Effectiveness of Data Mining in Predicting Heart Diseases

Shahria Afrin
13201066

Ashique Sikder
13201024

Supervisor: Dr. Amitabha Chakrabarty

Department of Computer Science and Engineering

BRAC University

Submitted on: 21st August, 2017

DECLARATION

This is a submission to the Department of Computer Science and Engineering for the purpose of obtaining a Bachelor's degree Bachelor of Science in Computer Science Engineering. We, hereby declare that the results of this research are solely dependent on our research. Credits for any resources used are provided in the research section. This paper was not used for any other academic or non-academic purpose and was not submitted for any other degree.

Signature of Supervisor:

_____

Dr. Amitabha Chakrabarty

Assistant Professor

Department of Computer Science and Engineering

BRAC University

Signature of Authors:

1. _____

Shahria Afrin

2. _____

Ashique Sikder

ABSTRACT

Heart Diseases affect a large population in today's world, where the lifestyle is moved from active to comfort-oriented. We live in era of fast foods. Which build up cholesterol, diabetes and many more factors which in turn affects the heart in some way or the other. According to the World Health Organization Cardiovascular Diseases (CVD) or Heart Diseases cause more death than any other diseases globally [1]. The amount of data in medical sectors is quite large and computerized as well. They are not utilized or put to any use. This data if studied and analyzed could be put to good use like prediction of diseases or even prevent them. Diseases such as cancer can be detected and the stage can also be predicted by training dataset with pictures of cancer cells. Similarly, heart disease can be predicted based on aspects like cholesterol, diabetes, heart rate etc. The prediction of heart diseases is a challenge and very risky. We observed that in some cases solutions of problems does not rely on a single method. It varies from situation to situation. It is also a challenge as most of the data are sparse or missing as they were not stored in the motive of analyzing. We therefore set out goal to finding which method would be best for predicting the diseases using data of four different hospitals from four different places. This is a comparative study on the efficiency of different data mining techniques such as DT, K-Nearest Neighbor and Support Vector Machine in predicting heart diseases. The Data Mining techniques are analyzed and the accuracy of prediction is noted for each method used. The result showed that heart diseases can be predicted with accuracy of above 90%.

Index Terms:

DT- Decision Tree, kNN- K-Nearest Neighbor, SVM- Support Vector Machine , Cardiovascular Diseases, Comparative Analysis, Cross validation error, dataset.

Acknowledgement

We are grateful to our supervisor Dr. Amitabha Chakrabarty, Assistant Professor of the School of Computer Science and Engineering of BRAC University for his immense support and guidance without which this thesis was not possible. We also thankful to Md. Saiful Islam, Lecturer of the School of Computer Science and Engineering of BRAC University for sharing his experience and providing us with ideas on our topic. We also express our gratitude to Dr. Sheikh Golam Raihan, Emergency Medical Officer of Ibn Sina Specialized Hospital who helped us clear our concept about cardiovascular diseases and encouraged us to work with it. And also to our parents and friends for their help and support. Lastly, we are thankful to all the faculty members of BRAC University who have inspired and motivated us throughout the entire undergraduate program.

Contents

List of Figures

List of Tables

Chapter 1

This chapter highlights the prediction of heart diseases and the different factors responsible for it. It briefly describes the different data mining techniques we used for comparison.

1.1 Introduction

Data mining is the extraction of new information by examining large amount of data which is previously unknown and important. It can be used to take certain decisions, estimate and predict using different algorithms. In a world like today's most of the data in medical sector is computerized but not utilized. It is stacked up in a database like old handwritten records and put to no use. This data can be harnessed to predict diseases such as Cancer, Cardiovascular Diseases and many more. Data mining techniques are used to predict different stages of cancer by using the different cancer cell photos for each stage. Similarly, heart disease can be predicted using different factors which includes family history, cholesterol, diabetes, exercise etc. According to the World Health Organization an estimation of 31% of global deaths are caused by Heart Diseases [1]. Over three quarters of which take place in low- and middle-income countries. Heart disease kills one in every 32 seconds in the United States of America [2]. The prediction of heart diseases is a very delicate factor and risky as well. If done properly it can be used by the medical administrative. This research is done to check the efficiency of certain data mining techniques on some chosen attributes as described later. It shows the effectiveness of DT, SVM and K- Nearest Neighbors. It was noticed that a technique does not always work for a given scenario. It differs due to the selected attributes of the data or even size. We observed some of the

techniques that works very efficiently with a scenario. The efficiency of each algorithm researched on the data is shown in the later sections.

The techniques used for comparison are Decision Tree, k- Nearest Neighbor and Support Vector Machine.

1.2 Motivation

The rate of heart diseases is increasing at an exponential rate. The busy lifestyle of people in this era with all the fast food in the lunch break and getting back to sitting and working has pushed as over the edge. Along with this people today have a lack of exercise and are less active. For most of them recreation is just another movie in bed or anything technology based. Physical activities have reduced drastically. These factors boosted the rate of heart diseases to an unfortunately high percentage. In a developing country like ours the rate of heart diseases has the same effect. The annual mortality rate per 100,000 people from cardiovascular diseases in Bangladesh has increased by 128.9% since 1990, an average of 5.6% a year [3].

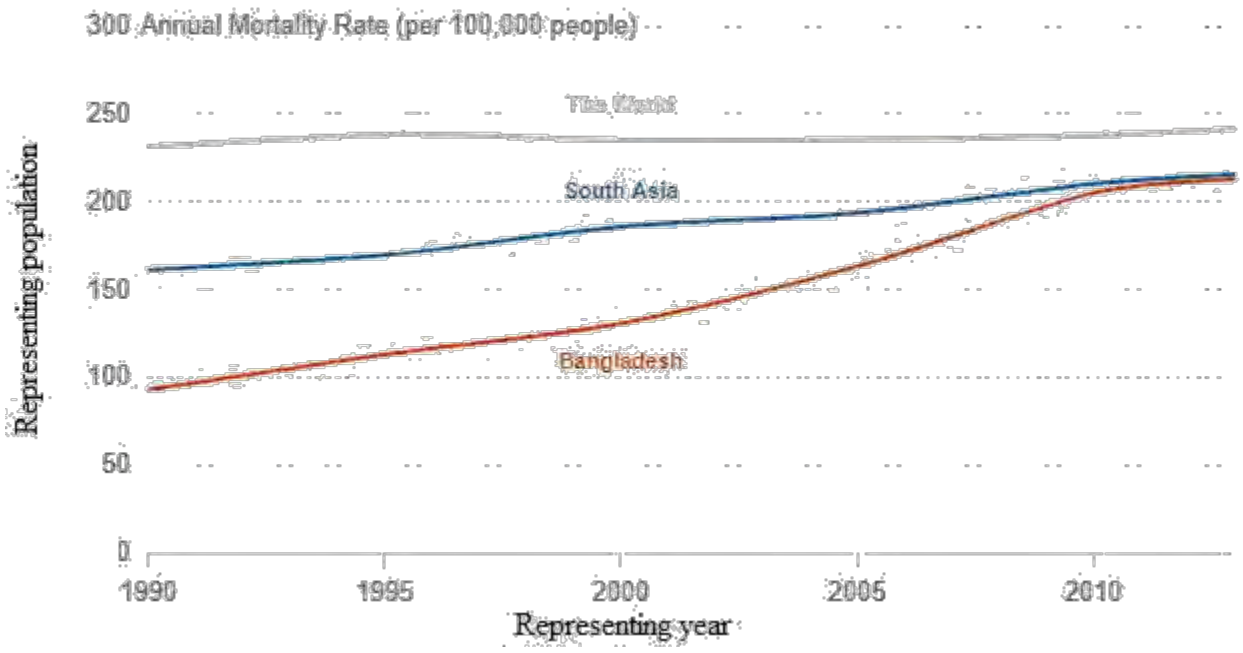


Figure 1.2: Annual Mortality Rate Graph of Bangladesh with the World. [3]

Prediction of heart diseases is a difficult and risky task. Since it is directly dependent on people's' health, accuracy is a major factor. If not predicted accurately it can be disastrous. This research therefore focuses on the comparison of different data mining techniques to predict it. It shows the comparative analysis of the different methods. Cross validation error is used to compare the techniques. We choose Decision Tree, k- Nearest Neighbor and Support Vector Machine as they are the most widely used techniques in determining diseases.

