A Model-Based Approach of Data Analysis and Prediction in Cardiovascular Disease

Rohini Shetty, Tanmayee Mandala, Roberto Salazar
Department of Systems Science and Industrial Engineering, The State University of New York at Binghamton, NY 13902

Abstract

During recent years, the number of machine learning and data mining algorithms applied into healthcare systems for disease's diagnosis and prediction has increased significantly. Their effective application can lead to significant costs savings related with treatment management, hospital admissions and drugs. This research is targeting cardiovascular disease, which is a serious chronic disease that affects the heart's performance and the individual's quality of life. However, cardiovascular diseases can be predicted based on the patient's health attributes and conditions. In the pursuit of improving the diagnosis process of cardiovascular disease, multiple data mining models have been proposed on the literature. The developed models in this study are based on the Cardiovascular Disease dataset published by Ulianova. This study suggests a model-based framework that utilizes multiple data mining techniques to be used as decision support tools in the diagnosis of cardiovascular disease. The proposed model, after the training and testing of multiple classification methods, utilizes a random forest as a prime classifier that can help researchers and physicians improve the prediction process of cardiovascular disease, and thus, to get more accurate results and better healthcare outcomes for patients. The validation of the developed models shows accuracy, sensitivity, recall and F-measure results.

Keywords

Classification, machine learning, supervised learning, cardiovascular disease

1. Introduction

Cardiovascular diseases (CDVs) are a disorder or malfunction in blood vessels and heart; examples include coronary artery disease (the most common type of heart disease in the United States), cerebrovascular diseases, deep vein thrombosis, pulmonary embolism (which create blood clots in the leg) and congenital heart disease [1]. Most cardiovascular diseases are the result of lifestyle followed by an individual that may include an unhealthy diet, tobacco usage, alcohol consumption, low physical activity and/or hypertension. According to the National Center for Chronic Disease Prevention and Health Promotion, heart diseases represent one of the top leading causes of Americans' deaths and disabilities and one of the leading drivers of the United States' $3.5 trillion in annual healthcare costs [2]. In the same way, according to [3, 4], about 647,000 Americans die from heart disease each year (corresponding to 1 in every 4 deaths) and about 18.2 million adults age 20 and older have a coronary artery disease. Numerous researches are being conducted to find out the correlation between the multiple causes of CVDs. Understanding them can result in the development of effective cardiovascular diseases' prevention plans before they reach a level where more costly treatments would be necessary to reduce health risks and mortality rates. However, the previous leads to the need for better diagnosis and prediction tools. The application of data mining classification tools in the healthcare industry has been effective for the detection of causes of diseases [5]. Thanks to the vast amount of data generated and collected by the healthcare industry about patients' information, in addition to recent advances in the field of data science and machine learning, nowadays it is possible to develop robust classification and prediction models that could lead to better health outcomes. The objective of this paper consists in developing multiple machine learning classification models (i.e. logistic regression, support vector machine, *k* nearest neighbor, decision tree, random forest and Gaussian Naïve Bayes classifier) capable of predicting and identifying the presence of cardiovascular disease in patients based on multiple medical features. The focus of this study will be to compare the models' performance by obtaining key process indicators after a 5-fold cross validation analysis, in order to identify the one with the highest accuracy rate, and thus, the one that can lead to better decision making and better medical outcomes. This will constitute a decision-making support system for physicians and researchers to help them predict cardiovascular diseases in patients in advance.

2. Related Work

The following literature review will discuss cardiovascular diseases, data mining and classification methods, which are the core contents of this study.

2.1 Cardiovascular Disease

Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels [6]. Within the most common cardiovascular diseases (i.e. diseases that involve the heart) are cardiomyopathy, hypertensive heart disease, heart failure, pulmonary heart disease and inflammatory heart disease. In the same way, within the most common vascular diseases (i.e. diseases that involve blood vessels) are coronary artery disease, peripheral arterial disease, cerebrovascular disease, renal artery disease and aortic aneurysm.

