ÁRBOLES DE DECISIÓN, REGLAS Y ENSEMBLE LEARNING

1. Descripción del Dataset

El dataset utilizado, **"Blood Transfusion Service Center"**, fue obtenido de OpenML (ID: 1464). Contiene información sobre donantes de sangre con el objetivo de predecir si una persona volverá a donar sangre.

1.1. Características del dataset:

- **Número de instancias:** 748.
- **Atributos:**
    - ULTIMA DONACION: Meses desde la última donación.
    - TOTAL DE DONACIONES: Número total de donaciones realizadas.
    - TOTAL SANGRE DONADA: Total de sangre donada en centímetros cúbicos.
    - MESES PRIMERA DONACION: Meses desde la primera donación.
    - DONADOR: Variable binaria objetivo (1 = No donará, 2 = Donará).
- **Distribución por clase:**
    - Clase 1 (No donará): *570*
    - Clase 2 (Donará): *178*

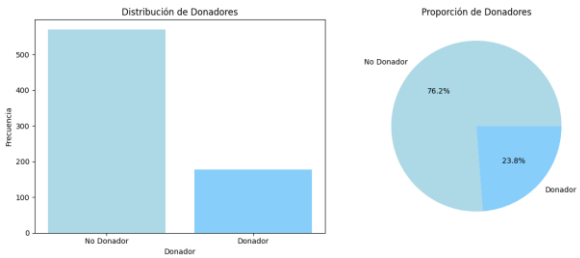


Ilustración 1. *Proporción y distribución de la variable objetivo*

2. Análisis Exploratorio

2.1. Visualizaciones:

Histogramas y distribución para cada atributo.

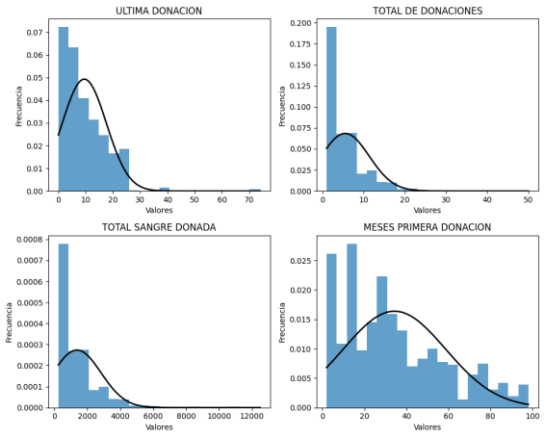


Ilustración 2. *Histogramas de frecuencias y distribuciones normales para cada atributo.*

Gráficos boxplot para detección de valores atípicos.

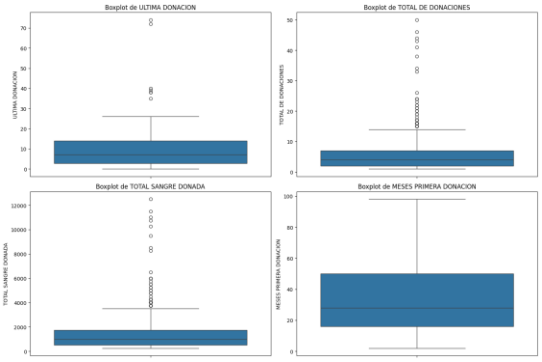


Ilustración 3. *Gráficos de boxplot para detección de valores atípicos y análisis de distribución.*

Matriz de correlación para observar relaciones.

Autor: Juan José Acevedo Dávila

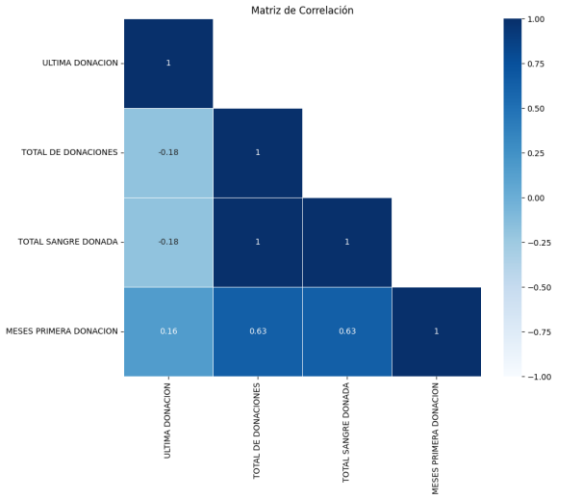


Ilustración 4. *Matriz de correlación para observar relaciones entre las variables.*

2.1.1. Hallazgos principales:

- En el análisis de la variable objetivo, se encontró un desbalance de clases, la cual se puede observar en la Ilustración 1.
- La mayoría de los donantes tienen un historial reciente y limitado, con pocas donaciones y cantidades moderadas de sangre, concentrándose en rangos bajos.
- Algunas variables presentan valores extremos que no se consideran atípicos dado el contexto del problema.
- Las variables están moderadamente relacionadas entre sí, excepto "Última Donación", que parece comportarse de manera más independiente. Esto puede indicar que los patrones de donación reciente no necesariamente reflejan el historial general de los donantes.

