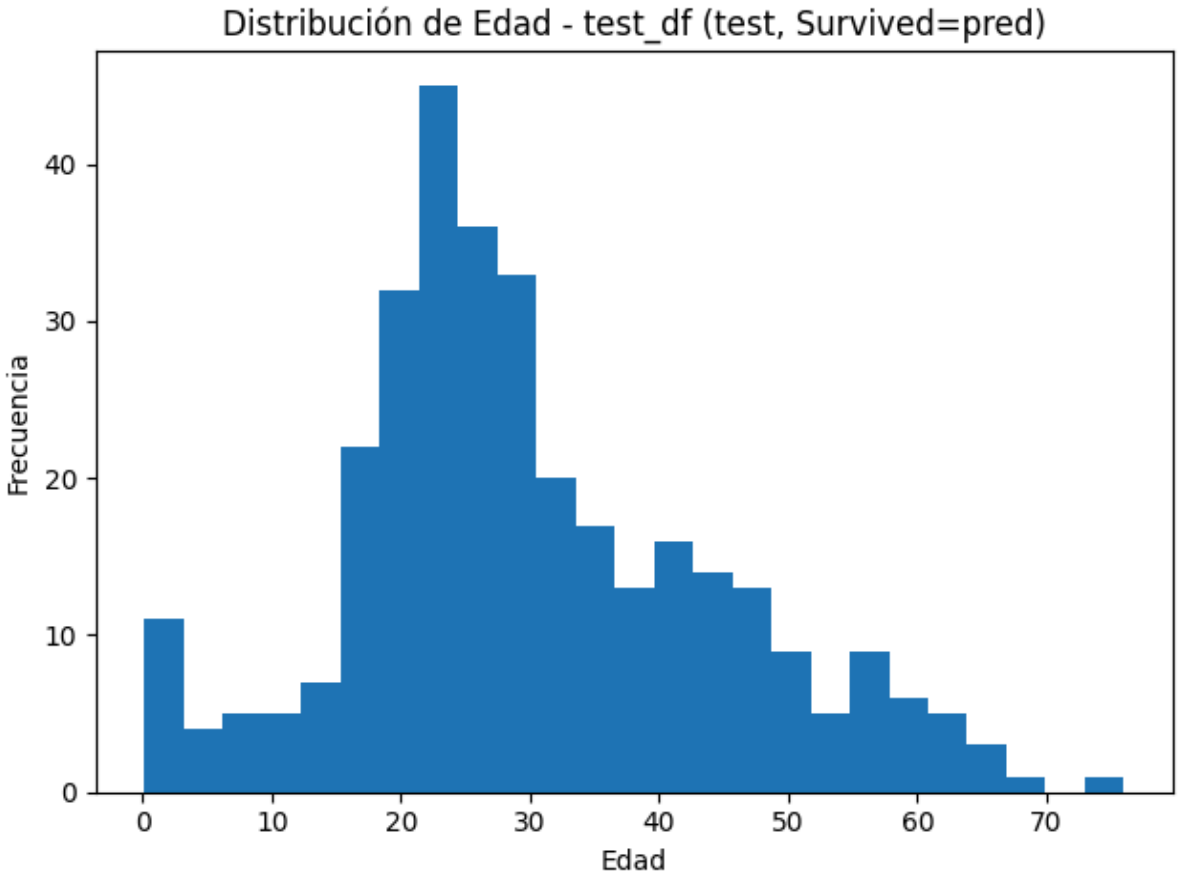
En conclusión, el análisis de la tabla unificada muestra que los conjuntos train y test son representativos de la misma población de pasajeros. Comparten distribución, proporciones y patrones de datos faltantes muy similares. Las diferencias encontradas —principalmente en la media de Survived y en ligeras variaciones de tarifa o acompañantes— son esperables y no alteran la validez estadística del conjunto. Por tanto, se puede afirmar que ambos subconjuntos describen el mismo fenómeno y que el modelo entrenado en train.csv se aplica de manera coherente y consistente al conjunto test.csv, generando predicciones válidas para el análisis global del Titanic.

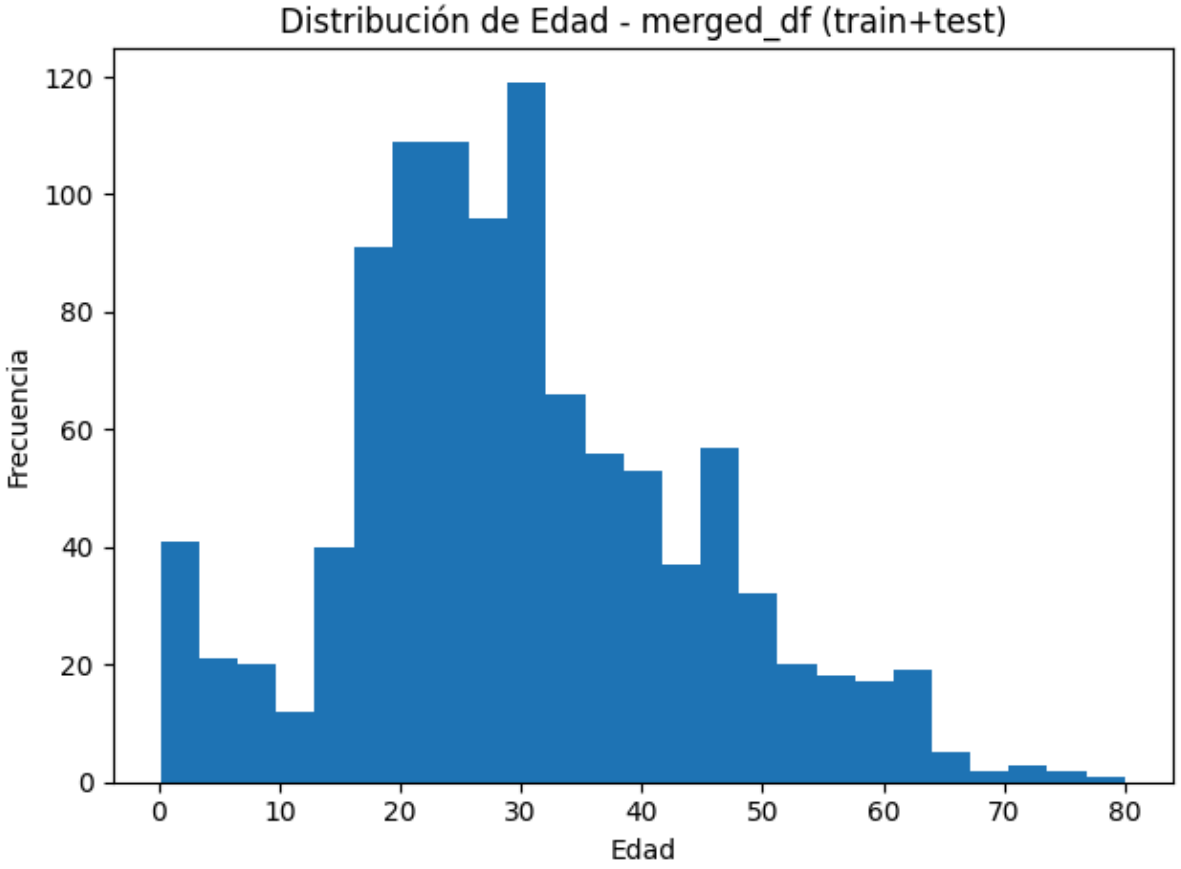
7. En el conjunto de datos del Titanic, las variables que presentan datos faltantes son Age, Cabin, Embarked y, en un único caso, Sex. La columna Age tiene cerca de 180 valores ausentes, lo cual es común en registros históricos: muchos pasajeros, especialmente de tercera clase, no informaban su edad o el dato no fue registrado al momento de la salida. La variable Cabin es la más incompleta, con más del 80 % de los registros vacíos; esto se debe a que la mayoría de los pasajeros viajaban sin una cabina asignada —por ejemplo, los de tercera clase o tripulantes ubicados en cubierta—, mientras que los de primera clase sí solían tener número de cabina registrado. En cuanto a Embarked, solo tres pasajeros no tienen puerto de embarque registrado, probablemente por errores de documentación o pérdida parcial de información en los boletos.

