

TITANIC: MACHINE LEARNING FROM DISASTER

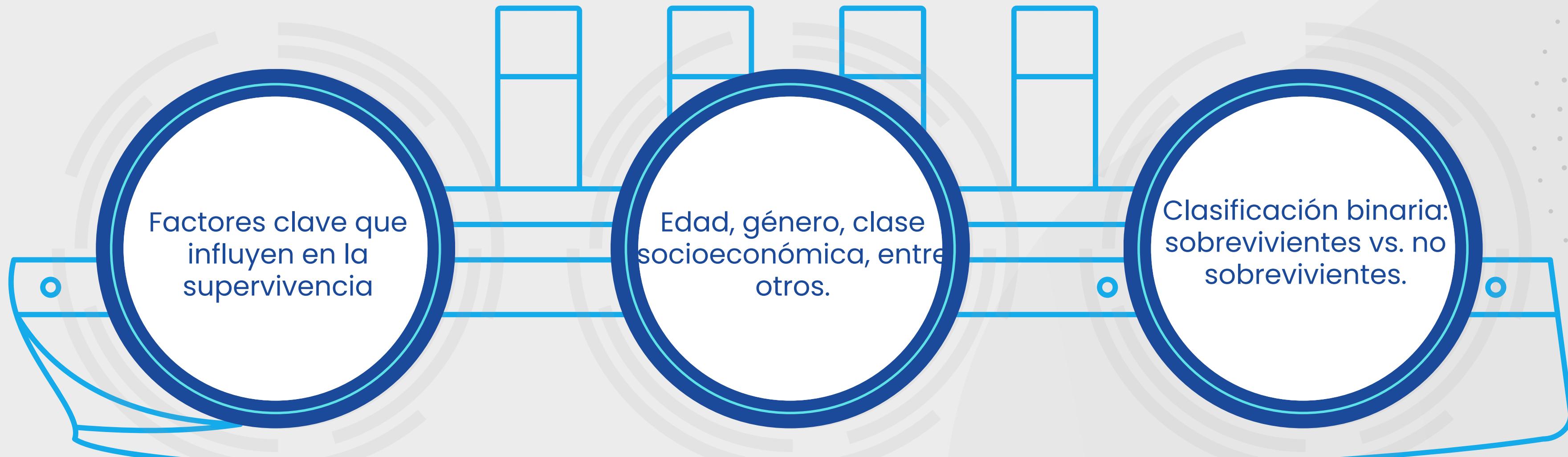
UN MODELO PREDICTIVO PARA LA SOBREVIVENCIA

Catherine Rojas
Adrian Pineda
Rogelio Lizárraga
Luis López
Rodolfo Cruz



NUESTRO RETO

Predecir la supervivencia de los pasajeros del Titanic, un transatlántico que se hundió en 1912, causando la muerte de más de 1500 personas



OBJETIVOS

La Venta es resultado de un proceso que exige rigor metodológico, atención a las métricas y la gestión adecuada de la primera línea comercial.”



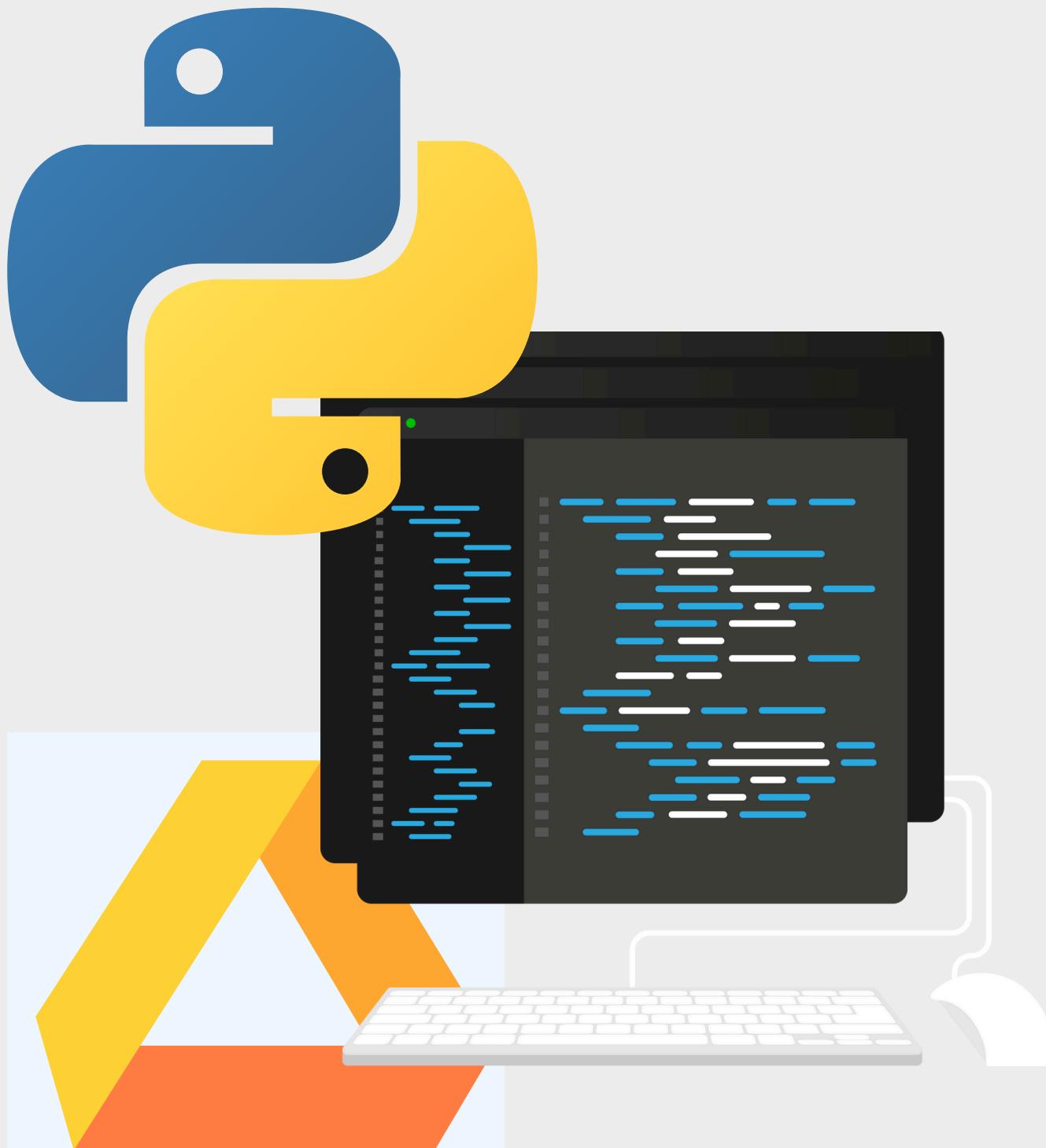
- Desarrollar un modelo predictivo
- Identificar factores clave
- Preparación de los datos



