
PREDICCIÓN DE EFICACIA EN DONANTES EN ASISTOLIA NO CONTROLADA

07/2025

Contents

1	Dataset inicial	2
2	Análisis variables pre ECMO	4
2.1	Descriptivo y análisis univariado	4
2.2	Análisis multivariado	5
2.2.1	Modelo de regresión logística	6
2.2.2	KNN	8
2.2.3	SVM	9
2.2.4	Random forest	11
2.2.5	Árbol de decision	12
2.2.6	Gradient Boosting	15
2.2.7	XGBoost	16
2.2.8	Comparativa final	17
3	Análisis global	18
3.1	Descriptivo y análisis univariado	18
4	Análisis multivariado	20
4.1	Correlación	20
4.2	PCA	21
4.3	Modelos de predicción	22

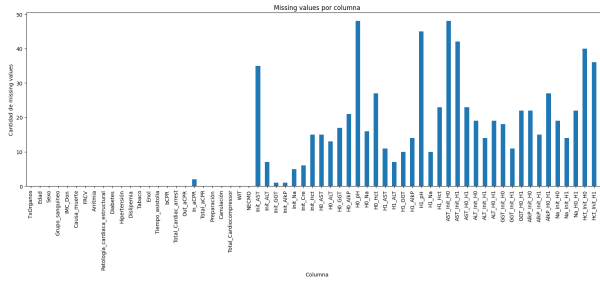


Figure 1: Missing values por columna

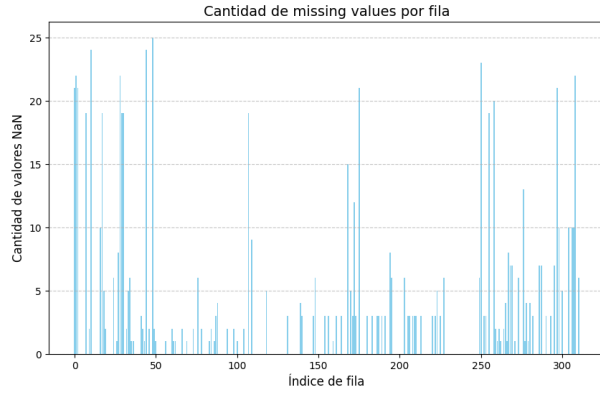


Figure 2: Missing values por filas

1 Dataset inicial

El dataset inicial, que incluye los donantes tras asistolia no controlada de entre 2002 y 2024, consta de 311 individuos, 20 de los cuales fueron desestimados por diversos motivos (Tabla 1), por lo que son eliminados. Se eliminan así mismo donantes para los cuales no constan los tiempos de parada ni los factores de riesgo cardiovascular. Así pues el dataset final cuenta con 279 donantes, de los cuales 205 son eficaces y 74 son no eficaces.

Este es el dataset usado en el **análisis de variables pre ECMO**, que consta de variables antropométricas, antecedentes y tiempos de parada.

En cambio, para el **análisis general**, se hacen algunas otras modificaciones.

Se eliminan las variables analíticas de hora 3 y 4, puesto que la gran mayoría de donantes no tienen medidos esos parámetros por ser la duración de PRN inferior. Se elimina también la creatinina, bilirrubina, potasio y LDH de las analíticas durante PRN por no estar presente en muchas analíticas (hemolización o no analizadas con el Piccolo). Se crean nuevas variables para estudiar la evolución de los parámetros analíticos (ASL, ALT, GGT, FA, Na, K, hematocrito). Donde se codifica con un 1 si se duplica el valor de referencia, 2 si se reduce a la mitad y 0 en cualquier otro caso.

Dada la abundancia de missing values (Figura 1 y Figura 2), se decide eliminar las filas con más de 10 NAs. Esto resulta en un dataset final de **267 filas y 64 columnas**, con 200 donantes eficaces y 67 no eficaces. De esta manera garantizamos que el análisis sea más robusto.

Motivo	N (%)
Neoplasia	7 (35)
Perforación-peritonitis	4 (20)
Sepsis	3 (15)
Chagas +	2 (10)
Contaminación	1 (5)
HIV +	1 (5)

Table 1: Descriptivo de motivos de desestimación de donantes (N = 20)

2 Análisis variables pre ECMO

2.1 Descriptivo y análisis univariado

Variable	General	Efectivo (N = 205)	No efectivo (N = 74)	p-valor
Edad	51 [44–58]	49 [41–57]	54.5 [48–60]	0.002
IMC	26.30 [24.22–29.40]	26.23 [24.22–29.32]	27.39 [24.76–30.84]	0.05
Tiempo sin maniobras	5 [1–10]	6 [1–10]	4.5 [0–10]	0.15
SVB	2 [0–8]	2 [0–8]	2 [0–7.75]	0.84
PCR-inicio SVA	10 [7–15]	10 [7–15]	10 [6–15]	0.52
SVA extrahospitalario	53 [43–67.5]	53 [43–67]	53 [39.25–67.5]	0.62
SVA intrahospitalario	5 [2–11]	5 [2–10]	5 [2–15.75]	0.63
Total SVA	60 [50–77]	60 [50–75]	61 [47.25–80]	0.9
Sexo (H), N (%)	255 (91.4)	187 (91.2)	68 (91.9)	1
Causa de la muerte (IAM), N (%)	168 (60.2)	125 (61.0)	43 (58.1)	0.28
FRCV, N (%)	207 (74.2)	149 (72.7)	58 (78.4)	0.42
Arritmia, N (%)	12 (4.3)	6 (2.9)	6 (8.1)	0.12
Patología cardíaca estructural, N (%)	50 (17.9)	28 (13.7)	22 (29.7)	0.003
HTA, N (%)	83 (29.7)	59 (28.8)	24 (32.4)	0.66
Diabetes, N (%)	21 (7.5)	9 (4.4)	12 (16.2)	0.002
DLP, N (%)	87 (31.2)	55 (26.8)	32 (43.2)	0.01
Tabaco, N (%)	134 (48.0)	102 (49.8)	32 (43.2)	0.41
Enol, N (%)	44 (15.8)	30 (14.6)	14 (18.9)	0.5

Table 2: Distribución de variables por tipo de donante

2.2 Análisis multivariado

Variable	Coefficiente	p-valor
Intercepto	1.3931	0.0028
Edad	-0.7279	0.0022
Diabetes	-1.0675	0.0090
Patología cardiaca estructural 1	-0.8722	0.0302
Causa muerte (4)	-0.7866	0.0446
Dislipemia	-0.6234	0.0621
Hipertensión	0.4932	0.1187
Enol	-0.5250	0.1187
FRCV	0.4179	0.1752
Arritmia	-0.5599	0.2225
Causa muerte (1)	-0.3429	0.3394
In aCPR	-0.2400	0.3437
Total aCPR	0.0399	0.3916
Out aCPR	-0.0029	0.3980
Causa muerte 8	-0.4098	0.4241
IMC Don	-0.1066	0.4387
Causa muerte 2	0.2775	0.6998
Sexo (H)	0.0336	0.8559
Tabaco	0.1697	0.8460
Causa muerte (3)	0.1196	1.0000
Tiempo asistolia	0.0193	1.0000
bCPR	-0.2582	1.0000
Total Cardiac arrest	-0.2064	1.0000

