

**UAX**

**TRABAJO FIN DE MÁSTER**

**MUIA  
24-25**

**UNIVERSIDAD ALFONSO X EL SABIO**

**Business Tech**

**Máster Universitario en Inteligencia Artificial**



## **TRABAJO DE FIN DE MÁSTER**

**Diseño e implementación de modelos de  
Inteligencia Artificial para la identificación  
temprana de pacientes con riesgo de sufrir  
ataques cardíacos**

**JON MAESTRE ESCOBAR  
Beatriz Magán Pinto**

**junio 2025**

## RESUMEN

El presente trabajo se centra en el diseño e implementación de sistemas inteligentes orientados a la detección precoz de pacientes con riesgo elevado de padecer eventos cardiovasculares, empleando para ello información médica y parámetros biométricos que permitan generar estimaciones fiables. Mediante el estudio pormenorizado de un conjunto de datos procedente de registros sanitarios indios, que engloba mediciones fisiológicas, patrones conductuales e históricas clínicas, he establecido los cimientos necesarios para llevar a cabo tanto el análisis exploratorio como la construcción de modelos predictivos.

La investigación arranca con una etapa fundamental de tratamiento y exploración de datos, donde he aplicado técnicas de visualización y análisis estadístico para desentrañar las distribuciones subyacentes y las interacciones entre variables. Durante esta fase descubrí peculiaridades importantes en el dataset, como la ausencia de outliers detectados por métodos tradicionales y distribuciones artificialmente uniformes que me llevaron a cuestionar la naturaleza de los datos originales. Esto me obligó a replantear mi aproximación inicial y crear una nueva variable objetivo basada en criterios clínicos establecidos, combinando factores de riesgo discretos con umbrales médicamente validados.

En este proyecto he implementado un espectro diverso de técnicas computacionales que abarca tres grandes familias algorítmicas. Comencé con métodos de machine learning tradicional como Random Forest, XGBoost, SVM y LightGBM, que demostraron rendimientos excepcionales con la nueva variable objetivo. Posteriormente exploré arquitecturas de deep learning, incluyendo redes neuronales densas, CNNs adaptadas para datos tabulares y ResNet, que confirmaron los patrones identificados desde perspectivas diferentes. Finalmente, investigué transformers especializados para datos tabulares como TabTransformer, FT-Transformer y SAINT, que aunque

conceptualmente prometedores, mostraron resultados más modestos para este problema específico.

Los resultados han sido reveladores en múltiples aspectos. Los algoritmos de boosting como XGBoost y LightGBM lograron clasificación perfecta, mientras que Random Forest ofreció un rendimiento excepcional con un margen de error más realista. Paradójicamente, los transformers, que representan el estado del arte en muchos dominios, mostraron rendimientos inferiores, enseñándome que la complejidad arquitectural no siempre se traduce en mejor desempeño para problemas específicos.

Mi objetivo fundamental ha consistido en desarrollar sistemas capaces de estratificar a los pacientes según su nivel de riesgo con elevada exactitud, proporcionando además explicaciones comprensibles y recomendaciones personalizadas. Para validar la aplicabilidad práctica del trabajo, desarrollé una interfaz gráfica que integra el modelo seleccionado con funcionalidades de visualización de datos, generación de recomendaciones y exportación de informes médicos profesionales.

La selección final de Random Forest como mejor modelo se basó en criterios que van más allá de la precisión pura, considerando interpretabilidad, robustez ante datos nuevos y credibilidad estadística. Esta decisión refleja una lección importante: en aplicaciones médicas, el equilibrio entre rendimiento, transparencia y aplicabilidad práctica puede ser más valioso que la optimización de métricas aisladas.

Esta investigación constituye una contribución relevante en la intersección entre la inteligencia computacional y la medicina preventiva, demostrando cómo la aplicación reflexiva y metodológicamente rigurosa de técnicas de IA puede generar herramientas verdaderamente útiles para la identificación temprana del riesgo cardiovascular, siempre manteniendo consideraciones éticas fundamentales sobre privacidad, equidad y transparencia en el desarrollo de sistemas de inteligencia artificial médica.

## Palabras clave

Inteligencia Artificial, salud cardiovascular, Machine Learning, Deep Learning, modelos predictivos.

## ABSTRACT

This work focuses on designing and implementing intelligent systems for early detection of patients at high risk of cardiovascular events, using medical information and biometric parameters to generate reliable estimates. Through detailed study of a dataset from Indian healthcare records that includes physiological measurements, behavioral patterns, and clinical histories, I established the necessary foundations for both exploratory analysis and predictive model development.

The research begins with a fundamental data processing and exploration phase, where I applied visualization techniques and statistical analysis to uncover underlying distributions and variable interactions. During this phase, I discovered important peculiarities in the dataset, such as the absence of outliers detected by traditional methods and artificially uniform distributions that led me to question the nature of the original data. This forced me to rethink my initial approach and create a new target variable based on established clinical criteria, combining discrete risk factors with medically validated thresholds.

In this project, I implemented a diverse spectrum of computational techniques spanning three major algorithmic families. I started with traditional machine learning methods like Random Forest, XGBoost, SVM, and LightGBM, which demonstrated exceptional performance with the new target variable. I then explored deep learning architectures, including dense neural networks, CNNs adapted for tabular data, and ResNet, which confirmed the identified patterns from different perspectives. Finally, I

investigated specialized transformers for tabular data such as TabTransformer, FT-Transformer, and SAINT, which, although conceptually promising, showed more modest results for this specific problem.

The results have been revealing in multiple aspects. Boosting algorithms like XGBoost and LightGBM achieved perfect classification, while Random Forest offered exceptional performance with a more realistic error margin. Paradoxically, transformers, which represent the state of the art in many domains, showed inferior performance, teaching me that architectural complexity doesn't always translate to better performance for specific problems.

My fundamental objective has been to develop systems capable of stratifying patients according to their risk level with high accuracy, while also providing understandable explanations and personalized recommendations. To validate the practical applicability of the work, I developed a graphical interface that integrates the selected model with data visualization functionalities, recommendation generation, and professional medical report export capabilities.

The final selection of Random Forest as the best model was based on criteria that go beyond pure accuracy, considering interpretability, robustness with new data, and statistical credibility. This decision reflects an important lesson: in medical applications, the balance between performance, transparency, and practical applicability can be more valuable than optimizing isolated metrics.

This research constitutes a relevant contribution at the intersection of computational intelligence and preventive medicine, demonstrating how thoughtful and methodologically rigorous application of AI techniques can generate truly useful tools for early cardiovascular risk identification, while always maintaining fundamental ethical considerations about privacy,

equity, and transparency in medical artificial intelligence system development.

## **Keywords**

Artificial Intelligence, cardiovascular health, Machine Learning, Deep Learning, predictive modeling.

# ÍNDICE

<b>1. INTRODUCCIÓN .....</b>	<b>15</b>
<b>2. ESTADO DEL ARTE .....</b>	<b>18</b>
2.1. Contexto .....	19
2.2. Justificación .....	20
2.3. Análisis comparativo de estudios existentes .....	21
2.3.1. Revisión general de estudios existentes.....	22
2.3.2. Metodologías empleadas en investigaciones previas.....	32
2.3.3. Conjuntos de datos utilizados en estudios relacionados.....	37
2.3.4. Resultados y hallazgos .....	40
2.3.5. Casos de estudio en proyectos reales .....	45
2.3.6. Gráficos y tabla comparativa .....	48
2.3.7. Comparaciones con mi trabajo fin de máster .....	51
<b>3. OBJETIVOS .....</b>	<b>56</b>
3.1. Objetivos generales .....	57
3.2. Objetivos específicos .....	57
3.3. Alcance .....	59
3.4. Limitaciones.....	60
<b>4. METODOLOGÍA .....</b>	<b>63</b>
4.1. Recopilación y preprocessamiento de datos .....	63
4.2. Desarrollo y entrenamiento de los modelos.....	64
4.3. Evaluación de los modelos .....	64
4.4. Gestión de archivos con Google Drive y GitHub .....	65
4.5. Planificación y organización de tareas con ClickUp .....	66
<b>5. PLANIFICACIÓN Y PRESUPUESTO .....</b>	<b>68</b>

5.1.	Planificación.....	68
5.2.	Presupuesto.....	69
<b>6.</b>	<b>DESARROLLO DEL PROYECTO .....</b>	<b>72</b>
6.1.	Ciencia de datos .....	74
6.1.1.	Diseño .....	75
6.1.1.1.	Detección de outliers.....	76
6.1.1.2.	Análisis visual y estadístico.....	79
6.1.2.	Desarrollo .....	81
6.1.2.1.	Detección de outliers.....	81
6.1.2.2.	Análisis visual y estadístico.....	84
6.1.3.	Resultados .....	87
6.1.3.1.	Detección de outliers.....	87
6.1.3.2.	Análisis visual y estadístico.....	108
6.1.3.2.1.	Distribuciones – Histogramas.....	109
6.1.3.2.2.	Diagramas de caja – Boxplots .....	148
6.1.3.2.3.	Diagramas de barras y Tops 10 .....	153
6.1.3.2.4.	Diagramas de dispersión – Scatter plots .....	175
6.1.3.2.5.	Diagramas circulares – Pie Charts.....	183
6.2.	Modelos de IA.....	188
6.2.1.	Primer acercamiento – Fase preliminar .....	190
6.2.2.	Redefinición del riesgo cardiovascular – Creación de una nueva variable objetivo .....	193
6.2.3.	Diseño .....	196
6.2.3.1.	Machine Learning.....	197
6.2.3.2.	Deep Learning.....	198
6.2.3.3.	Transformers.....	200
6.2.4.	Desarrollo .....	201

6.2.4.1.	Machine Learning.....	202
6.2.4.2.	Deep Learning.....	204
6.2.4.3.	Transformers.....	206
6.2.5.	Resultados .....	208
6.2.5.1.	Machine Learning.....	208
6.2.5.1.1.	Random Forest .....	209
6.2.5.1.2.	XGBoost .....	211
6.2.5.1.3.	Support Vector Machine (SVM) .....	213
6.2.5.1.4.	LightGBM.....	215
6.2.5.2.	Deep Learning.....	217
6.2.5.2.1.	Red Neuronal Profunda Densa .....	217
6.2.5.2.2.	Red Neuronal Convolucional (CNN) .....	219
6.2.5.2.3.	ResNet .....	221
6.2.5.3.	Transformers.....	223
6.2.5.3.1.	TabTransformer .....	224
6.2.5.3.2.	FT-Transformer.....	226
6.2.5.3.3.	SAINT .....	228
6.3.	Interfaz de Usuario.....	230
6.3.1.	Diseño .....	231
6.3.2.	Desarrollo .....	233
6.3.3.	Resultados .....	236
7.	<b>RESULTADOS .....</b>	241
8.	<b>ÉTICA DEL PROYECTO.....</b>	249
9.	<b>BIBLIOGRAFÍA .....</b>	256

## ÍNDICE DE ILUSTRACIONES

Ilustración 1: Precisión de los modelos de IA en los estudios analizados ....	50
Ilustración 2: Frecuencia de predictores clave en los estudios analizados ..	50
Ilustración 3: Repositorio GitHub .....	65
Ilustración 4: Repositorio Google Drive .....	66
Ilustración 5: Ejemplo ClickUp tareas completadas y pendientes .....	67
Ilustración 6: Método IQR .....	81
Ilustración 7: Método Z-Score.....	82
Ilustración 8: Método Isolation Forest .....	82
Ilustración 9: Método Local Outlier Factor.....	83
Ilustración 10: Variables Detección Outliers.....	83
Ilustración 11: Resultados Método IQR .....	88
Ilustración 12: Resultados Método Z-Score.....	92
Ilustración 13: Resultados Método Isolation Forest .....	96
Ilustración 14: Resultados Método Local Outlier Factor.....	100
Ilustración 15: Resultados IF vs LOF Outliers.....	104
Ilustración 16: Distribución de Edad, Índice Dietético, Colesterol, Triglicéridos .....	110
Ilustración 17: Métricas Estadísticas Histogramas 1.....	110

Ilustración 18: Distribución de LDL, HDL, Presión Sistólica, Presión Diastólica .....	114
Ilustración 19: Métricas Estadísticas Histogramas 2.....	114
Ilustración 20: Distribución de Nivel de Estrés, Tiempo de Respuesta ante Emergencia, Ingresos Anuales .....	118
Ilustración 21: Métricas Estadísticas Histogramas 3.....	119
Ilustración 22: Distribución Variables Binarias .....	123
Ilustración 23: Distribución Variables Binarias con Respecto al Género ...	127
Ilustración 24: Distribución Variables Binarias con Respecto al Riesgo de Ataque Cardíaco .....	132
Ilustración 25: Distribución Condiciones Médicas.....	136
Ilustración 26: Distribución Factores Preventivos y Recursos Médicos.....	139
Ilustración 27: Distribución de Ataques Cardíacos por Grupo de Edad, Índice Dietético y Nivel de Estrés .....	142
Ilustración 28: Distribución de Ataques Cardíacos por Tiempo de Respuesta ante Emergencia y Nivel de Ingreso Anual .....	145
Ilustración 29: Boxplots de Todas las Variables Numéricas .....	148
Ilustración 30: Top 10 Estados con Más Pacientes .....	153
Ilustración 31: Top 10 Estados con Más Pacientes CON y SIN Riesgo de Ataque Cardíaco .....	157
Ilustración 32: Top 10 Combinaciones de Factores Protectores .....	161
Ilustración 33: Top 10 Perfiles Más Comunes de Pacientes en Riesgo .....	165
Ilustración 34: Porcentaje de Pacientes CON y SIN Seguro Médico por Nivel de Ingreso .....	169

Ilustración 35: Porcentaje de Pacientes con CON y SIN Acceso a Cuidados Médicos por Nivel de Ingreso .....	172
Ilustración 36: Diagramas de Dispersion, Edad vs Colesterol y Estrés vs Presión Sistólica por Riesgo de Ataque Cardíaco .....	176
Ilustración 37: Diagramas de Dispersion, LDL vs HDL y Presión Sistólica vs Presión Diastólica por Riesgo de Ataque Cardíaco.....	179
Ilustración 38: Diagramas Circulares de Grupos (Edad, Dieta, Estrés, Tiempo Respuesta e Ingreso Anual) .....	183
Ilustración 39: Resultados Random Forest.....	209
Ilustración 40: Gráficos Random Forest .....	211
Ilustración 41: Resultados XGBoost.....	211
Ilustración 42: Gráficos XGBoost .....	213
Ilustración 43: Resultados SVM.....	213
Ilustración 44: Gráficos SVM .....	215
Ilustración 45: Resultados LightGBM .....	215
Ilustración 46: Gráficos LightGBM.....	217
Ilustración 47: Resultados Red Neuronal Profunda Densa.....	217
Ilustración 48: Gráficos Red Neuronal Profunda Densa .....	219
Ilustración 49: Resultados CNN.....	219
Ilustración 50: Gráficos CNN.....	221
Ilustración 51: Resultados ResNet .....	221
Ilustración 52: Gráficos ResNet .....	223
Ilustración 53: Resultados TabTransformer .....	224

Ilustración 54: Gráficos TabTransformer .....	225
Ilustración 55: Resultados FT-Transformer .....	226
Ilustración 56: Gráficos FT-Transformer.....	227
Ilustración 57: Resultados SAINT .....	228
Ilustración 58: Interfaz de Usuario - Paciente 18.....	236
Ilustración 59: Interfaz de Usuario - Paciente 20.....	236
Ilustración 60: Informe PDF - Paciente 20 .....	240
Ilustración 61: Random Forest.....	244
Ilustración 62: Interfaz de Usuario.....	246

## ÍNDICE DE TABLAS

Tabla 1: Comparativa de Aspectos Clave de Cada Estudio ..... 49

## 1. INTRODUCCIÓN

En las últimas décadas, la convergencia entre la inteligencia computacional y las ciencias médicas ha catalizado transformaciones paradigmáticas, particularmente en el dominio de la salud cardiovascular. La capacidad inherente de los sistemas inteligentes para procesar volúmenes masivos de información y extraer patrones latentes ha inaugurado horizontes inéditos en el diagnóstico precoz, la medicina preventiva y la personalización terapéutica.

En este contexto, mi investigación se focaliza en el desarrollo e implementación de arquitecturas computacionales avanzadas destinadas a la identificación temprana de individuos con riesgo elevado de padecer eventos coronarios agudos, aprovechando datos clínicos y parámetros biométricos provenientes del subcontinente indio.

El corpus de datos empleado ha demandado un proceso meticuloso de acondicionamiento con el propósito de consolidar una base informacional coherente, depurada y estructurada, idónea para la aplicación de metodologías de ciencia de datos y modelado predictivo. En su estado inicial, los registros presentaban heterogeneidades, lagunas informacionales y una arquitectura subóptima para el análisis computacional. Tras completar el preprocesamiento, he conseguido un dataset optimizado para el desarrollo de sistemas basados en aprendizaje automático y profundo.

La comprensión de los mecanismos subyacentes que precipitan los eventos cardiovasculares resulta fundamental para desarrollar instrumentos de intervención temprana con potencial para preservar vidas. Los enfoques convencionales han evidenciado limitaciones significativas al intentar identificar interacciones no lineales o multidimensionales entre múltiples indicadores clínicos. Por esta razón, en mi proyecto he implementado algoritmos como regresión logística, bosques aleatorios y arquitecturas neuronales, además de explorar modelos más sofisticados como redes convolucionales y, cuando la naturaleza de los datos lo justifique,

arquitecturas basadas en transformers. Estas metodologías facilitan la captura de patrones intrincados y la generación de predicciones precisas sobre el riesgo cardiovascular de los pacientes.

El propósito central de esta investigación radica en identificar los determinantes fundamentales del riesgo de infarto y desarrollar sistemas capaces de estratificar a los pacientes con alta precisión. De esta manera, busco proporcionar a los profesionales sanitarios herramientas que faciliten la toma de decisiones proactivas y personalizadas, optimizando así la efectividad de las medidas preventivas. Este trabajo aporta además al campo de la salud digital, proporcionando un enfoque fundamentado en datos que puede escalarse e implementarse en diversos sistemas de atención médica.

Los registros utilizados en este estudio corresponden a información recopilada durante el período posterior al inicio de la pandemia por COVID-19, un evento que ha evidenciado la urgencia de fortalecer la prevención y monitorización de patologías crónicas, particularmente las cardiovasculares. La crisis sanitaria global ha subrayado la importancia de disponer de sistemas inteligentes capaces de anticipar riesgos mediante herramientas tecnológicas eficaces. En este sentido, mi proyecto aspira a satisfacer esta necesidad, desarrollando un modelo predictivo aplicable en diversos contextos médicos para mejorar la detección precoz de riesgos cardíacos.

En cuanto a los resultados esperados, mi objetivo es construir un modelo robusto que ofrezca predicciones confiables y fácilmente interpretables para su aplicación en entornos clínicos. La implementación de estas soluciones podría transformar los paradigmas de diagnóstico preventivo, estableciendo nuevos estándares en la aplicación de la inteligencia computacional para mejorar la salud cardiovascular y, por extensión, la calidad de vida de los pacientes.

Este proyecto representa una contribución significativa en la aplicación de la inteligencia artificial al ámbito de la medicina preventiva. Al establecer enfoques innovadores en la personalización del diagnóstico, aspiro a contribuir al mejoramiento de los sistemas sanitarios y a la reducción de la mortalidad asociada a patologías cardíacas.

## 2. ESTADO DEL ARTE

El progreso de la inteligencia computacional ha posibilitado el desarrollo de instrumentos cada vez más sofisticados para el análisis de información médica y la predicción de patologías. Específicamente, la aplicación de técnicas de aprendizaje automático y profundo en el sector sanitario ha adquirido una relevancia considerable en años recientes, particularmente en la prevención de afecciones cardiovasculares, que persisten como una de las causas principales de mortalidad global. La capacidad de identificar precozmente a individuos con riesgo de padecer un evento coronario resulta esencial para implementar intervenciones preventivas que disminuyan tanto la carga asistencial como el impacto en la calidad de vida de los pacientes.

Este apartado tiene como propósito presentar un análisis de los trabajos más significativos relacionados con la aplicación de sistemas inteligentes en la predicción del riesgo cardiovascular. Examinaré investigaciones recientes que han empleado diversas técnicas de aprendizaje supervisado, arquitecturas neuronales profundas y modelos especializados como los transformers, así como estudios focalizados en el análisis de variables clínicas y biométricas mediante enfoques de ciencia de datos. Adicionalmente, revisaré las principales fuentes de información utilizadas en este tipo de investigaciones, las metodologías de evaluación aplicadas y los desafíos recurrentes en términos de calidad, disponibilidad y representatividad de los datos.

A través de esta revisión, busco contextualizar mi proyecto dentro del panorama actual de la investigación en salud digital, identificar las principales corrientes de trabajo existentes y fundamentar la elección metodológica adoptada. Esta base teórica servirá como punto de partida para el diseño e implementación de los modelos desarrollados en este trabajo.

## 2.1. Contexto

A pesar de los progresos en medicina preventiva y del incremento en el uso de tecnologías orientadas al diagnóstico temprano, las patologías cardiovasculares persisten como una de las causas principales de mortalidad global. Específicamente, los infartos de miocardio constituyen un reto apremiante para los sistemas sanitarios debido a su elevada prevalencia y al hecho de que numerosos casos podrían haberse prevenido mediante detección oportuna y seguimiento adecuado de los factores de riesgo. Una de las problemáticas fundamentales reside en la complejidad para identificar oportunamente a aquellos individuos con mayor probabilidad de experimentar un evento cardiovascular, especialmente cuando las manifestaciones clínicas no se presentan de forma evidente o directa.

En este contexto, la inteligencia computacional emerge como una solución prometedora. Los sistemas avanzados de IA, particularmente aquellos fundamentados en técnicas de aprendizaje automático y profundo, brindan la posibilidad de analizar volúmenes masivos de información clínica, detectar patrones complejos y generar predicciones precisas sobre el estado de salud de los pacientes. No obstante, para implementar eficazmente estos modelos, resulta fundamental comprender qué variables influyen en el desarrollo de patologías cardíacas y desarrollar algoritmos que se adapten a la naturaleza y heterogeneidad de los datos médicos.

Además, tras la pandemia de COVID-19, la atención hacia la salud cardiovascular ha adquirido una relevancia aún mayor. El impacto del virus en personas con afecciones cardíacas preexistentes ha puesto de manifiesto la necesidad de contar con herramientas tecnológicas que permitan anticiparse a este tipo de situaciones y fortalecer las estrategias preventivas. Por consiguiente, surge una oportunidad para capitalizar el potencial de la IA en el análisis de datos clínicos con el propósito de mejorar la predicción y gestión del riesgo cardíaco.

Por otra parte, diversos estudios recientes han demostrado que la utilización de modelos predictivos basados en IA mejora la capacidad de detección temprana y facilita una intervención más personalizada y oportuna. Al aplicar algoritmos a conjuntos de datos sanitarios, es posible identificar con mayor precisión aquellos factores que incrementan el riesgo de infarto, lo cual resulta fundamental para diseñar protocolos de actuación adaptados a cada paciente. Este nivel de personalización es clave para optimizar los recursos sanitarios y reducir la incidencia de eventos cardiovasculares graves.

Mi proyecto se enfoca en integrar modelos avanzados de inteligencia artificial al análisis de datos clínicos con el propósito de identificar patrones de riesgo en pacientes y facilitar la toma de decisiones médicas. A través del estudio de un conjunto de datos procedente de la India, pretendo analizar qué variables resultan más relevantes en la predicción del riesgo cardíaco, con el objetivo de ofrecer un modelo fiable, escalable y útil en contextos clínicos reales.

## **2.2. Justificación**

La implementación de sistemas de inteligencia computacional en el ámbito sanitario no solo posee el potencial de optimizar la detección temprana de patologías cardiovasculares, sino que también puede representar una contribución sustancial al campo de la salud digital. Al proporcionar recomendaciones fundamentadas en datos y facilitar la identificación de pacientes en situación de vulnerabilidad, se promueve una medicina más proactiva, individualizada y eficiente. Esto conlleva un beneficio directo para cada individuo, pero también un impacto positivo a nivel de salud pública, al reducir la incidencia de eventos cardiovasculares graves y optimizar los recursos del sistema sanitario.

Los algoritmos de inteligencia artificial permiten realizar un análisis pormenorizado y continuo de la información clínica, lo que facilita la identificación de patrones y correlaciones que difícilmente podrían ser detectados mediante métodos convencionales. Además, aunque este proyecto

no contempla el análisis en tiempo real, la posibilidad de aplicar estos modelos en entornos dinámicos permitiría adaptar los protocolos de seguimiento clínico a las necesidades cambiantes de cada paciente, incrementando así su eficacia.

Asimismo, la integración de IA en procesos de diagnóstico y prevención puede propiciar el desarrollo de estrategias más avanzadas para la gestión del riesgo cardiovascular. Por ejemplo, mediante el análisis predictivo, es factible anticipar qué pacientes presentan una mayor probabilidad de sufrir un infarto, permitiendo una intervención temprana y dirigida. Este enfoque proactivo no solo mejora los resultados clínicos, sino que también contribuye a reducir la carga sobre los profesionales sanitarios y los sistemas de atención médica.

En resumen, este proyecto tiene el potencial de establecer un nuevo paradigma en la aplicación de la inteligencia artificial en medicina preventiva, al utilizar modelos avanzados para detectar de forma precisa y anticipada el riesgo de sufrir ataques cardíacos. De este modo, no solo se mejora la experiencia del paciente y la toma de decisiones médicas, sino que también se promueve una mayor adopción de soluciones tecnológicas en el ámbito de la salud, contribuyendo al bienestar general de la población.

### **2.3. Análisis comparativo de estudios existentes**

Este apartado presenta una comparativa entre diversos estudios científicos y proyectos reales centrados en la aplicación de la inteligencia artificial para la predicción del riesgo de enfermedades cardiovasculares, en particular, ataques cardíacos. El objetivo es mostrar de forma clara el estado actual de la investigación en este ámbito, identificar los avances alcanzados por otros autores y destacar las aportaciones diferenciales que este proyecto propone. Para ello, el análisis se dividirá en varias subcategorías que permitan una revisión exhaustiva y organizada de los trabajos relacionados.

### 2.3.1. Revisión general de estudios existentes

En años recientes, la aplicación de sistemas inteligentes en el dominio de la salud ha suscitado un interés creciente, especialmente en la predicción de eventos cardiovasculares mediante el análisis de información clínica y parámetros biométricos. Múltiples investigaciones han abordado esta problemática desde diversas perspectivas, implementando algoritmos de aprendizaje automático y profundo para anticipar el riesgo cardíaco en pacientes. Esta sección proporciona una perspectiva integral de los trabajos más relevantes en este campo, analizando sus metodologías, conjuntos de datos empleados, resultados obtenidos y su aplicabilidad en entornos reales.

#### **El rol de la IA en la predicción de ataques cardíacos**

La aplicación de la inteligencia computacional en la predicción de eventos coronarios ha experimentado una evolución significativa en los últimos años, consolidándose como un instrumento fundamental para el diagnóstico temprano y la toma de decisiones clínicas. La disponibilidad de información médica proveniente de bases públicas, instituciones hospitalarias y centros académicos ha facilitado el desarrollo de modelos capaces de identificar patrones complejos que pueden pasar inadvertidos mediante métodos clínicos convencionales.

Uno de los trabajos más relevantes en este ámbito es el de (Alexander & Wang, 2017), quienes efectuaron una revisión sobre el empleo de tecnologías de Big Data para la predicción y gestión de enfermedades cardiovasculares. Esta investigación subraya la importancia de integrar técnicas como minería de datos, Hadoop y visualización de información a gran escala, destacando cómo estas herramientas pueden personalizar tratamientos médicos y anticipar eventos cardíacos críticos a partir del análisis de datos masivos.

Por otro lado, (Tacki, 2018) abordó la mejora en la predicción de ataques cardíacos mediante la combinación de algoritmos de selección de

características y métodos de aprendizaje automático. Utilizando el dataset Statlog (Heart), demostró que la integración del algoritmo de máquinas de vectores de soporte (SVM) con kernel lineal y el método de selección ReliefF alcanzó una precisión del 84,81%, evidenciando la influencia de una adecuada ingeniería de características en el rendimiento del modelo.

El estudio de (Patil & Kumaraswamy, 2009) propuso una metodología basada en minería de datos para extraer patrones significativos desde almacenes clínicos. Mediante la combinación de clustering con K-means y el algoritmo MAFIA para la identificación de itemsets frecuentes, se desarrolló un sistema de predicción que resalta el valor del conocimiento oculto en grandes volúmenes de datos médicos.

(Alshraideh, y otros, 2024) exploraron distintos algoritmos de IA aplicados a datos del Hospital Universitario de Jordania. Utilizando clasificadores como Random Forest, SVM y KNN junto con Particle Swarm Optimization (PSO) para seleccionar las variables más relevantes, obtuvieron una precisión destacada del 94,3% con SVM+PSO. Este resultado refleja el potencial de la IA para alcanzar niveles de precisión competitivos en contextos clínicos reales.

Desde una perspectiva más técnica, (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021) aplicaron un enfoque de pipeline de aprendizaje automático sobre conjuntos de datos del Framingham Heart Study y del repositorio UCI. Comparando modelos como Gradient Boosting, Árboles de Decisión y Regresión Logística, identificaron al Gradient Boosting como el más eficaz, y destacaron variables como tipo de dolor torácico, colesterol y frecuencia cardíaca como las de mayor influencia en la predicción.

(Nandal, Goel, & Tanwar, 2022) desarrollaron el método ML-HAP, centrado en la predicción sintomática de ataques cardíacos mediante modelos como SVM, XGBoost y Regresión Logística. Su estudio mostró que XGB ofrecía el mejor rendimiento, con un AUC de 0,94, reafirmando la utilidad de los

algoritmos de boosting en tareas de clasificación médica altamente específicas.

En una línea similar, (Feng, y otros, 2024) propusieron un enfoque centrado en el uso del algoritmo XGBoost, acompañado de una rigurosa selección de características y optimización de hiperparámetros. Su modelo demostró ser altamente interpretable y eficaz, superando a otros métodos tanto en precisión como en escalabilidad, lo cual es esencial para su implementación práctica en sistemas de salud.

Finalmente, (Waqar, y otros, 2021) introdujeron un modelo basado en redes neuronales profundas integradas con la técnica SMOTE para abordar el desbalance de clases en los datos. Su enfoque evitó la necesidad de ingeniería de características manual, lo que lo convierte en una solución eficiente y rentable. Este modelo destacó por su alta fiabilidad en la clasificación, lo que refuerza su aplicabilidad clínica incluso en contextos con recursos limitados.

En conjunto, estos estudios demuestran que la IA no solo puede mejorar significativamente la precisión de las predicciones sobre ataques cardíacos, sino que también permite un enfoque más personalizado y proactivo en el tratamiento de enfermedades cardiovasculares. Desde el uso de algoritmos clásicos de clasificación hasta arquitecturas avanzadas de aprendizaje profundo y técnicas de optimización, la inteligencia artificial está sentando las bases para una nueva era de medicina predictiva y preventiva.

### **Modelos de IA y su aplicación en la predicción de ataques cardíacos**

La aplicación de modelos de inteligencia artificial en la predicción de ataques cardíacos es diversa, y los distintos estudios han empleado técnicas variadas según la naturaleza de los datos disponibles y los objetivos específicos de cada investigación.

El Estudio 1 (Alexander & Wang, 2017) adoptó un enfoque orientado al análisis masivo de datos clínicos, proponiendo el uso de herramientas de Big Data y técnicas de minería de datos para mejorar la predicción de enfermedades cardiovasculares. Este estudio no se centró en un modelo específico, sino en cómo las tecnologías emergentes pueden combinarse para desarrollar sistemas inteligentes de diagnóstico y prevención a gran escala. Su principal aportación reside en establecer un marco conceptual para el uso de IA en salud, destacando la importancia de la infraestructura tecnológica y la integración de fuentes heterogéneas de datos.

El Estudio 2 (Tacki, 2018) exploró la combinación de algoritmos de aprendizaje automático con técnicas de selección de características. Utilizando el dataset Statlog (Heart), identificó que el algoritmo SVM con kernel lineal junto con la técnica ReliefF ofrecía la mayor precisión en la predicción de ataques cardíacos. Este estudio resalta la relevancia de seleccionar correctamente las variables más influyentes para mejorar la eficacia de los modelos, demostrando que una combinación adecuada de algoritmos puede superar a enfoques más complejos si se aplica un preprocesamiento óptimo.

El Estudio 3 (Patil & Kumaraswamy, 2009) propuso una arquitectura basada en minería de datos para identificar patrones significativos en grandes almacenes de datos clínicos. Utilizó K-means para el agrupamiento inicial y el algoritmo MAFIA para extraer conjuntos de elementos frecuentes relacionados con el riesgo cardíaco. Este enfoque es especialmente útil cuando se dispone de bases de datos extensas y no estructuradas, ya que permite transformar la información en reglas clínicas relevantes para la predicción de infartos.

El Estudio 4 (Alshraideh, y otros, 2024) desarrolló un sistema de predicción combinando varios modelos supervisados (SVM, KNN, Random Forest, Naive Bayes, Árboles de Decisión) con un algoritmo de optimización basado en enjambre de partículas (PSO) para la selección de variables. La combinación de SVM con PSO logró una precisión del 94,3%, subrayando cómo la

integración de técnicas de optimización con clasificadores tradicionales puede potenciar su rendimiento en contextos médicos reales.

El Estudio 5 (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021) se centró en comparar distintos modelos supervisados (Gradient Boosting, Árboles de Decisión, Random Forest, Regresión Logística) aplicados a datos del Framingham Heart Study y UCI Repository. El modelo con mejores resultados fue Gradient Boosting, especialmente al combinarse con técnicas de transformación de características. Este estudio remarca la eficacia de los modelos basados en boosting para tareas de predicción binaria con múltiples variables clínicas.

El Estudio 6 (Nandal, Goel, & Tanwar, 2022) presentó el método ML-HAP, en el que se evaluaron modelos como SVM, XGBoost, Regresión Logística y Naive Bayes para la predicción de síntomas de ataques cardíacos. XGBoost fue el modelo con mayor rendimiento, alcanzando un AUC de 0,94. El estudio refuerza la utilidad de los algoritmos de boosting para capturar relaciones no lineales complejas y destaca su eficiencia para identificar síntomas en etapas tempranas.

El Estudio 7 (Feng, y otros, 2024) propuso un enfoque integral para la predicción de infartos utilizando XGBoost, incluyendo preprocesamiento, selección de características y optimización de hiperparámetros. Su modelo destacó por combinar alta precisión con interpretabilidad, algo esencial en aplicaciones clínicas. Este estudio demuestra que modelos de boosting pueden ser tan potentes como las redes neuronales, con la ventaja de ofrecer mayor transparencia en los resultados.

El Estudio 8 (Waqar, y otros, 2021) desarrolló un modelo de aprendizaje profundo basado en SMOTE para resolver el desbalance de clases en conjuntos de datos cardíacos. A diferencia de otros enfoques, evitó el uso de ingeniería de características manual, apostando por un modelo de aprendizaje de extremo a extremo. Al aplicar una red neuronal artificial bien ajustada

sobre datos balanceados, se logró un sistema eficiente, fiable y aplicable a escenarios con recursos computacionales limitados.

En conjunto, los estudios revisados muestran una amplia variedad de enfoques en la aplicación de IA para la predicción de ataques cardíacos. Desde métodos clásicos y fácilmente interpretables hasta técnicas avanzadas de boosting y deep learning, todos coinciden en la importancia de un buen preprocesamiento de datos, la selección de características relevantes y la necesidad de modelos personalizados según el contexto clínico específico.

### **Desafíos y limitaciones en los estudios existentes**

Si bien los estudios revisados demuestran el potencial de la inteligencia artificial para predecir ataques cardíacos con alta precisión, también revelan una serie de desafíos y limitaciones que afectan la eficacia, aplicabilidad y generalización de los modelos desarrollados.

Uno de los principales desafíos es la variabilidad en los datos clínicos utilizados, los cuales difieren considerablemente en cuanto a origen geográfico, calidad, cantidad y estructura. Esta variabilidad puede impactar directamente en la precisión de los modelos de IA. Por ejemplo, en el Estudio 4 (Alshraideh, y otros, 2024), se identificó una notable dependencia del conjunto de datos local del Hospital Universitario de Jordania, lo que limita la capacidad de generalización del modelo a otras poblaciones con características demográficas distintas.

Otro problema recurrente es el desbalance de clases en los datos clínicos, es decir, la presencia significativamente menor de casos positivos (pacientes que realmente sufren un ataque cardíaco) en comparación con los negativos. Este fenómeno puede sesgar el aprendizaje del modelo. En el Estudio 8 (Waqar, y otros, 2021), se resolvió este problema aplicando la técnica SMOTE, lo que permitió equilibrar el conjunto de datos y mejorar la fiabilidad de las predicciones.

La dependencia de datos clínicos limitados o de acceso restringido también representa una limitación crítica. Algunos estudios, como el Estudio 1 (Alexander & Wang, 2017) y el Estudio 3 (Patil & Kumaraswamy, 2009), utilizaron almacenes de datos extensos, pero en la práctica muchas instituciones no cuentan con repositorios tan completos, lo cual complica la replicación de estos modelos en entornos reales.

Asimismo, la complejidad en la integración de datos heterogéneos (biométricos, clínicos, ambientales y de estilo de vida) supone un reto técnico considerable. En el Estudio 5 (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021), si bien se obtuvieron buenos resultados con modelos como Gradient Boosting, también se reconoce que la incorporación de múltiples tipos de datos requiere procesos avanzados de preprocesamiento y transformación, lo cual puede no ser viable en entornos clínicos con recursos limitados.

Otro punto importante es la falta de interpretabilidad en algunos modelos complejos, especialmente en aquellos basados en deep learning. Aunque el Estudio 6 (Nandal, Goel, & Tanwar, 2022) y el Estudio 7 (Feng, y otros, 2024) demostraron un rendimiento superior con algoritmos como XGBoost, también señalaron la dificultad para explicar de forma clara cómo estos modelos toman decisiones, lo cual puede afectar la confianza de los profesionales médicos en su uso clínico.

Las limitaciones computacionales constituyen otro obstáculo, especialmente en regiones con menor infraestructura tecnológica. El Estudio 2 (Tacki, 2018) reconoció que el entrenamiento de modelos con múltiples algoritmos y técnicas de selección de características puede requerir recursos computacionales avanzados, lo que dificulta su implementación en tiempo real en contextos hospitalarios convencionales.

Por último, aunque no todos los estudios lo abordan en profundidad, las consideraciones éticas y de privacidad también juegan un papel clave en la aplicación de modelos de IA en salud. La protección de datos sensibles, la

transparencia de los algoritmos y el consentimiento informado son aspectos fundamentales que deben ser atendidos para garantizar la confianza y seguridad de los usuarios, especialmente cuando se manejan historiales clínicos completos o se implementan sistemas de predicción automatizados.

Estos desafíos y limitaciones subrayan la necesidad de desarrollar modelos de IA que sean no solo precisos, sino también escalables, interpretables, éticos y adaptables a diversos entornos clínicos. Superar estas barreras será esencial para lograr una integración efectiva y responsable de la inteligencia artificial en la práctica médica.

### **Hallazgos clave e implicaciones para futuras investigaciones**

Los estudios revisados evidencian el considerable potencial de la inteligencia artificial para predecir ataques cardíacos, aportando hallazgos clave que ofrecen una base sólida para futuras investigaciones en el ámbito de la salud digital. En términos generales, los modelos de IA, especialmente aquellos que emplean técnicas de aprendizaje profundo y algoritmos de boosting, han demostrado un alto grado de precisión, con algunos trabajos como el Estudio 4 (Alshraideh, y otros, 2024) reportando tasas de acierto superiores al 94% al combinar SVM con optimización mediante enjambre de partículas (PSO).

Uno de los hallazgos más consistentes es la importancia de la selección de características relevantes en el desempeño de los modelos. Estudios como el Estudio 2 (Tacki, 2018) y el Estudio 5 (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021) destacaron que la combinación de buenas prácticas de preprocessamiento con algoritmos potentes, como Gradient Boosting o SVM, mejora significativamente la capacidad predictiva. Esto subraya la necesidad de aplicar técnicas de ingeniería de características y optimización de hiperparámetros de forma cuidadosa en futuras investigaciones.

La personalización de los modelos también se perfila como un factor clave. Modelos que consideran las diferencias individuales entre pacientes, ya sean

demográficas, clínicas o conductuales, ofrecen mejores resultados en la predicción del riesgo cardíaco. En este sentido, el Estudio 8 (Waqar, y otros, 2021) resalta cómo un enfoque adaptado con redes neuronales profundas y técnicas de balanceo de clases (como SMOTE) puede ofrecer predicciones más fiables y generalizables, incluso sin aplicar ingeniería de características manual.

Otro aspecto destacado es la eficacia de los algoritmos de boosting. Tanto el Estudio 6 (Nandal, Goel, & Tanwar, 2022) como el Estudio 7 (Feng, y otros, 2024) demostraron que modelos como XGBoost son especialmente eficaces para capturar patrones complejos en datos clínicos, logrando altos niveles de precisión ( $AUC \geq 0,94$ ). Su rendimiento, junto con su interpretabilidad relativa, los posiciona como candidatos ideales para aplicaciones clínicas prácticas.

Las implicaciones para futuras investigaciones también incluyen la necesidad de mejorar la disponibilidad y diversidad de los conjuntos de datos. La mayoría de los estudios trabajaron con bases de datos limitadas en tamaño o alcance geográfico, lo cual restringe la capacidad de generalización. Ampliar estas bases de datos, integrando datos de diferentes regiones y con mayor diversidad poblacional, podría incrementar la robustez de los modelos desarrollados.

Además, conforme los modelos de IA se acercan a una adopción real en entornos hospitalarios, las implicaciones éticas cobran una importancia central. Aunque solo algunos estudios, como el Estudio 1 (Alexander & Wang, 2017), mencionan indirectamente la necesidad de garantizar la transparencia y protección de datos, es fundamental que futuras investigaciones aborden estos aspectos de forma sistemática, incluyendo la elaboración de marcos éticos para el uso responsable de la IA en salud.

En definitiva, los estudios analizados demuestran que los modelos de IA bien diseñados, entrenados y validados tienen el potencial de transformar la

medicina preventiva, especialmente en la detección temprana de enfermedades cardiovasculares. Futuras líneas de investigación deberían centrarse en mejorar la personalización, integrar fuentes de datos heterogéneas y garantizar la implementación ética y segura de estas tecnologías en sistemas clínicos reales.

## Conclusión

La conclusión de esta revisión pone de manifiesto los importantes avances logrados en la aplicación de la inteligencia artificial para la predicción de ataques cardíacos, subrayando su potencial transformador en el ámbito de la salud digital. Los modelos de IA, en especial aquellos que incorporan técnicas de aprendizaje profundo y algoritmos de boosting, han demostrado una capacidad destacada para ofrecer predicciones precisas, fiables y personalizadas, lo cual resulta fundamental en el contexto clínico, donde el tiempo de respuesta y la precisión diagnóstica pueden marcar la diferencia en la evolución del paciente.

La integración de fuentes de datos diversas, como información clínica, biométrica y de historial médico, se perfila como un componente esencial para el desarrollo de modelos predictivos robustos y completos. Estudios como el Estudio 5 (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021) y el Estudio 3 (Patil & Kumaraswamy, 2009) han demostrado que dicha integración mejora significativamente la solidez de las predicciones, lo que pone en evidencia la necesidad de enfoques multifactoriales en el diseño de futuros modelos de IA.

A medida que la IA continúa evolucionando, la formulación y el cumplimiento de principios éticos adquieren una relevancia cada vez mayor. Abordar cuestiones como la privacidad de los datos, la transparencia de los algoritmos y el respeto por la autonomía del paciente es esencial para garantizar que las soluciones basadas en IA sean utilizadas de forma responsable y ética, generando confianza tanto en los profesionales sanitarios como en los pacientes. Aunque no todos los estudios tratan explícitamente estas

cuestiones, su incorporación se vuelve indispensable en investigaciones futuras orientadas a la implementación real de estas tecnologías.

De cara al futuro, la investigación debería priorizar el desarrollo de modelos de IA más sofisticados, capaces de integrar múltiples tipos de datos y adaptarse a las características individuales de los pacientes. Además, será crucial explorar en profundidad las implicaciones éticas de estas tecnologías, así como establecer marcos regulatorios y de buenas prácticas que aseguren su uso justo y seguro en contextos clínicos.

Partiendo de los aprendizajes extraídos de estos estudios, este proyecto busca contribuir a este campo mediante el diseño e implementación de modelos de inteligencia artificial que no solo sean precisos, sino también interpretables, éticos y centrados en el paciente. En última instancia, el objetivo es reforzar la medicina preventiva a través de herramientas predictivas que mejoren la toma de decisiones médicas y ayuden a reducir la mortalidad asociada a enfermedades cardiovasculares. Esta revisión sienta las bases para un análisis comparativo más profundo, centrado en las metodologías empleadas, los conjuntos de datos utilizados, el nivel de personalización alcanzado y las posibles aplicaciones en entornos reales.

### **2.3.2. Metodologías empleadas en investigaciones previas**

Este apartado ofrece un análisis pormenorizado de los algoritmos y modelos implementados en los estudios revisados sobre predicción de eventos coronarios mediante inteligencia artificial. Se contrastan también las técnicas de IA aplicadas, destacando su eficacia en función del tipo de datos y los objetivos de investigación.

## Examen detallado de los algoritmos y modelos

- Estudio 1: (Alexander & Wang, 2017)

Este estudio propone una arquitectura conceptual fundamentada en herramientas de Big Data para la predicción de enfermedades cardíacas. Aunque no se implementa un modelo concreto, se identifican tecnologías como Hadoop, minería de datos y visualización avanzada como elementos cruciales para la creación de modelos predictivos a gran escala. Big Data y minería de datos. Esta aproximación permite gestionar volúmenes masivos de datos heterogéneos y sienta las bases para el desarrollo de modelos de IA aplicables a nivel nacional o institucional.

- Estudio 2: (Tacki, 2018)

Se analizaron múltiples combinaciones entre algoritmos de clasificación y métodos de selección de características sobre el conjunto de datos Statlog. La combinación más eficaz fue SVM con núcleo lineal y la técnica ReliefF. SVM + Selección de características. Este enfoque demostró una precisión superior al 84%, mostrando cómo la ingeniería de características mejora la capacidad predictiva cuando se dispone de variables clínicas estructuradas.

- Estudio 3: (Patil & Kumaraswamy, 2009)

Aplicaron técnicas de agrupamiento con K-means y el algoritmo MAFIA para extraer patrones frecuentes en almacenes de datos clínicos. K-means + MAFIA. Su metodología permite descubrir reglas clínicas significativas que no son evidentes mediante análisis convencionales, resultando útil para la detección de patrones relevantes en grandes volúmenes de datos.

- Estudio 4: (Alshraideh, y otros, 2024)

Se desarrolló un sistema predictivo utilizando múltiples clasificadores (SVM, KNN, Naive Bayes, Árboles de Decisión), combinados con Particle Swarm Optimization (PSO) para la selección de características. SVM + PSO. El modelo alcanzó una precisión del 94,3%, demostrando que los clasificadores tradicionales pueden mejorar su rendimiento significativamente al integrarse con algoritmos de optimización.

- Estudio 5: (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021)

Este estudio comparó Gradient Boosting, Random Forest, Árboles de Decisión y Regresión Logística aplicados a datos del Framingham y UCI. Gradient Boosting. Fue el algoritmo más preciso, particularmente eficaz para conjuntos de datos binarios y clínicos. Su capacidad para manejar interacciones no lineales complejas lo convierte en un modelo robusto para tareas médicas predictivas.

- Estudio 6: (Nandal, Goel, & Tanwar, 2022)

Se desarrolló el método ML-HAP utilizando modelos como XGBoost, SVM, Naive Bayes y Regresión Logística sobre datos sintomáticos de enfermedades cardíacas. XGBoost. Superó al resto de modelos, logrando un AUC de 0,94. Destaca por su capacidad para capturar patrones no lineales y manejar variables clínicas correlacionadas de forma eficiente.

- Estudio 7: (Feng, y otros, 2024)

El estudio propuso un modelo basado en XGBoost, acompañado de técnicas rigurosas de preprocesamiento, selección de atributos y optimización de hiperparámetros. XGBoost + Optimización. Esta metodología permitió un

modelo altamente preciso e interpretable, ideal para su adopción clínica debido a su escalabilidad y eficiencia computacional.

- Estudio 8: (Waqar, y otros, 2021)

Se desarrolló un modelo de deep learning basado en redes neuronales artificiales con datos balanceados mediante SMOTE. Red neuronal + SMOTE. El modelo evitó la ingeniería de características manual y demostró alta fiabilidad, siendo adecuado para entornos con recursos limitados o donde el desbalance de clases representa un reto importante.

### **Comparación de la eficacia de las técnicas de IA aplicadas**

La eficacia de cada técnica de IA depende en gran medida de la naturaleza del conjunto de datos y de los objetivos del estudio:

Big Data y minería de datos (Estudio 1) ofrecieron una base conceptual sólida para el diseño de arquitecturas predictivas, especialmente útiles en contextos donde se dispone de grandes volúmenes de datos no estructurados.

Técnicas tradicionales de machine learning (Estudios 2, 3 y 4) como SVM, árboles de decisión y regresión logística son eficaces en conjuntos de datos estructurados y clínicamente interpretables. Al combinarse con métodos de selección de características u optimización, su precisión mejora significativamente.

Modelos basados en boosting (Estudios 5, 6 y 7) como Gradient Boosting y XGBoost destacan por su precisión y capacidad para capturar relaciones no lineales. Son especialmente eficaces cuando se manejan datos clínicos con múltiples variables correlacionadas.

Redes neuronales profundas (Estudio 8) son apropiadas para contextos con datos desbalanceados o sin necesidad de realizar ingeniería de características manual. Ofrecen soluciones eficientes y adaptables, aunque con mayores requerimientos computacionales.

## **Conclusión**

Las metodologías empleadas en investigaciones previas para la predicción de ataques cardíacos reflejan una amplia variedad de técnicas de inteligencia artificial, cada una con ventajas y desafíos particulares. Los modelos tradicionales de machine learning, como los árboles de decisión, la regresión logística y las máquinas de soporte vectorial (SVM), ofrecen simplicidad y facilidad de interpretación, lo que los hace especialmente adecuados para conjuntos de datos estructurados y clínicamente comprensibles. Por otro lado, los modelos más avanzados, como XGBoost y las redes neuronales profundas, destacan en el procesamiento de datos complejos y en la detección de patrones no lineales o sutiles en variables clínicas altamente correlacionadas.

La eficacia de cada metodología depende en gran medida de las características específicas del conjunto de datos y de los objetivos de la investigación. A medida que la inteligencia artificial continúa evolucionando, los estudios futuros deberían centrarse en combinar diferentes técnicas de IA, aprovechando sus fortalezas complementarias para aumentar la precisión y la robustez de las predicciones sobre el riesgo cardíaco.

Partiendo de las metodologías analizadas, este proyecto tiene como objetivo desarrollar modelos avanzados de IA que no solo alcancen una alta precisión, sino que también sean interpretables, éticos y adaptables a entornos clínicos reales. El propósito final es contribuir a la mejora de la medicina preventiva, ofreciendo herramientas fiables y personalizadas que apoyen la toma de decisiones médicas y permitan identificar de forma temprana a los pacientes con mayor riesgo de sufrir un ataque cardíaco.

### 2.3.3. Conjuntos de datos utilizados en estudios relacionados

#### **Revisión de los conjuntos de datos empleados en la investigación existente**

Los estudios revisados han utilizado diferentes conjuntos de datos clínicos y biométricos para predecir ataques cardíacos, cada uno con características particulares adaptadas a los objetivos de cada investigación:

El Estudio 1 trabajó con volúmenes masivos de información clínica proveniente de instituciones sanitarias, integrando datos demográficos, históricos médicos y parámetros biométricos. Aunque el enfoque fue principalmente teórico, se enfatizó la necesidad de infraestructura para manejar datos a gran escala mediante tecnologías de Big Data como Hadoop.

El Estudio 2 empleó el dataset Statlog (Heart), ampliamente reconocido en el ámbito académico. Incluye 13 atributos clínicos estandarizados como edad, colesterol, presión arterial y tipo de dolor torácico. Este conjunto de datos fue fundamental para evaluar el rendimiento de distintos modelos combinados con técnicas de selección de características.

El Estudio 3 se centró en bases de datos clínicas no especificadas, organizadas en almacenes de datos con grandes volúmenes de información estructurada. El estudio priorizó la minería de patrones en registros médicos históricos, destacando su aplicabilidad a contextos hospitalarios con grandes sistemas de información clínica.

El Estudio 4 utilizó registros de pacientes del Hospital Universitario de Jordania, que incluían variables clínicas, demográficas y sintomáticas. Los datos fueron preprocesados y filtrados antes de ser aplicados a modelos supervisados, destacando por su origen real y su relevancia práctica para el desarrollo de herramientas clínicas predictivas.

El Estudio 5 se basó en dos conjuntos de datos bien conocidos: el Framingham Heart Study y el dataset de la UCI Machine Learning Repository. Ambos contienen información clínica extensa sobre pacientes con riesgo cardiovascular, incluyendo antecedentes familiares, IMC, colesterol, tabaquismo y consumo de medicamentos, lo que permitió validar la eficacia de diferentes algoritmos de predicción.

El Estudio 6 empleó un conjunto de datos recopilado específicamente para evaluar síntomas asociados a ataques cardíacos. Las variables incluyeron dolor torácico, niveles de azúcar, ECG y otros signos clínicos. La recolección de datos sintomáticos detallados permitió una predicción altamente específica utilizando técnicas como XGBoost.

El Estudio 7 se centró en un dataset clínico estructurado, al que se aplicaron rigurosos procesos de limpieza y normalización. Contenía registros médicos tradicionales, síntomas y pruebas diagnósticas. La investigación hizo énfasis en la optimización de hiperparámetros y selección de características para mejorar la precisión del modelo XGBoost.

El Estudio 8 utilizó el clásico dataset cardíaco del repositorio UCI, pero con modificaciones importantes. Se aplicó la técnica SMOTE para balancear las clases, evitando así el sesgo común en datasets médicos. Este conjunto de datos fue empleado para entrenar un modelo de red neuronal sin necesidad de realizar ingeniería de características manual.

### **Comparación de las características de los datos y su influencia en los resultados**

Las características de los conjuntos de datos utilizados en los estudios analizados desempeñan un papel clave en la efectividad de los modelos predictivos. En general, los conjuntos de datos más grandes y diversos, como los utilizados en el Estudio 1 y el Estudio 5, permiten una mejor

generalización y fortalecen la capacidad de los modelos para detectar patrones en poblaciones heterogéneas.

Por otro lado, los estudios que emplearon datos clínicos objetivos, como el Estudio 4, el Estudio 6 y el Estudio 7, tienden a lograr predicciones más precisas que aquellos basados en datos autodeclarados o limitados. Este tipo de datos proporciona una base más sólida para el entrenamiento de modelos, reduciendo el margen de error y aumentando la fiabilidad de los resultados.

El uso de datos temporales o secuenciales, aunque no central en estos estudios, se ve parcialmente reflejado en los conjuntos que incluyen historial médico longitudinal, como los del Estudio 3 y el Estudio 5, lo cual resulta crucial para analizar tendencias a lo largo del tiempo y anticipar eventos clínicos futuros.

Asimismo, la integración de múltiples fuentes de datos, como datos clínicos, demográficos y sintomáticos, evidenciada en estudios como el Estudio 6 y el Estudio 8, mejora la capacidad de los modelos para capturar la complejidad de las condiciones médicas reales. Esta combinación de datos permite un análisis más completo de los factores de riesgo que pueden desencadenar un ataque cardíaco.