1.3 Thesis Contribution

The objective of this thesis was to prepare a comparative analysis on which data mining techniques have better accuracy in predicting Cardiovascular Diseases. The comparative analysis can be used by other researchers or by other software developers who are willing to work on developing software on heart diseases. This thesis gives an insight of the accuracy of Decision Tree, K- Nearest Neighbor and Support Vector Machine on the sparse data of Cardiovascular Disease.

1.4 Problem statement

The data in most of the hospitals are confidential in Bangladesh. We tried to collect data for research purposes but were denied. The data allowed to collect mostly contained name, age and sex which are not enough for prediction. Heart diseases depends on a lot of variables like hypertension, diabetes, exercise schedules, chest pain etc. Collection of these data was against the hospital rules. We therefore had to analyze the data from a dataset repository which had dataset from four different hospitals from four different states. This increases the diversity in the data. Although the data was finally collected, it has a lot of missing data. The data was filled by using median method. The data set was also sparse which made prediction difficult if the scale factor was set to get a refine and more accurate prediction.

1.5 Solutions

For the data, we chose to take it from the UCI Machine Learning Data Set as we could not collect t from the hospitals of Bangladesh.

To avoid over fitting the scale factors of Support vector Machine is reduced and the k for k- Nearest Neighbor is increased. The missing data problem was solved by median method.

1.6 Methodology

There are many algorithms that can be used to study data. The few we are using are as follows:

1.6.1 Decision Tree

DT are commonly used in operations research and operations management. In decision analysis, a DT and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. Another use of DTs is as a descriptive means for calculating conditional probabilities. The DT can be linearized into decision rules where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause.

Classification is an unsupervised learning used to predict the class of objects whose class label is

unknown. It is used for creating classification rules by means of decision trees from a given data set. Decision tree is used as a prognostic model. C4.5, C5.0, CART, ID3 are methods for building decision trees. It is an extension of the basic ID3 algorithm [4].

It is simply understandable and interpretable. Even the non-technical people are able to understand DT models after a brief explanation.

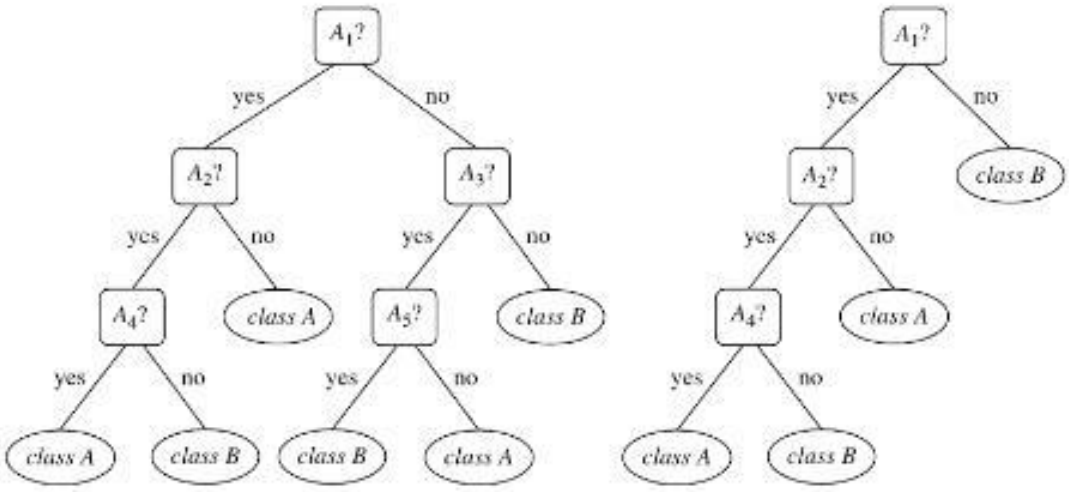


Figure 1.6.1: Basic Decision tree and its pruned version [5].

Moreover, these value even with little hard data. Furthermore, important insights can be generated based on experts describing a situation and their preferences for outcomes. It also allows the addition of new possible scenarios. Similarly, it helps to decide most exceedingly terrible, best and expected qualities for various situations.

1.6.2. K-Nearest Neighbor

KNN characterization is a standout amongst the most basic and straightforward arrangement strategies and ought to be one of the main decisions for an order study when there is almost no earlier learning about the dissemination of the information. KNN order was produced from the need to perform discriminated examination when dependable parametric assessments of likelihood densities are obscure or hard to decide. K-Nearest Neighbor is a also known as lazy learning classifier. [6]

Classification typically involves partitioning samples into training and testing categories.

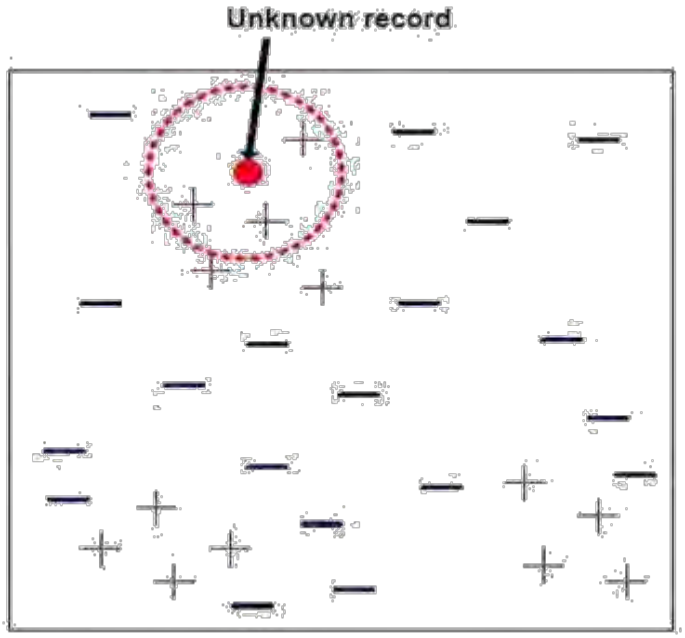


Figure 1.6.2(a): Unknown record in a Data set [6].

