Predicting Heart Disease Risk Using Bayesian Networks

Syeda Zobia Irshad
syedazobia.irshad@studio.unibo.it
Nadia Afzal
nadia.afzal2@studio.unibo.it
Syeda Anousha Hassan
anoushahassan.syeda@studio.unibo.it
Master's in Artificial Intelligence, University of Bologna

December 1, 2025

Abstract

This mini-project develops and evaluates a Bayesian Network for predicting heart disease using clinical and demographic variables such as age, sex, cholesterol level, chest pain type, maximum heart rate, and exercise-induced angina. Key features were discretized to improve interpretability. A manual network was constructed using domain expertise, while a Hill-Climb structure-learning method with BIC scoring generated a data-driven alternative. Inference using Variable Elimination and Likelihood-Weighted Sampling showed that age, sex, cholesterol, and symptom-related factors meaningfully influence heart-disease risk. Although the learned network introduced some differing dependencies, both models produced broadly similar inference patterns, with only modest variations across most probability estimates. Together, these expert-defined and data-driven approaches offer an interpretable and reliable framework for supporting clinical decision-making.

Introduction

Domain

Heart disease is a major public-health concern, and identifying key clinical and demographic risk factors is vital for early intervention. This project examines variables such as age, sex, chest pain type, cholesterol, maximum heart rate, and exercise-induced angina. A manual Bayesian Network captured clinically meaningful dependencies, while a Hill-Climb structure-learning approach with BIC scoring produced a complementary, data-driven network. Although the learned network revealed some additional dependencies, overall inference patterns were largely consistent with the expert-defined model. Combining expert knowledge with data-driven learning provides interpretable, evidence-based insights to support clinical decision-making.

Aim

The aim of this project is to develop a Bayesian Network model capable of predicting heart-disease risk using key clinical and demographic variables. The objective is to identify high-risk patient profiles and to provide an interpretable representation of how factors such as age, sex, cholesterol level, chest pain type, maximum heart rate, and exercise-induced angina interact to influence disease likelihood. By combining an expert-defined manual network with a data-driven structure-learning approach using Hill-Climb and BIC scoring, the project seeks to evaluate both perspectives and understand how domain knowledge and statistical patterns complement each other. Together, these models support risk prediction, explainability, and informed clinical decision-making.

Method

This project is following a structured 3-step process for the development and evaluation of Bayesian Networks to predict heart disease. Each step addresses a specific part of the workflow, from data preparation to model construction to inference analysis.

- **Data Preparation and Discretization.** The dataset was cleansed, continuous variables (age, cholesterol level, maximum heart rate) were discretized and key features (sex, type of chest pain, exercise induced angina) were selected to be modeled.
- **Model Construction and Parameter Learning.** A manual Bayesian network was constructed from domain knowledge; CPDs were estimated using `BayesianEstimator` with BDeu prior (ESS=10). Hill-Climb structure learning using BIC scoring led to a data network. Structural validation comprised DAG validation, independence validation, and Markov blanket validation.
- **Inference and Analysis.** Exact inference (Variable Elimination) and approximate inference (Likelihood Weighted Sampling, 10,000 samples) were carried out. Predictive, diagnostic and intercausal queries offered interpretable information Comparison of manual and learned networks confirmed some expert assumptions - and identified some dependencies determined by data itself.

Results

Using both the **manual Bayesian network** and the **Hill-Climb learned network** the analysis shows that the risk of the older patients developing heart disease is moderately high. Male patients also present with a high probability of the disease, and thus sex is an important factor. Cholesterol levels are also important, so the higher your cholesterol the more likely you are to get the disease.Intercausal analysis shows that in a population of individuals with high cholesterol and older age, the probability of heart disease increases. Moreover, male patients with high cholesterol exhibit a compounded risk, proving that age, sex and cholesterol interact to raise the likelihood of the disease further. Overall, the combination of expert-defined and data-driven network structures offers interpretive information about the combined impact of demographic and clinical factors on risk of heart disease.

Model

Figures 1 and 2 show the manual and learned Bayesian networks for heart disease prediction.

