Comparativa de técnicas en modelos predictivos de Machine Learning para enfermades cardiacas

Andres Felipe Quintero Zuluaga

Pregunta de investigación y objetivos

En el contexto actual de la inteligencia artificial y su implementación con los datos que se adquiere diariamente, se puede crear soluciones valiosas, sobre todo en áreas críticas como la salud, en donde existe un sesgo de diagnóstico dependiendo del profesional de la salud que evalúe el caso. Con los modelos de Machine Learning se buscará reducir el número de los malos diagnósticos, implementando los estadísticos de todos los datos ya acumulados, incrementando la probabilidad de acierto de estos. Adicionalmente es importante saber cuáles características fisiológicas y hábitos de los pacientes que se toman como parámetros inciden más en la toma de decisiones para hacer un especial enfoque a estos, en pro de realizar más estudios diagnósticos para evaluar a mas detalle este parámetro y su correlación con las fallas que puede presentar el corazón.

Por lo anterior, el principal objetivo con este proyecto es realizar una comparativa de técnicas para modelos predictivos de anomalías cardiacas basándose en la importancia de las características fisiologicas de entrada utilizando para ello las técnicas de visualización SHAP (SHapley Additive exPlanations).

Objetivos

Principal

- Comparar técnicas de preprocesamiento en modelos predictivos de Machine Learning para clasificar enfermedades cardiacas y analizar la relevancia de las variables fisiológicas mediante visualización con SHAP.

Secundarios

- Realizar análisis exploratorio de los datos
- Crear modelo base entrenando modelos con los datos crudos
- Realizar limpieza de los datos
- Implementar técnicas de preprocesamiento como normalización, balanceo y reducción de dimensionalidad

- Comparar las técnicas de preprocesamiento entrenando los modelos y evaluando métricas principales
- Validar modelos con un segundo dataset que presente una distribución de los datos diferente
- Visualización de la importancia de características con SHAP

Estado del arte

Las enfermedades cardiovasculares se definen como cualquier desviación del correcto funcionamiento del corazón, funciones como bombear oxígeno a todos los órganos por circulación sistémica y regulación hormonal para mantener un óptimo nivel de presión sanguínea (Ogunpola et al., 2024). Se estima que las enfermedades cardiacas causan a nivel global una suma de 17.5 millones de muertes, siendo esta causa quien lidera las muertes por enfermedades (El-Sofany et al., 2024).

Por lo anterior, las enfermedades del corazón es un tema de interés global. Se tiene que se ha intentado todos los modelos de clasificación posible para determinar el más adecuado y el que mejores métricas da. A continuación, se presentan en compendio de estudios relacionados con algoritmos de Machine Learning supervisados para clasificar y predecir posibles fallas del corazón.

En el estudio de "**Effectiveness of machine learning models in diagnosis of heart disease: a comparative study**" Waleed Alsabhan y Abdullah Alfadhly evaluaron hasta 12 modelos, clásicos como: Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, Random Forest; como avanzados: Gradient Boosting Machine, Light Gradient Boosting Machine, Catboost, Artificial Neural Network. Enfocándose en la comparación de técnicas de preprocesamiento como: Standadization, Minmax Scaling, Robust Scaling y Normalization, con el fin de evaluar el efecto de estas técnicas en la predicción de los modelos. Para ello implementaron diferentes métodos de validación como: accuracy, precision, recall, F1 score, Area Under Curve, Cohen's Kappa y Logloss**.** Se concluyó que los modelos con mejor eficiencia fueron los avanzados, específicamente Gradient Boosting Machine, Light Gradient Boosting Machine y Catboost. Aunque variaban su eficiencia dependiendo de la técnica de preprocesamiento implementada, siendo Light Gradient Boosting Machine el que presentó mayor estabilidad al cambio de técnica de preprocesamiento (Alsabhan & Alfadhly, 2025).

En otro estudio llamado "**Machine learning algorithms for heart disease diagnosis: A systematic review**" Yian Mao et al. Realiazan una revisión bibliográfica escogiendo 24 artículos sobre la aplicabilidad de modelos supervisados

de Machine Learning sobre la predicción y diagnóstico de enfermedades cardiacas y evaluar los modelos más usados. Los artículos revisados fueron recopilados desde el 2013 hasta el 2024, empleando Dataset como los de Cleveland o Statlog y modelos de clasificación como: Decision Tree, Random Forest, Naïve Bayes, Logistics Regression y Artificial Neural Network. Para ello los autores evaluaron y clasificaron cada artículo según: selección de proceso, fuente de la información, riesgo del sesgo de evaluación y métricas. Se determinó que el modelo de Machine Learning más usado es el de Decision Tree, debido a que refleja características esenciales en el Dataset, el trabajar con gran cantidad de datos con baja dimensionalidad, manejo eficiente datos sin transformar y robustes a valores atípicos, aunque presenta sus desventajas como el sobreajuste y el "greedy method" (Mao et al., 2025).

