ALTINBAS UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ESTIMATION OF HEART DISEASE BASED ON DATA MINING USING PATIENTS HEALTH DATA BASE

M. Sc. THESIS

AZHAR HATEM JEBUR AL-BAIDHANI

ISTANBUL, 2017

ESTIMATION OF HEART DISEASE BASED ON DATA MINING USING PATIENTS HEALTH DATA BASE

by

AZHAR HATEM JEBURH AL-BAIDANI

ALTINBAS UNIVERSITY

Submitted to the Graduate Faculty of Science and Engineering

In partial fulfillment of the requirements for the degree of

Master of Electrical and Computer Engineering

December 2017

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Adil Deniz Duru

Co-Supervisor

Asst. Prof. Yasa Ekşioğlu Özok

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

(Asst. Prof. Yasa Ekşioğlu Özok) _____

(Asst. Prof. Adil Deniz Duru) _____

(Asst. Prof. Dilek Göksel Duru) _____

(Asst.Prof. Çağatay AYDIN) _____

(Asst.Prof. Emrullah Fatih YETKİN) _____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Asst. Prof. Çağatay AYDIN

Head of Department

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

AZHAR HATEM JEBUR

ACKNOWLEDGEMENTS

I would like to give my special thanks to my major professor, Dr.Yasa Ekşioğlu Özok for her effective guidance. Also, special thanks to Dr. Adil Deniz Duru for His constant support for whom this thesis has not been approved . which has played a big role in my support and encouragement .and thanked the jury members. In addition, I would like to thank my family Especially the spirit of my father (the grace of god),my dear mother ,my husband , my children ,my brothers,my sisters and my friends for their support , encourage them and stand with me.

ABSTRACT

ESTIMATION OF HEART DISEASE BASED ON DATA MINING USING PATIENTS HEALTH DATA BASE

AZHAR HATEM JEBUR AL-BAIDHANI

M.S, Electrical and Computer Engineering, Altınbaş University,

Supervisor: Asst. Prof. Yasa Ekşioğlu Özok

Co-Supervisor: Asst. Prof. Adil Deniz Duru

Date: December, 2017

Data mining (DM) is the process of finding or extracting knowledge from information on a huge piece data. DM uses intelligent methods to find patterns in the process of knowledge discovery (KD) in a database. The appearance field of DM promises to give a new technique and good tools. Also, DM can help the person to understand, solve big amounts of data remains on complex and unsolved problem. The wide functions in DM practice includes: classification, clustering, regression rule generation, sequence analysis and discovering association.

The classification is one of the most important techniques of DM. As well as, many problems in various fields such as science, business, industry and medicine can be solved by using these approaches. Neural Networks (NN) have appeared as a good tool for classification. The study of Heart Diseases (HD) database is testing by using NN approach.

HD diagnosis is not easy work which demands to a lot of experience and acquaintance. The common way for predicting HD is a doctor's checkup or different medical examination like ECG, Heart MRI Stress Test and etc. Nowadays, 'Artificial Neural Network' (ANN) has been

commonly used to the technique for dissolving many problem clinical diagnoses. An ANN is the ‘simulation of the human brain’, it is a supervised learning.

This research aims to optimize or reduce the number of biomedical test which asked from patients.Correspondingly to do a classification approach using NN technique and a Feature Subset Selection (FSS) algorithm. FSS is a pre-processing phase used to reduce number of attribute and remove irrelevant data. HD values are used and originally 13 attributes are involved to classify the HD. To reduce or optimize the number of attributes, different evaluators and search methods are determined.

In this research the two studies are conducted on the STALOG data set. The first study used three algorithms: Naïve Bayes (NB) got on accuracy equal to 85.18%. While J84 obtained on 91.48% and NN algorithm got on 99.62%. However, in the second study the NB obtained on 85.92% and J84 got on 91.48%. Finally, NN obtained on 99.25%. Moreover, accuracy of NN algorithm in the two studies got on the best result when compared with the results of the other algorithms.

TABLE OF CONTENTS

LİST OF TABLE

LIST OF FIGURES

LIST OF BBREVIATIONS

DM : Data Mining.

KD : Knowledge Discovery

NN : Neural Network

KDD : Knowledge Discovery Database

HD : Heart Diseases

ANN : Artificial Neural network

FSS : Feature Subset Selection

NB : Naive Bayes

DT : Decision Tree

DB : Database

MLPNN : Multilayer Percepton Neural Network

BP : Back Propagation

ML : Machine Learning

UCI : University of california, Irvine

1. INTRODUCTION

1.1 Overview

Recently, living circumstances have become difficult with increasing the requirements of life and the cost of living. This led to increasing psychological pressure on people, which can be reflected negatively on their health. Many studies have shown an increase in cases of heart diseases.

In addition, some people are beginning to show signs of disease due to difficult conditions. Therefore, they are going to offer health care to check if they are candidate for the disease. This explains it is important to show why the number of people going to hospital is increasing each day. Each patient spent much time and money to conduct all tests. So, the researchers tried to solve this problem by using specific techniques to find the optimal solution.

1.2 Heart Disease (HD)

The heart is the prime member of our body. Life dependents on efficient working of heart. If working of this part is not convenient brain, kidney or any other parts of body will be affected. HD is a disease that effects on the processing of the heart [1, 2].

There are many operators which led risk of HD that are listed below [3]:-

(1) Family story of (HD).

(2) Smoking.

(3) High blood pressure.

(4) Cholesterol.

(5) Lack of sport exercise.

(6) Obesity.

Large amount of data remains on not resolved and complex. For this purpose the DM can carry an estimate of which ways of behavior confirm efficient [4] by comparing with estimating causes sign and methods of treatments. Works on HD patients DB are a kind of practical life applications. The disclosure of a disease from many operator or signs is a different problem and

could drive to incorrect result. It is sensible to try employing the knowledge of some particularity composed in DB in respect of support the diagnosis process [5] The specialists in the health side recognize and foresee the HD in addition preferring efficient care for the sick with the help of DM techniques. Information related with the HD, common in the form of (Electronic Clinical Records), gene expressions, therapy information; this work was applied in all those works. Hence, more efficient and practical ways of 'cardiac diseases' and 'cyclic examination' are of high importance. In this research our aim is an experiment to introduce a classification concept using NN and a Feature Subset Selection (FSS) using information earning from HD patients. Also, the results are obtained then compare with other DM techniques.

1.3 Problem statement:

The people who may be suffering from the HD, thirteen tests are done on them by specialized to ensure if (he or she) is normal or not. However, these tests can be caused to have a negative effect on the patients who are doing a great effort for these tests and wasting their time and their money.

1.4 Literature survey:

Our aim in this research is to apply Neural Network (NN) and (FSS) for prediction of HD. Large numbers of studies has been implemented to find the active methods of health detection for different diseases. This study is an effort to foresee active diagnosis with reducing number of tests (i.e. attributes) that subscribe is more towards the HD using classification.

M.Ahil,B.l Deek Shatulu & Priti applied FSS with chi-square, principal component analysis [6] and the accuracy was 98.14% in 2013. Anbarasiet.al [7] used FSS and genetic algorithm to enhance the prediction of HD in 2010. HD prediction employing 'associative classification' was suggested by M.A jabbaret.al. In recent studies, matrix based association rule mining and genetic algorithm was used [8, 9]. Genetic algorithm (GA) & Association rule was applied based on HD

prediction in [10, 11]. Cluster based association to detect diseases NN was implemented in [11]. Different tools were used the STATLOG dataset in [12].

NB, DT and NN were used to offer intelligent HD prediction system in [13]. K-nearest Neighbor and Genetic algorithm were implemented in 2013 [14]. They combined top group concept in diagram with "association" to detect diseases. FSS using FCBF in kind two Diabetes the sick data was offered by Sarojinibala and C4.5 was used before and after FSS in 2015 [15]. Associative classification and genetic algorithm were submitted HD prediction system in [16]. NB and J48 were used on the STATLOG data set and then compared the result [17].

NN was applied and add two attributes smoking & obesity to get a more accurate result. Three DM techniques were applied using decision tree, NB and NN. From that product it has been seen that NN provides highest accurate result when compared to NB &DT [18].

FSS filter method was used with techniques like Naïve Bayes, SMO, J48 and ZeroR in they considered lab test as features. Choosing a subset of these attribute would mean choosing the more important lab test to implement [19] in 2015.

1.5 Aim of study

Our goal is to reduce the number of biomedical tests which is asked from patients. Originally, 13 attributes are involved in the classification of HD. In the concept of this thesis different evaluators and search methods are determined the attributes to reduce the volume of attributes be taken from the sick people.

1.6 Data Mining (DM)

DM is a critical step in extraction of knowledge from big DB. In the last few years, we have big data pool knowledge. DM has instituted its importance in any domain including health care. Medical database (DB) consists of a many of medical tests which are primary to identify a particular disease [20]. Medical DB is elements of the scope where the operation of DM has been

expanded into a certain side because of the sequence decline of “Medical data”. It is the best for validity care manufacture to earn the feature of DM by employing the identical as a smart diagnostic agent. It is sensible to obtain information and knowledge about diseases from the patients given ‘stored measurements’ in so far as medicinal data are attentive. Hence, DM has progressed into a necessary range in validity care [21]. It is probable to foresee the worth of “medical treatments” through construction DM tools.

DM is an essence section of Knowledge Discovery Database (KDD). Some interested people treat DM as a equivalent for KDD since its main part of KDD process. KD consists of a series of the following points:

(1) Data Cleaning: irrelevant data should be removing.

(2) Data Integration: joint diverse data sources.

(3) Data Selection: from DB, data relevant to the job is analysis.

(4) Data Transformation: transformed or union the data turns into forms proper for mining by doing summaries or gathering operations.

(5) Data Mining (DM): a core style where smart methods are performed in order to find data patterns.

(6) Pattern Evaluation: for distinguishing the good patterns the state knowledge depends on some excited measures.

(7) Knowledge Presentation: the ‘mined knowledge’ are used to submit to the user [22].

Fig 1 shows detailed stages of KDD, at the beginning shows the problem definition, then data gathering and preparation for treatment after that begins to build the model and does an evaluation then the knowledge deployment.

