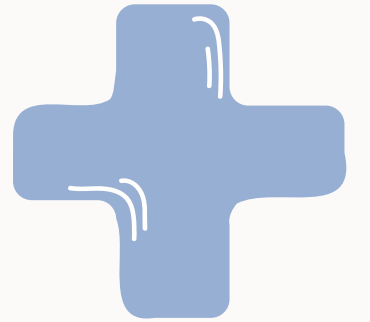


PREDICCIÓN DE ENFERMEDADES CARDÍACAS



CENTERS FOR DISEASE

ÍNDICE

1

Introducción

2

Objetivos

3

**Orígenes de
datos**

4

**Principales
indicadores**

5

**Algoritmos de
clasificación**

6

Tareas realizadas

7

Conclusiones

8

Integrantes

INTRODUCCIÓN



INTRODUCCIÓN

Según estudios realizados por los Centros para el Control y Prevención de Enfermedades (CDC) una de las principales causas de muerte en las personas en los EE.UU es debido a enfermedades cardíacas. La detección y prevención de factores de riesgo en una etapa temprana pueden salvar muchas vidas. La aplicación de métodos de aprendizaje automático para detectar patrones en base a los datos pueden incidir positivamente en la salud de la población.



CENTERS FOR DISEASE

OBJETIVOS



OBJETIVOS

El objetivo de esta investigación es poder predecir la probabilidad de que una determinada persona pueda contraer una enfermedad cardiaca, dependiendo de distintos factores tales como: la edad, el promedio de horas de descanso, la obesidad (IMC), si es fumador, si bebe alcohol, entre otras.



ORÍGENES DE DATOS



ORÍGENES DE DATOS

Encuesta anual 2020

El conjunto de datos proviene de los CDC y es una parte importante del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS), que realiza encuestas telefónicas anuales para recopilar datos sobre el estado de salud de los residentes de EE. UU. Fue establecido en 1984 con 15 estados, BRFSS ahora recopila datos en los 50 estados. El conjunto de datos más reciente (al 15 de febrero de 2022) incluye datos de 2020. Los datos con los que trabajaremos provienen de la encuesta anual de los CDC de 2020 de 319.795 adultos relacionados con su estado de salud.



PRINCIPALES INDICADORES



PRINCIPALES INDICADORES

Índice masa corporal (IMC).

Horas de descanso

Salud Física

Salud Mental

Actividad Física

Rango Edad

Sexo

Fumador

Alcoholismo

Actividad física

Antecedentes de
enfermedades



ALGORITMOS DE CLASIFICACIÓN



Árbol de decisión

El algoritmo de árboles de decisión es un algoritmo de clasificación para su uso en el modelado predictivo de atributos discretos y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos.

Random Forest

En Random Forest se ejecutan varios algoritmos de árbol de decisiones en lugar de uno solo. Para clasificar un nuevo objeto basado en atributos, cada árbol de decisión da una clasificación y finalmente la decisión con mayor "votos" es la predicción del algoritmo.

KNN (Vecinos próximos)

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual.

TAREAS REALIZADAS



TAREAS REALIZADAS

1

CARGA DEL SET DE DATASET

A lo largo de este análisis se realizó la carga del set de datos original. Se creó un DF el cual se analizó sus variables, las métricas que tenía. Se realizó un trabajo de limpieza y transformación sobre los datos.

2

APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING

Luego se aplicaron los algoritmos de machine learning los cuales fueron:

1. Árbol de decisión
2. Bosque aleatorio
3. Vecinos cercanos

3

PREDICCIONES

Una vez realizado el entrenamiento de los modelos con el 70% de los datos, se realizó una predicción con el 30% restante.

Para mejorar el desempeño de los algoritmos utilizamos el GridsearchCV. Modificando los Hiperparametros

4

CONCLUSIONES

A partir de los resultados obtenidos se compararon los resultados.

| CUADRO COMPARATIVO | accuracy | precision | recall | |
|-------------------------|--------------|-------------|-------------|-------|
| <i>Arbol decisión</i> | 0.9155171181 | 0.563568773 | 0.039721217 | train |
| | 0.913987012 | 0.532534246 | 0.037515078 | test |
| <i>Random Forest</i> | 0.99706507 | 0.996122778 | 0.969344442 | train |
| | 0.904616475 | 0.350468912 | 0.121712907 | test |
| <i>Vecinos cernanos</i> | 0.934931384 | 0.717980889 | 0.389823403 | train |
| | 0.89726805 | 0.323886639 | 0.173703256 | test |

CONCLUSIONES



CONCLUSIONES

Este trabajo tenía como finalidad aplicar modelos de inteligencia artificial para poder predecir el porcentaje de si una persona tenía probabilidad de contraer una enfermedad cardíaca. A lo largo del análisis pudimos observar las variables que más influencia tenía, como lo fue el caso de la edad. Se aplicaron 3 algoritmos de clasificación ya que el problema era una clasificación bivariable. Los algoritmos que se implementaron fueron:

1. Árbol de decisión
2. Bosque aleatorio
3. Vecinos cercanos



CONCLUSIONES

Concluimos que el algoritmo de clasificación que mejor rendimiento obtuvo fue el Árbol de decisión (decision tree) a que es muy baja la diferencia entre el promedio de entrenamiento y el promedio de evaluación. además, fue el algoritmo que menos casos falsos positivos obtuvo y teniendo en cuenta que se utilizará para predecir si una persona puede ser diagnosticada enferma o no puede ser un valor determinante.





PUNTOS A MEJORAR

Desbalanceo de carga

Uno de los desafíos que tuvimos con este dataset fue el desbalance que había entre la cantidad de casos que no presentaban enfermedad cardíaca con los casos que si presentaban enfermedad. Para ello se prodría aplicar la técnica del crossvalidation.

Agregar el modelo ensamble

Al ser un problema de clasificación bimétrica se podría incorporar el modelo de ensamble para comparar con los demás modelos utilizados.

INTEGRANTES

QUE REALIZARON ESTA INVESTIGACIÓN



EQUIPO ENCARGADO DE LA INVESTIGACIÓN



NICOLÁS HERRERA

Data Engineer



NICOLÁS BALBIANI

Data Sciences



MATÍAS VITOLA

Data Analytics

MUCHAS GRACIAS

