FEDERAL UNIVERSITY OF CEARÁ
CENTER OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE
POST-GRADUATION PROGRAM IN COMPUTER SCIENCE
DOCTORAL DEGREE IN COMPUTER SCIENCE

WESLLEY LIOBA CALDAS

IVS: INTERPRETATIVE VARIABLE SELECTION VIA PERFECT BIPARTITE MATCHING

FORTALEZA
2023

WESLLEY LIOBA CALDAS

IVS: INTERPRETATIVE VARIABLE SELECTION VIA PERFECT BIPARTITE MATCHING

Thesis submitted to the Post-Graduation Program in Computer Science of the Center of Science and Technology of the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Computer Science. Concentration Area: Machine Learning

Advisor: Prof. Dr. João Paulo Pordeus Gomes

Co-advisor: Prof. Dr. João Paulo do Vale Madeiro

FORTALEZA
2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C15i Caldas, Weslley Lioba.
Ivs : interpretative variable selection via perfect bipartite matching / Weslley Lioba Caldas. – 2023.
65 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação , Fortaleza, 2023.
Orientação: Prof. Dr. João Paulo Pordeus Gomes.
Coorientação: Prof. Dr. João Paulo do Vale Madeiro..

1. Seleção de atributos. 2. Modelos interpretáveis. 3. Feature selection. 4. Interpretive models. 5. Chagas. I. Título.

CDD 005

WESLLEY LIOBA CALDAS

IVS: INTERPRETATIVE VARIABLE SELECTION VIA PERFECT BIPARTITE MATCHING

Thesis submitted to the Post-Graduation Program in Computer Science of the Center of Science and Technology of the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Computer Science. Concentration Area: Machine Learning

Approved on:

EXAMINATION BOARD

Prof. Dr. João Paulo Pordeus Gomes (Advisor)
Federal University of Ceará (UFC)

Prof. Dr. João Paulo do Vale Madeiro (Co-advisor)
Federal University of Ceará (UFC)

Prof. Dr. José Maria da Silva Monteiro Filho
Federal University of Ceará (UFC)

Prof. Dr. Leonardo Ramos Rodrigues
Institute of Aeronautics and Space (IAE)

Dr. Roberto Coury Pedrosa
Federal University of Rio de Janeiro (UFRJ)

I’m profoundly grateful to my family, whose unwavering love and support have been my driving force. Your belief in me, even during challenges, has been my foundation. This achievement is as much yours as it is mine. Thank you for being my rock on this journey.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

To my beloved mother, Vanadia Silva Lioba, whose unconditional love and constant support have been the driving force that propelled me through every moment of this journey. Every step I took towards this doctoral degree was guided by the memory of your determination and dedication.

To my dear brothers, Asley Lioba Caldas and Odmir Fortes Menezes Caldas Filho, whose presence has always brought joy to my life and whose constant encouragement reminded me of the importance of building a path of success and achievements together.

To my advisor, João Paulo Pordeus Gomes, for his insightful guidance, profound knowledge, and belief in my abilities. Your wise words and valuable advice have been the compass that directed my research and reflections.

To my co-advisor, João Paulo do Vale Madeiro, for his indispensable collaboration, valuable insights, and dedication in sharing his knowledge. Your presence by my side enriched my work in immeasurable ways.

To the friends from Logia (Logic and Artificial Intelligence), Marcelo Veras, Diego Farias, and Alisson Alencar, whose stimulating discussions and exchange of ideas enriched my understanding and broadened my academic horizons.

To all those who, in some way, contributed to my academic journey, my heartfelt gratitude. This work would not have been possible without the support, encouragement, and inspiration I received along the way.

With gratitude,
Weslley Caldas

“The only source of knowledge is experience.”

(Albert Einstein)

ABSTRACT

Feature selection is a fundamental process in machine learning to identify the most relevant subset of features for a given problem. Among the various feature selection approaches, filter methods stand out for their simplicity and efficiency. However, these methods lack interpretability regarding the relationships between the selected and unselected features. To address this challenge, we propose a novel pairwise feature selection method based on Perfect Bipartite Matching, which establishes optimized linear relationships between features, thus facilitating the interpretation of feature connections. We also demonstrate how to incorporate domain knowledge, allowing users to exclude/include desirable patterns (e.g., pre-select specific features). Empirical evaluations using 17 datasets demonstrate the effectiveness of our approach compared to baseline methods. Furthermore, we present a case study on Chagas disease, showcasing detailed interpretation results and the significance of selected features in sudden cardiac death prevention.

Keywords: chagas disease; interpretability; feature selection; machine learning.

RESUMO

A seleção de características é um processo fundamental em aprendizado de máquina para identificar o subconjunto mais relevante de atributos para um determinado problema. Entre as várias abordagens de seleção de características, os métodos de filtro se destacam por sua simplicidade e eficiência. No entanto, esses métodos carecem de interpretabilidade em relação às relações entre as características selecionadas e não selecionadas. Para enfrentar esse desafio, propomos um novo método de seleção de características em pares baseado em Emparelhamento Bipartido Perfeito, que estabelece relações lineares otimizadas entre as características, facilitando assim a interpretação das conexões entre elas. Também demonstramos como incorporar conhecimento de domínio, permitindo aos usuários excluir/incluir padrões desejáveis (por exemplo, pré-selecionar características específicas). Avaliações empíricas utilizando 17 conjuntos de dados demonstram a eficácia de nossa abordagem em comparação com os métodos de referência. Além disso, apresentamos um estudo de caso sobre a doença de Chagas, mostrando resultados de interpretação detalhados e a importância das características selecionadas na prevenção da morte súbita cardíaca.

Palavras-chave: doença de chagas; interpretabilidade; seleção de atributos; aprendizagem de máquina.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS AND ACRONYMS

AP	Assignment Problem
CAD	Computer-aided diagnostics
FFS	Filter Feature Selection
FS	Feature selection
GFSH	Greedy Forward Search Heuristic
IVS	Interpretative Variable Selection
JMI	Joint Mutual Information
LVEF	Left ventricular ejection fraction
MI	Mutual Information
MIM	Mutual Information Maximization
ML	Machine Learning
mRMR	Minimum Redundancy Maximum Relevance
PBM	Perfect Bipartite Matching
PVC	Premature Ventricular Contraction
RAP	Restricted Assignment Problem
SAP	Symmetric Assignment Problem
SCD	Sudden Cardiac Death
SVM	Support Vector Machine
UAP	Unbalanced Assignment Problem
UCI	University of California, Irvine

LIST OF SYMBOLS

v_1	Random variable
v_2	Random variable
v_3	Random variable
a	Random variable
b	Random variable
C	Cost matrix representing the absolute value of the normalized covariance between pairs of features
C_{ij}	Element of the cost matrix C corresponding to the similarity between features i and j
C_{ii}	Element of the cost matrix C corresponding to the similarity between feature i and the target variable Y
D	Number of features
E	Edge set of a graph, connecting the vertices in sets ω and τ
G	Undirected graph with vertex set $V = \omega \cup \tau$ and edge set E
I	Mutual Information
J	Criterion function
J_{JMI}	Criterion for Joint Mutual Information
J_{MIM}	Criterion for Mutual Information Maximization
J_{ivs}	Objective function aiming to maximize similarity with the target variable while minimizing redundancy among features
J_{mRMR}	Criterion for Minimum Redundancy Maximum Relevance
$J'JMI$	Alternative formulation of Joint Mutual Information criterion
L	List of match rounds
M	Matching, a subset of edges in which no two edges share a node
M^r	Matching obtained in the r-th round of matches
S	Selected features set
V	Vertex set of a graph, where $V = \omega \cup \tau$

W	Feature
W'	Set of features associated with feature W
X	Complete feature set
X^r	Subset of features in the r-th round of matches
Y	Target variable
$\mathbb{R}$	Set of real numbers
ω	Set of workers
τ	Set of available tasks
$cov(i,j)$	Covariance between random variables i and j
$\sigma^2 i$	Variances of random variables i
$\sigma^2 j$	Variances of random variables j
corrcoefab	Correlation coefficient between variables a and b
$G(a,b)$	Cost function defined as $1-\lvert\text{corrcoef}ab\rvert$
mi,j	Represents a match between vertex i in set ω and vertex j in set

CONTENTS

1 INTRODUCTION

Feature selection is a crucial step in Machine Learning (ML) to identify the most relevant subset of features for a given problem. It offers several advantages, such as dimensionality reduction, elimination of irrelevant data, noise reduction, avoidance of overfitting, and accelerated ML algorithm training (CHANDRASHEKAR; SAHIN, 2014). Feature selection techniques can be broadly classified into supervised and unsupervised approaches. In supervised feature selection, the selection process considers the target variable or class labels to identify the most discriminative features contributing significantly to the prediction task. These methods aim to maximize the predictive power of the selected features. On the other hand, unsupervised feature selection techniques do not rely on the target variable and focus on capturing the data's underlying structure or intrinsic characteristics. They aim to discover relevant features based on statistical measures, such as variance or clustering analysis.

In addition to supervised and unsupervised classification, feature selection methods can be categorized into different strategies, such as filter, wrapper, and hybrid approaches (MIAO; NIU, 2016). Filter methods assess the relevance of features independently of a specific learning algorithm. They rely on statistical measures or information-theoretic criteria to rank features based on their characteristics. Filter methods are computationally efficient and provide a quick initial feature ranking. Wrapper methods, on the other hand, utilize a specific learning algorithm to evaluate feature subsets by considering their impact on the model's performance. They search through possible feature subsets, evaluating each subset's performance with the chosen learning algorithm. Although more computationally expensive than filter methods, wrapper methods can capture feature interactions and provide more accurate feature rankings. Hybrid approaches combine filter and wrapper elements, leveraging both strategies' advantages to improve feature selection outcomes.

Among the mentioned approaches, Filter Feature Selection (FFS) was widely studied and still has the community's attention, being applied in several scenarios as medical applications (REMESEIRO; BOLON-CANEDO, 2019), marketing strategies (ZHAO *et al.*, 2019), and general classification problems (BOMMERT *et al.*, 2020). Common FFS methods utilize feature correlation measures such as Pearson or Spearman correlation (FORMAN *et al.*, 2003), *chi-squared* function (ZHAI *et al.*, 2018), or Mutual Information (MI) measures, including Mutual Information Maximization (MIM) (LEWIS, 1992), mRMR (PENG *et al.*, 2005), and JMI (YANG; MOODY, 1999). The mentioned approaches are highly regarded for their effectiveness in enhancing accuracy and reducing dimensionality in various data analysis tasks. By evaluating the relevance of individual features through statistical measures or information-theoretic criteria, these techniques efficiently rank the most informative features for subsequent analyses. Beyond their performance benefits, filter methods offer an additional advantage - an explicit ranking system. The straightforward nature of filter-based approaches allows for a clear understanding of the selected features' impact on the model's performance (ZHAO *et al.*, 2019). This transparency enables researchers and practitioners to gain valuable insights into the underlying data patterns, fostering more informed decision-making and facilitating the interpretation of results, a crucial aspect in many real-world applications. However, since many of these algorithms employ greedy forward search heuristics that often consider the global relation between the variables, understanding the relationship between the selected and excluded features can be challenging (KHALID *et al.*, 2014).

To mitigate this issue, we propose a pairwise feature selection method based on the well-known Hungarian algorithm, the Assignment Problem (AP)(KUHN, 1955). Our method utilizes the AP to establish an optimized pairwise linear relationship between features, enabling us to construct a tree that facilitates the interpretation of feature connections. Besides, calculating the pairwise similarity allows us to explicitly identify why some feature was not selected, creating an easy way to connect the unselected and selected features.

Interpretative Variable Selection (IVS) also offers a unique advantage in the context of feature selection. This algorithm has the capability to incorporate specialist knowledge into the feature selection process by modifying the cost matrix used for AP. It can be used, for example, to influence the selection of specific variables, effectively ensuring their inclusion in the final selected set of features. By adjusting the cost matrix strategically, domain experts or researchers can prioritize certain variables based on their domain-specific knowledge and insights. This flexibility adds a valuable dimension to the feature selection process, allowing the integration of expert guidance and domain-specific constraints, which can be particularly beneficial in scenarios where certain variables are known to be critical or where domain expertise plays a crucial role in feature selection decisions.

To showcase the usability and efficacy of IVS, we conduct experiments using multiple datasets across various domains to demonstrate the comparable performance of our method against baseline approaches. Furthermore, we present a case study focused on Chagas disease (MARIN-NETO *et al.*, 2023). We provide detailed interpretation results by applying our feature selection method, shedding light on the connections between selected features and their significance in the context of Sudden Cardiac Death (SCD) prevention.

1.1 General Goal

This thesis aims to contribute to the field of machine learning by proposing a novel Feature Selection method that enhances interpretability and performance in various data analysis tasks.

1.2 Specific Goals

1. Investigate and compare the performance of various feature selection methods regarding accuracy, interpretability, and efficiency.
2. Develop a novel pairwise feature selection method using Perfect Bipartite Matching (PBM) to optimize feature relations and construct an interpretable graph for Chagas disease diagnosis.
3. Analyze the interpretability of the Computer-aided diagnostics (CAD) system by examining the relationships among selected features and providing insights into the diagnostic process.
4. Conduct empirical evaluations using multiple University of California, Irvine (UCI) datasets to assess the generalizability of the proposed method.
5. Validate the efficacy and reliability of the developed CAD system through extensive clinical validation studies in real-world healthcare settings.

1.3 Publications

The following publications were made in the development of this work:

1. CALDAS, W. L.; MADEIRO, J. P. V.; MATTOS, C. L. C.; GOMES, J. P. P. A new methodology for classifying qrs morphology in ecg signals. In: IEEE. **2020 International Joint Conference on Neural Networks (IJCNN)**. *[S.l.]*, 2020. p. 1–9.
2. PRIMO, P. E.; CALDAS, W. L.; ALMEIDA, G. S.; BRASIL, L. P.; CAVALCANTE, C. H.; MADEIRO, J. P.; GOMES, D. G.; PEDROSA, R. C. Auxílio ao diagnóstico para predição de morte súbita em pacientes chagásicos a partir de dados clínicos: uma abordagem baseada em aprendizagem de máquina. In: SBC. **Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde**. *[S.l.]*, 2021. p. 335–345.

3. CALDAS, W. L.; MADEIRO, J. P. do V.; PEDROSA, R. C.; GOMES, J. P. P.; DU, W.; MARQUES, J. A. L. Noise detection and classification in chagasic ecg signals based on one-dimensional convolutional neural networks. In: SPRINGER. **International Conference on Computer and Information Science**. *[S.l.]*, 2022. p. 117–129.
4. CAVALCANTE, C. H.; PRIMO, P. E.; SALES, C. A.; CALDAS, W. L.; SILVA, J. H.; SOUZA, A. H.; MARINHO, E. S.; PEDROSA, R. C.; MARQUES, J. A.; SANTOS, H. S. *et al.* Sudden cardiac death multiparametric classification system for chagas heart disease's patients based on clinical data and 24-hours ecg monitoring. **Mathematical Biosciences and Engineering**, v. 20, n. 5, p. 9159–9178, 2023.

1.3.1 Organization

The remainder of this thesis is structured as follows: Chapter 2 provides an overview of related work in the literature for Feature selection (FS) and describes the Perfect Bipartite Matching (PBM). Chapter 3 cover our proposed approach. Chapter 4 presents the experimental setup and the results obtained using our proposed method. Finally, Chapter 5 concludes the paper by summarizing our findings and outlining potential directions for future research.

2 LITERATURE REVIEW

2.1 Feature Selection.

This section will cover the state-of-art of feature selection algorithms, as the score of this work is supervised FFS; we will briefly cover the most critical supervised filter selections and then go deep into FFS methods. First, in classification and regression problems, FS is a crucial preprocessing step to eliminate irrelevant or redundant data, thereby enhancing learning accuracy and comprehensibility (CHANDRASHEKAR; SAHIN, 2014). It involves selecting a subset of relevant features, encompassing various approaches, like feature filtering, feature wrapping, and hybrid methods. Each approach has its advantages and limitations. Feature filtering is a simple and fast method, whereas feature wrapping tends to outperform feature filtering but is more prone to overfitting. Hybrid methods aim to find a balance between these two approaches (KHALID *et al.*, 2014).

Filter methods are widely used as a preprocessing step independent of the learning algorithm. They evaluate each feature individually or about others, based on statistical measures like MI or Chi-Square, and rank them according to their relevance or importance (KUMAR *et al.*, 2017). One advantage of these methods is their robustness to overfitting since they are not dependent on the specific learning algorithm being used and its simplicity, as most of the algorithms rely on the Greedy Forward Search Heuristic (GFSH) (KURSA, 2021).