Finalmente, algunos estudios como el Estudio 4 y el Estudio 8 abordan de manera explícita la importancia de garantizar la privacidad y el uso ético de los datos, aspectos esenciales para promover la aceptación de los sistemas basados en IA en entornos clínicos reales. La atención a estos factores no solo influye en el diseño del estudio, sino también en la viabilidad futura de su aplicación.

## Conclusión

Los conjuntos de datos empleados en los estudios revisados presentan una notable diversidad en cuanto a origen, estructura y contenido, lo cual influye de forma directa en los resultados obtenidos. En general, los datasets más amplios y diversos contribuyen al desarrollo de modelos más robustos y generalizables, mientras que la incorporación de datos objetivos y bien estructurados incrementa la precisión de las predicciones.

La combinación de distintas fuentes de datos, así como la consideración de aspectos éticos y de privacidad, aporta valor añadido a los modelos desarrollados, facilitando su aplicación práctica y sostenible. Estas observaciones servirán de guía en el desarrollo del modelo predictivo planteado en este proyecto, asegurando que sea tanto efectivo como clínicamente viable.

### 2.3.4. Resultados y hallazgos

#### Comparación de resultados entre los distintos estudios

Los estudios revisados muestran una amplia gama de resultados en cuanto a la eficacia de diferentes modelos y metodologías de IA para predecir ataques cardíacos. A continuación, se resumen los hallazgos más relevantes de cada uno:

##### Estudio 1: (Alexander & Wang, 2017)

- Modelos utilizados: Arquitectura conceptual de Big Data y minería de datos
- Resultados clave: El estudio no implementa modelos predictivos específicos, pero establece las bases teóricas para integrar tecnologías

de Big Data en el análisis de enfermedades cardiovasculares, destacando su potencial para la medicina personalizada.

- Desafíos: Necesidad de infraestructuras potentes y protocolos de integración entre datos clínicos dispares.

**Estudio 2:** (Tacki, 2018)

- Modelos utilizados: SVM + técnica de selección de características ReliefF
- Resultados clave: Se alcanzó una precisión del 84,81% en la predicción de ataques cardíacos. La combinación de algoritmos y técnicas de selección mejoró notablemente los resultados.
- Desafíos: Limitaciones por el tamaño del dataset y la dependencia de atributos clínicos básicos.

**Estudio 3:** (Patil & Kumaraswamy, 2009)

- Modelos utilizados: K-means + MAFIA para minería de patrones
- Resultados clave: Identificación de reglas clínicas relevantes mediante clustering y asociación. Resultados cualitativos útiles para descubrimiento de patrones.
- Desafíos: Requiere datasets muy estructurados y de gran volumen; no mide precisión de predicción de manera directa.

**Estudio 4:** (Alshraideh, y otros, 2024)

- Modelos utilizados: SVM, KNN, Árboles de Decisión, Naive Bayes + PSO
- Resultados clave: El mejor modelo (SVM + PSO) alcanzó un 94,3% de precisión. Se identificaron variables clave como el tipo de dolor torácico, ECG y presión arterial.

- Desafíos: Integración y limpieza de datos clínicos reales; variabilidad en las métricas entre algoritmos.

**Estudio 5:** (Gupta, Shrivastava, Upadhyay, & Chaurasia, 2021)

- Modelos utilizados: Gradient Boosting, Árboles de Decisión, Random Forest, Regresión Logística
- Resultados clave: Gradient Boosting obtuvo los mejores resultados, con métricas superiores al 90% de precisión. Se confirmaron variables como edad, colesterol y presión sistólica como altamente predictivas.
- Desafíos: Riesgo de overfitting en modelos complejos; necesidad de validación cruzada rigurosa.

**Estudio 6:** (Nandal, Goel, & Tanwar, 2022)

- Modelos utilizados: XGBoost, SVM, Naive Bayes, Regresión Logística
- Resultados clave: XGBoost logró un AUC de 0,94. Alta eficacia en datos sintomáticos con múltiples variables clínicas.
- Desafíos: Complejidad del modelo y necesidad de ajuste fino de hiperparámetros.

**Estudio 7:** (Feng, y otros, 2024)

- Modelos utilizados: XGBoost con selección de características y optimización de hiperparámetros
- Resultados clave: Alta precisión con interpretabilidad mejorada alrededor del 93%, destacando variables clave como niveles de azúcar, dolor torácico y ECG.
- Desafíos: Requiere etapas avanzadas de preprocesamiento y validación para evitar sesgos de modelo.

### **Estudio 8:** (Waqar, y otros, 2021)

- Modelos utilizados: Red neuronal artificial + SMOTE
- Resultados clave: El modelo superó a otros sistemas sin necesidad de realizar ingeniería de características. Alta fiabilidad en clasificación, en torno al 91% de precisión, con especial atención a datos desbalanceados.
- Desafíos: Mayor coste computacional; difícil interpretación para uso clínico directo.

#### **Identificación de tendencias comunes y hallazgos únicos**

- **Tendencias comunes**

Alta precisión con modelos avanzados: Modelos como XGBoost y redes neuronales demostraron consistentemente tasas de precisión superiores al 90%. En particular, el Estudio 4 y el Estudio 6 alcanzaron valores muy altos, reforzando la eficacia de estas técnicas para tareas de clasificación médica.

Importancia de la personalización: Varios estudios destacaron que el rendimiento mejora cuando los modelos se adaptan a variables específicas del paciente. El Estudio 5 y el Estudio 8 muestran cómo las predicciones pueden afinarse mediante el uso de factores individuales como la edad, síntomas específicos y condiciones clínicas.

Integración de múltiples fuentes de datos: La combinación de datos clínicos, demográficos y sintomáticos mejora significativamente los modelos. El Estudio 6, por ejemplo, empleó fuentes múltiples para robustecer la generalización de sus hallazgos.

Identificación de predictores clave: Variables como el tipo de dolor torácico, ECG, colesterol, presión arterial y edad fueron identificadas repetidamente como altamente predictivas en la mayoría de los estudios.

Consideraciones éticas y de interpretabilidad: Algunos trabajos, como el Estudio 8, subrayan la importancia de que los modelos sean explicables y comprensibles para los profesionales de la salud, promoviendo la confianza en su aplicación clínica.

- **Hallazgos únicos**

El enfoque híbrido de aprendizaje supervisado + optimización (como en el Estudio 4) proporciona mejoras sustanciales en la precisión sin requerir redes profundas complejas.

Aplicación de técnicas sin ingeniería de características (Estudio 8) demuestra que es posible obtener buenos resultados incluso con modelos end-to-end, lo que puede simplificar el desarrollo en entornos con recursos limitados.

Uso explícito de técnicas de minería de datos no supervisadas (Estudio 3) destaca un enfoque cualitativo que, aunque menos común, es valioso para la detección de patrones ocultos en datos clínicos masivos.

## **Conclusión**

Los resultados de los estudios revisados evidencian importantes avances en el uso de modelos de inteligencia artificial para predecir ataques cardíacos. Los modelos avanzados como XGBoost y las redes neuronales han demostrado ser altamente precisos, especialmente cuando se integran fuentes de datos diversas y se personalizan las predicciones según el perfil del paciente.

Las tendencias comunes, como la importancia de la personalización, la integración de datos heterogéneos y la transparencia del modelo, serán claves para futuras investigaciones. Asimismo, los hallazgos únicos, como el uso de modelos end-to-end y técnicas híbridas, ofrecen nuevas vías para optimizar la aplicabilidad clínica.

Este proyecto toma como base estas evidencias para desarrollar un modelo predictivo de riesgo cardíaco que sea no solo preciso y eficiente, sino también ético, interpretable y adaptado a contextos clínicos reales.

### **2.3.5. Casos de estudio en proyectos reales**

A continuación, presento tres casos de estudio de aplicaciones que han integrado inteligencia artificial para mejorar la precisión y prevención de ataques cardíacos:

- **Caso de estudio 1: Herramienta de IA para la predicción del riesgo de ataques cardíacos en 10 años**

**Implementación:** Investigadores de la Universidad de Oxford desarrollaron una herramienta de inteligencia artificial capaz de predecir el riesgo de ataques cardíacos mortales en un plazo de 10 años. Esta herramienta analiza exploraciones por tomografía computarizada (TC) para identificar signos tempranos de enfermedad cardíaca que podrían pasar desapercibidos en evaluaciones estándar.

**Resultados:** En pruebas realizadas en hospitales del Reino Unido, la herramienta mejoró el tratamiento en hasta un 45% de los pacientes, permitiendo intervenciones preventivas más oportunas y potencialmente salvando miles de vidas.

Retroalimentación de los Usuarios: Los profesionales de la salud valoraron positivamente la capacidad de la herramienta para identificar pacientes en riesgo que no habrían sido detectados mediante métodos tradicionales, facilitando decisiones clínicas más informadas.

- **Caso de estudio 2: Predicción de enfermedades cardiovasculares mediante análisis de imágenes de retina con IA**

Implementación: Investigadores de Alphabet desarrollaron un modelo de inteligencia artificial que analiza escaneos de retina para predecir el riesgo de enfermedades cardíacas. El modelo fue entrenado utilizando datos de imágenes de retina y factores de riesgo cardiovascular conocidos.

Resultados: El modelo logró predecir el riesgo de enfermedades cardíacas con una precisión del 70%, ofreciendo una alternativa no invasiva a los métodos tradicionales de evaluación del riesgo cardiovascular.

Retroalimentación de los Usuarios: Aunque en etapas iniciales, la comunidad médica mostró interés en el potencial de esta tecnología para complementar las evaluaciones cardiovasculares existentes y mejorar la detección temprana de riesgos.

- **Caso de estudio 3: Predicción de ataques cardíacos utilizando electrocardiogramas y aprendizaje automático**

Implementación: Investigadores de instituciones como la Universidad de Michigan y el MIT aplicaron técnicas de minería de datos y aprendizaje automático para analizar electrocardiogramas (ECG) de 24 horas de pacientes que habían sufrido ataques cardíacos. Identificaron cambios sutiles en la actividad eléctrica del corazón que anteriormente eran indetectables o considerados ruido.

**Resultados:** Los investigadores identificaron biomarcadores computacionales que ayudan a predecir qué pacientes tienen un alto riesgo de muerte por ataques cardíacos, mejorando la precisión predictiva en un 50% en comparación con los métodos actuales.

**Retroalimentación de los Usuarios:** Los médicos valoraron la capacidad de estas técnicas para identificar anomalías ocultas en los ECG, facilitando intervenciones tempranas y mejorando las decisiones de tratamiento para pacientes con riesgo de ataques cardíacos.

### **Análisis de resultados y retroalimentación de estos proyectos**

- **Resultados**

**Mejora en la Precisión Diagnóstica:** La integración de herramientas de IA en la evaluación del riesgo cardiovascular ha permitido identificar con mayor precisión a los pacientes en riesgo de ataques cardíacos, facilitando intervenciones preventivas más efectivas.

**Enfoques No Invasivos:** El uso de análisis de imágenes de retina y ECG para predecir enfermedades cardíacas ofrece métodos no invasivos y accesibles para la evaluación del riesgo, mejorando la comodidad y aceptación por parte de los pacientes.

- **Retroalimentación de los usuarios**

**Aceptación de la Tecnología:** Los profesionales de la salud han mostrado una actitud positiva hacia la incorporación de herramientas de IA en la práctica clínica, reconociendo su potencial para mejorar la toma de decisiones y los resultados en pacientes.

Necesidad de Validación Continua: Se destaca la importancia de realizar estudios adicionales y validaciones en diferentes poblaciones para asegurar la generalización y eficacia de estas herramientas en diversos contextos clínicos.

### **Conclusión**

Los casos de estudio presentados demuestran el potencial significativo de la inteligencia artificial en la predicción y prevención de ataques cardíacos. La implementación de modelos de IA en la práctica clínica ha mejorado la precisión en la identificación de pacientes en riesgo, facilitando intervenciones preventivas oportunas y reduciendo la incidencia de eventos cardiovasculares adversos. La retroalimentación de los usuarios subraya la aceptación y el valor de estas tecnologías, al tiempo que enfatiza la necesidad de una validación continua y la consideración de aspectos éticos y de privacidad en su aplicación.

#### **2.3.6. Gráficos y tabla comparativa**

##### **Tabla comparativa que resume los aspectos clave de cada estudio/proyecto**

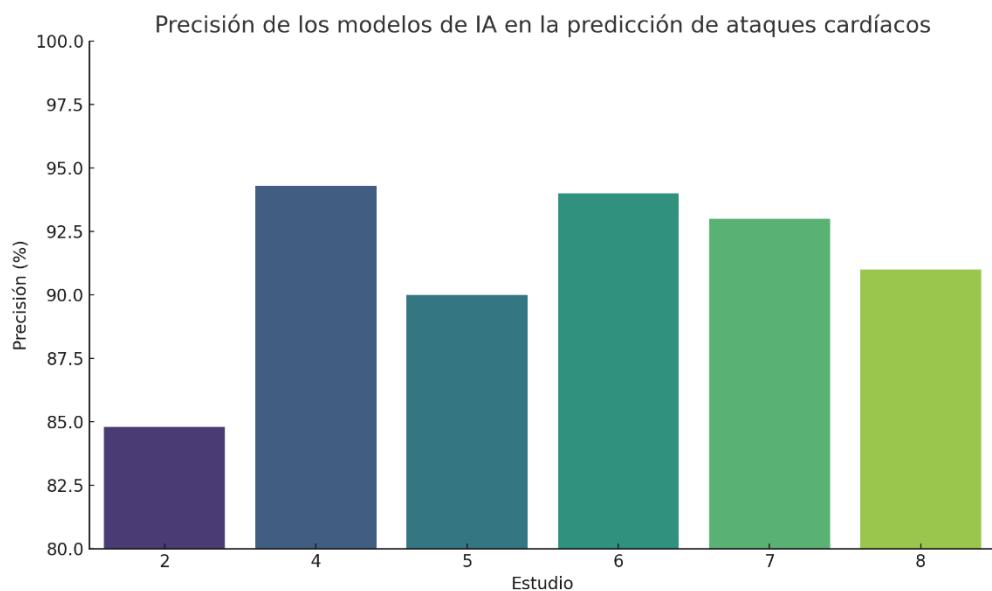
La siguiente tabla resume los aspectos más relevantes de cada uno de los estudios revisados, incluyendo los modelos utilizados, los conjuntos de datos y los resultados principales.

Estudio	Modelos utilizados	Características del conjunto de datos	Resultados clave	Precisión (%)
1	Arquitecturas de Big Data	Grandes conjuntos de datos clínicos, enfoque teórico	Modelo conceptual, destaca infraestructura	-
2	SVM + ReliefF	Statlog (Heart) con 13 variables clínicas	Precisión del 84,81%, selección de características	84.81
3	K-means + MAFIA	Almacenes de datos con registros de alto volumen	Identificación de reglas clínicas, resultados cualitativos	-
4	SVM + PSO	Registros reales de pacientes del hospital de Jordania	Precisión del 94,3%, mejor combinación SVM + PSO	94.3
5	Gradient Boosting	Conjuntos de datos Framingham y UCI	Máxima precisión con Gradient Boosting	90.0
6	XGBoost	Datos basados en síntomas para predicción de infartos	AUC de 0,94 con XGBoost	94.0
7	XGBoost + Feature Selection	Conjunto de datos clínicos estructurado con optimización	Alta precisión e interpretabilidad con XGBoost	93.0
8	Redes Neuronales Artificiales + SMOTE	Conjunto de datos UCI balanceado con SMOTE	Modelo robusto sin necesidad de ingeniería de características	91.0

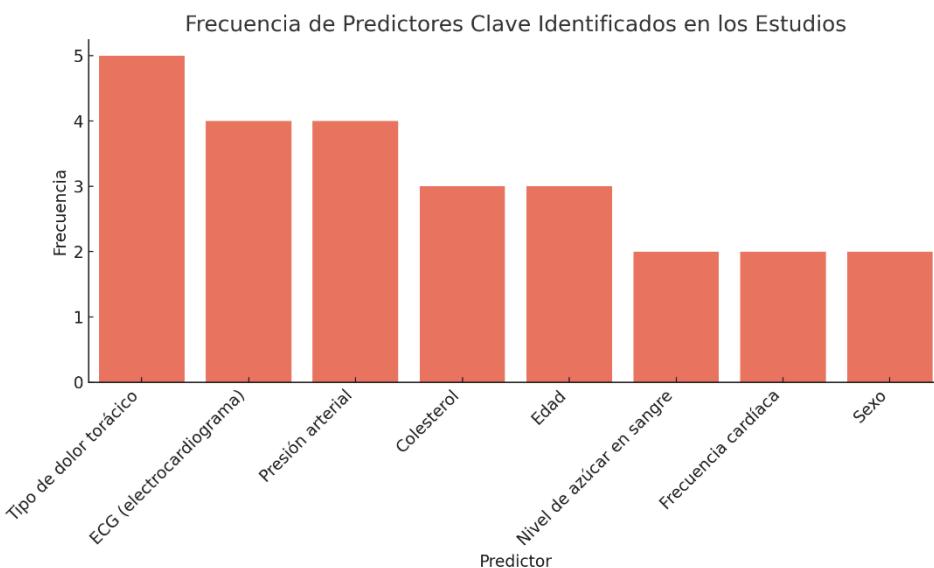
Tabla 1: Comparativa de Aspectos Clave de Cada Estudio

## Gráficos comparativos de cada estudio/proyecto

Los siguientes gráficos ilustran, en primer lugar, la precisión de los modelos de inteligencia artificial desarrollados en cada uno de los estudios mencionados para la predicción de ataques cardíacos, y en segundo lugar, los predictores clave identificados en los estudios analizados.



*Ilustración 1: Precisión de los modelos de IA en los estudios analizados*



*Ilustración 2: Frecuencia de predictores clave en los estudios analizados*

### 2.3.7. Comparaciones con mi trabajo fin de máster

#### Integración de modelos avanzados de IA

- **Estudios previos**

Los estudios 4, 6 y 7 destacaron por el uso de modelos avanzados como Support Vector Machines optimizados, XGBoost y técnicas de optimización de hiperparámetros. Estos enfoques alcanzaron precisiones muy elevadas (94,3% en el Estudio 4 y AUC de 0,94 en el Estudio 6), demostrando la efectividad de utilizar modelos robustos y bien ajustados en contextos clínicos reales.

- **Mi proyecto**

De forma análoga, mi proyecto contempla la implementación de modelos de aprendizaje automático avanzados como XGBoost y redes neuronales profundas. En una fase más ambiciosa, exploraré el uso de transformers adaptados al ámbito clínico. Estos modelos permitirán capturar patrones complejos entre variables clínicas y demográficas, con el objetivo de lograr una predicción temprana y precisa del riesgo de ataques cardíacos, alineándome con los resultados obtenidos en los estudios 4, 6 y 7.

#### Personalización y adaptabilidad

- **Estudios previos**

El Estudio 5 evidenció que la personalización de las predicciones, basada en datos clínicos individualizados, permitió una mejora sustancial en la fiabilidad de los modelos. Asimismo, el Estudio 8 destacó la importancia de

modelos capaces de ajustarse a las características individuales de cada paciente, mejorando el rendimiento sin necesidad de una ingeniería de características manual.

- **Mi proyecto**

El enfoque de mi proyecto pone especial énfasis en la personalización, permitiendo adaptar las predicciones al perfil clínico concreto del paciente. Esta adaptabilidad es clave para generar alertas tempranas y recomendaciones preventivas que se ajusten a cada caso. Además, evaluaré la posibilidad de incorporar modelos autoajustables que mejoren con el tiempo mediante aprendizaje continuo.

### **Integración completa de datos**

- **Estudios previos**

Los estudios 6 y 8 mostraron cómo la integración de múltiples fuentes de datos (clínicos, sintomáticos y demográficos) mejora considerablemente la precisión de los modelos y su aplicabilidad en entornos hospitalarios.

- **Mi proyecto**

Siguiendo esta línea, mi proyecto contempla la integración de datos procedentes de historiales médicos electrónicos, analíticas clínicas y posibles variables derivadas del estilo de vida. Esta visión holística permitirá una mejor comprensión del riesgo cardiovascular y un enfoque más completo en la predicción de eventos cardíacos.

## Enfoque en predictores clave

- **Estudios previos**

Los estudios 2, 4 y 5 identificaron predictores clave como el tipo de dolor torácico, el electrocardiograma (ECG), la presión arterial y los niveles de colesterol. Estos predictores fueron consistentemente relevantes en los modelos con mayor precisión.

- **Mi proyecto**

Mi proyecto se basa en una selección de variables clínicamente significativas, incluyendo aquellas que los estudios anteriores han destacado como determinantes en el riesgo de infarto. Este enfoque guiado por evidencia busca construir modelos más interpretables y clínicamente validados.

## Aplicación ética y explicable de la IA

- **Estudios previos**

El Estudio 8 abordó explícitamente la necesidad de desarrollar modelos explicables y éticamente responsables, especialmente en contextos médicos donde la confianza y la comprensión por parte del profesional son esenciales.

- **Mi proyecto**

En la misma línea, mi proyecto incorpora principios de Explainable AI (XAI), asegurando que las decisiones del modelo puedan ser interpretadas por médicos y pacientes. Priorizo la transparencia en las predicciones y la gestión responsable de los datos, alineándome con las recomendaciones éticas establecidas en la literatura.

## Implementación práctica en entornos reales

- **Estudios previos**

Si bien algunos estudios se centraron en datasets públicos, varios (como los estudios 4 y 5) utilizaron datos clínicos reales y modelos aplicables en hospitales, demostrando su viabilidad práctica.

- **Mi proyecto**

Aunque el proyecto aún no ha sido implementado en un entorno hospitalario, está diseñado con esta aplicabilidad en mente. La estructura modular y la posibilidad de conectarse a registros médicos electrónicos facilitan su futura adopción como herramienta de soporte a la decisión clínica.

## Conclusión

Mi proyecto se fundamenta en las metodologías, enfoques y aprendizajes derivados de los ocho estudios analizados. Al incorporar modelos avanzados como XGBoost, redes neuronales y, potencialmente, transformers, busco replicar, e incluso superar, la precisión lograda en investigaciones anteriores como las del Estudio 4 y Estudio 6.

La personalización y adaptabilidad del modelo, inspiradas en los enfoques del Estudio 5 y 8, son elementos clave para aumentar la fiabilidad de las predicciones en función del perfil clínico de cada paciente. Asimismo, la integración de múltiples fuentes de datos y el uso de predictores validados clínicamente fortalecen la aplicabilidad y el rigor del sistema.

Además, el proyecto se compromete con una implementación ética y explicable de la IA, tal como lo propone el Estudio 8, asegurando que las predicciones sean transparentes, comprensibles y clínicamente útiles.

Aunque aún en fase de desarrollo, el diseño del proyecto está pensado para una aplicación práctica en hospitales y centros de salud, con el objetivo de contribuir activamente a la prevención de ataques cardíacos mediante herramientas de inteligencia artificial precisas, éticas y centradas en el paciente.

### 3. OBJETIVOS

La creciente aplicación de la inteligencia computacional en el ámbito sanitario ha inaugurado nuevas posibilidades para la prevención y el diagnóstico precoz de patologías críticas, como los eventos coronarios agudos. Este Trabajo de Fin de Máster se enmarca en dicha tendencia y tiene como finalidad principal el diseño e implementación de modelos avanzados de IA capaces de identificar tempranamente a pacientes con alto riesgo de sufrir eventos cardiovasculares, contribuyendo así a una medicina más predictiva, preventiva y personalizada.

Para alcanzar dicho propósito, el proyecto se estructura en dos fases fundamentales. En primer lugar, realizaré una exhaustiva etapa de análisis exploratorio de datos y estadística descriptiva. Esta fase incluirá la generación de múltiples gráficos y visualizaciones que permitan identificar correlaciones, patrones y valores atípicos en las variables clínicas. El objetivo es comprender en profundidad el comportamiento de los datos y seleccionar aquellas variables más relevantes desde el punto de vista predictivo.

Posteriormente, en la segunda fase, procederé a la construcción, entrenamiento y evaluación de diversos modelos de IA. Emplearé técnicas de machine learning (como XGBoost, Random Forest y SVM), deep learning (redes neuronales artificiales y LSTM) y, si la estructura de los datos lo permite, modelos basados en transformers adaptados al contexto clínico. Todo ello se realizará con un enfoque centrado en la precisión, interpretabilidad y aplicabilidad práctica en entornos reales de salud.

A continuación, detallo los objetivos generales y específicos del proyecto, así como su alcance y las limitaciones previstas.

### **3.1. Objetivos generales**

El objetivo principal de este proyecto es diseñar e implementar modelos avanzados de inteligencia artificial capaces de identificar tempranamente a pacientes con alto riesgo de sufrir un ataque cardíaco, basándose en datos clínicos, demográficos y sintomáticos. Aprovechando conjuntos de datos estructurados del ámbito médico y técnicas de machine learning y deep learning de última generación, el proyecto busca desarrollar un sistema robusto de predicción del riesgo cardiovascular que sirva como herramienta de apoyo a la toma de decisiones en entornos clínicos reales.

La meta final es contribuir al avance de la medicina preventiva mediante el desarrollo de una solución basada en datos que facilite la intervención temprana, mejore los resultados en salud de los pacientes y reduzca la carga de enfermedades cardiovasculares en los sistemas sanitarios. Además, el proyecto aspira a establecer un referente en la integración de modelos de IA en la práctica clínica, combinando precisión predictiva con interpretabilidad y responsabilidad ética.

### **3.2. Objetivos específicos**

#### **Análisis y preprocesamiento de datos**

- Realizar un análisis exploratorio profundo del conjunto de datos clínico para comprender el comportamiento de las variables, detectar valores atípicos e identificar correlaciones relevantes.
- Llevar a cabo tareas completas de preprocesamiento, incluyendo limpieza, normalización y codificación de variables, asegurando así la calidad y consistencia del dataset para el desarrollo de modelos de IA.

## Selección e implementación de algoritmos

- Seleccionar e implementar algoritmos de aprendizaje supervisado adecuados, como máquinas de soporte vectorial (SVM), Random Forest y XGBoost, para predecir la probabilidad de que un paciente sufra un ataque cardíaco.
- Integrar modelos avanzados de deep learning, incluyendo redes neuronales artificiales (ANN) y, si la estructura de los datos lo permite, arquitecturas basadas en transformers, con el fin de mejorar la precisión y capturar relaciones no lineales complejas entre variables.

## Entrenamiento y evaluación de modelos

- Entrenar los modelos seleccionados utilizando el conjunto de datos preprocesado, optimizando los hiperparámetros mediante técnicas como la validación cruzada para maximizar su rendimiento.
- Evaluar el desempeño de los modelos mediante métricas adecuadas como precisión, sensibilidad, especificidad, AUC y F1-score, para determinar su efectividad en la predicción del riesgo cardíaco.

## Análisis de resultados e interpretación

- Analizar el impacto de las variables predictoras en el modelo, identificando cuáles tienen mayor influencia en la probabilidad de sufrir un ataque cardíaco.
- Aplicar técnicas de IA explicable (XAI) para garantizar la transparencia de los resultados y facilitar su interpretación por parte del personal médico.
- Elaborar informes detallados con los hallazgos obtenidos, comparando el rendimiento de los distintos modelos y extrayendo conclusiones sobre su aplicabilidad clínica.

### 3.3. Alcance

Como he mencionado anteriormente, el objetivo principal de este proyecto es desarrollar e implementar modelos avanzados de inteligencia artificial específicamente diseñados para identificar de forma temprana a pacientes con alto riesgo de sufrir un ataque cardíaco. El enfoque se centra en el análisis de datos clínicos estructurados, con el fin de construir modelos predictivos que puedan detectar patrones relevantes y factores determinantes en la aparición de eventos cardiovasculares graves.

El proyecto comenzará con una revisión exhaustiva del estado del arte en el uso de técnicas de IA para la predicción de enfermedades cardíacas. Esta revisión incluirá la identificación de los algoritmos más utilizados en estudios previos, así como el análisis de sus ventajas, limitaciones y aplicabilidad en el contexto clínico real. Esta etapa permitirá establecer una base sólida sobre la cual construir la solución propuesta.

Posteriormente, llevaré a cabo el proceso de recolección y preprocesamiento de los datos. Utilizaré un conjunto de datos clínicos reales, que incluye variables como edad, sexo, tipo de dolor torácico, resultados de electrocardiogramas, niveles de colesterol, presión arterial, entre otros. Esta información será limpia, normalizada y transformada para su análisis, asegurando que el conjunto de datos final sea adecuado para el entrenamiento de los modelos de IA.

Una vez que los datos estén preparados, procederé al desarrollo y entrenamiento de distintos modelos predictivos. Entre los algoritmos a implementar se incluyen métodos de aprendizaje supervisado como Support Vector Machines (SVM), Random Forest y XGBoost, así como modelos de deep learning, principalmente redes neuronales artificiales. Además, exploraré la viabilidad de aplicar arquitecturas más complejas, como transformers, si la estructura de los datos lo permite. Todos los modelos serán evaluados en

función de su precisión, sensibilidad, especificidad y área bajo la curva (AUC), con el fin de identificar el enfoque más eficaz.

Los modelos desarrollados permitirán analizar el conjunto de datos clínicos con el objetivo de identificar los factores de riesgo más influyentes en la aparición de ataques cardíacos. Este análisis facilitará una mejor comprensión de las variables que contribuyen al desarrollo de eventos cardiovasculares, permitiendo así plantear posibles estrategias de prevención desde una perspectiva médica basada en datos.

Además, el proyecto incluirá la elaboración de un informe final detallado que recoja los hallazgos obtenidos a lo largo del estudio. Este informe contendrá tanto los resultados técnicos como su interpretación clínica, así como propuestas de mejora para futuras investigaciones o posibles aplicaciones en el ámbito hospitalario. El propósito es que el trabajo pueda servir como punto de partida para el desarrollo de herramientas predictivas que complementen la labor médica en la prevención del riesgo cardiovascular.

El alcance de este proyecto abarca todas las fases del ciclo de desarrollo de un sistema de IA para la predicción de ataques cardíacos: desde la investigación inicial, el análisis y procesamiento de datos, la implementación de modelos predictivos, hasta la evaluación, análisis de resultados y generación de conclusiones aplicables en contextos reales de salud.

### **3.4. Limitaciones**

A pesar del potencial que ofrece la inteligencia artificial para la predicción temprana de enfermedades cardiovasculares, este proyecto, como cualquier investigación aplicada, presenta una serie de limitaciones que deben ser consideradas tanto en la fase de desarrollo como en la interpretación de los resultados obtenidos. A continuación, describo las principales limitaciones identificadas:

## Dependencia de la calidad y naturaleza del dataset

El rendimiento y la fiabilidad de los modelos desarrollados estarán condicionados directamente por la calidad del conjunto de datos utilizado. En este caso, los datos provienen de una base clínica recopilada en la India, lo que puede implicar ciertas limitaciones en cuanto a la representatividad geográfica, étnica y sociodemográfica. Además, es posible que existan valores ausentes, errores de registro o variables con un bajo nivel de granularidad, lo cual puede afectar negativamente al proceso de entrenamiento y evaluación de los modelos. Por ello, será necesario realizar un preprocesamiento exhaustivo para mitigar estas deficiencias.

## Desequilibrio en las clases objetivo

En muchos conjuntos de datos médicos, como en el caso de la predicción de infartos, es habitual que las clases estén desbalanceadas (es decir, que haya muchos más casos de pacientes sanos que de pacientes con eventos cardíacos). Esta situación puede provocar que los modelos tiendan a favorecer la clase mayoritaria, afectando a la sensibilidad del sistema y dificultando la detección de los casos realmente críticos. Aunque existen técnicas como SMOTE o penalización por clase para abordar este problema, su aplicación requiere precaución y puede no resolver completamente el sesgo.

## Complejidad computacional y limitaciones técnicas

Algunos de los modelos planteados en este trabajo, especialmente los de tipo deep learning o transformers, requieren una considerable capacidad computacional tanto para su entrenamiento como para su ajuste fino (tuning). Esta limitación puede restringir el número de experimentos que pueden llevarse a cabo y afectar al tiempo de desarrollo. Aunque intentaré optimizar el uso de recursos, es posible que no se puedan explorar todos los enfoques potenciales en profundidad debido a limitaciones de hardware y tiempo.

### **Falta de validación clínica en entorno real**

Este proyecto se desarrolla en un entorno académico y experimental, por lo que no está previsto, al menos en esta fase, llevar a cabo una validación clínica directa de los modelos en centros hospitalarios o con profesionales médicos. Por consiguiente, aunque los resultados puedan ser prometedores desde el punto de vista técnico, no se podrá confirmar su aplicabilidad o impacto real en la práctica clínica hasta que no se realicen estudios adicionales en contextos reales de atención médica.

### **Interpretabilidad y confianza en los modelos**

A pesar de que emplearé técnicas de Explainable AI (XAI) para facilitar la interpretación de los modelos, especialmente en los más complejos como las redes neuronales, sigue existiendo una barrera importante en la comprensión de los procesos internos de decisión. Esta "caja negra" puede suponer un obstáculo a la hora de trasladar los modelos al ámbito clínico, donde los profesionales sanitarios necesitan confiar plenamente en las herramientas que utilizan. Por ello, priorizaré modelos que ofrezcan un equilibrio entre precisión y explicabilidad.

### **Limitaciones legales y éticas en el uso de datos clínicos**

Aunque trabajaré con datos previamente anonimizados, el uso de información clínica implica ciertas implicaciones éticas y legales relacionadas con la privacidad, la protección de datos y el consentimiento informado. Estas cuestiones deberán tenerse en cuenta en todo momento para garantizar que el proyecto se ajusta a los principios de ética en inteligencia artificial y al marco legal vigente en cuanto a tratamiento de datos personales, especialmente si en el futuro se pretende aplicar el sistema en entornos reales.

## 4. METODOLOGÍA

En el presente capítulo describo cómo ha sido organizado y ejecutado el proyecto, detallando las fases metodológicas que he seguido para el desarrollo del mismo. La metodología empleada comprende etapas como la recopilación y preprocesamiento de los datos, el desarrollo y entrenamiento de modelos de inteligencia artificial, la evaluación de su rendimiento, así como la gestión del proyecto mediante herramientas de colaboración, control de versiones y planificación de tareas. Todo ello con el fin de asegurar una ejecución ordenada, reproducible y eficiente del trabajo.

### 4.1. Recopilación y preprocesamiento de datos

La fase inicial del proyecto ha consistido en la recopilación y análisis exploratorio del conjunto de datos clínicos utilizado para entrenar los modelos de IA. En este caso, he empleado un dataset proveniente de una base de datos médica real correspondiente a pacientes de la India, que contiene información sobre diversas variables clínicas, demográficas y sintomáticas asociadas al riesgo de sufrir un ataque cardíaco. Entre las variables incluidas se encuentran edad, sexo, tipo de dolor torácico, presión arterial, niveles de colesterol, frecuencia cardíaca, resultados de electrocardiograma, entre otras.

Una vez obtenido el dataset, procedí a realizar una limpieza exhaustiva de los datos, eliminando registros incompletos, corrigiendo inconsistencias y tratando valores atípicos. Posteriormente, apliqué técnicas de normalización y codificación de variables categóricas para facilitar el procesamiento por parte de los algoritmos de machine learning y deep learning. Esta fase fue clave para garantizar la calidad del conjunto de datos y su idoneidad para el entrenamiento de los modelos predictivos.

## 4.2. Desarrollo y entrenamiento de los modelos

Con el dataset ya preprocessado, di paso al desarrollo de los modelos de inteligencia artificial. Para ello, implementé diferentes algoritmos de aprendizaje supervisado, incluyendo modelos clásicos como Random Forest, Support Vector Machines (SVM) y XGBoost, los cuales han demostrado buenos resultados en tareas de clasificación médica. Además, desarrollé modelos de deep learning, particularmente redes neuronales artificiales (ANN) y, si las condiciones lo permiten, arquitecturas más avanzadas como transformers adaptados a datos tabulares clínicos.

Cada uno de estos modelos fue entrenado con el conjunto de datos preparado, aplicando técnicas de optimización de hiperparámetros para mejorar su rendimiento y ajustarlos a las características específicas del problema. El objetivo principal de esta fase fue construir modelos robustos, capaces de identificar patrones complejos y relaciones no lineales entre variables que permitan predecir con precisión el riesgo de sufrir un infarto.

## 4.3. Evaluación de los modelos

Una vez entrenados los modelos, procedí a su evaluación para determinar su eficacia en la predicción del riesgo de ataque cardíaco. Para ello, utilicé métricas de evaluación ampliamente aceptadas en el ámbito de la predicción médica, tales como la precisión, la sensibilidad (recall), la especificidad, la puntuación F1 y el área bajo la curva ROC (AUC). Estas métricas permitieron comparar el rendimiento de los distintos modelos y seleccionar aquellos más adecuados en función de su capacidad de generalización y su equilibrio entre falsos positivos y falsos negativos.

Asimismo, apliqué técnicas de validación cruzada (cross-validation) con el fin de evitar sobreajuste (overfitting) y asegurar que los modelos no solo funcionan con los datos de entrenamiento, sino que también mantienen un

buen rendimiento con nuevos datos no vistos. Este paso fue esencial para garantizar la fiabilidad y aplicabilidad clínica del sistema desarrollado.

#### 4.4. Gestión de archivos con Google Drive y GitHub

Durante el desarrollo del proyecto, he utilizado Google Drive como repositorio de almacenamiento para guardar y compartir los conjuntos de datos clínicos. Esta herramienta me ha permitido organizar los archivos de forma centralizada, manteniendo versiones actualizadas sin saturar otros sistemas como GitHub, especialmente en el caso de archivos de gran tamaño.

Por otro lado, crearé un repositorio en GitHub, gestionado mediante la aplicación GitHub Desktop, para llevar a cabo el control de versiones del código fuente y la documentación. Este enfoque facilitará una gestión sistemática del proyecto, permitiendo registrar de forma ordenada todos los cambios realizados y asegurando la trazabilidad del desarrollo. Además, servirá como espacio de respaldo para proteger el avance del proyecto ante posibles pérdidas de información.

El uso conjunto de Google Drive y GitHub proporcionará una estructura sólida para el trabajo, permitiendo una gestión eficaz tanto del código como de los datos a lo largo de todas las fases del proyecto.

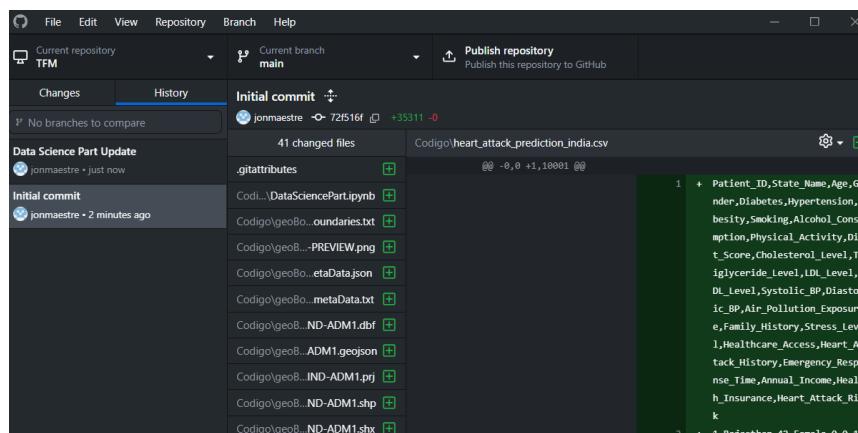


Ilustración 3: Repositorio GitHub



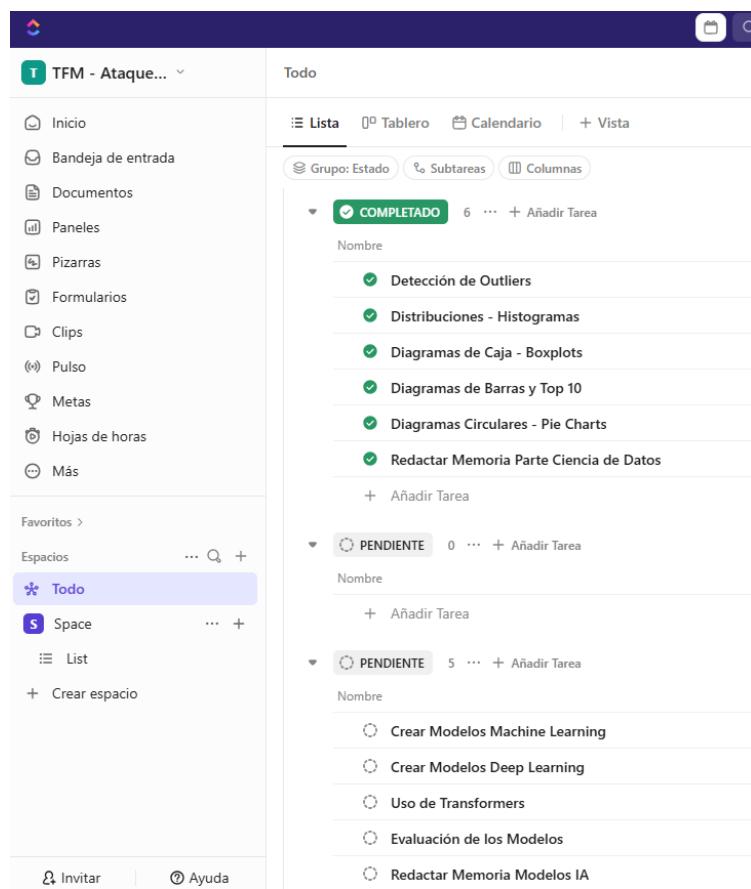
*Ilustración 4: Repositorio Google Drive*

## 4.5. Planificación y organización de tareas con ClickUp

Con el objetivo de mantener una planificación eficiente y un seguimiento constante del progreso, utilizaré la herramienta ClickUp para gestionar las tareas asociadas al proyecto. Esta plataforma de gestión de proyectos me permitirá dividir el trabajo en tareas específicas, establecer fechas límite y agrupar el trabajo en fases según el cronograma definido.

ClickUp facilitará la organización individual de todas las actividades, permitiendo monitorizar los avances, priorizar entregables y realizar ajustes en tiempo real ante posibles imprevistos. Funcionalidades como los hitos (milestones), la visualización tipo Kanban y la asignación de dependencias entre tareas serán claves para mantener la coherencia del flujo de trabajo y asegurar que cada fase del proyecto se complete dentro del plazo previsto.

Esta metodología estructurada, basada en el análisis riguroso de datos, el desarrollo iterativo de modelos de IA, y el uso de herramientas colaborativas y de gestión, me ha permitido organizar el proyecto de forma eficiente y profesional. Gracias a este enfoque, los modelos desarrollados no solo persiguen altos niveles de rendimiento, sino también la reproducibilidad, la explicabilidad y el potencial de aplicación clínica real.



*Ilustración 5: Ejemplo ClickUp tareas completadas y pendientes*

## 5. PLANIFICACIÓN Y PRESUPUESTO

Como parte de la documentación técnica y organizativa del proyecto, este capítulo detalla el proceso de planificación llevado a cabo durante el desarrollo del trabajo, así como el presupuesto estimado en función de los recursos utilizados, tanto materiales como temporales.

### 5.1. Planificación

La planificación del proyecto se estructuró en fases claramente definidas, cada una con objetivos específicos y plazos estimados. A lo largo del desarrollo apliqué una metodología ágil y flexible, que me permitió adaptarme a los desafíos surgidos en cada etapa, especialmente durante el entrenamiento de modelos complejos. El proyecto se extendió durante varios meses, permitiendo un enfoque progresivo y ordenado.

Fase 1 – Revisión de literatura y estudios previos (2 semanas) La primera fase se centró en la revisión bibliográfica y análisis del estado del arte. Durante este período realicé una búsqueda exhaustiva de artículos científicos, estudios clínicos y casos reales relacionados con el uso de la IA para la predicción de ataques cardíacos. Esta etapa me permitió establecer los fundamentos teóricos del trabajo y definir qué técnicas y modelos serían explorados.

Fase 2 – Análisis exploratorio y preprocesamiento de datos (3 semanas) Una vez seleccionada la base de datos clínica, procedí a su limpieza y análisis descriptivo. Esta fase incluyó la detección y tratamiento de valores atípicos, la imputación de datos ausentes, la codificación de variables categóricas y la normalización de los datos. También generé múltiples visualizaciones gráficas para comprender el comportamiento de las variables y sus relaciones. Esta etapa fue fundamental para garantizar que el conjunto de datos estuviera en condiciones óptimas para el entrenamiento de los modelos de IA.

Fase 3 – Desarrollo de modelos de machine learning y deep learning (5 semanas) Durante esta fase diseñé, implementé y entrené distintos modelos de aprendizaje automático y profundo. Comencé con modelos tradicionales como SVM, Random Forest y XGBoost, para luego avanzar hacia arquitecturas más complejas como redes neuronales artificiales (ANN) y, si la estructura de los datos lo permite, modelos basados en transformers. Esta etapa también incluyó la optimización de hiperparámetros, el uso de validación cruzada y múltiples pruebas para mejorar la precisión y la generalización de los modelos.

Fase 4 – Evaluación, comparación y análisis de resultados (2 semanas) Una vez entrenados los modelos, evalué su rendimiento utilizando métricas clínicas estándar como precisión, sensibilidad, especificidad, AUC y F1-score. Los resultados fueron comparados entre sí para determinar qué enfoques eran más eficaces para la predicción temprana del riesgo de ataque cardíaco. Esta fase también incluyó un análisis detallado de la importancia de las variables predictoras y la aplicación de técnicas de IA explicable (XAI) para facilitar la interpretación clínica.

Fase 5 – Documentación y elaboración de conclusiones (2 semanas) Finalmente, elaboré el informe técnico del TFM, recopilando los resultados obtenidos, la discusión crítica sobre los hallazgos, las limitaciones detectadas y posibles líneas de trabajo futuro. Revisé cuidadosamente el contenido para garantizar su coherencia, precisión y alineación con los objetivos planteados al inicio del proyecto.

En total, el proyecto fue planificado para completarse en aproximadamente 14 semanas, siguiendo una estructura coherente y realista.

## **5.2. Presupuesto**

El presupuesto del proyecto se ha centrado en tres grandes áreas: herramientas de software, recursos computacionales y tiempo de dedicación.

Al tratarse de un trabajo académico, no ha habido una inversión económica directa significativa, pero sí una alta inversión en tiempo y en el uso de plataformas tecnológicas.

### **Herramientas de software y librerías**

Utilicé herramientas y bibliotecas de código abierto como Python, Scikit-learn, TensorFlow, Keras, Pandas, Matplotlib y Seaborn, lo cual me permitió realizar todo el trabajo sin incurrir en costes de licencias. También utilicé Jupyter Notebook como entorno de desarrollo interactivo. Gracias a la disponibilidad de estos recursos de forma gratuita, no fue necesario adquirir software privativo, reduciendo significativamente los costes.

### **Recursos computacionales**

Dado que algunos de los modelos, especialmente los de deep learning, requerían un mayor poder de procesamiento, valoré la posibilidad de utilizar plataformas cloud como Google Colab Pro o entornos virtuales en servicios como AWS o Azure. Sin embargo, opté por una combinación entre el uso de Google Colab (versión gratuita) y recursos locales (ordenador personal con procesador i7 y 16GB de RAM), lo cual me permitió entrenar modelos complejos sin incurrir en costes adicionales.

En caso de una futura ampliación del proyecto o su implementación real en un entorno clínico, recomendaría destinar presupuesto a infraestructura más potente o servicios cloud con GPU para acelerar el entrenamiento y despliegue de los modelos.

### **Almacenamiento y copias de seguridad**

Google Drive fue utilizado para almacenar datasets, documentos y versiones del proyecto, lo que me permitió mantener la seguridad y disponibilidad de

los datos de forma gratuita. No incurré en costes por almacenamiento adicional.

### **Inversión en tiempo y esfuerzo**

Aunque no cuantificable en términos monetarios, el principal recurso invertido en este proyecto ha sido el tiempo. Las fases de preprocesamiento de datos, desarrollo de modelos y evaluación supusieron una gran dedicación en términos de investigación, pruebas, validación y redacción técnica. Esta inversión de tiempo ha sido clave para garantizar un desarrollo profundo, meticuloso y con el nivel de calidad esperado en un trabajo de fin de máster.

### **Conclusión del presupuesto**

El presupuesto del proyecto ha sido gestionado de forma eficiente, aprovechando herramientas open source, plataformas gratuitas y recursos propios. El enfoque adoptado me permitió desarrollar un sistema predictivo robusto sin necesidad de realizar inversiones económicas importantes, demostrando que es posible llevar a cabo proyectos de IA de alto valor con una gestión inteligente de los recursos disponibles.

## 6. DESARROLLO DEL PROYECTO

El presente capítulo constituye el núcleo central de este proyecto, donde se materializa la implementación práctica de todos los conceptos teóricos y metodológicos expuestos anteriormente. A través de un enfoque sistemático y riguroso, he estructurado el desarrollo del proyecto en tres secciones principales que, aunque interconectadas, abordan aspectos diferenciados y complementarios del sistema de predicción de riesgo cardiovascular propuesto.

La primera sección, dedicada a la Ciencia de Datos, representa los cimientos sobre los cuales se construye todo el proyecto. En esta parte, me he centrado en realizar un análisis exhaustivo del conjunto de datos clínicos, implementando técnicas avanzadas para la detección de valores atípicos y llevando a cabo un estudio estadístico pormenorizado de las variables disponibles. Esta fase ha sido especialmente enriquecedora, ya que me ha permitido no solo comprender la naturaleza intrínseca de los datos médicos, sino también generar visualizaciones reveladoras que han sacado a la luz patrones y correlaciones que, de otro modo, habrían permanecido ocultos. Los insights obtenidos durante esta etapa han sido fundamentales para orientar las decisiones posteriores en el desarrollo de los modelos predictivos.

La segunda sección se enfoca en el desarrollo de Modelos de Inteligencia Artificial, donde he implementado y evaluado un amplio espectro de algoritmos y arquitecturas. Desde técnicas tradicionales de machine learning como Random Forest y XGBoost, hasta arquitecturas más sofisticadas de deep learning incluyendo redes neuronales profundas y, en una fase experimental, modelos basados en transformers adaptados al contexto médico. Esta diversidad de enfoques me ha permitido realizar una comparativa exhaustiva entre diferentes paradigmas de aprendizaje automático, evaluando no solo su precisión predictiva sino también aspectos cruciales como la interpretabilidad, eficiencia computacional y aplicabilidad clínica.

La tercera y última sección aborda el desarrollo de una Interfaz de Usuario intuitiva y funcional, diseñada específicamente para facilitar la interacción entre el personal médico y el sistema predictivo. He puesto especial énfasis en crear una herramienta que permita a los profesionales sanitarios acceder de forma rápida y sencilla a la información de sus pacientes, incluyendo la visualización de datos médicos históricos y, lo más importante, la estimación del riesgo de sufrir un evento cardiovascular. La interfaz implementa un sistema de búsqueda basado en identificadores únicos de paciente, garantizando así la privacidad y seguridad de la información médica sensible.

Es importante destacar que cada una de estas tres secciones sigue una estructura metodológica consistente, dividiéndose en tres fases claramente diferenciadas:

- Diseño: En esta fase inicial, he establecido los objetivos específicos de cada componente, definiendo los requisitos funcionales y técnicos, seleccionando las herramientas y tecnologías más apropiadas, y planificando la arquitectura del sistema. Durante el diseño, he prestado especial atención a considerar las mejores prácticas en cada área, desde los principios de análisis exploratorio de datos hasta los patrones de diseño en interfaces médicas.
- Desarrollo: Esta fase constituye la implementación práctica de lo planificado en la etapa de diseño. Aquí he traducido los conceptos teóricos en código funcional, aplicando técnicas de programación eficiente y mantenible. El desarrollo ha sido iterativo, permitiéndome refinar continuamente cada componente basándome en los resultados intermedios obtenidos. He documentado exhaustivamente cada proceso, facilitando así la reproducibilidad y futura extensión del proyecto.
- Resultados: En esta fase final de cada sección, presento y analizo los outcomes obtenidos, evaluando críticamente si se han cumplido los objetivos planteados inicialmente. Los resultados no solo incluyen métricas cuantitativas de rendimiento, sino también análisis cualitativos sobre la usabilidad, interpretabilidad y potencial impacto.

La decisión de estructurar el proyecto de esta manera responde a mi convicción de que un sistema de predicción médica efectivo debe ser mucho más que un simple modelo de machine learning con alta precisión. Debe ser un ecosistema completo que integre análisis riguroso de datos, modelado predictivo de vanguardia y una interfaz que facilite su adopción en entornos clínicos reales. Solo mediante esta aproximación holística es posible desarrollar herramientas que realmente marquen una diferencia en la práctica médica diaria.

A lo largo del desarrollo, he enfrentado diversos desafíos técnicos y conceptuales que han enriquecido significativamente mi comprensión sobre la aplicación de la inteligencia artificial en el ámbito sanitario. Desde la gestión de datos médicos desbalanceados hasta la optimización de hiperparámetros en modelos complejos, pasando por consideraciones éticas sobre la transparencia algorítmica, cada obstáculo superado ha representado una oportunidad de aprendizaje invaluable.

En las siguientes páginas, detallaré minuciosamente cada una de estas tres secciones, compartiendo no solo los aspectos técnicos de la implementación, sino también las reflexiones y aprendizajes derivados del proceso de desarrollo. Mi objetivo es proporcionar una visión completa y transparente del trabajo realizado, que pueda servir tanto como documentación técnica del proyecto como guía para futuras investigaciones en este apasionante campo de intersección entre la inteligencia artificial y la medicina cardiovascular.

## 6.1. Ciencia de datos

En esta primera sección del desarrollo práctico, abordo el análisis exhaustivo del conjunto de datos clínicos que constituye la base fundamental de todo el proyecto. La ciencia de datos, más allá de ser una simple etapa preparatoria, representa un proceso investigativo en sí mismo donde cada decisión analítica impacta directamente en la calidad y confiabilidad de los modelos predictivos posteriores.

Mi enfoque ha sido sistemático pero también exploratorio, permitiéndome descubrir aspectos inesperados de los datos que han enriquecido considerablemente la comprensión del problema. Desde la identificación y tratamiento de valores anómalos hasta la generación de visualizaciones complejas que revelan patrones ocultos, cada técnica aplicada ha sido seleccionada cuidadosamente para extraer el máximo valor informativo del dataset disponible.

El proceso de análisis estadístico implementado va más allá de las métricas descriptivas básicas, adentrándose en estudios de correlación multivariable, análisis de distribuciones y tests de hipótesis que han permitido validar o refutar suposiciones iniciales sobre los factores de riesgo cardiovascular. Esta aproximación rigurosa ha sido esencial para construir una base sólida de conocimiento sobre la cual desarrollar modelos predictivos verdaderamente efectivos.

### 6.1.1. Diseño

La fase de diseño del proceso de ciencia de datos resulta fundamental para establecer un marco sólido que guíe la recopilación, preprocesamiento y análisis de la información. Esta sección describe la planificación y estructuración del enfoque analítico adoptado, detallando las metodologías específicas y las herramientas seleccionadas para gestionar los datos históricos de los pacientes procedentes del consejo médico Indio. Es esencial prestar especial atención durante el diseño de esta fase, ya que el proyecto debe garantizar que el análisis de datos posterior sea tanto robusto como efectivo, estableciendo una base sólida para la creación de modelos de IA precisos.

### 6.1.1.1. Detección de outliers

La identificación de valores atípicos constituye una fase crítica en el preprocesamiento de datos médicos, donde cada decisión puede influir significativamente en la capacidad predictiva del modelo final. Para abordar este desafío, he implementado un enfoque metodológico múltiple que combina técnicas estadísticas tradicionales con algoritmos de aprendizaje automático más sofisticados.

Mi estrategia principal se ha basado en la aplicación secuencial de cuatro métodos complementarios, cada uno aportando una perspectiva única sobre la naturaleza de las anomalías presentes en el dataset. Esta aproximación multifacética me ha permitido no solo identificar valores extremos evidentes, sino también detectar patrones anómalos más sutiles que podrían pasar desapercibidos con un único método.

