ISSN 1870-4069

Evaluation of Models for Estimating Cardiovascular Risk based on the Framingham Table

Norma Karen Valencia[1], Edgar Corona Organiche[2],
Abraham J. Jiménez Alfaro[2], Griselda Cortés Barrera[2]

[1] Tecnológico de Estudios Superiores de Chimalhuacán,
Mexico

[2] Tecnológico de Estudios Superiores de Ecatepec,
Mexico

karenvalencia@teschi.edu.mx,
{ecorona,ajimenez,gcortes}@tese.edu.mx

Abstract. Cardiovascular diseases (CVD) are responsible for the majority of deaths in the world, approximately 30%; which makes it important to estimate the risk of developing cardiovascular disease in order to prevent and reduce this index. The main purpose of this work is to evaluate the performance of different algorithms to select which of them is the most appropriate to estimate cardiovascular risk. The present work presents the comparison of the estimators developed through decision trees, Naive Bayes algorithm, a vector support machine (SVM) and a Neural Network composed of six inputs, 8 hidden layers and an output to estimate Cardiovascular Risk; considering seven risk factors as input and three types of risk as output: high, intermediate, and low. The confusion matrix of each implemented algorithm shows the precision of each of the models, taking as a reference the cardiovascular risk stratification by means of the Framingham function, allowing to establish the degree of error of each one of them. The analysis of the results allows us to observe that the performance of the multilayer perceptron was below the Decision Tree and the Naive Bayes Algorithm, which provide very similar results. Therefore, decision trees are a good choice to determine the level of cardiovascular risk. Efficiently estimating cardiovascular risk will allow health centers to establish strategies to reduce this factor in their attended population.

Keywords: Decision trees, Framingham table, vector support machine, naive Bayes, neural network.

1 Introduction

The World Health Organization (WHO) defines Cardiovascular Disease (CVD) as all those conditions that generate disorders of the heart and blood vessels, including: coronary heart disease, which is described as a disease of the blood vessels that supply the heart muscle. Also included are cerebrovascular diseases: diseases of the blood vessels supplying the brain; peripheral arteriopathies: diseases of the blood vessels;

Table 1. Risk of the Cardiovascular Event in a period of 10 years.

Risk level	10-year risk estimate percentage.	Years of possible cardiovascular event
Under	<15%	Older than 7 years
Intermediate	15-20%	Over 4 years
High	> 20%	Less than 5 years

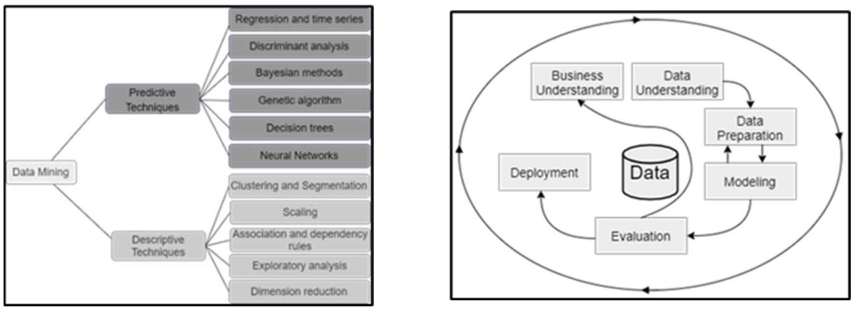


Fig. 1. a) Classification of Data Mining techniques, b) Phases of the CRISP-DM Methodology (GBD 2015 Risk Factors Collaborators., 2018).

rheumatic heart disease, congenital heart disease caused by malformations of the heart present from birth, adding to deep vein thrombosis and pulmonary emboli, which can detach (emboli) and lodge in the vessels of the heart and lungs [1].

In Mexico, it is estimated that CVD represents 20% of all deaths in adults. According to the National Institute of Statistics and Geography (INEGI), in 2016 136,342 deaths were reported due to heart disease, an increase of 7,611 deaths compared to 2015. Causes of death include ischemic heart disease, cerebrovascular diseases, hypertensive, among others [2].

