UC | Facultad de Ciencias

Data-Driven Insights into Ischaemic Heart Disease: Exploring Individual, Environmental and Social Influences through machine learning

(Análisis de datos sobre la cardiopatía isquémica: Exploración de influencias individuales, ambientales y sociales mediante el aprendizaje automático)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autor: Juan Miguel Cano Mazón

Director\es: Jon Zubiaur Zamacola
Joaquín Bedia Jiménez

Julio 2024

Acknowledgements

Me gustaría agradecer a mis dos directores, Jon y Joaquín, por su infinita paciencia, dedicación y apoyo durante estos meses. Su orientación y compromiso han sido fundamentales para el desarrollo de este proyecto.

También quiero expresar mi agradecimiento a mi familia y amigos por brindarme su constante apoyo durante este trayecto.

Contents

List of Figures

Abstract

This Master's Thesis presents a pioneering approach in the field by utilizing a unique, predictor-rich database to advance our understanding of ischaemic disease. Despite the database's limited size, it offers a wealth of predictors, including underexplored socio-health descriptors. More precisely, the study analyzes data from approximately 2400 patients at the Hospital Universitario Marqués de Valdecilla's Cardiology Unit from June 2016 to March 2020. The data includes a broad spectrum of variables, from medical history and treatments to social factors and lifestyle behaviors, often overlooked in similar studies.

The research focuses on predicting three critical determinants of a patient's clinical condition: cardiovascular death, heart attack, and hemorrhage. We employ Bayesian Networks, supplemented with random forest algorithms, to enhance model robustness and interpretability. These results are further expanded with additional classification algorithms that provide further support to the validity of the network models built.

The primary goal is to assess the influence of various factors on patients' cardiovascular health, leading to improved risk understanding and the development of more personalized interventions. Although some of the target variables attain low predictive accuracy, our results have been generally positive and promising, indicating the potential of our approach in improving the understanding of cardiovascular health risks and developing more personalized interventions.

Keywords: Ischemic heart disease, Bayesian Networks, Machine Learning, Random Forests, Healthcare, Cardiovascular Disease Prediction

Resumen

Esta tesis de máster presenta un enfoque pionero en el campo de la investigación de la cardiopatía isquémica al utilizar una base de datos única y rica en predictores para avanzar en nuestra comprensión de la enfermedad isquémica. A pesar del tamaño limitado de la base de datos, ofrece una gran cantidad de predictores, incluidos descriptores sociosanitarios poco explorados. En concreto, el estudio analiza datos de aproximadamente 2400 pacientes de la Unidad de Cardiología del Hospital Universitario Marqués de Valdecilla desde junio de 2016 hasta marzo de 2020. Los datos incluyen un amplio espectro de variables, desde la historia clínica y los tratamientos hasta factores sociales y comportamientos de estilo de vida, a menudo pasados por alto en estudios similares.

La investigación se centra en la predicción de tres determinantes críticos del estado clínico de un paciente: muerte cardiovascular, infarto de miocardio y hemorragia. Empleamos redes bayesianas, complementadas con bosques aleatorios (*random forests*), para mejorar la robustez y la interpretabilidad del modelo. Estos resultados se amplían con algoritmos de clasificación adicionales que respaldan aún más la validez de los modelos de red construidos.

El objetivo principal es evaluar la influencia de diversos factores en la salud cardiovascular de los pacientes, lo que permitirá comprender mejor los riesgos y desarrollar intervenciones más personalizadas. Aunque algunas de las variables objetivo obtienen una precisión predictiva baja, nuestros resultados han sido en general positivos y prometedores, lo que indica el potencial de nuestro enfoque para mejorar la comprensión de los riesgos para la salud cardiovascular y desarrollar intervenciones más personalizadas y efectivas.

Palabras clave: Cardiopatía isquémica, Redes Bayesianas, Aprendizaje Automático, Bosques Aleatorios, Asistencia Sanitaria, Predicción de Enfermedades Cardiovasculares

1

Introduction

1.1 Cardiovascular diseases: the leading cause of death

Cardiovascular diseases are the leading cause of death and morbidity worldwide, accounting for approximately 17.9 million deaths in 2022, representing 32% of all global deaths (WHO, 2023). This alarming figure underscores the severity of the problem and the need for continued efforts in the prevention, diagnosis, and treatment of these diseases (Vaduganathan et al., 2022). Among the various cardiovascular diseases, ischemic heart disease, also known as coronary artery disease, occurs when there is a reduction in blood flow to the heart muscle due to the obstruction of the coronary arteries, usually caused by atherosclerosis (Rafieian-Kopaei et al., 2014).

In addition to their considerable impact on health, cardiovascular diseases also impose a significant economic burden. Direct costs include medical expenses related to hospitalizations, medications, procedures, and specialist visits. These are compounded by indirect costs arising from loss of productivity, long-term disabilities, and the need for prolonged care. According to a recent study, the global economic cost of cardiovascular diseases amounts to billions of dollars annually, reflecting both healthcare expenses and associated economic losses (Weintraub, 2023).

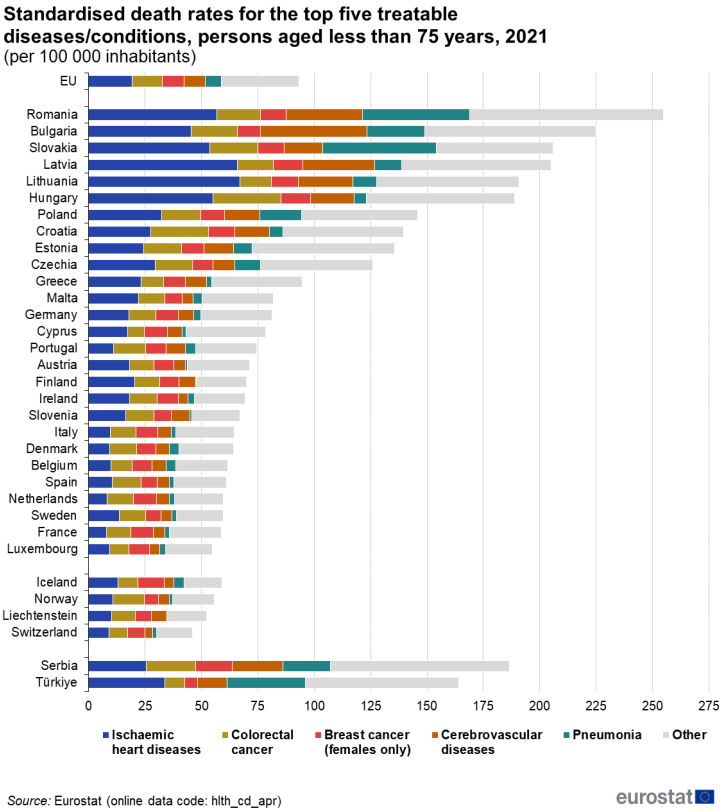

Figure 1.1: *Death rates standardized for the top five treatable diseases/conditions in individuals under 75 years, 2021 (Source: Eurostat)*

1.2 Factors contributing to cardiovascular risk

The risk factors for ischemic heart disease can be classified as modifiable and non-modifiable. Among the non-modifiable factors, risk increases with age. Women experience a delayed onset of ischemic heart disease compared to men, partly due to the protective effects of estrogen before menopause (Bhupathy et al., 2010). Genetic predisposition also plays a role, increasing the likelihood of hypertension, heart attacks, or arrhythmias (Bachmann et al., 2012).

Among the modifiable risk factors, high blood pressure is one of the most significant. This condition forces the heart to work harder, potentially thickening the heart muscle. Effective hypertension treatment can significantly reduce the risk of cardiovascular events, emphasizing the importance of medication, diet, and exercise (Wallace, 2003). Lower blood pressure levels also reduce the risk of kidney failure (Cubrilo-Turek, 2003). High choles-

terol significantly contributes to the formation of plaques on arterial walls, reducing blood flow. Elevated cholesterol levels can lead to dyslipidemia, characterized by an imbalance of lipids in the blood. To counteract this, a diet low in saturated fats and regular exercise are recommended (Stefanick et al., 1998). Quitting smoking is crucial for improving cardiovascular health and reducing the risk of dyslipidemia. This can lead to immediate health improvements, as carbon monoxide from tobacco smoke binds to hemoglobin, reducing oxygen transport and blood lipid levels (Chen and Boreham, 2002). Smoking also alters cholesterol, worsening dyslipidemia (López García-Aranda and Rubira, 2004). Obesity contributes to the development of dyslipidemia. People with obesity tend to have elevated LDL cholesterol levels and low HDL levels, increasing the risk of cardiovascular diseases. Obesity is also closely related to insulin resistance and is a significant risk factor for type 2 diabetes, often associated with alterations in the lipid profile (Katta et al., 2021). Lifestyle plays an essential role in cardiovascular health. Physical inactivity increases the risk of obesity, alters cholesterol levels, and promotes the development of hypertension (Cassiano et al., 2020; Alpsoy, 2020). A low-salt diet is crucial for prevention, as high salt intake is associated with hypertension. Excessive alcohol consumption can also raise blood pressure and negatively affect blood cholesterol levels (Whitman et al., 2017). The Spanish Heart Foundation provides recommendations to control risk factors and prevent coronary events, in line with other international authorities such as the European Society of Cardiology (Visseren and ... et al., 2021). These include quitting smoking, controlling blood pressure, and maintaining an ideal body weight. It is essential to reduce LDL cholesterol by 50%, aiming for levels equal to or lower than 55 mg/dl. Adopting a healthy diet rich in fish, fruits, and vegetables is also vital. Moreover, physical exercise plays a crucial role, with cardiac rehabilitation programs focused on controlled physical exercise for patients who have suffered an acute coronary syndrome (Fundación Española del Corazón, 2023).

Moreover, social determinants are important factors seldom analyzed in the literature (Sun et al., 2023). By analyzing these factors, a model could potentially predict the risk of ischaemic heart disease more accurately (Ohm et al., 2018; Freak-Poli et al., 2021), enabling early intervention and better patient care, and that is the reason this Master's Thesis will make an emphasis in this type of variables, leveraging the socio-health profile individual information gathered by the SCS, as detailed in Chapter 2.

1.3 Machine Learning applications in Medicine

Currently, precision medicine, also known as personalized medicine, has gained importance. Although doctors have always sought individualized treatments, modern tools now allow for a detailed view of patients, thus improving medical decision-making (Shameer et al., 2017; Steinhubl and Topol, 2015). The integration of artificial intelligence (AI) and machine learning is revolutionizing medical practice.

Advances in AI in the medical field are providing significant benefits. Using machine learning techniques, AI analyzes large volumes of medical data, identifying patterns and relationships that may go unnoticed by humans. Machine learning can be supervised, unsupervised, or reinforcement-based, adapting to various needs and types of data.

For example, supervised learning is applied in the diagnosis of ischemic heart diseases using labeled data, improving the accuracy and speed of diagnoses (Churpek et al., 2016). In contrast, unsupervised learning discovers patterns without labels, helping to identify new patient subgroups and personalize treatments. These machine learning algorithms can predict cardiovascular risk more accurately than traditional methods (Weng et al., 2017). Numerous studies have demonstrated their potential and effectiveness in improving medical diagnoses and prognoses (Miotto et al., 2018; Ogunpola et al., 2024).

1.4 Objectives and structure of the document

The objective of this work is to employ machine learning algorithms to study the relationships between various cardiovascular risk factors and see how they influence ischemic disease. To this end, the document is divided into 6 chapters and 3 appendices. The distribution of the chapters is as follows:

- **Chapter 1. Introduction:** The context and objectives of the study are presented.
- **Chapter 2. Dataset Overview:** Detailed description of the dataset used for the analysis.
- **Chapter 3. Bayesian Networks:** Explanation and application of Bayesian networks in the study.

- **Chapter 4. Random Forests:** Description and use of random forest algorithms in the analysis.
- **Chapter 5. Conclusions and Future Research Directions:** Summary of the main findings and proposals for future research based on the results obtained.

1.5 Legal and ethical considerations. Reproducibility

Due to the sensitive nature of medical data, it is important to handle it with care. For this reason, it has not been made public in this Master's Thesis. However, the executed notebooks showing the obtained results have been uploaded. The codes used for these analyses are available in the following public repository:
https://github.com/JuanMiguelCano/TFM

While the medical data used in this study is sensitive and cannot be disclosed due to privacy and ethical considerations, we have ensured that all other aspects of our research adhere to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson and ... et al., 2016). This includes the methodologies, algorithms, and code used in our analysis.

This way, the Master's Thesis codebase is thoroughly documented and structured in a way that promotes reproducibility. We have provided clear instructions for setting up the necessary computational environment, along with detailed comments explaining the purpose and functionality of each section of the code. This ensures that other researchers can understand, replicate, and build upon our work, even if they do not have access to the original dataset. Furthermore, we have made every effort to ensure that our results are robust and reliable. This includes rigorous testing of our code, cross-validation of our models, and sensitivity analyses to assess the impact of potential sources of error or bias. In summary, while the sensitive nature of our data prevents us from sharing it publicly, we have taken all possible steps to ensure that our research is as open, transparent, and reproducible as possible within these constraints.

2

Dataset Overview

The study sample includes data from 2399 patients of the Servicio Cántabro de Salud (SCS) who underwent procedures within the public health system of Autonomous Community of Cantabria (Spain) from June 12, 2016, to March 9, 2020, before the onset of COVID-19. The data were collected through follow-up medical questionnaires, which contain some redundancies and require preprocessing before starting the analyses. They are originally presented in an *xlsx* file that includes an explanatory dictionary with the variable categories. Most of these variables are categorical. As a result, the focus of this Master's Thesis is on specific ML classifiers that are designed to handle categorical data, such as discrete Bayesian Network models (Chapter 3), Random Forests (Chapter 4) and SVM, KNN, Naive Bayes ... (Appendix C).

In their original format, the variables are grouped into five categories:

1. **Background**: Includes information such as age, sex, weight, height, family history of ischemic heart disease, smoking, diabetes, hypertension, dyslipidemia, chronic kidney disease, hemoglobin levels, history of myocardial infarction, arteries treated with angioplasty and previous bypasses, peripheral arterial disease, and history of stroke. This set of variables holds significant importance as they represent *a priori* information about the patient. This means that these variables contain pre-existing knowledge or conditions of the patient before the onset of the disease or prior the patient's admission. The remaining subsets are data collected after the patient's admission.

