AI-Based Diabetes Prediction System

Problem Descrption:

The problem is to bulid an - powered diabetes prediction system that uses machine learning algorithms to analyse the medical data and predict the likelihood an individuals developing diabetes.The system aims to provide early risk assessment and - proactive actions to manage their Health.

Process Involved:

Step 1: Start

Step 2: Upload and read the medical dataset.

Step 3: After uploading the dataset Preprocessing the data.If it is ready for processing, then it divides into two parts as training and testing.

Step 4: If it is not ready, then again preprocess the data.

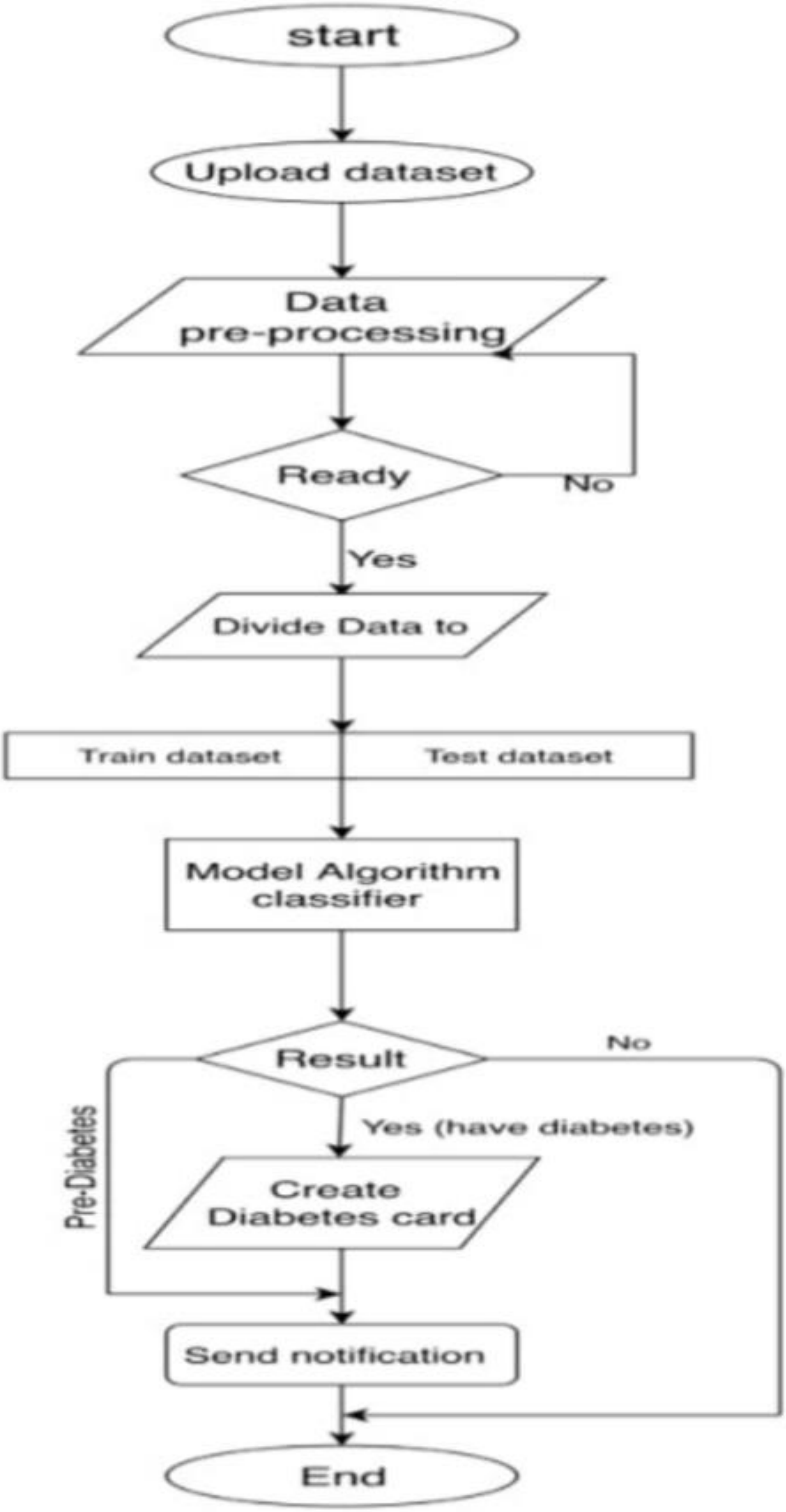
Step 5: For predicting the result apply the Macihne learning algorithms such as logistic regression, Random forest, Gradient Descent Algorithms.

Step 6: After predicting the result, the user will be notified**.**

Step 7: If the person have diabetes positive, then it creates the diabetes card and send the notification to the user. Otherwise it stops the process.

Step 8: Stop

Flow Chart:


start
Upload dataset
Data pre-processing
Ready
No
Yes
Divide Data to
Train dataset
Test dataset
Model Algorithm classifier
Result
No
Pre-Diabetes
Yes (have diabetes)
Create Diabetes card
Send notification
End

Uploading Dataset:

The datasets consists of several medical predictor(independent) variables and one target (dependent) variable Outcome.Independent variables include the BMI, Insulin level, age, Glucose, Blood pressure, Skin Thickeness, Diabetes pedigree. These dataset are obtained from **PIMA INDIANS DIABETES DATABASE** from **kaggle.com**

Data Preprocessing:

In the merged dataset, we discovered a few exceptional zero values. For example, skin thickness and Body Mass Index (BMI) cannot be zero. The zero value has been replaced by its corresponding mean value. The training and test dataset has been separated using the holdout validation technique, where 80% is the training data and 20% is the test data. First, the Pima Indian dataset was divided into an 8:2 ratio and three supervised regression models, extreme gradient boosting technique (XGB), support vector regression (SVR), and Gaussian process regression (GPR), have been employed to predict the selected outcome, that is, insulin of the validation samples of the Pima Indian dataset.

Machine learning Classifiers:

Various machine learning and ensemble techniques have been employed to implement the automatic diabetes prediction system.

Decision tree:

A decision tree represents the learning function provided by a set of rules. The decision tree learning technique performs a method for approximating discrete-valued target functions. Gini or entropy [7] are used to determine information gain, and each node is chosen based on these coefficients, which are expressed as

$Ginii =1-\sum k=1n(pi, k)2$ **(or)** $ENTROPY=\sum i=1n-pilog2pi$

KNN classifier:

A discrete-valued function can be approximated by K number of nearest classifiers [8]. To categorize, it creates a plane with the available training points and calculates the distance between the query and trained points. It determines the K number of neighbours (depending on the dataset) and classifies them using majority voting. In our research, we used K = 5 for the binary classification.

Random forest:

Random forest is a machine learning system that averages the predictions of several decision trees. As a result, the random forest can be considered an ensemble learning model [7]. In this research, we have applied random forest with estimators = 400, minimum samples leaf = 5, and ‘Gini’ impurity metrics utilizing hyperparameter tuning.

Support vector machine:

SVM performs supervised classification by choosing the best hyperplane [11]. In this study, we experimented with various SVM kernels in the training set. Finally, we discovered the SVM with a linear kernel, parameters C = 10 and gamma = 1, produces the best results in this dataset.

Logistic regression:

Logistic regression can be used to predict a binary class. To predict the outcome, it fits an ‘S’ shaped function [8]. The hyperparameter optimization technique obtained the maximum number of iterations for the convergence of the logistic regression model to be 150.