PRÁCTICA 3 - INTERPRETABILIDAD

DOCUMENTACIÓN

Inteligencia Artificial
Yago Iglesias Díaz

Índice

1. Contexto

1.1 Descripción general del proyecto

El término “enfermedad cardíaca” se refiere a varios tipos de afecciones cardíacas, y estas enfermedades del corazón describen una gran variedad de condiciones que afectan a este músculo. Las enfermedades del corazón incluyen:

- Enfermedad de los vasos sanguíneos, como la enfermedad de las arterias coronarias
- Problemas del ritmo cardíaco (arritmias)
- Defectos cardíacos con los que nace (defectos cardíacos congénitos)
- Enfermedad de las válvulas del corazón
- Enfermedad del músculo cardíaco
- Infección del corazón

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial, cobrando un estimado de 17,9 millones de vidas cada año, lo que representa el 31% de todas las muertes en todo el mundo. Cuatro de cada 5 muertes por ECV se deben a ataques cardíacos y accidentes cerebrovasculares, y un tercio de estas muertes ocurren prematuramente en personas menores de 70 años.

2. Problema a resolver

Muchas formas de enfermedades cardíacas se pueden prevenir o tratar con opciones de estilo de vida saludables. Vamos a analizar datos sobre el conjunto de datos de predicción de enfermedades cardíacas para descubrir las causas y las características importantes que afectan considerablemente las posibilidades de enfermedades cardíacas. Mejorar estos índices puede ayudarnos a reducir la prevalencia de enfermedades del corazón.

2.1 Cuestiones objetivo

En este estudio, se intentará resolver los siguientes cuestiones:

- ¿Cuáles son los principales indicadores de que una persona pueda padecer una de estas enfermedades?
- ¿Qué es la interpretabilidad y cuales son sus modelos para realizar la interpretabilidad en este estudio?
- ¿Qué modelos de machine learning son los mejores y más precisos a la hora de aplicar interpretabilidad en este conjunto de datos?

2.2 Variable fundamental

Dado que se dispone de gran variedad de datos recolectados de personas referentes a la salud cardiovascular, se pretende predecir lo expuesto con anterioridad. Para ello, hay una variable objetivo, la cual es "HeartDisease" que funciona de la siguiente manera, si una persona tiene insuficiencia cardíaca se representa con un 1 y si no la tiene se hace con un 0. Esta variable es el objetivo del estudio, averiguar la probabilidad de padecer problemas del corazón.

3. Estrategia seguida

La estrategia seguida se ha compuesto de varios puntos, empezando por una breve definición de lo que es la interpretabilidad seguido de los métodos, algoritmos, y librerías que se van a utilizar. Mas adelante se comienza con el proceso de analizar el conjunto de datos y codificarlo para un correcto funcionamiento de este, como por ejemplo en la variable “Colesterol” la cual tenía una gran cantidad de datos inicializados a 0, siendo nulos, por lo que se eliminaron esta información. Y por último, se empieza a analizar los correspondientes modelos de machine learning para comprobar su interpretabilidad.

3.2 Análisis del Conjunto de Datos

En este caso de estudio, se cuenta con datos extraídos de sitio web Kaggle, donde a través de esta url “https://www.kaggle.com/fedesoriano/heart-failure-prediction” se han podido adquirir la información necesaria para desarrollar este análisis, contando con 918 personas con diversos datos sobre ellos. Y todos ellos serán usados para la predicción de que una persona pueda tener problemas de insuficiencia cardíaca.

Este conjunto de datos, creado por el usuario "fedesoriano", contiene un total de 35.92 kB de almacenamiento de información en formato CSV, 145 códigos públicos creados por la comunidad en base a este, y fue subido a la plataforma Kaggle en septiembre del año 2021.

