Data Mining Algorithms in Healthcare

Anjali Dwivedi
B. Tech, Computer Science
SIET, Allahabad

Kulsoom Rehman
B. Tech, Computer Science
SIET, Allahabad

Mayuri Ghosh
B. Tech, Computer Science
SIET, Allahabad

R. Raman
Professor
Head of Department
Computer Science & Engineering
SIET, Allahabad

ABSTRACT
Data mining is the process of examining large pre-existing databases in order to generate new information. It discovers patterns in large datasets using various data mining algorithms to extract information. These data mining algorithms are extensively used in healthcare industry.

Our healthcare industry is becoming inconsistent with respect to diagnosis of diseases because of the increasing complexity of the diseases. Many algorithms are used for prediction of this disease like Naive Bayes, Support Vector Machine, Decision Tree, K-Nearest Neighbour etc. Data mining algorithms have proven to successfully predict such diseases using datasets. In this paper we have presented a general study of the most frequently used data mining algorithms in health care specially in predicting heart diseases.

General Words
Algorithms, medical field

Keywords
Cardiovascular disease, data mining, machine learning and classification.

1. INTRODUCTION
Modern medical field generates a great deal of information stored in medical database. So, extracting the useful information and making decision for diagnosis of diseases becomes necessary therefore, comes into the play the data mining algorithms. Most hospitals today employ some sort of medical information systems to manage their healthcare or patient data, which generate huge amounts of data which if used judicially, can contribute a lot to decision support system in healthcare. It would be highly advantageous if the techniques like data mining will be integrated with the medical information system.

A computer based information or decision support systems can facilitate accurate diagnosis that too at reduced cost. The automation of such system will help the physicians to do better diagnosis and treatment. The main aim is to study and analyse the application of various data mining algorithm that are being used in healthcare.

2. DATA MINING
Data mining is an intricate process of discovering and analysing meaningful data patterns that exist in large raw datasets, and it also seeks to establish relationships among the data. The main aim of the data mining process is to procure relevant information from a data set and transform it into a comprehensive construct, which is more feasible for further use. The technical basis for Data Mining is the Machine Learning; other basis includes statistics and database systems.

Data mining constitutes the exploration step in the process of "knowledge discovery in databases" i.e. KDD.KDD comprise of a repetitive process of data cleaning, data integration, data selection, data transformation, data pattern searching and knowledge representation. The KDD aims at the development of method and techniques for making sense of data, that is stored in large data warehouses and data marts.

2.1 Knowledge Discovery in Databases-
The basic model of the knowledge discovery in databases (KDD) process is defined with the following five stages:

Step (i) Selection

Step (ii) Pre-processing

Step (iii) Transformations

Step (iv) Data mining

Step (v) Interpretation/evaluation [12].

2.2 Data Mining Techniques:
Data mining involves six common classes of tasks [12]:

a) **Association rule learning** (also known as dependency modelling) – Association Rule learning seeks to establish relationships between variables.
b) **Clustering** – Clustering is the process of locating groups and structures in the data that are similar among them in any way, without using known structures in the data.
c) **Classification** – Classification is the task of assigning items in a collection to target categories or classes. For example, a tumour may be classified as benign or malignant depending.
d) **Regression**– Regression tries to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
e) **Anomaly detection** (outlier/change/deviation detection) – Anomaly Detection is the segregation of unusual data records, that might be interesting or data errors that require further investigation.
f) **Summarization**–Summarization provides a more compact representation of the data set, including visualization and report generation.

3. HEART DISEASE-

Heart diseases are disorders that affect the heart. This term is interchangeably used with cardio vascular disease. CVD or cardio vascular disease is generally related to blood vessels in particular. According to WHO a cardio vascular disease (CVD) takes 17.7 million lives every year globally which are 31% of all global deaths. The reason for the sharp increase in CVD is tobacco smoking, unhealthy diet, physical inactivity, harmful use of alcohol etc. This results in blood pressure fluctuation, obesity, elevated blood glucose which are the main reasons of CVD.