2. **Procedure**: Includes the arteries treated during the procedure, the type of access

(femoral or radial), and the treatment decided, with ejection fraction (EF) being one of the most important measures to assess cardiac function.

3. **Admission**: Contains information on complications and medications such as Adiro, ADP, and ACO.

4. **Follow-up**: It includes events of interest such as cases of cardiovascular death, myocardial infarction, and hemorrhage observed during the follow-up.

5. **Socio-Health Profile (PSS)**: Gathers information on clinical aspects, habits, cardiovascular risk factors, self-care, and social aspects. This includes knowledge about the disease, family habits such as smoking, alcohol or drug consumption, dietary preferences, exercise practice, frequency of medical check-ups, educational level, type of occupation, marital status, family support, and the use of social networks.

2.1 Data Curation and Preprocessing

Data preprocessing is fundamental to ensure the consistency of information. The variables were categorized into three main groups: MH (medical histories), TRT (treatments), and SOC (social variables), along with the events of interest: *cardiovascular death*, *acute myocardial infarction*, and *hemorrhage*. These target variables were named CVdeath, AMI and HEMORRHAGE respectively.

During pre-processing of data, various types of errors were corrected, including erroneous data and outliers. Additionally, null rows and those with a significant number of missing variables were eliminated.

The dataset contained many columns with repeated information, which allowed for filling in missing values in some cases. However, due to inconsistencies in the duplicated information, priority was given to the data provided in the Medical Histories (MH). Although the number of missing values was not excessively high, they were imputed using the Scikit-learn library in Python. During this preprocessing, binning discretization was used for some continuous variables, as it facilitates their handling and inclusion in Bayesian networks (Dimitrova et al., 2010). Finally, all data was saved in a CSV file named `DATA.csv`.

Regarding the variables in the MH group (14 variables), age (MH_AGE) was categorized into three groups: under 60 years, between 60 and 72 years, and over 72 years. A new variable (MH_BMI) was created to represent the Body Mass Index, using the ranges 0-18.5, 18.5-25, 25-30, and over 30, calculated from the continuous variables of weight and height. Hemoglobin levels (MH_HGB) were categorized according to clinical guidelines: 12.1-15.1 g/dL for women and 13.8-17.2 g/dL for men, classifying them as 1 for normal levels and 2 for abnormal levels. The sex variable (MH_SEX) was maintained as 1 for female and 2 for male. The variable MH_FHxIHD indicated a family history of ischemic heart disease, with 0 for no and 1 for yes. Smoking status (MH_SMK) was categorized as 1 for current smoker, 2 for ex-smoker, and 3 for never smoked. The presence of diabetes mellitus (MH_DM), hypertension (MH_HTN), dyslipidemia (MH_DLP), chronic kidney disease (MH_CKD), previous myocardial infarction (MH_PMxMI), previous percutaneous coronary intervention (MH_PMxPCI), peripheral arterial disease (MH_PAD), and history of stroke (MH_STK) were all represented with 0 for 'no' and 1 for 'yes'.

For the variables in the procedures and treatments group TRT (8 variables), the procedure indication (TRT_IND) was categorized as 1 for stable angina, 2 for ST-elevation myocardial infarction, and 3 for others. The type of access for treatment (TRT_ACC) was maintained as 1 for femoral and 2 for radial. The variables TRT_LCA, TRT_LAD, TRT_RCA, and TRT_LCX were categorized as 0 for no and 1 for yes. The treatment decision variable (TRT_DEC) was classified as 1 for PCI, 2 for surgery, and 3 for conservative treatment. The ejection fraction (TRT_EF) was categorized into two groups: 0 for good (50-70) and 1 for poor (different).

For the social variables group SOC (13 variables), marital status (SOC_MAR_ST) was categorized into 2 groups: 1 for married and 2 for single, divorced, or widowed. The variable SOC_LIV_ALN (living alone) was simplified from 4 categories to 2, categorizing it as 1 for no and 2 for yes. Educational level (SOC_EDU) was classified as 0 for none and 1 for minimal education. Employment status (SOC_ACT_EMP) was divided into 2: unemployed or retired (1) and active (2). Work type (SOC_WOR_TYPE) was categorized as 1 for white-collar and 2 for blue-collar. Family support (SOC_SUPP) was categorized as 0 for no and 1 for yes. Place of residence (SOC_RES) was classified as 1 for rural and 2 for urban. Exercise (SOC_EX) was categorized as 0 for no and 1 for yes. Salt

diet (SOC_SALT_DIET) was classified as 1 for no salt diet and 2 for salt diet. Alcohol consumption (SOC_ALC) was categorized as 1 for none, 2 for weekends, and 3 for daily. Social media use (SOC_SOC_MED) was categorized as 0 for no and 1 for yes. Mobile phone use (SOC_MOB_PH) was classified as 0 for no and 1 for yes. Treatment adherence (SOC_A_TRT) was categorized as 0 for good and 1 for poor.

Appendix A provides a summary table with the variables and their categories. Appendix B shows the initial analysis of the correlations between the different variables.

The code used for data preparation can be found in `NB1_datacuration`. Additionally, the notebooks `NB2_univariateanalysis` and `NB3_bivariateanalysis` were used to observe the distributions and relationships of the variables. Significant imbalances were observed in the variables for cardiovascular death (CVdeath), acute myocardial infarction (AMI), and hemorrhage (HEMORRHAGE), with percentages of 95.75-4.25, 91.66-8.34, and 96.25-3.75, respectively, which are taken into account when applying the algorithms.

3

Bayesian Networks

In this chapter, Bayesian networks are employed with the objective of discovering knowledge through the analysis of relationships among various variables, with particular interest in how they interact with the target variables CVdeath, AMI, and HEMORRHAGE (Sec. 2.1). Bayesian networks are probabilistic graphical models that represent a set of variables and their conditional dependencies through a graph (Castillo et al., 1997; Scutari and Denis, 2014). Each node in the graph corresponds to one of these variables, and direct dependency relationships are represented by arcs between pairs of variables. An example of a Bayesian network is shown in Fig. 3.1. Indirect dependency relationships are not explicitly represented, but they can be visualized as a sequence of arcs connecting one variable to another through one or more variables, forming a path. These paths in Bayesian networks cannot have cycles, as they are directed acyclic graphs (DAGs). Neapolitan et al. (2004). It is important to understand that the presence of an arc in a Bayesian network does not necessarily imply direct causality between the connected variables; rather, it indicates a probabilistic dependency between them. Bayesian network models can be particularly advantageous for investigating ischaemic heart disease. The Bayesian Network approach in this context has the following specific objectives for the modelling and prediction of the target variables:

1. Input a patient's data into the model (evidence) to estimate their risk (probability) of the above factors (through propagation and inference). This can guide preventive measures and early interventions.

2. Diagnostic Aid: If a patient presents with symptoms of ischaemic heart disease, the

model can help determine the likelihood of the specified outcomes. This can assist in making a more accurate diagnosis.

3. Treatment Planning: By updating the model with new information (such as the background information, socio-health profile etc. –Chapter 2–, clinicians can predict the likely outcomes of different treatment options. This can inform treatment planning.

4. Patient Education: The model can be used to explain to patients how different factors contribute to their risk of disease. This can help patients understand why certain lifestyle changes or treatments are recommended.

5. Research: Clinicians and researchers can use the model to investigate the relationships between different risk factors and ischaemic heart disease. This can contribute to a better understanding of the disease and the development of new treatments and targeted prevention campaigns.

Throughout this chapter, the R language and the `bnlearn` (Scutari, 2010) and `gRain` (Højsgaard, 2012) packages will be used for exact inference, as well as the `Rgraphviz` package for network visualization (Hansen et al., 2024). The codes used in this chapter are available in `NB4_BNanalysis` of the Github repo[1].

3.1 Definitions and essential concepts

Formally, a Bayesian network is defined as a pair $G = (V, E)$, where:

- V is a set of nodes, each of which corresponds to a random variable.
- E is a set of directed arcs between the nodes, representing the conditional dependency relationships between the variables.

Each node $X_i \in V$ is associated with a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$, where $\text{Parents}(X_i)$ are the nodes that have directed arcs towards X_i.

[1] https://github.com/JuanMiguelCano/TFM

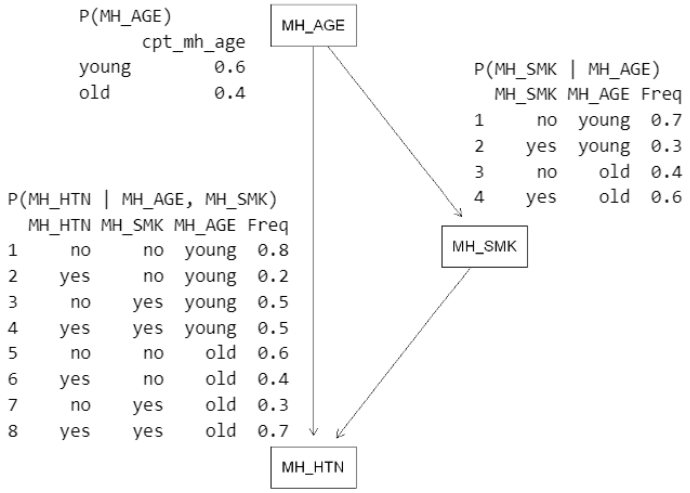

Figure 3.1: *Example of a Bayesian Network with its Conditional Probability Tables*

The joint distribution of all the variables in the Bayesian network can be expressed as the product of the conditional probability distributions of each variable given its parents in the graph:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i)).$$

(Scutari and Denis, 2014)

An important aspect of Bayesian network analysis is examining the fundamental structures that compose these networks (Fig. 3.2). These structures determine how variables are interrelated and how information flows through the graph. Understanding these basic structures is essential as they can be classified and extended to larger networks.

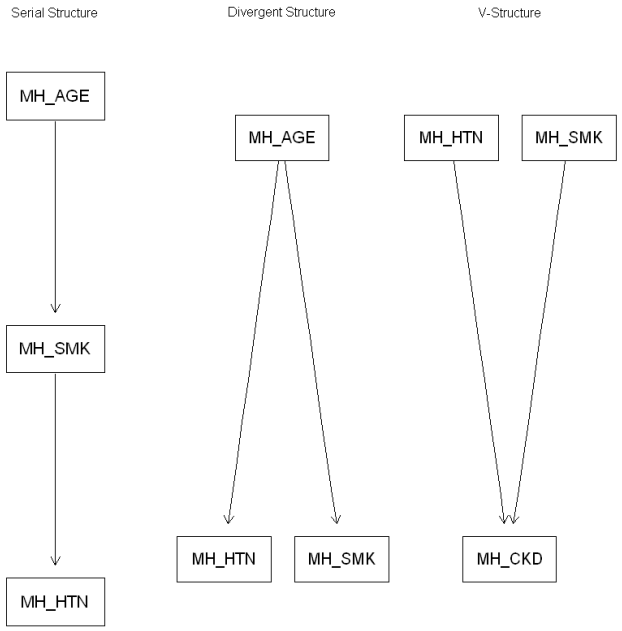

Figure 3.2: *Examples of the three fundamental connections in Bayesian Networks*

- **Series or cascade structure:** The cascade structure with the variables MH_AGE, MH_SMK, and MH_HTN satisfies the condition that information flows from MH_AGE to MH_SMK and from MH_SMK to MH_HTN. MH_HTN is conditionally independent of MH_AGE given MH_SMK, meaning that once the intermediate variable MH_SMK is known, the final variable MH_HTN does not provide additional information about the initial variable MH_AGE. That is,

$$MH\_AGE \perp\!\!\!\perp MH\_HTN \mid MH\_SMK.$$

Without knowing the smoking status, we know that age (MH AGE) may be related to the smoking habit (MH SMK), and that smoking increases the likelihood of hypertension (MH HTN). Therefore, it seems that age and hypertension are indirectly related. However, if we know that a person smokes, the probability of hypertension directly depends on the smoking habit, and we do not need to consider age, meaning the relationship between age and hypertension disappears once the smoking habit is known.

- **Divergent structure:** The divergent structure with the variables MH_AGE, MH_HTN, and MH_SMK satisfies the condition that information flows from MH_AGE to MH_HTN and from MH_AGE to MH_SMK. Additionally, MH_HTN is conditionally independent of MH_SMK given MH_AGE, meaning that once the initial variable MH_AGE is known, the variable MH_HTN does not provide additional information about the variable MH_SMK. This is expressed as

$$MH\_HTN \perp\!\!\!\perp MH\_SMK \mid MH\_AGE.$$

  Without knowing the age, it might seem that smoking MH_SMK and MH_HTN are related, as both can be influenced by MH_AGE. However, if we know a person's age, both the probability of MH_HTN and MH_SMK directly depend on MH_AGE, meaning the relationship between MH_SMK and MH_HTN disappears once the person's MH_AGE is known.

- **V-structure:** The V-structure with the variables MH_HTN, MH_SMK, and MH_CKD satisfies the condition that information flows from MH_HTN to MH_CKD and from MH_SMK to MH_CKD. Additionally, MH_HTN and MH_SMK are conditionally dependent given MH_CKD, meaning that once the final variable MH_CKD is known, MH_HTN provides additional information about MH_SMK. Without knowing the status of MH_CKD, MH_HTN and MH_SMK are independent, but knowing the status of MH_CKD creates a dependency between MH_HTN and MH_SMK.

$$MH\_HTN \not\perp\!\!\!\perp MH\_SMK \mid MH\_CKD.$$

  Without knowing whether a person has MH_CKD, MH_HTN and MH_SMK are independent because they do not influence each other directly. However, if we know that a person has MH_CKD, the probabilities of MH_HTN and MH_SMK become dependent, meaning the relationship between MH_HTN and MH_SMK appears once we know the status of MH_CKD.

In the series and divergent structures, the intermediate node acts as a separator of the probabilistic dependence between the other nodes when it is observed. This means that by knowing the state of the intermediate node, the initial and final variables become independent. In contrast, in the V structure, the separation occurs when the intermediate node is not observed. In this case, the variables that influence the intermediate node are independent of each other. However, when the intermediate variable is observed, the variables that influence it become conditionally dependent on each other.