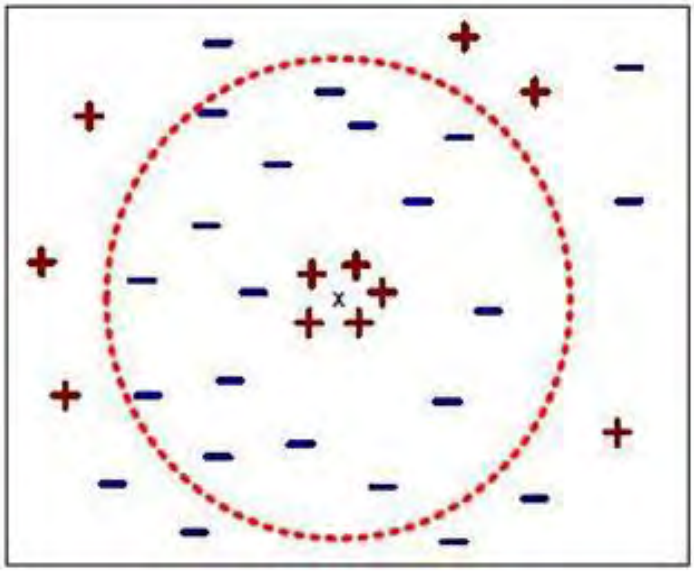


Figure 1.6.2(b): Choosing the value of K [6].

During the training process, we use only the true class ω of each training sample to train the classifier, while during testing we predict the class ω of each test sample. It warrants noting that KNN is a "supervised" classification method in that it uses the class labels of the training data. Unsupervised classification methods, or "clustering" methods, on the other hand, do not employ the class labels of the training data.

With 1-nearest neighbor rule, the predicted class of test sample x is set equal to the true class ω of its nearest neighbor, where mi is a nearest neighbor to x if the distance d(mi,x)=minj{d(mj,x)}.
For KNNs, the predicted class of test sample x is set equal to the most frequent true class among k nearest training samples. This forms the decision rule D: x→ω^n.

1.6.3 SVM

SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features we have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper plane that differentiate the two classes very well. SVM is very effective in high dimensional spaces. Also, it is effective in cases where number of dimensions is greater than the number of samples. Moreover, it uses a subset of training points in the decision function as a result it is also memory efficient. It has also some versatility like different Kernel functions can be

specified for the decision function. There are also some common kernels provided, but it is also possible to specify custom kernels. [4]

1.7 Data

The dataset we used contains data from four different hospitals.

Our algorithms were run only 14 attributes from the database of 76 raw attributes [7].

The data is taken from the following locations:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D

The attributes used from the data set to train our algorithm are Age, Sex, Chest Pain, Hypertension, Cholesterol, Diabetes, Resting Electrocardiographic Results, Maximum Heart Rate, Exercise Induced Angina, ST Depression, Slope of the peak exercise ST Segment, Number of Major Vessels Colored by Fluoroscopy and Thal.

Chapter 2

RELATED WORK

Data mining is widely used for prediction in many fields, medical being one of them. The more the documented data the more it can be manipulated in predictions. Many diseases are being predicted by data mining techniques with high efficiency. A study on prediction of breast cancer shows that. It can be predicted with high accuracy using different techniques [8]. Imagers of cancer cells at different stages help to show the patients conditions and even see if any particular drug works on the person. Heart disease prediction have also been studied in many cases.

Jyoti Soni in her paper showed how the performance of decision tree and Bayesian classification can be improved after using genetic algorithm. The paper first shows the analysis before using genetic algorithm and the analysis after [9]. The method used was classification via clustering. The classification was performed based on clustering.

Sellappan Palaniappan and Rafiah Awang built a model of Intelligent Heart Disease Prediction System (IHDS) with the use of data mining techniques namely, Neural Network, Naïve Bayes and Decision Tree [10]. Using medical profiles such as age, sex, blood pressure and blood sugar the system can predict the likelihood of patients getting a heart disease. IHDPS was capable of responding to "what if" queries that the usual decision support systems were not able to. It exploits the data using knowledge such as patterns, relationships amid medical factors connected with heart disease. The IHDS is a Web-based, user-friendly, scalable, reliable and expandable system.

In 2013, another paper was published which shows the diagnosis of lung cancer prediction using data mining classification techniques One Dependency Augmented Naïve Bayes classifier (ODANB) and naive creedal classifier 2 (NCC2) which are extensions of Bayesian classifier improved to work on data set which are small or incomplete [11]. Using generic lung cancer symptoms such as age, sex, Wheezing, Shortness of breath, Pain in shoulder, chest, arm, it can

predict the likelihood of patients getting a lung cancer disease. The paper mainly aims to make a model to provide early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient.

In 2014, M.A.Nishara Banu B.Gomathy Professor, Department of Computer Science and Engineering has published a research paper "Disease Forecasting System Using Data Mining Methods" [12]. In this paper data was clustered using clustering algorithms like K-means to cluster relevant data in database. Maximal Frequent Itemset Algorithm (MAFIA) is used for mining maximal frequent patterns in heart disease database. Their result showed that the designed system was capable of predicting heart attacks successfully.

The many systems used in predicting diseases inspired this research. Some of the motivated works are stated above. These prove that diseases can be predicted successfully using various data mining techniques.

Chapter 3

LITERATURE REVIEW AND WORKING PRINCIPLES

3.1 Literature Review

This sector contains the mechanisms used in existing works and also our reason to do this kind of research. Many researchers have contributed in this sector in using different techniques that helped to predict heart diseases in efficient ways. However, this field is still new to our type of research since more contributions can be made using different techniques, finding new patterns, improving accuracy etc. A survey says that heart disease is the most common reason for death in UK, USA, Canada and England [14]. Three different algorithms were proposed by Jyoti Soni et al those are Naïve Bayes, K-nearest neighbor and Decision tree [18]. These algorithms were used to predict heart disease. Among these, Naïve Bayes performed better compared to other algorithms. The tool they used for medical data classification was named Tanagra. Moreover, those data were calculated using 10 fold cross validation [18]. Sometimes there were cases when there were kinds of attributes which didn't perform to predict heart disease or to classify data. In those cases, Genetic Algorithm was used to reduce the definite data size to obtain the best possible subset of attributes [16]. Decision Tree, Naïve Bayes and Classification via clustering are the three classifiers used to analyze the occurrence of heart disease for the patients [16]. Shekar et al proposed new algorithm to mine association rules from medical data based on digit sequence and clustering for heart attack prediction [15]. The entire dataset is divided into partitions of equal size, each partition will be called cluster. Since it considers only a small cluster at a time and it is scalable and efficient [15]. That is why this approach reduces main memory requirements [15]. Niti Guru et al introduced the diagnosis of Heart Disease, Blood Pressure and diabetics with the help of neural networks [17]. They used valid patient's records to carry out a number of experiments. The algorithm neural

networks was trained and tested with 13 input variables such as age, diabetics, blood pressure, angiography report, cholesterol etc. Whenever an unfamiliar attribute was inputted in the system by the doctor, the algorithm itself founded the unknown data and created a catalogue of probable diseases the patient might have [17]. In year 2006, Kiyong Noh, et. al. [6] performed a work, *"Associative Classification Approach for Diagnosing Cardiovascular Disease*". In this research, FP-growth was the algorithm used which was an associative. The algorithm worked upon 670 valid patient data, which was divided into two groups, normal patients and patients with coronary artery disease. These two groups were employed to take out the experiment for the algorithm [18].

3.2 Decision Tree

A decision tree uses a graphs or models of decisions and the outcomes of a particular scenario. An algorithm can be shown in this graphical way. Decision tree is mainly used to get to a final objective as efficiently as possible. Moreover, it is also used as a powerful machine learning tool. DT is used as a prognostic model. C4.5, C5.0, C&RT, ID3 are methods for building DTs. It is an extension of the basic ID3 algorithm [16].

