e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

EMAIL SPAM DETECTION USING SUPERVISED ALGORITHMS

Sunidhi Pandey*1, Shantanu Singh Chandel*2, Prof. Kunal Kumar*3

*1,2,3Department Of Information Technology Govt. Engineering College, Bilaspur, India.

DOI : https://www.doi.org/10.56726/IRJMETS36685

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 Neighbor, Gradient Boosting, Extreme Gradient Boosting.

I. INTRODUCTION

Nowadays, Internet is a big part of our daily life. With the increased use of the internet, the spam using Email or electronic mail is also increased. Email spam is really diverse which contains several types of concepts like advertisement of the product or their sites, fast money-making schemes, pornography, etc. Due to that unwanted spam emails, users storage, time and speed to work get affected. It also blocks and contains misfortunes for the system. Machine Learning algorithms are one of the best approaches to deal with that issue. Specially, when we consider Supervised Learning we get a broad range of algorithms as regression, Support Vector Machine (SVM), Decision Tree, and many others. Using these several algorithms we can create a model with can help us to identify the email classification as Spam or Ham.

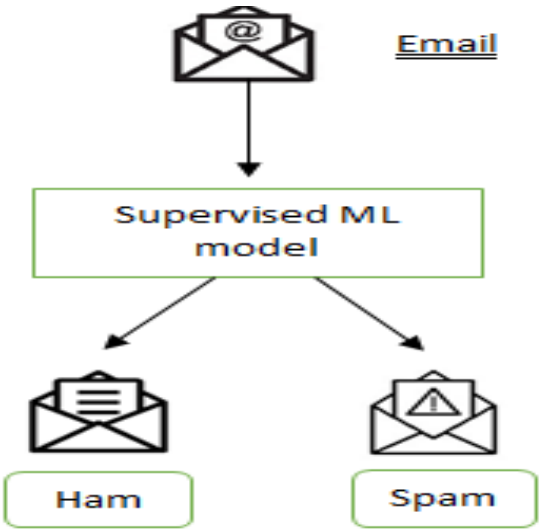


Fig. 1. Classification into Spam and Ham mails

According to search, the term 'ham' was originally coined by the Spam Byes sometime around 2001 and is currently defined and understood to be "E-mail that is generally desired and isn't considered spam." So, "Ham" is e-mail that is not Spam. In other words, "non-spam", or "good mail". It should be considered a shorter, snappier synonym for "non-spam".

II. LITERATURE SURVEY

In the paper[1], 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.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

In the paper[2], author used the deep learning technique approach to create model. Experimental result of the author shows that Bert-base-cased transformer model is best fit model with high accuracy and F1 score. For that author uses NLP task methodology with its five main phases as: data collection, data pre-processing, feature extraction, model training, and model evaluation.

In the paper[3], author work on the strategy of frequently repeated words in spam word cloud as well as in ham word cloud. Using NLKT and proposed system of Machine learning as Naïve bayes algorithms, author create model to detect ham or spam email. For it, author used to calculate Term frequency and Inverse Document frequency which help model to built in much better way.

In the paper[4], author use a broad way to classify. Author used Stop word, Tokenization, and Bag of words for Data preprocessing. Also work with many different ML algorithms as SVM, KNN, Naïve Bayes, Decision Tree, Random Forest, AdaBoost and Bagging Classifier. Compare all the model and conclude that the Multinominal Naïve Bayes gives the best outcome out of all.

III. METHODOLOGY

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:

