Decision Trees and Random Forest Assignment#5

Submitted to - Professor Sam Plati

Submitted by - Syeda Jaffer

Student ID- 100944094

Rational Statement

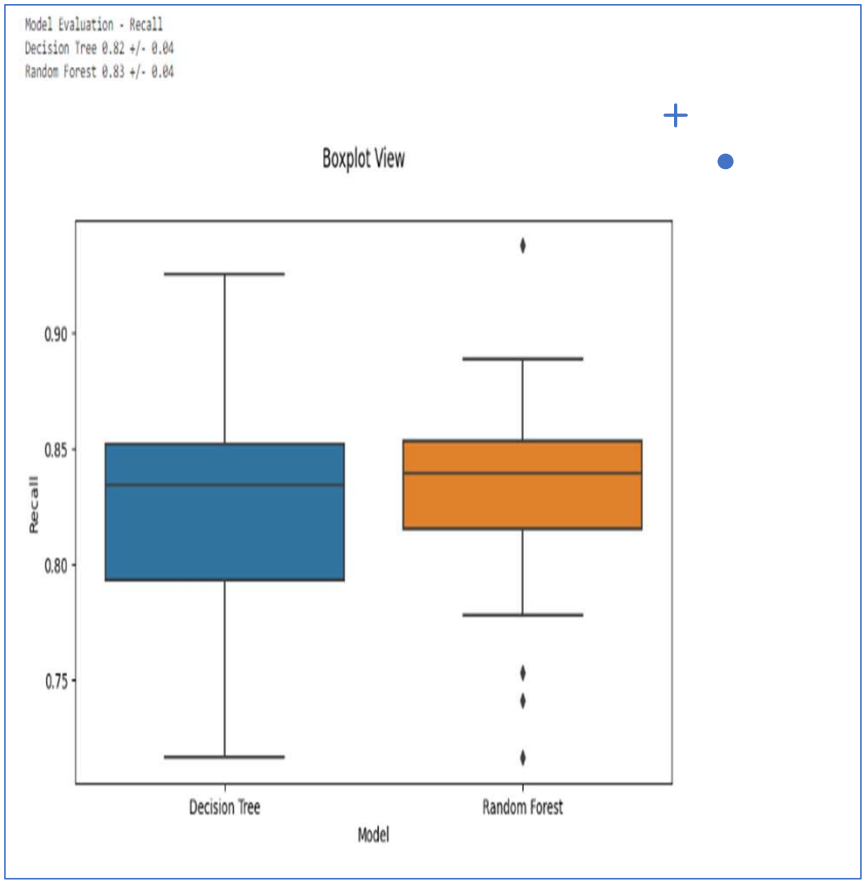
John Hughes is interested in developing an optimized decision tree and random forest model. These predictive models hold the potential to evaluate risk levels during pregnancy, allowing for early intervention and customized care tailored to the patient's specific needs. This, in turn, could potentially reduce the incidence of complications. The evaluation results of these models could also help manage and customize pregnancy care according to the patient's specific needs, potentially reducing the occurrence of complications.

Model Analysis Based on Box Plots

The boxplot of the Random Forest model are longer, indicating a wider range of recall values with more variability compared to the Decision Tree model. This suggests that although the Random Forest may generally perform better, its performance is less consistent.

The Random Forest model is more effective in terms of recall with a median recall score of about **0.83**. But it has a wider interquartile range (IQR) and outliers, indicating higher variability and less consistency compared to the Decision Tree model..

The Decision Tree algorithm exhibits a recall metric with a relatively narrow interquartile range (IQR) centered around a median value of **0.82**, indicating a more consistent and predictable performance. However, the presence of outliers in the data suggests that there are occasional instances of significant deviations from the expected performance levels.