The assessment of cardiovascular risks has generated various studies, since 2015 the Federal Government provided the "Risk Factors Questionnaire" to be used as an instrument to determine cardiovascular risk based on the risk stratification of the Framingham table. This work aims to verify if the decision trees generate better results for the Cardiovascular Risk Estimation.

2 Problematic

On September 25, 2015, the document entitled "Transforming Our World: the 2030 Agenda for Sustainable Development" was developed, adopted by member countries of the United Nations, including Mexico, and one of the objectives of this document is to develop strategies to prevent Non-Communicable Diseases, like CVD [3].

The study entity has historical records of more than 10 years, information concentrated in the Micro diagnosis certificates of its population. However, it does not have an information analysis tool that allows it to estimate or quantify morbidity, as well as mortality caused by cardiovascular diseases.

In addition to the above, the Health Institution does not have a model that allows it to estimate the risk of developing cardiovascular diseases in the population it serves.

Table 2. Description of Phases: CRISP-DM Methodology.

Phase	Description
Business understanding	It brings together the tasks of understanding the objectives and requirements of the project from a business or institutional perspective, in order to turn them into technical objectives and a project plan[11].
Understanding the Data	It includes the initial collection of data, with the aim of establishing a first contact with the problem[12].
Data Preparation	It consists of some tasks such as data selection by choosing a subset of the data collected in the previous stage. Cleaning the data, preparing them for the modeling phase, either by applying normalization techniques, discretization of numerical fields, treatment of null values, among others [13].
Modeling	It is suggested to make a comparison of the modeling techniques that are most appropriate to solve the problem[14].
Evaluation	The model is evaluated based on compliance with the success criteria of the problem, review the process followed taking into account the results obtained[15].
Implementation	This task takes the results of the evaluation and concludes a strategy for its implementation[15].

Likewise, it lacks the level of correlation that exists between the different risk factors that influence the development of this type of disease.

When prevention campaigns are started, the population is not fully captured, since the exhaustive analysis of data and risk factors for everyone is omitted, which generates an imprecise report of the population vulnerable to said diseases.

3 Theoretical Framework

The classic Framingham function estimates the risk of suffering an event in the next 10 years, considering death of coronary origin, non-fatal acute myocardial infarction, stable angina, or unstable angina (coronary insufficiency) as an event. The National Cholesterol Education Program in its latest document Adult Treatment Panel III (ATP-III) has modified this function to exclusively calculate the risk of so-called "hard" events, that is, non-fatal acute myocardial infarction and coronary death, excluding diabetics from this estimate by considering them directly at high risk. [4, 5].

Derived from the above, Table 1 summarizes the risks studied for cardiovascular events, the "Very High" risk level is not considered since it is included within the high risk, all these risks belong to the stratification based on Framingham table that is used in the Health Sector through the Clinical Practice Guide: Detection and Stratification of Cardiovascular risk factors [6].

The "Risk Factors Questionnaire" consists of the following variables to detect risk factors: Diabetes, Glycemia, Arterial Hypertension, Cancer (Colorectal, pulmonary, Oral, Gastric), cardiovascular diseases (heart disease, embolism, and hypertension), for the present study, those shown in Table1 are taken. The applications commonly developed with predictive analytics are predicting risks, predicting activation of new clients, predicting sales, among others. This type of analysis is characterized by

Table 3. Risk factors used in the models.

Risk factor	Neural Network	Naive bayes	Decision trees
Age	X	X	X
Diastolic pressure	X	X	X
Systolic pressure	X	X	X
Diabetes	X	X	X
Hypertension	X	X	X
Inheritance	X	X	X

Table 4. Cardiovascular risk probabilities using Bayes' theorem.

