Data mining and cardiac health: predicting heart attack risks

Inoc Rubio Paucar[1], Laberiano Andrade-Arenas[2]
[1]Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú
[2]Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú

Article Info

Article history:

Received Mar 2, 2024
Revised Nov 3, 2024
Accepted Nov 11, 2024

Keywords:

Cardiac health
Data mining
Heart attacks
Prediction
Risks

ABSTRACT

In a context where heart attacks continue to be a global health concern, the lack of precision in predicting who is at higher risk poses a critical challenge due to the variability of risk factors and complex interactions among them. The research aims to develop predictive models for heart attack risks using data mining techniques, employing the knowledge discovery in databases methodology (KDD) and the k-means algorithm with RapidMiner studio. The primary objective is to identify patterns and risk profiles, allowing for early identification of at-risk individuals, considering factors like obesity, diabetes, alcoholism, and stress, to reduce preventable deaths and improve cardiac healthcare. This innovative approach combines cardiac health, data mining, and KDD methodology to address the challenge of predicting heart attack risks and has the potential to enhance medical care and save lives. The predominant results obtained were that cluster 1 with a fraction of 0.312 and a percentage of 31.2% of the attribute diabetes was one of the most prevalent causes of cardiac risk. Finally, the research concluded that people with diabetes are more likely to have cardiac risk associated with dietary factors or consumption of other substances.

Corresponding Author:

Laberiano Andrade-Arenas
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener
Lima-Perú
Email: landrade@uch.edu.pe

1. INTRODUCTION

In the global health context, cardiovascular disorders have emerged as an urgent concern. Heart attacks remain one of the leading causes of morbidity and mortality worldwide [1], [2]. The World Health Organization (WHO) consistently reports that these cardiovascular events impose a significant burden on both public health and healthcare systems worldwide [3]. However, it is crucial to recognize that the risk of suffering a heart attack is significantly influenced by factors such as diabetes, obesity, alcoholism, and stress.

The issue at hand lies in the complexity of accurately predicting who is at the highest risk of suffering a heart attack due to the intricate interplay of multiple risk factors. Diabetes, a widespread metabolic disease, contributes to the risk by affecting cardiovascular health [4], [5]. Obesity, another significant global health concern, is closely related to heart attack risks. Excessive alcohol consumption can substantially elevate these risks. Furthermore, the detrimental impact of chronic stress on heart health is well-documented. These risk factors, which often coexist in individuals, make prediction a multifaceted challenge [6], [7]. Current approaches often fall short of providing precise predictions, resulting in delayed diagnoses and less effective healthcare responses, leading to a significant number of preventable deaths.

The justification for this research is robust. Improving the prediction of heart attack risks in the context of diabetes, obesity, alcoholism, and stress is of paramount importance for public health and individual well-being [8]. The application of data mining techniques presents a promising avenue to address this issue, as it can unveil hidden patterns in extensive clinical and biomedical datasets, enabling early and precise identification of at-risk individuals. This research stands to benefit not only patients but also healthcare professionals and policymakers by providing more effective healthcare, reducing mortality rates, and alleviating the financial burden on healthcare systems.

In the realm of public health, a critical challenge is posed by the high mortality rate associated with cardiovascular diseases (CVD), specifically heart attacks [9], [10]. This health issue is compounded by the growing prevalence of well-established risk factors, including obesity, diabetes, alcoholism, and stress. Preventing premature deaths related to these diseases has become a public health priority. Obesity has become a global epidemic, affecting individuals of all ages and demographics. This condition significantly contributes to the risk of CVD, including heart attacks. Similarly, diabetes, a chronic metabolic disease, is closely linked to heart disease and can dramatically increase mortality rates [11]. Alcoholism, when it evolves into excessive and chronic consumption, can substantially elevate the risk of heart attacks and other cardiac conditions. Lastly, chronic stress, stemming from the pressures of daily life, has been identified as a significant risk factor for CVD. To address this challenge and prevent the number of premature deaths related to these risk factors, a research proposal is put forth based on the application of data mining techniques. The knowledge discovery in databases (KDD) methodology will serve as the framework to unearth valuable patterns and insights from clinical and biomedical data. The K-means algorithm will be applied using RapidMiner studio to cluster and classify individuals, identifying profiles of patients with a higher likelihood of experiencing heart attacks.

This study aims to develop predictive models of heart attack risk using data mining techniques, taking into account the underlying causes leading to death in cardiac health. We aim to accurately predict who is most at risk in this complex network of risk factors, including diabetes, obesity, alcoholism, and stress, and to provide a basis for proactive medical decision-making. By achieving this goal, we will contribute to reducing heart attack-related deaths and improving the quality of life of those at risk. This scientific article will explore the construction and evaluation of these models, enriching the body of knowledge in cardiac health and data mining.

2. LITERATURE REVIEW

The present literature review focuses on the exciting field of data mining applied to cardiac health, specifically in the prediction of heart attack risks. This section aims to analyze research conducted by various experts and scientists in this field, highlighting their significant contributions while identifying limitations and opportunities for advancing this crucial aspect of healthcare. The combination of data mining technology and cardiac health has proven to be a promising approach for the early and accurate identification of risk factors, enabling more personalized and effective care for patients at risk of CVD.

The primary aim of this research was to assess the relationship between changes in the behavior of smoking patients and the risk of fatal incidence of CVD in individuals with type 2 diabetes mellitus (T2DM). The study encompassed a significant cohort of 349,137 smokers who were categorized into five distinct groups: those who quit smoking, reducers I with a reduction of less than 50%, reducers II with a moderate reduction of 20-50%, those who maintained their habit within a variability range of ±20%, and those who increased their cigarette consumption by a minimum of 20%. Importantly, it was revealed that among T2DM patients, quitting smoking was significantly associated with a decrease in both the incidence of CVD and the overall mortality rate from all causes [12]. These findings underscore the significance of smoking cessation as a fundamental preventive measure in managing cardiovascular health in individuals with type 2 diabetes.

The study examined 151 patients who were at risk of experiencing acute myocardial infarction according to the evaluation of the ST-segment elevation myocardial infarction (STEMI) after primary percutaneous coronary intervention (PCI). The study was conducted in a single-center fashion. Among the 151 STEMI patients who underwent primary PCI, 71 were subjected to an analysis of major adverse cardiovascular events (MACE) that occurred during their hospitalization. The predictive model yielded an area under the curve of 0.778 (95% CI: 0.690-0.865). Notably, this model demonstrated good calibration and clinical utility through decision and calibration curves [13]. These results emphasize the effectiveness and clinical relevance of the predictive model in assessing and managing cardiovascular events in STEMI patients undergoing primary PCI.

On a global scale, cardiovascular incidents rank among the leading causes of morbidity and mortality. Hence, the study aimed to examine 520 individuals who had experienced at least one cardiovascular event, assessing the associated risk factors related to the frequency and behavior of the ankle-brachial index (ABI). The study led to the conclusion that one cardiac event often paves the way for subsequent cardiovascular incidents. However, it is noteworthy that after a stroke, the likelihood of experiencing another stroke or a cardiac event is comparable [14]. These findings underscore the significance of continuous monitoring and effective management of risk factors in individuals who have undergone cardiovascular events to prevent future complications.

