EVALUATING THE PERFORMANCE OF NAIVE BAYES CLASSIFICATION ALGORITHM FOR BLOOD DONORS PROBLEM

[1]Anil Kewat, [2]P.N.Srivastava, [3]Arvind Kumar Sharma
[1]Department. Of Maths & Comp. Appl. B.U., Jhansi, India,
[2] Prof., Department. of Maths & Comp. App., B.U., Jhansi, India
[3] Prof., Department of CS, Kota University, Kota India

Abstract : ***The term like intelligent systems, knowledge based systems, expert systems etc. are meant to express message that it is possible to construct machines that can exhibit intelligence just like people in doing a little easy tasks. In these tasks we search for the final result of the performance of the machine for evaluation with the overall performance of a human being. We characteristic intelligence to the machine if the overall performance of the machine and human being are the identical. In the recent trends soft computing algorithms with the data mining techniques are applied in the different application domain for the prediction, knowledge extraction and performance evaluation tasks. Healthcare is one of them. In this paper a Naive Bayes soft computing algorithm is used with the data mining technique for investigating the performance of the blood bank and blood donors in a particular city on the idea of real-world datasets. Naive Bayes computing algorithm has the capability of supervised learning in addition to the statistical learning. Performances of Naive Bayes algorithm on the idea of varied parameters are evaluated and results are collected.***

Keywords: Naive-Bayes, soft computing, Blood-Donors, Neural Network, dataset.

---

1 INTRODUCTION

Recently using of soft computing techniques as an excellent tool for knowledge discovery in huge amounts of data. These hybrid combinations have the potential to handle large amount of data in a very quick and effective manner. Since the data to be analyzed is having with inexact and uncertainty. Therefore traditional techniques are not adequate. Properties of the same kind are typical of soft computing. Therefore the application of soft computing techniques results in systems that have high device ratio. Recently most widely used soft computing techniques are as follows:

1.1 **Genetic Algorithm**

Genetic algorithms are adaptive seek algorithms based totally on the evolutionary thoughts of natural choice and genetics. As such they constitute a sensible exploitation of a random are seeking used to resolve optimization hassle. The simple techniques of the genetic algorithm are designed to simulate manner in herbal structures vital for evolution, in particular people who examine the principles of nice survival. Genetic algorithms are higher than the tough computing algorithms in that they are more robust. In looking a large nation location a genetic algorithm may additionally offer substantial blessings over greater widespread optimization techniques. In famous, genetic set of rules starts as follows. An initial population is created which consist of randomly generated regulations. Every rule can be represented via a string of bits. Based at the belief of survival of the fittest, a ultra-modern populace is common to encompass the fittest regulations within the modern-day populace, as well as offspring of those suggestions. Normally, the fitness of a rule is classed through the usage of its class accuracy on set of education samples. Offspring are created by means of making use of genetic operators including crossover and mutation. In crossover, substrings from pairs of tips are swapped to shape new pairs of rules. In mutation, randomly selected bits in a policies string are inverted. The technique of producing new populations primarily based on in advance populations of guidelines keeps till a populace, P, evolves in which each rule in P satisfies a pre-particular health threshold. Genetic algorithms are without troubles parallelizable and had been used for class as well as extraordinary optimization troubles. In information mining, they will be used to evaluate the health of other algorithms [1].