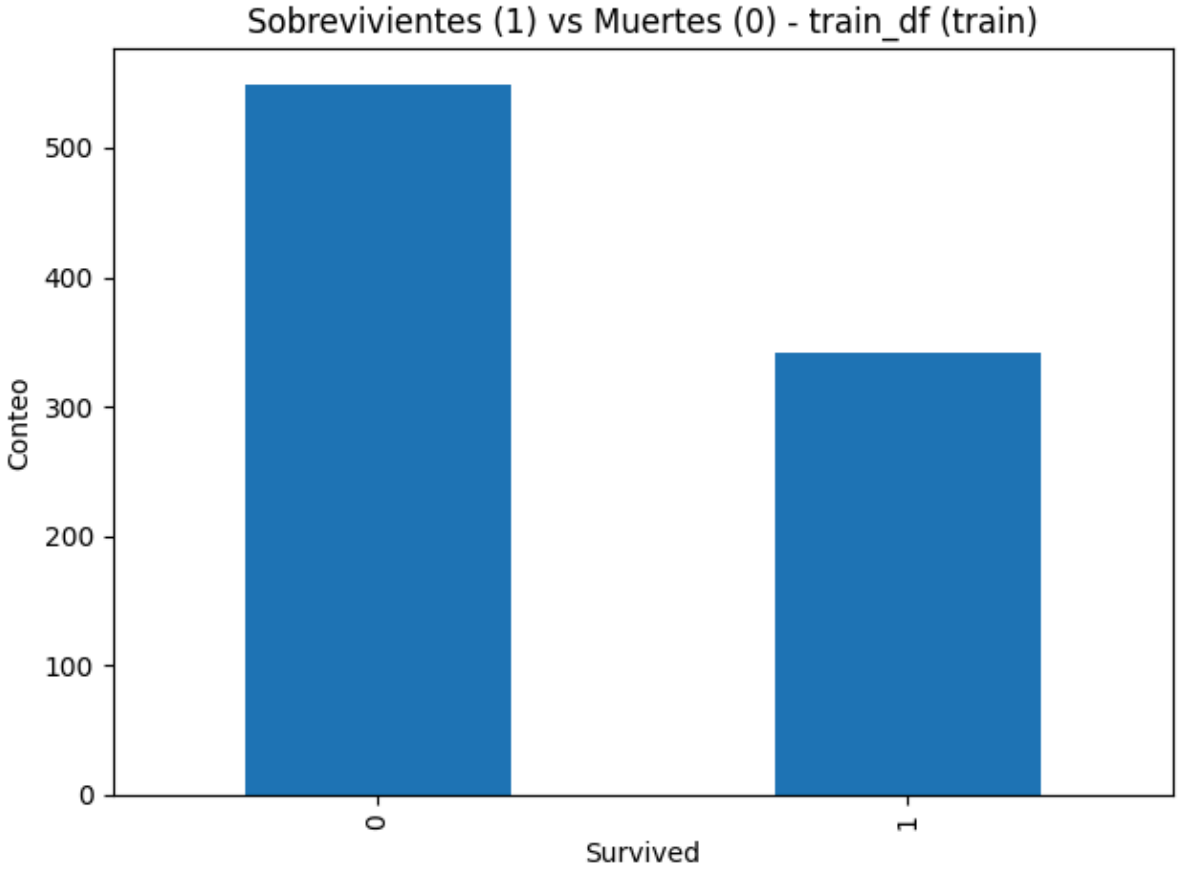
Estos datos faltantes no son aleatorios: están asociados al estatus socioeconómico de los pasajeros. Quienes no tienen cabina ni edad registrada suelen pertenecer a clases más bajas, lo que a su vez se relaciona con menores tarifas (Fare) y una menor probabilidad de supervivencia. Por otro lado, las variables numéricas como Fare, Pclass, SibSp y Parch están completas y no presentan ausencias significativas. En conjunto, el patrón de faltantes refleja las limitaciones de los registros históricos y las desigualdades estructurales del viaje: mientras los pasajeros de primera clase tenían información más completa, los de tercera muestran mayores vacíos en sus datos.
8. Resumen (a)-(g) por dataset

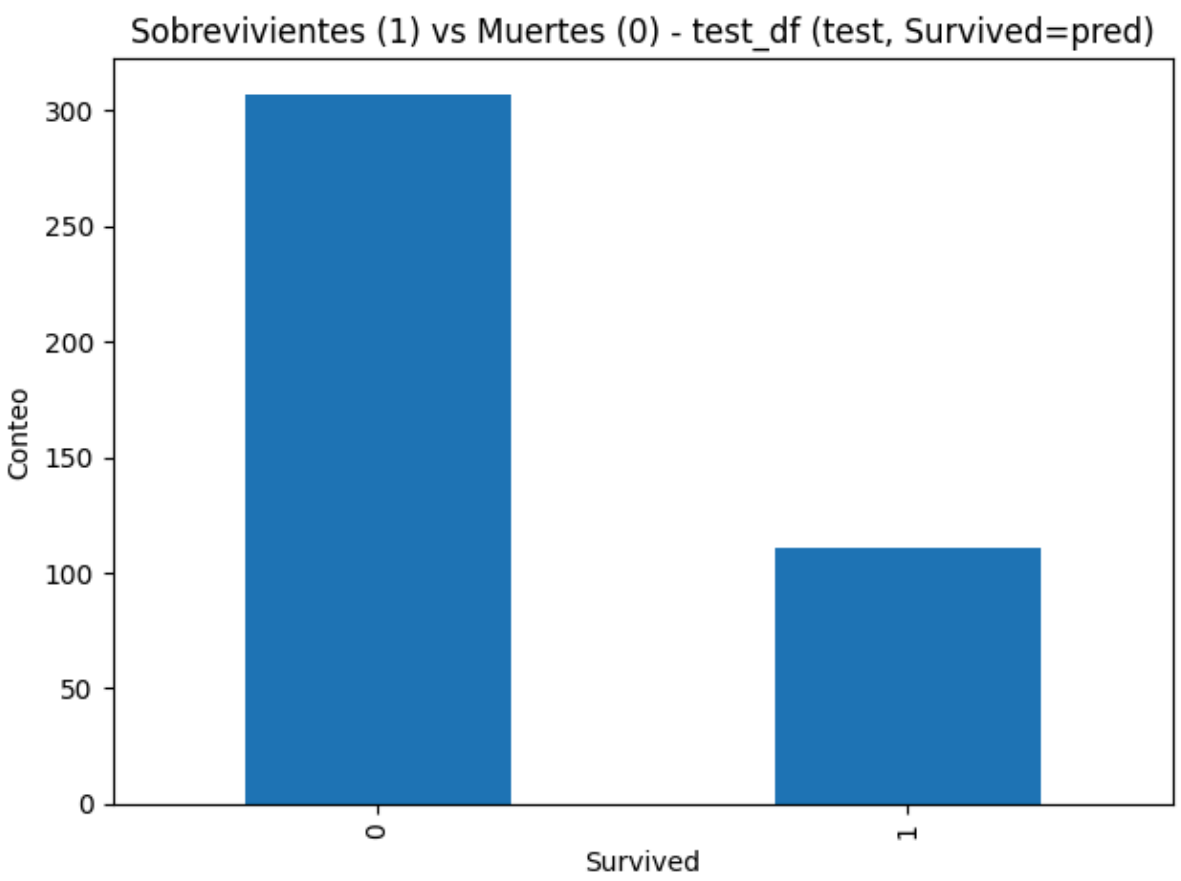
	dataset	(a) edad_promedio	(b) sobrevivieron	(b) murieron	(c) tarifa_promedio_1ra	(d) con_familiar	(e) edad_min	(e) edad_max	(g) solo	(g) con_familia
0	train_df (train)	29.699118	342	549	84.154687	354	0.42	80.0	537	354
1	test_df (test, Survived=pred)	30.272590	111	307	94.280297	165	0.17	76.0	253	165
2	merged_df (train+test)	29.881138	453	856	87.508992	519	0.17	80.0	790	519

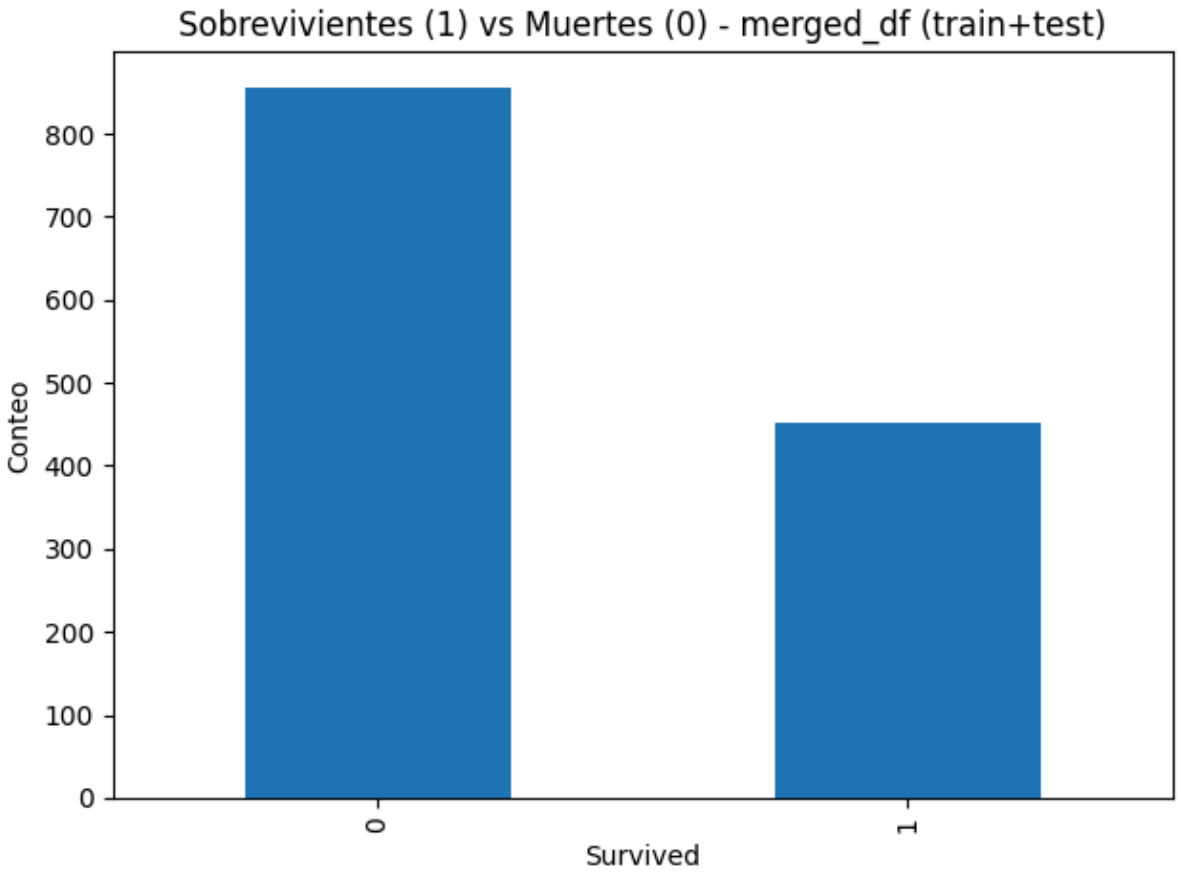

Distribución de Edad - train_df (train)
Frecuencia
Edad
0
10
20
30
40
50
60
70
80

