16-11-2021

Predicción de enfermedad cardiaca

**Ciencia de datos aplicada**

Andrés Felipe Vega Chacón - 201910197 Geovanny Andrés González - 201719528 Wilson Camilo Sánchez Forero - 201820379

**Universidad de los andes**

Tabla de contenido

[Introducción 2](#_Toc87896801)

[Objetivos 2](#_Toc87896802)

[Descripción del problema 3](#_Toc87896803)

[Enfoque Analítico 3](#_Toc87896804)

[Entendimiento de los datos 4](#_Toc87896805)

[Descripción de los datos 4](#_Toc87896806)

[Análisis de los datos 5](#_Toc87896807)

[Construcción del modelo 5](#_Toc87896808)

[Extreme Gradient Boosting (XGBoost) 6](#_Toc87896809)

[Perceptron Multi Capa 6](#_Toc87896810)

[Support Vector Classification - SVC 6](#_Toc87896811)

[Random Forest 6](#_Toc87896812)

[Conclusiones 8](#_Toc87896813)

[Despliegue del modelo 9](#_Toc87896814)

[Repositorio de código 10](#_Toc87896815)

[Referencias 10](#_Toc87896816)

## **Introducción**

De acuerdo con la Organización Mundial de la salud (OMS), las enfermedades cardiovasculares (ECV) son un conjunto de desórdenes del corazón que se presentan por lo general debido a la obstrucción en el sistema cardiovascular que impide la libre circulación de la sangre hacia el corazón o el cerebro.

En el año 2015 las enfermedades cardiovasculares fueron la principal causa de muerte, dejando 17.7 millones de personas muertas, lo que representa un 31% del total de muertes.

Generalmente estos problemas se presentan en personas que tienen malos hábitos alimenticios o que abusan de sustancias como el alcohol y el cigarrillo, aunque también factores como pobreza, estrés y componentes hereditarios deben ser considerados.

En su mayoría estos problemas se presentan en países con ingresos medios y bajos. Las causas obedecen por lo general a factores tales como: precariedad en los servicios de salud pública, un bajo control en la prevención del consumo en exceso de alimentos con elevados niveles de colesterol, azúcar y sal, así como una infraestructura pública que carece de espacios para la práctica de deportes como parques, senderos peatonales, espacios para practica de actividades aeróbicas entre otros.

Ante la preocupante cifra de muertes, la OMS presentó en el año 2013 una serie de mecanismos para tratar de reducir el nivel de personas con este tipo de enfermedades; uno de los principales propósitos de estas políticas se centra en la reducción de pacientes con niveles anormales de hipertensión, que es uno de los principales detonantes de fallas cardiacas a través del impulso de medidas para reducir la obesidad, consumo de alcohol y tabaco.

No menos importante, se presenta también un plan de acción mundial para asesoramiento y tratamiento en pacientes con elevados niveles de glucosa, dado que estas enfermedades se presentan en su mayoría por eventos que pueden ser atenuados y detectados a tiempo, resulta de vital importancia ver y analizar las causas que las producen, los factores de riesgo que incrementan las posibilidades de morir, es de vital importancia realizar un monitoreo y análisis constante de cualquier dato asociado que ayude a tomar decisiones sobre cómo aspectos como políticas púbicas, hábitos laborales, familiares y alimenticios podrían reducir o aumentar los casos de ECV.

## **Objetivos**

* Identificar cuáles son los factores con mayor incidencia en el padecimiento de la enfermedad cardiaca
* Aplicar modelos de machine learnig que permitan determinar si un paciente tiene alto o bajo riesgo de presentar enfermedad cardiaca

## **Descripción del problema**

La enfermedad cardiaca avanza lenta y silenciosamente en muchos de los pacientes, quienes consultan a los especialistas solo en casos críticos o en la existencia de una emergencia médica y es en ese momento en los que las probabilidades de superar las complicaciones son bajas.

