**CHAPTER 1**

1. **INTRODUCTION**

**1.1 OVERVIEW OF DATA MINING**

Data mining, the extractionof thehiddenpredictiveinformationfromlarge

databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyse massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?" This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today’s business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

* **Automated prediction of trends and behaviours**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
* **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, they can analyse massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. High speed makes it practical for users to analyse huge quantities of data. Larger databases, in turn, yield improved predictions.

Databases can be larger in both depth and breadth:

* **More columns**. Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.
* **More rows**. Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

**1.2 DATAMINING TECHNIQUES**

* **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
* **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
* **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
* **Nearest neighbour method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbour technique.
* **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

**1.3 DATA MINING PROCESSES**

**Data understanding**

* First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data. Some important activities must be performed including data load and data integration in order to make the data collection successfully.
* Next, the “gross” or “surface” properties of acquired data needs to be examined carefully and reported.
* Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting and visualization.
* Finally, the data quality must be examined by answering some important questions such as “Is the acquired data complete?”, “Is there any missing values in the acquired data?”

**Data preparation**

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

**Modelling**

* First, modelling techniques have to be selected to be used for the prepared dataset.
* Next, the test scenario must be generated to validate the quality and validity of the model.
* Then, one or more models are created by running the modelling tool on the prepared dataset.
* Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

**Evaluation**

In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that has been discovered in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

**Deployment**

The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it. Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization. In the deployment phase, the plans for deployment, maintenance and monitoring have to be created for implementation and also future supports. From the project point of view, the final report of the project needs to summary the project experiences and review the project to see what need to improved created learned lessons.

**1.4 TASKS OF DATA MINING**

Data mining as a term used for the specific set of six activities or tasks as follows:

1. Classification

2. Estimation

3. Prediction

4. Affinity grouping or association rules

5. Clustering

6. Description and visualization

**Classification**

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of pre-classified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:

* Classification of credit applicants as low, medium or high risk
* Classification of mushrooms as edible or poisonous
* Determination of which home telephone lines are used for internet access

**Estimation**

Estimation deals with continuously valued outcomes. Given some input data, we

use estimation to come up with a value for some unknown continuous variables

such as income, height or credit card balance. Some examples of estimation tasks

include:

* Estimating the number of children in a family from the input data of mothers’

Education.

* Estimating total household income of a family from the data of vehicles in the

Family.

* Estimating the value of a piece of a real estate from the data on proximity of

that land from a major business centre of the city.

**Prediction**

Any prediction can be thought of as classification or estimation. The difference is

one of emphasis. When data mining is used to classify a phone line as primarily

used for internet access or a credit card transaction as fraudulent, we do not expect

to be able to go back later to see if the classification was correct. Our classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit cards transaction is or is not fraudulent. With enough efforts, it is possible to check. Predictive tasks feel different because the records are classified according to some predicted future behaviour or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include:

* Predicting the size of the balance that will be transferred if a credit card

prospect accepts a balance transfer offer.

* Predicting which customers will leave within next six months.
* Predicting which telephone subscribers will order a value–added service such

as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be

adopted for use in prediction by using training examples where the value of the

variable to be predicted is already known, along with historical data for those

examples. The historical data is used to build a model that explains the current

observed behaviour. When this model is applied to current inputs, the result is a

prediction of future behaviour.

**Association Rules**

An association rule is a rule which implies certain association relationships among

a set of objects (such as “occur together” or “one implies the other”) in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

**Clustering**

Clustering is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes. In clustering, there are no predefined classes. The records are grouped together on the basis of self-similarity. Clustering is often done as a prelude to some other form of data mining or modelling. For example, clustering might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster.

**Description and Visualization**

Data visualization is a powerful form of descriptive data mining. It is not always

easy to come up with meaningful visualizations, but the right picture really can be

worth a thousand association rules since the human beings are extremely practiced at extracting meaning from visual scene

**CHAPTER 2**

**LITERATURE SURVEY**

1. **TITLE: Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction**

**AUTHOR:** Md. Tahmid Rahman Laskar, Md. Tahmid Hossain, Nafiul Rashid

According to this paper an Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable. The System can discover and extract hidden knowledge associated with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can answer complex queries for diagnosing disease and thus, assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs.

**Methodology:**

1. Naive Bayes
2. Neural Network
3. Decision Trees

**2. TITLE:** **DISEASE PREDICTING SYSTEM USING DATA MINING TECHNIQUES**

**AUTHOR:** M.A. Nishara Banu, B Gomathy

This paper deals with successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. The medical environment is still information rich but knowledge weak. There is a wealth of data possible within the medical systems. However, there is a lack of powerful analysis tools to identify hidden relationships and trends in data. Heart disease is a term that assigns to a large number of heath care conditions related to heart. These medical conditions describe the unexpected health conditions that directly control the heart and all its parts. Medical data mining techniques like association rule mining, classification, clustering is implemented to analyze the different kinds of heart-based problems. Classification is an important problem in data mining. Given a database contain collection of records, each with a single class label, a classifier performs a brief and clear definition for each class that can be used to classify successive records. A number of popular classifiers construct decision trees to generate class models. The data classification is based on MAFIA algorithms which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree. The heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database.

