ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

PREDICCIÓN DE ATAQUES CARDÍACOS MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Profesor
Ing. Manuel Eugenio Morocho

Autor
María del Cisne Herrera

2024

RESUMEN

La predicción de ataques cardíacos mediante técnicas de aprendizaje automático es un enfoque prometedor para mejorar la detección y prevención de esta enfermedad crítica. Este estudio se centra en el uso de algoritmos de bosques aleatorios y redes neuronales para analizar un conjunto de datos amplio y diverso de pacientes, incluyendo variables demográficas, médicas, de estilo de vida y socioeconómicas. Los objetivos principales son implementar y comparar estos modelos para identificar los factores de riesgo más significativos y desarrollar predicciones precisas del riesgo de ataque cardíaco. Los resultados indican que ambos modelos son efectivos para predecir casos sin riesgo, pero presentan desafíos al identificar casos de alto riesgo. Las técnicas de balanceo de datos y optimización de hiperparámetros mejoraron el rendimiento, pero se necesitan enfoques adicionales para incrementar la precisión en la predicción de casos de riesgo. Este estudio destaca la importancia de las técnicas de aprendizaje automático en la mejora de la práctica clínica, permitiendo intervenciones preventivas más precisas y personalizadas para reducir la mortalidad y mejorar la calidad de vida de los pacientes.

ABSTRACT

Heart attack prediction using machine learning techniques is a promising approach to improve detection and prevention of this critical disease. This study focuses on the use of random forest algorithms and neural networks to analyze a large and diverse dataset of patients, including demographic, medical, lifestyle, and socioeconomic variables. The main objectives are to implement and compare these models to identify the most significant risk factors and develop accurate predictions of heart attack risk. The results indicate that both models are effective in predicting no-risk cases, but present challenges in identifying high-risk cases. Data balancing and hyperparameter optimization techniques improved performance, but additional approaches are needed to increase accuracy in predicting at-risk cases. This study highlights the importance of machine learning techniques in improving clinical practice, enabling more accurate and personalized preventive interventions to reduce mortality and improve patients' quality of life.

ÍNDICE DEL CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La predicción de ataques cardíacos es un desafío crítico en el campo de la medicina y la salud pública debido a la alta mortalidad asociada con esta enfermedad. Según (Katarya & Meena, 2021) a nivel mundial, las enfermedades cardíacas representan una de las principales causas de muerte, y la capacidad de predecir y prevenir estos eventos puede salvar innumerables vidas. En este contexto, las técnicas de aprendizaje automático han emergido como herramientas poderosas para analizar grandes volúmenes de datos médicos y detectar patrones complejos que no son evidentes mediante métodos tradicionales.

Por tanto, este proyecto se enfoca en la utilización de algoritmos de aprendizaje automático, como bosques aleatorios y redes neuronales, para predecir el riesgo de ataques cardíacos. A través de un análisis detallado de diversas variables demográficas, médicas, de estilo de vida y socioeconómicas, se busca identificar los factores de riesgo más significativos y desarrollar modelos predictivos precisos. Los datos utilizados provienen de un conjunto extenso y diverso que incluye registros de pacientes de todo el mundo.

Los objetivos específicos de este estudio incluyen la implementación y comparación de diferentes modelos de aprendizaje automático, la evaluación de su desempeño y la identificación de las variables más influyentes en la predicción del riesgo cardíaco. Los resultados de esta investigación pueden tener importantes implicaciones para la práctica clínica, permitiendo a los profesionales de la salud intervenir de manera proactiva y personalizar los tratamientos según el perfil de riesgo de cada paciente. En última instancia, este trabajo busca contribuir al mejoramiento de la precisión diagnóstica y a la optimización de recursos en el sector de la salud, promoviendo una atención médica más eficiente y efectiva.

REVISIÓN DE LITERATURA

Una de las principales causas de mortalidad a nivel mundial es el riesgo de ataque cardíaco. Según Ansari et al. (2021) en la actualidad las enfermedades cardíacas se han convertido en una de las enfermedades más mortíferas, siendo uno de los mayores desafíos para los profesionales médicos debido a la complejidad de desarrollar un diagnóstico a tiempo de enfermedades cardíacas. Este padecimiento se encuentra influenciado por una serie de factores que incluyen antecedentes genéticos, estilo de vida y condiciones médicas preesxistentes (Ramesh et al., 2022). Según Mir & Sunanda (2023) el avance en las tecnologías de la información ha permitido la aplicación de técnicas de aprendizaje automático (machine learning) para el análisis de grandes volúmenes de datos médicos, el autor considera que estas técnicas ofrecen un potencial significativo para mejorar la precisión en la predicción del riesgo de ataque cardíaco al identificar patrones complejos y relaciones entre múltiples variables que pueden no ser evidentes mediante métodos tradicionales.

Debido a lo anterior, esta revisión de literatura se centra en los estudios más recientes que exploran los determinantes del riesgo de ataque cardíaco, haciendo hincapié en cómo el aprendizaje automático puede ser utilizado para mejorar las predicciones. Se analizan estudios que han investigado los antecedentes familiares, las condiciones médicas preexistentes como la hipertensión y la diabetes y los factores de estilo de vida como el tabaquismo y la actividad física. Y finalmente se discuten las diferentes metodologías y modelos de aprendizaje automático aplicados en estos estudios, así como los hallazgos y las implicaciones para la práctica clínica.

De acuerdo con Rahim et al. (2021) existe una gran variedad de métodos tradicionales para predecir este tipo de enfermedades como los algoritmos de minería de datos, sin embargo, tras algunos estudios se ha evidenciado que este tipo de modelos no son suficientes para predecir la presencia de enfermedades cardíacas, específicamente el ataque cardíaco. Según Reddy et al. (2021) las técnicas de aprendizaje automático como la máquina de vectores de soporte,

regresión logística, bayes ingenuos, bosque aleatorio y árbol de decisión han emergido como una de las herramientas principales para mejorar la precisión en la predicción del riesgo de ataque cardíaco. En los trabajos de Katarya & Meena (2021) y Ansari et al. (2023) se emplearon técnicas de aprendizaje supervisado como redes neuronales artificiales, KNN, árboles de decisión y Naive Bayes para desarrollar modelos predictivos, teniendo como resultado que las técnicas de aprendizaje automático son efectivas en la predicción y detección temprana de enfermedades cardíacas, siendo Naive Bayes y Bosque Aleatorio los más precisos respectivamente. Asimismo, los estudios de Ramesh et al.(2022), Nusinovici et al. (2020), Rubini et al. (2021) y Krittanawong et al. (2020) utilizaron las técnicas de regresión logística, máquinas de vectores de soporte (SVM) y bosque aleatorio cuyos resultados demostraron la eficacia de los algoritmos de aprendizaje automático en la predicción de enfermedades cardíacas. Por otra parte, en la investigación de Bailly et al. (2022) se emplearon las técnicas de regresión logística y redes neuronales, teniendo como resultado que las redes neuronales tienen una mayor precisión en la predicción del riesgo de ataque cardíaco. En estos estudios las métricas utilizadas fueron precisión, sensibilidad (recall), especificidad y puntuación F1.

Uno de los factores de riesgo más significativos que se ha identificado son los antecedentes familiares, estudios como el de Ramesh et al. (2022) han utilizado modelos de aprendizaje automático para analizar datos genéticos, teniendo como resultado que la presencia de ciertos marcadores genéticos tienden a aumentar en gran medida el riesgo de tener un ataque cardíaco. Asimismo, el estudio de Jindal et al. (2021) destaca la importancia de incluir a los antecedentes familiares como variable pues se evidenció mayor precisión en los modelos utilizados al considerar este factor de riesgo. Por otro lado, se identificaron otras variables que juegan un papel fundamental en el riesgo de tener un ataque cardíaco como el tabaquismo, el consumo de alcohol, la dieta y la actividad física. En la investigación de Mir & Sunanda (2023) se realizó una revisión exhaustiva sobre la manera en que los cambios en el estilo de vida pueden reducir el riesgo de enfermedades cardíacas, este estudio destaca la eficacia de las intervenciones basadas en el comportamiento y sugiere que la

incorporación de este tipo de datos en los modelos predictivos puede mejorar las predicciones y guiar intervenciones más efectivas.

En el caso de las condiciones médicas preexistentes, en el estudio de Ishaq et al. (2021) se utilizó modelos de aprendizaje automático para analizar registros clínicos y determinar la influencia de la hipertensión y diabetes en el riesgo de ataque cardíaco, dando como resultado que la combinación de estas enfermedades aumentan significativamente el riesgo. Por otra parte, Ansari et al., (2023) utilizaron diversas técnicas de aprendizaje automático para analizar cómo la hipertensión y diabetes afectan el riesgo de enfermedades cardíacas concluyendo que la presencia de estas condiciones son un predictor robusto del riesgo de ataque cardíaco.

Finalmente, se ha evidenciado que el uso del aprendizaje automático para predecir el riesgo de ataque cardíaco tiene varias implicaciones importantes para la práctica clínica, la investigación médica y la gestión de la salud pública. Los autores Awan et al. (2019), Ganesan & Sivakumar (2019), Wu et al. (2019) y Rani et al. (2021) expresan que utilizar un modelo preciso de predicción de ataque cardíaco mejora la precisión diagnóstica, la personalización de los tratamientos y la optimización de recursos.

FUENTES PRIMARIAS Y SECUNDARIAS

La fuente primaria que se va a utilizar en este proyecto es el conjunto de datos sobre el riesgo de ataque cardíaco, el cual es una base de datos que contiene 8,763 registros de pacientes de todo el mundo, con 26 variables detalladas que incluyen factores demográficos, médicos, de estilo de vida y socioeconómicos.

Las fuentes secundarias que se van a utilizar son:

- Estudios previos que identifican y analizan los factores de riesgo y predictores de ataques cardíacos, como niveles de colesterol, presión arterial, hábitos de fumar, actividad física, dieta, estrés, etc.

- Publicaciones en revistas médicas y de salud pública que hayan investigado la relación entre los factores de riesgo y los ataques cardíacos utilizando técnicas de aprendizaje automático.

- Estudios de caso específicos que muestren la implementación de modelos predictivos y sus resultados en la práctica clínica.

IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

El estudio de los determinantes del riesgo de ataque cardíaco es fundamental en el campo de la salud pública y la medicina preventiva. De acuerdo con Ansari et al. (2021) los ataques cardíacos, también conocidos como infartos de miocardio, son una de las principales causas de muerte en todo el mundo. Según la Organización Mundial de la Salud (2024) aproximadamente 17.9 millones de personas mueren cada año por enfermedades cardiovasculares, representando el 31% de todas las muertes globales. Bajo este contexto, este proyecto se centra en la utilización de técnicas de aprendizaje automático como redes neuronales y bosque aleatorio para identificar y analizar los factores que contribuyen al riesgo de ataque cardíaco. A diferencia de los enfoques revisados en la literatura, esta metodología busca comparar únicamente los modelos de Bosque Aleatorio y Redes Neuronales, los cuales han mostrado una alta precisión en estudios anteriores, además de integrar una mayor cantidad de variables.

El objeto de estudio de este proyecto es un conjunto de datos de pacientes que incluye una variedad de variables demográficas, médicas, de estilo de vida y socioeconómicas siendo la presencia de riesgo de ataque cardíaco la variable objetivo. Lo que este estudio busca es explorar cómo estas variables interaccionan y contribuyen al riesgo de ataque cardíaco mediante el uso de técnicas avanzadas de aprendizaje automático. Según Rahim et al. (2021) estas técnicas permiten la identificación de patrones y relaciones complejas que no son evidentes mediante análisis tradicionales.

La relevancia de este estudio se enmarca en el contexto de la alta prevalencia de enfermedades cardiovasculares por lo que la identificación precisa de los factores de riesgo es esencial para el desarrollo de intervenciones preventivas eficaces. De acuerdo con Ramesh et al. (2022) el análisis predictivo de ataques cardíacos tiene implicaciones significativas tanto para la salud pública como para la práctica clínica. Asimismo, la capacidad de predecir con precisión el riesgo de ataque cardíaco permite a los profesionales de la salud intervenir de manera proactiva, implementando estrategias de prevención y tratamiento personalizadas que pueden salvar vidas y mejorar la calidad de vida de los pacientes.

Además, estudios previos han demostrado que factores como la hipertensión, el colesterol elevado, el tabaquismo, la diabetes y la obesidad son predictores significativos de ataques cardíacos, sin embargo, la interacción entre estos factores y su influencia relativa en diferentes poblaciones aún no se comprende completamente. De esta manera, la implementación de técnicas de aprendizaje automático en la predicción de riesgos de salud puede optimizar los recursos médicos y financieros, al dirigir las intervenciones hacia aquellos individuos con mayor probabilidad de beneficio, esto es particularmente relevante en entornos con recursos limitados, donde la eficiencia y efectividad de las intervenciones son cruciales para maximizar el impacto positivo en la salud pública.

PLANTEAMIENTO DEL PROBLEMA

La problemática organizacional a ser estudiada se centra en la gestión ineficaz y la prevención de los ataques cardíacos dentro de la organización de la salud. Según Ansari et al. (2021) esta situación se deriva de la ausencia de un sistema predictivo robusto que pueda identificar con precisión a los individuos en alto riesgo de sufrir un ataque cardíaco. Actualmente, el área de la salud depende de métodos tradicionales de evaluación de riesgo a lo que Mir & Sunanda (2023) afirman que no son lo suficientemente precisos ni personalizados, como resultado, las intervenciones preventivas y los tratamientos suelen ser reactivos en lugar de proactivos.

De acuerdo a Krittanawong et al. (2020) la falta de un sistema predictivo avanzado significa que la organización no puede anticipar eficazmente los eventos cardíacos, lo que conlleva una respuesta tardía y menos efectiva. En particular, los métodos tradicionales no consideran la interacción compleja de múltiples factores de riesgo, como los demográficos, médicos, de estilo de vida y socioeconómicos, limitando así su capacidad para proporcionar una evaluación de riesgo precisa. Esta situación no solo afecta a los pacientes que experimentan eventos cardíacos, sino que también repercute negativamente en la eficiencia y la calidad de los servicios de salud que la organización ofrece, por lo que este proyecto busca superar estas limitaciones mediante la implementación de modelos predictivos avanzados de aprendizaje autimático que puedan ofrecer diagnósticos más rápidos y precisos, reduciendo así la mortalidad y mejorando la calidad de vida de los pacientes.

OBJETIVO GENERAL

- Implementar un modelo de aprendizaje automático capaz de analizar diversas variables demográficas, médicas, de estilo de vida y socioeconómicas para predecir con precisión la probabilidad de ataques cardíacos en individuos.

OBJETIVOS ESPECÍFICOS

- Implementar algoritmos de aprendizaje automático como bosque aleatorio y redes neuronales para identificar los factores de riesgo más significativos de ataques cardíacos.

- Comparar el desempeño de cada modelo para seleccionar el más adecuado para la predicción del riesgo de ataque cardíaco.

- Realizar un análisis comparativo del riesgo de ataque cardíaco entre diferentes subgrupos de la población tales como por país y niveles socioeconómicos. Este análisis ayudará a entender cómo varían los factores de riesgo y la prevalencia del riesgo de ataque cardíaco en diversas poblaciones y contextos geográficos.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

RECOLECCIÓN DE DATOS

Se utilizará una base de datos que fue obtenida del sitio web Kaggle (2024) este conjunto de datos incluye 8763 registros de pacientes de todo el mundo con 26 variables de diferente tipo.

LIMPIEZA, PRE-PROCESAMIENTO Y/O TRANSFORMACIÓN DE DATOS

El primer paso que se relizó fue cargar la base de datos a Colab, posteriormente en la base de datos se identificó una columna denominada "Patient ID" que no contribuye significativamente al proceso analítico, por lo que se ha tomado la decisión de eliminarla del conjunto de datos. Una vez elimidada la variable se procedió a cambiar el nombre de las variables que inicialmente estaban en inglés a español para un mejor entendimiento y posteriormente se realizó un análisis exploratorio de los datos teniendo los siguientes resultados.

