

**PROJECT REPORT ON:**

**“Emails Spam Classifier”**

**SUBMITTED BY**

**KRISHNA TEJA DINAVAHI**

**FLIPROBO SME:**

**GULSHANA CHAUDHARY**

**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms.Gulshana Chaudhary (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

**Contents:**

1. **Introduction**
   * Business Problem Framing
   * Conceptual Background of the Domain Problem
   * Review of literature
   * Motivation for the Problem undertaken

**2.Analytical Problem Framing**

* + Mathematical/ Analytical Modelling of the Problem
  + Data Sources and their formats
  + Data Pre-processing Done
  + Data Input – Logic – Output Relationships
  + Hardware, Software and Tools Used

**3. Data Analysis and Visualization**

**4. Model Developments and Evaluation**

* + The model algorithms used
  + ROU AUC curve
  + Interpretation of the result
  + Hyperparameter tuning

**5. Conclusions**

* + Key Finding and conclusions
  + Limitation of this works and scope for future works

**1.INTRODUCTION**

1.1Business Problem Framing:

The SMS Spam Collection is a set of SMSs tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.

What is a Spam Filtering? Spam Detector is used to detect unwanted, malicious and virus infected texts and helps to separate them from the no spam texts. It uses a binary type of classification containing the labels such as ‘ham’ (no spam) and spam. Application of this can be seen in Google Mail (GMAIL) where it segregates the spam emails in order to prevent them from getting into the user’s inbox.

The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

* 1. Conceptual Background of the Domain Problem

Predictive modelling, Classification algorithms are some of the machine learning techniques used along with the various libraries of the NLTK suite for Classification of comments. Using NLTK tools, the frequencies of malignant words occurring in textual data were estimated and given appropriate weightage, whilst filtering out words, and other noise which do not have any impact on the semantics of the comments and reducing the words to their base lemmas for efficient processing and accurate classification of the comments.

* 1. Review of Literature

Nowadays, email messaging is a way with which people communicate important messages to each other using Internet. It’s a very common way through which clients communicate among themselves formally. Now a days the extent to which these emails are sent has been increasing rapidly. Along with these emails, Spam emails are also sent in bulk through different platforms. These spam emails are usually difficult to recognize and it is the major problem that is being faced by the users. Spam consumes almost 98% of billions of emails sent every day. Due to the presence of different email filtering systems already present in the market, Spammers have become aware of these systems. Therefore, Spammers are trying different ways to send spam or junk mails to a number of users. One of them is by sending spam images and pdfs. For this kind of spam emails, presently there are not very effective systems present in the market. This paper illustrates a survey of different existing email classification system which can classify the email as ham or spam.

* 1. Motivation for the Problem Undertaken

Email has become one of the most important forms of communication. In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent each day worldwide. Spam is one of the major threats posed to email users. In 2013, 69.6% of all email flows were spam. Links in spam emails may lead to users to websites with malware or phishing schemes, which can access and disrupt the receiver’s computer system. These sites can also gather sensitive information from. Additionally, spam costs businesses around $2000 per employee per year due to decreased productivity. Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society.

So that we need to do spam filtering so user more user friendly. From above model building we got the Complement Naive Bayes Classifier is a best model deciding whether the emails have spam or not**.**

**2.Analytical Problem Framing**

* 1. Mathematical/ Analytical Modelling of the Problem

Various Classification analysis techniques were used to build predictive models to understand the relationships that exist between user emails and the corresponding user label.

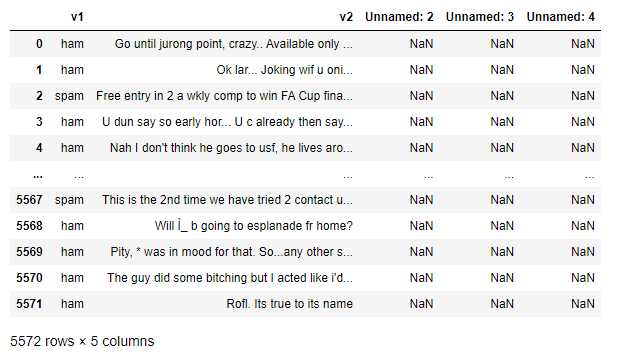
The user emails are collected, processed and normalized. Based on the context of the reviews on various items, with similar label, prediction of the label for a given email can be made based on similar email which already have corresponding label.

