Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.7485
This work is licensed under a Creative Commons Attribution 4.0 International License.

Pipeline for Discovering the Comorbidity Profile of Hypertensive Patients with Cardiovascular Diseases in Brazil

Beethoven M. Andrade [**Pontifical Catholic University of Minas Gerais (PUC-Minas)** | ***beethoven.m.a@gmail.com***]
Camila Grandini [**Pontifical Catholic University of Minas Gerais (PUC-Minas)** | ***camila.grandini.pro@gmail.com***]
Caio Davi Rabelo Fiorini [**Pontifical Catholic University of Minas Gerais (PUC-Minas)** | ***caiodavi442@gmail.com***]
Gustavo Costa Pontifical Catholic University of Minas Gerais (PUC-Minas) | ***gustavocosta.ds09@gmail.com***]
Luis Enrique Zárate [**Pontifical Catholic University of Minas Gerais (PUC-Minas)** | ***zarate@pucminas.br***]

Data Science and Computer Science Department, Applied Computational Intelligence Laboratory - LICAP, Pontifical Catholic University of Minas Gerais (PUC-Minas), Belo Horizonte, MG, 30140-100, Brazil.

Received: 27 January 2026 • **Published:** 27 November 2026

Abstract This work applies a knowledge discovery process to characterize the profiles of healthy individuals and those affected by hypertension and cardiovascular diseases as comorbidities. For this purpose, the database of the National Health Survey (PNS)/IBGE of 2019 is used. The pipeline considers two stages of feature selection, one based on understanding the problem and the other based on selection by a parallel genetic algorithm. The models were built using Decision Tree, Random Forest, and Bayesian classifier. All models showed satisfactory and similar performance with an average of 85%. The results highlight that household conditions, health care, self-perception aspects, presence/absence of diabetes, and the demographic profile are relevant characteristics to describe the profiles

Keywords: Knowledge discovery, Health informatics, Public health, Feature Selection, Genetic Algorithm

1 Introduction

Nowadays, Cardiovascular Diseases (CVDs) are the leading cause of death around the world. During 2008, it was estimated that these diseases caused 17.3 million deaths, 7.3 million due to heart attacks and 6.2 million due to strokes. The World Health Organization (WHO) projects that by 2030, more than 23 million people will die from these diseases that affect the cardiovascular system [WHO, 2011].

Arterial Hypertension (AH), broadly known as high blood pressure, is also a chronic disease, affecting the individual's cardiovascular system, and is currently one of the main public health issues in the world. WHO estimates that around 600 million people suffer from AH and that there will be a global increase of 60% in cases by 2025, in addition to the number of 7.1 million deaths per year Alwan [2011]. In Brazil, data from the National Health Survey (PNS) of 2019 reveal that 23.9% of adults reported a positive diagnosis for AH [Malta *et al*., 2022; Sousa and Zarate, 2024], indicating that there is a need to prevent the progression of this condition and its complications, since it is one of the main risk factors for cardiovascular diseases.

The relevance of studying the AH comorbidity, along with cardiovascular diseases, comes not only from the high prevalence of these conditions in the population but also from their significant impact on the quality of life, mortality rates, and financial costs related to their management. Arterial hypertension is a central risk factor for cardiovascular diseases, which impose an economic burden on healthcare systems due to hospitalizations, long-term treatments, and preventable complications. In Brazil, cardiovascular diseases, often associated with hypertension, generated an estimated cost of approximately R$50 billion between 2010 and 2020, considering direct expenditures on hospitalizations and treatments by the Unified Health System (SUS). Furthermore, indirect costs, such as loss of productivity and premature mortality, are estimated to further increase this economic burden in the country [Stevens *et al*., 2018; de Araújo *et al*., 2022].

The importance of studying this topic lies in the significant impact of hypertension and cardiovascular diseases on both public health and economic costs in Brazil. Preventing and diagnosing these conditions at an early stage is essential to reduce mortality and associated complications through public investments, the adoption of healthier lifestyles, and the use of medications.

Recent studies have explored the use of data mining and machine learning algorithms to improve the diagnosis and prediction of these diseases. For example, the scientific article AlKaabi *et al*. [2020] uses data from 987 individuals to apply techniques such as logistic regression and decision trees to predict hypertension based on non-invasive data, demonstrating the potential to reduce costs and optimize risk screening in vulnerable populations.

The work by Bhatt *et al*. [2023] uses a real-world dataset containing 70,000 instances to develop predictive models that classify the occurrence of cardiovascular diseases. The researchers applied algorithms such as Random Forest, Decision Tree, XGBoost, and Multilayer Perceptron (MLP). Risk factors such as poor diet, obesity, and smoking were identified as relevant variables for prediction, highlighting how data-driven approaches can support screening systems.

This article aims to describe the profile of Brazilian individuals who present hypertension and cardiovascular diseases, as well as to identify relevant variables associated with

this comorbidity. The study applies a knowledge-discovery process with two approaches for feature selection, applied in two different moments of the pipeline. Both of them have the objective of identifying the main factors that characterize this comorbidity in the Brazilian population and reducing dataset dimensionality for the construction of learning models.

The first approach, named conceptual selection, is based on understanding the problem's domain to obtain models that better represent the problem at hand. This process is applied as the first step on the knowledge discovery pipeline and aims to initially reduce the dimensionality of the original dataset. The second one corresponds to a typical feature selection process, based on an evolutionary algorithm, carried out using the Non-dominated Sorting Genetic Algorithm II (NSGA-II), which performs a systematic exploration of the search space using principles of the Darwinian genetic evolution to identify attribute subsets that most balance predictive performance and representativeness of the learning models. This strategy enables a broader search in the solution space, increasing the potential discovery of relevant patterns.

The learning models are built using three distinct approaches: Decision Tree, Random Forest, and Naive Bayes, considering two populations: (a) healthy individuals and (b) individuals with comorbidity. To characterize the Brazilian population, the most recent National Health Survey (PNS) from IBGE was used [IBGE, 2020]. It gathers demographic information, with clinical, behavioral and health condition data from the year 2019. It is expected that the obtained results reveal consistent patterns and contribute to the contextualized understanding of this comorbidity in Brazil, supporting initiatives oriented to prevention, risk stratification, and improvement of public policies focused on chronic cardiovascular diseases.

This article is divided into five sections. In Section 2, works related to the problem domain addressed are presented. In Section 3, the methodology, involving the subsections of materials and methods adopted in the pipeline, is described. In Section 4, the experiments and analysis of results are explored, and finally, the conclusions and future work are presented. This work is an extended version of the paper Costa and Zárate [2025] presented in SBBD 2025, Fortaleza, CE, Brazil.

Compared to the SBBD 2025 version, this extended manuscript introduces: (i) a more detailed knowledge discovery pipeline, including an additional feature-selection step using a parallel implementation of NSGA-II; (ii) a more comprehensive experimental evaluation with parameter tuning and statistical validation; and (iii) an extended analysis of the results, including rule interpretation and feature importance. The main contribution of this work is the integration of a dual-stage feature selection strategy, combining domain-driven conceptual selection with parallel NSGA-II, which enables an effective reduction of dimensionality while preserving predictive performance, supported by a distributed master–slave architecture that improves computational efficiency.

