South Health and Policy. 2026; 5:364
doi: 10.56294/shp2026364

REVIEW

Prediction of Cardiovascular Diseases Using Machine Learning Models

Predicción de Enfermedades Cardiovasculares mediante Modelos de Aprendizaje Automático

Michael Rafael Rodríguez Rodríguez[1], Claudia Alejandra Delgado Calpa[1], Héctor Andrés Mora Paz[1]

[1]Universidad CESMAG, Facultad de Ingeniería, Ingeniería de Sistemas. Pasto, Colombia.

Cite as: Rodríguez Rodríguez MR, Delgado Calpa CA, Mora Paz HA. Prediction of Cardiovascular Diseases Using Machine Learning Models. South Health and Policy. 2026; 5:364. https://doi.org/10.56294/shp2026364

Submitted: 09-02-2025 **Revised:** 08-06-2025 **Accepted:** 24-12-2025 **Published:** 01-01-2026

Editor: Dr. Telmo Raúl Aveiro-Róbalo

Corresponding Author: Michael Rafael Rodríguez Rodríguez

ABSTRACT

The study addressed the global problem of cardiovascular diseases, which were one of the leading causes of mortality and morbidity according to the World Health Organisation. Multiple risk factors, both modifiable and non-modifiable, were identified, and the need to implement technologies that would enable early and accurate detection was emphasised. Given this scenario, the use of machine learning algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN), combined with traditional and alternative kernel functions, was proposed. A comparative approach was developed to validate the hypothesis that under-explored kernel functions could improve predictive performance in terms of accuracy and response time. To this end, models were trained with data extracted from recognised platforms such as Kaggle and UCI, and metrics such as accuracy, recall and F1-score were applied. The models were adjusted with hyperparameter optimisation techniques using random search. The results demonstrated that certain alternative kernel functions offered improvements in the error-time ratio, in some cases outperforming conventional kernels. The research not only contributed methodological advances in the development of predictive models, but also provided a support tool for clinical decision-making, particularly useful in contexts where timely diagnosis is crucial. Finally, the project contributed to strengthening artificial intelligence in public health, promoting well-being through the prevention and proactive management of cardiovascular diseases.

Keywords: Cardiovascular Diseases; Machine Learning; Kernel Functions; SVM; Preventive Diagnosis.

RESUMEN

El estudio abordó la problemática global de las enfermedades cardiovasculares, las cuales representaron una de las principales causas de mortalidad y morbilidad según la Organización Mundial de la Salud. Se identificaron múltiples factores de riesgo, tanto modificables como no modificables, y se enfatizó en la necesidad de implementar tecnologías que permitieran una detección temprana y precisa. Frente a este panorama, se propuso el uso de algoritmos de aprendizaje automático como Máquinas de Soporte Vectorial (SVM) y Redes Neuronales Artificiales (ANN), combinados con funciones kernel tradicionales y alternativas. Se desarrolló un enfoque comparativo para validar la hipótesis de que funciones kernel poco exploradas podrían mejorar el rendimiento predictivo en cuanto a exactitud y tiempo de respuesta. Para ello, se entrenaron modelos con datos extraídos de plataformas reconocidas como Kaggle y UCI, y se aplicaron métricas como precisión, recall y F1-score. Los modelos fueron ajustados con técnicas de optimización de hiperparámetros mediante búsqueda aleatoria. Los resultados demostraron que ciertas funciones kernel alternativas ofrecieron mejoras en la relación error-tiempo, superando en algunos casos a los kernel convencionales. La investigación no solo aportó avances metodológicos en el desarrollo de modelos predictivos, sino que también ofreció una

herramienta de apoyo para la toma de decisiones clínicas, especialmente útil en contextos donde el diagnóstico oportuno es crucial. Finalmente, el proyecto contribuyó al fortalecimiento de la inteligencia artificial en salud pública, promoviendo el bienestar mediante la prevención y el manejo proactivo de enfermedades cardiovasculares.

Palabras clave: Enfermedades Cardiovasculares; Aprendizaje Automático; Funciones Kernel; SVM; Diagnóstico Preventivo.

INTRODUCTION

Cardiovascular diseases, according to the WHO, are one of the significant public health problems worldwide, with cerebrovascular disease being the leading cause of mortality and morbidity. According to estimates, it claims 17,9 million lives each year.[1,2,3,4,5,6]

In addition, the main problems that occur in people are heart attacks, stroke, metabolic syndrome, heart disease, and hypertension, are caused by overweight, obesity, dyslipidemia, smoking, physical inactivity, unhealthy diet, these are existing risk factors that can be modified, and some cannot be altered, but are a causal factor for the development of cardiovascular disease such as age, gender, personal history of cardiovascular disease and history of family members with premature CVD only when they have occurred in the first degree.[7,8,9,10]

There are many challenges facing people with this type of pathology, ranging from the quality of health care to the technologies involved in monitoring and care. A large number of users with cardiovascular diseases require frequent hospital stays, since the restricted technology concerning health is concentrated in medical centers, mainly because acquiring them has a high cost, as does their maintenance, without leaving aside the fact that specialized knowledge is needed to interpret and analyze the data captured by this highly complex equipment.[4] Likewise, the absence of technologies aimed at the detection and prevention of cardiovascular events that are dangerous to health contributes to the high rate of deaths and hospitalizations.[11,12,13,14,15]

A precedent to be taken into account is the application of machine learning in the healthcare field, where it has been increasing by approximately 30 % to 40 % year after year, thanks to the statistical and probabilistic analysis of other research. Even so, the challenge of data collection and manipulation is complex, as is predicting death when suffering from any disease. The challenge is even greater when it comes to human beings and organs as complex as the heart or the system that comprises it. But being diagnosed is not a decisive end for people, as there are several methods to maintain a healthy heart.[8]

To address the issues mentioned above, it is essential to tackle the problem at an earlier stage to prevent the onset of significant pathology; therefore, there is a need to utilize AI models that provide timely diagnostic support for cardiovascular diseases. Even so, the machine learning models known so far, which are used to predict the diagnosis of heart failure, achieve results with variable accuracy, ranging from 80 % to 90 % in their best cases.[8]

Thus, various machine learning techniques are suitable for identifying the most critical features after processing large amounts of data, thereby achieving prediction and improving specific systems.[9]

Therefore, the dataset to be used has been employed to train models using various machine learning techniques; however, some methods have not yet been utilized, which could yield significant results.[16,17,18]

On the other hand, people should generate awareness to detect chronic diseases promptly, before symptoms manifest, and to receive a timely diagnosis at any stage of life.[10]

Therefore, with this study we intend to validate the hypothesis that it is possible to find improvements in cardiovascular disease prediction models by inspecting the performance of SVM and ANN algorithms, introducing kernel functions that have not been developed and used for this case, determining that this final result will be helpful in studies related to the object of study.[11]

Indeed, machine learning has demonstrated its value in medical contexts, serving as a novel and alternative tool that supports complex tasks, such as disease diagnosis. These technologies can ensure safety more effectively through rapid and reliable disease detection, and may even alleviate concerns for entire populations.

Therefore, if the application is not implemented and the comparison study of kernel functions in the framework of supporting medical diagnostics is not carried out, the opportunity for people to become aware of the significance of the effects of their habits and lifestyles would be lost.[19,20,21]

It is worth noting that public awareness of cardiovascular risk prevention is likely to be lost, which could lead to an increase in mortality and the premature mortality rate due to these diseases.

In short, maintaining contact between physicians and patients from a wellness and health status approach allows for greater and better closeness, as mentioned in some articles, the use of these technologies based on Artificial Intelligence, complement the knowledge of physicians and will enable them to spend more time with their patients and improve the shared decision-making process.[12] If the predictive model is not implemented,

the possibility of applying these new techniques in time would also be lost, it would increase the risks and the possibility of making a timely diagnosis would be lost, on the other hand, there would be no probability of analyzing large amounts of data quickly with consistency and accuracy, it would not be possible to create that model of study so important for knowledge. Additionally, it cannot contribute to the overall state of health of each person, and without good health, any individual would find it very difficult to achieve success.[22,23,24]

Cardiovascular diseases are diseases of interest in public health, which mainly produce deaths and progressive deterioration of health, and are not diagnosed promptly. Still, when these are presented in an advanced stage, it is of special interest to obtain a timely diagnosis of these diseases, and what percentage of cardiovascular risk ranges are explained in greater frequency, which are linked to factors that can be modified in the future, allowing timely measures to be taken and also contributes to promote health and prevent the disease.[25,26,27,28]

Early diagnosis of CVD is decisive to reduce mortality and morbidity figures, as well as the risks that can alert in a more timely manner, for this reason several studies have been proposed to detect cardiovascular diseases and dangers, some from the processing of the electrocardiography (ECG) signal and others from general public health measurements, in both cases supported by machine learning techniques (Machine Learning).

