COMPSCI 589 Homework 3 - Spring 2024

By Noshitha Padma Pratyusha Juttu

1 Instructions to run the file

The below are the details of python files for each question:

1. For Q1: filename: `Random_Forest_Wine_Info_gain_Numerical.py`
2. For Q2: filename: `Random_Forest_Votes_Info_gain_Categorical.py`
3. EXTRA #1: filename: `Random_Forest_gini.py`
4. EXTRA #2: filename: `Random_Forest_Cancer_Info_gain.py`

Programming Section (100 Points Total)

In this homework you will be analyzing two datasets:

(1) The Wine Dataset. The goal, here, is to predict the type of a wine based on its chemical contents. The dataset is composed of 178 instances. Each instance is described by 13 *numerical* attributes, and there are 3 classes.

Accuracy :

- The accuracy of the model consistently improves with an increase in the number of trees.
- It reaches a peak of approximately 97.6% at 20 trees and remains relatively stable thereafter.
- Beyond 20 trees, there's little change in accuracy.

Precision and Recall :

- Both precision and recall also show improvements with the number of trees.
- Precision, which measures the proportion of true positive predictions among all positive predictions, reaches a peak of approximately 97.3% at 20 trees.
- Recall, which measures the proportion of true positive predictions among all actual positive instances, also reaches a peak of approximately 98.3% at 20 trees.

F1 Score :

- The F1-score, which is the harmonic mean of precision and recall, follows a similar pattern to precision and recall.