2 Related Work

The application of machine learning (ML) techniques to the prediction of arterial hypertension and cardiovascular diseases has grown in recent years. Studies such as AlKaabi *et al*. [2020] have used models like logistic regression to identify risk factors for AH, while others, such as Bhatt *et al*. [2023] and Gárate-Escamila *et al*. [2020], have focused on CVDs, achieving high accuracy values with algorithms such as XGBoost and Random Forest, respectively.

In the Brazilian scenario, which makes use of data from PNS survey, the approach is equally relevant. For example, de Carvalho *et al*. [2024] also used PNS 2019 to diagnose hypertension, obtaining an F1-Score of 75% with Random Forest, highlighting similar risk factors identified in this study. Similarly, Gonçalves and Gálvez [2024] used the same dataset (PNS 2019) to characterize the profile of individuals with stroke, reinforcing the potential of this dataset for public health studies in Brazil.

Gonçalves *et al*. [2025] analyzed the 2019 PNS with a focus on identifying risk profiles for stroke in middle-aged adults. The study employed a feature selection process that resulted in the reduction of three variables, enabling the construction of a Decision Tree–based model with an accuracy of 65%. These findings reinforce the role of attribute selection as a tool to refine predictive models and facilitate the interpretation of complex datasets.

Different from these studies, which focus on AH or CVDs (such as strokes) in an isolated manner, the present study advances the field by investigating the comorbidity of hypertension and cardiovascular diseases simultaneously. The objective is to not only predict, but also characterize these individuals' profiles, filling a gap in the understanding of factors associated with this high-risk group in the Brazilian population based on PNS data.

3 Methodology

The proposed study followed a series of steps, depicted in the flowchart contained in Figure 1.

In addition to following the knowledge discovery process, the proposed pipeline introduces methodological contributions. Firstly, it incorporates a dual-stage feature selection strategy, combining an initial conceptual selection based on domain knowledge with a subsequent optimization stage using a parallel implementation of the NSGA-II algorithm. This approach enables a more effective reduction of the feature space while preserving interpretability and predictive performance.

Finally, the adoption of a distributed master–slave architecture for the execution of the genetic algorithm significantly improves computational efficiency, making it feasible to explore large search spaces within a reasonable time. These combined aspects differentiate the proposed pipeline from traditional approaches that rely solely on sequential feature selection or purely data-driven preprocessing.

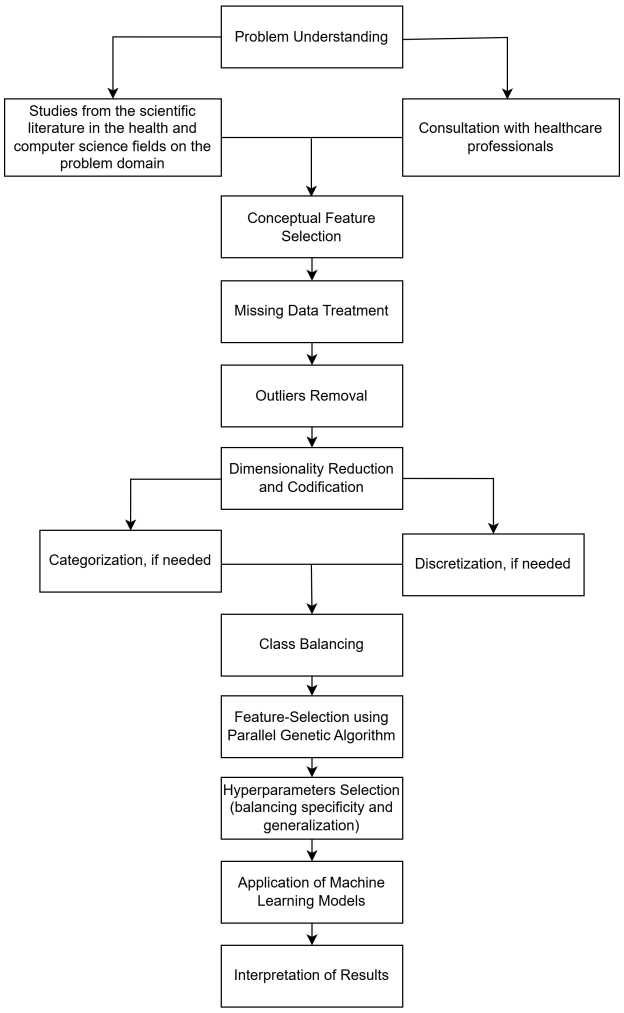

Figure 1. Proposed Knowledge Discovery Pipeline

3.1 Materials

The National Health Survey (PNS) 2019, used as the data source in this study, was conducted by IBGE, in partnership with the Ministry of Health. The PNS is considered an important data source on the Brazilian population's health conditions. PNS collects data from chronic diseases, health and dietary habits, living conditions, and access to healthcare, among other aspects. PNS possesses a significant sample from all Brazilian states, with sociodemographic information, lifestyle, and risk factors such as smoking and obesity, distributed across 26 question modules. The PNS dataset contains 293,726 records and 1,088 attributes.

3.2 Methods

3.2.1 Problem Understanding

Understanding the problem domain is fundamental to this study, since it means comprehending the context, mapping the most important factors to the issue, and the possible solutions. This process involves a systematic literature review and consultations with domain experts, in this case, healthcare professionals, aiming at a broader understanding of the main factors that influence the problem at hand. For this purpose, the CAPTO Gonçalves *et al*. [2024] method was used, enabling the visualization of the problem and its context from multiple perspectives. In the Figure 2, a CAPTO-based conceptual model built on Hypertension and Chronic Cardiovascular Diseases is presented.

3.2.2 Conceptual Feature Selection

After building the conceptual model using the CAPTO method, the next step consisted of conceptually selecting the identified attributes by the model, based on PNS 2019 data. After this process, some of these were renamed to improve interpretability during the construction of the learning models, as described in Table 1.

Table 1. Conceptual Feature Selection from PNS 2019

Attribute	Description	Renamed Name
P00104	Weight of the individual in Kg.	Weight
P00404	Height of the individual in Cm.	Height
C006	Gender of the individual.	Gender
C008	Age of the individual.	Age
C009	Race/Ethnicity of the individual.	Race_Ethnicity
P02601	Salt consumption of the individual.	Salt_Consumption
P034	Exercised in the last three months?	Physical_Activities
P035	How many days per week did they exercise?	Physical_Activity_Frequency
P050, P052, P05401, P05404, P05407, P05410, P05413, P05416, P05419	Amount smoked per day/week and smoking history.	Tobacco_Usage
P027	Frequency of alcohol consumption?	Alcohol_Frequency
P02801	Consumes alcohol, how many days a week?	Alcohol_Week_Frequency
P029	When you drink, how many doses are consumed?	Alcohol_Doses
J037	Hospitalizations in the last 12 months?	Hospitalized
H001	Last medical appointment?	Last_Consultation
N001	Health self-assessment.	Perception_Health_Status
I00102	Has health-assurance?	Has_Health_Insurance
N004	Feels chest pain when climbing stairs/hills?	Tiring_Uphill
N005	Feels chest pain when walking normally?	Tiring_Plan
D001	Knows how to read and write?	Literacy
VDD004A	Highest level of education achieved.	Education
V0026	Type of census situation.	Housing_Area
A01501	Household sewage system.	Sewage
A016010	Garbage destination.	Garbage_Collection
A01901	Has internet access?	Internet_Access
A005010	Household water supply.	Water_Supply
VDF004	Household income per capita.	Salary_Range
Q03001	Diagnosis of diabetes by a doctor?	Has_Diabetes

The attributes that were not renamed are those that later underwent processes of categorization, discretization, or merging and will be presented throughout the document in subsequent stages, as is the case for attributes related to smoking (P050, P052, P054019, etc).