Table 3: Regresión logística: coeficientes y valores p ordenados

2.2.1 Modelo de regresión logística

Métrica	Valor
Exactitud (Accuracy)	0.738
Sensibilidad (Recall)	0.919
Especificidad	0.227
Precisión (Precision)	0.770
F1 Score	0.837

Table 4: Métricas de desempeño del modelo a partir de la matriz de confusión

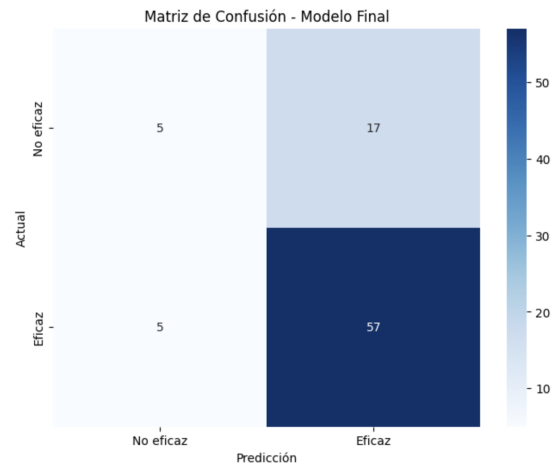


Figure 3: Matriz de confusión - modelo de regresión logística

Variable	Coefficiente	Valor p	Odds ratio (OR)
Constante	5.0331	0.0000	153.4028
Edad	-0.0643	0.0022	0.9377
bCPR	-0.0635	0.0183	0.9385
Diabetes	-1.2865	0.0197	0.2762
Patología cardiaca estructural 1	-0.8974	0.0312	0.4076
Causa muerte 4	-0.9378	0.0896	0.3915

Table 5: Coeficientes del modelo, valores p y odds ratio

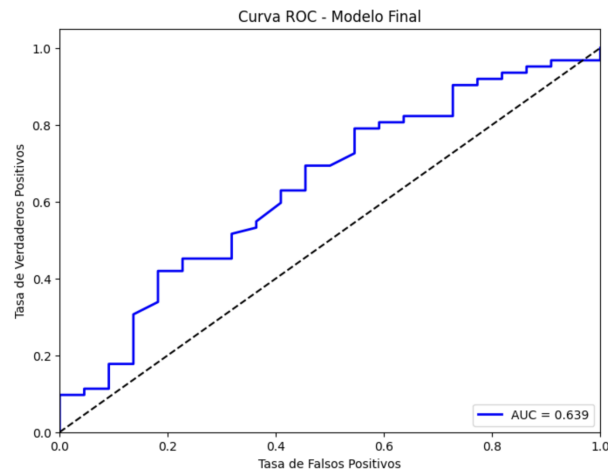


Figure 4: Curva ROC del modelo de regresión logística

Interpretación del modelo de regresión logística

La siguiente tabla resume los coeficientes, valores p y razones de momios (Odds Ratios, OR) del modelo de regresión logística. Se interpretan los resultados más relevantes a continuación:

- **Edad:** El coeficiente negativo (-0.0643 , $p = 0.0022$) indica que por cada año adicional de edad, las probabilidades de que el donante sea eficaz disminuyen en aproximadamente un 6.2% ($OR = 0.94$). Este efecto es estadísticamente significativo.
- **bCPR:** Cada minuto adicional de RCP básica está asociado con una disminución del 6.2% en la probabilidad de eficacia del donante ($OR = 0.94$, $p = 0.0183$).
- **Diabetes:** La presencia de diabetes reduce significativamente la probabilidad de eficacia en un 72.4% ($OR = 0.28$, $p = 0.0197$). Este resultado es estadísticamente significativo.
- **Patología cardíaca estructural:** Tener esta condición se asocia con una disminución del 59.2% en la probabilidad de eficacia ($OR = 0.41$, $p = 0.0312$).
- **Causa de muerte tipo 4:** Esta causa muestra una asociación con menor eficacia ($OR = 0.39$), lo que sugiere una reducción del 60.8% en la probabilidad, aunque no alcanza significancia estadística ($p = 0.0896$).

En conjunto, el modelo identifica varias variables clínicamente relevantes que afectan la probabilidad de que un donante sea eficaz. En particular, edad avanzada, mayor tiempo de bCPR, diabetes y patología cardíaca estructural se asocian con menor probabilidad de éxito.

2.2.2 KNN

Este modelo clasifica un caso nuevo observando los K casos más parecidos del conjunto de datos. Es decir, si un nuevo donante se parece mucho a varios donantes que fueron eficaces, el modelo asumirá que también será eficaz.

- No realiza un “entrenamiento”, sino que compara directamente con los datos previos.
- Es fácil de entender, pero puede verse afectado por variables irrelevantes o escalas distintas.
- Suele tener mejor rendimiento con pocos datos y bien distribuidos.

