Modelling cardiovascular risk using Bayesian Networks

Matteo Tarenzi

May 5, 2025

Introduction

Coronary heart disease is one of the leading causes of death worldwide. It occurs when the coronary arteries, which supply blood to the heart, narrow or block, leading to heart attacks and other cardiovascular complications. Research has shown that the risk of developing cardiovascular disorders is influenced by a combination of genetic, lifestyle, and environmental factors. In recent years, statistical models have become increasingly valuable tools for the prevention of such diseases. The aim of this project is to develop a Bayesian network that models the key factors that affect the risk of developing coronary heart disease. Thanks to the probabilistic and graphical characteristics of this type of model, the results can be used to assess the influence of multiple variables on cardiovascular risk and allow a better understanding of the relationships between these features.

1 The Dataset

The data for the project were obtained from *Kaggle*. *Kaggle* is a popular website and platform for machine learning where data science enthusiasts can download dataset, models, and also share their project. Data were downloaded and then loaded using pandas into a dataframe. Originally, it contained 16 features related to the medical information of the patients. The projects decided to utilize only nine of them, keeping the most important ones and eliminating the ones that were less interesting. In the following, they are presented with a brief description.

Cardiovascular risk dataframe:

- **Age** : the age of the patient;
- **TotChol**: the level of the total cholesterol;
- **SysBP**: the measured systolic blood pressure;
- **DiaBP**: the measured diastolic blood pressure;
- **BMI**: the Body Mass Index;
- **CigsPerDay**: the amount of smoked cigarettes per day;
- **Male**: the gender of the patient;
- **Diabetes**: whether the patient is diabetic or not;
- **TenYearCHD**: the prediction of coronary heart disease.

Data come from 4000 patients who are not under prescription for blood pressure medications. The features describe the cardiovascular condition of each individual. In particular, the variable "*TenYearCHD*" represents a prediction made by doctors regarding the likelihood of developing coronary heart disease within the next ten years.
The dataset includes both continuous and categorical variables. The categorical features are all binary, indicating gender, the presence of diabetes, and the prediction of coronary heart disease. Continuous variables are measured using their respective medical units.

2 Methodology

This section outlines the methodology adopted in the project. First, it introduces the ideas behind the data engineering process, including both the modification of existing features and the creation of new ones. Next, it provides an in-depth discussion of the construction of the Bayesian network and the estimation of the conditional probability distributions (CPDs). Finally, it presents the classification phase, focusing on predictions related to the risk of coronary heart disease.

Data Engeneering

As previously mentioned, the dataset includes both categorical and continuous variables. For categorical features, no major modifications were made, except for converting numeric labels back into natural language for improved interpretability. For example, in the case of *Gender*, the value '1' was mapped to '*Male*' and '0' to '*Female*'. This choice was made to simplify the interpretation of the conditional probability distributions (CPDs) during the variable elimination process.
In contrast, continuous variables required more careful preprocessing. **Figure 1** illustrates the relative frequency distributions, comparing the distributions of individuals labelled as "*Safe*" with those labelled as "At-*Risk*".

