EMAIL SPAM DETECTION

A Project Phase - II Report
Submitted in
partial fulfilment
for the award of the degree
of

Bachelor of Technology
in
Information Technology

By

SUNIDHI PANDEY
SHANTANU SINGH CHANDEL

Under the Guidance

Prof. Kunal Kumar

to the

GOVERNMENT ENGINEERING COLLEGE, BILASPUR

CHHATTISGARH SWAMI VIVEKANAND TECHNICAL UNIVERSITY, BHILAI

Session 2022-23

DECLARATION

We the undersigned solemnly declare that the report of the project work entitled "**Email Spam Detection**" is based on our own work carried out during the course of our study under the supervision of **Kunal Kumar**. We assert that the statements made and conclusions drawn are an outcome of the project work.

Signature of the Student
Sunidhi Pandey
300703319036

Signature of the Student
Shantanu Singh Chandel
300703319034

CERTIFICATE

It is certified that the work contained in the report entitled "**Email Spam Detection**" by **Sunidhi Pandey** (300703310936) and **Shantanu Singh Chandel** (300703319034) has been carried out under the supervision of **Kunal Kumar** and this work has been submitted for award of the degree of Bachelor of Technology in Information Technology.

Signature of Project Incharge
Dr. Awanish Kumar Upadhyay
Professor
Information Technology
Government Engineering College, Bilaspur

Signature of the Supervisor
Kunal Kumar
Assistant Professor
Information Technology
Government Engineering College, Bilaspur

Signature of the Head of Department
Dr. Awanish Kumar Upadhyay
Professor
Information Technology
Government Engineering College, Bilaspur

CERTIFICATE BY THE EXAMINERS

The report entitled **"Email Spam Detection"** submitted by **Sunidhi Pandey** (300703319036) and **Shantanu Singh Chandel** (300703319034) has been examined by the undersigned as a part of the examination and is hereby recommended for the award of the degree of Bachelor of Technology in Information Technology to Chhattisgarh Swami Vivekanand Technical University.

Internal Examiner

External Examiner

ACKNOWLEDGEMENT

I would like to express my deep gratitude to my project guide **Kunal Kumar**, Assistant Professor, Department of Information Technology, GECB, for his guidance with unsurpassed knowledge and immense encouragement. We are grateful to **Dr. Awanish Kumar Upadhyay**, Head of Department, Information Technology, as well as our project in charge for providing us with the required facilities for the completion of the project work.

I am very much thankful to the **Dr. B. S. Chawla, Principal and Management, GEC Bilaspur**, for their encouragement and cooperation to carry out this work.

I express my thanks to Project Coordinator, for his continuous support and encouragement. I thank **Samiksha Shukla, Priyanka Sahu, Himanshu Mokashe**, all teaching faculty of Department of Information Technology whose suggestions during reviews helped us in accomplishment of my project.

I would like to thank our parents, friends, and classmates for their encouragement throughout my project period. At last, but not least, I thank everyone for supporting us directly or indirectly in completing this project successfully.

Sunidhi Pandey

Shantanu Singh Chandel

ABSTRACT

Nowadays, Email spam is very common and easy illegal phishing technique, which is used in various ways. It is one of the biggest threats to the Internet. There are so many anti-spam tools already out in the market but still its lacking due to the personalized spams. This paper help to identify the mail as ham or spam as it is a complete spam EDA with Supervised algorithms.

This paper aims to compare different Supervised algorithms on the Spam email dataset to classify different Machine Learning techniques. We can evaluate them on the basis of accuracy, precision, recall, F1-score, as well as with AUC-ROC score.

Keywords: Machine Learning, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, K Nearest Neighbour, Gradient Boosting, Extreme Gradient Boosting.

Table of Contents

List of Figures

List of Abbreviations

ML – Machine Learning

AI – Artificial Intelligence

HTML – Hypertext Markup Language

ASCII – American Standard Code for Information Interchange

NLTK – Natural Language Toolkit

MNIST – Modified National Institute of Standards and Technology Database

EDA – Exploratory Data Analysis

UCI – University of California Irvine machine learning repository

SMOTE – Synthetic Minority Over-sampling Technique

VIF – Variance Inflation Factor

RFE – Recursive Feature Elimination

PCA – Principle Component Analysis

LR – Logistic Regression

DT – Decision Tree

RF – Random Forest

NB – Naïve Bayes

SVM – Support Vector Machine

KNN – K-Nearest Neighbours

GB – Gradient Boosting

XGB – Extreme Gradient Boosting

AUC – Area Under the (ROC) Curve

ROC – Receiver Operating Characteristic

1.1 EMAIL

Email (electronic mail) is the exchange of Computer stored messages by telecommunication. E-Mail messages are usually encoded in ASCII text. However, you can also send non-text files, such as graphic images and sound files, as attachments Sent in binary streams. E-mail was one of the first uses of the Internet and is still the most popular use. E-mail is a message that may contain text, files, images, or other attachments sent through a network to a Specified individual or group of individuals.

An email address is required to receive email, and that address is unique to the user. Some people use Internet-based applications and some use programs on their computer to access and store emails. Companies that are fully computerized make extensive use of e-mail because it is fast, flexible, and reliable. Email communication is not only used in lieu of letter writing, it has also replaced telephone calls in many social situations and in professional environments.

Some electronic mail systems are confined to a single computer system or network, but others have gateways to other computer systems, enabling users to send electronic mail anywhere in the world. It was one of the first methods of person-to-person communication made available through the Information superhighway.

In the early days of e-mail, simple text messages were sometimes difficult to manage, and adding pictures or documents was possible only if other software was available to make transmission from e-mail to computer possible. Current e-mail software generally provides easy-to-use options for attaching photos, sounds, video clips, complete documents, and Hypertext Markup languages (HTML) code. Even with attachments, however, e-mail messages continue to be text messages -- we'll see why when we get to the section on attachments.

1.2 EMAIL SPAM

Junk email or unsolicited bulk emails sent to a large list of email users through the email system are referred to as email spam. Typically, they are misleading ads that promote low-quality services and, in some instances, include images with content that is inappropriate for children.

Whether commercial or not, many of them are really dangerous since they may contain links that appear to be legitimate and recognizable, but they lead to phishing websites that host malware or include malware in the form of file attachments.

Spam emails that advertise products, such as miraculous weight loss pills or sexual enhancers; scams such as advance fees, current events, or tech support scams that try to trick you into paying money or giving away personal information; phishing emails that attempt to harvest sensitive information from unsuspecting victims, such as usernames, passwords, and credit card details; blank spam– this is an empty email, sometimes without a subject line, used by cybercriminals to test the validity of the email address so they can then target that address with malware-laden spam.

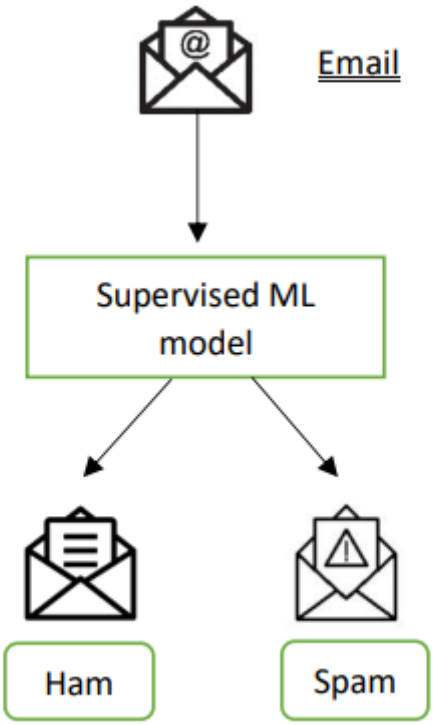

Fig 1.1 Classification into Spam and Ham mails

Malware messages that can deceive users into sharing private information, paying money, or doing things they would not normally do antivirus alerts. These notifications "notify" the user about virus infection and provide a "fix" for it. The threat actor will be able to obtain access to the victim's system if they fall for the lure and click on a link included in the email. The email may also contain a malicious file that will be downloaded to the device,

“you won” email messages that spammers send out claiming that the target has won something like a prize. The recipient has to click on a link in the email to get the prize promised in the message. The link is malicious and is frequently used to steal sensitive data from users.

1.2.1 SPAM DETECTION

The "spam" concept is diverse: advertisements for products/websites, make money fast schemes, chain letters, pornography, etc. Our collection of spam e-mails came from our postmaster and individuals who had filed spam.

Our collection of non-spam e-mails came from field work and personal e-mails, and hence the word 'George' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general-purpose spam filter.

Literature Review:

With growing development in the field of Data Analysis with big data, Data Science and Machine Learning, data mining, analytics, decision making, etc., various researches and experiments have been carried out in recent years releasing the relevant significant papers.

- Rajesh Kumar J, Sudarshan P and Mahalakshmi G proposed a paper "Email Spam Detection using Machine Learning Techniques". Author have mentioned two different machine learning algorithms ie. Naïve bayes classification and SVM and compare their results to prove that Naïve bayes are much better than SVM. Algorithms are compared using different parameters like accuracy, precision, recall and F-measures. Also work to help machines to understand human phrases and conversation.

- Isra'a AbdulNabi and Qussai Yaseen proposed a paper "Spam Email Detection Using Deep Learning Techniques". Author used the deep learning technique approach to create models. Experimental results of the author shows that Bert-base-cased transformer model is the best fit model with high accuracy and F1 score. The author uses NLP task methodology with its five main phases: data collection, data pre-processing, feature extraction, model training, and model evaluation.

- Thashina Sultana, K A Sapnaz, Fathima Sana and Mrs. Jamedar Najath proposed a paper "Email based Spam Detection". The author works on the strategy of frequently repeated words in spam word cloud as well as in ham word cloud. Using NLTK and the proposed system of Machine learning as Naïve Bayes algorithms, the author creates a model to detect ham or spam email. For it, the author used to calculate Term frequency and Inverse Document frequency which help the model to build in a much better way.

- Nikhil Kumar, Sanket Sonowal and Nishant proposed a paper "Email Spam Detection Using Machine Learning Algorithms". The author uses a broad way to classify. Author used Stop word, Tokenization, and Bag of words for Data pre-processing. Also work with many different ML algorithms such as SVM, KNN, Naïve Bayes, Decision Tree, Random Forest, AdaBoost and Bagging Classifier. Compare all the models and conclude that the Multinomial Naïve Bayes gives the best outcome out of all.

- V.Christina, S.Karpagavalli and G.Suganya proposed a paper “Email Spam Filtering using Supervised Machine Learning Techniques”. Author generated spam and legitimate corpus from the latest mails and employed ML techniques to build the model. The performance of the model is evaluated using 10-fold cross validation and observed that Multilayer Perceptron classifier out performs other classifiers and the false positive rate also very low compared to other algorithms. Email spam filters using this approach can be adopted either at mail server or at mail client side to reduce the amount of spam messages and to reduce the risk of productivity loss, bandwidth and storage usage.

- Mangena Venu Madhavan, Sagar Pande, Pooja Umekar, Tushar Mahore and Dhiraj Kalyankar proposed a paper “Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches”. Author tries to justifies the working and functionality of the algorithms along with their advantages and disadvantages based on numerous considered parameters.

This normally involves various steps, like choosing a sample, collecting data from this sample, and interpreting this data. The study of methods involves a detailed description and analysis of these processes. It includes evaluative aspects by comparing different methods to assess their advantages and disadvantages relative to different research goals and situations.