1.2 **Neural Networks**

New models of computing to participate in pattern recognition tasks are influenced with the aid of the constitution and performance of our organic neural community. A set of processing models when assembled in a intently interconnected community, offers a wealthy structure exhibiting some features of the organic neural community. This kind of constitution is known as an artificial neural network. Considering ANNs are applied on computer systems, it is valued at evaluating the processing capabilities of a computer with these of the brain [Simpson, 1990]. Neural networks are sluggish in processing understanding, on the grounds that cycle time akin to a neural event promoted by using an outside stimulus happens in milliseconds range therefore the pc method expertise virtually one million time faster. Neural networks can participate in massively parallel operations for the reason that prompted from organic networks where mind operates with hugely parallel operations each and every of them having comparatively fewer steps. Neural networks have gigantic number of computing elements and the computing isn't restrained to within neurons. Neural networks retailer expertise within the strengths of the interconnections. In a neural community new understanding is added via adjusting the interconnections strengths, without destroying the historic information. Consequently expertise in the brain is adaptable whereas in the laptop it's strictly replaceable. Neural networks exhibit fault tolerance considering that the expertise is dispensed within the connection throughout the network. There's no significant manage for processing expertise within the brain. In a neural network each and every neuron acts established on the neurons connected to it. Accordingly there is not any specified manipulate mechanism outside to the computing undertaking [2].

1.3 Support Vector Machine

Support vector machine, a promising new approach for the category of each linear and nonlinear data. An SVM is a set of rules that works as follows. It uses a nonlinear mapping to transform the authentic training records into a higher measurement. Within this new dimension, it searches for the linear greatest keeping apart hyper aircraft. With the ideal nonlinear mapping to a sufficiently excessive measurement, information from instructions can usually be separated through a hyper plane. The SVM reveals this hyper plane using help vectors and margins. the first paper on SVM became offered in 1992 through Vladimir Vapnik and colleagues, even though the ground work for SVM has been around because the Sixties. Although the training time of even the fastest SVM may be extremely sluggish, they're noticeably correct, due to their capacity to model complex nonlinear decision obstacles. They are a lot less susceptible to over fitting than different methods. The guide vector additionally offer a compact description of the discovered version. SVMs can be used for prediction as well as class. They had been carried out to some of areas, consisting of handwritten digit popularity, object recognition, and speaker identity, in addition to benchmarks time series prediction checks.

1.4 Fuzzy Logic

The concept of fuzzy sets was first introduced by L. Zadeh in 1965[Zadeh, 1965] to represent vagueness present in human reasoning. Fuzzy sets can be considered as a generalization of the classical set theory. In a classical set an element of the universe either belongs to or does not belong to the set. Thus the belongingness of an element is crisp. In a fuzzy set the belongingness of an element can be a continuous variable. Mathematically, a fuzzy set is a mapping from the universe of discourse to [0,1]. The higher the membership value of an input pattern to a class, the more is the belongingness of the pattern to the class [3].The membership function is usually designed by taking into consideration the requirements and constraints of the problem. Fuzzy logic deals with reasoning with fuzzy sets and fuzzy numbers. it's far to be noted that fuzzy uncertainty isn't the same as probabilistic uncertainty [Klir and Folger, 1993; Klir and Yuan, 1995]. In [Sarkar et al, 1998; Pal and Mitra, 1992] the network outputs are interpreted as fuzzy membership values. Learning laws are derived by minimizing a fuzzy objective function in a gradient descent manner. In [Sarkar and Yegnanarayana, 1997d] the concept of cross entropy was extended to incorporate fuzzy set theory. Incorporation of fuzziness in the objective functions led to better classification in many cases. In [Tsao et al, 1994] Kohonen's clustering network has been generalized to its fuzzy counterpart. The merits of this approach is that the final weight vectors do not depend on the sequence of presentation of the input vectors. The method uses a systematic approach to determine the learning rate parameter and size of the neighbourhood.

