UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Graduado en Ingeniería de la salud

Evaluación de Modelos de Aprendizaje Automático para la Predicción de Enfermedades Cardíacas

Evaluation of Machine Learning Models for Predicting Heart Disease

Realizado por
Pablo Molina Sánchez

Tutorizado por
Cristóbal Barba González
Sandro José Hurtado Requena

Departamento
Lenguajes y Ciencias de la Computación

MÁLAGA, junio 2025

Abstract

Cardiovascular diseases remain the leading cause of mortality worldwide, prompting the development of predictive tools for early detection. This study evaluates and compares five machine learning models—Support Vector Machine (SVM), Random Forest, Decision Tree, Multi-Layer Perceptron (MLP), and XGBoost—using a multi-class classification approach applied to the Cleveland Heart Disease dataset. The models were trained using advanced preprocessing techniques, class balancing through SMOTE, and variable selection via ANOVA F-test. Their predictive performance was assessed using F1-macro, accuracy, and classification reports. To enhance interpretability, SHAP (SHapley Additive exPlanations) was used to analyze feature importance across classes. Results highlight XGBoost and MLP as top-performing models, and underscore the clinical value of engineered features such as careful and emergency_risk_score. These findings support the application of interpretable machine learning models in clinical decision-making

Keywords: Heart disease, machine learning, classification

Resumen

Las enfermedades cardiovasculares siguen siendo una de las principales causas de muerte a nivel mundial, lo que motiva la búsqueda de herramientas predictivas eficaces para su detección temprana. Este trabajo analiza y compara cinco modelos de aprendizaje automático —SVM, Random Forest, Árbol de Decisión, Red Neuronal Artificial y XGBoost— mediante un enfoque de clasificación multiclase aplicado al conjunto de datos Cleveland Heart Disease. Los modelos se entrenaron aplicando técnicas de preprocesamiento avanzado, balanceo de clases con SMOTE y selección de variables mediante ANOVA F-test. El rendimiento se evaluó a través de métricas como F1-macro, exactitud y reportes de clasificación. Para mejorar la interpretabilidad, se empleó SHAP, permitiendo analizar la importancia de las variables por clase. Los resultados respaldan el uso de modelos interpretables de aprendizaje automático en la toma de decisiones médicas

Palabras clave: **Enfermedad cardíaca, aprendizaje automático, clasificación**

Índice

1 Introducción

1.1. Motivación

Las enfermedades cardíacas representan una de las principales causas de mortalidad en el mundo [1], afectando a millones de personas cada año. Factores como la edad, el estilo de vida y ciertas condiciones médicas pueden aumentar el riesgo de desarrollar patologías cardiovasculares. Debido a su alta incidencia y gravedad, la detección temprana de estas enfermedades es fundamental para mejorar el pronóstico y reducir la tasa de mortalidad [2]. En este estudio, se utilizará el conjunto de datos de enfermedades cardíacas de Cleveland, el cual contiene información relevante sobre distintos factores de riesgo, como la presión arterial, el colesterol, la frecuencia cardíaca y otros indicadores clínicos [3]. A partir de estos datos, se busca desarrollar un modelo de predicción que permita identificar la presencia de enfermedad cardíaca en un paciente, lo que podría contribuir a mejorar los diagnósticos y la toma de decisiones en el ámbito médico. El análisis de estos datos no solo permitirá evaluar la relación entre distin- tas variables y el desarrollo de enfermedades cardíacas, sino que también abrirá la posibilidad de optimizar herramientas de inteligencia artificial aplicadas a la salud. Con ello, se espera aportar conocimiento valioso que ayude a una mejor comprensión de este problema y facilite estrategias preventivas más eficaces.

1.2. Objetivos

Analizar los principales factores de riesgo asociados a la presencia de enfermedades cardíacas, como la presión arterial, el colesterol, la frecuencia cardíaca y otros indicadores clínicos. Desarrollar un modelo de predicción basado en técnicas de aprendizaje automático

que per- mita identificar la probabilidad de que un paciente padezca una enfermedad cardíaca. Evaluar la relación entre distintos atributos del conjunto de datos y la presencia de enfermedad cardíaca, con el fin de identificar patrones y correlaciones significativas. Comparar la efectividad de diferentes algoritmos de clasificación en la detección de enfermedades cardíacas, optimizando su precisión y capacidad de generalización. Contribuir a la mejora de herramientas de diagnós- tico médico mediante la aplicación de modelos predictivos, facilitando la detección temprana y la toma de decisiones clínicas.

1.3. Estructura del documento

Este Trabajo de Fin de Grado se organiza en los siguientes cinco bloques principales. Cada uno de ellos se desarrolla en uno o varios capítulos, garantizando una exposición clara y coherente del análisis realizado: Este Trabajo de Fin de Grado se organiza en los siguientes cinco bloques principales. Cada uno de ellos se desarrolla en uno o varios capítulos, garantizando una exposición clara y coherente del análisis realizado:

1. **Contexto y Objetivos**

   En esta primera parte se introduce el contexto en el que se enmarca el proyecto. Se presenta la motivación del estudio, el problema de la clasificación del nivel de emergencia y su relevancia en aplicaciones reales. Además, se definen con claridad los objetivos del trabajo, tanto a nivel teórico como práctico.

2. **Análisis y Preprocesamiento de Datos**

   Esta sección describe en detalle el conjunto de datos utilizado, incluyendo la naturaleza de las variables (demográficas, clínicas, etc.) y la variable objetivo (num). Se llevan a cabo varias tareas de preprocesamiento:

   - Tratamiento de valores ausentes
   - Creación de variables derivadas (feature engineering)
   - Análisis exploratorio de datos (EDA) con resúmenes estadísticos y visualizaciones
   - Detección y tratamiento de valores atípicos (outliers)

- Codificación de variables categóricas
- Estudio de correlaciones

Estas tareas permiten una comprensión más profunda de los datos y aseguran un con- junto limpio y adecuado para el modelado.

3. **Desarrollo y Evaluación de Modelos**

   Este bloque aborda el diseño, entrenamiento y evaluación de modelos de aprendizaje automático para predecir el nivel de emergencia. Se detallan los siguientes pasos:

   - División del conjunto de datos en entrenamiento y prueba mediante muestreo es- tratificado
   - Selección de variables mediante métodos estadísticos (SelectKBest con ANOVA F- test)
   - Tratamiento del desbalanceo de clases utilizando SMOTE
   - Escalado de variables con StandardScaler
   - Entrenamiento y ajuste de modelos mediante GridSearchCV en clasificadores como Decision Tree, Random Forest, SVM y XGBoost
   - Evaluación del rendimiento con métricas como accuracy, F1-score macro y reportes de clasificación
   - Análisis de importancia de características mediante SHAP. Esta técnica permite identificar qué variables contribuyen más a cada predicción y proporciona explicaciones tanto globales como locales del comportamiento del modelo

4. **Selección del mejor modelo**

   Se llevará a cabo un análisis comparativo del rendimiento de los distintos modelos de aprendizaje automático entrenados previamente. Para ello, se examinan las métricas de evaluación obtenidas por cada clasificador, tales como el *accuracy*, el *F1-score macro* y las puntuaciones individuales por clase. Tras revisar los resultados, se selecciona como mejor modelo aquel que logra el mayor compromiso entre rendimiento general, robustez

y capacidad de generalización sobre el conjunto de test.

5. **Conclusiones y Trabajo Futuro**
   Finalmente, se sintetizan los principales hallazgos del proyecto, reflexionando sobre los resultados alcanzados y los retos encontrados. Además, se proponen posibles líneas de mejora y posibilidades de expansión del modelo a futuro.

1.4. Tecnologías usadas

Para el correcto desarrollo del proyecto, hemos utilizado el entorno Python y sus librerías más importantes para un correcto análisis (Scikit-Learn, Matplotlib, Seaborn, Pandas,etc), Git para control de versiones y Latex para el desarrollo de la memoria.

2

Análisis y Preprocesamiento de Datos

En esta sección se describen en detalle las tareas realizadas sobre el conjunto de datos original con el objetivo de preparar la información para su uso en modelos de aprendizaje automático. Este proceso incluye la exploración inicial, el tratamiento de valores atípicos y ausentes, la codificación de variables categóricas, la selección de características y en definitiva, la preparación de los datos para el modelado.

2.1. Tratamiento de Valores Faltantes

Durante el análisis exploratorio de los datos, se detectó la presencia de valores perdidos (NaN) en varias variables del conjunto de datos. Primeramente, debemos definir qué son los valores faltantes. Los valores faltantes (missing values) se producen cuando no se dispone de un dato válido para una determinada variable en una observación específica, lo que puede deberse a errores de medición, problemas de registro o decisiones de omisión en la recogida de datos[4]. En Python (pandas), se representan como NaN (Not a Number). Su presencia puede deberse a errores de medición, problemas en la recolección de datos o simplemente a que no fue posible obtener el valor en cuestión.

En este conjunto de datos, se calculó el porcentaje de valores faltantes por variable como podemos ver en la tabla 1. Como se observa, las variables ca y thal presentan más de un 50 % de valores ausentes, por lo que se decidió eliminarlas del análisis para evitar sesgos y mantener la calidad del modelo.

Para el resto de variables con datos ausentes en menor proporción (trestbps, tchol, fbs, restecg, thalch, exang, oldpeak, slope), se aplicaron diferentes técnicas de imputación según el tipo de dato:

Variable	**Valores faltantes**
trestbps	6.41 %
chol	3.26 %
fbs	9.78 %
restecg	0.22 %
thalch	5.98 %
exang	5.98 %
oldpeak	6.74 %
slope	33.59 %
ca	66.41 %
thal	52.83 %

Tabla 1: Porcentaje de valores faltantes por variable en el conjunto de datos.

- Para las variables numéricas, se utilizó la media de los valores observados.
- Para las variables categóricas, se imputaron los valores faltantes respetando la distribución de frecuencias del resto de observaciones.

Para aplicar estas técnicas, se implementó una función personalizada en Python denominada na_filler, encargada de rellenar los valores faltantes de acuerdo con el tipo de cada variable. A continuación, se muestra el código de la función:La lógica de la función es la siguiente:

```
def na_filler(df):
    for i in df.columns:
        if df[i].dtype in ['int64', 'float64']:
            df[i] = df[i].fillna(df[i].mean())

        elif df[i].dtype == 'object':
            frecuencias = df[i].value_counts(normalize=True)
            df[i] = df[i].apply(lambda x: np.random.choice(
                frecuencias.index, p=frecuencias.values) if pd.
   isna(x) else x)
```

- Para variables numéricas (int64 o float64), los valores perdidos se rellenan con la media de la columna.

- Para variables categóricas (object), los valores perdidos se rellenan manteniendo la distribución original de frecuencias, usando una imputación aleatoria ponderada.

Gracias a este procedimiento se logró eliminar valores nulos de forma eficiente, permitiendo continuar con el análisis sin eliminar observaciones completas ni introducir sesgos estadísticos relevantes.

2.2. Descripción del Conjunto de Datos

El conjunto de datos utilizado contiene observaciones relacionadas con casos clasificados según su nivel de emergencia médica. Cada instancia incluye variables predictoras tanto cuantitativas como cualitativas (como datos demográficos, clínicos, etc.). Estas variables que conforman nuestro conjunto de datos se describen a continuación:

Variables de identificación y demográficas

2.2.1. Id

Identificador único de cada paciente u observación. Esta variable no tiene valor predictivo y se utiliza únicamente para fines de seguimiento o referencia.

2.2.2. Dataset

La variable dataset indica el centro hospitalario o la base de datos de origen de cada observación. En este conjunto de datos se integran registros provenientes de cuatro fuentes distintas: Cleveland, Hungary, VA Long Beach y Switzerland, lo cual convierte al dataset en un estudio multicéntrico.

Centro	Frecuencia	Porcentaje (%)
Cleveland	304	33.04
Hungary	293	31.85
VA Long Beach	200	21.74
Switzerland	123	13.37

Tabla 2: Distribución de frecuencias por centro en la variable dataset.

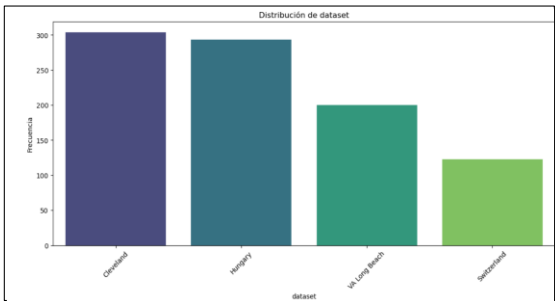

Figura 1: Distribución de la variable dataset.

Tal como se muestra en la **Tabla 2**, el mayor número de observaciones proviene del centro de Cleveland (33.04 %), seguido de Hungary (31.85 %). En menor proporción se encuentran VA Long Beach (21.74 %) y Switzerland (13.37 %).

La **Figura 1** representa gráficamente esta distribución. Esta información es relevante, ya que la procedencia de los datos puede introducir variabilidad relacionada con diferencias en protocolos clínicos, técnicas de medición o características poblacionales.

En consecuencia, podría ser necesario aplicar técnicas de normalización o incluir la variable dataset como covariable en los modelos predictivos para controlar posibles efectos de confusión entre centros.

2.2.3. Age

Edad del paciente, expresada en años. Es una variable continua que puede estar relacionada con el riesgo cardiovascular.

Estadístico	Valor
Media	53.51
Mediana	54.0
Moda	54
Mínimo	28
Máximo	77
Rango	49
Desviación estándar	9.42
Varianza	88.82
Curtosis	-0.38
Asimetría	-0.20
Percentil 25	47.0
Percentil 50 (Mediana)	54.0
Percentil 75	60.0
IQR (Rango Intercuartílico)	13.0

Tabla 3: Estadísticos descriptivos de la variable age.

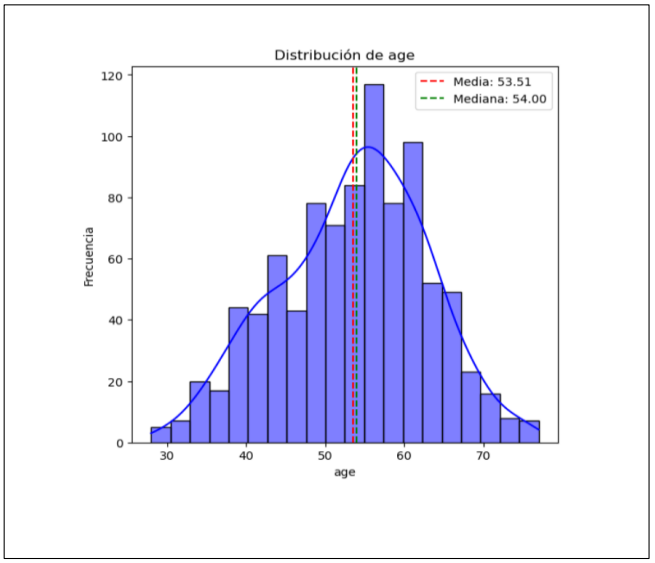

Figura 2: Distribución de valores en la variable age.

Como se observa en la **Tabla 3**, la media (53.51), la mediana (54.0) y la moda (54) son prácticamente idénticas. Esta coincidencia sugiere que la distribución de la variable edad es aproximadamente simétrica, lo cual es coherente con el valor de asimetría obtenido (-0.20), que está muy cercano a cero. La ligera asimetría negativa indica que hay una pequeña tendencia a acumular valores en la parte derecha de la distribución, aunque no de forma significativa como bien podemos ver en la **Figura 2** .

La curtosis de -0.38 sugiere que la distribución presenta colas algo más ligeras que una distribución normal, sin una concentración excesiva de valores extremos. El rango total de edades es de 49 años, con valores mínimos y máximos de 28 y 77 años, respectivamente, lo cual indica que la muestra cubre un espectro amplio de edades adultas.

El rango intercuartílico (IQR) es de 13 años, lo que indica que el 50 % central de la muestra se encuentra entre los 47 y los 60 años. Esto, junto con una desviación estándar de 9.42, señala una dispersión moderada en la variable.

En términos clínicos, estos resultados muestran que la mayoría de los pacientes incluidos en el conjunto de datos son adultos de mediana edad (**Figura 2**), siendo esta una población típicamente asociada a un riesgo creciente de enfermedades cardiovasculares. Por tanto, la edad se perfila como una variable predictiva clave a considerar en el desarrollo de modelos de clasificación del nivel de emergencia médica.

2.2.4. Sex

La variable sex representa el sexo biológico del paciente, codificado como Male o Female. Esta es una variable categórica de tipo binario, que puede tener implicancias relevantes desde el punto de vista clínico, ya que existen diferencias fisiológicas y epidemiológicas entre hombres y mujeres en relación con la enfermedad cardiovascular.

Sexo	Frecuencia	Porcentaje (%)
Male	726	78.91
Female	194	21.09

Tabla 4: Distribución de frecuencias para la variable sex.

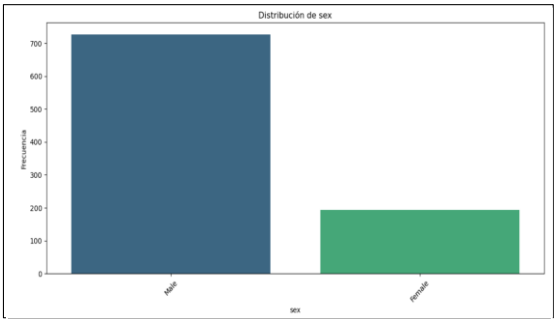

Figura 3: Distribución de la variable sex.

En la **Tabla 4** se muestra la distribución de frecuencias. Observamos que el 78.91 % de los registros corresponden a pacientes varones, mientras que el 21.09 % corresponde a mujeres. Esta desproporción podría estar relacionada con una mayor prevalencia o detección de enfermedad coronaria en hombres en la población estudiada, o bien con un sesgo de muestreo en el

conjunto de datos original.

La **Figura 3** ilustra gráficamente esta distribución, permitiendo apreciar visualmente el desequilibrio entre ambas categorías. Esta asimetría debe tenerse en cuenta durante el análisis, ya que puede influir en el rendimiento de los modelos predictivos, especialmente si se desea garantizar equidad entre subgrupos poblacionales. En este sentido, podría considerarse aplicar técnicas de balanceo o realizar análisis estratificados por sexo para evaluar posibles sesgos.

Variables clínicas

2.2.5. Cp

La variable cp (chest pain type) representa el tipo de dolor torácico que presenta el paciente, un síntoma clave en el diagnóstico de enfermedades cardíacas. Esta variable categórica puede tomar uno de los siguientes valores:

- Typical angina: dolor torácico clásico asociado con isquemia.
- Atypical angina: dolor no característico pero posiblemente de origen cardíaco.
- Non-anginal pain: dolor en el pecho no relacionado con el corazón.
- Asymptomatic: ausencia de dolor torácico.

