

ANÁLISIS PREDICTIVO DE SUPERVIVENCIA EN EL TITANIC: UN VIAJE A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO

Aplicación de Algoritmos de Machine Learning para la Clasificación de Pasajeros

Autores: Juan J. Bonilla
Ricardo Muñoz
Valentina Isaza
Nelcy L. Zapata



INTRODUCCIÓN DEL PROBLEMA

Un caso representativo ampliamente utilizado con fines educativos es el desastre del RMS Titanic, ocurrido en 1912, donde más de 1,500 personas perdieron la vida tras el hundimiento del transatlántico.

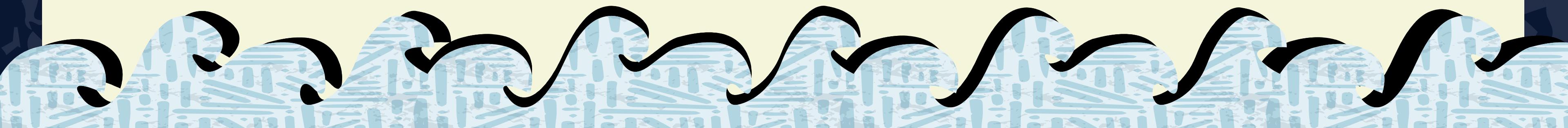
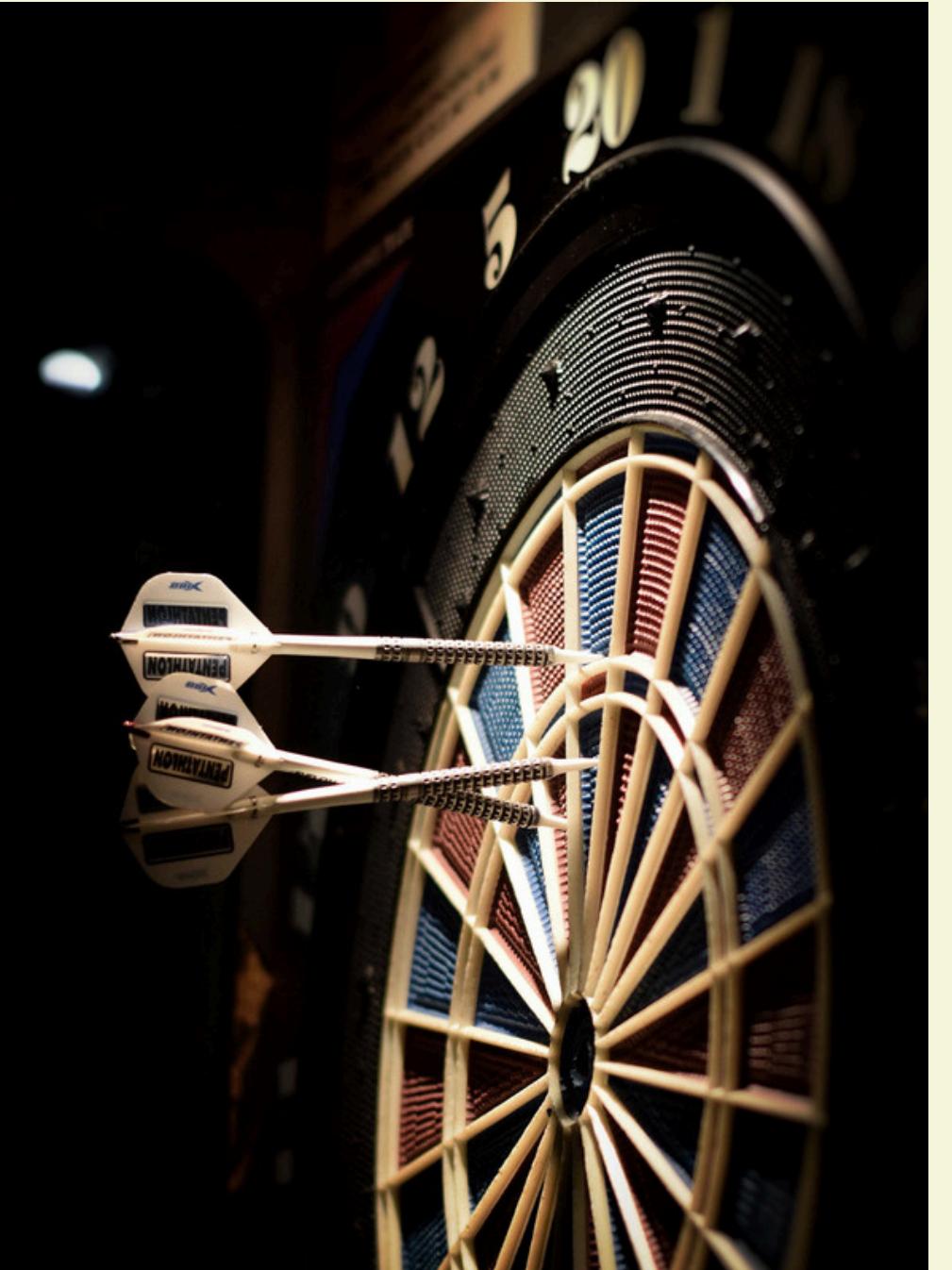
¿Qué tipo de personas tenían mayor probabilidad de sobrevivir al hundimiento del Titanic?