3.1 Introduction to Machine Learning

Machine learning (ML) is a field devoted to understanding and building methods that let machines "learn" – that is, methods that leverage data to improve computer performance on some set of tasks.

Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

In this project it is used for spam filtering in emails to save time from seeing the unwanted spam emails.

Machine learning programs can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand; on the computer's part, no learning is needed. For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.

The discipline of machine learning employs various approaches to teach computers to accomplish tasks where no fully satisfactory algorithm is available. In cases where vast

numbers of potential answers exist, one approach is to label some of the correct answers as valid. This can then be used as training data for the computer to improve the algorithm(s) it uses to determine correct answers. For example, to train a system for the task of digital character recognition, the MNIST dataset of handwritten digits has often been used.

Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

3.1.1 Supervised Learning

A support-vector machine is a supervised learning model that divides the data into regions separated by a linear boundary. Here, the linear boundary divides the black circles from the white.

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

Types of supervised-learning algorithms include active learning, classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. As an example, for a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email.

Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

3.1.2 Unsupervised Learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labelled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function. Though unsupervised learning encompasses other domains involving summarizing and explaining data features.

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters. Other methods are based on estimated density and graph connectivity.

3.1.3 Semi-supervised Learning

Semi-supervised learning falls between unsupervised learning (without any labelled training data) and supervised learning (with completely labelled training data). Some of the training examples are missing training labels, yet many machine-learning researchers have found that unlabelled data, when used in conjunction with a small amount of labelled data, can produce a considerable improvement in learning accuracy.

In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

3.1.4 Reinforcement Learning

Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. Due to its generality, the field is studied in many other disciplines, such as game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms. In machine learning, the environment is typically represented as a Markov decision process (MDP). Many reinforcement learning algorithms use dynamic programming techniques. Reinforcement learning algorithms do not assume knowledge of an exact mathematical model of the MDP and are used when exact models are infeasible. Reinforcement learning algorithms are used in autonomous vehicles or in learning to play a game against a human opponent.

3.1.5 Dimensionality Reduction

Dimensionality reduction is a process of reducing the number of random variables under consideration by obtaining a set of principal variables. In other words, it is a process of reducing the dimension of the feature set, also called the "number of features". Most of the dimensionality reduction techniques can be considered as either feature elimination or extraction. One of the popular methods of dimensionality reduction is principal component analysis (PCA). PCA involves changing higher-dimensional data (e.g., 3D) to a smaller space (e.g., 2D). This results in a smaller dimension of data (2D instead of 3D), while keeping all original variables in the model without changing the data. The manifold hypothesis proposes that high-dimensional data sets lie along low-dimensional manifolds, and many dimensionality reduction techniques make this assumption, leading to the area of manifold learning and manifold regularization.

3.2 Strategic Plan of Action

An action plan provides responsibilities, tasks, and the necessary resources to align the efforts with strategy and make them feel relevant, impactful, and engaging. Large companies like Amazon and Walmart use data science and ML in various fields: from marketing to supply chain management, from forecasting demands to Human Resources. Many companies have become dedicated data mining companies dealing with only data science and ML. To create a model we need to work with different Supervised machine learning algorithms. To identify the spam or ham email, we used a strategic plan of action which consist of 6 steps as:

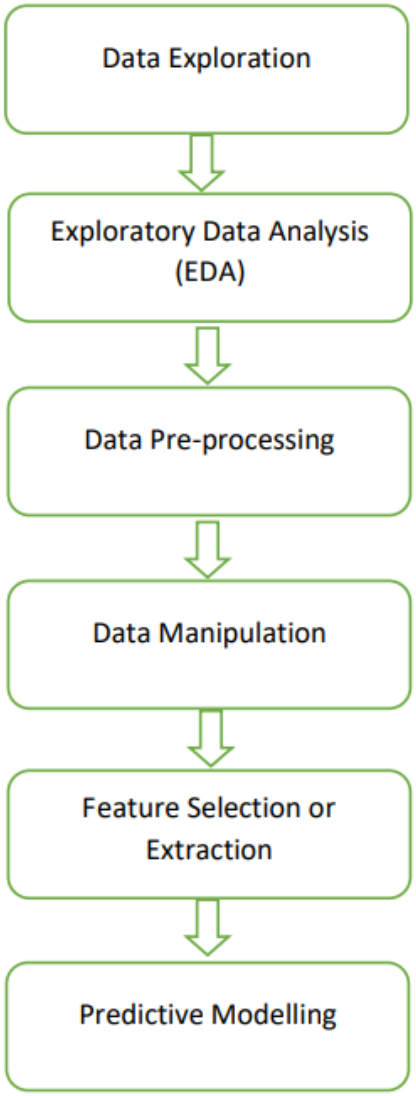

Fig 3.1 Strategic Plan of Action Steps

3.2.1 Data Exploration

The Data collection of our spam email came from postmasters and individuals who had filled spam. And collection of ham emails came from filed work and personal e-mails. The dataset taken from the UCI ML repository, contains about 4600 emails labelled as spam or ham.

It is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more. Using interactive dashboards and point-and-click data exploration, users can better understand the bigger picture and get to insights faster.

During exploration, raw data is typically reviewed with a combination of manual workflows and automated data-exploration techniques to visually explore data sets, look for similarities, patterns and outliers and to identify the relationships between different variables.

It is important because Humans process visual data better than numerical data, therefore it is extremely challenging for data scientists and data analysts to assign meaning to thousands of rows and columns of data points and communicate that meaning without any visual components.

Data visualization in data exploration leverages familiar visual cues such as shapes, dimensions, colors, lines, points, and angles so that data analysts can effectively visualize and define the metadata, and then perform data cleansing. Performing the initial step of data exploration enables data analysts to better understand and visually identify anomalies and relationships that might otherwise go undetected.

Few steps which our project includes while exploring the data are:

- Importing the dataset
- Checking the datatypes of all the columns
- Checking number of unique rows in each feature
- Checking the stats of all the columns, etc.

3.2.2 Exploratory Data Analysis (EDA)

It is an approach to analyse the dataset using visualize techniques. It helps to discover patterns, trends, characteristics, assumptions, etc. The overall objective of exploratory data analysis is to obtain vital insights and hence usually includes the following sub-objectives:

- Identifying and removing data outliers
- Identifying trends in time and space
- Uncover patterns related to the target
- Creating hypotheses and testing them through experiments
- Identifying new sources of data

A. Heatmap for Null values in Dataset –

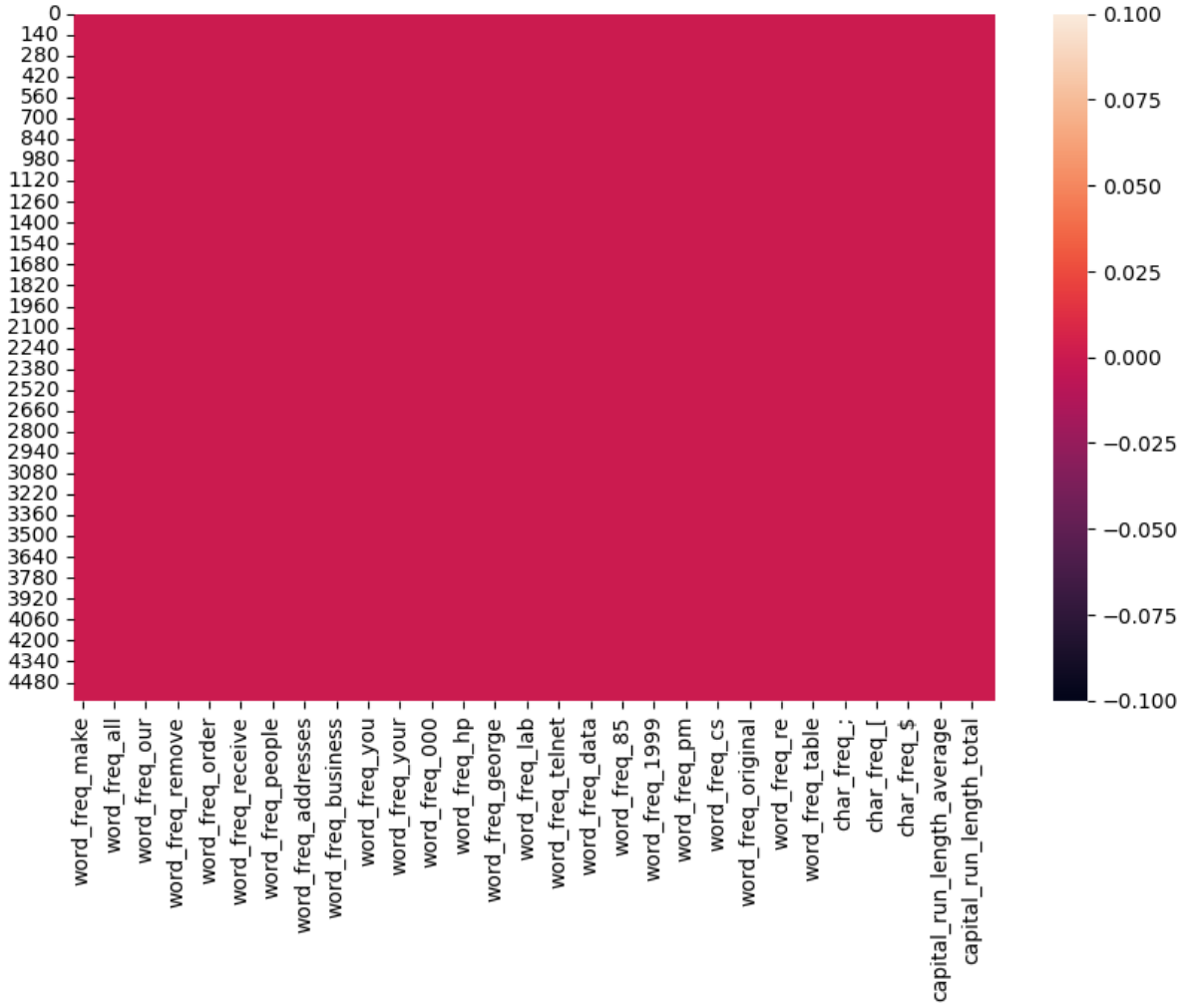

Fig 3.2 Heatmap of null values

A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

Here, we are using the seaborn library to use the heat map function within it. Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

Seaborn aims to make visualization the central part of exploring and understanding data. One of the ways to visualize the missing data is make a heatmap of the data coded as Boolean for missing-ness.

It shows that there are no null values or missing data in the used dataset.

