Predicción y diagnóstico de enfermedades cardíacas UCI.

Kimy Agudelo Jaramillo*
Instituto de Física, Universidad de Antioquia,
Calle 70 # 52-21, Apartado Aéreo 1226, Medellín, Colombia
(Dated: 17 de diciembre de 2021)

En este trabajo se presenta la predicción y diagnostico de enfermedades cardíacas, siendo está la principal causa de muerte en la actualidad. Esto fue logrado mediante la implementación de algoritmos supervisados de Machine Learning, por los cuales se logró realizar una clasificación de pacientes con mayor probabilidad de contraer enfermedades cardíacas. Se contaron con 304 datos de pacientes con diferentes características (un total de 14) que fueron analizados al procesar los datos con los respectivos algoritmos implementados. Dichos algoritmos son presentados en el lenguaje de *Python*.
Keywords: Machine Learning, enfermedades cardíacas,
Usage: Tarea del curso Computación Avanzada 2021-2.

I. INTRODUCCIÓN

El Ministerio de Protección Social de Colombia reportó para el año 2011 una epidemia de Enfermedades Cardiovasculares (ECV), dentro de las cuales se destaca la isquémica cardiaca o infarto considerándolas responsables del mayor número de casos fatales [1]. En dicho año el número de muertos reportados alcanzó una cifra de 29.000 personas, lo cual implica que en promedio 5 personas perdieron la vida por cuenta de enfermedades cardíacas y cardiovasculares [2]. Hay diferentes causas por las cuales las personas pueden poseer problemas con enfermedades cardíacas, entre las cuales se mencionan los factores hereditarios y presión alta en mayor porcentaje, además de factores como la diábetes o los desarrollados por malos hábitos alimenticios o por fumar. Por lo tanto, este hecho es una fuerte motivación para realizar investigaciones alrededor de dicha problemática, donde las soluciones que se pretenden encontrar rondan en los diagnósticos, como en el seguimiento de la enfermedad y posibles mejorías.

El objetivo de este trabajo es el análisis de datos reales de pacientes a los cuales se les tomaron una cierta cantidad de datos médicos (presián, diábetes, etc.) así como datos personales (edad, sexo) para realizar una determinación de qué personas eran/son más propensas a padecer alguna enfermedad cardíaca, esto se logró realizar a partir de algortimos Machine Learning. Este documento se estructura de la siguiente está compuesto por la siguiente estructura: en la sección II se presenta el marco teórico, donde se aborda temas como enfermedades cardíacas y máquinas de aprendizaje, en la sección III se muestran y discuten los resultados obtenidos al realizar la implementació de los algoritmos utilizados. Finalmente, en la sección IV se presentan las conclusiones del trabajo.

* kimy.agudelo@udea.edu.co

II. MARCO TEÓRICO

A. Machine Learning

Machine Learning (aprendizaje automático) hace referencia a cualquier tipo de programa (o algoritmo) que pueda .ªprender"por sí mismo sin tener que ser programado explícitamente por un humano [3]. En 1950 Alan Turing publicó un artículo titulado Çomputing Machinery and Intelligence.en el cual se mencionaban las "Máquina de aprendizaje"que podían engañar a los humanos, esto lo lograría haciéndole creer que es verdadero. En estos momentos Machine Learning hace referencia al análisis de macrodatos y la minería de datos, por lotato existen diferentes enfoques.

Estos algoritmos de aprendizaje pueden presentarse en tres tipos, donde cada uno dependerá de los requisitos de cada problema a resolver, y los tipos son los siguientes: algoritmos supervisados, no supervisados y por refuerzo.

B. Enfermedades cardíacas

La enfermedad cardíaca incluye una gran variedad de enfermedades que afectan el corazón, entre las cuales comprende las siguientes: Enfermedad de los vasos sanguíneos, como enfermedad de las arterias coronarias, problemas en el ritmo cardíaco (arritmias), defectos cardíacos de nacimiento (defectos cardíacos congénitos), enfermedad de las válvulas cardíacas, enfermedad del músculo cardíaco e infección del corazón, estás son algunas de las enfermedades cardíacas [5].

