Infraestructura del proyecto de clasificacion

Machine Learning

Addison Amin Reyes Cedano, 2021-2026

Mi proyecto de clasificacion tine como proposito identificar si un paciente es fumador o no en base a su edad, altura, peso, presion sistólica, presion en sangre, glucosa en ayunas, colesterol, hemoglobina, proteina en orina, caries dentales, etc...

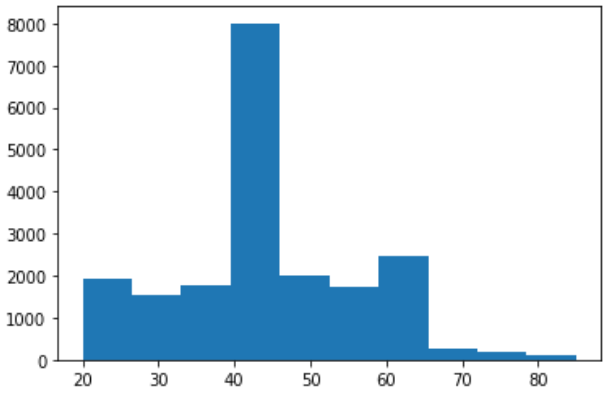
Lo primero en el proyecto fue buscar un dataframe en kaggle, este fue el dataframe que mas me gusto y mas se adecua a mis necesidades, luego de tener instalado el dataframe inicie el proyecto importando e instalando las librerias y dependencias necesarias para el proyecto.

Una vez escritas todas las librerias incia el proceso de recoleccion de data y preparacion o preprocesammiento en donde utilice pandas para poder extraer el csv, una vez extraido limpio el dataframe eliminando las columnas innecesarias y revisando que no existan valores nulos.

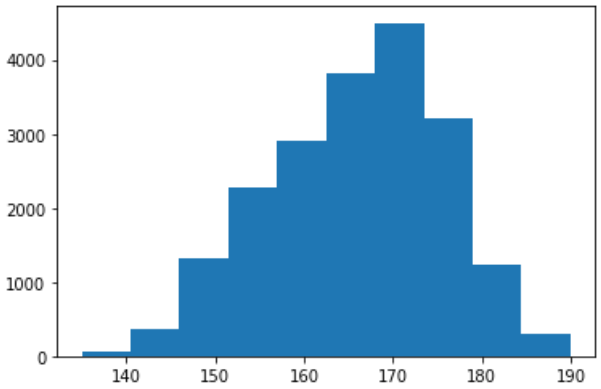
Despues de limpiar el dataframe divido los datos en dos dataframes a 10,000 fumadores y 10,000 no fumadores, concateno los dos para crear un solo dataframe con el que trabajare durante todo el proyecto de ahora en adelante.

Para el análisis descriptivo de la data (EDA) utilice dataprep y matplotlib, aquí algunos graficos relacionados con la data:

