

Universidad Peruana de Ciencias Aplicadas



INFORME DEL TRABAJO FINAL

CURSO DE BIG DATA

Alumnos:

Francesca Nicole Bances Torres

Marsi Figueiroa Larragan

Profesor: Andres Gibu La Torre

Julio 2025

Resumen

Este proyecto presenta el desarrollo de un sistema predictivo para evaluar el riesgo de enfermedad cardíaca utilizando aprendizaje automático con H2O. A partir de un conjunto de datos clínicos y de estilo de vida, se entrenaron modelos supervisados como Random Forest, Gradient Boosting, GLM y AutoML. Tras comparar su rendimiento, se implementó una aplicación web interactiva con Streamlit que permite a los usuarios ingresar su información y recibir una evaluación de riesgo personalizada. Esta propuesta busca demostrar el potencial de las tecnologías de datos para apoyar la prevención en salud de forma accesible y educativa.

Índice

1. Descripción del caso de uso.....	4
2. Descripción del conjunto de datos obtenidos de kaggle.....	5
2.1. Origen y contexto de los datos.....	5
2.2. Cantidad de registros y variables disponibles.....	6
2.3. Variables del conjunto.....	6
2.4. Limitaciones y consideraciones éticas.....	9
3. Análisis exploratorio de los datos (EDA).....	10
3.1. Inspección Inicial del Conjunto de Datos.....	10
3.2. Preprocesamiento y Limpieza de Datos.....	11
3.3. Análisis Univariado.....	11
3.4. Análisis Bivariado.....	12
3.5. Information Value.....	12
3.6. Visualización de Datos.....	14
4. Modelización.....	14
4.1. Modelos supervisados seleccionados.....	15
4.1.1. Random Forest.....	15
4.1.2. Gradient Boosting Machine (GBM).....	15
4.1.3. Generalized Linear Model (GLM).....	16
4.1.4. AutoML (Stacked Ensemble).....	16
4.2. Métricas de evaluación.....	16
5. Resultados.....	17
6. Implementación Aplicativa del Modelo Predictivo.....	18
6.1. Evidencias de funcionamiento.....	21
7. Conclusiones.....	22
8. Recomendaciones.....	23
9. Referencias.....	24

1. Descripción del caso de uso

Las enfermedades cardiovasculares (ECV) siguen siendo la principal causa de muerte a nivel mundial, cobrando aproximadamente 17,9 millones de vidas cada año, lo que representa un 32 % del total de defunciones globales. De estas muertes, más del 75 % se producen en países de ingresos bajos y medios, donde el acceso al diagnóstico temprano y a tratamientos efectivos suele ser limitado (WHO, 2023). Estas cifras reflejan no solo la magnitud del problema, sino también la urgencia de adoptar enfoques preventivos y tecnologías que puedan mitigar su impacto.

En el caso particular de América Latina y el Caribe, la situación no es menos alarmante. Según Carrillo-Larco et al. (2019), en esta región existe una importante carga de ECV asociada al aumento sostenido de factores de riesgo como el tabaquismo, el sedentarismo, la diabetes y la hipertensión, además de la falta de acceso a herramientas predictivas adecuadas para poblaciones locales. Los autores señalan que la mayoría de modelos de predicción del riesgo cardiovascular han sido desarrollados y validados en contextos extranjeros, lo que limita su aplicabilidad directa en países latinoamericanos debido a diferencias culturales, genéticas y de estilo de vida. Esto evidencia la necesidad de contar con soluciones adaptadas, accesibles y fáciles de interpretar para los usuarios finales y el personal sanitario de la región.

En respuesta a esta problemática, el presente proyecto propone el desarrollo de una aplicación interactiva que permite evaluar el riesgo de padecer enfermedades cardíacas mediante técnicas de aprendizaje automático. Utilizando H2O AutoML, se ha entrenado un modelo predictivo a partir de datos públicos que combinan variables demográficas, hábitos de salud y condiciones clínicas comunes. La herramienta ha sido diseñada pensando tanto en profesionales de la salud como en usuarios no especializados, facilitando una experiencia intuitiva para la autoevaluación. La interfaz, desarrollada con Streamlit, presenta un formulario en español y devuelve una predicción acompañada de recomendaciones generales.

