https://doi.org/10.15623/ijret.2018.0707026 Received: 15-05-2018, Accepted: 26-06-2018, Published: 17-07-2018

ADVANCED DETECTION OF BLOCKED CORONARY ARTERIES USING MACHINE LEARNING ALGORITHMS

Sujesh Shankar Dinesh Mandal[1]

[1]*School of Information Technology & Engineering, Vellore Institute of Technology, Tamil Nadu, India*

Abstract

The healthcare industry is observing a tremendous advancement along with upcoming innovations in Information Technology and Computer Science. A common task in Machine Learning is to classify data. An essential task for extracting knowledge from large databases is done by Data Mining. Data Mining in the Healthcare industry is an upcoming field of interest not only for Data Scientists but for Medical Experts for providing deeper understanding and prognosis of medical data. A majority of data mining methods depend on a set of features that define the learning algorithm directly or indirectly and influence the complexity of the resulting models. In the last 10 years, heart disease has been the leading cause of deaths in the world. A lot of researchers have been using data mining techniques to diagnose heart diseases using machine learning algorithms. Coronary Artery Disease [CAD] is a chronic disease that occurs when there is usual cause is the buildup of plaque. This causes coronary arteries to narrow, limiting blood flow to the heart. Coronary artery disease can range from no symptoms, to chest pain, to a heart attack. Treatments include lifestyle changes, medication, angioplasty and surgery. To reduce the large scale of deaths from Coronary Artery Disease, an efficient and quick detection technique is to be researched. Various steps are taken to handle the outburst of information related to medical sciences and acquisition of valuable knowledge. Data Analytics and Machine Learning Algorithms play a vital role in this area. This paper presents the Naïve Bayes algorithm, Decision Tree Algorithm using Entropy function, Support Vector Machines and Logistic Regression algorithm for early detection of blocked coronary arteries. The purpose is to enhance the accuracy and enhance the flexibility of the algorithm.

Keywords: *Data Mining, Coronary Artery Disease, Machine Learning, Naïve Bayes Method, Support Vector Machines, Big data analytics and Predictive analytics.*

--***--

1. INTRODUCTION

Researches have been using several data mining techniques in the diagnosis of Coronary Artery Disease. This chronic disease has been on the top of the list for most human deaths in the world for the past 10 years. Data mining and machine learning can help with the prediction of factors that lead to artery blockage and help prevent risks associated with this disease. Naïve Bayes and Support Vector Machines for the classification purpose are being used by most machine learning systems and have successfully employed such methods. However, other machine learning algorithms like Decision Tree and Logistic Regression have proven to be equally effective. Big data analytics can be seen as one of the major emerging sides in the field of medical sciences. Big data along with medical data with the help of predictive analytics is also being used for providing predictive intuitions in the healthcare field and it is also playing an important role in the analysis of chronic diseases. This eventually led the researchers and scientists towards applying their technical revolutions as well as inventions such as “predictive analytics”, “machine learning”, “big data analytics” and “learning algorithms” for gathering worthwhile understanding and support in better decision making. In health care industries, prediction can turn out to be useful as well as successful when the knowledge can be conveyed as action. In this case, we propose a method that gives real-time analyzed report about predicting heart related diseases with the help of historical data and real-time data. Data mining algorithms have a useful role in health care industries in the prediction cum diagnosis of diseases. These data mining applications relate to medical device industries and pharmaceutical industries including hospital management, with the main aim of finding useful and hidden information from the database. The process of knowledge discovery includes developing, understanding, selection, the creation of data and preprocessing of data.