III. RESULTADOS Y ANÁLISIS

Los datos utilizados en este trabajo fueron recopilados de Kaggle Datasets. Este conjunto de datos está compuesto por 303 instancias (303 pacientes, ver figura 2, donde se muestra la cantidad de pacientes que presentan enfermedades al corazón y los que no) con 75 atributos,

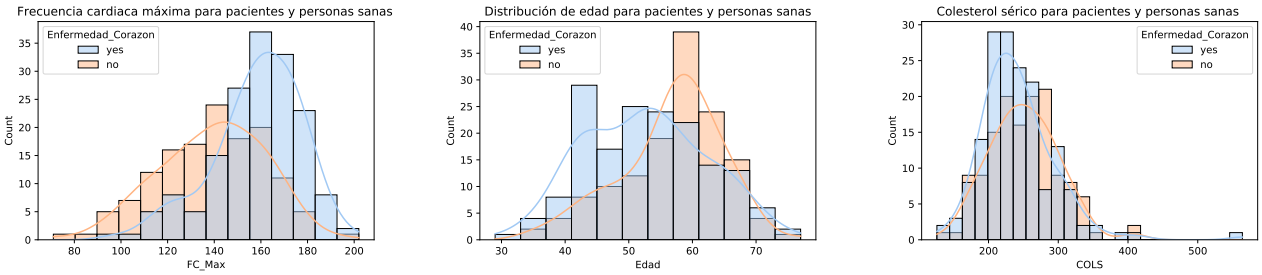

Figura 1: Distribuciones de los tres atributos más relevantes con respecto a la variable objetiva que corresponde a la enfermedad cardíaca

donde solo fueron utilizados 14, ya que según investigaciones son los que más presentan relevancia a la hora de realizar análisis sobre la predicción y diagnostico de enfermedades cárdiacas. Cabe resaltar que está base de datos ha sido utilizada para otras investigaciones de Machine Learning.

Los 14 atributos principales del dataset están dados por:

- Sexo, Edad (los cuales hacen parte de los datos personales)
- Presión arterial en reposo, Colesterol sérico en mg/dl, Tipo de dolor de pecho, Resultados electrocardiográficos en reposo, Máximo ritmo cardíacom alcanzado, Depresión inducida por el ejercicio relativo al descanso y Pendiente del segmento de ejercicio pico (los cuales hacen parte de los datos médicos)

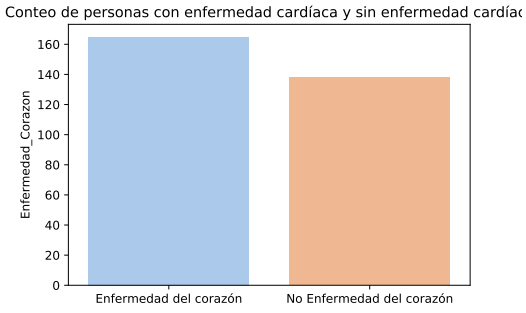

Figura 2: Conteo de pacientes en el dataset con enfermedad cardíaca y sin enfermedad cardíaca

Se tuvo en cuenta la cantidad de pacientes que sufren de algún tipo de enfermedad cardíaca fue importante en este análisis considerar el sexo de cada paciente, para realizar una comparación entre hombres y mujeres que presentan está afección, con el fin de obtener resultados sobre los posibles factores para cada uno de los sexos; en la gráfica (3) presentamos está parte del análisis de los datos, en donde se puede observar que en el conjunto de datos hay más mujeres que hombre (doblando la cantidad de hombres).

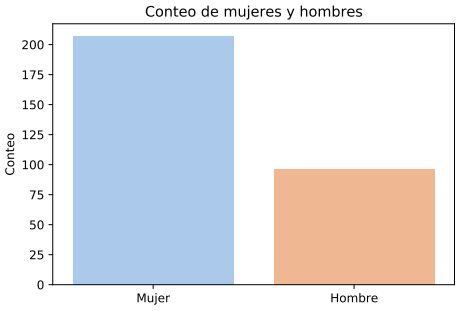

Figura 3: Conteo de pacientes hombres y mujeres que conforman el dataset

En este de análisis previo al entrenamiento de los datos para la implementación de árboles de decisiones se gráfico los atributos relevantes a considerar, tales como la frecuencia cardíaca máxima para pacientes que presentan enfemedades cardíacas como para pacientes que no presentan dicha enfermedad (ver figura 1), otro de los atributos relevantes fue el de la edad para pacientes que presentan este tipo de afección y los pacientes que no la presentan y finalmente se presenta el último atributo considerado y es el colesterol sérico para estos dos tipos de pacientes, como se observa en la figura ya referenciada la cantidad de pacientes enfermos del corazón que presentan frecuencia cardíaca máxima y colesterol sérico es casi la misma, mientras que las personas (contrario a lo que se llegaría a pensar) con mayor edad son los pacientes que menos presentan alguna enfermedad cardíaca o al corazón.

