DOMAIN : ARTIFICIAL INTELLIGENCE

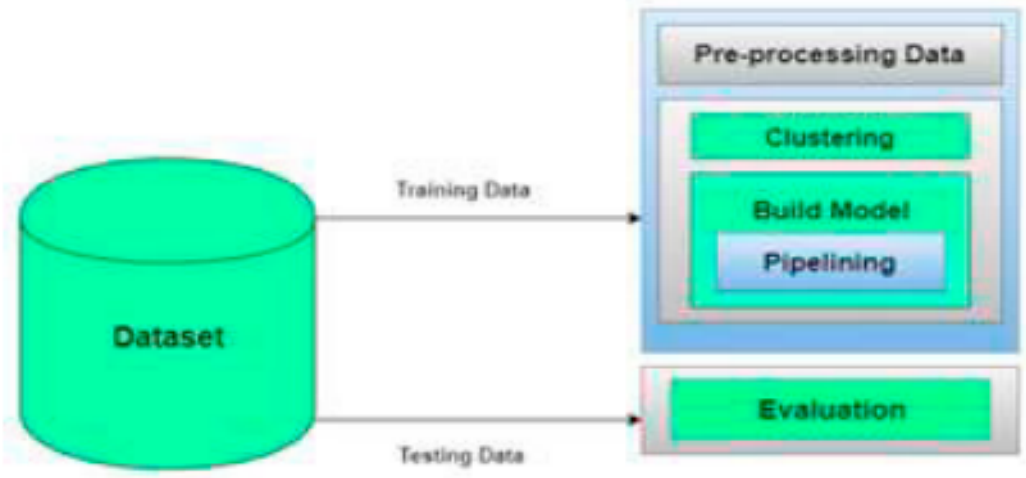
PROJECT: DIABETES PREDICTION USING AI

PHASE 3: DEVELOPMENT PART 1

Introduction:

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis.

- **In existing method, the classification and prediction accuracy is not so high. , we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.**



The diabetes development has five different modules .the modules are

- **Dataset Collection**
- **. Data Pre-processing**
- **. Clustering**
- **. Build Model**
- **. Evaluation**

Data set collection:

This Diabetes dataset contains 800 records and 10 attributes.

Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Job Type(Office-work/Field-work/Machine-work)	No

ii.Data Pre-processing This phase of model handles inconsistent data in order to get more

accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

iii.Clustering:

In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our record.

ALGORITHM:

- **Choose the number of clusters(K) and obtain the data points·**

- Place the centroids c_1, c_2, ... c_k randomly
- Steps 4 and 5 should be repeated until the end of a fixed number of iterations
- For each data point x_i:· -find the nearest centroid(c_1, c_2 ,..c_k) -assign the point to that cluster
- for each cluster j = 1..k· new centroid = mean of all points assigned to that cluster
- End·

iv.Model Building:

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression,

K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier iv.Model Building This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier.

Algorithm 1: Diabetes Prediction using various machine learning algorithms

- Generate training set and test set randomly
- . Specify algorithms that are used in model

- mn=[KNN( ), DTC( ), GaussianNB( ), LDA(),SVC(),LinearSVC(),AdaBoost(), RandomForestClassifier(), Perceptron(),
- ExtraTreeClassifier(), Bagging(), LogisticRegression(), GradientBoostClassifier()]

v.Evaluation:

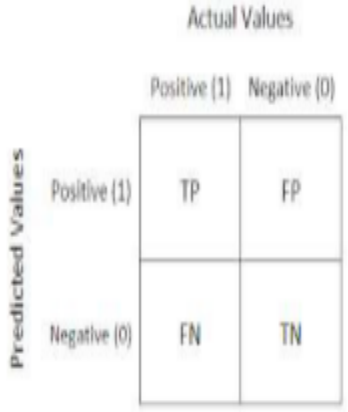
This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as

Accuracy=

Number of Correct Predictions/ Total number of predictions made

Confusion Matrix:- It gives us gives us a matrix as output and describes the complete

performance of the model



Where, TP: True Positive
FP: False Positive
FN: False Negative
TN: True Negative

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as-

Accuracy =tp+fn/n

Where, N:Total number of samples

F1 score-It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. .The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as- $FF1 = 2 * 1 (1 pppppppppppppppppppp) + (1 rrrrrrrrllll)$

F1 Score tries to find the balance between precision and recall. Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.