A DATA MINING APPROACH FOR SECURE CLOUD USING ENHANCED RANDOM FOREST

A DISSERTATION

BY

Shikha Pathania

To

Department of Computer Science and Engineering

In partial fulfillment of requirement for the

Award of the degree of

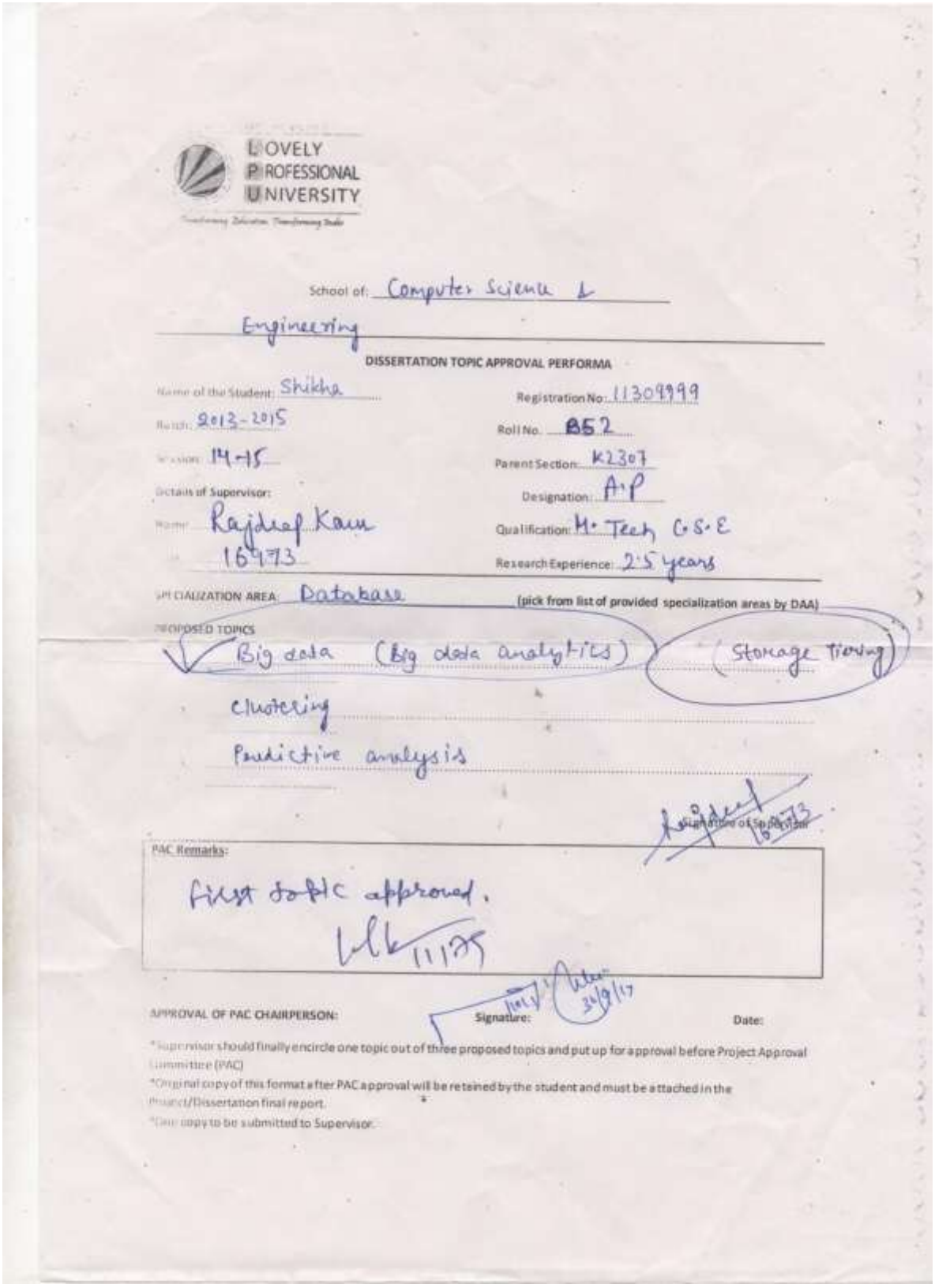
Master of Technology in Computer Science

Under the guidance of

Rajdeep Kaur

(May 2015)

PAC APPROVAL



LOVELY PROFESSIONAL UNIVERSITY

School of: Computer Science & Engineering

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: Shikha — Registration No: 11309999

Batch: 2013-2015 — Roll No: B52

Session: 14-15 — Parent Section: K2307

Details of Supervisor:

Name: Rajdeep Kaur — Designation: A.P

UID: 16473 — Qualification: M. Tech C.S.E

Research Experience: 2.5 years

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

PROPOSED TOPICS

1. Big data (Big data analytics) (Storage tiering)
2. clustering
3. Predictive analysis

Signature of Supervisor 16473

PAC Remarks:

first topic approved.

11/9/17

APPROVAL OF PAC CHAIRPERSON: Signature: 21/9/17 Date:

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.

*One copy to be submitted to Supervisor.

ABSTRACT

Random forest is an ensemble method which is widely used is application having large datasets because of its interesting features like handling imbalanced data, identifying variable importance and detecting DOB error rate. It consists of large number of decision trees. For building random forest randomness is established in two ways: Firstly by creating samples from original datasets randomly and Secondly at the time of creation of each tree, randomly selecting subsets of attributes at each node for best splitting decisions. But with the randomness in the bagging and feature selection Random forests are likely to have uninformative attributes which will lead to poor accuracy results and bad performance of the algorithm. In this paper we are providing an improved Feature selection Random Forest. In this first we are selecting the good features by applying the consistency on attributes after that we are combining this consistency based feature with the Random forest. Also most of the organizations are moving to the cloud data ,so we are performing the mining operation on the cloud based data. To protect the data from the unauthorized user we are securing the cloud data using AES algorithm through this no unauthorized user can access the data.

ACKNOWLEDGEMENT

First and foremost I would like to thank almighty for giving me courage to bring up this pre dissertation. Before getting into thick and thin of this pre dissertation I would like to show my gratitude to some of the people who have helped me in this project. Firstly I would like to purpose a word thanks to my mentor RAJDEEP KAUR who has encouraged me to get through this pre dissertation. Secondly I would like to thanks my friends who gave me unending support and helped me in numerous ways from the stage when the idea of the thesis was conceived. I am very thankful to all of them for making my work complete successfully under their guidance.

DECLARATION

I hereby declare that the dissertation proposal entitled, **Data mining approach for secure cloud using Enhance Random Forest** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date:

Shikha Pathania

RegNo.:11309999

CERTIFICATE

This is to certify that Shikha Pathania has completed M.Tech dissertation proposed titled **A Data mining approach for secure cloud using Enhance Random Forest** under my guidance and supervision. To the best of my knowledge the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.

The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech computer science and Engineering.

Date: _____

Signature of supervisor:

Name:

UID:

Table of Contents

TOPIC **PAGE NO.**

LIST OF TABLES

TABLE NAME **PAGE NO.**

LIST OF FIGURES

FIGURE NAME **PAGE NO.**

CHAPTER 1

INTRODUCTION

1.1 DATA MINING: INTRODUCTION

Data mining is the process of extracting and analyzing large datasets to find out various hidden relationships, patterns and much useful information. Although data mining is also called KDD (Knowledge discovery in databases) it is actually considered as a part in knowledge discovery process[28]. Through the KDD process we can discover useful knowledge from the raw data collection. This process consists of some steps which are discussed as follow:

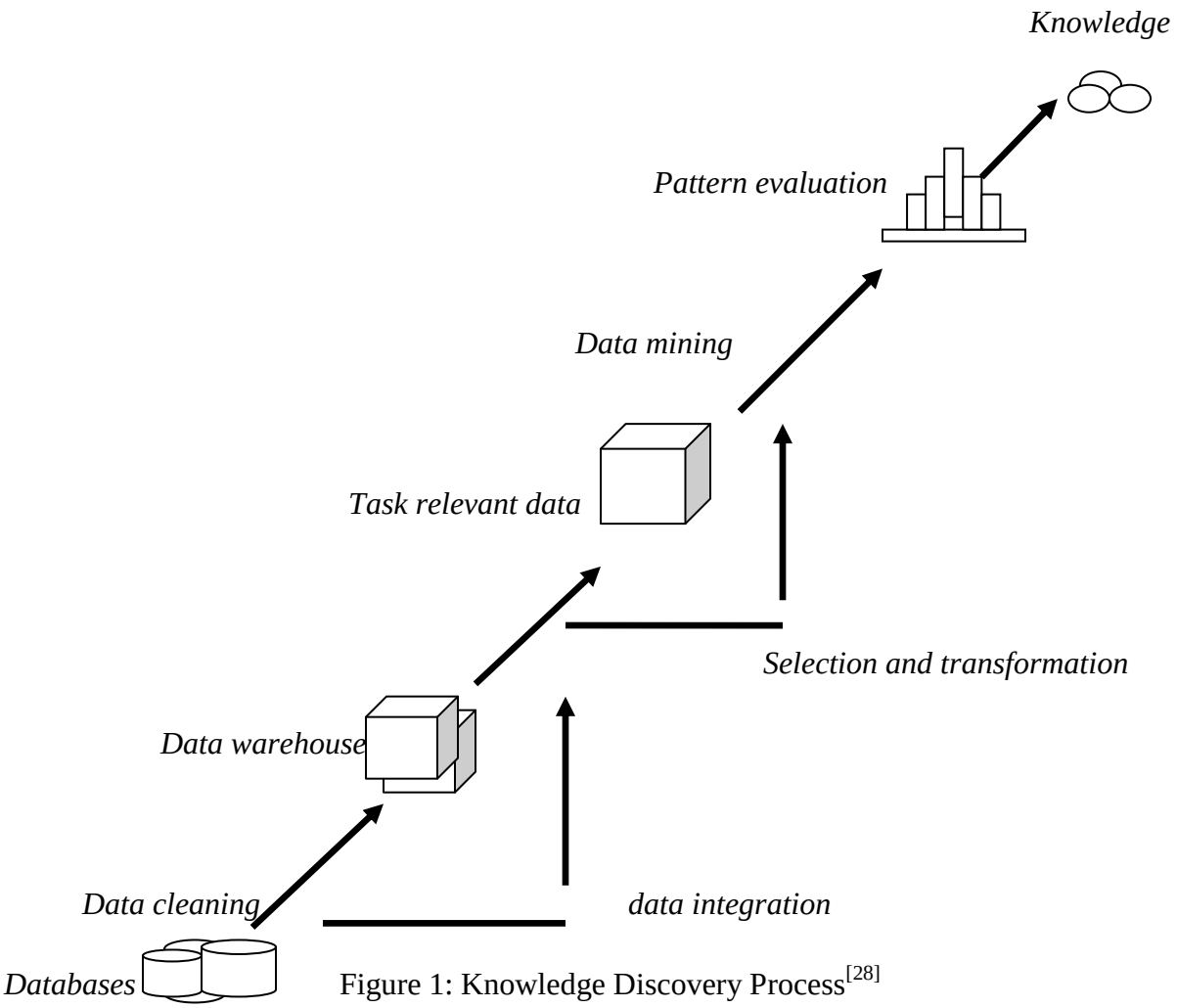


Figure 1: Knowledge Discovery Process[28]

- **Data cleaning:** In this step impure and irrelevant data is removed from the databases.
- **Data integration:** At this step various data sources that are in different form are converted into a common data type or form.
- **Data selection:** Now, the data that is relevant for the analysis is selected from the data collection.
- **Data transformation:** Data transformation is the process of transforming the data into the forms that are appropriate for data mining.
- **Data mining:** It is a process in which various data mining techniques are applied on the collected data to extract useful pattern or information.
- **Pattern evaluation:** In this step various patterns that are extracted from the data mining techniques are evaluated to find out the interesting patterns that are representing knowledge.
- **Knowledge representation:** It is the final step of KDD process. In this step the knowledge is represented to the user in such ways that is easily understandable by them. Various visualization techniques are used to represent the knowledge.

1.1.1 TYPES OF DATA THAT CAN BE MINED

There are various types of data in the information repository so data mining should be able to applied on all kind of data, it should not be limited to one type of data[39]. Different algorithms are used for different type of data each having their own difficulty level. Some of the data on which mining can be performed are discussed below:

- **Flat files:** Flat files are the most common data source for data mining .These are the simple data file which are either in text or binary format.
- **Relational databases:** In this the data is stored in the form of tables also called relation[28]. In a table there are rows and columns in which column represents attributes and rows represents the value for those attributes.
- **Data warehouse:** Data warehouse is a type of data store where the data from different sources are collected and kept in a unified form. Often the data we collected are heterogeneous data so first we need to transformed it into a common data type

after that we store them in data warehouse .The data stored in the warehouse is pure data.

- **Transaction databases:** Transaction database is a database in which each record represents transactions. In this each transaction is representing a timestamp, an identifier and set of items.
- **Spatial databases:** Spatial databases store geographical information like maps.

Time series databases: This type of data store time related information like logged activities or stock market information.

