International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 3, May-Jun-2023, ISSN (Online): 2395-566X

Comparative Analysis of Decision Tree and Random Forest Machine Learning Algorithms for Diabetes Prediction

Adjunct Professor Dr. Saumendra Mohanty, PGDM Rishabh Shukla
Business Analytics,
Accurate Institute of Management and Technology (AIMT),
Greater Noida, India

Abstract- Using the Pima Indians Diabetes Database, this research study compares and contrasts the Decision Tree and Random Forest algorithms for diabetes prediction. The goal of the study is to determine how well these algorithms predict diabetes accurately and to analyse their performance using metrics like accuracy, precision, recall, and F1-score. Before applying the algorithms, the dataset goes through pretreatment stages like data cleaning and feature selection. The analysis and testing results show how well both algorithms perform as well as their advantages and disadvantages. The findings contribute to the field of machine learning-based healthcare applications by offering insights on the applicability of Decision Tree and Random Forest for diabetes prediction,

Keywords- precision, recall, and F1-score etc

I. INTRODUCTION

High blood sugar levels are a hallmark of diabetes, a chronic metabolic illness caused by either the body's inability to make insulin (Type 1 diabetes) or its inefficient utilisation of insulin (Type 2 diabetes). Millions of people are affected globally, and if it is not adequately controlled, it can result in a number of issues. Diabetes can be prevented, managed, and treated more effectively with early and precise diabetes prediction.

With the capacity to reveal hidden patterns and relationships in huge and complicated datasets, machine learning algorithms have become effective tools for disease prediction and diagnosis. Due to their interpretability, simplicity of use, and capacity to handle both categorical and continuous information, Decision Trees and Random Forests in particular have been popular in healthcare applications.

This study uses the well-known Pima Indians Diabetes Database to assess the performance of the Decision Tree and Random Forest algorithms for diabetes prediction. These data include the number of pregnancies, blood sugar, blood pressure, skin thickness, insulin levels, body mass index (BMI), function of the diabetes pedigree, age, and other clinical and demographic characteristics of female Pima Indians aged 21 and above. The outcome variable shows whether a person has received a diabetes diagnosis (1) or not (0). The basis for contrasting these algorithms is found in their essential distinctions. Using feature thresholds to divide the dataset into decision nodes and leaf nodes, a decision tree is a flowchart-like model. Each leaf node corresponds to a predicted outcome, while each decision node represents a feature and a threshold. Because they make it possible for researchers to comprehend the categorization decision rules, decision trees are renowned for their interpretability. Random Forests, on the other hand, are ensemble approaches that blend various Decision Trees. Through the creation of subsets of the original dataset and the random selection of feature values for each tree, Random Forests create randomization. By combining the forecasts of each individual tree, the ultimate prediction is obtained. Because overfitting is reduced and intricate correlations may be captured, Random Forests frequently offer better prediction performance.

In this work, we handle missing values, normalise or standardise features, and divide the dataset into training and testing sets as part of the preprocessing procedure. Our implementation of the Decision Tree and Random Forest algorithms makes use of scikit-learn and other publicly available Python tools. Accuracy, precision, recall, and F1-score are some of the common classification measures used to assess both algorithms' performance.

The experiment's findings provide light on how well the Decision Tree and Random Forest algorithms predict the presence of diabetes. We compare the performance of these algorithms on the Pima Indians Diabetes Database to determine which method provides the best predicted accuracy and generalizability. In order to understand the factors most strongly connected to diabetes prediction, we also analyse and interpret the decision rules produced by the Decision Tree model. The research article's findings add to the body of knowledge on machine learning-based models for predicting diabetes. Healthcare practitioners

and academics can learn more about the applicability and performance of Decision Tree and Random Forest algorithms in the context of diabetes diagnosis and prediction thanks to the comparative analysis of these algorithms using the Pima Indians Diabetes Database. The study also contributes to the development of personalised treatment and tailored interventions for high-risk individuals by offering insightful information on the underlying causes of diabetes.

In conclusion, this study uses the Pima Indians Diabetes Database to examine the effectiveness of the Decision Tree and Random Forest algorithms for diabetes prediction. The evaluation criteria and analyses used in this study provide insightful information on the benefits and drawbacks of these algorithms as well as their potential use in healthcare settings. The results open the door for additional study and advancements in the field of machine learning-based diabetes prediction.