Este tipo de soluciones no busca reemplazar el diagnóstico médico, sino actuar como complemento en estrategias de promoción de la salud, educación preventiva y detección temprana de posibles riesgos. Tal como advierte la Organización Mundial de la Salud (WHO, 2023), es fundamental identificar a tiempo a las personas con alto riesgo cardiovascular para ofrecerles un tratamiento temprano y efectivo. La prevención y la concienciación pueden evitar muchas muertes prematuras, especialmente en contextos donde el sistema de salud presenta limitaciones estructurales.

En síntesis, este sistema representa un esfuerzo por integrar ciencia de datos, accesibilidad digital y enfoque preventivo en la lucha contra las enfermedades cardiovasculares, con especial énfasis en contextos latinoamericanos. Su valor radica no solo en la capacidad de predicción, sino en su potencial para empoderar al usuario, promover el autocuidado y contribuir a cerrar brechas en el acceso a información útil para la toma de decisiones en salud.

2. Descripción del conjunto de datos obtenidos de kaggle

2.1. Origen y contexto de los datos

El conjunto de datos utilizado para este proyecto fue obtenido desde Kaggle, una reconocida plataforma de ciencia de datos que reúne competiciones, datasets públicos y herramientas para análisis exploratorio. Este dataset en particular fue desarrollado a partir de una muestra de la encuesta Behavioral Risk Factor Surveillance System (BRFSS), llevada a cabo por los Centers for Disease Control and Prevention (CDC) en Estados Unidos. La BRFSS recopila información anual sobre comportamientos de salud, condiciones médicas y factores de riesgo en población adulta.

El propósito de este conjunto de datos es permitir la construcción de modelos predictivos que ayuden a determinar la probabilidad de que una persona padezca o no enfermedad cardíaca, a partir de sus características demográficas, hábitos de vida y condiciones autodeclaradas de salud. Su enfoque es preventivo, educativo y clínico, y se presta para la aplicación de modelos de inteligencia artificial por su riqueza de variables y volumen.

El dataset se encuentra en el siguiente enlace:
<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

2.2. Cantidad de registros y variables disponibles

El dataset contiene 319795 registros únicos, cada uno correspondiente a una persona encuestada. Las observaciones se presentan en formato tabular y cada fila representa un individuo. En total, hay 18 variables, entre categóricas y numéricas, que capturan información personal, hábitos, condiciones médicas preexistentes y percepciones subjetivas del estado de salud.

La variable objetivo (HeartDisease) es binaria y expresa si la persona ha sido diagnosticada con una enfermedad cardíaca o no. Esta variable fue derivada a partir de respuestas relacionadas con infartos, anginas y otras condiciones coronarias.

2.3. Variables del conjunto

El conjunto de datos cuenta con 18 características que describen aspectos clave de la salud, hábitos y condiciones clínicas autodeclaradas por los individuos encuestados. Estas características permiten construir un modelo de predicción binaria que indica si una persona presenta o no un diagnóstico previo de enfermedad cardíaca.

A continuación, se presenta una tabla detallada con cada una de las columnas disponibles, indicando su significado, tipo de dato y, en los casos correspondientes, los posibles valores que puede adoptar.

Nombre	Descripción	Tipo de dato	Valores posibles
--------	-------------	--------------	------------------

HeartDisease	Indica si la persona fue diagnosticada con enfermedad cardíaca (objetivo)	Categórico	Yes / No
BMI	Índice de masa corporal (peso/altura ²)	Numérico	Valores reales positivos
Smoking	Si la persona fuma actualmente o ha fumado	Categórico	Yes / No
AlcoholDrinking	Si la persona consume alcohol en exceso	Categórico	Yes / No
Stroke	Si la persona ha sufrido un derrame cerebral	Categórico	Yes / No
PhysicalHealth	Número de días con salud física no buena en los últimos 30 días	Numérico entero	0 a 30
MentalHealth	Número de días con salud mental no buena en los últimos 30 días	Numérico entero	0 a 30
DiffWalking	Si tiene dificultad para caminar o subir escaleras	Categórico	Yes / No
Sex	Género biológico	Categórico	Male / Female

