Volume 5, Issue 6, June 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Comparative Study of Data Mining Techniques for Predicting Disease Using Statlog Heart Disease Database

Shrawan Ram
Assistant Professor
Department of Computer Science and Engineering,
M.B.M Engineering College, J.N.V. University,
Jodhpur, India

Amit Doegar
Assistant Professor
Department of Computer Science,
National Institute of Technical Teachers Training and Research, Chandigarh, India

Abstract:- ***Data Mining (DM), frequently treated as synonymous to Knowledge Discovery in Databases (KDD) is actually a part of knowledge discovery process and is the process of extracting information including hidden patterns, trends and relationships between variables from large databases in order to make the information understandable and meaningful. The ultimate goal of data mining is prediction of unknown patterns and predictive data mining is the most common type of that which has the most direct real life applications. The process basically consists of three stages: (1) the initial data exploration, (2) model building or pattern identification with validation/verification process and (3) deployment of the data mining model. Therefore, in this research paper data mining techniques will be compared using the benchmark datasets. The different types of data classification methods and techniques are available such as Statistics, Visualization, Clustering, Decision Tree, Association Rule, Neural Networks, K-Nearest Neighbor Method and Genetic algorithms. The objective of this research paper is to do the comparative study and evaluation of decision tree, artificial neural network with the help of Statlog Heart Diseases Database collected from UCI machine learning repository. The advantages and disadvantages, of the data mining techniques depend on the capability and efficiency of the data mining techniques or algorithms to classify the large volume of database and predicting the relevant patterns for decision making process. The consequences of choosing any technique and the methods of implementation is very important factor. Data mining techniques such as Decision Tree and Artificial Neural Networks are used for the classification of Statlog heart disease datasets. These supervise machine learning algorithms are compared on the basis of classification accuracy and performance matrices.***

Keywords- ***Data Mining, Knowledge Discovery in Databases, Statlog Heart Disease Database, K-nearest Neighbor Method, Genetic Algorithm.***

I. INTRODUCTION

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavour, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details, and government statistics and Electronic Health Records) to the more exotic (such as images of astronomical bodies, molecular databases, and medical test records). With the rapid development of advanced computing resources, Internet technology and information processing tools and techniques in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in gigantic storage devices or owing into and out of the system in the form of data streams.

Data mining is an essential step in the knowledge discovery in databases (KDD) [18]. The terms of KDD and data mining are different; KDD refers to the overall process of discovering useful knowledge from data. Data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge [18]. In information era, knowledge is becoming a crucial organizational resource that provides competitive advantage and giving rise to knowledge management (KM) initiatives. The goal of pattern mining is to find item sets, sub sequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers.

