AI-Based Diabetes Prediction System Project By Loading And Preprocessing The Dataset

PHASE 3 DOCUMENTATION

PROJECT:AI-Based Diabetes Prediction System

ABSTRACT:

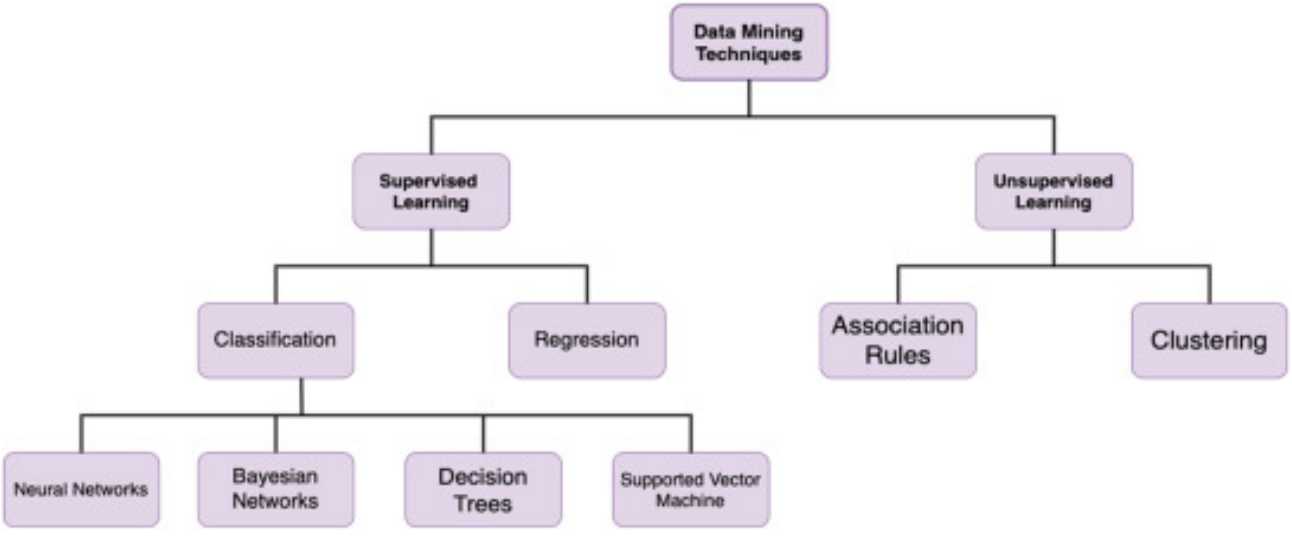
Diabetes is the leading cause of death in the world, and it also affects kidney disease, loss of vision, and heart disease. Data mining techniques contribute to health care decisions for accurate disease diagnosis and treatment, reducing the workload of experts. Diabetes prediction is a rapidly expanding field of research. Early diabetes prediction will result in improved treatment. Diabetes causes a variety of health issues. Therefore, it is critical to prevent, monitor, and raise awareness about it.

Type 1 and Type 2 diabetes can cause heart disease, renal problems, and eye difficulties. In this paper, we propose a diabetes prediction model using data mining techniques. We apply four data mining techniques such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. The proposed mechanism is trained using Python and analysed with a real dataset, which is collected from Kaggle. Furthermore, the performance of the proposed mechanism is analysed using the confusion matrix, sensitivity and accuracy performance metrices. In logistic regression, the accuracy is high, i.e., 82.46%, in comparison to other data mining technique

Attributes used in dataset:

Attributes	Description
Glucose	Plasma glucose concentration over 2 h in an oral glucose tolerance test.
Pregnancies	It shows how many times patient is pregnant.
Blood Pressure	It indicates BP of patient.
Skin Thickness	It shows skin fold thickness.
Diabetes Pedigree Function	It shows family history of patient.
BMI	It indicates Body mass index.
Insulin	2-Hour serum insulin (mu U/ml)
Age	It shows age of patient. The age group to be used is 21–81 for analysis.
Outcome	1 for diabetes and 0 for non-diabetes

Diabetes Prediction Model:



Data Collection:

- Gather a dataset that includes relevant information about individuals, such as age, gender, BMI, family history, blood pressure, glucose levels, and other health-related parameters.
- Ensure the dataset is diverse and representative of the population you intend to make predictions for.

Data Preprocessing:

- Handle Missing Data: Deal with missing values, either by imputing them or removing rows/columns with too many missing values.
- Data Cleaning: Check for outliers and errors in the data and decide whether to correct or remove them.