- Evaluación del modelo
- Optimización del modelo

HERRAMIENTAS Y RECURSOS UTILIZADOS

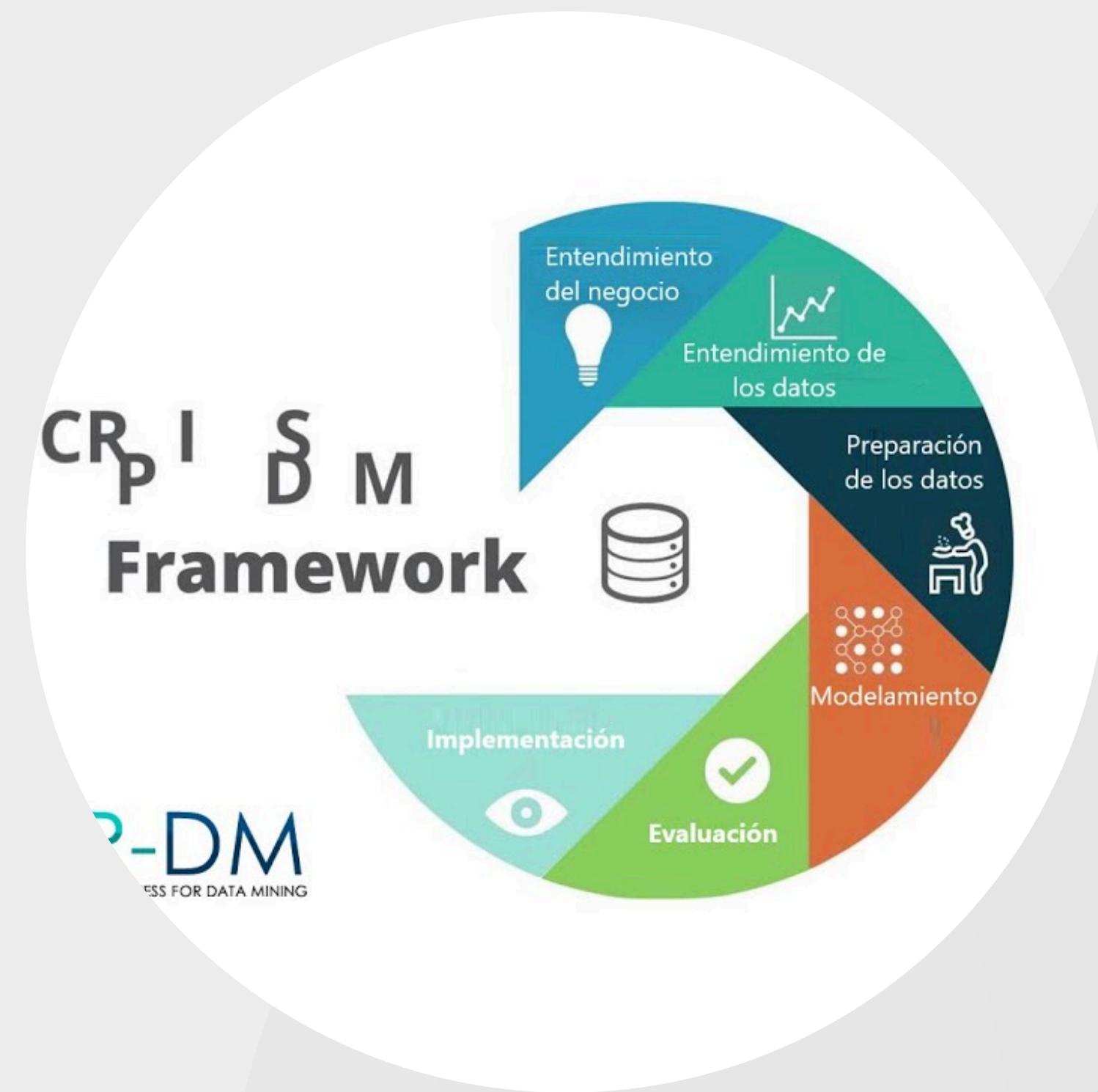
SOFTWARE Y FRAMEWORKS



- **Google Colab:** Plataforma para ejecución de notebooks en Python.
- **Pandas:** Manipulación y análisis de datos.
- **NumPy:** Cálculos numéricos y arrays.
- **Matplotlib & Seaborn:** Visualización de datos y gráficos.
- **Optuna:** Optimización automática de hiperparámetros de los algoritmos utilizados.
- **Scikit-learn:** Algoritmos de machine learning y optimización de modelos.