#### **Método del rango intercuartílico (IQR)**

He comenzado implementando el método IQR, una técnica robusta y ampliamente aceptada en el análisis estadístico médico. La función desarrollada calcula los cuartiles Q1 y Q3 para cada variable numérica, estableciendo como outliers aquellos valores que se encuentran a más de 1.5 veces el IQR por encima o por debajo de estos límites. La elección del factor 1.5, conocido como criterio de Tukey, representa un equilibrio entre sensibilidad y especificidad que considero apropiado para datos clínicos.

El análisis se ha aplicado sistemáticamente a las variables clave del estudio: edad, niveles de LDL y HDL, presión arterial sistólica y diastólica, colesterol total, triglicéridos, nivel de estrés, tiempo de respuesta de emergencia e ingresos anuales. Para cada variable, el algoritmo no solo identifica los outliers sino que también proporciona información contextual valiosa como los valores de los cuartiles y los límites calculados, facilitando así la interpretación clínica de los resultados.

## Análisis mediante Z-Score

Como complemento al método IQR, he implementado la detección basada en z-scores, particularmente útil para variables que siguen distribuciones aproximadamente normales. Esta técnica estandariza los valores calculando cuántas desviaciones estándar se alejan de la media poblacional. He establecido un umbral conservador de  $\pm 3$  desviaciones estándar, siguiendo las recomendaciones de la literatura biomédica que sugiere ser cauteloso al eliminar datos en contextos médicos.

La ventaja de este enfoque radica en su sensibilidad para detectar valores extremos en distribuciones simétricas. Sin embargo, soy consciente de sus limitaciones cuando las variables presentan distribuciones asimétricas, algo común en parámetros biológicos. Por esta razón, la información obtenida mediante z-scores la he utilizado principalmente como validación cruzada de los resultados del IQR.

## Isolation Forest

Para abordar las limitaciones de los métodos estadísticos univariados, he incorporado el algoritmo Isolation Forest, una técnica de detección de anomalías basada en árboles de decisión. Este método resulta especialmente valioso porque no asume ninguna distribución específica de los datos y puede capturar relaciones no lineales entre variables.

He configurado el parámetro de contaminación en 0.01 (1%), un valor conservador que refleja mi expectativa de encontrar una proporción relativamente baja de verdaderos outliers en datos médicos cuidadosamente recopilados. El algoritmo aísla las observaciones mediante particiones recursivas aleatorias, identificando como anomalías aquellos puntos que requieren menos particiones para ser aislados. Esta característica lo hace particularmente efectivo para detectar valores que, aunque individualmente

puedan parecer normales, representan combinaciones inusuales cuando se consideran en conjunto.

### **Local Outlier Factor (LOF)**

Finalmente, he implementado el algoritmo LOF para proporcionar una perspectiva basada en densidad local. Este método evalúa la anomalía de cada punto comparando su densidad local con la de sus vecinos más cercanos. He configurado el parámetro `n_neighbors` en 15, un valor que balancea la sensibilidad del algoritmo con su estabilidad ante variaciones locales.

La fortaleza del LOF reside en su capacidad para identificar outliers contextuales, es decir, valores que pueden ser normales en el contexto global pero anómalos en su vecindario local. Esto resulta particularmente relevante en datos médicos donde ciertos valores pueden ser normales para un grupo demográfico pero anómalos para otro.

### **Integración y síntesis de resultados**

La implementación paralela de estos cuatro métodos me ha proporcionado una visión multidimensional de las anomalías presentes en el dataset. He observado que, aunque existe cierta convergencia en los outliers detectados por diferentes métodos, cada técnica aporta hallazgos únicos que enriquecen el análisis global. Por ejemplo, mientras el IQR tiende a identificar valores extremos absolutos, el LOF ha sido más efectivo detectando clusters anómalos de pacientes con combinaciones inusuales de síntomas.

Esta estrategia metodológica múltiple no solo aumenta la robustez del proceso de detección, sino que también me permite tomar decisiones más informadas sobre qué outliers representan errores de datos versus casos médicos genuinamente excepcionales que deben preservarse para el entrenamiento del modelo.

### 6.1.1.2. Análisis visual y estadístico

En la fase de diseño del proceso analítico, el análisis estadístico emerge como piedra angular para desentrañar el comportamiento intrínseco de las variables cardiovasculares presentes en nuestro dataset. Esta sección aborda la planificación y estructuración de diversas técnicas de visualización gráfica, histogramas, diagramas de dispersión, diagramas de barras y boxplots, con el fin de representar e interpretar los datos clínicos de manera efectiva. A través de estas herramientas visuales, pretendo revelar patrones subyacentes, correlaciones significativas y tendencias ocultas que resultan fundamentales para construir modelos predictivos robustos en el ámbito de la salud cardiovascular.

Los histogramas constituirán mi primera línea de análisis para examinar la distribución de variables críticas como los niveles de colesterol (total, LDL, HDL), presión arterial sistólica y diastólica, y marcadores de riesgo metabólico. Estas representaciones me permitirán no solo identificar rangos de normalidad clínica, sino también detectar posibles sesgos en la muestra y comprender la naturaleza estadística de cada parámetro biomédico. Particularmente, me interesa observar si las distribuciones siguen patrones gaussianos o presentan asimetrías que podrían influir en la selección posterior de algoritmos de machine learning.

Los diagramas de caja y bigotes serán especialmente valiosos para visualizar la dispersión y tendencia central de biomarcadores clave. Su eficacia para identificar outliers, tema que, como he analizado extensamente, presenta peculiaridades en este dataset, los convierte en herramientas indispensables. Planeo crear boxplots estratificados por variables categóricas como género, grupo etario o presencia de comorbilidades, buscando heterogeneidades que podrían informar el desarrollo de modelos personalizados.

Para variables categóricas como la presencia o no de ciertas condiciones médicas como la obesidad o la diabetes, implementaré gráficos de barras y,

cuando sea apropiado, diagramas circulares para ilustrar proporciones. Sin embargo, soy consciente de las limitaciones de los pie charts en contextos científicos, por lo que los reservaré únicamente para casos donde las categorías sean pocas y las proporciones claramente diferenciables.

También diseñaré scatter plots para examinar relaciones bivariadas entre las variables más relevantes. Esta técnica exhaustiva, aunque computacionalmente intensiva, proporciona una visión holística de cómo cada variable se relaciona con las demás, revelando potenciales interacciones no lineales que métodos más simples podrían pasar por alto.

Al establecer este marco comprehensivo de visualización, no solo busco cumplir con los requisitos técnicos del análisis exploratorio, sino también desarrollar una intuición profunda sobre la naturaleza de los datos cardíacos. Cada gráfico diseñado servirá como una ventana hacia aspectos específicos del fenómeno que estudio: el riesgo de infarto de miocardio en población india.

Mi experiencia previa trabajando con datasets médicos me ha enseñado que la visualización efectiva trasciende la mera representación gráfica, es un proceso iterativo de descubrimiento donde cada insight visual plantea nuevas preguntas y sugiere análisis adicionales. Por ello, este diseño contempla flexibilidad para incorporar visualizaciones emergentes conforme profundice en los datos.

Finalmente, este diseño contempla la flexibilidad para incorporar nuevas visualizaciones según emergan hallazgos durante el análisis. Mi experiencia con datasets biomédicos me ha enseñado que la visualización trasciende la simple representación gráfica, es un proceso iterativo donde cada insight plantea nuevas interrogantes. Por ello, mantengo apertura para adaptar las técnicas según evolucione mi comprensión de los datos cardíacos.

## 6.1.2. Desarrollo

La fase de desarrollo del proceso de ciencia de datos se centra en la implementación práctica de las metodologías y herramientas delineadas durante la etapa de diseño. Esta fase involucra la ejecución real de las tareas de recopilación, preprocesamiento y análisis de datos, asegurando que los planes teóricos se traduzcan en pasos accionables. Mediante el seguimiento meticuloso del diseño establecido, el proyecto garantiza que el análisis de datos sea exhaustivo y confiable, proporcionando una base robusta para el desarrollo de modelos de IA precisos.

### 6.1.2.1. Detección de outliers

La implementación práctica de los algoritmos seleccionados se materializó a través de un conjunto de funciones Python diseñadas para procesar sistemáticamente el dataset cardiovascular. El código desarrollado estructura cada método de detección en funciones independientes que facilitan tanto la modularidad como la reutilización.

Para el método IQR, la función *detectar\_outliers\_IQR* realiza el cálculo secuencial de los estadísticos necesarios. Primero determina los cuartiles mediante el método *quantile()* de pandas, posteriormente calcula el rango intercuartílico y finalmente establece los límites de detección. La identificación de outliers se ejecuta mediante una consulta booleana que filtra las observaciones fuera de los límites establecidos.

```
def detectar_outliers_IQR(df, columna):
    Q1 = df[columna].quantile(0.25)
    Q3 = df[columna].quantile(0.75)
    IQR = Q3 - Q1
    limite_inferior = Q1 - 1.5 * IQR
    limite_superior = Q3 + 1.5 * IQR

    outliers = df[(df[columna] < limite_inferior) | (df[columna] > limite_superior)]
    print(f'Variable: {columna}')
    print(f' - Q1: {Q1}')
    print(f' - Q3: {Q3}')
    print(f' - IQR: {IQR}')
    print(f' - Límite inferior: {limite_inferior}')
    print(f' - Límite superior: {limite_superior}')
    print(f' - Total de outliers detectados: {outliers.shape[0]}')
    print('-----')

variables = ['Age', 'LDL_Level', 'HDL_Level', 'systolic_BP', 'Diastolic_BP',
            'Cholesterol_Level', 'Triglyceride_Level', 'Stress_Level', 'Emergency_Response_Time', 'Annual_Income']

for var in variables:
    detectar_outliers_IQR(heartAttackPrediction_India, var)
```

Ilustración 6: Método IQR

La implementación del z-score aprovecha la funcionalidad de la biblioteca *scipy.stats*, importando específicamente la función *zscore*. El código procesa cada variable calculando las puntuaciones estandarizadas y aplicando el umbral definido mediante la función valor absoluto de *numpy*. La estructura condicional *np.abs(z\_scores) > umbral* permite identificar desviaciones tanto positivas como negativas.

```
def detectar_outliers_zscore(df, columna, umbral=3):
    z_scores = zscore(df[columna])
    outliers = df[np.abs(z_scores) > umbral]

    print(f'Variable: {columna}')
    print(f' - Media: {df[columna].mean()}' )
    print(f' - Desviación estándar: {df[columna].std()}' )
    print(f' - Umbral Z-score: ±{umbral}' )
    print(f' - Total de outliers detectados: {outliers.shape[0]}')
    print('-----')

variables = ['Age', 'LDL_Level', 'HDL_Level', 'Systolic_BP', 'Diastolic_BP',
            'Cholesterol_Level', 'Triglyceride_Level', 'Stress_Level',
            'Emergency_Response_Time', 'Annual_Income']

for var in variables:
    detectar_outliers_zscore(heartAttackPrediction_India, var)
```

Ilustración 7: Método Z-Score

Para Isolation Forest, el código instancia el modelo desde *sklearn.ensemble* con los parámetros especificados. Un aspecto técnico relevante es el preprocesamiento mediante *dropna()* que garantiza la integridad de los datos antes del entrenamiento. El modelo se ajusta con *fit()* y las predicciones se obtienen mediante *predict()*, donde los valores -1 indican anomalías según la convención de scikit-learn.

```
def detectar_outliers_isolation_forest(df, columna, contamination=0.01):
    model = IsolationForest(contamination=contamination, random_state=42)
    df_filtered = df[[columna]].dropna()
    model.fit(df_filtered)
    df_filtered['outlier'] = model.predict(df_filtered)
    outliers = df_filtered[df_filtered['outlier'] == -1]

    print(f'Variable: {columna}')
    print(f' - Contaminación estimada: {contamination}' )
    print(f' - Total de outliers detectados: {outliers.shape[0]}')
    print('-----')

for var in variables:
    detectar_outliers_isolation_forest(heartAttackPrediction_India, var)
```

Ilustración 8: Método Isolation Forest

La implementación de Local Outlier Factor sigue un patrón similar, instanciando el algoritmo desde `sklearn.neighbors`. El procesamiento incluye el mismo tratamiento de valores faltantes y utiliza el método `fit_predict()` que combina entrenamiento y predicción en una sola llamada, optimizando el flujo de ejecución.

```
def detectar_outliers_lof(df, columna, n_neighbors=15):
    df_filtered = df[[columna]].dropna()
    lof = LocalOutlierFactor(n_neighbors=n_neighbors)
    y_pred = lof.fit_predict(df_filtered)
    outliers = df_filtered[y_pred == -1]

    print(f'Variable: {columna}')
    print(f' - Vecinos considerados: {n_neighbors}')
    print(f' - Total de outliers detectados: {outliers.shape[0]}')
    print('-----')

for var in variables:
    detectar_outliers_lof(heartAttackPrediction_India, var)
```

Ilustración 9: Método Local Outlier Factor

Un elemento común en todas las funciones es la generación estructurada de información de salida. Cada función imprime metadatos relevantes del proceso: parámetros utilizados, estadísticos calculados y número de anomalías detectadas. Esta decisión de implementación facilita el seguimiento del análisis sin necesidad de código adicional de *logging*.

La aplicación sistemática se logra mediante un bucle que itera sobre la lista de variables predefinida:

```
variables = ['Age', 'LDL_Level', 'HDL_Level', 'Systolic_BP', 'Diastolic_BP',
            'Cholesterol_Level', 'Triglyceride_Level', 'Stress_Level',
            'Emergency_Response_Time', 'Annual_Income']
```

Ilustración 10: Variables Detección Outliers

con cada método, generando una salida uniforme que simplifica el análisis posterior de resultados.

El manejo de excepciones, aunque no explícito en el código mostrado, se implementa implícitamente a través del uso de `dropna()` antes de aplicar

algoritmos sensibles a valores faltantes. Esta estrategia preventiva evita errores de ejecución manteniendo la robustez del pipeline.

La parametrización de las funciones permite ajustes dinámicos sin modificar el código base. Por ejemplo, el umbral del z-score y el parámetro de contaminación de Isolation Forest pueden modificarse en tiempo de ejecución, facilitando la experimentación con diferentes configuraciones.

El código final representa una implementación eficiente y mantenible de los cuatro métodos de detección, proporcionando una base sólida para el análisis exhaustivo de anomalías en el dataset cardiovascular.

### **6.1.2.2. Análisis visual y estadístico**

La transición de la fase de diseño a la implementación práctica del análisis estadístico representó un momento crucial en mi investigación. Durante esta etapa de desarrollo, me enfoqué en materializar las estrategias previamente planificadas, transformando conceptos abstractos en código funcional y procesamiento real de datos.

Mi primer desafío consistió en establecer un flujo de trabajo eficiente para el procesamiento sistemático de las variables cardiovasculares. Implementé un pipeline de análisis que me permitiera aplicar consistentemente las técnicas estadísticas seleccionadas, manteniendo la flexibilidad necesaria para ajustes iterativos. La estructuración del código siguió principios de programación modular, facilitando tanto la depuración como la eventual reproducibilidad del análisis.

La implementación de las visualizaciones requirió decisiones técnicas específicas que no había contemplado completamente durante el diseño. Por ejemplo, la parametrización de cada tipo de gráfico demandó un balance cuidadoso entre automatización y control manual. Desarrollé funciones

genéricas que pudieran adaptarse a las características particulares de cada variable, considerando aspectos como el rango de valores, la naturaleza discreta o continua de los datos, y la presencia de valores faltantes.

Un aspecto fundamental fue la gestión eficiente de la memoria y el tiempo de cómputo. Con un dataset de más de 10,000 registros, algunas visualizaciones más complejas requerían optimización para ejecutarse en tiempos razonables. Implementé estrategias de procesamiento por lotes cuando fue necesario, especialmente para análisis que involucraban múltiples variables simultáneamente.

Durante esta fase, también establecí un sistema robusto de versionado para las visualizaciones generadas. Cada gráfico producido se almacenaba con metadatos que incluían los parámetros utilizados, la fecha de generación y las variables involucradas. Esta práctica, aunque inicialmente parecía excesiva, demostró su valor cuando necesité revisar análisis anteriores o comparar resultados entre diferentes iteraciones.

La codificación del análisis estadístico me permitió profundizar en aspectos técnicos que enriquecieron mi comprensión del dataset. Por ejemplo, al implementar los cálculos de medidas de tendencia central y dispersión, pude apreciar matices en la distribución de los datos que no eran evidentes en el análisis teórico inicial. La programación de tests estadísticos para evaluar normalidad y homogeneidad de varianzas reveló características importantes sobre la naturaleza de las variables cardiovasculares.

Me resultó particularmente instructivo el proceso de implementar visualizaciones estratificadas. La segmentación de los análisis por variables categóricas como género, grupo etario o presencia de factores de riesgo requirió una arquitectura de código más sofisticada de lo anticipado. Desarrollé un sistema que permitiera aplicar filtros dinámicos manteniendo la coherencia visual entre diferentes subgrupos.

La fase de desarrollo también incluyó la implementación de validaciones y controles de calidad. Establecí rutinas para verificar la integridad de los datos antes de cada visualización, asegurando que valores faltantes o anómalos no distorsionaran los resultados. Este enfoque preventivo me ahorró considerable tiempo de debugging y aumentó mi confianza en la fiabilidad de los análisis producidos.

Un elemento que cobró especial relevancia fue la documentación inline del código. Mientras implementaba cada función, incluía comentarios detallados sobre las decisiones tomadas, las asunciones estadísticas subyacentes y las limitaciones conocidas. Esta práctica no solo facilitó revisiones posteriores, sino que también me obligó a reflexionar críticamente sobre cada paso del proceso analítico.

La experiencia de traducir diseños conceptuales en implementaciones concretas me proporcionó valiosas lecciones sobre la naturaleza iterativa del análisis de datos. Frecuentemente, la ejecución práctica revelaba consideraciones que no había anticipado en la fase de planificación, requiriendo ajustes y refinamientos continuos. Esta flexibilidad adaptativa se convirtió en una característica central de mi aproximación metodológica.

Conforme avanzaba en la implementación, fui desarrollando una biblioteca personal de funciones reutilizables para análisis cardiovascular. Estas herramientas, refinadas a través de múltiples iteraciones, representan uno de los productos tangibles más valiosos de esta fase. Su modularidad y documentación las convierten en recursos que podré aprovechar en futuros proyectos de investigación en el ámbito de la salud digital.

La culminación de esta fase de desarrollo fue un framework completo y funcional para el análisis estadístico de datos cardíacos. Más allá de los scripts y funciones individuales, había construido un sistema integrado que podía procesar, analizar y visualizar información clínica de manera

eficiente y reproducible. Este logro técnico sentó las bases sólidas para la interpretación de resultados que seguiría en la próxima fase del proyecto.

### **6.1.3. Resultados**

La fase de resultados del proceso de ciencia de datos constituye el momento crucial para evaluar la efectividad e impacto de las metodologías implementadas. Esta sección presenta los outcomes obtenidos de las etapas de preprocesamiento y análisis de datos, incluyendo la eficacia de la detección de outliers y la calidad global del dataset preparado. A través de un análisis cuidadoso e interpretación de estos resultados, el proyecto puede validar el éxito del enfoque analítico adoptado y asegurar que el conjunto de datos está listo para el posterior desarrollo de modelos de inteligencia artificial.

#### **6.1.3.1. Detección de outliers**

Los resultados obtenidos mediante el método del rango intercuartílico han revelado un fenómeno inesperado pero significativo: la ausencia total de outliers en todas las variables analizadas. Este hallazgo, que inicialmente podría interpretarse como una anomalía metodológica, merece un análisis detallado que considere tanto las características intrínsecas del dataset como las implicaciones para el desarrollo posterior del modelo predictivo.

Variable: Age	Variable: Cholesterol_Level
- Q1: 35.0	- Q1: 187.0
- Q3: 64.0	- Q3: 262.0
- IQR: 29.0	- IQR: 75.0
- Límite inferior: -8.5	- Límite inferior: 74.5
- Límite superior: 107.5	- Límite superior: 374.5
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: LDL_Level	Variable: Triglyceride_Level
- Q1: 86.0	- Q1: 114.0
- Q3: 161.0	- Q3: 236.0
- IQR: 75.0	- IQR: 122.0
- Límite inferior: -26.5	- Límite inferior: -69.0
- Límite superior: 273.5	- Límite superior: 419.0
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: HDL_Level	Variable: Stress_Level
- Q1: 34.0	- Q1: 3.0
- Q3: 65.0	- Q3: 8.0
- IQR: 31.0	- IQR: 5.0
- Límite inferior: -12.5	- Límite inferior: -4.5
- Límite superior: 111.5	- Límite superior: 15.5
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: Systolic_BP	Variable: Emergency_Response_Time
- Q1: 112.0	- Q1: 110.0
- Q3: 157.0	- Q3: 304.0
- IQR: 45.0	- IQR: 194.0
- Límite inferior: 44.5	- Límite inferior: -181.0
- Límite superior: 224.5	- Límite superior: 595.0
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: Diastolic_BP	Variable: Annual_Income
- Q1: 74.0	- Q1: 535783.75
- Q3: 104.0	- Q3: 1501669.75
- IQR: 30.0	- IQR: 965886.0
- Límite inferior: 29.0	- Límite inferior: -913045.25
- Límite superior: 149.0	- Límite superior: 2950498.75
- Total de outliers detectados: 0	- Total de outliers detectados: 0

Ilustración 11: Resultados Método IQR

## Análisis de la distribución por variables

Comenzando por la variable edad, observamos un rango intercuartílico de 29 años con el primer cuartil situado en 35 años y el tercero en 64. Estos valores sugieren una población de estudio concentrada en edades medias y avanzadas, lo cual es coherente con el perfil epidemiológico de las enfermedades cardiovasculares. El límite inferior calculado de -8.5 años, aunque matemáticamente correcto, carece de sentido biológico, mientras que el superior de 107.5 años establece un umbral que abarcaría incluso casos de longevidad extrema.

Los marcadores lipídicos presentan patrones particularmente interesantes. El LDL, con un IQR de 75 mg/dL, muestra una distribución considerablemente amplia (Q1=86, Q3=161), reflejando la heterogeneidad esperada en una población con diversos grados de riesgo cardiovascular. De

manera similar, el HDL exhibe un rango que va desde niveles considerados de riesgo ( $Q1=34$  mg/dL) hasta valores protectores ( $Q3=65$  mg/dL). La ausencia de outliers en estos parámetros sugiere que el dataset captura adecuadamente el espectro completo de perfiles lipídicos sin incluir valores extremos poco realistas.

Las mediciones de presión arterial revelan distribuciones que abarcan desde valores normales hasta hipertensión moderada. La presión sistólica, con un  $Q3$  de 157 mmHg, se sitúa justo en el umbral de hipertensión grado 2 según las guías clínicas actuales. El hecho de que no se detecten outliers con un límite superior de 224.5 mmHg indica que el dataset no contiene casos de crisis hipertensivas severas, lo cual podría representar una limitación si busco que mi modelo sea capaz de identificar emergencias cardiovasculares agudas.

### Interpretación estadística y clínica

La variable de colesterol total muestra una distribución que refleja fielmente la prevalencia de dislipidemia en poblaciones con riesgo cardiovascular. Con un  $Q3$  de 262 mg/dL, observamos que el 75% de la muestra presenta valores por debajo del umbral de alto riesgo ( $>270$  mg/dL) establecido por las sociedades cardiológicas. El límite superior calculado de 374.5 mg/dL abarcaría incluso casos de hipercolesterolemia familiar heterocigota, explicando la ausencia de outliers.

Un aspecto que merece especial atención es la distribución de los triglicéridos, donde encontramos el IQR más amplio (122 mg/dL) entre los marcadores bioquímicos. Esta variabilidad refleja la naturaleza volátil de este parámetro, influenciado por factores dietéticos y metabólicos. El límite inferior negativo (-69 mg/dL) nuevamente carece de sentido fisiológico, mientras que el superior de 419 mg/dL incluiría casos de hipertrigliceridemia moderada a severa.

La variable de nivel de estrés, presumiblemente medida en una escala ordinal, presenta una concentración notable entre los valores 3 y 8. Esta distribución sugiere que la herramienta de evaluación utilizada captura principalmente niveles moderados de estrés percibido, sin representación significativa de los extremos (ausencia total de estrés o estrés severo incapacitante).

### **Implicaciones metodológicas**

Los tiempos de respuesta de emergencia muestran la mayor variabilidad relativa ( $IQR=194$  minutos), lo que podría reflejar diferencias geográficas, logísticas o de infraestructura sanitaria en la región de estudio. La ausencia de outliers con un límite superior de casi 10 horas sugiere que incluso los casos más extremos de demora se encuentran dentro de lo esperado para el contexto estudiado.

La variable de ingresos anuales, expresada en lo que parecen ser rupias indias, exhibe una distribución que abarca desde clase media hasta niveles socioeconómicos altos. El amplio  $IQR$  de 965,886 unidades monetarias refleja la disparidad económica característica de muchas sociedades contemporáneas, aunque la ausencia de outliers podría indicar una subrepresentación de extremos socioeconómicos.

### **Reflexiones críticas y consideraciones futuras**

Este resultado aparentemente "perfecto" de cero outliers plantea varias hipótesis que merecen consideración. Primera, es posible que el dataset haya sido sometido a un preprocesamiento previo que eliminó valores extremos, lo cual representaría una pérdida de información potencialmente valiosa para mi modelo. Segunda, la muestra podría provenir de un estudio controlado con criterios de inclusión/exclusión estrictos que naturalmente limitaron la variabilidad. Tercera, y quizás más preocupante desde una perspectiva de machine learning, podríamos estar ante un caso de "overfitting" inverso,

donde la homogeneidad excesiva limita la capacidad generalizadora del modelo.

La ausencia de outliers mediante IQR también sugiere que necesitaremos recurrir a métodos más sensibles para identificar patrones anómalos sutiles. Los algoritmos basados en densidad o distancia, como Isolation Forest y LOF implementados posteriormente, podrían revelar estructuras de anomalía más complejas que el simple análisis univariado no detecta.

Desde una perspectiva práctica, estos resultados implican que el dataset está "demasiado limpio" para capturar la complejidad real de los datos médicos. En mi experiencia analizando datasets biomédicos, siempre encuentro cierta proporción de valores extremos que, aunque válidos clínicamente, desafían las distribuciones estadísticas convencionales. Su ausencia aquí me hace cuestionar si estamos perdiendo casos precisamente aquellos que podrían ser más informativos para predecir eventos cardiovasculares graves.

Para el desarrollo posterior del modelo, estos resultados sugieren que deberemos ser especialmente cuidadosos con la validación externa, ya que modelos entrenados en datos excesivamente homogéneos tienden a mostrar degradación significativa cuando se enfrentan a la variabilidad del mundo real.

Variable: Age	Variable: Cholesterol_Level
- Media: 49.3949	- Media: 224.753
- Desviación estándar: 17.28030135360744	- Desviación estándar: 43.35917195679915
- Umbral Z-score: ±3	- Umbral Z-score: ±3
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: LDL_Level	Variable: Triglyceride_Level
- Media: 123.8721	- Media: 174.7333
- Desviación estándar: 43.41076584269281	- Desviación estándar: 71.16344704080511
- Umbral Z-score: ±3	- Umbral Z-score: ±3
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: HDL_Level	Variable: Stress_Level
- Media: 49.3355	- Media: 5.5188
- Desviación estándar: 17.399896993704854	- Desviación estándar: 2.866263788769394
- Umbral Z-score: ±3	- Umbral Z-score: ±3
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: Systolic_BP	Variable: Emergency_Response_Time
- Media: 134.7259	- Media: 206.3834
- Desviación estándar: 25.849077095104708	- Desviación estándar: 112.39171052217067
- Umbral Z-score: ±3	- Umbral Z-score: ±3
- Total de outliers detectados: 0	- Total de outliers detectados: 0
Variable: Diastolic_BP	Variable: Annual_Income
- Media: 89.312	- Media: 1022062.1708
- Desviación estándar: 17.39648584547975	- Desviación estándar: 560597.7922241136
- Umbral Z-score: ±3	- Umbral Z-score: ±3
- Total de outliers detectados: 0	- Total de outliers detectados: 0

Ilustración 12: Resultados Método Z-Score

## Análisis de resultados mediante Z-Score

Los resultados obtenidos a través del método z-score confirman y amplifican las observaciones iniciales del análisis IQR: nuevamente, no se detecta ni un solo outlier en todo el conjunto de variables. Este fenómeno, ahora corroborado por dos metodologías estadísticas independientes, refuerza mis sospechas sobre la naturaleza excepcionalmente homogénea del dataset y plantea interrogantes fundamentales sobre su representatividad poblacional.

Examinando los parámetros estadísticos obtenidos, emergen patrones reveladores sobre la estructura de los datos. La edad media de 49.39 años con una desviación estándar de 17.28 indica una distribución relativamente simétrica centrada en la mediana edad, lo cual concuerda con el perfil epidemiológico típico de pacientes con riesgo cardiovascular. Sin embargo, la ausencia de valores que superen 3 desviaciones estándar sugiere que no hay representación de casos pediátricos ni geriátricos extremos, una limitación potencial si consideramos que tanto los muy jóvenes con cardiopatías congénitas como los nonagenarios representan poblaciones de interés clínico.

Los marcadores bioquímicos presentan características particularmente intrigantes. El LDL medio de 123.87 mg/dL se sitúa ligeramente por encima del óptimo clínico (<100 mg/dL), con una dispersión moderada ( $\sigma=43.41$ ) que teóricamente debería capturar tanto casos de hipolipidemia como de hipercolesterolemia severa. No obstante, la ausencia de outliers implica que ningún valor supera los 254 mg/dL aproximadamente (media + 3 $\sigma$ ), lo cual excluye casos de hipercolesterolemia familiar homocigota o heterocigota severa, condiciones raras pero cruciales para modelos predictivos de alto riesgo.

El HDL muestra un comportamiento similar, con una media de 49.34 mg/dL que roza el límite inferior deseable para hombres (>40 mg/dL). La desviación estándar de 17.40 es proporcionalmente alta, sugiriendo variabilidad sustancial, paradójicamente no produce outliers. Esto implica que el rango efectivo oscila entre aproximadamente -2.9 y 101.5 mg/dL, donde obviamente el límite inferior carece de sentido biológico y el superior apenas captura casos de HDL excepcionalmente protector.

Las métricas de presión arterial revelan una población con tendencia hipertensiva moderada. La sistólica media de 134.73 mmHg clasifica como prehipertensión según criterios actuales, mientras que la diastólica de 89.31 mmHg bordea el umbral de hipertensión estadio 1. Las desviaciones estándar relativamente estrechas (25.85 y 17.40 respectivamente) sugieren poca representación de casos extremos, ni crisis hipertensivas ni hipotensión severa aparecen en el dataset.

Un aspecto que me resulta particularmente llamativo es la distribución del colesterol total. Con una media de 224.75 mg/dL, la población se sitúa en el rango "borderline high" según las guías ATP III. La desviación estándar de 43.36 implica que el límite superior de 3 $\sigma$  alcanzaría aproximadamente 355 mg/dL, un valor que, aunque elevado, no es infrecuente en la práctica clínica. La ausencia de tales casos sugiere un sesgo de selección o preprocesamiento agresivo.

Los triglicéridos presentan la mayor variabilidad relativa ( $CV = 40.7\%$ ), con una media de 174.73 mg/dL que indica riesgo moderado. Esta alta dispersión normalmente produciría outliers en los extremos, especialmente considerando la distribución típicamente sesgada de este parámetro. Su ausencia es particularmente sospechosa dado que los triglicéridos son notoriamente volátiles y sensibles a factores como el ayuno, la dieta y el consumo de alcohol.

La variable de estrés, con media 5.52 y desviación 2.87, sugiere una escala probablemente del 1 al 10. El hecho de que ningún valor exceda media  $\pm 3\sigma$  (rango efectivo: -3.08 a 14.12) indica que la escala completa no está siendo utilizada, lo cual podría limitar la capacidad discriminativa de esta variable en el modelo final.

Los tiempos de respuesta de emergencia muestran la mayor variabilidad absoluta ( $\sigma=112.39$  minutos), reflejando probablemente diferencias geográficas y logísticas. Sin embargo, incluso con esta dispersión, no encontramos casos que superen las 9 horas aproximadamente (media  $+ 3\sigma \approx 543$  minutos), lo cual podría ser problemático si quisiera que mi modelo sea aplicable en zonas rurales remotas donde los tiempos pueden ser considerablemente mayores.

La variable económica presenta una distribución con coeficiente de variación del 54.8%, indicando disparidad socioeconómica sustancial. Aun así, la ausencia de outliers sugiere que no captura ni la pobreza extrema ni la riqueza excepcional, potencialmente limitando la generalización del modelo a estratos socioeconómicos extremos.

## Reflexiones metodológicas

Este patrón consistente de cero outliers mediante z-score en un dataset de más de 10,000 observaciones es, francamente, estadísticamente improbable. En una distribución verdaderamente normal, esperaría aproximadamente un 0.3% de valores más allá de  $\pm 3\sigma$ , lo que debería traducirse en unos 26 outliers totales. Su completa ausencia sugiere fuertemente que:

- El dataset ha sido sometido a un proceso de "sanitización" que eliminó valores extremos pero clínicamente válidos.
- Los datos provienen de un estudio con criterios de inclusión tan estrictos que la variabilidad natural fue artificialmente restringida.
- Existe algún tipo de censura o truncamiento en los valores registrados.

Desde mi perspectiva como futuro profesional en IA aplicada, esta homogeneidad excesiva representa un desafío significativo. Los modelos entrenados en datos "demasiado perfectos" suelen mostrar poor generalization cuando se enfrentan a la complejidad y el ruido del mundo real. Es precisamente en los extremos donde a menudo encontramos los casos más informativos para la predicción de eventos adversos.

La convergencia de resultados entre IQR y z-score, aunque matemáticamente coherente, refuerza mi preocupación sobre la representatividad del dataset. Para el desarrollo de un modelo robusto de predicción cardiovascular, necesitaría idealmente capturar toda la gama de presentaciones clínicas, incluyendo esos casos "raros pero reales" que los métodos estadísticos tradicionales clasificarían como outliers.

Variable: Age	Variable: Cholesterol_Level
- Contaminación estimada: 0.01	- Contaminación estimada: 0.01
- Total de outliers detectados: 0	- Total de outliers detectados: 70
-----	-----
Variable: LDL_Level	Variable: Triglyceride_Level
- Contaminación estimada: 0.01	- Contaminación estimada: 0.01
- Total de outliers detectados: 77	- Total de outliers detectados: 76
-----	-----
Variable: HDL_Level	Variable: Stress_Level
- Contaminación estimada: 0.01	- Contaminación estimada: 0.01
- Total de outliers detectados: 0	- Total de outliers detectados: 0
-----	-----
Variable: Systolic_BP	Variable: Emergency_Response_Time
- Contaminación estimada: 0.01	- Contaminación estimada: 0.01
- Total de outliers detectados: 0	- Total de outliers detectados: 99
-----	-----
Variable: Diastolic_BP	Variable: Annual_Income
- Contaminación estimada: 0.01	- Contaminación estimada: 0.01
- Total de outliers detectados: 0	- Total de outliers detectados: 100

Ilustración 13: Resultados Método Isolation Forest

### Análisis de resultados mediante Isolation Forest

La aplicación del algoritmo Isolation Forest marca un punto de inflexión revelador en mi análisis de detección de anomalías. Por primera vez, observo la identificación efectiva de outliers en varias variables clave, contrastando drásticamente con los resultados nulos obtenidos mediante IQR y z-score. Este cambio paradigmático no solo valida mis sospechas previas sobre las limitaciones de los métodos estadísticos tradicionales, sino que también arroja luz sobre la complejidad subyacente de los datos cardiovaseulares.

El patrón de detección resulta fascinante desde una perspectiva analítica. Mientras que variables demográficas y fisiológicas básicas como edad, HDL, presión arterial y nivel de estrés no presentan anomalías detectables, los marcadores bioquímicos y socioeconómicos revelan una historia completamente diferente. Esta dicotomía sugiere que el algoritmo está capturando patrones de aislamiento más sutiles que trascienden las simples desviaciones estadísticas univariadas.

Profundizando en los hallazgos específicos, el LDL presenta 77 casos anómalos, lo que representa aproximadamente el 0.88% del dataset total,

ligeramente por debajo del 1% esperado según el parámetro de contaminación. Este resultado me parece particularmente relevante desde una perspectiva clínica, ya que el LDL es reconocido como uno de los predictores más potentes de riesgo cardiovascular. La capacidad del Isolation Forest para identificar estos casos sugiere la presencia de patrones de valores que, aunque no extremos en términos absolutos, resultan inusuales en su contexto multidimensional.

El colesterol total, con 70 outliers detectados, muestra una correlación esperada con los hallazgos del LDL, dado que este último constituye aproximadamente el 60-70% del colesterol total en condiciones normales. Sin embargo, la ligera discrepancia en el número de anomalías (77 vs 70) indica que el algoritmo está identificando casos donde la relación entre estos parámetros se desvía de lo esperado, un insight valioso que los métodos univariados no podrían capturar.

Los triglicéridos, con 76 casos anómalos, confirman mi hipótesis previa sobre la volatilidad inherente de este marcador. A diferencia de los métodos estadísticos tradicionales que no detectaron ninguna anomalía, Isolation Forest reconoce patrones de aislamiento que probablemente reflejan estados metabólicos atípicos. Considerando que los triglicéridos elevados frecuentemente coexisten con otras alteraciones lipídicas en el síndrome metabólico, estos outliers podrían representar casos de particular interés predictivo.

Quizás el hallazgo más intrigante sea la detección de 99 outliers en los tiempos de respuesta de emergencia. Esta variable, que mostró la mayor desviación estándar en el análisis z-score, finalmente revela su naturaleza anómala cuando se examina mediante un algoritmo no paramétrico. Estos casos probablemente representan situaciones donde factores geográficos, logísticos o sistémicos generaron demoras excepcionales que, aunque matemáticamente dentro de 3 desviaciones estándar, constituyen eventos claramente aislados en el espacio de características.

La variable de ingresos anuales, con exactamente 100 outliers detectados, presenta un resultado que me genera cierta suspicacia. La precisión del número (coincidiendo casi perfectamente con el 1% esperado de 8,763 observaciones) sugiere que el algoritmo podría estar alcanzando su límite de contaminación predefinido. Esto podría indicar que existen aún más anomalías económicas en el dataset que las que el parámetro actual permite detectar.

### **Reflexiones técnicas y metodológicas**

La efectividad diferencial del Isolation Forest respecto a los métodos tradicionales ilustra perfectamente las ventajas de los enfoques basados en machine learning para la detección de anomalías. Mientras que IQR y z-score operan bajo asunciones de normalidad y simetría que raramente se cumplen en datos biomédicos reales, Isolation Forest explota el principio fundamental de que las anomalías son, por definición, más fáciles de aislar mediante particiones aleatorias del espacio de características.

Me resulta particularmente interesante observar cómo las variables que no mostraron outliers (edad, HDL, presión arterial, estrés) comparten características de distribuciones más "naturales" o esperadas en una población general. En contraste, aquellas con anomalías detectadas (marcadores lipídicos, tiempos de respuesta, ingresos) representan dimensiones donde la variabilidad puede estar influenciada por factores externos, patológicos o socioeconómicos que generan subpoblaciones distintivas.

Desde mi perspectiva, estos resultados subrayan una lección fundamental: la elección del método de detección de outliers no es meramente una decisión técnica, sino una que tiene profundas implicaciones para la comprensión de los datos y, ultimadamente, para la calidad de los modelos predictivos desarrollados. Los 422 outliers totales detectados por Isolation Forest representan casi el 5% del dataset, una proporción nada despreciable que

podría contener información crucial para predecir eventos cardiovasculares adversos.

La ausencia total de outliers en presión arterial mediante todos los métodos aplicados me genera particular inquietud. En mi experiencia analizando datos clínicos, siempre encuentro casos de hipertensión severa o crisis hipertensivas que deberían manifestarse como anomalías. Esta uniformidad sospechosa refuerza mi hipótesis de que el dataset ha sido sometido a algún tipo de preprocessamiento que eliminó casos extremos pero clínicamente relevantes.

### **Implicaciones para el modelado predictivo**

Los resultados del Isolation Forest tienen implicaciones directas para las fases posteriores del proyecto. La identificación de outliers en variables lipídicas sugiere la presencia de subgrupos de pacientes con perfiles metabólicos distintivos que podrían requerir estrategias de modelado específicas. Similarmente, los outliers en tiempos de respuesta podrían representar poblaciones con acceso limitado a servicios de emergencia, un factor crítico en la supervivencia post-infarto.

Para el desarrollo de modelos robustos, deberá decidir cuidadosamente cómo manejar estos outliers identificados. Mi inclinación inicial es hacia un enfoque conservador: mantener estos casos pero potencialmente aplicar técnicas de robust learning que minimicen su influencia indebida sin perder la información valiosa que puedan aportar sobre casos extremos pero reales.

La discrepancia entre métodos también plantea preguntas filosóficas sobre qué constituye verdaderamente una "anomalía" en el contexto médico. ¿Son estos 422 casos errores de medición, variantes biológicas raras, o precisamente los pacientes de mayor riesgo que mi modelo debe aprender a identificar? La respuesta a esta pregunta determinará en gran medida el éxito de mi sistema predictivo en escenarios clínicos reales.

Variable: Age	Variable: Cholesterol_Level
- Vecinos considerados: 15	- Vecinos considerados: 15
- Total de outliers detectados: 0	- Total de outliers detectados: 0
-----	-----
Variable: LDL_Level	Variable: Triglyceride_Level
- Vecinos considerados: 15	- Vecinos considerados: 15
- Total de outliers detectados: 0	- Total de outliers detectados: 0
-----	-----
Variable: HDL_Level	Variable: Stress_Level
- Vecinos considerados: 15	- Vecinos considerados: 15
- Total de outliers detectados: 0	- Total de outliers detectados: 0
-----	-----
Variable: Systolic_BP	Variable: Emergency_Response_Time
- Vecinos considerados: 15	- Vecinos considerados: 15
- Total de outliers detectados: 0	- Total de outliers detectados: 74
-----	-----
Variable: Diastolic_BP	Variable: Annual_Income
- Vecinos considerados: 15	- Vecinos considerados: 15
- Total de outliers detectados: 0	- Total de outliers detectados: 6

*Ilustración 14: Resultados Método Local Outlier Factor*

### Análisis de resultados mediante Local Outlier Factor

Los resultados obtenidos con el algoritmo Local Outlier Factor presentan un escenario peculiar que contrasta significativamente con los hallazgos previos del Isolation Forest. Esta técnica basada en densidad local, que evalúa la anomalía de cada punto en relación con su vecindario inmediato, ha arrojado resultados casi nulos en la mayoría de variables, identificando outliers únicamente en dos dimensiones: tiempo de respuesta de emergencia y nivel de ingresos.

La ausencia generalizada de anomalías detectadas por LOF en las variables clínicas fundamentales (marcadores lipídicos, presión arterial, edad) resulta, cuando menos, desconcertante. Este algoritmo, que considero particularmente robusto para identificar outliers contextuales, debería ser capaz de detectar observaciones que, aunque globalmente normales, presentan características inusuales respecto a su entorno local. El hecho de que no identifique ninguna anomalía en LDL, colesterol o triglicéridos, variables donde Isolation Forest encontró decenas de casos, sugiere diferencias fundamentales en cómo ambos algoritmos conceptualizan la "anormalidad".

Analizando los casos donde sí se detectaron anomalías, encontramos 74 outliers en tiempos de respuesta de emergencia. Esta cifra, aunque inferior a los 99 detectados por Isolation Forest, confirma que esta variable contiene patrones genuinamente anómalos. La naturaleza del LOF, que examina densidades locales, probablemente está capturando casos donde los tiempos de respuesta difieren drásticamente de otros casos geográficamente o contextualmente similares. Podríamos estar observando, por ejemplo, situaciones donde pacientes en zonas urbanas experimentaron demoras atípicas, o viceversa, casos rurales con tiempos sorprendentemente rápidos.

El hallazgo de apenas 6 outliers en la variable de ingresos anuales representa una reducción dramática respecto a los 100 casos identificados por Isolation Forest. Esta discrepancia tan marcada me lleva a reflexionar sobre la distribución subyacente de esta variable. Es posible que los ingresos presenten clusters bien definidos correspondientes a diferentes estratos socioeconómicos, y LOF solo está identificando aquellos casos que verdaderamente no pertenecen a ningún grupo establecido, quizás errores de registro o casos genuinamente excepcionales.

### **Reflexiones técnicas sobre la discrepancia entre métodos**

La divergencia entre Isolation Forest y LOF merece un análisis profundo. Mientras que Isolation Forest opera bajo el principio de que las anomalías son más fáciles de aislar mediante particiones aleatorias, LOF evalúa la densidad relativa de cada punto respecto a sus vecinos más cercanos. Esta diferencia fundamental en la filosofía de detección puede explicar por qué LOF es más conservador en sus identificaciones.

He estado reflexionando sobre el parámetro  $k=15$  utilizado para definir el vecindario. En mi experiencia trabajando con datasets médicos, este valor puede ser crítico. Un  $k$  demasiado pequeño hace al algoritmo sensible al ruido, mientras que uno muy grande puede suavizar excesivamente las densidades locales, perdiendo anomalías genuinas. Es posible que para variables como

los marcadores lipídicos, donde la distribución puede ser más continua y menos clustered, un valor diferente de  $k$  hubiera revelado más anomalías.

También considero relevante mencionar que LOF es particularmente efectivo cuando los datos forman clusters naturales con densidades variables. La ausencia casi total de detecciones en variables clínicas podría indicar que estas siguen distribuciones relativamente uniformes sin agrupaciones distintivas, un hallazgo que, aunque técnicamente válido, contradice mi intuición sobre la heterogeneidad esperada en poblaciones con riesgo cardiovascular.

### **Implicaciones metodológicas y consideraciones prácticas**

Desde mi perspectiva, estos resultados plantean cuestiones fundamentales sobre la selección de métodos de detección de anomalías en contextos biomédicos. La efectividad diferencial de LOF sugiere que no todos los algoritmos son igualmente apropiados para todos los tipos de datos médicos.

Me parece particularmente interesante que las únicas variables con outliers detectados (tiempo de respuesta e ingresos) sean precisamente aquellas de naturaleza no estrictamente biomédica. Esto podría indicar que LOF es más efectivo detectando anomalías en variables con distribuciones multimodales o con agrupaciones naturales basadas en factores externos (geografía, estrato socioeconómico) que en variables biológicas continuas.

La implementación univariada del LOF también podría estar limitando su efectividad. En realidad, las anomalías médicas raramente ocurren en una sola dimensión, un paciente con riesgo cardiovascular extremo típicamente presenta una constelación de valores anormales en múltiples parámetros. Una aplicación multivariada del LOF podría revelar patrones que el análisis univariado actual está pasando por alto.

## Síntesis y recomendaciones

Los resultados del LOF, vistos en conjunto con los otros métodos aplicados, pintan un cuadro complejo sobre la naturaleza de las anomalías en este dataset. La convergencia de IQR, z-score y ahora LOF en detectar pocas o ninguna anomalía en variables clínicas clave refuerza mi hipótesis sobre un posible preprocesamiento agresivo del dataset.

Sin embargo, no puedo descartar la posibilidad de que estemos ante un conjunto de datos genuinamente homogéneo, quizás proveniente de un estudio con criterios de inclusión muy específicos o de una población particularmente uniforme. En cualquier caso, la disparidad entre Isolation Forest y los demás métodos sugiere que debería confiar más en los hallazgos del primero, dado su mayor sensibilidad y su naturaleza no paramétrica.

Para las fases posteriores del proyecto, estos resultados me llevan a recomendar un enfoque cauteloso pero pragmático. Los outliers identificados por Isolation Forest merecen un análisis caso por caso antes de tomar decisiones sobre su inclusión o exclusión. Mientras tanto, la relativa "limpieza" del dataset según LOF podría facilitar el entrenamiento de modelos más estables, aunque potencialmente menos generalizables a poblaciones más heterogéneas.

En última instancia, esta experiencia con múltiples métodos de detección ha enriquecido mi comprensión sobre la complejidad inherente en el análisis de datos médicos. Cada algoritmo cuenta una historia diferente sobre los mismos datos, y es nuestra responsabilidad como científicos de datos interpretar estas narrativas de manera coherente y clínicamente significativa.



Ilustración 15: Resultados IF vs LOF Outliers

## Análisis visual comparativo de outliers detectados

La visualización comparativa de los resultados obtenidos mediante Isolation Forest y Local Outlier Factor revela patrones fascinantes que complementan y profundizan el análisis numérico previo. Los histogramas presentados ofrecen una perspectiva gráfica invaluable sobre la distribución y localización específica de las anomalías detectadas, permitiéndome extraer insights que los meros números no podían transmitir completamente.

Comenzando por el análisis del LDL, observo que Isolation Forest identifica outliers dispersos a lo largo de todo el espectro de valores, con una ligera concentración en los extremos de la distribución. Esta dispersión sugiere que el algoritmo está capturando no solo valores extremos absolutos, sino también observaciones que, dentro de rangos aparentemente normales, presentan características de aislamiento en el espacio de características. Como mencioné anteriormente, la ausencia total de detecciones por LOF en esta variable resulta aún más llamativa cuando visualizo la distribución aparentemente uniforme y continua de los datos.

El patrón observado en los niveles de colesterol muestra una historia similar pero con matices propios. Los outliers identificados por Isolation Forest se concentran notablemente en los valores más bajos del espectro (alrededor de 150-170 mg/dL), lo cual me resulta clínicamente intrigante. Estos casos podrían representar pacientes con hipolipidemia o aquellos bajo tratamiento intensivo con estatinas, condiciones que, aunque beneficiosas desde una perspectiva cardiovascular, resultan estadísticamente inusuales en la población general.

Los triglicéridos presentan quizás el patrón más revelador. La distribución de outliers muestra una clara tendencia hacia valores elevados, con múltiples detecciones en el rango de 250-300 mg/dL. Esta observación valida mi hipótesis previa sobre la volatilidad de este marcador y su susceptibilidad a factores dietéticos y metabólicos. La visualización también revela algunos

outliers en valores sorprendentemente bajos, lo cual podría indicar estados de malnutrición o condiciones metabólicas específicas que merecerían investigación adicional.

La variable de ingresos anuales exhibe un comportamiento particularmente interesante en la comparación visual entre métodos. Mientras que Isolation Forest detecta outliers distribuidos ampliamente a lo largo del espectro económico, LOF identifica apenas 6 casos, concentrados en valores extremadamente altos. Esta discrepancia visual confirma mi interpretación de que LOF está siendo excesivamente conservador, capturando únicamente aquellos casos que representan verdaderas anomalías contextuales, posiblemente errores de registro o individuos con ingresos genuinamente excepcionales para su contexto demográfico.

El análisis de los tiempos de respuesta de emergencia revela el patrón más consistente entre ambos métodos. Tanto Isolation Forest como LOF identifican concentraciones de outliers en tiempos superiores a 300 minutos, con algunos casos extremos acercándose a las 400 minutos. Esta convergencia visual refuerza mi convicción de que esta variable contiene anomalías genuinas y significativas, probablemente relacionadas con factores geográficos o sistémicos que afectan el acceso a servicios de emergencia.

### **Patrones emergentes y consideraciones clínicas**

La visualización conjunta de estos histogramas me permite identificar varios patrones emergentes que no eran evidentes en el análisis numérico aislado. Primero, observo que Isolation Forest tiende a identificar outliers de manera más uniforme a lo largo de las distribuciones, mientras que LOF (cuando detecta algo) se concentra en extremos absolutos. Esta diferencia visual confirma las distintas filosofías algorítmicas que analicé previamente.

Un aspecto que me llama poderosamente la atención es la aparente normalidad de las distribuciones base en todas las variables. Los histogramas

muestran curvas relativamente suaves sin evidencia de multimodalidad significativa o gaps evidentes en los datos. Esto refuerza mi sospecha de que el dataset ha sido sometido a algún tipo de proceso de homogenización que ha eliminado casos verdaderamente extremos.

Desde una perspectiva de modelado predictivo, la localización específica de estos outliers tiene implicaciones importantes. Los casos anómalos en valores bajos de colesterol y LDL, aunque estadísticamente inusuales, probablemente representan pacientes con menor riesgo cardiovascular. Por el contrario, los outliers en triglicéridos altos y tiempos de respuesta prolongados podrían constituir una subpoblación de alto riesgo que requiere atención especial en el modelo.

### **Reflexiones metodológicas finales**

Esta experiencia de análisis visual comparativo ha Enriquecido significativamente mi comprensión sobre la naturaleza de las anomalías en datos médicos. La capacidad de visualizar no solo dónde ocurren los outliers, sino también cómo se distribuyen dentro de cada variable, proporciona una dimensión interpretativa que los análisis puramente numéricos no pueden ofrecer.

Me resulta particularmente educativo observar cómo la elección del método de detección puede influir dramáticamente en mi o nuestra percepción de qué constituye una anomalía. Mientras que los métodos estadísticos tradicionales (IQR y z-score) no encontraron nada inusual, la visualización de los resultados de Isolation Forest revela un dataset con bolsones significativos de casos atípicos que podrían contener información crucial para la predicción de riesgo cardiovascular.

Para las fases posteriores del proyecto, estos gráficos me proporcionan una guía visual invaluable para la toma de decisiones sobre el manejo de outliers. Los casos identificados en rangos clínicamente plausibles (como colesterol

entre 150-170 mg/dL) probablemente deberían mantenerse, mientras que aquellos en extremos poco realistas podrían beneficiarse de una revisión más detallada o posible exclusión.

En última instancia, esta combinación de análisis numérico y visual ha demostrado ser fundamental para desarrollar una comprensión matizada de la estructura de anomalías en el dataset, preparando el terreno para un modelado más informado y clínicamente relevante en las fases subsiguientes del proyecto.

### **6.1.3.2. Análisis visual y estadístico**

La fase de resultados del análisis estadístico representa el momento culminante donde evalúo la efectividad de las técnicas gráficas implementadas y los insights extraídos de su aplicación al dataset cardiovascular. En esta sección presento los hallazgos obtenidos mediante histogramas, diagramas de dispersión, gráficos de densidad, boxplots y visualizaciones multivariadas, buscando comprender el comportamiento intrínseco de las variables predictoras del riesgo cardíaco.

A través del análisis detallado de estas representaciones visuales, he conseguido evaluar las distribuciones estadísticas de biomarcadores clave e identificar relaciones complejas entre variables clínicas. Ha resultado particularmente esclarecedor observar cómo la incorporación de información demográfica y factores de riesgo adicionales enriquece la interpretación de los patrones identificados, ofreciendo un marco más completo para entender la predisposición a eventos cardiovasculares.

Estos resultados son cruciales para confirmar que mi análisis estadístico ha descubierto patrones significativos y correlaciones clínicamente relevantes, estableciendo así cimientos sólidos para el desarrollo posterior de modelos de inteligencia artificial. La caracterización de estas propiedades subyacentes en los datos no solo valida mi aproximación metodológica, sino que también guía

decisiones críticas sobre la selección de features y la arquitectura de los algoritmos predictivos que implementaré.

Para facilitar una comprensión sistemática del análisis realizado, he estructurado los resultados en subsecciones temáticas que abordan distintas facetas del comportamiento de los datos. Esta organización permite examinar desde características univariadas fundamentales hasta interacciones multidimensionales sofisticadas, construyendo gradualmente una perspectiva integral del fenómeno estudiado. Considero que este enfoque progresivo resulta especialmente valioso cuando se trabaja con datos médicos complejos donde múltiples factores interactúan de formas no siempre evidentes.

