Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Ciudad de México

Tecnológico de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

Evidencia 2. Portafolio de implementación

Equipo 3:

Rosa Vanessa Palacios Beltrán	| A01652612
Irving Yael López Solis	|A01664809
Diego Aguilar Torres	|A01657884
Santiago Calderón Ortega	|A01663888
Cynthia Amador Santiago	|A01737854
David Alberto Padrón Sánchez	|A01663806
Katia Geraldine Vidals Estrever	|A01657587

Profesor Cesar David Betancourt Adame

Grupo 100

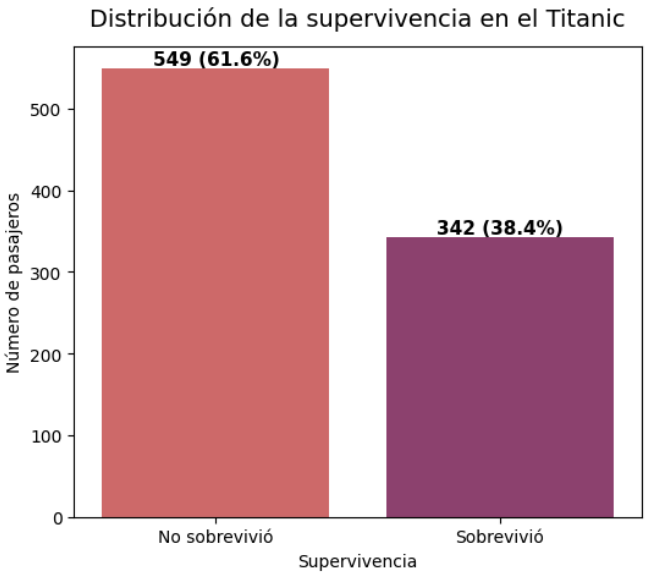
9 sept 2025

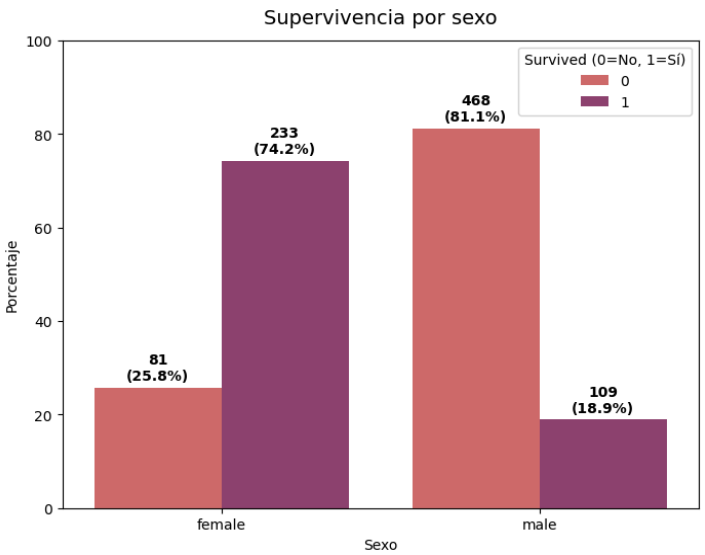
Introducción

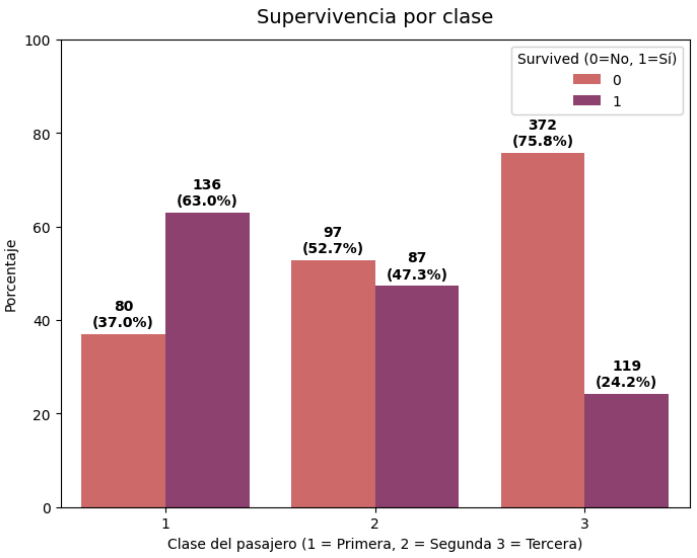
Este proyecto tiene como objetivo predecir la supervivencia de pasajeros del Titanic mediante un análisis de los datos que incluyó un proceso de **ETL** (extracción, transformación y carga) para limpiar el dataset, imputar valores nulos, crear nuevas variables y codificar variables categóricas. También un análisis exploratorio de datos (EDA) para identificar patrones en sexo, edad, clase y tarifas. Se realizó la aplicación de modelos de **clasificación supervisada**: Regresión Logística, Random Forest, K-Nearest Neighbors (KNN) y Support Vector Machine (SVM) como línea base. Para el **análisis no supervisado** con K-means (K=4) para segmentar pasajeros e interpretar arquetipos de supervivencia. Finalmente la evaluación de estrategias híbridas, tanto la Ruta A (Mixture of Experts) que combina modelos, como la Ruta B (Cluster como feature adicional) que integra los clústeres como nueva variable predictiva, con el fin de entender qué factores influyeron en la supervivencia y comparar distintos enfoques de modelado.