Various types of cardiovascular diseases are: -

1. **Cardiac Arrhythmia**: -They are problems with heart rhythms. They occur when the heart's electric impulse that coordinates our heartbeat does not work properly making it fast, slow or erratic. If our heart does not work properly then it does not pump blood properly and in return our organs do not get proper supply of blood and does not work properly and can get damaged [14].
2. **Angina**:-Angina is not a disease in its own right but a probable symptom of coronary artery disease. It is a tightness, pain, or discomfort in the chest that occurs when an area of the heart muscle receives less blood oxygen than usual. This usually happens because one or more of the coronary arteries is narrowed or blocked, also called ischemia[14].

3. **Congenital Heart Disease: -**This is a general term for birth defects that effect how the heart works. The word "congenital" means existing at birth. The heart ailment is a defect or abnormality, not a disease. It occurs when the heart or blood vessels near the heart do not develop normally before birth [14].

4. **Coronary Artery Disease**:-It is caused when cholesterol deposits in the coronary arteries, which results in narrowing of the arteries and heart gets less amount of oxygen.

5. **Cardiomyopathy**: - It is caused due to coronary heart diseases, because of which heart chambers get damaged and cannot pump blood properly.

6. **Myocardial Infarction**:-It is also known as heart attack, which is caused when heart muscles are damaged due to interrupted blood flow or lack of oxygen.

7. **Congestive Heart Failure**: - It is caused due to coronary artery disease or hypertension due which heart does not pump blood around the body efficiently.

8. **Mitral Valve Disease**:-It refers to irregular conditions of the mitral valve. In this condition, defective valve allows blood to flow backward into the left atrium and the heart is not able to pump enough blood out of the left ventricular chamber to supply the body with oxygen-filled blood [15]. If left untreated it can be life-threatening.

9. **Pulmonary Stenosis**: - It is caused when pulmonary valve is too tight and it is heart for the heart to pump blood from right ventricle to pulmonary artery, thus the right ventricle has to do more work to overcome this obstruction.

4. DATA MINING IN PREDICTION OF HEART DISEASE: -

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve healthcare and reduce costs. Healthcare, however, has been slow to incorporate the latest research into everyday practice.

Data mining mainly uses the two methodologies of Machine Learning:

- Supervised learning.
- Unsupervised learning.

4.1 Supervised Learning: -

In supervised learning, the system is trained with input output pairs that constitute the training set through which the system learns to generate result.

4.2 Unsupervised Learning: -

In unsupervised learning, there is no training set. Rather, it finds hidden structure and relation among the data.

The two applications of data mining are: -

- Classification
- Regression.

***4.2.1 Classification*:** - Classification models classify discrete data, unordered values. Examples of classification models include Decision trees, Neural Networks, Naïve bayes.

4.2.2 Regression: -Regression model predicts about continuous values. [5]. Examples of regression include Clustering, Association rule etc.

5. DATA MINING ALGORITHMS

In the health care industry, data mining and machine learning is mainly used for Disease Prediction. In this prediction of heart disease, we will analyse the following classification models of data mining:

1. Decision trees
2. Artificial Neural Networks
3. Naïve Bayes Classifier
4. Support Vector Machines
5. K-Nearest Neighbour

5.1. Decision trees

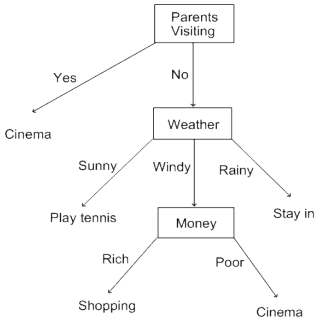

FIG. 1: Process of Decision Making.

It is one of the simplest and powerful techniques in classification in data mining. It is a knowledge representation structure that makes use of decision support tool. Decision trees uses a tree-like graph, consisting of nodes and branches such that, the internal nodes represent the attribute, the edges coming out from nodes represents the values of the attributes and the leaf node represents a class i.e. the goal attribute in the process of classification.

The process of classification in decision trees basically consists of a series of questions starting from the root node and moving down the tree, till we reach an output. (Fig. 1)

Decision tree algorithms: There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT.

Most preferred is ID3 i.e. Iterative Dichotomized, Developed by J. Ross Quinlan in 1980. ID3 uses information gain and entropy to classify data.

5.1.1 Advantages:

- It is simple and easy to comprehend and interpret.
- It allows for addition of new possible scenarios
- It can be combined with other decision techniques

5.1.2 Disadvantages:

- It may suffer from over fitting.
- It is unstable and often relatively inaccurate.
- It may lead to complex decision making situation
- Categorical data is difficult to handle; it has biased information gain in favour of those attributes with more levels.