Tipo de dolor	Frecuencia	Porcentaje (%)
asymptomatic	496	53.91
non-anginal	204	22.17
atypical angina	174	18.91
typical angina	46	5.00

Tabla 5: Distribución de frecuencias para la variable cp.

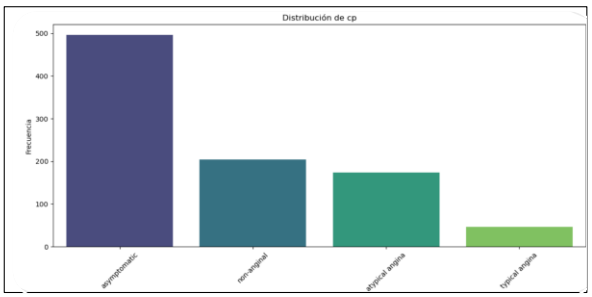

Figura 4: Distribución de la variable cp.

Como se observa en la **Tabla 5**, más de la mitad de los pacientes (53.91 %) se presentan como asintomáticos, lo cual puede dificultar el diagnóstico temprano. Le siguen los casos con

dolor no anginoso (22.17 %) y angina atípica (18.91 %). Solo un 5 % de los pacientes presenta angina típica, que suele ser el indicador más directo de una posible cardiopatía.

La **Figura 4** representa gráficamente esta distribución. Esta variable puede tener un gran valor predictivo, ya que distintos tipos de dolor torácico están asociados a diferentes grados de probabilidad de enfermedad cardíaca. Por ello, su correcta codificación e interpretación es crucial en la etapa de modelado.

2.2.6. Trestbps

La variable trestbps representa la presión arterial en reposo medida en milímetros de mercurio (mm Hg). Es una variable cuantitativa continua de tipo numérico y clínicamente relevante, ya que niveles elevados de presión arterial son un factor de riesgo reconocido en enfermedades cardiovasculares.

Estadístico	Valor
Media	132.13
Mediana	130.00
Moda	120
Mínimo	0.00
Máximo	200.00
Rango	200.00
Desviación estándar	18.44
Varianza	340.18
Curtosis	3.37
Asimetría	0.22
Percentil 25	120.00
Percentil 50 (Mediana)	130.00
Percentil 75	140.00
IQR	20.00

Tabla 6: Estadísticos descriptivos de la variable trestbps.

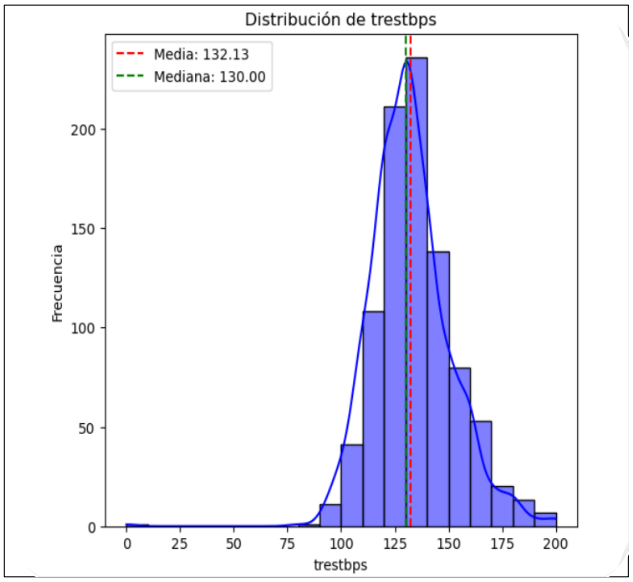

Figura 5: Distribución de la variable trestbps.

En la **Tabla 6** se presentan sus principales estadísticos descriptivos. La media es de 132.13 mm Hg y la mediana de 130.00 mm Hg, lo cual indica una distribución aproximadamente simétrica (confirmado también por la asimetría de 0.22). La moda es 120 mm Hg, valor que suele considerarse dentro del rango normal.

Sin embargo, se observa un valor mínimo atípico de 0 mm Hg, el cual probablemente corresponda a un error de medición o registro, ya que fisiológicamente es inviable. Este outlier podría afectar las estadísticas y sugiere la necesidad de una limpieza o tratamiento de datos, lo cual abordaremos más adelante.

La distribución, representada en la **Figura 5**, muestra una ligera curtosis (3.37), lo que indica colas algo más pesadas que una distribución normal. La dispersión de los datos es moderada, con una desviación estándar de 18.44 mm Hg y un rango intercuartílico (IQR) de 20 mm Hg.

En conjunto, estos datos reflejan que la mayoría de los pacientes tienen una presión en reposo entre 120 y 140 mm Hg, aunque existen valores extremos que deben considerarse durante el preprocesamiento para evitar distorsiones en los modelos predictivos.

2.2.7. Chol

La variable chol representa la concentración de colesterol sérico total en sangre, medida en miligramos por decilitro (mg/dl). Es una variable cuantitativa continua de alta relevancia clínica, ya que el colesterol elevado es uno de los principales factores de riesgo asociados a enfermedades cardiovasculares.

Estadístico	Valor
Media	199.35
Mediana	221.00
Moda	0
Mínimo	0.00
Máximo	603.00
Rango	603.00
Desviación estándar	108.82
Varianza	11841.46
Curtosis	0.18
Asimetría	-0.63
Percentil 25	178.50
Percentil 50 (Mediana)	221.00
Percentil 75	267.00
IQR	88.50

Tabla 7: Estadísticos descriptivos de la va- riable chol.

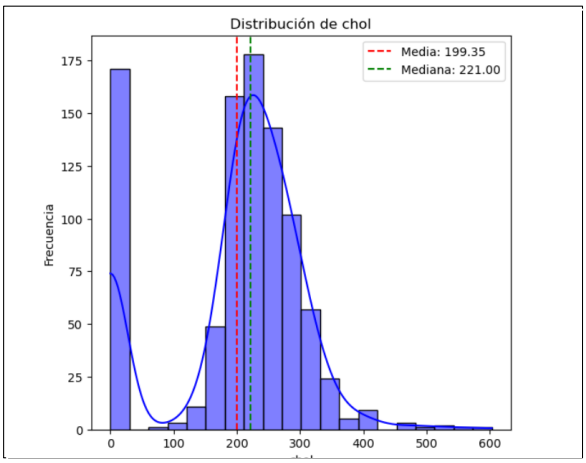

Figura 6: Distribución de la variable chol.

Como se observa en la **Tabla 7**, la media de colesterol es de 199.35 mg/dl, mientras que la mediana es de 221.00 mg/dl, lo que sugiere una asimetría negativa moderada en la distribución (asimetría de -0.63). La curtosis cercana a cero (0.18) indica una distribución relativamente mesocúrtica.

El valor mínimo registrado es 0 mg/dl, un dato fisiológicamente implausible y que proba-

blemente corresponde a errores de registro o valores faltantes mal codificados. Esta observación debe ser tratada adecuadamente en la fase de preprocesamiento para evitar que afecte negativamente el análisis y los modelos.

El rango es muy amplio (603 mg/dl), con un máximo extremo de 603 mg/dl. Esta gran dispersión también se refleja en la desviación estándar de 108.82 mg/dl y un IQR de 88.50 mg/dl.

La **Figura 6** muestra visualmente esta distribución, en la que pueden observarse tanto la presencia de valores atípicos como la ligera concentración hacia niveles más bajos.

Este tipo de variable puede beneficiarse de técnicas de imputación para los valores anómalos y, en algunos casos, transformaciones para mejorar su distribución en los modelos predictivos.

2.2.8. Fbs

La variable fbs (fasting blood sugar) indica si el nivel de glucosa en sangre en ayunas del paciente supera los 120 mg/dl. Se trata de una variable categórica binaria, donde True representa una glucemia en ayunas anormalmente elevada (>120 mg/dl), y False representa un valor normal (≤120 mg/dl).

Valor	Frecuencia	Porcentaje (%)
False	766	83.35
True	153	16.65

Tabla 8: Distribución de frecuencias de la variable fbs.

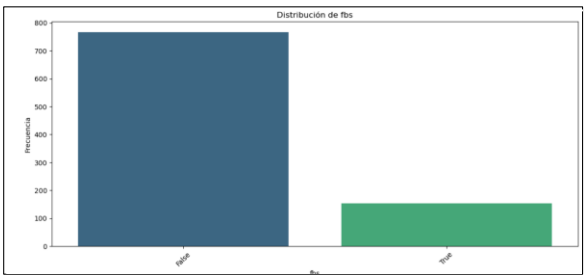

Figura 7: Distribución de la variable fbs.

Tal como se observa en la **Tabla 8**, la gran mayoría de los pacientes presenta niveles normales de glucosa en ayunas (83.35 %), mientras que solo un 16.65 % tiene niveles elevados. Esta desproporción sugiere un posible desequilibrio en la variable, lo cual podría ser relevante al momento de modelar, especialmente si se emplean técnicas sensibles a la distribución de

clases.

En la **Figura 7** se representa gráficamente esta distribución. Dado que niveles elevados de glucosa están asociados con diabetes o prediabetes, y estos a su vez con riesgo cardiovascular, esta variable puede ser de interés clínico en el desarrollo de modelos predictivos de enfermedad cardíaca.

Además, su binarización directa puede facilitar la interpretación, aunque podría explorarse un tratamiento adicional si se desea modelar más finamente la relación con otras variables fisiológicas o de laboratorio.

2.2.9. Restecg

La variable restecg describe el resultado del electrocardiograma en reposo del paciente. Es una variable categórica con tres posibles valores:

- **normal**: sin anomalías eléctricas observadas.
- **lv hypertrophy**: signos de hipertrofia del ventrículo izquierdo.
- **st-t abnormality**: anomalías en las ondas ST-T (elevaciones o depresiones), que pueden sugerir isquemia o infarto.

Resultado	Frecuencia	Porcentaje
normal	552	60.07
lv hypertrophy	188	20.46
st-t abnormality	179	19.48

Tabla 9: Distribución de frecuencias de la variable restecg.

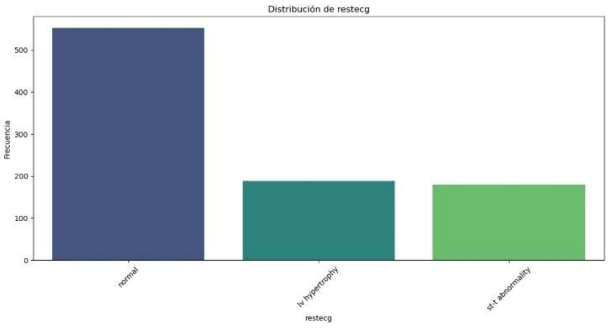

Figura 8: Distribución de la variable restecg.

Como se muestra en la **Tabla 9**, la mayoría de los pacientes presenta un electrocardiograma

en reposo normal (60.07 %). El resto se distribuye entre signos de hipertrofia ventricular izquierda (20.46 %) y anomalías en el segmento ST-T (19.48 %).

La **Figura 8** muestra esta distribución de forma gráfica. Este tipo de información es clínicamente relevante, ya que ciertas anomalías electrocardiográficas pueden estar asociadas con un mayor riesgo cardiovascular. Por tanto, la variable restecg puede aportar valor predictivo en los modelos de clasificación de enfermedad cardíaca.

Además, la inclusión de esta variable puede ayudar a reflejar alteraciones estructurales o eléctricas que no son evidentes a través de otras mediciones más básicas como la presión o el colesterol.

2.2.10. Thalch

La variable thalch representa la **frecuencia cardíaca máxima alcanzada durante una prueba de esfuerzo**. Es una medida fisiológica relevante, ya que permite evaluar la respuesta del corazón al ejercicio físico, aspecto clave en la detección de enfermedad coronaria.

A continuación, se resumen sus principales estadísticas descriptivas:

Estadístico	Valor
Media	137.53
Mediana	138.00
Moda	137.55
Mínimo	60.00
Máximo	202.00
Rango	142.00
Desviación estándar	25.15
Varianza	632.30
Curtosis	-0.32
Asimetría	-0.22
Percentil 25	120.00
Percentil 50 (Mediana)	138.00
Percentil 75	156.00
IQR	36.00

Tabla 10: Estadísticos descriptivos de la variable thalch.

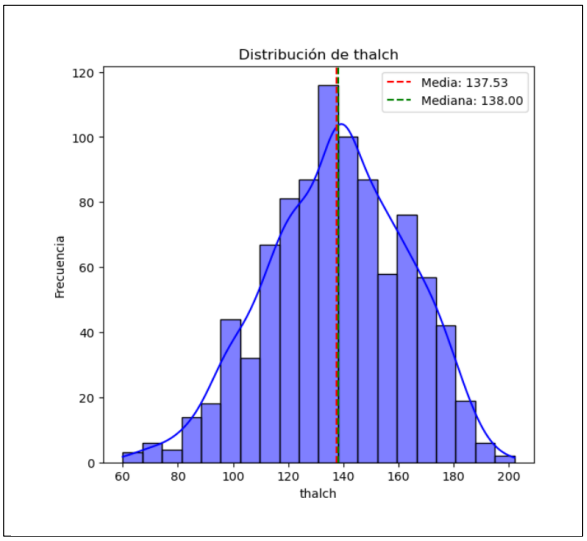

Figura 9: Distribución de la variable thalch.

La distribución de esta variable, mostrada en la **Figura 9**, presenta una forma aproximadamente simétrica, con una ligera asimetría negativa y curtosis cercana a la de una distribución normal, apoyándonos también numéricamente con la **Tabla 10.**

Este tipo de variable puede ser útil tanto para la detección de patrones clínicos como para la clasificación predictiva de pacientes con mayor riesgo cardíaco. Valores inusualmente bajos pueden indicar limitaciones físicas severas, mientras que valores altos pueden asociarse con un buen rendimiento cardíaco bajo esfuerzo.

2.2.11. Exang

La variable exang indica si el paciente presentó **angina inducida por ejercicio**, una condición que puede reflejar obstrucción de las arterias coronarias. Se trata de una variable binaria, donde True representa la presencia de angina inducida y False su ausencia.

Valor	Frecuencia	Porcentaje (%)
False	560	60.94
True	359	39.06

Tabla 11: Distribución de frecuencias de la variable exang.

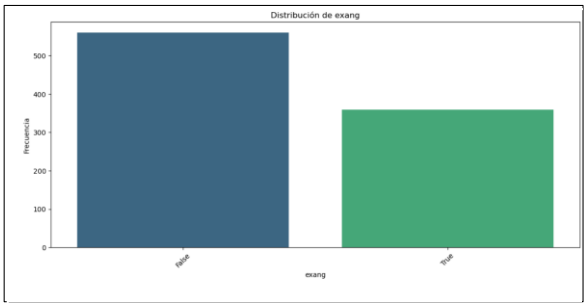

Figura 10: Distribución de la variable exang.

Tal como se muestra en la **Tabla 11**, aproximadamente el 60.94 % de los pacientes no presentaron angina durante el esfuerzo físico, mientras que el 39.06 % sí lo hicieron. Esta distribución, representada también en la **Figura 10**, refleja una ligera predominancia de casos negativos, pero con una proporción relevante de positivos que debe considerarse en los modelos predictivos.

Esta variable puede tener un impacto significativo en el diagnóstico y pronóstico de enfermedad cardíaca, ya que la aparición de síntomas durante el esfuerzo es un indicador clínico clásico de isquemia.

2.2.12. Oldpeak

La variable oldpeak representa la **depresión del segmento ST** inducida por el ejercicio en relación al estado de reposo. Esta medida es un indicador clave en pruebas de esfuerzo cardíaco, ya que una mayor depresión puede ser indicativa de isquemia miocárdica.

Estadístico	Valor
Media	0.88
Mediana	0.80
Moda	0
Mínimo	-2.60
Máximo	6.20
Rango	8.80
Desviación estándar	1.05
Varianza	1.11
Curtosis	1.43
Asimetría	1.08
Percentil 25	0.00
Percentil 50 (Mediana)	0.80
Percentil 75	1.50
IQR	1.50

Tabla 12: Estadísticos descriptivos de la variable oldpeak.

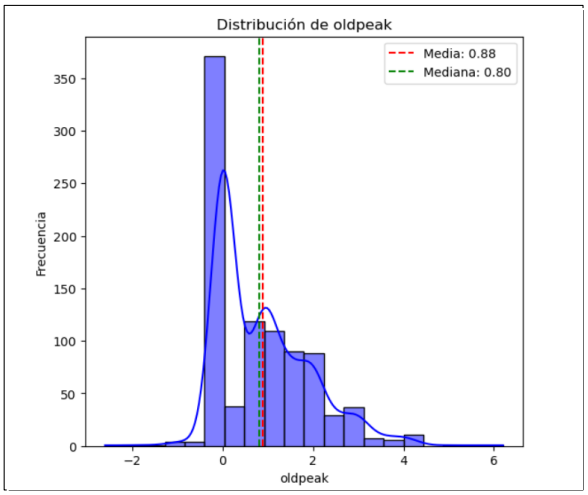

Figura 11: Distribución de la variable oldpeak.

En la **Tabla 12** se muestran los principales estadísticos descriptivos de esta variable. La media es de 0.88 y la mediana de 0.80, con una moda de 0. El rango va desde -2.6 hasta 6.2, siendo la desviación estándar de 1.05.

La distribución, la cual vemos en la **Figura 11**, presenta una **asimetría positiva** (1.08) y una curtosis de 1.43, lo que sugiere una forma levemente apuntada y sesgada hacia valores altos. Esta variable es importante en los modelos predictivos, ya que niveles elevados de oldpeak suelen asociarse con mayor riesgo de enfermedad coronaria.

2.2.13. Slope

La variable slope describe la **pendiente del segmento ST** durante el esfuerzo, un parámetro relevante en pruebas de esfuerzo para identificar anomalías cardíacas. Las categorías posibles son:

- **upsloping**: pendiente ascendente.
- **flat**: pendiente plana.
- **downsloping**: pendiente descendente.

Categoría	Frecuencia	Porcentaje (%)
flat	500	54.41
upsloping	326	35.47
downsloping	93	10.12

Tabla 13: Distribución de frecuencias de la variable slope.