1.5 Rough Sets

In many type tasks the aim is to shape lessons of objects which might not be considerably unique. These indistinguishable objects are beneficial to construct knowledge base concerning the task. For instance if the objects are classifieds consistent with color (red, black) and shape (triangle, square and circle) then the indiscernible classes are red triangles, black squares, red circles, and so on. As a result these attributes make a partition in the set of objects. Now if red triangles with distinct regions belong to different classes, then it is not possible for anybody to classify these two red triangles primarily based on the given attributes. This form of uncertainty is known as rough uncertainty [Pawlak, 1982; Pawlak et al, 1995]. Pawlak formulated the rough uncertainty in terms of rough sets. The rough uncertainty is absolutely prevented if we will successfully extract all of the important capabilities to represent distinct objects. But it may now not be feasible to guarantee this as our knowledge about the system generating the records is limited. It should be stated that rough uncertainty is different from fuzzy uncertainty [Dubois and Prade, 1992]. Using rough sets it could be feasible to lower the dimensionality of the input without dropping any statistics. A set of features is enough to categorize all of the input patterns if the rough ambiguity, for this set of capabilities is equal to 0. The use of this quantity it's far possible to pick a right set of features from the given data [Pawlak et al, 1988].

2 Literature Review

Nowadays there is a huge amount of information being collected and confine databases everywhere across the globe. There are valuable information and information "hidden" in such databases; and while not automatic ways in which for extracting this information it's much impossible to mine for them. Throughout the year's many algorithms were created to extract what is referred to as nuggets of knowledge from huge sets of information. There are several methodologies to approach this drawback.

W. Boonyanusith and P. Jittamai [4] on this studies the sample of blood donors' behaviours based on elements influencing blood donation choice is accomplished the usage of on line questionnaire. The surveyed records are used for device studying techniques of synthetic intelligence to classify the blood donor company into donors and non-donors. the accuracy finding out of the surveyed facts is achieved the usage of the synthetic neural network (ANN) and choice tree techniques on the way to are looking ahead to from a series of individual Blood conduct information whether or not or now not each character is a donor. the consequences suggest that the accuracy, precision, and do not forget values of ANN method are better than those of the choice tree method [4].

Classification is an information analysis technique to extract models describing necessary knowledge classes and predict future values. Processing uses classification techniques with machine learning, image process, language method, applied mathematics and visualization techniques to seek out and gift info in a clear format.

Most of the classification algorithms in literature are memory resident, usually presumptuous a little info size.Recent processing analysis has designed on such techniques, developing ascendable and durable classification techniques capable of handling huge disk-resident knowledge. The classification has varied applications in addition to flight classification, fraud detection, target promoting, performance prediction, manufacturing, and identification. The performance of the classification techniques is measured by the metrics like accuracy, speed, robustness, quality, comprehensibility, time and interpretability. Classification technique depends on the inductive learning principle that analyzes and finds the patterns from the knowledge. If the character of an environment is dynamic, then the model ought to be adaptive i.e. it got to be able to learn and map with efficiency.

Ankit et al [5] focuses on data mining and trends associated with it. In this paper, the main purpose of the system is to increase blood donor's rate as well as to attract more blood donors to donate blood. The work has been made to classify and predict the number of blood donor's according to their age and blood group. In this work, the WEKA data mining tool and the J48 algorithm is used to classify the data and evaluation of the data.

Limère et al. (2004) bestowed a model for firm growth with call tree induction principle. It offers fascinating results and fits the model to economic info like growth competence and resources, growth potential and growth ambitions.

