UNCOVERING CARDIOVASCULAR DISEASES RISK FACTORS: A GRAPHICAL MODEL APPROACH

Lumia Giulia, Bongiovanni Sabrina Anna, D’Antoni Marco

ABSTRACT

The cardiovascular diseases risk factors were explored by the mean of a graphical model. As a result it was possible to obtain a diagnostical tool with about 70% accuracy, capable of predicting whether a patient has a cardiovascular disease on the base of its characteristics.

INTRODUCTION

This report is aimed to show the results of an analysis carried out in order to highlight the relationships among possible risk factors and produce some assessments about the probability of having a cardiovascular disease.

The available data refer to:

- **Age**;
- **Gender**;
- **Weight**;
- **Height**;

PATIENT FEATURES

- **Physical activity**;
- **Alcohol intake**;
- **Smoking**;

LIFESTYLE

- **Systolic blood pressure**;
- **Diastolic blood pressure**;
- **Cholesterol**;
- **Glucose**;
- **Cardiovascular disease** (presence or absence).

MEDICAL FEATURES

The analysis is developed in four stages:

1. Data preprocessing;
2. Exploratory analysis;
3. Graphical model adaptation and selection;
4. Performance test.

1. DATA PREPROCESSING

The first step of the analysis aims to make the data suitable for the subsequent adaptation of a graphical model. For this purpose continuous variables were discretized on the base of some existing criteria:

- **Blood pressure.**
  Since all the available data refer to adults (the age ranges between 29 and 65 years), it was possible to summarize the information about the systolic and diastolic blood pressure according to the American college of cardiology guidelines:

BLOOD PRESSURE	SYSTOLIC		DIASTOLIC
normal	<120	AND	<80
elevated	[120-130)	AND	<80
hypertension stage 1	[130-140)	OR	[80-90)
hypertension stage 2	>=140	OR	>=90
hypertensive Crisis	>180	AND/OR	>120

- **BMI index.**
  The information about the weight of a person was summarised taking into account also his/her height and sex. The weight value considered alone, in fact, doesn't allow to assess whether a person is underweight or obese.
  As instance, given a weight of 60 kg a 1.75 meters woman can be considered with normal weight while a 1.50 meters woman would be classified as overweight.
  On the base of this well-known index patient can be classified as:

BMI	MAN	WOMAN
underweight	<20,1	< 18,7
normal	[20,1-25,1)	[18,7-23,9)
overweight	[25,1-30)	[23,9-28,7)
obese S1	[30-35,1)	[28,7-35,1)
obese S2	[35,1-40]	[35,1-40]
obese S3	>40	>40

- **Age.**
  Since no unique age cut-off exists, patients were classified on the base of a customised classification as ≤45, (45-55] and >55.

- **Height.**
  The height classification is based on that one proposed by Martin and Saller, taking into account the gender of the patient:

HEIGHT	MAN	WOMAN
dwarf	<1,30	<1,21
short	[1,30-1,60)	[1,21-1,49)
medium	[1,60-1,70)	[1,49-1,59)
tall	[1,70-2,00]	[1,59,1,87]
giant	>2,00	>1,87

Moreover, since data contained explicit errors and medical anomalies, some corrections were needed:
- the minimum value of BMI is 3.47 while the max is 298,67;
- there can't exist cases in which the systolic blood pressure is lower than the diastolic;
- blood pressure can't be negative.