Some of the major causes that contribute to the development of CVDs include high doses of sugar consumption (i.e. diabetes), overweight and obesity, unhealthy or unbalanced diets, excessive smoking, alcohol intake, minimal or no physical activity, hypertension, high blood pressure and high blood cholesterol levels. According to the World Health Organization [1], most CVDs can be prevented by addressing these behavioral risk factors. In the same way, [1] states that people with CVD or who are at high CVD risks due to the presence of one or more of the risk factors mentioned above, need early detection and management using counselling and medicines, as appropriate. The causes, symptoms, and treatment related with cardiovascular disease remain an active field for research between physicians, biomedical research and experimental medicine.

Currently, CVD is the leading cause of death in the United Sates and constitutes to 17% of the overall national health expenditures [7, 8, 9]. During the past decade, the medical costs associated with CVD have grown at an average annual rate of 6% and have accounted for approximately 15% of the increases in medical spending [10]. According to a research conducted to forecast the future of CVD in the United States [11], by 2030, 40.5% of the United States population is projected to have some form of CVD; between 2010 and 2030, real total direct medical costs of CVD are projected to triple from $273 billion to $818 billion while real indirect costs (due to lost productivity) for all CVD are estimated to increase from $172 billion to $276, an increase of 61%. Despite these increasing and alarming trends, there are multiple opportunities for improving CDV diagnosis, prediction and prevention while controlling medical costs through data analytical and machine learning models, respectively.

2.2 Data Mining and Classification Methods

Data mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models [12]. It can be considered as a relatively recent field with prominence in the year 1994 [12]. The healthcare industry has been historically considered as data-rich information-poor. One of its main challenges consists in transforming raw data into information, knowledge and business insights to generate added value to the services provided. During recent years, the application of data mining tools has been expanding within the healthcare industry, leading to more effective treatments, better healthcare management, better customer relationship management, and a more effective identification of potential frauds and abuses. Some of these tools include regression and classification models, respectively.

Classification models sort data into specific categorical classes through supervised learning (i.e. knowing in advance the outcome for particular data points). Their development consists in two main steps: training and testing. During the training process, the classification model “learns” from a subset of the original data set (i.e. training set) and creates rules within. During the testing set, once the classification model has completed its “learning” process, it is tested and evaluated on the remaining data from the original data set (i.e. testing set) to determine how capable it is to accurately classify new data entries. Since the testing set tends to contain fewer data points than the training test (usually by a 0.7/0.3 ratio), it requires less computational expense. While there are multiple classification techniques and algorithms for a diverse number of applications, the ones used and developed for this study were logistic regression, support vector machine, *k* nearest neighbors, decision tree, random forest and Gaussian Naïve Bayes, respectively.

2.2.1 Logistic Regression

Logistic regression is used for diagnosis for prediction of any disease presence or for prognosis and disease outcomes. Standard (maximum likelihood) logistic regression has been used by sixty-four studies, while penalized logistic regression (lasso, ridge, or elastic net) has been used in nine studies; bagged logistic regression classified as machine learning classifier have also being used [13]. It enables to establish a probabilistic relationship with either of

the classes in the data. In logistic regression the decision boundary enables a clear visualization of the observations in either of the two classes in a two-dimensional space.

2.2.2 Support Vector Machine

The crucial point of SVM is to handle the objects which are to be classified as points in a high-dimensional space and to find a line (hyperplane) that separates them. Molecular descriptors help molecules present in the space. The distance from the separating hyperplane to the nearest data point is known as margin of the hyperplane while SVM locates the maximum margin separating the hyperplane. The ability of SVM increases to predict the accurate classification of new compounds with selection of hyperplane [14].

2.2.3 *K* Nearest Neighbors

KNN classification model classifies each observation in the dataset based on its nearest neighbor. The algorithm works based on calculating the distance for the training and testing dataset, finding the closest neighbor and mapping the input data point to the closest class. Distances can be measured in different ways (e.g. Euclidean, Manhattan, Hamming or Minkowski distance). Some advantages of using KNN are that the training time is shorter and easier to implement, while a disadvantage for this method is that testing process might consume preferably more time than training and requires large storage space [15].

