CSE422 Lab Project Report

Diabetes Prediction Using Machine Learning

Name	ID	Email
Nouruzzaman Niloy	22141004	nouruzzaman.niloy@g.bracu.ac.bd
MD. Atiq Mahbub	22141013	atiq.mahbub@g.bracu.ac.bd

Table of contents

1. Introduction:

Our project focuses on predicting diabetes based on a set of clinical attributes using 3 machine learning models. The dataset that we used has 100,000 records from clinical data such as age, glucose levels, blood pressure and smoking history among others. The goal is to predict whether a patient is diabetic or non- diabetic and the dataset provides these attributes along with explicit labels for diagnosis, making it an ideal candidate for supervised classification tasks. We have implemented and compared the performance of 3 machine learning models which are Logistic Regression, Decision Trees, K- Nearest Neighbors (KNN). The project addresses challenges such handling missing data, encoding categorical features and managing imbalance data sets.

2. Dataset description:

Link:- https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/data

Reference**:** Choksi, Priyam. 100,000 Diabetes Clinical Dataset. Kaggle, https://www.kaggle.com/datasets/priyamchoksi/100000-diabetes-clinical-dataset/data, July, 2024.

Dataset Description:-

The dataset contains a total 16 features which can be categorized as quantitative and categorical. This is a classification problem which aims to predict whether a patient is diabetic or non-diabetic. The dataset includes 100,000 rows which means the number of data points is 100,000. There are two types of features as mentioned earlier. Which are:

Quantitative:- age, bmi, race:AfricanAmerican, race:Asian, race:Caucasian, race:Hispanic, blood_glucose_levels, hbA1c_levels.

Categorical:- gender, smoking history, location.

A correlation matrix (Fig. 1) was generated to examine the relationships between features and target variable (diabetes). Strong correlations were found between HbA1c level, blood glucose levels and diabetes status.