In order to predict label for user emails, models such as Logistic regression, Random Forest Classifier Boost Classifier, Extreme Gradient Boost Classifier, and Complement Naïve Bayes Classifier.

2.2 Data Sources and their formats

The dataset is provided by Flip Robo which is in the format csv. This dataset give use for exercise. The SMS Spam Collection is a set of SMSs tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.

**Data Descriptions**

****

2.3 Data Pre-processing Done

* Rows with null values were removed.
* Columns: Unnamed: 0(just a series of numbers) was dropped since it doesn't contribute to building a good model for predicting the target variable values.
* The train and test dataset contents were then converted into lowercase.
* Punctuations, unnecessary characters etc were removed, currency symbols, phone numbers, web urls, email addresses etc were replaced with single words
* Tokens that contributed nothing to semantics of the messages were removed as Stop words. Finally retained tokens were lemmatized using WordNetLemmatizer().
* The string lengths of original comments and the cleaned comments were then compared.

2.4 Data Inputs - Logic - Output Relationships

The comment tokens so vectorized using TfidVectorizer are input and the corresponding rating is predicted based on their context as output by classification models. State the set of assumptions related to the problem under consideration

The emails content made available in Dataset is assumed to be written in English Language in the standard Greco-Roman script. This is so that the Stopword package and WordNetLemmatizer can be effectively used.

2.5 Hardware, Software and Tool Used

**Hardware Used:**

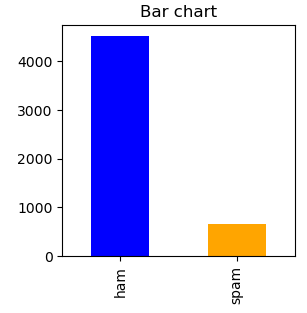
Processor – Intel core i3

Physical Memory – 8 GB

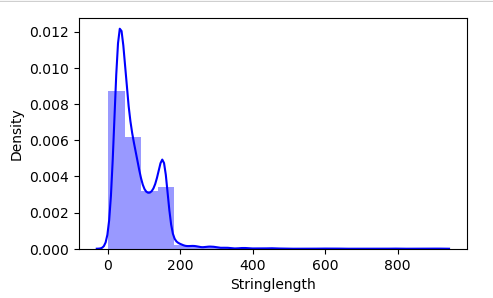
**Software Used:**

* Windows 10 Operating System
* Anaconda Package and Environment Manager
* Jupyter Notebook
* Python Libraries used: In Which Pandas, Seaborn, Matplotlib, Numpy and Scipy
* sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

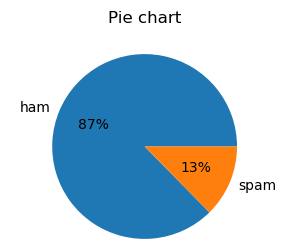
**3.Data Analysis and Visualization**



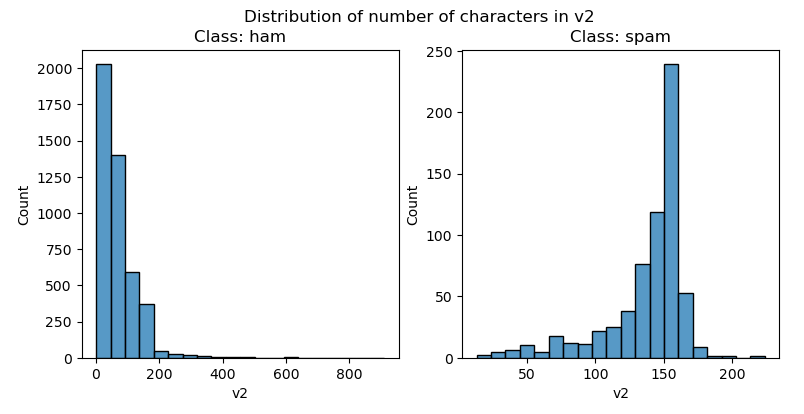
* We can see the label is not balanced wo we need balanced it.