2.2.4 Decision Tree

Decision tree is a representation similar to that of a flowchart, which consists of root node which represents features/attributes from the dataset and the branch from the root node is expected to make a decision (yes/no) and the leaf node from the decision branch indicates the outcome of that root node. The principle to be applied in the selection process is to choose those variables that best split the root node into distinct groups. If the samples in a node belong to the same group, then that node becomes a leaf node. Each path from root node to leaf node represents an if-then situation called a rule which has been described by [16].

2.2.5 Random Forest

The concept of bagging of decision tree is called a random forest classification model. The standard principle random forests is to build binary sub-trees using the training bootstrap samples coming from the learning sample L and selecting randomly at each node a subset of X. The classification with highest votes among all the trees in the forest opts by decision forest. It provides variable importance for classification tasks [17].

2.2.6 Gaussian Naïve Bayes

It applies Bayes theorem with considering assumptions of Naïve of independence between every pair of features. It calculates the probability that a given instance belongs to a certain class. Gaussian Naïve Bayes implements taking the considerations of likelihood of features [18]. Gaussian Naïve Bayes model computes the standard deviation and mean of the input values for every class.

3. Data Description

The Cardiovascular Disease dataset contains information about 70,000 patients with 11 features (ID not considered) and a single target, as detailed in Table 1. The dataset was published in [19]. This paper focuses on studying the cardiovascular disease presence target in sick patients.

Table 1: Features and their types

Feature	Measure	Classification	Type
ID	NA	Identification feature	Integer
Age	days	Objective feature	Integer
Height	cm	Objective feature	Integer
Weight	kg	Objective feature	Float
Gender	1 – women 2 – men	Objective feature	Categorical
Systolic blood pressure	ap_hi	Examination feature	Integer
Diastolic blood pressure	ap_lo	Examination feature	Integer
Cholesterol	1 – normal 2 – above normal	Examination feature	Categorical

	3 – well above normal		
Glucose	1 – normal 2 – above normal 3 – well above normal	Examination feature	Categorical
Smoking	0 – does not smoke 1 – smokes	Subjective feature	Binary
Alcohol intake	0 – does not consume 1 – consumes	Subjective feature	Binary
Physical activity	0 – does not exercise 1 – exercises	Subjective feature	Binary
Presence or absence of cardiovascular disease	0 – healthy 1 – sick	Target variable	Binary

4. Methodology

This section will provide the framework used throughout the development of the study, which consisted of three main steps (i.e. data cleaning and pre-processing, classification methods selection and key performance indicators) to reach the end goal of the project: developing and identifying the best machine learning classification model capable of predicting and classifying cardiovascular disease in patients.

4.1 Data Cleaning and Pre-Processing

The cardiovascular disease data set used for the development of this study did not contain any missing values or information regarding the patients' features. For this reason, data cleaning was not required. However, the first column from the data set (i.e. ID column) contained a redundant value for purposes of this study, reason why it was removed. On the other hand, taking into consideration that data pre-processing is a significant process in every data mining project [20] that leads to better performances (i.e. higher accuracy, precision, sensitivity rates) in some machine learning classification models (e.g. *k* nearest neighbors and artificial neural networks), it was applied in the development of the *k* nearest neighbor classifier through data scaling. Data scaling is highly recommended for when the predictor variables may have significant different ranges and when features need to be unit-independent (i.e. not reliant on the original scale of the measurements involved). It is worth to specify that since the number of features contained in the data set was not large, no feature selection methods were applied.

4.2 Classification Methods Selection

In the current literature, information about multiple machine learning classification methods can be found, which continues to increase with the introduction of new methods and variations/improvements of existing ones. For this study, logistic regression, support vector machine, *k* nearest neighbor, decision tree, random forest and Gaussian Naïve Bayes were selected based on their popularity, good performance in the literature and their previous applications in medical and healthcare systems fields.

4.3 Key Performance Indicators

Six classification methods were selected. However, taking into consideration that their performances–most likely–would vary between them, four key performance indicators (i.e. accuracy, sensitivity (recall), precision and F-measure) were calculated, as shown in Table 3.

