Classification and Clustering of Amazon Review Text

A. Introduction

In machine learning, binary and multiple classifications are commonly tackled by models such as neural networks, naive Bayes, and random forests. Alongside these techniques, in the context of clustering, the k-means algorithm has proven to be an effective way to segment and analyze data in an unsupervised manner [5]. Currently, the most widely used text classification techniques include Decision Tree (DT), Sup- port Vector Machine (SVM), K Nearest Neighbors (KNN), Naıve Bayes (NB), and hidden Markov model (HMM). These five algorithms are acknowledged as straightforward and effective methods for text classification [1].

The pipeline followed is same for any machine learning model. The data was already provided, hence, data collection was already there. Data processing was done on the data based on certain observations followed by splitting of training data into validation and training data. This was done to evaluate the models based on the given metrics. The metrics used for clustering were - silhouette score and adjusted rand score. On the other hand, the metrics used for clustering models were - F1 macro score, accuracy score, confusion matrix, and AUROC score. These have been discussed in detail in the following sections.

B. Literature Review

In the context of low-dimensional data (i.e. when the number of covariates is small compared to the sample size), logistic regression is considered a standard approach for binary classification [3]. In some papers, researchers have conducted experiments to analyse the performance of Random Dorests and Logistic Regression models and found that Random forests tend to perform better in majority of the cases. In this experiment, similar analysis has been conducted alongwith with SVM. SVM learning is one of many ML methods. Compared to the other ML methods SVM is very powerful at recognizing subtle patterns in complex datasets [2].

Cross-validation type of methods have been widely used to facilitate model estimation and variable selection. In this work, we suggest a new K-fold cross validation procedure to select a candidate ‘optimal’ model from each hold-out fold and average the K candidate ‘optimal’ models to obtain the ultimate model [4].

C. Methodology

Dataset : In the case of classification, we were provided with a training dataset of 16.7 MB and a test dataset of 2.4 MB. We split the training data into training (80%) and validation (20%) sets for the training and validation tasks during the cross-validation process. However, for clustering, we used the test dataset instead, where the overall(star) data present in the training dataset was withheld.

The following three models were used for the binary classification.

- Logistic Regression
- Support Vector Machine
- Random Forest

The following three models were used for the multi class classification.

- Logistic Regression
- Linear SVM
- Random Forest

For clustering problem, K-Means Algorithm was used.

The following pipeline was followed for the experimentation-

1. Data Collection: The data was already provided to us in CSV format which was directly imported into the workspace. This data was pre-divided into training and test data.
2. Data Preprocessing: In this step, we clean our data where we remove the irrelevant features and fill the NaN values with some default numbers. Alongwith this vectorization of text data has been carried out using TFIDFvectorizer from SkLearn. As a result, some columns were added into the data. This step has been done in separate functions for test data, training data for binary classification and training data for multi class classification.
3. Data splitting: In this process, we have divided the training data into validation and training test. This is further used to calculate the performance metrics.
4. Hyperparameter Tuning: The hyperparameter tuning has been carried out using GridSearchCV from SkLearn where a combination of alteast 7 parameters have been fed to the grid.
5. Training: After this training has been carried out for all 3 models for all 4 cutoffs. This results in a total of 12 models for binary classification and 3 models for multi class classification and one model for clustering.
6. Results: The results have been displayed given some performance metrics.

Performance Metrics: For the analysis part, we have focused on the F1 Macro score and ROC AUC score and additionally we have also calculated confusion matrix. Since hyperparameter tuning was carried out, we have also displayed the best hyperparameters that give us these scores for both binary and multiclass classification. Likewise, for clustering we referred to Silhouette score and Rand index to analyze the quality of clustering. Most of the analysis here are done on these scores.