2. METHODOLOGY

2.1 Data Mining Technologies

Extracting non trivial information from medical databases is what data mining technology is best used for. It is the intelligent computational analysis of large sets of data by using a combination of machine learning, statistical analysis and database technology, with the objective to discover patterns and rules useful for guiding decisions about future activities [2], [3] The goal of data mining is predicting and generalizing a pattern to other data. Medical data mining is becoming increasingly important in health care. Data mining with great potential has the powerful and the technology to help organizations focus on the most important information in their data warehouses [9]. Data mining tools predict

https://doi.org/10.15623/ijret.2018.0707026 Received: 15-05-2018, Accepted: 26-06-2018, Published: 17-07-2018

future trends and behaviours, help organizations to make proactive knowledge driven decisions [10]. Various data mining techniques are available with their suitability dependent on the domain application. Applications on Data Mining in health can have tremendous potential and usefulness. The process of finding predictive information in large databases is automated. Data mining classification technology consists of two models such as classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. Testing data set is used for testing the classification efficiency. Patient dataset is collected from cardiac healthcare institutes who have patients showing symptoms of heart disease. Classification algorithms like Naive bayes and Support vector machine are used for prediction to find whether the patient is suffering from heart disease with indicating levels. The following table shows the attributes required during experimentation.

Table 1: Attributes used in the experimentation

Attribute	**Description**
Sex	Classification of Sex
Age	Age of the patient
Family Heredity	Previous History (Father/Mother)
Weight	Weight of the patient
BP	Blood Pressure count
HDL	High Density Lipoprotein
LDL	Low Density Lipoprotein
Triglycerides	Triglyceride fat blood test
LP total Cholesterol	Total cholesterol level

2.2 Factors that Influence in Causing Heart Related Issues

The heart attack occurs when the arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed. Various types of heart diseases are :
1) Coronary heart disease
2) Cardiomyopathy
3) Cardiovascular disease
4) Ischaemic heart disease
5) Heart failure
6) Hypertensive heart disease

Some of the common factors that influence in causing heart issues are as follows:

- Hypertension: It increases the blood pressure.
- Smoking: It can have an effect on the blood pressure and thereby resulting in the to weaken
- Obesity: Obesity can lead to high blood sugar or high blood pressure can eventually result in weakening or damaging arteries
- Excess Dietary Sodium Intake : Increases blood volume which means more work for the heart and more pressure on blood vessels. Over time, the extra work and pressure can stiffen blood vessels, leading to high blood pressure and eventually heart attack.
- Heredity from family: Heart diseases can also be a result of family heredity.

Table 2: Attributes Used for Diagnosis

Attribute Name	**Attribute Role**	**Attribute Type**	**Description**
Regular	Sex	Binomial	Sex of the patient. Takes the following values: Male, Female
Regular	Age	Integer	Age of the patient
Regular	Fam/Heri	Polynomial	Indicates whether the patient’s parents were affected by Coronary Artery Disease. Takes the following values: Father, Mother, Both
Regular	Weight	Numeric	Weight of the patient
Regular	BP	Polynomial	Blood Pressure of the patient
Regular	Natriuretic peptides	Integer	Brain natriuretic peptide, B-type natriuretic peptide (BNP), is a protein that your heart and blood vessels produce.
Regular	Lipoprotein (a)	Integer	Lipoprotein (a), or Lp(a), is a type of LDL cholesterol. Your Lp(a) level is determined by your genes
Regular	HDL	Numeric	High Density Lipoprotein
Regular	LDL	Integer	Low Density Lipoprotein
Regular	VLDL	Integer	Very Low Density Lipoprotein
Regular	Vulnerability	Nominal	Indicates the vulnerability of the patients to heart disease. Takes the following values: High, Low

https://doi.org/10.15623/ijret.2018.0707026 Received: 15-05-2018, Accepted: 26-06-2018, Published: 17-07-2018

2.3 Problem Definition

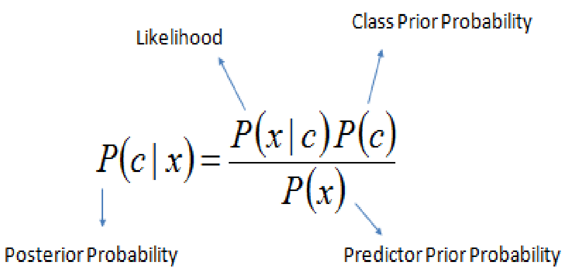
Ischemic heart disease (IHD) is also known as Coronary artery disease (CAD). This condition is within the group of cardiovascular diseases, of which it is the most common type. It refers to a group of diseases which includes stable angina, unstable angina, myocardial infarction, and sudden cardiac death. Coronary artery disease occurs when part of the smooth, elastic lining inside a coronary artery (the arteries that supply blood to the heart muscle) develops atherosclerosis. Due to atherosclerosis, the artery's lining becomes hardened, stiffened, and accumulates deposits of calcium, fatty lipids, and abnormal inflammatory cells to form a plaque. Plaques are like large "pimples" that protrude into the channel of an artery, causing a partial obstruction to blood flow. People with coronary artery disease might have just one or dozens distributed throughout their coronary arteries.