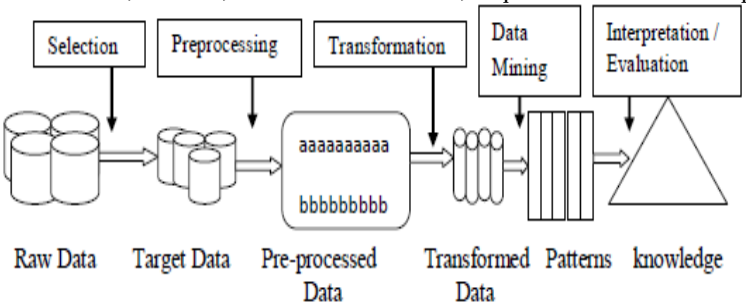

Figure 1.1: Data Mining Process

II. TYPES OF DATA MINING TECHNIQUES

Data mining plays a vital role in various applications such as business organizations, e-commerce, healthcare industry, scientific and engineering. Various data mining techniques are available for predicting hidden patterns namely Classification, Clustering, Association rules and Regressions. In the healthcare industry, the data mining is mainly used for diagnosis of the diseases from the clinical datasets. It is not very easy to define the data mining process but on the basis of processing the data and generating the knowledge, Data mining is defined as the Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns for decision making process [14]. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification process, and (3) deployment of the model. Classification is an important problem in data mining, which identifies essential features of different classes based on a set of training data and then classifies new unseen instances into appropriate classes. A classification of the data mining methods, would greatly simplify the understanding of the whole space of available methods. Furthermore, most data mining products either do not provide intelligent assistance for addressing the data mining process or tend to do so in the form of rudimentary "wizard-like" interfaces that make hard assumptions about the user's background knowledge. So the very simple and user friendly interface should be developed for the most of the data mining process [4].

III. DATA MINING MODELS AND TASKS

Data mining tasks are classified as predictive data mining and descriptive data mining, both are very useful tasks and used for many real life applications. Predictive data mining is more important tasks used for the classification of complex and multi valued datasets. Classification maps data into predefined groups or classes and the supervised learning algorithms are used for the classification of data sets. The classification, regression analysis, time series analysis and prediction are also predictive data mining tasks.

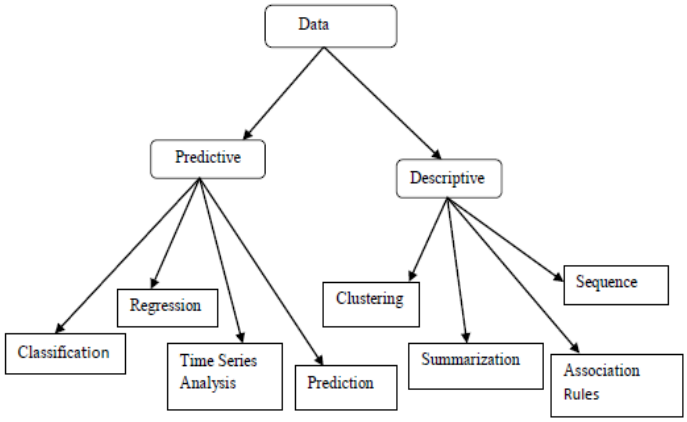

Figure 3.1 Data Mining Tasks

A variety of mathematical algorithms are used in data mining to analyze historical data. The results of this analysis are then used to build models based on real world behavior, which are in turn used to analyze incoming data and make predictions about future behavior [12]. Data mining as defined is the process of extracting hidden patterns from large databases. The classification and prediction models with the help of training datasets, models are trained to mine the data and find the relevant patterns.

IV. LITERATURE REVIEW

Different data mining techniques are used for finding the hidden and novel information for decision making purpose. During the literature survey it is found that there are many types of data mining techniques and tools available for research and commercial purpose. Many enterprises are using these techniques to provide the quality of services to their customers because the time and quality of service is more important. The literature reviews are followings:

1. Mai Shouman, Tim Turner and Rob Stocker, "Using data mining techniques in heart Disease Diagnosis and Treatment, March", 2012 [1].
2. A. Ling Kock Sheng and Teh Ying Wah, "A Comparative study of data mining techniques in predicting consumers' credit card risk in banks", 2011 [2].
3. Benish Fida, Muhammad Nazir, Nawazish Naveed and Sheeraz Akram, "Heart Disease Classification Ensemble Optimization Using Genetic Algorithm", 2011 [3].
4. Wei-Pin, Changa, Der-Ming and Liouc, "Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data", 2006 [4].
5. Henley and Hand, "Statistical Classification Methods in Consumers Credit Scoring, A Review", (1996) [5].
6. Yavuz Del_Cani, Lale Ozyilmaz and Tulay Yildirim, "Evolutionary Algorithms Based RBF Neural Networks for Parkinson's disease Diagnosis", 2011 [6].
7. Berardi, Patuwo and Hu, "A principled approach for building and evaluating neural network classification models", 2004 [7].
8. Qeethara Kadhim Al-Shayea , "Artificial Neural Networks in Medical Diagnosis", 2011 [8].