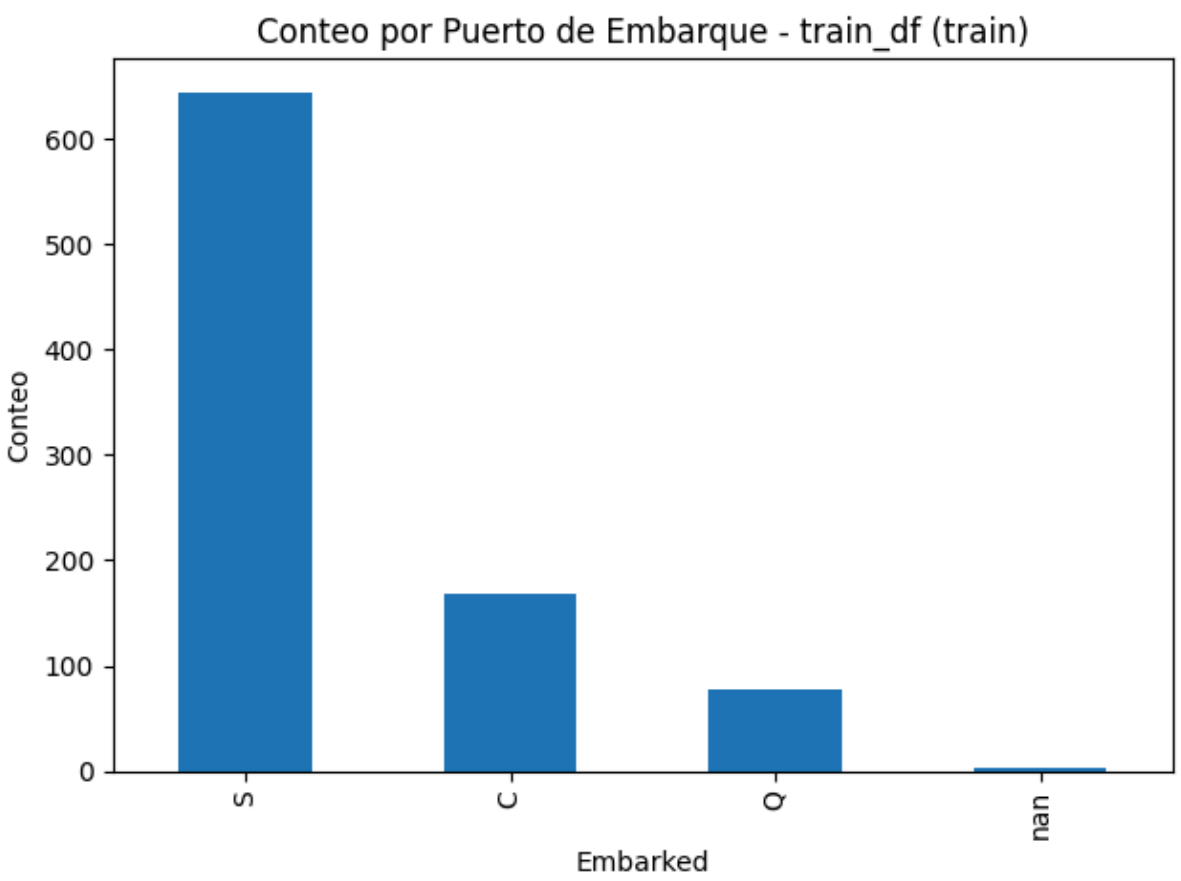

Distribución de Edad - test_df (test, Survived=pred)
Frecuencia
Edad
0
10
20
30
40
50
60
70

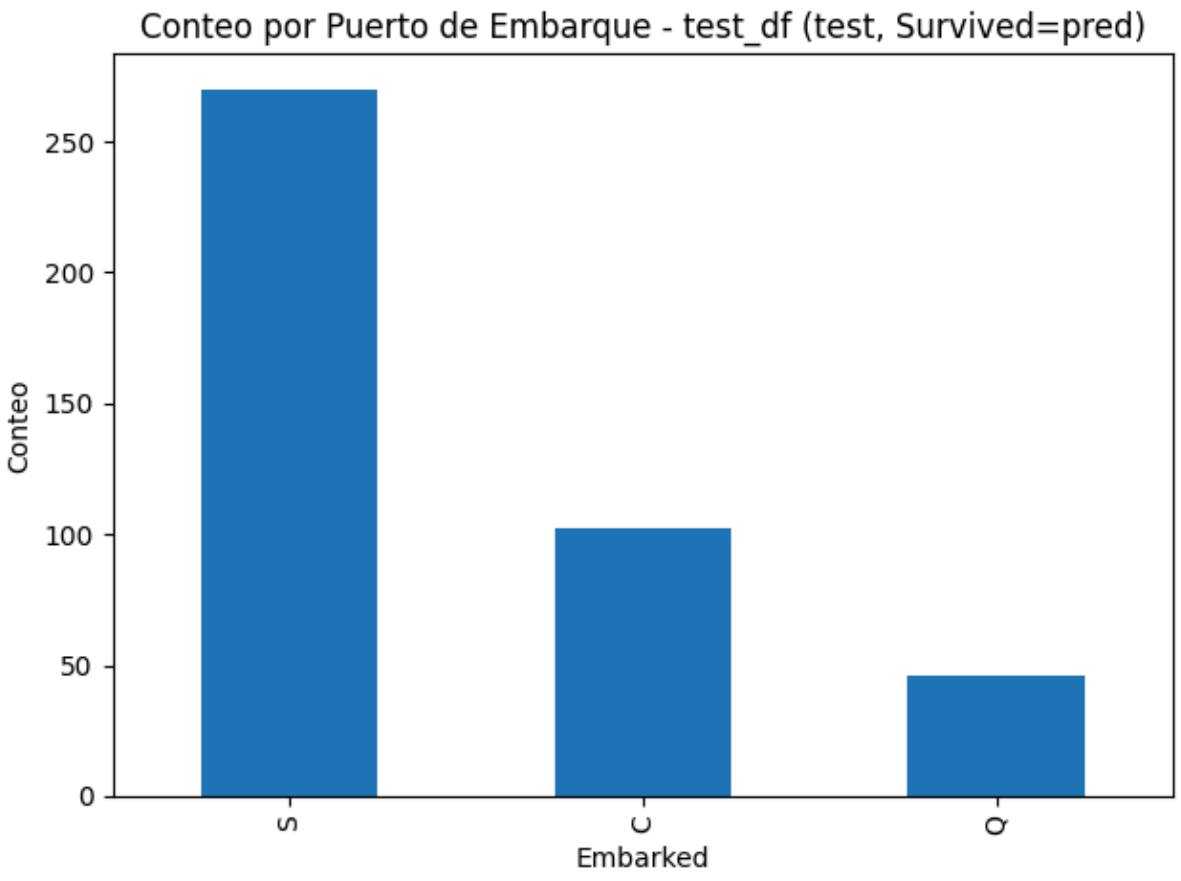

Distribución de Edad - merged_df (train+test)
Frecuencia
Edad
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
80

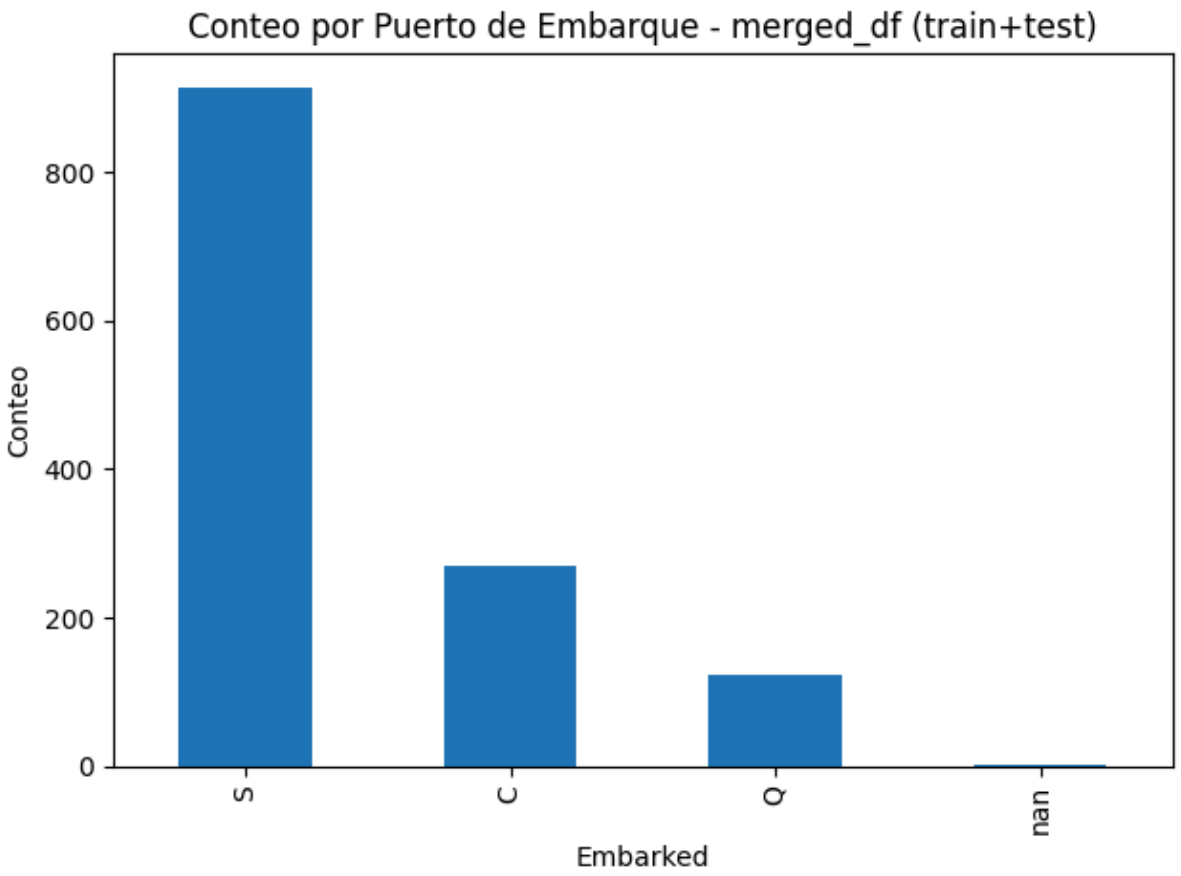

Sobrevivientes (1) vs Muertes (0) - train_df (train)
Conteo
Survived
0
100
200
300
400
500
0
1