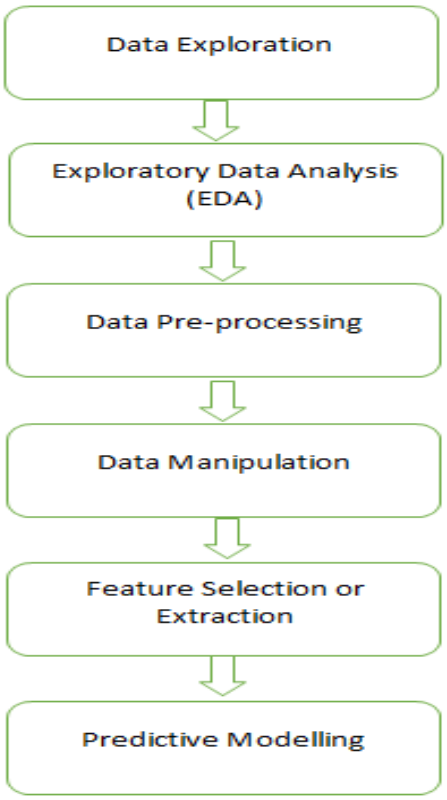


Fig. 2. Strategic Plan of Action Steps

1. Data Exploration:

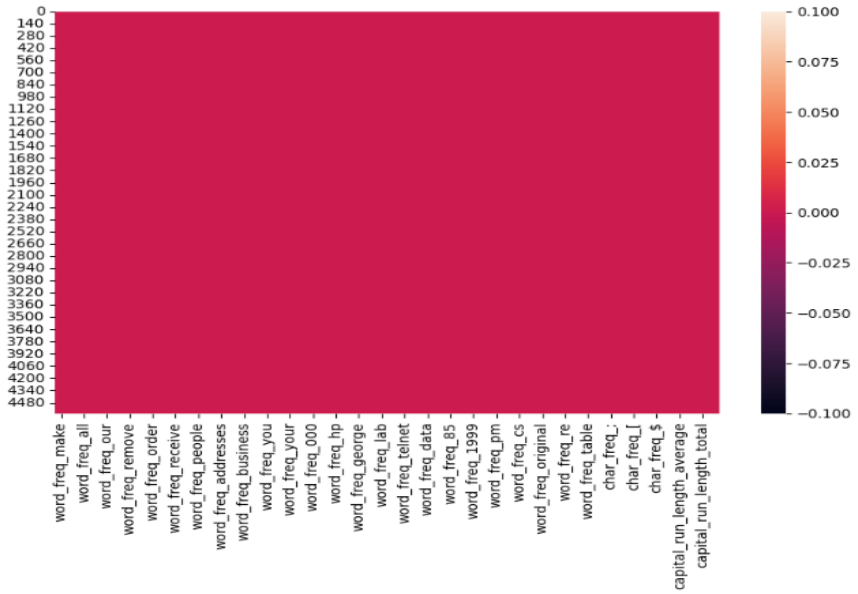
The Data collection of our spam email came from postmasters and individuals who had filled spam. And collection of ham emails came from filed work and personal e-mails. The dataset taken from the UCI ML repository, contains about 4600 emails labelled as spam or ham.

2. Exploratory Data Analysis (EDA):

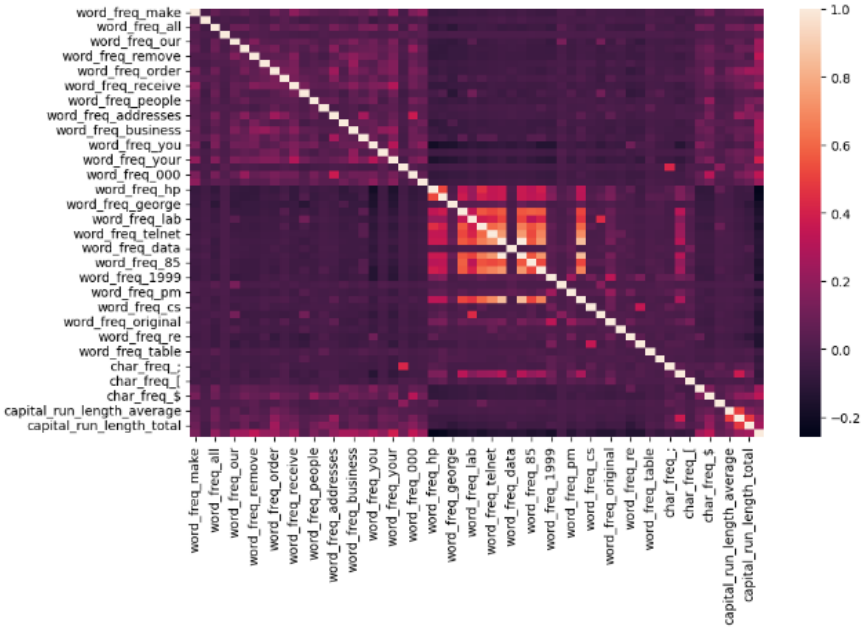
It is an approach to analyze the dataset using visualize techniques. It helps to discover patterns, trends, characteristics, assumptions, etc. For the spam dataset we have few EDA as:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