AgeCategory	Grupo etario de la persona	Categórico	“18-24”, “25-29”, ... “75-79”, “80 or older”
Race	Grupo étnico autodeclarado	Categórico	White, Black, Asian, American Indian/Alaska Native, etc.
Diabetic	Si ha sido diagnosticado con diabetes o prediabetes	Categórico	Yes, No, No (borderline diabetes), Yes (during pregnancy)
PhysicalActivity	Si ha realizado actividad física fuera del trabajo en los últimos 30 días	Categórico	Yes / No
GenHealth	Salud general autoperceptiva	Categórico	Excellent, Very good, Good, Fair, Poor
SleepTime	Número promedio de horas de sueño por noche	Numérico entero	Valores entre 0 y 24
Asthma	Si ha sido diagnosticado con asma	Categórico	Yes / No

KidneyDisease	Si ha sido diagnosticado con enfermedad renal	Categórico	Yes / No
SkinCancer	Si ha sido diagnosticado con cáncer de piel	Categórico	Yes / No

2.4. Limitaciones y consideraciones éticas

Si bien el dataset utilizado ofrece un gran potencial para el entrenamiento de modelos predictivos, presenta ciertas limitaciones que deben tenerse en cuenta al interpretar sus resultados.

En primer lugar, la información proviene exclusivamente de población adulta residente en Estados Unidos. Esto implica que sus patrones de salud, comportamiento y contexto socioeconómico pueden diferir significativamente de los de poblaciones latinoamericanas. Por tanto, aplicar directamente los resultados del modelo a otras regiones debe hacerse con precaución y considerando las diferencias culturales y estructurales del sistema de salud.

Otro punto importante es que muchas de las características del dataset son autodeclaradas. Esto significa que los datos podrían estar sujetos a sesgos personales, errores de recuerdo o deseabilidad social. Por ejemplo, variables como el consumo de alcohol, el estado general de salud o la actividad física pueden no reflejar con precisión la realidad clínica.

Además, el dataset no incluye información médica directa como resultados de análisis clínicos, electrocardiogramas, niveles de colesterol o antecedentes familiares detallados, lo cual limita la profundidad diagnóstica del modelo.

No obstante, a pesar de estas limitaciones, el conjunto de datos resulta muy valioso desde una perspectiva educativa. Su estructura simple y accesible permite demostrar cómo a partir de información básica es posible desarrollar herramientas

de apoyo a la toma de decisiones médicas preliminares, siempre y cuando se utilicen con responsabilidad y en combinación con criterios profesionales.

3. Análisis exploratorio de los datos (EDA)

El Análisis Exploratorio de Datos permite obtener una comprensión inicial del comportamiento del conjunto de datos. Antes de entrenar cualquier modelo predictivo, resulta fundamental revisar la estructura general de los datos, identificar valores atípicos o inconsistencias, y conocer cómo se distribuyen las variables tanto de forma individual como en relación con la variable objetivo. Este proceso no solo orienta las decisiones sobre qué variables podrían ser útiles para el modelo, sino que también ayuda a evitar errores posteriores derivados de datos mal comprendidos o mal tratados.

3.1. Inspección Inicial del Conjunto de Datos

El análisis comenzó con una inspección general del dataset `heart_2020_cleaned.csv`, que incluye un total de 319,795 registros correspondientes a personas adultas encuestadas por el sistema BRFSS (Behavioral Risk Factor Surveillance System) en Estados Unidos. Este archivo contiene 18 columnas, de las cuales 17 son predictoras y una corresponde a la variable objetivo, que indica si una persona ha sido diagnosticada con enfermedad cardíaca.

Se utilizó la función `.head()` para visualizar las primeras filas del dataset y verificar que los datos tuvieran una estructura coherente. A su vez, con funciones como `.info()` y `.dtypes`, se pudo distinguir entre las variables de tipo categórico (como `Smoking`, `Diabetic`, `SkinCancer`, etc.) y las variables numéricas discretas (por ejemplo, `PhysicalHealth`, `MentalHealth` y `SleepTime`).

También se revisó la variable objetivo HeartDisease y se encontró que presenta una distribución desbalanceada: una mayor proporción de personas sin diagnóstico (clase "No") frente a una menor cantidad de personas con diagnóstico (clase "Yes"). Este dato es crucial porque podría influir en el entrenamiento de modelos si no se maneja adecuadamente.