This investigation delves into the assessment of physical activity over the past 12 months and its relation to heart disease in breast cancer survivors. For this study, assessments were conducted on cohorts of individuals who had successfully battled breast cancer, with an average age ranging from forty to fifty years. The study encompassed 599 participants who had triumphed over their cancer treatment, with a median age of 55.5 years and a median time since treatment of 10.2 years. Significantly, an increase in physical activity was found to correlate with an improvement in the syndrome of superior vena cava (SVC) in individuals grappling with long-term conditions. This discovery highlights that boosting physical activity can enhance cardiovascular health, particularly for less active survivors [15]. These findings underscore the potential benefits of physical activity in improving cardiovascular health among breast cancer survivors.

Nattokinase has shown promising effects on heart health, as indicated by the research findings. The study, involving 546 participants, revealed that a relatively low dose of nattokinase hurt blood cholesterol, including both high-density lipoprotein (HDL) cholesterol and total cholesterol levels. The results from this study affirm that nattokinase can be used as an effective complementary treatment for hypertension. However, it is worth noting that nattokinase supplements in relatively low doses may not have a significant hypocholesterolemic effect [16]. These findings underscore the potential of nattokinase in managing hypertension, though higher doses may be necessary for a substantial impact on lipid levels.

The current study employed a machine learning (ML) based approach to predict cardiovascular health risk in individuals with coronary heart disease. For this research, the random forest (RF) model algorithm, encompassing four genetic loci and four epigenetic loci, was utilized, and data from a total of 1,180 individuals and 524 subjects were considered. As a result, the analysis demonstrated a sensitivity of 0.70 and a specificity of 0.74, indicating the model's ability to accurately identify risks. However, it is important to note that the sensitivity of the cardiovascular atherosclerotic risk estimator (ASCVD) test was 0.20, while the Framingham risk estimator yielded a sensitivity of 0.38 [17]. These findings highlight the utility of the RF model in assessing cardiovascular risk in coronary heart disease patients, despite variations in sensitivity among risk estimators.

The objective of this research is to establish a long-term pattern model of cardiovascular health (CVH) from childhood and assess its association with subclinical atherosclerosis in middle age. The cohort study utilized data from five cardiovascular cohort studies and included a total of 9,388 individuals aged 8 to 55 years who underwent a minimum of three examinations. Within this sample, five trajectory groups were identified, among which 5,146 [55%] were female, 6,228 [66%] were of Caucasian ethnicity, and the baseline mean age was 17.5 [7.5] years. These groups encompassed high-late decline (1,518 participants [16%]), high-moderate decline (2,403 [26%]), high-early decline (3,066 [32%]), and intermediate-late decline (1,475 [16%]). According to the study, CVH showed a decline from childhood to adulthood. The promotion and preservation of ideal CVH in early life may be linked to a reduced risk of future cardiovascular events [18]. This investigation sheds light on the importance of maintaining cardiovascular health from childhood to mitigate the risk of CVD in later years.

The primary aim of this research is to determine whether living with a higher chronic valvular heart disease CVH score in midlife is correlated with a reduced risk of hypertension, diabetes, chronic kidney disease, cardiovascular events, and its subtypes (such as coronary heart disease, stroke, congestive heart failure, and peripheral artery disease), as well as all-cause mortality in later stages of life. To conduct this prospective cohort study, data from 1,445 participants in the framingham heart study offspring, collected from 1991 to 2015, were analyzed. A composite score was created using various variables, including body mass index, fasting blood glucose levels, total serum cholesterol levels, dietary habits, physical activity, resting blood pressure, and smoking status. The findings from this investigation suggest that spending a longer duration of time with improved CVH in midlife may yield healthy cardiometabolic benefits and could be associated with reduced mortality in later life [19].

In the review of the eight articles, several limitations have been identified, including the lack of precision in predicting cardiac risks and the underutilization of data mining techniques for clinical data analysis. An important improvement proposal would be to effectively incorporate data mining and ML into future studies, enabling better prediction of cardiac risks from larger and more detailed datasets. This could help identify more subtle patterns and risk factors, leading to more personalized and effective interventions to reduce cardiac risks in patients with type 2 diabetes and other cardiovascular conditions.

3. METHOD

3.1. Definition of the KDD methodology

The KDD methodology is a process that allows predictions to be made in data mining through a series of stages. These processes include selection, processing, transformation, and interpretation [20] as mentioned in Figure 1. The management of this process is iterative and interactive, which means that it is possible to return to previous stages without affecting the already established processes. This technique allows the identification of significant patterns in the large volumes of data handled by the database on the topic of heart attack risks. In this database, important criteria are taken into account that trigger the most frequent causes that lead patients to develop heart attacks and even to die.

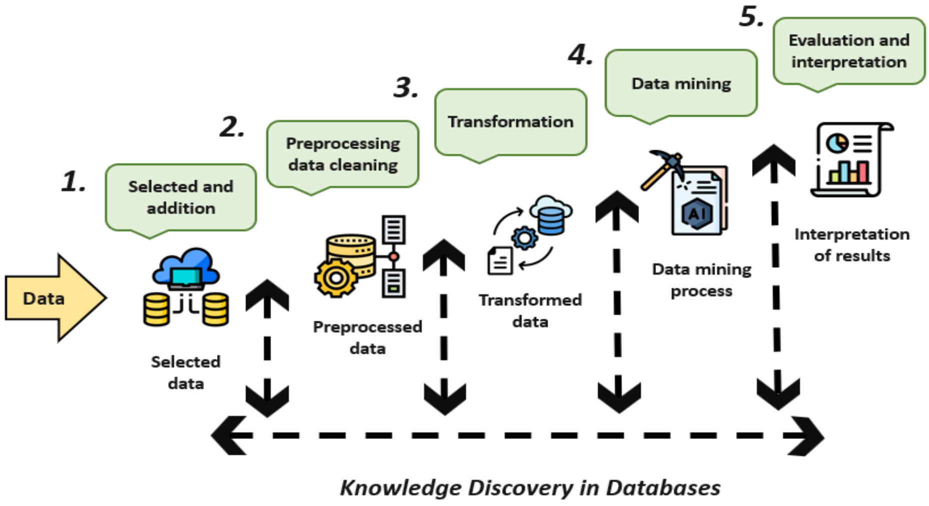

Figure 1. KDD methodology process

3.2. KDD methodology stages

This section will explain the stages of the methodology selected in the research. Each stage will be explained by developing the concepts according to the proposed topic and clarifying certain criteria in the development of the data mining model.