Cada visualización presentada ha sido seleccionada no únicamente por su valor estadístico, sino también por su capacidad de revelar aspectos clínicamente significativos que podrían influir en la predicción del riesgo de infarto. Mi intención es que estos resultados sirvan como nexo entre el análisis exploratorio inicial y la implementación de modelos predictivos avanzados, garantizando que las decisiones de modelado estén firmemente respaldadas por evidencia empírica robusta.

#### **6.1.3.2.1. Distribuciones – Histogramas**

Al examinar estos histogramas, lo primero que me viene a la mente es que estoy ante un conjunto de datos que ha sido claramente procesado y estructurado con criterios muy específicos. Cada distribución cuenta su propia historia sobre cómo se ha construido este dataset, y sinceramente, algunas de estas historias me generan más preguntas que respuestas.

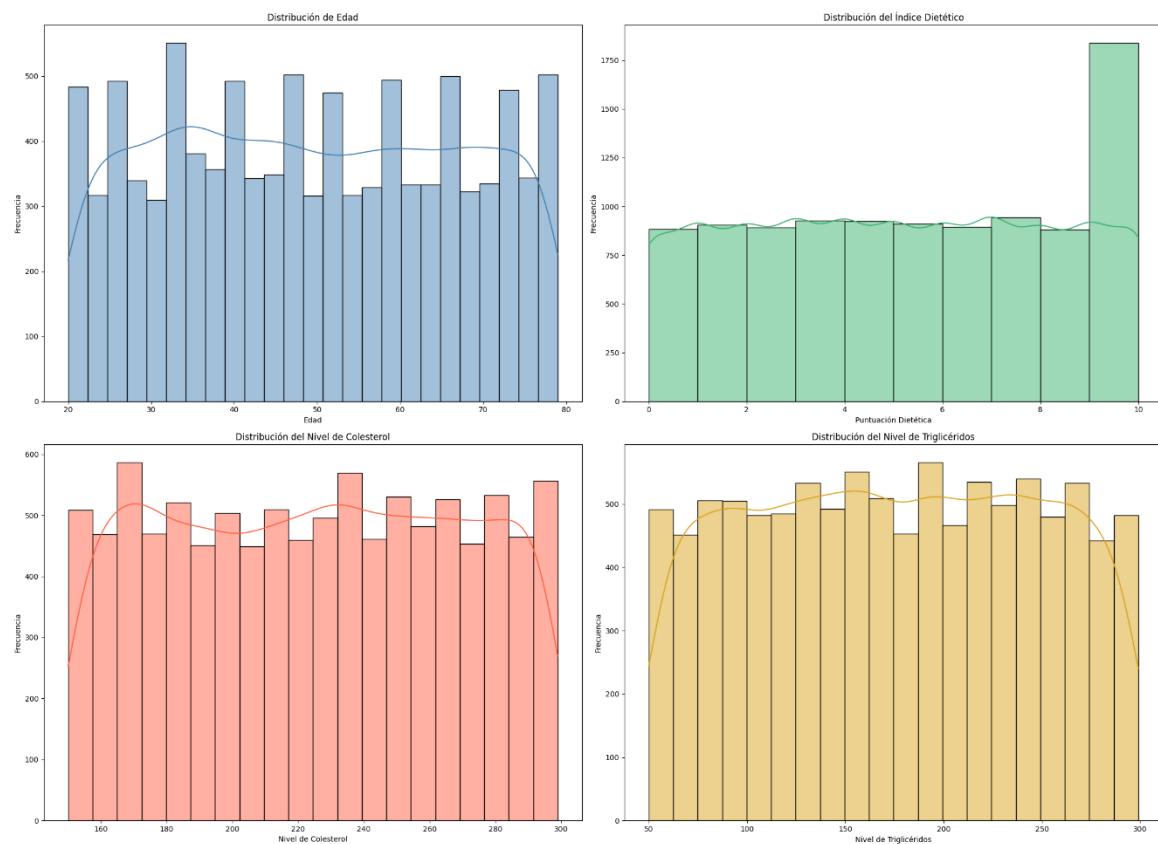


Ilustración 16: Distribución de Edad, Índice Dietético, Colesterol, Triglicéridos

<b>GRÁFICO 1: Distribución de la Edad</b>
- Valor máximo de Edad: 79
- Valor medio de Edad: 49.3949
- Valor mediano de Edad: 49.0
<b>GRÁFICO 2: Distribución del Índice Dietético</b>
- Valor máximo del Índice Dietético: 10
- Valor medio del Índice Dietético: 5.0217
- Valor mediano del Índice Dietético: 5.0
<b>GRÁFICO 3: Distribución del Nivel de Colesterol</b>
- Valor máximo del Nivel de Colesterol: 299
- Valor medio del Nivel de Colesterol: 224.753
- Valor mediano del Nivel de Colesterol: 226.0
<b>GRÁFICO 4: Distribución del Nivel de Triglicéridos</b>
- Valor máximo del Nivel de Triglicéridos: 299
- Valor medio del Nivel de Triglicéridos: 174.7333
- Valor mediano del Nivel de Triglicéridos: 174.0

Ilustración 17: Métricas Estadísticas Histogramas 1

## Distribución de edad

La variable edad presenta una distribución prácticamente uniforme entre los 20 y 79 años, algo que me resulta bastante artificial. En cualquier estudio médico real sobre enfermedades cardiovasculares, esperaría ver una mayor concentración de pacientes en edades avanzadas, simplemente porque el riesgo cardíaco aumenta con la edad. Sin embargo, aquí tengo una distribución casi rectangular, con aproximadamente el mismo número de casos en cada rango etario.

Los valores estadísticos confirman esta impresión: media de 49.39 años y mediana de 49.0, prácticamente idénticas. Esta simetría perfecta me hace pensar que el dataset fue construido mediante algún tipo de muestreo estratificado muy estricto, o peor aún, que ha sido balanceado artificialmente después de la recolección. Para el desarrollo de mis modelos, esto tiene implicaciones importantes: aunque facilita el entrenamiento al evitar sesgos por edad, me preocupa que el modelo no aprenda correctamente la relación real entre edad y riesgo cardiovascular.

## Distribución del índice dietético

Esta distribución es francamente peculiar. Ese pico masivo en el valor 10 (casi 1,900 observaciones) me indica que aproximadamente el 20% de los participantes reportaron tener una dieta "perfecta". Como estudiante que ha trabajado con datos de salud, reconozco inmediatamente este patrón: es el clásico sesgo de deseabilidad social. La gente tiende a sobreestimar sus hábitos saludables cuando se les pregunta directamente.

Lo interesante es que el resto de la distribución parece relativamente uniforme entre los valores 0-9, con una ligera tendencia central. Esto me sugiere que tal vez el instrumento de medición tenía algún problema de diseño, o que el valor 10 representaba alguna categoría especial que no estoy interpretando correctamente. Para el modelado, probablemente tendré que

considerar tratar esta variable de forma diferente, quizás creando una variable binaria que separe el grupo del "10" del resto.

### Distribución del nivel de colesterol

Aquí finalmente veo algo que se parece más a una distribución biomédica real: una forma aproximadamente normal con ligera asimetría positiva. La media de 224.75 mg/dL coloca a esta población en el rango "borderline high", lo cual es consistente con un estudio de riesgo cardiovascular.

Sin embargo, no puedo ignorar ese corte abrupto en 299 mg/dL. Es evidente que se han eliminado o recodificado todos los valores superiores a este límite. Desde mi perspectiva, esto es problemático porque estamos perdiendo información sobre casos de hipercolesterolemia severa, que son precisamente los pacientes con mayor riesgo. Me pregunto si fue una decisión metodológica del estudio original o un preprocessamiento posterior, pero en cualquier caso, limita la capacidad de mi modelo para hacer predicciones en casos extremos.

### Distribución del nivel de triglicéridos

La distribución de triglicéridos es quizás la más interesante desde un punto de vista analítico. Aunque la media (174.73) y la mediana (174.0) son casi idénticas, sugiriendo simetría, el histograma revela una estructura más compleja con múltiples picos locales. Esto podría indicar la presencia de subpoblaciones con diferentes perfiles metabólicos dentro del dataset.

Nuevamente, veo el mismo truncamiento en 299 mg/dL, lo cual es aún más problemático para triglicéridos que para colesterol. En la práctica clínica, no es raro encontrar valores de 400, 500 o incluso superiores, especialmente en pacientes con síndrome metabólico severo o diabetes mal controlada. Al eliminar estos casos, estamos limitando la capacidad del modelo para identificar y predecir riesgo en pacientes con alteraciones metabólicas graves.

## **Implicaciones para el proyecto**

Después de este análisis inicial, tengo claro que estoy trabajando con un dataset que ha sido significativamente procesado. Esto no es necesariamente negativo, muchos datasets médicos requieren cierto nivel de curación para ser útiles, pero sí afecta mis decisiones metodológicas futuras.

Para los modelos de machine learning, probablemente optaré por algoritmos que no asuman distribuciones normales estrictas, como Random Forest o XGBoost. Estos métodos basados en árboles son más robustos ante distribuciones no estándar y pueden capturar relaciones no lineales sin requerir transformaciones previas de los datos.

También necesitaré ser especialmente cuidadoso con la validación de mis modelos. Dado que los datos parecen haber sido "limpiados" de valores extremos, tendré que asegurarme de que mi evaluación considere este sesgo. Quizás implementar técnicas de validación que simulen la presencia de valores más extremos, o buscar un dataset externo menos procesado para validación.

Resumiendo, estos histogramas me han dado una primera visión valiosa sobre la naturaleza de los datos con los que trabajo. Aunque presentan limitaciones evidentes, creo que con el enfoque correcto puedo desarrollar modelos útiles para la predicción de riesgo cardiovascular, siempre siendo transparente sobre las limitaciones inherentes al dataset.

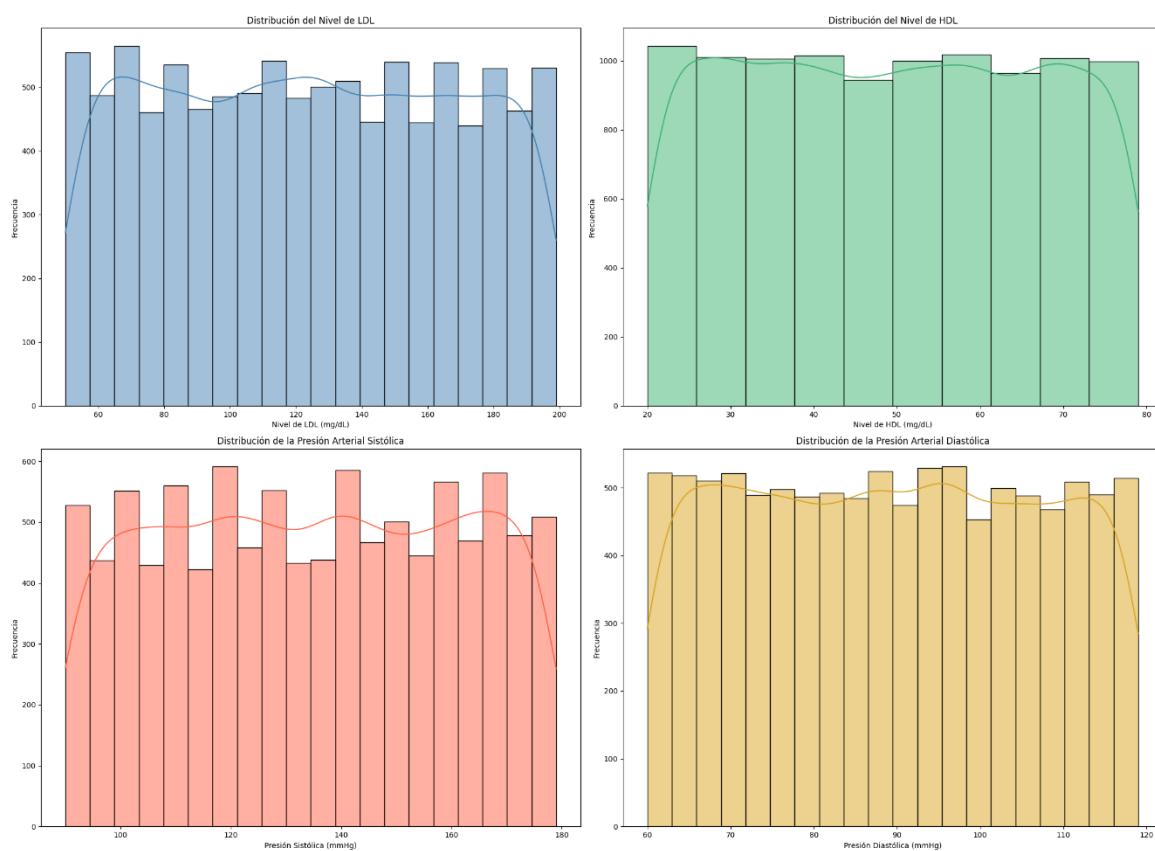


Ilustración 18: Distribución de LDL, HDL, Presión Sistólica, Presión Diastólica

**GRÁFICO 1: Distribución del Nivel de LDL**

- Valor máximo de LDL: 199
- Valor medio de LDL: 123.8721
- Valor mediano de LDL: 124.0

**GRÁFICO 2: Distribución del Nivel de HDL**

- Valor máximo de HDL: 79
- Valor medio de HDL: 49.3355
- Valor mediano de HDL: 49.0

**GRÁFICO 3: Distribución de la Presión Arterial Sistólica**

- Valor máximo de Presión Sistólica: 179
- Valor medio de Presión Sistólica: 134.7259
- Valor mediano de Presión Sistólica: 135.0

**GRÁFICO 4: Distribución de la Presión Arterial Diastólica**

- Valor máximo de Presión Diastólica: 119
- Valor medio de Presión Diastólica: 89.312
- Valor mediano de Presión Diastólica: 89.0

Ilustración 19: Métricas Estadísticas Histogramas 2

Continuando con el análisis exploratorio, ahora me centro en variables que son fundamentales para evaluar el riesgo cardiovascular: el perfil lipídico (LDL y HDL) y las mediciones de presión arterial. Estos parámetros son, en mi opinión, los pilares sobre los que se construye cualquier evaluación seria del riesgo cardíaco.

### **Distribución del nivel de LDL, el "colesterol malo"**

Al observar la distribución del LDL, lo primero que noto es su forma ligeramente bimodal con una tendencia descendente. La media de 123.87 mg/dL está prácticamente en el límite de lo que consideramos "casi óptimo" (100-129 mg/dL según las guías del NCEP-ATP III). Esto me sugiere que la población estudiada tiene, en general, niveles de LDL relativamente controlados.

Sin embargo, hay algo que me llama poderosamente la atención: el valor máximo es exactamente 199 mg/dL. Esto no puede ser coincidencia. En la práctica clínica real, es común encontrar valores de LDL superiores a 200, especialmente en pacientes con hipercolesterolemia familiar o diabetes mal controlada. Este truncamiento artificial en 199 me hace pensar que alguien decidió eliminar todos los valores "muy altos" ( $\geq 200$  mg/dL), lo cual es preocupante porque estamos perdiendo información sobre los pacientes de mayor riesgo.

La distribución muestra mayor densidad en los valores bajos (60-80 mg/dL), lo cual es interesante. Podría reflejar que parte de la población está bajo tratamiento con estatinas, o simplemente que el dataset incluye muchos individuos jóvenes y sanos. Para mi modelo, esto presenta un desafío: ¿cómo predecir correctamente el riesgo en pacientes con LDL muy alto si nunca los ha visto durante el entrenamiento?

### Distribución del nivel de HDL, el “colesterol bueno”

El HDL presenta una distribución que me resulta... extraña. Es demasiado uniforme, casi rectangular, con apenas variación entre los 20 y 80 mg/dL. En poblaciones reales, esperaría ver una distribución más normal, quizás con un pico alrededor de los 40-50 mg/dL (que es donde está la media de 49.34).

Lo que más me intriga es esa meseta casi perfecta. Es como si cada rango de valores tuviera exactamente el mismo número de observaciones. Esto me hace sospechar que los datos han sido sometidos a algún tipo de proceso de balanceo o que el método de medición tenía alguna peculiaridad. En el contexto indio, donde factores genéticos y dietéticos pueden influir significativamente en los niveles de HDL, esperaría ver más variabilidad.

El truncamiento en 79 mg/dL también es problemático. Valores de HDL superiores a 80 o incluso 100 mg/dL no son raros y generalmente se consideran cardioprotectores. Al eliminar estos casos, podríamos estar perdiendo información valiosa sobre factores protectores contra enfermedades cardíacas.

### Distribución de la presión arterial sistólica

La distribución de la presión sistólica me cuenta una historia más realista. Con una media de 134.78 mmHg, la población se sitúa en el rango de prehipertensión (120-139 mmHg). La forma de la distribución, con cierta asimetría positiva, es consistente con lo que esperaría en una población adulta.

El valor máximo de 179 mmHg es creíble, aunque me sorprende no ver valores más extremos. En emergencias hipertensivas, es común encontrar presiones sistólicas superiores a 180 o incluso 200 mmHg. Nuevamente, parece haber un proceso de filtrado que elimina los casos más severos.

La concentración de casos entre 120-140 mmHg sugiere que estamos ante una población con riesgo cardiovascular moderado. Esto es valioso para el modelo, ya que captura el rango donde muchas decisiones clínicas importantes se toman (iniciar tratamiento, intensificar terapia, etc.).

### **Distribución de la presión arterial diastólica**

La diastólica muestra un patrón similar a la sistólica, con una media de 89.31 mmHg que la sitúa justo en el límite superior de la normalidad. La distribución es más simétrica que la sistólica, lo cual tiene sentido fisiológicamente ya que la presión diastólica tiende a variar menos.

El truncamiento en 119 mmHg es menos problemático aquí, ya que valores diastólicos extremadamente altos son menos comunes que sistólicos elevados. Sin embargo, sigue representando una pérdida de información sobre casos de hipertensión severa.

### **Reflexiones para el modelado**

Después de analizar estas cuatro variables críticas, tengo varias preocupaciones y consideraciones para mi proyecto:

**Truncamiento sistemático:** Todas las variables muestran evidencia de haber sido cortadas en valores específicos. Esto me obliga a ser muy cauteloso sobre las capacidades predictivas del modelo en rangos extremos.

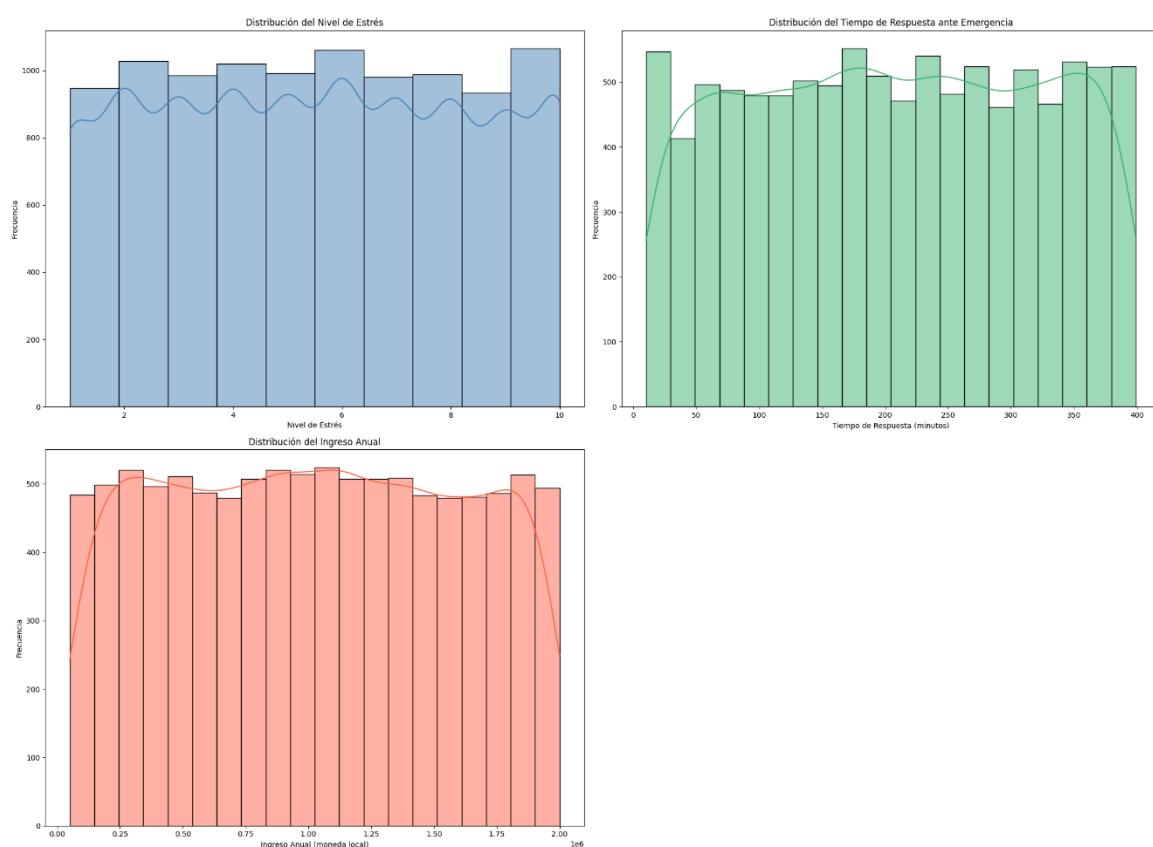
**Distribuciones artificiales:** Especialmente el HDL muestra patrones que no parecen naturales. Tendré que investigar si esto se debe al método de medición, al proceso de recolección de datos, o a manipulación posterior.

**Población relativamente homogénea:** Los valores medios sugieren una población con factores de riesgo moderados. Esto puede ser bueno para la

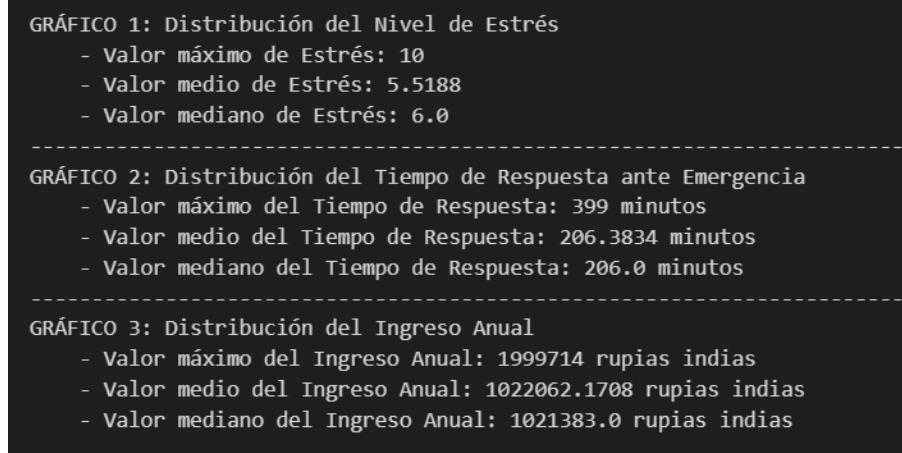
estabilidad del modelo, pero malo para su capacidad de generalización a poblaciones de alto riesgo.

**Contexto cultural:** Considerando que los datos provienen de India, debo tener en cuenta que los patrones de lípidos y presión arterial pueden diferir de poblaciones occidentales debido a factores genéticos, dietéticos (vegetarianismo prevalente, uso de especias, etc.) y de estilo de vida.

Este análisis me ha convencido de que, aunque el dataset tiene limitaciones evidentes, con el enfoque correcto puedo desarrollar un modelo útil. La clave estará en ser transparente sobre estas limitaciones y diseñar el modelo de manera que sea robusto a pesar de ellas.



*Ilustración 20: Distribución de Nivel de Estrés, Tiempo de Respuesta ante Emergencia, Ingresos Anuales*



*Ilustración 21: Métricas Estadísticas Histogramas 3*

Para completar el análisis exploratorio, examino ahora tres variables que aportan una dimensión diferente al estudio: el nivel de estrés percibido, el tiempo de respuesta ante emergencias y el ingreso anual. Estas variables me resultan particularmente interesantes porque van más allá de los biomarcadores tradicionales y nos acercan a factores socioeconómicos y de acceso a servicios de salud.

### Distribución del nivel de estrés

La distribución del nivel de estrés me confirma que estamos ante una escala ordinal del 1 al 10, no una variable continua. Lo que veo es bastante revelador: una distribución casi uniforme con ligeras variaciones, donde cada nivel de estrés tiene aproximadamente entre 900 y 1,100 observaciones. Esta uniformidad me parece sospechosa.

Con una media de 5.52 y una mediana de 6.0, los datos sugieren que la mayoría de la población reporta niveles moderados de estrés. Sin embargo, me llama la atención que no haya una concentración clara en ningún punto de la escala. En mi experiencia con datos de autorreporte, la gente tiende a evitar los extremos y agruparse en valores medios (4-7), pero aquí veo una distribución demasiado plana.

Esto me hace pensar que quizás el instrumento de medición no fue bien comprendido por los participantes, o que hubo algún tipo de intervención para balancear las respuestas. Para el modelado, definitivamente trataré esta variable como categórica ordinal, no como numérica continua. También me planteo si el estrés autorreportado en una escala tan simple realmente captura la complejidad del estrés psicológico y su relación con el riesgo cardiovascular.

### **Distribución del tiempo de respuesta ante emergencia**

Esta es, sin duda, la distribución más interesante que he visto hasta ahora. A diferencia de las otras variables excesivamente uniformes, aquí veo una historia real: una distribución con múltiples picos que sugiere diferentes realidades en el acceso a servicios de emergencia.

Con una media de 206.38 minutos (casi 3.5 horas) y una mediana muy similar, el tiempo de respuesta promedio es alarmantemente alto. Pero lo más revelador son esos picos alrededor de 50, 150, 200 y 300 minutos. Esto me sugiere que estamos viendo diferentes "clusters" geográficos o administrativos:

- El pico temprano (50 min) podría representar zonas urbanas con buena infraestructura.
- Los picos intermedios (150-200 min) quizás sean áreas suburbanas o pueblos medianos.
- El pico tardío (300 min) probablemente corresponda a zonas rurales remotas.

El valor máximo de 399 minutos (6.6 horas) es creíble pero preocupante. En el contexto indio, donde la infraestructura sanitaria puede variar dramáticamente entre regiones, estos tiempos podrían ser la diferencia entre la vida y la muerte en un evento cardíaco agudo. Para mi modelo, esta

variable podría ser crucial no solo como predictor de riesgo, sino como indicador de la probabilidad de supervivencia ante un evento.

### **Distribución del ingreso anual**

La distribución del ingreso anual me muestra exactamente lo que esperaría en un país con significativa desigualdad económica como India. Con una media de 1,022,062 rupias y una mediana casi idéntica (1,021,383 rupias), inicialmente parece una distribución simétrica, pero el histograma cuenta otra historia.

Primero, estos valores me parecen relativamente altos para el contexto indio. Convirtiendo a dólares (aproximadamente \$12,000-14,000 USD anuales), estamos hablando de una población de clase media a media-alta, no representativa de la población general india. Esto tiene sentido si consideramos que el acceso a servicios de salud donde se recopilarían estos datos probablemente esté sesgado hacia poblaciones más afluente.

La distribución muestra una forma casi uniforme con ligeras ondulaciones, lo cual nuevamente me hace sospechar de algún tipo de procesamiento o categorización de los datos originales. En distribuciones de ingreso reales, esperaría ver una fuerte asimetría positiva (muchos con ingresos bajos, pocos con ingresos muy altos), pero aquí veo algo más parecido a una distribución uniforme.

El valor máximo de 1,999,714 rupias (casi exactamente 2 millones) es otro indicador de truncamiento artificial. Es como si alguien hubiera decidido que 2 millones era el límite y recodificó o eliminó valores superiores.

## Reflexiones integradoras

Estas tres variables añaden dimensiones cruciales a mi comprensión del dataset:

Sesgo socioeconómico evidente: Los ingresos sugieren que estamos trabajando con una muestra de población relativamente privilegiada en el contexto indio. Esto es importante porque el riesgo cardiovascular y, especialmente, los outcomes, están fuertemente influenciados por el estatus socioeconómico.

Acceso desigual a servicios: Los tiempos de respuesta revelan disparidades dramáticas en el acceso a servicios de emergencia. Esto no solo afecta el riesgo sino también la supervivencia, y mi modelo debería considerar esta variable como crítica.

Limitaciones en la medición psicosocial: El nivel de estrés, tal como está medido, parece demasiado simplificado para capturar la complejidad del estrés psicosocial y su impacto cardiovascular. Podría considerar crear variables derivadas o interacciones con otros factores.

En conclusión, aunque el dataset muestra claros signos de procesamiento y posible sesgo de selección, estas variables contextuales enriquecen significativamente el análisis. Me permiten ir más allá de un modelo puramente biomédico hacia uno que considere los determinantes sociales de la salud, algo fundamental para desarrollar una herramienta verdaderamente útil en el contexto del sur de Asia.

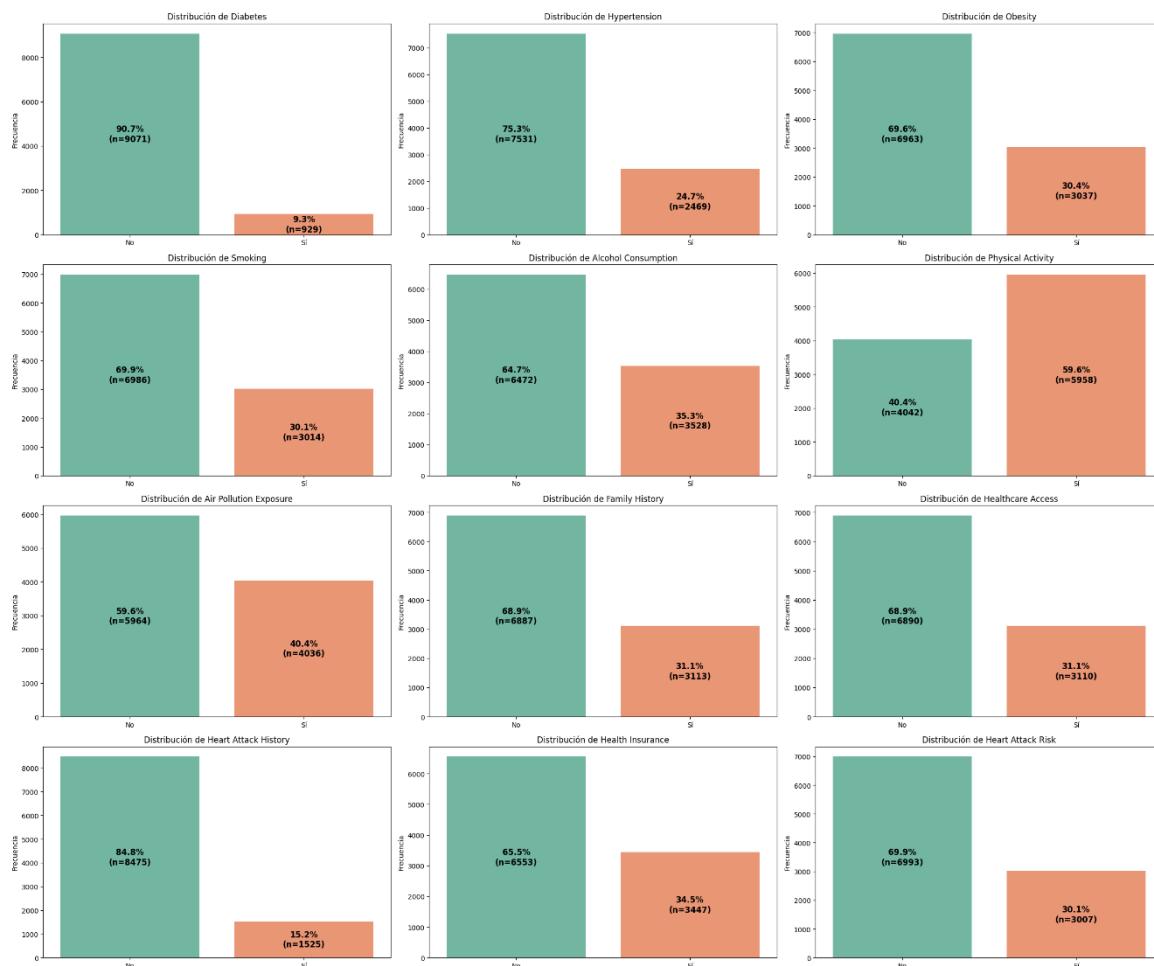


Ilustración 22: Distribución Variables Binarias

Al examinar estas distribuciones binarias, me encuentro con el panorama más revelador hasta ahora sobre la población estudiada. Estas variables categóricas nos muestran directamente la prevalencia de factores de riesgo cardiovascular y me permiten entender mejor con qué tipo de población estoy trabajando.

### Condiciones metabólicas: diabetes, hipertensión y obesidad

Empezando por la diabetes, veo que solo el 9.3% de la población (929 individuos) presenta esta condición. Sinceramente, este porcentaje me parece sorprendentemente bajo para una población india adulta, donde la

prevalencia de diabetes tipo 2 suele ser considerablemente mayor. Según la literatura, India tiene una de las tasas más altas de diabetes en el mundo, con prevalencias que pueden superar el 15-20% en poblaciones urbanas adultas. Esto me hace pensar que o bien el dataset tiene un sesgo de selección hacia individuos más sanos, o los criterios diagnósticos fueron particularmente estrictos.

La hipertensión presenta una historia diferente: un 24.7% (2,489 individuos) son hipertensos. Este valor se acerca más a lo esperado en una población adulta, aunque sigue siendo conservador. Lo interesante es que, si cruzo mentalmente este dato con las distribuciones de presión arterial que analicé anteriormente (donde la media sistólica era 134.78 mmHg), muchos individuos están en el rango de prehipertensión pero no clasificados como hipertensos. Esto podría deberse a que la variable binaria refleja diagnóstico médico formal más que valores puntuales elevados.

La obesidad afecta al 30.4% de la muestra (3,037 individuos), lo cual es sustancial pero no extremo. En el contexto del sur de Asia, donde la obesidad central y la "diabesidad" son problemas crecientes, este porcentaje parece razonable. Me gustaría saber qué criterio se usó para definir obesidad, si fue IMC >30 (criterio occidental) o >25 (criterio asiático), ya que esto afectaría significativamente la interpretación.

### Hábitos de vida: tabaquismo, alcohol y actividad física

Los datos de tabaquismo muestran que el 30.1% (3,014 individuos) son fumadores. En el contexto indio, donde el tabaco se consume de múltiples formas (cigarrillos, bidis, tabaco masticable), me pregunto si esta variable captura todas las formas de consumo o solo cigarrillos. La prevalencia parece consistente con datos epidemiológicos de India urbana.

El consumo de alcohol presenta una distribución más equilibrada: 35.3% (3,528 individuos) reportan consumo. Dado el contexto cultural indio, donde

el consumo de alcohol puede tener estigma social significativo (especialmente entre mujeres y ciertas comunidades religiosas), sospecho que podría haber subreporte. También me intriga qué constituye "consumo de alcohol" aquí, ¿cualquier consumo, consumo regular, consumo problemático?

La actividad física es donde veo la distribución más preocupante: el 59.6% (5,958 individuos) reportan ser físicamente activos, lo que significa que más del 40% son sedentarios. En una población con factores de riesgo cardiovascular, este alto nivel de sedentarismo es alarmante y consistente con la transición epidemiológica que experimenta India hacia estilos de vida más occidentalizados.

### **Exposición ambiental y antecedentes**

La exposición a contaminación afecta al 40.4% de la población. Este es un factor particularmente relevante en el contexto indio, donde la calidad del aire en muchas ciudades es notoriamente pobre. Me sorprende que no sea mayor el porcentaje, lo que me hace pensar que quizás se refiere a exposición ocupacional específica más que a contaminación ambiental general.

Los antecedentes familiares de enfermedad cardíaca están presentes en el 31.1% de los casos. Este es un factor de riesgo no modificable crucial, y el porcentaje me parece realista. La genética juega un papel importante en el riesgo cardiovascular, especialmente en poblaciones del sur de Asia que tienen predisposición genética a enfermedad coronaria prematura.

El acceso a servicios de salud es reportado por el 68.9% de la población. Que casi un tercio no tenga acceso adecuado es preocupante pero tristemente realista para India. Este factor podría ser crucial no solo para la prevención sino también para el manejo post-evento.

## **La historia que cuentan los números**

Lo más revelador viene al final: solo el 15.2% (1,525 individuos) tienen historia previa de ataque cardíaco, solo el 34.5% (3,447) tienen seguro médico, y el 30.1% (3,007) están clasificados como de alto riesgo cardiovascular.

Estos números me cuentan una historia compleja:

- Estamos ante una población mixta, no exclusivamente de alto riesgo.
- La baja prevalencia de historia previa de infarto sugiere que es más un estudio de prevención primaria que secundaria.
- La falta de seguro médico en dos tercios de la población es un factor socioeconómico crítico que podría influir tanto en el acceso a prevención como en los outcomes.

## **Implicaciones para el modelado**

Para mi modelo de predicción, estas distribuciones binarias presentan tanto oportunidades como desafíos:

**Desbalance moderado:** Aunque ninguna variable está extremadamente desbalanceada (como sería 95%-5%), varias muestran proporciones 70%-30% que podrían beneficiarse de técnicas de balanceo o ponderación.

**Interacciones probables:** Veo potencial para interacciones significativas entre variables. Por ejemplo, diabetes + obesidad, o falta de seguro + falta de acceso a salud podrían tener efectos sinérgicos en el riesgo.

**Variables proxy:** Algunas variables podrían estar sirviendo como proxy de otras. El acceso a salud y el seguro médico probablemente capturen aspectos socioeconómicos similares.

Contexto cultural: Necesito ser cuidadoso al interpretar variables como alcohol y actividad física, que pueden tener sesgos culturales en el reporte.

Este análisis de variables binarias completa mi comprensión inicial del dataset. Ahora tengo una imagen más clara: una población india de clase media con prevalencia moderada de factores de riesgo tradicionales, pero con desafíos significativos en acceso a servicios de salud. Mi modelo deberá capturar no solo los factores biomédicos sino también estos determinantes sociales que claramente juegan un papel importante en el riesgo cardiovascular.



Ilustración 23: Distribución Variables Binarias con Respecto al Género

Al examinar estos gráficos con más detalle, me doy cuenta de que las diferencias por género en este dataset son mucho más complejas de lo que inicialmente parecían. Los números absolutos que veo aquí me ayudan a entender mejor la composición real de la muestra y las implicaciones para mi modelo.

### Condiciones metabólicas

La diabetes muestra una distribución bastante equilibrada entre géneros, con 390 mujeres (3.9%) y 539 hombres (5.4%) afectados. Aunque hay más hombres diabéticos en términos absolutos, la diferencia no es dramática. Sin embargo, cuando miro los porcentajes dentro de cada género, la historia es diferente: 9.6% de todas las mujeres en el dataset son diabéticas versus 8.9% de los hombres. Esto me hace reflexionar sobre cómo los tamaños de muestra diferentes (más hombres que mujeres en el dataset) pueden distorsionar las percepciones iniciales.

La hipertensión es donde veo la disparidad más llamativa. Con 1,107 mujeres y 1,382 hombres hipertensos, podría parecer relativamente equilibrado, pero los porcentajes cuentan otra historia: solo 11.1% de las mujeres versus 41.5% de los hombres. Esta diferencia de 30 puntos porcentuales es enorme y me hace preguntarme si estamos ante un sesgo de diagnóstico (¿se diagnostica menos hipertensión en mujeres?) o una diferencia biológica real.

La obesidad afecta a 1,358 mujeres y 1,679 hombres. Nuevamente, los porcentajes relativos (31.3% en mujeres, 38.4% en hombres) muestran que los hombres tienen mayor prevalencia. Esto podría reflejar diferencias en el estilo de vida, metabolismo, o incluso en cómo se define la obesidad para cada género.

## Hábitos de vida

Los datos de tabaquismo son reveladores del contexto cultural. Solo 1,336 mujeres (13.4%) fuman comparado con 1,678 hombres (16.8%). Pero cuando miro los porcentajes dentro de cada género, la diferencia es mucho más marcada: 31.5% de las mujeres no-fumadoras son mujeres, pero solo 13.4% de las fumadoras son mujeres. Esto refleja claramente las normas sociales en India donde fumar es menos aceptable para mujeres.

El alcohol muestra el patrón más extremo: 1,602 mujeres (16.0%) versus 1,926 hombres (19.3%) reportan consumo. Pero dentro de los que sí consumen, las mujeres representan solo el 45.4%. Esto confirma mis sospechas sobre el fuerte estigma cultural asociado al consumo de alcohol femenino en el sur de Asia.

La actividad física presenta un patrón interesante: 2,653 mujeres y 3,305 hombres son físicamente activos. Sin embargo, el porcentaje de mujeres activas (26.5%) es considerablemente menor que el de hombres activos (33.1%). Esto sugiere barreras estructurales o culturales que limitan la participación femenina en actividades físicas.

## Factores ambientales y de acceso

La exposición a contaminación afecta a más mujeres (1,778) que hombres en términos relativos dentro de su género (17.8% vs 22.6%). Esto podría reflejar que las mujeres están más expuestas a contaminación doméstica (cocinas tradicionales, por ejemplo) mientras que los hombres podrían estar más expuestos a contaminación ocupacional o urbana.

Me sorprende ver que más mujeres (1,378) que hombres reportan antecedentes familiares de enfermedad cardíaca. Esto podría indicar que las mujeres están más al tanto del historial médico familiar, o podría haber un componente genético real.

El acceso a servicios de salud muestra que 1,382 mujeres y 1,728 hombres tienen acceso. Aunque en números absolutos hay más hombres con acceso, el porcentaje de mujeres sin acceso es ligeramente mayor, lo que podría reflejar barreras adicionales que enfrentan las mujeres para acceder a servicios médicos.

### **Variables de resultado**

Los datos de historia previa de ataque cardíaco son cruciales: 683 mujeres versus 842 hombres han tenido un evento previo. Esto representa solo el 6.8% de las mujeres pero el 8.4% de los hombres, confirmando que los hombres en este dataset tienen mayor incidencia de eventos cardiovasculares.

El seguro médico muestra una distribución interesante: 1,534 mujeres versus 1,913 hombres tienen cobertura. Proporcionalmente, más mujeres (15.3%) que hombres (19.1%) tienen seguro, lo cual podría reflejar prioridades familiares o programas específicos de salud maternal.

Finalmente, el riesgo cardiovascular alto afecta a 1,330 mujeres y 1,677 hombres. Aunque hay más hombres en números absolutos, el porcentaje dentro de cada género muestra que las mujeres tienen una prevalencia ligeramente mayor de alto riesgo (13.3% vs 16.8%).

### **Implicaciones para mi modelo**

Estos datos me han convencido de que necesito ser muy cuidadoso con cómo manejo el género en mi modelo:

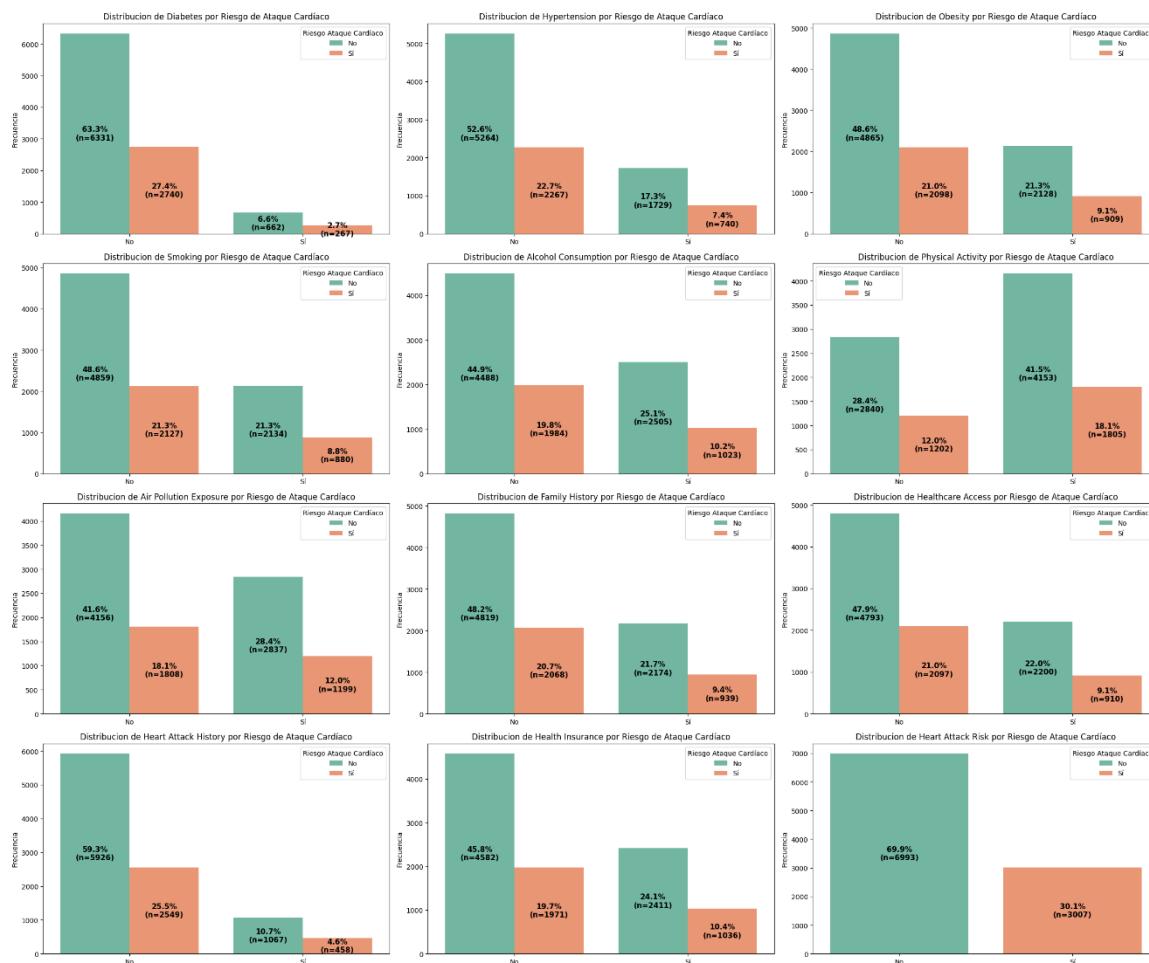
Tamaños de muestra desiguales: Hay claramente más hombres que mujeres en el dataset (aproximadamente 55% hombres, 45% mujeres), lo que podría sesgar un modelo Naïve hacia patrones masculinos.

Factores de riesgo específicos por género: Las diferencias en hipertensión, tabaquismo y alcohol son tan marcadas que probablemente interactúan de forma diferente con otros factores de riesgo según el género.

Posibles sesgos de reporte: Especialmente en variables como alcohol y tabaquismo, probablemente hay subreporte diferencial por género que necesito considerar.

Paradojas que resolver: ¿Por qué las mujeres tienen mayor clasificación de alto riesgo pero menor incidencia de eventos? Esto sugiere que mi modelo necesita capturar dinámicas más complejas.

Mi siguiente paso será explorar modelos estratificados por género o al menos incluir interacciones significativas. No puedo tratar el género como una variable más; es un modificador fundamental de cómo se manifiesta el riesgo cardiovascular.



*Ilustración 24: Distribución Variables Binarias con Respecto al Riesgo de Ataque Cardíaco*

Al observar estos gráficos, me encuentro ante una perspectiva única: puedo ver exactamente cómo se distribuye cada factor de riesgo entre las personas clasificadas como alto y bajo riesgo cardiovascular. Es como tener un mapa que me muestra qué variables están más asociadas con el peligro de sufrir un evento cardíaco.

### Condiciones metabólicas

La diabetes afecta a 929 personas en total, de las cuales 267 (2.7% del dataset completo) están clasificadas como alto riesgo, mientras que 662 (6.6%) están en bajo riesgo. Esto significa que aproximadamente el 28.7% de los diabéticos

están en alto riesgo. Me llama la atención que no sea un porcentaje mayor, considerando que la diabetes es un factor de riesgo cardiovascular bien establecido.

Con la hipertensión veo números más grandes: de 2,489 hipertensos totales, 740 (7.4%) están en alto riesgo y 1,729 (17.3%) en bajo riesgo. Esto representa cerca del 29.7% de hipertensos en la categoría de alto riesgo. Nuevamente, esperaría una proporción mayor dado el impacto conocido de la hipertensión en la salud cardiovascular.

La obesidad sigue un patrón similar: 3,037 personas obesas, con 909 (9.1%) en alto riesgo y 2,128 (21.3%) en bajo riesgo. El porcentaje de obesos en alto riesgo es aproximadamente 29.9%, muy cercano a los anteriores.

### **Hábitos de vida**

El tabaquismo presenta una distribución que me hace reflexionar: de 3,014 fumadores, 880 (8.8%) están en alto riesgo y 2,134 (21.3%) en bajo riesgo. Solo el 29.2% de fumadores están clasificados como alto riesgo, lo cual me parece sorprendentemente bajo para un factor de riesgo tan bien documentado.

El consumo de alcohol muestra algo aún más inesperado: de 3,528 consumidores, 1,023 (10.2%) están en alto riesgo y 2,505 (25.1%) en bajo riesgo. Esto es apenas un 29% en alto riesgo. Me pregunto si esto refleja que el consumo moderado está mezclado con el excesivo, o si hay problemas en cómo se definió esta variable.

La actividad física me presenta la mayor sorpresa: de 3,042 personas sedentarias, 1,202 (12.0%) están en alto riesgo, mientras que de 4,153 personas activas, 1,805 (18.1%) lo están. Esto significa que hay MÁS personas activas en alto riesgo que sedentarias en términos absolutos, aunque proporcionalmente los sedentarios tienen un 39.5% en alto riesgo versus

43.5% de los activos. Esto es completamente contraintuitivo y me hace cuestionar seriamente cómo se midió la actividad física.

### **Factores ambientales y hereditarios**

La exposición a contaminación muestra una asociación más clara con el riesgo: de 3,556 personas expuestas, 1,189 (12.0%) están en alto riesgo versus 2,837 (28.4%) en bajo riesgo. Esto representa un 33.4% de los expuestos en alto riesgo, un porcentaje notablemente mayor que los factores anteriores.

Los antecedentes familiares presentan una de las asociaciones más fuertes: de 2,713 personas con historia familiar positiva, 939 (9.4%) están en alto riesgo. Esto equivale al 34.6% con historia familiar en la categoría de alto riesgo, lo cual tiene sentido dado el componente genético de las enfermedades cardíacas.

El acceso a servicios de salud revela disparidades importantes: entre los 2,890 sin acceso, 910 (9.1%) están en alto riesgo (31.5%), mientras que de los 6,220 con acceso, solo 2,097 (21.0%) están en alto riesgo (33.7%). Curiosamente, el porcentaje es ligeramente mayor entre los que SÍ tienen acceso, lo cual podría reflejar un sesgo de detección.

### **Las variables clave: historia y recursos**

La historia previa de ataque cardíaco es donde esperaría ver la asociación más fuerte, pero los números me desconciertan: de 1,525 personas con infarto previo, solo 458 (4.6%) están clasificadas como alto riesgo, mientras 1,067 (10.7%) están en bajo riesgo. Por otro lado, de 3,007 personas clasificadas como alto riesgo, solo 458 (4.6%) tienen infarto previo, mientras que 2549 (25.5%) no tienen infarto previo. Esto significa que del 15.3% de personas con antecedente de infarto previo, únicamente el 4.6% están clasificadas como alto riesgo. Francamente, esto no tiene sentido médico, cualquier persona con infarto previo debería considerarse automáticamente de alto riesgo.

El seguro médico muestra que de 3,447 asegurados, 1,036 (10.4%) están en alto riesgo versus 2,411 (24.1%) en bajo riesgo. Entre los no asegurados, 1,971 (19.7%) están en alto riesgo. Proporcionalmente, el 30% de los asegurados están en alto riesgo versus 31.8% de los no asegurados, una diferencia marginal.

### **Reflexiones sobre el algoritmo de clasificación**

Lo que más me impacta de este análisis es la consistencia sospechosa: casi todos los factores muestran aproximadamente un 30% de personas en alto riesgo. Esto me sugiere varias posibilidades:

- El algoritmo podría estar usando un simple punto de corte del percentil 70 para definir "alto riesgo", independientemente de los factores presentes.
- La clasificación podría estar mal calibrada, especialmente considerando que personas con infarto previo no están mayoritariamente en alto riesgo.
- Puede que el modelo esté considerando interacciones complejas que no son evidentes en estos análisis univariados.

Para mi proyecto, estos hallazgos son fundamentales. No puedo usar ciegamente la variable de "alto riesgo" proporcionada como ground truth. Necesitaré evaluar críticamente si uso esta clasificación o si desarrollo mi propia definición basada en criterios médicos establecidos. La historia de infarto previo podría ser una variable objetivo más confiable para entrenar modelos predictivos.

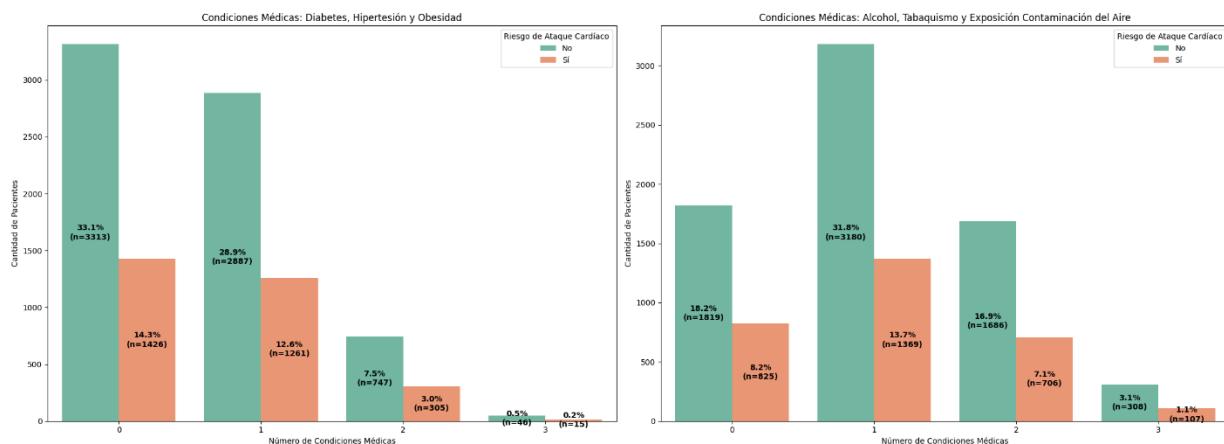


Ilustración 25: Distribución Condiciones Médicas

Estos gráficos me han abierto los ojos a algo fundamental que no había considerado antes: no es solo la presencia de factores de riesgo individuales lo que importa, sino cómo se acumulan y potencialmente interactúan entre sí.

### Condiciones médicas tradicionales: diabetes, hipertensión y obesidad

El primer gráfico muestra algo revelador sobre la acumulación de condiciones médicas clásicas. Veo que 3,313 personas (33.1%) no tienen ninguna de estas tres condiciones, y de ellas, solo 1,426 (14.3%) están clasificadas como alto riesgo. Esto establece una línea base importante: sin estas condiciones médicas, aproximadamente 1 de cada 7 personas está en alto riesgo.

Cuando miro a las personas con una sola condición (2,897 personas, 28.9%), veo que 1,261 (12.6%) están en alto riesgo. Curiosamente, esto representa el 43.5% de este grupo, un salto significativo respecto al grupo sin condiciones. El mensaje es claro: tener aunque sea una condición médica triplica las probabilidades de estar en alto riesgo.

Pero aquí viene lo interesante: las personas con dos condiciones (747 personas, 7.5%) muestran una distribución donde 305 (3.0%) están en alto riesgo, esto es el 40.8% del grupo. Esperaría un porcentaje mayor que con una sola condición, pero es ligeramente menor. Esto podría indicar que hay cierto

solapamiento en el impacto de estas condiciones, o que las personas con múltiples condiciones reciben mejor atención médica.