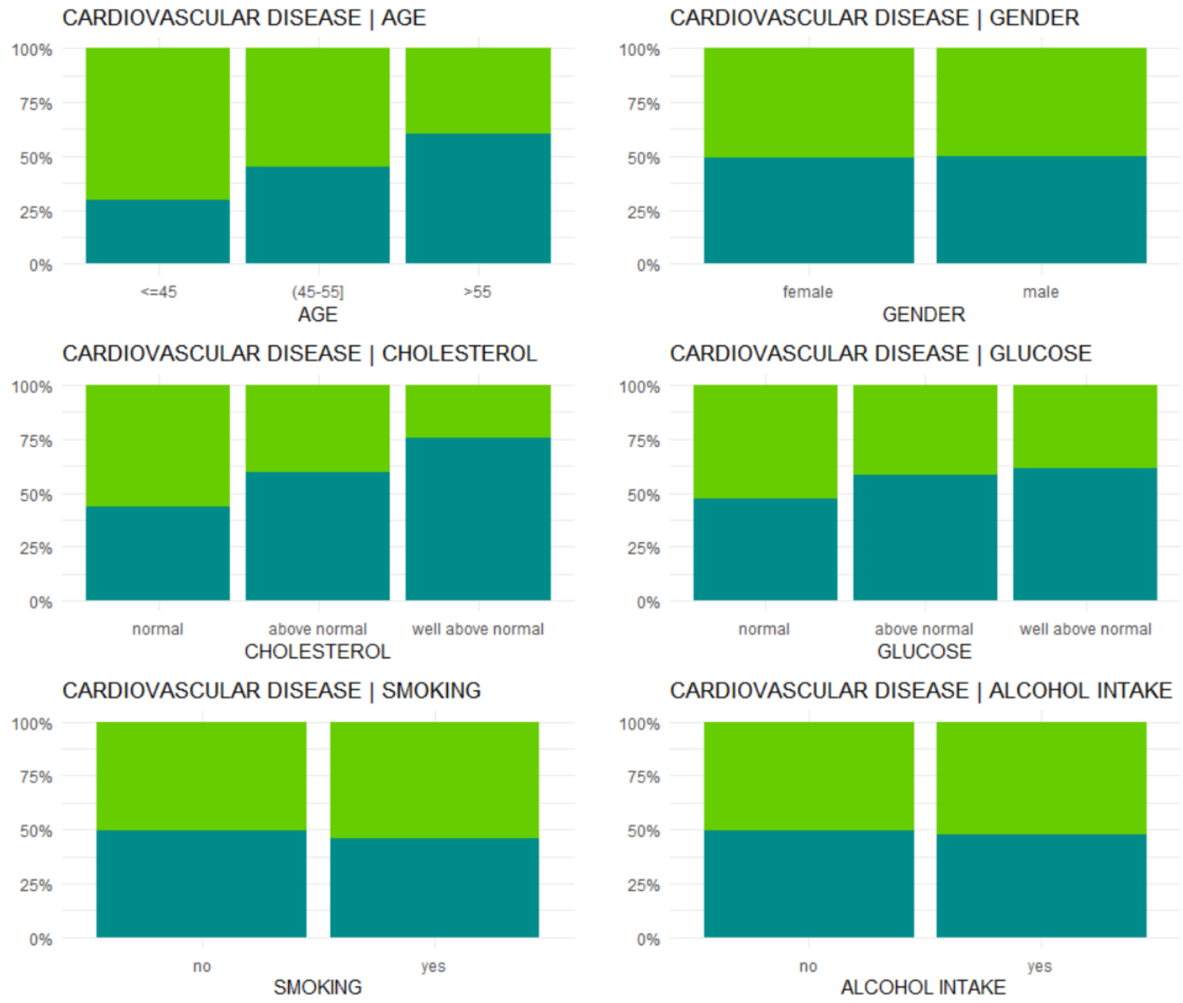
2. EXPLORATORY ANALYSIS

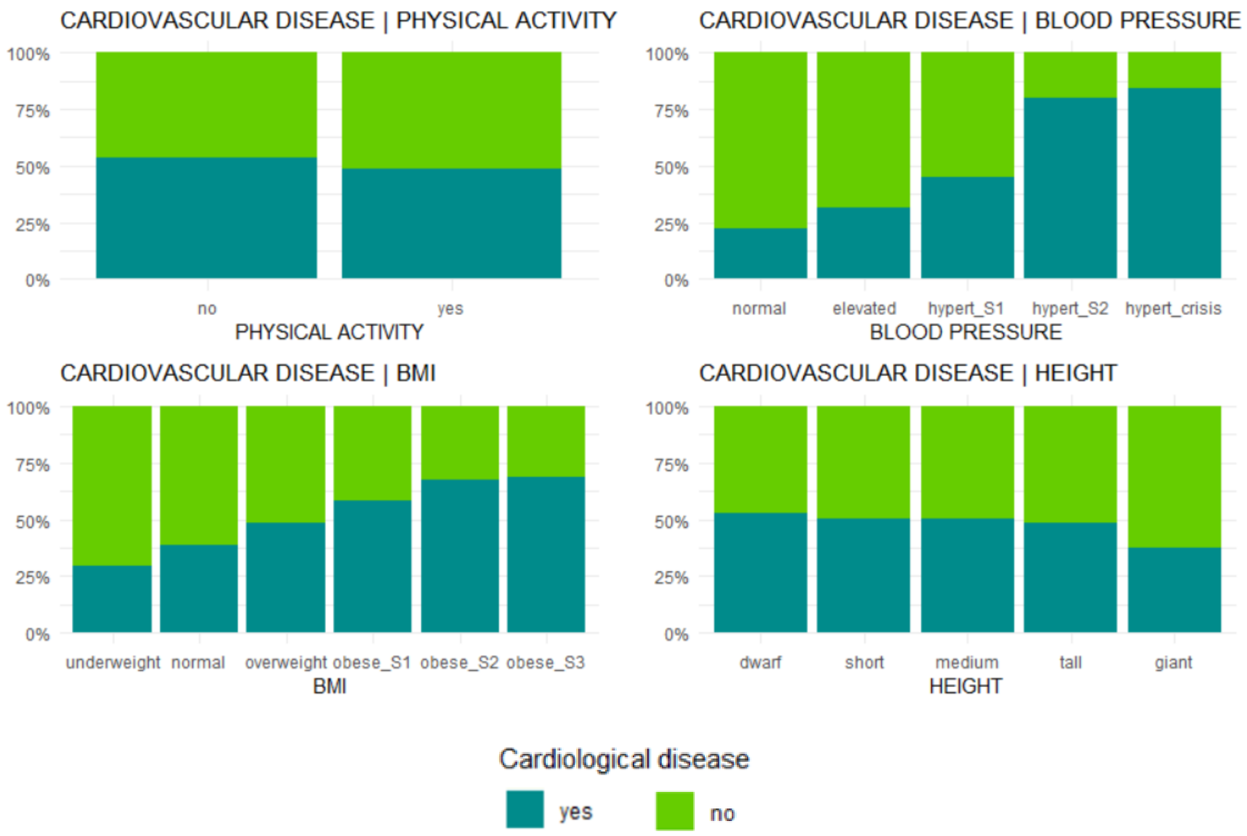
The main aim of the analysis is to develop a diagnostic tool with respect to the presence of cardiovascular diseases.
The original dataset was randomly partitioned in:
- **training set** (80% of the patients), used in order to train the classificator;
- **test set** (20% of the patients), used in order to test the performance of the obtained tool.

The partition was carried out stratifying with respect to the response variable to classify (cardiovascular disease), in order to maintain the original balance between the two categories (presence or absence of the disease).

With an exploratory purpose it may be useful to highlight the discriminant power of each of the variables available, with respect to the presence of a cardiovascular disease.

PLOT 2.1: DISCRIMINANT POWER OF EACH VARIABLE





PLOT 2.1 shows the conditional distribution of the presence of cardiovascular diseases with respect of each of the remaining variables.

