RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Classification Algorithms Used for Disease Prediction in Data Mining

Amit Tate[1], Bajrangsingh Rajpurohit[2], Jayanand Pawar[3], Ujwala Gavhane[4]

[1,2,3,4] Computer Engineering (Bachelor of Engineering), MESCOE PUNE-411001, INDIA

Gopal B. Deshmukh[5]

Department of Computer Engineering, MESCOE PUNE-411001, INDIA

Abstract:

Classification of data is a data mining technique based on machine learning is used to classification of each item set in as a set of dataset into a set of predefined labelled as classes or groups. Classification is tasks for different application such as text classification, image classification, class's predictions, data Classification etc. In this paper, we presenting the major classification techniques used for prediction of classes using supervised learning dataset. Several major types of classification method including Random Forest, Naive Bayes, Support Vector Machine (SVM) techniques. The goal of this review paper is to provide a review, accuracy and comparative between different classification techniques in data mining.

Keywords — **ML, Classification, Naive Bayes, (Support Vector Machine) SVM, Random Forest classifier (RF).**

I. INTRODUCTION

Disease or medical is based on doctor experience and patients database which people's survived from disease. Medical Misdiagnoses is a serious risk to our healthcare environment. If misdiagnosis is continue, then people will fear going to the hospital for treatment.

Disease prediction mean the some number of symptoms are selected for processing and using this symptoms as input we can predict the disease with help of many kind of classification algorithms.

In this paper we used Weka Explorer version 3.6.13 for prediction of diabetes disease using input training dataset. Analysed all experimental result among time taken, recall.

Data classification is divided into two process, including a learning step and a classification step. Learning step constructing the classification model and classification step where the dataset is used to predict the class labels for predefined dataset. In the learning step, a classification algorithm is construct defining a programmed set of data classes. This is the learning step (or training phase), Classification algorithm builds the classifier by analysing or "learning from" a training set dataset or database tuples and their related class labels.

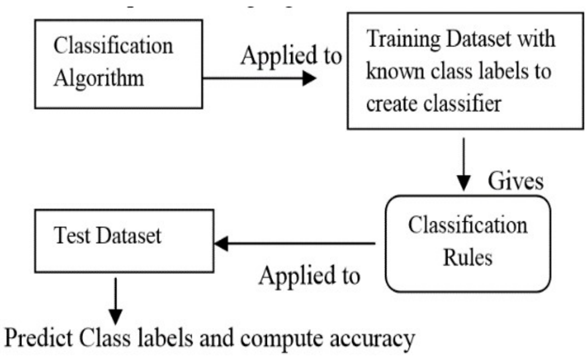
Classification Algorithm
Applied to
Training Dataset with known class labels to create classifier
Gives
Classification Rules
Applied to
Test Dataset
Predict Class labels and compute accuracy

Fig 1. Classification process

In the second step of Classification: Test data from given dataset are used to compute the accuracy of the classification procedures. If the correctness is predicted and acceptable. Then this procedures can be applied to the classification of new dataset in databases. [9]

In the classification process classification algorithms is directly applied to training dataset with known class labels. Those classification algorithms gives the classification rules and rules are applied to test dataset. Finally, Class labels and compute accuracy will predicted.

II. RELATED WORK

Disease prediction using Random Forest is very easy to implement. Powerful algorithms for classification and regression.

Author described in this paper major kinds of algorithms such as Support Vector Machine, K-nearest neighbour and Neural network.[1]

Comparative study of algorithms used for classification in web usage mining described in this paper. Pattern discovery and analysis is identified in Web usage mining using classification algorithms. [2]

In this paper author proposed dynamic characterization for disease diagnosis using ensemble classifier. In this paper author proved how Random Forest algorithms is powerful and best for classification. [3]

This paper involve survey of text mining using classification techniques and also proposed experimental result with error rate of classification for given dataset. [4]

Author introduced supervised machine learning techniques and analysed experimental result of accuracy. Tree based algorithm is best as compare to other classification is proved by author. [7]

Survey and future research work for random forest classifier. Author mentioned improvement required for Random Forest algorithm based on their performance. [8]

This author introduced survey on kinds of machine learning techniques in data mining. Paper include decision tree based algorithm and comparison among the classification techniques. [10]

Author proposed random forest algorithm in cloud environment for big data. Different Algorithms is optimized using dimension reduction and weighted vote approach. [11]