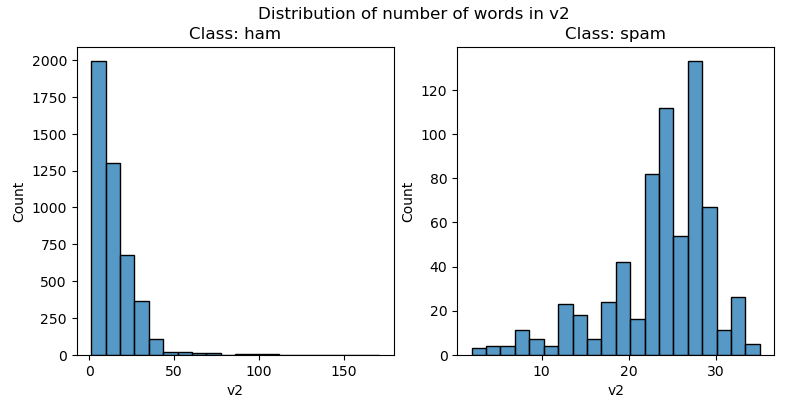
* we can see most of the emails are lies between 0 to 200 words.



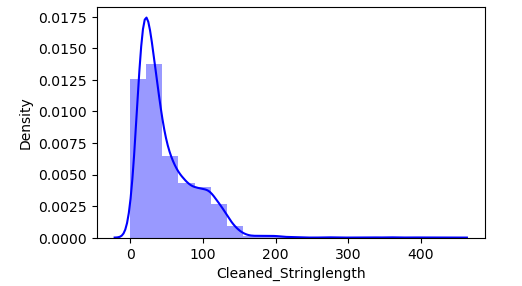
* We can see, the 13.4% of emails is spam but 86.6.
* 86.6% emails are non-spam(ham)



* We can see, that non-spam emails are lies between the 0 to 200 characters mostly.
* Spam emails characters mostly lies between the 110 to 160.

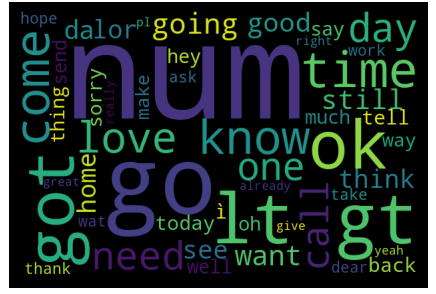


* The non-spam words in emails mostly lies between 0 to 30.
* But in spam mostly number of words lies in between 20 to 30.

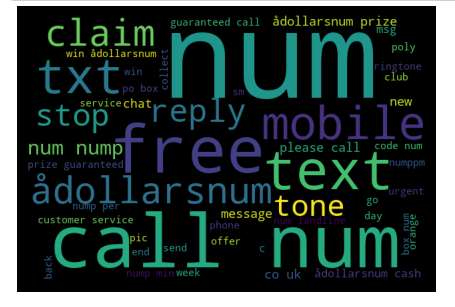


* We can see, word density has been reduced.

###### **World clouds of the most frequent words in emails.**

****

* We can see, so many emails are not spam.
* num, ok, love, need and come these are the word are mostly used.

****

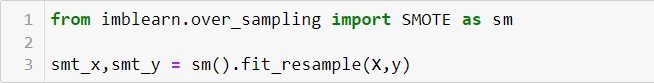
* We can see, spam emails words that are mostly used as show above.
* txt, call, claim, adollarsnum and mobile these are the words are mostly used.

