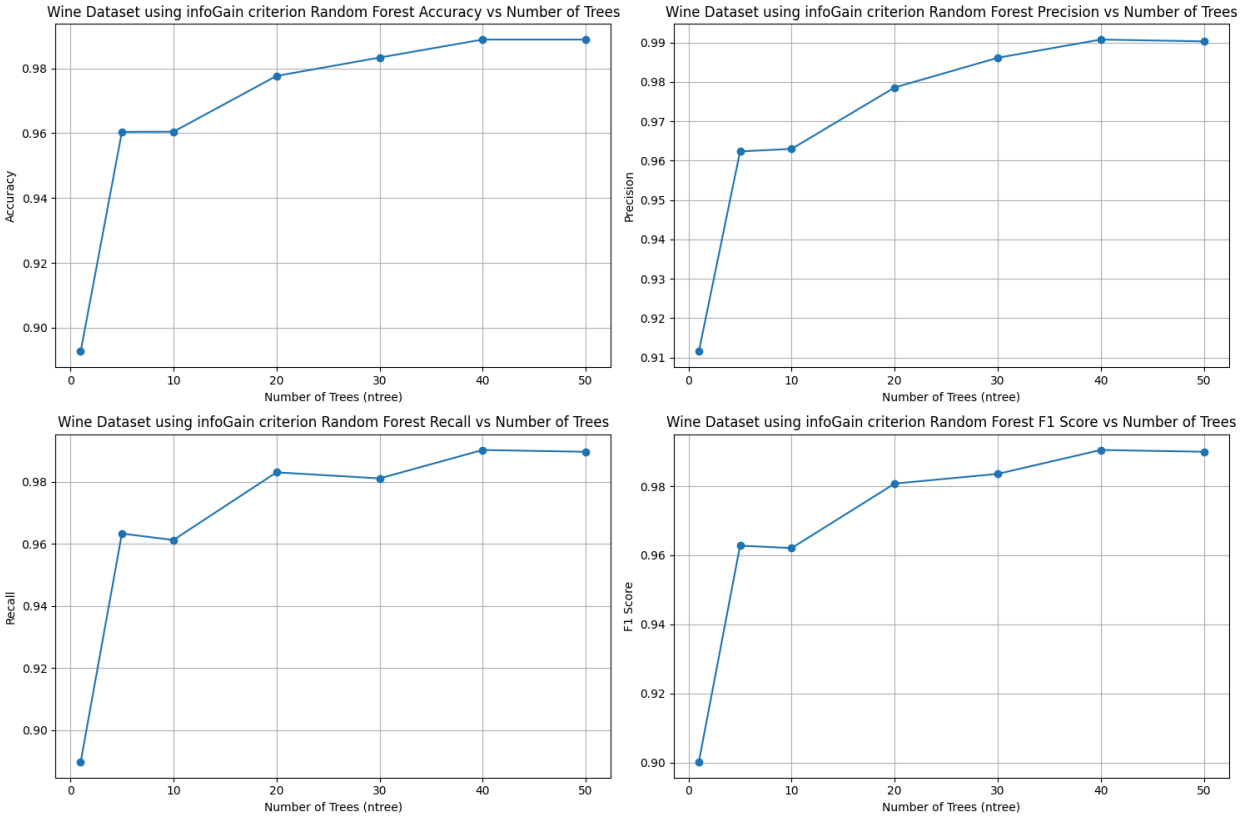
CS589 HW3
SHREYA BIRTHARE
34060222

For all my datasets for building a decision tree, I used a mixed stopping criteria. I determined by trying out different values for gain, max depth and minimum size for split. For my calculations I have used min_size=2, max_depth=7, min_gain=0.0001 as I was getting good results with these values.