OBJETIVO



El propósito de este trabajo es desarrollar e implementar un modelo de aprendizaje automático supervisado que permita predecir la probabilidad de supervivencia de los pasajeros del RMS Titanic, a partir de las variables proporcionadas en el conjunto de datos original distribuido por Kaggle.



ANTECEDENTES



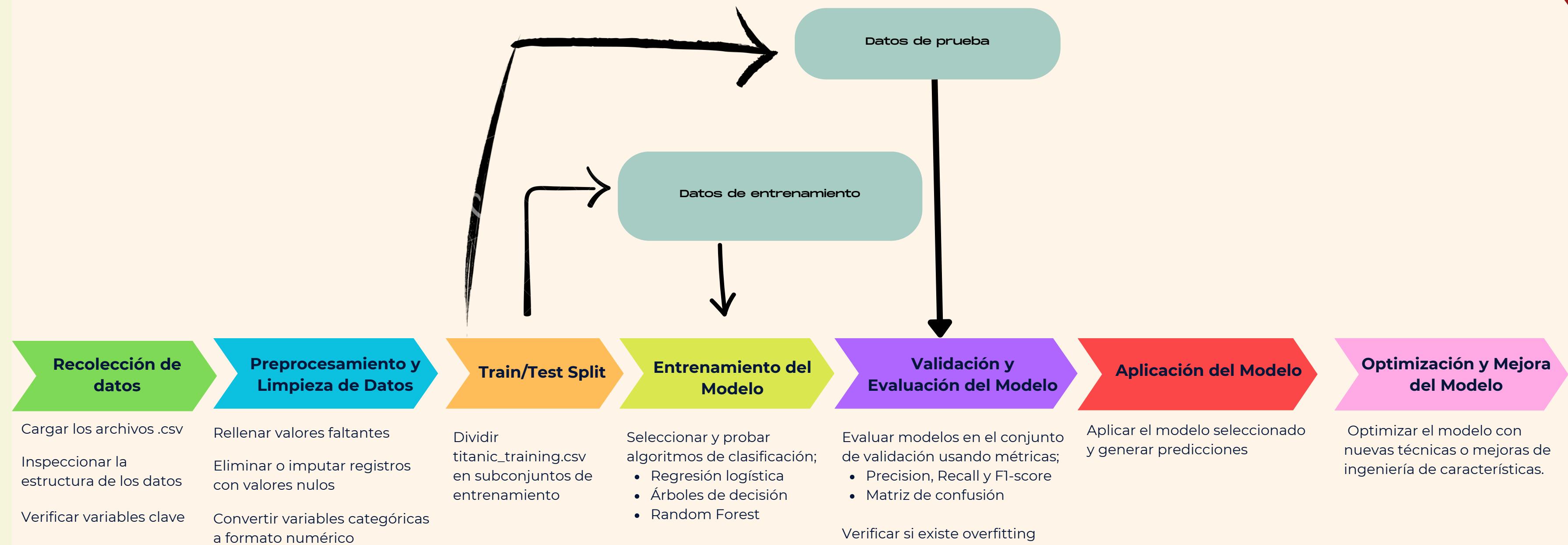
- **Classification of Titanic Passenger Data and Chances of Surviving the Disaster:** Usando Weka, identifican factores clave (clase, edad, embarque) relacionados con la supervivencia mediante minería de datos tradicional.
- **Titanic Disaster Prediction Based on Machine Learning Algorithms:** Implementa árboles de decisión y bosques aleatorios, destacando sexo, edad y clase como predictores principales.

ANTECEDENTES



- **Comparison of Machine Learning Classification Models in Predicting The Titanic Survival Rate:** Compara múltiples algoritmos (logística, Random Forest, XGBoost), concluyendo que Random Forest es el más preciso.
- **Research on Titanic Survival Prediction Based on Machine Learning Method:** Crea un modelo de votación dura con tres algoritmos, alcanzando 87.64% de precisión en la predicción de supervivencia.