METODOLOGIA

- **Comprendión del problema:** Predecir la supervivencia de los pasajeros del Titanic.
- **Comprendión de los datos:** Analizamos el conjunto de datos proporcionado, identificando variables clave para la solución.
- **Preparación de los datos:** Limpieza, imputación y transformación del dataset seleccionado.
- **Modelado:** Implementamos diversos algoritmos de Machine Learning, ajustando hiperparámetros y aplicando técnicas de optimización.
- **Evaluación:** Evaluamos los modelos usando métricas de desempeño.



HERRAMIENTAS Y RECURSOS UTILIZADOS

HARDWARE Y SERVICIOS

- CPU
- Almacenamiento en Google Drive



Conjunto de Entrenamiento:

- 891 pasajeros.
- 12 variables diferentes incluyendo el resultado de supervivencia.

Conjunto de Prueba:

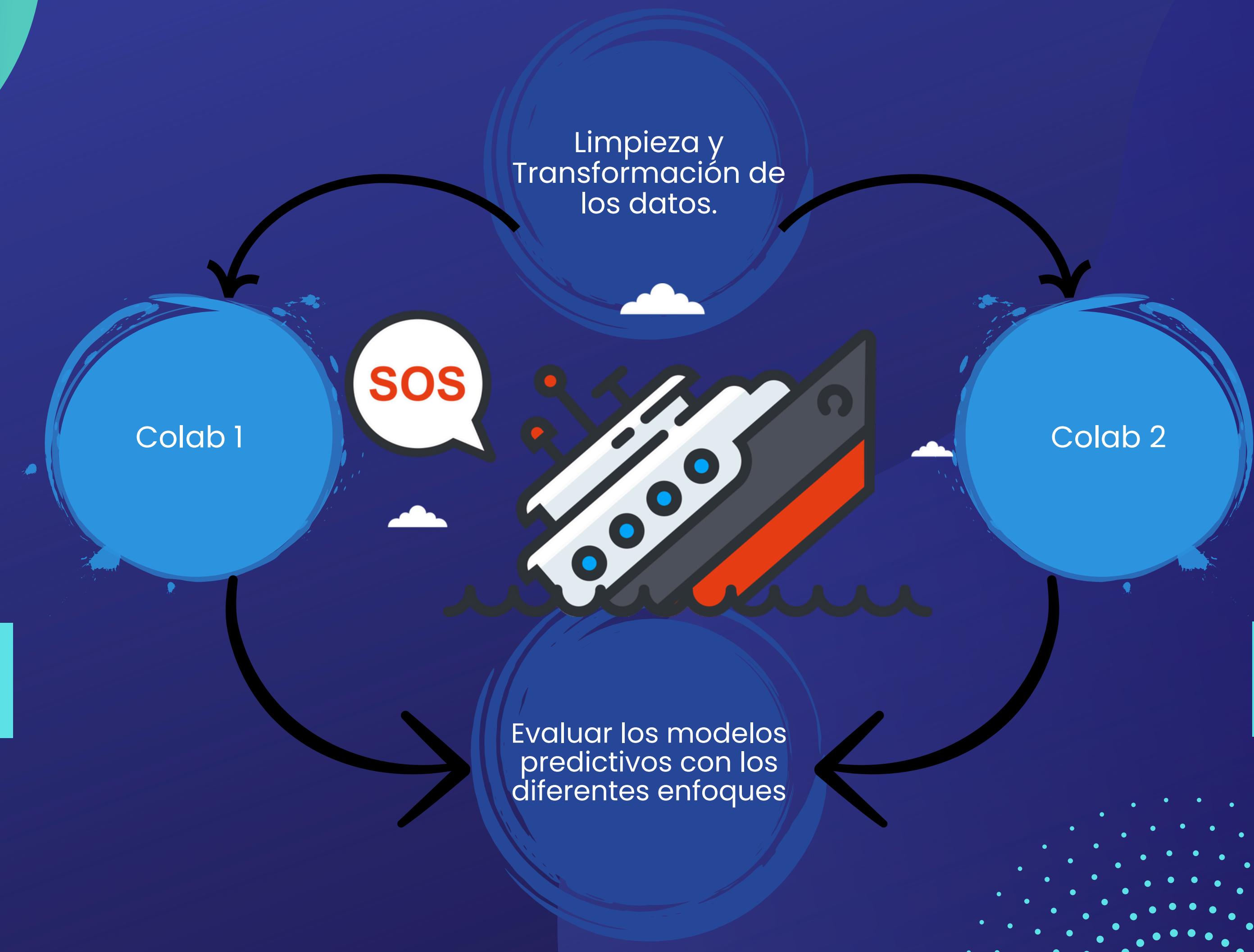
- 418 pasajeros.
- Sin etiqueta de supervivencia.

