7.0 Analysis and Recommendations

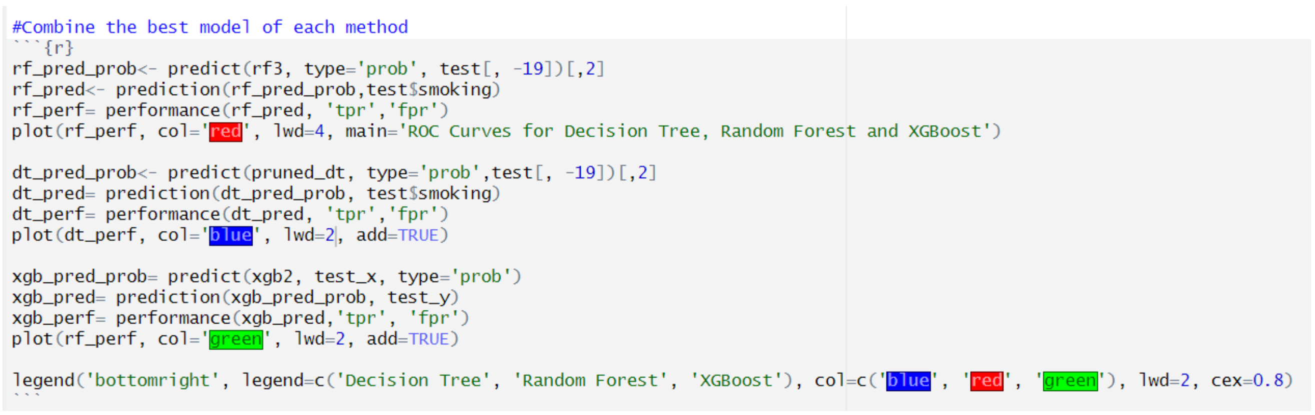
In this study, three tree-based model Decision Tree, Random Forest and XGBoost were built, trained, tuned, and tested on training and test sets respectively. Both Random Forest and XGBoost models are ensemble learning methods. Random Forest works on the concept of bagging while XGBoost works on the concept of boosting.

Specifically, Random Forest uses subsampling with replacement to create training samples, a group of models is trained on these subsets of samples in parallel. Then, the final output is generated from taking the average of the output of all models. This method is used to reduce variance on a noisy dataset. On the other hand, XGBoost combines a set of weak models into a strong one to minimize error. In this process, each model is trained sequentially and try to compensate for the weakness of the previous model. In turn, this will increase the model performance. Hence, it is no doubt that the performance of Random Forest and XGBoost is better than Decision Tree alone.

Model	Hyperparameter	Data	Accuracy	AUC	Sensitivity	Specificity
Decision Tree	-tuned best cp=0.000664	**Training**	75.76%	0.825	76.86%	73.86%
		Test	74.55%	0.814	76%	72.04%
Random Forest	-ntree=2000, mtry=3	**Training**	100%	1	100%	100%
		Test	82.55%	0.911	84.90%	78.51%
XGBoost	nround=1000, max_depth=12, eta=0.001, gamma=3, colsample_bytree=0.5, min_child_weight=5, subsample=0.5	**Training**	83.61%	0.912	84.05%	82.85%
		Test	77.31%	0.855	79.26%	73.96%

Table 1 Decision Tree, Random Forest, XGBoost Model Performance Comparison.

In the summary table below, each model represents the best model selected in each machine learning method. In general, the model performance on training data is slightly higher than the test set and no overfitting or underfitting issues were observed in these models. Based on the results on test data, Random Forest yields the best results, Accuracy: 82.55%, AUC: 0.911, Sensitivity: 84.9% and Specificity: 78.51%.



~~~
#Combine the best model of each method
```{r}
rf_pred_prob<- predict(rf3, type='prob', test[, -19])[,2]
rf_pred<- prediction(rf_pred_prob,test$smoking)
rf_perf= performance(rf_pred, 'tpr','fpr')
plot(rf_perf, col='red', lwd=4, main='ROC Curves for Decision Tree, Random Forest and XGBoost')