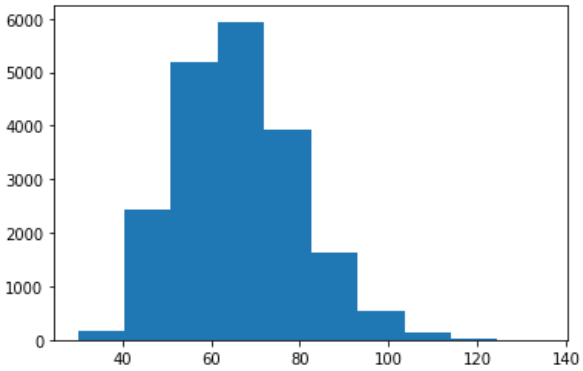
- Edad de todos los pacientes



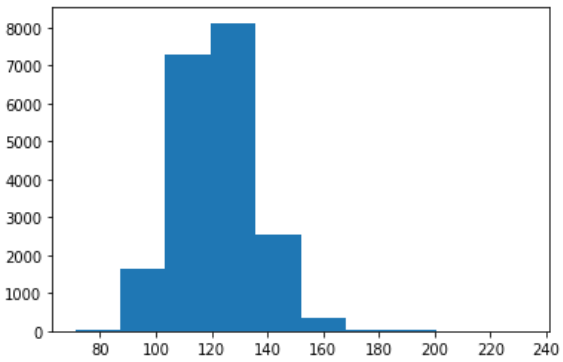
- Altura de todos los pacientes



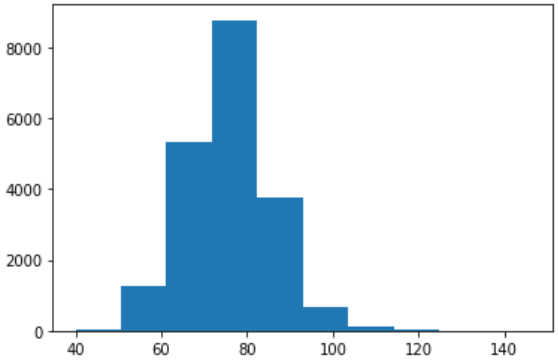
- Peso de todos los pacientes



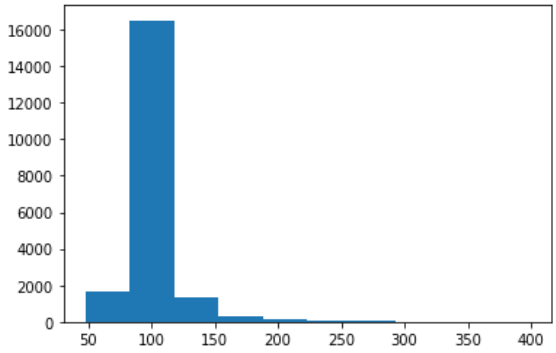
- Sistolica de todos los pacientes



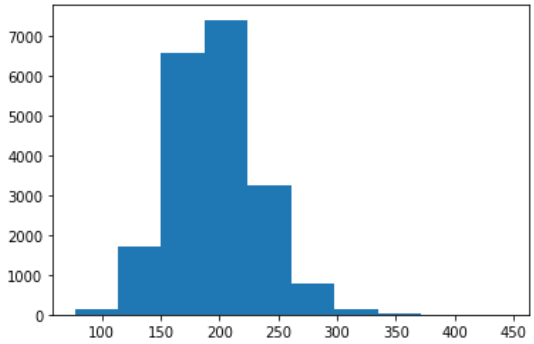
- Presion en la sangre de todos los pacientes



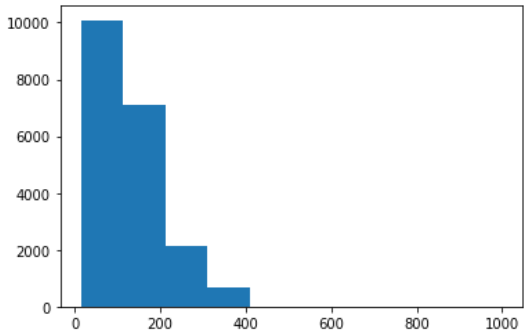
- Azucar en ayunas de todos los pacientes



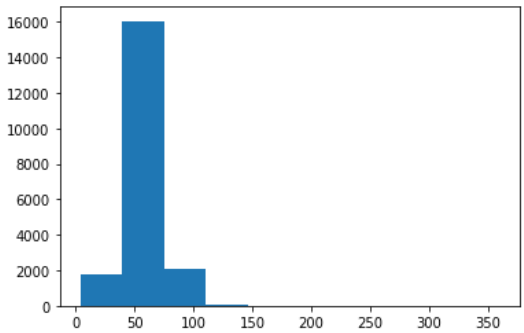
- Colesterol de todos los pacientes



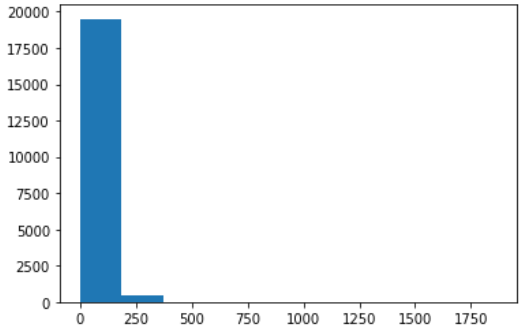
- Triglicelidos de todos los pacientes



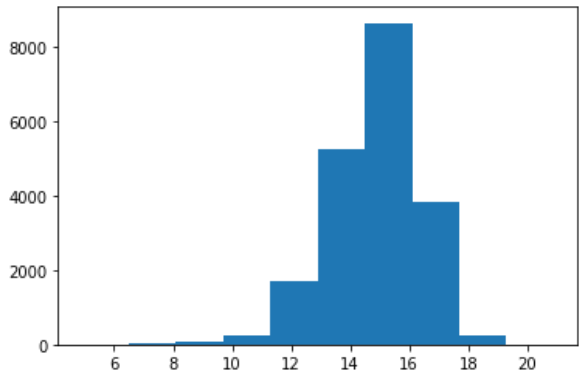
- Colesterol bueno de todos los pacientes



