A Survey on Various Supervised Classification Algorithms

Uma Narayanan
Research Scholar: Dept. of I.T
SOE, CUSAT
Kochi, India
uma@cusat.ac.in

Athira Unnikrishnan
M.Tech Student : Dept. of I.T
SOE, CUSAT
Kochi, India
aathira571993@gmail.com

Dr. Varghese Paul
Professor : Dept. of I.T
Rajagiri School of Engineering and Technology
Kochi, India
vp.itcusat@gmail.com

Dr. Shelbi Joseph
Assistant Professor: Dept. of I.T
SOE, CUSAT
Kochi, India
shelbi@cusat.ac.in

Abstract— **Machine learning is concerned with the development, the analysis and the applications of algorithms that allow computers to learn. Supervised learning is the machine learning task of inferring a function from labelled training data. In this paper, we present a survey of various supervised classification techniques. The goal of this survey is to provide an inclusive review of different supervised classification techniques such as decision tree, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Neural Network.**

Keywords— Decision tree; K-Nearest Neighbour; Neural Network; Naive Bayes; Support vector machine

I. INTRODUCTION

Machine learning is used to make prediction and better understand the system. Machine learning has a wide range of application such as retail, finance, Manufacturing, Medicine, Finance, Telecommunication, and Bioinformatics. Machine learning is mainly categorized into two as shown in fig.1, supervised learning and unsupervised learning. Here we are considering the supervised learning methods. The five classification techniques that we have chosen are Support vector machine, Decision tree, Naive Bayes, K-Nearest Neighbour, and Neural Network.

The rest of the articles will be as follows: The literature survey conducted on the basis of the above classifiers is given in section II. Section III presents Taxonomy. Research information is given in section IV. Conclusions and References are given in the Section V and VI.

II. LITERATURE SURVEY

Ximeng Liu, proposed Privacy-Preserving Patient-Centric Clinical Decision Support System on Naive Bayesian Classification [1]. In this paper, the past patients historical data are gathered in cloud which is then used to train naïve Bayesian classifier without disclosing any individual patient medical data, then the trained classifier can be applied to calculate the disease risk for new coming patients and these patients are given privileges to get their top-k disease names according to their own desire.

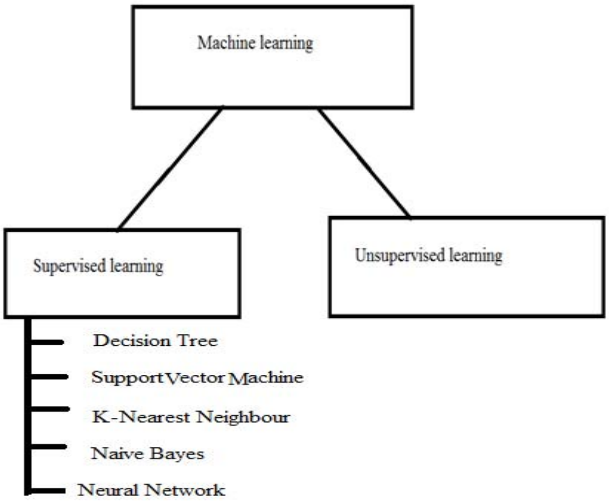

Figure 1: Overview of Machine learning.

To protect the privacy of past patients' historical data, the authors introduced a new cryptographic tool called additive holomorphic proxy aggregation scheme and also brought a privacy-preserving top k disease name retrieval protocol in the system. Naive Bayesian classifier is based on applying Bayes theorem with strong independent assumption between features. The Bayes theorem provides a way for calculating the posterior probability. Naïve Bayesian Classifier has a very good accuracy in classification for large set of data. The problem with this method is that it take the entire attribute independently thus if the attributes are independent then only Naïve Bayesian classifier gives it full accuracy. Naive Bayesian classifier is a very strong statistical classifier when it comes to accuracy. But as the name suggests it is naive and takes the presumption that all attributes are independent of each other.

