Graphical Model For Health Risk Assessment Using Bayesian Networks

Bayesian Statistics for Machine Learning Course Project

Halmstad University
HT2024

Eirini Ornithopoulou

Table of Contents

Introduction

This project involves constructing a Bayesian Network (BN) to model the dependencies among various health-related variables, such as hypertension, heart disease, smoking history, and diabetes. Bayesian Networks are powerful probabilistic graphical models that represent complex relationships between variables through directed acyclic graphs (DAGs), where nodes symbolize variables and edges signify conditional dependencies [1]. These networks are widely used for representing uncertain knowledge, allowing for the analysis of complex systems with incomplete data[2].

In healthcare, Bayesian Networks offer valuable insights into the conditional dependencies and potential causal interactions between risk factors and disease outcomes [3]. For instance, the relationship between hypertension and heart disease or between smoking and diabetes can be visually represented and quantitatively analyzed within the network framework. By capturing these relationships, Bayesian Networks help highlight how various health conditions interact, providing a robust framework for reasoning under uncertainty—a common challenge in medical contexts [4].

This approach is particularly beneficial in healthcare applications where understanding complex interactions among multiple risk factors is crucial for diagnosis, prognosis, and decision support. Bayesian Networks not only allow healthcare professionals to assess individual risk profiles more accurately but also enable the integration of prior domain knowledge with data, thus offering a more comprehensive and interpretable risk assessment tool [5].

The aim of this project is to investigate some of the methods for developing a well-structured graphical Bayesian model that effectively captures probabilistic relationships among health-related variables. By visually and mathematically modeling these dependencies, the interpretability of health risk factors is enhanced and decision-making processes are improved in clinical settings. This project could further demonstrate the applicability of Bayesian Networks as tools for personalized medicine, enabling individualized risk assessments based on a person’s unique profile of risk factors.

Methodology

Dataset

The selected dataset[6] is open and was downloaded from Kaggle. It has 100000 data points and 9 features: gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level and diabetes.

Data Preprocessing

Before beginning the analysis the dataset is preprocessed. This involved encoding the categorical variable smoking history and removing the variable gender to simplify associations.

Definition of Network Structure

The next step is to define the network structure. Two main approaches have been used here. One of proposing the network based on domain knowledge. And the other is by utilising commonly used network discovery methods.

Proposed network

The proposed Bayesian Network structure is designed based on established medical knowledge regarding the interrelationships among various health conditions and risk factors. Each directed edge in the structure represents a conditional dependency, where the first variable influences the second. This approach allows for a probabilistic model that captures known pathways of influence and risk in healthcare.

The proposed structure is as follows:

Hypertension → Heart Disease: There is substantial evidence linking hypertension as a significant risk factor for heart disease. Elevated blood pressure strains the cardiovascular system, increasing the likelihood of coronary artery disease, heart failure, and other cardiovascular conditions [7].

Heart Disease → Diabetes: Heart disease and diabetes frequently co-occur, with cardiovascular issues potentially exacerbating insulin resistance and other diabetic symptoms. Additionally, metabolic syndrome often links these conditions[8].

Smoking History → Heart Disease: Smoking is a well-documented risk factor for cardiovascular disease. It contributes to arterial damage, inflammation, and increased risk of plaque buildup, making it a critical variable affecting heart health [9].

Age → Diabetes: Age is a prominent risk factor for diabetes, particularly Type 2 diabetes, due to the decrease in glucose tolerance and insulin sensitivity that often occurs as people age [10].

BMI → Diabetes: Body Mass Index (BMI) is closely linked with diabetes risk. Higher BMI is associated with increased adiposity, which is known to decrease insulin sensitivity and contribute to the development of Type 2 diabetes [11].

BMI → Heart Disease: Elevated BMI, especially in the form of central obesity, is associated with higher risks of hypertension, atherosclerosis, and other cardiovascular conditions that contribute to heart disease [12].

HbA1c Level → Diabetes: Hemoglobin A1c (HbA1c) is a primary biomarker used to diagnose and monitor diabetes. Higher HbA1c levels are indicative of poor glucose control and are directly linked with diabetes management and prognosis [13].

Blood Glucose Level → Diabetes: Blood glucose levels are central to diabetes diagnosis and management, with elevated levels indicating hyperglycemia, a hallmark of diabetes [14].