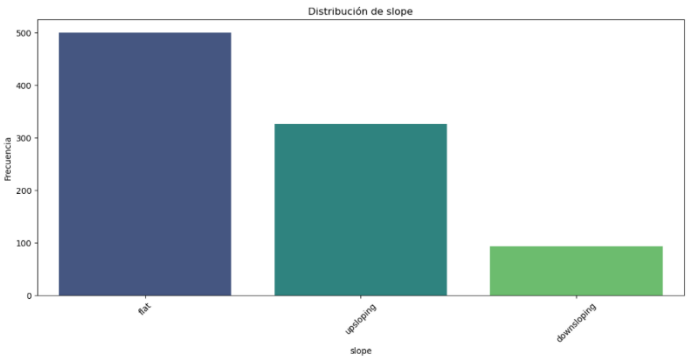

Figura 12: Distribución gráfica de la variable slope.

Tal como se observa en la **Tabla 13**, la mayoría de los pacientes presentan una pendiente flat (54.41 %), seguida de upsloping (35.47 %) y, en menor medida, downsloping (10.12 %).

La representación gráfica en la **Figura 12** permite visualizar claramente esta distribución. Esta variable es relevante en la predicción de enfermedad coronaria, dado que una pendiente descendente se asocia más frecuentemente con anomalías cardíacas.

Pendiente del segmento ST durante el esfuerzo. Puede ser upsloping, flat o downsloping, lo que puede asociarse a distintos grados de severidad.

Variable objetivo

2.2.14. Num

La variable num representa el diagnóstico de enfermedad cardíaca del paciente, con valores entre 0 y 4, donde:

- **0**: No se detecta enfermedad cardíaca.
- **1 a 4**: Distintos grados de presencia de enfermedad.

Esta variable actúa como **variable objetivo** del estudio, y su correcta interpretación y tratamiento es crucial para el rendimiento de los modelos de predicción.

Estadístico	Valor
Media	0.99
Mediana	1.00
Moda	0
Mínimo	0
Máximo	4
Rango	4
Desviación estándar	1.14
Varianza	1.30
Curtosis	-0.09
Asimetría	0.97
Percentil 25	0.00
Percentil 50 (Mediana)	1.00
Percentil 75	2.00
IQR	2.00

Tabla 14: Estadísticos descriptivos de la variable num.

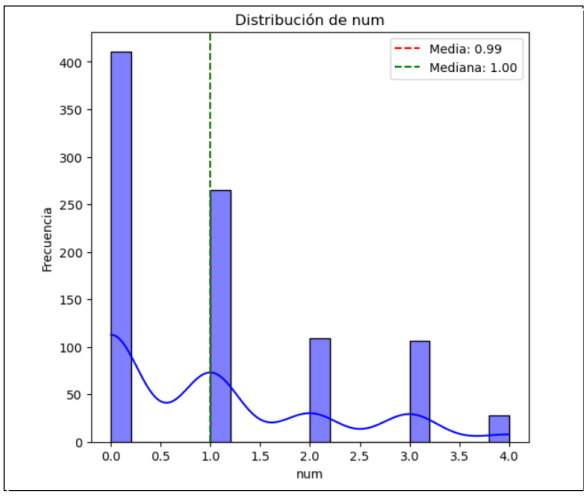

Figura 13: Distribución de la variable num.

Como se observa en la **Tabla 14** y la **Figura 13**, la distribución de clases está claramente desbalanceada, con una mayor proporción de pacientes sin enfermedad (clase 0). Este desbalance puede llevar a que los algoritmos predictivos se sesguen hacia la clase mayoritaria, disminuyendo así la sensibilidad hacia los casos con presencia de enfermedad.

Por ello, en la fase de preprocesamiento se planteará una transformación de esta variable para abordar el problema como una clasificación binaria (enfermo / no enfermo), y se aplicarán técnicas de balanceo como **SMOTE** para mejorar la capacidad generalizadora de los modelos entrenados. No obstante, con el fin de no perder información relevante durante esta simplificación, se incorporarán variables derivadas mediante técnicas de **Feature Engineering**, como una puntuación de riesgo y una etiqueta binaria adicional que permita captar diferentes niveles de gravedad. Esta estrategia híbrida permite conservar la complejidad del problema original, al tiempo que se facilita la implementación y evaluación de modelos de clasificación.

2.3. Feature Engineering

Durante esta fase del trabajo se han generado nuevas variables derivadas de las características originales del conjunto de datos, con el objetivo de mejorar la capacidad de los modelos predictivos para identificar patrones clínicamente relevantes. El proceso de feature engineering no solo enriquece la representación de los datos, sino que también permite modelar relaciones no lineales, destacar atributos clave y facilitar la interpretación de los resultados [5].

Las transformaciones más significativas aplicadas han sido las siguientes:

- **Variable binaria careful**: se ha creado una nueva característica denominada careful, la cual toma el valor 1 si la variable num (indicador de severidad de enfermedad cardíaca) es mayor o igual a 2, y 0 en caso contrario. Esta variable auxiliar permite realizar análisis complementarios centrados en la presencia de enfermedad significativa, sin alterar la naturaleza multiclase del problema original. Su incorporación facilita la generación de visualizaciones más interpretables y contribuye a una mejor comprensión de los perfiles de riesgo en los pacientes.

- **Índice de riesgo de emergencia (emergency_risk_score)**: se ha diseñado una variable compuesta que resume el nivel estimado de riesgo clínico de cada paciente. Esta variable se construyó mediante la normalización y truncamiento de varias características médicas relevantes normalizadas, tales como la edad, la presión arterial en reposo (trestbps), el nivel de colesterol sérico (chol), la frecuencia cardíaca máxima alcanzada (thalach), la depresión del segmento ST (oldpeak) y el número de vasos coloreados mediante fluoroscopia (ca). Además, se incorporaron factores binarios como el sexo del paciente, la glucemia en ayunas (fbs) y la presencia de angina inducida por el ejercicio (exang).

  Para el cálculo del score, cada variable fue transformada a una escala comparable (mediante normalización al rango [0, 1] o truncamiento en umbrales clínicos razonables) y posteriormente combinada mediante una fórmula ponderada. Los pesos asignados a cada componente se definieron de forma heurística, en función de su relevancia clínica estimada. El resultado es una métrica sintética que permite clasificar rápidamente a los pacientes según su riesgo potencial, mejorando la interpretabilidad de los datos.

Estas transformaciones permiten a los modelos posteriores trabajar con variables más limpias, informativas y balanceadas, favoreciendo una mejor generalización. Además, la creación del índice de riesgo proporciona una representación unificada de múltiples factores clínicos, lo cual resulta útil tanto en tareas de predicción como en la exploración y visualización de los datos, facilitando así la toma de decisiones médicas asistidas por datos.

2.4. Detección y tratamiento de valores atípicos: Outliers

Para asegurar la calidad del análisis, se realizó una detección de valores atípicos (outliers) en las variables numéricas. Se empleó el método del rango intercuartílico (IQR), una técnica muy robusta basada en los cuartiles de la distribución [6]. Según este criterio, se considera un valor atípico aquel que se encuentra fuera del rango definido por:

$$[Q1 - 1{,}5 \cdot IQR,\ Q3 + 1{,}5 \cdot IQR]$$

donde $Q1$ es el primer cuartil, $Q3$ el tercer cuartil e $IQR = Q3 - Q1$. Este método permite identificar valores extremos sin verse influenciado por distribuciones no normales.

Se procedió a contar la cantidad de outliers por variable con el objetivo de tomar decisiones informadas sobre su tratamiento. Dependiendo de la naturaleza de los datos y su contexto clínico, los valores atípicos pueden ser eliminados, corregidos o conservados si se considera que contienen información relevante.

Para facilitar esta evaluación, se generó una visualización que muestra la cantidad de valores atípicos por columna, como se observa en la **Figura 14**.

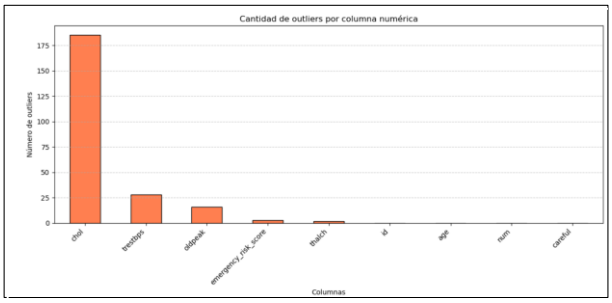

Figura 14: Cantidad de valores atípicos por variable numérica

Como puede observarse, la variable chol (colesterol sérico) presenta una gran cantidad de valores atípicos. Para investigar este comportamiento, se graficó la distribución de sus valores,

tal como se muestra en la **Figura 6**.

La figura revela una alta concentración de valores igual a 0, lo cual es clínicamente inviable, ya que un nivel de colesterol igual a cero no es compatible con la vida [7]. Esto sugiere la presencia de errores de registro o valores faltantes codificados incorrectamente. Para abordar este problema, se reemplazaron todos los valores iguales a 0 por valores nulos (NaN), y posteriormente se aplicó la función na_filler para imputar los valores faltantes mediante la media de la variable.

Un análisis similar se realizó sobre la variable trestbps (presión arterial en reposo), que también presenta una cantidad considerable de valores atípicos. Al examinar sus valores, se detectaron también registros con valor 0 (**Figura 5**), lo que, al igual que en el caso anterior, carece de validez clínica[8]. Por tanto, se aplicó la misma estrategia: los ceros fueron tratados como valores faltantes e imputados con la media correspondiente.

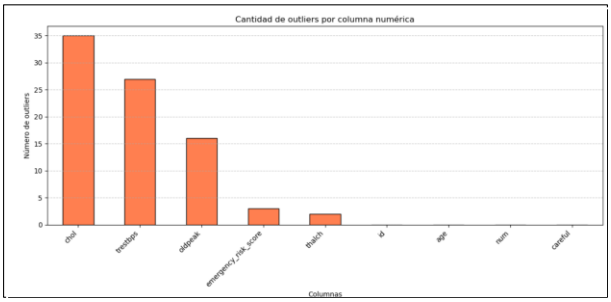

Figura 15: Cantidad de valores atípicos por variable numérica tras tratamiento

Así, obtenemos la cantidad de outliers reflejada en la **Figura 15** la cual esta reducida considerablemente. Este proceso de detección y corrección de outliers es fundamental para evitar sesgos en el modelo y mejorar la fiabilidad de los resultados obtenidos.

2.5. Codificación de variables categóricas

Con el objetivo de preparar los datos para su uso en algoritmos de aprendizaje automático, se llevó a cabo una codificación de las variables categóricas. Dado que muchos modelos no pueden trabajar directamente con datos no numéricos, fue necesario transformar estas variables en valores numéricos representativos. Para ello, se aplicó una técnica de codificación que asigna un número entero distinto a cada categoría presente en las variables de tipo categórico (Label Encoding). Esta transformación permite que los modelos puedan procesar correctamente la información sin que se pierda el significado de las categorías originales.

La elección de Label Encoding frente a otras alternativas, como One-Hot Encoding, se basó en la naturaleza de los algoritmos empleados y en la estructura de los datos. En particular, se optó por esta técnica debido a que el número de categorías por variable era reducido y no se requería una expansión significativa del espacio de características.

2.6. Matriz de correlación de variables

La matriz de correlación permite analizar las relaciones lineales entre las distintas variables numéricas del conjunto de datos. Cada celda de la matriz muestra el coeficiente de correlación de Pearson entre dos variables, cuyo valor oscila entre -1 y 1. Este coeficiente mide la intensidad y dirección de la relación lineal entre dos variables:

- Un valor cercano a 1 indica una **correlación positiva fuerte**: a medida que una variable aumenta, la otra también lo hace.
- Un valor cercano a -1 refleja una **correlación negativa fuerte**: al aumentar una variable, la otra tiende a disminuir.
- Un valor próximo a 0 indica **ausencia de relación lineal significativa** entre las variables.

En la **Figura 16** se presenta la matriz de correlación del conjunto de datos tras el preprocesamiento y la generación de nuevas variables.

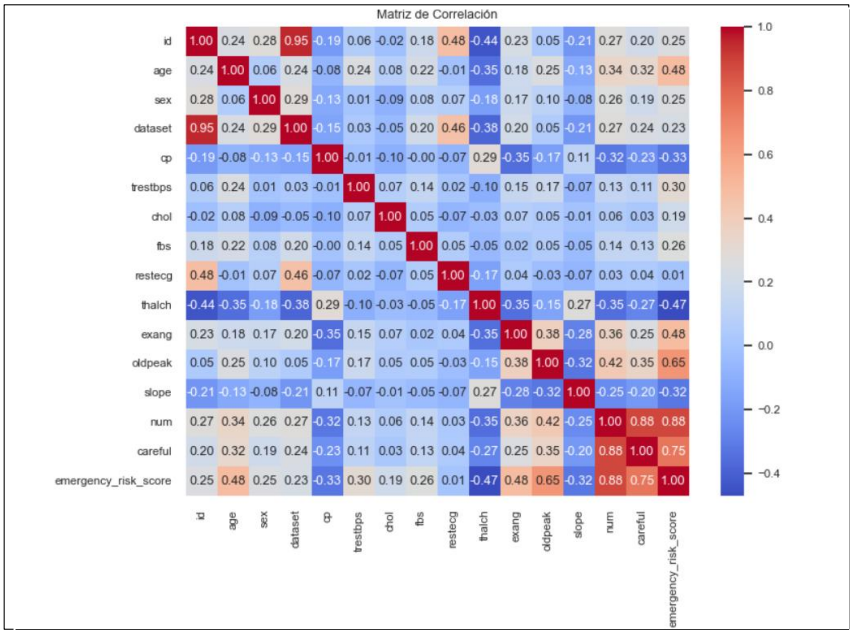

Figura 16: Matriz de correlación entre variables

Observaciones relevantes del análisis

- Las variables num, careful y emergency_risk_score presentan una **alta correlación positiva entre sí** (valores cercanos a 0,88), lo cual es consistente, dado que las dos últimas fueron derivadas en parte a partir de la variable original num, mediante técnicas de feature engineering.
- La variable oldpeak muestra una correlación positiva significativa con emergency_risk_score ($r = 0{,}65$), lo que sugiere que esta variable podría tener un papel importante como predictor en la estimación del riesgo clínico.
- Variables como thalach (frecuencia cardíaca máxima), exang (angina inducida por ejercicio) y slope (pendiente del segmento ST) también exhiben correlaciones moderadas con las variables objetivo derivadas, lo cual respalda su relevancia clínica en el diagnóstico de enfermedades cardíacas.
- Por otro lado, variables como fbs (nivel de azúcar en sangre en ayunas), chol (colesterol sérico) o restecg (resultado del electrocardiograma en reposo) muestran **correlaciones**

bajas con el resto de variables, lo que sugiere una menor influencia lineal directa, aunque no se descarta su utilidad combinada o no lineal en algunos modelos.

- Finalmente, variables como id o dataset, aunque presentan cierta correlación con otras, **no aportan información clínica directa** y no deben considerarse en el análisis predictivo. Su uso principal es técnico, y serán descartadas para el entrenamiento de modelos.

En conjunto, la matriz de correlación nos proporciona una visión inicial útil para comprender las relaciones existentes entre variables, detectar posibles redundancias y guiar la selección de características en las siguientes fases del análisis.

3

Desarrollo, Evaluación y Análisis de Modelos

3.1. División del conjunto de datos

Una vez realizado el preprocesamiento inicial, se procedió a dividir el conjunto de datos en dos subconjuntos: **entrenamiento (80 %)** y **test (20 %)**, garantizando la representatividad de las clases mediante el parámetro stratify. Esta separación permite entrenar los modelos de forma controlada y evaluar su capacidad de generalización sobre datos no vistos previamente.

3.2. Selección de variables

Para reducir la dimensionalidad del problema y seleccionar únicamente las variables más relevantes, se utilizó el método SelectKBest con la función de puntuación f_classif. Este método selecciona las k variables que presentan mayor poder predictivo con respecto a la variable objetivo.

En este caso, se seleccionaron las 10 variables más informativas, entre las que destacan age, thalach, oldpeak, exang y emergency_risk_score, entre otras. Esta selección se basa en criterios estadísticos univariantes, facilitando el entrenamiento de modelos más simples y eficientes.

3.3. Balanceo de clases con SMOTE

Dado que el conjunto de entrenamiento presenta un claro desbalance de clases, se aplicó la

técnica SMOTE (Synthetic Minority Over-sampling Technique). Este método genera nuevas instancias sintéticas de las clases minoritarias a partir de combinaciones lineales de muestra existentes, lo que permite equilibrar la distribución de clases sin pérdida de información [9].

Antes del balanceo, algunas clases contaban con muy pocas muestras en comparación con otras, como podemos ver en la **Figura 13**. Tras aplicar SMOTE, todas las clases presentes en el entrenamiento quedaron con el mismo número de instancias (329), favoreciendo así el entrenamiento de modelos más justos y con mejor capacidad de generalización.

3.4. Normalización de los datos

Finalmente, se realizó una **normalización estándar** (StandardScaler) sobre las variables seleccionadas del conjunto de entrenamiento, escalando las características para que tengan media cero y desviación típica uno. Este paso es especialmente importante para modelos sensibles a la escala de los datos, como redes neuronales o máquinas de vectores soporte (SVM) [10].

3.5. Entrenamiento de modelos

3.5.1. Modelo Support Vector Machine (SVM)

El modelo Support Vector Machine (SVM) fue entrenado mediante una búsqueda exhaustiva de hiperparámetros (Grid Search) con validación cruzada de 5 pliegues, evaluando un total de 24 combinaciones y realizando 120 ajustes. Los mejores hiperparámetros encontrados fueron:

- **C:** 20
- **gamma:** auto
- **kernel:** poly

El modelo alcanzó un valor óptimo de **F1-macro** de **0.9397** durante la validación cruzada, lo que refleja un equilibrio sólido entre precisión y recall en todas las clases. Posteriormente, el

modelo se evaluó sobre el conjunto de test, obteniendo un **accuracy** del **83.70 %**.

Clase	Precisión	Recall	F1-score	Soporte
0	0.90	0.89	0.90	82
1	0.83	0.85	0.84	53
2	0.77	0.77	0.77	22
3	0.72	0.62	0.67	21
4	0.67	1.00	0.80	6
Accuracy	0.837			
Macro Promedio	0.78	0.83	0.80	
Ponderado Promedio	0.84	0.84	0.84	

Cuadro 2: Reporte de clasificación del modelo SVM sobre el conjunto de test.

El **recall**, también conocido como exhaustividad, es una métrica fundamental en clasificación que mide la proporción de instancias positivas correctamente identificadas por el modelo respecto al total de positivos reales. Su interpretación es especialmente relevante en contextos donde los falsos negativos tienen un coste elevado, como en el diagnóstico de enfermedades.