In this paper author described comparative study of classification algorithm. Classification algorithms used for rainfall forecasting and predicted result in form of features. [13]

Disease diagnosis system is proposed using random forest algorithm with help of dynamic characterization. Medical images used for classification as input training dataset, then accuracy, precision and recall is computed and result is predicted. [14]

This author introduced prediction for heart disease using data mining classification algorithms. In this paper author described Random Forest algorithm used for heart disease prediction. [15]

III. MACHINE LEARNING

Machine learning indicates how computers can learn or improve their performance using data. Computer programs to automatically learn to identify patterns and make intelligent based decisions on data. Machine learning is a fast-growing discipline. Here, we illustrate classic problems in machine learning that are highly related to data mining. [9]

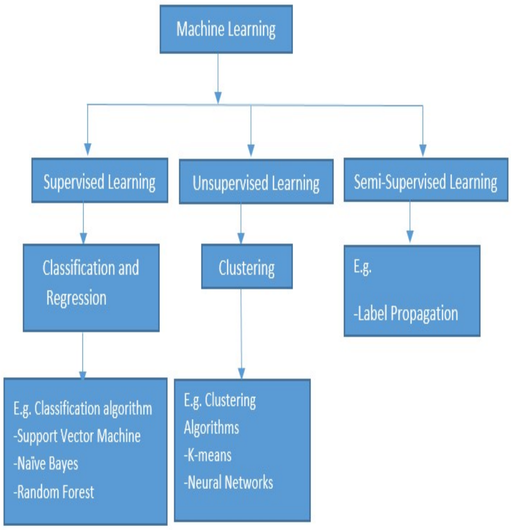

Fig 2. Classification of Machine Learning concepts

There are three main types of machine learning concept as following:

A. Supervised learning: Supervised learning consist of all data is labelled and algorithm learn to predict the output from training dataset.
E.g.: Support Vector Machine (SVM), Random Forest, Naive Bayes. [9]

B. Unsupervised learning: Unsupervised learning is used for clustering based algorithm. In this learning all the data is unlabelled and algorithm learn to essential structure from the input dataset. We can use clustering to discover classes within the dataset. E.g. K-means, KNN. Neural Networks [9]

C. Semi-supervised: Semi-supervised learning is a combination of supervised learning and unsupervised learning. In Semi-supervised learning some data is labelled and some data is not labelled. In this approach, labelled training dataset are used to learn class models and unlabelled training dataset are used to define the boundaries between classes. [9]

VI. CLASSIFICATION ALGORITHMS

A. Support Vector Machine (SVM):

SVM is one of the technique from supervised learning based algorithm which is used for classification and regression analysis. This algorithm is used for classification using training dataset. In Support vector Machine algorithm, it will design each data item set as a point in n-dimensional space. \In this space n is used for number of features in] training dataset and with the value of each feature being the value of a specific coordinate. [6]

Then, we achieve classification by finding and constructing the hyper-plane on dataset that divides the dataset into two classes. Support Vectors are basically the co-ordinates of individual reflection. Support Vector Machine is a border which best segregates the two classes. [6]

This algorithm is classified into linear data and non- linear data. Linear classification is implemented using hyperplane. Non-linear classification some kinds of transformation to given training dataset and then after transformation various methods are trying to use linear classification for separation.

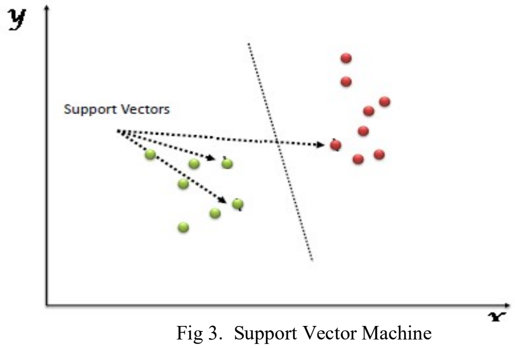

Fig 3. Support Vector Machine

There are 2 key implementations of SVM technique that are mathematical programming and kernel function. Hyper plane separates those data point of different classes in a high dimensional space. Support vector classifier (SVC) searching hyper plane. But SVC is outlined so kernel functions are introduced in order to non-line on decision surface. [10].