Network based on conditional dependency testing

This methodology involves creating a Bayesian Network structure by learning the network directly from the data. Instead of relying on domain knowledge to define the structure, this approach uses data-driven search algorithms—Hill Climbing and Greedy Equivalence Search (GES)—to find an optimal network structure. Bayesian Network structure learning is particularly useful when the relationships between variables are unknown or need validation.

A. Structure Learning Using Greedy Equivalence Search (GES)

Greedy Equivalence Search (GES) is a search-based algorithm for learning the structure of a Bayesian Network. GES is a two-phase algorithm. It starts with an empty network and iteratively adds edges to maximize a scoring function, improving the model fit with each step. After reaching a local maximum, it iteratively removes edges to refine the network structure.

The **Bayesian Information Criterion (BIC)** score is used as the scoring method, which evaluates the network's goodness-of-fit while penalizing complexity. Lower BIC scores

indicate better models. BIC starts by calculating how well the model explains the observed data (the likelihood). Higher likelihood means a better fit to the data. Then, it penalizes models with more parameters to discourage overly complex models. This penalty grows with the number of parameters and the sample size. Here's the formula for it:

$$BIC = -2 \times \ln(\text{Likelihood}) + k \times \ln(n)$$

Where k: Number of parameters in the model and n: Number of data samples.

After the GES algorithm finds an optimal structure (i.e., the set of edges defining dependencies between variables), the structure is converted into a BayesianNetwork object. This step ensures that we can treat the learned structure as a standard Bayesian Network in pgmpy, allowing us to proceed with parameter estimation and other operations. GES can provide an effective structure learning approach but may occasionally miss including certain nodes, like bmi, in the optimal structure.

B. Structure Learning Using Hill Climbing

Hill Climbing is a greedy algorithm that iteratively adds, removes, or reverses edges to maximize a scoring function (e.g., BIC or K2 scores). While fast, it can converge prematurely to a local maximum, potentially leading to suboptimal structures.

The HC algorithm begins with a basic network. Then adds, removes, or reverses edges. Using a scoring function (e.g., BIC) it evaluates each new structure. It adopts the change with the highest score improvement. And stops when no change improves the score.

(C. PC Algorithm, named after "Peter and Clark", is a constraint-based method for learning the structure of a Bayesian Network or causal graph from data. It identifies conditional independencies in the data to build a DAG that represents dependencies among variables. However, I couldn't get it to work. The PC algorithm is computationally expensive for large numbers of variables because it involves testing all pairs of variables across multiple conditioning sets. So, potentially it could have worked for less variables.)

Train the network

With the structure defined, the next step is parameter learning. Maximum Likelihood Estimation (MLE) is applied to estimate the conditional probability tables (CPTs) for each variable given its parent variables. MLE is a common approach for parameter estimation in Bayesian Networks, as it finds the parameters that maximize the likelihood of observing the given data [1].

Specifically, for each variable, MLE counts how often each value occurs with each combination of its parent values. Then, it divides these counts by the total occurrences of the parent combinations to estimate conditional probabilities. For example, if 30 out of 40 smokers have

heart disease, MLE would estimate P(HeartDisease|Smoking)=0.75. MLE does not incorporate prior knowledge; instead, it relies purely on the observed data to estimate probabilities.

Inference and Queries

After the model parameters are learned, inference is performed using the Variable Elimination method. Variable Elimination is an algorithm that enables probabilistic inference by systematically removing variables to compute marginal probabilities of interest. This method is computationally efficient for Bayesian Networks and well-suited for calculating conditional probabilities, such as the likelihood of diabetes given specific values of heart_disease, hypertension, or other health variables [5].

For example, using this inference approach, one can calculate the probability of diabetes given high blood pressure and a smoking history. Such insights can be valuable for assessing individual health risks and aiding decision-making in healthcare.

Model visualization

To effectively interpret the relationships and dependencies in the Bayesian Network model, visualizing the network structure is essential. A visual representation can aid in understanding complex dependencies among health variables, providing a clear overview of the interconnected risk factors in the healthcare domain.

Using networkx, a Python package for the creation, manipulation, and study of complex networks, a directed graph (DiGraph) is created to represent the Bayesian Network structure. Matplotlib is also necessary for plotting.