Xu et al. (2008) projected a reproducing kernel metric space framework for information theoretical learning. The framework uses the radial plus definite kernel operate i.e. cross-information potential. though this framework provides the best result than the previous RKHS frameworks, still there is a drawback to come to a decision on an appropriate kernel operate for a particular domain.
Shilton and Palaniswami (2008) printed a unified approach to support vector machines. This unified approach is developed for binary classification and after extended to one-class classification and regression. Kumar et al. (2012) explored a binary classification framework for two stage multiple kernel learning. The distinct advantage of this binary classification framework is that it's easier to leverage analysis in binary classification and to develop ascendible and durable kernel-based algorithms.
Takeda et al. (2012) projected a unified durable classification model that optimizes the prevailing classification models like SVM, minimax likelihood machine, and Fisher discriminate analysis. It provides several blessings like well-defined theoretical results, extends the prevailing techniques and clarifies relationships among existing models.
Yee and Nursingd Haykin (1993) viewed the pattern classification as an ill-posed disadvantage, it is a demand to develop a unified theoretical framework that classifies and solves the unwell expose problems. Recent literature on classification framework has reportable higher results for binary class datasets alone. For multiclass datasets, there's an aSbsence of accuracy and lustiness. So, evolving an economical classification framework for multiclass datasets remains an open analysis downside.
The evaluation of the parameters which influence the psychology of blood donors has been conducted largely because of the numerous effect of blood insufficiency at the continuance of patients [6]. The approach in discovering new styles of huge statistics units is recorded processing. It is able to be accustomed extract information from a present information set and redecorate into a character's perceivable structure for any use [7]. It utilizes techniques on the intersection of information, facts systems, gadget mastering, and computing. ANN might be a way of facts processing it really is accustomed predict or classify records inside the area of ideas or emotions and behaviors of customers efficiently [8]. It is fashioned in getting to know styles of the statistics [9]. To resolve the difficulty of category and grouping records are effective to investigate the promoting databases [10]. Multi-layer Perceptron may be a giant and useful feed-forward ANN version, which might be accustomed examine dataset to categories the focused cluster [11]. Moreover, choice Tree is one many of the useful strategies in type by way of getting to know patterns of the dataset. it will display end result diagrammatically as a tree model a good way to factor every step of concluding process from input to output [12].
Borkar and Deshmukh [13] planned using Naïve Thomas Bayes classifier for detection of swine-flu disease. The method starts with finding likelihood for every attribute of swine flu against all output. The probabilities of every attribute are then increased. Choosing the most likelihood from all the possibilities, the attributes belong to the category variable with the most worth. The promising results of the planned theme is used for investigation more the swine flu disease in patients using info technology. Patil et al [14] worked within the direction of diagnosis whether or not a patient along with his given info relating to age, sex, pressure, blood glucose, chest pain, electrocardiogram reports etc will have a cardiovascular disease later in life or not.
The experiments involve taking the parameters of the medical tests as inputs. The proposal is effective enough in getting used by nurses and medical students for training functions. the data mining technique used is Naive Thomas Bayes Classification for the event of decision network in cardiovascular disease Prediction System (HDPS). The performance of the proposal is additional improved employing a smoothing operation.
Kharya et al [15] proposed detecting in patients the chances of having Breast Cancer later in life. The severity of Breast Cancer is necessary seeing it becoming the second most cause of death among women. A Graphical User Interface (GUI) is designed for entering the patient‘s record for the prediction. The records are mined through the data repository. Naïve Bayes classifier, being simple and efficient is chosen for the prediction. The results obtained by the Naïve Bayes classifier are accurate, have low computational effort and fast. Implementation of the proposal is done through Java and the training of data is done using from UCI machine repository [16]. Another advantage of the proposed system is that the system expands according to the dataset used.
Stephanie J. Hickey [17] proposed using Naïve Bayes soft Classifier for public health domain. The public health data are used as an input and the purpose of the study was to analyze one or several attributes that predict a target attribute without the need for searching the input space exhaustively. The proposal achieved its goal with the increase in accuracy of classification. The target attributes were related to diagnosis or procedure codes.
Ambica et al [18] developed an efficient decision support system for Diabetes disease by using Naïve-Bayes soft computing algorithm. The developed classification system contains two steps. The first step explains analysis of optimality of the dataset and accordingly extraction of the optimal feature set from the training dataset.
The second step create the new dataset as the optimal training dataset and the developed classification scheme is now applied on the optimal feature set. The mismatched features from the training data are ignored and the dataset attributes are used for the calculation of posterior probability. The proposed procedure, therefore, shows elimination of unavailable features and document wise filtering.

3 Proposed Methodology

Methodology used to accomplish various task is shown by following figure1.