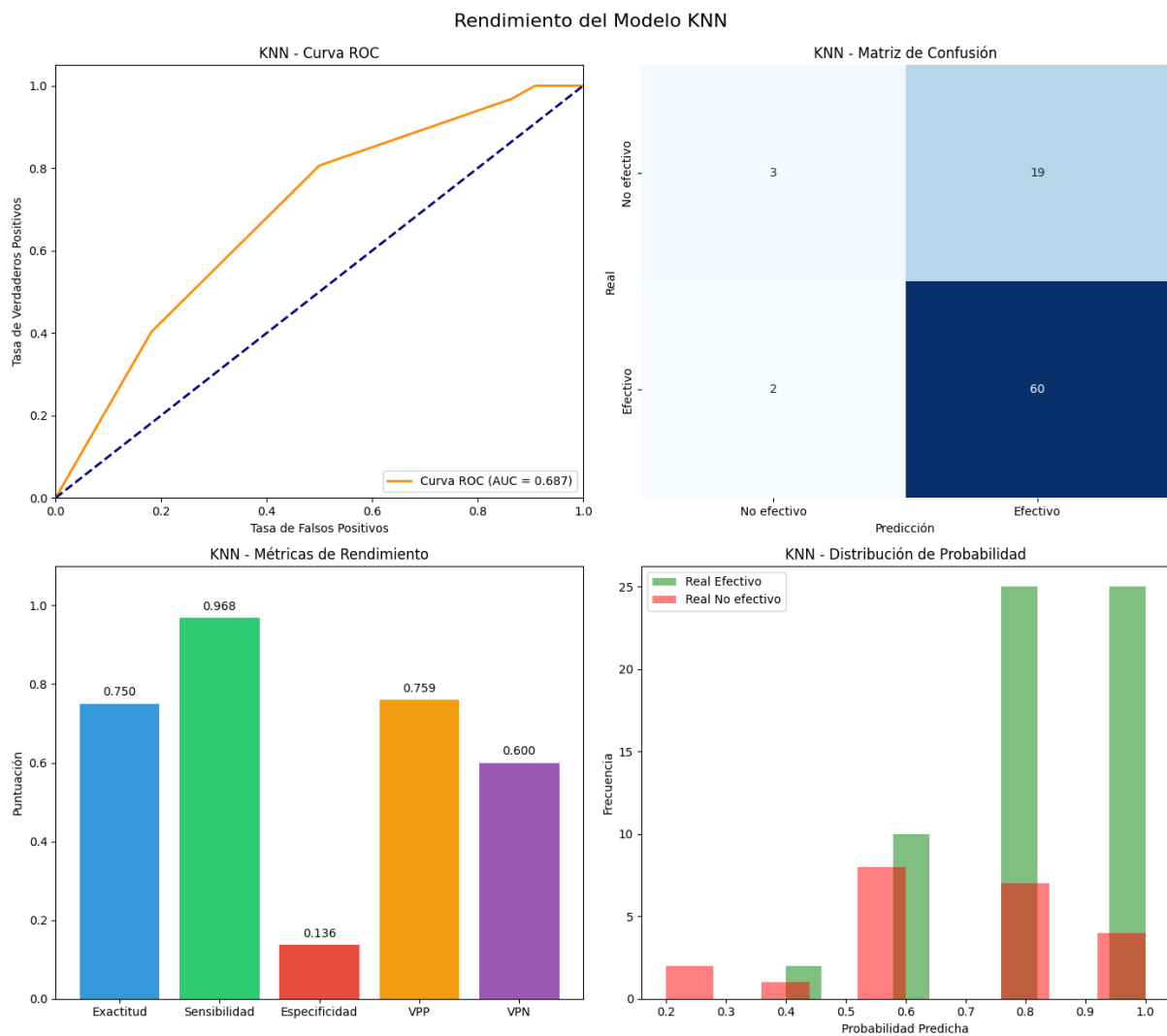


Figure 5: Resumen modelo KNN

2.2.3 SVM

Este modelo busca una **línea (o frontera)** que separe lo mejor posible los casos eficaces de los no eficaces.

- Intenta maximizar la distancia entre los dos grupos.
- Puede adaptarse a relaciones no lineales mediante funciones especiales (kernels).
- Es útil cuando los datos no están claramente separados, pero puede ser sensible a valores atípicos.

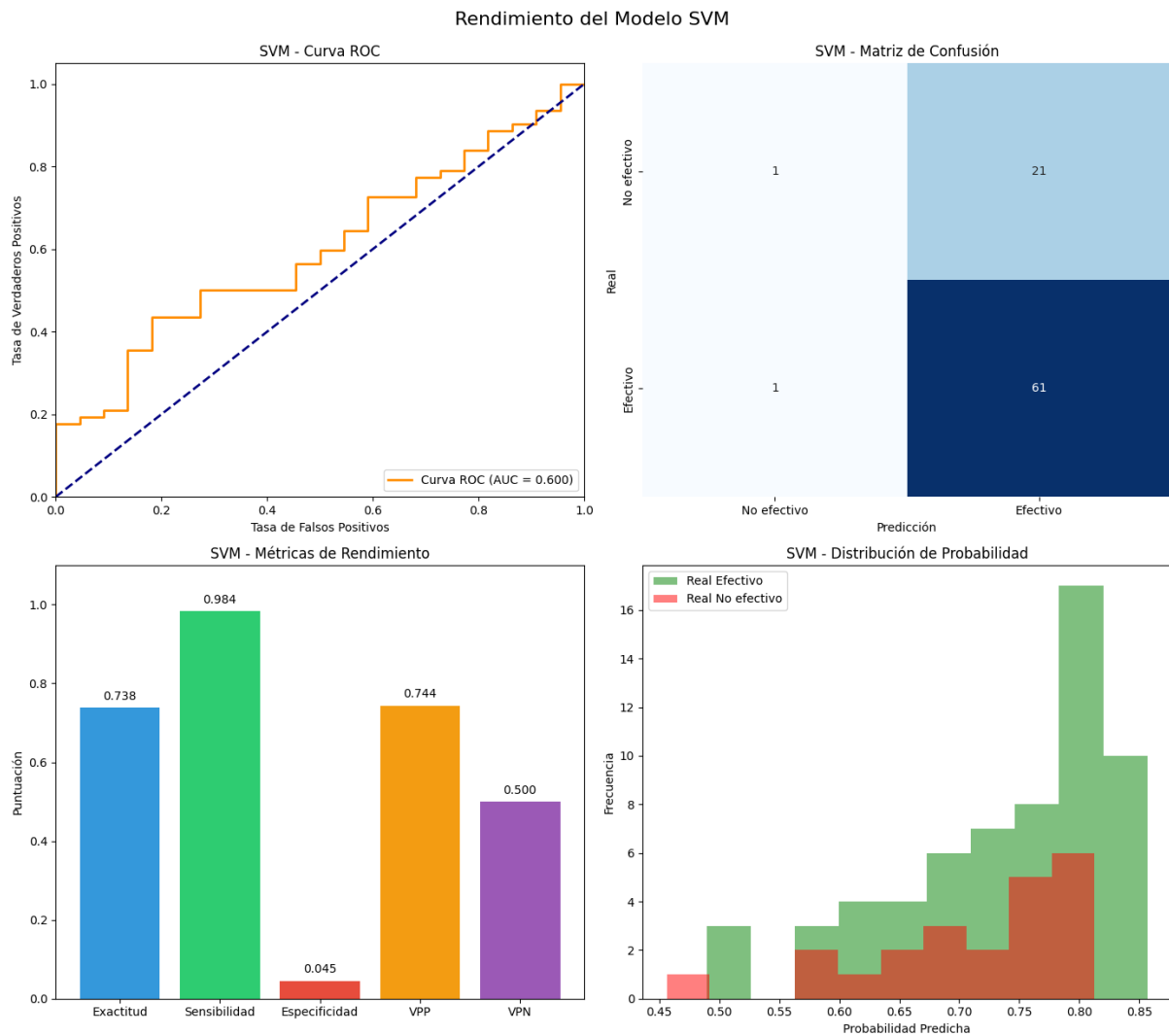


Figure 6: Resumen modelo SVM

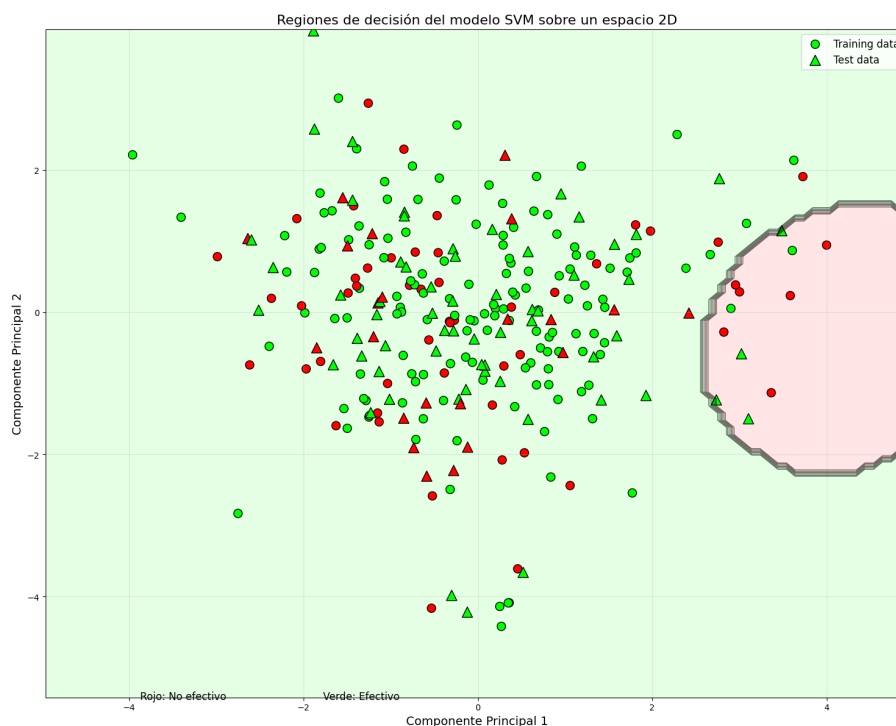


Figure 7: Regiones de decisión del modelo

Como podemos observar, el modelo ha trazado una frontera (no lineal) que intenta separar ambas clases. Sin embargo, los datos están muy mezclados: hay muchos puntos rojos en zona verde y viceversa. Esto significa que el modelo tiene dificultad para encontrar una separación clara entre donantes efectivos y no efectivos en este espacio. Hay alta sensibilidad (detecta muchos efectivos) pero baja especificidad (confunde no efectivos con efectivos).