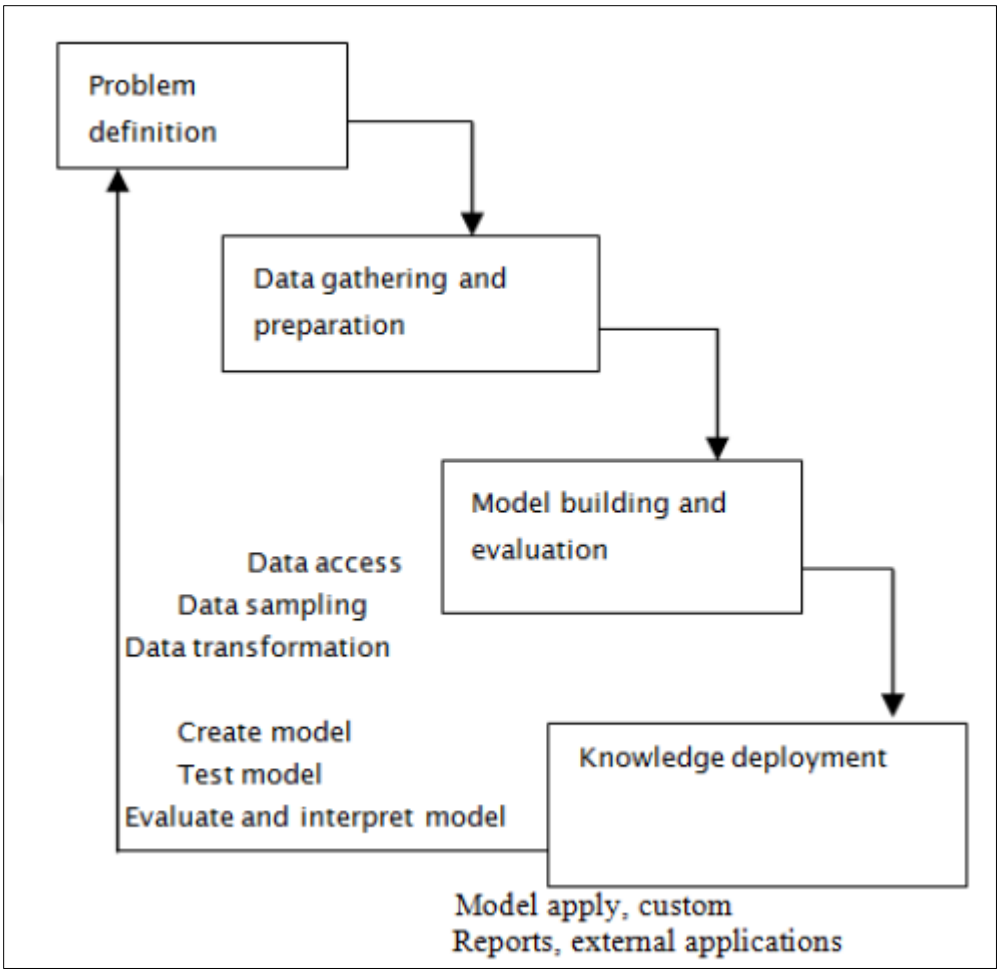

Figure. 1 steps of KDD [22]

1.7 Algorithms

1.7.1 "*Artificial Neural Networks" (ANN):-* NN is a "mathematical model" that depends on 'biological' NN. ANN is setup on supervision of a 'human brain'. It is very complex net of nodes. NN is an interconnected group of three simple layers the first one is input layers, hidden layer is second and output layer is the third layer. The column that are entered as input from the first layer.

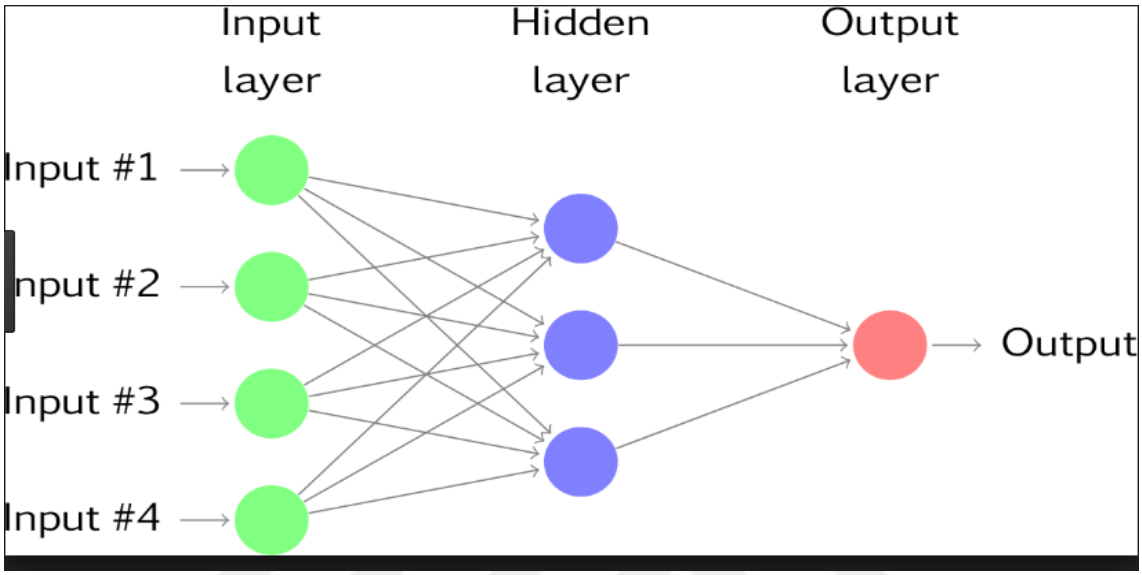

Figure. 2 Neural Networks (NN) [23]

For better achievements NN can be used to show in the Fig 2

The advantages of (NN) for classification are [23]:

(1) The weights give strong to the NN.

(2) Performance of NN blossom by learning.

(3) Degree of accuracy is high and a low error rate that lead the moment the convenient training performed.

(4) NN is hard in the noisy environment.

1.7.2. "*Feature Subset Selection" (FSS)*: FSS is a pre-processing operation usually applied with 'machine learning' (ML). Its idea is to remove irrelevant data in order to increase accuracy. It indicates to the problem which recognizes the columns (tests) that are more important in divine "class". Attributes may be nominal, continuous or discrete. Whatever the case features kind are of three:

1) Relevant.
2) Irrelevant.
3) Redundant.

FSS partitioned into two sections:

1) Attribute Evaluator.
2) Search Method.

Every type has different techniques. The first section each feature in the DB is evaluated in the situation of the output variable (e.g. the class). The second section' tries to move different groups of attributes in the DB to reach on a small list of subset features. Some Attributes Evaluators order the use of appointed second section. For example, the Ranker search method used with the Correlation Attribute Eval technique that evaluates each attribute and order the attributes in a rank order. A good feature subset is one that includes attributes closed correlated with 'class', so far uncorrelated with not prediction of each other.

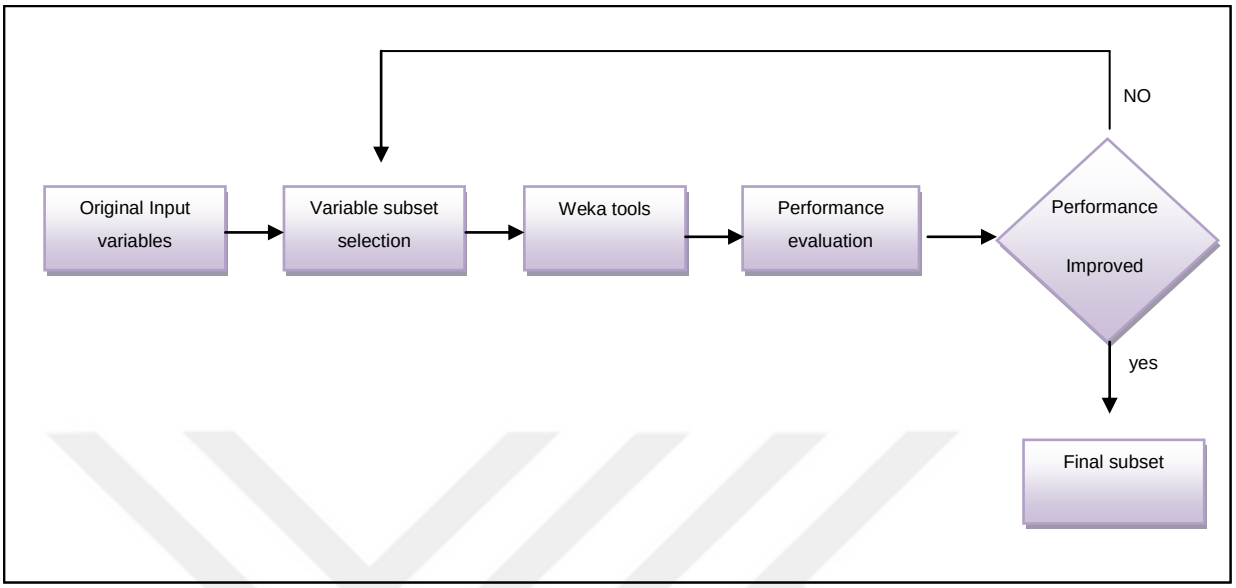

Figure. 3 Feature Subset Selections (FSS) [23]

Fig 3 shows the stage of FSS, all the original variables input after that chose the subset selection from these variables in pre-processing stages then applies Weka tools to check performance evaluation if it appropriate or not.

1.7.3. *Naive Bayes (NB):* The NB algorithm is a smooth classifier that compute a collection of probabilities by counting the combination and frequency values. NB classifier depends on Bayes theorem. Condition independence can be use in this classifier algorithm; in another words it supposed that features value on a given class does not build on the values of other features. NB is depending on supervised learning and it is a statistical classifier. Fig 4 shows the stages of Naïve Bayes and how can read data and then spilt training set data then use Naïve Bayes.

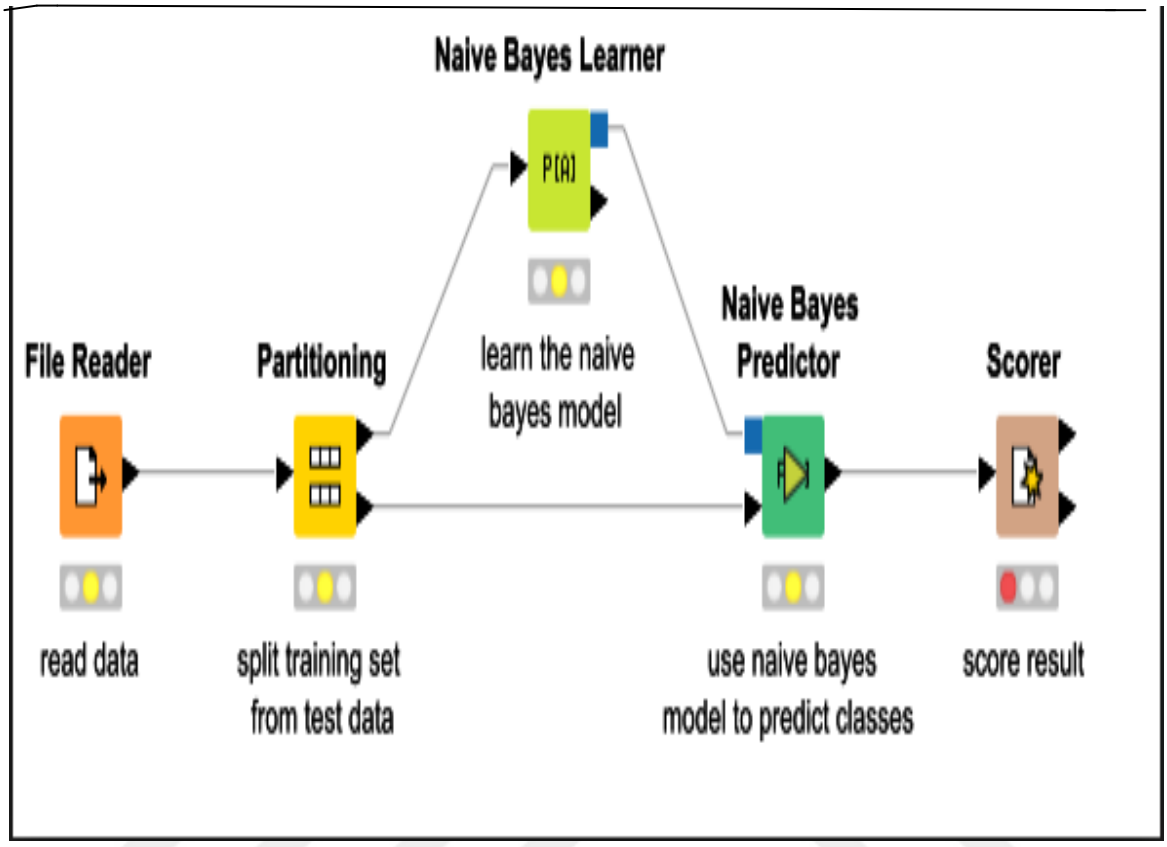

Figure. 4 Naïve Bayes (NB) [30]

1.7.4. *J48 classifier*: is a plain C4.5 Decision Tree (DT) for 'classification'. The DT process is the most useful in classification a problem. With this algorithm tree is a binary construct to form the classification operation. Fig 5 illustrates how tree is built.

