Quest Journals
Journal of Research in Environmental and Earth Science
Volume 3~ Issue 4 (2017) pp: 10-16
ISSN(Online) : 2348-2532
www.questjournals.org

Research Paper

Migraine (Headaches) Disease Data Classification Using Data Mining Classifiers

[1]Rahul Deo Sah, [2]Dr. Jitendra Sheetlani, [3]Dharam Raj Kumar,
[4]Indra Nath Sahu
[1.3.4]Research Scholar, Cse, Ssstums, Bhopal
[2]Associate Professor, Cse, Ssstums, Bhopal

Received; 30 Jan. 2017 Accepted; 18 Feb. 2017;

ABSTRACT: *The Paper has analyzed migraine headache using four classification technique K-NN, SVM, Random Forest, & Naïve Bays. Overall in 30 days a month getting migraine pain which intensity of pain between 1-10 which may be occurred in some hours, including pain of location, environmental factors and their associated symptoms. The data is being collected from headache diary. In this database, with 114 different data's having 6 different categories have been considered to conduct the experimentation with proposed migraine headache classification technique. On the whole on hundred and fourteen queries have been fired. The average success rate of classification determination for each of the categorical attributes has been classified using four classifiers. The results explicitly reveal performance improvement with proposed migraine headache classification compare to intensity of pain, environment factors and their associated symptoms which is the best results Naïve Bays Classification technique among four.*

Keywords*: SVM, Random Forest, K-NN, Naïve Bays, Headaches*

I. INTRODUCTION

Data Mining is a non-trivial extraction of implicit previously unknown and potential useful information and data. It is process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different application for examples health care industry. Now a day's health care industry generates large amount of data about patient's disease diagnosis. A major challenge facing health care industry is quality of service. Quality of services implies diagnosis disease correctly and provides effective treatments to patients. Poor diagnosis can be load to disastrous consequences which are inacceptable.

According to headache disease in many countries occur due to work overload, mental stress and many more reasons behind it. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made based on doctor's experience and knowledge. This leads to unwanted results and excessive medical loss of treatments provided to patients. Therefore the prediction of different types of headache blood pressure, blood sugar.

The data mining provides lots of technique for mine data is several field the technique of mining as association rule mining, clustering technique, classification technique. Emerging technique such as called ensemble classification technique. The process of ensemble classifier increase the classification rate and improved the majority voting of classification technique for individual classification algorithms KNN, decision tree and support vector machine. The new paradigms of ensemble classifiers are cluster oriented ensemble technique for classification of data.

II. LITERATURE SURVEY

Large number of work is carried out in finding out efficient methods of medical diagnosis. Our work is an attempt to predict efficiently diagnosis with reduced of factor (i.e attributes)that contributes more towards the symptoms of Migraine disease using classification Kids get headaches and migraines Many adults with headaches started having them as kids -- in fact, 20% of adult headache sufferers say their headaches started before age 10, and 50% report their headaches started before age 20. Headaches are very common in children and adolescents. In one study, 56% of boys and 74% of girls between the ages of 12 and 17 reported having had a headache within the past month. By age 15, 5% of all children and adolescents have had migraines and 15% have had tension headaches. Many parents worry that their child's headache is the sign of

a brain tumor or serious medical condition, but most headaches in children and adolescents are not the result of a serious illness.

All headaches are considered primary headaches or secondary headaches. Primary headaches are not associated with other diseases. Examples of primary headaches are migraine headaches, tension headaches, and cluster headaches. Secondary headaches are caused by other diseases. The associated disease may be minor or major. There are many types of headaches. Tension headaches are the most common type of primary headache. As many as 90% of adults have tension headaches. Tension headaches are more common among women than men.

Migraine headaches are the second most common type of primary headache. An estimated 28 million people in the U.S. have migraine headaches. Migraine headaches affect children as well as adults. Before puberty, boys and girls are affected equally by migraine headaches, but after puberty more women than men have them. Migraine often goes undiagnosed or is misdiagnosed as tension or sinus headaches.

