AN ANALYSIS ON THE IMPACT OF HIGH FLUORIDE LEVELS IN POTABLE WATER IN HUMAN HEALTH USING CLASSIFICATION DATAMINING TECHNIQUE

T. Balasubramanian
Department of Computer Science,
Sri VidyaMandir Arts and Science College,
Uthangarai(PO)-635 207, Krishnagiri(Dt),
Tamilnadu, India.
balaeswar123@gmail.com

R. Umarani
Department of Computer Science,
Sri Saradha College for Women (Autonomous),
Salem-636 016, Tamilnadu, India.
umainweb@gmail.com

Abstract

Data Mining is the process of extracting information from large data sets through using algorithms and Techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems. Traditional data analysis methods often involve manual work and interpretation of data which is slow, expensive and highly subjective Data Mining, popularly called as knowledge discovery in large data, enables firms and organizations to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System.[4]

This article explores data mining techniques in health care. In particular, it discusses data mining and its application in areas where people are affected severely by using the under- ground drinking water which consist of high levels of fluoride in Krishnagiri District, Tamil Nadu State, India. This paper identifies the risk factors associated with the high level of fluoride content in water, using classification algorithms and finds meaningful hidden patterns which give meaningful decision making to this socio-economic real world health hazard.

Keywords: Data mining, Fluoride affected people, Classification algorithms, J48. Naïve Bayes.

1. Introduction

Fluoride ion in drinking water ingestion is useful for Bone and Teeth development, but excessive ingestion causes a disease known as Fluorosis. The prevalence of Fluorosis is mainly due to the consumption of more Fluoride through drinking water. Though the different forms of fluoride exposure is important, if it exceeds or decrease from the required level it is risk of fluoride – prone diseases. [8]

Fluorosis was considered to be a problem related to Teeth only. But it now has turned up to be a serious health hazard. It seriously affects Bones and problems like Joint pain, Muscular Pain, etc. are its well-known manifestations. It not only affects the body of a person but also renders them socially and culturally crippled.

The goal of this paper by using the classification algorithms as a tool of data mining technique to find out the volume of people affected by the high fluoride content of potable water.

Fig.1 Skeletal Osteoroposis by Fluoride

2. MATERIALS AND METHODS

2.1 Literature Survey of the Problem

To understand the health hazards of fluoride content on living beings, discussions were held 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 in water from World Wide Web [8]. By

ISSN: 2349 - 6363

analyzing all these we came to know that the increased fluoride level in ground water create dental, skeletal and neuro problems. In this analysis we focus only on skeletal hazards by high fluoride level in drinking water. Level of fluoride content in water in different regions of Krishnagiri District was obtained from Water Analyst from TWAD. Based on the recommendations of WHO which released a water table[6], the Tamil Nadu Water And Drainage Board (TWAD) suggested the normal content of fluoride in drinking water should not be above 1.5 mg/L.[6]

The water table also shows the contents of minerals and associated health hazards. We found out that Krishnagiri District of Tamil Nadu in India is most affected by fluoride level in drinking water by naturally surrounded hills in the District. TWAD have analyzed the sample ground potable water from various regions of Krishnagiri District and maintained a table of High level fluoride (1.6mg/L to 2.4mg/L) contaminated ground drinking water of panchayats and villages list in this District. We have concludedthat many village people of Krishnagiri District are severely affected by ground potable water. So we have decided to make a survey and to find out the combination of diseases which are possibly affected mostly by high fluoride content indrinking water.

2.2 Data Preparation

Based on the information from various physicians and water analyst of TWAD, we have prepared questionnaire to get raw data from too many villagers who were affected with high level fluoride in drinking water from 1.6mg/L to 2.4mg/L.[6] People of different age groups with different ailments were interviewed based on the questionnaire prepared in our mother tongue i.e. Tamil since the people in and around the district are maximum illiterate and not studied upto the level of understanding other languages.