978-1-5386-1887-5/17/$31.00 ©2017 IEEE

Prof.(Dr.) Kanak Saxena come up with an Efficient Heart Disease Prediction System using Decision Tree [2]. The authors have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given details about the patients' health. The rules can be organized or can be set up based on the necessity of the user. The performance of the system is alleviated in terms of classification accuracy and the results show that the system has the potential to predict the heart disease risk level precisely.

Stephen R.Alty et al, present Cardiovascular Disease Prediction using Support Vector Machines [3]. The authors presented a model for assessing the patient's arterial stiffness and finding the possibility of occurring the cardiovascular disease without resorting to the blood test. In the model authors predict the risk of cardiovascular disease based on the simple measurement of a patients volume pulse measured at the fingertip using an infrared light absorption detector placed on the index finger. Then the necessary features are extracted from the waveform and a Support Vector Machine classifier has been found to make accurate prediction of high or low arterial stiffness as indicated by the Aortal Pulse Wave Velocity

Yevhenii Udovychenko presented Ischemic Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps [4]. In their paper authors used k-Nearest Neighbour algorithm. The algorithm is practiced for binary classification of myocardium current density distribution maps. The classification characteristics where optimized by Selection of number of neighbours for k-NN classifier.

Jayshril S. Sonawanel comes up with a Prediction of Heart Disease Using Multilayer Perceptron Neural Network [5]. In their work they had made a prediction system for heart disease using multilayer perceptron neural network. The neural network in this system takes thirteen clinical features as input and it is trained to predict that there is a presence or absence of heart disease in the patient with highest accuracy of 98% comparative to other systems.

Dana AL-Dlaeen[6] come up with an Using Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease developed an Alzheimer's disease prediction model that can assist medical professionals in predicting the status of the disease based on medical data about patients. The authors used decision tree induction to create decision tree that corresponds to the sample data. To determine which attribute to branch on at each level of the tree entropy or information gain is used.

Sumana Ghosh [7] presented Application of Decision Tree for understanding Indian Educational Scenario analyses the most important factor that will result in the improved education level of India using Decision tree. In the work the input data needed is taken from government portal, the experiment is done based on classifying the data on the literacy rate of each states by dividing them further to Low, Medium and High. The authors used an open source tool called Rapid Miner to generate the results

Mohammed Aashkaar [8] proposed Performance Analysis using J48 Decision Tree for Indian Corporate world analyses the most important factor that will result in understanding the Public-Private sector companies' growth in India. The input data needed is taken from the actual companies' data of India available in government portal and Weka tool is used for the analysis. The experiment provides analysis of confusion matrix and its results.

Li Dongming [9] in their work The Application of Decision Tree C4.5 Algorithm to Soil Quality Grade Forecasting Model, considered the decision tree C4.5 algorithm to construct the model for predicting the soil quality grade. The decision tree generated through the authors experiment can intuitively show the relationship between the composition of soil and soil quality grade.

B. Giraldo [10] in their work Support Vector Machine Classification Applied on Weaning Trials Patients proposes a method for the study of the differences in respiratory pattern variability in patients on weaning trials. The work proposed by authors is based on a support vector machine using 35 features extracted from the respiratory flow signal.

Argyro Kampouraki [11] come up with a Robustness of Support Vector Machine-based Classification of Heart Rate Signals studied the use of Support Vector Machine learning to classify heart rate signals. The database the author chooses consists of twenty young whose age ranges from 21 to 34 years old and twenty elderly whose age ranges from 68 to 85 years old.

Cesar Seijas [12] proposed Estimation of Action Potential of the Cellular Membrane using Support Vectors Machines, presented in the problem of the estimate of the action potential of the cellular membrane. The paper concludes a promissory future to the SVM in the modelling area of biochemical and electrophysiological processes.

U Ravi Babu [13] presented Handwritten Digit Recognition Using K-Nearest Neighbour Classifier presents a new approach to off-line handwritten digit recognition based on structural features. The author has used k-nearest neighbour classifier to classify the MNIST digit images in test database using the feature vector of training database. In K-nn object is classified to a particular class which has majority of vote.

