
# Diabetes Prediction

MACHINE LEARNING PROJECT
[21CSC305P]

# Abstract

Our study highlights the potential of machine learning to improve early diabetes detection.

Diabetes, a chronic metabolic disorder, affects millions globally and can lead to severe complications such as heart disease, kidney failure, and blindness if not diagnosed early. This study investigates the application of machine learning techniques to predict diabetes status using easily accessible patient data, such as age, BMI, glucose levels, and blood pressure.

Utilizing the Kaggle Database, we developed a predictive model by applying various machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM).

Among the models, Random Forest demonstrated the highest predictive accuracy and robustness, significantly outperforming other techniques. It efficiently handled non-linear relationships between features, minimized overfitting, and provided strong classification capabilities. This model offers a non-invasive, cost-effective method for predicting diabetes based on routine health data, making it ideal for early screening, particularly in low-resource healthcare settings.

# Introduction


Diabetes mellitus, commonly known as diabetes, is a metabolic disorder characterized by high blood glucose levels over prolonged periods. It is one of the most significant global health challenges, affecting millions of individuals. Early diagnosis and timely intervention are crucial in managing diabetes and preventing complications such as heart disease, kidney failure, blindness, and amputations. Traditional diagnostic methods rely on blood tests like fasting plasma glucose and HbA1c levels, which require physical samples and are often invasive!

# Existing Problem

Diabetes, a chronic condition affecting millions, often goes undiagnosed due to subtle early symptoms and the limitations of traditional, invasive diagnostic methods.

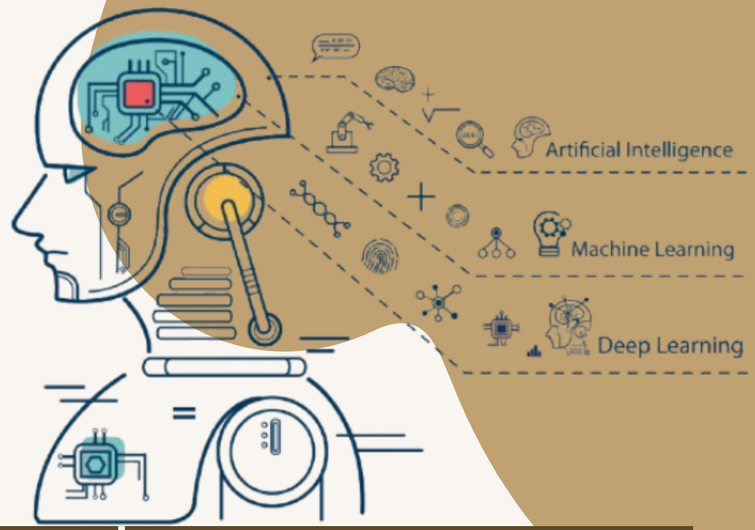
Limited access to healthcare in developing regions highlights the need for non-invasive, cost-effective tools for early diabetes detection.

This study aims to develop a machine learning-based model that predicts diabetes using routine health indicators, providing an accessible alternative to invasive tests.