Wrapper methods select subsets of features by incorporating feature selection as an integral part of the model-building process. These methods treat a learning algorithm (such as a classifier or regression model) as a black box to evaluate various feature subsets and optimize the model's performance. Wrapper methods often yield more accurate feature subsets than filter methods, but they can be computationally expensive and prone to overfitting, mainly when dealing with small datasets. Some noteworthy examples of wrapper methods include Lasso regression (ZHANG; HUANG, 2008), Recursive Feature Selection (CHEN; JEONG, 2007), Ridge regression (NG, 2004) and Random Forests (KURSA; RUDNICKI, 2010).

Hybrid methods aim to leverage the strengths of both filter and wrapper approaches. In a two-step process, they first employ a filter method to reduce the feature space by selecting the most relevant features. Subsequently, a wrapper method is applied to the filtered feature set to further fine-tune the feature selection process. This approach balances computational efficiency and performance, addressing potential overfitting issues associated with pure wrapper methods while enhancing performance compared to filter methods alone. Notable hybrid methods include the independent component analysis in conjunction with MI criteria (STONE, 2002), and Genetic Algorithm with MI criteria (HUANG *et al.*, 2007).

In summary, feature selection is a critical step in data preprocessing, and choosing the appropriate method depends on the specific problem, dataset characteristics, and computational resources available. Wrapper methods are effective but computationally expensive; filter methods are efficient but might not capture feature interactions. Hybrid methods compromise these two approaches, aiming for improved performance and efficiency, but are much more complex. Table 1 summarizes all the described approaches.

Table 1 – Comparison between feature selection strategies of relevance and redundancy estimations. Relevance metrics measure how important a feature is in relation to the specific prediction or classification task at hand. Redundancy metrics evaluate how much one feature is correlated with or redundant with respect to other features.

Method	**Type**	**Relevance Measure**	**Redundancy Measure**
JMI	Filter	Mutual Information	Mutual Information
MIM	Filter	Mutual Information	-
MRMR	Filter	Mutual Information	Mutual Information
chi^2	Filter	Chi-Square	-
Lasso Regression	Wrapper	Coefficient Magnitude	-
Ridge Regression	Wrapper	Square of the Coefficient Magnitude	-
Random Forest	Wrapper	Coefficient Magnitude	-
Hybrid GA (Genetic Algorithm)	Hybrid	Fitness Function	Genetic Diversity
mRMR-ICA	Hybrid	Mutual Information	Independent Component Analysis

Source: Author.

As mentioned before, the FFS on table 1 relies on GFSH, and, before going deep into these methods, we will first formally introduce the GFSH. First, consider a system (X,Y) consisting of D features denoted by X_i and a target variable Y. The algorithm starts with an empty list of selected features, denoted as S, and adds features step-by-step based on the maximal value of a specific criterion function J. Once a feature is selected and added to S, it is not considered again during the selection process (KURSA, 2021). Below, we will cover the most common criteria used with this approach.

Correlation-based criteria, such as Pearson or Spearman correlation, are the simplest criteria for feature selection. These approaches select features that exhibit the highest correlation with the target label and work well when the relationship between the features is linear (FORMAN *et al.*, 2003). Another commonly used approach replaces covariance with the chi-squared function, a statistical test employed to determine if there is a significant difference between observed categorical variables and expected frequencies. According to (ZHAI *et al.*, 2018), this approach has yielded competitive results in text mining. Based on the previous criteria, we can easily formulate a generic optimization function named Top_k where the algorithm evaluates each feature's relevance by scoring them against a similarity measure F and selecting the best k features. Below we present the equation:

$$J_{top_k} = \sum_{W \in S} F(W;Y) \tag{2.1}$$

Another well-known family of feature selection algorithms is based on MI(BENNASAR *et al.*, 2015), which measures the statistical dependence between two random variables. MI is a non-parametric technique capable of handling non-linear relationships, making it widely applicable in various fields (SIDDIQI *et al.*, 2020; ZHOU *et al.*, 2020; ZHOU *et al.*, 2022). Some commonly used methods within this category include MIM(LEWIS, 1992), mRMR (PENG *et al.*, 2005), and JMI(YANG; MOODY, 1999).

Let‘s start describing the simplest method of MI family, the MIM, which is considered the top_k version based on MI, where the criteria focus solely on the mutual information between a feature and the decision:

$$J_{MIM} = \sum_{W \in S} I(W;Y) \tag{2.2}$$

Even though the MIM present robust results, it disregards inter-feature interactions (PENG *et al.*, 2005). As an alternative, the mRMR algorithm aims to identify a subset of features relevant to the target variable while also having low correlation among themselves. To achieve this, the algorithm computes the MI between each feature and the target variable, as well as the MI between each feature and all other features in the dataset (PENG *et al.*, 2005). The features are then ranked based on a combination of these two values to maximize relevance and minimize redundancy, as shown in the following equation:

$$J_{mRMR} = I(X_i;Y) - \frac{1}{|S|} \sum_{W \in S} I(X_i;W) \tag{2.3}$$

Here, $I(v_1;v_2)$ represents the MI between random variables v_1 and v_2. Unfortunately, mRMR is unable to detect more complex interactions as it assumes that $I(X_i;Y)$ is an upper bound of feature significance (KURSA, 2021). In contrast, JMI iteratively selects features, taking advantage of the already selected features. The JMI criterion is formulated as follows:

$$J_{JMI} = \sum_{W \in S} I(X_i, W; Y) \tag{2.4}$$

Here, $I(v_1, v_2; v_2)$ represents the MI between random variables A and B given the variable C. Alternatively, because $I(v_1, v_2; v_3) = I(v_1; v_3 | v_2) - I(v_3; v_2)$, the criteria can also be formulated as:

$$J'_{JMI} = \sum_{W \in S} I(X_i; Y | W) \tag{2.5}$$

While the abovementioned methods yield good results in many problems, they lack interpretability and visualization tools to help users understand why certain features are selected while others are removed. The interpretability of these methods is normally based on the similarity between selected variables and the target variable; Theoretical frameworks were also proposed to offer another level of interpretation based on the statistical assumptions (BROWN *et al.*, 2012). Still, the optimizations equations 2.1 2.2, do not take into account the similarity between variables, and the equations 2.3, 2.4 calculate the global similarity between the selected features and the unselected ones; Unfortunately, this strategy makes hard to determine individual relationships; and because that makes the interpretation between a specific selected and unselected feature complicated.

To address this issue, we propose a pairwise feature selection method based on the assignment problem, which allows users to identify the relationship between selected and unselected features one-by-one while keeping competitive results.

2.2 Perfect Bipartite Matching Problem

The PBM is a well-known problem in graph theory that involves finding a maximum cardinality set of pairwise non-adjacent edges in a bipartite graph. This problem has found applications in various fields of machine learning and data mining, including multiview learning, recommendation systems, and object tracking (HASHEMI *et al.*, 2021; BEIRANVAND *et al.*, 2022; LI; CHEN, 2013; ZHANG *et al.*, 2016; WANG *et al.*, 2021). Several algorithms can be employed to solve the PBM, such as the Hopcroft-Karp Algorithm, Ford–Fulkerson algorithm, Hungarian algorithm, linear programming, auction algorithm, and branch and bound method (GERARDS, 1995). Table 2 summarizes the complexity and characteristic for each one of the methods.

Table 2 – Methods for solving the Assignment Problem. Where O represents the asymptotic complexity, V is the number of vertices or nodes, E the umber of edges, F is the Maximum flow value (used in Ford–Fulkerson algorithm), and n is the number of nodes.

Method	**Complexity**	**Algorithm Type**	**Description**
Hopcroft-Karp Algorithm	$O(\sqrt{V}E)$	Augmenting Path	Finds maximum matching in bipartite graphs.
Ford–Fulkerson Algorithm	$O(EF)$	Augmenting Path	Finds maximum flow in a network.
Hungarian Algorithm	$O(n^3)$	Augmenting Path	Solves the assignment problem for square cost matrices.
Linear Programming	$O(n^3)$	Optimization	Can be used for solving assignment problems with non-square cost matrices.
Auction Algorithm	$O(n^2 \log n)$	Combinatorial Auction	Iteratively finds prices and allocations in auction-like manner.
Branch and Bound Method	Exponential	Optimization	Systematically explores the search space for optimal assignment.

Source: (GERARDS, 1995).

Formally, let G be an undirected graph with vertex set $V = \omega \cup \tau$ and edge set E, where ω and τ are two disjoint sets of vertices, and E represents the set of edges connecting the vertices in ω and τ. A matching $M \subseteq E$ is a subset of edges in which no two edges share a node. A matching M is considered maximal if its cardinality is maximal among all matchings. The PBM aims to find a maximum cardinality set of pairwise non-adjacent edges, where each edge connects a vertex in ω to a vertex in τ. The problem can be mathematically formulated as follows:

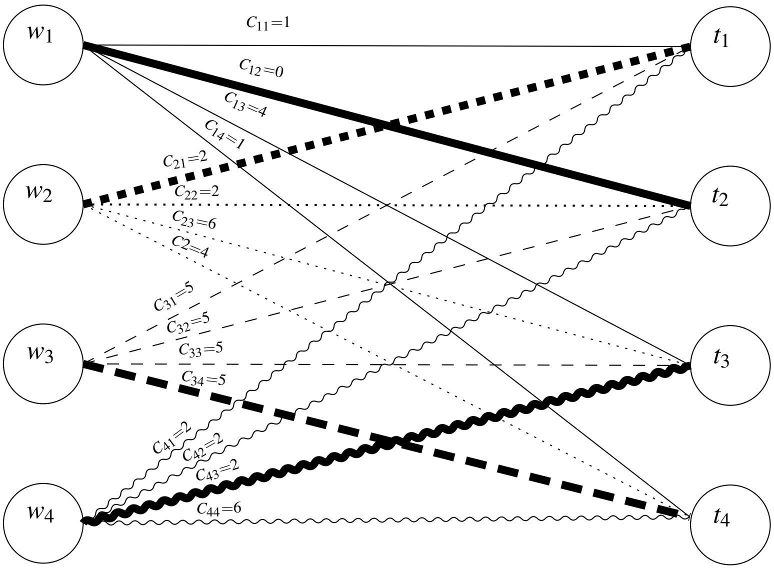
$$M = \max\{|M'| : M' \subseteq E, (\forall w \in \omega)(\exists! t \in \tau)(u,t) \in M'\} \tag{2.6}$$

A specific case of the PBM incorporates a weight function $w : E \to \mathbb{R}$ assigning weights to the edges. In this case, the objective is to find a subset $M \subseteq E$ that minimizes the total weight of the edges. This variant is often referred to as the AP, a well-known optimization theory problem. The AP involves assigning a set of resources to a set of tasks while minimizing the overall cost or maximizing the overall profit. In the formulation of the AP, ω represents the set of workers, τ represents the available tasks to be completed, and E represents the cost matrix for each worker to perform a task. The mathematical formulation is as follows:

$$\min\left\{\sum_{(u,t)\in M} w(u,t) : M \subseteq E, (\forall u \in \omega)(\exists! t \in \tau)(u,t) \in M\right\} \tag{2.7}$$

A solution to the AP is illustrated in Figure 1. While the Hungarian method is often used to solve the AP, we chose to use the Hopcroft-Karp Algorithm due to its better asymptotic time complexity of $\mathcal{O}(|E|\sqrt{V})$ compared to the Hungarian method's $\mathcal{O}(|V|^3)$. It should be noted that in the worst-case scenario, $|E| = |V|^2$.

Figure 1 – Solution for the Balanced Assignment example for $n = 4$. Each one of the wavy edges represents a different match between a worker and a task.



Source: Author

2.2.1 *Assignment variants*

In addition to the traditional AP, several variants are commonly encountered. Three notable variants are the Unbalanced Assignment Problem (UAP), the Restricted Assignment

Problem (RAP), and the Symmetric Assignment Problem (SAP).

The unbalanced assignment problem arises when the sets ω (representing the workers) and τ (representing the available tasks) in the bipartite graph have different cardinalities, making finding a match that connects all elements in both sets impossible. To address this issue, dummy nodes can be introduced with zero cost in the smaller set to balance the sizes of the sets. Note that including dummy nodes does not affect the optimization costs since their costs are constant (PENTICO, 2007).

On the other hand, the restricted assignment problem deals with situations where certain workers from ω cannot perform specific tasks in τ (WANG; SITTERS, 2016). This can be represented by removing edges in the bipartite graph or increasing the value to a high number to prevent the solver algorithms from picking up specific relations.

Another variant worth mentioning is the SAP, where the cost matrix is symmetric. In the SAP, if a match $m_{i,j}$ belongs to the matching set M, then the symmetric match $m_{j,i}$ must also be in M. This particular case of the SAP, in which the diagonal cells of the cost matrix have infinite costs, is equivalent to the perfect matching problem in graph theory (MURTY, 1967). It is important to note that perfect matches can only occur in graphs with an even number of vertices, thus requiring an even number of assignments. While there are specialized approaches to solving the SAP, they often have higher time complexity compared to the previous algorithms (DERIGS, 1978). As an alternative to the existing approach, we propose an efficient solution that leverages the Hopcroft-Karp algorithm. Our method aims to improve the matching process while maintaining a manageable time complexity. Here's how our approach works:

1. **Removing Non-Symmetric Matches:** We begin by analyzing the matches in the set M. We identify and eliminate any non-symmetric matches, which are represented by $m_{i,j}$ in M but $m_{j,i}$ is not present in the set. By removing these non-symmetric matches, we ensure that each remaining match is bidirectional and symmetrical.
2. **Adding Symmetric Matches Greedily:** After rectifying the matches, we proceed to add symmetric matches in a greedy procedure. This involves iteratively selecting the best pairs of entities and adding them to the set M. The process continues until no more symmetric matches can be found.

Importantly, this additional step of adding symmetric matches does not introduce a significant increase in the overall time complexity of the assignment algorithm. The time complexity of this operation is $\mathcal{O}(|E|)$, meaning it scales linearly with the number of entities(edges of the graph) in the matching pool. As a result, the overall time complexity of our proposed solution remains the same as the original algorithm.

Considering these variations, exploring different approaches, and adopting appropriate algorithms, the AP and its variants can be effectively solved and optimized in various real-world scenarios, providing valuable insights and practical solutions.

3 INTERPRETATIVE VARIABLE SELECTION VIA PERFECT BIPARTITE MATCHING (IVS)

This section presents a novel feature selection method based on the AP. We begin by briefly introducing feature selection and then present the general idea of the proposed method, along with its pseudocode. Finally, we discuss the potential interpretation of the results obtained through this technique.

Formally, given a standard classification/regression problem $f : X \rightarrow Y$, where $X \in \mathscr{R}^{NxM}$ represents the input data with n rows and D features, and Y is the target variable, our objective is to identify the optimal subset of features S that best represent the data for constructing the predictive function f.

In order to construct the mapping function f, we propose a feature selection algorithm based on the AP, which involves identifying the most similar pairs of variables iteratively and discarding the less relevant ones. We also demonstrate how we can adapt the AP to the context of feature selection.

First, the AP is primarily concerned with matching a set of workers with tasks while minimizing a cost matrix. Based on this formulation, we introduce a slight modification where the sets of workers and tasks correspond to the complete feature set; note that although they conceptually represent the same objects, they are still separate sets as needed in the assignment problem, the idea behind this modification is to associate feature pairs rather than worker/task pairs.

To construct the cost matrix C we can use any similarity metric, such as covariance or mutual information. Our approach uses a common variant of normalized covariance: the Pearson product-moment correlation coefficient due to its simplicity and faster implementation than other similarity measures. It represents the covariance ratio between two random variables to the square root of the product of their variances. Mathematically, one can express it as:

$$corrcoef_{ij} = \frac{cov(i,j)}{\sqrt{\sigma_i^2 * \sigma_j^2}} \tag{3.1}$$

Here, i and j denote two random variables, $cov(i,j)$ is their covariance, and σ_i^2 and σ_j^2 represent their variances. The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 1 represents a perfect positive correlation, and 0 implies no association between the variables.

From the perspective of feature selection, positive and negative correlations hold equal importance. Therefore, we take the absolute value of Equation 3.1 and subtract it from 1 (converting the weight minimization into weight maximization) to define our cost function $G(a,b) = 1 - abs(corrcoef_{ab})$. The cost matrix C can be defined as follows:

$$C_{ij} = \begin{cases} G(i,Y) & \text{if } i = j \\ G(i,j) & \text{otherwise} \end{cases} \tag{3.2}$$

Since the sets of workers and tasks correspond to the features, the cost matrix C represents the absolute value of the normalized covariance between all pairs of features in the complete feature set. Additionally, the diagonal elements of C are replaced with $G(i,Y)$, where Y denotes the target variable.