3. Modelos Implementados

3.1. Random Forest:

- Se usó GridSearchCV para optimizar los hiperparámetro: El número de árboles que se construirán (n-estimators), profundidad de cada árbol (max_depth) y el número mínimo de muestras necesarias (min_samples_split)
- Métricas de evaluación: Precisión, Recall, F1-Score.

3.2. RIPPER:

- Se entrenó una versión sin manejo de desbalance clases y otra con SMOTE (que se encarga del desbalance).
- Las reglas generadas se interpretaron para entender la lógica del modelo.

4. Resultados

4.1. Random Forest

4.1.1. Mejores parámetros encontrados:

- Número de árboles: 100
- Profundidad máxima: 10.
- Mínimo de muestras necesarias: 20

4.1.2. Matriz de confusión:

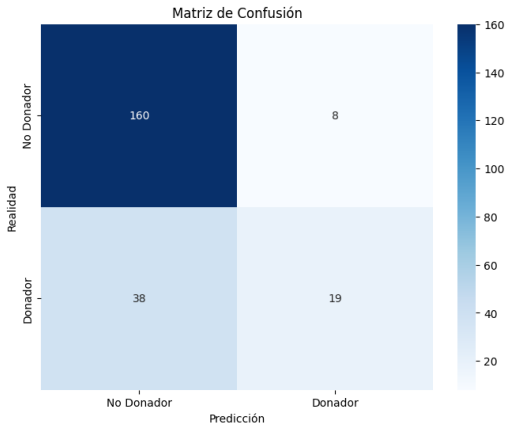


Ilustración 5. *Matriz de confusión de bosques aleatorios.*

4.1.3. Tasa de verdaderos y falsos positivos (TP Rate y FP Rate):

INSTANCIA	VALOR
TP RATE	33.33%
FP RATE	4.76%

Tabla 1. *Tasa TP y tasa FP del modelo Random Forest*

4.2. RIPPER

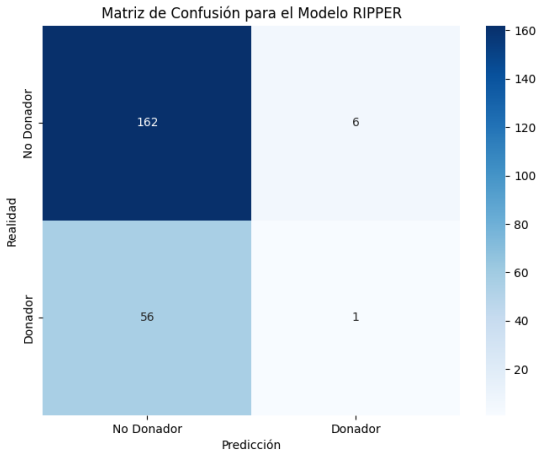
4.2.1 Sin SMOTE

Reglas generadas

Si se cumplen las siguientes reglas, se clasifica como donador (Sí Donará):

- Regla 1: [ULTIMADONACION: < 2 ^ TOTALDEDONACIONES: > 12 ^ MESESPRIMERADONACION: 28-35].
    - ULTIMADONACION: La última donación fue hace menos de 2 meses.
    - TOTALDEDONACIONES: El total de donaciones es mayor o igual a 12 centímetros cúbicos.
    - MESESPRIMERADONACION: Los meses que pasaron desde la primera donación se sitúa entre los 28 y 35 meses.
- Regla 2: [ULTIMADONACION: 2-4 ^ TOTALDEDONACIONES: 3-5 ^ MESESPRIMERADONACION: 28-35].
    - ULTIMADONACION: La última donación se realizó entre 2 y 4 meses.
    - TOTALDEDONACIONES: El total de donaciones se encuentra entre 3 y 5 centímetros cúbicos.
    - MESESPRIMERADONACION: Los meses que pasaron desde la primera donación se sitúa entre los 28 y 35 meses.
- Regla 3: [ULTIMADONACION: < 2 ^ TOTALDEDONACIONES: > 12].
    - ULTIMADONACION: La última donación fue hace menos de 2 meses.
    - TOTALDEDONACIONES: El total de donaciones es mayor o igual a 12 centímetros cúbicos.