Among the different types of machine learning, kernel-based learning (Kernel Learning) is a method for pattern analysis, whose best-known application algorithm is Support Vector Machines (SVM). The general task of pattern analysis is to identify and examine common patterns or relationships in datasets.[13]

For example, using different deep learning models, such as RNN (Recurrent Neural Network), LSTM (Short-Term Memory Network) and even CNN (Convolutional Neural Network) achieved results of up to 96 % accuracy in detecting different types of heart diseases, among which are ventricular tachycardia, atrial fibrillation and sinus tachycardia,[4,14] another example in which they propose a novel method based on a hybrid Random Forest model with a linear component that achieved an accuracy of 88,7 % in the prediction of cardiovascular diseases.[4,15]

This study arises from the need to apply a predictive model based on an exhaustive analysis of the configurations in the SVM and ANN algorithms, considering that more kernel characteristics approach the optimal machine learning model, making use of previously analyzed datasets, and data collection, for training prediction algorithms, to know the percentage in the range of risk of cardiovascular disease, and which factors can be intervened promptly by the physician.

The research will be beneficial at the social level because it provides tailored information on the state of health of the population and support to medical professionals, enabling them to identify whether any external or internal factors influence the quality of life and avoid a deterioration that may be preventable.

The result will be a contribution to artificial intelligence in the field of machine learning, particularly in the area of comparative algorithm studies. This will include a comparison path, the implementation of kernel functions, models trained to make predictions about cardiovascular disease, and the visualization of such data.

Finally, this project is in tune with the dynamics of health promotion and disease prevention that Colombia has been working on in its different contextual frameworks, in addition to the fact that cardiovascular disease does not catalog and can be suffered by any person in their majority age, that is why it is essential to study and analyze possible variant factors such as body mass index, abdominal perimeter, socioeconomic level, habits, history of personal and family diseases, among others, in addition to obtaining an improvement in kernel functions, to promote integrity in the predictive models.

Delimitation

This project was developed through experiments on databases obtained from repository tools such as Kaggle datasets and UCI datasets. The models will be trained using Neural Networks (ANN) and Support Vector Machines (SVM). To evaluate the models, classification metrics associated with accuracy and time will be used.

The Kernels to be evaluated will be selected from a set of transcendental functions that meet the Karush-Kuhn-Tucker (KKT) conditions.

The configuration of the hyperparameters will be performed using random search techniques. The project will be developed in an estimated time of 18 months, starting in period A of the year 2023 and ending in period B of 2024.

DEVELOPMENT

Topics of the theoretical framework

Background

We consulted bibliographic sources from the last 5 years related to cardiovascular risk prediction or heart disease prediction using artificial intelligence. We found articles at both international and national levels, which in turn correlate with conceptual aspects and associated techniques for predicting risk and the type of data. Another review carried out was to consult sources with machine learning algorithms using kernel functions and support vector machines, where they were obtained:

International Background

According to L. Yang et al.[11] in the project entitled "Study of random forest-based cardiovascular disease prediction model in eastern China", published in 2020, selected 29930 subjects with high risk of cardiovascular disease (CVD) from 101056 people in 2014; regular follow-up was conducted using an electronic health record system. Logistic regression analysis revealed that nearly 30 indicators were associated with CVD, including gender, age, household income, smoking, alcohol consumption, obesity, excessive waist circumference, abnormal cholesterol levels, abnormal low-density lipoprotein levels, abnormal fasting blood glucose levels, and others. Several methods were employed to construct the prediction model, including multivariate regression models, classification and regression trees (CART), Naïve Bayes, Bagged trees, AdaBoost, and Random Forests. They used the multivariate regression model as a benchmark for performance evaluation, with an area under the curve (AUC) of 0,7143. The results showed that Random Forest was superior to other methods with an AUC of 0,787 and achieved a significant improvement over the benchmark. They developed a CVD prediction model for assessing 3-year CVD risk. It was based on a large population at high CVD risk in eastern China using the Random Forest algorithm, which would provide a benchmark for CVD prediction and treatment work in China. Accordingly,

more population-based studies of the CVD prediction model proposed in this research are needed, with a larger population, longer follow-up time, covering more locations in China with external validation.[4,16] This contribution and development of the project allow the comparison of classical techniques. How these act on various parameters to obtain similar or different results to those already studied, in addition to determining the variables that most contribute or affect the response variable, which for this occasion the significant risk factors in cardiovascular disease prediction are obtained, likewise the comparison of methods used for performance evaluation is achieved thus having that result as a reference and determining how significant is the choice of the variables to work for a given result.

On the other hand, according to Chávez Olivera O et al.[17] in their study entitled "Mobile Application to Predict the Probability of Belonging to the Cardiovascular Risk Group Using Machine Learning," published in 2022., Lima, Peru the present study focuses on creating a mobile application with the primary functionality of predicting cardiovascular risk group membership in individuals over 50 years old. To achieve this, research has been conducted on different variables and machine learning algorithms that allow this task to be accomplished. Thus, it was decided that the inference engine would be an ensemble model, where the final metaclassifier is a Naïve Bayes model and the base models are Random Forest and Logistic Regression. The application validation process was conducted by a cardiology specialist, who verified the model's accuracy level. It was observed that the Support Vector Machine, Random Forest, Naïve Bayes, and Logistic Regression models achieved accuracies of 87,00 %, 88,00 %, 87,00 %, and 86,00 %, respectively, with stability levels of 8,00 %, 6,00 %, 2,00 %, and 8,00 %. However, the Random Forest model has better accuracy, but it is more unstable. Therefore, the Naïve Bayes model was chosen, as it has a similar accuracy and is more stable than the others. The contribution provided by this research is that it shows the training results of different machine learning models, and the combination of these, the ensemble allows the construction of more accurate and stable algorithms, in addition to using Random Forest and Naïve Bayes models can determine the best accuracy, and have a point of comparison with the research.

According to Scavino M et al.[18], in their report entitled "Informe final publicable de proyecto Creación de algoritmos utilizando técnicas de clasificación supervisada y no supervisada para el diagnóstico de enfermedades cardiovasculares en una población de adultos mayores de bajos recursos en Uruguay" published in 2022, Uruguay. The present research aims to develop machine learning algorithms for identifying cardiac pathology, specifically atrial fibrillation, from single-lead electrocardiographic signal data using an electronic mobile device. Deep learning algorithms with the considered architectures did not show good performance. A larger amount of data could lead to an improvement in the classification capability of these algorithms. On the other hand, statistical learning techniques applied to a set of features extracted from the raw ECG signal showed better performance. It should be noted that the construction of these algorithms allows understanding how they work and how the final diagnosis is reached, the ability to interpret the internal mechanisms of the methods to provide a result generates a wide source of knowledge with the possibility of future development, in addition to finding possible causes for the use of different algorithms to reach the variation in diagnoses.

In a fourth study, according to Polero L et al.[19] in their project entitled "Prediction of risk of suffering an acute coronary syndrome using a Machine Learning algorithm (ANGINA)" published in 2020, Buenos Aires, Argentina. The present research aims to demonstrate the ability of machine learning classifiers to diagnose and predict an ACS in patients spontaneously consulting EMS with chest pain of unidentified etiology, during a 30-day follow-up period. A total of 161 patients consulting EMS with chest pain were analyzed. Objective and subjective pain characterization variables were recorded using a machine learning classifier. Thus, it was obtained that the average age was 57 ± 12, 72,7 % were male, and 17,4 % presented a previous coronary event. Acute coronary syndrome was present in 57,8 % with an incidence of AMI of 29,8 %, of which 35 %

required revascularization by CTA, and 9,9 % required RCA during the 30-day follow-up period. A Random Forest Classifier was employed as the classification model, yielding an area under the ROC curve of 0,8991, a sensitivity of 0,8552, a specificity of 0,8588, and a precision of 0,8441. The most influential predictor variables were weight (p = 0,002), age (p = 5,011e-07), pain intensity (p = 3,0679e-05), systolic blood pressure (p = 0,6068), and subjective pain characteristics (p = 1,590e-04). This project enables us to evaluate models using specific metrics, providing quantitative measures that indicate their performance and identify crucial variables for their development. It also reveals the behavior of these variables regarding the problem being solved.