Next, we propose the following objective function:

$$j_{ivs} = \sum_{W \in S} G(W,Y) - \sum_{W \in S} \sum_{V \in W'} G(W,V) \tag{3.3}$$

Unfortunately, like most filtering methods, maximizing the cost function in Equation 3.3 is quadratic in terms of D, as it requires evaluating all feature permutations to find the optimal subset S. Instead of employing a greedy forward approach and evaluating features individually, we propose an algorithm based on multiple rounds of the assignment problem to determine S.

The core idea is to find the best global pairwise matches recursively. Initially, we compute the cost matrix C and apply the SAP variant instead of the classic AP to X, a necessary condition to avoid cycle dependency between features (see subsection 3.1). The output consists of a set of pairs that precisely corresponds to a total of M pairs, because we used the SAP variant a match between feature i and j (represented as m_ij) is essentially redundant when compared to a match between j and i (represented as m_ji), as they both capture the relationship between features i and j, based on this assumption, we can simplify and assume that there are only $|M|/2$ distinct pairs in the output. In other words, half of the features are matched with the other half. For each pair of matched features, we choose the feature with a higher C^r_{ii} value (indicating greater similarity to the target variable) forming a new set, denoted as X^r, where r represents the number of rounds of SAP performed. We repeat the SAP again in the r-th round of matches (starting from 0, i.e., $X = X^{r=0}$). We repeat this process recursively until $|X^r| > 2$ and store M^r in a list of match rounds denoted as L.

After several rounds of matches, the last feature in the final match round exhibits high similarity to Y while leaving out its correlated features from the previous steps. To select the feature set S, we sort the features in reverse order based on their appearance in L. For nodes at the same level (number of rounds of assignment or the height on the tree; see section 3.2), we use their similarity with Y as a tiebreaker (higher similarity is favored). Finally, we select the first $|S|$ features. We present the pseudocode of this algorithm in 1.

Algorithm 1 IVS Pseudocode

```
Input: X, Y, $|S|$
Output: subset of features S
$C \leftarrow G(X,Y)$
$X^r \leftarrow X$
$L \leftarrow []$
while $|X^r| \leq 2$ do
    if $|X^r|/2$ is odd then
        $X' \leftarrow \textsc{AddFakeNode}(X')$
    end if
    $C' \leftarrow \textsc{CalculateCurrentCosts}(X^r, C)$
    $C' \leftarrow \textsc{SetInfinityDiagonal}(C')$
    $M^r \leftarrow \textsc{FindAssigment}(C')$
    $X^r \leftarrow \textsc{SelectNextCandidates}(X', M^r)$
    $L.append(M^r)$
end while
$T \leftarrow \textsc{sortNodes}(L, X, C)$
$S \leftarrow \textsc{getBestFeatrues}(T)$
```

3.1 Dealing with the cycle dependency

As described before, for each round of the AP with M matches, it's necessary to ensure that we going to have only $|M|/2$ distinct pairs, otherwise, some features may have duplicate matches (despite the symmetric ones) forming a cycle dependency between them. For example, suppose the presence of a triple of matches m_ij, m_jk, m_ki. It's not possible to select which feature propagates because we have multiple relations for each one of them. To address this, we must ensure that if m_{ij} represents a match from feature i to feature j in the match set M^r, then m_{ji} must also be a match. To enforce this condition, we must use SAP variant applying the following conditions:

1. The value of $|X^r|/2$ must be even for each round.
2. The diagonal of C is set to infinity.

The first condition is necessary because an odd value of $|X^r|/2$ implies that it is impossible to have all matches for both m_{ij} and m_{ji} simultaneously, as this requires an even cardinality of matches. To address this, a dummy node with zero cost can be added to the feature set, as described in the 2.2.1 subsection. This modification does not interfere with the optimization problem and does not require any changes to the current algorithm. The second condition is easily fulfilled by setting the diagonal of C^r to infinity. Even for non-symmetric assignment problems, this step is necessary to prevent matches between the same features. Note that m_{ii} represents a match between worker i and task i, but since workers and tasks correspond to features, a match between the same feature is redundant.

3.2 Interpretation and visualization

Traditional filter methods such as mRMR, MIM, JMI, and Top-k chi-square provide users with a certain level of interpretation, typically based on the similarity level between the target variable and the selected features. While valuable and easy to interpret, these methods lack explainability regarding the intra-relationship among the features. In contrast, wrapper methods like tree-based feature selection offer better interpretability by revealing the structure that governs the feature relationships. However, as noted in (JOVIĆ *et al.*, 2015), wrapper methods are more prone to overfitting. In this work, we propose a simple approach to interpret the results of *IVS* while maintaining the simplicity of filter methodologies. We achieve this through constructing a hierarchical graph, which elucidates the relationships between all features and explains why they are selected.

To construct the hierarchical graph G, we create a node for each feature in X. The value of each node is computed using Equation 3.1, which represents the normalized covariance between the feature and the target variable. In each round of the assignment problem, we add an edge between the matched nodes on the edges set E. Using the last selected node as the root, a tree is formed where each edge denotes the linear relationship between the parent and child nodes. Notably, the tree exhibits a unique property where the value of each parent node must be greater than or equal to that of its child node. This property holds because a node with a higher value will continue to be propagated in the subsequent assignment rounds.

The tree structure of the graph facilitates an intuitive interpretation of the results. Starting from the root node, representing the feature with the highest linear similarity to the target variable, one can traverse the edges downward to explore a sequence of linearly related features. Parent nodes, with higher values than their child nodes, indicate features more strongly related to the target variable. By examining the value of each edge, it becomes possible to understand why certain features were not selected (e.g., they may be relevant but highly similar to other selected features). This sequence of features can serve as a feature subset for further analysis, such as model training or feature engineering. Section 4 presents an illustrated example for SCD, including how to build the graph and how to interpret it.

In summary, the hierarchical graph visually captures the linear relationships between features and the target variable, enabling a straightforward interpretation of the results and identification of feature subsets for subsequent analysis.

3.3 Incorporating Domain Knowledge

This section explains integrating domain-specific knowledge into feature selection by adjusting the cost matrix. It offers strategies to enhance feature relevance, ensure feature propagation, handle variable relationships, and strike a balance between domain knowledge and algorithmic objectivity.

The approach to incorporate domain-specific knowledge is to manipulate the values in the cost matrix (C), which influences feature similarity and is crucial in the selection process. Here are strategies for modifying these values:

3.3.1 Emphasizing Feature Relevance

In scenarios where certain features are known to be more crucial due to domain expertise, we can increase their feature relevance value. By increasing the similarity values between these features and the target variable Y, we encourage their inclusion in the final feature subset. This adjustment effectively elevates their importance in the feature selection process, making it more likely for these features to be selected.

3.3.2 Reducing Redundancy

In cases where redundancy among features is a concern, we can set the cost matrix values to enforce matches between related features. This ensures that important relationships between variables are preserved in the selected subset.

For example, if we know that features A and B are highly correlated and both contain valuable information, we can set C_{AB} and C_{BA} to encourage their simultaneous exclusion.

3.3.3 Preserving Relevance

Conversely, if we want to preserve features with the highest relevance by retaining multiple representative features from a group of correlated variables, we can set low-cost relation values to discourage matches between these variables. This ensures that the algorithm propagates the redundant features in different branches, preventing them from being removed.

3.3.4 Forcing Feature Propagation

If we increase the relevance for a single variable for the highest value, we can guarantee the variable will be selected because of the propagation ability of IVS. This property does not hold if we increase the feature relevance for multiple variables. It is insufficient to guarantee that all features will be present in the final sub-selected feature set because some may be removed because of redundancy.

As a countermeasure, we can strategically set their cost matrix values to ensure that specific features are propagated through the selection process. For example, if we have prior knowledge that features A and B are highly relevant, we can set high values for C_{AA}, C_{BB} while we set C_{AB} and C_{BA} to lower values compared to other feature pairs, making the algorithm favor these features to match other pairs and avoiding one of them being out of the final subset of selected features.

In other words, to guarantee a subset of features k to be inserted on the final subset, we must ensure:

1. for each pair ij we set $C_{ij} = 0$
2. for each variable i, we normalize $C_{ii} = C_{ii} + c$, where $c = 1$ Since the values of cost matrix C fall into a range between 0 and 1, adding a constant value c bigger or equal than one will force the k variables to be propagated once they will never match each other because

of the previous clause.

3.3.5 Balancing Act

It’s important to strike a balance between domain knowledge and algorithmic objectivity. While injecting domain expertise can improve the quality of feature selection, overly biased adjustments may lead to suboptimal results. Experimentation and validation should guide the fine-tuning process to ensure that domain knowledge enhances, rather than hinders, the feature selection algorithm.

Incorporating domain knowledge into the cost matrix empowers us to steer the feature selection process toward more informed and context-aware decisions. It enables us to leverage our understanding of the problem domain to select the most relevant and valuable features while managing redundancy effectively.

3.4 Running Example

This section will present how the algorithm works and how to interpret the results, taking an artificial data set in a controlled scenario as an example.

3.4.1 Description of the Toy Data Set

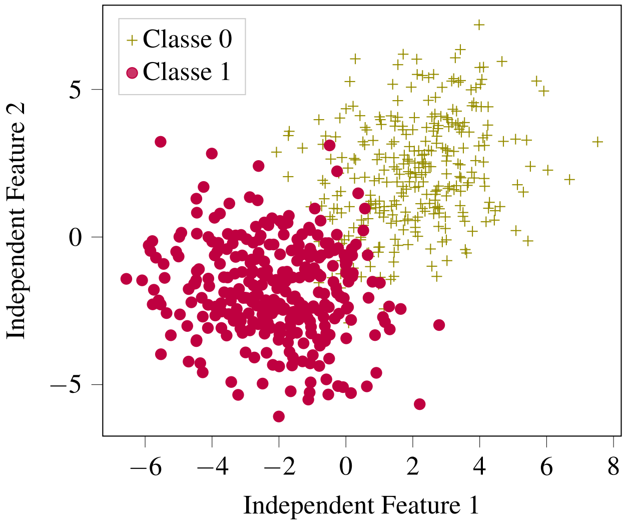
This subsection describes an experiment based on an artificial data set designed to facilitate a classification experiment to explore the interpretability of the feature selection algorithm. The data set consists of two classes: Class 0 and Class 1. Each class represents a distinct category that the classification algorithm aims to identify based on the input features. It also contains 600 samples evenly distributed across the two classes. The data generation process description can be consulted as follows:

1. **Informative Features (2):** Two informative features are generated using multivariate normal distributions. Class 0 data points are generated with a mean vector of $[2,2]$ and a covariance matrix of $\begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$. In contrast, class 1 data points are generated with a mean vector of $[-2,-2]$ and a covariance matrix of $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$. These features contribute significantly to the separation of the two classes.

2. **Redundant Features (4):** Four redundant features are created to investigate how the presence of such features affects the interpretability of the classification algorithm. Redundant features are obtained by combining the informative features linearly with adjustable coefficients and adding random noise using a normal distribution with a mean of 0 and a standard deviation of 1. This process ensures that the redundant features are related to the informative ones but do not provide any additional discriminatory information.
3. **Repeated Feature (1):** A single repeated feature is introduced to the data set. This feature is identical to the first informative feature, allowing one to examine the algorithm's sensitivity to repeated information and its impact on interpretability.
4. **Non-informative Feature (1):** One non-informative feature is included in the data set, which is purely random noise. This feature allows for investigating the algorithm's ability to differentiate between informative and uninformative variables.

The final data set has 8 features (2 informative + 4 redundant + 1 repeated + 1 non-informative). Figure 2 shows the 2D projection of the data set is visualized using scatter plots.

Figure 2 – Scatter plot for the synthetic data set. On the X label, we have the informative feature 1; On the axis Y, we have the informative feature 2.

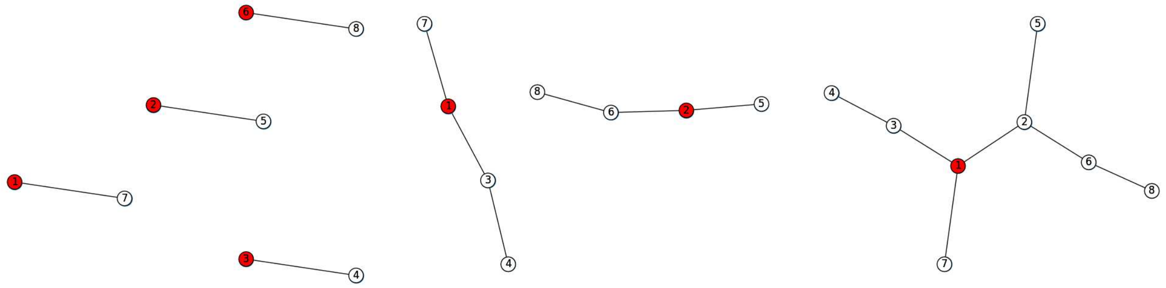


Source: Author

3.4.2 How to interpret the results

The first step in the algorithm execution is to build the cost matrix using the function 3.2. Note that for $n = 8$, the matrix will have $n^2 - n$ cells representing the similarity of each feature and n cells on the diagonal expressing the relevance of each feature (similarity with the target variable). We will then define the value $k = 2$, for the number of selected features. Then we will apply rounds of the assignment problem to form pairs of similar attributes.

Figure 3 – Aggregation rounds of the proposed method for the synthetic data set



(a) First round: we performed the SAP; The red node is more significant and will be sent to the next round.

(b) Second round: the double are matched, forming quartets of features.

(c) Third round: Feature 1 was selected as the root of the model.

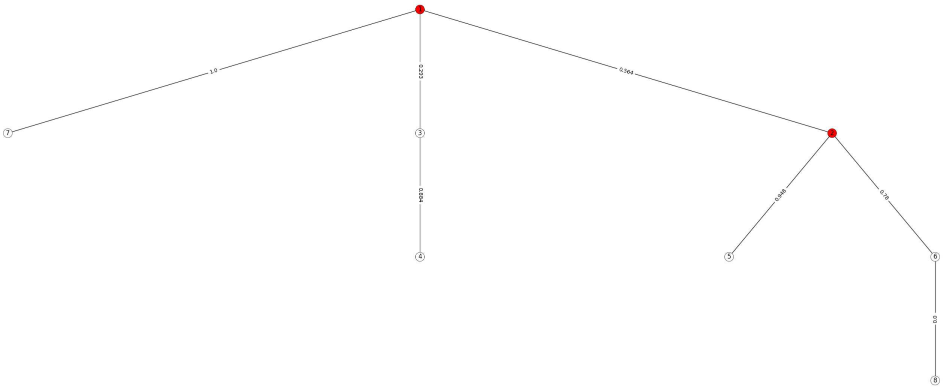
Source: Author

Figure 3a shows the first round of the assignment. Note that we will have eight pairs for eight features, of which four are redundant. This happens because we are using the symmetric variant of the AP (pairs (a,b) and (b,a) represent matches between feature a and feature b). Each of the formed pairs represents the best possible global match given the cost matrix, and as expected, the repeated variable represented by the number 7 matched with the informative variable represented by the number 1. There is a perfect degree of similarity between these two variables. Variables 3 and 4, 5 and 6 are linear compositions of variables 1 and 2, and therefore, they share a certain degree of information. This explains the matches $(3,4)$ and $(2,5)$. The match $(6,8)$ is different because variable 8 is composed of random noise. This match did not happen due to the similarity between 6 and 8 but rather to form the best combination that maximizes the global similarity of pairs. Despite this connection not representing a similarity between the variables, propagating the more relevant variable mitigates the possible loss of information.

Note that conforming figure 3a, only variables 1, 2, 3, and 6 were selected to compose the second round. This happened because they have higher relevance in their respective pairs. The final result of the second round can be seen in figure 3b, where the formed matches were $(2,6)$ and $(1,3)$, and variables 2 and 1 were propagated. The explanation for this is simple: variables 1 and 2 are independent variables used to form the output variable and, therefore contain much information regarding the output variable. Their matches with variables 6 and 3 occurred again because the latter are linear combinations of the former and thus share some information.

In the third and final round, only variables 1 and 2 remain, with only the pair $(1,2)$ left. In Figure 3c, we have the last round of assignments where feature 1 is propagated due to its higher relevance. With no more features left for a new round of assignments, the first part of the algorithm comes to an end.

Figure 4 – Hierarchical graph of features for $k = 2$. The selected nodes are in red. The edges represent the cost similarity between nodes. The unselected features are represented for the first selected node in their path to the root.



Source: Author.

In conclusion, based on the number of rounds of AP, the proposed algorithm selected the most relevant variables, which turned out to be features 1 and 2. The final structure, shown in Figure 4, represents a simplified model that captures the essential information needed for further analysis or modeling.

Starting from the root represented by feature 1, we explored the similarity relationships among nodes at lower levels. We analyzed the edges and found that some pairs like $(1,7)$ exhibited perfect similarity, allowing us to remove redundant information without compromising the model's accuracy. Additionally, the algorithm's approach of propagating the most relevant variables helped mitigate potential information loss, even when pairs with lower similarity values were present, like $(1,3)$.

