

Tec de Monterrey

Septiembre 2025



TITANIC DATASET

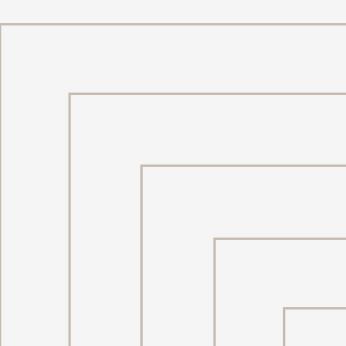
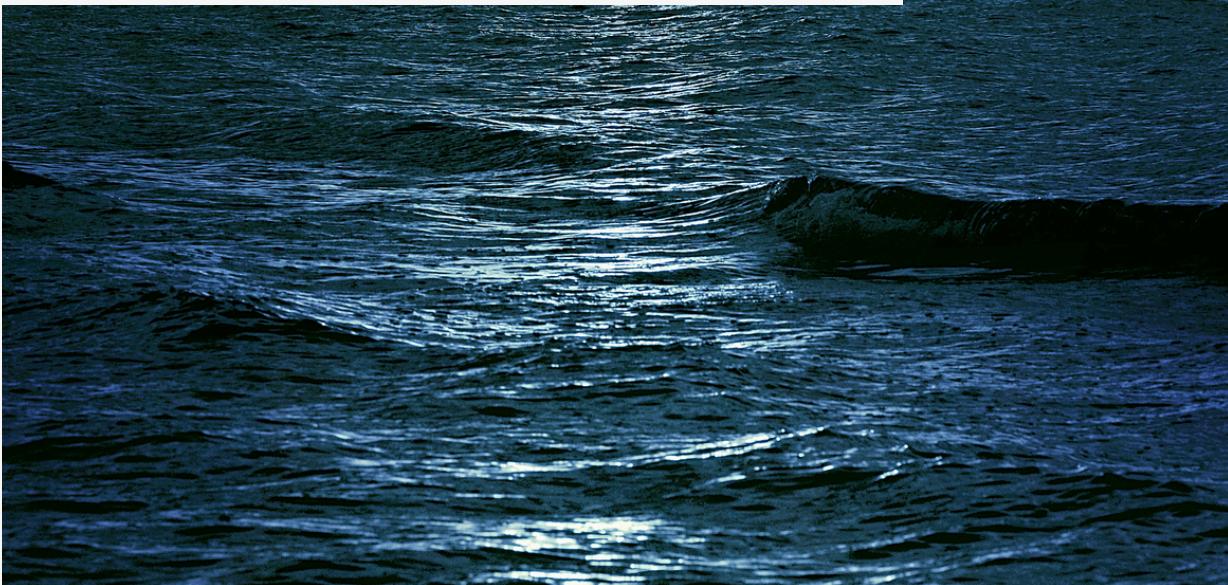
Equipo 3



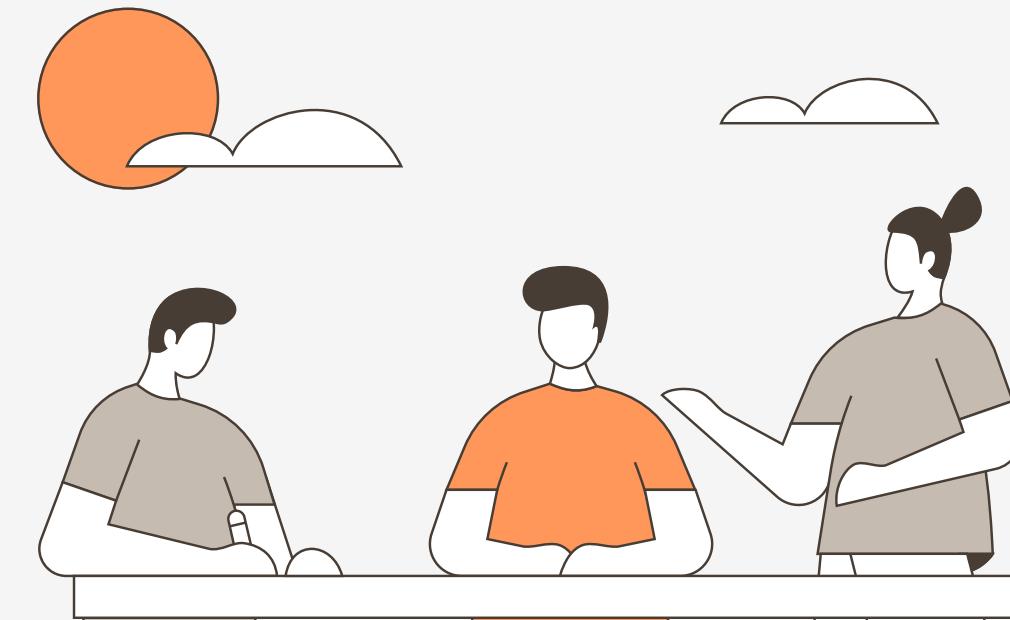
Pensamos que el Titanic se hundió en 1912. Pero en realidad, todavía estamos a bordo: navegamos en un mar de datos, donde algoritmos deciden quién avanza y quién se queda atrás.