WINE DATASET USING INFOGAIN

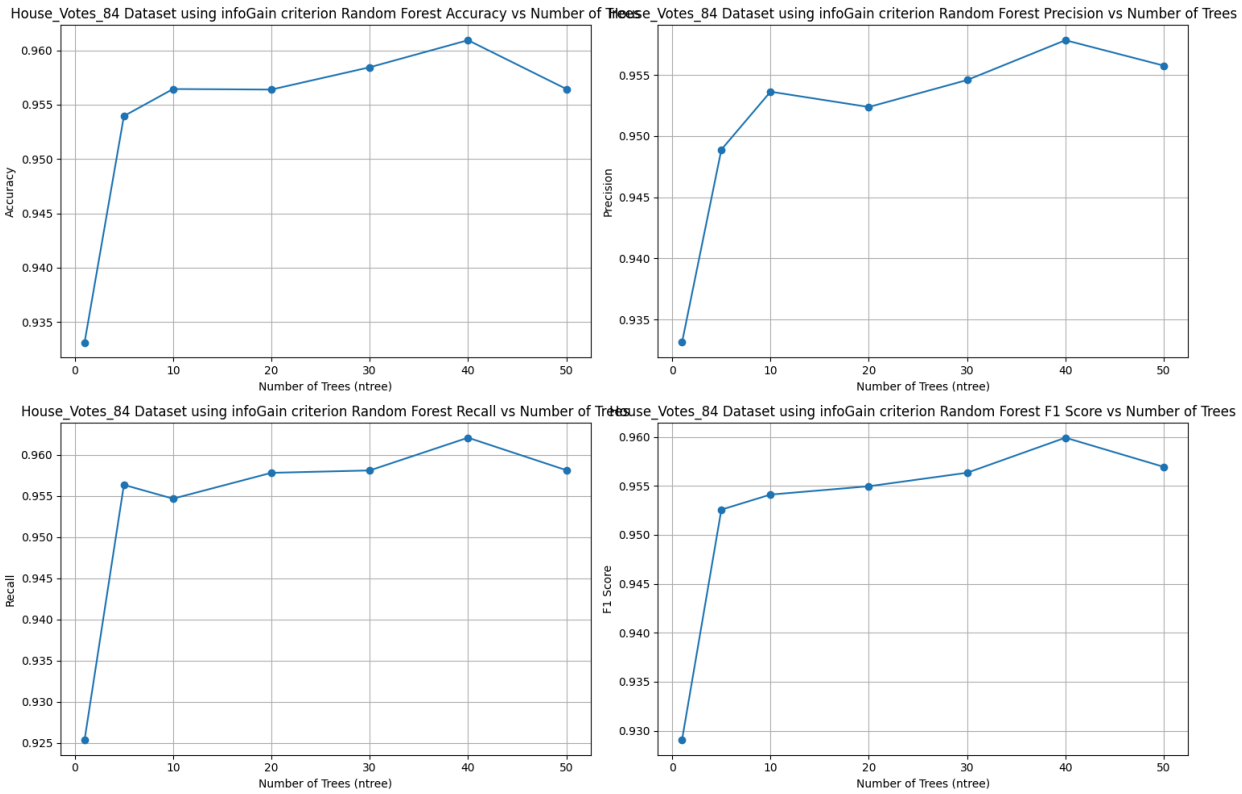
1. See code and output directories
2. See code and output directories
3. 

4. Accuracy: We observe the highest accuracy with ntree=40 & 50. This indicates that the classifier makes the correct prediction most of the time.
   Precision: Highest Precision observed with ntree=40 meaning that when the classifier predicts a label, it is very likely to be correct.
   Recall: Highest Recall observed with ntree=40 as it reflects the proportion of actual positives that were correctly identified
   F1 Score: Highest F1 Score observed with ntree=40 indicating a good balance between precision and recall.
   We can say that for ntree=40, we get the most optimal and well balanced results therefore.

5. We see that the accuracy,recall, precision, f1 score increase drastically from ntree = 1 to 10 and keep increasing until ntree=40. For ntree=20 to 30, the recall slightly decreases indicating that the model is identifying fewer of the actual positive cases relative to the total number of actual positives. For ntree=40, we see highest values of these parameters which is a good balance and is expected as we increase the number of trees to get better or optimal results as we will have more data points to evaluate our model. However, from ntree=40 to 50, there is no significant improvement in the performance. It very very slightly decreases but this decrease is almost negligible indicating that it has reached a plateau meaning that increasing ntree does not affect the performance of our model.