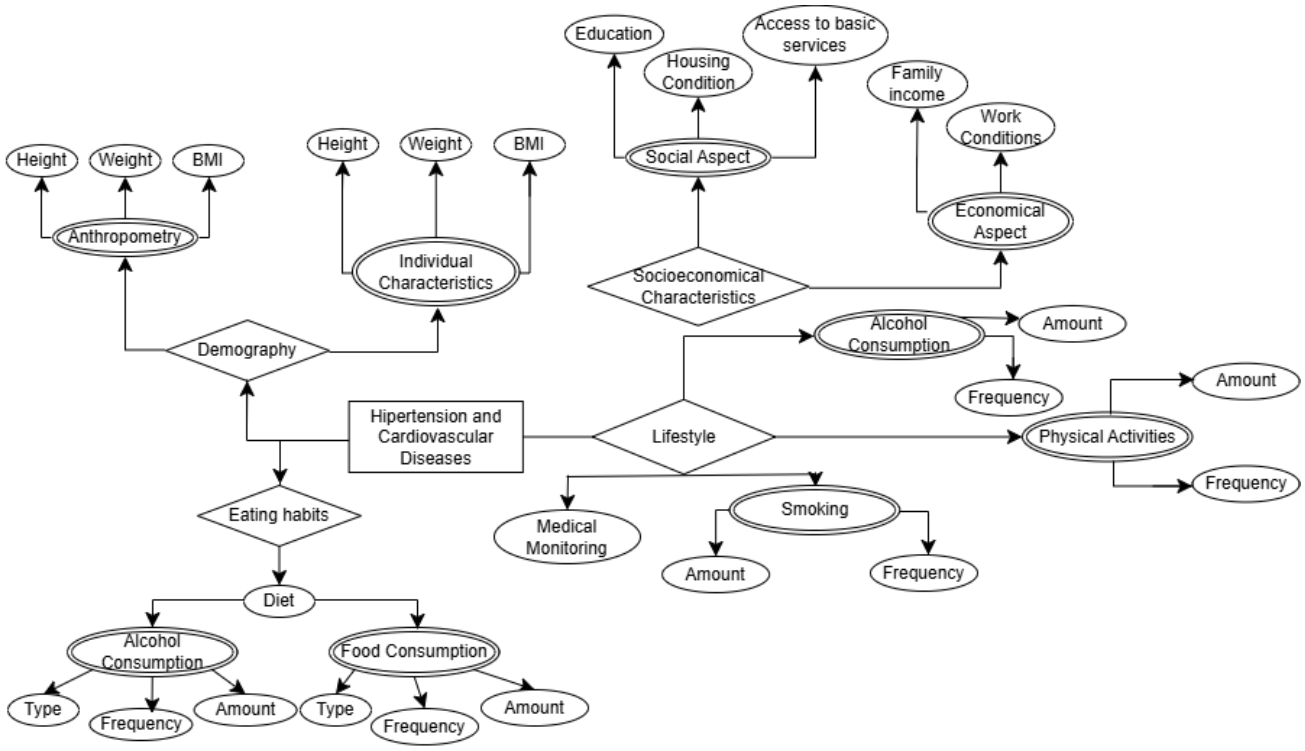

Figure 2. Conceptual Model – Hypertension with Cardiovascular Diseases

With the feature pre-selection completed, a filter was also applied to the ages of the survey participants, including only individuals aged 18 years or older. In the end, the dataset consisted of 45,757 instances and 36 attributes.

3.2.3 Missing Data Treatment

Missing data was handled in a specific manner for each aspect of the conceptual model (Figure 2), except "Eating Habits". First, in the <Anthropometric Aspect>, 575 records (573 healthy individuals and 2 hypertensive individuals with CVD) were removed due to missing values in one of the two attributes, "Weight" and "Height", since the absence of one variable made a reliable estimation of the other impossible. In addition, imputation using techniques such as KNN was considered unfeasible. In the <Physical Activity Aspect> 26,358 null values in "Physical_Activity_Frequency" were replaced with 0, as they corresponded to individuals who had not practiced exercise or sports in the previous three months.

For the <Smoking Aspect>, attributes related to current consumption (ex: P05401) had 39,127 null values imputed with 0, since the individuals associated with these records reported never having smoked. In attribute P052 (smoking history), 5,268 null values were filled with 1 (former daily smokers) and 655 with 2 (former occasional smokers), according to responses linked to other indicators. In the <Alcohol Consumption Aspect>, 25,273 null values in "Alcohol_Week_Frequency" and "Alcohol_Doses" were assigned 0, corresponding to individuals who never consumed alcohol. The remaining 6,136 null values, after class-based analysis (mean close to 0 days per week), were also imputed with 0, considering reported consumption of once per month or less.

In the <Medical Monitoring Aspect>, 537 records with missing values in "Tiring_Uphill" and "Tiring_Plan" as well as 3,746 records in "Has_Diabetes," were removed because they belonged to the majority class (healthy individuals). For the Social Aspect, 560 null values in "Sewage" were filled with 4 (septic tank), based on the socioeconomic profile (60% with income less than or equal to one minimum wage and rural residence). Finally, in the <Economical Aspect>, 13 instances with missing values in "Salary_Range" were excluded.

3.2.4 Outliers Removal

Outliers, which correspond to values that are highly discrepant compared to the distribution of the remaining data, were removed to avoid introducing noise due to the potential distortion they cause.

On the <Individual Characteristics Aspect> (Figure 2), an analysis using the Interquartile Range (IQR) was performed. This method consists of identifying the lower and upper bounds of the data distribution, calculated according to the following expressions: $LowerBound = Q_1 - 1.5 \times IQR$, $UpperBound = Q_3 + 1.5 \times IQR$, where $IQR = Q_3 - Q_1$. By default, this method computes the lower and upper bounds using a multiplier of 1.5 times the IQR. Thus, values above the upper bound or below the lower bound are considered outliers.