Another important concept is the Markov blanket, which is the minimal set of nodes that can isolate a specific node from the rest of the graph. The Markov blanket of a node contains all the information necessary to predict the state of that node, and once it is known, the node is conditionally independent of the rest of the graph.

Given a node X in a directed acyclic graph (DAG), the **Markov blanket** of X, denoted as M, is the set of nodes composed of:

1. **Parents of** X: The nodes that have an arrow pointing directly to X.
2. **Children of** X: The nodes to which X directly points.
3. **Parents of the children of** X: The nodes that directly point to the children of X.

Formally, for a node X in a graph G, the Markov blanket M is defined as:

$$M = \text{Parents}(X) \cup \text{Children}(X) \cup \text{Parents}(\text{Children}(X)) \setminus \{X\}$$

The main property of the Markov blanket is that any node X is conditionally independent of all other nodes in the graph given its Markov blanket M.

3.1.1 Structural learning

The structural learning of Bayesian networks involves constructing an appropriate graph that represents a dataset. This process can be carried out using expert knowledge or by learning the graph structure directly from the dataset. Algorithms that perform this task explore the possible graph configurations and are classified according to the statistical criterion employed to find the optimal configuration.

Some algorithms use conditional independence tests to determine the presence of arcs in the graph. They focus on identifying probabilistic dependencies at a local level, often starting by finding the Markov Blanket. If a dependency is supported by the data, the corresponding arc is included in the DAG. Examples of these algorithms are PC, Grow-Shrink, and Incremental Association Markov Blanket (IAMB). Other algorithms focus on maximizing a global score that measures the fit of the DAG to the dependency structure implicit in the data, rather than focusing on individual nodes or arcs. One of the most common scores is the Bayesian Information Criterion (BIC). Among these algorithms are Tabu and Hill Climbing (HC). The Hill Climbing algorithm starts with an empty DAG and proceeds by adding, reversing, and deleting arcs successively, retaining the changes that maximize the score at each step. Additionally, there are hybrid algorithms that combine constraint-based and score-based approaches to leverage the advantages of both methods, improving accuracy and efficiency in constructing the graph.

3.1.2 Parametric learning

In most cases, the parameters of the local distributions in a Bayesian network are estimated from an observed sample. The parameters to be estimated are the conditional probabilities in the local distributions. These probabilities can be calculated using the corresponding empirical frequencies in the dataset.

Once the graph and the distribution of local probabilities for each variable are defined, they are combined to create the actual Bayesian network. The parameters to be estimated are the conditional probabilities of the local distributions of each node, and they can be determined simply by computing the empirical frequency tables from the dataset. A common method for this estimation is Bayesian estimation.

3.1.3 Inference

Inference in Bayesian networks is fundamental for probabilistic calculation and can be performed using exact or approximate methods. Exact inference relies on the junction tree algorithm, which transforms the Bayesian network into a junction tree, facilitating the calculation of conditional probabilities. By reorganizing the network into a more

manageable format, the junction tree allows for more efficient exact inference.

The gRain package in R implements exact inference, providing fast and accurate results. However, its main disadvantage is that as the network grows in size, the computational cost increases exponentially.

On the other hand, approximate inference is based on generating random observations from the Bayesian network. Observations that match the given evidence are used to estimate the conditional probability of the event of interest. In R, the cpquery package allows for this approximate inference, returning the probability of an event given certain evidence.

3.2 Practical implementations

Using the CSV file `DATA.csv`, an analysis and construction of four Bayesian networks have been carried out. These networks have been configured and analyzed as follows:

1. **Social variables and cardiovascular death analysis**
2. **Medical history, treatments, social variables and cardiovascular death analysis**
3. **Medical history, treatments, social variables and acute myocardial infarction analysis**
4. **Medical history, treatments, social variables and hemorrhage analysis**

The structural learning of the Bayesian networks was conducted using the hill-climbing (hc) algorithm. Prior to this, a bootstrap test with 200 resamples (Efron and Tibshirani, 1994) as carried out to compare the hc, tabu, mmhc, and rsmax2 algorithms. The best overall net log-likelihood was obtained with the hc algorithm (Fig. 3.3), which is why it was selected for structural learning, confirming previous findings in the intercomparison of algorithms for discrete bayesian network (Scutari et al., 2019).

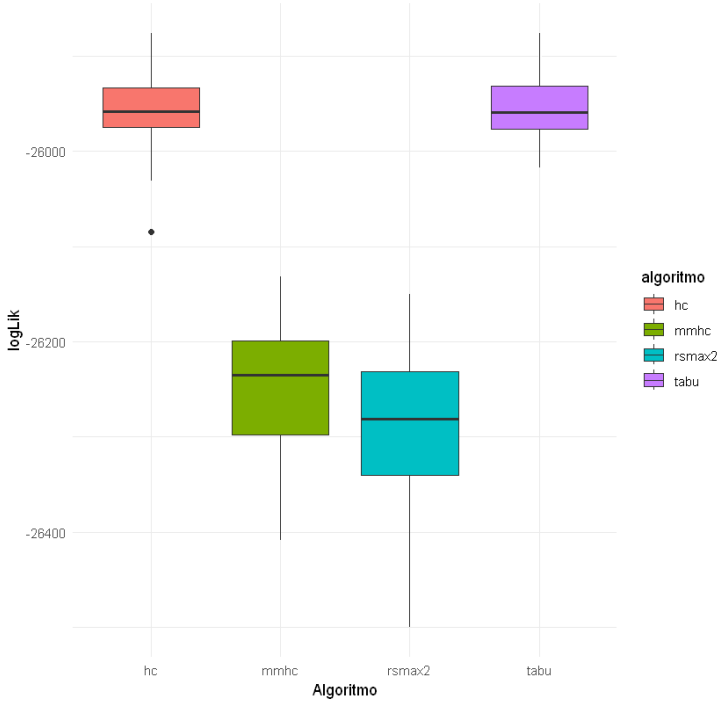

Figure 3.3: *Comparative Analysis of Log-Likelihood Scores: A Study on Hill-Climbing (hc), Tabu, MMHC, and RSMax2 algorithms for structural network learning using bootstrap resampling (R = 200).*

3.2.1 Social variables and cardiovascular death analysis

A first Bayesian network was constructed using a threshold that included edges appearing at least 100 times out of 200 bootstrap iterations, considering only the social variables. This approach was used to observe the relationships between the different social variables before analyzing the complete networks with all variables (Fig. 3.4).

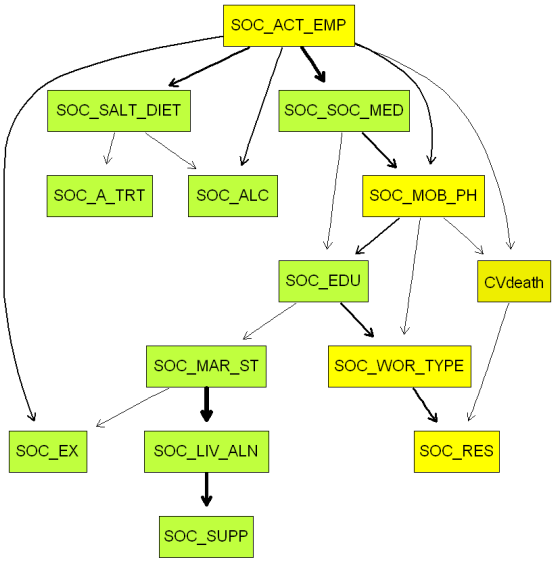

Figure 3.4: *Bayesian network constructed with social variables and the target node CVdeath. The description of the nodes can be found in Appendix A.*

According to data obtained from this network, when SOC_MOB_PH is "0" (no mobile phone), the probability of not dying from cardiovascular causes is 0.90, and the probability of dying is 0.10. In contrast, when SOC_MOB_PH is "1" (having a mobile phone), the probability of not dying increases to approximately 0.9643, while the probability of dying decreases to 0.0357. Possessing a mobile phone (SOC_MOB_PH = "1") is associated with a higher probability of survival compared to not having one (SOC_MOB_PH = "0").

The strength of the arcs was analyzed, revealing that the strongest connection is between the variables SOC_MAR_ST and SOC_LIV_ALN, with a BIC of -405.00. Other significant connections include SOC_ACT_EMP and SOC_SOC_MED (-98.31), SOC_LIV_ALN and SOC_SUPP (-79.79), and SOC_SOC_MED and SOC_MOB_PH (-44.42).

When SOC_MAR_ST is 1 (married), the probability of not living alone is 0.9904, while the probability of living alone is 0.0096. In contrast, when SOC_MAR_ST is 2 (another marital status, single, divorced, or widowed), the probability of not living alone decreases to 0.5352, and the probability of living alone increases to 0.4648. This reflects that married individuals have a much higher probability of not living alone compared to those who are not married, which is an obvious result that supports the coherence of the network (Fig. 3.5).

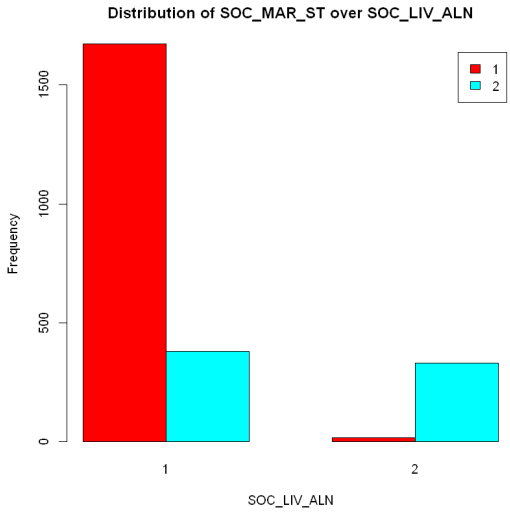

Figure 3.5: *Distribution of SOC_MAR_ST (marital status) over SOC_LIV_ALN (lives alone). The frequencies change drastically between levels, suggesting a strong correlation. SOC_MAR_ST has the states 1 (Married) and 2 (Others), while SOC_LIV_ALN has the states 1 (No) and 2 (Yes). This information comes from the conditional probability tables (CPTs) associated with each node, which are updated when new evidence is presented.*

For individuals with SOC_ACT_EMP at 1 (inactive), the probability of not using social media is 0.8053, while the probability of using social media is 0.1947. In contrast, when SOC_ACT_EMP is 2 (active), the probability of not using social media decreases to 0.5031, and the probability of using social media increases to 0.4969. Active individuals have a higher probability of using social media compared to inactive individuals.

In cases where SOC_SOC_MED is 0 (does not use social media), the probability of not using the mobile phone (SOC_MOB_PH = 0) is 0.1488, while the probability of using the mobile phone (SOC_MOB_PH = 1) is 0.8512. In contrast, when SOC_SOC_MED is 1 (uses social media), the probability of not using the mobile phone is 0.0064, and the probability of using it is 0.9936. This shows that almost all people who use social media also use the mobile phone.

Next, we illustrate the three Bayesian networks obtained using the variables MH, TRT, and SOC, each with one of the target nodes: CVdeath (Fig. 3.6), AMI (Fig. 3.7), and HEMORRHAGE (Fig. 3.8).

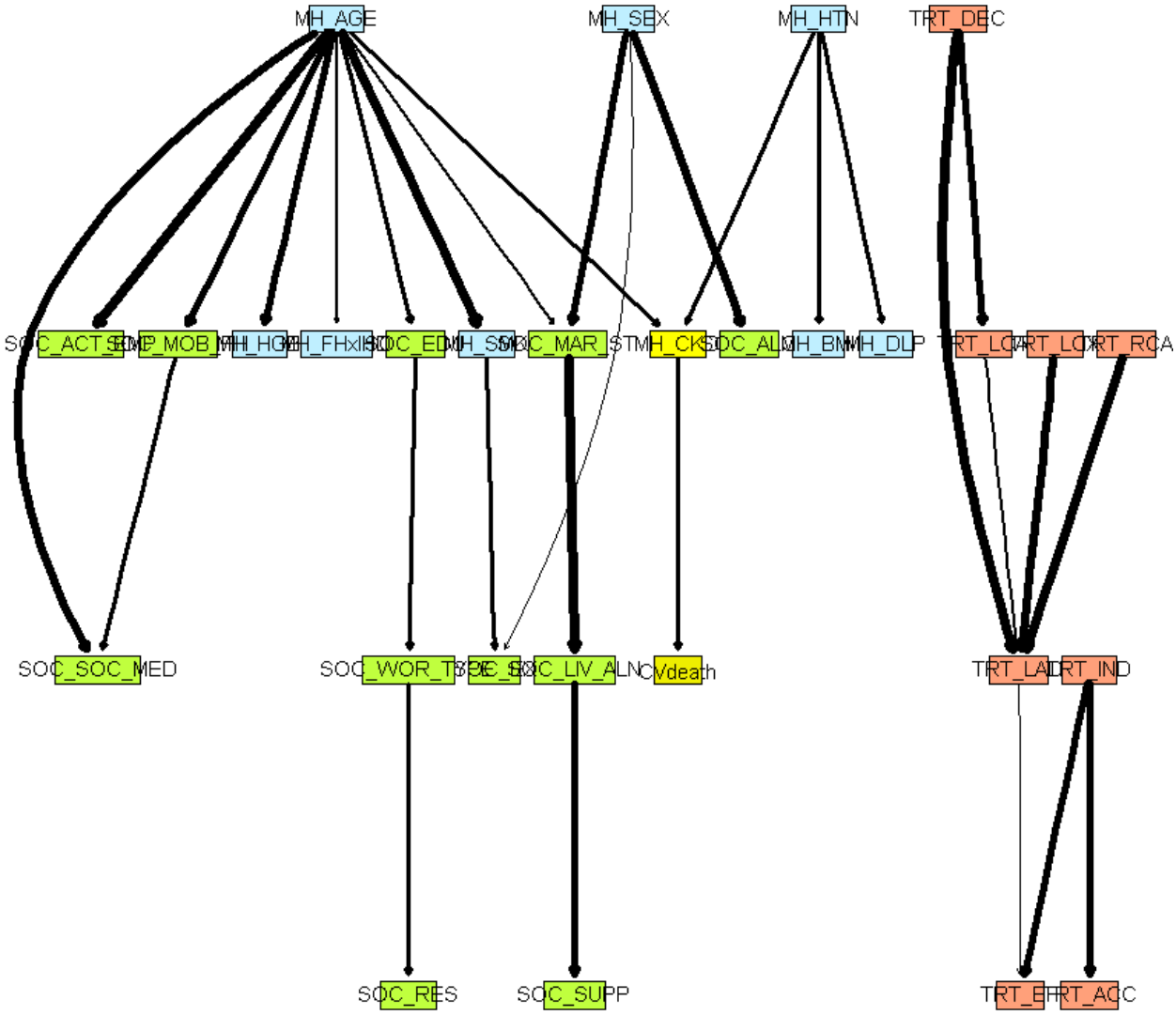

Figure 3.6: *Bayesian network obtained with MH, TRT, and SOC variables, and the target node CVdeath. Threshold: 150 out of 200 bootstrap, 113 parameters. The description of the nodes can be found in Appendix A.*