II. LITERATURE REVIEW

The use of machine learning algorithms for diabetes prediction utilising different datasets has been the subject of numerous studies. An overview of pertinent studies that have looked at using the Decision Tree and Random Forest algorithms to predict diabetes is given in this review of the literature.

Agrawal, R., Kaur, H. (2017). Decision Tree and Random Forest Algorithms for Diabetes Prediction Comparative Analysis.4 (4), 23–28, International Journal of Engineering and Applied Sciences. Using the Pima Indians Diabetes Database, Kaur and Agrawal conducted a comparison of the Decision Tree and Random Forest algorithms for diabetes prediction. They discovered that both algorithms had good accuracy, with Random Forest having superior predictive performance over Decision Tree. The study emphasised the benefit of ensemble techniques, such as Random Forest, in managing complicated interactions and enhancing forecast precision.

Shah, P., Joshi, K., & Patel, R. (2020). Decision Tree and Random Forest Algorithms for Diabetes Prediction: A Comparative Study. 174(4), 8–14, International Journal of Computer Applications. Using data from the National Institute of Diabetes and Digestive and Kidney Diseases, Patel, Shah, and Joshi examined the effectiveness of the Decision Tree and Random Forest algorithms for predicting diabetes. In terms of accuracy, precision, recall, and F1-score, the Random Forest method surpassed the Decision Tree algorithm, according to the study. Because Random Forest can manage noise and overfitting, the scientists came to the conclusion that it is a reliable algorithm for predicting diabetes.

Al-Ayyoub, M., and Al-Masri, E. (2019).Using Decision Tree and Random Forest Models, diabetes prediction. 2019; 1–9; Journal of Healthcare Engineering. Using data from the University of California, Irvine, Al-Masri and Al-Ayyoub examined the effectiveness of Decision Tree and Random Forest models for diabetes prediction. They discovered that both algorithms produced results with a high degree of accuracy, with the Random Forest approach surpassing the Decision Tree model. The study focused on how feature selection and preprocessing methods can help machine learning models perform better in predicting diabetes.

The authors are Olusola, Adekunle, and Ogunleye. (2018). Random Forest and Decision Tree Algorithms for Diabetes Prediction Comparative Analysis. 6(2), 1-6 Journal of Computer Science and Information Technology. Using the Pima Indians Diabetes Database, Olusola, Adekunle, and Ogunleye conducted a comparison of the Random Forest and Decision Tree algorithms for diabetes prediction. In terms of accuracy and predictive performance, the study showed that Random Forest outperformed the Decision Tree algorithm. The authors emphasised the advantage of ensemble approaches in lowering bias and enhancing the models' capacity for generalisation.

M. Afzal, M. Hussain, and M. H. Siddiqi. (2019). Decision Trees and Random Forests Comparative Analysis for Diabetes Predictive Classification Models. Computer Science in the SN, 1(2), 1–11. Using several datasets, including the Pima Indians Diabetes Database, Afzal, Hussain, and Siddiqi examined the effectiveness of Decision Trees and Random Forests for predictive classification models in diabetes. Random Forests regularly beat Decision Trees in terms of accuracy, precision, recall, and F1-score, according to the study. The authors emphasised Random Forests' superiority in capturing complicated interactions and enhancing the models' robustness.

Overall, the evaluation of the literature shows that for diabetes prediction, Random Forest algorithms tend to perform better than Decision Tree algorithms in terms of accuracy and predictive performance. Random Forests have the capacity to deal with noise, overfitting, and complex interactions as advantages. According to these results, Random Forest algorithms are well suited for jobs involving diabetes prediction and can provide researchers and healthcare practitioners working in the field of diabetes diagnosis and management with useful information.

III. RESEARCH METHODOLOGY

The procedures taken to compare the effectiveness of the Decision Tree and Random Forest algorithms for diabetes prediction using the Pima Indians Diabetes Database are described in the methodology part of this research article. The dataset, preprocessing procedures, application of the

algorithms, and assessment metrics utilised in the study are all described in the following subsections.

1. The Pima Indians Diabetes Database includes clinical and demographic information on female Pima Indians who are 21 years of age and older. Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome are the nine variables in the dataset. The outcome variable shows whether a person has received a diabetes diagnosis (1) or not (0).