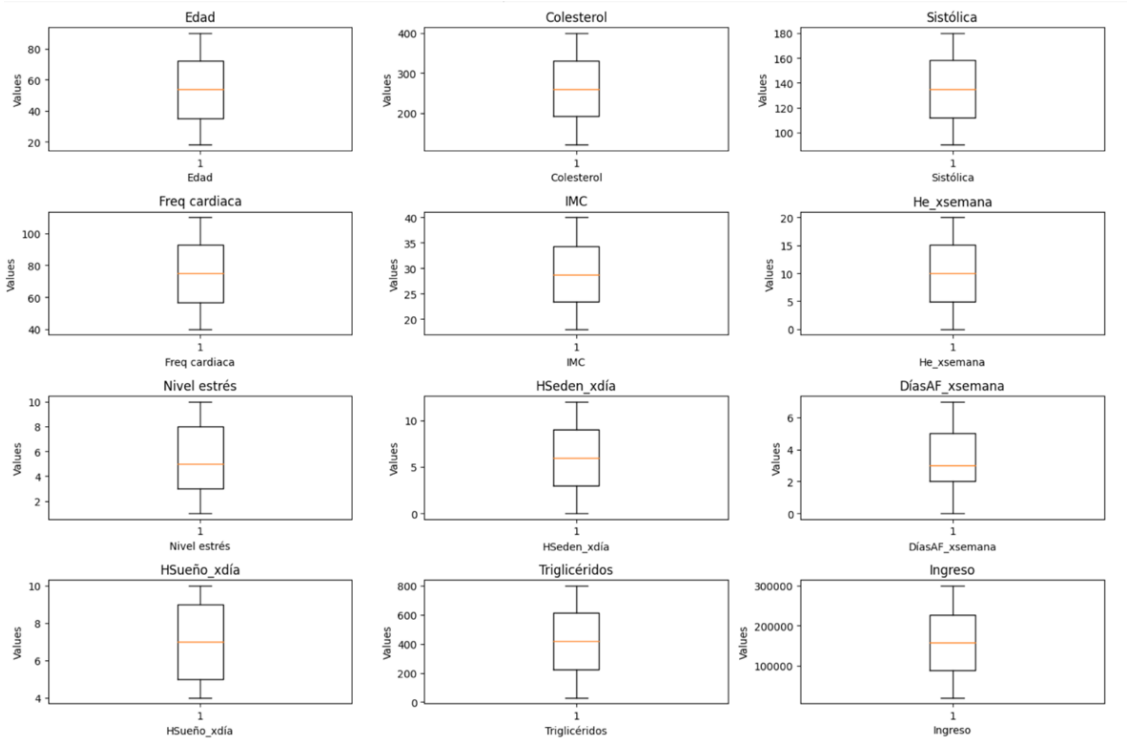
Tabla 1 ANÁLISIS EXPLORATORIO Y TRANSFORMACIÓN DE VARIABLES A INT

Base original			
#	Column	Non-Null Count	Dtype
0	Edad	8763 non-null	int64
1	Género	8763 non-null	object
2	Colesterol	8763 non-null	int64
3	Presión arterial	8763 non-null	object
4	Freq cardíaca	8763 non-null	int64
5	Diabetes	8763 non-null	int64
6	Hist_familiar	8763 non-null	int64
7	Fumador	8763 non-null	int64
8	Obesidad	8763 non-null	int64
9	Cons_alcohol	8763 non-null	int64
10	He_xsemana	8763 non-null	float64
11	Dieta	8763 non-null	object
12	Prob_cor_ant	8763 non-null	int64
13	Uso medicina	8763 non-null	int64
14	Nivel estrés	8763 non-null	int64
15	HSeden_xdía	8763 non-null	float64
16	Ingreso	8763 non-null	int64
17	IMC	8763 non-null	float64
18	Triglicéridos	8763 non-null	int64
19	DíasAF_xsemana	8763 non-null	int64
20	HSueño_xdía	8763 non-null	int64
21	País	8763 non-null	object
22	Continente	8763 non-null	object
23	Hemisferio	8763 non-null	object
24	Riesgo AC	8763 non-null	int64

Base transformada			
#	Column	Non-Null Count	Dtype
0	Edad	8763 non-null	int64
1	Género	8763 non-null	int64
2	Colesterol	8763 non-null	int64
3	Freq cardíaca	8763 non-null	int64
4	Diabetes	8763 non-null	int64
5	Hist_familiar	8763 non-null	int64
6	Fumador	8763 non-null	int64
7	Obesidad	8763 non-null	int64
8	Cons_alcohol	8763 non-null	int64
9	He_xsemana	8763 non-null	float64
10	Dieta	8763 non-null	int64
11	Prob_cor_ant	8763 non-null	int64
12	Uso medicina	8763 non-null	int64
13	Nivel estrés	8763 non-null	int64
14	HSeden_xdía	8763 non-null	float64
15	Ingreso	8763 non-null	int64
16	IMC	8763 non-null	float64
17	Triglicéridos	8763 non-null	int64
18	DíasAF_xsemana	8763 non-null	int64
19	HSueño_xdía	8763 non-null	int64
20	País	8763 non-null	int64
21	Continente	8763 non-null	int64
22	Hemisferio	8763 non-null	int64
23	Riesgo AC	8763 non-null	int64
24	Sistólica	8763 non-null	int64
25	Diastólica	8763 non-null	int64

Se evidencia que la base de datos no contiene valores nulos, sin embargo, se identificaron algunas columnas con datos tipo object. Para aplicar funciones estadísticas y de análisis se requiere que los datos sean de tipo numérico. Por lo que se debe convertir los datos object a int. Una vez transformado las variables se procedió a realizar gráficos de cajas para identificar valores atípicos.

Ilustración 1 GRÁFICOS DE CAJA



Los gráficos de caja muestran que la edad se concentra entre 40-80 años, el colesterol entre 200-350 mg/dL y la presión sistólica entre 120-160 mmHg. La frecuencia cardíaca está entre 60-90 lpm y el IMC entre 25-35. Las horas de ejercicio semanal se centran en 5-15 horas. Los triglicéridos varían entre 200-600 mg/dL y los ingresos entre 100,000-250,000 unidades monetarias. No se identificaron valores atípicos significativos.

Tabla 2 ESTADÍSTICOS DESCRIPTIVOS

	Edad	Género	Colesterol	Freq cardíaca	Diabetes	Hist familiar	Fumador	Obesidad	Cons alcohol	He_xsemana
Count	8763	8763	8763	8763	8763	8763	8763	8763	8763	8763
Mean	53.70	0.30	259.87	75.02	0.65	0.49	0.89	0.50	0.59	10.01
Std	21.24	0.45	80.86	20.55	0.47	0.49	0.30	0.50	0.49	5.78
Min	18.00	0.00	120.00	40.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	35.00	0.00	192.00	57.00	0.00	0.00	1.00	0.00	0.00	4.98
50%	54.00	0.00	259.00	75.00	1.00	0.00	1.00	1.00	1.00	10.06
75%	72.00	1.00	330.00	93.00	1.00	1.00	1.00	1.00	1.00	15.05
Max	90.00	1.00	400.00	110.00	1.00	1.00	1.00	1.00	1.00	19.99

	IMC	Triglicéridos	DíasAF xsemana	HSueño xdía	País	Continente	Hemisferio	Riesgo AC	Sistólica	Diastólica
Count	8763	8763	8763	8763	8763	8763	8763	8763	8763	8763
Mean	28.89	417.67	3.48	7.02	9.38	2.42	0.35	0.35	135.07	85.15
Std	6.31	223.74	2.28	1.98	5.78	1.59	0.47	0.47	26.34	14.67
Min	18.00	30.00	0.00	4.00	0.00	0.00	0.00	0.00	90.00	60.00
25%	23.42	225.50	2.00	5.00	4.00	1.00	0.00	0.00	112.00	72.00
50%	28.76	417.00	3.00	7.00	9.00	3.00	0.00	0.00	135.00	85.00
75%	34.32	612.00	5.00	9.00	14.00	4.00	1.00	1.00	158.00	98.00
Max	39.99	800.00	7.00	10.00	19.00	5.00	1.00	1.00	180.00	110.00

En la ilustración 3 se observa una visión general de la distribución de variables en donde la edad promedio es de aproximadamente 54 años con un rango que va de 18 a 90 años, los niveles de colesterol tienen una media de 259 mg/dL y una desviación estándar de 80, con valores que oscilan entre 120 y 400 mg/dL.Por otra parte, la frecuencia cardíaca promedio es de 75 latidos por minuto, el IMC medio es de 28.89, sugiriendo un peso promedio en la categoría de sobrepeso con un rango de 18 a 40. También se observa que las personas realizan ejercicio en promedio 3.5 días a la semana y duermen aproximadamente 7 horas al día.

IDENTIFICACIÓN Y DESCRIPCIÓN DE VARIABLES

Se utilizará una base de datos que fue obtenida del sitio web Kaggle (2024) este conjunto de datos incluye 8763 registros de pacientes de todo el mundo con 26 variables de diferente tipo, que incluyen factores demográficos, médicos, de estilo de vida y socioeconómicos. A continuación se presenta cada variable con su clasificación y tipo de datos.

Tabla 3 CLASIFICACIÓN DE VARIABLES POR NATURALEZA Y TIPO DE DATOS

FACTORES ESPECÍFICOS DE LOS PACIENTES		
Variable	**Significado**	**Tipo de variable**
1. Patient ID	Identificador único para cada paciente	Alfanumérica
2. Age	Edad del paciente	Numérica
3. Sex	Sexo	Dicotómica (1: mujer, 0: hombre
4. Cholesterol	Niveles de colesterol del paciente	Numérica
5. Blood pressure	Presión arterial (sistólica/diastólica)	Numérica
6. Heart rate	Frecuencia cardíaca del paciente	Numérica

7. Diabetes	Si el paciente tiene diabetes	Textual (Sí/No)
8. Family history	Historia familiar de problemas relacionados con el corazón	Dicotómica (1: Sí, 0: No)
9. Smoking	Condición de fumador	Dicotómica (1: fumador, 0: no fumador)
10. Obesity	Estado de obesidad	Dicotómica (1: obeso, 0: no obeso)
11. Alcohol consumption	Nivel de consumo de alcohol	Categórica (Ninguno/Leve/Moderado/Intenso)
12. BMI	Índice de masa corporal (IMC)	Numérica
FACTORES DE ESTILO DE VIDA		
Variable	**Significado**	**Tipo de variable**
Exercise hours per week	Número de horas de ejercicio por semana	Numérica
Diet	Hábitos dietéticos del paciente	Categórica
Stress level	Nivel de estrés informado por el paciente (1-10)	Numérica
Sedentary hours per day	Horas de actividad sedentaria por día	Numérica
Physical activity days per week	Días de actividad física por semana	Numérica
Sleep hours per day	Horas de sueño por día	Numérica
ASPECTOS MÉDICOS		
Variable	**Significado**	**Tipo de variable**
Previous heart problems	Problemas cardíacos previos del paciente	Dicotómica (1: Sí, 0: No)
Medication use	Uso de medicamentos por parte del paciente	Dicotómica (1: Sí, 0: No)
Triglycerides	Niveles de triglicéridos del paciente	Numérica
FACTORES SOCIOECONÓMICOS		
Variable	**Significado**	**Tipo de variable**
Income	Nivel de ingresos del paciente	Numérica
ATRIBUTOS GEOGRÁFICOS		
Variable	**Significado**	**Tipo de variable**
Country	País del paciente	Categórica
Continent	Continente donde reside el paciente	Categórica
Hemisphere	Hemisferio donde reside el paciente	Categórica
VARIABLE OBJETIVO		
Variable	**Significado**	**Tipo de variable**
Heart attack risk	Presencia de riesgo de ataque cardíaco	Dicotómica (1: Sí, 0: No)

Tabla 4 MATRIZ DE CORRELACIÓN

	Edad	Género	Colesterol	Freq cardíaca	Diabetes	Historial familiar	Fumador	Obesidad	Consumo alcohol	He xsemana
Edad	1.00	-0.02	-0.00	-0.00	-0.01	0.00	0.39	-0.00	-0.00	0.00
Género	-0.02	1.00	-0.00	0.01	-0.00	-0.00	-0.51	-0.00	-0.00	0.00
Colesterol	-0.00	-0.00	1.00	0.00	-0.01	-0.02	0.01	-0.01	-0.00	0.02
Freq cardíaca	-0.00	0.01	0.00	1.00	0.00	-0.01	-0.01	0.01	0.00	0.00
Diabetes	-0.01	-0.00	-0.01	0.00	1.00	-0.01	0.00	0.01	0.00	-0.00
Historial familiar	0.00	-0.00	-0.02	-0.01	-0.01	1.00	0.01	-0.00	0.01	-0.00
Fumador	0.39	-0.51	0.01	-0.01	0.00	0.01	1.00	0.00	0.01	-0.00
Obesidad	-0.00	-0.00	-0.01	0.01	0.01	-0.00	0.00	1.00	-0.02	0.00
Cons_alcohol	-0.00	-0.00	-0.00	0.00	0.00	0.01	0.01	-0.02	1.00	-0.00

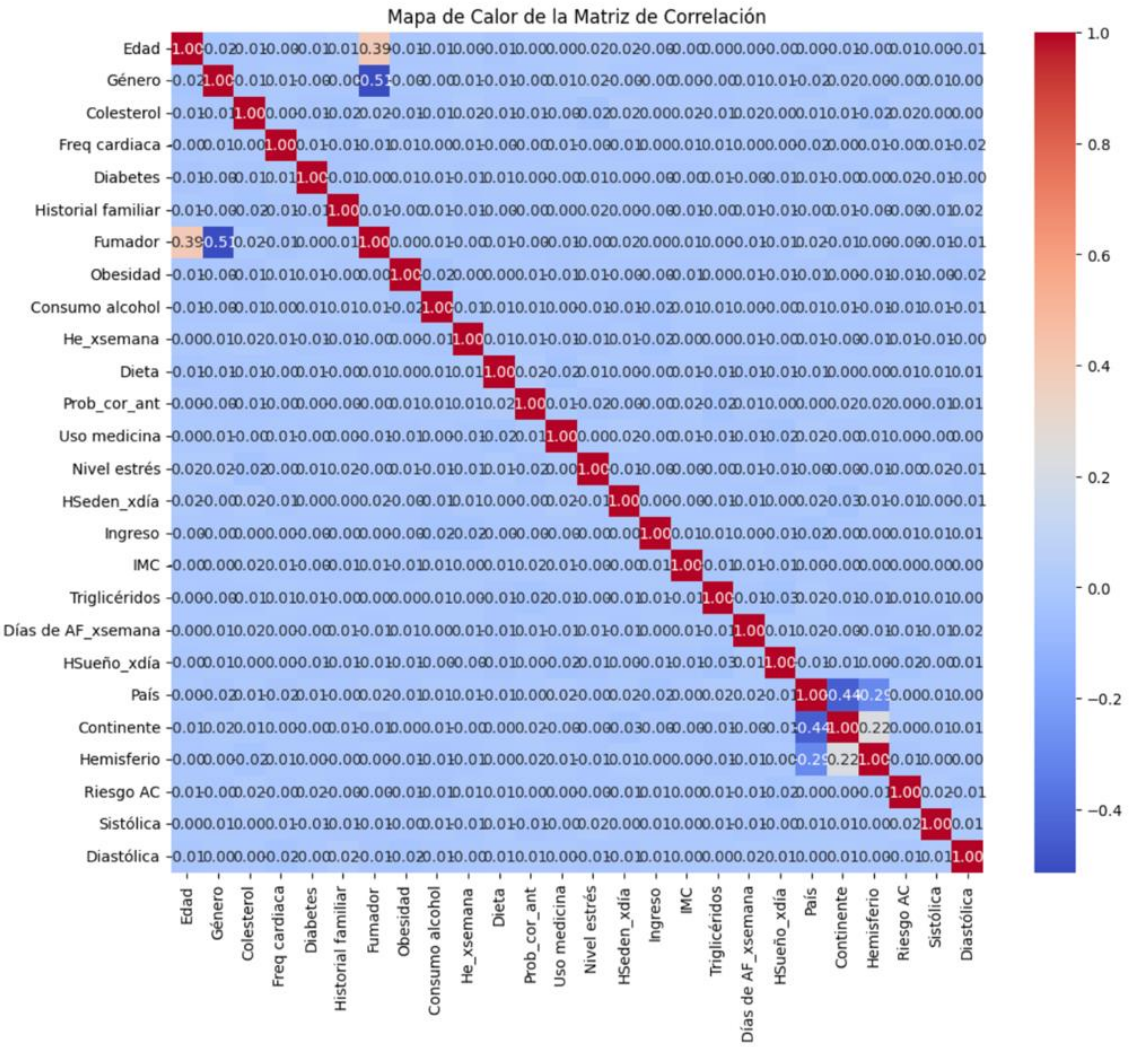
He_xsemana	0.00	0.00	0.02	0.00	-0.00	-0.00	-0.00	0.00	-0.00	1.00
Dieta	-0.01	-0.00	-0.01	-0.00	0.00	-0.00	0.00	0.00	0.00	0.00
Prob_cor_ant	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.01	0.00
Uso medicina	0.00	0.00	-0.00	0.00	-0.00	0.00	-0.01	-0.00	0.00	-0.00
Nivel estrés	0.01	0.02	-0.02	-0.00	0.00	0.01	-0.00	0.01	-0.00	-0.00
Hseden_xdía	0.01	-0.00	0.01	-0.01	0.00	0.00	0.01	-0.00	-0.01	0.00
Ingreso	-0.00	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.02	-0.02
IMC	-0.00	0.00	0.01	0.00	-0.00	-0.01	0.00	-0.00	0.01	0.00
Triglicéridos	0.00	-0.00	-0.00	0.01	0.01	-0.00	0.00	0.00	0.00	0.00
DAF xsemana	0.00	0.00	0.01	0.00	-0.00	0.00	-0.00	0.00	0.00	0.00
HSueño_xdía	-0.00	0.00	0.00	0.00	-0.01	-0.01	-0.00	-0.00	-0.00	-0.00
País	0.00	-0.01	0.01	-0.01	0.01	-0.00	0.02	-0.00	0.01	0.00
Continente	-0.01	0.01	0.00	0.00	-0.00	0.00	-0.01	0.00	0.00	-0.00
Hemisferio	-0.00	0.00	-0.01	0.01	0.00	-0.00	0.00	-0.00	-0.01	-0.00
Riesgo AC	0.00	-0.00	0.01	-0.00	0.01	-0.00	-0.00	-0.01	-0.01	0.01
Sistólica	0.00	0.00	0.00	0.00	-0.00	-0.00	-0.00	-0.00	0.01	-0.00
Diastólica	-0.00	0.00	0.00	-0.01	-0.00	0.01	-0.01	-0.02	-0.00	-0.00

