Module Five Lab: Ensemble Models

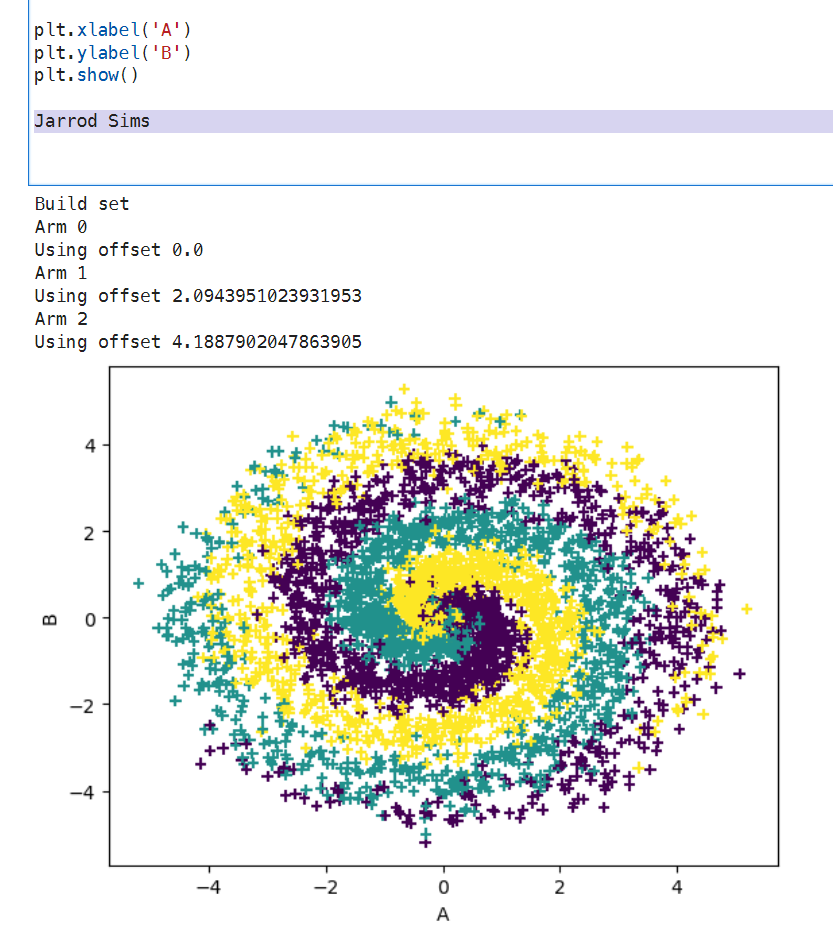
Author: Jarrod Sims

Date: 11/19/2023

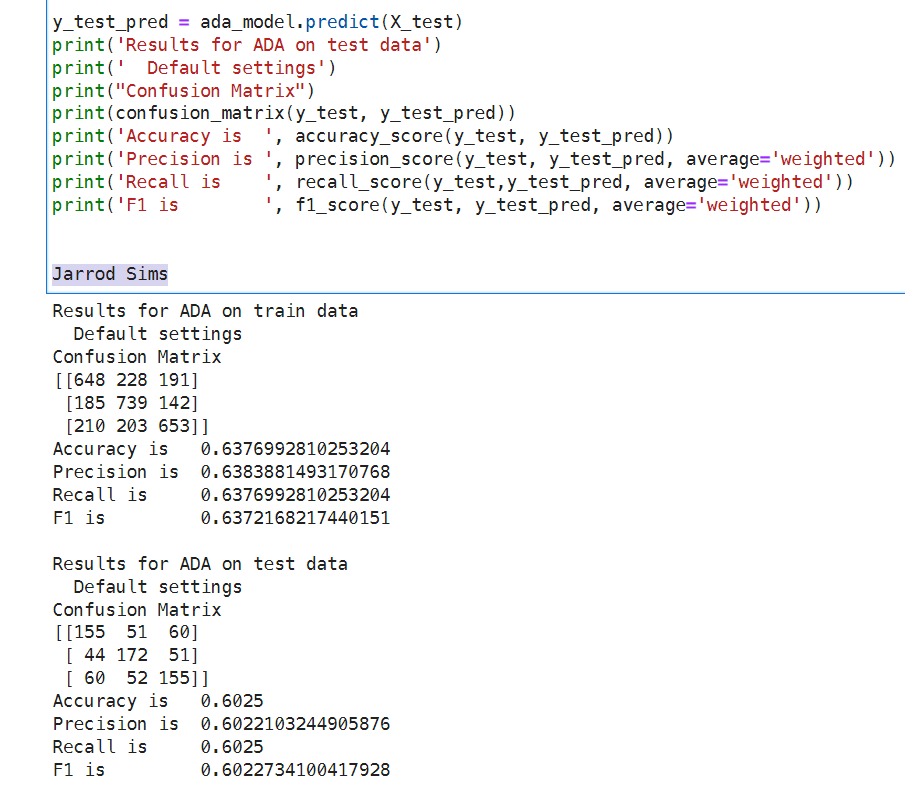
Purpose: This lab demonstrates different combinations of models to (potentially) improve the performance. In particular, we will look at

1. Boosted decision trees – We have multiple decision trees lined up in series. The first decision tree trains on the target with the input data. The second decision tree trains on the errors of the first tree. We continue training new trees on the error of the previous tree until we are satisfied with the performance.