A decision tree contains a flowchart type structure in which each node has a corresponding test item. The implementation of Decision tree starts with the construction of it. A feature test is selected for the root node that helps to partition the training data in a way that causes maximum disambiguation of the class label associated with the data. This feature test will cause maximum reduction in class entropy while going through the training data. Then we drop to a set of child node from the root node, one for each partition of the training data created by the feature test at the root node. The number of child we will get depends on how symbolic the features are. When

our features are numeric we find decision thresholds that bipartitions the data and we drop to child nodes. Then, for each child node we pose the same question that we posed for the root node that would maximally disambiguate the class labels associated with the training data corresponding to that child node. A decision tree selects the best attribute that splits the data with the strategy, entropy calculation also known as "information gain" as a heuristic. Since decision tree algorithm will always try to create short trees, the result comes out from the information gain will try to be as distinguishable as possible and choose the highest information gain. If split point is well chosen then the decision would be close to our desired outcomes.

3.2.1 Information Gain

The challenging part in decision trees is the selection of the best test attribute. The information gain measure is use to select the test attribute at each node in the tree. There is a tern called entropy that is an important fact in this strategy. Entropy is a measurement of how pure a collection of arbitrary examples are. Let, S be a set which consists of data samples s. Class label attribute has m distinct values where m defines numerous categories, Ck. Si be the number of samples consisting in S in Ck class. Classification is done by the given equation:

I(S1,S2,….,Sm) = log2Pk ……

Pk is the probability that an arbitrary sample belongs to class Ck and is estimated by Sk / s. Let, A have v individual values, {a1,a2, …. av}. This attribute A is used to divide S into v subsets where

Sj contains samples in S that have value aj of A.

3.3 KNN

KNN is one of the most straightforward data mining techniques. It stores all cases and classify new cases based on similarity measure. KNN in mostly used as statistical estimation and pattern recognition. It is a non-parametric classification method which is broadly classified into two types a) Structure based NN techniques and b) Structure less NN techniques. In structure less NN techniques, data is classified into training and test sample. Distance is measured from training point to sample point. The point with lowest distance is called nearest neighbor. Structures of data like orthogonal structure tree (OST), ball tree, k-d tree etc. lie under structured based KNN techniques. KNN works efficiently when all the attributes are continuous [9].

An example is given below to describe how this algorithm works in general.

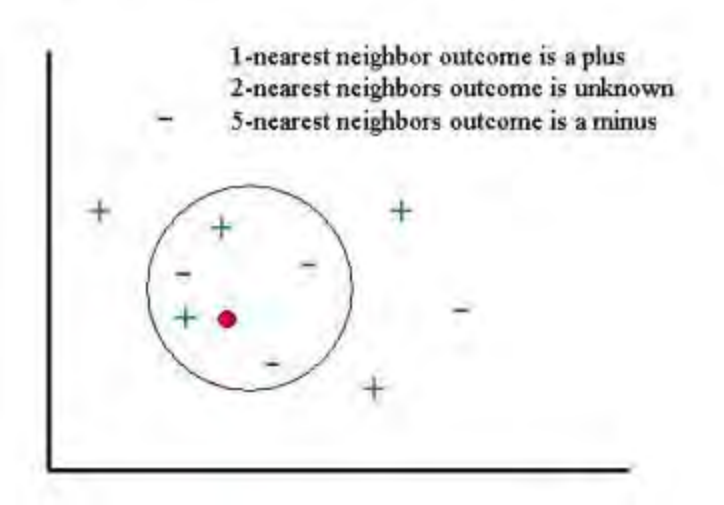


Figure 3.3(a): Basic KNN and its procedure [13]

The figure shows some plus and minus sign with a red circle that indicates a query point. We will try to figure out what will be the outcome of the red circle. Our goal is to find whether the red

point will be a plus or a minus. If we consider 1-Nearest Neighbor we can say that the output will be a plus. If we consider 2-Nearest Neighbor, we cannot figure out what will be the output now. If we consider 5-Nearest Neighbor, we can come up with the output minus since there are three minuses and two pluses.

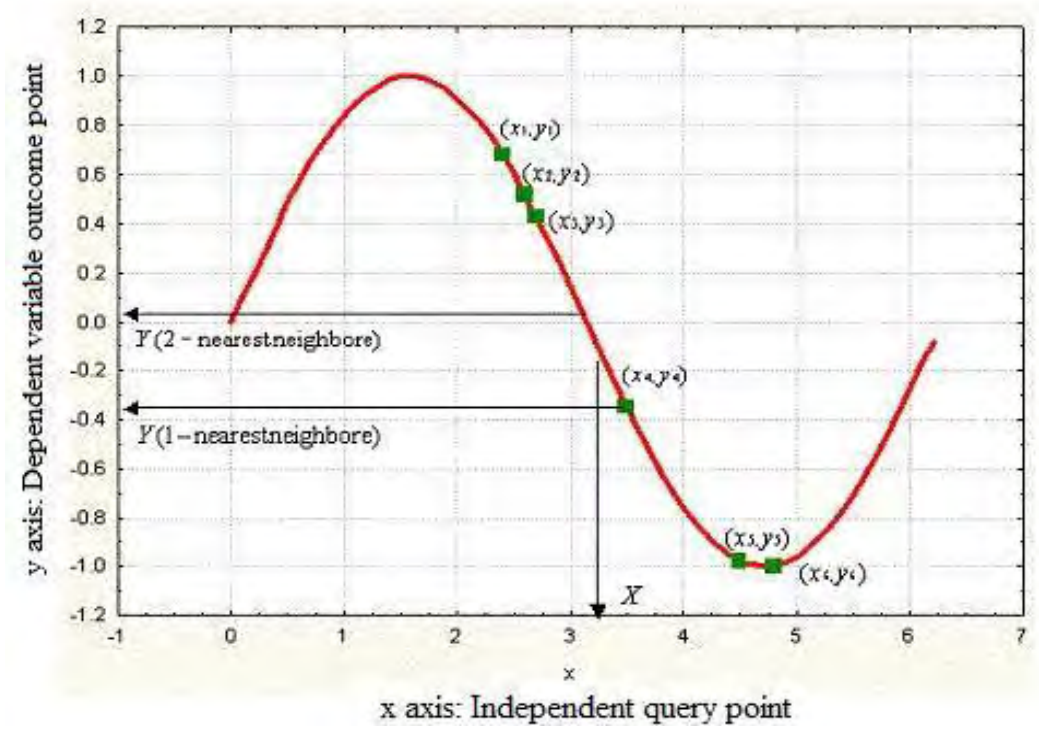


Figure 3.3(b): Nearest possible Neighbors (KNN) [13]

3.3.1 KNN Regression

In this section description of K-nearest neighbors with regression problem is given. This works by predicting a dependent variable with the help of some independent variables [13]. Firstly, we consider the schematic shown above, where a green squares are drawn from the relationship between the independent variable x and dependent variable y [13].

When we consider the 1- nearest neighbor, we search for the closest objects near the red circle and set the query point to X. For this case the point will be x4. So, for 1-nearest neighbor we write

Y = y4

For our own data, the dependent variable will be variable 14 names as

“num” y= num

When we consider 2-nearest neighbor method as an example, we locate the two closest points to X. Then points will be y3 and y4. The average of those two points comes; Y =y3+y4/2

For our data, the independent variable would be all the variables except variable

“num”. Then the equation forms like;

num = (age+sex+cp.....+thal)/13

To summarize, in a *k*-nearest neighbor method, the outcome *Y* of the query point *X* is taken to be the average of the outcomes of its *k*-nearest neighbors [13].