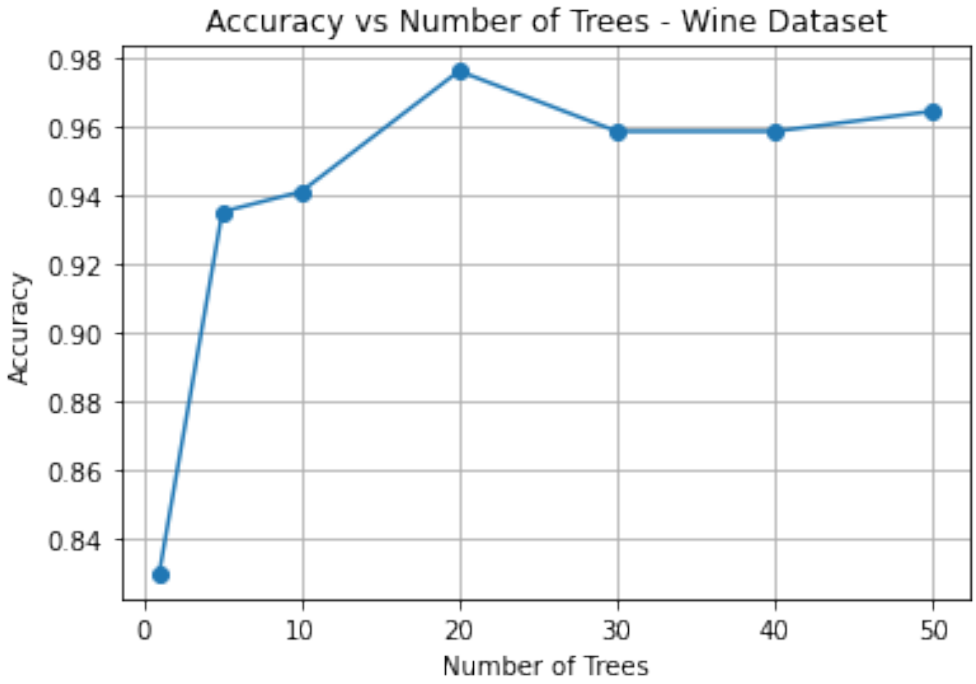

Figure 1: Accuracy vs No. of trees -Wine Data set - Information Gain

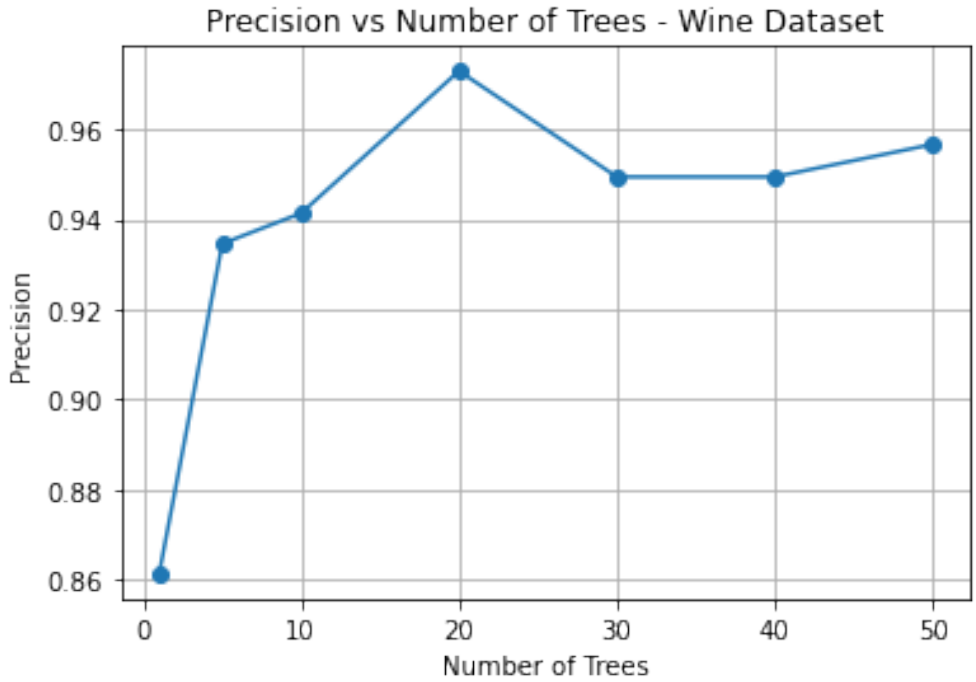

Figure 2: Precision vs No. of trees -Wine Data set - Information Gain

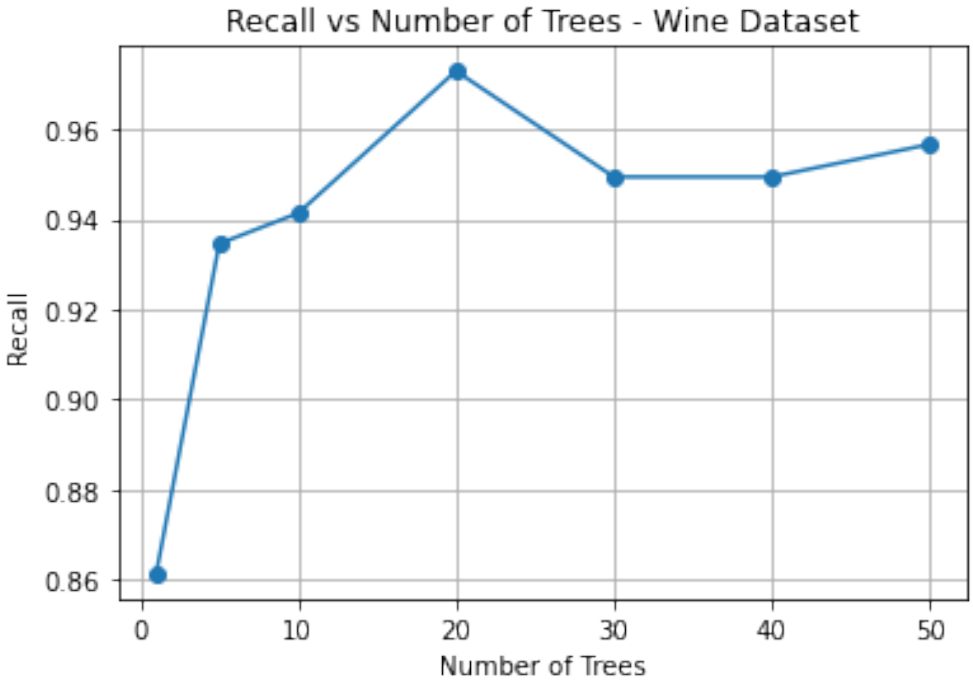

Figure 3: Recall vs No. of trees -Wine Data set - Information Gain

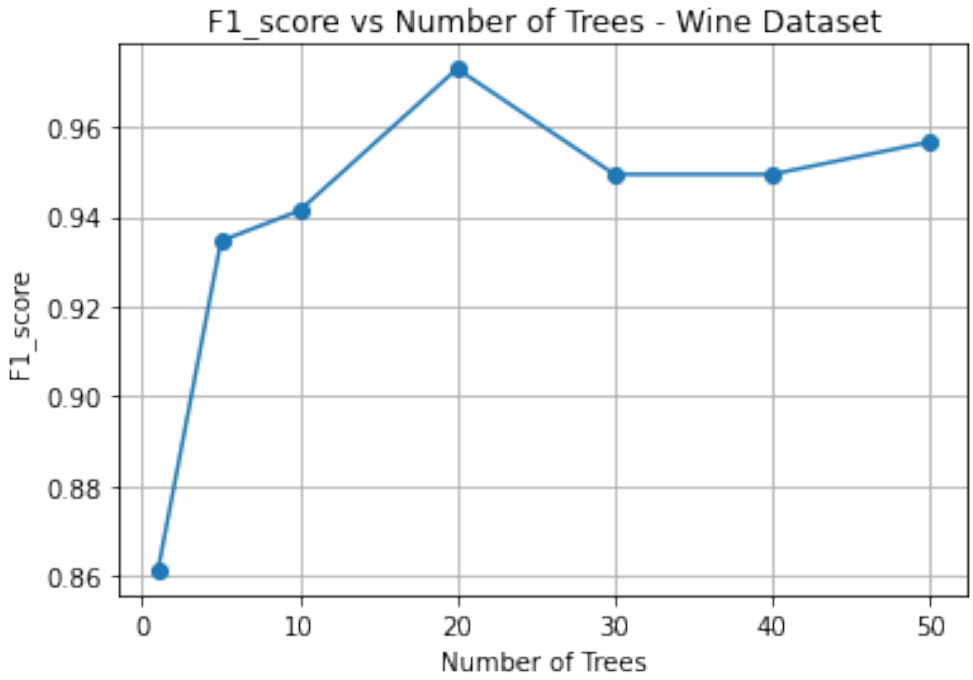

Figure 4: F1 Score vs No. of trees - Wine Data set - Information Gain

- It reaches a peak of approximately 97.6% at 20 trees.

Analysis:

- Increasing the number of trees generally improves these performance metrics up to a certain point (around 20 trees), beyond which there's little gain or even a slight decrease in performance.
- This suggests that the model's ability to capture complex patterns in the data improves with more trees initially, but beyond a certain threshold, adding more trees does not significantly enhance performance and may introduce unnecessary computational overhead.

(2) The 1984 United States Congressional Voting Dataset. The goal, here, is to predict the party (Democrat or Republican) of each U.S. House of Representatives Congressperson. The dataset is composed of 435 instances. Each instance is described by 16 *categorical* attributes, and there are 2 classes.

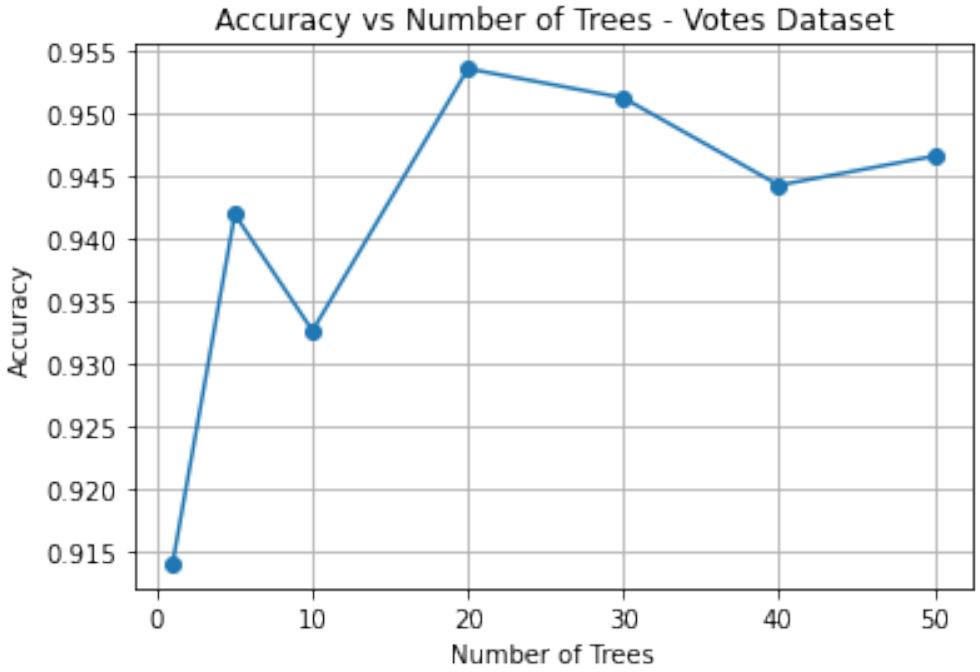

Figure 5: Accuracy vs No. of trees - Voting Data set- Information Gain

Accuracy :