Feature Selection:

- Feature selection is a critical step to choose the most relevant variables for your model. You can use various methods, including:
  - **Univariate Feature Selection**: This method involves selecting features based on statistical tests like chi-squared, ANOVA, or mutual information scores.
  - **Recursive Feature Elimination (RFE)**: It recursively fits the model and eliminates the least important features.
  - **Feature Importance from Trees**: If you're using decision tree-based algorithms, you can use the feature importance scores provided by these models.
  - **Principal Component Analysis (PCA)**: A dimensionality reduction technique that can be used to reduce the number of features while preserving the most important information.

Data Splitting:

- Split your dataset into a training set and a testing set. The training set is used to train your model, and the testing set is used to evaluate its performance.

Feature Scaling:

- Standardize or normalize the feature values so that they have a similar scale. Common methods include z-score scaling or min-max scaling.

Model Selection:

- Choose an appropriate machine learning algorithm for your problem. Common algorithms for classification tasks like diabetes prediction include logistic regression, decision trees, random forests, support vector machines, and deep learning models.

Model Training:

- Train your chosen model on the training data.

Model Evaluation:

- Evaluate the model's performance on the testing dataset using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Iterate and Fine-Tune:

- Based on the evaluation results, you may need to fine-tune hyperparameters, try different algorithms, or collect more data if the model's performance is not satisfactory.

Deployment:

- Once you are satisfied with your model's performance, deploy it as a real-time or batch prediction system. You can use various tools and platforms for deployment, depending on your infrastructure and requirements.

Monitoring and Maintenance:

- Continuously monitor the model's performance in the production environment and update it as needed to adapt to changing data patterns.

Ethical Considerations:

- Ensure that your data collection, model development, and deployment processes are conducted in an ethical and privacy-conscious manner

Types of Diabetes:

In this section types of diabetes are discussed. There are broadly four types of diabetes. These are:

•**Type 1 DM**: Historically, the terms “insulin-induced diabetes mellitus” (IDDM) and “juvenile diabetes” were employed. The cause is unknown. Diabetes affects youth and those under the age of 20. Type 1 will harm pancreatic cells, rendering them dysfunctional [7]. Due to a lack of insulin secretion, type 1 diabetes patients have been afflicted throughout their entire lives and are insulin dependent. Patients with type 1 diabetes should constantly engage in physical activity and have a balanced diet.

•**Type 2 DM**: Insulin resistance, which occurs when cells do not respond appropriately to insulin, is the underlying cause of type 2 diabetes. Insulin insufficiency may develop as the condition advances. The phrase ‘non-insulin-based diabetes mellitus' or ‘adult-induced diabetes' has been used previously. Obesity and lack of exercise are the leading causes [8]. Typically, it occurs at the age of four.

•**Gestational Diabetes**: This is the third basic kind of gestational diabetes, which occurs when pregnant women develop elevated blood sugar levels without a history of diabetes. Approximately 18% of pregnant women have diabetes, according to the most recent study on diabetes. Possibility of increased gestational diabetes in elder women. The third major type is frequently induced by excessive blood sugar levels in pregnant women.

•**Pregestational Diabetes**: Pregestational diabetes occurs prior to the onset of insulin-dependent diabetes during pregnancy. A guy with prediabetes is more likely to receive a score of 2 under such settings or measurements.

Effects OF Diabetes:

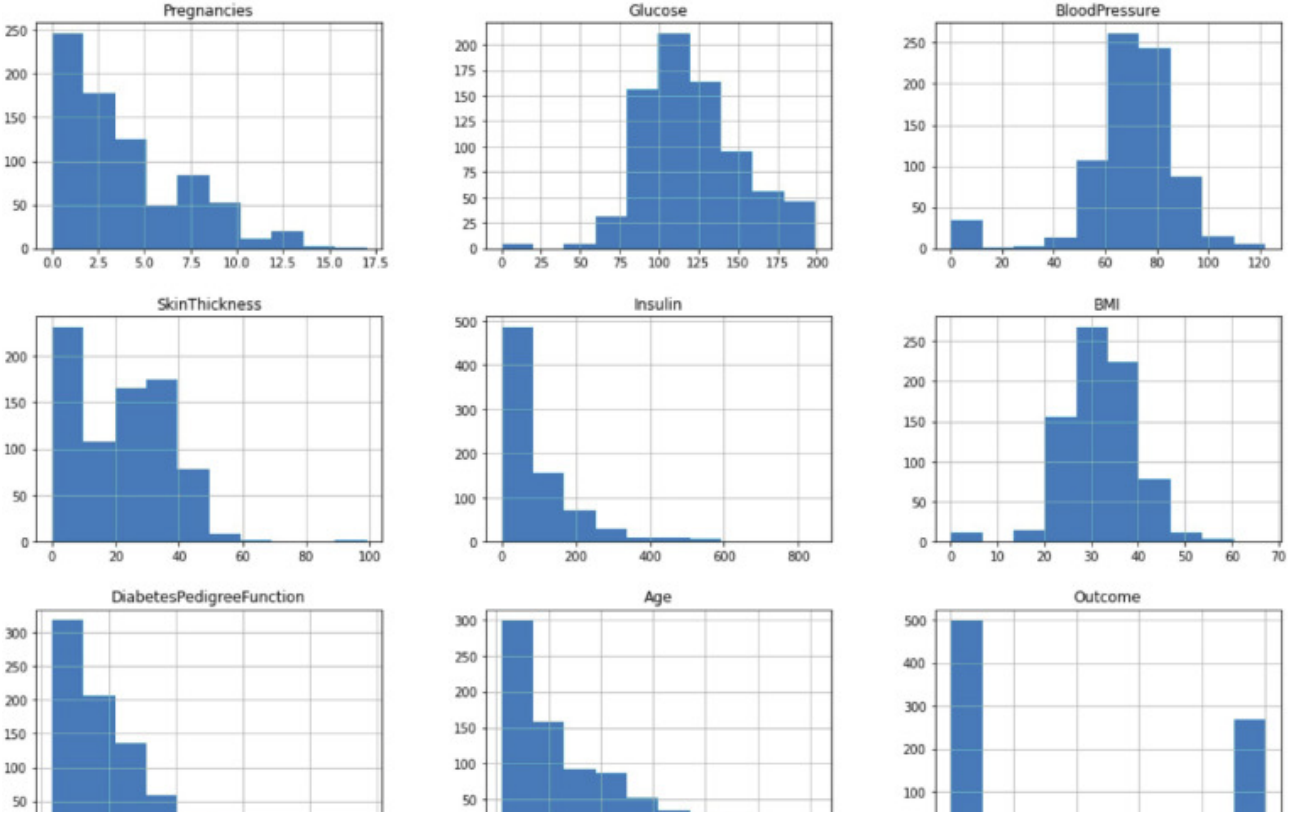
Diabetes can affect human body such as loss of vision, kidney neuropathy, liver and on heart [9]. In this section effects of diabetes have been discussed. These are:

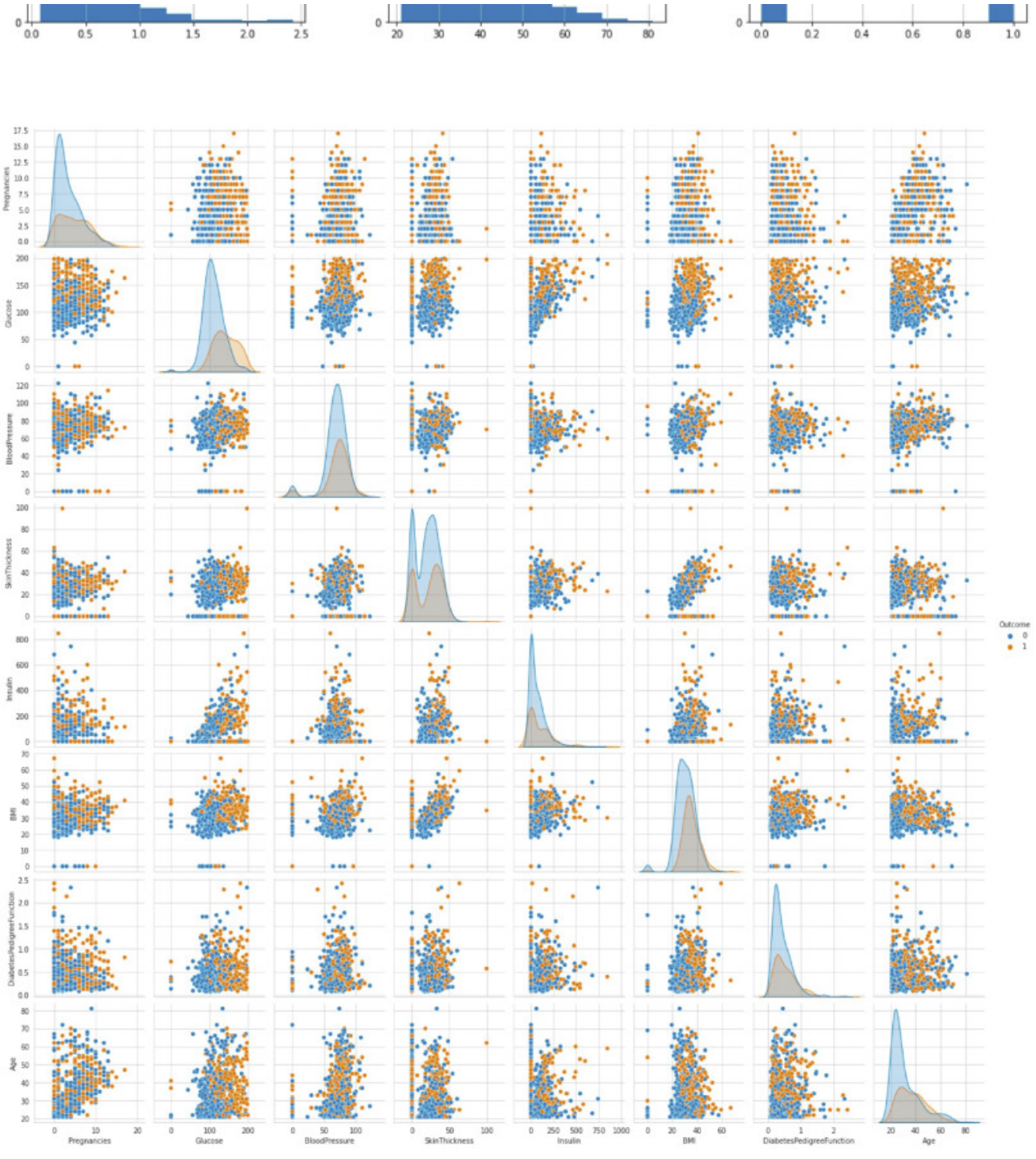
•**Loss of vision**: It is a condition that affects the retina and optic nerve of the eye. Due to night-time vision issues and swelling in the retina, mental contact may be diminished. A diabetic's eye vision can be restored with a few diagnostic procedures or pharmaceuticals.

•**Kidney Neuropathy**: Higher amounts of blood sugar harm the renal arteries, resulting in chronic kidney disease or diabetic neuropathy. The kidney is effective in transporting waste and large quantities of water into the blood.

•**Liver Problems**: Through the breakdown of starch via glucogenesis or glycogenolysis, the liver plays a crucial role in regulating the quantity of blood glucose in the circulation. Diabetes type 2 increases the chance of developing liver problems. A fatty liver is responsible for the development of a liver tumour.

•**Heart Disorders Cardiovascular Ailments**: It's continuous damage to blood vessels and neurons leads to the deception of the circulatory system or organ frame. Cardiovascular disease risk factors include hypertension, abnormally high cholesterol and triglyceride levels, obesity, and lack of physical activity. Multiple clinical characteristics, such as poor glycaemic management and insulin resistance in diabetes, influence cardiac issues.