Lianmeng Jiao [14] come up with An Evidential K-Nearest Neighbor Classification Method with Weighted Attributes extends the classical K-nearest neighbour rule within the framework of evidence theory. In the work a new evidential K-nearest neighbor classification method with weighted attributes is proposed to overcome the limitations of EK-NN.

N. Z. Supardi. M. Y[15] in their work Classification of Blasts in Acute Leukemia Blood Samples Using K-Nearest Neighbour paper presents the study on blasts classifying in acute leukemia into two major forms which are acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) by using k-NN. TheExperiment by using k-

NN produced good performance in classifying both AML and ALL with high percentage of accuracy up to 86 percentages.

Ching-man Au Yeung [16] proposed A k-Nearest-Neighbour Method for Classifying Web Search Results with Data in Folksonomies attempted to provide a solution to the problem of keyword ambiguity in Web search using a knearest-neighbour approach to classify documents returned by a search engine, by building classifiers using data collected from collaborative tagging systems.

Abu Nowshed Chy [17] in their work Bangla News Classification using Naive Bayes classifier proposed an approach that provides a user to find out news articles which are related to a specific classification. This system proposed by the authors provides users with efficient and reliable access to classified news from different sources. It achieves a high accuracy of classifications with the possibility that one story be classified into more than one category. Naive Bayes algorithm has been used which is based on probabilistic framework to handle the classification problem.

Anand Shanker Tewari [18] come up with an Opinion Based Book Recommendation Using Naive Bayes Classifier extracted, summarizes and categorizes all the customer reviews of a book. The work proposes a book recommendation technique based on opinion mining and Naïve Bayes classifier to recommend top ranking books to buyers. Naive Bayesian classifier is probabilistic, simple and effective method to categorize text and shows a very good performance..

Rohit Kumar Solanki [19] in their work Spam Filtering Using Hybrid Local-Global Naive Bayes Classifier propose a novel learning framework for classification of messages into spam and legit. Naive Bayesian classifier works on the basis of Bayes theorem with independent assumption between predictors. In the work authors used Enron Spam dataset. The experiment result shows that nearly 95% of the sham messages are correctly classified for each user.

Muhammad Hassan in [20] presented a Gas Classification Using Binary Decision Tree Classifier. The authors used binary decision tree approach for gas classification. The paper proposes a gas classification algorithm for an electronic nose system. The authors targeted five gases to evaluate the performance of the algorithm in two different scenarios. The first scenario corresponds to using the same concentration data in the training and the testing sets. The next scenario corresponds to using different concentrations data in the training and the testing sets.

Animesh Giri in [21] presented A Placement Prediction System Using K-Nearest Neighbours Classifier. The work propose a Placement Prediction System which predicts the probability of a undergrad student getting placed in an IT company by applying the machine learning model of k-nearest neighbour's classification. The authors considered two classes for classification, a yes or a no. The tests were conducted with the help of the placement data of the college dividing the data into training set and cross validating testing set.

Qiaowei Jiang in [22] researched on Deep Feature Weighting in Naïve Bayes for Chinese Text Classification. In the work authors proposed a high efficient method called deep feature weighting Naive Bayes. The data needed for the experiment is taken from the Chinese Natural Language Processing Open Platform which is provided by Ronglu Li. The deep feature weighting naive Bayes to Chinese text classifiers and obtain a better performance than ordinary feature weight naive Bayes.

Sagar K. Bhakre in [23] presented Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier. The proposed system evaluates and classifies statistical features of energy, pitch, MFCC, ZCR. Naïve Bayes classifier is trained by the extracted feature such as pitch, energy, MFCC

