Cardiovascular Disease Analysis Using Supervised and Unsupervised Data Mining Techniques

Fabio Mendoza Palechor*, Alexis De la Hoz Manotas, Paola Ariza Colpas , Jorge Sepulveda Ojeda, Roberto Morales Ortega, Marlon Piñeres Melo
Universidad de la Costa, Barranquilla, Atlantico, Colombia.

* Corresponding author. Tel.: 318929611; email: fmendoza1@cuc.edu.co
Manuscript submitted September 25, 2016; accepted November 2, 2016.
doi: 10.17706/jsw.12.2.81-90

Abstract: Cardiovascular diseases are the main cause of death around the world. Every year, more people die from these diseases than from any other cause. According to World Health Organization data, in 2012 more than 17,5 million people died from this cause, and that represents 31% of all deaths registered worldwide. Data mining techniques are widely used for the analysis of diseases, including cardiovascular conditions, and the techniques used in the proposed method in this research are decision trees, support vector machines, bayesian networks and k-nearest neighbors. Apart from the previous techniques, it was necessary to use a clustering method for data segmentation according to their diagnosis. As a result, the Simple K-Means clustering method and the support vector machines technique obtained the best levels in metrics such as precision (97%), coverage (97%), true positive rate (97%) and false positive rate (0.02%), and this can be taken as evidence that the proposed method can be used assertively as decision making support to diagnose a patient with cardiovascular disease.

Key words: Bayesian networks, cardiovascular disease, K-Nearest neighbor, data mining, decision trees, support vector machines.

1. Introduction

Cardiovascular diseases are the main cause of death around the world. Every year, more people die from these diseases than from any other cause. According to World Health Organization data, in 2012 more than 17,5 million people died from this cause, and that represents 31% of all deaths registered worldwide. Based on those statistics, 7,4 million deaths were caused by coronary cardiopathy and 6,7 million deaths from cardiovascular conditions, the WHO manifest that three quarters of deaths due CVDs can be found in countries with low and medium income. For people with CVD or high cardiovascular risk (presence of one or more risk factors, such as arterial hypertension, diabetes, hyperlipidemia or any other CVD confirmed) is extremely important early detection and treatment [1].

Currently, several studies have focused in cardiovascular disease diagnosis, with the objective of predicting or detecting in advance the risks of contracting some type of CVD. An example of this can be found in [2], they propose a study called Coronary Risk Factors and plaque morphology in men with coronary disease who died suddenly. Current research has implemented several techniques or methods configured to achieve the best precision levels in classification, guaranteeing a reliable system to be used for decision making to diagnose in assertive manner the presence of CVD.

In [3] its mentioned the naked ear and the stethoscope were big help in classification of most heart diseases, especially the ones related to valves. They suggest that sounds can be a diagnose technique for cardiac diseases

using Hidden Markov Models (HMM).

In [4] several computational intelligence techniques are studied for heart disease detection, and their study is based in comparing six known classifiers with the data provided by the University of Cleveland. In that research the method with better precision levels was support vector machines (89,49%) surpassing Decision Trees, K-Nearest neighbors, AdaBoots, and Part. These results can be confirmed in [5], through experimental results obtained a classification precision of 77% using the method of logistic regression.

According to [6], diagnosis of cardiac disease it is a relevant issue and many researchers have developed intelligent decision support systems to improve the capacity of medical staff. In that research, a novelty methodology is presented using SAS 9.1.3 software. The results had 89.01% of precision based on the heart disease dataset of the University of Cleveland, and also obtained 80,95% in sensitivity and 95,91% in specificity in the diagnosis of heart conditions.

In [7] a new diagnosis intelligent system is introduced, based on the fuzzy inference system (ANFIS) for heart valve disease. The accuracy of classification of this intelligent system PCA-ANFIS was 96% for normal subjects and 93,1% for abnormal subjects.