National Antecedents

According to Peres[20] in his study entitled "Optimización De Un Modelo De Clasificación De Enfermedades Cardiovasculares Utilizando Técnicas De Aprendizaje Profundo Supervisado Y Despliegue De Dashboard Web," published in the year 2021, Cartagena. The development of this project is based on the optimization of a previous hypertension prediction model, from which the data analysis process has been improved. The initial model failed to generate an accurate prediction of cardiovascular disease classification due to misinterpretation and inadequate data cleaning. Likewise, optimization involved making adjustments to the model, improving the construction of the neural network, and refining the training process, as well as identifying activation techniques and optimal times to achieve the best results. Finally, data preparation allowed us to formulate, train and test a cardiovascular disease classification model built in the TensorFlow environment developed by Google, which resulted in a recurrent neural network composed of an input layer with 18 nodes, two hidden layers with 128 nodes each layer and an output layer of 3 nodes in which were the most optimal for the prediction which obtained an accuracy of 97 % in the validation of the model, and subsequently deployed making use of a web application for consultation of medical personnel the model developed by the above mentioned yielded as results, an accuracy greater than 86 %, seeing the percentages in the evaluation metrics that were used as: Accuracy, Recall, F1 Score and accuracy, which yielded results exceeding 80 % in the classification of risks. This research is valuable for the development of the project because when comparing the evaluation metrics, they show considerable results, which are:

This research is valuable for the project›s development because, upon comparing the evaluation metrics, it presents considerable results that can be attributed to the quality of the data obtained, as well as a tour of the TensorFlow environment, which enables the visualization of data across various areas of knowledge.

As a second research, according to Martinez[21] in his study entitled "Predicción De Enfermedades Cardiovasculares Mediante Algoritmos De Inteligencia Artificial," published in 2020, Málaga. This project focuses on implementing five different classification algorithms, analyzing how they fit the available data, and then creating a genetic algorithm that detects the optimal combination of parameters for each algorithm, yielding the best results in terms of accuracy. The algorithms were implemented using the Python scikit-learn library; no additional libraries were used for the genetic algorithm. The results showed that some algorithms are better adapted to evolution, i.e., accuracy increases over generations. Other algorithms showed a decrease in this value, suggesting that it is necessary to study for each type of algorithm the impact of each parameter, in addition to the values that in this project were considered constant: number of generations, number of individuals per generation, mutation and crossover probability and the size of the data set and the training, validation and testing subsets. This research is significant for the project due to the comparison it makes of classification algorithms and how they achieve better results through increased accuracy, contributing to one of the specific objectives in establishing a broad theoretical framework for the research.

As a third research, according to Florez[22] in his study entitled "Model of artificial intelligence as diagnostic support for the estimation of cardiovascular risk in patients attended under the modality of telemedicine in an IPS of the department of Sucre 2021," published in the year 2021, Sucre. The present project aims to investigate, through an experimental study, the incorporation of Artificial Intelligence as a diagnostic support in the care process for patients with cardiovascular risk. The data obtained from patients who attended the department of Sucre between 2018 and 2019 will be used. The problem of increased mortality in people with cardiovascular diseases, which can be prevented if diagnoses are made promptly, is presented. For this, the project seeks, through an experimental study, to incorporate Artificial Intelligence as diagnostic support in the care process of patients with cardiovascular risk. Among the variables to be monitored are age, sex, weight, height, body mass index, blood pressure, presence of diabetes, and dyslipidemia, among others. In turn, under the concept of Machine Learning, an intelligent algorithm will be developed with the ability to learn without being explicitly programmed, evaluating the potential risks of the individual and thus virtually assisting medical personnel in promoting, preventing, and diagnosing actions thoroughly and accurately for the modality of telemedicine care. Based on the characterization of the variables, we can identify the risk factors with the most significant impact on the incidence of cardiovascular disease in the study population. With the selection of a predictive model that best fits the characteristics of the population and the quality of the data provided, it will be possible to define a cardiovascular risk classification that can be adapted to the telemonitoring service

in both primary and secondary prevention. By utilizing these algorithms in a specific manner, we can view the prediction and these algorithms from a different perspective, as their intention is for the algorithm to learn without being explicitly programmed.

The study of this project allows us to confirm and corroborate the kernel functions used and those that better fit the data, depending on the use to which they are put. In addition to allowing improvements in certain aspects of the development, this approach offers different options.[11]

Statement of theoretical assumptions of the investigation

In this chapter, the theoretical bases of the present work, of a conceptual nature, are examined.

Cardiovascular diseases

These are diseases of the circulatory system, of diverse etiology and location. They are classified into four general types: ischemic heart disease, cerebrovascular disease, peripheral vascular disease, and other diseases.[23]

The identification of traditional risk factors has facilitated essential advances in the treatment of CVD. Still, despite the accumulated clinical evidence, the implementation of strategies to prevent cardiovascular disease remains far from optimal.[24]

Cardiovascular risk

Cardiovascular risk is defined as the probability of suffering a cardiovascular event in a specific period, which is usually established in 5 or 10 years, especially in patients who do not have cardiovascular disease, that is, in primary prevention, it is fundamental to demonstrate the intensity of the intervention, the need to develop pharmacological treatment and the periodicity of follow-up visits.[24]

Cardiovascular risk factors

Risk factors are those biological signs or acquired habits that occur more frequently in patients with a particular disease. Cardiovascular disease is multifactorial in origin, and one risk factor must be considered in the context of the others. The classic or traditional cardiovascular risk factors are divided into two large groups: non-modifiable (age, sex, and family history) and modifiable (dyslipidemia, smoking, diabetes, arterial hypertension, obesity, and sedentary lifestyle).

Although the impact of individual risk factors such as hypertension, dyslipidemia, smoking, and diabetes, among others, is well established and improves the prediction of cardiovascular risk.[24]

The higher the level of each risk factor, the higher the risk of having atherosclerotic cardiovascular disease, such as coronary heart disease.

- Obesity: The relationship between weight and heart disease comes as no surprise, as an obese person will have more fat and therefore a greater chance of CVD. "Regardless of metabolic health, overweight and obese people are at higher risk of coronary heart disease than lean people".[25]
- Physical activity: consistent exercise and a healthy lifestyle can, in general, very positively impact disease management: prevent or delay the onset of
- Type 2 diabetes, lower blood pressure, and help reduce the risk of heart attack and stroke.[25] The more vigorous the activity, the greater the benefit. However, even moderate-intensity activities can be beneficial when done regularly and over the long term. Exercise can help control cholesterol, diabetes, and obesity, as well as lower blood pressure in some people. Physical activity should be a daily activity. Walking 30 to 40 minutes, as many days as possible, but not fewer than 3 days, is a good form of exercise and has few contraindications.[26]
- Cholesterol levels: cholesterol accumulation is a major cause of atherosclerosis. It has been consistently shown that higher long-term LDL-cholesterol levels and non-high-density lipoprotein cholesterol concentrations are associated with an increased risk of CVD.[25]
- Glucose / Diabetes: diabetes is not only an alteration of blood sugar levels, but it also affects the overall system. Studies report a positive association between hypertension and insulin resistance.[25]
- Smoking: There is evidence that smoking causes approximately 1 in 10 deaths from cardiovascular disease. Tobacco smoke contributes to cardiovascular disease by increasing atherosclerotic plaque and the likelihood of thrombosis.[25] Tobacco smoke is the major risk factor for sudden cardiac death, and smokers are two to four times more at risk than nonsmokers. Cardiovascular risk decreases rapidly with smoking cessation.[26]
- Family history: children of parents with ischemic heart disease, especially if this has been premature (fathers before the age of 65 years, mothers before the age of 55 years) or with arterial hypertension, are more likely to develop it. There are minority forms of very high cholesterol (above 350 mg/dl) called familial hypercholesterolemia, which are due to hereditary disorders and carry a very high risk, even before menopause. In these cases, aggressive medical treatment with lipid-lowering agents is required.[26]

There are two methods for calculating cardiovascular risk: qualitative and quantitative. The qualitative ones are based on the sum of risk factors or the measurement of their level and classify the individual in: mild, moderate, high and very high risk; the quantitative ones, on the other hand, are based on risk prediction equations that give us a number that is the probability of presenting a cardiovascular event in a specific time, and the form of calculation is through computer programs or the so-called cardiovascular risk tables, which are handy tools for decision making in routine clinical practice.[24]

- Framingham study algorithm: the probability of occurrence of cardiovascular disease for a given variable, thus we find:
  - Men and women have different probabilities of developing cardiovascular disease.
  - Age is another determinant; the older the age, the greater the cardiovascular risk.
  - Tobacco use is a variable that increases cardiovascular risk independently of the other variables.
  - Cholesterol, HDL, and LDL levels are all variables that increase or decrease cardiovascular risk independently of the others.
  - High blood pressure levels and whether or not one has pharmacological treatment for hypertension.