HOUSE OF VOTES DATASET USING INFO GAIN

1. See code and output directories
2. See code and output directories

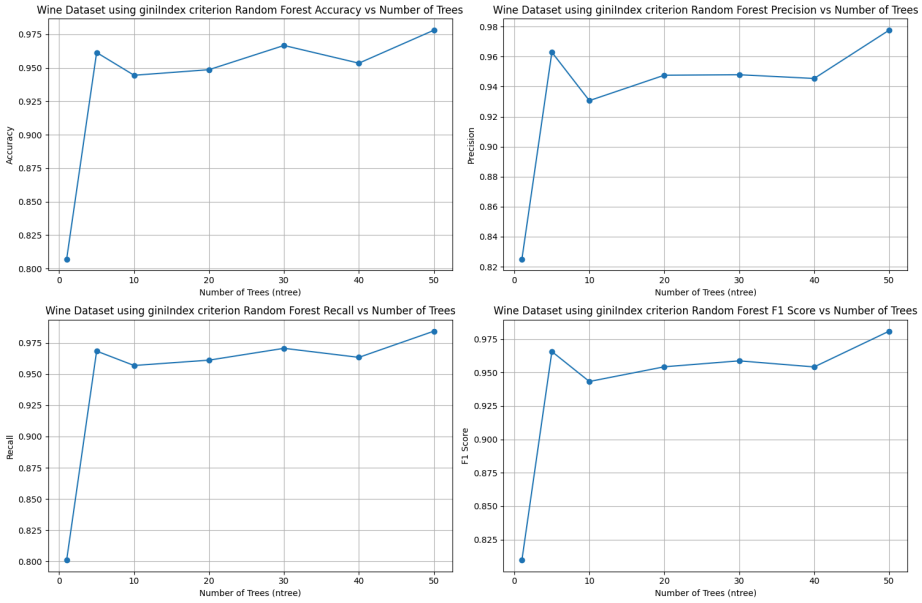


3.
4. We observe the highest values of ntrees for
   Accuracy: ntree=40 indicating that the classifier makes the correct prediction most of the time.
   Precision: ntree=40. Precision measures the ratio of true positives to all predicted positives, so improvements here mean the model is getting better at correctly labeling instances as positive, which in this context means correctly identifying a Congressperson's party.

Recall: ntree=40 indicating that the model is getting better at finding all relevant instances (i.e., correctly identifying as many Congresspersons' parties as possible). F1 Score: ntree=40 providing good balance between precision and recall to correctly identify a congressperson's party and also correctly identifying many of such parties.

5. We can see accuracy increases with the number of trees, but seems to level off after ntree=40. This suggests that the majority of the gains in correct classifications are obtained by this point, and adding more trees has a marginal or no significant effect. It rises from ntree=1-10, staying almost the same from ntree=10-20 Rising from ntree=20-40 with the highest being at ntree=40 and then decreases from ntree=40-50 might be because of potential overfitting or random variance. For Precision we see a hike from 1-10 then decreases slightly from 10-20 and increases from 20-40, with highest at 40 and decreasing again from 40-50. This indicates that from 40-50 those trees are redundant leading to potential overfitting. From 10-20 we see precision slightly decreasing indicating it is also making more false positive predictions. For recall n=1-5 it increases drastically, 5-10 decreases a bit, 10-20 increases slightly, 20-40 becomes a plateau, 30-40 increases slightly with the highest at 40 and decreases from 40-50. The overall trend shows that initially, the performance benefits from a higher number of trees, but beyond a certain point, the benefit plateaus, and further increases might not be cost-effective or could reduce the model's generalization ability. We can say that the returns are diminishing, where after a certain point, more trees doesn't yield proportional improvements and might lead to overfitting.

