**Diabetes Monitoring System Capstone Project**

**Technical Progress Report**

**Introduction**

The goal of this capstone project is to build a diabetes monitoring system that assesses the status of diabetes based on user-input parameters and subsequently generates an electronic health record (EHR) through a conversational interface.

This part of the project involves data collection, analysis, feature engineering, and the development of algorithms to determine diabetes status.

**Data collection and EDA**

Source ([link](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)):

We used the “***diabetes\_012\_health\_indicators\_BRFSS2015.csv”*** dataset available in Kaggle, to train the classifier model. This is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes.

Exploratory data analysis:

A screenshot of a computer screen

Description automatically generated

Figure 1: EDA result

This dataset has 21 features that are used to determine the diabetes status of a person. From the above analysis, there are 253680 observations and none of the features has any missing value. This analysis also provides insight on number of unique values, maximum value, most frequent value and so on.

We built a custom package to perform a preliminary statistical analysis of the raw data. Following images show the code:

A screen shot of a computer program

Description automatically generatedA screen shot of a computer program

Description automatically generated

1. (b)

A screen shot of a computer program

Description automatically generated

(c)

Figure 2: (a), (b), (c) Custom class for analysis

Next step is to perform visualization. We used plotly library for this purpose.

A screen shot of a computer

Description automatically generated

Figure 3: Distribution of our target variable

Here we have a tabular as well as visual representation of how different cases of diabetes are distributed across the dataset. There is severe class imbalance, and this is one of the most important things to take care of before training any classifier model.

After target variable, we also analysis a few other variables as follows:

A screen shot of a computer

Description automatically generated

Figure 4: BMI analysis

The above histogram shows distribution of BMI. We know from our analysis that there are 84 unique values.

We also visualize correlations between our Diabetes status indicator and all the 21 features. More details can be seen on hovering over the plot. We have a correlation heatmap for all the variables (including target). We even generated a bar plot that shows correlation of the target variable with all the features.

A screenshot of a computer

Description automatically generated

Figure 5: Correlation heatmap for all variables (including target variable)

A screenshot of a graph

Description automatically generatedA screen shot of a computer program

Description automatically generated

(b)

(a)

Figure 6: (a) Code snippet for bar plots; (b) Bar plot showing correlation between target variable and features