Total data collected from Villages

$$\left.\begin{array}{ll}\textbf{Men} & \textbf{251 (48\%)} \\ \textbf{Women} & \textbf{269 (52\%)}\end{array}\right\} \textbf{520}$$

As per the opinion and findings ofmedical practitioners, while analyzing the data for classification, the degree of symptoms of diseases are placed in several compartments as under

Mild Skeletal Victims
Moderate Skeletal Victims
Osteoroposis Victims

From the above, the status and degree of diseases classified as under with sample table.

Those who are found with one to three low symptoms are grouped as Mild victims skeletal disease.

Those who are found with four low symptoms or one to three medium and one high symptoms are grouped as Moderate victims of skeletal disease.

Those who are foundwith more than two medium symptoms are grouped as osteoporosis victims of skeletal disease.

Table 1: Sample classification of symptoms of diseases

Neck pain	Joint pain	Body Pain	FootNeck Pain	Class
Low	Low	--	--	Mild Skeletal
Low	Low	Low	--	Mild Skeletal
Low	Low	Low	Low	Mildto Moderate Skeletal
Low	Medium	Low	Medium	Moderate Skeletal
Low	Medium	Low	High	Moderate Skeletal
Low	Medium	Medium	-Medium	Osteoroposis

2.3 Classification as the Data mining application

Classification is a form of data analysis that can be used to extract models describing important data classes. Such analysis will provide us a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels. Classification have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis.[2]

2.4 WEKA tool

In this paper we have usedWEKA(to find interesting patterns in the selected dataset)**,** a Data Mining tool for classification 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. “WEKA” stands for the Waikato Environment of Knowledge Analysis. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA expects the data to be fed into be in ARFF format (Attribution Relation File Format).[9]

WEKA has two primary modes: experiment mode and exploration mode .The exploration mode allows easy access to all of WEKA's data preprocessing, learning, data processing, attribute selection and data visualization modules in an environment that encourages initial exploration of data. The experiment mode allows larger-scale experiments to be run with results stored in a database for retrieval and analysis.

2.5 Classification in WEKA

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. Classification is done to know exactly how the data is being classified. The Classify Tab is also supported which shows the list of machine learning tools. These tools in general operate on a classification 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. [5][10]

The first simply splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a classifier works.

2.6 Manifold Machine Learning algorithm

The main motivation for different supervised machine learning algorithms are accuracy improvement. Different algorithms are used different rule for generalizing different representations of the knowledge. Therefore, they tend to error on different parts of the instance space. The combined use of different algorithms could lead to the correction of the individual uncorrelated errors. As a result the error rate and time taken to develop the algorithm is compared with different algorithm [7].

2.7 Experimental Setup

The data mining method used to build the model is classification. The data analysis is processed using WEKA data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 520 instances with 15 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of fluoride affected persons. The performances of the classifiers areevaluated and their results are analyzed. The results of comparison are based on ten-fold cross-validations. According to the attributes, the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing [10].

2.8 Learning Algorithms

This paper consists of three different supervised machine learning algorithms derived from the WEKA data mining tool. Which include:

- J48 (C4.5)
- Naive Bayes,
- CART

The above algorithms were used to predict the accuracy of Fluoride Skeletal diseases affected persons.

3.DISCUSSIONS

3.1 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 record of data base consists of 15 attributes, from which 10 attributes were selected based on attribute selection in explorer mode of WEKA 3.6.4.

Table 2: classification of attributes

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 10 attributes . The other attributes Name, Pregnancy state, Sex, Known status of fluoride, profession omitted for the convenience of analysis of finding impaction among peoples in the district.

Table 3: Selected attributes for analysis

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

3.2 Classifier chosen using Ranker testing in WEKA

```
=== 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
```

Fig.2 Attribute selection in WEKA Explorer

The Classify option in WEKA has many learning tools for finding hidden patterns based on classification. We can choose the best learning tool for the created learning data base from the ranking test in WEKA Experimenter option. Randomly we have chosen six learning algorithms and applied in Experimenter.

