Comparative analysis of data mining for diabetes prediction

Varnasri K- 71762131058

Problem Statement

- Diabetes is a chronic and potentially debilitating condition affecting millions of people worldwide, leading to severe health complications if not managed effectively.
- Early detection and timely intervention are critical to mitigating the adverse effects of diabetes and improving patient outcomes.
- This project addresses the primary challenge of developing a reliable predictive model that accurately determines the likelihood of an individual having diabetes based on various health indicators such as glucose level, blood pressure, BMI, and other relevant features.
- By leveraging data mining techniques and machine learning algorithms, this project aims to provide a robust tool for individuals to assess their risk of diabetes, thereby facilitating early detection and intervention.

Abstract

- The rising prevalence of diabetes necessitates effective prediction tools for early diagnosis and treatment.
- This project aims to develop a predictive model for diabetes based on a dataset containing medical parameters like glucose levels, BMI, age etc.
- By preprocessing data, normalizing features, and applying logistic regression, decision tree, and K-nearest neighbors classifiers, we identify the best model for predicting diabetes and also identifying the risk factor.
- The project also includes a real-time prediction function based on user input. and a risk prediction component to categorize diabetes risk levels based on glucose values.

Dataset

Pregnancies: More frequent pregnancies may affect health conditions such as diabetes due to hormonal changes.

Glucose: Higher glucose levels are a direct indicator of diabetes.

Blood Pressure: Hypertension is often associated with diabetes.

Skin Thickness: It can be an indicator of body fat, which is related to diabetes.

Insulin: Insulin levels are critical in the diagnosis and management of diabetes.

BMI (Body Mass Index): Higher BMI values are associated with obesity, which is a risk factor for diabetes.

Diabetes Pedigree Function: A higher value indicates a higher risk of diabetes.

Age: Age is a significant risk factor for diabetes. Older individuals are more likely to develop diabetes.

HbA1c Levels: Higher HbA1c levels indicate poorer blood sugar control and a higher risk of diabetes.

Stress Levels: Higher stress levels can increase the risk of developing diabetes.

Sleep Quality: Poor sleep quality can negatively affect metabolic processes, including glucose metabolism, potentially leading to diabetes.

Family History: A positive family history (having relatives with diabetes) increases the likelihood of developing diabetes due to genetic factors.

Outcome: This is the target variable indicating whether the individual has diabetes (1) or not (0).

Data Cleaning and Preprocessing:

- Summary Statistics:
  - print(data.describe()) generates a summary of the dataset's statistical properties, including measures like mean, median, standard deviation, and percentiles for numerical columns.
  - Inference: The dataset contains a mix of continuous and binary variables.
  - Several columns have 0 values that likely indicate missing data (e.g., Insulin, Skin Thickness, BMI).
  - Most variables show some degree of skewness, with notable differences between means and medians.
  - The distribution of binary variables (Family History and Outcome) shows a clear divide.

Output:

```
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin
count   616.000000  615.000000     614.000000     617.000000  615.000000
mean      4.579545  126.442276      70.568404      21.317666   70.147967
std       3.314934   36.918382      21.185833      15.987705  116.916822
min       0.000000   71.000000       0.000000       0.000000    0.000000
25%       1.000000  101.000000      64.000000      10.000000    0.000000
50%       5.500000  122.000000      70.000000      23.000000    0.000000
75%       7.000000  159.000000      84.000000      32.000000   82.000000
max      13.000000  197.000000     110.000000      60.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age  HbA1c Levels  \
count  613.000000                607.000000  609.000000    613.000000
mean    30.597227                  0.585501   35.119869      6.899511
std     10.142220                  0.435021   11.178151      0.372436
min      0.000000                  0.134000   21.000000      5.800000
25%     24.200000                  0.294000   25.000000      6.700000
50%     29.700000                  0.491000   31.000000      6.800000
75%     39.100000                  0.696000   44.000000      7.100000
max     46.800000                  2.288000   60.000000      7.800000

       Stress Levels  Sleep Quality  Family History     Outcome
count     615.000000     617.000000      617.000000  617.000000
mean        3.473171       7.267423        0.377634    0.290113
std         0.903583       0.689526        0.485189    0.454183
min         2.000000       6.000000        0.000000    0.000000
25%         3.000000       7.000000        0.000000    0.000000
50%         3.000000       7.000000        0.000000    0.000000
75%         4.000000       8.000000        1.000000    1.000000
max         6.000000       8.000000        1.000000    1.000000
```