Cluster headaches are a rare but important type of primary headache, affecting mainly men. The average age of cluster headache sufferers is 28-30 years of age, although headaches may begin in childhood.

Secondary headaches may result from innumerable conditions, ranging from life-threatening ones such as brain tumors, strokes, meningitis, vacuities, and subarachnoid hemorrhages to less serious but common conditions such as withdrawal from caffeine, sinus infection (sinusitis), and discontinuation of analgesics (painkillingmedication). Pregnancy sometimes causes headaches. Many people suffer from "mixed" headache disorders in which tension headaches or secondary headaches may trigger migraine.

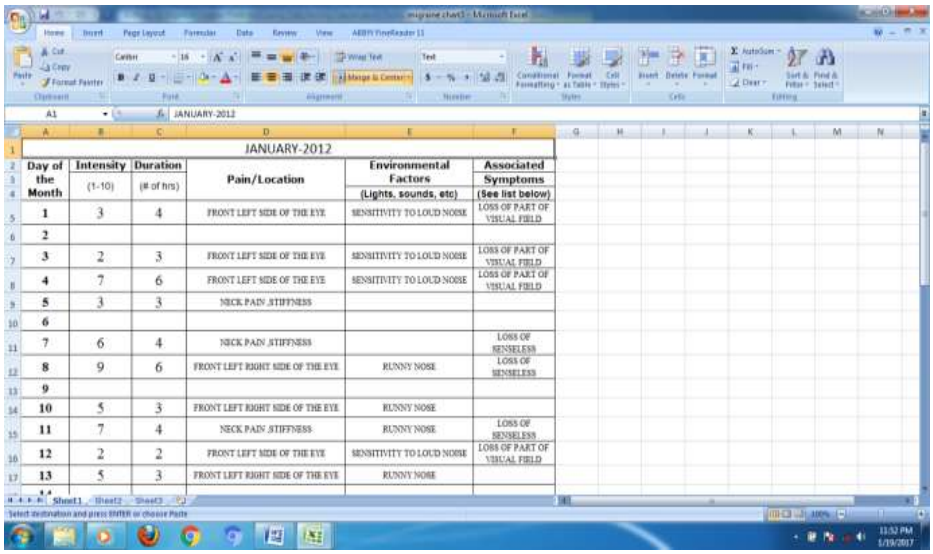
The treatment of the headache depends on the type and severity of the headache and on other factors, such as the age of the patient. According to month of disease database I can classify the dataset using different types of data mining technique. Data Mining presents many challenges as knowledge is automatically extracted from datasets, especially when data are complex in nature with a many of hundreds of variables and relationships among those variables that vary in time, space or both often with measure of uncertainty as is common within medicine Cios and Moore (2001) indentified a numbers of unique features of medical data mining, including the use of imaging amounts of unstructured nature of free text within records. Data Township and the distributed nature of data. The privacy and security concern of patients requiring anonymous data used, where possible together with the difficulty in making mathematical characterization of the domain. Data Mining are better describing as exercise in machine learning where the main issue are for example, discovering the complexity of relationship among data items or making prediction in light of uncertainty ,rather than data mining in large possibly distributed volumes of data that are also highly complex.

III. RELATED WORK

Now the research concern the data is being collected from headache diary and it's a lot of variable and their attribute. In the database having information about headaches datasets for analysis of different attributes and their variables using selection technique, visualization technique to classify the dataset using classification technique. Support vector machine , k nearest neighbor , Random Forest Classification, Naïve Bayes, Classification Tree.[4][7][10]In this regards some specific prediction for headache (Migraine). For that making questionnaire:-

1. What was the date/time of the week when I had the headache?
2. What did I do just before the headache
3. Did I eat something?
4. Where I physically active?
5. Where I stressed?
6. Did I have less or more sleep than usual?
7. Did I take medication for something other than headache
8. What else did feel during headache?
9. Did my sight changes?
10. Did I see bright lights or blind sports?
11. Did I feel nauseated?
12. Did I vomit?
13. Did I feel dizzy or confused?
14. How much my head hurt during my headache?
15. Where does my head hurt during my headache?
16. Did I have any changes in my vision?
17. What did I do to make myself feel better?
18. What date/time did the headache end?