Introducción

Análisis integral del dataset del Titanic para identificar factores determinantes en la supervivencia de pasajeros durante el naufragio del 15 de abril de 1912. Aplicación de metodologías avanzadas de ciencia de datos. ¿Como salvaríamos mas vidas?



Objetivos Generales



Objetivos Analíticos

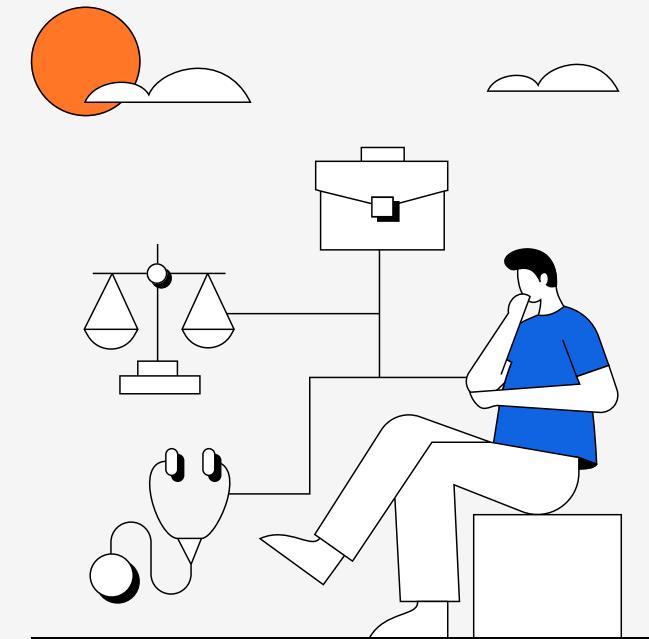
Identificar patrones de supervivencia y validar hipótesis históricas mediante análisis exploratorio y estadístico riguroso.

Objetivos Técnicos

Desarrollar modelos predictivos robustos y analizar sesgos sociales reflejados en los datos del desastre marítimo más famoso.