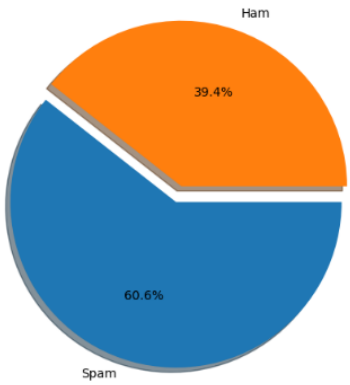
A. Heatmap for Null values in Dataset -



B. Heatmap for Correlation of Dataframe -



C. Analyze the target variable distribution -



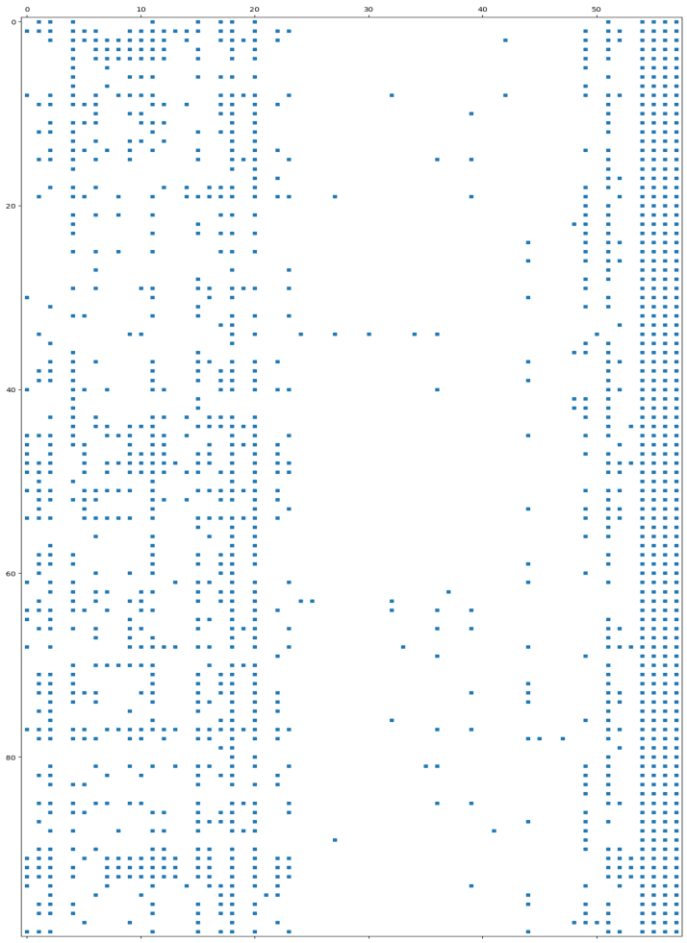
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

Here,

Spam = 60.6%

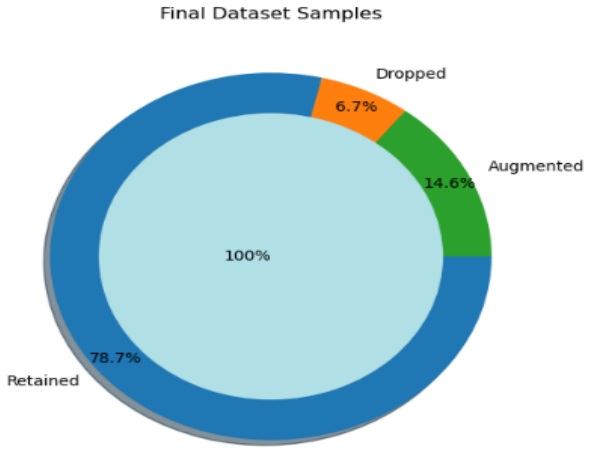
Ham = 39.4%

D. Visualizing the Sparse Matrix -



3. Data pre-processing:

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 preprocessing is:



In final dataset sample:

6.7% - Dropped

14.6% - Augmented

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

And 78.7% - Retained

So, the final dataset after cleanup has 58 samples & 4601 rows.