B. Heatmap for Correlation of Dataframe –

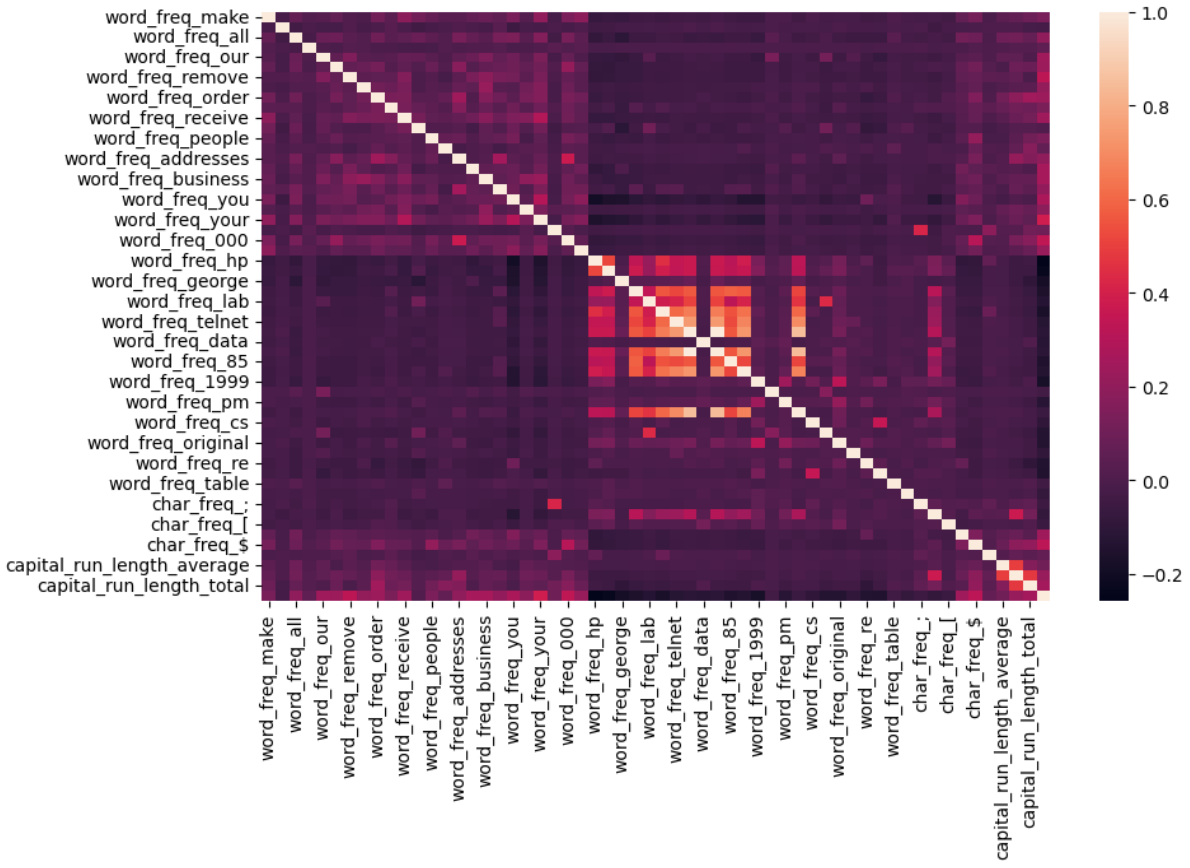

Fig 3.3 Heatmap of correlation

A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1. Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

C. Analyse the target variable distribution –

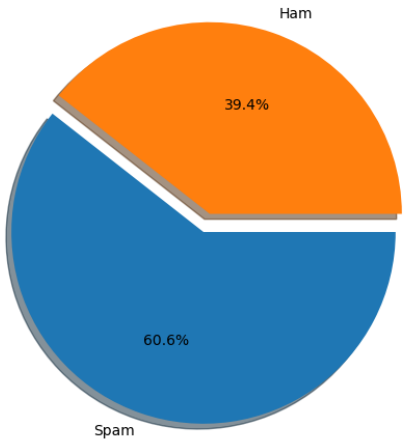

Fig 3.4 Distribution of the target variable

Here,

Spam = 60.6%

Ham = 39.4%

The target variable seems to be slightly imbalanced.

Hence, we try to perform data argumentation.

D. Visualizing the Sparse Matrix –

A sparse matrix is a special case of a matrix in which the number of zero elements is much higher than the number of non-zero elements.

A sparse matrix is a matrix that is comprised of mostly zero values. Sparse matrices are distinct from matrices with mostly non-zero values, which are referred to as dense matrices. A matrix is sparse if many of its coefficients are zero.

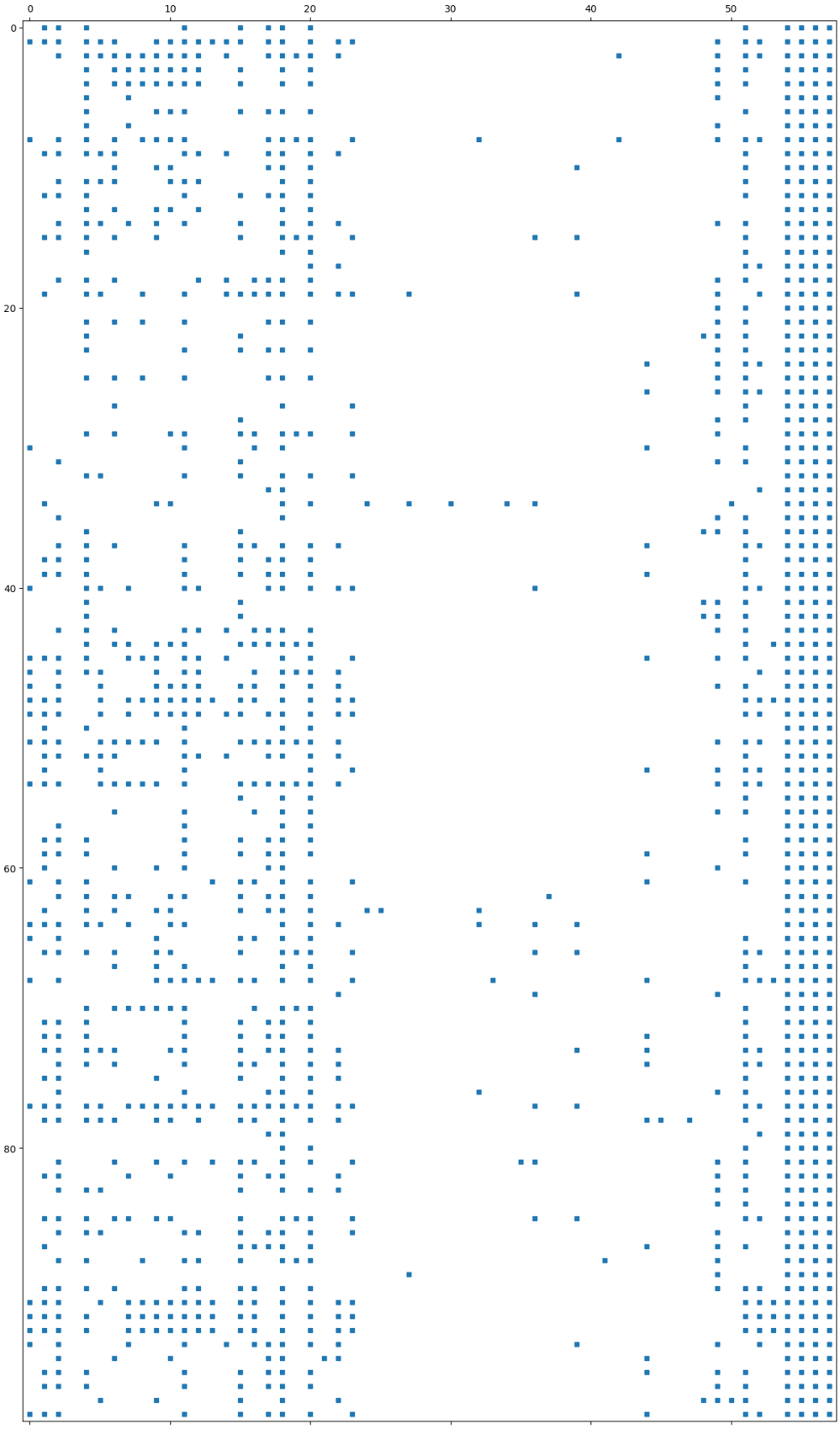

Fig 3.5 Sparse Matrix Visualization

3.2.3 Data Pre-processing

Data pre-processing, a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process.

Data pre-processing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance.

In this step, we work for the removal of duplicate rows, check for the empty elements and fixed the imbalance using SMOTE technique. The Final Dataset size after performing pre-processing is:

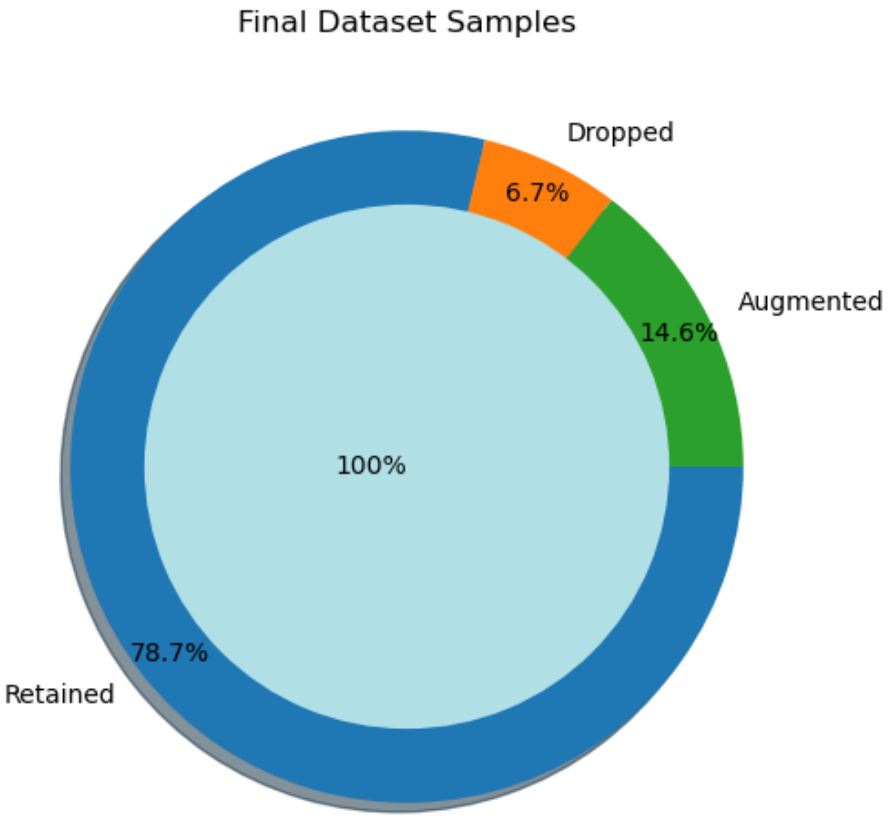

Fig 3.6 Final Dataset samples distribution

In final dataset sample:

6.7% - Dropped

14.6% - Augmented

And 78.7% - Retained.

So, the final dataset after clean-up has 58 samples & 4601 rows.

3.2.4 Data Manipulation

For data manipulation, we split the data into training and testing sets and perform feature scaling for standardization.
The dataset gets split as:
Original set ---> (5062, 57) (5062,)
Training set ---> (4049, 57) (4049,)
Testing set ---> (1013, 57) (1013,)

After that we perform feature scaling (standardization).

Feature scaling is a data pre-processing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model. Feature scaling is essential when working with datasets where the features have different ranges, units of measurement, or orders of magnitude.

Standardization is another scaling method where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

3.2.5 Feature Selection/Extraction

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features. Feature selection can significantly increase the performance of a learning algorithm (both accuracy and computation time).

Correlation plot between the variables convey lot of information about the relationship between them. There seems to be strong multicollinearity in the dataset.

With different techniques, we can improve the model's performance by performing Feature Selection/Extraction to take care of these multi-collinearity.

We can fix these multicollinearities with three techniques:

1. Manual Method - Variance Inflation Factor (VIF)
2. Automatic Method - Recursive Feature Elimination (RFE)
3. Decomposition Method - Principle Component Analysis (PCA)

Manual Method - Variance Inflation Factor (VIF)

Variance inflation factor measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression.