D. Experiment and Results

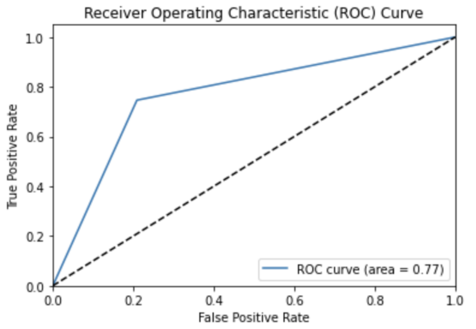
D.1. Binary Classification

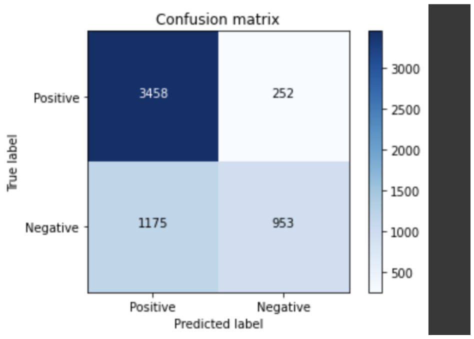
D.1.1 Cutoff - 1

1. Model - Logistic Regression

The results for this model are as follows -

F1 score for Logistic Regression is 0.7004077969825029
AUROC for Logistic Regression is 0.7686280006412592
Accuracy Score for Logistic Regression is 0.7578262494169419
Best hyperparameters: ’C’: 1, ’penalty’: ’l2’



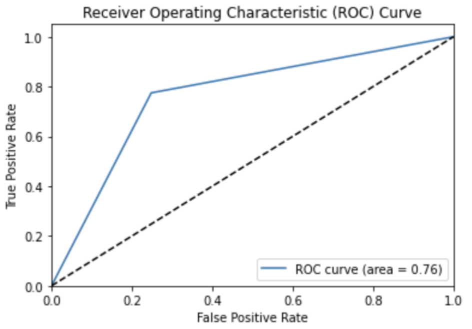


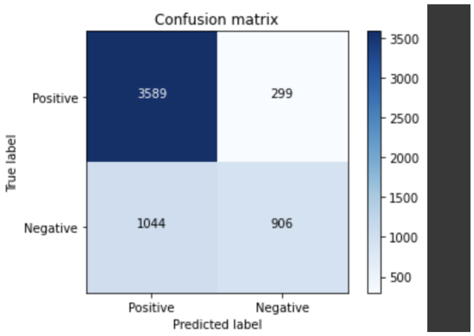
2. Model - SVM

The results for this model are as follows -

F1 score for SVM is 0.7083579284218294

AUROC for SVM is 0.7632636337012215 Accuracy Score for SVM is 0.7609952088988673 Best hyperparameters: ’C’: 10, ’loss’: ’squaredHinge’, ’penalty’: ’l2’

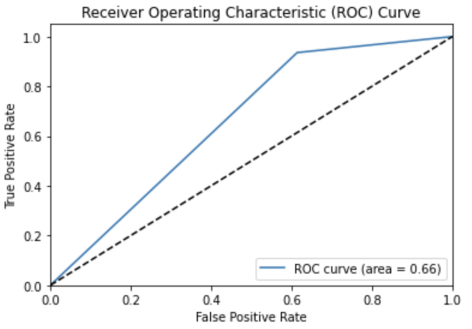


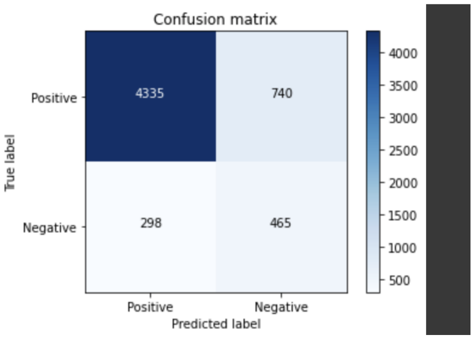


3. Model - Random Forest

The results for this model are as follows -

F1 score for Random Forest is 0.682819424764087
AUROC for Random Forest is 0.6607854709986897
Accuracy Score for Random Forest is 0.8332404137811977
Best hyperparameters: ’maxDepth’: None, ’nEstimators’: 200



