CREATION OF AN ADAPTIVE CLASSIFIER TO ENHANCE THE CLASSIFICATION ACCURACY OF EXISTING CLASSIFICATION ALGORITHMS IN THE FIELD OF MEDICAL DATA MINING

A Dissertation Submitted

By

Sneha Chandra (11309529)

To

Department of

Computer Science & Engineering

In partial fulfilment of the Requirement

for the Award of the Degree of

Master of Technology in Computer Science & Engineering

Under the Guidance of

Ms. Maneet Kaur (15709)

(May 2015)

INDIA'S LARGEST UNIVERSITY

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education, Transforming India

School of: COMPUTER SCIENCE & ENGINEERING

DISSERTATION TOPIC APPROVAL PERFORMA

Name of the Student: SNEHA CHANDRA

Registration Number: 11309529

Batch: 2013-15

Roll Number: RK2307B50

Session: AUTUMN 2014-15

Parent Section: K2307

Details of Supervisor:

Designation: ASSISTANT PROFESSOR

Name: MANEET KAUR

Qualification: M.TECH

UID: 15709

Research Experience: 3 YEARS

SPECIALIZATION AREA: DATABASE & INFORMATION SYATEMS **(pick from list of provided specialization ar as by DAA)**

PROPOSED TOPICS

1. Creation of Ensemble of Classifiers to enhance the efficiency of existing Classification Algorithms in the field of Medical Data Mining
2. Classification Data Mining
3. Clustering Data Mining

Signature of Supervisor

PAC Remarks:

first topic is approved, publication expected.

APPROVAL OF PAC CHAIRPERSON: Signature: 30/9/14 Date:

*Supervisor should finally encircle one topic out of three proposed topics and put up for approval before Project Approval Committee (PAC)

*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report

*One copy to be submitted to Supervisor

ABSTRACT

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. One of the most important techniques of Data Mining is classification which is used for prediction purposes. In this paper, a novel classifier is presented for classification in the field of Medical Data Mining. The idea is to apply the Adaptive Classifier on the sample medical dataset, and compare its results with the results obtained through the individual classification techniques. This approach has been implemented and tested to show higher classification accuracy for the Adaptive Classifier. The results obtained have lit a great spark for future investigation.

Keywords – Adaptive Classifier, Bayesian Classification, Decision-Tree Classification, Diameter Measurement, Ensemble Classification, Euclidean Distance, Image Segmentation, K-means Clustering, Laplacian Correction, Mean Gray Level Algorithm, Medical Data Mining, Model Evaluation, Rule-Based Classification, Texture Analysis

CERTIFICATE

This is to certify that Sneha Chandra (11309529) has completed M.Tech dissertation titled Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation has ever been submitted for any other degree or diploma.

The dissertation is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech in Computer Science & Engineering.

Date: 28-04-2015

Signature of Advisor

Name: Ms. Maneet Kaur

ACKNOWLEDGMENT

It takes a lot of endeavour to bring a research work to success. Each and every step involved in a research work is the result of sincere and honest efforts on the part of the researcher. However, it is not only the dedication of the researcher that helps him to achieve success but the valuable contributions of those who support him in his work, as well.

Hence, it is time to thank all those who have made significant contributions towards my research work, and deserve special recognition.

First of all, I would like to thank the Almighty GOD who has guided me through all phases of my life. Next, I would like to extend my gratitude to my Dissertation Mentor, Ms. Maneet Kaur, who has imparted valuable assistance to my research work. I would also like to thank all the members of Lovely Professional University who have assisted me in my research work, either directly or indirectly.

Last but not the least, I would like to thank my family members and my friends who have showed unflinching loyalty and support for the successful completion of this research work.

Sneha Chandra

DECLARATION

I, hereby, declare that the dissertation entitled, Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining, submitted for the M.Tech Degree, is entirely my original work, and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: 28-04-2015

Investigator: Sneha Chandra
Registration Number: 11309529

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Over the past several years, there has been a great inflow of medical data at the health information systems. As health issues have been a major area of concern, challenges exist in detecting and finding possible cures to the deadly diseases. Many of these diseases remain unnoticed for a longer duration, which leads to their unchecked progression. However, a person has higher chances of survival if he or she is detected to be suffering from a disease at an early stage, because early diagnosis facilitates early management.

Time and then, experts have worked towards developing an intelligent agent that could help in early detection of the disease that a person is suffering from. Various algorithms have been applied to solve the problem. However, as concluded in [2], no single classification algorithm is best for all kinds of dataset. Hence, there arose a need to develop an intelligent agent that could combine many classification algorithms and produce more certain, precise and accurate results.

Moreover, since the past few years, image processing [1,8] has achieved significant importance in the field of Medical Data Mining [4,5]. As health issues have remained a major area of concern, early detection of a disease requires careful investigation. In most of the cases, an investigation requires the use of certain medical tests to reach to a common conclusion. With the advancement of technology, a large number of image processing techniques have been developed to aid in the diagnosis of the disease.

Hence, this research work proposes a novel Data Mining approach by applying an Adaptive Classifier for classification in the field of Medical Data Mining. The objective is to generate more certain, precise and accurate results, as stated by Kesavaraj and Sukumaran [2], which would help in better detection of diseases.

In the advanced version of our Adaptive Classifier, an innovative image processing technique is used which works upon the externally, as well as, internally observable features of the diseases to classify the diseases with greater classification accuracy [2]. This technique is designed to overcome the limitations of our initial Adaptive Classifier [10], which was designed to work only upon the externally observable features of diseases. Further, the advanced version of our Adaptive Classifier has been generated using the techniques of Clustering Data Mining in conjunction with Classification Data Mining, to handle one class sample medical datasets instead of two class sample medical datasets handled by our initial Adaptive Classifier [10].

The objective is to achieve higher classification accuracy in classifying all types of diseases, obtained from different classes of sample medical datasets, which may or may not always be characterized by the observable features of the diseases [2,10].

CHAPTER 2

REVIEW OF LITERATURE

2.1 Review of Concepts & Theories

Medical Data Mining has been described as the most challenging field of Data Mining since the past few years [4,5,9]. The challenge lies in the detection and the treatment of diseases, some of which may progress to cause profound morbidity (ill health), and even death. Further, analysis of medical images is the most critical aspect of Medical Data Mining [1,8]. Many different forms of image analysis techniques have been developed to gain an insight into the human body. These techniques aid in the diagnosis of diseases to a large extent, which may otherwise, prove detrimental for the human health. Some of the major diseases affecting the present world population are: Diabetes, Liver Cirrhosis and Tuberculosis.

2.1.1 Diabetes Mellitus

Figure 2.1: Foot ulcer in a patient with Diabetes Mellitus

Diabetes Mellitus is a non-infectious metabolic disease which has characteristic high blood sugar levels over a long period of time. The characteristic symptoms of Diabetes Mellitus include frequent urination (*polyuria*), increased thirst (*polydipsia*), and increased hunger (*polyphagia*). If Diabetes Mellitus is left untreated, it can cause many serious long-term complications, such as, heart disease, stroke, kidney failure, foot ulcers and damage to the eyes [4,5].

2.1.2 Liver Cirrhosis

Figure 2.2: Swollen feet of a patient with Liver Cirrhosis

Liver Cirrhosis is defined as a chronic form of liver disease [9]. It is characterized by the replacement of liver tissue by fibrosis (scar tissue) and regenerative nodules (lumps that occur due to attempted repair of damaged tissue), leading to an alteration in the architecture of liver and loss of liver functions. Liver Cirrhosis is most commonly caused by alcoholism, Hepatitis B, Hepatitis C, and non-alcoholic fatty liver disease (NAFLD). However, it has many other possible causes some of which are unknown.

2.1.3 Tuberculosis

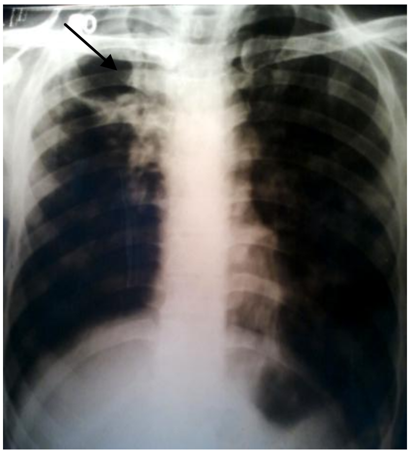

Figure 2.3: A cavity in the right lung as seen in the chest X-ray of a patient with advanced Pulmonary Tuberculosis infection

Figure 2.4: A lymph node swelling as seen in a patient with Lymph Node Tuberculosis

Tuberculosis is a widespread infectious disease caused by various strains of mycobacterium, usually *Mycobacterium tuberculosis*. The infection of tuberculosis occurs in almost one-third of the world's population out of which 10% go on to develop the disease. Tuberculosis infection is known to attack the lungs, causing Pulmonary Tuberculosis. The typical symptoms of an active Pulmonary Tuberculosis infection are chronic cough, cough with blood-tinged sputum, fever, night sweats and weight loss. It is fatal in many cases. However, it can affect other parts of the body, as well. The infection of other organs causes Extra Pulmonary Tuberculosis, which is characterized by a wider range of symptoms, depending upon the organ infected by tuberculosis. In this paper, research work has been carried out on Pulmonary Tuberculosis and Lymph Node Tuberculosis, a form of Extra Pulmonary Tuberculosis. Lymph Node Tuberculosis is characterized by the swelling of lymph nodes, which vary in size and number.

Hence, to contain the above major health problems, experts have, time and then, tried to build intelligent agents that could help in early diagnosis, thus facilitating early treatment of these deadly diseases. They have tried their utmost to prevent morbidity and save human lives. After several years of tough endeavour, they were able to obtain a major breakthrough in the field of medical diagnosis, by the development of classifiers for prediction purposes. A classifier is either built using a single classification algorithm or a combination of classification algorithms (ensemble classifier) to obtain higher classification accuracy.

2.1.4 Clustering

Clustering is an important technique of Data Mining that groups a set of data objects into multiple groups or clusters so that objects within a cluster are highly similar to each other but are highly dissimilar to objects in other clusters [1,2].

a) K-means Clustering

K-means Clustering is a centroid-based partitioning technique used in cluster analysis [3].

Let D, the dataset, contain n objects in Euclidean space. The objects in D are distributed into k clusters, $C_1,\ldots\ldots\ldots\ldots, C_k$, where $C_i \subset D$ and $C_i \cap C_j = \emptyset$ for $1 \leq i, j \leq k$. The centroid-based partitioning technique uses the *centroid* of a cluster, C_i, to represent that cluster. The centroid can be defined in various ways such as by the mean or the medoid of the objects assigned to the cluster. The difference between an object $\boldsymbol{p} \in C_i$ and $\boldsymbol{c_i}$, the representative of the cluster, is measured by $dist(\boldsymbol{p},\boldsymbol{c_i})$, where $dist(\boldsymbol{x},\boldsymbol{y})$ is the Euclidean distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$. The quality of cluster C_i can be measured by the within-cluster variation, which is the sum of *squared error* between all objects in C_i and the centroid $\boldsymbol{c_i}$, defined by:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2 \qquad Eq.(2.1)$$

where E is the sum of the squared error for all objects in the dataset; $\boldsymbol{p}$ is the point in space representing a given object; and $\boldsymbol{c_i}$ is the centroid of cluster C_i (both $\boldsymbol{p}$ and $\boldsymbol{c_i}$ are multidimensional). 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 objective function tries to make the resulting k clusters as compact and as separate as possible.

2.1.5 Classification

Classification is an important technique of Data Mining that uses classifiers to predict categorical (discrete, unordered) class labels [2,11]. It is one of the main issues in the field of Data Mining Research [7]. For example, a classification model can be constructed to predict whether a patient has a particular disease or not. Here, the task of data analysis is classification in which a model or classifier is constructed to predict categorical class label for the patient.

General Approach to Classification:

The task of data classification is defined as a two step process, as shown in Figure 2.1, consisting of a learning step (where a classification model is constructed), and Figure 2.2, a classification step (where the model is used to predict class labels for the given data).

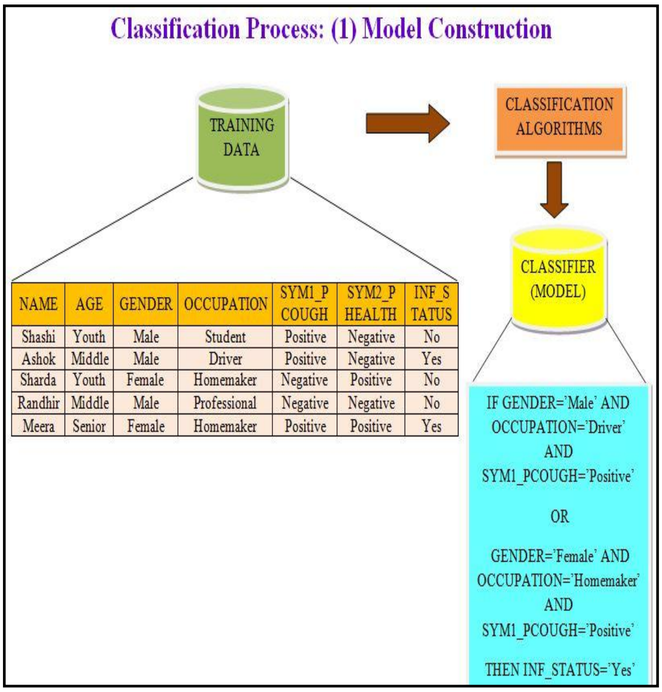

Figure 2.5: Classification Process: (1) Model Construction

Step 1: In this step, a classifier is built from a classification algorithm. The classification algorithm first performs the analysis, and then, trains itself from a *training set* consisting of predetermined class labels.

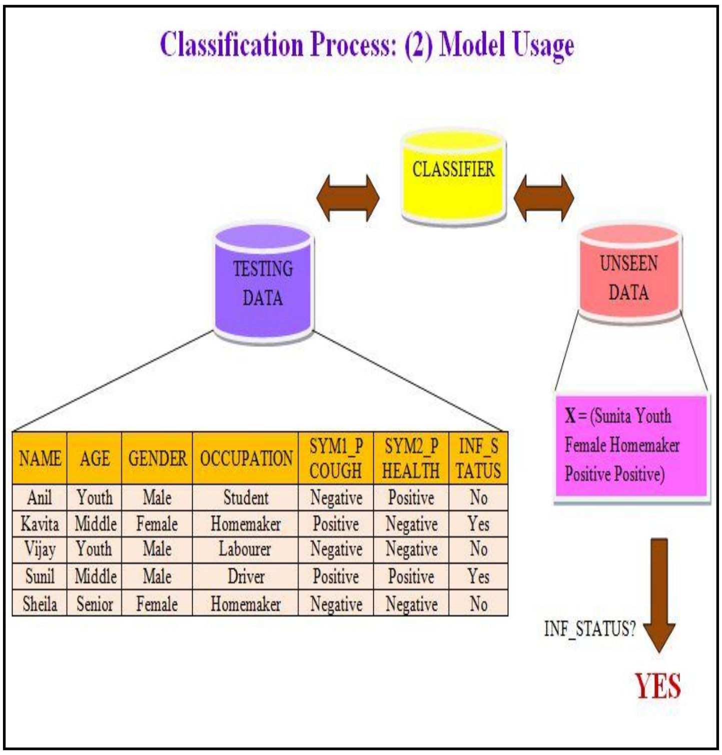

Figure 2.6: Classification Process: (2) Model Usage

Step 2: In this step, the newly built classifier is used for classification. First of all, we estimate the predictive accuracy of the classifier, by using it on a *test set*. Next, we evaluate the classification accuracy of the classifier, by comparing the actual class label of each test tuple with the predicted class label for that tuple. If the classification accuracy falls within the standard limits, the classifier is considered fit for classification purposes.

a) Rule-Based Classification

Rule-Based Classification uses a set of IF-THEN rules for classification [3]. An IF-THEN rule is an expression of the form:

IF *condition* THEN *conclusion*

In this research work, I have generated rules based on the features of patients to predict whether a patient has Diabetes Mellitus or Liver Cirrhosis or either forms of Tuberculosis or not.