**METHODOLOGY:**

1. MAFIA Algorithm
2. C4.5 Algorithm
3. K-Means Clustering Algorithm

**3.TITLE: Efficient Heart Disease Prediction System**

**AUTHOR:** Purushottam, Prof. (Dr.) Kanak Saxena, & Richa Sharma

This paper deals with a framework that can proficiently find the tenets to foresee the risk level of patients in view of the given parameter about their health. The main contribution of this study is to help a non-specialized doctor to make correct decision about the heart disease risk level. The rules generated by the proposed system are prioritized as Original Rules, Pruned Rules, Rules without duplicates, Classified Rules, Sorted Rules and Polish. The execution of the framework is assessed as far as arrangement precision and the outcomes demonstrates that the framework has extraordinary potential in anticipating the coronary illness risk level all the more precisely.

**METHODOLOGY:**

1. SVM Algorithm
2. C4.5 Algorithm
3. MLP Algorithm
4. **TITLE: Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review**

**AUTHOR:** Animesh Hazra & Amit Gupta

This paper summarizes some of the current research on predicting heart diseases using data mining techniques, analyse the various combinations of mining algorithms used and conclude which technique(s) are effective and efficient. Also, some future directions on prediction systems have been addressed. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death.

**METHODOLOGY:**

1. Decision Trees
2. K-Means Clustering
3. C4.5 Algorithm
4. **TITLE: Early Prediction of Heart Diseases Using Data Mining Techniques**

**AUTHOR:** Vikas Chaurasia

The main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease survivability. We used three popular data mining algorithms CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and decision table (DT) extracted from a decision tree or rule-based classifier to develop the prediction models using a large dataset. We also used 10-fold cross validation methods to measure the unbiased estimate. This paper proposes data classification is based on MAFIA algorithms which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree. The heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database

**METHODOLOGY:**

1. ID3 Algorithm
2. K-Means Clustering
3. C4.5 Algorithm

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

The raw diabetic big data or data set is given as input to the system. The unstructured voluminous input data can be obtained from various Electronic Health Record (EHR) / Patient Health Record (PHR), Clinical systems and external sources (government sources, laboratories, pharmacies, insurance companies etc.). Algorithms are Naive bayes classification, SVM algorithm, J48 Decision tree.

To explore Big Data, proposed system analysed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies.

**DISADVANTAGES OF EXISTING SYSTEM**

There are only few decision support systems available in medical industry whose functionalities are very limited. As mentioned earlier, medical decisions are made with doctor’s intuition and not from the rich data from the medical database. Wrong treatment due to misdiagnosis causes serious threat in medical field. In order to solve these issues data mining solution was with help of medical databases was introduced. Apart from the benefits like probabilistic approaches and fast reliable algorithm of Naive Bayes, the serious shortcoming of the algorithm is its ability in handling small datasets. naive Bayes classifier requires relatively large dataset to obtain best results. Yet, studies showed that Naive Bayes algorithm outperforms other algorithms in accuracy and efficiency. Notable limitation of this paper is the usage of small dataset. This dataset can be used for training or testing purpose only. Also the dataset could include more attributes for a more effective prediction in supporting clinical decisions.

**3.2 PROPOSED SYSTEM**

The Proposed system consists of the following modules:

**Module – 1:** - Registration and Data collection

* The patient needs to have his/her account. For that the patient needs to enter his personal details and create an account.
* The patient will have to fill the forms containing questionnaires. After the completion of forms, the proposed system database will be updated with the patient records.

**Module – 2:** - Finding the Probability

* After the collection of the data entered by the patient, the proposed system will generate the probability whether the patient is in risk of a heart disease or not.

**Module – 3:** - Generating the Remedy

* The symptoms will be entered by the patients and depending upon the probability and the symptoms, the remedy will be generated and will be displayed.

In this study, a prototype of a system that comprises of a binary classification model to predict the risk factor of an individual based on his/her medical data is proposed. The system is well equipped with a comprehensive graphic user interface that is easy to use and understand. The classification follows a supervised learning wherein the dataset used was obtained from University of California, Irvine's machine learning repository. The reference is limited to the Cleveland's dataset which was collected as unstructured data in the form of medical reports and converted to a structured dataset. The dataset represents a binary classification problem. The dataset comprises of 14 attributes in total, out of which 13 are predictor variables and one feature is a binary response variable. Therefore, the dataset represents a binary classification problem. Naïve Bayes was used for the classification process. In the end, an easy to use web interface was developed so that the system can be used by persons with little or no technical knowledge, thereby completely abstracting the core functionality and implementation details of the system.

**Advantages of proposed system**:

1. Better result accuracy.
2. Reduced time complexity.

**3.3 Architecture of the proposed System**

The proposed system is done as Java Application. The patients need to answer various questions and the data is collected and then stored in the database. The data is classified based on the different possibilities. The patient can

check whether he is prone to heart attack or not. The data entered by the patient is considered and the probability is generated whether the patient is in risk of getting the heart attack or not. Further questions are asked to the patient and depending upon the answers, remedy is generated. The medicine corresponding to a particular symptom is displayed. The doctors can view the patient details as well as the medicine database.