1.1.2 DATA MINING TASKS

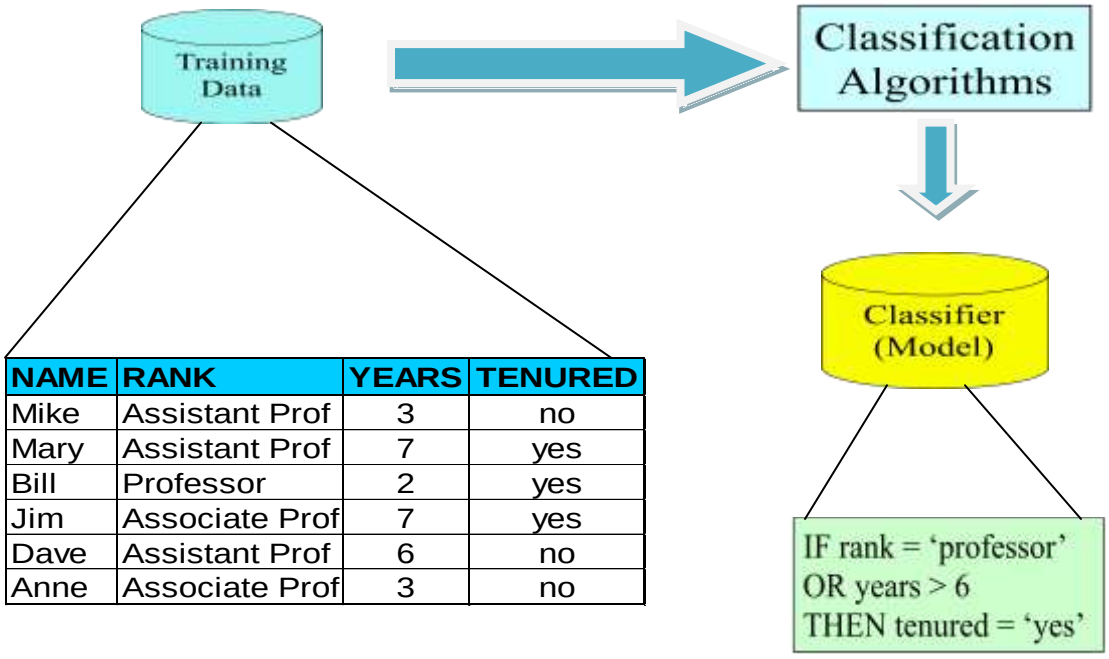
As it is mentioned above that data mining analyses data and find out the patterns[35] .The kind of patterns that are discovered depend upon the type of data mining task we used. There are two types of data mining tasks:

1. **Descriptive data mining tasks:** This data mining task find out the patterns that describes the general properties of existing data.
    - Clustering
    - Association rule discovery
    - Sequential pattern discovery
2. **Predictive data mining tasks:** This mining task performs predictions on the available data to find out the future values of other variables.
    - Classification
    - Regression
    - Deviation detection

1.1.3 CLASSIFICATION

Classification is a data mining technique which is used to categorize the data for its most effective and efficient use. In simple words we classify a data based on the class labels. In the classification technique the classes are already known, since the classes are already known it is called a supervised learning. Classification is done in two steps:

- **Building the classifier:** By analyzing the training set using the algorithm a classifier will be build. The algorithm will analyze the relationship between the attributes which will help in making classification rules.



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Figure 2: Building the classifier[34]

- **Using the classifier for the classification**

  After this the test data is used for testing the accuracy of classification rules. If the accuracy is acceptable then the classifier is applied on the new data tuples through which we can predict the outcome.

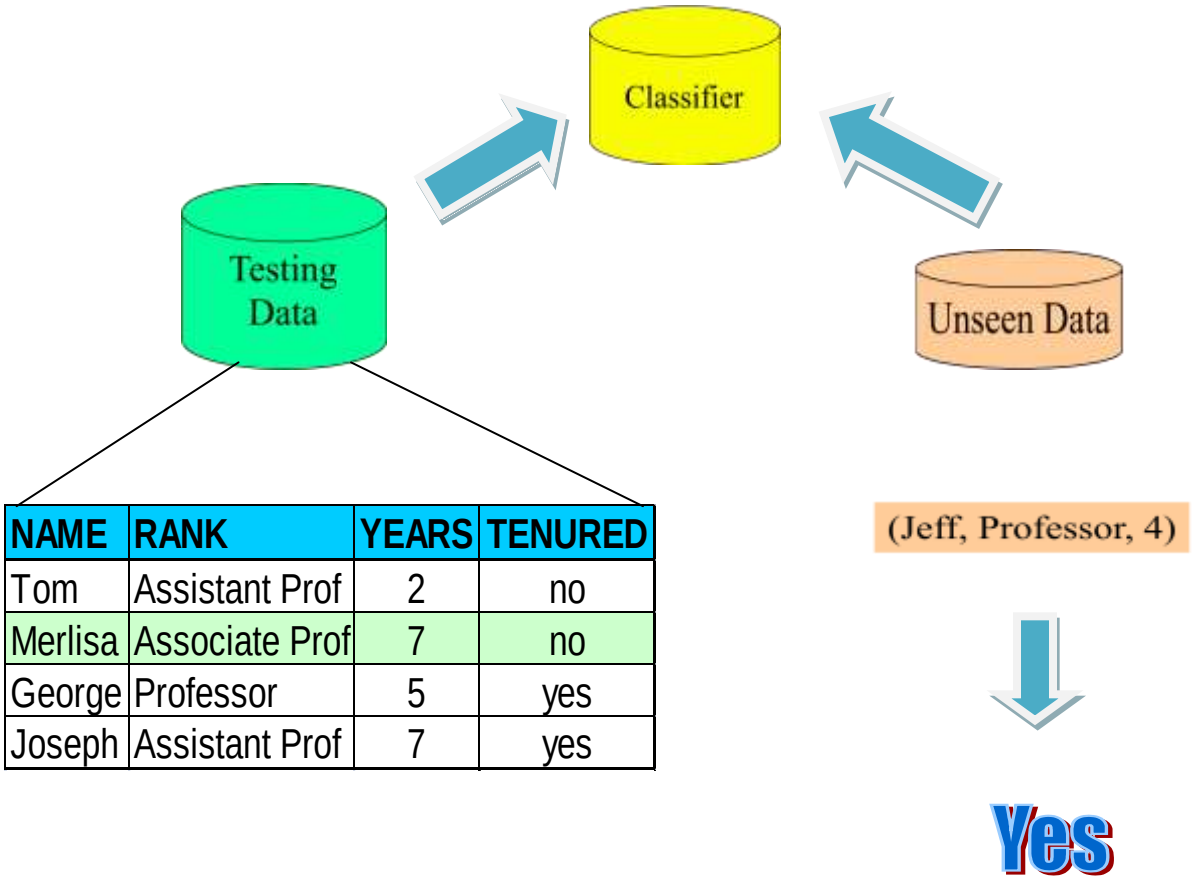


Figure 3: Process of using classifier for classification[34]

1.1.3.1 CLASSIFICATION TECHNIQUES

- Naive Bayesian classification
- Rule based classification
- SVM
- Decision trees

Other classification methods are:

- Genetic algorithms
- Rough set approach
- Fuzzy set approach

A. NAÏVE BAYESIAN CLASSIFICATION

Bayesian classification is based on bayes theorem and the classifiers build in this classification are statistical classifier. In this the data is classified on the basis of the probability like the probability of a given tuple that belongs to particular class [37].

Bayes Theorem

Two types of probability are there in this theorem:

- Posteriori probability:[P(C/X)]
- Prior Probability[P(H)/P(X)]

According to this theorem:

$$P(C/X) = P(X/C)\,P(C)/P(X)$$

P (C/X): probability of instance X being in class C.

P(X/C): probability of generating X instance in the given class C.

P(C): probability of occurring of class C.

P(X): probability of occurrence of instance X

Consider an example where we need to find out whether the persons working in a library whose name is Taylor is a male or female. We have a database with name and sex information.

NAME	SEX
Taylor	male
Sarah	female
Taylor	female
Taylor	female
Aiden	male
Kelly	female
Samantha	female
Trevor	male

P (male/Taylor) = (1/3 * 3/8) / (3/8)

=0.125/ (3/8) = 0.33333

P (female/Taylor) = (2/5 * 5/8) / (3/8)

=0.250/ (3/8) = 0.66666

So the person working in a library is likely to be a female.

B. RULE BASED CLASSIFICATION

In rule based classification records are classified using if –then rules [37]. Rules are written in the following way:

IF condition THEN conclusion

IF part is called the antecedent while the THEN part of rule is called consequent. Antecedents are the attributes and the consequent are the class variables. Consider an example of classification rule:

NAME	AGE	STUDENT	Buy_computer
Shikha	old	no	No
Shiny	youth	yes	Yes
Kshitiz	youth	yes	Yes

Rule 1: IF age=youth AND student=yes THEN buy_computer =yes

Rule 2: IF age=old AND student=no THEN buy_computer =no

Now for the following test data find out on which class labels the person belong.

NAME	AGE	STUDENT	Buy_computer
Rahul	youth	yes	?

In this test data Rule1 is applied so Buy_computer=yes for Rahul.

C. SVM

SVM is a machine learning algorithms which belongs to kernel method. In this they classify the data by finding out the hyper plane between points of two classes in multidimensional space[38]. Attributes that are closer to the hyper plane is called support vector.

D. DECISION TREES

Decision trees are the predictive models which uses tree like structure to conclude to a particular decision[35]. These trees are consists of root node, internal node and leaf node.

Each internal node denotes the test on the attribute, the branch indicates the outcome of test and the leaf node denotes the classes that a particular attribute belongs to. The topmost node is called the root node.

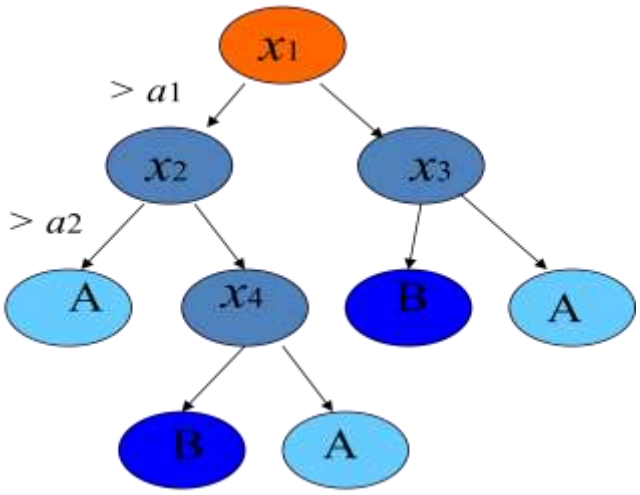


Figure 4: Structure of tree[33]

In this x1 is a root node x2, x3 and x4 are internal nodes whereas A and B are the leaf nodes.

If x1>a1 AND x2>a2 THEN class A

CONSTRUCTION OF DECISION TREES

Steps to create a decision tree:

1. First select the best feature from the dataset and then make it your root node of the tree.
2. After selecting the best feature perform partition on that feature. Partitioning can be done on the basis of three types of features that are:
    - **Categorical feature**: In this type of partitioning the number of division done on that dataset is equal to the number of values of that feature.
    - **Numerical feature:** In this partitioning is done on the basis of the numerical values.
    - **Nominal feature**

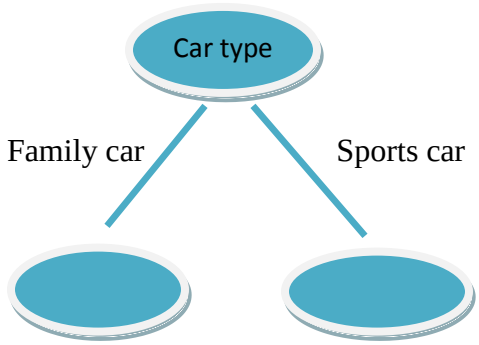


Figure 5: Nominal feature[33]

3. Perform the above steps repeatedly until the partitioning become pure.

Consider an example to explain the construction of the trees.

Outlook	Temp	Humidity	Windy	class
Sunny	75	70	true	Play
Sunny	80	90	true	Don't
Sunny	85	85	false	Don't
Sunny	72	95	true	Don't
Sunny	69	70	false	Play
Overcast	72	90	true	Play
Overcast	83	78	false	Play
Overcast	64	65	true	Play
Overcast	81	75	false	Play
Rain	71	80	true	Don't
Rain	65	70	true	Don't
Rain	75	80	false	Play
Rain	68	80	false	Play
Rain	70	96	false	Play

Figure 6: Dataset for classification [33]

9 play and 5 don't play samples are there in this following dataset. So total samples in this dataset are 14. Now we are creating a partitioning on the basis of categories and numeric feature.

