Phase-2: Innovation

Title: AI Based Diabetes Prediction System

Introduction:

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Diabetes mellitus (DM) is classified as-

Type 1:

Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient.

Type 2:

Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.

Type 3:

Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person.

Objectives:

1. Predict if person is diabetes patient or not.
2. Find most indicative features of diabetes.
3. Try different classification methods to find highest accuracy.

Installing Libraries:

In this first step I have imported most common libraries used in python for machine learning such as Pandas, Seaborn, Matplitlib etc.

I am using Python because if very flexible and effective programming language i ever used. I used Python in software development field too.

Importing Data:

The diabetes data set was originated from **https://www.kaggle.com/datasets/mathchi/diabetes-data-set.** Diabetes dataset containing 768 instances with 9 features.The objective is to predict if the patient is diabetic or not. The “Outcome” is the feature we are going to predict ,0 means No diabetes, 1 means diabetes.

Missing Value Analysis:

Next, i will cleanup the dataset which is the important part of data science. Missing data can lead to wrong statistics during modeling and predictions.

Feature Engineering:

Till now, i explored the dataset, did missing value corrections and data visualization. Next, i have started feature engineering. Feature engineering is useful to improve the performance of machine learning algorithms and is often considered as applied machine learning.

Outlier Detection:

In this part i removed all the records outlined in dataset. Outliers impacts Model accuracy. I used *Tukey Method* used for outlier detection.

Modeling:

In this sections, i tried different models and compare the accuracy for each. Then, i performed Hyperparameter Tuning on Models that has high accuracy.

Before i split the dataset i need to transform the data into quantile using `sklearn.preprocessing` .

Data Splitting:

Next, i split data in test and train dataset. Train dataset will be used in Model training and evaluation and test dataset will be used in prediction. Before i predict the test data, i performed cross validation for various models.

Prediction:

Till now, i worked on EDA, Feature Engineering, Cross Validation of Models, and Hyperparameter Tuning and find the best working Model for my dataset. Next, I did prediction from my test dataset and storing the result in CSV.

Conclusion:

1. Diabetes is one of the ricks during Pregnancy. It has to be treat to avoid complications.
2. BMI index can help to avoid complications of diabetes a way before
3. Diabetes start showing in age of 35 – 40 and increase with person age.