**Project Name:** Diabetes Prediction

**Github Link:** https://github.com/projectsforstudents2022/Diabetes\_Prediction.git

**Why was this project created?**

A vital and important task for healthcare workers is medical diagnosis. Diabetes classification is particularly difficult. The key to managing diabetes is early detection of the disease. A patient must undergo a number of tests, and after that it is quite problematic for specialists to keep track of several elements during the diagnosis process, which can result in false results and make detection very difficult. Machine learning algorithms in particular are highly helpful for the quick and accurate prediction of sickness in the healthcare industries thanks to the most advanced technologies.

**What problem is it solving?**

The major goal is to determine a patient's likelihood of having diabetes using the diagnostic parameters gathered in the database. All patients are girls aged 21 and older who are of Pima Indian ancestry.

**Entire explanation of project**

* **PROPOSED APPROACH**

Dataset will be loaded into a pandas dataframe named df as our first step after importing Pandas and NumPy into our Python environment. Eight feature variables and one target variable make up the dataset's 768 observables. Understanding the existence of missing values in any dataset is crucial before beginning the data analysis and coming to any conclusions. The output indicates that there are no null values. However, there are five characteristics, including zero values for glucose, blood pressure, skin thickness, insulin, and BMI, which is impossible given the patient's medical history. These values will be regarded as missing values. The zero values will be changed to NaN, and their mean value will be imputed.

We divided the target (y) and feature (X) dataframes into training (80%) and testing (20%) sets. The training set is used to create the classification model, and the testing set is used to assess the model's effectiveness. We'll employ a Random Forest Classifier machine learning basic model. Using the training dataset, we train the model with the default settings. Using a test dataset, we assess the performance of our model. We split the training (80%) and testing (20%) sets of the target (y) and feature (X) dataframes. The classification model is developed using the training set, and its performance is evaluated using the testing set. We'll use a fundamental machine learning model called the Random Forest Classifier. We train the model using the training dataset and the default parameters. We evaluate our model's performance using a test dataset.

Algorithm for creating next word prediction model :

**Step 1:** Dataset is imported

**Step 2:** The data is preprocessed, and the dataset is split into training and testing.

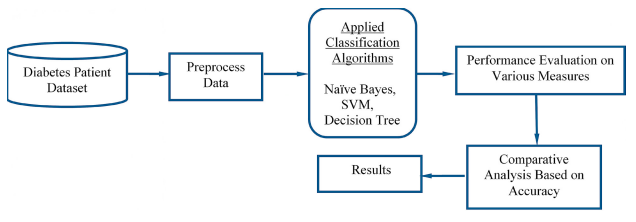
**Step 3:** Feature Extraction

**Step 4:** Predict Best Model

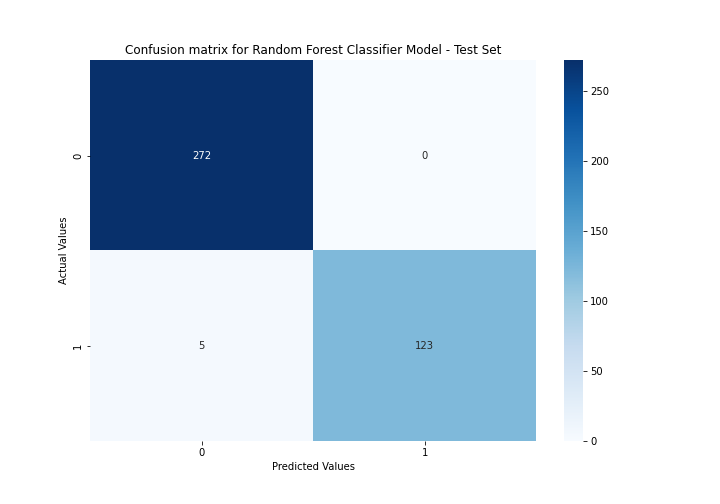
**Step 5:** Create Random Forest Model and Train the model

**Step 6:** Calculate Accuracy

* **DATA FLOW DIAGRAM**



* **RESULT**



* **CONCLUSION**

Using data from the database, we developed a classifier based on machine learning that can determine if a patient has diabetes or not. We discovered the fundamentals of preprocessing techniques like feature scaling and missing value imputed while developing this predictor. Using the accuracy score, we constructed the Random Forest method and compared the results between the train and test sets of data. This gives us an accuracy of 98.87%.