Enfermedades cardiacas

Principal causa de muerte en España

Cristina Hidalgo Palacios

Contenido

Contexto

La idea de este proyecto surgió de la hipótesis de diseñar un producto que pudiera ser utilizado tanto por las entidades sanitarias como por las empresas aseguradoras. La salud y el dinero son dos aspectos fundamentales en la vida y en la sociedad consumista actual, y este producto busca ofrecer una solución que aborde ambos aspectos de manera efectiva.

Por ello, se quiere realizar un análisis exploratorio de datos que permita desarrollar un algoritmo de predicción de la enfermedad escogida.

Introducción

Para seleccionar la enfermedad que definirá el proyecto se han descargado datos del INE que contienen el total de muertes de la población española por causas y edad desde 2010-2021. Están disponibles públicamente en: https://www.ine.es/jaxiT3/Datos.htm?t=49075#!tabs-tabla.

Dos de las tres hipótesis planteadas en este proyecto son comprobar si la principal causa de muerte en España es debida a tumores cancerosos y si la principal causa de muerte en España la sufren más los hombres o las mujeres.

A partir de los datos arriba mencionados se ha obtenido la principal causa de muerte en España.

La principal causa de muerte en los principales países desarrollados y en España se debe a enfermedades cardiovasculares. Las cardiopatías son desde hace 20 años la causa principal de mortalidad en todo el mundo.

El pronóstico temprano de las enfermedades cardiovasculares puede ayudar en la toma de decisiones sobre cambios en el estilo de vida en pacientes de alto riesgo y/o a reducir las complicaciones de la enfermedad.

Esta investigación tiene como hipótesis principal la intención de identificar los factores de riesgo más relevantes de la enfermedad cardíaca.

El conjunto de datos con los que se ha realizado el proyecto ha sido recogido por la organización *Framingham Heart Study* y está disponible públicamente en el sitio web: https://www.framinghamheartstudy.org/ y en https://ocw.mit.edu/courses/15-071-the-analytics-edge-spring-2017/resources/framingham/. Estos datos provienen de un estudio cardiovascular en curso en los residentes de la ciudad de Framingham, Massachusetts. El objetivo de la clasificación es predecir si el paciente tiene un riesgo de enfermedad cardiaca (CHD) en los próximos 10 años (CHD10). Se ha escogido este conjunto de datos ya que se trata de datos muy completos y actualizados y pertenecen a una población del primer mundo con un estilo de vida cercano a las condiciones en España.

El conjunto de datos proporciona la información de los pacientes e incluye más de 4.000 registros y 15 atributos (factores de riesgo potencial). Existen factores de riesgo tanto demográficos como de hábitos y médicos. Las variables analizadas son: Sexo, Edad, Educación, Fumador Actual, Cigarrillos al día, Medicación para presión arterial,

Ictus previo, Hipertensión, Diabetes, Colesterol, Presión Arterial Sistólica, Presión Arterial Diastólica, Índice de Masa Corporal, Ritmo Cardiaco y Glucosa. La variable de destino es CHD10, que indica si una persona tiene un riesgo de desarrollar una enfermedad coronaria en los próximos 10 años.

Análisis Exploratorio de Datos

Para la realización del proyecto se ha seguido la siguiente línea de trabajo:

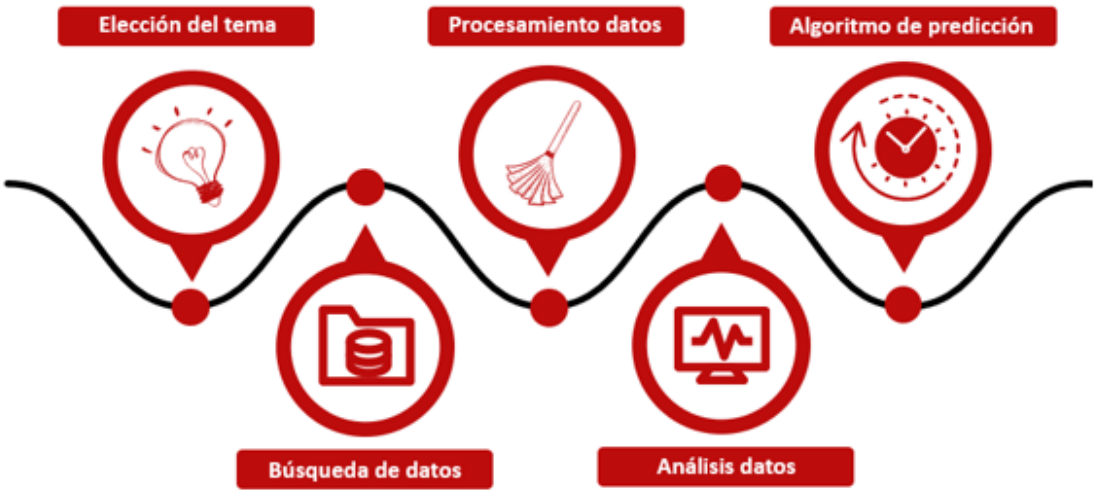

Figura 1 - Línea de trabajo del proyecto.

Una vez elegido el tema y encontrado la fuente de datos con la que se quería trabajar, se importaron las librerías y los datos necesarios.

A continuación, se procedió al procesamiento de datos. Para ello se comprobó si había datos nulos y/o duplicados y cómo afectaban a los mismos. Como no se podía sacar un resultado concluyente de aquellos pacientes con datos nulos, se procedió a eliminar dichas filas.

A continuación, se analizaron los outliers de las variables implicadas. Pese a tener outliers significativos, para el EDA no se han querido desestimar dichas variables ya que tampoco afectaban en gran medida a los resultados obtenidos. Pero sí se procederá a su eliminación para el proyecto de Machine Learning.

Una vez concluida con la limpieza de datos se procedió al análisis de los mismos. Se observó la correlación existente entre las variables y se decidió eliminar la variable “PAD” ya que estaba altamente correlacionada con la variable “PAS” y a su vez en la definición de las mismas. La presión arterial sistólica es la presión que ejerce la sangre sobre las paredes de las arterias cuando el corazón se contrae y bombea sangre hacia el cuerpo. La presión arterial diastólica es la presión que ejerce la sangre sobre las

paredes de las arterias cuando el corazón está en reposo entre latidos. Es decir, una PAS elevada conlleva una PAD elevada.

El resto de variables analizadas no sufrieron más cambios antes de proceder al análisis entre variables y su distribución se representa en la siguiente figura.