Within the <Anthropometric Aspect>, the application of outlier detection for the "Height" attribute, based on limits calculated from the IQR, identified, in general, many potential outliers for both sexes, but none that significantly distorted the overall data distribution. This is because differences in height preserve the variability of the original data. Therefore, a multiplier of 3 times the IQR was used instead of the standard 1.5 threshold. Increasing this threshold makes

the outlier detection process more conservative and less stringent, with an estimated removal of 0.7% of the data relative to the total available data when dealing with a normal distribution [Yang *et al*., 2019].

Regarding the "Weight" attribute, the standard boxplot using the 1.5 threshold identified 4 potential outliers below the lower bound and 847 above the upper bound. However, the values above the upper bound were not discarded, as they correspond to individuals classified as overweight or obese and are relevant to the context of this study. Therefore, only the outliers below the lower bound were removed.

On the matter of alcohol consumption, instances were identified in which individuals consumed more than 15 alcoholic drinks per day. According to NIAAA [2022], one alcoholic drink is generally defined as containing 14 grams of pure alcohol, which corresponds to a standard drink. Consuming 15 drinks, therefore, implies ingesting 210 grams of alcohol, a value way beyond the daily limits recommended for health. Such levels of consumption can quickly result in alcohol intoxication or long-term permanent damage. Consequently, these instances were removed from the dataset as they were considered outliers.

3.2.5 Feature Fusion

This stage aims to reduce the dimensionality of attributes that share the same relevant information on the problem. For example, this approach was applied to the <Smoking Aspect>, where attributes related to the type of tobacco product consumed and the frequency of consumption were grouped into categories that combine these two characteristics, resulting in the following groups: Heavy Smoker, Moderate Smoker, Light Smoker, and Non-Smoker. Thus, nine tobacco-related attributes were merged and categorized into a single attribute called *Tobacco_Usage*, as shown in Table 2.

Table 2. Smoking Categories and Logic Conditions

Smoking Category	Logical Condition
Heavy smoker	If "P050" = 1 (Smokes currently every day).
Moderate smoker	If "P050" = 3 and "P052" = 1 (Does not currently smoke, but has smoked in the past).
Light Smoker	If "P050" = 3 and "P052" = 2 (Does not currently smoke and smokes less than he has in the past).
Light Smoker	If "P050" = 2 and ("P05401" = 4 or "P05404" = 4 or "P05410" = 4) (Smokes currently less than daily and rarely uses tobacco products).
Moderate smoker	If "P050" = 2 and ("P05401" = 2 or "P05404" = 2 or "P05410" = 2) (Smokes currently less than daily and uses tobacco products reasonably).
Heavy smoker	If "P050" = 2 and ("P05401" = 1 or "P05404" = 1 or "P05410" = 1) (Smokes currently less than daily and uses tobacco products often).
Non-Smoker	If "P050" = 3 and "P052" = 3 (Does not currently smoke and has never smoked in the past).
Non-Smoker	If there are ignored values in "P050" or "P052" (Ignored values are treated as non-smokers).

To describe the individual's alcohol consumption characteristics, a fusion encoding was also created with the aim of reducing the dimensionality of alcohol-related variables. Thus, the attributes "Alcohol_Doses," "Alcohol_Frequency" and "Alcohol_Week_Frequency" were merged and categorized into a single attribute, called *Cat_Alcohol* (see Table 3).

Table 3. Alcoholism Categories and Logical Conditions

Alcohol Consumption Category	Logical Condition
Non-alcoholic	Frequency of alcoholism = 1
Social drinker	Frequency of alcoholism = 2
Frequent drinker	Frequency of alcoholism = 3 and Weekly amount of alcohol ≤ 3
Excessive drinker	Frequency of alcoholism = 3 and Weekly amount of alcohol > 4.

The "Race_Ethnicity" attribute was also categorized into three groups: White, Black, and Mixed. Minority race/ethnicity categories, such as Asian, Indigenous, and individuals who did not respond to this question, were included in the White category. This grouping into three major groups was adopted due to the genetic predisposition of Black and Brown individuals to develop arterial hypertension at higher rates compared to White individuals [Zilbermint *et al*., 2019].

Finally, it was necessary to create a new attribute to calculate the Body Mass Index (BMI) of each individual in the dataset, since obesity and overweight are risk factors for the development of arterial hypertension and cardiovascular diseases [Powell-Wiley *et al*., 2021]. After calculating BMI, these values were categorized into four classes: Underweight, Ideal Weight, Overweight, and Obese [WHO, 2021], as shown in Table 4. Table 5 presents the set of attributes (23) resulting from the preliminary data preparation steps, including the newly created attribute *Class = {Healthy, AH_CVD}*.

Table 4. BMI Classification According to WHO Guidelines

Category	BMI Intervals (kg/m²)
Underweight (1)	BMI $<$ 18.5
Ideal Weight (2)	18.5 $\leq$ BMI $<$ 25.0
Overweight (3)	25.0 $\leq$ BMI $<$ 30.0
Obese (4)	BMI $\geq$ 30.0

3.2.6 Feature Selection via Genetic Algorithm

Conceptual attribute selection and dataset preprocessing are essential steps in building representative models, which are particularly relevant when dealing with high dimensionality. Conceptual selection aims to describe the problem domain, consequently reducing the search space and the computational effort typically required by traditional feature selection techniques, especially bio-inspired methods such as genetic algorithms (GAs). Continuing along the pipeline, a new feature selection stage based on a genetic algorithm was applied to further reduce dimensionality by identifying the features that best represent the problem domain.

As it is well known, the sequential execution of genetic algorithms (GAs) for feature selection is limited by high computational costs, mainly due to the exponential search space as a function of the initial number of attributes. Finding satisfactory solutions may require days or even weeks of processing, which makes parallelization an essential strategy to accelerate the process and enable its practical application. From this need arose Parallel Genetic Algorithms (PGAs), which exploit the independence among candidate solutions to distribute computation across multiple nodes. This ap-

