Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification

Prashasti Kanikar
Assistant Professor (Computer Engg.)
NMIMS University
Mumbai, India

Disha Rajeshkumar Shah
M Tech (Computer Engg.)
NMIMS University
Mumbai, India

ABSTRACT
Cardiovascular disease is a broad team for a range of diseases affecting heart and blood vessels. Cardiovascular disease are the number one cause of death globally. The health care industry contains lots of medical data, therefore data mining techniques are required to discover hidden patterns and to make decision effectively in prediction of heart diseases. By applying data mining techniques, valuable knowledge can be extracted from health care systems. Data mining classification techniques like Naïve Bayesian and Support vector machine (SVM) are explained in this paper with their benefits and limitations. Data mining will help doctors to extract useful information from a huge dataset. In proposed research pre-processing uses techniques like noise removal, discarding records with missing data, filling default values if applicable and classification of attributes for decision making at different levels. This paper has predicted accuracy, specificity and sensitivity using a classifier. A classifier will predict whether a person has heart disease or not by using machine learning techniques like Support Vector Machine (SVM) and Naïve Bayes.

General Terms
Data Mining; Cardiovascular Diseases

Keywords
Classification; Support Vector Machine (SVM); Naïve Bayes

1. INTRODUCTION
In medical industry huge amount of data is available but people are not able to extract the important information about the factors that causes cardiovascular disease. Data mining will assist doctors and patients to carry out their diagnosis. It will help doctors to emphasize on some informative knowledge to predict the disease more quickly. According to the Indian National Commission on Macroeconomics and Health, it shows that in India in the year 2015 there are 61.5 million cases of cardiovascular diseases [17]. A challenging task in medical industry is to provide effective treatments to patients and to diagnose the disease correctly. The occurrence of coronary artery disease is calculated to be up to 7 percent in rural India as measured in urban areas where it is up to 12 percent [19]. The major reasons for this epidemic are changes in lifestyle such as inactive jobs, increase in socioeconomic status that leads to unhealthy diets, increase in job stress and dependence on smoking and tobacco.

"Data mining is defined as non-trivial extraction of implied, unidentified and hidden information about data" [18]. Mining hidden patterns in Data mining is done through techniques like classification and clustering. Decision Tree, Naïve Bayesian, Neural Network (NN) and Support Vector Machine (SVM) are various classification technique. Fuzzy C-means, K-means and Apriori algorithm are various Clustering techniques [13]. In data mining, the challenge is to extract patterns which are earlier unknown and hidden.

Tasks in data mining are split into two type's i.e. predictive and descriptive task [16]. Predicting the value of an individual attribute on the basis of another attribute is done in Predictive tasks. Predictive tasks includes classification technique. Descriptive tasks summarize the relationship between data and it determine patterns. Descriptive tasks includes clustering techniques.

Data pre-processing is an essential step in the data mining process. Data-gathering methods are generally loosely controlled, occurring in out-of-scope values, impractical combinations of data, missing values, etc. [30] Evaluating data that has not been carefully hidden for such problems can produce inaccurate results. Thus, the representation and data quality is first and most important before running an analysis. If there is much inappropriate and redundant data present or noisy and unreliable data, then knowledge discovery at the time of the training phase is also challenging.

Data pre-processing tasks includes data cleaning, data integration, data transformation, data reduction and data discretization [16]

1. Data Cleaning: Data cleaning process fill's in missing values, smooth noisy data, classify or eliminate outliers, and clarify inconsistencies.
2. Data Integration: Data integration is done using different databases, records, or data cubes.
3. Data Transformation: Data transformation task includes techniques like normalization and aggregation.
4. Data Reduction: Data reduction reduces the volume but produces the same or similar analytical results.
5. Data Discretization: Data discretization is a part of data reduction. Its task is to replace numerical attributes with nominal ones.