2.2.4 Random forest

Es un modelo basado en muchos **árboles de decisión**. Cada árbol hace una predicción y el modelo final toma la decisión por mayoría.

- Es robusto y funciona bien con muchos tipos de datos.
- Reduce el riesgo de sobreajuste que puede ocurrir con un solo árbol.
- Permite identificar la **importancia de cada variable** en las decisiones del modelo.

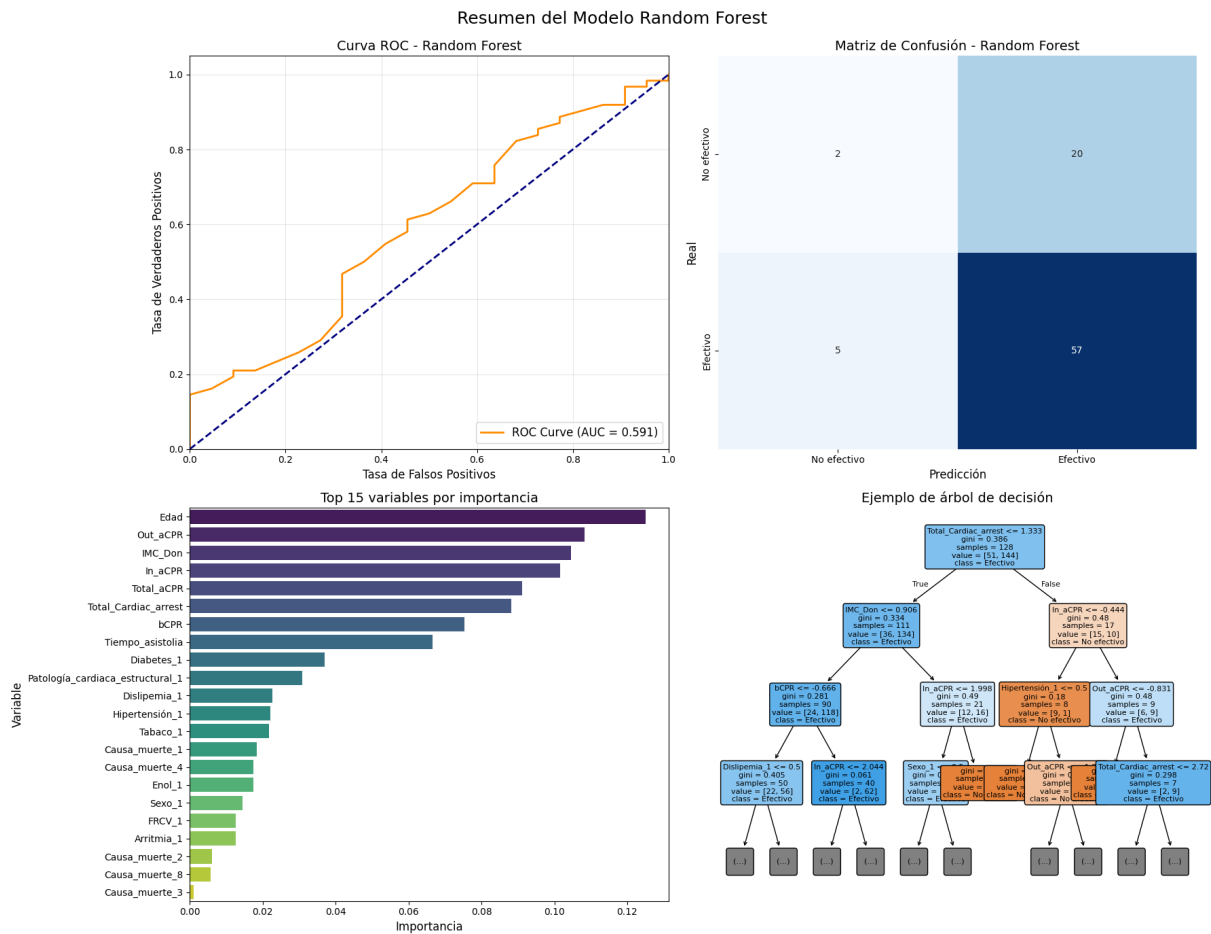


Figure 8: Resumen modelo Random forest

2.2.5 Árbol de decision

Un **árbol de decisión** es un modelo que se utiliza para tomar decisiones o hacer predicciones a partir de los datos de los que se disponen.

Se parece a un *diagrama de preguntas*: en cada paso se evalúa una condición (por ejemplo, “¿la edad es menor a 50 años?”) y, según la respuesta (sí o no), se sigue una rama diferente del árbol.

- El árbol comienza con una **pregunta inicial** sobre una variable clínica, por ejemplo: “¿bCPR es menor a 21.5 minutos?”.
- Según la respuesta, se pasa a otra pregunta, como: “¿el paciente tiene diabetes?”.
- Este proceso se repite hasta llegar a una **conclusión**: por ejemplo, “el donante es eficaz” o “no eficaz”.

Cada nodo del árbol representa una decisión basada en una variable clínica, y permite seguir un camino lógico hasta una predicción final.

En este caso se ha diseñado un árbol de decisión con una profundidad de 5 para que no resulte excesivamente complejo.