Examples of Classification Rules:

R1: IF *polyuria* = *yes* AND *polydiphagia* = *yes* OR *polyuria* = *yes* AND *polydiphagia* = *yes* AND *poor_health* = *yes* THEN *diabetes_disease* = *yes* (where, *polydiphagia* = *polydipsia* + *polyphagia*)

R2: IF *liver_size* = small AND *liver_texture* = altered OR *liver_size* = medium AND *liver_texture* = altered OR *liver_size* = large AND *liver_texture* = altered OR *liver_texture* = altered AND *poor_health* = positive THEN *liver_cirrhosis_disease* = yes

R3: IF *persistent_cough* = *yes* AND *poor_health* = *yes* THEN t*uberculosis_disease* = *yes* (where, *persistent_cough* = *chronic_cough* + *blood_tinged_sputum_in_cough*)

R4: IF *lymph_node_size* = medium AND *poor_health* = positive OR *lymph_node_size* = large AND *poor_health* = positive OR *mantoux_test* = positive THEN *ln_tuberculosis_disease* = yes

In all the above cases, *poor_health* refers to the non-distinctive or common features of the diseases, such as fever, sweating, tiredness, loss of appetite, loss of weight, and so on.

b) Decision-Tree Classification

Decision-Tree Classification is based on Decision Tree Induction which is defined as the learning of decision trees from class-labeled training tuples. A decision tree is represented as a graphical tree structure in which each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node is considered as the root node [3].

In this research work, I have chosen *Information Gain* as the attribute selection measure. The notation used herein is as follows:

Let D, the data partition, be a *training set* of class-labeled tuples. Let the class label attribute have m distinct values defining m distinct classes, C_i (for i = 1,2,……,m). Let $C_{i,D}$ be the set of tuples of class C_i in D. Let $|D|$ and $|C_{i,D}|$ denote the number of tuples in D and $C_{i,D}$, respectively. Let node N represent or hold the tuples of partition D.

The expected information needed to classify a tuple in D is given by:

$$Info(D) = -\sum_{i=1}^{m} p_i \, log_2(p_i) \qquad Eq.(2.2)$$

where p_i is the nonzero probability that an arbitrary tuple in D belongs to class C_i and is estimated by $\frac{|C_{i,D}|}{|D|}$. *Info*(D), also known as the entropy of D, is the average amount of information needed to identify the class label of a tuple in D.

Since we are concerned with discrete-valued attributes, the amount of information still needed (after the partitioning) to arrive at an exact classification, for each attribute A, is measured by:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad Eq.(2.3)$$

where the term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. $Info_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A.

The difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A) is measured by:

$$Gain(A) = Info(D) - Info_A(D) \qquad Eq. (2.4)$$

Gain(*A*) is the expected reduction in the information requirement caused by knowing the value of *A*. The attribute *A* with the highest information gain, *Gain*(*A*), is chosen as the splitting attribute at node *N*.

c) Bayesian Classification

Bayesian Classification is based on Bayes' Theorem which is used to predict class membership probabilities [3]. The Bayes' Theorem is given by:

$$P(H/\boldsymbol{X}) = \frac{P(\boldsymbol{X}/H)\, P(H)}{P(\boldsymbol{X})} \qquad Eq. (2.5)$$

In this research work, I have used Naïve Bayesian Classification or Simple Bayesian Classification which assumes that the effect of an attribute value on a given class is independent of the values of the other attributes (*class conditional independence).*

d) Laplacian Correction

Laplacian Correction or Laplace Estimator is the method of probability estimation used to avoid the cases of probability values of zero [3].

e) Ensemble Classification

An ensemble for classification is a composite model, made up of a combination of classifiers. An ensemble returns a class label prediction based on the collection of votes of its individual classifiers. Hence, an ensemble is considered to be more accurate than its individual classifiers [3].

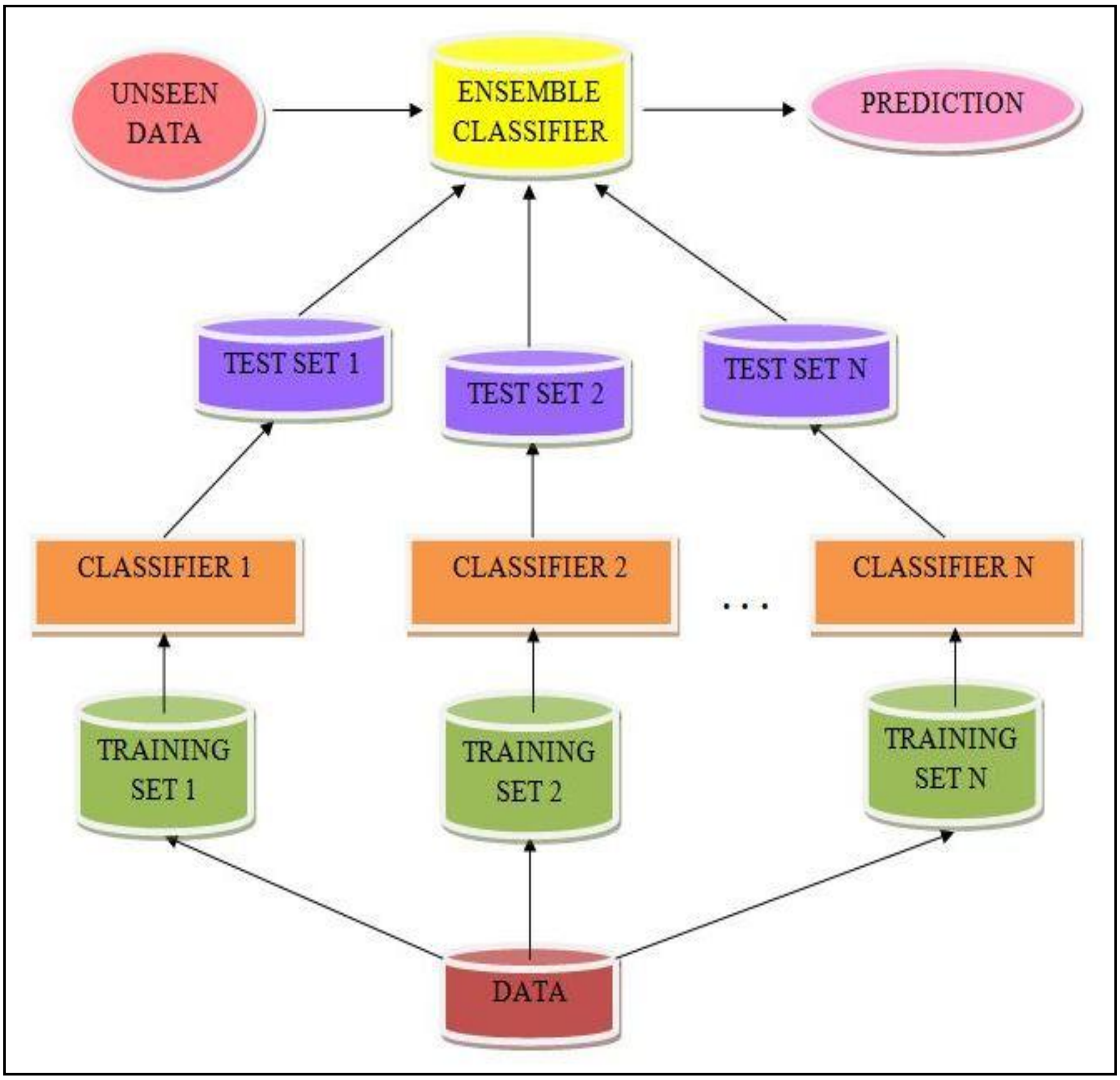

Figure 2.7: An Ensemble Classifier

The given dataset under consideration undergoes a few preprocessing techniques so as to obtain better results, and to improve the performance of the proposed model [6]. The given dataset is used to generate n *training sets* for n classifiers. Each *training set* contains the same training data which are fed into the individual classifiers. The most commonly used individual classifiers include Rule-Based Classifier, Decision-Tree Classifier, K-means Classifier, Naïve Bayesian Classifier, and Neural Networks Classifier. After the training phase is over, n *test sets* are generated for n classifiers. Here as well, each *test set* contains the same test data which are fed into the individual classifiers. The class label prediction for each *test set* tuple, by each individual classifier, is taken into consideration. An ensemble classifier is generated by combining the votes of each individual classifier, for each *test tuple* under consideration. After the combination of votes of individual classifiers, an ensemble method is used to obtain the final class label prediction returned by the ensemble classifier. Some of the ensemble methods include *Bagging*, *Boosting*, and *Random Forests*. If the classification accuracy of the ensemble classifier is found to be within the standard limits, it is used for classification purposes on new or unseen data. Generally, an ensemble has higher classification accuracy than its individual classifiers. Hence, it is widely used for classification purposes.

2.1.6 Euclidean Distance

Euclidean Distance is defined as the displacement between two points in Euclidean Space.

2.1.7 Mean Gray Level Algorithm

Mean Gray Level Algorithm is defined as the algorithm that sums up all the pixel values in the image and then takes the mean of the total to obtain the Threshold Value.

2.2 Review of Previous Research Findings

Mohammad Saraee, George Koundourakis, Babis Theodoulidis, “EasyMiner: Data Mining in Medical Databases”, The Institution of Electrical Engineers, Savoy Place, London WC2R OBL, UK, Pg 7/1 – 7/3, IEEE, 1998.

In this paper, Data Mining technology has been used to discover relationships in large clinical units. Emphasis has been laid on the evaluation of accumulated clinical data which may lead to the discovery of trends and patterns contained within the data. These trends and patterns could significantly broaden our horizon about disease progression and management. EasyMiner, the data mining system, was designed and developed for interactive mining of interesting patterns in time-oriented medical databases. This system implements a plethora of data mining functions, including generalization, relevance analysis, classification and discovery of association rules.

At the very first step, EasyMiner uses the generalization rules, to generalize the values of the attributes at multiple concept levels. Next, relevance analysis is performed which helps to determine the relevance of an attribute to the required data mining task. Now, we have the most important step of classification, which groups large quantities of data on the basis of common characteristics and properties. EasyMiner uses the classification technique based on the decision tree structure. The idea was implemented on the clinical database of stroke patients from East Lancashire. Finally, association rules were generated that help in analyzing the patterns.

EasyMiner provides us with an efficient technique of Data Mining, to investigate large quantities of clinical data for hidden patterns and relationships. It has been successful in implementing the quality and cost-effectiveness of patients care. However, other Data Mining approaches must be tested to give more accurate results, either individually, or in combination with other approaches.

Su-Lan Zhang and Ji-Fu Zhang, “A New Classification Mining Model Based on the Data Warehouse”, Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi’an, 2-5 November, 2003, ISBN: 0-7803-7865-2/03/$17.00, Pg 168 – 171, IEEE, 2003

The motivation for this research work arose from the field of decision-making management. In this paper, a new classification mining model is developed which helps to manage the data present in the Data Warehouse. The theories of Concept Lattice and Rough Set have been merged to enhance the efficiency and reliability of the knowledge obtained from the Data Warehouse.

The uncertain data obtained from the data warehouse has been handled using attribute reduction measures on Rough Set Theory. Rough Set requires the use of specific upper and lower approximate set for data classification. However, certain data has been handled by constructing the Concept Lattice directly, and then, using a reduction algorithm for reduction of the lattice. Concept lattice uses the relation of concept intension and concept extension for data classification. Finally, the classification knowledge is extracted from the Data Warehouse.

The researchers have laid an emphasis on the fact that a single method cannot be perfect for all kinds of test sets. Hence, it is better to build a model which combines different concepts and theories. Hence, the new classification mining model has been designed with the aim of guiding industry and commerce management departments to make better and more effective decisions.

Rafal Rak, Lukasz Kurgan, Marek Reformat, "Multi-label Associative Classification of Medical Documents from MEDLINE", Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA 05), ISBN: 0-7695-2495-8/05/$20.00, IEEE, 2005

This research work aims to eliminate the extensive manual work at health information systems by the development of an automated system for the classification of articles present at these information systems. Articles from the largest medical repository, MEDLINE, have been used in this research work. OHSUMED, a subset of the MEDLINE database, was used as the source of documents and the MESH tree was taken as the class labels.

The proposed approach involves many classification algorithms, including recurrent and non-recurrent associative classification, in which rules obtained from association rule mining processes are used by classification algorithms to classify the objects. The most important area of this research work is multi-label classification in which each medical document can be assigned to several classes. A modified version of ACRI tool was used to enable multi-label classification. The algorithms were able to obtain a high level of classification accuracy.

The result analysis showed that recurrent item based associative classification achieved a higher degree of performance than its counterparts. Further, different measures of classification quality were investigated. The researchers have presented three alternative setups that allow the user to obtain different desired classification qualities.

Winnie W. M. Lam and Keith C. C. Chan, "A Structural Data Mining Approach for the Classification of Secondary RNA Structure", Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, Shanghai, China, September 1-4, 2005, ISBN: 0-7803-8740-6/05/$20.00, Pg 4759 – 4762, IEEE, 2005

In this paper, research work has been carried out in the field of Medical Data Mining. The researchers aim to classify the genomic data present in different living organisms by exploiting the structures of the genes rather than their primary sequences.

The proposed approach works upon transfer RNA (tRNA) which is the key molecule in protein biosynthesis in all living organisms. The algorithm designed for this purpose is the Random Multi-Level Attributed (RMLA) graph algorithm, which is able to perform data mining on structural data, and represent the secondary genomic structures from such biomolecules.

The proposed approach was able to classify tRNA secondary structures into different classes, effectively. Moreover, it was able to identify the dominant patterns shared by tRNA secondary structures of the same class. The results of RMLA graph algorithm were compared against SUBDUE, another structural classification algorithm. RMLA graph algorithm was found to achieve higher classification accuracy than SUBDUE, amongst all the classes of living organisms. Apart from the classification of genomic data, RMLA graphs were found to be useful for discovering patterns in the databases.

Christopher F. Barnes, Senior Member, IEEE, "Image-Driven Data Mining for Image Content Segmentation, Classification, and Attribution", IEEE Transactions on Geoscience and Remote Sensing, Vol. 45, No. 9, September 2007, 0196-2892/$25.00, Pg 2964 – 2978, IEEE, 2007

In this research work, Image-Driven Data Mining (IDDM) methods are described which deal in image content segmentation, classification, and attribution. Sample results are depicted through synthetic aperture radar images and with high resolution pan-sharpened satellite images of the Payagala, Sri Lanka area before its devastation by the 2004 Asian Tsunami.

Each pixel location of an image under consideration is taken as the center point of a pixel-block query which returns an estimated class label. The researchers have proposed novel methods for pixel-block mining, pattern similarity scoring, class label assignments, and attribute mining. These novel methods have been developed on a direct sum tree structure known as the sigma-tree which is absorbed in near-neighbour similarity scoring. Nowadays, sigma-trees are built inside the Data Warehouse Systems, which provide referential capabilities for feature attribute data. This has laid a foundation for Data Mining called Source Optimized, Labeled, DIgital Expanded Representations (SOLDIER).