The above are the classic variables analyzed by the study, but other researchers also seek to add some new ones:

- Ancestry, marital status, and education.
- Type of work, work rhythm and schedule, support from colleagues or supervisor.
- Abdominal circumference in people with diabetes or metabolic syndrome.[27]

Artificial Intelligence (AI) could be defined as the combination of algorithms proposed with the purpose of creating machines that present the same capabilities as human beings, either: thinking, feeling, solving problems, making decisions and even learning, AI comprises the area of Machine Learning, Deep Learning, Big Data and data science.

- Scikit Learn: Scikit-learn is a free machine learning library for Python. It features various algorithms, such as support vector machines, random forests, and k-nearest neighbors, and also supports Python numerical and scientific libraries, including NumPy and SciPy.[28]

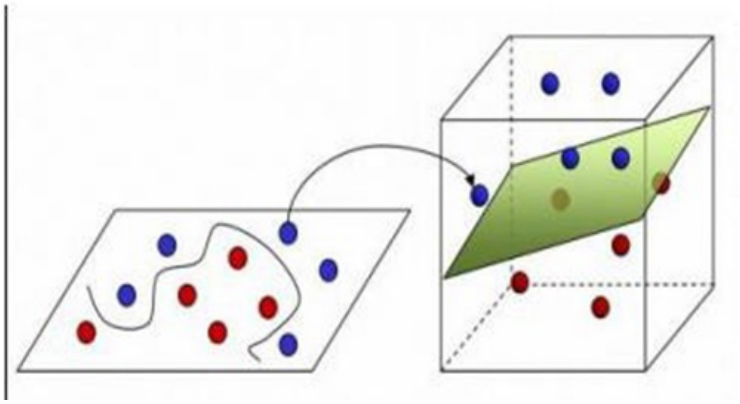
Support Vector Machines, Support Vector Algorithm (SVM)

It is a supervised learning algorithm used in various classification and regression problems, including medical applications, signal processing, natural language processing, and image and speech recognition.

The goal of the SVM algorithm is to find a hyperplane that best separates two different classes of data points. "Best possible way" implies the hyperplane with the widest margin between the two classes, represented by the plus and minus signs in the figure below. The margin is defined as the maximum width of the region parallel to the hyperplane that has no interior data points. The algorithm can only find this hyperplane in problems that allow linear separation; in most practical issues, the algorithm maximizes the flexible margin, allowing a small number of misclassifications.[29,30,31]

Kernel functions

Kernel Functions. These are mathematical functions used in Support Vector Machines. These functions are the ones that allow you to convert what would be a nonlinear classification problem in the original dimensional space into a simple linear classification problem in a higher-dimensional space. This M-dimensional space is known as Hilbert space.[32,33,34]



Source: https://www.researchgate.net/figure/260283043_fig13_Figure-A15-The-non-linear-SVM-classifier-with-the-kernel-trick

Figure 1. Two-dimensional to three-dimensional kernel trick

In figure 1, it can be seen how kernel functions work, taking a two-dimensional data distribution to three dimensions, this operation is generally called kernel trick. This trick allows to reduce the complexity of a function that separates the classes of a data distribution as demonstrated by Peluffo-Ordóñez, figure 1 for example in two dimensions could be separated by non-linear or segmented functions (Ellipse, Circle, Rectangle), while in three dimensions it could be separated by a linear function (Hyperplane) as demonstrated by Baudat & Anouar in the paper.[11]

For kernel functions to be considered kernel candidates, they must satisfy three fundamental initial conditions; they must be:

- Continuous.
- Symmetric.
- Positive.

These are the basic requirements to be expressed as a scalar product in a high dimensional space.[30,35,36]

There are several kernels commonly employed in machine learning libraries such as linear, RBF, polynomial and hyperbolic tangent whose definitions can be expressed in figure 2.[11]

Función kernel	Ecuación	Condición
Lineal	$k(x,x') = \langle x,x' \rangle$	$x,x \in \mathbb{R}$
RBF	$k(x,x') = \exp\left(-\sum_{i=1}^{d} \lambda_i (x_i x_i')\right)$	$\lambda_i > 0, \beta \in (0,2]$
Polinomial	$k(x,x') = (\alpha\langle x,x' \rangle + 1)^m$	$m \in \mathbb{N}, \alpha > 0$
Tangente hiperbólica	$k(x,x') = tanh(\alpha\langle x,x' \rangle + b)$	$a > 0, b < 0$

Source: Comparative kernel functions on supply prediction of alternative energy sources

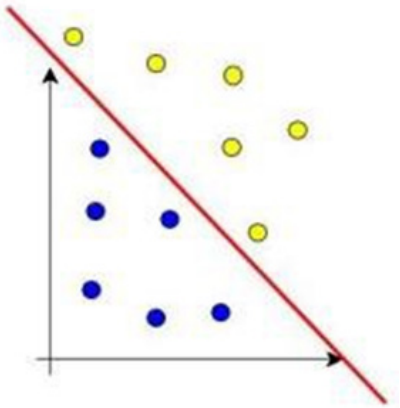
Figure 2. Formal definition of kernel functions

Although the functions in figure 2 are commonly used functions, there are many more kernel functions. These functions are used in supervised algorithms for regression and classification; and unsupervised for anomaly detection, analysis, clustering and feature extraction. Among the most prominent algorithms are Gaussian processes, Spectral clustering, Kernel Linear Discriminant Analysis, Kernel Principial Components Analysis, Kernel Canonical Correlation Analysis, Kernel Independent Component Analysis, SVM, ANN and many more.[11]

Pattern identification using linear models

Pattern detection is the process of finding in a set of chaotic data models capable of generalizing the behavior of the data to obtain classifications, predictions or anomaly detection. In order to obtain these patterns, the synergy of knowledge provided by several sciences such as mathematics, statistics, probability, computation, among others, is used. Currently, these techniques are encompassed in a branch called machine learning, divided into supervised, unsupervised and reinforcement learning techniques.[11]

- Linear classifiers: a linear classifier is one capable of finding the discrete class (y) to which a dataset belongs based on a linear combination of its attributes (x) as shown in figure 3.[11]



Source: https://docplayer.es/192405564-Comparativo-de-kernels-sobre-prediccion-de-oferta-de-fuentes-energy-alternatives.html

Figure 3. Example of linear classifier

As shown in figure 3 the linear classifier has drawn a separator, in this case a line that allows to deduce to which class it belongs (yellow or blue dots) so a new record, if it is positioned on the upper side of the line would be classified as a yellow dot, otherwise as blue.[11]

If the input of the classifier is a vector of real features $\vec{x}$, then the output result is:

$$y = f(\vec{w}.\vec{x}) = \left(f \sum_j w_j x_j \right),$$

Source: https://docplayer.es/192405564-Comparativo-de-kernels-sobre-prediccion-de-oferta-de-fuentes-energy-alternatives.html

Figure 4. Linear Classifiers

Where $\vec{w}$ is a real vector of weights and f is a function that converts the dot-to-dot product of the two vectors into the desired output. The vector of weights $\vec{w}$ learns from a set of training samples. Often f is a simple function that maps all values above a certain threshold to the first class and the rest to the second class.[31,37,38,39]

Some of the most commonly used linear classification algorithms for finding these classification patterns include linear discriminant analysis, linear Bayes classifiers, and regression.