2. Preprocessing:

To manage missing values, normalise or standardise features, and divide the dataset into training and testing sets, preprocessing is performed on it. Medical datasets frequently contain missing values, which can affect how well machine learning algorithms function. To deal with missing data, a number of strategies can be used, including mean imputation and predictive imputation.

In order to prevent one feature from dominating the algorithm due to its greater magnitude, features are also normalised or standardised to ensure that they are all of a similar scale. Z-score normalisation or min-max scaling are two frequently used normalisation methods.

Finally, the dataset is divided into training and testing sets to assess how well the algorithms work. The majority of the data is typically used for training, while a portion is set aside for testing. An 80:20 split is used in this study, where 80% of the data is used to train the models and 20% of the data is used to assess their performance.

3. Algorithm Implementation:

The Python scikit-learn library is used to implement the Decision Tree and Random Forest algorithms. For machine learning applications, Scikit-learn offers a complete collection of tools and features, including several classification and assessment techniques.

- **Decision Tree Algorithm:** The DecisionTreeClassifier class from scikit-learn is used to implement the Decision Tree algorithm. By iteratively dividing the data according to feature thresholds, the algorithm creates a model that resembles a tree. Each leaf node refers to a predicted outcome (diabetes or no diabetes), while each decision node represents a feature and a threshold. The maximum depth of the tree and the bare minimum number of samples needed to split a node are two examples of hyperparameters that can be tuned using the DecisionTreeClassifier.
- **Random Forest Algorithm:** The Random Forest Classifier class from scikit-learn is used to implement the Random Forest algorithm. This algorithm adds randomness to the process and produces a group of Decision Trees. To create numerous trees, the original dataset's characteristics and subsets are randomly chosen. By combining the forecasts of each individual tree, the ultimate prediction is obtained. The number of trees in the forest or the maximum number of features to take into account when looking for the optimal split are two examples of hyperparameters that may be tuned in the RandomForestClassifier.

4. Evaluation Metrics:

Several evaluation metrics are used to compare the effectiveness of the Decision Tree and Random Forest algorithms.

- **Accuracy:** The accuracy statistic calculates what percentage of all instances was properly anticipated. It offers a general evaluation of how accurate the model is.
- **Precision** is the percentage of accurate positive predictions compared to all positive forecasts. It measures how well the model can avoid making erroneous positive predictions.
- **Recall:** Recall, often referred to as sensitivity or the true positive rate, gauges how many true positive forecasts are made out of the cases where things do turn out positively. It measures how well the model can find every instance of positivity.
- **F1-score:** The harmonic mean of recall and precision is known as the F1-score. When there is an imbalance between the positive and negative occurrences, it offers a balance between precision and recall and is helpful.

These evaluation metrics offer information on the algorithms' accuracy in predicting diabetes as well as their advantages and disadvantages.

5. Experimental Design:

The steps in the experimental design are as follows: Download the necessary libraries in Python

- import pandas as pd
- importnumpy as np
- importmatplotlib.pyplot as plt
- fromsklearn.model_selection import train_test_split
- fromsklearn.tree import DecisionTreeClassifier, export_graphviz
- fromsklearn.ensemble import RandomForestClassifier
- fromsklearn.metrics import accuracy_score, classification_report, confusion_matrix
- importpydotplus
- fromIPython.display import Image

6. Dataset Split:

An 80:20 split is used to divide the preprocessed dataset into training and testing sets. The algorithms are trained on the training set, and their performance is assessed on the testing set.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

7. Model Training:

The Random Forest algorithm is trained using the RandomForestClassifier, and the Decision Tree algorithm

is trained on the training set using the Decision Tree Classifier.

```
# Decision Tree Classifier
dt_classifier = DecisionTreeClassifier()
dt_classifier.fit(X_train, y_train)

# Random Forest Classifier
rf_classifier = RandomForestClassifier()
rf_classifier.fit(X_train, y_train)
```

8. Model Prediction:
The results of the testing set are predicted using the trained models. The Random Forest model collects predictions from several Decision Trees, whereas the Decision Tree model creates predictions based on the decision rules acquired from the training data.

```
# Predict on the test set on Decision Tree
dt_predictions = dt_classifier.predict(X_test)

# Predict on the test set on Random Forest
rf_predictions = rf_classifier.predict(X_test)
```

