Concentración en Inteligencia Artificial

Titanic

Fernanda Vásquez A00837164
Luis Gerardo Juárez García A00836928
Julio César Madrigal John A01737106
Joseph Elí Pulido A01831526
Sergio David Laverde A01831525

Titanic

Tu vida decidida por

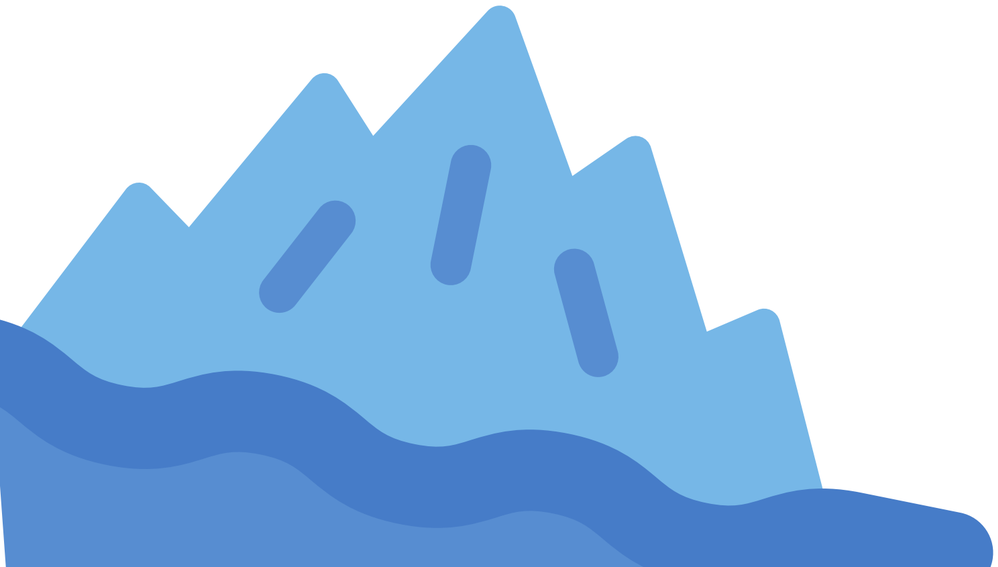
Genero

Desigualdad social

Familia

Problema

El reto que abordamos fue **analizar** este dataset histórico con técnicas de **Machine Learning.**



Motivación

Entender qué **variables** fueron más determinantes para **sobrevivir**

Reflexionar sobre como **sesgos** del pasado **influyen** a sistemas de **inteligencia artificial** actualmente

Preguntas de investigación

¿Qué factores fueron más determinantes en la supervivencia de los pasajeros?

¿Cómo se manifiestan los sesgos sociales género, clase y edad en los datos?

¿Qué compromisos existen entre accuracy y fairness en los modelos de ML?

¿Qué lecciones nos deja este caso histórico para el diseño de algoritmos más responsables hoy?

Hipótesis

1

Los pasajeros con familiares a bordo del barco tienen mayor posibilidad de sobrevivir.

2

Los hombres de primera clase tuvieron una tasa de supervivencia más alta en comparación a las mujeres que viajaban en tercera clase.

3

El puerto de embarque influyó en la probabilidad de supervivencia en el Titanic.