Preguntas de investigación



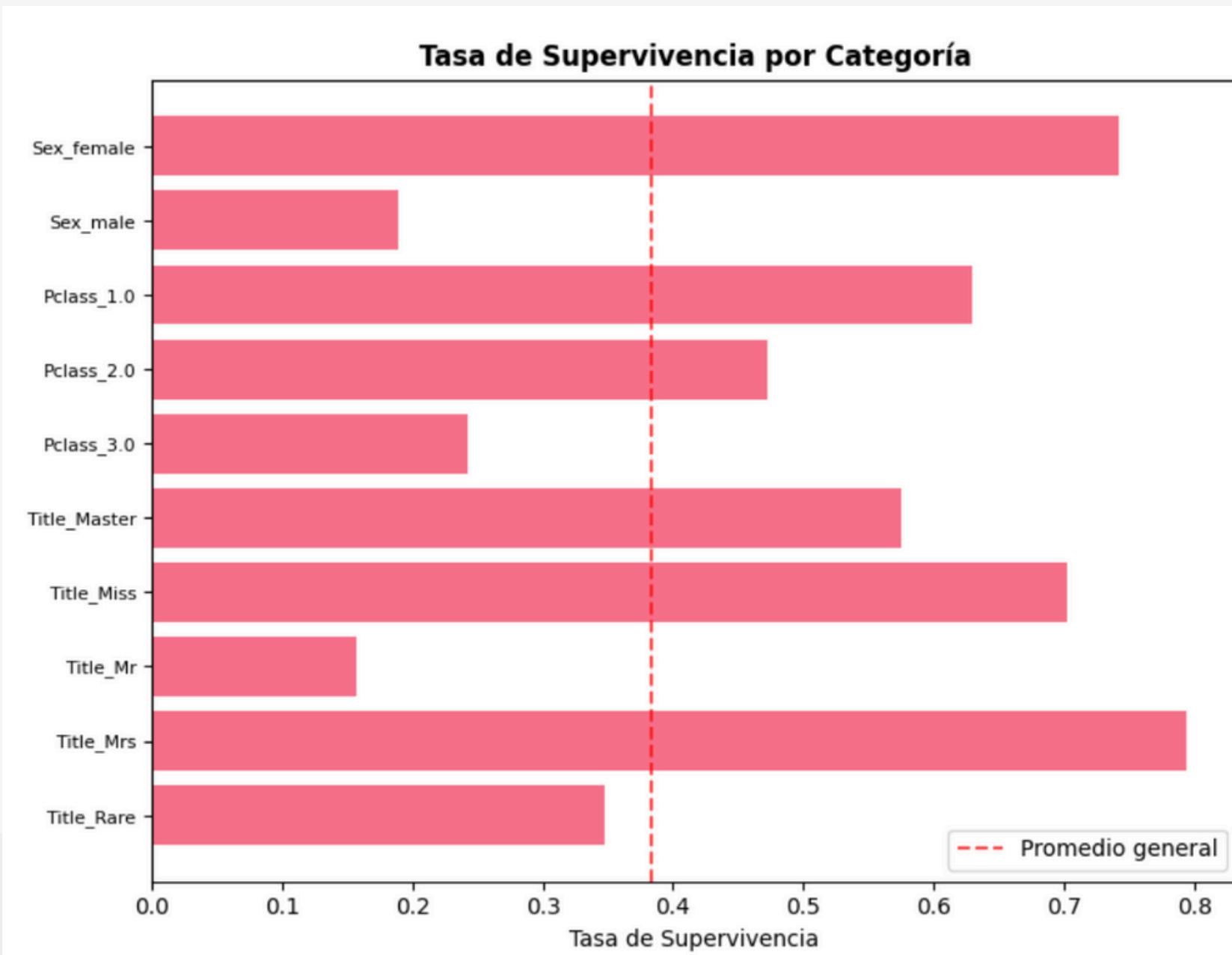
RQ1

¿Qué factores demográficos, socioeconómicos y situacionales fueron más determinantes para la supervivencia en el Titanic, y cómo se pueden cuantificar estas influencias utilizando técnicas modernas de interpretabilidad?

RQ2

¿Cómo se manifiestan los sesgos sociales históricos en los patrones de supervivencia del Titanic, y qué métricas de fairness son más apropiadas para evaluar estas disparidades? :::::

Hipótesis de investigación



H1: Protocolo de género aumentó supervivencia femenina.



H2: Viajar acompañado favoreció supervivencia.



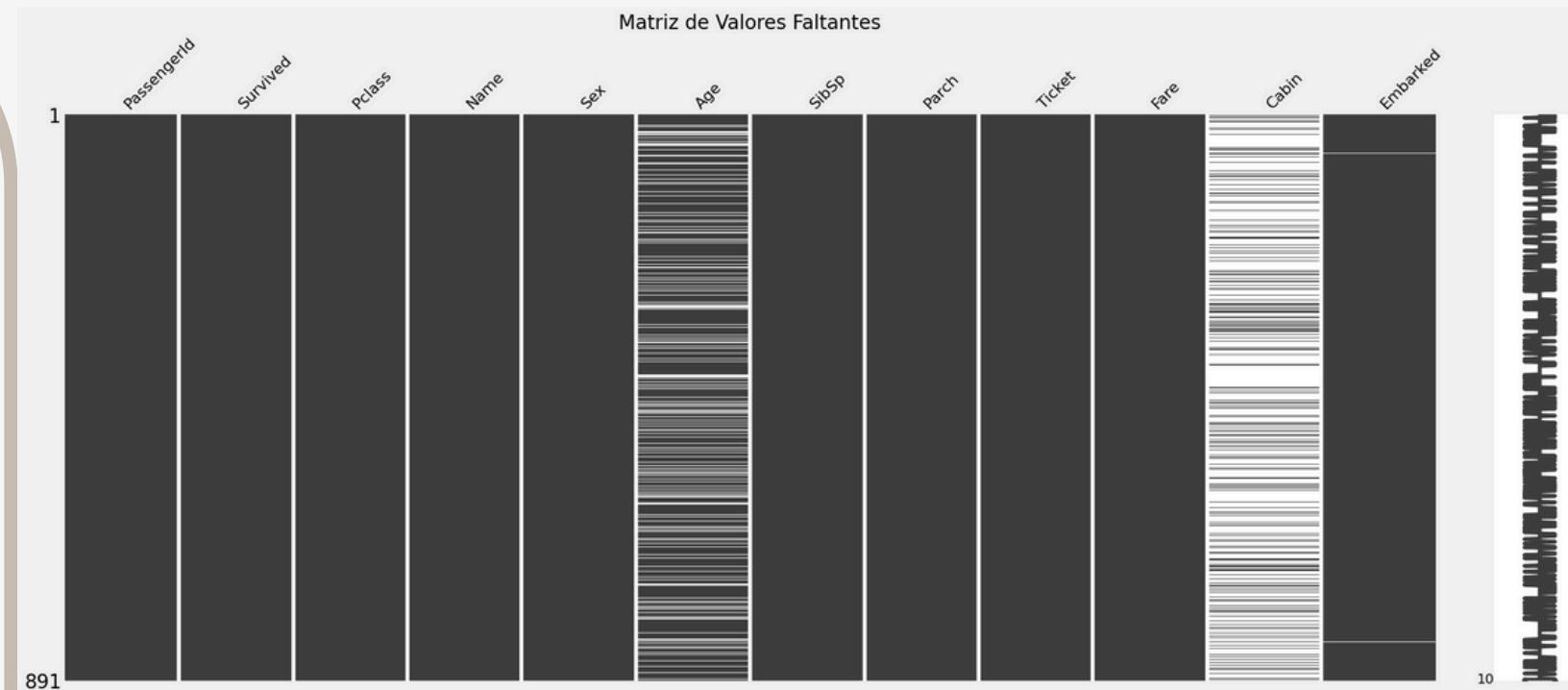
H3: Gestión inadecuada de recursos de salvamento redujo tasas de supervivencia general.



