Consensus Based Ensemble Model for Spam Detection

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Information Security

Submitted By

Paritosh Pantola
(801333015)

Under the supervision of:
Dr. Anju Bala
Assistant Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled "*CONSENSUS BASED ENSEMBLE MODEL FOR SPAM DETECTION*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Information Security* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Anju Bala and refer other researcher's work which are duly listed in reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Paritosh Pantola

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge

14/7/15

Dr. Anju Bala
Assistant professor
Computer science and
Engineering department

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. S. Bhatia)
Dean (Academic Affair)
Thapar University
Patiala

Acknowledgement

The successful completion of any task would be incomplete without acknowledging the people who made it possible and whose constant guidance and encouragement secured the success.

First of all I wish to acknowledge the benevolence of omnipotent God who gave me strength and courage to overcome all obstacles and showed me the silver lining in the dark clouds. With the profound sense of gratitude and heartiest regard, I express my sincere feelings of indebtedness to my guide **Dr. Anju Bala**, Assistant Professor, Computer Science and Engineering Department, Thapar University for her positive attitude, excellent guidance, constant encouragement, keen interest, invaluable co-operation, generous attitude and above all her blessings. She has been a source of inspiration for me.

I am grateful to **Dr. Deepak Garg**, Head of Department and **Dr. Prashant Singh Rana**, Assistant Professor, Computer Science and Engineering Department, Thapar University for the motivation and inspiration for the completion of this thesis.

I will be failing in my duty if I do not express my gratitude to **Dr. S.S. Bhatia**, Senior Professor and Dean of Academics Affairs in the University, for making provisions of infrastructure such as library facilities, computer labs equipped with internet facility, immensely useful for the learners to equip themselves with latest in the field.

Last but not the least I would like to express my heartfelt thanks to my parents and my friends who with their thought provoking views, veracity and whole hearted co-operation helped me in doing this thesis.

Paritosh Pantola

(801333015)

Abstract

In the era of Internet, Email is the best way for communication. Email spam is increasing day by day. Spam is the main reason for the financial loss on the internet, steal valuable information (bank account detail, password etc.), slow down internet bandwidth etc. Email spam is the mail that user does not want to receive. There are various existing algorithms for filtering the spam such as Naïve Bayes, Support vector machine, K-nearest neighbor and Decision tree etc. Email spam is also increasing day by day so existing algorithms will not be efficient in the future, hence, there is a need to combine two or more algorithms to enhance the performance of existing models using ensemble methods. Thus, various existing machine learning models have been explored and analyzed. The models have been evaluated based on various performance metrics. Further, the models have been compared with existing models to evaluate the accuracy.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

This chapter describes the basics of machine learning, its background and types of machine learning approaches. This chapter also discusses ensemble methods and its applications. This chapter also describes the main dataset used for the current work.

1.1 Background

Machine learning is a branch of computer science which basically deals with the recognition, categorization, and learning. It is applicable in all areas of artificial intelligence. Machine learning was first developed in 1950. It is a branch of artificial intelligence. It is basically used for studying big data. It is the study of how the computer realizes the behavior of the human being [1]. Machine Learning is the study of how computer simulation or realize human learning behavior, to get new knowledge or skills, to form the knowledge structure of the existing, to constantly improve their performance. It is the core of artificial intelligence, and its application used in all areas of artificial intelligence; it mainly uses inductive, comprehensive approaches [2].

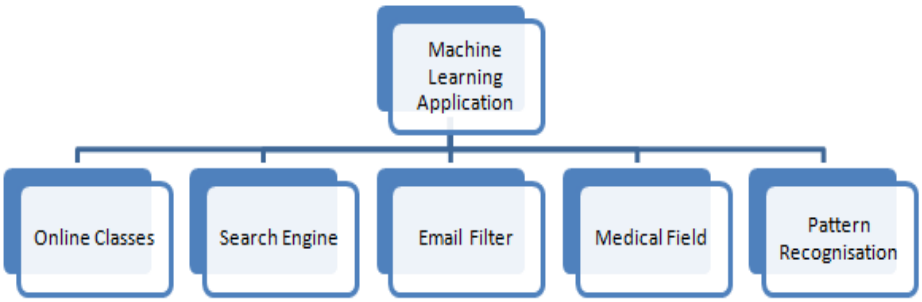

Figure 1.1: Application of Machine Learning

Figure 1.1 describes the various applications of machine learning. One main application of machine learning is email filtration. The email can be filtered by two methods one is machine learning another one is non-machine learning. The categorization of email filtration describes in Figure 1.2. Non-machine learning approach further divided into four categories i.e. heuristic approach, signature approach, hash-based, and traffic analysis while the machine learning approach divided into two types complete and complementary. The complete approach is further divided into unified model and ensemble model. Algorithms under

the unified model are the Bayesian algorithm, Support Vector machine, and neural network while algorithms under ensemble model stacked generalization, Boosting tree, and Hidden Markov model. The algorithms under complementary approach are Adaptive and trust network.

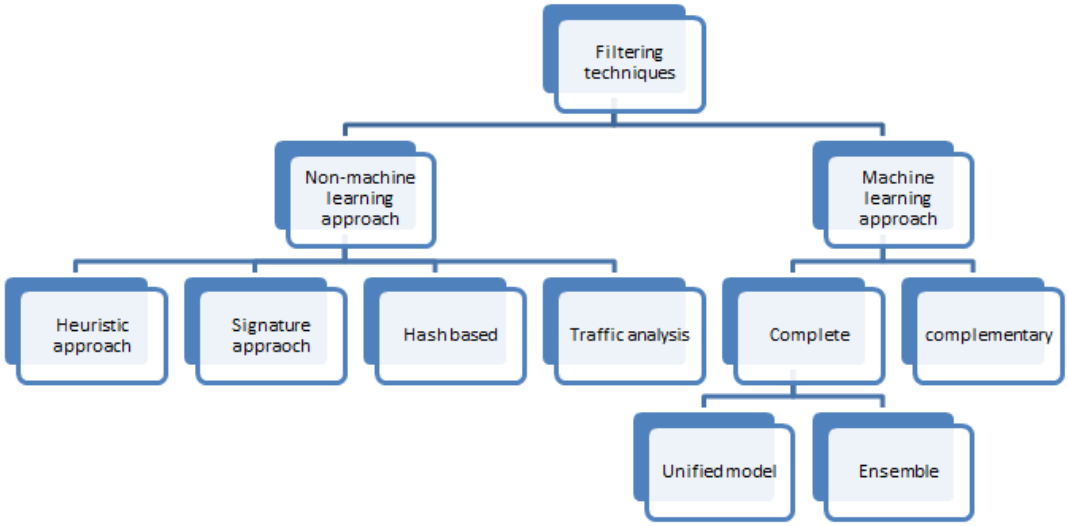

Figure 1.2: Email Filter Approach

Machine learning is divided into four categories [3].

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

1.2 Types of Machine Learning

1.2.1 Supervised Learning

Supervised learning uses some specific rules to predict the output of the given function. In supervised learning, the main aim is to predict the target features with the help of test features. Supervised learning is further divided into two categories.

- Classification
- Regression

Classification

Classification is the approach used for classifying the data. In this approach, target value is set either true (1) or false (0). This type of approach basically used when classification approach is used in the problem i.e. disease is malignant or not, email is spam or not. There are various applications for classification algorithm such as speech recognition, handwriting recognition, biometric, pattern identification, classification

of the document, search engine etc. Most common example of classification is email is spam or not.

List of algorithms supported by classification approach

- Flexible discriminant analysis
- K-nearest neighbor
- Neural network
- Support vector machine
- Decision Tree
- Naïve Bayes classifier

Regression

Regression is just opposite of classification. In classification approach, target variable are set either true or false while in regression target variable is set into real value. In regression, the output variable is continuous or real value. Regression is further divided into ten parts i.e. linear regression, logistic regression, ridge regression, lasso regression, ecologic regression, regression in unusual space, logic regression, bayesian regression, quantile regression, LAD regression and jack-knife regression.

List of algorithms supported by regression approach

- Neural network
- Generalized linear model
- Linear regression
- Decision tree
- Nonlinear regression

1.2.2 Un-Supervised Learning

Unsupervised learning is just opposite of supervised learning. In unsupervised learning, there is no output required. Unsupervised learning basically deals with clustering of data. Applications of unsupervised learning are compression of image, bio-informatics, association analysis etc.

1.2.3 Semi-Supervised Learning

It is combination of learning of data from the label as well as unlabeled data. It basically uses the strength of both the learning method, i.e. supervised as well as unsupervised. Application of semi-supervised learning is face recognition.

1.2.4 Reinforcement Learning

Reinforcement learning basically deals with agent to maximize the reward continuously .It basically interacts with environment.

Ensemble Model

Ensemble model is combination of two or more machine learning models (algorithms). Ensemble model is more efficient than the individual models. In machine learning, ensemble model is combining two or more models to get a better prediction, accuracy and robustness as compared to individual model separately. While performing, ensemble model initially put training dataset into different models after that, select the best model suited for the dataset. In this work, the analysis is done by using six machine learning methods i.e. Accuracy, Receiver operating characteristics (ROC) curve, Confusion matrix, Sensitivity, Specificity and Kappa value. After that implementation of k fold validation is done on best five models.

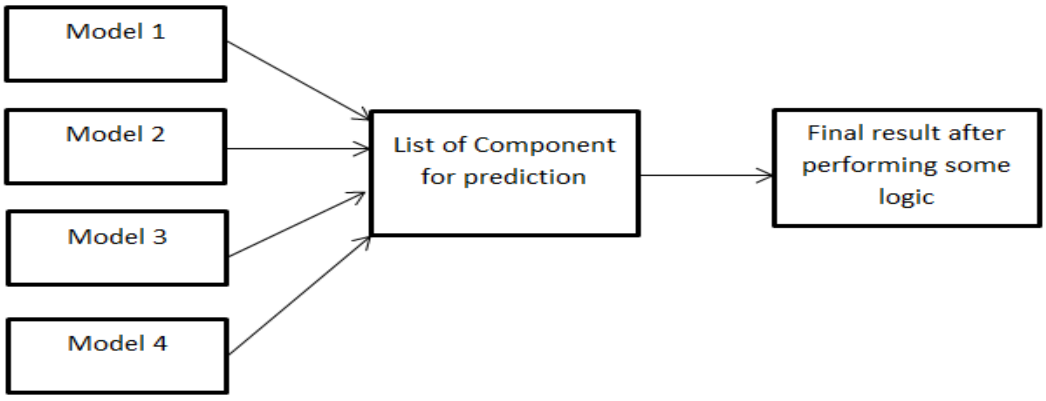

Figure 1.3: Process for Ensemble Model

1.4 Application of Ensemble model

1.4.1 Kernel Factory

Kernel factory is basically an ensemble method on kernel machine. A kernel is a method used in machine learning. It is basically used for pattern recognitions. In kernel method, there is a function called kernel function. Kernel is basically a similarity function which is used for determining, that how similar the two machine learning algorithms are. For example, suppose there is a task for text classification. This can be done by two ways, first one is take textual data as training data after that, calculate feature and put those features into machine learning algorithms and compute the performance of individual text and after that compute the similarity between each textual data, second method is kernel method. In this case, define kernel function and with respect to that kernel function compute the similarity between two textual data.

In this work there is one kernel related algorithm that is kernel support vector machine.

1.4.2 Improving Algorithms Performance

Ensemble model not only combine algorithms but also increases the performance of the algorithm. Suppose naïve Bayes algorithm is implemented on some dataset. Suppose accuracy for a particular dataset is very less, so accuracy can be increased by two methods. First one is to use another algorithm and the second one, combine this algorithm with other algorithms for improving the performance. In this work, second method is used for improving the performance of algorithms. This application is also very useful to increase the efficiency, the accuracy of the algorithms.

1.5 R Programming Language

R is a programming platform, free open source software, mainly used for statistics computation. Beside statistics application R also act as the programming language (same as C or JAVA). Code written in R is very much similar to python (Basic algebraic calculation) R released at 1993. It is a GNU project. R is rich in libraries. Number of package present in R is approx. 5000. So it avoids a great deal of line of code and saves time (as in C++, JAVA). Out of all the software present such as weka, java, C, R is considered to be the best platform for machine learning because of its simplicity. R also displays graphical computation very impressively as compared to Mathematica and Matlab. One important feature of R is that it is platform independent hence R can be used in any operating system. R can also integrate with language such as JAVA, C++ etc. Inside R, there are about 200 machine learning models present. R is also used for finding missing values. R is also used in business. Popular companies which use R, are Bing (Mainly used for increasing the awareness in social search), Google (making online add effectively), Facebook (Facebook status analysis, Prediction of friends or colleague interaction).