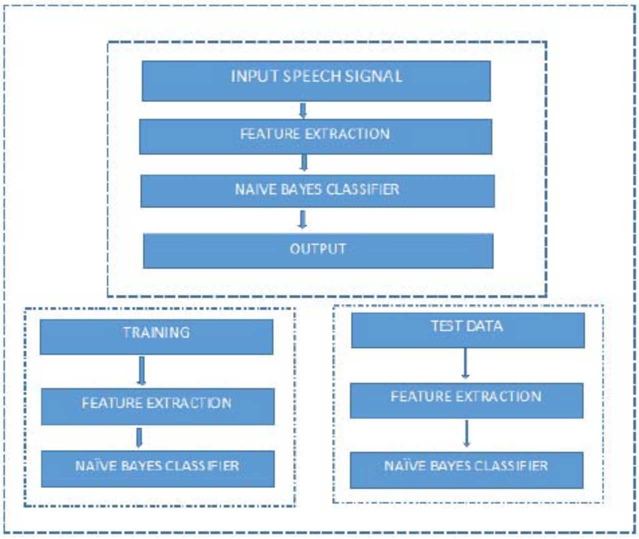

Figure 2: System based on Naïve bayes[24].

Jianxiao Liu in their work [24] come up with An Approach of Semantic Web Service Classification Based on Naive Bayes. The ontology concepts of Web service interface in the experiment use the data set in OWLSTC. The work uses the concept semantic relationships of service interface and capability. The proposed methods can help to enhance service classification and discovery efficiency and accuracy.

Wei Huang in their work [25] Polynomial Neural Network Classifiers Based on Data Preprocessing and Space Search Optimization proposed a novel architecture of polynomial neural network classifier (PNNC) with the aid of data preprocessing technique and space search optimization, which adopts accelerated convergence mechanism instead of purely random search. Polynomial neural network classifier emerges from six components such as inputs, outputs, preprocessing part, premise part, consequence part, and aggregation part.

Shuangrong Liu in their work [26] Prediction of Share Price Trend Using FCM Neural Network Classifier proposed a novel method called Floating Centroids Method. FCM is used to establish the share price trend model, the algorithm fits law of share price trend by finding the optimal neural network.

FCM algorithm has higher average accuracy and better generalization ability.

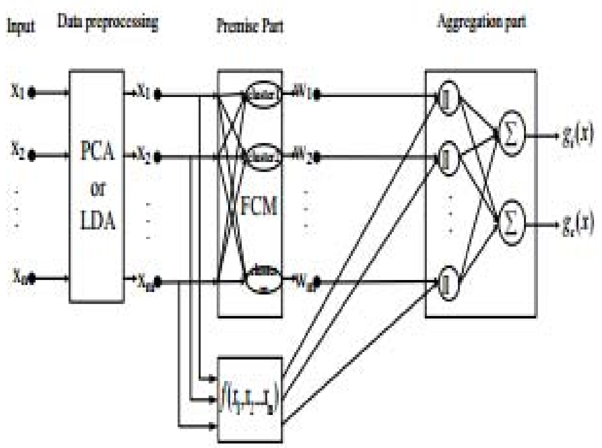

Figure 3: Architecture0 of polynomial neural network classifiers[25].

Manpreet Singh in their work [27] Decision Tree Classifier for Human Protein Function Prediction researched on developing a new decision tree induction technique in which uncertainty measure is used for best attribute selection. The data needed for the experiment is taken from the from Human Protein Reference Database.

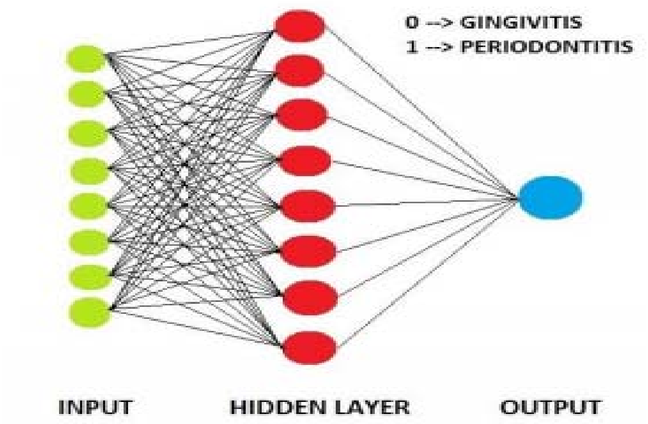

Figure 4: Three layer neural network architecture [28]

Anita Thakur in their work [28] Symptom & risk factor based diagnosis of gum diseases using neural network researched on a non-invasive method to detect whether a patient is suffered from gum disease. The authors used 11 inputs which are combination of symptoms and risks factors. Hidden layer automatically extracts the features of the input and reduces its dimensionality further. The output prepared in such a manner that if 1 periodontal disease is present & if it is 0 gingivitis disease is present.