4. Recursos usados

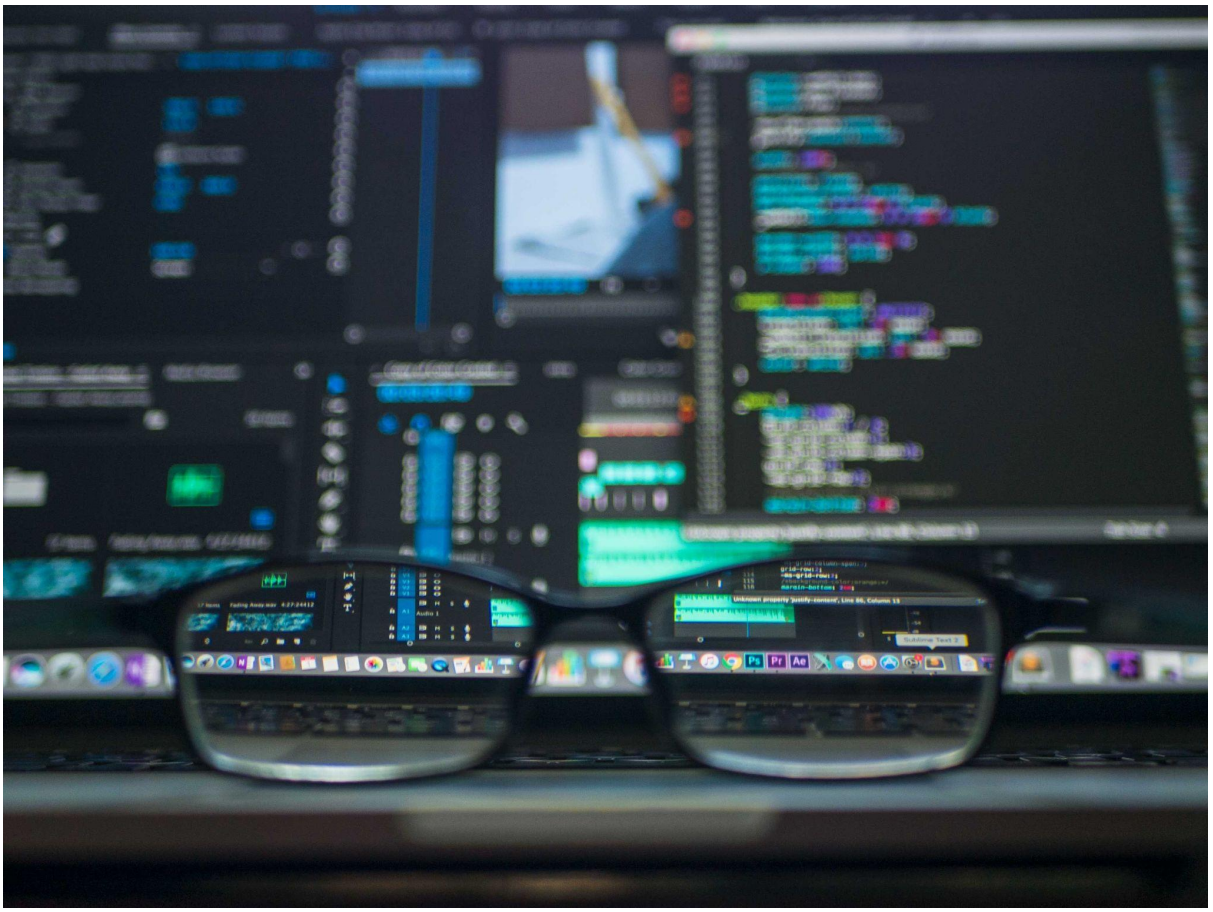
En este estudio se han utilizado alguna variedad de recursos para la correcta realización de este, ya que sin estos recursos, como pueden ser librerías y modelos de machine learning no es posible hacer el análisis. En primer lugar se presentarán las librerías y más tarde los modelos utilizados.