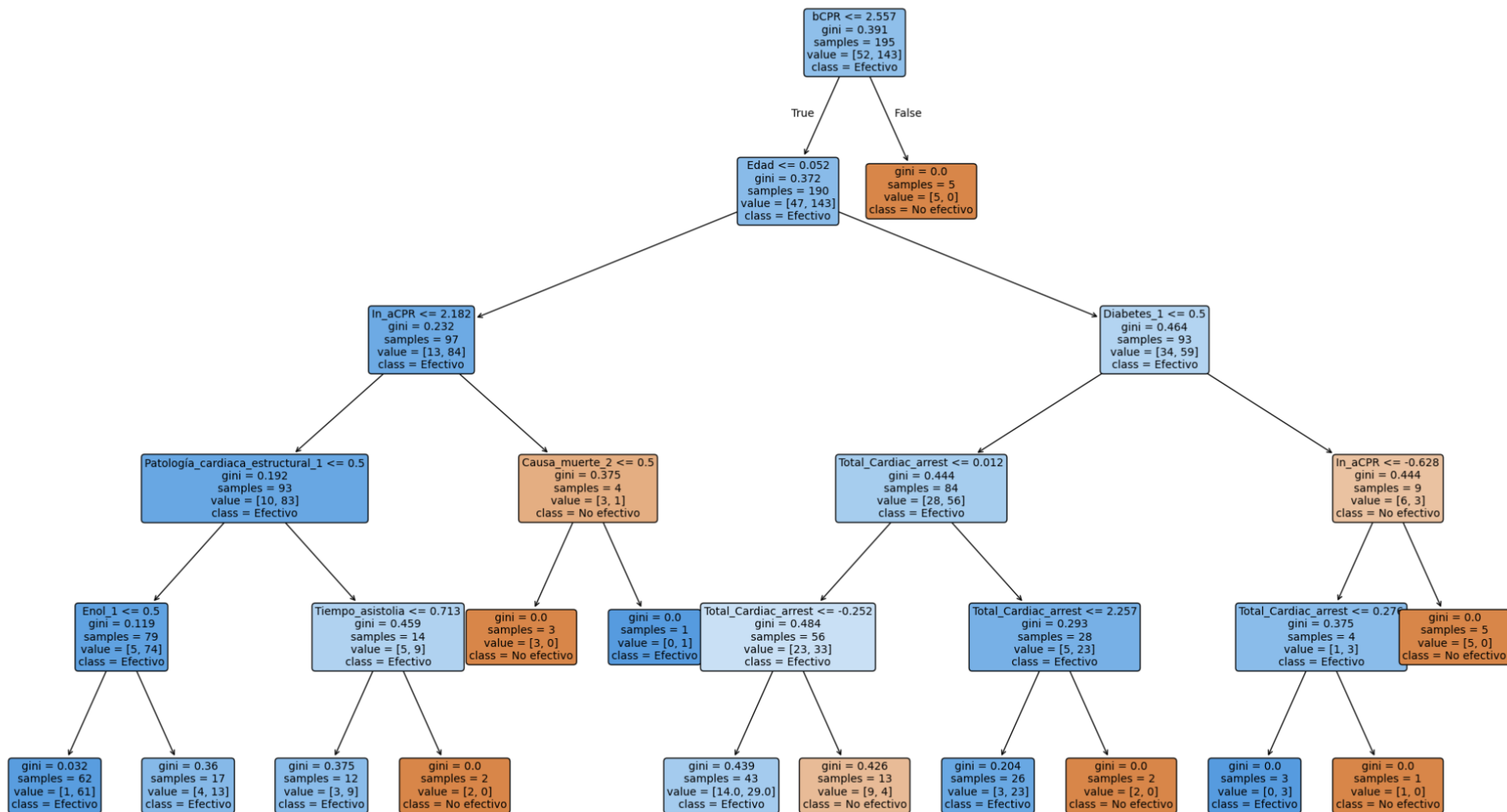


Figure 9: Árbol de decisión (variables estandarizadas)

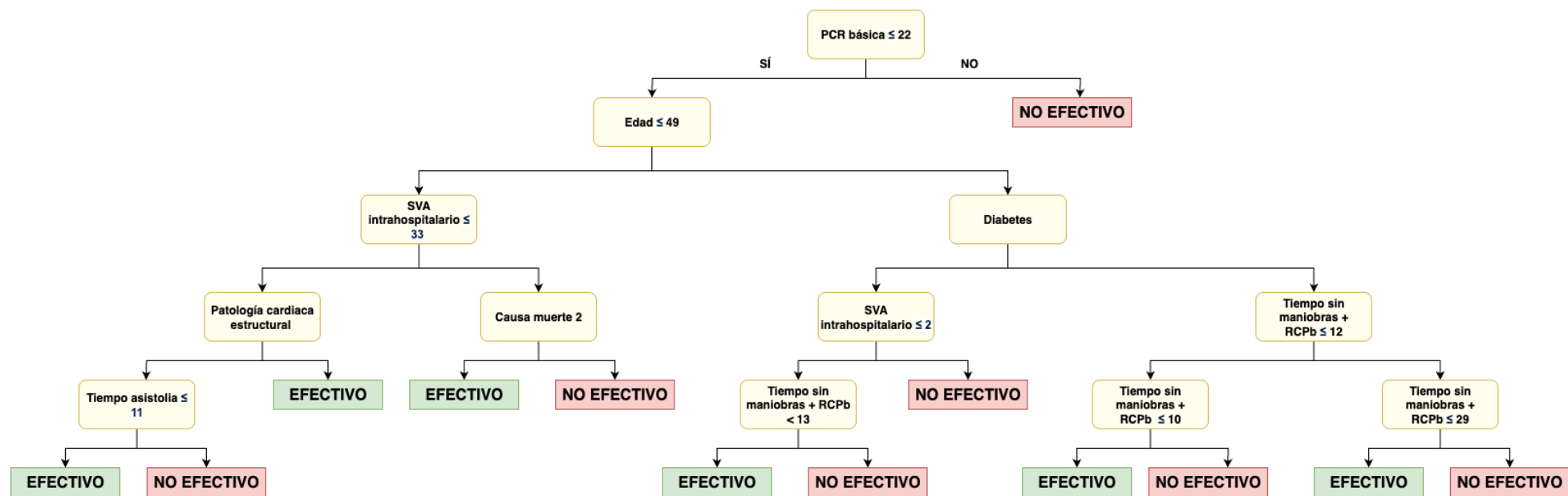


Figure 10: Árbol de decisión (variables en unidades normales)

2.2.6 Gradient Boosting

Gradient Boosting es un método de aprendizaje automático que se utiliza para hacer predicciones. Gradient Boosting no usa un solo modelo, sino que combina muchos modelos pequeños llamados *árboles de decisión*. Cada árbol aprende de los errores del árbol anterior. Así, el modelo mejora paso a paso.

- Se crea un primer árbol que intenta predecir el resultado (por ejemplo, si un donante es eficaz).
- Luego, se crea un segundo árbol que intenta corregir los errores del primero.
- Este proceso se repite muchas veces, mejorando la predicción en cada paso.

Suele tener una alta precisión, incluso con datos complejos. Se adapta bien a distintos tipos de variables (números, categorías, etc.). Permite identificar qué variables tienen más peso en la predicción.