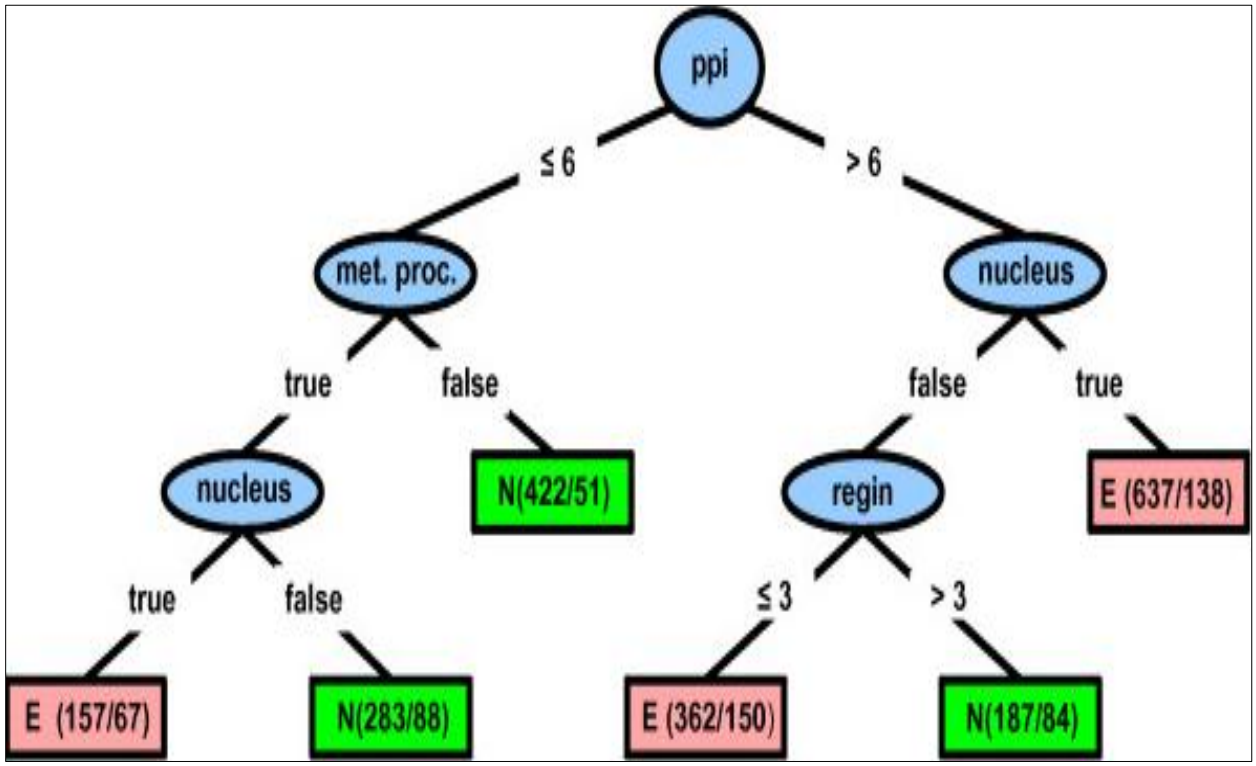

Figure. 5 J48 [31]

2. METHODS

2.1 Normalization of Attributes: Normalization is a process to convert input variables into the data range [0, 1] that the sigmoid activation functions lie in [24, 25].Better classifier will be promoted the training of the network.

$$N.V = \frac{X - Xmin}{Xmax - Xmin} \qquad (1)$$

Where

N.V: value after normalization.

X: value before normalization.

"X max: the maximum value of attributes".

"X min: the minimum value of attributes".

If Input variables are not normalize, we will get in accurate result.

2.2 Algorithms

2.2.1 Artificial Neural Networks (ANN)

ANN is a "arithmetical model" that depends on biological NN.ANN which is established on the supervision of a "philanthropic brain". The brain of human is very complex set of nodes. As shown in Fig 6.

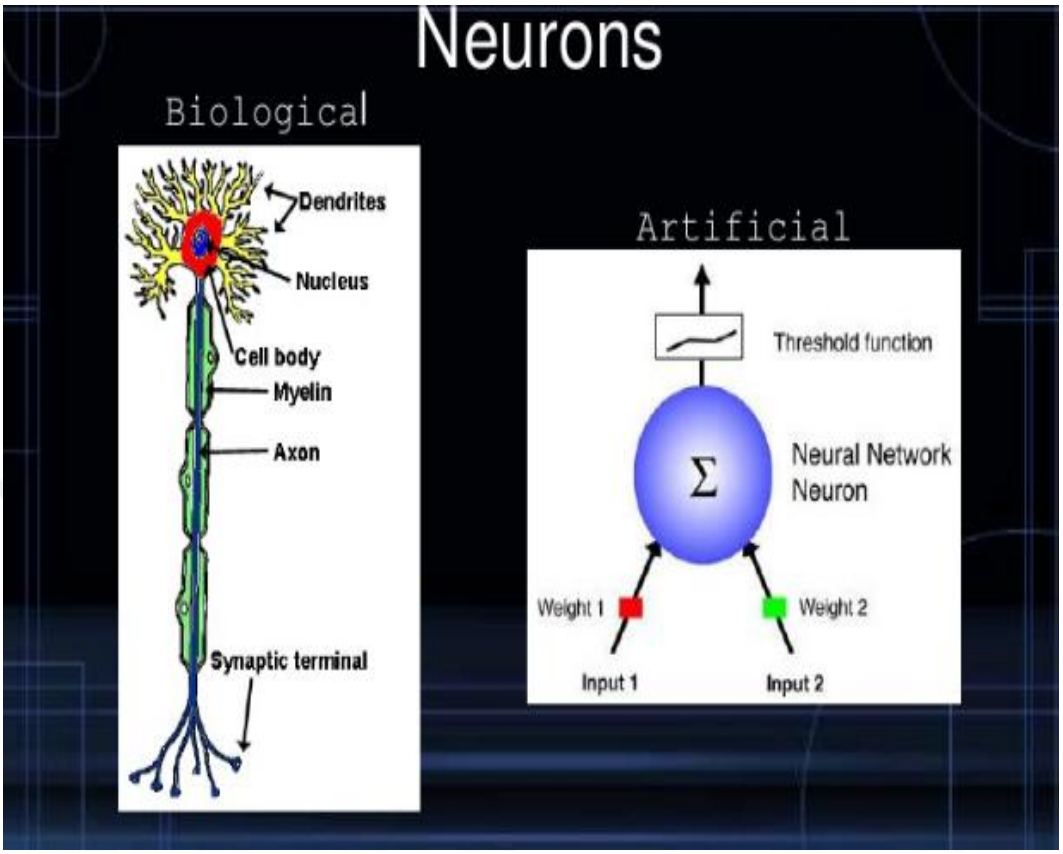

Figure. 6 biological and ANN [23]

NN is an interconnected set of three simple layers input layer, hidden layer and output layer. The attributes that are entered as input to form a first layer. In health diagnosis the sick risk operator are treated as input to the ANN, Fig 7 illustrates the structure and illustrates the three layers (input, hidden and output).

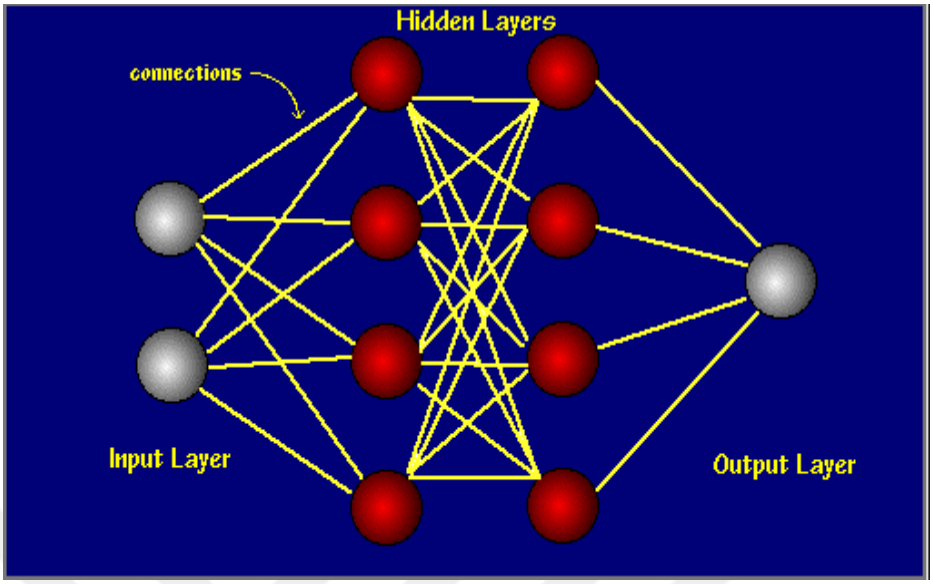

Figure.7 Neural Networks (NN) [23]

2.2.1.1 Multi-Layer Perceptron Neural Network (MLPNN):-

Literature anatomy reveals a persistent implementation of Feed Forward NN, from among the several categories of link for 'artificial neurons'. A type of feed forward NN technicality is the MLPNN. The design of MLPNN is shown in Fig 8.