3.2.1. Data selection

The information is contained in a database with 8763 records that deal with the causes that lead patients to get heart problems and the classification of their lifestyle. The data that are selected during the information search process will be used for the knowledge discovery process which involves defining the relevant data for the implementation of the model [21]. The following processes will focus on common problems with databases, as a specific database is expected to contain some noisy information.

a. Data processing techniques

In this section, we will specify some criteria and data mining techniques for the selection process specified below.

b. Filtering criteria by absolute value

|X| > T, where |X| s is the absolute value of the variable and T is the threshold. This criterion is used to select data based on the absolute value of a variable. Data is selected if the absolute value of variable X is greater than a threshold T.

c. Filter criteria by range

The formula states that Min < X < Max, where Min *and* Max are the minimum and maximum values set. It is used to select data that falls within a given range. A variable *X* is selected if it is within the *Min* y *Max* limits.

d. Frequency filtering criteria

The formula is as follows Count(X) > N, *where* Count(X) is the frequency of a variable X *and* N *is* the minimum number of occurrences needed. This criterion is used to select data based on the frequency of a variable. Variable *X* is selected if it occurs at least *N* times.

e. Filtering criteria by percentage

The formula says that the % of X is greater than P, where % of X is the percentage of occurrence of a variable X and P is the minimum percentage needed. It is used to select data based on the percentage of occurrence of a variable. It is chosen if the percentage of occurrence of *X* is greater than *P*.

f. Correlation filtering criteria

The formula states that Corr (X, Y) is the correlation between variables X and Y, and C is the minimum correlation value required. This criterion is used to choose data based on the relationship between two variables. It is chosen if the correlation between *X* and *Y* is higher than *C*.

g. Filtering criteria by variability

The formula says that *Var(X)* is the variance of a variable *X* and *V* is the minimum variance value required. It is used to choose data according to the variability of a variable. *X* is chosen if the variance of *X* is greater than *V*.

h. Formulate the correlation criterion

Data are selected if the correlation between variables exceeds a specific threshold as mentioned in (1). The fundamental formula for this criterion is as (1),

$$|Corr(X,Y) > C| \tag{1}$$

where:

- | *Corr (X, Y)* | is the absolute value of the correlation between variables *X* and *Y*.
- *C* is a required minimum correlation value. If |*Corr (X, Y)* | is greater than *C*, it is selected.

Pearson's correlation coefficient, a measure ranging from -1 to 1, is often used to determine the correlation between two variables. A value of one indicates a perfectly positive correlation, a negative value indicates a perfectly negative correlation and a value of zero indicates no correlation. The general formula for calculating the pearson correlation coefficient as specified in (2) between *X* and *Y* variables is as follows:

$$Corr(X,Y) = \Sigma\,[(Xi - \bar{Y})]/[\sqrt{\Sigma\,(Xi - \bar{X})^2 * \Sigma\,(Yi - \bar{Y})^2}\,] \tag{2}$$

where:

- The observation values of *Xi* and *Yi* are *X* and *Y*, respectively.
- $\bar{X}$ and $\bar{Y}$ are the averages of *X* and *Y*.

In Figure 2, you can see the different steps of the workflow that have been designed to accomplish this task. Each step of the process is carefully set up to ensure that the specific requirements of the data analysis are met. The operators are connected in a logical manner, ensuring that the data is handled and processed appropriately at each step of the process.

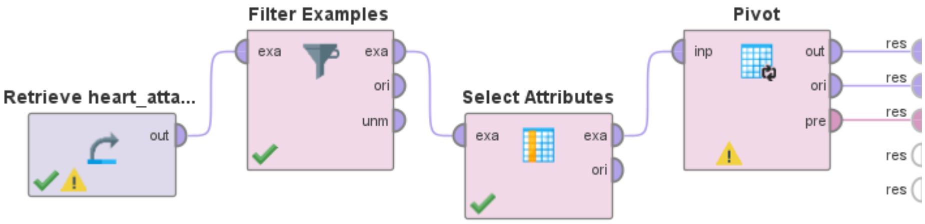

Figure 2. Data selection stage

3.2.2. Data preprocessing

At this stage, most of the information in a database presents noise, which requires cleaning to be prepared for the next stage [22]. In the case of the database found, according to the topic raised, the information was analyzed and empty fields and outliers were found, which allows us to perform a normalization that allows us to solve certain problems with the information.

a. Techniques for data preprocessing

In this phase, a series of steps are followed to consolidate the information in this process. For this purpose, the selection of operators within the RapidMiner studio tool is taken into account to carry out the established process.

b. Missing value cleaning

Formulas such as mean, median or a value predetermined by a ML model can be used to impute missing values.

- Outlier elimination: This may include identifying values that are above or below a specific statistical threshold, but are generally not expressed in a single formula.
- Data transformation: To normalize the, use a formula such as (X - X_min)/(X_max - X_min), where X is the original value and X_min and X_max are the minimum and maximum values of a range. This involves converting categorical variables into numerical variables using techniques such as one-hot coding.

Figure 3 shows how operators perform data preprocessing using established methods. This removes noise from the selected research database. On the other hand, to avoid errors in the subsequent model processes, the clean data is prepared for the following process.

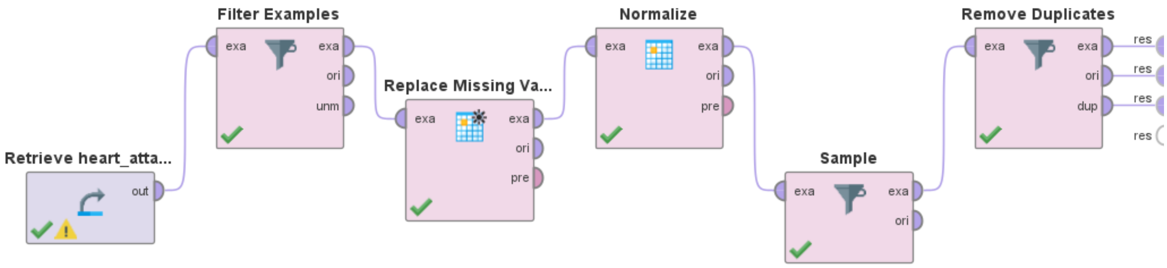

Figure 3. Data preprocessing stage

3.2.3. Data transformation

In this stage, the data preprocessed in previous phases is converted into a more accurate representation for the desired analysis according to the objectives outlined within the research [23]. This includes reducing the dimensionality of the data by creating new features for the established model. In this context, the concepts of the data transformation stage will be applied and explained in detail using mathematical formulas in each process.

a. MIN-MAX normalization

Normalization-Max is a method to modify the values of a variable to be within a given range, generally between 0 and 1. This facilitates the comparison between different variables by allowing the values to have a uniform scale, as specified in (3)

$$Form:\ X_{norm} = \frac{Xi - X_{min}}{X_{max} - X_{min}} \tag{3}$$

b. Z-score (standardization)

The process of, also known as the Z-score, consists of adjusting the values of a variable so that they have a mean of 0 and a standard deviation of 1. This facilitates the comparison and analysis of variables in the same context by eliminating differences in the scale of the variables, as conformed to form (4).

$$Form:\ X_Z = \frac{X - \mu}{\sigma} \tag{4}$$

c. Logarithmic transformation

The natural logarithm is applied to the values of a variable during the logarithmic transformation. It is commonly used to stabilize variance and reduce skewness in data that show a right-skewed distribution, such as financial data or exponential growth data, as specified in form (5).

$$Form:\ X_{log} = \log{(X)} \tag{5}$$

In RapidMiner, there is no specific operator for this, but you can use the "Generate Attributes" operator to apply the logarithmic transformation.

d. One-hot coding

For categorical variables, the formula creates a binary column for each category, with a value of 1 if the category is present and a value of 0 if it is not present. Operator in RapidMiner: "Nominal to Numerical" to convert categorical variables to numerical with multiple binary columns.

e. Dimensionality reduction (PCA)

The PCA formula involves matrix calculations and spectral decomposition; there is a unique formula. The operator uses "PCA" to perform principal component analysis in RapidMiner.

f. Imputation of missing values (average)

Missing-value imputation replaces the missing values of a variable with the average of that variable in the data set. This is a common way to deal with missing values and avoid data loss.