V. RESEARCH METHODOLOGIES

PROPOSED CLASSIFICATION MODEL

Data mining techniques usually fall into two categories, predictive and descriptive. Predictive DM uses historical data to infer something about future events. Predictive mining tasks use data to build a model to make predictions on unseen future events. Descriptive DM aims to find patterns in the data that provide some information about internal hidden relationships. Descriptive mining tasks characterize the general properties of the data and represent it in a meaningful way. Classification and regression models are predictive and clustering and association rules models are descriptive [12]. Predictive data mining techniques are Decision Tree, Neural Network and K-Nearest Neighbor Methods.

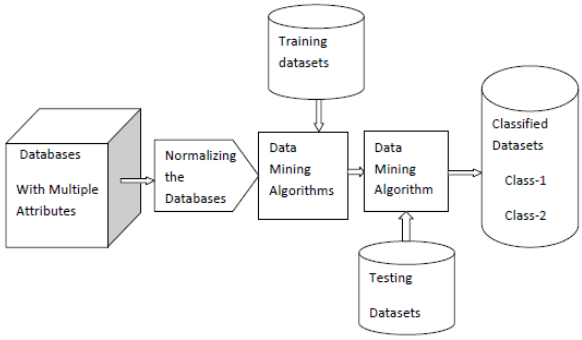

Figure 5.1 Data Mining Model

EMPIRICAL DATA COLLECTION

Heart Disease Data sets are available on UCI machine learning repository, centre for machine learning and intelligent systems. The data used for the implementation of proposed data mining techniques is downloaded from the Statlog dataset UCI Machine learning repository. The datasets contain the data of 810 patients collected from UCI machine learning repositories. The Statlog database has 76 raw attributes However; all of the published experiments only referred 13 of them. These 13 attributes are shown in the table 5.1 [15]. Table 5.1 shows the detailed description of Statlog Heart Disease database.

DATA MINING TECHNIQUES USED FOR CLASSIFICATION

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection, Disease Diagnosis and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In machine learning algorithms the training data are analyzed by the classification algorithms. In classification of data sets the test data are used to estimate the accuracy of the classification algorithm. If the accuracy of training stage of the model is acceptable than test datasets are used for the testing of the model and finally validation process is used for the actual performance of the model.

For the classification of Statlog Heart Disease Datasets following data mining techniques are mainly used.

1. Decision Trees 2. Artificial Neural Networks

Table 5.1 Statlog Heart Disease Database Attributes

S.NO.	NAME	DESCRIPTION OF ATTRIBUTE
1	Age	age in years
2	Sex	1 = male ; 0 = female
3	Chest pain type (4 values)	chest pain type (1 = typical angina; 2 = atypical angina ; 3 = non-anginal pain; 4 = asymptomatic)
4	Resting blood pressure	resting blood pressure (in mm Hg on admission to the hospital)
5	Serum cholesterol in mg/dl	serum cholesterol in mg/dl
6	Fasting blood sugar > 120 mg/dl	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	Resting electrocardiographic results (values 0, 1, 2)	resting electrocardiographic results (0 =normal; 1 = having ST-T wave abnormality; 2 = showing probable or define left ventricular hypertrophy by Estes' criteria)
8	Maximum heart rate achieved	maximum heart rate achieved
9	Exercise induced angina	exercise induced angina (1 = yes; 0 = no)
10	Oldpeak = ST depression induced by exercise relative to rest	ST depression induced by exercise relative to rest
11	The slope of the peak exercise ST segment	the slope of the peak exercise ST segment (1 = up sloping; 2 = flat ; 3= down sloping)
12	Number of major vessels (0-3) colored by fluoroscopy	number of major vessels (0-3) colored by fluoroscopy
13	Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect	(3 = normal; 6 = fixed defect; 7 = 19 eversible defect)

DECISION TREES

Decision tree is a tree-shaped structure that represents sets of decisions. These decisions generate rules for the classification of a dataset. Decision Tree is a popular classifier which is
simple and easy to implement [16]. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes.