En el reporte anterior, se observa que el modelo logra un recall perfecto (1.00) en la clase 4, pero con una precisión más baja (0.67), lo que indica que aunque identifica todos los casos reales de esa clase, también incurre en varios falsos positivos. En contraste, las clases con mayor frecuencia (como la 0 y la 1) mantienen tanto alta precisión como recall, con valores de F1-score cercanos o superiores a 0.84.

La métrica **F1-score** es especialmente útil en problemas con clases desbalanceadas, ya que representa la media armónica entre la precisión y el recall, proporcionando una visión más equilibrada del rendimiento del modelo cuando una métrica puede estar inflada a costa de la otra.

El análisis del rendimiento por clase pone en evidencia la presencia de cierto **desbalanceo de clases** en la variable objetivo, ya que algunas clases, como la 4, tienen un número significativamente menor de instancias que otras (solo 6 casos frente a 82 de la clase 0). Este desbalance puede dificultar que el modelo generalice adecuadamente para clases minoritarias.

3.5.2. Modelo Random Forest

El modelo Random Forest fue ajustado utilizando búsqueda de hiperparámetros mediante validación cruzada con 5 particiones, evaluando 96 combinaciones posibles, lo que resultó en un total de 480 ajustes. Los mejores parámetros encontrados fueron:

- **bootstrap:** False
- **max_depth:** None
- **min_samples_leaf:** 1
- **min_samples_split:** 5
- **n_estimators:** 200

El modelo obtuvo un **F1-macro** de **0.9348** durante la validación cruzada, mostrando una alta capacidad de generalización entre todas las clases. Al evaluarse sobre el conjunto de test, alcanzó un **accuracy** del **86.96 %**, lo que indica un rendimiento superior al de otros modelos previamente entrenados.

Clase	Precisión	Recall	F1-score	Soporte
0	0.94	0.89	0.91	82
1	0.84	0.91	0.87	53
2	0.85	0.77	0.81	22
3	0.72	0.86	0.78	21
4	1.00	0.67	0.80	6
Accuracy	0.870			
Macro Promedio	0.87	0.82	0.84	
Ponderado Promedio	0.88	0.87	0.87	

Cuadro 3: Reporte de clasificación del modelo Random Forest sobre el conjunto de test.

Una observación relevante es el comportamiento del modelo respecto al **recall**, métrica que evalúa la capacidad del modelo para identificar correctamente los ejemplos positivos reales.

Por ejemplo, en la clase 1, el modelo alcanza un recall del 91 %, lo que implica que prácticamente todos los casos reales de esa clase fueron correctamente detectados.

En contraste, aunque la clase 4 logra una precisión perfecta (1.00), su recall es más bajo (0.67), lo cual sugiere que el modelo fue muy conservador al clasificar esta clase: los pocos casos que predijo como clase 4 eran correctos, pero no logró identificar todos los existentes. Esto se debe, en gran parte, a que dicha clase está subrepresentada en el conjunto de datos (sólo 6 instancias), un claro ejemplo del **desbalanceo de clases**.

El **F1-score** es particularmente útil en contextos donde las clases están desbalanceadas, como en este caso. El modelo muestra un buen rendimiento global, con un F1-macro de 0.84, y un F1 ponderado de 0.87, lo que indica que mantiene un desempeño consistente incluso teniendo en cuenta la proporción desigual de las clases.

En resumen, el modelo Random Forest ha demostrado una excelente capacidad de clasificación multiclase, con un rendimiento particularmente bueno en clases frecuentes y aceptable en clases minoritarias. El desbalanceo en la variable objetivo puede generar una tendencia del modelo a favorecer las clases más frecuentes, aunque Random Forest es robusto ante este tipo de problemas

3.5.3. Modelo Decision Tree

El modelo Decision Tree fue ajustado utilizando validación cruzada con 5 particiones, evaluando 72 combinaciones de hiperparámetros, lo que dio lugar a un total de 360 ajustes. Los mejores parámetros obtenidos fueron:

- **criterion:** gini
- **max_depth:** 10
- **min_samples_leaf:** 1
- **min_samples_split:** 2

Durante la validación cruzada, se alcanzó un **F1-macro** de **0.8871**, lo cual sugiere un rendimiento razonablemente equilibrado en las distintas clases. En el conjunto de prueba, el modelo logró un **accuracy** del **79.89 %**.

Clase	Precisión	Recall	F1-score	Soporte
0	0.89	0.87	0.88	82
1	0.80	0.83	0.81	53
2	0.84	0.73	0.78	22
3	0.58	0.71	0.64	21
4	0.25	0.17	0.20	6
Accuracy	0.799			
Macro Promedio	0.67	0.66	0.66	
Ponderado Promedio	0.80	0.80	0.80	

Cuadro 4: Reporte de clasificación del modelo Decision Tree sobre el conjunto de test.

El rendimiento del modelo Decision Tree es aceptable en general, especialmente para las clases más representadas como la clase 0 y la clase 1, donde se observa un **recall** alto (0.87 y 0.83, respectivamente). Esto indica que el modelo es capaz de recuperar una buena proporción de los verdaderos positivos en estas clases, lo cual es positivo. Sin embargo, la situación cambia cuando se consideran las clases minoritarias. En particular, la clase 4, con solo 6 ejemplos en el conjunto de prueba, presenta un **recall** de apenas 0.17, lo que implica que el modelo solo identifica correctamente 1 de cada 6 casos reales de esta clase. Esto es una manifestación clara del **desbalanceo de clases**.

En este contexto, el uso del **F1-score** como métrica cobra especial relevancia, ya que considera tanto la precisión como la exhaustividad (recall). El F1-score bajo para la clase 4 (0.20) resalta la dificultad del modelo para manejar adecuadamente clases con pocos ejemplos, en contraste con los buenos resultados obtenidos en clases con mayor representación. El promedio macro del F1-score (0.66) pone de manifiesto esta desigualdad en el rendimiento, mientras que el promedio ponderado (0.80) suaviza este efecto al dar mayor peso a las clases más frecuentes.

En definitiva, el modelo Decision Tree presenta una interpretación directa y una buena capacidad para capturar relaciones en los datos, pero muestra limitaciones importantes en la clasificación de clases poco representadas.

3.5.4. Modelo Red Neuronal Artificial

Se implementó un clasificador MLPClassifier (perceptrón multicapa) con una búsqueda exhaustiva de hiperparámetros mediante validación cruzada con 5 particiones. Se evaluaron 64 combinaciones distintas, totalizando 320 entrenamientos. Los mejores parámetros encontrados fueron:

- **activation**: tanh
- **hidden_layer_sizes**: (100, 50)
- **learning_rate_init**: 0.001
- **max_iter**: 300
- **solver**: adam

El mejor valor de F1-macro obtenido durante la validación cruzada fue de **0.9415**. Posteriormente, el modelo se evaluó en el conjunto de test independiente, obteniendo los siguientes resultados:

Clase	Precisión	Recall	F1-score	Soporte
0	0.96	0.90	0.93	82
1	0.86	0.94	0.90	53
2	0.85	0.77	0.81	22
3	0.74	0.67	0.70	21
4	0.60	1.00	0.75	6
Accuracy	0.875			
Macro Promedio	0.80	0.86	0.82	
Ponderado Promedio	0.88	0.88	0.88	

Cuadro 5: Reporte de clasificación del modelo de Red Neuronal sobre el conjunto de test.

Se observa un rendimiento muy competitivo en comparación con otros modelos evaluados. Las clases mayoritarias (como la clase 0 y 1) presentan valores altos de F1-score, lo que indica un buen balance entre precisión y recall. La clase minoritaria (clase 4), aunque con un soporte muy reducido, logra un recall perfecto (1.00), aunque con menor precisión (0.60), lo cual repercute en un F1 moderado (0.75).

El **recall** es especialmente importante en el contexto médico, pues mide la capacidad del modelo para identificar correctamente los casos positivos reales, como explicamos en la sección 3.5.1 . En modelos multiclase con distribución desequilibrada como este, un recall alto en las clases con menor frecuencia puede ser crucial si estas representan mayor gravedad clínica.

Por otro lado, el **F1-score** ofrece una medida equilibrada que considera tanto la precisión como el recall. En este modelo, tanto la macro como la media ponderada del F1-score muestran un comportamiento robusto, lo cual sugiere que el clasificador está gestionando adecuadamente el desbalanceo presente en la variable objetivo num.

En definitiva, el modelo Redes Neuronales ha mostrado un desempeño sólido, destacando por

su capacidad para capturar patrones complejos en los datos. A pesar del desbalanceo entre clases, logra mantener un equilibrio razonable en las métricas de evaluación, con especial eficacia en las clases mayoritarias y un comportamiento sorprendentemente bueno en las minoritarias. Esto sugiere que, aunque más complejo y con menor interpretabilidad directa que otros modelos, su aplicación puede ser especialmente útil en contextos donde se prioriza el rendimiento predictivo.

3.5.5. Modelo XGBoost

Se implementó un clasificador basado en el algoritmo XGBoost, conocido por su eficiencia, capacidad de generalización y alto rendimiento en tareas de clasificación. Se realizó una búsqueda exhaustiva de hiperparámetros utilizando validación cruzada con 5 particiones. En total, se exploraron múltiples combinaciones hasta encontrar los parámetros óptimos, que resultaron ser:

- **colsample_bytree**: 1.0
- **learning_rate**: 0.1
- **max_depth**: 5
- **n_estimators**: 750
- **subsample**: 0.8

El mejor valor de F1-macro obtenido durante la validación cruzada fue de **0.9409**. Posteriormente, el modelo se evaluó en el conjunto de test independiente, obteniendo los siguientes resultados:

Clase	Precisión	Recall	F1-score	Soporte
0	0.91	0.88	0.89	82
1	0.82	0.87	0.84	53
2	0.89	0.73	0.80	22
3	0.73	0.90	0.81	21
4	1.00	0.83	0.91	6
Accuracy	0.859			
Macro Promedio	0.87	0.84	0.85	
Ponderado Promedio	0.87	0.86	0.86	

Cuadro 6: Reporte de clasificación del modelo XGBoost sobre el conjunto de test.

El modelo XGBoost demuestra un rendimiento competitivo, con una precisión y F1-score elevados especialmente en las clases más representadas (clases 0 y 1), lo cual refuerza su capacidad para identificar correctamente patrones frecuentes. Destaca también el excelente desempeño en la clase minoritaria (clase 4), con una precisión perfecta (1.00) y un F1-score de 0.91, lo que sugiere una buena capacidad para no solo detectar estos casos, sino hacerlo con alto grado de certeza.

Al igual que en otros modelos, el **recall** cobra especial relevancia en aplicaciones médicas, donde la identificación de todos los casos positivos es crítica. En este sentido, el modelo mantiene valores sólidos de recall en la mayoría de clases, incluyendo las menos frecuentes (como la clase 3, con un recall de 0.90).

La media **F1-score** tanto macro como ponderada indica un equilibrio adecuado entre clases, incluso ante el desbalance presente en la variable objetivo. Esto evidencia la robustez del modelo, que ha logrado una buena generalización sin sobreajustarse a las clases mayoritarias. XGBoost es especialmente adecuado para problemas con clases desbalanceadas debido a su compatibilidad con funciones de pérdida personalizadas y métricas centradas en recall o F1-score. Estas características permiten al modelo mantener un buen rendimiento incluso en clases minoritarias, como se observa en los resultados obtenidos.

En resumen, el modelo XGBoost ha alcanzado resultados muy satisfactorios, ofreciendo un

equilibrio excelente entre precisión, recall y capacidad de generalización. Su rendimiento en las clases minoritarias y su precisión general lo posicionan como una de las alternativas más eficaces para este problema, particularmente en entornos donde la exactitud del diagnóstico resulta esencial.

3.6. Análisis de Importancia de Características mediante SHAP

Con el objetivo de interpretar el comportamiento del modelo de clasificación multiclase entrenado, se ha empleado el método SHAP (SHapley Additive exPlanations), una técnica basada en la teoría de juegos que permite descomponer las predicciones del modelo en contribuciones aditivas asociadas a cada variable de entrada [11]. Este enfoque resulta especialmente valioso en contextos donde la interpretabilidad del modelo es un requisito esencial, como ocurre en el ámbito clínico.

3.6.1. Modelo Support Vector Machine (SVM

Las Figuras 17 a 21 muestran la importancia media de las características para las clases 0 a 4, respectivamente, basada en los valores absolutos medios de SHAP obtenidos para el modelo SVM.

Estos gráficos permiten interpretar cómo las diferentes variables contribuyen a las decisiones del modelo para cada clase.

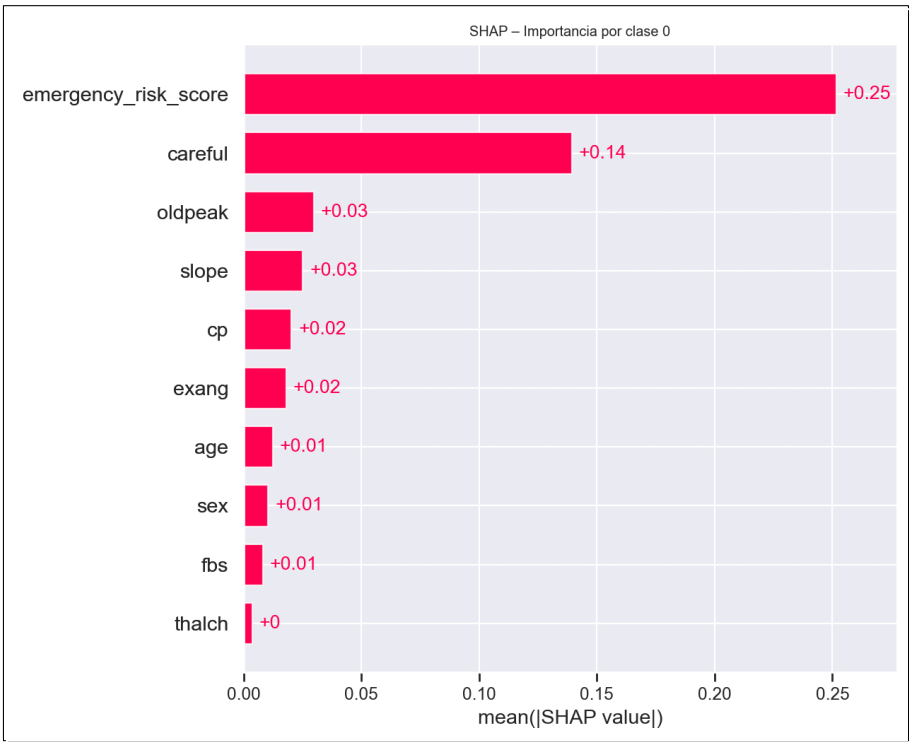

Figura 17: Importancia media de las características para la Clase 0, según los valores SHAP con SVM.

En la **Figura 17**, la variable emergency_risk_score (+0.25) aparece como la más relevante, seguida por careful (+0.14), lo cual destaca el valor predictivo de las variables sintéticas creadas mediante feature engineering. Variables como oldpeak, slope y cp también contribuyen a la predicción, aunque con menor peso.

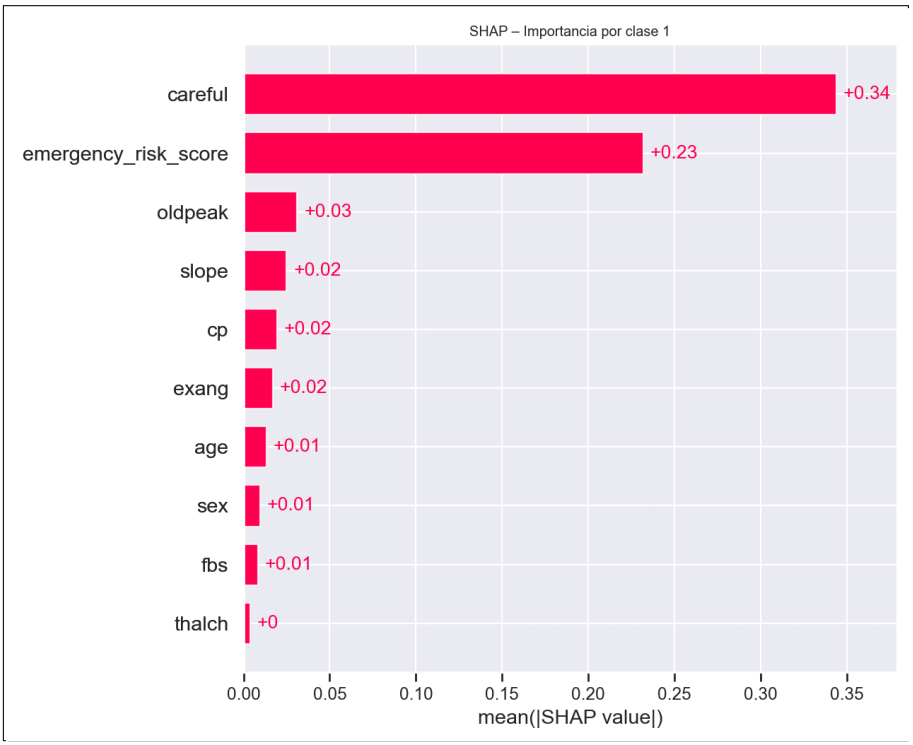

Figura 18: Importancia media de las características para la Clase 1, según los valores SHAP con SVM.

La **Figura 18** revela que careful (+0.34) y emergency_risk_score (+0.23) son nuevamente las variables más influyentes, lo que refuerza su relevancia en los perfiles de riesgo intermedio. El resto de las variables presentan contribuciones mucho más bajas, destacando ligeramente oldpeak (+0.03) y slope (+0.02).