Categorical partioning:

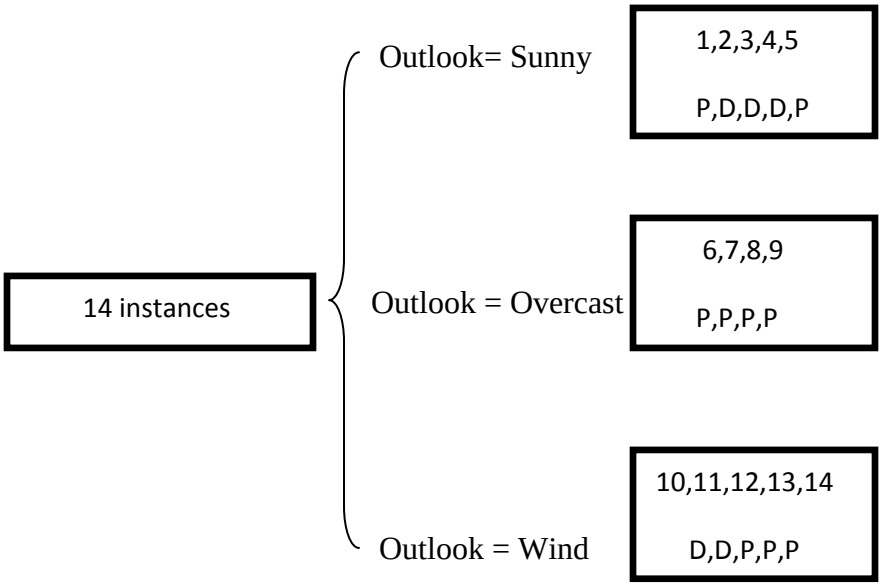


Figure 7: Categorical partitioning of the dataset[33]

Numerical partitioning:

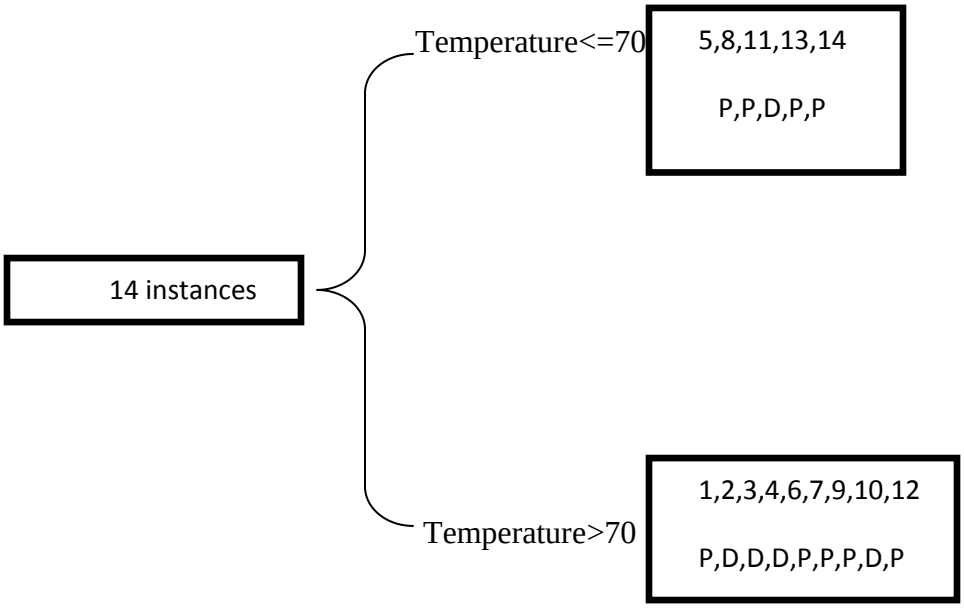


Figure 8: Numerical partitioning of the dataset[33]

We are creating a decision tree on the basis of categorical partitioning.

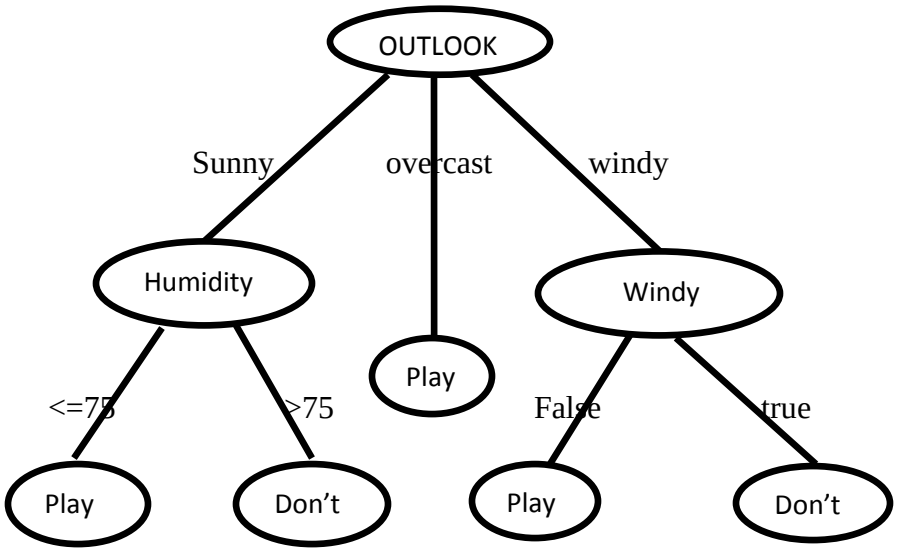


Figure 9 : Decision tree on the basis of categorical partitioning[33]

CART tree

CART are Classification and regression tree. Classification gives the response value in the categorical form while the regression tree provides the response dat in continous integer type[36]. Example of classification problem is like we want to predict that which students are going to graduate this year from stanford university and which students are not going to graduate this year. Similarly the example of regression problem is like predicting the selling prices of the house.

TREE PRUNING

Decision trees that are very large and complex leads to the over-fitting of training set because trees with too many branches return anomalies. Over-fitting means that the models that are generated describe the anomalies and noises instead of the relationships[37].

So in order to solve this problem tree pruning technique is used. This technique reduces the size and complexity of decision trees which in turn helps in removing the anomalies and noises from the training set. Two approaches are used for tree pruning

- **Pre Pruning:** In this approach the construction of the tree is stopped early[37].
- **Post Pruning:** In this approach the sub-tree is removed from the fully grown tree.

1.1.4 DECISION TREE ENSEMBLES

Ensemble is a supervised technique that uses many weak learners in order to produce a strong learner. It improves the accuracy. Some of the decision trees ensembles are:

- Bagging
- Boosting
- Random forest

A. BAGGING

Bagging is an ensemble method of combining multiple predictors and is proposed by Breiman in 1996. Bagging means bootstrap aggregating[26]. It constructs a large number of

trees with bootstrap sample from a dataset and provides prediction on the basis of voting of each tree. Following are the steps used in bagging algorithm:

- Let the original sample L= (X1, X2….. Xn)
- Create K samples from the original samples L by random sampling with replacement.
- Lastly train a classifier using each bootstrap sample
- Provide predictions on the basis of voting.

B. BOOSTING

Boosting is an ensemble method that is used in the area of data mining. In this it creates multiple classifiers each of which is generating some weighted observation[29]. Following are the steps used in boosting algorithm:

- Apply each classifier on the training dataset which will give some observation .Assign the weight to each of the observation.
- After that apply these observations on each of the sample. Apply the greater weight to those observations that are difficult to classify and the lower weight to those that are easy to classify[29].
- Sequentially combine all predictions of each of the classifier to provide a single best prediction.

C. RANDOM FOREST

Random forest is a supervised machine learning algorithm originally proposed by Leo Breiman in 1999. It is an ensemble method in which multiple trees are used as base classifiers and the classification with the majority of votes by each tree is chosen[36]. This algorithm develops set of trees by random selection of data or by random selection of variables. Since the trees in this are generated randomly and you have many trees it is called random forest algorithm. The classes that are generated are dependent on the variables.

Steps that are used in Random forest algorithm:

1. First select the data randomly from the training set.
2. Then generate the decision tree by using any of the two following ways:

- Bootstrap samples: In this trees are built using the different random samples of data with replacement.
- Random feature selection of input attributes of training set.

3. After creating the decision trees choose the classification with the majority of vote.

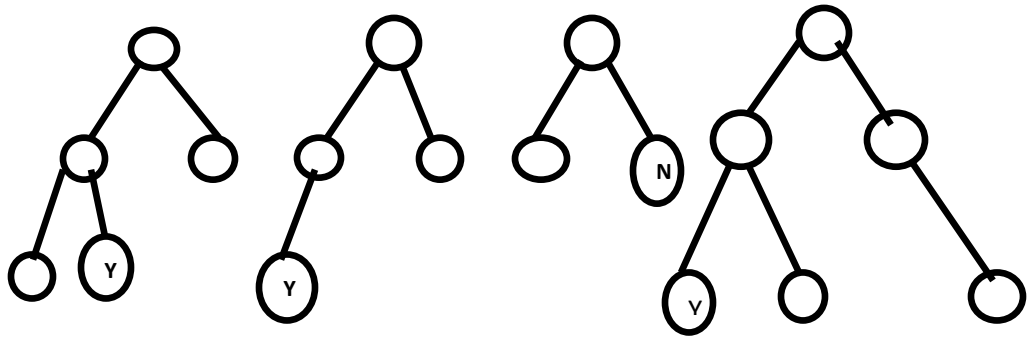


Figure 10: Structure of Random Forest

In this there are four decision trees and the majority of votes are given to the Y class so the classification result is Y.

The error rate of the Random forest is based on two things:

- Correlation between the trees: If the correlation between the increases then the error of the forest also increases[27].
- Strength of individual tree: The error rate of tree is low if the tree is a strong classifier.

ADVANTAGES AND DISADVANTAGES OF RANDOM FOREST

Advantages

- Random forest is one of the best classification algorithms as it provides best accuracy to classify the data.
- It runs smoothly on large datasets.
- It handles large number of input variables.
- It works well on the datasets that are imbalanced in nature.
- It always provide low error rate for classification.
- It also tells the features that are important for the classification purpose.

- It provides interesting methods in identifying the relationships between the variables and classes.

Disadvantages

- Random forests do not work well with the regression technique because in this the ways in which trees are created they are not able to predict beyond the range of the respond value in the training set.

HOW RANDOM FOREST WORKS

- Let N be the cases in the training set T and M be the variables in that training set.
- To create S number of trees you need to create S bootstrap samples by random sampling with replacement. This will result in T=(T1, T2, …..TS) Samples.
- Due to sampling with replacement every sample can have duplicate values. The collection of bootstrap samples is called bagging.
- Create S trees from each of the bootstrap samples that have created in the previous steps.
- At each node of the trees, split the node by randomly selecting m subset of attributes where m<<M. The number of attributes selected at a time is m= √M.
- **OOB error estimate**: When creating samples from original dataset some cases are not left out of the bootstrap sample. They are not used in the construction of the trees. This data is called out of bag data. Then we use these cases and run down on the created trees for predicting their classification. Lastly aggregate the results of these OOB predictions then calculate the error rate which is called OOB estimate error rate[26].

1.2 CLOUD COMPUTING: INTRODUCTION

Cloud computing refers to the process of providing on-demand services over the internet[40]. In this the data is kept on the internet instead of your hard drives. In cloud computing the hardware and software are provided to each individuals and business organizations. These hardware and software are kept on the remote locations and are managed by the third parties.

Since the data is managed by the third parties concept of privacy is major concerns in this. This model allows you to access the resources from anywhere if there is a network connection. Some examples of cloud computing are emails, social networking sites and online file storage.

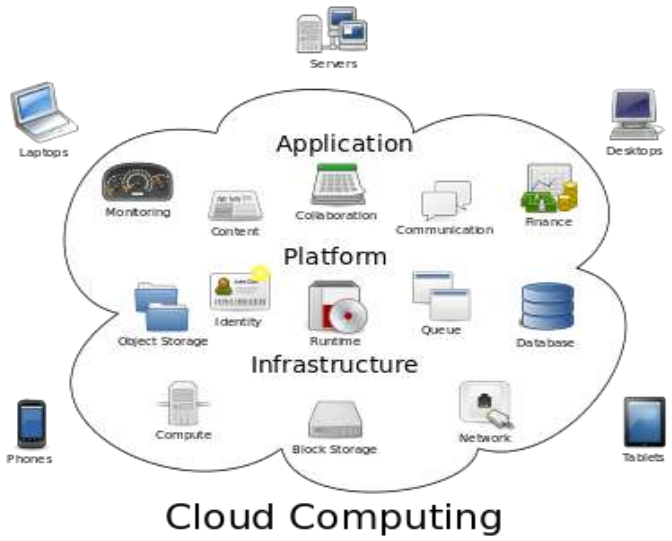


Figure 11: Cloud computing[38]

1.2.1 CHARACTERISTICS OF CLOUD

Five essential characteristics of cloud:

- **On demand self service:** This characteristic means that customer can request the computing resources whenever they needed without any human intervention.
- **Broad network access:** It allows us to access the services via the network or internet from a variety of client platforms such as mobile, laptops.
- **Resource pooling:** A cloud has large and flexible resource pool to meet the customer's requirements. From this pool the resources like compute, storage and network are provided dynamically to the customer. The customers are unaware of the location of the resources[40].
- **Rapid elasticity:** IT resources provided by the cloud can be expand and reduced according to the needs of customers.

- **Metered services:** This service provides the billing information of the resources consumed by the customers.

1.2.2 SERVICE MODELS

There are three service models in cloud:

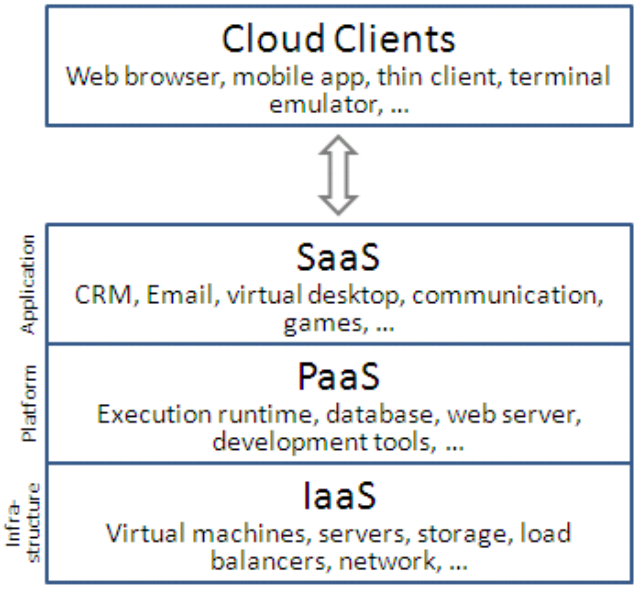


Figure 12: Service models[38]

- **IAAS(Infrastructure as a service):** IAAS is the base layer for the cloud stack which act as a foundation for the above two layers. Only the hardware and network services are provided by the IAAS provider, customer can install any software or operating system.
- **PAAS(Platform as a service):** In this model PAAS provider provides application development environment to the customer[40]. In other words operating system, network and hardware is provided by the provider in this service model.
- **SAAS(Software as a service):** In this model SAAS provider provides the applications running on the cloud infrastructure, along with the operating system, hardware and network.

1.2.3 DEPLOYMENT MODELS

Three deployment models in the cloud:

- **Public cloud:** In public cloud IT resources are made available to everyone (organizations, user) via internet. Examples of public cloud are Google app, Amazon elastic compute cloud.