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes are called leaves (also known as terminal or decision nodes).

BASIC DECISION TREE BUILDING ALGORITHM

Algorithm Learn Decision Tree (examples, attributes, default) returns a decision tree
inputs: examples, a set of examples
attributes, a set of attributes
default, default value for goal attribute
if examples is empty then return default
else if all examples have same value for goal attribute then return value
else
best = ChooseAttribute(attributes, examples)
tree = a new decision tree with root test best
for each value vi of best do
examplesi = {elements of examples with best = vi}
subtree = LearnDecisionTree(examples, attributes – best,
MajorityValue (examples))
add a branch to tree with label vi and subtree , return tree.

If we analyze the algorithm with m be the number of attributes, n be the number of instances than the Depth of tree will be O(log n) and for each level of the tree all n instances are considered (best = vi) than O(n log n) work for a single attribute over the entire tree. The total cost will be O(mn log n) since all attributes are eventually considered**.** With total number of m attribute the possible number of tree will be 2^{2m}.

INFORMATION GAIN

Measures how well a given attribute separates the training examples according to their target classification. This measure is used to select among the candidate attributes at each step while growing the tree. The entropy (Information Gain) approach selects the splitting attribute that minimizes the value of entropy, thus maximizing the Information Gain [13]. To identify the splitting attribute of the Decision Tree, one must calculate the Information Gain for each attribute and then select the attribute that maximizes the Information Gain.

The entropy for each attribute is calculated using the Following formula

$$E(S) = - p(P)\log2\ p(P) – p(N)\log2\ p(N) \qquad (5.1)$$

Where p(P) is the probability of positive responses, and p(N) is the probability of the negative responses. Information gain measures the expected reduction in entropy, or uncertainty.

CONFUSION MATRIX FOR PERFORMANCE

In the field of machine learning, a confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, typically used in supervised learning algorithms to show the predicted results (in unsupervised learning it is usually called a matching matrix) [18]. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another). It is an important performance matrix used in machine learning.

In machine learning, the confusion matrix is often called the contingency table or the error matrix.

Table 5.2 A Confusion Matrix for Prediction [11]

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANN's) have been used widely in many application areas in recent years. Most applications use feedforward ANN's and the backpropagation (BP) training algorithm. There are numerous variants of the classical BP algorithm and other training algorithms. All these training algorithms assume a fixed ANN architecture. They only are training weights in the fixed architecture that includes both connectivity and node transfer functions [3].

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the biological nervous systems, such as the brain, processes information [6]. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected data processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process or training process.

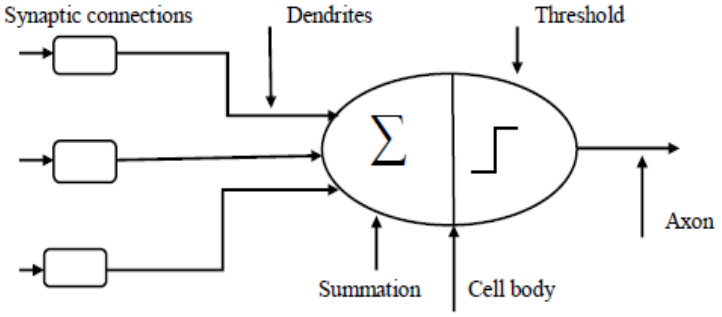

Figure 5.1: simple Neuron Model

VI. RESULTS AND CONCLUSION

RESULTS GENERATED USING DECISION TREE DATA MINING TECHNIQUE

Classification and Regression Tree algorithm is implemented using IMB SPSS 20 Tool boxes for the classification of Statlog heart disease datasets. In this classification Numerical Values of the predictors are used and the response values in the both cases are Numeric and classified in to two categories as 1 and 2.