Dataset

Características Generales:

- **891 pasajeros** registrados en el RMS Titanic
- **12 variables** que incluyen datos demográficos, socioeconómicos y del viaje
- **Variable objetivo:** Supervivencia (Survived: 0=No, 1=Sí)



Variables Numéricas (7)

- PassengerId
- Survived: Variable objetivo (binaria)
- Pclass: (1^a, 2^a, 3^a clase)
- Age
- SibSp: Número de hermanos/cónyuges a bordo
- Parch: Número de padres/hijos a bordo
- Fare

Variables Categóricas (5)

- Name: Nombre completo del pasajero
- Sex: Género (male/female)
- Ticket: Número de ticket
- Cabin: Número de cabina
- Embarked: Puerto de embarque (C/Q/S)



Tratamiento de valores faltantes

Age (19.9% faltantes): imputación por grupos Pclass-Sex.

Cabin (77.1%): variable Has_Cabin + imputación por clase.

Embarked (0.2%): imputación por moda. Método por grupos seleccionado.



Diseño experimental

Formulación del Problema

- $f:X \rightarrow \{0,1\}$
- X representa el vector de características de un pasajero
- f es la función que predice la supervivencia (1 = sobrevivió, 0 = falleció)

Estrategia de Validación

- Train (60%) Validation (20%) Test (20%)
- Cross validation

Algoritmos Implementados y justificación

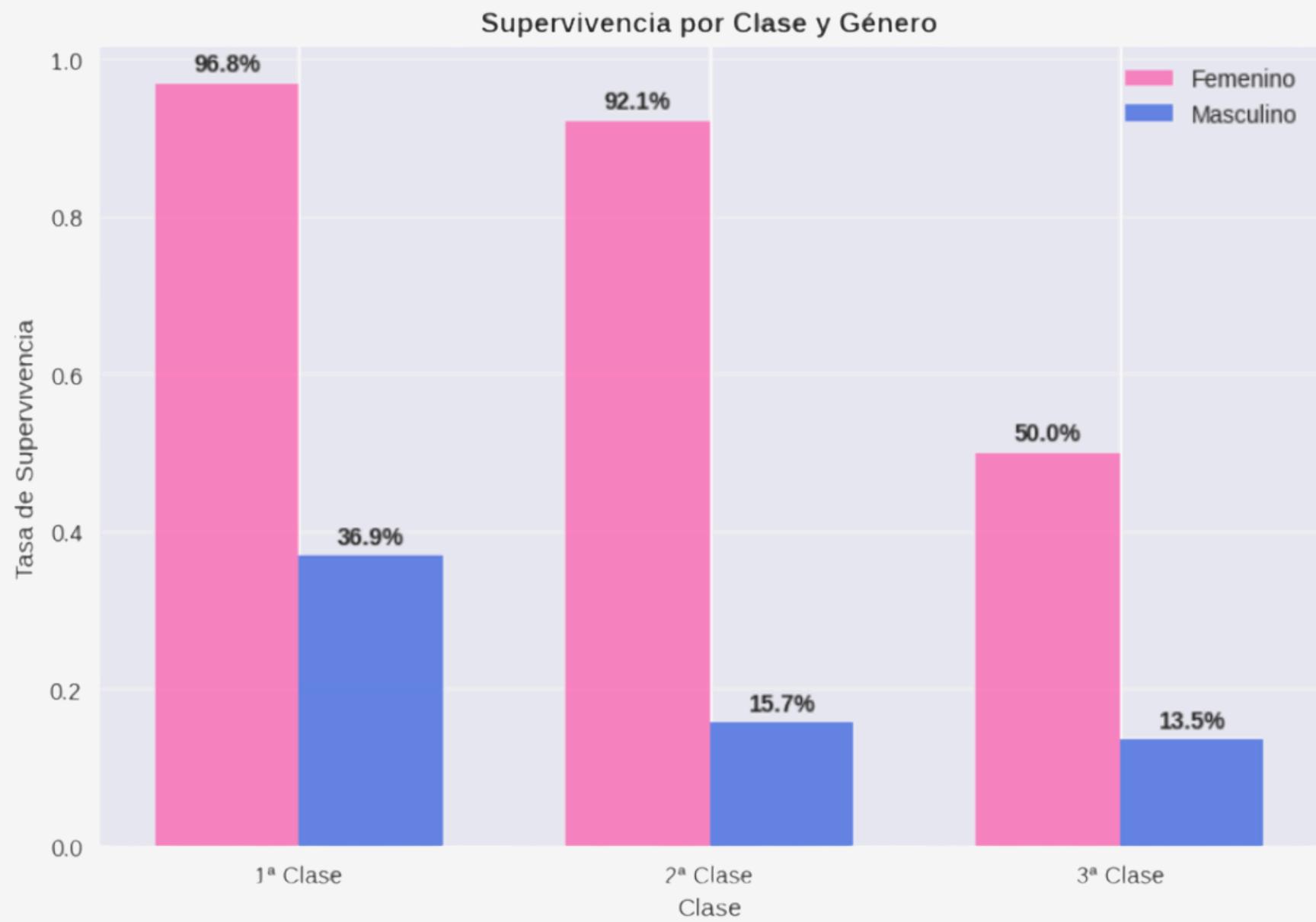
- Regresión Logística
 - Modelo de clasificación base
- XG Boost
 - Previene el overfitting con early stopping
- SVM
 - Bueno generalizando y capturando relaciones no lineales
- Random Forest
 - Puede modelar relaciones no lineales y feature importance
 - No es sensible a valores extremos.
- Red Neuronal
 - Modelo más complejo
 - Tiende a overfitting por falta de datos