Risk factor	Risk of cardiovascular disease (RECV)
Age	$P(RECV\|E)$
Diastolic pressure	$P(RECV\|E \cap Pd)$
Systolic pressure	$P(RECV\|E \cap Pd \cap Ps)$
Diabetes	$P(RECV\|E \cap Pd \cap Ps \cap D)$
Hypertension	$P(RECV\|E \cap Pd \cap Ps \cap D \cap Hip)$
Inheritance	$P(RECV\|E \cap Pd \cap Ps \cap D \cap Hip \cap Her)$

Table 5. Review of variables in various cardiovascular risk stratification tables.

Variables	Identification card Microdiagnosis	SCORE	Framingham	PCE
Age	*	*	*	*
Sex	*	*	*	*
Cholesterol level	Only valued people	*	*	*
Pressure Systolic mmg	Only valued people	*	*	*
Smoking	*	*	*	*
Diabetes	*	*	*	*
Diet	*	*	*	*
Physical activity	*	*	*	*
Weight	*	*	*	*

requiring a training set, which is made up of a data log. Discrete prediction and continuous prediction tasks can be performed in predictive analytics [7].

Within data mining, classifiers can obtain the incidence of cardiovascular events. Bayesian classifiers allow to classify discrete and limited events (independent variables) in a certain number of classes by defining a statistical function for each class. In Fig. 1a, the graph is showing its classification. Bayesian networks are a probabilistic model by means of which it is feasible to construct a graph between the causes of an event (independent variables) and its consequences (dependent variables) [8].

4 Methodological Framework

There are several methodologies that provide a series of steps to follow in order to carry out a proper implementation of data mining. According to polls published in

KDnuggets [9], the most used methodologies are CRISP-DM (Cross Industry Standard Process for Data Mining), SEMMA, KDD and Catalyst. This information can be consulted in the link https://www.kdnuggets.com/2014/10/c risp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

The CRISP-DM is a free distribution methodology that can work with any tool to develop any project, it structures the life cycle of a Data Mining project, its phases described in Table 2, interact with each other iteratively during the development of the project as shown in Fig. 1b designed in a neutral way to the tool used for the development of the project [10].

5 Methodological Development

5.1 Business Understanding or Understanding

The Framingham table works with nine risk factors as well as other risk stratification tables such as the Systematic Coronary Risk Evaluation (SCORE) that measures the risk of mortality due to cardiovascular disease at 10 years in a European population aged 40 to 65 years [16], the present study focuses on the first table and on the practice of clinical guidelines for cardiovascular risk stratification.

5.2 Understanding the Data

Nine risk factors that are presented in Table 3 were considered, three characteristics were discarded in the first place, a non-modifiable factor such as sex and two modifiable factors which are weight, smoking and physical activity. The risk factors that were not considered have a close relationship with those that were used as characteristics, in the case of the risk factor "weight" it is linked to systolic and diastolic pressure, as well as hypertension and diabetes.

5.3 Data Preparation

For the model to work optimally, the data must be standardized complying with the characteristics, three categories High, Intermediate and Low are established as output, from inputs six characteristics are taken containing a total of 248 records, the input data as well as the output are shown in Table 5 and Fig 2.

5.4 Modeling

The Neural network architecture consists of 6 inputs, 1 hidden layer with 8 neurons and three outputs as shown in Fig. 2a.

5.4.1 Neural Network

The essential characteristics of a neural network are the nodes (organized in layers), the network architecture, which describes the connection between the nodes, and the