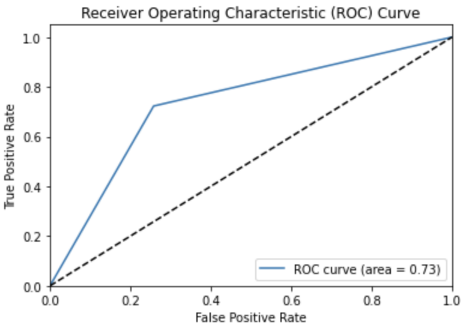


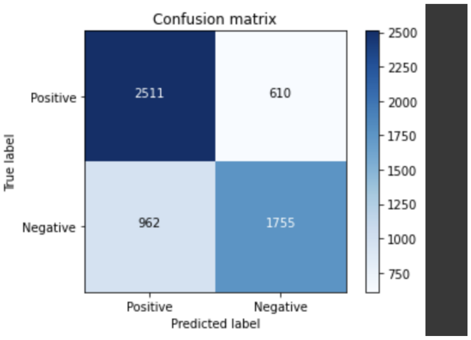
D.1.2 Cutoff - 2

1. Model - Logistic Regression

The results for this model are as follows -

F1 score for Logistic Regression is 0.7261372096346039
AUROC for Logistic Regression is 0.7325389641261583
Accuracy Score for Logistic Regression is 0.7350009091678585
Best hyperparameters: ’C’: 1, ’penalty’: ’l2’

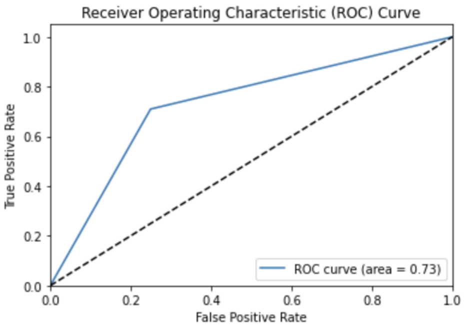


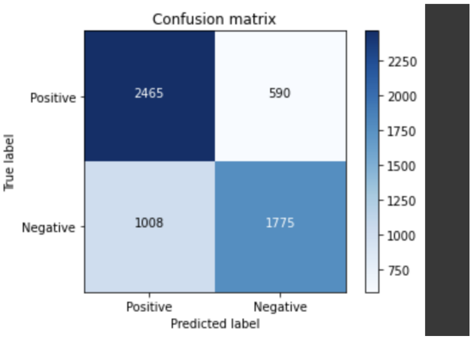


2. Model - SVM

The results for this model are as follows -

F1 score for SVM is 0.7223982614607614
AUROC for SVM is 0.7301447773795923
Accuracy Score for Logistic Regression is 0.7350007881215448
Best hyperparameters: ’C’: 10, ’loss’: ’squaredHinge’, ’penalty’: ’l2’

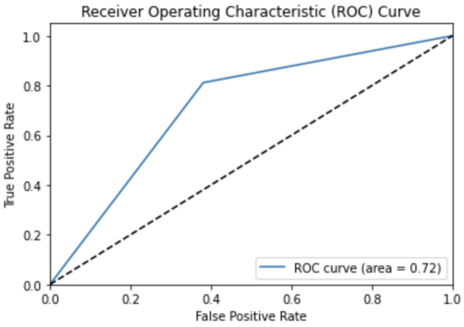


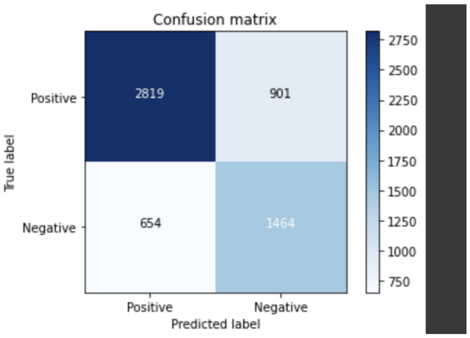


3. Model - Random Forest

The results for this model are as follows -

F1 score for Random Forest is 0.7184758312284611
AUROC for Random Forest is 0.715358832771565
Accuracy Score for Logistic Regression is 0.7350435889976735
Best hyperparameters: ’maxDepth’: None, ’nEstimators’: 200