2. Decision tree forest. – We have multiple decision trees lined up in parallel. Each tree trains on a subset of the data. We then combine the results together.
3. Combined heterogenous models. – We use different kinds of models in parallel and combine the results.
4. Cross validation – Create different instances of the same kind of model using all of the data for validation. Each data fold will act as validation once.

## Progress 1 - Progress mark (10) Screen shot of graph of spiral data frame.



## Progress 2- Progress mark (10) Screen shot of train/test performance for ada.



## Progress 3 - Analysis (5)

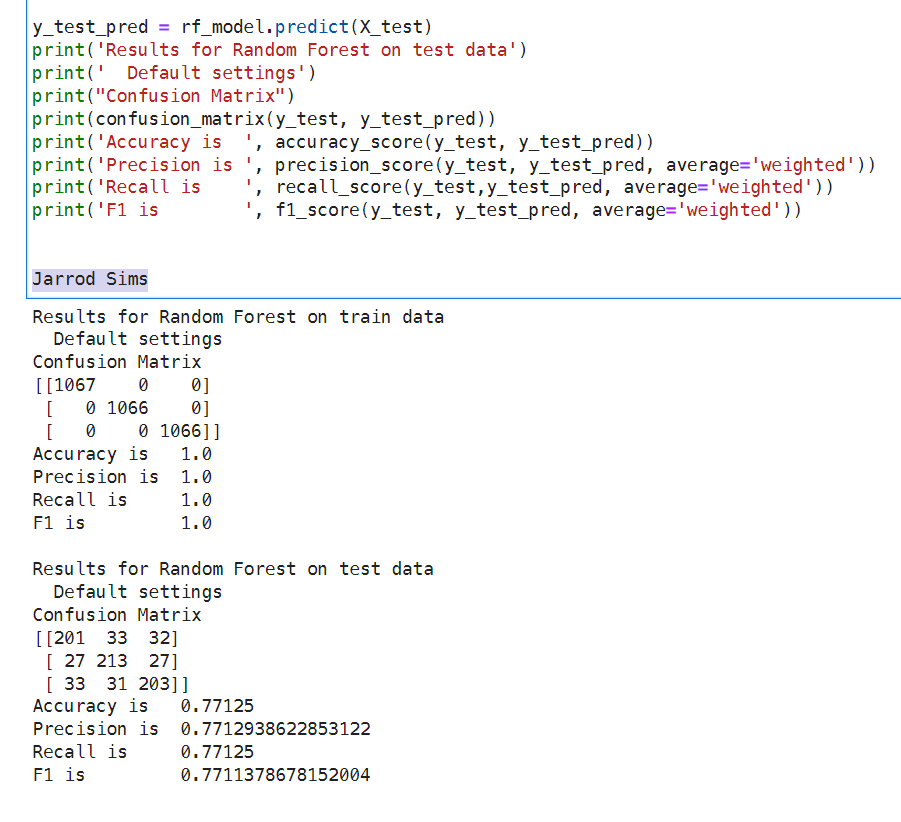
### How does the ADA model compare with the other models?

