ISSN 1870-4069

A Hybrid Intelligent System for Improving a Health Model Associated with Cardiovascular Disease

A. Martínez, C. Alberto Ochoa, Jorge R. Rodríguez

Universidad Autónoma de Ciudad Juárez, Mexico
al164654@alumnos.uacj.mx

Abstract. In some individuals, cardiovascular disease (CVD) is a congenital defect. However, factors such as stress, caffeine, alcohol, tobacco, and certain medications are the prevailing causes of CVDs. These factors are considered by studies such as the Framingham Heart Study, which is a reference point today to determine cardiovascular risk as a preventive measure. However, this study was conducted on US-based individuals, whose genetics, customs and lifestyle are different from the Latin American population. Due to the above, in our research we'll use the PhysioBC database, which contains 114 registries of inhabitants of Mexicali, Baja California. Each registry has its own fact sheet, electrocardiogram (ECG) record, and doctor's diagnosis, with ages ranging from 18 to 68 years. To process data, we employed data mining techniques for extraction and preprocessing (cleaning). To analyze and interpret data, we used the Waikato Environment for Knowledge Analysis (WEKA), and the classification algorithms Naive Bayes, Multilayer Perceptron, and J48. While testing these algorithms, we obtained the best results with the Naive Bayes classifier.

Keywords: cardiovascular risk, data mining, WEKA, J48, naive Bayes, multilayer perceptron.

1 Introduction

Any abnormality or irregularity in the natural heart rhythm can be defined as arrhythmia, and there are several factors that can cause it. Arrhythmias can be found in individuals with cardiovascular diseases (CVD), which are one of the leading causes of death in the entire world, representing about 30% of total deaths from heart disease [1].

CVDs are a global problem. Estimates suggest that in 2020, CVD deaths will increase from 15% to 20%, and by 2030, approximately 23.6 million people will die due to heart attacks and strokes [2]. In Western countries, CVDs are the leading cause of death and an important source of disability, which, cost-wise, means a huge burden for the healthcare sector [3,4]. To speed up the response of the health sector to CVD, the WHO Global Strategy as well as the Pan American Health Organization (PAHO) Regional Strategy, establish that health systems should focus on promotion and primary health care by increasing prevention and improving medical care [5].

In the literature review we found that risk factors for ECVs are widely identified as: age, diabetes, smoking (treated and untreated), systolic blood pressure, total cholesterol, HDL cholesterol, and Body Mass Index (BMI) [6–10]. There are several studies that can determine the risk of CVD, such as the Framingham Heart Study,

whose factors are mentioned above; however, this analysis is based on the US population, with a CVD risk and prevalence different than ours[6], as well as a different lifestyle, socioeconomic status and genetics. Although the Framingham Study is one of the most widely used ones to determine risk factors for cardiovascular diseases, there are several studies that analyze the applicability of this model for a different population than the American. [6,10]. The study of genetic variation compares the data obtained through genome sequencing of different individuals, which allows finding genes linked to certain diseases. In this sense, [11] highlights that caution must be taken when applying genetic CVD risk prediction models based on Single Nucleotide Polymorphism (SNP) that do not belong to the group of ancestors from which it derived. Due to the aforementioned, our work analyzed the CVD risk factors of the Baja California population. For this, we used the open-access PhysioBC database [12], which has the registry of 114 individuals, men and women aged between 18 and 68. This database contains ECG records, fact sheets, and diagnosis by a specialist doctor.

Having a CVD affects the quality of life of the individual, and it has both social and economic repercussions. CVDs are considered costly diseases, and the government of Mexico allocates a part of its budget to the Fund for Protection against Catastrophic Health Expenditure (FPGC) in response to the expenses related to this type of disease [13]. In Mexico, the population pyramid determines that 75% of the adults have less than 55 years old, and despite the fact that prevalence of cardiovascular risk factor is higher after 40, a large amount of the carriers of these risk factors are located in an economically active population [2]. In the same manner, these data can be observed in the population pyramid of Baja California, as shown in figure 1, data obtained from the National Population Council (CONAPO), projections of population for 2010-2050 [14].