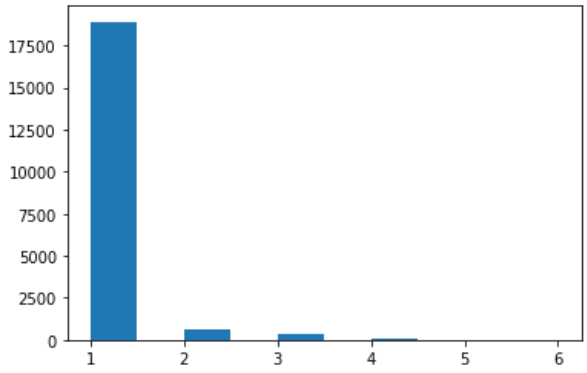
- Colesterol malo de todos los pacientes



- Hemoglobina de todos los pacientes



- Proteina en orina de todos los pacientes



Despues de analizar los graficos hechos con matplotlip y dataprep empece a entrenar el modelo con sklearn, primero cree los valores de testeo y entrenamiento con train_test_spli(), despues de tener los valores inicie los modelos KNeighborsClassifier, RandomForestClassifier, SVC, MLPClassifier, GaussianNB. Después entrene los modelos con mis datos de entrenamiento.

1. {'KNC': array([0.706 , 0.69933333, 0.71066667, 0.71933333, 0.709])}
2. {'RFC': array([0.77366667, 0.77266667, 0.76933333, 0.77533333, 0.767])}
3. {'SVC': array([0.75133333, 0.73966667, 0.74266667, 0.76066667, 0.75466667])}
4. {'MLP': array([0.745 , 0.74033333, 0.743 , 0.758 , 0.74333333])}
5. {'GNB': array([0.66133333, 0.65366667, 0.66366667, 0.66666667, 0.66966667])}

De los cinco entrenados el que mayor precision me devolvio fue **‘RFC’**, es decir, RandomForestClassifier.

Ya con mi modelo en mano cree su classification_report y matrix de confusión, los resultados fueron estos:

Reporte de clasificacion

	precision	recall	f1-score	support
0	0.82	0.69	0.75	2433
1	0.74	0.85	0.79	2567
accuracy			0.77	5000
macro avg	0.78	0.77	0.77	5000
weighted avg	0.78	0.77	0.77	5000

Matriz de confusion

[[1681 752]
[378 2189]]