2. LITERATURE SURVEY
Cardiovascular diseases (CVDs) are a set of disorders of the heart. Coronary heart disease, cerebrovascular disease, peripheral disease, rheumatic and congenital heart diseases are the different types of cardiovascular diseases [1]. Heart attack syndromes are chest pain, pain in the arms and in left shoulder, elbow pain or back pain. Other symptoms are troublesome in breathing or breathe shortness, vomiting or feeling faint. Risk factors that causes heart attack are commonly tobacco, hypertension, obesity, diabetes, unhealthy diets and physical inactivity [4]. In India, coronary heart disease is very peculiar. In Coronary heart disease (CHD), a waxy substance develops inside the coronary arteries. Various tests to diagnose coronary heart disease includes an MRI scan, CT scan, Coronary angiography, Stress testing. Electrocardiogram (ECG) and

nuclear isotope imaging [21]. Coronary heart disease has various precautionary measures like no smoking, exercise regularly, maintaining blood cholesterol and weight of the body, properly maintaining diabetes and high blood pressure levels [22]. Early detection is needed for the people who are at very high cardiovascular risk or people with cardiovascular disease. Hence, more efficient methods of cardiovascular disease are of great concern.

3. PROPOSED SYSTEM

The main objective of this research is to predict heart disease using risk factors like Age, Sex, Chest pain type, Resting blood sugar, Cholesterol, Resting Electrographic results, Fasting blood sugar, Thalach, Exang, Oldpeak, Slope, Number of major vessels colored by Flourosopy, Thal, Height and Weight. In proposed research pre-processing techniques includes noise removal, discarding records with missing data, filling default values if applicable and classification of attributes for decision making at different levels. In this paper, classification techniques are used. Classification technique forecasts class membership for data samples. The data mining classification techniques like Support Vector Machine (SVM) and Naïve Bayes are used.

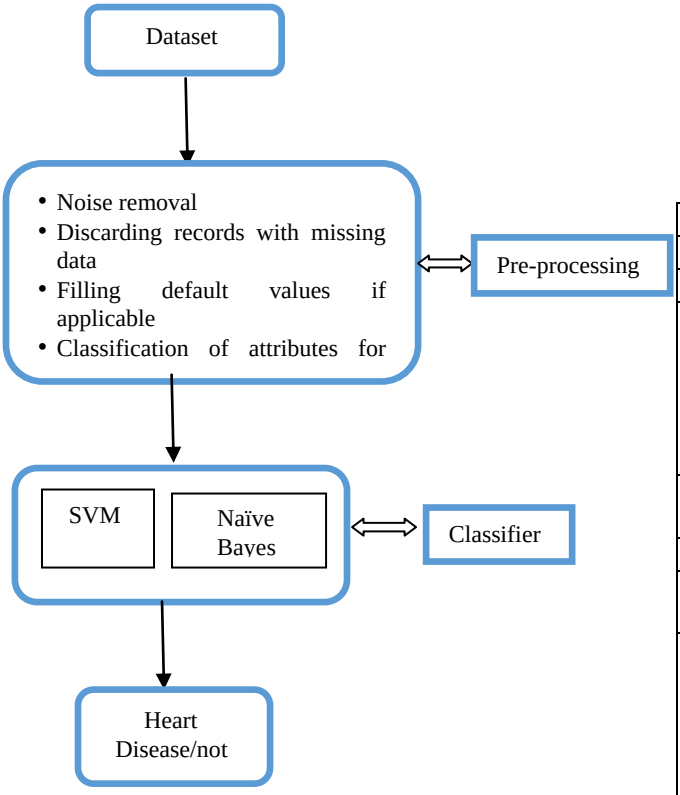

Fig 1: Proposed system for prediction of cardiovascular diseases

Figure 1 describes the flow chart of the proposed system. The initial step of the process is to collect dataset then technique like pre-processing is done. Classification algorithms like Support Vector Machine (SVM) and Naïve Bayes is use as classifier to predict the final result i.e. whether a person has heart disease or not.