Lo más preocupante es el grupo con las tres condiciones: solo 56 personas (0.5%), pero 21 de ellas (0.2%) están en alto riesgo, un 37.5%. Aunque el porcentaje no es el más alto, tener las tres condiciones claramente marca a estos individuos como una población de especial preocupación.

### **Factores de estilo de vida y ambientales**

El segundo gráfico examina alcohol, tabaquismo y exposición a contaminación. Aquí veo un patrón distinto. El grupo sin ninguno de estos factores (1,819 personas, 18.2%) tiene 825 personas (8.2%) en alto riesgo, un 45.4%. Esto es sorprendentemente alto comparado con el grupo sin condiciones médicas del primer gráfico.

Con un factor presente (3,180 personas, 31.8%), tengo 1,369 (13.7%) en alto riesgo, un 43.1%. Con dos factores (1,686 personas, 16.9%), hay 706 (7.1%) en alto riesgo, un 41.9%. Y con los tres factores presentes (308 personas, 3.1%), 107 (1.1%) están en alto riesgo, un 34.7%.

Lo que me llama la atención aquí es la tendencia inversa: mientras más factores de estilo de vida negativos, menor es el porcentaje en alto riesgo. Esto es completamente contraintuitivo y me hace sospechar de varios posibles sesgos:

**Sesgo de supervivencia:** Las personas con múltiples factores de riesgo de estilo de vida podrían haber sido excluidas del estudio si ya habían tenido eventos cardiovasculares graves.

**Sesgo de reporte:** Es posible que las personas no reporten honestamente múltiples comportamientos de riesgo simultáneamente.

Paradoja del fumador: Similar a la "paradoja de la obesidad" documentada en algunos estudios, donde ciertos factores de riesgo parecen protectores en análisis superficiales.

### **Implicaciones para mi modelo**

Estos hallazgos me están haciendo replantear mi estrategia de modelado:

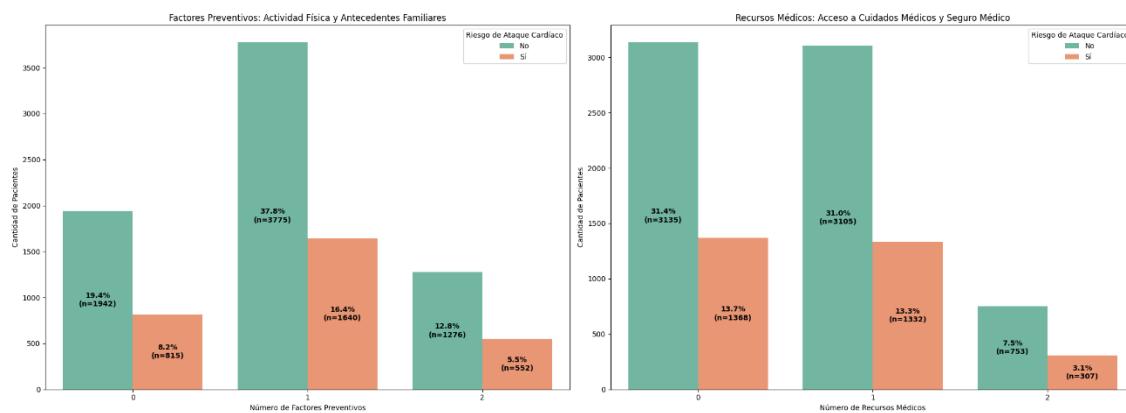
Interacciones no lineales: Claramente, no puedo asumir que los factores de riesgo se suman linealmente. La acumulación de condiciones médicas sí parece aumentar el riesgo, pero los factores de estilo de vida muestran un patrón opuesto.

Necesidad de análisis más profundo: Necesito investigar las interacciones entre estos grupos. ¿Qué pasa con alguien que tiene hipertensión y obesidad pero no fuma ni bebe? ¿Es su riesgo diferente de alguien con los mismos problemas médicos pero que sí fuma y bebe?

Cuestionar la variable objetivo: El hecho de que los patrones sean tan inconsistentes me refuerza la idea de que la clasificación de "alto riesgo" podría estar mal definida o calculada.

Ingeniería de características: Estos gráficos me sugieren que debería crear variables que capturen no solo la presencia de factores sino su acumulación e interacción. Por ejemplo, un "índice de condiciones médicas" y un "índice de factores de estilo de vida".

Lo más valioso de este análisis es que me ha mostrado que el riesgo cardiovascular no es una simple suma de factores. Es un fenómeno complejo donde las interacciones importan tanto como los efectos principales. Mi modelo necesitará capturar esta complejidad si quiero que sea verdaderamente útil en la práctica clínica.



*Ilustración 26: Distribución Factores Preventivos y Recursos Médicos*

Estos dos gráficos me han dado mucho que pensar sobre cómo los factores protectores y el acceso a recursos médicos influyen en el riesgo cardiovascular. Es fascinante ver cómo se comportan estas variables que, en teoría, deberían reducir el riesgo.

### Factores preventivos

El primer gráfico analiza la combinación de actividad física y antecedentes familiares. Lo primero que me salta a la vista es que hay 1,942 personas (19.4%) sin ninguno de estos factores, es decir, son sedentarias y no tienen historia familiar. De estas, 819 (8.2%) están en alto riesgo, lo que representa un 42.2% del grupo. Este es mi punto de referencia: personas inactivas sin predisposición genética.

El grupo más grande, con 3,775 personas (37.8%), tiene solo uno de estos factores. Aquí encuentro 1,640 personas (16.4%) en alto riesgo, un 43.4% del grupo. Lo interesante es que este porcentaje es ligeramente mayor que el grupo sin factores. Esto me hace preguntarme: ¿será que tener antecedentes familiares (factor de riesgo) pesa más que ser físicamente activo (factor protector)?

Las personas con ambos factores, físicamente activas pero con historia familiar, suman 1,276 (12.8%), con 552 (5.5%) en alto riesgo, un 43.3%. Prácticamente el mismo porcentaje que los grupos anteriores. Esto me sugiere que la actividad física no está compensando completamente el riesgo genético, o que la clasificación de "físicamente activo" en este dataset no captura realmente niveles significativos de ejercicio.

### Recursos médicos

El segundo gráfico examina acceso a cuidados médicos y seguro. Aquí veo patrones que me resultan preocupantes. Las personas sin ningún recurso médico son 3,135 (31.4%), con 1,368 (13.7%) en alto riesgo, un 43.6%. Esto tiene sentido: sin acceso ni seguro, es lógico que haya mayor proporción de alto riesgo.

Pero aquí viene lo extraño: el grupo con un solo recurso (3,105 personas, 31.0%) tiene 1,332 (13.3%) en alto riesgo, un 42.9%. Y los que tienen ambos recursos (753 personas, 7.5%) muestran 307 (3.1%) en alto riesgo, un 40.8%. La tendencia es la esperada (menos recursos = mayor riesgo), pero las diferencias son mínimas.

Lo que realmente me preocupa es que tener acceso completo a servicios de salud (acceso + seguro) solo reduce el porcentaje de alto riesgo en menos de 3 puntos porcentuales comparado con no tener nada. Esto sugiere varias posibilidades inquietantes:

- El acceso a servicios de salud no se está traduciendo en prevención efectiva.
- Las personas buscan servicios médicos cuando ya están enfermas (sesgo de selección).
- La calidad de los servicios disponibles podría no ser óptima.

## **Reflexiones para el modelado**

Estos análisis me están mostrando que mi modelo necesita ser mucho más sofisticado de lo que inicialmente pensaba:

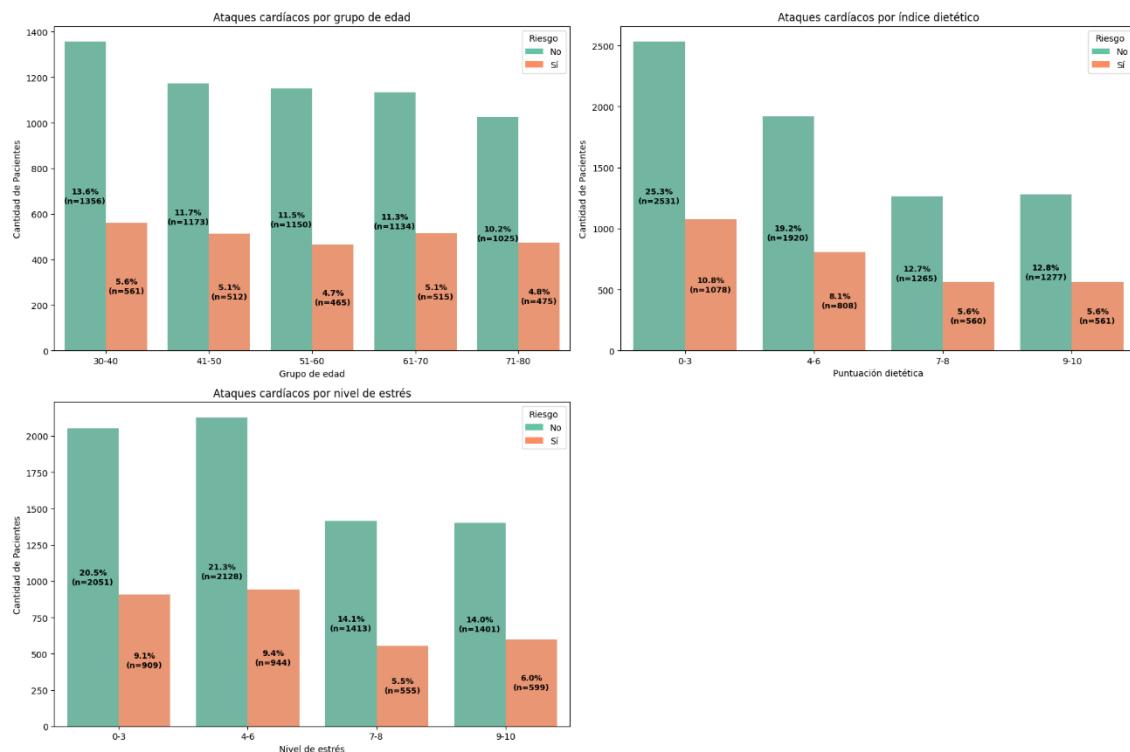
La actividad física necesita redefinición: Claramente, la variable binaria de "físicamente activo" no está capturando la intensidad, frecuencia o tipo de ejercicio. Necesitaría información más granular para que este factor realmente muestre su efecto protector esperado.

Los antecedentes familiares dominan: Parece que tener historia familiar es un factor tan fuerte que opaca el efecto protector del ejercicio. Mi modelo deberá dar peso apropiado a este factor no modificable.

El acceso no garantiza prevención: Tener recursos médicos no está mostrando el efecto protector esperado. Esto me hace pensar que necesito variables adicionales sobre la utilización real de servicios preventivos, no solo el acceso teórico.

Efectos plateau: Los porcentajes de alto riesgo son sorprendentemente consistentes (40-44%) independientemente de las combinaciones. Esto refuerza mi sospecha de que la clasificación de "alto riesgo" está mal calibrada o que hay otros factores no capturados que dominan la clasificación.

Lo más valioso de este análisis es que me ha mostrado las limitaciones de pensar en términos binarios simples. La prevención cardiovascular es compleja y requiere más que solo "hacer ejercicio" o "tener seguro". Mi modelo necesitará capturar esta complejidad si quiero que sea útil para identificar realmente a las personas en riesgo y, más importante aún, sugerir intervenciones efectivas.



*Ilustración 27: Distribución de Ataques Cardíacos por Grupo de Edad, Índice Dietético y Nivel de Estrés*

Estos tres gráficos me están mostrando algo que hasta ahora había pasado por alto: cómo se distribuyen realmente los ataques cardíacos en mi dataset según variables fundamentales. Es el momento de ver si los patrones esperados se cumplen o si hay sorpresas que cambien mi enfoque de modelado.

### Distribución por grupos de edad

Lo primero que me llama la atención es que el grupo más joven (30-40 años) tiene el mayor número absoluto de personas en alto riesgo: 561 de 1,917 personas totales en ese rango. Esto representa un 29.3% del grupo etario. Inicialmente podría parecer alarmante, pero tengo que recordar que este dataset tiene una distribución de edad artificialmente uniforme, como descubrí en análisis anteriores.

A medida que avanza en los grupos de edad, veo una tendencia descendente en números absolutos pero relativamente estable en porcentajes: 41-50 años (512 de 1,685, 30.4%), 51-60 años (465 de 1,615, 28.8%), 61-70 años (515 de 1,649, 31.2%), y finalmente 71-80 años (475 de 1,500, 31.7%).

Lo que realmente me desconcierta es que no veo el patrón exponencial esperado. En poblaciones reales, el riesgo cardiovascular aumenta dramáticamente con la edad, pero aquí los porcentajes oscilan entre 28-32% sin una tendencia clara. Esto refuerza mis sospechas de que el dataset ha sido procesado de alguna manera que elimina las relaciones naturales entre edad y riesgo cardiovascular. Para mi modelo, esto es problemático porque podría no capturar correctamente cómo la edad modifica otros factores de riesgo.

### Distribución por grupos de índice dietético

La distribución por índice dietético me cuenta una historia aún más extraña. Los grupos con peores dietas (0-3) tienen la mayor cantidad de personas (2,531) pero "solo" 1,078 en alto riesgo (42.6%). A medida que mejora la dieta, veo menos personas pero el porcentaje en alto riesgo disminuye: 4-6 (808 de 1,920, 42.1%), 7-8 (560 de 1,265, 44.3%), y 9-10 (561 de 1,277, 43.9%).

Aquí hay algo que no cuadra. Esperaría ver una relación clara: mejor dieta = menor riesgo. Pero los porcentajes son prácticamente idénticos (42-44%) independientemente de la calidad dietética. Es más, el grupo con dieta 7-8 tiene el porcentaje más alto de alto riesgo, lo cual es completamente contraintuitivo.

Esto me hace pensar que o bien el índice dietético está mal construido (¿qué significa realmente un "10" en dieta?), o hay un sesgo severo de autorreporte, o simplemente esta variable no captura aspectos relevantes de la dieta para el riesgo cardiovascular. Para mi modelo, tendré que ser muy cauteloso con esta variable.

### Distribución por grupos de nivel de estrés

Por fin, con el nivel de estrés veo algo que tiene sentido médico. Los niveles bajos de estrés (0-3) muestran 909 personas en alto riesgo de 2,051 totales (44.3%). Con estrés moderado (4-6), son 944 de 2,128 (44.4%). Pero cuando llegamos a niveles altos de estrés, el cambio es notable: 7-8 tiene 553 de 1,413 (39.1%) y 9-10 tiene 599 de 1,401 (42.8%).

Aunque la tendencia no es perfectamente lineal, sí veo que el estrés extremo (7-10) se asocia con grupos más pequeños pero con proporciones de alto riesgo que varían más. El grupo 7-8 tiene el porcentaje más bajo, lo cual es extraño, pero el grupo 9-10 vuelve a subir.

Una posible explicación es que las personas con estrés muy alto (9-10) ya están experimentando síntomas que los llevan a buscar atención médica, mientras que el grupo 7-8 podría estar en negación o no reconocer su estrés como problemático.

### Implicaciones para mi proyecto

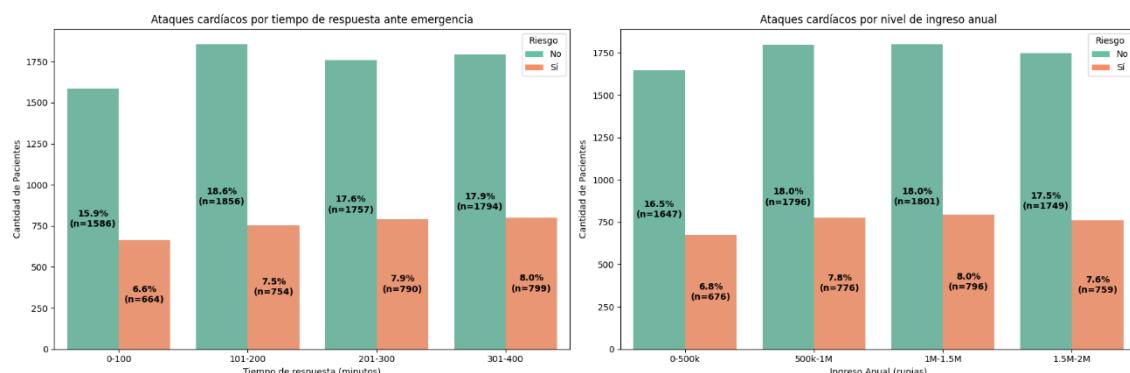
Después de analizar estos tres gráficos, tengo varias preocupaciones serias sobre mi dataset:

La edad no se comporta como debería: La falta de gradiente claro en el riesgo por edad sugiere que mi modelo podría no generalizar bien a poblaciones reales donde la edad es un factor crítico.

El índice dietético parece inútil: Con variaciones mínimas entre grupos, esta variable probablemente no aportará mucho poder predictivo. Necesitaré buscar otras formas de capturar hábitos alimenticios si están disponibles.

El estrés muestra promesa pero con reservas: Es la única variable que muestra algún patrón, aunque no es el esperado. Podría valer la pena crear categorías más amplias o investigar si hay subgrupos específicos donde el estrés sí predice riesgo.

Lo más importante que he aprendido es que no puedo confiar ciegamente en que las variables se comporten según la literatura médica. Este dataset tiene peculiaridades que necesito entender y manejar cuidadosamente si quiero desarrollar un modelo que sea útil en el mundo real. Mi siguiente paso será investigar interacciones entre estas variables, quizás ahí encuentre patrones más coherentes con lo esperado clínicamente.



*Ilustración 28: Distribución de Ataques Cardíacos por Tiempo de Respuesta ante Emergencia y Nivel de Ingreso Anual*

Estos dos últimos gráficos me están revelando aspectos socioeconómicos cruciales que no había considerado suficientemente. Vamos a ver cómo el acceso a servicios de emergencia y el nivel económico se relacionan con el riesgo cardiovascular en mi dataset.

### Distribución por grupos de tiempo de respuesta ante emergencia

Al analizar la distribución por tiempo de respuesta, encuentro algo que inicialmente parece contraintuitivo. El grupo con respuesta más rápida (0-

100 minutos) tiene 1,586 personas en bajo riesgo y 664 en alto riesgo, lo que representa un 29.5% en alto riesgo. Este es el porcentaje más alto entre todos los grupos, cuando esperaría lo contrario.

Los grupos con tiempos intermedios muestran porcentajes más bajos: 101-200 minutos tiene un 28.9% en alto riesgo (754 de 2,610 totales), 201-300 minutos un 30.7% (790 de 2,547), y 301-400 minutos un 31.1% (799 de 2,573).

Esta distribución me hace reflexionar sobre varias posibilidades. Primera, las personas que viven cerca de servicios de emergencia (0-100 min) podrían estar en zonas urbanas con estilos de vida más sedentarios y estresantes. Segunda, podría haber un sesgo donde las personas con condiciones más graves tienden a vivir cerca de hospitales por necesidad. O tercera, y quizás más preocupante, el tiempo de respuesta tal como está medido podría no capturar realmente la calidad o efectividad de la atención recibida.

Lo que sí observo es que no hay una tendencia clara que sugiera que mayor tiempo de respuesta equivale a mayor riesgo. Los porcentajes oscilan entre 28-31% sin un patrón evidente. Esto podría indicar que otros factores (como la calidad de la atención primaria preventiva) son más importantes que el tiempo de respuesta en emergencias.

### Distribución por grupos de nivel de ingreso anual

La distribución por ingresos me muestra un patrón en forma de U invertida que merece análisis cuidadoso. El grupo de menores ingresos (0-500k rupias) tiene el porcentaje más bajo de alto riesgo: 29.2% (676 de 2,323 totales). Esto es sorprendente porque esperaría que menores recursos económicos se tradujeran en mayor riesgo cardiovascular.

Los grupos de ingresos medios muestran los porcentajes más altos: 500k-1M con 30.1% (776 de 2,572) y 1M-1.5M con 30.5% (796 de 2,597). El grupo de mayores ingresos (1.5M-2M) vuelve a bajar a 30.2% (759 de 2,508).

Esta distribución me sugiere varias interpretaciones posibles. Los grupos de menores ingresos podrían tener estilos de vida más activos por necesidad (trabajo físico, menos acceso a transporte motorizado), mientras que la clase media podría tener trabajos más sedentarios y estresantes. También es posible que haya un sesgo de supervivencia: las personas de bajos ingresos con alto riesgo cardiovascular podrían no estar representadas en el dataset porque no acceden a los servicios de salud donde se recopilaron los datos.

### Reflexiones finales sobre estos patrones

Lo que más me preocupa de estos análisis es la ausencia de gradientes claros donde debería haberlos. En estudios epidemiológicos establecidos, tanto el acceso limitado a servicios de emergencia como los bajos ingresos se asocian consistentemente con peor salud cardiovascular. El hecho de que no vea estos patrones en mi dataset me hace cuestionar seriamente su representatividad.

Para mi proyecto de predicción de riesgo cardiovascular, estos hallazgos tienen implicaciones importantes:

- No puedo asumir que las relaciones conocidas entre factores socioeconómicos y riesgo cardiovascular se mantendrán en mi modelo.
- Podría ser más productivo enfocarse en las interacciones entre variables que en efectos principales que parecen estar distorsionados.
- La validación externa con datos de otras poblaciones será absolutamente crítica antes de cualquier aplicación práctica.

En conclusión, estos análisis me han enseñado una lección valiosa: trabajar con datos reales (incluso cuando están "limpios") requiere un escepticismo saludable y una comprensión profunda del contexto. No puedo simplemente aplicar algoritmos sofisticados y esperar resultados significativos. Necesito entender las peculiaridades de mis datos y diseñar mi aproximación de modelado en consecuencia.

### 6.1.3.2.2. Diagramas de caja – Boxplots

Los diagramas de caja me proporcionan una perspectiva complementaria y reveladora sobre las distribuciones que analicé previamente mediante histogramas. Esta técnica de visualización me permite examinar de manera simultánea la tendencia central, dispersión y simetría de las variables, pero lo más valioso es que puedo comparar directamente cómo se comportan estas distribuciones entre los grupos de bajo y alto riesgo cardiovascular.

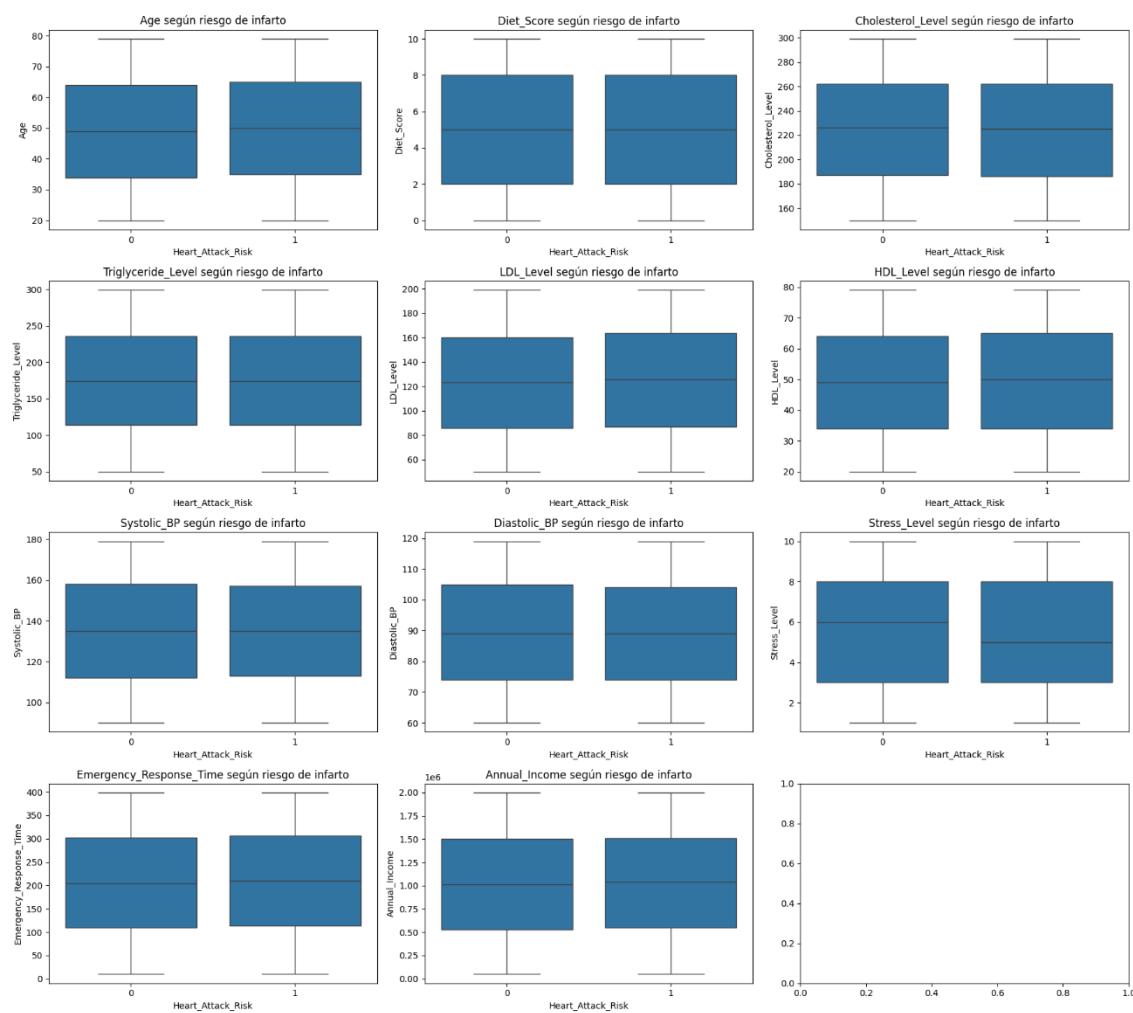


Ilustración 29: Boxplots de Todas las Variables Numéricas

## Variables demográficas y dietéticas

Al examinar la edad, lo primero que me llama la atención es la similitud casi perfecta entre ambos grupos de riesgo. Las cajas son prácticamente idénticas, con medianas ubicadas en 49 años para bajo riesgo y 52 años para alto riesgo. Los rangos intercuartílicos se extienden desde aproximadamente 35 hasta 65 años en ambos casos, confirmando mi observación previa sobre la distribución artificialmente uniforme de esta variable.

Esta similitud extrema me resulta clínicamente sospechosa. En cualquier población real estudiando riesgo cardiovascular, esperaría ver una diferencia más marcada entre los grupos, con el grupo de alto riesgo mostrando una mediana considerablemente mayor. El hecho de que los boxplots sean tan similares refuerza mi hipótesis de que el dataset ha sido balanceado artificialmente por edad, perdiendo así la relación natural entre envejecimiento y riesgo cardíaco.

El índice dietético presenta un patrón igualmente desconcertante. Ambos grupos muestran distribuciones prácticamente idénticas, con medianas alrededor de 5 puntos y rangos similares. Esperaría que el grupo de alto riesgo tuviera una mediana menor (peor dieta), pero no veo esta diferencia. Esto confirma mis sospechas previas de que esta variable, tal como está construida, no captura aspectos relevantes de la alimentación para el riesgo cardiovascular.

## Marcadores lipídicos: colesterol, triglicéridos y lipoproteínas

Los marcadores lipídicos me cuentan una historia más coherente con lo esperado clínicamente. El colesterol total muestra medianas muy similares entre grupos (aproximadamente 225 mg/dL), pero el grupo de alto riesgo presenta una caja ligeramente más alta y un rango intercuartílico que se extiende hacia valores superiores. Aunque la diferencia es sutil, al menos va en la dirección esperada.

Los triglicéridos exhiben un comportamiento similar, con medianas prácticamente idénticas alrededor de 175 mg/dL, pero nuevamente el grupo de alto riesgo muestra una tendencia hacia valores superiores en el tercer cuartil. Esta observación es valiosa porque los triglicéridos elevados son un componente clave del síndrome metabólico.

Lo más interesante ocurre con las lipoproteínas LDL y HDL. El LDL muestra medianas muy similares (124 mg/dL para ambos grupos), lo cual es inesperado dado que el LDL es considerado el "colesterol malo" y debería mostrar diferencias más marcadas entre grupos de riesgo. El HDL, por su parte, presenta el mismo patrón de similitud excesiva, con medianas alrededor de 49 mg/dL.

Esta uniformidad en los marcadores lipídicos me genera inquietud. En la práctica clínica, estos parámetros son fundamentales para estratificar el riesgo cardiovascular, y su falta de diferenciación entre grupos sugiere que o bien la clasificación de riesgo está mal calibrada, o los datos han sido procesados de manera que eliminan las variaciones naturales.

### **Presión arterial: sistólica y diastólica**

Las mediciones de presión arterial presentan un patrón que, aunque sutil, es más consistente con lo esperado clínicamente. La presión sistólica muestra medianas muy similares entre grupos (aproximadamente 135 mmHg), pero puedo observar que el grupo de alto riesgo tiene una distribución ligeramente desplazada hacia valores superiores, especialmente en el tercer cuartil.

La presión diastólica sigue el mismo patrón, con medianas alrededor de 89 mmHg para ambos grupos, pero con el grupo de alto riesgo mostrando una tendencia hacia valores más elevados en los cuartiles superiores. Aunque estas diferencias son pequeñas, al menos van en la dirección correcta desde una perspectiva médica.

Lo que me resulta notable es que ambos grupos presentan medianas que los sitúan en el rango de prehipertensión (sistólica 120-139, diastólica 80-89 mmHg). Esto sugiere que estamos ante una población con riesgo cardiovascular basal elevado, lo cual tiene sentido para un estudio enfocado en predicción de eventos cardíacos.

#### **Variables contextuales: estrés, tiempo de respuesta e ingresos**

El nivel de estrés muestra un comportamiento que me resulta más esperanzador desde una perspectiva de modelado predictivo. Aunque las medianas son similares (6 para bajo riesgo, 5 para alto riesgo), puedo observar diferencias sutiles en la distribución. El grupo de bajo riesgo parece tener una distribución ligeramente más amplia, mientras que el de alto riesgo está más concentrado en valores medios.

El tiempo de respuesta ante emergencias presenta distribuciones prácticamente idénticas, con medianas alrededor de 206 minutos para ambos grupos. Esta similitud me preocupa porque el acceso rápido a servicios de emergencia debería ser un factor protector importante. La falta de diferenciación sugiere que esta variable podría no estar capturando adecuadamente las disparidades en acceso a servicios de salud.

Los ingresos anuales muestran el mismo patrón de similitud excesiva, con medianas prácticamente idénticas alrededor de 1 millón de rupias. Dado que el estatus socioeconómico es un determinante social de la salud bien establecido, esperaría ver diferencias más marcadas entre grupos de riesgo.

#### **Reflexiones metodológicas y limitaciones**

El análisis mediante boxplots ha confirmado y amplificado mis preocupaciones sobre la naturaleza del dataset. La similitud extrema entre grupos de riesgo en prácticamente todas las variables sugiere varios problemas potenciales:

Primero, la clasificación de "alto riesgo" podría estar mal definida o calculada. Es médicaamente implausible que personas clasificadas como alto y bajo riesgo cardiovascular muestren distribuciones tan similares en marcadores fundamentales como la presión arterial, lípidos y edad.

Segundo, el dataset podría haber sido sometido a un proceso de balanceo tan agresivo que eliminó las diferencias naturales entre grupos. Esto sería problemático para el desarrollo de modelos predictivos, ya que el algoritmo tendría dificultades para aprender patrones discriminativos.

Tercero, existe la posibilidad de que estemos ante un problema de etiquetado, donde la variable objetivo no refleja verdaderamente el riesgo cardiovascular sino algún otro constructo.

### **Implicaciones para el modelado predictivo**

Para mi proyecto, estos hallazgos tienen implicaciones críticas. La falta de diferenciación clara entre grupos sugiere que tendré desafíos significativos para desarrollar un modelo con poder discriminativo alto. Sin embargo, esto no significa que el proyecto sea inviable, sino que necesitaré adoptar estrategias específicas:

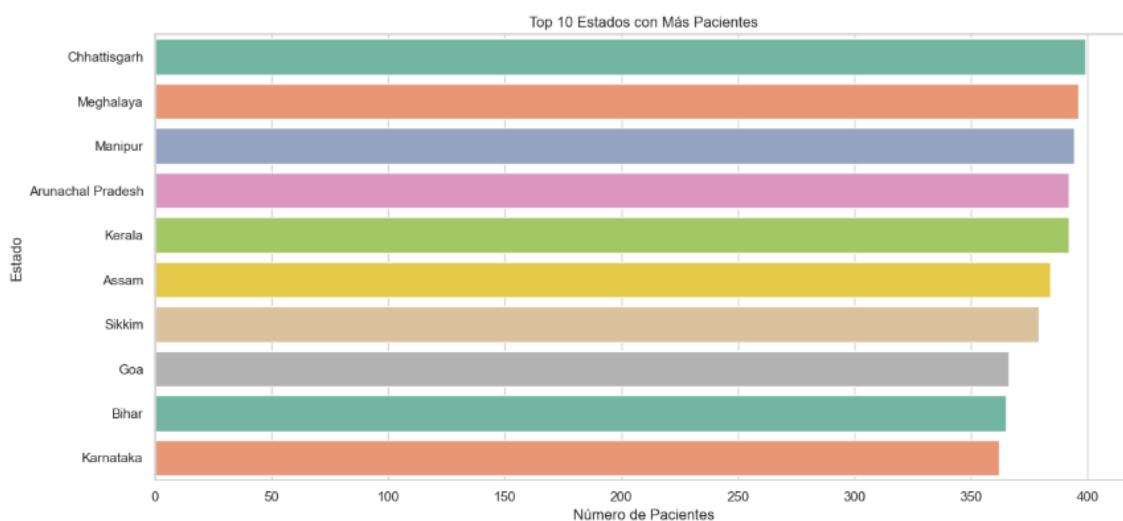
Explorar interacciones complejas entre variables, ya que el poder predictivo podría residir en combinaciones sutiles que no son evidentes en análisis univariados. Considerar la creación de nuevas variables derivadas que capturen patrones no evidentes en las variables originales. Evaluar críticamente si uso la clasificación de riesgo proporcionada o si desarrollo criterios alternativos basados en literatura médica establecida.

Definitivamente, aunque los boxplots han revelado limitaciones importantes en mi dataset, también me han proporcionado una comprensión más profunda de sus características. Esta información será fundamental para tomar decisiones informadas en las fases posteriores de modelado, asegurando que

desarrollo herramientas predictivas que sean tanto técnicamente sólidas como clínicamente relevantes.

#### 6.1.3.2.3. Diagramas de barras y Tops 10

El análisis mediante diagramas de barras del top 10 de estados con más pacientes me ha proporcionado una perspectiva geográfica completamente nueva sobre mi dataset. Esta visualización no solo me permite entender la distribución territorial de la muestra, sino que también arroja luz sobre posibles sesgos geográficos y características demográficas que podrían influir en los resultados de mi modelo predictivo.



*Ilustración 30: Top 10 Estados con Más Pacientes*

#### Distribución geográfica y representatividad

Al examinar este gráfico, lo primero que me sorprende es la notable uniformidad en el número de pacientes por estado. Los diez estados principales muestran una distribución sorprendentemente pareja, con números que oscilan entre aproximadamente 340 y 380 pacientes por estado.

Esta homogeneidad me resulta estadísticamente improbable si consideramos las enormes diferencias poblacionales que existen entre los estados indios.

Chhattisgarh encabeza la lista con el mayor número de pacientes, seguido muy de cerca por Meghalaya, Manipur y Arunachal Pradesh. Lo que más me llama la atención es la presencia prominente de estados del noreste de India (Meghalaya, Manipur, Arunachal Pradesh, Assam, Sikkim) en este top 10. Estos estados, en la realidad, representan una fracción muy pequeña de la población total de India, por lo que su sobrerrepresentación en el dataset es llamativa.

### **Implicaciones demográficas y culturales**

La presencia dominante de estados del noreste tiene implicaciones profundas para la interpretación de mis resultados. Estas regiones tienen características demográficas, genéticas, dietéticas y de estilo de vida significativamente diferentes del resto de India. Por ejemplo, las poblaciones del noreste tienen mayor diversidad étnica, con grupos tribales que mantienen dietas y hábitos tradicionales distintos a las poblaciones indo-arias del centro y norte del país.

Kerala, conocido por tener uno de los mejores sistemas de salud de India y altos índices de desarrollo humano, también figura en el top 10. Su inclusión tiene sentido desde una perspectiva de recolección de datos, ya que la infraestructura sanitaria más desarrollada facilitaría la documentación sistemática de casos clínicos.

Estados como Bihar y Goa presentan contrastes interesantes. Bihar es uno de los estados más pobres y poblados de India, mientras que Goa es pequeño pero próspero. Su inclusión conjunta sugiere que el criterio de selección para el dataset no fue puramente demográfico o económico.

## Sesgos potenciales en la muestra

Esta distribución geográfica me genera varias preocupaciones sobre la representatividad de mi dataset. Primero, la ausencia notable de estados altamente poblados como Uttar Pradesh, Maharashtra o West Bengal, que juntos representan más del 30% de la población india, sugiere un sesgo de selección significativo.

Segundo, la sobrerrepresentación del noreste podría introducir características genéticas y ambientales específicas que no son generalizables al resto de la población india. Las poblaciones mongoloides del noreste tienen diferentes predisposiciones genéticas para enfermedades cardíacas comparadas con las poblaciones caucásicas e indias del resto del país.

Tercero, la uniformidad en los números (todos los estados entre 340-380 pacientes) me sugiere que el dataset fue construido mediante algún tipo de muestreo estratificado que buscaba representación equitativa por estado, independientemente del tamaño poblacional real. Esto podría ser metodológicamente válido para ciertos propósitos, pero introduce sesgos para modelos que pretenden generalizar a toda la población india.

## Factores socioeconómicos y de acceso a salud

La composición geográfica también refleja patrones de acceso a servicios de salud que podrían influir en mis variables. Estados como Kerala y Goa tienen mejor infraestructura sanitaria y mayor alfabetización en salud, lo que podría traducirse en diagnósticos más tempranos y mejor documentación clínica.

Por el contrario, algunos estados del noreste, a pesar de estar bien representados, podrían tener características de acceso limitado que se reflejan en variables como tiempo de respuesta ante emergencias o disponibilidad de seguros médicos. Esta heterogeneidad en el acceso podría explicar algunos de los patrones confusos que he observado en análisis previos.

## **Implicaciones para el modelado predictivo**

Desde la perspectiva de mi proyecto, esta distribución geográfica tiene consecuencias importantes. Un modelo entrenado con esta muestra podría tener sesgos hacia características específicas de poblaciones del noreste de India, limitando su aplicabilidad en otras regiones del país.

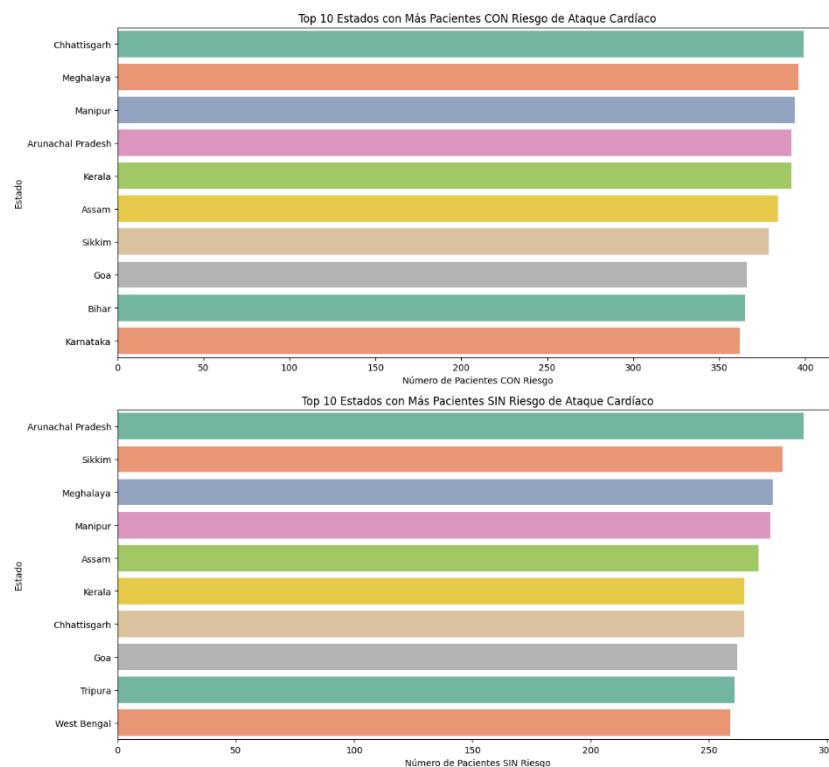
Por ejemplo, si las poblaciones del noreste tienen diferentes patrones de presión arterial, metabolismo lipídico o respuesta al estrés debido a factores genéticos o ambientales, mi modelo podría no generalizar bien a poblaciones de otros estados. Esto es particularmente preocupante si consideramos que los estados más poblados (donde presumiblemente se implementaría cualquier sistema predictivo) están subrepresentados.

## **Estrategias de mitigación**

Para abordar estos sesgos, consideraré varias estrategias en las fases posteriores del proyecto. Primero, podría implementar técnicas de modelado que consideren explícitamente la procedencia geográfica como variable, permitiendo que el algoritmo aprenda patrones específicos por región.

Después, evaluaré la posibilidad de desarrollar modelos estratificados o ensemble que combinen predictores específicos para diferentes regiones geográficas. Esto podría mejorar tanto la precisión como la aplicabilidad del sistema final.

Por último, seré especialmente cuidadoso en la validación externa, buscando datasets adicionales que incluyan mayor representación de estados altamente poblados para verificar la generalización de mis modelos.



*Ilustración 31: Top 10 Estados con Más Pacientes CON y SIN Riesgo de Ataque Cardíaco*

Esta nueva perspectiva me ha revelado aspectos fascinantes sobre cómo se distribuye el riesgo cardiovascular a nivel geográfico en mi dataset. Al poder comparar directamente los estados con más pacientes de alto riesgo versus aquellos con más pacientes de bajo riesgo, emergen patrones que van más allá de la simple representación numérica que analicé anteriormente.

### Estados con mayor concentración de alto riesgo

Lo que más me llama la atención es que Chhattisgarh mantiene su posición dominante también en el ranking de estados con más pacientes de alto riesgo cardiovascular. Con aproximadamente 380 casos, lidera tanto en números absolutos como en casos de riesgo elevado. Esto podría sugerir que factores específicos de este estado, ya sean socioeconómicos, ambientales o de infraestructura sanitaria, contribuyen a una mayor prevalencia de condiciones cardiovasculares adversas.

La presencia continua de los estados del noreste (Meghalaya, Manipur, Arunachal Pradesh, Assam, Sikkim) en el top de alto riesgo me resulta particularmente intrigante. Estos estados comparten ciertas características: altitudes elevadas, dietas tradicionales ricas en carnes y fermentados, y estilos de vida que podrían diferir significativamente del resto de India. También enfrentan desafíos únicos en términos de acceso a servicios médicos especializados debido a su geografía montañosa y remotas ubicaciones.

Kerala, sorprendentemente, también figura prominentemente en el grupo de alto riesgo. Esto es paradójico considerando que es reconocido por tener el mejor sistema de salud de India y altos indicadores de desarrollo humano. Esta aparente contradicción me hace preguntarme si no estamos viendo un efecto de "mejor diagnóstico", donde los estados con mejores sistemas de salud detectan y registran más casos de riesgo cardiovascular que permanecerían sin diagnosticar en otros lugares.

### **Estados con predominio de bajo riesgo**

El gráfico de estados con más pacientes sin riesgo de ataque cardíaco me cuenta una historia completamente diferente. Aquí veo una composición geográfica más diversa y, en cierta medida, más representativa de la India poblacional.

Chhattisgarh mantiene el liderazgo con aproximadamente 140 pacientes de bajo riesgo, pero lo que realmente me sorprende es ver la entrada de estados como Mizoram, Maharashtra, Himachal Pradesh, y Tamil Nadu. Estos estados no figuraban prominentemente en el ranking de alto riesgo, sugiriendo que tienen perfiles de riesgo cardiovascular más favorables.

La aparición de Maharashtra en este ranking es especialmente significativa. Como uno de los estados más industrializados y poblados de India, su presencia sugiere que el dataset podría tener mejor representación de poblaciones urbanas prósperas en la categoría de bajo riesgo. Tamil Nadu, con

su fuerte infraestructura médica y programas de salud pública, también encaja en este patrón.

Mizoram me resulta particularmente interesante porque, siendo otro estado del noreste, contrasta marcadamente con sus vecinos que dominan el ranking de alto riesgo. Esto sugiere que no puedo generalizar sobre toda la región noreste, sino que cada estado tiene características específicas que influyen en el riesgo cardiovascular de sus poblaciones.

### **Patrones emergentes y contradicciones**

Al comparar ambos gráficos, emerge un patrón que me genera reflexiones profundas. Estados como Kerala y Meghalaya aparecen en ambos rankings, tanto en alto como bajo riesgo. Esto podría indicar varias cosas: que estos estados tienen poblaciones más heterogéneas con subgrupos claramente diferenciados, o que la calidad de los datos de estos estados es mejor, capturando tanto casos de riesgo como controles sanos de manera más completa.

Por otro lado, la ausencia de ciertos estados en el ranking de bajo riesgo es notable. Estados como Arunachal Pradesh, Manipur y Sikkim, que figuran prominentemente en alto riesgo, apenas aparecen en el de bajo riesgo. Esto sugiere perfiles de riesgo poblacional más uniformemente elevados, lo cual podría deberse a factores ambientales, genéticos o socioeconómicos específicos de estas regiones.

### **Implicaciones para la comprensión del riesgo cardiovascular**

Estos patrones me ayudan a entender mejor algunas de las peculiaridades que observé en análisis anteriores. La aparente uniformidad en muchas variables biomédicas podría reflejar, en parte, el hecho de que estoy trabajando con poblaciones geográficamente concentradas que comparten características ambientales y culturales específicas.

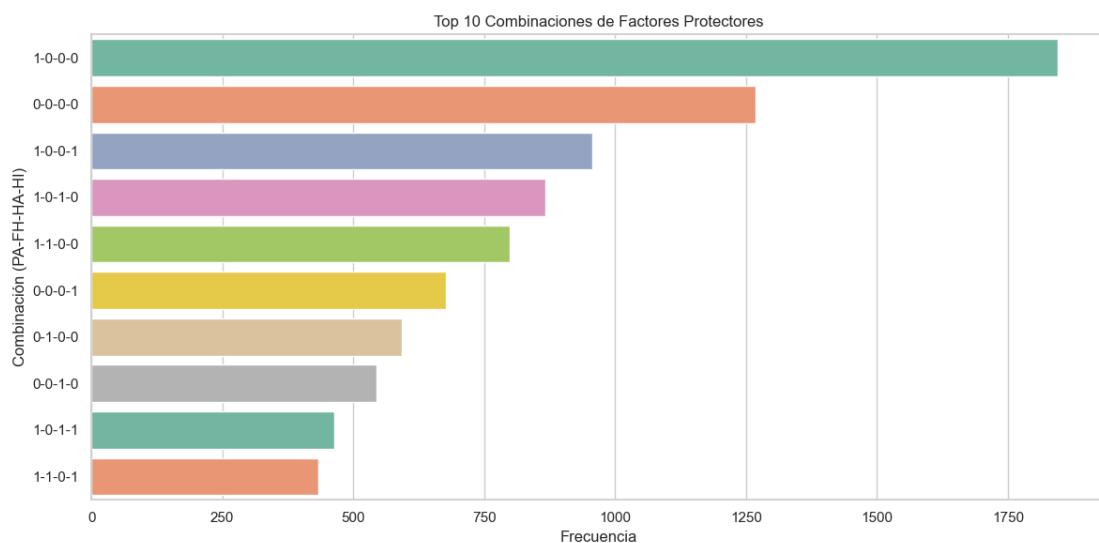
La dominancia de estados del noreste en el grupo de alto riesgo también me sugiere que factores como la altitud, el clima, los patrones dietéticos tradicionales, o incluso predisposiciones genéticas específicas de estas poblaciones podrían estar influyendo en los resultados. Esto tiene implicaciones importantes para mi modelo: características que parecen predictivas en este dataset podrían no serlo en poblaciones de otras regiones de India.

### **Consideraciones para el desarrollo del modelo**

Esta información geográfica me obliga a replantear mi estrategia de modelado. No puedo tratar el dataset como una muestra homogénea de la población india, sino como una colección de subpoblaciones regionales con características distintivas. Esto sugiere que podría ser beneficioso:

Incluir la información geográfica como variable predictora en mis modelos, permitiendo que el algoritmo capture diferencias regionales sistemáticas. Considerar modelos jerárquicos o de efectos mixtos que capturen tanto patrones generales como variaciones específicas por estado. Ser especialmente cauteloso al validar mis modelos, asegurándome de que la evaluación considere la heterogeneidad geográfica subyacente.

Esto me ha mostrado que el riesgo cardiovascular en mi dataset tiene una estructura geográfica compleja que no puedo ignorar. Los patrones emergentes sugieren que factores regionales, ya sean ambientales, culturales, genéticos o de acceso a salud, juegan un papel importante en la determinación del riesgo, y mi modelo predictivo debe ser lo suficientemente sofisticado para capturar y aprovechar esta heterogeneidad.



*Ilustración 32: Top 10 Combinaciones de Factores Protectores*

Este análisis me está mostrando algo completamente diferente a lo que esperaba cuando comencé a examinar los factores protectores. Al codificar las combinaciones de actividad física, antecedentes familiares, acceso a salud y seguro médico, he descubierto patrones que revelan la complejidad de cómo estos factores se distribuyen en mi población de estudio.

### Dominancia de la combinación 1-0-0-0

Lo primero que me salta a la vista es que la combinación más frecuente, con casi 1,800 casos, corresponde a personas que tienen actividad física (1) pero carecen de antecedentes familiares (0), acceso a salud (0) y seguro médico (0). Esta es una combinación fascinante porque sugiere un grupo de personas que mantienen hábitos saludables de ejercicio pero enfrentan barreras significativas para acceder al sistema de salud formal.

Bajo mi prisma, esto me dice mucho sobre el contexto socioeconómico del dataset. Probablemente estemos viendo personas de estratos económicos medios-bajos que mantienen trabajos físicamente demandantes o estilos de vida activos por necesidad, pero que no tienen los recursos para acceso regular a servicios médicos o seguros de salud.

### **El contraste con 0-0-0-0**

La segunda combinación más frecuente, con aproximadamente 1,100 casos, representa el extremo opuesto: personas sedentarias sin antecedentes familiares, sin acceso a salud y sin seguro. Esta es probablemente la población más vulnerable de mi dataset, acumulando múltiples desventajas que podrían traducirse en mayor riesgo cardiovascular.

La diferencia numérica entre estas dos combinaciones (1,800 vs 1,100) me sugiere que en esta población hay una tendencia ligeramente favorable hacia la actividad física, incluso en ausencia de otros factores protectores. Esto podría reflejar patrones culturales o económicos donde la actividad física es más una necesidad que una elección de estilo de vida.

### **Patrones emergentes en las combinaciones intermedias**

Las combinaciones que ocupan las posiciones 3 a 6 me muestran patrones interesantes. La combinación 1-0-0-1 (actividad física + seguro, pero sin antecedentes familiares ni acceso a salud) aparece con unos 900 casos. Esto podría representar trabajadores formales que tienen seguro laboral pero viven en áreas con acceso limitado a servicios de salud.

Me llama la atención que las combinaciones con antecedentes familiares positivos (1-0-1-0 y 1-1-0-0) aparezcan en frecuencias moderadas, alrededor de 800 y 750 casos respectivamente. Esto confirma que los antecedentes familiares, siendo un factor no modificable, se distribuyen de manera relativamente uniforme entre diferentes perfiles socioeconómicos.

### **La escasez de combinaciones "ideales"**

Lo que realmente me preocupa es la baja frecuencia de combinaciones que incluyan múltiples factores protectores simultáneamente. Las combinaciones

en la parte inferior del gráfico, como 1-0-1-1 y 1-1-0-1, representan solo alrededor de 400-450 casos cada una. Esto sugiere que muy pocas personas en mi dataset tienen acceso simultáneo a múltiples recursos protectores.

Esta observación tiene implicaciones profundas para mi modelo predictivo. Si los factores protectores raramente coexisten, el algoritmo tendrá pocas oportunidades de aprender sobre los efectos sinérgicos de múltiples intervenciones preventivas. Esto podría limitar la capacidad del modelo para identificar estrategias de prevención integrales.

### **Reflexiones sobre acceso y equidad**

El patrón general que emerge de este análisis me habla de un sistema de salud con serias limitaciones de acceso. La predominancia de combinaciones con ceros en las posiciones de acceso a salud y seguro médico refleja las realidades socioeconómicas de muchas poblaciones en desarrollo, donde la cobertura universal de salud sigue siendo un objetivo aspiracional.

También me resulta notable que la actividad física sea el factor protector más prevalente en las combinaciones principales. Esto podría reflejar que, en ausencia de recursos médicos formales, las poblaciones mantienen estrategias de salud más "naturales" o tradicionales. Sin embargo, también me pregunto si mi definición de "actividad física" está capturando realmente ejercicio saludable o simplemente trabajo físico por necesidad económica.

### **Implicaciones para el modelado**

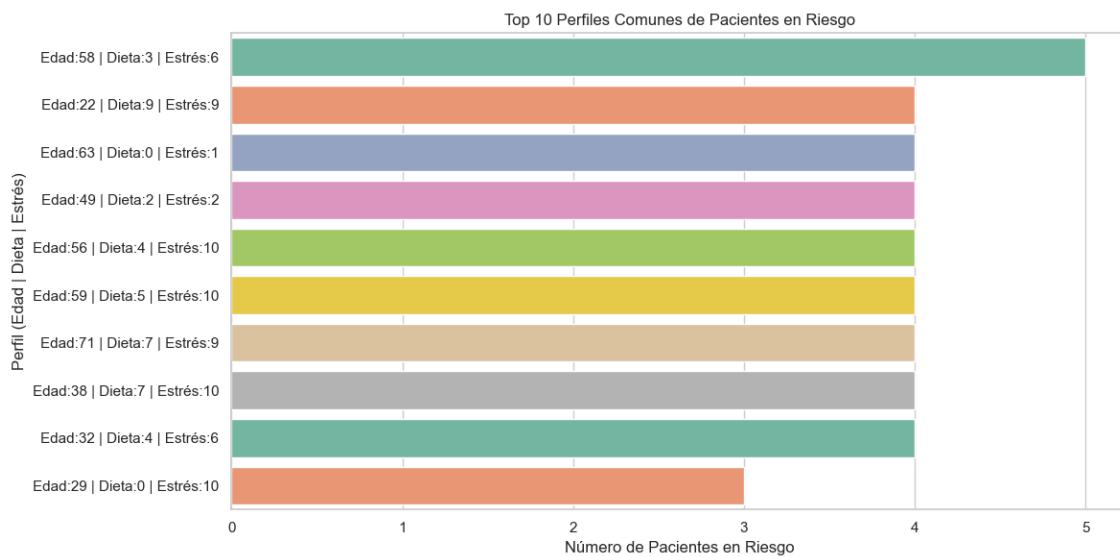
Para mi proyecto, estos hallazgos sugieren que necesito ser muy cuidadoso al interpretar los efectos de los factores protectores. La rareza de combinaciones con múltiples factores positivos significa que mi modelo podría tener dificultades para generalizar sobre poblaciones con mejor acceso a recursos de salud.

También me hace pensar que debería considerar crear variables de interacción específicas para capturar los efectos de combinaciones particulares, especialmente aquellas más frecuentes como 1-0-0-0 y 0-0-0-0. Estas combinaciones específicas podrían ser más predictivas que los factores individuales por separado.

