Machine Learning para predecir enfermedades del corazón

Para cumplir con parte de los requerimientos establecidos para obtener el título de Magíster en Ciencia de Datos

Autor:

Anthony Mora (ID: 1097694)

Asesor:

Renato González Disla

Instituto Tecnológico de Santo Domingo

Área de ingenierías

Maestría en Ciencia de Datos (MCD)

Santo Domingo, RD

12/18/22

Machine Learning para predecir enfermedades del corazón

Anthony Mora, Renato González Disla

Resumen

Los eventos cardiovasculares (ECV) son de las principales causas de fallecimientos a lo largo del mundo. Al ser la salud una de las prioridades a garantizar para las personas, resulta de interés la elaboración de herramientas para la pronta detección de enfermedades del corazón. Este trabajo busca mostar el desempeño de algoritmos de machine learning ya posicionados como estándar en la industria, para resolver una tarea de clasificación y así predecir la probabilidad de un paciente padecer de una enfermedad cardiovascular. Los datos a utilizar provienen del Heart Disease dataset obtenido del repositorio de datos de la Universidad de California, Irvine (UCI).

Palabras clave: Algoritmos predictivos, aprendizaje de máquina, evento cardiovascular

Abstract

Cardiovascular events (CVDs) are one of the leading causes of death throughout the world. Since health is one of the priorities to guarantee for people, it is of interest to develop tools for the early detection of heart disease. This work seeks to show the performance of machine learning algorithms already positioned as a standard in the industry, to solve a classification task and thus predict the probability of a patient suffering from cardiovascular disease. The data to be used come from the Heart Disease dataset obtained from the University of California, Irvine (UCI) data repository.

Keywords: Predictive algorithms, machine learning, cardiovascular event

Contenido

[1. Introducción 4](#_Toc122370494)

[2. Planteamiento del problema 4](#_Toc122370495)

[2.1 Objetivo general 4](#_Toc122370496)

[2.2 Objetivos específicos 4](#_Toc122370497)

[2.3 Revisión de Literatura 5](#_Toc122370498)

[2.4 Metodología 6](#_Toc122370499)

[3. Resultados 7](#_Toc122370500)

[4. Conclusiones 7](#_Toc122370501)

[Bibliografía 7](#_Toc122370502)

# 1. Introducción

El siglo XXI se ha visto marcado por el acelerado desarrollo de las tecnologías de información, esto ha implicado avances en el poder de procesamiento y la capacidad de almacenamiento de nuestras computadoras. Estas nuevas herramientas en manos de la humanidad han permitido el descubrimiento de los datos como uno de los recursos más preciados y abundantes de la actualidad, ya que el día a día del hombre moderno se encuentra muy relacionado con la constante generación de datos.

Los investigadores de inteligencia artificial (IA) y el machine learning (ML) se han dado a la tarea de aprovechar los conocimientos de distintas ramas como las ciencias estadísticas, computacionales y cognitivas para resolver problemas con la ayuda de las computadoras. Esto por medio de la elaboración de algoritmos que identifican patrones en los datos para ser aplicados en tareas como predicción o clasificación, estas siendo las más populares.

Según la Organización Mundial de la Salud (OMS), las enfermedades cardiovasculares son la principal causa de muertes a nivel mundial, con un estimado de 17.9 millones de personas fallecidas por eventos cardiovasculares (ECV) en el 2019, mayormente por cardiopatía coronaria y accidentes cerebrovasculares (ACV).

Las rápidas transformaciones económicas (industrialización, mercantilización, urbanización, globalización e informacionalización) han contribuido a prolongar la vida de las personas, los estilos de vida poco saludables y los cambios medioambientales. Predisponiendo a aumentar los factores de riesgo cardiovascular y la pandemia de ECV. Crecientes tasas de ECV han tenido un impacto económico, que amenaza al sistema de medicina y a la sociedad (Wu et al., 2016).

En décadas recientes, se han realizado grandes avances en investigación y práctica cardiovascular con el objetivo de mejorar el diagnóstico y tratamiento de enfermedades cardiacas, así como la reducción de la mortalidad por ECV (Chen et al., 2020).

El machine learning es cada vez más utilizado dentro del campo de la medicina cardiovascular (Krittanawong et al., 2020), donde vemos aplicaciones como medicina de precisión, predicciones clínicas, análisis de imágenes cardíacas y robots inteligentes (Yan et al., 2019). La promesa de la IA y el machine learning en cardiología es proveer una serie de herramientas para aumentar y extender la efectividad del cardiólogo (Johnson et al., 2018).

# 2. Planteamiento del problema

## 2.1 Objetivo general

Nuestro principal objetivo es predecir si un paciente sufre de alguna enfermedad del corazón utilizando algoritmos de machine learning.

