**Project Title**

**Diabetes Prediction Model with Health Indicators**

**Team Members**

1. ABCDED
2. ASDASDAS
3. Sadasdasdsa
4. Sadsadsa
5. SADASDASd

**Idea Description:**

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.

**Goals and Objectives:**

In this project, we want to identify important health indicators which can predict the a person is diabetics or not. We will do the cleaning, balancing of the datasets and build the Machine Learning (ML )model which will be able to predict and do the report of the model classification and try to get most accuracy of the model with given datasets.

**Motivation:**

**Significance:**

**Literature Survey:**

**Objectives:**

**Features:**

**Expected outcome :**

**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.

Basically there are three types of diabetes which is found across people.

* **Type I diabetes** : it is found in children who have deficient insulin production and requires daily administration of insulin.
* **Type II diabetes** : It is the most common form and it’s prevalence varies by age, education, income, location, race and other social determinants of health. More than 95% of people with diabetes have type 2 diabetes. This type of diabetes is largely the result of excess body weight and physical inactivity.
* **Gestational diabetes** : Gestational diabetes occurs during pregnancy. Women with gestational diabetes are at an increased risk of complications during pregnancy and at delivery

In this project, we have built a model which can help us identify the root cause on determining the factors which leads them to becoming diabetes.

**Dataset :**

For this project, we have chosen the [dataset](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset) given on Kaggle “Diabetes Health Indicators Dataset”, which had gathered the data from the user on health-related survey named “The Behavioral Risk Factor Surveillance System” (BRFSS) that is collected annually by the “Central of Disease Control” (CDC). Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services. We have select the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.

This dataset contains 3 files:

1. diabetes\_012\_health\_indicators\_BRFSS2015.csv: is a clean dataset of 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 a clean dataset of 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.

1. 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.

We have selected this dataset, as each survey file is not balanced, that could be provides us the real world scenario, as diabetes varies race and location-wise differently. We were thinking of choosing around 22 variables from the datasets which has features which can help us built the model for the projects.

* Diabetes\_binary
* HighBP
* HighChol
* CholCheck
* BMI
* Smokes
* Stroke
* HeartDiseaseorAttack
* PhysActivity
* Fruits
* Veggies
* HvyAlcoholConsump
* AnyHealthcare
* NoDocbcCost
* GenHlth
* MentHlth
* PhysHlth
* DiffWalk
* Sex
* Age
* Education
* Income

We have already identified the all the three sample files to provide us the real world scenario of samples.

**Analysis Plan :**

1. Understanding of features from analysis and plotting i.e. correlation, or dividing the datasets by age wise to selecting outcomes.
2. Selecting of the risk factors from datasets which could real impact us model outcomes for prediction of early diabetes
3. Machine kernel which identify and works well according the data plane.
4. If there are many factors changing the model accuracy, then study and implement the deep-leaning model building

**Project Timeline:**

**Step 1**: Perform Analysis and Study the risk factors of the dataset.

**Step 2:** Built the model and perform prediction analysis with accuracy.

**Step 3:** Change the way dataset is used for prediction to illustrates the results segment-wise i.e. age-wise, income-wise, location-wise

**References:**

**HDS 5960: Capstone Proposal (1-2 pages)**

**Please submit this capstone proposal prior to starting your project. Once you receive approval, you may begin your project.**

**Please provide sufficient detail for complete understanding of your project. This should include the following information:**

1. Name of organization and preceptor
2. Brief organization background
3. Introduction/Background of Topic
4. Data Source (name, years of data, type of data)
5. Variables and operationalization (outcome, exposure, control, confounding)
6. Sample and inclusion criteria
7. IRB approval if necessary
8. Analysis plan (statistics you will run)
9. Timeline for project completion

Proposals should be well-organized, thoughtfully written, and generally free of grammatical and spelling errors. Any literature, websites, or other resources should be properly referenced using APA or another standard reference format. Documents should be submitted as a PDF.