Data Mining tools and techniques in Medical Science

Anu Bura[1], DR. Parkash Pathak[2]
[1]*M. Tech Student, World College of Technology and Management*
[2]*Professor, World College of Technology and Management*

***Abstract*- Our research is based on the Study of various data mining techniques and tools that are used and in Medical Science. In this study we have conclude the techniques , breast cancer and Chronic and kidney Disease. The major goal of the classification technique is to predict the target class accurately for each case in the data. The present study is focused on the usage of classification techniques in the field of medical science and bioinformatics.The goal of data mining application is to turn that data are facts, numbers, or text which can be processed by a computer into knowledge or information. The main purpose of data mining application in healthcare systems is to develop an automated tool for identifying and disseminating relevant healthcare information.Breast cancer is one of the leading cancers for women in developed countries including India. It is the second most common cause of cancer death in women. The high incidence of breast cancer in women has increased significantly in the last years.Chronic- Kidney-Disease prediction using weka data mining tool and its usage for classification in the field of medical bioinformatics. It firstly classifies dataset and then determines which algorithm performs better for diagnosis and prediction of Chronic- Kidney-Disease. Prediction begins with identification of symptoms in patients and then identifying sick patients from a lot of sick and healthy ones. presents a comparative study of different data mining applications, techniques and different methodologies applied for extracting knowledge from database generated in the healthcare industry. Finally, the existing data mining techniques with data mining algorithms and its application tools which are more valuable for healthcare services are discussed in detail.**

I. INTRODUCTION:

In present days, computers have brought significant improvements to technology that lead to the creation of huge volumes of data. Moreover, the advancement of the healthcare database management systems creates a huge number of medical databases. Creating knowledge and management of large amounts of heterogeneous data has become a major field of research, namely data mining. Data Mining is a process of identifying novel, potentially useful, valid and ultimately understandable patterns in data [1]. Data mining techniques can be classified into both unsupervised and supervised learning techniques. Unsupervised learning technique is not guided by variable and does not create a hypothesis before analysis.In the present study, we have focused on the usage of classification techniques in the field of medical science and bioinformatics. Classification is the most commonly applied data mining technique, and employs a set of pre-classified examples to develop a model that can classify the population of records at large.

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides [1]. This study mainly discusses the Data Mining applications in the scientific side. Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from traditional market driven data mining applications. In this work, a detailed survey is carried out on data mining applications in the healthcare sector, types of data used and details of the information extracted. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. There are a large number of data mining applications are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management.

II. METHODOLOGY:

In order to carry out experiments and implementations WEKA is used as the data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches in the field of bioinformatics. In this work

we have used the data mining techniques to predict the survivability of Chronic-Kidney disease through classification of different algorithms accuracy.

Explorer: The explorer interface has several panels like pre-process, classify, cluster, associate, select attribute and visualize. But in this interface our main focus is on the Classification Panel.

Experimenter: This interface provides facility for systematic comparison of different algorithms on basis of given datasets. Each algorithm runs 10 times and then the accuracy gets reported.

2.1 Literature Survey of The Problem

To understand the health hazards of fluoride content on living beings, discussions were made with medical practitioners and specialists like General Dental, Neuro surgeons and Ortho specialists. We have also gathered details about the impact of high fluoride content water from World Wide Web. By analyzing all these we came to know that the increased fluoride level in ground water creates dental, skeletal and neuro problems. In this analysis we focus only on skeletal hazards by high fluoride level in Drinking water.

2.2 Data Preparation

Based on the information from various physicians and water analyst, we have prepared questionnaires to get raw data from the various fluoride impacted villages and panchayats, having fluoride level in water from 1.6mg/L to 2.4mg/L. People of different age groups with different ailments were interviewed with the help of questionnaires prepared in our mother tongue, Tamil since the people in and around the district are illiterate.

Total data collected from Villages and PanchayatsMild Skeletal Victims Moderate Skeletal Victims Osteoporosis Victims With the following classification, those who are found with one to three low symptoms are grouped as Mild victim of skeletal disease.