3.3.2 KNN Cross-Validation

Cross-validation is a technique that can be used to obtain estimates of model parameters that are unknown. Here we discuss the applicability of this technique to estimating *k* [13]. Here, we divide the data sample into v number of randomly taken folds. We apply KNN for the vth segment and evaluate the error. We generally use sum-of-squared. Moreover for classification, this method gives the best results in percentage of correctly classified cases [13]. The same thing we do for other segments of v. After measuring for all the segments, an average is done for checking the stability of the model. For various k, we check the error and we take the k that has the lowest error.

3.3.3 KNN Distance Matrix

To make prediction with KNN we define some equations. These equations helps us to find the distance between the query point and the example samples. We generally use Euclidean but others are also popular like Euclidean squared, City-block, and Chebyshev:

$D(x,p) = \sqrt{(x-p)_2}$ Euclidean

$D(x,p) = (x-p)_2$ Euclidean squared

$D(x,p) = Abs(x-p)$ Cityblock

$D(x,p) = Max(|x-p|)$ Chebyshew

Here, x is the query point.

p is the example point.

3.4 SVM (Support Vector Machine)

Support Vector Machine is a concept of defining decision boundaries. A decision planes separate different class items [13]. In the below figure the plane divides the example points to either red or green section. Any new example (data) that will fall right to the plane will be labeled as green otherwise it will be labeled as red [13].

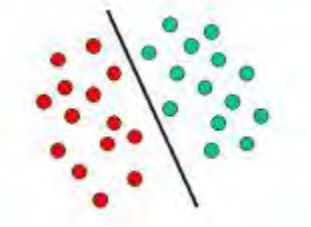


Figure 3.4(a): General SVM classification for linear hyper plane [13]

The above example was a linear type classification. But most of the classification examples are not simple and needs complex techniques to reach an optimal separation. For the below example, we can guess that there needs a curve plane to distinguish between objects [13]. The drawing of these planes for classification task is called hyper plane classifiers [13].

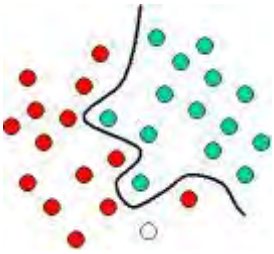
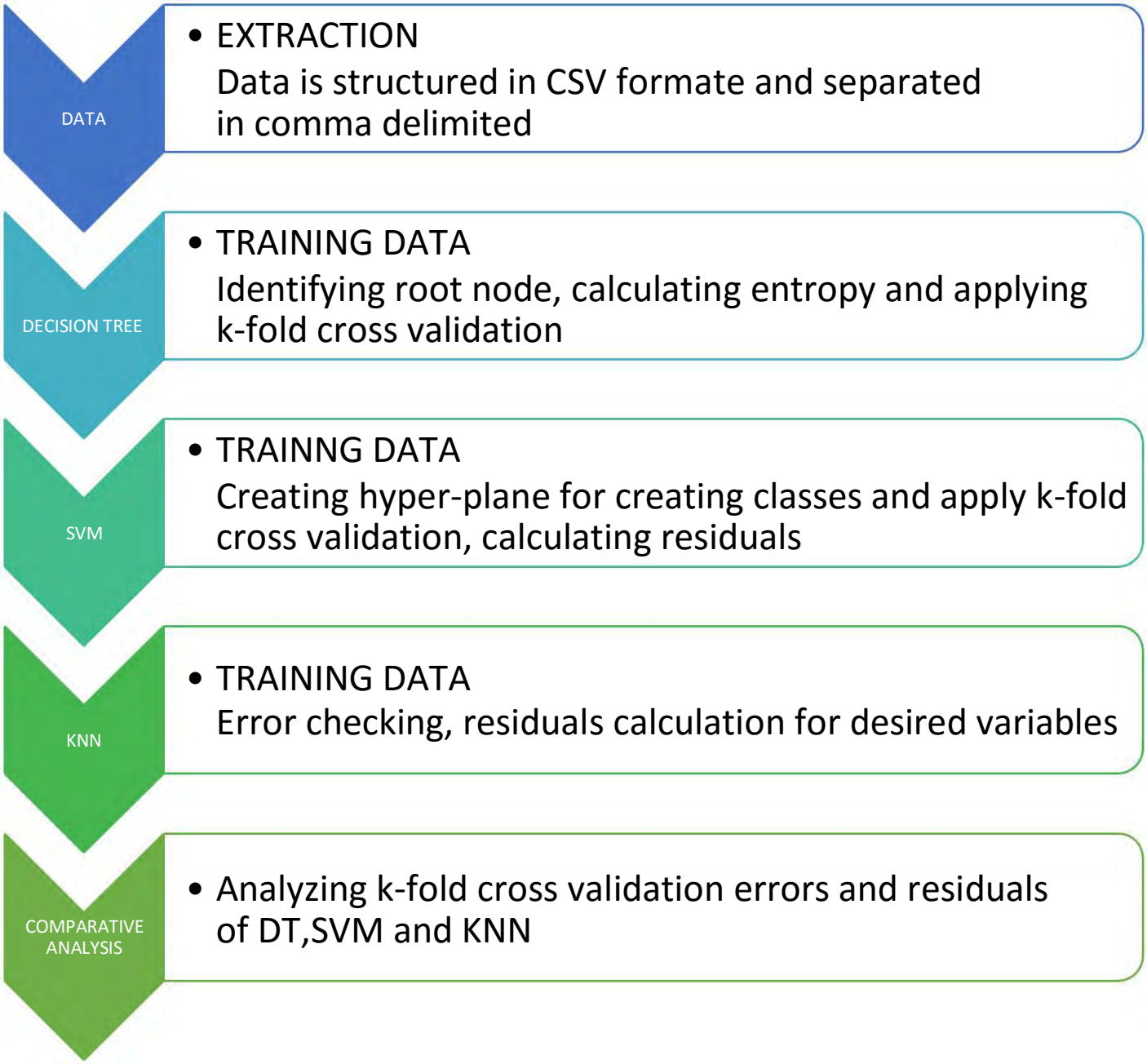


Figure 3.4(b): SVM classification for curve hyper plane

3.5: WORKFLOW


DATA
• EXTRACTION
Data is structured in CSV formate and separated in comma delimited
DECISION TREE
• TRAINING DATA
Identifying root node, calculating entropy and applying k-fold cross validation
SVM
• TRAINNG DATA
Creating hyper-plane for creating classes and apply k-fold cross validation, calculating residuals
KNN
• TRAINING DATA
Error checking, residuals calculation for desired variables
COMPARATIVE ANALYSIS
• Analyzing k-fold cross validation errors and residuals of DT,SVM and KNN

Chapter 4

RESULT ANALYSIS

Our data like most of the datasets contains noise. Since our data was collected from four different databases there were missing data in many cases. This noise and missing data leads to over fitting. Since, we used Radial Basis Function (RBF) kernel with a scale factor of 0.077 while implementing SVM our cross-validation error was reduced as compared to the others. We received an error of only 0.199 the lower the scale factor used the lower the chance of over fitting.

In case of KNN, over fitting was avoided by choosing 5 nearest neighbors instead of 1. Hence giving an error of 1.09.

Decision tree in our case does not give an efficient result compared to others due to the sparse data and the missing values. It gives an error of 21.2.