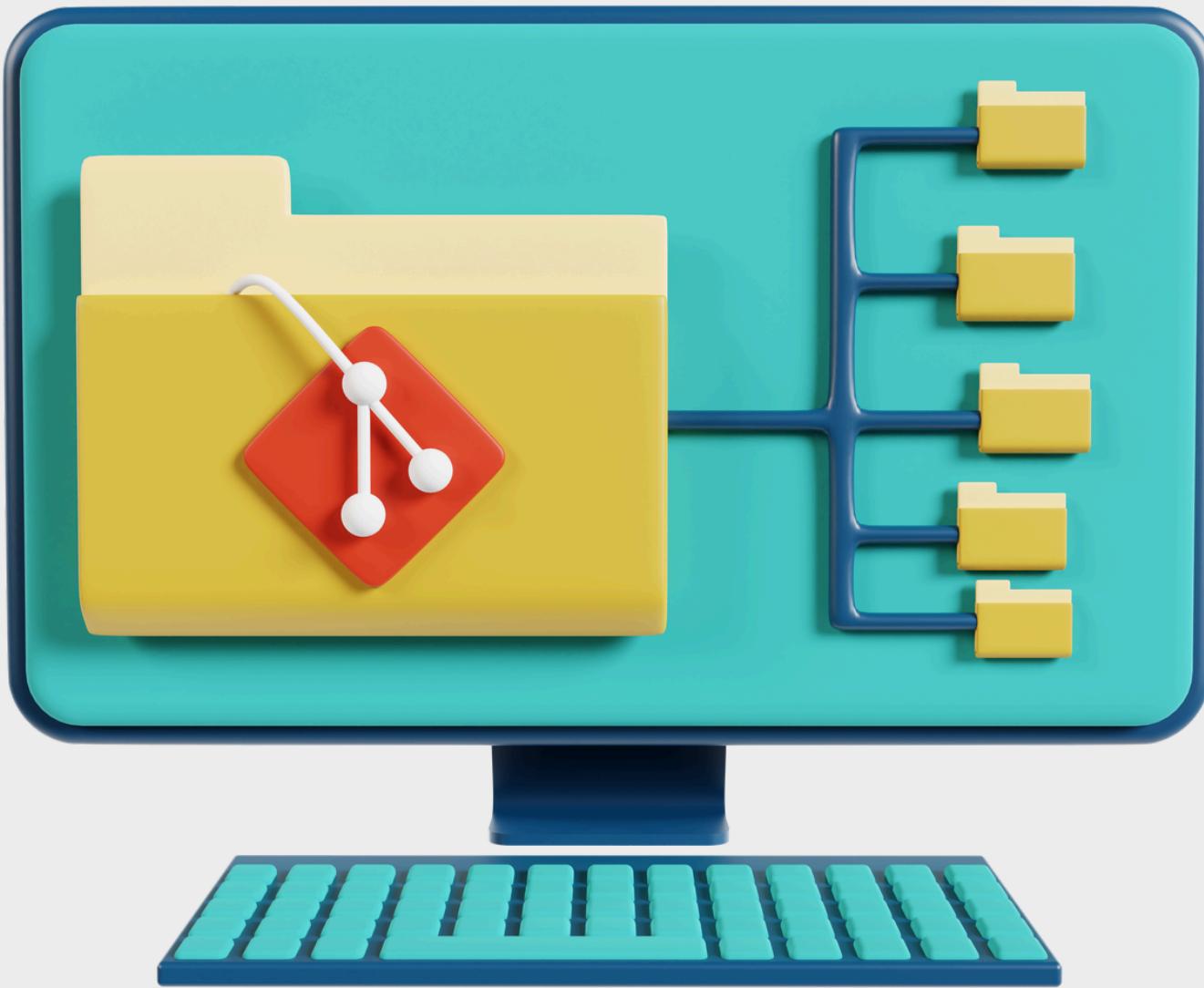
CALIDAD Y ESTRUCTURA

- **Calidad:** Datos limpiados y procesados.
- **Cantidad:** 891 en entrenamiento, 418 en prueba.
- **Formato:** Datos tabulares con diferentes variables: ID, Sobrevivio, Clase Ticket, Nombre, Sexo, Edad, Hermanos_Esposos, Padres_Hijos, Boleto, Tarifa, Cabina y Embarcación.



ETAPAS





EXTRACCIÓN DE DATOS

- Se cargaron los archivos train.csv y test.csv con sus columnas ya traducidas al español desde un repositorio público de GitHub.
- Se realizó una inspección inicial.

EXPLORACIÓN DE DATOS

Columna	Train Nulos	Test Nulos
ID	0	0
Sobrevivio	0	N/A
Clase Ticket	0	0
Nombre	0	0
Sexo	0	0
Edad	177	86
Hermanos_Esp osos	0	0
Padres_Hijos	0	0
Boleto	0	0
Tarifa	0	1
Cabina	687	327
Embarcacion	2	0

Las variables que presentan valores nulos en nuestro dataset, como **"Edad"**, **"Cabina"**, **"Embarcación"** y **"Tarifa"**, serán clave para definir la estrategia adecuada en el preprocesamiento de los datos.