The sigma-tree based IDDM system has been designed to manage system complexity by the introduction of structural constraints on the high-dimensional template subsystem. The resulting system was successful in achieving good results. However, the present research is still in its initial stage of development. As such, more research work needs to be carried out in this area.

Syed Zahid Hassan and Brijesh Verma, “A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases”, Seventh International Conference on Intelligent Systems Design and Applications, ISBN: 0-7695-2976-3/07/$25.00, Pg 503 - 508, IEEE, 2007.

In this paper, a hybrid classifier has been used for knowledge management, in the field of Medical Data Mining. Self Organizing Map (SOM), K-Means, and Naïve Bayes classifiers are used in conjunction with a neural network based classifier, to produce the hybrid classifier. Neural and statistical clustering is used to cluster the data into soft clusters. Serial and parallel fusion has been combined with a neural classifier, to fuse the clusters.

The idea was put forth for implementation. The classification accuracy of the Proposed Hybrid Approach was found to be higher than that of the individual algorithms. Next, classes were formed, and accuracies of the algorithms were compared, taking each class into consideration. Here, as well, the Proposed Hybrid Approach showed a higher accuracy rate.

The classifier was able to achieve over 92% classification accuracy on the *test set*, which has lit a spark for future analysis. Moreover, it provided for data visualization which further, helped in the interpretation of results. However, this paper presented the results using serial fusion only. The results from parallel fusion still remain to be collected and analyzed.

Sarojini Balakrishnan, Nickolas Savarimuthu, Ramaraj Narayanaswamy, Rita Samikannu, "SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases", 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC 2008), ISBN: 1-4244-2384-2/08/$20.00, Pg 2628 – 2633, IEEE, 2008

This research work deals with the data mining technique of classification analysis for healthcare applications that aid in improving the quality of life. The researchers aim to increase the predictive accuracy of the classifiers through feature selection techniques since medical databases involve highly dimensional datasets.

This paper lays emphasis on the need of pre-processing the data for Data Mining and Machine Learning. The proposed approach involves improving the classification accuracy of Naïve Bayesian Classifier through feature selection approach of finding an optimum feature subset. Further, emphasis is laid on reducing the False Negative Rate, which plays a crucial role in Medical Data Mining.

Research work was carried out on Pima Indian Diabetes Dataset using SVM Ranking with Backward Search approach. A number of feature subsets were extracted from the main dataset. The experiment showed promising results as 37.5% feature reduction led to an increase of 1.88% in the classification accuracy. The research work aims to achieve even higher classification accuracy on the sample medical datasets.

Quoc-Nam Tran, “Mining Medical Databases with Modified Gini Index Classification”, Fifth International Conference on Information Technology: New Generations, ISBN: 978-0-7695-3099-4/08/$25.00, Pg 195 – 200, IEEE, 2008.

In this paper, focus has been on Decision Tree classification, using Gini Index. The Gini indexes have been normalized to consider the information related to the splitting status of all attributes. The approach taken is to use the ratios of Gini indexes along with their splitting values, instead of Gini indexes for attribute selection.

The idea was put forth for implementation. The different Data Sets taken were Heart Disease Data Set, Breast Cancer Wisconsin Data Set, Liver Disorders Data Set, and Dermatology Data Set. The accuracy of the classifier was measured for each Data Set.

Modified Gini Index Classification helped in the reduction of biases. However, when this method was tried on several benchmark medical databases, it performed well for some databases, but may not be the best for others.

Lin Zhang, Yan Chen, Yan Liang, Nan Li, “Application of Data Mining Classification Algorithms in Customer Membership Card Classification Model”, 2008 International Conference on Information Management, Innovation Management and Industrial Engineering, ISBN: 978-0-7695-3435-0/08/$25.00, Pg 211 – 215, IEEE, 2008.

In this paper, C5.0 and CART Decision Tree classification algorithms have been used to obtain useful information that would help in decision-making out of customers' transaction behaviours. First of all, the results of the two algorithms were obtained, and then compared. Next, an application was built and analyzed for achieving successful results in business enterprises.

The idea was implemented and tested on a customer database. By Business Understanding, Data Understanding and Data Preparing, Modeling, and Evaluating, the results of the two algorithms were obtained. Both the algorithms produced near similar production rules. Next, the results were compared using the target database. The accuracy of the two algorithms was found to be above 80% each, while the combined accuracy was found to be above 95%.

Hence, the classification algorithms used in the Customer Membership Card Classification Model were successful in grouping the customers, and then advising the customer to use the corresponding card. This helped the enterprise in meeting the interests of different card users, thereby, adding to the growth of the enterprise.

Sedigheh Khajouei Nejad, Hamed Ahmadi, Farid Seifi, Nima Seifi, “Applying Data Mining in Prediction and Classification of Urban Traffic”, 2009 World Congress on Computer Science and Information Engineering, ISBN: 978-0-7695-3507-4/08/$25.00, Pg 674 – 678, IEEE, 2008.

The ultimate focus of this research work is traffic prediction. In this paper, Data Mining technology has been applied in the field of Traffic Management. These methods have been chosen to generate dependable patterns for traffic prediction. The algorithms used herein are the Decision Tree algorithms.

First of all, the training dataset was taken. Records of this dataset consisted of two parameters, Time and Temperature. Time denotes the time of sampling, while Temperature denotes the temperature of environment at sampling time. The classes formed for this dataset have the values of 1, 2, and 3 which indicate the traffic condition at that time. Number 1 is the indicator for very light traffic, 2 is the indicator for light traffic and 3 is the indicator for heavy traffic.

Using Decision Tree classification for Traffic Management and Traffic Prediction produced promising results. However, in this research work, we found that only two parameters were taken into consideration. Hence, a greater number of parameters must be taken into account to present a real life scenario. Moreover, other algorithms can be implemented in this area, and these algorithms can be combined to get more accurate predictions.

Hamidah Jantan, Abdul Razak Hamdan, Zulaiha Ali Othman, "Potential Data Mining Classification Techniques for Academic Talent Forecasting", 2009 Ninth International Conference on Intelligent Systems Design and Applications, ISBN: 978-0-7695-3872-3/09/$26.00, Pg 1173 - 1178, IEEE, 2009.

In this paper, a study is presented to solve the problems related to talent management at different organizations. The proposed approach uses classification and prediction techniques of Data Mining. The idea is to predict the talent performance, by using the knowledge of historical data, discovered from the existing databases. Decision Tree, Neural Network and Nearest Neighbour algorithms have been used to discover the talent pattern.

The idea was put forth for implementation on a human resource database. At the start, all the attributes for Academic Talent were presented. Then, the accuracy for each classifier algorithm was evaluated on the basis of full attributes. Secondly, few attributes were selected from amongst all the attributes. Then, the accuracy for each classifier algorithm was evaluated on the basis of attribute reduction. The accuracy in either case showed promising results.

It was found that C4.5 classification algorithm showed the highest accuracy amongst all the algorithms considered. However, other Data Mining techniques should also be applied, and checked for results. In some cases, the relevancy of the attributes selected, affect the accuracy of the classifier.

Alireza Kajabadi, Mohamad Hosein Saraee, Sedighe Asgari, "Data Mining Cardiovascular Risk Factors", ISBN: 978-1-4244-4740-4/09/$25.00, IEEE, 2009.

The main focus of this research work is to manage the vast amount of data present at medical units. In this paper, techniques have been used to investigate the data at medical centers, and obtain useful conclusions about different diseases. The focus has been to obtain useful relationships among the major risk factors of cardiovascular diseases. This research work uses decision tree as the classification technique and CART as the software.

The idea was put forth for implementation. LDL (Low Density Lipoprotein) was taken as the class variable. 27 predictor variables were involved in the purpose of classification. The results show that the most important variables with regards to LDL are: level of cholesterol, age, body mass index, APOB, triglyceride, (APOB/APOA) and smoking.

The results obtained after mining the cardiovascular dataset have lit a spark for further analysis. The selection of the target variable and the predictor variables plays an important role in determining the accuracy of the classifier. However, the research work was localized to a particular area. The usage of greater number of records could cause better results. Also, the datasets should be refined further to obtain higher accuracy for the classifier.

Yi Mao, Yixin Chen, Gregory Hackmann, Minmin Chen, Chenyang Lu, Marin Kollef, Thomas C. Bailey, "Medical Data Mining for Early Deterioration Warning in General Hospital Wards", 2011 11th IEEE International Conference on Data Mining Workshops, ISBN: 978-0-7695-4409-0/11/$26.00, Pg 1042 – 1049, IEEE, 2011.

The ultimate focus of this research work is to enhance the treatment quality of hospitals and increase the survival rate of patients. An Early Warning System (EWS) has been designed to identify the signs of clinical deterioration and to provide an early warning for serious clinical events at General Hospital Wards (GHWs). It predicts whether a patient should be transferred to the ICU or not based on the existing electronic medical record of the patient.

The sample medical dataset was obtained from a renowned medical hospital. As the sample medical dataset was highly dimensional in nature, it underwent preprocessing techniques to remove irregularity, multi-scale data gaps, measurement errors, outliers, and to solve the class imbalance problem. The proposed approach uses bucketing techniques, bootstrap aggregating schemes and smoothing schemes to handle the multi-dimensional medical dataset. Some of the major classification algorithms such as Logistic Regression and SVM are supported in this framework.

Next, the system was applied in a real-time clinical trial which showed promising results. After analysis, it was found that the system can give the alert for ICU transfers atleast 4 hours before the transfer time, and for death situations, atleast 30 hours before the death of the patient. Hence, the research work was successful in the use of Data Mining technology on digitized medical dataset.

Nazanin Shahrokhi, Roxana Dehzad, Soheila Sahami, “Targeting Customers with Data Mining Techniques: Classification”, 2011 International Conference on User Science and Engineering (i-USEr), ISBN: 978-1-4577-1655-3/11/$26.00, Pg 212 – 215, IEEE, 2011.

In this paper, classification technique has been applied to marketing concepts. The data mining approach has targeted customers, considering their past performance. Few classification algorithms were used to discover performance patterns from an existing dataset of passengers at an airport. Amongst them, C5.0 was considered to be the best suited for the scenario.

Various attributes related to the airports, as well as, the passengers were taken into account. A classification was performed on the basis of the airlines to determine the features of each airline, and results were obtained to suggest airlines for the passengers.

C5.0 presented 49% classification accuracy. Without using any classification technique, the probability that a customer picks up his or her best airline, from a set of 22 airlines, is 1/22. However, after using the classification technique, the accuracy rate increased to 49%. However, some problems exist with the quality of data. The data was collected through interviews, because of which the customers were not accessible for feedback. Also, some unknown errors might have affected the interviewing process. Moreover, other data mining techniques could also be applied, and on a larger dataset.

S. Yasodha, P. S. Prakash, "Data Mining Classification Technique for Talent Management using SVM", 2012 International Conference on Computing, Electronics and Electrical Technologies [ICCEET], ISBN: 978-1-4673-0210-4/12/$31.00, Pg 959 – 963, IEEE, 2012

This research work aims for Human Resource Management (HRM). It deals with the management of organizational talents which is a major challenge for HR professionals. In this study, the classification technique of Data Mining has been used to solve the talent management problem.

The researchers have suggested a new hybrid approach CACC-SVM for the classification of HR data. CACC discretization algorithm maps continuous values into known intervals, thereby, performing concise summarization before classification. SVM, which is a powerful generalized classifier and works mostly on numeric data, has been chosen for the purpose of classification. Hence, the proposed model has been generated using CACC discretization algorithm in conjunction with Sequential Minimal Optimization algorithm, a type of classification algorithm under the SVM classifier.

In this paper, a new hybrid approach was taken for the purpose of classification which was compared with the other standard classification techniques. This research work lays an emphasis on the need of a hybrid classifier for the purpose of classification. Focus has been laid on the best possible classification of HR data and its usage for future predictions.

Dr. R. Geetha Ramani, S. Vinodh Kumar, Shomona Gracia Jacob, "Predicting Fault-Prone Software Modules Using Feature Selection and Classification through Data Mining Algorithms", 2012 IEEE International Conference on Computational Intelligence and Computing Research, ISBN: 978-1-4673-1344-5/12/$31.00, IEEE, 2012.

The main focus of this research work has been software defect detection. In this paper, the performances of supervised machine learning techniques have been evaluated through Data Mining algorithms. This paper lays emphasis on the performances of classification algorithms in categorizing seven datasets under two classes. In order to conduct this research work, publicly available datasets were gathered from different organizations. This helped to monitor the impact of data from different sources on different processes for finding appropriate classification models.

A computational framework is proposed using Data Mining techniques to detect the existence of defects in software components. The framework comprises of data pre-processing, data classification and classifier evaluation. The performance of twenty classification algorithms on seven publicly available datasets from the NASA MDP Repository is reported. Random Tree classification algorithm produced 100% accuracy in classifying the datasets. Hence, the features selected by this technique were considered to be the most significant features.

Hence, Random Tree classification algorithm categorized the software modules with utmost accuracy. The detection of defective software can save the workers in the specific field, a lot of time and capital. Other classification algorithms must be taken into account for earlier software defect detection.

C. M. Velu and K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", 2012 3rd IEEE International Advance Computing Conference (IACC), ISBN: 978-1-4673-4529-3/12/$31.00, Pg 1070-1075, IEEE 2012.

The main focus of this research work is the classification of patients suffering from diabetes. The sample medical dataset was obtained from Pima Indian Diabetes (PID) dataset while the simulation tests were performed through the WEKA software tool.

In this research work, different clustering techniques have been used on the sample medical dataset for finding important patterns that would help in medical analysis. The different clustering algorithms used are Expectation-Maximization Algorithm, H-means Clustering, and Genetic Algorithm. Different attributes were taken into consideration for simulation tests. Patients with similar symptoms were grouped into the same clusters.

After analyzing the results, it was found that H-means Clustering and double crossover genetics process based techniques performed better than the other algorithms. This research work aims to use artificial intelligent tools for diabetes detection. The research claims that the causes of diabetes are still inexplicit in nature. The researchers have identified genetics and environmental factors as some of the diabetes causing factors.

M. Akhil Jabbar, Priti Chandra, B. L. Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection", 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), ISBN: 978-1-4673-5119-5/12/$31.00, Pg 628 – 634, IEEE, 2012.

In this paper, the focus has been on scoring the risk value of a patient suffering from heart disease. The algorithm used herein is Associative Classification. The focus has been on class association rules, which would help the physicians to make a judgment on the heart disease of a patient.

Associative Classification is a combination of Association Rule Mining and Classification techniques. The class association rules were generated using feature subset selection. Feature subset selection helped in the identification of relevant attributes. Next, Genetic Algorithm was applied for enhanced prediction of heart disease. Finally, the pattern of heart disease was analyzed.

Associative Classification was found to be fit in the field of Medical Data Mining. The feature selection methods helped to determine the attributes that shared a greater contribution in the prediction of heart disease. This helped to reduce the number of diagnostic tests which should be taken by a patient. However, the research work was limited to a small population. It should be carried out over a larger population.

T. John Peter and K. Somasundaram, “An Empirical Study on Prediction of Heart Disease using Classification Data Mining Techniques”, IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012) March 30, 31, 2012, ISBN: 978-81-909042-2-3, Pg 514 – 518, IEEE, 2012.