	IMC	**Triglicéridos**	**Días AF xsemana**	**HSueño xdía**	**País**	**Continente**	**Hemisferio**	**Riesgo AC**	**Sistólica**	**Diastólica**
Edad	-0.00	0.00	0.00	-0.00	0.00	-0.01	-0.00	0.00	0.00	-0.00
Género	0.00	-0.00	0.00	0.00	-0.01	0.01	0.00	-0.00	0.00	0.00
Colesterol	0.01	-0.00	0.01	0.00	0.01	0.00	-0.01	0.01	0.00	0.00
Freq cardíaca	0.00	0.01	0.00	0.00	-0.01	0.00	0.01	-0.00	0.00	-0.01
Diabetes	-0.00	0.01	-0.00	-0.01	0.01	-0.00	0.00	0.01	-0.00	-0.00
His_familiar	-0.01	-0.00	0.00	-0.01	-0.00	0.00	-0.00	-0.00	-0.00	0.01
Fumador	0.00	0.00	-0.00	-0.00	0.02	-0.01	0.00	-0.00	-0.00	-0.01
Obesidad	-0.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.01	-0.00	-0.02
Cons_alcohol	0.01	0.00	0.00	-0.00	0.01	0.00	-0.01	-0.01	0.01	-0.00
He_xsemana	0.00	0.00	0.00	-0.00	0.00	-0.00	-0.00	0.01	-0.00	-0.00
Dieta	0.01	-0.01	-0.01	-0.01	-0.00	0.00	0.00	0.00	0.01	0.00
Prob_cor_ant	0.01	-0.01	0.00	0.00	0.00	0.02	0.01	0.00	-0.01	0.00
Uso medicina	0.00	-0.01	-0.01	-0.02	0.01	-0.00	0.01	0.00	-0.00	0.00
Nivel estrés	-0.00	-0.00	0.00	-0.01	-0.00	-0.00	-0.00	-0.00	0.01	-0.00
Hseden_xdía	-0.00	-0.00	-0.00	0.00	0.01	-0.02	0.01	-0.00	0.00	-0.00
Ingreso	0.00	0.01	0.00	-0.00	-0.01	-0.00	0.00	0.00	0.01	0.00
IMC	1.00	-0.00	0.00	-0.01	0.00	-0.00	0.00	0.00	0.00	0.00
Triglicéridos	-0.00	1.00	-0.00	-0.02	0.01	-0.01	-0.01	0.01	0.00	0.00
DAF_xsemana	0.00	-0.00	1.00	0.01	0.01	-0.00	-0.01	-0.00	-0.00	0.01
HSueño_xdía	-0.01	-0.02	0.01	1.00	-0.00	-0.01	0.00	-0.01	-0.00	0.01
País	0.00	0.01	0.01	-0.00	1.00	-0.44	-0.29	0.00	0.00	0.00
Continente	-0.00	-0.01	-0.00	-0.01	-0.44	1.00	0.22	0.00	0.00	0.01
Hemisferio	0.00	-0.01	-0.01	0.00	-0.29	0.22	1.00	-0.01	0.00	0.00
Riesgo AC	0.00	0.01	-0.00	-0.01	0.00	0.00	-0.01	1.00	0.01	0.00
Sistólica	0.00	0.00	-0.00	-0.00	0.00	0.00	0.00	0.01	1.00	0.01
Diastólica	0.00	0.00	0.01	0.01	0.00	0.01	0.00	-0.00	0.01	1.00

En la ilustración 4 se muestra la matriz de todas las variables, en donde se identificó que los problemas de corazón anteriores tienen la correlación más alta con el riesgo de ataque cardíaco (0.021855), lo que indica que tener antecedentes de problemas coronarios está ligeramente asociado con un mayor riesgo de AC. Además, se determinó que el colesterol con un valor de (0.019462) tiene una ligera correlación positiva, es decir, mientras niveles más altos de colesterol se tiene un mayor riesgo de AC. También se evidenció otros niveles que muestran una ligera correlación con el riesgo de ataque cardíaco como el nivel de estrés (0.017879), historial familiar (0.017489), obesidad (0.017306), sistólica (0.018585), edad (0.002463) y IMC (0.001138).

Por otro lado, se identificaron variables con una correlación negativa al riesgo de ataque cardíaco como la frecuencia cardíaca (-0.014633), hemisferio (-0.004446), horas de sueño por día (-0.003860) y días de actividad física por semana (-0.007046).

Ilustración 2 MAPA DE CALOR DE LA MATRIZ DE CORRELACIÓN

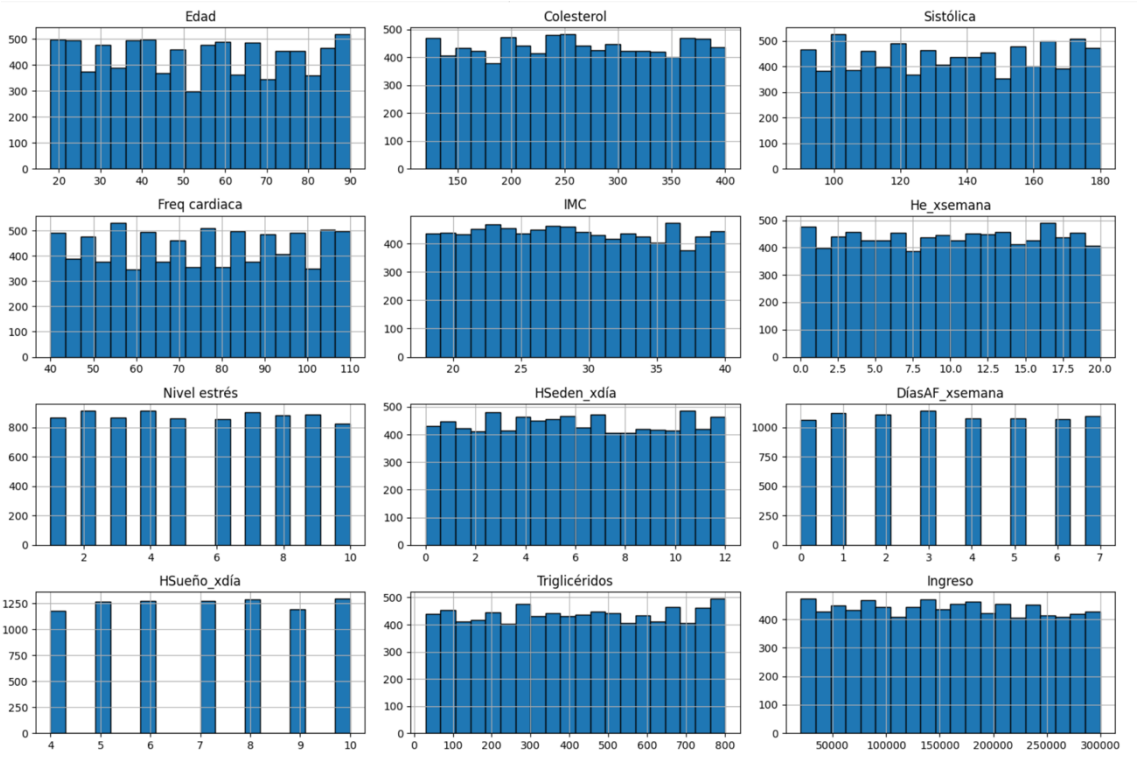


En la ilustración 5 se observa que la mayoría de las correlaciones con el riesgo de ataque cardíaco son bastante bajas, lo que se refleja en los colores predominantemente claros alrededor de esta variable en el mapa de calor. Las correlaciones más destacadas, aunque no fuertes, son con Prob_cor_ant, Colesterol, Nivel estrés, Historial familiar, Obesidad y Sistólica, todas con correlaciones positivas leves. Además se destaca una alta correlación entre condición de fumador con la edad, continente con país y país con hemisferio.

VISUALIZACIÓN DE VARIABLES

En primer lugar se realizó un gráfico de la distribución de las variables, con el objetivo de analizar cada una de ellas por separado y de esta manera entender sus características individualmente sin considerar las relaciones entre diferentes variables.

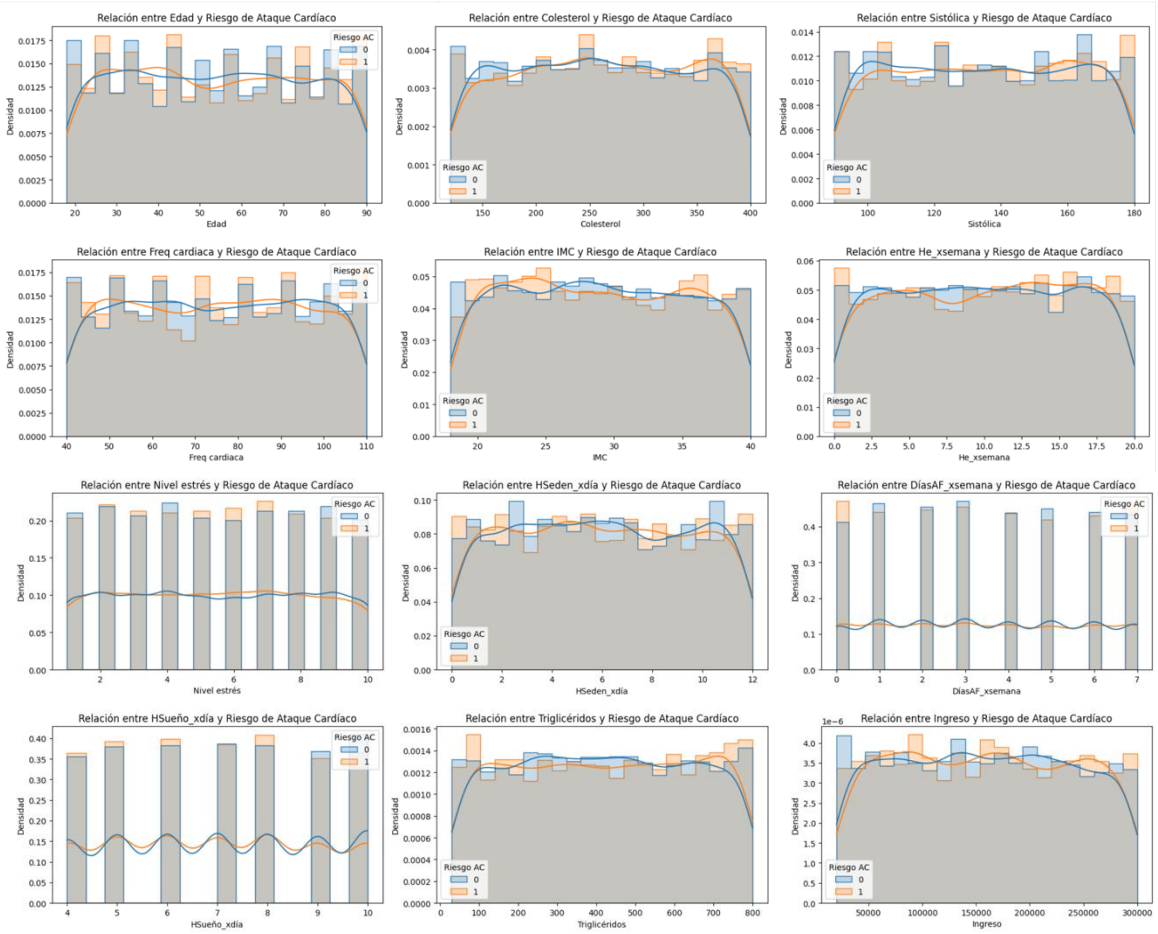
Ilustración 3 DISTRIBUCIÓN DE VARIABLES CLAVE



En la ilustración 6 se evidencia que la distribución de las edades y los niveles de colesterol entre los individuos es relativamente uniforme, sin un grupo o rango dominante. La frecuencia cardíaca muestra una distribución dispersa y las horas de ejercicio por semana tienden ligeramente hacia menos horas. Además, los niveles de estrés están distribuidos uniformemente entre 1 y 10, las horas sedentarias diarias presentan una ligera concentración entre 4 y 6 horas, el Índice de Masa Corporal (IMC) varía uniformemente entre 20 y 40, con una ligera tendencia hacia los valores más bajos, mientras que los ingresos muestran una distribución dispersa con una ligera inclinación hacia los ingresos más altos. En el caso de los triglicéridos se evidencia que tiene una distribución uniforme al igual que la cantidad de días de actividad física por semana y las horas de sueño por día. Finalmente, la presión arterial sistólica se distribuye uniformemente en el rango de 100 a 180 mmHg, tendiendo ligeramente hacia los valores más altos.

En segundo lugar se realizó un análisis bivariado, el cual consiste en analizar de manera simultánea dos variables con el objetivo de determinar que relación existe entre ellas o de que manera una variable afecta a la otra. En este caso, el gráfico muestra la relación de las variables con el riesgo de ataque cardíaco.

Ilustración 4 ANÁLISIS BIVARIADO CON EL RIESGO DE ATAQUE CARDÍACO



En la ilustración 7 se evidencia que el riesgo de ataque cardíaco varía ligeramente de acuerdo con la edad, sin embargo, no se ha podido identificar una tendencia clara de aumento o disminución en casos específicos. En el caso del colesterol, no se presentan tendencias significativas en las distribuciones de los grupos de con y sin riesgo de ataque cardíaco lo que indica que esta variable por sí sola no puede ser un indicador importante del riesgo de ataque cardíaco.

En la variable de presión arterial sistólica se muestra que el riesgo de ataque cardíaco se concentra ligeramente en los niveles más altos de presión arterial, lo que indica que las personas que tienen una mayor presión sistólica son más propensas a sufrir un ataque cardíaco. Además se evidencia que la frecuencia cardíaca, las horas de ejercicio por semana, los días de actividad física por semana y el ingreso presentan distrubuciones similares en ambos grupos lo que

sugiere que estas variables no son factores determinantes en el riesgo de ataque cardíaco.

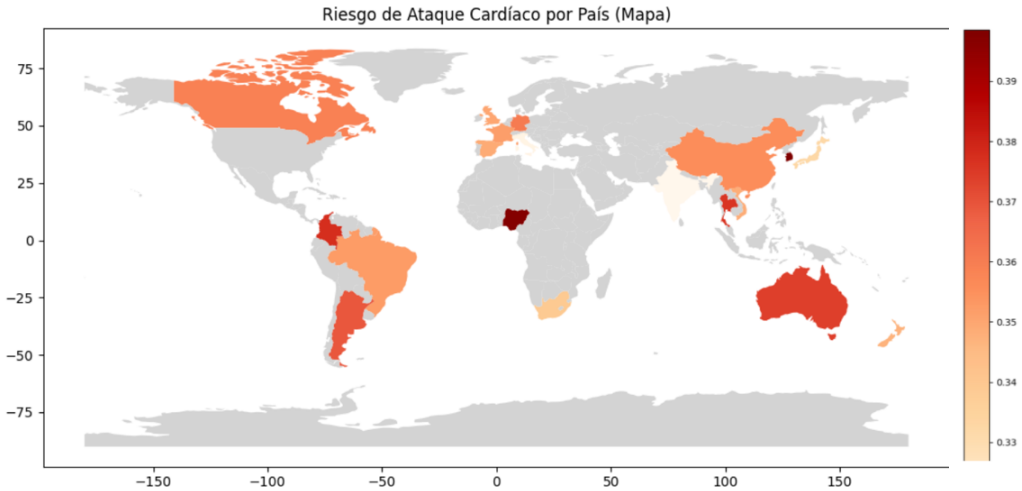
Por otra parte, en base a la gráfica se evidencia que en variables como el índice de masa corporal, el nivel de estrés, las horas de sendentarismo por día y los triglicéridos se tiene una mayor densidad en el grupo con riesgo de ataque cardíaco, es decir, a niveles altos de estas variables más riesgo se tiene de sufrir un ataque cardíaco. Finalmente, en el caso de las horas de sueño por día se infiere que tanto la falta de sueño como el exceso de sueño podrían estar asociados con un mayor riesgo de ataque cardíaco, las personas que duermen menos de 6 horas o más de 9 horas al día presentan una mayor densidad en el grupo con riesgo de ataque cardíaco.

Tabla 5 RIESGO DE ATAQUE CARDÍACO POR PAÍS

País	RA promedio
Nigeria	0.40
Corea del Sur	0.40
Colombia	0.38
Tailandia	0.38
Argentina	0.37
Australia	0.37
Canadá	0.36
China	0.36
Alemania	0.36
Brasil	0.35

País	RA promedio
Francia	0.35
Nueva Zelanda	0.35
España	0.35
Reino Unido	0.35
Vietnam	0.35
África del Sur	0.34
Japón	0.33
Italia	0.32
India	0.31

Ilustración 5 RIESGO DE ATAQUE CARDÍACO POR PAÍS



Se realizó una tabla con el riesgo de ataque cardíaco promedio por país, en base a la tabla se generó un mapa en el cual se utiliza una escala de colores para diferenciar el nivel de riesgo de ataque cardíaco por país, los colores más oscuros indican un mayor riesgo, mientras que los colores claros indican un menor riesgo. Esta gráfica fue generada utilizando los datos de la base del proyecto, la cual incluye registros detallados de pacientes de diversos países, permitiendo visualizar la distribución del riesgo de ataque cardíaco por país. La información fue analizada y representada gráficamente para identificar las áreas geográficas con mayor y menor riesgo, utilizando una escala de colores donde los tonos más oscuros indican un mayor riesgo y los más claros un menor riesgo.

Mediante el mapa se puede determinar que el riesgo de ataque cardíaco no es uniforme y varía significativamente entre regiones y países. En el caso de Nigeria y Korea del Sur, el riesgo de ataque cardíaco es muy elevado a comparación de otros países como Argentina, Tailandia, India, Australia, Colombia y Reino Unido los cuales no presentan un riesgo tan elevado. Por otra parte, los países con un riesgo moderado son Canadá, Francia, Alemania, Brasil, Vietnam, China, Italia, España y Nueva Zelanda. Finalmente los países que tienen el menor riesgo de tener un ataque cardíaco son Japón y África del Sur.

SELECCIÓN DE MODELO ESTADÍSTICO

Para este estudio, la metodología de aprendizaje automático es seleccionada debido a que mediante la revisión de estudios anteriores se evidenció su capacidad para manejar grandes volúmenes de datos y su eficacia en la identificación de patrones complejos entre múltiples variables que no son evidentes mediante métodos tradicionales. Según Reddy et al. (2021) las técnicas de aprendizaje automático como la máquina de vectores de soporte, regresión logística, bayes ingenuos, bosque aleatorio y árbol de decisión han emergido como una de las herramientas principales para mejorar la precisión en la predicción del riesgo de ataque cardíaco.