LIMPIEZA

Aspecto	Primer Google Colab	Segundo Google Colab
Variables	ID, Clase Ticket, Nombre, Sexo, Edad, Hermanos_Esposos, Padres_Hijos, Tarifa, Sobrevidió	ID, Sobrevidió, Clase Ticket, Nombre, Sexo, Edad, Hermanos_Esposos, Padres_Hijos, Tarifa, Embarcación
Cabin	Eliminada por la alta cantidad de valores faltantes y baja relevancia en la predicción de supervivencia.	Eliminada por las mismas razones.
Ticket	Eliminada porque no se identificó un patrón útil en los números de boleto y por su naturaleza categórica que requiere una ingeniería de características compleja.	Eliminada por las mismas razones.
Embarcación	Eliminada porque no se consideró relevante para la predicción de la supervivencia.	Se prueba una imputación de datos para esta variable con el fin de observar si aporta información adicional significativa para los modelos.

TRANSFORMACIÓN

Aspecto	Primer Google Colab	Segundo Google Colab
Codificación de Variables Categóricas	One Hot Encoding en columna Sexo (male: 0, female: 1)	One Hot Encoding en Clase Ticket, Embarcación, Familiares, Título, y Sexo
Creación de Nuevas Variables	Clasificación de Edad (Bebé, Niño, Adolescente, Adulto, Adulto mayor, Viejo) y Familiares (suma de Hermanos/Espouses y Padres/Hijos)	Clasificación de Edad (Bebé, Niño, Adolescente, Adulto, Adulto mayor, Viejo), Familiares (suma), Solo Viaje (binaria)
Escalamiento de los Datos	Se utilizó MinMaxScaler en Edad, Familiares y Clase Ticket.	MinMaxScaler en Edad, Hermanos_Espouses, Padres_Hijos y Tarifa
Análisis de Componentes Principales (PCA)	-	PCA para reducir a 6 componentes principales, explicando el 90 % de la variabilidad

IMPUTACIÓN DE DATOS

Aspecto	Primer Google Colab	Segundo Google Colab
Imputación de Edad	Títulos extraídos (Mr, Mrs, Miss, Master). Imputación probabilística basada en media y desviación estándar. Manejo especial para título Ms.	Títulos extraídos y normalizados (Mlle, Mme, Ms). Función assign_age con imputación probabilística. Manejo especial para títulos raros.
Extracción de Títulos	Extraídos directamente de la columna Nombre.	Expresión regular utilizada para extraer títulos (([A-Za-z]+)).
Manejo de Casos Especiales	Título Ms imputado con promedio ponderado y desviación estándar ponderada de Mrs y Miss	Títulos raros (e.g., Ms) combinados con medias de títulos similares para imputar edad
Imputación de Tarifa	Imputación basada en la clase socioeconómica. Media y desviación estándar por clase.	Similar al primer colab, pero con generador de números aleatorios (RandomState(22))
Imputación de Tarifa	-	Imputación de valores nulos en la columna Embarcación utilizando la moda.

CONSIDERACIONES ESPECIALES

ETAPA PREPARACION DATOS

Primer Google Colab

- **Titulos:**
 - 1.Mr: $\mu = 32.37, \sigma = 12.71$
 - 2.Mrs: $\mu = 35.90, \sigma = 11.43$
 - 3.Master: $\mu = 4.57, \sigma = 3.62$
 - 4.Miss: $\mu = 21.77, \sigma = 12.91$
 - 5.Dr: $\mu = 42.00, \sigma = 12.02$
- **Clasificacion Edad:**
 - 1.Bebé: 0 y 5 años.
 - 2.Niño: 5 y 14 años.
 - 3.Adolescente: 14 y 18 años.
 - 4.Adulto: 18 y 30 años.
 - 5.Adulto mayor: 30 y 60 años.
 - 6.Viejo: 60 y 100 años.

Segundo Google Colab

- **Titulos:**
 - 1.Mr: $\mu = 32.37, \sigma = 12.71$
 - 2.Mrs: $\mu = 35.79, \sigma = 11.44$
 - 3.Master: $\mu = 4.57, \sigma = 3.62$
 - 4.Miss: $\mu = 21.85, \sigma = 12.87$
 - 5.Otro: $\mu = 45.55, \sigma = 11.78$
- **Clasificacion Edad:**
 - 1.Bebé: 0 y 5 años.
 - 2.Niño: 5 y 14 años.
 - 3.Adolescente: 14 y 18 años.
 - 4.Adulto: 18 y 30 años.
 - 5.Adulto mayor: 30 y 60 años.
 - 6.Viejo: 60 y 100 años.
- **Columnas Utilizadas en PCA:**
 - Edad Escalada
 - Clase Ticket (3 col)
 - Embarcacion(3)
 - Familiares (3)
 - Titulo(5)
 - Sexo(2)

DATASET POSTERIOR A SU PROCESAMIENTO

COLAB 1

ID	Sobrevivió	Clase Ticket	Nombre	Sexo	Edad	Hermanos_Esposos	Padres_Hijos	Tarifa	Título	Clasificación_Edad	Familiares
0	1	3	Braund, Mr. Owen Harris	0	22.0	1	0	7.2500	Mr	Adulto	1
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	1	38.0	1	0	71.2833	Mrs	Adulto mayor	1
2	3	1	Heikkinen, Miss. Laina	1	26.0	0	0	7.9250	Miss	Adulto	0
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	53.1000	Mrs	Adulto mayor	1
4	5	0	Allen, Mr. William Henry	0	35.0	0	0	8.0500	Mr	Adulto mayor	0