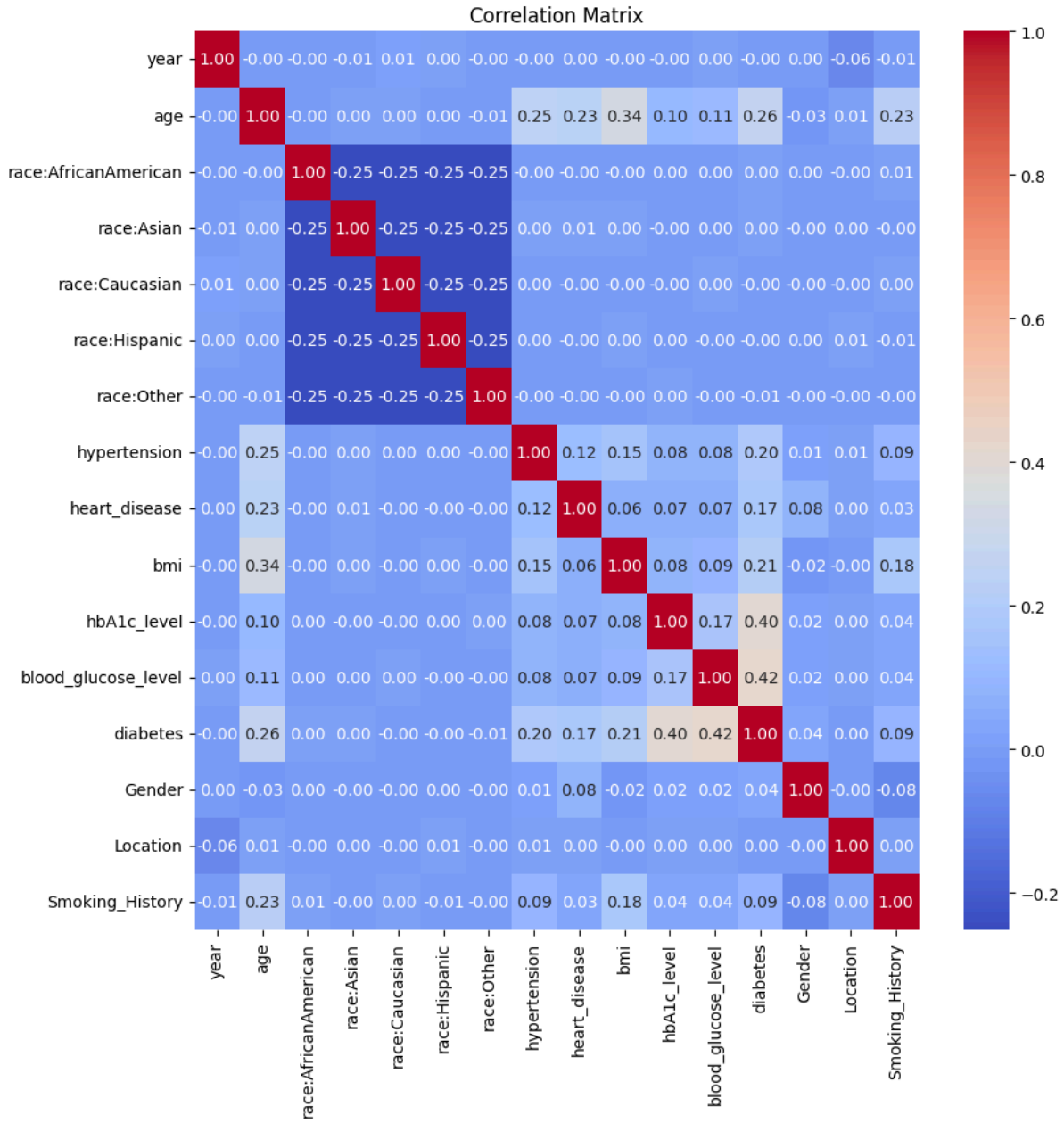

Fig. 1. Correlation matrix

The output feature (diabetes) exhibits imbalanced class distribution as there are a higher number of non-diabetic individuals compared to the diabetic individuals. A bar chart (Fig. 2) was plotted to display the frequency distribution of diabetic (1) and non-diabetic (0) patients, emphasizing the need for stratified splitting.