TABLE I
TAXONOMY

Sl No	Title	Merit	Demerit
1.	Privacy-Preserving Patient-Centric Clinical Decision Support System on Naive Bayesian Classification[1]	Improve Diagnosis accuracy, Reduce diagnosis time	It take the entire attribute independently thus if the attributes are independent then only Naive Bayesian classifier gives it full accuracy
2	Efficient Heart Disease Prediction System using Decision Tree.[2]	It is simple and gives an accurate result based on the cases, which has been used in building it.	Building the decision tree with huge amount of data might be time consuming.
3	Cardiovascular Disease Prediction using Support Vector Machines.[3]	SVM can, classify the data into binomial class or multilevel class	Performance degrades when the data set has a large amount of noisy data
4	Ischemic Heart Disease Recognition by k-NN Classification of Current Density Distribution Maps. [4]	k-nn is a simple classifier that works well on basic recognitoin problems	k-nn can be slow for real-time prediction if there are a large number of training examples and is not robust to noisy data
5	Prediction of heart disease using multilayer perceptron neural network.[5]	highest accuracy of ninety-eight percentage compared to the other systems	Neural networks are not probabilistic.
6	Application of Decision Tree for understanding Indian Educational Scenario.[7]	Analysed the most important factor that will result in the improved education	Once a mistake is made at higher level, any sub-tree is wrong
7	The Application of Decision Tree C4.5 Algorithm to Soil Quality Grade Forecasting Model.[9]	Constructed the model for predicting the soil quality grade using decision tree C4.5 algorithm.	Training time is relatively expensive.

8	Robustness of Support Vector Machine-based Classification of Heart Rate Signals.[11]	Signals studied the use of Support Vector Machine learning to classify heart rate signals.	Performance degrades when the data set has a large amount of noisy data.
9	An Evidential K-Nearest Neighbor Classification Method with Weighted Attributes [14]	Proposed a method with weighted attributes to overcome the limitations of EK-NN.	Distance based learning is not clear. Which type of distance to use and which attribute to use to produce the better results
10	Classification of Blasts in Acute Leukemia Blood Samples Using K-Nearest Neighbor.[15]	The experiment done by using k-NN produced good performance in classifying both AML and ALL with percentage of accuracy up to 86 %.	Computational cost is quite high.
11	A k-Nearest-Neighbor Method for Classifying Web Search Results with Data in Folksonomies. [16]	Provided a solution to the problem of keyword ambiguity in Web search using a k-nearest neighbor classifier.	Distance based learning is not clear. Which type of distance to use and which attribute to use to produce the better results
12	Bangla News Classification using Naive Bayes classifier[17]	Proposed system uses an approach that provides user to find out news articles using Naive Bayes classifier.	Computation takes longer time.
13	Opinion Based Book Recommendation Using Naive Bayes Classifier.[18]	The work proposes a book recommendation technique based on opinion mining and Naïve Bayes classifier.	It takes the entire attribute independently.
14	Spam Filtering Using Hybrid Local-Global Naive Bayes Classifier.[19]	The work proposed a novel learning framework for classification of messages into spam and legit.	Only independent attributes are taken.
15	Gas Classification Using Binary Decision Tree Classifier.[20]	The paper proposes a gas classification algorithm for an electronic nose system.	Once a mistake is made at higher level, any sub-tree is wrong.
16	A Placement Prediction System Using K-Nearest Neighbor Classifier.[21]	Proposed a system which predicts the probability of an undergraduate student getting placed in an IT company.	Distance based learning is not clear.
17	Deep Feature Weighting In Naive Bayes For Chinese Text Classification.[22]	A new method for Chinese text classifications based on weighted Naïve Bayes.	Error was not mentioned.
18	Prediction of Share Price Trend Using FCM Neural Network Classifier.[26]	Proposed a novel method called Floating Centroids Method. FCM is used to establish the share price trend model	Require high processing time for large Neural Network.
19	Decision Tree Classifier for Human Protein Function Prediction.[27]	Researched on developing a new decision tree induction technique in which uncertainty measure is used for best attribute selection	Computationally intensive, especially when the size of training set grows.
20	Symptom & risk factor based diagnosis of gum diseases using neural network.[28]	A noninvasive method to detect whether a patient is suffered from gum disease	Require high processing time for large Neural Network.