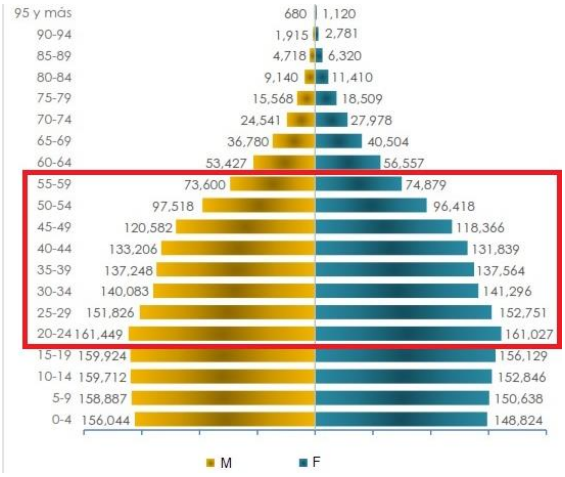

Fig. 1. Population distribution 2017 [14].

At present, there are vast databases that allow us to analyze the trend of the population in regards to public health, in order to find out their current situation and implement strategies to prevent diseases. There are several techniques for processing large amounts of data, among them is Knowledge Discovery in Databases (KDD), a pro-

cess that allows us to identify valid, novel, potentially useful and easy-to-understand patterns. The process begins with the understanding of the field of study and establishment of specific goals, followed by selection and integration of data from different sources. Due to this, there may be a need to clean up information noise, missing data and other forms of inconsistencies. Once data reprocessing is done, the next step is data mining, in which algorithms are used to find patterns or relationships between databases. The last step of this process consists in interpretation of the obtained information, where knowledge is finally acquired, which allows decision-making analyses [15]. This process can be observed in figure 2.

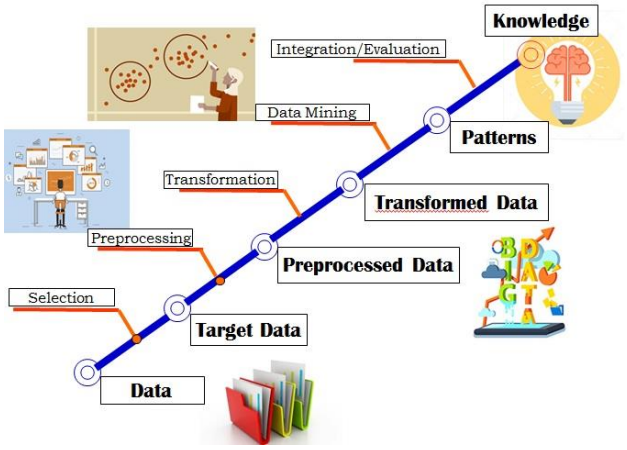

Fig. 2. KDD process partially obtained from: https://mnrva.io/kdd-platform.html

In our study, we used data mining techniques and the open-source software Waikato Environment for Knowledge Analysis (WEKA), written in Java, and developed at the University of Waikato, in New Zealand. WEKA is a collection of machine learning algorithms that contains tools for data preparation, classification, regression, clustering, data mining, association rules mining, and visualization rules [16]. There are several researches that use WEKA for classification, highlighting the use of the Naive Bayes, Multilayer Perceptron and J48 Algorithm for its high efficiency for observing results [17–22].

2 Related Research

In their research, [20] used the J48 Classifier Algorithm of the WEKA platform to predict cancer recurrence. The authors worked with a database of patients that underwent treatment for breast cancer provided by the UCI Machine Learning Repository. The dataset consists of 286 instances and 9 attributes, such as the patient's age at time of diagnosis, menopause, tumor size, inv-nodes (number of lymph glands that contain metastatic breast cancer), node caps, degree of malignancy, breast (whether left or right is diagnosed with tumor), breast quadrants, and irradiation. The dataset was in an ARFF file format. The selected algorithm was the J48, which analyzes data

through decision trees. The study was conducted for experimental purposes, and a decision tree was generated by taking the degree of malignancy as root node, so after interpreting results of the experiment, the authors concluded that patients with a specific value range of this attribute have higher chances of recurrent cancer.