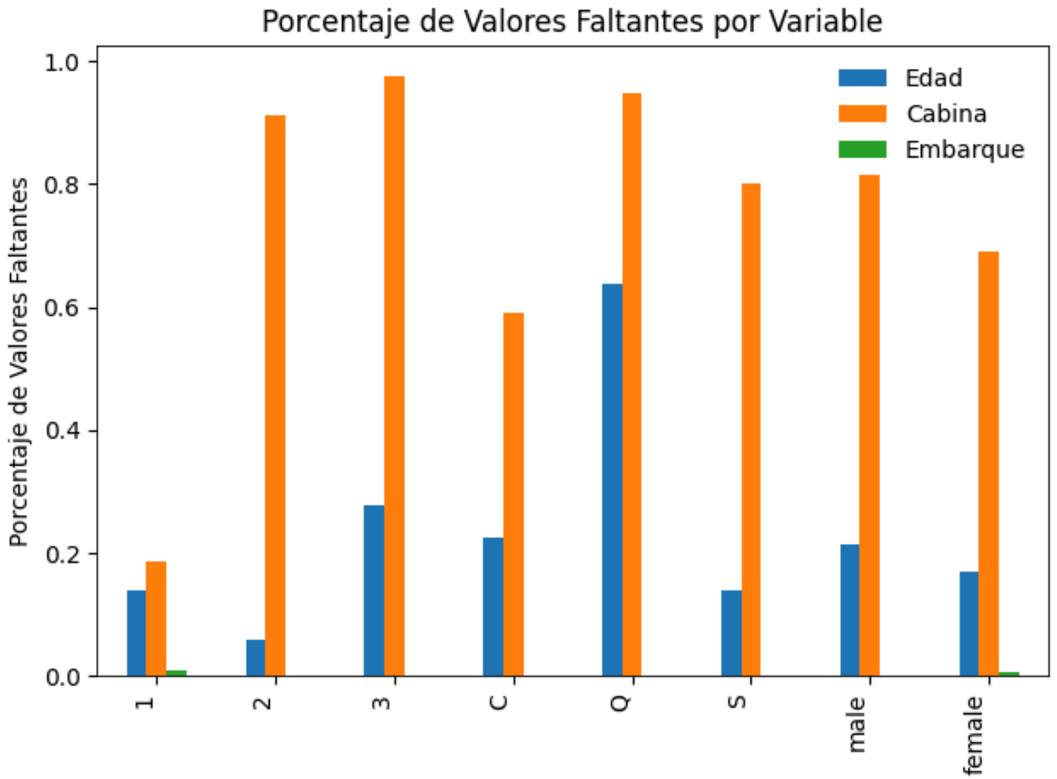
Variable	Definition	Key
PassengerId	Identificador del pasajero	Número único asignado a cada pasajero
Survived	Supervivencia	0 = No, 1 = Sí
Pclass	Clase del ticket	1 = Primera, 2 = Segunda, 3 = Tercera
Name	Nombre del pasajero	Incluye título (Sr., Sra., etc.)
Sex	Sexo	male = Hombre, female = Mujer
Age	Edad en años	Dato númerico que mide la edad del pasajero
SibSp	Hermanos / cónyuge a bordo	Número de hermanos y/o cónyuge que viajaban
Parch	Padres / hijos a bordo	Número de padres y/o hijos que viajaban
Ticket	Número de ticket	Puede repetirse entre pasajeros
Fare	Tarifa del pasajero	Precio pagado por el boleto
Cabin	Número de cabina	Contiene el identificador de la cabina asignada
Embarked	Puerto de embarque	C = Cherbourg, Q = Queenstown, S = Southampton

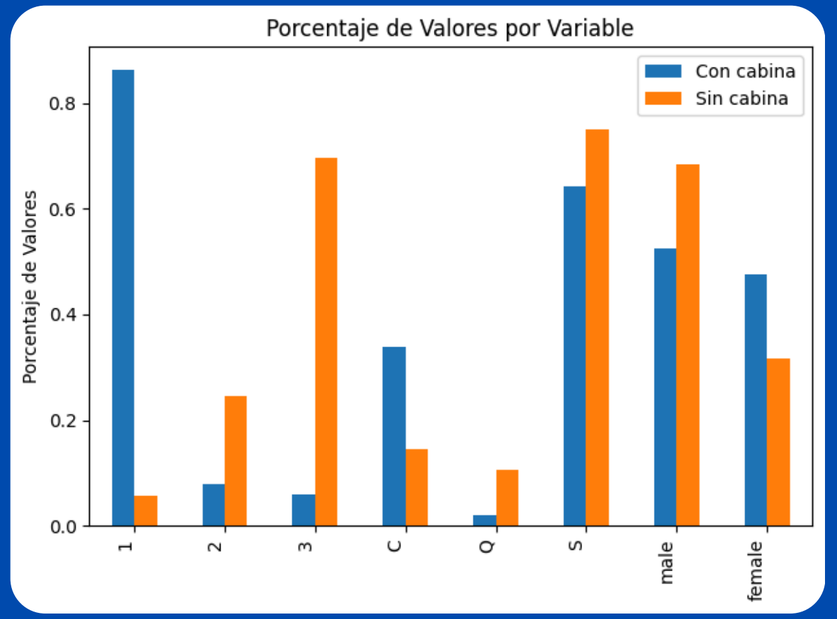
Data Set

Análisis Preliminar

- Valores faltantes:

	missing_count	missing_%
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

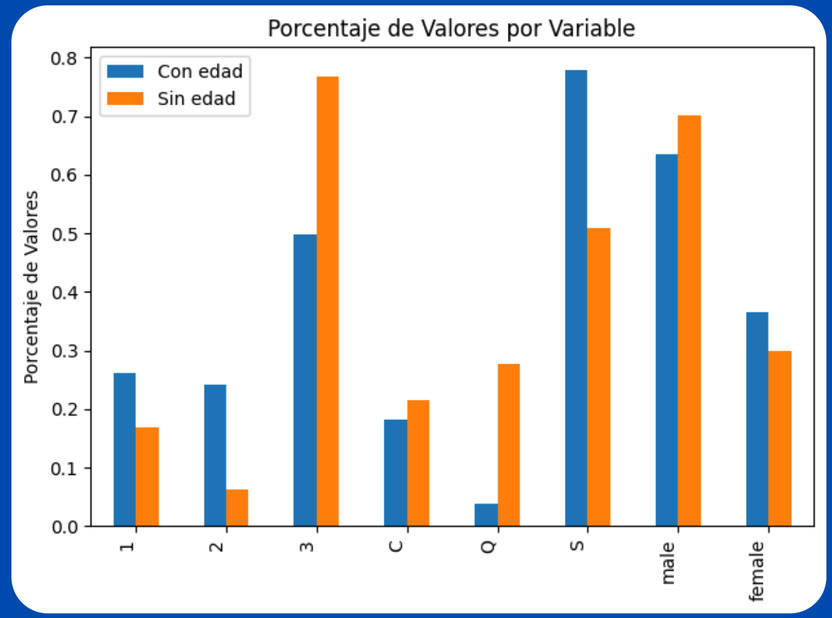




Promedio de clase:

- **Con cabina: 1.2**
- **Sin cabina: 2.6**

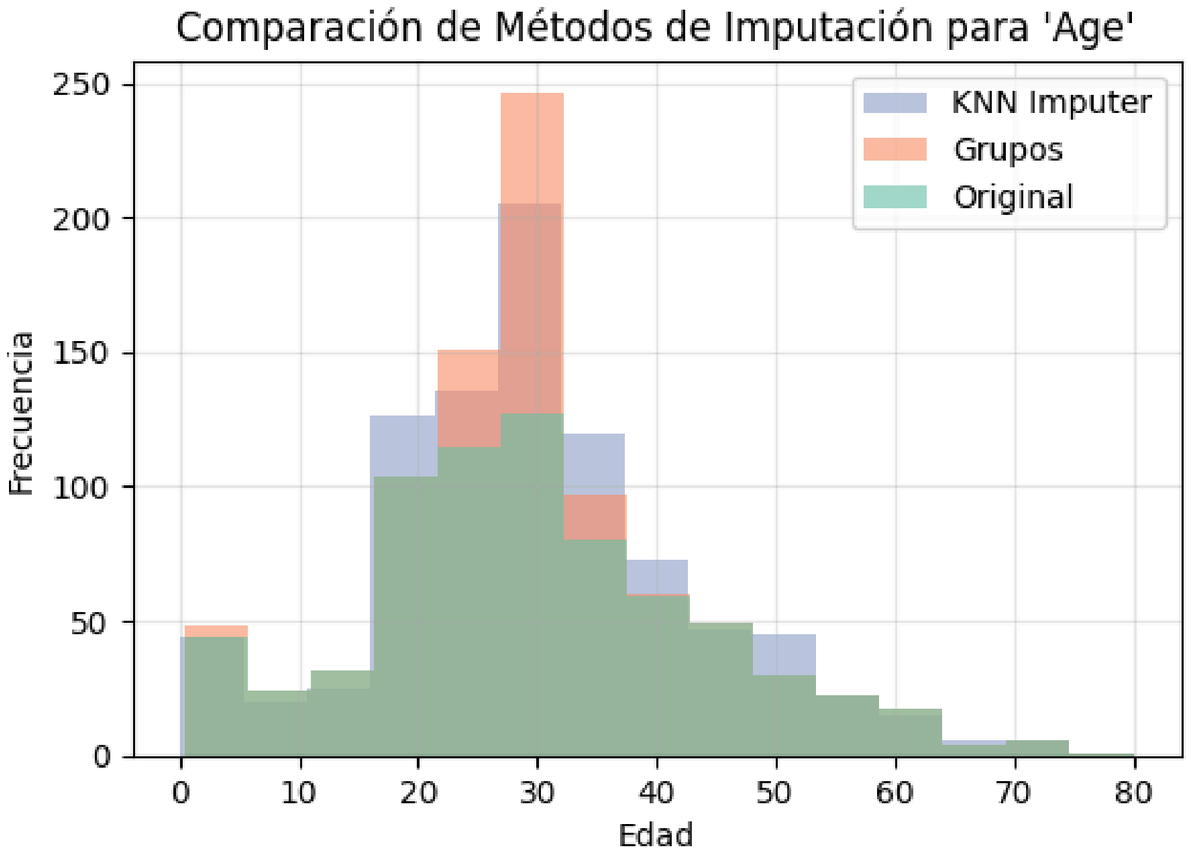
Difieren en cuanto a las variables clase y puerto de embarque