## 2.2 Objetivos específicos

* Realizar un análisis de correlación entre las variables explicativas
* Comparar el desempeño de los distintos algoritmos implementados, a modo de elaborar un marco de referencia
* Mostrar técnicas estándar de preprocesamiento para mejorar el desempeño de los algoritmos

## 2.3 Revisión de Literatura

Para resolver la tarea de clasificación de este Heart Disease dataset se han implementado varios modelos de Machine Learning y Deep Learning, y estos reportan altas métricas de accuracy en sus predicciones.

Podemos encontrar aplicaciones recientes de técnicas de Machine Learning establecidas como multilayer perceptron (MLP) y K-Nearest Neighbour (KNN) implementados por Pal et al. (2022) con resultados de accuracy de 82.3% y área bajo la curva ROC de 86.4%. Ahmad et al. (2022) implementaron 11 modelos de clasificación, donde con métodos como Random Forest Classifier (RFC), Decision Trees (DT) y Gradient Boosted Classifier (GBC) obtienen métricas de accuracy entre el 97%-100%. El desempeño de estos modelos es mejorado por medio de un Stacking CV Classifier, donde además de implementar k-5 Cross Validation para medir el desempeño de los algoritmos, tambien combinan los resultados de varios clasificadores por medio de este Ensamble Model.

Budholiya et al. (2020) proponen un sistema de diagnóstico que utiliza XGBoost Classifier (Extreme Gradient Boosting) optimizado con Bayesian Optimization como método de selección de hiperparámetros, además de técnicas de One Hot Encoding para facilitar el aprendizaje de las variables categóricas al modelo, obteniendo un accuracy de 91.8%.

Spencer et al. (2020) determinan que la variable más relevante para este ejercicio de clasificación es el dolor en el pecho del paciente, seguida por la evaluación de angina luego de ejercitarse, los niveles de colesterol y pruebas de talasemia, con esta información aplican diversos modelos apoyados en técnicas de selección de variables predictoras para alcanzar métricas de accuracy entre 81%-85%.

Gao et al. (2021) utilizan técnicas de Ensamble Learning como Boosting y Bagging, en combinación con métodos de extracción de predictores como Análisis de Discriminantes Lineales (LDA) y Análisis de Componentes Principales (PCA). Sus resultados indican que la aplicación de bagging con Decision Trees y PCA obtiene los mejores resultados con un accuracy de 98.6%.

Al momento de las técnicas de estado del arte de la Inteligencia Artificial (AI) son las arquitecturas de Redes Neuronales propuestas por la rama del Deep Learning. Zhang et al. (2021) combinan selección de variables incrustadas con un algoritmo de Linear Support Vector Classifier (SVC) y norma L1 como penalidad para elegir las variables relacionadas significativamente con la variable a predecir de enfermedad cardíaca, para alimentar estas características identificadas a una Red Neuronal Profunda que muestra resultados de accuracy de 98.6%.

Mehmood et al. (2021) logran predecir la presencia de una enfermedad cardiovascular con un accuracy del 97%, al implementar Redes Neuronales Convolucionales, apoyándose en selección de variables a través de un análisis de correlación y preprocesamiento de datos como normalización de variables continuas. Das et al. (2009) introducen su metodología de Redes Neuronales con ensamblado, donde crean nuevos modelos al combinar las probabilidades posteriores de múltiples redes neuronales, mejorando así la eficiencia de este modelo final y obteniendo un accuracy de clasificación del 89%.

Mienye et al. (2020) proponen un método de dos etapas donde primero entrenan un autocodificador disperso, el cual es una red neuronal no supervisada que aprende una nueva representación de los datos de entrenamiento, luego en la segunda etapa utilizan una red neuronal con esta nueva representación de datos para predecir la condición de enfermedad cardiovascular en los pacientes analizados.

## 2.4 Metodología

Para este trabajo se ha utilizado el Heart Disease dataset obtenido del repositorio de datos de la Unversidad de California, Irvine (UCI), el cual ha sido utilizado para la predicción de enfermedades cardíacas por medio de diversos algoritmos de clasificación.

El mismo está compuesto por 303 observaciones de las cuales identificamos 6 con valores faltantes que fueron sustraídos del dataset a utilizar. El dataset original dispone de 76 variables por sujeto, de las cuales estudios pasados indican las 13 que han sido más efectivas para la detección de enfermedades cardíacas. La descripción de las variables se encuentra en la Tabla 1.

El propósito de este documento es mostrar el poder predictivo del Machine Learning en tareas del sector salud, como la identificación del riesgo de padecer una enfermedad cardíaca.

Como comentábamos anteriormente, los datos a utilizar corresponden al Cleveland Heart Disease dataset. Luego de realizar un análisis exploratorio de datos identificamos registros con información desconocida, los cuales fueron suprimidos. Adicional, el análisis de correlación entre las variables mostró que los niveles de colesterol y azúcar en la sangre presentan una baja correlación con nuestra variable dependiente.

Como parte del preprocesamiento de los datos realizados la segmentación de nuestros datos de entrenamiento y validación, utilizando una proporción de 80/20. Las variables categóricas a utilizar fueron transformadas por medio de técnicas como one hot encoding, donde las variables categóricas se convierten en un vector que indica 1 para la categoría que corresponda; y normalización de las variables numéricas, que implica escalarlas para que mientras que las numéricas pasaron por un proceso de normalización.