III. Research Information

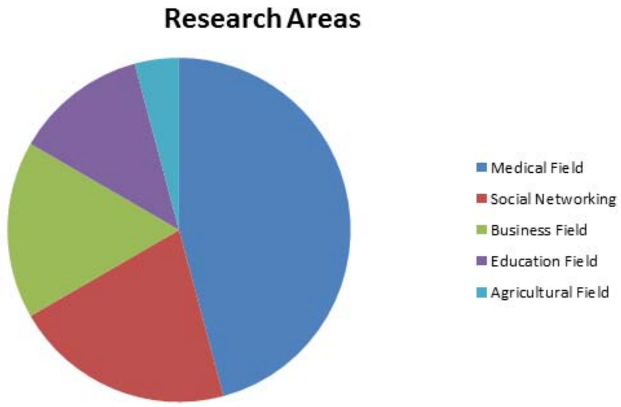

Figure 5: Classification of papers according to area of research

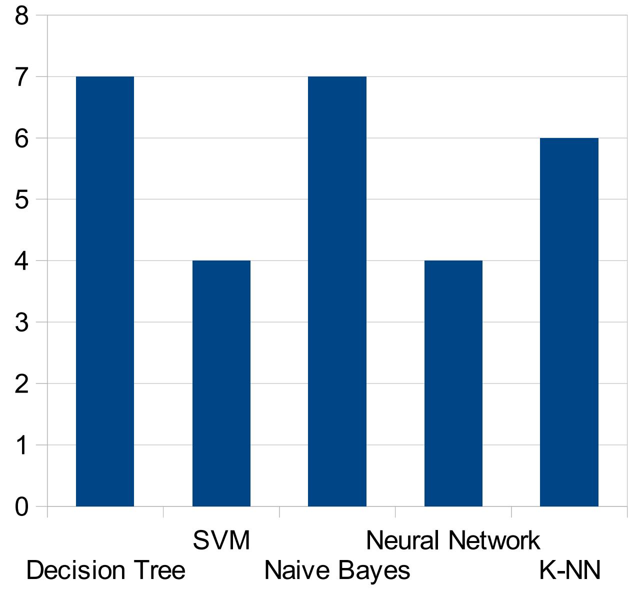

Figure 6: Classification of papers according to classification algorithm used.