Adicionalmente, la falta de acceso a especialistas, falta de controles periódicos y un deficiente sistema de salud hace que los pacientes no puedan consultar a tiempo o simplemente por percepción personal no deseen acercarse a los hospitales.

En todos los casos es necesario: i) agilizar los diagnósticos de los pacientes y ii) realizar seguimientos para tener diagnósticos a tiempo y que poder acceder a tratamientos que permitan prevenir o mejorar las condiciones de salud evitando complicaciones médicas.

Por lo anterior, proponemos implementar métodos de machine learning a la información del historial medico de los pacientes para poder tener un presunto diagnostico de que paciente puede padecer o estar padeciendo de una enfermedad cardiaca; esta herramienta le permitirá a los especialistas tener un apoyo en el diagnostico y a los usuarios un resultado prematuro para que tomen medidas de corrección y ajuste a tiempo.

## **Enfoque Analítico**

Los hospitales universitarios de Basel, Zurich, el Instituto Húngaro de Cardiología, el centro médico de Veteranos de Long Beach y la clínica de Cleveland han recolectado información relacionada con variables asociadas a los problemas cardiacos para caracterizar a los pacientes que pueden llegar a padecer o padecen enfermedades cardiacas.

Con esta información se desea implementar modelos de machine learning para obtener una recomendación de la probabilidad que puede tener un paciente de padecer una enfermedad cardiaca y así mismo mejorar el dictamen de los profesionales de la salud.

Dentro de los datos se tiene la variable “target”, la cual es una variable dicotómica que relaciona: *0 = ninguna o menos probabilidad de ataque cardíaco* y *1 = más probabilidad de ataque cardíaco*

Cada registro corresponde con la histórica clínica de un paciente. En este data set encontramos información de pacientes con características normales y pacientes con problemas cardiacos, por lo cual es un insumo importante para entrenar modelos de machine learnig para realizar la clasificación de ser propenso o no de padecer enfermedad cardiaca.

## Entendimiento de los datos

El data set de *Heart Disease* contiene con 14 atributos y 303 registros, donde cada registro corresponde con un paciente.

La información contenida en el data set corresponde con pacientes sin enfermedad, con enfermedad moderada y avanzada. Esto permite entrenar los modelos de machine learning en los diferentes casos que se puedan presentar.

### Descripción de los datos

* **Edad**: la edad mínima es 29, el máximo es 77 y el promedio es de 54 años.
* **Sex**: el 68% de la población es hombre y el 32% son mujeres
* **Cp** (tipo de dolor de pecho): es categórica y contiene:
  + Angina típica 143 registros
  + Angina atípica 50 registros
  + Sin dolor 87 registros
  + Asintomático 23 registros
* **Trestbps**: valor mínimo 94, valor máximo 200 y valor promedio 131 presión arterial en reposo
* **Chol**: valor mínimo 126, valor máximo 564 y valor promedio 246 mg/dl
* **Fbs**: el 85% de los datos es 0 y el 15% es 1
* **Restecg**: el 50% de los datos son 1, el 49% son 0 y el resto es 2
* **Thalach**: valor mínimo 71, valor máximo 202 y valor promedio 149
* **Exang**: el 67% es 0
* **Oldpeak**: valor mínimo 0, valor máximo 6.2 y valor promedio 1
* **Slope**: el 47% es 2 y el 46% es 1
* **Ca**: contiene 5 valores, donde 0 corresponde con el 58% de los datos
* **Tal**: índice de talasemia, tiene 4 valores, donde lo mas prominente es 2 es el 55% y 3 es el 39% de la población
* **Target**: el 54% de la población presenta enfermedad cardiaca

### Análisis de los datos

Realizando un análisis de cada uno de los campos dispuestos en el data set encontramos lo siguiente:

1. En su mayoría, los pacientes dentro del conjunto de datos son de 58 años con un 63% de participación.
2. La mayoría de los pacientes del conjunto de datos son hombres, con un 68,7% de participación. Esto se debe a que los hombres son más propensos a sufrir obesidad, tabaquismo e hipertensión, factores esenciales para aumentar el riesgo de padecer enfermedad cardiaca.
3. La presión arterial más frecuente es de 120.
4. De acuerdo con los análisis de las correlaciones, hemos encontrado que los valores que más influyen en el diagnóstico podrían ser:
   1. Tipo de dolor en el pecho
   2. Frecuencia cardiaca máxima lograda
   3. Pendiente del segmento ST

## Construcción del modelo

la clasificación de si una persona presenta un riesgo alto de sufrir un ataque cardiaco o no, se realiza implementando los siguientes modelos de machine learning:

* Máquina de soporte vectorial (SVM)
* XGBoost
* Perceptrón multicapa (MLP)
* Random Forest.

Las métricas empleadas son exactitud, precisión, f-score y entropía cruzada. Para seleccionar el mejor modelo nos vamos a enfocar en el que mayor f-score posea y menor entropía cruzada.

Modelos de clasificación implementados:

### Extreme Gradient Boosting (XGBoost)

Texto

Descripción generada automáticamente

Resultado:

Texto

Descripción generada automáticamente

### Perceptron Multi Capa

Texto

Descripción generada automáticamente

Resultado:

Texto

Descripción generada automáticamente

### Support Vector Classification - SVC



Resultado:

Texto

Descripción generada automáticamente

### Random Forest

Texto

Descripción generada automáticamente

Resultado:

Texto

Descripción generada automáticamente

## Conclusiones

1. Al generar variables dummies empleando una codificación One Hot Encoding a partir de las variables categóricas los modelos de clasificación quedan con un score de predicción más alto
2. El modelo de Random Forest es el que nos ha generado un score más alto, siendo para el ejercicio uno de los modelos con poder de predicción más alto obteniendo un F-score de 0.88
3. Con base en el F-score podemos ordenar la efectividad de los modelos de la siguiente manera:

* Random Forest: 0.87
* XGBoost: 0.79
* Red neuronal MLP: 0.71
* Máquina de soporte vectorial: 0.69

1. Con respecto a la distribución de los datos, es relevante mencionar nuevamente que, tras observar las distribuciones en los tres subconjuntos para cada una de las características de los datos podemos concluir que las tres particiones mantienen la misma distribución.
2. Según la información encontrada, las investigaciones se centran en los catorce campos objeto de este estudio. Se hizo la revisión del dataset original donde se evidencia que tal vez estas observaciones, a los ojos de los expertos fueron las más relevantes, y desde el punto de vista de la calidad de los datos, se observa, que el dataset original, de sus 76 columnas, una es el ID, y 24 columnas tienen un valor constante.

## **Despliegue del modelo**

Para realizar el despliegue del modelo vamos a emplear una aproximación mediante contenedores. Para la interfaz de usuario, vamos a emplear una página web de una vista desarrollada con el framework: Django en el cual podremos ingresar desde la vista del navegador un archivo con los datos requeridos de cada una de las variables predictoras.

A continuación, encontrará el enlace a un video explicativo del funcionamiento de la aplicación: <https://youtu.be/95cSBOLrVVA>

De igual modo, dejamos disponible un archivo de prueba si desea probar el funcionamiento de la aplicación. A continuación, encontrará el enlace de la aplicación para acceder mediante navegador web.

Aplicación: <http://34.133.250.118:8000/>

## **Repositorio de código**

A continuación, se encuentra disponible el repositorio con el desarrollo del análisis descriptivo e inferencial sobre los datos del problema:

Enlace: <https://github.com/CDA-2021-20/cda-grupo-5>

Asimismo, en el siguiente enlace se encuentra disponible el código de la aplicación del proyecto: <https://github.com/CDA-2021-20/cda-aplicacion>

## **Referencias**

Enfermedades Cardiovasculares (2017, 17 Mayo). OMS. <https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>

Heart Disease Dataset (1988, 07 Enero) Donald Bren School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Heart+Disease