2.4 Proposed System

In this System, our main focus is on the machine learning techniques that help in predicting the chronic coronary artery disease. The predictive models that we will be using for the above scenarios are for example Naïve Bayes, Decision tree, Support vector machine models and Logistic regression.

2.4.1 Naïve Bayes Algorithm

Bayes theorem provides a way of calculating the posterior probability, *P*(*c*|*x*), from *P*(*c*), *P*(*x*), and *P*(*x*|*c*). Naive Bayes classifier assume that the effect of the value of a predictor (*x*) on a given class (*c*) is independent of the values of other predictors. This assumption is called class conditional independence.



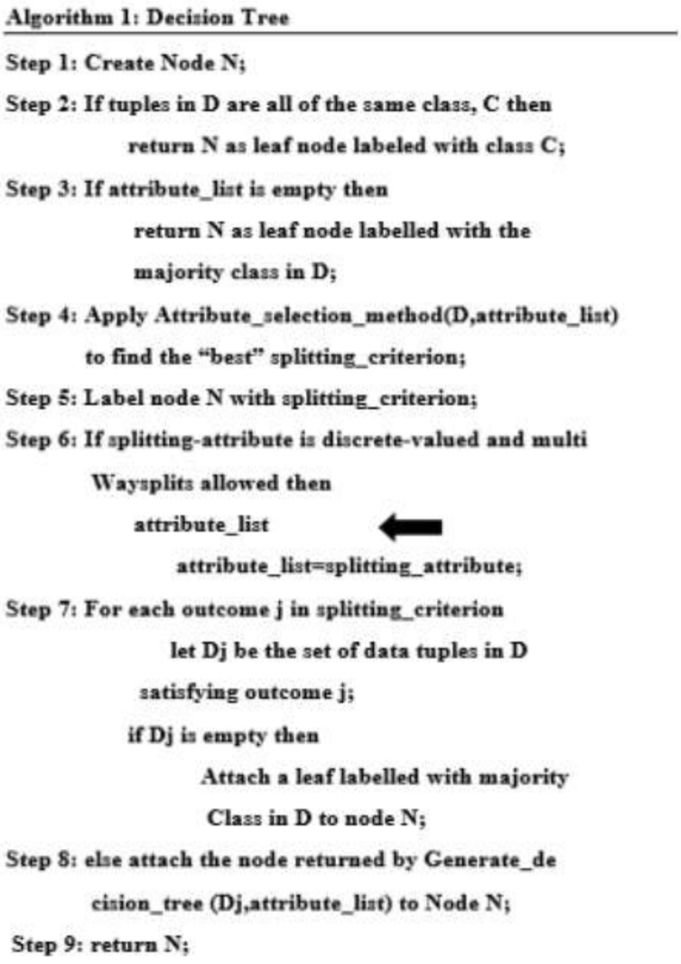
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