En el estudio de "**Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance**" M Darshan Teja y G Mokesh Rayalu integraron y evaluaron varios repositorios como los de Cleveland, Switzerland, Hungary, Long Beach y Statlog. Para comparar la eficiencia de 15 modelos de Machine Learning, clásicos y avanzados, entre ellos: Random Forest, K-Nearest Neighbors, Logistic Regression, Naïve Bayes, Gradient Boosting, AdaBoost, XGBoost y Bagged Trees. Para evaluar su rendimiento usaron varias métricas como: precision, recall, F1-score y ROC-AUC; usando la técnica de validación de K-fold cross-validation. El desempeño de cada modelo vario dependiendo de la métrica usada, por ejemplo, XGBoost y Bagged Trees obtuvieron el mayor accuracy con un 93%, Random Forest y Bagged Trees lo obtuvieron para el ROC-AUC. Y que las métricas como precision, recall y F1-score mostraron sensibilidad y especificidad para los modelos de Bagged Trees y Random Forest (Teja & Rayalu, 2025).

En el estudio de "**Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment**" Moiz Ur Rehman et al. Destacan las limitaciones que han mostrado los estudios sobre la predicción de enfermedades cardiacas usando modelos de Machine Learning. Haciendo énfasis en:

- Selección de características: La selección de todas las características como entradas y no la selección de las características más relevantes para optimizar la exactitud de las predicciones.
- Desbalance de los datos clínicos: La aplicación de balanceos de clases sin enfocarse en la importancia de este y no sacarle el potencial de técnicas de balanceo de clases como el SMOTE.

- Mal desempeño en grandes datasets: Algunos modelos como Random Forest y K-Nearest Neighbor, presentan un mal desempeño en datasets grandes y complejos con múltiples clases.
- Inconsistencia en los resultados: Análisis pobre a la hora de comparar los resultados en de los diferentes modelos aplicados.
- No aplicabilidad al mundo real: Se estudian modelos muy teóricos, sin tener en cuenta su aplicación en un entorno real.

(Rehman et al., 2025)

Metodología de investigación

Se escogió un dataset llamado “Cardiovascular Disease Dataset” desde la plataforma de Mendeley Data. A este dataset se le realizará un análisis exploratorio de los datos; implementando la matrix de correlación, visualización de la distribución de las variables y medición de outliers con gráficos unidimensionales como los boxplots. Posteriormente se implementará un modelo base con los datos en crudo utilizando modelos de clasificación supervisados como: Logistic Regresion, Decision Tree, Random Forest, Naive Bayes y XGBoost, sacados de las librerías de scikit-learn y xgboost. Luego se realizará una comparativa con métricas de clasificación entre los datos originales filtrados y preprocesados, con técnicas de normalización y balanceo como MinMaxScaler y SMOTE, contra los datos originales, pero aplicándoles técnicas de reducción de dimensionalidad como PCA en 2 y 3 dimensiones. Lo anterior se ejecutará acompañado de una técnica de visualización de importancia de características como SHAP. Por último, se hará un tuneo de hiperparámetros únicamente para el modelo XGBoost utilizando gridSearchCV de la librería scikit-learn. En cada paso se realizará una validación con un segundo dataset, sacado de la plataforma de Kaggle, el cual presenta las mismas características de entrada que el primer dataset, pero con una distribución de los datos diferente.

Análisis de los datos

El primer dataset fue otorgado por un hospital de la India, este cuenta con 14 columnas entre ellas la variable objetivo y posee 1000 observaciones. La validación se realizará con segundo dataset, que es una recopilación de diferentes datasets independientes como: Cleverland, Hungarian, Switzerland, Long Beach y Stalog; los

cuales cuentan con 303, 294 123, 200 y 270 observaciones tomadas respectivamente, para un total de 1190 observaciones.

A continuación, se describen las características fisiológicas cardiacas que se usarán para el proyecto.

