Ciências ULisboa | Informática

Advanced Machine Learning

PROJECT 1: Bayesian Networks

INTRODUCTION

Cardiovascular diseases (CVD) are the leading cause of mortality in Europe, accounting for over 3.9 million deaths each year, or 45% of all fatalities, and annual treatment costs for CVD exceed 210 billion euros across Europe [1]. Accurate individual CVD risk prediction is essential to identify high-risk patients, assist clinicians in developing effective intervention strategies, support research hypothesis generation, and improve patient adherence to treatments.

This project aims to examine the relationships between established cardiovascular risk factors – such as age, sex, physical activity, tobacco use, and body mass index – and additional factors, including depression, sleep duration, and socioeconomic status, that may also contribute to CVD. To accomplish this, we will construct a discrete Bayesian network where each node represents a cardiovascular risk factor or medical condition, with structure and probability distributions learned from a dataset of annual health assessments.

DATASET

Data were collected from annual health assessments of adult workers covered by a private insurance provider in Spain between 2012 and 2016. These records were anonymised and securely stored. Additional socioeconomic data were derived from census information linked to postal codes, enabling to infer socioeconomic status (based on occupation, economic activity, and professional context averaged within each code) and education level (ranging from 0, indicating no formal education, to 4, representing a university degree or higher).

The current data for this project has two datasets: a **training dataset** ("cardiovascular_train.csv") with 3845 patients and a **test set** ("cardiovascular_test.csv") with 1282 patients. All outliers, duplicates, misrecorded and missing values were removed, and variables considered relevant for the study were selected. The relevant variables are grouped as follows:

- **Non-modifiable CV risk factors:** sex, age, education level, and socioeconomic status
- **Modifiable CV risk factors:** body mass index, physical activity, sleep duration, smoking profile, anxiety, and depression
- **Medical conditions:** hypertension, hypercholesterolemia, and diabetes

The data was already discretised, and Table 1 presents the values considered for each variable. Age was categorised into six groups following the coding of the Spanish National Statistical Institute (INE) in their National Sport Habits survey [2]. Education and socioeconomic levels were discretised to 1, 2, and 3 (with a larger index indicating a higher education or socioeconomic level). WHO BMI guidelines were followed: underweight (<18.5 kg/m^2), normal weight ([18.5, 25) kg/m^2), overweight ([25, 30) kg/m^2), and obese (≥ 30 kg/m^2) classes.

Medical conditions were assessed to determine if patients had diabetes (i.e., medicated or glycaemia ≥125mg/dL), hypercholesterolemia (medicated or total cholesterol ≥240mg/dL), or hypertension (medicated or systolic/diastolic blood pressure ≥140/90mm Hg).

Table 1 Variables for the model.

Variable	Levels
Sex	{female, male}
Age	(24,24], (34,44], (44,54], (54,64], (64,74]
Education level	{1, 2, 3}
Socioeconomic status	{1, 2, 3}
Body mass index	{underweight, normal, obese, overweight}
Physical activity	{insufficiently active, regularly active}
Sleep duration	{<6 hours, 6-9 hours, >9 hours}
Smoker profile	{non-smoker, ex-smoker, smoker}
Anxiety	{yes, no}
Depression	{yes, no}
Hypertension	{yes, no}
Hypercholesterolemia	{yes, no}
Diabetes	{yes, no}

WORK PLAN

The main goal of this project is to implement a Bayesian network that better captures the dependencies of this problem. Thus, you need to learn a Bayesian network for the dataset described above using the techniques learned in lectures. You are free to explore more advanced techniques of Bayesian networks that were not taught in the lectures. However, you need to explain the method briefly in the report.

This project has **five mandatory parts**:

1. **Exploratory data analysis.** Your first step is to do a brief exploration of your data to understand how each variable is distributed by the categories considered in Table 1.

2. **Learn a Bayesian network from the data.** You should use the techniques learnt in classes for structure and parameter learning. Present your explanations for the decisions made. Beware that the network might link variables in a way that does not make much sense (e.g., Body Mass Index -> Age, when it probably makes more sense the other way around). In such cases, you can force the network to include certain links. For instance, we know that physical activity may influence hypertension or hypercholesterolemia, so we can force those links in the network. However, do not force many links because you might end up doing the network by hand, which is burdensome and unnecessary. Let the algorithms learn the structure!
3. **Network and problem analysis.** After you have your Bayesian model, you can analyse relationships between variables. So, you can draw conclusions (meaning, make inferences) about:
    a. How age affects sleep duration.
    b. How does the smoker profile affect the three medical conditions (diabetes, hypertension and hypercholesterolemia.
    c. How body mass index affects the three medical conditions.
    d. You might think of other analyses that you considered interesting to analyse.
4. **Classification of the three medical conditions.** Using the test dataset provided and the Bayesian model trained, assess the classification results when you separately predict each of the target variables (medical conditions).
5. **Improving classification results.** Note that from the exploratory data analysis (step 1), one of the three medical conditions is very imbalanced. Let's explore this a little bit by using SMOTE to balance that medical condition. So, you need to apply SMOTE [3] (there are many variants of SMOTE; you need to choose the one suitable for your data) to the medical condition with a higher imbalance (and only to that condition), and train your Bayesian model to predict the three medical conditions. In this step, you only need to do classification (similar to step 4) and NOT network and problem analysis (as in step 3). Have your predicted results improved? Why?

SUBMISSION

Each group should submit the report written in Jupyter Notebook with the name **AAA2425_P1_xx.ipynb**, where the group number registered in Moodle should replace xx. The report should include (but not limited to):

- The identification of the members of the group (number and name of each element)
- All the sections described in the work plan
- The **code and corresponding outputs** obtained

- Explanations and justifications for your decisions
- Discussion of the model and results obtained

Beware that **I will not run the code of all groups, but I might run some randomly selected groups, so I need** explicit outputs in the report to confirm your conclusions.

DEADLINE

The deadline for this project is **November 22nd at 23:59 in Moodle**.

REFERENCES

[1] "European Heart Disease," 2024. [Online]. Available: https://ehnheart.org/. [Accessed 5 November 2024].

[2] P. A. M. T. &. L. V. Fernandez-Navarro, "Leisure-time physical activity and prevalence of non-communicable pathologies and prescription medication in Spain.," *PLoS One, 13(1),* p. e0191542, 2018.

[3] "Imbalanced-learn Documentation," October 2024. [Online]. Available: https://imbalanced-learn.org/stable/references/over_sampling.html. [Accessed 5 November 2024].