On the other hand, [21] proposed an improved J48 Classification Algorithm to predict risk factors of diabetes. By using an interface between WEKA and MATLAB, they introduced an improved J48 Algorithm. The authors compared the Naive Bayes Classifier Algorithm, the multilayer perceptron (MLP), and the improved J48 Algorithm for the analysis of the same database; and came to the conclusion that the J48 Algorithm performed better at classification, with an accuracy rate of up to 99.87%. The Pima Indians Diabetes Data Set was used for experimental purpose.

For their part, [22] used several classifiers, such as the J48, Decision Tree, Random Trees, Random Forests, and Naive Bayes to analyze the relationship between people with diabetes and the risk of having a heart disease or not. In their experiments, they found that the J48 Algorithm showed the highest accuracy (95%) in comparison to the other classifiers, and that the Naive Bayes Algorithm spent less time in classification (0,00 seconds).

Below we present the results obtained in the application of the Naive Bayes, Multilayer Perceptron and J48 algorithms as an approach for the prediction of CVDs from risk factors obtained from the PhysioBC database.

3 Techniques and Tools Used

In the current work, we conducted experiments by using the PhysioBC database, which contains 114 ECG reports and 91 fact sheets of different patients. From these fact sheets, we generated a database with 89 instances (removing those that had missing, null or insufficient values) and 17 attributes, such as: patient, age, sex, 10-year and 30-year Framingham Risk Score, Body Mass Index (BMI), systolic blood pressure, diastolic blood pressure, smoking, drinking, exercise, diabetes, Arterial Hypertension (AH), CVD diagnosis, treatment, respiratory rate, and heart rate. Of the 89 patients, 50 were women and 39 men, aged between 18 and 68 years. The patients were volunteers from the health sector and the textile manufacturing sector in the municipality of Mexicali, Baja California.

In order to analyze the risk factors of suffering a CVD the database was prepared in a comma-separated values (.csv), making use of the different extensions and formats supported by WEKA, including .arff, .data, .names, .data, and .csv, among others. Figure 3 illustrates WEKA's Graphical User Interface (GUI), which was used along with the Explorer application shown in the interface menu.

The WEKA platform has several classification algorithms which we work with:

Naive Bayes, which is one of the most widely used classifiers for its simplicity and speed. It is a supervised classification and prediction technique, building models that predict the probability of possible outcomes.

Multilayer Perceptron (MLP), which is a neural network connecting multiple layers in a directed graph, the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses

backpropagation as a supervised learning technique, is widely used in research into computational neuroscience and parallel distributed processing.

Algorithm J48, is a version of C4.5 and builds decision trees from a set of training data using the concept of information entropy. At each tree node, choose the attribute that most effectively divides the set of samples into subsets. Its criteria is the normalized for information gain (obtained from the entropy difference) that results in the choice of an attribute to divide the data. The attribute with the highest information gain is chosen as the main node from which the branches are derived.

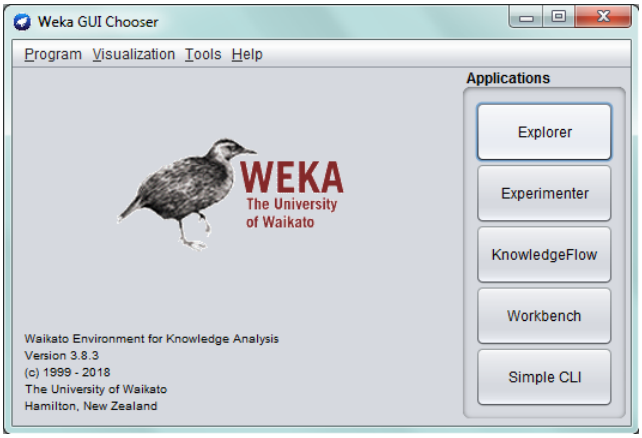

Fig. 3. WEKA graphical user interface (GUI).

Figure 4 shows the GUI of WEKA at the moment of opening the database, in which appears the list of attributes and a graphic representation of the distribution of patients according to the selected attribute, in this case the graph of the attribute "age" is observed.