Results and Discussion

The proposed structured network

First, the proposed network was trained. The details of the logic behind the structure are in the methodology section. In Fig. 1, the visual representation using the proposed network for the DAG visualization can be seen. Now, let's have a look at the discovered networks using GES and HC respectively.

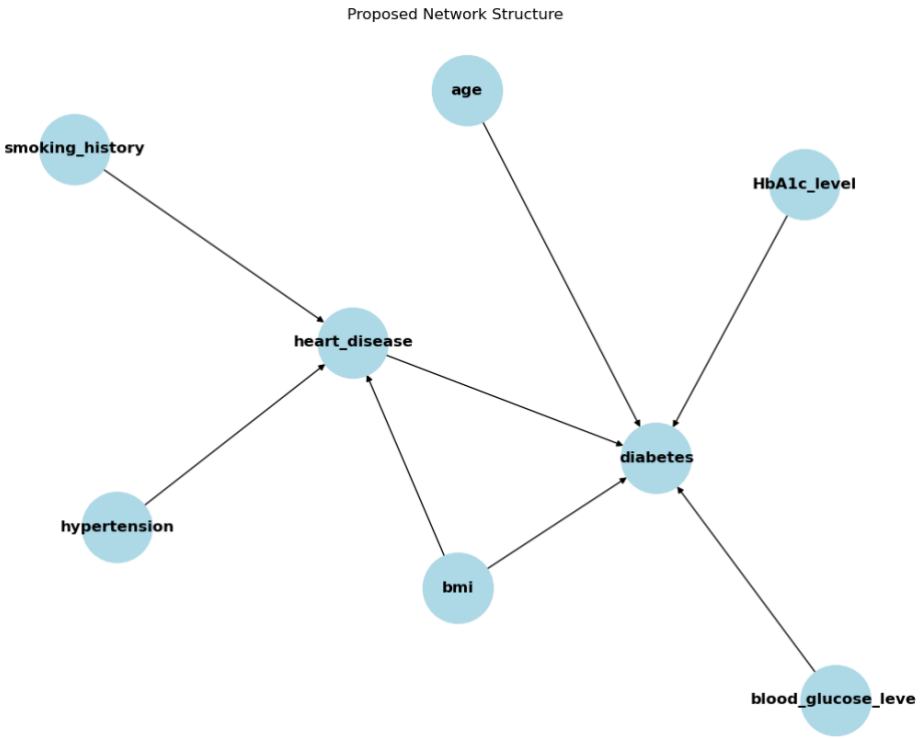

Figure 1The Proposed DAG using networkx.

The GES structured network

The GES-generated network, as shown in Fig. 2, reveals several notable differences from the initially proposed structure. For instance, it suggests that diabetes influences age, and while HbA1c levels impact diabetes, it is diabetes that affects blood glucose levels, highlighting an unexpected direction of causation. This raises questions about the causal interpretations in the learned network and suggests that the proposed model may need a reassessment of edge directions to better align with domain knowledge.

Another point of interest, that is not shown here but was important when choosing to remove the gender from the calculations, is that heart disease and smoking history appeared to affect

gender, an implausible dependency that likely reflects limitations in the data or the structure learning algorithm rather than a real causal effect.

Furthermore, the exclusion of BMI from the network is significant—whether this exclusion represents a genuine lack of dependency in the dataset or a limitation of the algorithm's ability to detect complex interactions remains unclear. These observations suggest that the network complexity, given the current dataset and learning method, might benefit from simplification or the use of fewer variables, as the inclusion of too many features may have introduced noise or led to unstable dependencies.