TABLE.1. Table showing the Decision Tree parameter values

Predicting Class	Diagnosis of Heart Disease
Features	Age, Sex, Chest Pain, Hypertension, Cholesterol, Diabetes, Resting Electrocardiographic Results, Maximum Heart Rate, Exercise Induced Angina, ST Depression, Slope of the peak exercise ST Segment ,Number of Major Vessels Colored by Fluoroscopy, Thal
Entropy Threshold	0.01
Depth of the tree	8

TABLE.2. Table showing the node considered as the root node and the entropy for the root node

Root Node	ca
Node Creation Entropy	1.915

TABLE.3: Table showing the KNN parameter values

Number of nearest neighbors	5
Distance measure	Euclidean
Input standardization	on
Averaging	uniform
Cross-validation error	1.09571

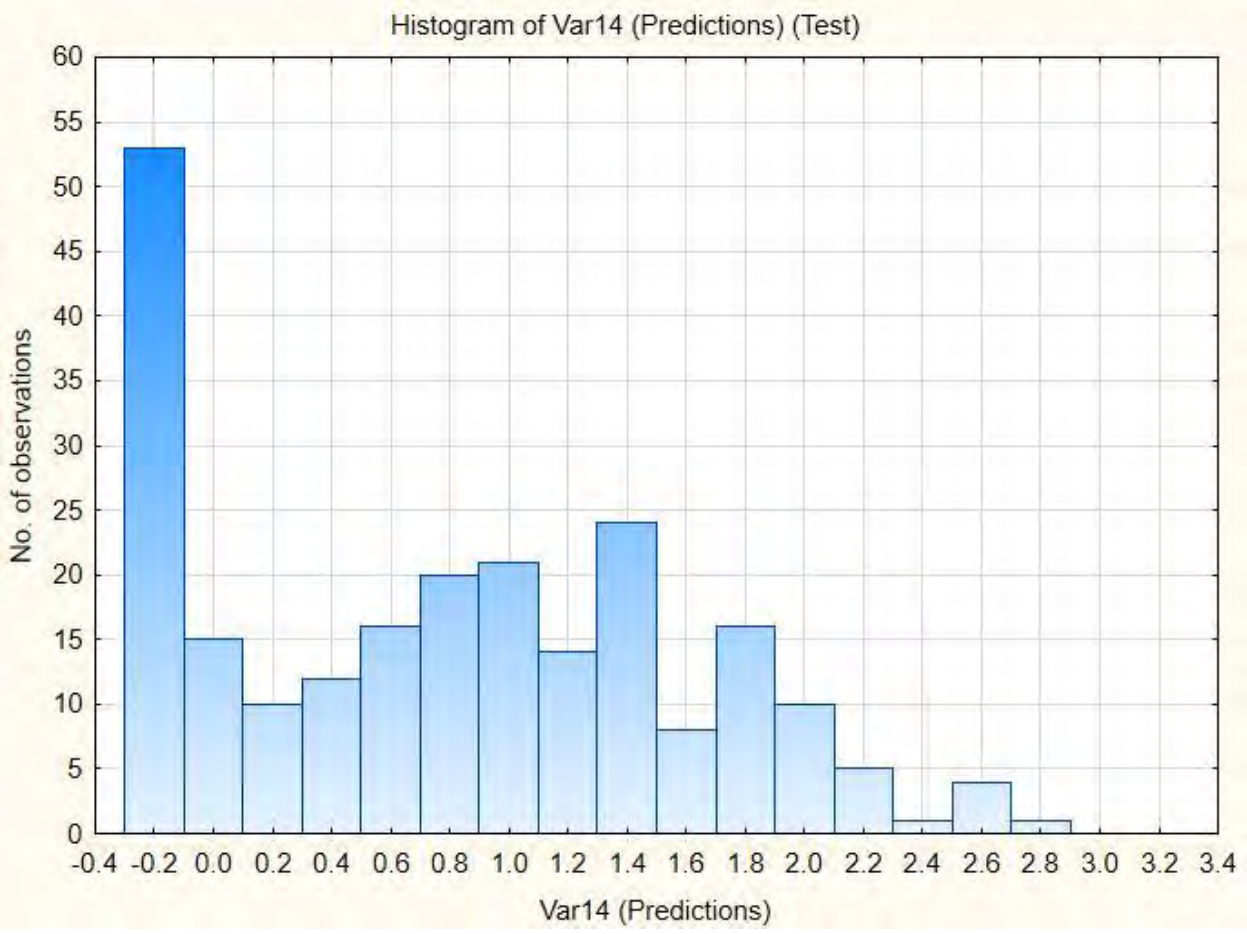


Figure 4(a): KNN histogram of num(Prediction test)

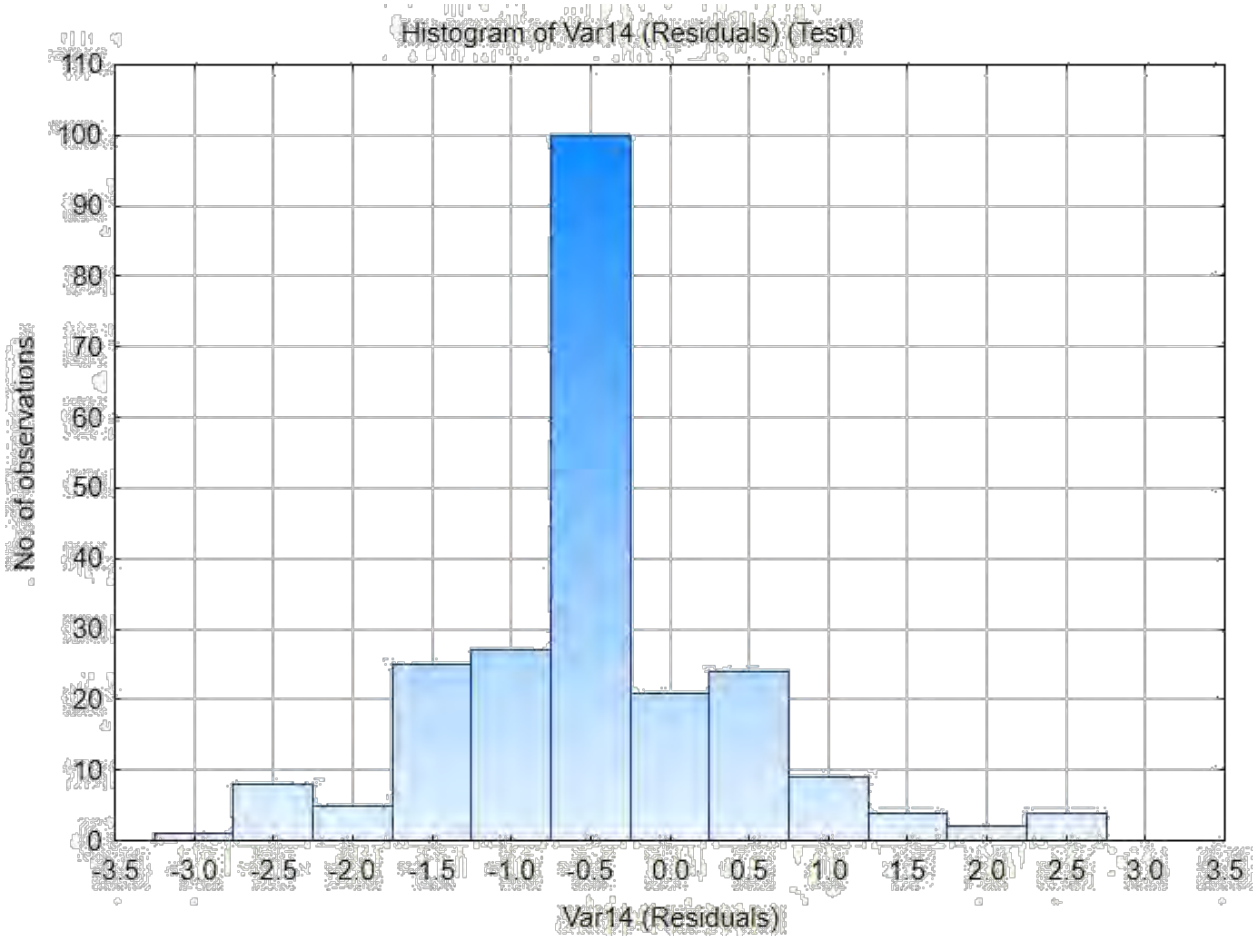


Figure 4(b): Histogram of num(Residual test)