Furthermore, we observed that some edges ($3-4$, $2-5$, and $2-6$) showed high linear similarity between parent and child nodes, indicating that the higher-level variable accurately represented the information of its corresponding lower-level node. This insight allowed us to remove the child nodes, further simplifying the model while retaining the necessary information.

Lastly, we identified an edge with a similarity value of 0, corresponding to variable

8, which was composed solely of random noise and lacked helpful information. As expected, we safely discarded this variable without negatively impacting the model's performance.

Overall, the algorithm's application successfully selected the most relevant variables, simplified the model's structure, and preserved crucial information, making it a valuable tool for feature selection and interpretation.

4 EXPERIMENTS AND RESULTS

This section presents the empirical evaluation of the proposed feature selection method through two experiments. The first experiment investigates the accuracy of the algorithms as the number of features increases, while the second examines the general classification accuracy.

4.1 General data sets

In order to evaluate and compare IVS with the state-of-art approaches, we carefully selected 17 datasets from the UCI Repository (ASUNCION; NEWMAN, 2007) that exhibit variations in features, sizes, and domains. These datasets are commonly used in similar studies (BROWN *et al.*, 2012; BENNASAR *et al.*, 2015).

To assess the difficulty level of the feature selection task for each dataset, we computed the example-feature ratio $\frac{N}{mc}$, where N represents the number of data points, m denotes the median arity of the features, and c signifies the number of classes. A lower value of the example-feature ratio indicates a more challenging feature selection problem (BROWN *et al.*, 2012). The table describing the selected datasets can be found in 3.

Table 3 – Data sets used in experiments. The final column indicates the difficulty level of performing feature selection, where a smaller value indicates the hardest problem.

Dataset Name	Size	Attributes	Classes	Ratio
COIL20	1440	1024	20	14
USPS	9298	256	10	186
arcene	200	10000	2	1
colon	62	2000	2	10
madelon	2600	500	2	9
lung_discrete	73	325	7	3
connectionist-bench-sonar	208	60	2	21
optical-recognition-handwritten-digits	3823	64	10	22
ozone-level-detection-one	1848	72	2	185
ozone-level-detection-eight	1847	72	2	185
libras-movement	360	90	15	5
hill-valley-noise,	606	100	2	61
hill-valley	606	100	2	61
adult,	32561	107	2	8140
mushroom,	8124	111	2	2031
horse-colic,	300	121	2	75
lung-cancer,	32	146	3	5

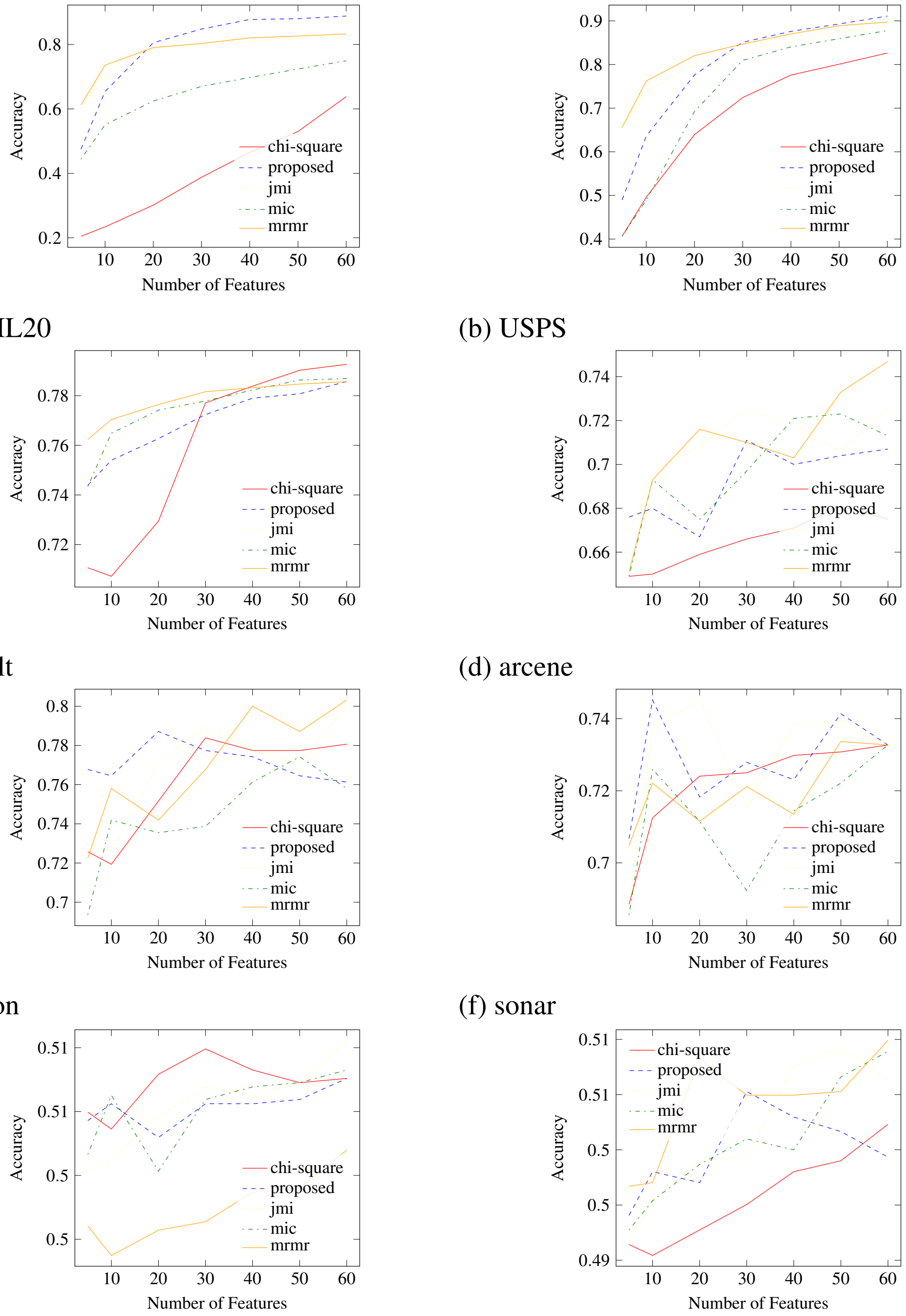
Source: Author.

Our data processing pipeline begins by applying a standardization function to all datasets. Following the approach of (BROWN *et al.*, 2012), we perform a discretization step on continuous features using an equal-width strategy with five bins.

In line with existing literature, we utilize the widely known Support Vector Machine (SVM) as the learning model for classification. We employ a 10-fold cross-validation setup with a linear kernel and set the hyperparameter C to 1. We repeat the experiment 10 times without performing any hyperparameter tuning. The primary focus of the experiment is to evaluate the feature selection effectiveness of the algorithms rather than the predictive capability of the learning model.

For the first experiment, we conduct a classification task using five well-known filter algorithms as baselines for comparison with our proposed method. We vary the number of features across different datasets, and the final results are presented in Figures 5 and 6. As expected, in most datasets, increasing the number of features also increases the accuracy, except 6a and 6e, where redundant features appear to have negligible impact. Some datasets, such as 5g, 5h, 6b, 6c, exhibit poor results for all methods, our hypothesis is this happened due to the nonlinearity of the datasets. It is worth noting that the chi-square algorithm performs poorly when a low number of features is selected in almost all datasets, even in 5c, which has the highest example ratio (indicating an easier feature selection task). We can expect this behavior since chi-square is one of the simplest baseline filter methods.

Figure 5 – Performance of all methods regarding classification accuracy and the number of features. Data sets 1-9.



(a) COIL20

(b) USPS

(c) adult

(d) arcene

(e) colon

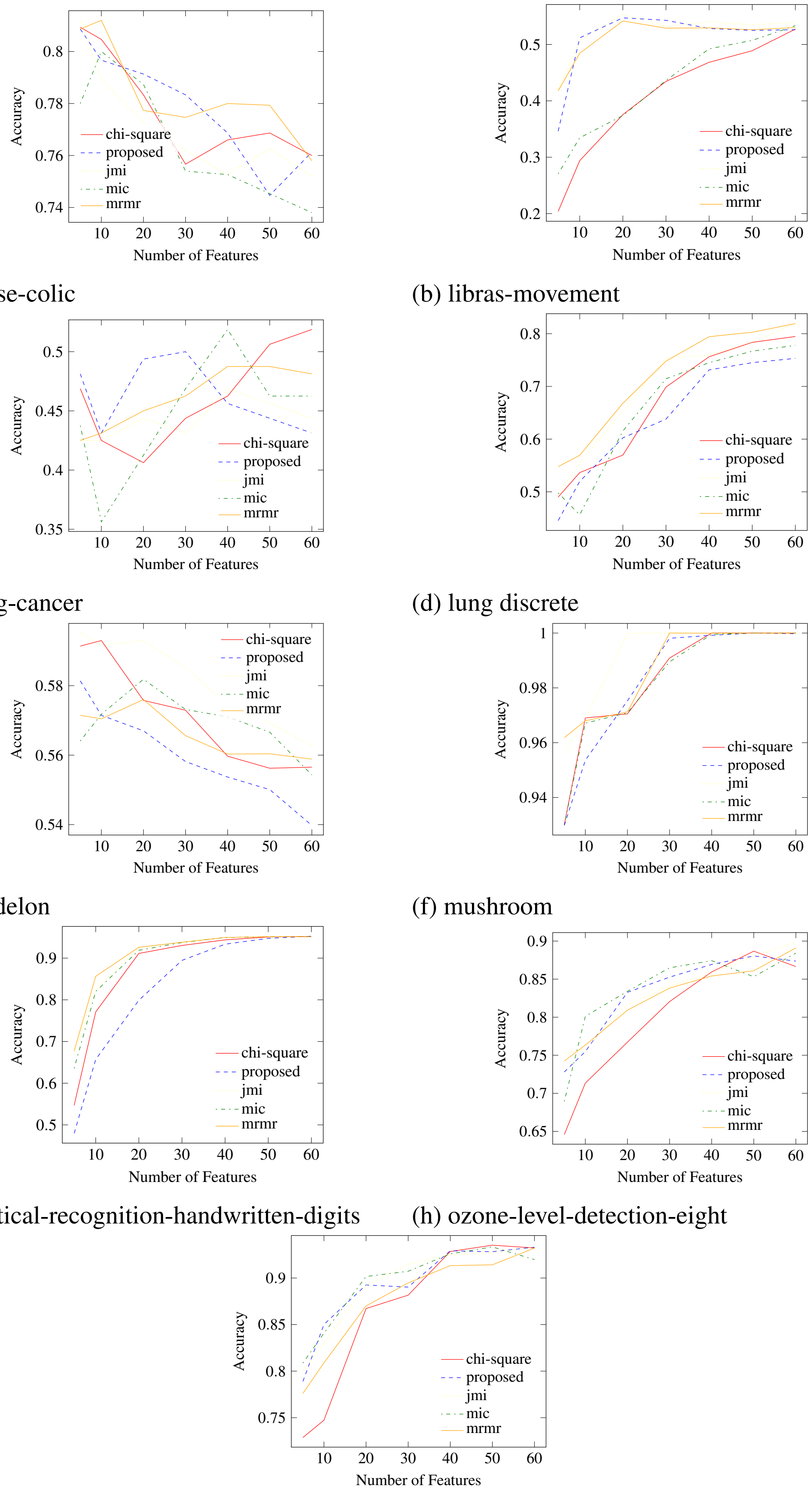
(f) sonar

(g) hill-valley

(h) hill-valley-noise

Source: Author

Figure 6 – Performance of all methods regarding classification accuracy and the number of features. Data sets 10-18.



(a) horse-colic

(b) libras-movement

(c) lung-cancer

(d) lung discrete

(e) madelon

(f) mushroom

(g) loptical-recognition-handwritten-digits

(h) ozone-level-detection-eight

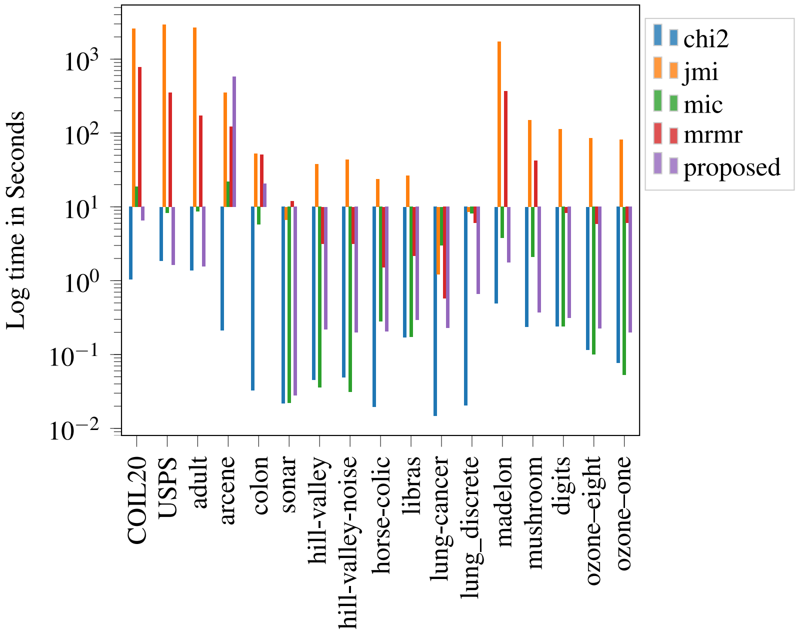
(i) ozone-level-detection-one

Source: Author

An interesting observation is that the proposed method achieves better results on 5a and 5c, where both datasets have similar example ratios and relatively large numbers of features/examples compared to the other datasets. The proposed method benefits from having more samples to measure the similarity between features accurately. High-dimensional datasets often contain more redundant features, and the proposed method's ability to aggregate similar features and propagate the most relevant ones becomes advantageous.

The mutual information-based filters (JMI, mRMR, MIM) demonstrate the best performance on the most challenging datasets, such as 5d and 6d. Unlike covariance-based methods, mutual information can capture nonlinear information between features leading to better results. However, as shown in Figure 7, these methods exhibit slower performance times compared to the proposed method, especially for JMI and mRMR. This is expected because calculating mutual information involves the computation of multiple integrals, which is computationally expensive. This suggests a trade-off between accuracy and time when comparing covariance-based and mutual information methods. Nonetheless, in the upcoming experiment, we will demonstrate that the proposed method achieves statistically similar results to JMI and mRMR.

Figure 7 – Expended time in seconds for classification experiment. As expected, the proposed method is slower than simple mic and chi2 methods but faster than JMI and mRMR. The upper part of the figure represents values greater than 1, while the inferior part represents values between 0 and 1.



Source: Author.

In the second experiment, we conducted the classification problem and examined which filter would provide better accuracy for each dataset. The goal was to determine if there was a significant difference in the accuracy of the outcomes among the methods. Table 4 provides a summary of the accuracy achieved by each filter for each dataset.

From the table, we can observe that the proposed method achieved better accuracy in 7 datasets, even though the graphs in Figures 5 and 6 show high values at certain peaks for the other methods, we can hypothesize that our approach was more effective in avoiding overfitting in these experiments.

To assess the statistical significance of the results and compare the proposed method with the method that achieved higher accuracy (beyond the proposed method), we conducted a corrected resampled t-test (BOUCKAERT; FRANK, 2004). This t-test checks the hypothesis that the difference between the outcomes of the two classifiers comes from a distribution with a mean of zero.

Table 5 provides the p-values and statistics of the t-tests comparing the proposed method with the alternative methods for each dataset. The p-values indicate the probability of observing the obtained difference in accuracy if the true difference in performance is zero. A small p-value suggests that the observed difference is statistically significant, indicating that the proposed method performs similarly to the alternative method.

By examining the table, we can conclude that for most datasets, the p-values are relatively high, indicating no significant difference in performance between the proposed method and the alternative method with higher accuracy. This finding supports the claim that the proposed method achieves statistically similar results to the baseline methods.

The results of the previous experiments indicate that the proposed method performs comparably to the baseline methods while demonstrating faster execution times, particularly in comparison to traditional methods like JMI and mRMR, for most data sets as can be seen in Figure 7. These findings suggest that the proposed method is a practical and effective alternative to existing feature selection filter algorithms in the literature.

Overall, the experiments provide evidence that the proposed method is a viable and promising approach for feature selection, offering a balance between accuracy and computational efficiency compared to traditional filter algorithms.