dt_pred_prob<- predict(pruned_dt, type='prob',test[, -19])[,2]
dt_pred= prediction(dt_pred_prob, test$smoking)
dt_perf= performance(dt_pred, 'tpr','fpr')
plot(dt_perf, col='blue', lwd=2, add=TRUE)

xgb_pred_prob= predict(xgb2, test_x, type='prob')
xgb_pred= prediction(xgb_pred_prob, test_y)
xgb_perf= performance(xgb_pred,'tpr', 'fpr')
plot(rf_perf, col='green', lwd=2, add=TRUE)

legend('bottomright', legend=c('Decision Tree', 'Random Forest', 'XGBoost'), col=c('blue', 'red', 'green'), lwd=2, cex=0.8)
```
~~~

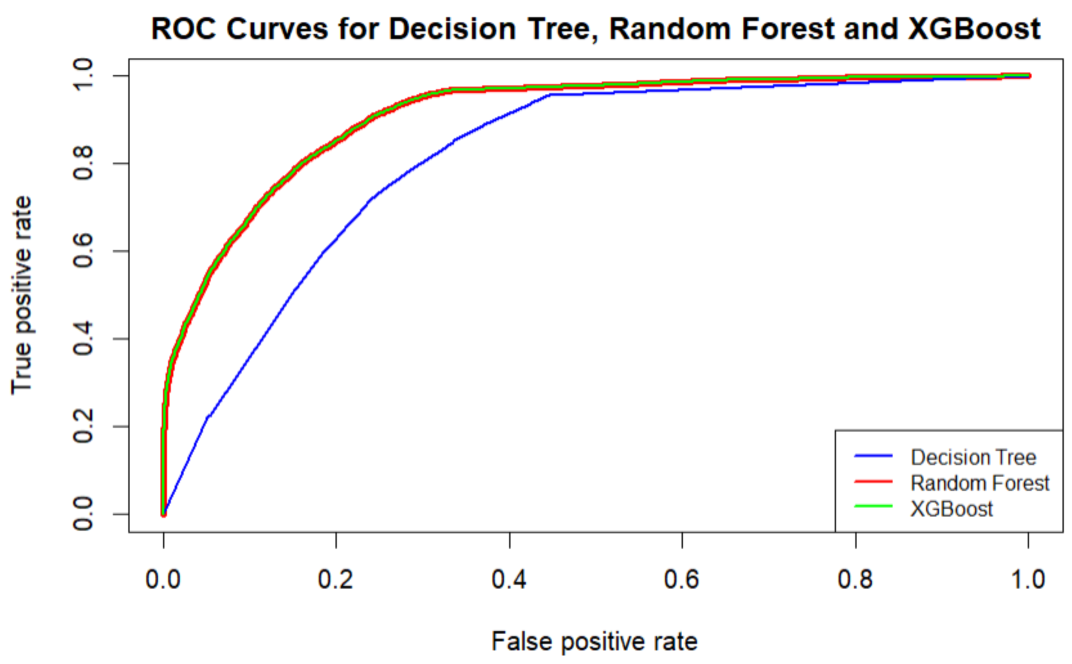


Figure 1 ROC Curve of Decision Tree, Random Forest and XGBoost Models.

From the ROC curve, it is found that Random Forest and XGBoost have a better performance than Decision Tree.

In comparison to the results of three relevant work, which used the same body of signal smoking dataset obtained from Kaggle, it is found that the best performing model in these studies is also Random Forest. Among all the studies, Mustanger (2022) model shows the best performance, with accuracy 84.2%, which is slightly higher than the performance in this study (82.55%), whereas the Al-btoush (2022) and Titas (2022) showed similar accuracy as the current model performance, which is 82.6% and 82.56% respectively. However, in these studies, besides accuracy, no other metrics were analysed. Hence, the model performance from other perspectives is lacking. For example, the model fitness is not known.

Work	**Best Model performance**
Mustanger (2022)	Random Forest Accuracy: 84.2%
Titas (2022)	Random Forest accuracy:82.56%
Al-btoush (2022)	Random Forest accuracy 82.6%
Current Study	Random Forest Accuracy: 82.55% AUC: 0.911 Specificity: 78.51% Sensitivity: 84.90%

Table 2 Model Comparison with Relevant Work using Same Dataset