4. Models Development and Evaluation

4.1 The model algorithms used

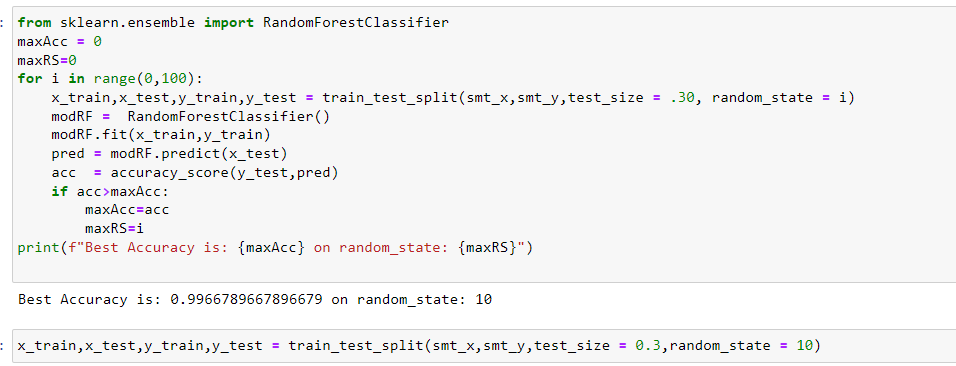
The model algorithms used were as follows:

* Logistic Regression: It is a classification algorithm used to find the probability of event success and event failure. It is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. It supports categorizing data into discrete classes by studying the relationship from a given set of labelled data. It learns a linear relationship from the given dataset and then introduces a nonlinearity in the form of the Sigmoid function. It not only provides a measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).
* XGBClassifier: XGBoost uses decision trees as base learners; combining many weak learners to make a strong learner. As a result, it is referred to as an ensemble learning method since it uses the output of many models in the final prediction. It uses the power of parallel processing and supports regularization.
* RandomForestClassifier: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A random forest produces good predictions that can be understood easily. It reduces overfitting and can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm
* Complement Naïve Bayes Classifier: Complement Naive Bayes is somewhat an adaptation of the standard Multinomial Naive Bayes algorithm. Complement Naive Bayes is particularly suited to work with imbalanced datasets. In complement Naive Bayes, instead of calculating the probability of an item belonging to a certain class, we calculate the probability of the item belonging to all the classes.

**Balancing out the classes by using SMOTE technique**

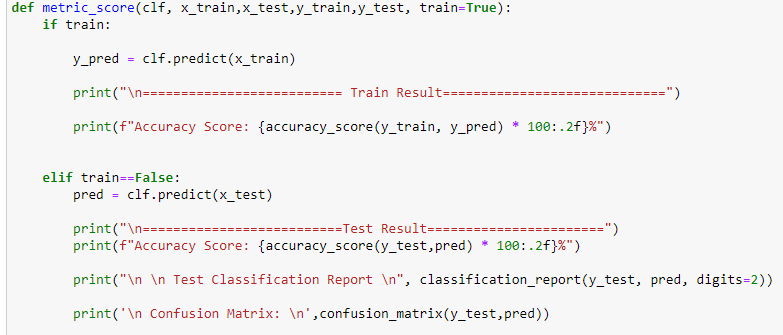
**Train Test Split**

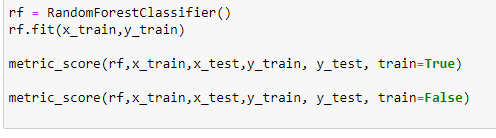
Best random state was found to be 10.