- In this case, the accuracy generally increases with the number of trees but starts to plateau around 20 trees.
- Beyond this point, there's little improvement or even a slight decrease in accuracy.
- For real-life deployment, a balance between computational resources and performance is crucial. With an accuracy of around 95.3 percent at 20 trees, this would be a reasonable

Precision and Recall :

- Both precision and recall show similar trends to accuracy, increasing with the number of trees

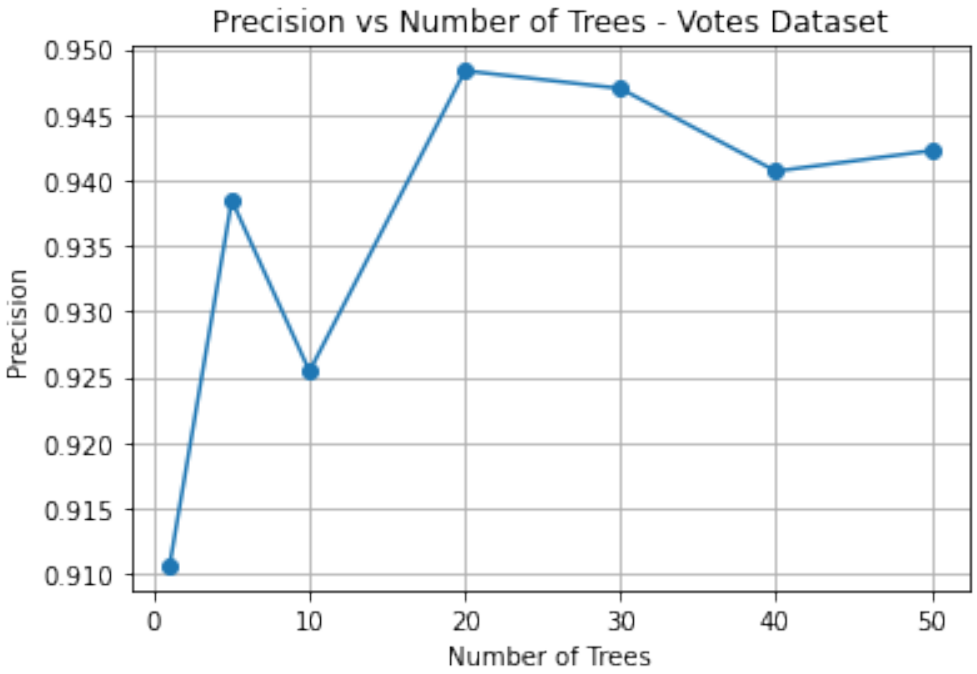

Figure 6: Precision vs No. of trees - Voting Data set- Information Gain

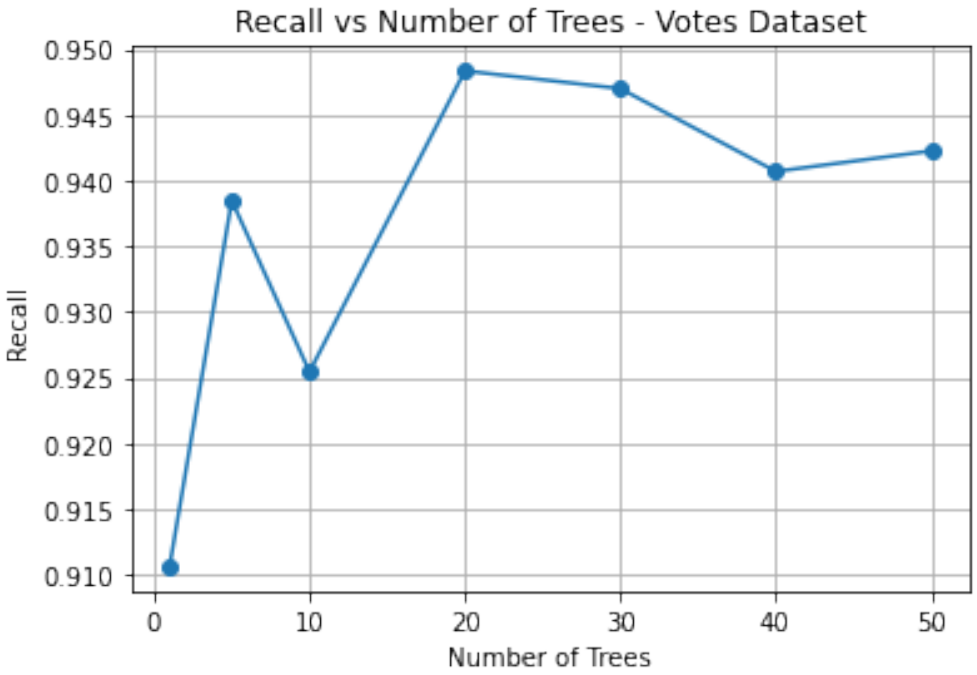

Figure 7: Recall vs No. of trees - Voting Data set- Information Gain

and stabilizing around 20 trees.

- A balance between precision and recall is important depending on the specific application. However, since precision and recall are relatively high and stable from 20 trees onward, selecting 20 trees would be reasonable for deployment.