Table 1.1: Functions used in R

Functions	Explanation
Rpart	This is the function used for implementing decision tree
Ada	This is the function used for implementing Ada boost model.
Rf	This is the function used for implementing random forest. Package used for this model Randomforest
Ksvm	This is the function used for implementing

	support vector machine. Package used for this model Kernlab
Glm	This is the function used for implementing generalized linear model
Nnet	This is the function used for implementing neural network
Fda	This is the function used for implementing flexible discriminant analysis.
oblique.tree	This is the function used for implementing oblique tree
Evtree	This is the function used for implementing Tree model for genetic algorithm
Earth	This is the function used for implementing Multivariate adaptive regression spline
C5.0	This is the function used for implementing C5.0
J48	This is the function used for implementing J48.Package used for this model is Rweka
LMT	This is the function used for implementing Logical model tree.Package used for this model is Rweka
M5P	This is the function used for implementing M5P.Package used for this model is Rweka
Mda	This is the function used for implementing mixed discriminant analysis. Package used for this model is Mda
read.csv	This is used for reading csv file.
set.seed	This is used for setting the seed value
Predict	This is used for calculating the prediction value
write.csv	This is used for writing the output into csv file
Library	This is used for importing different library into R. for example library(ROCR)
confusionMatrix	This is used for calculating machine learning parameter such as sensitivity, specificity, kappa value
Setwd	This is used for setting the current working directory.
Setdiff	This is used for removing desired attribute from the dataset.

R is considered to be best programming language for machine learning as compared to Python, SQL, SAS, JAVA, MATLAB, C++, Mathematica. Poll is conducted by KDnugget in 2013 described in Figure 1.3. Although there are some other languages such as Julia which is much more efficient than R, but R is still considered as the best programming language for statical and graphical computation and best platform for implementing machine learning models or algorithms.

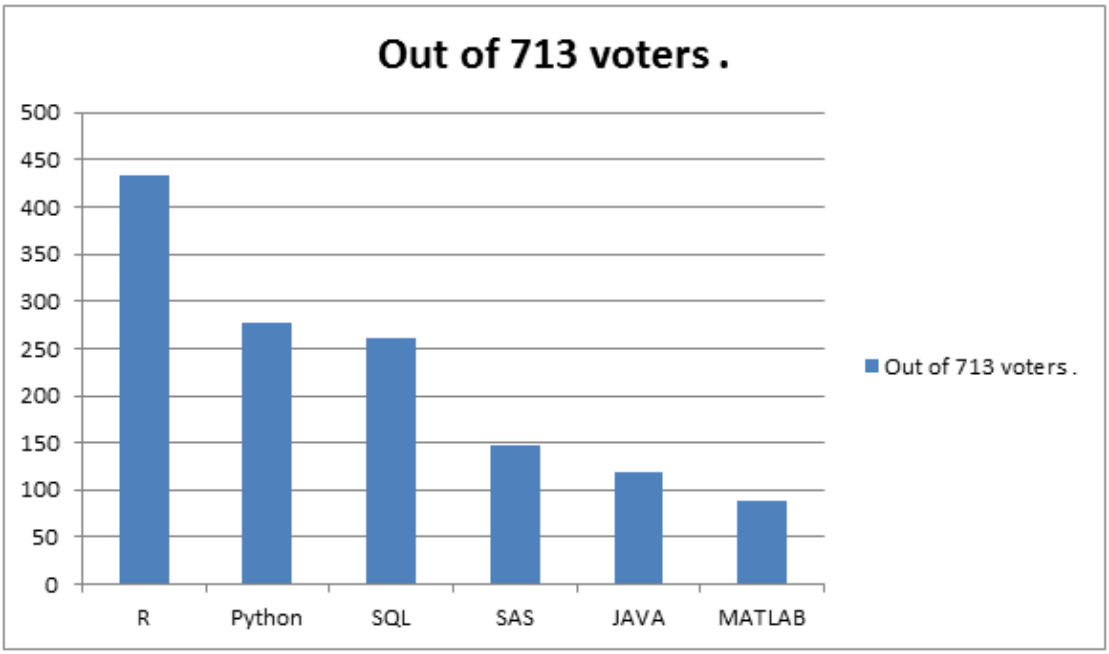

Figure 1.4: Popularity of R

Table 1.1 represents functions used in R for calculation. These functions are rpart, ada, rf, ksvm, glm, nnet, fda, oblique.tree, evtree, earth, c5.0, j48, lmt, m5p, mda, read.csv, set.seed, Predict, write.csv, Library, confusionMatrix, Setwd, Setdiff. The use of each function is described in Table 1.1. Table 1.2 describes how to use functions in R programming language

Table 1.2: Functions and its Formula

Functions	Function formulae
Ada	ada(class ~ .,data=crs$dataset[crs$train,c(crs$input, crs$target)],cp=0.01,minsplit=20, xval=10),iter=50)
M5P	M5P(class ~ .,data = training_data)
Rpart	rpart(class ~ .,data=crs$dataset[crs$train, c(crs$input, crs$target)])
J48	J48(as.factor(class) ~ .,data = training_data)
C5	C5.0(as.factor(class) ~ ., training_data)
Ksvm	ksvm(as.factor(class) ~ .data=crs$dataset[crs$train,c(crs$input, crs$target)])
GLM	glm(class ~ .,data=crs$dataset[crs$train, c(crs$input, crs$target)],family=binomial(link="logit"))
M5P	M5P(class ~ .,data = training_data)
Mda	mda(as.factor(class) ~ .,data = training_data)
Neural network	nnet(as.factor(class) ~

	.,data=crs$dataset[crs$sample,c(crs$input,)
Oblique.tree	oblique.tree(as.factor(class) ~ .,training_data,oblique.splits = "on")
read.csv("C:/User/Machine/ Desktop")	This is used for reading the csv file present in Desktop
set.seed(201210)	This is used for setting the seed value to 201210.
predict	predict(oblique_tree, testing_data,type="class")
write.csv(file="C:/User/Machine?Desk top")	This is used for writing the csv file into dektop.
Setdiff	setdiff(seq_len(nrow(my_data)),training_index)
ConfusionMatrix	confusionMatrix(table(my_data[testing_index,predic, dnn=c("Actual", "Predicted")))

1.6 Microsoft Excel

Microsoft excel is another platform used for this work. It is spreadsheet application. Excel has many features. Some of the features present in Excel are Pivot table, Basic math calculation (Sum, multiply, subtract, division, average etc.), Creation of Graph, Sorting etc. Excel is a very good analysis tool. For this work, excel is used because after calculating models performance in R next task is the categorization of the model on the basis of their accuracy.

Table 1.3 displays functions used in excel for the work. These functions are round, average and If. Explanation of each function is described in Table 1.3.

Table 1.3: Function used in Microsoft Excel

Excel Function	Explanation
Round	Used for round up the decimal values
Average	Used for calculating the average
If	Used for checking some specific conditions

1.6.1 If Function in Excel

'If' is condition statement present in Excel. Excel's "if" functions returns only one value either for the true condition or the false condition. It is basically used for checking the condition. The syntax for if condition in Microsoft Excel is

=If (condition, statement 1, statement 2)

- Here condition refers the condition that is to be tested.
- If condition is true statement 1 is executed.
- If condition is false statement 2 is executed.

Figure 1.5 displays how If function is used in the work. Figure 1.5 describes, if actual and predicted value are equal than output is 1 else 0 here actual value is actual target value present in the dataset and predicted value is predicted target value calculated by the individual model. In Figure 1.5 there is one another condition, which is used for checking decimal value. If predicted value i.e. model outputted target value, is in decimal then it is compared with 0.5. If predicted value is greater than 0.5 than, the predicted value is equals to 1 otherwise 0. The data is classification data so the categorization is done into 0 or 1.

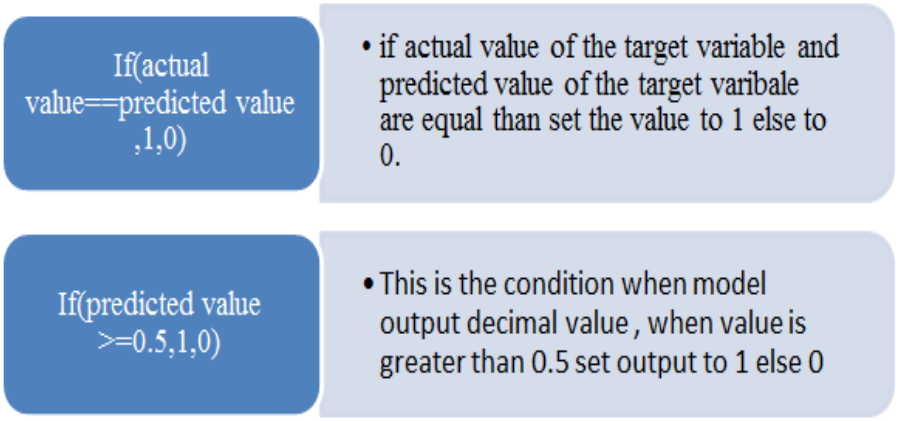

Figure 1.5: If Function in Excel

1.6.2 Average Function in Excel

Average is another function used in this work. Average is used for calculating the average of numbers. Average function returns the arithmetic of the argument passed into average function. Syntax for average is:

=Average (Number1: Number 2)

- Number 1 is the first number from where formula start.
- Number 2 is the range from where formula end.

Figure 1.6 displays how Average function is used in the work. Described in Figure 1.6.Average (R1: R4601) calculate the average of 4601 rows. Average is done after the categorization of data into 1 and 0 by using "If" function which is described in Figure 1.6. Here average is calculated in order to find the accuracy. So after

calculating average multiply the value unto 100. Hence, the proper syntax become. Average (R1: R4601) *100. There is another syntax discussed in Figure 1.6 i.e. Average (B2: C2).This syntax used when machine learning models are combined.

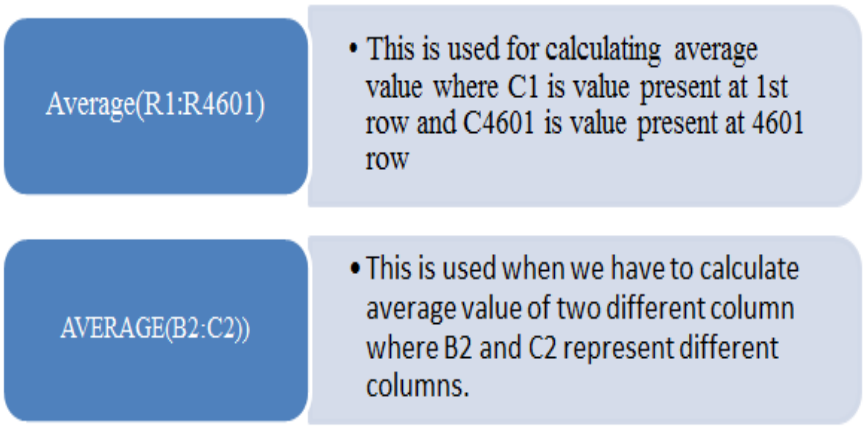

Figure 1.6: Average Function in Excel

1.6.3 Round Function in Excel

Round is another function used in this work. Round is used for rounding up the values with specific number of digit i.e. if digit has to round up to the specific decimal places than round function is used. Syntax for round function is:

=Round (number, Number of digit)

- Number is that decimal number which has to round up.
- Number of digit is that number which has to round. Number of digits is categorized into three parts.
  - (1) If number is greater than 0 :here number is rounded up to specific digit
  - (2) Number is equal to 0: Number rounded up to nearest number.
  - (3) Number is less than 0: Number is rounded from left side of decimal point.

Figure 1.7 displays how to use the round function. For this work round function is not used directly instead it is used with the average function. Here first calculate the average of two different columns and after that round up that value into the nearest integer. This function is used to find out the best ensemble model for the dataset.

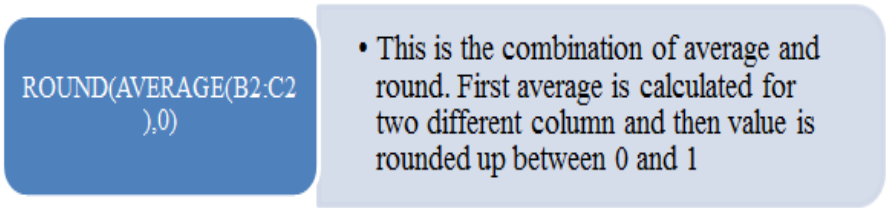

Figure 1.7: Round Function in Excel

1.7 Dataset and Features

Data set for spam is found at https://archive.ics.uci.edu/ml/datasets/Spambase. Number of instance present in the dataset is 4601.Number of attribute present in the dataset is 58 out of which 57 are continuous and one is used for categorization of the data set, i.e. weather it is spam or not where 1 is considered as spam and 0 is considered as not-spam. Characteristics of dataset is multivariate i.e. it include observation of more than one statistical outcome.