2.3 Clustering as the Data mining application

Clustering is one of the central concepts in the field of unsupervised data analysis, it is also a very controversial issue, and the very meaning of the concept “clustering” may vary a great deal between different scientific disciplines. However, a common goal in all cases is that the objective is to find a structural representation of data by grouping (in some sense) similar data items together. A cluster has high similarity in comparison to one another but is very dissimilar to objects in other clusters.

2.4 Weka as a data miner tool

In this paper we have used WEKA (to find interesting patterns in the selected dataset), a Data Mining tool for clustering techniques.. The selected software is able to provide the required data mining functions and methodologies. The suitable data format for WEKA data mining software are MS Excel and ARFF formats respectively. Scalability-Maximum number of columns and rows the software can efficiently handle. However, in the selected data set, the number of columns and the number of records were reduced. WEKA is developed at the University of Waikato in New Zealand.

2.5 Clustering in WEKA

The classification is based on supervised algorithms. This algorithm is applicable for the input data. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.. The Cluster tab is also supported which shows the list of machine learning tools. These tools in general operate on a clustering algorithm and run it multiple times to manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with WEKA.

2.6 Learning Algorithm

This paper consists of an unsupervised machine learning algorithm for clustering derived from the WEKA data mining tool. Which include :

III. K-MEANS

The above clustering model was used to cluster the group of people who are affected by skeletal fluorosis at different skeletal disease levels and to cluster the different water sources using by the people which are causes for skeletal fluorosis in krishnagiri district.

PRELIMINARY Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires classes to be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier

training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. In Classification, training examples are used to learn a model that can classify the data samples into known classes. The Classification process involves following steps:

- Create training data set.
- Identify class attribute and classes.
- Identify useful attributes for classification (Relevance analysis).
- Learn a model using training examples in Training set.
- Use the model to classify the unknown data samples.

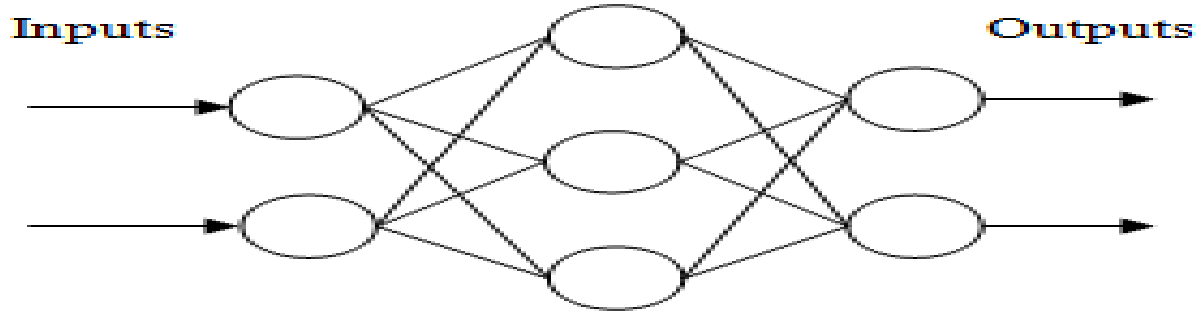
Classifiers Used In this work six classification algorithms have been used for classification task to study their classification accuracy and performance over the Chronic-Kidney-Disease data set. The classifiers in Weka have been categorized into different groups such as Bayes, Functions, Lazy, Rules, Tree based classifiers etc. A good mix of algorithms has been chosen from these groups which are used in distributed data mining. They include Naive Bayes (from Bayes), Multilayer Perceptron, SVM, J48, Conjuctive rule and Decision Table. The following sections explain a brief about each of these algorithms.