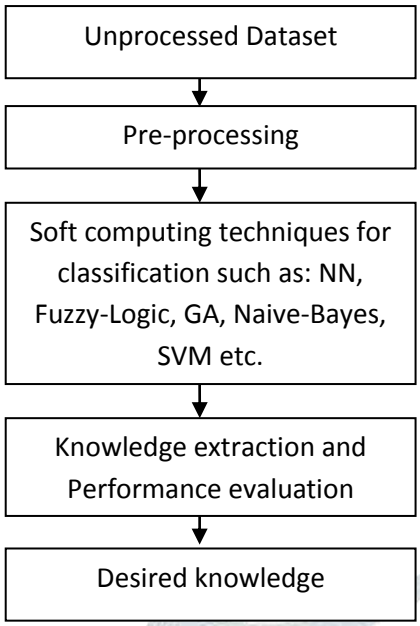

Fig.1.knowledge extraction process with Soft Computing techniques

4 Experimental Evaluations

As long as the underlying assumption of independence is true, the Naïve Bayes classifier works fine. The independence here refers to the idea that the underlying category should be a better predictor of the options, the features that are independent given the class. The other benefits of the classifier embody simplicity, quick to classify, not sensitive to extraneous options and simple implementation that create it a promising technique to be tried for a brand new classification issue.

However, the underlying assumption isn't each and every time feasible as per the real world situation. In this paper it's been applied to classify and predict dataset. Performances of naive bayes classification technique on the idea of varied parameters are evaluated and results collected. There are a lot of merits of this algorithm some of them are as follows:

1. When the input variables are unconditional this algorithm plays nicely.
2. This classifier converges quicker requiring fairly less training data than different discriminative models such as logistic regression.
3. It is less difficult to expect the class of the test dataset in this algorithm. This classifier is an excellent guess for multi-class predictions also.
4. This algorithm has offered excellent performance in numerous application areas in spite of conditional independence assumption.
5. There are different flavours of Naive Bayes algorithms such as Gaussian naïve-bayes, Multinomial naïve-bayes, Bernoulli naive-bayes.
6. It is best suited for text classification problems. Generally it is used for spam email classification problem
7. This algorithm can also be used to train small dataset.

There are numerous areas where naive bayes algorithm are used some of them are as follows.

1. To check whether your email is junk mail or not.
2. For characterizing news articles about entertainment, politics, sports, technology etc. this algorithm is used.
3. It is used by social sites such as face book to break down announcements communicating positive or negative feelings.
4. It is also used as a document classification for indexing the document in a database.

4.1 Working Methodology of the Naive Bayes Classifier

Naïve Bayes Classification method starts with text document as an input. For measuring the relative degree of association between the class-word pairs, the classifier makes a log-linear decision rule that assigns an independent parameter to each class-word pair. The two steps of the classifier include Calculation of class conditional probability and Calculation of classification or posterior probability. For each term t and class c_j, the class conditional probability $(ti\,|cj)$ taking into consideration only one training set is represented as follows:

$$\hat{p}(t_i/c_j) = \frac{\Sigma tf\,(t_i, d \in c_j) + \alpha}{\Sigma N_{d \in c_j} + \alpha.M} \qquad (1)$$

Where $\Sigma tf\,(t_i, d \in c_j)$: the total sum of the term frequencies of the word from all documents in the training samples belonging to a class C_J, α , is a smoothing parameter.

$\Sigma N_{d \in c_j}$: Is the sum of all term frequencies in the training dataset for class C_j, and M is the number of terms.

Once the conditional probability is calculated for each term and class, the trained classifier is able to predict the class of any upcoming new document. Let the document to be queried query is with feature vectors represented by term frequencies. The posterior probability of a document which belongs to a class c_j is the product of individual class conditional probabilities of all terms contained in the query document.

$$\hat{p}(d/c_j) = \hat{p}(t1/c_j).\hat{p}(t2/c_j)...\hat{p}(t_m/c_j)$$
$$= \prod_{i=1}^{M} \hat{p}(t_i/c_j)^{tf(ti,d)}$$

(2)

After the calculation of both the probabilities, the highest probability of class c_k which show that the queried document d belongs to class c_k is given by k= argmax_j