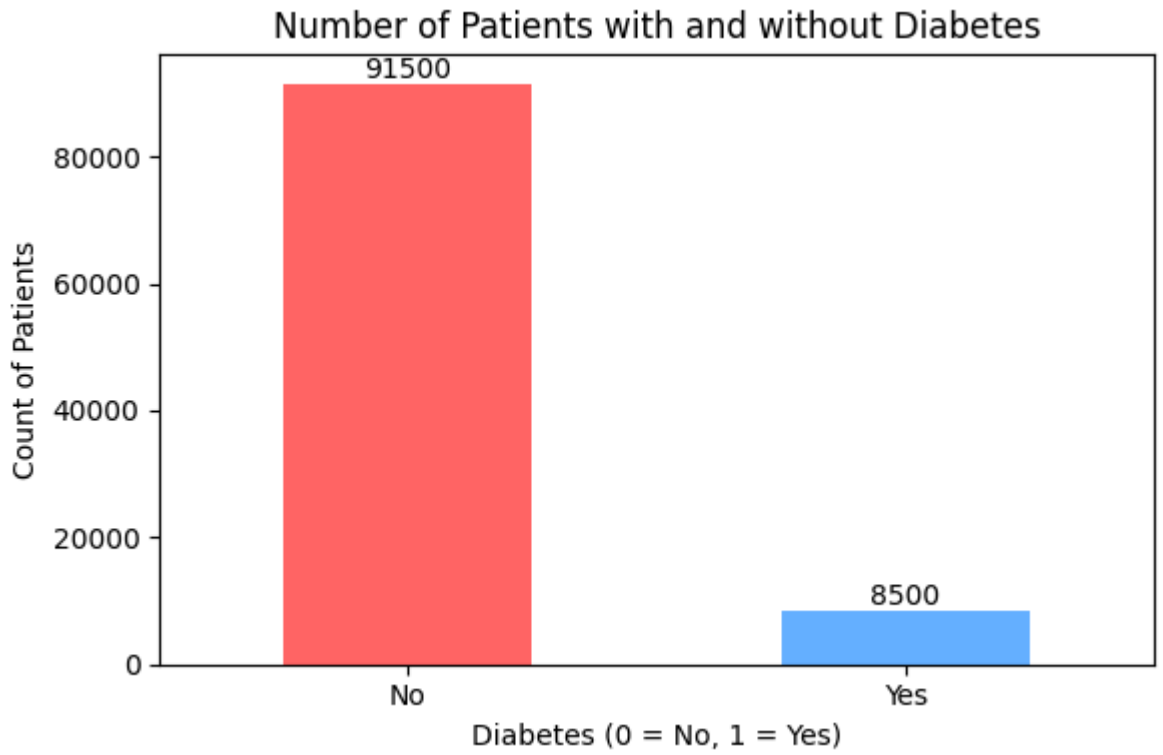

Fig. 2. Frequency distribution of diabetic and nondiabetic patients

3. Dataset Preprocessing:

During the preprocessing of the dataset, null values and duplicate values were identified and addressed. Initially, rows with null values in the "diabetes" column were dropped, as this information was critical for analysis. Subsequently, missing values in other columns were handled through imputation. Specifically:

- Null values in the "location" column were replaced with "Unknown".
- Null values in the "gender" column were replaced with "Other".
- Null values in the "smoking" column were replaced with "No Info".

Following the initial preprocessing steps, null values in the "year," "age," "blood glucose level," "hypertension," and "heart disease" columns were replaced with their respective mean values, which were rounded. This rounding was necessary because these columns represent categorical or discrete values that should be integers. Also, null values in the "bmi" and "hba1c" columns were replaced with their mean values without rounding, as these measurements can be expressed as floating-point numbers and do not require integer values. After that we removed the duplicate values from the dataset which removed the duplicate rows.

4. Feature scaling:

Feature scaling was performed using `StandardScaler` to standardize numerical data. This was to ensure that all features contributed equally to the model, avoiding bias from larger numerical ranges.

5. Dataset Splitting:

For each model, we have splitted dataset into 2 parts with 7:3 ratio, meaning 70% of the dataset will be train data that will be used to train every mode. Then on the trained model, 30% of test data will be used to predict clusters. This method guaranteed that both the training set and the testing set maintained the original sample population or original data's characteristics.

6. Model Training and Testing:

Logistic Regression:

Logistic regression is a widely used model for binary classification tasks, and in the case of diabetes prediction, it is particularly well-suited.Our goal is to classify individuals as either "diabetic" or "non-diabetic," which is a binary classification problem. Logistic regression is specifically designed for binary outcomes, making it an ideal model for predicting diabetes. It produces probabilities that represent the likelihood of an individual having diabetes (a value between 0 and 1), which can easily be converted into a binary classification. This resulted in model accuracy of 96%.

The Logistic Regression model's performance in predicting diabetes among the patients demonstrates high accuracy and precision, presented in the confusion matrix (Fig.3). The model accurately classified all 27115 no-diabetes cases (true negatives) with 236 false positives. Additionally, it correctly identified 1598 out of 2544 actual diabetes cases (true positives), with 946 instances misclassified as no-diabetes (false negatives).

Class	Precision	Recall	F1-Score	Support
0.0	0.97	0.99	0.98	27,351
1.0	0.87	0.63	0.73	2,544

Accuracy			0.96	29,895
Macro Avg	0.92	0.81	0.85	29,895
Weighted Avg	0.96	0.96	0.96	29,895