DATASET CON POSTERIOR AL PROCESAMIENTO

COLAB 2

**One Hot Encoding
(30 columnas)**

Clase Ticket

- Primera Clase
- Segunda Clase
- Tercera Clase

Embarcación

- S
- Q
- C

Familiares

- FamPequeña
- FamMediana
- FamGrande

Título

- Mr: Señor
- Miss: Señorita
- Mrs: Señora
- Master: Joven (para varones menores de edad)
- Otro: Otros títulos no especificados

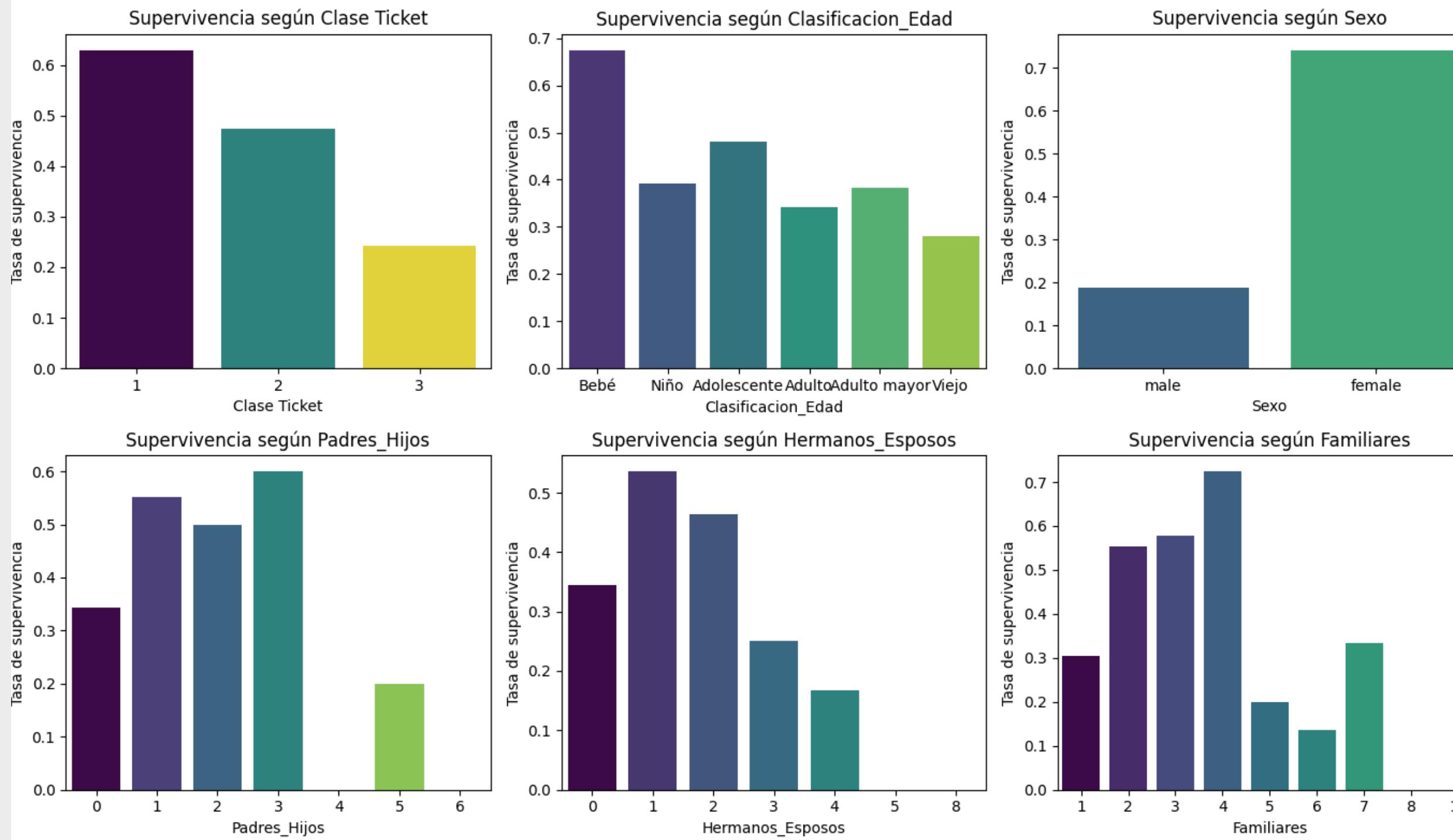
Sexo

- Femenino
- Masculino

ID	Sobrevivió	Clase Ticket	Nombre	Sexo	Edad	Hermanos_Espouses	Padres_Hijos	Tarifa	Embarcación	FamPequeña
0	1	3	Braund, Mr. Owen Harris	male	22.0	1	0	7.2500	s	1	
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	71.2833	c	1	
2	3	1	Heikkinen, Miss. Laina	female	26.0	0	0	7.9250	s	1	
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	53.1000	s	1	
4	5	0	Allen, Mr. William Henry	male	35.0	0	0	8.0500	s	1	