5.2. Artificial neural network

Artificial neural network is a model to process information. This model is inspired by the way biological nervous system (specifically, the brain), process information. Artificial neural network learns by examples like the human brain. It consists of a collection of simple, highly interconnected processing units i.e. the neurons (key element in ANN) operating in parallel. Neurons process information as a response to external stimuli. Stimuli are transmitted from one processing element to another via synapses or inter-connection, which can be excitatory or inhibitory[2]

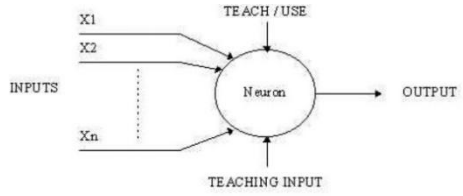

FIG. 2: A Simple Neuron

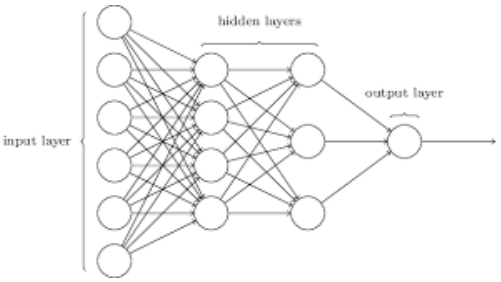

FIG. 3: A Simple Neural Networks

Neural network comprises of several layers consisting of large no of elements. Each network has three basic layers – (refer Fig. 3)

1.**Input layer**: It is the raw information that is fed into network

2. **Hidden layer**: It maps the input information to the output class.

3. **Output layer**: It outputs value of the target class as predicted by the trained neural net.

The connection between the layers is determined by the weights. Learning occurs by changing the weights at interconnections, so that influence of one neuron on another changes. Therefore, Learning is simply the determination of these weights. Learning is broadly classified is into two types:

- Supervised Learning.
- Unsupervised Learning.

Supervised Learning: - In **supervised learning**, during the process of training the network is provide with training set that consists of input-output patterns. The hidden units develop their own representation of the input, adjusting the weights on interconnections to match its output with the actual output in an iterative process until a desirable result is reached.

Unsupervised Learning:- In **unsupervised learning**, network makes use of self-organising maps by clustering the input data and find features inherent to the problem, since there are no known answers and the network is provided only with inputs.

5.2.1 *Advantages-*

- It has the ability to learn how to do tasks based on training or initial experience.
- It can handle missing or noisy data, in case of fault or partial destruction of network.
- It can easily work with large number of datasets.

5.2.2 Disadvantages-

- As network cannot be retrained, it is very difficult to modify an existing network.

5.3. Naïve Bayes classifier

Naive Bayes Classifier describes the probability of an event based on the prior knowledge of conditions that may be related to the event. It is assumed that the predictors or the conditions are independent of each other. In simple words it assumes that the presence or absence of any feature is independent of other features in that class.[16]

For example a fruit is a grape if it is oval in shape, green or violet in colour and 1-inch in diameter. Even if these features are related to each other they independently contribute to the probability of the fruit being a grape.

Bayes theorem states the following:-

P(a|c)=(P(c|a)*P(a))/P(c)

- **P(a|c)** – posterior probability of class **c** given the predictor **c**.
- **P(c|a)** – likelihood which is the probability of predictor given the class.
- **P(a)** – prior probability of class.
- **P(c)-** prior probability of predictor.
  Where: -
  - **Posterior probability** means probability which is calculated after taking into consideration the new information.
  - **Prior probability** means probability calculated before taking into consideration the new information.

5.3.1 Algorithm: -

- Given a set of variables $\mathbf{c=\{c_1,c_2,c_3.....c_n\}}$, we want to construct a posterior probability for the event $\mathbf{A_i}$

with the possible outcomes **A={A_1,A_2,A_3, ...A_i};** where **c** is set of predictors and **A_i** is set of possible outcomes and The set of attributes are independent of each other.