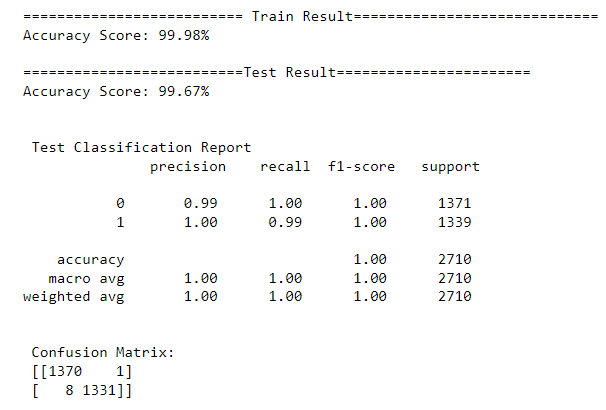


**Model Building**

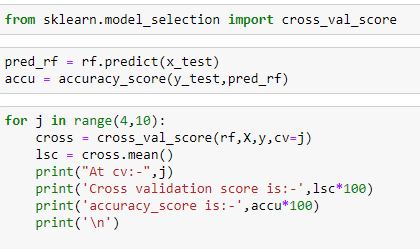
**Random Forest Classifier**

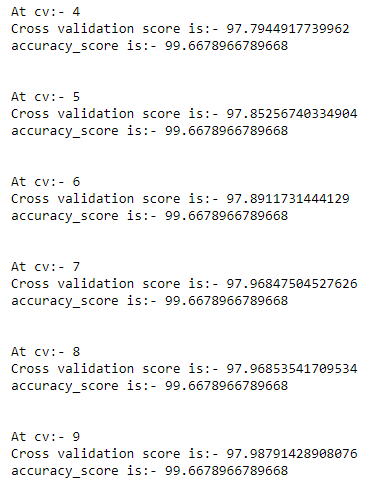
****

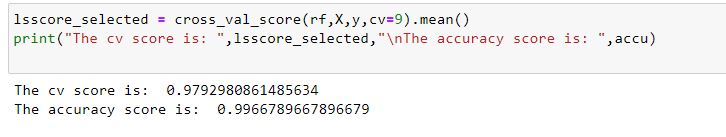
****

****

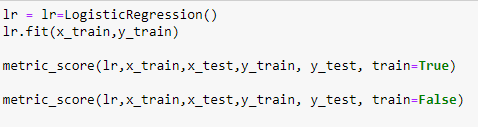
**Cross Validation for random forest classifier**

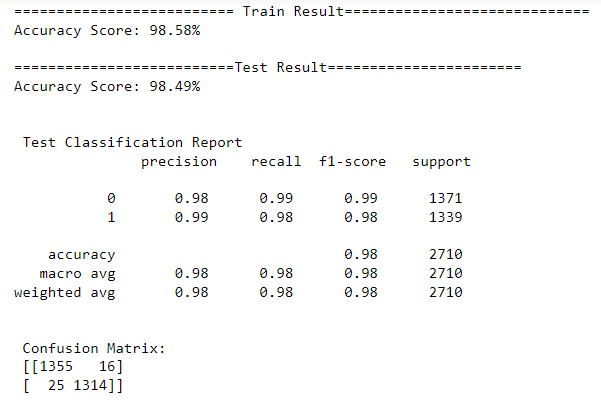
****

****

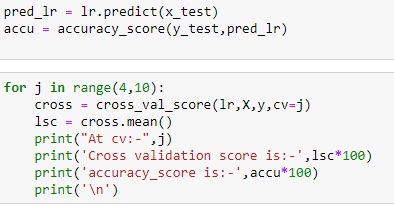
****

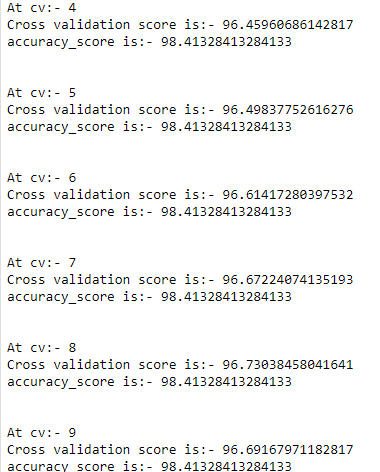
**Logistic Regression**

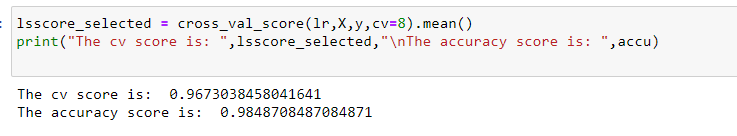
****

****

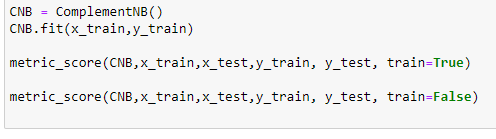
**Cross Validation for Logistic Regression**