3.1 Naive Bayes Classifier

In this case, we used WEKA's 'Use training set' option, and obtained 86 instances classified correctly, which represents a 96.62%, where 4 instances were classified as 'Yes CVD' and 82 as 'No CVD', as observed in the main diagonal of the confusion matrix, shown in figure 5. The Kappa coefficient was 0.7095 which indicates a considerable degree of agreement according to the Landis and Koch scale [23].

To visualize the results in a graphical format, we right-click on the Results list and select the option 'Visualize Classifier Error' in the panel. Figure 6 shows the graph obtained, where it can be seen that points in the upper right corner represent instances that had 'No CVD' and were classified correctly. In the same way, those that had 'Yes CVD' and were classified correctly can be observed in the lower left corner. In this part, we can visualize the attributes of each instance individually by selecting the point in the graph, which makes easier to analyze instances that were not classified correctly. Figure 7 portrays an example of the attributes of an instance that had 'Yes' on CVD diagnosis and was not classified correctly, which are the 2 cases observed in the upper left corner of the graph.

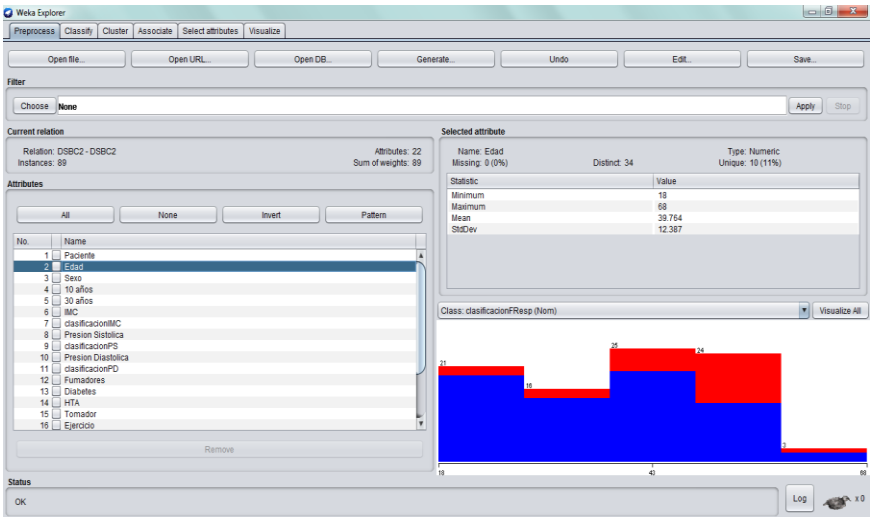

Fig. 4. List of attributes shown when opening the database and the graphical representation of the age attribute distribution.