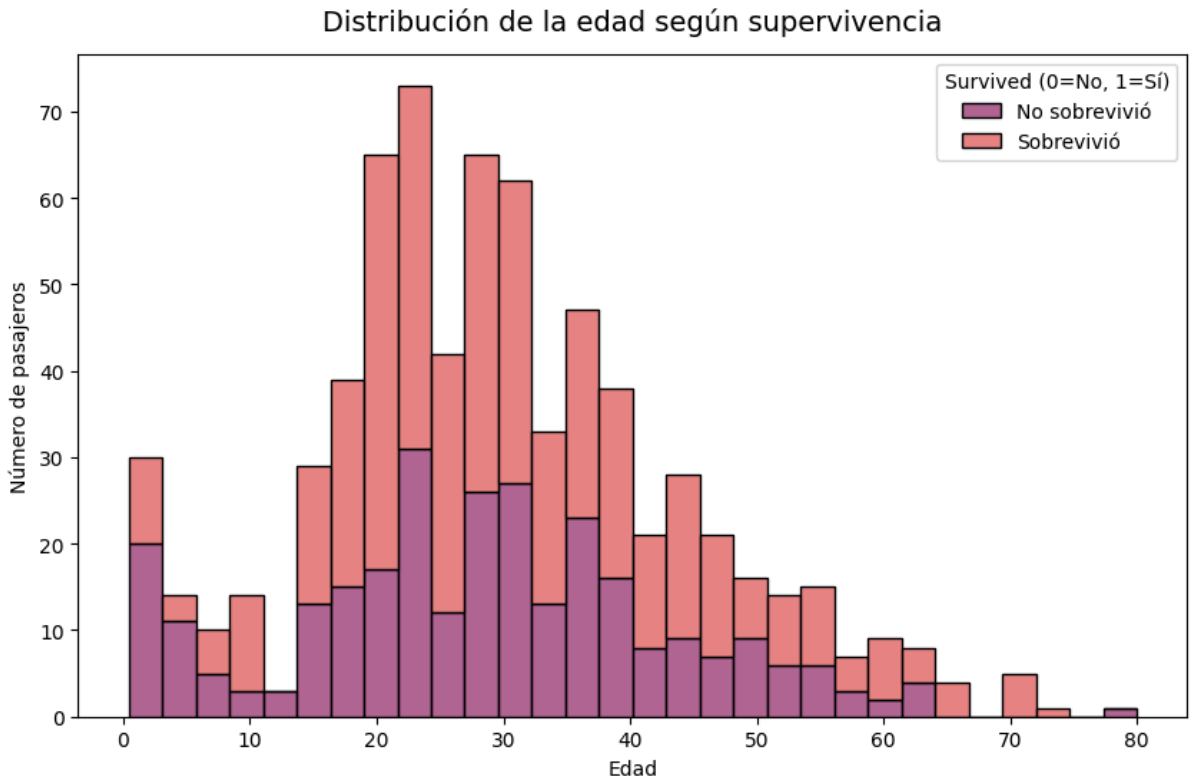
Análisis Exploratorio

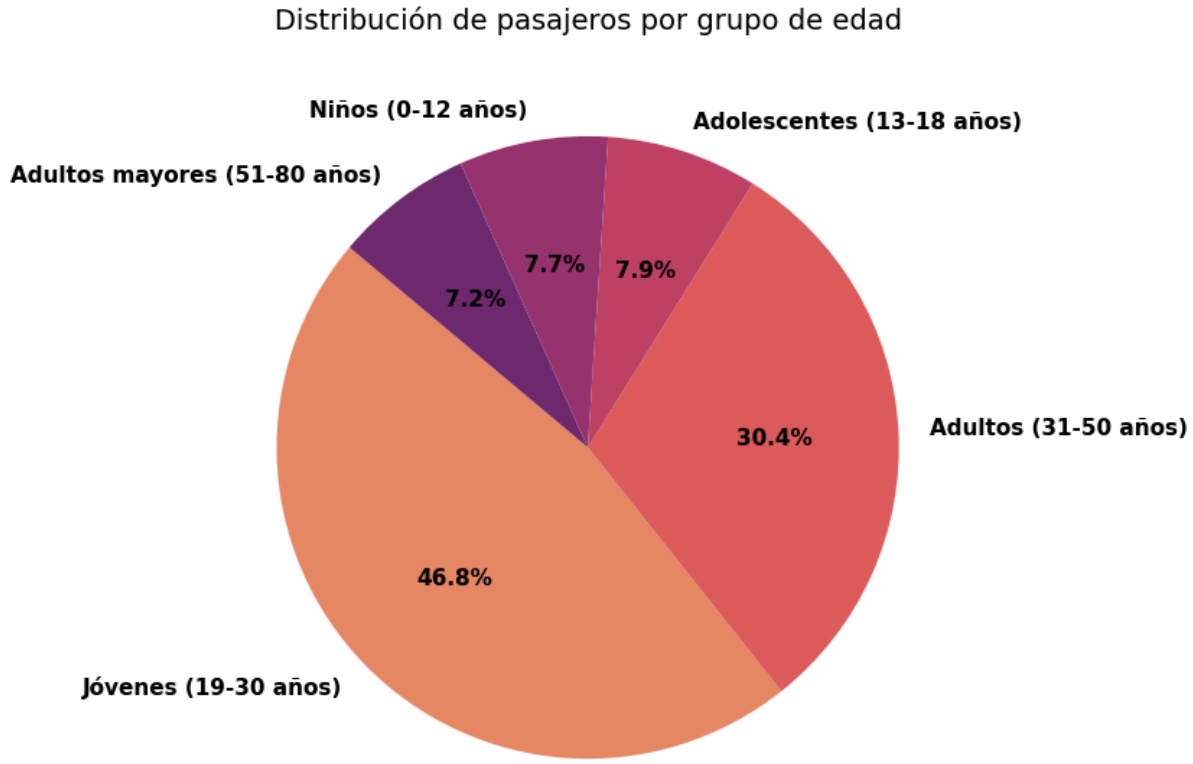
Se realizó un análisis inicial para conocer la distribución de los datos. Entre los hallazgos: el 38.3% de los pasajeros sobrevivió, las mujeres tuvieron mayor probabilidad de supervivencia y los pasajeros de 1ª