Daren Saenz
Tai Nguyen

Project Proposal: Predicting heart disease

1. Significance of the problem:
   In the US, a person dies from heart disease every 33 seconds. Heart disease remains the leading cause of death for both men and women across all racial and ethnic groups. In 2021, it accounted for 20% of total deaths, amounting to 695,000 people. Therefore, early diagnosis of heart disease is crucial to prevent fatalities. In this project, we will focus on detecting individuals at risk of heart disease, helping them become aware of their risk, and taking action to improve their health.

2. Expected outcomes:
   The expected outcomes from this project involve having patients detect and predict health problems earlier. This allows them to respond and take action in a timely manner. Another outcome would be lowering costs and improving the efficiency of resource allocation.

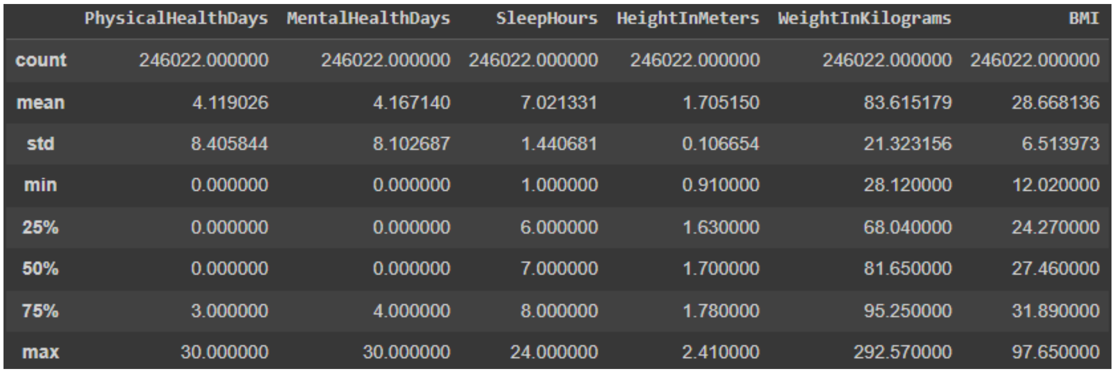
3. Variable descriptions:

   - Categorical Variables
     - RaceEthnicityCategory - Lists the patient's ethnicity.
     - Sex - List of patient are male or female.
     - GeneralHealth - Lists patient's general state of health ranging from "very good" to "very poor".
     - PhysicalActivities - States whether the patient has done physical activities.
     - HadAngina - States if patient has ever been diagnosed with Angina.
     - HadStroke - States if patient has ever had a stroke.
     - HadAsthma - States if patient has ever been diagnosed with Asthma.
     - HadCOPD - States if patient has ever been diagnosed with COPD.
     - HadDepressiveDisorder - States if patient has ever been diagnosed with any depressive disorder.
     - HadKidneyDisease - States if patient has ever been diagnosed with kidney disease.
     - HadDiabetes - States if patient has ever been diagnosed with diabetes.
     - DifficultyWalking - States if patient has ever had difficulty walking.
     - DifficultyDressingBathing - States if patient has ever had difficulty dressing and bathing.
     - SmokerStatus - States if patient is a smoker.
     - ECigaretteUsage - States if patient uses ECigarettes.
     - AlcoholDrinkers - States if patient drinks alcohol.
     - HighRiskLastYear - States if patient has been at high risk of heart disease last year.

    - CovidPos - States if patient has ever been positive for covid.

- Numerical Variables
    - AgeCategory - Lists the patient’s age range.
    - PhysicalHealthDays - Lists days when patient did physical therapy.
    - MentalHealthDays - Lists days when patient did mental health therapy.
    - Sleep Hours- Lists hours patients had sleep.
    - HeightInMeters - Lists patient’s height in meters.
    - WeightInKilograms - Lists patient’s weight in kilograms.
    - BMI: Body mass index (BMI) is a person's weight in kilograms divided by the square of height in meters.

Numerical Variables statistic



	PhysicalHealthDays	MentalHealthDays	SleepHours	HeightInMeters	WeightInKilograms	BMI
count	246022.000000	246022.000000	246022.000000	246022.000000	246022.000000	246022.000000
mean	4.119026	4.167140	7.021331	1.705150	83.615179	28.668136
std	8.405844	8.102687	1.440681	0.106654	21.323156	6.513973
min	0.000000	0.000000	1.000000	0.910000	28.120000	12.020000
25%	0.000000	0.000000	6.000000	1.630000	68.040000	24.270000
50%	0.000000	0.000000	7.000000	1.700000	81.650000	27.460000
75%	3.000000	4.000000	8.000000	1.780000	95.250000	31.890000
max	30.000000	30.000000	24.000000	2.410000	292.570000	97.650000

- Categorical variables (top 10):

    1. RaceEthnicityCategory - Lists the patient’s ethnicity.
    2. Sex - List of patient are male or female.
    3. GeneralHealth - Lists patient’s general state of health ranging from “very good” to “very poor”.
    4. PhysicalActivities - States whether the patient has done physical activities.
    5. HadAngina - States if patient has ever been diagnosed with Angina.
    6. HadStroke - States if patient has ever had a stroke.
    7. HadDiabetes - States if patient has ever been diagnosed with diabetes.
    8. SmokerStatus - States if patient is a smoker.
    9. AlcoholDrinkers - States if patient drinks alcohol.
    10. HighRiskLastYear - States if patient has been at high risk of heart disease last year.

4. Method of analysis
   Due to the data structure, the people without heart disease outnumber the people without heart disease. Also, the dependent variable is categorical (Yes/No). The method used in

this project is supervised learning Logistic Regression. Using this method, we will be able to know the weight of each variable through the coefficients From that, we can predict the percentage of heart disease for each datapoint.
Here is the detail on the dependent and independent variables:

- Dependant variable: HadHeartAttact
- Independent variable: After filtering many unnecessary variables, we narrow down some independent variables:
    - RaceEthnicityCategory
    - Sex
    - AgeCategory
    - BIM
    - GeneralHealth
    - PhysicalHealthDays
    - MentalHealthDays
    - PhysicalActivities
    - HadAngina
    - HadStroke
    - HadAsthma
    - HadCOPD
    - HadDepressiveDisorder
    - HadKidneyDisease
    - HadDiabetes
    - DifficultyWalking
    - DifficultyDressingBathing
    - SmokerStatus
    - ECigaretteUsage
    - AlcoholDrinkers
    - HighRiskLastYear
    - CovidPos

Citation

"Heart Disease Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 May 2023, www.cdc.gov/heartdisease/facts.htm.