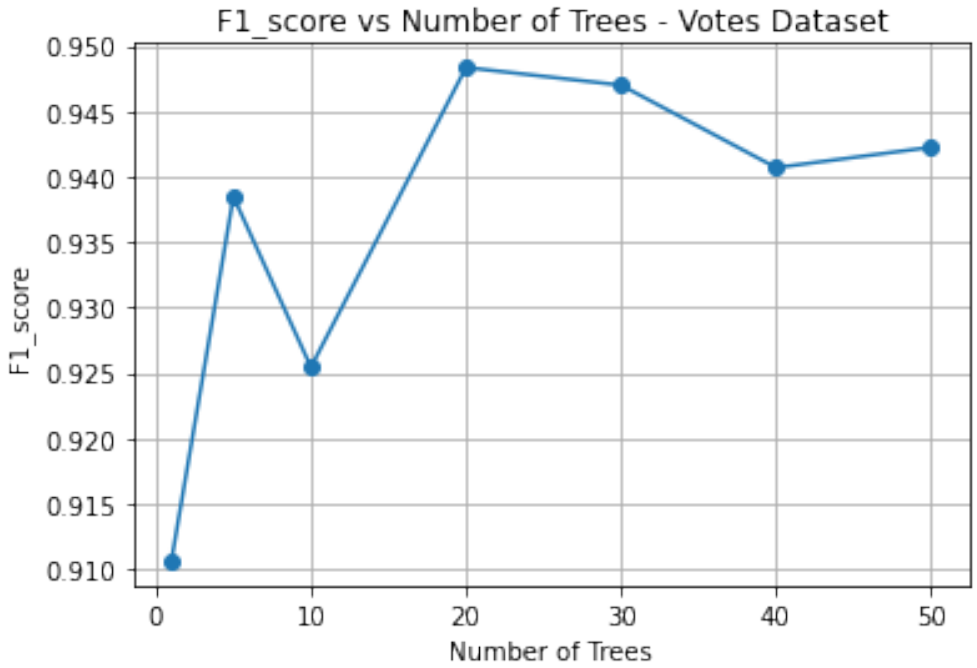

Figure 8: F1 Score vs No. of trees - Voting Data set- Information Gain

F1 Score : The F1-score is the harmonic mean of precision and recall and provides a balance between them. Like precision and recall, the F1-score improves with the number of trees and stabilizes around 20 trees. Since the F1-score reflects the balance between precision and recall, selecting the number of trees that maximizes the F1-score (around 20 trees) would ensure a good balance between precision and recall in real-life deployment.

Analysis :

- Accuracy: Initially, increasing the number of trees leads to improved accuracy, but beyond a certain point, there's little gain. This suggests that the accuracy of the random forest becomes less sensitive to increasing ntree beyond a certain threshold.
- Precision, Recall, and F1-score: Similar to accuracy, precision, recall, and F1-score show improvements with increasing ntree up to a certain point.
- This indicates that the model's ability to capture complex patterns in the data improves as the number of trees increases.
- However, beyond a certain point (around 20 trees), the marginal improvement diminishes, suggesting that adding more trees doesn't significantly enhance performance and may even introduce overfitting or computational overhead without substantial gains in predictive power.

There are three ways in which we may receive extra credit in this homework.

(Extra Points #1: 8 Points) Reconstruct the same graphs as above, but now using the Gini criterion. You should present the same analyses and graphs mentioned above. Discuss whether (and how) different performance metrics were affected (positively or negatively) by changing the splitting criterion, and explain why you think that was the case.

Dataset- 1: The Wine Dataset

- Accuracies: [0.9647058823529411, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
- Precisions: [0.9541666666666666, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
- Recalls: [0.94625, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
- F1-scores: [0.9662698412698413, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]
- Num_trees: [1, 5, 10, 20, 30, 40, 50]

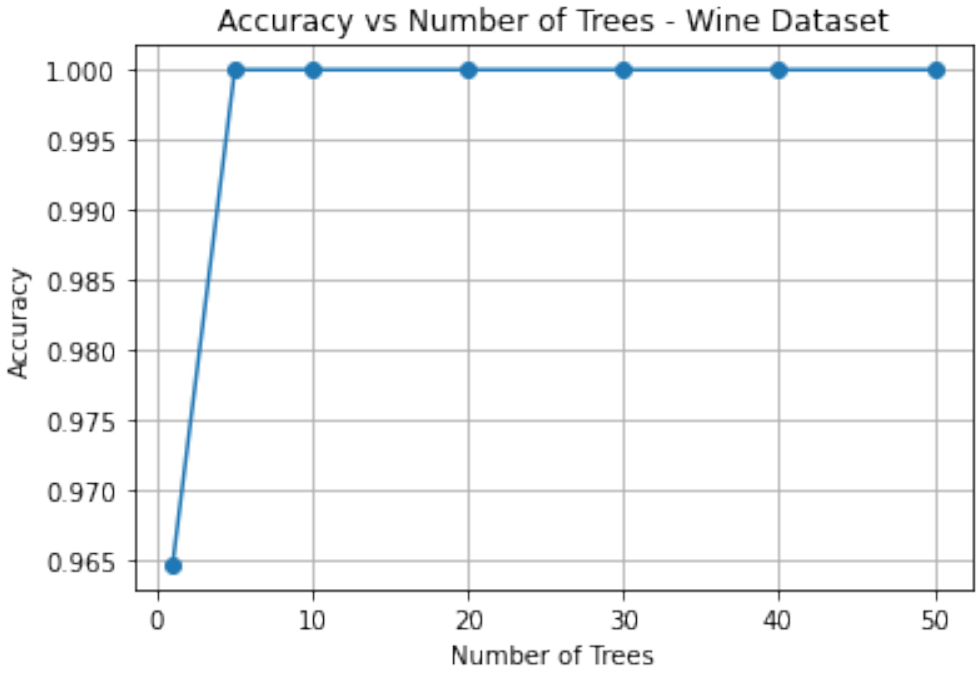

Figure 9: Accuracy vs No. of trees - Wine - gini criteria

Analysis :

- The performance metrics are not particularly sensitive to increasing the number of trees beyond 1, and adding more trees does not improve performance further.
- The F1-score, along with other metrics, quickly reaches its maximum value with just 1 tree, indicating that this dataset does not require a significant number of trees in the ensemble to achieve optimal performance.