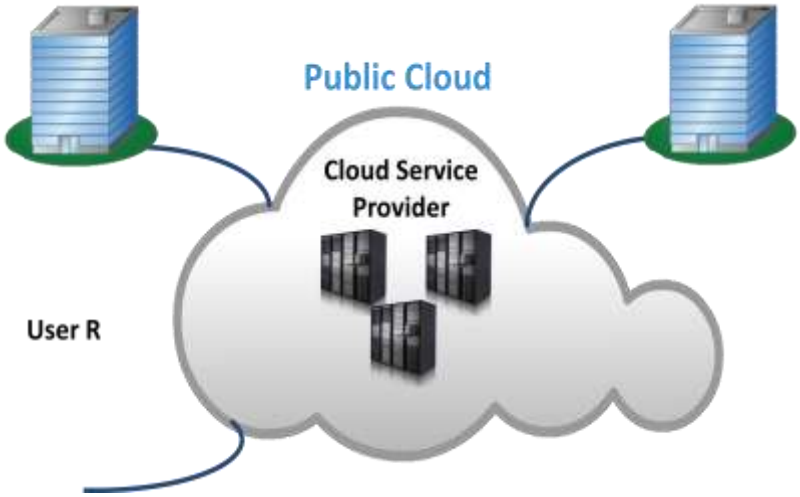


Figure 13: Public cloud[38]

- **Private cloud:** In this model the resources are available to only that one organization for which the cloud infrastructure is operating. Resources are not provided to the users that are outside of the organization. In this cloud works for only one organization. Resources are not available to anyone.

Figure14: Private cloud[38]

- **Hybrid cloud:** This model is a combination of public and private cloud. In this model organization consume the resources from both clouds. The organization uses private model for their normal works but uses public cloud for the works that require high load requirements.

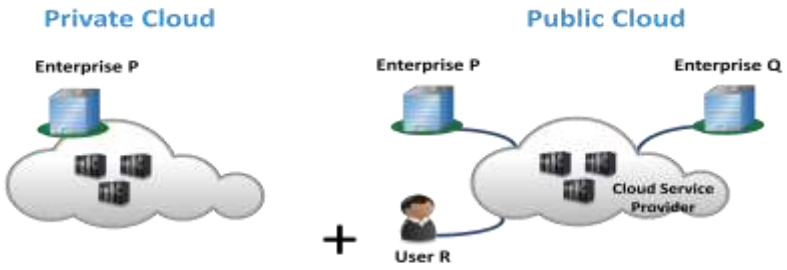


Figure 15: Hybrid cloud[38]

1.2.4 BENEFITS OF CLOUD COMPUTING

- **Reduced IT cost:** Cloud services can be hired on the basis of the needs of the customer. If he wants to use expensive software for performing some task then instead of buying the software he can just hire that software and after finishing the task give the services back to the cloud provider. Also buying large datacenters require cooling, lots of power and management. So organization can just hire these services instead of buying, by this their real estate cost can be minimized.

- **Flexible scaling:** Services provided by cloud are very flexible meaning that the services can be expanded and reduced on the basis of the customers demand. If the customer who have hired some services to do its task is finished before the specified time, then he can give the services back to the cloud provider and the billing charges is only till he have those service. The charges will not be for the same specified time.
- **Less energy consumption:** Many organizations are using cloud services due to which less power is consumed by them because most of the software and hardware that requires power consumptions are just hired by the cloud.
- **High availability:** The applications provided by the cloud providers have higher availability as they have some policies regarding the applications.

1.2.5 CHALLENGES OF CLOUD COMPUTING

- **Security and regulations:** The customer's data are very sensitive so it requires continuous monitoring. When storing the data in cloud it may lose its sensitivity. For example a data is stored by the customer in cloud and the customer do not know at which country his data is stored and may violate some national data protection.
- **Network latency:** The cloud services can be accessed from anywhere in the world. As the data in cloud is stored in the distributed way sometimes it may happen that the resources consumed by the consumer are not close to the consumer location which results in network latency.
- **Supportability:** Some applications may not be supported by cloud. For example if customer wants to access some platform for its application but cloud does not have the compatible operating system for that application.
- **Interoperability:** If consumer wants to move from one cloud service to another cloud service then they may face some complexity and cost issues because of the lack of interoperability of API of different cloud services.

1.3 AES ALGORITHM

AES (Advanced encryption standard) is a symmetric block cipher which uses 128 bit of block size and key size of 128,192 and 256 bits. The number of parameters to be used in this

algorithm depends upon the size of the key. For example if the key size is 128 bit then 10 rounds will be used in this algorithm and if the key size is 192 then 12 rounds and for 256 key size 14 rounds will be used.

It does not used the feistel structure which means that at each round the processing is done on the entire block instead of just on the half of the data block. 128 bit is the most common key size that is used. The input for the encryption and decryption is single 128 bit block. This block is depicted into 4x4 square matrix called in matrix[38]. The block is then copied to state array. The state array is modified at each stages of the algorithm and is finally copied to the output matrix.

Similarly the key is also depicted in form of matrix and is then expanded in each stage using 128 bit key expansion. The algorithm consists of N rounds. The first N-1 round consist of four transformation function and the last round consist of three transformation functions.

The four functions that are used in each round are:

- Substitute byte: In this byte to byte substitution is performed on the block using S-box.
- Shift rows: These are just simple permutation.
- Mix columns: This operation operates on each column individually.
- AddRoundKey: It is a simple bit wise XOR operation performed with the expanded key on the current block.

CHAPTER 2
LITERATURE REVIEW

Ke Sun, Wansheng Miao, Xin Zhang, Ruonan Rao "An improvement to feature selection of Random forest on Spark.", (2014) In this paper they focused on improving the feature selection of random forest and then implementing this strategy on Spark. They are improving the feature selection by eliminating the noisy and redundancy features from the data that is to be analyzed. In this paper they have proposed a three phase method to get the important features during feature selection.

Phase1: In this they are removing those features which are of no importance to their classification problem. T-test is used for binary classification and ANOVA is used for multi classification.

Phase2: In this they are using a recursive feature selection. At each iteration the feature that is of less importance is eliminated until the sizes decreases to a given threshold value.

Phase3: SFS (Sequential forward selection) is applied on the remaining features to create sub-RF of low error value. The sub-RF that leads to smallest low error is selected and then the features used in that sub-RF are selected.

Then lastly they have implemented this strategy on the SPARK which is a fast distributed system for big data[11].

S.Bharathidason, C.Jothi Venkataeswaran "Improving Classification accuracy based on Random Forest model with uncorrelated high performing trees.", (2014) In this paper they are improving the accuracy of Random forest by using the high performance uncorrelated trees in their forest. Firstly, they are evaluating the accuracy of each tree after

that they are selecting the good trees based on the accuracy from the large trees. Then they are making the clusters of the trees based on the correlations. Now trees in intra clusters have high correlation among them and inter cluster trees have low correlation among them. Lastly select high performing tree from each cluster which results in high performance uncorrelated trees.

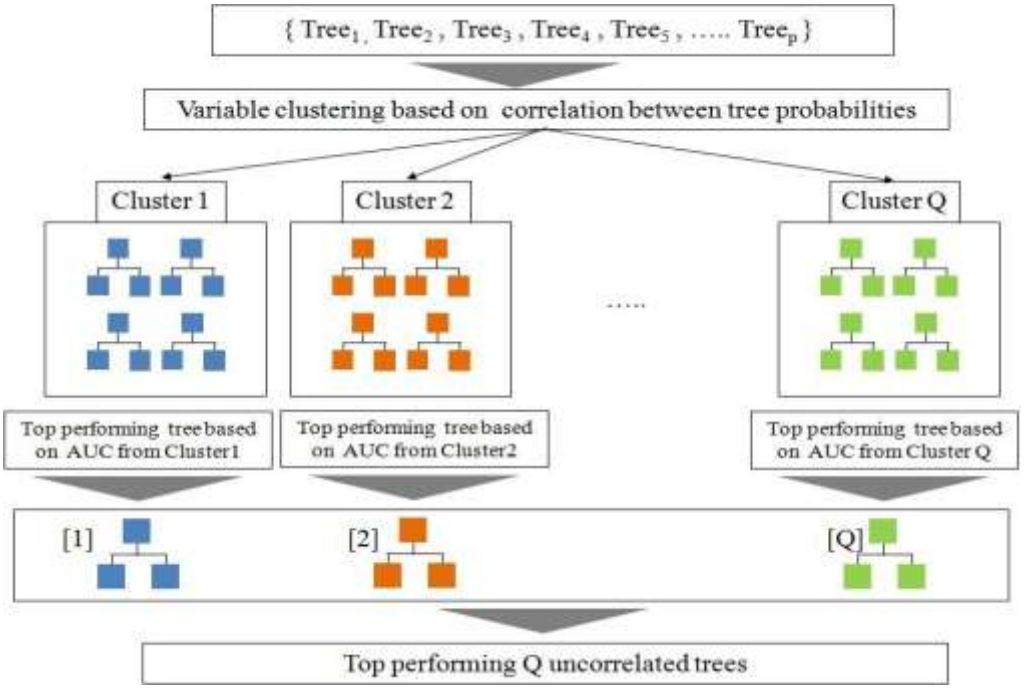


Figure 16: Enhanced Random forest building[15]

They have performed their improved random forest on four datasets Heart disease, Telecom churn, Credit risk factor and bank marketing. Heart disease dataset is collected from the diabetic research institute, Chennai. Telecom churn dataset is collected from the telecom company, Credit risk factor is collected from the MV diabetic lab, Chennai and bank marketing dataset is used from the UCI repository[15].

Vrushali Y Kulkarni, Pradeep K Sinha "Effective learning and classification using Random Forest algorithm." , (2014) This paper proposed five approaches to improve the performance of random forest classifiers based on the accuracy and to reduce the learning time of classification. The five approaches are:

- Disjoint partitioning approach: In this approach the diversity of trees are increased by using the disjoint sets of original dataset. This means that for generating each tree fixed number of disjoint samples are used without replacement which in turn will increase the diversity among individual tree. Also the less correlated attributes are taken to increase the diversity.
- Weighted hybrid decision tree model: In this a hybrid model is integrated with the weighted voting instead of majority voting. In the hybrid model three measures are taken at each node split. These three split measures are Information gain, Gain ratio, and Gini index. Weight of individual tree is calculated on the basis of the OOB error. The tree which has OOB error less than the average OOB error is assigned a higher weight.
- Optimal subset of random forest: In this the optimal subset of random forest is selected using the dynamic programming algorithm. In this approach the problems are solved using the solutions of sub-problem. Each of the sub-problems is independent. The algorithm solves the sub-problem only once and then stores it in the table in order to avoid the re-computing every time the sub-problem is appeared. In this the subsets are created for each datasets and then 10 cross validation is performed on each of the sets. The accuracy results of subsets are then compared with the accuracy results of the 10 cross validation of RF. If the subsets have accuracy greater than the 10 cross validation of RF then it is stored.
- Diversity based dynamic pruning: In this they are using an approach which is dynamic as well as parallel in nature. The idea is to perform pruning in random forest but without affecting the parallelism of tree as well as its accuracy. In this they are creating only those trees that are diverse in nature and to achieve these ranking of bootstrap samples are performed based on the diversity.
- Parallel Random forest: As we know parallelism leads to efficient working, in this approach they are increasing the degree of parallelism by not only generating the trees parallels but also each tree in this approach are parallelized[20].

YIN XiaoHong, DIAO Zhijian "Research and implementation of the data mining algorithm based on cloud platform.", (2014) In this paper they are studying the data

mining algorithms and Hadoop technology. Analyzing the large data using the mining algorithms take long time in processing but by integrating these mining algorithms with the cloud platform the mining time is decreased because of its distributed nature. The parallelism will help in mining the data faster. Hadoop, HDFS, Map-reduce are discussed in this paper and logistic regression of the mining algorithms is also introduced in this paper[24].

Diego Marron, Albert Bifet , Gianmarco De Francisci Morales "Random Forest of very fast decision trees on GPU for mining evolving big data streams." , (2014) In this paper they are providing a method for building a random forest for data streams on GPU using very fast decision trees. They are implementing the very fast decision tree in the GPU thus calling it a GVFDT. The main structure of GVFDT is tree and its main operation is traversing. In this they are identifying the decision tree algorithm execution then increasing them horizontally to increase the degree of parallelism. The random forest algorithm uses per-tree mechanism to identify the changes so that they can react to it. These two methods reduce the communication gap between the CPU and GPU by directly building the tree inside the GPU. Because in the previous study decision tree used in GPU uses a pre-processing in CPU to build the data structure and then passing that data structure to GPU for parallel processing[6].

Florian Baumann, Fangda Li, Arne Ehlers, Bodo Rosenhahn "Thresholding a Random Forest classifier." ,(2014) This paper proposed threshold variant method for improving the accuracy of Random forest. They have introduced this idea by using the AdaBoost and linear combination of weak classifier to create strong classifier. This method eliminates the fluctuate path which results in better accuracy of random forest. First each node is weighted on the basis of uncertainty. The bad node is punished while the good node is weighted high. Next the weight of each tree is summed up which makes more robust decision. By summing up all the weights a threshold point is created. The threshold point will tell the percentage of weights that need to be available for classifying the class[8].

Thnah-Tung Nguyen, Joshua Zhexue Huang, ThuyThi Nguyen "Unbiased feature selection in learning Random forest for High dimension data." ,(2014) In this paper they

have proposed a new improved random forest called xRF on the basis of unbiased feature selection. This algorithm is also best for high dimensional data. Firstly they are removing the uninformative features using the statistical wilcoxon rank-sum test. This test separates the informative features from the uninformative features. After removing the uninformative features the remaining feature set is partitioned into two feature subset. The one subset contains highly informative features and the other one contains weakly informative features. The sampling technique is then used to select the sample features from the two subsets and then merge them to create a new sub space. Since subspace always contains informative features which will produce best split at each node the sampling feature selection will not take biased information which leads to higher accuracy[19].