The AdaBoost model has the worst performance of any of the models with the lowest accuracy (0.638) of any of the models.

### Is the ADA boosted classifier overfitting?

While the ADABoost model does have high accuracy, the performance on the test set is very similar to the training set indicating that there doesn’t appear to be an significant overfitting issue.

## Progress 4 - Progress mark (10) Screen shot of train/test performance for random forest classifier.



## Submission 5 - Analysis (5)

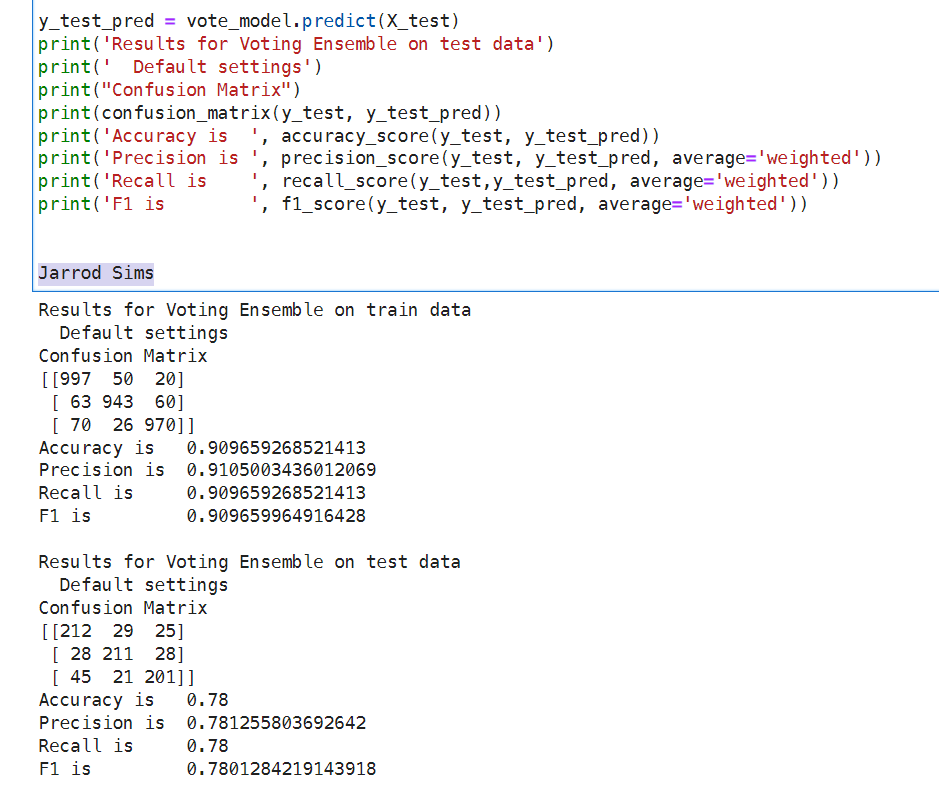
### How does the RF model compare with the other models?

The random forest model performs better than the AdaBoost model with an accuracy of 0.771 compared to 0.6025 for the AdaBoost.

### Is the RF classifier overfitting?

The RF model performs perfectly on the training set but that performance drops in the test set indicating that there may be an overfitting issue.