DESCRIPCIÓN SOLUCIÓN/PROCESO





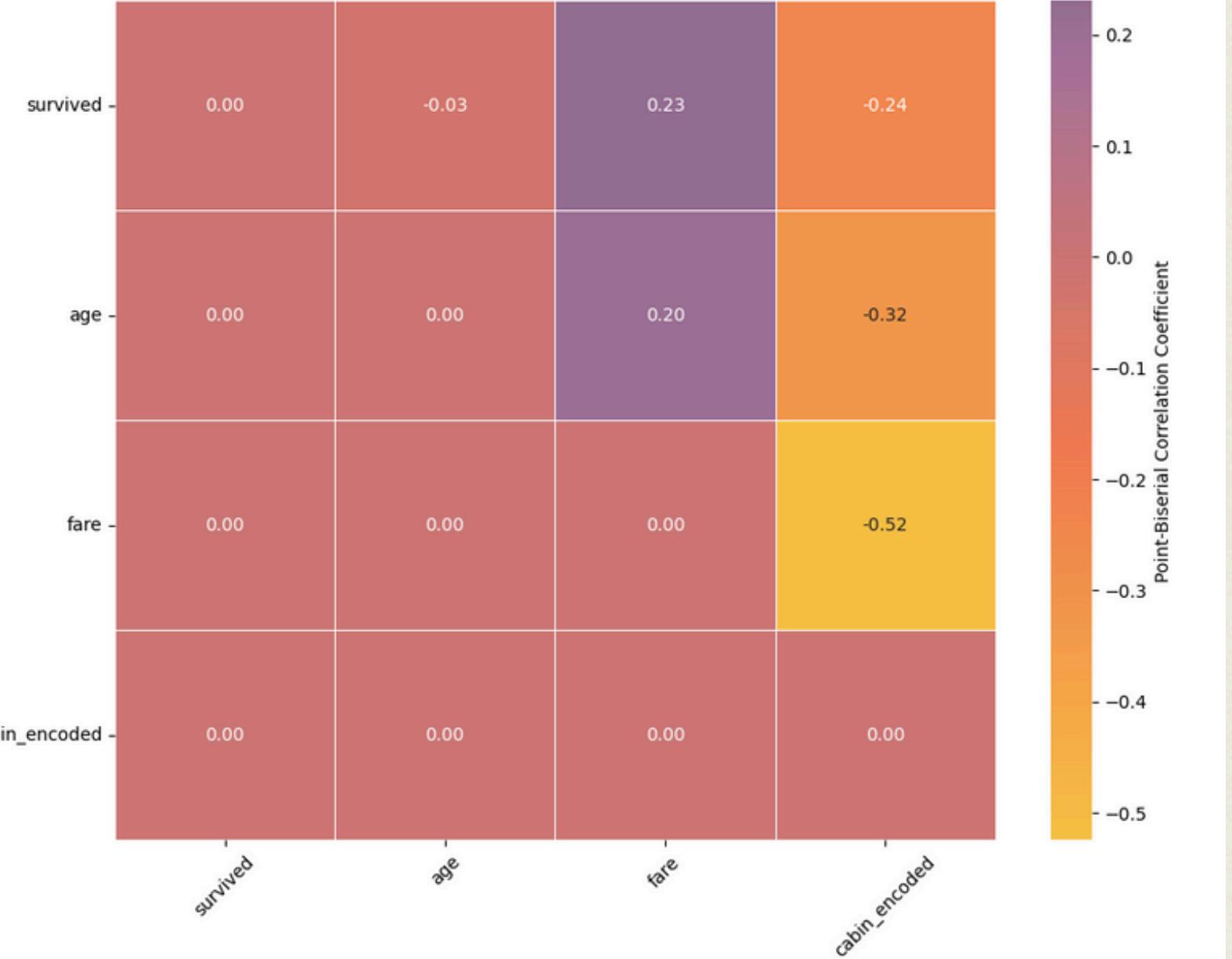
- Dataset: titanic_trainig.csv, 1000 registros, 10 columnas: survived, pclass, sex, age, sibsp, parch, ticket, fare, cabin, embarked.

- Se eliminó un registro (survived nulo).
- Codificación ordinal de variables categóricas.
- Punto biserial para variables continuas y survived.
- Crammer's V para variables categóricas y survived.
- Dataset desbalanceado: pocos sobrevivientes.

