Project Proposal

The dataset that you plan to use. Please include its source, a list of the features and categories included in the dataset, and how many records are in the dataset.

Data I’m using: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_health_indicators_BRFSS2015.csv>

Is taken from below data, but specific for diabetes

original data: <https://www.cdc.gov/brfss/annual_data/annual_2015.html>

description:

What you plan to do with the dataset (e.g., predict the category x) and how you plan to do it (e.g., use four supervised learning techniques including the following …).

1. Binary diabetes classification (yes or no)
2. Multiclass diabetes classification (no, pre, yes)
3. Exploratory analysis (remove na’s/duplicates)
4. Feature selection based on tests (chi squared)
5. Compare random forest, gradient boosted trees
6. Use 10 fold cross val, tune hyper params