**Smart Diabetes Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds**

**Abstract :**

Type 2 diabetes mellitus (DM) is a chronic condition whose prevalence has been gradually growing across the world. About 30 million people in India have Diabetes, and are many more at risk. Therefore, to preclude diabetes and its symptoms related to it, so early diagnosis is required. The idea behind utilizing different methods for hypothetical determination of Type-2 Diabetes dependent on indicative analysis the enhancement of the diagnosis period of the disease through the way toward assessing suggestive qualities and day by day habits, permitting the estimating of Type 2 Diabetes without the need of clinical tests through prescient analysis. An immense measure of clinical information is accessible today with respect to the infection, their manifestations, purposes behind sickness, and their consequences for wellbeing. Using various machine learning techniques, Since these algorithms are so accurate, the threat of Type-2 Diabetes can be prevised, which is crucial for the medical field.

## Motivation:

Diabetes Mellitus, also called as diabetes, is a category of metabolism disorders characterized by high glucose levels in the blood (DM). Stages are a set of events that occur over time. One of the most defining characteristics of diabetes is the presence of a family history of the disease .Diabetes is caused by the pancreas' failure to produce sufficient insulin, as well as the body's ineffective use of insulin. Diabetes mellitus is divided into three types .Insulin-subordinate diabetes mellitus disease (IDDM) is a form of diabetes mellitus characterized by a pancreas that produces less insulin than present in the body requires ().Individuals suffering from DM type-1 .To make up for the least insulin provided by the pancreas, the external insulin dose is required..

**Existing:**

Researchers discovered that among all the features in PIDD, BMI and the most important variables for diabetes prediction using logistic regression on PIDD are age and glucose level. Huang et. al. in [18] To select and identify diabetes features, researchers used the NaIve-bayes, IB1, and C4.5 algorithms. Patient age, diagnosis time, insulin need, and diet control were found to be the most important variables for blood sugar regulation in the study. Treatment type, home monitoring, and the significance of smoking are all factors that influence the result..

**Disadvantages:**

A wide scope of viewable signs are likewise empowered normally by the Random Forest including shading, shape, surface and profundity. Arbitrary Forests are viewed as broadly useful vision apparatuses and considered as effective. Irregular Forest as characterized in is a conventional rule of classifier mix that utilizes L tree-organized base classifiers, where X indicates the information and is a group of indistinguishable and ward disseminated arbitrary vectors. Each Decision Tree is made by arbitrarily choosing the information from the accessible information

# Proposed System:

Python was used as the programming language for the analysis, and Python idle was used for implementation. On the collected data set and the Pima dataset, machine learning methods such as logistic regression and random forest classifications were used to predict diabetes. They compared all of the classifier predictions to one another. The diagram depicts the steps for implementing the machine learning algorithm..

# Advantages of proposed system :

# The proposed approach in this paper has three steps in its methodology. In step-1, load the diabetes dataset into RStudio for the purpose of pre-processing. Further Data pre-processing is done on loaded dataset with cross validation method with 10 folds and this process is repeated 3 times. This is a common configuration or standard method for comparing different models. Next to that, the preprocessed data is randomly divided into two sets namely training set and test set with the ratio of 80: 20 respectively which is commonly used ratio in literature. Apply different machine learning algorithms such as RF, LDA, CART and k-NN to learn the data patterns and train the data to get predictions. Then learn about the model to test the predictions with test dataset. After this step, analysis is performed based on accuracy and kappa metrics.

**Software Requirements**

* **Operating System:** Windows
* **Coding Language**: Python 3.7
* **Script:**

# Hardware Requirements:

* **Processor** - Pentium –III
* **Speed** – 2.4 GHz
* **RAM** - 512 MB (min)
* **Hard Disk** - 20 GB
* **Floppy Drive** - 1.44 MB
* **Key Board** - Standard Keyboard
* **Monitor** – 15 VGA Colour