The motivation for this research work came from the field of Cardiovascular Science. In this paper, pattern recognition has been clubbed with data mining techniques, to generate risk prediction models in the field of Medical Data Mining. Classification data mining techniques have been chosen to model and classify the data. Naïve Bayes, Decision Tree, K-NN and Neural Network classifiers are used to generate risk prediction models for heart disease datasets.

The idea was tested and implemented on a medical database. First of all, the entire dataset was considered, and the classifiers were used to classify the samples. Next, few attributes were selected using the attribute selection methods, and the classifiers were used on the reduced features. The accuracy in either case showed promising results.

It was observed that Naïve Bayesian classifier showed the highest accuracy amongst all the classifiers considered. However, other techniques must be considered to further improve the classifier accuracy. Also, attribute selection methods play a critical role in deciding the reduced features.

Emanuel Weitschek, Giovanni Felici, Paola Bertolazzi, "Clinical Data Mining: Problems, Pitfalls and Solutions", 2013 24th International Workshop on Database and Expert Systems Applications, ISBN: 1529-4188/13/$26.00, Pg 90 – 94, IEEE, 2013

This research work aims to handle the common problems and pitfalls of Clinical Data Mining and to provide for their suitable solutions. The data for this research work has been collected from several clinical departments of Italy.

Huge amounts of clinical data are gathered at the health information systems every day. With the advancement of technology, most of the clinical data is stored in an electronic format in the computerized devices. The major activities of these information systems involve the collection, management, integration, and analysis of the medical data. The researchers have brought to light the problems of incompleteness, the different adopted measure scales, and the integration of disparate collection procedures as the major issues of medical data analysis. In this research work, the researchers have focused on the main challenges of managing clinical data, discovering patient interactions, and integrating the different data sources. The proposed approach deals with the field of logic classification in which a data model is computed in the form of propositional logic formulae, and investigated for clinical data mining. This approach when compared with other methods was found to be successful in computation of a compact data model for clinical knowledge discovery.

Hence, the main aim of this research work is to extract relevant information from huge amounts of, structured and unstructured, medical data. This study aims to develop an intelligent decision support system which would perform a completely automatic diagnosis of medical data.

M. Akhil Jabbar, Dr. B. L. Deekshatulu, Dr. Priti Chandra, “Heart Disease Prediction using Lazy Associative Classification”, ISBN: 978-1-4673-5090-7/13/$31.00, Pg 40 – 46, IEEE, 2013.

In this paper, research work has been implemented in the field of Medical Data Mining. A new rule based approach, known as Associative Classification is used, which combines Association Rule Mining and Classification techniques. A hybrid classifier has been generated by combining the Decision Tree classifier (J 48) and Naïve Bayesian classifier, which would help in the prediction of heart disease.

The idea was tested and implemented on a medical database. First of all, the risk factors of Coronary Heart Disease were identified. Next, the prevalence of these risk factors was evaluated. The individual and the hybrid classifiers were applied on the non-medical data sets, as well as, the medical data sets. The accuracy of each classifier was judged. The attributes related to heart disease were identified.

The Proposed Hybrid Classifier achieved greater accuracy than the individual classifiers. The generation of class association rules plays a major role in the prediction of heart disease. Also, this system is designed for the population of Andhra Pradesh. It can be further refined, so that it can be applied to other states, as well.

Sina Bahramirad, Aida Mustapha, Maryam Eshraghi, “Classification of Liver Disease Diagnosis: A Comparative Study”, ISBN: 978-1-4673-5256-7/13/$31.00, Pg 42 – 46, IEEE, 2013.

In this paper, real liver patient datasets have been investigated for building classification models in order to predict liver diagnosis. A total of eleven Data Mining classification algorithms were applied to the datasets and their performances were compared against each other.

Two liver patient datasets were chosen, one from India (AP Dataset) and the other one from USA (BUPA Dataset), for the purpose of building and testing the classification models. The attributes in each dataset were identified separately. Next, the different classification algorithms were applied, and their results were compared on the basis of the performance parameters: Accuracy, Precision and Recall. Next, Brute Force Optimization and Bayesian Boosting were applied to improve the classification accuracy.

The research showed promising results in diagnosing liver disease during the earlier stages. However, there were considerable differences in some of the performance parameters of the two datasets, which should be taken into account. Other optimization techniques should be applied to further improve the classification accuracy.

S. Udhaya Kumar, H. Hannah Inbarani, S. Senthil Kumar, "Bijective Soft Set Based Classification of Medical Data", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22, ISBN: 978-1-4673-5845-3/13/$31.00, Pg 517 – 521, IEEE, 2013.

The motivation for this research work arose from the field of Medical Data Mining. In this paper, classification rules are generated from the datasets, using Bijective Soft Set theory. In addition, the generated rules have been compared with the Decision Tree classification algorithms and Naïve Bayesian classification algorithms.

The idea was put forth for implementation on a sample medical database. The performance of the three algorithms was analyzed on the basis of the accuracy measures: Precision, Recall and F-Measure. In each of the cases, Bijective Soft Set achieved the highest accuracy.

Hence, Bijective Soft Set theory has been found to be efficient in discovering data dependencies, as well as, performing data reduction, classification and rule generation from databases. Further research work should be carried out in this field to make it valuable for inductive learning and for building expert systems.

G. Kesavaraj and Dr. S. Sukumaran, “A Study on Classification Techniques in Data Mining”, 4th ICCCNT - 2013, July 4 - 6, 2013, Tiruchengode, India, IEEE, 2013.

In this paper, the basic classification techniques are presented. Major kinds of classification methods, such as, Decision Tree Induction, Rule-Based Classification, Neural Networks, Bayesian Networks, and Support Vector Machines, have been reviewed.

The algorithms reviewed under Decision Tree Induction are Hunt’s Algorithm, Iterative Dichotomiser Algorithm, C4.5 Algorithm, and Rnd Tree (Random Forest) Algorithm. The Rule-Based Method has been considered by using examples of classification rules. The algorithm examined under Neural Networks section is the Feed Forward Neural Network. Next section describes the Bayesian Networks. Finally, we have the description of Support Vector Machines.

The algorithms were applied on the data of Facebook and Twitter. The accuracy of each individual algorithm was evaluated. The average accuracy rates of Supervised and Unsupervised Algorithms was also evaluated. However, we find that no single classification algorithm is best suited for all kinds of dataset. Classification algorithms are specific in their problem domain. Hence, we must focus our attention towards the creation of ensemble of classifiers.

Satej Wagle, J. Alamelu Mangai, V. Santosh Kumar, “An Improved Medical Image Classification Model using Data Mining Techniques”, 2013 IEEE GCC Conference and Exhibition, November 17-20, Doha, Qatar, ISBN: 978-1-4799-0724-3/13/$31.00, Pg 114-118, IEEE, 2013.

The motivation for this research work arose from the field of Medical Data Mining. Owning to the vast amounts of medical data flowing in at the health information systems, appropriate technology is required to handle the data and to retrieve useful information from them. With the advancement of technology, analysis of medical data involves images to a large extent.

This paper presents a modified KNN Classifier for prediction purposes. The modified KNN Classifier has been generated by using weighting techniques on the KNN classification algorithm. Apart from this, image pre-processing techniques are used to select the best representative features to classify an image, thereby, reducing the number of dimensions required for analysis. Retinal fundus images have been used as the sample medical dataset on this classification model.

The performance of the modified KNN Classifier was compared against the original KNN Classifier and with the other classification algorithms, on the basis of metrics such as classification accuracy and area under the ROC curve. The proposed classifier was found to perform better than most of the classifiers included in the study. Hence, the research work was successful in the generation of an improved method to classify the medical images.

CHAPTER 3

SCOPE OF THE STUDY

The scope of this study encompasses the field of Medical Data Mining. This study is aimed at the development of an intelligent agent which would classify the infectious and the non-infectious diseases with greater classification accuracy. Our research work is centered on machine learning techniques, such as clustering and classification, which handle all classes of sample medical datasets. Moreover, our research work uses image processing techniques, whenever required, to aid in the diagnosis of the diseases. Here, we build an Adaptive Classifier which works upon the sample medical datasets and produces more certain, precise and accurate results than the individual classifiers.

CHAPTER 4

OBJECTIVE OF THE STUDY

The objective of this study is to enhance the classification accuracy of existing classification algorithms, by using an Adaptive Classifier, developed from an ensemble of modified classifiers, in the field of Medical Data Mining. With an enhancement in the classification accuracy, we aim to achieve more certain, precise and accurate results on the sample medical datasets. We aim for an early detection of diseases at the health information systems, which would, ultimately, facilitate the survival rates of the patients. Overall, this research work is focused to improve the quality of life and the standard of living of the patients.

CHAPTER 5

RESEARCH METHODOLOGY

5.1 Research Problem Definition

Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining

5.2 Formulation of Hypothesis

μ_{H0} = Classification Accuracy of a Single Classifier

μ = Classification Accuracy of an Ensemble Classifier

$H_0 : \ \mu = \mu_{H0}$

$H_a : \ \mu > \mu_{H0}$

For the successful completion of this research work, H_0 must be rejected or H_a must be accepted, after the testing of hypothesis.

5.3 Research Design

For two class sample medical datasets, Laplacian Correction was applied to Decision Tree Classification algorithm, as well as, Naïve Bayesian Classification algorithm. This helped to generate an improved version of the two classifiers, namely, Decision-Tree Classifier (modified) and Naïve-Bayesian Classifier (modified).

The next step involved the use of *Bagging* as the ensemble method to improve the classification accuracy of our initial Adaptive Classifier.

Hence, the initial Adaptive Classifier was created by combining the techniques of Rule-Based Classifier, Decision-Tree Classifier (modified) and Naïve-Bayesian Classifier (modified).

However, for one class sample medical datasets, Laplacian Correction was applied to Naïve Bayesian Classification algorithm only. This helped to generate an improved version of the Naïve Bayesian Classifier, namely, Naïve Bayesian Classifier (modified).

As an alternative to the Decision-Tree Classifier (modified), K-means Clustering algorithm was used to generate the K-means Classifier.

The next step involved the use of *Bagging* as the ensemble method to improve the classification accuracy of our advanced Adaptive Classifier.

Hence, the advanced Adaptive Classifier was created by combining the techniques of Rule-Based Classifier, K-means Classifier and Naïve-Bayesian Classifier (modified).

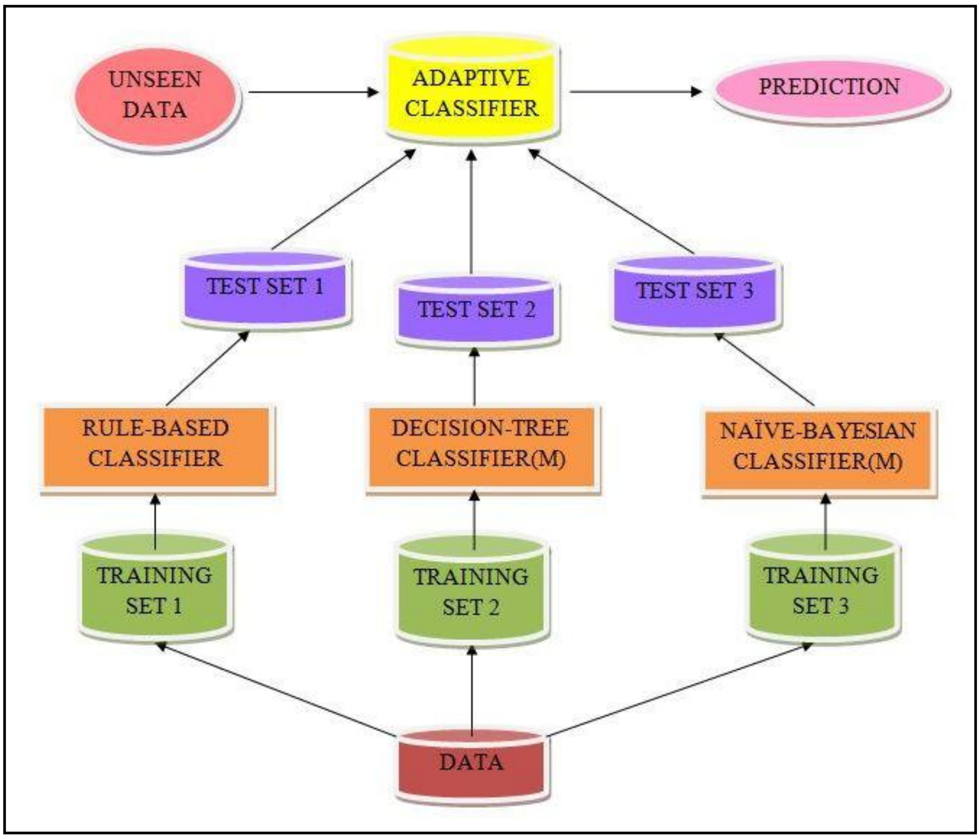

Figure 5.1: Initial Adaptive Classifier

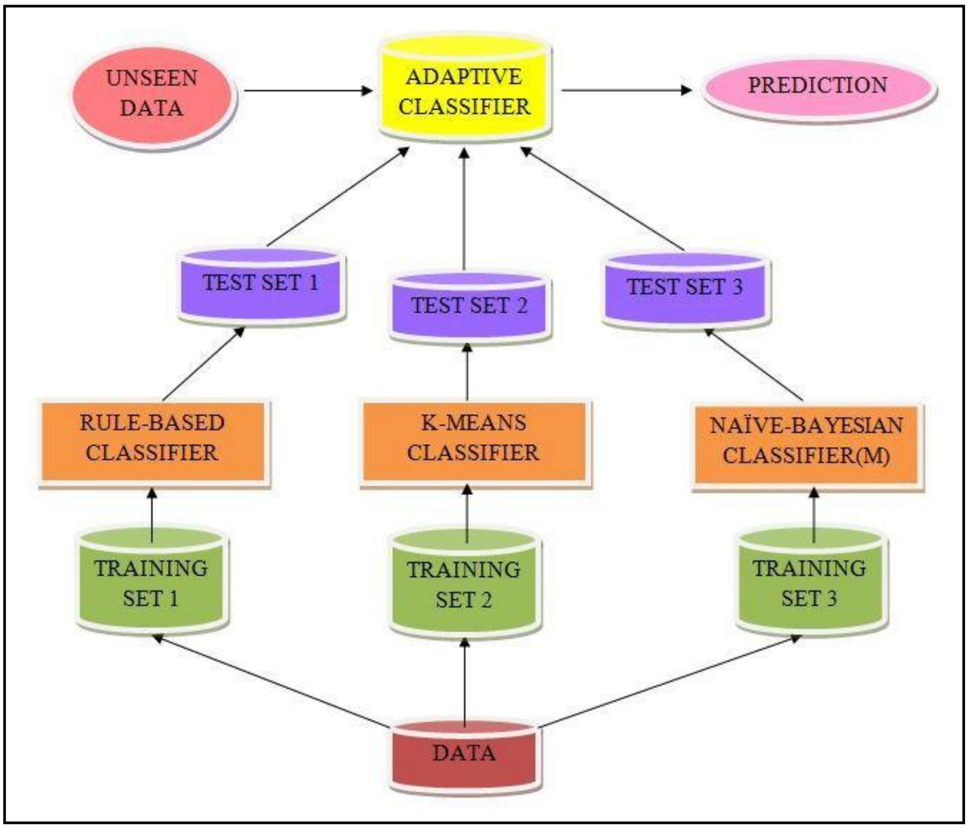

Figure 5.2: Advanced Adaptive Classifier

Many supportive medical tests aid in the diagnosis of Lymph Node Tuberculosis. Some of these tests include, Mantoux, ELISA and Tuberculin Tests. In Mantoux Test, the diameter of induration on the forearm is taken as a measure of positive Tuberculosis skin test. However, in this research work, lymph node diameter has been considered, along with Mantoux Test, as an important factor in the diagnosis of Tuberculosis. Image processing has been carried out on the lymph nodes of patients to measure their diameters (in pixels). The diameters so obtained helped to classify the lymph nodes into different classes, namely, Small, Medium and Large, based upon their measurements. The data were then fed into our Adaptive Classifier.