Sunita, Prachi "Efficient cloud mining using RBAC concept.",(2013) This paper is focused on making the data secure from unauthorized user by using RBAC .RBAC is the role based access control which restricts the user from using the unauthorized data. For RBAC architecture SVM algorithm is used**.** In cloud, security of data is of main importance and access control to multi-tenancy requirements do not scale well because they are mostly based on individual user IDs each having different granularity level. So handling and managing the security of number of users can be hectic. For this purpose RBAC is well suited because users having similar security levels can provide one role. In this the roles are less and users can be classified according to their roles. This paper proposes RBAC method for multi-tenancy architecture in clouds[17].

Vrushali Y Kulkarni, Pradeep K Sinha "Efficient learning of Random Forest Classifier using disjoint partitioning approach.", (2013) This paper proposes an approach which builds the random classifier efficiently and provides better accuracy then the original Random forest. In this they are increasing the accuracy by making trees diverse .The trees are diverse by using the disjoint sets of original dataset. This means that for generating each tree fixed number of disjoint samples are used without replacement which in turn will increase the diversity among individual tree. Less correlated attributes is another measure that they are using to increase the diversity. They have tested this approach on six datasets all of which are collected from the UCI repository. By testing the approach on these dataset they came to

know that this approach works well for those datasets that are imbalanced in nature. They have used weka tool with 10 cross validation[21].

Supraja.Y, T.V.Sai Krishna, P.VenkatasubbaReddy, Dr. M.A.D.Swamy, Dr.P.Srinivasulu "Random Forest machine learning algorithm to detect abnormal behavior in cloud-based mobile services.", (2013) This paper proposed a prototype application which detect the abnormal behavior in mobile-based cloud infrastructure.

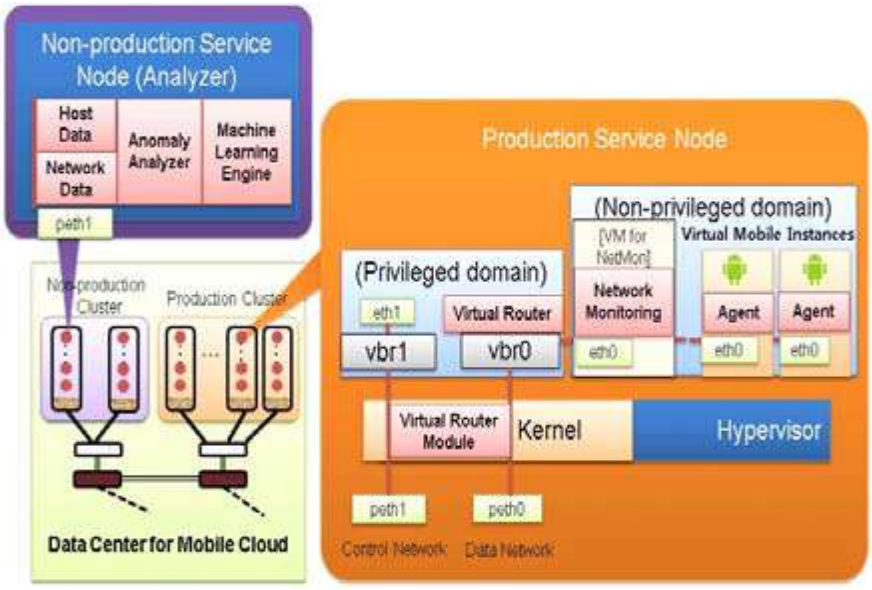


Figure 17: Architecture of mobile cloud infrastructure[18]

In this architecture there is a data center, production service node and non-production service node. In data center there is production and non production cluster. Non-production service node is analyzer which has host, machine learning engine and network data. Production service node has privileged domain which is used for routing and non privileged domain which is used for network monitoring. Hypervisor is used for virtualization concept which is in the production service node. To train the data for abnormal behavior random forest algorithm is used. They classified the data in three states inactive, active and abnormal. In case of inactive no instance is used by the application. When user starts the application the state moves from inactive to active. When a node is affected by some malware then abnormal state will appear[18].

Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena "Machine learning approach for intrusion detection on cloud." ,(2013) This paper proposed an intrusion detection system to detect the intrusion in the virtual machines on cloud. Naïve bayes and hybrid approach of Naïve bayes and Random forest random forest are two popular approaches that are used in their detection system. Naïve Bayes classifier is used for classifying the intrusion connections in intrusion and normal connection labels. Then hybrid approach is used for balancing the datasets and building the detection method by analyzing the patterns from the training set. This system is tested on KDD'99 dataset.

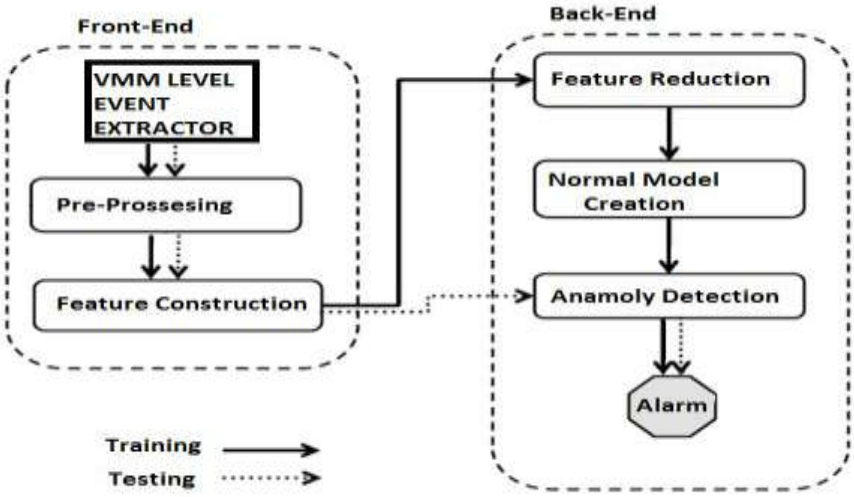


Figure 18: VMM-IDS[1]

In the above figure VMM monitors the events in the virtual machines and extracts them for the feature selection purpose. Then pre-processing component is used to remove the noise from the VMM events. After that in the feature construction they convert the continuous data into nominal data. The output of feature selection provides the set of features. The feature reduction then reduced the feature set for the creation of model. The anomaly detector distinguishes which is the anomalous event and which is the normal event by matching the features with the classes identified by the classifier. If it is an anomalous event then alarm will generate[1].

Kashish Ara Shakil, Mansaf Alam "Data management in cloud based environment using K-Median clustering technique", (2013) In this paper they proposed an approach

which provides efficient management of data in cloud as the growth rate of data is increasing. They are using K-mean clustering technique to manage the data.

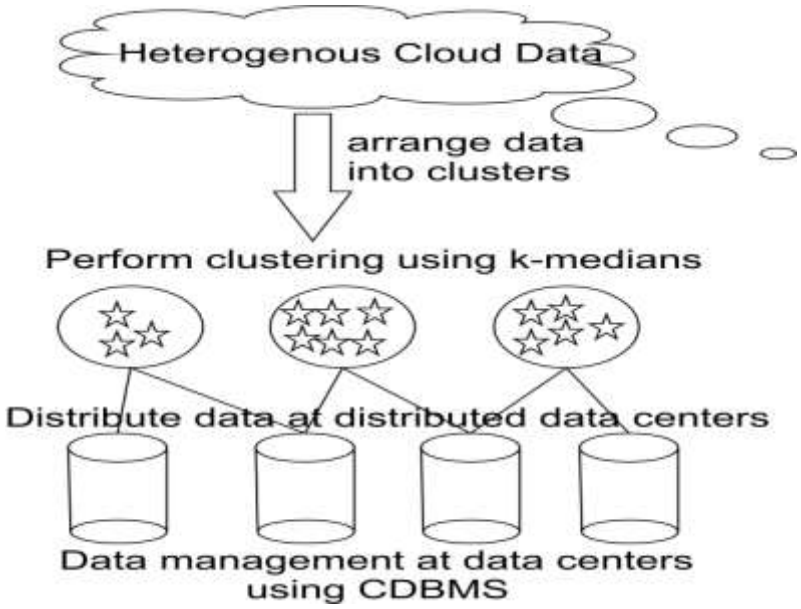


Figure 19: Managing data through clustering in cloud[10]

Firstly they are organizing the cloud data into clusters based on the similarities. For this they are using K-mean clustering. After the cloud data is organized in the clusters they are distributed across various data centers at different geographical locations[10].

Bhagyashri U. Gaikwad, P.P.Halkarnikar "Spam E-mail detection by Random Forest algorithm.", (2013) In this paper they want to detect the spam emails as they increases the load on the server and also increases the bandwidth. In order to automatically detect the spam email they are using the Random forest for classifying the emails. If the classified category is 0 then it is a non spam email but if the category is 1 then email is marked a spam email.

First they are performing pre-processing on the body of the email message. In this pre-processing steps three tasks are performed that is tokenization in which symbols are eliminated, next is stop word they are those words that do not have any information theses words are eliminated in this step and the last task is stemming in which prefixes and postfixes are removed. After performing pre-processing, feature selection is done by using TF selection method. In feature selection only those features that are useful for the classification

are selected. Then random forest is used which will generate the decision trees based on the features. Lastly the voting of each tree is summed up and the class with high votes is selected. If the category is 0 it is considered as a non spam email and if it is 1 then it is a spam email[4].

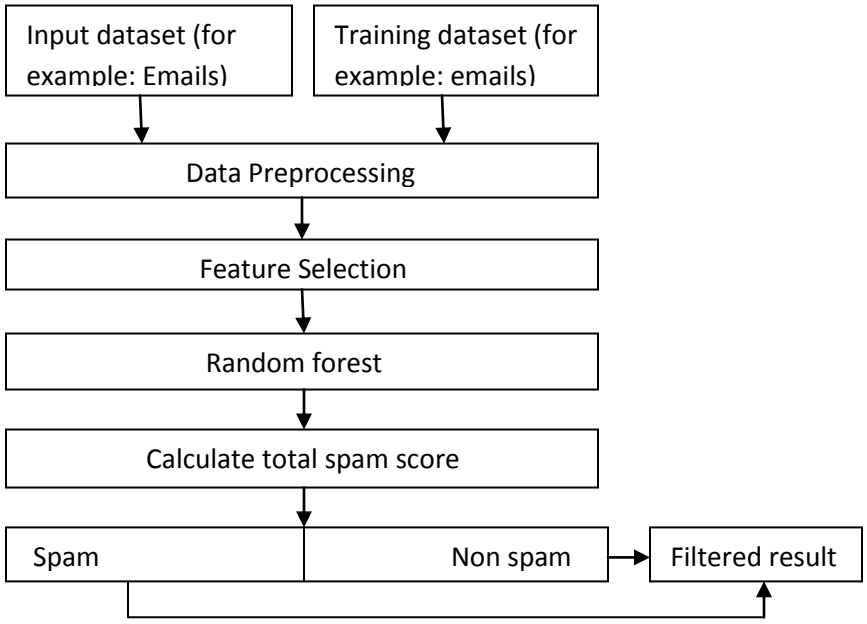


Figure 20: Framework of proposed system[4]

Vrushali Y Kulkarni, Ashu Singh, Pradeep K Sinha "An approach towards optimizing Random forest using Dynamic Programming algorithm.", (2013) In this paper an approach is proposed in which the optimal subset of random forest is selected using the dynamic programming algorithm. Testing the experiment on various UCI dataset they obtain the optimal subset. In this Dynamic programming approach the problems are solved using the solutions of sub-problem. Each of the sub-problems is independent. The algorithm solves the sub-problem only once and then stores it in the table in order to avoid the re-computing every time the sub-problem is appeared. In this the subsets are created for each datasets and then 10 cross validation is performed on each of the sets. The accuracy results of subsets are then compared with the accuracy results of the 10 cross validation of RF. If the subsets have accuracy greater than the 10 cross validation of RF then it is stored[22].

Cuong Nguyen, Yong Wang, Ha Nam Nguyen "Random Forest classifier combined with Feature selection for breast cancer.", (2013) In this they are proposing a diagnostic system by which they can distinguish the benevolent breast cancer from the malignant one by using the random forest machine learning algorithm with the feature selection . In the first phase they are ranking the features based on the Bayesian probability. The ranking is done in ascending order. After that a backward elimination approach is used. In this approach each feature is evaluated to find out which feature is best for using the classification process. In the second phase the random forest is applied on the selected features. It will reduce the prediction time and also provide the better accuracy[5].

Xiao Liu, Mingli Song, Dacheng Tao, Zicheng Liu, Luming Zhang, Chun Chen, Jiajun Bu "Semi supervised node splitting for Random forest construction.", (2013) In this paper they have studied that the node splitting is the major issue in RF as it requires large number of training samples. Existing solutions fail to accurately splitting the nodes and most of the performance bottleneck is at the node splitting. So to overcome this problem they are providing a semi supervised node splitting approach. In this they are splitting the node with the proper guidance of labeled and unlabeled data. They are using unlabeled data because some time labeled data is insufficient. Since the unlabeled data is used for learning purpose it is called semi supervised learning. So they are using the guidance of semi supervised learning for splitting the node. Next they are using a non-parametric algorithm so that the unlabeled data can be properly utilized for the better accuracy at each node splitting. Finally they experimented that their method improves the performance of their random forest[23].

D.L. Gupta, A.K.Malviya , Satyendra Singh "Performance analysis of classification tree learning algorithms.",(2012) In this paper they have used four classification methods to classify the weather nominal dataset. These four classification methods are: J48, Random forest, Reduce error pruning, Logistic model tree. By performing classification on this dataset they came to know that Random forest provides best accuracy and lowest error rate. Weka tool is used in this research work. Each of the algorithms is evaluated using 5 cross validation testing[7].