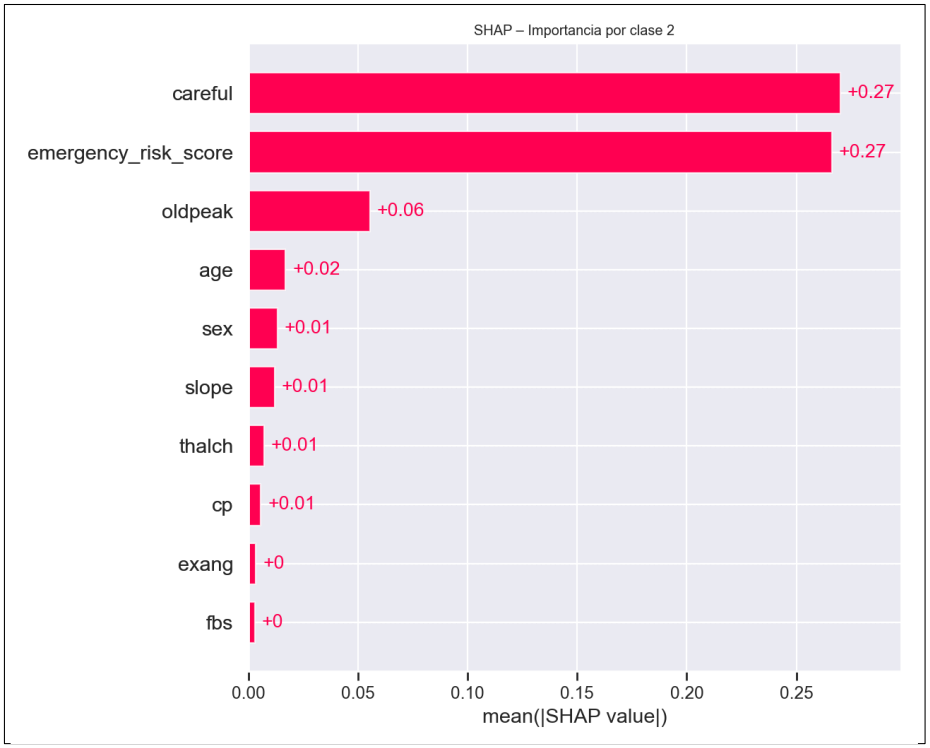

Figura 19: Importancia media de las características para la Clase 2, según los valores SHAP con SVM.

La **Figura 19** muestra un equilibrio entre careful (+0.27) y emergency_risk_score (+0.27) como principales variables. Oldpeak (+0.06) destaca por encima de otras variables fisiológicas como age, sex o thalach, cuyas contribuciones son moderadas o bajas.

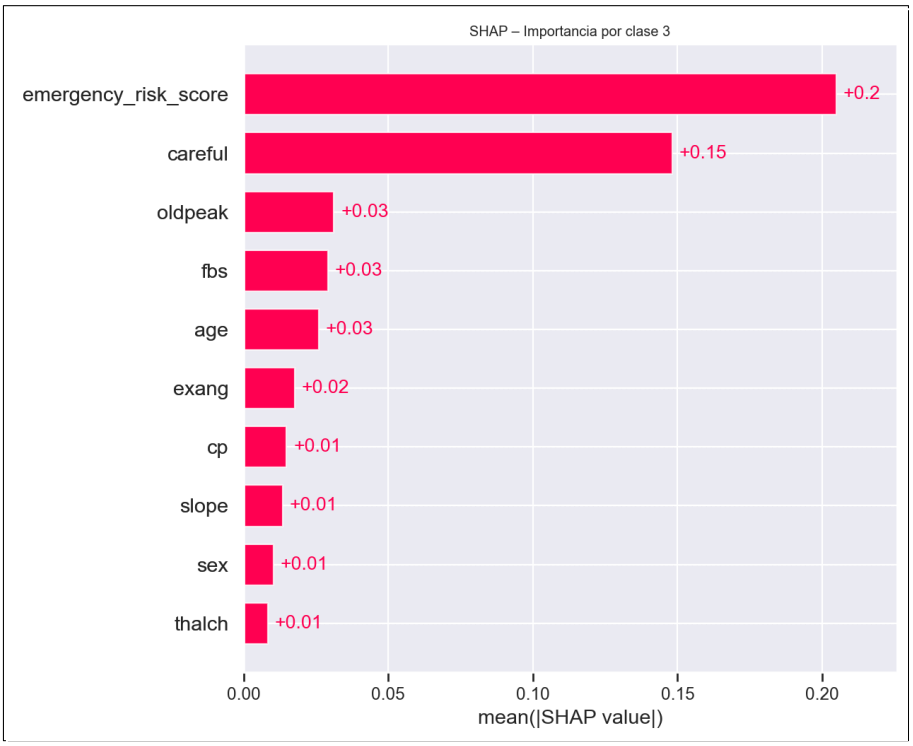

Figura 20: Importancia media de las características para la Clase 3, según los valores SHAP con SVM.

En la **Figura 20**, emergency_risk_score lidera con +0.20, seguida por careful (+0.15) y oldpeak (+0.03). Este patrón parece indicar una mayor dependencia del modelo hacia variables que integran múltiples factores clínicos, más que en variables estructurales como cp o thalach.

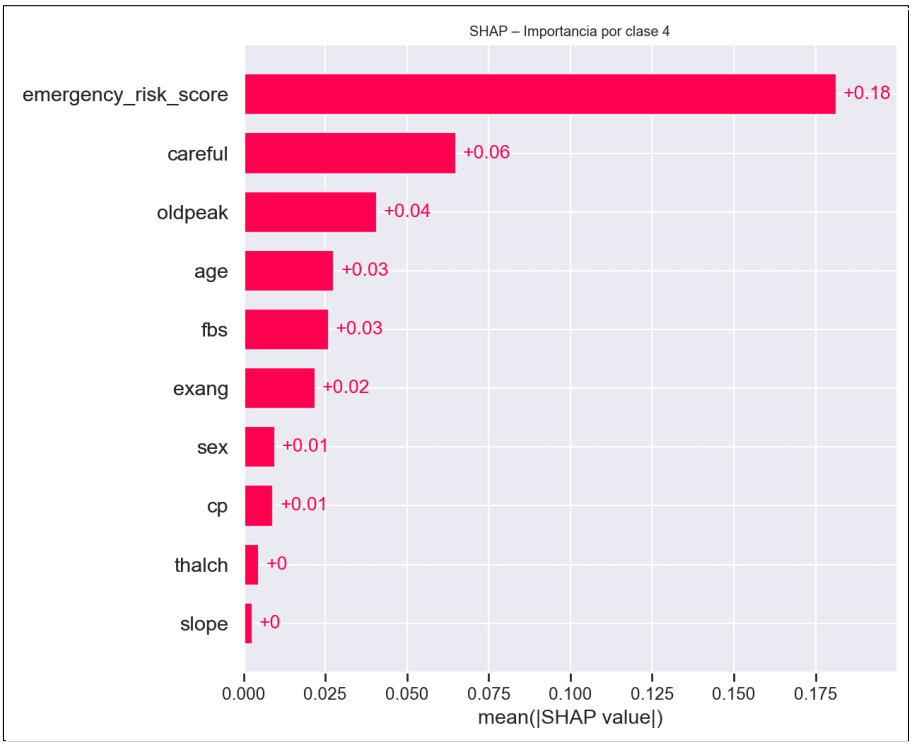

Figura 21: Importancia media de las características para la Clase 4, según los valores SHAP con SVM.

En la **Figura 21**, la variable emergency_risk_score (+0.18) mantiene su protagonismo, mientras que careful desciende a +0.06. Aquí también ganan algo de peso variables como age, fbs y oldpeak, sugiriendo que el perfil de riesgo en esta clase está más vinculado con factores de salud crónicos y edad.

En el caso concreto del modelo SVM, la variable careful muestra una fuerte contribución en las Clases 1 y 2, lo que refuerza su papel como marcador anatómico compuesto, capaz de captar patrones clínicos vinculados a una afectación coronaria significativa. Este comportamiento sugiere que el modelo es sensible a fenotipos clínicos isquémicos, donde la carga aterosclerótica y los síntomas asociados se expresan de manera más evidente.

A medida que se avanza hacia las Clases 3 y 4, la importancia de careful disminuye de forma considerable, mientras que emergency_risk_score cobra mayor protagonismo. Este cambio en la relevancia de las variables sugiere una transición del modelo hacia una interpretación más centrada en el riesgo funcional y clínico global, posiblemente reflejando escenarios como los síndromes

coronarios sin elevación del ST, la disfunción microvascular o los pacientes con comorbilidades múltiples, donde la anatomía coronaria ya no es el principal factor discriminante.

De forma coherente con esta interpretación, variables clínicas clásicas como oldpeak, thalach y age mantienen una importancia intermedia pero constante en varias clases, lo que avala su utilidad como predictores de riesgo cardiovascular. En particular, oldpeak (descenso del segmento ST inducido por ejercicio) y thalach (frecuencia cardíaca máxima alcanzada) aparecen como marcadores fisiológicos de respuesta al esfuerzo, mientras que age refuerza su papel como indicador de riesgo basal, especialmente en los extremos del espectro clínico.

Por el contrario, variables como sex, slope o exang presentan un impacto mucho menor en la explicación del modelo, lo cual podría deberse a su limitada variabilidad dentro del conjunto de datos o a su solapamiento con otras variables más complejas. Este hallazgo refuerza la idea de que el modelo SVM, incluso sin una estructura explícitamente jerárquica, logra identificar patrones clínicos relevantes y priorizar la información útil en función de la clase objetivo.

3.6.2. Modelo Random Forest

Las Figuras 22 a 26 presentan la importancia media de las características para las clases 1 a 4, respectivamente, según los valores absolutos medios de SHAP obtenidos con el modelo Random Forest. En estos gráficos de barras horizontales, las variables se encuentran ordenadas por su contribución promedioa la predicción de cada clase.

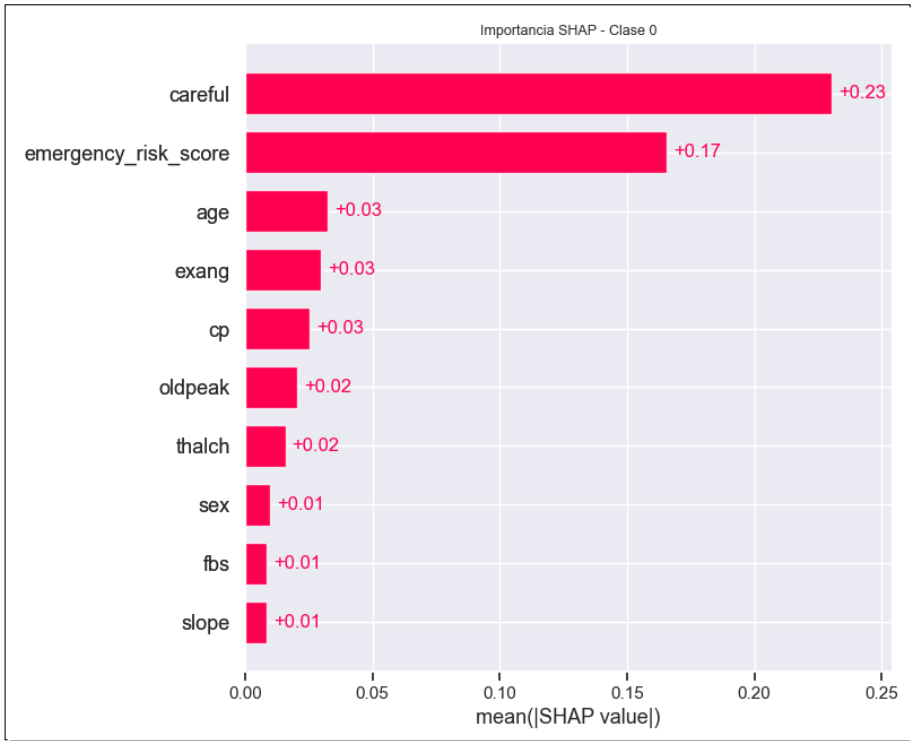

Figura 22: Importancia media de las características para la Clase 0, según los valores SHAP con Random Forest.

En la **Figura 22**, correspondiente a la Clase 0, las variables careful (+0.23) y emergency_risk_score (+0.17) destacan claramente como las de mayor importancia en la explicación del modelo. A considerable distancia se encuentran age, exang y cp, todas con valores SHAP medios de aproximadamente +0.03. Otras variables como oldpeak y thalach presentan una contribución moderada (+0.02), mientras que sex, fbs y slope muestran una importancia menor (+0.01). Estos resultados sugieren que, para la predicción de la Clase 0, el modelo se apoya principalmente en variables derivadas y de riesgo

funcional, manteniendo un peso reducido para el resto de las características clásicas.

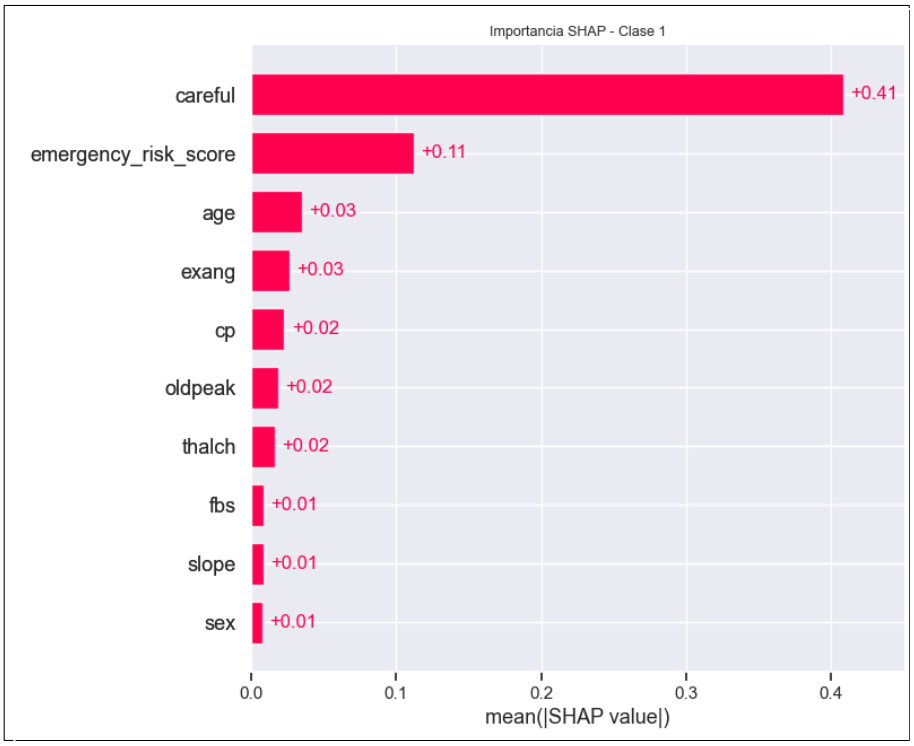

Figura 23: Importancia media de las características para la Clase 1, según los valores SHAP con Random Forest.

En la **Figura 23**, correspondiente a la Clase 1, las variables careful y emergency_risk_score dominan la explicación del modelo, con valores SHAP medios de aproximadamente +2.85 y +1.94, respectivamente. A mayor distancia aparecen age, exang y cp, lo que sugiere que el modelo identifica componentes anatómicos y de riesgo funcional como clave en esta clase.

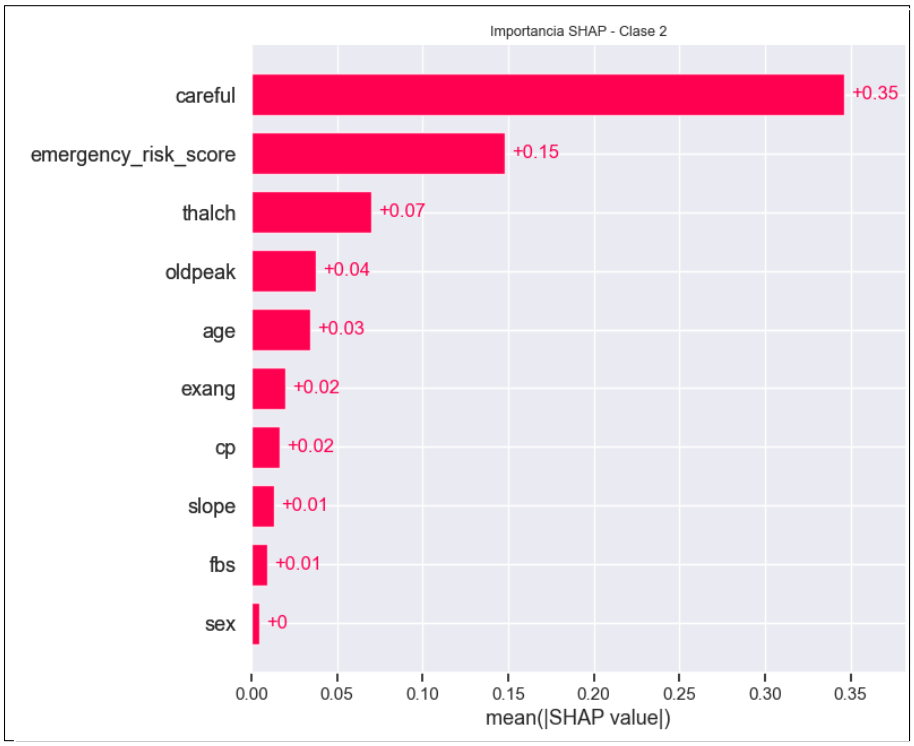

Figura 24: Importancia media de las características para la Clase 2, según los valores SHAP con Random Forest.

En la **Figura 24**, careful (+2.43) y emergency_risk_score (+1.76) continúan siendo las variables más relevantes. Se incorpora thalach (+1.25) como tercera variable en importancia, lo que refuerza la idea de un fenotipo clínico más ligado a la capacidad funcional cardíaca.

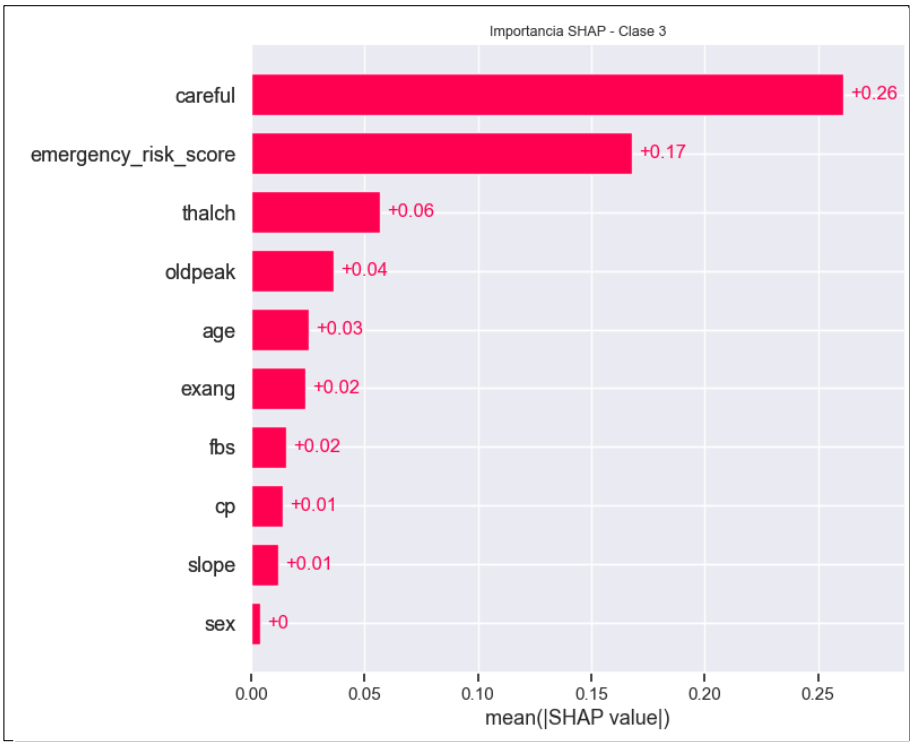

Figura 25: Importancia media de las características para la Clase 3, según los valores SHAP con Random Forest.