1.7.1 Feature Description

In spam dataset, majority of features represent whether a word frequently present in the email. F1 to F48 represent attribute type word_frequency_word. F49 to F54 represent attributes type char_frequency_char. F55 represent attribute of type capital_run_length_average, F56 attribute of type capital_run_length_longest, F57 attribute of type capital_run_length_total.F58 represent weather email is spam or not. In this data set there are 1813 spam (39.4%) and 2788 non-spam (60.6%) and the misclassification error is 6.91 %. Final dataset contain eight features i.e. F7, F16, F21, F25, F27, F46, F52 and F53 where F7 is frequency of the number of remove words, F16 is frequency of the number of free words, F21 is frequency of the number of time your is present, F25 is frequency of the word hp, F27 is frequency of the word george, F46 is frequency of the word edu, F52 is frequency of the character ! (Exclamation), F53 is frequency of the character @ .All the features explained in Table 1.4

Table 1.4: Descriptions of the Features

S.no	Features	Features detail
1	F7	word_frequency_remove
2	F16	word_frequency_free
3	F21	word_frequency_your
4	F25	word_frequency_hp
5	F27	word_frequency_george
6	F46	word_frequency_edu
7	F52	char_frequency_!
8	F53	char_frequency_@

1.8 Structure of the Thesis

The thesis is organized as follows:

Chapter 2: This chapter discuss the detail analysis of spam i.e. types of spam, effect of spam. This chapter also describes work done by the researchers in the field of spam filtering.

Chapter 3: This chapter discusses the problem definition along with research gaps and also describes the objectives of this research work.

Chapter 4: This chapter discusses the implementation details with results of the experiments from the different existing algorithms .This chapter also describes the graph and comparative analysis of the proposed techniques.

Chapter 5: This chapter discusses the conclusion, contribution of work done and the future scope of the work.

Chapter 2
Literature Review

This chapter discusses the types of email spam, analysis of spam, work done by different researchers in the field of spam filtering. This chapter also describes the analysis of R programming language.

2.1 Introduction

As the internet is growing the amount of spam is also increasing day by day. Spam is the main reason for financial loss on internet, steal valuable information (bank account detail, password etc.), reduces the network bandwidth etc. Email spam is categorized into various types. This categorization is described in Figure 2.1.

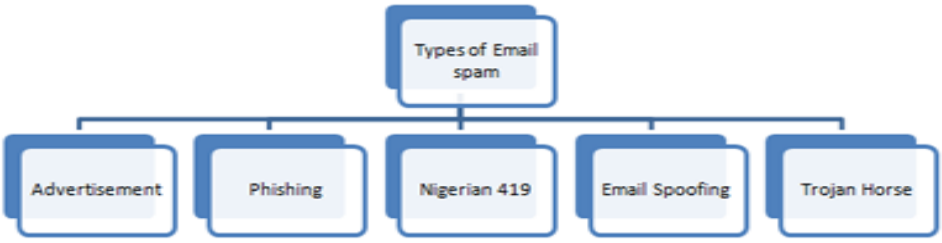

Figure 2.1: Types of Email Spam

A Survey was done by New York Technology industry. In that survey they mentioned, only email spam is the reason for losses of 40 billion (U.S dollars) all over the world. Popularity of spam is increasing so drastically that one study shows that about 95 percent of email received by a single company is only spam. According to Microsoft researcher's study and Google researcher' s study, every day 1.8%-3% of the 50 billion pieces of spams are sent out by anti-spam defenses and each spam takes 5 seconds to delete, and hence user's time is approximately $25 an hour. Email spam is increasing day by day so huge amount of complains are registered every day. Table 2.1 describes the FBI report regarding spam complain.

Table 2.1: Complain for Email Spam

City effected by spam	Complains
California	35,170
Florida	21,035
Texas	19,478
New York	16,056
Ohio	13,662

January 2013 EMC report, they calculated that the phishing attack in 2012 was 59% higher than 2011. The loss occurred due to phishing attack in 2012 was $ 1.5 billion which was 22% higher as compared to 2011 statistics. Various countries were affected by phishing attack. The list of some countries affected by phishing attack:

- United state
- United Kingdom
- Brazil
- Japan
- France
- Canada
- Russia
- Poland
- Netherlands

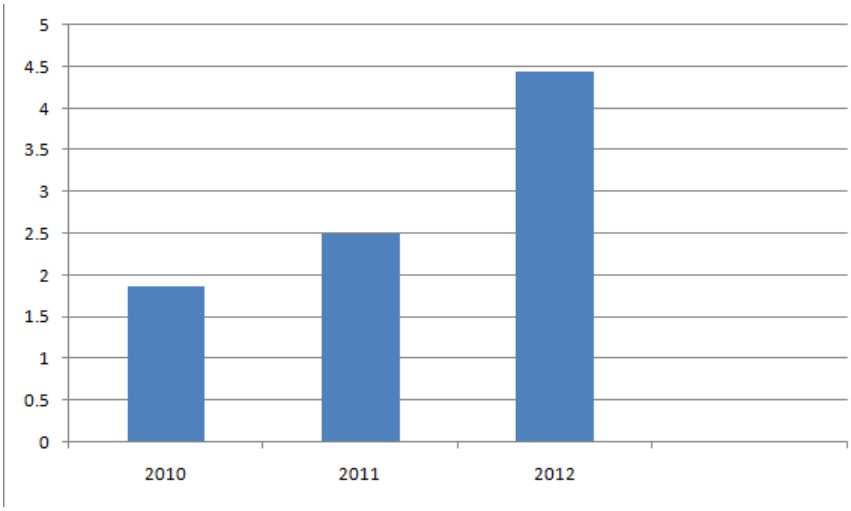

Figure 2.2: Phishing Attack

Phishing attack detected by RSA Anti-Fraud Command Centre (AFCC) on the yearly basis is described in Figure 2.2 (1 unit is equals to 100000). There is another type of spam, which is called internal spam. Internal spam is the spam created by the employee present inside the company. Survey was conducted by National Technology Readiness Survey in 2004 and according to them spam causes the economic loss of $21 billion every year, only in United States

2.2 Related Work

Problem which is done by spam is discussed by Anish and Geerthika on their work they used both the approaches i.e. machine learning filtering approach and non-machine learning filtering approach. In non-machine filtering approach, they

proposed heuristic based method, which is rule based approach; secondly they also proposed signature based method, which is generally based upon hash value. Blacklisting is another method proposed by them in non-machine learning approach, traffic analysis is last non-machine learning approach proposed by them. While in machine learning approach they proposed methods such as unified model filters, filter based on likeness, ensemble filter and complementary filter [4]. Dr.Prabahkar and Basvaraju also worked in the area of spam filtering. On their work they also proposed new spam detection technique called test clustering which is basically based upon vector space model. Vector space model is also used for data representation and clustering for data reduction [5]. P. Sudhakar, R. Kishore Kumar G. Poonkuzhali, used data mining technique for spam filtering in their work, "Comparative Study on Email Spam Classifier using Data Mining Techniques". In their work they used data mining tool called TANAGRA, to find out efficient machine learning model for the dataset [6]. Beside text spam, work was also done in the area of image spam. Vandana Jaswal, presented her work to detect stem word or repeated word present in the dataset and also utilized markov rule to detect image spam [7]. Naïve bayes is considered as one of the best spam filtering and spam detection algorithm. Asmeeta mali used this algorithm for spam detection on her work, "Spam Detection using Bayesian with Pattern Discovery". In this work she proposed an effective technique for spam filtering using Bayesian classification [8]. Islam and Xiang introduced the new filtering technique with the help of instance selection method using WEKA [9]. Besides spam filtering various works had been done in the field of biology. Neha singh had worked on," Dendritic Cell algorithm and Dempster Belief Theory Using Improved Intrusion Detection System" [10], while Greensmith had worked on Dendritic Cell algorithm [11].

2.3 R Programming Language

R was formed in 1993. It is open source software used for statistical and graphical computation. R is available in many different interfaces. These interfaces are:

- R studio: It is an integrated development environment. For running R studio first install R. Latest version of R is 3.2.1. This is stable version which was released on 18 June 2015.
- Rattle: It is also a graphical user interface. It is basically used for data mining (feature selection. Feature extraction etc.)

- R commander: It is also a graphical user interface. Stable version of R commander is 2.0.0, released on 21 August 2013.
- RKWard: This is the extended graphical user interface and integrated development environment for R. It is basically written in C++ and ECMA script. Stable version of this software is 0.6.3, released on 7 march 2015

Chapter 3
Research Problem

Previous chapter discusses work done by the researchers in the field of spam filtering. This chapter has been focused on the problem definition along with research gaps and describes the objectives of this research work.

3.1 Problem Statement

The use of internet is increasing exponentially day by day. So, the amount of spam data is also increasing day by day. There are many algorithms proposed for spam filter such as multinomial naive bayes classifier, support vector machine, and genetic algorithm etc. The use of internet data is increasing exponentially, so the individual algorithm would not be efficient. So, to overcome this problem one algorithm can be combined with another algorithm, thus it can work works more efficiently as compared to individual one. The combination of the different algorithms into the single algorithm is called ensemble model.

Ensemble modeling gains popularity because many organizations deployed computer software to run such a model. Some of the software are SAS predictive analysis, IBM predictive analysis etc. Thus, the use of ensemble the model plays an important role today.

Due to increasing amount of spam data in internet, efficient spam filtering approach has always been the most researched area. But with increasing amount of data, format of data is also varying for example data may be present in image format, text format or video format. So it is always a challenging task to propose such a model which works efficiently for all type of data format. R programming language is one of the important platforms for testing the machine learning models for the given dataset. Machine learning is the statistics approach so the use of Microsoft Excel plays an important role. Only ensembling the model is not enough for a dataset, feature selection also plays an important role while ensemble the model.

3.2 Research Gaps

As many algorithms are proposed for spam filtering. Different algorithm works efficiently for the different dataset and different data format (text, image or video). For proposing the algorithm that suited best for all data format ensemble model plays an important role. Although many data types are the combination of two or more

different data format i.e. text and image or text and video or text, image and video. For that type of dataset, ensemble model plays an important role.

3.3 Objectives

The objectives used for this work are discussed below:

- Analyze the various machine learning models on spam dataset for different partitions, seed value, features etc.
- To perform the comparative analysis for different machine leaning models on the basis of accuracy, sensitivity, specificity, cohen coefficient (Kappa value) and confusion matrix.
- To propose the ensemble model for the dataset on the basis of performance metrics.

3.4 Research Methodology

Machine learning model is implemented on R programming language where dataset is spam dataset other platform used for this purpose is Microsoft Excel .Research methodology as follow:

- Collect the spam dataset from UCI machine repository.
- Model Implementation on RAW data i.e. no feature selection. Number of features is fifty eight.
- Feature selection using GINI index. GINI index is a method available in random forest , mainly used for best feature selection
- Compare the performance of machine learning models between RAW data and feature selection data.
- Compare the performance of data for different partition of training and testing data.
- Analysis of Result on the basis of accuracy, ROC, confusion matrix, k-value, sensitivity, specificity.
- After analyzing the result, ensemble top five models and draw the graph for best ensemble model for the dataset.

Chapter 4
Implementation and Results

This chapter introduces implemented models on spam dataset, optimization of the dataset which includes optimizing spam dataset into important features. This chapter also describes validating the dataset. This chapter also describes comparison with respect to different partitions for Accuracy and ROC values. Model comparison and ensemble models are also described in this chapter. This chapter also describes K-fold validation for top five models.

4.1. Implementation Environment

Data used for this work is obtained from UCI machine repository. It is spam data set. Environment used for this work is R programming language and Microsoft excel. R is used for implementing different machine learning models on the spam dataset and Microsoft excel is used for comparing the performance of different models. The machine learning models are developed in R. The specifications of the machine which is used for implementation purpose are – Intel (R) Core (TM) i5-3210M CPU @ 2.50GHz with 4 GB RAM and 32-bit Windows operating system.