- Supervised learning is a branch of Machine Learning, a method of data analysis that uses algorithms that iteratively learn from data to allow computers to find hidden information without having to program where to look explicitly. Supervised learning is one of three methods of how machines "learn": supervised, unsupervised, and optimization. Supervised learning addresses known problems by utilizing a set of labeled data to train an algorithm to perform specific tasks.[32,40]
- Unsupervised learning is a type of Machine Learning that is used to identify new patterns and detect anomalies. The data fed into unsupervised learning algorithms is not labeled. The algorithm (or models) attempt to make sense of the data themselves by searching for features and patterns.[32,41,42] It is important to note that for this project, we will focus specifically on supervised learning techniques for classification and prediction using linear models.
- Linear classifiers: a linear classifier is one capable of finding the discrete class (y) to which a dataset belongs based on a linear combination of its attributes (x).[11]

Database

A database is a set of information belonging to the same context, arranged systematically for later retrieval, analysis, and transmission. Nowadays, databases come in different shapes and sizes, depending on their application, such as in a library or a company's accounts. Databases originated to meet the need to store large amounts of information, that is, to preserve it over time and against deterioration, so that it could be accessed later. In this sense, the emergence of electronics and computing provided the indispensable digital element to store vast amounts of data in limited physical spaces, thanks to their conversion into electrical or magnetic signals.[33,43,44]

Measurement metrics

Regression metrics in machine learning and each Machine Learning model that uses it tries to solve a problem with another objective using a different dataset.[34,45,46]

MÉTRICAS DE REGRESIÓN EN APRENDIZAJE AUTOMÁTICO		
NOMBRE	**DEFINICIÓN**	**FÓRMULA**
(MSE) Error cuadrático medio	Es una métrica utilizada en estadística y aprendizaje automático para evaluar el rendimiento de un modelo de predicción. Se utiliza para medir la diferencia entre los valores predichos por el modelo y los valores reales observados. El MSE se calcula tomando la diferencia entre cada valor predicho y el valor real correspondiente, elevando al cuadrado esta diferencia y luego calculando el promedio de todos estos errores cuadráticos. En otras palabras,	$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ **Error Cuadrático Medio** Donde n es el número de observaciones en el conjunto de datos, y_i es el valor real observado y $\hat{y}_i$ es el valor predicho por el modelo.

	errores cuadráticos. En otras palabras, se obtiene la media de los errores cuadrados. Cuanto menor sea el valor del MSE, mejor será el modelo, ya que indica que las predicciones se acercan más a los valores reales. Un MSE de cero indicaría un modelo perfecto en el que las predicciones coinciden exactamente con los valores observados [34].	

(RMSE) Raíz del error cuadrático medio	Es una medida estadística comúnmente utilizada para evaluar la precisión de un modelo de regresión. Es una métrica que se deriva del Error Cuadrático Medio (MSE) y se calcula tomando la raíz cuadrada del MSE [34].	$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} = \sqrt{\mathrm{MSE}}$$ **RMSE** Donde MSE es el Error Cuadrático Medio. Al igual que el MSE, el RMSE mide la diferencia promedio entre los valores predichos por el modelo y los valores reales observados. La principal diferencia es que el RMSE tiene la misma unidad de medida que la variable objetivo, lo que lo hace más interpretable y fácil de comparar con los valores reales. El RMSE se utiliza ampliamente en problemas de regresión para evaluar qué tan cerca están las predicciones del modelo de los valores reales. Cuanto menor sea el valor del RMSE, mejor será la capacidad predictiva del modelo. Por ejemplo, si se tiene dos conjuntos de predicciones, A y B, y se dice que el MSE de A es mayor que el MSE de B, entonces se puede estar seguro de que RMSE de A es mayor que RMSE de B. Y también funciona en la dirección opuesta.
(MAE) -Error absoluto medio	Se calcula como la media de las diferencias absolutas entre los valores predichos por el modelo y los valores reales observados. El MAE proporciona una medida de la magnitud promedio de los errores de predicción sin tener en cuenta su dirección. Es una métrica comúnmente utilizada debido a su simplicidad y facilidad de interpretación. El MAE se expresa en las mismas unidades que la variable objetivo y	$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$$ **Error Absoluto Medio** Donde: • MAE: Mean Absolute Error • N: Número de observaciones en el conjunto de datos. • ∑: Sumatoria. • y_i: Valor real observado.

	cuanto menor sea su valor, mayor será la precisión del modelo en términos de predicción. Es importante tener en cuenta que el MAE no considera la magnitud relativa de los errores, por lo que todos los errores se tratan por igual en la métrica. Esto puede ser adecuado en ciertos escenarios donde se desea penalizar de manera uniforme todos los errores de predicción [34].	• $\hat{y}_i$: Valor predicho por el modelo.
(R^2) – R al cuadrado	El coeficiente de determinación, o R^2 (a veces leído como R-dos), es una medida estadística utilizada en el análisis de regresión para evaluar qué tan bien se ajusta un modelo a los datos observados. El R-cuadrado es un valor que varía entre 0 y 1, y se interpreta como el porcentaje de la variabilidad de la variable dependiente que es explicada por el modelo. un valor de R^2 cercano a 1 indica que el modelo explica una gran proporción de la variabilidad de la variable dependiente y se ajusta bien a los datos. Por otro lado, un valor de R^2 cercano a 0 indica que el modelo no explica de manera adecuada la variabilidad de la variable dependiente y no se ajusta bien a los datos [34].	$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$ **R Cuadrado**
R cuadrado ajustado (R^2)	Es una medida estadística relacionada con el coeficiente de determinación (R-cuadrado) que tiene en cuenta la cantidad de variables predictoras en un modelo de regresión y ajusta el R-cuadrado en función del número de	$$R^2_{adj} = 1 - [\frac{(1 - R^2)(n - 1)}{n - k - 1}]$$ **R Cuadrado Ajustado** Donde: • R^2 ajustado: Coeficiente de determinación

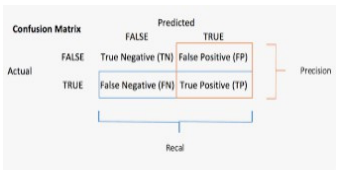
	predictores utilizados. El R-cuadrado ajustado se utiliza para evaluar y comparar modelos de regresión que tienen diferentes números de variables predictoras. El R-cuadrado ajustado penaliza el uso de variables predictoras adicionales que no aportan información significativa al modelo. A medida que se agregan más variables predictoras al modelo, el R-cuadrado ajustado disminuirá si esas variables no mejoran de manera	ajustado. • R^2: Coeficiente de determinación (R-cuadrado). • n: es el número total de observaciones • k: es el número de regresores independientes, es decir, el número de variables en su modelo, excluyendo la constante.

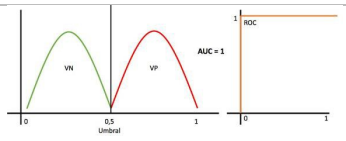
	sustancial la capacidad de explicación del modelo. Por lo tanto, el R-cuadrado ajustado tiende a ser más conservador que el R-cuadrado estándar y proporciona una medida más realista del ajuste del modelo [34].	
(MSPCE) – Error porcentual cuadrático medio	Es una medida de error utilizada para evaluar la precisión de un modelo de predicción en relación con los valores reales. Es una métrica que combina la magnitud de los errores y la proporción relativa de los errores en relación con los valores reales. El MSPE se expresa como un porcentaje, lo que facilita la interpretación de la magnitud del error en relación con los valores reales. Un valor más bajo de MSPE indica una mejor precisión del modelo, mientras que un valor más alto indica una mayor discrepancia entre las predicciones y los valores reales [34].	$$\text{MSPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2$$ **Error de Porcentaje Cuadrático Medio (MSPCE)** Donde: • n es el número total de muestras o datos. • y_i: son los valores reales o verdaderos. • $\hat{y}_i$: son los valores predichos por el modelo.
(MAPE) – Error porcentual absoluto medio	Es una medida de error comúnmente utilizada para evaluar la precisión de un modelo de predicción en relación con los valores reales. El MAPE calcula el promedio del porcentaje absoluto de error entre las predicciones y los valores reales. El MAPE se expresa como un porcentaje, lo que facilita la interpretación de la magnitud del error en relación con los valores reales. Un valor más bajo de MAPE indica una mejor precisión del modelo, mientras que un valor más alto indica una mayor discrepancia entre las predicciones y los valores reales [34].	$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$$ **Error Porcentual Absoluto Medio (MAPE)** Donde: • n es el número total de muestras o datos. • y_i: son los valores reales o verdaderos. • $\hat{y}_i$: son los valores predichos por el modelo.
(RMSLE) – Error logarítmico cuadrático medio	Es solo un RMSE calculado en escala logarítmica. Es una métrica útil cuando los valores tienen una amplia gama y hay una gran variabilidad en los datos. El RMSLE toma el logaritmo de los valores reales y los valores predichos antes de calcular el error cuadrático medio. Esto es útil cuando los valores objetivo abarcan un rango amplio y se desea penalizar de manera más equitativa los errores en diferentes magnitudes. El RMSLE se calcula como la raíz cuadrada del promedio de los errores cuadráticos de los logaritmos. Al tomar la raíz cuadrada, se obtiene una medida de error en la misma escala que los valores originales [34].	$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2} =$$ $$= \text{RMSE}(\log(y_i + 1), \log(\hat{y}_i + 1)) =$$ $$= \sqrt{\text{MSE}(\log(y_i + 1), \log(\hat{y}_i + 1))} =$$ **Error Logarítmico Cuadrático Medio (RMSLE)** Por lo tanto, esta métrica se usa generalmente en la misma situación que MSPCE y MAPE, ya que también conlleva errores relativos más que errores absolutos. RMSLE penaliza una estimación poco predicha mayor que una estimación sobre pronosticada. RMSLE se puede calcular sin la operación raíz, pero la versión rooteada se usa más ampliamente.