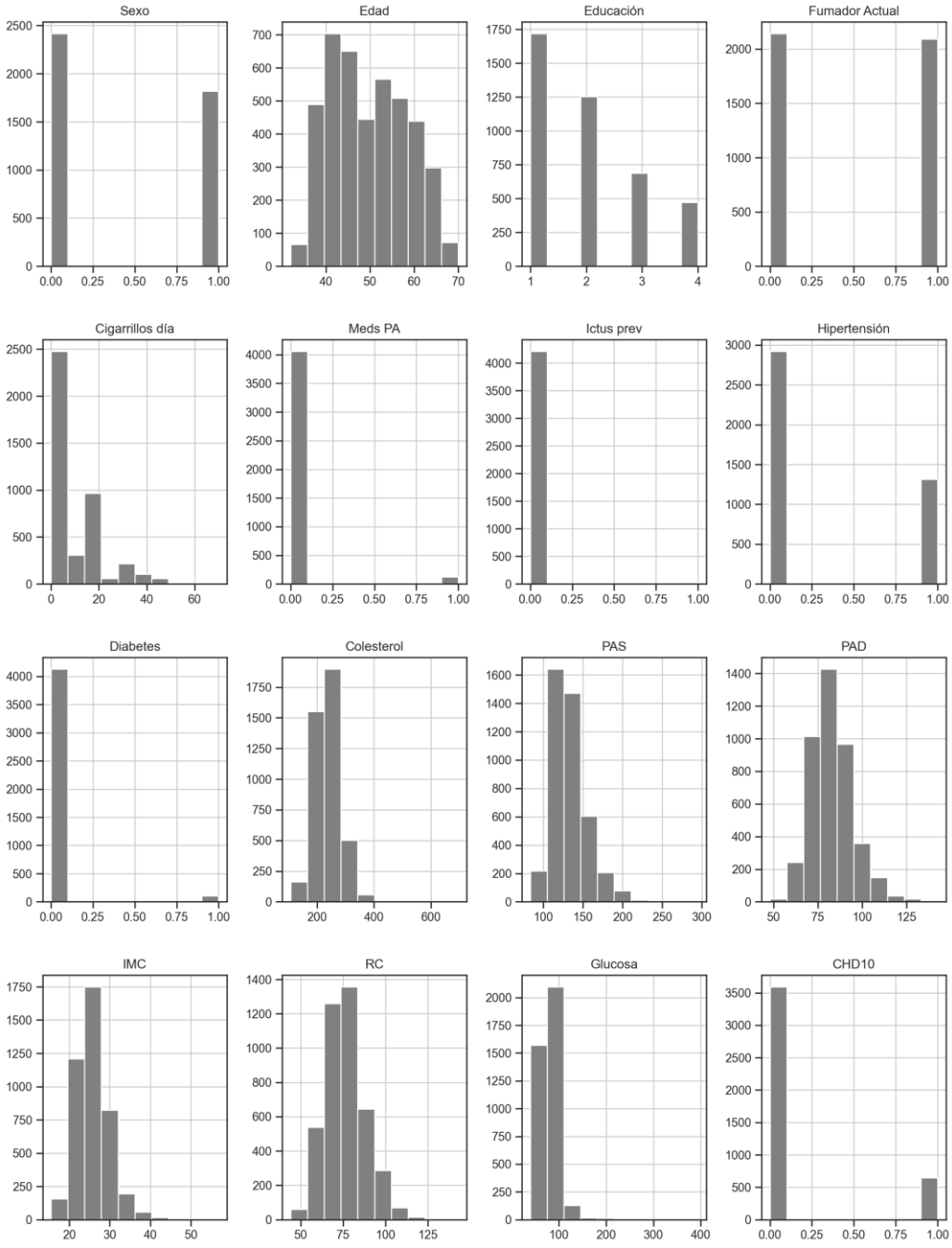

Figura 2 - Distribución de las variables implicadas.

Para conocer qué variables tienen mayor relación con la probabilidad de sufrir una enfermedad cardiaca se importó 'SelectKBest' y 'chi2' de la librería scikit-learn.

La prueba de chi-cuadrado se utiliza para determinar si dos variables categóricas están relacionadas o no. La prueba de chi-cuadrado se usa comúnmente en el análisis de datos para evaluar la relación entre dos variables.

Se separaron las variables independientes y dependientes de data. Las variables independientes son todas las columnas excepto CHD10, que se asignan a la variable X. La variable dependiente es la última columna, que se asigna a la variable Y.

A partir de este estudio se obtuvo el top 10 de las variables con mayor relación en la probabilidad de sufrir una enfermedad cardiaca.

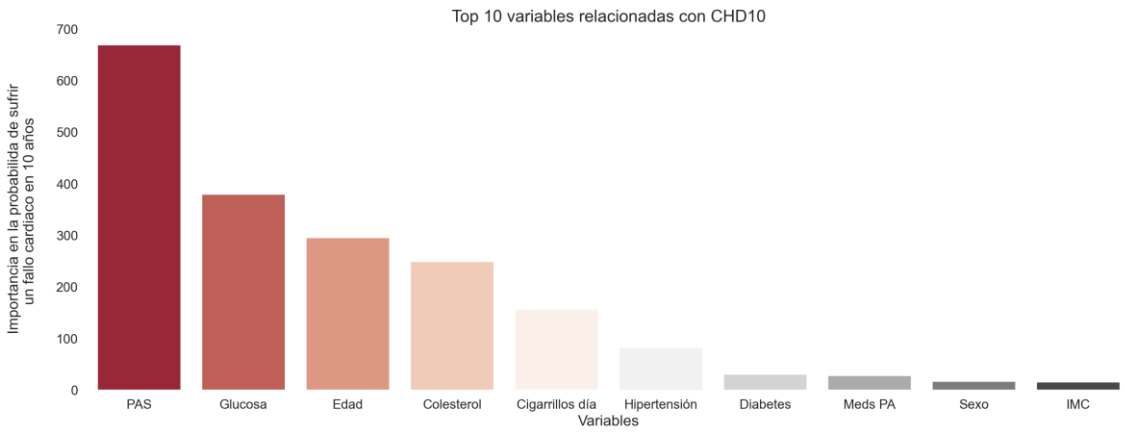

Figura 3 - Top 10 variables relacionadas con CHD10.

Para terminar, se graficó la relación entre las 10 variables del top con respecto a la probabilidad de sufrir una enfermedad cardiaca y se sacaron conclusiones sobre su comportamiento con la variable de salida. Finalmente, se comprobó si las hipótesis planteadas son válidas o no.