B. Random Forest (RF):

Random Forest algorithm are an ensemble supervised learning method which is used as predictor of data for classification and regression. In the classification process algorithm build a number of decision trees at training time and construct the class that is the mode of the classes output by using each single trees. (Random Forests is introduced by Leo Breiman and Adele Cutler for an ensemble of decision trees). [5]

Random Forest algorithm is a grouping of tree predictors where each tree based on the values of a random vector experimented independently with the equal distribution for all trees in the forest. The basic principle is that a group of “weak learners” can come together to form a “strong learner”. Random Forests are a perfect tool for making predictions considering they do not over fit. Presenting the accurate kind of randomness makes them accurate classifiers and regression. [5]

Single decision trees often have high variance or high bias. Random Forests trying to moderate the high variance problems and high bias by averaging to find a natural balance between the two extremes.

Considering that Random Forests have few parameters to tune and can be used simply with default parameter settings, they are a simple tool to use without having a model or to produce a reasonable model fast and efficiently. [5]

Random Forests are easy to learn and use for both professionals and lay people - with little research and programming required and may be used by folks without a strong statistical background. Simply put, you can safely make more accurate predictions without most basic mistakes common to other methods.

Random Forests produces several classification for given trees. Each tree is grown as follows:

1. If number of circumstances in the training data set is D, sample D cases at random state but with replacement, from the original dataset. This sample testing set will be the training set for increasing the tree.
2. If there are I_n input variables from training dataset, a number I_n is indicated such that at each node of the tree, m variables are selected at random available for the I_n and the best splitting on these I_n is used to splitting the node. The value of I_n is used as constant during entire the forest growing.
3. Each tree is grown to the largest size as possible. There is no pruning an overall grownup tree.

The random forest algorithm is an ensemble classifier algorithm based on the decision tree model. It generates k different training data subsets from an original dataset using a bootstrap sampling approach, and then, k decision trees are built by training these subsets. A random forest is finally constructed from these decision trees. Each sample of the testing dataset is predicted by all decision trees, and the final classification result is returned depending on the votes of these trees. [11] The original training dataset is formalized as

$$S = \{(ai, bj), i = 1, 2... D; j = 1, 2... I_n\},$$

Where a is a sample and b is a feature variable of S. Namely, the original training dataset contains D samples, and there are I_n feature variables in each sample.

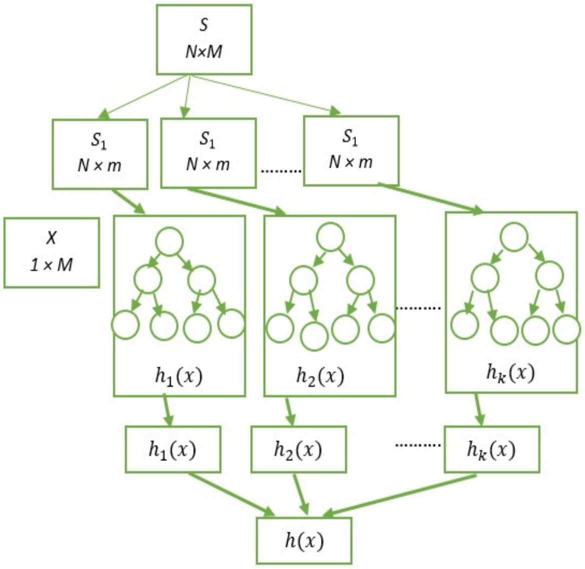

Fig 4. Process of the construction of the RF algorithm

The main process of the construction of the RF algorithm is presented in Fig. 4[11]

The steps of the construction of the random forest algorithm are as follows:

Step 1. Sampling k training subsets.

In the first step, k training datasets are experimented from the original training dataset S in a bootstrap selection manner. Namely, N records are selected from S by a random sampling and replacement method in each sampling time. After the current step, k training subsets are constructed as a collection of training subsets S Train:

S Train = {S1, S2,...,Sk}.

At the same time, the records that are not to be selected in each sampling period are composed as an Out-Of-Bag (OOB) dataset.

In this way, k OOB sets are constructed as a collection of SOOB:

SOOB = {OOB1, OOB2... OOBk},

Where k ≪ N, Si ∩ OOBi = ϕ and Si ∪ OOBi = S.

To obtain the classification accuracy of each tree model, these OOB sets are used as testing sets after the training process. [11]

Step 2. Constructing each decision tree model.