Some considerations can be made:

- Variables like gender, smoking, alcohol and physical activity, taken alone, doesn’t seem to have any discriminant power with respect to the cardiological disease.
  In fact, as an example, knowing that a person is a female doesn’t facilitate the classification because not only the risk of having a disease is the same for both male and female, but also the two categories are equally likely;
- Variables like age, cholesterol, BMI index and blood pressure seem to have a relevant discriminant power. As instance, a person with a well above normal level of cholesterol is far more likely to have a cardiovascular disease than a person with a normal level.
  It is also possible to notice that the risk of having a cardiovascular disease increases as age, BMI index, cholesterol level and blood pressure increase.

3. GRAPHICAL MODEL ADAPTATION AND SELECTION

Due to the nature of the problem the chosen model was a graphical one, based on a directed acyclic graph. In fact, DAGs are often used in the context of diagnostics problems and turn out to be very useful in highlighting the causal relationships among variables and their particular patterns.

The DAG selection was carried out by the mean of the hill-climbing algorithm, in order to optimize the value of the specified scoring function (BIC).

A particular aspect of DAG models is that they allow the researcher to include knowledge of the system under study into the model selection process: in the case under study it means that some implausible relationships can be forbidden a priori, including some constrain to the adjacent matrix. PLOT 3.1 and 3.2 show the result of a hill-climbing selection, without and with the constraints specified in TABLE 3.1.