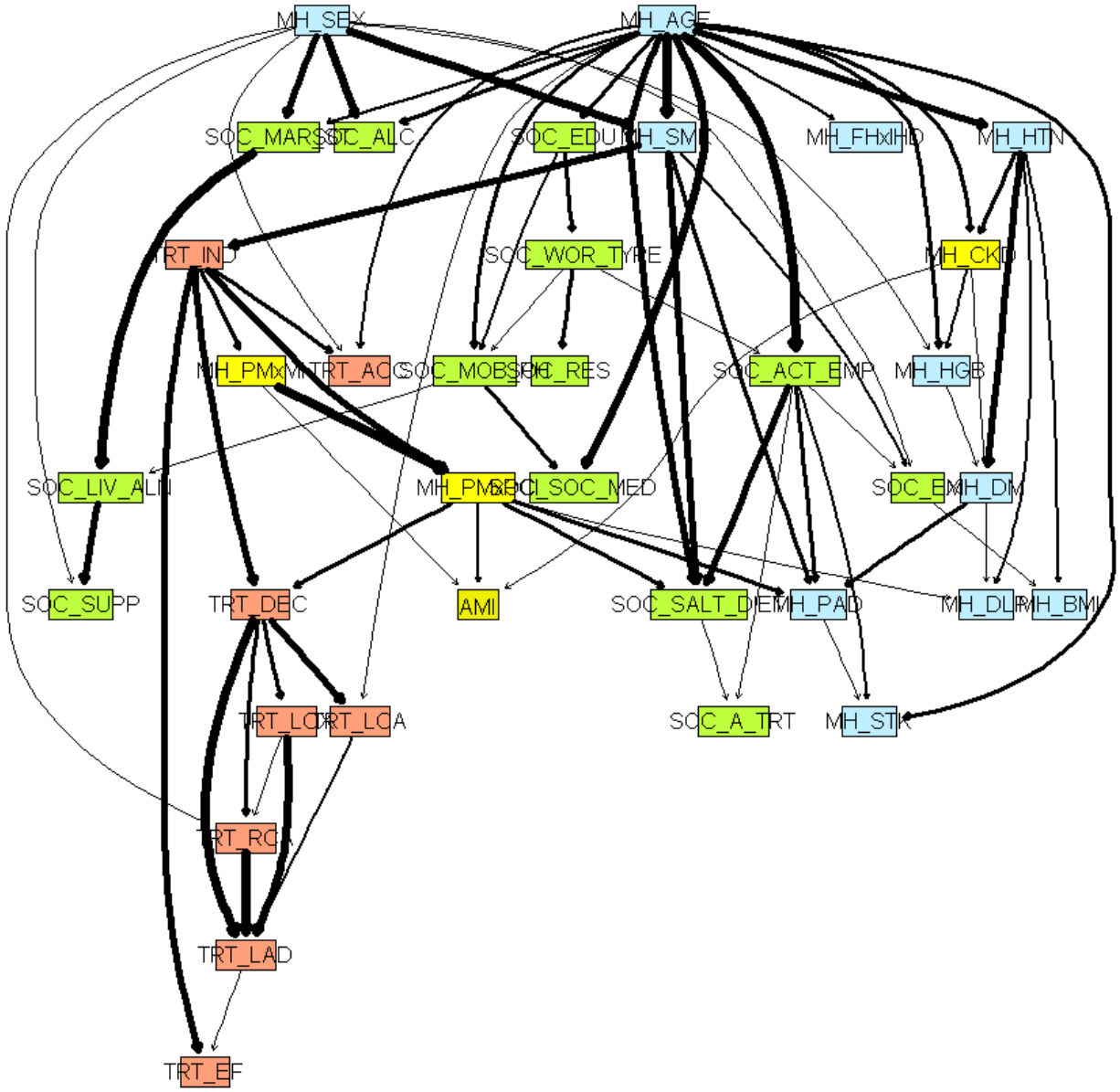

Figure 3.7: *Bayesian network obtained with MH, TRT, and SOC variables, and the target node AMI. Threshold: 50 out of 200 bootstrap, 319 parameters. The description of the nodes can be found in Appendix A.*

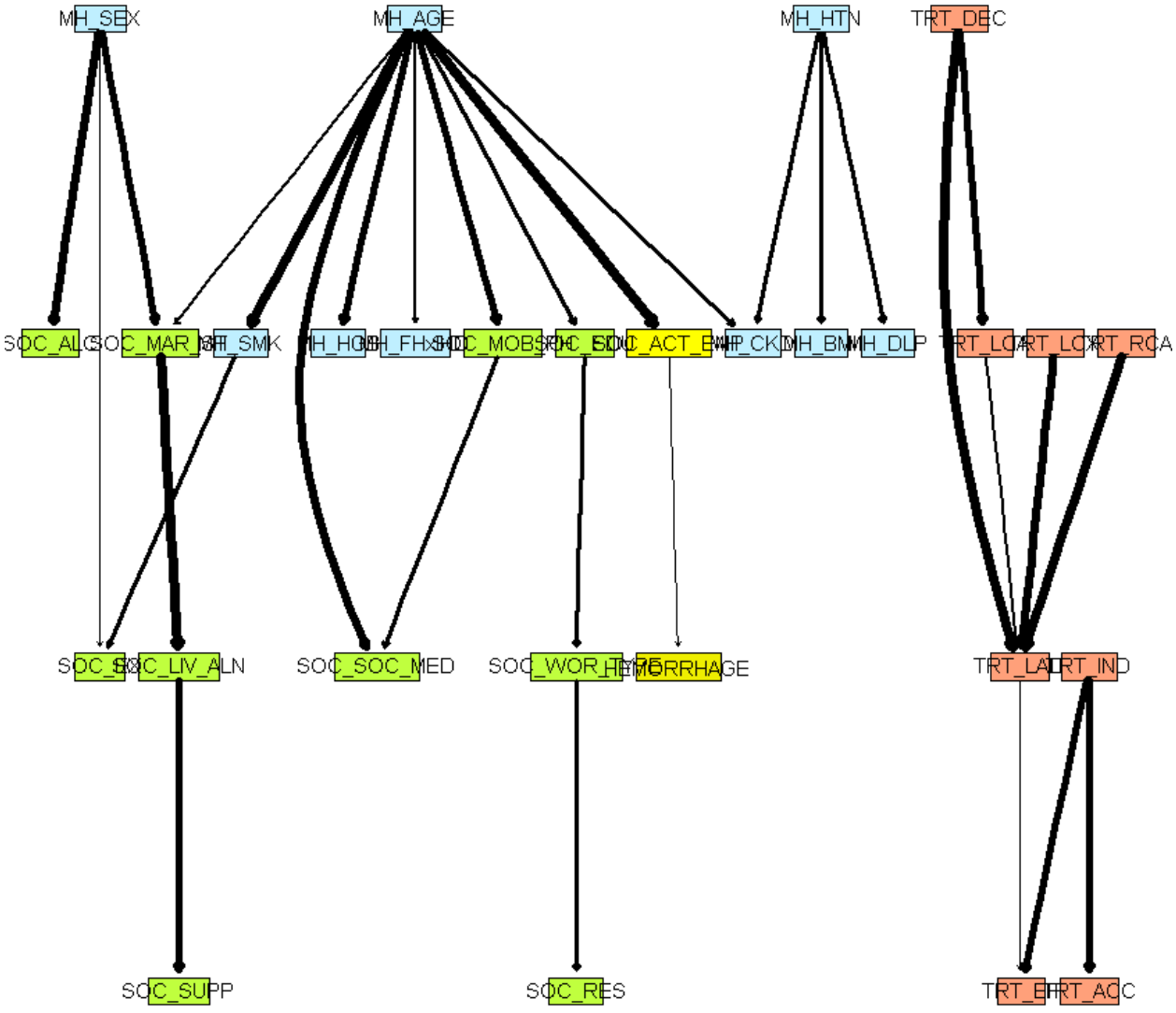

Figure 3.8: *Bayesian network obtained with MH, TRT, and SOC variables, and the target node HEMORRHAGE. Threshold: 150 out of 200 bootstrap, 113 parameters. The description of the nodes can be found in Appendix A.*

3.2.2 Medical history, treatments, social variables and cardiovascular death analysis

Another Bayesian network is constructed for analyzing the node CVdeath (cardiovascular death) using a threshold that includes the edges that appear at least 150 times out of the 200 iterations in the bootstrap. We observe that the edge connecting the nodes MH_CKD and CVdeath is quite robust, appearing 183 times out of 200, and has a BIC strength of -26.779768. In fact, it is among the top 20 strongest edges (Fig. 3.9).

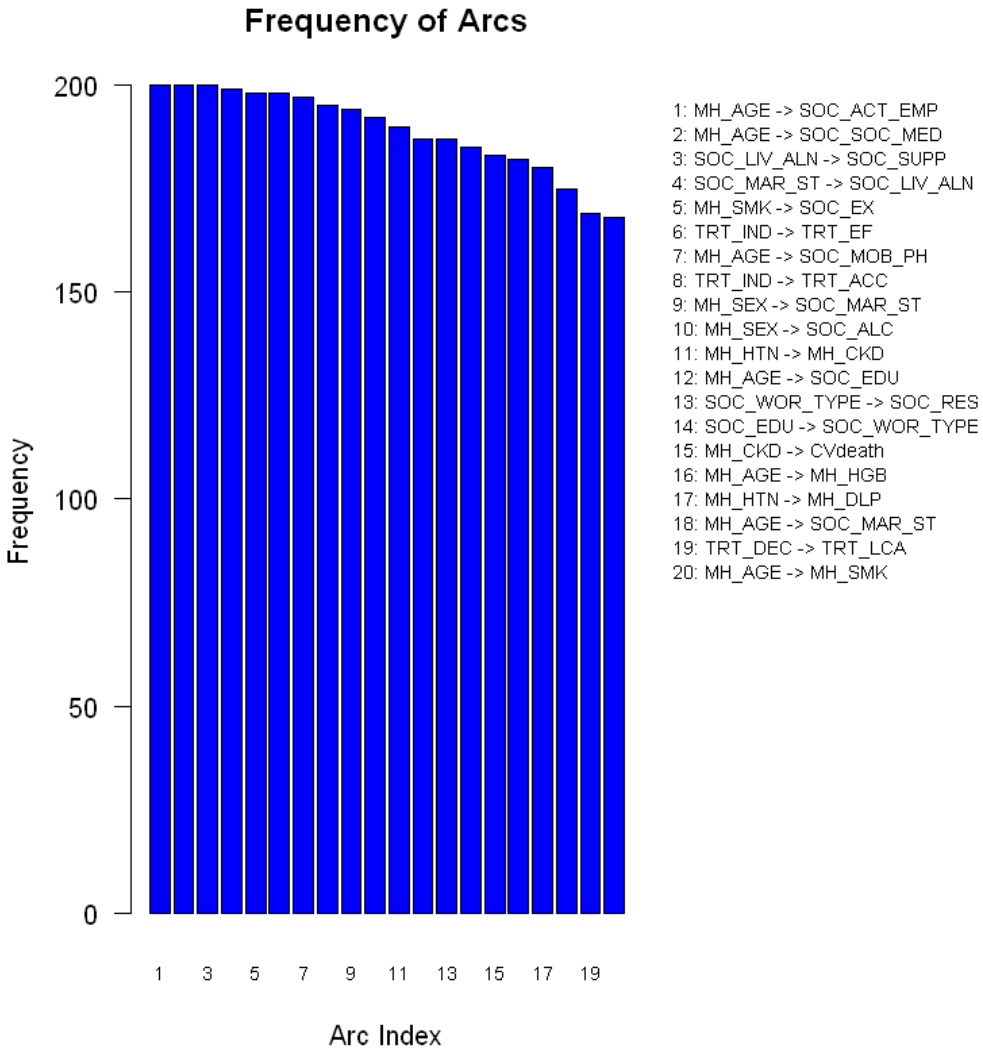

Figure 3.9: *Strength of the top 20 arcs in the Bayesian network analysis of medical history, treatments, social variables and cardiovascular mortality.*

When MH_CKD is 0 (without chronic kidney disease), the probability of not suffering a cardiovascular death is 0.9716, while the probability of suffering it is 0.0284. In contrast, when MH_CKD is 1 (with chronic kidney disease), the probability of not suffering a cardiovascular death decreases to 0.8488, and the probability of suffering it increases to 0.1512. The presence of chronic kidney disease significantly increases the risk of cardiovascular death.

The three strongest connections identified are: the relationship between MH_AGE and SOC_ACT_EMP, with a value of -498.482094; the connection between SOC_MAR_ST and SOC_LIV_ALN, with a value of -405.001224; and the relationship between TRT_DEC and TRT_LAD, with a value of -241.608179.

For the youngest group (MH_AGE = 1), the probability of not being active (SOC_ACT_EMP = 1) is 0.3355, while the probability of being active (SOC_ACT_EMP = 2) is 0.6645. In the intermediate age group (MH_AGE = 2), the probability of not being active rises to 0.8643 and the probability of being active drops to 0.1357. In the oldest age group (MH_AGE = 3), the probability of not being active is 0.9854 and the probability of being active is 0.0146. This reflects that the intermediate and older age groups are the least active, as they include retired individuals.

The second relationship was already analyzed in the previous network.