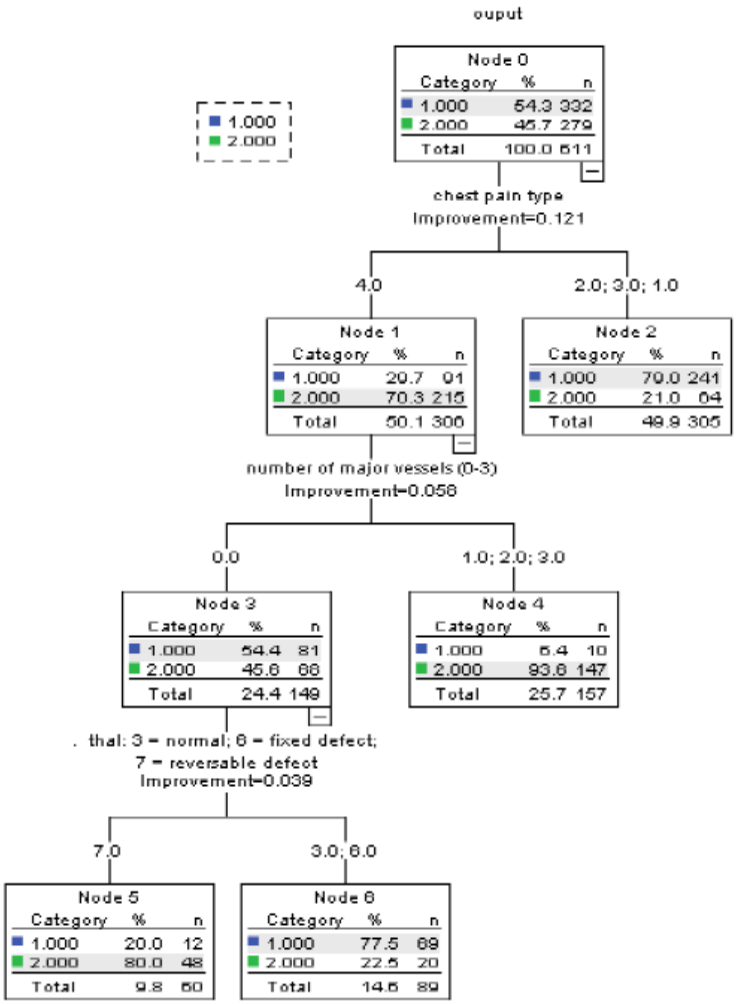

Figure 6.1 Decision Tree for Training data Sets

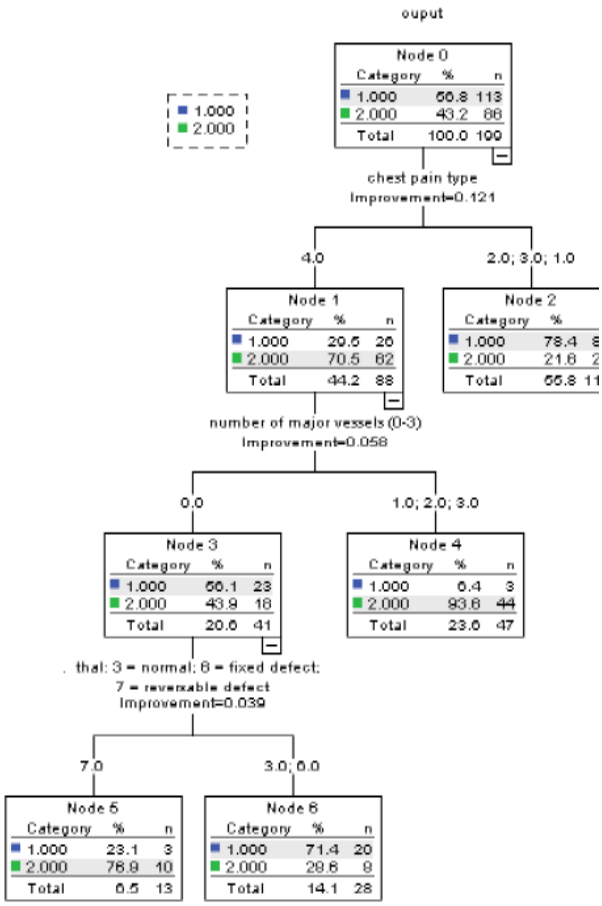

Figure 6.2 Decision Tree for Test Dataset

Table 6.1 Risk Factors for the Classification

Sample	Estimate	Std. Error
Training	.173	.015
Test	.191	.028

Growing Method: CRT and Dependent Variable: output

Table 6.2 Classification of Statlog Datasets

Sample	Observed	Predicted		
		1	2	Percent Correct
Training	1	310	22	93.4%
	2	84	195	69.9%
	Overall Percentage	64.5%	35.5%	82.7%
Test	1	107	6	94.7%
	2	32	54	62.8%
	Overall Percentage	69.8%	30.2%	80.9%

Growing Method: CRT and Dependent Variable: output

NEURAL NETWORK FOR THE CLASSIFICATION OF STATLOG HEART DISEASE DATABASE

Neural Network for the classification of Statlog Heart disease database is shown in the Figure 6.3 is displayed with help of IBM SPSS 20 data mining tool. It shows the input layer with 13 values of input data with bias and output generated on the basis of classification categories as 1 and 2 as the indicating the absence of disease and the presence of disease respectively.