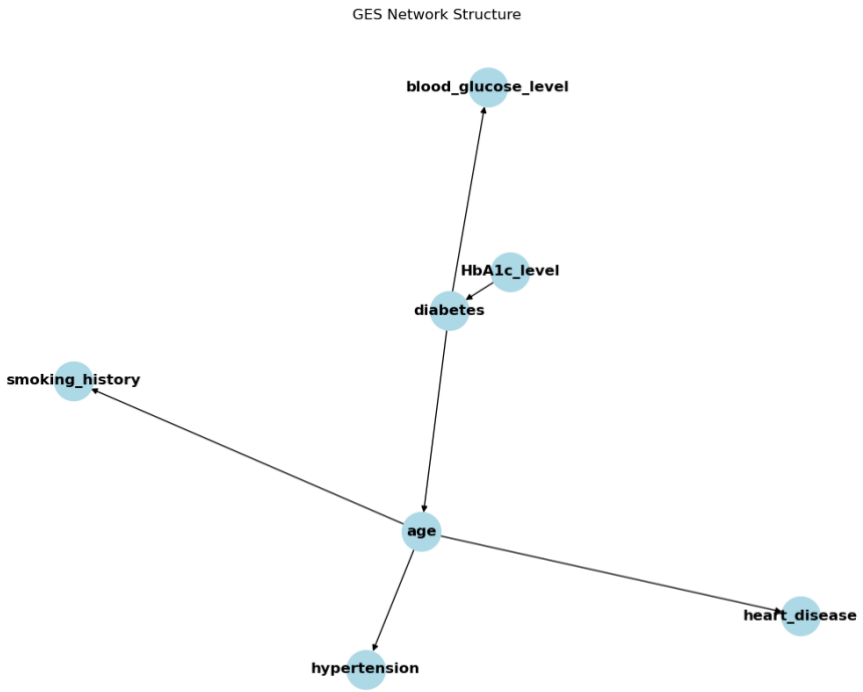

Figure 2The GES structured network.

The HC structured network

In Fig. 3, the HC structured network can be viewed. It has exactly the same structure and directionality as the GES network, and has excluded BMI.

This is likely due to the absence of statistically significant dependencies between BMI and other health-related variables in the dataset. This outcome suggests that BMI may be conditionally independent of the other variables given the available data, or that other variables, such as age or smoking history, provide sufficient predictive power for the model without needing BMI. This is not proof of the insignificance of BMI, but rather it could point to issues in the chosen dataset.

Additionally, the Bayesian Information Criterion (BIC) scoring function used by both algorithms penalizes model complexity, and the inclusion of BMI may not have sufficiently improved model fit to offset this penalty. This again indicates that either the relationships involving BMI are not strong in this dataset, or the dataset may lack enough data quality or quantity to capture BMI's impact effectively.

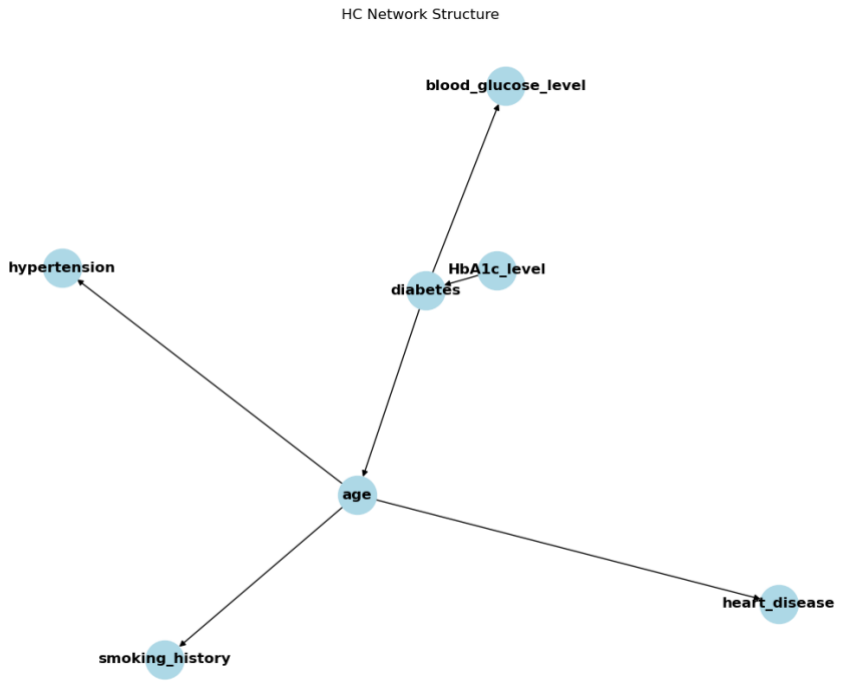

Figure 3The HC structured network.

Inference

When evaluating the responses to the queries, both the GES and Hill Climb networks produced identical results, as they share the same learned structure, see Table 1. This consistency is expected given that both networks excluded the BMI variable, likely due to limitations in the dataset or the scoring criteria, despite BMI's known connections to both diabetes and heart disease. The absence of BMI in these structures poses a limitation, as it prevents us from evaluating queries that would otherwise incorporate BMI's influence on health outcomes.