En la **Figura 25** la variable emergency_risk_score toma el liderazgo (+3.88), seguida por careful (+1.29) y oldpeak (+1.04). Este cambio en el orden sugiere un perfil clínico menos anatómico y más funcional en la predicción de esta clase.

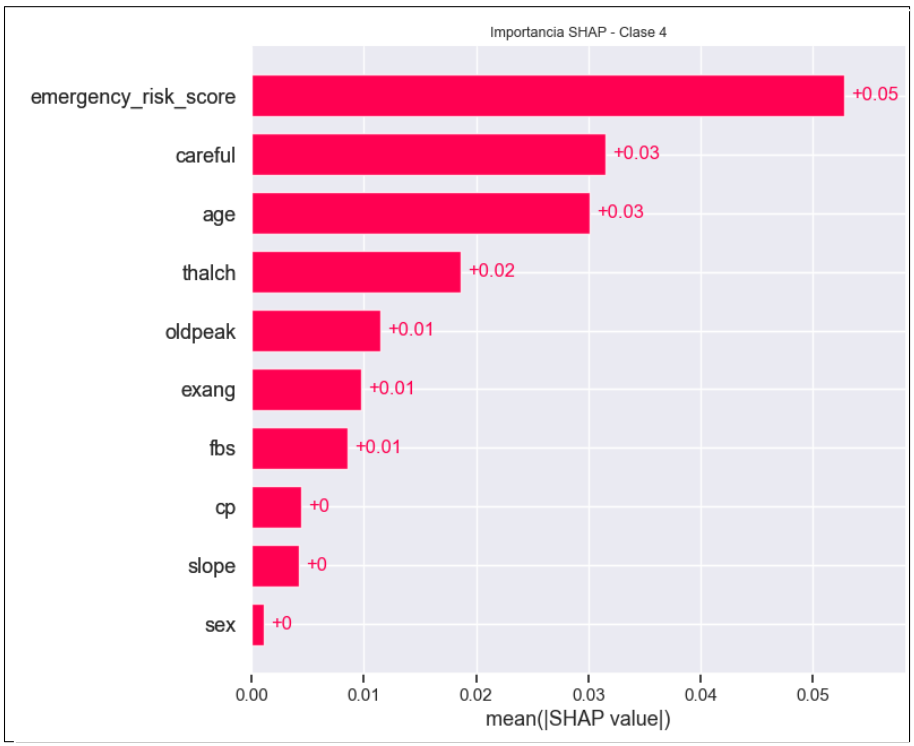

Figura 26: Importancia media de las características para la Clase 4, según los valores SHAP con Random Forest.

En la **Figura 26**, se observa que emergency_risk_score (+2.34) y age (+1.68) lideran la contribución explicativa del modelo. Careful disminuye notablemente su peso hasta +0.19, lo que podría reflejar que en esta clase final la afectación anatómica pierde protagonismo frente a variables funcionales y de riesgo basal.

Los resultados del Random Forest confirman el papel central de las variables derivadas careful y emergency_risk_score como principales motores explicativos del modelo, aunque su peso relativo varía entre clases. El aumento de la importancia de age en clases avanzadas podría estar asociado con patrones clínicos propios de pacientes mayores, mientras que variables clásicas como thalach y oldpeak mantienen una presencia relevante en varias clases, apuntalando la validez del modelo desde una perspectiva fisiopatológica.

3.6.3. Modelo Decision Tree

Las **Figuras 27** a **31** muestran la importancia media de las características para las Clases 1 a 3, respectivamente, según los valores absolutos medios de SHAP obtenidos con el modelo Decision Tree. En estos gráficos de barras horizontales, las variables están ordenadas de mayor a menor según su contribución promedio a la predicción de cada clase.

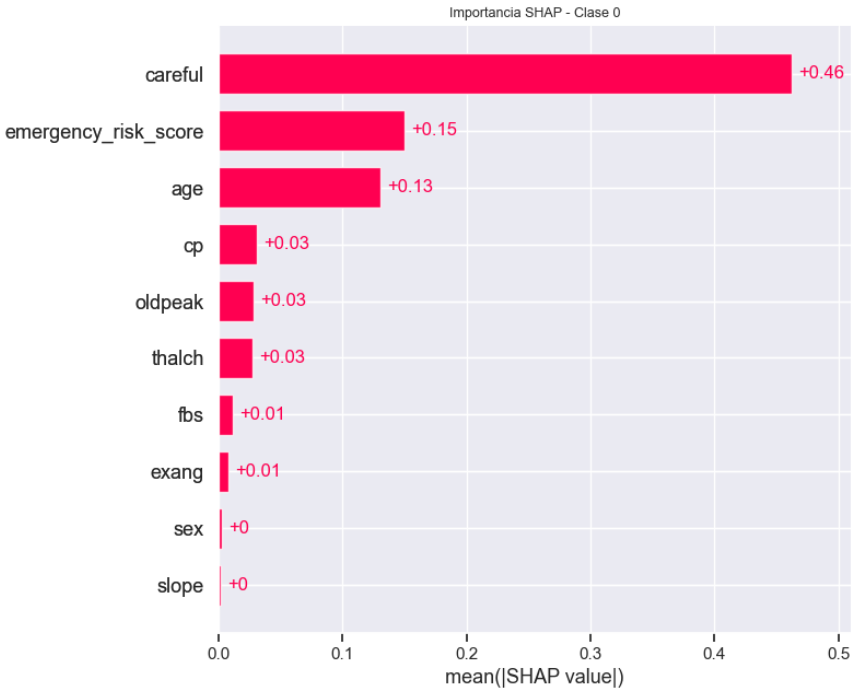

Figura 27**:** Importancia media de las características para la Clase 0, según los valores SHAP con Decision Tree.

En la **Figura 27,** correspondiente a la Clase 0, la variable careful destaca como la más relevante para el modelo, con una contribución media de +0,46. Le siguen emergency_risk_score (+0,15) y age (+0,13), lo que indica que tanto las variables derivadas como la edad del paciente tienen un papel clave en la predicción de esta clase. Otras variables como cp, oldpeak y thalach presentan una importancia moderada (+0,03), mientras que fbs y exang tienen un peso menor (+0,01). Variables como sex y slope no presentan impacto significativo (valor SHAP medio = 0). Este patrón sugiere que incluso en pacientes sin manifestaciones graves, el modelo se apoya en indicadores de riesgo funcional y compuestos para clasificar correctamente.

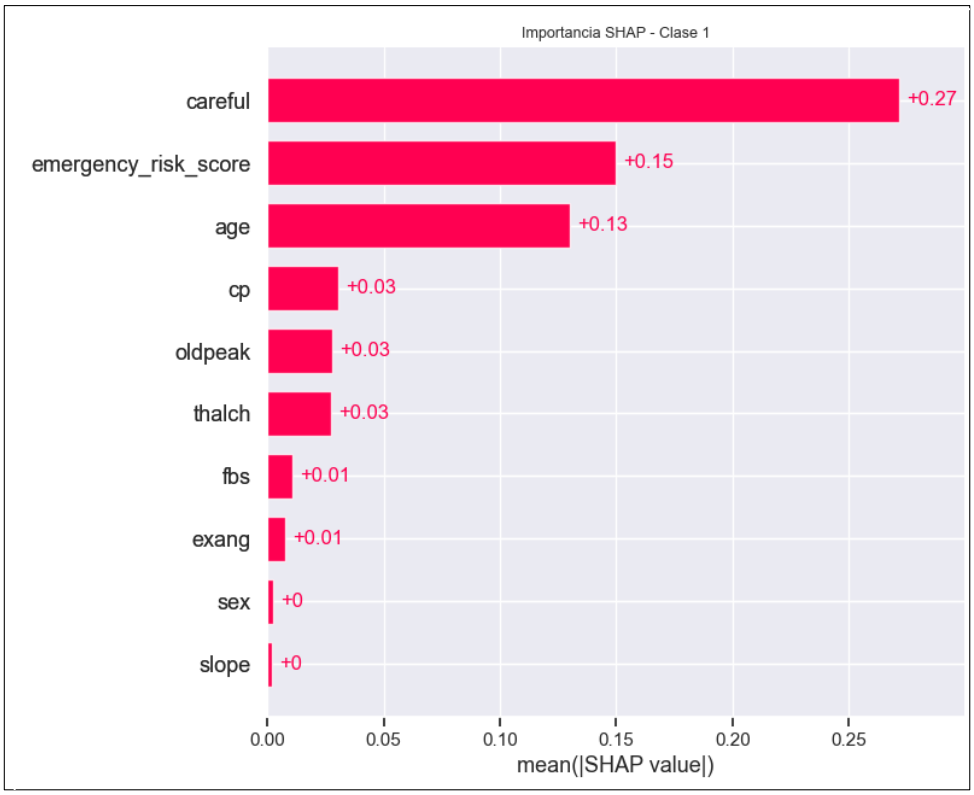

Figura 28: Importancia media de las características para la Clase 1, según los valores SHAP con Decision Tree.

En la **Figura 28**, correspondiente a la Clase 1, las variables careful (0.27), emergency_risk_score (0.15) y age (0.13) destacan como las más influyentes. Les siguen cp, oldpeak y thalach, todas con una contribución media aproximada de 0.03. En contraste, variables como fbs y exang tienen una influencia muy baja (0.01), y otras como sex y slope no presentan una contribución significativa.

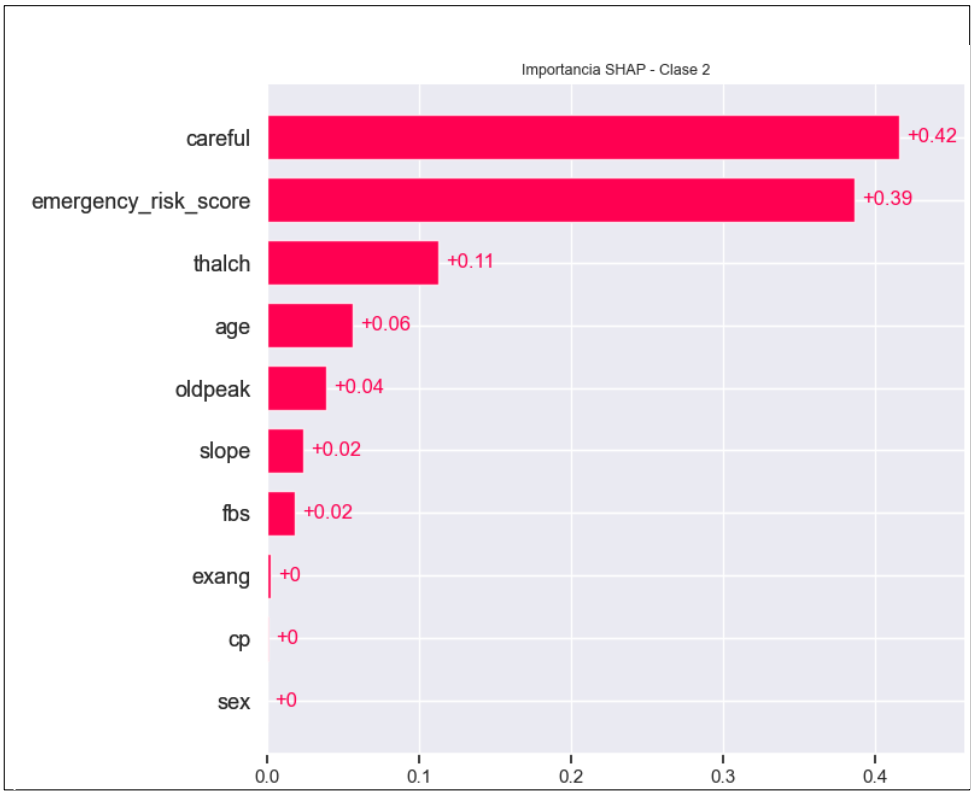

Figura 29: Importancia media de las características para la Clase 2, según los valores SHAP con Decision Tree.

La **Figura 29** muestra que para la Clase 2, las variables más relevantes son nuevamente careful (0.42) y emergency_risk_score (0.39), seguidas de thalach (0.11). También destacan age (0.06), oldpeak (0.04) y slope (0.02), aunque con menor peso. El resto de las variables tienen contribuciones cercanas a cero.

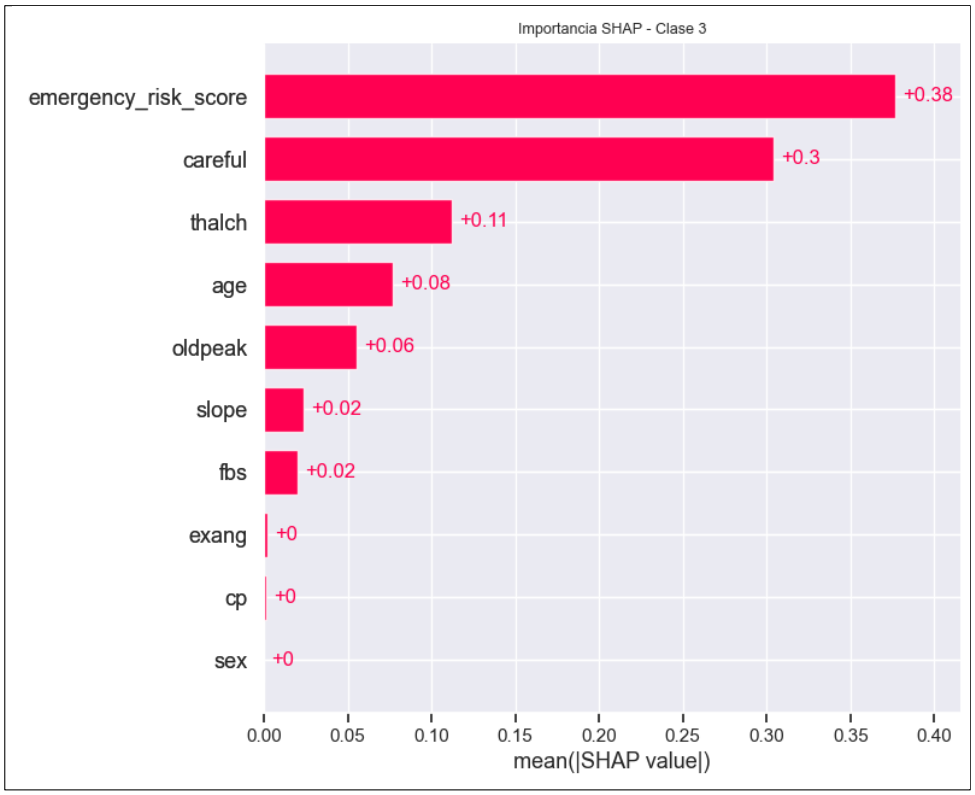

Figura 30: Importancia media de las características para la Clase 3, según los valores SHAP con Decision Tree.

En la **Figura 30**, correspondiente a la Clase 3, emergency_risk_score (0.38), careful (0.30) y thalach (0.11) son las variables más relevantes. También se observa una influencia moderada de age (0.08) y oldpeak (0.06), mientras que variables como slope y fbs tienen menor impacto (0.02). Otras variables, como cp, exang y sex, no aportan significativamente.

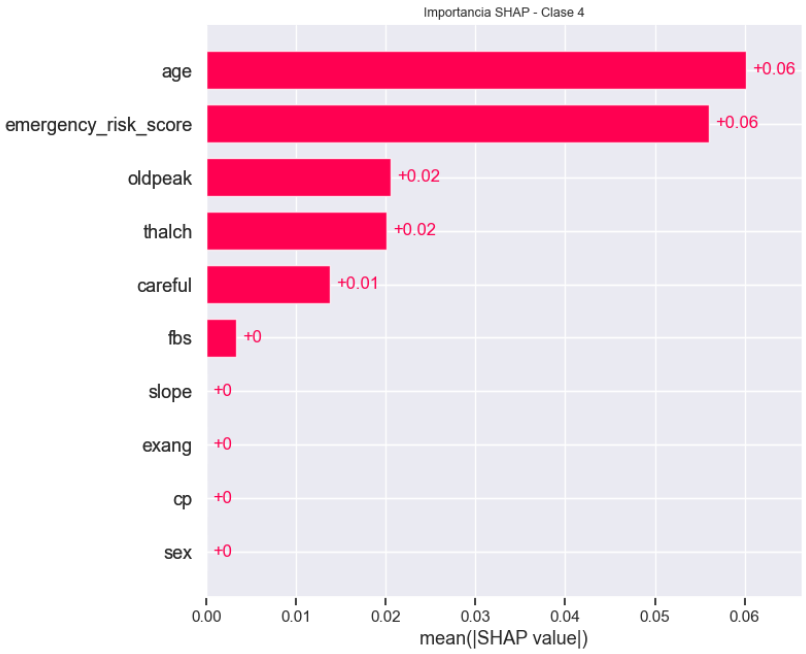

Figura 31: Importancia media de las características para la Clase 4, según los valores SHAP con Decision Tree.

La **Figura 31** muestra que, para la Clase 4, el modelo cambia su foco explicativo: la variable más relevante pasa a ser age (+0,06), seguida de emergency_risk_score con un valor SHAP similar. A continuación, aparecen oldpeak y thalach (ambas +0,02), lo que indica una mayor influencia de factores fisiológicos relacionados con la respuesta al esfuerzo. La variable careful pierde gran parte de su protagonismo, y variables como cp, sex o slope dejan de ser relevantes (valor SHAP medio = 0). En esta clase, asociada a perfiles más crónicos o avanzados, el modelo se apoya en variables de tipo funcional y demográfico, más que en indicadores anatómicos.

3.6.4. Modelo Red Neuronal Artificial

Las **Figuras 32** y **36** muestran la importancia media de las características para las clases 0 y 1, respectivamente, de acuerdo con los valores absolutos medios de SHAP obtenidos a partir del modelo MLPClassifier. Las variables se ordenan de forma descendente según su contribución promedio a la predicción de cada clase.