En la figura (4) se presenta la distribucción entre el sexo de los pacientes y el hecho de que presenten una enfermedad cardíaca o no, como se observa en dicho gráfico los hombres son los pacientes que más pesentan enfermedades cardíacas, alcanzando a las mujeres en un 75 % aproximadamente, aunque es muy importante considerar

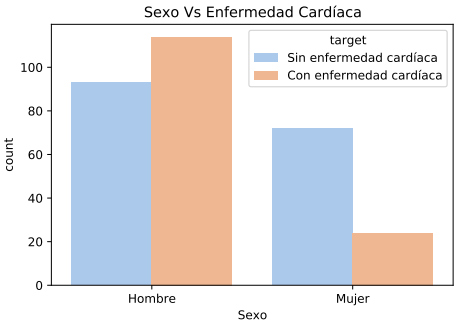

Figura 4: Distribucción entre el sexo de los pacientes y el hecho de que presenten una enfermedad cardíaca o no.

que en el conjunto de datos no hay la misma cantidad de hombres que de mujeres.

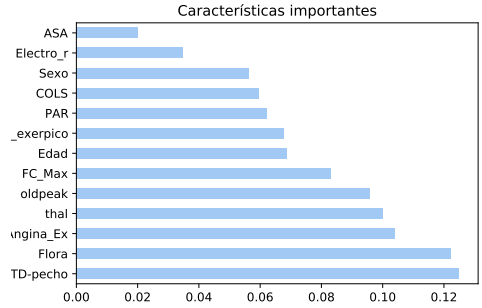

Figura 5: Importancia de cada característica del conjunto de datos utilizado

En la figura (4) se presenta la importancia de cada característica del conjunto de datos utilizado, El valor de la característica le da una puntuación para cada función de sus resultados; cuanto más alta sea la puntuación, más significativa o apropiada será la variable de rendimiento, por lo tanto como se observa en la figura, la variable con mayor importancia es la correspondiente *TD-Pecho: tipo de dolor en el pecho (4valores)*, mientras que la que mejor relevacia presenta es la correspondiente a *ASA: azúcar en sangre en ayunas* $> 120mg/dl$. La importancia de la característica es la clase incorporada que viene con los clasificadores basados en árboles.

La matriz de correlación con mapa de calor, representada en la figura (6), relacionan las características entre sí o con la variable de destino. La correlación puede ser positiva (el aumento de un valor de la característica aumenta el valor de la variable de destino) o negativa (el aumento de un valor de la característica disminuye el valor de la variable de destino). El mapa de calor facilita la clasificación de las características que son más relevantes a la variable de destino, para ello fue necesario utilizar la librería *seaborn*.

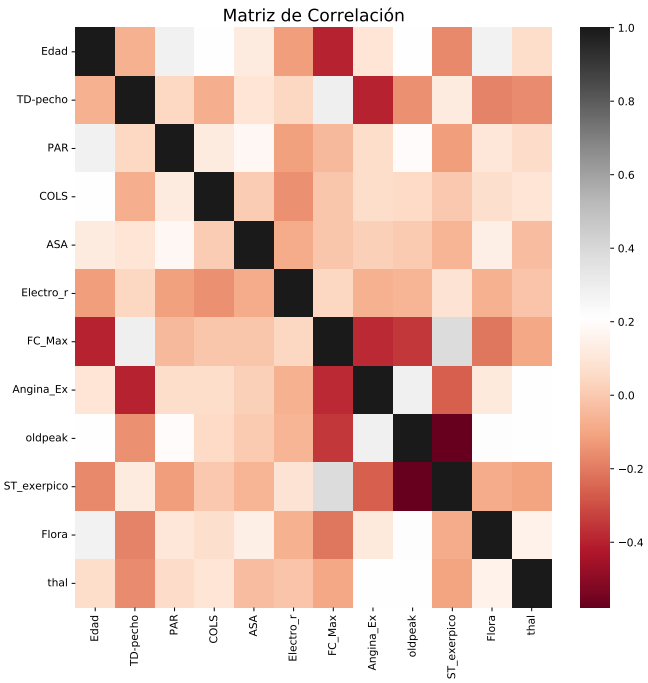

Figura 6: Matriz de correlación con mapa de calor

Finalmente se prepararon los datos para el entrenamiento y para poder obtener el árbol de decisiones, para ello se utilizaron las siguientes factores:

- Sexo
- Frecuencia Cardíaca Máxima
- Ritmo cardíaco.
- Diagnóstico de enfermedad del corazón
- Colesterol Sérico

Estos datos usados de entrenamiento suministran una base confiable para la ejecución del mismo. Usando esta distribución para la implementación inicial, obtuvimos un 79,5 % de precisión por parte del algoritmo implementado, en contra posición al obtenido inicialmente que fue de 73,8 %, es importante mencionar que en contraposición a los obtenidos por otras investigaciones el cual fue de 90,2 %, está diferecia puede ser atribuida a la implementación del algoritmo. Se debe considerar queel conjunto de datos (Dataset) no es lo suficientemente grande como para realizar un entrenamiento más exhaustivo del modelo, sin embargo, este como se ha visto proporciona resultados interesantes (ver figura 8)

En la figura 9 presentamos el árbol de decisión, el cual puede manejar datos continuos, pero no puede manejar datos categóricos con múltiples valores, como es nuestro

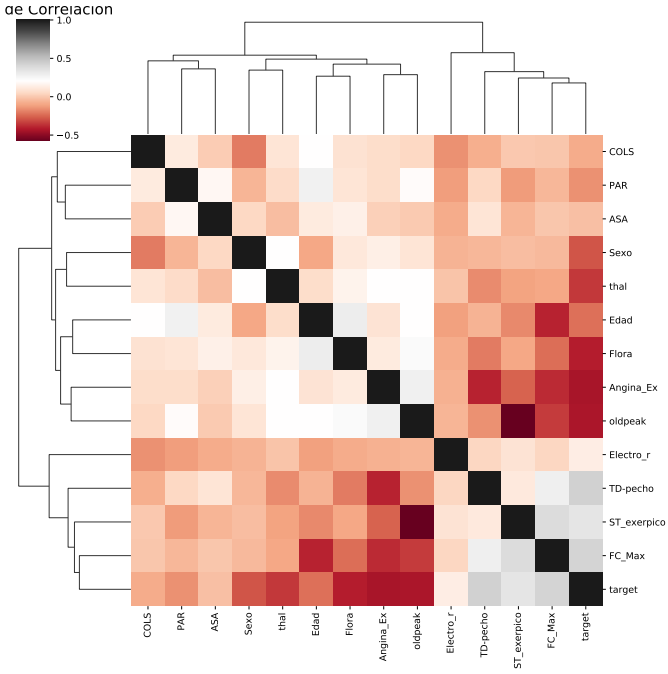

Figura 7: Matriz de correlación con mapa de calor

caso, por lo tanto se hizo necesario codificarlos. Estos árboles de decisión son modelos predictivos, los cuales se encuentran formados por reglas binarias tales como: si y no; por lo tanto lo que se busca es repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta.

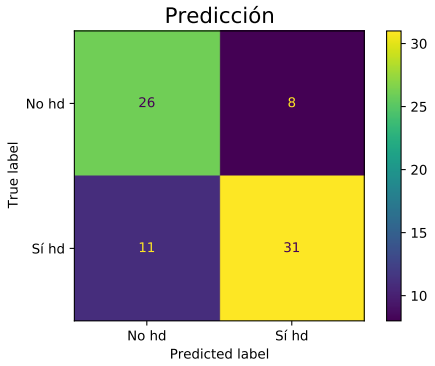

Figura 8: Predición

IV. CONCLUSIONES

En este trabajo, se se realiza un análisis a partir de algoritmos de Machine Learning para la clasificación de pacientes que sufren de enfermedades cardíacas y se compara su eficiencia y efectividad en predición con algortimos compartidos por investigaciones en el área de médicia, estos algoritmos se pueden encontrar en Kaggel Datasets. los resultados obtenidos en este trabajo muestra que el algoritmo no fue lo suficientemente eficiente o pedictivo, esto se debe a posibles errores en la implementación del código en *Python* o debido a que el conjunto de datos es muy pequeño como para realizar alguna predición a partir de este.

[1] Javier O Rodríguez V, Signed E Prieto B, Sandra C Correa H, Diagnóstico de la dinámica cardiaca durante 16 horas desde los sistemas dinámicos aplicable en UCI, Rev. Univ. Ind. Santander. Salud vol.49 no.1 Bucaramanga Jan./Mar. 2017.