- Using Bayes Theorem:
  P(Ai|c_1,c_2,c_3,.....c_n)αP(c_1,c_2,c_3,.....c_n|A_i).P(A_i)
- Where **P (Ai | c_1, c_2, c_3..., cn)** is the posterior probability i.e., the probability that **c** belongs to **Ai.** Since Naive Bayes assumes that the conditional probabilities of the independent variables are independent of each other we can reduce the probability to a product of terms:

  P (c| A_i) $\alpha\pi^n_{k=1}$P (c_k|A_i)

  We rewrite the posterior probability as:

  P (A_i|c) α P (A_i) $\pi^n_{k=1}$P(c_k|A_i)

Using Bayes' rule above, we label a new case **c** with a class level **Ai** that achieves the highest posterior probability.

- Although the assumption that the independent variables are independent is not always true, it does simplify the classification task extensively.

5.3.3 Advantages: -

- It is fast and easy to use and predict class of test data set.
- When assumption of independence holds then it works better than other classification algorithms.
- It works well in case of categorical input variables in place of numerical variables.

5.3.4 Disadvantages: -

- It is limited to the independence of predictor's assumption.
- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction.

5.4. SUPPORT VECTOR MACHINE

SVM is a supervised machine learning algorithm which is mainly used for the classification of the data sets. It was developed by Vapnik in 1963. It classifies the dataset by building a hyper plane so that two classes are separated as wide as possible [9].

Support Vectors: The points/vectors from each class that participate in building a hyper plane are known as support vectors.

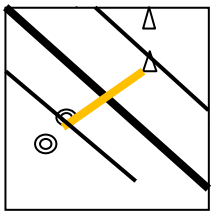

Fig4. Classification in SVM

In the figure 4, the separation (decision boundary/hyper plane) is the perpendicular bisector of the line joining two critical points (support vectors) from each class building a hyper plane. The two lines beside the decision boundary are the margins which need to be as wide as possible. General formula for support vector machine in terms of the diagram is as follows

$$Y = w^t x + b$$

$$w^t y + b = 1$$

$$w^t z + b = -1$$

Where,

Y: classification label**, y∈ {-1, +1}**

w: weight vector

x and b: parameters of plane.

y and z: triangle and circle respectively

If the points occurs on the decision boundary $w^t x + b = 0$

There are two types of kernel:

- **Linear kernel**: when the two classes are separated by a single line.
- **Gaussian kernel**: for complex boundaries we use Gaussian Kernel.

5.4.1 Advantages

- Effective in high dimensional space.
- Resistant to over fitting [18].
- Memory efficient as it uses only support vectors to build the hyper plane.

5.4.2 Disadvantages

- Used for small datasets
- Premature optimization

5.5. K- Nearest Neighbour

K-Nearest neighbours is used for classification problems. It is based on distance based algorithms. It is an algorithm that stores all available classes and classifies the new unknown class based on its distance from the known stored classes which is calculated with the help of distance function [17].

K-NN algorithm is used in statistical estimation, pattern recognition, classification problems etc.

For example: -

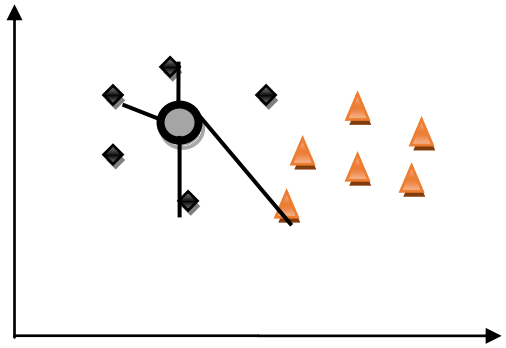

FIG 5.Classification in K-NN

In the above figure the diamonds and the triangles are the known classes and the oval is the unknown class.

- We have to find the class to which this oval belongs.
- We then generally use distance function to calculate the distance between the known and unknown class.
- The unknown class is then classified among the known class by a majority of vote by the neighbours.

The distance functions used are: -

Table 1. Distance function and formula

	DISTANCE FUNCTION & FORMULA
1	$\sqrt[2]{\sum_{i=1}^{k}(x_i - y_i)^2}$ Euclidean distance function(for continuous variables)
2	$\sum_{i=1}^{k}\|x_i - y_i\|$ Manhattan distance function(for continuous variables)
3	$\left(\sum_{i=1}^{k}(\|x_i - y_i\|)^q\right)^{1/q}$ Minkowski distance function(for continuous variables)
4	$D_H = \sum_{i=1}^{k}\|x_i - y_i\|$ Hamming distance function(for categorical variables)

Choosing the optimal value for K is best done by inspecting the data.