TABLE.4. Table showing the SVM parameter values

SVM type	Regression type 1 (capacity=6.000,epsilon 0..100)
Kernel type	Radial Basis Function (gamma 0.077)
Number of support vectors	446 (352 bounded)
Cross-validation error	0.199
Mean error squared	0.661(Overall)
S.D. ratio	0.701(Overall)
Correlation coefficient	0.713(Overall)

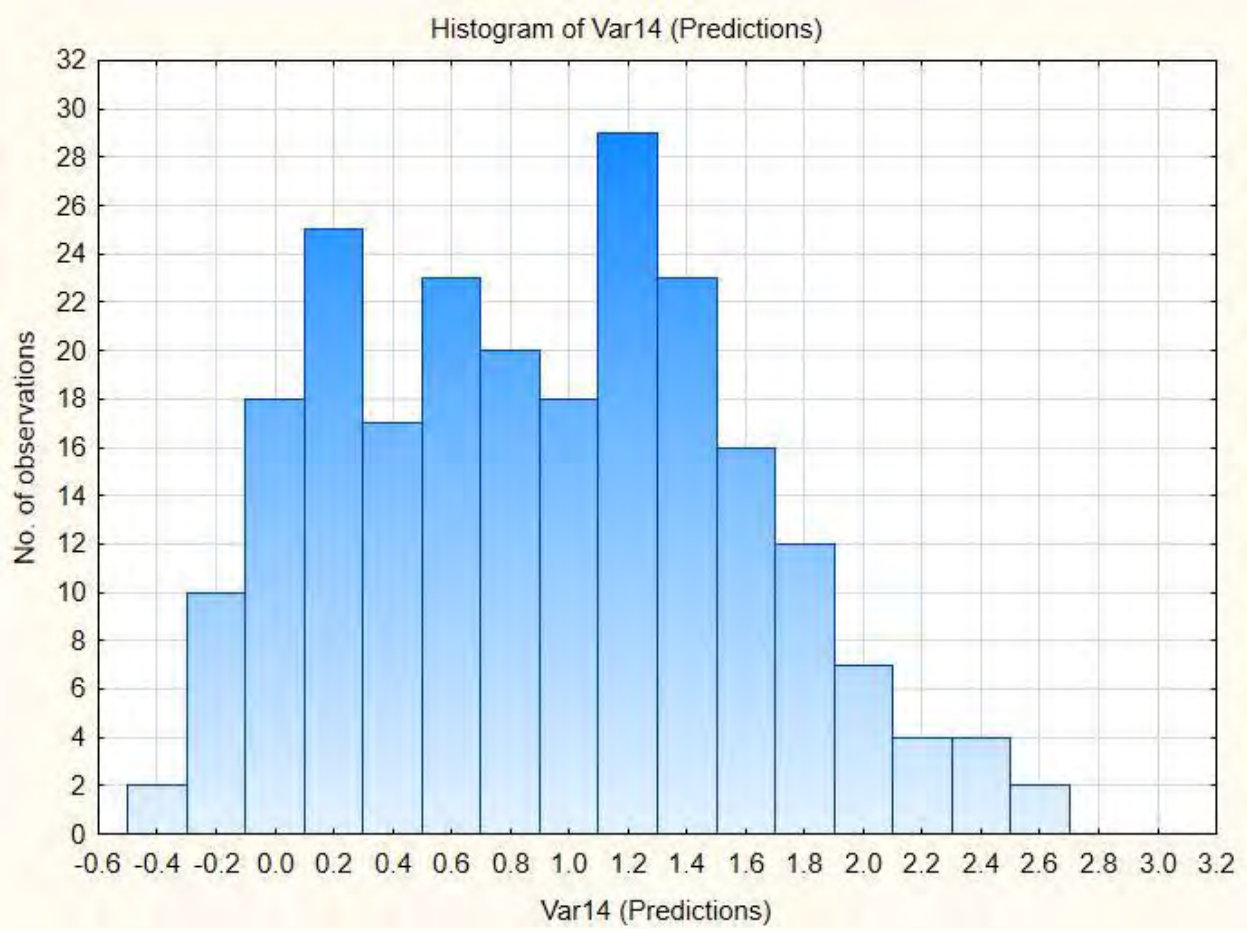


Figure 4(c): SVM histogram of num (Prediction)