Table 6.3 Case Processing Summary

		N	Percent
Sample	Training	551	69.2%
	Testing	245	30.8%
Valid		796	100.0%
Excluded		14	
Total		810	

The case processing summery shows the data partitioned in training, testing and validation process of the model. It shows that 69.2% data are used for the training of the model and 30.8%, for the testing of the model. 14 data records are excluded during the validation of the model.

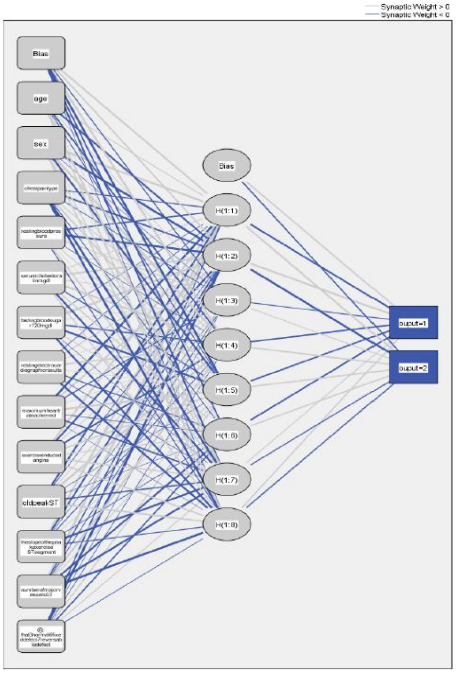

Figure 6.3 Neural Networks for Statlog Datasets

Table 6.4 Model Summary

Training	Cross Entropy Error	26.356
	Percent Incorrect Predictions	0.7%
	Stopping Rule Used	1 consecutive step(s) with no decrease in error[a]
	Training Time	0:00:00.16
Testing	Cross Entropy Error	36.796
	Percent Incorrect Predictions	2.4%

Dependent Variable: output

a. Error Computations use both the training and testing samples.

