**HDS 5960: Capstone Progress Report on**

**Diabetes Prediction Model with Health Indicators**

**Ajith**

**Dataset :** Diabetes Health Indicators Dataset

**Demographic Information:** Survey conducted in America by Behavioral Risk Factor Surveillance System (BRFSS)

**Year of Survey** : 2015

**People participated** : 441,455 individuals

**Questions answered are known as Features** : 330

**Reason for Selecting this datasets:** We have found many datasets on diabetes but mostly are has very limited data compiled with less number of column (features i.e. BMI, Blood pressure). We want to explore more with high number of people participated in studies, due to that we found this datasets which has so many individuals who participated and they also have divided the dataset into three category.

**Answer 1:**

In this analysis part, We have done few more analysis on Bi-viriate Analysis and perform the Chi-Square test on the individual features to understand if it has p-value more than significance to understand the Hypothesis result to consider Null Hypothesis or Alternate Hypothesis. After that we have perform the Machine learning classifiers to understand predict data and get accuracy of these five classifiers i.e. Logistic Regression, Decision Tree, Gaussian Naïve Bayes classifier, Random Forest Classifier, XGB boost classifier. In the last, we have published the evaluation anaylsis part.

**Answer 2:**

I have perform all the analysis which I think, requires for understand the diabetes health markers on dataset and now going to optimize it , if I can.

* Checking my results and understanding more about data.
* Check If any other classifier works well.

**Answer 3:**

Yes, Almost Project is finished, need to just tune it a little bit to write my final conclusion as result.

**Summary of Activity :**

* Done Bi-variate analysis on health-markers
* Perfom Chi-Square Test for understanding the health-markers to know if they are correlated with Diabetes.
* Choose five classifier which is listed above and evaluate the accuracy of each

**Plan for Next Steps:**

* I will check few more machine learning kernel, which can works on predicting the target for Diabetes.
* Checks if I have missed something, while performing the analysis to increase the accuracy of the classifiers.
* Need to validate the similar results from other datasets (df2, df3). Currently performed analysis on the df1 (balanced dataset only)

**Evaluation of Proposal TimeLine:**

* *Loading datasets, Cleaning datasets (Completed)*
* *Analysis and Study the risk factors of the dataset. (Completed)*
* Built the model and perform prediction analysis with accuracy. (Completed)
* Change the way dataset is used for prediction to illustrates the results segment-wise i.e. age-wise, income-wise, location-wise. (Next Phase to check the results)

**References:**

1. Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.
2. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Emerging Risk Factors Collaboration