The Experimenter has given above the accuracy over the created learning data base. So that we have chosen two high accuracy and one medium accuracy learning algorithms which have highlighted in the above table to find the hidden pattern of the classification.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   1
Resultsets: 6
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       4/27/11 1:58 AM


Dataset            (1) meta.Bag | (2) bayes (3) trees (4) trees (5) trees (6) trees
--------------------------------------------------------------------------------
bbb               (10)  90.90 |  92.95   96.36 v  83.57   90.04   95.96 v
--------------------------------------------------------------------------------
                   (v/ /*) |  (0/1/0)  (1/0/0)  (0/1/0)  (0/1/0)  (1/0/0)


Key:
(1) meta.Bagging '-P 100 -S 1 -I 10 -W trees.REPTree -- -M 2 -V 0.0010 -N 3 -S 1 -L -1' -
5058799622371997O3
(2) bayes.NaiveBayes '' 5995231201785697655
(3) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) trees.RandomTree '-K 0 -M 1.0 -S 1'
8934314652175299374
(5) trees.REPTree '-M 2 -V 0.0010 -N 3 -S 1 -L -1' -9216785998198681299
(6) trees.SimpleCart '-S 1 -M 2.0 -N 5 -C 1.0' 4154189200352566053
```

Fig.3 Ranking test in WEKA Experimenter

Table 4: Experimenter accuracy on Data set

Classifier tool	Experimenter accuracy
Simple Cart	**95.96**
REPTree	90.04
Random Tree	83.57
J48	**96.36**
Bagging	90.90
Naïve Bayes	**92.95**

3.3 J48 algorithm in WEKA

The J48 decision tree in WEKA is based on the C4.5 decision tree algorithm. The C4.5 algorithm is a part of the multi-way split decision tree. C 4.5 yields a binary split if the selected variable is numerical, but if there are other variables representing the attributes it will result in a categorical split. That is, the node will be split into C nodes where C is the number of categories for that attribute . The learning algorithm J48 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 no need of preprocessing for further process [2].

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 which gives the following results.

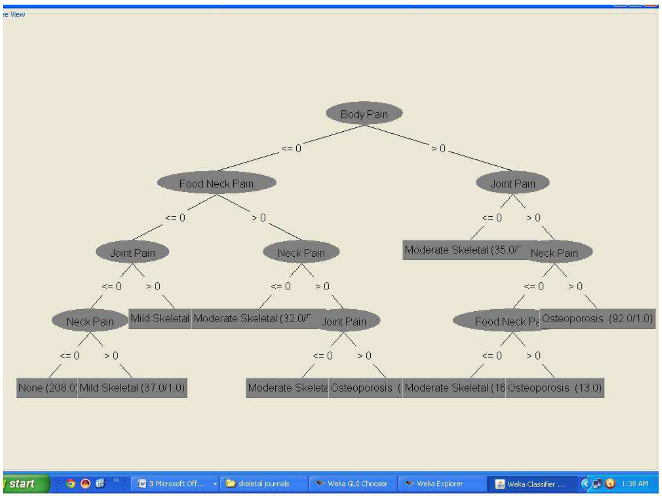

Fig.4 Tree Visualization of J48 in WEKA Explorer

If – then rules of the above implementation

IF (Body pain)
IF(Joint pain)
IF (Neck pain)
Osteoporosis
Else
IF(Foot Neck pain)
Osteoporosis
Else
Moderate Skeletal
Else
Moderate Skeletal
Else
IF(Foot Neck pain)
IF(Neck pain)
IF(Joint pain)
Osteoporosis
Else
Moderate Skeletal
Else
Moderate Skeletal
Else
Mild Skeletal

From the WEKA 3.6.4 classifier Confusion matrix confirms that the Krishnagiri district people are impacted by Moderate Osteoroposis disease.

3.4 Classification And Regression Tree(CART) algorithm in WEKA

It builds a binary decision tree by splitting the records at each node, according to a function of a single attribute. CART uses the Gini index for determining the best split.