In [8] a classification technique study was proposed, comparing them to select the one with best prediction levels for coronary artery disease (CAD). The analysis was conducted with 1245 subjects (865 CAD present and 380 CAD absent), and the techniques used were: logistic regression (RL), regression and classification trees (CART), multilayer perceptron (MLP), radial base function (RBF) and self-organizing feature map (SOFM). The best levels were achieved with MLP with a 93.4% in sensitivity.

Based on the previous results of current literature, it can be established a model to be used as a tool for pattern recognition and prediction on patients that could suffer from cardiovascular disease, and certainly it can be also applied to other health care sectors.

2. Materials and Methods

2.1. Dataset Preparation and Analysis

For this study, the dataset selected was “Heart Disease Dataset” and it’s located in the Machine Learning Repository UCI [9]. This dataset has 14 attributes and 303 records. The description of all attributes of the dataset are:

- Age: values of the age of a person in years.
- Sex: male takes the value of 1 and female is 0.
- Chest Pain Type: In case of typical angina the value is1, atypical angina is 2, other kind of pain is 3 and asymptomatic is 4.
- Resting Blood pressure: Value calculated in Hg at the time of hospital admission.
- Cholesterol: mg/dl.
- Blood sugar > 120 mg/dl: 1, in case is true, 0 otherwise.
- Electrocardiogram Result: In case of normal value, is 0, anomaly is 1 and ventricular hypertrophy is 2.
- Maximum heart rate achieved
- Exercise induced angina: In case of negative is 0, otherwise is 1.
- Induced depression.
- Slope peak exercise:
- Number of major vessels (0 - 3) colored by fluoroscopy.
- Thal: in case of normal value is 3, by default is 6, in case of reversible defects is 7.
- Heart disease diagnosis (angiographic disease status): 0 in case of less than 50%, 1 otherwise.

2.2. Decision Trees

It’s considered one of the most widely used data mining algorithms in classification and prediction problems. Each interior node maps to an input variable and its divided in children nodes based on the input variable values.

Each leaf node represents a particular value of an output variable. When a decision tree is executed, samples in each interior node are divided in subsets based on an attribute, and this process is repeated in each subset of a recursive partition. In every step, during growth of the decision tree, one of the input variables is selected for division samples. With the chosen variable, the new partitioning point is determined through a value test of the attribute, and the most common tests are impurity and entropy [10].

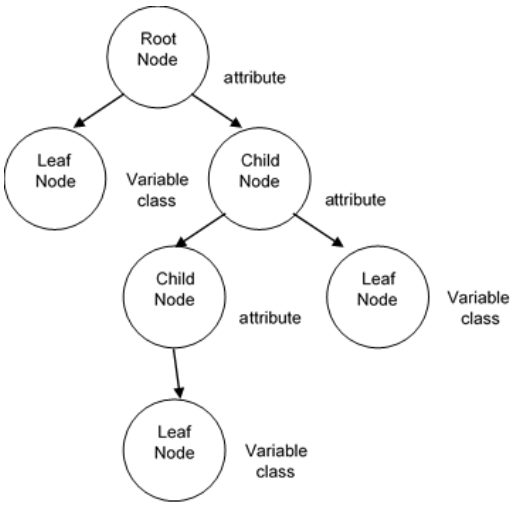

Fig. 1. Decision tree structure. Source: [11]

2.3. Support Vector Machines (SVM)

Commonly used to solve prediction and classification problems in efficient way due to its automatic learning system. They are based in the statistic learning system developed by [12], when a mathematic model is proposed for regression and classification problems [13].

Other authors mention that SVM is a margin classifier that gets trained by a dataset with feature vectors. SVM tries to find an optimal limit that separates two classes with different feature vectors with a maximal margin (distance between optimum hyperplane and the nearest vector). To make classification of an inseparable dataset, a nonlinear SVM projects a feature vector in a high dimensional space using a kernel function such as radial basis kernel function [14].

The construction of support vector machines (SVM) is based in transforming or projecting a dataset in a given n dimension to higher dimension space applying a kernel function – kernel trick. From this new space created, the data is operated as a linear problem, solving it without considering the data dimensionality [15].