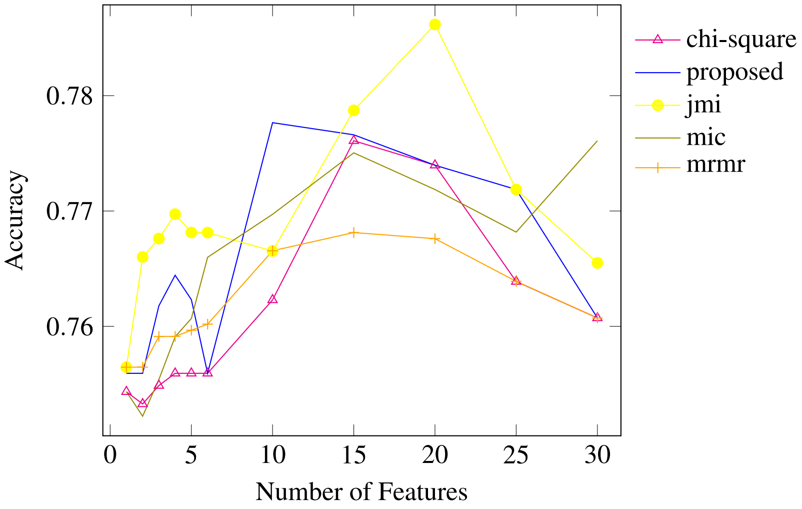
4.2 Application in real-word case.

Sudden cardiac death (SCD) is a significant adverse outcome of Chagas Disease. Some approaches have been proposed to identify clinical and laboratory features that can assist in early diagnosing high SCD propensity. Some of these approaches include the Rassi score (JR *et al.*, 2006), multivariate analysis (SOUZA *et al.*, 2015), and Heart Rate Variability (HRV) analysis (ALBERTO *et al.*, 2017; ALBERTO *et al.*, 2020).

In this study, we aim to reduce the number of variables in existing systems, which can help lower the costs of exams or minimize the need for invasive procedures. Identifying a subset of relevant features can also contribute to the early diagnosis of high SCD propensity in individuals with Chagas Disease. This process can help streamline and optimize the diagnostic process, potentially enhancing outcomes and resource utilization in the clinical setting. The data set used in this study is the same as the one used in (CARVALHO *et al.*, 2019), and Table 6 summarizes of the features included in the dataset.

This section will demonstrate the proposed method's usability while comparing it with the mentioned baseline approaches. We will conduct a classification + feature selection experiment to identify patients with a record of SCD; We performed a similar experiment in the previous section using SVM with $10X10$-fold cross-validation, linear kernel, and the hyper-parameter $C = 1$. Figure 8 shows the accuracy comparison between the baseline filter and the proposed work, varying the number of features.

Figure 8 – Performance of all methods regarding classification accuracy and the number of features for Chagas Dataset.



Source: Author

The Chagas Disease data set analysis reveals interesting findings regarding the accuracy of different feature selection methods. The accuracy of the methods generally shows an increase in the range of 1-20 features, followed by a decrease in accuracy in the range of 20-30 features. This pattern suggests the presence of linear redundancy among the features.

Among the methods evaluated, JMI with $k = 20$, the proposed method with $k = 10$, and MIM with $k = 30$ achieved the highest accuracy results, with slight differences. Specifically, JMI obtained an accuracy of 0.786, the proposed method achieved 0.778, and MIM also achieved 0.778. While the differences in accuracy are minimal, the proposed method stands out for its ability to significantly reduce the number of features compared to JMI and MIM.

A classification experiment was conducted by tuning the number of features to further investigate the findings. The classification results are summarized in Table 7. Interestingly, the MIM filter, which is the simplest method among the mutual information (MI) family, achieved the highest accuracy. This result could be attributed to the greedy search criteria used by the other methods, which may make them more susceptible to overfitting.

To determine the significance of the results, a corrected t-test was performed between MIM and the proposed method. The test resulted in a p-value of 0.8171 and a statistic of 0.2381, which indicates that the null hypothesis is accepted. This means that there is no significant difference between the accuracy of the two methods.

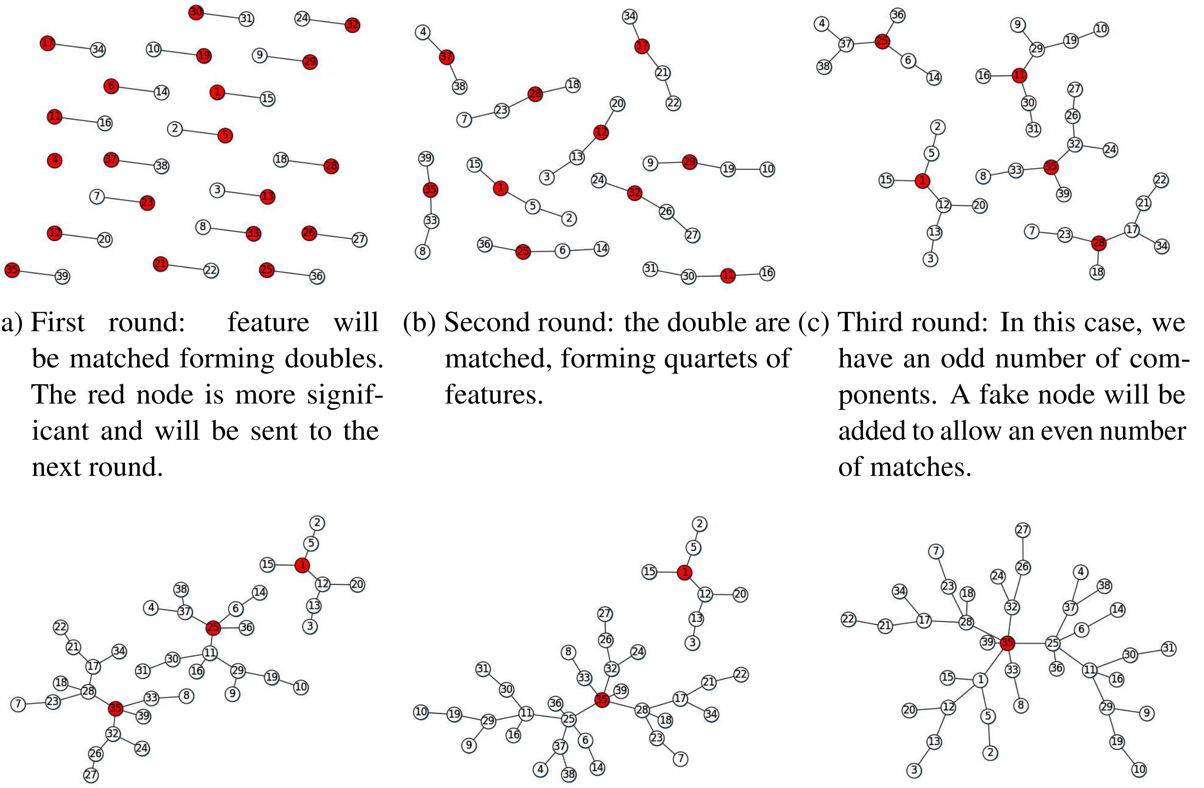
Overall, these findings suggest that despite its simplicity, our approach achieves comparable accuracy to other feature selection methods while significantly reducing the number of features required, being valuable in optimizing costs and the need for invasive procedures to diagnose high SCD propensity in Chagas Disease.

Despite the comparable accuracies to other filters, IVS also offers a high degree of interpretation. As explained in Section 3.2, the assignment iterations can be used to construct a tree representing the feature selection process. The process is visualized in Figure 9, and the final tree for $k = 10$ is shown in Figure 10.

In the tree, the selected features are represented by nodes in red. By disconnecting the parent edges of the red nodes, we obtain k disconnected graph components, each taking the form of a tree. These components can be interpreted as clusters of features formed through pairwise similarity (e.g., Feature A matches with B, B matches with C, and so on). The root node of each cluster represents the entire cluster. This hierarchical structure explains the selection or removal of each feature.

Examining the tree structure clarifies why each feature was selected or discarded during the selection process. This interpretability aspect of the proposed method adds value and enhances the understanding of the underlying relationships between features.

Figure 9 – Aggregation rounds of the proposed method.



(a) First round: feature will be matched forming doubles. The red node is more significant and will be sent to the next round.

(b) Second round: the double are matched, forming quartets of features.

(c) Third round: In this case, we have an odd number of components. A fake node will be added to allow an even number of matches.

(d) Fourth round: the component with root 0 was matched with the fake node, so no aggregation was performed. Still, we move this component to the next round.

(e) Fifth round: again, the component with root 0 was assigned to the fake node. We repeat the process.

(f) Sixth round: the final graph was built. Node number 31 is the most significant node across all other nodes.

Source: Author

To illustrate the relationship between selected and non-selected features, we present a table summarizing the similarity of the features (Table 8). On the left side of the table, we can see the selected feature names, their relevance, and the number of iterations in which the feature participated. On the right side, we list the removed features, their relevance, the number of iterations, and their similarity with the selected feature. The relevance column indicates the importance of the variable according to the similarity function mentioned in 3.2, representing the similarity between the feature and the target label. The iteration column represents the number of times the features were propagated for another round of the assignment problem. The selected features are chosen based on their number of iterations and relevance, with a higher number of iterations indicating greater importance. The similarity column represents the similarity between the selected and non-selected features.

Firstly, it's important to note that all selected features have more significance than their corresponding matches. This is expected since the proposed method propagates highly relevant features while removing those with lower relevance. The most significant variable selected, serving as the model's root, is **Count Premature Ventricular Contraction (PVC)**'s, which represents the number of PVCs for 24 hours. PVCs are considered triggering factors for SCD according to the classic biological model of sudden death (ROBERT *et al.*, 1989). This triggering factor may also apply to chronic Chagas' heart disease (JR *et al.*, 2001). However, **PVC**'s was connected to **Syncope**, resulting in the removal of this variable from the model. The literature supports this similarity, which indicates that progressive PVCs induced by exercise or stress can cause syncope or sudden death (PEREZ-SILVA; MERINO, 2011).

Another important finding is that the second most important variable is **Gender**. This aligns with the studies presented on (KEEGAN *et al.*, 2020), which suggest a strong predisposition for SCD in Chagas disease among males. The literature also shows that variables such as **Systemic Arterial Hypertension**, **Type 2 Diabetes Mellitus**, and **Sedentary Lifestyle** are influenced by gender (KAUTZKY-WILLER *et al.*, 2016; BRUNO *et al.*, 2016). In this context, it is plausible to hypothesize that the gender variable encapsulates the information of the other mentioned clinical variables in the classification of SCD in Chagas disease. An alternative hypothesis worth considering is that the observed differences may stem from survey selection and information biases. It is possible that men, if they have underlying health conditions or disabilities, might be hesitant to engage in survey participation or disclose pertinent health-related information. (OKSUZYAN *et al.*, 2010).

Other relevant features, such as **Left ventricular ejection fraction (LVEF)**, are also associated with SCD in Chagas disease (KEEGAN *et al.*, 2020). It's worth noting that frequent idiopathic PVCs can result in a reduced LVEF (BAMAN *et al.*, 2010). However, as described in Tables 11 and 10, the relation between PVCs and LVEF is weaker than between PVCs and syncope. Consequently, some related variables may appear in different branches as other strong relations suppress them. Despite this, it's also possible to note trivial relations, such as **Cardiac insufficiency** and **Average Heart Rate**.

The literature supports the findings presented by our method. However, it is important to acknowledge that important variables may be excluded from the model due to their similarity with the selected ones. This occurs because there is a threshold between the relevance and redundancy of features, and sometimes it is reasonable to retain certain variables even if they are redundant. An example is the **Syncope** variable, well-known as a factor for sudden cardiac death in Chagas disease (KEEGAN *et al.*, 2020). In such cases, incorporating specialist knowledge into the model can be beneficial. Constraints can be applied to the assignment connections by removing edges between specific nodes, a variant known as the constrained assignment problem (described in Section 2.2.1).

To explore this, we replicated the experiment by removing the connection between **Syncope** and the variables in its path according to Figure 10, namely **Count PVC's** and **Non-Sustained Ventricular Tachycardia (Holter)**. We also removed the connections between **Left ventricular ejection fraction (LVEF)** and **Count PVC's**, and **Syncope**. The rationale behind this modification was to disregard the similarity of well-known important features for sudden cardiac death in Chagas disease. By doing so, we compelled the algorithm to explore alternative connection paths between variables while propagating relevant variables in different branches. Additionally, instead of using the relevance value of **Syncope** calculated by Equation 3.2, we assigned it the same value as **Count PVC's**. This manual adjustment incorporated more importance to the **Syncope** variable, ensuring its selection in the final model.

We observed no significant changes in terms of accuracy compared to the previous version. The accuracy obtained was 0.781 ± 0.018 compared to $0.772 \pm .025$ in the previous experiment, with a p-value of 0.7348 and a statistic of 0.3495, indicating statistical similarity between the two experiments. However, in medical applications, understanding the outcome for classification is as important as the correctness of the result. We argue that determining which clinical variables should be included in the model without compromising interpretability or accuracy is a valuable toolset.

The interpretation output of this experiment is presented in Table 9. The variable **LVEF** had some connections modified. While its strong connection with **Diastolic Dysfunction** was preserved, other variables such as **Left Ventricular Diastolic Diameter (LVDD)** and **Left Ventricular Systolic Diameter (LVSD)** are now associated with **Syncope**, which is now part of the selected feature set. This behavior is expected as the high linear relationships between **LVEF** and **Syncope** can also be shared with **LVDD** and **LVSD**. Moreover, there is a high similarity between the feature **Syncope** and **Amiodarone**, a medication that significantly increases the risk of fall-related injuries and syncope (DALGAARD *et al.*, 2019). We can observe similar changes in connections with **Average Heart Rate**, where **Cardiac insufficiency** is now associated with **Pacemaker** instead, and the former is associated with **Other Heart Diseases**. These alterations in connections are also expected since many features share information due to being based on similar exams. However, the root of the model remains the same, and most of the branches are fully preserved, such as **Gender**, **Atrial Fibrillation/Flutter (Holter)**, **Implantable Cardioverter Defibrillator**, and **Premature Ventricular Contraction**.

As evident, IVS facilitates users to seamlessly integrate specialized knowledge through minor modifications. This process guarantees the discernment of crucial variables and the establishment of fresh interpretative correlations, all while minimizing substantial alterations to the preexisting structure.

Table 4 – Performance of proposed method compared with baseline methods regarding classification accuracy and time in milliseconds.

Dataset	chi2 (acc)±(std)	chi2 (time)	JMI (acc)±(std)	JMI (time)	mic (acc)±(std)	mic (time)	mRMR (acc)±(std)	mRMR (time)	proposed (acc)±(std)	proposed (time)
COIL20	0.638±0.064	0.001	0.828±0.022	2.591	0.750±0.014	0.019	0.833±0.013	0.780	**0.888±0.017**	0.007
USPS	0.826±0.004	0.002	0.889±0.003	2.941	0.878±0.008	0.008	0.897±0.004	0.350	**0.911±0.003**	0.002
adult	0.793±0.004	0.001	**0.795±0.005**	2.683	0.791±0.004	0.009	0.787±0.004	0.172	0.785±0.006	0.002
arcene	0.676±0.050	0.000	0.713±0.043	0.354	0.705±0.031	0.022	**0.728±0.048**	0.123	0.696±0.037	0.582
colon	0.748±0.068	0.000	0.742±0.059	0.053	0.735±0.062	0.006	0.758±0.055	0.051	**0.768±0.076**	0.021
connectionist bench sonar	**0.733±0.050**	0.000	0.726±0.044	0.007	0.728±0.042	0.000	0.727±0.052	0.012	**0.733±0.048**	0.000
hill valley	0.509±0.013	0.000	**0.509±0.012**	0.038	0.505±0.017	0.000	0.500±0.011	0.003	0.507±0.012	0.000
hill valley noise	0.502±0.009	0.000	0.506±0.014	0.044	0.512±0.018	0.000	**0.517±0.019**	0.003	0.504±0.009	0.000
horse colic	0.759±0.054	0.000	0.754±0.048	0.024	0.739±0.041	0.000	0.761±0.038	0.002	**0.766±0.041**	0.000
libras movement	0.527±0.036	0.000	0.531±0.033	0.026	**0.534±0.036**	0.000	0.529±0.045	0.002	0.532±0.022	0.000
lung cancer	0.406±0.085	0.000	0.412±0.084	0.001	0.438±0.118	0.003	0.438±0.106	0.001	**0.450±0.117**	0.000
lung discrete	0.567±0.073	0.000	0.569±0.115	0.009	0.558±0.130	0.008	**0.591±0.082**	0.006	0.572±0.094	0.001
madelon	0.556±0.009	0.000	0.565±0.017	1.727	**0.566±0.014**	0.004	0.559±0.010	0.369	0.549±0.013	0.002
mushroom	**1.000±0.000**	0.000	**1.000±0.000**	0.150	**1.000±0.000**	0.002	**1.000±0.000**	0.042	**1.000±0.000**	0.000
optical recognition handwritten digits	0.951±0.006	0.000	0.951±0.006	0.113	0.951±0.006	0.000	0.951±0.007	0.008	**0.952±0.006**	0.000
ozone level detection eight	0.891±0.012	0.000	**0.899±0.014**	0.086	0.897±0.012	0.000	0.891±0.010	0.006	0.896±0.008	0.000
ozone level detection one	0.936±0.008	0.000	0.934±0.011	0.082	**0.940±0.009**	0.000	0.932±0.012	0.006	0.936±0.009	0.000

Source: the author.