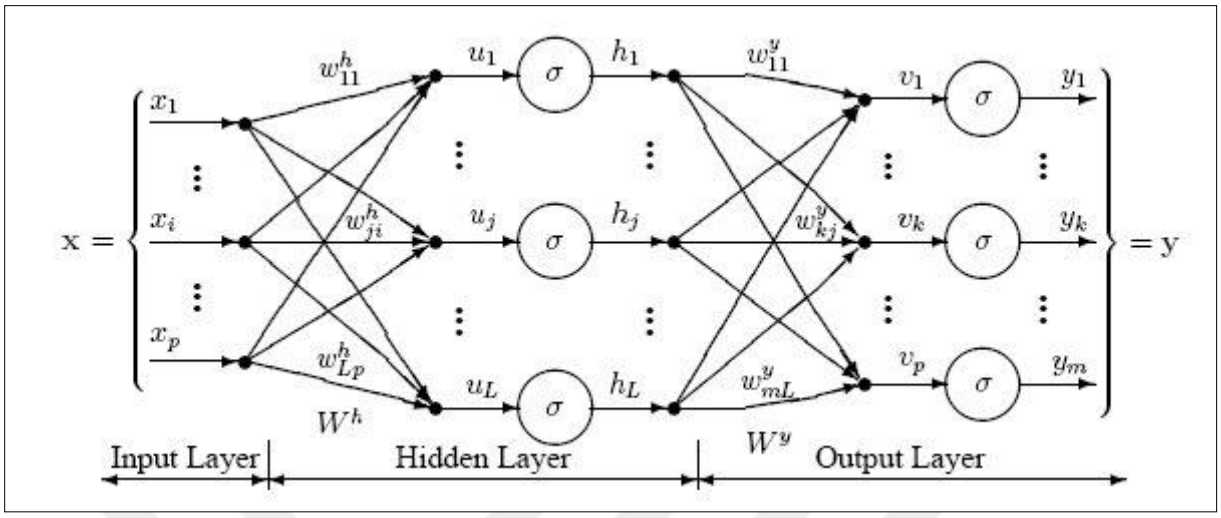

Figure. 8 Structure of MLPNN [26].

One of the most famous models in ANN is MLPNN.

Three Types of layers in the 'MLPNN': the first type is an input layer, the second one is a hidden layers. The third one is output layer type (several nodes in each layer). Information passes from one node to the next node. Outer nodes transform signals to the input layer. The 'output of first layer' passes to second layer over "weighted connection line". It accomplishes computations and sends the result to final layer (output) over weighted links. After that the output of second layer is send to last layer, for a final result [26].

The main duty of the nodes for the first layer in the MLPNN is the split all "input signal Xi" among nodes in the second layer (hidden). Every node J in the 'second layer' adds its 'input signals' Xi once it weights them with the intensity of the expert connections $W_{j\,i}$ from the first layer and fixing its output Y_j as a function F of the aggregate, specified as:

$$Y_i = F\left(\sum W_{J\,i}\, X_i\right) \qquad (2)$$

At this point it is probable for F to be a plain Activation function. An activation function is a ‘mathematical operation ‘has been done on the ‘signal output’.

Types of activation function:

(1) Eq(1) is Threshold function; when it is a positive output (1), otherwise(0). Fig 9 explains the work of the function.

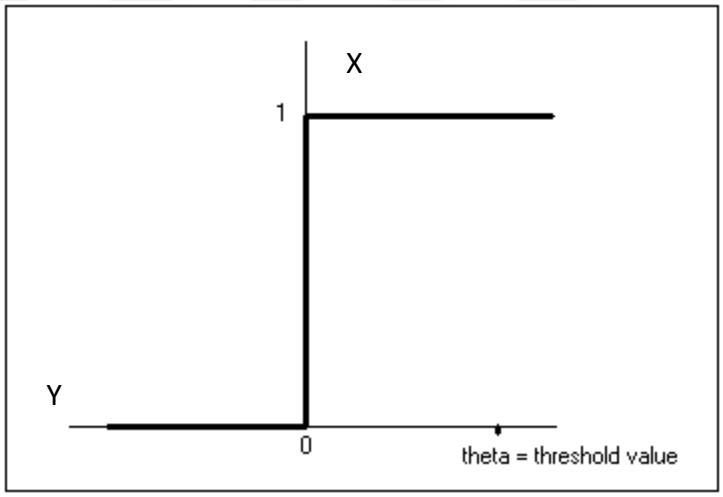

Figure. 9 Threshold functions [26]

(2) Eq (2) Sigmoid function:

$$F(x) = \frac{1}{1+e^{-x}}$$

Fig 10 shows work of the function.

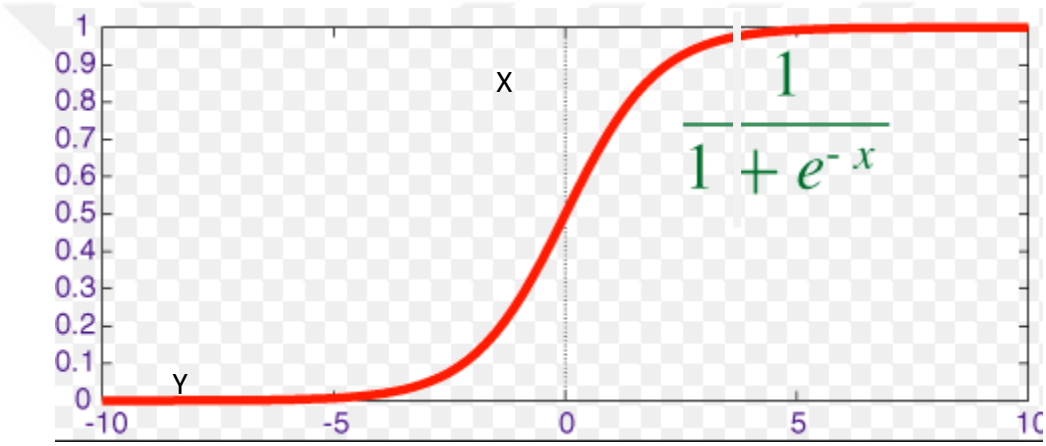

Figure. 10 sigmoid function [26]

The working steps of (MLPNN) are given in the following:

(1) Input data is given to the first layer for treatment which leads a divine output.
(2) The error is calculated by the difference between the predicted output & actual output.
(3) Back-Propagation algorithm has used to adapt the weights.
(4) For adjusting weights, it begins from weights between last layer nodes and last hidden layer nodes and works backwards through the network.
(5) When BP is finished, the operation starts again.
(6) The operation is renewed until the error between actual & predicted output to reach the least difference.

2.2.1.2. Back-Propagation Training (BP):

The BP can be appointed to practice NN. It is more familiar for applications MLPN [27]. The BP algorithm can be split into two stages:

(1) Propagation.
(2) Weight update.

Stage (1): Propagation

(a) Forward propagation of a training pattern's input through the NN so as to the propagation's output activations.

(b) Back propagation of the propagation's output activations through the NN using the training pattern's target to produce the deltas of all output and hidden nodes.

Stage (2): Weight update

For each (weight-synapse): synapse means a junction between two cells:

(a) Multiply each output by input activation to obtain the gradient of the weight.

(b) Fetch the weight in the inverse direction of the slop by laying a ratio of it from the weight [28].

Repeat stage (1) and (2) until the achievement of the grid is good.

Simple steps in NN are listed below:

(1) Receives several input.
(2) Each input is multiplied by its weight.
(3) Apply activation function (sigmoid function for example) to the sum of results.

Algorithm: Back Propagation Neural Network (BNN) for classification:

Input: - data set consisting of target values and training tuples.

I: learning rate.

Output: A trained neural network.

2.2.1.3 Methods:

Biases also start as small random values [29].

Algorithm: Back Propagation. Neural Network (BNN) for classification or numeric prediction, using the Back propagation algorithm.

Input:

D, a data set consisting of the training tuples and their associated target values;

L, the learning rate;

Network, a multilayer feed-forward network.

Output: A trained neural network.

Method:

(1) Initialize all weights and biases in *network*;

(2) While terminating condition is not satisfied f

(3) For each training tuple *X* in *D* f

(4) // propagate the inputs forward:

(5) For each input layer unit *j* f

(6) O_j D I_j; // output of an input unit is its actual input value

(7) For each hidden or output layer unit *j* f

(8) $I_j = \sum w_{ij} O_i + \Theta$; //compute the net input of unit *j* with respect to the previous layer *i*

(9) $O_j = 1/((1+(e^{-1})j)$;} // compute the output of each unit *j*

(10) // Back-propagate the errors:

(11) For each unit *j* in the output layer

(12) Err_j D $O_j(1-O_j)(T_j - O)$/; // compute the error

(13) For each unit *j* in the hidden layers, from the last to the first hidden layer

(14) $Err_j O_j$. $(1-O_j)\sum_k$ Err $_k$ w_{jk} ; // compute the error with respect to the next higher layer, *k*

(15) For each weight *w I j* in *network* f

(16) $\Delta w\ I\ j = (L) Err\ j\ Qi$; // weight increment

(17) $w_{ij} = w_{ij} + \Delta w_{ij}$; // weight update

(18) For each bias _*j* in *network* f

(19) $\Delta\Theta\ j = (l) Err\ j$; // bias increment

(20) $\Theta_j = \Theta_j + \Delta\Theta_j$; // bias update

(21)}}

2.2.1.3.1 Propagate input to forward:

First, training row is enter to the first layer (input) , then inputs enter to the input unit, without any modification ,that is $O_j = I_j$, after the input and output for each unit in second and third layers are calculate . The input to unit j is output from the last (previous) layer. Each overlap has a weight. Input is multiplied by the weight to compute the next input. And this is summed and given unit j in second or third layer.

$$I_j = \sum_I W_{Ij} O_i + \omega_j \qquad (3)$$

$W_{ij:}$ is the weight of the connection from layer I to the layer j.

$O_{i:}$ is the output of unit I from the previous layer.

ω_j: is the bias of the unit, the bias represents the threshold in that it helps to change the action of the unit.

Every unit in the hidden and output layers presented the input and then applies activation function (a sigmoid function). Fig 11 illustrates this point.

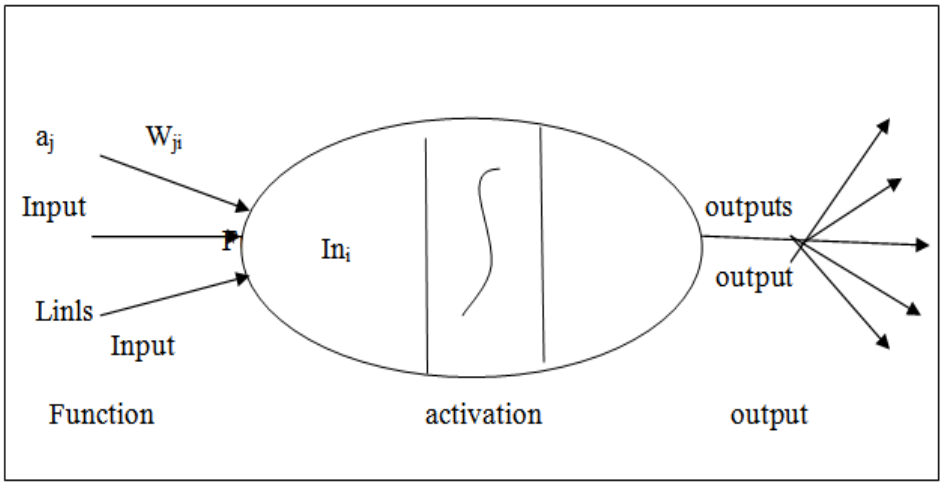

Figure. 11 Neural Network (input, activation & output) [29]

The sigmoid function is given as:

$$O_j = \frac{1}{1+e-Ij} \qquad (4)$$

I_j the next input to unit j.

O_j the output of unit j.