[2] Ministerio de Salud y Protección Social. Colombia enfrenta epidemia de enfermedades cardiovasculares y diabetes. Boletín de Prensa No 077 de 2014.

[3] Agnieszka ŁAWRYNOWICZ aand Volker TRESP, Introducing Machine Learning, Institute of Computing Science, Poznan University of Technology, Poznán, Poland, January 2014.

[4] C. Nicholson, "A beginner's guide to neural networks and deep learning," https://pathmind.com/wiki/neural-network.

[5] Enfermedad Cardíaca, Mayo Clinic. https://www.mayoclinic.org/es-es/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

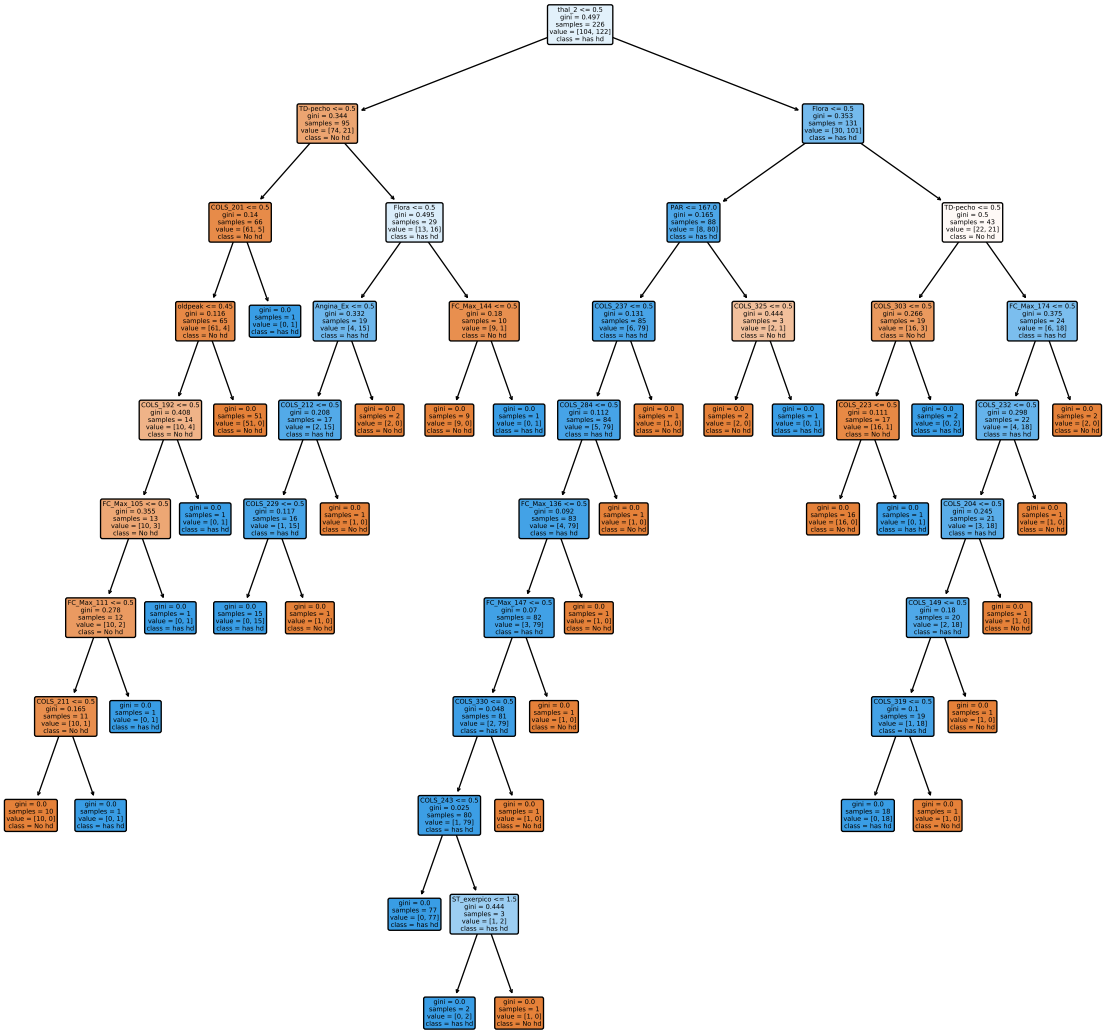

Figura 9: Árbol de decisión