Integrantes
Jaime Alexis Herrera Ruiz
Arley Fuentes Arenales

1. Problema predictivo

Según el Centro para el Control y la Prevención de Enfermedades (CDC), las enfermedades cardíacas son una de las principales causas de muerte en los EE.UU. Aproximadamente la mitad de todos los estadounidenses tienen al menos uno de los tres factores de más riesgo clave de enfermedad cardíaca: presión arterial alta, colesterol alto y el tabaquismo. Además, a estos factores se incluyen también, la diabetes, la obesidad, el sedentarismo o la ingesta de alcohol.
A partir de los datos obtenidos, se pretende construir un modelo, a través de métodos de aprendizaje automático que permitan detectar patrones que puedan predecir la posibilidad de una persona de padecer enfermedades cardíacas.

2. Dataset

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

El conjunto de datos proviene del CDC y hacen parte 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.
Consta de 319795 filas y 18 columnas. La gran mayoría de las columnas son preguntas que se hacen a los encuestados sobre su estado de salud:

- **HeartDisease:** encuestados que alguna vez informaron haber tenido una enfermedad cardíaca coronaria (CHD) o un infarto de miocardio (IM).
- **BMI:** Índice de Masa Corporal (IMC).
- **Smoking:** ¿Ha fumado al menos 100 cigarrillos en toda su vida? (La respuesta Sí o No).
- **AlcoholDrinking:** bebedores empedernidos (hombres adultos que toman más de 14 tragos por semana y mujeres adultas que toman más de 7 tragos por semana)
- **Stroke:** (Alguna vez le dijeron) (usted tuvo) un accidente cerebrovascular?
- **PhysicalHealth:** Pensando en su salud física ¿durante cuántos días durante los últimos 30 días su salud física no fue buena? (0-30 días).
- **MentalHealth:** Pensando en su salud mental, ¿durante cuántos días durante los últimos 30 días su salud mental no fue buena? (0-30 días).
- **DiffWalking:** ¿Tiene serias dificultades para caminar o subir escaleras?
- **Sex:** ¿Eres hombre o mujer?
- **AgeCategory:** Categoría de edad de catorce niveles.
- **Race:** Valor de raza/etnicidad
- **Diabetic:** (Alguna vez le dijeron) (usted tenía) diabetes?
- **PhysicalActivity:** Adultos que informaron haber realizado actividad física o ejercicio durante los últimos 30 días además de su trabajo habitual.
- **GenHealth:** ¿Diría usted que, en general, su salud es...
- **SleepTime:** en promedio, ¿cuántas horas duermes en un período de 24 horas?
- **Asthma:** (Alguna vez le dijeron) (usted tenía) asma?
- **KidneyDisease:** sin incluir cálculos renales, infección de la vejiga o incontinencia, ¿alguna vez le dijeron que tenía una enfermedad renal?
- **SkinCancer:** (Alguna vez le dijeron) (usted tenía) cáncer de piel?

3. Notebooks implementados

Se han realizado 4 notebooks así:

3.1. Modificación del dataset

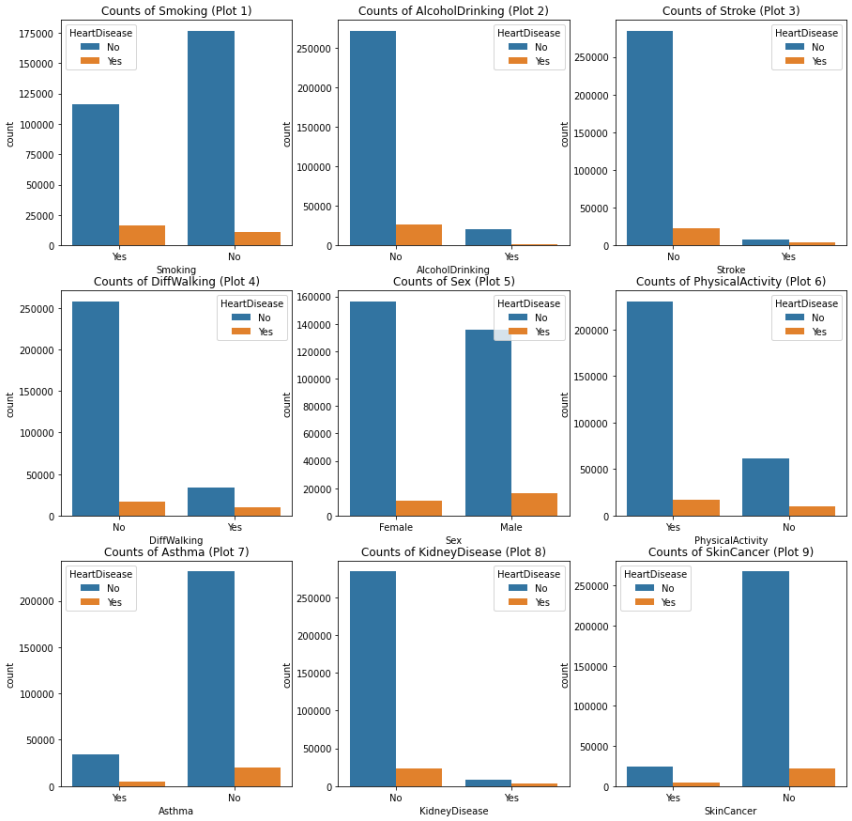
Este notebook se utiliza para eliminar aleatoriamente cierto porcentaje de datos (entre 5% y 10%) en 4 columnas escogidas también de manera aleatoria por código.

```
# Eliminación de datos

# Elección aleatoria de 4 columnas
def elegir_columnas():
    lista_cols = []
    df_cols = list(df.columns)
    df_cols.remove('HeartDisease')
    n = 0
    while n < 4:
        col_alea = random.choice(df_cols)
        lista_cols.append(col_alea)
        df_cols.remove(col_alea)
        n += 1
    return lista_cols

# Eliminación aleatoria de valores
num_datos = int(df.shape[0])-1
for col in elegir_columnas():
    n = 0
    # Porcentaje aleatorio
    porc_datos = int(num_datos*random.randint(5,10)/100)

    while n < porc_datos:
        # Elección aleatoria de la fila
        index = random.randint(0, num_datos)
        # Cambio del valor a vacío
        if df.loc[index, col] == df.loc[index, col]:
            df.loc[index, col] = np.NaN
            n += 1
```