4.2 Data Set Used

The blood donor's information collected from the Kota blood bank having 5656 instances with 12 attributes. In principle, the usage of big data set to construct the classifier version will increase the performance while classifying new statistics due to the fact it would be less complicated to assemble an extra trendy model and subsequently finding a suitable match for our dataset. The dataset used to construct the classifier model is dependent on a variety of things which include the scale of the type of problem, the classifier algorithm used and the statistics set. The blood donor classification model was evaluated using a Naive-Byes classification technique [25]. There are two categories of the blood donor' male and female. There are 5656 Instances of the blood donors dataset and there are seven attributes which are Bag-no, Age, Date Group, Available, Tested and Sex. The Testing mode is set at 10-fold cross-validation. The total execution time to build model is 0.02 seconds. The results of blood donors dataset for Naïve Bayes Classification technique is shown in the table-1.

Table 1.

Truly classified instances	5515	97.5588%
Badly classified instances	138	2.44412%
Unknown instances	03	
Total instances	5526	

In the following step of the experiment we have calculated the classification accuracy in the table-2.

Table 2.

MAE	0.04890
RMSE	0.1574
RAE	102.9244%
RRSE	102.3834%

Where MAE, RMSE, RAE, RRSE are Mean absolute, Root mean squared, Relative absolute and Root relative squared errors respectively.

According the male and female class accuracy is given in the following table-3.

Table 3.

CLASS	MALE	FEMALE
True Positive Rate	1	0
False Positive Rate	1	0
Precision	0.976	0
Recall	1	0
F-Measure	0.988	0
ROC-Area	0.634	0.63

The following graph shows the Accuracy of Male and Female class.

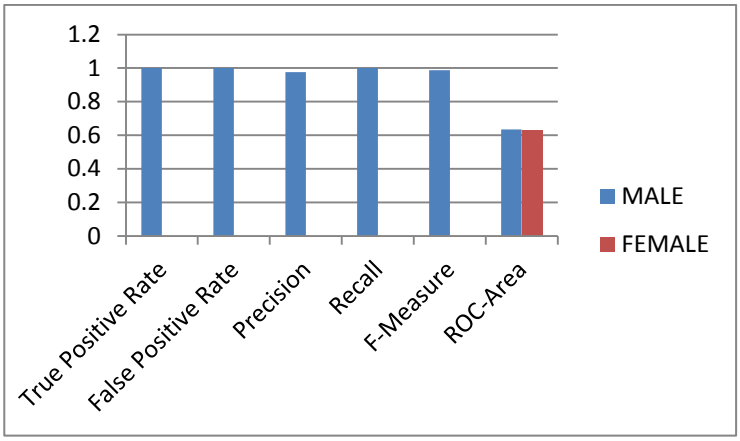

Fig.2. Accuracy of Male and Female class

The following output screen generated when we run Naive Bayes computing algorithm is shown in the following Fig.3.