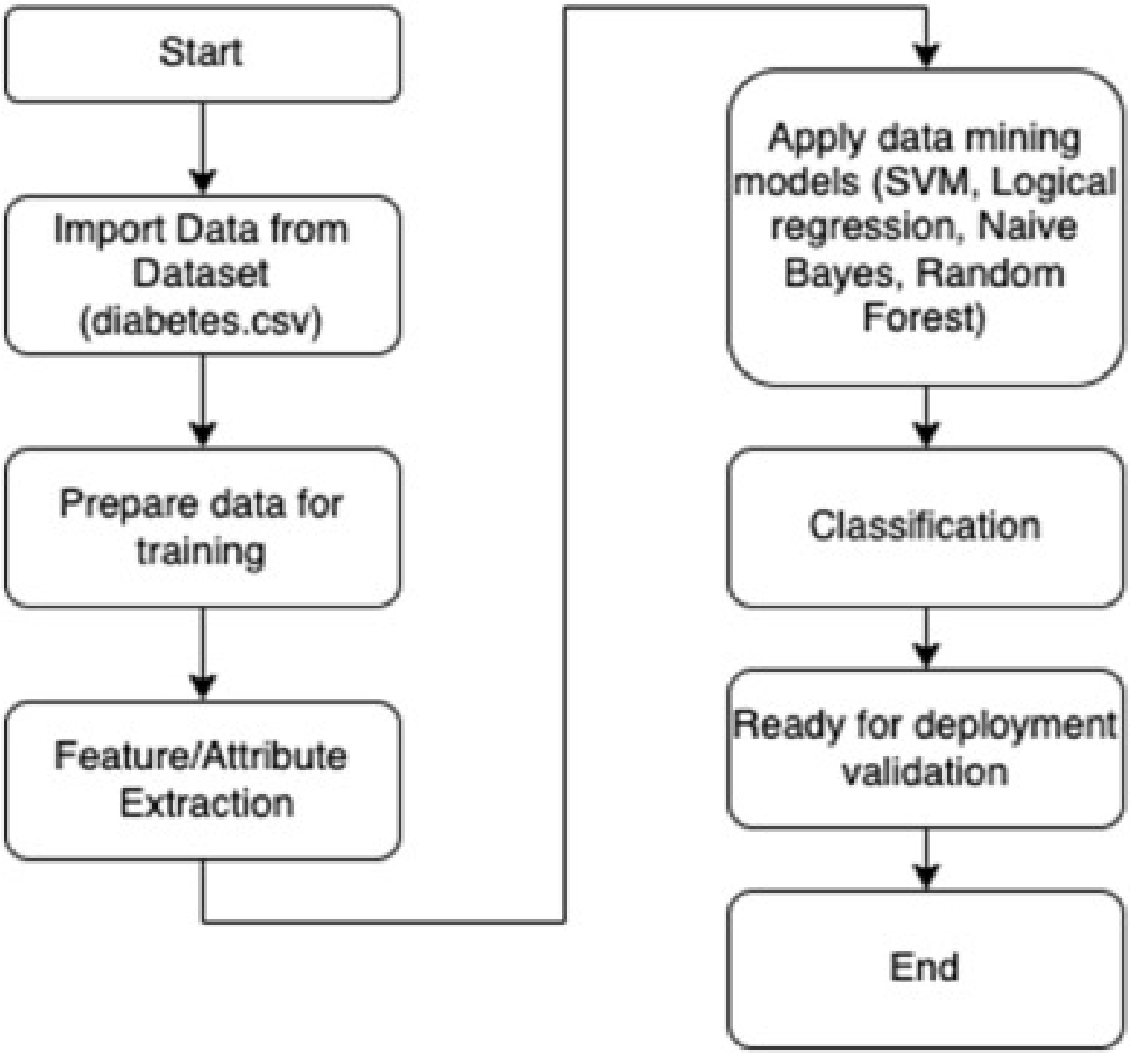
DATASET LINK:

LINK: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

PROPOSED ALGORITHM:

1. INITIALIZE: x, dataset, data_mining_model,
2. IMPORT: dataset = diabetes.csv
3. PREPARE: training dataset: training_data
4. EXTRACT: features -> from training_data
5. APPLY: data mining model ->new_data = data_mining_model(training_data)

5. APPLY: data_mining_model -> new_data = data_mining_model(training_data)
6. CLASSIFY: new_data
7. DEPLOY new_data -> dataset
8. END



Program:

```
from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression
```

```
X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.2)

model = LogisticRegression()

model.fit(X_train, y_train)

from sklearn.metrics import accuracy_score, classification_report


y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)

print("Accuracy:", accuracy)

print(classification_report(y_test, y_pred))

new_data = [feature_values]  # Replace with actual feature values

prediction = model.predict(new_data)

if prediction[0] == 1:

   print("The patient may have diabetes.")

else:

   print("The patient may not have diabetes.")
```

CONCLUSION:

Nowadays, data mining plays a crucial role in diabetes prediction in the healthcare system. Diabetes is a major health challenge in the world. Early prediction of diabetes will result in improved results. This paper presents a diabetes prediction model with the help of data mining techniques. We apply Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine techniques to predict diabetes disease. The proposed mechanism is implemented using Python. To analyse the proposed mechanism, a real dataset is collected from Kaggle. Accuracy, confusion, and sensitivity matrices are used to assess performance. In the logistic regression model, the accuracy is high, i.e., 82.46% as compared to other models. whereas in SVM the accuracy is low, i.e., 79.22% as compared to other models. In the future, it is intended to continue working on it and apply more classification algorithms to predict diabetes datasets. It is also meant to suggest a new way to make predictions about diabetes outcomes more accurate.