Variables independientes

- *patientid:* Identificación del paciente (Numérica)
- *age:* Edad del paciente (Numérica)
- *gender:* Sexo del paciente (Binaria)
- *chestpain:* Clasificación del dolor de pecho por poca oxigenación del mismo (Numérica)
    - *Angina típica:* 0
    - *Angina Atípica:* 1
    - *Sin dolor anginoso:* 2
    - *Asintomático:* 3
- *restingBP:* Medida de la presión sistólica/diastólica con el paciente en reposo (Numérica) (94 a 200) [mm Hg]
- *serumcholestrol:* Medida del colesterol en sangre (Numérica) (126 a 564) [mg/dl]
- *fastingbloodsugar:* Medida de la glucosa en sangre (Binaria) (mayor a 120) [mg/dl]
    - *Falso:* 0
    - *Verdadero:* 1
- *restingrelectro:* Diagnostico en el electrocardiograma del paciente en reposo (Numérica)
    - *Normal:* 0
    - *Anomalía en la onda ST-T:* 1
    - *Probable hipertrofía del ventrícvulo izquierdo:* 2
- *maxheartrate:* Máxima cantidad de latidos alcanzados (Numérica) (71 a 202) [Lat/min]
- *exerciseangia:* Angina producida por realizar ejercicio (Binaria)
    - *No:* 0
    - *Si:* 1
- *oldpeak:* Depresión de la onda ST (Numérica) (0 a 6.2)
- *slope:* Cambio del segmento ST durante una prueba de ejercicio (Numérica)
    - *Ascendente:* 1
    - *Horizontal:* 2
    - *Descendente:* 3
- *noofmajorvessels:* Número de vasos mayores coloreados por fluoroscopia (Numérica)

Variable dependiente

- **target:** Clasificación de los pacientes (Binaria)
    - *Ausencia de enfermedad cardiaca:* 0
    - *Presencia de enfermedad cardiaca:* 1

En las Figura 1 y Figura 2 se muestra la distribución y correlaciones respetivamente del primer dataset. En las distribuciones, quitando las variables categóricas, se muestra una distribución uniforme en la edad. Para el colesterol, omitiendo las 53 observaciones con colesterol 0 (algo imposible), contamos con una tendencia a una distribución normal. Para la medida de la presión sistólica/diastólica notamos que los individuos que pueden presentar alguna enfermedad cardiaca se agrupan mas en valores altos de presión. Lo mismo ocurre con la máxima cantidad de latidos alcanzados, en donde para las personas a las que no se le identificaron alguna falla cardiaca, se distribuyen uniformemente a lo largo del rango, mientras que los individuos que si se identifica falla cardiaca tienden a tener una distribución normal sesgada hacia la derecha. Por último, para la depresión de la onda ST, tanto en la presencia y ausencia de enfermedad cardiaca tienen una tendencia descendente a altos valores. Para la gráfica de correlación no hay demasiada colinealidad entre las variables independientes, pero si se tiene una alta correlación entre la variable slope y la variable objetivo, con un 0.8, con algo menos de correlación tenemos el dolor de pecho con la variable objetivo con un 0.53. Estos dos hallazgos también se presentan en al utilizar la técnica SHAP, la cual para la mayoría de modelos las tomaron como las características que más influye para las predicciones.