Data Information

- print(data.info()) provides a concise summary of the DataFrame, including the number of entries, data types of each column, and memory usage.
- The output shows that the dataset is a DataFrame with 617 entries, indexed from 0 to 616.It contains 13 columns.
- Inference:Dataset Completeness: The dataset is mostly complete, but some columns have missing values that need to be handled during preprocessing.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 617 entries, 0 to 616
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               616 non-null    float64
 1   Glucose                   615 non-null    float64
 2   BloodPressure             614 non-null    float64
 3   SkinThickness             617 non-null    int64  
 4   Insulin                   615 non-null    float64
 5   BMI                       613 non-null    float64
 6   DiabetesPedigreeFunction  607 non-null    float64
 7   Age                       609 non-null    float64
 8   HbA1c Levels              613 non-null    float64
 9   Stress Levels             615 non-null    float64
 10  Sleep Quality             617 non-null    int64  
 11  Family History            617 non-null    int64  
 12  Outcome                   617 non-null    int64  
dtypes: float64(9), int64(4)
```

Data Cleaning and Preprocessing:

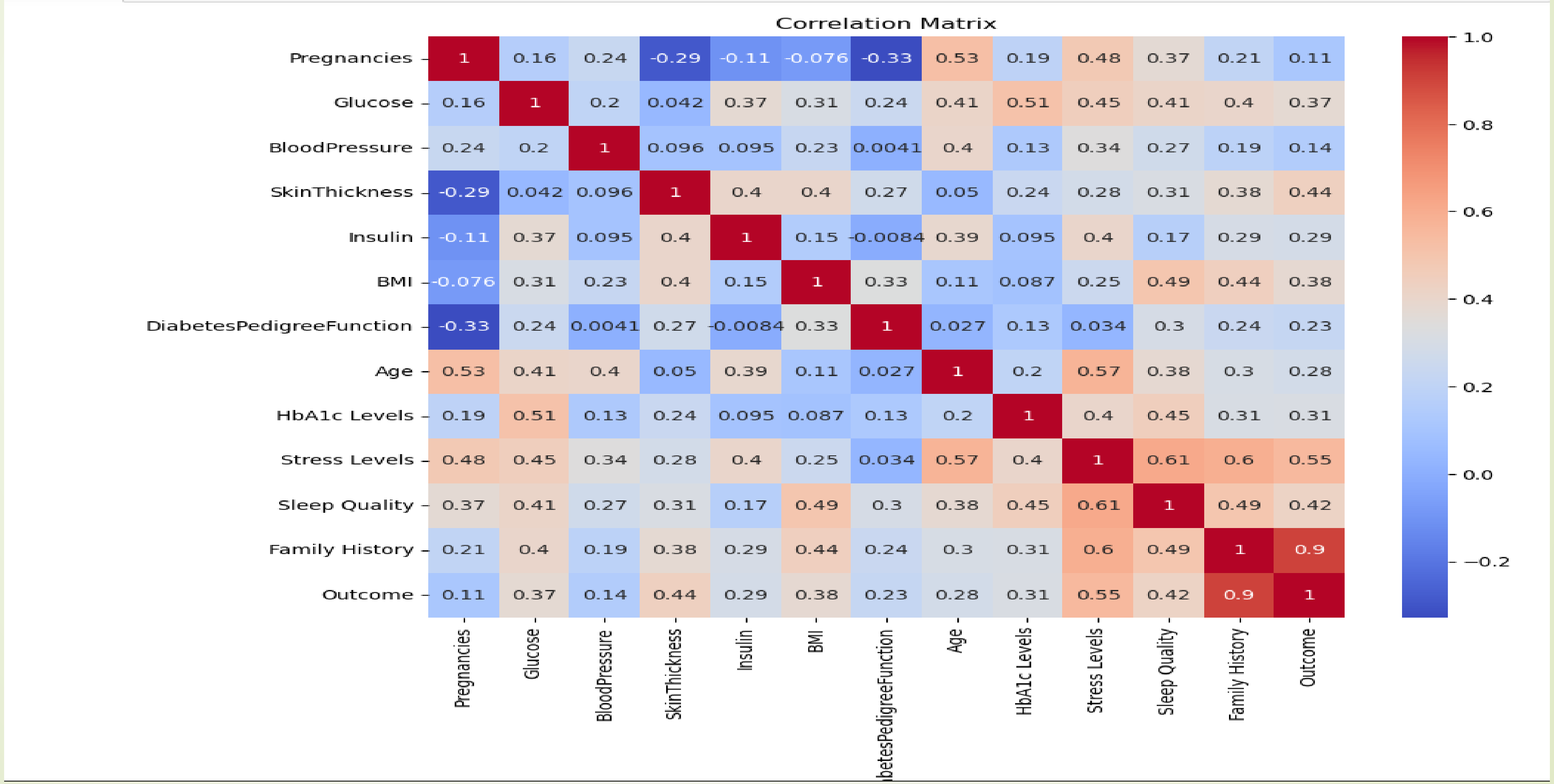
- Fill Missing Values: Datafolha(data.mean(), inplace=True) replaces missing values in the dataset with the mean of the respective columns.
- This is a common strategy for handling missing data, especially for numerical columns, as it helps to retain the dataset's size and structure while minimizing potential bias introduced by missing values.
- Removing Duplicate Rows:
- Drop Duplicates: data.drop_duplicates(inplace=True) removes any duplicate rows from the dataset.
- Duplicates can skew analysis and model training, so removing them helps to ensure the integrity and accuracy of the data.

Outlier Detection:

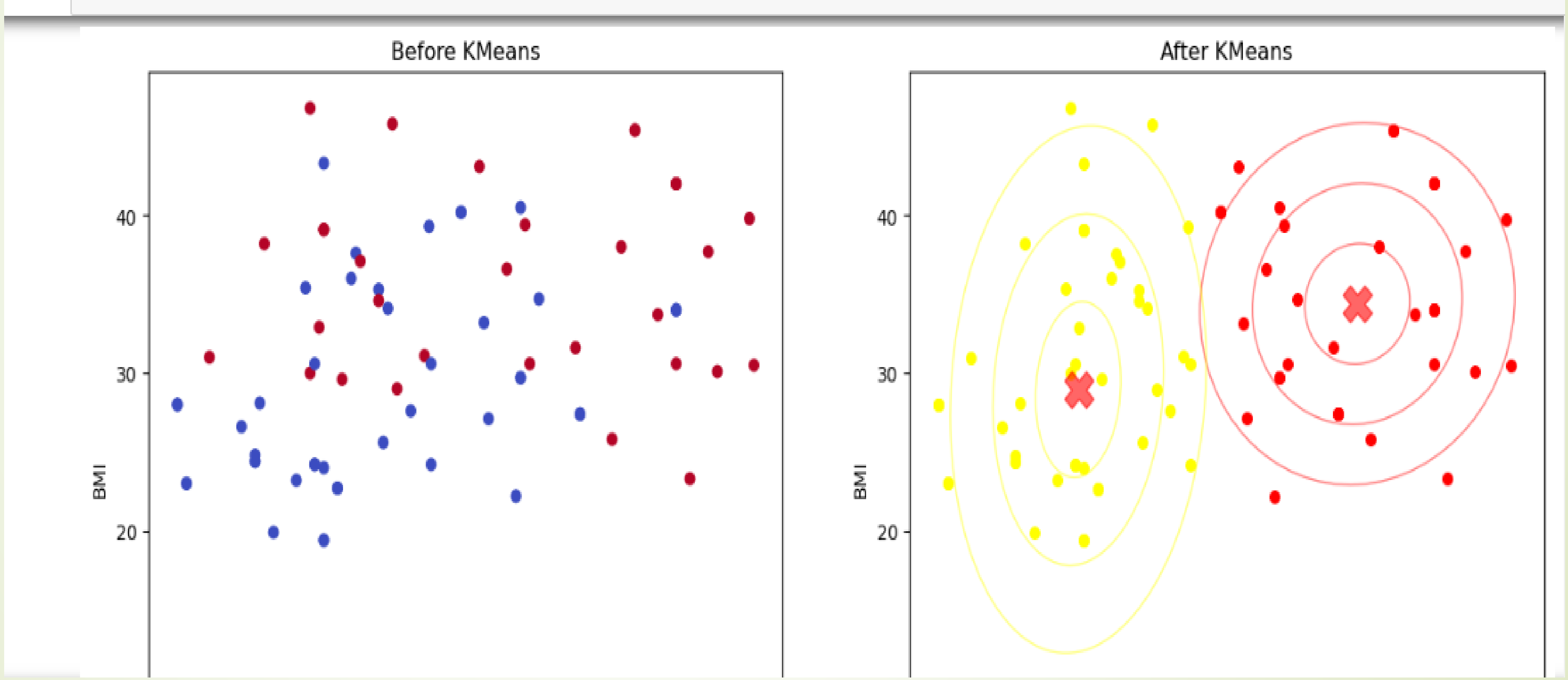
- Z-scores help identify outliers in the data.
- A data point with a Z-score greater than 3 or less than -3 is often considered an outlier.
- Outliers can skew results and affect the performance of machine learning models, so detecting and addressing them is crucial.