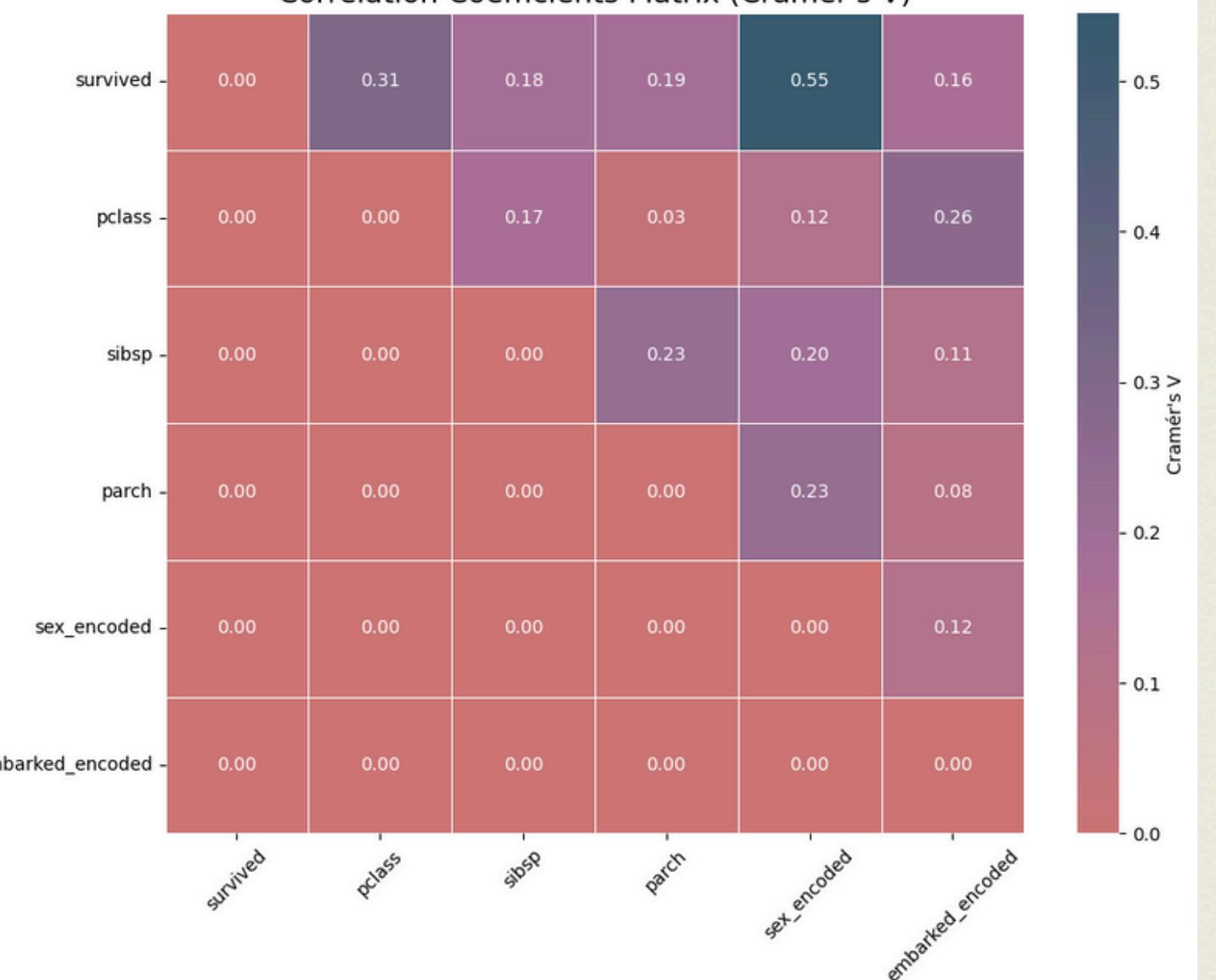
Nº	Columna	Valores no nulos	Tipo de dato	Observaciones
0	survived	999	float64	Variable objetivo (0 = No, 1 = Sí)
1	pclass	999	float64	Clase del boleto (1, 2, 3)
2	sex	999	object	Género (male, female)
3	age	804	float64	Edad (valores faltantes)
4	sibsp	999	float64	Nº de hermanos/cónyuge a bordo
5	parch	999	float64	Nº de padres/hijos a bordo
6	ticket	999	object	Número de boleto
7	fare	998	float64	Tarifa pagada
8	cabin	227	object	Cabina asignada (muchos nulos)
9	embarked	997	object	Puerto de embarque (C, Q, S)



Point-Biserial Correlation Coefficient Matrix



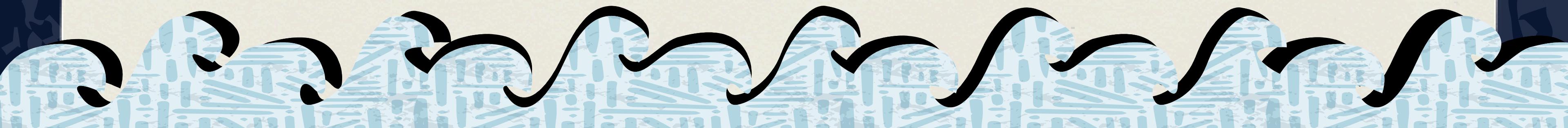
Correlation Coefficients Matrix (Cramér's V)





Las variables predictoras se organizarían en orden de importancia de la forma:

- Sexo
- Clase del boleto/Tarifa de viaje
- Habitación (Está muy asociada a la tarifa de viaje debido a que las habitaciones se organizaban en las secciones A, B, C, D y E, siendo la A la más costosa y la E la menos costosa. Se escogió el codificador ordinal para mantener de alguna forma este comportamiento).
- Cantidad de padres o hijos a bordo.
- Cantidad de hermanos o cónyugues a bordo.
- Lugar de donde embarcó el pasajero.
- Edad (Este es un caso especial porque el valor P del coeficiente de asociación es bastante alto, así que a futuro podríamos probar incluyendola y quitándola para ver qué resultados obtenemos).



MODELLAMIENTO

Con el objetivo de evaluar el rendimiento de distintos enfoques de clasificación para predecir la supervivencia de pasajeros del Titanic, se implementaron y compararon seis modelos supervisados: Regresión Logística, Random Forest, XGBoost, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) y Naive Bayes. Siendo Survived = 1 de mayor interés.

	MODELO	EXACTITUD (TEST)	RECALL (CLASE 1)	PRECISION (CLASE 1)	F1-SCORE (CLASE 1)
0	Regresión Logística Original	0.795	0.7042	0.7143	0.7092
1	Random Forest (sin GS)	0.79	0.7183	0.6986	0.7083
2	XGBoost (sin GS)	0.805	0.7606	0.7105	0.7347
3	SVM (sin GS)	0.81	0.6761	0.7619	0.7164
4	KNN (sin GS)	0.785	0.6761	0.7059	0.6906
5	Gaussian Naive Bayes (sin GS)	0.78	0.7465	0.6709	0.7067