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,)

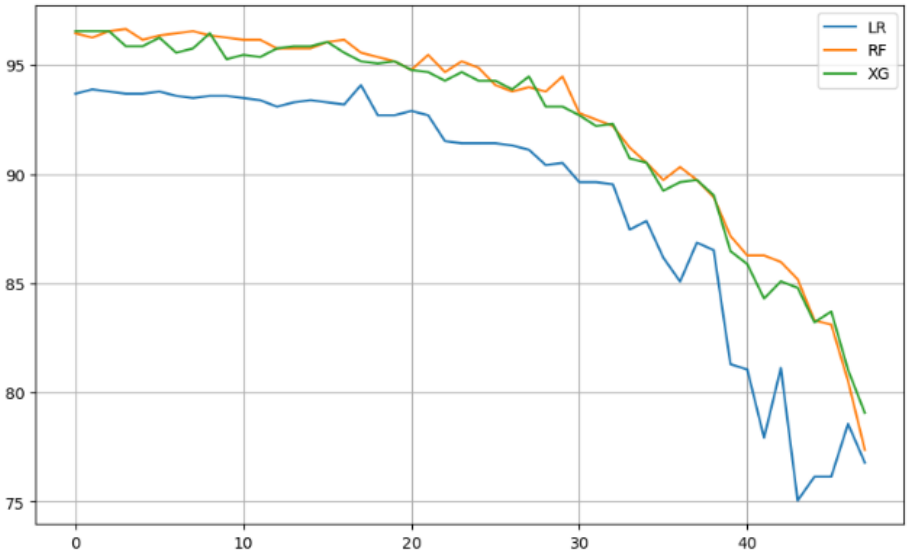
Testing set ---> (1013, 57) (1013,)

5. Feature Selection/Extraction:

Initially, we check the correlation of the dataset. We find some multicollinearity. To fix multicollinearity we use 3 different techniques as:

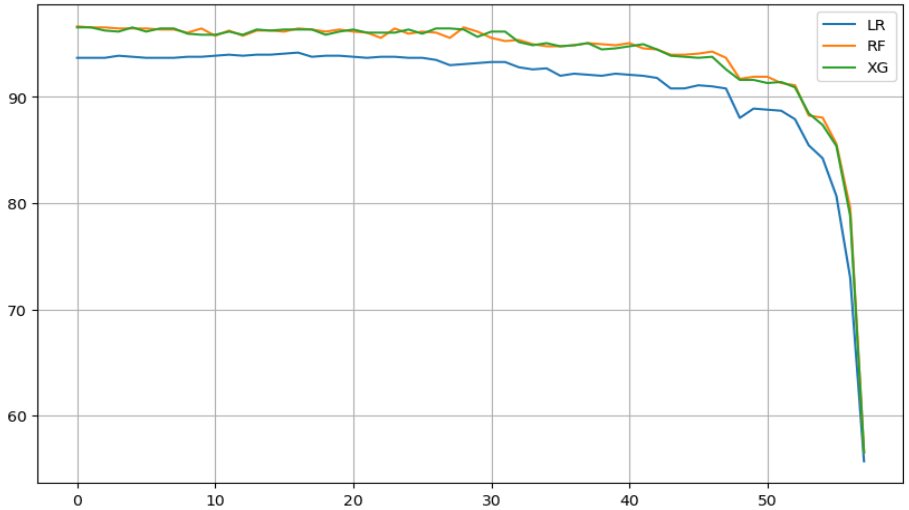
A. Manual Method – VIF

VIF is the abbreviation of Variance Inflation Factor. VIF helps to measure the strength of correlation between the independent variables.