3.2. Preprocesamiento y Limpieza de Datos

Este conjunto ya se encuentra limpio y depurado, lo que implica que no contiene valores nulos ni errores evidentes en su estructura. No obstante, se llevaron a cabo las siguientes tareas de preprocesamiento para optimizar su uso:

- Conversión de tipos de datos: las variables categóricas fueron transformadas a tipo category, lo cual reduce el uso de memoria y permite ejecutar operaciones estadísticas más rápidas y precisas.
- Revisión de valores únicos: se exploraron los valores posibles de cada variable categórica para asegurar que no existieran errores de digitación, duplicados mal codificados (por ejemplo, "yes" y "Yes") o categorías irrelevantes.
- Búsqueda de outliers: se generaron gráficos de caja (boxplots) para variables numéricas como BMI, SleepTime y PhysicalHealth. Aunque se encontraron algunos valores extremos (como 0 horas de sueño o más de 25 días de salud física deficiente), estos no fueron eliminados al considerar que podrían corresponder a casos reales dentro del contexto de salud pública.

3.3. Análisis Univariado

El análisis univariado tuvo como objetivo examinar la distribución individual de cada variable. Para las variables categóricas, como Smoking, Diabetic, PhysicalActivity o SkinCancer, se calcularon frecuencias relativas y absolutas mediante .value_counts() y se representaron visualmente con gráficos de barras.

Algunos hallazgos relevantes fueron:

- Cerca del 45% de los encuestados se identificó como fumador actual o exfumador.
- La mayoría reportó realizar algún tipo de actividad física en los últimos 30 días.
- Un número considerable de personas reportó haber sido diagnosticado con condiciones como diabetes o cáncer de piel.

En el caso de las variables numéricas, como PhysicalHealth, se calcularon estadísticas como media, mediana, desviación estándar, mínimo y máximo, identificando que la mayoría de personas reportaba entre 0 y 5 días de mal estado físico, lo que concuerda con una población mayormente funcional.

3.4. Análisis Bivariado

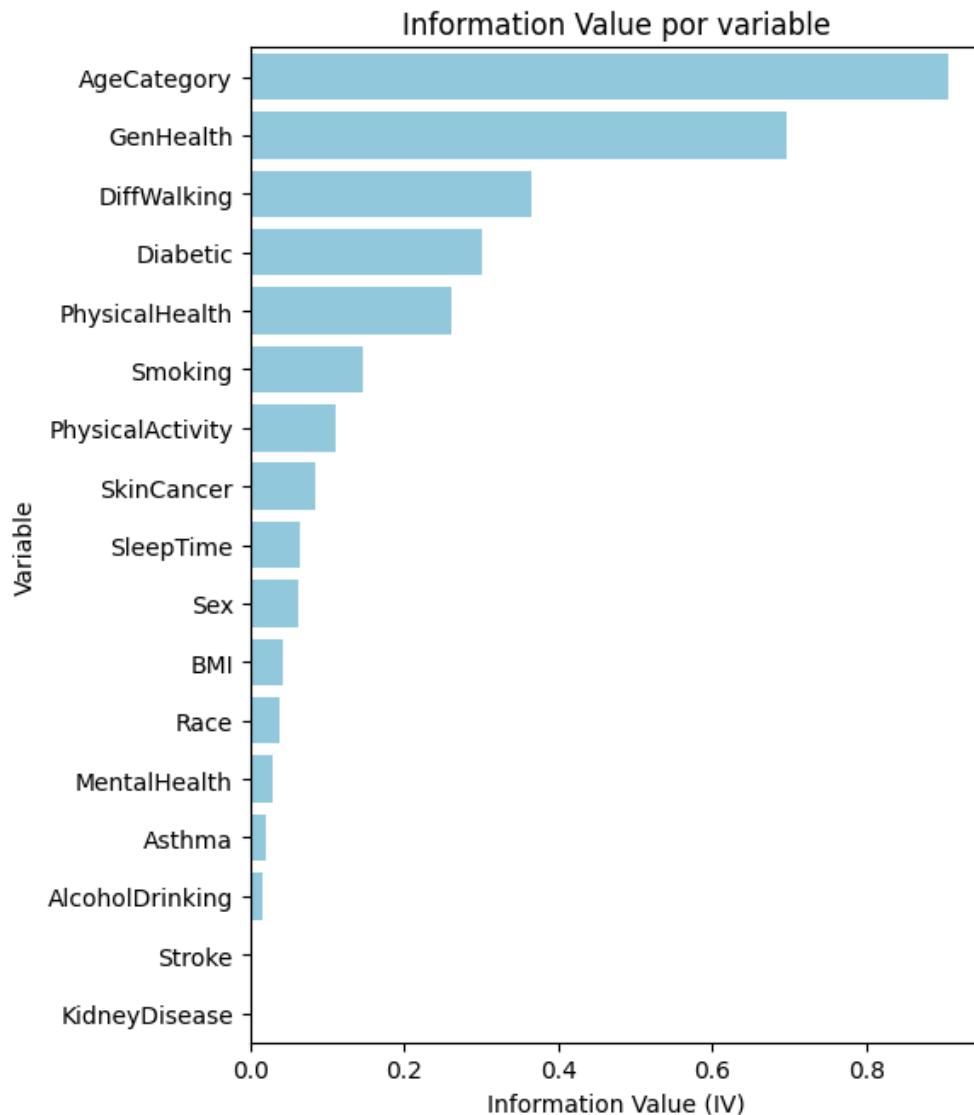
Esta etapa consistió en analizar la relación entre las variables predictoras y la variable objetivo HeartDisease. Se distinguió entre dos enfoques, según el tipo de variable:

- Para las variables categóricas, se utilizaron tablas cruzadas (tablas de contingencia) y se graficaron comparaciones entre categorías. Por ejemplo, se observó que la proporción de enfermedad cardíaca era más alta entre personas que indicaron tener dificultades para caminar (DiffWalking = Sí) o que habían sido diagnosticadas previamente con diabetes (Diabetic = Sí).
- Para las variables numéricas, se emplearon gráficos de caja (boxplots) que permitieron comparar la distribución de valores según la presencia o ausencia de enfermedad. En PhysicalHealth, por ejemplo, se identificó una diferencia notable: las personas diagnosticadas con enfermedad cardíaca reportaban más días de mala salud física en los últimos 30 días.

3.5. Information Value

Además de los métodos tradicionales de análisis bivariado, se incorporó una métrica cuantitativa conocida como Information Value (IV), la cual permite estimar el poder predictivo de cada variable respecto a la variable objetivo

HeartDisease. Esta métrica se utiliza comúnmente en modelos de riesgo crediticio, pero también es altamente útil en contextos médicos donde se busca medir el grado de asociación entre variables categóricas y un resultado binario.



El IV se interpreta según las siguientes escalas:

- $IV < 0.02$ ----- No predictiva
- $0.02 \leq IV < 0.1$ ----- Débil
- $0.1 \leq IV < 0.3$ ----- Moderada
- $IV \geq 0.3$ ----- Fuerte

En este análisis se calculó el IV para las principales variables categóricas del dataset. Entre los resultados más destacados:

- DiffWalking obtuvo un IV superior a 0.4, clasificándose como una variable fuertemente predictiva. Esto refuerza lo observado visualmente: las personas con dificultad para caminar presentan una incidencia significativamente mayor de enfermedad cardíaca.
- GenHealth y Diabetic alcanzaron valores de IV entre 0.2 y 0.3, indicando una capacidad predictiva moderada.
- Variables como PhysicalActivity, SkinCancer y Smoking mostraron un IV entre 0.05 y 0.1, clasificadas como predictoras débiles pero relevantes en un contexto multivariable.

El uso del IV permitió confirmar que muchas de las variables seleccionadas para el modelo no solo tienen sentido clínico, sino también un respaldo estadístico claro como factores de riesgo.

3.6. Visualización de Datos

Para facilitar la interpretación y mejorar la comunicación de los hallazgos, se incorporaron diversas visualizaciones generadas con matplotlib y seaborn:

- Gráficos de barras para representar frecuencias de variables categóricas.
- Gráficos de caja (boxplots) para evaluar diferencias en variables numéricas entre grupos con y sin enfermedad.
- Histogramas para analizar la distribución de variables continuas.

Estas visualizaciones no solo ayudaron a entender mejor la distribución de los datos, sino también a detectar posibles relaciones no evidentes a simple vista.