Some advantages de support vector machines are: First, it has a solid mathematic foundation. Second, it has the concept of structural risk minimization [16], [17], that translates into the minimization of the probability of a wrong classification on new examples. This case is very common when there are too few data for training. The third advantage relies in the availability of powerful tools and algorithms to find the solution in fast and efficiently [18].

2.4. Naive Bayes

Bayesian networks are considered an alternative to classic expert systems oriented to decision making and prediction under uncertainty in probabilistic terms [19]. In [20], [21] a structure composed by four levels is shown. In the highest level would be a set of variables mapped by nodes and arrows that relate with influence terms. In the next level you would find the levels or states, also known as state space [22], [23] that can take each of the model variables. Third place, you can find a set of conditional probability functions, one for each node, and represents the probability of occurrence of each state of the variable conditioned to possible values. Lately, in the lowest level, would be a set of algorithms that would allow to the network to recalculate the probabilities

assigned to each of the levels when some evidence from the model is known.

2.5. K- Nearest Neighbors

It's an algorithm used for classification and data regression. The algorithm saves all known cases and makes a classification or assign a property to new cases bases on similar features [22]. This method must be one of the first choices for a classification study when there are no previous, or any, knowledge about the data distribution. The method was developed due to the need of performing discriminant analysis when it´s unknown of difficult to establish reliable parametric estimates of probability densities [23].

The main difficult of this method is to set the value of k, since if k takes a higher value than it should, the risk of classifying according to the majority of the data is too high, and if the k value is too low, there can be a lack of precision in classification due to too few data selected as comparison instances [24].

One of most important advantages of the k-NN method is that can change radically its classification results without modifying its structure, only changing the metric used to find the distance. The metric must be selected according to the problem at hand. The great advantage of being able to vary metrics, is that you can obtain different results without transforming the algorithm of the method, only changes are made in the procedure of measuring distances.

3. Experimentation

The present research was based in the dataset "Heart Disease Dataset" located in the Machine Learning Repository UCI [9]. The dataset has 14 attributes and 303 instances. The purpose of the study is to compare different data mining techniques, that are commonly used in task such as classification, prediction and segmentation.

First, the data preparation was conducted to train and evaluate the proposed method, then an analysis of data mining tools was made, taking into account the availability of the methods of the Weka tool and their license and terms of use.

After tool selection and data preparation, the selection of the clustering method to use was decided, based on the need of grouping users if they have a heart disease or not. To do that, the method Simple K-Means was used, and the results are found in Table 1.

Table 1. Data Aggregation Results with the Chosen Clustering Method
Source: Created by Author

Cluster	Number of records
Cluster 0	166 (55%)
Cluster 1	137 (45%)

In table 1, you can see the amount of users that belong to each cluster, where cluster 0 are the patients with no diagnosis of CVD, while cluster 1 are those patients with a heart disease.

With the data prepared and segmented, next step was to select the classification methods to be used and compared later on, the methods chosen were: Decision Trees, Support Vector Machines, Naïve Bayes and Lazy IBK. The selected metrics were True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision and Recall, using the equations shown next.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$TpRate = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falso\ Negativo} \quad (3)$$

$$FpRate = \frac{Falsos\ Positivos}{Falsos\ Positivos + Verdaderos\ Negativos} \quad (4)$$

The data distribution for the training and testing process was made using crossed validation, the tool selects a percentage of the data for training and another for testing with the proposed method. Then, each method is tested and compared with the metrics mentioned previously.

Finally, with the results of the previous phase, the best method was selected and a test file was used to check the results with the best classifier selected. The steps of the research process can be found in Fig. 2.

Fig. 2. Methodology used in the study.
Source: Created by Author.

4. Result

In the present research, the starting point was data preparation, then it was very important to choose the clustering technique, given that the segmentation selection strongly affects the results of the classifiers subject to analysis. In the next graphs, the results of each classifier studied are presented.