For TRT_DEC = 1 (ACTP), the probability of not performing treatment on the left anterior descending artery (TRT_LAD = 0) is 0.3371, while the probability of performing treatment (TRT_LAD = 1) is 0.6629. In the case of TRT_DEC = 2 (surgery), the probability of not performing treatment is 0.3017 and the probability of performing treatment is 0.6983. Finally, for TRT_DEC = 3 (conservative), the probability of not performing treatment is 0.6485 and the probability of performing treatment is 0.3515. This shows that the probability of performing treatment on the left anterior descending artery significantly decreases in conservative approaches compared to ACTP and surgery.

3.2.3 Medical history, treatments, social variables and acute myocardial infarction analysis

The third Bayesian network is used to analyze the variable AMI (acute myocardial infarction). A threshold of 50 is set for the edges since the most frequently occurring edge

involving the AMI node appears only 76 times out of 200 iterations, indicating it is less robust. This results in a network with more connections and a greater number of parameters (319).

The edges that most frequently involved the AMI node were MH_PMxPCI (previous percutaneous coronary intervention) - AMI, appearing 76 times; MH_PMxMI (previous myocardial infarction) - AMI, appearing 61 times; and MH_CKD - AMI, appearing 57 times. We also observe that the obtained strengths are much lower compared to the previous network.

For MH_PMxPCI = 0 (without previous percutaneous coronary intervention), the probability of not having an acute myocardial infarction (AMI = 0) is 0.9301, while the probability of having one (AMI = 1) is 0.0699. In contrast, for MH_PMxPCI = 1 (with previous percutaneous coronary intervention), the probability of not having an acute myocardial infarction is 0.8820, and the probability of having one is 0.1180. A previous percutaneous coronary intervention is associated with a higher probability of suffering an acute myocardial infarction.

When MH_PMxMI = 0 (without a history of myocardial infarction), the probability of not having an acute myocardial infarction (AMI = 0) is 0.9278, while the probability of having one (AMI = 1) is 0.0722. In contrast, for MH_PMxMI = 1 (with a history of myocardial infarction), the probability of not having an acute myocardial infarction is 0.8711 and the probability of having one is 0.1289. This reflects that a history of myocardial infarction also increases the probability of suffering another subsequent infarction.

Finally, when MH_CKD = 0 (without chronic kidney disease), the probability of not having an acute myocardial infarction (AMI = 0) is 0.9237, while the probability of having one (AMI = 1) is 0.0763. In contrast, if MH_CKD = 1 (with chronic kidney disease), the probability of not having an acute myocardial infarction decreases to 0.8655 and the probability of having one increases to 0.1345. Chronic kidney disease is also associated with a higher probability of suffering an acute myocardial infarction.

3.2.4 Medical history, treatments, social variables and hemorrhage analysis

The fourth Bayesian network is used to analyze the variable HEMORRHAGE. A threshold of 150 is set, similar to that for CVdeath. This results in a Bayesian network with 113

parameters.

It is observed that the edge connecting SOC_ACT_EMP and HEMORRHAGE appears 167 out of 200 times, indicating it is quite robust. However, the strength is not very high, with a BIC of -9.691784.

For individuals with SOC_ACT_EMP = 1 (inactive), the probability of not experiencing hemorrhage (HEMORRHAGE = 0) is 0.9319, while the probability of experiencing it (HEMORRHAGE = 1) is 0.0681. In contrast, for active individuals (SOC_ACT_EMP = 2), the probability of not experiencing hemorrhage is 0.9169 and the probability of experiencing it is 0.0831. It is observed that the incidence of hemorrhage is slightly higher in active individuals.

3.2.5 Factors influencing cardiovascular mortality

The risk of death from cardiovascular causes increases with age. For individuals older than 72 years, this probability is 0.0524, while it is 0.0396 for those aged 60-72 years, and 0.0342 for individuals under 60 years old. Regarding body mass index (BMI), obese individuals exhibit the highest rate at 0.0435, followed by overweight individuals at 0.0416. Those with normal weight have a slightly lower risk of 0.0402, and underweight individuals have a probability of 0.0429.

The presence of hypertension significantly raises the risk of death, with a probability of 0.0468 compared to 0.0321 for those without hypertension. Similarly, the presence of dyslipidemia is associated with an increased probability, rising from 0.0402 in individuals without dyslipidemia to 0.0429 in those with it. Additionally, hemoglobin levels play a crucial role: individuals with abnormal levels have a probability of 0.0443, while those with normal levels have a slightly lower risk of 0.0406.

In terms of social variables, individuals without formal education have a probability of cardiovascular death of 0.0478, whereas those with education have a reduced probability of 0.0415. Similarly, unemployed individuals have a probability of 0.0443, compared to 0.0354 for those employed. Regarding the use of social networks, non-users have a probability of 0.0437, while users have a probability of 0.0373. As for mobile phone usage, non-users have a probability of 0.0478, whereas users have a probability of 0.0412.

4

Random Forests

The primary aim of this chapter is to examine the outcomes of a classifier that is fundamentally different from Bayesian networks, specifically, random forests. This analysis will provide insights into the consistency of variable importance within the model, helping us to build trust in the Bayesian Network models developed and their consistency as compared to other classifiers in case that both models have similar "important" variables. Moreover, the advancement of random forests will enable a comparative analysis with Bayesian Networks, focusing on the predictive accuracy of the variables of interest. This comparison will provide further insights into the efficacy and potential of our Bayesian network models, reinforcing their role as expert systems for clinical diagnosis. It's important to note that the predictive accuracy experiment is not only limited to the comparison between random forests and Bayesian Networks. It is, in fact, further enriched by the inclusion of additional classification algorithms. However, to maintain conciseness in this discussion, the results derived from these additional algorithms are presented separately in Appendix C.

4.1 Overview of random forests

Random Forests are a supervised learning method composed of multiple decision trees, which significantly improve prediction accuracy and reduce the risk of overfitting (Breiman, 2001). The construction of a Random Forest involves generating numerous independent decision trees, each trained with a random sample of the original dataset. This sample is taken with replacement, meaning some data points may be repeated in the sample used

to train a tree, while other data points may not be used at all for that particular tree.

A key feature of Random Forests is the use of the "Out of Bag" (OOB) error (Matthew et al., 2011). During the formation of each tree, the data not selected for training that specific tree is referred to as "Out of Bag." These OOB data points are used to evaluate the tree's performance, providing an internal and unbiased estimate of the model's generalization error.

The calculation of the OOB error is done by predicting the labels of the OOB data points using the tree that did not see them during its training. By averaging the prediction errors of all the trees, an accurate estimation of the model's performance on unseen data is obtained, without the need for a separate test set. This methodology not only improves the efficiency of the validation process but also increases the reliability of error estimates, as it is based on a large number of individual evaluations.

To construct a decision tree, various criteria are applied to determine how and where to split the data at each node. One of the most common methods is the Gini index, used in classification problems to measure the impurity of a node. The Gini index is calculated by assessing the probability that a randomly selected sample would be incorrectly classified based on the class distribution in the node. During tree construction, the feature and threshold that maximize the reduction in Gini index between the current node and its resulting child nodes are chosen. Additionally, the number of predictors selected randomly as candidates in each division, denoted as 'mtry', has been set to the square root of the total number of predictors. The number of trees used in the forest, 'ntree', has been set to 100. Furthermore, 'minsize', the minimum size of terminal nodes has been set to 1.

4.2 Variable importance measure

During the construction of a decision tree, variable selection is performed by evaluating each available feature and choosing the one that provides the greatest reduction in the Gini index. For each possible split point of a feature, the Gini impurity is calculated for the resulting nodes and compared with the impurity of the original node. The feature and threshold that result in the greatest reduction in the Gini index are selected to make the split at the current node. Variable importance is measured by accumulating the reduction

in the Gini index that each variable provides in all the splits it participates in throughout the tree. Variables that generate the largest reductions in the total impurity of the tree are considered the most important, indicating that these features have a greater impact on the correct classification of the samples.

4.3 Dataset Balancing

In order to make a fair comparison of classifiers, we have undertaken the data balancing, and identical balanced datasets have been used to prepare the out-of-sample predictions. Balancing classes is important because machine learning algorithms can be biased towards the majority class, leading to poor performance on the minority class. This can be problematic if the minority class is of particular interest, as it is the case in our database for some of the key variables such as CVdeath, AMI and HEMORRHAGE, that are rare outcomes among all the ischaemic disease cases recorded.

Data balancing is typically not required for Bayesian Network models. This is because Bayesian Networks are probabilistic models that learn the joint probability distribution of the data. They are less sensitive to class imbalance as they do not rely on the same assumptions as many other machine learning algorithms and can perform well in this case (Flores and Gámez, 2015). However, a balanced dataset has been specifically prepared for this chapter's comparison against random forests. For this purpose, we have used the *ovun.sample* function from the R package ROSE (Lunardon et al., 2014), which randomly selects an equal number of instances from each class (0/1). Additionally, it has been used for the classifier inter-comparison in Appendix C, since in that case we are using the ROC area and other ROC-based performance scores, which are sensitive to class imbalance.

4.4 Practical implementations

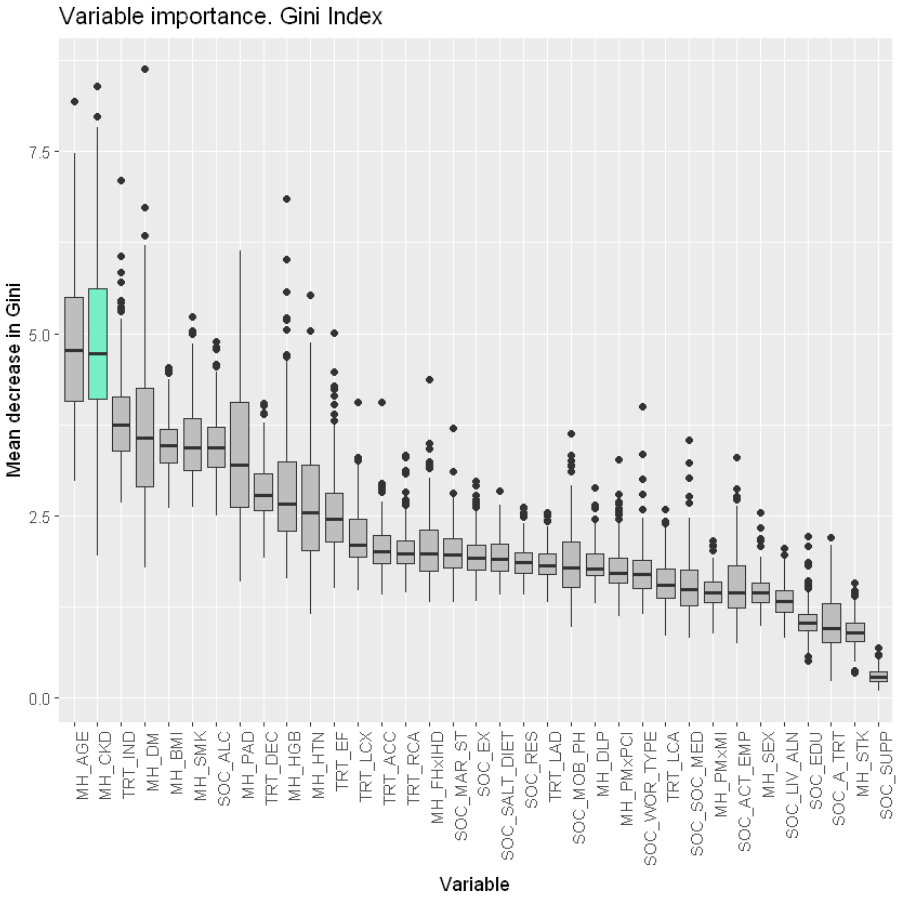

Figure 4.1: *The Gini index is used to assess the importance of variables in explaining cardiovascular death (CVdeath). A bootstrap method with 200 iterations has been used to ensure a robust estimation of variable importance.*

In the previous chapter, it was observed that the Bayesian network identified MH_CKD as the most important node for predicting cardiovascular death (CVdeath). When comparing this information with the variable importance provided by the Gini index, MH_CKD also stood out as a significant variable, tying in importance with MH_AGE. Additionally, MH_AGE was located very close to the CVdeath node in the Bayesian network, further reinforcing its relevance in predicting CVdeath. Despite the two models being completely different, they share similar results, highlighting the importance of MH_CKD and MH_AGE in the predictive model.

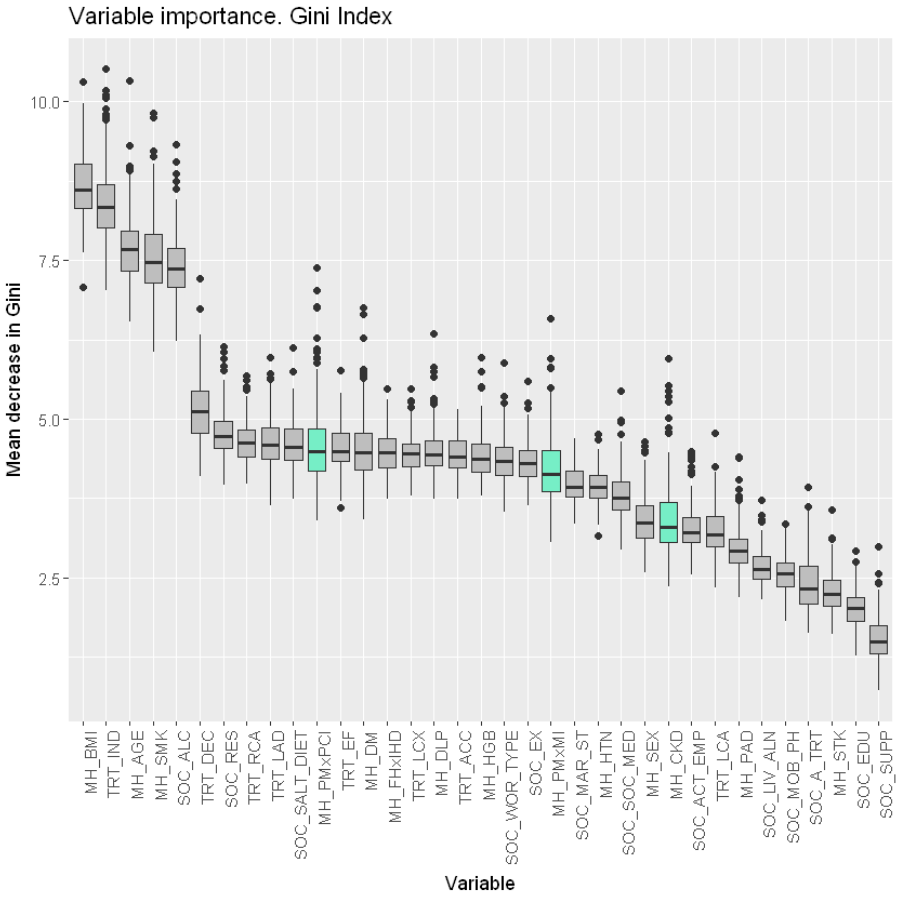

Figure 4.2: *The Gini index is used to assess the importance of variables in explaining acute myocardial infarction (AMI). A bootstrap method with 200 iterations has been used to ensure a robust estimation of variable importance.*

For the case of AMI, the nearby nodes were MH_PMxPCI, MH_PMxMI, and MH_CKD. However, it was observed that the robustness of these connections was considerably lower compared to the previous network, with these links appearing 76, 61, and 57 times respectively. Additionally, the strength obtained using the arc.strength function and the BIC was low (11.154624, 10.803513, and 10.222042 respectively). When checking the results of the ranking obtained by the Gini variable importance index, it is observed that in this case, the values give more importance to the variables MH_BMI, TRT_IND, MH_AGE, MH_SMK, and SOC_ALC. The three variables ordered by robustness and strength remain in the same order, although this time in positions 11, 21, and 26.