- Form $X_{imputed} = \mu$ (6), where μ is the mean of the variable.
- In RapidMiner, the operator must "replace unnecessary values" using the imputation strategy set to "mean.".

Figure 4 shows the operators used in the data transformation process. These operators play a crucial role in allowing the data to be properly prepared for the next process in the workflow. Data transformation is a fundamental step in data analysis, as it ensures that the data are in the correct format and contain the relevant information needed to consolidate the model proposed in the project objectives.

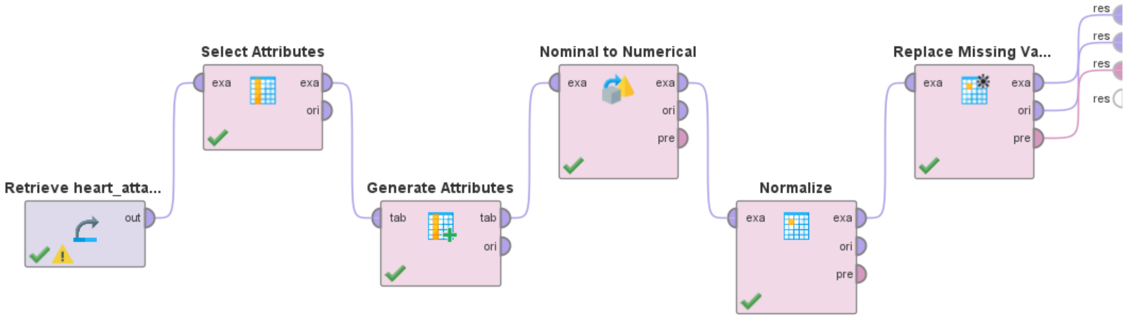

Figure 4. Data transformation stage

3.2.4. Data mining

The process of data mining is the application of techniques that allow the detection of patterns related to the raised topic [24]. This includes the application of automatic learning algorithms that have the function of predicting these patterns within the model in the data analysis. Within the research framework, the K-means algorithm belonging to the group of classification algorithms will be applied, taking into account the proposed objective.

– K-means algorithms

The mean or the mean between its points, which refers to the centroids of the environment, represents this algorithm as a group. The advantage of this representation lies in the fact that it has an immediate graphical and statistical meaning through its centroids [25], [26]. The group technique in data mining is a ML algorithm that aims to divide data sets into groups such that the points in each group are similar.

When data have no prior label, clustering (as opposed to classification) divides data into groups based on their similar attributes. Partitioning methods, such as k-means, hierarchical (network analysis map), density-based (DBSCAN) [27], [28], and grid-based, are among the clustering techniques. To achieve the research objectives, the partitioning algorithm, also known as k-means, was used. The most effective partial clustering algorithm is K-means clustering. This method uses a partitioning strategy during the clustering process to gradually reduce the data gap between each clustering kernel [29], [30]. For the application of the K-means algorithm, certain processes are applied that have the function of grouping the data in clusters to consolidate the results of the prediction, as shown in Figure 5. On the other hand, Figure 6 shows the structure and operation of the K, with concepts established for its application within the K-means process logic to reach a result that proposes the objective to be achieved in the research.

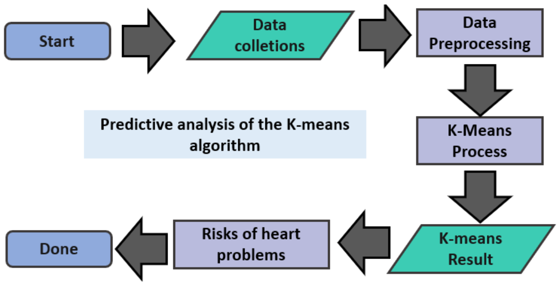

Figure 5. K-means algorithm process

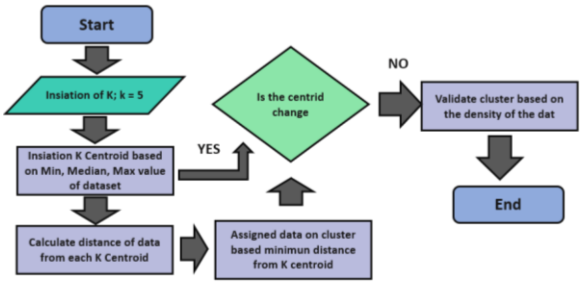
Start
Insiation of K; k = 5
Insiation K Centroid based on Min, Median, Max value of dataset
Calculate distance of data from each K Centroid
Assigned data on cluster based minimun distance from K centroid
YES
Is the centrid change
NO
Validate cluster based on the density of the dat
End

Figure 6. Structure of the K-means algorithm

– Euclidean distance

The term Euclidean “distance” is given between the distances of two points in a triangle of Euclidean shape. A Euclidean that provides concepts (two-dimensional space or of higher dimensions) is used to dimension a specific [31]. Also, it refers to a metric related to the K-means algorithm and in other contexts. The formula for the Euclidean distance between two n-dimensional space points is stated as follows, as shown in (7):

$$D(p,q) = \sqrt{((q_1 - p_1)^2 + \cdots + (qn - pn)^2)} \tag{7}$$

where:
D (p, q) es la distancia euclidiana entre los puntos *p* y *q*.

p₁, p₂, ..., pn are the coordinates of point *p* in *n*-dimensional space.
q₁, q₂, ..., qn are the coordinates of point *q* in *n*-dimensional space.

– Concepts about centroids

The centroids are representative points of certain groups or clustering represented within an algorithm such as K-means, where the formula for this concept is the following:

– Centroid of a cluster (K-means)

As indicated in the formula, the centroid represents a point in that cluster that represents all instances of that cluster as indicated in (8).

$$C_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \tag{8}$$

Where:
C_j is the cluster centroid *j*.
n_j is the number of instances in the cluster *j*.
n_j are the coordinates of the instance i in the cluster *j*.

– Centroid update

To perform the centroid update, the centroids are updated in each iteration as shown in (9).

$$C_j^{(t+1)} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \tag{9}$$

Where:
$C_j^{(t+1)}$ is the centroid number of the cluster *j* in the iteration $t + 1$.
n_j is an instance number in the cluster J.
x_i are the coordinates of the instance i in the cluster j.

– Euclidean distance between a point and a centroid

The application of the Euclidean distance is used to terminate the closeness of a point to a centroid as shown in (10).

$$d(x, C_j) = \sqrt{\sum_{i=1}^{n}(x_i - C_{j_i})^2} \tag{10}$$

Where:
$d(x, C_j)$ is the distance between the point x and the centroid C_j.
n is the number of dimensions (features) in the data.
x_i represents the coordinates of x.
C_{j_i} are the coordinates of the centroid C_j.