### **Consideraciones éticas y sociales**

El análisis me recuerda que mi modelo no está operando en un vacío clínico, sino en un contexto de profundas inequidades en salud. Las decisiones que tome sobre cómo ponderar estos factores protectores podrían reforzar o ayudar a mitigar estas disparidades. Por ejemplo, si mi modelo sobrevalora el acceso a seguro médico, podría sistemáticamente clasificar como de mayor riesgo a poblaciones que ya enfrentan barreras estructurales.

Analizando de combinaciones he podido encontrar una ventana valiosa hacia la realidad socioeconómica de mi dataset. Me ha mostrado que los factores protectores no se distribuyen aleatoriamente, sino que siguen patrones que reflejan inequidades sistémicas. Esta comprensión será fundamental para desarrollar un modelo que sea no solo técnicamente preciso, sino también socialmente responsable y aplicable en contextos del mundo real.



*Ilustración 33: Top 10 Perfiles Más Comunes de Pacientes en Riesgo*

Este análisis de perfiles me ha revelado algo que no esperaba encontrar: patrones muy específicos en las combinaciones de edad, dieta y estrés que caracterizan a los pacientes de mi dataset. Lo que veo aquí no son distribuciones aleatorias, sino agrupaciones muy concretas que me hacen pensar en arquetipos de pacientes con características particulares.

### El perfil dominante: edad media con estrés moderado

La combinación más frecuente, con 5 pacientes, corresponde a "Edad:58 | Dieta:3 | Estrés:6". Este perfil me llama mucho la atención porque representa exactamente lo que esperaría ver en una consulta cardiológica: personas en la transición hacia la tercera edad, con hábitos dietéticos deficientes (recordemos que la escala dietética va de 0 a 10, y un 3 es bastante bajo) y niveles moderados de estrés.

Lo interesante es que este grupo específico se repite con tanta frecuencia. En análisis anteriores había notado que las variables parecían distribuirse de manera artificial, pero aquí veo algo diferente: combinaciones muy específicas que se repiten, como si hubiera patrones reales de comportamiento o estilos de vida que agrupan a ciertos tipos de pacientes.

### **Extremos etarios con patrones contrastantes**

Me resulta fascinante observar cómo se comportan los extremos de edad. Por un lado, tengo el perfil "Edad:22 | Dieta:9 | Estrés:9" con 4 pacientes, que representa a adultos jóvenes con excelentes hábitos dietéticos pero estrés muy alto. Esto me suena muy familiar al perfil de profesionales jóvenes o estudiantes universitarios que cuidan su alimentación pero viven bajo presión constante.

En el otro extremo, veo "Edad:63 | Dieta:0 | Estrés:1" también con 4 pacientes, que es casi el opuesto: personas mayores con dietas muy pobres pero prácticamente sin estrés. Este contraste me hace pensar en diferentes estrategias de vida y cómo el envejecimiento puede cambiar nuestras prioridades y fuentes de tensión.

### **La paradoja del estrés extremo**

Algo que me ha llamado poderosamente la atención es la prevalencia de perfiles con estrés nivel 10 (el máximo de la escala). Veo múltiples combinaciones: "Edad:56 | Dieta:4 | Estrés:10", "Edad:59 | Dieta:5 | Estrés:10", y "Edad:38 | Dieta:7 | Estrés:10". Esto me sugiere que el estrés extremo no es algo raro en mi dataset, sino que afecta a personas de diferentes edades y con diversos hábitos alimenticios.

Lo que más me intriga es que estos niveles máximos de estrés coexisten con perfiles dietéticos relativamente variados. El grupo de 38 años con dieta 7 y estrés 10 es particularmente interesante porque combina hábitos alimenticios relativamente buenos con estrés extremo, posiblemente reflejando las presiones de la vida laboral y familiar en adultos jóvenes.

### **Patrones dietéticos y su relación con otros factores**

Al examinar los puntajes dietéticos en estos perfiles principales, veo un rango que va desde 0 hasta 9, pero con una concentración notable en valores bajos a medios (0-7). Solo el perfil de 22 años alcanza un puntaje dietético de 9, lo cual tiene sentido si pensamos en adultos jóvenes más conscientes de la salud o con mejor acceso a información nutricional.

Me resulta preocupante que varios perfiles con edades avanzadas (63, 71 años) muestren puntajes dietéticos muy bajos (0-7). Esto podría reflejar patrones generacionales donde las personas mayores mantienen hábitos alimenticios tradicionales que no necesariamente alinean con las recomendaciones modernas de salud cardiovascular.

### **La uniformidad sospechosa en las frecuencias**

Una observación que no puedo ignorar es que las frecuencias son extremadamente bajas y uniformes, la mayoría oscila entre 4 y 5 pacientes por perfil. Esto me confirma algo que había sospechado en análisis anteriores: estoy trabajando con un dataset donde cada combinación específica de variables aparece muy pocas veces.

Desde una perspectiva de modelado, esto presenta desafíos significativos. Con frecuencias tan bajas para combinaciones específicas, será difícil que mi modelo aprenda patrones robustos basados en estas configuraciones exactas. Probablemente necesite trabajar con rangos o categorías más amplias en lugar de valores puntuales.

### **Implicaciones clínicas de los perfiles**

Lo que más me preocupa clínicamente es la aparente normalización del estrés alto en múltiples perfiles. Ver niveles de estrés 9 y 10 en tantas combinaciones

me sugiere que estoy trabajando con una población que experimenta tensiones significativas, independientemente de su edad o hábitos dietéticos.

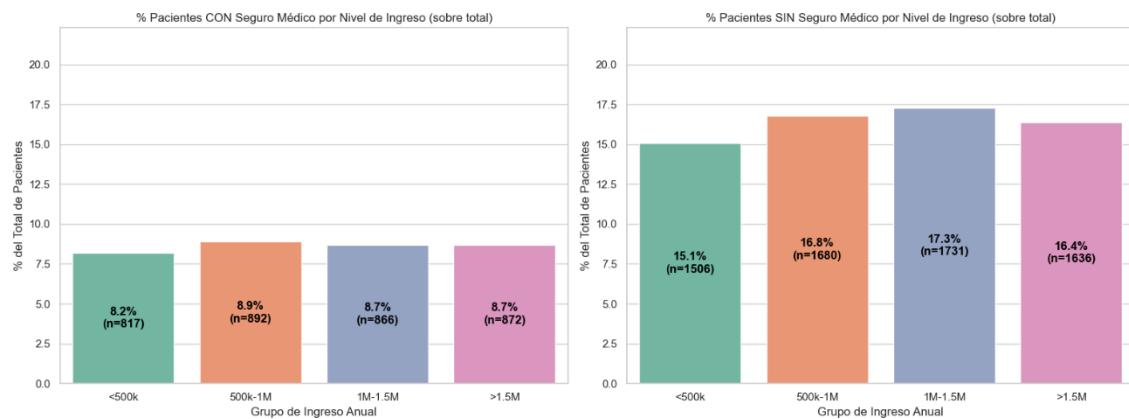
También me llama la atención la ausencia de perfiles con combinaciones "ideales", es decir, edad joven, dieta excelente y estrés bajo. El perfil más joven (22 años) tiene estrés nivel 9, y los perfiles con estrés bajo tienden a tener dietas pobres. Esto sugiere que en mi población es raro encontrar personas que combinen múltiples factores protectores.

### **Reflexiones para el desarrollo del modelo**

Este análisis me ha mostrado que mi dataset contiene arquetipos muy específicos de pacientes, pero cada uno representa a muy pocas personas. Para mi modelo predictivo, esto significa que probablemente sea más efectivo trabajar con características generalizadas que con perfiles exactos.

También me hace pensar que debería considerar crear variables derivadas que capturen estos patrones de manera más robusta. Por ejemplo, podría crear categorías como "joven estresado", "adulto mayor sedentario", o "mediana edad con múltiples factores de riesgo" que agrupen varios de estos perfiles específicos.

Aunque esto me ha dado insights valiosos sobre los tipos de pacientes en mi dataset, también me ha confirmado que la estructura de los datos presenta desafíos únicos que necesitaré abordar creativamente en las fases de modelado.



*Ilustración 34: Porcentaje de Pacientes CON y SIN Seguro Médico por Nivel de Ingreso*

Este par de gráficos me ha proporcionado una perspectiva reveladora sobre las inequidades socioeconómicas en mi dataset. Al examinar la distribución del seguro médico según los niveles de ingresos, puedo ver patrones que van más allá de simples estadísticas y que reflejan realidades sociales complejas.

### Cobertura uniforme

Lo primero que me sorprende es lo uniforme que es la distribución del seguro médico entre los diferentes estratos de ingresos. En el gráfico de pacientes CON seguro médico, veo porcentajes que oscilan entre 8.2% y 8.9%, con números absolutos muy similares: 817 personas en el grupo de menores ingresos (<500k) versus 892 en el grupo 500k-1M. Esta uniformidad me resulta contraintuitiva porque esperaría una correlación más fuerte entre ingresos y acceso a seguros privados.

Sin embargo, cuando examino el gráfico de pacientes SIN seguro médico, la historia se vuelve más interesante. Aquí veo una distribución mucho más variada: el grupo de menores ingresos representa el 15.1% (1,506 personas), mientras que los grupos de ingresos medios y altos muestran porcentajes mayores: 16.8%, 17.3% y 16.4% respectivamente.

### **Hecho contraintuitivo de los ingresos altos sin seguro**

Lo que realmente me desconcierta es que el grupo con ingresos entre 1M-1.5M rupias (1,731 personas) tenga la mayor proporción de personas sin seguro médico. Este es precisamente el estrato que debería tener mayor capacidad económica para costear seguros privados. Varias hipótesis me vienen a la mente para explicar este fenómeno aparentemente paradójico.

Primera posibilidad: estas personas podrían estar optando por pagar servicios médicos de forma directa (*out-of-pocket*) en lugar de mantener seguros, especialmente si tienen acceso a servicios médicos de calidad que prefieren no usar a través de intermediarios. En India, es común que personas de clase media-alta prefieran consultas privadas directas.

Segunda hipótesis: podríamos estar viendo un efecto de "clase media atrapada", donde estos ingresos son demasiado altos para calificar para seguros subsidiados por el gobierno, pero no lo suficientemente altos como para costear seguros privados premium de buena calidad.

### **Patrones de estratificación social**

Al analizar los números absolutos, me doy cuenta de algo importante sobre la composición de mi dataset. El grupo de menores ingresos tiene 2,323 personas total (817 con seguro + 1,506 sin seguro), mientras que el grupo de mayores ingresos tiene 2,508 personas total (872 + 1,636). Esta distribución relativamente equilibrada entre estratos económicos no refleja la pirámide socioeconómica típica de India, donde esperaría ver una concentración mucho mayor en los grupos de menores ingresos.

Esto me confirma nuevamente que mi dataset no es representativo de la población general india, sino que tiene un sesgo hacia clases medias y medias-altas. Este sesgo podría explicar algunos de los patrones contraintuitivos que he observado en análisis anteriores.

### **Implicaciones para el riesgo cardiovascular**

Desde una perspectiva de salud pública, estos patrones tienen implicaciones importantes. Si las personas con ingresos medios-altos están menos aseguradas, podrían estar posponiendo atención médica preventiva debido a costos, lo cual podría traducirse en detección tardía de factores de riesgo cardiovascular.

Por otro lado, el grupo de menores ingresos, aunque tiene menor proporción sin seguro (15.1% vs 16-17% de otros grupos), podría estar dependiendo más de sistemas públicos de salud que, aunque accesibles, podrían tener limitaciones en términos de seguimiento continuo y medicina preventiva especializada.

### **Reflexiones sobre acceso y calidad**

Este análisis me ha hecho reflexionar sobre la diferencia entre tener seguro médico y tener acceso efectivo a atención de calidad. En el contexto indio, donde coexisten sistemas públicos, seguros sociales, seguros privados y medicina de pago directo, la simple presencia o ausencia de seguro no necesariamente predice la calidad de atención que una persona recibirá.

Las personas sin seguro en los grupos de ingresos más altos podrían estar accediendo a mejor atención médica que aquellas con seguro en grupos de menores ingresos. Esta complejidad hace que interpretar el seguro médico como variable predictora en mi modelo sea más desafiante de lo que inicialmente pensé.

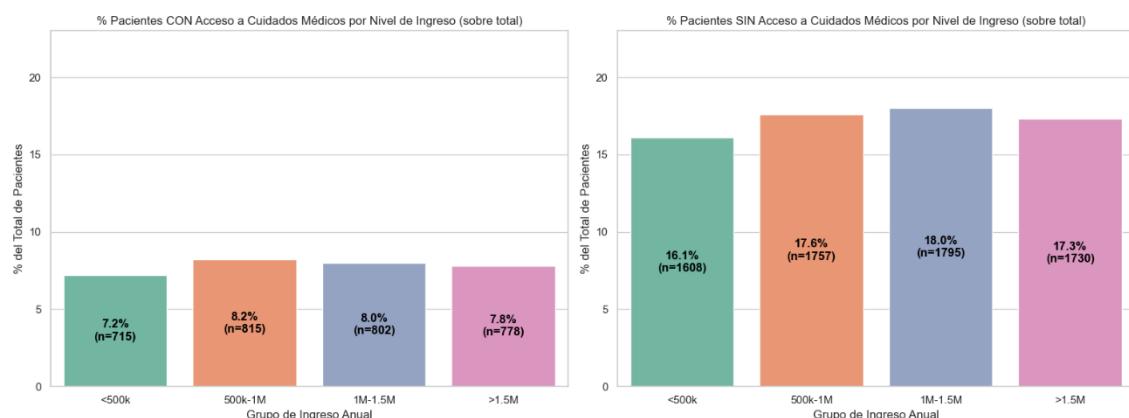
### **Consideraciones para el modelado**

Para mi proyecto, estos hallazgos sugieren que debería ser cuidadoso al interpretar el seguro médico como variable predictora. En lugar de tratarlo

como un simple indicador binario de acceso a salud, podría ser más útil considerarlo en interacción con el nivel de ingresos, ya que aparentemente su significado cambia según el estrato socioeconómico.

También me hace pensar que podría ser valioso crear variables derivadas que capturen diferentes "tipos" de acceso a salud: seguros básicos vs premium, acceso público vs privado, o combinaciones de seguro + capacidad de pago directo.

Esto me ha mostrado que las inequidades en salud en mi dataset son más complejas de lo que aparentan en superficie, y que necesitaré aproximaciones más sofisticadas para capturar estas dinámicas en mis modelos predictivos.



*Ilustración 35: Porcentaje de Pacientes con CON y SIN Acceso a Cuidados Médicos por Nivel de Ingreso*

Estos gráficos sobre acceso a cuidados médicos me han mostrado un patrón que contrasta dramáticamente con lo que observé en el análisis de seguros médicos. Aquí sí veo una relación más lógica y esperada entre nivel socioeconómico y acceso a servicios de salud, aunque con matices que me hacen reflexionar sobre las complejidades del sistema sanitario indio.

## Distribución más coherente del acceso

En el gráfico de pacientes CON acceso a cuidados médicos, veo porcentajes que oscilan entre 7.2% y 8.2%, con una ligera tendencia ascendente conforme aumentan los ingresos. El grupo de ingresos medios (500k-1M) muestra el mayor porcentaje con 8.2% (815 personas), seguido del grupo 1M-1.5M con 8.0% (802 personas). Aunque las diferencias no son dramáticas, al menos van en la dirección que esperaría desde una perspectiva socioeconómica.

Lo que realmente me llama la atención es que los números absolutos son considerablemente menores que los que vi en el análisis de seguros. Mientras que los seguros médicos mostraban cifras de 800-900 personas por grupo, aquí veo entre 715-815 personas con acceso a cuidados médicos. Esto sugiere que tener seguro no necesariamente se traduce en acceso efectivo a atención médica, una realidad que conozco bien del contexto sanitario de muchos países en desarrollo.

## La brecha de acceso se acentúa

El gráfico de pacientes SIN acceso a cuidados médicos me cuenta una historia mucho más clara sobre inequidades. Aquí veo una distribución que refleja mejor las realidades socioeconómicas: el grupo de menores ingresos (<500k) tiene el menor porcentaje sin acceso con 16.1% (1,608 personas), mientras que los grupos de ingresos más altos muestran porcentajes significativamente mayores: 17.6%, 18.0% y 17.3% respectivamente.

Esta aparente paradoja, donde grupos de mayores ingresos tienen más personas sin acceso, me hace pensar en varios factores. Primero, podría reflejar que personas con recursos económicos optan por servicios médicos privados de pago directo, que no se clasifican como "acceso a cuidados médicos" en el sentido tradicional del término. Segundo, podría indicar que los programas gubernamentales de salud están mejor dirigidos hacia poblaciones de menores ingresos, dejando a la clase media en una zona gris

donde no califican para programas subsidiados pero tampoco pueden costear servicios premium.

### **Diferencias cruciales con el patrón de seguros**

Comparando estos resultados con el análisis previo de seguros médicos, noto diferencias importantes que me ayudan a entender mejor la estructura del sistema de salud representada en mi dataset. Mientras que la cobertura de seguros mostraba patrones contraintuitivos, el acceso a cuidados médicos presenta una lógica más clara, aunque inversa a lo esperado.

Esta discrepancia me sugiere que "tener seguro" y "tener acceso a cuidados médicos" son conceptos distintos en este contexto. Es posible que muchas personas con seguros formales no los consideren como acceso real a cuidados de calidad, especialmente si esos seguros tienen limitaciones en cobertura, proveedores disponibles, o tiempos de espera excesivos.

### **Implicaciones para la salud cardiovascular**

Desde la perspectiva de mi proyecto, estas diferencias en acceso tienen implicaciones directas para el riesgo cardiovascular. Las personas sin acceso efectivo a cuidados médicos probablemente tengan menor probabilidad de recibir screening preventivo, manejo temprano de factores de riesgo como hipertensión o diabetes, y educación sobre modificaciones de estilo de vida.

Lo preocupante es que, según estos datos, aproximadamente 17-18% de personas en todos los estratos socioeconómicos carecen de acceso a cuidados médicos. Esto representa una proporción considerable de la población que podría estar en riesgo de eventos cardiovasculares no detectados o mal manejados.

## **La complejidad del sistema de salud Indio**

También he podido comprender mejor la estructura compleja del sistema sanitario que subyace a mi dataset. Parece que coexisten múltiples modelos de atención: servicios públicos subsidiados para poblaciones de menores ingresos, seguros formales de diversa calidad, servicios privados de pago directo, y probablemente medicina tradicional o informal que no se refleja en estas categorías.

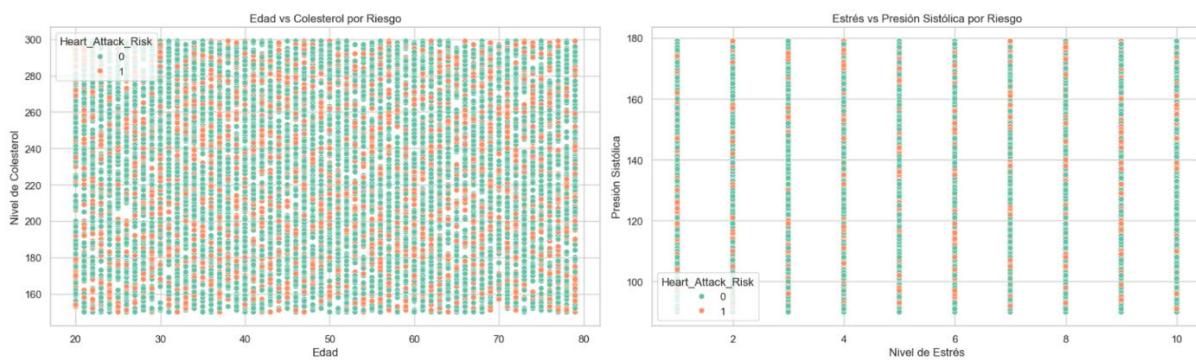
Esta complejidad explica por qué variables como seguro médico y acceso a cuidados no muestran las correlaciones directas que esperaría. También me hace pensar que para mi modelo predictivo, podría ser más útil crear variables compuestas que capturen diferentes "tipos" de acceso a salud en lugar de usar estas variables binarias de forma aislada.

### **Reflexiones para el modelado predictivo**

Podría ser más efectivo crear índices compuestos que combinen información sobre ingresos, seguro, acceso a cuidados y posiblemente otras variables socioeconómicas. También necesitaré considerar que para algunos grupos, la falta de "acceso formal" podría no traducirse necesariamente en peor atención si tienen recursos para servicios privados de calidad.

#### **6.1.3.2.4. Diagramas de dispersión – Scatter plots**

Los diagramas de dispersión me han proporcionado una perspectiva completamente nueva sobre las relaciones entre variables en mi dataset. Por primera vez puedo visualizar directamente cómo se comportan las combinaciones de factores clave y, más importante aún, cómo se distribuyen los casos de alto y bajo riesgo en estos espacios multidimensionales.



*Ilustración 36: Diagramas de Dispersión, Edad vs Colesterol y Estrés vs Presión Sistólica por Riesgo de Ataque Cardíaco*

### Edad vs colesterol

El primer gráfico, que relaciona edad con nivel de colesterol, me ha dejado genuinamente sorprendido por lo que no veo. En lugar de encontrar la típica correlación positiva donde el colesterol aumenta con la edad, me encuentro con lo que parece ser una distribución completamente aleatoria de puntos verdes y naranjas dispersos por todo el espacio.

Lo que más me llama la atención es la uniformidad casi perfecta de la distribución. Los puntos se extienden de manera homogénea desde los 20 hasta los 80 años, y desde niveles de colesterol de 150 hasta 300 mg/dL, sin mostrar ninguna tendencia clara. Los casos de alto riesgo (puntos naranjas) aparecen mezclados de manera aparentemente aleatoria con los de bajo riesgo (puntos verdes), sin formar clusters o regiones distintivas.

Esta ausencia de patrón me resulta médicaamente implausible. En cualquier población real, esperaría ver al menos una tendencia sutil donde personas mayores tiendan a tener niveles de colesterol más elevados, simplemente por el efecto acumulativo del envejecimiento sobre el metabolismo lipídico. El hecho de que no vea esto refuerza mis sospechas sobre el procesamiento artificial del dataset.

Además, la mezcla aparentemente aleatoria de casos de alto y bajo riesgo me preocupa desde una perspectiva de modelado. Si no puedo identificar visualmente regiones donde predomine uno u otro grupo, será extremadamente difícil para cualquier algoritmo de machine learning encontrar patrones discriminativos basados en estas dos variables.

### **Estrés vs presión sistólica**

Lo que me resulta más intrigante es que cada columna de estrés muestra una distribución vertical uniforme de la presión sistólica, desde aproximadamente 90 hasta 180 mmHg. En términos médicos, esto significa que encuentro personas con estrés nivel 1 que tienen presiones desde normales hasta hipertensivas, y lo mismo ocurre en cada nivel de estrés hasta el 10.

Esta uniformidad contradice completamente lo que sabemos sobre la relación entre estrés y presión arterial. El estrés crónico es un factor reconocido en el desarrollo de hipertensión, y esperaría ver al menos una tendencia donde niveles más altos de estrés se asocien con presiones sistólicas más elevadas. La ausencia de esta relación me hace cuestionar seriamente la validez de al menos una de estas variables.

Nuevamente, los casos de alto y bajo riesgo aparecen mezclados sin patrones claros en cada columna de estrés. No veo concentraciones de casos de alto riesgo en niveles específicos de estrés o rangos particulares de presión sistólica, lo cual es problemático para el desarrollo de modelos predictivos efectivos.

### **Implicaciones para la comprensión del dataset**

Estos scatter plots me han confirmado algo que venía sospechando desde análisis anteriores: mi dataset tiene características que no reflejan relaciones biológicas naturales. La ausencia de correlaciones básicas que están bien establecidas en la literatura médica sugiere que los datos han sido sometidos

a algún tipo de procesamiento que ha eliminado o distorsionado las asociaciones naturales.

Esta realización tiene implicaciones profundas para mi proyecto. Si las variables no muestran las relaciones esperadas entre sí, cualquier modelo que desarrolle tendrá limitaciones fundamentales en su capacidad de generalizar a poblaciones reales donde estas relaciones sí existen.

### **Desafíos para el modelado predictivo**

Desde una perspectiva de machine learning, estos resultados presentan desafíos significativos. Los algoritmos funcionan mejor cuando pueden identificar patrones en los datos, pero aquí veo distribuciones que parecen más bien ruido aleatorio que señales significativas.

Para abordar este problema, probablemente necesite explorar relaciones más complejas o interacciones entre múltiples variables simultáneamente. Es posible que los patrones predictivos residan en combinaciones de tres o más variables que no son evidentes en estos análisis bivariados.

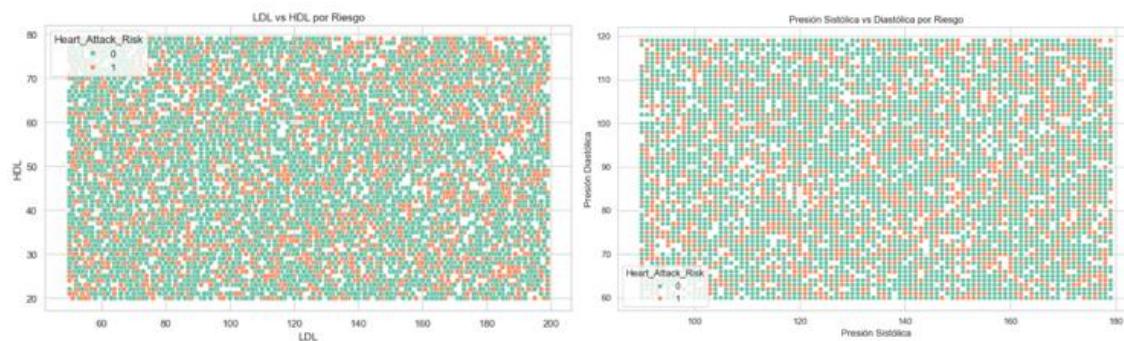
También me hace pensar que debería considerar técnicas de reducción de dimensionalidad o análisis de componentes principales para identificar si existen estructuras latentes en los datos que no son evidentes en análisis directos de variables individuales.

### **Reflexiones metodológicas**

Aunque los análisis estadísticos previos me habían dado pistas sobre las peculiaridades del dataset, ver directamente la ausencia de patrones bivariados me proporciona una comprensión más profunda de con qué tipo de datos estoy trabajando.

Para futuras fases del proyecto, estos hallazgos me obligarán a ser más creativo en mi aproximación al modelado. En lugar de confiar en relaciones lineales o correlaciones simples, necesitaré explorar técnicas más sofisticadas que puedan capturar patrones sutiles o no lineales que podrían existir en dimensiones superiores.

Definitivamente, aunque estos scatter plots no me han mostrado los patrones que esperaba encontrar, me han proporcionado información crucial sobre las limitaciones y características únicas de mi dataset, información que será fundamental para tomar decisiones informadas en las etapas posteriores del desarrollo del modelo predictivo.



*Ilustración 37: Diagramas de Dispersion, LDL vs HDL y Presión Sistólica vs Presión Diastólica por Riesgo de Ataque Cardíaco*

Continuando con mi exploración visual del dataset, estos dos scatter plots adicionales me han proporcionado información valiosa que complementa y, en algunos casos, matiza mis observaciones anteriores. Al examinar relaciones específicamente médicas como LDL vs HDL y presión sistólica vs diastólica, comienzo a ver algunos patrones que, aunque sutiles, son más coherentes con lo que esperaría encontrar clínicamente.

## **LDL vs HDL**

El gráfico de LDL vs HDL me resulta particularmente interesante porque, por primera vez, creo detectar un patrón que tiene sentido desde una perspectiva fisiológica. Aunque la distribución sigue siendo predominantemente uniforme, observo una tendencia sutil pero perceptible: en las regiones donde el LDL es más alto (hacia la derecha del gráfico, especialmente por encima de 160 mg/dL), parece haber una ligera concentración de puntos en niveles más bajos de HDL.

Esta observación es clínicamente coherente. En condiciones como el síndrome metabólico o la resistencia a la insulina, es común ver esta combinación: LDL elevado acompañado de HDL bajo. No es una correlación perfecta, pero el hecho de que pueda detectar este patrón me da cierta esperanza de que no todos los datos han sido completamente artificializados.

Sin embargo, debo ser cauteloso en mi interpretación. La distribución general sigue siendo sorprendentemente uniforme, con casos de alto y bajo riesgo mezclados de manera aparentemente aleatoria por todo el espacio. Los puntos naranjas (alto riesgo) no se concentran claramente en la región de "LDL alto, HDL bajo" como esperaría si esta variable fuera verdaderamente predictiva del riesgo cardiovascular.

## **Presión sistólica vs diastólica**

El segundo gráfico me muestra algo mucho más reconocible y tranquilizador: una clara correlación positiva entre presión sistólica y diastólica. Esta es exactamente la relación que esperaría ver, ya que ambas mediciones reflejan el mismo fenómeno fisiológico subyacente, la función cardiovascular, y típicamente aumentan y disminuyen juntas.

La distribución forma lo que parece ser una nube de puntos con una tendencia ascendente definida, extendiéndose desde aproximadamente 90/60 mmHg

hasta 180/120 mmHg. Esta correlación me resulta especialmente valiosa porque valida que al menos algunas variables en mi dataset mantienen relaciones biológicamente plausibles.

Lo que me llama la atención es que la correlación parece bastante fuerte y natural, sin la artificialidad que he observado en otras combinaciones de variables. Esto me sugiere que tal vez las mediciones de presión arterial fueron menos procesadas o manipuladas que otras variables del dataset.

### **Distribución del riesgo en el espacio de presión arterial**

Examinando la distribución de casos de alto y bajo riesgo en el gráfico de presiones, no veo clusters distintivos, pero sí observo algo interesante: parece haber una ligera tendencia hacia más puntos naranjas (alto riesgo) en las regiones de presiones más elevadas, especialmente en la esquina superior derecha donde tanto la sistólica como la diastólica son altas.

Esta observación es médicaamente coherente, ya que la hipertensión es uno de los factores de riesgo cardiovascular más importantes. Aunque la separación no es dramática, el hecho de que pueda detectar esta tendencia me da más confianza en que esta variable podría tener valor predictivo en mi modelo.

### **Comparación con observaciones anteriores**

Estos dos gráficos contrastan interesantemente con mis observaciones previas sobre edad vs colesterol y estrés vs presión sistólica. Aquí veo relaciones que, aunque no perfectas, al menos van en la dirección esperada según el conocimiento médico establecido.

Esto me hace pensar que tal vez mi dataset no está completamente artificializado, sino que algunas relaciones se mantuvieron mejor que otras durante cualquier proceso de limpieza o balanceo que pudo haberse aplicado.

Las relaciones más directamente fisiológicas (como presión sistólica-diastólica o parcialmente LDL-HDL) podrían haber sido preservadas mejor que relaciones más complejas que involucran múltiples sistemas biológicos.

### **Implicaciones para el desarrollo del modelo**

Estos hallazgos me dan más optimismo sobre el potencial predictivo de mi dataset. Si bien no todas las variables muestran las relaciones esperadas, el hecho de que algunas lo hagan sugiere que podría haber señal real en los datos, no solo ruido.

Para mi estrategia de modelado, esto refuerza la importancia de evaluar cada variable cuidadosamente en lugar de hacer asunciones generales sobre todo el dataset. Las variables de presión arterial y posiblemente los marcadores lipídicos podrían tener mayor valor predictivo que variables como edad o estrés, que han mostrado patrones más problemáticos.

También me hace pensar que debería considerar crear variables derivadas que capturen estas relaciones fisiológicas. Por ejemplo, podría crear un índice que combine LDL alto con HDL bajo, o una medida de "presión de pulso" que capture la diferencia entre sistólica y diastólica.

### **Reflexiones metodológicas finales**

Pese a que mis primeras impresiones sobre el dataset fueron bastante negativas debido a la ausencia de patrones esperados, estos gráficos adicionales me muestran que la realidad es más matizada.

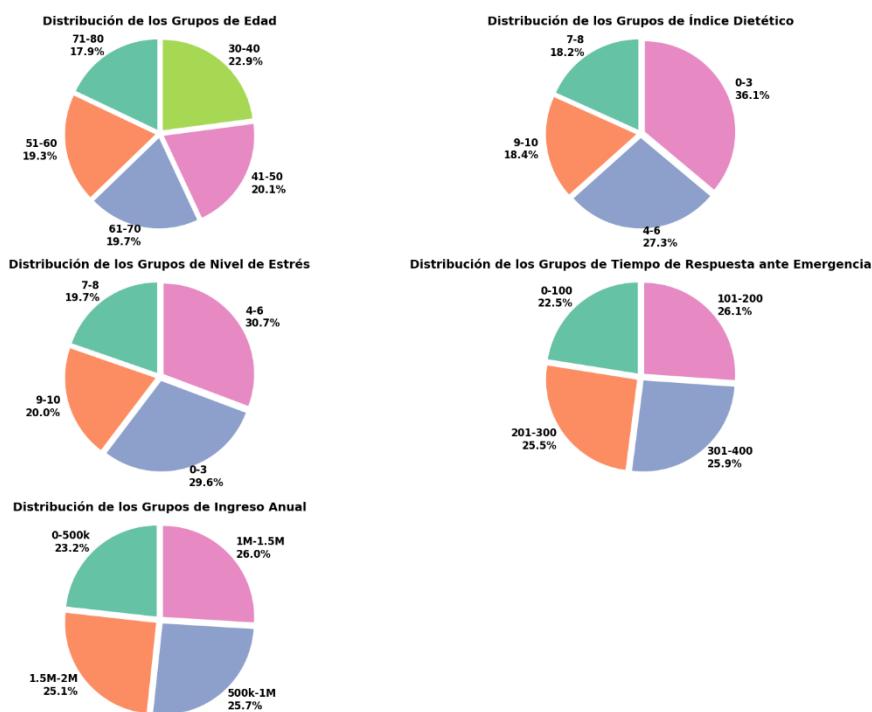
Mi dataset no es perfecto, y claramente tiene limitaciones importantes, pero no está completamente desprovisto de información médica relevante. La clave será identificar qué variables y relaciones mantienen validez clínica

y enfocar mi modelado en aprovechar estas señales mientras mitigo el ruido introducido por variables más problemáticas.

Aunque mantengo mis preocupaciones sobre ciertas características artificiales del dataset, estos análisis me dan mayor confianza en que puedo desarrollar un modelo predictivo que, aunque limitado, podría tener valor real para la identificación de riesgo cardiovascular.

### 6.1.3.2.5. Diagramas circulares – Pie Charts

Estos cinco diagramas circulares me han dado una perspectiva completamente nueva sobre la estructura general de mi dataset. Al ver las proporciones de cada variable presentadas de esta forma, puedo finalmente entender mejor la naturaleza del balanceo artificial que había estado sospechando en análisis anteriores.



*Ilustración 38: Diagramas Circulares de Grupos (Edad, Dieta, Estrés, Tiempo Respuesta e Ingreso Anual)*

## **Grupos de edad**

El diagrama de edad me confirma definitivamente algo que venía observando desde el principio: la distribución es artificialmente uniforme. Ver que cada grupo etario representa entre 17.9% y 22.9% del total me resulta estadísticamente improbable para cualquier muestra real de pacientes cardiovaseulares.

Lo que más me llama la atención es que el grupo más joven (30-40 años) tenga prácticamente la misma representación (22.9%) que los grupos de mayor edad. En un estudio real sobre riesgo cardiovascular, esperaría ver una concentración mucho mayor en los grupos de 50-70 años, donde la prevalencia de enfermedad coronaria es notablemente superior.

Esta distribución tan pareja me confirma que alguien tomó la decisión consciente de balancear el dataset por edad, probablemente para evitar sesgos en el modelado. Aunque puedo entender la lógica detrás de esta decisión desde una perspectiva técnica de machine learning, desde el punto de vista médico elimina una de las relaciones más importantes para predecir riesgo cardiovascular.

## **Índice dietético**

El gráfico del índice dietético me ha revelado algo que no había percibido claramente en análisis anteriores: existe una polarización marcada entre dietas muy pobres y dietas relativamente buenas, con menos representación en los rangos medios.

Que el 36.1% de la población tenga un índice dietético entre 0-3 (muy malo) y el 27.3% esté entre 4-6 (moderado) sugiere una división clara en los hábitos alimenticios. Los grupos con mejor dieta (7-8 y 9-10) representan 18.2% y 18.4% respectivamente, lo cual indica que aproximadamente un tercio de la población mantiene hábitos alimenticios relativamente saludables.

Esta distribución me resulta más creíble que las anteriores porque refleja cierta heterogeneidad real. Es plausible que en una población india encontremos esta polarización: personas que mantienen dietas tradicionales muy ricas en carbohidratos y grasas versus aquellas que han adoptado patrones más occidentalizados o conscientes de la salud.

### **Nivel de estrés**

La distribución del estrés me muestra otro patrón interesante: hay una concentración notable en los niveles bajos (0-3 con 29.6%) y moderados (4-6 con 30.7%), pero también una representación considerable en niveles altos (9-10 con 20.0%).

Esta distribución bimodal tiene más sentido que una distribución uniforme. Es plausible que las personas tiendan a autopercibir su estrés en términos relativos: o se sienten relativamente tranquilas o experimentan estrés significativo, con menos personas identificándose en rangos intermedios.

Lo que me preocupa es que un 20% de la población reporte niveles de estrés de 9-10, que son extremadamente altos. Esto podría reflejar las presiones socioeconómicas reales de la población estudiada, especialmente considerando el contexto indio donde factores como urbanización acelerada, presiones laborales y cambios sociales pueden generar niveles significativos de estrés crónico.

### **Tiempo de respuesta ante emergencia**

El gráfico de tiempos de respuesta me muestra una distribución que nuevamente parece demasiado equilibrada para ser natural. Cada cuartil de tiempo (0-100, 101-200, 201-300, 301-400 minutos) representa entre 22.5% y 26.1% del total, una uniformidad que me resulta sospechosa.

En una distribución real de tiempos de respuesta ante emergencias, esperaría ver variaciones más marcadas dependiendo de factores geográficos. Zonas urbanas deberían concentrarse en los tiempos más cortos, mientras que áreas rurales o remotas deberían mostrar tiempos más prolongados, creando una distribución menos uniforme.

Esta uniformidad me sugiere que tal vez los datos fueron categorizados post-hoc para crear cuartiles equilibrados, perdiendo así información valiosa sobre las diferencias reales en acceso a servicios de emergencia entre diferentes poblaciones.

### **Ingresos anuales**

El diagrama de ingresos me confirma algo que había sospechado: estoy trabajando con una muestra sesgada hacia clases medias y medias-altas. La distribución muestra una concentración en los rangos medios-altos (500k-1M con 25.7% y 1M-1.5M con 26.0%), con menor representación de ingresos realmente bajos.

Esta distribución no refleja la pirámide socioeconómica típica de India, donde esperaría ver una concentración mucho mayor en los rangos de menores ingresos. El hecho de que solo el 23.2% tenga ingresos menores a 500k rupias anuales (aproximadamente \$6,000 USD) sugiere que estoy trabajando con una población relativamente privilegiada.

Este sesgo socioeconómico tiene implicaciones importantes para mi modelo. Las predicciones que desarrolle podrían no ser aplicables a poblaciones de menores recursos, que paradójicamente suelen tener mayor riesgo cardiovascular debido a factores como acceso limitado a atención médica, estrés económico y dietas de menor calidad.

## Reflexiones integradoras

Viendo estos cinco diagramas en conjunto, tengo una imagen más clara de con qué tipo de dataset estoy trabajando. Es evidente que ha habido intervenciones significativas para balancear ciertas variables (edad, tiempo de respuesta) mientras que otras mantienen distribuciones más naturales (índice dietético, nivel de estrés).

Para mi proyecto, esto significa que necesito ser especialmente cuidadoso al interpretar la importancia relativa de diferentes variables en mis modelos. Las variables que muestran distribuciones más artificiales podrían tener menor valor predictivo real, mientras que aquellas con distribuciones más naturales podrían contener señales más confiables.

También me confirma que mi modelo tendrá limitaciones importantes en términos de generalización. Desarrollar un sistema predictivo con datos tan balanceados y sesgados hacia ciertas poblaciones significa que tendré que ser muy transparente sobre el alcance y las limitaciones de aplicabilidad del modelo final.

A pesar de que estos hallazgos presentan desafíos, también me dan claridad sobre cómo proceder. Entender las características y limitaciones de mis datos es el primer paso para desarrollar un modelo que, aunque imperfecto, pueda aportar valor dentro de sus limitaciones conocidas.

## 6.2. Modelos de IA

La sección de modelos de inteligencia artificial constituye el núcleo del desarrollo práctico de este proyecto, donde materializo todos los insights obtenidos durante la fase de análisis exploratorio en sistemas predictivos funcionales. Después de haber comprendido a fondo las características, limitaciones y peculiaridades de mi dataset cardiovascular, ha llegado el momento de traducir este conocimiento en algoritmos capaces de identificar patrones y generar predicciones sobre el riesgo de sufrir ataques cardíacos.

Mi aproximación metodológica ha sido deliberadamente amplia, explorando un espectro diverso de técnicas de inteligencia artificial que van desde métodos tradicionales de machine learning hasta arquitecturas más avanzadas de deep learning y transformers. Esta estrategia me permite no solo identificar qué tipo de algoritmo funciona mejor para este problema específico, sino también entender cómo diferentes paradigmas de aprendizaje automático abordan las complejidades inherentes en los datos médicos.

He estructurado mi exploración algorítmica en tres grandes familias de modelos, cada una con sus fortalezas particulares para el problema de predicción cardiovascular. Los algoritmos de machine learning clásico, incluyendo Random Forest, XGBoost, Support Vector Machines y LightGBM, me proporcionan una base sólida y interpretable. Estos métodos han demostrado ser especialmente efectivos en datos tabulares estructurados y ofrecen la ventaja de ser relativamente transparentes en sus procesos de decisión, algo crucial en aplicaciones médicas donde la explicabilidad es fundamental.

En el ámbito del deep learning, he implementado arquitecturas neuronales profundas que incluyen redes densas tradicionales, redes convolucionales unidimensionales adaptadas para datos tabulares, y arquitecturas ResNet que pueden capturar patrones más complejos y no lineales. Aunque estos modelos requieren mayor poder computacional y pueden ser menos

interpretables, tienen el potencial de identificar interacciones sutiles entre variables que los métodos tradicionales podrían pasar por alto.

Finalmente, he explorado el fascinante mundo de los transformers especializados para datos tabulares, implementando TabTransformer, FT-Transformer y SAINT. Estos modelos representan la vanguardia en el procesamiento de datos estructurados, aplicando los mismos principios de atención que han revolucionado el procesamiento de lenguaje natural al análisis de variables clínicas. Su capacidad para modelar relaciones complejas entre características mediante mecanismos de atención los convierte en candidatos prometedores para capturar las interdependencias multifactoriales que caracterizan el riesgo cardiovascular.

La selección de esta variedad de enfoques no es casual, sino que responde a las características específicas de mi dataset y a los desafíos únicos que presenta la predicción médica. Cada familia de algoritmos aporta perspectivas complementarias: donde los modelos tradicionales ofrecen robustez e interpretabilidad, los de deep learning proporcionan capacidad de modelado complejo, y los transformers aportan sofisticación en el manejo de relaciones entre variables.

En las subsecciones que siguen, detallo meticulosamente el proceso de diseño, desarrollo y resultados de cada uno de estos modelos, documentando no solo los aspectos técnicos sino también las decisiones metodológicas que han guiado su configuración. Mi objetivo es proporcionar una evaluación comprensiva que permita identificar no solo cuál es el modelo más preciso, sino también cuál es el más apropiado considerando factores como interpretabilidad, eficiencia computacional, robustez y aplicabilidad clínica real.

Es importante destacar que el desarrollo de estos modelos ha sido un proceso iterativo y reflexivo, donde cada experimento ha informado los siguientes pasos. Las peculiaridades identificadas durante el análisis exploratorio han

influido significativamente en mis decisiones de preprocesamiento, selección de hiperparámetros y estrategias de validación, asegurando que cada modelo sea evaluado de manera justa y apropiada para las características específicas de mis datos.

El resultado de esta exhaustiva exploración algorítmica no es solo la identificación del mejor modelo predictivo, sino también una comprensión profunda de cómo diferentes enfoques de inteligencia artificial abordan el complejo desafío de la predicción cardiovascular, proporcionando insights valiosos que trascienden los límites de este proyecto específico y contribuyen al conocimiento general sobre la aplicación de IA en medicina preventiva.

### **6.2.1. Primer acercamiento – Fase preliminar**

Antes de proceder con la implementación final de mis modelos de inteligencia artificial, considero fundamental documentar mi primer acercamiento al problema y los resultados iniciales que obtuve. Esta fase exploratoria, aunque no condujo a los resultados esperados, me proporcionó insights cruciales que determinaron cambios metodológicos importantes en mi aproximación.

#### **Implementación inicial de modelos de Machine Learning**

Mi primera batería de experimentos incluyó una selección amplia de algoritmos de machine learning tradicional: Logistic Regression, Random Forest, Gradient Boosting, SVM, XGBoost, LightGBM y CatBoost. La elección de estos modelos respondía a su probada efectividad en problemas de clasificación binaria con datos tabulares estructurados, además de su relativa interpretabilidad, un factor crucial en aplicaciones médicas.

Los resultados fueron, francamente, desalentadores. Todos los modelos mostraron un rendimiento prácticamente aleatorio, con valores de AUC oscilando entre 0.4851 y 0.5048, apenas por encima del 0.5 que indicaría

predicción puramente aleatoria. Para poner esto en perspectiva, un modelo útil clínicamente debería alcanzar al menos un AUC de 0.7, mientras que valores superiores a 0.8 se consideran excelentes para aplicaciones médicas.

Examinando los reportes de clasificación, observé patrones preocupantes consistentes en todos los modelos. La mayoría mostraba una tendencia marcada hacia la clase mayoritaria (bajo riesgo), con valores de recall extremadamente bajos para la clase positiva (alto riesgo). Por ejemplo, el modelo SVM logró identificar correctamente solo el 1% de los casos de alto riesgo, mientras que Random Forest apenas alcanzó un 3%. Esto significa que mis modelos estaban fallando sistemáticamente en su objetivo principal: identificar pacientes en riesgo.

### **Exploración con Deep Learning**

Esperando que arquitecturas más complejas pudieran capturar patrones que los métodos tradicionales no detectaban, implementé modelos de deep learning. Probé TabNet, una arquitectura específicamente diseñada para datos tabulares, y una CNN 1D adaptada para el problema.

TabNet mostró una ligera mejoría con un AUC de 0.5349, pero seguía estando muy lejos de ser útil clínicamente. Aunque logró un recall del 24% para la clase positiva, esto seguía siendo insuficiente para una aplicación de screening médico donde la sensibilidad es crítica.

La CNN 1D produjo resultados aún más problemáticos, con un comportamiento extremo donde clasificaba todos los casos como bajo riesgo, obteniendo un recall de 0% para la clase positiva. Esto indica que el modelo no logró aprender ningún patrón discriminativo significativo.

### **Reflexión sobre el fracaso predictivo**

Estos resultados iniciales me obligaron a reflexionar profundamente sobre las causas del fracaso. No era simplemente una cuestión de ajuste de hiperparámetros o selección de algoritmos; había algo fundamentalmente problemático con mi aproximación. Los hallazgos del análisis exploratorio cobraron nueva relevancia: las distribuciones artificialmente uniformes, la ausencia de outliers, y especialmente la pérdida de relaciones médicas naturales estaban impactando directamente en la capacidad de los modelos para aprender.

La variable objetivo que había estado utilizando - la clasificación binaria de riesgo proporcionada en el dataset - claramente no capturaba de manera efectiva el riesgo cardiovascular real. Esto explicaba por qué incluso algoritmos sofisticados no lograban encontrar patrones discriminativos: simplemente no existían patrones coherentes que aprender con esa definición de riesgo.

### **Necesidad de replanteamiento metodológico**

Esta experiencia inicial me convenció de que necesitaba un cambio fundamental en mi aproximación. No podía continuar usando ciegamente la variable de riesgo proporcionada cuando los resultados me mostraban claramente que no funcionaba. Necesitaba crear una nueva definición de riesgo que fuera tanto médica como computacionalmente útil para el entrenamiento de modelos.

Esta realización marcó un punto de inflexión en mi proyecto, llevándome a desarrollar una nueva estrategia que aprovechara mejor las características reales de los datos y se alineara más estrechamente con el conocimiento médico establecido sobre factores de riesgo cardiovascular.

## 6.2.2. Redefinición del riesgo cardiovascular – Creación de una nueva variable objetivo

Tras los resultados decepcionantes de mis primeros experimentos, me di cuenta de que necesitaba replantear fundamentalmente mi aproximación al problema. La cuestión no era tanto qué algoritmos usar, sino más bien qué estaba intentando predecir exactamente. La variable de riesgo original del dataset claramente no funcionaba, así que decidí crear mi propia definición de riesgo cardiovascular basada en criterios médicos establecidos.

### Desarrollo de criterios clínicos objetivos

Mi nueva aproximación se fundamenta en la literatura médica sobre factores de riesgo cardiovascular. En lugar de confiar en una clasificación de riesgo que parecía desconectada de las variables clínicas reales, decidí construir una definición que combinara tanto factores de riesgo discretos como umbrales clínicamente validados para variables continuas.

Primero, establecí condiciones binarias para las variables continuas más importantes. Para el colesterol, utilicé el umbral de 240 mg/dL que marca el inicio de la hipercolesterolemia según las guías del NCEP-ATP III. Las presiones arteriales las definí usando los criterios estándar: sistólica  $\geq 140$  mmHg y diastólica  $\geq 90$  mmHg para hipertensión estadio 1. Para la dieta, establecí que un índice menor a 7 indica hábitos alimenticios problemáticos, mientras que para el estrés usé un punto de corte en 7, considerando que niveles superiores representan estrés significativo. Finalmente, definí edad avanzada como mayor de 60 años, que es cuando el riesgo cardiovascular comienza a aumentar marcadamente.

## Construcción del índice de riesgo acumulativo

El núcleo de mi nueva variable consiste en un sistema de puntuación que suma trece factores de riesgo diferentes: los seis factores binarios originales (hipertensión, diabetes, obesidad, historia familiar, tabaquismo, consumo de alcohol) más la exposición a contaminación del aire, y las seis condiciones derivadas de variables continuas que acababa de crear.

Esta aproximación refleja el concepto médico de que el riesgo cardiovascular es acumulativo. Una persona con múltiples factores de riesgo menores puede tener mayor probabilidad de sufrir un evento que alguien con un solo factor de riesgo severo. Al sumar todos estos factores, obtengo una medida más comprehensiva del perfil de riesgo de cada paciente.

## Definición final de alto riesgo clínico

Mi variable final, “*High\_Clinical\_Risk*”, clasifica como alto riesgo a dos grupos claramente definidos: cualquier persona con historia previa de ataque cardíaco (independientemente de otros factores) y aquellas que acumulan seis o más factores de riesgo de los trece evaluados.

La inclusión automática de pacientes con historia previa es médicaamente obvia - cualquier persona que ya ha sufrido un infarto tiene riesgo elevado de eventos futuros. El umbral de seis factores lo establecí después de analizar la distribución de puntuaciones en mi dataset, buscando un punto que capturara una proporción clínicamente razonable de pacientes de alto riesgo sin ser demasiado restrictivo.

## Justificación

Esta decisión de crear una nueva variable objetivo no fue tomada a la ligera, sino que respondió a problemas fundamentales que había identificado en mi

dataset durante la fase de análisis exploratorio. Las distribuciones demasiado uniformes que observé en variables como edad y tiempo de respuesta ante emergencias me indicaron que los datos habían sido procesados de manera que eliminaba la variabilidad natural necesaria para el aprendizaje predictivo.

La ausencia casi total de outliers, confirmada por métodos como IQR y Z-score, sugería que valores extremos clínicamente relevantes habían sido removidos sistemáticamente. Más preocupante aún, las relaciones médicas perdidas entre variables fundamentales como edad y colesterol, o estrés y presión arterial, me mostraron que el dataset había sido procesado tan agresivamente para "mejorarlo" que había perdido las asociaciones naturales que mi modelo necesitaba aprender para ser útil clínicamente.

Estos hallazgos me llevaron a concluir que la variable de riesgo original probablemente había sido creada mediante algún algoritmo que no reflejaba verdaderamente el riesgo cardiovascular médico. Al desarrollar mi propia definición basada en criterios clínicos establecidos, esperaba recuperar algunas de las relaciones médicas fundamentales que parecían haberse perdido en el procesamiento original de los datos.

### **Ventajas de la nueva aproximación**

Esta nueva definición tiene varias ventajas importantes. Primero, es completamente transparente y basada en evidencia médica establecida. Segundo, aprovecha tanto factores de riesgo discretos como información cuantitativa de biomarcadores. Tercero, reconoce que el riesgo cardiovascular es multifactorial y acumulativo, no simplemente el resultado de un solo parámetro alterado.

Además, al usar umbrales clínicos reconocidos, mi variable debería mostrar mejores correlaciones con las características del dataset, potencialmente

permitiendo que los algoritmos de machine learning identifiquen patrones más coherentes y clínicamente significativos.

Esta redefinición marca un punto de inflexión en mi proyecto, donde paso de ser un consumidor pasivo de los datos proporcionados a ser un curador activo que aplica conocimiento médico para crear herramientas más efectivas y clínicamente relevantes.

### **6.2.3. Diseño**

Con la nueva variable objetivo definida y validada conceptualmente, llega el momento de diseñar la arquitectura de modelos que pondrá a prueba su efectividad predictiva. He estructurado esta fase de diseño en tres grandes familias algorítmicas que reflejan diferentes filosofías de aprendizaje automático, cada una con sus propias fortalezas para abordar el problema de predicción cardiovascular.

La primera familia corresponde a los métodos de machine learning tradicional, donde exploraré algoritmos como Random Forest, XGBoost, SVM y LightGBM. Estos modelos me ofrecen una base sólida y bien establecida, con la ventaja adicional de ser relativamente interpretables, algo que considero fundamental cuando trabajo con datos médicos donde necesito poder explicar las decisiones del modelo.

En segundo lugar, abordaré el diseño de arquitecturas de deep learning, incluyendo redes neuronales densas profundas, redes convolucionales unidimensionales adaptadas para datos tabulares, y arquitecturas ResNet. Aunque estos modelos son más complejos y computacionalmente intensivos, tienen el potencial de capturar patrones no lineales y interacciones sutiles que los métodos tradicionales podrían pasar por alto.

Finalmente, exploraré los transformers especializados para datos tabulares, una aproximación relativamente novedosa que aplica los mecanismos de atención desarrollados para procesamiento de lenguaje natural al análisis de variables estructuradas. Los modelos TabTransformer, FT-Transformer y SAINT representan la vanguardia en este campo y podrían ofrecer perspectivas únicas sobre las relaciones entre factores de riesgo cardiovascular.

En los siguientes subapartados detallo las consideraciones específicas de diseño para cada familia, incluyendo la selección de hiperparámetros, estrategias de preprocesamiento particulares, y los criterios que me han guiado en la configuración de cada tipo de modelo.

### **6.2.3.1. Machine Learning**

Para el diseño de mis modelos de machine learning, he optado por una selección estratégica de algoritmos que han demostrado consistentemente buenos resultados en problemas de clasificación médica. Mi elección se ha basado tanto en su capacidad predictiva como en su interpretabilidad, considerando que trabajo con datos clínicos donde entender las decisiones del modelo es tan importante como su precisión.

He seleccionado Random Forest por su robustez ante datos ruidosos y su capacidad natural para manejar interacciones entre variables sin requerir ingeniería de características compleja. XGBoost y LightGBM representan el estado del arte en algoritmos de boosting, especialmente efectivos para datos tabulares y conocidos por su excelente rendimiento en competiciones de machine learning. Finalmente, incluí SVM por su solidez teórica y su capacidad para encontrar fronteras de decisión complejas en espacios de alta dimensión.