4. DATA SET

The data set of total 303 records are used in prediction of cardiovascular diseases with 15 attributes (risk factors) are obtained from machine learning repository of UCI [31]. The attribute Patient Id is use as patient identification number. Table 1 lists all the factors. The records are divided into two datasets: training and testing dataset. The records for each set are selected randomly, to avoid favoritism.

A	B	C	D	E	F
Patient Id	P1	P2	P3	P4	P5
Age	63	44	60	55	66
Sex	1	1	1	1	1
Chest pain type(CP)	4	4	4	4	3
TrestBps	140	130	132	142	110
Cholestrol	260	209	218	228	213
FBS	0	0	0	0	1
RestECG	1	1	1	1	2
Thalach	112	127	140	149	99
Exang	1	0	1	1	1
Old peak	3	0	1.5	2.5	1.3
Slope of the peak exercise ST segment	2	2	3	1	2
Ca	3	0	0	2	3
Thal	3	6	3	3	3
Height(feet)	7'6"	5'4"	6'7"	7'5"	6'10"
Weight(Kgs)	75	51	61	63	52

Fig 2: Sample data set

Figure 2 shows 303 records with 15 risk factors from machine learning repository of UCI and patient Id attribute is use as patient identification number.

Table 1. Description of attributes

Name	Description
Age	age: age in years
Sex	sex: sex (1 = male; 0 = female)
Cp	cp: chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
Trestbps	trestbps: resting blood pressure (in mm Hg on admission to the hospital)
Chol	chol: serum cholestoral in mg/dl
Fbs	fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	restecg: resting electrocardiographic results Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalach	thalach: maximum heart rate achieved
Exang	exang: exercise induced angina (1 = yes; 0 = no)
Oldpeak	oldpeak = ST depression induced by exercise relative to rest
Slope	slope: the slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: downsloping
Ca	ca: number of major vessels (0-3) colored by

	flourosopy
Thal	thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
Height	Height in Feet
Weight	Weight in Kgs

5. DATA MINING TECHNIQUES

Data mining techniques are adopted to find unknown patterns from the data and an analysis of data is performed by data mining techniques.

5.1 Classification Techniques Used in Data Mining

5.1.1 Naïve Bayesian

Naïve Bayes algorithm is depended on Bayes theorem [7]. Naïve Bayesian algorithm is named after Thomas Bayes, who was an eighteenth-century theologian. Naïve Bayesian classification is an analytical classifier which classifies between attributes and assumes no dependency between them. Algorithm start out with the simplest probabilistic classifier and then make a few assumptions and learn the naïve Bayes classifier. It's called naïve because the formulation makes some naïve assumptions [27].

Naive Bayes model is easy to build and especially effective for very large data sets. Naïve Bayes is known to outrun even exceptionally sophisticated methods of classification, along with homogeneity [26]. Naive Bayes classifiers are extremely flexible, demanding a number of parameters continuous in the number of features in a learning issue. Maximum-likelihood training can be completed by estimating a closed-form expression, which yields linear time, relatively than by costly iterative approximation as needed for many other category of classifiers. Formula for Bayes theorem is given by: [29]

$$P(H/X) = \frac{P(X/H).\ P(H)}{P(X)}$$

Above,

- P(H/X) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(H) is the prior probability of class.
- P(X/H) is the likelihood which is the probability of predictor given class.
- P(X) is the prior probability of predictor.

5.1.2 Support Vector Machine(SVM) Algorithm

Support Vector Machine (SVM) make good decisions for data points that are outside the training set. There are two classes of data in SVM. The data points are isolated in such a way that they could draw a straight line on the figure. The line is made in a way that it separates all the points on one side of one class and all the points on the other side of the other class. When such situation occurs, then the data are linearly separable. The line used to separate the dataset is called a separating hyperplane. The points closest to the separating hyperplane are known as support vectors. Kernels are used to extend SVMs to a larger number of datasets. Mapping of one feature space to another is done by kernel [29].