Ultrasound is an important medical imaging technique that aids in the diagnosis of Liver Cirrhosis. The disease is characterized by an enlarged liver for most of the stages, and by a shrunken liver in the last stage of its progression. Apart from this, liver has a tawny surface and its texture varies greatly during the progression of the disease. An ultrasound report provides for an exact measurement of the size of liver. However, the report provides a qualitative analysis of its texture only. In this research work, Mean Gray Level algorithm has been used to analyze the texture of liver presented in the ultrasound report, and provide for its quantitative measurement (in pixels). The texture of liver was categorized into different classes, namely, Normal and Altered, based upon its measurements. The data were then fed into our Adaptive Classifier.

The software tools used in this research work are as follows:

- Microsoft Visual Studio 2010
- Microsoft SQL Server Management Studio 2012
- MATLAB R2013a
- Microsoft Office Excel 2007

5.4 Collection of Data

The data for our research work have been collected from Nalanda Medical College Hospital (NMCH), and Patna Medical Hospital (PMCH), Patna, Bihar, India. The data were collected during the months of July and August, 2014 for the non-infectious disease Diabetes Mellitus and the infectious disease Pulmonary Tuberculosis, and during the months of November and December, 2014 for the non-infectious disease Liver Cirrhosis and the infectious disease Lymph Node Tuberculosis. The attributes of the patients included in the sample medical dataset are: Name, Age, Gender, Location, Occupation, Economic Status, Disease Symptoms, Medical Test Results and Disease Status. The raw data collected underwent a few pre-processing techniques, such as Data Cleaning and Data Discretization, so as to obtain better results, and to improve the performance of the proposed model [6].

The sample medical dataset for the non-infectious diseases Diabetes Mellitus and Liver Cirrhosis, and the infectious diseases Pulmonary Tuberculosis and Lymph Node Tuberculosis, is presented below. The names of the patients have been substituted with their initials, in order to maintain the confidentiality of the patients.

TABLE 5.1: Sample Medical Dataset (Diabetes Mellitus Dataset)

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PURIA	SYM2_PDIPHAGIA	SYM3_PHEALTH	DIS_STATUS
1	RD	Middle	Female	Urban	Homemaker	Medium	Negative	Positive	Positive	No
2	SD	Middle	Female	Urban	Homemaker	Medium	Positive	Negative	Positive	No
3	RPS	Middle	Male	Urban	Professional	High	Positive	Positive	Negative	Yes
4	DK	Middle	Male	Rural	Labourer	Low	Negative	Positive	Positive	No
5	UD	Middle	Female	Urban	Professional	Medium	Positive	Positive	Negative	No
6	PD	Middle	Female	Urban	Homemaker	High	Positive	Positive	Positive	Yes
7	CR	Senior	Male	Rural	Labourer	Low	Negative	Negative	Negative	No
8	JR	Middle	Male	Urban	Professional	Medium	Negative	Negative	Negative	No
9	LD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	No
10	BD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
11	MD	Middle	Female	Rural	Homemaker	High	Negative	Negative	Negative	No
12	FK	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
13	SRG	Middle	Male	Urban	Professional	Medium	Positive	Positive	Positive	Yes
14	RP	Senior	Male	Urban	Professional	Medium	Positive	Negative	Positive	No
15	ID	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
16	KD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	Yes
17	UD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
18	OP	Middle	Male	Urban	Professional	High	Negative	Negative	Negative	No
19	CD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
20	MD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
21	BD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Positive	No
22	ID	Senior	Female	Rural	Homemaker	Medium	Positive	Negative	Positive	No
23	RD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
24	MR	Middle	Male	Rural	Driver	Low	Negative	Negative	Negative	No
25	BD	Middle	Female	Urban	Homemaker	High	Negative	Negative	Negative	No
26	PD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	No
27	SK	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
28	MD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
29	GD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
30	NP	Senior	Male	Urban	Professional	High	Positive	Positive	Negative	Yes
31	MD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
32	DR	Senior	Male	Urban	Homemaker	Medium	Negative	Positive	Positive	No
33	JD	Middle	Female	Rural	Labourer	Low	Positive	Positive	Positive	Yes
34	NR	Middle	Male	Urban	Professional	Medium	Negative	Negative	Negative	No
35	RD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
36	KD	Senior	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
37	TD	Senior	Female	Urban	Homemaker	Medium	Positive	Negative	Positive	No
38	BRS	Middle	Male	Urban	Professional	High	Negative	Positive	Positive	No
39	LPK	Senior	Male	Urban	Professional	Medium	Positive	Positive	Negative	No
40	RKD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PURIA	SYM2_PDIPHAGIA	SYM3_PHEALTH	DIS_STATUS
41	KD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
42	MD	Middle	Female	Rural	Homemaker	Medium	Negative	Negative	Positive	No
43	LD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	No
44	KD	Senior	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
45	MJ	Senior	Male	Urban	Professional	High	Positive	Negative	Positive	No
46	CD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
47	AS	Middle	Male	Urban	Professional	High	Negative	Positive	Positive	No
48	GD	Middle	Female	Rural	Labourer	Low	Positive	Positive	Positive	Yes
49	MK	Middle	Female	Urban	Homemaker	Medium	Positive	Negative	Negative	No
50	PS	Middle	Male	Urban	Professional	High	Negative	Negative	Negative	No
51	KD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
52	RP	Middle	Male	Urban	Professional	High	Positive	Positive	Positive	Yes
53	MA	Senior	Male	Urban	Professional	Medium	Positive	Negative	Positive	No
54	NP	Middle	Male	Urban	Driver	Low	Negative	Positive	Positive	No
55	KK	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	No
56	SP	Senior	Male	Rural	Homemaker	High	Positive	Positive	Positive	Yes
57	DD	Middle	Female	Urban	Homemaker	High	Negative	Negative	Negative	No
58	BP	Middle	Male	Urban	Professional	High	Positive	Positive	Negative	Yes
59	LD	Senior	Female	Rural	Homemaker	Medium	Positive	Negative	Positive	No
60	AD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
61	SK	Middle	Male	Rural	Professional	High	Negative	Negative	Negative	No
62	ND	Middle	Female	Rural	Labourer	Low	Positive	Positive	Positive	Yes
63	UD	Senior	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
64	RD	Senior	Female	Rural	Labourer	Low	Negative	Positive	Positive	No
65	KD	Senior	Female	Rural	Homemaker	Medium	Negative	Negative	Negative	No
66	DD	Senior	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
67	MS	Senior	Male	Urban	Professional	High	Positive	Negative	Positive	No
68	KD	Middle	Female	Urban	Professional	Medium	Negative	Negative	Negative	No
69	SD	Middle	Female	Rural	Labourer	Low	Positive	Positive	Negative	No
70	PD	Middle	Female	Urban	Homemaker	High	Positive	Positive	Negative	Yes
71	RS	Middle	Male	Urban	Professional	High	Negative	Negative	Negative	No
72	DD	Middle	Female	Rural	Homemaker	Medium	Negative	Positive	Positive	No
73	MT	Middle	Male	Urban	Driver	Low	Negative	Negative	Negative	No
74	MD	Middle	Female	Rural	Homemaker	Medium	Positive	Positive	Positive	Yes
75	MD	Middle	Female	Urban	Professional	Medium	Negative	Negative	Negative	No
76	LP	Middle	Male	Urban	Professional	High	Negative	Positive	Negative	No
77	KD	Middle	Female	Rural	Labourer	Low	Positive	Negative	Positive	No
78	NT	Middle	Male	Urban	Professional	Medium	Positive	Positive	Positive	Yes
79	CD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
80	SP	Middle	Male	Urban	Professional	Medium	Positive	Positive	Negative	No

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PURIA	SYM2_PDIPHAGIA	SYM3_PHEALTH	DIS_STATUS
81	UD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
82	MD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Negative	Yes
83	RYS	Middle	Male	Urban	Driver	Low	Negative	Negative	Negative	No
84	MD	Middle	Female	Rural	Homemaker	Medium	Positive	Positive	Negative	No
85	SK	Middle	Male	Urban	Professional	High	Negative	Negative	Negative	No
86	NK	Senior	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
87	LD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	Negative	No
88	PD	Middle	Female	Urban	Professional	High	Negative	Positive	Positive	No
89	SD	Middle	Female	Rural	Labourer	Low	Negative	Negative	Negative	No
90	A	Middle	Male	Urban	Professional	High	Positive	Positive	Positive	Yes
91	YS	Senior	Male	Urban	Homemaker	Medium	Positive	Negative	Positive	No
92	CD	Middle	Female	Rural	Homemaker	Medium	Negative	Positive	Positive	No
93	SK	Middle	Male	Urban	Driver	Low	Negative	Negative	Negative	No
94	RKD	Middle	Female	Rural	Homemaker	Medium	Positive	Positive	Negative	Yes
95	SP	Middle	Male	Urban	Driver	Low	Negative	Negative	Negative	No
96	SD	Senior	Female	Rural	Homemaker	Medium	Positive	Positive	Negative	No
97	KD	Middle	Female	Rural	Labourer	Low	Positive	Positive	Positive	Yes
98	KP	Senior	Male	Urban	Professional	High	Positive	Negative	Negative	No
99	MD	Middle	Female	Urban	Homemaker	Medium	Positive	Positive	Positive	Yes
100	AP	Middle	Male	Rural	Labourer	Low	Negative	Negative	Negative	No

INDEX:

PID = Patient ID

NAME = Patient_Name

ECO_STATUS = Economic _Status

SYM1_PURIA = Symptom1_Polyuria

SYM2_PDIPHAGIA = Symptom2_Polydipsia
+ Symptom2_Polyphagia

SYM3_PHEALTH = Symptom3_Poor_Health

DIS_STATUS = Diabetes_Mellitus_Disease_Status

TABLE 5.2: Sample Medical Dataset (Pulmonary Tuberculosis Dataset)

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PCOUGH	SYM2_PHEALTH	DIS_STATUS
1	SD	Middle	Female	Rural	Homemaker	Medium	Negative	Positive	No
2	SNP	Middle	Male	Urban	Professional	High	Negative	Negative	No
3	MD	Youth	Female	Rural	Labourer	Low	Positive	Positive	Yes
4	AS	Youth	Male	Rural	Driver	Low	Negative	Negative	No
5	SK	Middle	Male	Urban	Labourer	Low	Negative	Positive	No
6	TK	Youth	Male	Rural	Labourer	Low	Positive	Negative	Yes
7	BR	Senior	Male	Rural	Labourer	Low	Negative	Positive	No
8	CD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	No
9	SM	Middle	Male	Urban	Driver	Low	Negative	Negative	No
10	KS	Senior	Male	Rural	Homemaker	Medium	Positive	Positive	Yes
11	RBS	Senior	Male	Urban	Professional	High	Negative	Positive	No
12	KD	Youth	Female	Rural	Homemaker	Low	Positive	Negative	No
13	SSS	Middle	Male	Rural	Professional	High	Negative	Positive	No
14	CS	Middle	Male	Rural	Homemaker	Medium	Negative	Negative	No
15	PK	Youth	Female	Urban	Student	Medium	Positive	Positive	Yes
16	SKM	Senior	Male	Rural	Homemaker	Low	Negative	Negative	No
17	PK	Middle	Male	Rural	Driver	Low	Negative	Positive	No
18	SD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	Yes
19	GK	Youth	Female	Urban	Professional	Medium	Negative	Negative	No
20	LM	Middle	Male	Urban	Labourer	Low	Negative	Positive	No
21	ND	Middle	Female	Rural	Labourer	Low	Positive	Negative	No
22	UD	Middle	Female	Rural	Homemaker	Low	Negative	Negative	No
23	BP	Middle	Male	Urban	Driver	Low	Positive	Positive	Yes
24	MS	Senior	Male	Rural	Homemaker	Medium	Negative	Negative	No
25	BM	Middle	Male	Urban	Driver	Low	Positive	Negative	No
26	KK	Youth	Female	Rural	Labourer	Low	Negative	Negative	No
27	USC	Middle	Male	Rural	Labourer	Low	Positive	Positive	Yes
28	MD	Middle	Female	Rural	Homemaker	Medium	Negative	Negative	No
29	SD	Senior	Female	Urban	Homemaker	Medium	Negative	Positive	No
30	SK	Youth	Male	Rural	Driver	Low	Positive	Negative	No
31	MD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	Yes
32	LD	Middle	Female	Urban	Homemaker	Medium	Negative	Negative	No
33	SK	Youth	Male	Rural	Student	Medium	Negative	Positive	No
34	AM	Middle	Male	Urban	Labourer	Low	Negative	Negative	No
35	PC	Middle	Male	Rural	Driver	Low	Positive	Positive	Yes
36	RK	Youth	Female	Urban	Student	Medium	Negative	Negative	No
37	MP	Middle	Male	Rural	Labourer	Low	Positive	Negative	No
38	SD	Middle	Female	Rural	Homemaker	Low	Negative	Positive	No
39	RSY	Senior	Male	Rural	Homemaker	Medium	Negative	Negative	No
40	SD	Youth	Female	Rural	Labourer	Low	Positive	Negative	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PCOUGH	SYM2_PHEALTH	DIS_STATUS
41	RK	Youth	Female	Urban	Student	Medium	Negative	Negative	No
42	PK	Youth	Female	Rural	Labourer	Low	Positive	Negative	No
43	SK	Youth	Female	Rural	Student	Medium	Negative	Positive	No
44	MD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	No
45	AKS	Youth	Male	Rural	Student	Medium	Positive	Positive	Yes
46	LP	Senior	Male	Urban	Homemaker	Medium	Negative	Negative	No
47	SD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	No
48	AK	Youth	Male	Rural	Driver	Low	Negative	Negative	No
49	TP	Middle	Male	Urban	Driver	Low	Positive	Positive	Yes
50	RP	Middle	Male	Rural	Labourer	Low	Negative	Negative	No
51	KD	Middle	Female	Rural	Homemaker	Medium	Negative	Negative	No
52	SP	Middle	Male	Urban	Driver	Low	Negative	Positive	No
53	DK	Youth	Male	Rural	Labourer	Low	Positive	Positive	Yes
54	SR	Middle	Male	Rural	Professional	Medium	Negative	Negative	No
55	SD	Youth	Female	Rural	Homemaker	Low	Positive	Negative	No
56	RK	Youth	Female	Rural	Labourer	Low	Positive	Negative	Yes
57	PK	Youth	Male	Rural	Driver	Low	Negative	Negative	No
58	SP	Middle	Male	Rural	Professional	Medium	Negative	Positive	No
59	AK	Middle	Male	Rural	Driver	Low	Positive	Positive	Yes
60	MD	Middle	Female	Rural	Homemaker	Low	Negative	Negative	No
61	SK	Youth	Male	Urban	Driver	Low	Negative	Positive	No
62	US	Middle	Male	Urban	Professional	Medium	Negative	Negative	No
63	AK	Youth	Male	Rural	Labourer	Low	Positive	Positive	Yes
64	KD	Youth	Female	Urban	Homemaker	Medium	Negative	Negative	No
65	R	Youth	Male	Rural	Labourer	Low	Positive	Negative	No
66	RD	Middle	Female	Rural	Homemaker	Low	Negative	Positive	No
67	CK	Youth	Female	Rural	Labourer	Low	Positive	Negative	Yes
68	RV	Middle	Male	Urban	Professional	Medium	Negative	Negative	No
69	SK	Youth	Male	Rural	Driver	Low	Positive	Negative	No
70	SD	Senior	Female	Urban	Homemaker	Medium	Negative	Positive	No
71	TB	Middle	Male	Rural	Labourer	Low	Negative	Negative	No
72	DNT	Middle	Male	Urban	Driver	Low	Positive	Positive	Yes
73	SP	Youth	Male	Rural	Student	Medium	Negative	Negative	No
74	MK	Youth	Male	Rural	Driver	Low	Positive	Negative	No
75	MD	Youth	Female	Rural	Labourer	Low	Negative	Negative	No
76	SD	Middle	Female	Rural	Homemaker	Low	Positive	Negative	Yes
77	MK	Middle	Female	Urban	Professional	Medium	Negative	Negative	No
78	KK	Youth	Male	Rural	Labourer	Low	Negative	Positive	No
79	BP	Middle	Male	Rural	Driver	Low	Positive	Negative	No
80	RD	Middle	Female	Rural	Homemaker	Medium	Negative	Positive	No