```
[81 rows x 11 columns]
Original data shape: (81, 13)
Cleaned data shape: (71, 13)
```

Correlation Analysis



K-Means Clustering



Mutual Information

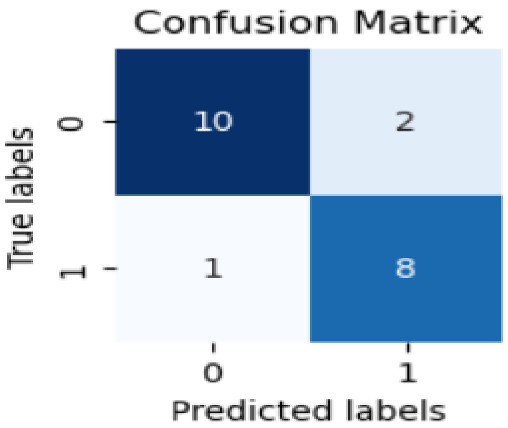
- Mutual information measures the dependency between each feature and the target variable. A higher mutual information score indicates a stronger relationship between the feature and the target variable.

```
Mutual information scores:
 Family History               0.523271
BMI                          0.236591
SkinThickness                0.189547
Age                          0.180113
HbA1c Levels                 0.125514
Stress Levels                0.105741
Sleep Quality                0.101695
BloodPressure                0.061102
Insulin                      0.058033
DiabetesPedigreeFunction     0.054335
Cluster                      0.046899
Glucose                      0.015636
dtype: float64
```

Logistic Regression

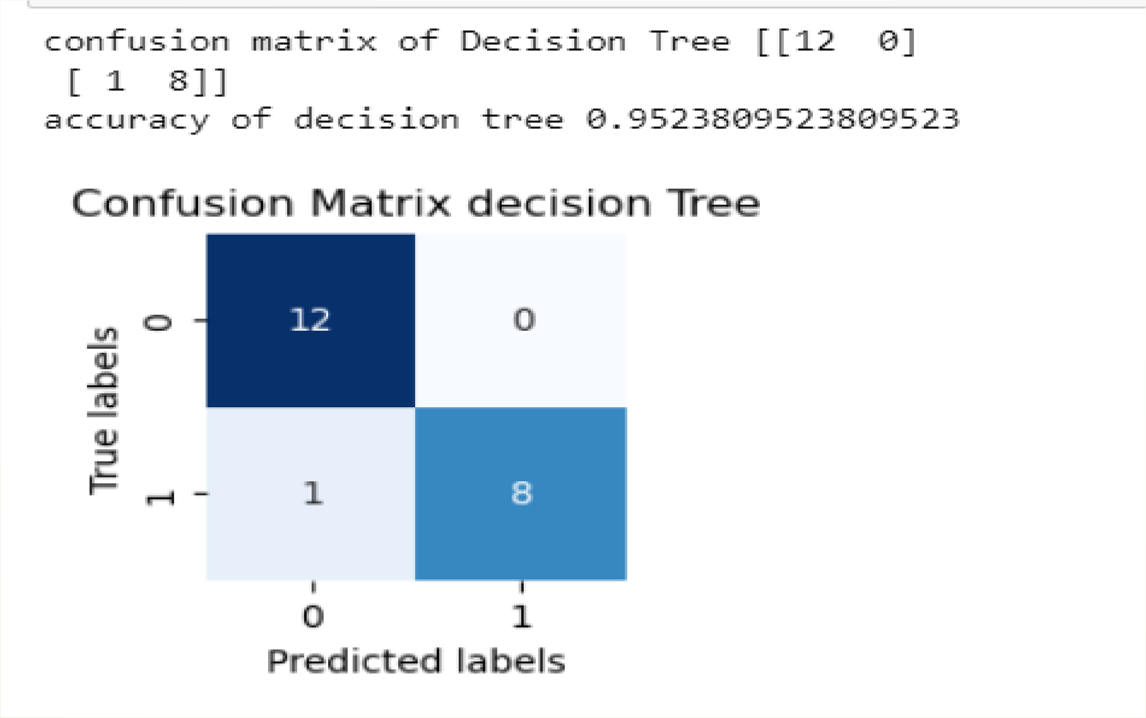
- Logistic regression is well-suited for binary classification problems like predicting whether an outcome is 0 or 1.
- It provides coefficients that can be easily interpreted to understand the influence of each feature on the probability of the outcome.
- Logistic regression is computationally efficient and performs well when the relationship between the features and the target variable is approximately linear.