3.1 Naïve Bayes Classifier

It is one of the fastest statistical classifier algorithm works on probability of all attribute contained in data sample individually and then classifies them accurately. It is used to predict class membership probabilities i.e. probability about the tuple that belongs to the particular class or not. Bayesian classification is based on Bayes theorem. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector X= (x1, x2, ... ,xn) representing some n features (independent variables), it assigns to this instance probabilities p(Ck | x1, ..., xn) for each of k possible outcomes or classes. The problem with the above formulation is that if the number of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable. Using Bayes' theorem, the conditional probability can be decomposed as $pCkX = (X|Ck)(X)$ In other words, using Bayesian probability terminology, the above equation can be written as Posterior = $prior \times likelihoodevidence$

3.2 Multilayer Perceptron

It is the most popular network architecture in today's world. Each unit performs a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are arranged in a layered feed forward topology. The network has a simple input-output model, with the weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptron are the design specification of the number of hidden layers and the number of units in these layers. Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a back propagation neural network with one or more layers between input and output layer. The following diagram illustrates a perceptron network with three layers.



3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The standard SVM takes a set of input data and predicts, for each given input, which of two

possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier.

3.4 J48

J48 classifier is a simple C4.5 decision tree for classification. It is supervised method of classification. It creates a small binary tree. It is univarient decision tree. It is an extension of ID3 algorithm. In this classifier Divide and Conquer approach is used to classify the data. It divides the data into range based on the attribute value for that value that are found in training sample. As this approach is range based and univarient [11], it does not perform better than multivarient approach. As this is decision tree approach it is very much useful in predicting the values. J48 accuracy of correctly classified instance is much more than that of the other algorithms which are univarient in nature [10].

3.5 Conjuctive Rule

It is a decision-making rule in which the intending buyer assigns least values for a number of factors and discards any result which does not meet the bare minimum value on all of the factors i.e. a superior performance on one factor cannot recompenset for deficit on another. Conjuctive rule uses the AND logical relation to correlate stimulus attributes. Conjuctive rule is a simple well interpretable 2-class classifier.

IV. DECISION TABLE

A decision table is a predictive modeling tool that performs classification. It incorporates an inducer (an algorithm for generating decision table models), and a visualizer. Unlike the evidence model, the Decision Table model does not assume that the attributes are independent. A decision table is a hierarchical breakdown of the data, with two attributes at each level of the hierarchy. The Decision Table inducer identifies the most important attributes (columns) for classifying the data, and the accompanying visualizer displays the resulting model graphically. It summarizes the dataset with a decision table which contains the same number of attributes as the original dataset. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table.

4.1 Characteristics required for Classification

Algorithm In this work, we have focused on the following three measures namely correctly classified instances, incorrectly classified instances, and accuracy. (i) Correctly classified instance: These are the instances which are correctly classified by any classification algorithm. Percentage of correctly classified instances is called as accuracy. (ii) Incorrectly classified instances: These instances are not correctly classified by the algorithm. Sometimes it is observed that the data which is incorrectly classified may contain inconsistent data, noisy data or data out of scope. (iii) Accuracy: Accuracy is how a measured value is closed to the true value. The general formula is given below: Accuracy = $Tp+TnP+N$ (1) where, Tp indicates True positive, Tn indicates True negative, P indicates total positive, N indicates total negative. And P = Tp + Fn , N = Fp + Tn. In classification system, the algorithm with highest accuracy will be selected for the prediction. Accuracy of the algorithm varies according to the dataset used. So before using the algorithms for prediction system, we must check the accuracy of the algorithm. So it will reduce the cost of doing trial and error of using algorithms in the prediction system.

4.2 Performance Evaluation

10-fold cross validation technique is used to evaluate the performance of classification methods, Data set is randomly sub divided into ten equal sized partitions. Among the partitions nine of them are used as training set and the remaining one is used as a test set. Evaluation of performance is compared using Mean absolute error, Rootmean squared error, Receiver Operating Characteristic (ROC) Area and Kappa statistics. Large test sets gives a good assessment of the classifier's performance and small training sets which result in a poor classifier.

4.3 Kappa Statistics

Kappa Statistics measure degree of agreement between two sets of categorized data. Kappa result varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement. Kappa is a normalized value of agreement for chance of agreement. K = $PA\ -(E)1-P(E)$ Where P(A) = percentage of agreement P(E) = chance of agreement. If K =1 agreement is perfect between the classifier and ground truth. If K=0 indicates there is a chance of agreement.