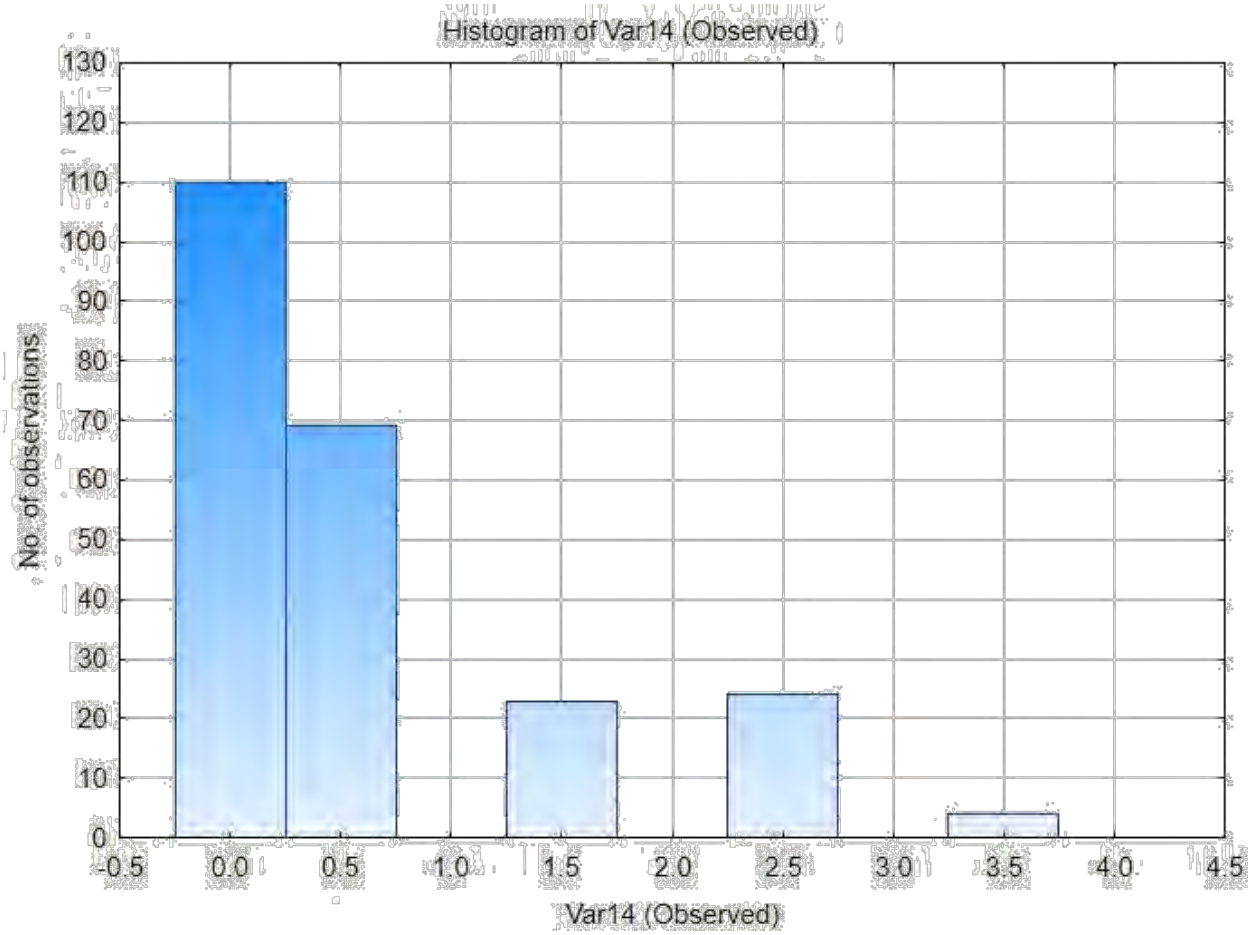


Figure 4(d): SVM Histogram of num (observed)

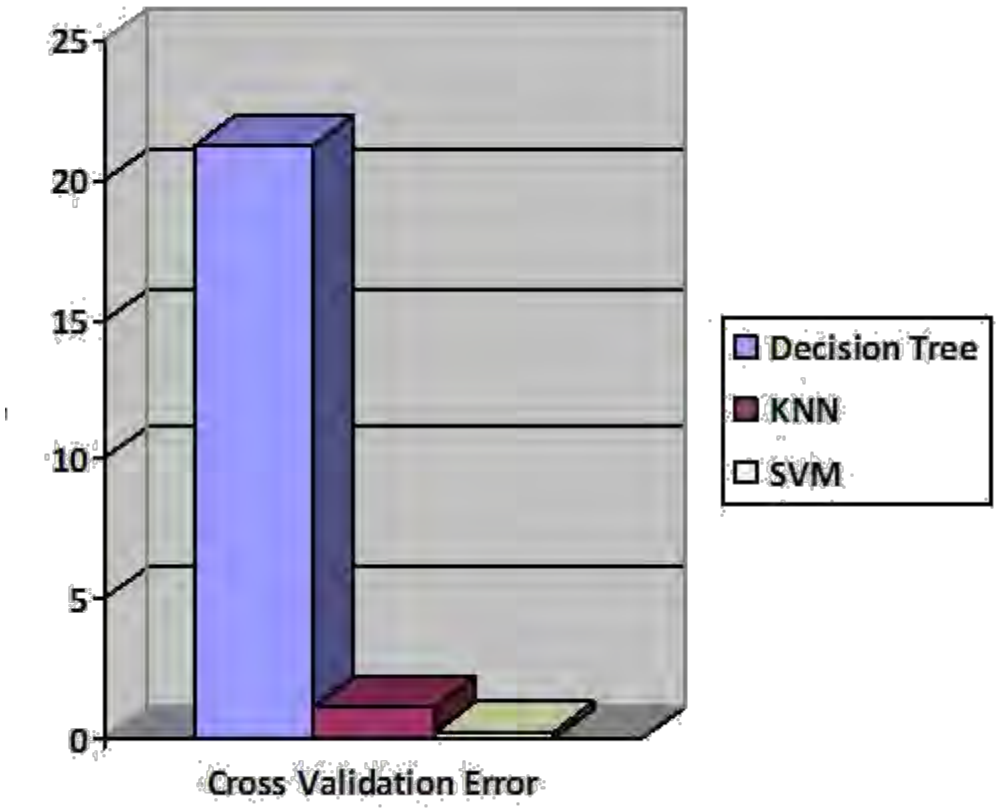


Figure 4(f). Histogram of Cross Validation Error of Decision Tree, KNN and SVM.

TABLE.5: Table of the cross validation errors

Index	Method	Cross Validation Error
1	Decision Tree	21.210
2	KNN	1.090
3	SVM	0.199

Chapter 5

CONCLUSION AND FUTURE WORK

Most of the data in today's world is computerized, these are usually scattered and not properly utilized. These data if analyzed for hidden patterns can be put to good use. Therefore, it has become a wide area for research with increasing importance and facilities. The main motive of this research was to create a basic data mining which can be used to predict heart diseases and also to find the efficiency of the data mining in this particular data set by the three chosen algorithm i.e. Decision tree, SVM and KNN.

Data mining can be used to build a software which help predict heart diseases. This software could be used by any non-medical employee of the hospital making it easy for the patients and saving time for the doctors.

REFERENCES

[1] WHO, "Cardiovascular diseases (CVDs)," Published by WHO, 2013. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 30-May-2017].

[2] V. Manikantan and S. Latha, "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods," Int. J. Adv. Comput. Theory Eng., vol. 2, no. 2, pp. 5–10, 2013.

[3] HealthGrove, "Cardiovascular Diseases in Bangladesh," by Graphiq. [Online]. Available: http://global-disease-burden.healthgrove.com/l/41188/Cardiovascular-Diseases-in-Bangladesh#Overview&s=3D2kvJ. [Accessed: 03-Aug-2017].

[4] G. Karthiga, C. Preethi, and R. D. H. Devi, "Heart Disease Analysis System Using Data," vol. 3, no. 3, pp. 3101–3105, 2014.

[5] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 3rd ed. Elsevier Inc., 2011.

[6] Peterson, Leif. "K-Nearest Neighbor". N.p., 2017. Print.

[7] W. David, "Heart Disease Data Set," Published by UCI, 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[8] S. Gupta, D. Kumar, and A. Sharma, "DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR," vol. 2, no. 2, pp. 188–195, 2011.

[9] J. Soni, "Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction," vol. 17, no. 8, pp. 43–48, 2011.

[10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115. doi: 10.1109/AICCSA.2008.4493524

[11] V. Krishnaiah, G. Narsimha, and N. S. Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," vol. 4, no. 1, pp. 39–45, 2013.

[12] M. A. N. Banu and B. Gomathy, "Disease Forecasting System Using Data Mining Methods," 2014 International Conference on Intelligent Computing Applications, Coimbatore, 2014, pp. 130-133. doi: 10.1109/ICICA.2014.36

[13]StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.

[14] Shadab Adam Pattekari and Alma Parveen,"Prediction system for heart disease using Naïve Bayes", International Journal of Advanced Computerand Mathematical Sciences, vol.3,pp 290-294,2012.

[15] K.Shekar, N.Deepika and D.Sujatha,"Association rule for classification of heart-attack patients", International Journal of Advanced Engineering
Sciences and Technologies, vol.11, no. 2, pp.253-257, 2011.

[16] M. Anbarasi, E. Anupriya and N.Iyengar, "Enhanced prediction of heart disease with feature subset selection using Genetic algorithm",
International Journal of Engineering Science and Technology vol.2, pp.5370- 5376, 2010.

[17] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. I (January - June 2007.

[18] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.