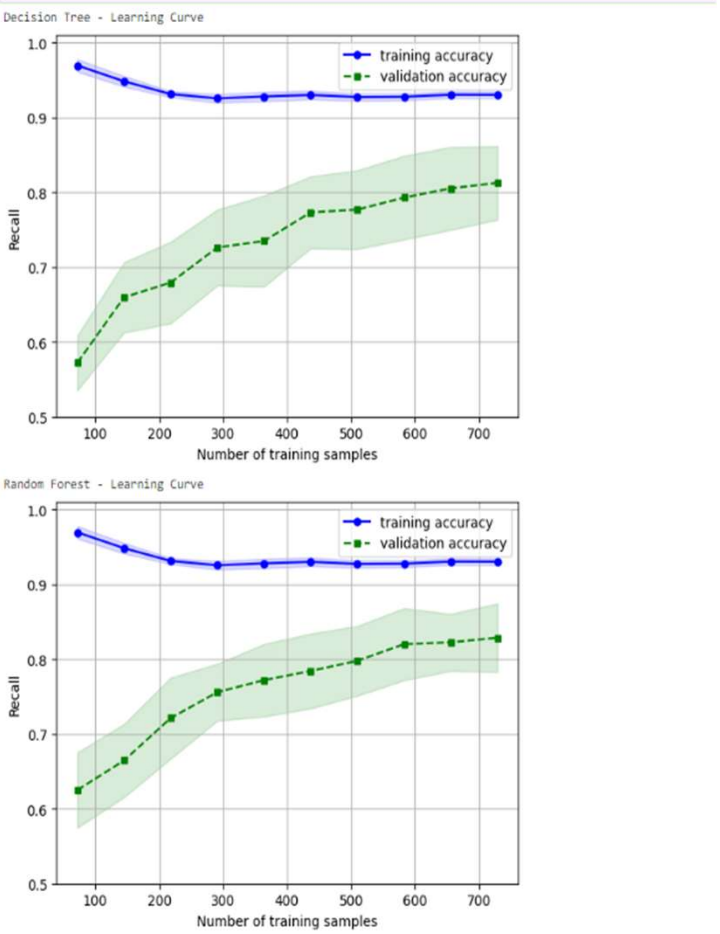


Model Analysis Based on Learning Curves

Based on the learning curves, it can be observed that the Random Forest model performs almost perfectly on the training set, but the validation score is comparatively lower and exhibits more variance, which may imply overfitting.

The Random Forest model shows good performance on training data with consistent high accuracy. However, there is more variability in accuracy when tested on validation data which could indicate overfitting to the noise in the training data instead of underlying patterns.

As the number of training samples increases, the Decision Tree model demonstrates a consistent rise in validation accuracy, which is a positive indicator of its ability to generalize well. Furthermore, the difference between training and validation accuracy is smaller in comparison to the Random Forest model, indicating that the Decision Tree model is less prone to overfitting.



Classification Report

Random Forest Model

1. The model exhibits high precision With a precision of **0.92**, particularly for the "High risk" category. This means that when it predicts a patient as "High risk," it's highly likely to be correct.

2. The recall for the "High risk" category is **0.89** which also indicates a strong ability to correctly identify a significant number of genuine high-risk cases.

3. The F1-score for the "High risk" category is **0.91** further underscores the tool's effectiveness, striking a fine balance between precision and recall - a crucial feature for any medical diagnostic tool.

A good balance of precision and recall is critical in a medical diagnostic context.

Decision Tree Model

1. While the model shows good precision **0.87**, it is slightly lower compared to the Random Forest model. This could result in more false-positive results.

2. The recall for the "High risk" category is **0.85** which is also lower than that of the Random Forest model, indicating that it might miss some high-risk cases.

3. The F1-score for "High risk" is **0.86** lower than the Random Forest model, which suggests that it may not balance precision and recall as effectively.

.

Based on the support values of the Decision Tree model, the dataset is reasonably balanced. There are 55 instances of "High risk", 81 instances of "Medium risk", and 67 instances of "Low risk". This is great news for model training, as it indicates that each class had a sufficient number of instances for the model to learn from.

Feature Importance Insights

Random Forest Model

1. Systolic BP (0.18) and Diastolic BP (0.14) have significant importance, which aligns with medical knowledge that blood pressure is a critical factor during pregnancy.
2. The Blood Sugar (BS) (0.35) feature stands out as the most crucial factor, highlighting the significance of monitoring blood glucose levels during pregnancy.

Decision Tree Model

1. Glucose levels play a critical role in assessing risk, as indicated by the high importance score (0.43) in the Decision Tree Model.

2. Age (0.16) and Systolic BP (0.19) are also significant factors with similar importance, suggesting that both patient age and systolic blood pressure are almost equally important in risk assessment.

Recommendations

1. **Model stability**

Given the variability in the recall of the Random Forest model, as shown by the boxplot, it may be beneficial to explore strategies to stabilize the model's performance. This could be achieved through further parameter tuning or additional training data.

2. **Model Evaluation**

It is important to continuously assess the model's performance using new patient data as it becomes available. It should be used to retrain and update the models to maintain their accuracy and reliability. We can also explore ensemble methods that can combine the strengths of both Decision Trees and Random Forests to improve overall reliability.

3. **Over Fitting Consideration**

The learning curves show that the Random Forest model's training accuracy is very high compared to its validation accuracy, which is a classic sign of overfitting to mitigate this technique like pruning or adding more training data to ensure better generalization to new data.

By following these recommendations, Mr. John Hughes can work towards improving the predictive performance and reliability of the models, ensuring they provide accurate risk assessments for pregnant women.