Los algoritmos de aprendizaje automático que se aplicarán son el bosque aleatorio y redes neuronales. En primer lugar, se eligió a bosque aleatorio debido

a que según Rahim et al. (2021) los bosques aleatorios han demostrado ser una técnica robusta y eficiente para la predicción de enfermedades cardíacas, además, su capacidad para manejar grandes conjuntos de datos y múltiples variables los hace ideales para problemas complejos de salud. Asimismo, Ramesh et al. (2022) y Rahim et al. (2021) utilizaron bosques aleatorios para predecir enfermedades cardíacas, encontrando que esta técnica superaba a otros métodos como la regresión logística en términos de precisión y capacidad de interpretación, además se evidenció su capacidad para identificar patrones complejos y relaciones entre múltiples variables. De acuerdo con (Breiman, 2001) los árboles de decisión se representan por medio de las siguientes ecuaciones:

Tabla 6 ECUACIONES DEL BOSQUE ALEATORIO

Para clasificación	**Para regresión**
$\hat{y} = mode\{h_1(x), h_2(x), \dots, h_B(x)\}$	$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} h_i(x)$
En donde: ▪ $\hat{y}$: predicción final de la clase. ▪ $h_i(x)$: predicción del i-ésimo árbol para la entrada xxx. ▪ B: número total de árboles en el bosque. ▪ Mode: denota la clase más frecuente entre las predicciones de los árboles.	En donde: ▪ $\hat{y}$: predicción final. ▪ $h_i(x)$: predicción del i-ésimo árbol para la entrada x. ▪ B: número total de árboles en el bosque.

Por otro lado, se escogió a las redes neuronales pues Mir y Sunanda (2023) y Bailly et al. (2022) subrayan que las redes neuronales son extremadamente eficaces para modelar relaciones no lineales complejas entre las variables de entrada, lo que es esencial en la predicción de enfermedades cardíacas donde múltiples factores interactúan de manera compleja. Además, los estudios de Katarya & Meena (2021) y Ansari et al. (2023) demostraron que las redes neuronales artificiales superan a otros algoritmos como KNN y Naive Bayes en términos de precisión predictiva para enfermedades cardíacas, esta técnica captura mejor las complejidades y no linealidades inherentes en los datos de salud y se destacó su capacidad para manejar grandes volúmenes de datos y aprender patrones complejos. Según Haykin (1998) las ecuaciones de redes neuronales son:

Tabla 7 ECUACIONES DE LAS REDES NEURONALES

Capa individual	**Capa completa**
$a_j = f\left(\sum_{i=1}^{n} W_{ij}x_i + b_j\right)$	$a = f(Wx + b)$
▪ x_i: entradas a la neurona. ▪ w_{ij}: pesos asociados a las entradas x_i de la neurona j. ▪ b_j: sesgo (bias) de la neurona j. ▪ f: función de activación.	▪ a: vector de salidas de la capa. ▪ W: matriz de pesos de la capa. ▪ x: vector de entradas a la capa. ▪ b: vector de sesgos de la capa. ▪ F: se aplica de manera elemento a elemento al vector resultante.
Red neuronal completa	**Red neuronal multicapa**
$a^{(l)} = f^{(l)}\left(W^{(l)}a^{(l-1)} + b^{(l)}\right)$	$Entrada: a^{(0)} = x$ $Capa\ 1: a^1 = f^{(1)}\left(W^{(1)}a^{(0)} + b^{(1)}\right)$ $Capa\ 2: a^2 = f^{(2)}\left(W^{(2)}a^{(1)} + b^{(2)}\right)$ $Capa\ L: a^L = f^{(L)}\left(W^{(L)}a^{(L-1)} + b^{(L)}\right)$
▪ $a^{(l)}$: vector de activaciones de la capa l. ▪ $W^{(l)}$: matriz de pesos de la capa l. ▪ $a^{(l-1)}$: vector de activaciones de la capa l−1l. ▪ $b^{(l)}$: vector de sesgos de la capa l. $f^{(l)}$: función de activación de la capa l.	

RESULTADOS

ANÁLISIS DE MODELO ESTADÍSTICO

A continuación, mediante la herramienta Colab y utilizando el lenguaje de programación de Phyton, se aplicarán los modelos de aprendizaje escogidos para evaluar el rendimiento de ambos modelos mediante métricas adecuadas, con el objetivo de seleccionar el modelo más preciso. Para el desarrollo de los modelos predictivos para la detección de ataques cardíacos, se va a utilizar un enfoque exhaustivo que involucra el uso de todas las variables disponibles en el conjunto de datos, esta decisión de incluir todas las variables fue motivada por la necesidad de capturar la mayor cantidad de información posible, con el fin de maximizar la precisión y robustez de los modelos predictivos.

1. BOSQUE ALEATORIO

Para la aplicación de este modelo los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%). Posteriormente se configuró el modelo de bosque aleatorio con la elección de los parámetros como el número de árboles en el bosque de 100, esto debido a que según Probst et al. (2019) este valor proporciona un buen equilibrio entre el rendimiento del modelo y el costo computacional. Adicional se incluyó el parámetro de la profundidad máxima de

cada árbol para evitar el sobreajuste, en este caso el valor es de 10 para mantener el modelo manejable, manteniendo la complejidad para modelar adecuadamente las relaciones de los datos.

Tabla 8 MODELO DE BOSQUE ALEATORIO

Accuracy: 0.6411865373645179				
Classification Report:				
	Precision	**Recall**	**F1-score**	**Support**
0	0.64	1.00	0.78	1125
1	0.33	0.00	0.00	628
Accuracy			0.64	1753
Macro avg	0.49	0.50	0.39	1753
Weighted avg	0.53	0.64	0.50	1753

En base a los resultados obtenidos se puede evidenciar que el modelo identifica correctamente todos los casos de la clase 0 con un recall de 1.00, pero con una precisión moderada de (0.64) lo que sugiere la existencia de algunos falsos positivos. Por otro lado, el modelo tiene dificultades significativas para identificar correctamente los casos de la clase 1, pues tiene un recall de 0.00 y un F1-score de 0.00, lo que quiere decir que el modelo no está capturando los casos de la clase 1 en absoluto. Este caso podría deberse a un desbalance en las clases del conjunto de datos, donde la clase 1 está subrepresentada, o a la necesidad de ajustar más los parámetros del modelo. Por lo que se van a aplicar técnicas de balanceo de datos y ajuste de hiperparámetros para mejorar el rendimiento del modelo en la clase minoritaria.

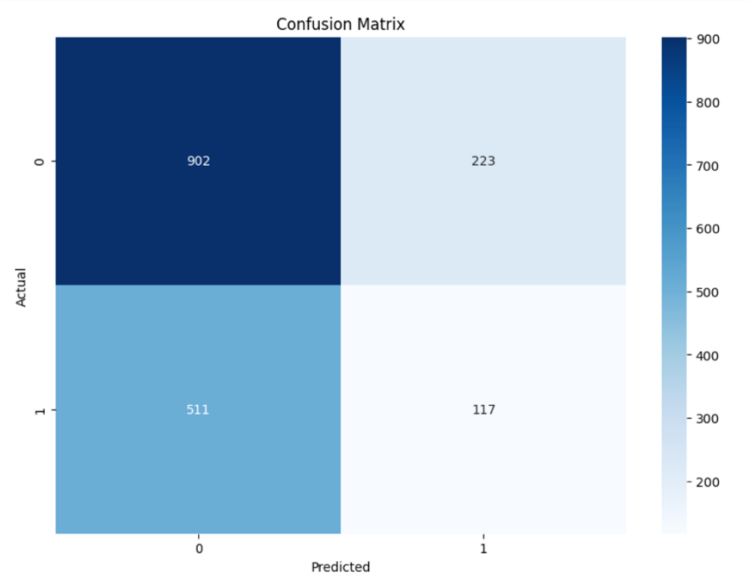
Tabla 9 MODELO DE BOSQUE ALEATORIO AJUSTADO

Fitting 3 folds for each of 216 candidates, totalling 648 fits Best Parameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}				
Best Grid Search Score: 0.6437299097975914 **Accuracy:** 0.5812892184826013				
Classification Report:				
	Precision	**Recall**	**F1-score**	**Support**
0	0.64	0.80	0.71	1125
1	0.34	0.19	0.24	628
Accuracy			0.58	1753
Macro avg	0.49	0.49	0.48	1753
Weighted avg	0.53	0.58	0.54	1753

Para balancear las clases del modelo y aumentar la representación de la clase minoritaria se utilizó SMOTE, además, se optimizaron los hiperparámetros a través de GridSearchCV en donde se ajustaron los valores de número de árboles, profundidad máxima, tamaño mínimo de las hojas y divisiones, y uso de bootstrap. Estos ajustes se realizaron con el propósito de optimizar el modelo de Random Forest para obtener el mejor desempeño posible en la predicción de "Riesgo AC", asegurando al mismo tiempo que el modelo maneje adecuadamente el desbalance en las clases.

Los resultados del modelo muestran que su desempeño es moderado para la clase '0' (no riesgo de ataque cardíaco), con una precisión del 64% y un recall del 80%, lo que indica que identifica correctamente la mayoría de los casos sin riesgo. Sin embargo, el rendimiento es significativamente peor para la clase '1' (riesgo de ataque cardíaco), con una precisión del 34% y un recall del 19%, reflejando una capacidad limitada para identificar correctamente a los individuos en riesgo. La exactitud general del modelo es del 58,12%, lo que sugiere que más de la mitad de las predicciones totales son correctas, pero aún deja mucho que desear en términos de confiabilidad y efectividad, especialmente para los casos críticos de riesgo. Al comparar estos resultados con la literatura se confirma que el modelo es poco confiable pues en el estudio de Rubini et al. (2021) en donde el objetivo de estudio fue similiar el modelo tuvo una precisión del 84,81%.

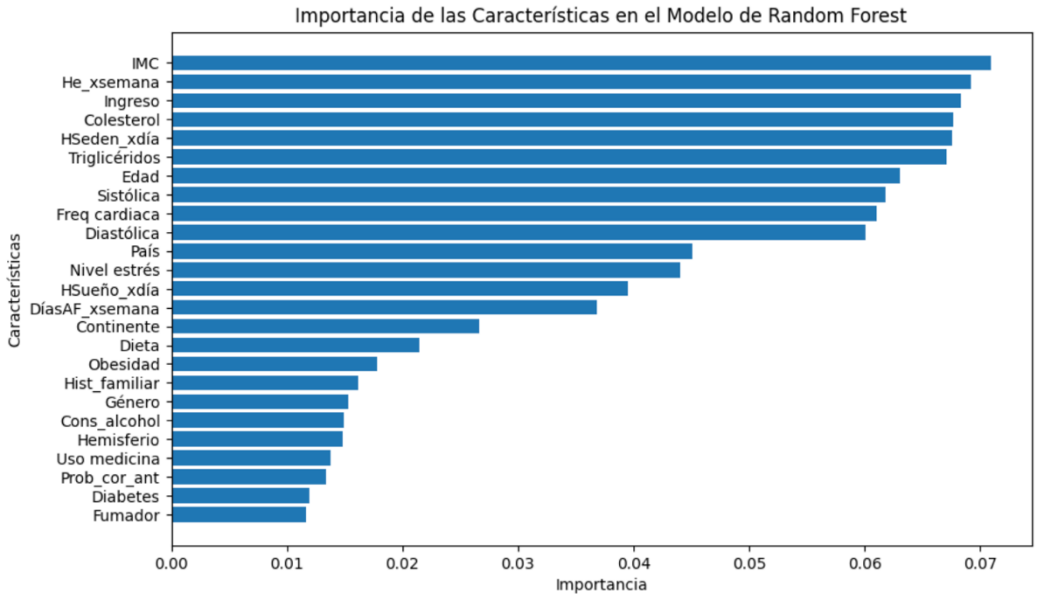
Ilustración 6 MATRIZ DE CONFUSIÓN



Se realizó la matriz de confusión que es utilizada para describir el rendimiento de un modelo de clasificación. En este caso, la matriz es de 2x2 ya que se trata de un problema de clasificación binaria (dos clases: 0 y 1). Los resultados de esta grpafica muestran que el modelo tiene una alta cantidad de verdaderos negativos (902) comparado con los verdaderos positivos (117), lo que indica que es bastante bueno identificando correctamente los casos de la clase 0. Sin embargo, tiene un número significativo de falsos negativos (511), lo que indica que el modelo tiene dificultades para identificar correctamente los casos de la clase 1.

Posterior a la aplicación del modelo se consideró esencial evaluar la contribución individual de cada variable en el desempeño del modelo. Para ello, se generó un gráfico por modelo donde se visualiza la importancia de las características para cada modelo. Este análisis permitió identificar cuáles variables tenían un mayor impacto en las predicciones del riesgo de ataques cardíacos, lo que es crucial para comprender el funcionamiento interno de los modelos y para la interpretación de sus resultados.

Ilustración 7 IMPORTANCIA DE LAS CARACTERÍSTICAS EN EL MODELO RANDOM FOREST



El gráfico obtenido muestra la importancia relativa de cada característica utilizada en el modelo de Random Forest. En primer lugar se evidencia que variables como el IMC, H. y He_xsemana son las características más

importantes, indicando que tienen el mayor impacto en las predicciones del modelo, mientras que las variables como Fumador y Diabetes tienen una menor importancia, lo que sugiere que influyen menos en las decisiones del modelo.

2. REDES NEURONALES

La aplicación de este modelo comenzó con la separación de las características de la variable objetivo (riesgo de ataque cardíaco). Luego, se escalaron las características y se dividieron en conjuntos de entrenamiento y prueba (80% y 20%) y se normalizaron los datos. Además se especificaron el número de épocas y el tamaño del lote, en primer lugar se definió el número de épocas de 50 pues a partir de estudios anteriores se ha evidenciado que este parámetro permite a la red ajustar sus pesos iterativamente para minimizar la función de pérdida y asegurar que el modelo converge sin sobreajustarse, también se estableció un tamaño de lote de 32 pues este valor es comúnmente utilizado debido a que proporciona un equilibrio entre eficiencia computacional y capacidad de actualización frecuente del modelo, esto ayuda a encontrar un buen mínimo de la función de pérdida.

Se implementó la validación de prueba el cual ayuda a monitorear el rendimiento del modelo y detectar sobreajuste. El modelo se compiló con el optimizador Adam el cual según García (2022) es adecuado para entrenar modelos de redes neuronales, permitiendo que los desarrolladores logren buenos resultados con menos necesidad de ajuste fino de hiperparámetros.

Una vez establecido los valores anteriores se procedió a definir la red neuronal desde cero con una capa oculta de 64 neuronas y una capa de salida con una neurona. Se decidió correr el modelo con 64 neuronas puesto que según la literatura este valor es una configuración estándar que ha demostrado ser eficaz en muchos problemas de clasificación y regresión. Adicional se utilizó la función de activación ReLU la cual introduce no linealidades, la capa de Dropout que previene el sobreajuste. Finalmente, la capa de salida con activación sigmoide es útil en problemas de clasificación binaria pues produce una probabilidad entre 0 y 1.

Tabla 10 MODELO DE RED NEURONAL CON UNA CAPA

Loss: 0.6644670963287354		**Accuracy:** 0.6371933817863464		
	Precision	**Recall**	**F1-score**	**Support**
0	0.64	1.00	0.78	1125
1	0.33	0.00	0.00	628
Accuracy			0.64	1753
Macro avg	0.49	0.50	0.39	1753
Weighted avg	0.53	0.64	0.50	1753

Los resultados muestran que se tiene una pérdina de 0.664, lo que quiere decir que el modelo no está prediciendo correctamente las etiquetas verdaderas. Por otra parte, se tiene una precisión general del 63.7% con un buen rendimiento para la clase 0 (sin riesgo de ataque cardíaco) con una precisión de 64% y un recall de 100%. Sin embargo, el modelo no tiene un buen rendimiento para la clase 1 (riesgo de ataque cardíaco) pues tiene una precisión de 33% y un recall de 0%, esto indica que el modelo no está identificando correctamente ningún caso positivo de riesgo de ataque cardíaco. Estos resultados sugieren la necesidad de mejorar el modelo para detectar con mayor precisión la clase 1, por lo que se va a realizar ajustes en la arquitectura del modelo.

Tabla 11 MODELO DE RED NEURONAL CON DOS CAPAS

Loss: 0.658257782459259		**Accuracy:** 0.6423274278640747		
	Precision	**Recall**	**F1-score**	**Support**
0	0.64	1.00	0.78	1125
1	1.00	0.00	0.00	628
Accuracy			0.64	1753
Macro avg	0.82	0.50	0.39	1753
Weighted avg	0.77	0.64	0.50	1753

Para el segundo modelo de red neuronal se añadió una segunda capa oculta con 32 neuronas y activación ReLU, se utilizó Dropout con una tasa de 0.5 para prevenir el sobreajuste y se mantuvo al optimizador Adam.

Los resultados del segundo modelo indican que se tiene una precisión general del 64.2% y una pérdida de 0.658. El modelo sigue siendo efectivo en la predicción de la clase 0 (sin riesgo de ataque cardíaco), con una precisión del

64% y un recall del 100%, logrando un buen equilibrio entre precisión y recall el cual se refleja en un F1-Score de 0.78. Sin embargo, el rendimiento del modelo en la predicción de la clase 1 (riesgo de ataque cardíaco) sigue siendo deficiente, con un recall de 0%, es decir, el modelo sigue sin identificar correctamente ningún caso positivo de riesgo de ataque cardíaco. Aunque la precisión para la clase 1 es del 100%, el F1-Score y el recall es de 0.00 lo que resulta ser engañoso. En téminos generales, a pesar de la incorporación de una segunda capa oculta el modelo sigue sin poder detectar correctamente los casos de riesgo de ataque cardíaco. En base a los resultados obtenidos de los primeros dos modelos de red neuronal se va a ir aumentando capas y neuronas con el objetivo de encontrar al modelo que tenga más balance entre clases y a su vez una mayor precisión.

Tabla 12 MODELOS DE RED NEURONAL

Modelo/métricas	Neuronas				
	64	64-32	128-64-32	128-64-32-16	128-64-32-16-8
	1 capa	2 capas	3 capas	4 capas	5 capas
P. General	0,63	0,64	0,57	0,57	0,58
Precisión (0)	0,64	0,64	0,63	0,64	0,64
Precisión (1)	0,33	1,00	0,33	0,36	0,33
Recall (0)	1,00	1,00	0,80	0,76	0,83
Recall (1)	0,00	0,00	0,17	0,24	0,15
F1 Score (0)	0,78	0,78	0,71	0,70	0,72
F1 Score (1)	0,00	0,00	0,23	0,29	0,21

En cada modelo se utilizó la activación ReLu, Sigmoid para cada capa de salida, optimizador Adam y en el caso de Dropout se empezó a usar una tasa de 0.3 a partir del tercer modelo para encontrar un equilibrio más adecuado entre regularización y capacidad de aprendizaje bajo el contexto de tener conjunto de datos más equilibrado.