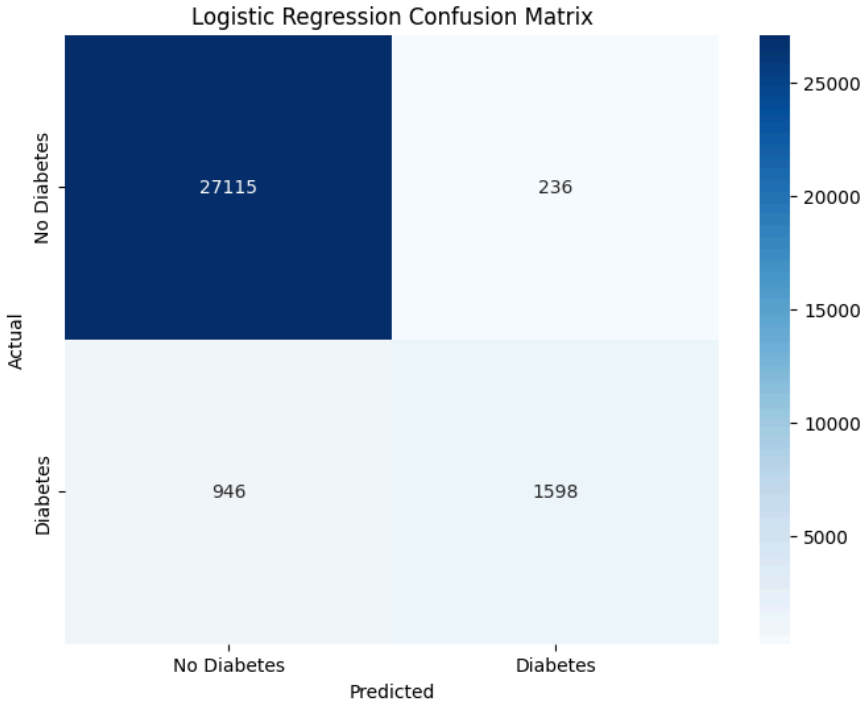

Fig. 3. Logistic regression confusion matrix

Decision tree:

Decision Tree is one of the most popular and powerful models for binary classification tasks. In this case, diabetes prediction. It works by step by step division of data based on the most important features, leading to a final prediction at the leaf nodes. This helps the model easily handle complex patterns and relationships between features. Decision Trees will classify whether a person is diabetic or not from the given health data for diabetes prediction. This model resulted in an accuracy of 95%.

The Decision Tree model's performance in predicting diabetes among the patients demonstrates also gives high accuracy and precision, presented in the confusion matrix (Fig. 4). The model accurately classified all 26543 no-diabetes cases (true negatives) with 808 false positives. Additionally, it correctly identified 1855 out of 2544 actual diabetes cases (true positives), 689 with instances misclassified as no-diabetes (false negatives).

Class	Precision	Recall	F1-Score	Support
0.0	0.97	0.97	0.97	27351
1.0	0.70	0.73	0.71	2544
Accuracy			0.95	29895
Macro avg	0.84	0.85	0.84	29895
Weighted avg	0.95	0.95	0.95	29895

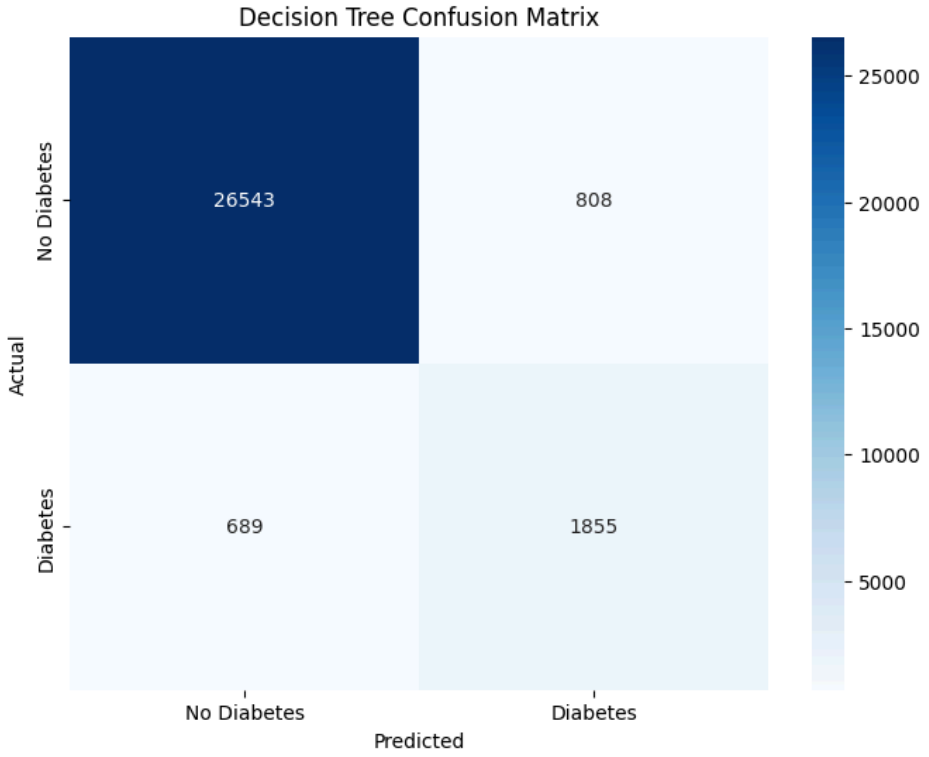

Fig. 4. Decision Tree confusion matrix

K- Nearest Neighbors (KNN):

KNN is another popular model to execute a classification task for binary classes. Here, the aim is to classify a patient either as diabetic or nondiabetic, similar to the logistic regression model. KNN will look for the nearest neighbors in which a given person locates and predict the majority class of those nearest neighbors. A person is classified here by matching his health records to the records of patients closer to him. By this it predicts whether a person is diabetic or not from the given health data for diabetes prediction. Just like the decision tree this resulted in model accuracy of 95%.

