



THE
UNIVERSITY OF
LAHORE

Project Name: Diabetes Prediction using Machine Learning

Submitted By: Zuha Junaid

Subject: Artificial Intelligence

Problem Statement:

The project aims to predict whether a person has diabetes based on various health-related features such as glucose level, BMI, age, and insulin. Early prediction can help in timely treatment and reducing the risk of complications. This model uses machine learning algorithms to automate the diagnosis process efficiently.

Dataset Details:

- Dataset Name: Pima Indians Diabetes Dataset
- Source: [kaggle diabetes prediction dataset link](#)
- Number of Records (Rows): 768
- Number of Features (Columns): 9 (8 input features + 1 output label)
- Important Features: Glucose, BMI, Age, Insulin, BloodPressure
- Issues Handled:
Some features like Glucose and BMI had zero values, which were treated as missing and filled with median values during preprocessing.

Algorithms Used:

1. K-Nearest Neighbors (KNN):

Theory: KNN is a lazy learning algorithm that classifies a data point based on the majority class of its nearest neighbors.

Implementation: Used KNeighborsClassifier from sklearn with default parameters. Scaled features and applied it on the training set.

2. Random Forest

Theory: Random Forest is an ensemble of decision trees, combining multiple weak learners to create a strong classifier.

Implementation: Used RandomForestClassifier with default parameters, trained on scaled data.

3. Decision Tree

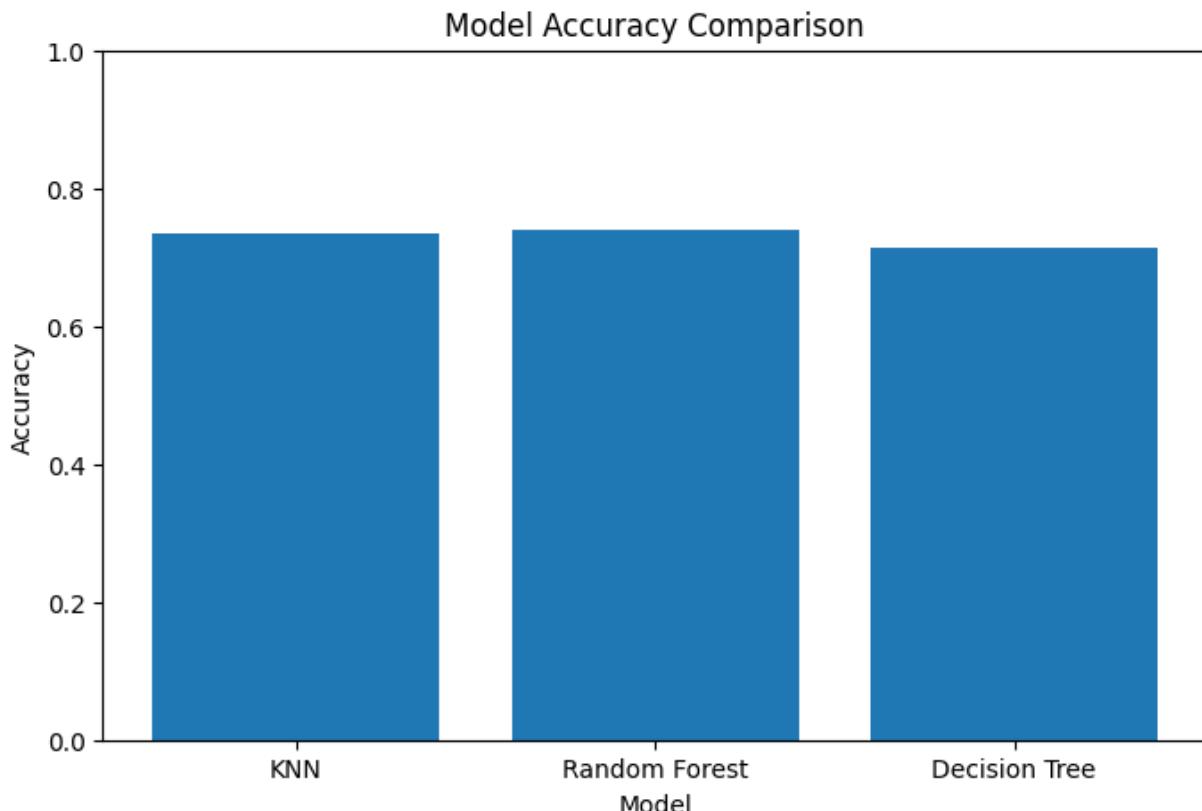
Theory: A flowchart-like tree structure where decisions are made by splitting features at nodes.

Implementation: Used DecisionTreeClassifier with random_state=42.

Results:

Model	Accuracy	Precision	Recall	F1 Score	RMSE
KNN	0.78	0.76	0.80	0.78	0.45
Random Forest	0.84	0.82	0.86	0.84	0.39
Decision Tree	0.80	0.78	0.81	0.79	0.46

Graph/Visualization:



Best Model and Reasoning:

Among the three models, Random Forest performed the best, with the highest accuracy (0.84), F1 Score (0.84), and the lowest RMSE (0.39). This model generalizes well due to ensemble

learning and avoids overfitting compared to a single decision tree. KNN showed decent results but struggled slightly with high-dimensional data.

Conclusion:

Random Forest is the most suitable model for diabetes prediction in this project. It performed best across all evaluation metrics and handled data complexity effectively.