Back propagate errors: update the weight and biases the error is propagated to be inverted the.

$$Err_j = O_j(1-O_j)(T_j - O_j) \qquad (5)$$

O_j actual output of unit j.

T_j target value of given training tuple, connected to unit j in the next layer.

$$Err_j = O_j(1-O_j)\sum_k Err_k W_{jk} \qquad (6)$$

W_{jk} weight of connection from unit j to k in the next layer.

Err_k error of unit k.

Weights are updated by the following equation:

$$W_{Ij} = (I)\, Err_j\, O_i \qquad (7)$$

$$W_{Ij} = W_{Ij} + \Delta W_{Ij} \qquad (8)$$

I is the learning rate (the time to update the weight) having value between (0 - 1).

Biases update by the following equation:

$$\Delta\omega_j = (I)\, Err\, j \qquad (9)$$

$$\omega_j = \omega_j + \Delta\omega_j \qquad (10)$$

$\Delta\omega_j$ change in bias ω_j.

After presentation of each tuple there are updating weights and biases. Weights and biases increments could be put in variables; after all of the tuples the biases and weights are updated. This is known as epoch updating.

2.2.1.4 Types of Neural Network (NN)

(NN) types can be classified on following attributes:-

(1) Connection type:
   (a) Static (feed forward).
   (b) Dynamic (feedback).

(2) Topology:
   (a) Single layer Fig 12 shows Single layer. Just input and output layer without hidden layer.

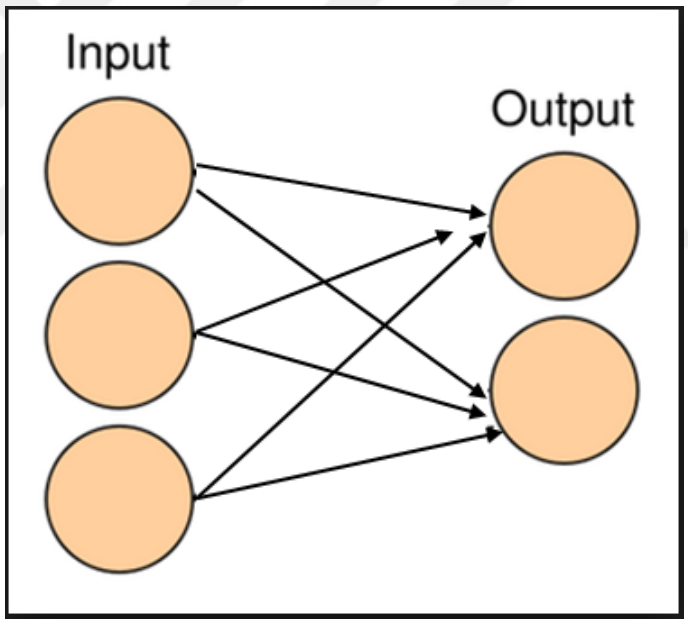

Figure .12 Single layers [29]

   (b) Multilayer. Fig 13 shows Multilayer structure. Three layers can be shown in this Figure.

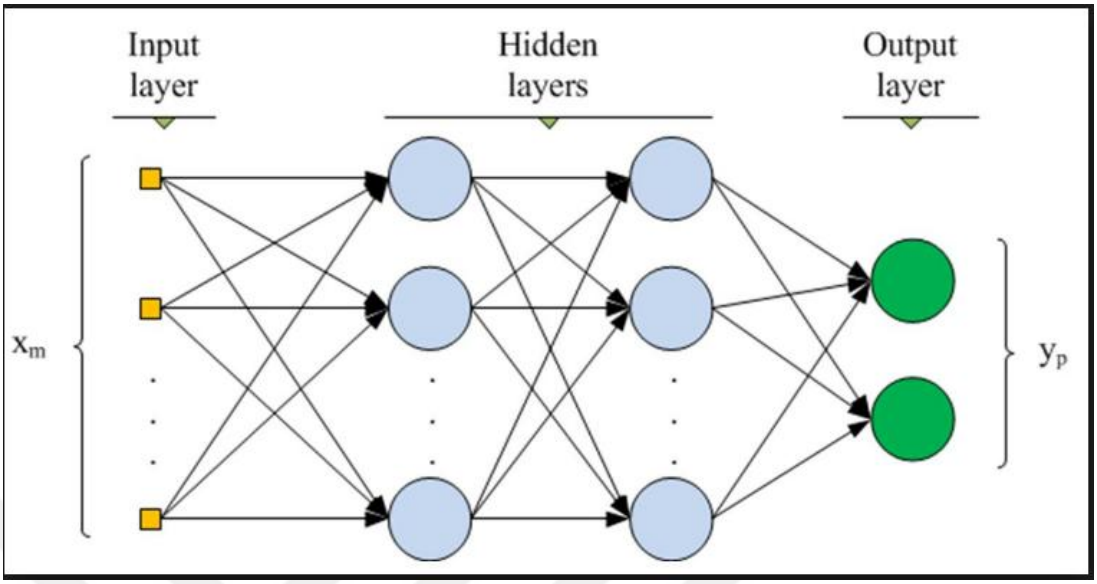

Figure. 13 Multilayer [29]

(a) Recurrent. Fig 14 shows recurrent structure three layers (input, multilayer's in hidden layers and last one output layer).

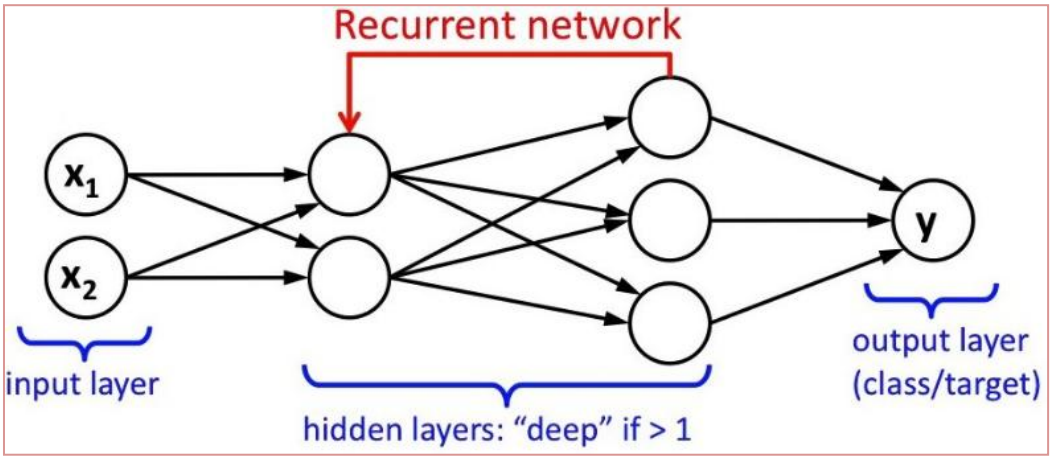

Figure. 14 Recurrent [29]

(3) Learning methods:

(a) Supervised (1) each training pattern =input +desire output.

(2) Weights must be ready at each presentation.

(b)Unsupervised: no training data, no help from the outside, no information available on the desired output.

(c) Reinforcement: teacher: training data, the teacher sign (scores) the execution of the training examples then uses these scores to mix (shuffle) weights randomly.

2.2.2 Feature Subset Selection (FSS):

FSS is a pre-processing phase usually used in ML Its concept is to remove unrelated data thus gets more accuracy.

Three interest keys of implement FSS on your dataset are:

(1) Reduces Over fitting space.
(2) Improves Accuracy..
(3) Reduces Training Time.

2.2.2.1 Sections of Feature Selection Subset (FSS):

FFS is divided into two sections:

(1) Attribute Evaluator.
(2) Search Method.

Every type has multiple techniques from which to choose.

Attribute evaluator: - Is the technique by which each feature or attribute in your dataset is evaluated in the state of the output variable (e.g. the class).

Search method: - is the technique by which to try to move to different set of attributes in the dataset to reach on a small list of subset input variables.

Although graph search algorithms are famous, some features evaluators techniques impose the use of specific search Methods. For example, the Correlation Attribute Eval technique use with ranker that evaluates each attribute and lists the results in a rank order. Sometime message appears to ask you to change the search method when selecting different Attribute Evaluators, to compatible with the chosen technique. Fig 15 shows the work of FSS.

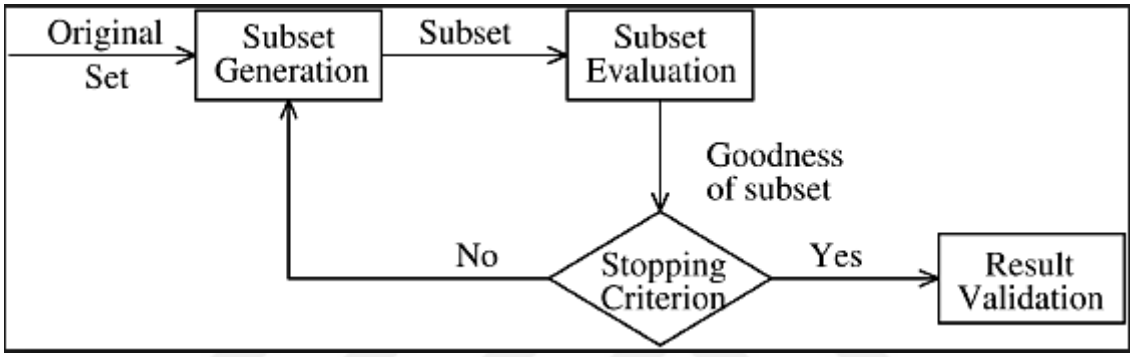

Figure. 15 Feature Subset Selections (FSS) [23]

2.2.3 Naive Bayes (NB):

The (NB) algorithm is an easy probabilistic classifier that computes a set of "probabilities" by computing the combination and frequency values in the DB. The NB classifier depends on Bayes theorem. Condition independence can be used in this classifier algorithm; in another words, it suppose that an features value on a specified class does not depend on the values of other attributes. NB is depends on supervised learning and it is a statistical classifier [30].

NB depends on Bayes theorem

$$P(A\backslash B) = \frac{P(B\backslash A)p(A)}{P(B)} \qquad (11)$$

Where

A & B are events.

P (A) probabilities of A.

P (B) probabilities of B.

P (A\B) is the probabilities of observing event A given that B is true.

P (B\A) is the probabilities of observing event B given that A is true.

We prefer (NB) algorithm:

(1) When the data is large.

(2) When the features (attributes) are in depended of each other.

(3) When we look forward to the efficient output.

2.2.4 J48 classifier

J48 classifier is an easy C4.5 DT for classification. The DT approach is important in classification problem. With this algorithm a binary tree is built to model the "classification" [31]. At the point the tree is built, it is implement to each row in the DB.

NB based on probability.

J48 is based on the DT.

2.3 WEKA Tool: -WEKA is a DM system improved and developed by the "University of Waikato in New Zealand" that testes DM techniques by using the "JAVA language". WEKA is a collection of ML algorithms for DM tasks. The algorithms are conducted directly to DB. Develop WEKA tests DM techniques for data pre-processing new ML schemes that can also with this package. WEKA is "open source" SW issued. The type of data normally applied by WEKA is in 'ARFF' file format, which contains particular tags to point out several things in the data. WEKA is an invention tool in the story of the DM and machine learning communities. By

putting efforts since 1994, this tool was evolved by WEKA team. WEKA contains many available algorithms for DM and ML. It is open source and complimentary. The people who do not having much knowledge of DM can also use this software.