	Accuracy
	Precision
	Recall
	F1-Score
	ROC-AUC

Hallazgos Principales



- **Supervivencia por género:** mujeres 74.2% vs hombres 18.9%.
- **Por clase:** primera 63%, segunda 47%, tercera 24%.
- **Familias medianas** (2-4 miembros) mostraron mejores tasas de supervivencia.



Elección del modelo

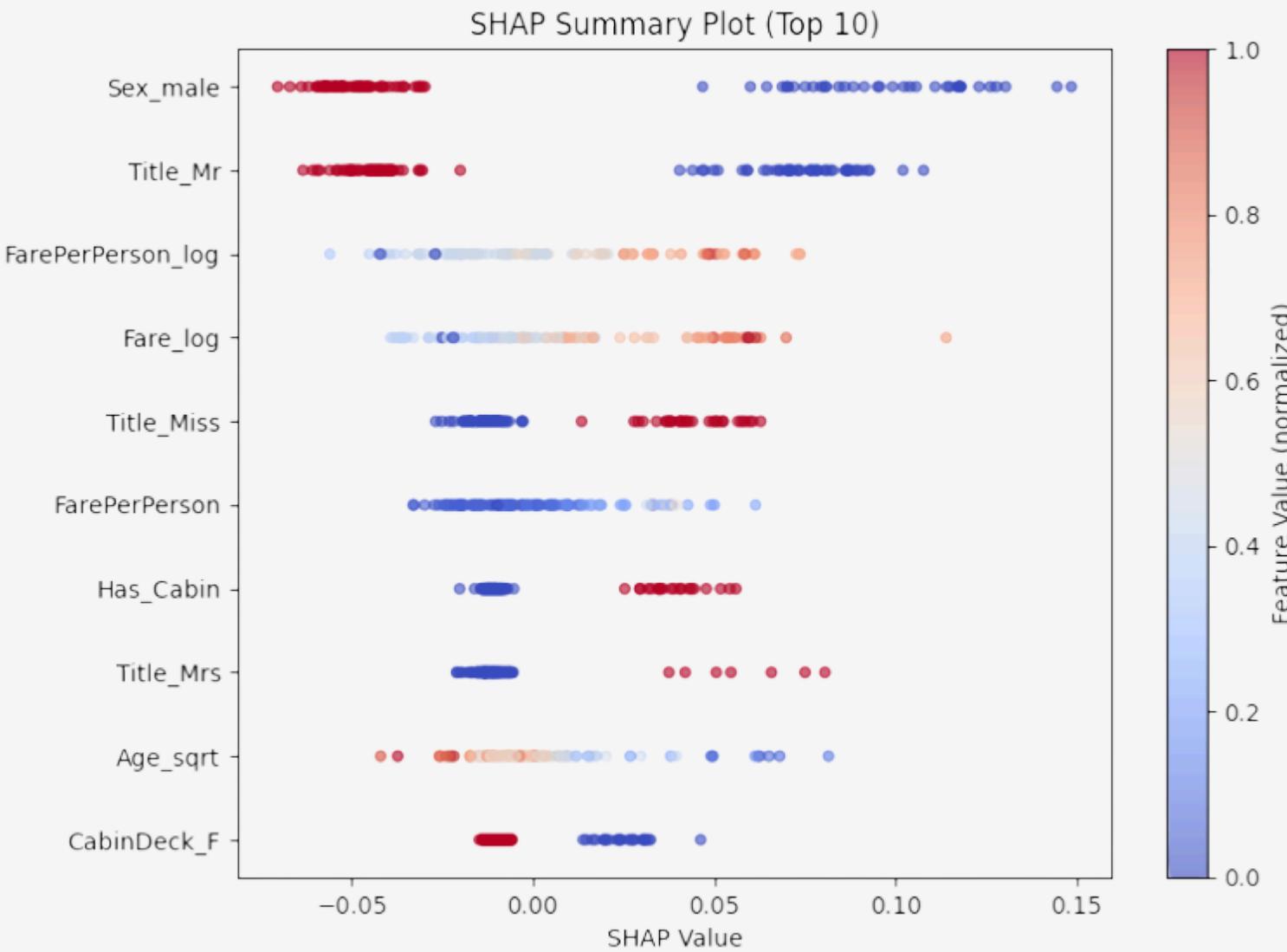
Tabla 1: Métricas Comparativas de Modelos (Conjunto de Prueba)