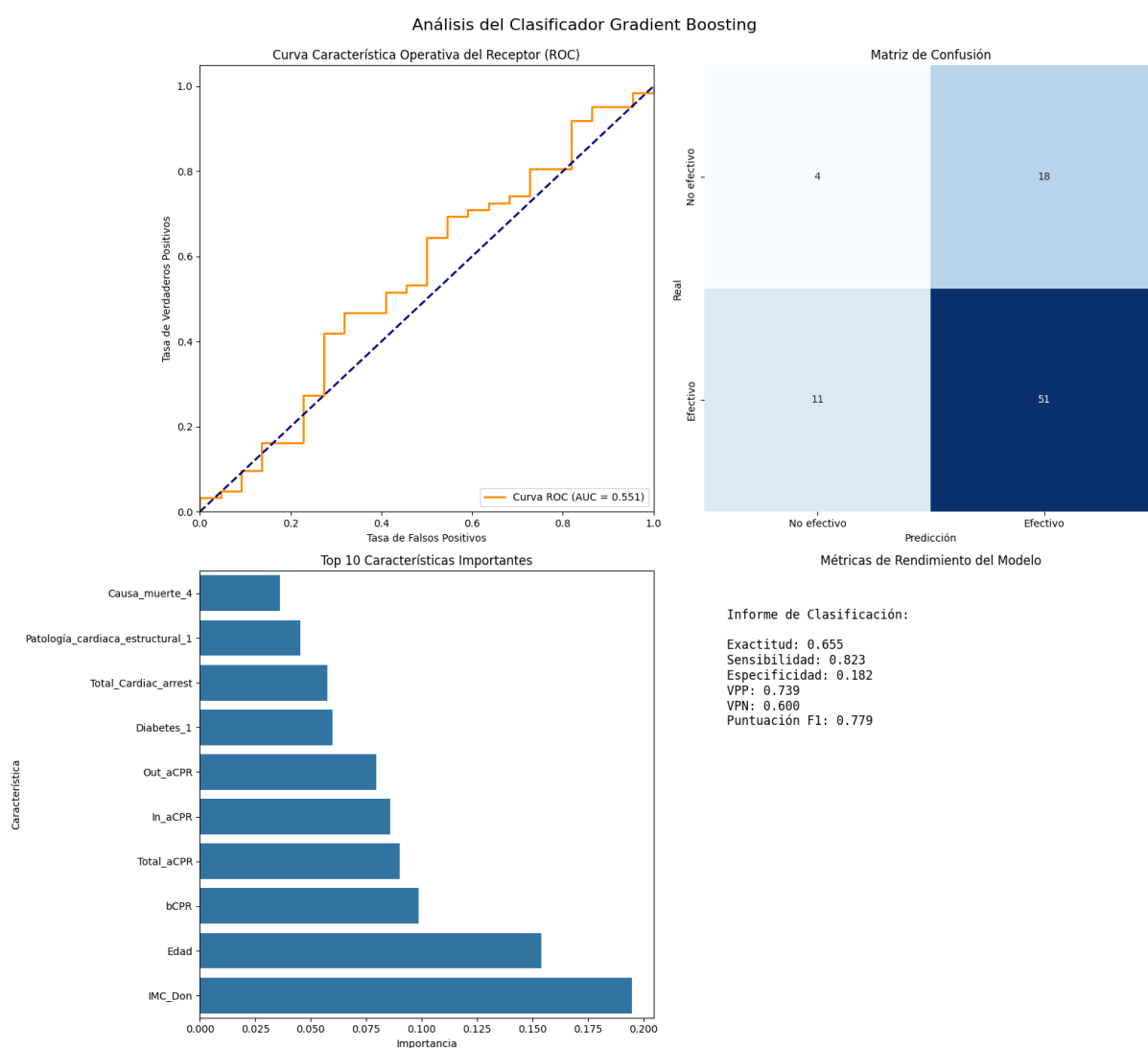


Figure 11: Gradient boosting

2.2.7 XGBoost

XGBoost es un modelo de aprendizaje automático que se utiliza para hacer predicciones a partir de datos. En este caso, lo usamos para predecir si un donante será eficaz o no, en función de sus características clínicas.

XGBoost construye una serie de pequeños árboles de decisión, uno tras otro. Cada árbol intenta corregir los errores que cometió el anterior, aprendiendo de los casos en los que el modelo falló. Al final, se combinan todos los árboles para tomar una decisión más precisa.

Es una **version optimizada y mejorada de Gradient boosting** Este modelo es especialmente útil porque:

- Es muy preciso y ha demostrado buen rendimiento en muchos problemas médicos y clínicos.
- Funciona bien incluso cuando hay muchas variables involucradas.
- Nos permite saber qué variables son más importantes para predecir si un donante será eficaz.

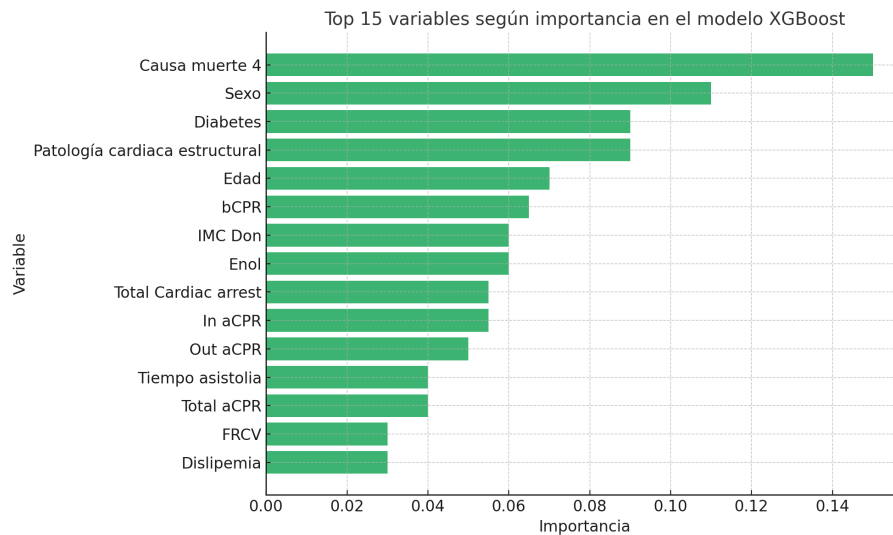


Figure 12: Importancia de variables en XGBoost

2.2.8 Comparativa final

Modelo	AUC Test	Exactitud	Sensibilidad	Especificidad	VPP	VPN
Regresión Logística	0.738	0.738	0.919	0.227	0.770	0.500
KNN	0.687	0.750	0.968	0.136	0.759	0.600
SVM	0.600	0.738	0.984	0.045	0.744	0.500
Random Forest	0.591	0.702	0.919	0.091	0.740	0.286
Árbol de decisión	0.587	0.607	0.629	0.545	0.796	0.343
Gradient Boosting	0.551	0.655	0.823	0.182	0.739	0.267
XGBoost	0.625	0.690	0.871	0.182	0.750	0.333

Table 6: Comparación del desempeño de distintos modelos de clasificación

En este estudio se comparan varios modelos de aprendizaje automático para predecir la eficacia de los donantes en función de sus características clínicas. Dado que en este contexto es más importante identificar correctamente a los donantes eficaces, se da prioridad a la **sensibilidad** del modelo, incluso si esto conlleva un mayor número de falsos positivos.

- El modelo de **regresión logística** muestra el mejor equilibrio entre sensibilidad (0.919), precisión (0.770) y área bajo la curva ($AUC = 0.738$), por lo que se considera el más adecuado en términos generales.
- El modelo **K-Nearest Neighbors (KNN)** presenta la **mayor sensibilidad** (0.968) y el mejor valor predictivo negativo (0.60), lo que resulta útil si se desea minimizar al máximo los falsos negativos.
- El modelo **SVM** también alcanza una sensibilidad muy alta (0.984), pero con una especificidad extremadamente baja (0.045), lo que puede generar un exceso de falsos positivos.
- Modelos como **Random Forest**, **Árbol de decisión** y **Gradient Boosting** muestran desempeños inferiores en general y menor capacidad discriminativa.

En resumen, la **regresión logística** se posiciona como el modelo más balanceado y clínicamente útil para esta tarea, ya que mantiene una alta sensibilidad sin comprometer significativamente la precisión ni la interpretabilidad.