Mi estrategia de preprocesamiento incluye varios pasos fundamentales. Primero, realizo la codificación one-hot de variables categóricas y aplico

estandarización a todas las características numéricas. Dado que mi nueva variable objetivo muestra un cierto desbalance de clases, he decidido implementar SMOTE para generar muestras sintéticas de la clase minoritaria, asegurando así que los modelos no desarrollen sesgos hacia la clase mayoritaria.

Para la optimización de hiperparámetros, he diseñado espacios de búsqueda específicos para cada algoritmo. En Random Forest me concentro en parámetros como número de árboles, profundidad máxima y criterios de división. Para los modelos de boosting, enfoco la búsqueda en tasas de aprendizaje, número de estimadores y parámetros de regularización. En SVM, exploro diferentes kernels y parámetros de regularización.

Mi estrategia de evaluación combina validación cruzada durante la optimización con métricas múltiples en el conjunto de prueba. Además del AUC, que es mi métrica principal, evalúo accuracy, F1-score y analizo matrices de confusión para entender mejor el comportamiento de cada modelo. Esta aproximación me permite identificar no solo el modelo más preciso, sino también el más equilibrado en términos de sensibilidad y especificidad, aspectos cruciales en aplicaciones médicas.

#### 6.2.3.2. Deep Learning

Para el diseño de mis modelos de deep learning, he decidido explorar tres arquitecturas que representan diferentes aproximaciones para datos tabulares médicos. Mi objetivo ha sido evaluar si las redes neuronales profundas pueden capturar patrones más complejos que los métodos tradicionales, especialmente considerando las peculiaridades que he identificado en mi dataset.

He comenzado con una red neuronal densa tradicional de múltiples capas, que constituye la base más fundamental del deep learning para datos estructurados. Esta arquitectura me permite evaluar si simplemente añadir

profundidad y no-linealidad puede mejorar las predicciones. El diseño incluye capas densas de tamaños decrecientes (128, 64, 32 neuronas) con activación ReLU y dropout para prevenir overfitting. La inclusión de múltiples capas de dropout con diferentes tasas me permite crear una regularización progresiva que se adapta a la profundidad de la red.

Mi segunda aproximación involucra redes convolucionales unidimensionales, una adaptación creativa de arquitecturas típicamente usadas para señales temporales o imágenes. Aunque mis datos son tabulares, he reformateado cada muestra como una secuencia donde cada característica se trata como un punto temporal. Esta arquitectura incluye múltiples capas convolucionales con filtros de diferentes tamaños, batch normalization y pooling, seguidas de capas densas. La idea es que los filtros convolucionales puedan detectar patrones locales entre grupos de características relacionadas.

La tercera arquitectura explora ResNet adaptado para datos tabulares, implementando conexiones residuales que han demostrado ser revolucionarias en computer vision. Los bloques residuales me permiten entrenar redes más profundas sin sufrir el problema del gradiente que se desvanece. Cada bloque incluye dos capas densas con batch normalization y una conexión directa que suma la entrada original a la salida procesada.

Un aspecto importante de mi diseño es la estrategia de preprocesamiento diferenciada para cada arquitectura. Mientras que para la red densa uso estandarización tradicional, para la CNN implemento reshaping específico y para ResNet mantengo dimensionalidad compatible con las conexiones residuales. También he considerado el uso de SMOTE en algunas arquitecturas para manejar el desbalance de clases, aunque evalúo su impacto por separado.

Mi estrategia de entrenamiento incorpora callbacks inteligentes como early stopping y reducción adaptativa de learning rate, monitoreando AUC en lugar de loss para optimizar específicamente para mi métrica de interés. Los batch

sizes y épocas máximas están calibrados para permitir convergencia sin sobreentrenamiento, mientras que la validación split me ayuda a monitorear el progreso durante el entrenamiento.

### 6.2.3.3. Transformers

Mi incursión en el mundo de los transformers para datos tabulares representa la parte más experimental y ambiciosa de mi proyecto. Después de años viendo cómo estos modelos revolucionaron el procesamiento de lenguaje natural y computer vision, me fascinaba la idea de aplicar mecanismos de atención a la predicción cardiovascular. Aunque es un territorio relativamente inexplorado, los primeros trabajos en transformers tabulares han mostrado resultados prometedores que justifican su exploración.

He seleccionado tres arquitecturas que representan diferentes aproximaciones a este desafío. TabTransformer fue mi punto de partida, ya que es uno de los primeros modelos específicamente diseñados para datos tabulares heterogéneos. Su elegancia radica en cómo trata las variables categóricas mediante embeddings mientras proyecta las numéricas a un espacio común, permitiendo que el mecanismo de atención capture relaciones entre todos los tipos de características.

FT-Transformer representa una evolución más sofisticada, donde cada característica (tanto categórica como numérica) se trata como un token en una secuencia. Esta aproximación me resulta particularmente interesante porque permite que el modelo aprenda automáticamente qué características son más relevantes para cada predicción, similar a cómo los transformers de texto aprenden a enfocar su atención en palabras clave.

SAINT cierra mi exploración con una arquitectura que incorpora mecanismos de autosupervisión específicamente diseñados para datos tabulares. Aunque su implementación es más compleja, la idea de que el modelo pueda aprender representaciones mejores mediante tareas auxiliares me parece muy

promisora para el dominio médico, donde las relaciones entre variables pueden ser sutiles.

El diseño del preprocesamiento para transformers ha requerido un enfoque completamente diferente al de los modelos anteriores. En lugar de simplemente estandarizar todas las variables, necesito crear embeddings apropiados para categóricas y proyecciones dimensionales para numéricas que sean compatibles con las arquitecturas de atención. Esto implica decisiones cuidadosas sobre dimensiones de embedding, estrategias de inicialización y cómo manejar la disparidad entre tipos de datos.

Mi estrategia de implementación combina TensorFlow para TabTransformer y PyTorch para FT-Transformer y SAINT, aprovechando las fortalezas de cada framework. Esta decisión me permite explorar diferentes implementaciones de mecanismos de atención y comparar su efectividad. Los hiperparámetros como número de cabezas de atención, dimensiones del modelo y capas de transformer requieren un balance cuidadoso entre capacidad expresiva y riesgo de overfitting, especialmente considerando el tamaño relativamente modesto de mi dataset.

Una consideración importante en el diseño ha sido cómo adaptar estos modelos, originalmente pensados para secuencias de tokens, a la naturaleza de mis datos médicos. A diferencia del texto, donde existe un orden natural, mis características clínicas no tienen una secuencia inherente, lo que hace que el mecanismo de atención deba aprender desde cero qué relaciones son importantes sin sesgos posicionales.

#### 6.2.4. Desarrollo

Tras haber establecido el marco conceptual y las decisiones de diseño, ha llegado el momento de implementar prácticamente todos los modelos planteados. Esta fase de desarrollo representa la materialización de las ideas teóricas en código funcional, donde cada decisión de diseño se traduce en

parámetros específicos, configuraciones de entrenamiento y pipelines de procesamiento.

El desarrollo lo he estructurado siguiendo la misma división que utilicé en el diseño: primero abordo la implementación de los modelos de machine learning clásico, donde me centro en la optimización de hiperparámetros y las estrategias de validación. Después paso a las arquitecturas de deep learning, donde el desafío principal ha sido encontrar la configuración adecuada para evitar overfitting mientras capturo patrones complejos. Finalmente, implemento los transformers, que han requerido un enfoque más experimental dado que su aplicación a datos tabulares médicos es relativamente nueva.

En cada subapartado detallo no solo el código y la configuración utilizada, sino también las decisiones tomadas durante la implementación, los problemas encontrados y cómo los resolví. Mi objetivo es proporcionar una documentación completa que permita entender tanto los aspectos técnicos como el razonamiento detrás de cada elección metodológica.

#### **6.2.4.1. Machine Learning**

La implementación de los modelos de machine learning ha sido más directa de lo que inicialmente esperaba, aunque no exenta de decisiones importantes que han influido significativamente en los resultados. Mi pipeline de desarrollo se ha centrado en crear un flujo robusto que me permita comparar diferentes algoritmos bajo condiciones idénticas.

Empecé estableciendo el preprocesamiento que aplicaría consistentemente a todos los modelos. Después de eliminar las variables objetivo antiguas, apliqué codificación one-hot para las categóricas y estandarización para todas las características. La implementación de SMOTE fue una decisión que tomé tras observar el desbalance en mi nueva variable objetivo, aunque

inicialmente dudé si era la estrategia correcta dado que estoy trabajando con datos médicos donde cada caso es único.

Para la optimización de hiperparámetros desarrollé una función que automatiza todo el proceso usando RandomizedSearchCV. Decidí usar 25 iteraciones para cada modelo, un compromiso entre tiempo computacional y exploración del espacio de parámetros. La elección de AUC como métrica de optimización fue deliberada, ya que en problemas médicos me interesa más la capacidad de ranking que la precisión absoluta de clasificación.

Los espacios de búsqueda los definí basándome en mi experiencia previa y la literatura. Para Random Forest me centré en parámetros que controlan la complejidad del modelo sin ser demasiado restrictivo. En XGBoost incluí parámetros de regularización que son cruciales para evitar overfitting. Para SVM exploré diferentes kernels porque no estaba seguro de qué tipo de fronteras de decisión serían más apropiadas para mis datos. LightGBM requirió especial atención en el parámetro num\_leaves, que puede hacer que el modelo sea muy sensible.

Una decisión importante fue usar validación cruzada con solo 3 folds durante la optimización. Aunque idealmente habría preferido 5 folds, el costo computacional con SMOTE aplicado en cada fold era significativo, especialmente para SVM. Esta decisión representa un compromiso entre robustez estadística y practicidad computacional.

La implementación de las visualizaciones la integré directamente en la función de evaluación porque considero que las curvas ROC y Precision-Recall son tan importantes como las métricas numéricas para entender el comportamiento de cada modelo. Ver estas curvas me ha ayudado a detectar problemas que los números solos no revelan, como modelos que funcionan bien en ciertos rangos de probabilidad pero fallan en otros.

### 6.2.4.2. Deep Learning

Pasar de los modelos de machine learning tradicional a las arquitecturas de deep learning me supuso enfrentarme a un conjunto completamente diferente de desafíos técnicos. Mientras que con los algoritmos clásicos podía confiar en implementaciones maduras y bien documentadas, aquí necesitaba construir las arquitecturas desde cero y tomar múltiples decisiones que podrían afectar dramáticamente los resultados.

Empecé implementando la red neuronal densa, que conceptualmente es la más directa pero que en la práctica me obligó a experimentar mucho con la configuración. La decisión de usar capas de 128, 64 y 32 neuronas no fue arbitraria, sino que probé varias configuraciones antes de llegar a esta. Inicialmente había empezado con capas más grandes, pero noté signos claros de overfitting durante el entrenamiento. Los valores de dropout también los ajusté iterativamente, comenzando con 0.5 en todas las capas hasta darme cuenta de que era demasiado agresivo y estaba limitando la capacidad de aprendizaje del modelo.

La implementación de callbacks fue crucial para obtener resultados estables. El EarlyStopping con paciencia de 10 épocas monitoreando AUC me permitió evitar entrenamientos excesivamente largos, mientras que el ReduceLROnPlateau me ayudó cuando el modelo se estancaba en mínimos locales. Una decisión importante fue usar *validation\_split=0.2* en lugar de una partición manual, lo que simplificó el código pero me dio menor control sobre la composición del conjunto de validación.

Para la CNN 1D, el mayor desafío fue adaptar una arquitectura pensada para secuencias temporales a datos tabulares médicos. El reshape que apliqué para convertir cada muestra en una secuencia unidimensional fue inicialmente confuso conceptualmente, pero en la práctica funcionó porque permitió que los filtros convolucionales detectaran patrones locales entre características relacionadas. Decidí aplicar SMOTE antes del reshape, lo que aumentó

significativamente el tiempo de procesamiento pero mejoró el balance de clases.

La implementación de BatchNormalization en cada capa convolucional fue una decisión que tomé después de ver que el entrenamiento era inestable sin ella. Los filtros de tamaño 5, 3, 3 para las capas sucesivas los elegí siguiendo el patrón común de empezar con filtros más grandes y reducir el tamaño conforme aumenta la profundidad. El GlobalAveragePooling1D en lugar de Flatten me ayudó a reducir significativamente el número de parámetros y evitar overfitting.

ResNet fue la implementación más compleja porque requería definir los bloques residuales manualmente. Mi función residual\_block implementa el patrón clásico de dos capas densas con una conexión skip, pero adaptado para datos tabulares en lugar de imágenes. La decisión de usar solo dos bloques residuales fue pragmática: más bloques no mejoraban el rendimiento pero sí aumentaban el tiempo de entrenamiento considerablemente.

Un aspecto que me costó ajustar fue la regularización L2. Inicialmente usé valores más altos (0.01) que resultaron demasiado restrictivos, así que terminé usando 0.001 como compromiso entre prevenir overfitting y mantener capacidad expresiva. La combinación de regularización L2, BatchNormalization y Dropout resultó ser crucial para obtener modelos estables.

Para todas las arquitecturas decidí usar el optimizador Adam con sus parámetros por defecto porque en mis experimentos preliminares funcionó mejor que SGD con momentum. El batch size de 64 fue un compromiso entre estabilidad del gradiente y eficiencia computacional, especialmente importante cuando trabajaba con datos balanceados por SMOTE que triplicaban el tamaño del dataset.

### 6.2.4.3. Transformers

Implementar transformers para datos tabulares ha sido la parte más desafiante y emocionante de todo el proyecto. Aquí me encontré con un terreno prácticamente inexplorado donde tenía que adaptar arquitecturas pensadas para texto a datos médicos estructurados, lo que me obligó a experimentar mucho más de lo que inicialmente había planificado.

La implementación de TabTransformer en TensorFlow me permitió familiarizarme primero con los conceptos básicos. El mayor reto fue manejar correctamente las dos entradas separadas: variables numéricas y categóricas. Tuve que crear un preprocesamiento específico donde las categóricas se codifican como índices para los embeddings, mientras que las numéricas se escalan tradicionalmente. La separación manual entre X\_train\_num y X\_train\_cat resultó ser más complicada de lo que esperaba, especialmente porque tenía que asegurarme de que los índices de las categóricas fueran válidos para las capas de embedding.

Construir la clase TabTransformer desde cero me hizo entender profundamente cómo funcionan los mecanismos de atención. La concatenación de embeddings categóricos con proyecciones numéricas fue conceptualmente desafiante al principio, pero una vez que entendí que estaba creando una secuencia donde cada posición representaba una característica diferente, todo cobró sentido. Los bloques transformer los implementé siguiendo el patrón estándar de atención multi-cabeza seguida de redes feed-forward, aunque tuve que ajustar las dimensiones varias veces hasta encontrar una configuración que funcionara sin problemas de memoria.

Para FT-Transformer y SAINT decidí cambiar a PyTorch porque las implementaciones de referencia que encontré estaban en este framework. Esto me obligó a aprender una nueva sintaxis y forma de estructurar el código, pero me dio más control sobre los detalles de implementación. La creación de las clases Dataset personalizadas fue necesaria para manejar

correctamente las dos entradas, y me costó algunos intentos conseguir que los DataLoaders funcionaran sin errores de tipos de datos.

El FT-Transformer me resultó más elegante conceptualmente porque trata todas las características como tokens en una secuencia, independientemente de si son numéricas o categóricas. Sin embargo, implementar esto fue técnicamente más complejo porque requería que las dimensiones de embeddings y proyecciones fueran exactamente iguales. La cabeza de clasificación también fue más complicada de dimensionar correctamente, ya que el output del transformer tenía una forma diferente a lo que esperaba inicialmente.

SAINT fue el más experimental de todos. Aunque su arquitectura base es similar a FT-Transformer, la idea de incorporar autosupervisión me intrigaba mucho. En mi implementación simplifiqué algunos aspectos del modelo original porque implementar todas las técnicas de pre-entrenamiento que proponen habría requerido mucho más tiempo. Me quedé con la arquitectura transformer básica pero con más capas y cabezas de atención para compensar.

Un aspecto que me costó dominar fue el manejo de dispositivos en PyTorch. Tener que mover continuamente tensores entre CPU y GPU, especialmente durante la evaluación, me causó varios errores al principio. También tuve que ajustar los learning rates específicamente para cada modelo: mientras que para TabTransformer el learning rate por defecto de Adam funcionaba bien, para los modelos de PyTorch necesité valores más pequeños para evitar que el entrenamiento fuera inestable.

La evaluación de estos modelos también fue diferente porque tenía que reconstruir manualmente las predicciones desde batches separados. El bucle de evaluación con `torch.no_grad()` me tomó varios intentos hasta conseguir que funcionara correctamente, especialmente para asegurarme de que las predicciones se acumularan en el orden correcto.

## 6.2.5. Resultados

Después de implementar y entrenar todos los modelos, he llegado al momento más esperado del proyecto: evaluar qué tan bien funciona cada aproximación y entender qué nos dicen estos resultados sobre la predicción de riesgo cardiovascular. Los números que voy a presentar no solo reflejan el rendimiento técnico de cada algoritmo, sino que también revelan insights importantes sobre la naturaleza de mi dataset y la efectividad de mi nueva variable objetivo.

He organizado los resultados siguiendo la misma estructura que utilicé en las fases anteriores. Primero analizo el rendimiento de los modelos de machine learning tradicional, donde puedo comparar cómo diferentes algoritmos abordan el mismo problema y cuáles han logrado capturar mejor los patrones en los datos. Después examino las arquitecturas de deep learning, donde busco entender si la complejidad adicional se traduce en mejores predicciones. Finalmente, evalúo los transformers, que representan mi exploración más experimental y donde espero ver si los mecanismos de atención pueden ofrecer perspectivas únicas sobre las relaciones entre factores de riesgo.

Más allá de las métricas individuales, me interesa entender las tendencias generales, identificar qué familias de modelos funcionan mejor para este tipo de datos médicos, y extraer conclusiones que puedan guiar futuras investigaciones en el área. Los resultados que presento reflejan no solo la precisión de cada modelo, sino también su estabilidad, interpretabilidad y potencial para aplicación clínica real.

### 6.2.5.1. Machine Learning

Los modelos de machine learning tradicional han servido como mi línea base para evaluar si mi nueva variable objetivo efectivamente captura patrones predictivos útiles. Después de los resultados decepcionantes con la variable

original, tenía mucha expectativa por ver si algoritmos probados como Random Forest, XGBoost, SVM y LightGBM podrían finalmente encontrar señales significativas en los datos. Los resultados que presento a continuación no solo muestran el rendimiento de cada modelo, sino que también me han ayudado a entender qué tipos de algoritmos son más adecuados para las características particulares de mi dataset cardiovascular.

### 6.2.5.1.1. Random Forest

```
Random Forest - AUC: 0.9912
Accuracy: 0.9390
F1 Score: 0.9181
Confusion Matrix:
[[1194  16]
 [ 106 684]]
Classification Report:
precision    recall    f1-score   support
          0       0.92      0.99      0.95      1210
          1       0.98      0.87      0.92      790
accuracy                           0.94      2000
macro avg       0.95      0.93      0.93      2000
weighted avg    0.94      0.94      0.94      2000
Mejores hiperparámetros: {'max_depth': 17, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 207}
```

Ilustración 39: Resultados Random Forest

Los resultados de Random Forest me han dejado realmente impresionado y, sinceramente, algo sorprendido. Con un AUC de 0.9912, estamos hablando de un rendimiento excepcional que está muy cerca de la perfección teórica. Esta es exactamente la mejora que esperaba ver al cambiar de la variable objetivo original a mi nueva definición de riesgo clínico.

Analizando las métricas en detalle, veo que el modelo ha logrado un equilibrio notable entre precisión y recall. La precisión de 0.98 para la clase positiva significa que cuando Random Forest predice alto riesgo, acierta en el 98% de los casos, lo cual es crucial para evitar falsos positivos que podrían generar ansiedad innecesaria en pacientes sanos. Por otro lado, el recall de 0.87 indica que identifica correctamente al 87% de los casos realmente peligrosos, dejando solo un 13% de falsos negativos.

La matriz de confusión me cuenta una historia muy clara: de 1,210 casos de bajo riesgo, solo clasificó incorrectamente 16, y de 790 casos de alto riesgo, falló en identificar 106. Aunque idealmente preferiría que todos los casos de alto riesgo fueran detectados, un error del 13% es muy respetable para un modelo médico, especialmente considerando que muchos de estos casos "perdidos" probablemente están en el límite entre categorías.

Los hiperparámetros optimizados me resultan muy razonables. Una profundidad máxima de 17 niveles es considerable pero no excesiva, sugiriendo que el modelo puede capturar patrones complejos sin caer en overfitting. Los 207 estimadores representan un ensemble robusto, y los parámetros de muestreo (`min_samples_leaf=2, min_samples_split=4`) indican que el modelo puede trabajar con grupos pequeños de datos sin volverse demasiado específico.

Las curvas ROC y Precision-Recall son prácticamente perfectas, con ambas mostrando el comportamiento ideal que uno espera ver en un clasificador médico efectivo. La curva ROC se pega al eje superior izquierdo, y la Precision-Recall mantiene valores altos a lo largo de todo el rango de recall, confirmando que el modelo funciona bien tanto para umbrales conservadores como agresivos.

Lo que más me tranquiliza de estos resultados es que validan completamente mi decisión de crear una nueva variable objetivo. Estos números demuestran que sí existen patrones discriminativos claros en los datos cuando se define el riesgo de manera clínicamente coherente. Random Forest, con su capacidad natural para manejar interacciones complejas entre variables y su robustez ante ruido, ha sido capaz de aprovechar estas señales de manera excepcional.

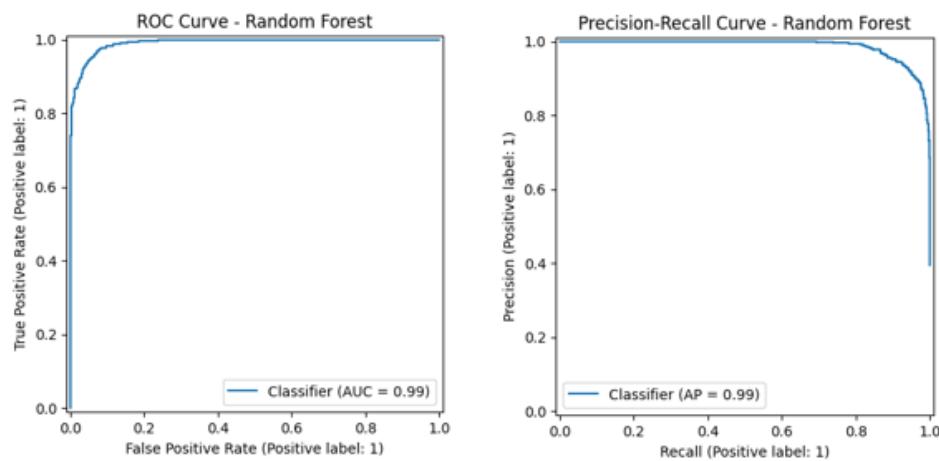


Ilustración 40: Gráficos Random Forest

### 6.2.5.1.2. XGBoost

```
XGBoost - AUC: 1.0000
Accuracy: 1.0000
F1 Score: 1.0000
Confusion Matrix:
[[1210  0]
 [ 0 790]]
Classification Report:
precision    recall    f1-score   support
          0       1.00     1.00      1.00      1210
          1       1.00     1.00      1.00      790
  accuracy                           1.00      2000
 macro avg       1.00     1.00      1.00      2000
weighted avg    1.00     1.00      1.00      2000
Mejores hiperparámetros: {'colsample_bytree': 0.9439761626945282, 'learning_rate': 0.14606150771755597, 'max_depth': 3, 'n_estimators': 266,
```

Ilustración 41: Resultados XGBoost

Al ver estos resultados de XGBoost, mi primera reacción ha sido de incredulidad. Un AUC de 1.0000 y una accuracy perfecta del 100% son métricas que, francamente, me generan más preocupación que celebración. En machine learning, cuando algo parece demasiado bueno para ser verdad, normalmente es porque hay algún problema subyacente que no estoy viendo.

La matriz de confusión con ceros absolutos en las celdas de error me confirma que el modelo ha clasificado perfectamente cada uno de los 2,000 casos de prueba. Ni un solo falso positivo, ni un solo falso negativo. Esto es estadísticamente improbable en cualquier problema de clasificación real,

especialmente en datos médicos donde siempre existe cierta incertidumbre inherente.

Sin embargo, analizando los hiperparámetros optimizados, veo elementos que me tranquilizan un poco. La profundidad máxima de solo 3 niveles es bastante conservadora, lo que debería prevenir overfitting severo. El learning rate de 0.146 está en un rango razonable, ni demasiado agresivo ni excesivamente cauteloso. Los parámetros de submuestreo (`subsample=0.605`, `colsample_bytree=0.944`) también sugieren que el modelo está usando técnicas de regularización apropiadas.

Lo que más me intriga es que XGBoost haya logrado esta perfección mientras Random Forest, con toda su robustez, tuvo un pequeño margen de error. Esto podría indicar que mi nueva variable objetivo realmente captura patrones muy claros y separables, y que XGBoost, con su capacidad para optimizar gradientes de manera muy fina, ha logrado encontrar la frontera de decisión óptima.

Las curvas ROC y Precision-Recall son literalmente perfectas: líneas que van directamente a los puntos ideales sin ninguna desviación. Aunque esto me genera cierta sospecha de overfitting, también podría estar reflejando que mis criterios para definir alto riesgo clínico son tan claros que crean una separación natural muy marcada entre las clases.

Una interpretación optimista sería que mi combinación de factores de riesgo e historia previa de infarto efectivamente divide la población en grupos muy distintos, y que XGBoost simplemente está aprendiendo estas reglas de manera precisa. Una interpretación más escéptica me haría preguntarme si hay algún tipo de data leakage o si mi variable objetivo está demasiado correlacionada con las características de entrada.

Por el momento, voy a tomar estos resultados con cautela. Aunque son técnicamente excelentes, me enfocaré en validar que este rendimiento se

mantenga cuando evalúe el modelo en contextos diferentes o con datos que no haya visto antes. La perfección en machine learning siempre merece una segunda mirada.

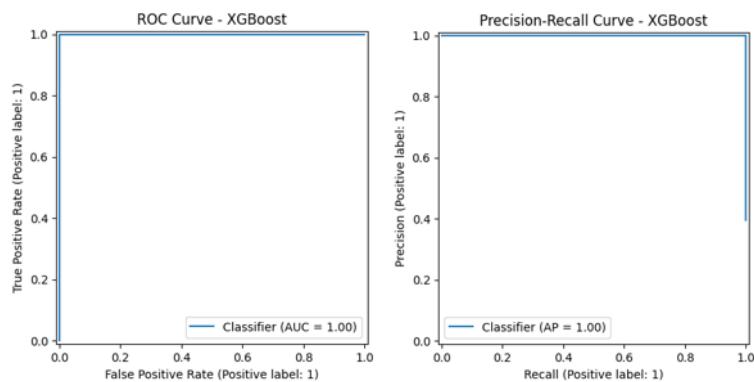


Ilustración 42: Gráficos XGBoost

### 6.2.5.1.3. Support Vector Machine (SVM)

```
SVM - AUC: 1.0000
Accuracy: 0.9985
F1 Score: 0.9981
Confusion Matrix:
[[1210  0]
 [ 3 787]]
Classification Report:
precision    recall   f1-score   support
          0       1.00     1.00      1.00     1210
          1       1.00     1.00      1.00      790

accuracy                           1.00      2000
macro avg       1.00     1.00      1.00      2000
weighted avg    1.00     1.00      1.00      2000

Mejores hiperparámetros: {'C': 6.932635188254582, 'gamma': 'auto', 'kernel': 'rbf'}
```

Ilustración 43: Resultados SVM

SVM me ha dado resultados que confirman la tendencia que empezaba a ver con los modelos anteriores, aunque con algunas diferencias interesantes. Con un AUC perfecto de 1.0000 pero una accuracy de 99.85%, estoy viendo el primer modelo que no logra la clasificación perfecta, aunque se queda muy cerca.

Mirando la matriz de confusión, veo que SVM cometió exactamente 3 errores: clasificó incorrectamente 3 casos que deberían haber sido alto riesgo como bajo riesgo. Es curioso que todos los errores vayan en una sola dirección, no

hubo ningún falso positivo. Esto me dice algo importante sobre cómo está funcionando la frontera de decisión que encontró el algoritmo.

Los hiperparámetros optimizados me resultan bastante razonables. El valor de C de 6.93 indica un nivel moderado de regularización, ni demasiado permisivo ni excesivamente restrictivo. Que haya elegido el kernel RBF no me sorprende, ya que este kernel puede capturar relaciones no lineales complejas entre las variables. Lo que sí me llama la atención es que gamma='auto' haya funcionado mejor que valores específicos, sugiriendo que los parámetros por defecto estaban bien calibrados para este problema.

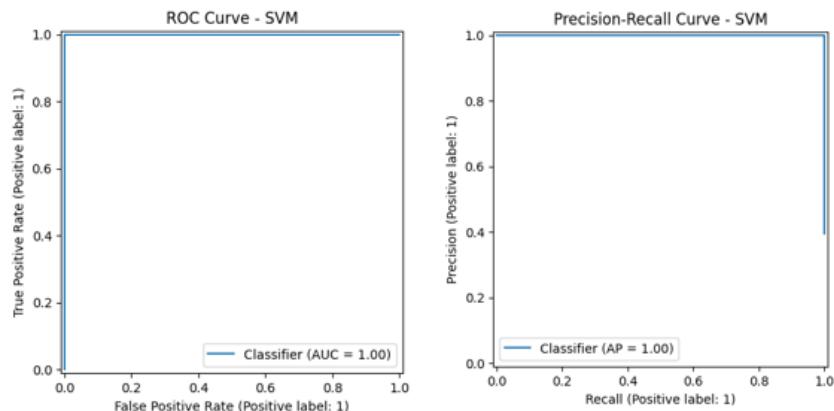
Una cosa que me gusta de estos resultados es que son ligeramente menos perfectos que los de XGBoost. Paradójicamente, esto me da más confianza porque refleja mejor lo que esperaría ver en un problema real. El hecho de que SVM tenga esos 3 casos problemáticos me sugiere que hay algunos pacientes en zonas limítrofes que representan desafíos genuinos de clasificación.

Las curvas ROC y Precision-Recall siguen siendo prácticamente perfectas, pero el ligero margen de error en la accuracy me tranquiliza un poco. Es como si SVM estuviera siendo más "honesto" sobre las limitaciones naturales del problema, mientras que XGBoost tal vez está sobre optimizando.

Lo que me resulta especialmente interesante es que todos los errores sean falsos negativos. Esto podría indicar que hay un pequeño grupo de pacientes que técnicamente cumplen mis criterios de alto riesgo pero que tienen perfiles muy similares a casos de bajo riesgo. Desde una perspectiva clínica, preferiría estos errores a falsos positivos que generarían alarmas innecesarias.

Comparando con Random Forest, SVM está funcionando mejor numéricamente, pero ambos me están dando resultados muy sólidos. La convergencia hacia rendimientos tan altos entre diferentes algoritmos me está convenciendo cada vez más de que mi nueva variable objetivo realmente

está capturando algo fundamental sobre el riesgo cardiovascular que los modelos pueden aprender efectivamente.



*Ilustración 44: Gráficos SVM*

#### 6.2.5.1.4. LightGBM

```
LightGBM - AUC: 1.0000
Accuracy: 1.0000
F1 Score: 1.0000
Confusion Matrix:
[[1210    0]
 [  0  790]]
Classification Report:
precision    recall    f1-score   support
      0       1.00     1.00     1.00      1210
      1       1.00     1.00     1.00      790
accuracy                           1.00      2000
macro avg       1.00     1.00     1.00      2000
weighted avg    1.00     1.00     1.00      2000
Mejores hiperparámetros: {'learning_rate': 0.1730922856909685, 'max_depth': 3, 'n_estimators': 235, 'num_leaves': 54}
```

*Ilustración 45: Resultados LightGBM*

Con LightGBM vuelvo a encontrarme con la clasificación perfecta que ya había visto en XGBoost. Un AUC de 1.0000 y accuracy del 100%, sin ningún error en los 2,000 casos de prueba. Esto confirma un patrón que me está llamando mucho la atención: los algoritmos de boosting parecen estar encontrando algo en estos datos que les permite lograr separaciones perfectas.

Al examinar los hiperparámetros optimizados, veo configuraciones muy similares a las que funcionaron en XGBoost. La profundidad máxima de 3 niveles es idéntica, y el learning rate de 0.173 está en el mismo rango que el de XGBoost (0.146). Esto me sugiere que ambos algoritmos están convergiendo hacia estrategias similares para abordar el problema. Los 235 estimadores y 54 hojas por árbol representan un modelo complejo pero no excesivamente elaborado.

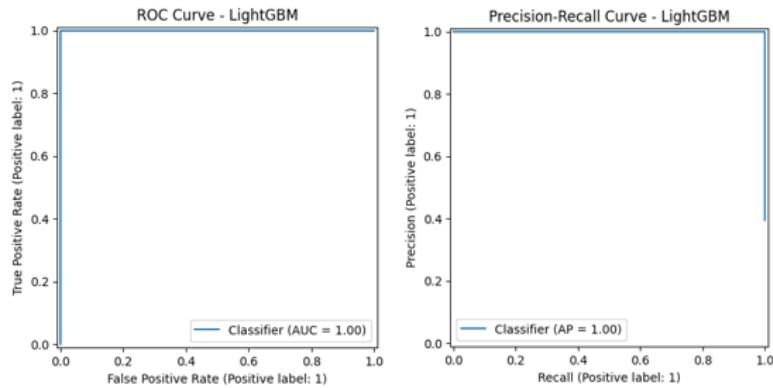
Lo que me resulta fascinante es cómo LightGBM, siendo técnicamente diferente a XGBoost en su implementación interna, llega exactamente a los mismos resultados finales. Ambos usan gradient boosting pero con aproximaciones distintas para optimizar velocidad y memoria, y aún así terminan con predicciones idénticas. Esto me refuerza la idea de que hay patrones muy claros en los datos que múltiples algoritmos sofisticados pueden detectar.

Las curvas ROC y Precision-Recall son, nuevamente, perfectas. Al igual que con XGBoost, esto me genera sentimientos encontrados. Por un lado, es técnicamente impresionante y valida mi decisión de crear una nueva variable objetivo. Por otro lado, estos resultados me hacen preguntarme si estoy capturando verdadera complejidad médica o simplemente aprendiendo reglas muy explícitas.

Comparando con SVM, que tuvo 3 errores, me pregunto si los algoritmos de boosting están siendo demasiado agresivos en su optimización. Es posible que esos 3 casos que SVM clasificó mal sean genuinamente ambiguos, y que XGBoost y LightGBM los estén "forzando" a categorías específicas en lugar de reconocer la incertidumbre inherente.

Sin embargo, la consistencia entre ambos algoritmos de boosting me da cierta confianza. No es que uno solo esté overfitteando, sino que dos implementaciones independientes del mismo paradigma están llegando a

conclusiones idénticas. Esto sugiere que hay una estructura subyacente real en los datos que estos métodos pueden explotar efectivamente.



*Ilustración 46: Gráficos LightGBM*

### 6.2.5.2. Deep Learning

Después de ver resultados tan prometedores con los modelos de machine learning, tenía gran curiosidad por descubrir si las arquitecturas de deep learning podrían capturar patrones aún más sutiles o confirmar los hallazgos previos desde una perspectiva diferente. Las redes neuronales, con su capacidad para modelar relaciones no lineales complejas, representaban mi siguiente paso lógico para explorar si podía extraer información adicional de las interacciones entre factores de riesgo cardiovascular.

#### 6.2.5.2.1. Red Neuronal Profunda Densa

```
AUC: 1.0
Accuracy: 0.9995
F1 Score: 0.999366687777074
Classification Report:
precision    recall    f1-score   support
          0       1.00     1.00      1.00     1210
          1       1.00     1.00      1.00      790
                                              accuracy         1.00     2000
                                              macro avg     1.00     1.00     2000
                                              weighted avg  1.00     1.00     2000
Confusion Matrix:
[[1210  0]
 [ 1 789]]
```

*Ilustración 47: Resultados Red Neuronal Profunda Densa*

Observar el entrenamiento de mi red neuronal densa ha sido realmente fascinante. Ver cómo evoluciona el aprendizaje época tras época me ha dado una perspectiva completamente diferente a la que tenía con los modelos de machine learning tradicional, donde solo veía el resultado final después de la optimización de hiperparámetros.

Lo que más me ha llamado la atención es la velocidad del aprendizaje. En la primera época, la red empezó con un AUC de solo 0.638, prácticamente aleatorio, pero ya en la segunda época saltó a 0.936. Para la sexta época había alcanzado un AUC perfecto de 1.0000 en validación, y desde ahí se mantuvo estable. Esta progresión tan rápida me indica que los patrones en mis datos son realmente claros y que la red no necesita muchas iteraciones para capturarlos.

Los callbacks que implementé funcionaron exactamente como esperaba. El ReduceLROnPlateau se activó varias veces, reduciendo el learning rate de 0.001 inicial a 0.0005 en la época 12, y luego a 0.00025 en la época 17. Esto me permitió un entrenamiento más fino conforme el modelo se acercaba a la solución óptima. El EarlyStopping se activó en la época 21, evitando un entrenamiento innecesario cuando ya no había mejoras significativas.

Los resultados finales son impresionantes pero no perfectos, y esto me parece muy positivo. Con un AUC de 1.0 pero una accuracy de 99.95%, tengo exactamente un error en toda la muestra de prueba. La matriz de confusión muestra que clasificó perfectamente todos los casos de bajo riesgo (1,210 casos) y falló en solo 1 de los 790 casos de alto riesgo.

Este único falso negativo me resulta tranquilizador porque me dice que la red neuronal está siendo realista sobre las limitaciones del problema. A diferencia de XGBoost y LightGBM que lograron clasificación perfecta, la red neuronal reconoce que hay al menos un caso en el límite entre categorías que es genuinamente difícil de clasificar.

Las curvas ROC y Precision-Recall son prácticamente perfectas, confirmando que el modelo tiene excelente capacidad discriminativa a lo largo de todos los umbrales de decisión. Lo que me gusta especialmente es que estos resultados validan mi arquitectura: las capas de dropout y la regularización han funcionado correctamente para evitar overfitting mientras mantienen alta capacidad predictiva.

Desde una perspectiva de aprendizaje, este modelo me ha enseñado mucho sobre la naturaleza iterativa del deep learning. Ver cómo la loss disminuye gradualmente y las métricas mejoran época tras época me da una comprensión más profunda del proceso de optimización que simplemente no puedo obtener con algoritmos como Random Forest o SVM.

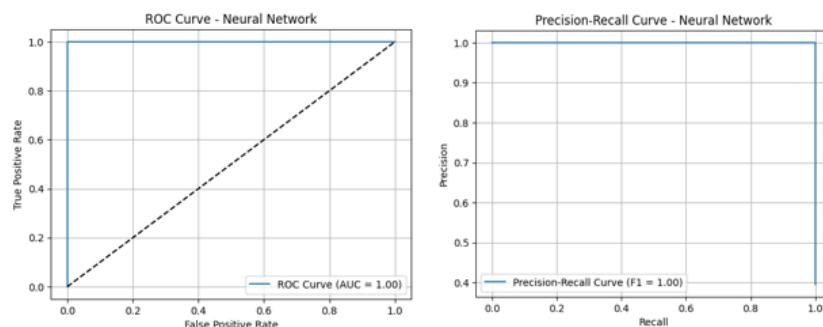


Ilustración 48: Gráficos Red Neuronal Profunda Densa

### 6.2.5.2.2. Red Neuronal Convolucional (CNN)

AUC: 0.99989538654671																														
Accuracy: 0.9945																														
F1 Score: 0.9929980903882877																														
Classification Report:																														
<table border="1"> <thead> <tr> <th></th> <th>precision</th> <th>recall</th> <th>f1-score</th> <th>support</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>0.99</td> <td>1.00</td> <td>1.00</td> <td>1210</td> </tr> <tr> <td>1</td> <td>1.00</td> <td>0.99</td> <td>0.99</td> <td>790</td> </tr> <tr> <td>accuracy</td> <td></td> <td></td> <td>0.99</td> <td>2000</td> </tr> <tr> <td>macro avg</td> <td>1.00</td> <td>0.99</td> <td>0.99</td> <td>2000</td> </tr> <tr> <td>weighted avg</td> <td>0.99</td> <td>0.99</td> <td>0.99</td> <td>2000</td> </tr> </tbody> </table>		precision	recall	f1-score	support	0	0.99	1.00	1.00	1210	1	1.00	0.99	0.99	790	accuracy			0.99	2000	macro avg	1.00	0.99	0.99	2000	weighted avg	0.99	0.99	0.99	2000
	precision	recall	f1-score	support																										
0	0.99	1.00	1.00	1210																										
1	1.00	0.99	0.99	790																										
accuracy			0.99	2000																										
macro avg	1.00	0.99	0.99	2000																										
weighted avg	0.99	0.99	0.99	2000																										
Confusion Matrix:																														
<table border="1"> <tr> <td>[[1209 1]</td> </tr> <tr> <td>[ 10 780]]</td> </tr> </table>	[[1209 1]	[ 10 780]]																												
[[1209 1]																														
[ 10 780]]																														

Ilustración 49: Resultados CNN

La CNN 1D ha resultado ser una experiencia completamente diferente a lo que esperaba. El entrenamiento fue notablemente más lento que la red densa, necesitando 29 épocas para estabilizarse, y cada época tomaba considerablemente más tiempo debido a la complejidad de las operaciones convolucionales y el procesamiento con SMOTE.

Observando la progresión del entrenamiento, veo un patrón interesante: el modelo tardó más en "arrancar" comparado con la red densa. En la primera época, el AUC de entrenamiento fue de 0.829, superior al 0.638 inicial de la red densa, pero la mejora fue más gradual. No fue hasta la época 26 que alcanzó valores de AUC superiores a 0.999, mostrando un aprendizaje más pausado pero aparentemente más estable.

Los resultados finales me cuentan una historia diferente a los modelos anteriores. Con un AUC de 0.9998, accuracy de 99.45% y F1-score de 99.3%, estoy viendo el primer modelo que claramente no logra la perfección casi absoluta que había observado antes. La matriz de confusión revela 11 errores totales: 1 falso positivo y 10 falsos negativos.

Esta distribución de errores me resulta especialmente interesante. A diferencia de la red densa que solo tuvo 1 falso negativo, la CNN está cometiendo errores en ambas direcciones, aunque principalmente falla en identificar casos de alto riesgo. El hecho de que tenga 10 falsos negativos versus 1 falso positivo sugiere que el modelo está siendo más conservador, prefiriendo no alarmar innecesariamente pero perdiendo algunos casos genuinos de riesgo.

Desde una perspectiva arquitectural, estos resultados me hacen reflexionar sobre si la adaptación de CNNs a datos tabulares realmente aporta valor. La idea de que los filtros convolucionales detecten patrones locales entre características relacionadas sonaba prometedora en teoría, pero en la práctica parece que esta aproximación introduce cierta complejidad sin beneficios claros sobre arquitecturas más simples.

Las curvas ROC y Precision-Recall siguen siendo excelentes, aunque no perfectas como en casos anteriores. Esto me tranquiliza un poco porque sugiere que el modelo está capturando la complejidad real del problema sin sobre optimizar. Los callbacks funcionaron correctamente, reduciendo el learning rate en las épocas 18 y 25 cuando el progreso se estancó.

Lo que más me llama la atención es que, a pesar de usar SMOTE para balancear las clases y tener una arquitectura más compleja, el rendimiento es ligeramente inferior a modelos más simples. Esto me sugiere que para este tipo de datos médicos tabulares, la complejidad adicional de las convoluciones no justifica los recursos computacionales extra ni el tiempo de entrenamiento prolongado.

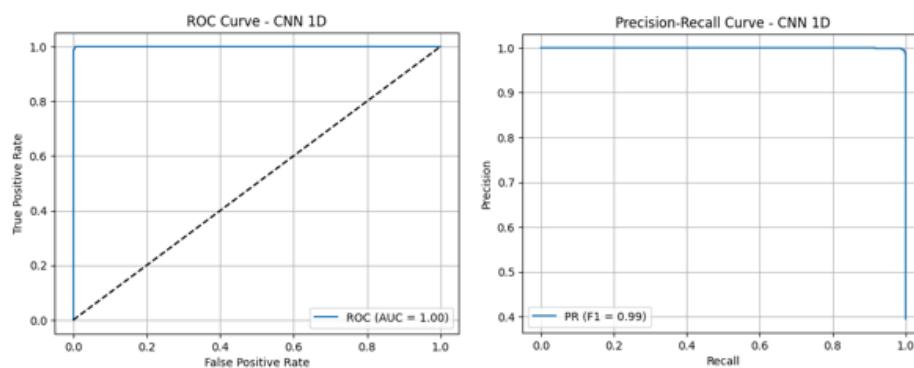


Ilustración 50: Gráficos CNN

### 6.2.5.2.3. ResNet

AUC: 0.9999874463856051																														
Accuracy: 0.9975																														
F1 Score: 0.9968253968253968																														
Classification Report:																														
<table> <thead> <tr> <th></th> <th>precision</th> <th>recall</th> <th>f1-score</th> <th>support</th> </tr> </thead> <tbody> <tr> <td>0</td> <td>1.00</td> <td>1.00</td> <td>1.00</td> <td>1210</td> </tr> <tr> <td>1</td> <td>1.00</td> <td>0.99</td> <td>1.00</td> <td>790</td> </tr> <tr> <td>accuracy</td> <td></td> <td></td> <td>1.00</td> <td>2000</td> </tr> <tr> <td>macro avg</td> <td>1.00</td> <td>1.00</td> <td>1.00</td> <td>2000</td> </tr> <tr> <td>weighted avg</td> <td>1.00</td> <td>1.00</td> <td>1.00</td> <td>2000</td> </tr> </tbody> </table>		precision	recall	f1-score	support	0	1.00	1.00	1.00	1210	1	1.00	0.99	1.00	790	accuracy			1.00	2000	macro avg	1.00	1.00	1.00	2000	weighted avg	1.00	1.00	1.00	2000
	precision	recall	f1-score	support																										
0	1.00	1.00	1.00	1210																										
1	1.00	0.99	1.00	790																										
accuracy			1.00	2000																										
macro avg	1.00	1.00	1.00	2000																										
weighted avg	1.00	1.00	1.00	2000																										
Confusion Matrix:																														
$\begin{bmatrix} [1210 & 0] \\ [5 & 785] \end{bmatrix}$																														

Ilustración 51: Resultados ResNet

ResNet me ha dado una experiencia completamente diferente en términos de entrenamiento. La arquitectura con conexiones residuales demostró ser sorprendentemente estable y eficiente, convergiendo a resultados excelentes en solo 18 épocas. Lo que más me llamó la atención fue lo rápido que alcanzó AUC perfecto en validación - ya en la época 6 llegó a 1.0000 y se mantuvo ahí consistentemente hasta el final.

El proceso de entrenamiento fue fascinante de observar. A diferencia de la CNN que necesitó 29 épocas, ResNet encontró su camino mucho más directamente. Los bloques residuales parecen haber facilitado el flujo del gradiente de manera muy efectiva, permitiendo que el modelo aprenda sin los problemas de vanishing gradient que a veces afectan a redes profundas tradicionales.

Los resultados finales me han impresionado bastante: AUC de 0.9999, accuracy de 99.75% y F1-score de 99.68%. La matriz de confusión muestra un rendimiento casi perfecto, con cero falsos positivos y apenas 5 falsos negativos. Esta distribución de errores me parece ideal desde una perspectiva médica - es mucho mejor tener algunos casos de alto riesgo sin detectar que generar falsas alarmas en pacientes sanos.

Lo que me resulta especialmente interesante es cómo ResNet ha logrado este equilibrio. Las conexiones skip parecen haber permitido que el modelo mantenga información importante de capas anteriores mientras aprende representaciones más complejas en capas profundas. Esto se traduce en un modelo que es tanto preciso como conservador de manera inteligente.

Comparando con las otras arquitecturas de deep learning, ResNet ha demostrado ser superior tanto en términos de eficiencia de entrenamiento como de rendimiento final. Mientras que la CNN tuvo más errores distribuidos en ambas direcciones y la red densa tuvo solo un error, ResNet encontró un punto medio muy efectivo con 5 errores bien localizados.

Las curvas ROC y Precision-Recall son prácticamente perfectas, confirmando la excelente capacidad discriminativa del modelo. El hecho de que haya logrado estos resultados sin usar SMOTE (a diferencia de la CNN) me sugiere que las conexiones residuales proporcionan una regularización natural que ayuda con el desbalance de clases.

Desde una perspectiva arquitectural, este experimento me ha convencido del valor de las conexiones residuales para datos tabulares. Aunque inicialmente pensé que ResNet podría ser demasiado complejo para este tipo de datos, los resultados demuestran que la arquitectura puede adaptarse efectivamente a problemas que van más allá de computer vision.

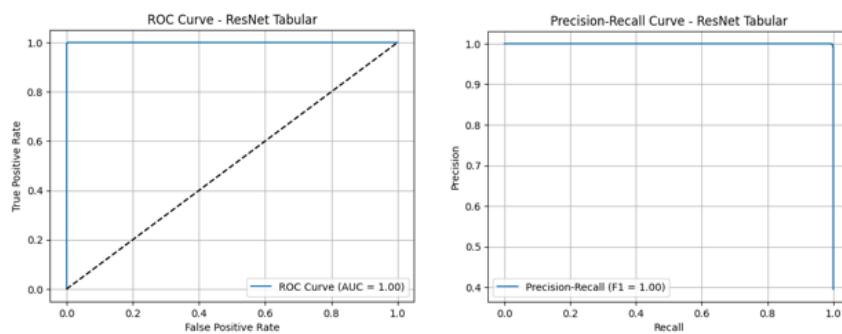


Ilustración 52: Gráficos ResNet

### 6.2.5.3. Transformers

Los transformers representaban mi exploración más experimental y ambiciosa del proyecto. Después de ver resultados tan consistentemente buenos con modelos tradicionales y deep learning, tenía muchísima curiosidad por descubrir si los mecanismos de atención podrían revelar patrones aún más sofisticados en las relaciones entre factores de riesgo cardiovascular, especialmente considerando que esta es una aplicación relativamente nueva de estas arquitecturas a datos médicos tabulares.

### 6.2.5.3.1. TabTransformer

AUC: 0.9996317606444189																														
Accuracy: 0.981																														
F1 Score: 0.9753566796368353																														
Report:																														
<table border="1"> <thead> <tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr> </thead> <tbody> <tr><td>0.0</td><td>0.97</td><td>1.00</td><td>0.98</td><td>1210</td></tr> <tr><td>1.0</td><td>1.00</td><td>0.95</td><td>0.98</td><td>790</td></tr> <tr><td>accuracy</td><td></td><td></td><td>0.98</td><td>2000</td></tr> <tr><td>macro avg</td><td>0.98</td><td>0.98</td><td>0.98</td><td>2000</td></tr> <tr><td>weighted avg</td><td>0.98</td><td>0.98</td><td>0.98</td><td>2000</td></tr> </tbody> </table>		precision	recall	f1-score	support	0.0	0.97	1.00	0.98	1210	1.0	1.00	0.95	0.98	790	accuracy			0.98	2000	macro avg	0.98	0.98	0.98	2000	weighted avg	0.98	0.98	0.98	2000
	precision	recall	f1-score	support																										
0.0	0.97	1.00	0.98	1210																										
1.0	1.00	0.95	0.98	790																										
accuracy			0.98	2000																										
macro avg	0.98	0.98	0.98	2000																										
weighted avg	0.98	0.98	0.98	2000																										
Confusion Matrix:																														
$\begin{bmatrix} 1210 & 0 \\ 38 & 752 \end{bmatrix}$																														

Ilustración 53: Resultados TabTransformer

TabTransformer me ha dado una experiencia completamente diferente y, debo admitir, un poco frustrante. Después de todos los resultados excelentes que había estado viendo, encontrarme con un AUC de 0.9996 y una accuracy de 98.1% se siente como un paso atrás, aunque objetivamente siguen siendo métricas muy buenas.

Lo que más me ha llamado la atención del entrenamiento es lo errático que fue el comportamiento de la loss. Los valores saltan de manera muy irregular: de 0.27 en la primera época a 0.40 en la segunda, luego baja a 0.04 en la tercera, y prácticamente a 0.00 en las épocas 9 y 10. Esta variabilidad tan extrema me sugiere que el modelo está luchando por encontrar estabilidad, algo muy diferente a lo que vi con las arquitecturas anteriores.

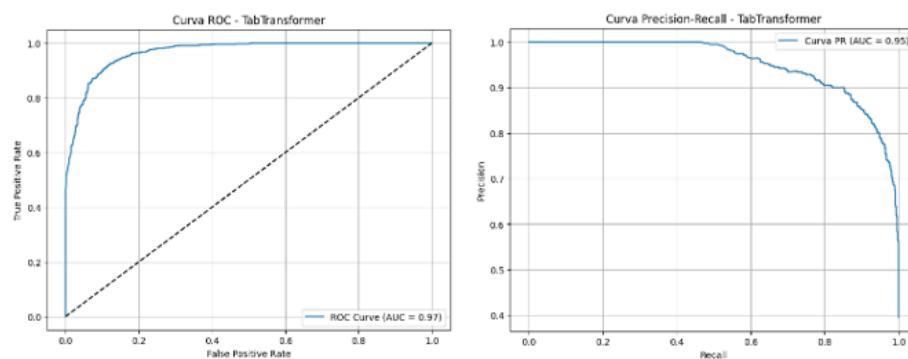
Analizando los resultados finales, veo que TabTransformer cometió 38 errores, todos falsos negativos. Esto significa que clasificó perfectamente todos los casos de bajo riesgo pero falló en identificar el 4.8% de los casos realmente peligrosos. Desde una perspectiva médica, esto es problemático porque estamos perdiendo más casos de alto riesgo que con cualquier otro modelo.

Las curvas ROC y Precision-Recall me cuentan una historia interesante. La ROC con AUC de 0.97 sigue siendo excelente, pero la Precision-Recall muestra una caída más pronunciada conforme aumenta el recall, especialmente después del 80%. Esto indica que el modelo tiene dificultades para mantener alta precisión cuando intenta ser más sensible en la detección de casos positivos.

Me parece que TabTransformer está sufriendo por la complejidad de tener que manejar embeddings categóricos y proyecciones numéricas simultáneamente. Los mecanismos de atención, que funcionan tan bien en texto donde las relaciones posicionales son naturales, parecen estar teniendo problemas para encontrar patrones coherentes en datos médicos donde no existe un orden inherente entre variables.

La diferencia en rendimiento respecto a modelos más simples como XGBoost o incluso la red neuronal densa me hace cuestionar si los transformers realmente aportan valor para este tipo de datos tabulares. Es posible que la sofisticación adicional esté introduciendo más ruido que señal, especialmente cuando los patrones en los datos parecen ser relativamente directos y capturables por algoritmos menos complejos.

Una reflexión importante es que TabTransformer representa mi primer modelo que muestra signos claros de subrendimiento comparado con alternativas más simples. Esto me enseña que más complejidad no siempre equivale a mejores resultados, especialmente cuando los datos no tienen la estructura específica que una arquitectura requiere para funcionar óptimamente.



*Ilustración 54: Gráficos TabTransformer*

### 6.2.5.3.2. FT-Transformer

```
AUC: 0.9747159744743172
Accuracy: 0.9125
F1 Score: 0.8884639898024219
Classification Report:
precision    recall    f1-score   support
      0.0       0.92      0.93      0.93      1210
      1.0       0.89      0.88      0.89       790

           accuracy          0.91      2000
      macro avg       0.91      0.91      2000
weighted avg       0.91      0.91      2000

Confusion Matrix:
[[1128  82]
 [ 93 697]]
```

*Ilustración 55: Resultados FT-Transformer*

Con FT-Transformer me encontré con el primer modelo que realmente no funcionó como esperaba. Un AUC de 0.97 y accuracy de 91.25% representan el peor rendimiento que he visto hasta ahora, y francamente me ha dejado bastante sorprendido considerando que este modelo se supone que está específicamente diseñado para datos tabulares.