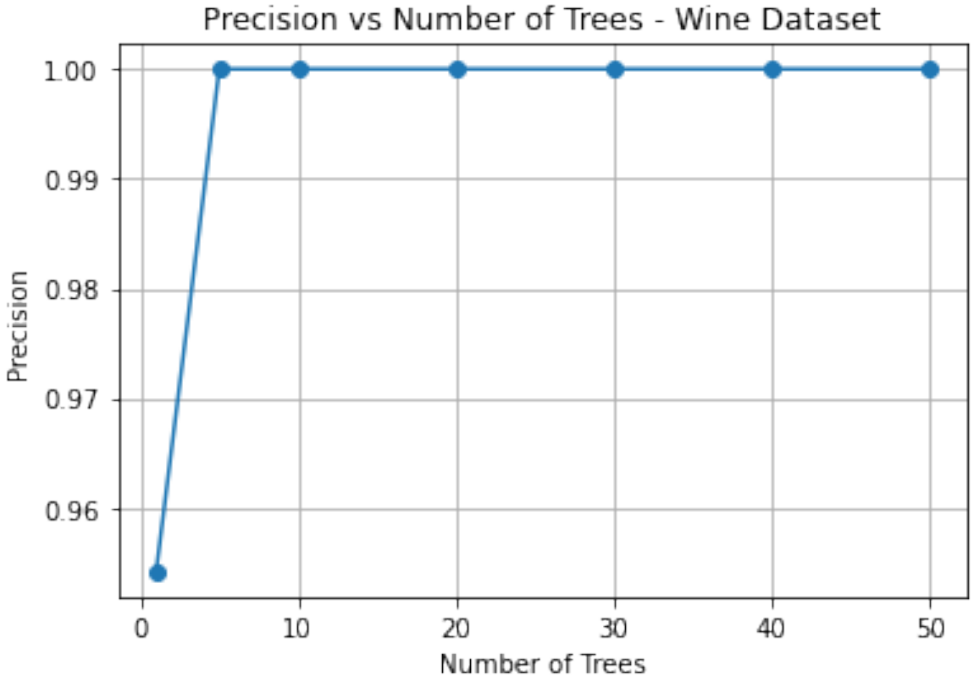

Figure 10: Precision vs No. of trees -Wine - Gini criteria

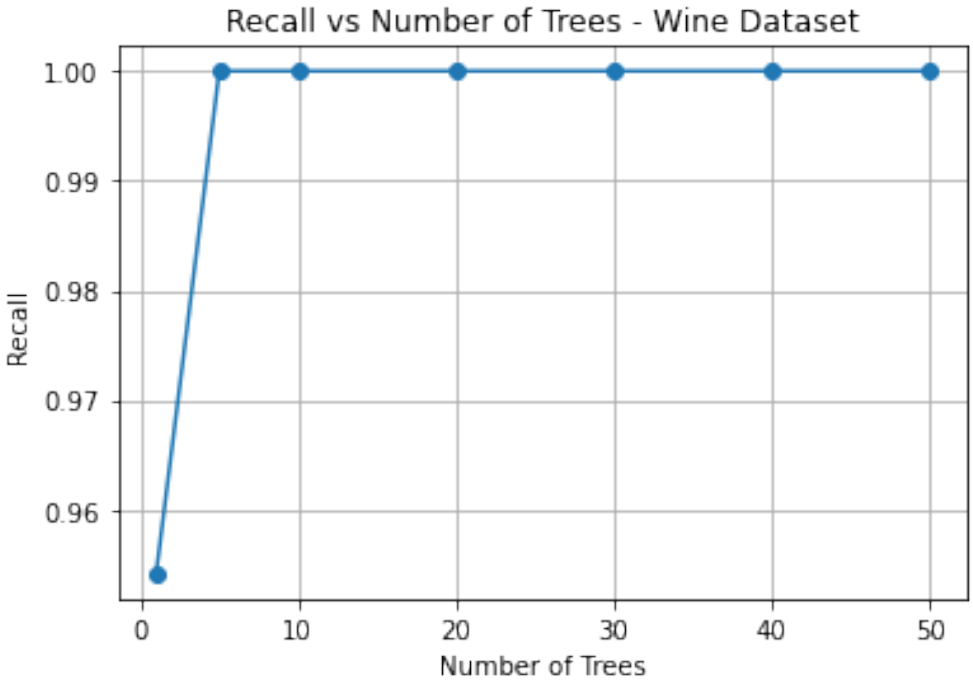

Figure 11: Recall vs No. of trees - Wine Data set - Gini criteria

- Gini impurity as the splitting criterion generally led to improvements in all performance metrics, including accuracy, precision, recall, and F1-score.
- Gini impurity resulted in better separation of classes, fewer false positives, and improved balance between precision and recall.
- Overall, Gini impurity was more effective for this dataset compared to InfoGain.

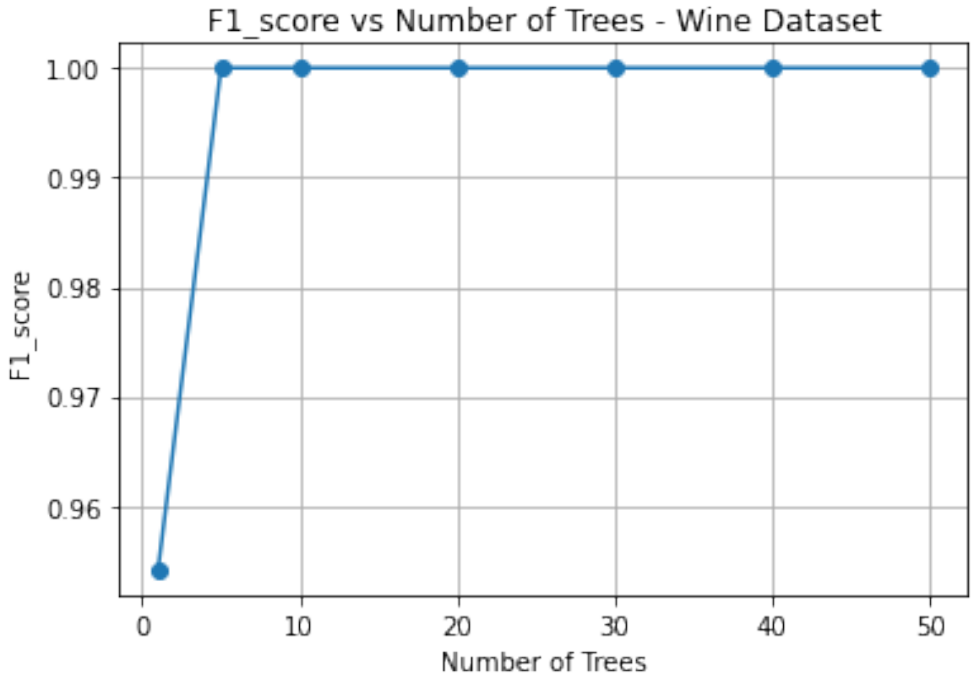

Figure 12: F1 Score vs No. of trees - Wine Data set - Gini criteria

Dataset- 2: The 1984 United States Congressional Voting Dataset. The goal, here, is to predict the party (Democrat or Republican) of each U.S. House of Representatives Congressperson. The dataset is composed of 435 instances. Each instance is described by 16 *categorical* attributes, and there are 2 classes.