The model summary Table 6.4 shows the different parameters of the model with cross entropy error, incorrect prediction stopping rules and training time

Table 6.5 Classification of Statlog Database

Sample	Observed	Predicted		
		1	2	Percent Correct
Training	1	302	0	100.0%
	2	4	245	98.4%
	Overall Percent	55.5%	44.5%	99.3%
Testing	1	131	2	98.5%
	2	4	108	96.4%
	Overall Percent	55.1%	44.9%	97.6%

Dependent Variable: output

The Sensitivity and the Specificity are statistical measures of the performance of a binary classification test, also known in Statistics as classification functions.

$$Sensitivity = \frac{\text{Number of True Positives}}{\text{Number of Positives}}$$

It shows the probability of positive test as presence of disease.

$$Specificity = \frac{\text{Number of True Nagatives}}{\text{Number of Nagatives}}$$

It shows the probability of a negative test, as absence of the Disease.

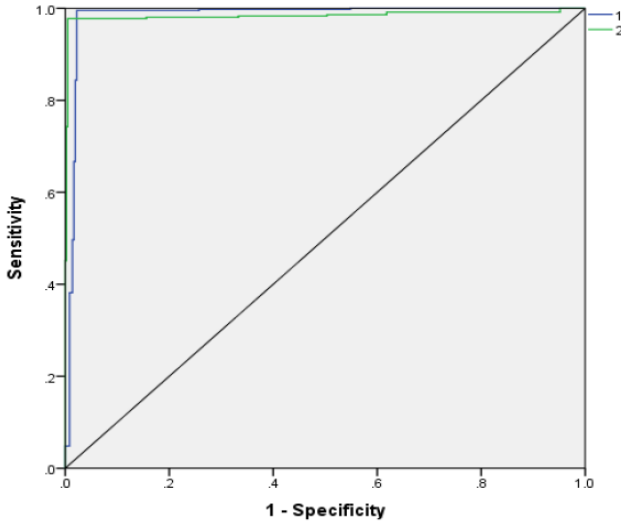

Figure 6.4 Sensitivity and Specificity Graph

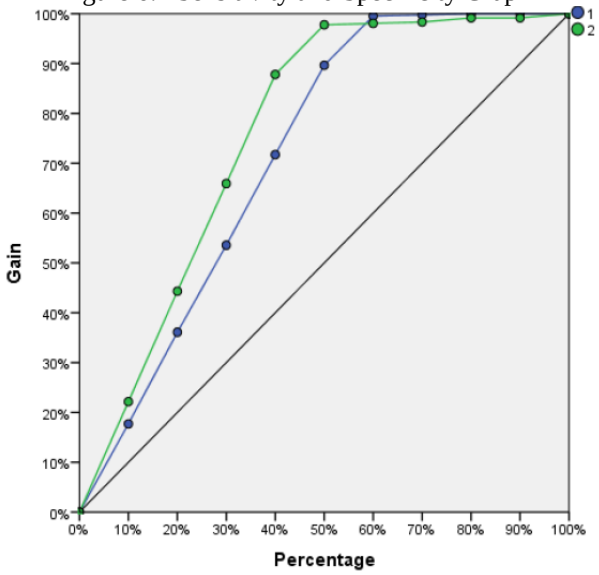

Figure 6.5.Percentage and Gain for the Classification

Gains chart shows the percentage of the overall number of cases in a given category "gained" by targeting a percentage of the total number of cases. For example, the first point on the curve for the presence of disease category is approximately at (10%, 20.2%). The diagonal line is the "baseline" curve;
if we select 10% of the cases from the scored dataset at random, we would expect to "gain" approximately 10.8% and 20.2% of all of the cases that actually take any given category.

VII. CONCLUSIONS

Machine learning techniques are more efficient and accurate. It has been shown through the Classification of Statlog Heart Disease datasets that Neural Networks and Decision tree machine learning algorithm are can be applied for classification of medical databases as well as other databases for prediction of hidden patterns.

REFERENCES

[1] Mai Shouman, Tim Turner and Rob Stocker, "*Using Data Mining Techniques in Heart Disease Diagnosis and Treatment*", Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), pp. 173-177, 2012.