3. RESULTS AND DISCUSSIONS

3.1 Data source

The STATLOG (HD) is used dataset on DM tools such as WEKA. This DB is taken from the repository (UCI) [32]. It includes (270) rows of healthy persons and the sick with HD problems. It comprises class with13 columns as listed down in Table 1. This table illustrates a set of "features or attributes" with descriptions and their values. For example age mean the age of the patient in years like 60 or 50 years, sex=1 if male or 0 if female, chest pain has four types and so on for other attributes.

Table. 1 Description of 13 Attribute used [32]

No.	Attributes	Description	Value
1	age	Age in years	continuous
2	sex	Male or female	Male=1 female=0
3	cp	Chest pain type	Typical type 1=1 Typical type again=2 Non-again pain=3 Asymptomatic =4
4	thestbps	Resting blood pressure	Continuous in mm hg
5	Chol	serum	Continuous in mm/dl
6	restecg	Resting electro graphic result	Normal=0 Having ST_T wave Abnormal=1 left ventricular
7	fbs	Fasting blood sugar	1>= 120mg/dl 0<120 mg/dl
8	Thalach	Maximum heart rate achieved	Continuous value
9	exang	Exercise induced angina	No=0 Yes=1
10	Old peak	St depression induced by exercise relative to rest	Continuous value
11	slope	Slope of the peak exercise St segment	Un sloping=2 Flat=2 Down sloping=3
12	ca	Number of major vessels colour by fluoroscopy	Value(0-3)
13	thal	Defect type	Normal=3,fixed=6, Reversible defect=7

In order to find most accurate with less attributes the NN with FSS is applied in this thesis.

Three DM techniques with FSS are applied and different evaluators with different search methods are selected.

For example ‘cfsset eval’ evaluator with best first search method is applied. then implemented with correlation attribute evaluator with ranker search method. After that accuracy is recorded. Table 2 shows confusion matrix which shows the number of records true or false.

Confusion matrix includes information about actual and predicted classification done by a classification system.

TP “True Positive”: It means the number of records classified as true while they were really true.

FN “False Negative”: It means the number of records classified as false while they were really true.

FP “False Positive”: It means the number of records classified as true while they were really false.

TN “True Negative”: It means the number of records classified as false while they were really false.

Table .2 Confusion Matrix

	A(has HD)	B(no HD)
A(has HD)	TP	FN
B(no HD)	FP	TN

Table 3 illustrates neural network (NN) and FSS with different evaluators and search methods. According to Table 3 the accuracy has the maximum value (99.2593) when “Correlation” attribute’ and ‘Ranker’ are chosen together. However, the accuracy has minimum value (96.293) when ‘Cfs set Eval’ and ‘Best first’ are chosen together.

-

Table 3 Accuracy of NN and FSS with different evaluator and search methods

Evaluator	Search method	Accuracy (%)
Cfs set Eval	Best first	96.2963
Correlation attribute	Ranker	99.2593
Gain	Ranker	97.77
One AttributeEval	Ranker	98.5185

In Table .4 the three DM techniques (NN, J48 and NB) are shown before implementation of FSS. The NN algorithm gives highest accuracy with (99.6296). The J48 algorithm with (91.4815) and Naïve Bayes with (85.182).

Table 4 different DM techniques before implement (FSS)

DM Techniques	Accuracy (%)
NB	85.182
J48	91.4815
NN	99.6296

X-axes corresponds the three DM technique, Y-axes corresponds accuracy.

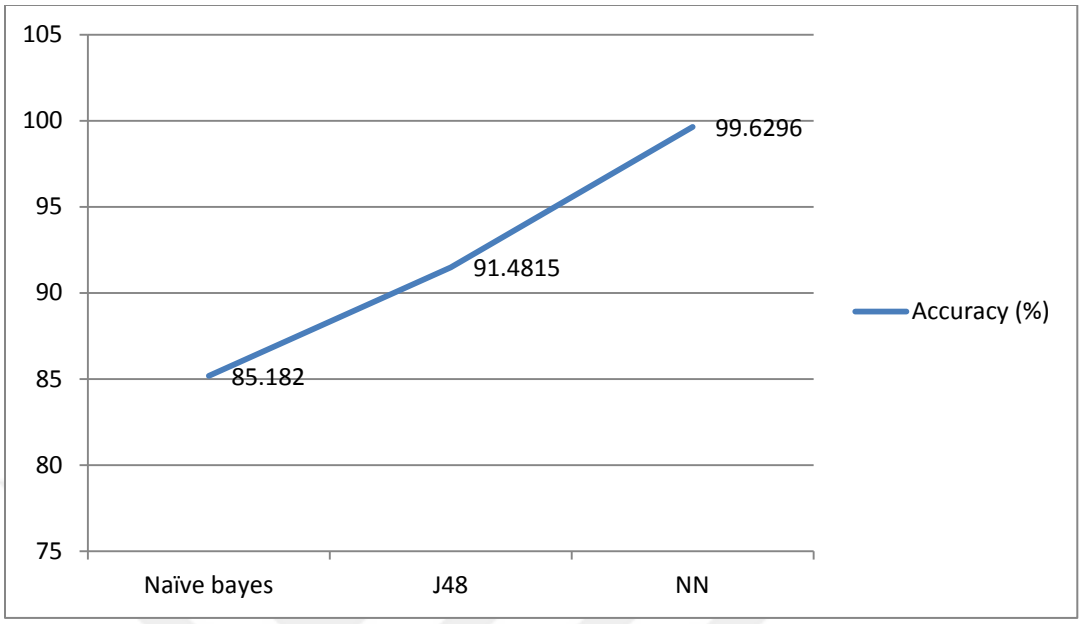

Figure .16 the comparison of the accuracy of the three algorithms (NB, J48, NN)

When WEKA tool is applied, the output of true and false classifications can be seen through confusion matrix. In case of (NB) in Fig 17 confusion matrix explains the true classify 230 from 270 which means 40 records are false. The calculation of accuracy is given with this formulation = (true classify record/total)*%100 = (230/270)*%100=85.185%. In Fig 17 “a” means actual no HD (absent), “b” means actual yes has HD (present).a=133+23=156 (no HD) absent, b=17+97=114 (yes HD) present.

Total =270

True classify 133+97=230 (record)

False classify 17+23=40 (record)

Accuracy =85.182.

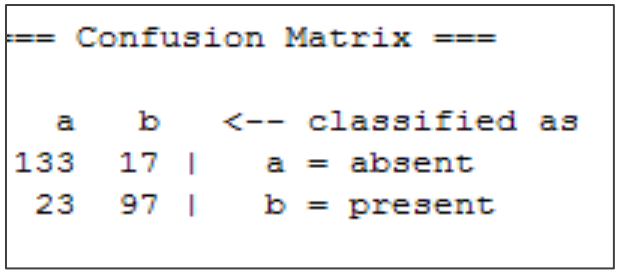

Figure. 17 shows the confusion matrix for (NB)

Similarly, confusion matrix for (J48) in Fig 18 the accuracy = (247/270)*%100=91.4815:-

True classify 141+106=247 (record)

False classify 9+14 =23 (record)

Accuracy =91.4815.

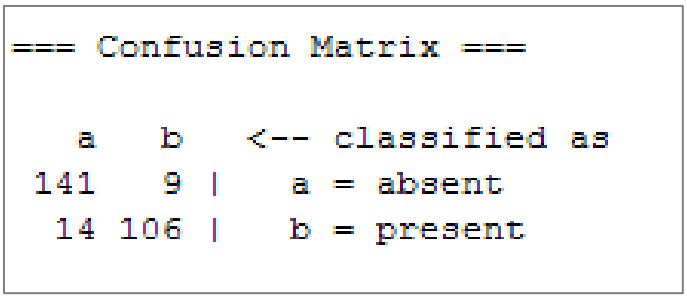

Figure . 18 confusion matrix (J48)

Finally, confusion matrix for NN in Fig 19 the accuracy = (269/270)*%100=99.6296 %:-

True classify 149+120=269 (record)

False classify 1+0=1 (record)

Accuracy = (99.6296)

```
=== Confusion Matrix ===

   a   b   <-- classified as
 149   1 |   a = absent
   0 120 |   b = present
```

Figure . 19 Confusion matrixes for (NN)

Table .5 shows accuracy of three DM Techniques (NN, J48 and NB) in second study and the time is taken to build the system after applying FSS (a cfs and best first). The attributes reduce from 13 to 7 are shown. The selected attributes is the same for all techniques and they are explained in Fig 20. The selected attributes is shown in Fig 20.

Table 5 After FSS when apply cfs set eval and best first are applied

DM Techniques	Accuracy (%)	Time taken to built the system	No. of attribute
NB	85.9259	0.02 sec	(7)
J48	87.7778	0.06 sec	(7)
NN	96.2963	0.59 sec	(7)

As it is clear in Table. 5 the same attributes were chosen in all techniques. The Fig. 20 shows the (7) selected attributes from 13 attributes mentioned in Table 1. The chosen attributes such as (chest, resting, maximum heart rate achieved etc) have more effect to diagnose the HD.

```
Selected attributes: 3,7,8,9,10,12,13 : 7
                     chest
                     resting_electrocardiographic_results
                     maximum_heart_rate_achieved
                     exercise_induced_angina
                     oldpeak
                     number_of_major_vessels
                     thal
```

Figure 20 selected attributes are shown

Table. 6 shows the three DM techniques (NN, J48, and NB) with accuracy and model time after using FSS (second study) (correlation and Ranker). All 13 attributes are used.

Table 6 After FSS when correlation attribute and Ranker are applied

DM Techniques	Accuracy (%)	Model time	No. of attribute
NB	85.1852	0.01 sec	13
J84	91.4815	0.01 sec	13
NN	99.2593	0.72 sec	13

Fig 21 shows the order of attributes according to diagnose HD, the first attribute (thal) with Rank (0.525) is greater importance for diagnosing the disease followed by (number of_major_vessls) with Rank (0.4553) which has less effect than it down to the last one (fasting_blood_sugar) with Rank (0.0163) which has the least impact on diagnosis of the disease.