Table 5. Set of Features after preliminary data preparation steps

Feature	Renamed name (category)	Description
C006	Gender (PD)	1:Man 2:Woman
C009	Race_Ethnicity (PD)	1:White 2:Black 3:Yellow 4:Brown 5:Indigenous
P02601	Salt_Consumption (HA)	1:Very High 2:High 3:Adequate 4:Low 5:Very Low
P035	Physical_Activity_Frequency (EV)	1 to 7 days 0: Never or less than once a week
J037	Hospitalized(HI)	1:Yes 2:No
Cat_Age	Age_Group (PD)	1:Young adult 2:Adult 3:Elderly 4:Very Elderly 5:Extreme old age
Cat_BMI	Cat_BMI (CS)	1:Underweight 2: Ideal Weight 3: Overweight 4: Obese
	Tobacco_Usage (EV) Cat_Smoking	0: Does not smoke 1: Smokes a little 2: Moderate smoker 3: Heavy smoker
Cat_Alcohol	Cat_Alcohol (EV)	
H001	Last_Medical_Appointment (CS)	1: Up to 15 days 2: More than 15 days up to 1 month 3: More than 1 month up to 6 months 4: More than 6 months up to 1 year 5: More than 1 year ago
N001	Perception_health_status (APS)	1: Very good 2: Good 3: Regular 4:Bad 5:Very bad
I00102	Has_Health_Insurance (CS)	1:Yes 2:No
N004	Gets_tired_climbing_up (APS)	1:Yes 2:No
N005	Gets_tired_walking (APS)	1:Yes 2:No
D001	Literacy (PSD)	1:Yes 2:No
VDD004A	Education (PSD)	1: No formal education 2: Incomplete primary education 3: Complete primary education 4: Incomplete secondary education 5: Complete secondary education 6: Incomplete higher education 7: Complete higher education
V0026	Area_Housing (CD)	1:Urban 2:Rural
A01501	Sewage (CD)	1: Public sewer system 2: Septic tank connected to the sewer system 3: Septic tank not connected to the sewer system 4: Rudimentary septic pit 5: Open ditch 6: River, lake, stream, sea 7: Other
A016010	Garbage_Collection (CD)	1: Public collection 2: Collected in a dumpster 3: Burned 4: Buried 5: Dumped on vacant land 6: Other
A01901	Internet_Access (CD)	1:Yes 2:No
A005010	Water_supply (CD)	1: Public water supply 2: Artesian well 3: Shallow well (water table well) 4: Spring or natural source 5: Stored rainwater 6: Other
VDF004	Salary_Range (PSD)	1: Up to 1/4 of the minimum wage 2: More than 1/4 up to 1/2 of the minimum wage 3: More than 1/2 up to 1 minimum wage 4: More than 1 up to 2 minimum wages 5: More than 2 up to 3 minimum wages 6: More than 3 up to 5 minimum wages 7: More than 5 minimum wages
Q03001	Has_diabetes (CO)	1:Yes 2:No
Class		1:Healthy 2:AH_CVD

PD: Demographic profile HA: Dietary habits EV: Lifestyle CS: Health care APS: Self-perceived health PSD: Socio-demographic profile CD: Household characteristics CO: Comorbidity

Table 6. Dataset Reduction Throughout the Pipeline

Stage	Number of Instances	Number of features
Original dataset (PNS 2019)	293,726	1081
After preliminary processings	45,757	22
After all processings	4,738	16

Table 7. Experimental Parameters for the Genetic Algorithm

Parameter	Configuration / Interval	Details
Dataset	`SBBD_PNS_HiperCardio.csv`	Entry file
Condition	1	Identifier
Backup	`backup_geracao.txt`	Recovery log
Seed	2156, 2, 885463215	Seeds
Population	204 a 504	Step: 200
Generations	200 a 504	Step: 200
CrossOverFactor	0.7 a 0.91	Step: 0.2
TournamentSize	2	Fixed value
MutationRate	0.1 a 0.41	Step: 0.2
ElitismFactor	1	Fixed value
DecisionTree	`max_depth=5`	Hyperparameter

Table 8. Genetic Algorithm Experiments

Exp.	F1	Before	After	Selected Attributes	Gr.
1	0.863	23	10	H001, N001, N004, D001, VDD004A, A01901, Q03001, Cat_Alcohol, Cat_Age	G1
2	0.864	23	8	H001, N001, N004, D001, VDD004A, A01901, Q03001, Cat_Age	G2
3	0.863	23	13	H001, N001, I00102, N004, N005, D001, VDD004A, A01901, Q03001, Cat_Alcohol, Cat_BMI, Age_Group	G3
4	0.862	23	16	C006, P035, H001, N001, I00102, N004, N005, VDD004A, V0026, A01501, A01901, A005010, Q03001, Cat_Alcohol, Cat_BMI, Age_Group	G4
5	0.863	23	9	C006, C009, J037, H001, N001, N004, A01901, Q03001, Age_Group	G5
6	0.864	23	10	C006, P035, H001, N001, N004, N005, D001, A005010, Q03001, Age_Group	G6
7	0.859	23	6	I00102, N004, D001, Q03001, Cat_Alcohol, Age_Group	G7
8	0.864	23	9	C006, H001, N001, N004, D001, VDD004A, A005010, Q03001, Age_Group	G8
9	0.863	23	9	C006, H001, N001, I00102, N004, D001, Q03001, Cat_BMI, Age_Group	G9
10	0.862	23	12	C006, C009, H001, N001, I00102, N004, D001, VDD004A, Q03001, Cat_BMI, Age_Group	G10

Gr (Group):
G1: CS+APS+PSD+CD+CO+PD+EV;
G2: CS+APS+PSD+CD+CO+PD;
G3: CS+APS+CO+EV+PD+PF+PSD+CD;
G4: CS+APS+PSD+CD+CO+PD+EV+PD+PF;
G5: CS+APS+CD+CO+PD+PF+HI;
G6: CS+APS+CD+CO+PD+PF+EV;
G7: CS+APS+PSD+CO+EV+PD;
G8: CS+APS+PSD+CO+PD+PF;
G9: CS+APS+CD+CO+PD+PF
G10: CS+APS+CD+CO+EV+PD+PF).

proach requires a balance between computational efficiency, hardware architecture, and communication structure. In this study, the parallel version of NSGA-II was applied.

The execution of the experiments was made by parameters defined for this study, which establish ranges for population size, number of generations, mutation, and crossover rates. Table 7 details the parameters and configurations adopted for the experiments. The calibration of these parameters is critical, as it directly influences the algorithm's convergence and the fitness of the final solution. The total number of experiments results from the combination of all specified ranges. The fitness function considered both the minimization of model error and the reduction in the number of features.

The experiments employed a distributed master–slave architecture to assign 40 tasks independently. The number of repetitions was set to 10 to ensure statistical representativeness. The master node distributed the tasks among the worker nodes, allowing each server to process a subset of tasks in parallel and independently. At the end of each execution, each worker node sent its results to the master node, which stored them in a specific directory for subsequent consolidation and ranking of the best solutions.

The cluster used in this work consists of one master node and two worker nodes, all running *Ubuntu Server LTS*. The essential configurations of each machine are presented below:

- **Master Node**: *Intel Core i7-4770* processor (4 cores / 8 threads), **32 GB RAM** and operating system **Ubuntu Server 22.04** (kernel 5.x, architecture x86_64).
- **Worker Node 1**: *Dell PowerEdge R430* server with two

Intel Xeon Silver 4114 processors (20 cores / 40 threads total), **192 GB RAM** and operational system **Ubuntu Server 22.04** (kernel 5.x, architecture x86_64).

Despite hardware heterogeneity, particularly in memory capacity and server models, the cluster exhibited stable behavior during the parallel experiments.

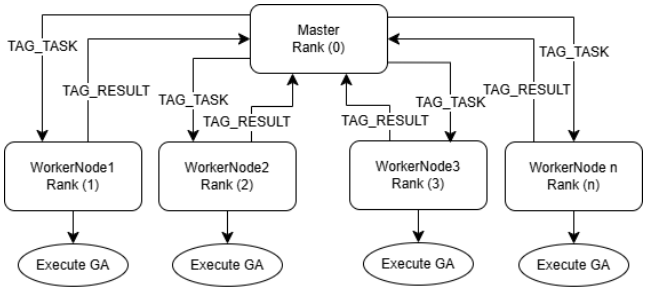

Figure 3. Experiments execution in the master-slave distributed architecture

The Figure 3 illustrates the operation and the distribution of experiments between the master server and the worker servers for conducting the experiments. Figure 4 shows how the experiments were arranged.