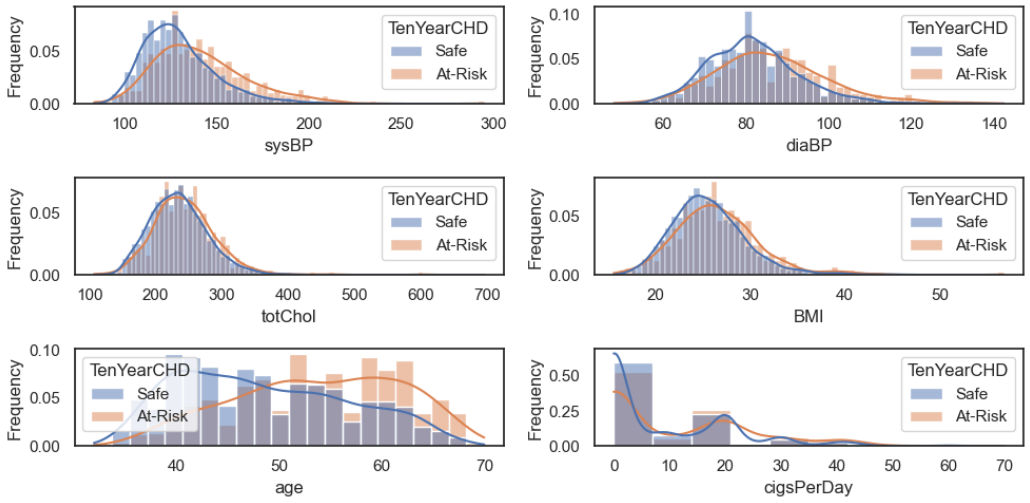


Figure 1: Continuous variables frequency distributions

From the plots, it is already possible to infer some insights about the influence of these variables on heart health. From an exploratory point of view, it is also noteworthy that, except for *Age* and *CigsPerDay*, all other variables exhibit a pronounced long tail. To determine whether these represent actual distribution tails rather than outliers, a further detailed analysis was performed using box plots. **Figure 2** displays the box plots for the four continuous variables under consideration.

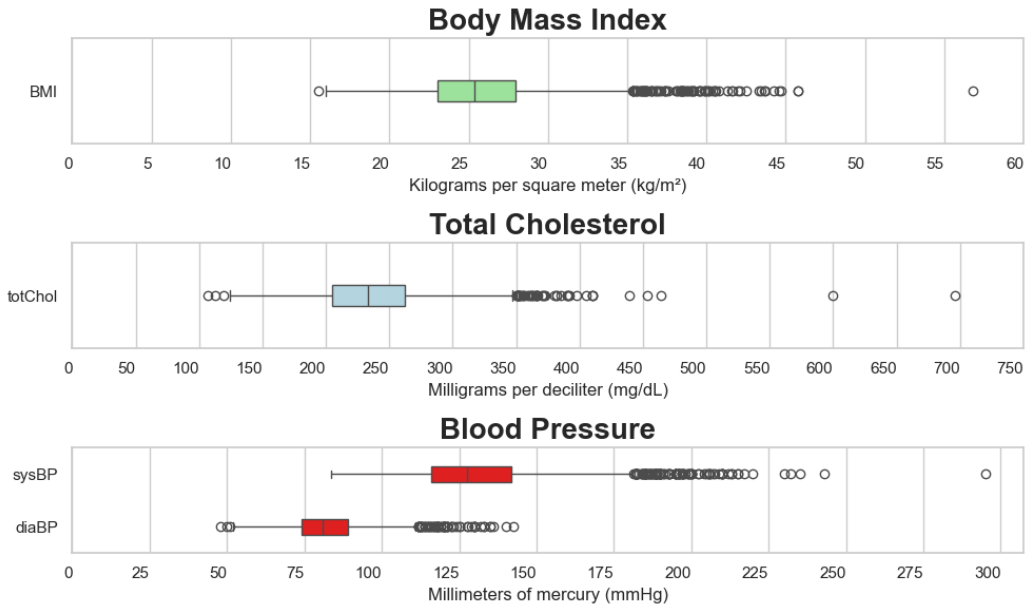


Figure 2: Continuous variables box plots

It is possible to observe a significant number of data points outside the whiskers, particularly on the right side. These points represent the long tails of the distributions. However, some extend even farther and can reasonably be classified as outliers. Based on a combination of visual inspection and general medical reasoning, these outliers were removed by manually setting upper thresholds for each variable.
For Body Mass Index (BMI), a threshold of 50 $\mathrm{kg/m^2}$ was applied. This corresponds to approximately 144 kg for an individual who is 1.70 m tall. For total cholesterol, the threshold was set at 425 mg/dL, and for systolic blood pressure, the threshold was defined at 225 mmHg.

Based on both medical knowledge and the distribution shapes observed before, data points exceeding these values are considered extreme and were excluded from the analysis.
The goal of the project is to create a discrete Bayesian network. For this reason, the continuous variables within the dataset needed to be discretized. For each of them, a categorical variable was created using binning, based on medical charts and information. Similarly to the categorical variables already present in the dataset, the newly created features use labels that are easy to understand, which facilitates a clearer interpretation of the variable elimination results.