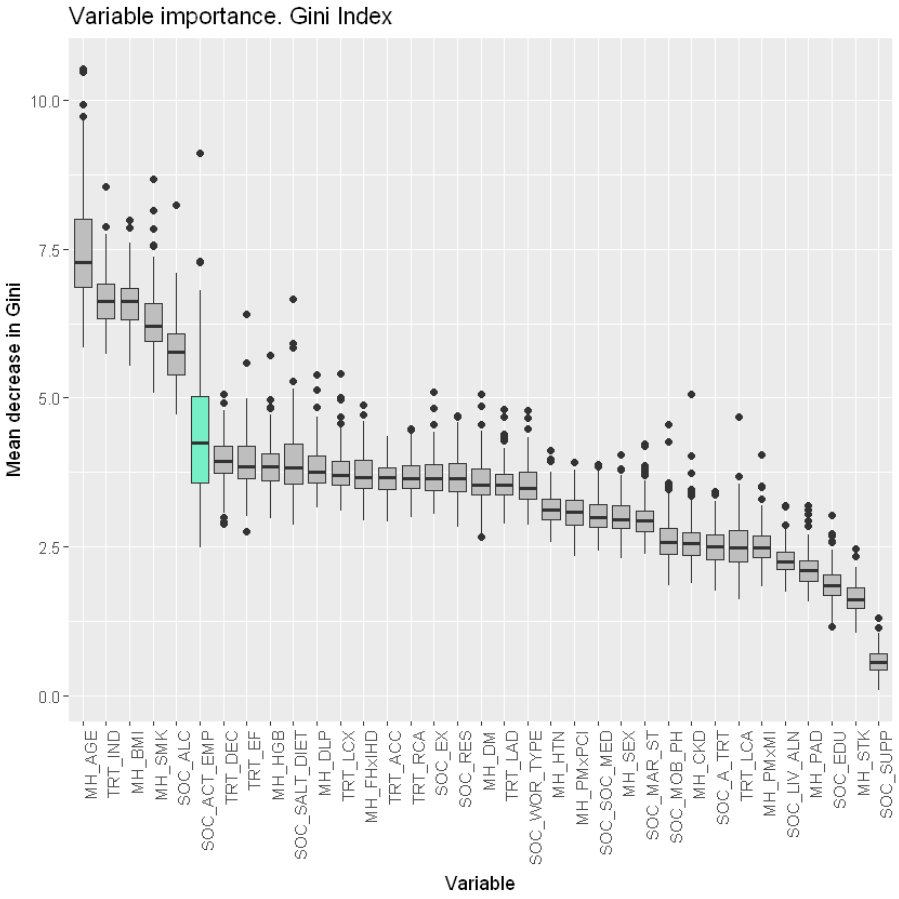

Figure 4.3: *The Gini index is used to assess the importance of variables in explaining hemorrhage. A bootstrap method with 200 iterations has been used to ensure a robust estimation of variable importance.*

Finally, in the case of hemorrhage, we observed that the edge SOC_ACT_EMP - HEMORRHAGE appeared in 167 out of 200 instances, indicating notable robustness. However, when evaluating its strength, the value of -9.691784 was not very high. Now, comparing these results with random forest, we see that it also does not rank this edge among the most important variables. Random forest gives more importance to age and the indicated treatment among the three variables studied.

5

Conclusions and Future Research Directions

Cardiovascular diseases often appear suddenly; however, in reality, they have a long asymptomatic course before manifesting. Understanding the relationships between various factors and variables dependent on a given variable becomes very useful in these cases. In this work, it has been shown, for example, that for a fatal outcome of cardiovascular death (variable CVdeath) , the Bayesian network has correctly identified the importance of the variable chronic kidney disease (MH_CKD), as it is linked to many others related to ischemic heart disease.

The fact that both Bayesian Network models and its random forests counterpart identify the same groups of variables as important provides a form of cross-validation. It strengthens the confidence in these variables' true significance in predicting the outcome, confirming the validity of the models developed. The consistency indicates that both models are robust, as they agree on the influential variables despite their different underlying methodologies. The overlap in important variables could open up possibilities for integrating the two models, potentially leveraging the strengths of both to improve predictive performance. These are some possibilities of such multimodel integration:

Model Stacking: Random Forest and a Bayesian Network model are trained separately, as we did in this work; then another third model can be trained to learn how to best combine their predictions, for example using a weighting scheme. This second-level model is trained to effectively capture the strengths of each model based on the patterns in the data.

Feature Engineering: The feature importance from the Random Forest can be used to

guide the structure learning in the Bayesian Network (or vice versa). This can help to force some nodes and/or arcs to be present/absent in the bayesian network (via whitelist/blacklist arguments), or simply discarding features that have low importance in the random forest algorithm.

Hybrid Models: Develop a hybrid model that uses the Bayesian Network for probabilistic reasoning and the Random Forest for handling high-dimensional data. This could be as simple as taking a weighted average of the two predictions, or it could involve a more complex scheme like training another model to learn how to best combine the predictions from both classifiers.

The predictive models for *acute myocardial infarction* (AMI) and *hemorrhage* did not yield good predictive results, with AUC values below 0.6 in most cases. This can be due to a low sample size (there are few positive cases among the database) or to the absence of some key predictors (although the database use is quite comprehensive, well beyond the number of predictors usually handled in the literature of cardiovascular disease research). The prediction of cardiovascular death, however, suggests an accurate classifier that could be useful in real applications. Nevertheless, it has been demonstrated that, although it is not their primary purpose, when Bayesian networks are used as classification models, the results obtained are comparable to those of widely supported classification models like random forest or SVM, resulting in very competitive classifiers for ischaemic disease analysis. Bayesian networks facilitate the processing of large volumes of data to generate informed decisions or recommendations, often exceeding human abilities in the analysis of intricate information. This proves particularly beneficial in the context of ischemic disease prevention, diagnosis, and treatment. Clinicians often face the challenge of making –quick– decisions based on multifaceted problems involving numerous variables, which often interact in non-linear and complex ways. As demonstrated in our work, the capabilities of Bayesian networks significantly aid in navigating these complexities.

One of the greatest strengths of Bayesian networks is their ability to understand and represent the relationships between variables. They allow for the incorporation of knowledge at any time, based on more or less complete evidence, making them extremely flexible and adaptive. This characteristic has proven to be especially useful in this work.

Regarding future lines of research, the use of continuous data could be explored, which would allow for better utilization of data variability and improve the accuracy of predictions. This improvement in the representation of continuous data could lead to significant advances in the accuracy and utility of Bayesian networks in various fields of study.

Furthermore, with an expanded database, we could leverage the power of deep learning models to gain insights into ischemic disease, complementing the knowledge derived from our existing Bayesian network models. Deep learning models, with their ability to learn complex patterns and relationships from large volumes of data, can provide a different perspective and potentially uncover novel insights. These models excel at identifying intricate structures within high-dimensional data, making them particularly suited for tasks such as feature extraction and pattern recognition. This could be invaluable in the context of ischemic disease, where numerous variables and their complex interactions play a role. However, it's important to note that the effectiveness of deep learning models is directly proportional to the size of the available dataset, and that is the main reason that prevented us from the application in this work. Larger datasets allow these models to better learn and generalize, thereby improving their predictive performance. Therefore, having access to a larger database would enable us to tap into the potential of deep learning, providing a more comprehensive understanding of ischemic disease alongside our Bayesian network models.

Bibliography

Alpsoy, Ş., 2020: Exercise and hypertension. *Physical exercise for human health*, 153–167.

Bachmann, J. M., B. L. Willis, C. R. Ayers, A. Khera, and J. D. Berry, 2012: Association between family history and coronary heart disease death across long-term follow-up in men: the cooper center longitudinal study. *Circulation*, **125**, 3092–3098.

Bedia, J., J. Busqué, and J. M. Gutiérrez, 2011: Predicting plant species distribution across an alpine rangeland in northern Spain: a comparison of probabilistic methods. *Applied Vegetation Science*, **14**, 415–432.

Bhupathy, P., C. D. Haines, and L. A. Leinwand, 2010: Influence of sex hormones and phytoestrogens on heart disease in men and women. *Women's health*, **6**, 77–95.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5–32, doi:10.1023/A:1010933404324.

Broecker, J., 2011: Probability Forecasts. *Forecast Verification*, I. T. Jolliffe and D. B. Stevenson, eds., John Wiley & Sons, Ltd, 119–139.

Cassiano, A. d. N., T. S. d. Silva, C. Q. d. Nascimento, E. M. Wanderley, E. S. Prado, T. M. d. M. Santos, C. S. Mello, and J. A. Barros-Neto, 2020: Effects of physical exercise on cardiovascular risk and quality of life in hypertensive elderly people. *Ciencia & saude coletiva*, **25**, 2203–2212.

Castillo, E., J. M. Gutiérrez, and A. S. Hadi, 1997: *Expert Systems and Probabilistic Network Models*. Springer, New York, NY.

Chen, Z. and J. Boreham, 2002: *Smoking and cardiovascular disease*. Thieme Medical Publishers, Inc., New York.

Churpek, M. M., T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, 2016: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine*.

Cortes, C. and V. Vapnik, 1995: Support-vector networks. *Machine Learning*, **20**, 273–297.

Cubrilo-Turek, M., 2003: Hypertension and coronary heart disease. *Ejifcc*, **14**, 67.

Dimitrova, E. S., M. P. V. Licona, J. McGee, and R. Laubenbacher, 2010: Discretization of Time Series Data. *Journal of Computational Biology*, **17**, 853–868, doi:10.1089/cmb.2008.0023.

Efron, B. and R. J. Tibshirani, 1994: *An introduction to the bootstrap*. Chapman and Hall/CRC.

Fielding, A. H. and J. F. Bell, 1997: A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Flores, M. J. and J. A. Gámez, 2015: Impact on Bayesian Networks Classifiers When Learning from Imbalanced Datasets:. *Proceedings of the International Conference on Agents and Artificial Intelligence*, SCITEPRESS - Science and and Technology Publications, Lisbon, Portugal, 382–389.

Freak-Poli, R., J. Ryan, J. T. Neumann, A. Tonkin, C. M. Reid, R. L. Woods, M. Nelson, N. Stocks, M. Berk, J. J. McNeil, et al., 2021: Social isolation, social support and loneliness as predictors of cardiovascular disease incidence and mortality. *BMC geriatrics*, **21**, 1–14.

Fundación Española del Corazón, 2023: Ficha del paciente. Para el uso de pacientes y profesionales de la salud. Fecha de actualización: 2023.
URL https://fundaciondelcorazon.com

Hansen, K. D., J. Gentry, L. Long, R. Gentleman, S. Falcon, F. Hahne, and D. Sarkar, 2024: *Rgraphviz: Provides plotting capabilities for R graph objects*. R package version

2.48.0.
URL https://bioconductor.org/packages/Rgraphviz

Højsgaard, S., 2012: Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, **46**, 1–26, doi:10.18637/jss.v046.i10.

Katta, N., T. Loethen, C. J. Lavie, and M. A. Alpert, 2021: Obesity and coronary heart disease: epidemiology, pathology, and coronary artery imaging. *Current problems in cardiology*, **46**, 100655.

Kohavi, R., 1995: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, 1137–1143.

López García-Aranda, V. and J. G. Rubira, 2004: Smoking and cardiovascular disease. *Adicciones*.

Lorenz, E., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, **26**, 636–&.

Lunardon, N., G. Menardi, and N. Torelli, 2014: Rose: a package for binary imbalanced learning. *R journal*, **6**.

Mandrekar, J. N., 2010: Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, **5**, 1315–1316, doi:10.1097/JTO.0b013e3181ec173d.

Matthew, W. et al., 2011: Bias of the random forest out-of-bag (oob) error for certain input parameters. *Open Journal of Statistics*, **2011**.

Miotto, R., F. Wang, S. Wang, X. Jiang, and J. T. Dudley, 2018: Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*.

Neapolitan, R. E. et al., 2004: *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River.

Ng, A. and M. Jordan, 2001: On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, MIT Press, volume 14.

Ogunpola, A., F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, 2024: Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, **14**, 144.

Ohm, J., P. H. Skoglund, A. Discacciati, J. Sundström, K. Hambraeus, T. Jernberg, and P. Svensson, 2018: Socioeconomic status predicts second cardiovascular event in 29,226 survivors of a first myocardial infarction. *European Journal of Preventive Cardiology*, **25**, 985–993, doi:10.1177/2047487318766646.

Rafieian-Kopaei, M., M. Setorki, M. Doudi, A. Baradaran, and H. Nasri, 2014: Atherosclerosis: process, indicators, risk factors and new hopes. *International journal of preventive medicine*, **5**, 927.

Scutari, M., 2010: Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, **35**, 1–22, doi:10.18637/jss.v035.i03.