4. Modelización

En este trabajo, se utilizaron algoritmos de aprendizaje automático implementados en la plataforma H2O, una herramienta open source optimizada para alto rendimiento y manejo de grandes volúmenes de datos. Se trabajó con un enfoque supervisado, dado que la

variable objetivo (HeartDisease) es binaria, lo cual permite predecir la probabilidad de padecer enfermedad cardíaca en base a variables clínicas y de estilo de vida.

4.1. Modelos supervisados seleccionados

Con el objetivo de comparar distintos enfoques de predicción, se seleccionaron cinco tipos de modelos ampliamente reconocidos por su desempeño y versatilidad:

4.1.1. Random Forest

El modelo de Random Forest se basa en un conjunto de árboles de decisión que trabajan de manera colaborativa. Cada árbol es entrenado con una muestra aleatoria del conjunto de datos y realiza una predicción independiente; luego, se realiza una votación para obtener la predicción final. Esta técnica mejora la precisión del modelo y reduce el riesgo de sobreajuste, aprovechando la diversidad de los árboles individuales. En este trabajo, se ajustaron hiperparámetros clave como la cantidad de árboles (`ntrees`), la profundidad máxima (`max_depth`) y el número mínimo de muestras por hoja (`min_rows`) para optimizar su rendimiento.

4.1.2. Gradient Boosting Machine (GBM)

El GBM es un método de ensamble basado en boosting secuencial. A diferencia de Random Forest, en GBM cada nuevo árbol se construye corrigiendo los errores cometidos por los árboles anteriores. Esto permite capturar relaciones complejas en los datos y mejorar la precisión del modelo de forma iterativa. Aunque puede ser más sensible al sobreajuste si no se controla adecuadamente la profundidad de los árboles o la tasa de aprendizaje, su alto rendimiento lo convierte en una opción muy competitiva. En este análisis se evaluaron combinaciones de parámetros como `ntrees`, `max_depth` y `learn_rate`.

4.1.3. Generalized Linear Model (GLM)

El GLM, específicamente en su versión logística (regresión logística binaria), fue utilizado como modelo base. Este enfoque permite estimar la probabilidad de pertenecer a una clase (presencia o ausencia de enfermedad) en función de las variables predictoras, bajo el supuesto de una relación lineal entre éstas y la variable respuesta. Aunque es un modelo más simple comparado con métodos de ensamble, tiene la ventaja de ser interpretable y rápido de entrenar. Su inclusión permitió establecer una línea base clara sobre la cual comparar modelos más complejos.

4.1.4. AutoML (Stacked Ensemble)

Finalmente, se utilizó el módulo AutoML de H2O, el cual automatiza el proceso de entrenamiento y selección de modelos. Esta herramienta ejecuta múltiples algoritmos —como GBM, GLM, Deep Learning y XGBoost— y combina los mejores mediante un modelo ensamblado llamado Stacked Ensemble. Este enfoque aprovecha lo mejor de cada modelo base, optimizando la capacidad predictiva del conjunto. En este caso, el modelo final generado por AutoML fue un ensemble que logró el mejor rendimiento en términos de precisión y área bajo la curva (AUC), convirtiéndose en el modelo más destacado del análisis.

4.2. Métricas de evaluación

Para evaluar el desempeño de los modelos supervisados utilizados en este estudio, se emplearon diversas métricas que permiten medir tanto la capacidad predictiva como la calidad de la clasificación. A continuación, se describen las métricas seleccionadas y su relevancia en el contexto del problema:

- AUC (Área bajo la curva ROC): Esta métrica evalúa la capacidad del modelo para discriminar correctamente entre las dos clases (riesgo de enfermedad cardíaca: sí o no). Un valor de AUC cercano a 1 indica una excelente capacidad de diferenciación, mientras que valores cercanos a 0.5 reflejan un desempeño similar al azar.

- LogLoss (Pérdida logarítmica): Mide el grado de incertidumbre de las predicciones probabilísticas. Penaliza más fuertemente las predicciones incorrectas que están muy alejadas del valor real. Cuanto menor sea el LogLoss, mayor será la precisión y la calibración del modelo.
- F1-score: Es el promedio armónico entre la precisión (exactitud de los positivos predichos) y la sensibilidad o recall (capacidad del modelo para identificar correctamente los casos positivos). Esta métrica resulta especialmente útil en escenarios con clases desbalanceadas, como suele ocurrir en datos de salud pública.
- Coeficiente de Gini: Derivado del AUC, este indicador cuantifica la desigualdad en la distribución de las predicciones. Un mayor valor del coeficiente de Gini refleja una mejor capacidad del modelo para separar correctamente las clases.