Jehad Ali, Rehanulla Khan, Nasir Ahmad, Imraan Masqood "Random Forest and decision trees.",(2012) In this paper they have compared the classification results of Random forest and J48 algorithms. They have performed theses algorithms on 20 different datasets. Different classification parameters like precision, recall, F-measure, accuracy are used to compare the classification results. There research find that Random forest algorithm provides best accuracy results and is considered best for large datasets whereas J48 is best for small datasets only[9].

Baoxun Xu, Xiufeng Guo, Yunming Ye, Jiefeng Cheng "An improved Random Forest classifier for Text categorization.",(2012) In this paper an improved Random forest algorithm is used for Text classification. They have used weighted method and tree selection method in their algorithm. In weighted method they are measuring how much informative each feature is and how much it is correlated to its class. The higher weighted feature means that it is informative and is best for using prediction. Previously only two samples T-test method is used for weighting due to which it can weight only two class data. But in this paper they are using chi-square test for solving multi-class problem. By using these two features the subspace size is reduced which increases the performance without increasing the error bound. They have tested their improved algorithm on six different datasets each one having diverse characteristics[3].

Mohamed Bader-El-Den , Mohamed Gaber "GARF: Towards self –optimized Random Forest.", (2012) In this paper a self-optimized random forest is proposed by using the genetic algorithm. They have named this self optimized algorithm as GARF. Self-optimized RF can be capable of dynamically changing the trees if necessary. Multiple RF can be obtained from a large RF. In this approach they are first building a large RF of n decision tree. From this large RF they are building small RF. Within each small RF number of trees are denoted as n_i <N where i=1,2,3….,S and where S is the number of trees in small RF. In genetic algorithm this S is considered population. After that good population is selected this is then used to produce new generation. In such a way they are using the genetic algorithm with the Random forest. After this they compared the results and find out that it gives better performance than the original random forest[14].

Shengqiao Li, E James Harner, Donald A Adjeroh "Random KNN feature selection- A fast and stable alternative to Random forest.", (2011) In this paper they have proposed a RKNN-FS feature selection procedure for the improvement of the Random forest. RKNN-FS is Random K nearest neighbor model in which k-nearest neighbor are the classifier. The classifier in the RKNN-FS model is built from the random selection of the input variable. In this they are ranking the importance of the variables by criteria called support. Then they compare the results of Random forest and RKNN-FS model. After comparing the results they find out that RKNN-FS is east to implement and understand then. Also it is more stable and very much faster than Random Forest[16].

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

Random forest is a supervised machine learning algorithm. It is an ensemble method whose performance depends upon the accuracy, diversity and the correlation. But the randomization used in this can provides us bad features which will lead to the generation of bad trees[9]. So improving the Random forest with best feature selection can lead us to the better performance and accuracy. Also the data used for the algorithms needs to be protected so that no one can do any mining attack on it. To protect the data cloud environment is used. With this only authorized user is able to use the data for the mining purpose.

3.2 OBJECTIVES

Objectives of the research work are:

- Creating virtual environment using cloud simulation.
- Enhancing the security of data using cloud partitioning.
- To implement secure Mining using Random Forest and Enhanced Random forest.
- Evaluate parameters like accuracy, FP rate, TP rate, precision and compare the performance of existing and proposed technique.

3.3 METHODOLOGY

The research work proceeds with the following steps:

1. Collect the dataset from the UCI repository to implement it on the mining algorithm.
2. After the dataset is collected it is partitioned into 3 parts. Also the partitioned data is encrypted with the AES algorithm.

3. These encrypted files are then store in the cloud environment. Single cloud is used in this environment but the cloud is partitioned into three parts each one storing a single encrypted file.
4. Now the data is decrypted so that the mining can be performed.
5. Lastly the results of random forest and improved random forest are compared.

3.3.1 FLOWCHART

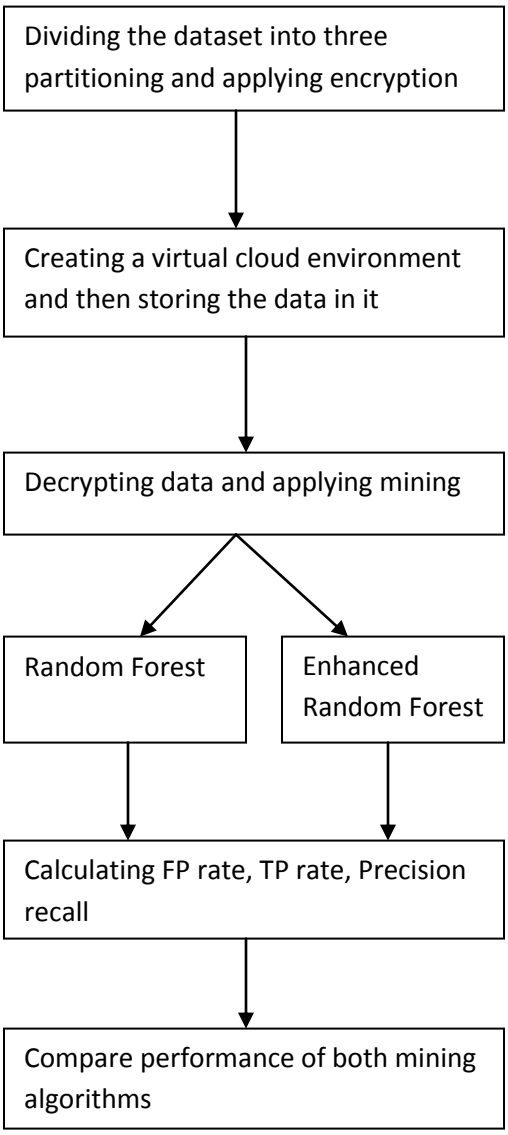


Figure 21: Basic design of proposed method

3.3.1 Generation of dataset

The dataset used for testing the RF algorithm is collected from the UCI repository. We are using ionosphere dataset in which radars returning from ionosphere are classified in good or bad classes.

3.3.2 CLOUD DATA STORAGE

The collected dataset is partitioned into three parts and are then encrypted using the AES algorithm in order to provide security on the data. The data is then stored in the virtual cloud environment. In this single cloud three partitions are created and then three encrypted files are stored in each of these partitions. Only the authorized user can access this data.

3.3.3 RANDOM FOREST ALGORITHM

The data stored in the cloud is then decrypted so that mining task can be performed to mine the data. The Random forests algorithm steps are as followed:

Step1: Create Tn bootstrap samples randomly with replacement from the original dataset T. Let us assume that the number of attributes in this dataset is M.

Step2: From each bootstrap sample generate a tree such that at each node rather than choosing the best split from all the features, select random subset of attributes. The number of attributes to be chosen randomly is m, where $m = \sqrt{M}^{[2]}$.

Step3: Classify the data by aggregating the classification of each tree. Means the class with the majority of the votes will be the classification result.

3.3.4 FEATURE SELECTION

Feature selection is the machine learning technique that selects the relevant feature subset used in the model construction. In classification feature selection is the process of finding optimal subset feature to increase overall accuracy and to decrease the data size so that learning time can be increased. When performing the classification task by feature selection, there will be some features that are irrelevant for our work. So the main goal of feature selection is to extract the feature that is relevant for the classification task.

Consistency based feature selection is one of the category of feature selection. This feature selection selects the feature that yields good results on the basis of their consistency[22]. A feature subset is said to be inconsistent if there exist at least two instances that have same feature values but different class labels. The consistency measured for feature selection is inconsistency rate.

Inconsistency rate = Number of inconsistent examples / Total number of examples

The feature selection process has three steps:

1. **Generation procedure:** In this step the search method is used to create feature subset. In this paper we are using Exhausting search. In exhausting search it looks for the minimal subset whose consistency is equal to the full set of attributes.
2. **Evaluation procedure:** In this they measure the goodness of the feature subset on the basis of the inconsistency rate.
3. **Stopping criteria:** If stopping criteria is not selected then the feature selection process will run unnecessarily longer. So stopping criteria is needed. We can use a predefined iteration or predefined features as stopping criteria[24].

3.3.5 IMPROVED RANDOM FOREST

In improved random forest, the consistency based feature selection is integrated with original random forest. In this, first the consistency based feature selection will select the optimal feature subset. After this we are constructing random forest using the selected features which will reduce the classification time as well as increase the performance.

3.3.6 TOOL

To test the classification model we have use WEKA tool with 10 cross validation. With this tool we are able to measure the accuracy of the mining.

10 Cross validation is a testing method that is used for validating the models. In this dataset is divided into 10 parts out of which 9 folds are used as the training set and 1 fold is used for the testing set. Then the error rate of the model is tested by calculating the average error rate of each of the 10 folds.

CHAPTER 4
RESULTS AND DISCUSSIONS

4.1 PERFORMANCE MEASURE

Following are some performance measures used in comparing original RF with the improved RF:

- **Accuracy:** Accuracy determines the overall correctness of the constructed model. It is defined as the ratio of correctly classified instances to incorrectly classified instances.

  Accuracy= (TP+FP) / (TP+FP+FP+FN)

- **Precision:** Precision is retrieving the relevant records from both relevant and irrelevant records. It is defined as the ratio of positive predicted value to the predicted instances that are relevant.

  Precision= TP / (TP+FP)

- **Recall:** Recall is the fraction of relevant instances that are retrieved. It is defined as the ratio of true predicted values to the true predicted and false predicted values[32].

  Recall= TP / (TP+FN)

- **F-measure:** It is defined as the combination of both precision and recall.

  F-Measure= (2 * Precision * Recall) / (Recall + Precision)

4.2 GRAPHICAL USER INTERFACE OF PROPOSED METHOD

Following are the snapshots of the GUI of how data is stored in the cloud and then the mining that is performed on that data.

1. On running the register java file following window will appear in which user is going to register for performing the cloud mining.

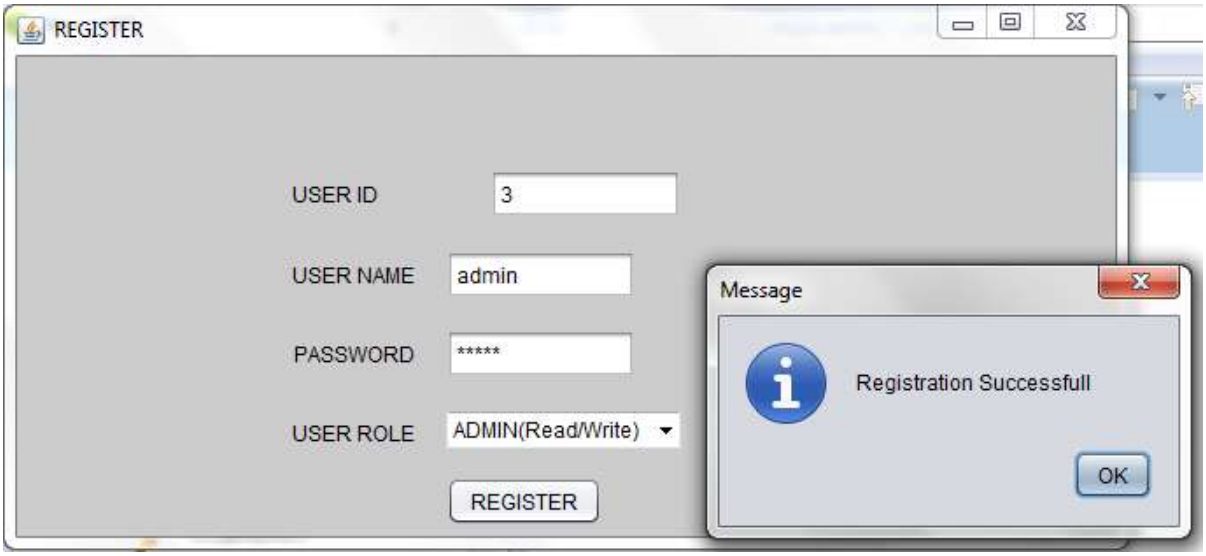


Figure 22: Snapshot of user registration interface

2. Now the register user will login using her username and password and a ticket will be generated to that user. After clicking on ok button a new window will be opened.