Figure 5. Regression metrics in machine learning

Classification metrics

MÉTRICAS DE EVALUACIÓN		
NOMBRE	**DEFINICIÓN**	**FÓRMULA**
Precision (Precisión)	Es una medida de evaluación utilizada en problemas de clasificación para medir la exactitud de un modelo al predecir correctamente las instancias positivas. Representa la proporción de predicciones positivas que son verdaderamente positivas en comparación con todas las predicciones positivas realizadas por el modelo. La precisión se expresa como un valor entre 0 y 1, donde 1 representa una precisión perfecta y 0 indica una precisión nula. La precisión es una métrica útil cuando el objetivo principal es minimizar los falsos positivos, es decir, evitar clasificar incorrectamente instancias negativas como positivas. Es especialmente importante en casos donde los falsos positivos tienen un impacto significativo o costoso. Es importante tener en cuenta que la precisión no tiene en cuenta los casos negativos que fueron correctamente clasificados como negativos (verdaderos negativos) [35].	$$\text{precision} = \frac{TP}{TP + FP}$$ **Fórmula de Precisión (Presicion)** Donde: • Verdaderos positivos (TP) son los casos positivos que fueron correctamente identificados por el modelo. • Falsos positivos (FP) son los casos negativos que fueron incorrectamente clasificados como positivos por el modelo.
RECALL (Exhaustividad)	El recall, también conocido como sensibilidad o tasa de verdaderos positivos, es una medida de evaluación utilizada en problemas de clasificación para medir la capacidad de un modelo de identificar correctamente las instancias positivas. El recall se expresa como un valor entre 0 y 1, donde 1 representa un recall perfecto y 0 indica un recall nulo. El recall es una métrica importante cuando el objetivo principal es minimizar los falsos negativos, es decir, evitar clasificar incorrectamente instancias	$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ **Exhaustividad (recall)** Donde: • Verdaderos positivos (TP): son los casos positivos que fueron correctamente identificados por el modelo. • Falsos negativos (FN): son los casos positivos que fueron incorrectamente clasificados como negativos por el modelo.

	positivas como negativas. Es especialmente relevante en casos donde los falsos negativos tienen un impacto significativo o costoso [35].	
F1-SCORE (Valor-F)	El F1-score también conocido como puntuación F1, es una medida de evaluación utilizada en problemas de clasificación que combina la precisión y el recall en una sola métrica. Representa el equilibrio entre la precisión y el recall de un modelo. El F1-score es una medida que varía entre 0 y 1, donde 1 representa un F1-score perfecto y 0 indica un rendimiento nulo. Un F1-score más alto indica un mejor equilibrio entre la precisión y el	$F1 = 2.\frac{precision.recall}{precision + recall}$ **F1-Score** • Precision es la proporción de verdaderos positivos (TP) sobre el total de instancias clasificadas como positivas por el modelo, es decir, la capacidad del modelo para evitar falsos positivos. • Recall es la proporción de verdaderos positivos (TP) sobre el total de instancias positivas reales, es decir, la capacidad del modelo para evitar falsos negativos.
	recall. El F1-score es especialmente útil cuando hay un desequilibrio entre las clases en los datos de entrenamiento [35].	
Accuracy (Exactitud)	La exactitud (accuracy) representa la proporción de instancias clasificadas correctamente sobre el total de instancias en el conjunto de datos [35]. La exactitud se calcula dividiendo el número de predicciones correctas (verdaderos positivos y verdaderos negativos) entre el número total de instancias en el conjunto de datos. La exactitud es la cantidad de aciertos del modelo sobre el total de instancias. Una exactitud de 1 indica que el modelo clasifica todas las instancias correctamente, mientras que una exactitud de 0 indica que el modelo no acierta ninguna instancia [35].	$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$ **Fórmula Accuracy**
CONFUSION MATRIX (Matriz de confusión)	Confusión o error Matrix es una tabla que describe el rendimiento de un modelo supervisado de Machine Learning en los datos de prueba, donde se desconocen los verdaderos valores. Se llama "matriz de confusión" porque hace que sea fácil detectar dónde el sistema está confundiendo dos	 Fig. 4. Matriz de Confusión

	• True Positives (TP): cuando la clase real del punto de datos	
	era 1 (Verdadero) y la predicha es también 1 (Verdadero). • Verdaderos Negativos (TN): cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso). • False Positives (FP): cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True). • False Negatives (FN): Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso) [34].	
ESPECIFICIDAD O TNR (Tasa negativa real)	Es el número de ítems correctamente identificados como negativos fuera del total de negativos. La especificidad se calcula dividiendo el número de verdaderos negativos (TN) entre la suma de los verdaderos negativos (TN) y los falsos positivos (FP). La especificidad proporciona información sobre la capacidad del modelo para evitar clasificar incorrectamente los casos negativos. Un valor alto de especificidad indica que el modelo tiene una alta tasa de aciertos en la identificación de los casos negativos [34].	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$ **Fórmula Especificidad Área**
ÁREA BAJO LA CURVA DE FUNCIONAMIENTO	Es una métrica utilizada en problemas de clasificación binaria	
DEL RECEPTOR (ROC) (AUC)	para evaluar la capacidad de un modelo para discriminar entre clases positivas y negativas. La curva ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación. El área bajo esta curva (AUC) proporciona una medida de la capacidad de discriminación del modelo: cuanto mayor sea el valor del AUC, mejor	 Fig. 5. Área Bajo la Curva

	será la capacidad del modelo para distinguir entre las clases. El AUC se calcula integrando el área bajo la curva ROC, que varía de 0 a 1. Un valor de AUC de 0.5 indica un rendimiento aleatorio o equivalente al azar, mientras que un valor de AUC cercano a 1 indica un rendimiento excelente del modelo en la clasificación. El AUC es una métrica popular en la evaluación de modelos de clasificación, especialmente cuando los conjuntos de datos están desequilibrados. Proporciona una medida agregada de la precisión del modelo en todos los umbrales de clasificación posibles y es independiente del umbral de decisión específico utilizado [34].	
PÉRDIDA LOGARÍTMICA	Es una métrica utilizada en problemas de clasificación para evaluar la calidad de las predicciones de un modelo de aprendizaje automático. La pérdida logarítmica mide la discrepancia entre las probabilidades predichas por el modelo y las etiquetas verdaderas de los ejemplos de entrenamiento. Cuanto más baja sea la pérdida logarítmica, mejor será el modelo en términos de su capacidad para asignar probabilidades adecuadas a las clases correctas. La pérdida logarítmica penaliza de manera más pronunciada las predicciones incorrectas con probabilidades cercanas a 0 o 1. En problemas de clasificación binaria, se busca minimizar la pérdida logarítmica para obtener un modelo con mejores predicciones [36].	$-(y\log(p)+(1-y)\log(1-p))$ **Pérdida Logarítmica** Donde: • y es la etiqueta verdadera del ejemplo (0 o 1). • p es la probabilidad predicha por el modelo de que el ejemplo pertenezca a la clase positiva.