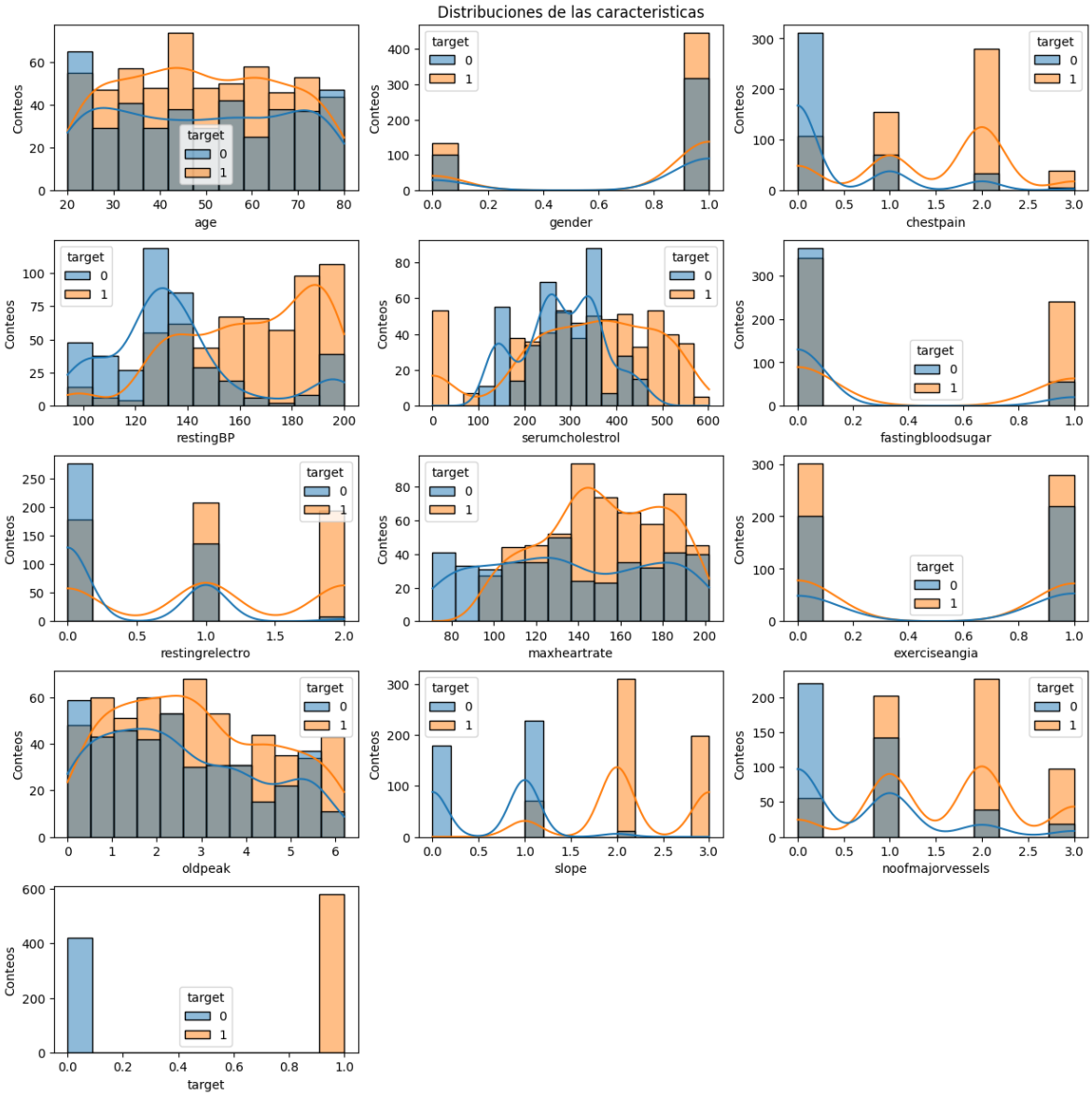

Figura 1 Distribución de las características del 1er dataset

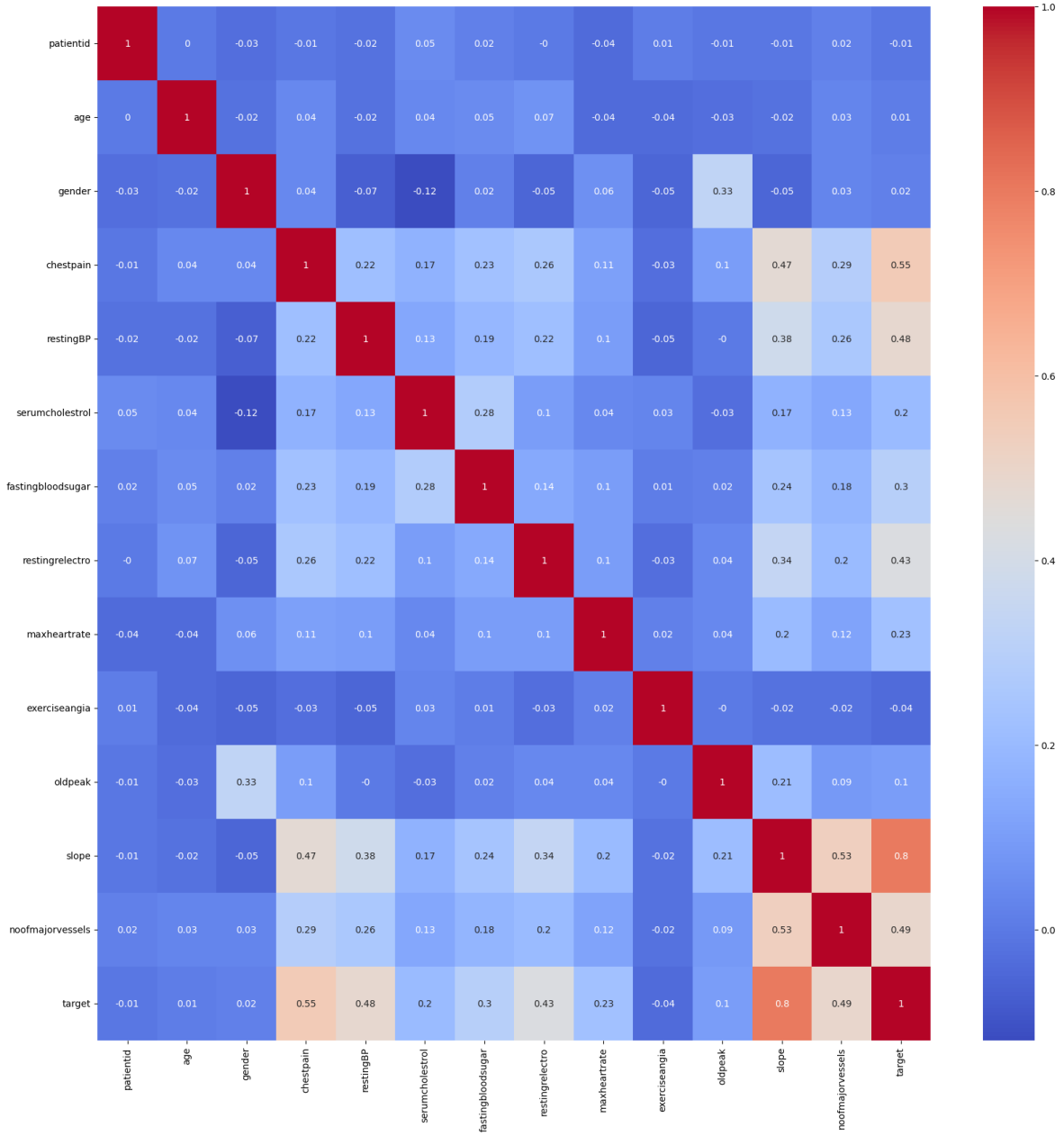

Figura 2 Matriz de correlación 1er dataset

En la Figura 3 y Figura 4 se muestran la distribución y correlaciones respectivamente del segundo dataset. Lo primero que se nota es una diferencia entre ambos datasets, se ve una distribución diferente de las variables numéricas en el segundo dataset, con una tendencia a una distribución normal para ambos grupos objetivos. Se identifica que las personas con una edad mayor presentan mayores casos de falla cardiaca. Que la distribución del colesterol en sangre es muy similar en ambos grupos objetivos, aunque hay presencia de observaciones con colesterol 0, unas 172 observaciones en total. También se nota una diferencia con

respecto al grupo en la máxima cantidad de latidos, en el primer dataset los que presentaban mayor ratio de latidos por minutos eran las personas con presencia de enfermedad cardiaca, sin embargo, para este segundo dataset cambia, las personas con mayor ratio de latidos por segundo serán las personas que no presentan alguna enfermedad cardiaca. Además, que este segundo dataset no cuenta con la variable vasos sanguíneos mayores marcados con fluoroscopia. Con respecto a la matriz de correlación notamos que hay mas cantidad de variables independientes medianamente correlacionadas entre si y que la correlación entre las variables independientes con la variable dependiente no es tan fuerte como la obtenida para el primer dataset.