Variable	**Threshold Range**	**Label**
Age	30–44	30–44
	45–54	45–54
	55–64	55–64
	65–75	65–75
Total Cholesterol	100–199	Optimal
	200–239	Elevated
	240–425	High
Blood Pressure (sysBP/diaBP)	≤ 120 and ≤ 80	Optimal
	≤ 140 and ≤ 90	Elevated
	≤ 180 and ≤ 120	Hypertension
	> 180 or > 120	Crisis
BMI	14–18.4	Underweight
	18.5–24.9	Normal
	25–29.9	Overweight
	30–50	Obese
CigsPerDay	0	NonSmoker
	1–5	LightSmoker
	6–20	ModerateSmoker
	21–50	SevereSmoker

Table1: Construction of the new categorical variables

Except for *Age* and *CigsPerDay*, the other variables follow precise binning schemes suggested by researchers and medical institutions. The binning process for these two variables, instead, took a different approach. For *Age*, the

binning is straightforward: the data range is divided into smaller intervals of fifteen years. The binning for *CigsPerDay*, on the other hand, is based on the number of cigarettes in a package. Assuming a package contains 20 cigarettes, the project defines a light smoker as someone who smokes a quarter of a package per day, a moderate smoker as someone who smokes one package per day, and a severe smoker as someone who smokes more than one package per day.
During the binning process, another important variable was created: *Blood Pressure*. This feature combines systolic and diastolic measurements to describe the blood pressure health of patients. A critical label is *Hypertension*, which, according to doctors, is one of the leading causes of heart disease among patients. The binning process has made all the variables much easier to read and understand, allowing even people without medical knowledge to grasp the content of each variable. For example, not everyone can distinguish different BMI levels simply by looking at the raw values. However, once the variable is transformed using the new binning, it becomes much clearer and more accessible to a broader audience.

Model creation and variable elimination

In order to define the best structure for the Bayesian network, three different methods were tested and then compared.
The first method implements a score-based approach. For evaluating candidate networks, the Bayesian Information Criterion (BIC) was selected as the scoring function, while Hill Climb was chosen as the search strategy. Hill climb search implements a greedy local search that starts from an initial DAG (by default, a disconnected graph) and iteratively performs single-edge modifications that most improve the score. The search ends once a local maximum is reached. To guide the process and improve the resulting structure, a whitelist and a blacklist of edges were defined before starting the search.
The second method manually defines the Bayesian network, relying on domain knowledge. This structure was constructed by consulting academic literature, discussing with a doctor (a friend of mine), and conducting conditional independence tests. The goal was to build a network that aligns with established medical understanding.
The third approach applies a hybrid technique that combines automatic conditional independence testing with score-based search. The structure is learned in two main steps. First, an undirected graph skeleton is learned us-

ing the constraint-based MMPC (Max-Min Parents and Children) algorithm, which identifies candidate parents and children based on conditional independence tests. Then, the edges are oriented using a score-based optimization technique. Specifically, a modified hill climbing algorithm is employed in combination with the Bayesina Information Criterion (BIC) score to determine the optimal directed structure.
Finally, the three resulting networks were compared using the BIC. The model with the lowest score was selected and used for inference. To obtain meaningful information about the risk of coronary disease, variable elimination was performed. This inference algorithm allows one to compute marginal probabilities for a target variable given observed evidence on other variables in the network. Systematically, it eliminates irrelevant variables through marginalization and conditioning, focusing the computation only on those that influence the target. By providing evidence on specific variables, such as smoking habits, blood pressure, or BMI, the network is able to update the belief about the probability of developing coronary disease. This makes it possible to simulate and analyze a variety of clinical scenarios and better understand how combinations of risk factors affect the likelihood of heart conditions.