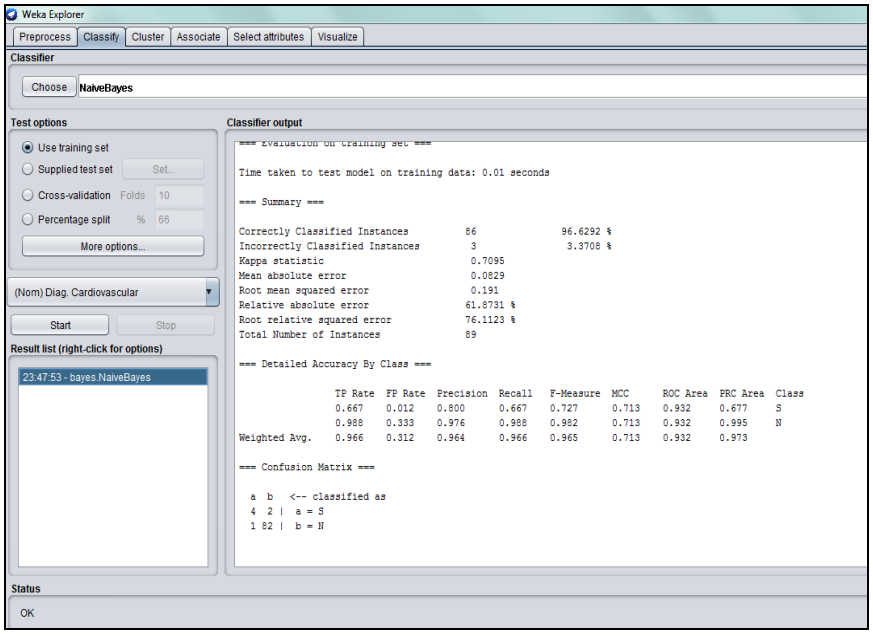

Fig. 5. Results of the Naive Bayes classifier.

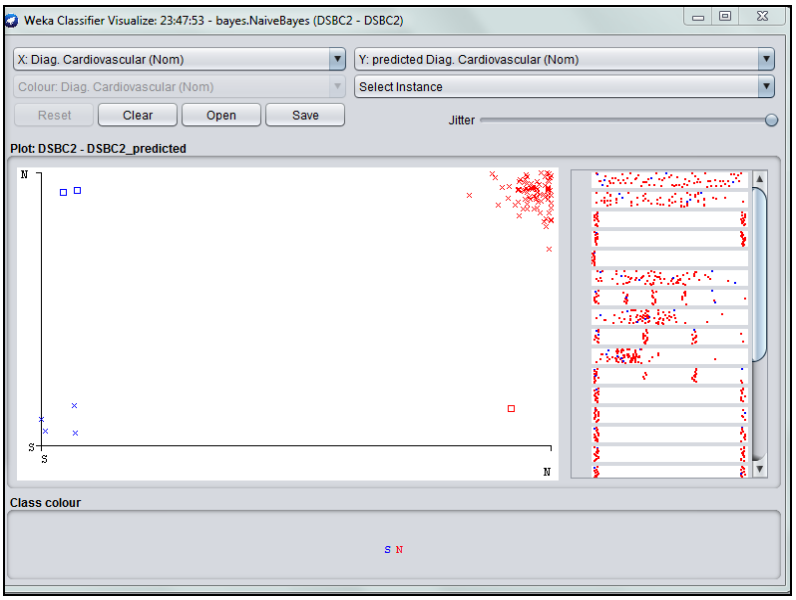

Fig. 6. Graphic visualizer of classifier errors with Naive Bayes.

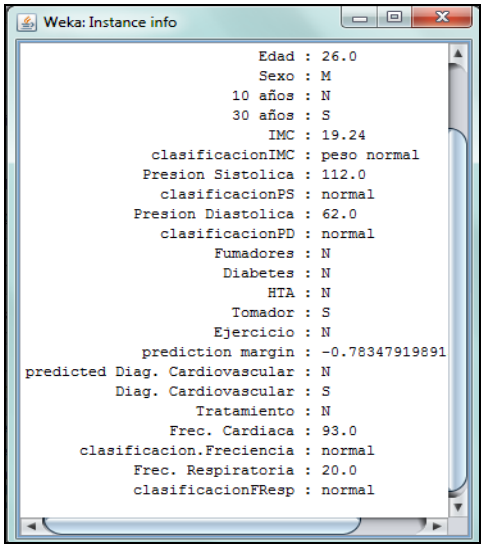

Fig. 7. Attributes of an instance that had cardiovascular diagnosis (Yes) and was not classified correctly.

3.2 Multilayer Perceptron

In this case we also used the use training set option and obtained 88 instances correctly classified which represents 98.87%, of which 5 instances were classified as Yes

CVD and 83 as No CVD as observed in the main diagonal of the confusion matrix, see figure 8. The Kappa coefficient in this case was 0.9032 which indicates a total or almost perfect concordance according to the Landis and Koch scale [23].

