1. First import all the libraries that needed in the project.
2. Download the data and read the csv file.
3. Read the data and put it in a data frame (df).
4. Dropped the column subject because it’s a text and it wont affect the classification models.
5. I took sample of 8000 rows of each class where there is 13 class so the total sample is 104000 of the original data set 1,200,000 rows.
6. Cleaned the data from duplicates and nulls,
7. The distribution of the classes wasn’t balance so I took 8k rows of every class to make the data balanced.
8. So I shuffled the data to start learning the models.
9. Splitted the data into train and test (80,20) to train the models.
10. KNN:

Accuracy: 0.963173076923077

Precision: 0.9635487488501266

Recall: 0.963173076923077

F1 Score: 0.9598927822450087

Cross-validation scores : [0.95]

1. Linear Regression:

Mean squared error: 10.65

Root mean squared error: 3.26

1. Neural Network:

Accuracy: 0.95

Precision: 0.95

Recall: 0.95

F1-score: 0.95

1. SVM:

Accuracy: 0.96

Precision: 0.96

Recall: 0.96

F1 Score: 0.96

1. Logistic Regression:

accuracy: 0.54

precision: 0.52

recall: 0.54  
f1: 0.52

In conclusion The best results was from the SVM model with the highest accuracy.

It’s the most suitable model for this data because it can learn complex data and understand the relationship between it and adapt which will give us an accurate classification and regression.