Scutari, M. and J. B. Denis, 2014: *Bayesian networks: with examples in R*. Chapman & Hall.

Scutari, M., C. E. Graafland, and J. M. Gutiérrez, 2019: Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, **115**, 235–253, doi:10.1016/j.ijar.2019.10.003.

Shameer, K., M. A. Badgeley, R. Miotto, B. S. Glicksberg, J. W. Morgan, and J. T. Dudley, 2017: Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. *Briefings in bioinformatics*.

Stefanick, M. L., S. Mackey, M. Sheehan, N. Ellsworth, W. L. Haskell, and P. D. Wood, 1998: Effects of diet and exercise in men and postmenopausal women with low levels of hdl cholesterol and high levels of ldl cholesterol. *New England Journal of Medicine*, **339**, 12–20.

Steinhubl, S. R. and E. J. Topol, 2015: Moving from digitalization to digitization in cardiovascular care: why is it important, and what could it mean for patients and providers? *Journal of the American College of Cardiology*.

Sun, F., J. Yao, S. Du, F. Qian, A. A. Appleton, C. Tao, H. Xu, L. Liu, Q. Dai, B. T. Joyce, D. R. Nannini, L. Hou, and K. Zhang, 2023: Social Determinants, Cardiovascular Disease, and Health Care Cost: A Nationwide Study in the United States Using Machine Learning. *Journal of the American Heart Association*, **12**, e027919, doi:10.1161/JAHA.122.027919.

Vaduganathan, M., G. A. Mensah, J. V. Turco, V. Fuster, and G. A. Roth, 2022: The Global Burden of Cardiovascular Diseases and Risk. *Journal of the American College of Cardiology*, **80**, 2361–2371, doi:10.1016/j.jacc.2022.11.005.

Visseren, F. L. J. and ... et al., 2021: 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal*, **42**, 3227–3337, doi:10.1093/eurheartj/ehab484.

Wallace, J. P., 2003: Exercise in hypertension: a clinical review. *Sports medicine*.

Weintraub, W. S., 2023: High costs of cardiovascular disease in the European Union. *European Heart Journal*, **44**, 4768–4770, doi:10.1093/eurheartj/ehad587.

Weng, S. F., J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, 2017: Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, **12**, e0174944.

Whitman, I. R., V. Agarwal, G. Nah, J. W. Dukes, E. Vittinghoff, T. A. Dewland, and G. M. Marcus, 2017: Alcohol abuse and cardiac disease. *Journal of the American College of Cardiology*, **69**, 13–24.

WHO, 2023: Cardiovascular diseases - world health organization. Accessed: 2023-07-01. URL https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Wilkinson, M. D. and ... et al., 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018, doi:10.1038/sdata.2016.18.

Appendix A: Curated Database Variable Description

Variable	Description	Categories	Percentages
MH_AGE	Age	1 (Less than 60) 2 (60-72) 3 (More than 72)	33.06% 35.06% 31.89%
MH_SEX	Patient sex	1 (Female) 2 (Male)	23.09% 76.91%
MH_BMI	Body Mass Index (BMI)	1 (Underweight) 2 (Normal) 3 (Overweight) 4 (Obesity)	0.17% 20.47% 49.40% 29.97%
MH_FHxIHD	Family history of ischemic heart disease	0 (No) 1 (Yes)	60.53% 39.47%
MH_SMK	Smoking status	1 (Current) 2 (Ex-smoker) 3 (Never smoked)	26.97% 45.94% 27.09%
MH_DM	Diabetes mellitus	0 (No) 1 (Yes)	70.53% 29.47%
MH_HTN	Hypertension	0 (No) 1 (Yes)	33.56% 66.44%
MH_DLP	Dyslipidemia	0 (No) 1 (Yes)	37.89% 62.11%
MH_CKD	Chronic kidney disease	0 (No) 1 (Yes)	88.37% 11.63%
MH_HGB	Hemoglobin	1 (Normal levels) 2 (Abnormal levels)	65.53% 34.47%
MH_PMxMI	History of myocardial infarction	0 (No) 1 (Yes)	80.70% 19.30%
MH_PMxPCI	Previous percutaneous coronary intervention	0 (No) 1 (Yes)	72.57% 27.43%
MH_PAD	Peripheral arterial disease	0 (No) 1 (Yes)	89.83% 10.17%
MH_STK	Stroke	0 (No) 1 (Yes)	94.25% 5.75%

Table A1: *Description of variables in the medical history dataset (MH) and their adapted categories. Percentages indicate the prevalence of each case within the database.*

Variable	Description	Categories	Percentages
TRT_IND	Indication for procedure	1: Stable angina 2: ST elevation myocardial infarction 3: Others	27.47% 27.76% 44.77%
TRT_ACC	Type of access for treatment	1: Femoral 2: Radial	53.90% 46.10%
TRT_LCA	Treatment of the left coronary artery	0: No 1: Yes	89.29% 10.71%
TRT_LAD	Treatment of the left anterior descending artery	0: No 1: Yes	40.64% 59.36%
TRT_RCA	Treatment of the right coronary artery	0: No 1: Yes	49.69% 50.31%
TRT_LCX	Treatment of the circumflex artery	0: No 1: Yes	61.19% 38.81%
TRT_DEC	Treatment decision	1: ACTP 2: Surgery 3: Conservative	76.20% 5.59% 18.22%
TRT_EF	Ejection fraction	0: Good: 50-70 1: Bad: Different from 50-70	57.52% 42.48%

Table A2: *Same as Table A1 but for the variables in the treatment dataset (TRT).*

Variable	Description	Categories	Percentages
SOC_MAR_ST	Marital status (has a partner)	1: Married 2: Others	70.40% 29.60%
SOC_LIV_ALN	Lives alone	1: No 2: Yes	85.58% 14.42%
SOC_SUPP	Family support	0: No 1: Yes	2.54% 97.46%
SOC_RES	Place of residence	1: Rural 2: Urban	56.90% 43.10%
SOC_EDU	Education	0: None 1: With education	6.59% 93.41%
SOC_ACT_EMP	Active employment	1: No 2: Yes	72.82% 27.18%
SOC_WOR_TYPE	Type of work	1: White-collar 2: Blue-collar	30.80% 69.20%
SOC_EX	Exercise	0: No 1: Yes	35.43% 64.57%
SOC_SALT_DIET	Salt diet	1: Salt-free diet 2: Diet with salt	52.40% 47.60%
SOC_ALC	Alcohol consumption	1: No 2: Weekend 3: Daily	50.90% 12.88% 36.22%
SOC_SOC_MED	Social media	0: No 1: Yes	72.32% 27.68%
SOC_MOB_PH	Mobile phone	0: No 1: Yes	10.92% 89.08%
SOC_A_TRT	Adherence to treatment	0: Good 1: Bad	88.70% 11.30%

Table A3: *Same as Table A1 but for the variables in the social dataset (SOC) and their categories.*

Appendix B: Associations between variables

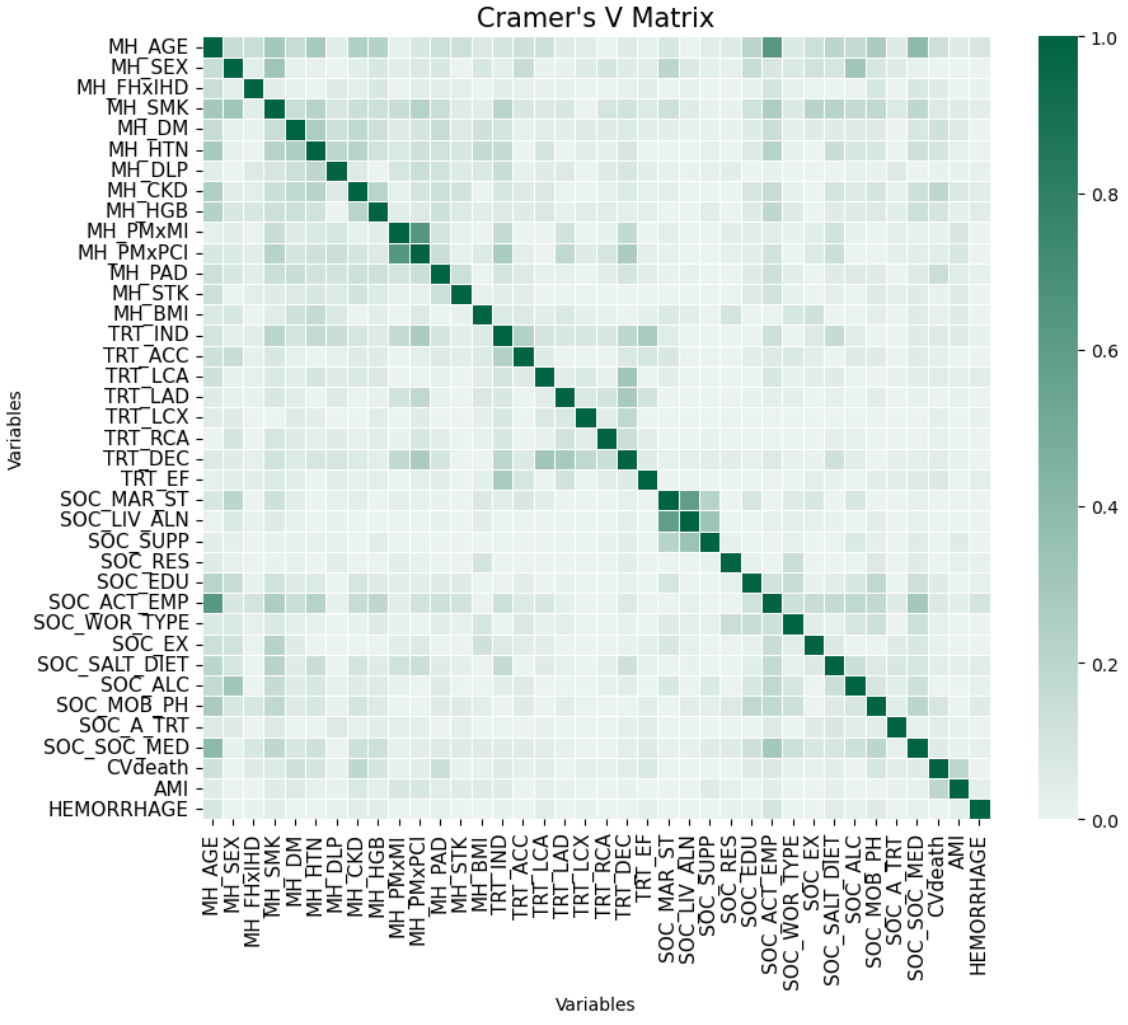

Figure B1: *Cramer's V Matrix describing the association between variables.*

To represent the correlation matrix (Fig. B1), Cramer's V coefficient is used, a measure of association for categorical variables that ranges between 0 and 1.

The graph highlights a strong association between a history of previous heart attack

and percutaneous coronary intervention. Additionally, it shows a significant relationship between marital status and living alone, as well as a connection between age and active employment.

A contingency table is constructed, showing the observed frequencies of category combinations for each pair of variables. For example, for `MH_AGE` and `SOC_ACT_EMP`:

	SOC_ACT_EMP = 1	**SOC_ACT_EMP** = 2
MH_AGE = 1	266	527
MH_AGE = 2	727	114
MH_AGE = 3	754	11

The chi-square statistic is then calculated for the contingency table:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where O_{ij} are the observed frequencies and E_{ij} are the expected frequencies under the null hypothesis of independence.

In this case, the chi-square statistic is $\chi^2 = 953.12$.

Subsequently, Cramér's V coefficient is calculated:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

Where n is the total sample size (2399), k is the number of rows in the contingency table (3), and r is the number of columns in the contingency table (2).

Therefore, Cramér's V coefficient is:

$$V = \sqrt{\frac{953.12/2399}{\min(3-1, 2-1)}} = \sqrt{\frac{0.397}{1}} = 0.63$$

Appendix C: Other classification models

Model intercomparison and evaluation

Evaluating the quality of a predictive model isn't a one-score-fits-all scenario; various metrics shed light on different facets of the correlation between actual and forecasted values. This complexity escalates when dealing with probabilistic predictions (Broecker, 2011). Compared to deterministic predictions, quantitative probabilistic predictions offer several benefits, including the ability to convey valuable information by introducing a concept of probability or associated cardiovascular risk. In terms of management applications, probabilistic metrics offer greater adaptability to end users of the model, enabling them to establish probability thresholds that align best with their specific objectives. Consequently, in our experiment to assess predictive accuracy performance, we will evaluate the model's classification capability using the Area Under the ROC Curve (AUC) for model intercomparison.

In the case of probabilistic predictions, the receiver operating characteristics (ROC) curve is commonly used as a generalization of the above validation procedure to describe the accuracy of the model (Mandrekar, 2010). Probabilities above/below a certain probability threshold u are set to positive/negative. The ROC curve describes the predictive ability of the system under the whole range of probability thresholds, thus representing a global measure of model performance. The area enclosed under the ROC curve (AUC), which ranges from 1 (perfect prediction) to 0 (completely inverted prediction), passing through 0.5 (random prediction), provides a quantitative measure of model performance. This curve is defined by plotting the sensitivity(u) versus 1-specificity(u) values for the deterministic prediction. Thus AUC is to be preferred as a measure of model accuracy when interest is focused in comparing and ranking the performance of different classifiers (see e.g. Fielding and Bell, 1997; Bedia et al., 2011).

Note that all the classifiers have been trained with the same balanced datasets, prepared as detailed in Sec. 4.3, in order to obtain a fair comparison in terms of their classification ability.

Ultimately, the models were evaluated ten times, with each iteration involving a different subsampling of the balanced dataset and applying Leave-One-Out Cross-Validation (LOOCV) to calculate the AUC (Kohavi, 1995). In each LOOCV sub-iteration, the dataset was divided into a training set and a test set, and each model was trained and evaluated. After completing all the sub-iterations for each subsampling iteration, the AUC was calculated. Once all iterations were completed, graphs were generated to visualize the distribution of the AUC and the average ROC curves.