Table 5 – Statistical hypothesis to verify accuracy differences for every dataset compared to the proposed approach.

dataset	statistic	p-value
connectionist-bench-sonar	0.1262	0.9024
optical-recognition-handwritten-digits	0.4451	0.6667
ozone-level-detection-one	0.2974	0.7729
ozone-level-detection-eight	0.1609	0.8757
libras-movement	0.0959	0.9257
hill-valley-noise	0.8550	0.4147
hill-valley	0.2109	0.8376
adult	0.9686	0.3580
mushroom	0.0000	1.0000
horse-colic	0.1517	0.8828
lung-cancer	0.0667	0.9483
USPS	2.0878	0.0664
arcene	0.5104	0.6220
colon	0.1011	0.9217
COIL20	2.0066	0.0757
$lung_discrete$	0.1627	0.8743
madelon	1.2931	0.2282

Source: Author

Table 6 – Chagas data set description.

Attributes Group	Feature Number	Variables	Type
Personal Data	1	Gender	Categorical
	2	Body Mass Index	Quantitative
Clinical History	3	Cancer	Categorical
	4	Systemic Arterial Hypertension	Categorical
	5	Type 2 Diabetes Mellitus	Categorical
	6	Other Heart Diseases	Categorical
	7	Pacemaker	Categorical
	8	Syncope	Categorical
	9	Atrial Fibrillation/Flutter	Categorical
	10	Chronic Kidney Failure	Categorical
	11	Cardiac insufficiency	Categorical
	12	Ventriculoperitoneal Shunt	Categorical
	13	Tabagism	Categorical
	14	Alcoholism	Categorical
	15	Sedentary Lifestyle	Categorical
ECG	16	Inactive Electrical Area	Categorical
	17	Ventricular Extrasystole	Categorical
	18	Supraventricular Extrasystole	Categorical
	19	Non-Sustained Ventricular Tachycardia	Categorical
	20	Pause $>$ 3s	Categorical
	21	Primary Change	Categorical
	22	Interventricular Conduction Disturbance	Categorical
	23	Atrioventricular Conduction Disturbance	Categorical
ECO	24	Diastolic Dysfunction	Categorical
	25	Left Atrial Diameter	Quantitative
	26	Left Ventricular Diastolic Diameter	Quantitative
	27	Left Ventricular Systolic Diameter	Quantitative
	28	Left ventricular ejection fraction (LVEF)	Quantitative
	29	Segmental Deficit	Categorical
Holter	30	Atrial Fibrillation/Flutter	Categorical
	31	Average Heart Rate	Quantitative
	32	Sinus Node Dysfunction	Categorical
	34	Non-Sustained Ventricular Tachycardia	Categorical
	35	Premature Ventricular Contraction (PVC)	Categorical
	36	Count PVS‘s	Quantitative
	37	Atrioventricular Conduction Disturbance	Categorical
Medicine	38	Implantable Cardioverter Defibrillator	Categorical
	39	Ablations	Categorical
	40	Amiodarone	Categorical

Source: Author.

Table 7 – Chagas classification Experiment.

Dataset	chi2 (acc)±(std)	chi2 (time)	JMI (acc)±(std)	JMI (time)	mic (acc)±(std)	mic (time)	mRMR (acc)±(std)	mRMR (time)	proposed (acc)±(std)	proposed (time)
chagas	0.767±0.017	0.000	0.765±0.019	0.003	**0.777±0.021**	0.000	0.756±0.022	0.000	0.772±0.025	0.000

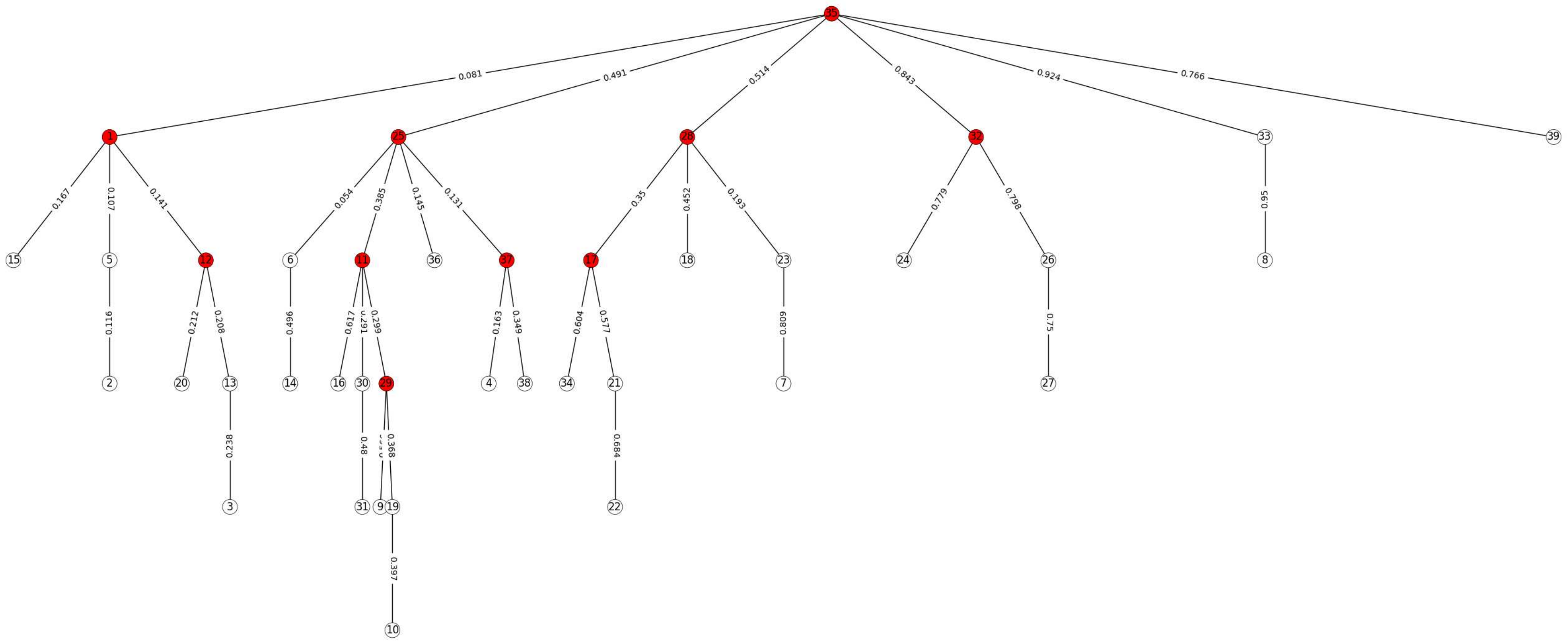
Source: the author

Table 8 – Linear table relation.

Selected Feature	Relevance	Round	Non-selected Feature/Replaced Feature	Relevance	Round	Similarity
Gender	0.1473	5	Body Mass Index	0.0885	0	0.0324
	0.1473	5	Type 2 Diabetes Mellitus	0.1049	1	0.1073
	0.1473	5	Sedentary Lifestyle	0.0035	0	0.1674
Cardiac insufficiency	0.1988	3	Inactive Electrical Area	0.1718	0	0.6168
	0.1988	3	Average Heart Rate	0.0014	1	0.2906
	0.1988	3	Sinus Node Dysfunction	0.0013	0	0.2823
Ventriculoperitoneal Shunt	0.1102	2	Cancer	0.0597	0	0.016
	0.1102	2	Tabagism	0.0981	1	0.2078
	0.1102	2	Pause $>$ 3s	0.0896	0	0.2125
Premature Ventricular Contraction	0.2028	2	Primary Change	0.0964	1	0.5773
	0.2028	2	Interventricular Conduction Disturbance	0.0837	0	0.5586
	0.2028	2	Premature Ventricular Contraction (Holter)	0.189	0	0.6043
Left Atrial Diameter	0.2976	4	Other Heart Diseases	0.1806	1	0.054
	0.2976	4	Alcoholism	0.0896	0	0.1142
	0.2976	4	Atrioventricular Conduction Disturbance(Holter)	0.0034	0	0.145
Segmental Deficit	0.3771	3	Pacemaker	0.0203	0	0.1829
	0.3771	3	Supraventricular Extrasystole	0.1415	0	0.452
	0.3771	3	Atrioventricular Conduction Disturbance	0.022	1	0.1933
Atrial Fibrillation/Flutter(Holter)	0.1029	2	Atrial Fibrillation/Flutter	0.052	0	0.5661
	0.1029	2	Chronic Kidney Failure	0.0259	0	0.1256
	0.1029	2	Non-Sustained Ventricular Tachycardia	0.0843	1	0.3675
Count PVC’s	0.7257	6	Syncope	0.6715	0	0.9312
	0.7257	6	Non-Sustained Ventricular Tachycardia(Holter)	0.7087	1	0.9243
	0.7257	6	Amiodarone	0.461	0	0.7661
Implantable Cardioverter Defibrillator	0.1001	2	Systemic Arterial Hypertension	0.091	1	0.1626
	0.1001	2	Ablations	0.0044	0	0.3489
Left ventricular ejection fraction (LVEF)	0.5648	2	Diastolic Dysfunction	0.4254	0	0.7793
	0.5648	2	Left Ventricular Diastolic Diameter	0.5126	1	0.798
	0.5648	2	Left Ventricular Systolic Diameter	0.4028	0	0.7662

Source: Author.

Figure 10 – Hierarchical graph of features for $k = 10$. The selected nodes are in red. The edges represent the cost similarity between nodes. The unselected features are represented for the first selected node in their path to the root. Any dummy nodes are removed after the final round of assignments.



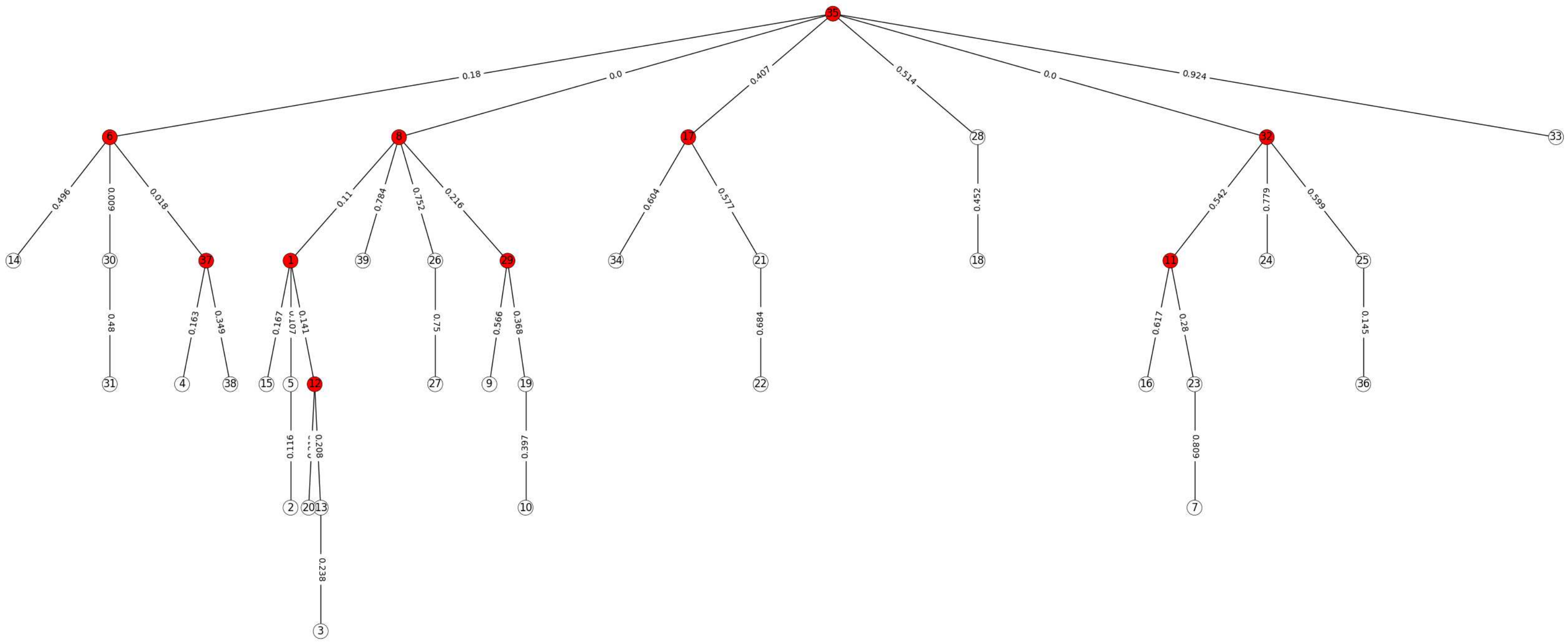
Source: Author.

Table 9 – Linear table relation with specialist knowledge.

Selected Feature	Relevance	Round	Non-selected Feature/Replaced Feature	Relevance	Round	Similarity
Gender	0.1473	3	Body Mass Index	0.0885	0	0.0324
	0.1473	3	Type 2 Diabetes Mellitus	0.1049	1	0.1073
	0.1473	3	Sedentary Lifestyle	0.0035	0	0.1674
Other Heart Diseases	0.1806	3	Alcoholism	0.0896	0	0.4962
	0.1806	3	Average Heart Rate	0.0014	1	0.0086
	0.1806	3	Sinus Node Dysfunction	0.0013	0	0.0457
Syncope	0.6715	4	Left Ventricular Diastolic Diameter	0.5126	1	0.752
	0.6715	4	Left Ventricular Systolic Diameter	0.4028	0	0.6714
	0.6715	4	Amiodarone	0.461	0	0.7844
Cardiac insufficiency	0.1988	2	Pacemaker	0.0203	0	0.2679
	0.1988	2	Inactive Electrical Area	0.1718	0	0.6168
	0.1988	2	Atrioventricular Conduction Disturbance	0.022	1	0.2805
Ventriculoperitoneal Shunt	0.1102	2	Cancer	0.0597	0	0.016
	0.1102	2	Tabagism	0.0981	1	0.2078
	0.1102	2	Pause $>$ 3s	0.0896	0	0.2125
Premature Ventricular Contraction	0.2028	2	Primary Change	0.0964	1	0.5773
	0.2028	2	Interventricular Conduction Disturbance	0.0837	0	0.5586
	0.2028	2	Premature Ventricular Contraction (Holter)	0.189	0	0.6043
Atrial Fibrillation/Flutter(Holter)	0.1029	2	Atrial Fibrillation/Flutter	0.052	0	0.5661
	0.1029	2	Chronic Kidney Failure	0.0259	0	0.1256
	0.1029	2	Non-Sustained Ventricular Tachycardia	0.0843	1	0.3675
Count PVC’s	0.7257	6	Supraventricular Extrasystole	0.1415	0	0.3331
	0.7257	6	Segmental Deficit	0.3771	1	0.5137
	0.7257	6	Non-Sustained Ventricular Tachycardia(Holter)	0.7087	0	0.9243
Implantable Cardioverter Defibrillator	0.1001	2	Systemic Arterial Hypertension	0.091	1	0.1626
	0.1001	2	Ablations	0.0044	0	0.3489
Left ventricular ejection fraction (LVEF)	0.5648	5	Diastolic Dysfunction	0.4254	0	0.7793
	0.5648	5	Left Atrial Diameter	0.2976	1	0.5987
	0.5648	5	Atrioventricular Conduction Disturbance(Holter)	0.0034	0	0.1583

Source: the author.

Figure 11 – Hierarchical graph of features for $k = 10$ after the insertion of knowledge specialist. As expected, most low-level branch structure was preserved. Only the variables with a high number of iterations changed.



Source: Author.

Table 10 – Matrix of values. Columns 20-39. The data set description can be consulted in the table.