## Submission 6 – Progress mark (10) Screen shot of train/test performance for voting classifier.



## Progress 7 - Analysis (5)

### How does the voting model compare with the other models?

The Voting Model performs very well on the training set (accuracy = 0.91) and also exhibits good performance on the training set with accuracy, precision, recall, and F1 scores around 0.78 indicating that the model is generalizing to unseen data well. This

### Is the voting classifier overfitting?

There is a slight decline in performance from the training to the test set indicating a mild problem with overfitting but not as severe as the Decision Tree or Random Forest models.

## Submission 8 - Analysis (5) Does the cross validation give us a good idea of how well the decision tree will perform on the test set?

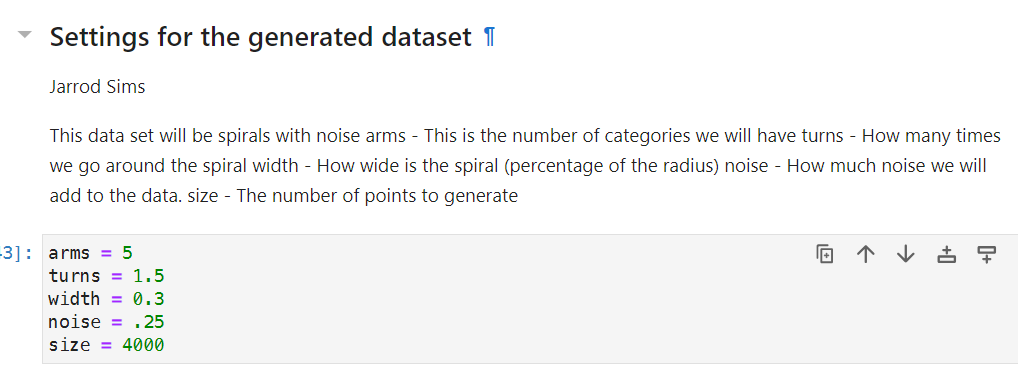
The cross-validation provided a good predictor of the performance of the Decision Tree model with a predicted accuracy of 0.744 vs. 0.736 test set performance.

## Progress 9 - Progress (5) Screen shot of results table.

## Submission 10 - Prediction (5)

### Value being changed and how much

I am going to change the number of arms from 3 to 5.