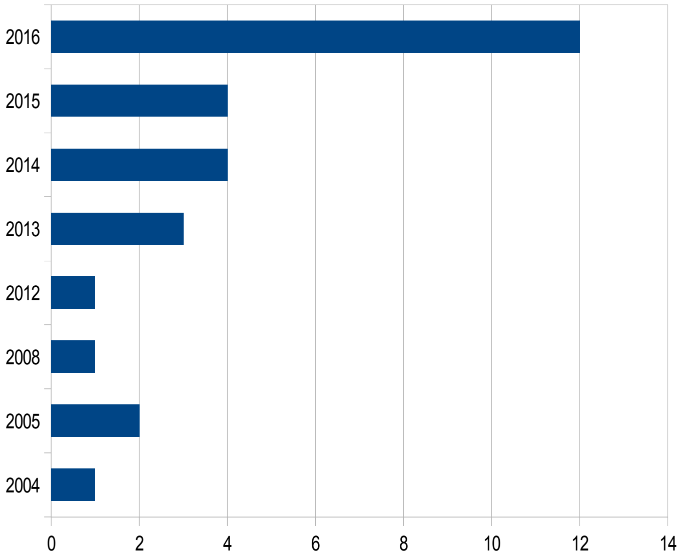

Figure 7: Classification of papers according to year of publication

IV. Conclusion

These classification techniques help us to understand the way in which data can be determined and grouped when a new set of data is available. This survey paper covers with various supervised classification techniques. Advantages and Disadvantages of each paper have been discussed. Decision Tree and Naïve Bayes are the most commonly used classifiers. The survey contains paper from different fields like Medical, Social Networking, Education, Business and agriculture. Through the survey it was analyzed that most of the application of classification Technique is in Medical Field. It has been observed that classification Technique plays an important role in medical analysis. Classification Techniques is widely used in every field and it is one of the hot research areas.

References

[1] Ximeng Liu, Student Member, IEEE, Rongxing Lu, Member, IEEE, Jianfeng Ma, Le Chen, and Baodong Qin, “Privacy-Preserving Patient-Centric Clinical Decision Support System on Naive Bayesian Classification”, 2168-2194 © 2015 IEEE.

[2] Purushottam, Prof. (Dr.) KanakSaxena, Richa Sharma, “Efficient Heart Disease Prediction System using Decision Tree”, ISBN:978-1-4799-8890-7/15/$31.00 ©2015 IEEE.

[3] Stephen R. Alty*, Sandrine C. Millasseaut, Philip J Chowienczykt and Andreas Jakobssont “Cardiovascular Disease Prediction Using Support Vector Machines”, 0-7803-8294-3/04/$20.00 ©2004 IEEE.

[4] YevheniiUdovychenko, Anton Popov, IllyaChaikovsky, “Ischemic Heart Disease Recognition by *k-NN* Classification of Current Density Distribution Maps”, **978-1-**4673-6534-5/15/$31.00 ©2015 IEEE.

[5] Jayshril S. Sonawane, D. R. Patil., "Prediction of Heart Disease Using Multilayer Perceptron Neural Network" ISBN No.978-1-4799-3834-6/14/$31.00©2014 IEEE.

[6] Dana AL-Dlaeen, Abdallah Alashqur, "Using Decision Tree Classification to Assist in the Prediction of Alzheimer's Disease", 978-1-4799-3999-2/14/$31.00©2014 IEEE.

[7] Sumana Ghosh, Shweta Shukla, Dr. Deepti Mehrotra, "Application of Decision Tree for understanding Indian Educational Scenario".,978-1-4673-9939-5/16/$31.00 ©2016 IEEE.

[8] Mohammed Aashkaar, Purushottam Sharma, Naveen Garg , "Performance Analysis using J48 Decision Tree for Indian Corporate world", 978-1-4673-8819-8/16/$31.00 ©2016 IEEE.

[9] Li Dongming, Li Yan, Yuan Chao, Li Chaoran, Liu Huan, Zhang Lijuan, "The Application of Decision Tree C4.5 Algorithm to Soil Quality Grade Forecasting Model" , 978-1-4673-8515-2/16/$31.00 ©2016 IEEE.

[10] B. Giraldo, Member, IEEE, A. Garde, C. Arizmendi,Member, IEEE, R. Jané, Member, IEEE, S. Benito, I. Diaz, D. Ballesteros , "Support Vector Machine Classification Applied on Weaning Trials Patients" 1-4244-0033-3/06/$20.00 ©2006 IEEE.

[11] Argyro Kampouraki, Christophoros Nikou∗ and George Manis, "Robustness of Support Vector Machine-based Classification of Heart Rate Signals". 1-4244-0033-3/06/$20.00 ©2006 IEEE.

[12] Cesar Seijas, Antonino Caralli , Member, IEEE and Sergio Villazana ,"Estimation of Action Potential of the Cellular Membrane using Support Vectors Machines"1-4244-0033-3/06/$20.00 ©2006 IEEE.