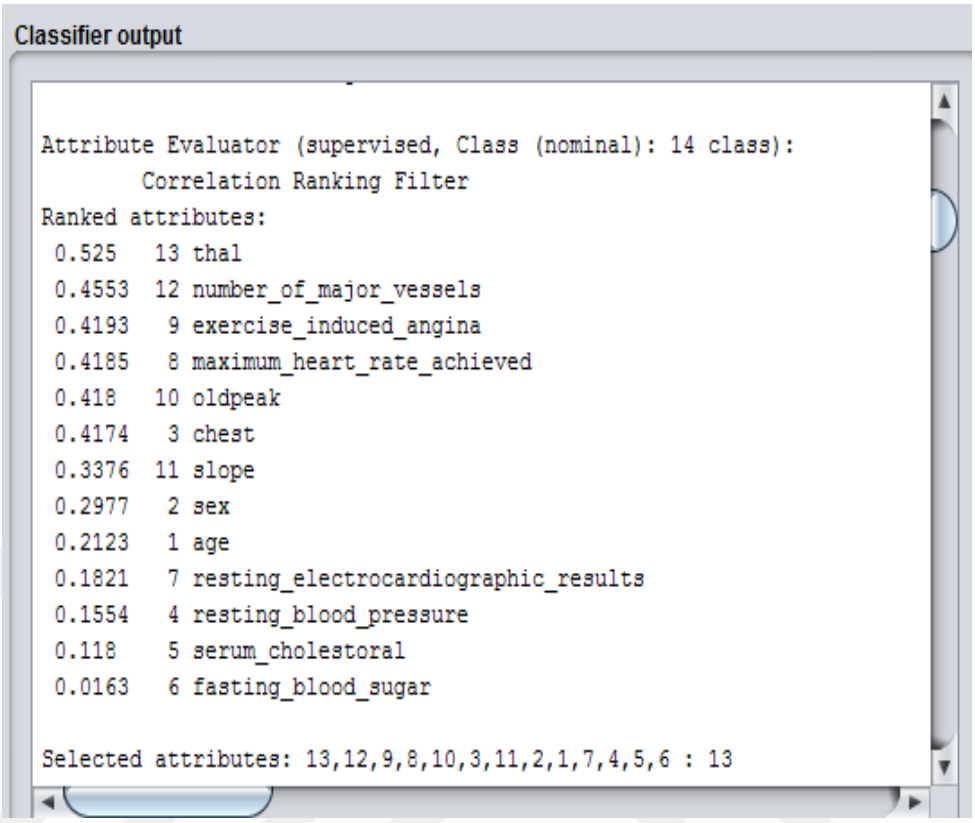

Figure 21 classifier output

Fig 22 shows the WEKA interface and how to choose the classifier, evaluator and search method.

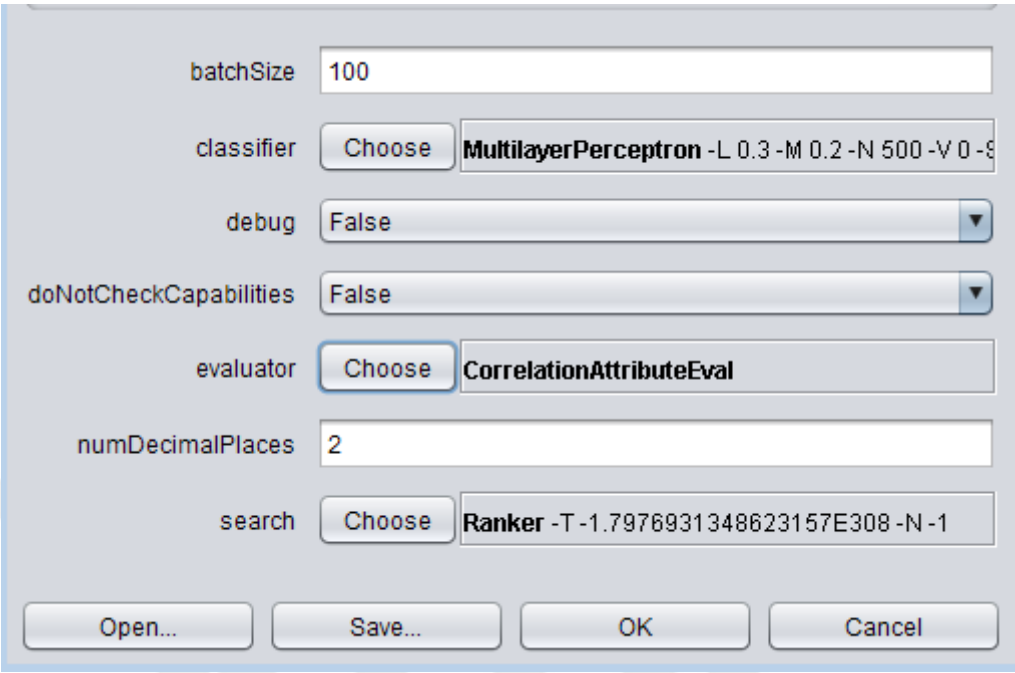

Figure. 22 shows the graphic to choose classifier

As shown in Table 5 Naïve Bayes has reached the best accuracy after FSS is used when (csf set evel and best first search method is considered). However, it has the same accuracy when (correlation attribute and ranker search are applied).J48 gives the same accuracy before and after FSS and when (CSF set Eval and best first search method are chosen) we get less accuracy .NN has higher accuracy before applying FSS than after applying FSS.

Number of node in hidden layers changes the accuracy value. Different no. of "node in hidden layers" are applied, we reach to the best result when the number of nodes in hidden layers = 9. See Fig 23: thirteen green rectangles mean the attributes (patient tests) and nine red balls mean the node in hidden layers two yellow balls mean output layers.

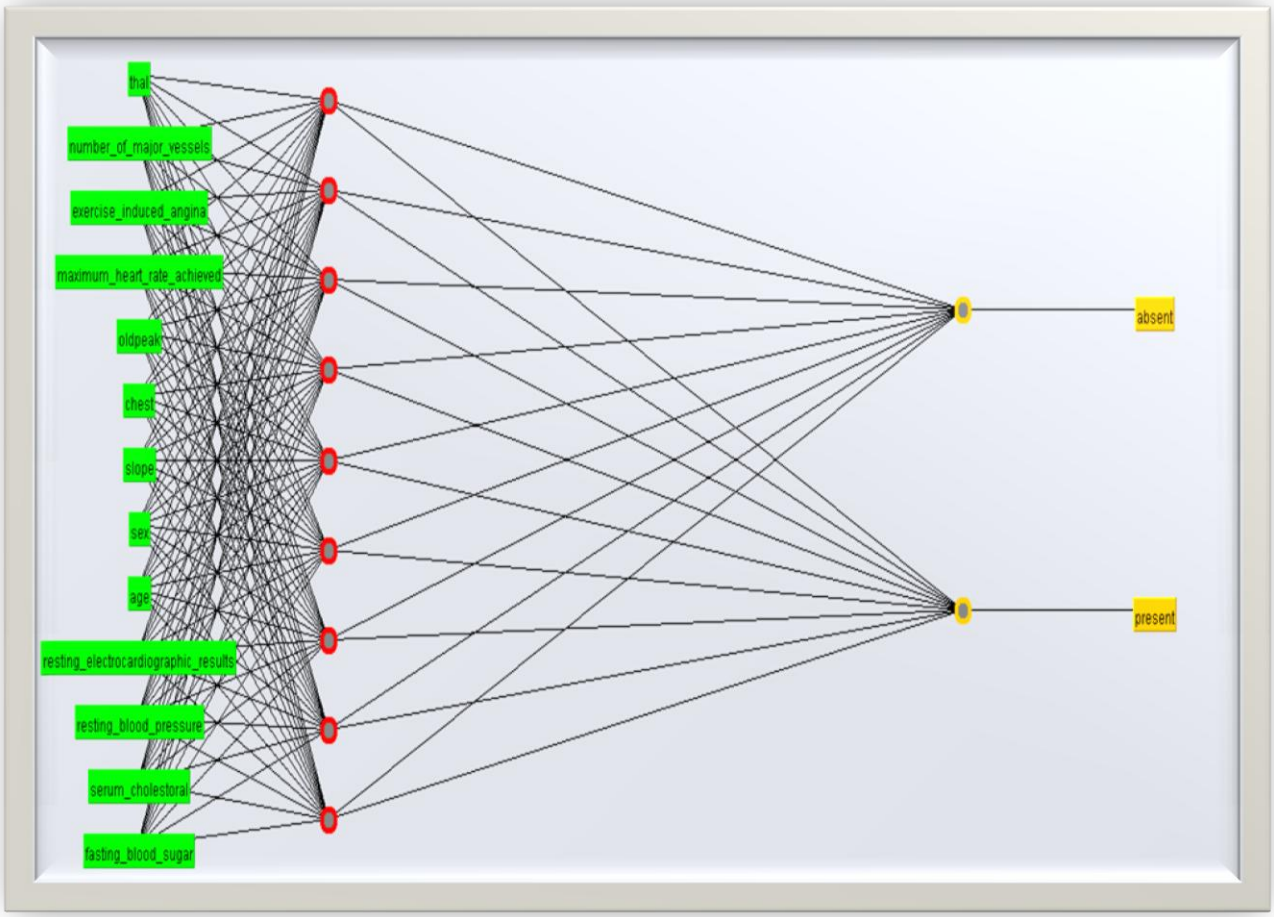

Figure .23 Graphic shows the 9 nodes in hidden layers

Fig 24 shows the Accuracy when different attributes are selected according to ranker as explained previously in Fig 21.

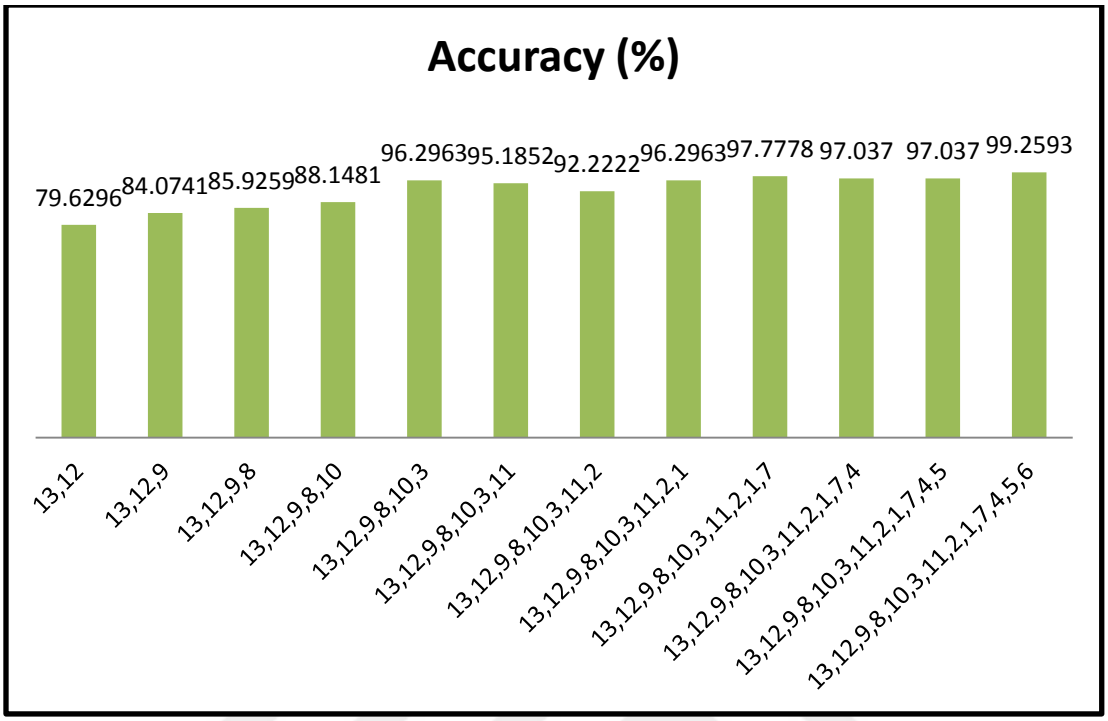

Figure. 24 shows the accuracy attributes

In Fig 24, two attributes chosen (12 and 13) in the beginning. They are selected because they have the higher rank and accuracy was (79.629%). After that three attributes (9, 12, and 13) are chosen with accuracy (84.074%). This procedure continues until we arrive to last experiment when we choose 13 attributes with higher accuracy (99.259%).