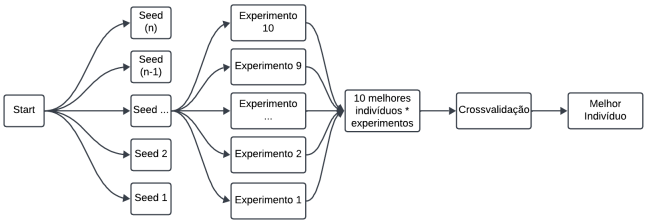

Figure 4. Experiments - GA pipeline for feature selection

Each experiment is executed on a different core, and the final output is the subset of attributes that provides the lowest model error. Table 8 presents the 10 best final experiments from the selection process. Note that the experiments exhibit similar performance in terms of the F1 measure; however, the number of selected attributes varies between 6 and 16, indicating that the attribute selection process via genetic algorithms is sensitive to the chosen parameters. For this work, the solution from experiment 4 (16 attributes) was selected because it involves the largest number of attribute categories (CS+APS+PSD+CD+CO+PD+EV+PD+PF), which may provide useful and non-obvious knowledge about the problem under study.

3.2.7 Class Balancing

A total of 37,507 instances belonging to the healthy individuals class and 2,693 instances belonging to the class of individuals with a positive diagnosis of chronic hypertension and cardiovascular diseases were identified.

Given the class imbalance, the random under-sampling technique (Random_Under_Sampling) was applied to reduce the majority class and randomly balance it with respect to the minority class. As a result, 2,369 random instances were selected from the 30,005 instances belonging to the healthy class, resulting in a training dataset of 4,738 instances with 16 features, plus the class attribute *Class = {Healthy, AH_CVD}*.

To provide a clear overview of the impact of the preprocessing steps on the dataset size, Table 6 summarizes the number of instances and features remaining before and after all stages of the data preparation pipeline. The dataset is available at: https://github.com/licapLaboratory/DataBase-PNS-Hipertensao-Cardio

3.2.8 Application of Machine Learning Models

The distinction between black-box and white-box (interpretable) models is discussed in Loyola-González [2019]. Based on this, for the scope of this study, the Decision Tree, Random Forest, and Naive Bayes models were chosen due to their complementary characteristics: the Decision Tree was selected for its high interpretability; Random Forest for its robust predictive performance; and Naive Bayes as a computationally efficient baseline for categorical data.

The next step consisted of splitting the dataset into training and testing sets, with 30% of the data reserved for testing and 70% allocated to the training set.

4 Experiments and Results Analysis

The experiments were conducted to evaluate the performance of different machine learning models within the proposed domain. In this section, we present the details of the algorithm parameterization and the analysis of the results obtained. For the experiments, a Dataflow implemented in KNIME was used.

4.1 Algorithm Parameterization and Evaluation Metrics

To select the best decision tree–based learning model, different models were built by varying the minimum number of instances per node (MNIN). The selection criteria considered MNIN, model accuracy (ACC), and the average number of antecedents in the rules generated by each model. Models with greater tree depth (lower MNIN values) correspond to higher specificity, while models with lower depth (higher MNIN values) correspond to greater generalization. For this study, MNIN = 12, ACC = 85.7%, and an average of 6 antecedents were considered. Note that this setting represents a balance between specificity and generalization, as shown in Figure 5.

Table 9 shows the confidence intervals (with a significance level of 5%) generated for MNIN = 12 to evaluate the statistical representativeness of the model, which was subjected to a 10-fold cross-validation process using 70% of the training records. The results demonstrate a satisfactory and consistent performance across all metrics considered, since, as presented in Figure 6, the variation among the models does not exceed 2%, evidencing low inter-model variability. This convergence of results across algorithms of distinct natures—such as Naive Bayes and Random Forest—indicates that the selected feature set is representative of the problem domain.

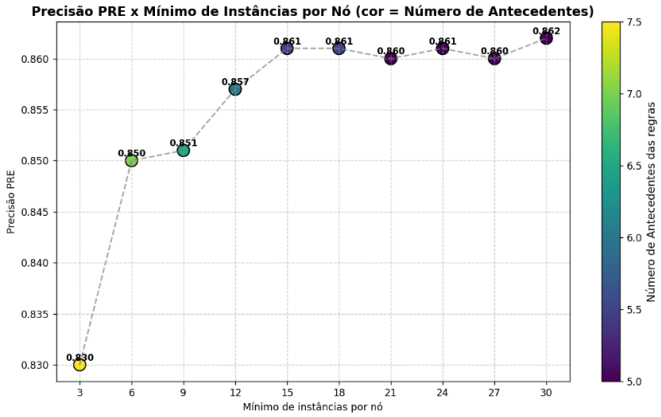

Figure 5. Precision PRE x Minimum instances by node (color = number of predecessors

Such proximity between the values, observed within a balanced dataset of 4,738 instances, suggests that the identified health patterns are not dependent on the specific classifier architecture.

Table 9. Confidence Intervals for the Decision Tree–Based Model

Measure	Lower Bound	Upper Bound
Precision	0.838	0.869
Sensitivity	0.838	0.868
Specificity	0.838	0.868
F-measure_F1	0.838	0.868
Accuracy	0.838	0.868

For the construction of the Random Forest-based model, the main hyperparameters were *min_samples_leaf* = 12 (equivalent to the Decision Tree) and *n_estimators* = 150. For the Naive Bayes-based model, the maximum number of values per attribute was set to 12. All models were built using *k-folds* = 10. Figure 6 presents the results obtained in the testing phase. The models achieved similar performance levels (around 85% accuracy), which are consistent with those reported in related studies, such as Bhatt *et al*. [2023] and de Carvalho *et al*. [2024], that also employ Random Forest models. In addition, the pHi coefficient remained above 0.70, further indicating the robustness and good performance of all three models.

4.2 Results Analysis

Given the satisfactory results presented by the three models, this study considered the Decision Tree for its interpretability and the Random Forest for its robustness and for allowing an analysis of attribute importance within the models.

The Decision Tree-based model generated 125 rules. The Table 10 shows 14 rules that presented the biggest coverage (above 50) and the highest accuracy rate.