Classification algorithm description

The models used included Bayesian networks, random forest, and Naive Bayes, a classification algorithm based on Bayes' theorem that assumes independence between features. This algorithm calculated the probability that an instance belonged to a given class based on the joint probability of the features and assigned the class with the highest posterior probability (Ng and Jordan, 2001). SVM was also utilized, which sought to find the optimal hyperplane that maximized the separation (margin) between classes in a feature space, using data points close to the margin (support vectors) to define and orient the hyperplane (Cortes and Vapnik, 1995). Additionally, the K-nearest neighbors method (KNN, Lorenz, 1969) was employed, which assigned a class to an instance based on the classes of its K nearest neighbors in the feature space (Lorenz, 1969).

In Bayesian Networks, the hill-climbing (hc) algorithm was used to learn the model structure (see Sec. 3.2). The Random Forest was configured with 100 trees (Sec. 4.1). SVM with a radial kernel was also employed to classify data points and predict class probabilities for new data, converting these probabilities into binary class labels based on a threshold.

The K-Nearest Neighbors (KNN) model, which assigns the class of an unknown sample based on the classes of its k nearest neighbors, was configured with 5 neighbors. Naive Bayes, assuming independence between features, calculates the posterior probability of

each class using Bayes' theorem, assigning the new data to the class with the highest posterior probability based on the likelihood of observing the features given each class and the prior probability of the class (Ng and Jordan, 2001).

Intercomparison results

Model	**CVdeath**	**AMI**	**HEMORRHAGE**
Bayesian Network	0.7073	0.5354	0.5857
Random Forest	0.7087	0.5625	0.5687
SVM	0.7042	0.5700	0.5756
KNN	0.6467	0.5189	0.5424
Naive Bayes	0.6899	0.5726	0.6004

Table C1: *AUC Results for the 5 Models*

Model	**CVdeath (5th-95th)**	**AMI (5th-95th)**	**HEMORRHAGE (5th-95th)**
Bayesian Network	0.6562 - 0.7560	*0.4022 - 0.6296*	0.5145 - 0.6329
Random Forest	0.6797 - 0.7413	0.5318 - 0.5990	0.5303 - 0.6253
SVM	0.6733 - 0.7340	0.5116 - 0.6152	0.5375 - 0.6115
KNN	0.6097 - 0.6771	*0.4966 - 0.5447*	0.5087 - 0.5829
Naive Bayes	0.6729 - 0.7170	0.5362 - 0.6105	0.5687 - 0.6272

Table C2: *AUC Results (Bootstrapped confidence interval of 90%). Essentially random predictions (containing 0.5 in their AUC interval) are indicated in italics. These AUC interval results were obtained from the ROC curved depicted in Figs. C2, C4 and C6 for the classification of death (CVdeath), stroke (AMI) and hemorrhage respectively.*

The AUC values obtained for the five models have been compiled. It was found that, for cardiovascular death, the best results were achieved with Random Forest, SVM, and Bayesian networks, all showing very similar results. This suggests that Bayesian networks could effectively compete with more popular algorithms like Random Forest and SVM. Regarding the AMI variable, the models did not demonstrate significant discriminative capability. The best results were obtained with Naive Bayes classifier and SVM. Lastly, for hemorrhage, slightly better generalization capabilities were observed, with Naive Bayes again performing the best followed by SVM (Table C1).

The AUCs obtained for predicting AMI and HEMORRHAGE (Table C1) are low compared to those for predicting CVdeath. This poor performance in predicting AMI and HEMORRHAGE is likely due to the limited amount of data available for these events.

The small dataset size can negatively impact the models' ability to generalize and accurately predict these outcomes. Notably, Naive Bayes is one of the models that performs comparatively better under these conditions.

To reinforce the validity of predictive models, bootstrapped confidence intervals at 90% were obtained (Table C2). For cardiovascular death, Bayesian networks performed with a range of 0.6562 to 0.7560, while Random Forest showed more consistent and higher performance, ranging from 0.6797 to 0.7413. SVM exhibited performance similar to Random Forest but with a slightly wider interval, ranging from 0.6733 to 0.7340. KNN had the lowest performance, with an interval between 0.6097 and 0.6771, whereas Naive Bayes showed a range of 0.6729 to 0.7170, also demonstrating consistency (Figure C1).

In terms of AMI, both Bayesian networks and KNN included the value 0.5 in their accuracy, indicating they did not outperform a random classifier. For Random Forest, SVM, and Naive Bayes, results did not significantly surpass this benchmark (Figure C3).

Regarding hemorrhage, KNN performed similarly to a random classifier, while the other four models achieved slightly higher results compared to those obtained for AMI (Figure C5).

When evaluating the ROC curves across the five models, it is observed that Random Forest tends to produce ROC curves that are less steep or flatter compared to the other models. This is due to its ability to combine multiple decision trees into a final prediction, providing a more robust classification that is less sensitive to small variations in the decision threshold (Figures C2, C4, C6).

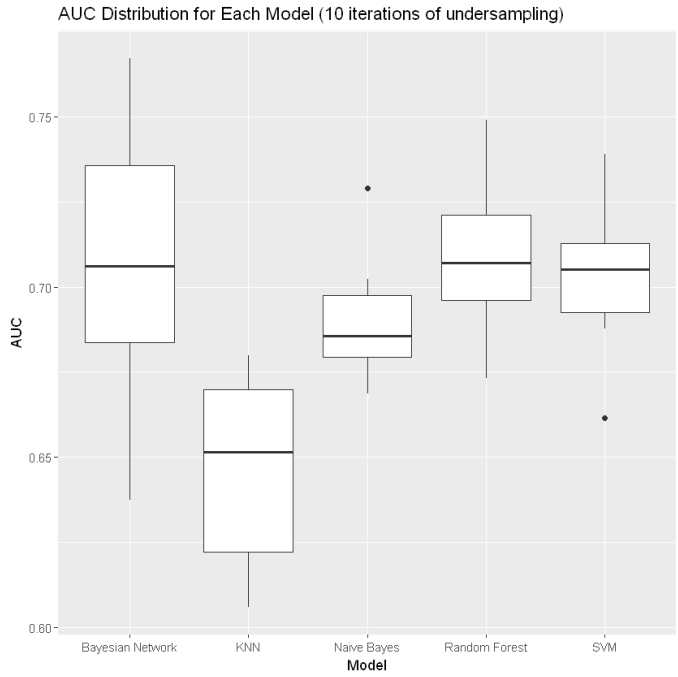

Figure C1: *AUC distribution for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting CVdeath. Bootstrap resampling with 10 iterations.*

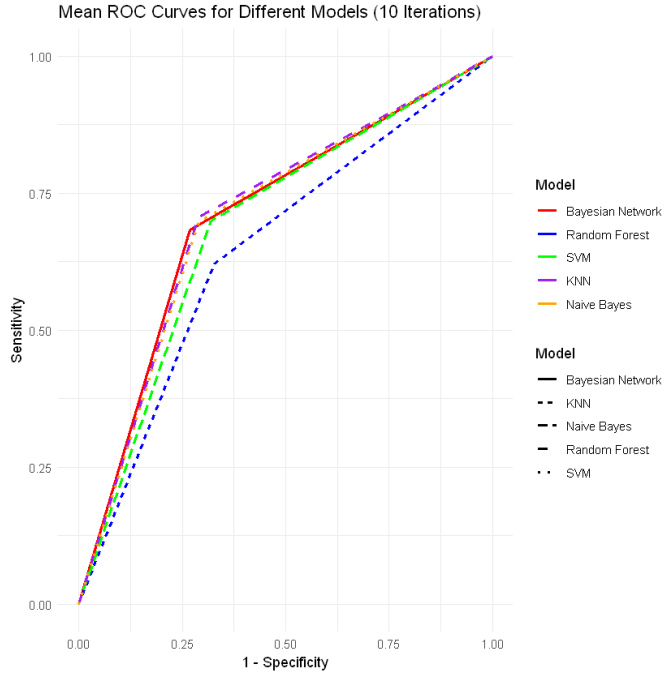

Figure C2: *Mean ROC Curves for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting CVdeath*

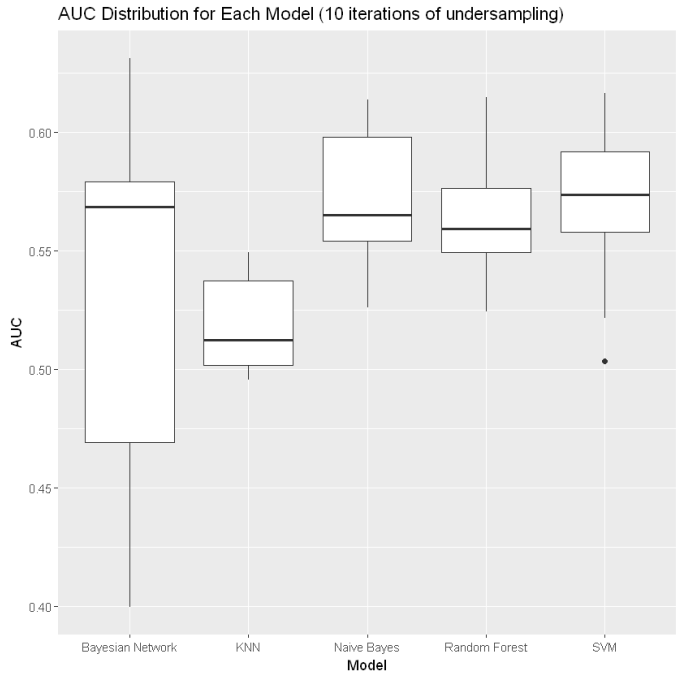

Figure C3: *AUC distribution for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting AMI. Bootstrap resampling with 10 iterations.*

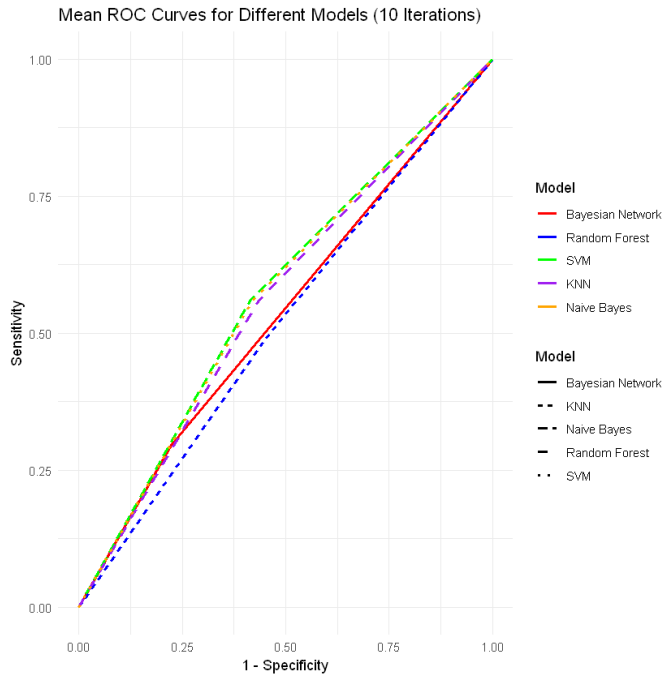

Figure C4: *Mean ROC Curves for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting AMI*

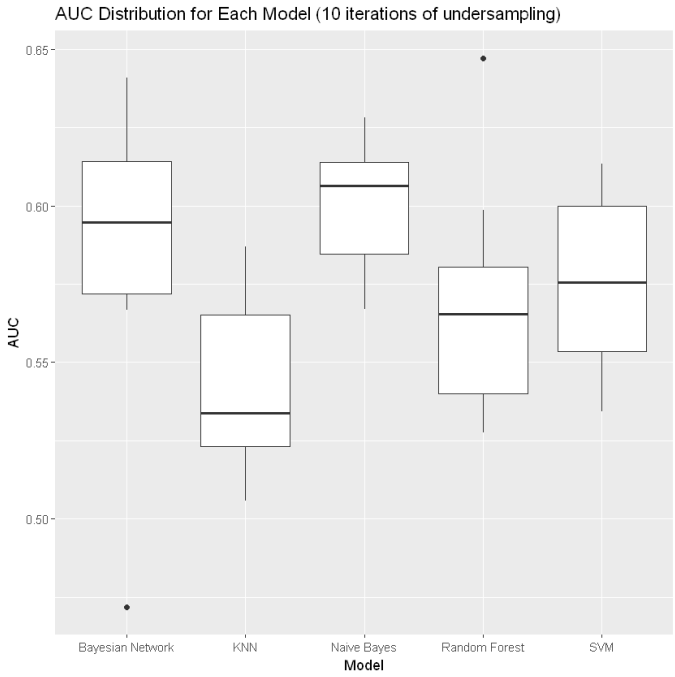

Figure C5: *AUC distribution for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting Hemorrhage. Bootstrap resampling with 10 iterations.*

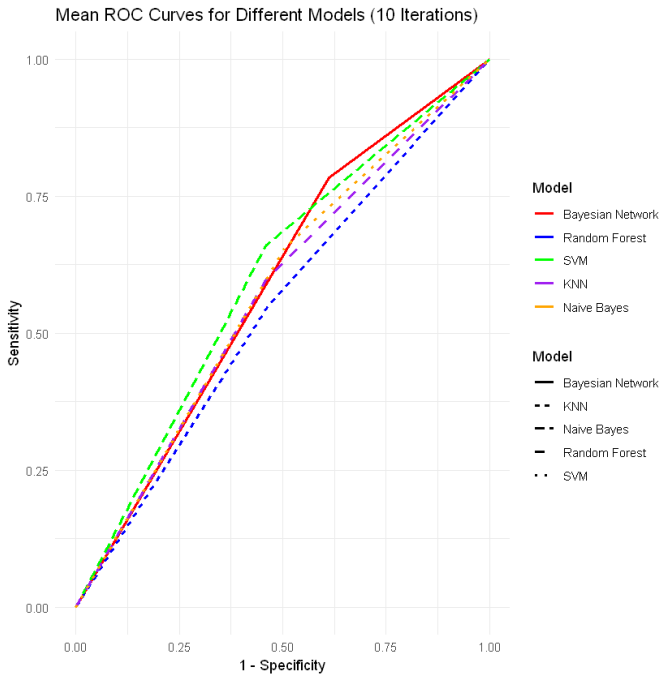

Figure C6: *Mean ROC Curves for Bayesian Network, Random Forest, SVM, KNN, and Naive Bayes Models Predicting Hemorrhage*