2.4.2 Decision Tree

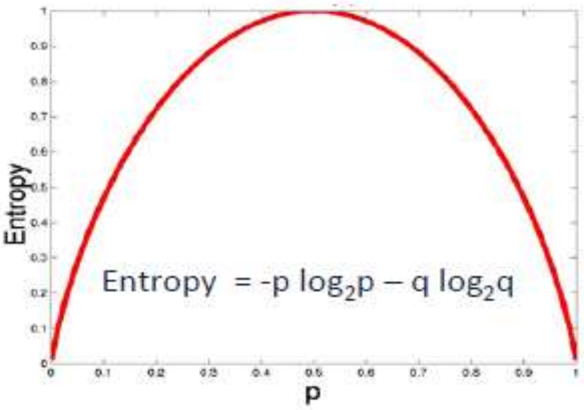
The decision tree algorithm builds regression or classification models in the form of a structured tree. It starts by breaking down a dataset into smaller subsets. An associated decision tree is incrementally developed at the same time when a dataset is being broken down. The final result is a tree with decision and leaf nodes. A decision node is one which contains two or more branches. Leaf node represents a classification or decision. A root node is the topmost decision node in a tree which corresponds to the best predictor. Categorical as well as numerical data can be handled by the decision tree algorithm. ID3 is the core algorithm for building decision trees. A top-down, greedy search through the space of possible branches with no backtracking is employed. ID3 algorithm is often used to construct a decision tree.



```
Algorithm 1: Decision Tree
Step 1: Create Node N;
Step 2: If tuples in D are all of the same class, C then
            return N as leaf node labeled with class C;
Step 3: If attribute_list is empty then
            return N as leaf node labelled with the
            majority class in D;
Step 4: Apply Attribute_selection_method(D,attribute_list)
        to find the "best" splitting_criterion;
Step 5: Label node N with splitting_criterion;
Step 6: If splitting-attribute is discrete-valued and multi
        Waysplits allowed then
            attribute_list ⬅
                attribute_list=splitting_attribute;
Step 7: For each outcome j in splitting_criterion
                let Dj be the set of data tuples in D
            satisfying outcome j;
            if Dj is empty then
                Attach a leaf labelled with majority
                Class in D to node N;
Step 8: else attach the node returned by Generate_de
        cision_tree (Dj,attribute_list) to Node N;
Step 9: return N;
```

A top-down approach using a root node leads to building an entropy decision tree. Partitioning the data into subsets that contain instances with similar values (homogenous) is involved. The homogeneity of a sample is calculated by using the entropy ID3 algorithm. If the sample is completely homogeneous then the entropy value is zero. If the sample is an equally divided sample, then it has entropy of one.

https://doi.org/10.15623/ijret.2018.0707026 Received: 15-05-2018, Accepted: 26-06-2018, Published: 17-07-2018



We need to calculate two types of entropy using frequency table formulae to build an effective decision tree.

A) Entropy using the frequency table of one attribute:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

B) Entropy using the frequency table of two attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

2.4.3 Support Vector Machines

Support Vector Machines helps in predicting the class of the data. It is one of the most commonly used data mining techniques for classification problem.

Using the trained data finding an optimal hyperplane amongst two classes, hyperplane is discovered by Support Vector Machine with the help of optimal problem solution:

$$\max Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

where $0 \leq \alpha_i \leq C$ for $i = 1, 2, \ldots, n$

The digital function f(x) which is defined as a “kernel function” is used by Support Vector Machine in order to calculate the output.

$$f(x) = \operatorname{sign}\left[\sum_{i=1}^{l} \alpha_i d_i K(x, x_i) + b\right]$$

where $K(x, x_i)$ is the kernel function.

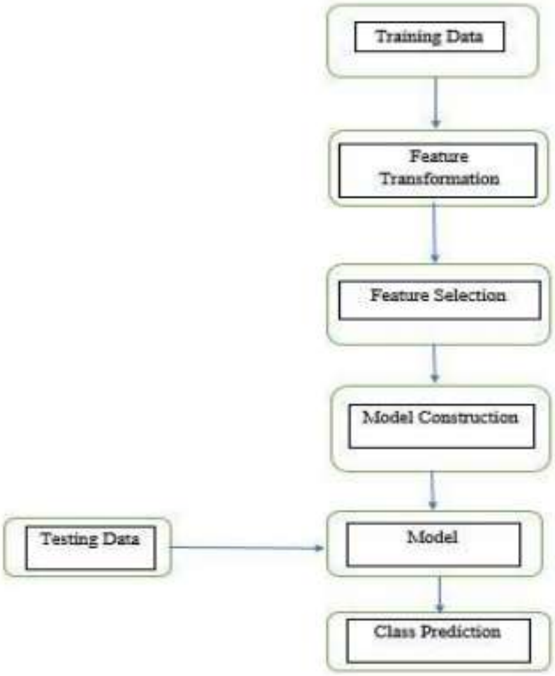
2.4.4 Logistic-Regression

This is an advanced version of a linear regression model. Logistic Regression is used to compute the distribution between an example X and a bool class tag Y via P(X|Y). It classifies the Boolean class label Y as follows:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