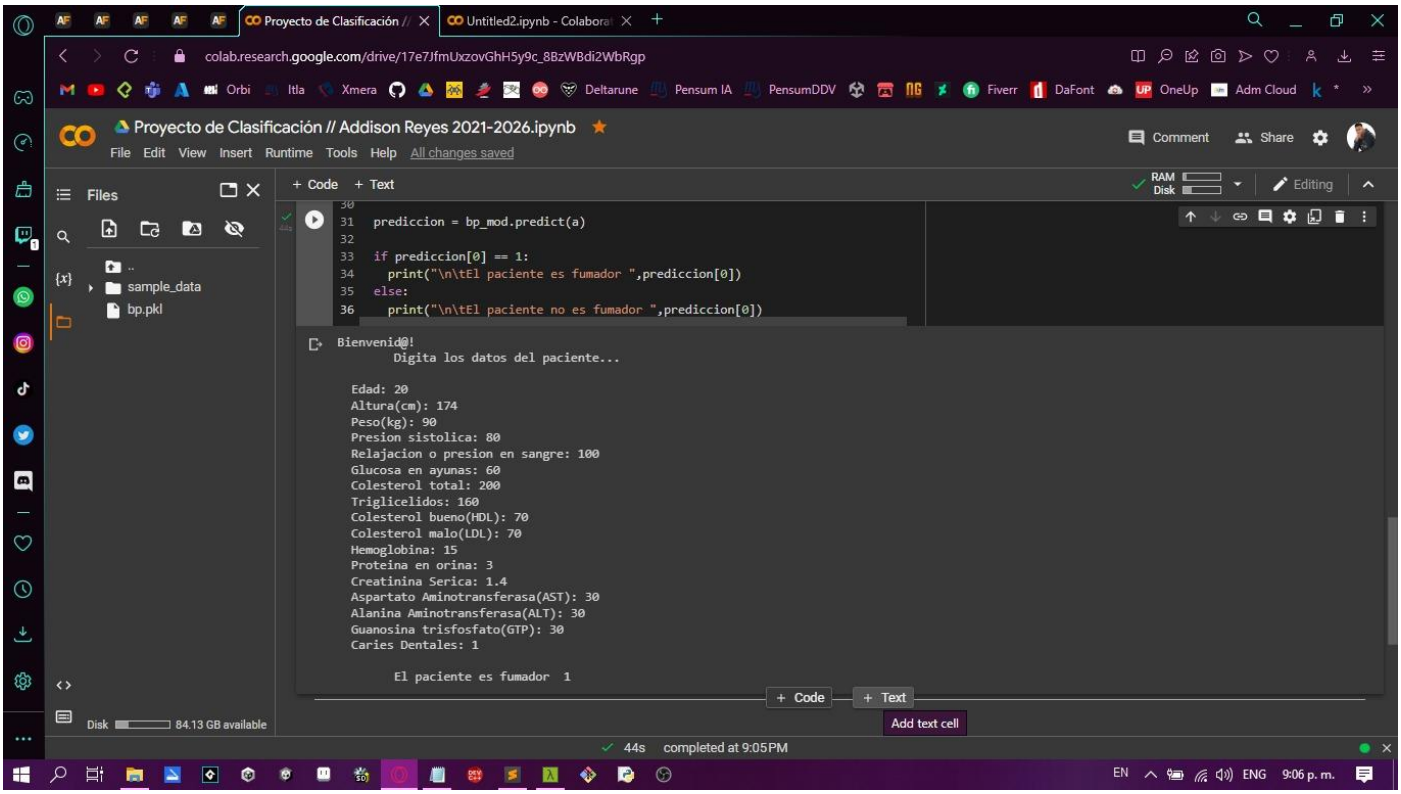
Una vez hecho el testeo y entrenamiento de la data, junto con el analisis de la matriz de confusion y el reporte de clasificacion exporte el modelo con pickle con el nombre de bp.pkl. Por ultimo cree un programa externo en donde se extraia el modelo con pickle y el cual te pide los datos de un paciente para convertir toda la informacion en un dataframe de pandas y predecir si el paciente clasifica como un fumador o no en base a todos los parametros que le entres, los parametros son los siguientes:

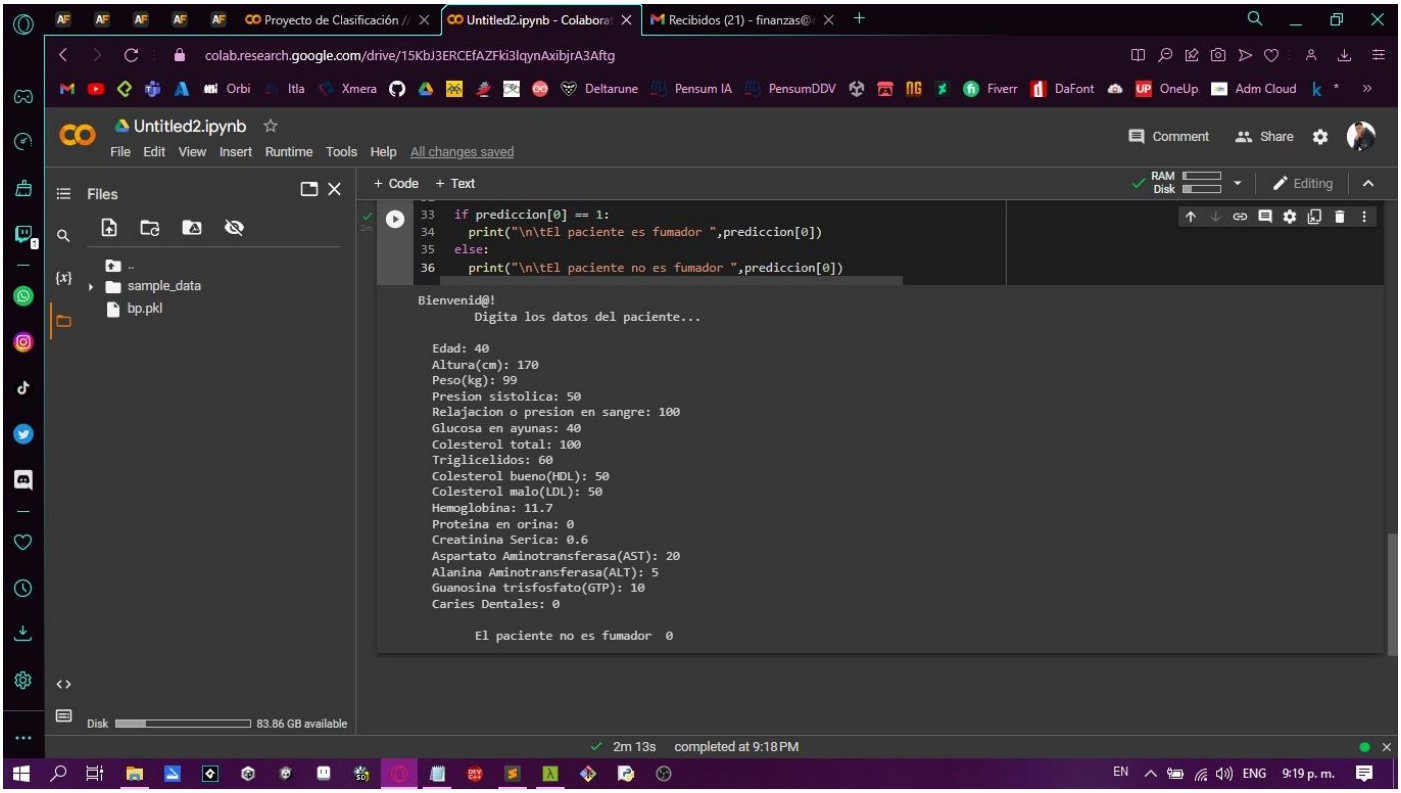
- **Edad:** Edad del pacient
- **Altura(cm):** Altura del paciente
- **Peso(kg):** Peso del paciente
- **Presion sistólica:** presión máxima que ejerce el corazón cuando late
  - Persona sana: <90
  - Fumador: 130>
- **Relajacion o presion en sangre:** La fuerza que la sangre ejerce contra las paredes arteriales, de 0 a 100
- **Glucosa en ayunas:** Medida de concentración de azúcar libre en sangre
  - Persona sana: <100
  - Fumador: 100>
- **Colesterol total:** Cantidad total de colesterol en la sangre
  - Persona sana: <200
  - Fumador: 200>
- **Triglicelidos:** Grasa que se encuentra en la sangre
  - Persona sana: <150
  - Fumador: 150>
- **Colesterol bueno(HDL):** El colesterol de lipoproteína de alta densidad.
  - Persona sana: <50
  - Fumador: 50>
- **Colesterol malo(LDL):** El colesterol de lipoproteína de baja densidad.
  - Persona sana: <70
  - Fumador: 70>
- **Hemoglobina:** Encargada de transportar el oxigeno a los órganos y tejidos.
  - Persona sana:
    - Hombres: 13.2 a 16.6
    - Mujeres: 11.6 a 15
  - Fumador: 15>
- **Proteina en orina:** Proteína liberada de los riñones involuntariamente, 1 a 5
- **Creatinina Serica:** Análisis que mide el nivel de la creatinina en la sangre, se hace para ver que tan bien funcionan los riñones.
  - Persona sana: 0.6 a 1.3
  - Fumador: 1.3>