Imputación de variables faltantes

1. **Cabina**: Dado que a la variable Cabina le hacen falta el 77% de los datos, no se considerará la columna completa.
2. **Embarked**: Dado que para la variable embarque sólo faltan 2 valores, los asociaremos a un error aleatorio (MCAR). De 'Encyclopedia Titanica' obtenemos que los valores faltantes corresponden al puerto de Southampton.
3. **Edad**: Se analizan 2 métodos de imputación:

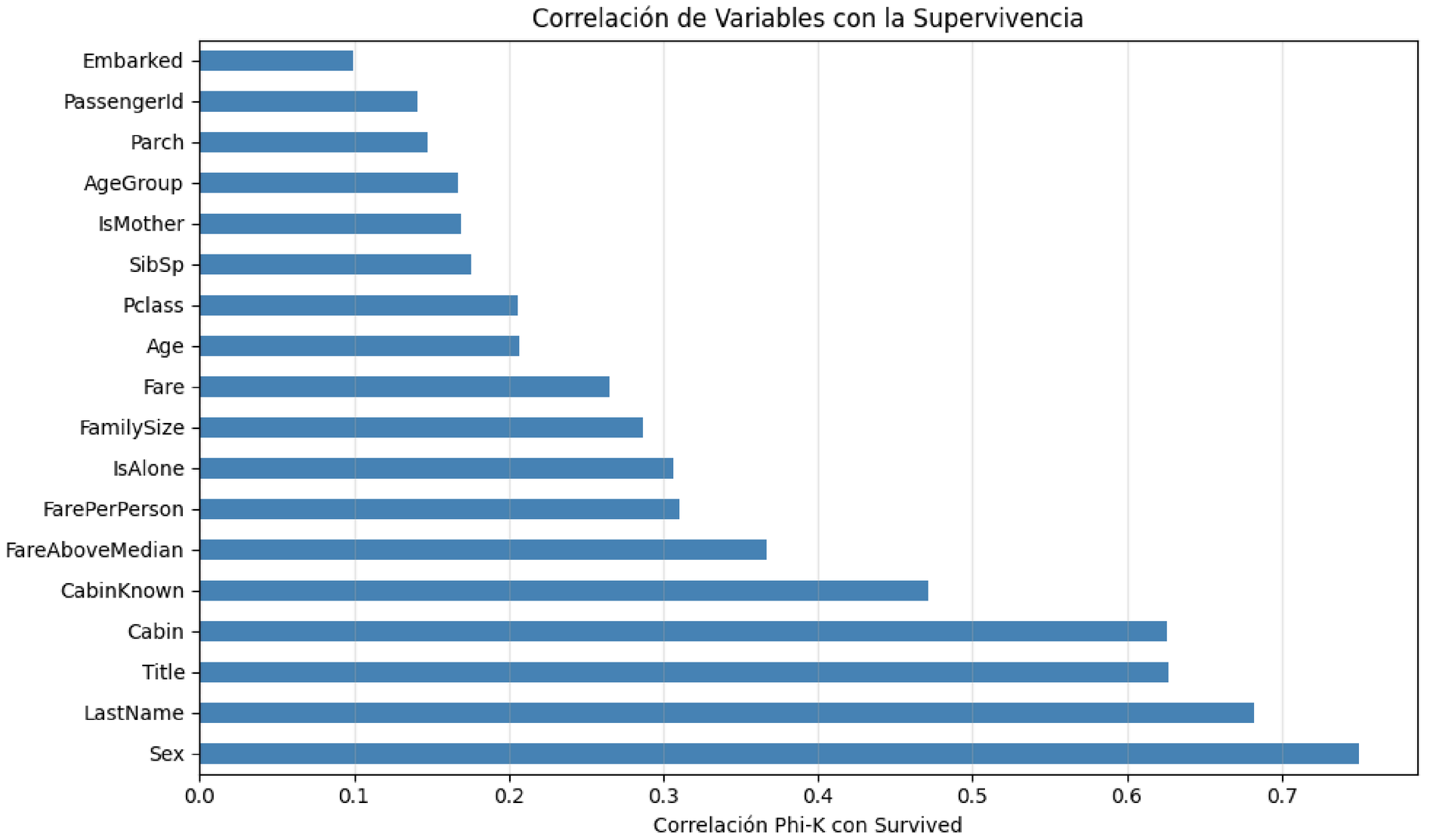
- Método Basado en Grupos: edad promedio por Título.
- Método KNN imputer.