PLOT 3.1: UNCONSTRAINED SELECTED DAG MODEL

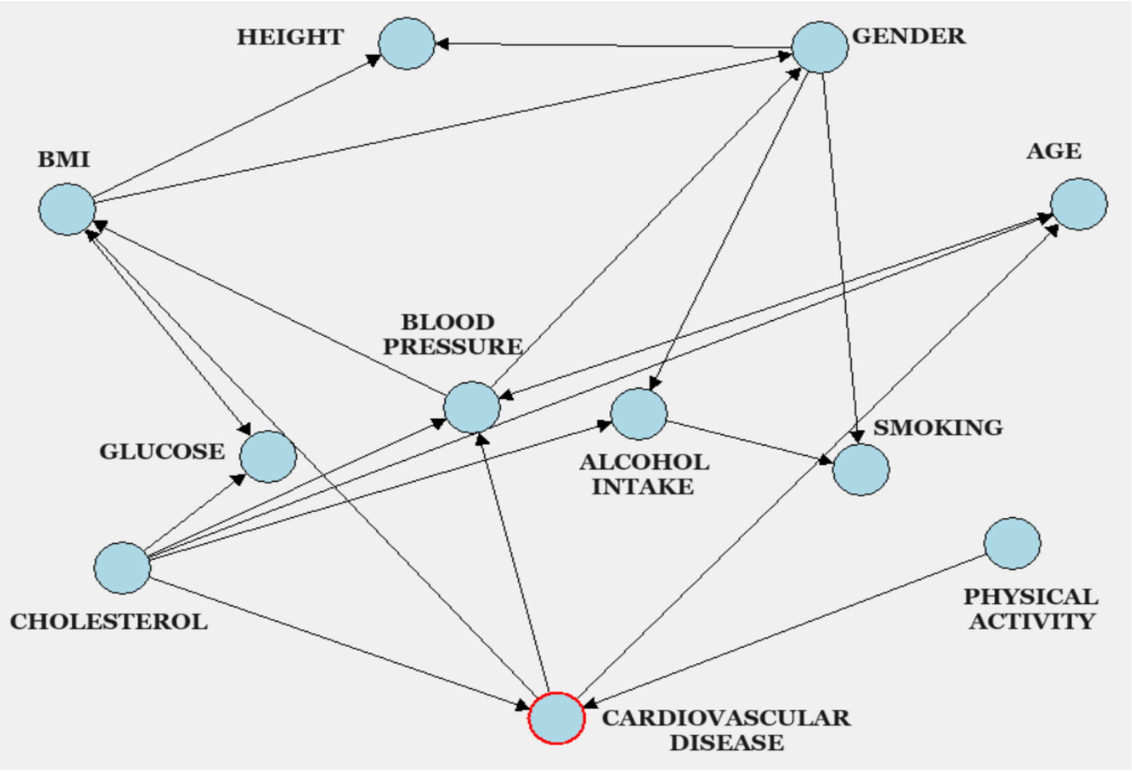
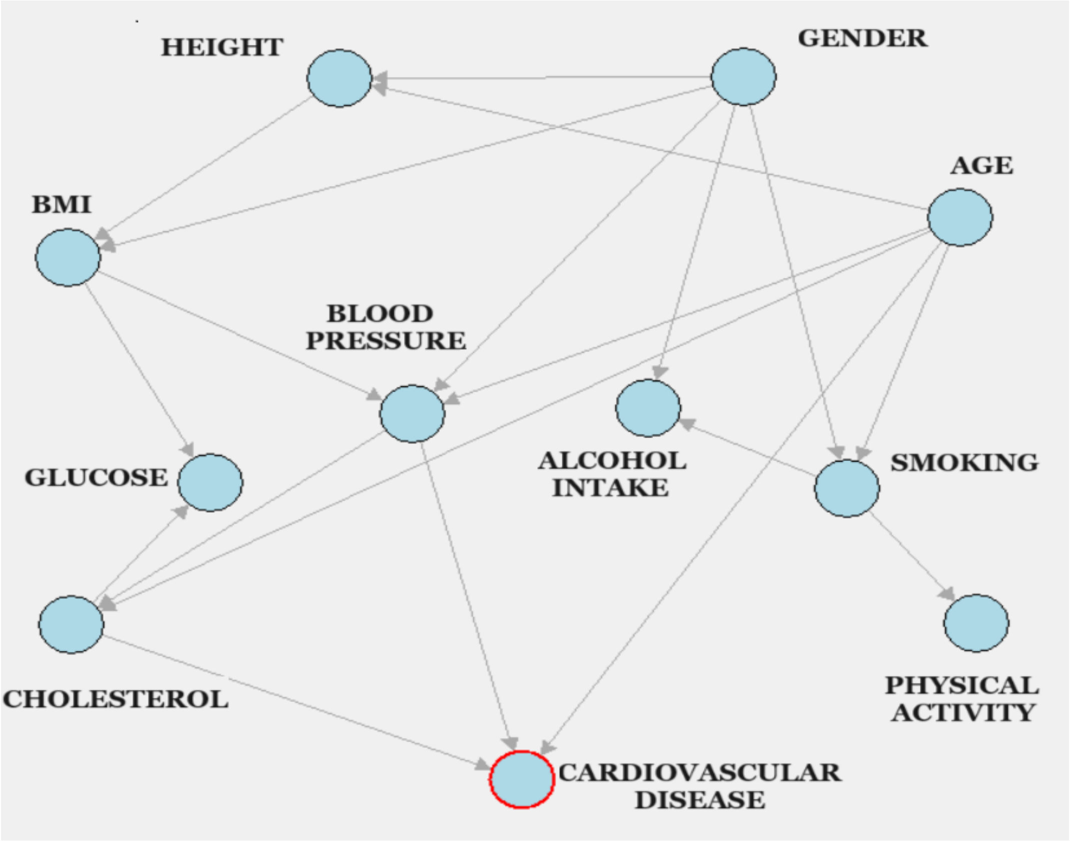