Additionally, the responses from the GES and Hill Climb networks exhibit a striking level of certainty, particularly for Query 1, where the probability of diabetes given high HbA1c and hypertension was calculated as 100%. Such deterministic results are unusual in probabilistic models and raise questions about the validity of the data, model assumptions, or parameter estimation. These results, especially in cases like Queries 1 and 3, where the probabilities differ significantly from those in the proposed network, suggest that the learned structures may not fully capture the complexities of real-world dependencies and could benefit from additional refinements or constraints. Query 2 gives low and similar probability in both cases, therefore it is harder to discuss on it.

About Query 5, hypertension and diabetes are both well-established, significant risk factors for heart disease. When these conditions coexist, they substantially increase the risk of heart disease due to their combined impact on the cardiovascular system. A probability of 35% better reflects this elevated risk, while a 10% probability underestimates the severity of these conditions as a combined predictor of heart disease.

Table 1 Query comparison for all three networks.

Query	Proposed network	GES network	Hill Climb network
1.Probability of diabetes given high hBA1c and hypertension	51%	100%	100%
2.Probability of heart disease given smoking history and high blood glucose	11%	3%	3%
3. Probability of diabetes given older age and high HbA1c level	51%	100%	100%
4. Probability of heart disease given hypertension and diabetes	35%	10%	10%

Conclusion

This project demonstrates the application of Bayesian Networks for modeling dependencies among health-related variables to assess disease risk and support decision-making in healthcare. The proposed network, grounded in medical knowledge, provided more realistic and interpretable results than the data-driven structures generated by Greedy Equivalence Search (GES) and Hill Climbing (HC). The proposed model captured nuanced dependencies between variables, reflecting the probabilistic nature of health outcomes. In contrast, the GES and HC networks exhibited more deterministic behavior, yielding overly confident probabilities sometimes, and omitted the BMI variable, potentially due to limitations in the dataset and scoring function constraints.

These findings highlight the importance of integrating domain knowledge when constructing Bayesian Networks in healthcare, as purely data-driven approaches may miss clinically relevant dependencies or produce implausible causal directions. Further research could enhance model accuracy by refining dataset quality, exploring alternative scoring methods, and incorporating expert-validated structures. Overall, this study reinforces the value of Bayesian Networks as an interpretable and flexible framework for personalized risk assessment in healthcare, especially when supported by robust data and domain-informed design.

References

1. Neapolitan, R.E., *Learning bayesian networks*. Vol. 38. 2004: Pearson Prentice Hall Upper Saddle River.
2. Pearl, J., *Probabilistic reasoning in intelligent systems: networks of plausible inference*. 2014: Elsevier.
3. Lucas, P.J., L.C. Van der Gaag, and A. Abu-Hanna, *Bayesian networks in biomedicine and health-care*. 2004, Elsevier. p. 201-214.
4. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine learning, 1997. **29**: p. 131-163.
5. Koller, D., *Probabilistic Graphical Models: Principles and Techniques*. 2009: The MIT Press.
6. Mustafa, T.Z., *Diabetes Prediction Dataset [Data set]*. 2023, Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset.
7. Chobanian, A.V., et al., *Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure.* hypertension, 2003. **42**(6): p. 1206-1252.
8. Kannel, W.B. and D.L. McGee, *Diabetes and cardiovascular disease: the Framingham study.* Jama, 1979. **241**(19): p. 2035-2038.
9. Services, U.D.o.H.a.H., *The health consequences of smoking—50 years of progress: a report of the Surgeon General*. 2014, Atlanta, GA: US Department of Health and Human Services, Centers for Disease ….
10. Schmidt, M.I.s., et al., *Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study.* Diabetes care, 2005. **28**(8): p. 2013-2018.
11. Chan, J.M., et al., *Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men.* Diabetes care, 1994. **17**(9): p. 961-969.
12. Kenchaiah, S., et al., *Obesity and the risk of heart failure.* New England Journal of Medicine, 2002. **347**(5): p. 305-313.
13. Selvin, E., et al., *Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults.* New England Journal of Medicine, 2010. **362**(9): p. 800-811.
14. Association, A.D., *Standards of medical care in diabetes—2014.* Diabetes care, 2014. **37**(Supplement_1): p. S14-S80.