Parkinson's Disease Classification

Deepesh Sahu(2022148) Anikait Agarwal(2022072)

I. Motivation

Classification models for Parkinson's Disease (PD) are motivated by the need for early and accurate diagnosis, which can lead to better management and potentially slower disease progression. These machine-learning approaches offer objective assessments, reducing subjective variability in clinical evaluations. Additionally, predictive models can assist in tailoring personalized treatment plans and interventions, improving patient outcomes and quality of life. Furthermore, analyzing large datasets using machine learning can aid in identifying biomarkers and risk factors, contributing valuable insights for research and drug development.

II. Dataset

We have collected our dataset from a renowned paper: "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection,
Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM.
BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)"

III. Proposed Processing

We conducted certain technique to see were does our model predictor will fit the best.

1. LDA: We divided our dataset in two different ways first fixed splitting in which initial 80% is kept for training and 20% is left for testing and second split is random but training and test proportion is the same. To remove any redundancy and inaccuracy, we try to use inbuild function and methods to create our model. We plotted confusion matrix graph , actual vs predicted graph for reaching our conclusion

2. QDA:We divided the dataset in fixed split only because of some unexpected result we didn't perform the second split which is random. Similar to LDA, we plotted confusion matrix graph and actual vs predicted graph for analysis.

3. Decision Tree:We tried to follow the same mechanism as in LDA.Nothing new has been done except of the function call name

4. ADABoost: We applied AdaBoost, a popular ensemble learning technique, to our dataset. First, we divided the data into training, validation, and test sets. The training set comprised 60% of the data, while 20% each was allocated to the validation and test sets.

We used a decision tree as the base estimator for AdaBoost. Initially, we trained the AdaBoost classifier with a varying number of trees (estimators) ranging from 1 to 300. For each iteration, we evaluated the accuracy of the classifier on the validation set. By monitoring the accuracy on the validation set, we selected the optimal number of trees that yielded the highest accuracy. We then trained a new AdaBoost classifier using this optimal number of trees and evaluated its performance on the test set. This process allowed us to assess the generalization performance of the AdaBoost model on unseen data. Finally, we reported the test accuracy achieved by the best-performing AdaBoost model.

5. XBOOST: Our approach involved leveraging the powerful XGBoost algorithm for classification. We meticulously divided the dataset into training and test sets, utilizing various evaluation metrics to assess model performance. By conducting validation set analysis, we determined the optimal number of trees (estimators) for the XGBoost classifier, ensuring the highest accuracy on unseen data. Subsequently, we trained a new XGBoost classifier with this optimal configuration and evaluated its performance on the test set. This comprehensive analysis allowed us to ascertain the effectiveness of the XGBoost model in accurately predicting outcomes

IV. Results

1. LDA:
   Using Fixed Splitting : 51.28%
   Using Randomized Splitting: 87.18%
   This is a Strong indication that the data might have a collinear relationship.

2. QDA:
   Using Fixed Splitting : 38.46%
   With this running we get that our data is collinear so doing this with randomized splitting might generate the same model as LDA which is of no use

3. Decision Tree:
   Using Fixed Splitting : 51.28%
   Using Randomized Splitting : 89.74%

4. AdaBoost
   Using Randomized Splitting: 92.31%
   These results suggest that AdaBoost demonstrates promising performance. Further analysis is warranted to determine the optimal configuration and assess its robustness against overfitting.

5. XGBoost:

   Using Randomized Splitting: 93.51%

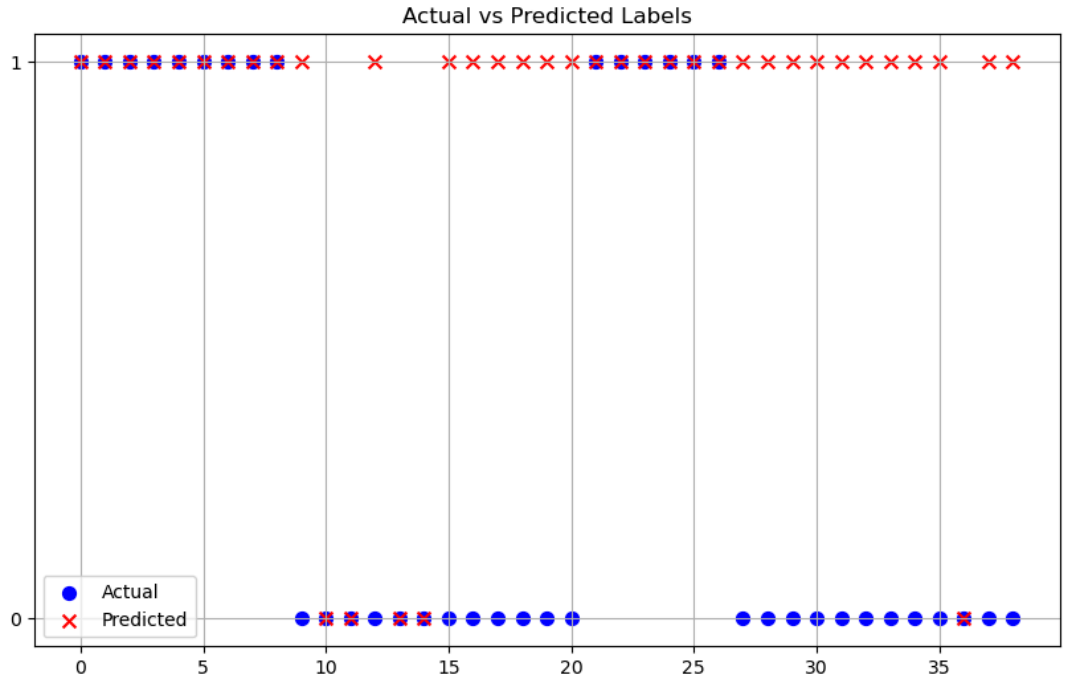
   The performance of XGBoost indicates its capability to handle complex relationships within the data.