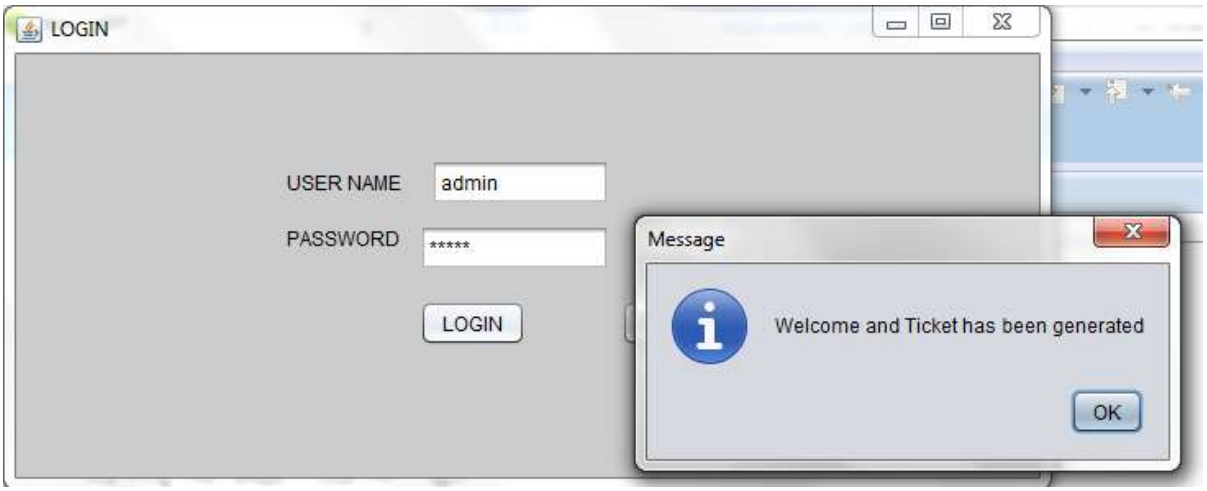


Figure 23: Snapshot of Login interface

3. In this window there are three buttons

- Upload dataset
- Decrypt dataset
- Apply mining

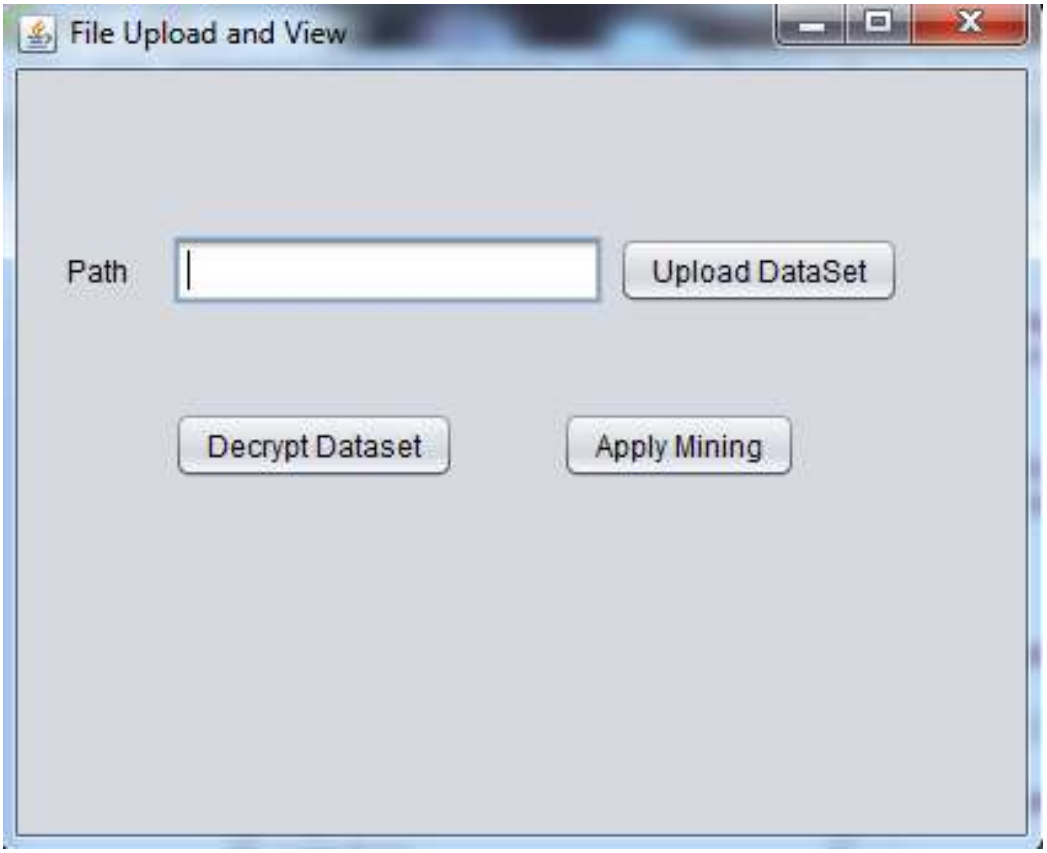


Figure 24: Snapshot of dataset upload

On clicking the upload dataset button a pop up window for selecting the dataset will open. Select the desired dataset and then click open.

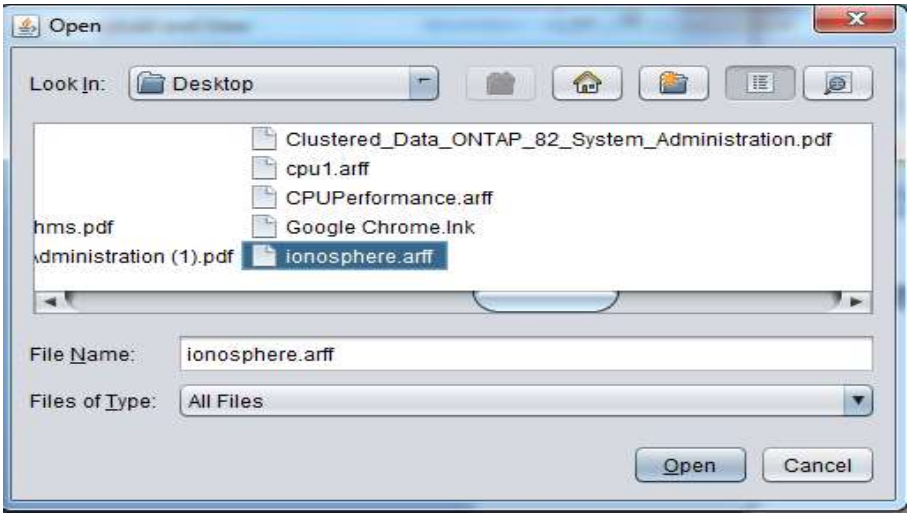


Figure 25: Snapshot for selecting the dataset

4. Then by clicking on the open button of the database pop up window a message will be displayed saying that the data has been migrated to cloud in encrypted form.

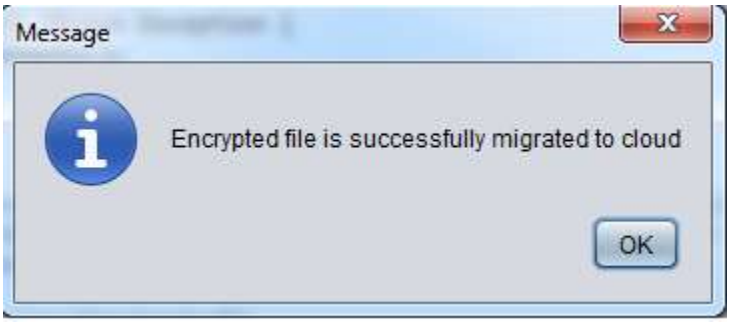


Figure 26: Snapshot of message telling about data storage

The data is stored in the cloud in three partitioning.

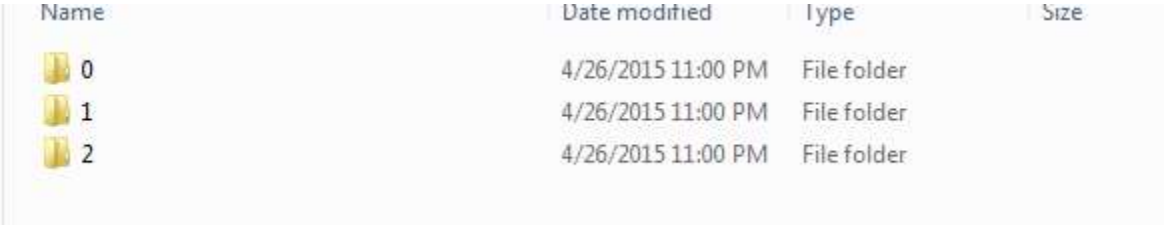


Figure 27: Snapshot of data partitioning on cloud

5. In this window by clicking on the decrypt dataset button the dataset stored into the cloud will be decrypted for the mining purpose and a message that the data that data has been decrypted and is stored in new folder will be displayed is stored.

Figure 28: Snapshot of dataset decryption

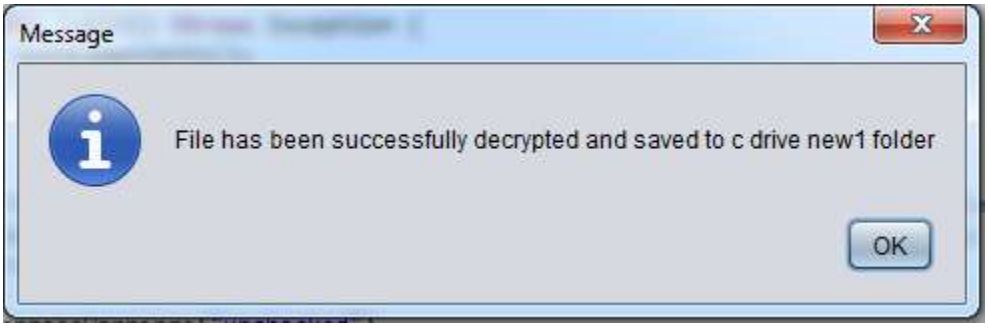


Figure 29: Snapshot of message telling about successful data decryption

6. Now after the data is decrypted, click on the apply mining button and a window to perform classification mining will be appeared.

Figure 30: Snapshot of Data mining

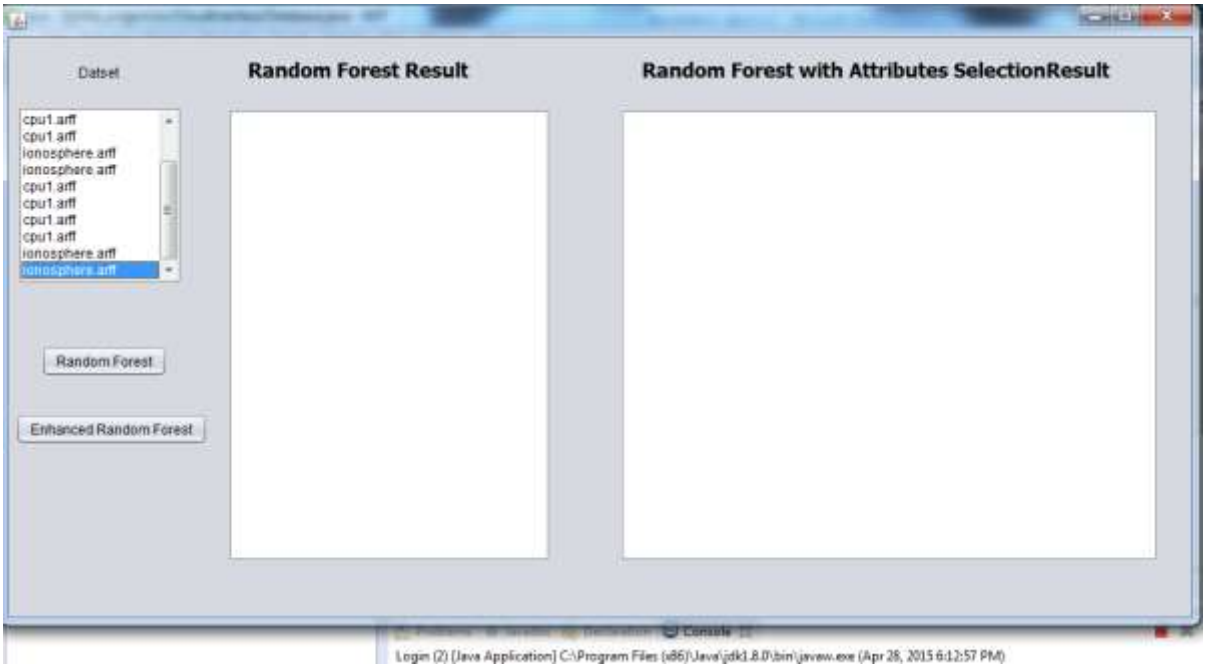


Figure 31: Snapshot of RF and Improved RF Mining interface

In this window there are two buttons named Random forest and enhanced random forest. After selecting the dataset perform mining using both methods.

7. In this the results of both the algorithms are shown. Random forest is showing 92.59% of accuracy and enhanced random forest is showing 98.29%.