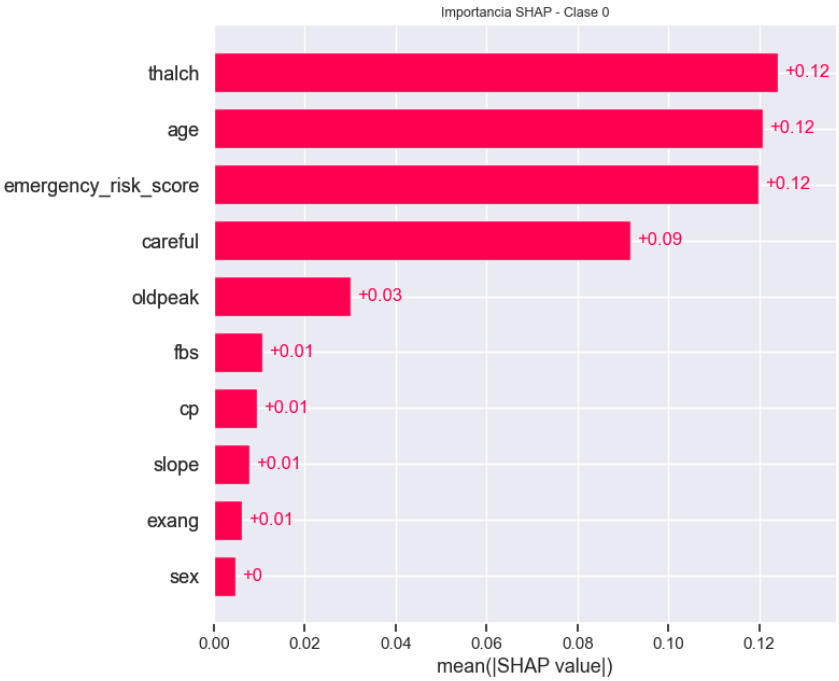

Figura 32: Importancia media de las características para la Clase 0, según los valores SHAP con Red Neuronal

En la **Figura 32**, correspondiente a la clase 0 (pacientes sin enfermedad), las variables thalach, age y emergency_risk_score comparten un nivel similar de relevancia, con valores SHAP medios cercanos a +0.12. Este triple empate en el liderazgo sugiere que el modelo no se apoya en un único marcador determinante, sino en una combinación equilibrada de factores fisiológicos (frecuencia cardíaca máxima), demográficos (edad) y compuestos (riesgo clínico estimado). A continuación, careful también destaca con una contribución de +0.09, reflejando que incluso en pacientes sanos, los patrones de afectación anatómica pueden tener cierto peso. Variables clásicas como oldpeak y fbs mantienen una influencia moderada, mientras que otras como sex apenas presentan impacto en esta clase.

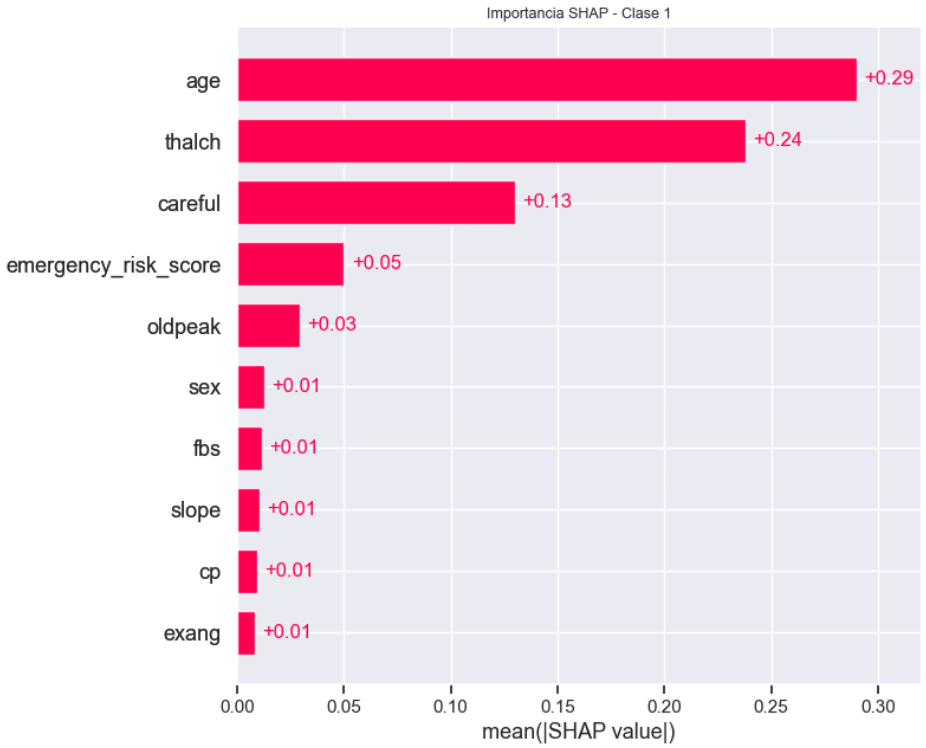

Figura 33: Importancia media de las características para la Clase 1, según los valores SHAP con Red Neuronal

En la **Figura 33**, correspondiente a la Clase 1 (primer nivel de presencia de enfermedad), se aprecia un cambio significativo en la distribución de importancia. La variable age destaca como la más relevante con un valor SHAP medio de +0.29, seguida por thalach (+0.24) y careful (+0.13). Este patrón refuerza la relevancia de la edad como factor de riesgo basal, y de thalach como indicador de reserva funcional cardiovascular. La menor contribución de emergency_risk_score (+0.05) en esta clase indica que en estadios tempranos de la enfermedad, el modelo da más peso a variables individuales y menos a métricas compuestas. Las variables oldpeak, sex, fbs y cp muestran una participación baja, aunque no despreciable, lo cual indica que el modelo considera un amplio abanico de señales clínicas, incluso cuando su efecto individual es reducido.

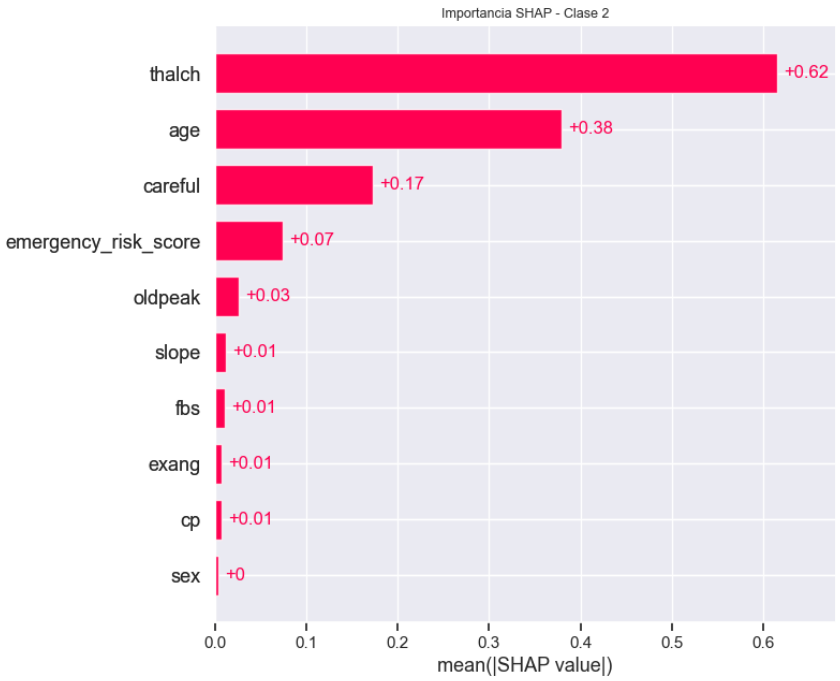

Figura 34: Importancia media de las características para la Clase 2, según los valores SHAP con Red Neuronal Artificial

En la **Figura 34**, correspondiente a la Clase 2, la variable thalach (frecuencia cardíaca máxima alcanzada) se sitúa como la más relevante (+0,62), seguida por age (+0,38) y careful (+0,17). La variable emergency_risk_score, que en otros contextos había mostrado una fuerte influencia, tiene en este caso un impacto más discreto (+0,07).

Este patrón sugiere que, para la Red Neuronal, los factores relacionados con la respuesta funcional al esfuerzo (como thalach) y la edad son determinantes clave para identificar esta clase. La variable careful mantiene presencia, lo que refuerza su papel como marcador anatómico útil en contextos isquémicos. En contraste, variables clásicas como slope, fbs, exang, cp y sex presentan una contribución marginal.

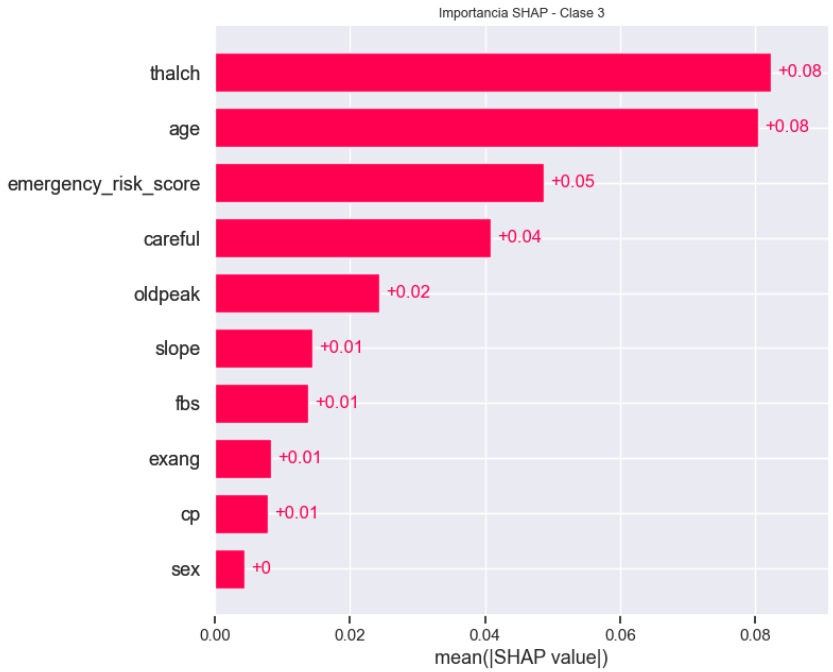

Figura 35: Importancia media de las características para la Clase 3, según los valores SHAP con Red Neuronal Artificial

En la **Figura 35**, correspondiente a la Clase 3, se observa una distribución más homogénea de las contribuciones. thalach y age comparten el primer lugar con una contribución media de +0,08, seguidos de cerca por emergency_risk_score (+0,05) y careful (+0,04). A partir de ahí, las demás variables tienen un impacto limitado (≤ +0,02).

Este perfil indica que el modelo de Red Neuronal emplea una combinación equilibrada de marcadores funcionales, factores de riesgo clínico compuestos y variables anatómicas para definir esta clase. La pérdida de protagonismo de careful frente a otras clases sugiere que esta categoría podría representar fenotipos más heterogéneos o menos dependientes de la afectación coronaria estructural directa.

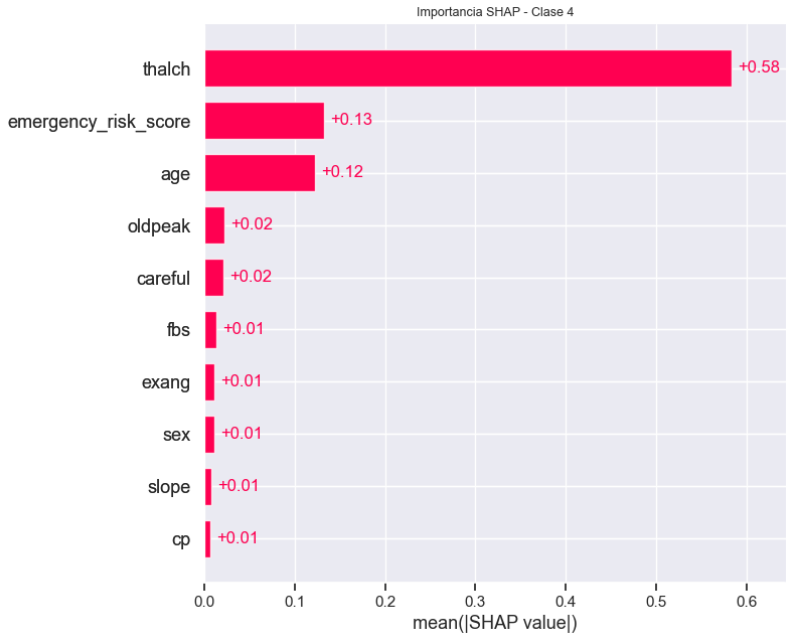

Figura 36: Importancia media de las características para la Clase 4, según los valores SHAP con Red Neuronal Artificial

Finalmente, en la **Figura 36**, correspondiente a la Clase 4, la variable más importante vuelve a ser thalach (+0,58), consolidándose como una de las principales variables explicativas en las clases más severas. Le siguen emergency_risk_score (+0,13) y age (+0,12), lo que sugiere una combinación de capacidad funcional y riesgo basal como principales motores de predicción.

En este caso, la variable careful prácticamente desaparece del ranking (valor SHAP medio = +0,02), lo que coincide con los hallazgos en modelos anteriores para esta misma clase. Esto refuerza la hipótesis de que, en escenarios clínicos avanzados (como los de Clase 4), los marcadores anatómicos pierden relevancia, y las predicciones se basan más en factores clínicos globales y funcionales.

3.6.5. Modelo XGBoost

Las **Figuras 37** a **41** muestran la importancia media de las características seleccionadas para las clases 1 a 4, respectivamente, de acuerdo con los valores absolutos medios de SHAP. En dichos gráficos de barras horizontales, las variables se ordenan según su contribución promedio a la predicción de cada clase.

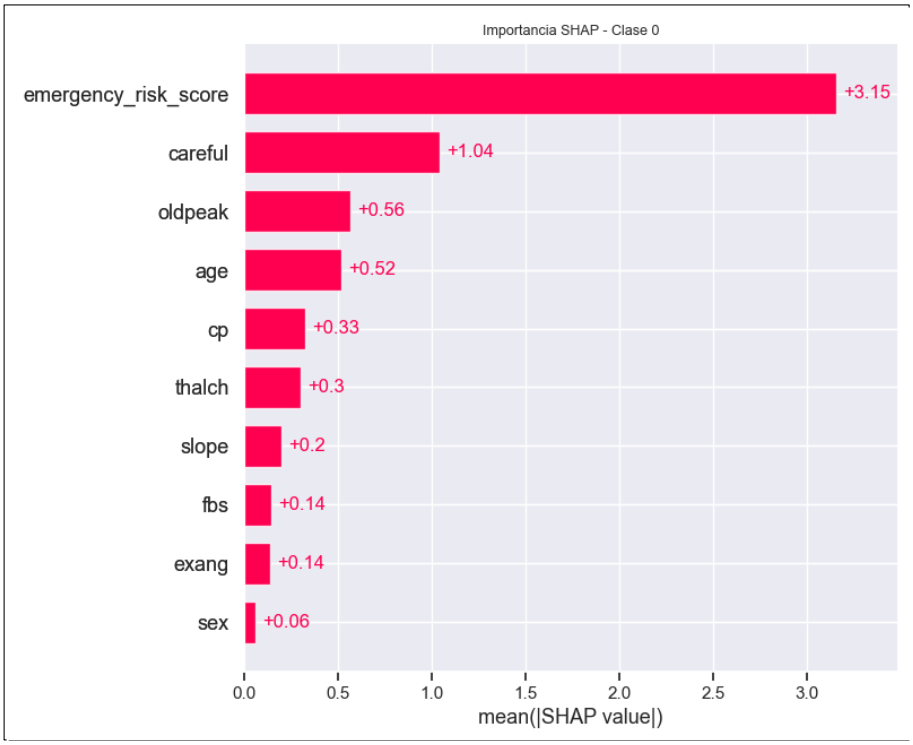

Figura 37: Importancia media de las características para la Clase 0, según los valores SHAP con XGBoost.

En la **Figura 37**, correspondiente a la Clase 0, se observa que la variable emergency_risk_score presenta una contribución media de +3,15, lo que la sitúa como el principal factor explicativo del modelo en esta clase. Le sigue careful, con un valor SHAP medio de +1,04, manteniéndose como una variable derivada de gran relevancia. A continuación, se encuentran oldpeak (+0,56) y age (+0,52), lo cual refleja la influencia de factores tanto funcionales como demográficos en la clasificación, aunque de bastante menor influencia.

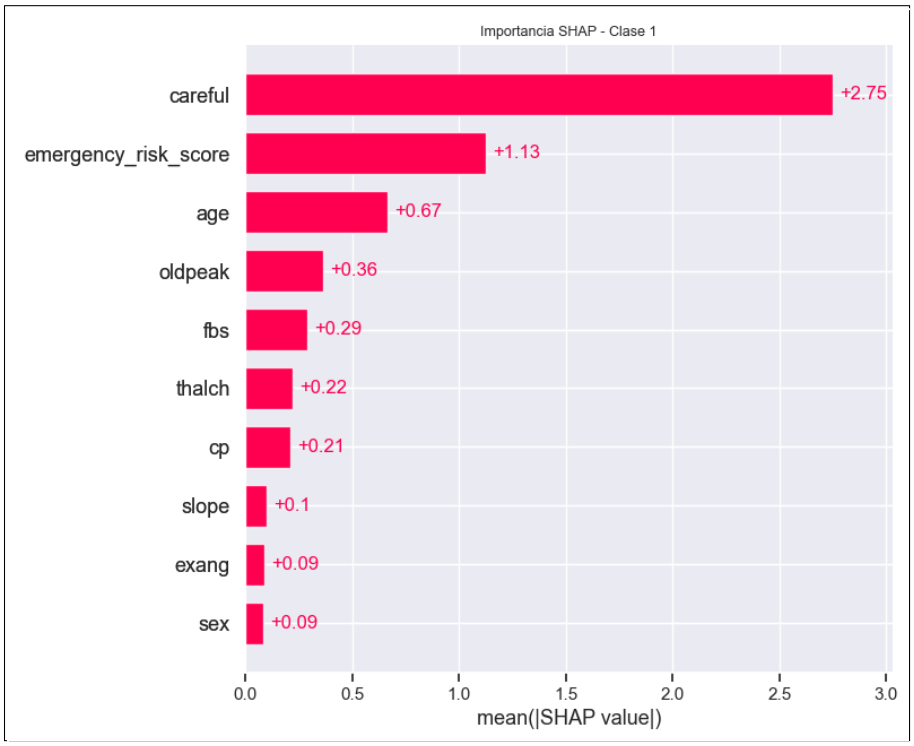

Figura 38: Importancia media de las características para la Clase 1, según los valores SHAP con XGBoost

En la **Figura 38**, correspondiente a la Clase 1, se observa que la variable careful presenta una influencia destacada, con un valor SHAP medio de aproximadamente +3,47, muy por encima del resto de características. Le siguen emergency_risk_score y oldpeak, con contribuciones de +1,63 y +0,92, respectivamente. Este patrón sugiere que dichas variables sintetizadas tienen una gran capacidad de discriminación en la predicción de esta clase.