3.2. Exploración de los datos

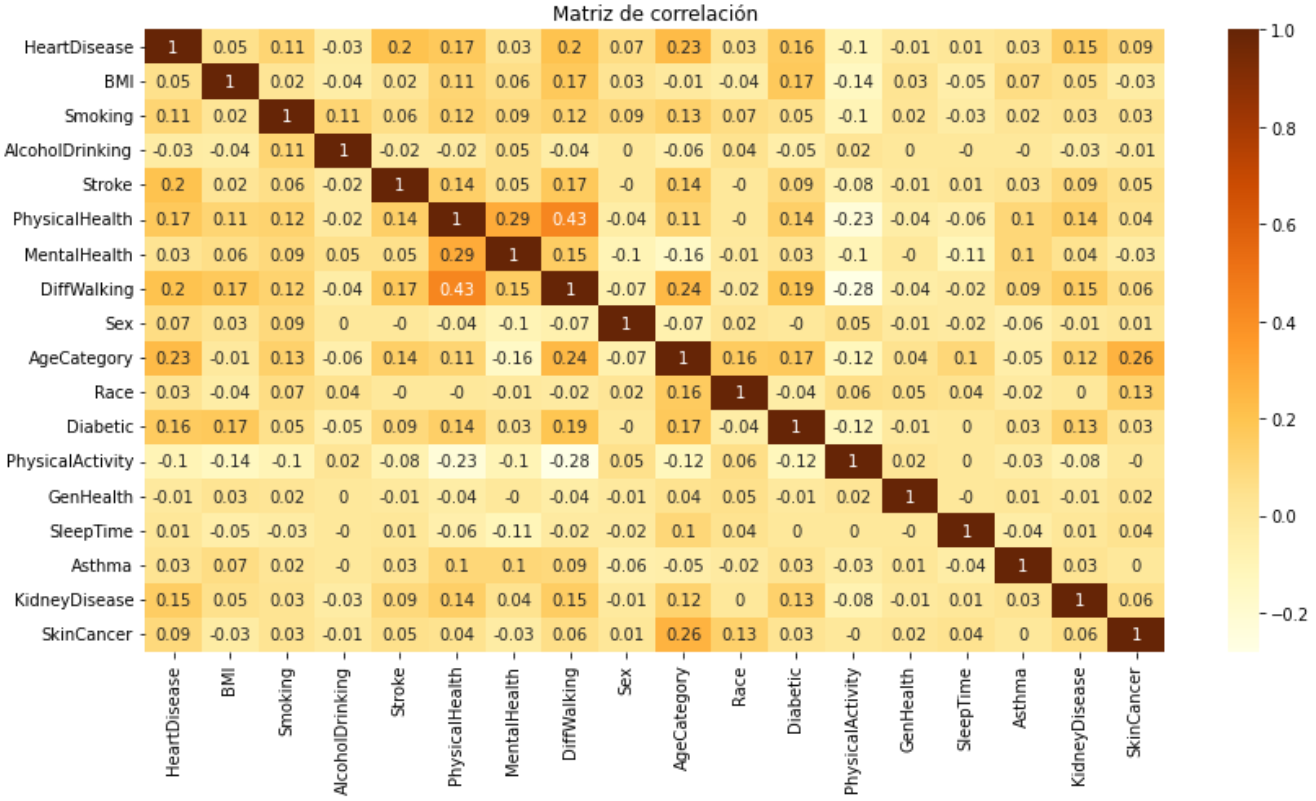
En este notebook mediante el uso de las librerías para graficar, se visualizan los datos para tratar de entender más fácil cómo se está comportando la información, en este caso todas las columnas se comparan con la variable respuesta que es la de enfermedades cardíacas (HeartDisease).

```
# Esta lista de listas contiene todas las columnas que tienen valores categóricos binarios
colRange = [['Smoking','AlcoholDrinking','Stroke'],['DiffWalking','Sex','PhysicalActivity'],['Asthma',
'KidneyDisease','SkinCancer']]
# Esta función imprime los gráficos de conteo contando el número de personas en cada categoría
def printCount(cols):
    fig, axes = plt.subplots(3, 3, figsize=(15, 15))
    row=0
    col=0
    p_count=1
    for row in range(3):
        for col in range(3):
            # reads column name from the list
            column = colRange[row][col]
            # plots the counts of the particular column
            sns.countplot(ax=axes[row,col],x=df[column],hue=df['HeartDisease'])
            # sets the title of the corresponding plot along with plot number
            axes[row,col].set_title("Counts of {} (Plot {})".format(column,p_count))
            p_count += 1
# Calling the function
printCount(colRange)
```