Table. 7 shows the results of the previous studies using the same dataset (STATLOG.ARFF) in different DM techniques and compares with the results of this thesis. As we see in Table 7 three DM techniques (NB, J48 and NN) are used in the first study and NN gets the highest Accuracy (99.62%) and second study considers the same three DM techniques (NB,J48 and NN) with FSS and NN also show the highest accuracy (99.25%). Pritam H,patil used Naïve Bayes with the same data set with different tool named (KNIME) and showed accuracy =83.70%. Chaitrali S, and Sulabhs S.Apt used the same data set on WEKA with (13) attributes and added two new attributes (obesity and smoking) resulting in an increasing accuracy of NB to (94.44%) and DT to (96.66%). In this thesis FSS and NN also get the highest Accuracy (99.25) when compare with other two techniques(J48 and NB).

Pritam Hpatil used different DM tools named (KNIME) is somewhere near to WEKA. But in this thesis same data set are apply on (WEKA) therefore the result are different. Chaitrali S,and Sulabhs S.Apt added two new attributes (15) attributes applied on WEKA tool which led to increased the accuracy.

In this thesis first study (13) attributes have used on NN and accuracy with (99.629%), however

Chaitrali S, and Sulabhs S.Apt used (15) attributes on NN and accuracy with (99.25%).

In this thesis, when we use NN in (first study) compare with (second study) NN &FSS accuracy has been reduced very little value from (99.62%) to (99.25%), but on the other hand we helped to decrease the number of tests asked from the patients who will help to reduce the time, cost and effort.

Table.7 comparison with previous studies

Author	Year	Previous studies	Accuracy rate (%)
Ebenezer[33]	2015	Support vector machine	87.5
Pritam H,patil[34]	2014	NB	83.70
ChaitraliS.Dangare[35] Sulabha S. Apte, PhD.	2012	NB	94.44
		DT	96.66
		NN	99.25
In this thesis	2017	NN J48 NB	99.6296 91.4815 85.182
		NN & FSS J48 & FSS NB &FSS	99.25 91.48 85.9259

4. CONCLUSION AND FUTURE WORK

4.1 Conclusion

Recently the percentage of people who suffered from heart diseases has increased as a result of the pressures of life and unhealthy habits. Therefore, the hospitals suffered from increasing the number of the sick and the number of tests which asked from them. The 13 tests are conducted for each patient. This is wasting time and effort for patients and specialists as well as the cost. For this reason, the researchers have been interested to solve this problem by using many techniques, one of them DM techniques. DM is a critical step in extraction of knowledge from huge database.

In this thesis, there are two studies that are conducted on STATLOG.ARFF dataset which contains on 13 attributes by using WEKA software. The first study means using three DM techniques (NN, J48 and NB). Second study means using the same three DM techniques with FSS.

The first study, the three DM algorithms (NB, J48 and NN) are applied and compared with each other using all the attributes. According to the results from the chapter 3, NN algorithm shows the best accuracy comparison with other (J48 and NB). Moreover, the accuracy of NN algorithm shows better result when it is compared with previous studies in Table 7. While the second study used the same three DM techniques (NB, J48 and NN) with FSS algorithm which is an important affair in classification . It decreases the number of dimensions of the DB from 13 to 7 attributes; so that, the processor and memory usage minimize; the data becomes more sensible and easier to study on. Although the accuracy was reduced slightly in (NN with FSS) comparison with first study . Accuracy in (J48 & FSS) not changes when compared with first study. Accuracy in (NB & FSS) increased slightly when compared with first study. The Second study used different evaluators and search methods. According to the results in Chapter 3.

The second study solved the problem of the large number of tests, by reducing the number the attributes from 13 to 7. The aim of this study provides the time, cost and effort for the patients after they suffered from spending a long time in hospital and lose much money.

4.2 Future Work

In the future, it is recommended to use many datasets on the same the algorithms and compared with results to get on the highest accuracy.

Reference:-

[1] Dr. Kusha Rani," Analysis of heart diseases dataset using neural network approach ", International journal of data mining & knowledge management process (ijdkp) volume1, no.5, September 2011.

[2] Suvma Thube, Bhaki,Bhaki ratnaparkhi, K.Rajeswari, "Analysis of different data mining tools using classification, clustering and association rule mining", International journal of computer applications (0975 – 8887) volume 93, no.8, May 2014.

[3] Anamika Gupta Naveen Kumar Vasudha Bhatnagar "Analysis of medical data using data mining and formal concept analysis" proceedings of world academy of science engineering and technology volume 6, June 2005.

[4] S Stilou P D Bamidis N Maglaveras C Pappas "Mining association rules from clinical databases: an intelligent diagnostic process in healthcare" Stud Health Techno Inform volume 84, no. Pt 2 pp. 1399-1403 2001.

[5] Hian Chye Koh Gerald Tan "Data mining applications in healthcare" Journal of healthcare information management volume.19, no. 2 pp. 64-72 2005.

[6] MA Jabbar, B.L Deekshatulu & Priti Chandra, "Classification of heart disease using artificial neural network ", Global Journal of computer science and technology neural &artificial intelligence, 2013.

[7] Anbarasi et al. " Enhanced prediction of heart disease with feature subset selection using genetic algorithm" International Journal of Engineering Science and Technology volume 2, no. 10 pp. 5370-5376 2010.

[8] MA.Jabbar, B.L Deekshatulu & Priti Chandra, "An evolutionary algorithm for heart disease prediction", CCIS pp 378-389 Springer (2012).

[9] MA. Jabbar et.al," Knowledge discovery using associative classification for heart disease prediction", AISC 182 pp29-39, Springer (2012).

[10] MA. Jabbar et.al, "Knowledge discovery from mining association rules for heart disease prediction", JAJIT, Volume 41, (2) pp 45-51 (2012).

[11] MA. Jabbar, B.L Deekshatulu & Priti Chandra, "Cluster based association rule mining for heart Disease prediction", JATIT volume .32, no. 2 October (2011).

[12] Pritam h.Patil, Suvarna Thube Bhakti Ratnaparkhi. K.rajeswa "Analysis of different data mining tools using classification, clustering and association rule mining "International Journal of Computer Applications (0975 – 8887) Volume 93, No.8, May 2014.

[13] Sellappan Palaniappan " intelligent heart disease prediction system using data mining techniques" , International Journal of Computer Science and Network Security in 2008.

[14] M.akhil jabbar ,b.l deekshatula " Classification of heart disease using k- nearest neighbor and genetic algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications in 2013.

[15] Sarojinibala Krishnan "Comparative study on the performance of mmifs and dmifs feature selection algorithms on medical datasets", Indian Journal of Science & Technology in 2015.

[16] MA.jabbar, B.L Deekshatulu & Priti Chandra, "An evolutionary algorithm for heart disease prediction", ccis pp 378-389springer (2012).

[17] John Shafer, Rakesh Agarwal, & Manish Mehta " Sprint:a scalable parallel classifier for data mining", in proc. of the vldb Conference, Bombay, India.1996.

[18] Chaitrali & Sulabha " A data mining approach for prediction of heart disease using neural networks", International Journal of Computer Engineering and Technology (ijcet), volume 3, issue 3, October-December 2012.

[19] Noura Alnuaimi, Mohammad m Masud and Farhan Mohammed "Examining the effect of feature selection on improving patient deterioration prediction", International Journal of Data Mining & Knowledge Management process (ijdkp) volume 6, no.6, November 2015.

[20] Beant Kaur, Williamjeet Singh "Review on heart disease prediction system using data mining techniques", International Journal on Recent and Innovation trends in Computing and Communication issn: 2321-8169 volume 2, issue: 10.

[21] Frank lemke Johann-Adolf Mueller “Medical data analysis using self-organizing data mining technologies”, Systems Analysis Modeling Simulation volume 43, no. 10 pp. 1399-1408 2003.

[22) Kurgan a. Lukasz cios J. Krzysztof Tadeusiewicz Ryszard “ Knowledge discovery approach to automated cardiac spect diagnosis” Artificial Intelligence in Medicine pp. 149169 2001.

[23] Dr.K. Usha Rani “ Analysis of heart diseases dataset using neural network approach “ International Journal of Data Mining & Knowledge Management process (ijdkp) volume1, no.5, September 2011.

[24] E. O.Olaniyi, K. Adnan., “ Onset diabetes diagnosis using artificial neural network”, International Journal of Scientific and Engineering Research, volume 5, issue 10,October (2014).

[25] N. Kilic, B. Ekici, S. Hartomacioglu, “Determination of penetration depth at high velocity impact using finite element method and artificial neural network tools,” Defense Technology xxpp.1-13, 2015.[online]: http://dx.doi.org/10.1016/j.dt.2014.12.001.

[26] Chaitrali, Sulabha, “A data mining approach for prediction of heart disease using neural network “ International Journal of Computer Engineering and Technology volume 3, October -December 2012.

[27] Dhanashree s. Medhekar1 , Mayur p. bote2 , Shruti d. Deshmukh3 " Heart disease prediction system using Naive Bayes", International Journal of Enhanced Research in Science Technology & Engineering, 2013.

[28] A.t Sayad,P,P Halkarnik " Diagnosis of heart disease using neural network approach", International Journal of Advances in Science Engineering and Technology, issn: 2321-9009, volume- 2, issue-3, July-2014.

[29] Jiawei han & Micheline Kamber "Data mining concepts and techniques second edition" 2012.

[30] N. AL-Milli, " Back propagation neural network for prediction of heart disease," Journal of Theoretical and applied Information Technology, volume. 56, pp.131-135, October10, 2013.

[31] Lakshmi Devasena c "Proficiency comparison of random forest and J48 classifiers for heart disease prediction ", International Journal of Computing Academic Research (ijcar) issn 2305-9184, volume 5, no. 1 (February 2016), pp.46-55.

[32] Http.seasr.org/datasets/u://repository ci/arff.

[33] Ebenezer Obaloluwa Olaniyi & Oyebade, Khashman Adnan, "Heart diseases diagnosis using neural networks arbitration" , i.j. Intelligent Systems and Applications, 2015, 12, 75-82 published online November 2015 in mecs (http://www.mecs-press.org/) doi: 10.5815/ijisa.2015.12.08.

[34] Pritam h. Suvrna Thube, Bhakti, Krajeswari, “Analysis of different data mining tools using classification, clustering and association rule mining”, International Journal of Computer Applications (0975 – 8887) volume 93, no.8, May 2014.

[35] Chaitrali s. Dangare Sulabha s. Apte, Phd. “ Improved study of heart disease Prediction system using data mining classification techniques”, International Journal of Computer Applications (0975 – 888) Volume 47, No.10, June 2012.