**DATA WRANGLING**

**Summary:**

The purpose of this notebook is to clean and prepare the BRFSS data into a useable format to predict the accuracy if a person might have diabetes or is at risk of diabetes by finding out what risk factors are the most predictive of diabetic risk.

The following tasks were performed. Starting with the loading of the chosen dataset. A file structure was created, pushed to GitHub and the glob library was downloaded.

Reviewed the BRFSS2015 codebook to understand the dataset more in depth. Check the column names, shape, info, summary statistics for each of the columns and unique values of the dataset. Started cleaning the data for missing values or NaN values and duplicates. The data did not have any missing values, NaN's or any duplicates.

Looked into the value counts for each feature to understand the dataset and made a histogram for the data to see the distribution of features and see any useful values that may need to be investigated further. The value counts for each of the features gave a picture on how many people were diabetic or not, how many male or female, how many had high BP or high cholesterol and so on which may be of use in later analysis.

Also look at the mean, median and mode of the data to see if it showed any use on categorical and non-categorical features of the data. The mode of the data is of interest in that most of the features will likely help in analyzing which features would be a good predictor for diabetic risk. Lastly, group the data by diabetes\_binary to further see if the features will be of use in predicting risk factors for diabetes. In here, the data showed some association with the different risk predictors for diabetes, but not conclusive.

There is further need for investigation and data analysis to reach a more concrete conclusion.

**EDA ANALYSIS**

**SUMMARY**

We revisited the hypothesis we want to find the answers to. Data Wrangling was performed. We start our EDA analysis by assessing the quality of the data. We do this to understand what information is in the data and make sure what the data stands for and what is missing.

We concluded from the data wrangling that we have no missing data and have a tidy data.

We assess the datasets statistical summary and followed by creating a visualization of our attributes or features on histogram. We can see that the majority of the data are binary and are categorical data.

We generated a Pandas profiling report which shows us not much since our data are mostly categorical and would need a different method to represent the data relationships.

Other visualization like the boxplot and heatmap were created and only more questions come up to see if the variables would show some more interaction. This leads to the need for further exploration.

Since the question to answer is finding the best predictors for diabetes, we checked the Diabetes\_binary which would be the possible response variable to use in predicting the risk factors. This variable is explored from histogram to summary statistics and counts.

Correlation of binary and categorical data are not valid and performing a correlation method would not provide us with any value or clues.

Next step is to use statistical methods to show statistical relationship strengths.

Decided to create histograms for the data segregated by the Diabetes\_binary. It showed the interaction of diabetics and non-diabetics with the different attributes in the dataset. We simplified the categories within the single data field to <10 categories as shown in the histogram charts.

The BMI data undergone min/max scaling to make it consistent with the other categorical data and performed one hot encoding for categorical data.

Hypothesis testing was performed to decide if there is a difference between diabetics and non-diabetics. We used the two-sided test, z-test and p-value calculation and found that most of the attributes are significantly different between the diabetics and non-diabetics which would need more investigation.

We also performed z-test and p-value calculation for the BMI attribute (quantitative variable) and results showed that the non-diabetics have lower BMI values than the diabetics.

We have tidied the data and explored the relationship between our dependent/response variable and independent/explanatory variables. It seems that almost all the features would still likely have some impact in the modeling based on the relationships between the features and the response variable.

**MODELING**

**SUMMARY**