In addition to performing linear classification, SVMs can conveniently implement a non-linear classification using the kernel trick, essentially mapping their inputs into high dimensional feature spaces. This mapping from one feature space to another is done by a kernel. Assume kernel as a wrapper or interface for the data to convert it from a tough formatting to a simple formatting. The Radial Bias Function (RBF) is a kernel that's generally needed with support vector machines (SVM). A radial bias function is an operation that takes a vector and outputs a scalar positioned on the vector's distance. This space can be one from 0, 0 or from another vector.

Kernel method, maps the data (sometimes also called as nonlinear data) from a small dimensional space to a large dimensional space. In a larger dimension, it determines linear problem that's nonlinear in smaller-dimensional space. The Radial Bias Function (RBF) is a prominent kernel that measures the distance among two vectors. The execution of an SVM is also responsive to optimization parameters and specifications of the kernel used. Support vector machines (SVM) are a binary classifier and other methods can be continued to classification of classes greater than two.

5.2 Benefits and Limitations of Classification Algorithm [14][30]

	Advantages	Disadvantages
Support Vector Machine (SVM)	1. Easy to interpret results	1. Sensitive to tuning parameters and kernel choice
	2.Low generalization error and Computationally inexpensive	2. Natively only handles binary classification
Naïve Bayes	1. It provides minimum error rate and easy to build	1.Not much accurate because assumptions are made due to class conditional dependencies
	2. Works with a small amount of data and handles multiple classes	2. Sensitive to how the input data is prepared

6. RESULTS

Training & Testing Ratio	SVM			Naïve Bayes		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
50:50	57%	43%	38%	61%	50%	50%
70:30	53%	85%	25%	49%	65%	23%
75:25	55%	80%	20%	40%	84%	19%
80:20	57%	87%	35%	52%	86%	28%

Fig 3: Results of classification algorithms

Figure shows accuracy, specificity and sensitivity of SVM & Naïve Bayes approach for different training and testing data.

Graph of SVM &Naïve Bayes approach

Fig 4: Performance Evaluation Measure of SVM and Naïve Bayes algorithm for different testing and training data

On y-axis, figure shows the percentage and on x-axis it shows accuracy, specificity and sensitivity of SVM and Naïve Bayesian algorithm. Blue, green, yellow and red lines indicates 50:50, 70:30, 75:25, 80:20 ratio respectively of training and testing data.

Graph of accuracy for SVM & Naïve Bayes algorithm

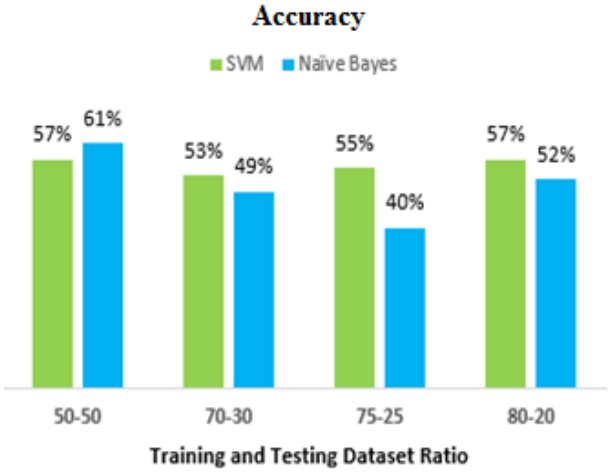

Fig 5: Accuracy graph for SVM and Naïve Bayes algorithm for different testing and training data

Green and blue bar in the figure indicates accuracy of SVM and Naïve Bayes algorithm respectively with percentages of accuracy mentioned above the bar chart with dataset of 50:50, 70:30, 75:25, 80:20 as training and testing ratio.