De esta manera, se corrió el modelo hasta 5 capas, en donde se determinó que el modelo que presentó un mayor equilibrio en las métricas de rendimiento para ambas clases es el modelo con 4 capas de 128, 64, 32 y 16 neuronas, pues se tuvo una precisión de 0.36 y un recall de 0.24 para la clase 1, resultando en el F1 Score más alto con 0.29 para esta clase en comparación con los otros

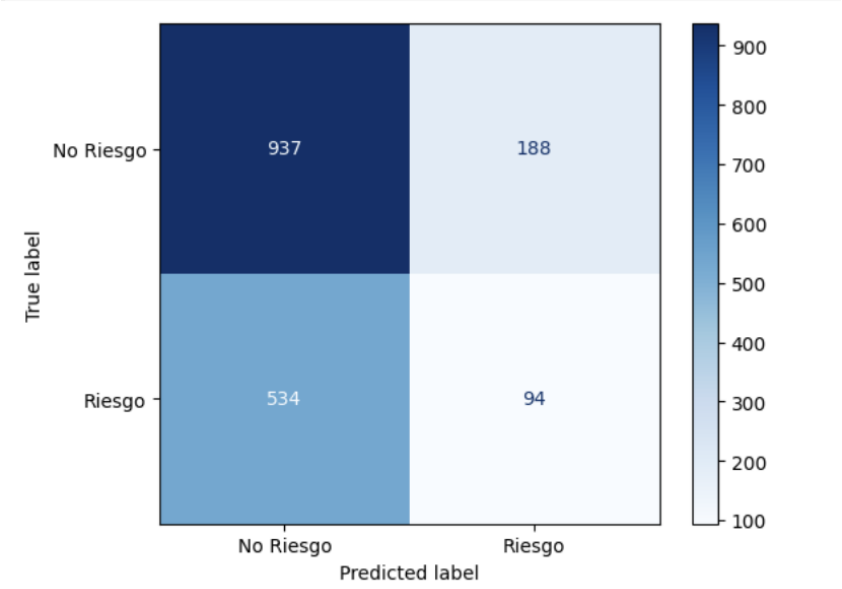
modelos. Además a comparación de los modelos anteriores se tuvo una mejora en el recall y F1 Score para la clase 1. En un estudio similar de Sakshi et al. (2019) la precisión de esta técnica fue del 98,58% evidenciando que el modelo tiene problemas debido al desbalance de clases y la complejidad de los datos.

Ilustración 8 MODELO DE RED NEURONAL CON CUATRO CAPAS

Loss: 0.7118188738822937		**Accuracy:** 0.5761551856994629		
	Precision	**Recall**	**F1-score**	**Support**
0	0.64	0.76	0.70	1125
1	0.36	0.24	0.29	628
Accuracy			0.58	1753
Macro avg	0.50	0.50	0.49	1753
Weighted avg	0.54	0.58	0.55	1753

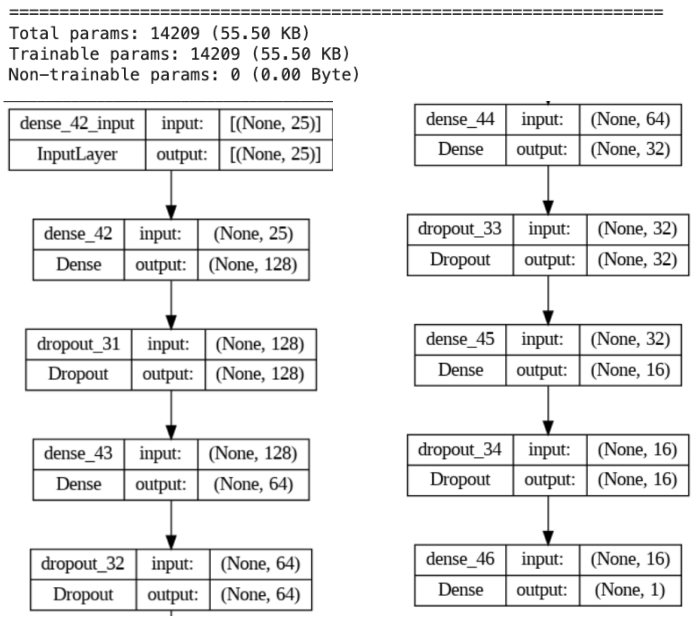
Al elegir al modelo de red neuronal con cuatro capas ocultas se realizó una matriz de confusión para describir el rendimiento del un modelo.

Ilustración 9 MATRIZ DE CONFUSIÓN



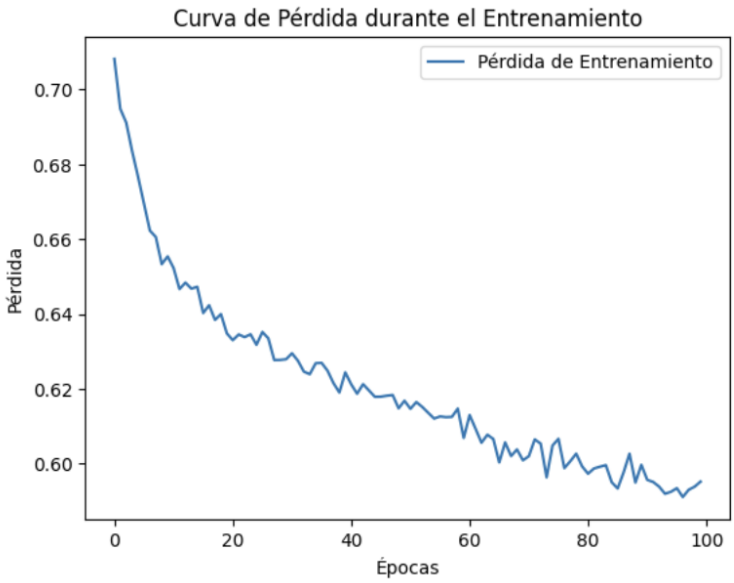
La matriz de confusión a partir del modelo de red neuronal con 4 capas ocultas evidencia que el modelo tiene un rendimiento aceptable en la predicción de la clase 0 (sin riesgo) con 937 verdaderos negativos y 188 falsos positivos. Sin embargo, se confirma que el modelo tiene dificultades significativas para detectar a la clase 1 (con riesgo) con 94 verdaderos positivos y 534 falsos negativos.

Ilustración 10 ESTRUCTURA DE LA RED NEURONAL



La gráfica de la estructura de red neuronal muestra que el modelo tiene múltiples capas densas con un número decreciente de neuronas (128, 64, 32, 16) y capas de Dropout intercaladas con una tasa de 0.3 para prevenir el sobreajuste y la capa de salida tiene una sola neurona con activación Sigmoid para la clasificación binaria. Se tiene 14.209 parámetros entrenables, lo que indica que el modelo tiene una complejidad moderada.

Ilustración 11 CURVA DE PÉRDIDA DE ENTRENAMIENTO



La curva de pérdida durante el entrenamiento muestra una disminución constante, lo que quiere decir que el modelo está aprendiendo y ajustando bien sus parámetros.

DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

En la investigación realizada se emplearon técnicas de aprendizaje automático para predecir el riesgo de ataques cardíacos utilizando un conjunto de datos diverso y extenso. Se implementaron dos modelos principales: Bosque Aleatorio y Redes Neuronales, cada uno con sus respectivos resultados y análisis de rendimiento.

S. En primer lugar se debería seguir utilizando técnicas de balanceo de datos como SMOTE para asegurar una representación más equitativa de ambas clases, además de realizar ajustes más finos en los hiperparámetros de los modelos utilizando técnicas avanzadas como Random Search o Bayesian Optimization para encontrar configuraciones óptimas.

ELECCIÓN DEL MEJOR MODELO

Se realizó una tabla comparativa sobre los principales resultados del modelo para identificar, cuál de los dos modelos es el que más se ajusta al obejtivo del estudio.

Tabla 13 COMPARACIÓN DE MODELOS

Bosque Aleatorio ajustado				**Redes Neuronales con cuatro capas**			
Clase 0 (sin riesgo de ataque cardíaco)		Clase 1 (con riesgo de ataque cardíaco)		Clase 0 (sin riesgo de ataque cardíaco)		Clase 1 (con riesgo de ataque cardíaco)	
Precisión	**Recall**	**Precisión**	**Recall**	**Precisión**	**Recall**	**Precisión**	**Recall**
64%	80%	34%	19%	64%	76%	36%	24%
Exactitud general: 58.12%				**Exactitud general:** 57.61%			
Problemas identificados: Alto número de falsos negativos y moderada precisión para la clase 0				**Problemas identificados:** Dificultad en identificar correctamente los casos con riesgo, alto número de falsos negativos.			

Ambos modelos presentan resultados similares en términos de precisión y recall para la clase 0, sin embargo, ambos tienen dificultades significativas para identificar correctamente los casos de riesgo (clase 1). A continuación se resaltan algunas comparaciones clave:

El modelo de Bosque aleatorio tiene un mejor recall para la clase 0 con el 80%, además este modelo tiene una exactitud ligeramente mejor que Redes Neuronales con el 58.12%. Por otra parte, el modelo de Redes Neuronales presenta un mejor recall para la clase 1 con el 24%. Considerando que la prioridad bajo el contexto de predicción de ataques cardíacos es identificar casos de alto riesgo, se elige al modelo de Redes Neuronales con cuatro capas pues a comparación del modelo de Bosque Aleatorio presenta una mayor capacidad para identificar casos de alto riesgo con una mayor precisión y recall. Sin embargo, ambos modelos requieren ajustes adicionales para mejorar la precisión y reducir el número de falsos negativos.

CONCLUSIONES Y RECOMENDACIONES

Teniendo en cuenta que el problema de estudio se centra en la necesidad de un sistema predictivo robusto y preciso para identificar individuos en alto riesgo de sufrir un ataque cardíaco superando las limitaciones de los métodos tradicionales que no consideran la compleja interacción de múltiples factores de riesgo, se puede concluir que la implementación de modelos de aprendizaje automático, como Bosque Aleatorio y Redes Neuronales aborda esta necesidad al proporcionar herramientas avanzadas para el análisis y la predicción del riesgo cardíaco.

Los resultados de la investigación muestran que, aunque los modelos actuales tienen dificultades para predecir con precisión los casos de alto riesgo, existen oportunidades significativas para mejorar su rendimiento mediante el balanceo de datos, optimización de hiperparámetros, y la incorporación de datos adicionales. La propuesta de soluciones avanzadas y personalizadas basadas en estos modelos puede contribuir de manera efectiva a la prevención y tratamiento de enfermedades cardíacas, mejorando la calidad de vida de los pacientes y optimizando los recursos en el sector de la salud.

Implementar estas mejoras no solo permitirá una predicción más precisa del riesgo de ataques cardíacos, sino que también facilitará la intervención temprana y personalizada, lo que es crucial para reducir la mortalidad y mejorar la calidad de vida de los pacientes. La capacidad de predecir con precisión el riesgo de ataques cardíacos permite a los profesionales de la salud intervenir de manera proactiva, implementando estrategias de prevención y tratamiento personalizadas que pueden salvar vidas y mejorar la calidad de vida de los pacientes.

En el contexto de Ecuador, según el Instituto Nacional de Estadísticas y Censos, (2024) las enfermedades cardiovasculares son una de las principales causas de mortalidad, por lo que la implementación de modelos predictivos basados en aprendizaje automático como Bosque Aleatorio y Redes Neuronales puede transformar la prevención y el tratamiento de los ataques cardíacos. Se recomienda utilizar estos modelos en el sistema de salud ecuatoriano, entrenándolos con datos locales para mejorar su precisión y relevancia. Además se propone la capacitación del personal médico para su uso e interpretación y a

su vez establecer un sistema de monitoreo y evaluación continua para ajustar y mejorar los modelos. La adopción de estas técnicas de aprendizaje automático en la predicción de ataques cardíacos no solamente aumentaría la eficiencia y efectividad del sistema de salud, sino que también ayudaría a reducir la mortalidad de la población ecuatoriana.

Para futuras investigaciones sobre la predicción de ataques cardíacos, específicamente en Ecuador, se considera importante incorporar una variedad de variables que reflejen tanto características globales como particulares de la población ecuatoriana. Aparte de las variables ya estudiadas en este trabajo se recomienda incluir datos demográficos adicionales como etnia, nivel de educación y ocupación, además, los factores médicos deberían abarcar el historial de hipertensión familiar, frecuencia cardíaca en reposo y medicación actual. Incluir otros factores socioeconómicos también es importante el acceso a servicios de salud, seguro de salud y condiciones de vivienda. En el caso de las variables geográficas se debería incluir la región de residencia (Costa, Sierra, Amazonía, Galápagos) y altitud de residencia. Por último, para un análisis más profundo se debería aumentar factores culturales y psicosociales como el apoyo social y familiar, así como las creencias y prácticas relacionadas con la salud.

Por otra parte, en base al estudio realizado se considera importante para reducir el riesgo de tener un ataque cardíaco las personas deberían adoptar un estilo de vida saludable y gestionar adecuadamente los factores de riesgo. Se recomienda mantener una dieta equilibrada, realizar actividad física regularmente, mantener un peso corporal saludable, dejar de fumar, controlar la presión arterial, gestionar el y controlar los niveles de colesterol y triglicéridos las cuales son medidas esenciales para reducir el riesgo de enfermedades cardíacas.

Finalmente, para mejorar la precisión y eficacia de los modelos predictivos, se proponen las siguientes recomendaciones. En primer lugar, se recomienda continuar investigando y aplicando algoritmos de aprendizaje automático más avanzados, como modelos de ensamble y redes neuronales convolucionales, para capturar mejor la complejidad y no linealidades de los datos e implementar métodos de validación cruzada y técnicas de evaluación robustas para asegurar que los modelos generalicen bien a nuevos datos y no estén sobreajustados a los datos de entrenamiento.

REFERENCIAS

Ansari, G., Bhat, S., Ansari, M. D., Ahmad, S., Nazeer, J., & Eljialy, A. E. M. (2023). Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction. *Computational and Mathematical Methods in Medicine*, *2023*. https://doi.org/10.1155/2023/8191261

Ansari, M., Alankar, B., & Kaur, H. (2021). A prediction of heart disease using machine learning algorithms. *Advances in Intelligent Systems and Computing*, *1200 AISC*, 497–504. https://doi.org/10.1007/978-3-030-51859-2_45

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC Heart Failure*, *6*(2), 428–435. https://doi.org/10.1002/ehf2.12419

Bailly, A., Blanc, C., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models. *Computer Methods and Programs in Biomedicine*, *213*. https://doi.org/10.1016/j.cmpb.2021.106504

Breiman, L. (2001). *Random Forests* (Vol. 45). https://doi.org/https://doi.org/10.1023/A:1010933404324

Ganesan, M., & Sivakumar, N. (2019). *IoT based heart disease prediction and diagnosis model for healthcare using machine learning models*. https://doi.org/10.1109/ICSCAN.2019.8878850

García, R. (2022). *El perceptrón: Una red neuronal artificial para clasificar datos*. http://www.economicas.uba.ar/institutos_y_centros/revista-modelos-matematicos/

Haykin, S. (1998). *Neural Networks - A Comprehensive Foundation*.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access*, *9*, 39707–39716. https://doi.org/10.1109/ACCESS.2021.3064084

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*(1). https://doi.org/10.1088/1757-899X/1022/1/012072

Katarya, R., & Meena, S. K. (2021). Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health and Technology*, *11*(1), 87–97. https://doi.org/10.1007/s12553-020-00505-7

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H. J., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-72685-1

Mir, S., & Sunanda, D. (2023). *A Comprehensive Review of Recent Advances in Heart Disease Prediction using Machine Learning Algorithms with Optimization Techniques and Feature Selection*.