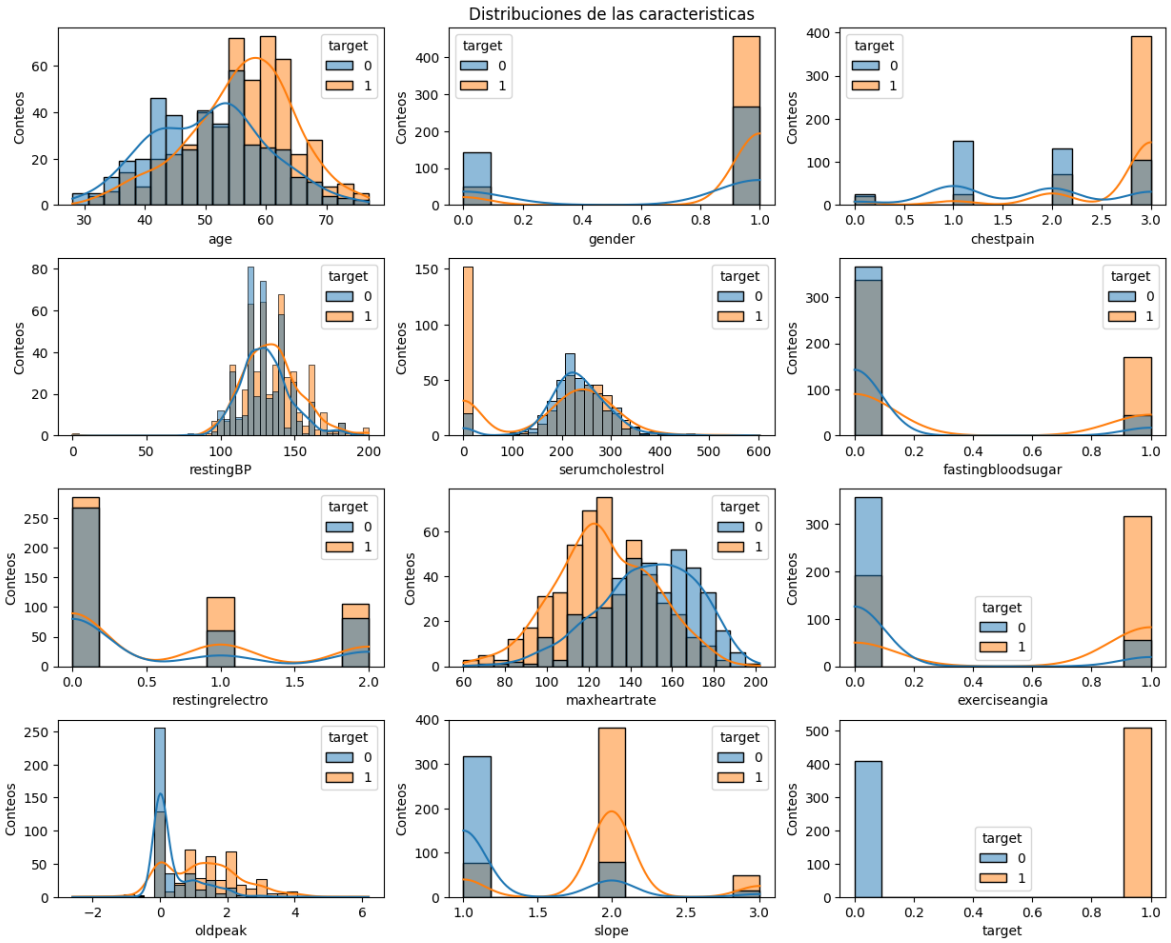

Figura 3 Distribución de las características del 2do dataset

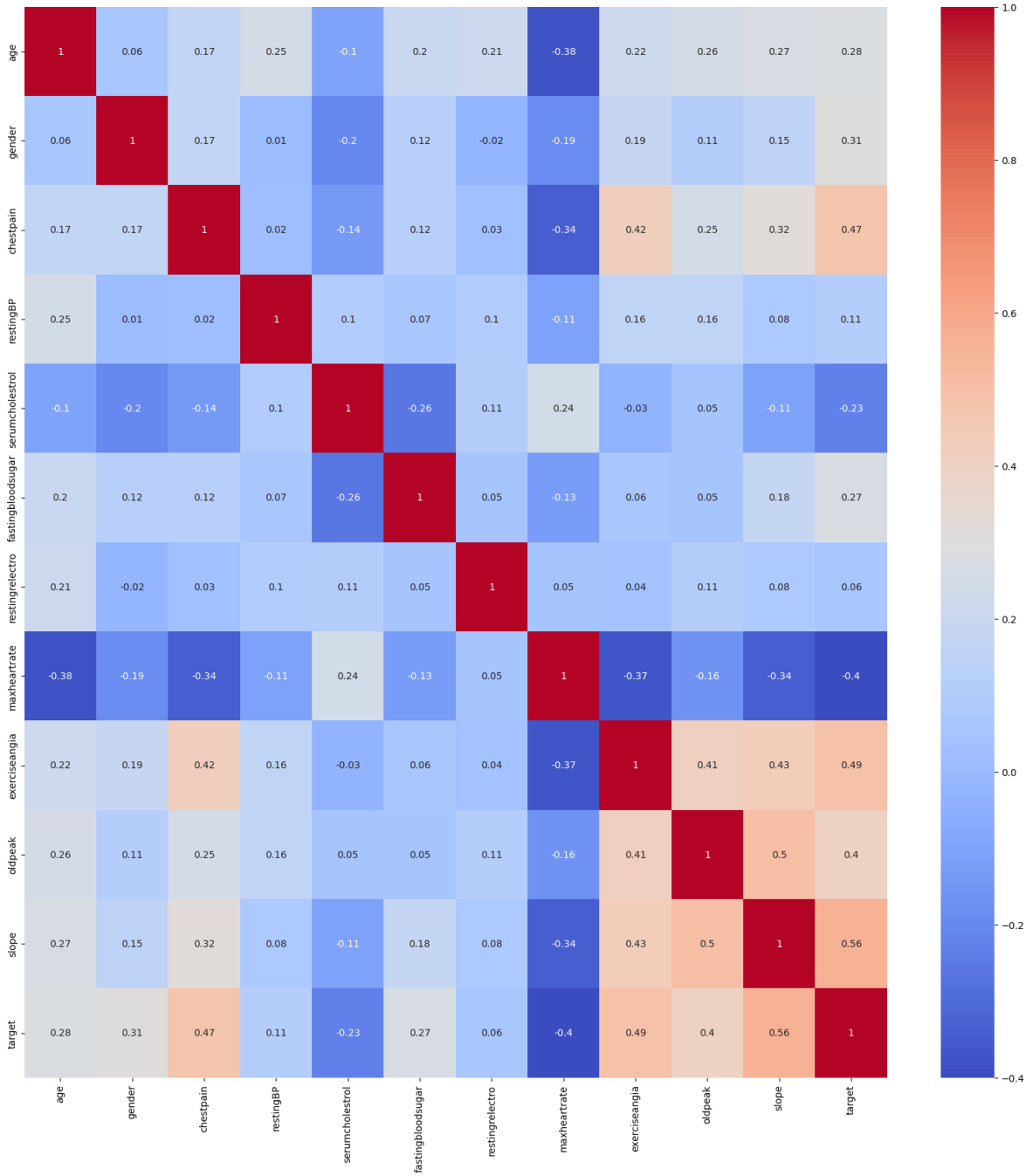

Figura 4 Matriz de correlación 2do dataset

Uso de la metodología y herramientas de aprendizaje estadístico

Las herramientas usadas fueron las que se vieron a lo largo del diplomado, comenzando con Pandas para el tratamiento de datos tabulares, Matplotlib y Seaborn para la visualización de los datos. Se hizo un amplio uso de la librería Scikit-Learn tanto para el encoding, preprocesamiento de los datos, separación de los datos, normalización y principalmente la importación de los modelos de clasificación convencionales; Logistic Regression, Decision Tree, Random Forest y Naive Bayes. Se eligieron estos debido a que fueron los que se vieron en el diplomado. Adicionalmente se hizo uso del algoritmo de XGBoost importada de la librería de xgboost, para implementar un algoritmo mas avanzado y comparar sus resultados con los algoritmos convencionales. En cada paso se realizó una validación con un segundo dataset, para visualizar el comportamiento del modelo entrenado con datos con una distribución diferente. A su vez, en cada paso se realizó una visualización con SHAP, de la librería shap, las cual nos ayuda a darle la importancia a cada una de las características de entrada según su aporte a la predicción del modelo, este se basa en la teoría de juegos cooperativos. Por último, se realizó un tuneo de hiperparámetros y se selecciónó el modelo de XGBoost debido a que fue el que a lo largo de todas las pruebas mostró mejores resultados, los hiperparámetros se eligieron con base a la documentación leída.

Conclusiones y trabajo futuro

El objetivo de este trabajo fue analizar, paso a paso, distintas técnicas de preprocesamiento y su influencia en la exactitud de las predicciones de diversos modelos de clasificación. También se buscaba observar cómo variaba la importancia de las características fisiológicas en la predicción a medida que se aplicaban distintas transformaciones al conjunto de datos.

Inicialmente, se observó que los modelos se ajustaban excesivamente al primer dataset, alcanzando una exactitud cercana al 100%, incluso en los datos de prueba. Sin embargo, al validar los modelos con un segundo dataset, con una distribución diferente, las métricas de desempeño disminuían considerablemente, con un acierto hasta del 70%.