Table 10. Identified Rules According to the Implementation in KNIME

No.	Rule description	Coverage	Accuracy	Rate of hits
1	A005010 = "1" AND Cat_BMI = "3" AND A01901 = "1" AND I00102 = "2" AND N001 = "2" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(CD+CS+APS+PD+CO)*	182	167	91.7
2	Cat_BMI = "2" AND A01901 = "1" AND I00102 = "2" AND N001 = "2" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(CD+CS+PD+APS+CO)*	210	200	95.2
3	A01901 = "2" AND I00102 = "2" AND N001 = "2" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(CD+CS+APS+PD)*	117	114	97.4
4	H001 = "4" AND I00102 = "1" AND N001 = "2" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(CS+APS+PD+CO)*	86	77	89.5
5	H001 = "4" AND I00102 = "2" AND N001 = "3" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(CS+APS+PD+CO)*	133	113	84.9
6	N001 = "1" AND Cat_Age = "2" AND N004 = "2" AND Q03001 = "2" => "2" *(APS+PD+CO)*	296	284	95.9
7	I00102 = "1" AND Cat_Age = "4" AND N004 = "2" AND Q03001 = "2" => "1" *(CS+PD+APS+CO)*	61	60	98.3
8	H001 = "4" AND N001 = "2" AND Cat_Age = "3" AND N004 = "2" AND Q03001 = "2" => "2" *(CS+APS+PD+CO)*	53	40	75.4
9	A01901 = "1" AND N001 = "3" AND Cat_Age = "3" AND N004 = "2" AND Q03001 = "2" => "1" *(CD+APS+PD+APS+CO)*	121	108	89.2
10	Cat_Age = "1" AND N004 = "2" AND Q03001 = "2" => "2" *(PD+APS+CO)*	202	199	98.5
11	N001 = "4" AND Cat_Age = "2" AND N004 = "1" AND Q03001 = "2" => "1" *(APS+PD++CO)*	83	76	91
12	A005010 = "1" AND N001 = "3" AND Cat_Age = "2" AND N004 = "1" AND Q03001 = "2" => "1" *(CD+APS+PD+CO)*	136	111	81
13	A005010 = "1" AND Cat_Age = "3" AND N004 = "1" AND Q03001 = "2" => "1" *(CD+PD+APS+CO)*	200	188	94
14	Q03001 = "1" => "1" *(CO)*	522	522	100

1: Healthy 2: AH + CVD (Arterial Hypertension + Cardiovascular Disease) PD: Demographic profile HA: Dietary habits EV: Lifestyle CS: Health care APS: Self-perception of health CD: Household CO: Comorbidity

For cases in the healthy profile, the rule with the highest coverage and highest accuracy rate (Rule 3) considers adequate household conditions, positive health care and self-perception aspects, and the absence of diabetes as relevant characteristics for a healthy condition. For cases in the profile of individuals with hypertension and cardiovascular problems, the rule with the highest coverage and highest accuracy rate (Rule 13) identifies, in addition to adequate household conditions, a demographic profile corresponding to older age, self-perception of getting tired when climbing, and the presence of diabetes as characteristics of the comorbidity. Rule 14 highlights that 100% of individuals with diabetes also have hypertension and cardiovascular problems, as previously noted in the medical literature [Bassi *et al*., 2018].

In the variable importance analysis, Figure 7, age feature was the most relevant factor, corroborating the relationship between age and the risk of chronic diseases. Perception of fatigue when climbing, a diabetes diagnosis, chest pain when walking normally, and self-assessment of health also stood out, indicating that subjective symptoms and associated comorbidities are early indicators. Highest level of education attained, frequency of medical visits, alcohol consumption, and even household infrastructure aspects, such as access to a sewage system, are factors that characterize these profiles. Contextual interpretation of the data is essential, especially in population studies aiming to outline health profiles based on socioeconomic and behavioral factors.

5 Conclusions

The proposed knowledge discovery pipeline introduces innovative aspects compared to traditional approaches, particularly the integration of a dual-stage feature selection strategy. This includes an initial conceptual selection guided by domain knowledge, followed by a parallel implementation of NSGA-II. Additionally, using a distributed master–slave architecture for feature selection significantly reduces computational cost while maintaining model performance. The pipeline proved to be effective for the problem domain under consideration.

The first stage, based on conceptual attribute selection, enabled an initial reduction of dataset dimensionality guided by domain understanding. The second stage, based on a genetic algorithm (GA), further refined the feature set by identifying the most relevant attributes after a rigorous data preparation process. As a result, the machine learning models evaluated in this study demonstrated consistent and competitive performance in classifying healthy individuals and those with hypertension and cardiovascular diseases (AH + CVD), in line with related studies.

The variable importance analysis revealed that attributes such as age group, subjective health perception, fatigue, chest discomfort during exercising, and prior diabetes diagnosis were key determinants for classification, corroborating scientific evidence on cardiovascular risk factors. These findings, in turn, provide incentives for practical public health actions: the relevance of age reinforces the need for active screening in older populations, while the importance of self-reported symptoms suggests the potential of simple questionnaires as low-cost initial screening tools. However, the descriptive nature of the dataset limited the ability to identify asymptomatic individuals, highlighting the importance of complementing population data with detailed clinical information and longitudinal follow-up to capture the complexity of chronic diseases.

To further improve model accuracy, especially in classifying AH + CVD, future studies should prioritize the incorporation of longitudinal data, such as serial medical exams, monitoring of habits, and genetic markers, as well as exploring techniques like deep learning to model nonlinear interactions among variables. Validation on datasets with greater socioeconomic diversity and diagnostic detail is also crucial to ensure the generalizability of the results. Finally, the study emphasizes that predictive modeling in public health requires an integrated approach, combining quantitative data, clinical contexts, and behavioral factors to more effectively address the challenges posed by chronic diseases.

Acknowledgments

The authors acknowledge the support received from the National Council for Scientific and Technological Development (CNPq), Process No. 303133/2021-0, and the Research Support Foundation of the State of Minas Gerais (FAPEMIG), Process No. PCE-00349-25.

References

AlKaabi, L., Ahmed, L., Al Attiyah, M., and Abdel-Rahman, M. (2020). Predicting hypertension using machine learning: Findings from qatar biobank study. *PLOS ONE*, 15(10):e0240370. DOI: 10.1371/journal.pone.0240370.

Alwan, A. (2011). *Global status report on noncommunicable diseases 2010*. World Health Organization, Geneva. 176 pp.

Bassi, D., Cabiddu, R., Mendes, R., Tossini, N., Arakelian, V., Caruso, F., Bonjorno, J. J., Arena, R., and Borghi-Silva, A. (2018). Efeitos da coexistência de diabetes tipo 2 e hipertensão sobre a variabilidade da frequência cardíaca e capacidade cardiorrespiratória. *Arq. Bras. Cardiol.*, 111:64–72. DOI: 10.5935/abc.20180105.

Bhatt, C., Patel, P., Ghetia, T., and Mazzeo, P. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms*, 16(2):88. DOI: 10.3390/a16020088.

Costa, G. and Zárate, L. E. (2025). Mineração de dados para caracterizar indivíduos hipertensos com doenças cardiovasculares no brasil. In *Proceedings of the 40th Brazilian Symposium on Data Bases (SBBD)*, pages 168–181, Fortaleza, CE, Brasil. SBC.

de Araújo, J., de Alencar Rodrigues, R., da Costa Pereira de Arruda Neta, A., *et al*. (2022). The direct and indirect costs of cardiovascular diseases in brazil. *PLOS ONE*, 17(12):e0278891. DOI: 10.1371/journal.pone.0278891.

de Carvalho, N., Gomes, M., and Zárate, L. (2024). Mineração de dados no diagnóstico de hipertensão baseado na pesquisa nacional em saúde 2019. *J Health Inform*, 16(Especial).