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

Organización Mundial de la Salud (OMS). (2024). *Cardiovascular diseases*.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 9, Issue 3). Wiley-Blackwell. https://doi.org/10.1002/widm.1301

Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases. *IEEE Access*, *9*, 106575–106588. https://doi.org/10.1109/ACCESS.2021.3098688

Ramesh, T. R., Lilhore, U. K., Poongodi, M., Simaiya, S., Kaur, A., & Hamdi, M. (2022). Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science*, *2022*(Special Issue 1), 132–148. https://doi.org/10.22452/mjcs.sp2022no1.10

Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263–275. https://doi.org/10.1007/s40860-021-00133-6

Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., & Pranavanand, S. (2021). Heart disease risk prediction using machine learning classifiers with attribute evaluators. *Applied Sciences (Switzerland)*, *11*(18). https://doi.org/10.3390/app11188352

Rubini, P., Subasini, C., Katharine, V., Kumaresan, V., Gowdhamkumar, S., & Nithya, T. (2021). *A Cardiovascular Disease Prediction using Machine Learning Algorithms* (Vol. 25, Issue 2). http://annalsofrscb.ro

Wu, C. S., Badshah, M., & Bhagwat, V. (2019). Heart disease prediction using data mining techniques. *ACM International Conference Proceeding Series*, 7–11. https://doi.org/10.1145/3352411.3352413

ANEXOS

ANEXO No. 1: Matriz de investigaciones anteriores

Autor	Título	Objetivos de la investigación	Problema abordado	Fuente de datos utilizados	Metodología implementada	Técnica utilizada	Métricas de precisión	Resultados	Modelo más preciso	Implicaciones
Ramesh T, Lilhore U, Poongodi M, Simaiya S, Kaur A, Hamndi M (2022)	Análisis predictivo de enfermedades cardiacas con enfoques de Machine Learning	Utilizar enfoques de machine learning para realizar un análisis predictivo de enfermedades cardíacas.	Necesidad de mejorar la detección temprana y la precisión en la predicción de enfermedades cardíacas.	Conjunto de datos UCI en línea con 303 filas y 76 propiedades.	Preprocesamiento de datos, evaluación, selección de características, entrenamiento de modelos, comparación de resultados.	Regresión logística, máquinas de vectores de soporte (SVM), árboles de decisión, bosques aleatorios y k-vecinos más cercanos (KNN).	Precisión, recuperación (Sensibilidad), puntuación F1, AUC (Área bajo la curva ROC).	Se identificaron características críticas que influyen en la predicción de enfermedades cardíacas, como la presión arterial alta, la diabetes, el colesterol alto, la obesidad, el tabaco y los antecedentes familiares de enfermedades cardíacas.	Bosque Aleatorio y el KNN	Mejora en la detección temprana, personalización de la atención médica, optimización de recursos, investigación futura, impacto en la salud pública
Sobia Mir, Dr. Sunanda (2023)	Una revisión completa de los avances recientes en la predicción de enfermedades cardíacas utilizando algoritmos de aprendizaje automático	Explorar la detección temprana de enfermedades cardíacas mediante técnicas de aprendizaje automático.	Los métodos convencionales de detección y predicción de enfermedades cardíacas pueden ser costosos, y pueden consumir mucho tiempo y producir resultados inexactos.	Conjunto de datos de Z Alizadeh Sani, que consta de 303 pacientes con variables de entrada relacionadas con el perfil del paciente, datos demográficos, examen físico y pruebas de laboratorio.	Aplicación de técnicas de aprendizaje automático, algoritmos de selección de características y modelos de predicción.	Máquina de Vectores de Soporte (SVM), modelo de apilamiento, clasificador de refuerzo de gradiente (GBC) y técnicas de selección de características.	Precisión de entrenamiento y prueba, valor AUC (Área bajo la curva ROC).	Se destacan varios modelos y enfoques propuestos que han demostrado ser eficaces en la detección y predicción de enfermedades cardíacas con altos niveles de precisión.	Clasificador de Refuerzo de Gradiente (GBC)	Detección temprana, reducción de la mortalidad, mejora en la atención médica.
Abid Ishaq, Salma Sadiq , Muhammad Umer , Saleem Ullah, Seyedali Mirjalili, Valbhav Ruparara, Michelle Nappi (2021)	Mejora de la predicción de la insuficiencia cardíaca. Supervivencia de los pacientes utilizando SMOTE y Técnicas efectivas de minería de datos	Analizar características importantes y técnicas de extracción de datos eficaces para aumentar la precisión de la predicción de supervivencia de pacientes cardiovasculares.	Dificultad de identificar la enfermedad cardiovascular debido a diversos factores contribuyentes, como la presión arterial alta, el nivel de colesterol, la diabetes y otros.	Registros clínicos de pacientes con insuficiencia cardíaca recopilados previamente en el Instituto de Cardiología y un hospital aliado en Faisalabad, Pakistán.	Se emplearon nueve modelos de aprendizaje automático, se comparó el rendimiento de estos modelos utilizando la técnica SMOTE para abordar el desequilibrio de clases en los datos.	Sobremuestreo de minorías sintéticas (SMOTE), árboles de decisión, modelo de impulso adaptativo (AdaBoost), regresión logística y descenso de gradiente estocástico (SGD)	Precisión (Accuracy), Recuerdo (Recall), Puntuación F (F Score)	La investigación muestra que el uso de técnicas avanzadas de minería de datos y aprendizaje automático, junto con el empleo de modelos como árboles de decisión, AdaBoost, regresión logística, entre otros, ha mejorado significativamente la predicción de la supervivencia de pacientes con insuficiencia cardíaca.	Random Forest	Impacto positivo en la atención médica, la reducción de la mortalidad y el avance en la investigación médica en el campo de las enfermedades cardiovasculares.
Gufran Ahmad Ansari, Salliah Shafi Bhat, Mohd Dilshad Ansari , Sultan Ahmad , Jabeen Nazeer , and A. E. M. Eljialy (2023)	Evaluación del desempeño de técnicas de aprendizaje automático (MLT) para la predicción de enfermedades cardíacas	Mejorar la precisión en la predicción de enfermedades cardíacas mediante la comparación y evaluación exhaustiva de diferentes algoritmos de aprendizaje automático.	Necesidad de mejorar la precisión en la predicción de enfermedades cardíacas utilizando técnicas de aprendizaje automático. Dado que las enfermedades cardíacas representan una de las principales causas de mortalidad a nivel mundial.	Se usó el "Dataset for Cleveland Heart" de la Universidad de California en Irvine (UCI), que contiene atributos relevantes para el estudio de enfermedades cardíacas, como la edad, el sexo, la presión arterial, el colesterol, entre otros	El conjunto de datos se dividió en 70% para entrenamiento y 30% para pruebas, utilizando diversas tecnologías de aprendizaje automático como el algoritmo de bosque aleatorio, regresión logística, Naive Bayes, redes neuronales, entre otros.	Naive Bayes Redes Neuronales Árboles de Decisión Vecinos más Cercanos (KNN) Máquinas de Vectores de Soporte (SVM) Bosques Aleatorios (Random Forest)	Precisión Recuperación (Sensibilidad) Medida F Coeficiente de correlación de Matthews (MCC)	Los resultados demostraron la eficacia de los algoritmos de aprendizaje automático en la predicción de enfermedades cardíacas, con el NB destacando como el más preciso en este contexto	Naive Bayes	Mejora en la precisión del diagnóstico, la personalización de tratamientos según el riesgo individual de cada paciente y la reducción de costos en la atención médica mediante medidas preventivas efectivas.
Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath (2021)	Predicción de enfermedades cardíacas mediante algoritmos de aprendizaje automático	Mejorar la detección, diagnóstico y tratamiento de enfermedades cardíacas mediante tecnologías avanzadas de análisis de datos y aprendizaje automático.	Necesidad de mejorar la detección temprana, el diagnóstico preciso y el tratamiento efectivo de enfermedades cardiovasculares.	Historiales médicos de pacientes con datos como la edad, presión arterial, nivel de azúcar en sangre, antecedentes familiares, síntomas, entre otros.	Recopilación de historiales médicos, preprocesamiento de datos, la extracción de variables relevantes, entrenamiento de algoritmos, evaluación del modelo y comparación de la eficacia de los algoritmos.	Regresión Logística, K-Nearest Neighbors (KNN), Clasificador de Bosque Aleatorio	Precisión, sensibilidad (Recall), especificidad, valor F1	Se logró una precisión satisfactoria en la predicción de enfermedades cardíacas en pacientes utilizando técnicas como regresión logística, KNN y clasificador de bosque aleatorio. Se destacó que el algoritmo KNN fue el más eficiente, con una precisión del 88.52%.	K-Nearest Neighbors (KNN)	Impacto positivo en la prevención, diagnóstico y tratamiento de estas afecciones, mejorando la calidad de la atención médica y la salud cardiovascular de la población.
Mohd Faisal Ansari, Bhavya AlankarKaur, and Harleen Kaur (2021)	Una predicción de enfermedades cardíacas mediante algoritmos de aprendizaje automático	Mejorar la capacidad de los profesionales de la salud para identificar y tratar enfermedades cardíacas de manera más efectiva mediante el desarrollo de métodos más precisos y eficientes para el diagnóstico temprano de estas enfermedades.	Se busca superar las limitaciones de los métodos tradicionales de predicción, como los algoritmos de minería de datos, mediante el uso de algoritmos de aprendizaje automático.	Registros médicos de pacientes obtenidos de bases de datos médicas, específicamente de la Historia Clínica Electrónica. Estos datos incluyen información relevante como la edad, presión arterial, niveles de colesterol, frecuencia cardíaca y otros atributos característicos.	Máquinas de vectores de soporte, regresión logística, bayes ingenuos, bosque aleatorio, árbol de decisión.	Análisis de Componentes Principales (PCA), Regresión Logística, Máquinas de Vectores de Soporte (SVM)	Precisión, Sensibilidad (Recall), Especificidad, Curva ROC (Receiver Operating Characteristic)	El modelo de Análisis de Componentes Principales (PCA) mostró una precisión del 86%, lo que lo convierte en el modelo más preciso en la clasificación y predicción de enfermedades cardíacas, lo que la posiciona como un modelo efectivo en la predicción de enfermedades cardíacas utilizando los atributos considerados en el estudio.	Análisis de Componentes Principales (PCA)	Los resultados de esta investigación tienen implicaciones importantes para la práctica clínica, la prevención de enfermedades cardíacas y el avance de la medicina personalizada, destacando el potencial de los algoritmos de aprendizaje automático en la mejora de la atención médica y la salud cardiovascular.

Rahul Katarya, Sunit Kumar Meena (2020)	Técnicas de aprendizaje automático para la predicción de enfermedades cardíacas: estudio y análisis comparativos	Proporcionar una predicción precisa y eficiente de la ocurrencia de enfermedades cardíacas utilizando diferentes algoritmos y modelos de aprendizaje automático.	Mejorar la predicción precisa y temprana de enfermedades cardíacas, dada su alta tasa de mortalidad a nivel mundial.	Conjuntos de datos clínicos que contienen información relevante sobre pacientes, como datos demográficos, antecedentes médicos, resultados de pruebas diagnósticas, factores de riesgo, entre otros.	Análisis comparativo de diferentes enfoques y modelos, empleo de técnicas de aprendizaje automático.	Redes Neuronales Artificiales (ANN), K Vecino Más Cercano (KNN), Arboles de Decisión, Máquinas de Vectores de Soporte (SVM), Naïve Bayes, Bosque Aleatorio (Random Forest)	Precisión, Recall (Recuperación o Sensibilidad), RMSE (Error Cuadrático Medio de la Raíz), MAE (Error Absoluto Medio).	Los modelos predictivos basados en redes neuronales artificiales, KNN, árboles de decisión, SVM y Naïve Bayes han mostrado resultados prometedores en la clasificación de casos y la identificación de patrones relevantes.	Bosque Aleatorio (Random Forest)	El uso de técnicas de aprendizaje automático para la predicción de enfermedades cardíacas tiene el potencial de impactar positivamente en la prevención, diagnóstico y tratamiento de estas afecciones, mejorando la calidad de vida de los pacientes.
A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar (2022)	Un marco integrado de aprendizaje automático para la predicción eficaz de enfermedades cardiovasculares	Abordar los desafíos específicos relacionados con la predicción de enfermedades cardiovasculares mediante un enfoque integrado de aprendizaje automático que busca mejorar la precisión y eficacia de los modelos predictivos.	Desequilibrio de clases en los datos de entrenamiento, la selección de características óptimas, la gestión de valores faltantes y la mejora de la precisión de los clasificadores.	Conjunto de datos de Framingham que consta de 16 características. Este conjunto de datos incluye información recopilada de personas de tres generaciones diferentes.	Desarrollo de un marco integrado de aprendizaje automático denominado "MaLCaDD" para la predicción eficaz de enfermedades cardiovasculares.	Regresión logística, Algoritmo de vecino más cercano (KNN)	Precisión	El marco integrado de aprendizaje automático MaLCaDD, especialmente con la combinación de regresión logística y KNN, es altamente fiable y eficaz en la predicción de enfermedades cardiovasculares.	Combinación de regresión logística (LR) y el algoritmo de vecino más cercano (KNN).	El uso de MaLCaDD puede mejorar la detección temprana de enfermedades cardiovasculares, permitiendo intervenir antes de la manifestación de síntomas, la capacidad de predecir enfermedades cardiovasculares de forma precisa permite adaptar la atención médica a las necesidades individuales de cada paciente.
Simon Nusinovicia , Yih Chung Thama, Marco Yu Chak Yana , Daniel Shu Wei Tinga, Jialiang Lia, Charumathi Sabanayagama , Tien Yin Wonga, Ching-Yu Chenga (2020)	La regresión logística fue tan buena como el aprendizaje automático para predecir las principales enfermedades crónicas	Evaluar el desempeño de algoritmos de aprendizaje automático (ML) y compararlos con la regresión logística para la predicción del riesgo de enfermedades cardiovasculares (ECV), enfermedad renal crónica (ERC), diabetes (DM) e hipertensión (HTA)	Evaluación del desempeño de la regresión logística en comparación con los modelos de aprendizaje automático para predecir el riesgo de enfermedades crónicas importantes, utilizando un conjunto limitado de predictores clínicos simples.	Estudio de cohorte poblacional en adultos asiáticos que incluyó 6,762 participantes, los datos se recopilaron a lo largo de un período de seis años.	Selección aleatoria de participantes de tres grupos étnicos al inicio del estudio mediante un método de muestreo estratificado por edad. Se realizaron exámenes de seguimiento aproximadamente seis años después de la visita inicial, con una tasa de respuesta del 78% entre los participantes elegibles.	Regresión logística, Red neuronal de una sola capa oculta, Máquina de vectores de soporte, Bosque aleatorio, Máquina de aumento de gradiente K vecino más cercano.	Bajo la Curva, ROC, Precisión, Sensibilidad, Especificidad, Precisión, Valor F1	Los resultados del estudio mostraron que la regresión logística fue tan efectiva como las técnicas de aprendizaje automático en la predicción del riesgo de enfermedades crónicas importantes, como enfermedades cardiovasculares, enfermedad renal crónica, diabetes e hipertensión.	Regresión logística	Simplicidad y eficacia de la regresión logística, su costo-efectividad al no requerir grandes cantidades de datos, y su interpretabilidad al proporcionar coeficientes directamente interpretables.
Rubini PE, Dr.C.A.Subasini, Dr.A.Vanitha Katharine, V.Kumaresan,S.GowdhamKumar, T.M. Nithya (2021)	Una predicción de enfermedades cardiovasculares mediante algoritmos de aprendizaje automático	Desarrollar un sistema de predicción preciso y confiable que utilice 14 características relevantes para mejorar la detección temprana y el tratamiento de enfermedades cardíacas.	Necesidad de mejorar la precisión y confiabilidad en la detección temprana de enfermedades cardíacas.	Conjunto de características relevantes relacionadas con la salud cardiovascular de los pacientes. Estas características incluyen información como la edad, el sexo, la frecuencia del pulso, la presión arterial en reposo, los niveles de colesterol y azúcar en sangre en ayunas.	Se recopiló los datos de salud cardiovascular, se limpian y se convierten en valores numéricos. Se utilizaron algoritmos para clasificar y predecir enfermedades cardiovasculares y se comparó la eficacia de los algoritmos.	Random Forest, Regresión Logística, Support Vector Machine (Máquinas de Vectores de Soporte), Naïve Bayes.	Precisión, Recuperación, Puntuación F1	Los resultados del estudio de predicción de enfermedades cardiovasculares mediante algoritmos de aprendizaje automático mostraron que el algoritmo Random Forest fue el más preciso y confiable en la clasificación y predicción de enfermedades cardiovasculares.	Random Forest	El uso de algoritmos de aprendizaje automático en la predicción de enfermedades cardiovasculares tiene el potencial de transformar la práctica clínica, mejorar los resultados de salud y promover un enfoque más proactivo y preventivo en el cuidado de la salud cardiovascular.
Chayakrit Krittanawong, Hafeez Ul HassanVirk , Sripal Bangalore , ZhenWang, KippW. Johnson, Rachel Pinotti, HongJu Zhang, Scott Kaplin. (2020)	Predicción del aprendizaje automático en enfermedades cardiovasculares: un metanálisis	Mejorar la capacidad predictiva en el diagnóstico y tratamiento de enfermedades cardiovasculares mediante el uso de algoritmos de aprendizaje automático, identificando tanto las fortalezas como las limitaciones de estas técnicas en el contexto clínico.	La necesidad de mejorar la capacidad predictiva y el tratamiento de enfermedades del corazón mediante el uso de algoritmos de aprendizaje automático debido a la falta de métricas de evaluación completas en varios estudios.	La fuente de datos utilizada fue una estrategia de búsqueda integral en las bases de datos MEDLINE, Embase y Scopus desde el inicio de la base de datos hasta el 15 de marzo de 2019. Esta búsqueda exhaustiva permitió identificar un total de 344 estudios, de los cuales 103 cohortes, con un total de 3.377.318 individuos, cumplieron con los criterios de inclusión establecidos para el metanálisis.	La metodología utilizada en la investigación involucró una búsqueda exhaustiva, una selección rigurosa de estudios, una evaluación de calidad, un análisis estadístico y una interpretación detallada de los resultados para evaluar la capacidad predictiva de los algoritmos de aprendizaje automático en enfermedades cardiovasculares.	Máquina de Vectores de Soporte (SVM), Redes Neuronales Convolucionales (CNN), Regresión Logística Lineal, Árbol de Decisión	Área bajo la curva (AUC), Sensibilidad, Especificidad y Precisión	Los resultados de la investigación sugieren que los algoritmos de aprendizaje automático son precisos en la predicción de enfermedades cardiovasculares, con especial énfasis en la CAD y los accidentes cerebrovasculares. Se identificaron algoritmos prometedores como el refuerzo y las SVM para estas predicciones, lo que destaca el potencial de la inteligencia artificial en la medicina cardiovascular.	Máquina de Vectores de Soporte (SVM)	Mejora en la predicción de enfermedades cardiovasculares, los algoritmos precisos de aprendizaje automático podrían integrarse en la práctica clínica para mejorar la identificación temprana, diagnóstico y gestión de enfermedades cardiovasculares, potencialmente mejorando los resultados clínicos y la atención al paciente.