Figure 6. Evaluation metrics

Spiral Methodology

Spiral development is a software life cycle model first defined by Barry Boehm in 1986,1 and is generally used in software engineering.[47,48,49]

The activities in this model consist of committing in a spiral, with each loop or iteration representing a set of activities. The activities are not prioritized, but the following activities are selected based on risk analysis,

starting with the inner loop. A spiral life cycle model takes into account the risks that arise when developing software.[50,51]

To do this, it begins by examining the possible development alternatives, selects the one with the most acceptable risk, and cycles through the spiral.[52,53,54] Suppose the customer wants to continue improving the software. In that case, the various new alternatives and risks are re-evaluated, and another round of the spiral is performed, until the software product developed is accepted and no further improvements are needed through another new cycle.[37]

Study variables

Nominal definition of the variables

The following tables show the dependent and independent variables present in the prediction study.

Nombre	Descripción	Tipo de Variable
Riesgo Cardiovascular	El riesgo cardiovascular es la probabilidad que tiene cada persona de sufrir una enfermedad cardiovascular, es decir, un infarto de miocardio, hemorragias cerebrales, embolias, etcétera.	Dependiente
Tiempo de entrenamiento	Es el tiempo que se toma un algoritmo en detectar los patrones de un conjunto de datos.	Dependiente
Tiempo de predicción	Conjunto de actuaciones que, siguiendo una metodología determinada y a través de los resultados de los modelos numéricos de predicción, van dirigidas a definir el valor más probable de los parámetros de tiempo.	Dependiente
Exactitud	La exactitud (accuracy) mide el porcentaje de casos que el modelo ha acertado [35].	Dependiente

Figure 7. Dependent study variables

Nombre	Descripción	Unidad de Medida	Tipo de Variable	Valores Normales
Edad	Edad de la persona	Años	Independiente	0-120 años
Peso	El peso es una medida de la fuerza gravitatoria que actúa sobre un objeto.	kilogramos	Independiente	IMC es entre 18.5 y 24.9, está dentro de los valores "normales" o de peso saludable. Si su IMC es entre 25.0 y 29.9, está dentro de los valores

				correspondientes a "sobrepeso". Si su IMC es 30.0 o superior, está dentro de los valores de "obesidad".
Talla	Medida de la persona en estatura	Centímetros	Independiente	1-200 centímetros
Género	El género se refiere a los roles, las características y oportunidades definidos por la sociedad que se consideran apropiados para los hombres, las mujeres, los niños, las niñas y las personas con identidades no binarias.	No aplica	Independiente	No Aplica
Tensión Arterial	Medida de la fuerza o presión de la sangre sobre las arterias cuando el corazón bombea.	Milímetro de mercurio	Independiente	Presión sistólica de menos de 120 y una presión diastólica de menos de 80.
Nivel de colesterol en sangre	Mide la cantidad colesterol y de ciertos lípidos en la sangre.	Miligramos de colesterol por decilitro de sangre	Independiente	Menos de 200 mg/dL
Nivel de glucosa en sangre	Mide los niveles normales de azúcar en sangre	Miligramos de glucosa por decilitro de sangre	Independiente	Menor que 100 mg/dl

Figure 8. Independent study variables

Operational definition of the variables

The independent variables correspond to variables that have already been measured in data sets provided by expert researchers of academic and research importance, in which it is confirmed in the SEA 2022 document for the global control of cardiovascular risk.[55,56,57] To obtain the results of this study, a dataset obtained through the Kaggle platform, a Data Science platform that allows users to acquire and publish datasets, was analyzed.

The dependent variables can be measured after each model is trained, based on which the accuracy represents the percentage of correct predictions compared to the total. Thus, it is the ratio between the cases well classified by the model (true positives and true negatives), and the sum of all cases.[20]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} - \text{TN} + \text{FP} + \text{FN}}$$

Ecuación 1. Accuracy

Precision refers to how close the result of a prediction is to the true value. Thus, it is the ratio between the positive cases well classified by the model and the total positive predictions. It is prioritized in cases where having a large number of false positives has a higher cost.[20,58]

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Ecuación 2. Precision

Sensitivity (recall) represents the true positive rate (TP). It is the ratio between the positive cases well classified by the model, with respect to the total number of positives.[20,59]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Ecuación 3. Recall

The F-value (F1-score) is used to combine the precision and recall measures into a single value, where 0 is the worst score and 1 is the best. It is used when it is best to reduce both false positives and false negatives.[20,60]

$$\text{F1score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Ecuación 4.F1-Score

$$\text{F1score} = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Ecuación 5. Modelo matemático de la evaluación de la medida-F

Hypothesis formulation

Research hypothesis

Alternative kernel functions improve the error-time relationship in cardiovascular disease prediction.

Null hypothesis

Alternative kernel functions do not improve the error-time relationship in cardiovascular disease prediction.

Alternate Hypothesis

Kernel functions decrease error, but take longer to give a prediction.

CONCLUSIONS

Cardiovascular diseases represent one of the leading causes of death worldwide, being a critical public health problem due to their high incidence and difficulty of early diagnosis. Despite advances in understanding risk factors, challenges persist in prevention, timely detection, and effective treatment. In this context, the use of emerging technologies, such as Artificial Intelligence (AI), particularly machine learning and kernel functions, is positioned as a promising alternative to support the diagnosis and monitoring of these pathologies.

This study proposes the use of classification algorithms, such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN), combined with new non-traditional kernel functions, to enhance the accuracy, efficiency, and utility of predictive models in cardiovascular risk identification. The validation of this proposal, based on reliable databases and rigorous metrics, demonstrates that it is possible to optimize the performance of the models through appropriate adjustments in the hyperparameters and the choice of suitable kernels, which can be translated into valuable tools for medical personnel in real scenarios.

Thus, it is concluded that the integration of machine learning models in the healthcare field not only provides a technical advantage in analyzing large volumes of clinical data but also has a social value by facilitating more accurate diagnoses, promoting prevention, and contributing to more timely and personalized medical care. The practical implementation of these technologies can have a positive impact on reducing morbidity and mortality associated with cardiovascular disease, as well as promoting healthy lifestyles and raising public awareness.

BIBLIOGRAPHIC REFERENCES

1. Patiño Zambrano CF. Dispositivo vestible inteligente para la generación de alertas tempranas de eventos cardiovasculares de riesgo. Envigado; 2022.

2. Sanofi Campus. Machine Learning y predicción de enfermedades cardiovasculares. 2022. https://campus.sanofi.es/es/noticias/machine-learning-prediccion-enfermedades-cardiovasculares

3. Gómez LA. Las enfermedades cardiovasculares: un problema de salud pública y un reto global. SciELO. 2011;1.

4. Pérez Leal LE, Buitrago Cárdenas JA. Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático. Bogotá: Universidad Antonio Nariño; 2021.

5. Gallego Valcárcel DA, Lucas Monsalve DF. Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardíaca con datos clínicos. Bogotá: Universidad Antonio Nariño; 2021.

6. Álvarez Vega M, Quirós Mora LM, Cortés Badilla MV. Inteligencia artificial y aprendizaje automático en medicina. Rev Méd Sinergia. 2020;5(8):12.

7. Mora Paz HA. Comparativo de Kernels sobre predicción de oferta de fuentes alternativas de energía. Pasto: UNIR La Universidad en Internet; 2019.

8. Friedman PA, Kapa S, López Jiménez F, Noseworthy PA. Inteligencia artificial en cardiología. Mayo Clinic; 2023. https://www.mayoclinic.org/es-es/departments-centers/ai-cardiology/overview/ovc-20486648

9. Aprende IA. Kernel y máquinas de vectores de soporte. https://aprendeia.com/kernel-maquinas-vectores-de-soporte-clasificacion-regresion/

10. Sowmya V, Sanjana K, Gopalakrishnan E, Soman KP. Inteligencia artificial explicable para la variabilidad de la frecuencia cardíaca en la señal de ECG. Health Technol Lett. 2020;7(6):146.

11. Wu H, Yang L, Jin X, Zheng P. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020;10(1):5245.

12. Chavez Olivera O, Galindo Honores L, Barrientos Padilla A, Cuadros Galvez M. Aplicación móvil para predecir la probabilidad de pertenecer al grupo de riesgo cardiovascular utilizando machine learning. En: XII Conf Iberoamericana de Complejidad, Informática y Cibernética. Lima; 2022.