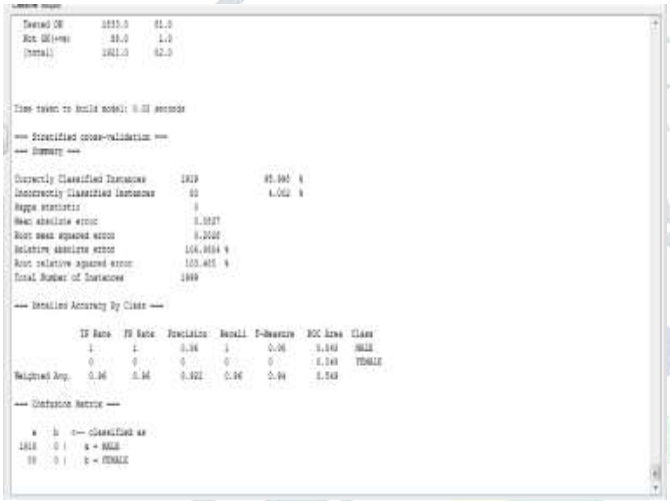

Fig.3. Classified and Unclassified Instances

5 Conclusions

There are so many parameters comparing the performance and accuracy of an algorithm. The objective of this research paper is for the classification and prediction of blood donors according their sex and blood group. In this paper we have discussed that how a Naive-Bayes soft computing algorithm can be used in knowledge discovery for classification and prediction. During this work a data mining model is developed and tested for extracting knowledge of blood donor's classification which can be used to support certain kind of decisions in blood bank organization. The blood donor's dataset collected from an authentic government blood bank centre. The experimental outcomes represents that the generated classification rules carried out perfectly with accuracy rate 97.5588%. In the next paper the soft computing techniques with KDD will be implemented on the real world dataset for predicting the blood donor's conduct and mindset.

References

[1]. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (2004).

[2]. Mitchell, T.M.: Machine Learning. McGraw Hill International Editions, New York (1997)

[3]. Jang, J.S.R., Sun, C.T., Mizutani, E.: Neuro-Fuzzy and Soft Computing. Pearson Education, London (2004)

[4]. W. Boonyanusith and P. Jittamai, "Blood Donor Classification Using Neural Network and Decision Tree Techniques," in *Proceedings of the World Congress on Engineering and Computer Science*, 2012.

[5]. A. Bhardwaj, A. Sharma, and V. K. Shrivastava, "Data Mining Techniques and Their Implementation in Blood Bank Sector – A Review," *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. August, pp. 1303–1309, 2012.

[6]. K. Smith and J. Gupta, "Neural networks in business: Techniques and applications for the operations researcher," *Computers and Operations Research*, vol. 27, pp. 1023–1044, 2000

[7]. S. McKechnie, "Integrating intelligent systems into marketing to support market segmentation decisions," *Intelligent Systems in Accounting, Finance and Management*, vol. 14, pp. 117–127, 2006.

[8]. C. Bishop, *Neural networks for pattern recognition*. New York: Oxford University Press, 1999.

[9]. R. Sharda, "Neural networks for the MS/OR analyst: An application bibliography," *Interfaces*, vol. 24, pp. 116–130, 1994.

[10]. J. Zahavi and N. Levin, "Applying neural computing to target marketing," *Journal of Direct Marketing*, vol. 11, pp. 5–22, 1997.

[11]. T. Reutterer and B. Natter, "Segmentation-based competitive analysis with MULTICLUS and topology representing networks," *Computer sand Operations Research*, vol. 27, pp. 1227–1247, 2000.

[12]. J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol.1, no. 1, pp. 81-106, 1986.

[13]. Ankita R. Borkar and Dr. Prashant R. Deshmukh , ―*NaïveBayes Classifier for Prediction of Swine Flu Disease*‖, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5,Issue 4, pp. 120-123, 2015.

[14]. Ms.Rupali R.Patil, ―*Heart Disease Prediction System using Naive Bayes and Jelinek-Mercer smoothing*‖, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6787-6792,2014.

[15]. Shweta Kharya, Shika Agrawal and Sunita Soni, *Naive Bayes Classifiers:A Probabilistic Detection Model for Breast Cancer*‖,International Journal of Computer Applications (0975 – 8887) Volume 92 – No.10, pp.26-31, 2014

[16]. UCI Machine Learning Repository, http://ics.uci.edu/mlearn/MLRepository.html

[17]. Stephanie J. Hickey, ―*Naive Bayes Classification of Public Health Data with Greedy Feature Selection* Communications of the IIMA, Volume 13, Issue 2 Article ,pp. 87-98, 2013.

[18]. Ambica, Satyanarayana Gandi and Amarendra Kothalanka, ―*An Efficient Expert System For Diabetes By Naïve Bayesian lassifier*‖, International Journal of Engineering Trends and Technology (IJETT) –Volume 4 Issue 10, pp.4634-4639, Oct 2013