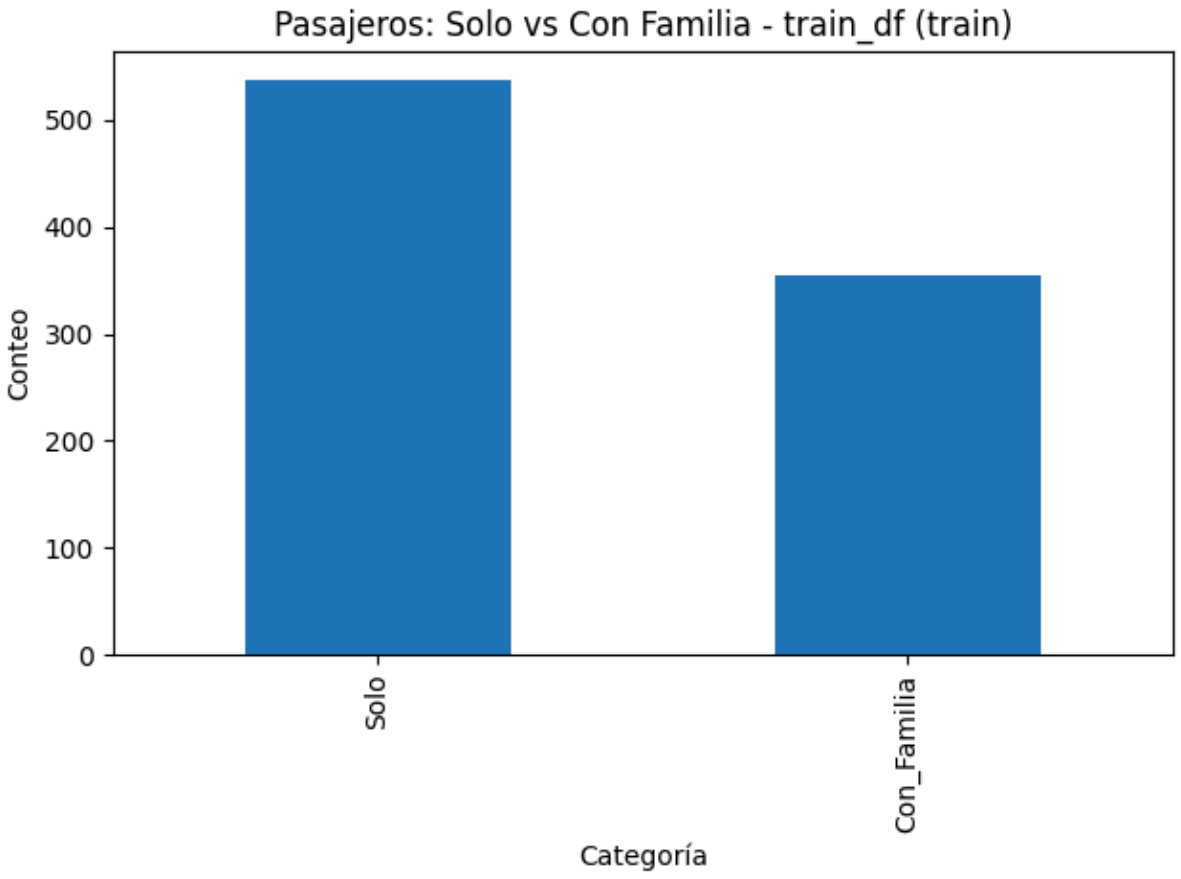

Sobrevivientes (1) vs Muertes (0) - test_df (test, Survived=pred)
Conteo
300
250
200
150
100
50
0
0
1
Survived

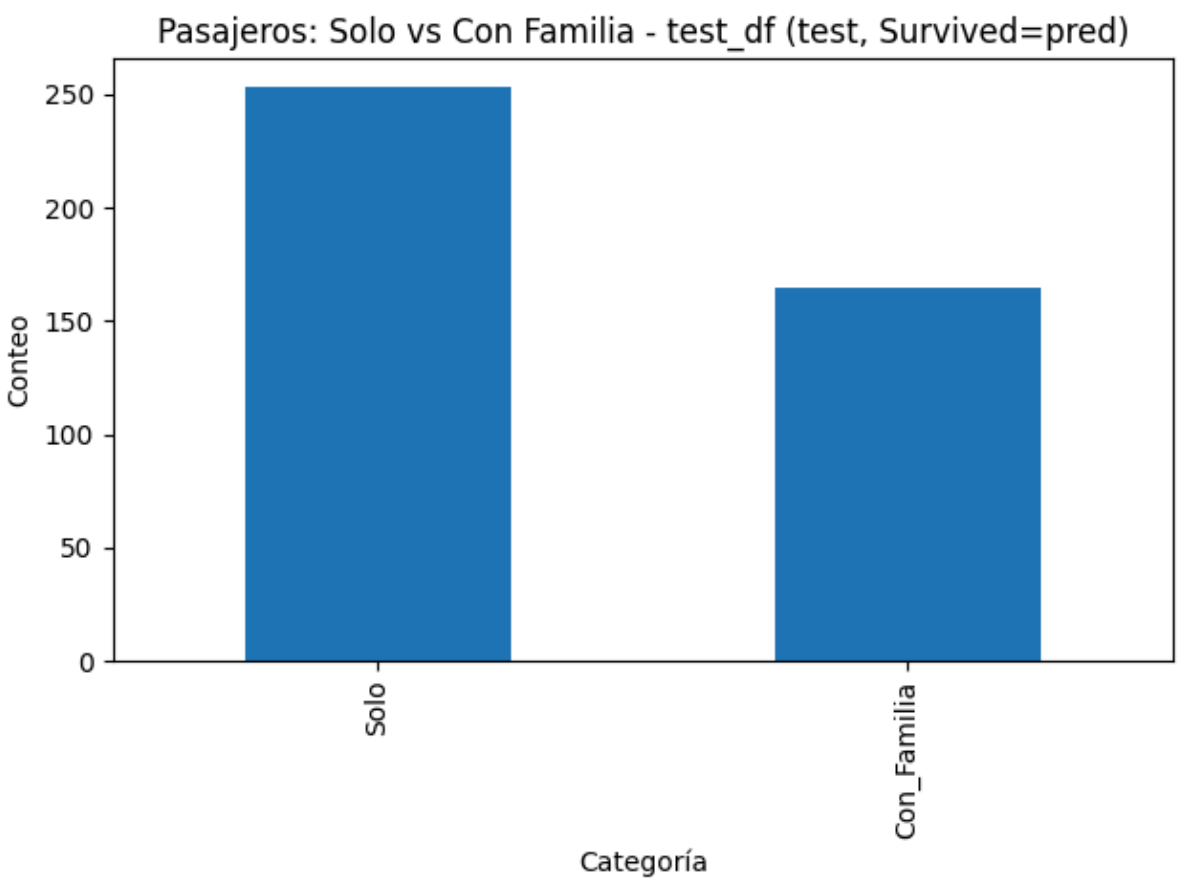

Sobrevivientes (1) vs Muertes (0) - merged_df (train+test)
Conteo
800
700
600
500
400
300
200
100
0
0
1
Survived