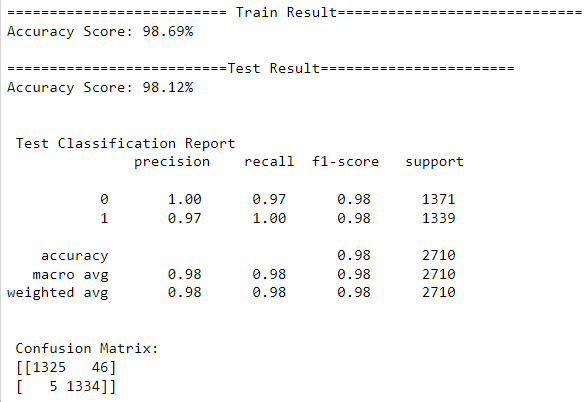
****

****

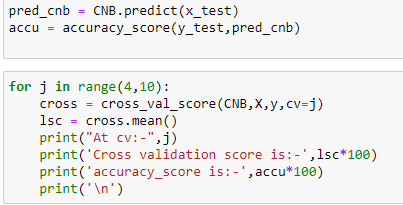
****

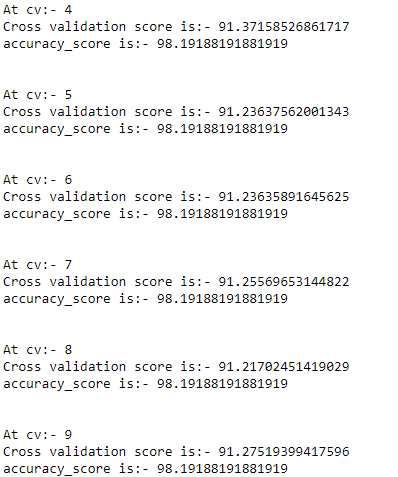
**Complement Naive Bayes**

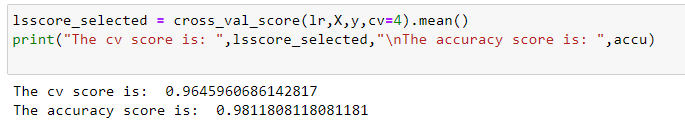
****

****

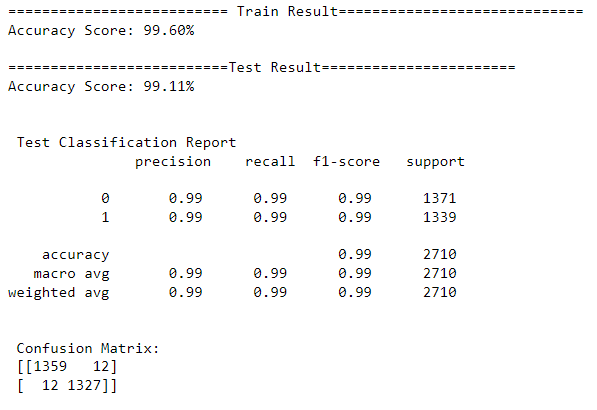
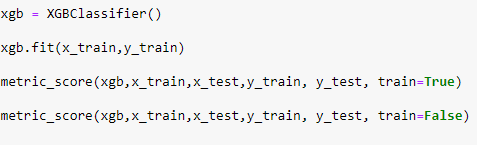
**Cross validation for Complement Naïve Bayes**

****

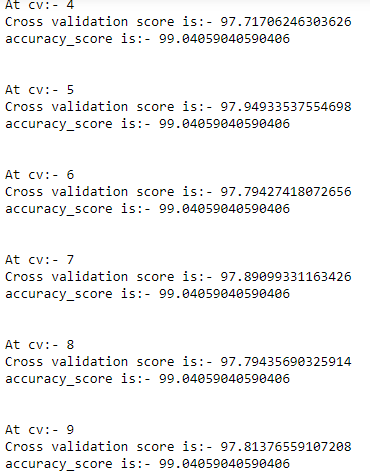
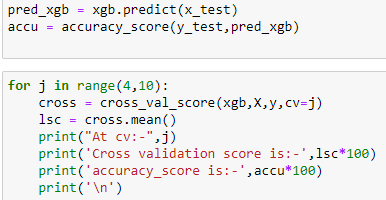
****

