[CM3]

Decision trees are a series of sequential steps designed to answer a question and provide probabilities, costs, or other consequence of making a particular decision. They are simple to understand, providing a clear visual to guide the decision making progress. However, this simplicity comes with a few serious disadvantages, including overfitting, error due to bias and error due to variance.

Random forest is a collection of independently built decision trees with a single, aggregated result, produced at the end of the process (by averaging or "majority rules"). Random forests reduce the variance seen in decision trees by using different samples for training, specifying random feature subsets, and building and combining small (shallow) trees.

A single decision tree is a weak predictor, but is relatively fast to build. More trees give you a more robust model and prevent overfitting. However, the more trees you have, the slower the process.

Gradient boosting is a set of decision trees built one tree at a time combining results along the way. This additive model (ensemble) works in a forward stage-wise manner, introducing a weak learner to improve the shortcomings of existing weak learners.

Random forests and gradient boosting each excel in different areas. Random forests perform well for multi-class object detection and bioinformatics, which tends to have a lot of statistical noise. Gradient Boosting performs well when you have unbalanced data such as in real time risk assessment.

Seeds Dataset:

We see Seeds dataset is a balanced dataset (as observed from the count plot) of 3 different varieties of wheat: Kama, Rosa and Canadian. It is a small dataset (with 198 entries). We can expect all the three classifiers (Decision tree, Random forest, Gradient tree boosting) to yield similar results/accuracies for the dataset as it is relatively small and well balanced. But since decision tree is highly sensitive to changes in the training data, Random forest will produce a more robust model as they combine many independent weak decision tree learners to get a single, aggregated result. Also Gradient tree though may produce similar results, would be computationally complex than Random forest. Random forest tends to perform well for multi-class object detection.

Observing the model performances on the processed dataset:

In Decision tree, we observe maximum accuracy of 91% at maximum depth of 5 on the training set with 10 fold cross validation. The model gives an accuracy of 92.5% on the test set. We see that the accuracy increases from depth 3, maximizing at 5 and further decreases at depth 10. As the value of max_depth increases, the tree becomes deeper with unnecessary conclusions, which leads to overfitting of the model. The lesser the depth of a tree, the more reasonable the model becomes.

In Random Forest, we observe maximum accuracy of 92.95% at the maximum depth of 5 and 50 independent tress on the training set with 10 fold cross validation. The model also gives an accuracy of 92.5% on the test set. A greater number of trees and their depth will increase the computational complexity.

In Gradient Tree Boosting, we observe maximum accuracy of 94.25% with 150 tress on the training set with 10 fold cross validation. The model also gives an accuracy of 92.5% on the test set. We observed that the accuracy of the model increased as the number of trees increased reaching a maximum at 150. This is because as the number of trees increases, the cumulative performance of all the weak learner trees produce a strong learner.

We observe Gradient tree performing better on the train set followed by Random forest and then Decision tree. Whereas, all the models yield similar accuracy on the test set. This can be attributed to the size and balance of the dataset. Comparing the computation complexity and time, we see decision tree is the fastest, followed by random forest and Gradient tree.

Covid Dataset:

We see Covid dataset is a balanced dataset (as observed from the count plot) of 3 different Outcomes: Resolved, Not Resolved and Fatal. It is a large dataset (with 14851 entries). We can expect Gradient Boosting to perform well on the dataset compared to Random forest and Decision tree as this combines a series of weak learner to improve the shortcomings of existing weak learners

Observing the model performances on the processed dataset:

In Decision tree, we observe maximum accuracy of 65.7% at maximum depth of 5 on the training set with 10 fold cross validation. The model gives an accuracy of 65.9% on the test set. We see that the accuracy increases from depth 3, maximizing at 5 and further decreases at depth 10. As the value of max_depth increases, the tree becomes deeper with unnecessary conclusions, which leads to overfitting of the model. The lesser the depth of a tree, the more reasonable the model becomes.

In Random Forest, we observe maximum accuracy of 66.5% at the maximum depth of 5 and 150 independent tress on the training set with 10 fold cross validation. The model gives an accuracy of 66.4% on the test set. A greater value of trees and their depth will increase the computational complexity.

In Gradient Tree Boosting, we observe maximum accuracy of 66.7% with 200 tress on the training set with 10 fold cross validation. The model also gives an accuracy of 67.2% on the test set. We observed that the accuracy of the model increased as the number of trees increased reaching a maximum at 200. This is because as the number of trees increases, the cumulative performance of all the weak learner trees produce a strong learner.

We observe Gradient tree performing better followed by Random forest and then Decision tree. Comparing the computation complexity and time, we see decision tree is the fastest, followed by random forest and Gradient tree. This observation aligns with the expected results, as Gradient Boots yields a better performance on the test set.