[13] U Ravi Babu, Dr. Y Venkateswarlu, Aneel Kumar Chintha "Handwritten Digit Recognition Using K-Nearest Neighbour Classifier", 978-1-4799-2876-7/13 $31.00 © 2013 IEEE DOI 10.1109/WCCCT.2014.7.

[14] Lianmeng Jiao, Quan Pan, Xiaoxue Feng, and Feng Yang, "An Evidential K-Nearest Neighbor Classification Method with Weighted Attributes" 978-605-86311-1-3 ©2013 ISIF.

[15] Z. Supardi. M. Y. Mashor. N. H. Harun. F. A. Bakri, R. Hassan, "Classification of Blasts in Acute Leukemia Blood Samples Using K-Nearest Neighbour N" 978-1-4673-0961-5/12/$31.00 ©2012 IEEE

[16] Ching-man Au Yeung Nicholas Gibbins Nigel Shadbolt, "A k-Nearest-Neighbour Method for Classifying Web Search Results with Data in Folksonomies" 978-0-7695-3496-1/08 $25.00 © 2008 IEEE DOI 10.1109/WIIAT.2008.269.

[17] Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das ," Bangla News Classification using Naive Bayes classifier" 978-1-4799-3497-3/13/$31.00 ©2013 IEEE

[18] Anand Shanker Tewari, Tasif Sultan Ansari, Asim Gopal Barman, "Opinion Based Book Recommendation Using Naive Bayes Classifier" 978-1-4799-6629-5/14/$31.00c 2014 IEEE.

[19] Rohit Kumar Solanki, Karun Verma, Ravinder Kumar, "Spam Filtering Using Hybrid LocalGlobal Naive Bayes Classifier" 978-1-4799-8792-4/15/$31.00 © 2015 IEEE

[20] Muhammad Hassan, Amine Bermak "Gas Classification Using Binary Decision Tree Classifier", 978-1-4799-3432-4/14/$31.00 ©2014 IEEE..

[21] Animesh Giri, M Vignesh V Bhagavath, Bysani Pruthvi, Naini Dubey ,"A Placement Prediction System Using K-Nearest Neighbors Classifier" 978-1-5090-1025-7/16/$31.00 ©2016 IEEE.

[22] Qiaowei Jiang, Wen Wang, Xu Han, Shasha Zhang, Xinyan Wang, Cong Wang, "Deep Feature Weighting In Naive Bayes For Chinese Text Classification", 978-1-5090-1256-5/16/$31.00 ©2016 IEEE.

[23] Sagar K. Bhakre, Prof.Arti Bang, "Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier", 978-1-5090-2029-4/16/$31.00 @2016 IEEE..

[24] Jianxiao Liu, Zonglin Tian, Panbiao Liu, Jiawei, Zhao Li, "An Approach of Semantic Web Service Classification Based on Naive Bayes" , 978-1-5090-2628-9/16 $31.00 © 2016 IEEE DOI 10.1109/SCC.2016.53.

[25] Wei Huang, Sung-Kwun Oh , "Polynomial Neural Network Classifiers Based on Data Preprocessing and Space Search Optimization" 978-1-5090-2678-4/16 $31.00 © 2016 IEEE DOI 10.1109/SCIS&ISIS.2016.173

[26] Shuangrong Liu; Bo Yang; Lin Wang; Xiuyang Zhao; Jin Zhou; Jifeng Guo, "Prediction of share price trend using FCM neural network classifier" 978-1-5090-3367-6/16/$31.00 ©2016 IEEE DOI: 10.1109/ICCSS.2016.7586428..

[27] Manpreet Singh, Parvinder Singh, Dr. Hardeep Singh"Decision Tree Classifier for Human Protein FunctionPrediction" 1-4244-0716-8/06/$20.00 ©2006 IEEE.

[28] Anita Thakur , Payal Guleria2, Nimisha Bansal "Sympton & Risk factor based diagnosis of Gum diseases using Neural Network", 978-1-4673-8203-8/16/$31.00 c 2016 IEEE