Gárate-Escamila, A., El Hassani, A., and Andrès, E. (2020). Classification models for heart disease prediction using feature selection and pca. *Informatics in Medicine Unlocked*, 19:100330. DOI: 10.1016/j.imu.2020.100330.

Gonçalves, L., Franca, D., and Zarate, L. (2024). Relevância do entendimento do domínio de problema na construção de modelos computacionais de aprendizado. In *Anais do XVIII Brazilian e-Science Workshop*, pages 135–142, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/bresci.2024.240233.

Gonçalves, L. and Gálvez, L. (2024). Aplicação de técnicas de mineração de dados e aprendizado de máquina na comparação de perfis de avc entre idosos e adultos de meia-idade: Estudo da pns 2019. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Da-*

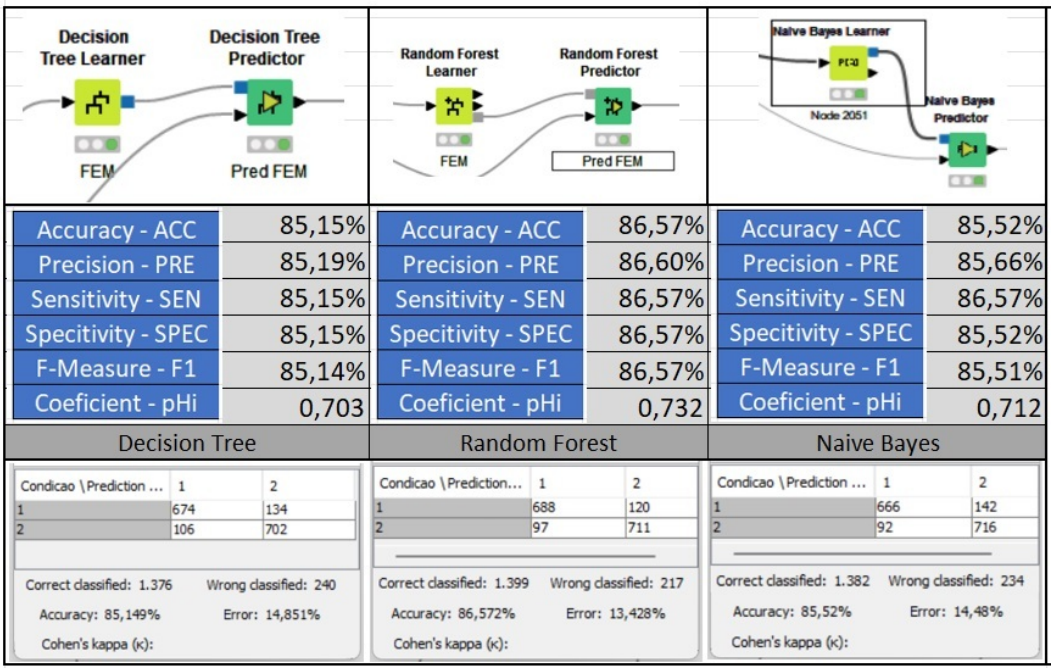

Figure 6. Models Comparison

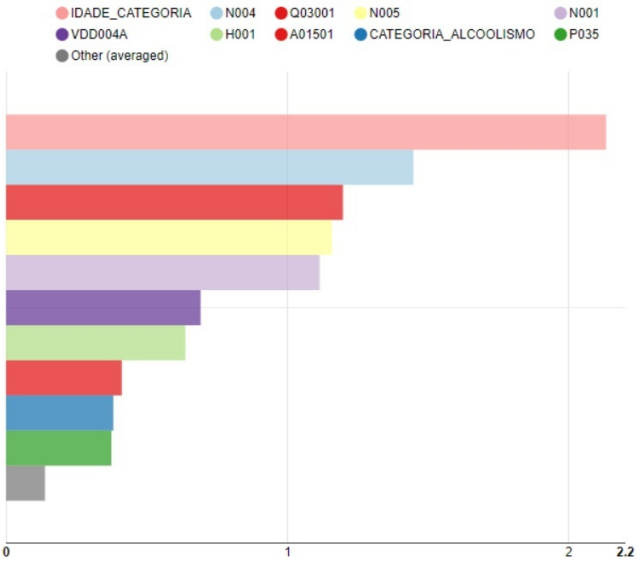

Figure 7. Feature Importance of the Used Attributes

dos, pages 308–317, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd_estendido.2024.243520.

Gonçalves, L. F., Fiorini, C. D. R., França, D. R., Noronha, M. D. M., Song, M., and Galvez, L. E. Z. (2025). Understanding stroke risk profiles in middle-aged adults: A genetic algorithm-based feature selection aproach. In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (HEALTHINF), Volume 2*, pages 623–630. SciTePress. DOI: 10.5220/0013184600003911.

IBGE (2020). Pesquisa nacional de saúde 2019 - instituto brasileiro de geografia e estatística. `https://www.ibge.gov.br/estatisticas/sociais/saude/9160-pesquisa-nacional-de-saude.html?edicao=25921&t=resultados`. Acesso em: 2024-07-15.

Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113. DOI: 10.1109/ACCESS.2019.2949286.

Malta, D. *et al*. (2022). Hipertensão arterial e fatores associados: Pesquisa nacional de saúde, 2019. *Revista de Saúde Pública*, 56:122. DOI: 10.11606/s1518-8787.2022056004177.

NIAAA (2022). Standard alcohol guidelines. National Institute on Alcohol Abuse and Alcoholism.

Powell-Wiley, T., Poirier, P., Burke, L., *et al*. (2021). Obesity and cardiovascular disease: A scientific statement from the american heart association. *Circulation*, 143(21):e84–e118. DOI: 10.1161/CIR.0000000000000973.

Sousa, M. and Zarate, L. (2024). A epidemia silenciosa: Explorando os determinantes comportamentais e socioeconômicos da deficiência renal crônica no brasil. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 318–327, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd_estendido.2024.243624.

Stevens, B., Pezzullo, L., Verdian, L., Tomlinson, J., George, A., and Bacal, F. (2018). The economic burden of heart conditions in brazil. *Arq. Bras. Cardiol.*, 111(1):29–36. DOI: 10.5935/abc.20180104.

WHO (2011). *Global Atlas on Cardiovascular Disease Prevention and Control*. World Health Organization, Geneva.

WHO (2021). Obesity and overweight.

Yang, J., Rahardja, S., and Fränti, P. (2019). Outlier detection: how to threshold outlier scores? In *Proc. of the Int. Conf. on Artificial Intelligence, Information Processing and Cloud Computing*, pages 37–42. DOI: 10.1145/3371425.3371427.

Zilbermint, M., Hannah-Shmouni, F., and Stratakis, C. (2019). Genetics of hypertension in african americans and others of african descent. *Int. J. Mol. Sci.*, 20(5):1081.

DOI: 10.3390/ijms20051081.