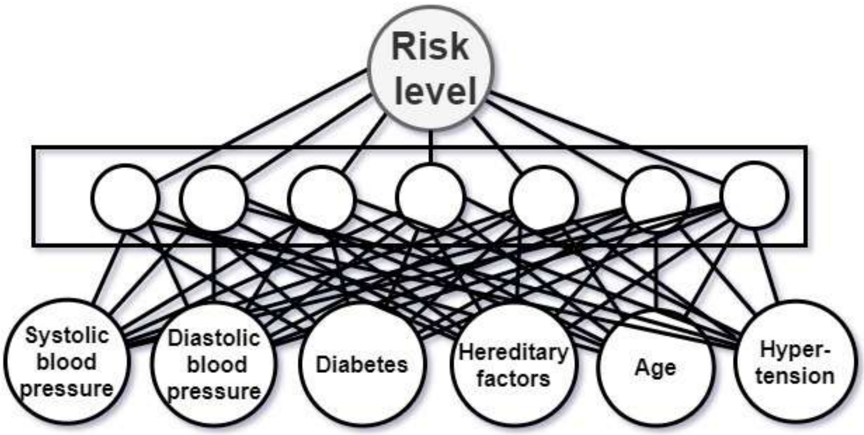


Fig. 2. a) Neural Network Architecture.

Table 6. Partial sample of the data entered the models (Hered. Mean Hereditary and Hyper. Hypertension).

TARGET	AGE	GLUCOSE	PD	PS	HERED.	HYPER.
HIGH	47	101	134	81	0	1
LOW	56	98	126	86	0	0
INTERMEDIATE	66	113	115	80	0	0
LOW	35	97	122	83	0	0

algorithm used to find the values of the network parameters (weights). [17] The layers of a network can be: a) To begin with, made up of neurons that introduce the six risk factors, no processing is performed in these neurons, b) Hidden made up of neurons whose inputs come from the previous layer and c) Output, formed by neuron that indicates the level of risk.

The optimization method used is Adam, which only requires first-order gradients with low memory requirements. The method calculates the individual adaptive learning rates for different parameters from the estimates of the first and second moment of the gradients; the name Adam is derived from the adaptive estimation of the moment [18]. The Orange Datamining software (version 3.26.0) was used as a data analysis tool. The Fig 3, shows the widgets used to work with the models.

5.4.2 Naive Bayes Algorithm

The proposed model uses the Naïve Bayes algorithm based on age, then other factors, generating the probability of suffering from cardiovascular diseases. The following table shows the probabilities obtained. Where: *P (RECV | E)* = Probability of cardiovascular risk-taking Age as a risk factor. *P (RECV | E⋂Pd)* = Probability of cardiovascular risk-taking Age and diastolic pressure as risk factors. *P (RECV | E⋂Pd⋂Ps)* = Probability of cardiovascular risk-taking Age, distolic pressure and systolic pressure as risk factors. *P (RECV | E⋂Pd⋂Ps⋂D)* = Probability of cardiovascular risk-taking Age, diastolic pressure, systolic pressure and Diabetes as a risk factor. *P (RECV | E⋂Pd⋂ Ps⋂D⋂Hip)* = Probability of cardiovascular risk-taking Age, diastolic pressure, systolic pressure, Diabetes and Hypertension as a risk factor. *P (RECV |*

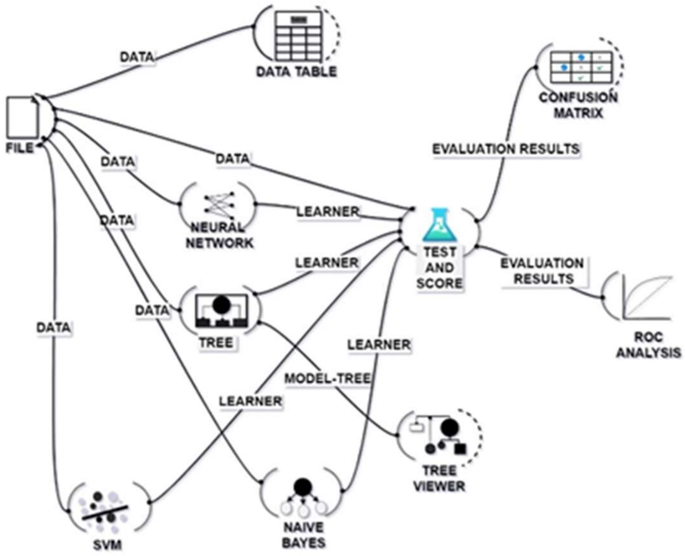


Fig 3. The RN configuration handles 8 hidden layers with logistic activation function and Adam algorithm with a maximum of 200 iterations.