```
=== Run information ===

Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    bbb
Instances:   520
Attributes:  9
          Age
          FL
          Drinking water type
          Duration of  drinking water used in years
          Neck Pain
          Joint Pain
          Body Pain
          Foot Neck Pain
          Disease Level
Test mode:   10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Body Pain <= 0
|   Foot Neck Pain <= 0
|   |   Joint Pain <= 0
|   |   |   Neck Pain <= 0: None (208.0)
|   |   |   Neck Pain > 0: Mild Skeletal (37.0/1.0)
|   |   Joint Pain > 0: Mild Skeletal (73.0)
|   Foot Neck Pain > 0
|   |   Neck Pain <= 0: Moderate Skeletal (32.0/5.0)
|   |   Neck Pain > 0
|   |   |   Joint Pain <= 0: Moderate Skeletal (3.0)
|   |   |   Joint Pain > 0: Osteoporosis (11.0/1.0)
Body Pain > 0
|   Joint Pain <= 0: Moderate Skeletal (35.0/6.0)
|   Joint Pain > 0
|   |   Neck Pain <= 0
|   |   |   Foot Neck Pain <= 0: Moderate Skeletal (16.0/2.0)
|   |   |   Foot Neck Pain > 0: Osteoporosis (13.0)
|   |   Neck Pain > 0: Osteoporosis (92.0/1.0)

Number of Leaves  : 10

Size of the tree : 19


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         503               96.7308 %
Incorrectly Classified Instances        17                3.2692 %
Kappa statistic                          0.9544
Mean absolute error                      0.0262
Root mean squared error                  0.1231
Relative absolute error                  7.3338 %
Root relative squared error             29.126  %
Total Number of Instances              520

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1        1          1        None
                 0.916     0.005      0.982      0.916    0.948      0.984    Mild Skeletal
                 0.973     0.029      0.847      0.973    0.906      0.967    Moderate Skeletal
                 0.958     0.005      0.983      0.958    0.97       0.985    Osteoporosis
Weighted Avg.    0.967     0.006      0.97       0.967    0.968      0.988

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 208   0   0   0 |   a = None
   0 109   9   1 |   b = Mild Skeletal
   0   1  72   1 |   c = Moderate Skeletal
   0   1   4 114 |   d = Osteoporosis
```

Fig.5 J48 Implementation in WEKA Explorer

The initial split produces two nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine all the input fields to find candidate splitters. If no split can be found that significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of over fitting [1][3]

At the end of the tree growing process, every record of the training set has been assigned to some leaf of the full decision tree. Each leaf can now be assigned a class and an error rate. The error rate of a leaf node is the percentage of incorrect classification at that node. The error rate of an entire decision tree is a weighted sum of the error rates of all the leaves. Each leaf's contribution to the total is the error rate at that leaf multiplied by the probability that a record will end up in there.

We have trained the training data by using the 10 Fold Cross Validated testing. The CART decision tree classifier Confusion matrix too confirms the same result obtained in the J48 decision tree. That is the Krishnagiri District are impacted by Skeletal Osteoroposis.

3.5 Naive Bayes algorithm in WEKA

Bayesian classification is quite different from the decision tree approach. In Bayesian classification we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data.

The expression P(A) refers to the probability that event A will occur. P(A/B) stands for the probability that event A will happen given that event B has already happened. In other words p(A/B) is the conditional probability of A based on the condition that B has already happened. For example, A and B may be probability of passing a course A and passing another course B respectively. P(A/B) then is the probability of passing A when we know that B has been passed [1, 3].

If we consider X to be an object to be classified then Bayes theorem may be read as giving the probability of it belonging to one of the classes C_1,C_2,C_3, etc by calculating $P(C_i/X)$. Once these probabilities have been computed for all the classes, we simply assign X to the class that the highest conditional probability.