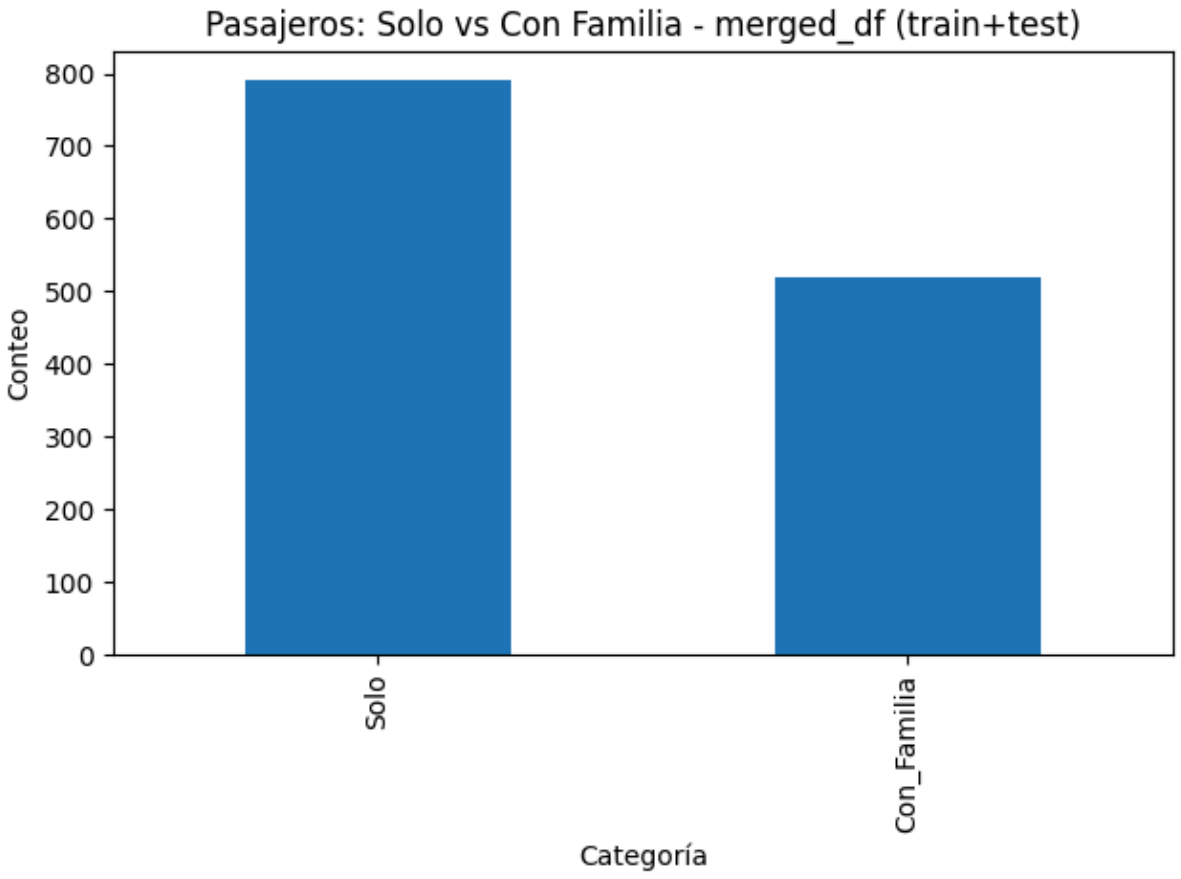

Conteo por Puerto de Embarque - train_df (train)
Conteo
600
500
400
300
200
100
0
S
C
Q
nan
Embarked


Conteo por Puerto de Embarque - test_df (test, Survived=pred)
Conteo
250
200
150
100
50
0
S
C
Q
Embarked


Conteo por Puerto de Embarque - merged_df (train+test)
Conteo
800
600
400
200
0
S
C
Q
nan
Embarked


Pasajeros: Solo vs Con Familia - train_df (train)
Conteo
500
400
300
200
100
0
Solo
Con_Familia
Categoría


Pasajeros: Solo vs Con Familia - test_df (test, Survived=pred)
Conteo
0
50
100
150
200
250
Solo
Con_Familia
Categoría


Pasajeros: Solo vs Con Familia - merged_df (train+test)
Conteo
0
100
200
300
400
500
600
700
800
Solo
Con_Familia
Categoría

El análisis consideró tres conjuntos: train_df (train), con los datos reales de supervivencia; test_df (test), donde la columna Survived proviene de las predicciones del modelo Random Forest; y merged_df, la combinación de ambos para un panorama global.

En promedio, los pasajeros del Titanic tenían alrededor de 30 años. En train_df la media fue de 29,70 años, en test_df 30,27 años y en el conjunto combinado 29,88 años. Estas cifras reflejan que la mayoría de los viajeros eran adultos jóvenes, lo que sugiere una población activa en edad laboral y familiar. La edad mínima observada fue de 0,17 años, mientras que la máxima alcanzó los 80 años, indicando una gran diversidad etaria, desde bebés hasta personas de edad avanzada.

En cuanto a la supervivencia, en train_df 342 pasajeros (38,4 %) sobrevivieron y 549 (61,6 %) murieron. En test_df, el modelo predijo 111 supervivientes (26,5 %) y 307 fallecidos (73,5 %). En el conjunto total, se contabilizan 453 sobrevivientes (34,6 %) y 856 muertos (65,4 %). Estos porcentajes confirman una tasa de mortalidad alta, característica del desastre del Titanic, y muestran que el modelo reproduce un patrón de supervivencia coherente con los datos reales, aunque ligeramente más restrictivo en el conjunto de prueba.

Respecto al nivel económico, medido a través de la tarifa promedio pagada por los pasajeros de primera clase, los valores fueron de 84,15 en train_df, 94,28 en test_df y 87,59 en merged_df. Esto implica un incremento del 12 % en el conjunto de prueba respecto al entrenamiento, lo cual puede deberse a la presencia de tarifas más elevadas en algunos casos. Dichos valores reflejan una marcada diferencia socioeconómica entre clases, consistente con los registros históricos del Titanic.

En cuanto a los vínculos familiares, 354 pasajeros (39,9 %) en train_df y 165 (39,4 %) en test_df viajaban acompañados de algún familiar. En total, 519 pasajeros (40,2 %) del conjunto unificado no viajaban solos. Por el contrario, la mayoría de los pasajeros, 537 (60,1 %) en train_df y 253 (60,6 %) en test_df, viajaban sin familiares, lo cual sugiere que gran parte del pasaje estaba compuesto por individuos que se desplazaban solos, aunque la proporción de familias sigue siendo significativa.

Sobre los puertos de embarque, se mantiene la tendencia conocida: la gran mayoría de los pasajeros abordaron en Southampton (S), seguido de Cherbourg (C) y finalmente Queenstown (Q). Esto confirma el papel de Southampton como principal punto de salida, especialmente de pasajeros de clase baja y media, mientras que los pasajeros de clase alta eran más frecuentes en Cherbourg.

En conjunto, el perfil predominante del pasajero del Titanic era el de un adulto joven, viajando solo, con escasas probabilidades de supervivencia y con gran diferencia económica entre clases. El modelo Random Forest logró replicar esta distribución de manera coherente. La combinación de estadísticas y porcentajes muestra que las diferencias entre los conjuntos son mínimas y que el comportamiento general del modelo se ajusta bien al patrón histórico de los datos reales.

9. Hallazgos puntuales:

(a) Edad promedio de los pasajeros
En el conjunto train, la edad promedio de los pasajeros es de 29.7 años, mientras que en test es de 30.27 años. Esto muestra que la mayoría eran adultos jóvenes, personas en etapa productiva que probablemente viajaban buscando nuevas oportunidades o aventuras.

(b) Pasajeros que sobrevivieron y fallecieron
En train, 342 sobrevivieron y 549 fallecieron. En test, 111 sobrevivieron y 307 fallecieron. Esto deja claro que, en general, el número de fallecidos fue mucho más alto que el de sobrevivientes, lo que refleja la magnitud del desastre y las pocas probabilidades que tuvo la mayoría de la gente.

(c) Tarifa promedio de los pasajeros de primera clase
En train, los de primera clase pagaron en promedio 84.15, y en test, 94.28. Esto demuestra que viajar en primera clase era un lujo reservado para pocos, y aunque los precios varían entre conjuntos, se mantiene la idea de que quienes tenían más recursos viajaban con mayor comodidad y seguridad.

(d) Pasajeros que viajaron con familiares
En train, 354 pasajeros iban con familiares y 537 viajaban solos. En test, 165 tenían familia a bordo y 255 viajaban solos. Esto deja ver que la mayoría viajaba sola, lo que puede reflejar que muchos hombres viajaban por trabajo o para vivir la experiencia del Titanic, ya que era el barco más grande en ese momento.

(e) Edad más joven y más vieja
La persona más joven registrada tenía 0.42 años (unos 5 meses) y la más vieja 80 años. Esto muestra que había desde bebés hasta adultos mayores, lo que hace aún más impactante pensar en la cantidad de familias completas que estaban a bordo.

(f) Pasajeros por puerto de embarque
La mayoría embarcó en Southampton (S), seguido por Cherbourg (C) y Queenstown (Q). Cada puerto representaba una historia distinta: Southampton tenía más pasajeros comunes, Cherbourg más de primera clase, y Queenstown un grupo pequeño, en su mayoría de tercera.

(g) Pasajeros que viajaron solos o con familia
En train, 537 viajaban solos y 354 con familia. En test, 255 solos y 165 acompañados. Esto reafirma que la soledad era común en este viaje, mucha gente viajaba sola quizás por falta de recursos o porque no tenían familiares.