7. CONCLUSION & FUTURE SCOPE

Prediction of cardiovascular disease is a major challenge in health care systems. The objective of proposed work is to provide a study of different data mining classification techniques with their pros and cons. Data set of 303 records and 15 attributes is collected from UCI. Results shows accuracy, specificity and sensitivity for SVM & Naïve Bayes algorithm with different number of training dataset and testing dataset. Accuracy graph shows that SVM algorithm is better than Naïve Bayes because accuracy of SVM is not below 50% in any training and testing dataset. SVM algorithm performs better for large dataset using Radial Bias Function (RBF). Based on literature review only two algorithms namely Support Vector Machine (SVM) and Naïve Bayes classification have been implemented so far. There is still scope for improvement in accuracy, specificity and sensitivity. So other classification approaches can be implemented and tested.

8. ACKNOWLEDGMENTS

This research was supported by my guide Prof. Prashasti Kanikar, for providing excellent guidance, encouragement, inspiration, suggestions and support from an early stage of this research and providing me extraordinary experiences throughout the project work. Her involvement with originality has triggered and nourished my intellectual maturity that will help me for a long time to come. I would also like to thank the Head of the Department, Dr. Dhirendra Mishra, for their kind support and would also like to express my gratitude to the Dean of the college, Dr.S.Y.Mhaiskar and also to Asst. Prof. Avinash Tandle at MPSTME, NMIMS University for their support and cooperation.

9. REFERENCES

[1] K Raj Mohan, Ilango Paramasivam and Subhashini Sathya Narayan, "Prediction and Diagnosis of Cardio Vascular Disease– A Critical Survey", published on Computing and Communication Technologies (WCCCT), 2014 World Congress on, pp.246-251, Feb. 27 2014-March 1 2014

[2] Minas A. Karaolis, Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE Transactions On Information Technology In Biomedicine, VOL. 14, NO. 3, MAY 2010.

[3] T.John Peter, K. Somasundaram, "An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", IEEE, International conference on Advances in engineering, science and management, pp.514-518, 2012.

[4] Sulabha S.Apte and Chaitrali S.Dangare, "Improved Study of Heart Disease prediction System using Data Mining Classification Technique", published in International Journal of Computer Applications(0975-888), Vol.47-No.10, June 2012

[5] Lovepreet Kaur, "Predicting Heart Disease Symptoms using Fuzzy C-Means Clustering", published in International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 12, December 2014.

[6] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", IJCA, Vol.68- No.16 April 2013.

[7] Ranganatha S, Pooja Raj H.R., Anusha C and Vinay S.K., "Medical data mining and analysis for heart disease dataset using classification techniques", published in IEEE National Conference on Challenges in Research & Technology in the Coming Decades (CRT 2013), , pp.1 – 5, 27-28 Sept. 2013.

[8] Alireza Kajabadi, Mohamad Hosein Saraee, and Sedighe Asgari., "Data Mining Cardiovascular Risk Factors", published in Application of Information and Communication Technologies, 2009.AICT 2009.International Conference on, pp. 1-5,14-16 Oct.2009

[9] Yanwei Xing ,Jie Wang, Zhihong Zhao, and Yonghong Gao, "Combination Data Mining Methods with New

Medical Data to Predicting Outcome of Coronary Heart Disease", published in Convergence Information Technology, 2007. International Conference on, pp.868 – 872, 21-23 Nov. 2007.

[10] Chen, A.H., Huang, S.Y.; Hong, P.S.; Cheng, C.H.and Lin, E.J., "HDPS: Heart disease prediction system", published in Computing in Cardiology, 2011, pp. 557 – 560, 18-21 Sept. 2011.

[11] Eman AbuKhousa, Piers Campbell, "Predictive Data Mining to Support Clinical Decisions: An Overview of Heart Disease Prediction Systems", published in 2012 International Conference on Innovations in Information Technology (IIT),pp. 267 – 272, 18-20 March 2012.