clase presentan ventajas claras frente a 2ª y 3ª clase. La edad también fue un factor, siendo los niños los más beneficiados.







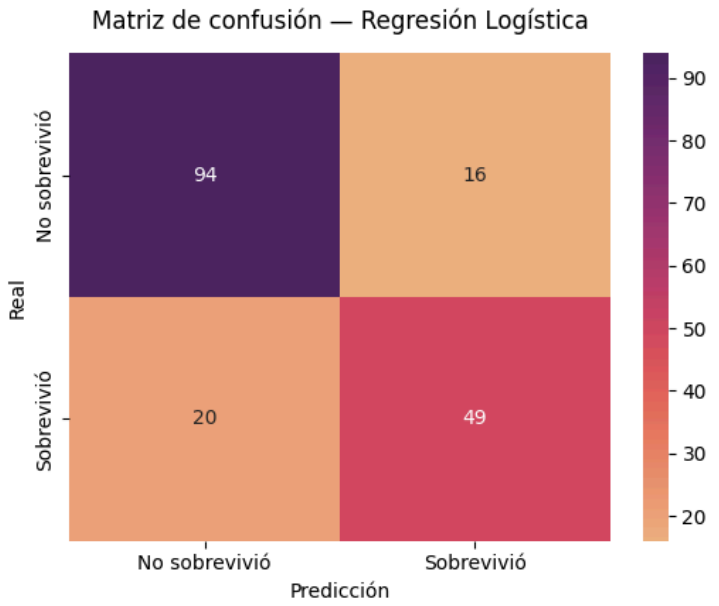

Distribución de la edad según supervivencia
Survived (0=No, 1=Sí)
No sobrevivió
Sobrevivió
Número de pasajeros
Edad
0
10
20
30
40
50
60
70
80