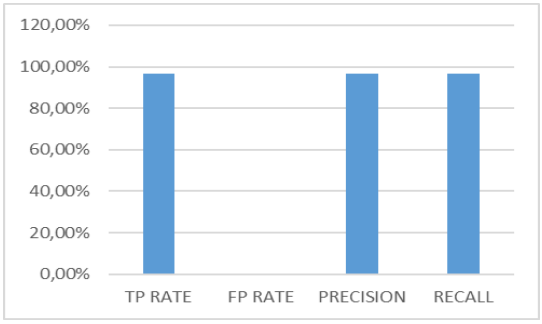

Fig. 3. J48 algorithm results.
Source: Created by Author.

In Fig. 3 the results of the J48 decision trees algorithm can be found, with each metric the results were: 96.70% (TpRate), 0.34% (FpRate), 96.70% (Precision), 96.70% (Recall).

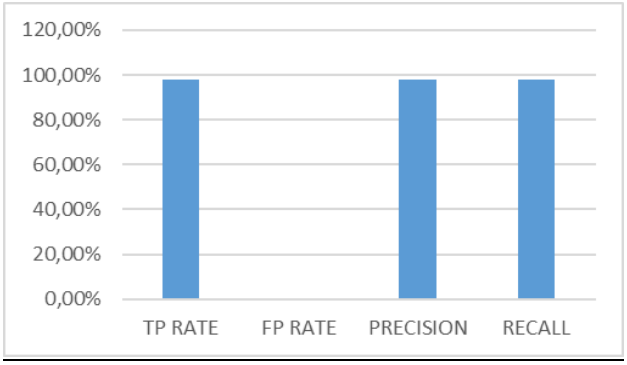

Fig. 4. SMO algorithm results.
Source: Created by Author.

In Fig. 4 the results of the support vector machines algorithm can be found, with each metric the results were: 97.70% (TpRate), 0.02% (FpRate), 97.70% (Precision), 97.70% (Recall).

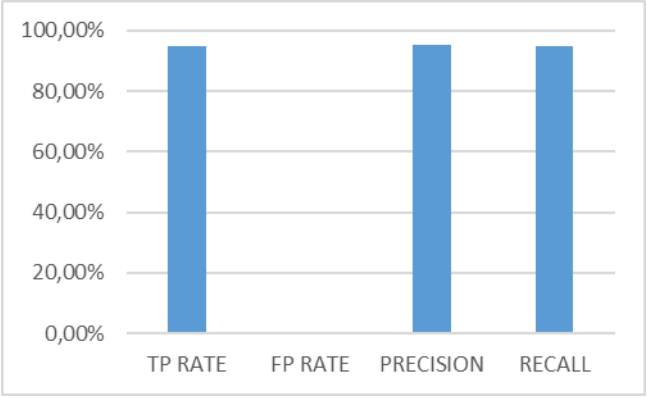

Fig. 5. Naive bayes algorithm results.
Source: Created by Author.

In Fig. 5 the results of the Naïve Bayes algorithm can be found, with each metric the results were: 95% (TpRate), 0.54% (FpRate), 95.10% (Precision), 95% (Recall).

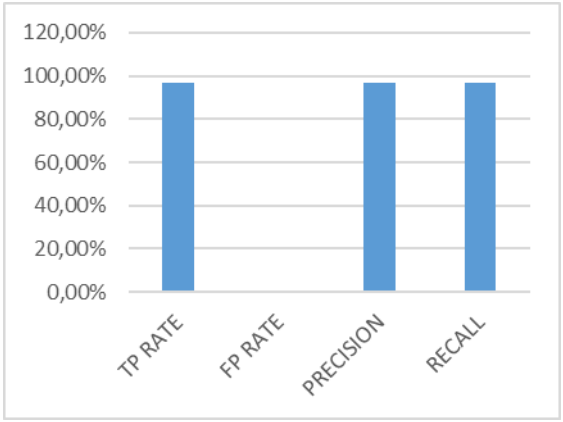

Fig. 6. Lazy IBK.
Source: Created by Author.