Mentioned above question are making database. Some specific table and its given different type of tables attributes and their variables for analysis migraine headaches



JANUARY-2012

Day of the Month	Intensity (1-10)	Duration (# of hrs)	Pain/Location	Environmental Factors (Lights, sounds, etc)	Associated Symptoms (See list below)
1	3	4	FRONT LEFT SIDE OF THE EYE	SENSITIVITY TO LOUD NOISE	LOSS OF PART OF VISUAL FIELD
2					
3	2	3	FRONT LEFT SIDE OF THE EYE	SENSITIVITY TO LOUD NOISE	LOSS OF PART OF VISUAL FIELD
4	7	6	FRONT LEFT SIDE OF THE EYE	SENSITIVITY TO LOUD NOISE	LOSS OF PART OF VISUAL FIELD
5	3	3	NECK PAIN ,STIFFNESS		
6					
7	6	4	NECK PAIN ,STIFFNESS		LOSS OF SENSELESS
8	9	6	FRONT LEFT RIGHT SIDE OF THE EYE	RUNNY NOSE	LOSS OF SENSELESS
9					
10	5	3	FRONT LEFT RIGHT SIDE OF THE EYE	RUNNY NOSE	
11	7	4	NECK PAIN ,STIFFNESS	RUNNY NOSE	LOSS OF SENSELESS
12	2	2	FRONT LEFT SIDE OF THE EYE	SENSITIVITY TO LOUD NOISE	LOSS OF PART OF VISUAL FIELD
13	5	3	FRONT LEFT RIGHT SIDE OF THE EYE	RUNNY NOSE	

Table-I

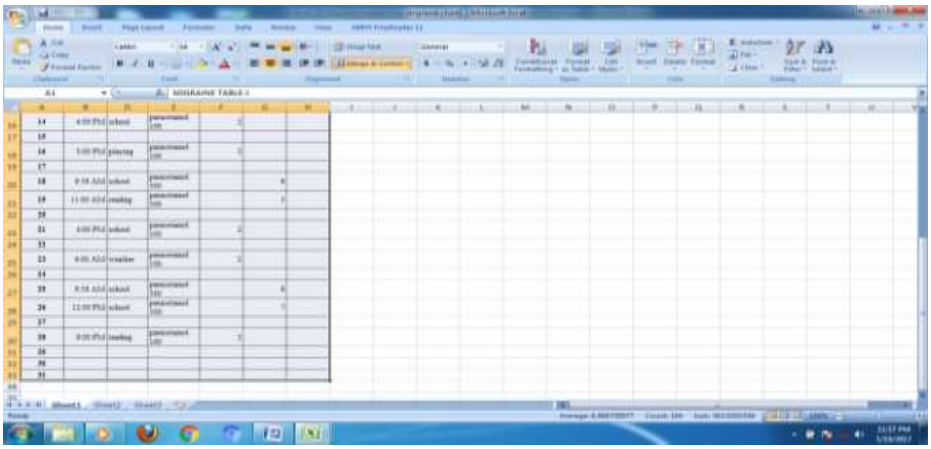
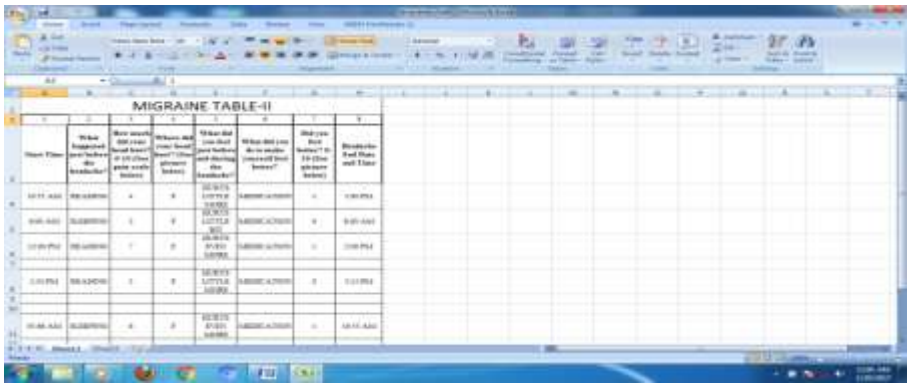