WINE DATASET USING GINI INDEX

1. See code and output directories
2. See code and output directories
3. 

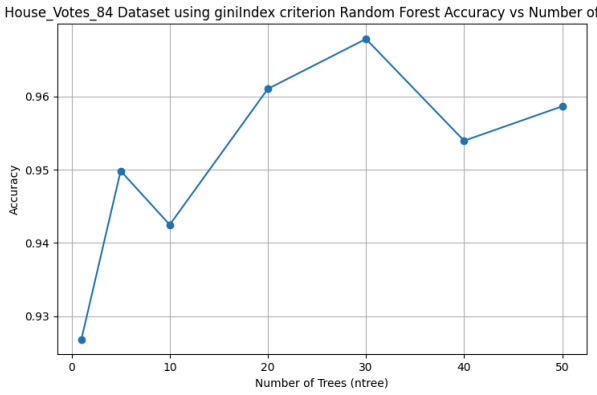
4. We see highest values of parameters with ntree: Accuracy: ntree=50 Precision: ntree=50 Recall: ntree=50 F1 Score: ntree=50

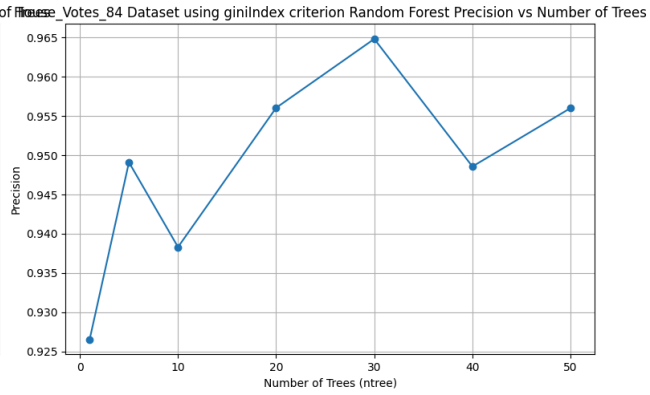
However instead of ntree=50, we can take ntree=30 for all the 4 metrics if we want to use the classifier in real life. This is because from ntree=10-40 we sort of see a plateau with very insignificant increase or decline. From ntree=10-40, the most optimal results are observed with ntree=30. Although it increases from ntree=40-50 but the increase is not that significant and there is no point of adding more trees. So we can use ntree=30 for all the 4 parameters.

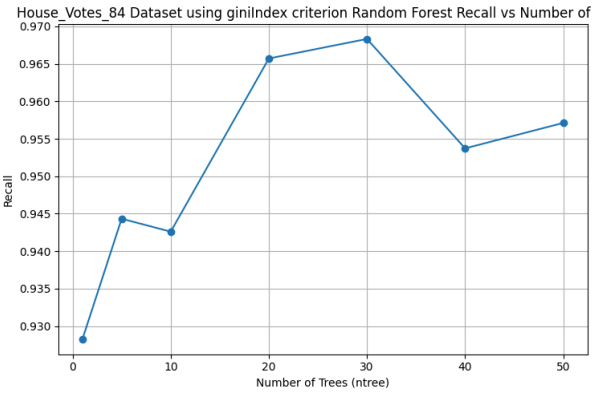
5. The accuracy seems to be relatively sensitive to the increase in ntree. Initially, there is a substantial improvement as ntree increases from 1-10, indicating that a single tree is not sufficient for optimal performance. The accuracy continues to improve overall as more trees are added to the ensemble, albeit with diminishing returns. The substantial gains seem to plateau after ntree=30, and especially after ntree=40, suggesting that beyond this point, the benefit of adding more trees diminishes and is not as cost-effective. Precision fluctuates initially but then stabilizes and improves consistently as ntree increases. There are periods of instability, such as the drop between ntree=5 and ntree=10, but precision generally trends upward. This suggests that precision benefits from the ensemble effect, with more trees providing a more robust consensus on classification, which improves precision. It increases slightly from ntree=10-20, becomes almost constant from 20-40 and then rises again from 40-50. The recall metric sees an initial jump from ntree=1 to ntree=5, followed by a more gradual increase with additional trees. Recall tends to be less sensitive to the number of trees past a certain point (ntree=20/30 in this case), indicating that after a certain threshold, most of the true positive cases are already being captured by the ensemble. The F1 score, which balances precision and recall, generally improves with more trees, though it does show some volatility in the lower ntree values. This improvement suggests that both precision and recall are being enhanced by the addition of more trees to some extent, though the gains in F1 score after ntree=30 are more gradual. Overall, the addition of trees to the random forest tends to improve performance across all metrics, but the rate of improvement decreases as the number of trees increases after ntree=30. There seems to be a point of diminishing returns. This can be due to overfitting, where the model's increased complexity starts capturing noise instead of representing the underlying data distribution more accurately.