KNN imputer es más próximo a la distribución original.

Features

Title: Es el título honorifico de cada pasajero.
FamilySize: Tamaño de la familia.
IsAlone: 0 si viajaba solo, 1 si no.
AgeGroup: Se categorizó según la edad, "Infant"(0-12) años, "Teenager" (13-18) años, "Adult" (19-64)años y "Third age" (mayor de 65 años).
FarePerPerson: Se encontró el costo del ticket individual.
IsMother: 1 si es madre, 0 si no.
FareAboveMedian: 1 si el precio del ticket esta por encima de la media, 0 si no.
FamilySurvivalID: Agrupa a los pasajeros que compartían el billete o apellido.
LastName: El apellido del pasajero.
CabinKnown: 1 si se conoce la cabina, 0 si no.


Correlación de Variables con la Supervivencia
Embarked
PassengerId
Parch
AgeGroup
IsMother
SibSp
Pclass
Age
Fare
FamilySize
IsAlone
FarePerPerson
FareAboveMedian
CabinKnown
Cabin
Title
LastName
Sex
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Correlación Phi-K con Survived

Variables seleccionadas para los modelos

Se seleccionaron las siguientes variables debido a su correlación con la supervivencia o para comprobar/negar las hipótesis antes planteadas:

Sex_male: La variable más significativa. 1 si es hombre, 0 si es mujer.
IsAlone: Para poder comprobar/negar la hipótesis 1. 0 si viajaba solo, 1 si no.
FareAboveMedian: Para observar si el factor económico fue significativo. 1 si el precio del tiket esta por ensima de la media, 0 si no.
CabinKnown: Para ver si el missingness influye o no en la supervivencia. 1 si se conoce la cabina, 0 si no.
Pclass_2, Pclass_3: Para determinar si la clase fue significativa o no. One-Hot Encoding de ‘Pclass’.
Embarked_Q, Embarked_S: Para poder comprobar/negar la hipótesis 3. One-Hot Encoding de ‘Embarked’.

Modelos implementados

Regresión Logística

Modelo lineal interpretable. Permite odds ratios ajustadas, base transparente para fairness y benchmarking.

Random Forest

Ensamble de árboles: captura interacciones no lineales y es robusto a outliers y a datos faltantes binarizados.

XGBoost

Boosting graduado: excelente para datos tabulares, optimiza log-loss y maneja relaciones complejas.