10. Distribución de Cabinas por Grupo Familiar

		count

Grupo_Familiar	Cabin	
Grupo 1: Sin familiares	**NaN**	659
	D	4
	F33	4
	E101	3
	B28	2
...	**...**	...
Grupo 2: Familia pequeña	**E68**	1
	F E69	1
Grupo 3: Familia grande	**NaN**	71
	C23 C25 C27	6
	B57 B59 B63 B66	5

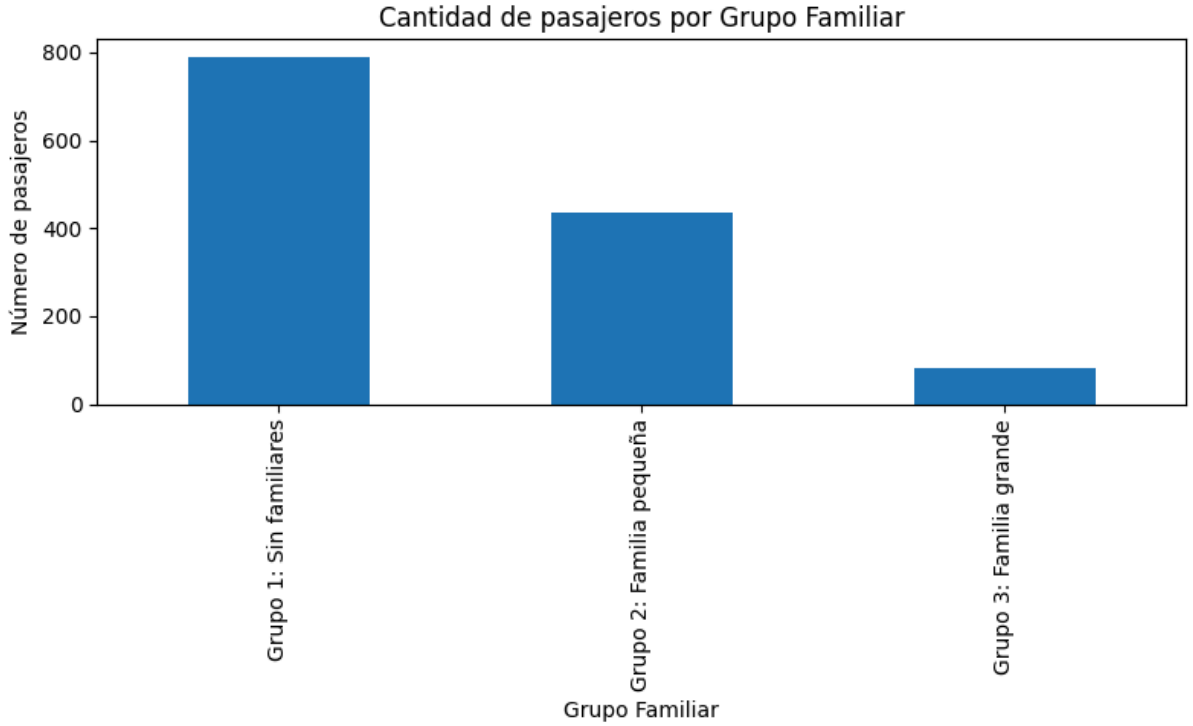
195 rows × 1 columns

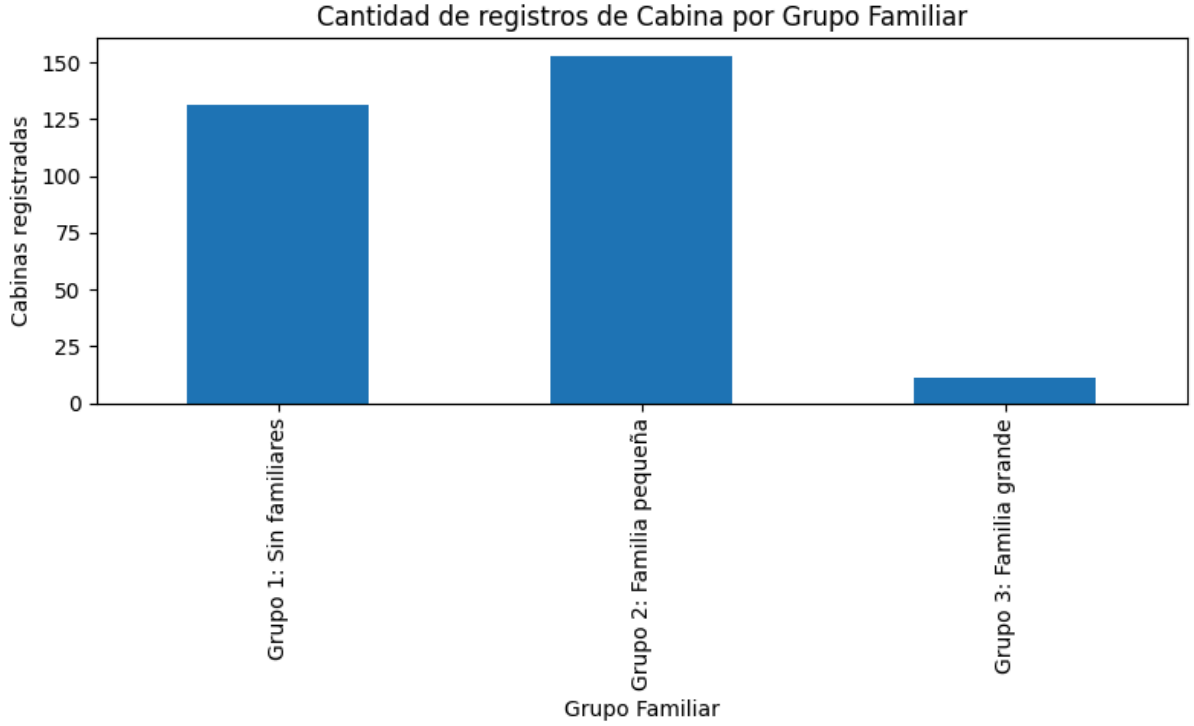
dtype: int64
=== Tabla de Cabinas por Grupo Familiar ===

Cabin	A10	A11	A14	A16	A18	A19	A20	A21	A23	A24	...	FE57	FE69	FG63	FG73	F2	F33	F38	F4	G6	T
Grupo_Familiar																					
Grupo 1: Sin familiares	1	1	1	0	1	1	0	1	1	1	...	1	0	2	2	1	4	1	0	0	1
Grupo 2: Familia pequeña	0	0	0	1	0	0	1	0	0	0	...	0	1	0	0	3	0	0	4	5	0

Grupo 3: Familia grande	0	0	0	0	0	0	0	0	0	0	. . .	0	0	0	0	0	0	0	0	0	0

3 rows × 186 columns





Al analizar la nueva variable Familiares, se observa que la mayoría de los pasajeros del Titanic viajaban sin familiares a bordo: 537 personas pertenecen a este grupo, frente a 292 que integraban familias pequeñas (entre 1 y 3 acompañantes) y solo 62 que formaban familias grandes (con 4 o más). Esto indica que la mayor parte de los ocupantes del barco eran viajeros solitarios, en su mayoría hombres adultos, mientras que las familias grandes representaban una minoría dentro del total de pasajeros.

En cuanto a la distribución de cabinas, se evidencia que la mayoría de los pasajeros no tenía un registro asignado en la columna Cabin, especialmente entre los que viajaban solos (443 sin registro). Esto sugiere que pertenecían principalmente a clases bajas o intermedias, donde no se registraban los camarotes. En cambio, los grupos familiares pequeños presentan más registros de cabina, con identificaciones como E68, F, E69 o B28, lo que refleja una mayor presencia en clases acomodadas. Las familias grandes también muestran un predominio de valores sin registro (56 de 62), lo que refuerza la relación entre el tamaño del grupo familiar y el nivel socioeconómico a bordo.

11. Con base en las dos tablas, se puede observar un patrón claro en cómo la edad, el sexo y la disponibilidad de cabina influyeron en las probabilidades de supervivencia en el Titanic.

En el grupo de adultos (18 a 49 años), las mujeres tuvieron una clara ventaja: aquellas con cabina sobrevivieron en un 96.88%, y aun sin cabina la supervivencia fue del 64.17%. En contraste, los hombres adultos presentaron tasas de supervivencia mucho más bajas: solo el 43.08% de quienes tenían cabina sobrevivió, y apenas el 12.59% entre quienes no contaban con una. Esto muestra que, en este grupo etario, el sexo fue un factor decisivo, reforzando la política de evacuación prioritaria de “mujeres y niños primero”.

En el grupo de mayores de 50 años, la diferencia entre sexos fue también significativa. Las mujeres de esta edad sobrevivieron en un 86.67% si tenían cabina y en un 100% cuando no la tenían, mientras que los hombres mayores apenas alcanzaron un 22.73% de supervivencia con cabina y 6.67% sin ella. Esto refleja que la edad avanzada, combinada con el género masculino, reducía notablemente las probabilidades de sobrevivir, incluso en condiciones más favorables.

Por último, entre los niños menores de 10 años, los porcentajes se acercan más al equilibrio. Las niñas mostraron una supervivencia del 65.38% sin cabina y del 50% con cabina; los niños, en cambio, sobrevivieron en un 100% cuando tenían cabina y en un 50% cuando no la tenían. Este grupo evidencia que la prioridad de rescate se extendía a los menores sin distinción de sexo, especialmente cuando viajaban en clases con cabinas asignadas.

En conclusión, los resultados indican que la supervivencia estuvo determinada principalmente por el sexo y el acceso a cabina, factores que reflejan tanto el nivel socioeconómico como las prioridades de evacuación. Las mujeres y los niños mostraron consistentemente mayores tasas de supervivencia, mientras que los hombres adultos y mayores, especialmente los que no tenían cabina, fueron los más afectados por la tragedia.