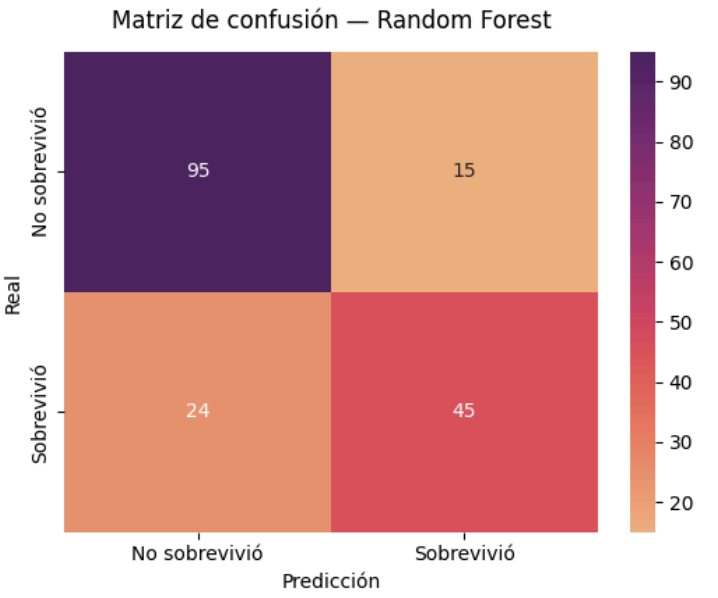

Distribución de pasajeros por grupo de edad
Niños (0-12 años)
Adolescentes (13-18 años)
Adultos mayores (51-80 años)
7.7%
7.9%
7.2%
30.4%
Adultos (31-50 años)
46.8%
Jóvenes (19-30 años)

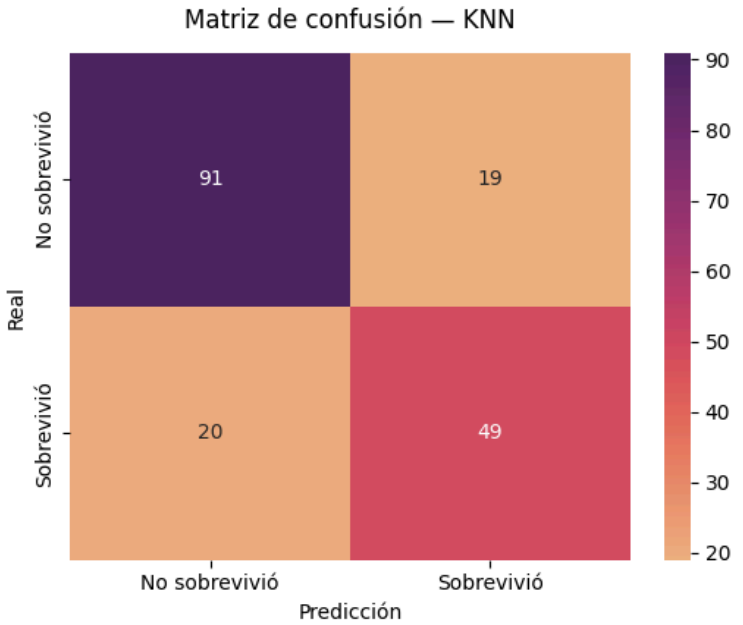
Modelado Supervisado (Baseline)

