Words Count: 1018

1.0 Introduction

This report presents a dynamic Shiny app designed to visualise and compare Decision Tree and Random Forest classifiers for predicting diabetes, a binary classification problem. The app lets users explore how different parameters influence model accuracy and classification boundaries. The Pima Indians Diabetes dataset, sourced from the “mlbench” package in R, was used for this task. The dataset was randomly sampled to include 400 observations with 8 predictive features (e.g., Glucose, Mass (BMI), Age), with the target variable indicating a diabetes diagnosis.

The objective is to predict whether a patient has diabetes based on medical diagnostic measurements. Accurate classification is crucial, as false positives (incorrectly diagnosing a healthy person) can lead to unnecessary anxiety and treatment, while false negatives (failing to detect diabetes) pose serious health risks. Both models are evaluated against different metrics and visually supported by a classification boundary grid, which illustrates how predictive features interact and contribute to classification decisions.

1.1 Data Preprocessing

The dataset was divided into 85% training and 15% test sets, ensuring the model had sufficient data to learn patterns while preserving a portion for evaluating generalisation performance. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. However, excessive over-sampling of the minority class introduced bias, requiring further control using the "themis" package.

To ensure robust evaluation, a 10-fold cross-validation was implemented, which enhances statistical significance while reducing variance and preventing overfitting. Both models were assessed using accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC).

2.0 Classification Model

2.1 Decision Tree

Decision Trees classify data by recursively splitting features based on importance, optimising criteria like Gini impurity, the default criteria of the caret package. They are interpretable and efficient but prone to overfitting, requiring hyperparameters such as maxdepth (controls tree growth) and cp (prunes unnecessary branches for generalisability). Due to high variance, they often underperform compared to ensemble methods like Random Forest, which aggregate multiple trees for stability. The bias-variance trade-off governs performance: high-bias, low-variance models underfit, missing essential patterns, while low-bias, high-variance models overfit, limiting generalisation in diabetes diagnosis.

2.2 Random Forest

Random Forest is an ensemble learning method that builds multiple Decision Trees and combines their predictions to improve accuracy and reduce overfitting. It is particularly effective in high-dimensional data, capturing feature interactions and reducing the impact of noisy variables. Unlike a single Decision Tree, it provides more stable predictions by averaging multiple models.

A key advantage is its built-in feature importance analysis, using:

- Mean Decrease in Accuracy – Measures performance drop when a feature is removed.
- Mean Decrease in Gini – Indicates a feature's contribution to node purity.

From the analysis, Glucose, BMI (Mass), and Age were the most influential predictors, with Glucose being the strongest, aligning with medical findings on blood sugar and diabetes risk (Valerie et al., 2004). In contrast, Insulin levels and Triceps skinfold thickness contributed minimally, suggesting lower discriminative power. These insights reinforce Random Forest's role in healthcare diagnostics, offering both predictive power and interpretability.

3.0 App Creation

3.1 User Interface (UI)

The Shiny app was designed with multiple interactive panels, including:

- Slider inputs to adjust model parameters and visualise changes in real time.
- Decision Tree Tab, displaying the model structure using fancyRpartPlot.
- Random Forest Feature Selection, allowing users to examine key predictors.
- Decision Boundaries, visualised using ggplot2, highlighting classification regions.

