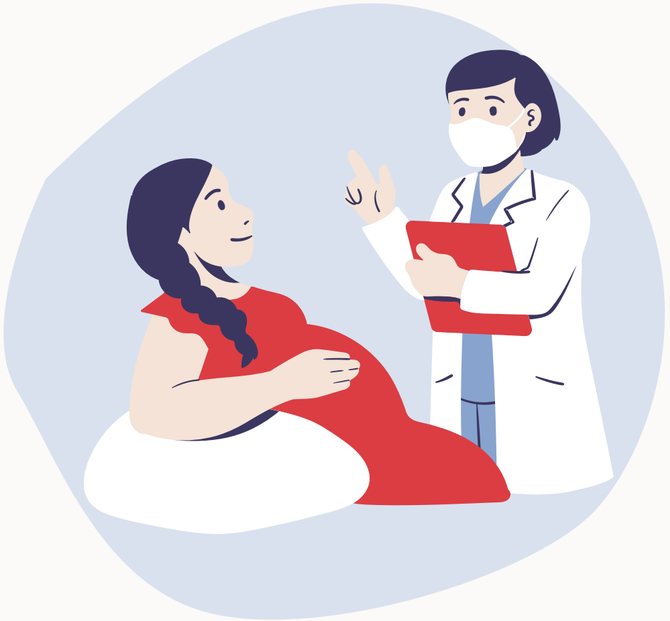
PREDICCIÓN DE ENFERMEDADES CARDÍACAS



CENTERS FOR DISEASE

CODER HOUSE

ÍNDICE

CODER HOUSE

CODER HOUSE

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.

CODER HOUSE

CODER HOUSE

ORÍGENES DE DATOS

ORÍGENES DE DATOS

Encuesta anual 2020

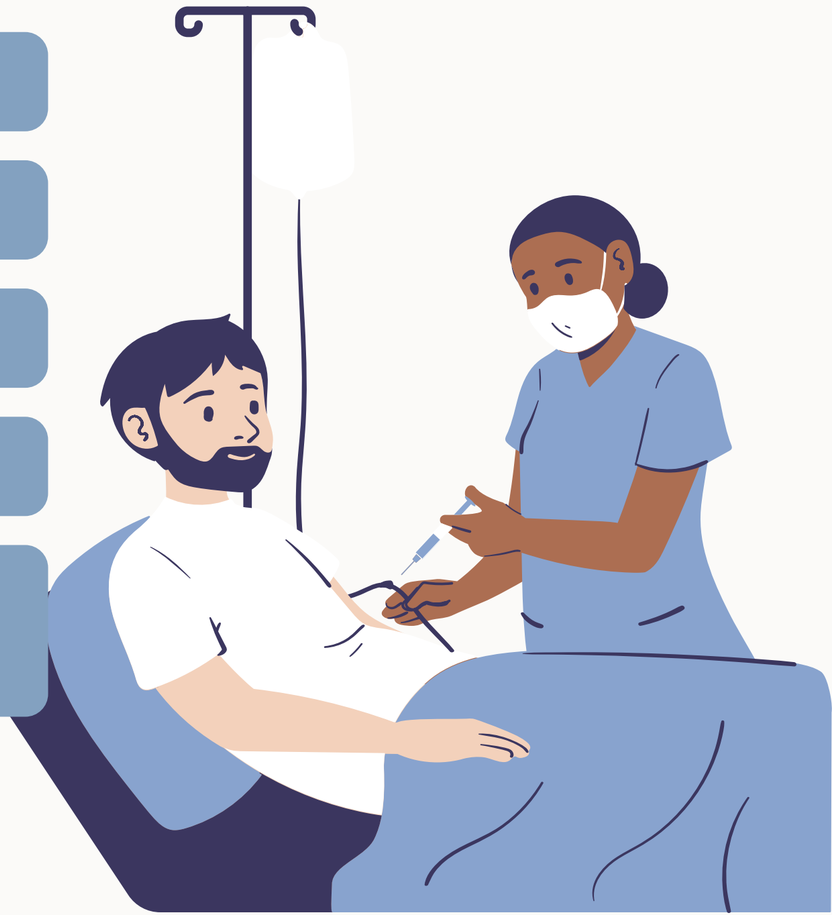
El conjunto de datos proviene de los CDC y es una parte importante del Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS), que realiza encuestas telefónicas anuales para recopilar datos sobre el estado de salud de los residentes de EE. UU. Fue establecido en 1984 con 15 estados, BRFSS 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.