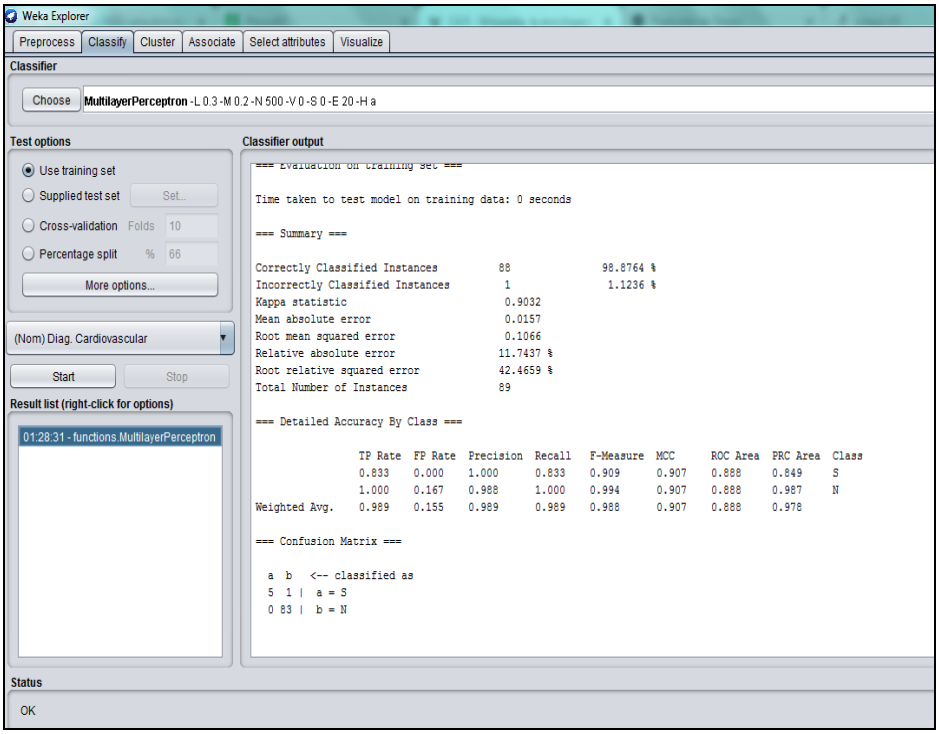

Fig. 8. Results of the Multilayer Perceptron classifier.

To visualize the results in a graphical form, we right-click on the results list and select the option "Visualize the errors" in the panel. Figure 9 shows the graph obtained, in this case only one instance was not classified correctly, as observed in the upper left corner of the graph.

3.3 J48 Algorithm

In this case we also used the ´use training set´ option and obtained 83 instances classified correctly, which represents 93.25%, of which 0 instances were classified as ´Yes CVD´ and 83 as ´No CVD´, as observed in the main diagonal of the confusion matrix in figure 10. Here, the Kappa coefficient was 0.0 which indicates a poor concordance according to the Landis and Koch scale [23].

To visualize the results in a graphical format, we followed the same procedure as in other cases. Figure 11 illustrates the graph obtained. In this experiment, only those instances that had 'No CVD' were classified correctly, while the algorithm failed to classify instances that had 'Yes CVD'. The J48 algorithm is a decision tree classifier,

and there is an option to visualize results as a tree; however, in this case, a clear tree was not obtained to determine whether a patient had a CVD or not.

