**CS 677 Project Description**

My project is about diabetes detection using machine learning. I have used the dataset from Kaggle which is a CDC telephonic health survey dataset with features that are considered as important risk factors for diabetes. As the dataset is from a survey, most of the features have values Yes or No reflected by 1 and 0 in the dataset. The dataset has 21 features and based on these features we will train our supervised learning classifiers and test them to calculate the accuracy of each of these classifiers.

The data is contained in one csv file named: “Diabetes CDC BRFSS 2015”. The python script to be run is named: “Muhammad Osama U41261655 CS 677 Semester Project.py”. The csv file and the python script should be in the same folder for the script to run successfully. The graphs that I have used in the presentation will be plotted by the python script and saved as well. I will also be including these graphs with the python script. The program will take like 3-4 minutes to run depending on the machine being used.

Below are the graphs plotted by different supervised learning classifiers that I have ran on my dataset:

Chart, line chart

Description automatically generated

Decision Tree Classifier Accuracy for different Depths. Max Accuracy is 73.04% for depth 7.

Chart, line chart

Description automatically generated

Adaboost classifier with Base Estimator = Logistic Regression. Max Accuracy is 72.92% for learning rate 1 and number of estimators 11.

Chart, line chart

Description automatically generated

Adaboost classifier with Base Estimator = Naïve Bayesian. Max Accuracy is 71.90% for learning rate 1 and number of estimators 7.

Chart, line chart

Description automatically generated

Adaboost classifier with Base Estimator = Random Forest. Max Accuracy is 75.54% for learning rate 0.5 and number of estimators 13.

Chart

Description automatically generated

K Mean Clustering applied on the dataset to see if we could achieve pure clusters meaning clusters with only diabetes class or non-diabetes class. But we can see that this is not feasible as the inertia is very high and to reduce it we need to have a very large number of clusters which would not be of any use.

Chart, line chart

Description automatically generated

K Nearest Neighbor for P = 1.5. Highest accuracy is 71.48% for 11 neighbors.

Chart, line chart

Description automatically generated

K Nearest Neighbor for P = 2. Highest accuracy is 71.34% for 11 neighbors.

Chart, line chart

Description automatically generated

K Nearest Neighbor for P = 1. Highest accuracy is 71.34% for 11 neighbors.

Chart, line chart

Description automatically generated

K Nearest Neighbor for P = 3. Highest accuracy is 71.22% for 11 neighbors.

Diagram

Description automatically generated

Pair plot for the 21 features in the dataset. This is included in separately with the graphs as well so that we can zoom in and see the patterns.

Chart, line chart

Description automatically generated

Accuracy for Random Forest Classifier with different number of estimators and different depths. Max accuracy is 74.58% for 15 estimators and depth: 7.