4.2. Proposed Approach

The approach is described in Fig 4.1. In the first phase, spam data set is collected from the UCI machine repository. In the second phase, implementation of five machine learning models (Decision Tree, random forest, KSVM, neural network and Ada boost model) is done on RAW data. After implementing these models, performance metrics have been evaluated. Feature selection is the third step of this approach. Feature selection means removing irrelevant features (done by GINI index) such that only important features are present in the dataset. Feature selection is basically used for working the machine learning models more efficiently. Feature selection is done with the help of performance metrics. In the fourth phase, comparison is done between the, performance of feature selection data and raw data. In the fifth phase, implementation of fifteen machine learning models has to be performed on the data set (after feature selection). In the sixth phase, evaluations of the models have been performed w.r.t. performance metrics. In the seventh phase, k-fold validation have been performed for ensembling the models

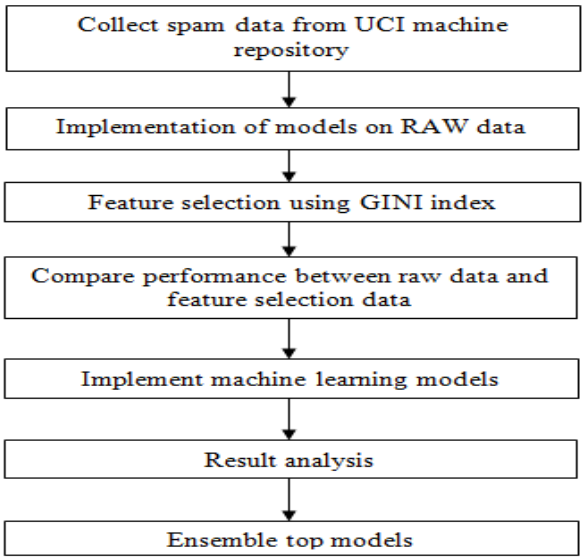

Figure 4.1: Used Methodology

Figure 4.2 shows the flow chart of Consensus Based Ensemble Model for Spam detection. Initially, data is collected from the UCI machine repository. After that models are implemeted for two cases first one is when features are selected and second one is when features are not selected. After that the performance has been compared for these two different cases. If performances are equal than data optimaztion is possible otherwise data optimization is not possible. After data optimization, dataset should be validate w.r.t models. Validation is basically done to validate the dataset. If dataset is not validate than this means approach is wrong. If dataset is validate then the performance is compared for ROC and accuracy on different partitions. If performance are equal for accuracy, ROC at different partitions that means approach is right otheriwse approach is wrong. After that the models have been implemented on the spam dataset and performance of the models has been compared on the basis of accuracy, ROC, kapp value and confusion matrix. After model comparison, top five models for the dataset have been selected. After that k-fold validation is performed on these top five selected models and create the graph. If graph displays straight line for every models this means that models are robust for the datset and if the models are robust than and only models can be ensemble. Hence ensembling the model, which is last step of this work.

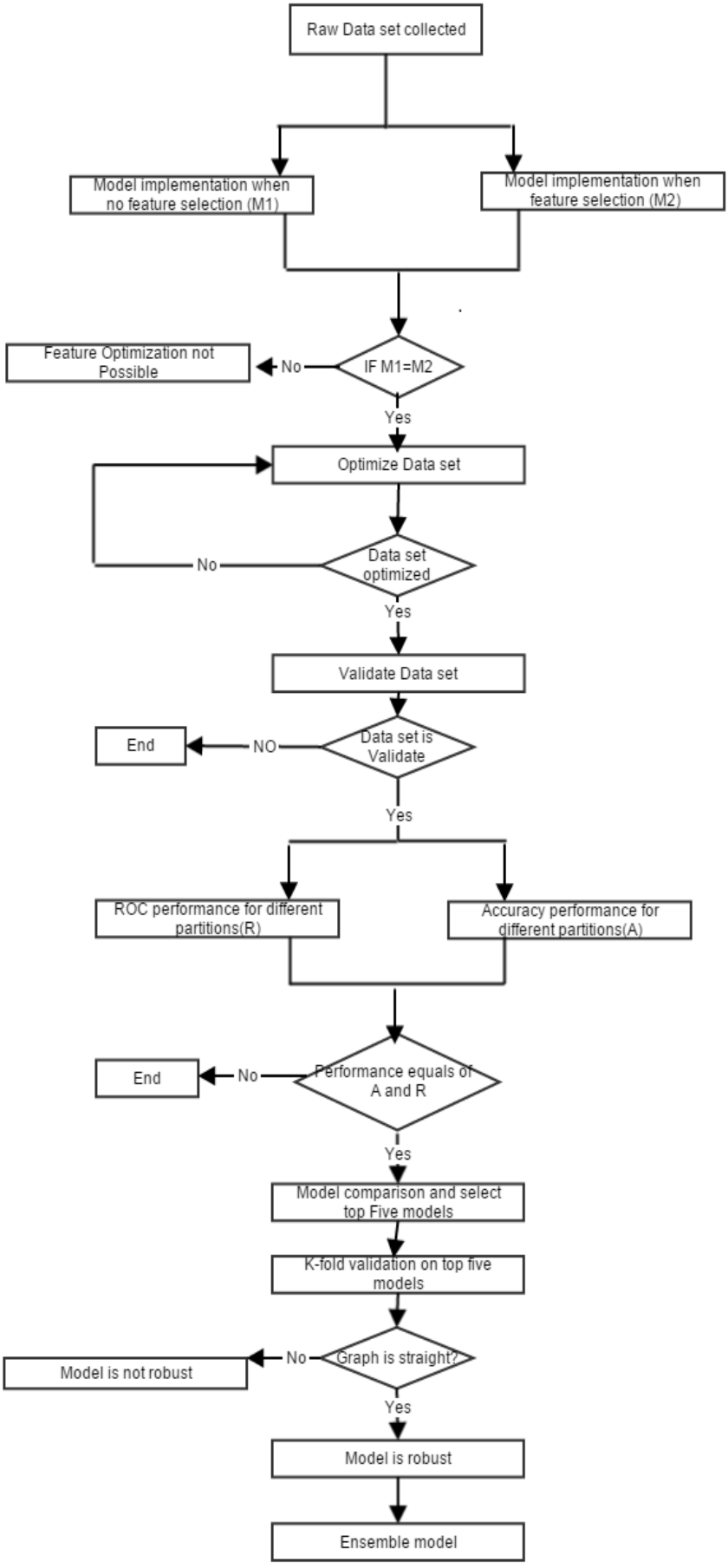

Figure 4.2: Flow Chart

Eq.1, eq. 2, eq. 3 and eq. 4 discuss the formula used for ensemble the model. First, the average is calculated for those models to which ensembling are performed. This work is based upon classification problem so output value (here it is X) should be either 0 or 1. This can be done by round function. After that actual value of the target variable is compared with output value (here it is X). If both the values are equals then the predicted value of the ensemble model will be 1 otherwise 0. After that accuracy is calculated for the ensemble model, which is described in following equations.

$$X_i = \sum_{1}^{n} (model_i + model_{i+1}) \div k \qquad \text{Eq. 1}$$

$$\text{X} = \begin{matrix} \{If X > 0.5; X = 1\} \\ \{If X <= 0.5; X = 0\} \end{matrix} \qquad \text{Eq. 2}$$

$$\begin{matrix} If(A_i == X_i) \\ then P_i = 1 \\ else P_i = 0 \end{matrix} \qquad \text{Eq. 3}$$

$$Acuuracy = (\sum_{1}^{\kappa} (p_i + p_{i+1})/k) \times 100 \qquad \text{Eq. 4}$$

X_i = Average value after ensemble the models

K= Number of instances present in the dataset.

N= Number of models.

P_i = Predicted target variable value

A_i = Actual target variable value

4.3. Implemented Algorithms

After defining the problem statement, the next step is implementation of machine learning algorithms for solving the problem. There are various algorithms available in machine learning algorithm. These algorithms are Decision tree, Boosted classification tree, Random forest, kernel support vector machine, Generalized linear model, neural network, Flexible discriminant analysis, Oblique tree, Tree model for genetic algorithm, Multivariate adaptive regression spline, C5.0, C4.5 like tree, Logistic model tree, M5P, Mixture discriminant analysis (as shown in table 4.1).

- Decision Trees (rpart): This method is an extension of C4.5 classification algorithms described by Quinlan [12].

- AdaBoost: It is very popular and perhaps the most significant historically. It was the first algorithm that could adapt to the weak learners [13].
- Random Forest (randomForest): It is based on a forest of trees using random inputs [14].
- Support Vector Machine (SVM): SVMs yet represent a powerful technique for general (nonlinear) classification, regression and outliers' detection with an intuitive model representation [15].
- Linear Models (GLM): For regression it uses linear model, analysis of covariance and single analysis of variance [16].
- Neural Network (NN): Training of neural networks using back-propagation, resilient back propagation with or without weight or the modified globally convergent version [17].
- Flexible Discriminant analysis: It is generally used for multigroup classification [18].
- Oblique tree: It is a part of CART framework. It is used when node is being split for discovering oblique hyper plane[19]
- Tree model for genetic algorithm: It is used for optimal classification [20].
- Multivariate adaptive regression spline: It is an extension of generalize linear model. Here it includes the technique used by Freidman in Multivariate Adaptive Regression Splines [21] and in fast MARS [22].
- C 5.0: It is the improvement of C 4.5 algorithm.
- C4.5 algorithm: It is used to generate the classification and regression tree (CART).It is also used for generating decision tree [23].
- Logistic model tree: It is a combination of two different models i.e. decision tree learning and logistic regression [24][25].
- M5P: It is a tree learners classifier [26]
- Mixed discriminant analysis: It is an extended version of linear discriminant analysis. It is generally based on combination of models

Table 4.1: Implemented Models on Spam Dataset

SN	Model	Method	Package
1	Decision Tree (rpart)	Rpart	Rpart
2	Boosted classification tree	Ada	ada,plyr
3	Random Forest	Rf	Randomforest
4	KSVM	Ksvm	Kernlab
5	Generalized linear model	Glm	None
6	Neural network	Nnet	Nnet

7	Flexible Discriminant analysis	Fda	earth,mda
8	Oblique tree	oblique.tree	oblique.tree
9	Genetic algorithm	Evtree	Evtree
10	Adaptive regression spline	Earth	Earth
11	C5.0	C5.0	C5.0
12	C4.5 like tree	J48	RWeka
13	Logistic Model Tree	LMT	RWeka
14	M5P	M5P	RWeka
15	Mixture Discriminant Analysis	Mda	Mda

4.4. Implementation on Raw Data

As discussed above raw data contain fifty-eight features. So at first, implementation of machine learning models is performed on raw data for different seed values. These seed value are 99378, 201210 and 849287. The comparison is done on the basis of accuracy, Confusion matrix, ROC curve and risk. The comparison is performed on two different scenarios first is without feature selection and the second is with feature selection. Implemented models on raw data are decision tree, ADA boost model, random forest, kernel support vector machine, generalized linear model and neural network.

Accuracy

Accuracy is basically used for comparing the performance of the models or algorithm. The accuracy is calculated between the actual value and the predicted value. The Predicted value is predicted output of the respective model. Accuracy is calculated using following equation.

$$If(Actualvlaues == Predictedvalues) \\ then \\ Accuracy = 1 \\ else \\ Accuracy = 0 \qquad \text{Eq. 5}$$

ROC Curve

ROC (receiver operating characteristics) curve is also used for determining the performance of models. ROC graph is constructed between true a positive rate, which is also called sensitivity and false positive rate, which is also called specificity. The area under the ROC curve is called area under curve (AUC), ROC curve performance is determined by area under the curve (AUC).If the value of AUC is 1 than the model is assumed to be perfect for the dataset. Distribution of AUC is as follows:

- If AUC is between 0.9 to 1.0 than the test is excellent.

- If AUC is between 0.8 to 0.9 than the test is good.
- If AUC is between 0.7 to 0.8 than the test is fair.
- If AUC is between 0.6 to 0.7 than the test is poor.
- If AUC is between 0.5 to 0.6 than the test is failed.

In this work, there is an important role of ROC for determining the performance of the models. Performance is determined by the Area under Curve (AUC) which is one of the properties of ROC.

Confusion Matrix

The confusion matrix is also called error matrix. It is also used for determining the performance of the model (model can be classification or regression). Lower the value of confusion matrix better will be the model. For classification problem, it is confusion matrix and for regression problem, it is an error matrix. The confusion matrix is 2 x 2 matrixes categorized by false positive, true positive, true negative and false negative. A confusion matrix is a matrix that can be divided into true positive rate, false positive rate, false negative rate, true negative rate described in Figure 4.2. Now consider the situation for prediction of rainfall

- True positive rate : It is the scenario when the prediction is true i.e. prediction of rainfall is true
- False positive rate: It is the scenario when the prediction is true, but outcome is false i.e. prediction is rainfall will occur, but actually it's not.
- False negative rate: In this scenario prediction is false as compared to the real event. i.e. prediction is rainfall will not occur, but it occurs.
- True negative rate: In this scenario prediction is that rainfall will not occur and it will not occur.

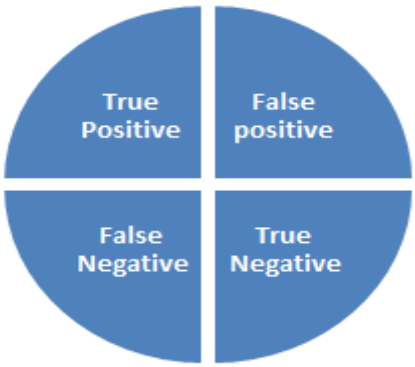

Figure 4.3: Confusion Matrix

Cohen Kappa Coefficient (k-value):

It is also used for determining the performance of the model .The value of K is always less than or equal to 1. Distribution of k-value is defined by Landis and Koch [27]. Distribution of k value is as follow:

- If k value is between 0.8 to 1.0 than it is very good agreement
- If k value is between 0.6 to 0.8 than it is a good agreement
- If k value is between 0.4 to 0.6 than it is moderate agreement
- If k value is between 0.2 to 0.4 than it is fair agreement
- If k value is less than 0.2 than it is poor agreement

4.4.1 Seed Value: 99378

a) No Feature Selection

This is the first scenario when seed value is 99378 and the numbers of features used in this case are 58. As the accuracy is above 90 percent that means these models are very good for the dataset, ROC curve is also above 90 and according to AUC distribution this is also good distribution. Less the value of confusion matrix better will be the model.