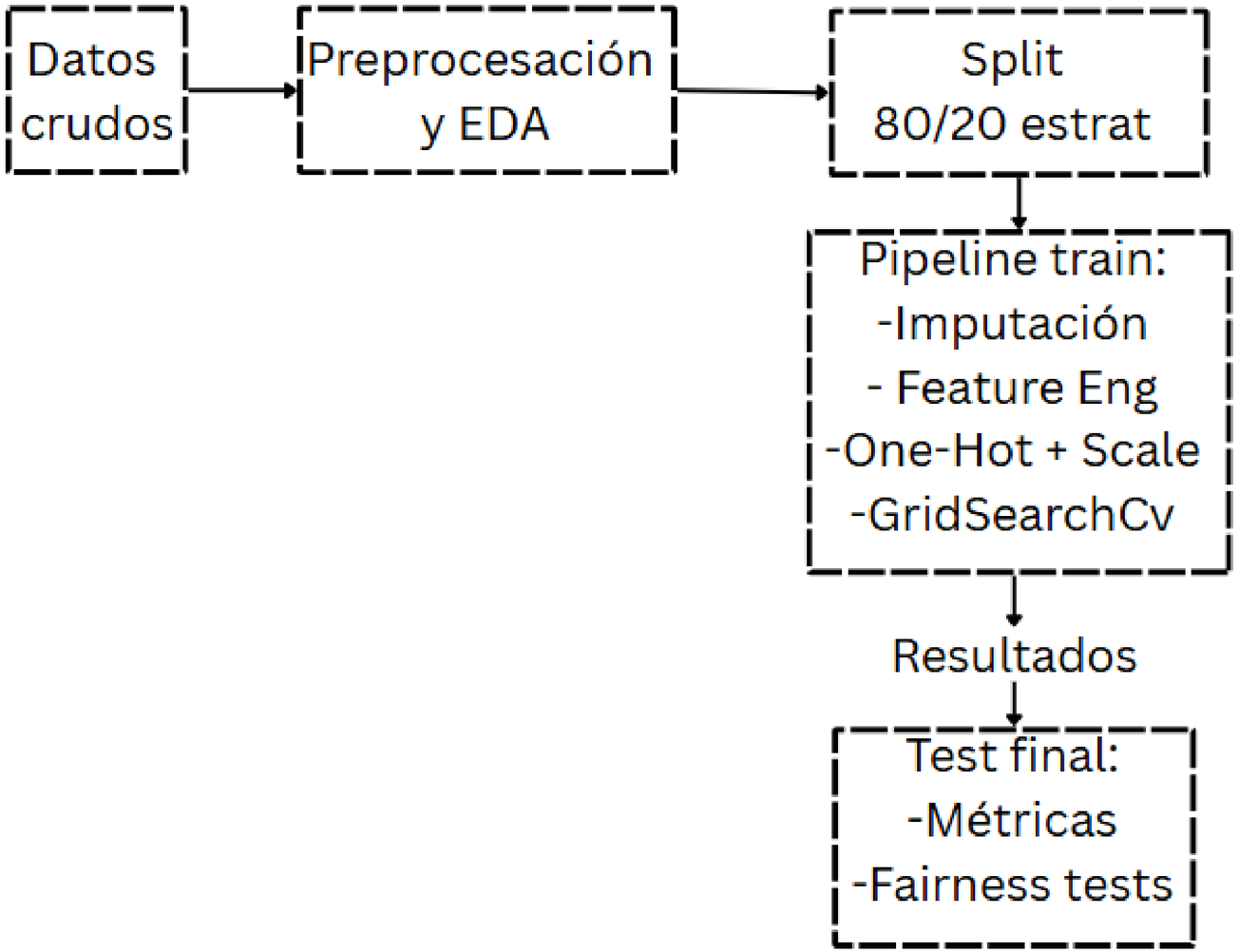
Support Vector Machine

Frontera de decisión en espacios de alta dimensión; útil para comparar no-linealidad con sombreado suave de kernels.

Neural Netwok

Captura patrones no lineales y dependencias de alto orden; prueba de capacidad de representación más allá de árboles y planos lineales.