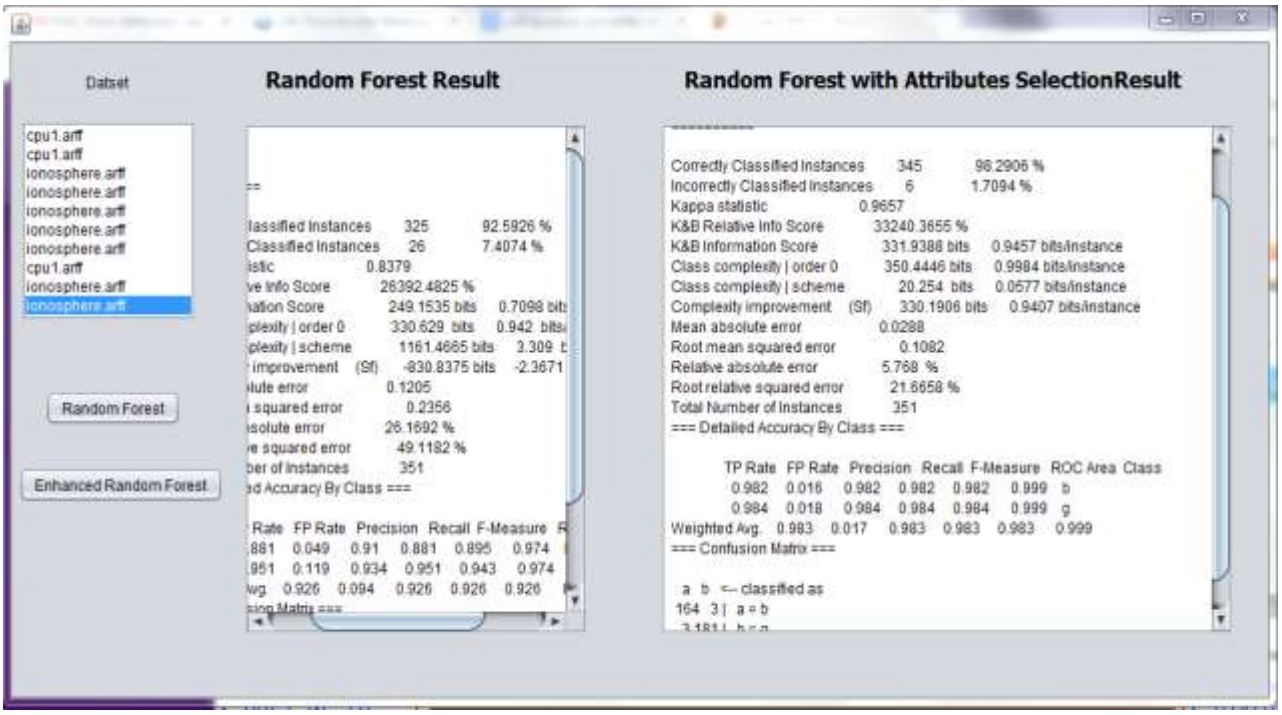


Figure 32:Snapshot of RF and improved RF result

4.3 EXPERIMENTAL EVALUATION

GRAPHS

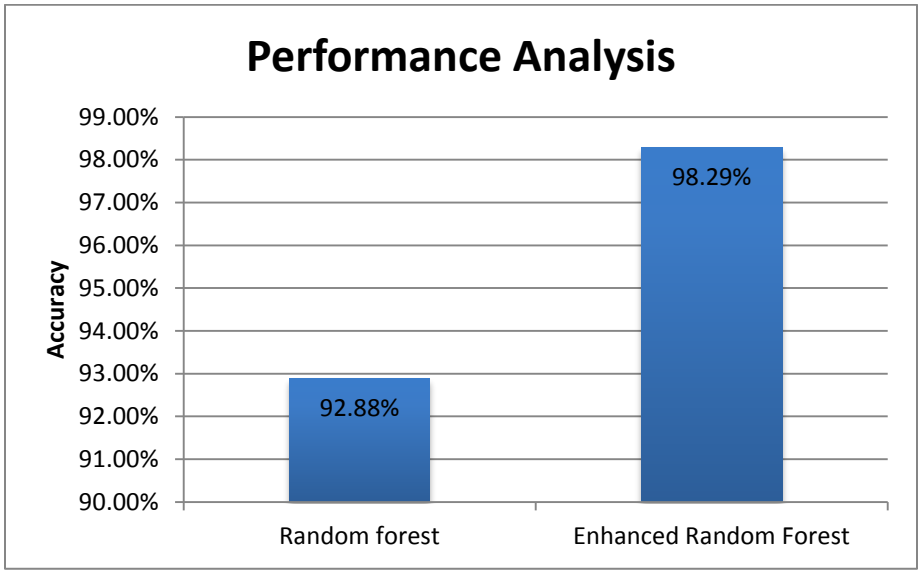


Figure 33: Comparison of performance analysis in terms of accuracy

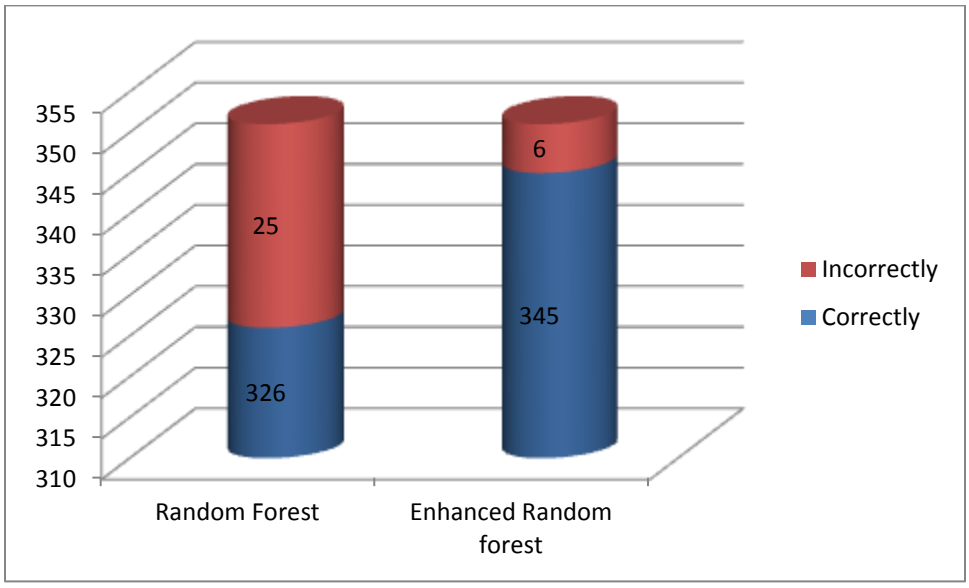


Figure 34: Comparison of correctly and incorrectly classified instances

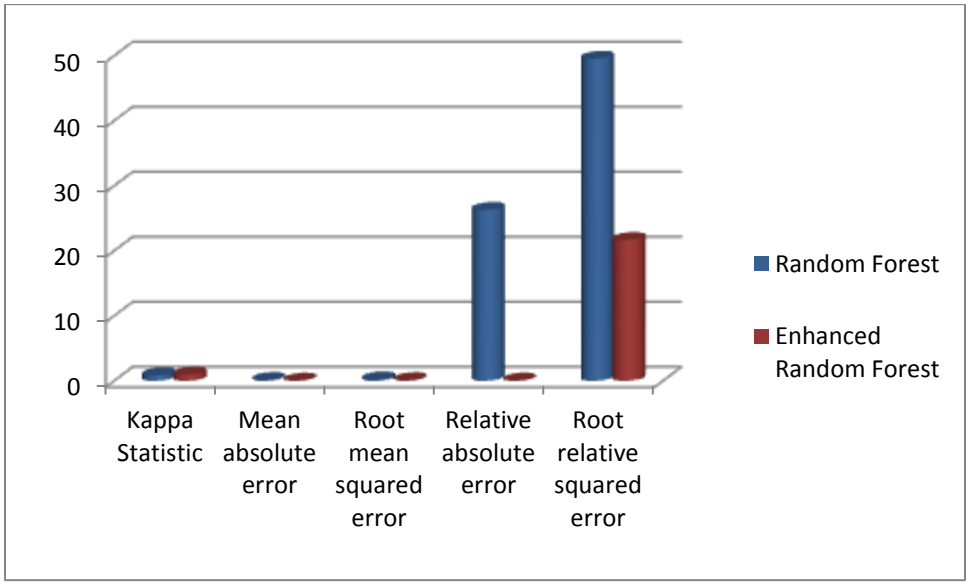


Figure 35: Comparison of various parameters

TABLES

Known Class	Predicted Class: Positive	Predicted Class: negative
Positive	111	15
Negative	10	215

Table 1: Confusion matrix of Random Forest

Known Class	Predicted Class: Positive	Predicted Class: negative
Positive	164	3
Negative	3	181

Table 2: Confusion matrix of Enhanced Random Forest

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area
positive	0.811	0.044	0.917	0.881	0.899	0.973
negative	0.956	0.119	0.935	0.956	0.945	0.973
Weighted Avg.	0.929	0.092	0.929	0.929	0.928	0.973

Table 3: Detailed accuracy by Random Forest

	TP rate	FP rate	Precision	Recall	F-measure	ROC Area
positive	0.982	0.016	0.982	0.982	0.982	0.999
negative	0.984	0.018	0.984	0.984	0.984	0.999
Weighted Avg.	0.983	0.017	0.983	0.983	0.983	0.999

Table 4: Detailed accuracy of Enhanced Random Forest

	Random Forest	Enhanced Random
Kappa Statistic	0.8439	0.9657
Mean absolute error	0.1211	0.0288
Root mean squared error	0.2377	0.1082
Relative absolute error	26.293	5.768
Root relative squared error	49.5448	21.6658

Table 5: Comparison of various parameters

	Random Forest	Enhanced Random
Correctly classified instances	326	345
Incorrectly classified instances	25	6

Table 6: Correctly and incorrectly instances

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

In this paper we have done data mining on the cloud based data. We have used a classification algorithm named Random forest and then also improve that algorithm by improving the feature selection of the RF. To protect our data we partitioned the data and then apply AES encryption on the data so if even some portion of the data is leaked it can't be understand by the intruder. To improve the feature selection of RF algorithm first we apply consistency on the features which will give us the optimal subsets of relevant feature after that we integrate this feature selection with the RF algorithm. After comparing the RF results on the cloud based data we find out that the accuracy of our improved RF is better than the original RF.

The future work can be done in improving the security of cloud data. Also now every other organizations our moving to cloud so further mining can be performed in the cloud based data.

CHAPTER 5
REFERENCES

I Research Papers

[1] Amjad Hussain Bhat, Sabyasachi Patra, Dr. Debasish Jena, (2013) "Machine learning approach for intrusion detection on cloud", International Journal of Innovation in Engineering & Management, Volume 2

[2] Andy Liaw and Matthew Wiener,(2002) "Classification and regression by random forest"

[3] Baoxun Xu, Xiufeng Guo, Yunming Ye, Jiefeng Cheng (2012)"An improved Random Forest classifier for Text categorization.", Journals of Computers, Volume 7

[4] Bhagyashri U. Gaikwad, P.P.Halkarnikar ,(2013)"Spam E-mail detection by Random Forest algorithm.", Computer Science & Technology, Department of Technology, Shivaji University, Kolhapur, Maharashtra, India.

[5] Cuong Nguyen1, Yong Wang1, Ha Nam Nguyen,(2013)"Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic." School of Business and Administration, Chongqing University, Chongqing, China ,College of Technology, Vietnam National University, Hanoi, Vietnam.

[6] Diego Marron, Albert Bifet , Gianmarco De Francisci Morales ,(2014)"Random Forest of very fast decision trees on GPU for mining evolving big data streams."

[7] D.L. Gupta, A.K.Malviya , Satyendra Singh (2012) "Performance analysis of classification tree learning algorithms.", International Journal of computer applications, Volume 55

[8] Florian Baumann, Fangda Li, Arne Ehlers, Bodo Rosenhahn ,(2014) "Thresholding a Random Forest classifier."

[9] Jehad Ali, Rehanulla Khan, Nasir Ahmad, Imraan Masqood(2012) "Random Forest and decision trees.", Computer Systems Engineering, International Journal of Computer Science Issues, volume 9.

[10] Kashish Ara Shakil, Mansaf Alam,(2013) "Data management in cloud based environment using K-Median clustering technique", International journal of Computer applications

[11] Ke Sun, Wansheng Miao, Xin Zhang, Ruonan Rao, (2014) "An improvement to feature selection of Random forest on Spark", IEEE 17th International conference on Computer Science and Engineering.

[12] Kilho Shin, Danny Fernandes and Seiya miyazaki, (2011) "Consistency measures for feature selection: A Formal Definition, relative sensitivity comparison and a fast algorithm" International joint conference on Artificial intelligence

[13] Manoranjan Dash, Huan Liu,(2003) " Consistency based search in Feature selection" Elsevier Computer science

[14]Mohamed Bader-El-Den , Mohamed Gaber,(2012) "GARF: Towards self –optimized Random Forest.", School of Computing,

[15] S.Bharathidason, C.Jothi Venkataeswaran, (2014) "Improving Classification accuracy based on Random Forest model with uncorrelated high performing trees.", International Journal of Computer Applications Volume 101- No.13, September 2014

[16] Shengqiao Li, E James Harner, Donald A Adjeroh ,(2011)"Random KNN feature selection- A fast and stable alternative to Random forest.", BMC bioinformetics

[17] Sunita, Prachi, (2013) "Efficient cloud mining using RBAC concept", International Journal of Advanced Research in Computer Science and Software Engineering.

[18] Supraja.Y, T.V.Sai Krishna, P.VenkatasubbaReddy, Dr. M.A.D.Swamy, Dr.P.Srinivasulu,(2013) "Random Forest machine learning algorithm to detect abnormal behavior in cloud-based mobile services." , International Journal of Computer Science and Information Technology and Security, Vol 3.

[19] Thanh-Tung Nguyen, Joshua Zhexue Huang, and Thuy Thi Nguyen,(2014) "Unbiased Feature Selection in Learning Random Forests for High – Dimensional Data", Hindwai Publishing corporation The scientific world journal

[20] Vrushali Y Kulkarni, Pradeep K Sinha, (2014) "Effective learning and classification using Random Forest algorithm", International Journal of Engineering and Innovative Technology(IJEIT) Volume 3, Issue11, May 2014.

[21] Vrushali Y Kulkarni, Pradeep K Sinha,(2013) "Efficient learning of Random Forest Classifier using disjoint partitioning approach." Proceeding of the world Congress on Engineering, London , U.K

[22] Vrushali Y Kulkarni, Aashu Singh, Pradeep K Sinha,(2013) "An Approach towards Optimizing Random Forest using Dynamic Programming Algorithm," International Journal of computer applications ,Volume75

[23] Xiao Liu, Mingli Song, Dacheng Tao ,Zicheng Liu,Chun Chen and Jiajun Bu."Semi-Supervised Node Splitting for Random Forest Construction." IEEE

[24] YIN XiaoHong, DIAO Zhijian, (2014) "Research and implementation of the data mining algorithm based on cloud platform", 2014 IEEE workshop on Electronics, Computer and applications.

II Website Links

[25] http://dissertation.laerd.com/simple-random-sampling.php

[26]http://faculty.washington.edu/fxia/courses/LING572/bagging.ppt

[27] https://github.com/ironman/Random-Forest

[28] http://in.mathworks.com/help/stats/classification-trees-and-regression-trees.html

[29]http://stat.psu.edu/~jiali/course/stat597e/notes2/bagging.pdf

[20] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/

[31] https://onlinecourses.science.psu.edu/stat857/node/181

[32] http://www.compumine.com/web/public/newsletter/20071/precision-recall

[33]http://www.comp.nus.edu.sg/~wongls/talks/pkdd04/2004-09-ECMLPKDD-tutorial-part2.ppt

[34]http://www.cs.nyu.edu/courses/spring05/G22.3033-010/7class.ppt classifier

[35] http://www.saedsayad.com/decision_tree.htm

[36] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#intro

[37] http://www.tutorialspoint.com/data_mining/dm_dti.htm

[38] http://www.wikipedia.com

[39] http://www.webopedia.com

III Books

[40] William Stalling, Cryptography and Network Security, 5th edition