Performing classification with the model

The second application tested for the Bayesian network is classification. The main objective was to evaluate the predictive capabilities of the model and to assess its practical applicability. In this task, *TenYearCHD* serves as the target variable, while all other features are used as evidence for the variable elimination algorithm.
The dataset was first divide into three subsets: training, validation, and test sets. Due to the highly imbalanced nature of the dataset, SMOTEN (Synthetic Minority Over-sampling Technique for Nominal features) was applied to the training set to address the class imbalance. SMOTEN is a variant of SMOTE specifically designed for categorical data; it generates synthetic samples for the minority class by interpolating between existing instances. The resulting training set was then used to fit the Bayesian network by estimating the Conditional Probability Distributions (CPDs).
Subsequently, the validation set was used to determine the optimal classification threshold. Each observation in the validation set was passed through the variable elimination algorithm, using *TenYearCHD* as the query vari-

able and all other features as evidence. Once the posterior probabilities were computed, the best threshold was selected by maximizing the *F1-score* for the "*At-Risk*" class, iterating over multiple values.
Finally, the test set was used to assess the model performance. Predictions were generated using the previously defined threshold and standard classification metrics such as accuracy, precision, recall, and the F1 score were calculated to evaluate the effectiveness of the Bayesian network. **Figure 3** below, shows the classification workflow with the operations of each subset.