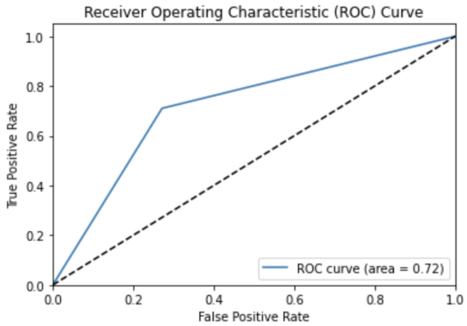


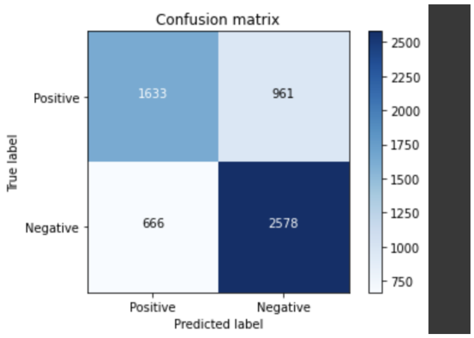
D.1.3 Cutoff - 3

1. Model - Logistic Regression

The results for this model are as follows -

F1 score for Logistic Regression is 0.7138098971235207
AUROC for Logistic Regression is 0.7193815977830331
Accuracy Score for Logistic Regression is 0.7284484411149519
Best hyperparameters: 'C': 0.1, 'penalty': 'l2'

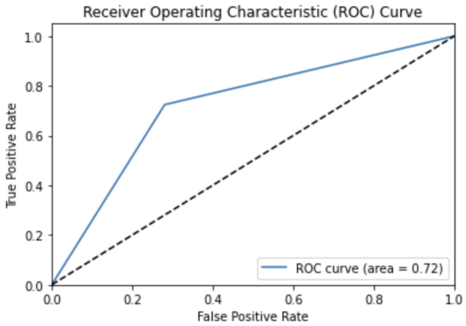


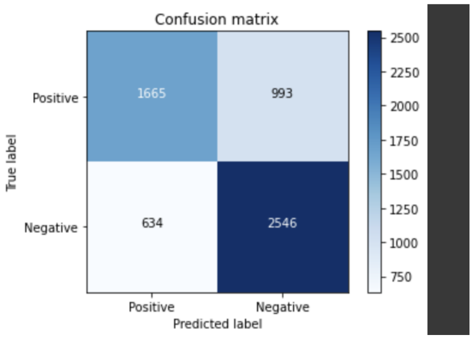


2. Model - SVM

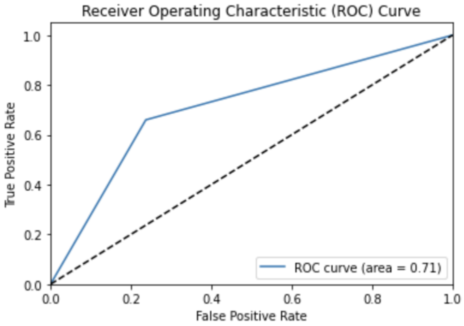
The results for this model are as follows -

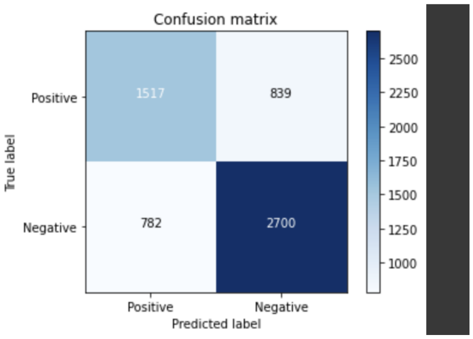
F1 score for SVM is 0.7148140776566252
AUROC for SVM is 0.721820094268046
Accuracy Score for Logistic Regression is 0.7265641021442497
Best hyperparameters: 'C': 0.1, 'loss': 'squaredHinge', 'penalty': 'l2'





