DIABETES PREDICTION USING MACHINE LEARNING

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is the main source of energy and comes from the food one consumes. Insulin or Hormone made by Pancreas helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough or any insulin or doesn’t use insulin well. Glucose then stays in the blood and doesn’t reach the cells.

Over time, having too much glucose in the blood can cause health problems. Although diabetes has no cure, one can take steps to manage the diabetes and stay healthy.

Sometimes people call diabetes “A touch of Sugar” or “Borderline Diabetes”. These terms suggest that someone doesn’t really have diabetes or has a less serious case, but every case of diabetes or has less serious case, but every case of diabetes is serious.

# **OBJECTIVE**

In this Project, I tried to build a machine learning model to accurately predict whether or not the patients in the dataset have diabetes or not

# **DETAILS ABOUT DIABETES**

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

* **Pregnancies**: Number of times pregnant
* **Glucose**: Plasma glucose concentration 2 hours in an oral glucose tolerance test
* **Blood Pressure**: Diastolic blood pressure (mm Hg)
* **Skin Thickness**: Triceps skin fold thickness (mm)
* **Insulin**: 2-Hour serum insulin (mu U/ml)
* **BMI**: Body mass index (weight in kg/(height in m)^2)
* **DiabetesPedigreeFunction**: Diabetes pedigree function
* **Age**: Age (years)
* **Outcome**: Class variable (0 or 1)

# **DEMONSTRATION**

The aim of this study was to create classification models for the diabetes data set and to predict whether a person is sick by establishing models and to obtain maximum validation scores in the established models. The work done is as follows:

1) Diabetes Data Set read.

2) With Exploratory Data Analysis; The data set's structural data were checked. The types of variables in the dataset were examined. Size information of the dataset was accessed. The 0 values in the data set are missing values. Primarily these 0 values were replaced with NaN values. Descriptive statistics of the data set were examined.

3) Data Preprocessing section; df for: The NaN values missing observations were filled with the median values of whether each variable was sick or not. The outliers were determined by LOF and dropped. The X variables were standardized with the rubost method..

4) During Model Building; Logistic Regression, KNN, SVM, CART, Random Forests, XGBoost, LightGBM like using machine learning models Cross Validation Score were calculated. Later Random Forests, XGBoost, LightGBM hyperparameter optimizations optimized to increase Cross Validation value.

5) Result; The model created because of XGBoost hyperparameter optimization became the model with the lowest Cross Validation Score value. (0.90)

Designed by Surmeet Mohanty