Table 4.2: Model Comparison when no Feature Selection and Seed Value 99378

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	90.59	0.09	0.96	0.94
2	ADA boost model	95.51	0.044	0.992	0.98
3	Random forest	95.00	0.049	0.991	0.98
4	Kernel support vector machine	99.261	0.073	0.981	0.97
5	Generalized linear model	90.58	0.094	0.931	0.90
6	Neural network	85.22	0.147	0.94	0.91

b) Feature Selection

This is the second scenario after feature selection when the seed value is 99378. Number of features used, in this case, is 10.As the accuracy is above 90 percent that means these models are very good for the dataset, ROC curve is also between 1 and .9 and according to AUC distribution this is also good distribution.

Table 4.3: Model Comparison when Feature Selection and Seed Value 99378

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	90.58	0.09	0.96	0.94
2	ADA boost model	94.13	0.058	0.99	0.98
3	Random forest	94.56	0.054	0.99	0.98
4	KSVM	91.74	0.082	0.97	0.96
5	GLM	91.09	0.089	0.97	0.96
6	Neural network	91.23	0.087	0.97	0.96

4.4.2 Seed Value: 201210

a) No Feature Selection

This is the first scenario when seed value is 201210 and no feature selection. The number of feature present here is 58. ROC curve and accuracy is above 90 percent that means for seed value 201210 these six models are good for the dataset. The most values of ROC are close to 1 that means features can be reduced. In Table 4.4 after analyzing accuracy, Confusion matrix, ROC curve (when no feature selected i.e. fifty-eight), and seed value is 201210 all the models perform very efficiently except decision tree, while considering ROC value, accuracy and confusion matrix value least efficient model for this dataset is decision tree. Decision tree has the least accuracy as compared to other models (88.99). The value of confusion matrix for decision tree is also high as compared to another model (.11). Similarly, ROC value is very less as all the models has ROC values close to 1 but the ROC value of decision tree is .90. Best models for the dataset is ADA boosts model and random forest after considering the values of accuracy, confusion matrix, and ROC curve. ROC curve value for ADA boost model and random forest models are 0.98, which is very close to 1. Maximum accuracy for the random forest which is 95.07

Table 4.4: Model Comparison when no Feature Selection and Seed Value 201210

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.99	0.11	0.92	0.90
2	ADA boost model	95.01	0.049	0.99	0.98
3	Random forest	95.07	0.049	0.99	0.98
4	KSVM	92.25	0.077	0.97	0.96
5	GLM	92.25	0.077	0.97	0.97
6	Neural network	92.32	0.076	0.97	0.97

b) Feature Selection

This is the second scenario after selection of important features and seed value is 201210. Performance is measured on the basis of accuracy, confusion matrix, and roc curve. The best model for this dataset after feature selection is ADA boost model and random forest. Considering the confusion matrix, the confusion matrix value for decision tree is 0.11, which is a bit high as compared to other models for seed value 201210 and number of features eight. Similarly for ADA boost model confusion matrix is 0.05 and it is very less that means, ADA boost model is best for the spam dataset, similarly random forest is also a good model for this dataset as confusion matrix is very less, which is 0.051, similarly for kernel support vector machine here value of confusion matrix is 0.084 that means it is also good for the dataset. Similarly, for generalized linear model and neural network both have confusion matrix value very less which means these both are good models for the dataset.

Table 4.5: Model Comparison when Feature Selection and Seed Value 201210

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.99	0.11	0.92	0.90
2	ADA boost model	94.27	0.05	0.98	0.98
3	Random forest	94.85	0.051	0.99	0.98
4	KSVM	91.52	0.084	0.97	0.96
5	GLM	91.23	0.087	0.97	0.96
6	Neural network	88.99	0.11	0.96	0.95

4.4.2 Seed Value: 849287

a) No Feature Selection

This is the first scenario when no feature is selected i.e. data is raw and the seed value is 849287. Accuracy for the models is greater than 90 that mean, these machine learning algorithms are good for spam dataset. While considering the value of confusion matrix, it is 0.12 for decision tree, which is a little bit high so that means, for seed value 849287, decision tree is not so efficient for spam dataset. Similarly for ADA boost model value of confusion matrix is 0.057 and it is very less so that means, ADA boost model is best model for the dataset, similarly random forest is also a good model for the dataset as confusion matrix is very less, which is 0.055, similarly for

kernel support vector machine here value of confusion matrix is 0.071 that means it is also good for the dataset. Similarly, for generalized linear model and neural network both have confusion matrix value very less that means; these both are good models for the dataset. While considering accuracy, the least accurate model for spam dataset is decision tree, which is 89.72 and most accurate model is kernel support vector machine, which is 94.83. So for seed value 849827 poor models for spam dataset is decision tree with accuracy 89.72, confusion matrix value 0.12, and ROC value 0.89.

Table 4.6: Model Comparison when no Feature Selection and Seed Value 849287

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	89.72	0.12	0.923	0.89
2	ADA boost model	94.28	0.057	0.986	0.98
3	Random forest	94.49	0.055	0.985	0.98
4	KSVM	94.83	0.071	0.974	0.96
5	GLM	92.61	0.073	0.981	0.97
6	Neural network	92.1	0.078	0.982	0.97

b) Feature Selection

This is the second scenario when the important feature is selected and the seed value is 849287. After feature selection, best models or algorithms for spam dataset is ADA boost model and random forest having AUC 0.97. For decision tree ROC value is 0.89 and according to AUC distribution decision tree is not efficient for spam dataset when seed value is 849287 and numbers of features are eight. After testing with different seed values, the conclusion is that, the decision tree is not a good model for this dataset. Consider Table 4.7 where accuracy is 89.71 for ADA boost model that is less than all six models. ROC value for ADA boost model is 0.89, which is less as compared to other models. Next model, which is inefficient for spam dataset for seed value 849827, is the neural network where accuracy is 89.35 and ROC value is 0.88. The ROC curve for ADA boost model and random forest is 0.97 that means these two models are the best models for the dataset, similarly ROC value for kernel support vector machine and generalized linear model ROC value is same i.e. 0.96 that means these two models are also good model for the dataset. While considering confusion matrix value maximum value for confusion matrix is .1 for decision tree and 0.106 for the neural network and higher the value of confusion matrix poor will be the model. For ADA boost model value of confusion matrix is 0.06 that means this model is

good for spam dataset. Confusion matrix value for the random forest is 0.055, which is also very less that means this model is also good for spam dataset. Similarly for kernel support vector machine and generalized linear model the value of confusion matrix is 0.073 and 0.09 that means these are also good model for spam dataset.

Table 4.7: Model comparison when no Feature Selection and Seed Value 849287

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	89.71	0.10	0.923	0.89
2	ADA boost model	93.92	0.06	0.984	0.97
3	Random forest	94.5	0.055	0.984	0.97
4	KSVM	92.68	0.073	0.971	0.96
5	GLM	90.94	0.09	0.975	0.96
6	Neural network	89.35	0.106	0.91	0.88

Implementation of machine learning models for feature selection and no feature selection, on spam dataset, for different seed values is already done. After using three different seed values, conclusion is that models that are inefficient for spam dataset is ADA boost model and neural network and efficient models for spam dataset are Random forest, kernel support vector machine, generalized linear model and ADA boost model. Section 4.5 discusses optimization of the dataset. Initially, there are fifty-seven features, so feature optimization done in such a way that unwanted feature should be removed in order to increase the efficiency of models.

4.5. Dataset Optimization

After implementing six machine learning algorithm on spam dataset with different seed value(with no feature selection and feature selection) the conclusion is that after performing feature selection, performance remain very close to raw data (i.e. when there is no feature selection). So feature optimization should be done in such a way that the performance of models remains good on the basis of accuracy, confusion matrix, ROC curve etc. Initially, feature selection starts when the number of features present on the dataset is ten and after that optimization is done on the basis of accuracy, ROC value, and confusion matrix. Feature selection is performed by GINI index. It is a property present inside the random forest. In the GINI index, random forest assign some integer value to every feature which is also called Mean Decrease accuracy after assigning the integer values remove those features which have less integer value (Mean Decrease Accuracy).

4.5.1 Top Ten Features

After performing six machine learning models on the spam dataset when features are ten, value of ROC curve is close to one for most of the models and the accuracy is above 90 percent that mean there are chances for optimizing the dataset by removing some features. In this case, top ten features are F52, F53, F7, F25, F55, F46, F16, F56, F21, and F27. Out of these ten features least important feature according to GINI index is F56. So for the next case remove F56 and then compare the performance on the basis of the confusion matrix, accuracy, ROC curve.

Table 4.8: Performance Comparison for Ten Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.99	0.11	0.92	0.90
2	ADA boost model	93.41	0.065	0.98	0.98
3	Random forest	93.41	0.065	0.98	0.97
4	KSVM	90.44	0.095	0.96	0.95
5	GLM	90.58	0.094	0.97	0.96
6	Neural network	91.52	0.084	0.97	0.97

4.5.2. Top Nine Features

After removing the F56 feature, nine features are present in the dataset. These nine features are F52 ,F25 ,F7 ,F53 ,F46 ,F16 ,F21 ,F27 and F55.On these remaining nine features, ROC curve is close to 1 and accuracy is above 90 that means machine learning models work efficiently for nine feature also. So next task is removing one more feature. Out of these nine features least important feature according to GINI index is F55.So next task is to remove F55 feature from the dataset and run machine learning models on eight features and compare the performance on the basis of accuracy, ROC curve, confusion matrix etc.

Table 4.9: Performance Comparison for Nine Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.99	0.11	0.92	0.90
2	ADA boost model	92.68	0.073	0.98	0.97
3	Random forest	93.12	0.068	0.97	0.96
4	KSVM	89.93	0.1	0.96	0.95

5	GLM	90.51	0.094	0.96	0.95
6	Neural network	91.45	0.085	0.97	0.96

4.5.3. Top Eight Features

After removing the F55 feature, eight features are present in the dataset. These features are F52, F25, F7, F53, F46, F16, F21, and F27. For these eight features, ROC curve is very close to 1 and accuracy is also above 90. Here the value of confusion matrix is higher for two models i.e. Decision tree and generalized linear model and accuracy for these two models is also below 90 that means when the number of features is eight, two models i.e. Decision tree and generalized linear model are not performing efficiently. Out of these eight features least important feature is F21.So next task is to remove F21 and after that measure the performance of the models

Table 4.10: Performance Comparison for Eight Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.77	0.112	0.90	0.89
2	ADA boost model	91.6	0.083	0.97	0.98
3	Random forest	91.45	0.086	0.95	0.98
4	KSVM	90.22	0.097	0.94	0.96
5	GLM	89.34	0.1	0.95	0.97
6	Neural network	91.16	0.088	0.96	0.97

4.5.4. Top Seven Features

After finding least important feature, i.e. F21 with the help of GINI index remaining features present in spam dataset are F53, F7, F25, F52, F46, F16, and F27. In this case the value of ROC is again close to 1 for ADA boost model, generalized linear model, and neural network. The accuracy of ADA boost model, random forest, and the neural network is above 90. Values of confusion matrix is high for decision tree, kernel support vector machine, and generalized linear model, So with the help of Accuracy, confusion matrix and ROC curve conclusion is that worse model for spam dataset is decision tree. So next task is to remove one more feature and compare the performance of all six models. In this case least important feature is F27.

Table 4.11: Performance Comparison for Seven Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.77	0.112	0.92	0.90
2	ADA boost model	91.38	0.086	0.97	0.96
3	Random forest	91.38	0.086	0.96	0.94
4	KSVM	89.86	0.101	0.95	0.94
5	GLM	89.35	0.106	0.96	0.95
6	Neural network	91.01	0.089	0.96	0.95

4.5.5. Top Six Features

With the help of GINI index least important feature is F27. So after removing F27 remaining features present in spam dataset are F16, F53, F7, F25, F52, and F46. Here the value of ROC is close to 1 for the neural network and ADA boost model. The best model for spam dataset on the basis of accuracy is ADA boost model and random forest (as accuracy is greater than 90). The confusion matrix is lower for ADA boost model and random forest that means these two models are good for the dataset. So next task is to remove one more feature and compare the performance on the basis of accuracy, confusion matrix and ROC curve. With the help of GINI index least important feature is F16.