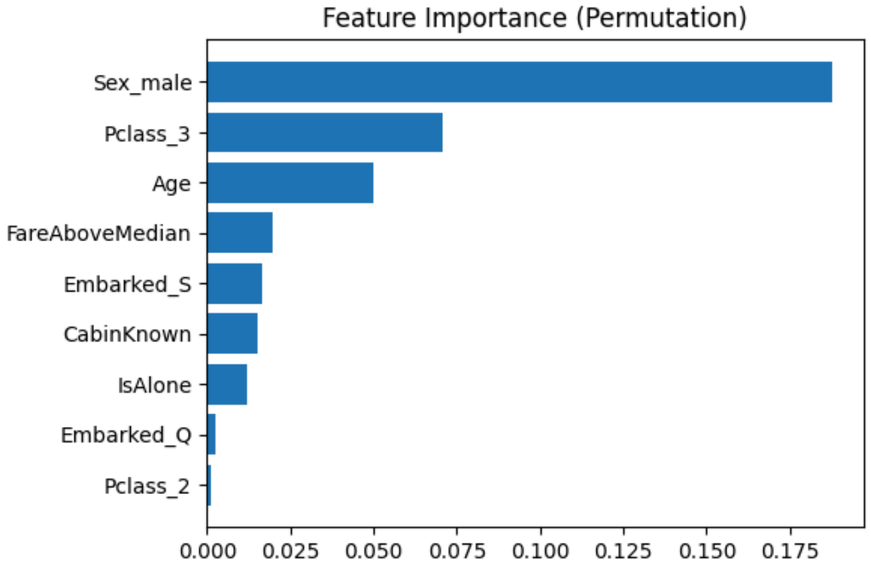
Pipeline

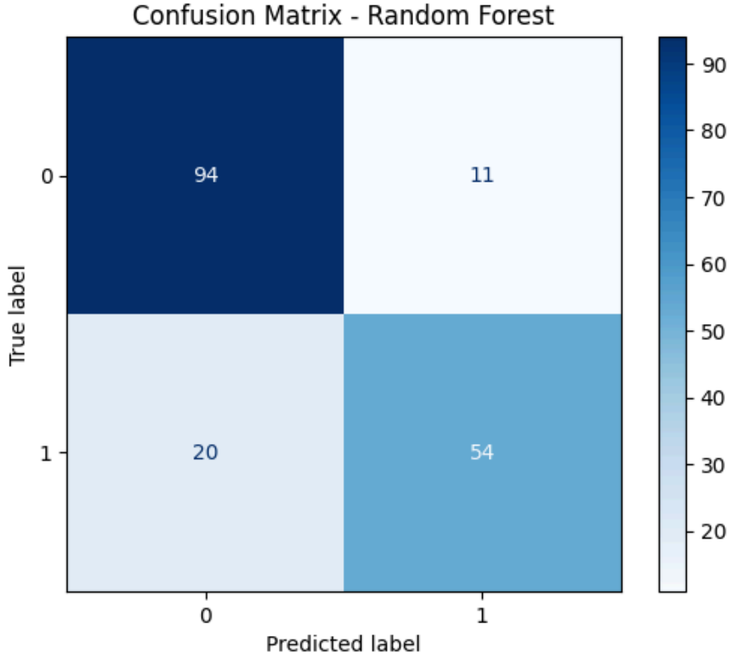

Datos crudos
Preprocesación y EDA
Split 80/20 estrat
Pipeline train:
-Imputación
- Feature Eng
-One-Hot + Scale
-GridSearchCv
Resultados
Test final:
-Métricas
-Fairness tests

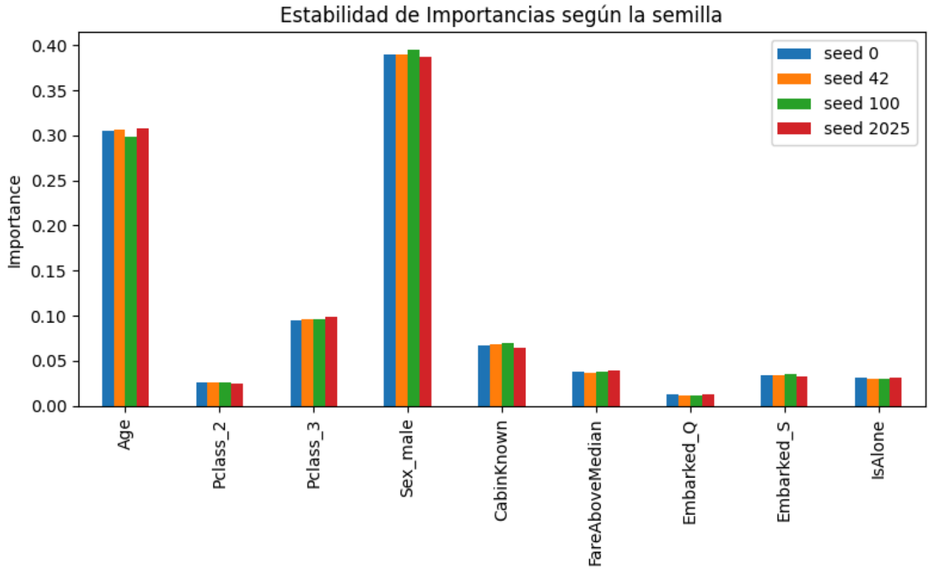
Comparación de Modelos - Métricas Redondeadas