In Fig. 6 the results of the Lazy IBK algorithm can be found, with each metric the results were: 97% (TpRate), 0.32% (FpRate), 97% (Precision), 97% (Recall).

Table 2. Classifier Algorithm Results
Source: Created by Author

Method	TPRATE	FPRATE	Precision	Recall
J48	96.70%	0.34%	96.70%	96.70%
SMO	97.70%	0.02%	97.70%	97.70%
Naïve Bayes	95.00%	0.54%	95.10%	95.00%
Lazy IBK	97.00%	0.32%	97.00%	97.00%

Fig. 7. Classifiers algorithm results.
Source: Created by Author.

In Fig. 7 you can find the comparison of techniques used in the study, and it's quite easy to see that the best methods achieve the best percentages in all selected metrics, which led us to conclude that these methods can be used for classification processes in cardiovascular diseases based on the evaluated data. This structure of this method is found in Fig. 8.

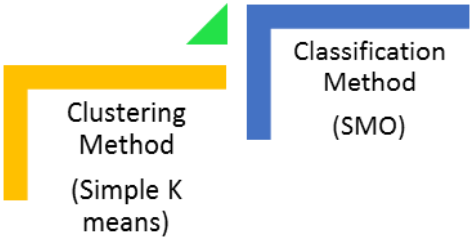

Fig. 8. General Structure of the proposed method.
Source: Created by Author

5. Conclusion

This study proposed a method for classification processes in the diagnosis of patients that can suffer from cardiovascular disease (CVD), using as source the "Heart Disease Dataset", starting with the segmentation phase of the patient data through the Simple K-Means method, and finally, different data mining techniques were applied such as decision trees, support vector machines, bayesian networks and k-nearest neighbors. These techniques were compared and the best results obtained came from the method support vector machines, based on the selected metrics, TpRate 97.70%, FpRate 0.02%, Precision 97.70%, and Recall 97.70. These results clearly show that the propose method has a high percentage in the evaluated metrics which its evidence to indicate that it´s an efficient and accurate method for the diagnosis of cardiovascular diseases, improving the results obtained in previous studies such as [4]-[6].

Acknowledgment

We recognize the great contribution of the repository machine learning uci, (center machine learning and intelligent systems) for providing data sets of information that allows submit a study different methods or techniques of data mining,

References

[1] Centro, P. (2015). Organizacion Mundial de la Salud. Retrieved from the website: http://www.who.int/mediacentre/factsheets/fs317/es/

[2] Burke, A. P., Farb, A., Malcom, G. T., Liang, Y. H., Smialek, J., & Virmani, R. (1997). Coronary risk factors and plaque morphology in men with coronary disease who died suddenly. *New England Journal of Medicine*, *336(18)*, 1276-1282.

[3] El-Hanjouri, M., Alkhaldi, W., Hamdy, N., & Alim, O. A. (2002). Heart diseases diagnosis using HMM. *Proceedings of the Electrotechnical Conference on Mediterranean* (pp. 489-492).

[4] Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*, *40(1)*, 96-104.

[5] Detrano, R., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American. *Journal of Cardiology*, *64(5)*, 304-310.

[6] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert systems with applications*, *36(4)*, 7675-7680.

[7] Avci, E., & Turkoglu, I. (2009). An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. *Expert Systems with Applications*, *36(2)*, 2873-2878.

[8] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34(1)*, 366-374.

[9] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. *Machine Learning Repository: Fertility Data Set. Recupedao*.

[10] Kyoungok, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. 157-163.

[11] Barrientos, R. E., Cruz, N., Acosta, H. G., Rabatte, I., Gogeascoechea, M. C., Pavón, & Blázquez, S. L. (2009). Árboles de decisión como herramienta en el diagnóstico médico. Revista médica de la Universidad Veracruzana, *9(2)*, 19-24.