Table 4.12: Performance Comparison for Six Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.77	0.11	0.92	0.90
2	ADA boost model	90.65	0.093	0.96	0.95
3	Random forest	90.65	0.093	0.95	0.94
4	KSVM	88.77	0.112	0.95	0.93
5	GLM	86.69	0.123	0.95	0.94
6	Neural network	89.86	0.101	0.96	0.95

4.5.6. Top Five Features

After removing F16 with the help of GINI index remaining features present in the dataset are F53, F7, F25, F52, and F46. The value of ROC curve is above .9, but not close to one and accuracy is also less than 90 that means this dataset, which contain five features, is not feasible for the model. So next task is to remove some more

feature and check whether the less featured dataset is feasible or not. So with the help of GINI index less important feature is F52. Next task is to eliminate the feature F52. With this experiment, one conclusion is that F16 is an important feature so, by removing this feature models become less efficient.

Table 4.13: Performance Comparison for Five Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.99	0.12	0.92	0.90
2	ADA boost model	89.5	0.086	0.96	0.94
3	Random forest	89.35	0.076	0.95	0.93
4	KSVM	89.28	0.089	0.94	0.92
5	GLM	84.47	0.09	0.95	0.93
6	Neural network	89.71	0.088	0.95	0.94

4.5.7. Top Four Features

After removing F52 remaining features present in the dataset are F25, F7, F53, and F46. After observing Table 4.14 one prediction is that, the value of ROC curve is going down and the accuracy of all the models becomes less and the value of confusion matrix is quite high that means F16 is important feature in the spam dataset. So next task is to remove some more feature and check whether the performance of the model gets affected or not.

Table 4.14: Performance Comparison for Four Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	87.11	0.128	0.92	0.89
2	ADA boost model	87.83	0.121	0.93	0.91
3	Random forest	88.05	0.119	0.88	0.85
4	KSVM	86.96	0.13	0.91	0.89
5	GLM	86.38	0.136	0.93	0.91
6	Neural network	85.73	0.142	0.93	0.91

4.5.8. Top Three Features

With the help of GINI index least important feature present in the dataset is F46. Next task is to remove feature F46. So after removing F46, remaining features present in the dataset are F53, F27, and F7. After removing feature F46 value of ROC is going down. The value of Confusion matrix is also high (higher the confusion matrix less

the model will be efficient). It is also observable that accuracy of all the models is less than 90 that means model becomes less accurate after reducing the features in the dataset. So this also means F46 is also an important feature. So next task is to remove one more feature and compare the performance of the models.

Table 4.15: Performance comparison for three features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	87.03	.129	.92	.89
2	ADA boost model	87.40	.125	.92	.90
3	Random forest	87.40	.125	.91	.89
4	KSVM	86.82	.131	.91	.88
5	GLM	85.66	.143	.92	.89
6	Neural network	87.11	.128	.92	.89

4.5.9. Top Two Features

With the help of GINI index next least important feature is F53. So after removing F53 remaining features present in the dataset are F7 and F25. It is described in Table 5.15 accuracy is very poor for the dataset that means for two features these six models are not efficient. The value of ROC is also very less. The value of confusion matrix is also very high and higher the value of confusion matrix the model will be less efficient. So the conclusion is that F53 is also very important features for the dataset. Out of the fifty-seven optimization done up to two features and these two features are not feasible for the dataset. So, next task is to select best features which are best suited for the models.

Table 4.16: Performance Comparison for Two Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	76.9	0.23	0.77	0.70
2	ADA boost model	77.1	0.22	0.85	0.80
3	Random forest	77.1	0.22	0.85	0.80
4	Kernel support vector machine	77.1	0.22	0.76	0.69
5	Generalized linear model	76.1	0.23	0.85	0.80
6	Neural network	77.04	0.23	0.85	0.80

As from the above nine experiments, comparison of the performance of the models is done by removing features one by one. After observing Table 4.15, Table 4.14 and Table 4.13. F53, F46, and F16 are important features because after removing these features performance of models goes down. Next task is to select the dataset with best optimized features. If comparison is done between Table 4.8 (ten features) and Table 4.9 (nine features), there is not so much difference in performance (Accuracy, ROC curve, confusion matrix) that means after removing one features from table 4.1 the performance remain unchanged. Now if comparison is done between table 4.9 (nine features) and table 4.10 (eight features) AUC of decision tree, ADA boost model, kernel support vector machine and neural network is almost the same but accuracy is little bit different. Now if comparison is done between table 4.9 (nine features), table 4.10 (eight features) and table 4.11 (seven features) than there is vast difference in accuracy, ROC and confusion matrix so the best feature for the dataset is when number of features present in dataset is eight.

4.6. Validating Dataset

Validation is another method which is used to validate the dataset w.r.t. particular models. If validation data set and test data set values are equal than this means proposed approach is right. From Table 4.8 to Table 4.16 dataset is divided into two categories i.e. training dataset and testing dataset and for validating the dataset initially dataset was partitioned into three categories. First one is training dataset (Figure 4.4 line number 9), the second one is validation dataset (Figure 4.4 line number 10) and the third one is testing dataset (Figure 4.4 line number 11). Figure 4.4 describes the categorization of the dataset into training dataset, Validation dataset, and the testing dataset.

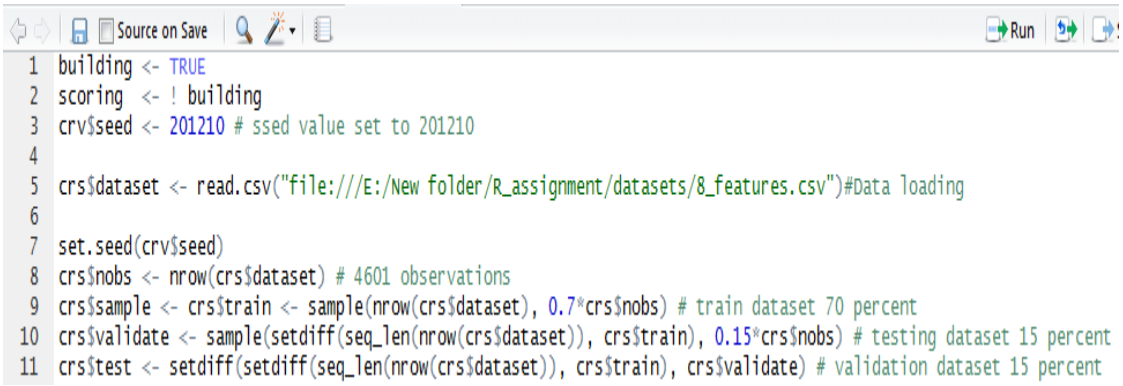

Figure 4.4: Dataset Division

So for validating the dataset two cases are compared, first one is when the numbers of features are four and second one is when the numbers of features are five.

4.6.1 Top Four Features

Decision tree ROC values are 0.89 and 0.88, for ADA boost model ROC values are 0.91 and 0.90, for random forest ROC values are 0.89 and 0.84, for kernel support vector machine ROC values are 0.91 and 0.88, for generalized linear model ROC values are 0.91 and 0.90, for Neural network ROC value is 0.91 and 0.90 .Similarly for decision tree accuracies are 87.11 and 85.65 , for ADA boost model accuracies are 87.83 and 86.95 , for random forest accuracies are 88.05 and 86.81 , for Kernel support vector machine accuracies are 86.96 and 85.79 , for generalized linear mode accuracies are 86.38 and 85.07 , for Neural network accuracies are 85.73 and 84.78 , Similarly for decision tree confusion matrix values are 0.128 and 0.127 , for ADA boost model confusion matrix values are 0.121 and 0.131 ,for random forest confusion matrix values are 0.119 and 0.125 , for Kernel support vector machine confusion matrix values are 0.13 and 0.14 , for generalized linear model confusion matrix values are 0.136 and 0.149 ,for Neural network confusion matrix values are 0.142 and 0.151 . All the above comparison is done between Table 4.14 and Table 4.17 respectively.

Table 4.17: Validation Dataset for Four Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	85.65	0.127	0.91	0.88
2	ADA boost model	86.95	0.13	0.93	0.90
3	Random forest	86.81	0.125	0.88	0.84
4	KSVM	85.79	0.142	0.91	0.88
5	GLM	85.07	0.149	0.92	0.90
6	Neural network	84.78	0.151	0.93	0.90

4.6.2 Top Five Features

For Decision tree ROC values are 0.90 and 0.90 , for ADA boost model ROC values are 0.94 and 0.94 , for random forest ROC values are 0.93 and 0.93 , for kernel support vector machine ROC values are 0.92 and 0.92 , for generalized linear model ROC values are 0.93 and 0.94 , for Neural network ROC values are 0.94 and 0.93 . The accuracies for decision tree are 88.99 and 88.84 , for ADA boost model accuracies are 89.50 and 89.55 , for random forest accuracies are 89.35 and 89.27 , for Kernel support vector machine accuracies are 89.28 and 88.69 , for generalized linear mode accuracies are 84.47 and 86.95 , for Neural network accuracies are 89.71

and 89.41 , Similarly for decision tree confusion matrix values are 0.12 and 0.11 , for ADA boost model confusion matrix values are 0.086 and 0.084 , for random forest confusion matrix values are 0.076 and 0.075 , for Kernel support vector machine confusion matrix values are 0.089 and 0.088 , for generalized linear model confusion matrix values are 0.09 and 0.089 , for Neural network confusion matrix values are 0.088 and 0.087. All the above comparison is done between Table 4.13 and Table 4.18.

Value of accuracy, confusion matrix and ROC values are close to each other between validation dataset and non-validation dataset that means approach performed in this work is correct.

Table 4.18: Validation Dataset for Five Features

S.no	Model Name	Accuracy	Confusion matrix	Risk	ROC curve
1	Decision tree(Rpart)	88.84	0.11	0.92	0.90
2	ADA boost model	89.55	0.084	0.96	0.94
3	Random forest	89.27	0.075	0.95	0.93
4	KSVM	88.69	0.088	0.94	0.92
5	GLM	86.95	0.089	0.95	0.94
6	Neural network	89.42	0.087	0.95	0.93

4.7. Comparison on Different Partition

As discussed above dataset is divided into two partitions, first one is testing data, and the second one is training data. For validating the dataset, the dataset is divided into three partitions (Table 4.17 and Table 4.18), the first one is training data, the second one is testing data and the third one is validation data. Here training and testing data is divided in the ratio of 70, 30 respectively. So, next task is to partition the dataset with different ratio and after that compare the result of all the partitions. For measuring the performance two parameters are proposed, the first one is accuracy and the second one is ROC value. This is used to check whether the models are working efficiently for all the partitions or not. There is no need to validate the dataset so here dataset is divided into only two parts, training data and testing data.

4.7.1 Accuracy on Different Partition

In table 4.19 dataset divided into five partitions i.e. 50-50, 60-40, 70-30, 80-20, 90-10.Now the comparison is done for different partitions. First for decision tree when both training and testing data is divided into half i.e.50-50 accuracy is 89.04, when

training and testing data is divided into 60-40 accuracy is 89.51 which is very much close to previous partition, when training and testing data is divided into 70-30 accuracy is 88.77 which is very much close to previous two partition, similarly when partition is 80-20 and 90-10 on decision tree accuracies are 88.70 and 89.15 which is close to all the previous partition. Similarly for ADA boost model when training and testing data divided into half accuracy is 91.69, when partition is 60-30 accuracy is 91.90 which is very much close to previous partition, similarly when training and testing data is divided into 70-30, 80-20, 90-10 accuracies are 91.60 ,91.10 and 91.10 respectively which is very much close to previous partitions. Similarly for Random forest for partition 50-50,60-40,70-30,80-20 and 90-10 accuracies are 91.82, 91.90, 91.45, 91.20, 90.88 respectively, where all the accuracies are close to one another .Similarly for kernel support vector machine ,generalized linear model and neural network (Table 4.19) for different partitions accuracies are closed to one another that means the models are performing well on the dataset for different partitions this also means more machine learning models can be implemented on this dataset and predict the best models for the dataset.

Table 4.19: Accuracy on Different Partition

S.no	Model Name	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
		50-50	60-40	70-30	80-20	90-10
1	Decision tree(Rpart)	89.04	89.51	88.77	88.70	89.15
2	ADA boost model	91.69	91.90	91.60	91.10	91.10
3	Random forest	91.82	91.90	91.45	91.20	90.88
4	KSVM	91.17	90.54	90.22	89.46	88.93
5	GLM	90.39	90.38	89.34	89.14	89.58
6	Neural network	91.69	91.58	91.16	90.33	90.67

4.7.2 ROC on Different Partition

In table 4.20 dataset is divided into five partitions i.e. 50-50, 60-40, 70-30, 80-20, 90-10.Now the comparison is done for different partitions. First for decision tree when both training and testing data is divided into half i.e.50-50 area under curve is 0.91, when training and testing data is divided into 60-40 area under curve is 0.92 which is very much close to previous partition, when training and testing data is divided into 70-30 area under curve is 0.90 which is very much close to previous two partition,

similarly when partition is 80-20 and 90-10 on decision tree area under curve is 0.90 and 0.90 which is close to all the previous partition. Similarly for ADA boost model when training and testing data is divided into half area under curve is .97, when partition is 60-30 accuracy is 0.97 which is equals to previous partitions, similarly when training and testing data is divided into 70-30,80-20,90-10 area under curve is 0.97, 0.97 and 0.96 respectively, which is almost equals to previous partitions. Similarly for Random forest for partitions 50-50,60-40,70-30,80-20 and 90-10 area under curve are 0.95, 0.96, 0.95, 0.96, and 0.95 where all the area under the curve is almost equals to one another. Similarly for kernel support vector machine, generalized linear model and neural network (Table 4.20) for different partitions area under curve are closed to one another that mean the models are performing well on this dataset for different partitions this also means more machine learning models can be implemented on this dataset and predict the best models for the dataset.

