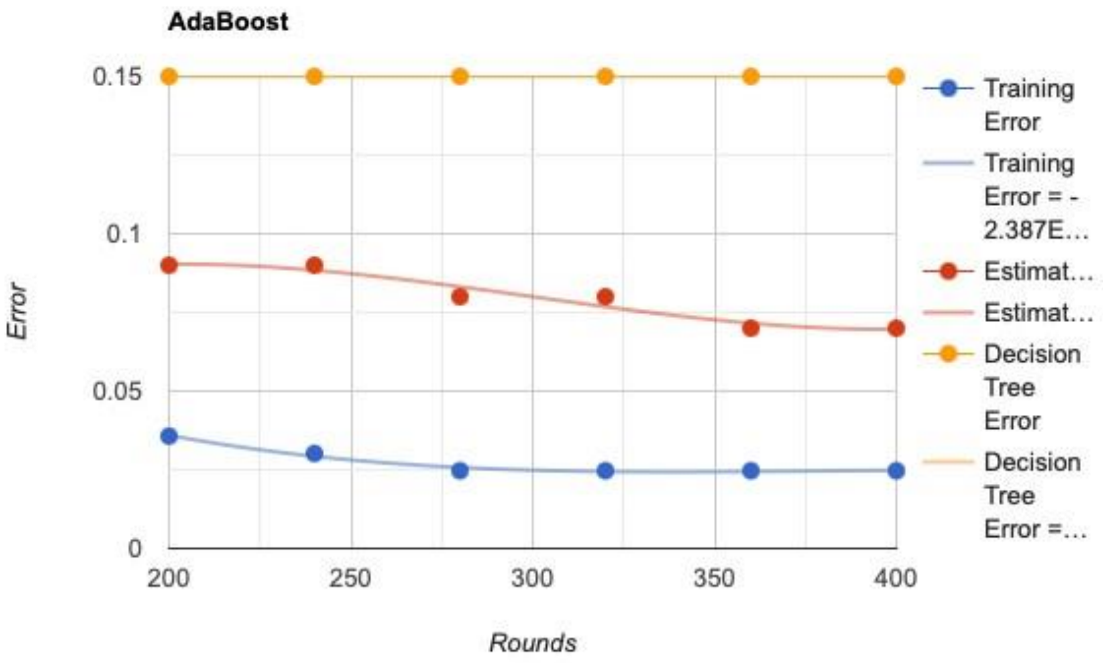
WRITTEN REPORT

OVERVIEW: The goal of this assignment was to use Gini Index to implement a decision stump which would be used as weak learners in the implementation of the boosting algorithm – AdaBoost.

Steps:

- ✦ Data Preprocessing / Importing ○ Converted the target values to 1 and -1 for AdaBoost.
  ○ Converted both to arrays
- ✦ Create a decision tree from scratch with depth = 1 (decision stump) ○ Used Gini Index as the backbone.
- ✦ I spent about 80% of my time implementing and making my decision stump work but it wasn't working as intended. I still wanted to give my best on AdaBoost algorithm, so I used sklearn's Decision Tree Classifier for that.
- ✦ Use the decision stump as a weak learner to implement AdaBoost.
- ✦ The AdaBoost was run on 6 different number of rounds, the results of the training/test errors that I received are shown in the plot below.



As can be seen from the plot above, both the training set error and test set error exhibit expected trends given the value of number of iterations AdaBoost was run. Both the training and test errors are better than the decision stump error which shows how boosting a weak classifier can result in a more accurate model. The results from the plot have been summarized in the table below:

	ROUNDS	CLASSIFICATION ERROR – TRAINING DATA	CLASSIFICATION ERROR – TEST DATA	TIME
1	200	0.0357	0.09	2s
2	240	0.0302	0.09	3s
3	280	0.0247	0.08	3s
4	320	0.0247	0.08	3s
5	360	0.0247	0.07	4s
6	400	0.0247	0.07	4s

- As the above table and plot shows, the classification error on both the training set and test set decrease as the number of AdaBoost rounds increases.
- After a certain number of rounds, the graph begins to level off. However, for the training error, if there are enough weak classifiers combined, then the training error would be observed to go reach 0 eventually.
- If AdaBoost is run on a larger number of rounds, test error will gradually start to increase if the model is overfitting. In this case, the test error would be seen decreasing at first but then increases after a certain point.
- Minimum number of rounds were 200, but I tried running AdaBoost for even fewer rounds and it was evident from the results that in the first few rounds, the training error decreases drastically. Compared to that, the training error for rounds after 200 decreased at a much slower rate.
- Nonetheless, it was clear from the results that combining several weak learners and boosting them massively increases the effectiveness and accuracy of a model.

Resources used
For concept: **https://hastie.su.domains/Papers/samme.pdf**
For the plot: **https://www.rapidtables.com/tools/scatter-plot.html**