ANALISIS ANTERIOR AL PREPROCESAMIENTO DE MODELOS

Primer Google Colab



Segundo Google Colab

GENERACIÓN DE MODELOS

Aspecto	Primer Google Colab	Segundo Google Colab
Variables Predictoras	Clase Ticket, Sexo, Edad y Familiares	PC1, PC2, PC3, PC4, PC5, PC6
Modelos utilizados	Regresión Logística, Árboles de decisión, Bosques Aleatorios, K Vecinos Más Cercanos, Naive Bayes, Regresión Logística sin framework, FNN, LSTM, XGBOOST	Regresión Logística, Árboles de decisión, Bosques Aleatorios, K Vecinos Más Cercanos, Naive Bayes.
Búsqueda de hiperparámetros	GridSearchCV y Optuna	GridSearchCV y Optuna

EVALUACIÓN DE MODELOS

Aspecto	Primer Google Colab	Segundo Google Colab
Mejor Modelo Convencional usando GridSearchCV *Hiperparámetros en el reporte	Modelo: Árboles de Decisión Accuracy en Validación: 83.05% Accuracy en Prueba: 76.56%	Modelo: K Vecinos Más Cercanos Accuracy en Validación: 82.38% Accuracy en Prueba: 73.68%
Mejor Modelo Convencional usando Optuna *Hiperparámetros en el reporte	Modelo: Árboles de Decisión Accuracy en Prueba: 78.94%	Modelo: K Vecinos Más Cercanos Accuracy en Prueba: 77.51%
FNN	Accuracy en Prueba: 77.99%	-
LSTM	Accuracy en Prueba: 76.55%	-
XGBOOST	Accuracy en Prueba: 75.87%	-
Mejor Modelo Complejo con Optuna	Modelo: FNN Accuracy en Prueba: 77.99%	

RESULTADOS

Se encontró que el mejor modelo realizado fue el de Árboles de decisión con un accuracy en prueba de 78.94 %.



Pofundidad
máxima de 6



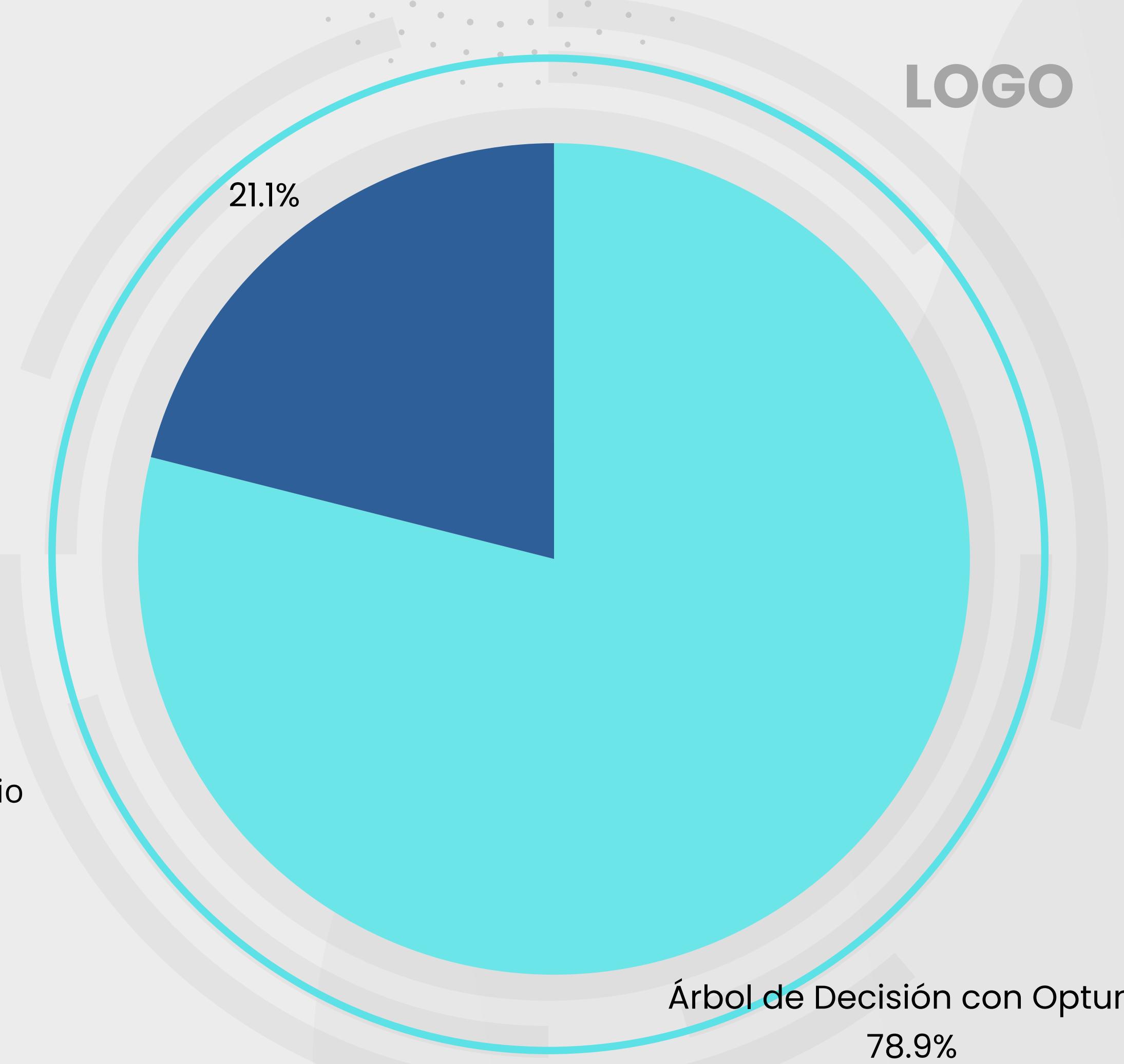
Mínimo de
muestras de
nodos externos
de 3 y un
máximo de
muestras de
nodos externos
de 27.