In an RF model, each Meta decision tree is created by a C4.5 or CART algorithm from each training subset Si. In the growth process of each tree, m feature variables of dataset Si are randomly selected from M variables. In each tree node's dividing process is done, then gain ratio of each feature variable is computed, and the best one or most priority node is chosen as the splitting node. This splitting process is repeated until a leaf node is generated. Finally, k decision trees are trained from k training subsets in the same way. [11]

Step 3. Collecting k trees into an RF model.

The k trained trees are collected into an RF model, which is defined in Eq. (1):

$$H(X, \Theta\text{j}) = \sum_{i=1}^{k} hi(x, \Theta\text{j}), (\text{j} \; = \; 1{,}2, \ldots, \text{m}),$$

where hi(x,Θj) is a meta decision tree classifier, X are the input feature vectors of the training dataset, and Θj is an independent and identically distributed random vector that determines the growth process of the tree. [11]

C. Naive Bayes

Naive Bayes algorithm is technique based on Bayes Theorem used for classification with an assumption of independence between class predictors. Naive Bayes classification algorithm assumes that the presence of an exact feature in a class is dissimilar to the presence of any other feature. [12]

Naive Bayes algorithm model is easy to implement and useful for large datasets. Naive Bayes is known to outclass even highly cultured classification algorithm. [12]

Bayes theorem describes a way of calculating posterior probability $P\,(C_x|X)$ from P(C), P(X) and $P\,(X|C_x)$. See the following equation.

$$P(C_x|X) = \frac{\text{P}(\text{x}|c_x)\text{P}(C_x)}{\text{P}(\text{X})}$$

$$P\,(C_x|X) = P(X_1|C_x) \times P(X_2|C_x) \times \ldots \times P(X_n|C_x) \times P(C_x)$$

Above,

- $P\,(C_x|X)$ is a posterior probability of *class* (C_x, *target*) given *predictor* (x, *attributes*).
- *P*(*C*) is called *class* prior probability.
- $P\,(X|C_x)$ is the likelihood which indicate the *predictor* probability of given *class*.
- *P*(*X*) is called prior probability of the *predictor*.

Following example are explaining implementation of Naive Bayes algorithms

E.g. Weather dataset having attributes Weather- Sunny, Overcast, and Rainy. Using training dataset Naïve Bayes algorithms will predict the value as you can play or not.

*T*able 1. Training dataset of Weather

Weather	**Play**
Sunny	Yes
Overcast	Yes
Overcast	Yes
Sunny	No
Rainy	No
Sunny	No
Overcast	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Sunny	Yes
Rainy	Yes
Overcast	Yes
Rainy	Yes
Rainy	No
Sunny	Yes
Overcast	Yes

Table 2. Frequency data of Weather

Frequency Table		
Weather \ Play	**Yes**	No
Rainy	3	2
Sunny	4	2
Overcast	6	0
Total	13	4

Now, using Naive Bayesian equation we can compute the posterior probability for each class. The class which has highest posterior probability is the result of prediction.

P (Yes | Sunny) = P (Sunny | Yes) * P (Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 4/13 = 0.30, P (Sunny) = 6/17 = 0.35, P (Yes) = 13/17 = 0.76

Now, P (Yes | Sunny) = 0.30 * 0.76 / 0.35 = 0.65, which has higher probability.

V. ADVANTAGES AND APPLICATION AREA OF ALGORITHMS: [7]

Table 1. Advantages and application of classification algorithms

Algorithms	Advantages	Application area
Naive Bayes	Ability to interpret problem in terms of structural relationship among predictors, takes less computational time for training, no free parameters to be set	Document classification, medical diagnostic systems
Random forest	Fast, scalable, robust to noise, does not over fit, offer explanation and visualization of its output without any parameters to worry about	To find cluster of patients. Classification of microarray data, object detection
Support Vector Machine (SVM)	High accuracy, avoid over fitting, flexible selection of kernels for nonlinearity, accuracy and performance are independent of number of features, good generalization ability	Text Classification

6. ANALYSIS OF ALGORITHMS

We have classified Diabetes demo of dataset using Weka Explorer version 3.6.13 and we have seen the following results:

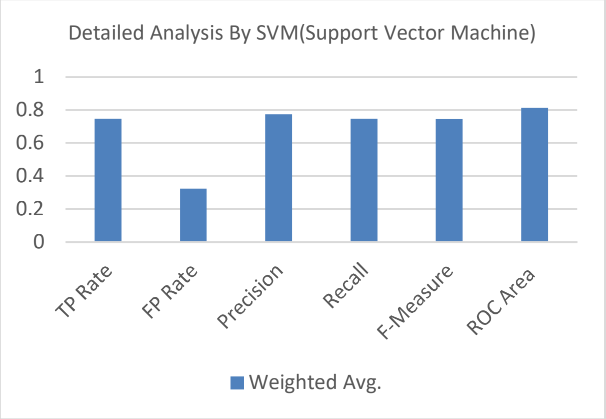

Fig 5. Detailed analysis of Support Vector Machine algorithm

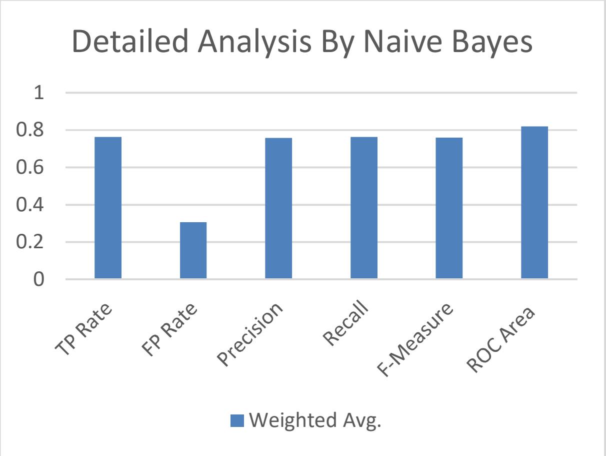

Fig 6. Detailed analysis of Naïve Bayes algorithm

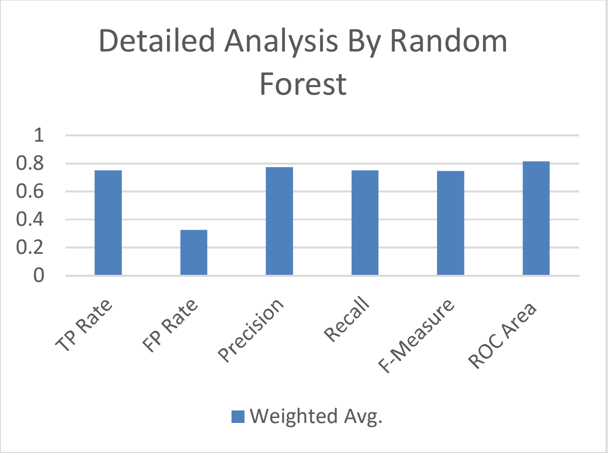

Fig 7. Detailed analysis of Random Forest

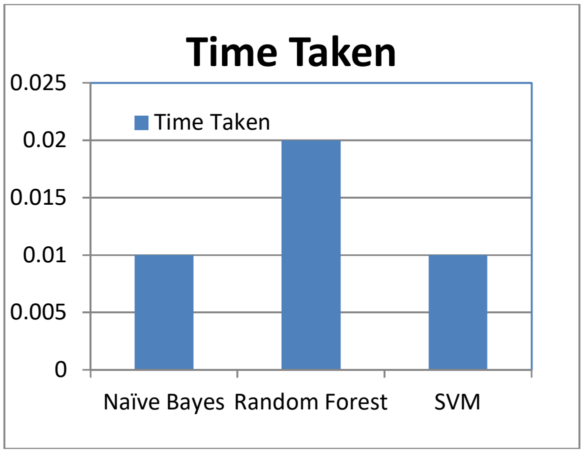

Fig 8. Time taken by algorithms in seconds

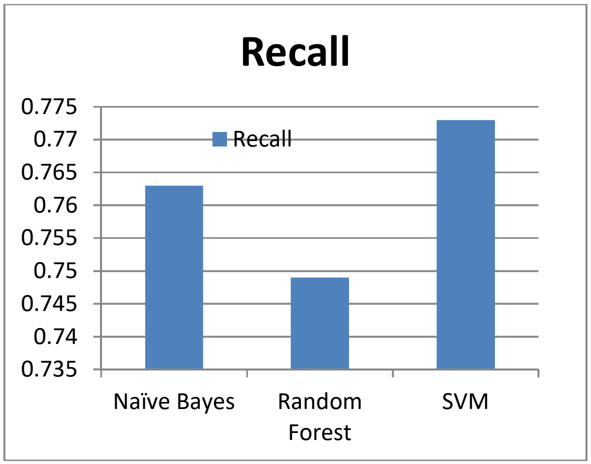

Fig 9. Comparison for Recall

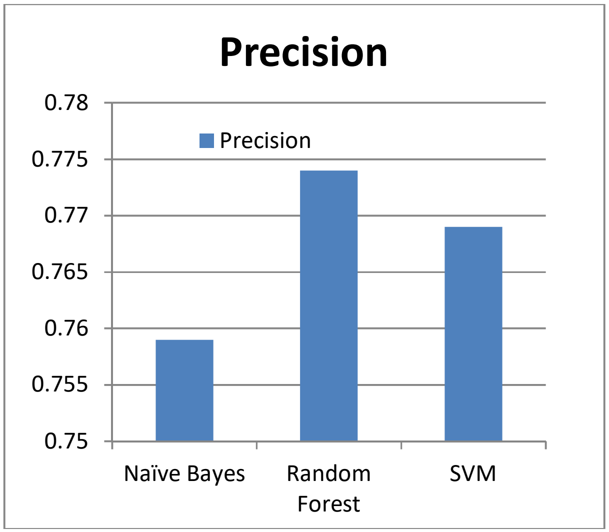

Fig 10. Comparison for Precision

VI. CONCLUSION

This paper presents a comparative study on the performance of various classifiers of data mining over high dimensional data. We observe that Random Forest algorithm performed well with respect to all the factors. Different classification algorithms gives different result on base of accuracy, training time, precision, recall. In future work, different data mining techniques can be used with Random Forest classification algorithm to improve performance and high accuracy.

REFERENCES

1. *Patel Pinky S, Raksha R. Patel, Ankita J. Patel, Maitri Joshi "Review on Classification Algorithms in Data Mining" (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015)*

2. *. Supreet Dhillon, Kamaljit Kaur "Comparative Study of Classification Algorithms for Web Usage Mining", Volume 4, Issue 7, July 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.*

3. *Sarika Pachange, Bela Joglekar "An Approach for Dynamic Characterization of Ensemble classifier for Disease Diagnosis".*

4. *S.Brindha, Dr.K.Prabha, Dr.S.Sukumaran, "A SURVEY ON CLASSIFICATION TECHNIQUES FOR TEXT MINING". 2016 3rd International Conference on Advanced Computing and Communication Systems.*

