**Stat Assignment 5**

Data processing

Remove the data with missing values >50

Normalize the Data

Univariate distribution should be done

Compare two sets of quantities using one another (Do scatter plot type of thing to compare the data) QQ plots or any other plots similar to it.

* Compare the levels of Hba1c in which race and age group it is high in range
* compare the effect of chronic kidney diseases and diabetes
* compare different age groups 45 -60 and 60 – 70, 70 -80 above HbA1c levels
* See the rate of readmission
* **(Readmitted** Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission)
* **Which race people readmitted more**

**Clustering:**

Run Hierarchial clustering method using cosine distance and compared the results with K-Means for both Race and Readmitted attributes.

Silhouette score for hierarchial clustering based on race & readmittance

Implement Models: Logistic Regression and Decision Trees

Do principal component analysis if possible.

**Code and report both are needed separately Write materials and methods along with results and discussion. (with graphs and images give the explanation) No Plagiarism in code and report.**

**Deadline: 12/9/2020 9pm Eastern standard Time**

I have attached the data in this zip file, if there is any problem with that file you can download the data from here. <https://www.kaggle.com/brandao/diabetes>

Name of the dataset: [Diabetes 130 US hospitals for years 1999-2008](https://www.kaggle.com/brandao/diabetes)