This project focuses on improving the accuracy of existing Machine Learning Data Models using the AdaBoost classifier of the Ensemble Method of boosting.

The aim is to classify emails as either "Spam" or "Ham". It uses data from the Spam Base Dataset provided through UC Irvine's Machine Learning Repository.

There are two datasets: spambase and spambase\_names.

Execution order is as follows:

1. Linear Regression
2. KNeighbours Classifier
3. Decision Tree
4. Logistic Regression+AdaBoost Classifier
5. Decision Tree +AdaBoost Classifier

Upload both datasets first and then run the files.

This project also does AdaBoost Classifier implementation in Python from scratch.

We first implemented a few ML algorithms like KNN, Decision Tree, Linear Regression, and Logistic Regression on our dataset to see the accuracy we obtain from each. We find the accuracies in this order:

Linear regression: 54%

KNN: 85%

Decision tree: 89%

Logistic regression: 91%

We then tried to apply the AdaBoost classifier to each of them and found that AdaBoost doesn’t apply to Linear regression and KNN models. This is because Linear regression works on Binary data that could not be segregated into weights.

So, then we applied AdaBoost to Logistic regression and Decision Tree and found that it reduced the accuracy of Logistic regression but increased the accuracy of Decision tree to up to 94%.

Therefore, we were finally able to increase the accuracy by experiment and implementation.

SUBMITTED BY:

ARUSHI 102003448

TANISHA PARKASH 102003447

2CO18