Table-II



MIGRAINE TABLE-II

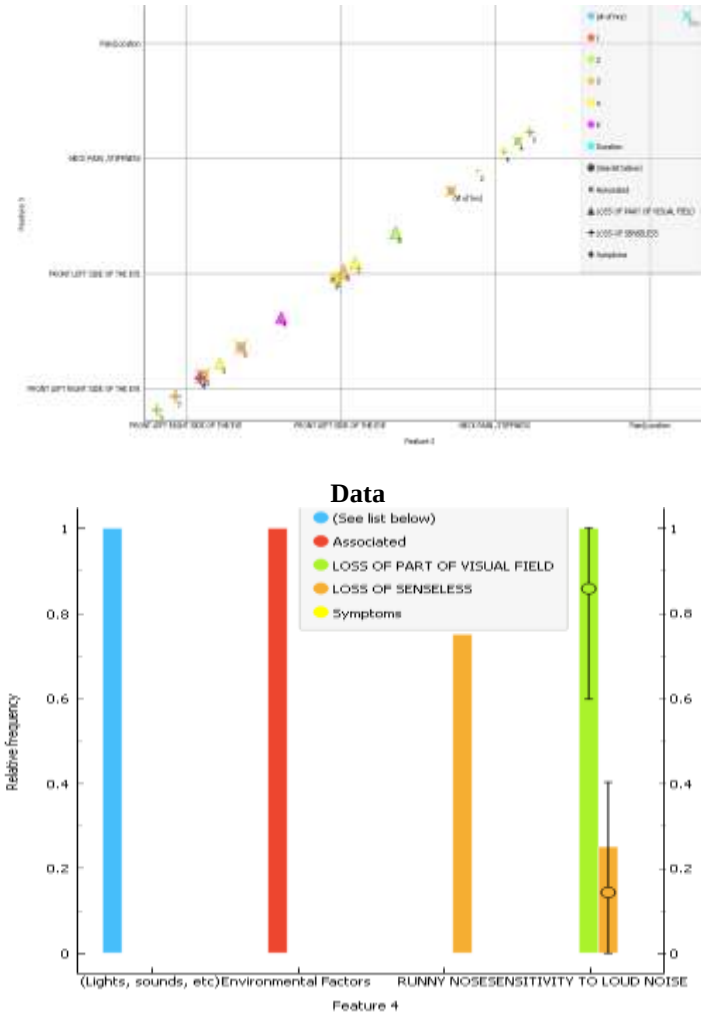
Tabl e-III

A classifier is a supervised function where the learned (target) attribute is categorical ("nominal").It is used after the learning process to classify new records (data) by giving them the best target attribute (prediction).The target attribute can be one of k class membership.[2][3][4][5].The closest neighbor (NN) rule distinguishes the classification of unknown data point on the basis of its closest neighbor whose class is already known. M. Cover and P. E. Hart purpose k nearest neighbor (KNN) in which nearest neighbor is computed on the basis of estimation of k that indicates how many nearest neighbors are to be considered to characterize class of a sample data point. It makes utilization of the more than one closest neighbor to determine the class in which the given data point belongs to and and consequently it is called as KNN.[9][10] These data

samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique. T. Bailey and A. K. Jain enhance KNN which is focused on weights. The training points are assigned weights according to their distances from sample data point. But at the same time the computational complexity and memory requirements remain the primary concern dependably. To overcome memory limitation size of data set is reduced. For this the repeated patterns which don't include additional data are also eliminated from training data set. To further enhance the information focuses which don't influence the result are additionally eliminated from training data set. The NN training data set can be organized utilizing different systems to enhance over memory limit of KNN. The KNN implementation can be done using ball tree, k-d tree, nearest feature line (NFL), principal axis search tree and orthogonal search tree. The tree structured training data is further divided into nodes and techniques like NFL and tunable metric divide the training data set according to planes. Using these algorithms we can expand the speed of basic KNN algorithm. Consider that an object is sampled with a set of different attributes. Assuming its group can be determined from its attributes; different algorithms can be used to automate the classification process. In pseudo code k-nearest neighbor classification algorithm can be expressed,

K ← number of nearest neighbors
For each object *X*in the test set do
calculate the distance D(*X*,*Y*) between *X* and every object *Y* in the training set
neighborhood ← the *k* neighbors in the training set closest to *X*
X.class ← Select Class (neighborhood)

End for



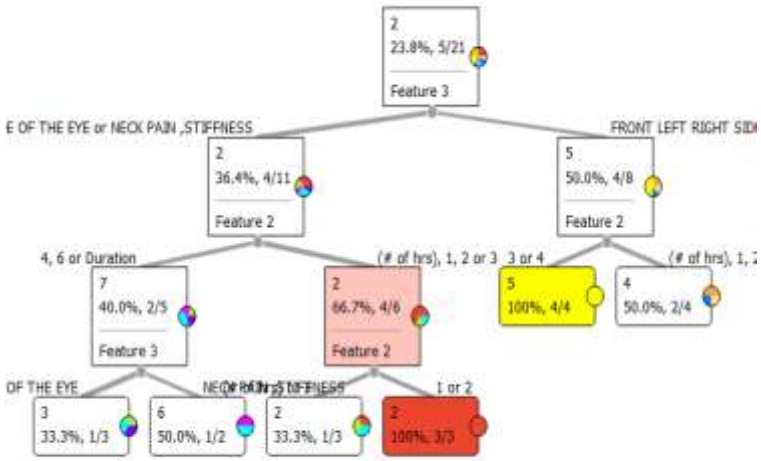
Name: kNN
Model parameters
Number of neighbours: 5 Metric: Mahalanobis Weight: Uniform
Data instances: 34 Features: Feature 2, Feature 3, Feature 4, Feature 5 Target: Feature 1

Settings
Sampling type: Stratified 2-fold Cross validation Target class: 1-10

Scores

Mehod	AUC	CA	F1	Precision
KNN	0.500	0.953	0.902	0.952
SVM	0.500	0.952	0. .907	0.952
Random Forest-	0.500	0.952	0.929	0.952
Classification				
Naive Bayes -	0.475	0.905	0.905	0.905

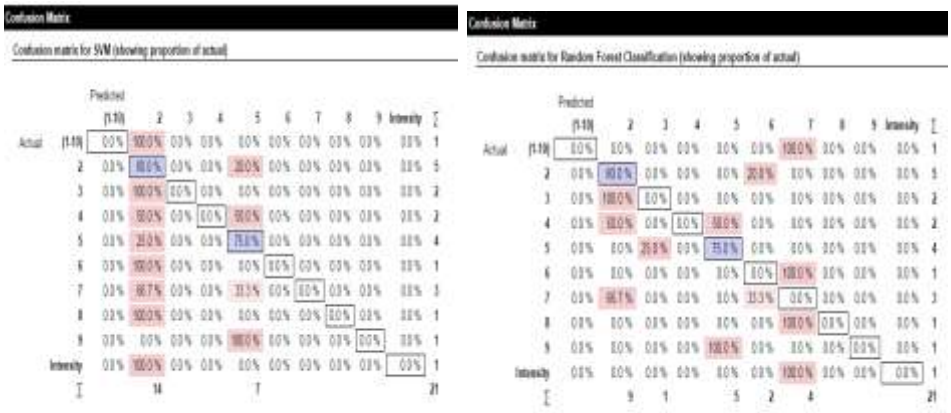
Classification is a supervised learning method to extract models describing important data classes or to predict future trends. Classification methods are largely used in machine learning, pattern recognition and artificial intelligence. Classification methods have numerous applications which includes risk analysis, credit card fraud detection, target marketing, manufacturing and medical diagnosis.[20] Our work intends to use three classifiers Decision Tree, Naive Bayes and Classification via clustering to diagnosis the presence of headache disease in patients Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Our work uses decision tree for classification. Observations show that Decision trees outperform the other two classifiers but take more time to build the model.



