* Progress 1 - Screen shot spiral graph

A screenshot of a computer screen

Description automatically generated

* Progress 2 - Screen shot ADA performance

A screenshot of a computer

Description automatically generated

* Analysis 3 - ADA Evaluation
* Compare with other models. Did ADA overfit?

The ADA Boosted Classifier doesn’t seem to be overfitting the data significantly, especially when compared with the Decision Tree model, which had perfect accuracy on the training data. It did not perform as well as the MLP or SVC models though in terms of accuracy and F1 score on training and test data, and so may need some additional tuning to boost performance.

* Progress 4- Random Forest performance

A screenshot of a computer program

Description automatically generated

* Analysis 5 - Evaluation Random Forest
* Compare with other models. Did RF overfit?

Like the Decision Tree model performance, the RF model had perfect accuracy and F1 scores on the training data, and had a significant drop in test data performance, indicating that it is likely overfitting the data. Despite this, the RF model did have the highest performance on the test data among all the models to this point, indicating that it may generalize better than the other models, if given a bit of tuning.

* Progress 6 - Screen shot of Voting performance

A screenshot of a computer

Description automatically generated

* Analysis 7 - Evaluation voting classifier
* Compare with other models. Did the Voting classifier overfit?

The Voting model does have a moderate difference between the test and training performance, with a drop of around 7% in test performance, indicating that it does have some level of overfitting. That said, it is much less severe than the overfitting seen in the Decision Tree and Random Forest models. The Voting model also had the highest test accuracy and F1 scores among all the models, which seems to suggest that it generalizes better than the other models, is less prone to overfitting, and has a strong test performance.

* Analysis 8 - How well did cross validation predict the test set performance?

The cross-validation was a very good predictor of the test set performance for the Decision Tree model, with an average cross-validation accuracy of 74.94% which was very close to the test set accuracy, and a low standard deviation of 0.020, which indicates there was also consistent performance across the different data splits that were generated. This indicates that cross-validation was a good estimate of how our model would perform on unseen data.

* Progress 9 - Screen shot of results table

A screenshot of a computer

Description automatically generated

* Prediction 10 - performance of random forest
* Data set parameter changed and amount
  + - I increased the noise parameter from 0.25 to 0.5.
* Prediction of change in random forest performance
  + - I predict that the increase in noise will decrease the accuracy of the Random Forest model (and performance in general) will decrease, as more introduced noise will make it harder for the model to accurately classify data points.
* Progress 11 - Reran cells

Screen shot of spiral graph

A screenshot of a computer screen

Description automatically generated

Performance of RF

A screenshot of a computer program

Description automatically generated

* Analysis 12 - Prediction results
* How well did you predict?

As predicted, the increased noise resulted in a significant decrease in the accuracy of the model, which fell to 56.38%. Other performance metrics, including precision, recall, and F1 score, also experienced an equally considerable drop. The issue of model overfitting, with perfect training scores again but even lower test scores this time, was exacerbated by the introduction of more noise.

* Explanation

By introducing more noise to the dataset, I increased the variability among the data points. This change is visually evident as well when comparing the spiral graph with noise set at 0.25 to the spiral graph with noise increased to 0.5; the points are spaced further apart after the adjustment, indicating greater randomness introduced to the model. This increase in noise complicates the model's ability to accurately identify patterns and make accurate classifications. As a result, the model's performance metrics decreased significantly. This result seems to highlight the importance of managing noise in a dataset to improve the model's generalization capability and avoid the issue of overfitting.