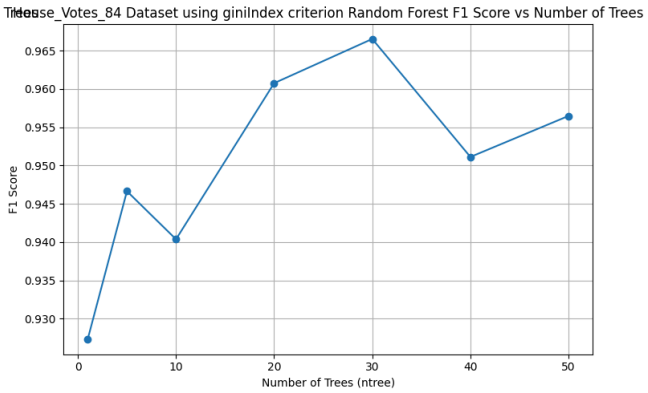
HOUSE OF VOTES DATASET USING GINI INDEX

1. See code and output directories
2. See code and output directories









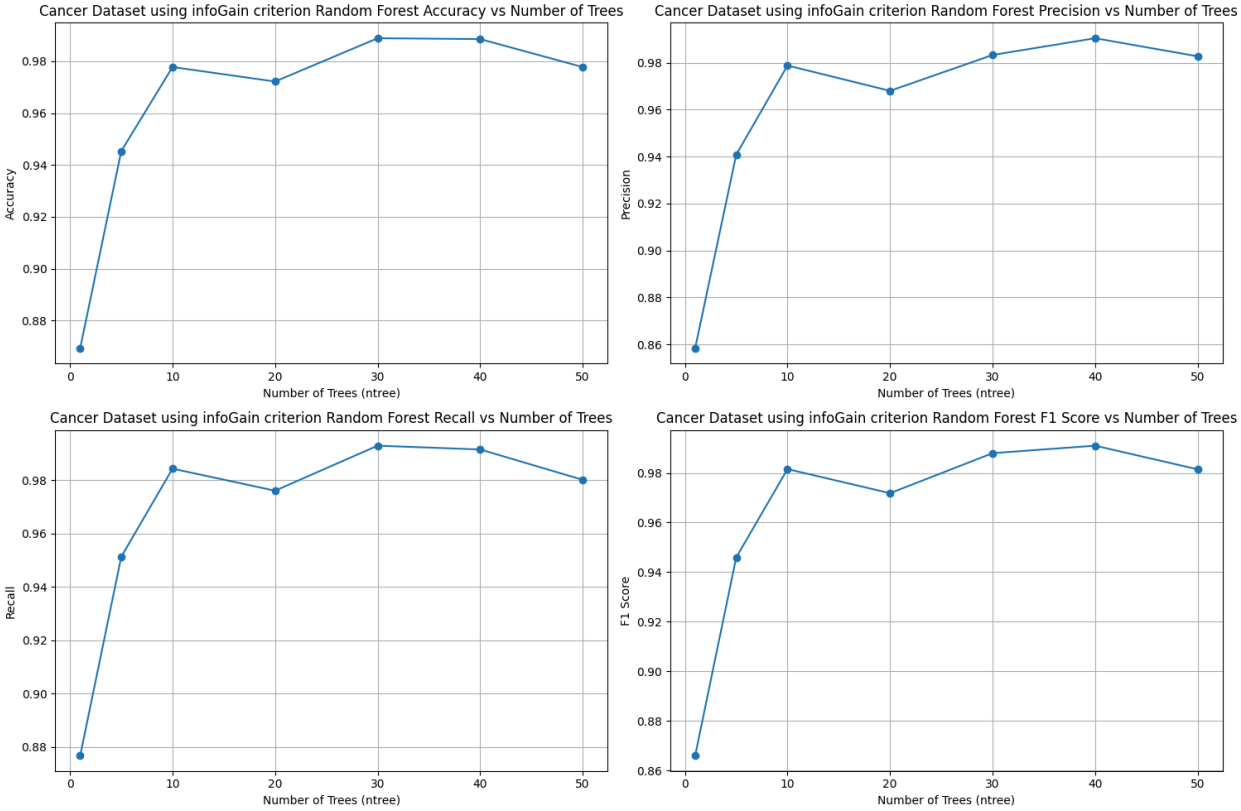
3.