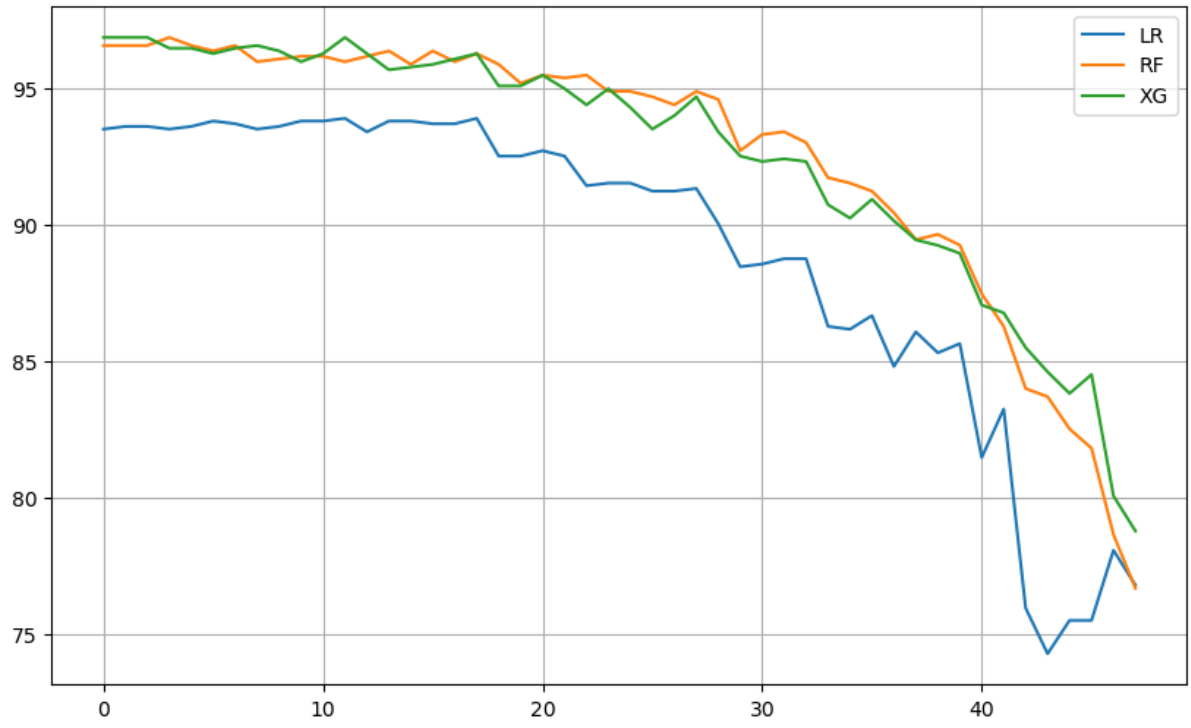

Fig 3.7 VIF plot

When significant multicollinearity issues exist, the variance inflation factor will be very large for the variables involved. After these variables are identified, several approaches can be used to eliminate or combine collinear variables, resolving the multicollinearity issue.

The formula for VIF is:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where,

R_i^2=Unadjusted coefficient of determination forregressing the ith independent variable on the remaining ones.

When R_i^2 is equal to 0, and therefore, when VIF or tolerance is equal to 1, the i^{th} independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.
In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

Automatic Method - Recursive Feature Elimination (RFE)

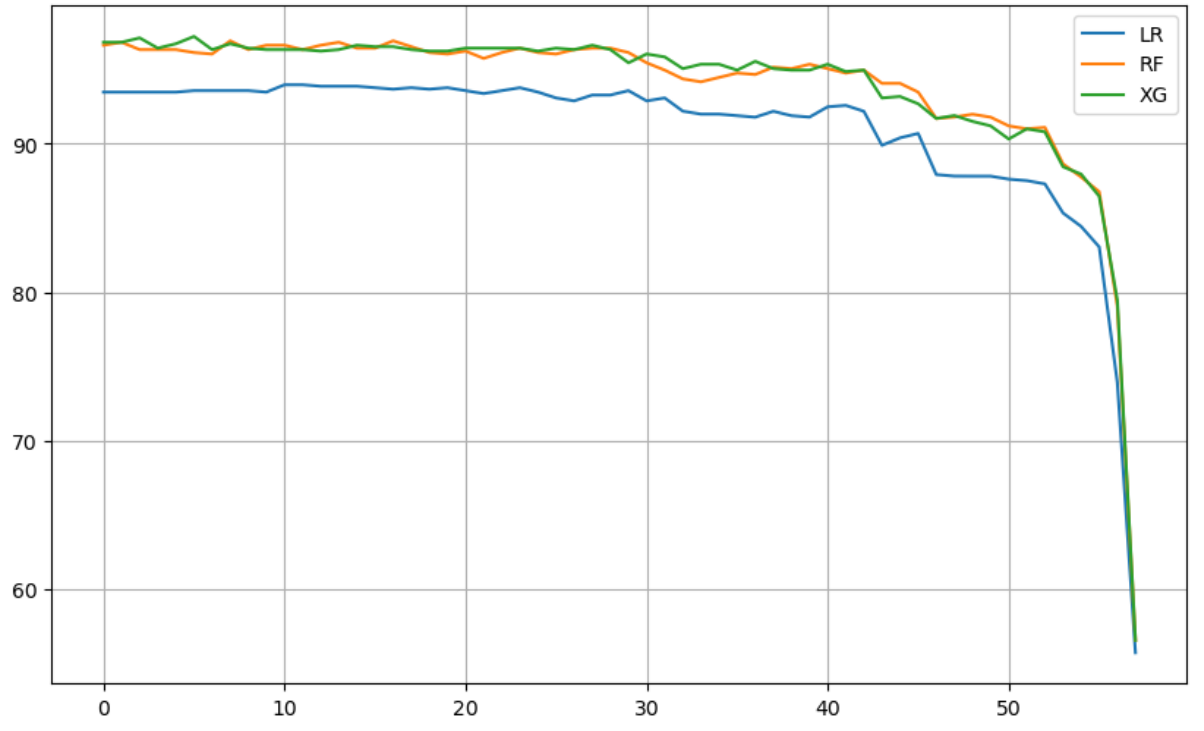

Fig 3.8 RFE plot

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

RFE requires a specified number of features to keep, however it is often not known in advance how many features are valid. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features.

It is a Wrapper method. It is a transformer estimator, which means it follows the familiar fit or transform pattern.

Decomposition Method - Principle Component Analysis (PCA)

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.

For concluding PCA, the Explained Variance of Components are:

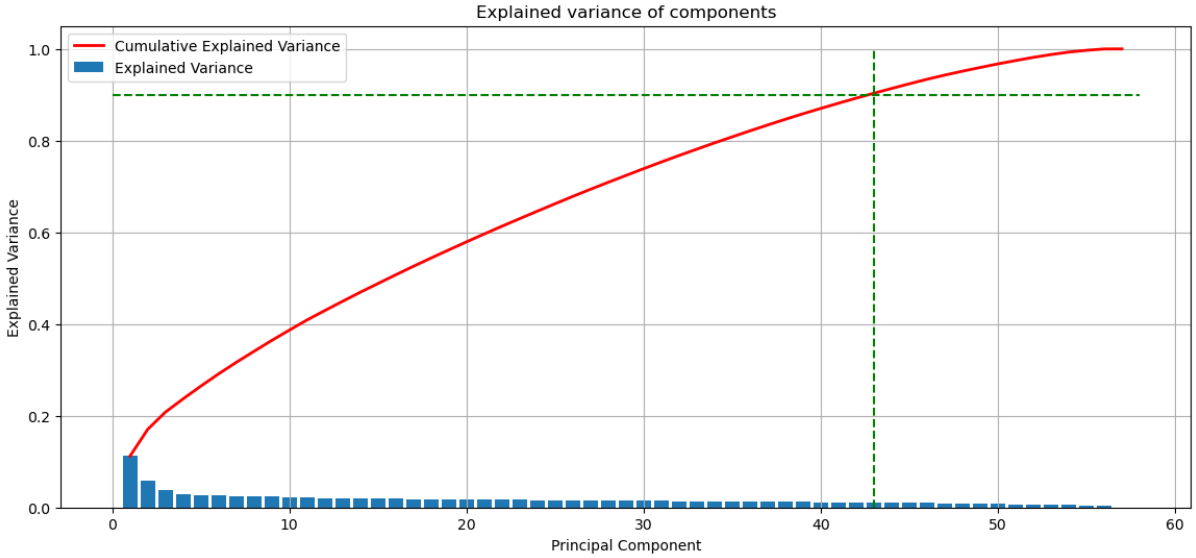

Fig 3.9 Explained variance of components

We shall avoid performing dimensionality reduction for the current problem.

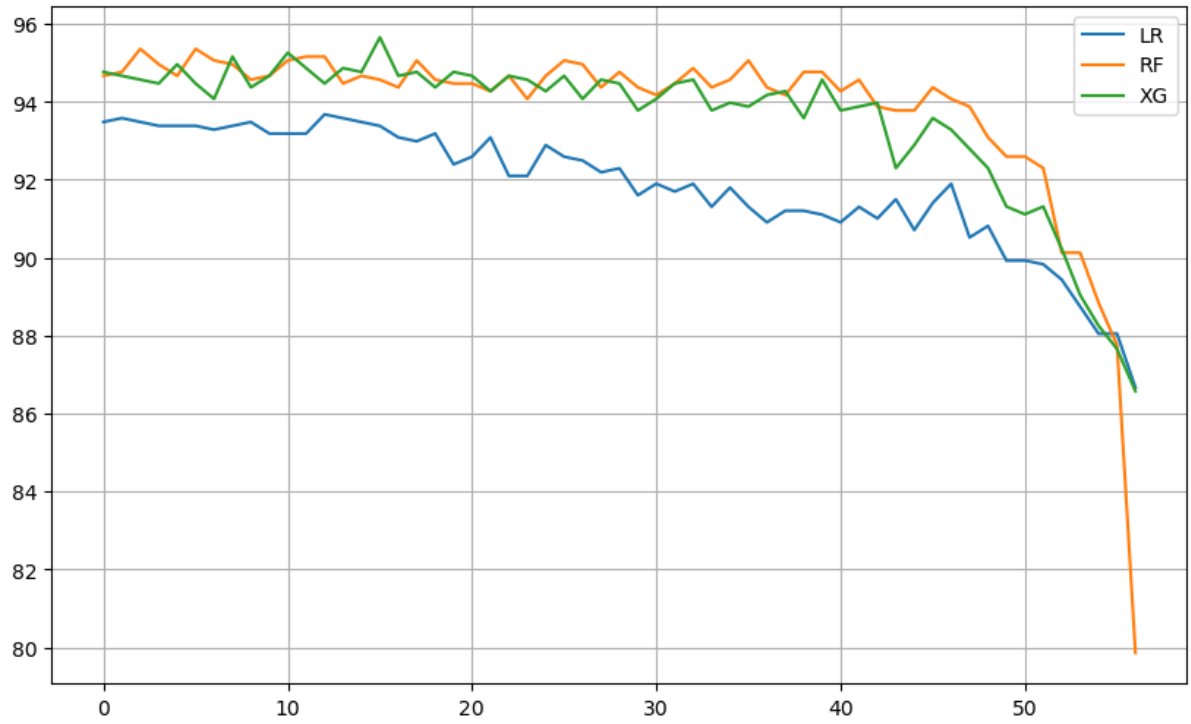

Fig 3.10 PCA plot

And, the shape of final transformed training feature set is (4049, 32).
Whereas, the shape of final transformed testing feature set is (1013, 32).