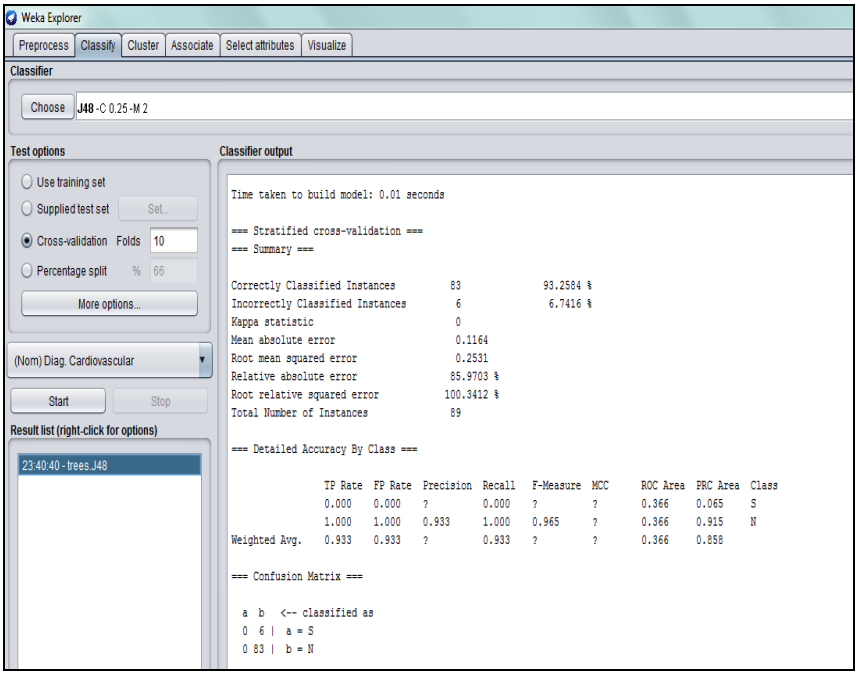

Fig. 10. Results of the J48 classifier.

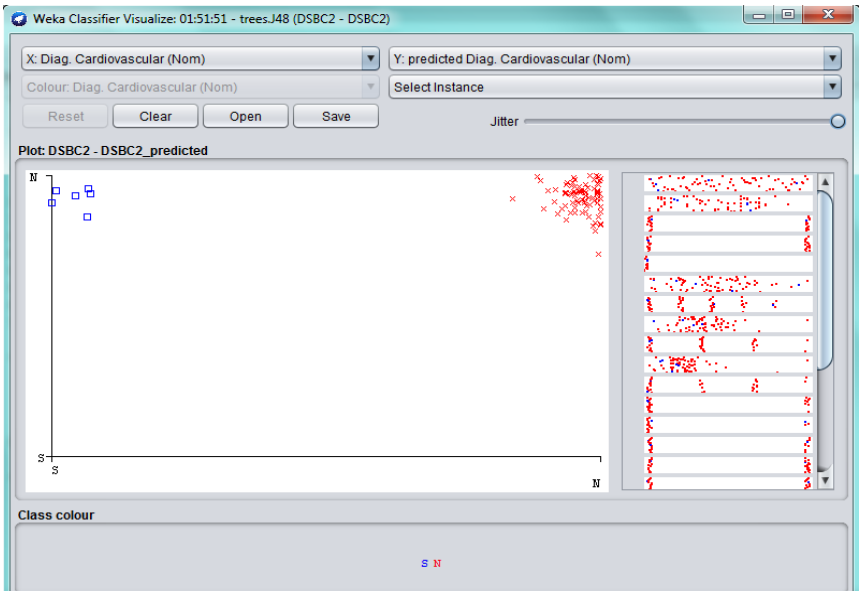

Fig. 11. Graphic visualizer of classifier errors with J48.

4 Conclusions

In this work, the Naive Bayes, Multilayer Perceptron and J48 classifiers were used as tools to determine the levels of accuracy in classification and prediction that a patient has a CVD by studying the presence or absence of widely identified risk factors, such as obesity (through BMI), diabetes, HA, smoking, alcoholism, among others. In the results obtained with each of these classifiers we found that the Multilayer Perceptron had better results, with 98.87% of the instances classified correctly, a Kappa coefficient of 0.9032, showing a complete or almost perfect concordance and a response time of 0.0 seconds. On the other hand, the results from the J48 classifier we obtained 93.25% of instances classified correctly, but only from those patients who did not present CVD; additionally, the Kappa coefficient was 0.0, which indicates a poor concordance. After these results, we came to the conclusion that for the J48 algorithm requires more instances that have CVD in order to make predictions correctly, since this data set only had 6 instances with diagnosed CVD. In the analysis of data to predict the risk factor of suffering or not a high risk disease such as CVDs there is still much to do, it is convenient to work on the creation of large databases that can identify the important aspects of a given population, in addition to making them available for use in research this would allow the development of accessible systems that support prevention plans in diseases such as Cancer, Diabetes and Cardiovascular Diseases.

Acknowledgments. We are very grateful to Dr. R. L. Avitia and N. Flores, members of the Department of Bioengineering and Environmental Health of the Autonomous University of Baja California, who kindly provided us with information from the database used in this work.