3 Análisis global

3.1 Descriptivo y análisis univariado

Variable	General (N = 267)	Eficaz (N = 200)	No eficaz (N = 67)	p-valor
Edad	51.00 [44.00–58.00]	50.00 [42.00–57.00]	54.00 [48.00–60.00]	0.003
IMC	26.30 [24.22–29.40]	26.23 [24.22–29.32]	27.68 [24.70–30.90]	0.05
Tiempo sin maniobras	5.00 [1.00–10.00]	5.00 [1.00–10.00]	4.50 [0.00–10.00]	0.21
bCPR	2.00 [0.00–8.00]	2.00 [0.00–8.00]	2.00 [0.00–8.00]	0.8
PCR - SVA	10.00 [7.00–15.00]	10.00 [7.00–15.00]	10.00 [5.75–15.00]	0.57
SVA extrahospitalario	53.00 [43.00–69.00]	53.00 [44.00–68.00]	53.00 [39.75–69.75]	0.66
SVA intrahospitalario	5.00 [2.00–11.00]	5.00 [2.00–10.00]	5.00 [2.00–15.50]	0.54
SVA total	61.00 [50.00–77.00]	61.00 [50.00–75.00]	62.00 [46.00–80.00]	0.96
Preparación	25.00 [16.00–33.00]	23.00 [15.00–32.00]	25.00 [20.00–35.00]	0.06
Canulación	31.00 [25.00–40.00]	30.00 [25.00–40.00]	34.50 [30.00–41.75]	0.10
Total Cardiocompressor	58.00 [48.00–70.00]	56.00 [47.00–70.00]	60.50 [50.75–75.25]	0.03
WIT	141.00 [124.00–154.00]	140.00 [123.00–152.00]	145.00 [128.75–155.00]	0.06
NECMO	202.00 [172.00–235.00]	202.00 [172.00–237.00]	202.50 [170.00–232.00]	0.81
AST Inicial	136.00 [75.25–269.25]	129.00 [75.00–270.00]	152.00 [81.00–267.00]	0.58
ALT Inicial	112.00 [62.00–255.75]	112.00 [60.50–263.00]	116.00 [65.00–228.00]	0.92
GGT Inicial	39.50 [25.75–74.00]	39.00 [26.00–69.00]	45.50 [23.75–85.50]	0.41
FA Inicial	120.00 [88.75–169.50]	120.00 [91.00–169.00]	120.00 [84.25–171.75]	0.64
Na Inicial	141.00 [138.00–145.00]	141.00 [138.00–145.00]	141.00 [139.00–145.00]	0.88
Cre Inicial	1.54 [1.30–1.79]	1.51 [1.30–1.73]	1.56 [1.35–1.80]	0.31
Hct Inicial	43.50 [38.00–48.00]	44.00 [39.00–48.00]	42.00 [38.00–46.00]	0.02
AST H0	104.50 [57.25–209.50]	100.00 [53.00–197.50]	116.00 [64.00–231.50]	0.47
ALT H0	99.00 [41.75–220.75]	92.00 [43.00–220.00]	111.00 [39.00–208.00]	0.78
GGT H0	23.50 [14.00–43.00]	23.00 [15.00–38.00]	25.00 [13.25–51.50]	0.61
FA H0	64.00 [44.00–81.00]	66.00 [44.00–82.00]	63.00 [46.00–73.00]	0.25
pH H0	7.12 [6.94–7.33]	7.07 [6.92–7.27]	7.29 [6.99–7.46]	0.001
Na H0	137.00 [131.00–148.00]	137.00 [131.00–148.00]	136.50 [131.00–145.25]	0.65
Hct H0	23.00 [17.00–28.00]	23.00 [18.00–28.00]	19.00 [15.00–26.00]	0.008
AST H1	205.00 [99.00–400.75]	203.00 [100.75–400.25]	213.00 [98.00–398.00]	0.93
ALT H1	172.00 [71.50–371.00]	164.00 [70.50–380.00]	174.00 [74.50–326.50]	0.96
GGT H1	38.00 [21.00–66.50]	39.00 [22.00–66.00]	32.00 [20.25–70.75]	0.33
FA H1	74.00 [53.00–99.00]	76.00 [55.00–100.00]	72.00 [50.25–94.75]	0.27
pH H1	7.25 [7.03–7.40]	7.24 [7.03–7.39]	7.30 [7.03–7.44]	0.19
Na H1	143.00 [136.50–150.00]	142.00 [136.50–151.00]	143.00 [136.75–147.25]	0.83-
Hct H1	26.00 [20.00–31.00]	27.00 [22.00–31.00]	22.00 [18.00–30.00]	0.004

Table 7: Análisis de variables numéricas

Variable	General (N = 269)	Eficaz (N = 201)	No eficaz (N = 68)	p-valor
Sexo (H), N (%)	244 (90.70)	183 (91.04)	61 (89.71)	1.0
Grupo Sanguineo (O-)	78 (29.00)	56 (27.86)	22 (32.35)	0.35
Causa muerte (IAM), N (%)	165 (61.80)	123 (61.19)	42 (63.64)	0.43
FRCV, N (%)	198 (74.16)	146 (72.64)	52 (76.47)	0.55
Arritmia, N (%)	12 (4.47)	6 (2.99)	6 (8.82)	0.08
Patología cardiaca estructural, N (%)	48 (17.84)	28 (13.93)	20 (29.41)	0.007
Diabetes, N (%)	20 (7.46)	9 (4.48)	11 (16.18)	0.003
Hipertensión, N (%)	80 (29.74)	58 (28.86)	22 (32.35)	0.66
Dislipemia, N (%)	84 (31.23)	55 (27.36)	29 (42.65)	0.02
Tabaco, N (%)	130 (48.51)	100 (49.75)	30 (44.12)	0.55
Enol, N (%)	40 (14.87)	28 (13.93)	12 (17.65)	0.56
AST Init-H0, (0) N (%)	139 (51.67)	106 (52.74)	33 (48.53)	0.37
AST Init-H1, (0) N (%)	125 (46.47)	94 (46.77)	29 (42.65)	0.04
AST H0-H1, (0) N (%)	107 (39.78)	74 (36.82)	33 (48.53)	0.24
ALT Init-H0, (0) N (%)	141 (52.43)	106 (52.74)	35 (51.47)	0.85
ALT Init-H1, (0) N (%)	133 (49.44)	101 (50.25)	32 (47.06)	0.23
ALT H0-H1, (0) N (%)	123 (45.72)	83 (41.29)	40 (58.82)	0.05
GGT Init-H0, (0) N (%)	144 (53.53)	110 (54.73)	34 (50.00)	0.38
GGT Init-H1, (0) N (%)	177 (66.29)	135 (67.16)	44 (64.71)	0.004
GGT H0-H1, (0) N (%)	134 (50.00)	90 (44.78)	44 (64.71)	0.01
FA Init-H0, (0) N (%)	113 (42.01)	81 (40.30)	32 (47.06)	0.70
FA Init-H1, (0) N (%)	164 (61.04)	120 (59.70)	44 (64.71)	0.80
FA H0-H1, (0) N (%)	200 (74.35)	145 (72.14)	55 (80.88)	0.62
Na Init-H0, (0) N (%)	227 (84.39)	185 (92.04)	62 (91.18)	0.57
Na Init-H1, (0) N (%)	253 (94.42)	190 (94.53)	63 (92.65)	1.0
Na H0-H1, (0) N (%)	253 (94.42)	190 (94.53)	63 (92.65)	1.0
Hct Init-H0, (0) N (%)	134 (49.81)	102 (50.75)	32 (47.06)	0.52
Hct Init-H1, (0) N (%)	167 (62.17)	126 (62.69)	41 (60.29)	0.28
Hct H0-H1, (0) N (%)	200 (74.35)	149 (74.13)	51 (75.00)	0.77

Table 8: Análisis de variables categóricas

4 Análisis multivariado

4.1 Correlación

Para el análisis multivariado, dada la cantidad de variables del dataset, se hace un pestudio previo de correlación, con el objetivo de eliminar multicolinealidad. Se muestra en la imagen clusters de correlación significativa entre variables, que parece encontrarse en las variables AST y ALT para los diferentes momentos de analítica, por ese motivo se hacen los gráficos de estas variables y se decide eliminar la ALT.