Para intentar mejorar el rendimiento en el segundo dataset, se implementaron técnicas de preprocesamiento como MinMaxScaler y SMOTE. Aunque los resultados en los datos de entrenamiento y prueba seguían siendo muy buenos, las métricas de validación en el segundo dataset apenas mejoraban en promedio un 1%. Lo mismo ocurrió al aplicar el ajuste de hiperparámetros: las métricas con el primer dataset fueron excelentes, pero en el segundo conjunto de datos no se observaron mejoras significativas.

También se evaluó el uso de PCA con 2 y 3 componentes. Los resultados iniciales fueron muy pobres al aplicar PCA directamente. Sin embargo, al aplicar una normalización previa, los resultados mejoraron considerablemente. Aun así, los modelos continuaron obteniendo mejores métricas sin el uso de PCA.

Para trabajos futuros, lo ideal sería mejorar la generalidad de los modelos, que podría hacerse usando modelos de clasificación mas avanzados y técnicas de preprocesamiento mas robustas, principalmente para datasets médicos, los cuales presentan un desbalanceo marcado de las clases.

Implicaciones éticas

Con respecto a la implicación ética, la más destacada sería si el modelo falla, como todo en la salud es crítico se pueden llegar resultados catastróficos. El modelo podría fallar de dos maneras. La primera seria dando una falsa alarma a paciente que no presentan ninguna posibilidad de falla cardiaca por lo que se perderían recursos en exámenes complementarios, y la segunda, no diagnosticar a los pacientes que sí podrían tener futuras complicaciones y por la confianza que generaría el modelo no prevenir fallas cardiacas.

Otra implicación ética sería que las predicciones podrían generar duda en los profesionales cuando la predicción del modelo sea contraría a la estimada por el médico, por lo que él médico podría tomar 2 decisiones en ese caso. La primera sería diagnosticar según lo que predijo el modelo por encima a su opinión, por lo que sentiría que está siendo relegado de su cargo. La segunda sería diagnosticar según su opinión sobre la predicha por el modelo, lo que generaría desconfianza hacia su diagnóstico debido a que está yendo en contra de un modelo que trabaja con una ingesta robusta de datos.

Aspectos legales y comerciales

La salud como un aspecto de interés social y la cantidad de centros de atención hospitalarios, se crearía una gran demanda por la implementación de nuevas tecnologías en pro al soporte de los diagnósticos médicos. Para lo anterior se tendría que implementar medidas regulatorias estrictas y evitar que cualquier persona cree su propio modelo solo con fines de lucro, sin tener en cuenta un correcto diagnóstico.

Con respecto al marco legal, siempre entra en juego las demandas por mal diagnósticos. Entraría en rigor las leyes nacionales que amparan al paciente por errores en el diagnóstico y las implicaciones que pudieron tener esos errores.

Bibliografía

Alsabhan, W., & Alfadhly, A. (2025). Effectiveness of machine learning models in diagnosis of heart disease: a comparative study. *Scientific Reports*, *15*(1), 1–19. https://doi.org/10.1038/S41598-025-09423-Y/METRICS

El-Sofany, H., Bouallegue, B., & El-Latif, Y. M. A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Scientific Reports 2024 14:1*, *14*(1), 1–18. https://doi.org/10.1038/s41598-024-74656-2

Mao, Y., Jimma, B. L., & Mihretie, T. B. (2025). Machine learning algorithms for heart disease diagnosis: A systematic review. *Current Problems in Cardiology*, *50*(8), 103082. https://doi.org/10.1016/J.CPCARDIOL.2025.103082

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics 2024, Vol. 14, Page 144*, *14*(2), 144. https://doi.org/10.3390/DIAGNOSTICS14020144

Rehman, M. U., Naseem, S., Butt, A. U. R., Mahmood, T., Khan, A. R., Khan, I., Khan, J., & Jung, Y. (2025). Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Scientific Reports*, *15*(1), 1–15. https://doi.org/10.1038/S41598-025-96437-1;SUBJMETA=114,166,631,639,705,985;KWRD=BIOMEDICAL+ENGINEERING,COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS,ENGINEERING,MATHEMATICS+AND+COMPUTING

Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. *BMC Cardiovascular Disorders*, *25*(1), 1–12. https://doi.org/10.1186/S12872-025-04627-6/TABLES/6