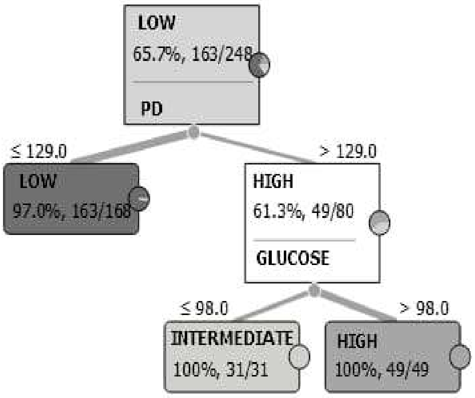
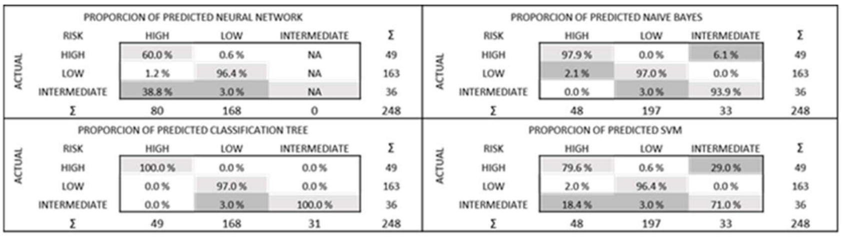


Fig. 4. a) Tree Architecture.

E⋂Pd⋂Ps⋂D⋂Hip⋂Her) = Probability of cardiovascular risk-taking Age, diastolic pressure, systolic pressure, Diabetes, Hypertension and Heredity as a risk factor.

5.4.3 Classification Trees

Classification trees, which are an alternative to more classical statistical techniques, such as multiple regression, ANOVA analysis, logistic regression, discriminant



PROPORCION OF PREDICTED NEURAL NETWORK

	RISK	HIGH	LOW	INTERMEDIATE	Σ
ACTUAL	HIGH	60.0 %	0.6 %	NA	49
	LOW	1.2 %	96.4 %	NA	163
	INTERMEDIATE	38.8 %	3.0 %	NA	36
	Σ	80	168	0	248

PROPORCION OF PREDICTED NAIVE BAYES

	RISK	HIGH	LOW	INTERMEDIATE	Σ
ACTUAL	HIGH	97.9 %	0.0 %	6.1 %	49
	LOW	2.1 %	97.0 %	0.0 %	163
	INTERMEDIATE	0.0 %	3.0 %	93.9 %	36
	Σ	48	197	33	248

PROPORCION OF PREDICTED CLASSIFICATION TREE

	RISK	HIGH	LOW	INTERMEDIATE	Σ
ACTUAL	HIGH	100.0 %	0.0 %	0.0 %	49
	LOW	0.0 %	97.0 %	0.0 %	163
	INTERMEDIATE	0.0 %	3.0 %	100.0 %	36
	Σ	49	168	31	248

PROPORCION OF PREDICTED SVM

	RISK	HIGH	LOW	INTERMEDIATE	Σ
ACTUAL	HIGH	79.6 %	0.6 %	29.0 %	49
	LOW	2.0 %	96.4 %	0.0 %	163
	INTERMEDIATE	18.4 %	3.0 %	71.0 %	36
	Σ	48	197	33	248

Fig. 6. Confusion matrix (misclassified is shown in a darker tone).