[2] Ling Kock Sheng and Teh Ying Wah , "*A Comparative Study of data Mining Techniques in Predicting Consumers*' Credit Card Risk in Banks", African Journal of Business management Vol. 5(20), pp. 8307-8312, 2011.

[3] Benish Fida, Muhammad Nazir, Nawazish Naveed and Sheeraz Akram, "*Heart Disease Classification Ensemble Optimization Using Genetic Algorithm*", IEEE 14th International Multitopic Conference, (INMIC), pp.19 –24, 2011.

[4] Wei-Pin, Changa,b Der-Ming, and Liouc,d, "*Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data*", Journal of Telemedicine and Telecare , pp.1-26, 2008.

[5] Henley and Hand, "*Statistical Classification Methods in Consumers Credit Scoring: A Review*", Journal of the Royal Statistical Society: Series A (Statistics in Society),Vol. 3, pp. 523-541, 1997.

[6] Yavuz Del_Can, Lale Özyilmaz and Tülay Yildirim, "*Evolutionary Algorithms Based RBF Neural Networks For Parkinson's disease Diagnosis*", 7th International Conference on Electrical and Electronics Engineering (ELECO), pp. 311-315, 2011.

[7] Mai Shouman, Tim Turner, and Rob Stocker, "*Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients*" International Journal of Information and Education Technology, Vol. 2, No. 3, pp.220-223, 2012.

[8] Berardi, Patuwo and Hu, "*A Principled Approach for Building and Evaluating Neural Network Classification Models*", Decision Support Systems Vol. 38, Issue 2, pp. 233-246, 2004.

[9] Qeethara Kadhim Al-Shayea , "*Artificial Neural Networks in Medical Diagnosis*", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 2, pp.150-154, 2011.

[10] Kemal Hakan, Gulkesen, İsmail Turker Koksal, Sebahat Ozdem,and Osman Saka, "*Prediction of Prostate Cancer Using Decision Tree Algorithm*", Turkish Journal of Medical Sciences,Vol. 40(5), pp. 681-686, 2010.

[11] Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka and Lampros K. Michalis, "*Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling*", IEEE Transactions on Information Technology in Biomedicine, Vol. 12, No. 4, pp. 447-457, 2008.

[12] Chazard E, Ficheur G, Bernonville S, Luyckx M and Beuscart R, "*Data Mining to Generate Adverse Drug Events Detection Rules*", IEEE Transaction on Informatin Technology in Biomedicine, Vol. 15, No. 6, pp. 823-830, 2011.

[13] Mu Zhua, Wenhong Chena, John P. Hirdesb and Paul Stoleeb, "*The K-nearest Neighbor Algorithm Predicted Rehabilitation Potential Better Than Current Clinical Assessment Protocol*", Journal of Clinical Epidemiology, pp. 1015-1021, 2010.

[14] Serhat Özekes and A.Yilmaz Çamurcu, "*Classification and Prediction in a Data Mining Application*", Journal of Marmara for Pure and Applied Sciences, Vol.18, pp. 159-174, 2002.

[15] Ishtake , and Sanap, "*Intelligent Heart Disease Prediction System Using Data Mining Techniques*", International J. of Healthcare & Biomedical Research, Vol.1, Issue 3, pp. 94-101, 2013.

[16] T. Smitha, and V. Sundaram, "*Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis*", International Journal of Advances in Engineering & Technology, Vol.4, Issue 2, pp. 173-178, 2012.

[17] S. M. Kamruzzaman, Ahmed Ryadh Hasan, Abu Bakar Siddiquee and Md. Ehsanul Hoque ,Mazumder ,"*Medical Diagnosis Using Neural Network*", Proc. 3rd International Conference on Electrical & Computer Engineering (ICECE), pp. 537-540, 2004.

[18] Antonia Vlahou, , John O. Schorge, Betsy W. Gregory,and Robert L. Coleman. "*Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data*", Journal of Biomedicine and Biotechnology, Vol. 5, Issue 5, pp. 308314, 2003.