4. We see the highest values of the parameters:
   Accuracy: ntree=30
   Precision: ntree=30
   Recall: ntree=30
   F1 Score: ntree=30
   We observe highest values of the parameters with these ntrees and post ntree=30, we observe performance starts decreasing quite a bit. Until ntree=30, there is an overall increasing trend in all the 4 parameters.
5. The accuracy trend shows some fluctuations as ntree increases, but overall, there is an upward trend until ntree=30. After this, the accuracy slightly decreases, which suggests that the model starts to overfit. The best accuracy is achieved with ntree=30, indicating that adding more trees beyond this number may not be beneficial for the accuracy of the model. Precision seems to fluctuate with the number of trees in the forest. The highest precision is achieved at ntree=30, and there's a notable dip at ntree=40. Precision reflects the model's ability to classify a true instance correctly and is particularly sensitive to false positives. The fluctuation suggests that certain ensembles of trees may be overfitting on the noise, which can decrease precision. Recall steadily increases up to ntree=30 and then starts to show less consistency. The peak at ntree=30 indicates that at this point, the model is able to capture most of the relevant cases without being too restrictive. After ntree=30, the performance in terms of recall starts to vary, indicating instability that could be due to overfitting or a lack of generalization due to the increased complexity of the model. The F1 score, which balances the precision and recall, is highest at ntree=30 and shows variations at other points. It requires a good balance

between recall and precision. The peak at ntree=30 followed by a decrease suggests that this is the optimal balance point for the current dataset. In general, all metrics improved as the number of trees increased up to a certain point, but after ntree=30, the performance gains were not consistent, and in some cases, the performance slightly worsened starting to overfit the data.