3. Model - Random Forest
The results for this model are as follows -
F1 score for Random Forest is 0.7104467478338348
AUROC for Random Forest is 0.7113897451144342
Accuracy Score for Logistic Regression is 0.7271636775489944
Best hyperparameters: 'maxDepth': 20, 'nEstimators': 200



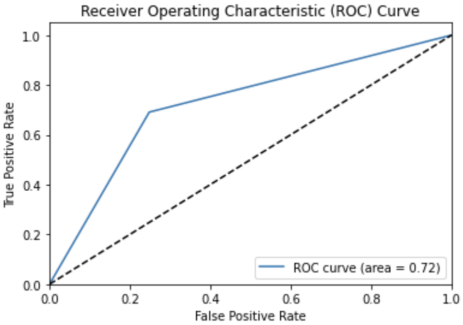


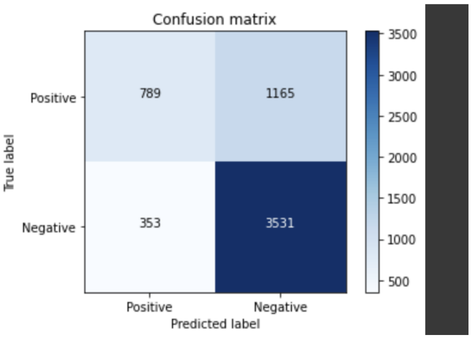
D.1.4 Cutoff - 4

1. Model - Logistic Regression

The results for this model are as follows -

F1 score for Logistic Regression is 0.6663834227787717
AUROC for Logistic Regression is 0.721404847289641
Accuracy Score for Logistic Regression is 0.7401400184527401
Best hyperparameters: ’C’: 0.1, ’penalty’: ’l1’

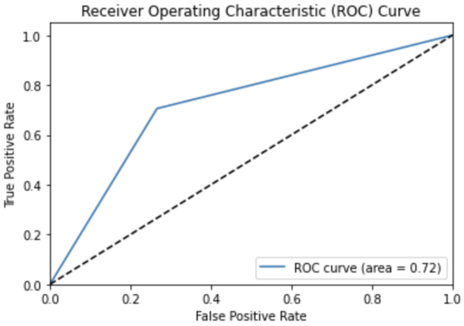


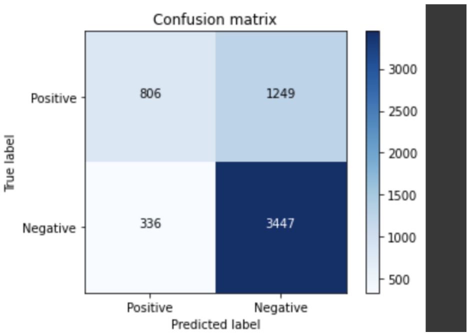


2. Model - SVM

The results for this model are as follows -

F1 score for SVM is 0.6586451437566981
AUROC for SVM is 0.7199041476592964
Accuracy Score for Logistic Regression is 0.7265643882537186
Best hyperparameters: ’C’: 10, ’loss’: ’squaredHinge’, ’penalty’: ’l2’