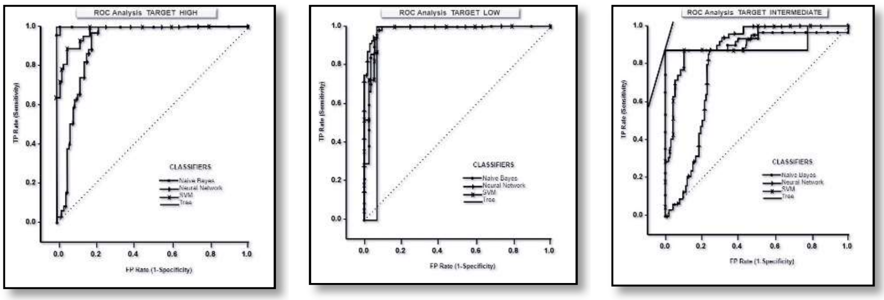


Fig. 7. Analysis of ROC Curves of the models: a) High Risk, b) Low Risk. c) ROC Curve Analysis of Intermediate risk models.

Table7. Model results.

Model	* AUC	Classification Accuracy	Precision	Specificity
Naive bayes	0.977	0.976	0.976	0.961
Neural Networt	0.944	0.843	0.745	0.907
SVM	0.943	0.875	0.871	0.930
Tree	0.955	0.980	0.980	0.961

* AUC = Area under ROC curve

analysis, and survival models. (Classification trees also seem to obtain better predictive rates than the rest of data mining techniques when mostly categorical information is used [19].

The configuration for SVM is Cost: 1.0, Regression Loss Epsilon 0.10, Kernel: Sigmoid ang Iteration Limit: 100. The parameters for the tree are Limit the maximal tree depth to: 100, Stop when majority reaches: 95 and Min. number of instances in leaves: 2.

5.4.4 Support Vector Machines

This model uses the technique of support vector machines (SVM), to make hyperplanes that divide the population between usual and unusual (in our study low, intermediate and high risk), from the training population, with which it is achieved classify the validation population [20].

6 Results

The data that were treated for this study were from 248 patients corresponding to a section of the studied Health Center. The configuration chosen with respect to the SVM is shown in Table 4, as it is observed the cost was chosen of 1 and the loss of 0.10. Access to training data is here[1].

The perceptron did not correctly identify the level of Intermediate Risk as seen in the confusion matrix Fig. 6, in such a way that it only worked on two levels of Risk: low and high, that is, its classification precision is 0.843 as seen in the data in Fig. 6.

The general results are shown in the table 6, in which the performance of the models is compared. A more global way of knowing the quality of the test in the full spectrum of cut-off points is using ROC curves (receiver operating characteristics), which constitute a fundamental and unifying tool in the evaluation process, see Fig. 7 [21]. The analysis had as parameters: Default threshold (0.5) point. Show performance line: FP Cost: 500, FN Cost: 500 and Prior probability: 66%.

From the above data, Table 6 is generated where the analysis of the ROC curves is interpreted numerically as well as the classification matrix corresponding to each model. The result of Analysis by means of the ROC curves shows that the trees generate better results.

7 Conclusions

Based on the results of the confusion matrices, it is concluded that the decision trees are the most suitable for estimating cardiovascular risk, followed by the Naive Bayes algorithm and the one with the lowest performance was the SVM. The efficient estimation of cardiovascular risk will allow health centers to establish strategies to reduce both this index and the mortality index in their attended population due to the development of cardiovascular disease.

References

1. OMS: Enfermedades cardiovasculares hechos clave (2021)
2. OMENT: Un panorama de las enfermedades cardiovasculares. Observatorio Mexicano de Enfermedades no Transmisibles (2018)
3. ONU: Informe de los objetivos de desarrollo sostenible 2020. Naciones Unidas (2020)
4. Pacheco, A. M., Jáquez, T. J.: Prevalencia de síndrome metabólico en la consulta externa. Rev Sanid Milit Mex, vol. 71, pp. 264–275 (2017)
5. Godala, M., Materek-Kuśmierkiewicz, I., Moczulski, D., Szatko, F., Gaszyńska, E., Kowalski, J.: Estimation of cardiovascular risk in patients with metabolic syndrome. Pol Merkur Lekarski; vol. 41, no. 246, pp. 275–278 (2016)
6. Secretaría de Salud. Intervenciones de enfermería para la prevención de complicaciones de Enfermedades Cardiovasculares en adultos en los tres niveles de atención. México, G.P.C. SS-365-16 (2016)
7. Lanzarini, L. C., Hasperué, W., Villa-Monte, A., Basgall, M. J., Molina, R., Rojas-Flores, L., Corvi, J. P., Jimbo-Santana, P., Fernández-Bariviera, A., Puente, C., Olivas-Varela, J.