5. Resultados

Tras aplicar los modelos supervisados sobre la plataforma H2O, se obtuvieron las siguientes métricas clave de rendimiento: el Área bajo la curva ROC (AUC) y el Coeficiente de Gini. Estas métricas permiten evaluar qué tan bien los modelos logran discriminar entre personas con y sin riesgo de enfermedad cardíaca.

A continuación, se muestra una tabla comparativa con los valores obtenidos:

Modelo	AUC	Gini
GLM	0.821539	0.643079
Boosting	0.820098	0.640196
Random Forest	0.814128	0.628255
AutoML	0.824738	0.649476

Como se puede observar, el modelo generado por AutoML obtuvo el mejor rendimiento general, alcanzando un AUC de 0.8247 y un coeficiente de Gini de 0.6495, lo cual indica

una mayor capacidad discriminativa en comparación con los otros modelos. El modelo GLM también mostró un desempeño competitivo, superando al Random Forest tanto en AUC como en Gini, a pesar de ser un modelo más simple y lineal.

El algoritmo de Boosting obtuvo métricas intermedias, ligeramente por debajo de GLM y AutoML, lo que sugiere que su combinación de árboles secuenciales fue efectiva, pero no la más óptima frente al conjunto de datos.

Estos resultados confirman la utilidad del enfoque automatizado de AutoML, ya que logra integrar múltiples modelos base en un ensamblado que optimiza el rendimiento sin intervención manual. Además, el hecho de que todos los modelos superen un AUC de 0.81 respalda la calidad del conjunto de datos y la viabilidad del problema para abordarlo con aprendizaje automático.

6. Implementación Aplicativa del Modelo Predictivo

Con el objetivo de demostrar el uso práctico del modelo desarrollado, se construyó una aplicación interactiva utilizando la biblioteca Streamlit. Esta herramienta permite desplegar modelos de machine learning mediante una interfaz sencilla y accesible para usuarios no técnicos.

La aplicación fue diseñada con una interfaz amigable en español, dividida en dos secciones principales:

- Sección introductoria: Explica brevemente la finalidad del sistema, su enfoque educativo, y brinda instrucciones de uso para el usuario final.

Sistema de Predicción de Enfermedad Cardíaca

Este proyecto académico utiliza modelos de aprendizaje automático para predecir el riesgo de padecer enfermedad cardíaca en base a información personal y hábitos de salud.

¿Cómo usar esta herramienta?

1. Completa el formulario con tus datos.
2. Haz clic en **Evaluar**.
3. Revisa tu nivel de riesgo y recibe una recomendación.

|  *Nota: Esta herramienta no reemplaza el diagnóstico médico profesional.*

- Formulario de predicción: Permite ingresar variables personales como edad, salud percibida, actividad física, diagnóstico de diabetes, entre otros factores. Internamente, estas variables son traducidas al formato esperado por el modelo entrenado en H2O, el cual genera una predicción sobre la probabilidad de padecer enfermedad cardíaca.

Evaluación de Riesgo Personal

Edad

18-24

Salud general

Pobre

Dificultad al caminar

Sí

Diabetes

Sí

Días de mala salud física (últimos 30)

0

Fuma

Sí

Actividad física

Sí

Cáncer de piel

Sí

Evaluar

El resultado es mostrado de manera interpretativa, clasificando el riesgo como alto o bajo, acompañado de una recomendación preventiva, como por ejemplo, visitar a un médico especialista o continuar con buenos hábitos.

Esta implementación no solo evidencia la capacidad predictiva del modelo, sino que también resalta su valor educativo y su posible utilidad como sistema de apoyo a la decisión en entornos de salud pública.

6.1. Evidencias de funcionamiento

Evaluación de Riesgo Personal

Edad

80 o más

Salud general

Pobre

Dificultad al caminar

Sí

Diabetes

Sí

Días de mala salud física (últimos 30)

0 29 30

Fuma

Sí

Actividad física

No

Cáncer de piel

Sí

Evaluar

⚠ Riesgo **alto** de enfermedad cardíaca. (Probabilidad: 50.67%)

Se recomienda consultar con un especialista médico.