4.1 Librerías

En cuanto a las librerías utilizadas para el funcionamiento del código han destacado las siguientes:

- Pandas: es una librería de Python especializada en el manejo y análisis de estructuras de datos. Con él se han realizado visualizaciones gráficas que nos permiten determinar entre otros la dependencia y el grado de correlación entre variables del modelo.
- Numpy: da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.
- Sklearn: es un conjunto de rutinas escritas en Python para hacer análisis predictivo, que incluyen clasificadores, algoritmos de clusterización, etc.

- Matplotlib: es una biblioteca para la generación de gráficos a partir de datos contenidos en listas o arrays en el lenguaje de programación Python y su extensión matemática NumPy.
- Seaborn: es una librería para Python que permite generar fácilmente elegantes gráficos. Seaborn esta basada en matplotlib y proporciona una interfaz de alto nivel que es realmente sencilla de aprender.
- XGBoost: es uno de los algoritmos de machine learning de tipo supervisado más usados en la actualidad. Este algoritmo se caracteriza por obtener buenos resultados de predicción con relativamente poco esfuerzo



4.2 Algoritmos

En cuanto a los algoritmos implementados en este estudio se han utilizado primero una serie de algoritmos de base (utilizando hiperparámetros de fábrica) antes de pasar a soluciones más sofisticadas.

Más tarde se realiza el desarrollo de las técnicas de machine learning para su interpretabilidad: ELI5 library, Partial Dependence Plots, Skater Model Interpretation, LIME, SHAP Values, SHAP Summary Plots, FairML.

- ELI5 library: es una librería de Python que ayuda a explicar predicciones de forma fácil de entender e intuitiva. Es un buen punto de partida y soporta modelos basados en árboles y paramétricos/lineales y también utilidades de procesamiento de texto y HashingVectorizer de scikit-learn, pero no soporta verdaderas interpretaciones agnósticas del modelo.
- Partial Dependence Plots: "El gráfico de dependencia parcial (PD plot) muestra el efecto marginal que tienen una o dos características en el resultado predicho de un modelo de aprendizaje automático (J. H. Friedman 200127). Un gráfico de dependencia parcial puede mostrar si la relación entre el objetivo y una característica es lineal, monótona o más compleja". (de la documentación de PDPbox).
- Skater Model Interpretation: es un marco unificado para permitir la Interpretación de Modelos (post-model) construye un sistema de aprendizaje automático interpretable a menudo necesario para los casos de uso del mundo real utilizando un enfoque agnóstico de modelo.
- LIME: las explicaciones LIME se basan en modelos sustitutos locales. Los modelos sustitutos son modelos interpretables (como un modelo lineal o un árbol de decisión) que se aprenden sobre las predicciones del modelo original de caja negra.
- SHAP: (SHapley Additive exPlanations) asigna a cada característica un valor de importancia para una predicción concreta. Sus componentes novedosos incluyen: la identificación de una nueva clase de medidas de importancia de características aditivas, y resultados teóricos que muestran que hay una solución única en esta clase con un conjunto de propiedades deseables.
- FairML: es una biblioteca de Python que audita modelos predictivos de caja negra. La idea básica de FairML (y de muchos otros intentos de auditar o interpretar el comportamiento de los modelos) es medir la dependencia de un modelo de sus entradas cambiándolas.

5. Objetivos esperados y logrados

5.1 Objetivos esperados

Los objetivos esperados en este proyecto eran varios, como ya se mencionó anteriormente en el punto uno.

