Group14 - SPAM-HAM Classification

Omkar Raghatwan	Harshdeep Telang	Flavin Dabre	Ishaan Shivhare
19070059	191070024	201070905	191070070
T.Y.B.Tech CS	T.Y.B.Tech CS	T.Y.B.Tech CS	T.Y.B.Tech CS
VJTI, matunga.	VJTI, matunga.	VJTI, matunga.	VJTI, matunga

MOTIVATION :

Detecting spam alerts in emails and messages is one of the main applications that every big tech company tries to improve for its customers. Apple's official messaging app and Google's Gmail are great examples of such applications where spam detection works well to protect users from spam alerts.

ABSTRACT:

Automatic filtering of spam emails becomes an essential feature for a good email service provider. To gain direct or indirect benefits organizations/individuals are sending a lot of spam emails. Such kind emails activities are not only distracting the user but also consume a lot of resources including processing power, memory and network bandwidth. The security issues are also associated with these unwanted emails as these emails may contain malicious content and/or links. Content based spam filtering is one of the effective approaches used for filtering.

Keywords:

ML: Machine Learning
SVC: Support Vector Classification,
NB: Naive Bayes,
DT: Decision Tree,
KNN: K-Nearest Neighbor.

INTRODUCTION:

Whenever you submit details about your email or contact number on any platform, it has become easy for those platforms to market their products by advertising them by sending emails or by sending messages directly to your contact number. This results in lots of spam alerts and notifications in your inbox. This is where the task of spam detection comes in.
Spam detection means detecting spam messages or emails by understanding text content so that you can only receive notifications about messages or emails that are very important to you. If spam messages are found, they are automatically transferred to a spam folder and you are never notified of such alerts. This helps to improve the user experience, as many spam alerts can bother many users.

Spam email may include malware as scripts or other executable file attachments. Spammers collect email addresses from chat rooms, websites, customer lists, newsgroups, and viruses,etc

Ham email is a term opposed to spam messages. Ham is then all "good" legitimate email messages, that is to say, all messages solicited by the recipient through an opt-in process

APPROACH:

PART 1:

- Exploring Data Source
- Data Preparation
- Exploratory Data Analysis

DATASET: Spam.csv

The dataset is taken from UCI Machine Learning.

The data set has 5 columns and 5572 rows. Out of these 5572 data points, *type* of 747 is labeled as spam and 4825 as 'ham' and contain 3 extra column as *Unnamed:2/3/4* which is redundant.

If the *type* has value ham, it means the text or message is not spam but if the value of *type* is spam then it means the *text* is spam and *text* are not in a chronological order.

EDA:

Using Visually attractive open source libraries like seaborn, matplotlib, scatterplot, violinplot, histograms,etc Before jumping into exploratory data analysis we've to make sure that data is ready to be used in ML algorithms and will work efficiently with mentioned libraries to give us insights about the data. For that we must follow given steps:

HANDLE missing data::

By deleting the particular row**:**

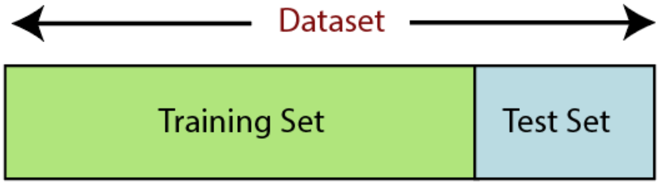
Used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.There are several methods but we have large data so this method is quick.

CLEAR VISION:

Make sure what is our target variable and proceed accordingly. You may add new variable which is somehow related to the target variable but is easy to understand, That way our model will work efficiently and we'll get desired results.

TRAIN/TEST Splitting:

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:



FEATURE SCALING

Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no variable dominates the other variable.

The steps includes:

Tokenization is the process of converting the normal text strings into a list of tokens (words that we actually want).
Example,

Raw message: "How are you doing?"
Tokenized message: [How, are, you, doing,?]

Vectorization is the process of converting each of the messages into a vector.

Steps for Vectorization:
1. Count how many times a word occurs in each message (Known as term frequency).
2. Weigh the counts, so that frequent tokens get lower weight (inverse document frequency).
3. Normalize the vectors to unit length, to abstract from the original text length
(L2 norm) by TF-IDF.

PART 2:

- Naïve Bayes Behind Spam or Ham
- Decision Tree algorithm
- SVC On Spam or Ham
- KNN on Spam or Ham
- Performance Measurement Criterion

Naive Bayes:

uses the Bayes’ Theorem and assumes that all predictors are independent. In other words, this classifier assumes that the presence of one particular feature in a class doesn’t affect the presence of another one.