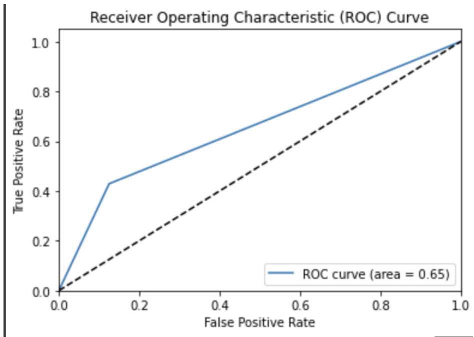


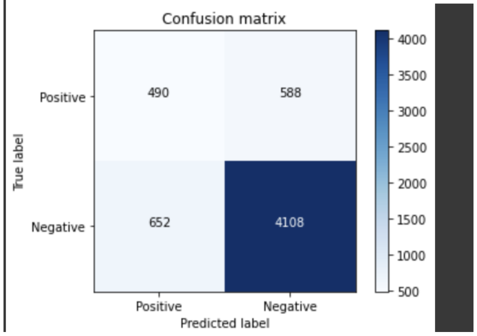


3. Model - Random Forest

The results for this model are as follows -

F1 score for Random Forest is 0.6551538848493164
AUROC for Random Forest is 0.6519294283318963 Accuracy Score for Logistic Regression is 0.786304694611815
Best hyperparameters: ’maxDepth’: None, ’nEstimators’: 200



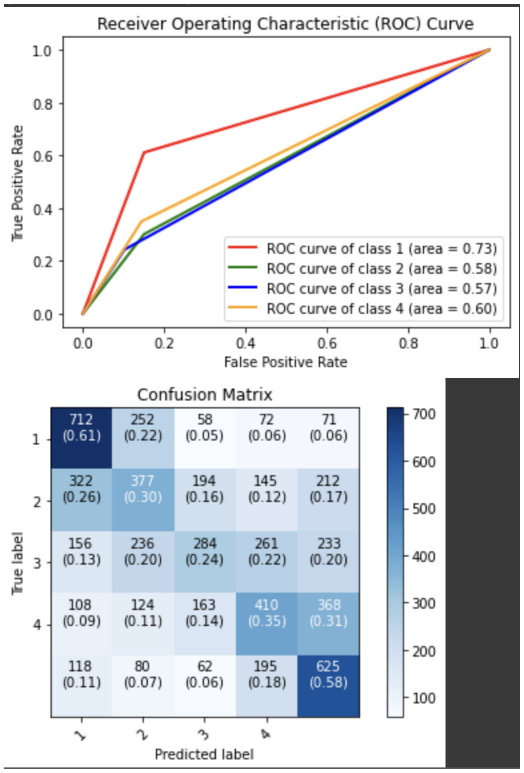


D.2. Multi Class Classification

1. Model - Logistic Regression

The results for this model are as follows -

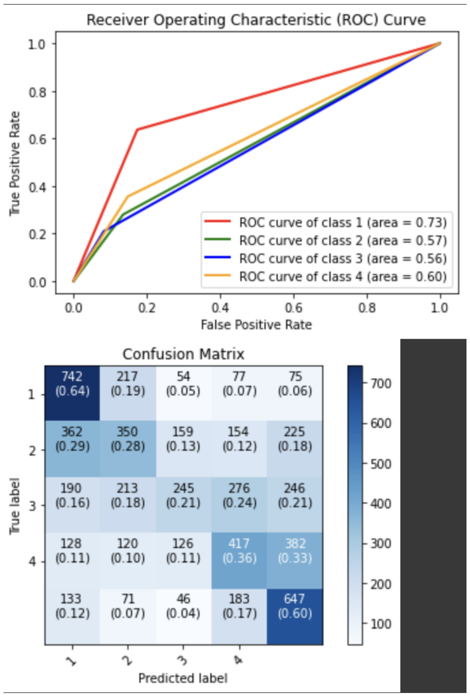
F1 score for Logistic Regression based Multi class classification is 0.40345989362474466
Score for Logistic Regression based Multi class classification is 0.4249069455389466
Best hyperparameters: ’$logisticregressionC$’: 1, ’$logisticregressionPenalty$’: ’l2’



	1	2	3	4	
1	712 (0.61)	252 (0.22)	58 (0.05)	72 (0.06)	71 (0.06)
2	322 (0.26)	377 (0.30)	194 (0.16)	145 (0.12)	212 (0.17)
3	156 (0.13)	236 (0.20)	284 (0.24)	261 (0.22)	233 (0.20)
4	108 (0.09)	124 (0.11)	163 (0.14)	410 (0.35)	368 (0.31)
	118 (0.11)	80 (0.07)	62 (0.06)	195 (0.18)	625 (0.58)