B. Automatic Method – RFE

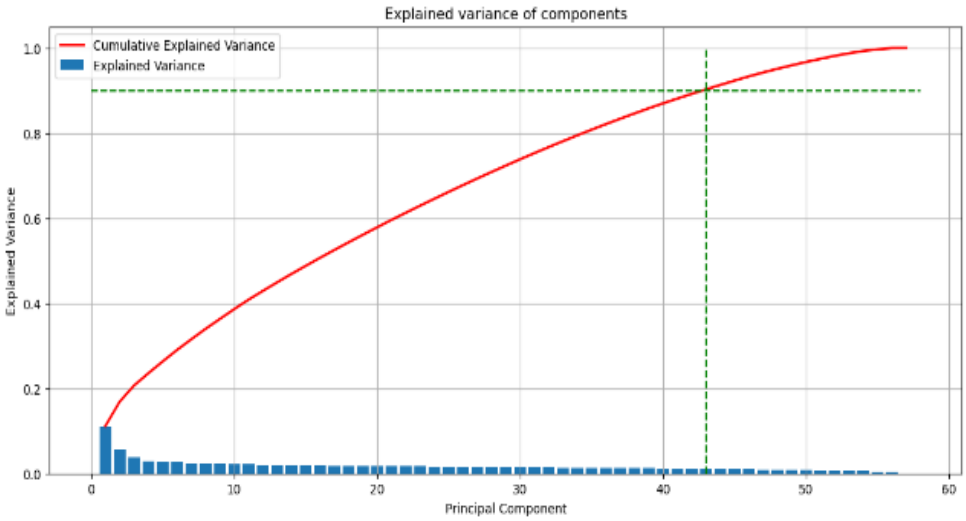
RFE is the abbreviation for Recursive Feature Elimination. It is a Wrapper method. It is a transformer estimator, which means it follows the familiar fit or transform pattern.



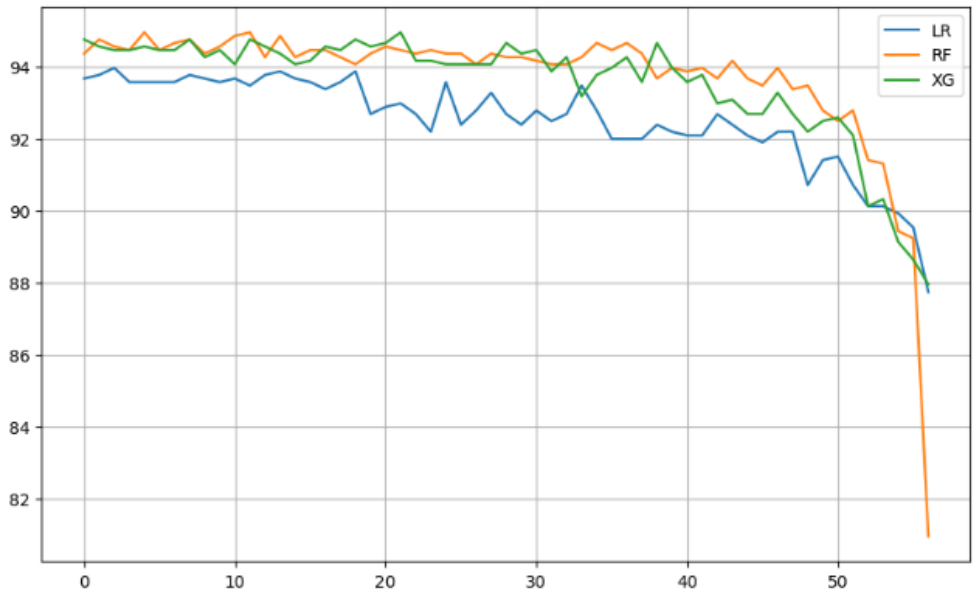
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

C. Decomposition method – PCA

PCA is the abbreviation of Principle Component Analysis. It is one of the popular unsupervised ML technique which is used for reducing the dimensionality of the data. For concluding PCA, the Explained Variance of Components are:



After performing the PCA transformation, we get the final graph as:



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

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

6. Predictive Modelling:

It is a mathematical process which is used to predict future behavior by using statistical techniques. It gives the solution for data-mining technologies by analyzing the past and current data and create a model which helps to predict the future outcomes.