The proposed system is divided into two parts

* Data collection
* Probability generation
* Generation of remedies ()

The heart dataset contains large volumes of which consumes more time for classification so by using attribute selection methods the dimensionality of data is reduced. In both cases Naïve Bayes classification technique produced enhanced results. This is observed when the performance of four classification algorithms: Naïve Bayes, Decision Tree, K-NN and Neural Network are investigated on complete heart disease dataset and reduced dataset. A Web based clinical decision support system which uses medical profiles like age, blood pressure, *etc*. is proposed to predict the prospect of patients attaining heart disease. In supervised learning the network computes a response to each input and then compares it with the target value. If the computed response differs from the target value, the weights of the network are adapted according to a learning rule. Examples of supervised learning are Single-layer perceptron and Multi-layer perceptron. In unsupervised learning the networks learn by identifying special features in the problems they are exposed to

**3.4Analysis**

Analysis on classification algorithms. Although it is a simple method, Naïve Bayes can outperform more sophisticated classification methods. In this experiment, Naïve Bayes outperforms Decision Tree and K-Nearest Neighbour. The reason for Naïve Bayes good performance is not because there are no attribute dependences in the data. In fact, its good performance is caused by the zero-one loss function used in the classification.

**Methodology**

The Heart Disease Prediction System (HDPS) is done using Naïve Bayes Data mining technique Record set with medical attributes was obtained from the Cleveland Heart Disease database. With the help of the dataset, the patterns significant to the heart attack prediction are extracted. Questionnaires are also sharply limited by the fact that respondents must be able to read the questions and respond to them.

The records are split equally into two datasets:

* Training dataset.
* Testing dataset.

The attribute “Diagnosis” is identified as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease. “PatientId” is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

**3.5 Attributes of Heart Data Sets**

Serial No. Attribute Description

**1.** **Sex value** 1: Male

value 0: Female

**2**. **Chest Pain Type**

value 1: typical type 1 angina

value 2: typical type angina

value 3: non-angina pain

value 4: asymptomatic

**3.** Fasting Blood Sugar value 1: > 120 mg/dl

value 0:< 120 mg/dl

**4.** **Rest ECG**

resting electrographic results

value 0: normal

value 1: 1 having ST-T wave abnormality

value 2: showing probable/definite left ventricular hypertrophy

**5**. **Exang**

exercise induced angina

value 1: yes

value 0: no

**6.** **Slope**

the slope of the peak exercise ST segment

value 1: unsloping

value 2: flat

value 3: down sloping

**7.** CA number of major vessels coloured by fluoroscopy (value 0 – 3)

**8.** **Thal**

value 3: normal

value 6: fixed defect

value 7: reversible defect

**9.** **Trest** **Blood Pressure** (mm Hg on admission to the hospital)

**10.** **Serum Cholesterol** (mg/dl)

**11.** **Thalach** maximum heart rate achieved

**12.** **Old peak** ST depression induced by exercise

**13.** **Age** In year

**14.** **Height** In cms

**15.** **Weight** In kgs

**Bayes Rule**

A conditional probability is the likelihood of some conclusion, *C*, given some evidence/observation, *E*, where a

dependence relationship exists between *C* and *E*.

This probability is denoted as P *(C* |*E)* where ......... (I)

For proposition A and evidence B

* P(A), the prior, is the initial degree of belief in A.
* P (A | B), the posterior, is the degree of belief having accounted for B.
* P (B | A) / P(B) represents the support B provides for A [1].

**Variables**

Each attribute is a potential risk factor. There are both demographic, behavioural and medical risk factors.

* **Demographic:** sex: male or female;(Nominal)
* age: age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
* **Behavioural**
* current Smoker: whether or not the patient is a current smoker (Nominal)
* cigsPerDay: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

**Medical(history):**

* BPMeds: whether or not the patient was on blood pressure medication (Nominal)
* prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
* prevalentHyp: whether or not the patient was hypertensive (Nominal)
* diabetes: whether or not the patient had diabetes (Nominal)

**Medical(current):**

* tot Chol: total cholesterol level (Continuous)
* sys BP: systolic blood pressure (Continuous)
* dia BP: diastolic blood pressure (Continuous)
* BMI: Body Mass Index (Continuous)
* heartrate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
* glucose: glucose level (Continuous)

**Predict variable (desired target):**

* 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

**3.6 ALGORITHMS IN PROPOSED SYSTEM**

**RANDOM FOREST ALGORITHM:**

Decision trees involve the greedy selection of the best split point from the dataset at each step. This algorithm makes decision trees susceptible to high variance if they are not pruned. This high variance can be harnessed and reduced by creating multiple trees with different samples of the training dataset (different views of the problem) and combining their predictions. This approach is called bootstrap aggregation or bagging for short. A limitation of bagging is that the same greedy algorithm is used to create each tree, meaning that it is likely that the same or very similar split points will be chosen in each tree making the different trees very similar (trees will be correlated). This, in turn, makes their predictions similar, mitigating the variance originally sought. We can force the decision trees to be different by limiting the features (rows) that the greedy algorithm can evaluate at each split point when creating the tree. This is called the Random Forest algorithm. Like bagging, multiple samples of the training dataset are taken and a different tree trained on each. The difference is that at each point a split is made in the data and added to the tree, only a fixed subset of attributes can be considered. For classification problems, the type of problems we will look at in this tutorial, the number of attributes to be considered for the split is limited to the square root of number of input features. The result of this one small change are trees that are more different from each other (uncorrelated) resulting predictions that are more diverse and a combined prediction that often has better performance that single tree or bagging alone.