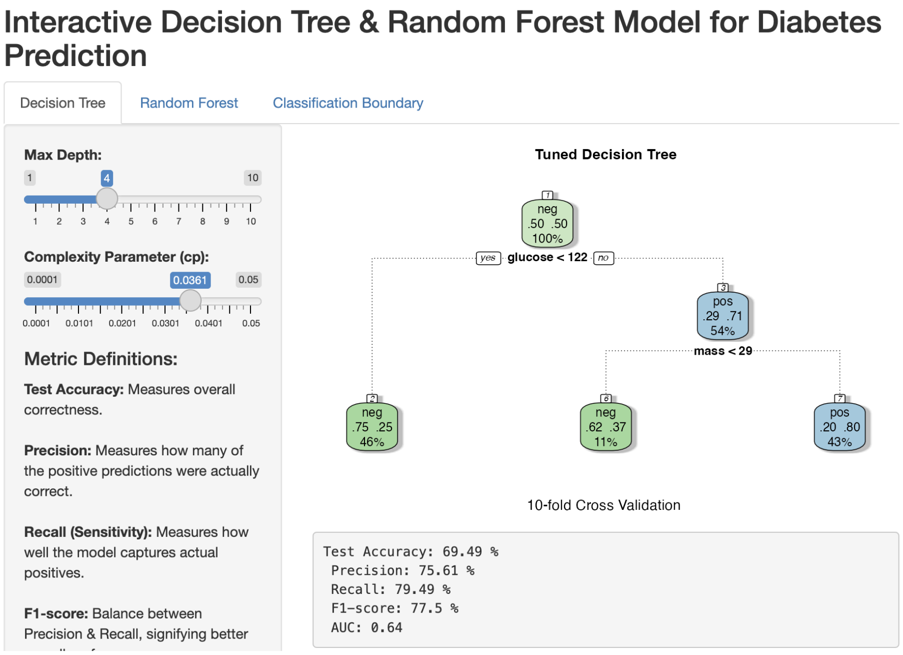
These elements enhance interpretability, enabling users to understand how predictive features contribute to classification.

3.2 Server Functionality

The server function dynamically updates models based on user input, providing real-time classification boundary visualisation and ensuring interactive exploration of model performance.

4.0 Results

4.1 Decision Trees



Interactive Decision Tree & Random Forest Model for Diabetes Prediction

Max Depth: 4 (1–10)

Complexity Parameter (cp): 0.0361 (0.0001–0.05)

Metric Definitions:

Test Accuracy: Measures overall correctness.

Precision: Measures how many of the positive predictions were actually correct.

Recall (Sensitivity): Measures how well the model captures actual positives.

F1-score: Balance between Precision & Recall, signifying better

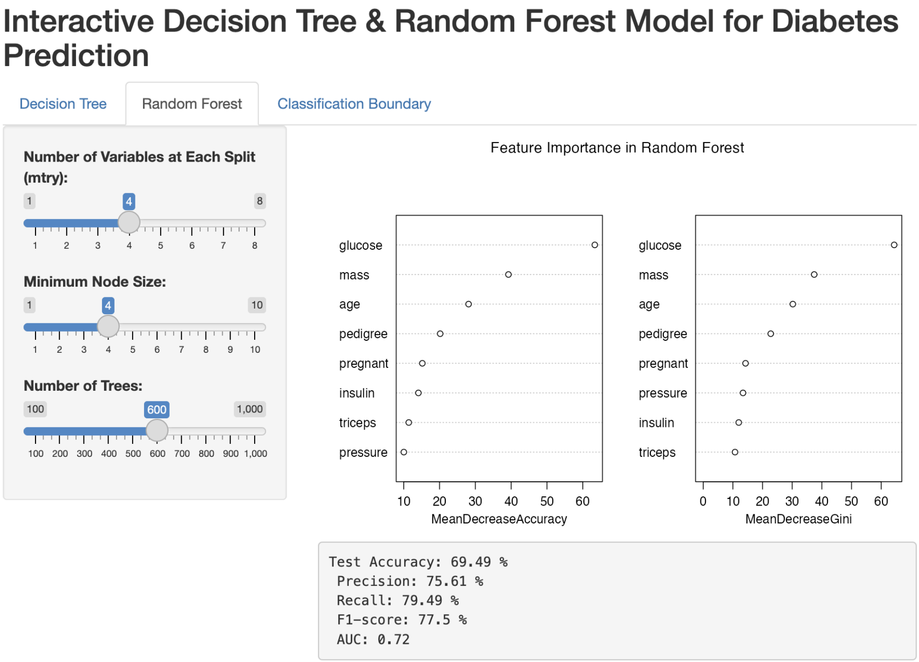
10-fold Cross Validation

```
Test Accuracy: 69.49 %
 Precision: 75.61 %
 Recall: 79.49 %
 F1-score: 77.5 %
 AUC: 0.64
```

Figure 1.0

The Decision Tree model follows a hierarchical structure, with glucose < 122 as the primary decision point, reflecting its strong predictive power. Further splits refine classification using BMI, improving differentiation between diabetic and non-diabetic patients. The model identifies 79.49% of actual diabetic cases, reducing missed diagnoses. However, false negatives remain a concern, as misclassifying diabetic patients may lead to delayed treatment and severe health complications.