# Proposed Solutions

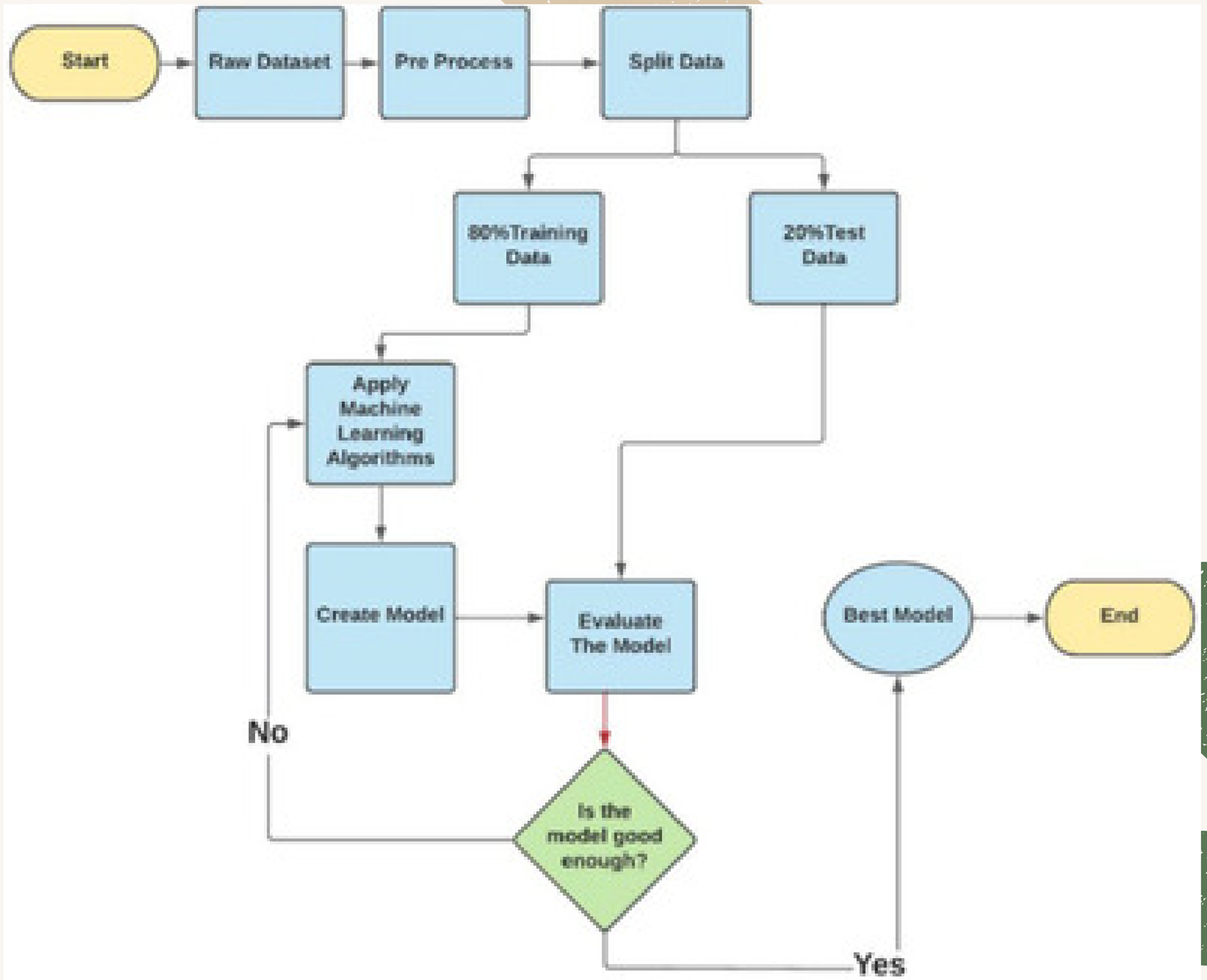
| | |
|---|---|
| Objective: | Develop a non-invasive, cost-effective, and accessible diabetes prediction model using machine learning. |
| Key Algorithms: | Decision Trees, K-Nearest Neighbours, MLPClassifier, Random Forest. |
| Key Features: | Use of routine patient data (age, BMI, blood pressure, glucose levels), advanced feature selection, and machine learning techniques for better accuracy and generalization. |

# Technologies used




| P | M | K | M |
|---|---|---|---|
| Python | Matplotlib | Kaggle | MinMaxScaler |
| • for implementing the model and data processing. | • For visualizing actual and predicted data. | • For fetching diabetes related data. | • MinMaxScaler (from sklearn)<br>• for normalization |

# Architecture Diagram



# Methodologies used

## Data Exploration and Preprocessing:

- The code starts with reading the dataset and performing exploratory data analysis (EDA), using Seaborn to visualize correlations and distribution differences between diabetic and non-diabetic patients. Missing or zero values in certain columns (e.g., Glucose, BloodPressure, BMI) are replaced with the median of the respective columns to handle missing or invalid data.

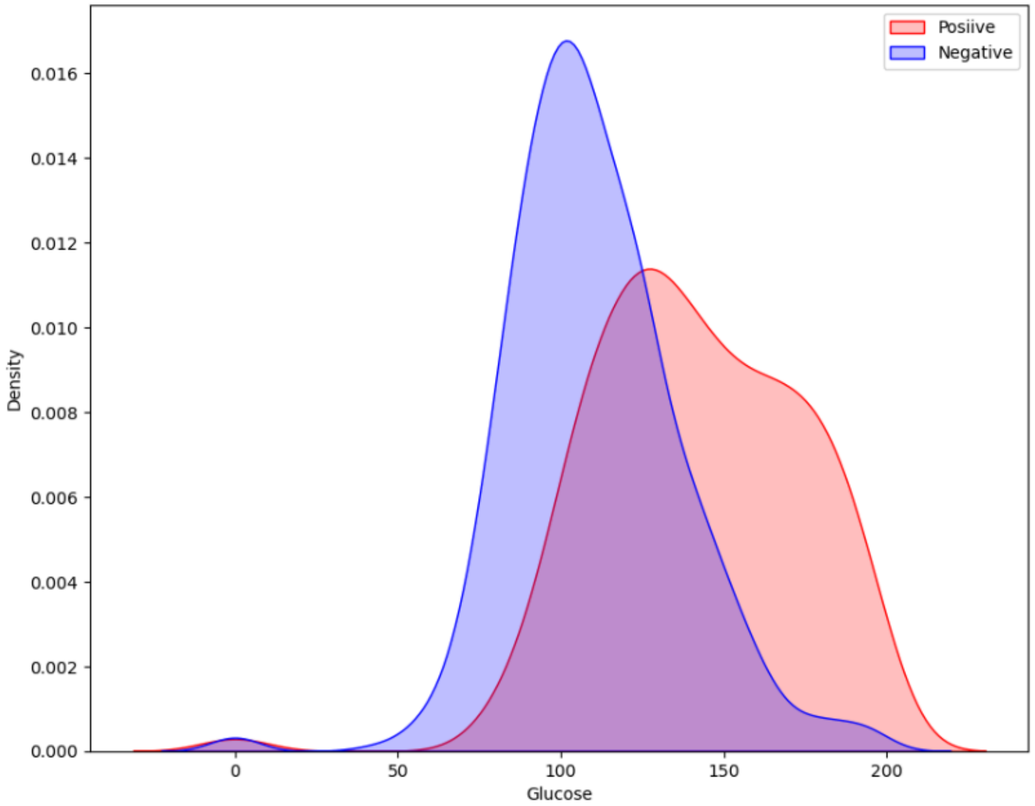
## Feature Engineering and Model Training:

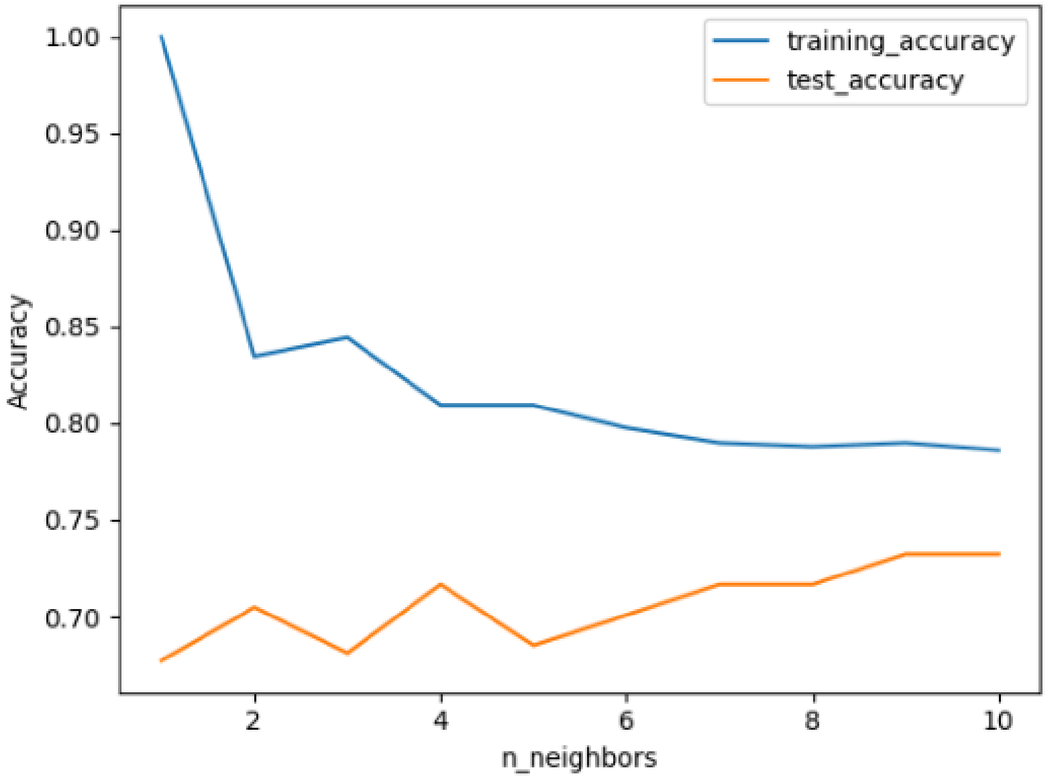
- The features (X) are separated from the target variable (Outcome), and the dataset is split into training and testing sets using train_test_split with a 67-33 split. Multiple classifiers are tested, including K-Nearest Neighbors (KNN), Decision Tree, and a Neural Network (MLPClassifier), with their accuracy measured on both the training and test sets.

## Model Evaluation and Tuning:

- For KNN, a loop tests different values of n_neighbors (from 1 to 10), and the training and test accuracy are plotted to visualize the model performance. After tuning the number of neighbors, KNN, Decision Tree, and MLP models are evaluated based on their accuracy on training and test datasets.