**NAIVE BAYES ALGORITHM:**

Naïve Bayes is a classification algorithm based on Bayes theorem, which calculates a probability by counting the frequency of values and combination of values n historical data. Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred.

Pro (B given A) = Pro (A and B)/Pro(A)

Advantage of this algorithm is it requires only a small amount of training data for estimating the parameters essential for classification.

The Naive Thomas Bayes rule relies on Bayesian theorem as given by on top of equation Steps in rule area unit as follows:

1. Every information sample is diagrammatical by associate n dimensional feature vector, X= (x1, x2. xn), portraying n measurements created on the sample from n attributes, severally A1, A2, An.
2. Suppose that there is a unit m category, C1, C2Cm. Given associate unknown information sample, X (i.e., having no category label), the category |the category iffier can predict that X belongs to the class having the very best posterior chance, conditioned if and solely if:

P(Ci/X) &gt;P(Cj/X) for all 1&lt;=j&lt;=m and j!=I therefore we tend to maximize P(Ci|X). The category Ci that P(Ci|X) is maximized is called the most posteriori hypothesis. By Thomas Bayes theorem.

1. As P(X) is constant for all categories, solely P(X|Ci) P(Ci) want be maximized. If the category previous chances don't look like to be proverbial, then it's ordinarily considered that the categories area unit equally possible, i.e. P(C1) =P(C2) =. =P(Cm), and we would thus maximize P(X|Ci). Otherwise, we tend to maximize P(X|Ci) P(Ci).

Note that the category previous chances could also be calculable by P(Ci)=si/s, where Si is that the range of coaching samples of sophistication Ci, and s is that the total range of coaching samples, On X. That is, the naive chance assigns associate unknown sample X to the category Ci.

**SVM CLASSIFIER ALGORITHM:**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning*),* the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. In addition to performing [linear classification](https://en.wikipedia.org/wiki/Linear_classifier), SVMs can efficiently perform a non-linear classification using what is called the [kernel trick](https://en.wikipedia.org/wiki/Kernel_trick), implicitly mapping their inputs into high-dimensional feature spaces. In the case of support-vector machines, a data point is viewed as a {\displaystyle p}two-dimensional vector (a list of {\displaystyle p} numbers), and we want to know whether we can separate such points with a {\displaystyle (p-1)}-dimensional [hyperplane](https://en.wikipedia.org/wiki/Hyperplane). This is called a [linear classifier](https://en.wikipedia.org/wiki/Linear_classifier). There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or [margin](https://en.wikipedia.org/wiki/Margin_(machine_learning)), between the two classes. So, we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized.

**KNN ALGORITHM:**

In [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), the k-nearest neighbors algorithm (k-NN) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis).[[1]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-1)In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether *k*-NN is used for classification or regression:

* In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
* In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

k-NN is a type of [instance-based learning](https://en.wikipedia.org/wiki/Instance-based_learning), or [lazy learning](https://en.wikipedia.org/wiki/Lazy_learning), where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known.

**CHAPTER 4**

**SYSTEM DESIGN**

**4.1 SYSTEM ARCHITECTURE**

Training Data

Testing Data

Implementation of algorithms

Naive-bayes

SVM

KNN

Figure

Random Forest

Comparison of accuracy

Result

DATASET

Data cleaning

Data splitting

**Figure 1**

**Disease Prediction System Architecture**

There are three tiers in the disease prediction system mining architecture:

1. **Data layer**: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.

2. **Application layer** is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.

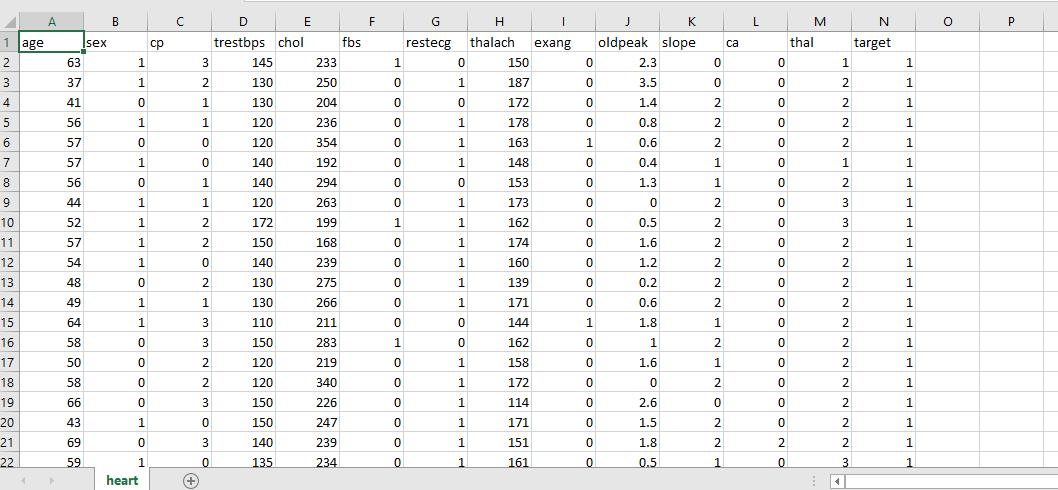
3. **Front-end layer** provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