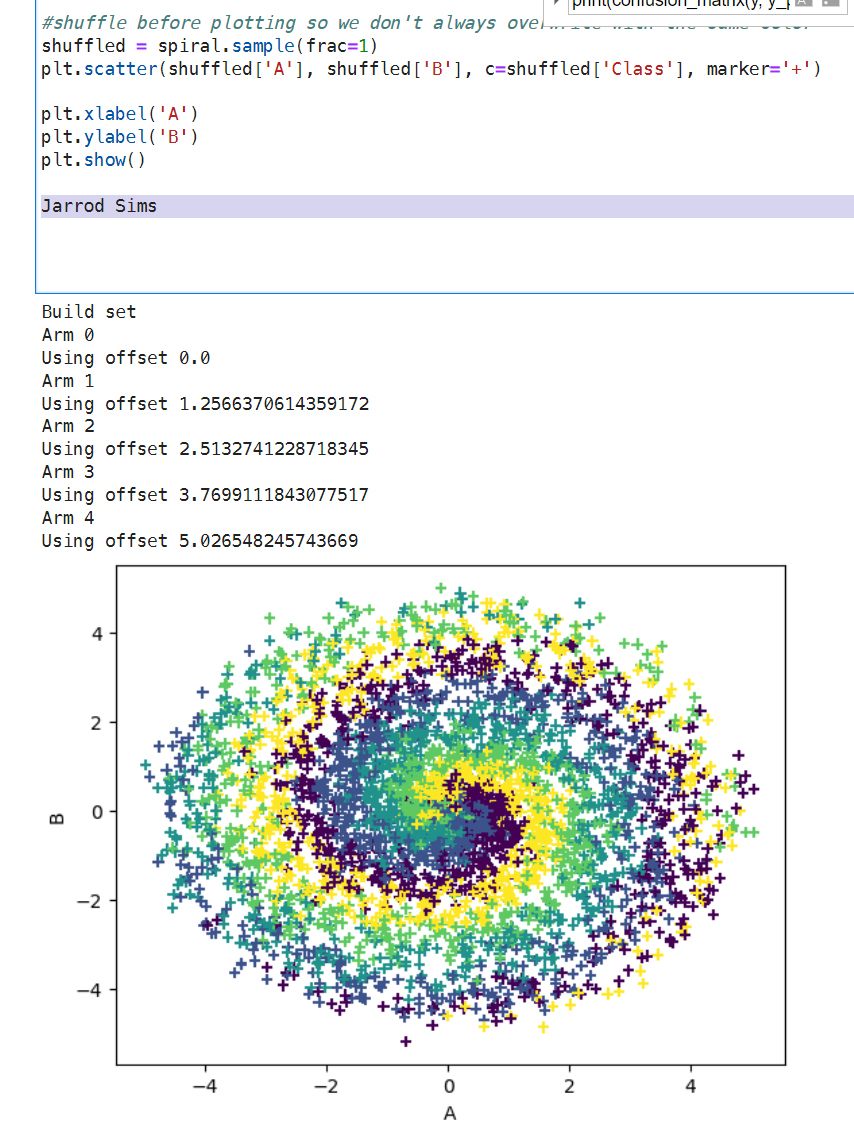


### Predicted change in accuracy

I predict that this will lower the accuracy of the RF model.

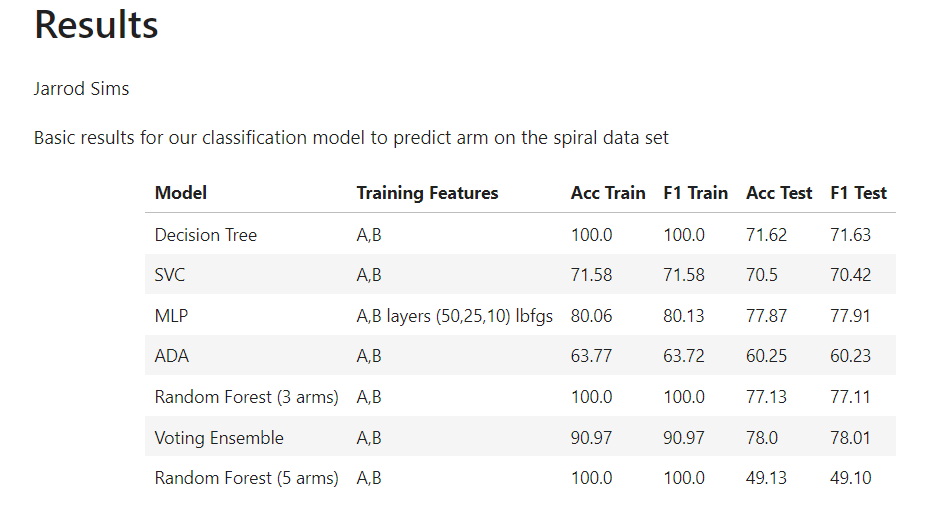
## Progress 11 - Progress (10)

### Screen shot data set graph



### Actual Random Forest performance

The Random Forest model based on the spiral set with 5 arms performed much worse than the model based on the spiral with 3 arms.



## Submission 12- Analysis (5)

### Did your prediction match the actual performance.

I predicted that the random forest model based on the spiral set with three arms would have a higher accuracy than the model based on the set with five and I was correct (Acc Test 77.13 vs. 49.13).

### Propose an explanation

Increasing the number of arms of the dataset increases the complexity of the model and makes the classifier’s job much more difficult. This is true because there is more overlap between the arms of the dataset and less definite boundaries between the arms.