**Diabetes prediction model by using Machine Learning techniques**

**Team Members**

1. Prudvi Siddanathi – rmz769

2. Aanathi Mosali – bat761

3. Bhanu Anish Akkineni – umb160

**Problem Statement**

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 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.

We want to develop a predictive Machine Learning (ML) 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 Diabetes Health Indicator datasets, which has sufficient number of training samples to validate our model.

**Methods**

* ​Logistic Regression
* Gaussian Naïve Bayes
* ​Decision Tree Classifier
* ​Random Forest Classifier
* XGBoost Classifier

**Dataset:** [Diabetes Health Indicator datasets](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset) (from Kaggle)

The dataset 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.

**Dataset Analysis and Outcomes:**

1. Does Age, Education and Sex have more impact on Diabetes?
2. Which features has more importance on Diabetes while performing predictive analysis?
3. Which predictive model is the best for your prediction analysis, in terms of time and accuracy?
4. What is the usages of the developed predictive Model?

**Project Goals and Tools:**

**Language** : Python

**Module**: Pandas, NumPy, Scikit-Learn, Matplotlib, Seaborn

**Editor**: Python-Notebook, Google-Colab