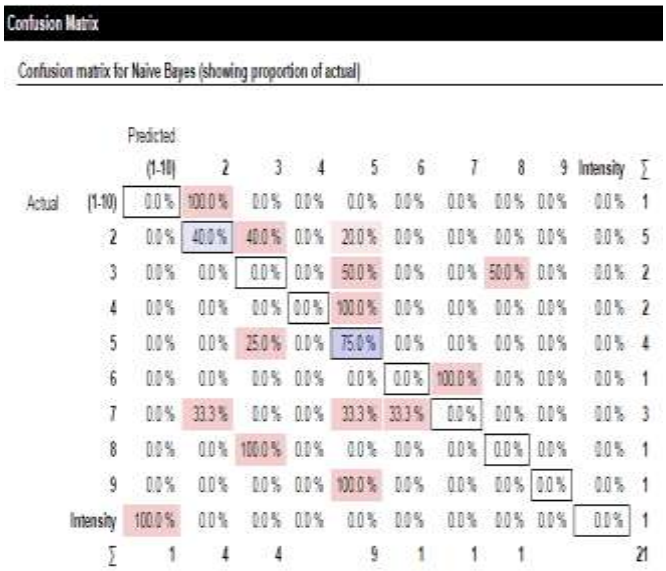
Classification Tree View

Naive Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naive Bayes performs consistently before and after reduction of number of attributes Classification via clustering: Clustering is the process of grouping similar elements.[1] This technique may be used as a preprocessing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering. Observations show that Classification via clustering performs poor even after reduction of number of attributes when compared to the other two methods. SVM have attracted a great deal of attention in the last decade and actively applied to various domains applications. SVMs are typically used for learning classification, regression or ranking function.[11][12][14][15][16][17][18]. SVM are based on statistical learning theory and structural risk minimization principal and have the aim of determining the location of decision boundaries also known as hyper plane that produce the optimal separation of classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyper plane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error. Efficiency of SVM based classification is not directly depends on the dimension of classified entities. Though SVM is the most

robust and accurate classification technique, there are several problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally expensive, as solving quadratic programming methods require large matrix operations as well as time consuming numerical computations. Training time for SVM scales quadratic ally in the number of examples, so researches strive all the time for more efficient training algorithm, resulting in several variant based algorithm.SVM can also be extended to learn non-linear decision functions by first projecting the input data onto a high-dimensional feature space using kernel functions and formulating a linear classification problem in that feature space. The resulting feature space is much larger than the size of dataset which are not possible to store in popular computers. Investigation on this issues leads to several decomposition based algorithms. The basic idea of decomposition method is to split the variables into two parts: set of free variables called as working set, which can be updated in each iteration and set of fixed variables, which are fixed at a particular value temporarily. This procedure is repeated until the termination conditions are met originally, the SVM was developed for binary classification, and it is not simple to extend it for multi-class classification problem.[20] The basic idea to apply multi classification to SVM is to decompose the multi class problems into several two class problems that can be addressed directly using several SVMs



Machine Learning is generally covers automatic computing procedures based on logical or binary operations that learn a task from a series of examples. Here we are just