[12] Vapnik, V. N. (1998). *Statistical Learning Theory*. Nueva York: Wiley-usa

[13] Vapnik, V. N. (1995). *The Nature of Statistical learning Theory*. Nueva York: Springer-Verlag.

[14] Bakhtiarizadeh, M. R. (2014). Neural network and SVM classifiers accurately predict lipid binding proteins. *Irrespective of Sequence Homology*.

[15] Gutierrez, M., & J, F. (2011). Pronóstico de incumplimiento de pago mediante máquinas de vectores de soporte.

[16] Kecman, V. (2001). Learning and soft computing. Londres: mit Press-uk.

[17] Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machine and Other Kernel-Based Learning Methods*. Nueva York: Cambridge University Press.

[18] Jimenez, L., & Rengifo, P. (2010). Al interior de una máquina de soporte vectorial. Revista de ciencias, *(14)*, 73-85.

[19] Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). Probabilistic networks and expert systems.

[20] Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist*, *53(4)*, 416

[21] Edwards, W., & Fasolo, B. (2001). Decision technology. *Annual review of psychology*, *52(1)*, 581-606.

[22] Nadkarni, S., & Shenoy, P. P. (2001). A Bayesian network approach to making inferences in causal maps. *European Journal of Operational Research*, *128(3)*, 479-498.

[23] Nadkarni, S., & Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. *Decision support systems*, *38(2)*, 259-281

[24] Sánchez, A. S., Iglesias-Rodríguez, F. J., Fernández, P. R., & De, C. J. F. J. (2016). Applying the K-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders. *International Journal of Industrial Ergonomics*, *52*, 92-99

[25] Fix, E., & Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: Consistency properties. California Univ Berkeley.

[26] Mucherino, A., Papajorgji, P. J., & Pardalos, P. M. (2009). *Data Mining in Agriculture*. Springer Science & Business Media.

Fabio Mendoza Palechor received his M.Sc. degree in engineering from Universidad Tecnologica de Bolivar in 2014. his B.S. in systems engineering from Universidad de la Costa in 2010. He received his audit specialist title in 2011 from Universidad de la Costa. He is an associate professor at Universidad de la Costa since 2012. His research interests are artificial intelligence, data mining and web applications design.

Alexis De la Hoz Manotas received his M.Sc. in systems engineering and computing in 2011. He graduated in pedagogical studies in 2011 from Universidad de la Costa. He received his B.S. in systems engineering from Universidad del Norte, in Barranquilla in 2000. Since 2002 he is an assistant professor in the Departament of Systems Engineering in Universidad de la Costa. His research interests are location based systems, intelligent agents, algorithm analysis and software processes. He has published technical papers in peer-reviewed conferences and journals, and coauthored the books "Talleres de Algoritmos y Estructuras de Datos basados en UML", "Arboles y Grafos" and "Programación Orientada a Objetos".

Paola Ariza Colpas received his M.Sc. in systems engineering and computing in 2011. She received his B.S. in systems engineering from Universidad Simón Bolívar, in Barranquilla in 2007. His research interests are software quality y data mining.

Jorge Sepulveda Ojeda received his M.Sc. in systems engineering and computing in 2009. He received his B.S. in systems engineering from Universidad Simón Bolívar, in Barranquilla, Colombia. 2003. His research interests are Requirements engineering, testing de Software, methodology de Software, development applications Web.

Roberto Morales Ortega received his M.Sc. in government of information technology in 2016, Barranquilla, Colombia. He received his B.S. in systems engineering from Universidad de la Costa, Barranquilla, Colombia. 2011. His research interests are development applications web, mobile application development on android and IOS operating systems, business intelligence.

Marlon Piñeres Melo received his M.Sc. in systems engineering and computing in 2009. She received his B.S. in systems engineering from Universidad Simón Bolívar, in Barranquilla in 2005. His research interests are requirements engineering, testing de software, methodology de Software, development applications web.