3.2.6 Predictive Modelling

It is a mathematical process which is used to predict future behaviour by using statistical techniques. It gives the solution for data-mining technologies by analysing the past and current data and create a model which helps to predict the future outcomes. In predictive modelling, data is collected, a statistical model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available.

Banking institutions, for example, may leverage predictive modelling to collect a customer's credit record and other historical data. They might then use this information to calculate a person's credit score and the odds of them making timely credit payments.

Organizations implement predictive analytics using predictive models, which assists them in making better business decisions. Predictive models let companies understand their customer

base better, predict future sales prospects, etc. Following are some of the ways in which predictive models benefit various businesses-

- Implement techniques to acquire a competitive advantage,
- Gain a better understanding of the consumer base and their demands,
- Assess and mitigate financial risks,
- Enhance existing products to boost revenue,
- Minimize time and expenses in predicting outcomes,
- Predict external elements that may have an impact on productivity, etc.

Predictive models are used for forecasting inventory, managing resources, setting ticket prices, managing equipment maintenance, developing credit risk models, and much more. They help companies reduce risks, optimize operations, and increase revenue.

Predictive modelling as a part of data analytics can take many different forms, depending on what type of data is available and what type of prediction is being made.

For this project, we use different predictive modelling techniques of supervised ML as:

LR	Logistic Regression	**SVM**	Support Vector Machine
DT	Decision Tree	**KNN**	K Nearest Neighbours
RF	Random Forest Classifier	**GB**	Gradient Boosting
NB	Naive Bayes Classifier	**XGB**	Extreme Gradient Boosting

Fig 3.11 Different Predictive Modelling Techniques

A. Logistic Regression (LR)

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.

It is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

LR predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1**.**

It is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems**.**

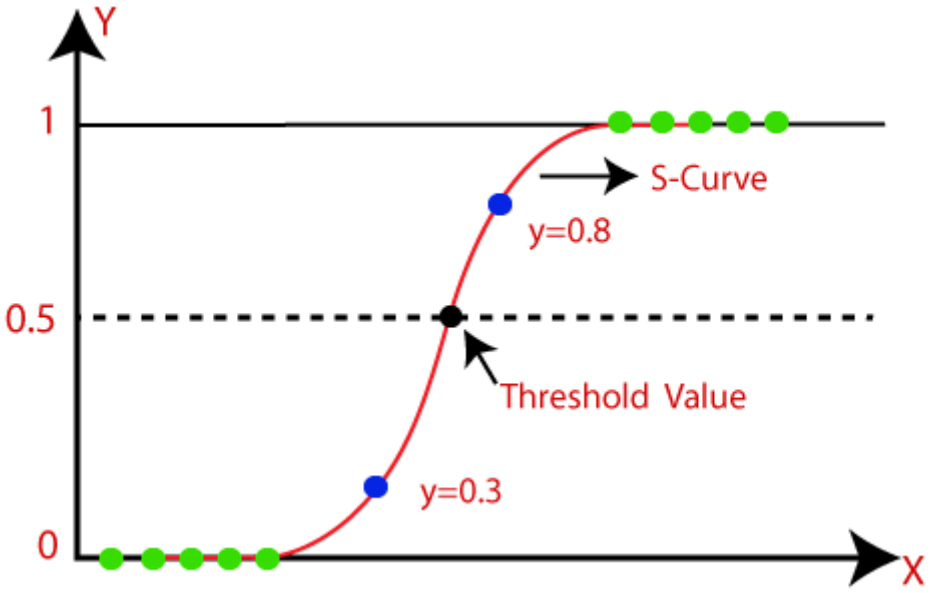

Fig 3.12 Sigmoid Function

Instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

LR is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Function:

It is a mathematical function used to map the predicted values to probabilities. It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.

In LR, the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions of LR includes the dependent variable must be categorical in nature; and the independent variable should not have multicollinearity.

Equation of LR:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

As LR,

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

So, the final equation of LR becomes

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In our project, the LR ROC curves looks like:

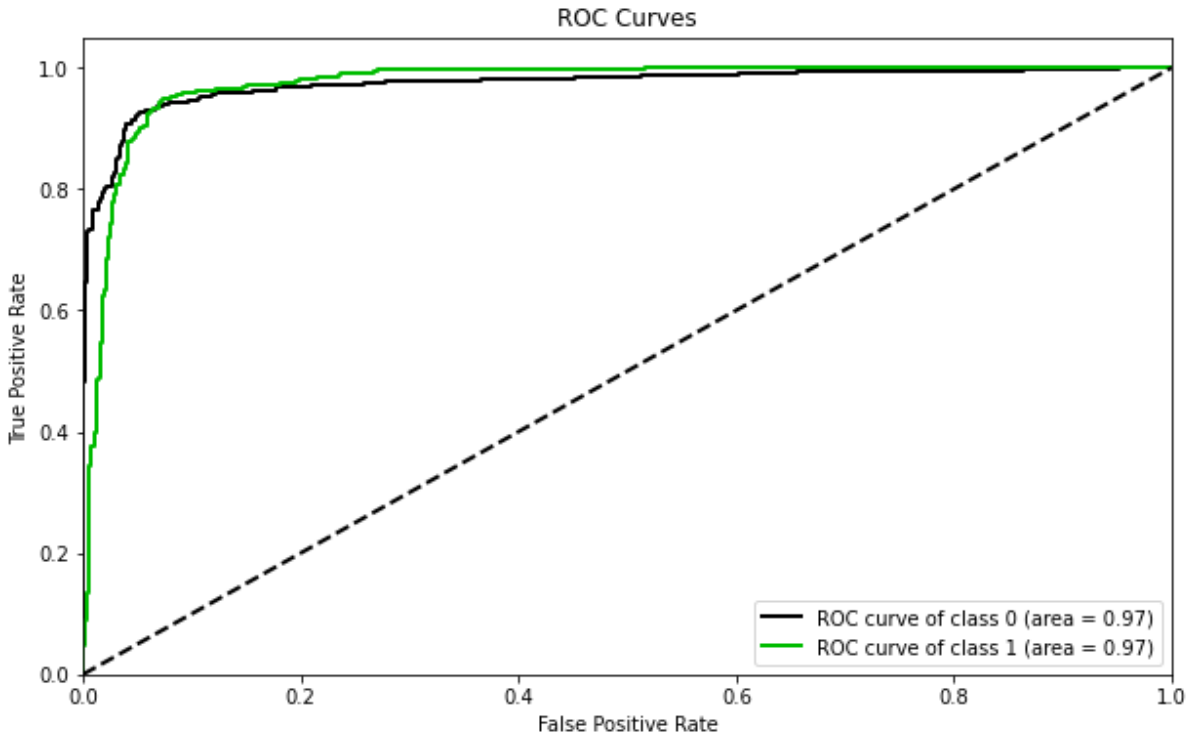

Fig 3.13 LR ROC Curves

B. Decision Tree (DT) Classifier

DT is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

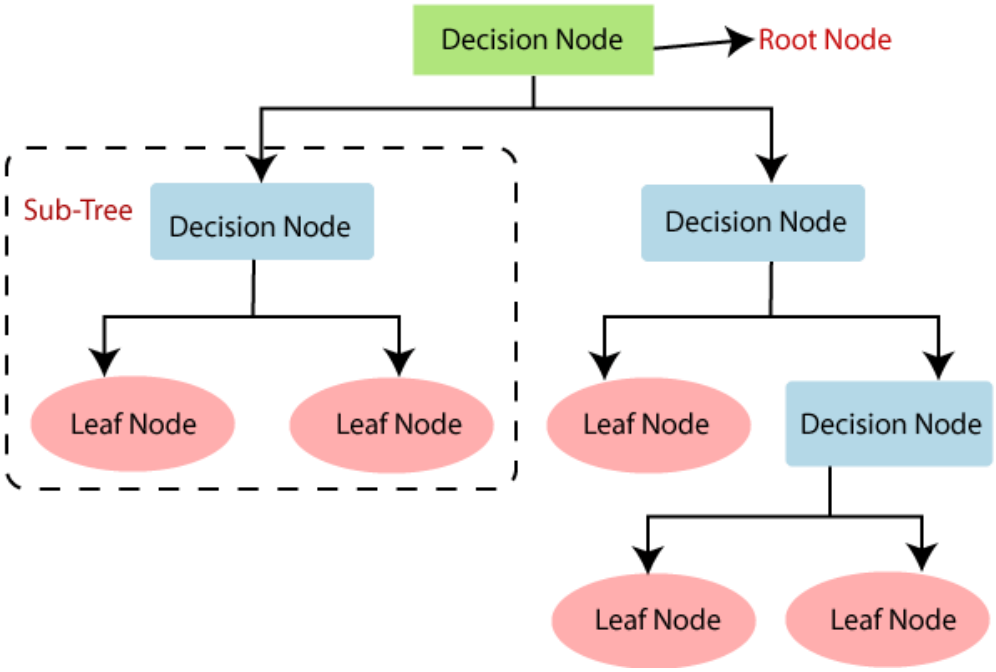

Fig 3.14 General Structure of Decision Tree

There are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The ROC Curves of the DT in our model is:

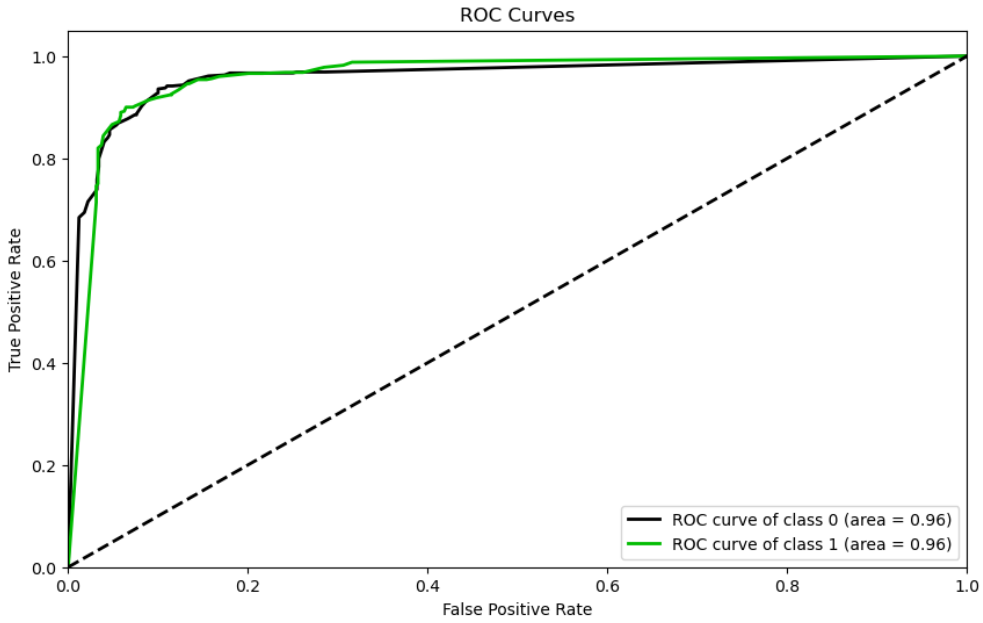

Fig 3.15 DT ROC Curves

DT is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

Interpreting the output of DT:

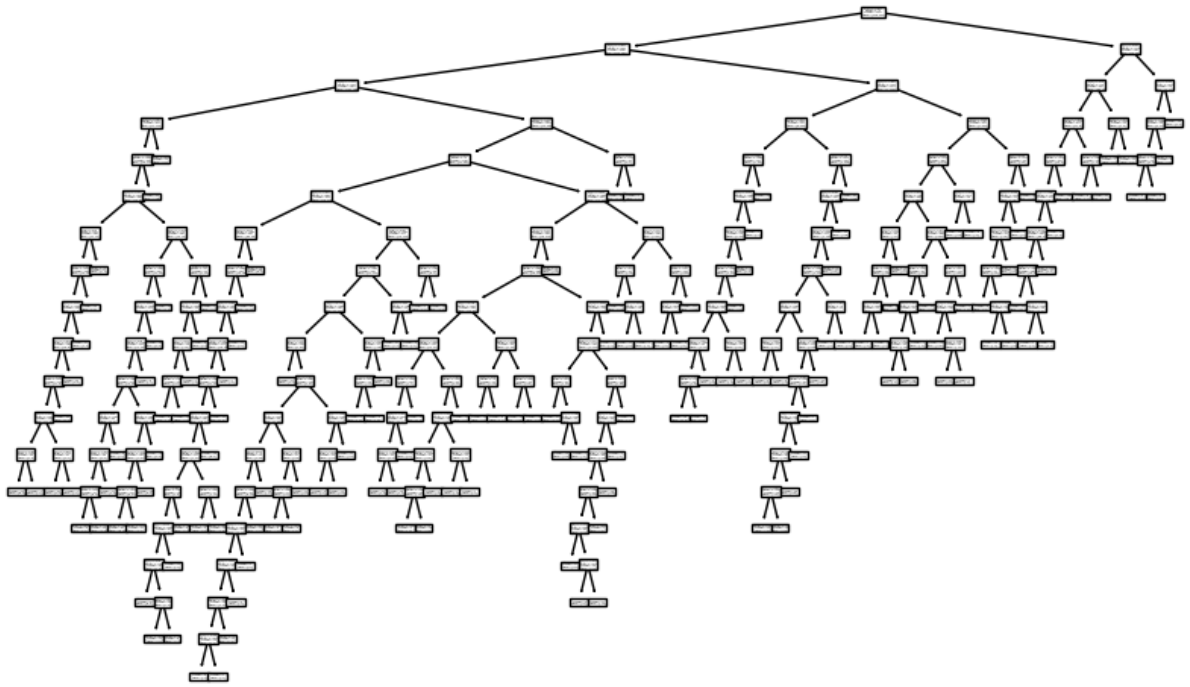

Fig 3.16 Interpreted Output of DT

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

C. Random Forest (RF) Classifier

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

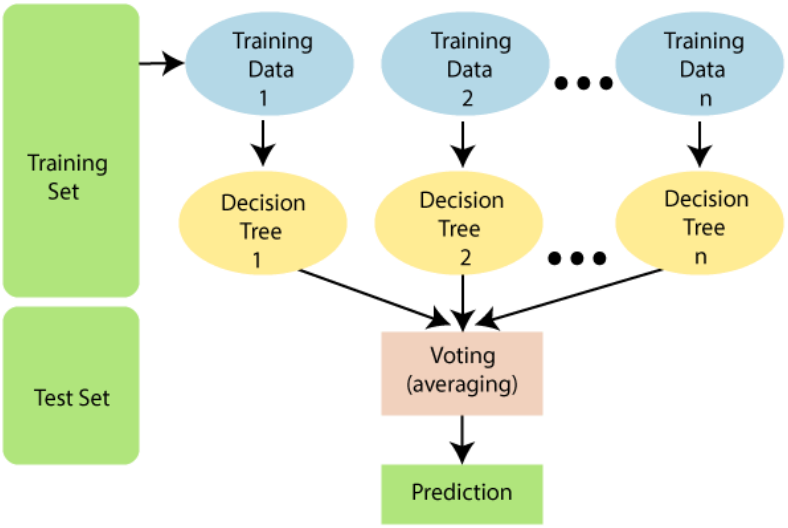

Fig 3.17 Working of RF Algorithm

The graph of ROC for random forest classifier is:

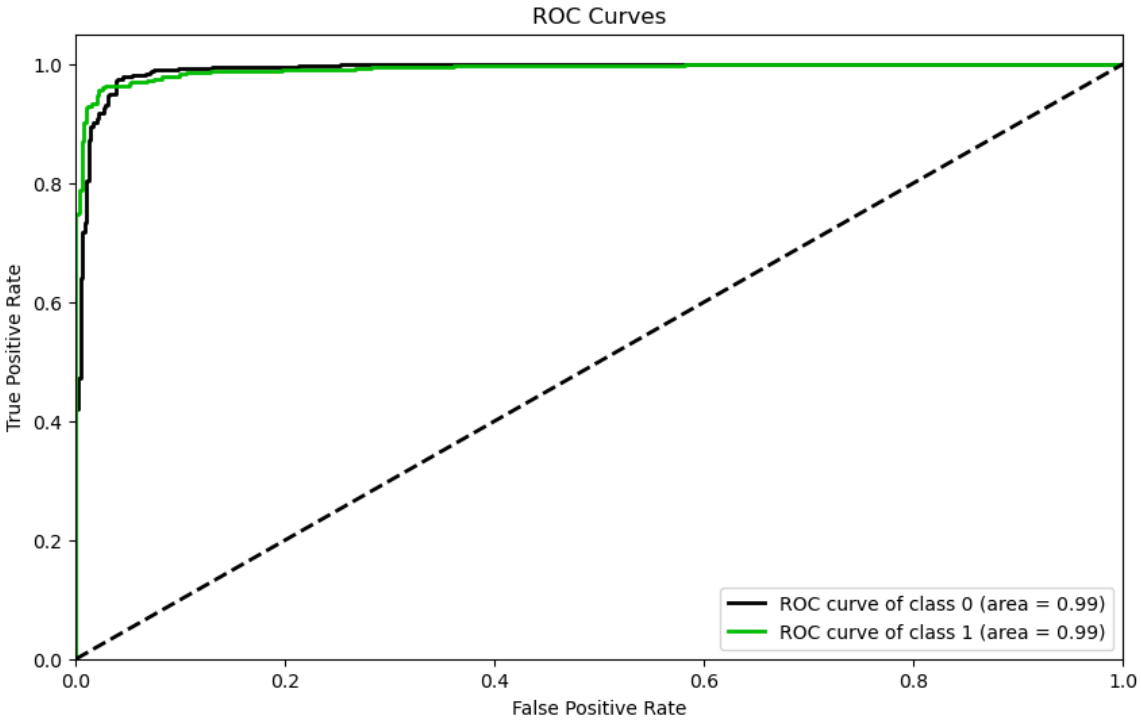

Fig 3.18 RF ROC Curves

Interpreting the output of RF as:

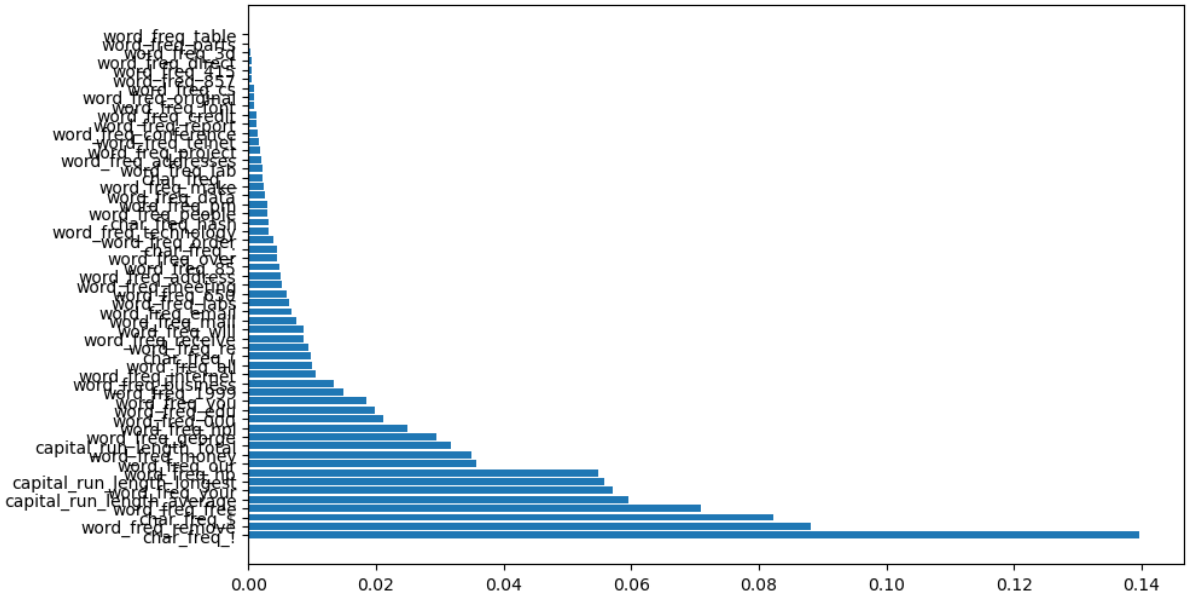

Fig 3.19 Interpreted Output of RF

D. Naïve Bayes (NB) Classifier

NB algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.

NB Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability. The formula for Bayes' Theorem is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Here, P(A|B) is Posterior Probability.

P(B|A) is Likelihood Probability.

P(A) is Prior Probability.

P(B) is Marginal Probability.

The ROC curves of NB look like:

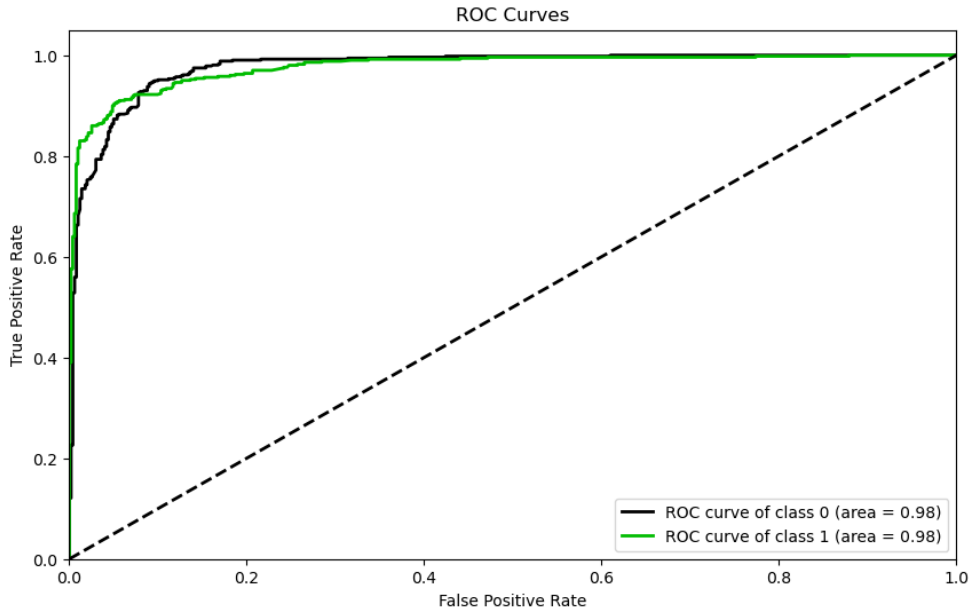

Fig 3.20 NB ROC Curves

E. Support Vector Machine (SVM)

SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

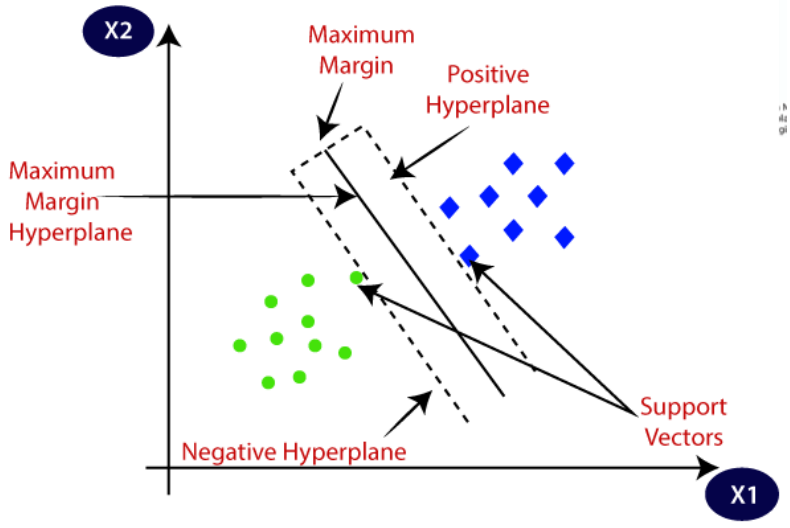

Fig 3.21 SVM Graph

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

In diagram, there are two different categories that are classified using a decision boundary or hyperplane.

SVM algorithm can be used for Face detection, image classification, text categorization, etc.

The ROC graph for the SVM is:

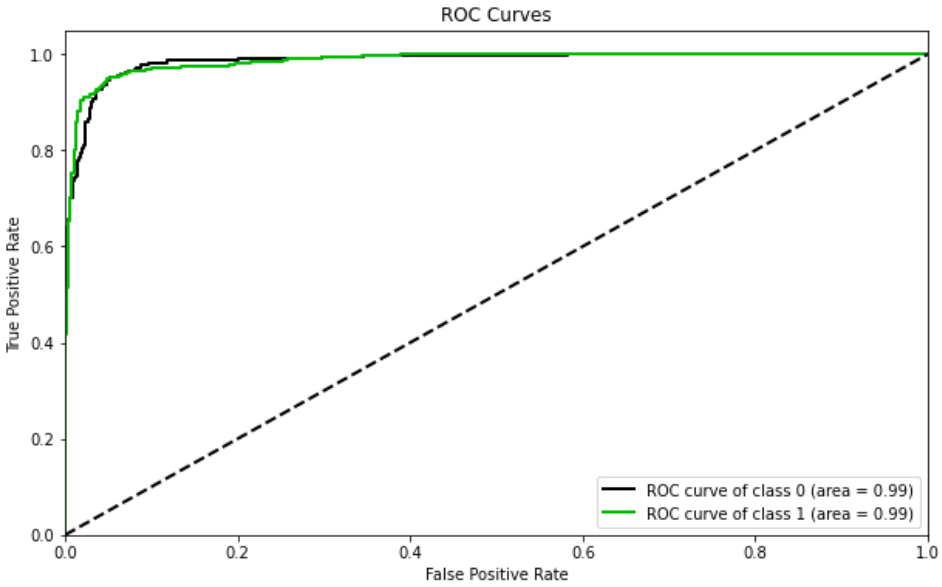

Fig 3.22 SVM ROC Curves

F. K-Nearest Neighbours (KNN)

KNN is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm.

KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a KNN algorithm. With the help of KNN, we can easily identify the category or class of a particular dataset.

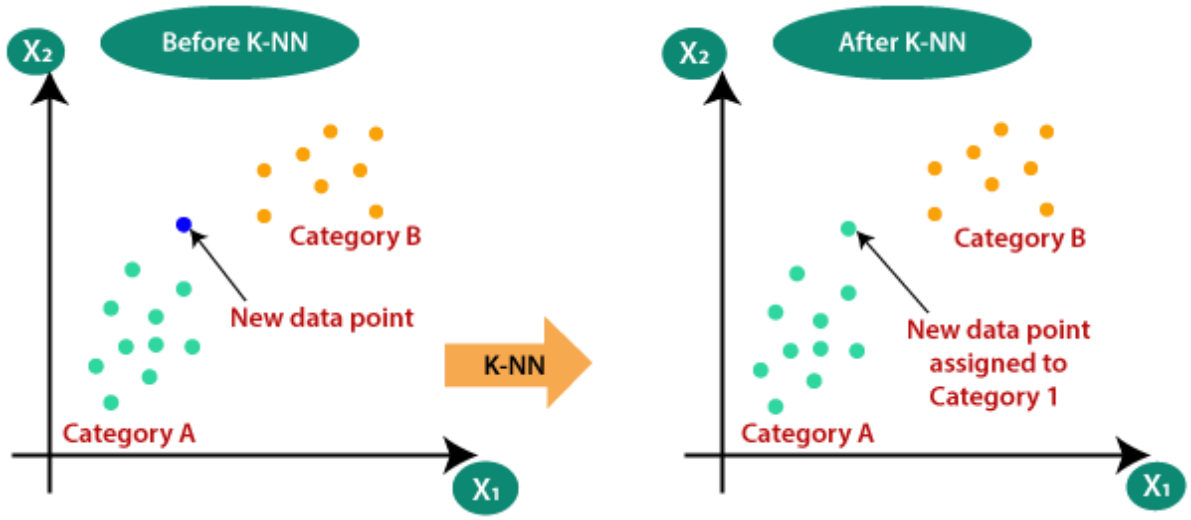

Fig 3.23 KNN Diagram

The ROC plot for that,

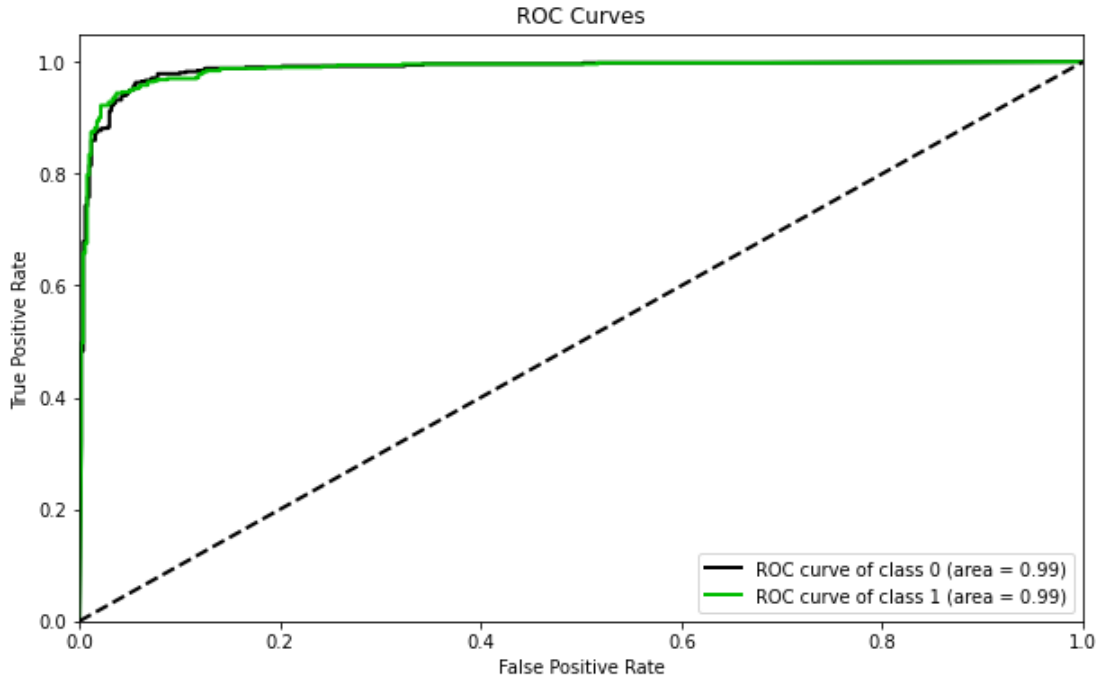

Fig 3.24 KNN ROC Curves

G. Gradient Boosting (GB)

Machine learning is one of the most popular technologies to build predictive models for various complex regression and classification tasks. GB is considered one of the most powerful boosting algorithms.

Although, there are so many algorithms used in machine learning, boosting algorithms has become mainstream in the machine learning community across the world. Boosting technique follows the concept of ensemble learning, and hence it combines multiple simple models (weak learners or base estimators) to generate the final output. GB is also used as an ensemble method in machine learning which converts the weak learners into strong learners.

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model.

Fig 3.25 GB Method

The ROC plot for GB algorithms:

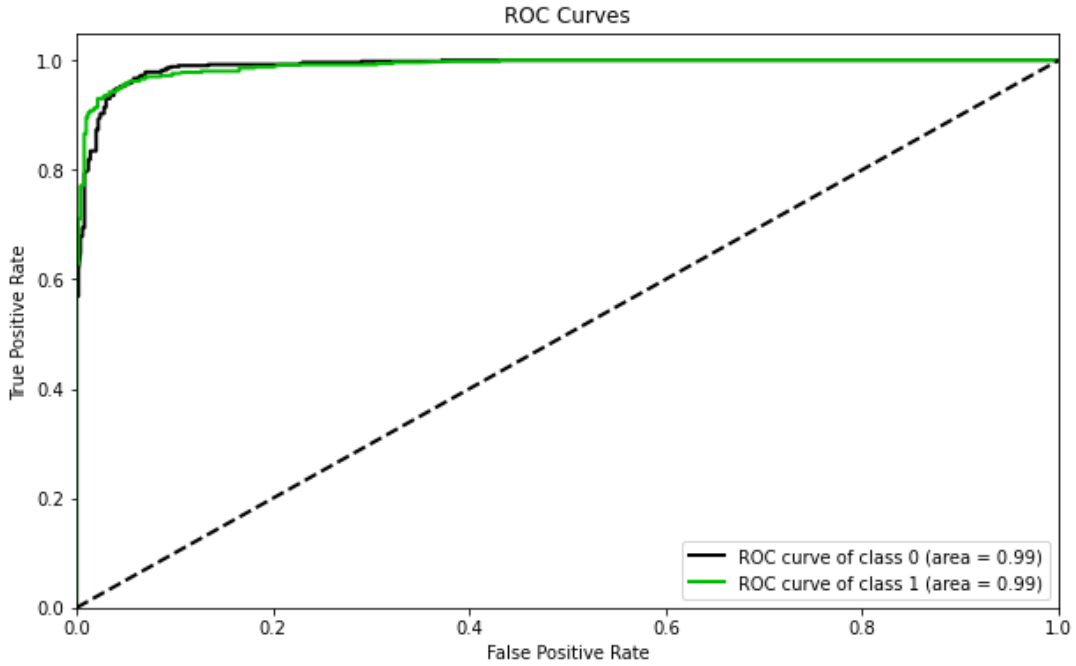

Fig 3.26 GB ROC Curves

H. Extreme Gradient Boosting (XGB)

XGB is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGB models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGB. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results.