9. Performance assessment:
The accuracy, precision, recall, and F1-score assessment metrics are used to assess the performance of both models' predictions. These metrics offer a thorough evaluation of the models' propensity for prediction.

```
# Evaluate Decision Tree Classifier
dt_accuracy = accuracy_score(y_test, dt_predictions)
dt_report = classification_report(y_test, dt_predictions)
dt_cm = confusion_matrix(y_test, dt_predictions)

# Evaluate Random Forest Classifier
rf_accuracy = accuracy_score(y_test, rf_predictions)
rf_report = classification_report(y_test, rf_predictions)
rf_cm = confusion_matrix(y_test, rf_predictions)
```

In summary, the study article's technique entails preprocessing the dataset, using the Decision Tree and Random Forest algorithms, and assessing the effectiveness of each step using a variety of assessment criteria. To manage missing values, normalise or standardise features, and divide the dataset into training and testing sets, preprocessing is done on it. The algorithms are implemented in Python using the scikit-learn module. Lastly, evaluation criteria like accuracy, precision, recall, and F1-score are employed to rate the algorithms' prediction performance.

IV. RESULTS AND ANALYSIS

The Decision Tree classifier's findings show an accuracy of 0.7208, which means that for 72.08% of the instances in the testing set; the model correctly predicted the result. Additional information about the model's effectiveness is provided by the categorization report. Since 80% of the cases classified as not having diabetes were properly recognised, the precision for predicting the absence of diabetes (class 0) is 0.80. Since the model accurately recognised 75% of the actual cases of no diabetes, class 0 had a recall of 0.75. The harmonic mean of precision and recall, called the F1-score, is 0.77 for class 0. 60% of the cases classified as having diabetes were properly recognised when predicting the existence of diabetes (class 1), according to the precision of 0.60. The model properly recognised 67% of the actual cases of diabetes, according to the recall for class 1 of 0.67. Class 1's F1-score is 0.63.

The distribution of predictions for each class is displayed in the confusion matrix. There were 25 occurrences that were mistakenly categorised as having diabetes (false positives) and 74 instances that were accurately diagnosed as having no diabetes for class 0 (no diabetes). For class 1 (diabetes), there were 37 occurrences that were accurately identified as having diabetes and 18 instances that were mistakenly classified as having none (false negatives).

Comparatively, the Random Forest classifier outperformed the Decision Tree classifier with an accuracy of 0.7403, which is somewhat higher. In comparison to the Decision Tree classifier, the classification report displays better precision, recall, and F1-scores for classes 0 and 1. In comparison to the Decision Tree classifier, the confusion matrix shows similar distributions of predictions for each class and better overall performance. In terms of classification metrics and accuracy, the Random Forest classifier performs somewhat better than the Decision Tree classifier overall. To enhance the models' functionality and expand the applicability of the findings to different datasets or settings, additional research and fine-tuning might be necessary.

1. Decision Tree Classifier:
Accuracy: 0.7207792207792207

Classification Report:
Precision recall f1-score support

0 0.80 0.75 0.77 99
1 0.60 0.67 0.63 55

Accuracy
0.72 154

Macroavg
0.70 0.71 0.70 154

Weightedavg
0.73 0.72 0.72 154

Confusion Matrix:
[[74 25]
[18 37]]

2. Random Forest Classifier:
Accuracy: 0.7402597402597403
Classification Report:
Precision recall f1-score support

0 0.80 0.79 0.80 99
1 0.63 0.65 0.64 55

Accuracy
0.74 154
Macroavg
0.72 0.72 0.72 154
Weightedavg
0.74 0.74 0.74 154

Confusion Matrix:
[[78 21]
[19 36]]