**Data Mining Applications in Medicine**

* Data mining enables to characterize patient activities to see incoming office visits.
* Data mining helps identify the patterns of successful medical therapies for different illnesses.

Data mining applications are continuously developing in various industries to provide more hidden knowledge that increases business efficiency and grows businesses.

**4.2 DATASETS USED**



**CHAPTER 5**

**SYSTEM REQUIREMENTS**

**5.1 Hardware and software details**

**Software Required**

* Anaconda

**Hardware Required**

* Server to connect to the database
* 4 GB Ram
* Intel core i3 processor

**HARDWARE REQUIREMENTS**

|  |  |
| --- | --- |
| * CPU SPEED * CPU TYPE | : 3.1 GHz  : 64 bit |
| * STORAGE | : 1 TB |
| * OS | : Microsoft Windows 10 64 bit |
| * RAM | : 8 GB |

**5.2 ABOUT THE SOFTWARE**

**ANACONDA NAVIGATOR**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS and Linux. Anaconda® is a package manager, an environment manager, a Python/R data science distribution, and a collection of over 1500+ open source packages. Anaconda is free and easy to install, and it offers free community support. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window.

**CONDA** Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux. Conda quickly installs, runs and updates packages and their dependencies. Conda easily creates, saves, loads and switches between environments on your local computer. It was created for Python programs, but it can package and distribute software for any language.

**APPLICATIONS AVAILABLE IN NAVIGATOR**

* Jupyter Lab
* Jupyter Notebook
* QT Console
* Spyder
* VS Code

**JUPYTER NOTEBOOK**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

**5.3 PYTHON**

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions- **Python 2 and Python 3**. Both are quite different. Some of the advantages of using python:

* Emphasis on code readability, shorter codes, ease of writing
* Programmers can express logical concepts in fewer linesof code in comparison to languages such as C++ or Java.
* Python supports multipl**e** programming paradigms, like object-oriented, imperative and functional programming or procedural.
* There exist inbuilt functions for almost all of the frequently used concepts.

**FEATURES**

* Interpreted
* Platform Independent
* Free and open source, Redistributable
* Embeddable
* Robust
* Rich Library support

**DATA ANALYSIS IN PYTHON**

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. Panda is one of those packages, and makes importing and analyzing data much easier.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORK**

Many DSS exists to predict the heart disease with various techniques. The World life expectancy statistics implies that heart disease is prevailing more in number. So, it is necessary to build an efficient intelligent trusted automated system which predicts the heart disease accurately based on the symptoms according to gender/age and domain knowledge of experts in the field at the lowest cost. It is concluded that data mining is the approach which extract useful information from the rough data. The prediction analysis is the technique which predicts future from the current information. In this paper, various techniques of prediction analysis are reviewed in terms of description and outcomes. Decision Support in Heart Disease Prediction System is a Java Application which is intended to be done using Data mining technique “Naive Bayesian Classification technique”. The system extracts hidden knowledge from the database. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. In this research paper, we have presented an Efficient Heart Disease Prediction System using data mining. This system can help medical practitioner in efficient decision making based on the given parameter. We have train and test the system using 10-fold method and find the accuracy of 86.3 % in testing phase and 87.3 % in training phase and because this model demonstrates the better results and helps the area specialists and even individual related with the field to get ready for a superior determine and give the patient to have early determination results as it performs sensibly well even without retraining.

**APPENDICES**

**APPENDIX I**

**SAMPLE CODE**

**Importing Data**

Import pandas as pd

Import numpy as NP

Import matplotlib.pieplot as PLT

Import Seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score

**Reading CSV File**

df=pd.read\_csv('heart.csv')

df.head()

**Visualizing the Data**

df.shape

df.info ()

sns.pairplot(df\_cancer,hue='target',vars=['mean radius','mean texture','mean perimeter','mean area', 'mean smoothness'])

plt.figure(figsize=(20,9))

sns.heatmap(df\_cancer.corr(),annot=True)

**Data Preprocessing**

cancer\_mapping={'Benign':0,'Malign':1}

df.Diagnosis=df.Diagnosis.map(cancer\_mapping)

df.isnull().any()

df.duplicated()

**Features Selection**

y=df['Diagnosis']

x=df[['Mean Radius','Mean Texture','Mean Perimeter','Mean Area','Mean Smoothness','Mean Compactness','Mean Concavity','Mean Concave points','Mean Symmetry','Mean Fractal dimension']

