**Project Ideas**

**Three Datasets of Interest**

Dataset 1:

**Stroke Prediction Dataset**

Source: Kaggle

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Description:

According to WHO, stroke is the leading cause of disability worldwide and the second leading cause of death. The Global Stroke Factsheet released in 2022 reveals that the lifetime risk of developing a stroke has increased by 50% over the last 17 years and now 1 in 4 people is estimated to have a stroke in their lifetime.

This dataset can be used to predict which gender is more likely to get a stroke and find out which of the parameters (gender, age, various diseases, smoking status) will highly predict having a stroke.

TITLE: “Which gender is more likely to get a stroke? Which parameters will highly predict having a stroke?”

Dataset 2:

**Diabetes Health Indicators Dataset**

Source: Created from BRFSS 2015 dataset already in Kaggle

<https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Description:

Diabetes is among the most prevalent chronic diseases in the United States, impacting millions of Americans each year and exerting a significant financial burden on the economy. While there is no cure for diabetes, strategies like losing weight, eating healthily, being active, and receiving medical treatments can mitigate the harms of this disease in many patients. Early diagnosis can lead to lifestyle changes and more effective treatment, making predictive models for diabetes risk important tools for public and public health officials.

This dataset has 3 files. I am interested in using this file.

The diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.

This dataset can be used to predict the top 5 risk factors that will predict diabetic risk.

TITLE: “Which of the risk factors is the most predictive of diabetic risk?”

Dataset 3:

**Heart Attack**

Source: Kaggle

<https://www.kaggle.com/datasets/pritsheta/heart-attack>

Description:

In the United States, someone has a heart attack every 40 seconds. Every year, about 805,000 people in the United States have a heart attack. Of these, 605,000 are first heart attacks.

This dataset can be used to predict which attributes will more likely be the best predictor for heart attack and whether men are more prone to heart attacks than women.

TITLE: “Which of the parameters will be the best predictor for heart attack? Are men more prone than women to heart attacks?”