```
=== Run information ===

Scheme:     weka.classifiers.trees.SimpleCart -S 1 -M 2.0 -N 5 -C 1.0
Relation:   bbb
Instances:  520
Attributes: 9
            Age
            FL
            Drinking water type
            Duration of  drinking water used in years
            Neck Pain
            Joint Pain
            Body Pain
            Foot Neck Pain
            Disease Level
Test mode:   evaluate on training data

=== Classifier model (full training set) ===

CART Decision Tree

Joint Pain < 0.5
|  Neck Pain < 0.5
|  |  Foot Neck Pain < 0.5
|  |  |  Body Pain < 0.5: None(208.0/0.0)
|  |  |  Body Pain >= 0.5: Moderate Skeletal(14.0/3.0)
|  |  Foot Neck Pain >= 0.5
|  |  |  Age < 24.5: Moderate Skeletal(12.0/0.0)
|  |  |  Age >= 24.5
|  |  |  |  Body Pain < 0.5
|  |  |  |  |  Duration of  drinking water used in years=(8.0)|(10.0)|(3.0)|(5.0)|(15.0): Mild Skeletal(4.0/2.0)
|  |  |  |  |  Duration of  drinking water used in years!=(8.0)|(10.0)|(3.0)|(5.0)|(15.0): Moderate Skeletal(6.0/1.
|  |  |  |  Body Pain >= 0.5: Moderate Skeletal(8.0/0.0)
|  Neck Pain >= 0.5
|  |  Body Pain < 0.5
|  |  |  Foot Neck Pain < 0.5: Mild Skeletal(36.0/1.0)
|  |  |  Foot Neck Pain >= 0.5: Moderate Skeletal(3.0/0.0)
|  |  Body Pain >= 0.5
|  |  |  FL < 1.7000000000000002: Moderate Skeletal(5.0/0.0)
|  |  |  FL >= 1.7000000000000002: Osteoporosis (3.0/0.0)
Joint Pain >= 0.5
|  Body Pain < 0.5
|  |  Foot Neck Pain < 0.5: Mild Skeletal(73.0/0.0)
|  |  Foot Neck Pain >= 0.5
|  |  |  Neck Pain < 0.5: Moderate Skeletal(9.0/0.0)
|  |  |  Neck Pain >= 0.5: Osteoporosis (10.0/1.0)
|  Body Pain >= 0.5
|  |  Neck Pain < 0.5
|  |  |  Foot Neck Pain < 0.5
|  |  |  |  Age < 54.0: Moderate Skeletal(9.0/0.0)
|  |  |  |  Age >= 54.0
|  |  |  |  |  FL < 1.7000000000000002: Moderate Skeletal(4.0/0.0)
|  |  |  |  |  FL >= 1.7000000000000002: Mild Skeletal(2.0/1.0)
|  |  |  Foot Neck Pain >= 0.5: Osteoporosis (13.0/0.0)
|  |  Neck Pain >= 0.5: Osteoporosis (91.0/1.0)

Number of Leaf Nodes: 18

Size of the Tree: 35

Time taken to build model: 0.28 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        510          98.0769 %
Incorrectly Classified Instances       10           1.9231 %
Kappa statistic                  0.9731
Mean absolute error                  0.016
Root mean squared error                0.0894
Relative absolute error              4.472  %
Root relative squared error           21.1509 %
Total Number of Instances            520

=== Detailed Accuracy By Class ===

         TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          1       0        1          1       1          1         None
          0.966   0.01     0.966      0.966   0.966      0.998     Mild Skeletal
          0.946   0.009    0.946      0.946   0.946      0.997     Moderate Skeletal
          0.983   0.005    0.983      0.983   0.983      0.998     Osteoporosis
Weighted Avg.  0.981   0.005   0.981   0.981   0.981   0.998