A screenshot of a mobile application titled "Evaluación de Riesgo Personal". The interface is dark-themed with white and light gray text. It contains several dropdown menus and input fields for user information: age (80 or older), general health (poor), difficulty walking (yes), diabetes (yes), physical health days (29), smoking (yes), physical activity (no), and skin cancer (yes). Below these is a red button labeled "Evaluar". A purple bar displays a warning: "⚠ Riesgo alto de enfermedad cardíaca. (Probabilidad: 50.67%)". At the bottom is a green bar with the recommendation: "Se recomienda consultar con un especialista médico.".

Evaluación de Riesgo Personal

Edad

18-24

Salud general

Pobre

Dificultad al caminar

Sí

Diabetes

No

Días de mala salud física (últimos 30)

3

0 30

Fuma

Sí

Actividad física

No

Cáncer de piel

No

Evaluar

Riesgo bajo de enfermedad cardíaca. (Probabilidad: 92.33%)

Sigue cuidando tu salud con buenos hábitos. 🌱🏃‍♂️💪

7. Conclusiones

- A partir del análisis del conjunto de datos del BRFSS y la implementación de modelos de aprendizaje automático sobre la plataforma H2O, se logró construir un sistema predictivo eficaz para detectar riesgo de enfermedad cardíaca utilizando variables fácilmente recolectables (edad, salud autopercibida, diabetes, actividad física, etc.).

- Entre los modelos evaluados, el enfoque de AutoML demostró el mejor rendimiento global ($AUC = 0.8247$ y $Gini = 0.6495$), superando ligeramente a modelos individuales como GLM y Boosting. Esto valida el potencial del ensamblado automático de modelos en tareas de clasificación médica.
- Las variables que mayor impacto tuvieron en la predicción fueron la percepción del estado de salud general, la edad, la presencia de diabetes, la actividad física y las dificultades para caminar, lo cual coincide con hallazgos clínicos previos en estudios epidemiológicos.
- El análisis exploratorio mostró que el conjunto estaba limpio y balanceado en su mayoría, permitiendo un preprocesamiento sencillo. Asimismo, el uso de métricas como AUC y Gini permitió evaluar adecuadamente la capacidad discriminativa de cada modelo.
- Finalmente, se comprobó que, incluso sin variables clínicas invasivas o costosas, es posible obtener predicciones razonables que podrían servir como apoyo inicial en entornos de salud pública o educación preventiva.

8. Recomendaciones

- Ampliar la validación en contextos latinoamericanos: Dado que el conjunto de datos proviene exclusivamente de población estadounidense, se recomienda probar el modelo en muestras locales o adaptar sus parámetros con datos de Perú u otros países de la región para mejorar su aplicabilidad.
- Incluir variables clínicas más detalladas: Aunque el modelo actual se basa en información autodeclarada, se sugiere incorporar indicadores más específicos como niveles de colesterol, presión arterial o antecedentes familiares, lo cual podría aumentar la precisión del sistema predictivo.
- Desarrollar una versión accesible al público: La aplicación construida podría evolucionar hacia una herramienta de consulta interactiva y amigable para usuarios sin conocimientos técnicos, brindando orientación preventiva en centros de salud o campañas comunitarias.
- Monitorear el sesgo y la equidad del modelo: Es importante evaluar el desempeño del sistema en diferentes subgrupos poblacionales (por edad, género, condición

socioeconómica, etc.) para garantizar decisiones justas y evitar disparidades en las recomendaciones.

- Continuar con el entrenamiento iterativo del modelo: A medida que se recolecten más datos locales o se disponga de nuevas variables, se recomienda actualizar el modelo con nuevas iteraciones de entrenamiento para mantener su vigencia y precisión.

9. Referencias

Carrillo-Larco, R. M., Pacheco-Barrios, N., & Altez-Fernández, C. (2019). Risk prediction tools for cardiovascular diseases in Latin America and the Caribbean: A systematic review. *Global Heart*, 14(3), 203–213.
<https://globalheartjournal.com/articles/10.1016/j.ghert.2019.05.001>

World Health Organization. (2023). Cardiovascular diseases (CVDs).
[https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))