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_PCOUGH	SYM2_PHEALTH	DIS_STATUS
81	CK	Youth	Female	Rural	Labourer	Low	Positive	Positive	Yes
82	MS	Youth	Male	Rural	Driver	Low	Negative	Negative	No
83	BS	Middle	Male	Urban	Professional	High	Positive	Negative	No
84	RD	Youth	Female	Rural	Homemaker	Low	Positive	Positive	Yes
85	SR	Middle	Male	Urban	Professional	Medium	Negative	Positive	No
86	JD	Youth	Male	Rural	Driver	Low	Positive	Negative	No
87	RS	Senior	Male	Rural	Homemaker	Medium	Negative	Negative	No
88	VK	Youth	Male	Rural	Driver	Low	Positive	Negative	Yes
89	MS	Middle	Male	Urban	Professional	Medium	Negative	Negative	No
90	BP	Middle	Male	Rural	Professional	Medium	Negative	Positive	No
91	SD	Middle	Female	Rural	Homemaker	Low	Negative	Negative	No
92	GD	Youth	Female	Urban	Labourer	Low	Positive	Positive	Yes
93	LD	Middle	Female	Rural	Homemaker	Low	Negative	Negative	No
94	D	Youth	Male	Rural	Student	Medium	Negative	Positive	No
95	RBS	Senior	Male	Urban	Professional	High	Positive	Negative	No
96	LD	Middle	Female	Rural	Homemaker	Low	Negative	Negative	No
97	ND	Middle	Male	Urban	Labourer	Low	Positive	Positive	Yes
98	BD	Youth	Female	Rural	Labourer	Low	Negative	Positive	No
99	BD	Youth	Female	Rural	Homemaker	Low	Positive	Negative	No
100	ND	Middle	Female	Rural	Homemaker	Low	Negative	Positive	No

INDEX:

PID = Patient ID

NAME = Patient_Name

ECO_STATUS = Economic _Status

SYM1_PCOUGH = Symptom1_Persistent_Cough

SYM2_PHEALTH = Symptom2_Poor_Health

DIS_STATUS = Pulmonary_Tuberculosis_Disease_Status

TABLE 5.3: Sample Medical Dataset (Liver Cirrhosis Dataset)

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	LIVER_SIZE	LIVER_TEXTURE	SYM_PHEALTH	DIS_STATUS
1	PS	Middle	Male	Rural	Professional	Medium	Large	Altered	Positive	Yes
2	AK	Youth	Male	Urban	Driver	Low	Large	Altered	Negative	Yes
3	BD	Youth	Female	Rural	Labourer	Low	Medium	Altered	Positive	Yes
4	PK	Youth	Female	Rural	Student	Low	Medium	Normal	Negative	Yes
5	KB	Middle	Female	Urban	Homemaker	High	Large	Altered	Positive	Yes
6	BD	Middle	Female	Rural	Labourer	Low	Small	Altered	Positive	Yes
7	PS	Middle	Male	Rural	Driver	Low	Large	Altered	Positive	Yes
8	SA	Middle	Male	Urban	Driver	Low	Small	Altered	Positive	Yes
9	GS	Middle	Male	Rural	Labourer	Low	Medium	Normal	Negative	Yes
10	MG	Middle	Male	Rural	Driver	Low	Large	Altered	Positive	Yes
11	GD	Middle	Female	Urban	Homemaker	Medium	Medium	Altered	Positive	Yes
12	SD	Middle	Female	Rural	Labourer	Low	Small	Altered	Positive	Yes
13	SKT	Middle	Male	Urban	Professional	High	Large	Altered	Positive	Yes
14	NS	Youth	Male	Rural	Driver	Low	Medium	Altered	Positive	Yes
15	NY	Middle	Male	Rural	Labourer	Low	Large	Altered	Negative	Yes
16	NS	Senior	Male	Urban	Homemaker	Medium	Small	Altered	Positive	Yes
17	SBP	Middle	Male	Rural	Professional	High	Large	Altered	Positive	Yes
18	SY	Middle	Male	Rural	Labourer	Low	Medium	Normal	Negative	Yes
19	PK	Middle	Male	Rural	Driver	Low	Large	Altered	Positive	Yes
20	MC	Middle	Male	Urban	Professional	High	Small	Altered	Positive	Yes
21	AP	Middle	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
22	SNJ	Middle	Male	Rural	Homemaker	High	Large	Altered	Positive	Yes
23	RK	Youth	Male	Urban	Professional	Medium	Medium	Altered	Negative	Yes
24	CK	Youth	Female	Rural	Labourer	Low	Medium	Altered	Positive	Yes
25	RRS	Middle	Male	Urban	Professional	Medium	Large	Altered	Positive	Yes
26	DK	Middle	Male	Rural	Driver	Low	Medium	Altered	Negative	Yes
27	PS	Middle	Male	Rural	Homemaker	High	Small	Altered	Positive	Yes
28	AK	Youth	Male	Rural	Labourer	Low	Medium	Normal	Positive	Yes
29	MU	Middle	Male	Rural	Driver	Low	Large	Altered	Positive	Yes
30	CDD	Middle	Male	Urban	Professional	High	Small	Altered	Positive	Yes
31	GP	Youth	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
32	CK	Youth	Male	Urban	Driver	Low	Medium	Altered	Negative	Yes
33	CMK	Youth	Male	Urban	Student	Medium	Large	Altered	Positive	Yes
34	SG	Middle	Male	Rural	Professional	Medium	Large	Altered	Positive	Yes
35	MM	Middle	Male	Rural	Labourer	Low	Small	Altered	Positive	Yes
36	WA	Middle	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
37	MD	Youth	Female	Urban	Homemaker	Medium	Medium	Normal	Negative	Yes
38	SK	Middle	Male	Urban	Professional	High	Medium	Altered	Positive	Yes
39	AKJ	Middle	Male	Rural	Driver	Low	Small	Altered	Positive	Yes
40	GD	Middle	Female	Rural	Homemaker	Low	Small	Altered	Positive	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	LIVER_SIZE	LIVER_TEXTURE	SYM_PHEALTH	DIS_STATUS
41	RKT	Middle	Male	Rural	Professional	High	Medium	Altered	Negative	Yes
42	SNM	Middle	Male	Urban	Driver	Low	Medium	Altered	Positive	Yes
43	VBS	Middle	Male	Rural	Professional	Medium	Small	Altered	Positive	Yes
44	AC	Middle	Male	Rural	Driver	Low	Large	Altered	Positive	Yes
45	AR	Middle	Male	Rural	Professional	High	Medium	Altered	Positive	Yes
46	CD	Middle	Female	Rural	Labourer	Low	Small	Altered	Negative	Yes
47	MK	Middle	Male	Urban	Driver	Low	Medium	Normal	Positive	Yes
48	MR	Middle	Female	Urban	Homemaker	High	Small	Altered	Positive	Yes
49	JY	Senior	Male	Rural	Labourer	Low	Large	Altered	Positive	Yes
50	MD	Middle	Female	Urban	Professional	Medium	Small	Altered	Negative	Yes
51	SS	Middle	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
52	RK	Youth	Male	Rural	Student	Medium	Small	Altered	Positive	Yes
53	IK	Youth	Female	Rural	Student	Medium	Medium	Altered	Positive	Yes
54	RC	Middle	Male	Urban	Professional	High	Small	Altered	Positive	Yes
55	DN	Youth	Male	Rural	Labourer	Low	Medium	Normal	Negative	Yes
56	RPM	Middle	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
57	TS	Middle	Male	Rural	Professional	Medium	Small	Altered	Positive	Yes
58	RNS	Senior	Male	Rural	Homemaker	High	Large	Altered	Positive	Yes
59	RK	Youth	Male	Urban	Professional	Medium	Medium	Altered	Negative	Yes
60	AKJ	Middle	Male	Rural	Professional	Medium	Medium	Altered	Positive	Yes
61	SR	Middle	Male	Rural	Driver	Low	Small	Altered	Positive	Yes
62	RS	Middle	Male	Urban	Homemaker	Medium	Large	Altered	Positive	Yes
63	KD	Youth	Female	Rural	Labourer	Low	Medium	Altered	Positive	Yes
64	YP	Middle	Male	Rural	Labourer	Low	Large	Altered	Positive	Yes
65	PKV	Youth	Male	Rural	Driver	Low	Medium	Altered	Negative	Yes
66	RK	Youth	Female	Urban	Student	Medium	Medium	Normal	Positive	Yes
67	SP	Senior	Male	Rural	Labourer	Low	Large	Altered	Positive	Yes
68	CC	Youth	Male	Rural	Professional	Medium	Medium	Altered	Positive	Yes
69	AS	Middle	Male	Rural	Labourer	Low	Small	Altered	Positive	Yes
70	AK	Youth	Male	Urban	Professional	Medium	Medium	Altered	Negative	Yes
71	KS	Senior	Male	Rural	Homemaker	High	Small	Altered	Positive	Yes
72	CD	Youth	Male	Urban	Professional	Medium	Small	Altered	Negative	Yes
73	HPS	Senior	Male	Rural	Labourer	Low	Small	Altered	Positive	Yes
74	RK	Youth	Male	Rural	Driver	Low	Medium	Normal	Positive	Yes
75	BR	Middle	Male	Rural	Labourer	Low	Medium	Altered	Positive	Yes
76	KD	Youth	Female	Urban	Labourer	Low	Small	Altered	Negative	Yes
77	MM	Middle	Male	Urban	Professional	Medium	Medium	Altered	Positive	Yes
78	LP	Middle	Male	Urban	Homemaker	High	Large	Altered	Positive	Yes
79	IDS	Middle	Male	Rural	Professional	Medium	Small	Altered	Positive	Yes
80	MK	Youth	Male	Rural	Student	Low	Medium	Altered	Negative	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	LIVER_SIZE	LIVER_TEXTURE	SYM_PHEALTH	DIS_STATUS
81	RKG	Middle	Male	Rural	Professional	Medium	Large	Altered	Positive	Yes
82	JM	Middle	Male	Rural	Labourer	Low	Small	Altered	Positive	Yes
83	SD	Middle	Female	Rural	Labourer	Low	Medium	Normal	Positive	Yes
84	AY	Middle	Male	Rural	Labourer	Low	Small	Altered	Positive	Yes
85	DK	Middle	Male	Urban	Driver	Low	Large	Altered	Positive	Yes
86	SR	Middle	Male	Urban	Professional	Medium	Medium	Altered	Positive	Yes
87	JS	Middle	Male	Urban	Professional	High	Large	Altered	Negative	Yes
88	MK	Youth	Female	Rural	Labourer	Low	Small	Altered	Negative	Yes
89	MR	Senior	Male	Rural	Professional	High	Medium	Altered	Positive	Yes
90	DK	Middle	Male	Rural	Driver	Low	Small	Altered	Negative	Yes
91	SK	Youth	Male	Rural	Labourer	Low	Large	Altered	Positive	Yes
92	SK	Youth	Female	Rural	Student	Medium	Medium	Normal	Negative	Yes
93	RIS	Middle	Male	Urban	Professional	Medium	Large	Altered	Positive	Yes
94	KMS	Middle	Male	Urban	Professional	High	Small	Altered	Negative	Yes
95	AS	Middle	Male	Rural	Homemaker	Medium	Medium	Altered	Positive	Yes
96	SD	Middle	Female	Rural	Labourer	Low	Medium	Altered	Positive	Yes
97	PS	Middle	Male	Rural	Labourer	Low	Large	Altered	Positive	Yes
98	VK	Youth	Male	Urban	Driver	Low	Medium	Normal	Negative	Yes
99	MS	Middle	Male	Rural	Professional	Medium	Large	Altered	Positive	Yes
100	MA	Youth	Male	Urban	Labourer	Low	Medium	Altered	Positive	Yes

INDEX:

PID = Patient ID

NAME = Patient_Name

ECO_STATUS = Economic _Status

LIVER_SIZE = Liver_Size

LIVER_TEXTURE = Liver_Texture

SYM_PHEALTH = Symptom_Poor_Health

DIS_STATUS = Liver_Cirrhosis_Disease_Status

TABLE 5.4: Sample Medical Dataset (Lymph Node Tuberculosis Dataset)