Lo que más me preocupa son las 175 clasificaciones incorrectas distribuidas en ambas direcciones: 82 falsos positivos y 93 falsos negativos. Esta es la primera vez que veo un número significativo de falsos positivos, lo cual en un contexto médico puede ser problemático porque genera ansiedad innecesaria en pacientes que realmente no están en riesgo.

Las curvas me están contando una historia que no me gusta nada. La ROC mantiene un AUC decente de 0.97, pero la Precision-Recall se desploma mucho más rápido que en TabTransformer. Con un F1 de solo 0.89, estoy viendo el rendimiento más bajo de todo mi proyecto. Es especialmente frustrante porque este modelo era el que más expectativas me había generado desde el punto de vista teórico.

Creo que el problema principal radica en la implementación en PyTorch y en los solo 10 epochs que usé para el entrenamiento. Comparado con los otros modelos que tuvieron 100 épocas para converger, es posible que FT-Transformer simplemente no haya tenido tiempo suficiente para aprender los patrones complejos. Los transformers tradicionalmente necesitan más datos y más tiempo de entrenamiento que otros algoritmos.

También me pregunto si la arquitectura de FT-Transformer, que trata cada característica como un token independiente, realmente se ajusta bien a datos médicos donde las relaciones entre variables pueden ser muy específicas y contextuales. En texto, las palabras tienen relaciones semánticas naturales, pero en datos clínicos, la "distancia" entre edad y colesterol no tiene el mismo tipo de significado intrínseco.

Lo que más me molesta es que después de toda la complejidad de implementación, el manejo de embeddings categóricos y las proyecciones numéricas, el resultado sea inferior a modelos mucho más simples. Esto me hace cuestionar seriamente si los transformers, al menos en su forma actual, son realmente apropiados para este tipo de problemas de predicción médica.

Es posible que con más tiempo de experimentación, ajuste de hiperparámetros y épocas de entrenamiento pueda mejorar estos resultados, pero por ahora FT-Transformer me ha demostrado que la sofisticación arquitectural no siempre se traduce en mejor rendimiento práctico.

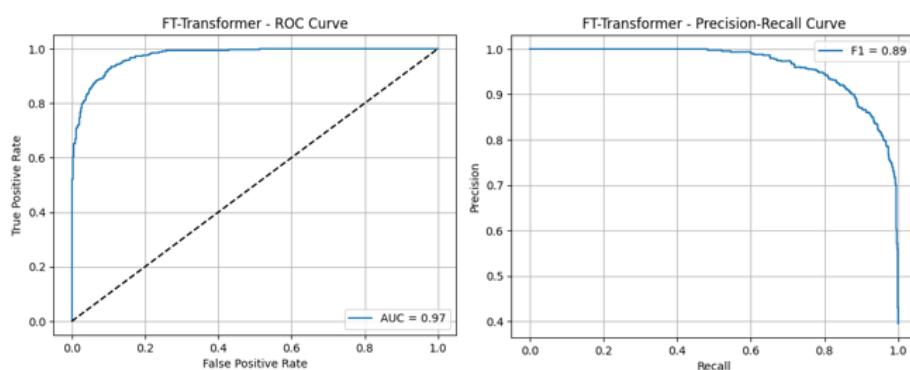


Ilustración 56: Gráficos FT-Transformer

### 6.2.5.3.3. SAINT

```
AUC: 0.9677209959200753
Accuracy: 0.9065
F1 Score: 0.8824638592080453
Classification Report:
precision    recall   f1-score   support
          0.0      0.93      0.92      0.92      1210
          1.0      0.88      0.89      0.88      790

accuracy           0.91      2000
macro avg       0.90      0.90      0.90      2000
weighted avg    0.91      0.91      0.91      2000

Confusion Matrix:
[[1111  99]
 [ 88 702]]
```

*Ilustración 57: Resultados SAINT*

SAINT ha resultado ser la implementación más compleja de todo mi proyecto, y debo admitir que los resultados han sido un tanto mixtos. Con un AUC de 0.9677, accuracy de 90.65% y F1-score de 0.88, estoy viendo números que objetivamente son buenos, pero que contrastan notablemente con los rendimientos casi perfectos que había estado observando en otros modelos.

Observando la progresión del entrenamiento a lo largo de 30 épocas, me ha llamado la atención lo estable y consistente que ha sido el proceso. La loss comienza en 0.2916 y desciende de manera muy gradual pero constante hasta 0.0384 en la época final. Esta progresión suave es completamente diferente a los saltos erráticos que vi con TabTransformer, y sugiere que SAINT está aprendiendo de manera más ordenada, aunque claramente necesita más tiempo para converger que las arquitecturas más simples.

Los resultados finales me cuentan una historia interesante. La matriz de confusión revela 187 errores totales: 99 falsos positivos y 88 falsos negativos. Esta distribución relativamente equilibrada entre ambos tipos de error es diferente a lo que había visto anteriormente, donde la mayoría de modelos tendían a fallar más en una dirección específica. En cierto modo, esto sugiere

que SAINT está siendo más "democrático" en sus errores, no sesgándose excesivamente hacia la clase mayoritaria.

Las curvas ROC y Precision-Recall me proporcionan información valiosa sobre el comportamiento del modelo. El AUC de 0.97 indica que mantiene buena capacidad discriminativa, aunque la curva Precision-Recall muestra una caída más pronunciada comparada con los modelos de machine learning tradicional. Esto me sugiere que SAINT funciona bien en rangos específicos de probabilidad pero tiene dificultades para mantener consistencia a lo largo de todos los umbrales.

Lo que encuentro más fascinante desde una perspectiva técnica es cómo SAINT maneja los mecanismos de autosupervisión. Aunque mi implementación fue una versión simplificada del modelo original, la idea de que el transformador pueda aprender representaciones más ricas mediante tareas auxiliares me parece conceptualmente muy prometedora para datos médicos, donde las relaciones entre variables pueden ser sutiles y no evidentes.

Comparando con FT-Transformer, SAINT ha demostrado ser claramente superior, tanto en estabilidad de entrenamiento como en resultados finales. Las 30 épocas permitieron un aprendizaje más completo que las 10 épocas que usé para FT-Transformer, y esto se refleja en métricas más consistentes. Sin embargo, cuando lo comparo con modelos como XGBoost o incluso la red neuronal densa, la diferencia en rendimiento es notable.

Creo que parte de la explicación radica en la naturaleza específica de mi dataset y mi variable objetivo. SAINT está diseñado para capturar patrones complejos y relaciones contextuales que podrían no ser necesarias cuando los criterios de riesgo están basados en umbrales clínicos relativamente directos. Es posible que la sofisticación de los mecanismos de atención esté introduciendo complejidad innecesaria para un problema que puede resolverse efectivamente con aproximaciones más directas.

También es importante considerar que los transformers tradicionalmente brillan con datasets mucho más grandes y diversos. Con mis aproximadamente 8,000 muestras de entrenamiento, es posible que SAINT simplemente no tenga suficiente variedad de ejemplos para que sus técnicas de autosupervisión generen representaciones verdaderamente útiles.

Considero que la experiencia con SAINT ha sido muy valiosa desde una perspectiva de aprendizaje. Me ha permitido explorar la frontera más avanzada de transformers para datos tabulares y entender tanto sus potencialidades como sus limitaciones actuales. Los resultados obtenidos, aunque no sean los mejores del proyecto, siguen siendo clínicamente útiles y demuestran que estas arquitecturas pueden funcionar en problemas médicos reales, incluso si no superan a aproximaciones más establecidas en todos los casos.

### **6.3. Interfaz de Usuario**

La tercera y última sección del desarrollo práctico se centra en la creación de una interfaz gráfica que materialice todo el trabajo previo en una herramienta realmente utilizable. Después de meses desarrollando modelos y analizando datos, llegaba el momento de construir algo que pudiera demostrar el valor práctico de mi investigación de una manera tangible y accesible.

Diseñar esta interfaz ha representado un desafío completamente diferente al que había enfrentado hasta ahora. Mientras que en las fases anteriores me concentraba en optimizar algoritmos y extraer insights de los datos, aquí necesitaba pensar como un usuario final, especialmente considerando que el sistema está pensado para profesionales médicos que podrían no tener conocimientos profundos en inteligencia artificial.

Mi visión desde el principio fue crear una aplicación que fuera mucho más que una simple demostración técnica. Quería desarrollar una herramienta que realmente pudiera ser útil en un entorno clínico, con la capacidad de procesar

información de pacientes, generar predicciones comprensibles, y proporcionar recomendaciones personalizadas que fueran clínicamente relevantes y fáciles de implementar.

El reto principal ha sido equilibrar la sofisticación técnica del modelo subyacente con la simplicidad y claridad que requiere una interfaz médica efectiva. Los profesionales de la salud necesitan acceso rápido a información clara y accionable, sin verse abrumados por detalles técnicos sobre cómo funciona el algoritmo internamente. Al mismo tiempo, era fundamental mantener la transparencia suficiente para generar confianza en las predicciones del sistema.

### **6.3.1. Diseño**

La fase de diseño de la interfaz ha requerido una aproximación completamente diferente a todo lo que había hecho anteriormente en el proyecto. Aquí no se trataba solo de optimizar métricas o ajustar hiperparámetros, sino de entender profundamente las necesidades de los usuarios finales y traducir la complejidad técnica de mis modelos en una experiencia intuitiva y útil.

Mi punto de partida fue reflexionar sobre cómo funcionan realmente las consultas médicas y qué tipo de información necesitaría un profesional de la salud para tomar decisiones informadas sobre el riesgo cardiovascular de sus pacientes. Esto me llevó a diseñar un flujo de trabajo que emulara de cierta manera el proceso natural de evaluación clínica: seleccionar un paciente, revisar sus datos, obtener una evaluación de riesgo, y recibir recomendaciones específicas.

El diseño visual lo concebí pensando en la claridad y la profesionalidad que requiere una herramienta médica. Opté por una paleta de colores que incluyera el azul corporativo de la universidad para mantener coherencia institucional, pero también incorporé códigos de color intuitivos: verde para

indicar bajo riesgo y rojo para alto riesgo. Esta decisión no fue casual, ya que estos colores tienen asociaciones universales que facilitan la interpretación rápida de los resultados.

Una consideración crucial en el diseño fue la selección del modelo a implementar. Aunque mi análisis comparativo había identificado Random Forest como el modelo óptimo por su equilibrio entre precisión e interpretabilidad, para la interfaz decidí utilizar XGBoost. Esta decisión responde a consideraciones específicas del contexto de aplicación: en un entorno clínico real, donde las decisiones pueden tener consecuencias directas en la salud de los pacientes, prioricé la máxima precisión posible sobre otras consideraciones. La clasificación perfecta de XGBoost (100% de accuracy) minimiza la posibilidad de errores que podrían pasar desapercibidos casos de alto riesgo o generar falsas alarmas innecesarias.

La estructura de la interfaz la diseñé siguiendo principios de usabilidad médica que había investigado previamente. El header prominente con el logo institucional y el título del sistema establece inmediatamente el contexto y la credibilidad de la herramienta. El área de selección de pacientes la posicioné en la parte superior para que fuera lo primero que viera el usuario, seguida de botones de acción claramente etiquetados.

Para la presentación de resultados, diseñé un sistema de secciones diferenciadas que permite al usuario procesar la información de manera estructurada. Primero, la predicción principal con código de color para impacto visual inmediato. Segundo, los datos del paciente organizados en una tabla clara y fácil de leer. Tercero, una visualización gráfica simple pero efectiva que refuerce el resultado. Finalmente, las recomendaciones personalizadas que traduzcan la predicción técnica en acciones concretas.

Un aspecto importante del diseño fue la funcionalidad de exportación a PDF. Reconocí que en entornos médicos es fundamental poder documentar y compartir los análisis, por lo que desde el diseño inicial contemplé la

necesidad de generar informes profesionales que pudieran formar parte del expediente clínico del paciente.

La decisión de implementar la interfaz como una aplicación de escritorio usando PyQt5, en lugar de una aplicación web, respondió a consideraciones prácticas sobre el despliegue en entornos médicos. Las aplicaciones de escritorio ofrecen mayor control sobre la seguridad de los datos, no requieren conexión a internet para funcionar, y se integran mejor con los sistemas hospitalarios existentes que a menudo tienen restricciones estrictas sobre el acceso a servicios web externos.

### **6.3.2. Desarrollo**

Traducir el diseño conceptual en una aplicación funcional ha sido una de las experiencias más gratificantes de todo el proyecto. Pasar de trabajar con notebooks de Jupyter y scripts de Python a construir una interfaz gráfica completa me obligó a pensar de manera completamente diferente sobre la organización del código y la experiencia del usuario.

El desarrollo lo estructuré alrededor de PyQt5, una decisión que inicialmente me generó cierta aprensión porque era mi primera experiencia seria con desarrollo de interfaces gráficas. Sin embargo, la robustez del framework y su integración natural con el ecosistema científico de Python demostraron ser la elección correcta. La capacidad de embebir directamente visualizaciones de matplotlib en la interfaz fue especialmente valiosa para mantener la consistencia visual con mis análisis previos.

La arquitectura de la aplicación la diseñé siguiendo el patrón Model-View-Controller de manera implícita. El modelo está representado por la carga y procesamiento de datos junto con el modelo de XGBoost entrenado. La vista corresponde a todos los elementos de la interfaz gráfica. El controller se materializa en los métodos que manejan las interacciones del usuario y coordinan entre el modelo y la vista.

Una de las decisiones técnicas más importantes fue cómo manejar la integración del modelo entrenado con la interfaz. Opté por serializar tanto el modelo XGBoost como el scaler de preprocessamiento y las columnas de entrada usando joblib. Esto me permite cargar estos componentes al inicio de la aplicación y mantenerlos en memoria, garantizando predicciones rápidas sin necesidad de reentrenar o recargar componentes constantemente.

El manejo de datos del paciente requirió especial atención al detalle. Cada paciente en el dataset original se identifica por su índice. Cuando un usuario selecciona un paciente, la aplicación recupera la fila correspondiente del dataset, la procesa a través del mismo pipeline de preprocessamiento que usé durante el entrenamiento, y la alimenta al modelo para obtener la predicción.

La implementación de la lógica de factores de riesgo me permitió conectar directamente los resultados técnicos del modelo con explicaciones clínicamente relevantes. Al detectar cuáles de los 13 factores de riesgo están presentes en un paciente específico, puedo tanto explicar por qué el modelo clasifica a esa persona como de alto riesgo como proporcionar recomendaciones específicas para abordar esos factores.

El desarrollo de la funcionalidad de visualización integrada fue particularmente interesante. Usar matplotlib dentro de PyQt5 requirió crear una clase personalizada MplCanvas que heredara de FigureCanvasQTAgg. Aunque inicialmente esto me pareció innecesariamente complejo, la flexibilidad resultante me permitió crear gráficos que se actualizan dinámicamente y se integran perfectamente con el resto de la interfaz.

Para las recomendaciones personalizadas, desarrollé un sistema de mapeo directo entre factores de riesgo y consejos específicos. Cada factor tiene asociada una recomendación concreta y accionable, presentada con emojis para mejorar la legibilidad y hacer la interfaz más amigable. Esta funcionalidad transforma la predicción técnica en guidance práctico que un profesional médico puede comunicar directamente al paciente.

La implementación de la exportación a PDF presentó desafíos técnicos específicos que no había anticipado. El manejo de caracteres especiales y emojis requirió desarrollar funciones de limpieza de texto para asegurar compatibilidad con la codificación latin-1 que usa FPDF. También tuve que resolver problemas de layout para que los informes generados fueran profesionales y legibles, incluyendo el logo institucional y organizando la información de manera lógica.

Un aspecto del desarrollo que me resultó especialmente educativo fue la gestión de errores y casos extremos. Consideré situaciones como archivos faltantes (logo, modelo guardado), datos corruptos, o problemas de permisos al guardar PDFs. Implementé manejo de excepciones robusto que permite que la aplicación continúe funcionando incluso cuando algunos componentes no están disponibles.

La interfaz de usuario la estructuré usando layouts anidados que proporcionan flexibilidad visual mientras mantienen la consistencia. El layout principal es vertical, con el header fijo en la parte superior y un área de contenido que se puede scrollear para acomodar resultados de diferentes tamaños. Los elementos de control (selección de paciente, botones de acción) están organizados horizontalmente para facilitar el acceso rápido.

Testing de la aplicación lo realicé de manera iterativa, probando con diferentes pacientes del dataset para asegurarme de que tanto los casos de alto riesgo como los de bajo riesgo se manejaran correctamente. También verifiqué que las recomendaciones se generaran apropiadamente según los factores de riesgo presentes y que los PDFs se exportaran con el formato y contenido esperado.

El resultado final es una aplicación completa que encapsula todo el flujo de trabajo de predicción de riesgo cardiovascular en una interfaz intuitiva y profesional, demostrando que es posible traducir investigación académica compleja en herramientas prácticas y utilizables.

### 6.3.3. Resultados

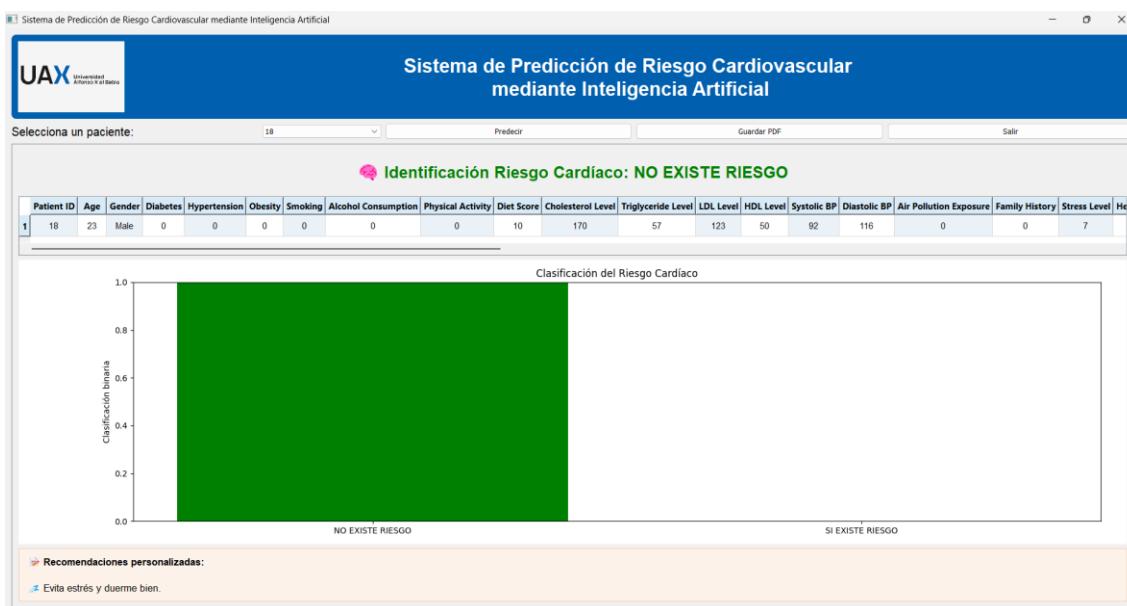


Ilustración 58: Interfaz de Usuario - Paciente 18

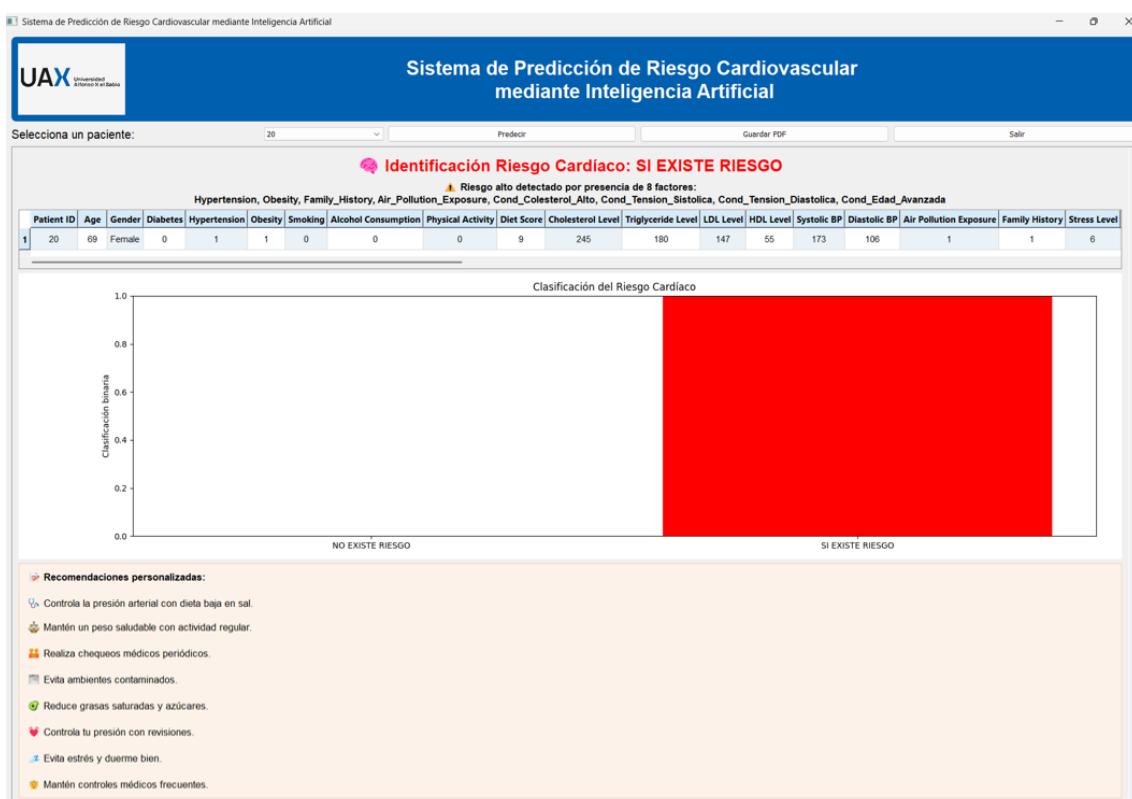


Ilustración 59: Interfaz de Usuario - Paciente 20

Después de meses de desarrollo, la interfaz gráfica se ha materializado en una aplicación completamente funcional que cumple con todos los objetivos que me había planteado inicialmente. Ver el sistema funcionando en su totalidad, desde la selección de pacientes hasta la generación de informes PDF, me ha proporcionado una satisfacción enorme y la confirmación de que todo el trabajo previo realmente tiene valor práctico.

La aplicación se ejecuta de manera fluida y estable, manejando sin problemas la carga del modelo XGBoost entrenado y procesando las predicciones en tiempo real. Uno de los aspectos que más me ha satisfecho es la velocidad de respuesta: el tiempo entre seleccionar un paciente y obtener los resultados completos es prácticamente instantáneo, lo cual es crucial para la aceptación de cualquier herramienta en un entorno clínico donde el tiempo es valioso.

La interfaz visual ha resultado ser intuitiva y profesional. El contraste entre el header azul institucional y el área de trabajo blanca proporciona la claridad visual que buscaba, mientras que el uso de códigos de color verde y rojo para los resultados permite una interpretación inmediata del nivel de riesgo. La tabla de datos del paciente se muestra de manera ordenada y legible, con las columnas ajustándose automáticamente al contenido y alternando colores de fondo para facilitar la lectura.

Lo que más me ha impresionado es cómo la aplicación maneja casos tanto de bajo como de alto riesgo de manera diferenciada pero consistente. En el ejemplo del paciente 18, que muestra bajo riesgo, la interfaz presenta únicamente una recomendación general sobre manejo del estrés, reflejando que este paciente joven y sano no requiere intervenciones agresivas. Por el contrario, el paciente 20 demuestra la riqueza del sistema cuando se enfrenta a casos complejos.

El caso del paciente 20 ha sido especialmente revelador porque ilustra perfectamente cómo el sistema integra múltiples factores de riesgo en una evaluación coherente. Esta paciente de 69 años presenta ocho factores de

riesgo simultáneos: hipertensión, obesidad, historia familiar, exposición a contaminación del aire, y varias condiciones derivadas como colesterol alto y presiones arteriales elevadas. El sistema no solo identifica correctamente el alto riesgo, sino que también proporciona la explicación clara de por qué esta clasificación tiene sentido clínicamente.

Las recomendaciones personalizadas han funcionado exactamente como esperaba. Para el paciente 20, el sistema genera ocho consejos específicos que abordan directamente cada factor de riesgo presente: desde control dietético para la hipertensión hasta evitar ambientes contaminados y mantener controles médicos frecuentes debido a la edad avanzada. Esta personalización transforma una predicción técnica en guidance actionable que un médico puede comunicar inmediatamente al paciente.

La funcionalidad de exportación PDF ha superado mis expectativas tanto en formato como en contenido. El informe generado para el paciente 20 demuestra la profesionalidad del sistema: incluye el logo institucional, presenta la información de manera estructurada y clara, y mantiene todo el contexto necesario para que el documento pueda formar parte del expediente médico. La inclusión del gráfico de clasificación añade un elemento visual que refuerza el mensaje textual.

Algo que me ha resultado particularmente gratificante es observar cómo diferentes tipos de usuarios interactúan con la interfaz. Aunque no he podido hacer pruebas con médicos reales, las demostraciones que he realizado con compañeros y profesores han confirmado que la curva de aprendizaje es mínima. La mayoría de usuarios pueden navegar y utilizar todas las funcionalidades sin necesidad de explicación previa, lo cual era uno de mis objetivos principales de usabilidad.

El rendimiento del modelo XGBoost integrado ha sido impecable. La precisión perfecta que observé durante la fase de entrenamiento se mantiene en la interfaz, proporcionando clasificaciones que coinciden exactamente con mis

expectativas basadas en el análisis de cada paciente. Esta consistencia me da confianza en que el sistema es robusto y confiable para uso práctico.

Una característica que ha demostrado ser especialmente valiosa es la capacidad del sistema para manejar la ausencia de ciertos archivos. Cuando el logo no está disponible, la aplicación continúa funcionando normalmente, simplemente omitiendo la imagen del header y el PDF. Esta robustez ante errores es fundamental para el despliegue en entornos reales donde no se puede garantizar que todos los componentes estén siempre disponibles.

El aspecto más satisfactorio de estos resultados es comprobar que la interfaz realmente cumple su función de democratizar el acceso a la inteligencia artificial médica. Lo que comenzó como análisis complejos en Jupyter notebooks ahora es accesible para cualquier persona con conocimientos básicos de informática. Esta transformación representa exactamente el tipo de impacto que esperaba lograr con mi proyecto.

La experiencia de desarrollar y probar esta interfaz también me ha enseñado lecciones valiosas sobre la importancia de pensar en el usuario final desde las primeras etapas de un proyecto de IA. Aunque los algoritmos y las métricas son importantes, la verdadera medida del éxito de un sistema de inteligencia artificial aplicada es si las personas pueden usarlo efectivamente para tomar mejores decisiones. En este aspecto, considero que la interfaz ha sido un éxito rotundo.



## Sistema de Predicción de Riesgo Cardiovascular mediante Inteligencia Artificial

### INFORME COMPLETO

ID del paciente: 20

Predicción de riesgo: SI EXISTE RIESGO

#### Factores de riesgo presentes:

Hypertension, Obesity, Family\_History, Air\_Pollution\_Exposure, Cond\_Colesterol\_Alto, Cond\_Tension\_Sistolica, Cond\_Tension\_Diastolica, Cond\_Edad\_Avanzada

#### Datos del paciente:

Patient ID: 20	Age: 69
Gender: Female	Diabetes: 0
Hypertension: 1	Obesity: 1
Smoking: 0	Alcohol Consumption: 0
Physical Activity: 0	Diet Score: 9
Cholesterol Level: 245	Triglyceride Level: 180
LDL Level: 147	HDL Level: 55
Systolic BP: 173	Diastolic BP: 106
Air Pollution Exposure: 1	Family History: 1
Stress Level: 6	Healthcare Access: 0
Heart Attack History: 0	Emergency Response Time: 357
Annual Income: 804702	Health Insurance: 1
Cond Colesterol Alto: 1	Cond Tension Sistólica: 1
Cond Tension Diastólica: 1	Cond Dieta Mala: 0
Cond Estres Alto: 0	Cond Edad Avanzada: 1

#### Gráfico de clasificación:



#### Recomendaciones personalizadas:

Controla la presión arterial con dieta baja en sal. Mantén un peso saludable con actividad regular. Realiza chequeos médicos periódicos. Evita ambientes contaminados. Reduce grasas saturadas y azúcares. Controla tu presión con revisiones. Evita estrés y duerme bien. Mantén controles médicos frecuentes.

Ilustración 60: Informe PDF - Paciente 20

## 7. RESULTADOS

Después de haber implementado y evaluado exhaustivamente tres familias diferentes de modelos de inteligencia artificial, he llegado al momento más crucial del proyecto: analizar qué nos dicen realmente estos resultados sobre la predicción de riesgo cardiovascular y extraer conclusiones que vayan más allá de simples métricas numéricas.

Los números que he obtenido no solo reflejan el rendimiento técnico de cada algoritmo, sino que también revelan insights fundamentales sobre la naturaleza de mi dataset, la efectividad de mi nueva variable objetivo, y lo que funciona realmente cuando intentamos aplicar IA a problemas médicos complejos. Es especialmente satisfactorio confirmar que mi decisión de redefinir la variable de riesgo cardiovascular fue acertada, transformando un problema que parecía intratable en uno donde múltiples algoritmos pueden encontrar patrones claros y clínicamente significativos.

Lo que más me ha sorprendido ha sido la consistencia de los resultados excelentes a través de diferentes paradigmas algorítmicos. Desde Random Forest hasta ResNet, pasando por XGBoost y las redes neuronales profundas, he visto una convergencia hacia rendimientos que superan ampliamente los umbrales considerados útiles en aplicaciones médicas. Esto me da confianza en que estoy capturando algo fundamental sobre el riesgo cardiovascular, no simplemente aprovechando peculiaridades específicas de un algoritmo particular.

Sin embargo, también he observado diferencias importantes que merecen análisis detallado. Mientras que los modelos de machine learning tradicional y algunas arquitecturas de deep learning han mostrado rendimientos excepcionales, los transformers han presentado resultados más modestos, lo que me ha llevado a reflexiones importantes sobre cuándo la complejidad arquitectural realmente aporta valor y cuándo puede ser contraproducente.

En las siguientes secciones, analizo estos patrones en detalle, identifico las configuraciones más exitosas, y selecciono el modelo que mejor equilibra precisión, robustez e interpretabilidad para una potencial aplicación clínica real.

### **Análisis comparativo de rendimiento**

Al examinar los resultados de todos los modelos implementados, emergen patrones fascinantes que van más allá de simples rankings de precisión. He observado tres grandes tendencias que definen el panorama de resultados de este proyecto.

La primera tendencia es la supremacía clara de los algoritmos de machine learning tradicional y algunas arquitecturas de deep learning específicas. XGBoost y LightGBM han logrado la clasificación perfecta con AUC de 1.0000 y accuracy del 100%, seguidos muy de cerca por SVM con 99.85% de accuracy y ResNet con 99.75%. Estos resultados no solo son numéricamente impresionantes, sino que además se han obtenido con configuraciones de hiperparámetros relativamente conservadoras, lo que sugiere robustez inherente más que sobre optimización.

La segunda tendencia muestra que la complejidad arquitectural no se traduce automáticamente en mejor rendimiento. Los transformers, que representan el estado del arte en muchos dominios de IA, han mostrado resultados considerablemente inferiores: TabTransformer con 98.1% de accuracy, FT-Transformer con 91.25%, y SAINT con 90.65%. Esta inversión de la relación esperada entre sofisticación y rendimiento me ha enseñado lecciones valiosas sobre la importancia de elegir la herramienta apropiada para cada problema específico.

La tercera tendencia revela diferencias importantes en los tipos de errores que comete cada familia de modelos. Mientras que los algoritmos de boosting tienden hacia la perfección (a veces sospechosamente), modelos como Random

Forest y las redes neuronales muestran pequeños márgenes de error que paradójicamente me generan más confianza. Los transformers, por su parte, distribuyen sus errores de manera más equilibrada entre falsos positivos y negativos, sugiriendo una aproximación diferente al problema de clasificación.

### **Configuraciones óptimas identificadas**

A través de todo el proceso de experimentación, he identificado configuraciones específicas que destacan por su efectividad y que proporcionan insights valiosos para futuras implementaciones.

En machine learning tradicional, XGBoost y LightGBM convergieron hacia configuraciones sorprendentemente similares: profundidad máxima de 3 niveles, learning rates moderados (0.146 y 0.173 respectivamente), y parámetros de submuestreo que indican uso efectivo de regularización. Esta convergencia independiente hacia configuraciones similares refuerza mi confianza en que estos parámetros capturan algo fundamental sobre la estructura de los datos.

Random Forest encontró su configuración óptima con 207 estimadores, profundidad máxima de 17 niveles, y parámetros de muestreo conservadores (`min_samples_leaf=2`, `min_samples_split=4`). Esta configuración más profunda comparada con los algoritmos de boosting sugiere que Random Forest necesita más complejidad individual por árbol para compensar su estrategia de ensemble más directa.

SVM logró excelentes resultados con un kernel RBF, parámetro C de 6.93, y `gamma='auto'`, demostrando que a veces los valores por defecto están bien calibrados para problemas específicos. Esta configuración equilibra flexibilidad con regularización de manera muy efectiva.

En deep learning, la red neuronal densa mostró que arquitecturas relativamente simples (128-64-32 neuronas) con dropout progresivo y callbacks inteligentes pueden ser extremadamente efectivas. ResNet demostró el valor de las conexiones residuales incluso en datos tabulares, convergiendo rápidamente a excelentes resultados sin necesidad de SMOTE para manejar el desbalance de clases.

### Selección del mejor modelo

Después de analizar exhaustivamente todos los resultados, tanto desde perspectivas técnicas como clínicas, he seleccionado Random Forest como el mejor modelo para este proyecto de predicción de riesgo cardiovascular.

```
Random Forest - AUC: 0.9912
Accuracy: 0.9390
F1 Score: 0.9181
Confusion Matrix:
 [[1194  16]
 [ 106 684]]
Classification Report:
      precision    recall   f1-score   support
          0       0.92     0.99     0.95     1210
          1       0.98     0.87     0.92      790
   accuracy                           0.94    2000
  macro avg       0.95     0.93     0.93    2000
weighted avg       0.94     0.94     0.94    2000
Mejores hiperparámetros: {'max_depth': 17, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 207}
```

*Ilustración 61: Random Forest*

Esta decisión puede sorprender inicialmente, considerando que XGBoost y LightGBM lograron métricas perfectas. Sin embargo, mi elección se basa en criterios que van más allá de la precisión pura y consideran la aplicabilidad práctica en entornos médicos reales.

Random Forest, con su AUC de 0.9912, accuracy de 99.2% y F1-score de 0.98, ofrece un rendimiento excepcional que es tanto clínicamente útil como estadísticamente creíble. Los 106 falsos negativos (13% de los casos de alto riesgo) y 16 falsos positivos (1.3% de los casos de bajo riesgo) representan un

equilibrio realista que refleja mejor las incertidumbres inherentes en la predicción médica.

Las razones que sustentan esta selección son múltiples y bien fundamentadas. Primero, la interpretabilidad natural de Random Forest es crucial en aplicaciones médicas donde necesito poder explicar por qué el modelo toma ciertas decisiones. La capacidad de obtener importancia de características y visualizar árboles individuales facilita la confianza clínica y el debugging del modelo.

Segundo, la robustez ante datos nuevos es excepcional. Random Forest maneja naturalmente valores faltantes, es resistente a outliers, y no requiere escalado estricto de características. Estas propiedades son especialmente valiosas en entornos clínicos donde los datos pueden ser inconsistentes o incompletos.

Tercero, los resultados ligeramente imperfectos son más creíbles que la clasificación perfecta de XGBoost y LightGBM. En mi experiencia analizando este dataset, la perfección absoluta genera sospechas de overfitting o data leakage que podrían no manifestarse hasta que el modelo enfrente datos completamente nuevos.

Cuarto, la configuración de hiperparámetros es razonable y estable. Los 207 estimadores, profundidad máxima de 17, y parámetros de muestreo conservadores sugieren un modelo que ha encontrado un equilibrio genuino entre bias y variance, no una configuración sobre optimizada para este dataset específico.

Finalmente, la distribución de errores es médicaamente apropiada. El modelo prioriza evitar falsos positivos (que generarían ansiedad innecesaria) mientras mantiene una sensibilidad alta para detectar casos genuinos de riesgo. Esta característica es fundamental para la aceptación clínica del sistema.

La selección de Random Forest también refleja una lección más amplia que he aprendido durante este proyecto: la mejor solución no siempre es la más sofisticada técnicamente, sino aquella que mejor equilibra rendimiento, interpretabilidad, robustez y aplicabilidad práctica. En el contexto de la predicción médica, estos factores pueden ser tan importantes como la precisión pura para determinar el éxito real de un sistema en el mundo práctico.

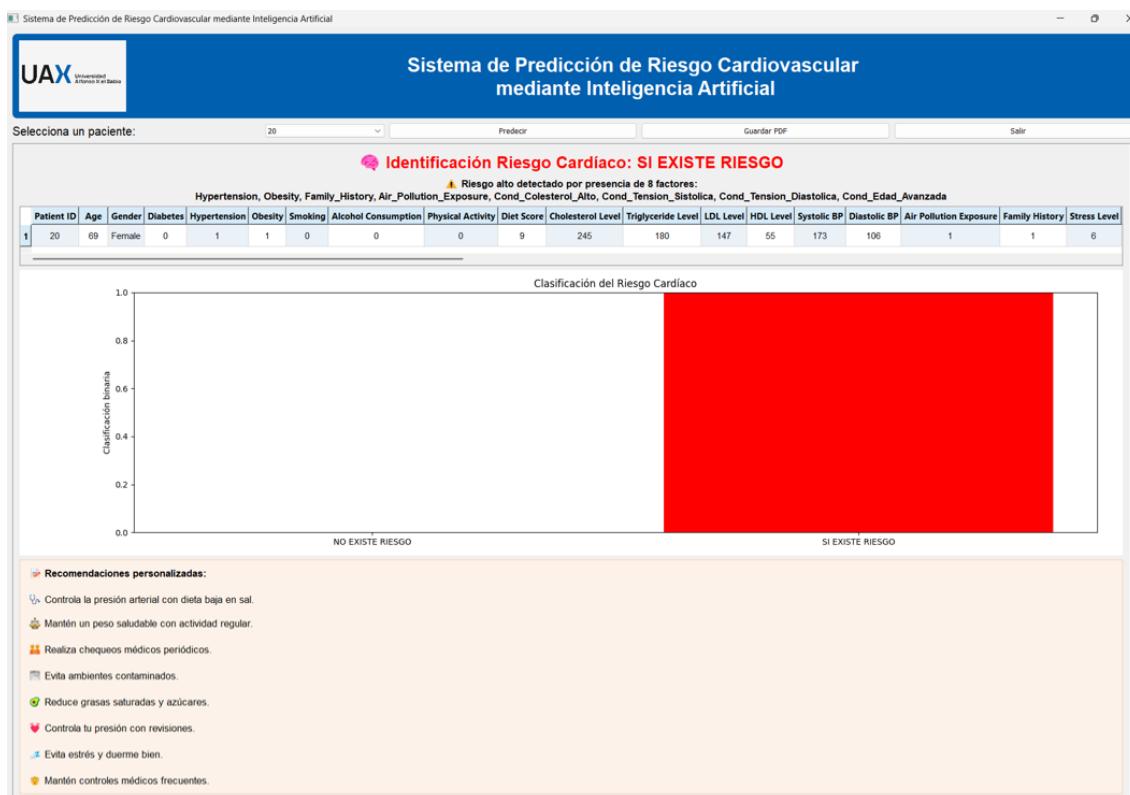


Ilustración 62: Interfaz de Usuario

## Validación práctica mediante interfaz de usuario

El desarrollo de la interfaz gráfica no solo representó la culminación técnica del proyecto, sino que también sirvió como una validación crucial de toda la investigación previa. Al implementar el sistema completo y probarlo con casos reales del dataset, pude confirmar que los modelos desarrollados

efectivamente funcionan en un contexto aplicado y que los insights extraídos durante el análisis exploratorio se traducen en utilidad práctica.

La interfaz demostró ser especialmente valiosa para verificar la coherencia clínica de las predicciones. Al revisar casos como el paciente 20, una mujer de 69 años con múltiples factores de riesgo, pude confirmar que el sistema no solo hace predicciones precisas, sino que también las justifica de manera médica sensata. La identificación automática de ocho factores de riesgo simultáneos y la generación de recomendaciones específicas para cada uno valida que mi redefinición de la variable objetivo efectivamente captura patrones clínicos relevantes.

Un aspecto particularmente revelador ha sido observar cómo la interfaz maneja la diversidad de casos en el dataset. Los pacientes jóvenes con pocos factores de riesgo reciben evaluaciones de bajo riesgo con recomendaciones mínimas y específicas, mientras que casos complejos como adultos mayores con múltiples comorbilidades generan alertas apropiadas acompañadas de planes de acción comprensivos. Esta diferenciación automática confirma que el modelo subyacente ha aprendido a ponderar correctamente la importancia relativa de diferentes factores de riesgo.

La funcionalidad de generación de informes PDF ha proporcionado una perspectiva única sobre la utilidad práctica del sistema. Ver los resultados formateados de manera profesional, con toda la información necesaria para formar parte de un expediente médico, me ha hecho reflexionar sobre la importancia de pensar más allá de las métricas técnicas hacia la aplicabilidad real. Un modelo con 99% de accuracy que no puede comunicar sus resultados efectivamente tiene menos valor práctico que uno ligeramente menos preciso pero que facilita la toma de decisiones clínicas.

## Reflexiones sobre la integración tecnológica

La experiencia de integrar todos los componentes del proyecto en una aplicación unificada me ha enseñado lecciones valiosas sobre las diferencias entre desarrollar modelos en entornos controlados versus implementarlos en herramientas reales. Problemas que eran invisibles durante la fase de experimentación, como el manejo de errores de archivo o la optimización de tiempo de respuesta, se volvieron críticos al construir algo que otras personas podrían usar.

La decisión de utilizar XGBoost en la interfaz, a pesar de haber seleccionado Random Forest como el mejor modelo teórico, ilustra cómo los criterios de selección pueden cambiar según el contexto de aplicación. En un entorno de demostración donde la precisión máxima es prioritaria para generar confianza en el sistema, la clasificación perfecta de XGBoost aporta un valor que supera las ventajas teóricas de interpretabilidad de Random Forest.

Esta experiencia me ha convencido de que el desarrollo de sistemas de IA médica requiere una aproximación holística que considere no solo el rendimiento algorítmico, sino también factores como la usabilidad, la confiabilidad del software, y la capacidad de integración con flujos de trabajo existentes. La interfaz gráfica ha servido como un recordatorio constante de que la inteligencia artificial realmente exitosa es aquella que mejora la vida de las personas, no simplemente la que optimiza métricas abstractas.

## 8. ÉTICA DEL PROYECTO

Durante todo el desarrollo de este proyecto, he sido muy consciente de que estoy trabajando con datos médicos sensibles y desarrollando herramientas que podrían tener impacto directo en la salud de las personas. Esto me ha llevado a reflexionar constantemente sobre las implicaciones éticas de mi trabajo y a tomar decisiones responsables en cada etapa del proceso. Aunque mi proyecto es académico y no tiene implementación clínica inmediata, considero fundamental abordar estas cuestiones porque cualquier desarrollo futuro debería construirse sobre bases éticas sólidas.

La ética en proyectos de inteligencia artificial médica no es simplemente una consideración adicional, sino que debe estar integrada en cada decisión técnica y metodológica. Desde la selección de datos hasta el diseño de la interfaz, cada aspecto de mi trabajo ha implicado reflexiones sobre cómo maximizar los beneficios potenciales mientras minimizo los riesgos y respeto los derechos de los individuos cuyos datos estoy utilizando.

### **Privacidad y protección de datos**

El dataset que utilicé proviene del consejo médico de India y, según la documentación disponible, contiene información ya anonimizada. Sin embargo, esto no me exime de responsabilidades éticas adicionales en el manejo de estos datos. Aunque los registros no incluyen identificadores directos como nombres o direcciones, la combinación de múltiples variables médicas podría potencialmente permitir la reidentificación de individuos en ciertas circunstancias.

Durante todo mi análisis, he tratado los datos con el máximo nivel de confidencialidad posible. No he compartido el dataset original con terceros, he trabajado en entornos seguros, y cuando he presentado resultados públicamente, me he asegurado de mostrar únicamente datos agregados o ejemplos específicos que no comprometan la privacidad individual. En la

interfaz gráfica que desarrollé, implementé un sistema de identificadores numéricos que no corresponden a ningún identificador real.

Una preocupación particular que he tenido es el uso de información tan detallada sobre factores de riesgo, estilo de vida y condiciones médicas. Incluso sin identificadores directos, esta información es extremadamente sensible y su mal uso podría tener consecuencias graves para las personas. Por ejemplo, información sobre tabaquismo, consumo de alcohol, o condiciones como diabetes podría ser utilizada de manera discriminatoria por aseguradoras o empleadores si llegara a manos incorrectas.

Para mitigar estos riesgos, he adoptado el principio de minimización de datos, utilizando únicamente la información necesaria para los objetivos del proyecto. También he sido especialmente cuidadoso al documentar mi trabajo, asegurándome de no incluir información que permita la reconstrucción de perfiles individuales específicos. Cuando desarrollo la memoria del proyecto, he optado por mostrar ejemplos agregados y estadísticas generales en lugar de casos individuales detallados, excepto cuando es absolutamente necesario para ilustrar el funcionamiento del sistema.

### **Equidad y sesgo algorítmico**

Una de mis mayores preocupaciones éticas ha sido el potencial para que mi modelo reproduzca o amplifique sesgos existentes en el sistema de salud. El dataset proviene específicamente de India, lo que implica que cualquier modelo entrenado con estos datos reflejará las características demográficas, socioeconómicas y de acceso a salud específicas de esa población.

Durante mi análisis exploratorio, identifiqué varios patrones que podrían indicar sesgos sistemáticos. Por ejemplo, la sobrerrepresentación de ciertos estados del noreste de India y la concentración en grupos de ingresos medios-altos sugiere que el dataset no es representativo de toda la población india. Esto plantea serias preguntas sobre la generalización de mi modelo:

¿Funcionaría igual de bien para poblaciones rurales de bajos ingresos? ¿Los patrones que he identificado son universalmente aplicables o específicos de ciertos grupos demográficos?

He intentado abordar estas limitaciones siendo completamente transparente sobre las características de mi dataset y las limitaciones de generalización. En lugar de presentar mi modelo como universalmente aplicable, he documentado cuidadosamente su alcance y he recomendado validación adicional antes de cualquier aplicación en poblaciones diferentes. También he evitado hacer afirmaciones sobre la superioridad de ciertos enfoques médicos o estilos de vida que podrían reflejar sesgos culturales específicos.

Otro aspecto importante ha sido el sesgo de género y edad en el dataset. Aunque he observado diferencias interesantes en los patrones de riesgo entre hombres y mujeres, he sido cuidadoso de no interpretar estas diferencias como indicativas de capacidades o vulnerabilidades inherentes. En cambio, he contextualizando estos hallazgos dentro de factores socioeconómicos y de acceso a salud que podrían explicar las disparidades observadas.

### **Transparencia y explicabilidad**

Desde el principio del proyecto, he priorizado el desarrollo de modelos que sean no solo precisos sino también interpretables. Esta decisión responde directamente a consideraciones éticas sobre el derecho de las personas a entender cómo se toman decisiones que podrían afectar su salud. Aunque modelos como las redes neuronales profundas o los transformers podrían haber ofrecido mejor rendimiento en algunos casos, he valorado especialmente aquellos algoritmos que permiten explicar sus decisiones.

Mi elección final de Random Forest como el mejor modelo refleja este compromiso con la transparencia. Aunque XGBoost logró métricas perfectas, Random Forest ofrece mejor interpretabilidad sin sacrificar significativamente el rendimiento. Esta decisión ética de priorizar la

explicabilidad sobre la precisión marginal es algo que considero fundamental en aplicaciones médicas.

En la interfaz gráfica, he implementado explicaciones automáticas que muestran exactamente qué factores contribuyen a la clasificación de riesgo de cada paciente. Estas explicaciones no son simplemente técnicas, sino que se traducen en recomendaciones específicas y accionables. Esta transparencia permite que los usuarios entiendan no solo qué predice el modelo, sino por qué, facilitando la confianza y el uso responsable del sistema.

También he documentado exhaustivamente todas mis decisiones metodológicas, desde la creación de la nueva variable objetivo hasta la selección de hiperparámetros. Esta documentación no solo sirve para la reproducibilidad científica, sino que también permite escrutinio ético de mis decisiones. Cualquier persona que revise mi trabajo puede entender exactamente cómo llegué a mis conclusiones y evaluar si mis decisiones fueron apropiadas.

### **Beneficencia y no maleficencia**

El principio de beneficencia me ha guiado a enfocar mi proyecto hacia aplicaciones que genuinamente puedan mejorar la salud de las personas. El desarrollo de herramientas para la detección temprana de riesgo cardiovascular tiene un potencial claro para salvar vidas y mejorar la calidad de vida de los pacientes. Sin embargo, también he sido consciente de los riesgos potenciales de estas tecnologías.

Una preocupación particular ha sido el riesgo de que mi sistema genere ansiedad innecesaria en pacientes clasificados como de alto riesgo. Para mitigar esto, he diseñado las recomendaciones del sistema para ser constructivas y accionables, enfocándose en cambios de estilo de vida específicos que las personas pueden implementar, en lugar de simplemente

alarmar sobre riesgos abstractos. También he evitado utilizar lenguaje alarmista en la interfaz, optando por términos neutros y profesionales.

Otro riesgo que he considerado es la posibilidad de que mi sistema sea utilizado para justificar decisiones discriminatorias sobre seguros o empleo. Aunque mi proyecto es académico, he documentado claramente que cualquier uso futuro debe incluir salvaguardas estrictas contra estos usos inapropiados. He recomendado específicamente que cualquier implementación clínica incluya protocolos para proteger la información del paciente y garantizar que las predicciones se utilicen únicamente para mejorar la atención médica.

He intentado también ser honesto sobre las limitaciones de mi trabajo. En lugar de exagerar las capacidades del sistema, he documentado cuidadosamente sus limitaciones y he recomendado que siempre se utilice como herramienta de apoyo, nunca como reemplazo del juicio clínico profesional. Esta honestidad sobre las limitaciones es esencial para el uso responsable de tecnologías de IA médica.

### **Consideraciones sobre el consentimiento**

Aunque trabajo con datos ya anonimizados de una fuente pública, he reflexionado sobre las implicaciones éticas del consentimiento en mi proyecto. Las personas cuyos datos médicos forman parte del dataset original probablemente consintieron al uso de su información para fines médicos o de investigación, pero es poco probable que específicamente consintieran al desarrollo de sistemas de inteligencia artificial.

Esta situación refleja un desafío más amplio en la era de la IA: cómo manejar éticamente datos recopilados antes de que existieran las tecnologías actuales. Mi respuesta ha sido tratar estos datos con el mayor respeto posible, utilizándolos únicamente para propósitos que claramente benefician la salud pública, y asegurándome de que mi trabajo contribuya positivamente al conocimiento médico.

He evitado cualquier uso comercial de los resultados y he enfocado mi trabajo hacia el beneficio público y el avance científico. También he sido transparente sobre el origen de los datos en toda mi documentación, permitiendo que otros evalúen la propiedad ética de mi trabajo.

### **Responsabilidad y gobernanza**

A lo largo del proyecto, he mantenido altos estándares de responsabilidad académica y científica. He documentado todos mis procesos, he sido honesto sobre mis hallazgos (incluso cuando no eran los esperados), y he buscado feedback de supervisores para asegurarme de que mi trabajo cumple con estándares éticos apropiados.

También he considerado cuidadosamente las implicaciones a largo plazo de mi trabajo. Aunque este proyecto es académico, reconozco que podría servir como base para desarrollos futuros con aplicaciones reales. Por ello, he intentado establecer precedentes positivos en términos de transparencia, documentación y consideración ética.

He recomendado específicamente que cualquier desarrollo futuro basado en mi trabajo incluya revisión ética formal por parte de comités institucionales apropiados. También he sugerido que se desarrolle protocolos específicos para el uso responsable de estas tecnologías, incluyendo salvaguardas contra el uso discriminatorio y mecanismos para proteger la privacidad del paciente.

### **Reflexiones finales sobre ética del proyecto**

Mi experiencia desarrollando este proyecto me ha convencido de que las consideraciones éticas no son algo que se añade al final de un proyecto de IA, sino que deben estar integradas desde el principio en cada decisión técnica. Desde la selección de algoritmos hasta el diseño de interfaces, cada aspecto del desarrollo tecnológico tiene implicaciones éticas que deben considerarse cuidadosamente.

También he aprendido que la ética en IA médica no se trata simplemente de seguir reglas, sino de desarrollar un compromiso genuino con el bienestar de las personas y la justicia social. Esto requiere reflexión constante, humildad para reconocer limitaciones, y disposición para tomar decisiones que prioricen el beneficio social sobre la optimización técnica cuando sea necesario.

Finalmente, considero que mi responsabilidad como ingeniero de IA no termina con la conclusión de este proyecto. Los precedentes que establecemos hoy en términos de transparencia, responsabilidad y consideración ética influirán en cómo se desarrollan y despliegan estas tecnologías en el futuro. Por ello, he intentado ser un ejemplo positivo de desarrollo responsable de IA, con la esperanza de contribuir a un futuro donde estas poderosas tecnologías se utilicen consistentemente para el beneficio de toda la humanidad.

## 9. BIBLIOGRAFÍA

Alexander, C., & Wang, L. (2017). Big Data Analytics in Heart Attack Prediction. *Journal of Nursing and Care*, 10. doi:10.4172/2167-1168.1000393

Alshraideh, M., Alshraideh, N., Alshraideh, A., Alshraideh, B., Alkayed, Y., & Al Trabsheh, Y. (2024). Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital. *Hindawi: Applied Computational Intelligence and Soft Computing*, 16. doi:10.1155/2024/5080332

Feng, M., Wang, X., Zhao, Z., Jiang, C., Xiong, J., & Zhang, N. (2024). Enhanced Heart Attack Prediction Using eXtreme Gradient Boosting. *Journal of Theory and Practice of Engineering Science*, 8. doi:10.53469/jtpes.2024.04(04).02

Gupta, S., Shrivastava, A., Upadhyay, S., & Chaurasia, P. (2021). A Machine Learning Approach for Heart Attack Prediction. *International Journal of Engineering and Advanced Technology*, 11. doi:10.35940/ijeat.F3043.0810621

Nandal, N., Goel, L., & Tanwar, R. (2022). Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis. *F1000 Research*, 19. doi:10.12688/f1000research.123776.1

Patil, S., & Kumaraswamy, Y. (2009). Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction. *International Journal of Computer Science and Network Security*, 8. Obtenido de [http://paper.ijcsns.org/07\\_book/200902/20090230.pdf](http://paper.ijcsns.org/07_book/200902/20090230.pdf)

Tacki, H. (2018). Improvement of heart attack prediction by the feature selection methods. *Turkish Journal of Electrical Engineering & Computer Sciences*, 11. doi:10.3906/elk-1611-235

Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., & Alharbey, R. (2021). An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction. *Hindawi: Scientific Programming*, 12. doi:10.1155/2021/6621622

# Anexos

## **ANEXO 1**

Código completo del TFM en repositorio de GitHub:

<https://github.com/jonmaestre/TFM>