In case of our proposed system, we stress on some of the machine learning techniques in order for the prediction of coronary heart diseases. The main agenda of the method that is proposed is to compare the classification performance of Naïve Bayes, decision tree, logistic regression, and support vector machine. The proposed process for constructing the predictive models are given as follows:



1) Training of Data: Transformation of nominal attributes to binary attributes takes place in the first step of the proposed algorithm.
2) Feature Transformation: We make use of the Best-First feature selection method in the second step to choose the feature subsets which will help reduce the training time as well as the number of attributes.
3) The Best-First is a feature selection method where it seeks feature subset space leveraging greedy hill climbing model which is amplified along with backtracking capability.

https://doi.org/10.15623/ijret.2018.0707026 Received: 15-05-2018, Accepted: 26-06-2018, Published: 17-07-2018

4) This step focuses on training the classifier model in order to generate a predictive model to predict the undetected data.
5) The final step focuses on the prediction of the class which includes coronary heart diseases as a result of testing the data.

5. CONCLUSION

Coronary Artery Blockage Detection for given data can be simulated using Naïve Bayes, Decision Tree and Support Vector Machine & Logistic regression algorithms respectively. The problem faced:- Data Accumulation and Segmentation is quite strenuous. The algorithm takes a lot of time to iterate over the large data set. Optimizing an individual algorithm for doing the whole task is crucial. However, predicting early coronary artery blockages only using one individual algorithm will not efficient. Every algorithm has its own pros and cons and to successfully identify artery blockages we have to combine algorithms and leverage their advantages individually to reduce errors.

REFERENCES

[1] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

[2] J. Han Kamber, M. 2006. Data Mining: Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufmann.

[3] U. Fayyad, G.Piatetsky-Shapiro, and P.Smyth, “From Data Mining to Knowledge Discovery in Databases”, AI Magazine, Vol.17, pp.37-54, 1996.

[4] Prerana T H M1, Shivaprakash N C2 , Swetha N3 ”Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS” International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200.

[5] Michael W. Berry et.al,”Lecture notes in data mining”,World Scientific(2006).

[6] A. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013.

[7] Y. C. T. Bo Jin, “Support vector machines with genetic fuzzy feature transformation for biomedical data classification.,” Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[8] World Health Organization. Available: http: // www.who.int/topics/ diabetes mellitus/en/.

[9] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation 10500 Falls Road, Potomac, MD 20854 (U.S.A.),1999.

[10] L.A.Rose, D.T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN O-471-66657-2, John Wiley & Sons, Inc, 2005.

[11] G.Parthiban, A.Rajesh, S.K.Srivatsa, “Diagnosis of Heart Disease for Patients using Naïve Bayes Method”, International Journal of Computer Applications (IJCA) Volume 24-No.3, June 2011, 0975- 8887.

[12] Elkan C. “Naive Bayesian Learning, Technical Report CS97-557”, Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.

[13] Kelly H. Zou, PhD; A. James O‟Malley, PhD; Laura Mauri, MD, M.Sc “ROC Analysis for Evaluating Diagnostic Test and Predictive Models.

[14] G.Suganya, D.Dhivya “Extracting Diagnostic rules from SVM” , Journal of Computer Applications (JCA), 2011.

[15] H.Barakat, Andrew P.Bradley and Mohammed Nabil H.Barakat (2009) “Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus”, IEEE Transactions on Information Technology in Bio Medicine, Volume 14, Issue 4, pp 1-7, 2009. Available: http://ieeexplore. ieee.org /xpls/ abs_ all.jsp?arnumber=5378519 Digital Object Identifier: 10.1109 /TITB. 2009.2039485.