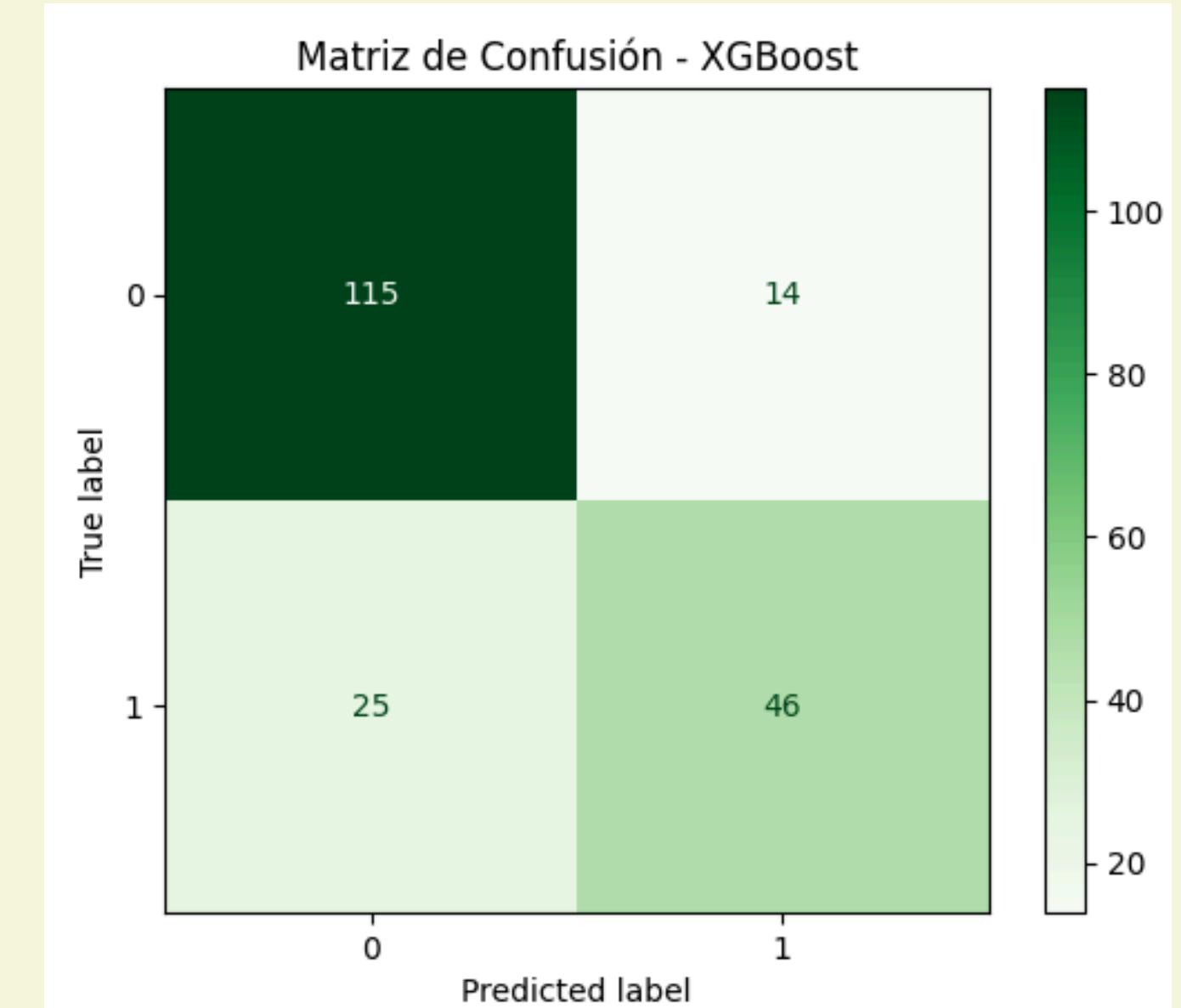
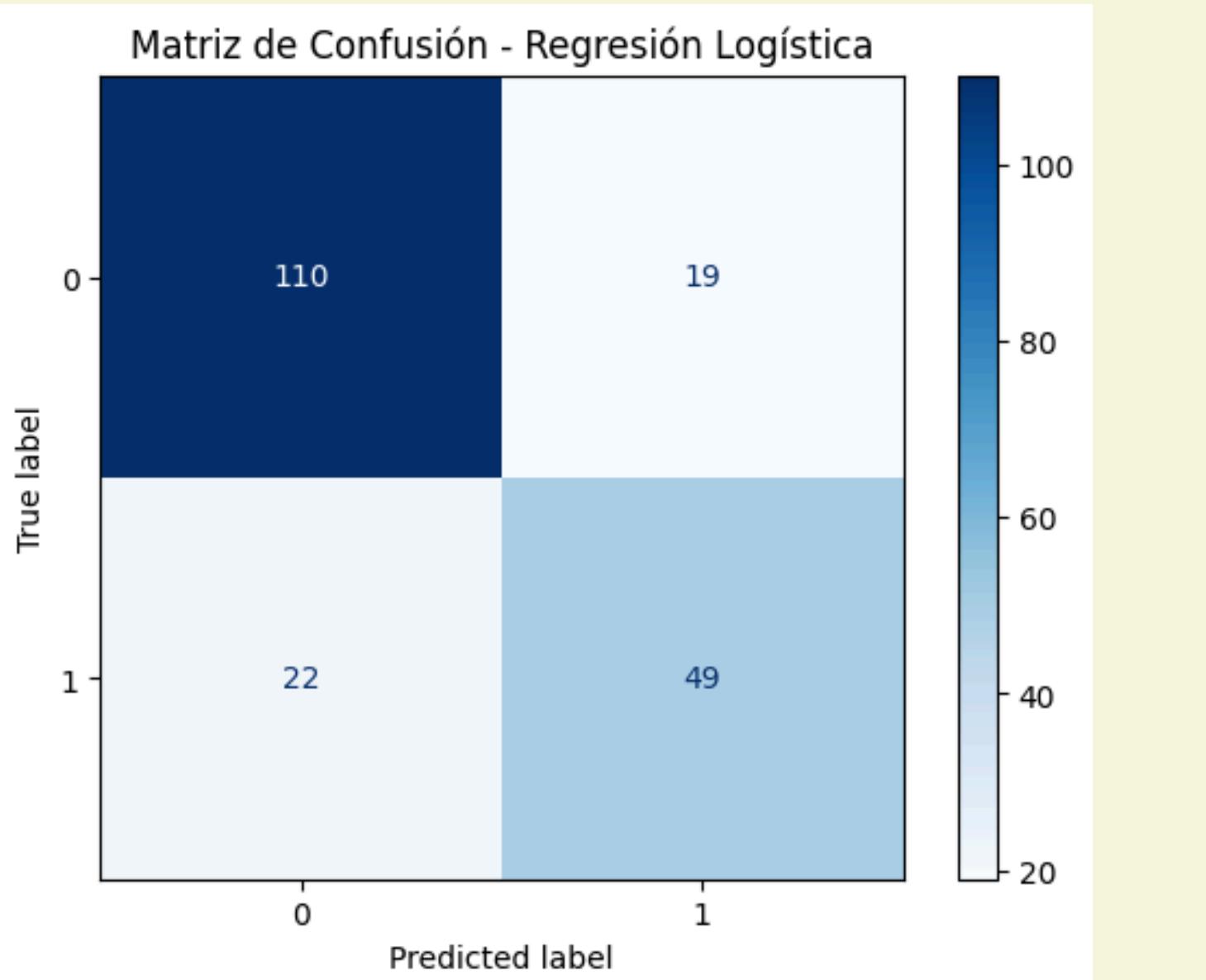
MODELLAMIENTO