TABLE 3.1: BLACKLIST ADJACENCE MATRIX

	AGE	GENDER	CHOLESTEROL	GLUCOSE	SMOKING	ALCOHOL	PHYSICAL ACTIVITY	CARDIOVASCULAR DISEASE	BLOOD PRESSURE	BMI	HEIGHT
AGE	0	1	0	0	0	0	0	0	0	0	0
GENDER	1	0	0	0	0	0	0	0	0	0	0
CHOLESTEROL	1	1	0	0	1	1	0	0	0	0	1
GLUCOSE	1	1	0	0	0	0	0	0	0	0	1
SMOKING	1	1	0	0	0	0	0	0	0	0	1
ALCOHOL	1	1	0	0	0	0	0	0	0	0	1
PHYSICAL ACTIVITY	1	1	0	0	0	0	0	0	0	0	1
CARDIOVASCULAR DISEASE	1	1	1	1	1	1	1	1	1	1	1
BLOOD PRESSURE	1	1	0	0	0	0	0	0	0	1	1
BMI	1	1	0	0	0	0	0	0	0	0	1
HEIGHT	1	1	0	0	0	0	0	0	0	0	0

The blacklist adjacence matrix (TABLE 3.1) summarizes all the forbidden relationships. A generic cell ij equal to 1 means that the directed link starting from i with destination j can't exists.

PLOT 3.2: CONSTRAINED SELECTED DAG MODEL



Starting from the observation of the selected graphical model some considerations can be made:

- The presence of a cardiovascular disease is directly influenced by the cholesterol level, the blood pressure and the age as suggested by the exploratory analysis;
- The presence of a cardiovascular disease appears to have a marginal relationship with alcohol, glucose, and physical activity.
  However, this relationship changes when other variables are observed:
  - knowing whether a person smoke the presence of a cardiovascular disease become independent from the physical activity;
  - knowing the gender and the smoking habit of the patient the presence of a cardiovascular disease become independent on the alcohol intake;
  - knowing the BMI and the cholesterol level of the patient the glucose level has no more influence on the presence of a disease.
- As concerns the relationship among risk factors it is possible to notice that the alcohol intake is directly influenced by gender and the smoking habits while smoking by both gender and age;

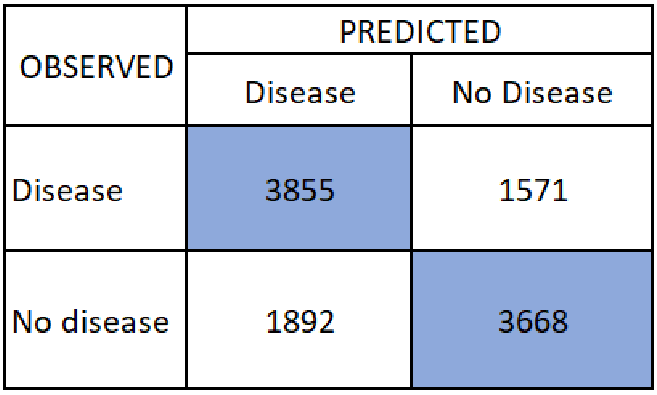
- The blood pressure status directly depends on age, gender and BMI;
- Knowing whether a person smoke or not age has no relationship with physical activity.

4. PERFORMANCE TEST

The resulting diagnostic tool works by considering the features of each patient but the presence of cardiovascular disease.
It obtains the distribution of the cardiovascular disease variable conditional on the other variable and classify a person as ill if the probability of having a disease is greater than 0.5.

In order to check whether the resulting diagnostic tool is reliable or not a performance test was performed on a sample of about 11000 patients.

TABLE 4.1: CONFUSION MATRIX



OBSERVED	PREDICTED	
	Disease	No Disease
Disease	3855	1571
No disease	1892	3668

ACCURACY: 68.40%

SENSITIVITY 65.97%

SPECIFICITY 71.00%

+ PREDICTIVE VALUE 70.00%

- PREDICTIVE VALUE 67.00%

TABLE 4.1 shows the results of the performance test.
The tool turned out to be capable of detecting about 68% of the true disease status.
It was capable of detecting the disease in the 66% of the cases where it was really present.
It was capable of detecting the absence of the disease in the 71% of the cases where it was really absent.
Among those classified as sick, the 70% was really sick whereas among those classified as healthy the 67% was really healthy.