=== Confusion Matrix ===

  a   b   c   d   <-- classified as
208   0   0   0 |   a = None
  0 115   3   1 |   b = Mild Skeletal
  0   3  70   1 |   c = Moderate Skeletal
  0   1   1 117 |   d = Osteoporosis
```

Fig.6 CART implementation in WEKA Explorer

```
=== Run information ===

Scheme:     weka.classifiers.bayes.NaiveBayes
Relation:   bbb
Instances:  520
Attributes: 9
            Age
FL
            Drinking water type
            Duration of drinking water used in years
            Neck Pain
            Joint Pain
            Body Pain
            Foot Neck Pain
            Disease Level
Test mode:  evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

                        Class
Attribute                None   Mild Skeletal Moderate Skeletal   Osteoporosis
                        (0.4)     (0.23)      (0.14)       (0.23)
================================================================================
Age
 mean                   24.9604   36.5014    36.8674    44.8512
 std. dev.              13.2702   15.7583    17.0054    16.5922
 weight sum                 208       119         74        119
 precision               1.3231    1.3231     1.3231     1.3231


=== Detailed Accuracy By Class ===

     TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Ar
      0.976   0        1          0.976   0.988      1       Noı
      0.908   0.017    0.939      0.908   0.923      0.965   Mild Sk
      0.892   0.034    0.815      0.892   0.852      0.972   Moderate Skeleta
      0.933   0.025    0.917      0.933   0.925      0.99    Osteopc
Weighted Avg.  0.938   0.014   0.941   0.938   0.93