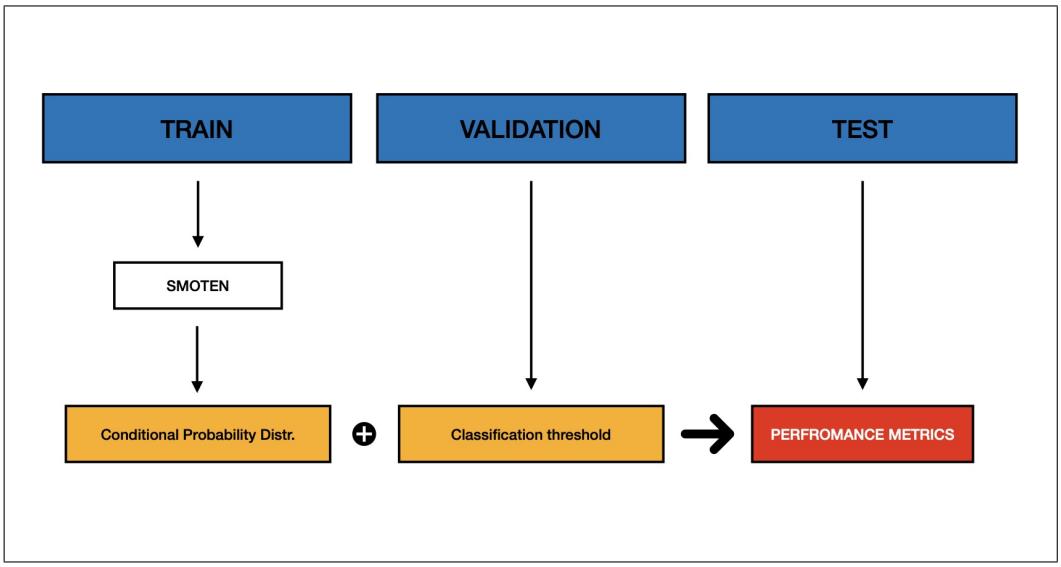


Figure 3: Prediction workflow

3 The Results

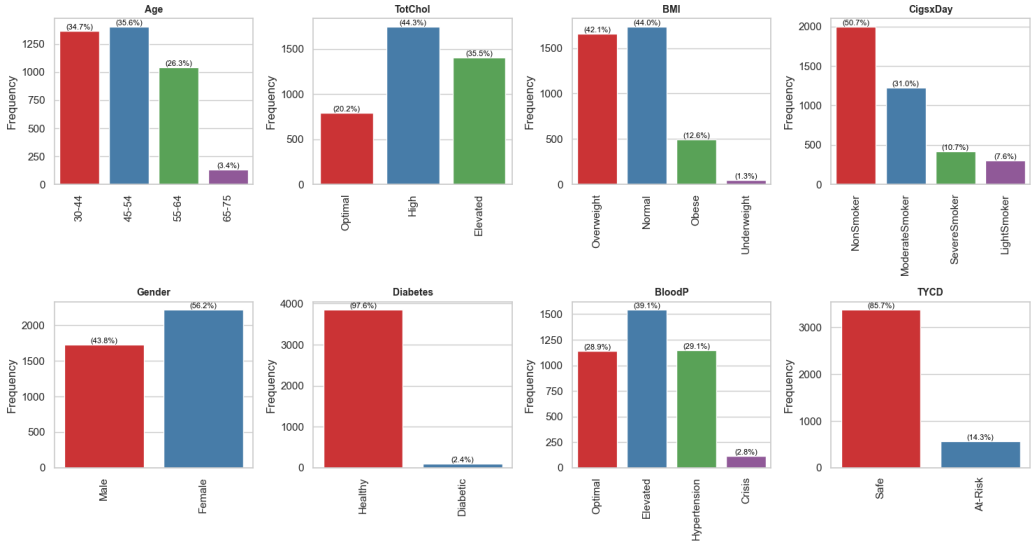


Figure 4: Data exploration for categorical features

Figure 4 shows frequency bar plots for the refined categorical variables in the dataset. Although many observations can be made from the plot, this dissertation will focus on the most significant ones.
Based on frequencies, the typical individual is a woman between the ages of 45 and 54 who does not smoke. However, this individual suffers from high cholesterol levels and elevated blood pressure. Getting into detail, we observe how unbalanced the dataset is for coronary disease prediction. Individuals diagnosed as at risk make up only 15%. Although this presents some challenges for the estimations, the predictions seem realistic given the type of feature. Another notable aspect is that, considering total cholesterol levels and blood pressure, the number of people with harmful conditions is higher than that of people with healthy conditions. In particular, when looking at blood pressure, the number of individuals suffering from hypertension exceeds that of those with normal blood pressure. The reason behind this could be attributed to the increase in poor eating habits, as seen in the BMI data, where overweight individuals are almost as numerous as those considered in good shape.

Network structure

As mentioned in the *Methodology* section, the project tested three different methods for the construction of the Bayesian network structure. From the criterion, the best one is the network defined using academic literature and medical notions. Here are the BIC evaluation for the three:

	Hill Climb search	Hybrid method	Custom method
BIC	-265198.28	-26248.75	-28221.91

The structure of the best model can be observed in **Figure 5**. *Age* and *Gender* are independent variables that influence both *Body Mass Index (BMI)* and *TotChol*. Cigarette consumption (measured as cigarettes per day) also plays a central role, directly affecting BMI and contributing to the risk of developing diabetes in conjunction with BMI. In addition, diabetes, BMI, total cholesterol, and cigarette consumption have an impact on blood pressure, a central variable in the project. Finally, the probability of developing coronary disease within ten years (TYCD) is influenced by blood pressure, diabetes, cigarette consumption, and age. This structure reflects a plausible chain of dependencies in which demographic and lifestyle factors affect intermediate health conditions, which in turn contribute to long term heart disease risk.

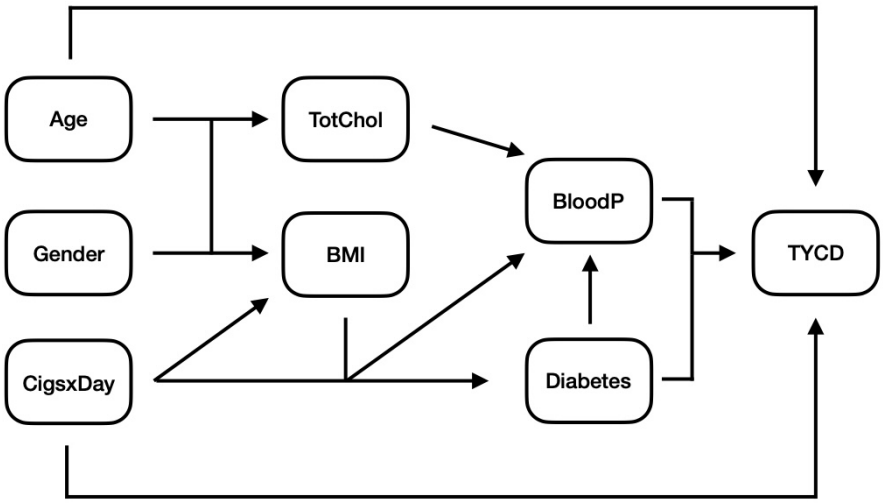


Figure 5: Bayesian Network structure

Variable elimination

In this subsection, the key findings obtained using the variable elimination algorithm will be discussed. For each variable of interest, a table containing the conditional probability distributions will be presented, followed by a detailed discussion.

