**DSCI 5260 - Project Proposal**

**Title:** Diabetes Prediction Model by utilizing the Health Indicator diabetes datasets

**Team Members:**

Mayank

ZSDA

ZASD

Zsad

**Problem statement:** We want to develop a predictive model, which can prompt certain well-defined questions from users and able to predict the as the person is diabetes or non-diabetes. These questions relevance will be considered from Health Indicator diabetes datasets, which has sufficient number of training samples to validate our model.

**Introduction:**

Diabetes is a chronic disease, which is becoming common and found in every third person near you. According to WHO, the Centres for Disease Control and Prevention (CDC) has indicated that as of 2018, 34.2 million Americans have diabetes and 88 million have prediabetes. Furthermore, the CDC estimates that 1 in 5 diabetics, and roughly 8 in 10 prediabetics are unaware of their risk.

There are three types of diabetes which is found across people.

* **Type I diabetes**: found in children due to deficient insulin production and requires daily administration of insulin
* **Type II diabetes**: found in more than 95% of people with diabetes have type 2 diabetes. It found in people with excess body weight and physical inactivity
* **Gestational diabetes**: It occurs during pregnancy. It has increased risk of complications during pregnancy and at delivery

In this project, we have built a predictive model which can help us identify the root cause on determine the factors which leads them to becoming diabetes. While exploring the diabetes dataset, we have found out only two diabetes dataset.

1. **PIMA Indians diabetes dataset:** It was collected from hospital and real patients dataset, which has 8 independent variables, 1 dependent variable, 768 patients sample only
2. **Health Indicator diabetes dataset:** It was the survey generated dataset by Central of Disease Control (CDC). It was collected by 441,455 individuals in 2015 by asking questions and generated 330 features. It was released as 3 dataset files:
   1. **diabetes\_012\_health\_indicators\_BRFSS2015.csv:** It was created by 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. This dataset has 21 feature variables and is imbalanced.
   2. **diabetes\_binary\_5050split\_health\_indicators\_BRFSS2015.csv:** is was created by 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.
   3. **diabetes\_binary\_health\_indicators\_BRFSS2015.csv:** is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.

For this project, we have selected to work with “**Diabetes Health Indicator dataset”** for the following reasons:

1. It has many independent variable (20 feature) and 1 dependent variable ( diabetes class)
2. It has many number of samples in each kind of dataset files (253680, 70692, 253680). It is good for predictive model to have sufficient number of samples for training and testing
3. It has the the features i.e. Diabetes\_binary, HighBP, HighChol, CholCheck, BMI, Smokes, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education and Income

**Project Goals and Tools:**

* **Language** : Python
* **Module**: pandas, numpy, scikit-learn, matplotlib, seaborn
* **Editor**: python-notebook, Google-colab

**Research Methods :**

1. Perform statistics methods i.e. null hypothesis/ alternate hypothesis by applying methods like Chi-Square test to prove the feature relevance
2. Learn each classifier and their way of segregation by hyperplane into different classes

**Research Question :**

**Question1:** Which features are more correlated towards Diabetes?

**Answer:** Understanding of features from dataset, we need to perform reading, cleaning, and exploring of the dataset by doing EDA (Exploratory Data Analysis)

**Question2:** Which feature variables has more importance on Diabetes while performing predictive analysis?

**Answer:** We need to read the dataset, clean the dataset and evaluate the feature importance by applying the analysis and plotting in relevant to Diabetes (Feature Importance)

**Question3:** Which predictive model is the best for your prediction analysis, in terms of time and accuracy?

**Answer:** Understanding the predictive analysis, we need to apply multiple classifier and see how eac classifier is able to segregate the datasets in hyperplane correctly to build the classifier. The better the classification by hyperplane, impact the accuracy of the model. While selecting the best predictive model, we need to observe the converge time of the model in term of training. These analysis is requires to come to conclusion of identifying the best predictive model.

**Question4:** What is the usages of this predictive model?

**Answer:**  By building this predictive model, we can prompt the user for survey based question and predict a person is diabetes or not? It will be final result of this analysis and predictive model building (Build a GUI form or text based inputs question)

**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.