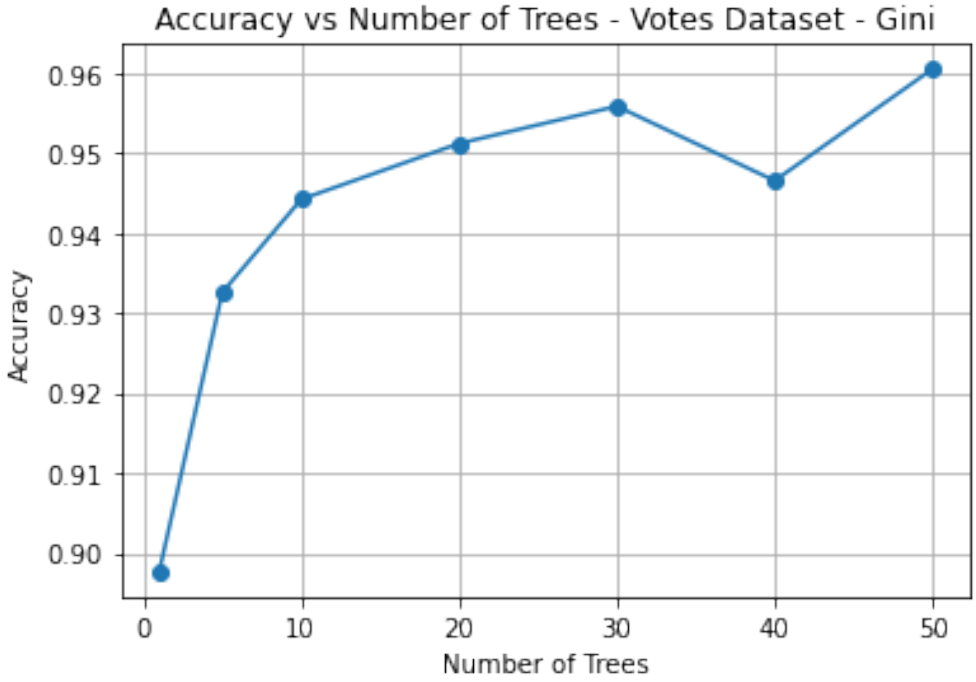

Figure 13: Accuracy vs No. of trees - Voting Data set - Gini

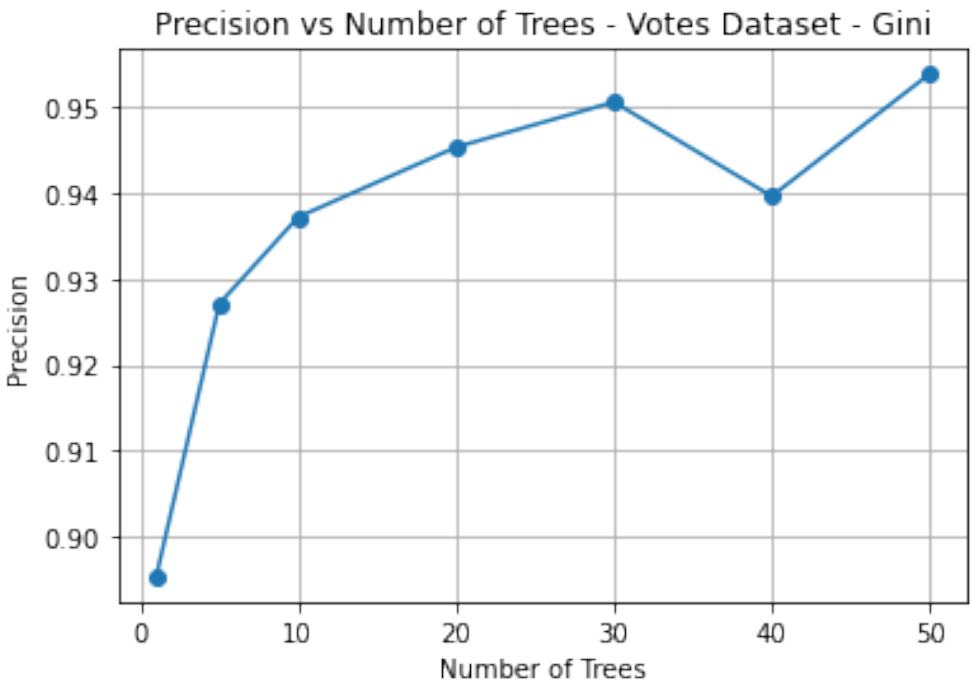

Figure 14: Precision vs No. of trees - Voting Data set

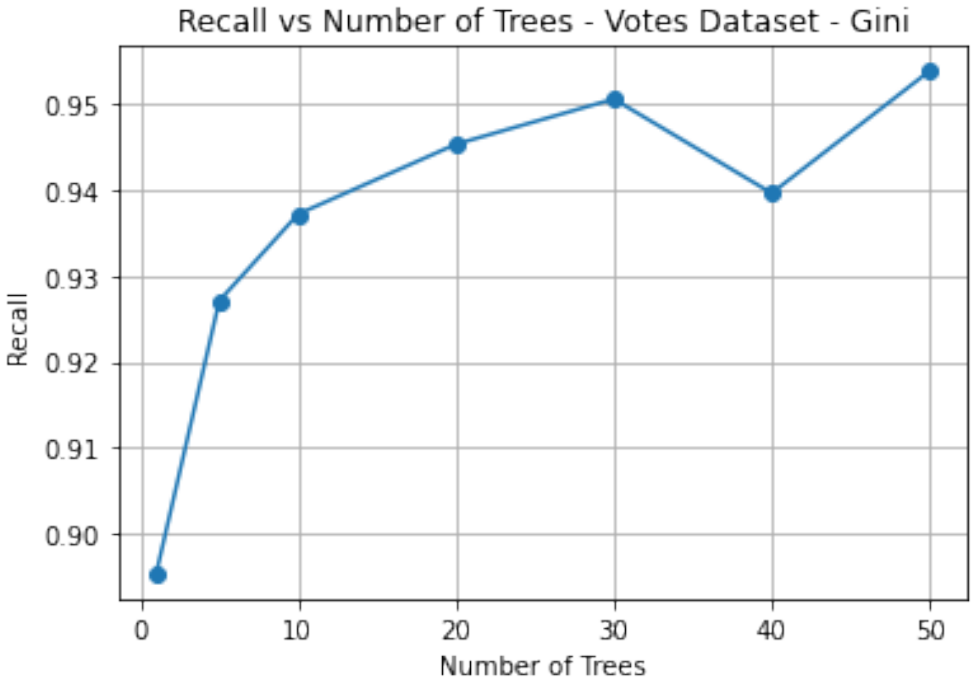

Figure 15: Recall vs No. of trees - Voting Data set

Analysis :

- The F1-score appears to be the most directly affected metric by changing the number of trees.
- While other metrics such as accuracy, precision, and recall also improve with more trees, they may plateau or show diminishing returns beyond a certain point.
- Therefore, selecting the number of trees that maximizes the F1-score would ensure a well-balanced performance of the classifier in real-life deployment.
- The optimal number of trees (ntrees) for this classifier would be 50.
- The choice of splitting criterion can significantly impact the performance of a random forest classifier.
- In this scenario, Gini impurity outperforms information gain in terms of accuracy, precision, recall, and F1-score.
- Gini impurity, by favoring more balanced splits, may require a larger number of trees to effectively capture the complexity of the data and achieve optimal performance, while information gain might achieve satisfactory performance with fewer trees but may be more prone to overfitting.