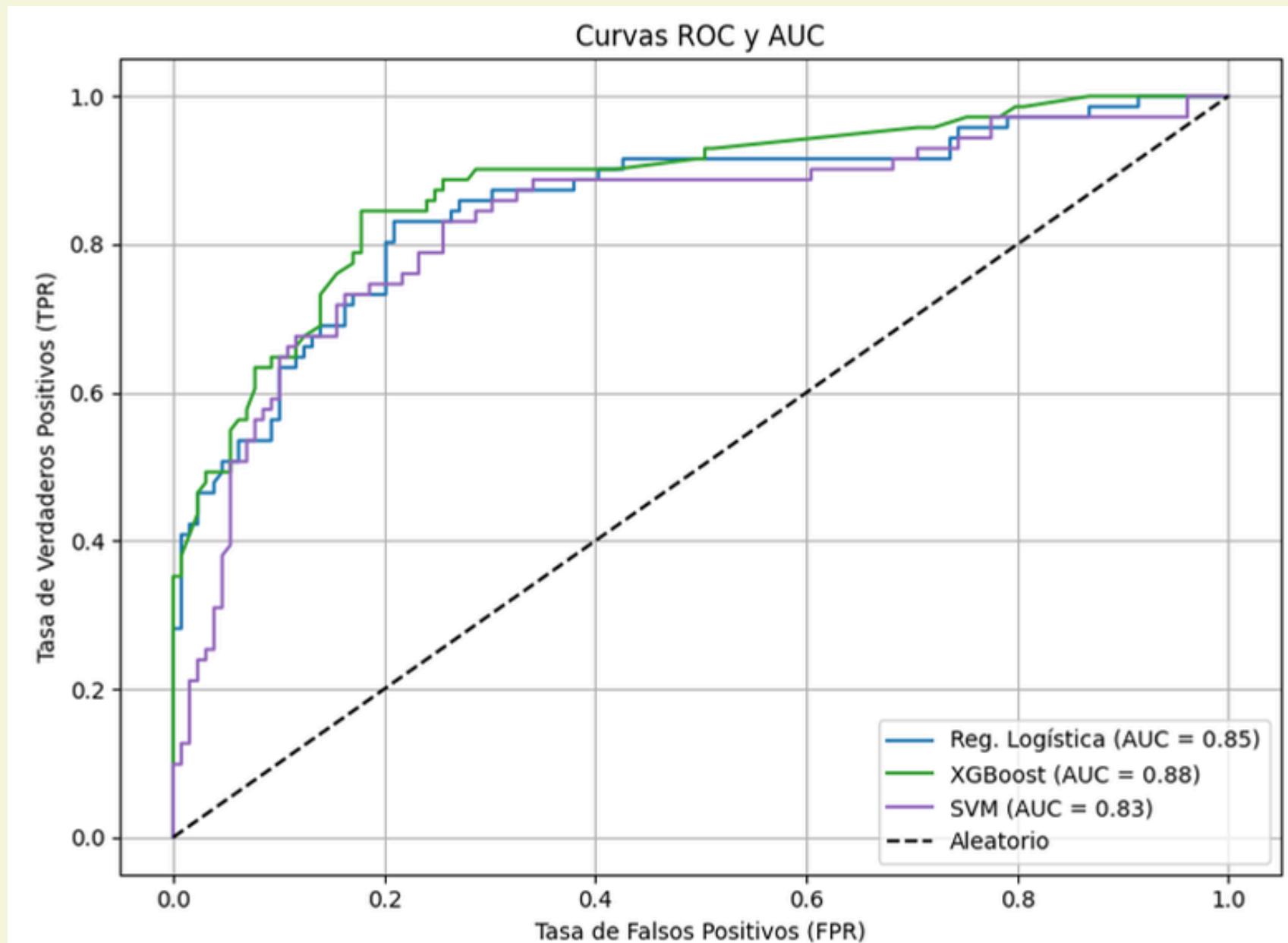
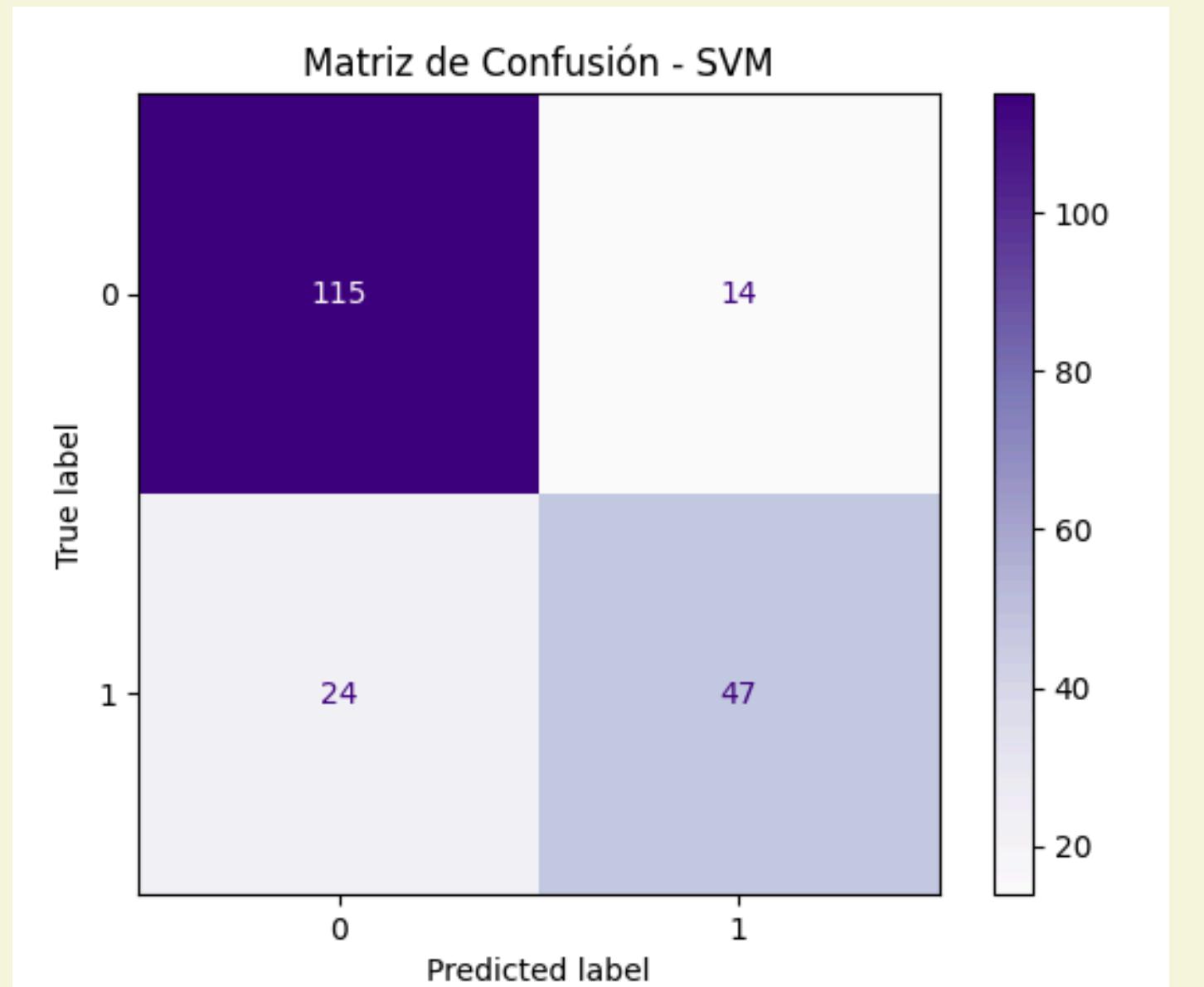
Adicionalmente, se hizo validación cruzada y optimización de los hiperparámetros de algunos modelos utilizados en la búsqueda de un mejor rendimiento para clasificar a los sobrevivientes del accidente, se observó que la variación de rendimiento de los nuevos modelos fue marginal.