4.4 Mean Absolute Error (MAE)

The mean absolute error (MAE) is a quantity used to measure predictions of the eventual outcomes. The mean absolute error is given by MAE = $1n$ $|fi-yi|ni$=1 = $1n$ $|ei|ni$=1 The mean absolute error is an average of the absolute errors ei = | fi – yi |, where fi = prediction ,yi = true value.

4.5 Root Mean Squared Error (RMSE)

Root mean squared error is the square root of the mean of the squares of the values. It squares the errors before they are averaged [18] and RMSE gives a relatively high weight to large errors. The RMSE Ei of an individual program i is evaluated by the equation:

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{P_{(i,j)}-T_j}{T_j}\right)^2}$$

where, P(i,j) = the value predicted by the individual program i = fitness case Tj =the target value for fitness case j.

4.6 Receiver Operating Characteristic (ROC) Area

ROC Area is defined as area under the ROC curve which is the probability of randomly chosen positive instance that is ranked above randomly chosen negative one. Receiver Operating Characteristic represents test performance guide for classifications accuracy of diagnostic test based on: excellent (0.90-1), good (0.80-0.90), fair (0.70-0.80), poor (0.60-0.70), fail (0.50–0.60).

CLASSIFICATION OF ATTRIBUTES

V. DISCUSSION AND RESULT

We have conducted two experiments based on the dataset with all above discussed classification algorithms; first without using feature selection and second with using Genetic search feature selection. First the results of the classification algorithms based on parameters such as accuracy of classification, kappa statistics, MAE, RMSE, model building time, model testing time, and ROC are shown in the following Table 2, where model building time and model testing time are generated by WEKA Tool itself during classification.

5.1 Attributes selection

First of all, we have to find the correlated attributes for finding the hidden pattern for the problem stated. The WEKA data miner tool has supported many in built learning algorithms for correlated attributes. There are many filtered tools for this analysis but we have selected one among them by trial.[5]

Totally there are 520 records of data base which have been created in Excel 2007 and saved in the format of CSV (Comma Separated Value format) that converted to the WEKA accepted of ARFF by using command line premier of WEKA.

The records of data base consist of 15 attributes, fromwhich 10 attributes were selected based on attribute selection in explorer mode of WEKA 3.6.4.

S.NO.	Attributes	Data Type
01.	Name	Text
02.	Age	Numeric(Integer)
03.	Education	Text
04.	Sex	Character
05.	Fluoride Level	Numeric(Real)
06.	Profession	Text
07.	Praganancy status	Boolean
08.	Drinking water	Text
09.	Duration	Numeric(Integer/Real)
10.	Known status of fluoride	Boolean
11.	Neck Pain	Numeric(Binary)
12.	Joint Pain	Numeric(Binary)
13.	Body Pain	Numeric(Binary)
14.	Foot Neck Pain	Numeric(Binary)
15.	Disease Level	Text

We have chosen Symmetrical random filter tester for attribute selection in WEKA attribute selector. It listed 14 selected attributes, but from which we have taken only 8 attributes. The other attributes are omitted for the convenience of analysis of finding impaction among peoples in the district

S.NO.	Attributes	Data Type
01.	Age	Numeric(Integer)
02.	Education	Text
03.	Fluoride Level	Numeric(Real)
04.	Drinking water	Text
05.	Duration	Numeric(Integer/Real)
06.	Neck Pain	Numeric(Binary)
07.	Joint Pain	Numeric(Binary)
08.	Body Pain	Numeric(Binary)
09.	Foot Neck Pain	Numeric(Binary)
10.	Disease Level	Text

5.2 K-Means Method

The k-Means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed the cluster's centroid or center of gravity.

The k –Means algorithm proceeds as follows

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterated until the criterion function converges. Typically, the square-error criterion is used, defined as [2] [3] [4]$E=\Sigma K\ \Sigma\ i\ |p-mi|2$

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster Ci . In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible

K-Means algorithm

Input;

= k:the number of clusters,

D:a data set containing n objects

Output: A set of k clusters.