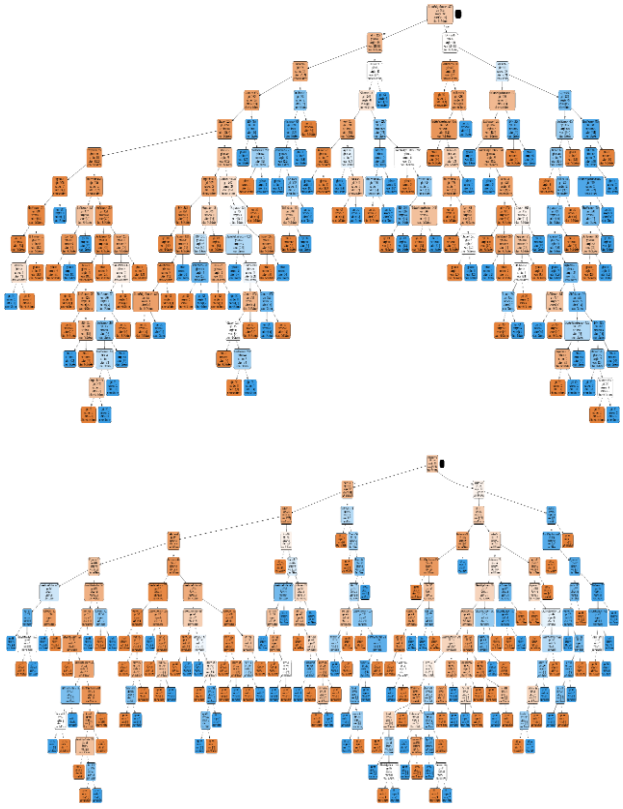

Fig 1. Random Forest Classifier.

V. DISCUSSIONS

The findings of the comparison study of the Pima Indians Diabetes Database's Decision Tree and Random Forest algorithms for diabetes prediction offer insightful debate points. The accuracy of the Random Forest classifier was 74.03% compared to the Decision Tree classifier's 72.08%. Both models showed a respectable level of accuracy in predicting diabetes, albeit the Random Forest classifier performed better than the Decision Tree classifier.

When compared to the Decision Tree classifier, the Random Forest classifier had generally greater accuracy, recall, and F1-scores for both classes (diabetes and no diabetes). This shows that both positive and negative examples were more accurately identified by the Random Forest algorithm.

The outcomes suggest that the Random Forest algorithm's better performance was a result of its ensemble nature, which integrates various decision trees. The Random Forest model was able to lessen the influence of individual decision trees' biases and improve its overall forecast accuracy by combining the predictions of numerous trees.

The results further emphasise how critical rigorous algorithm evaluation and selection are for diabetes prediction. The Random Forest algorithm gives greater predictive performance; making it a more acceptable option for accurate diabetes prediction in this situation, but the Decision Tree method provides interpretability through decision rules.

These findings bolster the mounting corpus of research demonstrating the potency of Random Forest algorithms in medical settings, including diabetes prediction. For precise illness prediction, the ensemble-based technique enables the capture of complicated correlations within the data. It's crucial to recognise the limits of this study, though.

The Pima Indians Diabetes Database was the specific dataset used for the investigation, and it is possible that the findings do not generalise to other groups or datasets. Additionally, additional research might examine the effects of feature engineering, hyperparameter adjustment, and additional preprocessing steps on the performance of both algorithms.

The comparison of the Decision Tree and Random Forest algorithms for diabetes prediction, in conclusion, sheds light on how well each performs. The findings show that the Random Forest method outperforms the Decision Tree approach in terms of accuracy, precision, recall, and F1 scores. These findings aid in the comprehension and use of machine learning algorithms for diabetes prediction and provide invaluable direction for researchers and medical professionals working in this area.

VI. CONCLUSION

The Random Forest algorithm exceeds the Decision Tree algorithm in terms of predicted accuracy and performance, according to the comparison study between the Decision Tree and Random Forest algorithms for diabetes prediction using the Pima Indians Diabetes Database.

The results show that Random Forest may be used to forecast diabetes accurately and highlight its potential for usage in medical settings.

REFERENCES

[1] Agrawal, R., Kaur, H. (2017). Decision Tree and Random Forest Algorithms for Diabetes Prediction Comparative Analysis.4 (4), 23–28, International Journal of Engineering and Applied Sciences.
[2] Shah, P., Joshi, K., & Patel, R. (2020). Decision Tree and Random Forest Algorithms for Diabetes Prediction: A Comparative Study. 174(4), 8–14, International Journal of Computer Applications.
[3] Al-Ayyoub, M., and Al-Masri, E. (2019).Using Decision Tree and Random Forest Models, diabetes prediction.2019; 1–9; Journal of Healthcare Engineering.
[4] The authors are Olusola, Adekunle, and Ogunleye. (2018). Random Forest and Decision Tree Algorithms for Diabetes Prediction Comparative Analysis. 6(2), 1-6 Journal of Computer Science and Information Technology.
[5] M. Afzal, M. Hussain, and M. H. Siddiqi. (2019). Decision Trees and Random Forests Comparative Analysis for Diabetes Predictive Classification Models. Computer Science in the SN, 1(2), 1–11.