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
1	0.0745	0.1726	0.0307	0.0609	0.0128	0.0774	0.0379	0.0181	0.0471	0.0799	0.1046	0.0712	0.1237	0.0429	0.0810	0.0194	0.0564	0.1669	0.0542
2	0.1626	0.0675	0.0854	0.1471	0.1246	0.0925	0.0545	0.1282	0.0025	0.0267	0.0129	0.1505	0.0898	0.0173	0.1055	0.1092	0.0842	0.0314	0.1202
3	0.0158	0.0355	0.0514	0.0608	0.1748	0.0912	0.1056	0.0055	0.0569	0.1174	0.1133	0.1166	0.0442	0.1229	0.0802	0.0183	0.0256	0.0476	0.0432
4	0.0462	0.0115	0.0536	0.0792	0.1029	0.0414	0.0720	0.0910	0.0218	0.0285	0.0417	0.0993	0.0913	0.1163	0.1196	0.0977	0.1626	0.0780	0.1101
5	0.0549	0.0067	0.0301	0.0161	0.0533	0.0928	0.0910	0.0829	0.1361	0.0706	0.0383	0.0621	0.1014	0.0369	0.0778	0.0729	0.0150	0.0408	0.1123
6	0.0111	0.0020	0.0151	0.0697	0.0540	0.0863	0.1239	0.0806	0.0398	0.0086	0.0457	0.1009	0.1492	0.0817	0.1803	0.0366	0.0179	0.0333	0.1206
7	0.1726	0.1428	0.8091	0.2611	0.1857	0.1107	0.1158	0.1829	0.1104	0.0467	0.1351	0.1638	0.0632	0.0252	0.0557	0.5671	0.0998	0.0028	0.0845
8	0.2237	0.2794	0.1619	0.6757	0.4839	0.7520	0.6714	0.4762	0.2163	0.1587	0.1269	0.8326	0.9496	0.3180	0.9312	0.1543	0.0498	0.2007	0.7844
9	0.1222	0.1502	0.0108	0.2203	0.1433	0.1266	0.1609	0.1294	0.5661	0.2051	0.2477	0.2134	0.1933	0.0774	0.1846	0.0364	0.0288	0.0535	0.1211
10	0.1697	0.2017	0.2987	0.3995	0.2809	0.4136	0.4922	0.2183	0.1256	0.1911	0.2407	0.3671	0.2240	0.1514	0.2314	0.3257	0.1964	0.0465	0.2796
11	0.2370	0.3848	0.2805	0.5743	0.3846	0.5488	0.5766	0.2511	0.2992	0.2906	0.2823	0.5423	0.4104	0.2229	0.3915	0.2168	0.0552	0.1440	0.3545
12	0.0556	0.0425	0.0624	0.1279	0.0712	0.0788	0.1415	0.0699	0.0379	0.0789	0.0684	0.1228	0.0792	0.0342	0.0730	0.0855	0.0418	0.0012	0.1055
13	0.0313	0.1049	0.0758	0.1138	0.1407	0.1351	0.1588	0.0856	0.1604	0.0021	0.0343	0.1297	0.1304	0.0188	0.1269	0.0829	0.0239	0.0444	0.0834
14	0.1352	0.0924	0.0282	0.0211	0.1142	0.0987	0.1204	0.0400	0.0198	0.0409	0.0227	0.0429	0.0740	0.0405	0.1055	0.0309	0.0089	0.0165	0.0973
15	0.1326	0.1427	0.3506	0.2789	0.2060	0.3028	0.2456	0.1936	0.1610	0.1852	0.2253	0.1702	0.1128	0.1546	0.1268	0.2475	0.0604	0.0344	0.1631
16	0.3206	0.3786	0.2255	0.4993	0.3163	0.3965	0.3806	0.3396	0.2332	0.2806	0.2843	0.4573	0.3079	0.2643	0.3230	0.1779	0.1115	0.1726	0.2744
17	0.5773	0.5586	0.0786	0.4445	0.3149	0.4076	0.3928	0.3505	0.2239	0.1346	0.1487	0.4760	0.3867	0.6043	0.4074	0.0933	0.1007	0.1399	0.3551
18	0.6324	0.5259	0.1982	0.3973	0.2466	0.3193	0.3042	0.4520	0.2311	0.2640	0.1934	0.4288	0.3316	0.4149	0.3331	0.1628	0.1039	0.1932	0.3108
19	0.1517	0.1292	0.0866	0.2934	0.2717	0.2369	0.3365	0.1607	0.3675	0.2564	0.4404	0.2868	0.2209	0.1629	0.2345	0.0862	0.0357	0.1080	0.2305
20	0.0377	0.0677	0.2205	0.2018	0.2308	0.0987	0.1989	0.0400	0.2582	0.1398	0.2252	0.1340	0.0740	0.0405	0.0571	0.1653	0.0089	0.0165	0.0973
21	0.0964	0.6835	0.2086	0.4247	0.2583	0.2438	0.1902	0.3674	0.1462	0.1952	0.1676	0.3067	0.2148	0.4157	0.2058	0.2283	0.0657	0.1222	0.2334
22	0.6835	0.0837	0.1586	0.4777	0.2853	0.2641	0.3075	0.2702	0.2033	0.3005	0.1954	0.3779	0.2472	0.3454	0.2766	0.2015	0.0560	0.2020	0.2908
23	0.2086	0.1586	0.0220	0.2571	0.1889	0.1276	0.1600	0.1933	0.0743	0.0008	0.1611	0.1721	0.1143	0.0731	0.1015	0.7517	0.0471	0.0108	0.1548
24	0.4247	0.4777	0.2571	0.4254	0.5505	0.6452	0.6898	0.4133	0.2265	0.2087	0.1811	0.7793	0.6868	0.3277	0.6553	0.2347	0.0495	0.1722	0.5798
25	0.2583	0.2853	0.1889	0.5505	0.2976	0.6074	0.5842	0.2713	0.1843	0.1052	0.1333	0.5987	0.4655	0.2312	0.4908	0.1450	0.1312	0.1433	0.4968
26	0.2438	0.2641	0.1276	0.6452	0.6074	0.5126	0.7499	0.4561	0.1967	0.1374	0.1258	0.7980	0.7400	0.3256	0.7391	0.1173	0.0421	0.1989	0.6703
27	0.1902	0.3075	0.1600	0.6898	0.5842	0.7499	0.4028	0.3266	0.1838	0.1761	0.1664	0.7662	0.6548	0.3174	0.6696	0.1459	0.1184	0.2117	0.6407
28	0.3674	0.2702	0.1933	0.4133	0.2713	0.4561	0.3266	0.3771	0.1548	0.1121	0.1362	0.5102	0.4925	0.4036	0.5137	0.1765	0.0696	0.1294	0.4110
29	0.1462	0.2033	0.0743	0.2265	0.1843	0.1967	0.1838	0.1548	0.1029	0.2760	0.3923	0.2802	0.2075	0.1570	0.2087	0.0110	0.0344	0.0640	0.2111
30	0.1952	0.3005	0.0008	0.2087	0.1052	0.1374	0.1761	0.1121	0.2760	0.0014	0.4801	0.1901	0.1687	0.1181	0.1857	0.1458	0.0471	0.0020	0.2834
31	0.1676	0.1954	0.1611	0.1811	0.1333	0.1258	0.1664	0.1362	0.3923	0.4801	0.0013	0.1892	0.1170	0.1389	0.1408	0.2500	0.0394	0.0733	0.1365
32	0.3067	0.3779	0.1721	0.7793	0.5987	0.7980	0.7662	0.5102	0.2802	0.1901	0.1892	0.5648	0.8421	0.3668	0.8431	0.1583	0.0914	0.2134	0.6995
33	0.2148	0.2472	0.1143	0.6868	0.4655	0.7400	0.6548	0.4925	0.2075	0.1687	0.1170	0.8421	0.7087	0.3111	0.9243	0.1130	0.0459	0.1933	0.7608
34	0.4157	0.3454	0.0731	0.3277	0.2312	0.3256	0.3174	0.4036	0.1570	0.1181	0.1389	0.3668	0.3111	0.1890	0.3588	0.0829	0.0706	0.1313	0.2680
35	0.2058	0.2766	0.1015	0.6553	0.4908	0.7391	0.6696	0.5137	0.2087	0.1857	0.1408	0.8431	0.9243	0.3588	0.7257	0.1202	0.0364	0.1739	0.7661
36	0.2283	0.2015	0.7517	0.2347	0.1450	0.1173	0.1459	0.1765	0.0110	0.1458	0.2500	0.1583	0.1130	0.0829	0.1202	0.0034	0.0601	0.0271	0.2087
37	0.0657	0.0560	0.0471	0.0495	0.1312	0.0421	0.1184	0.0696	0.0344	0.0471	0.0394	0.0914	0.0459	0.0706	0.0364	0.0601	0.1001	0.3489	0.0825
38	0.1222	0.2020	0.0108	0.1722	0.1433	0.1989	0.2117	0.1294	0.0640	0.0020	0.0733	0.2134	0.1933	0.1313	0.1739	0.0271	0.3489	0.0044	0.1695
39	0.2334	0.2908	0.1548	0.5798	0.4968	0.6703	0.6407	0.4110	0.2111	0.2834	0.1365	0.6995	0.7608	0.2680	0.7661	0.2087	0.0825	0.1695	0.4610

Source: Author.

Table 11 – Matrix of values. Columns 01-20. The data set description can be consulted in the table.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.1473	0.0324	0.0276	0.0545	0.1073	0.0894	0.0482	0.1100	0.0660	0.0156	0.0157	0.1411	0.1606	0.0803	0.1674	0.0094	0.0534	0.0463	0.0145	0.0803
2	0.0324	0.0885	0.0158	0.1127	0.1162	0.0222	0.1029	0.0843	0.0028	0.0630	0.0630	0.0848	0.0489	0.0572	0.1439	0.1222	0.0228	0.0436	0.0249	0.0220
3	0.0276	0.0158	0.0597	0.0259	0.0134	0.0296	0.0672	0.0895	0.0476	0.0717	0.0234	0.0160	0.2383	0.0147	0.0826	0.0492	0.0438	0.0371	0.0591	0.0147
4	0.0545	0.1127	0.0259	0.0910	0.0746	0.0658	0.0328	0.1044	0.0168	0.0193	0.0469	0.1090	0.0877	0.0326	0.0407	0.0348	0.0402	0.0267	0.0298	0.0326
5	0.1073	0.1162	0.0134	0.0746	0.1049	0.0896	0.0267	0.0979	0.1138	0.0421	0.0541	0.0219	0.0583	0.0445	0.0354	0.0459	0.0052	0.0319	0.0507	0.1148
6	0.0894	0.0222	0.0296	0.0658	0.0896	0.1806	0.0470	0.1538	0.0333	0.1565	0.0646	0.0708	0.0276	0.4962	0.0942	0.0062	0.0430	0.0472	0.0413	0.0103
7	0.0482	0.1029	0.0672	0.0328	0.0267	0.0470	0.0203	0.1078	0.0028	0.2790	0.2679	0.0610	0.0627	0.0233	0.3442	0.3059	0.0184	0.1682	0.0359	0.2186
8	0.1100	0.0843	0.0895	0.1044	0.0979	0.1538	0.1078	0.6715	0.1543	0.2362	0.4247	0.0896	0.1247	0.0763	0.1393	0.3040	0.3834	0.3480	0.2300	0.0763
9	0.0660	0.0028	0.0476	0.0168	0.1138	0.0333	0.0028	0.1543	0.0520	0.1148	0.3297	0.0754	0.2054	0.0165	0.1279	0.2769	0.1872	0.1462	0.1952	0.0165
10	0.0156	0.0630	0.0717	0.0193	0.0421	0.1565	0.2790	0.2362	0.1148	0.0259	0.5881	0.2165	0.0748	0.1832	0.3613	0.5075	0.2237	0.2649	0.3970	0.1832
11	0.0157	0.0630	0.0234	0.0469	0.0541	0.0646	0.2679	0.4247	0.3297	0.5881	0.1988	0.0743	0.0127	0.0320	0.4378	0.6168	0.3632	0.3196	0.3325	0.1592
12	0.1411	0.0848	0.0160	0.1090	0.0219	0.0708	0.0610	0.0896	0.0754	0.2165	0.0743	0.1102	0.2078	0.0240	0.0500	0.1751	0.0659	0.1443	0.1570	0.2125
13	0.1606	0.0489	0.2383	0.0877	0.0583	0.0276	0.0627	0.1247	0.2054	0.0748	0.0127	0.2078	0.0981	0.0137	0.0563	0.0546	0.0128	0.0489	0.0483	0.0137
14	0.0803	0.0572	0.0147	0.0326	0.0445	0.4962	0.0233	0.0763	0.0165	0.1832	0.0320	0.0240	0.0137	0.0896	0.0828	0.0433	0.0882	0.0855	0.0205	0.0051
15	0.1674	0.1439	0.0826	0.0407	0.0354	0.0942	0.3442	0.1393	0.1279	0.3613	0.4378	0.0500	0.0563	0.0828	0.0035	0.3565	0.1104	0.2198	0.2169	0.0828
16	0.0094	0.1222	0.0492	0.0348	0.0459	0.0062	0.3059	0.3040	0.2769	0.5075	0.6168	0.1751	0.0546	0.0433	0.3565	0.1718	0.3507	0.4604	0.3869	0.1177
17	0.0534	0.0228	0.0438	0.0402	0.0052	0.0430	0.0184	0.3834	0.1872	0.2237	0.3632	0.0659	0.0128	0.0882	0.1104	0.3507	0.2028	0.5470	0.1542	0.0578
18	0.0463	0.0436	0.0371	0.0267	0.0319	0.0472	0.1682	0.3480	0.1462	0.2649	0.3196	0.1443	0.0489	0.0855	0.2198	0.4604	0.5470	0.1415	0.2399	0.0597
19	0.0145	0.0249	0.0591	0.0298	0.0507	0.0413	0.0359	0.2300	0.1952	0.3970	0.3325	0.1570	0.0483	0.0205	0.2169	0.3869	0.1542	0.2399	0.0843	0.2488
20	0.0803	0.0220	0.0147	0.0326	0.1148	0.0103	0.2186	0.0763	0.0165	0.1832	0.1592	0.2125	0.0137	0.0051	0.0828	0.1177	0.0578	0.0597	0.2488	0.0896
21	0.0745	0.1626	0.0158	0.0462	0.0549	0.0111	0.1726	0.2237	0.1222	0.1697	0.2370	0.0556	0.0313	0.1352	0.1326	0.3206	0.5773	0.6324	0.1517	0.0377
22	0.1726	0.0675	0.0355	0.0115	0.0067	0.0020	0.1428	0.2794	0.1502	0.2017	0.3848	0.0425	0.1049	0.0924	0.1427	0.3786	0.5586	0.5259	0.1292	0.0677
23	0.0307	0.0854	0.0514	0.0536	0.0301	0.0151	0.8091	0.1619	0.0108	0.2987	0.2805	0.0624	0.0758	0.0282	0.3506	0.2255	0.0786	0.1982	0.0866	0.2205
24	0.0609	0.1471	0.0608	0.0792	0.0161	0.0697	0.2611	0.6757	0.2203	0.3995	0.5743	0.1279	0.1138	0.0211	0.2789	0.4993	0.4445	0.3973	0.2934	0.2018
25	0.0128	0.1246	0.1748	0.1029	0.0533	0.0540	0.1857	0.4839	0.1433	0.2809	0.3846	0.0712	0.1407	0.1142	0.2060	0.3163	0.3149	0.2466	0.2717	0.2308
26	0.0774	0.0925	0.0912	0.0414	0.0928	0.0863	0.1107	0.7520	0.1266	0.4136	0.5488	0.0788	0.1351	0.0987	0.3028	0.3965	0.4076	0.3193	0.2369	0.0987
27	0.0379	0.0545	0.1056	0.0720	0.0910	0.1239	0.1158	0.6714	0.1609	0.4922	0.5766	0.1415	0.1588	0.1204	0.2456	0.3806	0.3928	0.3042	0.3365	0.1989
28	0.0181	0.1282	0.0055	0.0910	0.0829	0.0806	0.1829	0.4762	0.1294	0.2183	0.2511	0.0699	0.0856	0.0400	0.1936	0.3396	0.3505	0.4520	0.1607	0.0400
29	0.0471	0.0025	0.0569	0.0218	0.1361	0.0398	0.1104	0.2163	0.5661	0.1256	0.2992	0.0379	0.1604	0.0198	0.1610	0.2332	0.2239	0.2311	0.3675	0.2582
30	0.0799	0.0267	0.1174	0.0285	0.0706	0.0086	0.0467	0.1587	0.2051	0.1911	0.2906	0.0789	0.0021	0.0409	0.1852	0.2806	0.1346	0.2640	0.2564	0.1398
31	0.1046	0.0129	0.1133	0.0417	0.0383	0.0457	0.1351	0.1269	0.2477	0.2407	0.2823	0.0684	0.0343	0.0227	0.2253	0.2843	0.1487	0.1934	0.4404	0.2252
32	0.0712	0.1505	0.1166	0.0993	0.0621	0.1009	0.1638	0.8326	0.2134	0.3671	0.5423	0.1228	0.1297	0.0429	0.1702	0.4573	0.4760	0.4288	0.2868	0.1340
33	0.1237	0.0898	0.0442	0.0913	0.1014	0.1492	0.0632	0.9496	0.1933	0.2240	0.4104	0.0792	0.1304	0.0740	0.1128	0.3079	0.3867	0.3316	0.2209	0.0740
34	0.0429	0.0173	0.1229	0.1163	0.0369	0.0817	0.0252	0.3180	0.0774	0.1514	0.2229	0.0342	0.0188	0.0405	0.1546	0.2643	0.6043	0.4149	0.1629	0.0405
35	0.0810	0.1055	0.0802	0.1196	0.0778	0.1803	0.0557	0.9312	0.1846	0.2314	0.3915	0.0730	0.1269	0.1055	0.1268	0.3230	0.4074	0.3331	0.2345	0.0571
36	0.0194	0.1092	0.0183	0.0977	0.0729	0.0366	0.5671	0.1543	0.0364	0.3257	0.2168	0.0855	0.0829	0.0309	0.2475	0.1779	0.0933	0.1628	0.0862	0.1653
37	0.0564	0.0842	0.0256	0.1626	0.0150	0.0179	0.0998	0.0498	0.0288	0.1964	0.0552	0.0418	0.0239	0.0089	0.0604	0.1115	0.1007	0.1039	0.0357	0.0089
38	0.1669	0.0314	0.0476	0.0780	0.0408	0.0333	0.0028	0.2007	0.0535	0.0465	0.1440	0.0012	0.0444	0.0165	0.0344	0.1726	0.1399	0.1932	0.1080	0.0165
39	0.0542	0.1202	0.0432	0.1101	0.1123	0.1206	0.0845	0.7844	0.1211	0.2796	0.3545	0.1055	0.0834	0.0973	0.1631	0.2744	0.3551	0.3108	0.2305	0.0973

Source: Author.