– Application of concepts in the K-means algorithm

The representation of the objects is called real vectors of d dimension $(x_1, x_2, \ldots, x_n)$. The K-means algorithm provides k groups where the sum of distances of the objects within each group $S = \{s_1, s_2 \ldots, s_n\}$ to its centroid is minimized which is mentioned in (11) and shown below:

$$\frac{min}{s} E\ (\mu_\iota) = \frac{min}{s} \sum_{i=1}^{n} \sum_{x_j\, \epsilon S_i} ||x_j - \mu_i\, ||^2 \tag{11}$$

S belongs to a data set which are elements x_j objects represented by vectors. Each element represents a certain characteristic or attribute. K groups represent the clusters with their centroid μ_ι as seen in (12).

$$\frac{\partial E}{\partial \mu_i} = 0 => \mu_i^{(t+1)} = \frac{1}{S_i^{(t)}} \sum_{x_j\, \epsilon S_i^{(t)}} x_j \tag{12}$$

This space visualizes the application of the K-means algorithm for different types of clustering, where the necessary operators are placed to perform clustering of diseases related to cardiac diseases, as mentioned in Figure 7. The use of the K-means algorithm is fundamental in data analysis and segmentation of unlabeled data into meaningful groups. By carefully selecting relevant attributes and adjusting the algorithm parameters, effective clustering of heart disease-related conditions can be achieved, providing valuable information for medical research and clinical decision making. The layout of the operators in RapidMiner studio reflects the workflow design process to perform this specific task, allowing for efficient implementation and interactive exploration of the clustering results.

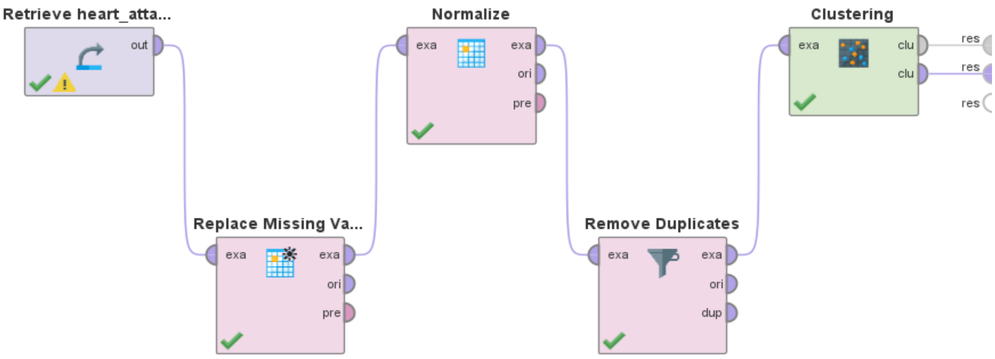

Figure 7. Data mining stage

4. RESULT

4.1. Evaluation of result

Figure 8 shows a graph showing the results of the model established according to the objectives set. In this sense, the is two axes. For this purpose, the Y-axis represents the measurement of the percentage of each cluster. On the other hand, the X represents the classification of the data by clusters that were grouped into 4 clusters, taking into account the criteria of the Rapid studio tool.

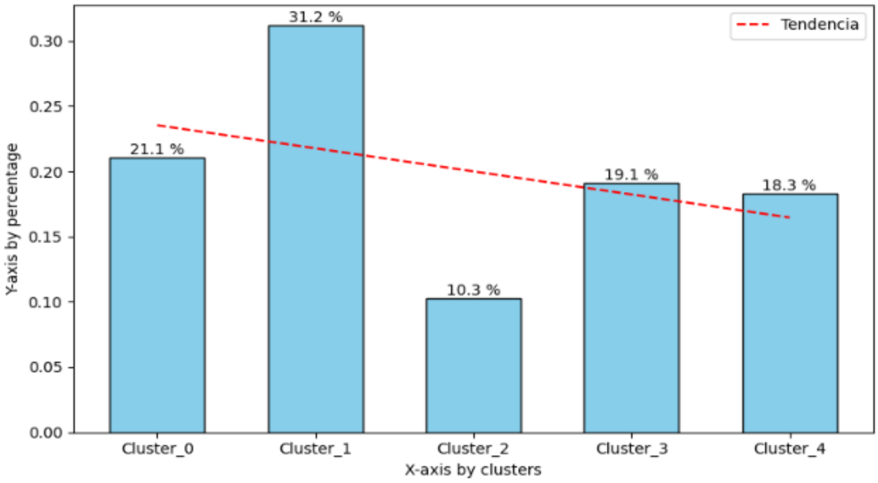

Figure 8. Statistical results on the model

These values show how many records have been divided into each cluster. The result of the clustering reveals that, according to their classification into clusters, the intended process has worked as shown in Table 1. Table 2 presents the clustering results according to the selected attributes. The central position of a set of points in a space, which is used to represent the distance between groups of data within a Cartesian plane. Finally, a grouping of labeled elements in each cluster, numbered from 0 to 4, was achieved, which established certain quantities of elements in each cluster and the total number of elements established in the database was obtained, as shown in Table 3.

Table 1. Results by clusters

	Classification of clusters		
Index	Nominal Value	Fraction	Percentage
1	Cluster_0	0.211	21,1 %
2	Cluster_1	0.312	31,2 %
3	Cluster_2	0.103	10,3 %
4	Cluster_3	0.190	19,1 %
5	Cluster_4	0.183	18,3 %

Table 2. Classification according to their centroids

Attributes	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Patient ID	4379.603	4440	4272.175	4407.016	4321.921
Age	0.121	0.156	1.164	0.122	0.124
Cholesterol	0.022	0.025	0.048	0.010	0.031
Sex	1.218	1.225	2	1.225	1.222
Diabetes	0.730	1.370	0.002	0.730	0.730
Family history	0.006	0.022	0.035	0.986	1.014
Smoking	0.339	0.339	2.948	0.339	0.339
Obesity	0.000	0.025	0.012	0.035	0.012
Alcohol consumption	0.004	0.002	0.038	0.022	0.044
Exercise hours per week	0.029	0.016	0.000	0.023	0.036
Diet	2.012	2.003	2.022	1.986	2.014
Strees level	0.013	0.011	0.005	0.034	0.035
Triglycerides	0.022	0.012	0.014	0.024	0.029

Table 3. Item classification by clustering

Cluster grouping	Number of items
Cluster 0	1852 items
Cluster 1	2732 items
Cluster 2	904 items
Cluster 3	1669 items
Cluster 4	1606 items
Total number of items	8763

The number of records represented by items makes a total of 8763, which was divided into groups called clusters. Cluster 0 has 1852 items with a fraction of 0.211, which is equivalent to 21.1%. Cluster 1 has 2732 items with a fraction of 0.312, which is equivalent to 31.2%. For cluster 2, there are 904 items with a fraction of 0.103, which is equivalent to 10.3%. Cluster 3 has 1669 items with a fraction of 0.190, equivalent to a percentage of 19.1%. Finally, cluster 4 has 1606 items with a fraction of 0.183, equivalent to a percentage of 18.3%.