The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

XGB Features The library is laser-focused on computational speed and model performance, as such, there are few frills.

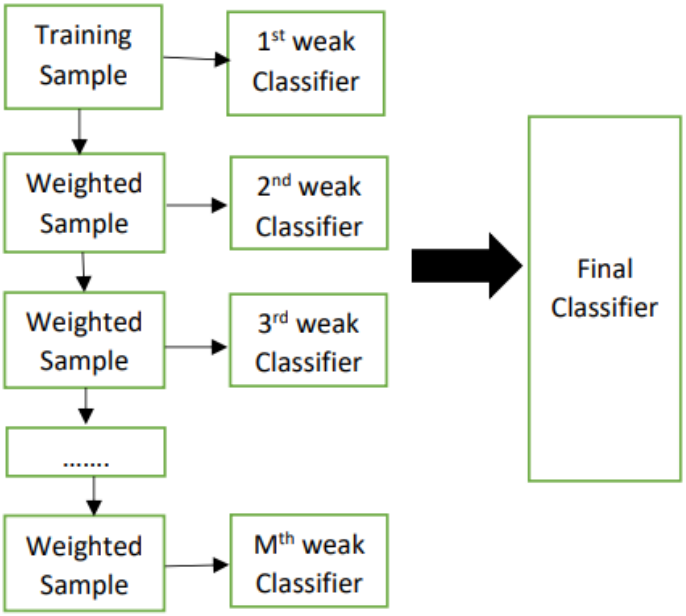

Fig 3.27 XGB Work

System Features:

- For use of a range of computing environments this library provides-
- Parallelization of tree construction
- Distributed Computing for training very large models
- Cache Optimization of data structures and algorithm

The ROC plot seems as:

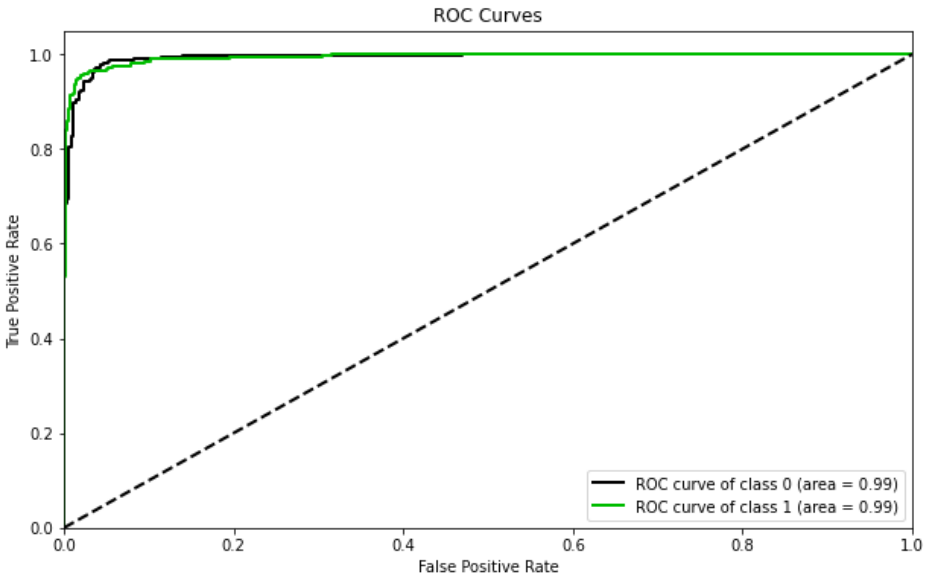

Fig 3.28 XGB ROC Curve

Plotting Confusion-Matrix of all predictive models:

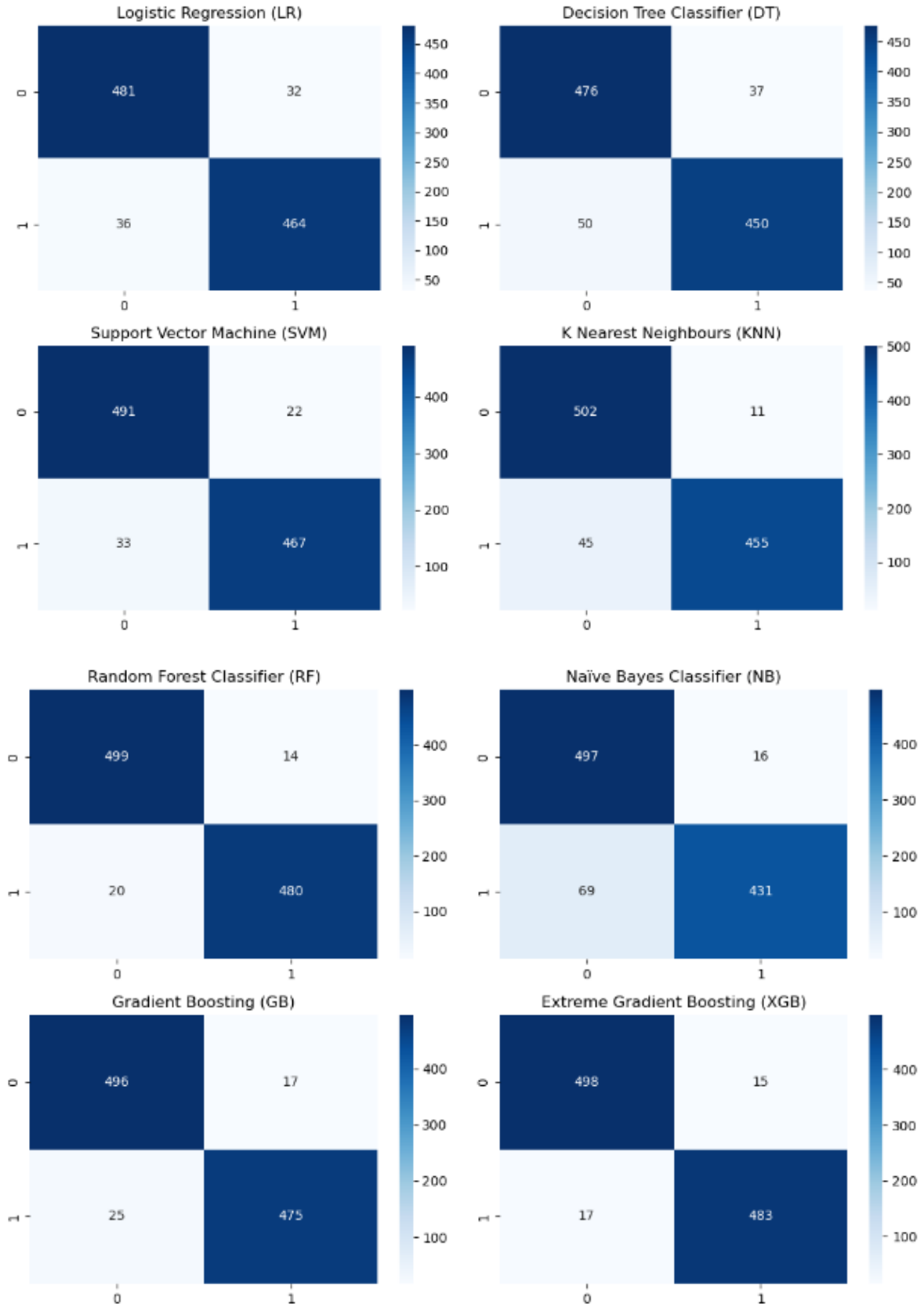

Fig 4.1 Confusion Matix of all models

Comparing all models scores using table as:

Comparison Table

Criteria	Accuracy	Precision	Recall	F1 – Score	AUC-ROC Score
LR	93.3	93.3	93.3	93.3	97.5
DT	91.4	91.4	91.4	91.4	95.9
RF	96.6	96.6	96.6	96.6	99.2
NB	91.6	92.1	91.6	91.6	97.6
SVM	94.6	94.6	94.6	94.6	98.5
KNN	94.5	94.7	94.5	94.5	98.5
GB	95.9	95.9	95.9	95.9	99.1
XGB	96.8	96.8	96.8	96.8	99.3

Email has been the most important medium of communication nowadays, through internet connectivity any message can be delivered to all over the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank, related to money or anything that causes destruction to single individual or a corporation or a group of people. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information.

Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company. In the future this system can be implemented by using different algorithms and also, more features can be added to the existing system.

From the above comparison table, we can conclude that Extreme Gradient Boosting gives the best model out of all with accuracy of 96.5 and AUC-ROC score of 99.4.

5.1 Future Scope

Review spam detection is essential since it can ensure justice for the sellers and retain the trust of the buyer on the online stores. The algorithms developed so far have not been able to remove the requirement of manual checking of the reviews. Hence there is scope for complete automation of spam detection systems with maximum efficiency. With the growing popularity of online stores, the competition also increases. The spammers get smarter day by day and spam reviews become untraceable. It is necessary to identify the spamming techniques in order to produce counter algorithms.

5.2 Conclusion

The Dataset was quite small totalling around 4600 samples & after pre-processing 14.6% of the data samples were dropped. The samples were slightly imbalanced after processing, hence SMOTE Technique was applied on the data to balance the classes, adding 16.7% more samples to the dataset. Visualizing the distribution of data & their relationships, helped us to get some insights on the relationship between the feature-set. Feature Selection/Elimination was carried out and appropriate features were shortlisted. Testing multiple algorithms with fine-tuning hyperparameters gave us some understanding on the model performance for various algorithms on this specific dataset. The Random Forest Classifier & XG-Boost performed exceptionally well on the current dataset, considering Precision Score as the key-metric. Yet it is wise to also consider a simpler model like Logistic Regression as it is more generalisable & is computationally less expensive, but comes at the cost of slight misclassifications.

REFERENCES

[1] Rajesh Kumar J, Sudarshan P and Mahalakshmi G. Email Spam Detection using Machine Learning Techniques. In June 2021, International Advanced Research Journal in Science, Engineering and Technology (IARJSET), Vol. 8, Issue 6.

[2] Isra'a AbdulNabi and Qussai Yaseen. Spam Email Detection Using Deep Learning Techniques. In 2021, The 2nd International Workshop on Data-Driven Security (DDSW 2021) March 23 - 26, 2021, Warsaw, Poland. Department of Computer Information Systems, Jordan University of Science and Technology, 3030, Irbid 22110, Jordan.

[3] Thashina Sultana, K A Sapnaz, Fathima Sana, Mrs. Jamedar Najath. Email based Spam Detection. In June 2020, International Journal of Engineering Research & Technology (IJERT), Vol. 9, Issue 06.

[4] Nikhil Kumar, Sanket Sonowal and Nishant. Email Spam Detection Using Machine Learning Algorithms. In July 2020, Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020) IEEE Xplore.

[5] V.Christina, S.Karpagavalli and G.Suganya. Email Spam Filtering using Supervised Machine Learning Techniques. In December 2010, International Journal on Computer Science and Engineering (IJCSE), Vol. 02, No. 09.

[6] Mangena Venu Madhavan, Sagar Pande, Pooja Umekar, Tushar Mahore and Dhiraj Kalyankar. Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches. In 2020, IOP Conf. Series: Materials Science and Engineering, ICCRDA 2020.

[7] Sunidhi Pandey, Shantanu Singh Chandel and Prof. Kunal Kumar. Email Spam Detection Using Supervised Algorithms. In April 2023, International Research Journal of Modernization in Engineering Technology and Science (IRJMETS). Volume:05/Issue:04/April-2023.