Monther Tarawneh and Ossama Embarak (2019)	Enfoque híbrido para la predicción de enfermedades cardíacas mediante técnicas de minería de datos	Investigar y comparar diversas técnicas de extracción de datos disponibles para predecir enfermedades cardíacas. Se busca combinar los resultados de estas técnicas para obtener el diagnóstico más preciso posible.	La complejidad y la importancia del diagnóstico de enfermedades cardíacas, que puede resultar en muerte o discapacidad si no se realiza de manera adecuada y oportuna. La escasez de médicos y expertos, junto con la posibilidad de que se pasen por alto síntomas importantes de los pacientes, plantea un desafío significativo en el campo de la salud.	Conjunto de datos de enfermedades cardíacas de Cleveland que consta de 909 registros con 15 características médicas relevantes.	Desarrollo de un prototipo de sistema inteligente de predicción de enfermedades cardíacas mediante técnicas de minería de datos y aplicación de algoritmos al conjunto de datos	Árbol de decisión, Naïve Bayes, Redes Neuronales	Precisión (Accuracy), Sensibilidad (Recall), Especificidad (Specificity, Valor F1 (F1 Score)	El modelo más eficaz para predecir pacientes con enfermedades cardíacas fue Naïve Bayes, con una precisión del 86,53%. Le siguió la red neuronal, con menos del 1% de diferencia en precisión, y el árbol de decisiones obtuvo la mejor precisión para predecir pacientes sin enfermedades cardíacas en un conjunto de datos que incluía pacientes con y sin enfermedades cardíacas	Naïve Bayes	Al contar con un modelo altamente preciso, los médicos pueden tomar decisiones informadas sobre el tratamiento y seguimiento de los pacientes con enfermedades cardíacas, lo que puede mejorar la calidad de la atención y los resultados para los pacientes.
Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, Ben Van Calster (2019)	La revisión sistemática no muestra ningún beneficio de rendimiento del aprendizaje automático sobre la regresión logística para los modelos de predicción clínica	Comparar el rendimiento del aprendizaje automático con la regresión logística en modelos de predicción clínica. Se busca determinar si el aprendizaje automático ofrece beneficios significativos en términos de rendimiento en comparación con la regresión logística en este contexto específico.	La falta de evidencia que respalde un rendimiento superior del aprendizaje automático sobre la regresión logística en modelos de predicción clínica.	Se utilizaron fuentes de datos como Medline para realizar búsquedas bibliográficas sensibles.	Identificación de estudios relevantes a través de búsquedas en Medline, revisión de resúmenes y textos completos, comparación del rendimiento de los algoritmos de regresión logística y aprendizaje automático en el desarrollo de modelos de predicción clínica de diagnóstico o pronóstico para resultados binarios basados en datos clínicos.	Se utilizaron técnicas de revisión sistemática y metaanálisis . Se aplicaron criterios de inclusión y exclusión para seleccionar los estudios pertinentes y se empleó estadística descriptiva para resumir los resultados y comparar el rendimiento de los diferentes enfoques de modelado.	Área bajo la curva (AUC) Sensibilidad y especificidad, Valor predictivo positivo y negativo,Exactitud, F1-score	Los resultados de la investigación indicaron que, en general, el área bajo la curva (AUC) de los modelos de regresión logística y aprendizaje automático para la predicción del riesgo clínico fue similar cuando las comparaciones tenían bajo riesgo de sesgo. Sin embargo, se observó que el rendimiento del aprendizaje automático fue mayor en las comparaciones que tenían un alto riesgo de sesgo.	No se encontró evidencia de un rendimiento superior del aprendizaje automático sobre la regresión logística en términos de precisión para la predicción del riesgo clínico.	Los resultados sugieren que no hay un modelo claramente superior entre regresión logística y aprendizaje automático para la predicción del riesgo clínico. Por lo tanto, los investigadores y profesionales de la salud deben considerar cuidadosamente las características específicas de cada problema de predicción al seleccionar el modelo más apropiado.
Saqib Ejaz Awan, Mohammed Bennamoun, Ferdous Sohel, Frank Mario Sanfilippo and Girish Dwivedi (2019)	Predicción basada en aprendizaje automático de reingreso o muerte por insuficiencia cardíaca: implicaciones de elegir el modelo correcto y las métricas correctas	Investigar diferentes modelos basados en Machine Learning para predecir el reingreso o muerte por insuficiencia cardíaca y comparar el rendimiento de modelos de regresión estándar con modelos basados en ML en términos de predicción de reingreso o muerte por insuficiencia cardíaca.	Necesidad de desarrollar modelos predictivos precisos y efectivos para identificar a los pacientes con mayor riesgo de reingreso o muerte por insuficiencia cardíaca.	Conjunto de datos administrativos de salud vinculados con el Programa de Beneficios de Medicare.	La metodología implementada en este estudio se centra en la aplicación de modelos de Machine Learning, el uso de datos administrativos de salud vinculados, la consideración del desequilibrio de clases y la evaluación exhaustiva de métricas de rendimiento.	Modelo basado en perceptrón multicapa (MLP), Regresión logística	Característica operativa del receptor (AUC), Área bajo la curva de precisión-recuperación, Sensibilidad y especificidad	Los resultados de la investigación respaldan la eficacia del enfoque basado en perceptrón multicapa (MLP) en la predicción de eventos de reingreso o muerte por insuficiencia cardíaca a corto plazo, lo que sugiere su utilidad potencial en la práctica clínica para identificar a los pacientes con mayor riesgo y mejorar la atención médica personalizada.	Modelo basado en perceptrón multicapa (MLP)	Elegir el modelo correcto y las métricas adecuadas en la predicción de eventos clínicos en pacientes con insuficiencia cardíaca puede tener importantes implicaciones en la mejora de la atención médica, la personalización de los tratamientos y la optimización de los recursos sanitarios.
Ching-seh (Mike) Wu, Mustafa Badshah, Vishwa Bhagwat (2019)	Predicción de enfermedades cardíacas mediante técnicas de minería de datos	Aprovechar diversas técnicas de extracción de datos y desarrollar modelos de predicción para mejorar la supervivencia de pacientes con enfermedades cardíacas.	La alta tasa de mortalidad debido a enfermedades cardíacas, que se han convertido en la principal causa de muerte a nivel mundial.	El conjunto de datos utilizado proviene de la Cleveland Clinic Foundation. Este conjunto de datos describe trece atributos que caracterizan a cada paciente, incluyendo información relacionada con la salud, edad, sexo, entre otros	Ciclo de vida completo del análisis de Big Data y la minería de datos se utilizaron técnicas de minería de datos para extraer patrones de grandes conjuntos de datos, combinando métodos de estadística y aprendizaje automático con conceptos de gestión de bases de datos.	Análisis de Componentes Principales (PCA), Regresión Logística y Naïve Bayes, Árbol de Decisión y Bosque Aleatorio	Matriz de Confusión, Precisión, Sensibilidad (Recall o True Positive Rate), Exactitud	Los resultados obtenidos en la investigación muestran que diferentes clasificadores, como la regresión logística, Naïve Bayes, árbol de decisión y bosque aleatorio, tienen un impacto significativo en la predicción de enfermedades cardíacas.	Bosque Aleatorio	La utilización del modelo de Bosque Aleatorio para predecir enfermedades cardíacas tiene implicaciones positivas en términos de precisión diagnóstica, asignación de recursos y resultados de salud de los pacientes.
Sakshi Goel, Abhinav Deep, Shilpa Srivastava, Aprna Tripathi (2019)	Análisis comparativo de diversas técnicas de predicción de enfermedades cardíacas	Evaluar y comparar la precisión de diferentes enfoques tecnológicos utilizados para predecir con exactitud la presencia de enfermedades cardiovasculares, se busca identificar cuáles son los métodos más efectivos para predecir enfermedades cardíacas basándose en factores de riesgo específicos.	Necesidad de mejorar la precisión en la predicción de enfermedades cardiovasculares. Dado que las enfermedades cardíacas representan una causa significativa de mortalidad a nivel mundial	Conjunto de datos de enfermedades cardíacas de la UCI. Estos conjuntos de datos contienen información relevante sobre factores de riesgo, síntomas, diagnósticos y otros datos clínicos que son fundamentales para entrenar y evaluar los modelos de predicción de enfermedades cardíacas.	Se aplicaron técnicas de minería de datos y aprendizaje automático, se emplearon los algoritmos genéticos para la optimización y selección de características en los modelos predictivos.	Redes Neuronales Artificiales (RNA), Aprendizaje Automático, Minería de Datos, Lógica Difusa, Algoritmos Genéticos	Precisión, Sensibilidad, Especificidad, Valor Predictivo Positivo y Negativo, F1Score	Algunos hallazgos destacados incluyen la mejora en la precisión de los clasificadores débiles mediante técnicas de conjunto como el embolsado, el uso prometedor de redes neuronales y la eficacia de algoritmos específicos como J48 en la predicción de enfermedades cardíacas, y las variaciones en los resultados al comparar tecnologías como árboles de decisión, regresión logística, bosque aleatorio, redes neuronales y máquinas de vectores de soporte.	Redes Neuronales Artificiales (RNA)	La implementación de modelos altamente precisos en la predicción de enfermedades cardíacas tiene el potencial de mejorar significativamente la atención médica, la calidad de vida de los pacientes y la eficacia de las estrategias de prevención y tratamiento en el campo de la cardiología.

M.Ganesan, Dr.N.Sivakumar (2019)	Predicción y diagnóstico de enfermedades cardíacas basado en IoT modelo para atención médica utilizando modelos de aprendizaje automático	Desarrollar un modelo de diagnóstico y predicción de enfermedades cardíacas basado en IoT para servicios de atención médica en línea y emplear algoritmos de aprendizaje automático para predecir enfermedades cardíacas y clasificar los datos de los pacientes para identificar la presencia de enfermedades.	La dificultad de predecir y diagnosticar enfermedades cardíacas de manera temprana y precisa y la gestión de la gran cantidad de datos generados por dispositivos IoT en el campo médico.	Conjunto de datos de referencia sobre enfermedades cardíacas del Repositorio UCI que incluye registros médicos antiguos de pacientes recopilados de instituciones médicas.	Utilización de algoritmos de aprendizaje automático, el entrenamiento y prueba del modelo con datos reales de pacientes, la implementación del sistema en la nube y la validación de los resultados para mejorar la predicción y diagnóstico de enfermedades cardíacas basado en IoT.	Algoritmos de clasificación Aprendizaje automático Validación cruzada Computación en la nube Modelado de datos	Exactitud (Accuracy) Precisión (Precision). Recuperación (Recall)	Los resultados de la investigación destacan el rendimiento superior del clasificador J48 en la predicción y diagnóstico de enfermedades cardíacas en comparación con otros algoritmos de clasificación utilizados.	Clasificador J48	Las implicaciones de utilizar un modelo preciso de predicción y diagnóstico de enfermedades cardíacas basado en IoT, como el clasificador J48, son prometedoras en términos de mejorar la precisión diagnóstica, optimizar recursos, mejorar la atención al paciente y avanzar en la telemedicina en el campo de la atención médica.
Karna Vishnu Vardhana Reddy , Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam, Hui Na Chua and S. Pranavanand (2021)	Predicción del riesgo de enfermedades cardíacas mediante el aprendizaje automático de clasificadores con evaluadores de atributos	Mejorar la predicción del riesgo de enfermedades cardíacas utilizando técnicas de aprendizaje automático y evaluadores de atributos. Se busca identificar atributos significativos en conjuntos de datos de enfermedades cardíacas	El problema abordado en la investigación se centra en la predicción del riesgo de enfermedades cardíacas utilizando técnicas de aprendizaje automático.	Conjunto de datos de enfermedades cardíacas de Cleveland, recopilado en formato .csv del repositorio de aprendizaje automático de la UCI. Este conjunto de datos contiene información relevante para el estudio y análisis de enfermedades cardíacas, incluyendo atributos como edad, sexo, presión arterial, colesterol, entre otros.	Recopilación del conjunto de datos, exploración de los atributos, tipos, rangos de valores y otra información estadística del conjunto de datos, preprocesamiento de los datos, identificación y manejo de valores faltantes en el conjunto de datos, selección de atributos significativos y ajuste de hiperparámetros en los clasificadores para optimizar su rendimiento en la predicción del riesgo de enfermedades cardíacas.	Selección de atributos significativos, Ajuste de hiperparámetros	Precisión Sensibilidad (Recuperación) Especificidad Medida F	Se logró una precisión del 85,148% con el modelo SMO basado en el conjunto completo de atributos y una precisión del 84,158% con el modelo NB basado en un conjunto óptimo de siete atributos obtenidos de la correlación. Estos resultados reflejan el rendimiento de los clasificadores en la predicción del riesgo de enfermedades cardíacas. Esto sugiere que la combinación de técnicas de evaluación de atributos y ajuste de hiperparámetros contribuyó a mejorar la precisión de las predicciones	Support Vector Machine (SVM)	Los hallazgos de esta investigación tienen el potencial de impactar positivamente la práctica clínica, la gestión de la salud y la investigación médica al mejorar la capacidad de predecir y abordar el riesgo de enfermedades cardíacas de manera más eficaz
Pooja Rani, Rajneesh Kumar, Nada M. O. Sid Ahmed, Anurag Jain (2021)	Un sistema de apoyo a la toma de decisiones para la predicción de enfermedades cardíacas basado en el aprendizaje automático	Desarrollar un sistema híbrido de apoyo a la toma de decisiones para diagnosticar enfermedades cardíacas con mayor precisión que los sistemas existentes.	Detección temprana y precisa de enfermedades cardíacas. La principal preocupación es el retraso en la detección de estas enfermedades, lo que puede llevar a una mayor mortalidad.	Conjunto de datos de enfermedades cardíacas de Cleveland. Este conjunto de datos está disponible en el repositorio de aprendizaje automático de la Universidad de California, Irvine (UCI).	Se seleccionaron varios algoritmos de aprendizaje automático, como máquinas de vectores de soporte, bayes ingenuos, regresión logística, bosque aleatorio y clasificadores adaboost, para ser utilizados en el sistema híbrido de predicción de enfermedades cardíacas.	Aprendizaje automático Imputación de valores faltantes Selección de características Validación cruzada	Precisión Sensibilidad (Recall) Especificidad F-Measure	Los resultados de la investigación demostraron que el sistema híbrido propuesto, basado en el aprendizaje automático y utilizando el clasificador de bosque aleatorio, logró una alta precisión en la predicción de enfermedades cardíacas.	Bosque Aleatorio	Las implicaciones de la investigación incluyen mejoras en la precisión del diagnóstico de enfermedades cardíacas, la posibilidad de aplicación en entornos con recursos limitados, el potencial de expansión a otras enfermedades crónicas y la integración con tecnologías innovadoras para mejorar la atención médica y la toma de decisiones clínicas.
Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, Pascal Roy (2022)	Efectos del tamaño del conjunto de datos y las interacciones sobre el rendimiento de predicción de la regresión logística y los modelos de aprendizaje profundo	Comparar la predicción de modelos de regresión logística y modelos de aprendizaje profundo en medicina, teniendo en cuenta el tamaño del conjunto de datos y la complejidad de las interacciones entre variables y resultados.	Influencia del tamaño del conjunto de datos y las interacciones entre variables en el rendimiento de los modelos de predicción en medicina.	Se utilizaron conjuntos de datos simulados para llevar a cabo la comparación del rendimiento de la regresión logística y los modelos de aprendizaje profundo en medicina.	Se crearon conjuntos de datos simulados con variaciones en el número de observaciones y la complejidad de las interacciones entre variables y resultados, se entrenaron modelos de regresión logística y modelos de aprendizaje profundo en los conjuntos de datos simulados, se evaluó el rendimiento de los modelos mediante validación cruzada.	Regresión logística Modelos de aprendizaje profundo (redes neuronales)	Precisión Puntuación F1 Área bajo la curva (AUC)	Los resultados indicaron que los modelos de aprendizaje automático fueron menos influenciados por el tamaño del conjunto de datos, pero necesitaban términos de interacción para lograr un buen rendimiento. Por otro lado, los modelos de aprendizaje profundo pudieron lograr un buen rendimiento sin necesidad de términos de interacción.	Redes Neuronales	Las implicaciones de este estudio resaltan la importancia de considerar el tamaño del conjunto de datos, las interacciones entre variables y la elección del modelo adecuado para mejorar la precisión en la predicción de enfermedades utilizando técnicas de aprendizaje automático en el campo de la medicina.