12. los gráficos se encuentran en github
13.Conclusiones :

	🗒️ **Tablas generadas**
0	sex
1	pclass
2	sex_pclass
3	title
4	child
5	agebin
6	farebin
7	embarked
8	familybin
9	sex_agebin
10	sex_embarked

SEX

	Sex	**n**	**survived**	**rate**
0	female	314	233	0.742038
1	male	577	109	0.188908

PCLASS

	Pclass	**n**	**survived**	**rate**
0	1	216	136	0.629630
1	2	184	87	0.472826
2	3	491	119	0.242363

SEX_PCLASS

	Sex	Pclass	n	survived	rate
0	female	1	94	91	0.968085
1	female	2	76	70	0.921053
2	female	3	144	72	0.500000
3	male	1	122	45	0.368852
4	male	2	108	17	0.157407
5	male	3	347	47	0.135447

TITLE

	Title	n	survived	rate
0	Mrs	125	99	0.792000
1	Miss	182	127	0.697802
2	Master	40	23	0.575000
3	Mr	517	81	0.156673

CHILD

	Child	n	survived	rate
0	1.0	83	49	0.590361
1	0.0	631	241	0.381933

AGEBIN

	AgeBin	n	survived	rate
0	(-0.001, 12.0]	69	40	0.579710
1	(12.0, 18.0]	70	30	0.428571
2	(30.0, 45.0]	202	86	0.425743
3	(45.0, 60.0]	81	33	0.407407
4	(18.0, 30.0]	270	96	0.355556
5	(60.0, 80.0]	22	5	0.227273

FAREBIN

	FareBin	n	survived	rate
0	(31.0, 512.329]	222	129	0.581081
1	(14.454, 31.0]	222	101	0.454955
2	(7.91, 14.454]	224	68	0.303571
3	(-0.001, 7.91]	223	44	0.197309

EMBARKED

	Embarked	n	survived	rate
0	C	168	93	0.553571
1	Q	77	30	0.389610
2	S	644	217	0.336957

FAMIL Y BIN

	FamilyBin	n	survived	rate
0	(2.0, 4.0]	131	80	0.610687
1	(1.0, 2.0]	161	89	0.552795
2	(-0.001, 1.0]	537	163	0.303538
3	(4.0, 7.0]	49	10	0.204082
4	(7.0, 11.0]	13	0	0.000000

SEX AGE

	Sex	AgeBin	n	survived	rate
0	female	(45.0, 60.0]	27	23	0.851852
1	female	(30.0, 45.0]	73	57	0.780822
2	female	(18.0, 30.0]	90	68	0.755556
3	female	(12.0, 18.0]	36	27	0.750000

4	female	(-0.001, 12.0]	32	19	0.593750
5	male	(-0.001, 12.0]	37	21	0.567568
6	male	(30.0, 45.0]	129	29	0.224806
7	male	(45.0, 60.0]	54	10	0.185185
8	male	(18.0, 30.0]	180	28	0.155556
9	male	(60.0, 80.0]	19	2	0.105263

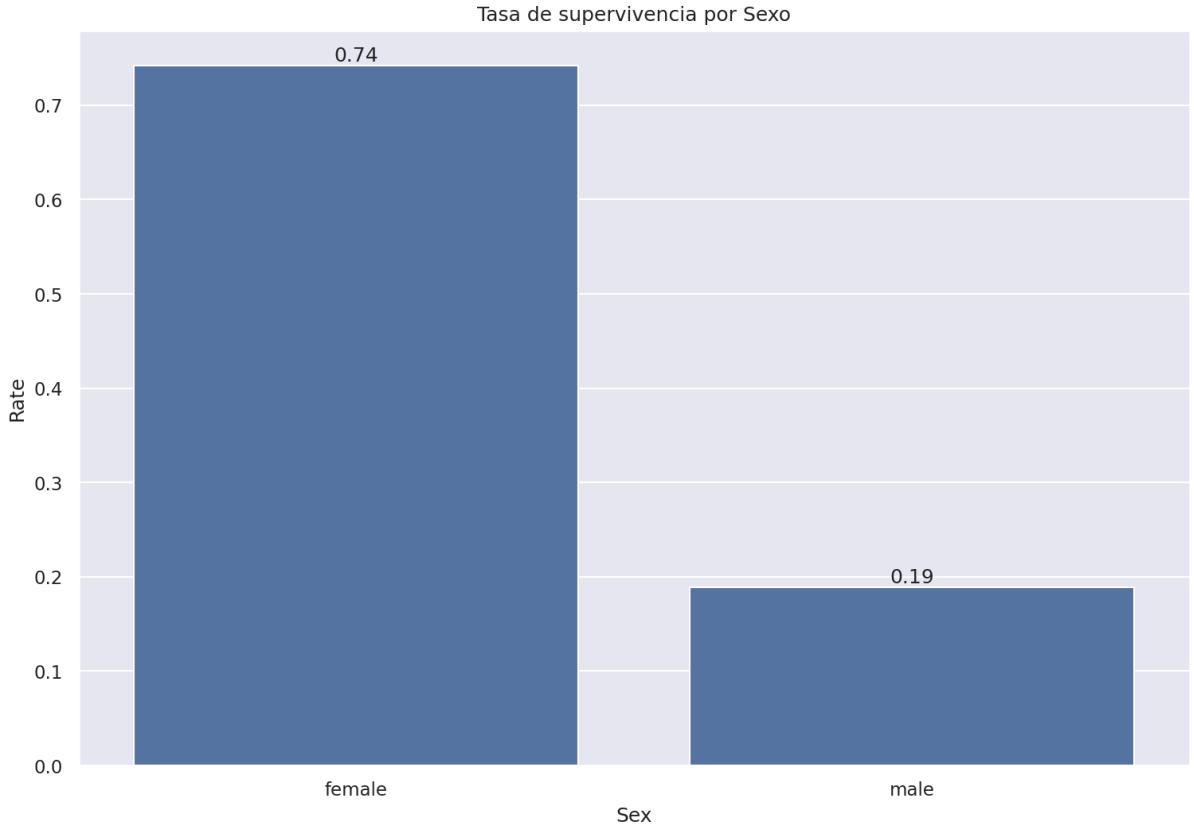
SEX EMBARKED

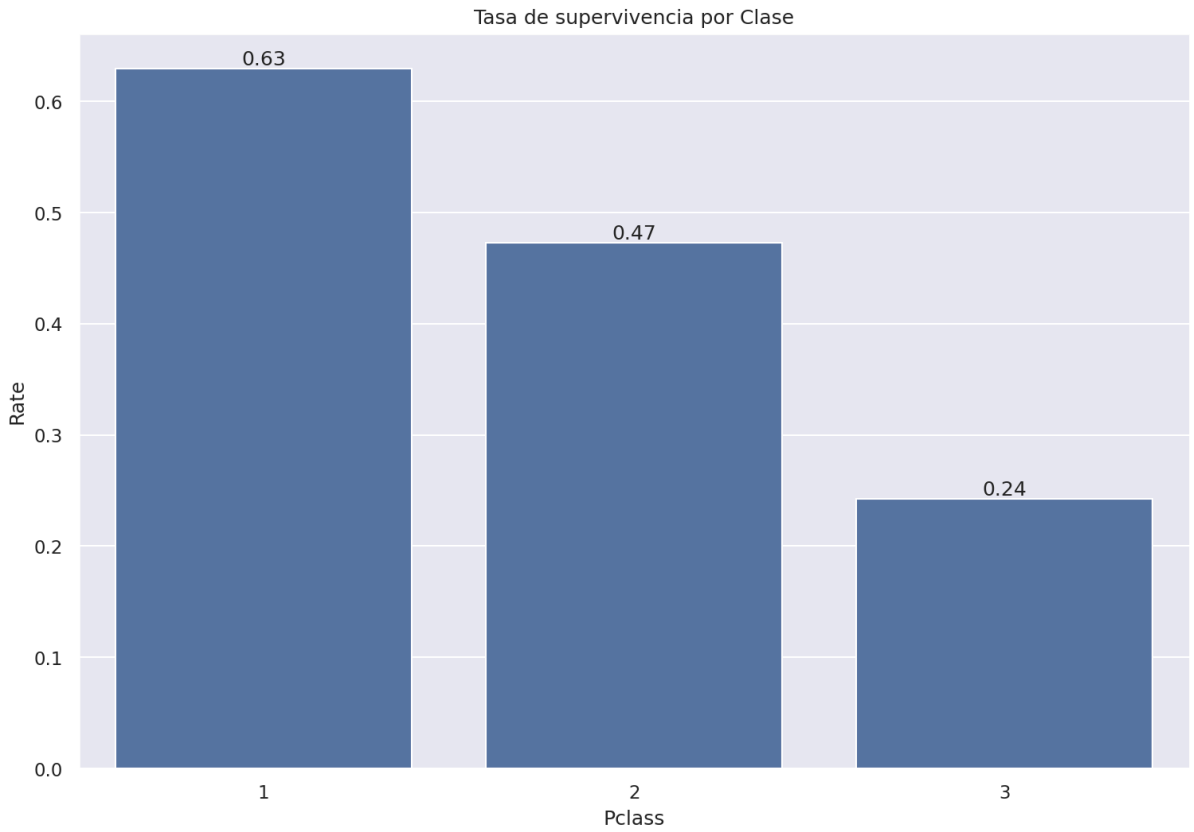
	Sex	**Embarked**	**n**	**survived**	**rate**
0	female	C	73	64	0.876712
1	female	Q	36	27	0.750000
2	female	S	203	140	0.689655
3	male	C	95	29	0.305263
4	male	S	441	77	0.174603
5	male	Q	41	3	0.073171