CODER HOUSE

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



CODER HOUSE

CODER HOUSE

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.

Regresión Logística

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica.

CODER HOUSE

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. Regresión Logística

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

CONCLUCIONES

A partir de los resultados obtenidos se compararon los resultados.

CUADRO COMPARATIVO	accuracy	precision	recall	
Arbol decisión	0.9155171181	0.563568773	0.039721217	**train**
	0.913987012	0.532534246	0.037515078	**test**
Random Forest	0.99706507	0.996122778	0.969344442	**train**
	0.904616475	0.350468912	0.121712907	**test**
Regresión Logística	0.9163748123	0.54645508	0.1118796834	**train**
	0.914476907	0.526545908	0.101688781	**test**

CODER HOUSE

CONCLUSIONES

CODER HOUSE

CONCLUSIONES

Este trabajo tenía como finalidad la aplicación 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 tenían, como por ejemplo, fue el caso de la edad.
Se aplicaron 3 algoritmos de clasificación ya que el problema era una clasificación binaria. Los algoritmos que se implementaron fueron:

1. Árbol de decisión
2. Bosque aleatorio
3. Regresión Logística

CONCLUSIONES

Concluimos que el algoritmo de clasificación que mejor rendimiento obtuvo fue el Árbol de decisión (desicion tree), ya 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.

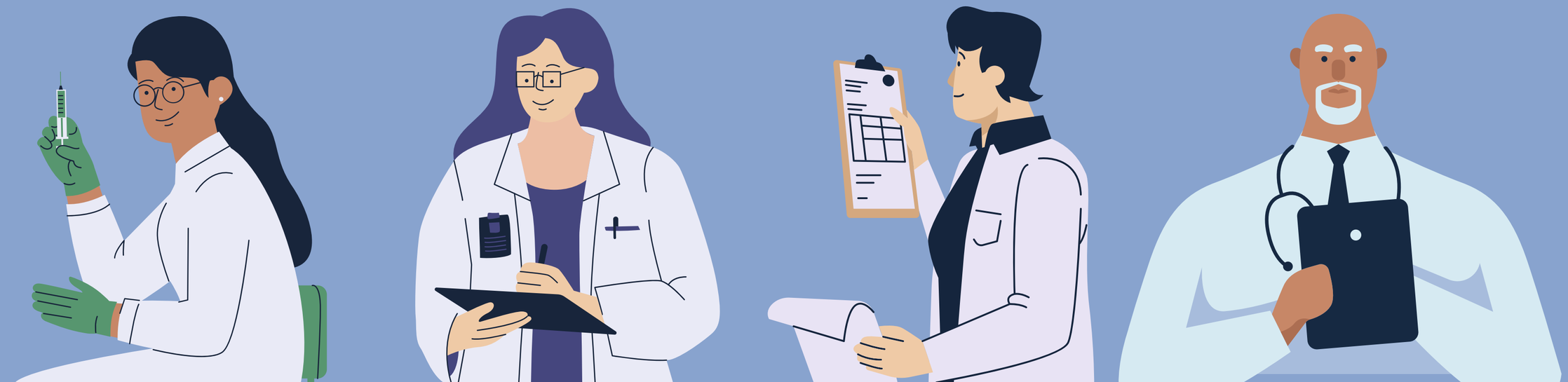
Agregar el modelo ensamble

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

CODER HOUSE

INTEGRANTES

QUE REALIZARON ESTA INVESTIGACIÓN



EQUIPO ENCARGADO DE LA INVESTIGACIÓN

NICOLÁS HERRERA

Data Engineer

NICOLÁS BALBIANI

Data Sciences

MATÍAS VITOLA

Data Analytics

CODER HOUSE

CODER HOUSE

MUCHAS GRACIAS