The attribute of diabetes in clusters 0 and 4 outperformed all other risk factors mentioned in the database; this indicates that, according to the research conducted, the attribute of diabetes is the most prevalent risk factor for contracting a heart attack. Obesity, on the other hand, is a less likely risk factor for developing a heart attack, as it is the result of obesity in clusters 3 and 4. Figure 9 shows a representation in RapidMiner studio of the application of the correlation matrix. In this section, the important attributes were selected to create a heat map with these concepts. In that sense, the application of this theory means implementing a heat map that allows for the comparison of the applied variables.

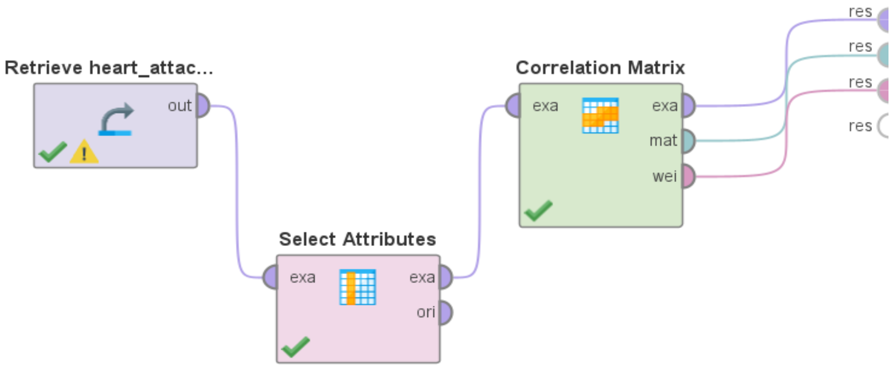

Figure 9. Correlation model

Table 4 shows the correlation matrix in which the values of each variable selected in the prediction are calculated. This matrix provides crucial information on the relationship between the variables and helps to identify possible patterns or dependencies between them. On the other hand, Figure 10 shows the corresponding heat map, which graphically visualizes the information contained in the correlation matrix. This heat map is generated according to the processes established in previous stages with the RapidMiner studio tool. By using colors to represent different levels of correlation, the heat map facilitates the identification of strong or weak relationships between variables, which is fundamental to understand the structure of the data and guide analysis and decision making in the predictive modeling process.

Table 4. Correlation matrix

Attributes	Heart rate	Previous heart problems	Medication use	Sedentary hours per day	Income	BMI	Physical activity days per week	Sleep hours per day
Heart rate	1	-0.005	0.009	-0.010	0.005	0.005	0.001	0.002
Previous heart problems	-0.005	1	0.005	-0.003	-0.003	0.016	0.009	0.004
Medication use	0.009	0.005	1	0.023	-0.003	0.010	-0.011	-0.020
Sedentary hours per day	-0.010	-0.003	0.023	1	0.004	-0.000	-0.006	0.005
Income	0.005	-0.003	-0.003	0.004	1	0.009	0.000	-0.007
BMI	0.005	0.016	0.010	-0.000	0.009	1	0.008	-0.010
Physical activity Days per week	0.001	0.009	-0.011	-0.006	0.000	0.008	1	0.014
Sleep hours per day	0.002	0.004	-0.020	0.005	-0.007	-0.010	0.014	1

4.2. Model comparison

This section shows a comparison of algorithms used in data mining to evaluate their effectiveness in approaching data mining solutions. To achieve this, it was decided to use a Cartesian plane, which is considered to be the most distinctive from one to the other. Figure 10 shows a comparison of models such as Naive Bayes (NB), decision tree (DT), and rule induction. With 1.0, rule induction is the most notable algorithm, while the others have lower results, which helps us to understand some of the algorithms that can be built in the tool.

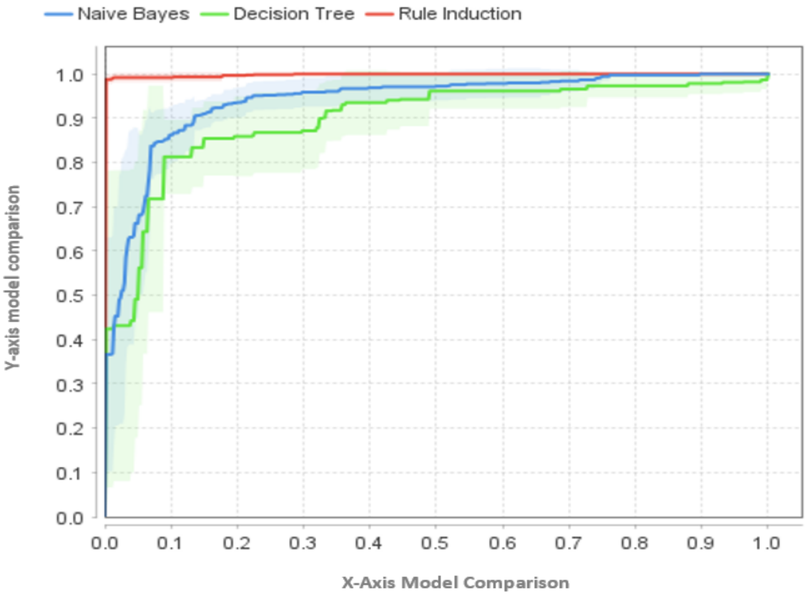

Figure 10. Model comparison

4.3. Comparison of methodologies

The KDD methodology was chosen by SEMMA and CRISP-DM because it was more suitable for the suggested data mining project. The KDD methodology stands out for its comprehensive approach, covering all phases of the data mining process, from data selection and preparation to model evaluation. In addition, KDD has proven to be extremely effective in identifying valuable patterns and insights in large data sets, which was critical to our project. KDD outperformed SEMMA and CRISP-DM in versatility and ability to address the specific challenges of our data mining project more effectively and completely, as visualized in Table 5.

Table 5. Comparison of methodologies

Comparison of attributes	Methodology KDD	Methodology CRIPS-DM	Methodology SEMMA
Structure and sequence	Other methodologies include data selection, cleaning, transformation, and extraction, as well as evaluation and application of knowledge, but their structure is not as rigorous [32].	This methodology consists of six steps: understanding the business, understanding the data, preparing the data, modeling, evaluation, and implementation [33].	The five steps of Semma's methodology are sampling, exploration, modification, modeling, and evaluation [34].
Entrepreneurial orientationl	Recognizes the importance of the company's business objectives and seeks to learn to gain a competitive advantage.	Understands business objectives from the outset and ensures that the results are actionable and valuable for decision-making.	Performs information analysis considering the company's objectives and how the results are used.
Flexibility	Through a broader and less structured approach, it provides a general framework for knowledge discovery.	It is adaptable to a variety of contexts and projects and can be expanded for commercial use.	Although it follows a predetermined sequence of steps, it is adaptable to various projects.
Interaction	It requires repetition. However, it lacks an obvious structure, as does the Semma or Crisp-DM methodology.	Learn to use iterative results review method. Adapts to constantly evolving projects.	This is a procedure that can be carried out in stages as required. Adjustments can be made throughout the process.