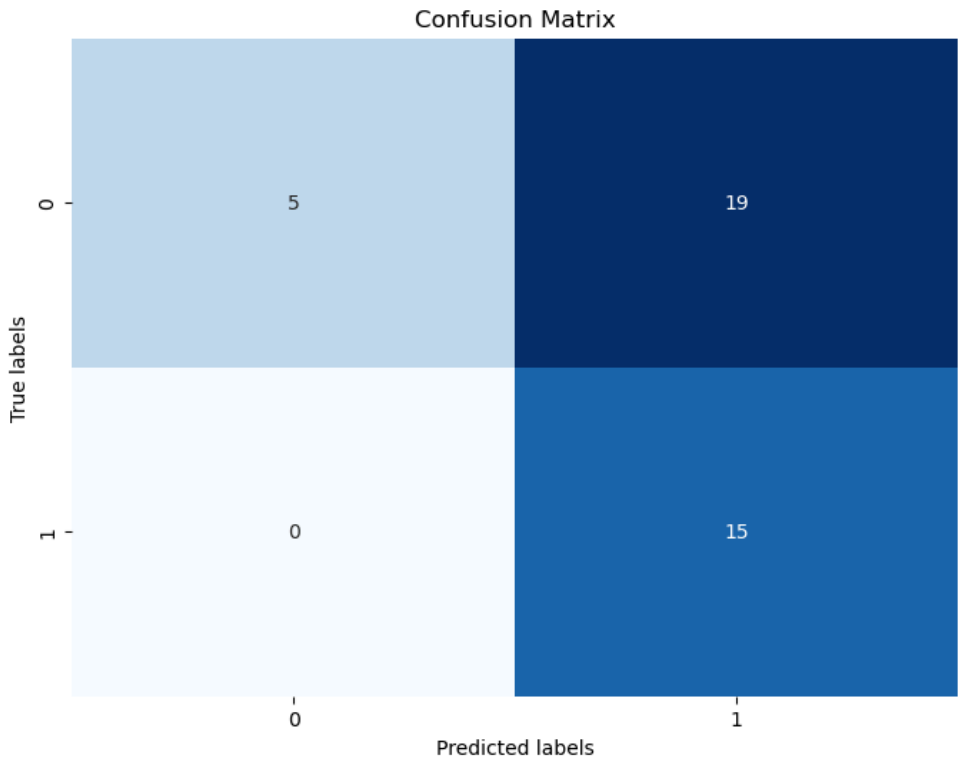
   Considering all methods and evaluation metrics, we conclude that XGBoost will have a higher chance of giving accurate results than any other; thus, the final model for Parkinson's disease is XGBoost

V. Graph Analysis

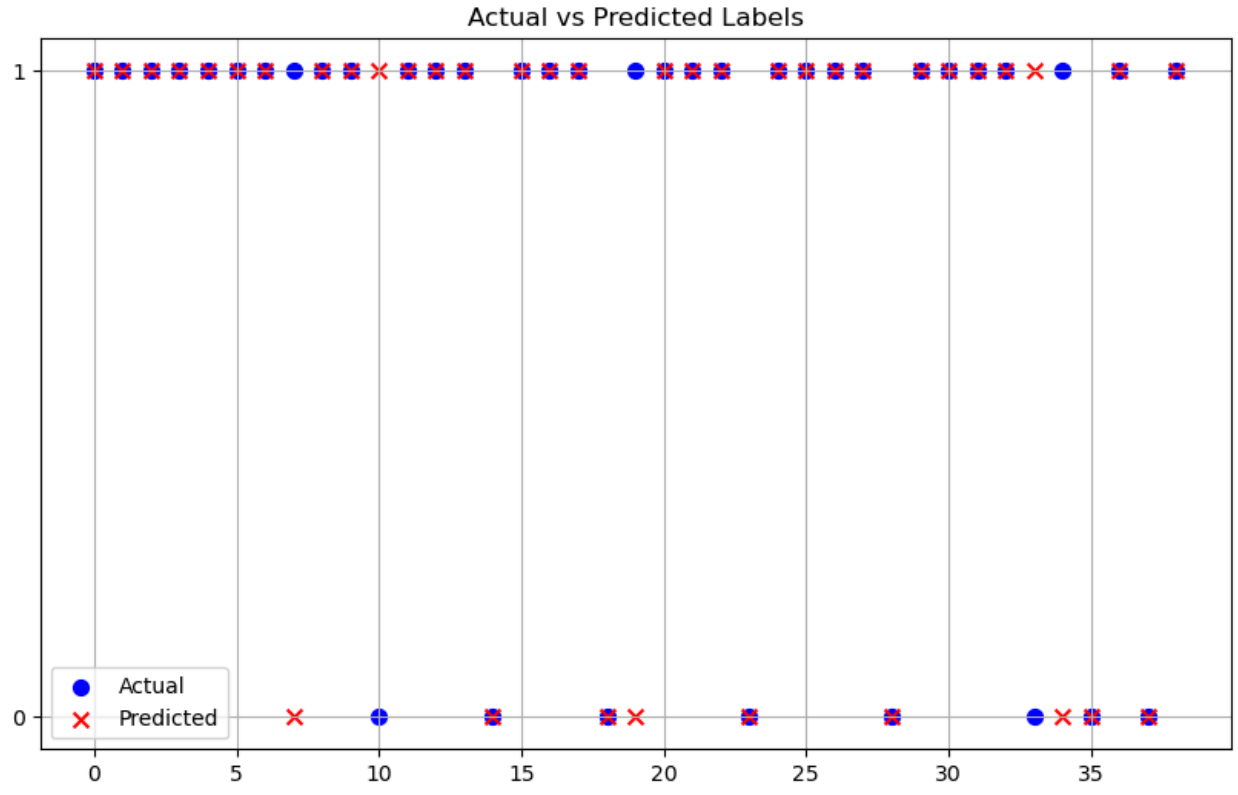
LDA:

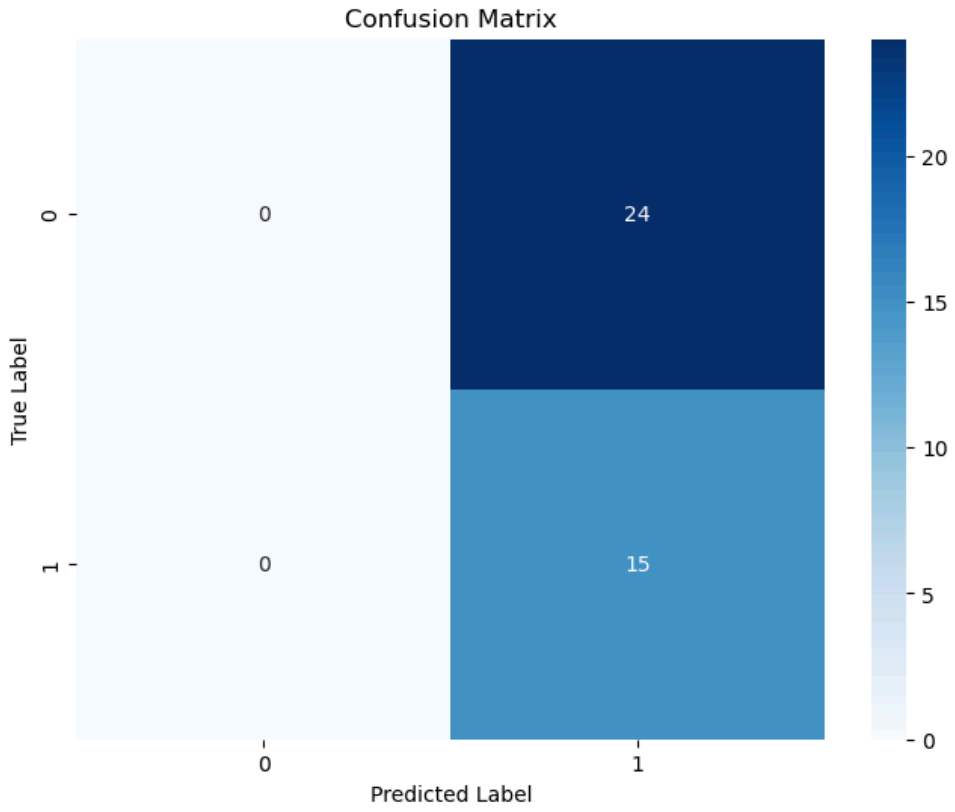
Randomised dataset:



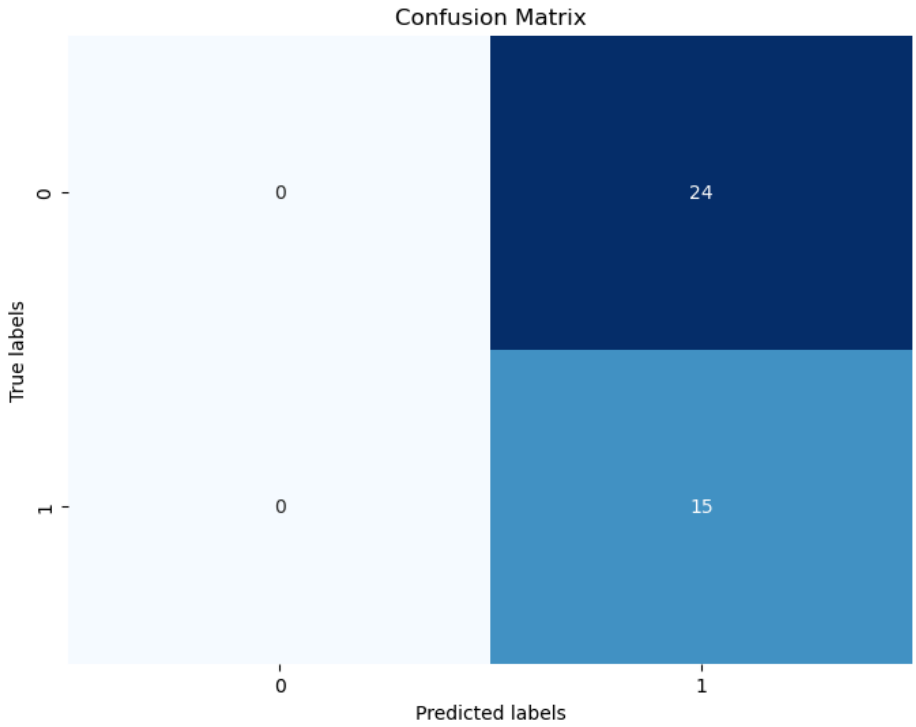

Confusion Matrix
5
19
0
15
True labels
0
1
0
1
Predicted labels

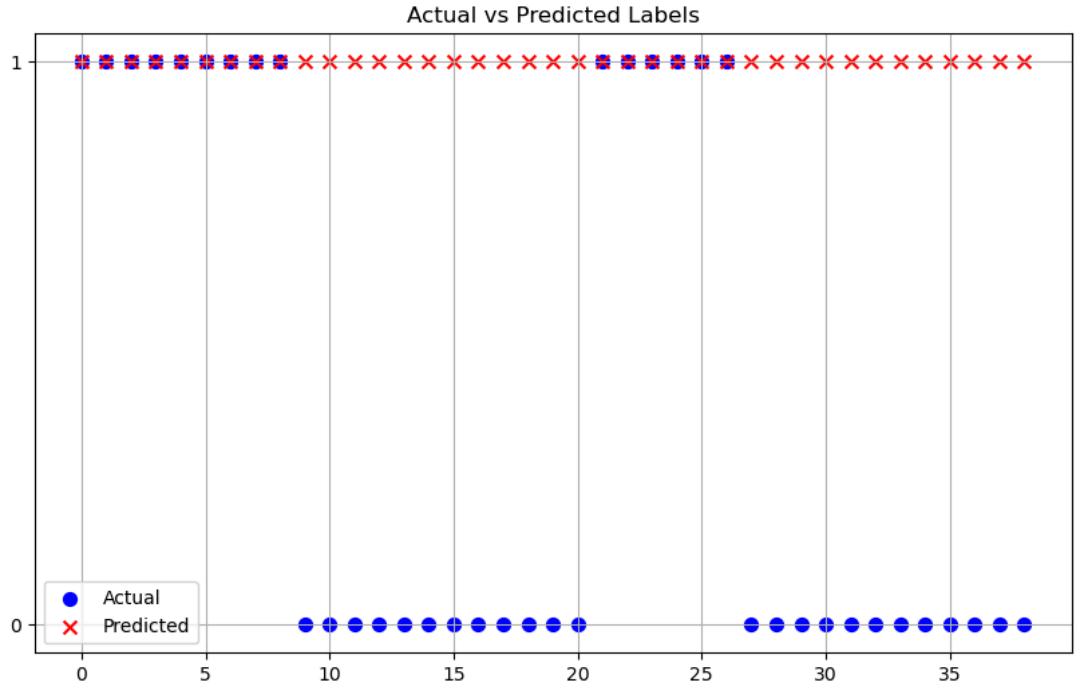
Fixed Split:


Actual vs Predicted Labels
1
0
Actual
Predicted
0
5
10
15
20
25
30
35


Confusion Matrix
True Label
Predicted Label
0
1
0
24
0
15
0
1
20
15
10
5
0

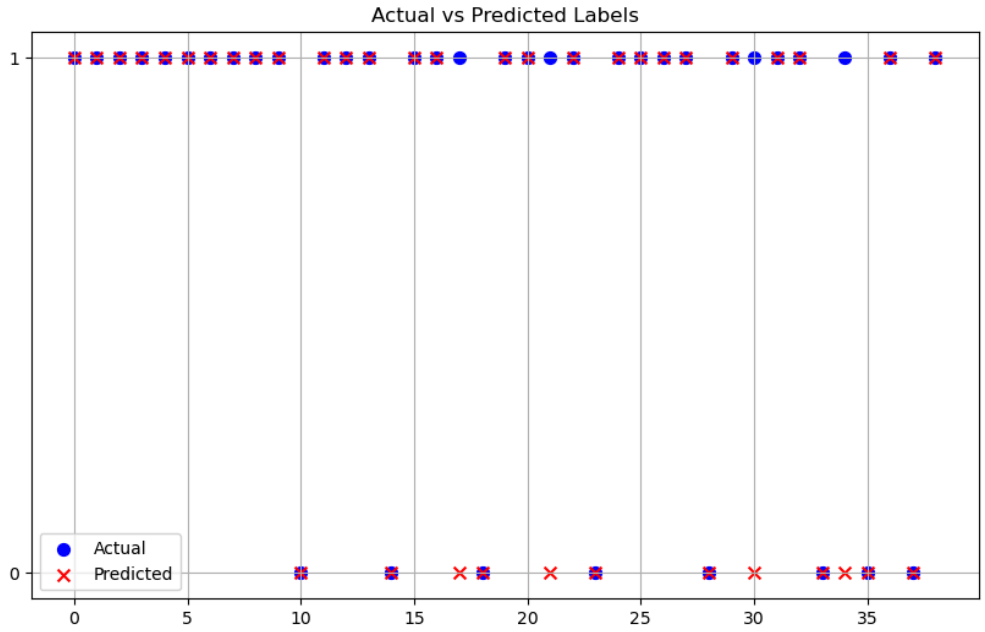
QDA:

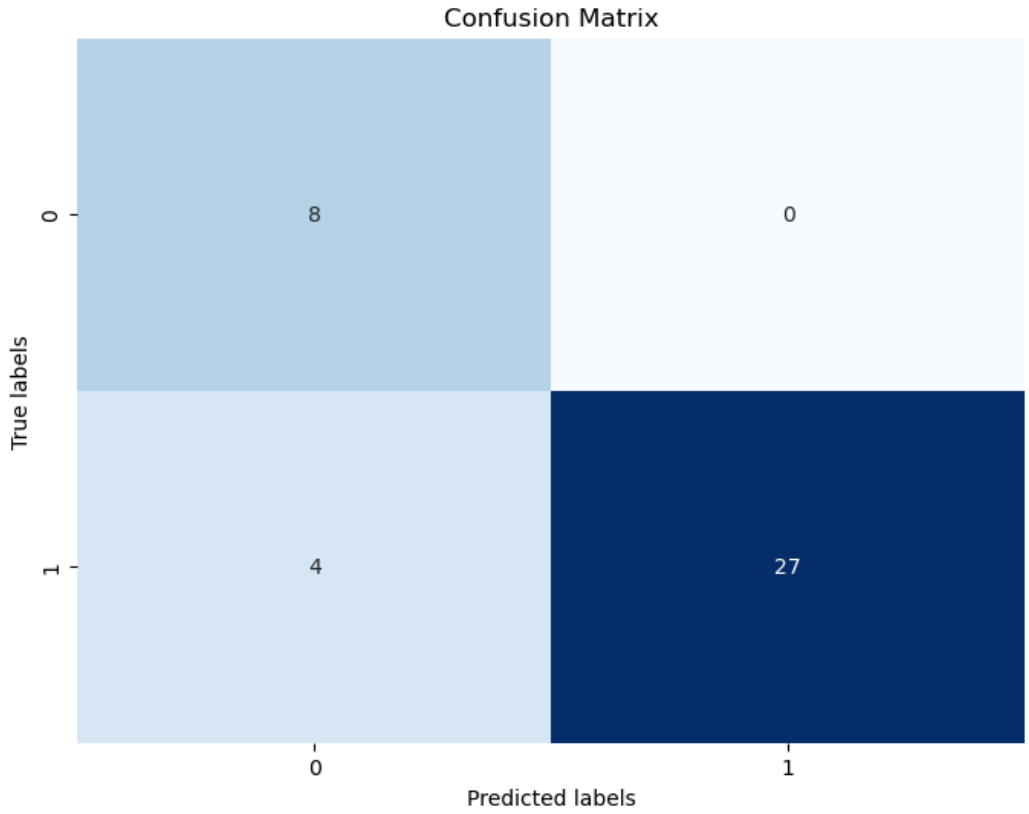




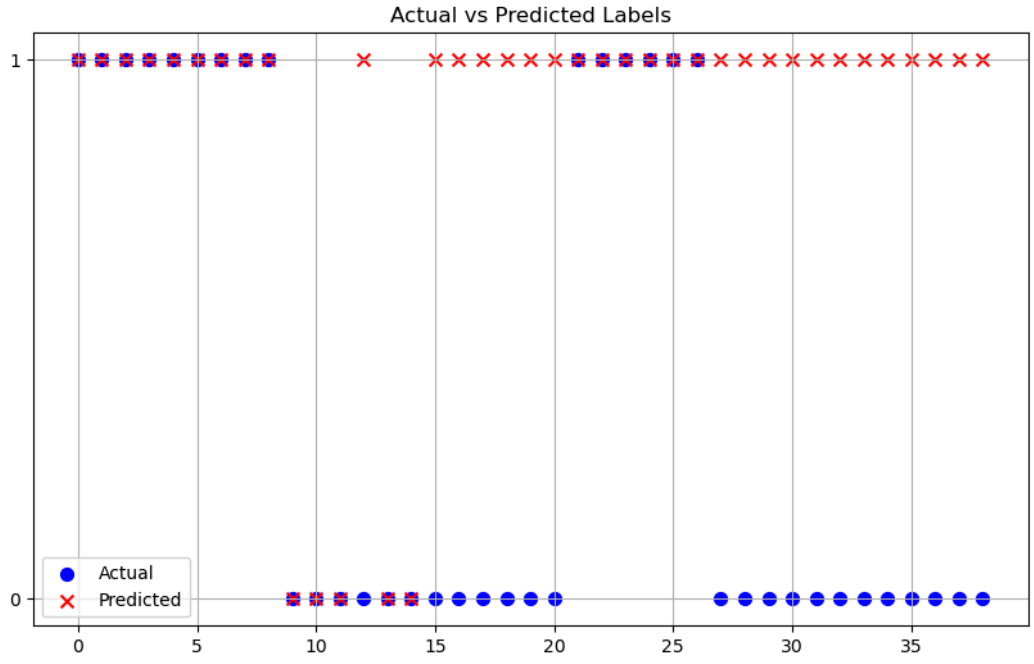
As the data is collinear, it doesn't make sense to continue with QDA analysis as in the best case it will give the same result as LDA. thus dropping the QDA analysis from here