=== Confusion Matrix ===

  a   b   c   d   <-- classified as
203   5   0   0 |  a = None
  0 108   8   3 |  b = Mild seletal
  0   1  66   7 |  c = Moderate Skeletal
  0   1   7 111 |  d = Osteoporosis
```

Fig.7 Naïve Bayes implementation in WEKA Explorer

$P(C_i/X)$ may be calculated as

$P(C_i/X) = [P(X/C_i)P(C_i)]/P(X)$

- $P(C_i/X)$ is the probability of lthe object X belonging to class C_r
- $P(X/C_i)$ is the probability of obtaining attribute values X if we know that it belongs to class C_r
- $P(C_r)$ is the probability of any object belonging to class C_i without any other information.
- $P(X)$ is the probability of obtaining attribute values X whatever class the object belongs to.

The Naive Bayes classifier Confusion matrix also declares the same result obtained in the J48 and CART decision trees. That is the Krishnagiri District are impacted by Skeletal Osteoroposis.

4. RESULT COMPARISION

The above implementation algorithm yields the same results that the Krishnagiri district residing people affected by the Osteoroposisdisease. However some key parameters which played important role in which algorithm works better.

Table 5: Comparison of classified Trees

Classification Algorithm Tree-Type	% of Coreectly Classified instances	Root mean Square error	Time take to build the model(In seconds)
J48(C4.5)	96.7308%	0.1231	0.00
Simple CART	98.0769%	0.0894	0.28
Naïve Bayes	93.8462%	0.1583	0.00

All the three classified learning algorithms train the data up to 98% so the error rate completely reduced. The time taken to build the algorithm relatively too small. However the J48(C4.5) build the model faster than other two algorithms.

Table 6:Comparision of Accuracy

CLASSIFIER	CLASS	TP RATE	FP RATE	PRECISION	RECALL	F MEASURE	ROC AREA
J48	None	1	0	1	1	1	1
	Mild Skeletal	0.916	0.005	0.982	0.916	0.918	0.984
	Moderate skeletal	0.973	0.029	0.847	0.973	0.906	0.967
	Osteoporosis	0.958	0.005	0.983	0.958	0.970	0.985
Simple CART	None	1	0	1	1	1	1
	Mild Skeletal	0.966	0.01	0.966	0.966	0.966	0.998
	Moderate skeletal	0.946	0.009	0.946	0.946	0.946	0.997
	Osteoporosis	0.983	0.005	0.983	0.983	0.983	0.998
Naïve Bayes	None	0.976	0	1	0.976	0.988	1
	Mild Skeletal	0.908	0.017	0.939	0.908	0.923	0.965
	Moderate skeletal	0.892	0.034	0.815	0.892	0.852	0.972
	Osteoporosis	0.933	0.025	0.917	0.933	0.925	0.990

The accuracy can be measured from true positive and false positive ratio. All algorithms vary in the range of 0.050 ratio by true positive and vary in the range of 0.005 ratio false positive in Dental Moderate class. So the accuracy among the algorithms also supports the results. From the accuracy comparison it is understood that the Krishnagiri district impacted by Osteoroporis.

5. CONCLUSION

Datamining applied in health care domain, by which the people get beneficial for their lives. As the analog of this research found the meaningful hidden pattern that from the real data set collected the people impacted in Krishnagiri district 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 diseases. However the Primary Health hazards of fluoride are Osteoroposis and Bone diseases which disturbed their daily meager life. It is primary duty of the Government to providing good hygienic drinking water to the people and reduces the fluoride content potable water with the latest technologies and creating awareness among the people in some way like medical camps and taking documentary films. Through this research the problem of fluoride in Krishnagiri come to light.

References

[1] Jiawei Han and MichelineKamber, "Data mining concepts and Techniques",Second Edition, Morgan Kaufmann Publishers second edition,2008.

[2] ArunK.Pujari, "Datamining Techniques", University Press, First edition, fourteenth reprint, 2009.

[3] G.K.Gupta, "Introduction to Datamining with case studies", PHI. 2009

[4] BerrryMjLinoff G, "Data mining Techniques: for Marketing, Sales and Customer support USA", Wiley, 1997.

[5] Weka3.6.4 data miner manual. 2010.

[6] Water Quality for Better Health – TWAD Released Water book. Published IEC, TWAD,Chennai.mail:twadboard@vsnl.in,2009.

[7] PlamenaAndreeva, Maya Dimibova and Petra Radeve, "Data mining Learning models and Algorithms for medical applications – White paper",page no.44, 2004.

[8] Professionals statement calling for an End to water Fluoridation – Conference Report NRC Review,2006.(www.fluoridealert.org)

[9] "Analysis of Liver Disorder Using Data mining algorithms", Global Journal of computer science and Technology, l.10 issue 14 (ver1.0) November 2010, pp. 48 - 52.

[10] Peter Reutemann, Ian H. Witten,"The WEKA Data Mining Software: An Update- White paper", Pentaho Corporation. SIGKDD Explorations Volume 11, Issue 1,pp. 10 - 18, 2005

T. Balasubramanian (Corresponding author) received his M. Sc Computer Science Degree from Jamal Mohamed College, Trichy affiliated with Bharathidasan University and M. Phil Degree from Periyar University. Now pursuing his Part time Ph. D research in Bharathiar University, Coimbatore. Now he is working As Asst. Professor, Department of Computer Science in Sri VidyaMandir Arts and Science College, Uthangarai, Krishnagiri Dt. His research area is of Data Mining application Techniques. He has published 9 research papers in various National, International conferences and 6 paper in various international journals.

Dr. R. Umarani has completed her M.C.A. from NIT, Trichy in 1989. She did her M.Phil. from Mother Teresa University, Kodaikanal. She received her Ph.D., from PeriyarUniverisity, Salem in 2006. Her area of interest includes Information Security, Data mining and Mobile communications. She has published about 50 papers in National and International conferences. She is also working as Associate Professor in Department of Computer Science, Sri Sarada College for women, Salem. She has published 35 papers in International and National journals.