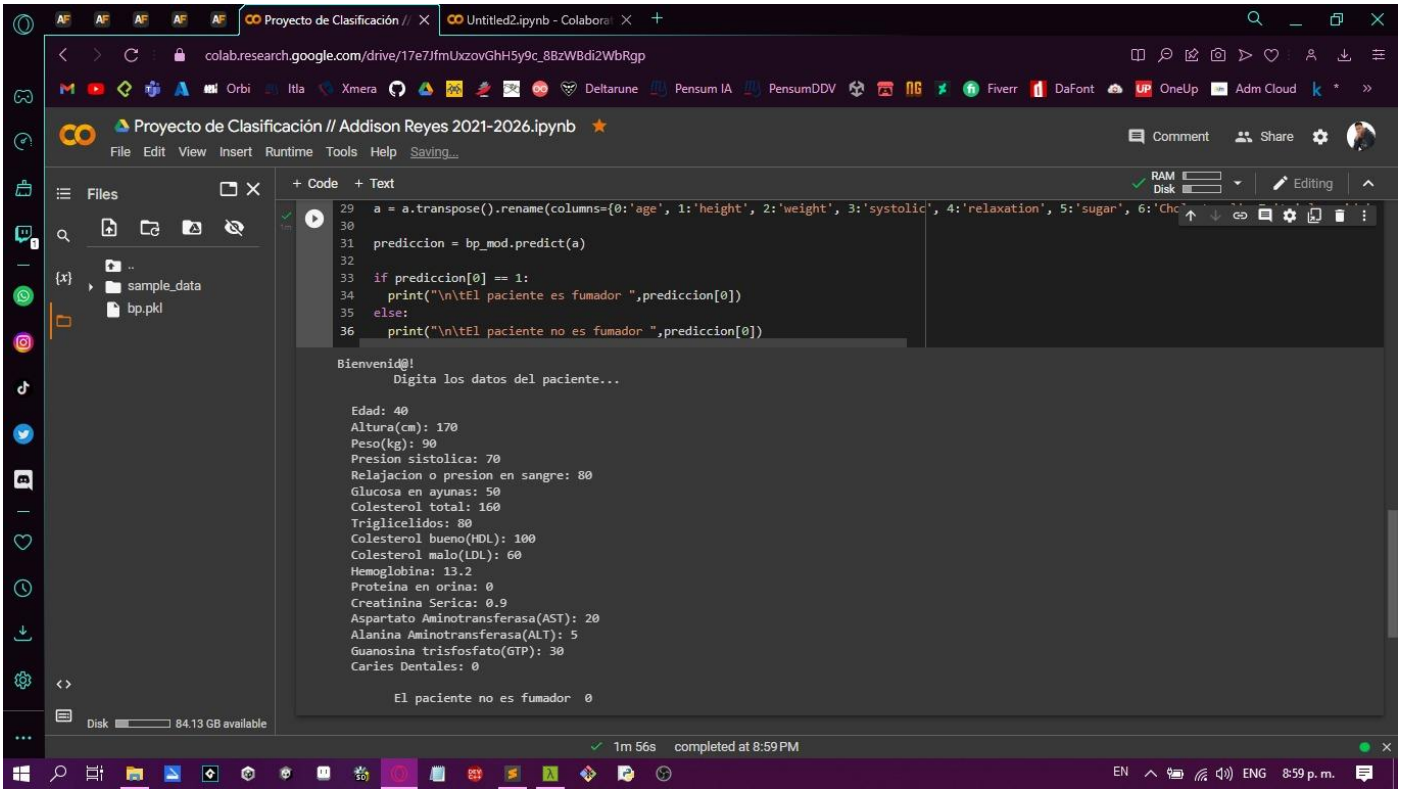
- **Aspartato Aminotransferasa(AST):** Enzima que se encuentra en el hígado y en los músculos y se libera en el torrente sanguíneo cuando hay células dananadas.
    - Persona sana: 8 a 33
    - Fumador: 33>
- **Alanina Aminotransferasa(ALT):** Enzima que se encuentra principalmente en el hígado.
    - Persona sana: 4 a 36
    - Fumador: 36>
- **Guanosina trisfosfato(GTP):** Es uno de los nucleótidos trifosfato usados en el metabolismo celular junto al ATP, CTP, TTP y UTP.
    - Persona sana: <40
    - Fumador: 40>
- **Caries Dentales:** La caries dental es el daño que le puede ocurrir a un diente cuando las bacterias que causan caries que se encuentran en la boca producen ácidos que atacan la superficie del diente o esmalte.
    - 1: Tiene caries.
    - 0: No tiene caries.

A partir de todos estos valores el programa devolvera **‘1’** si el paciente es fumador o **‘0’** si no lo es.

Capturas de ejemplo del programa externo:







Fuentes

https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction?select=train_dataset.csv

https://colab.research.google.com/drive/17e7JfmUxzovGhH5y9c_8BzWBdi2WbRgp?usp=sharing