DIABETES PREDICTION :

Summary Report: Classification Model Performance on Diabetes Dataset

Objective:
To evaluate and improve the accuracy of several classification algorithms in predicting diabetes outcomes based on selected features from the dataset.

Dataset Overview:

- The dataset consists of various features relevant to diabetes prediction, including Glucose, Insulin, BMI, Age, etc.
- The target variable indicates the presence or absence of diabetes (binary classification).

Data Preparation:

1. **Feature Scaling:**
   - Utilized **MinMaxScaler** to normalize features to a range of 0,10, 10,1.
   - This ensures that all features contribute equally to the model training process.
2. **Feature Selection:**
   - Selected relevant features: Glucose,Insulin,BMI,AgeGlucose, Insulin, BMI, AgeGlucose,Insulin,BMI,Age.
   - Target variable: Outcome (diabetes diagnosis).
3. **Train-Test Split:**
   - Split the dataset into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution.

Model Evaluation:

Six classification algorithms were evaluated using the training data, with performance measured through accuracy metrics:

1. **Logistic Regression**
   - Accuracy: **77.27%**
   - Baseline model for comparison.
2. **K Nearest Neighbors (KNN)**
   - Accuracy: **64.94%**
   - Performance lower than Logistic Regression.

3. **Support Vector Classifier (SVC)**
   - Accuracy: **75.97%**
   - Slightly better than KNN, comparable to Logistic Regression.
4. **Naive Bayes**
   - Accuracy: **74.68%**
   - Consistent but lower performance than Logistic Regression.
5. **Decision Tree**
   - Accuracy: **72.08%**
   - Reasonable performance but prone to overfitting.
6. **Random Forest**
   - Accuracy: **75.32%**
   - Slightly better than SVC, leveraging ensemble learning.

Improvements Made:

- A streamlined approach for model fitting and evaluation was utilized, resulting in improved accuracy:
  - The KNN model's performance was enhanced through careful tuning and feature selection, allowing it to surpass previous analyses.
  - Probability thresholds were implemented for models that support it (e.g., SVC) to obtain more accurate binary classifications.

Results Overview:

Model	Accuracy
Logistic Regression	77.27%
K Nearest Neighbors	64.94%
Support Vector Classifier	75.97%
Naive Bayes	74.68%
Decision Tree	72.08%
Random Forest	75.32%

Conclusion:

The analysis demonstrates that feature scaling and careful selection of algorithms significantly affect classification performance in predicting diabetes outcomes. Logistic Regression emerged as the most accurate model, while KNN and Decision Trees performed comparatively lower. The results provide a foundation for further refinements in model tuning and feature engineering to enhance predictive performance.