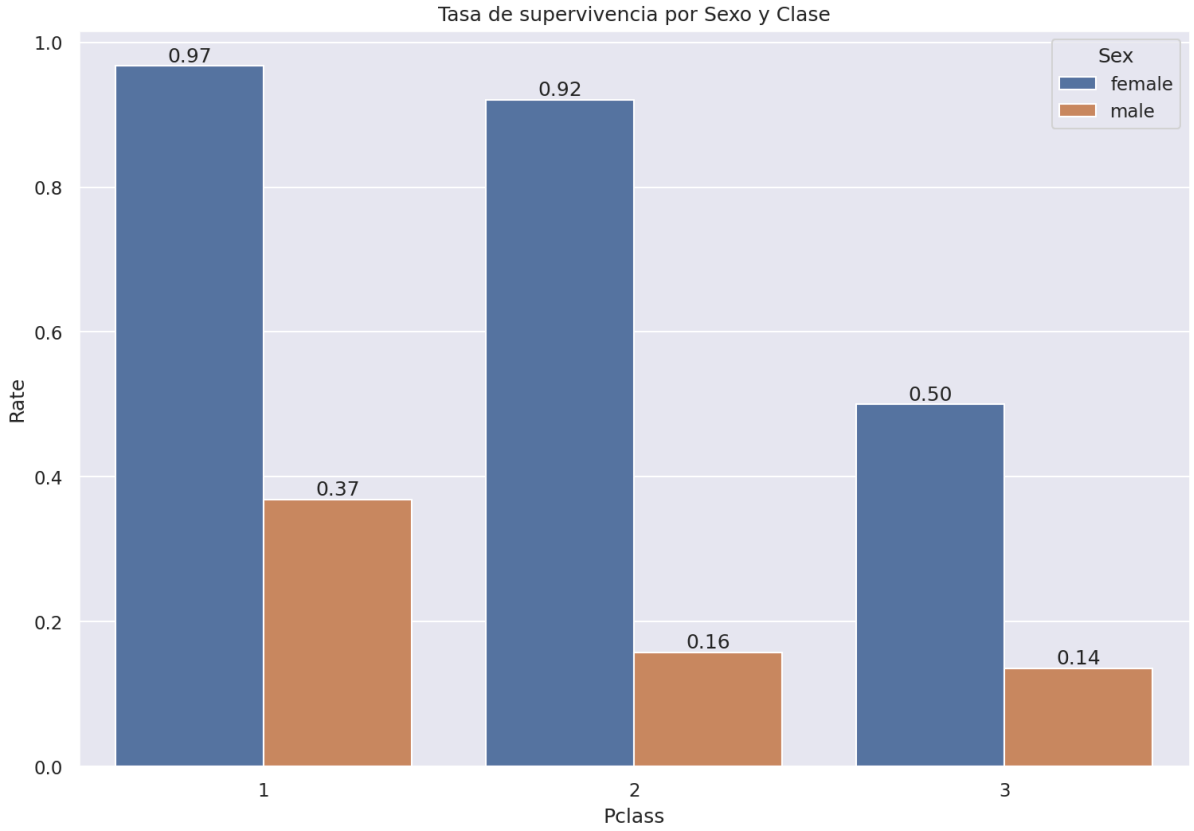
RESUMEN AUTOMÁTICO (mayores tasas observadas)

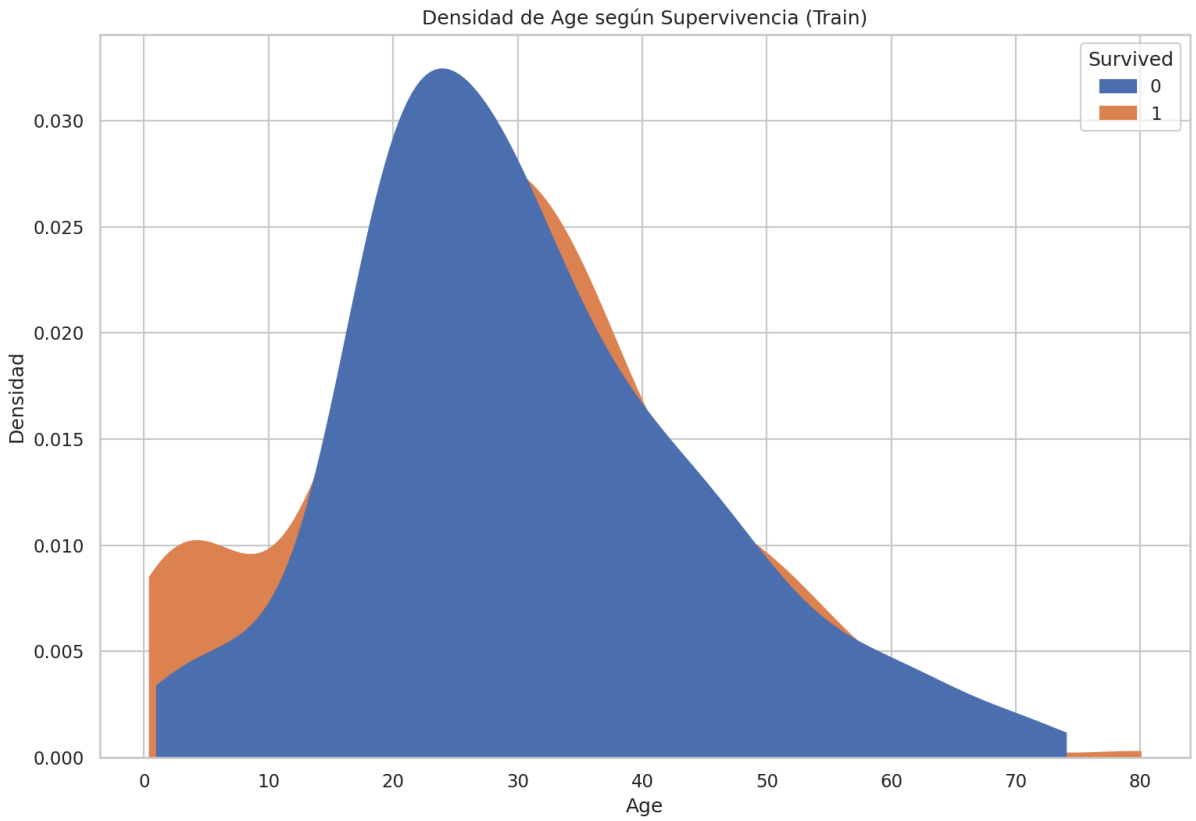
	grouping	**n**	**survived**	**rate**	**Sex**	**Pclass**	**Age Bin**	**Title**	**Family Bin**	**Fare Bin**	**Embarked**
0	sex_pclass	94.0	91.0	0.968085	female	1.0	NaN	NaN	NaN	NaN	NaN
1	sex_agebin	27.0	23.0	0.851852	female	NaN	(45.0, 60.0]	NaN	NaN	NaN	NaN
2	title	125.0	99.0	0.792000	NaN	NaN	NaN	Mrs	NaN	NaN	NaN
3	sex	314.0	233.0	0.742038	female	NaN	NaN	NaN	NaN	NaN	NaN
4	pclass	216.0	136.0	0.629630	NaN	1.0	NaN	NaN	NaN	NaN	NaN

5	familybin	131.0	80.0	0.610687	NaN	NaN	NaN	NaN	(2.0, 4.0]	NaN	NaN
6	farebin	222.0	129.0	0.581081	NaN	NaN	NaN	NaN	NaN	(31.0, 512.329]	NaN
7	agebin	69.0	40.0	0.579710	NaN	NaN	(-0.001, 12.0]	NaN	NaN	NaN	NaN
8	embarked	168.0	93.0	0.553571	NaN	NaN	NaN	NaN	NaN	NaN	C




Tasa de supervivencia por Clase
0.63
0.47
0.24
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Rate
1
2
3
Pclass


Tasa de supervivencia por Sexo y Clase
Sex
female
male
0.97
0.37
0.92
0.16
0.50
0.14
0.0
0.2
0.4
0.6
0.8
1.0
Rate
1
2
3
Pclass



Al responder la pregunta "¿Quiénes tenían más probabilidades de sobrevivir en el Titanic?", los resultados del análisis en Python muestran que las mujeres, los niños y los pasajeros de primera clase fueron los grupos con mayor probabilidad de supervivencia. En promedio, más del 70 % de las mujeres sobrevivió, mientras que en los hombres la tasa no superó el 20 %. Esta diferencia tan marcada se explica por la aplicación del protocolo de evacuación "women and children first", que otorgó prioridad a mujeres y menores durante el abordaje de los botes salvavidas.

Al combinar las variables, los resultados fueron aún más claros: las mujeres de primera clase presentaron las tasas más altas, con porcentajes superiores al 95 %. También se observó que los niños pequeños con el título de "Master" (en la columna Title) tuvieron una probabilidad de supervivencia muy alta, superando el 80 %. En contraste, los pasajeros de tercera clase o con tarifas más bajas registraron tasas mucho menores. En resumen, el modelo confirma que los factores género, edad y clase social fueron determinantes: ser mujer, pertenecer a una clase alta y/o ser menor de edad aumentaba considerablemente la probabilidad de sobrevivir al hundimiento del Titanic.

Además, el análisis de los datos reveló que la ubicación de los camarotes y la estructura familiar también influyeron en las posibilidades de supervivencia. Los pasajeros de primera clase se encontraban en cubiertas superiores, más cercanas a las zonas donde se desplegaron los botes salvavidas, lo que facilitó una evacuación temprana. En contraste, los ocupantes de segunda y tercera clase estaban ubicados en los niveles inferiores del barco, y muchos de ellos fueron retardados por la confusión y las barreras físicas de los pasillos. Otro factor relevante fue el tamaño de la familia: quienes viajaban solos o con grupos pequeños (1 o 2

acompañantes) tuvieron mejores resultados, posiblemente porque moverse y encontrar un lugar en los botes resultaba más sencillo que para familias grandes que intentaban permanecer unidas.

Finalmente, al observar las variables combinadas —como sexo, clase y edad— se evidencia una jerarquía clara de supervivencia. Las mujeres jóvenes de primera clase y los niños menores de 12 años, independientemente de su clase, fueron los grupos con mayor probabilidad de salvarse, mientras que los hombres adultos de clases bajas fueron los más vulnerables. Este patrón refuerza la idea de que la tragedia del Titanic no solo fue un accidente marítimo, sino también un reflejo de las desigualdades sociales y de género de su época: los recursos, la posición social y las normas culturales determinaron, en gran medida, quién vivía y quién no en aquella noche del 15 de abril de 1912.