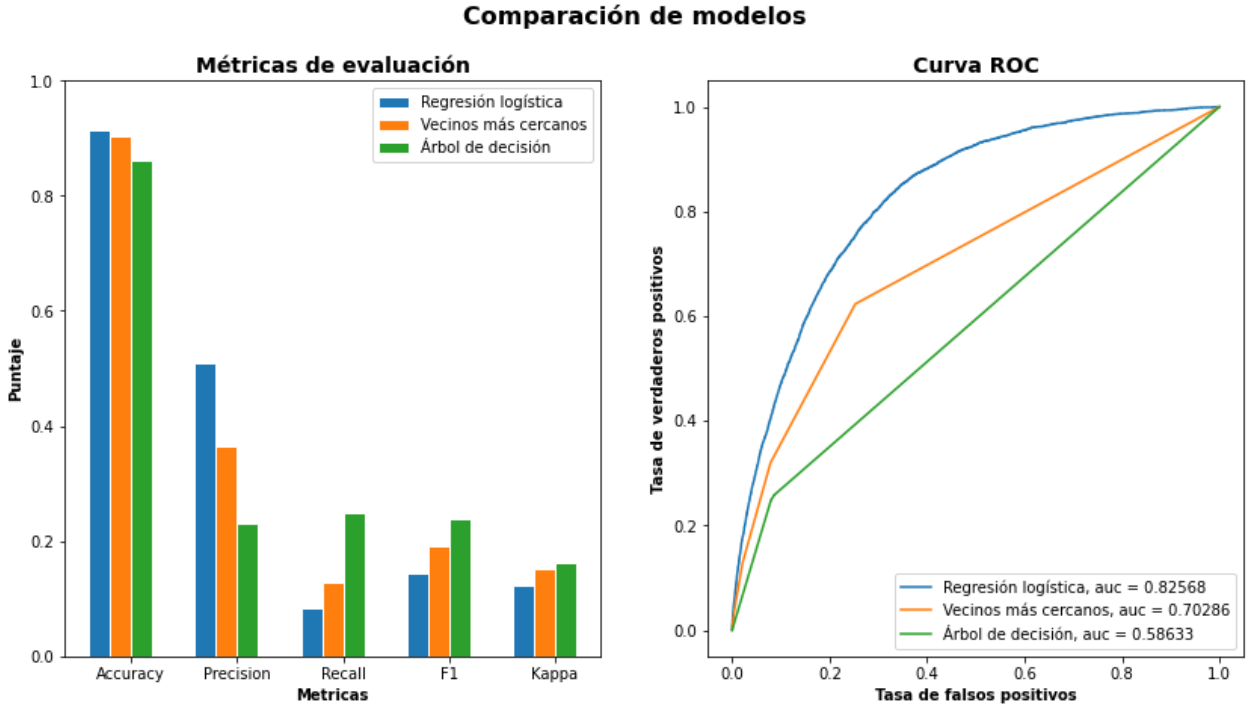
3.3. Limpieza de datos

En este notebook se desarrolla la limpieza de datos que consta de rellenar los datos faltantes utilizando algunos métodos como sacar la media o rellenar con datos aleatorios.
También se transforman los valores de texto a valores numéricos ya que los algoritmos de Machine Learning trabajan mejor de esta forma.



3.4. Creación de modelos

Ya habiendo hecho la limpieza de datos, procedimos a realizar algunos modelos de Machine Learning. Utilizamos: regresión logística, vecinos más cercanos y árbol de decisión.



4. Algunos problemas y dudas

- En la exploración de los datos hay que profundizar en el análisis de las gráficas haciendo más comparaciones y no exclusivamente con la columna HeartDisease.
- Hay dudas sobre el llenado de datos faltantes en columnas con datos categóricos ya que se debería tener en cuenta las categorías con más frecuencia.
- Hay verificar si la fusión de los datos de la columna Diabetic influye de manera negativa en los resultados de las predicciones que se realizarán.
- Hay que utilizar otros modelos como Redes Neuronales, Máquina de vectores de soporte, bosques aleatorios y Naive Bayes.

Fuentes

HeartDiseases. (2022, enero 28). https://www.cdc.gov/nchs/fastats/heart-disease.htm

Personal Key Indicators of Heart Disease. (s. f.). Recuperado 4 de julio de 2022, de https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Preprocesamiento de datasets con Python, Scikit-learn y Pandas (Parte #4)—Aprende con ejemplos. (2021, mayo 12). https://aprendeconejemplos.org/python/preprocesamiento-de-datasets-con-scikit-learn-y-pandas

¿Clasificación o Regresión? - IArtificial.net. (2018, diciembre 15). https://www.iartificial.net/clasificacion-o-regresion/

Prediction of Heart Disease (Easy). (s. f.). Recuperado 22 de agosto de 2022, de https://kaggle.com/code/arkalodh/prediction-of-heart-disease-easy

Heart Disease Prediction. (s. f.). Recuperado 22 de agosto de 2022, de https://kaggle.com/code/andls555/heart-disease-prediction