Heart Disease Risk Assessment using Bayesian Networks

Matteo Fasulo, Luca Tedeschini, Antonio Gravina, Luca Babboni
Master's Degree in Artificial Intelligence, University of Bologna
{ matteo.fasulo, luca.tedeschini3, antonio.gravina, luca.babboni2 }@studio.unibo.it

March 9, 2024

Abstract

Cardiovascular diseases (CVDs) remain the foremost global cause of mortality, claiming approximately 17.9 million lives in 2019. These diseases, which include heart attacks and strokes, accounted for 85% of CVD-related fatalities. In this study, Bayesian Networks (BNs) are employed for the early detection of CVDs, with a focus on validating the predictive efficacy of various BNs and exploring the intricate interactions among diverse risk factors. Our research emphasize the importance of balancing data-driven optimization with domain knowledge for meaningful and effective CVD prediction. The study found that the BN built upon domain knowledge demonstrates remarkable predictive performance, achieving a notable ROC AUC score of 0.85, highlighting its potential for meaningful and effective CVD prediction.

Introduction

Domain

In our study, we are modeling the risk of cardiovascular diseases (CVDs), a leading cause of death globally (WHO 2024). Our work draws inspiration from a paper (Ordovas et al. 2023) where the authors developed a Bayesian Network (BN) to predict CVD risk by means of CVD risk factors (CVRFs) divided into modifiable and non-modifiable CVRFs. Researchers have identified diverse CVD risk factors (Mahmood et al. 2014), such as: age, sex, chest pain type, resting blood pressure, total serum cholesterol and other meaningful features.

Our choice of using a BN as predictive model is motivated by its ability to handle complex, real-world data and its success in healthcare applications (Nielsen and JENSEN 2009). Furthermore, BNs let us analyze how various risk factors interact, providing valuable insights into the mechanisms of CVDs.

Aim

Our project aims at the creation of several BN classifiers built by using different methods, in order to predict the likelihood of CVDs based on a range of risk factors, such as the ones mentioned before. It prioritizes the maximization of the ROC AUC score, as it is proven to be particularly useful for prediction and evaluation of healthcare outcomes (Marcusson et al. 2020). Additionally, the project explores methods to transform continuous variables like cholesterol and heart rate into discrete categories by using established medical references, allowing the BN models to handle them more effectively. Ultimately, the project seeks to evaluate the performance of each model to get the best one in terms of results (accuracy and ROC AUC) and semantic meaningfulness.

Method

In order to build and test the BN classifiers, we use the 'pgmpy' library and we explore different model configurations: Naive Bayes, Hill Climbing (with all its possible scoring methods, both constrained and unconstrained, provided by the library), Domain Knowledge network (using scientific literature to establish the edges) and a reduced network with feature selection. For the latter, we use a library named 'PyImpetus' that implements a Markov Blanket based feature selection algorithm. Ultimately, this aims at the identification of the BN structure and scoring method combination that maximizes the ROC AUC score for accurate heart disease prediction. This chosen model is then subjected to further analysis, regarding the structure and the properties of the network.

Results

Our exploration reveals that the structure of the BN significantly impacts performances. Although the Naive Bayes and the Hill Climbing unconstrained models yield good results, they lack semantic meaning. Conversely, the Hill Climbing constrained and the Domain Knowledge network models maintain strong explanatory power and incorporate sensible relationships between features. This highlights the importance of balancing data-driven optimization with domain knowledge to achieve a meaningful and effective BN for heart disease prediction.

Model

In the BNs built for this study, each node embodies a unique random variable, each with a distinct interpretation and range. These variables include 'Age', 'Sex', 'ChestPainType', among others, all of which are pertinent to the study of CVDs. The continuous variables within the dataset

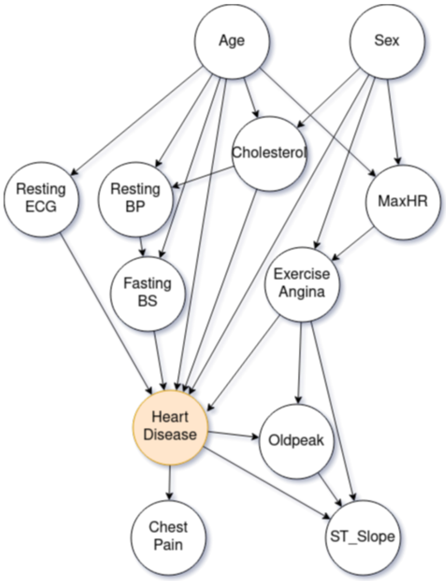

Figure 1: Domain Knowledge Bayesian Network structure

have been discretized, utilizing standard ranges derived from existing scientific literature. For models that do not grasp the semantic meaning of the variables, and consequently establish links without discerning causal relationships, the Maximum Likelihood Estimator is employed to estimate the Conditional Probability Distributions (CPDs). Conversely, for the Domain Knowledge BN, the Bayesian Estimator is used. For the construction of the Domain Knowledge network, we identify the causal relationships among the variables, substantiated by both scientific literature and domain knowledge. Among all the models, this particular one has been selected for the analysis of its structure and properties. We can also consider the network with reduced dimensionality. However, it is overly simplistic and it does not take into account the majority of the features.

Analysis

Experimental setup

To assess the results, we employ the ROC AUC score. In order to do this, we partition the data into training (85%) and testing (15%) sets. Additionally, we apply KFold Cross Validation to obtain a more realistic estimate of the classifier's performance on unseen data. The expectations are higher for the networks where we manually incorporate edges and constraints.

Results

Bayesian Model	ROC AUC
Naïve Bayes	0.84
Hill Climbing Unconstrained	0.86
Hill Climbing Constrained	0.83
Domain Knowledge	**0.85**
Reduced Network	0.86

Table 1: Results found for different BN construction methods

In terms of ROC AUC, all the results seem similar. However, the model that relies on domain knowledge performs slightly less effectively than both the unconstrained Hill Climbing model and the reduced network.

Conclusion

In this project, we explored various strategies for constructing a BN classifier. Our findings suggest that a fully explainable network is preferable to a high-performing network that lacks semantic meaning or does not exploit all the features. Even though these features may introduce some noise into the network, we believe that their inclusion is more beneficial than excluding them altogether. The final network achieved a ROC AUC score of 0.85, which is commendable. Any attempt to further improve this score could potentially lead to overfitting on the dataset.

Links to external resources

The notebook containing the project is available on GitHub. Refer (fedesoriano 2021) for the dataset.

References

fedesoriano. 2021. Heart failure prediction dataset. Retrieved: 25-02-2024.

Mahmood, S. S.; Levy, D.; Vasan, R. S.; and Wang, T. J. 2014. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet* 383(9921):999–1008.

Marcusson, J.; Nord, M.; Dong, H.-J.; and Lyth, J. 2020. Clinically useful prediction of hospital admissions in an older population. *BMC Geriatrics* 20.

Nielsen, T., and JENSEN, F. 2009. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer New York.

Ordovas, J.; Rios-Insua, D.; Santos-Lozano, A.; Lucia, A.; Torres, A.; Kosgodagan, A.; and Camacho, J. 2023. A bayesian network model for predicting cardiovascular risk. *Computer Methods and Programs in Biomedicine* 231:107405.

WHO. 2024. Cardiovascular diseases (cvds). Accessed: 25-02-2024.