Table 4.20: ROC on Different Partition

S.no	Model Name	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
		50-50	60-40	70-30	80-20	90-10
1	Decision tree(Rpart)	.91	.92	.90	.90	.90
2	ADA boost model	.97	.97	.97	.97	.96
3	Random forest	.95	.96	.95	.96	.95
4	KSVM	.95	.95	.94	.94	.93
5	GLM	.96	.95	.95	.95	.94
6	Neural network	.96	.96	.96	.96	.96

4.8. Model Comparison

After comparison the performance of the different partitions (Table 4.19 and Table 4.20), next task is to compare the dataset for different models. Models used for comparison are Decision tree, ADA boost model, Random Forest, Kernel support vector machine, Generalized linear models, Neural network, Bagged Flexible discriminant analysis, Oblique tree, Tree model for genetic algorithm, multivariate adaptive regression spline, C5, J48, Logistic model tree, M5P, Mixed discriminant analysis. The comparison is done on the basis of accuracy, ROC curve; Confusion Matrix, Sensitivity, Specificity, and Kappa value (Table 4.21). The comparison was done when number of feature are eight. Now out of fifteen models, selection should be done for top five models of the dataset. After analyzing

value of accuracy, ROC curve, Confusion matrix, Sensitivity, Specificity, kappa value for fifteen different models best five models for the spam dataset are oblique tree, Tree models for genetic algorithms, ADA boost model, Random forest, Neural network (Table 4.21). The new parameter for comparison of model performance is Kappa value. This distribution is defined by Landis and Koch [2].

Table 4.21: Comparison of Fifteen Models

S.no	Model Name	Accuracy	ROC curve	Confusion matrix	Sensitivity	Specificity	Kappa Value
1	Decision tree(Rpart)	88.77	0.90	0.112	0.8854	0.8921	0.7589
2	ADA boost model	91.60	0.97	0.083	0.9326	0.8857	0.8229
3	Random forest	91.45	0.95	0.086	0.9264	0.8952	0.8192
4	KSVM	90.24	0.94	0.097	0.9199	0.8741	0.7937
5	GLM	89.34	0.95	0.1	0.9015	0.8956	0.785
6	Neural network	91.16	0.96	0.088	0.923	0.8929	0.813
7	FDA	90.58	0.96	0.094	0.8943	0.9087	0.7838
8	Oblique Tree	91.57	0.96	0.085	0.9264	0.8952	0.8192
9	Tree model for genetic algorithm	91.45	0.94	0.085	0.9284	0.8922	0.8195
10	Multivariate adaptive regression spline	90.80	0.96	0.091	0.9083	0.9076	0.8036
11	C5	90.65	0.93	0.093	0.9326	0.8673	0.8041
12	J48	90.80	0.93	0.091	0.9317	0.8718	0.8069
13	Logistic model tree	90.94	0.96	0.093	0.9288	0.8792	0.8094
14	M5P	89.86	0.96	0.101	0.9165	0.8701	0.786
15	Mixed discriminant analysis	82.83	0.92	0.171	0.8146	0.861	0.621

4.9. K-fold Validation

After finding the top five models for the dataset, next task is k-fold validation on that dataset. K-fold validation is basically used for determining the robustness of the dataset. In k-fold validation, different seed values are used to find the accuracy for each machine learning model. If the graph is straight line then the model is robust otherwise not.

In k-fold validation k is variable, which is varied from number of seed values. For example if there are five seed values for the experiment than it is 5-fold validation and if there are eight seed values than it is 8-fold validation. In this experiment ten different seed values are used hence it is called 10-fold validation for top five model i.e. oblique tree, Tree models for genetic algorithms, ADA boost model, Random forest, Neural network.

4.9.1. Ten-Fold Validation on Oblique Tree

Table 4.22 shows 10-fold validation when the model is oblique tree. For different seed values, accuracy is calculated. In Table 4.22 all the accuracies are very close to each other. Maximum accuracy for the oblique tree is 91.745 when seed value is 451510 and minimum accuracy is 90.152 when seed value is 11111. For two seed values, i.e. 451510 and 88 the accuracies are equal i.e. 91.745.

Table 4.22: 10-Fold Validation for Oblique Tree

S.no	Seed Value	Accuracy
1	451510	91.745
2	908040	91.238
3	847	92.541
4	8568	90.514
5	11111	90.152
6	78965	91.527
7	45	91.021
8	87684	91.093
9	94387	91.383
10	888	91.745

Figure 4.22 is a graph for oblique tree for above ten seed values (described in Table 4.22). In this graph there is a straight line for oblique tree which shows the robustness of the model.

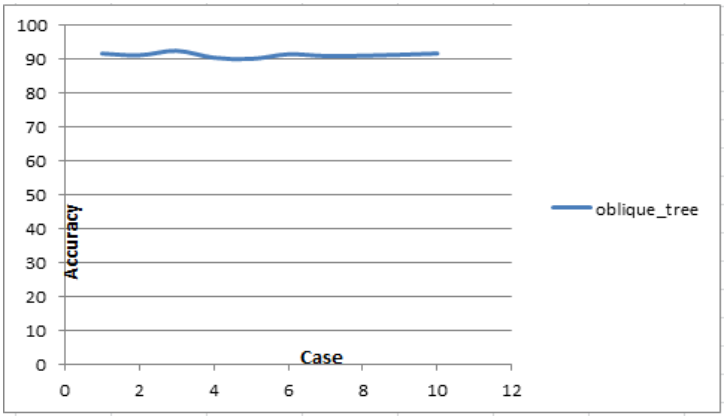

Figure 4.5: Performance for Oblique Tree

4.9.2 Ten-Fold Validation on Tree models for Genetic Algorithms

Table 4.23 shows 10-Fold validation on spam dataset when the model is Tree models for genetic algorithms. For different seed value all the accuracies are very close to each other. Maximum accuracy for Tree models for genetic algorithms is 92.759 when seed value is 11111 and least accuracy value for Tree models for genetic

algorithms is 90.441 when seed value is 8568. All the accuracies are close to each other that means model is working efficiently on every seed value

Table 4.23: 10-Fold validation for Tree Models for Genetic Algorithms

S.no	Seed Value	Accuracy
1	451510	92.252
2	908040	91.165
3	847	91.817
4	8568	90.441
5	11111	92.759
6	78965	91.310
7	45	91.745
8	87684	91.31
9	94387	90.514
10	888	90.731

Figure 4.5 shows the graph for Tree models for genetic algorithms for ten different seed value ev_tree is the package used for implementation of Tree models for genetic algorithms on R .As the graph shows the straight line that means the models (Tree models for genetic algorithms) is robust for every seed value. Maximum value of graph is obtained when x-axis value is five and it is clear from the Table 4.23 maximum accuracy is obtained when seed value is 11111 which is at number 5.

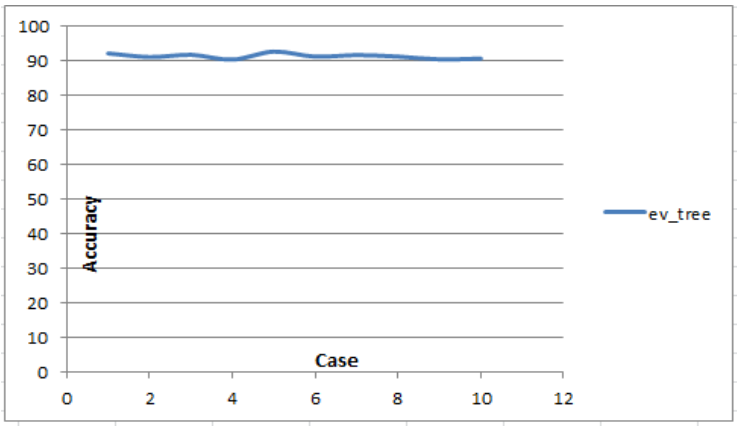

Figure 4.6: Performance for Tree models for Genetic Algorithms

4.9.3 Ten-Fold Validation on ADA boost model

Table 4.24 shows 10-fold validation when the model is ADA boost model. For different seed values accuracies are very close to each other. Maximum accuracy for ADA boost model is 92.541 when seed value is 11111 and least accuracy is 90.948 when seed value is 8568. For different seed value accuracy are close to each other that means ADA boost model work efficiently for all seed values.

Table 4.24: 10-Fold Validation for ADA Boost Model

S.no	Seed Value	Accuracy
1	451510	92.469
2	908040	92.324
3	847	92.469
4	8568	90.948
5	11111	92.541
6	78965	92.034
7	45	91.238
8	87684	91.889
9	94387	91.165
10	888	92.252

Figure 4.6 is a graph for ADA boost model for ten different seed values. The package used for ADA boost model is ada and plyr. The graph is again a straight line that means model (ADA boost model) works efficiently for each seed value. Maximum value obtained when the value of X-axis is five and at Table 4.24 this value is 11111. ADA boost model is also called adaptive boost (as discussed in Figure 4.6). With the help of this graph, conclusion is that all the accuracies are lying between 90 and 100 (see Figure 4.6)

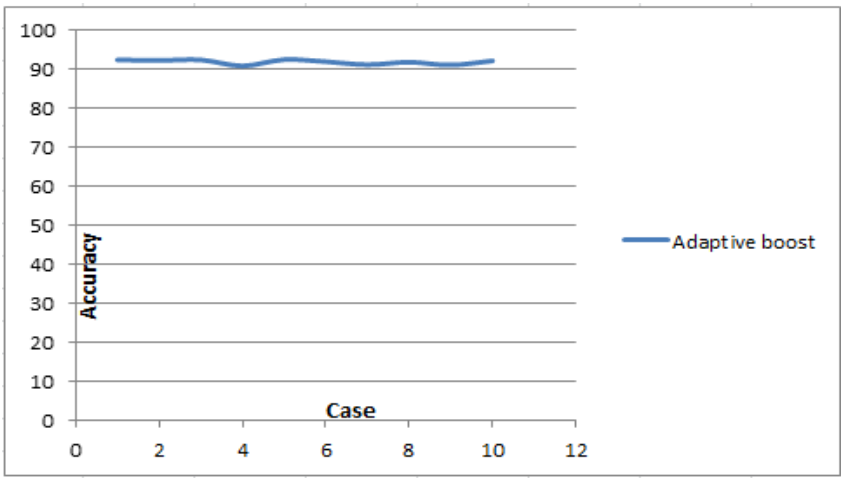

Figure 4.7: Performance for ADA boost model

4.9.4 Ten-Fold Validation on Random forest

Table 4.25 shows 10-fold validation when the model is the random forest. For different seed values, accuracies are close to each other that mean the model (random forest) is efficient for all other seed values. Maximum value of accuracy while performing 10-fold validation is 92.541 for seed value 11111. Minimum value of accuracy while performing 10-fold validation is 90.948 for seed value 8568, For two different seed values i.e. 451510 and 847 seed value are equal to 92.469.

Table 4.25: 10-Fold Validation for Random forest

S.no	Seed Value	Accuracy
1	451510	92.469
2	908040	92.324
3	847	92.469
4	8568	90.948
5	11111	92.541
6	78965	92.034
7	45	91.238
8	87684	91.889
9	94387	91.165
10	888	92.252

Figure 4.7 is graph for the random forest while performing 10-fold validation. The package used for implementing random forest is RandomForest. For random forest model, the graph shows a straight line that means the model (random forest) is robust. For seed value 451510 and 847, accuracies are equal and this scenario is easily noticed on Figure 4.7.