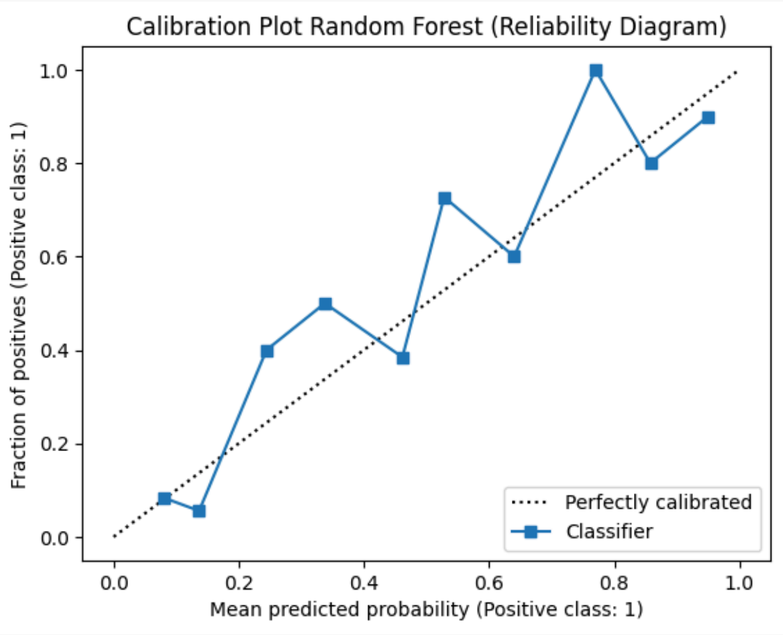
Modelo	Accuracy	Precision	Recall	F1	ROC-AUC	Tiempo Training	Tiempo Inference
XGBoost	0.81	0.823	0.689	0.75	0.88	0.079	0.005
Regresión Logística	0.804	0.741	0.811	0.774	0.863	0.007	0.003
Random Forest	0.827	0.831	0.73	0.777	0.889	0.221	0.03
Neural Network	0.827	0.841	0.716	0.774	0.873	0.151	0.0
Support Vector Machine	0.807	0.831	0.625	0.712	-	-	-

Gráficas


Feature Importance (Permutation)
Sex_male
Pclass_3
Age
FareAboveMedian
Embarked_S
CabinKnown
IsAlone
Embarked_Q
Pclass_2
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175


Confusion Matrix - Random Forest
True label
Predicted label
94
11
20
54
0
1
90
80
70
60
50
40
30
20


Estabilidad de Importancias según la semilla
Importance
seed 0
seed 42
seed 100
seed 2025
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Age
Pclass_2
Pclass_3
Sex_male
CabinKnown
FareAboveMedian
Embarked_Q
Embarked_S
IsAlone


Calibration Plot Random Forest (Reliability Diagram)
Fraction of positives (Positive class: 1)
Mean predicted probability (Positive class: 1)
Perfectly calibrated
Classifier
1.0
0.8
0.6
0.4
0.2
0.0

Hallazgos de Fairness

- **Demographic Parity:** compara si distintos grupos (ej. hombres/mujeres, clases sociales) reciben el mismo trato.
- **Equal Opportunity:** mide si las probabilidades de predecir bien son similares entre grupos.

Género
Clase social
Edad
Tarifa

Implicaciones Eticas

Decisión sobre vidas

Equidad vs. Precisión

Reproducción de sesgos sociales en modelos ML

Dashboard

Conclusiones