2. Model - SVM

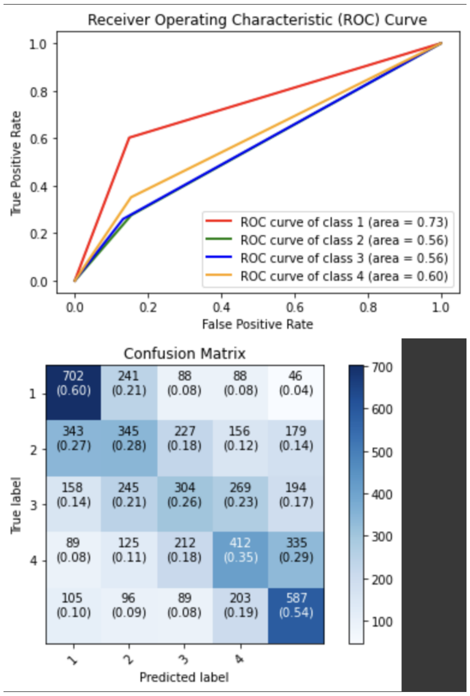
F1 score for Linear SVC based Multi class classification is 0.3972311148184553
Score for Linear SVC based Multi class classification is 0.4243502461999572
Best hyperparameters: ’linearsvcC’: 0.1, ’linearsvcPenalty’: ’l1’



	1	2	3	4	
1	742 (0.64)	217 (0.19)	54 (0.05)	77 (0.07)	75 (0.06)
2	362 (0.29)	350 (0.28)	159 (0.13)	154 (0.12)	225 (0.18)
3	190 (0.16)	213 (0.18)	245 (0.21)	276 (0.24)	246 (0.21)
4	128 (0.11)	120 (0.10)	126 (0.11)	417 (0.36)	382 (0.33)
	133 (0.12)	71 (0.07)	46 (0.04)	183 (0.17)	647 (0.60)

3. Model - Random Forest

F1 score for Linear SVC based Multi class classification is 0.396337314526243
Score for Linear SVC based Multi class classification is 0.409019073906747
Best hyperparameters: ’randomforestclassifierMaxDepth’: None, ’randomforestclassifierMinSamplesSplit’: 10, ’randomforestclassifierNEstimators’: 200



	1	2	3	4	
1	702 (0.60)	241 (0.21)	88 (0.08)	88 (0.08)	46 (0.04)
2	343 (0.27)	345 (0.28)	227 (0.18)	156 (0.12)	179 (0.14)
3	158 (0.14)	245 (0.21)	304 (0.26)	269 (0.23)	194 (0.17)
4	89 (0.08)	125 (0.11)	212 (0.18)	412 (0.35)	335 (0.29)
	105 (0.10)	96 (0.09)	89 (0.08)	203 (0.19)	587 (0.54)

D.3. Clustering Algorithm: KMeans

The Silhouette Score comes out to be 0.599190380028831 and the Adjusted Rand score comes out to be 0.006850490923692817 which is above the given baseline.

E. Conclusion

In this study, various models were used for classification and one for clustering. Overall, it was observed that all three algorithms give similar results. Baselines were crossed for the most part. This has been a great learning experience having various challenges.

References

[1] Ahmed Hussain Aliwy and Esraa H. Abdul Ameer. Comparative study of five text classification algorithms with their improvements. 2017. 1

[2] S Aruna and SP Rajagopalan. A novel svm based cssffs feature selection algorithm for detecting breast cancer. *International journal of computer applications*, 31(8):14–20, 2011. 1

[3] Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1):270, Jul 2018. 1

[4] Yoonsuh Jung and Jianhua Hu. A k-fold averaging cross-validation procedure. *J Nonparametr Stat*, 27(2):167–179, Feb. 2015. 1

[5] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020. 1