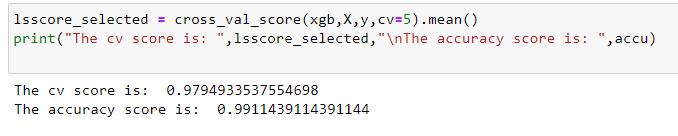
****

**XGBClassifier**

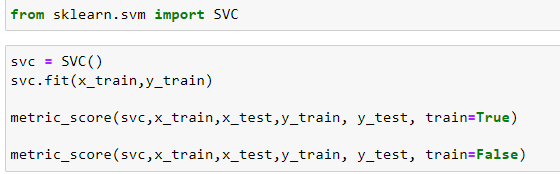
****

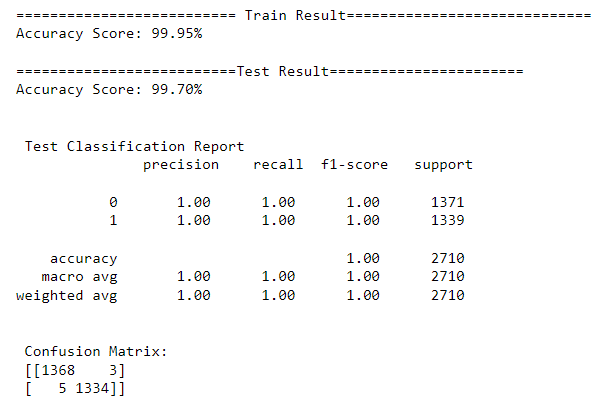
**Cross Validation for XGBClassifier**

****

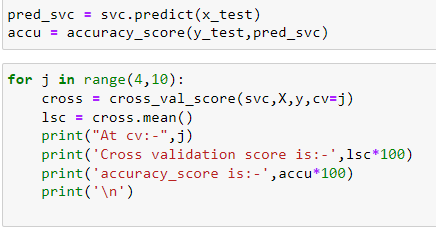
****

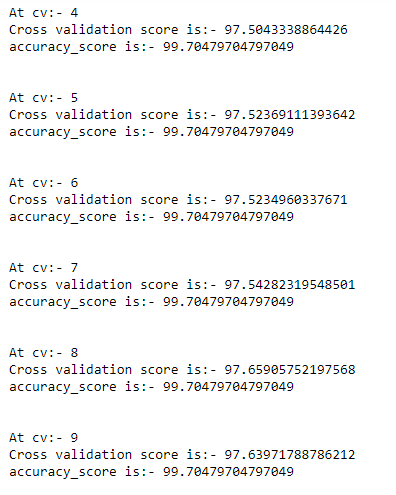
**SVC**

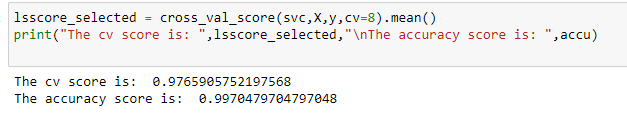
****

****

**Cross Validation for SVC**

****

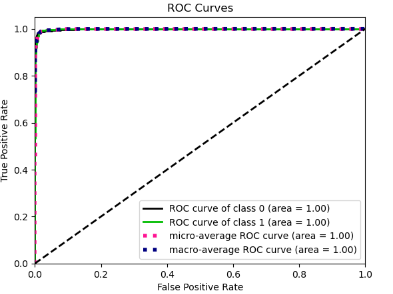
****

****

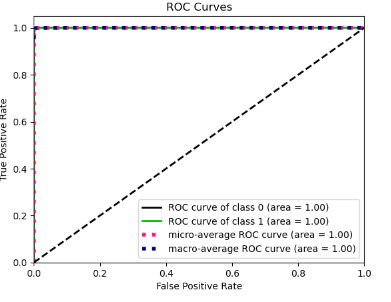
4.2 ROC AUC Curve

The AUC-ROC curve helps us visualize how well our machine learning classifier is performing. ROC curves are appropriate when the observations are balanced between each class.