5. DISCUSSION

The KDD methodology was chosen over SEMMA and CRISP-DM because it was more suitable for the suggested data mining project. The KDD methodology stands out for its comprehensive approach, covering all phases of the data mining process, from data selection and preparation to model evaluation. In addition, KDD has proven to be very effective in identifying valuable patterns and insights in large data sets, which was crucial for our project. KDD outperformed SEMMA and CRISP-DM in versatility and ability to address the specific challenges of our data mining project more effectively and comprehensively.

According to Broek *et al.* [12], one of the causes of heart problems is cigarette dependence, which shows worrisome results in his research. The findings of the investigated studies showed dimensions related to risk factors for heart attacks. On the other hand, breast cancer is another risk factor related to heart problems; however, this study has people of average age older than 55 years of age [15], in contrast to the other research in which the age of people who were dependent on cigarette smoking was not mentioned. However, in research different from the previous ones, the following author seeks to establish that living with a chronic valve score (CVH) is related to a lower risk of hypertension, chronic kidney disease, cardiovascular events, and their subtypes. In addition, these cardiac problems may be caused by factors such as smoking and dietary habits that cause long-term cardiac problems. The study on cardiovascular cohorts also used this technique, which included 9,388 people aged 8 to 55 who received three examinations [18]. In other cases, studies on acute myocardial infarction have been subjected to analysis of MACE by using predictive models using decision curves and calibration of these techniques. The results of this investigation emphasize the efficacy and clinical relevance of the model [13]. The type or tools that were used to make that prediction are not specified. However, other studies focused on a model based on ML using the RF algorithm to predict cardiovascular health risk in people with coronary heart disease [17]. This study details the data mining-based model used and the research findings. It is important to note that the proposed research does not contemplate the implementation of specific pharmacological treatments that may decrease the effects of cardiac risk diseases. This research established nattocin as a promising drug for heart health but found that during its administration it had a negative effect on blood cholesterol but was effective in treating high blood pressure [16]. The following study found that monitoring reduces the likelihood of stroke or heart attack, indicating that ignoring such problems leads to increased morbidity and mortality [14].

The comparison with the model established by our research has shown promising results in detecting the most common causes of the risk of heart attacks that most people present. In that sense, previous paragraphs have not considered technological models such as data mining to detect certain important patterns that aim to obtain positive results to prevent the triggering causes on heart problems, there is only one research that relates to data mining establishing prediction algorithms comparing it with our model was established RF algorithm while our research uses the K-means algorithm to group records called items.

6. CONCLUSION

Society is concerned about the risks of heart attacks. Many cases have been linked to certain factors, such as a healthy lifestyle and a poor diet, causing diseases that increase a person's likelihood of developing heart disease. Since many people are unaware of some symptoms that occur throughout a person's life, this problem of morbidity and mortality in patients is very relative. With this concept, I analyze some causes of heart disease. However, the KDD technique was used, which allows establishing stages of model development by schematizing the group model using the K-means algorithm. To achieve the objective, it was concluded that one of the underlying causes of the risk of infarction was people who contract diabetes, where certain factors such as poor diet or associated hereditary factors of the person develop. The need to use the established KDD methodology allowed to outline the realization of the predictive model based on clustering, where it is concluded that this methodology adjusts to the amount of data handled at each stage, finding important patterns for the discovery of relevant information. It was calculated that the application of the K-means algorithm maintained a prediction related to diabetes, using this information to emphasize treatments focused on diabetes and establish patterns for taking care of certain causes, such as healthy eating and factors associated with combating a sedentary lifestyle, among others, to prevent cardiac risks. In this sense, no limitations were found for the research since the model meets the expectations of the objectives set. Finally, data mining is a very powerful discipline in the field of artificial intelligence, and our research complements perfectly with other disciplines such as programming, such as the creation of software that allows medical predictions, which would help to make treatments and prevent heart disease. In conjunction with big data and the Internet of Things, significant added value is added that facilitates information management.

REFERENCES

[1] S. Dan, M. Pant, and S. K. Upadhyay, "The case fatality rate in COVID-19 patients with cardiovascular disease: global health challenge and paradigm in the current pandemic," *Current Pharmacology Reports*, vol. 6, no. 6, pp. 315–324, Dec. 2020, doi: 10.1007/s40495-020-00239-0.

[2] M. Escofet Peris *et al.*, "Long-term morbidity and mortality after first and recurrent cardiovascular events in the ARTPER cohort," *Journal of Clinical Medicine*, vol. 9, no. 12, Dec. 2020, doi: 10.3390/jcm9124064.

[3] S. Y. Jung *et al.*, "Cardiovascular events and safety outcomes associated with remdesivir using a World Health Organization international pharmacovigilance database," *Clinical and Translational Science*, vol. 15, no. 2, pp. 501–513, Feb. 2022, doi: 10.1111/cts.13168.

[4] T. J. Van Trier *et al.*, "Unexploited potential of risk factor treatment in patients with atherosclerotic cardiovascular disease," *(in Spanish) European Journal of Preventive Cardiology*, vol. 30, no. 7, pp. 601–610, May 2023, doi: 10.1093/eurjpc/zwad038.

[5] A. Al-Mrabeh, "β-Cell dysfunction, hepatic lipid metabolism, and cardiovascular health in type 2 diabetes: new directions of research and novel therapeutic strategies," *Biomedicines*, vol. 9, no. 2, Feb. 2021, doi: 10.3390/biomedicines9020226.

[6] P. Zhong *et al.*, "Metabolomic phenotyping of obesity for profiling cardiovascular and ocular diseases," *Journal of Translational Medicine*, vol. 21, no. 1, Jun. 2023, doi: 10.1186/s12967-023-04244-x.

[7] T. Medling *et al.*, "Relation of patient's opinion of alcohol's health effects and drinking habits among hospitalized patients with cardiovascular disease," *The American Journal of Cardiology*, vol. 179, pp. 31–38, Sep. 2022, doi: 10.1016/j.amjcard.2022.06.033.

[8] P. Ananda Selva Das, M. Dubey, R. Kaur, H. R. Salve, C. Varghese, and B. Nongkynrih, "WHO non-lab-based CVD risk assessment: A reliable measure in a North Indian population," *Global Heart*, vol. 17, no. 1, Sep. 2022, doi: 10.5334/gh.1148.

[9] C. Ke *et al.*, "Association of prior outpatient diabetes screening with cardiovascular events and mortality among people with incident diabetes: a population-based cohort study," *Cardiovascular Diabetology*, vol. 22, no. 1, Aug. 2023, doi: 10.1186/s12933-023-01952-y.

[10] L. Huang, J. Zhang, Q. Huang, R. Cui, and J. Chen, "In-hospital major adverse cardiovascular events after primary percutaneous coronary intervention in patients with acute ST-segment elevation myocardial infarction: a retrospective study under the China chest pain center (standard center) treatment system," *BMC Cardiovascular Disorders*, vol. 23, no. 1, Apr. 2023, doi: 10.1186/s12872-023-03214-x.

[11] S.-M. Jeong *et al.*, "Smoking behavior change and risk of cardiovascular disease incidence and mortality in patients with type 2 diabetes mellitus," *Cardiovascular Diabetology*, vol. 22, no. 1, Jul. 2023, doi: 10.1186/s12933-023-01930-4.