**Data splitting**

from sklearn.model\_selection import train\_test\_split

x\_train, x\_test, y\_train, y\_test=train\_test\_split(x, y, test\_size=0.2)

**Implement Algorithm**

**Support Vector Machine**

from sklearn.svm import SVC

model\_all=SVC ()

model\_all.fit(x\_train, y\_train)

y\_pred=model\_all.predict(x\_test)

y\_pred.shape

confusion\_matrix(y\_test,y\_pred)

accuracy\_score(y\_test,y\_pred)

**KNN**

from sklearn.neighbors import KNeighborsClassifier

model\_all=KNeighborsClassifier()

model\_all.fit(x\_train, y\_train)

y\_pred=model\_all.predict(x\_test)

y\_pred.shape

confusion\_matrix(y\_test,y\_pred)

accuracy\_score(y\_test,y\_pred)

**Naive Bayes**

from sklearn.naive\_bayes import GaussianNB

model\_all=GaussianNB()

model\_all.fit(x\_train, y\_train)

y\_pred=model\_all.predict(x\_test)

y\_pred.shape

confusion\_matrix(y\_test,y\_pred)

accuracy\_score(y\_test,y\_pre

**Random Forest**

from sklearn.ensemble import RandomForestClassifier

model\_all=RandomForestClassifier()

model\_all.fit(x\_train, y\_train)

y\_pred=model\_all.predict(x\_test)

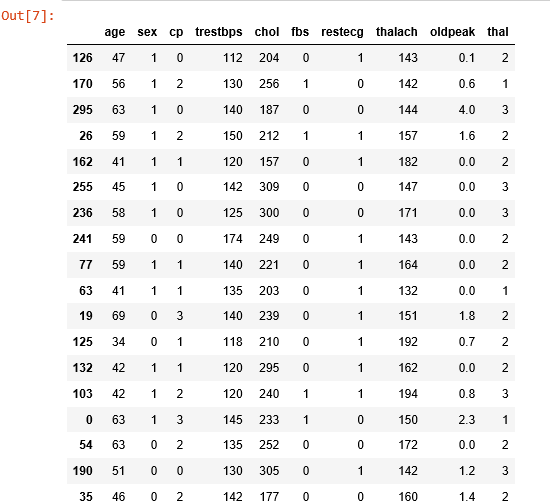
y\_pred.shape

confusion\_matrix(y\_test,y\_pred)

accuracy\_score(y\_test,y\_pred)