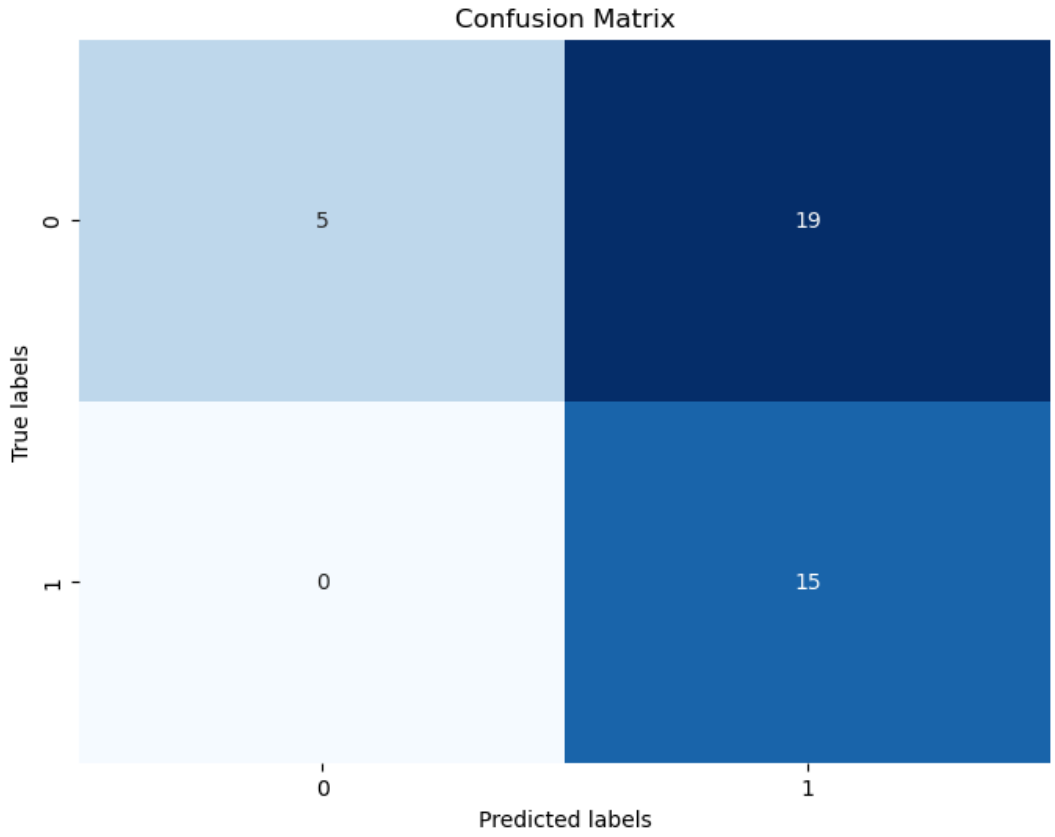
Decision Tree:
Using Randomized dataset:



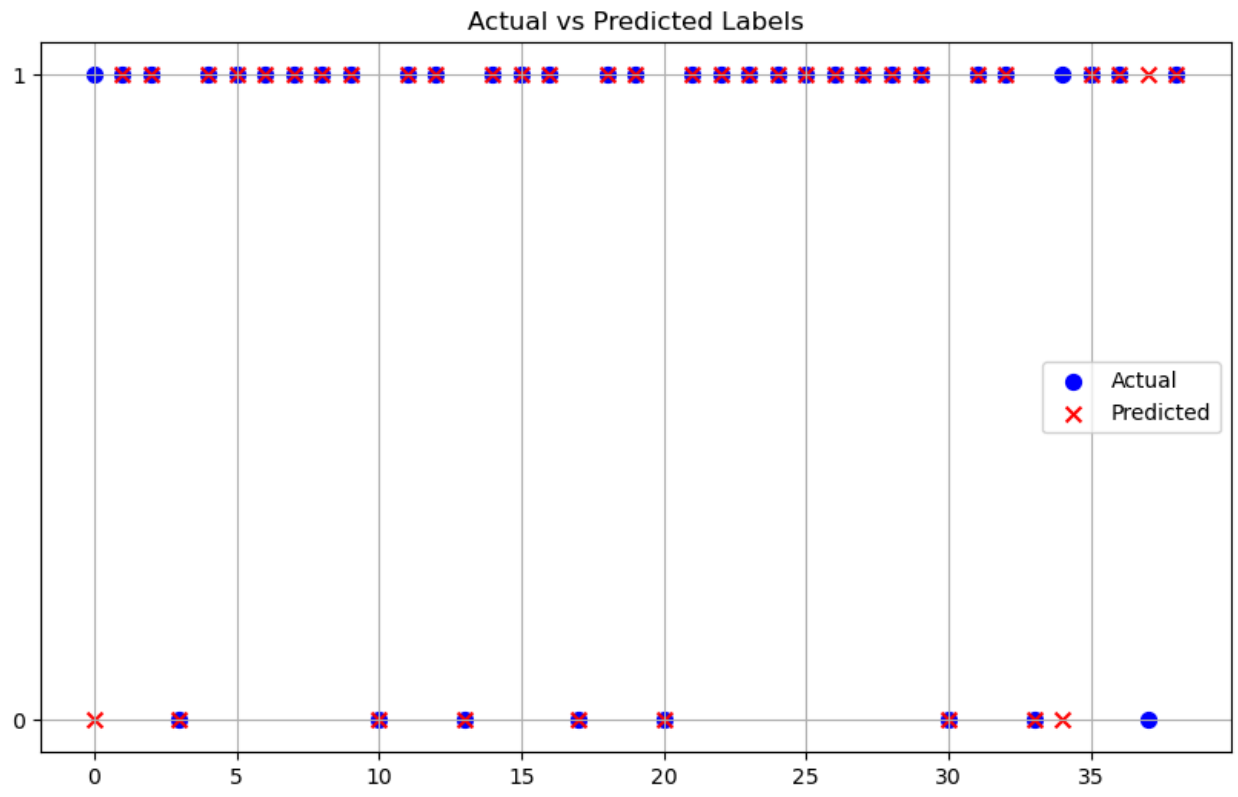


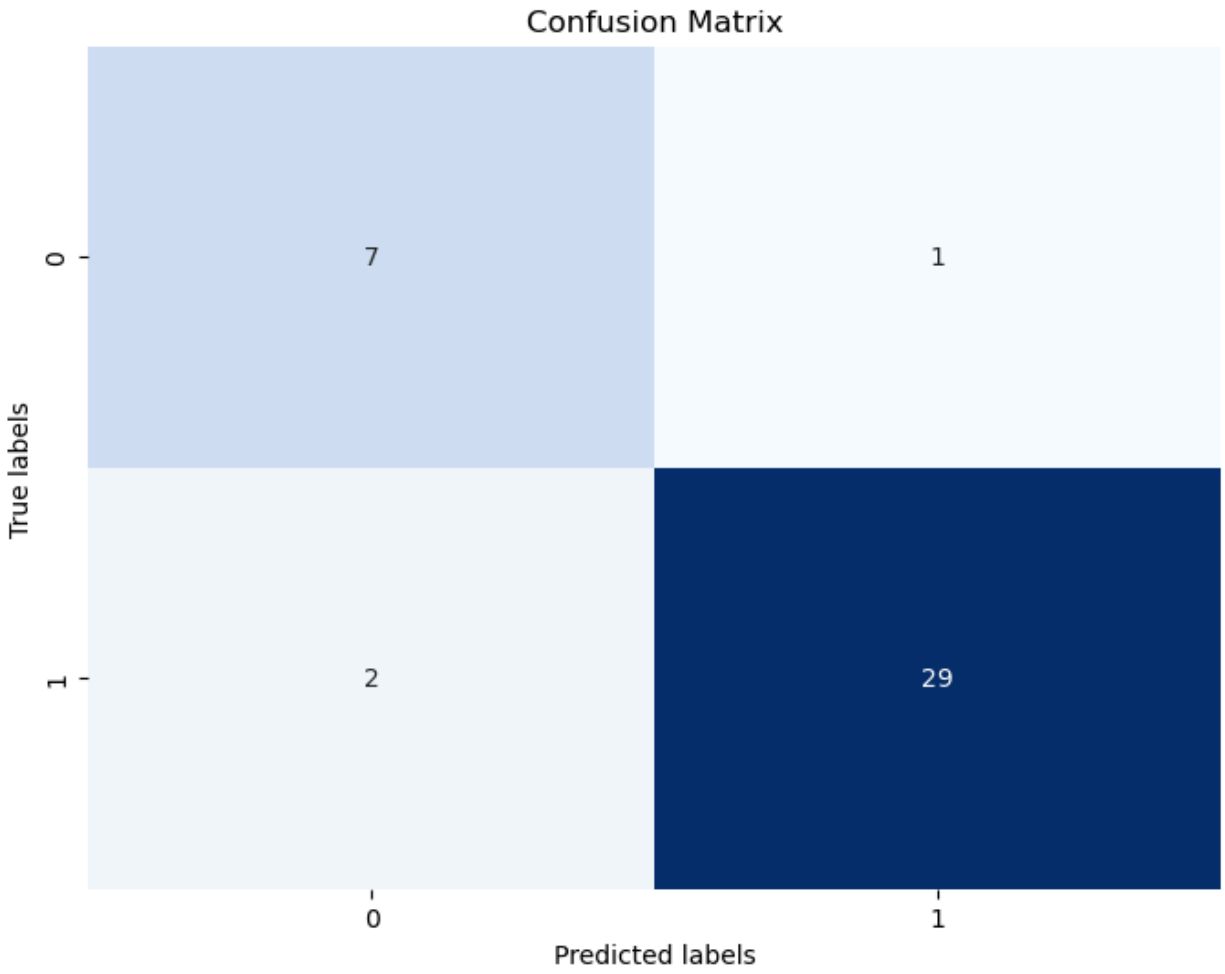
Using Fixed data:

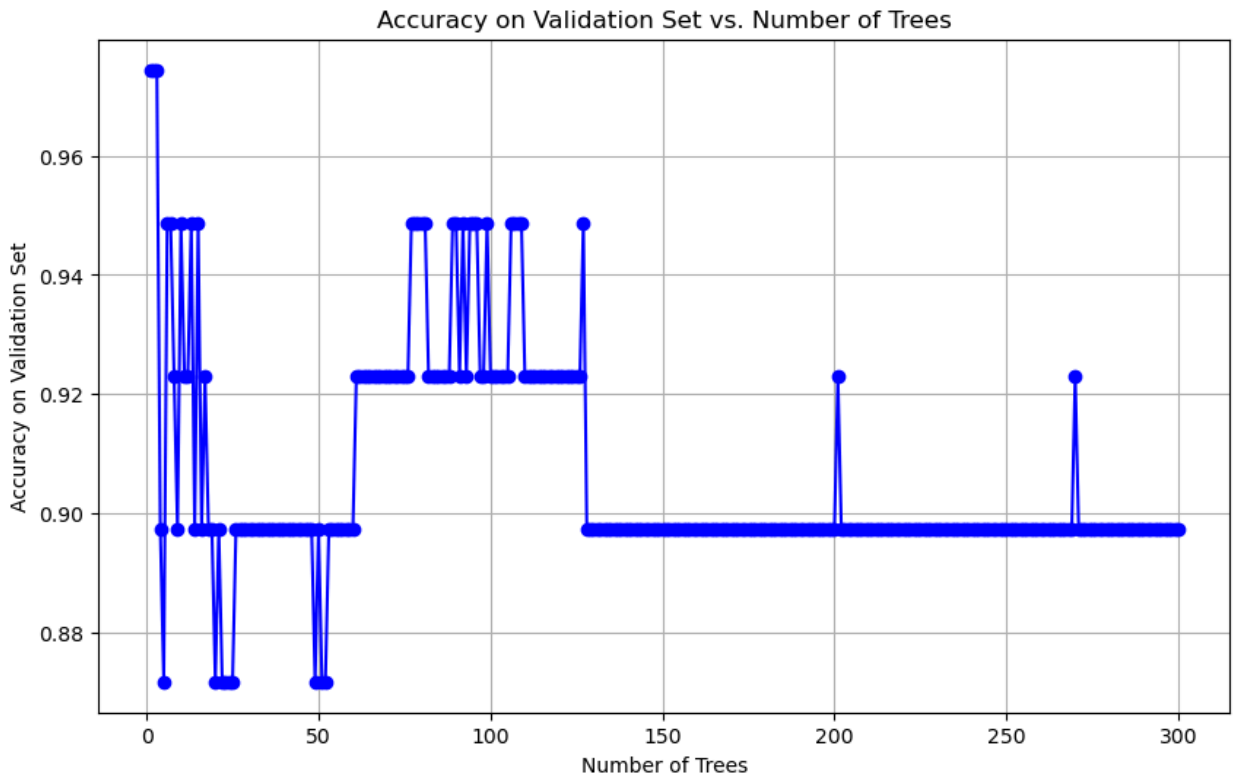




ADA Boost:

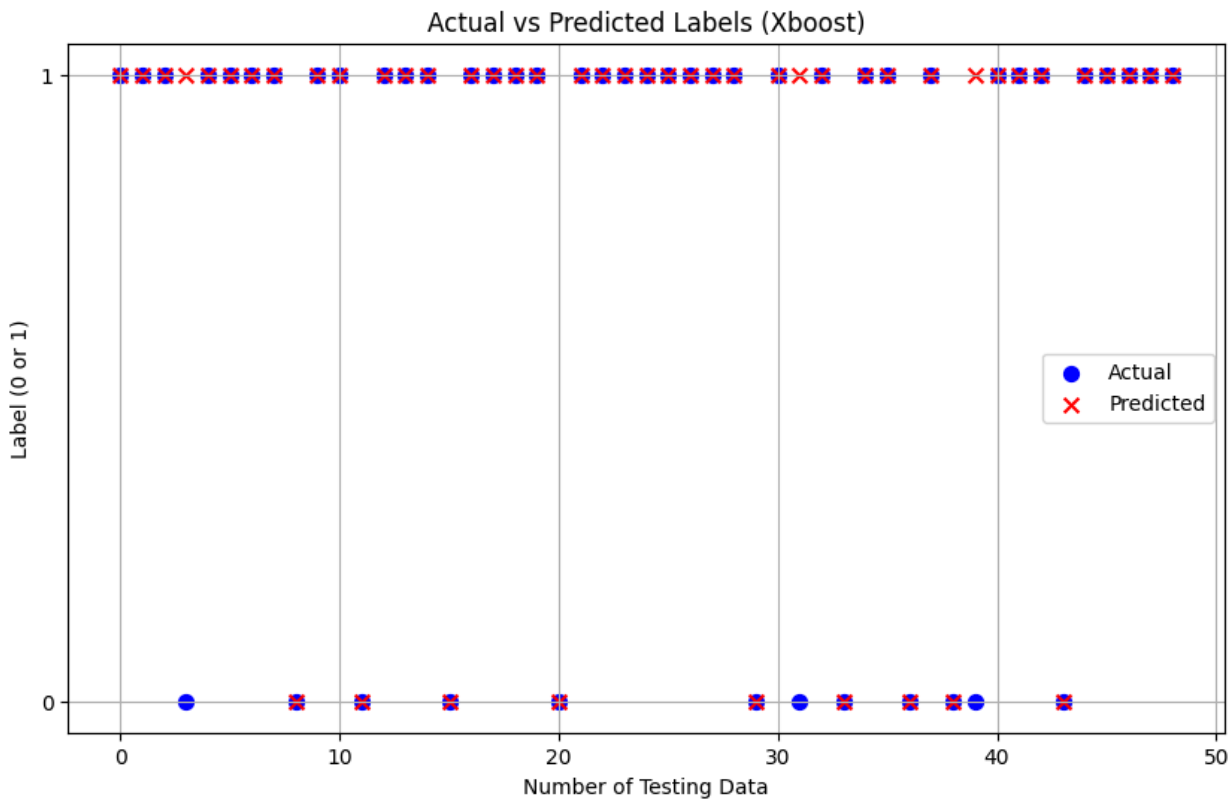

Actual vs Predicted Labels
1
0
0
5
10
15
20
25
30
35
Actual
Predicted

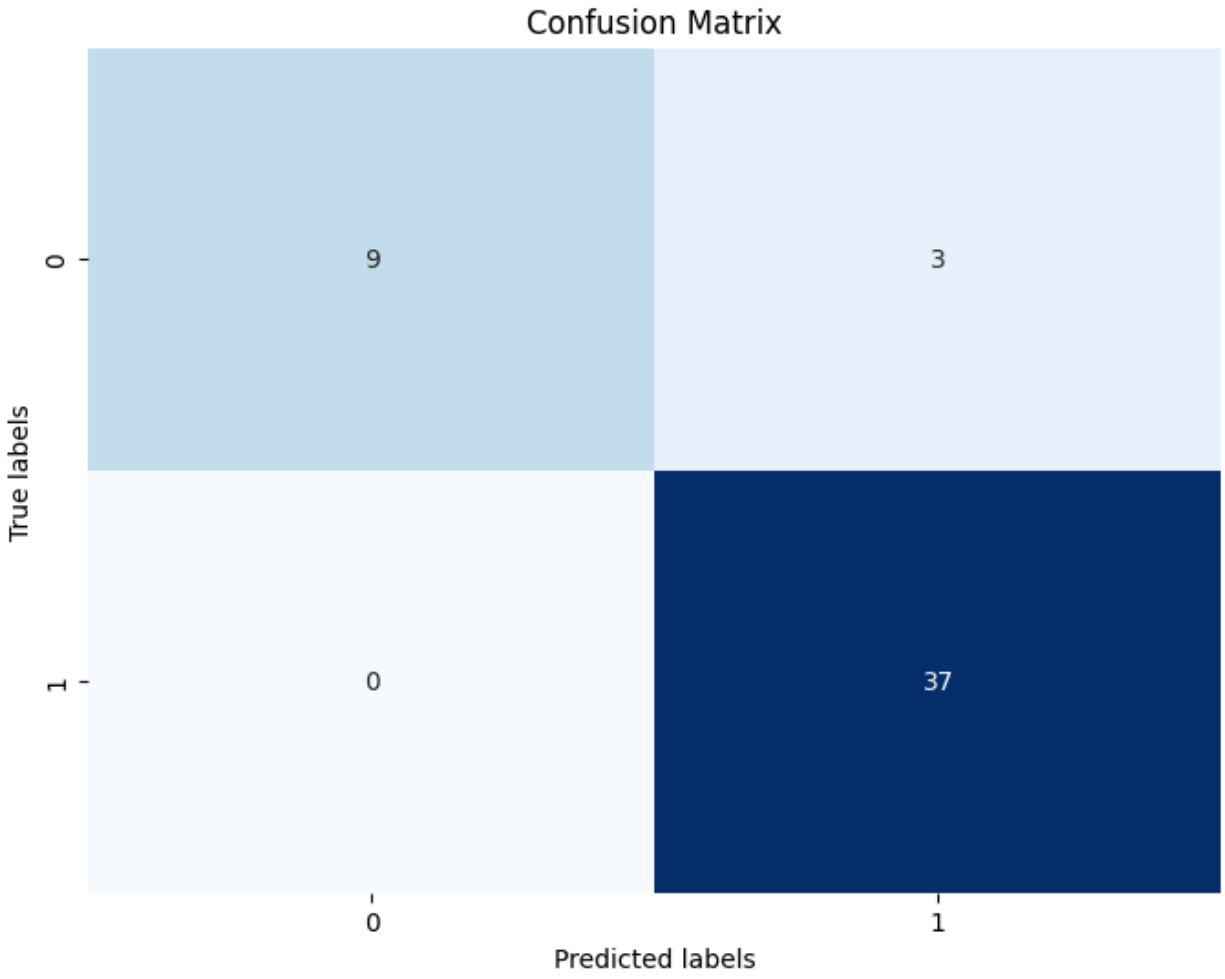

Confusion Matrix
True labels
Predicted labels
0
1
0
1
7
1
2
29

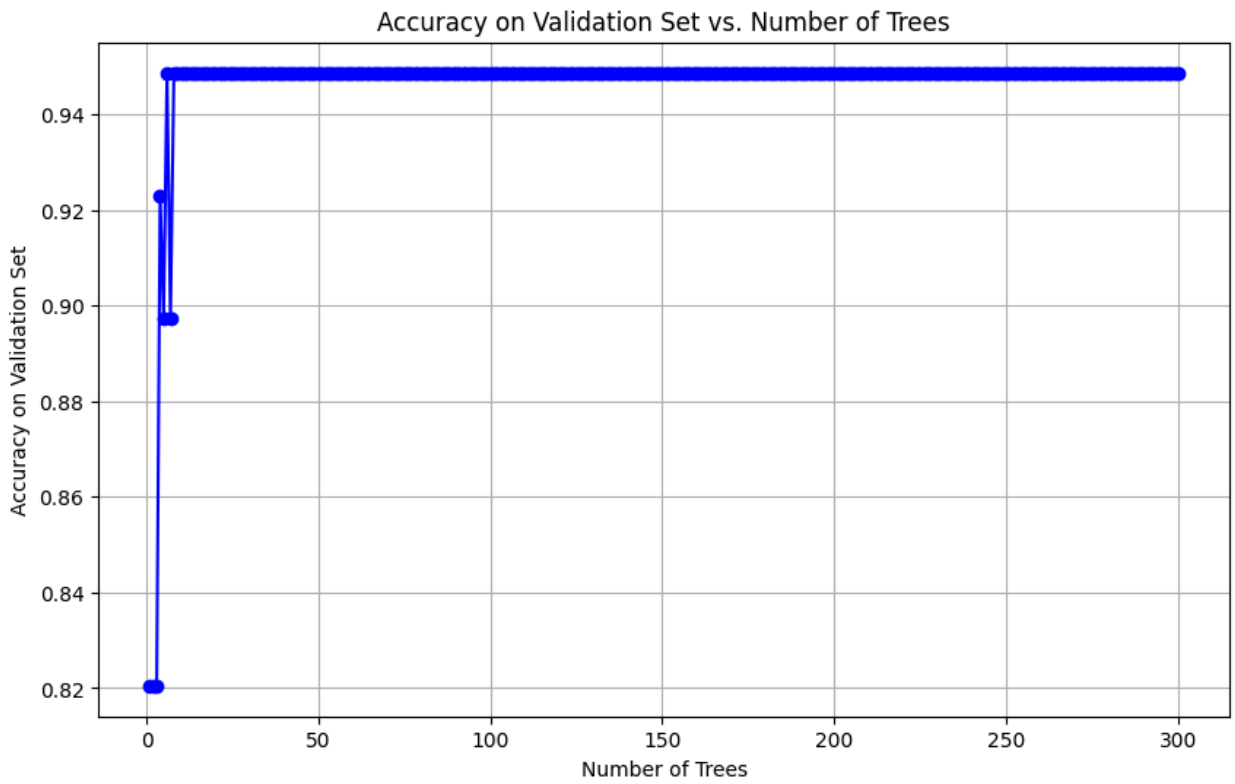


The spikes you're observing in the accuracy plot might be due to the fluctuation in performance as the number of trees in the AdaBoost model increases. we can say that as the complexity of the AdaBoost model increases (by adding more trees), it becomes more prone to overfitting, resulting in fluctuations in performance on the validation set.thus it is not a good choice to go with ADABOOST for Parkinson disease classification

XGBoost:


Actual vs Predicted Labels (Xboost)
Label (0 or 1)
1
0
0
10
20
30
40
50
Number of Testing Data
Actual
Predicted


Confusion Matrix
9
3
0
37
0
1
True labels
Predicted labels



This increasing trend suggests that the model effectively learns and adapts to the data over iterations. As more trees are added, the ensemble model becomes more robust and capable of capturing complex patterns in the data.

VI. Conclusion

After experimenting with various machine learning models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, and AdaBoost, we found that XGBoost consistently outperformed the other models in terms of accuracy on our dataset. Despite our efforts to optimize the parameters and explore different algorithms, XGBoost consistently demonstrated superior performance, achieving the highest accuracy among all models tested. This suggests that the dataset is well-suited for XGBoost, as it effectively leverages the strengths of gradient boosting to handle complex relationships and improve predictive accuracy. Therefore, for this particular dataset, XGBoost is the preferred choice for achieving the best predictive performance.