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_LNSIZE	SYM2_PHEALTH	MANTOUX_TEST	DIS_STATUS
1	SK	Youth	Male	Urban	Student	Medium	Small	Positive	Negative	Yes
2	DVP	Youth	Male	Urban	Driver	Low	Medium	Positive	Positive	Yes
3	AK	Youth	Male	Rural	Student	Medium	Medium	Negative	Positive	Yes
4	UD	Middle	Female	Urban	Homemaker	High	Large	Positive	Positive	Yes
5	SK	Youth	Male	Urban	Homemaker	High	Medium	Positive	Negative	Yes
6	MAF	Youth	Male	Urban	Student	Medium	Large	Positive	Positive	Yes
7	SK	Youth	Male	Urban	Labourer	Low	Small	Positive	Positive	Yes
8	SD	Youth	Female	Rural	Homemaker	Medium	Small	Positive	Positive	Yes
9	BK	Youth	Male	Rural	Labourer	Low	Medium	Negative	Positive	Yes
10	PK	Youth	Female	Rural	Homemaker	Medium	Medium	Positive	Positive	Yes
11	SK	Youth	Female	Rural	Homemaker	Medium	Large	Positive	Positive	Yes
12	AK	Youth	Male	Rural	Homemaker	Medium	Small	Negative	Negative	Yes
13	RK	Youth	Female	Rural	Student	Low	Medium	Positive	Positive	Yes
14	A	Youth	Female	Urban	Student	Medium	Medium	Positive	Positive	Yes
15	HP	Youth	Female	Urban	Student	Medium	Small	Positive	Positive	Yes
16	JK	Youth	Female	Rural	Homemaker	High	Medium	Positive	Positive	Yes
17	SP	Middle	Male	Urban	Professional	High	Medium	Positive	Negative	Yes
18	RJK	Youth	Male	Rural	Labourer	Low	Large	Negative	Positive	Yes
19	ID	Youth	Female	Rural	Homemaker	Medium	Medium	Positive	Positive	Yes
20	SK	Youth	Female	Rural	Labourer	Low	Small	Positive	Positive	Yes
21	SD	Middle	Female	Rural	Homemaker	Low	Medium	Positive	Positive	Yes
22	BS	Middle	Male	Rural	Professional	High	Large	Positive	Positive	Yes
23	KK	Youth	Female	Rural	Student	Medium	Small	Positive	Negative	Yes
24	SK	Youth	Male	Rural	Driver	Low	Medium	Negative	Positive	Yes
25	SM	Youth	Male	Rural	Driver	Low	Large	Positive	Positive	Yes
26	PLD	Middle	Female	Rural	Homemaker	Medium	Medium	Positive	Positive	Yes
27	SR	Youth	Male	Rural	Student	High	Medium	Negative	Negative	Yes
28	DP	Youth	Female	Rural	Student	Medium	Small	Positive	Positive	Yes
29	PK	Youth	Female	Rural	Labourer	Low	Small	Positive	Positive	Yes
30	RD	Youth	Female	Urban	Homemaker	High	Large	Positive	Positive	Yes
31	SS	Middle	Male	Rural	Professional	High	Medium	Negative	Positive	Yes
32	CK	Youth	Female	Urban	Student	Medium	Medium	Positive	Negative	Yes
33	PD	Youth	Female	Rural	Labourer	Low	Small	Positive	Positive	Yes
34	KK	Youth	Female	Rural	Labourer	Low	Medium	Positive	Positive	Yes
35	R	Youth	Male	Urban	Student	Medium	Medium	Negative	Positive	Yes
36	S	Youth	Male	Urban	Driver	Low	Large	Positive	Positive	Yes
37	RD	Youth	Female	Rural	Labourer	Low	Large	Positive	Positive	Yes
38	JD	Youth	Female	Rural	Labourer	Low	Small	Positive	Negative	Yes
39	CK	Youth	Female	Rural	Student	Medium	Medium	Positive	Positive	Yes
40	SK	Youth	Male	Rural	Student	Medium	Medium	Positive	Positive	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_LNSIZE	SYM2_PHEALTH	MANTOUX_TEST	DIS_STATUS
41	MA	Youth	Male	Rural	Driver	Low	Large	Positive	Positive	Yes
42	P	Youth	Female	Rural	Labourer	Low	Medium	Positive	Positive	Yes
43	LS	Youth	Female	Rural	Student	High	Medium	Negative	Positive	Yes
44	SS	Youth	Female	Urban	Professional	Medium	Small	Positive	Negative	Yes
45	SP	Youth	Female	Rural	Student	Medium	Medium	Positive	Positive	Yes
46	AK	Youth	Male	Rural	Homemaker	High	Medium	Negative	Positive	Yes
47	LB	Middle	Male	Rural	Professional	High	Large	Positive	Positive	Yes
48	SK	Youth	Male	Rural	Labourer	Low	Small	Positive	Positive	Yes
49	P	Youth	Female	Rural	Student	Low	Small	Positive	Negative	Yes
50	LD	Middle	Female	Rural	Professional	High	Small	Negative	Positive	Yes
51	TR	Middle	Male	Rural	Homemaker	Medium	Medium	Positive	Positive	Yes
52	PD	Youth	Female	Rural	Homemaker	Medium	Medium	Positive	Positive	Yes
53	K	Youth	Female	Urban	Student	High	Small	Positive	Negative	Yes
54	NJB	Middle	Female	Rural	Homemaker	Medium	Small	Positive	Positive	Yes
55	LD	Middle	Female	Urban	Professional	Medium	Large	Positive	Positive	Yes
56	RD	Youth	Male	Urban	Driver	Low	Large	Positive	Positive	Yes
57	AK	Youth	Male	Rural	Student	Medium	Small	Positive	Positive	Yes
58	G	Youth	Male	Urban	Driver	Low	Large	Positive	Negative	Yes
59	SK	Youth	Female	Urban	Labourer	Low	Medium	Positive	Positive	Yes
60	MD	Youth	Female	Urban	Homemaker	High	Medium	Positive	Positive	Yes
61	PK	Youth	Female	Rural	Student	Medium	Small	Negative	Positive	Yes
62	AK	Youth	Male	Rural	Student	Medium	Small	Positive	Negative	Yes
63	R	Youth	Male	Urban	Labourer	Low	Large	Positive	Positive	Yes
64	SK	Youth	Female	Rural	Student	Low	Medium	Positive	Positive	Yes
65	RK	Youth	Male	Urban	Homemaker	High	Medium	Negative	Positive	Yes
66	AK	Youth	Female	Urban	Student	Medium	Small	Positive	Positive	Yes
67	AK	Youth	Female	Urban	Student	Medium	Large	Positive	Positive	Yes
68	AD	Youth	Female	Urban	Homemaker	High	Medium	Positive	Negative	Yes
69	MK	Youth	Male	Urban	Professional	Medium	Medium	Positive	Positive	Yes
70	GD	Youth	Female	Urban	Professional	High	Large	Negative	Positive	Yes
71	SD	Middle	Female	Rural	Homemaker	Low	Large	Positive	Positive	Yes
72	SK	Middle	Male	Rural	Professional	High	Medium	Positive	Positive	Yes
73	PD	Middle	Female	Rural	Homemaker	Medium	Medium	Positive	Negative	Yes
74	NKM	Middle	Male	Urban	Driver	Low	Small	Positive	Positive	Yes
75	MD	Middle	Female	Urban	Homemaker	Medium	Large	Positive	Positive	Yes
76	NP	Youth	Female	Urban	Student	Medium	Large	Positive	Positive	Yes
77	SD	Youth	Female	Rural	Labourer	Low	Medium	Positive	Positive	Yes
78	AP	Middle	Male	Rural	Professional	Medium	Medium	Positive	Negative	Yes
79	RD	Middle	Female	Urban	Homemaker	Medium	Large	Positive	Positive	Yes
80	AKS	Middle	Male	Urban	Professional	Medium	Large	Positive	Positive	Yes

PID	NAME	AGE	GENDER	LOCATION	OCCUPATION	ECO_STATUS	SYM1_LNSIZE	SYM2_PHEALTH	MANTOUX_TEST	DIS_STATUS
81	KD	Youth	Female	Urban	Labourer	Low	Medium	Negative	Positive	Yes
82	SK	Middle	Male	Rural	Driver	Low	Medium	Positive	Positive	Yes
83	AK	Youth	Male	Rural	Student	Medium	Small	Positive	Negative	Yes
84	RS	Youth	Male	Rural	Driver	Low	Large	Positive	Positive	Yes
85	AK	Middle	Male	Rural	Labourer	Low	Large	Positive	Positive	Yes
86	AK	Youth	Female	Rural	Student	Low	Medium	Positive	Positive	Yes
87	GD	Middle	Female	Rural	Homemaker	Medium	Medium	Positive	Negative	Yes
88	PK	Youth	Male	Rural	Student	Low	Medium	Negative	Positive	Yes
89	RD	Youth	Female	Urban	Labourer	Low	Large	Positive	Positive	Yes
90	PK	Youth	Female	Rural	Student	Medium	Medium	Positive	Positive	Yes
91	SD	Middle	Female	Urban	Homemaker	Medium	Large	Positive	Positive	Yes
92	PD	Youth	Female	Urban	Professional	High	Medium	Positive	Positive	Yes
93	DK	Youth	Female	Rural	Student	Medium	Small	Negative	Negative	Yes
94	ID	Youth	Female	Rural	Labourer	Low	Medium	Positive	Positive	Yes
95	MK	Youth	Female	Urban	Labourer	Low	Medium	Positive	Positive	Yes
96	MK	Youth	Female	Rural	Student	Medium	Small	Positive	Positive	Yes
97	SK	Youth	Male	Urban	Student	Medium	Small	Positive	Negative	Yes
98	NK	Youth	Female	Urban	Homemaker	Medium	Medium	Negative	Positive	Yes
99	MA	Youth	Male	Urban	Labourer	Low	Medium	Positive	Positive	Yes
100	RK	Youth	Male	Urban	Student	Medium	Medium	Positive	Positive	Yes

INDEX:

PID = Patient ID

NAME = Patient_Name

ECO_STATUS = Economic _Status

SYM1_LNSIZE = Symptom1_Lymph_Node_Size

SYM2_PHEALTH = Symptom2_Poor_Health

MANTOUX_TEST = Mantoux_Test

DIS_STATUS = Lymph_Node_Tuberculosis_Disease_Status

CHAPTER 6

MODEL EVALUATION

Model Evaluation is defined as the assessment or the judgment of a classifier's performance. It involves measures for assessing or judging how accurate the classifier is in its predictive abilities. The evaluation is carried out on a *test set* of class-labeled tuples that were not used to train the model [3].

Hence, I implemented our Adaptive Classifier on the *test set* of our sample medical dataset, and evaluated its performance on the basis of the terminology given below:

P: The number of positive tuples

N: The number of negative tuples

True Positives (*TP*): The number of positive tuples that were correctly labeled by the classifier.

True Negatives (*TN*): The number of negative tuples that were correctly labeled by the classifier.

False Positives (*FP*): The number of negative tuples that were incorrectly labeled as positive.

False Negatives (*FN*): The number of positive tuples that were mislabeled as negative.

Confusion Matrix: The Confusion Matrix is a useful tool for analyzing how well the classifier can recognize tuples of different classes. *TP* and *TN* tell us the number of correct decisions taken by the classifier, while *FP* and *FN* tell us the number of incorrect decisions taken by the classifier.

TABLE 6.1: Confusion Matrix

Class	**Predicted Class**				
			yes	*no*	Total
Actual Class	*yes*		*TP*	*FN*	*P*
	no		*FP*	*TN*	*N*
	Total		*P'*	*N'*	*P + N*

classification accuracy: The *accuracy* of a classifier on a given *test set* is the percentage of *test set* tuples that are correctly classified by the classifier.

$$accuracy = \frac{TP + TN}{P + N} \qquad Eq. (6.1)$$

classification precision: The *precision* of a classifier on a given *test set* is a measure of exactness, i.e. what percentage of tuples labeled as positive are actually such.

$$precision = \frac{TP}{TP + FP} \qquad Eq. (6.2)$$

CHAPTER 7

RESULT AND DISCUSSION

The experimental results are shown below which help in further analysis of the sample medical datasets.

```
(-(age_v1c1 / age_v1) * Math.Log((age_v1c1 + var1) / (age_v1 + var2), 2)

(-(age_v2c1 / age_v2) * Math.Log((age_v2c1 + var1) / (age_v2 + var2), 2)

(-(age_v3c1 / age_v3) * Math.Log((age_v3c1 + var1) / (age_v3 + var2), 2)
```

Figure 7.1: Decision-Tree Classifier (modified) showing implementation of Laplacian Correction in Decision Tree Classification algorithm using MS Visual Studio 2010

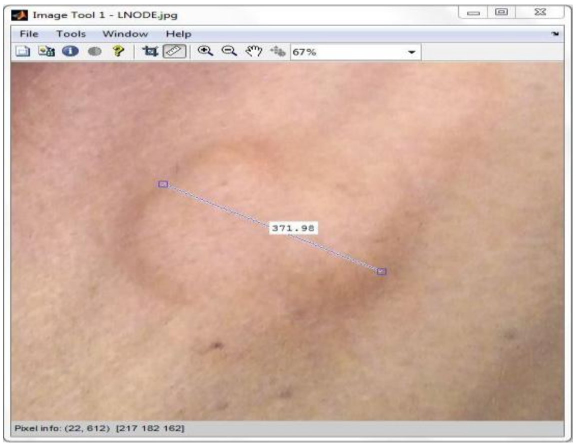

Figure 7.2: Measurement of diameter of lymph node of a patient by the calculation of Euclidean Distance (in pixels) using MATLAB R2013a

Figure 7.3: An ultrasound image of a patient suffering from the disease Liver Cirrhosis

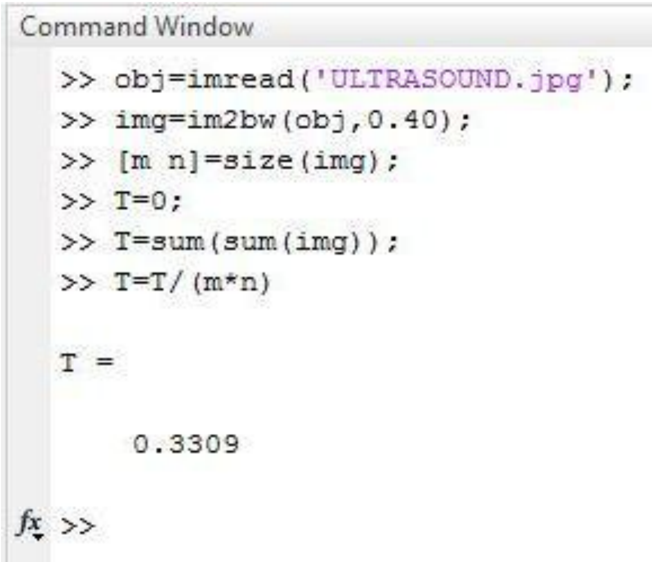

Figure 7.4: Texture Analysis of Liver by the implementation of Mean Gray Level Algorithm on the ultrasound image of a patient using MATLAB 2013a

First of all, I found out the number of correctly classified tuples by each classifier, including our Adaptive Classifier. This helped me to calculate the percentage of correctly classified tuples by each classifier. Next, I evaluated the *accuracy* and the *precision* of each classifier, as discussed earlier in *Eq. (6.1)* and *Eq. (6.2),* respectively. The *accuracy* and the *precision* of each classifier were represented in graphical format for better visualization.