[12] L. G. uit het Broek, B. B. A. Ort, H. Vermeulen, T. Pelgrim, L. C. M. Vloet, and S. A. A. Berben, "Risk stratification tools for patients with syncope in emergency medical services and emergency departments: a scoping review," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 31, no. 1, Sep. 2023, doi: 10.1186/s13049-023-01102-z.

[13] S. Kuila, N. Dhanda, and S. Joardar, "ECG signal classification to detect heart arrhythmia using ELM and CNN," *Multimedia Tools and Applications*, vol. 82, no. 19, pp. 29857–29881, Aug. 2023, doi: 10.1007/s11042-022-14233-9.

[14] A. E. S. Ahmed, Q. Abbas, Y. Daadaa, I. Qureshi, G. Perumal, and M. E. A. Ibrahim, "A residual-dense-based convolutional neural network architecture for recognition of cardiac health based on ECG signals," *Sensors*, vol. 23, no. 16, Aug. 2023, doi: 10.3390/s23167204.

[15] W. R. Naaktgeboren *et al.*, "Physical activity and cardiac function in long-term breast cancer survivors: a cross-sectional study," *JACC: CardioOncology*, vol. 4, no. 2, pp. 183–191, Jun. 2022, doi: 10.1016/j.jaccao.2022.02.007.

[16] X. Li *et al.*, "Nattokinase supplementation and cardiovascular risk factors: A systematic review and meta-analysis of randomized controlled trials," *Reviews in Cardiovascular Medicine*, vol. 24, no. 8, Aug. 2023, doi: 10.31083/j.rcm2408234.

[17] M. V. Dogan, S. R. H. Beach, R. L. Simons, A. Lendasse, B. Penaluna, and R. A. Philibert, "Blood-based biomarkers for predicting the risk for five-year incident coronary heart disease in the framingham heart study via machine learning," *Genes*, vol. 9, no. 12, Dec. 2018, doi: 10.3390/genes9120641.

[18] N. B. Allen *et al.*, "Cardiovascular health trajectories from childhood through middle age and their association with subclinical atherosclerosis," *JAMA Cardiology*, vol. 5, no. 5, May 2020, doi: 10.1001/jamacardio.2020.0140.

[19] L. Corlin, M. I. Short, R. S. Vasan, and V. Xanthakis, "Association of the duration of ideal cardiovascular health through adulthood with cardiometabolic outcomes and mortality in the framingham offspring study," *JAMA Cardiology*, vol. 5, no. 5, May 2020, doi: 10.1001/jamacardio.2020.0109.

[20] A. Dekhtyar and J. H. Hayes, "Automating requirements traceability: two decades of learning from KDD," in *2018 1st International Workshop on Learning from other Disciplines for Requirements Engineering (D4RE)*, Aug. 2018, pp. 12–15. doi: 10.1109/D4RE.2018.00009.

[21] N. Akhtar, M. Ramzan, and N. Kanwal, "Data mining techniques to construct a model: Cardiac diseases," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, 2018, doi: 10.14569/IJACSA.2018.090173.

[22] L. Al-Alawi, J. Al Shaqsi, A. Tarhini, and A. S. Al-Busaidi, "Using machine learning to predict factors affecting academic performance: The case of college students on academic probation," *Education and Information Technologies*, vol. 28, no. 10, pp. 12407–12432, Oct. 2023, doi: 10.1007/s10639-023-11700-0.

[23] S. D'Oca, S. Corgnati, and T. Hong, "Data mining of occupant behavior in office buildings," *Energy Procedia*, vol. 78, pp. 585–590, Nov. 2015, doi: 10.1016/j.egypro.2015.11.022.

[24] E. N. Houstis, A. C. Catlin, J. R. Rice, V. S. Verykios, N. Ramakrishnan, and C. E. Houstis, "PYTHIA-II: a knowledge/database system for managing performance data and recommending scientific software," *ACM Transactions on Mathematical Software*, vol. 26, no. 2, pp. 227–253, Jun. 2000, doi: 10.1145/353474.353475.

[25] E. L. Cahapin, B. A. Malabag, C. S. Santiago Jr., J. L. Reyes, G. S. Legaspi, and K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, Dec. 2023, doi: 10.11591/eei.v12i6.4849.

[26] Y. Guo, Z. Zhang, Y. Liu, Y. Wang, and P. Xue, "An advanced ensemble clustering approach for data partitioning and mining to optimize performance in variable refrigerant flow systems," *Journal of Building Engineering*, vol. 78, Nov. 2023, doi: 10.1016/j.jobe.2023.107716.

[27] M. Wang *et al.*, "Discovering causes of traffic congestion via deep transfer clustering," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 5, pp. 1–24, Oct. 2023, doi: 10.1145/3604810.

[28] M. A. Ahmed, H. Baharin, and P. N. Nohuddin, "Text clustering of tafseer translations by using k-means algorithm: an Al-Baqarah chapter view," *Annals of Emerging Technologies in Computing*, vol. 7, no. 4, pp. 27–34, Oct. 2023, doi: 10.33166/AETiC.2023.04.003.

[29] M. Loog, J. H. Krijthe, and M. Bicego, "Also for k-means: more data does not imply better performance," *Machine Learning*, vol. 112, no. 8, pp. 3033–3050, Aug. 2023, doi: 10.1007/s10994-023-06361-6.

[30] E. P. W. Mandala and D. E. Putri, "Data mining technique for grouping products using clustering based on association," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 2, pp. 835–844, Aug. 2023, doi: 10.11591/ijeecs.v31.i2.pp835-844.

[31] S. Spiers, H. T. Bui, and R. Loxton, "An exact cutting plane method for the Euclidean max-sum diversity problem," *European Journal of Operational Research*, vol. 311, no. 2, pp. 444–454, Dec. 2023, doi: 10.1016/j.ejor.2023.05.014.

[32] A. H. Azizan *et al.*, "A machine learning approach for improving the performance of network intrusion detection systems," *Annals of Emerging Technologies in Computing*, vol. 5, no. 5, pp. 201–208, Mar. 2021, doi: 10.33166/AETiC.2021.05.025.

[33] J. Bokrantz, M. Subramaniyan, and A. Skoogh, "Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM," *Production Planning & Control*, vol. 35, no. 16, pp. 2234–2254, Dec. 2024, doi: 10.1080/09537287.2023.2234882.

[34] S. López-Torres *et al.*, "IoT monitoring of water consumption for irrigation systems using SEMMA methodology," in *Intelligent Human Computer Interaction*, 2020, pp. 222–234. doi: 10.1007/978-3-030-44689-5_20.

BIOGRAPHIES OF AUTHORS

Inoc Rubio Paucar bachelor in Systems and Computer Engineering. He has a background in database management and computer system design, with a focus on artificial intelligence applications, machine learning, and data science. His research interests are in the area of computer science. He can be contacted at email: Enoc.Rubio06@hotmail.com.

Laberiano Andrade-Arenas doctor in Systems and Computer Engineering. Master in Systems Engineering. Graduated with a Master's Degree in University Teaching. Graduated with a Master's degree in accreditation and evaluation of educational quality. Systems Engineer. scrum fundamentals certified, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.