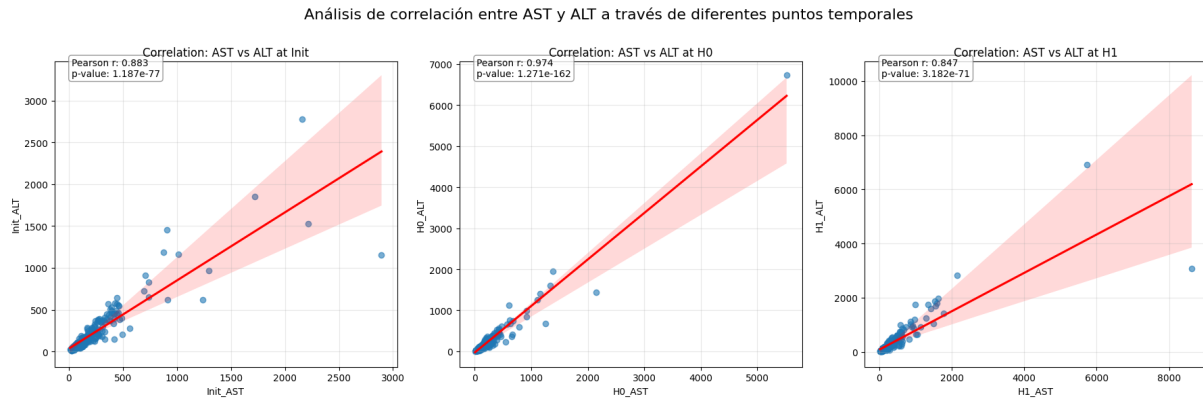


Figure 13: Estudio de correlación entre AST y ALT

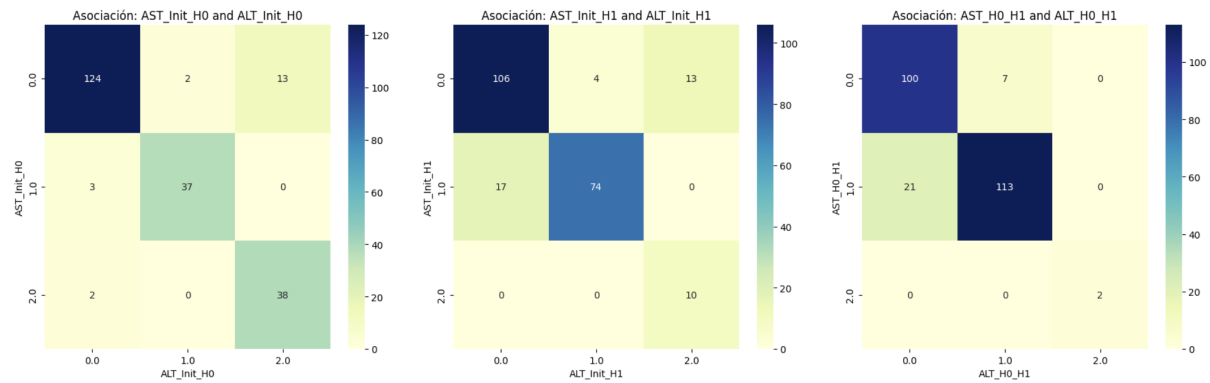


Figure 14: Estudio de asociación de las nuevas variables generadas para el estudio de la evolución de AST y ALT

4.2 PCA

Se realiza asimismo un *Principal Component Analysis* para intentar hacer una reducción de dimensionalidad. No obstante, la varianza que explica cada componente es muy pequeña, lo que obligaría a coger muchas variables, yendo en contra del objetivo de este método.

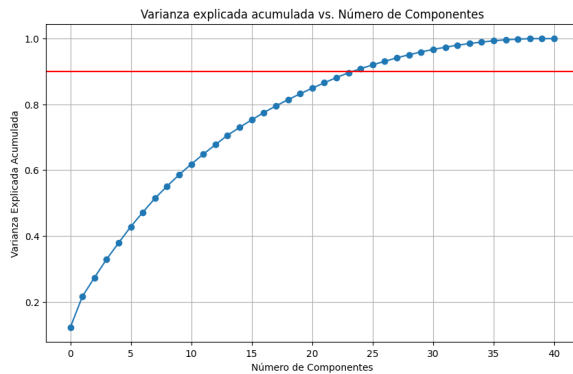


Figure 15: Gráfico de varianza acumulada

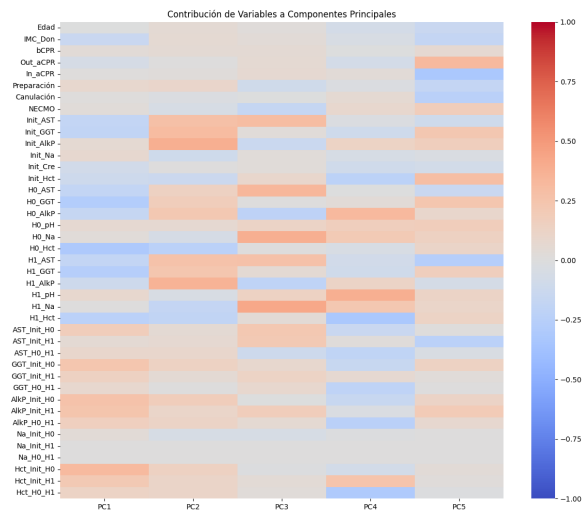


Figure 16: Matriz de contribución de variables a los componentes

En el gráfico de varianza explicada acumulada se observa que:

- Se requieren aproximadamente **23 componentes** para explicar el **90 % de la varianza total** (representado por la línea roja).
- Esto sugiere que, aunque el conjunto de datos contiene más de 40 variables, muchas de ellas no aportan información única o diferenciadora.
- Existe redundancia entre las variables, y el PCA permite resumirlas en un número menor de dimensiones, conservando la mayor parte de la información.

El mapa de calor de las cargas de las variables muestra lo siguiente:

- El **PC1** está dominado por variables bioquímicas como **GGT_Init_H0**, **AlkP_Init_H0**, **Na_H0_H1**, lo que sugiere un eje de información metabólica o hepática.
- El **PC2** está influenciado por variables dinámicas relacionadas con la progresión enzimática, como **ALT_H0_H1**, **H1_AST**, **AST_H0_H1**.
- Los componentes posteriores capturan variabilidad más específica o residual (como **pH**, **Hct**, o variables temporales como **Out_aCPR**, **In_aCPR**).

Así pues, la combinación de ambos análisis permite **reducir la dimensionalidad** del conjunto de datos de forma efectiva, explicando un alto porcentaje de la varianza con un número reducido de componentes. Algunas variables presentan **información redundante** y podrían ser eliminadas en análisis o modelos posteriores, especialmente si representan distintas mediciones de la misma magnitud. Este enfoque resulta útil para minimizar la **multicolinealidad** y facilitar interpretaciones más robustas en modelos predictivos.

4.3 Modelos de predicción