4.2 Random Forest



Interactive Decision Tree & Random Forest Model for Diabetes Prediction

Number of Variables at Each Split (mtry): 4

Minimum Node Size: 4

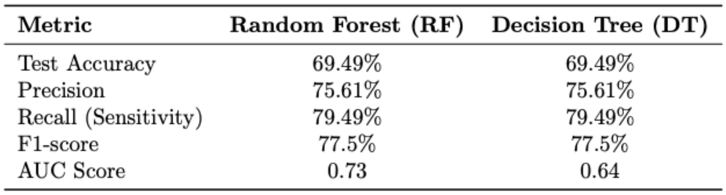
Number of Trees: 600

```
Test Accuracy: 69.49 %
 Precision: 75.61 %
 Recall: 79.49 %
 F1-score: 77.5 %
 AUC: 0.72
```

Figure 2.0

The Random Forest model leverages an ensemble of decision trees to enhance predictive stability and accuracy. It evaluates multiple decision pathways, reducing variance and mitigating the risk of overfitting. Feature importance analysis highlights glucose, BMI (mass), and age as the most influential predictors in determining diabetes, reinforcing well-established medical correlations.

The model correctly classifies 69.49% of cases, with a precision of 75.61%, meaning that when it predicts diabetes, it is correct in about three out of four cases. The recall rate of 79.49% suggests that the model successfully identifies nearly four out of five actual diabetic cases, reducing the likelihood of missed diagnoses.



Metric	Random Forest (RF)	Decision Tree (DT)
Test Accuracy	69.49%	69.49%
Precision	75.61%	75.61%
Recall (Sensitivity)	79.49%	79.49%
F1-score	77.5%	77.5%
AUC Score	0.73	0.64

Table 1: Comparison of Random Forest and Decision Tree Models

Figure 3.0

The higher AUC score of 0.73 for Random Forest compared to 0.64 for Decision Tree indicates that Random Forest is better at distinguishing between diabetic and non-diabetic cases across different classification thresholds. A higher AUC means the model ranks positive cases more effectively, reducing the likelihood of both false positives and false negatives. This makes Random Forest the preferred choice for automated diagnosis, as it provides more reliable and stable predictions.

However, despite its lower AUC, Decision Trees remain valuable for clinical decision-making due to their interpretability. They offer a clear, step-by-step logic for diagnosis, allowing doctors to understand and justify predictions. In contrast, Random Forest operates as a "black box" model, making it less transparent but more robust in prediction accuracy.

5.0 Conclusion

The results indicate that Random Forest is the preferred model due to its higher AUC score, making it more effective at distinguishing between classes. However, Decision Trees remain valuable for interpretability, offering a clear diagnostic logic for clinical decision-making.

Both models could be improved with a larger dataset, additional hyperparameter tuning, and additional predictive features. For higher classification accuracy, Extreme Gradient Boosting (XGBoost) presents a promising alternative, as it builds trees sequentially, reducing bias more effectively than Random Forest. However, given its computational complexity, XGBoost is better suited for larger datasets and was therefore not used in this analysis.

6.0 References

Valeri, C., Pozzilli, P. and Leslie, D., 2004. Glucose control in diabetes. Diabetes/metabolism research and reviews, 20(S2), pp.S1-S8.

7.0 Appendix

7.1 Ethical Considerations in Automated Medical Diagnosis

While machine learning improves diagnostic accuracy, bias in datasets may result in prediction disparities across different demographics. Over-reliance on automated predictions may also reduce human oversight, potentially leading to misdiagnoses. Ensuring fairness, transparency, and interpretability is crucial when deploying these models in real-world healthcare applications.