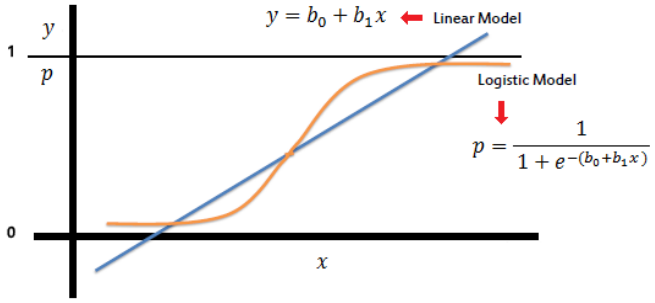
For our spam dataset we use different Supervised machine learning algorithms as:

A. Logistic Regression (LR):

It is one of the famous Supervised ML algorithms. It is used to predict the output of a categorical dependent variable. It can be either Yes or No, 1 or 0, True or False, etc. But it gives us a probabilistic value which lies between 0 and 1. It helps to solve the classification problems.

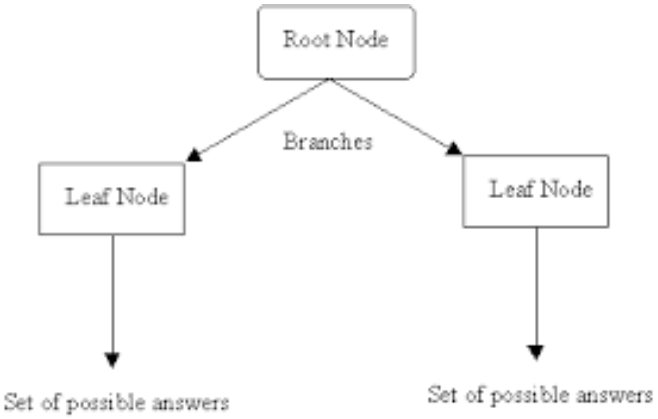
It is similar to Linear Regression but not same. It forms a "S" shapes logistic function in graph.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com



B. Decision Tree (DT):

Decision Tree also comes under the supervised algorithms. It is used for regression as well as classification. It forms tree like structure to create set of possible answers.



C. Random Forest (RF):

It builds different Decision tree samples and take the majority one for classification and average one for the regression. It helps to handle the dataset which have continuous variable. It also works for classification as well as for regression problems.

D. Naïve Bayes (NB):

It is a Supervised algorithms which is based on Bayes Theorem and help to work in classification problems. It uses the bayes law as:

Here,

P(A|B) is posterior probability.

P(A) is prior probability.

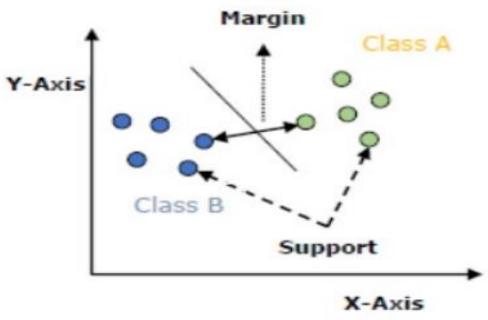
And, P(B) is marginal probability.

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

P(B|A) is likelihood probability.

E. Support Vector Machine (SVM):

It is one of the popular Supervised learning algorithms mainly used for classification. The main aim of this algorithm is to create a best line (margin) and hyperplane to support different types of vectors that helps to classify the points or vectors.

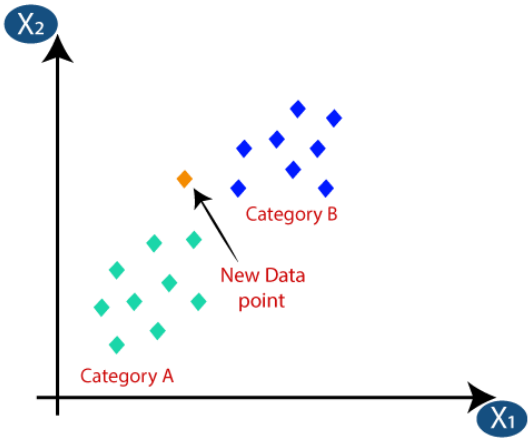


e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com

Here, Class A and Class B is Negative and Positive Hyperplane respectively.

F. K Nearest Neighbors (KNN):

It is a non-parametric algorithm which assumes to classify the new data point according to the similarity of the available or past case. It is mostly used for classification problems.



Here, this new data point goes with the Category A as it is similar and nearest to A.