Method:

arbitrarily choose k objects from from D as the initial cluster centers;

(2) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(3) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(4) until no change;

```
=== Run information ===

Evaluator:   weka.attributeSelection.SymmetricalUncertAttributeEval
Search:     weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:    FORMAT OF 1-520 SKELETAL-weka.filters.unsupervised.attribute.Remove-R1
Instances:   520
Attributes:  15
        Name
        Age
        Education
        Sex
        FL
        Profession
        Pragnancy status while interview
        Drinking water type
        Duration of  drinking water used in years
        Known status of fluoride impact
        Neck Pain
        Joint Pain
        Body Pain
        Food Neck Pain
        Disease Level
Evaluation mode:   evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
	Attribute ranking.

Attribute   Class (nominal): 15 Disease Level):
	Symmetrical Uncertainty Ranking Filter

Ranked attributes:
 0.42554  13 Body Pain
 0.39888  12 Joint Pain
 0.37011  11 Neck Pain
 0.29908   1 Name
 0.24185  14 Food Neck Pain
 0.11147   2 Age
 0.09357   6 Profession
 0.09249   3 Education
 0.07813   9 Duration of  drinking water used in years
 0.01282   7 Pragnancy status while interview
 0.01263  10 Known status of fluoride impact
 0.01133   8 Drinking water type
 0.00667   4 Sex
 0         5 FL

Selected a       12       11        1       14        2        6        3        9        7       10        8        4     5:14
```

Suppose that there is a set of objects located in space as depicted in the rectangle. Let k = 3; i.e., the user would like the objects to be partitioned into three clusters.

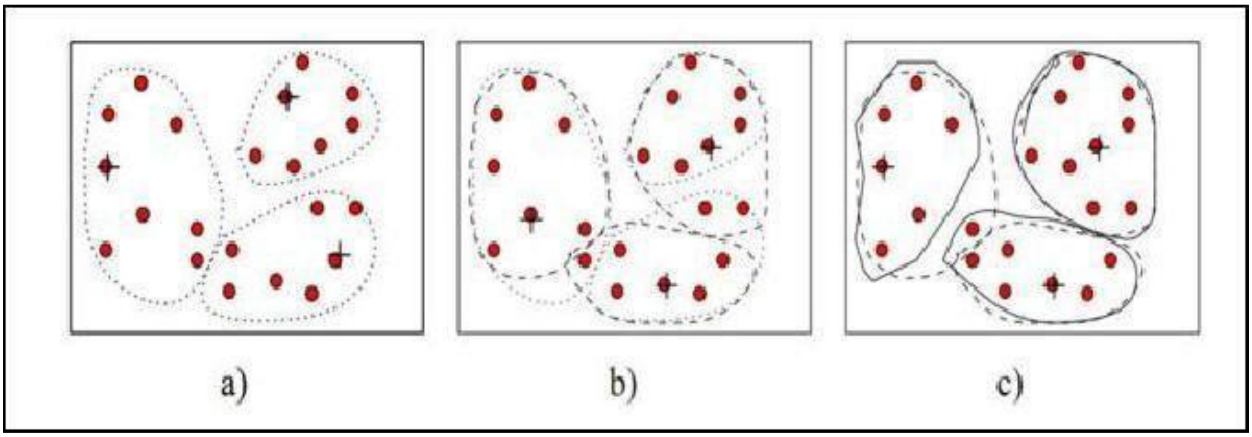
According to the algorithm above we arbitrarily choose three objects as the three initial cluster centers, where cluster centers are marked by a "+". Each objects is distributed to a cluster based on the cluster center to which it is the nearest. Such a distribution forms encircled by dotted curves.

Next, the cluster centers are updated. That is the mean value of each cluster which is recalculated

based on the current objects in the cluster. Using the new cluster
centers, the objects are redistributed to the clusters based on which cluster center is the nearest. Such a redistribution forms new encircled by dashed curves.This process iterates. The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as iterative relocation. Eventually, no redistribution of the objects in any cluster occurs, and so the process terminates. The resulting cluster is returned by the clustering process.

K-Means in WEKA

The learning algorithm k-Means in WEKA 3.6.4 accepts the training data base in the format of ARFF.