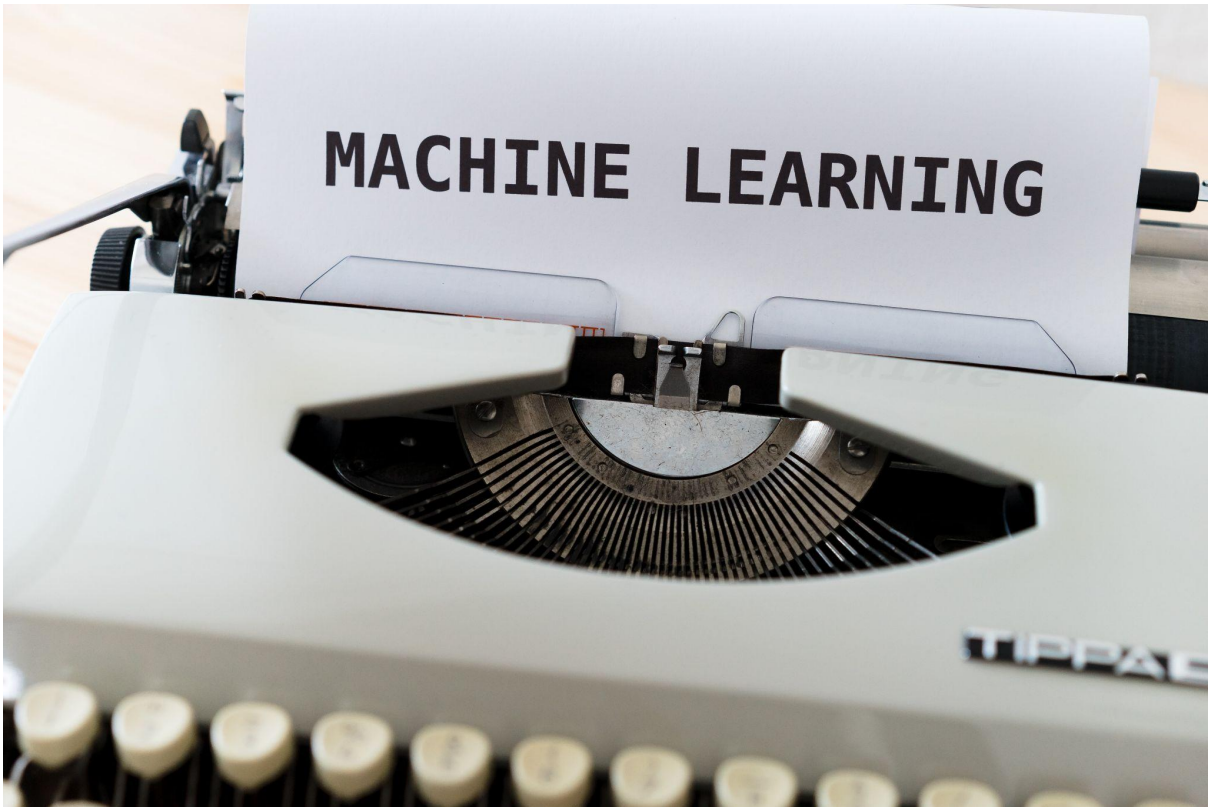
En primer lugar, se pretende analizar varios modelos de machine learning para saber su interpretabilidad, pero antes codificar los datos del conjunto de datos con el

fin que todas las características y su información esté lo menos sesgada posible para una interpretabilidad buena y fiable.

5.2 Objetivos logrados

En cuanto a los objetivos logrados, se puede afirmar que se han alcanzado casi la totalidad de los esperados, ya que se han explicado y realizado los siete modelos de machine learning correspondientes mencionados al principio del archivo jupyter. Por otro lado también se han podido realizar la visualización de los resultados mediante distintos gráficos, usando diferentes librerías para cada tipo de dato, como por ejemplo, el uso de js.

Pero por otro lado se han conseguido una gran cantidad de objetivos, así como averiguar la probabilidad de que una persona en específico pueda tener enfermedades cardiovasculares, o saber cuáles son los principales indicadores de que una persona pueda padecer una de estas enfermedades, como por ejemplo la prueba del esfuerzo, la cual es el mayor indicador según los algoritmos desarrollados de que una persona pueda llegar a tener una enfermedad cardíaca.



6. Conclusión

El estudio ha resultado exitoso ya que se han conseguido los objetivos esperados mencionados en el apartado anterior.

Por otro lado, también se ha apreciado que el conjunto de datos escogido de enfermedades del corazón ha resultado interpretable, basándose en los distintos modelos de machine learning probados.