Table 2: Nomenclature

Acronym	Meaning	Definition
TP	True positive	A sick patient identified as sick
TN	True negative	A healthy patient identified as healthy
FP	False positive	A healthy patient identified as sick
FN	False negative	A sick patient identified as healthy

Table 3: Performance measures

Key Performance Indicator	Formula	Definition
Accuracy	(TP + TN) / (P + N)	Test's rate of correct prediction for both classes

Sensitivity (Recall)	TP / (TP + FN)	Test's ability to correctly detect sick patients who have the disease
Precision	TP / (TP + FP)	Sick patients predicted value
F-measure	(2 * recall * precision) / (recall + precision)	Harmonic mean that combines precision and recall

The six machine learning classification models were programmed using Python programming language. The scikit-learn library [21], a machine learning library in Python with simple and efficient tools for predictive data analytics built on Numpy, SciPy and matplotlib Python libraries, was primarily used for the models' programming and testing. The computational environment for this was a processor 2.50 GHz, 8 GB RAM and 64-bit operating system, Windows 10.

5. Experimental Results

5.1 Performance Results (KPI's)

In order to determine the classification performance for each of the models, key performance indicators were calculated for each one, respectively. The results are shown in Table 4, which includes accuracy, sensitivity (recall), precision and F-measure for each classification method, using the average of five-folds of cross validations. Cross validation uses a limited data sample (i.e. one *k* group) to estimate the skill of machine learning models on unseen data. It results in less biased estimates than a regular train/test split, thus it gets a more accurate performance of the classification models.

Table 4. Models' performance evaluation

Performance Measure	**Logistic Regression**	**Support Vector Machine**	***K* Nearest Neighbor**	**Decision Tree**	**Random Forest**	**Gaussian Naïve Bayes**
Accuracy	69.5 %	60.4 %	71.0 %	63.3 %	71.6 %	59.2 %
Precision	70.9 %	61.4 %	73.6 %	63.3 %	72.3 %	73.0 %
Sensitivity (recall)	66.1 %	55.9 %	65.7 %	63.3 %	70.1 %	29.4 %
F-measure	68.4 %	58.5 %	69.4 %	63.3 %	71.2 %	41.9 %

Based on the results from Table 4, it can be stated that the random forest had the highest accuracy, sensitivity (recall) and F-measure levels, while the *k* nearest neighbor classifier had the best precision level. On the other hand, it can be stated that the Gaussian Naïve Bayes classifier got the worst results, which means it is not capable of classifying correctly patients with cardiovascular disease from patients without cardiovascular disease.

5.2 *K* Nearest Neighbors

In order to obtain the optimal number of *k* neighbors during the development of the *k* nearest neighbor classifier, the original dataset was split into a training set and a testing set using a 0.7/0.3 ratio. The *k* nearest neighbor classification algorithm was trained and tested using 40 different *k* values (i.e. from $k = 1$ to $k = 40$) and the error rate was plotted, obtaining the elbow plot on Figure 1. The elbow method is a heuristic method used in cluster analyses to help determine an appropriate number of clusters in a dataset by looking at the percentage of variance explained as a function of the number of clusters [22]. According to Figure 1, the error rate tends to decrease after incrementing the *k* value until $k = 27$, where it reaches its minimum in the range of from $k = 1$ to $k = 40$. For this reason, the optimal number of *k* neighbors selected that was used during the cross-validation analysis was 27.