4.2.2. Matriz de confusión



4.2.3. Tasa de verdaderos y falsos positivos (TP Rate y FP Rate)

INSTANCIA	VALOR
TP RATE	1.75%
FP RATE	3.57%

4.3. Con SMOTE:

4.3.1. Reglas generadas:

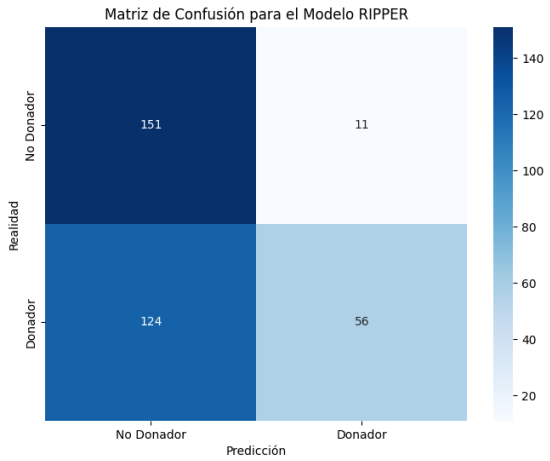
Si se cumplen las siguientes reglas, se clasifica como donador (Sí Donará):

- Regla 1: [ULTIMADONACION: < 2 ^ TOTALDEDONACIONES: 5-7].
    - ULTIMADONACION: La última donación fue hace menos de 2 meses.

- TOTALDEDONACIONES: El total de donaciones se encuentra entre 5 y 7 centímetros cúbicos.

- Regla 2: [ULTIMADONACION: < 2 -3].
  - ULTIMADONACION: La última donación se realizó entre 2 y 3 meses.
- Regla 3: [ULTIMADONACION: < 2 ^ TOTALSANGREDONADA: > 3250].
  - ULTIMADONACION: La última donación fue hace menos de 2 meses.
  - TOTALSANGREDONADA: El total de sangre donada es mayor o igual a 3250 centímetros cúbicos.
- Regla 4: [ULTIMADONACION: 4-8 ^ MESESPRIMERADONACION: 4-13]
  - ULTIMADONACION: La última donación se realizó entre 4 y 8 meses.
  - MESESPRIMERADONACION: La primera donación la realizó entre 4 y 13 meses.

4.3.2. Matriz de confusión

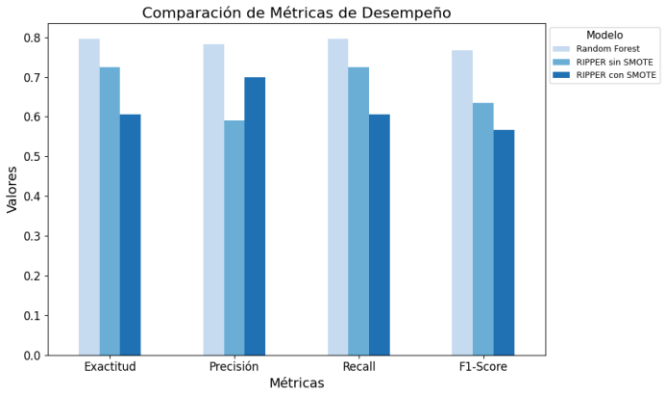


4.3.3. Tasa de verdaderos y falsos positivos (TP Rate y FP Rate)

INSTANCIA	VALOR
TP RATE	31.1%
FP RATE	6.79%

5. Comparación de Modelos

5.1. Gráfico de comparación de métricas:



Modelo	Exactitud	Precisión	Recall	F1-Score
Random Forest	0.795556	0.781639	0.795556	0.767426
RIPPER sin SMOTE	0.724444	0.591053	0.724444	0.634652
RIPPER con SMOTE	0.605263	0.700001	0.605263	0.566005

6. Discusión y Conclusión

Los resultados indican que el modelo Random Forest es el más eficaz para el conjunto de datos analizado, ya que logra el mejor desempeño en todas las métricas clave (Exactitud, Precisión, Recall y F1-Score). Esto lo convierte en la opción más robusta para predecir correctamente tanto los casos positivos como negativos.

Por otro lado, el modelo RIPPER, aunque menos efectivo en general, ofrece interpretabilidad a través de reglas claras. Sin embargo, su desempeño es inferior, especialmente en términos de equilibrio entre precisión y recall. Además, el uso de SMOTE para balancear las clases con RIPPER no resultó en mejoras significativas; de hecho, perjudicó el recall y el F1-Score, lo que podría

indicar que este método no es el más adecuado para optimizar el rendimiento de RIPPER en este caso.

Finalmente, se recomienda desarrollar nuevos modelos de Random Forest aplicando técnicas para tratar el desbalance de clases, ya que esto podría mejorar significativamente su rendimiento. Además, para ambos modelos, tanto Random Forest como RIPPER, se sugiere explorar la eliminación de datos atípicos, lo que podría contribuir a mejorar su desempeño, independientemente de si se utiliza un tratamiento para el desbalance de clases.