The KNN model's performance in predicting diabetes among the patients demonstrates high accuracy and precision just like the other two models, presented in the confusion matrix (Fig. 5). The model accurately classified all 27174 no-diabetes cases (true negatives) with 177 false positives. Additionally, it correctly identified 1365 out of 2544 actual diabetes cases (true positives), 1179 with instances misclassified as no- diabetes (false negatives).

Class	Precision	Recall	F1-Score	Support
0.0	0.96	0.99	0.98	27351
1.0	0.89	0.54	0.67	2544
Accuracy			0.95	29895
Macro avg	0.92	0.77	0.82	29895
Weighted avg	0.95	0.95	0.95	29895

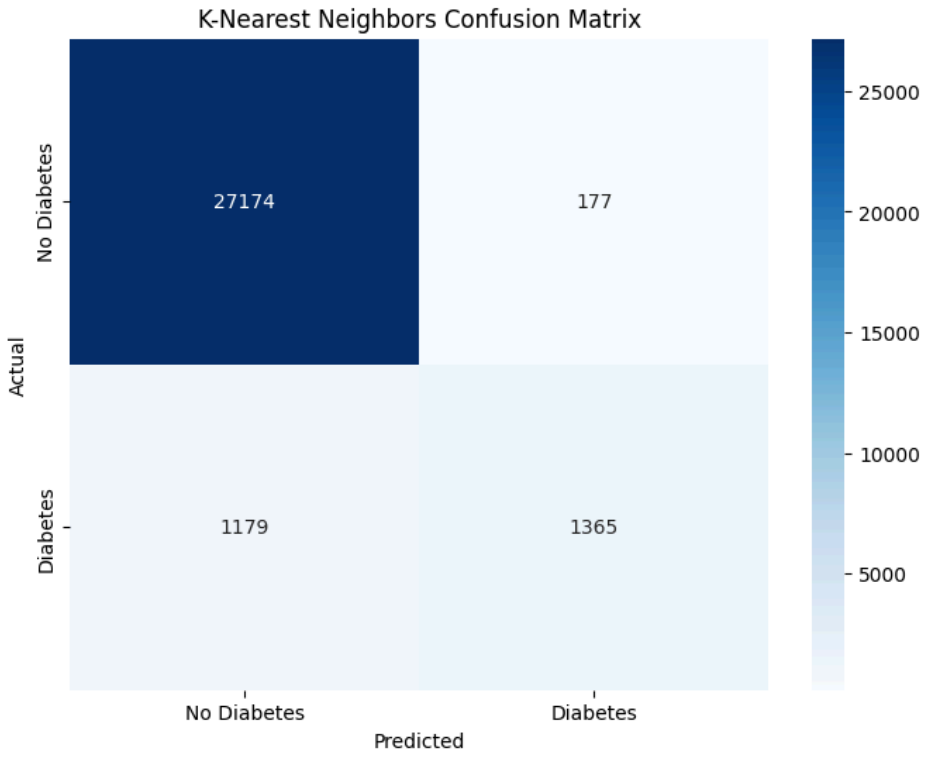

Fig. 5. K- Nearest Neighbors (KNN) confusion matrix

7. Model selection and comparison analysis:

In our project, we used three machine learning models: (1) Logistic Regression, (2) Decision Tree, and (3) K- Nearest Neighbors (KNN). The comparison of the accuracy scores(precision,recall, f1-score) can be found in Fig. 6.

- **Precision:** Precision indicates the accuracy of positive predictions. It gives us the proportion of predicted positive instances that are actually positive.

- **Recall:** Recall measures how accurately a model can identify actual positive instances. It gives us the proportion of actual positive cases that were correctly identified by the model.