References

1. WHO: "WHO | Cardiovascular diseases (CVDs)," *WHO* (2016)
2. Sanchez, A., Bobadilla, M., Altamirano, B., Ortega, M., Gonzalez, G.: Enfermedad cardiovascular: primera causa de morbilidad en un hospital de tercer nivel Heart. Rev. Mex. Cardiol. 27(s3), pp. 98–102 (2016)
3. Hidalgo, M.M.: Nuevos modelos multivariantes en la medición del riesgo cardiovascular. Universidad de Salamanca, Departamanto de Estadística (2015)
4. Schnabel, R.B. *et al.*: "50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: A cohort study," *Lancet*, vol. 386, no. 9989, pp. 154–162 (2015)
5. Gómez, L.A.: Las enfermedades cardiovasculares: un problema de salud pública y un reto global. Biomédica 31(4), (2011)
6. Jiménez-Corona, A., López-Ridaura, R., Williams, K., González-Villalpando, M.E., Simón, J., González-Villalpando, C.: Applicability of Framingham risk equations for studying a low-income Mexican population. Salud Publica Mex. 51(4), pp. 298–305 (2009)
7. Rosas-Peralta, M. *et al.*: Cardiovascular risk reduction : Past, present and future in Mexico. pp. 38–47 (2018)
8. de León, G.P. y P., Campoy, U.R., Bravo, A.C., Witrón, J.J.: Factores de riesgo cardiovascular y la percepción del estado de salud en profesores de tiempo completo de la

UABC, campus Mexicali / Cardiovascular disease risk factors and the perception of health in full professors of the UABC, campus Mexicali. RICS Rev. Iberoam. las Ciencias la Salud 5(10), pp. 98–120 (2016)
9. María, J., Cortes, M., Aragonés, N., Godoy, P., José, M., Moros, S.: Las enfermedades crónicas como prioridad de la vigilancia de la salud pública en España. 30(2), pp. 154–157 (2016)
10. Alvarez, A.:Las tablas de riesgo cardiovascular. Una revisión crítica. Medifam 11(3), pp. 122–139 (2001)
11. Carlson C.S. *et al.*: Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. PLoS Biol. 11(9), (2013)
12. Flores, N.: PhysioBC. [Online]. Available: http://www.physiobc.org/. [Accessed: 14-Nov-2018]
13. Comisión Nacional de Protección Social en Salud: INTERVENCIONES DEL FONDO DE PROTECCIÓN CONTRA GASTOS CATASTRÓFICOS 2018 (2018)
14. CONAPO: Projections of the Population of Mexico 2010-2050. p. 15 (2010)
15. Minerva Data Mining: KDD: Knowledge Discovery in Databases | Minerva Data. [Online]. Available: https://mnrva.io/kdd-platform.html. [Accessed: 16-Nov-2018].
16. The University of Waikato (NZ): Weka 3 - Data Mining with Open Source Machine Learning Software in Java."[Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 16-Nov-2018].
17. Bhargava, Sharma: Decision Tree Analysis on J48 Algorithm for Data Mining. IJARCSSE 3(6), pp. 1114–1119 (2013)
18. Arora, R.: Pxc3882492 54(13), pp. 21–25 (2012)
19. Kau,r H., Raghava, G.P.S.:A neural-network based method for prediction of γ-turns in proteins from multiple sequence alignment. Protein Sci. 12(5), pp. 923–929 (2003)
20. Sharma, S., Purohit, R., Rathore, P.S.:Prediction of Recurrence Cancer using J48 Algorithm.In:. Icces 2017, pp. 386–390 (2017)
21. Kaur G., Chhabra, A.: Improved J48 Classification Algorithm for the Prediction of Diabetes. Int. J. Comput. Appl. 98(22), pp. 13–17 (2014)
22. Gokilam G.G., hanthi, K.: Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset. Compusoft 5(3), pp. 2074–2079 (2016)
23. Landis, J.R. *et al.*: Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. Acta Trop. 33(1), pp. 54–58 (2015)