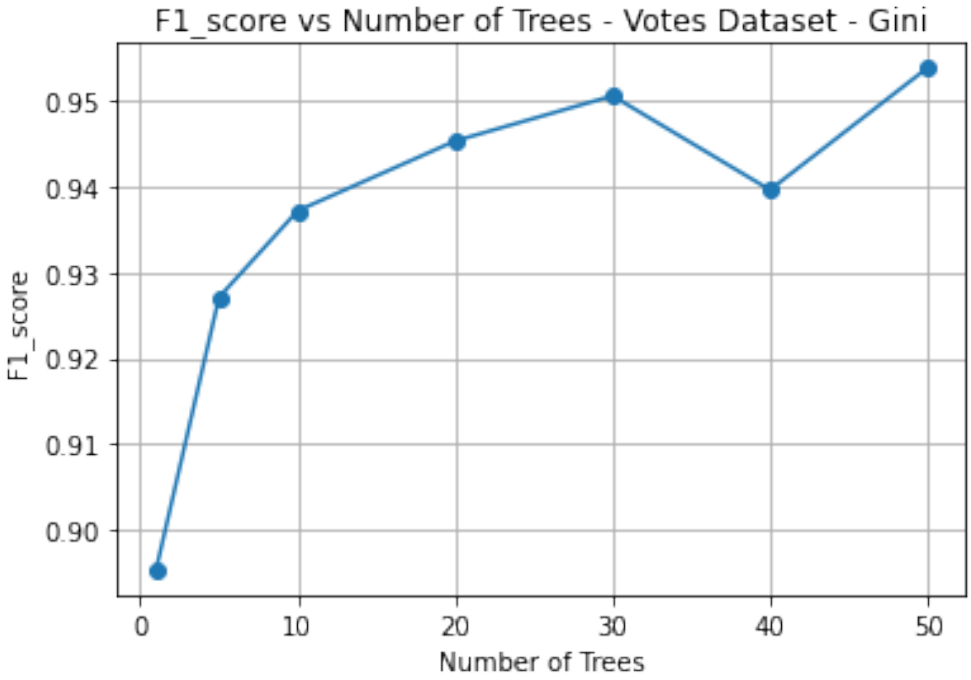

Figure 16: F1 Score vs No. of trees - Voting Data set

(Extra Points #2: 8 Points) Analyze a third dataset: the **Breast Cancer Dataset**. The goal, here, is to classify whether tissue removed via a biopsy indicates whether a person may or may not have breast cancer. There are 699 instances in this dataset. Each instance is described by 9 *numerical* attributes, and there are 2 classes. You should present the same analyses and graphs as discussed above. This dataset can be found in the same zip file as the two main datasets.

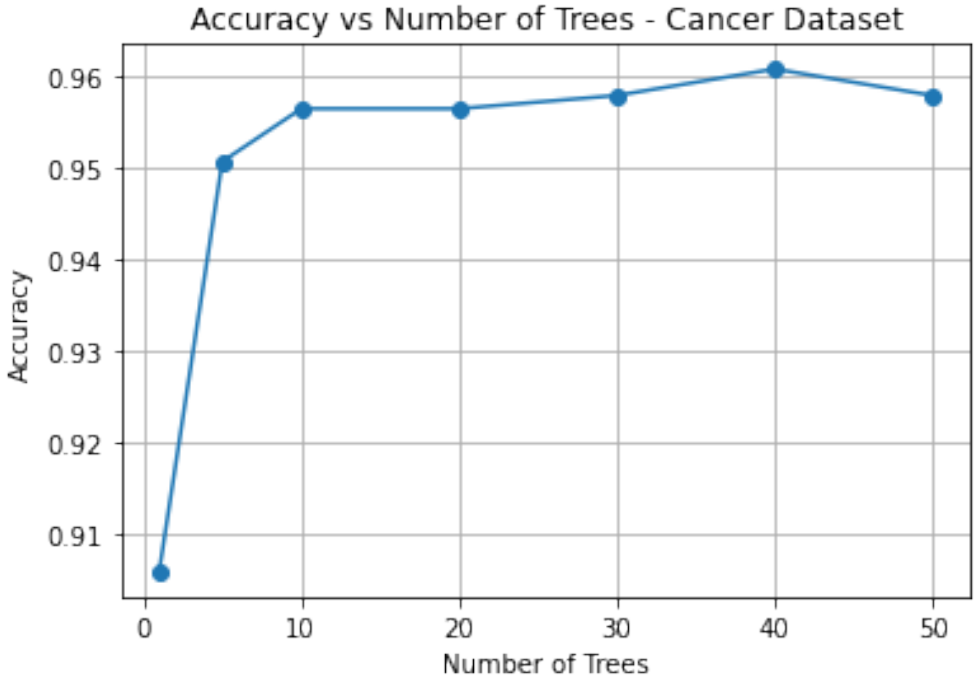

Figure 17: Accuracy vs No. of trees - Cancer dataset

F1 Score :

- The F1-score, being the harmonic mean of precision and recall, also follows a similar pattern to precision and recall.
- It reaches its peak of approximately 96.9% at 30 trees, indicating a good balance between precision and recall.

Analysis:

- *Optimal ntree value:* Around 40 to 50 trees would be suitable for deployment in real life, given the balanced performance across metrics.
- *Effect of ntree on Performance Metrics:*
  - *Accuracy:* Improves steadily with increasing ntree, reaching a peak around 40 to 50 trees.
  - *Precision and Recall:* Both metrics exhibit consistent improvement with increasing ntree, showing stability around 40 to 50 trees.
  - *F1-score:* Follows a similar trend to precision and recall, peaking around 40 to 50 trees.
- *Sensitivity of Metrics to ntree:*