5. *http://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm.*

6. *. https://www.analyticsvidhya.com/blog/2015*

7. *Amanpreet Singh, Narina Thakur, Aakanksha Sharma, "A Review of Supervised Machine Learning Algorithms" 2016 International Conference on Computing for Sustainable Global Development.*

8. *Vrushali Y Kulkarni, Dr Pradeep K Sinha. "Random Forest Classifiers: A Survey and Future Research Directions". International Journal of Advanced Computing, ISSN: 2051-0845, Vol.36, Issue.*

9. *Jiawei Han (Author), Micheline Kamber (Author), Jian Pei Professor (Author). "Data Mining: Concepts and Techniques"*

10. *“Machine Learning Techniques for Data Mining: A Survey” by Seema Sharma, Jitendra Agrawal, Shikha Agarwal, Sanjeev Sharma.*

11. *Jianguo Chen, Kenli Li, Senior Member, IEEE, Zhuo Tang, Member, IEEE, Kashif Bilal, Shui Yu, Member, IEEE, Chuliang Weng, Member, IEEE, and Keqin Li, Fellow, IEEE “A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment”.*

12. *https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/.*

13. *Deepti Gupta, Udayan Ghose “A Comparative Study of Classification Algorithms for Forecasting Rainfall”.*

14. *Mohammed E. El-Telbany, Mahmoud Warda“An Empirical Comparison of Tree-Based Learning Algorithms: An Egyptian Rice Diseases Classification Case Study”*

15. *Ankur Makwana, Jaymin Patel “Decision Support System for Heart Disease Prediction using Data Mining Classification Techniques” International Journal of Computer Applications (0975 8887) Volume 117 - No. 22, May 2015.*