BREAST CANCER DATASET USING INFO GAIN

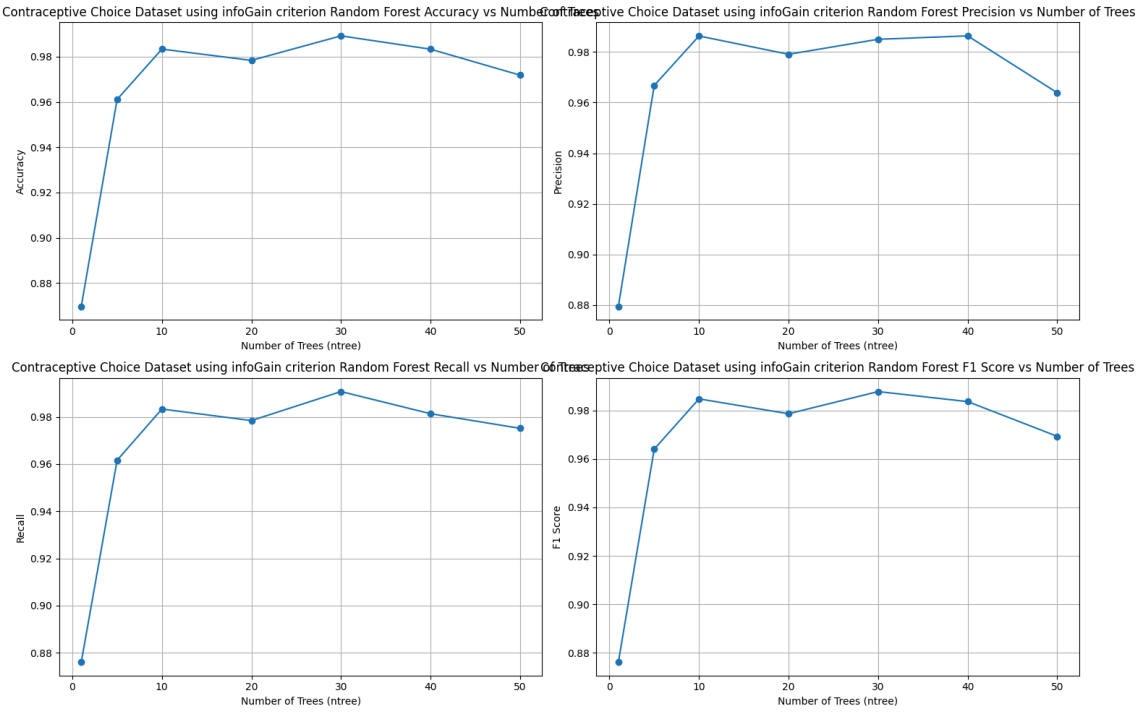
1. See code and output directories
2. See code and output directories



3.
4. Accuracy: ntree=30, precision: ntree=30, recall=30 , f1 score: ntree=30. For medical classifications, it's important that we get high values of these metrics. Accuracy is the highest at 30 i.e. it predicts the instances correctly for ntree=30. High precision means that when the model predicts cancer, it is correct most of the time. Between ntree=30&40, we can pick ntree=30 as the increase in precision is extremely insignificant from 30-40 High recall reduces the chance of missing a true positive (a person having cancer that goes undetected). The highest recall is also at ntree=30. F1 score measures balance between recall and precision. We will choose f1 score as 30 as we get a good balance between precision and recall for 30. There is extremely slight increase in precision and f1 score from 30-40 and extremely slight decline in recall from 30-40. Therefore we can take ntree=30 as our optimal value for all the parameters. Here increasing ntree from 30-40 will not offer us much performance improvement.

5. We can see accuracy, precision, recall, f1 score all increase from ntree=1-10 drops very slightly from 10-20, increases linearly from 20-30, changes very insignificantly from 30-40 and then decreases from 40-50. The increase in the value of the parameters from 1-10 are obvious as 1 tree is not sufficient to correctly classify the data. The gradual increase and decrease from 10-30 are slight and amongst these 30 is where we get a good deal between all the 4 metrics. The model gives almost the same performance from 30-40. And then decreases in performance from 40-50. Deciding f1 score and what parameters to choose from these was a bit hard as from 30-40 trees there was extremely slight increase in performance, so we need to consider if it is worth adding more trees and if the returns are worth or not. That's why since performance is almost the same for both, we choose ntree=30

CONTRACEPTIVE METHOD CHOICE DATASET USING INFO GAIN

1. See code and output directories
2. See code and output directories
3. 

4. For accuracy, precision, recall and f1 score we can select ntree=30 as the value. For accuracy, recall and f1 score we see the highest value at ntree=30. For Precision between ntree=30 and 40 there is very very slight increase, almost negligible and won't affect the performance much. That's why we can choose ntree=30 for the 4 parameters.
5. We see all the parameters rise from ntree=1-10 then from 10-20 there is slight decline, from 20-30, slight increase, 30-40 slight decrease. Deciding between ntree=10 and 30 was tough as all the 4 metrics have almost very comparable results. I decided to go with ntree=30 to rule out any underfitting/overfitting scenarios and noises. After ntree=30, there makes no point to keep increasing the trees as we can get optimal results with 30.