[12] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh, "Heart Disease Diagnosis System Using Apriori Algorithm", published in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, February 2015.

[13] Aqueel Ahmed, Shaikh Abdul Hannan,"Data Mining Techniques to Find out Heart Diseases: An Overview", published in International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-1, Issue-4, September 2012.

[14] Shashikant Ghumbre,Chetn Patil and Ashok Ghatol, "Heart Disease Diagnosis Using Support Vector Machine", International Conference on computer science and information Technology(ICCSIT 2011),Pattaya Dec.2011.

[15] Nidhi Bhatla,Kiran Jyoti, "An Analysis Of Heart Disease Prediction Using Different Data Mining Techniques", International journal of engineering Research and Technology(IJERT),ISSN:2278-0181,Vol.1 Issue 8,October2012.

[16] S.Sivagowry,M.Durairaj; A.Persia, " Am empirical study on applying data mining techniques for the analysis and prediction of heart diseases", Published in Information Communication and Embedded Systems (ICICES), 2013 International Conference, pp.265-270,21-22 Feb 2013.

[17] Rajeev Gupta, KD Gupta, "Coronary Heart Disease in Low Socioeconomic Status Subjects in India -An Evolving Epidemic", 2009. [Online]. Available: http://indianheartjournal.com/ihj09/july_aug_09/358-367.html. [Accessed: 24-Aug-2015].

[18] Frawley and G.Piatetsky-shapiroa, "knowledge discovery in databases: An Overview", published by the AAAI Press/ The MIT Press, Menlo Park, C.A.1996.

[19] Indian express news on heart disease. [Online]. Available: http://archive.indianexpress.com/news/india-set-to-be-heartdisease-capital-of-world--say-doctors/1009607. [Accessed: 24-Aug-2015]

[20] M. Bogl, W. Aigner, P. Filzmoser, T. Gschwandtner, T. Lammarsch, S. Miksch, and A. Rind, "Visual Analytics Methods to Guide Diagnostics for Time Series Model Predictions", published in Proceedings of the 2014 IEEE VIS Workshop on Visualization for Predictive Analytics.

[21] Cardiovascular disease [Online]. Available: http://www.nhlbi.nih.gov/health/healthtopics/topics/cad/diagnosis. [Accessed : 25-Aug-2015]

[22] Coronary heart disease in Indians. [Online]. Available:http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3028954 [Accessed: 25-Aug-2015]

[23] Jyoti Soni Ujma Ansari Dipesh Sharma and Sunita Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", published in International Journal of Computer Applications, Volume 17– on 8, March 2011.

[24] Zhifang He and shuiping Chen, "Application of spss software on mental health education for community resident", published in Computer Science & Education (ICCSE), 2015 10th International Conference on 22-24 July 2015, pp. 673 – 676.

[25] S.Florence1, Amma2, G.Annapoorani, K.Malathi, "Predicting the Risk of Heart Attacks using Neural Network and Decision Tree", published in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, November 2014.

[26] Ms. Priti V. Wadal, Dr. S. R. Gupta, "Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction", published in International Journal of Engineering Research and Applications and International Conference on Industrial Automation and Computing (ICIAC) on 12-13th April 2014.

[27] L. A. Muhammed, "Using data mining technique to diagnosis heart disease", published in in Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference, pp. 1-3,10-12 Sept. 2012.

[28] Carlos O., Edward O, Levien de Braal, and team "Mining Constrained Association Rules to Predict Heart Disease", IEEE, International Conference on Data Mining p.433-440, 2001.

[29] Peter Harrington, "Machine Learning in Actions", Published in April 16th 2012 by Manning Publications.

[30] Jiawei H. Micheline Kamber,"Data Mining,Concepts and techniques",SecondEdition,Elsevier,2006.

[31] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. [Accessed:27-April-2016]