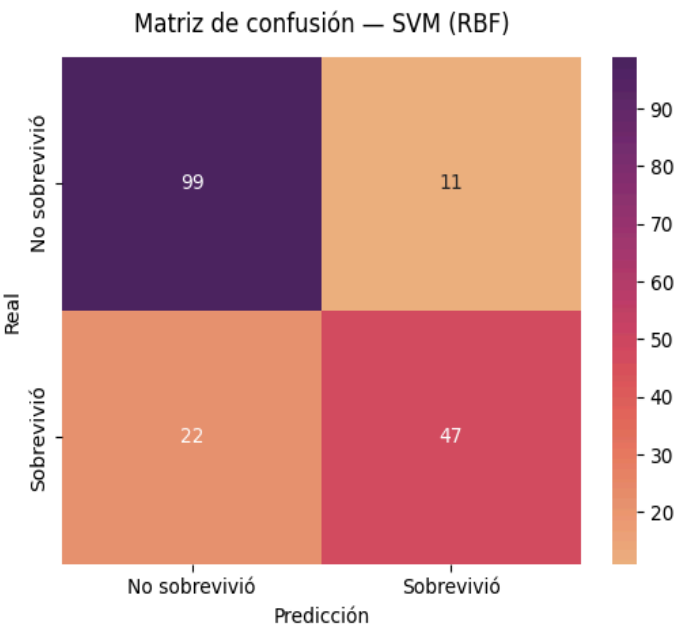
Se aplicaron cuatro modelos supervisados: Regresión Logística, Random Forest, K-Nearest Neighbors (KNN) y Support Vector Machine (SVM). Los resultados mostraron que SVM fue el más competitivo entre los modelos baseline.

Modelo	Accuracy	Precision	Recall	AUC
Regresión Logística	0.8338	0.7998	0.7571	0.8696
Random Forest	0.8114	0.7642	0.7369	0.8623
KNN	0.8159	0.7925	0.7047	0.8622
SVM	0.8100	0.8100	0.6800	0.8370









Comparación de matrices de confusión

En las figuras anteriores se muestran las matrices de confusión para los cuatro modelos supervisados evaluados (Regresión Logística, Random Forest, KNN y SVM). Estas matrices permiten visualizar el desempeño de cada clasificador al identificar correctamente a los pasajeros que **sobrevivieron** y los que **no sobrevivieron**.

- **Regresión Logística y KNN** presentan un equilibrio aceptable entre verdaderos positivos y negativos, aunque con algunos falsos negativos (pasajeros que sobrevivieron pero fueron clasificados como no sobrevivientes).
- **Random Forest** muestra un comportamiento similar, pero con una ligera mayor cantidad de falsos negativos respecto a la regresión logística.
- **SVM** es el que logra la mejor capacidad para identificar a los que no sobrevivieron (mayor número de verdaderos negativos), aunque aún mantiene falsos negativos en los sobrevivientes.

Las matrices evidencian las fortalezas y debilidades de cada algoritmo: mientras algunos priorizan la detección de los no sobrevivientes, otros buscan mayor balance entre clases, lo cual se refleja en las métricas de precisión, recall y AUC.