Healthy		**Diseased**	
Normal, Optimal, NonSmoker		**Obese, Crisis, SevereSmoker**	
TYCD(At-Risk)	0.0603	TYCD(At-Risk)	0.5668
TYCD(Healthy)	0.9397	TYCD(Healthy)	0.4332

Table 2: Good habits vs Bad habits

The first insight focuses on comparing good habits with bad habits. This table illustrates how significantly the choices people make can impact the probability of developing a cardiovascular complications. On the left, the table shows the conditional probability distribution (CPD) for an individual with a healthy BMI, normal blood pressure, and no smoking habits. On the right, the risk increases to almost 57% for an individual who is obese, in a hypertension crisis, and a severe smoker.

TYCD / Blood P	**Optimal**	**Elevated**	**Hypertension**	**Crisis**
At-Risk	0.103009	0.123039	0.195868	0.284912
Healthy	0.896991	0.876961	0.804132	0.715088

Table 3: The influence of blood pressure on TYCD

The most important feature is *Blood Pressure*. The table clearly shows that an increase in blood pressure, particularly when accompanied by conditions such as hypertension, is a key factor in the prevention of coronary disease. In fact, with blood pressure alone, we can observe a significant increase in the probability of risk. For individuals with optimal blood pressure, the probability of developing disorders is quite low, just 10%. As blood pressure increases, the probability of risk also increases, reaching nearly 20% for individuals with hypertension and 28% for those in a hypertension crisis.

Since blood pressure is one of the most important factors in the prevention of coronary heart disease, the next step was to investigate the variables that influence blood pressure. Based on the defined structure, blood pressure depends on four key features: *BMI, Diabetes, CigsxDay*, and *TotChol*. The dissertation presents the findings related to BMI.

BloodP / BMI	**Underweight**	**Normal**	**Overweight**	**Obese**
Optimal	0.510524	0.397638	0.215197	0.118598
Elevated	0.332267	0.397958	0.412098	0.320323
Hypertension	0.119174	0.189167	0.342849	0.491304
Crisis	0.038035	0.015237	0.029857	0.069775

Table 5: The influence of BMI on blood pressure

From **Table 5**, it is clear that an increase in the Body Mass Index (BMI) consistently results in a worsening of blood pressure status. Although individuals in the underweight category are not in optimal condition, they appear to have fewer problems than those in other categories. Although a hypertension crisis remains unlikely probable, overweight individuals (who are nearly as numerous as those with a normal BMI) are more likely to have elevated blood pressure or even hypertension.
Two aspects are particularly concerning. First, obese individuals have an almost 50% probability of having hypertension, a dangerous condition for the prevention of coronary heart disease. Secondly, for patients with normal blood pressure, the probability of developing elevated blood pressure is just as high as the probability of maintaining healthy levels. This highlights an emerging problem, where diet-related disorders are increasingly contributing to an increase in blood pressure levels and subsequently in the rates of heart disease and related complications.

TYCD / Smoke	**NonSmoker**	**Light**	**Moderate**	**Severe**
At-Risk	0.115254	0.107719	0.163482	0.240637
Healthy	0.884746	0.892281	0.836518	0.759363

Table 6: The influence of smoking levels on TYCD

The final insight focuses on smoking. It is already well documented in various studies that smoking has negative effects on health. **Table 6** clearly demonstrates the harmful impact of this habit on the risk of coronary heart

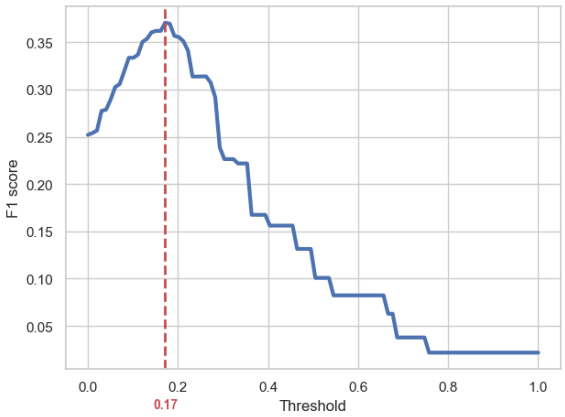
BloodP / Smoke	NonSmoker	Light	Moderate	Severe
Optimal	0.240764	0.290403	0.364909	0.281417
Elevated	0.391468	0.399165	0.387654	0.414368
Hypertension	0.330272	0.283249	0.229294	0.286951
Crisis	0.037496	0.027183	0.018143	0.017264