[1] https://mega.nz/folder/F5U1DKKI#Ucryqpb4KAaVa4 IDQlt3lA.

A.: Minería de datos y Big data: Aplicaciones en riesgo crediticio, salud y análisis de mercado. In: XX Workshop de Investigadores en Ciencias de la Computación, pp. 350–354 (2018)

8. Castrillón, O. D., Sarache, W., Castaño, E.: Sistema bayesiano para la predicción de la diabetes. Inf. Tecnol., vol. 28, no. 6, pp. 161–168 (2017)
9. Condori, S. E.: Modelo de minería de datos para la predicción de casos de anemia en gestantes de la provincia de Ilo. Tesis Universidad Nacional de Moquegua, Perú (2019)
10. Mejía, H. R.: Modelo de minería de datos para la identificación de patrones que influyen en la mora de la cooperativa de ahorro y crédito San José S.J., M. S. Tesis, Pontificia Universidad Católica de Ecuador. Ecuador (2018)
11. Betancourt, S., Gómez, M. C., Quintero, J. B.: Inteligencia de negocios aplicada al ecoturismo en Colombia: Un caso de estudio aplicando la metodología CRISP-DM. In: XIV Congreso Ibérico de Sistemas y Tecnologías de la Información, CISTI 2019, Coimbra, Portugal (2019)
12. Huber, S., Wiemer, H. Schneidera, D., Ihlenfeldt, S.: DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. In: 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, pp. 403–408 (2018)
13. Schäfer, F., Zeiselmair, C.: Sintetizar CRISP-DM y gestión de calidad: un enfoque de minería de datos para procesos de producción. In: Conferencia Internacional IEEE 2018 sobre Gestión de Tecnología, Operaciones y Decisiones, Marrakech, Marruecos (2018)
14. Espinosa, J. J.: Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. Ing. en inv. Tec., vol. XXI pp. 1–17 (2020)
15. Mancilla, G., Leal-Gatica, P., Sánchez-Ortiz, A., Vidal-Silva, C.: Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: Un análisis en minería de datos. Formación Universitaria, vol. 13, no. 6 (2020)
16. Marco, V., Jarauta, E.: Burden of disease calculation of cardiovascular risk and therapeutic objectives. Clínica e Investigación en Arteriosclerosis, vol. 33, pp. 10–17 (2021)
17. Millan-Solarte J. C., Cerezo, E. C.: Modelos para otorgamiento y seguimiento en la gestión del riesgo de crédito. Mét. Cuantitativos para la Eco. y la Emp., vol. 25, pp. 23–41 (2018)
18. Jalomo, J., Preciado, E.: Comparativa de desempeño de los optimizadores Adam vs SGD en el entrenamiento de redes neuronales convolucionales para la clasificación de imágenes ECG. Rev. Pistas Educativas, vol.42, pp. 313–325 (2020)
19. Ortíz, J. M. Rúa-Vieites, A., Bilbao-Calabuig, M. P.: Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas. Rev. Recta, vol. 18, pp. 177–201 (2017)
20. Gracia, M. E.: Máquinas de soporte vectorial y árboles de clasificación para la detección de operaciones sospechosas de lavado de activos. Lámpsakos, vol. 21, pp. 26–38 (2019)
21. Cobo, M. G.: Determinantes de malnutrición en pacientes en hemodiálisis: Efecto de la suplementación Proteica oral intradiálisis. Ph. D. Tesis, Departamento de Medicina, Universidad Complutense de Madrid, España (2018)