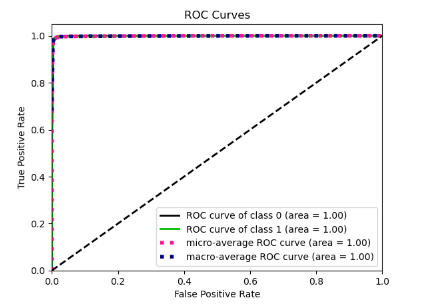
**Random Forest Classifier**

****

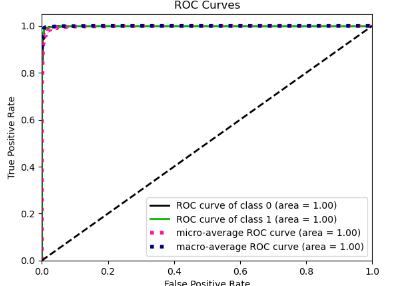
**Logistic Regression**

****

**XGBClassifier**

****

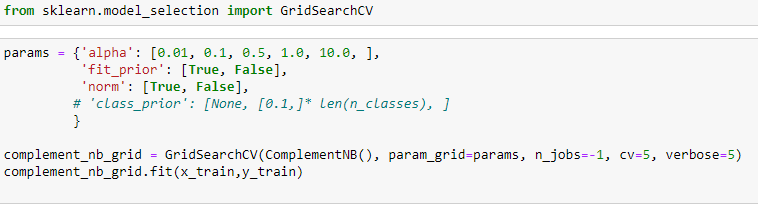
**Complement Naïve Bayes**

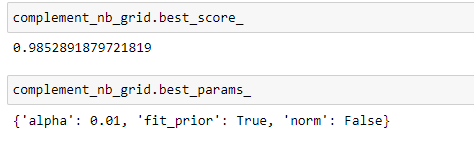
****

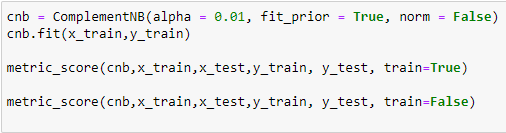
4.3 Interpretation of the results

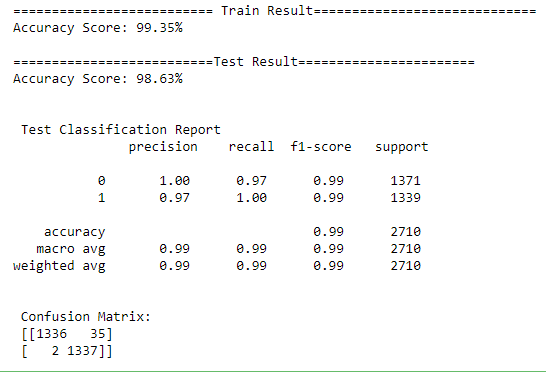
Based on comparing the above graphs, Precision, Recall, Accuracy Scores with Cross validation scores, it is determined that Complement Naive Bayes Classifier is the best model for the dataset.

4.4 Hyperparameter Tuning

****

****

****

****

**5. Conclusions**

5.1 Key Finding and Conclusions

* The final model performed with 99% accuracy, Recall score of 0.99. It means that the model is optimized better to predict label whether it is spam email or not.
* Spam emails have become a major concern for the internet community as it poses a threat to integrity and productivity of the users. Filtering of email is very much necessary for email communication. The accurate detection of spam emails is a big issue and many filtering methods have been proposed by various research
* Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose.
* So that we need to do spam filtering so user more user friendly. From above model building we got the Complement Naive Bayes Classifier is a best model deciding whether the emails have spam or not.

5.2 Limitation of this works and scope for future works

A small dataset to work with posed a challenge in building highly accurate models. By training the models on more diverse data sets, longer comments, and a more balanced dataset, more accurate and efficient classification models can be built.