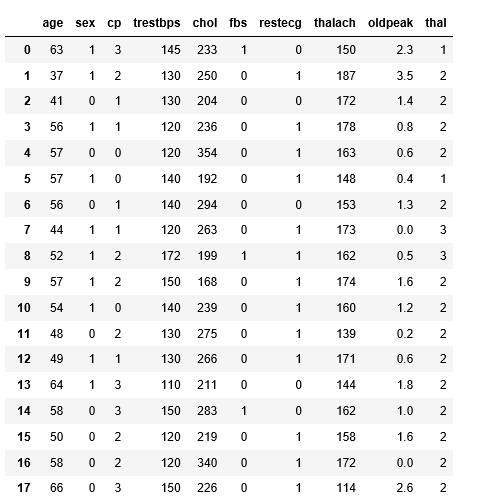
**APPENDIX II**

**EXPECTED OUTCOMES**

**IMPORTING DATA**

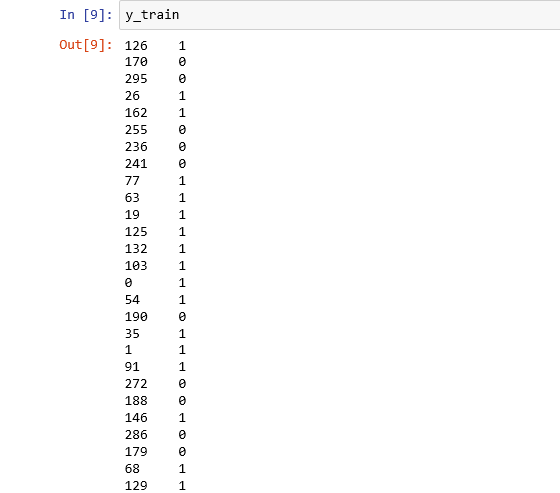
**FIGURE 2**

**READING THE CSV FILE**



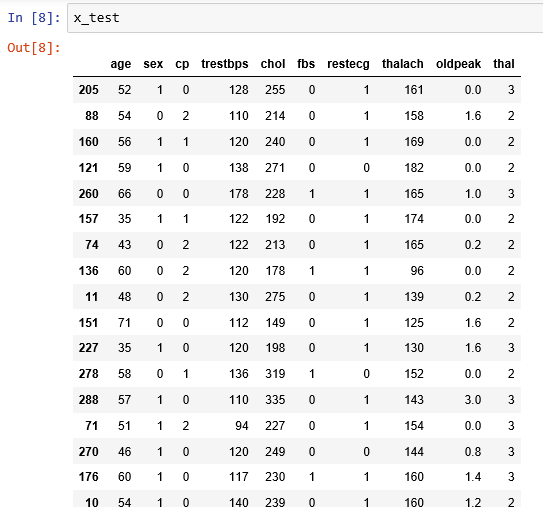
**Figure 3**

**TRAINING DATASETS**



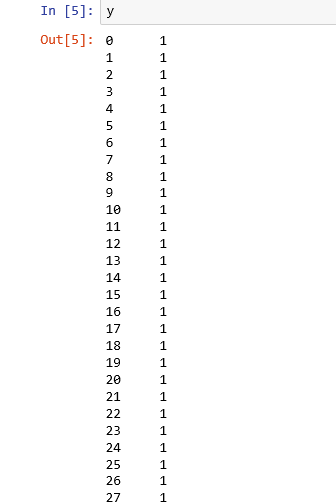
**Figure 4**

**TESTING DATASETS**



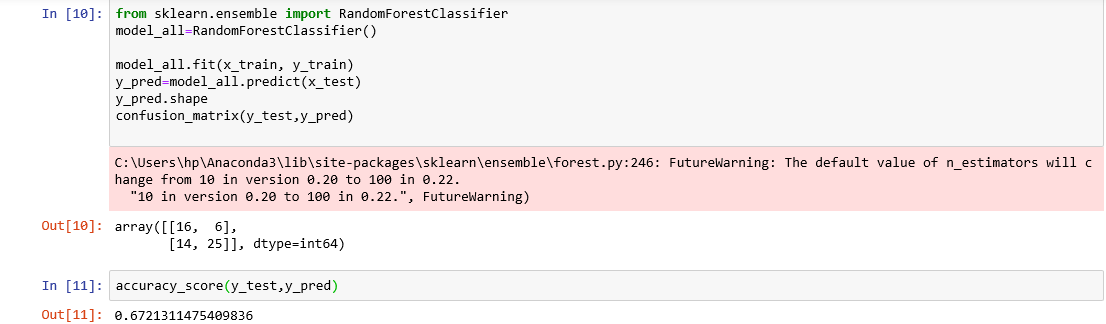
**Figure 5**

**TARGET DATA**



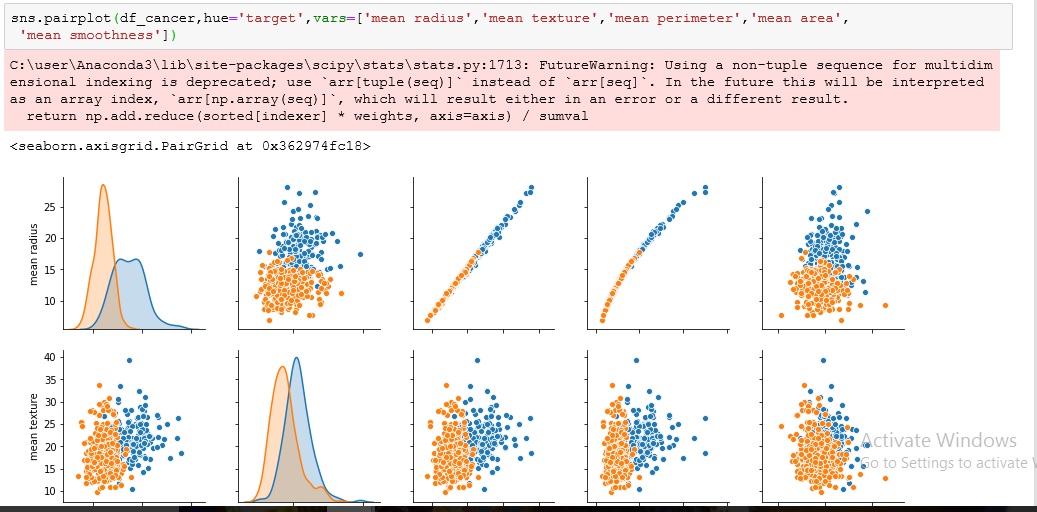
**Figure 6**

**ACCURACY DETECTION**



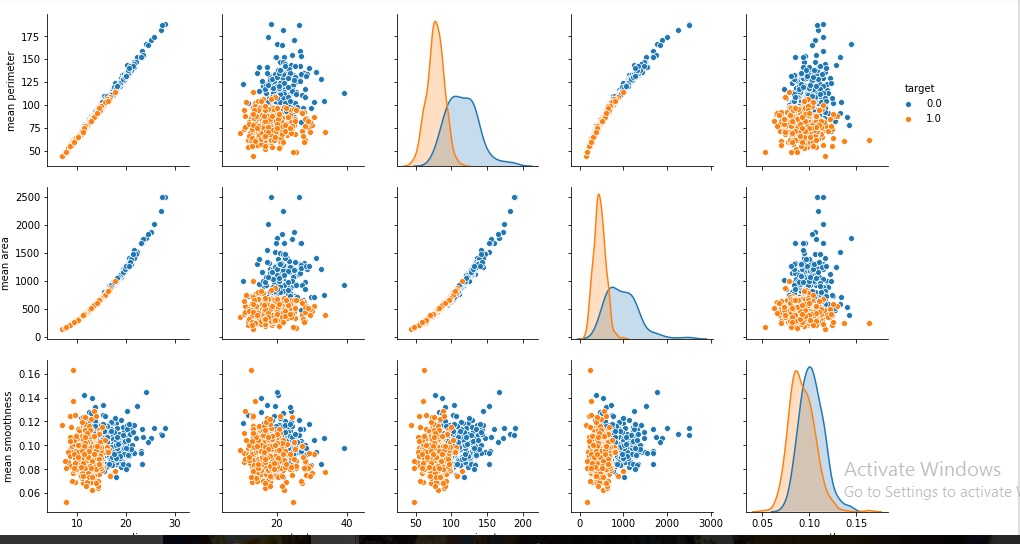
**Figure 7**

**DATA VISUALIZATION**



**Figure 8**

**SUPPORTING GRAPHS**



**Figure 9**

**REFERENCES**

[1] D. Ratnam, P. Him Bindu, V. Mallik Sai, S. P. Rama Devi and P. Raghavendra Rao, “Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm”, International Journal of Computer Science and Information Technologies, vol. 5, no. 2 **(2014)** pp.2384-2388.

[2] T. Santhanam and E. P. Hephzibah, “Heart Disease Prediction Using Hybrid Genetic Fuzzy Model”, Indian Journal of Science and Technology, vol.8, no. 9, **(2015)**, pp.797–803.

[3] G. Purusothaman and P. Krishna Kumari, “A Survey of Data Mining Techniques on Risk Prediction: Heart Disease”, Indian Journal of Science and Technology, vol. 8, no. 12, **(2015).**

[4] K. Srinivas, G. Raghavendra Rao and A. Govardhan, “Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques”, Proceedings of 5th International Conference on Computer Science & Education, China, **(2010)**, pp. 24–27.

[5] K. Thenmozhi and P. Deepika, “Heart Disease Prediction Using Classification with Different Decision Tree Techniques”, International Journal of Engineering Research and General Science, vol. 2, no. 6, **(2014)**, pp.6-11.

[6] S. Chitrali, Dungaree and S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications, vol. 47, no. 10, **(2012)**, pp. 0975 –888

[7] J Peter and K. Somasundaram, “An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques”, Proceedings of IEEE International Conference on Advances In Engineering, Science And Management (ICAESM), **(2012)**, pp. 514-518.

[8] H. D. Masethe and M. A. Masethe, “Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science (WCECS), San Francisco, USA. **(2014)**.

[9] S. J. Gnanasoundhari, G. Visalatchi and M. Balamurugan, “A Survey on Heart Disease Prediction System Using Data Mining Techniques”, International Journal of Computer Science and Mobile Application, vol. 2, no. 2 **(2014)**, pp. 72-77