It accepts the nominal data and binary sets. So our attributes selected in nominal and binary formats naturally. So there is no need of preprocessing for further process. We have trained the training data by using the 10 Fold Cross Validated testing which used our trained data set as one third of the data for training and remaining for testing.

After training and testing this gives the following results.

Euclidean distance

K-means cluster analysis supports various data types such as Quantitative, binary, nominal or ordinal, but do not support categorical data. Cluster analysis is based on measuring similarity between objects by computing thedistance between each pair. There are a number of methods for computing distance in a multidimensional environment. Distance is a well understood concept that has a number of simple properties. Distance is always positive Distance from point x to itself is always zero Distance from point x to point y cannot be greater than the sum of the distance from x to some other point z and distance from z to y. Distance from x to y is always the same as from y to x. It is possible to assign weights to all attributes indicating their importance. There are number of distance measures such as Euclidean distance, Manhattan distance and Chebychev distance. But in this analysis Weka tool used Euclidean distance. Euclidean distance of the difference vector is most commonly used to compute distances and has an intuitive appeal but the largest valued attribute may dominate the distance. It is therefore essential that the attributaes are properly scaled. Let the distance between two points x and y be D(x,y). D(x,y) (Σ(xi-yi)2)1/2

Clustering of Disease Symptoms

The collected disease symptoms such as Neck pain, Joint pain, Body \pain, Foot Neck as raw data, supplied to kmeans method is being carried out in weka using Euclidean distance method to measure cluster centroids. The result is obtained in iteration 12 after clustered. The centroid cluster points are measured based on the diseases symptoms and the water they are drinking. Based on the diseases symptoms in raw data the kmeans clustered two main clustering units. From the confusion matrix above we came to know that the district mainly impacted by skeletal osteoporosis.

VI. CONCLUSION

The main objective of this chapter is to predict chronic kidney disease. We have used six algorithms i.e. Naive Bayes, Multilayer Perceptron, SVM, J48, Conjuctive Rule and Decision Table for our experiments. These algorithms are implemented using WEKA data mining tool to analyze accuracy which is obtained after running these algorithms in the output window. These algorithms have been

compared with classification accuracy to each other on the basis of correctly classified instances, time taken to build model, time taken to test the model, mean absolute error, Kappa statistics and ROC Area. In the experiments Multilayer perceptron algorithm gives better classification accuracy and prediction performance to predict chronic kidney disease (CKD) using relevant dataset available at UCI machine learning repository.

Data mining applied in health care domain, by which the people get beneficial for their lives. As the analog of this research we found out that the meaningful hidden pattern from the real data set collected the people impacted in Krishnagiri district is by drinking high fluoride content of potable water. By which we can easily know that the people do not get awareness among themselves about the fluoride impaction. If it continues in this way, it may lead to some primary health hazards like Kidney failure, mental disability, Thyroid deficiency and Heart disease.

REFERENCES

[1] Jain, M. Murty, and . Flynn, "Data clustering: A review," A M Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[2] Jiawei Han and MichelineKamber – Data mining concepts and Techniques. - Second Edition –Morgan Kaufmann Publishers

[3] ArunK.Pujari –Datamining Techniques – University Press.

[4] Introduction to Datamining with case studies - G.K.Gupta PHI. Fuzzy Models for Social Scientists - W.B.VasanthaKandasamy (e-book :http:mit.iitm.ac.in)

[5] BerrryMjLinoff G Data mining [10] Professionals statement calling for an Techniques: for Marketing, Sales and Customer support USA.Wiley,1997.

[6] Weka3.6.4 data miner manual.

[7] Water Quality for Better Health – TWAD Released Water book.

[8] Data mining Learning models and Algorithms for medical applications – White paper - PlamenaAndreeva, Maya Dimibova, Petra Radeve

[9] Elementary Fuzzy Matrix Theory and End to water Fluoridation – Conference Report (www.fluoridealert.org)

[11] Analysis of Liver Disorder Using Data mining algorithms - Global Journal of computer science and Technology l.10 issue 14 (ver1.0) November 2010 page 48.

[12] The WEKA Data Mining Software: An Update, Peter Reutemann, Ian H. Witten, Pentaho Corporation, Department of Computer Science