Métrica	XGBoost	Regresión Logística	SVM	Random Forest	Red Neuronal
Accuracy	0.826	0.843	0.821	0.843	0.793
Precision	0.825	0.778	0.776	0.803	0.711
Recall	0.691	0.824	0.681	0.779	0.783
F1 Score	0.752	0.800	0.746	0.791	0.745
ROC AUC	0.886	0.884	0.855	0.894	0.830

De los modelos comparados (XG Boost, Regresión Logística, SVC, Random Forest y Red Neuronal), el **Random Forest** obtuvo el mejor desempeño global con un **ROC-AUC** de 0.894, acompañado de una **accuracy** de 0.843 y un **F1-Score** de 0.791.

Interpretabilidad del modelo

Patrones descubiertos (SHAP & LIME)



- Sexo “male” y Título “Mr” variables más influyentes en el modelo.
- Clase alta = mayor supervivencia (FarePerPerson, Pclass).
- Caso con mayor supervivencia: combinación de edad joven, clase alta y tarifa alta.

Hallazgos de fairness y trade-offs

TABLE V
MÉTRICAS DE FAIRNESS POR GRUPO DEMOGRÁFICO

Grupo	TPR	FPR	Precisión	Dem. Par.
Género: Mujer	0.90	0.56	0.83	0.82
Género: Hombre	0.47	0.04	0.69	0.12
Clase: 1	0.89	0.14	0.92	0.63
Clase: 2	0.90	0.12	0.90	0.54
Clase: 3	0.52	0.11	0.55	0.20
Adulto (≥ 18)	0.75	0.12	0.79	0.35
Menor (< 18)	1.00	0.09	0.90	0.50

El modelo actual prioriza la **precisión** y la **selección** para mujeres y clases altas, sacrificando equidad para hombres y clase baja.

Mejorar la equidad implica un compromiso con la precisión.