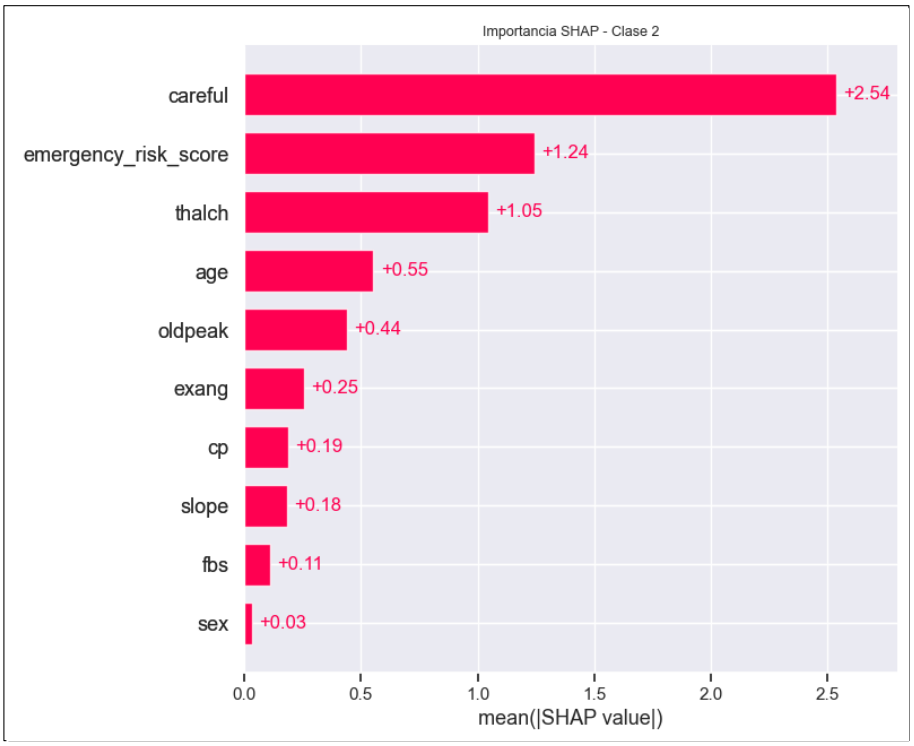

Figura 39: Importancia media de las características para la Clase 2, según los valores SHAP con XGBoost

Por otro lado, en la **Figura 39**, correspondiente a la Clase 2, la variable careful vuelve a situarse como la más relevante (+2,77), seguida por emergency_risk_score (+2,09) y thalach (+1,72). Aunque algunas variables mantienen su relevancia entre clases, la magnitud y el orden de importancia difieren, lo que indica que el modelo emplea distintos mecanismos de decisión en función de la clase objetivo.

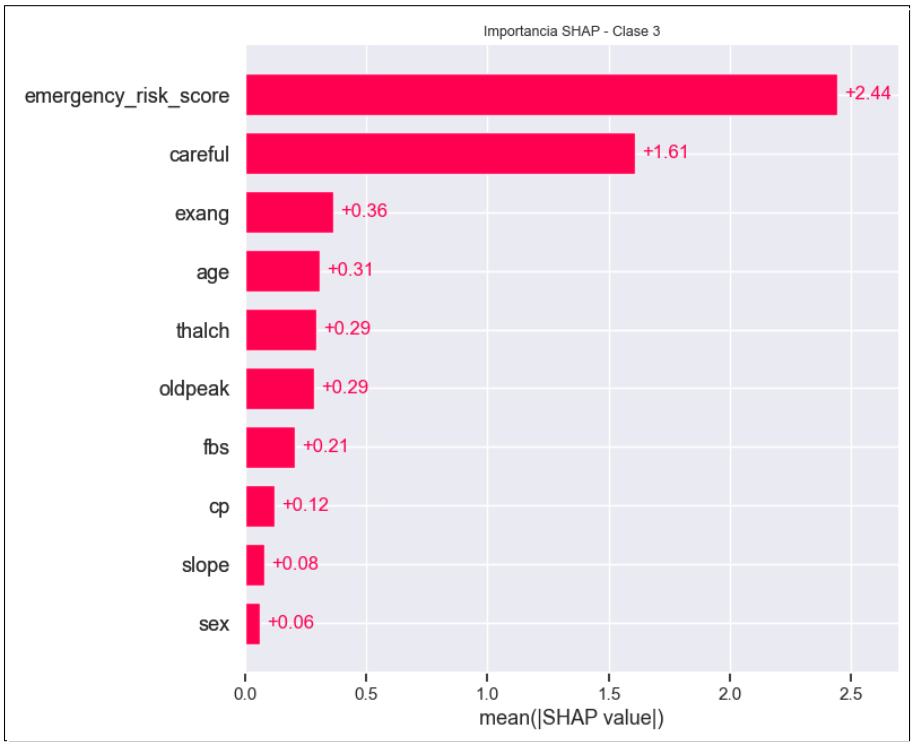

Figura 40: Importancia media de las características para la Clase 3, según los valores SHAP con XGBoost

En la **Figura 40**, correspondiente a la Clase 3, la variable emergency_risk_score pasa a ocupar el primer lugar con una contribución media de +4,05, notablemente superior al resto. Le siguen careful (+1,33) y oldpeak (+0,79). Esta redistribución sugiere que el modelo basa su predicción para esta clase principalmente en factores de riesgo integrados, en lugar de en marcadores anatómicos específicos.

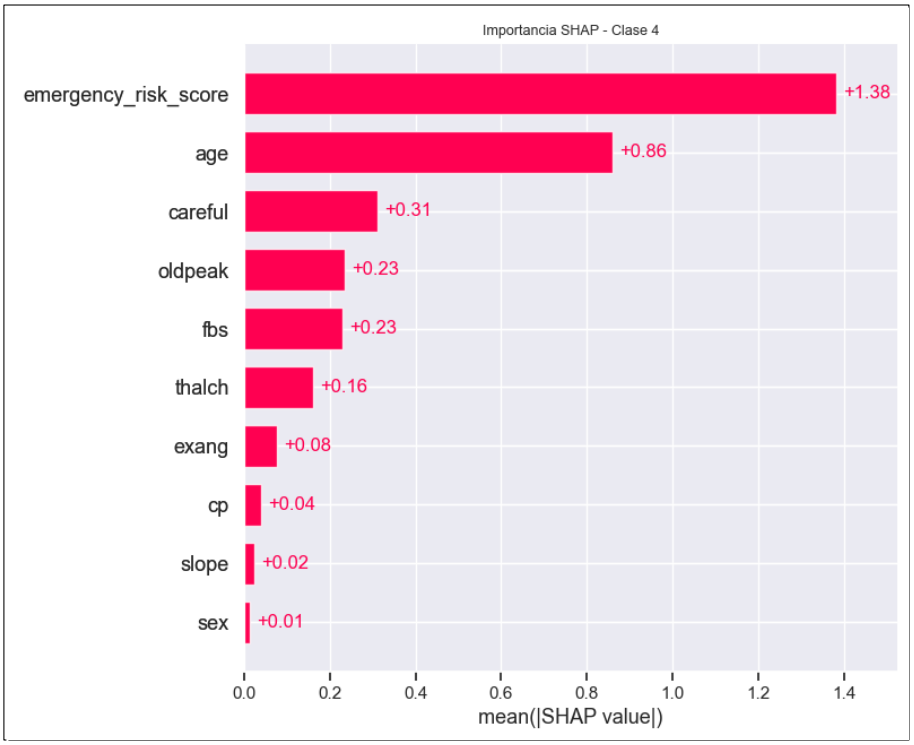

Figura 41: Importancia media de las características para la Clase 4, según los valores SHAP con XGBoost

Finalmente, en la **Figura 41**, correspondiente a la Clase 4, la variable más importante vuelve a ser emergency_risk_score (+1,83), seguida por age (+1,23) y oldpeak (+0,69). En este caso, careful desaparece completamente del ranking (valor SHAP medio = 0), lo que puede indicar que en esta clase la anatomía coronaria deja de ser un criterio diferenciador y el modelo se enfoca en el riesgo basal relacionado con la edad y la respuesta funcional al esfuerzo. Este es un patrón que se extiende al resto de modelos.

Desde una perspectiva clínica, los resultados evidencian la gran relevancia de dos variables derivadas mediante técnicas de feature engineering: careful y emergency_risk_score, cuyas contribuciones varían según la clase, pero en conjunto destacan como pilares de la capacidad explicativa del modelo.

La gran contribución de la variable careful en las Clases 1 y 2 sugiere que el modelo está captando adecuadamente la presencia de afectación anatómica relevante, reflejando un fenotipo

clínico más claramente isquémico. La pérdida de relevancia en la Clase 3 y su desaparición total en la Clase 4 podrían representar escenarios donde la carga aterosclerótica no es el determinante principal del riesgo, como en pacientes mayores o con disfunción microvascular.

Por su parte, el peso dominante de la variable emergency_risk_score en las Clases 3 y 4 sugiere que estas clases representan perfiles donde la interacción entre múltiples factores funcionales y clínicos cobra más peso que la enfermedad anatómica per se. Esto es coherente con escenarios clínicos como los síndromes coronarios sin elevación del ST, la enfermedad microvascular, o pacientes geriátricos con con múltiples comorbilidades.

Adicionalmente, variables tradicionales como oldpeak (descenso del segmento ST), thalach (frecuencia cardíaca máxima alcanzada) y age (edad) mantienen un papel destacado en distintas clases, reflejando la validez de estas variables como marcadores clásicos de riesgo cardiovascular. Por ejemplo, la edad gana protagonismo en la Clase 4, lo que refuerza su papel como factor de riesgo basal no modificable, particularmente en situaciones donde la edad avanzada condiciona el pronóstico incluso en ausencia de hallazgos estructurales claros.

En contraste, variables como sex, slope (pendiente del ST), o exang (angina inducida por esfuerzo) muestran un impacto menor, lo que podría deberse a su escasa variabilidad o a su redundancia explicativa frente a las variables compuestas y fisiológicas más integradoras.

4

Selección del mejor modelo

Con el objetivo de seleccionar el modelo más adecuado para el problema de clasificación multiclase, se realiza a continuación una comparativa integral basada en las métricas obtenidas por cada clasificador sobre el conjunto de prueba.

Se consideran como principales indicadores la accuracy, el F1-score macro (que evalúa de forma equitativa el rendimiento en todas las clases) y el F1-score ponderado (que pondera por la frecuencia de cada clase), además de analizar específicamente el comportamiento en las clases menos representadas (clase 3 y clase 4), que presentan un mayor desafío por el desbalance de la variable objetivo.

Modelo	Accuracy	F1 Macro	F1 Ponderado	Recall Clase 3	Recall Clase 4
Red Neuronal	**0.875**	0.82	**0.88**	0.67	**1.00**
XGBoost	0.859	**0.85**	0.86	**0.90**	0.83
Random Forest	0.870	0.84	0.87	0.86	0.67
SVM	0.837	0.80	0.84	0.62	**1.00**
Decision Tree	0.799	0.66	0.80	0.71	0.17

Cuadro 7: Comparativa de métricas por modelo en el conjunto de prueba.

Desde una perspectiva global, **XGBoost** y la **Red Neuronal Artificial (MLP)** presentan los mejores resultados generales. El modelo de red neuronal alcanza la mayor accuracy (87.5%) y el mayor F1 ponderado (0.88), lo cual indica una alta eficacia considerando la distribución real de las clases. Por su parte, **XGBoost** obtiene el mejor F1 macro (0.85), reflejando un rendimiento más balanceado entre clases, incluyendo aquellas con menor soporte.

En cuanto al comportamiento en clases minoritarias, el modelo **XGBoost** destaca en la clase 3 con un recall del 90 %, y un sólido rendimiento en la clase 4 (F1-score de 0.91), lo cual demuestra su capacidad para generalizar en escenarios de baja representatividad. El modelo de **Red Neuronal Artificial** logra un recall perfecto en la clase 4 (1.00), aunque con menor precisión (0.60), lo que puede aumentar los falsos positivos. En la clase 3 su desempeño es correcto (recall 0.67), aunque inferior al de **XGBoost** y **Random Forest**.

El modelo **Random Forest** mantiene métricas competitivas, con valores de F1-score cercanos a los de los modelos superiores, y un recall elevado en la clase 3 (0.86), aunque su cobertura en la clase 4 es más limitada (recall 0.67). En contraste, **SVM** presenta un rendimiento más modesto en términos generales, si bien logra un recall perfecto en la clase 4, lo cual no compensa sus menores valores en el resto de métricas.

Finalmente, el modelo **Decision Tree**, aunque interpretable y eficiente computacionalmente, muestra un rendimiento claramente inferior. Su F1-score macro de 0.66 y su bajo recall en la clase 4 (0.17) reflejan limitaciones importantes para capturar patrones en clases menos frecuentes, lo que limita su aplicabilidad en contextos clínicos o de diagnóstico.

Considerando tanto el rendimiento global como la capacidad del modelo para generalizar en presencia de clases desbalanceadas, el modelo **XGBoost** se posiciona como la mejor alternativa. Su combinación de alta precisión, equilibrio entre clases y sólido comportamiento en categorías minoritarias lo convierten en el clasificador más robusto para el problema abordado. Valorando el rendimiento de los modelos, nuestro ranking quedaría estratificado de la siguiente forma:

- ***1.*** **XGBoost**
- ***2.*** **Red Neuronal Artificial (MLP)**
- ***3.*** **Random Forest**
- ***4.*** **SVM**
- ***5.*** **Decision Tree**

5

Conclusiones y Líneas Futuras

5.1. Conclusiones

Este trabajo ha permitido demostrar el potencial del aprendizaje automático como herramienta eficaz para la predicción de enfermedades cardíacas, aportando no solo modelos predictivos competitivos, sino también un marco reproducible que combina limpieza de datos, ingeniería de características, balanceo de clases e interpretación de resultados.

Una de las principales conclusiones es que la calidad del preprocesamiento resulta tan crítica como la selección del modelo. La detección de errores clínicamente inviables (por ejemplo, valores nulos o ceros en variables como colesterol o presión arterial) y su correcta imputación fue fundamental para obtener modelos robustos. Asimismo, la generación de nuevas variables compuestas, como emergency_risk_score o careful, demostró que la combinación de variables clínicas en indicadores sintéticos puede mejorar significativamente la capacidad predictiva y la interpretabilidad de los modelos.

Otro punto destacable es la capacidad diferencial de los algoritmos frente al desbalance de clases. Aunque todos los modelos alcanzaron métricas satisfactorias, aquellos con mejor capacidad de generalización (XGBoost, Redes Neuronales) fueron también los que mostraron mayor sensibilidad ante clases minoritarias. Esto es especialmente relevante en el contexto clínico, donde detectar casos menos frecuentes puede tener un alto impacto en el diagnóstico temprano.

Por otra parte, el análisis SHAP ha puesto de relieve que la transparencia en modelos

complejos es posible. Lejos de ser una caja negra, los resultados demuestran que incluso modelos como XGBoost o SVM pueden ser interpretables si se acompañan de técnicas explicativas adecuadas, lo cual es indispensable en entornos médicos donde la comprensión de los motivos detrás de una predicción es tan importante como su precisión.

Finalmente, este trabajo pone en valor la importancia de integrar la visión estadística, clínica y computacional en un problema de salud pública. La capacidad de detectar patrones de riesgo de manera automatizada puede complementar el juicio clínico, reducir errores de diagnóstico y optimizar recursos, especialmente en sistemas sanitarios con alta carga asistencial.

.

5.2. Líneas Futuras

A partir del trabajo realizado, se proponen varias líneas de desarrollo futuro:

- **Validación externa**: Aplicar los modelos a datos de otros centros hospitalarios no presentes en el conjunto original para evaluar su capacidad de generalización y detectar posibles sesgos contextuales.

- **Desarrollo de una herramienta clínica**: Traducir el modelo entrenado en una interfaz sencilla para uso por personal sanitario permitiría llevar este trabajo a la práctica real, especialmente en centros con recursos limitados.

- **Estudio temporal**: En vez de usar datos estáticos, sería interesante analizar series temporales de pacientes para predecir no solo la presencia de enfermedad, sino su evolución.

- **Inclusión de variables socioeconómicas y estilo de vida**: Actualmente, el modelo se basa en variables clínicas. Incluir información sobre hábitos de vida, tabaquismo, alimentación o nivel socioeconómico podría ampliar el valor predictivo y preventivo del sistema.

Referencias

[1] E. J. Benjamin *et al.*, “Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association,” *Circulation*, vol. 139, no. 10, pp. e56–e528, Mar. 2019, doi: 10.1161/CIR.0000000000000659.

[2] D. Corella and J. M. Ordovas, “Genes, dieta y enfermedades cardiovasculares,” 2014. [Online]. Available: https://www.researchgate.net/publication/28182228

[3] M. D. Hungarian Institute of Cardiology. Budapest: Andras Janosi, L. B. and C. C. F. D. M. D. , Ph. D. V.A. Medical Center, B. S. M. P. M. D. University Hospital, and Z. S. W. S. M. D. University Hospital, “Heart Disease data,” 2020. Accessed: Jun. 03, 2025. [Online]. Available: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data

[4] Little and Rubin, “Wiley Series in Probability and Statistics Statistical Analysis with Missing Data,” 2002. doi: 10.1002/9781119013563.fmatter.

[5] A. Shmuel, O. Glickman, and T. Lazebnik, “Symbolic regression as a feature engineering method for machine and deep learning regression tasks,” *Mach Learn Sci Technol*, vol. 5, no. 2, Jun. 2024, doi: 10.1088/2632-2153/ad513a.

[6] “Exploratory-Data-Analysis-1977-John-Tukey”.

[7] I. Tabas, “Cholesterol in health and disease,” *Journal of Clinical Investigation*, vol. 110, no. 5, pp. 583–590, Sep. 2002, doi: 10.1172/jci200216381.

[8] A. Brunauer, A. Koköfer, O. Bataar, I. Gradwohl-Matis, D. Dankl, and M. W. Dünser, “The arterial blood pressure associated with terminal cardiovascular collapse in critically ill patients: A retrospective cohort study,” *Crit Care*, vol. 18, no. 1, Dec. 2014, doi: 10.1186/s13054-014-0719-2.

[9] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” 2002.

[10] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2011.

[11] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” May 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

UNIVERSIDAD DE MÁLAGA | uma.es