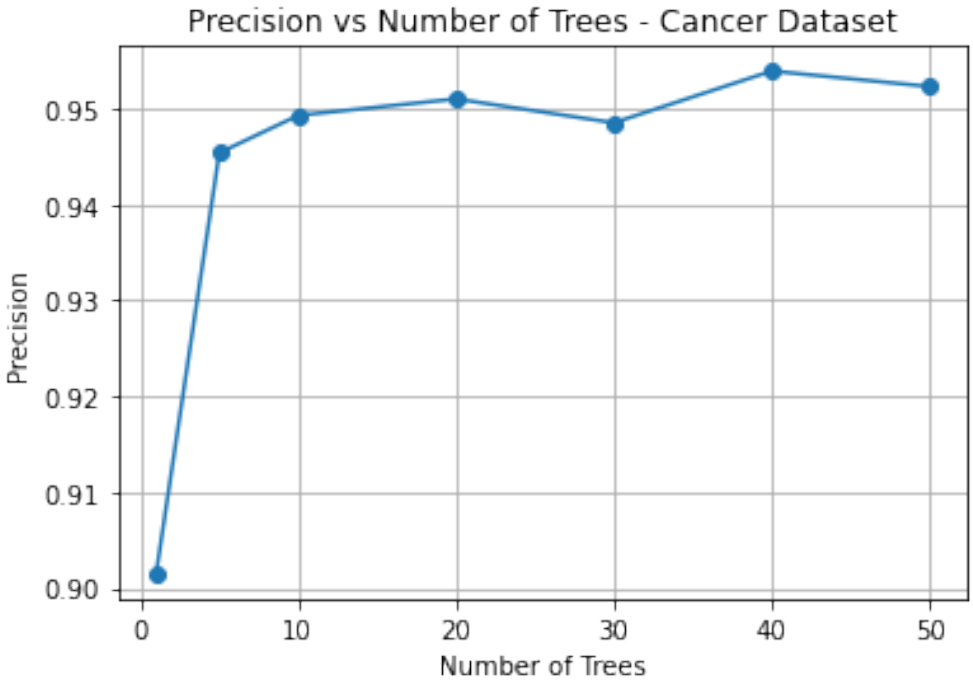

Figure 18: Precision vs No. of trees - Cancer dataset

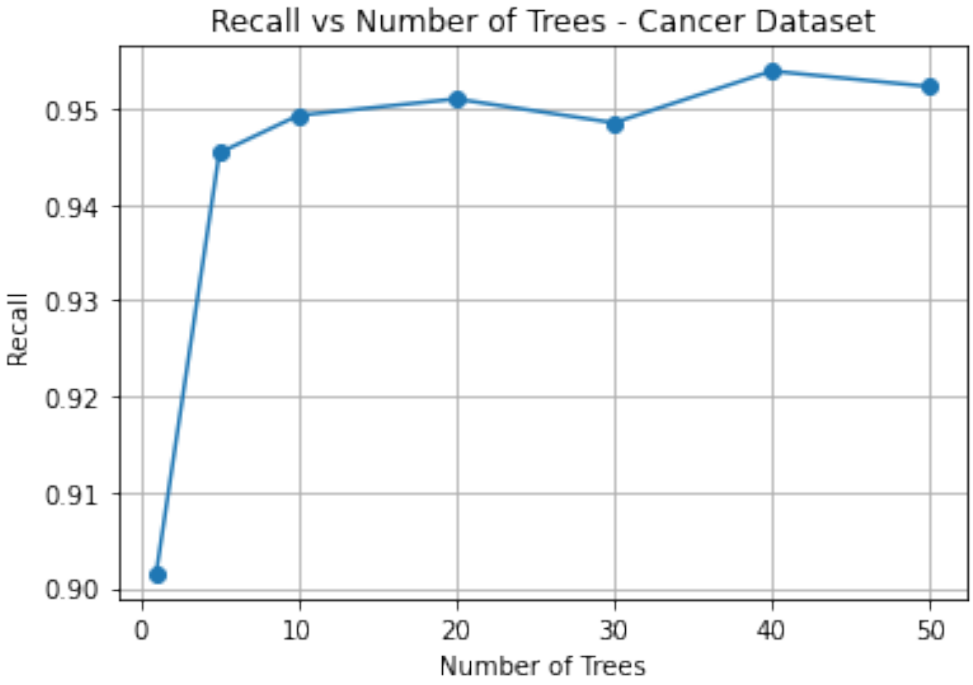

Figure 19: Recall vs No. of trees - Cancer dataset

  - Accuracy: Moderately sensitive to changes in ntree, with noticeable improvements up to around 40 to 50 trees.
  - F1-score: Requires a moderate number of trees to optimize, typically around 40 to 50 trees.

- In summary, deploying a random forest classifier with around 40 to 50 trees would be a reasonable choice for the cancer dataset. This selection provides a balance between achieving high performance metrics (accuracy, precision, recall, and F1-score) and minimizing computational overhead

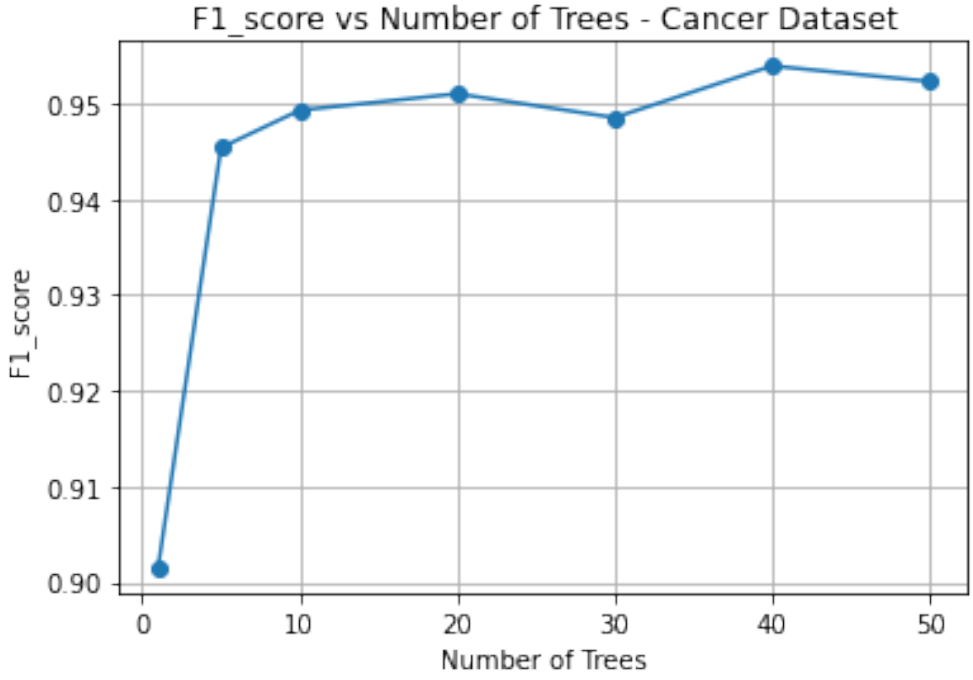

Figure 20: F1 Score vs No. of trees - Cancer dataset