G. Gradient Boosting (GB):

It is a popular boosting ML algorithm which is a kind of ensemble technique. It is used for classification and regression tasks. It helps to train the model sequentially and each model tries to correct the previous model.

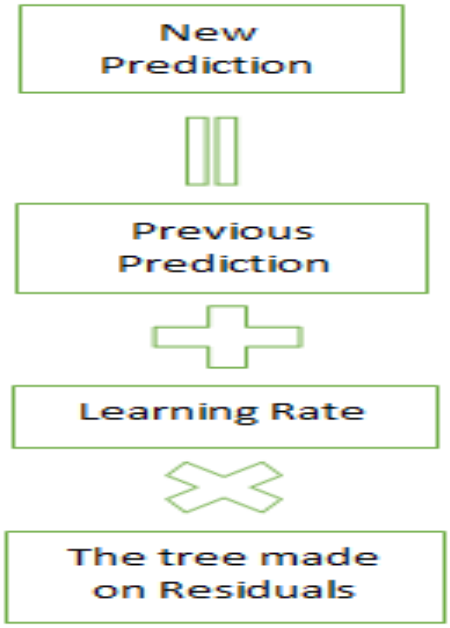
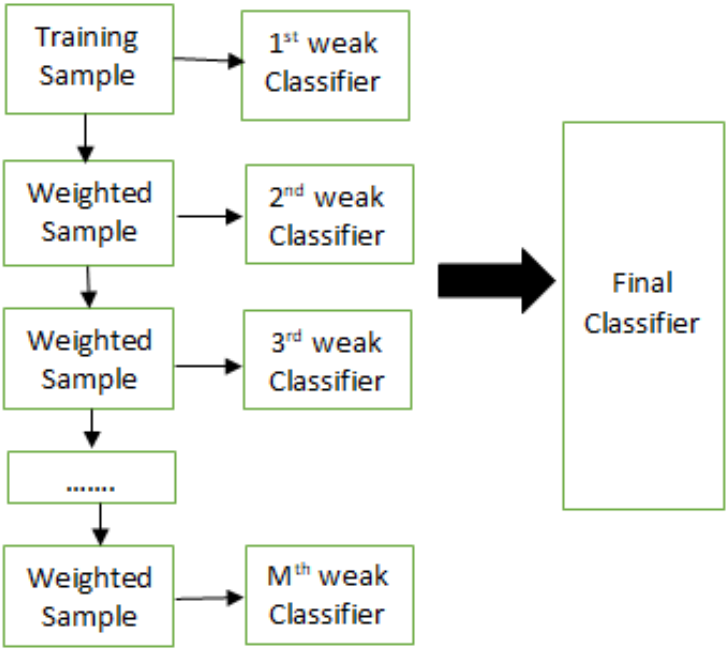


Fig 3. Gradient Boosting Method

H. Extreme Gradient Boosting (XGB):

It is the implementation of the Gradient Boosted Decision Tree. It was written in C++. It helps to create the sequential decision trees and improves the speed and performance of the model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:04/April-2023 Impact Factor- 7.868 www.irjmets.com



IV. RESULT

The created model is built using eight different supervised machine learning algorithms. We check and classify each model accuracy and results using parameters.

We get the comparison table as:

Comparison Table

Criteria	Acc.	Pre.	Rec.	F1	A-R
LR	93.8	93.8	93.8	93.8	97.4
DT	91.0	91.0	91.0	91.0	96.7
RF	96.8	96.9	96.8	96.8	99.1
NB	92.5	92.8	92.5	92.5	97.7
SVM	94.5	94.5	94.5	94.5	98.6
KNN	94.0	94.2	94.0	94.0	98.3
GB	96.1	96.1	96.1	96.1	99.2
XGB	96.5	96.5	96.5	96.5	99.4

Here,

Acc. – Accuracy

Pre. – Precision

Rec. – Recall

F1 – F1 Score

A-R – AUC-ROC Score

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.

V. 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 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.

VI. CONCLUSION

The Dataset was quiet 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. Visualising 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 wise to also consider simpler model like Logistic Regression as it is more generalisable & is computationally less expensive, but comes at the cost of slight misclassifications.

VII. 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 Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.