5 CONCLUSION AND FUTURE WORKS

In conclusion, this thesis addressed the challenge of interpretability for FFS by proposing a novel pairwise feature selection method using PBM. The proposed method offers valuable insights into the relationships between variables by optimizing feature relations and constructing an interpretable graph. The empirical evaluations conducted on 18 datasets and a case study on Chagas disease demonstrated the approach's efficacy in achieving competitive accuracy while enhancing interpretability. By selecting relevant features and clearly understanding their relationships, the proposed method improves the transparency and trustworthiness of Chagas disease diagnostic systems. This research opens up new avenues for developing interpretable machine learning techniques in medical applications and holds promise for advancing the field of computer-aided diagnostics for various diseases.

In addition to the findings and contributions of this study, there are several potential avenues for future research. One possible direction is to explore utilizing different cost functions in the pairwise feature selection method. While this study employed a specific cost function to measure the relevance and similarity between features, alternative cost functions could be investigated to further optimize the selection process. This could involve considering different feature importance measures or incorporating domain-specific knowledge to guide selection.

Furthermore, extending the proposed method to handle large-scale datasets and high-dimensional feature spaces would be another interesting area of investigation. The algorithm's scalability could be enhanced by exploring parallel computing. By addressing the challenges posed by big data in the context of feature selection, the method could be applied to more comprehensive datasets, potentially leading to improved accuracy and interpretability.

Moreover, integrating the pairwise feature selection method with other machine learning algorithms or ensemble methods could be explored. Combining the strengths of different techniques makes it possible to develop more robust and accurate models. Additionally, investigating the interpretability of the ensemble models and understanding the interactions between the selected features in the ensemble framework would provide valuable insights into the diagnostic process in other CAD tols.

Lastly, conducting extensive clinical validation studies on diverse patient populations and collaborating with medical experts would be essential to assess the real-world applicability and generalizability of the proposed method. This would involve evaluating the performance and interpretability of the method on larger and more diverse datasets, as well as validating the identified feature relationships against known clinical knowledge.

By pursuing these future research directions, we can further enhance the interpretability and effectiveness of feature selection methods in computer-aided diagnostics, ultimately advancing our understanding and diagnostic capabilities in diseases such as Chagas disease.

REFERENCES

ALBERTO, A. C.; LIMEIRA, G. A.; PEDROSA, R. C.; ZARZOSO, V.; NADAL, J. Ecg-based predictors of sudden cardiac death in chagas' disease. In: IEEE. **2017 Computing in Cardiology (CinC)**. *[S.l.]*, 2017. p. 1–4.

ALBERTO, A. C.; PEDROSA, R. C.; ZARZOSO, V.; NADAL, J. Association between circadian holter ecg changes and sudden cardiac death in patients with chagas heart disease. **Physiological Measurement**, IOP Publishing, v. 41, n. 2, p. 025006, 2020.

ASUNCION, A.; NEWMAN, D. **UCI machine learning repository**. *[S.l.]*: Irvine, CA, USA, 2007.

BAMAN, T. S.; LANGE, D. C.; ILG, K. J.; GUPTA, S. K.; LIU, T.-Y.; ALGUIRE, C.; ARMSTRONG, W.; GOOD, E.; CHUGH, A.; JONGNARANGSIN, K. *et al.* Relationship between burden of premature ventricular complexes and left ventricular function. **Heart rhythm**, Elsevier, v. 7, n. 7, p. 865–869, 2010.

BEIRANVAND, F.; MEHRDAD, V.; DOWLATSHAHI, M. B. Unsupervised feature selection for image classification: A bipartite matching-based principal component analysis approach. **Knowledge-Based Systems**, Elsevier, p. 109085, 2022.

BENNASAR, M.; HICKS, Y.; SETCHI, R. Feature selection using joint mutual information maximisation. **Expert Systems with Applications**, Elsevier, v. 42, n. 22, p. 8520–8532, 2015.

BOMMERT, A.; SUN, X.; BISCHL, B.; RAHNENFÜHRER, J.; LANG, M. Benchmark for filter methods for feature selection in high-dimensional classification data. **Computational Statistics & Data Analysis**, Elsevier, v. 143, p. 106839, 2020.

BOUCKAERT, R. R.; FRANK, E. Evaluating the replicability of significance tests for comparing learning algorithms. In: SPRINGER. **PAKDD**. *[S.l.]*, 2004. v. 3056, p. 3–12.

BROWN, G.; POCOCK, A.; ZHAO, M.-J.; LUJÁN, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. **The journal of machine learning research**, JMLR. org, v. 13, p. 27–66, 2012.

BRUNO, R. M.; PUCCI, G.; ROSTICCI, M.; GUARINO, L.; GUGLIELMO, C.; ROSEI, C. A.; MONTICONE, S.; GIAVARINI, A.; LONATI, C.; TORLASCO, C. *et al.* Association between lifestyle and systemic arterial hypertension in young adults: a national, survey-based, cross-sectional study. **High Blood Pressure & Cardiovascular Prevention**, Springer, v. 23, p. 31–40, 2016.

CALDAS, W. L.; MADEIRO, J. P. do V.; PEDROSA, R. C.; GOMES, J. P. P.; DU, W.; MARQUES, J. A. L. Noise detection and classification in chagasic ecg signals based on one-dimensional convolutional neural networks. In: SPRINGER. **International Conference on Computer and Information Science**. *[S.l.]*, 2022. p. 117–129.

CALDAS, W. L.; MADEIRO, J. P. V.; MATTOS, C. L. C.; GOMES, J. P. P. A new methodology for classifying qrs morphology in ecg signals. In: IEEE. **2020 International Joint Conference on Neural Networks (IJCNN)**. *[S.l.]*, 2020. p. 1–9.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. **Electronics**, MDPI, v. 8, n. 8, p. 832, 2019.

CAVALCANTE, C. H.; PRIMO, P. E.; SALES, C. A.; CALDAS, W. L.; SILVA, J. H.; SOUZA, A. H.; MARINHO, E. S.; PEDROSA, R. C.; MARQUES, J. A.; SANTOS, H. S. *et al.* Sudden cardiac death multiparametric classification system for chagas heart disease's patients based on clinical data and 24-hours ecg monitoring. **Mathematical Biosciences and Engineering**, v. 20, n. 5, p. 9159–9178, 2023.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 1, p. 16–28, 2014.

CHEN, X.-w.; JEONG, J. C. Enhanced recursive feature elimination. In: IEEE. **Sixth international conference on machine learning and applications (ICMLA 2007)**. *[S.l.]*, 2007. p. 429–435.

DALGAARD, F.; PALLISGAARD, J. L.; NUMÉ, A.-K.; LINDHARDT, T. B.; GISLASON, G. H.; TORP-PEDERSEN, C.; RUWALD, M. H. Rate or rhythm control in older atrial fibrillation patients: risk of fall-related injuries and syncope. **Journal of the American Geriatrics Society**, Wiley Online Library, v. 67, n. 10, p. 2023–2030, 2019.

DERIGS, U. On solving symmetric assignment and perfect matching problems with algebraic objectives. In: **Optimization and Operations Research**. *[S.l.]*: Springer, 1978. p. 79–86.

FORMAN, G. *et al.* An extensive empirical study of feature selection metrics for text classification. **J. Mach. Learn. Res.**, v. 3, n. Mar, p. 1289–1305, 2003.

GERARDS, A. Matching. **Handbooks in operations research and management science**, Elsevier, v. 7, p. 135–224, 1995.

HASHEMI, A.; DOWLATSHAHI, M. B.; NEZAMABADI-POUR, H. A bipartite matching-based feature selection for multi-label learning. **International Journal of Machine Learning and Cybernetics**, Springer, v. 12, n. 2, p. 459–475, 2021.

HUANG, J.; CAI, Y.; XU, X. A hybrid genetic algorithm for feature selection wrapper based on mutual information. **Pattern recognition letters**, Elsevier, v. 28, n. 13, p. 1825–1844, 2007.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. In: IEEE. **2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)**. *[S.l.]*, 2015. p. 1200–1205.

JR, A. R.; RASSI, A.; LITTLE, W. C.; XAVIER, S. S.; RASSI, S. G.; RASSI, A. G.; RASSI, G. G.; HASSLOCHER-MORENO, A.; SOUSA, A. S.; SCANAVACCA, M. I. Development and validation of a risk score for predicting death in chagas' heart disease. **New England Journal of Medicine**, Mass Medical Soc, v. 355, n. 8, p. 799–808, 2006.

JR, A. R.; RASSI, S. G.; RASSI, A. Sudden death in chagas' disease. **Arquivos brasileiros de cardiologia**, SciELO Brasil, v. 76, n. 1, p. 86–96, 2001.

KAUTZKY-WILLER, A.; HARREITER, J.; PACINI, G. Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. **Endocrine reviews**, Oxford University Press, v. 37, n. 3, p. 278–316, 2016.

KEEGAN, R.; YEUNG, C.; BARANCHUK, A. Sudden cardiac death risk stratification and prevention in chagas disease: a non-systematic review of the literature. **Arrhythmia & Electrophysiology Review**, Radcliffe Cardiology, v. 9, n. 4, p. 175, 2020.

KHALID, S.; KHALIL, T.; NASREEN, S. A survey of feature selection and feature extraction techniques in machine learning. In: IEEE. **2014 science and information conference**. *[S.l.]*, 2014. p. 372–378.

KUHN, H. W. The hungarian method for the assignment problem. **Naval research logistics quarterly**, Wiley Online Library, v. 2, n. 1-2, p. 83–97, 1955.

KUMAR, R. R.; REDDY, M. B.; PRAVEEN, P. A review of feature subset selection on unsupervised learning. In: IEEE. **2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)**. *[S.l.]*, 2017. p. 163–167.

KURSA, M. B. Praznik: High performance information-based feature selection. **SoftwareX**, Elsevier, v. 16, p. 100819, 2021.

KURSA, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. **Journal of statistical software**, v. 36, p. 1–13, 2010.

LEWIS, D. D. Feature selection and feature extraction for text categorization. In: **Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992**. *[S.l.: s.n.]*, 1992.

LI, X.; CHEN, H. Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. **Decision Support Systems**, Elsevier, v. 54, n. 2, p. 880–890, 2013.

MARIN-NETO, J. A.; JR, A. R.; OLIVEIRA, G. M. M.; CORREIA, L. C. L.; JÚNIOR, A. N. R.; LUQUETTI, A. O.; HASSLOCHER-MORENO, A. M.; SOUSA, A. S. d.; PAOLA, A. A. V. d.; SOUSA, A. C. S. *et al.* Diretriz da sbc sobre diagnóstico e tratamento de pacientes com cardiomiopatia da doença de chagas–2023. **Arquivos Brasileiros de Cardiologia**, SciELO Brasil, v. 120, p. e20230269, 2023.

MIAO, J.; NIU, L. A survey on feature selection. **Procedia computer science**, Elsevier, v. 91, p. 919–926, 2016.

MURTY, K. G. **The symmetric assignment problem**. *[S.l.]*, 1967.

NG, A. Y. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In: **Proceedings of the twenty-first international conference on Machine learning**. *[S.l.: s.n.]*, 2004. p. 78.

OKSUZYAN, A.; BRØNNUM-HANSEN, H.; JEUNE, B. **Gender gap in health expectancy**. *[S.l.]*: Springer, 2010. 213–218 p.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 27, n. 8, p. 1226–1238, 2005.

PENTICO, D. W. Assignment problems: A golden anniversary survey. **European Journal of Operational Research**, Elsevier, v. 176, n. 2, p. 774–793, 2007.

PEREZ-SILVA, A.; MERINO, J. L. Frequent ventricular extrasystoles: significance, prognosis and treatment. **ESC Coun. Cardiol. Pract**, v. 9, p. 17, 2011.

PRIMO, P. E.; CALDAS, W. L.; ALMEIDA, G. S.; BRASIL, L. P.; CAVALCANTE, C. H.; MADEIRO, J. P.; GOMES, D. G.; PEDROSA, R. C. Auxílio ao diagnóstico para predição de morte súbita em pacientes chagásicos a partir de dados clínicos: uma abordagem baseada em aprendizagem de máquina. In: SBC. **Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde**. *[S.l.]*, 2021. p. 335–345.

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in biology and medicine**, Elsevier, v. 112, p. 103375, 2019.

ROBERT, J. M.; KENNETH, M. K.; ARTHUR, L. B.; AGUSTIN, C. A biological approach to sudden cardiac death: structure, function and cause. **The American journal of cardiology**, Elsevier BV, v. 63, n. 20, p. 1512–1516, 1989.

SIDDIQI, U. F.; SAIT, S. M.; KAYNAK, O. Genetic algorithm for the mutual information-based feature selection in univariate time series data. **IEEE Access**, IEEE, v. 8, p. 9597–9609, 2020.

SOUZA, A. C. J. de; SALLES, G.; HASSLOCHER-MORENO, A. M.; SOUSA, A. S. de; BRASIL, P. E. A. A. do; SARAIVA, R. M.; XAVIER, S. S. Development of a risk score to predict sudden death in patients with chaga's heart disease. **International journal of cardiology**, Elsevier, v. 187, p. 700–704, 2015.

STONE, J. V. Independent component analysis: an introduction. **Trends in cognitive sciences**, Elsevier, v. 6, n. 2, p. 59–64, 2002.

WANG, C.; SITTERS, R. On some special cases of the restricted assignment problem. **Information Processing Letters**, Elsevier, v. 116, n. 11, p. 723–728, 2016.

WANG, H.; LI, H.; ZHOU, H.; CHEN, X. Low-altitude infrared small target detection based on fully convolutional regression network and graph matching. **Infrared Physics & Technology**, Elsevier, v. 115, p. 103738, 2021.

YANG, H.; MOODY, J. Data visualization and feature selection: New algorithms for nongaussian data. **Advances in neural information processing systems**, v. 12, 1999.

ZHAI, Y.; SONG, W.; LIU, X.; LIU, L.; ZHAO, X. A chi-square statistics based feature selection method in text classification. In: IEEE. **2018 IEEE 9th International conference on software engineering and service science (ICSESS)**. *[S.l.]*, 2018. p. 160–163.

ZHANG, C.-H.; HUANG, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. 2008.

ZHANG, Y.; JIANG, F.; RHO, S.; LIU, S.; ZHAO, D.; JI, R. 3d object retrieval with multi-feature collaboration and bipartite graph matching. **Neurocomputing**, Elsevier, v. 195, p. 40–49, 2016.

ZHAO, Z.; ANAND, R.; WANG, M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In: IEEE. **2019 IEEE international conference on data science and advanced analytics (DSAA)**. *[S.l.]*, 2019. p. 442–452.

ZHOU, H.; WANG, X.; ZHANG, Y. Feature selection based on weighted conditional mutual information. **Applied Computing and Informatics**, Emerald Publishing Limited, 2020.

ZHOU, H.; WANG, X.; ZHU, R. Feature selection based on mutual information with correlation coefficient. **Applied Intelligence**, Springer, v. 52, n. 5, p. 5457–5474, 2022.