5.1.1 Advantages: -

- Implementation is simple [1].
- Analytic method.
- Use local information.

5.1.2 Disadvantages: -

- It is slow process.
- High storage required
- Affected by dimensionality.

6. RELATED WORK: -

There has been much work done in the field of classification of cardiac arrhythmia. Most of the work is based on neural networks, Support vector machines, Naïve Bayes. Self-organising maps (SOMs) are used for the analysis of ECG signals.

There has also been developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Networks **[21].** IHDPS is web-based, user-friendly, scalable, reliable and expandable system which is implemented on the .NET platform. It displays the results both in tabular and graphical forms. This IHDPS is based on 15 attributes. All the related papers cited included various algorithms which are taken forward herein by showing direct comparisons between these algorithms.

Table 2. Some related works of various authors are summarized below:

1.	AUTHORS: THARA SOMAN PATRICK O. BOBBIE [3]. TOPIC: "Classification of Arrhythmia Using Machine Learning Techniques". OBJECTIVE: The aim of their study is to automatically classify cardiac arrhythmias and also to study the performance of various machine learning algorithms. ALGORITHM STUDIED: Naïve bayes, J48, OneR. RESULT: The result shows that naïve bayes and OneR have stable accuracy rate which is not true for J48.
2.	AUTHORS: Babak Mohammadzadeh Asl, Seyed Kamaled in Setarehda, Maryam Mohebbi[4].TOPIC:"Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal''. OBJECTIVE: It shows cardiac arrhythmia classification algorithm using the heart rate variability (HRV) signal. The algorithm which they propose is based on the generalized discriminant analysis (GDA) feature reduction scheme and the support vector machine (SVM) classifier. ALGORITHM STUDIED Support vector machine. RESULT: It is completely based on the HRV (R—R interval) signal which can be extracted from even a very noisy ECG signal with a relatively high accuracy rate.
3.	AUTHORS: Jeen-Shing Wanga,n , Wei-Chun Chiang a , Yu-Liang Hsu a , Ya-Ting C.Yang[5].TOPIC:"ECG arrhythmia classification using a probabilistic neural network with a feature reduction method". OBJECTIVE: The objective is to present an effective electrocardiogram (ECG) arrhythmia classification which consists of a feature reduction method combining principal component analysis (PCA) with linear discriminant analysis (LDA), and a probabilistic neural network (PNN) classifier to classify eight different classes of cardiac arrhythmia ALGORITHM STUDIED: Neural network. RESULT: Experimental results have successfully showed that the method used by them can achieve satisfactory accurate results.
4.	AUTHORS: Dayong Gao, Michael Madden, Des Chambers, and Gerard Lyons [6]. TOPIC: "Bayesian ANN Classifier for ECG Arrhythmia Diagnostic System: A Comparison Study". OBJECTIVE: It outlines a system for detection of cardiac arrhythmias within ECG signals, which is built by the use of a logistic regression model and the back propagation algorithm based on a Bayesian framework. ALGORITHM STUDIED: Naïve bayes, neural network, logistic regression, decision tree. RESULT: Bayesian ANN classifier appears to acquire arrhythmia properties from the underlying dynamics of the system, even when the dataset includes incomplete information, such as missing feature values and unclassified classes. This approach is potentially useful for generating a pattern recognition model based on given {input, output} sets to classify future input sets for arrhythmia diagnosis.
5.	AUTHORS: Luyang Chen, Qi Cao, Sihua Li, Xiao Ju[7].TOPIC:"Predicting Heart Attacks". OBJECTIVE: This paper aims at a better understanding and application of machine learning in medical domain. In this paper, they modify three classical models for multiclass problems: Logistic Regression, Naive Bayes and SVM, and then implement them to predict cardiac arrhythmia based on patients' medical records .ALGORITHM