Tendencia del TPR es arriba del 75% y los FPR abajo del 17%



Caso interesante en FPR de mujeres

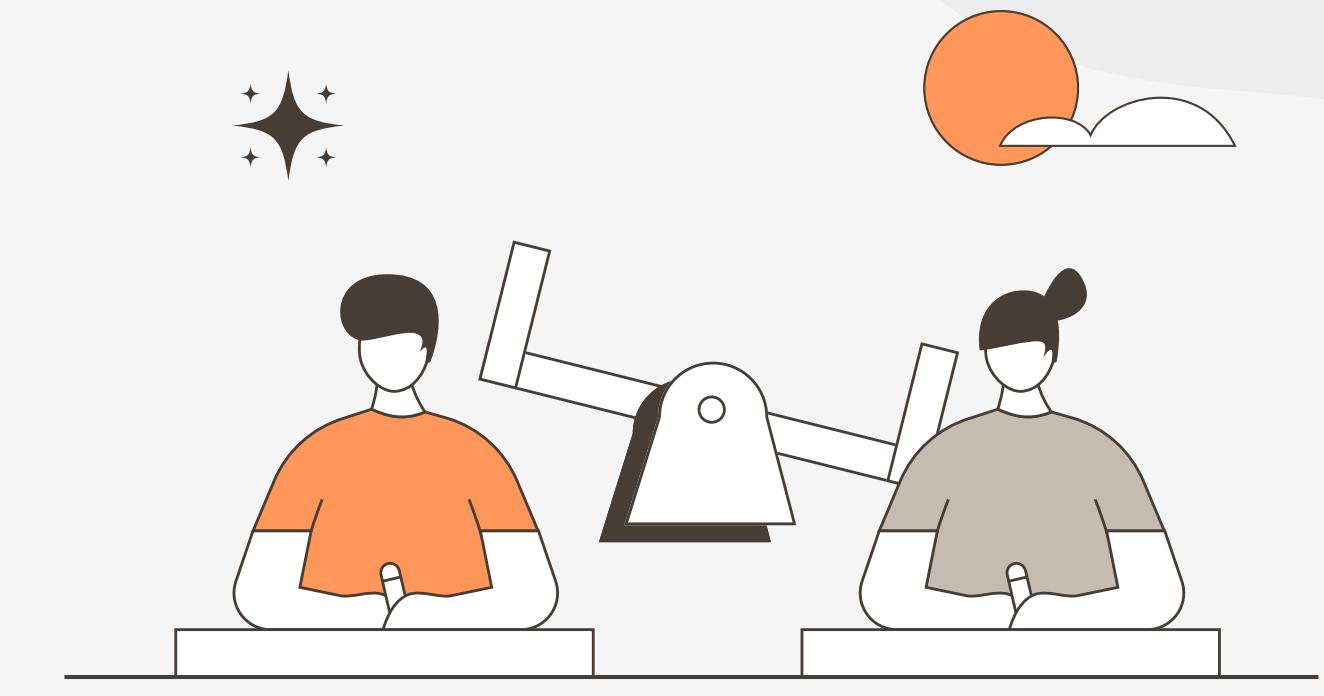


Validación de hipótesis

- H1: Fuertemente soportada - género determinante en supervivencia.
- H2: Confirmada - FamilySize aumenta probabilidad de supervivencia en 85% casos.
- H3: Validada - capacidad botes vs supervivientes reales evidencia mala gestión.

Reflexiones éticas principales, Dilemas & Aplicaciones modernas

- Los datos no son solo estadísticas: representan tragedias humanas.
- Dilema central: ¿es ético predecir quién “merece” sobrevivir?



Parallelismo actual: algoritmos en crédito, empleo o justicia también refuerzan desigualdades.

Sesgos reflejados:

- Código de honor (“mujeres y niños primero”).
- Clase social y género influyeron mucho en la supervivencia.

Métricas no son neutrales:

- precisión vs. recall → mejor priorizar recall para no excluir posibles sobrevivientes.

Limitaciones clave



- **Contexto Temporal Específico:** Las normas sociales de 1912 no son directamente aplicables al contexto actual.
- **Sesgo de Supervivencia:** Los registros disponibles pueden estar sesgados hacia pasajeros con documentación más completa. (ej. primera clase)
- **Información Faltante Sistématica:** El 77.1% de valores faltantes en la variable Cabin, refleja diferencias estructurales en el registro de información según la clase social.



Contribuciones principales

Ética y actualidad

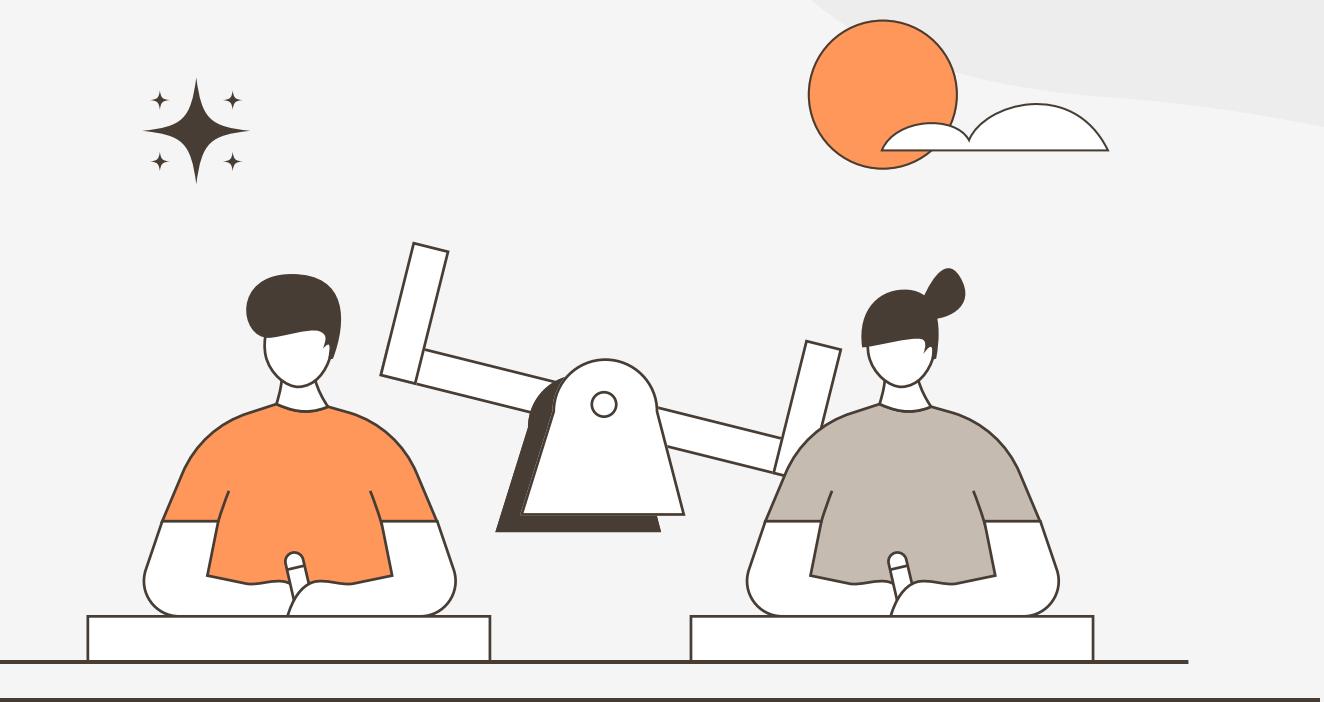
- Conectamos desigualdades históricas con problemas presentes.
- Priorizamos transparencia y análisis crítico.