Concentrating on classification and so attention has focused on decision-tree approaches in which classification results from a sequence of logical steps. These classification results are capable of representing the most complex problem given sufficient data. Other techniques such as genetic algorithms and inductive logic procedures (ILP) are currently under active improvement and its principle would allow us to deal with more general types of data including cases where the number and type of attributes may vary. Machine Learning approach aims to generate classifying expressions simple enough to be understood easily by the human and must mimic human reasoning sufficiently to provide insight into the decision process .Like statistical

approaches background knowledge may be used in development but operation is assumed without human interference.[14]

Experiments were conducted with Orange 3.3.0 tool. Data set of 34 records with 5 attributes is used. All attributes are made categorical and inconsistencies are resolved for simplicity. To enhance the prediction of classifiers, and diagnosis the migraine disease perfectly. We used some classifiers such as KNN , Naïve Bays, Random Forest and support vector machine find out the results accuracy of the classifier is naïve bays which is 0.475 AUC and Precision 0.905 so the Naïve bays is the best classifier among these.

Now we used Confusion matrixes such as SVM, Random Forest, Naive Bays among these confusion matrixes naïve bays is good classifiers for migraine headache disease.

IV. CONCLUSION

In medical field, Data Mining provides various techniques and have been widely used in clinical decision support systems that are useful for predicting and diagnosis of various diseases. These data mining classification technique used in migraine headache diseases takes less time and make process fast for the classify to Migraine diseases with good accuracy in order to improve their health. The average success rate of classification determination for each of the categorical attributes has been classified using four classifiers. The results explicitly reveal performance improvement with proposed migraine headache classification compare to intensity of pain, environment factors and their associated symptoms which is the best results Naïve Bays Classification technique among four. KNN, Naïve Bays, Random forest, SVM which help to headache is more seriously and dangerous for health which affected of any parts of the body which is symptoms associated with the headache.

REFERENCES

[1]. Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naive bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012

[2]. Venkatadri.M, Dr. Lokanatha C. Reddy a review on data mining from past to the future. International Journal of Computer Applications, 2011.

[3]. Abhishek taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[4]. Rashedur M. Rahman, Farhana Afroz, Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis, Journal of Software Engineering and Applications, 2013.

[5]. S. K. Yadav and Pal S., "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.

[6]. Breiman, L. Friedman, J.H., Olshen, R.A., Stone, C.J. Classification and Regression Trees. Chapman & Hall, New York, 1984.

[7]. Marcel A.J. van Gerven, Predicting carcinoid heart disease with the noisy-threshold classifier, Elsevier, 2007.

[8]. Mohammad Taha Khan, Dr. Shamimul Qamar and Laurent F. Massin, A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining, International Journal of Applied Engineering Research, 2012

[9]. M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu, Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2012

[10]. Ms. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte, "A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology",2012

[11]. Clark, P., Niblett, T. The CN2 Induction Algorithm. Machine Learning 1989, 3(4), pp.261-283

[12]. Dougherty, J., Kohavi, M., Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. Conf. on Machine Learning, San Francisco, 1995. pp.194-202

[13]. J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elevier, 2011.

[14]. V. Vapnik and C. Cortes , "Support Vector Network," Machine Learning, 20; 273-297, (1995).

[15]. C. J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, 2; (1998).

[16]. H. Bhavsar, A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 -2307, 2(4); (2012)

[17]. G. Wang, "A Survey on Training Algorithms for Support Vector Machine Classifiers", Fourth International Conference on Networked Computing and Advanced Information Management, 2008,IEEE.

[18]. G Madzarov, D. Gjorgievikj and I. Chorbev, " A Multi-class SVM Classifier Utilizing Binary Decision Tree", Informatica, pp. 233-241 (2009).

[19]. M. Aly, "Survey on Multiclass Classification Methods", November (2005).

[20]. Vapnik, "Statistical Learning Theory", Wiley, New York, (1998)