[10] K. Lokanayaki and A. Malathi, “Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis”, International Journal of Computer Applications, vol. 77, no. 5, **(2013)**, pp. 0975 – 8887.

[11] A. Khemphila and V. Boonjing, “Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients”, Proceedings of International conference on Computer Information Systems and Industrial Management Applications (CISIM), **(2010)**, pp. 193-198.

[12] M. Anbarasi, E. Anupriya, Iyengar N.Ch.S.N, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm”, International Journal of Engineering Science and Technology, vol. 2, no.10, **(2010)**, pp. 5370-5376.

[13] S. Vijayarani and S. Sudha, “A Study of Heart Disease Prediction in Data Mining”, International Journal of Computer Science and Information Technology & Security (IJCSITS), vol. 2, no. 5, **(2012)**, pp. 2249-9555.

[14] L. Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, International Journal of Biological and Life Sciences, vol. 3, no. 3, **(2007**), pp. 157-160.

[15] M. AkhilJabbar, B. L. Deekshatulu, P.Chandra, “Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection”, Global Journal of Computer Science and Technology Neural & Artificial Intelligence, vol. 13, no. 3 **(2013),** pp. 0975-4350.

[16] P. Gupta and B. Kaur, “Accuracy Enhancement of Heart Disease Diagnosis System Using Neural Network and Genetic Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 103, no. 13, **(2014)**, pp. 11-15.

[17] A. Wilson, G. Wilson and J. Likhiya, “Heart Disease Prediction using the Data Mining Techniques”, International Journal of Computer Science Trends and Technology (IJCST), vol. 2, no. 1, **(2014)**, pp. 2347-8578.

[18] O. Olabode and B. T. Olabode, “Cerebrovascular Accident Attack Classification Using Multilayer Feed Forward Artificial Neural Network with Back Propagation Error”, Journal of Computer Science, vol. 8, no. 1, **(2012)**, pp.18 – 25.

[19] U. Fayyad, “Data Mining and Knowledge Discovery in Databases: Implications for scientific databases”, Proceedings of 9th International Conference on Scientific and Statistical Database Management, Olympia, Washington, USA, **(1997)**, pp. 211.

[20]www.searo.who.int/india/cardiovascular\_diseases/Commision\_on\_Macroeconimic and Health

[21] www.worldlifeexpectancy.com

[22] www.world-heart-federation.org

[23] B. Thuraisingham, “A Primer for Understanding and Applying Data Mining, IT Professional, **(2000)**,