Accuracy: 98.2655 %

Decision Tree:

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data.

Accuracy: 97.3%

SVC:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector.

Distance between the vectors and the hyperplane is called Margin.

Goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

Accuracy: 83.43%

KNN:

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Accuracy: 86.03%

Bagging & Boosting:

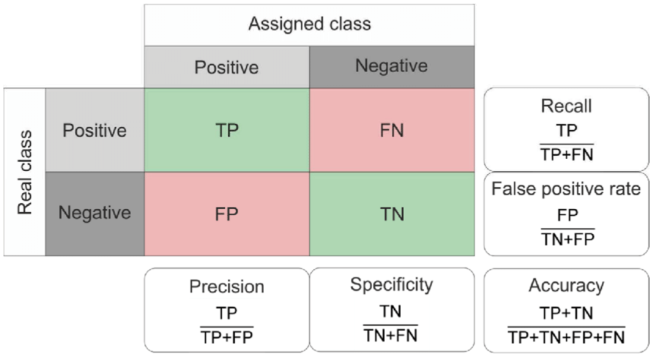
Bagging and Boosting are ensemble methods focused on getting N learners from a single learner.Bagging and Boosting make random sampling and generate several training data sets

Bagging works parallelly while Boosting works sequentially keeping record of previous iteration.

Accuracy: 97.5% and 96.5% respectively.

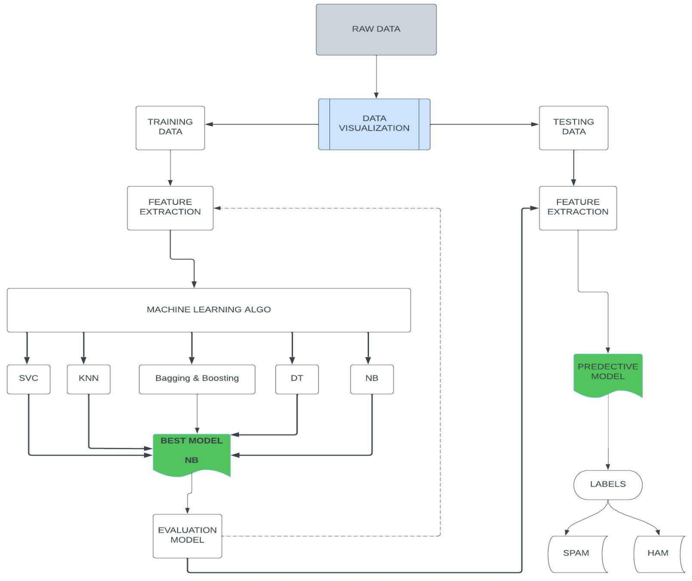
Confusion matrix:

A confusion matrix is a [2×2] matrix contains the number of *true positives*, *true negatives, false positives,* and *false negatives.* Using these 4 parameters we can get more precise information about the accuracy of our model.



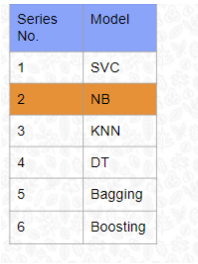
		Assigned class		
		Positive	Negative	
Real class	Positive	TP	FN	Recall $\frac{TP}{TP+FN}$
	Negative	FP	TN	False positive rate $\frac{FP}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Specificity $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

FLOWCHART :

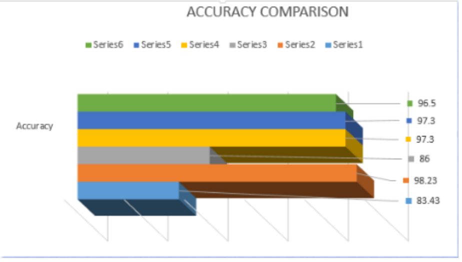


FLOW CHART OF OUR APPROACH

RESULTS:



Series No.	Model
1	SVC
2	NB
3	KNN
4	DT
5	Bagging
6	Boosting



REFERENCES:

- https://towardsdatascience.com/the-ultimate-guide-to-sms-spam-or-ham-detector-aec467aecd85
- ML Lecture slides(Seema Shrawne)
- Machine Learning
  Book by Tom M. Mitchell
- Pattern Recognition and Machine Learning by Christopher Bishop