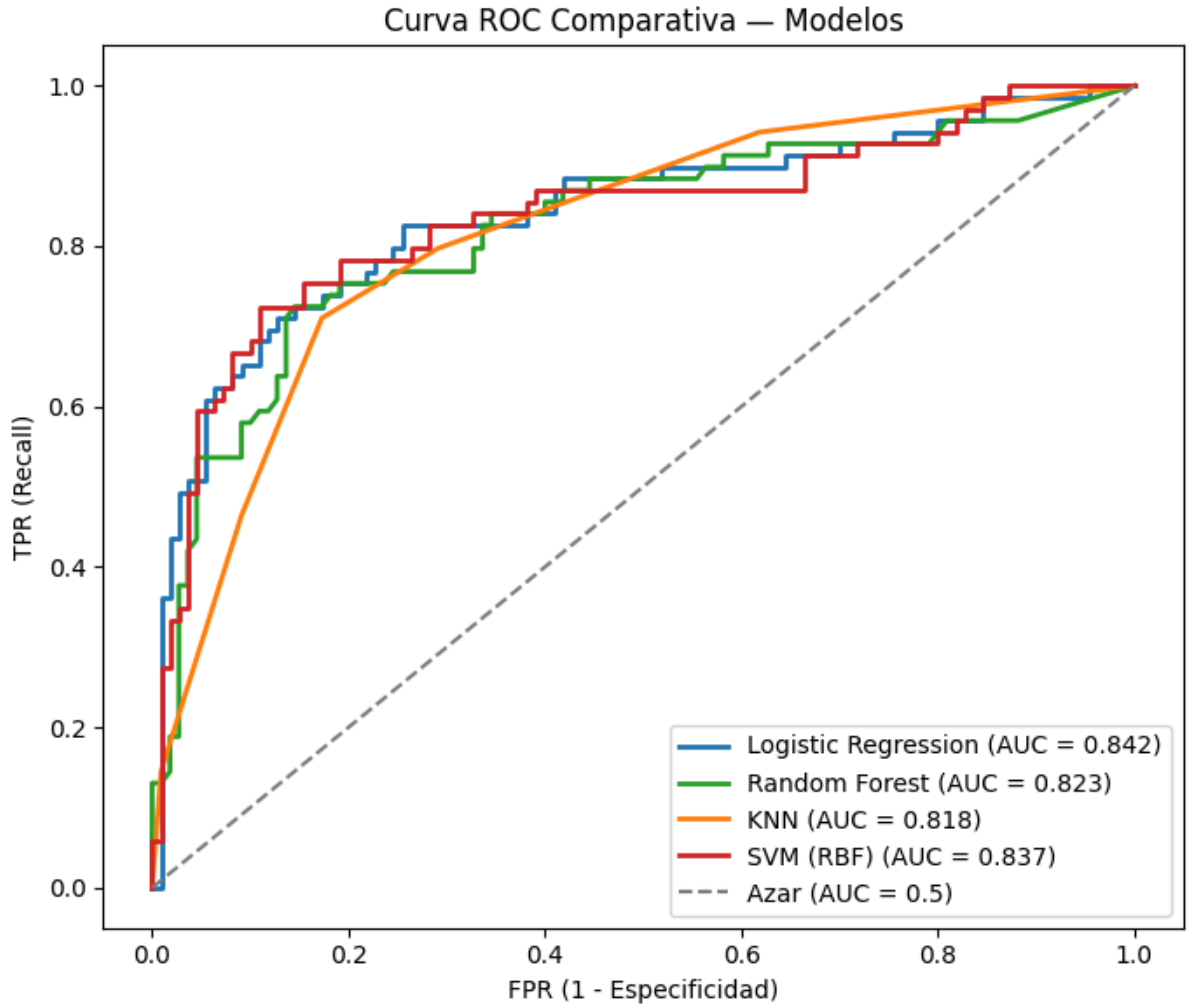
Comparación de curvas ROC

Las curvas ROC de los cuatro modelos supervisados: Regresión Logística, Random Forest, KNN y SVM. La curva ROC permite evaluar la capacidad de cada modelo para distinguir entre las dos clases (sobrevivió/no sobrevivió), mientras que el valor AUC resume este desempeño en una métrica global.

- **Regresión Logística** obtuvo el mayor AUC (0.842), mostrando un excelente balance entre sensibilidad (recall) y especificidad.
- **SVM** alcanzó un valor cercano (0.837), confirmando también un buen desempeño.
- **Random Forest** obtuvo un AUC de 0.823, ligeramente menor pero igualmente competitivo.
- **KNN** fue el que tuvo el valor más bajo (0.818), aunque se mantiene por encima del umbral de 0.8 considerado como buen rendimiento.

La línea diagonal representa el **modelo aleatorio** (AUC = 0.5), que sirve como referencia mínima.

En conclusión, todos los modelos presentan un desempeño adecuado, destacando la Regresión Logística como el mejor clasificador en términos de AUC, seguida de cerca por SVM.



Análisis No Supervisado (K-means)

Se aplicó K-means con K=4 para segmentar a los pasajeros en grupos. Cada clúster mostró diferentes tasas de supervivencia, revelando patrones como: mujeres jóvenes de clases altas con mayor supervivencia y hombres de 3ª clase con tasas muy bajas.

Interpretación de arquetipos por clúster (K-means)

A partir del análisis no supervisado con K-means (K=4) se identificaron cuatro arquetipos de pasajeros con comportamientos diferenciados en cuanto a supervivencia.

- **El Clúster 3** mostró la mayor tasa de supervivencia (≈79%), compuesto en su mayoría por mujeres (67%) de segunda clase, con tarifas moderadas

y familias pequeñas, lo que corresponde al perfil de “mujeres de clase media con alta probabilidad de supervivencia”.

- **El Clúster 1**, con una supervivencia intermedia (≈47%), estuvo conformado principalmente por adultos mayores, en su mayoría hombres de primera clase.
- **El Clúster 2** presentó una tasa de supervivencia cercana al 44% y agrupó a familias numerosas (mediana de 4 integrantes), mayormente de tercera clase y con presencia equilibrada de mujeres, lo que lo perfila como “familias numerosas de tercera clase con supervivencia baja-intermedia”.
- **El Clúster 0** mostró la menor tasa de supervivencia (≈26%), siendo el grupo más grande (≈481 pasajeros), caracterizado por hombres jóvenes (≈74%) de tercera clase.