- **F1 Score:** The F1 score calculates the harmonic mean of Precision and Recall. Since our dataset is imbalanced, It is used to balance the trade-off between Precision and Recall.

These scores can be calculated using the following equations:

Precision=TP/(TP+FP) TP=True positives TN=True Negatives

Recall= TP/(TP+FN) FP=False positives FN=False Negatives

F1 score=2×[(Precision+Recall)/(Precision×Recall)]

After conducting our tests, the logistic regression model came with highest accuracy(96%) and precision (96%). While we got high accuracy across all three models,the risk associated with false negatives where failing to detect diabetic individuals could lead to severe health risks. In that case, the Decision tree model gives us the best result with the least amount of false negative cases(689 out of 2544). Since our goal is to minimize the risk of misdiagnosis, the Decision Tree model is the most reliable in identifying both the diabetic and non-diabetic individuals.

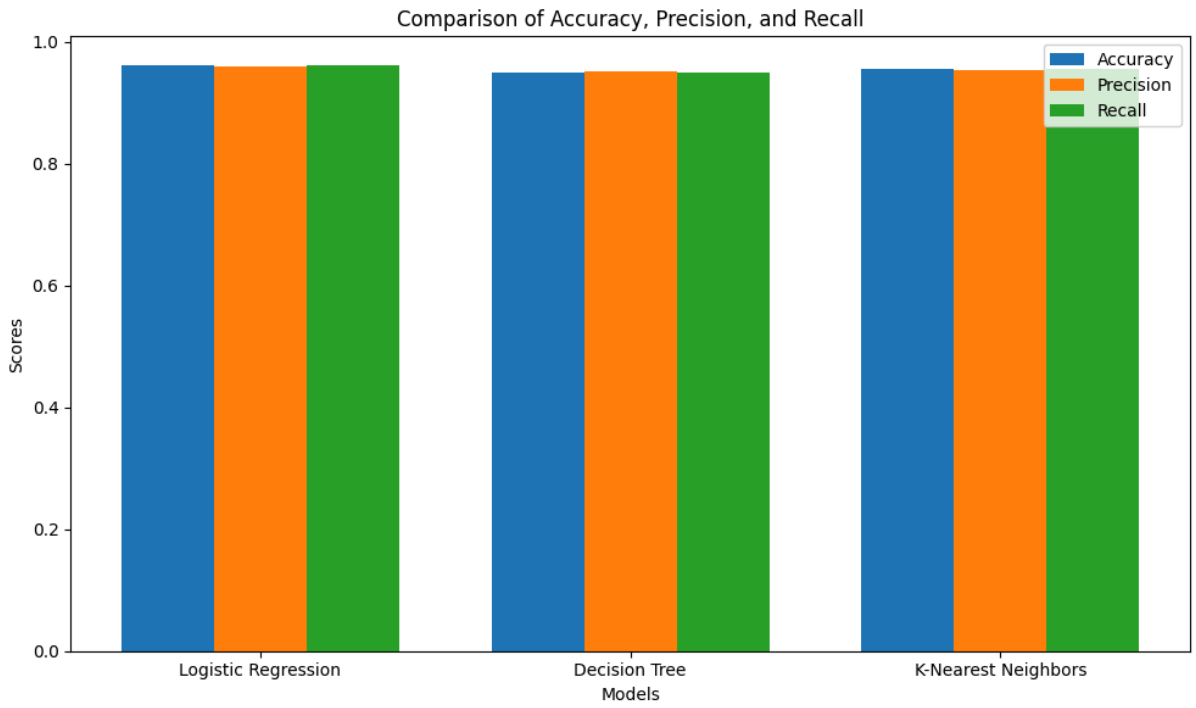

Fig. 6. Comparison between three models

8. Conclusion:

In conclusion, our analysis demonstrated that while all three models - Logistic Regression, Decision Tree, and K- Nearest Neighbors (KNN) - performed well, the Decision Tree model proved to be the most reliable for predicting diabetes. Although Logistic Regression achieved the highest accuracy and precision, the Decision Tree model provided the best balance, minimizing the risk of false negatives, which is crucial in a healthcare context where misdiagnosis could have severe consequences. With fewer false negatives, the Decision Tree ensures that more diabetic individuals are accurately identified, making it the most dependable choice for this task.