TABLE 7.1: Comparison of results of different classifiers (Diabetes Mellitus Dataset)[a]

Classifier	Total Number of Tuples	Number of Correctly Classified Tuples	Percentage of Correctly Classified Tuples
RBC	50	45	90%
DTC (m)	50	43	86%
NBC (m)	50	44	88%
AC	50	46	92%

TABLE 7.2: Evaluation of *accuracy*, *precision* of different classifiers (Diabetes Mellitus Dataset)[b]

Classifier	*accuracy*	*precision*
RBC	0.900	0.737
DTC (m)	0.860	1.000
NBC (m)	0.880	0.750
AC	0.920	0.778

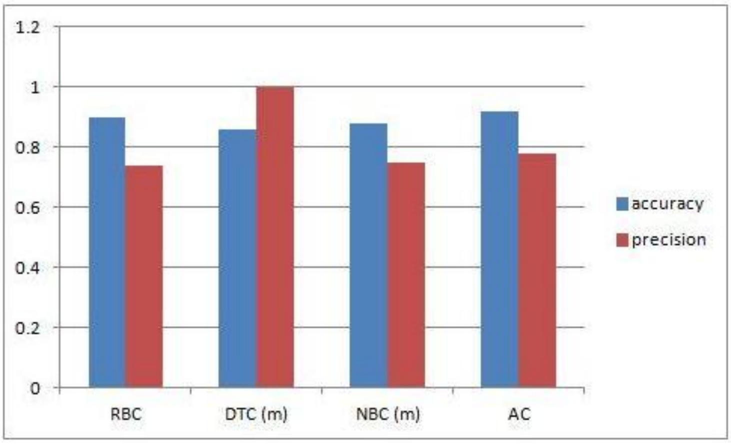

Figure 7.5: Analysis of *accuracy*, *precision* of different classifiers (Diabetes Mellitus Dataset)[c]

[a,b,c]

RBC = Rule-Based Classifier

DTC (m) = Decision-Tree Classifier (modified)

NBC (m) = Naïve-Bayesian Classifier (modified)

AC = Adaptive Classifier

TABLE 7.3: Comparison of results of different classifiers (Pulmonary Tuberculosis Dataset)[d]

Classifier	**Total Number of Tuples**	**Number of Correctly Classified Tuples**	**Percentage of Correctly Classified Tuples**
RBC	50	46	92%
DTC (m)	50	44	88%
NBC (m)	50	43	86%
AC	50	47	94%

TABLE 7.4: Evaluation of *accuracy*, *precision* of different classifiers (Pulmonary Tuberculosis Dataset)[e]

Classifier	***accuracy***	***precision***
RBC	0.920	1.000
DTC (m)	0.880	0.875
NBC (m)	0.860	0.647
AC	0.940	0.909

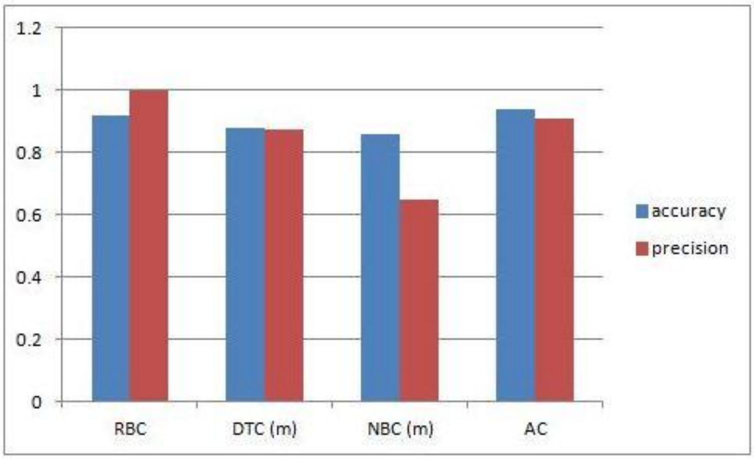

Figure 7.6: Analysis of *accuracy*, *precision* of different classifiers (Pulmonary Tuberculosis Dataset)[f]

[d,e,f]

RBC = Rule-Based Classifier

DTC (m) = Decision-Tree Classifier (modified)

NBC (m) = Naïve-Bayesian Classifier (modified)

AC = Adaptive Classifier

TABLE 7.5: Comparison of results of different classifiers (Liver Cirrhosis Dataset)[g]

Classifier	**Total Number of Tuples**	**Number of Correctly Classified Tuples**	**Percentage of Correctly Classified Tuples**
RBC	50	44	88%
KMC	50	29	58%
NBC (m)	50	44	88%
AC	50	45	90%

TABLE 7.6: Evaluation of *accuracy*, *precision* of different classifiers (Liver Cirrhosis Dataset)[h]

Classifier	***accuracy***	***precision***
RBC	0.880	1.000
KMC	0.580	1.000
NBC (m)	0.880	1.000
AC	0.900	1.000

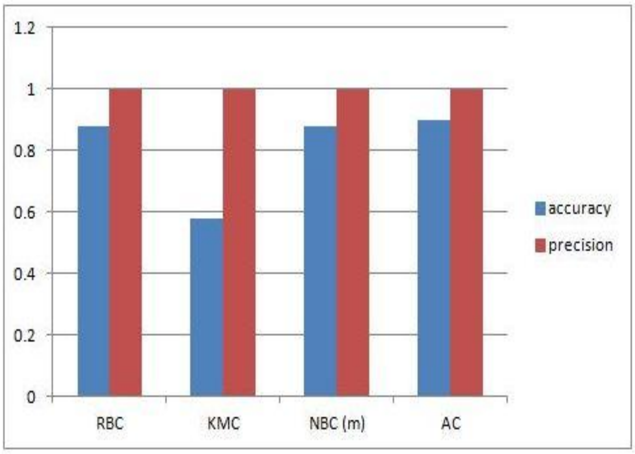

Figure 7.7: Analysis of *accuracy*, *precision* of different classifiers (Liver Cirrhosis Dataset)[i]

g,h,i

RBC = Rule-Based Classifier
KMC = K-means Classifier
NBC (m) = Naïve-Bayesian Classifier (modified)
AC = Adaptive Classifier

TABLE 7.7: Comparison of results of different classifiers (Lymph Node Tuberculosis Dataset)[j]

Classifier	Total Number of Tuples	Number of Correctly Classified Tuples	Percentage of Correctly Classified Tuples
RBC	50	45	90%
KMC	50	21	42%
NBC (m)	50	45	90%
AC	50	47	94%

TABLE 7.8: Evaluation of *accuracy*, *precision* of different classifiers (Lymph Node Tuberculosis Dataset)[k]

Classifier	*accuracy*	*precision*
RBC	0.900	1.000
KMC	0.420	1.000
NBC (m)	0.900	1.000
AC	0.940	1.000

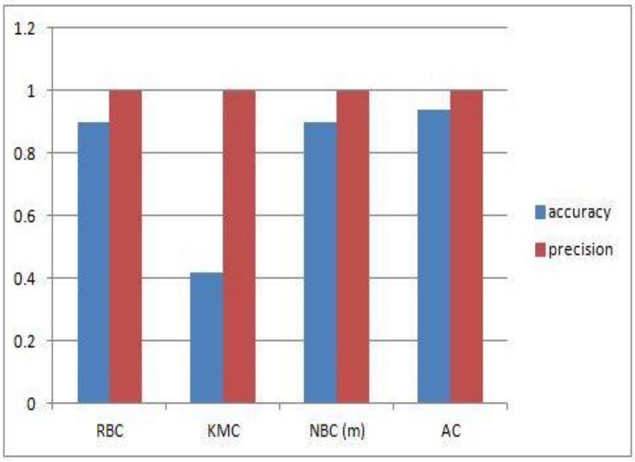

Figure 7.8: Analysis of *accuracy*, *precision* of different classifiers (Lymph Node Tuberculosis Dataset)[l]

[j,k,l]

RBC = Rule-Based Classifier
KMC = K-means Classifier
NBC (m) = Naïve-Bayesian Classifier (modified)
AC = Adaptive Classifier

While analyzing the Diabetes Mellitus Dataset, we found that the *precision* of Decision-Tree Classifier (modified) is 1.000 which is higher than that of our Adaptive Classifier (0.778). Similarly, while analyzing the Pulmonary Tuberculosis Dataset, we found that the *precision* of Rule-Based Classifier is 1.000 which is higher than that of our Adaptive Classifier (0.909). However, both these classifiers have lower classification accuracy than our Adaptive Classifier.

While analyzing our sample medical dataset, we were able to extract some other valuable trends of Diabetes Mellitus and Pulmonary Tuberculosis progression. Diabetes Mellitus was found to be more prevalent amongst the female members of the society. Diabetes Mellitus, generally affected the people of medium or high economic status, most of whom were homemakers or professionals. On the other hand, Pulmonary Tuberculosis was found to be more prevalent amongst the male members of the society. Pulmonary Tuberculosis, generally affected the lower socio-economic class, and those who lived in rural areas. The people who had a greater contact with the outside environment and those who suffered from poor health conditions, were more prone to infections, and thus showed higher chances of Pulmonary Tuberculosis infection.

While analyzing the Liver Cirrhosis Dataset, we found that Rule-Based Classifier and Naïve-Bayesian Classifier (modified) achieved very high classification accuracy (88% each). On the other hand, K-means Classifier did not perform too well and achieved average classification accuracy (58%). Similarly, while analyzing the Lymph Node Tuberculosis Dataset, we found that Rule-Based Classifier and Naïve-Bayesian Classifier (modified) were successful in achieving high classification accuracy (90% each). Here as well, the K-means Classifier did not perform too well and achieved even lower classification accuracy (42%). However, the results of our Adaptive Classifier, obtained by applying the *Bagging* technique on the individual classifiers, showed the highest classification accuracy in both the Datasets (90% for Liver Cirrhosis Dataset and 94% for Lymph Node Tuberculosis Dataset). Moreover, we find that the *precision* of each of the classifiers is 1.000, in either of the Datasets, since we worked upon one-class sample medical dataset (patients who suffered from Liver Cirrhosis or Lymph Node Tuberculosis).

While analyzing our proposed approach, we found that the calculation of lymph node diameter, using the concept of Euclidean Distance, was highly successful in generating categories for the lymph node sizes, in the case of Lymph Node Tuberculosis. Similarly, the ultrasound image segmentation, done in the case of Liver Cirrhosis, helped us to generate categories on the basis of texture analysis. Higher Mean Gray Level values were obtained for the ultrasound images that showed severely damaged liver textures than those images where the liver textures were in a healthier condition.

While analyzing our sample medical dataset, we were able to extract some other valuable trends of Liver Cirrhosis and Lymph Node Tuberculosis progression. A majority of the Liver Cirrhosis patients were in the middle age group. The male section of the society was largely affected by Liver Cirrhosis. An obvious reason for this phenomenon is the higher addiction of male members towards alcoholism than their female counterparts. Most of the Liver Cirrhosis patients belonged to the lower socio-economic class, and resided in rural areas. On the other hand, Lymph Node Tuberculosis patients belonged mostly to the youth section of the society. The number of female patients was slightly higher than the male patients. Moreover, a greater number of patients belonged to the rural areas. Lymph Node Tuberculosis, generally affected the lower socio-economic class. Its victims were mainly labourers and vehicle drivers who are in greater contact with the outside environment.

From the experimental results, we can see that our Adaptive Classifier has the highest classification accuracy in classifying each of the Datasets. Hence, H_0 is rejected or H_a is accepted, after the testing of hypothesis. The Adaptive Classifier has rightly justified its name by handling all types of diseases, which may or may not always be characterized by the observable features of the diseases, apart from handling single and multi-class sample medical datasets. It emerged as a useful tool for prediction purposes. Since our Adaptive Classifier was able to achieve an acceptable standard of classification accuracy on the sample medical dataset, it can be recommended for classifying future tuples in the field of Medical Data Mining.

CHAPTER 8

CONCLUSION AND FUTURE WORK

A novel classifier was presented for the purpose of medical diagnosis, which combined the different techniques of classification, along with Laplacian Correction, for better results. The proposed approach when implemented and tested on the sample medical dataset achieved over 90% classification accuracy. Hence, our proposed approach showed promising results which has lit a spark for future investigation. However, our initial Adaptive Classifier was found to be inefficient in handling those diseases for which analysis cannot be done only through the features perceived externally, but required the use of certain medical tests in conjunction with the perceived features.

As such, an innovative image processing technique was presented which was successful in classifying diseases that require the use of certain medical tests in conjunction with the observable features of the diseases. Moreover, our advanced Adaptive Classifier was built using the techniques of clustering in conjunction with classification which helped us to classify one-class sample medical datasets. Further, we found that it is better to use an ensemble classifier than an individual classifier since an individual classifier may not give highly accurate results (in our case, K-means classifier). Once again, our advanced Adaptive Classifier was able to achieve over 90% classification accuracy on the sample medical dataset which has rightly generated the curiosity required for further analysis. However, our advanced Adaptive Classifier has not yet been able to achieve 100% classification accuracy. Some of the obvious reasons for not achieving the 100% mark include multiplicity of lymph nodes at a single location, where the diameter of the largest node is only taken into consideration, as well as, the quality of ultrasound images taken for our research work.

Hence, our future work involves achieving even higher classification accuracy for the sample medical datasets, as well as, working upon other diseases in the field of Medical Data Mining.

CHAPTER 9

REFERENCES

[1] C. M. Velu and K. R. Kashwan, "Visual Data Mining Techniques for Classification of Diabetic Patients", 2012 3rd IEEE International Advance Computing Conference (IACC), ISBN: 978-1-4673-4529-3/12/$31.00, Pg 1070-1075, IEEE 2012.

[2] G. Kesavaraj and Dr. S. Sukumaran, "A Study on Classification Techniques in Data Mining", 4th ICCCNT - 2013, July 4 - 6, 2013, Tiruchengode, India.

[3] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining - Concepts and Techniques*, ISBN: 978-93-80931-91-3, Third Edition, Morgan Kaufmann Publishers, New Delhi, 2012.

[4] M. Akhil Jabbar, Priti Chandra, B. L. Deekshatulu, "Prediction of Risk Score for Heart Disease using Associative Classification and Hybrid Feature Subset Selection", 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), ISBN: 978-1-4673-5119-5/12/$31.00, Pg 628 - 634, IEEE, 2012.

[5] M. Akhil Jabbar, Priti Chandra, B. L. Deekshatulu, "Heart Disease Prediction using Lazy Associative Classification", ISBN: 978-1-4673-5090-7/13/$31.00, Pg 40 - 46, IEEE, 2013.

[6] Nazanin Shahrokhi, Roxana Dehzad, Soheila Sahami, "Targeting Customers with Data Mining Techniques: Classification", 2011 International Conference on User Science and Engineering (i-USEr), ISBN: 978-1-4577-1655-3/11/$26.00, Pg 212 - 215, IEEE, 2011.

[7] S. Udhaya Kumar, H. Hannah Inbarani, S. Senthil Kumar, “Bijective Soft Set based Classification of Medical Data”, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22, ISBN: 978-1-4673-5845-3/13/$31.00, Pg 517 - 521, IEEE, 2013.

[8] Satej Wagle, J. Alamelu Mangai, V. Santosh Kumar, “An Improved Medical Image Classification Model using Data Mining Techniques”, 2013 IEEE GCC Conference and Exhibition, November 17-20, Doha, Qatar, ISBN: 978-1-4799-0724-3/13/$31.00, Pg 114-118, IEEE, 2013.

[9] Sina Bahramirad, Aida Mustapha, Maryam Eshraghi, “Classification of Liver Disease Diagnosis: A Comparative Study”, ISBN: 978-1-4673-5256-7/13/$31.00, Pg 42 - 46, IEEE, 2013.

[10] Sneha Chandra and Maneet Kaur, “Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining”, INDIACom-2015, ISSN 0973-7529, ISBN: 978-93-80544-14-4, Pg 1.188-1.193

[11] T. John Peter and K. Somasundaram, “An Empirical Study on Prediction of Heart Disease using Classification Data Mining Techniques”, IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012) March 30, 31, 2012, ISBN: 978-81-909042-2-3, Pg 514 - 518, IEEE, 2012.

CHAPTER 10

LIST OF OWN PUBLICATIONS

[1] Sneha Chandra and Maneet Kaur, "Creation of an Adaptive Classifier to Enhance the Classification Accuracy of Existing Classification Algorithms in the Field of Medical Data Mining", INDIACom-2015, ISSN 0973-7529, ISBN: 978-93-80544-14-4, Pg 1.188-1.193

[2] Website: http://icacci-conference.org/web/

IEEE Conference ID: 35697

IEEE Conference Name: 2015 4th International Conference on "Advances in Computing, Communications and Informatics" (ICACCI 2015)

Date of Conference: 10th August-13th August, 2015

Venue of Conference:

SCMS Group of Educational Institutions
SCMS Campus, Prathap Nagar, Muttom, Aluva
Kochi-683106, Kerala, India

Full Paper Submission Deadline: 23rd April, 2015
Acceptance Notification: 25th May, 2015
Final Paper Submission Deadline: 25th June, 2015

Paper ID: 1570128991
Author 1: Sneha Chandra (Member ID: 1259343)
Author 2: Maneet Kaur Rainu (Member ID: 1259349)