En conclusión, los clústeres reflejan patrones claros donde la clase social, el género y el tamaño familiar resultaron factores determinantes en la probabilidad de supervivencia, confirmando la regla histórica del Titanic: ***“Mujeres y niños primero, especialmente de clases altas”***.

Estrategias Híbridas

Se probaron dos estrategias híbridas:

- **Ruta A (Mixture of Experts):** combinación de varios modelos.
- **Ruta B (Cluster como feature adicional):** incorporación de los clústeres como variable.

Modelo	**Accuracy**	**Precision**	**Recall**	**AUC**
Regresión Logística	0.8338	0.7998	0.7571	0.8696
Random Forest	0.8114	0.7642	0.7369	0.8623
KNN	0.8159	0.7925	0.7047	0.8622
SVM	0.8100	0.8100	0.6800	0.8370
Ruta A MoE	**0.7877**	**0.7460**	**0.6811**	**0.7678**
Ruta B Cluster (Regresión Logística)	**0.8305**	**0.7961**	**0.7513**	**0.8695**

Discusión

Se gana interpretabilidad al identificar arquetipos con diferencias reales de supervivencia (ejemplo: Cluster 2 ≈ 77.6% vs. Cluster 1 ≈ 23.8%), lo que permite explicar patrones por perfil (mujeres de 1ª clase acompañadas vs. hombres de 3ª clase solos); sin embargo, se pierde simplicidad y rendimiento: el pipeline se vuelve más complejo, aumenta el riesgo de sobreajuste si los clusters no son estables y, en este caso, no hubo mejora de métricas frente al baseline.

El modelo no mejoró. La Regresión Logística baseline sigue siendo la mejor por AUC (≈ 0.870 en CV); al usar el cluster como feature (Ruta B) las métricas quedan prácticamente iguales (AUC ≈ 0.8695), mientras que la estrategia Mixture of Experts por cluster (Ruta A) rinde peor (AUC ≈ 0.768). En síntesis, introducir clustering no superó al baseline y, en la variante MoE, lo degradó.

Si bien el enfoque MoE permite asignar un *VotingClassifier* específico a cada cluster, lo que en teoría podría captar patrones de manera más precisa, en la práctica cada cluster contiene una cantidad de datos considerablemente menor que el dataset original. Esto implica que los modelos entrenados en cada subgrupo disponen de menos información para aprender, lo que reduce su capacidad de generalización. Como consecuencia, los resultados obtenidos con el MoE híbrido no superan a los alcanzados por los modelos aplicados individualmente, ya que estos últimos cuentan con un mayor volumen de observaciones y, por lo tanto, con más evidencia para estimar relaciones entre variables y predecir correctamente.

Por otra parte, se detectó fuga al ajustar K-means con todo el dataset y reutilizar etiquetas en validación, también por imputaciones globales (Age por mediana de Title y Embarked por moda calculadas con todo el conjunto) y por features con conocimiento global (ejemplo: TicketGroupSize/InGroup) derivadas del dataset completo. Para mitigarlo, todo preprocesamiento y clustering deben ir en un pipeline y recalcular solo con el fold de entrenamiento en cada split/CV, evitando compartir estadísticas o transformaciones con la validación.

Conclusiones Finales

Nuestro análisis mostró que factores como sexo, edad y clase social fueron determinantes en la supervivencia del Titanic. Los modelos supervisados alcanzaron buen rendimiento (AUC > 0.8), destacando Regresión Logística en AUC y SVM en precisión.

La Regresión Logística se mantiene como el mejor modelo por AUC (~0.87), equilibrando rendimiento, el clustering aportó interpretabilidad al revelar arquetipos con diferencias claras de supervivencia, pero no mejoró las métricas: la Ruta B (clúster como feature) quedó esencialmente igual (AUC≈0.8695) y la Ruta A (Mixture of Experts) degradó el desempeño (AUC≈0.768). Esto indica que, con el tamaño y señal del dataset, la fragmentación por clúster no generaliza mejor que un baseline.

El análisis no supervisado con K-means permitió identificar arquetipos de pasajeros, confirmando la regla histórica de “mujeres y niños primero, especialmente de clases altas”.