Auditoría de fairness (Titanic)

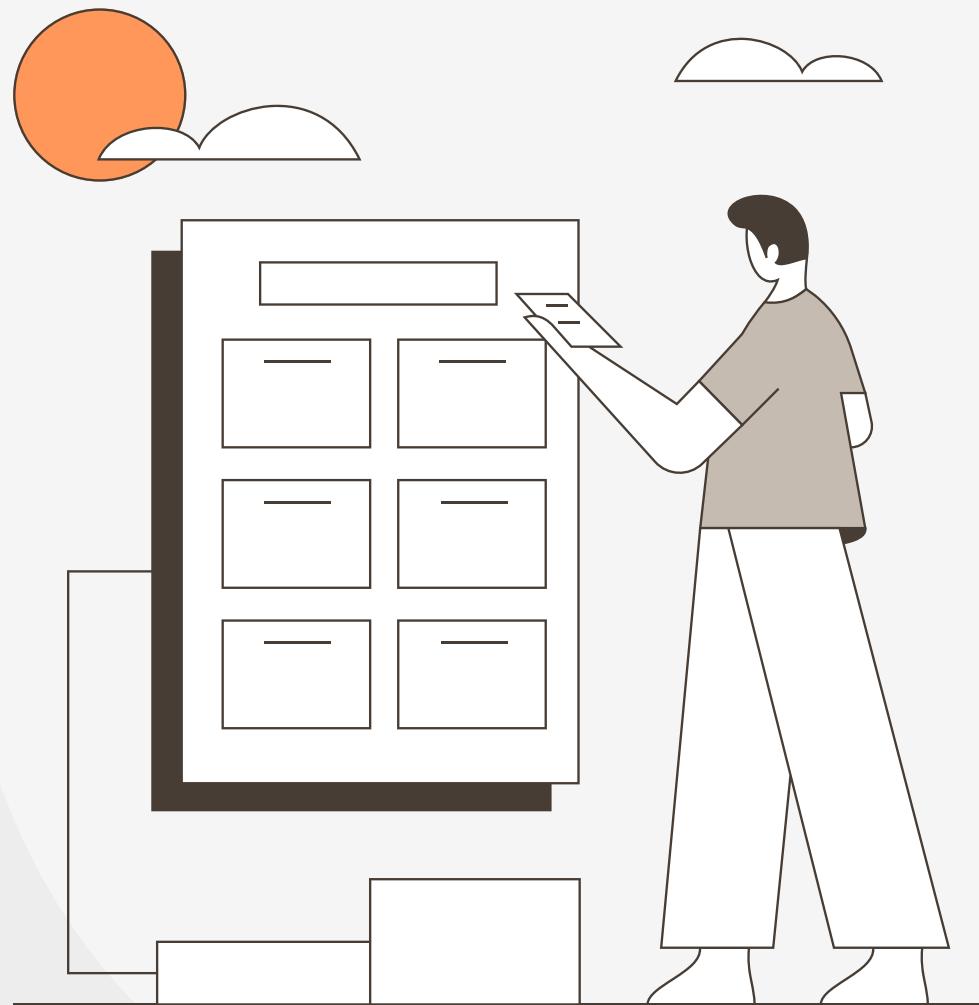
- Usamos métricas formales: paridad demográfica, igualdad de oportunidades y calibración.
- Mostramos disparidades entre grupos y cómo interactúan las variables.

Enfoque interseccional y explicable

- Combinamos interpretabilidad (SHAP, LIME) con análisis por subgrupos (género × clase × edad).
- Ofrecemos explicaciones locales y globales siguiendo buenas prácticas.



Dashboard



**Exploración de Datos
Avanzada**

01

Predictión Interactiva

02

Análisis de Modelo

03

What - If

04



EL TITANIC NOS RECUERDA QUE DETRÁS DE CADA
DATO HAY UNA VIDA HUMANA. NUESTRA
RESPONSABILIDAD COMO DESARROLLADORES DE AI
ES GARANTIZAR QUE LOS ALGORITMOS DEL FUTURO
HONREN ESTA LECCIÓN.

GRACIAS TOTALES