	MODELO	EXACTITUD (TEST)	RECALL (CLASE 1)	PRECISION (CLASE 1)	F1-SCORE (CLASE 1)
0	Regresión Logística (GridSearchCV)	0.795	0.6901	0.7206	0.705
1	XGBoost (GridSearchCV)	0.805	0.6479	0.7667	0.7023
2	SVM (GridSearchCV)	0.81	0.662	0.7705	0.7121

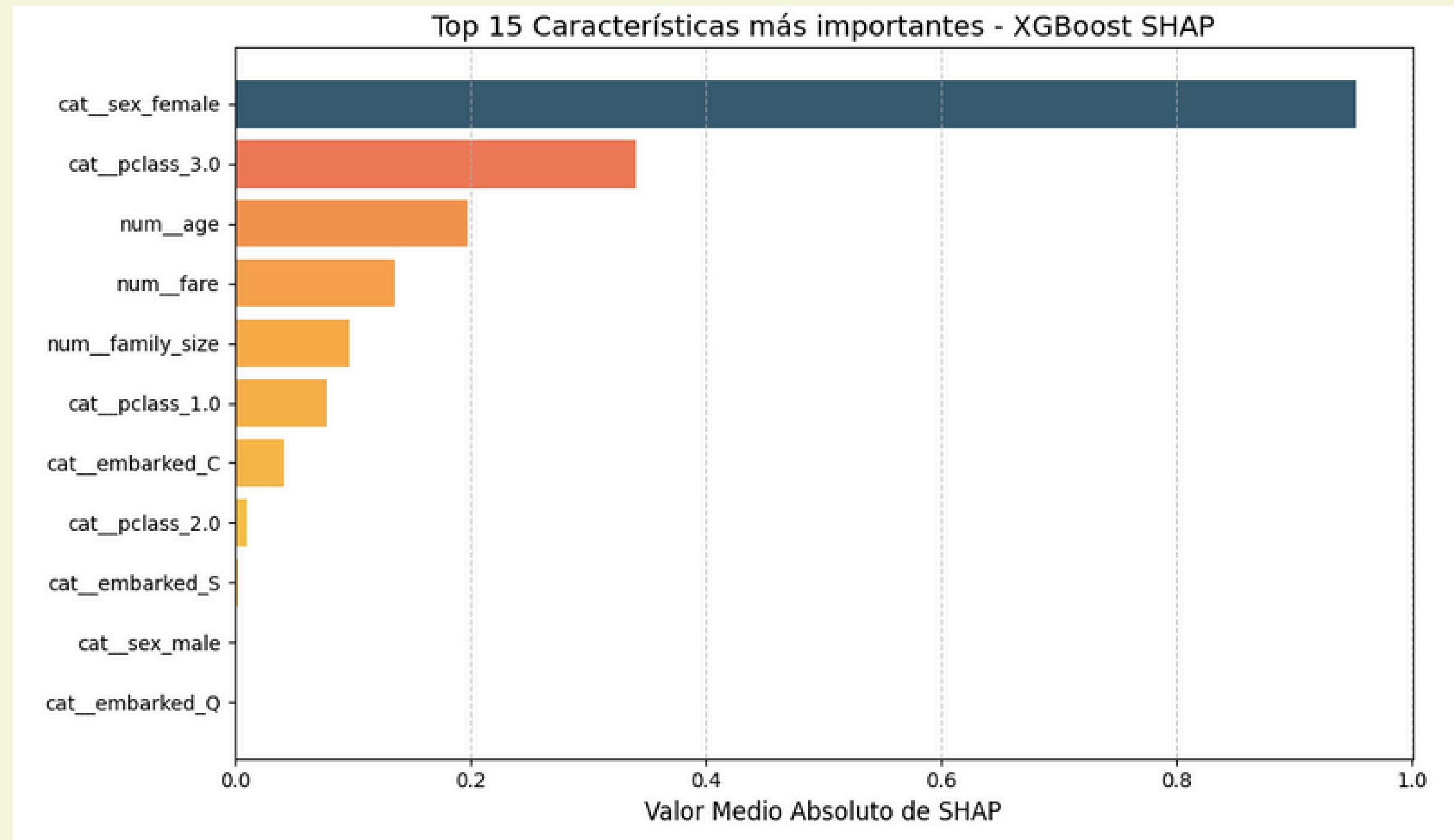
VISUALIZACIÓN E INTERPRETABILIDAD



VISUALIZACIÓN E INTERPRETABILIDAD



VISUALIZACIÓN E INTERPRETABILIDAD



CONCLUSIONES

- Los resultados refuerzan un hallazgo crítico: los pasajeros de tercera clase y los hombres tuvieron una probabilidad significativamente menor de sobrevivir, lo que sugiere un acceso desigual a los botes salvavidas.
- Tras optimización, los modelos XGBoost y SVM ofrecieron un equilibrio superior entre exactitud, precisión y sensibilidad, logrando predecir correctamente una mayor proporción de sobrevivientes.
- A pesar de su simplicidad, la Regresión Logística mostró un rendimiento competitivo y una interpretabilidad valiosa.
- Las técnicas de interpretabilidad (SHAP, matrices de confusión, curvas ROC) no solo ayudaron a entender cómo funcionan los modelos, sino que también permitieron conectar patrones numéricos con decisiones humanas históricas reales.



GRACIAS!