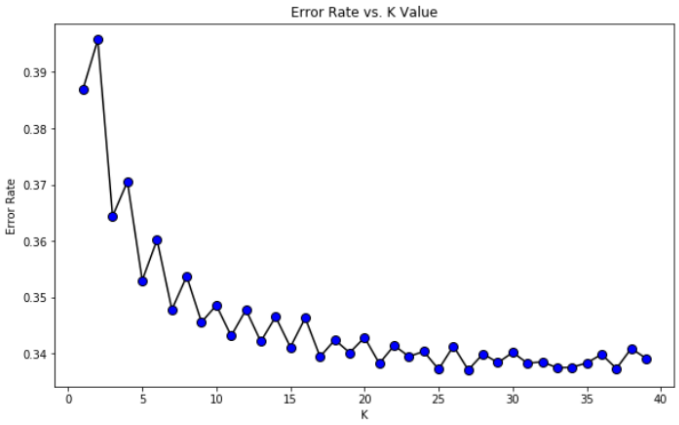

Figure 1: Elbow plot for KNN

6. Conclusion

This study consists in the application of classification techniques and algorithms to classify patients with cardiovascular disease given multiple health features. The data set used in this study contained information of 70,000 patients, which can be considered as appropriate to justify the outcomes from the study. Six classification algorithms were introduced: logistic regression, support vector machine, *k* nearest neighbors, decision tree, random forest and Gaussian Naïve Bayes classifier. As a result, the random forest model had the best performance in terms of accuracy, sensitivity (recall) and F- measure; in terms of precision, it got 1.3% less than the model with the highest precision rate (i.e. *k* nearest neighbor), but still quite close to it. An accuracy rate of 71.6% is quite acceptable for the prediction of cardiovascular diseases; however, due to their significant impact and risk on people's health, it should be improved. Obtaining machine learning classification models with higher accuracy rates could allow better decision making for improving patients' health and reducing future risks. As a feature research, it would be recommended to gather patients' data regarding new features since some of them might have significant correlation on the development of cardiovascular disease, such as race or ethnicity, family history, diet, stress and amount of sleep, respectively.

References

1. The World Health Organization, "Cardiovascular Disease (CVDs)". *The World Health Organization*, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP), "Heart Disease and Stroke", *National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP)*, https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm
3. Benjamin, Emelia J., Paul Muntner, and Márcio Sommer Bittencourt. "Heart disease and stroke statistics-2019 update: a report from the American Heart Association." *Circulation* 139.10 (2019): e56-e528.
4. Fryar, Cheryl D., Te-Ching Chen, and Xianfen Li. *Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010*. No. 103. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.
5. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5.5 (2013): 241-266.
6. Mendis, Shanthi, et al. *Global atlas on cardiovascular disease prevention and control*. Geneva: World Health Organization, 2011.
7. Trogdon, Justin G., et al. "The economic burden of chronic cardiovascular disease for major insurers." *Health promotion practice* 8.3 (2007): 234-242.
8. Roger, Véronique L., et al. "Heart disease and stroke statistics—2011 update: a report from the American Heart Association." *Circulation* 123.4 (2011): e18-e209.
9. Cohen, Joel W., and Nancy A. Krauss. "Spending and service use among people with the fifteen most costly medical conditions, 1997." *Health Affairs* 22.2 (2003): 129-138.
10. Roehrig, Charles, et al. "National Health Spending By Medical Condition, 1996–2005: Mental disorders and heart conditions were found to be the most costly." *Health Affairs* 28.Suppl1 (2009): w358-w367.
11. Heidenreich, Paul A., et al. "Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association." *Circulation* 123.8 (2011): 933-944.

12. Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." *Journal of healthcare information management* 19.2 (2011): 65.
13. Jie, M. A., et al. "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models." *Journal of clinical epidemiology* (2019).
14. Bikadi, Zsolt, et al. "Predicting P-glycoprotein-mediated drug transport based on support vector machine and three-dimensional crystal structure of P-glycoprotein." *PloS one* 6.10 (2011): e25815.
15. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5.5 (2013): 241-266.
16. Ramezankhani, Roghieh, et al. "Application of decision tree for prediction of cutaneous leishmaniasis incidence based on environmental and topographic factors in Isfahan Province, Iran." *Geospatial health* 13.1 (2018).
17. Kumar, Dinesh. "Evolving Differential evolution method with random forest for prediction of Air Pollution." *Procedia computer science* 132 (2018): 824-833.
18. Lou, Wangchao, et al. "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes." *PloS one* 9.1 (2014): e86703.
19. Ulianova, Svetlana. "Cardiovascular Disease Dataset". *Kaggle*, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset. Accessed 20 November 2019.
20. Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5.5 (2013): 241-266.
21. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12.Oct (2011): 2825-2830.
22. Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN." *International Journal of Computer Applications* 105.9 (2014).