13. Scavino M, Castrillejo A, Estragó Mérola VS, Luraghi López LE, Muñoz M, Álvarez Vaz R. Informe final publicable del proyecto de creación de algoritmos utilizando técnicas de clasificación supervisada y no supervisada para el diagnóstico de enfermedades cardiovasculares. Uruguay; 2022.

14. Polero LD, Garmendia CM, Echegoyen RE, Alves de Lima A, Bertón F, Lambardi F, et al. Predicción de riesgo de sufrir un síndrome coronario agudo mediante un algoritmo de Machine Learning (ANGINA). Rev Argent Cardiol. 2020;88(1).

15. Perez Tatis JD. Optimización de un modelo de clasificación de enfermedades cardiovasculares utilizando técnicas de aprendizaje profundo supervisado y despliegue de dashboard web. Cartagena; 2021.

16. Martínez EJ. Predicción de enfermedades cardiovasculares mediante algoritmos de inteligencia artificial. Málaga; 2020.

17. Carrascal Porras FL, Florez Prias LA. Modelo de inteligencia artificial como apoyo diagnóstico para la estimación de riesgo cardiovascular en pacientes atendidos bajo la modalidad de telemedicina. Sucre: UNAD; 2021.

18. Dolores C, Ordovás J. Genes, dieta y enfermedades cardiovasculares. Genética. 2007;5:71–118.

19. Martínez EJ. Predicción de enfermedades cardiovasculares. Málaga: Universidad de Málaga; 2020.

20. Sans Menéndez S. Enfermedades cardiovasculares. Barcelona: Institut d'Estudis de la Salut; 2011.

21. Perez Tatis JD. Optimización de un modelo de clasificación de enfermedades cardiovasculares utilizando técnicas de aprendizaje profundo supervisado y despliegue de dashboard web. Cartagena: Universidad del Sinú; 2021.

22. MathWorks. Máquina de soporte vectorial (SVM). https://es.mathworks.com/discovery/support-vector-machine.html

23. Ecured. Función kernel. https://www.ecured.cu/Funci%C3%B3n_Kernel

24. Wikipedia. Clasificador lineal. 2019. https://es.wikipedia.org/wiki/Clasificador_lineal

25. Tibco Data Science. ¿Qué es el aprendizaje supervisado? https://www.tibco.com/es/reference-center/what-is-supervised-learning

26. Montiel de Jesús A. Desarrollo de una aplicación para dispositivos móviles para la detección temprana de enfermedades cardiovasculares. Orizaba, México: TECNM; 2022.

27. Sitio BigData. Modelos de Machine Learning: Métricas de regresión. 2019. https://sitiobigdata.com/2019/05/27/modelos-de-machine-learning-metricas-de-regresion-mse-parte-2/

28. Martínez Heras J. Métricas de clasificación: precisión, recall, F1, accuracy. IArtificial; 2020. https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

29. Quintanilla L, Warren G, Kirsch S, Youssef V, Kershaw N, Killeen S, et al. Learn Microsoft: métricas de ML. 2023. https://learn.microsoft.com/es-es/dotnet/machine-learning/resources/metrics

30. Boehm B. Desarrollo en espiral. Wikipedia; 2012. https://es.wikipedia.org/wiki/Desarrollo_en_espiral

31. Ballina Ríos F. Paradigmas y perspectivas teórico-metodológicas en el estudio de la administración. UVMX; 2013.

32. Radrigán M. Método empírico-analítico. Wikipedia; 2022. https://es.wikipedia.org/wiki/M%C3%A9todo_emp%C3%ADrico-anal%C3%ADtico

33. IBM. Conceptos básicos de ayuda de CRISP-DM. 2021. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

34. Vallalta Rueda JF. CRISP-DM: una metodología para minería de datos en salud. https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

35. Toral Barrera JA. Redes neuronales. España: CUCEI; 2019.

36. Asunción A, Newman D. Repositorios. Rexa.info: Massachusetts Amherst; 2007.

37. Mora Paz H, Riascos JA, Salazar Castro JA, Mora G, Pantoja A. Comparación de funciones kernel para la predicción de la oferta energética fotovoltaica. RISTI. 2020;(E38):310-24.

38. Belanche LA, Villegas MA. Kernel functions for categorical variables with application to problems in the life sciences. 2023;(08034):1-3.

39. Esri. Spatial analysis in ArcGIS Pro. https://pro.arcgis.com/es/pro-app/latest/help/analysis/introduction/spatial-analysis-in-arcgis-pro.htm

40. Scriptología. Tutorial de Flask: desarrollando aplicaciones web en Python. 2024. https://scriptologia.com/tutorial-de-flask-desarrollando-aplicaciones-web-en-python/

41. Hunter J, Dale D, Firing E, Droettboom M. Introducción a pyplot. Matplotlib; 2012. https://es.matplotlib.

net/stable/tutorials/introductory/pyplot.html

42. Python. Biblioteca Pickle. 2001. https://docs.python.org/es/3/library/pickle.html

43. DataScientest. Uso de pandas en Python. 2023. https://datascientest.com/es/pandas-python

44. Manav N. Escribir bytes a archivo en Python. 2023. https://www.delftstack.com/es/howto/python/write-bytes-to-file-python/

45. Python. Biblioteca base64. https://docs.python.org/es/dev/library/base64.html

46. Navarro S. ¿Para qué sirve el train-test split? KeepCoding; 2024. https://keepcoding.io/blog/para-que-sirve-el-train-test-split/

47. Scikit Learn. Nystroem Kernel Approximation. 2007. https://scikit-learn.org/stable/modules/generated/sklearn.kernel_approximation.Nystroem.html

48. Li B, Lu P, chmccl v. Multiclass Neural Network. Microsoft Learn; 2023. https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/multiclass-neural-network?view=azureml-api-2

49. Imbert A, Lemaitre G. sklearn.svm.SVC. Scikit-Learn; 2024. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

50. Ruiz M. ¿Qué es Redux? OpenWebinars; 2018. https://openwebinars.net/blog/que-es-redux/

51. Moes T. ¿Qué es Windows? SoftwareLab; 2023. https://softwarelab.org/es/blog/que-es-windows/

52. D. A. Qué es Bootstrap. Hostinger; 2023. https://www.hostinger.mx/tutoriales/que-es-bootstrap

53. Monk R. What is Python used for. Coursera; 2023. https://www.coursera.org/mx/articles/what-is-python-used-for-a-beginners-guide-to-using-python

54. Valiente FT. Aprendizaje por refuerzo. IIC UAM. https://www.iic.uam.es/inteligencia-artificial/aprendizaje-por-refuerzo

55. Arceo Vilas AM. Estado nutricional y adherencia a la dieta mediterránea en población mayor de 40 años: IA vs estadística clásica. A Coruña: Universidad de Coruña; 2020.

56. Sánchez Santos JM, Sánchez Fernández PL. Predicción de eventos cardiovasculares y hemorrágicos en pacientes con doble antiagregación con modelos ML. CREDOS. Salamanca; 2020.

57. Gallego Valcárcel DA, Lucas Monsalve DF. Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardiaca con datos clínicos. Bogotá: Universidad Antonio Nariño; 2021.

58. Lozada J. Investigación aplicada. Dialnet UNIRIOJA. 2014;3(1):47-50.

59. Núñez Cárdenas FJ, Zavaleta Chi IDC, Felipe Redondo AM, Meléndez Hernández J. Aplicación de minería de datos para tipificación de ECV en alumnos universitarios. México; 2018.

60. Martínez J. Más allá del accuracy: precision, recall y F1. Datasmarts; 2019. https://datasmarts.net/es/mas-alla-del-accuracy-precision-recall-y-f1/

FINANCING

None.

CONFLICT OF INTEREST

None.

AUTHORSHIP CONTRIBUTION

Conceptualization: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Data curation: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Formal analysis: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Research: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Methodology: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Project Management: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Resources: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Software: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Supervision: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Validation: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Visualization: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Writing - original draft: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

Writing - proofreading and editing: Michael Rafael Rodríguez Rodríguez, Claudia Alejandra Delgado Calpa, Héctor Andrés Mora Paz.

https://doi.org/10.56294/shp2026364 ISSN: 3072-7677