Entre los algoritmos implementados se encuentran Support Vector Classifier (SVC), Regresión Logística, K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Gradient Boosted Classifier (GBC), además realizamos una implementación de una Artificial Neural Network (ANN) y el uso de herramientas optimizadas para trabajar con información tabular y redes neuronales como fastai.

Los resultados de los métodos tradicionales de machine learning muestran valores de accuracy por encima de 90%, siendo KNN y Random Forest Classifier los de mejor desempeño. Cabe destacar que aún los modelos basados en redes neuronales mostrando un desempeño entre el 81%-88%, logran minimizar la función de costo en mejor medida.

# 3. Resultados

| Clasificador | Accuracy (%) | Recall (%) | Precision (%) | F1-Score (%) | ROC-AUC (%) |
| --- | --- | --- | --- | --- | --- |
| Support Vector Classifier | 90.0% | 90.0% | 90.0% | 90.0% | 98.0% |
| Regresión Logística | 90.0% | 90.0% | 90.0% | 90.0% | 98.0% |
| K-Nearest Neighbors | 93.0% | 93.0% | 93.0% | 93.0% | 94.0% |
| Decision Tree Classifier | 83.0% | 83.0% | 83.0% | 83.0% | 83.0% |
| Random Forest Classifier | 92.0% | 92.0% | 92.0% | 92.0% | 96.0% |
| Gradient Boosted Classifier | 90.0% | 90.0% | 90.0% | 90.0% | 94.0% |
| ANN | 88.1% | 70.8% | 85.0% |  | 87.3% |
| fastai tabular | 81.4% | 79.3% | 82.1% | 80.7% | 88.1% |
| Pal et al. (2022) - MPL | 82.47% |  |  |  | 86.41% |

# 4. Conclusiones

Hemos visto como a través de modelos de machine learning y deep learning como podemos obtener predicciones superiores a la intuición, utilizando como insumo registros médicos en formato tabular.

El desempeño de estos algoritmos es mejorado al realizar técnicas de preprocesamiento como one hot encoding y normalización de las variables categóricas y numéricas respectivamente, en conjunto con selección de variables relevantes a partir de un análisis de correlación.

# Bibliografía

Ahmad, G., Shafiullah, Fatima, H., Abbas, M., Rahman, O., Imdadullah y Alqahtani, M. (2022). Mixed Machine Learning Approach for Efficient Prediction of Human Heart Disease by Identifying the Numerical and Categorical Features. *Applied Sciences*, *12*, 7449. <https://doi.org/10.3390/app12157449>

Budholiya, K., Shrivastava, S. y Sharma, V. (2020). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*, *34*. <https://doi.org/10.1016/j.jksuci.2020.10.013>

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W. y Rueckert, D. (2020). Deep Learning for Cardiac Image Segmentation: A Review. *Frontiers in Cardiovascular Medicine*, *7*. <https://doi.org/10.3389/fcvm.2020.00025>

Das, R., Turkoglu, I. y Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Syst. Appl.*, *36*(4), 7675-7680.

Gao, X.-Y., Amin Ali, A., Shaban Hassan, H. y Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*, *2021*, 1-10.

Johnson, K. W., Torres Soto, J., Glicksberg, B. S., Shameer, K., Miotto, R., Ali, M., Ashley, E. y Dudley, J. T. (2018). Artificial Intelligence in Cardiology. *Journal of the American College of Cardiology*, *71*(23), 2668-2679. https://doi.org/<https://doi.org/10.1016/j.jacc.2018.03.521>

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L. y Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci. Rep.*, *10*(1), 16057. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7525515/>

Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T. y Masood, M. (2021). Prediction of heart disease using deep convolutional neural networks. *Arab. J. Sci. Eng.*, *46*(4), 3409-3422.

Mienye, I. D., Sun, Y. y Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. *Inform. Med. Unlocked*, *18*(100307), 100307.

Pal, M., Parija, S., Panda, G., Dhama, K. y Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100-1113. <https://doi.org/doi:10.1515/med-2022-0508>

Spencer, R., Thabtah, F., Abdelhamid, N. y Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digit. Health*, *6*, 2055207620914777.

Wu, Y., Benjamin, E. J. y MacMahon, S. (2016). Prevention and control of cardiovascular disease in the rapidly changing economy of China. *Circulation*, *133*(24), 2545-2560.

Yan, Y., Zhang, J.-W., Zang, G.-Y. y Pu, J. (2019). The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine? *J. Geriatr. Cardiol.*, *16*(8), 585-591.

Zhang, D., Chen, Y., Chen, Y., Ye, S., Cai, W., Jiang, J., Xu, Y., Zheng, G. y Chen, M. (2021). Heart disease prediction based on the embedded feature selection method and deep neural network. *J. Healthc. Eng.*, *2021*, 6260022.