```
Logistic regression accuracy 0.8571428571428571
confusion matrix of logistic regression [[10  2]
 [ 1  8]]
prediction of logistic regression [1 1 1 1 1 0 0 1 1 0 0 0 0 0 1 1 1 0 0 0 0]
Intercept:  [-7.61590177]
Coefficients:  [[ 0.02046365 -0.03147864  0.05585701 -0.00270719  0.08718083 -0.33242102
   0.0550702  -0.14273518  0.8026947  -0.34932055  2.40770665  0.89538902]]
```



Decision Tree

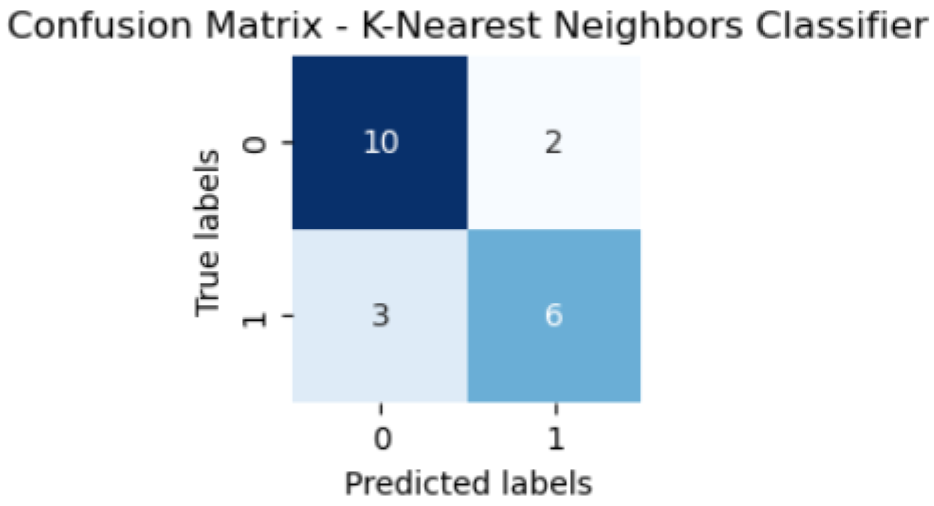
- Non-linear Relationships: Decision trees can capture non-linear relationships between features and the target variable.
- Interpretability: The model is easy to interpret as it mimics human decision-making processes and can show feature importance.



K-Nearest Neighbors (KNN):

- Instance-Based Learning: KNN is a non-parametric method that makes predictions based on the closest training examples in the feature space.
- Simplicity: It is simple to understand and implement.

```
confusion matrix of KNeighour Classification
[[10  2]
 [ 3  6]]
accuracy of KNeighour Classification 0.7619047619047619
KNeighour Classification prediction [0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 1 1 0 0 0 0]
```



Best Model

```
Best Model: Decision Tree
Enter Pregnancies: 0
Enter Glucose level: 120
Enter Blood Pressure: 98
Enter Skin Thickness: 30
Enter Insulin: 0
Enter BMI: 30
Enter Diabetes Pedigree Function: 0.790
Enter Age: 40
Enter HbA1c Levels: 8.6
Enter Stress Levels: 5
Enter Sleep Quality: 4
Enter Family History: 1
Prediction: Person has diabetes
Risk Level: Medium
```

Best Model –Decision Tree

- Decision Trees can capture non-linear relationships between the features and the target variable. In the case of diabetes prediction, the relationship between features such as glucose level, insulin, BMI, age, etc., and the outcome not be linear.
- Decision Trees can effectively model these complex interactions
- Decision Trees can automatically determine the most important features for making predictions.
- Features like glucose level, insulin, and BMI might have significant impacts on the prediction of diabetes, and the Decision Tree can prioritize these features in its splits.

Conclusion

- This project successfully developed a predictive model for diabetes diagnosis.
- After comparing multiple models, the Decision Tree model was identified as the most accurate.
- The project demonstrates the effectiveness of data mining in medical diagnostics and provides a practical tool for early diabetes detection.
- The final model can predict diabetes risk with a high degree of accuracy, potentially aiding healthcare professionals in making informed decisions.