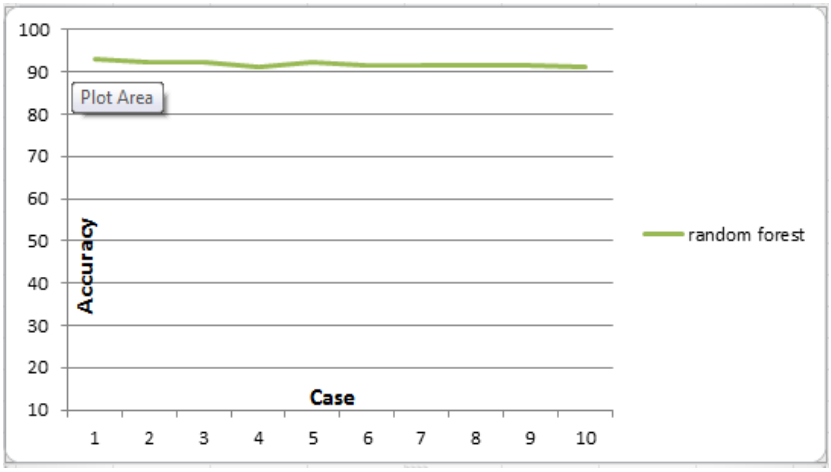

Figure 4.8: Performance for Random forest

4.9.5 Ten-Fold Validation on Neural Network

Table 4.26 shows 10 fold validation on dataset when the model is the neural network. The experiment performed on ten different seed values (as shown in Table 4.26). For different seed values, accuracies are very close to each other. Maximum accuracy obtained while performing 10-fold validation is 92.614, for seed value 847. Minimum accuracy obtained 91.456 for seed value 8568. For seed values 78965 and 888 accuracies are equal, which is 92.179 and for seed value 908040 and 87684 accuracies are again equal, which is 92.107 and for seed value 451510 and 11111 seed value accuracies are again equal, which is 92.324.

Table 4.26: 10-Fold validation for Neural Network

S.no	Seed Value	Accuracy
1	451510	92.324
2	908040	92.107
3	847	92.614
4	8568	91.456
5	11111	92.324
6	78965	92.179
7	45	92.034
8	87684	92.107
9	94387	91.527
10	888	92.179

Figure 4.8 is a graph for the neural network while performing 10-fold validation. The package used for the neural network is Nnet. The graph shows the straight line that means the model (neural network) is robust for every seed value.

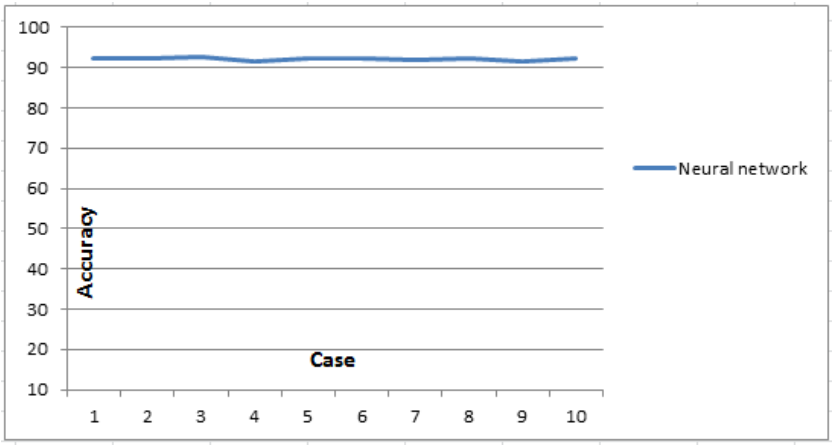

Figure 4.9: Performance for Neural network

4.10. Ensemble Top Five Models

After comparison of fifteen machine learning models, next task is to find out top five models for spam dataset. After performing comparison on the basis of Accuracy, Confusion matrix, Sensitivity, Specificity and ROC value (Table 4.21) best five models for spam dataset are oblique tree, Tree models for genetic algorithms, ADA boost model, Random forest, Neural network. After that, 10-fold validation on these top five models having seed values 451510, 908040, 847, 8568, 11111, 78965, 45, 87684, 94387 and 888.

Now next task is to ensemble these five models in such a way that it gives maximum accuracy. Discussed in Table 4.27 experiment performed on top five models and for

these top five models perform different permutation to calculate maximum accuracy. After performing twenty permutations, maximum accuracy is 91.89 which is for ADA boost model, random forest, neural network and oblique tree. Table 4.21 shows experiment done for seed value 201210. So now consider Table 4.27 and compare the accuracy w.r.t the accuracy obtained in Table 4.21. It is easily noticeable that the accuracy of ensemble model is greater than the individual model. So for the spam dataset maximum accuracy for ADA boost model, random forest, neural network and oblique tree is 91.89. Random forest, oblique tree, ev tree, ADA, nnet and ev tree and Random forest, nnet, ev tree have accuracy equals to 91.52 while ADA, random forest and nnet and ADA and random forest have accuracy equals to 91.67 while nnet and oblique tree, nnet and ev tree, ADA and nnet and Oblique tree and ev tree have accuracy equals to 91.3.

Table 4.27: Ensemble Top Five Models

S.No	Ensemble model	Accuracy
1	ADA and random forest	91.67
2	ADA,random forest and nnet	91.67
3	ADA,random forest,nnet,oblique tree	91.89
4	ADA,random forest,nnet,oblique tree,ev tree	91.45
5	ADA and nnet	91.31
6	ADA and oblique tree	91.45
7	ADA and ev tree	91.53
8	Random forest and nnet	91.16
9	Random forest and oblique tree	91.31`
10	Random forest and ev tree	91.38
11	nnet and oblique tree	91.31
12	nnet and ev tree	91.31
13	Oblique tree and ev tree	91.31
14	ADA,nnet,Oblique tree,ev tree	91.45
15	nnet,oblique tree,ev tree	91.45
16	Random forest,nnet,oblique tree and ev tree	91.31
17	Random forest,oblique tree,ev tree	91.52
18	ADA,nnet and ev tree	91.52
19	Random forest,nnet,ev tree	91.52
20	ADA,random forest,oblique tree,ev tree	91.38

After performing all the permutation on top five models i.e. oblique tree, Tree models for genetic algorithms, ADA boost model, Random forest, Neural network. According to Table 4.21 best ensemble model for the dataset is ADA boost model, random forest, Neural network, oblique tree with accuracy 91.89 (refer to Table 4.27). From Figure 4.4 to 4.8 there are individual graph for top five models for spam dataset. Figure 4.9 is combined graph that is obtained after performing model ensemble.

Figure 4.10 shows all the graph lines are almost straight that means models are robust for spam dataset.

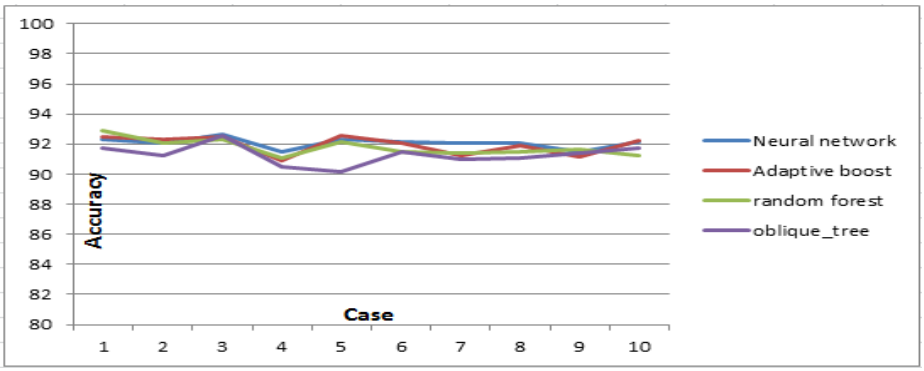

Figure 4.10: Performance for Ensemble Models

Chapter 5
Conclusion and Future Scope

This chapter describes conclusion of the work present in the thesis and discusses the contribution of the thesis and the future work of this thesis.

5.1. Conclusion

In this work is done on fifteen different machine learning models for the spam dataset. After implementing fifteen models on the spam dataset, top five models, which are best for the dataset are Boosted classification tree, Random forest, Neural Network, Oblique tree on the basis of accuracy, area under the curve, confusion matrix and kappa value. Different permutations were done on top five models (refer to Table 4.27). After performing the entire permutations, ensemble model for the spam dataset is proposed.

5.2. Summary of Contributions

The contributions made for the research work presented in this thesis are summarized as follows:

- UCI machine repository, which plays an important role for providing the dataset.
- Comparison of Model on the basis of Accuracy, ROC curve, Kappa value, Confusion matrix, Sensitivity, Specificity.
- Study of applications of dataset in real-world projects such as Visualization and Data Mining in 3D Immersive Environment: Summer Project 2003, Online Policy Adaptation for Ensemble Classifiers, Modeling for Optimal Probability Prediction.
- Methodology to convert Models predicted output into accuracy in Excel.
- Comparative analysis of different models on the basis of their different seed value and different partition value.

5.3. Future Scope

Ensembling the model plays an important role to increase the efficiency or accuracy of the model. This work is tested on spam dataset, so this ensemble model is applicable to another spam dataset (Gmail dataset, Yahoo dataset etc.). The various algorithms used for spam filtering are the Genetic algorithm, Support vector machine, Naïve Bayes classifier etc. Instead of the individual algorithm, if spam dataset uses proposed ensemble model approach than the spam filtering becomes much more efficient.

References

1. M. Xue, C. Zhu, ”A Study and Application on Machine Learning of Artificial Intelligence”, in International Joint Conference on Artificial Intelligence, April 2009.
2. W. Yuntian, ”Based on Machine Learning of Data Mining to Further Explore”, in International Conference on Computer Science and Information Processing (CSIP), August 2012.
3. P. Kulkarni, ”Reinforcement and Systemic Machine Learning for Decision Making, First Edition”, in Institute of Electrical and Electronics Engineers, Inc, John Wiley & Sons, Inc, 2012.
4. S. Geerthik and T.P Anish, “Filtering Spam: Current Trends and Techniques”, International Journal of Mechatronics, Electrical and Computer Technology Austrian E-Journals of Universal Scientific Organization, Volume. 3, pp 208-223, July 2013.
5. M. Basavaraju, “A Novel Method of Spam Mail Detection using Text Based Clustering Approach”, International Journal of Computer Applications, Volume 5, August 2010.
6. R. K. Kumar, G. Poonkuzhali, P. Sudhakar, ” Comparative Study on Email Spam Classifier using Data Mining Techniques”, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Volume 1, March 2012.
7. V. Jaswal, “Spam Detection System Using Hidden Markov Model”, International Journal of Advanced Research in Computer Science and Software Engineering, July 2013
8. A. Mali, “Spam Detection Using Baysian with Pattren Discovery”, International Journal of Recent Technology and Engineering (IJRTE), Volume 2, July 2013.
9. R. Islam, Y. Xiang, member IEEE, “Email Classification Using Data Reduction Method”, June 2010.
10. N. Singh,” Dendritic Cell Algorithm and Dempster Belief Theory Using Improved Intrusion Detection System“, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, July 2013.
11. J. Greensmith, “The Dendritic Cell Algorithm”, Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy, October 2007.
12. J. R. Quinlan, “Induction of decision trees”, Machine learning, Volume 1, pp 81–106, 1986.

13. Y. Freund, R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of Computer and System Sciences, Volume 55, pp 119-139, 1997.
14. A. Liaw, M. Wiener, “Classification and Regression by randomForest”. R News, Volume 2, pp18–22, 2002.
15. S. S. Keerthi, E. G. Gilbert, “Convergence of a generalized SMO algorithm for SVM classifier design”. Machine Learning, Volume 46, pp 351–360, 2002.
16. J. M. Chambers. “Computational methods for data analysis”. Applied Stat, Volume 1, pp 1–10, 1977.
17. M. Riedmiller, H. Braun, “A direct adaptive method for faster back propagation learning: The RPROP algorithm”. In IEEE International Conference on Neural Nets, pp 586–591, 1993.
18. T. Hastie, R. Tibshirani, A.buja, “Flexible Discriminant analysis by optimal scoring”, Journal of the American statistical association, Volume 89, December 1994.
19. L. Brieman, J. Friedman, R. Olshen, C. Stone. “Classification and Regression Trees”, Wadsworth, 1984.
20. T. Grubinger A. Zeileis, K. P. Pfeiffer, “evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R”, Journal of Statistical Software, Volume 64, September 2014.
21. J. H. Friedman, “Multivariate adaptive regression spline”, the annals of statistics, Volume 19, 1991.
22. J. H. Friedman, “Fast MARS”, Department of statistics Stanford University, Technical report No.1991, May 1993.
23. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
24. N. Landwehr, Mark Hall and Eibe Frank, “Logistic Model Trees”, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2003.
25. N. Landwehr, Mark Hall, Eibe Frank, “Logistic Model Trees”, Machine Learning, Volume 59, pp 161–205, 2005.
26. E Frank, Y Wang, S Inglis, G Holmes, IH Witten. “Using model trees for classification”, Machine Learning, Volume 32, pp 63–76, 2006.
27. J. R. Landis, G. G. Koch, “The measurement of observer agreement for categorical data”, Volume 33, pp 159-74, 1977.

List of Publications

1. Paritosh Pantola, Anju Bala and Prashant Singh Rana, "Consensus based ensemble model for spam detection", Fourth International Conference on Advances in Computing, Communications and Informatics (ICACCI-2015), Kochi, June 8, 2015**. [Accepted**].
2) Paritosh Pantola and Anju Bala, "Ensemble Model for Spam Detection", Third International Conference on Information System Design and Intelligent Applications, August 30, 2015.**[Communicated]**

Video Presentation

https://www.youtube.com/watch?v=UTYMqjci-VU