	STUDIED: Naïve bayes, Support vector machine, Logistic Regression. RESULT: In their paper, before feature selection, Naive Bayes achieves lower cross validation error than SVM. While after feature selection, SVM achieves lower cross validation error than Naive Bayes. The problem may lie in the lack of enough training examples (475) and excessive number of features (274).
6.	AUTHORS: R. Tamilarasi, Dr. R. Porkodi[8].TOPIC:"A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare". OBJECTIVE: The aim of paper is to find the performance of different classification methods of large datasets. ALGORITHM STUDIED: Decision tree, naïve bayes, IBK, J48, Neural network. RESULT: The results showed that the performance of each of the classification algorithm differs on type of problem are being considered. The best classification algorithm based on heart data is K Nearest Neighbour classifier. It has an accuracy of 100% and the total time taken to build the model is at 0.13 seconds. K Nearest Neighbour has the lowest error compared to others.

7. CONCLUSION

Data mining algorithms are providing huge success in the field of healthcare particularly for the diagnosis of diseases. Lot of work has been done and lots more to be accomplished in future. Every algorithm that has been worked upon has shown their individual capabilities in predicting diseases and there is lot of variation depending upon the attributes. The result of our analysis and after studying the work of various authors shows that the algorithms namely naïve bayes, SVM, decision tree, KNN and neural network are the most commonly used algorithms for diagnosis of diseases and thus hold grave importance in healthcare.

8. REFERENCES

[1] Satish Kumar David1*, Amr T.M. Saeb2, Khalid Al Rubeaa . "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics" www.iiste.org, ISSN 2222-2863 (Online) Vol.4, No.13, 2013.

[2] Harleen Kaur and Siri Krishan Wasan," Empirical Study on Applications on Data Mining Techniques in Healthcare" Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications.

[3] THARA SOMAN PATRICK O. BOBBIE. "Classification of Arrhythmia Using Machine Learning Techniques". Marietta Parkway, Marietta, GA 30060-ECGDiagnosis-ICOSSE-2005-V2.0.doc

[4] Babak M. Asl, Seyed K. Setarehda , M. Mohebbi."Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal'', Iran-Elsevier-Artificial Intelligence in Medicine (2008) 44, 51—64 accepted 28 April 2008.

[5] Jeen-Shing Wang a,n , Wei-Chun Chiang a , Yu-Liang Hsu a , Ya-Ting C. Yang."ECG arrhythmia classification using a probabilistic neural network with a feature reduction method". ROC.-Elsevier- Neurocomputing 116 (2013) 38–45

[6] D. Gao, Michael Madden, Des Chambers, and G. Lyons"Bayesian ANN Classifier for ECG Arrhythmia Diagnostic System: A Comparison Study". Ireland-Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005

[7] Predicting Heart Attacks Luyang Chen, Qi Cao, Sihua Li, Xiao Ju Stanford University lych, qcao, sihua, and xju@stanford.edu.

[8] R.Tamilarasi, Dr. R. Porkodi," A Study and Analysis of Disease Prediction Techniques in Data Mining for Healthcare " IJERMT ISSN: 2278-9359 (Volume-4, Issue-3)-March 2015

[9] Ismail Babaogˇlu a, *, Ogˇuz Fındık a, Mehmet Bayrak "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine" -Elsevier- Expert Systems with Applications Volume 37, Issue 3, 15 March 2010, Pages 2182-

[10] "Data mining Curriculum" ACM SIGKDD. http://www.kdd.org/curriculum/index.html

[11] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic "From Data Mining to Knowledge Discovery in Databases", Published 1996 in Lecture Notes in Computer Science- DOI: 10.1007/978-3-642-37456-2

[12] https://childrensnational.org/choose-childrens/conditions-and treatments/heart/pulmonary-stenosis.

[13] https://www.mdedge.com/ecardiologynews/article/41655 /acute-coronary-syndromes/five-types-mi-will-make-new-definition.

[14] http://www.heart.org.

[15] https://www.healthline.com/health/mitral-valve-disease.2185.

[16] http://www.statsoft.com/textbook/naive-bayes-classifier.

[17] http://www.saedsayad.com/k_nearest_neighbors.html

[18] https://gerardnico.com/wiki/data_mining/support_vector _machine

[19] https://www.analyticsvidhya.com/.../understanding-support-vector-machine-example

[20] https://en.wikibooks.org/wiki/Data_Mining_Algorithms_ In_R/Classification/SVM

[21] Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang https://www.scribd.com/doc/58095973/IHDPS.

[22]