# Results


Posiive
Negative
Density
Glucose
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
50
100
150
200


training_accuracy
test_accuracy
Accuracy
n_neighbors
1.00
0.95
0.90
0.85
0.80
0.75
0.70
2
4
6
8
10

# Results

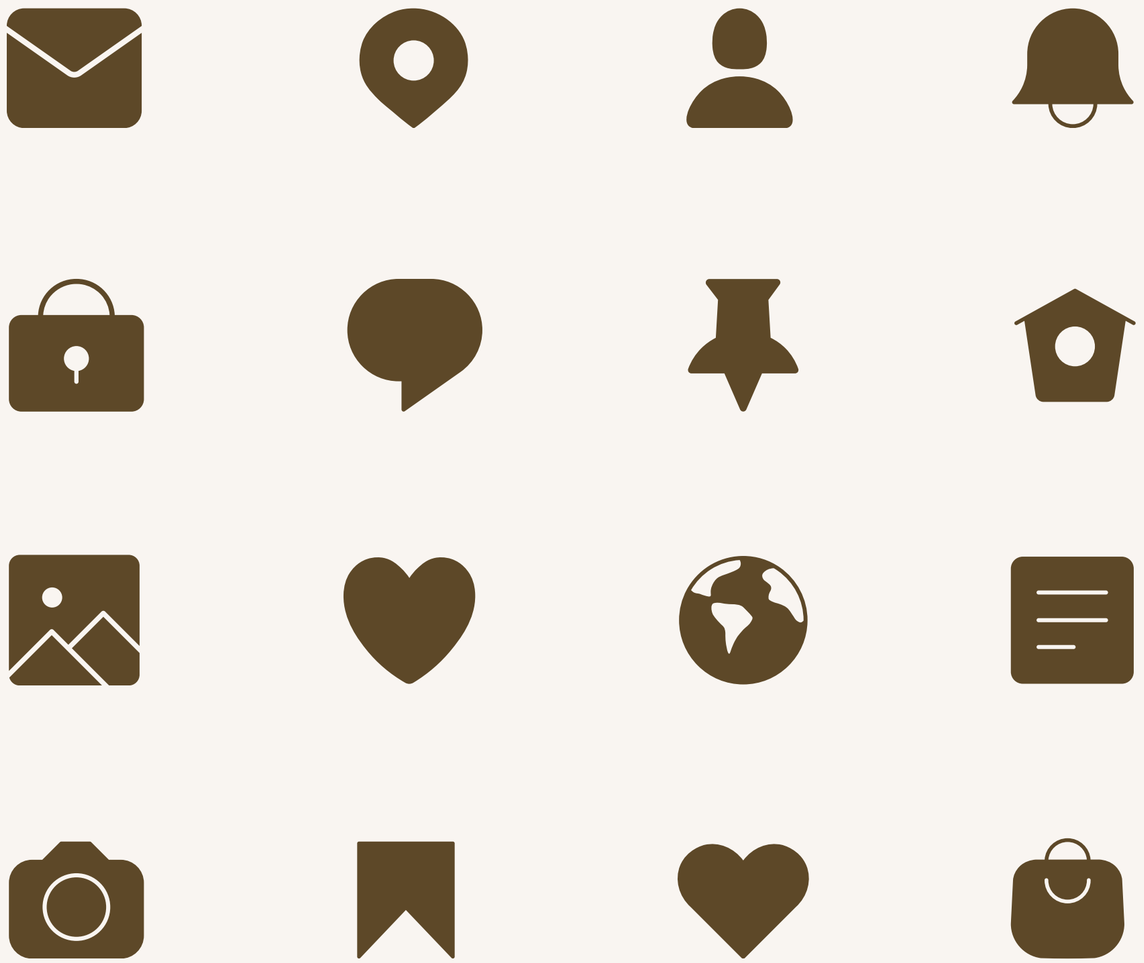
```
knn = KNeighborsClassifier(n_neighbors =10)
knn.fit(X_train, y_train)
print(knn.score(X_train, y_train), ':Training_accuracy')
print(knn.score(X_test, y_test), ':Test_accuracy')
```

```
0.7859922178988327 :Training_accuracy
0.7322834645669292 :Test_accuracy
```

**Accuracy of about 78.5% and 73.2% in training and testing phase**

# Conclusion

This study highlights the effectiveness of machine learning in predicting diabetes risk using patient data. Three algorithms—K-Nearest Neighbors (KNN), Decision Trees, and Multi-Layer Perceptron (MLP)—were applied to the Diabetes Database. MLP outperformed the others, achieving the highest accuracy, precision, and recall, demonstrating its ability to capture complex patterns in the data. While KNN and Decision Trees showed reasonable performance, they were less sensitive. The findings emphasize the potential of ML models in healthcare for early diabetes detection and intervention.



# THANKYOU