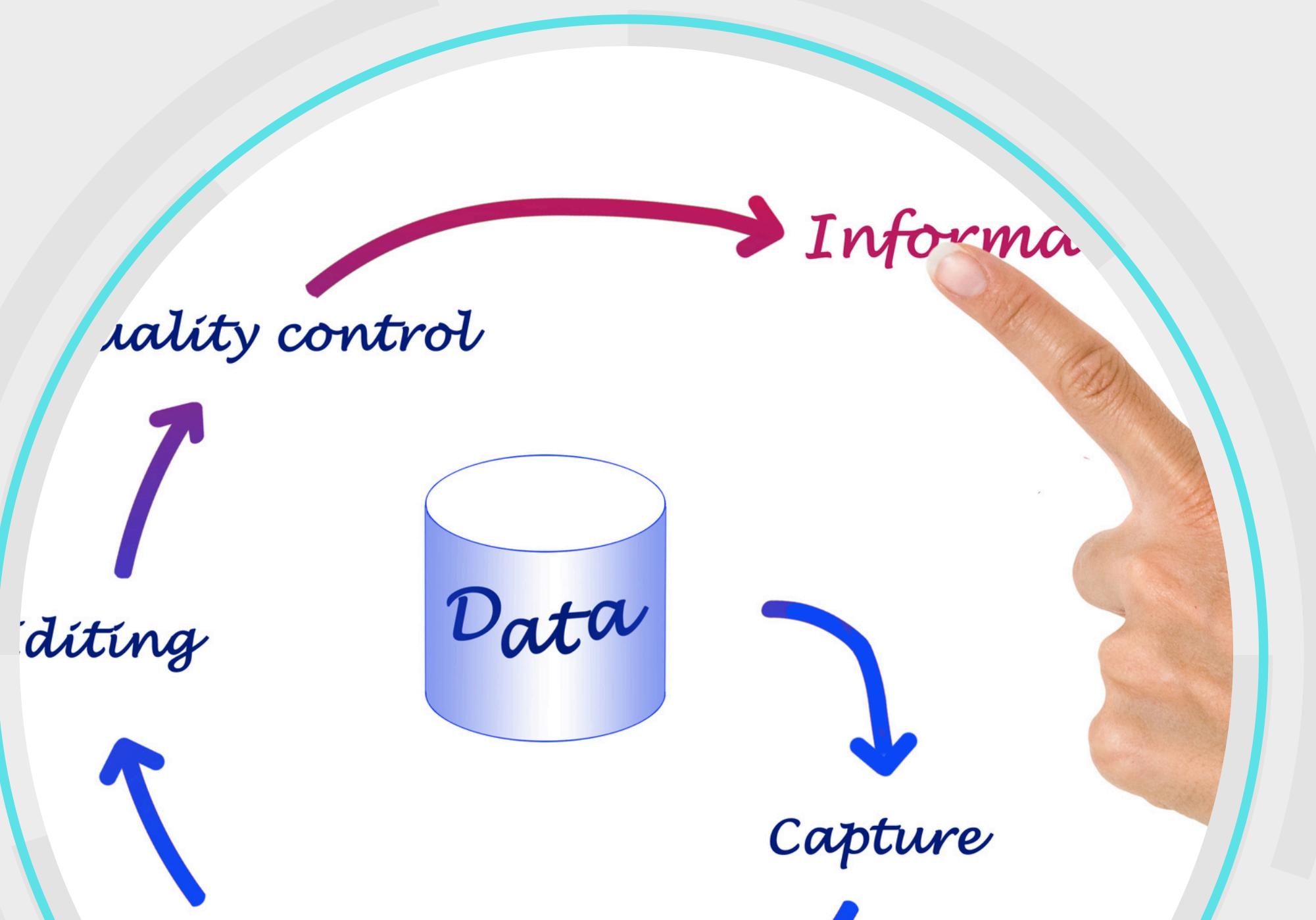


Split mínimo de
muestras de 17

CONCLUSIONES CLAVE

- **Desarrollo de modelos:** El Árbol de Decisión optimizado con Optuna alcanzó un 78.94% de precisión en el conjunto de prueba.
- **Factores clave:** Clase socioeconómica, Género, Edad.
- **Optimización de Modelos:** GridSearchCV y Optuna.
- **Desempeño:** Modelo interpretable, ideal para problemas de clasificación binaria. Mejor equilibrio entre simplicidad y precisión.





TRABAJO A FUTURO

- **Refinamiento de la Ingeniería de Características:** Explorar nuevas transformaciones y combinaciones para capturar relaciones más complejas.
- **Modelos Avanzados:** Implementar redes neuronales profundas para identificar patrones no lineales.
- **Generación de Datos Sintéticos:** Ampliar el dataset con técnicas de generación de datos.
- **Imputación Avanzada de Datos Faltantes:** Mejorar la precisión en la estimación de valores nulos.
- **Validación Cruzada Exhaustiva:** Aplicar validación cruzada más profunda



¡GRACIAS POR
SU ATENCIÓN!