Table 7: The influence of smoking levels on blood pressure

disease. Although the effects are not as pronounced as with other factors, it is evident that moderate and severe smokers are more likely to develop heart-related complications.
In addition, recent studies have shown that smoking negatively affects blood pressure by influencing both insulin production and insulin resistance. However, the overall effects of smoking on blood pressure remain unclear. This is highlighted in **Table 7**, where the data contradict our expectations. One possible explanation could be the reduction in BMI associated with smoking, but this alone is not sufficient to explain why smokers appear to have better blood pressure compared to non smokers.

3.1 Bayesian network classification

Once the Bayesian network model was created and the CPDs estimated using the training subset, the validation set was used to determine the optimal threshold for classification. By iterating over a range of values, the *F1 score* for the "*At-Risk*" class was computed for each candidate threshold and stored. The threshold that corresponded to the highest *F1 score* was then selected.
From the results, the project selected a classification threshold of 0,17 which corresponded to an F1 score of 0,3706 for the validation set. This threshold was then used to assign labels to the variable elimination predictions made using the test set. Subsequently, the results and the ground truth were used to compute the metrics necessary to evaluate the performance of the models. These metrics are summarized in **Table 8**.
The first point to note is the imbalance of the dataset. Of the 632 individuals, only 94 were identified as "*At-Risk*" by doctors. This imbalance has once again influenced the robustness of the estimates and the performance results. In general, the model demonstrates poor classification capabilities, with an accuracy of only 0,71. However, upon closer inspection of the metrics, we



Class	Precision	Recall	F1-Score	Support
Healthy	0.93	0.71	0.81	538
At-Risk	0.29	0.68	0.41	94
Accuracy				0.71
Macro avg	0.61	0.70	0.61	632
Weighted avg	0.83	0.71	0.75	632

Table 8: Classification Report

can observe that the recall for the "At-Risk" label reaches almost 0,70. This means that 70% of patients who could potentially develop heart disease are correctly classified by the model. This, combined with high precision in classifying healthy patients, represents the most important metric of the model. In a medical context, avoiding false negatives is more crucial than avoiding false positives, as patient health should be the top priority. This is why the classification threshold was selected based on the F1 score for the "*At-Risk*" class, rather than simply using accuracy. However, the low precision of the model could result in higher costs for the healthcare system. With more false positives, the system would require additional exams and visits, subsequently increasing national healthcare expenses.

4 Conclusions

In general, considering the nature of the project, the outcomes are satisfactory. The study successfully highlighted the relationships between patients' health features and their risk of developing coronary heart disease. One of the significant strengths of the project is the accessibility of cardiovascular risk distributions and how they are now available to everyone without requiring advanced medical knowledge. In this regard, the project succeeded in replicating the charts proposed by various health institutes.
However, the performance in classification is somewhat unsatisfactory. Although the model achieved good recall, other performance metrics like precision and accuracy revealed the limitations of this approach. Several factors contribute to these shortcomings.
First, the ground truth proposed by doctors may be flawed or biased, as it is based on their intuition rather than concrete evidence. Secondly, the imbalance in the dataset had a significant impact on the evaluation, affecting both the performance and the robustness of the results. Finally, for the simplicity of evaluation and interpretation, a basic network structure was chosen. This structure, while easy to assess, may not be suitable for capturing the complexity of the problem. Additionally, the Bayesian network itself might not be the ideal model, as in the medical field many variables tend to influence each other. Future studies could assess the classification performance of other models, including combinations of different variables, to improve outcomes.
Despite these limitations, the project remains valuable in its attempt to identify and express the relationships between general health features and the risk of developing coronary heart disease. The results provide useful insights for anyone interested in the issue or looking to establish a foundation for more complex and detailed studies in the future.