ANEXO No. 2: Carga de la base de datos

```
[ ] import pandas as pd
```

```
[ ] heart_a = pd.read_csv('/content/heart_attack_prediction_dataset.csv')
    heart_a.head()
```

	Patient ID	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	...	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Country
0	BMW7812	67	Male	208	158/88	72	0	0	1	0	...	6.615001	261404	31.251233	286	0	6	Argentina
1	CZE1114	21	Male	389	165/93	98	1	1	1	1	...	4.963459	285768	27.194973	235	1	7	Canada
2	BNI9906	21	Female	324	174/99	72	1	0	0	0	...	9.463426	235282	28.176571	587	4	4	France
3	JLN3497	84	Male	383	163/100	73	1	1	1	0	...	7.648981	125640	36.464704	378	3	4	Canada
4	GFO8847	66	Male	318	91/88	93	1	1	1	1	...	1.514821	160555	21.809144	231	1	5	Thailand

5 rows × 26 columns

ANEXO No. 3: Eliminación de la variable “PATIENT ID”

```
[ ] heart_a = heart_a.drop('Patient ID', axis=1)
    heart_a.head()
```

	Age	Sex	Cholesterol	Blood Pressure	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	...	Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day
0	67	Male	208	158/88	72	0	0	1	0	0	...	6.615001	261404	31.251233	286	0	6
1	21	Male	389	165/93	98	1	1	1	1	1	...	4.963459	285768	27.194973	235	1	7
2	21	Female	324	174/99	72	1	0	0	0	0	...	9.463426	235282	28.176571	587	4	4
3	84	Male	383	163/100	73	1	1	1	0	1	...	7.648981	125640	36.464704	378	3	4
4	66	Male	318	91/88	93	1	1	1	1	0	...	1.514821	160555	21.809144	231	1	5

5 rows × 25 columns

ANEXO No. 4: Análisis exploratorio y transformación de variables a int

```
[ ] heart_a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 25 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Edad             8763 non-null   int64
 1   Género           8763 non-null   object
 2   Colesterol       8763 non-null   int64
 3   Blood Pressure   8763 non-null   object
 4   Freq cardiaca    8763 non-null   int64
 5   Diabetes         8763 non-null   int64
 6   Hist_familiar    8763 non-null   int64
 7   Fumador          8763 non-null   int64
 8   Obesidad         8763 non-null   int64
 9   Cons_alcohol     8763 non-null   int64
 10  He_xsemana       8763 non-null   float64
 11  Dieta            8763 non-null   object
 12  Prob_cor_ant     8763 non-null   int64
 13  Uso medicina     8763 non-null   int64
 14  Nivel estrés     8763 non-null   int64
 15  HSeden_xdía      8763 non-null   float64
 16  Ingreso          8763 non-null   int64
 17  IMC              8763 non-null   float64
 18  Triglicéridos    8763 non-null   int64
 19  DíasAF_xsemana   8763 non-null   int64
 20  HSueño_xdía      8763 non-null   int64
 21  País             8763 non-null   object
 22  Continente       8763 non-null   object
 23  Hemisferio       8763 non-null   object
 24  Riesgo AC        8763 non-null   int64
dtypes: float64(3), int64(16), object(6)
memory usage: 1.7+ MB
```

```
heart_a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Edad             8763 non-null   int64
 1   Género           8763 non-null   int64
 2   Colesterol       8763 non-null   int64
 3   Freq cardiaca    8763 non-null   int64
 4   Diabetes         8763 non-null   int64
 5   Hist_familiar    8763 non-null   int64
 6   Fumador          8763 non-null   int64
 7   Obesidad         8763 non-null   int64
 8   Cons_alcohol     8763 non-null   int64
 9   He_xsemana       8763 non-null   float64
 10  Dieta            8763 non-null   int64
 11  Prob_cor_ant     8763 non-null   int64
 12  Uso medicina     8763 non-null   int64
 13  Nivel estrés     8763 non-null   int64
 14  HSeden_xdía      8763 non-null   float64
 15  Ingreso          8763 non-null   int64
 16  IMC              8763 non-null   float64
 17  Triglicéridos    8763 non-null   int64
 18  DíasAF_xsemana   8763 non-null   int64
 19  HSueño_xdía      8763 non-null   int64
 20  País             8763 non-null   int64
 21  Continente       8763 non-null   int64
 22  Hemisferio       8763 non-null   int64
 23  Riesgo AC        8763 non-null   int64
 24  Sistólica        8763 non-null   int64
 25  Diastólica       8763 non-null   int64
dtypes: float64(3), int64(23)
memory usage: 1.7 MB
```

ANEXO No. 5: Tranformación de datos tipo object a int

1. Sex

```
heart_a['Sex'] = heart_a['Sex'].replace({'Male':0, 'Female':1})
heart_a.head()
```

2. Blood Pressure

Es necesario convertir los datos a un formato numérico adecuado, lo que incluye dividir la presión arterial en sus componentes sistólicos y diastólicos.

```
# Se divide la columna 'Blood Pressure' en dos nuevas columnas 'Sistólica' y 'Diastólica'
heart_a[['Sistólica', 'Diastólica']] = heart_a['Blood Pressure'].str.split('/', expand=True)
```

```
# Se convierten las nuevas columnas a tipo numérico
heart_a['Sistólica'] = pd.to_numeric(heart_a['Sistólica'])
heart_a['Diastólica'] = pd.to_numeric(heart_a['Diastólica'])
```

```
# Se elimina la columna de Blood Pressure
heart_a = heart_a.drop('Blood Pressure', axis=1)
heart_a.head()
```

3. Diet

```
heart_a['Diet'].unique()
```

```
array(['Average', 'Unhealthy', 'Healthy'], dtype=object)
```

```
heart_a['Diet'] = heart_a['Diet'].replace({'Unhealthy':0, 'Average':1, 'Healthy':2})
heart_a.head().T
```

4. Country

Se asigna un número a cada país

```
heart_a['Country'].unique()
```

```
array(['Argentina', 'Canada', 'France', 'Thailand', 'Germany', 'Japan',
       'Brazil', 'South Africa', 'United States', 'Vietnam', 'China',
       'Italy', 'Spain', 'India', 'Nigeria', 'New Zealand', 'South Korea',
       'Australia', 'Colombia', 'United Kingdom'], dtype=object)
```

```
from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()

heart_a['Country']= label_encoder.fit_transform(heart_a['Country'])
heart_a.head()
```

5. Continent

Se asigna un número a cada continente

```
heart_a['Continent'].unique()
```

```
array(['South America', 'North America', 'Europe', 'Asia', 'Africa',
       'Australia'], dtype=object)
```

```
heart_a['Continent']= label_encoder.fit_transform(heart_a['Continent'])
heart_a.head()
```

6. Hemisphere

Se asigna un número a cada hemisferio

```
heart_a['Hemisphere'].unique()
```

```
array(['Southern Hemisphere', 'Northern Hemisphere'], dtype=object)
```

```
heart_a['Hemisphere']= label_encoder.fit_transform(heart_a['Hemisphere'])
heart_a.head()
```

ANEXO No. 6: Estadísticas descriptivas

	Edad	Género	Colesterol	Freq cardiaca	Diabetes	Hist_familiar	Fumador	Obesidad	Cons_alcohol	He_xsemana
count	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
mean	53.707977	0.302636	259.877211	75.021682	0.652288	0.492982	0.896839	0.501426	0.598083	10.014284
std	21.249509	0.459425	80.863276	20.550948	0.476271	0.499979	0.304186	0.500026	0.490313	5.783745
min	18.000000	0.000000	120.000000	40.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.002442
25%	35.000000	0.000000	192.000000	57.000000	0.000000	0.000000	1.000000	0.000000	0.000000	4.981579
50%	54.000000	0.000000	259.000000	75.000000	1.000000	0.000000	1.000000	1.000000	1.000000	10.069559
75%	72.000000	1.000000	330.000000	93.000000	1.000000	1.000000	1.000000	1.000000	1.000000	15.050018
max	90.000000	1.000000	400.000000	110.000000	1.000000	1.000000	1.000000	1.000000	1.000000	19.998709

	IMC	Triglicéridos	DíasAF_xsemana	HSueño_xdía	País	Continente	Hemisferio	Riesgo AC	Sistólica	Diastólica
count	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
mean	28.891446	417.677051	3.489672	7.023508	9.382175	2.428849	0.354102	0.358211	135.075659	85.156111
std	6.319181	223.748137	2.282687	1.988473	5.788964	1.597561	0.478268	0.479502	26.349976	14.676565
min	18.002337	30.000000	0.000000	4.000000	0.000000	0.000000	0.000000	0.000000	90.000000	60.000000
25%	23.422985	225.500000	2.000000	5.000000	4.000000	1.000000	0.000000	0.000000	112.000000	72.000000
50%	28.768999	417.000000	3.000000	7.000000	9.000000	3.000000	0.000000	0.000000	135.000000	85.000000
75%	34.324594	612.000000	5.000000	9.000000	14.000000	4.000000	1.000000	1.000000	158.000000	98.000000
max	39.997211	800.000000	7.000000	10.000000	19.000000	5.000000	1.000000	1.000000	180.000000	110.000000

ANEXO No. 7: Matriz de correlación

	Edad	Género	Colesterol	Freq cardiaca	Diabetes	Historial familiar	Fumador	Obesidad	Consumo alcohol	He_xsemana	IMC
Edad	1.000000	-0.020067	-0.009107	-0.003844	-0.014105	0.008353	0.394891	-0.008140	-0.006666	0.001206	-0.002612
Género	-0.020067	1.000000	-0.007614	0.010921	-0.003582	-0.002180	-0.514837	-0.002376	-0.002085	0.006976	0.003021
Colesterol	-0.009107	-0.007614	1.000000	0.000315	-0.013428	-0.021608	0.016342	-0.014843	-0.007261	0.021517	0.017292
Freq cardiaca	-0.003844	0.010921	0.000315	1.000000	0.006764	-0.013470	-0.012331	0.012725	0.003459	0.008276	0.005299
Diabetes	-0.014105	-0.003582	-0.013428	0.006764	1.000000	-0.013844	0.000527	0.012866	0.005551	-0.007014	-0.002852
Historial familiar	0.008353	-0.002180	-0.021608	-0.013470	-0.013844	1.000000	0.011748	-0.001444	0.012701	-0.006378	-0.011492
Fumador	0.394891	-0.514837	0.016342	-0.012331	0.000527	0.011748	1.000000	0.003969	0.012754	-0.000150	0.007670
Obesidad	-0.008140	-0.002376	-0.014843	0.012725	0.012866	-0.001444	0.003969	1.000000	-0.024195	0.002099	-0.006058
Consumo alcohol	-0.006666	-0.002085	-0.007261	0.003459	0.005551	0.012701	0.012754	-0.024195	1.000000	-0.008514	0.010562
He_xsemana	0.001206	0.006976	0.021517	0.008276	-0.007014	-0.006378	-0.000150	0.002099	-0.008514	1.000000	0.003777
Dieta	-0.013230	-0.005740	-0.010765	-0.003014	0.006156	-0.001401	0.006023	0.003743	0.005336	0.007667	0.011755
Prob_cor_ant	0.000868	-0.001964	-0.006070	-0.004956	0.000867	-0.004568	-0.000574	0.005159	0.010395	0.005253	0.015718
Uso medicina	0.000980	0.007148	-0.000905	0.009244	-0.002656	0.000981	-0.010877	-0.006267	0.003339	-0.007119	0.009514
Nivel estrés	0.018307	0.021835	-0.024487	-0.004547	0.006719	0.015637	-0.001757	0.010626	-0.005023	-0.009102	-0.003250
HSeden_xdía	0.017280	-0.002995	0.018914	-0.010232	0.004705	0.002561	0.015311	-0.001333	-0.012828	0.008756	-0.000024
Ingreso	-0.001733	-0.002660	0.000007	0.004873	-0.000759	-0.000401	0.003096	-0.003870	-0.022396	-0.023414	0.008836
IMC	-0.002612	0.003021	0.017292	0.005299	-0.002852	-0.011492	0.007670	-0.006058	0.010562	0.003777	1.000000
Triglicéridos	0.003415	-0.002933	-0.005454	0.012244	0.010431	-0.001904	0.004650	0.001467	0.006169	0.001717	-0.005964
Días de AF_xsemana	0.001384	0.007660	0.016056	0.000834	-0.002411	0.009561	-0.006465	0.005337	0.001593	0.007725	0.008110
HSueño_xdía	-0.002185	0.005329	0.004456	0.001811	-0.012457	-0.011199	-0.005424	-0.005314	-0.000843	-0.001245	-0.010030
País	0.002567	-0.016501	0.012962	-0.016436	0.011031	-0.003194	0.021095	-0.009631	0.013029	0.005799	0.002394
Continente	-0.010387	0.016902	0.008892	0.000975	-0.002445	0.006055	-0.013697	0.001377	0.005013	-0.002928	-0.001403
Hemisferio	-0.002795	0.003075	-0.019462	0.010145	0.001478	-0.004163	0.002439	-0.008555	-0.012095	-0.007485	0.001512
Riesgo AC	0.006403	-0.003095	0.019340	-0.004251	0.017225	-0.001652	-0.004051	-0.013318	-0.013778	0.011133	0.000020
Sistólica	0.003070	0.006037	0.000133	0.008482	-0.005306	-0.009762	-0.009534	-0.001918	0.010764	-0.009506	0.004279
Diastólica	-0.009826	0.002251	0.002083	-0.018113	-0.000512	0.017818	-0.012293	-0.020574	-0.007282	-0.003469	0.000806

	Triglicéridos	Días de AF_xsemana	HSueño_xdía	País	Continente	Hemisferio	Riesgo AC	Sistólica	Diastólica
Edad	0.003415	0.001384	-0.002185	0.002567	-0.010387	-0.002795	0.006403	0.003070	-0.009826
Género	-0.002933	0.007660	0.005329	-0.016501	0.016902	0.003075	-0.003095	0.006037	0.002251
Colesterol	-0.005454	0.016056	0.004456	0.012962	0.008892	-0.019462	0.019340	0.000133	0.002083
Freq cardiaca	0.012244	0.000834	0.001811	-0.016436	0.000975	0.010145	-0.004251	0.008482	-0.018113
Diabetes	0.010431	-0.002411	-0.012457	0.011031	-0.002445	0.001478	0.017225	-0.005306	-0.000512
Historial familiar	-0.001904	0.009561	-0.011199	-0.003194	0.006055	-0.004163	-0.001652	-0.009762	0.017818
Fumador	0.004650	-0.006465	-0.005424	0.021095	-0.013697	0.002439	-0.004051	-0.009534	-0.012293
Obesidad	0.001467	0.005337	-0.005314	-0.009631	0.001377	-0.008555	-0.013318	-0.001918	-0.020574
Consumo alcohol	0.006169	0.001593	-0.000843	0.013029	0.005013	-0.012095	-0.013778	0.010764	-0.007282
He_xsemana	0.001717	0.007725	-0.001245	0.005799	-0.002928	-0.007485	0.011133	-0.009506	-0.003469
Dieta	-0.013660	-0.013265	-0.014513	-0.005703	0.003620	0.002210	0.005908	0.013648	0.005636
Prob_cor_ant	-0.019029	0.008537	0.004460	0.001871	0.021525	0.017384	0.000274	-0.011926	0.008813
Uso medicina	-0.011095	-0.011139	-0.020393	0.019992	-0.002255	0.011279	0.002234	-0.001182	0.004607
Nivel estrés	-0.003921	0.007405	-0.014205	-0.004309	-0.001078	-0.008886	-0.004111	0.017848	-0.008445
HSeden_xdía	-0.005785	-0.006178	0.004792	0.016105	-0.027349	0.011850	-0.005613	0.003393	-0.006606
Ingreso	0.010739	0.000130	-0.006598	-0.018578	-0.000598	0.002447	0.009628	0.010414	0.008816
IMC	-0.005964	0.008110	-0.010030	0.002394	-0.001403	0.001512	0.000020	0.004279	0.000806
Triglicéridos	1.000000	-0.007556	-0.029216	0.019330	-0.010763	-0.013185	0.010471	0.005121	0.000545
Días de AF_xsemana	-0.007556	1.000000	0.014033	0.015435	-0.003918	-0.010397	-0.005014	-0.007574	0.016294
HSueño_xdía	-0.029216	0.014033	1.000000	-0.008157	-0.011078	0.001327	-0.018528	-0.004628	0.010679
País	0.019330	0.015435	-0.008157	1.000000	-0.444254	-0.292710	0.003550	0.007975	0.001683
Continente	-0.010763	-0.003918	-0.011078	-0.444254	1.000000	0.220963	0.004446	0.006940	0.013057
Hemisferio	-0.013185	-0.010397	0.001327	-0.292710	0.220963	1.000000	-0.012704	0.003969	0.003961
Riesgo AC	0.010471	-0.005014	-0.018528	0.003550	0.004446	-0.012704	1.000000	0.018585	-0.007509
Sistólica	0.005121	-0.007574	-0.004628	0.007975	0.006940	0.003969	0.018585	1.000000	0.013337
Diastólica	0.000545	0.016294	0.010679	0.001683	0.013057	0.003961	-0.007509	0.013337	1.000000

ANEXO No. 8: Modelo de Bosque Aleatorio

```
Accuracy: 0.6411865373645179
Classification Report:
              precision    recall  f1-score   support

           0       0.64      1.00      0.78      1125
           1       0.33      0.00      0.00       628

    accuracy                           0.64      1753
   macro avg       0.49      0.50      0.39      1753
weighted avg       0.53      0.64      0.50      1753
```

ANEXO No. 9: Modelo de Bosque Aleatorio ajustado

```
Fitting 3 folds for each of 216 candidates, totalling 648 fits
Best Parameters: {'bootstrap': False, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
Best Grid Search Score: 0.6437299097975914
Accuracy: 0.5812892184826013
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.80      0.71      1125
           1       0.34      0.19      0.24       628

    accuracy                           0.58      1753
   macro avg       0.49      0.49      0.48      1753
weighted avg       0.53      0.58      0.54      1753
```

ANEXO No. 10: Modelo de Red Neuronal con una capa

```
Loss: 0.6644670963287354, Accuracy: 0.6371933817863464
              precision    recall  f1-score   support

           0       0.64      1.00      0.78      1125
           1       0.33      0.00      0.00       628

    accuracy                           0.64      1753
   macro avg       0.49      0.50      0.39      1753
weighted avg       0.53      0.64      0.50      1753
```

ANEXO No. 11: Modelo de Red Neuronal con dos capas

```
Loss: 0.658257782459259, Accuracy: 0.6423274278640747
55/55 [==============================] - 0s 4ms/step
              precision    recall  f1-score   support

           0       0.64      1.00      0.78      1125
           1       1.00      0.00      0.00       628

    accuracy                           0.64      1753
   macro avg       0.82      0.50      0.39      1753
weighted avg       0.77      0.64      0.50      1753
```

ANEXO No. 12: Modelo de red neuronal con tres capas

```
Loss: 0.7176427245140076, Accuracy: 0.5767256021499634
55/55 [==============================] - 0s 4ms/step
              precision    recall  f1-score   support

           0       0.63      0.80      0.71      1125
           1       0.33      0.17      0.23       628

    accuracy                           0.58      1753
   macro avg       0.48      0.49      0.47      1753
weighted avg       0.52      0.58      0.54      1753
```

ANEXO No. 13: Modelo de red neuronal con cuatro capas

```
Loss: 0.706662118434906, Accuracy: 0.5784369707107544
55/55 [==============================] - 0s 2ms/step
              precision    recall  f1-score   support

           0       0.64      0.80      0.71      1125
           1       0.33      0.18      0.23       628

    accuracy                           0.58      1753
   macro avg       0.49      0.49      0.47      1753
weighted avg       0.53      0.58      0.54      1753
```

ANEXO No. 14: Modelo de red neuronal con cinco capas

```
Loss: 0.6955868601799011, Accuracy: 0.5881346464157104
55/55 [==============================] - 0s 2ms/step
              precision    recall  f1-score   support

           0       0.64      0.83      0.72      1125
           1       0.33      0.15      0.21       628

    accuracy                           0.59      1753
   macro avg       0.49      0.49      0.46      1753
weighted avg       0.53      0.59      0.54      1753
```