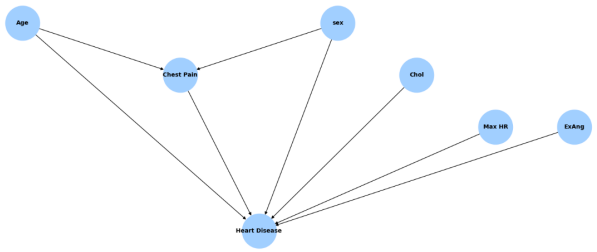

Figure 1: Manual Bayesian Network constructed using expert knowledge.

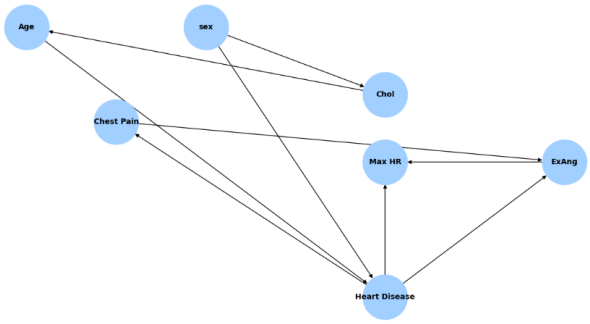

Figure 2: Bayesian Network learned using Hill-Climb structure learning (BIC scoring).

Both models were built using seven variables to capture the combined effects of demographic and clinical factors on heart disease risk. The variables were discretized to improve interpretability: *Age Group* represents the patient's life stage (Young <40 years, Middle 40–54 years, Older ≥55 years); *Sex* indicates gender (0 = Female, 1 = Male); *Chest Pain Type* categorizes symptoms (0 = Typical Angina, 1 = Atypical Angina, 2 = Non-Anginal Pain, 3 = Asymptomatic); *Cholesterol Level* is grouped into Desirable (≤200 mg/dl), Borderline (200–239 mg/dl), or High (≥240 mg/dl); *Max HR (Thalach Level)* categorizes maximum heart rate during stress testing as Low (≤120), Medium (120–159), or High (≥160); *Exercise-Induced Angina* indicates whether angina occurs during exertion (0 = No, 1 = Yes); and *Heart Disease* represents the final diagnosis (0 = No, 1 = Yes).

In the **manual network**, *Age Group* and *Sex* have an influence on both *Chest Pain* and *Heart Disease* - and *Chest Pain*, *Cholesterol Level*, *Max HR* and *Exercise-Induced Angina* have a direct influence on *Heart Disease* - which represents the expert knowledge.In contrast, the **Hill-Climb learned network** shows that *Sex* influences *Cholesterol Level* and *Cholesterol Level* influences *Age Group*, patterns absent in the expert model. Both *Sex* and *Age Group* directly affect *Heart Disease*, which in turn affects *Chest Pain (CP)*, *Max HR (Thalach Level)*, and *Exercise-Induced Angina (ExAng)*. The network also reverses certain directions, such as making *Heart Disease* a parent of *CP*. Additionally, *CP* affects *ExAng*, which influences *Thalach Level*, revealing new relationships not present in the manual model.

Analysis

Experimental Setup

To test the predictive power and interpretability of the Bayesian networks a number of inference queries were developed, including predictive, diagnostic and intercausal reasoning. Each query was understood in terms of patient risk to ensure that the models brought out coherent insights consistent with clinical understanding. The queries included:

1. What is the likelihood of heart disease among older individuals?
2. How likely is a patient to have heart disease if they are male?
3. If a patient has heart disease, what is the probability they have high cholesterol?
4. If a male patient has heart disease, how does this influence the likelihood of high cholesterol?
5. For older patients with heart disease, how likely is it that they have high cholesterol?

Model reliability for both the **manual** and **Hill-Climb learned networks** was confirmed by checking that both the CPTs in the network all summed to 1 and checked that both networks were valid DAGs. Local independence assertions as well as Markov blanket analyses confirmed their structures. Inference based on Variable Elimination and Likelihood Weighted Sampling (10,000 samples) resulted in consistent and comparable posterior distributions providing support for robust evaluation of predictive risk and interpretability of clinical and demographic factors. .

Results

The inference results from both the **manual Bayesian network** and the **Hill-Climb learned network** align with major clinical patterns, although some probability estimates differ between the two models. Both networks indicate that male patients have an elevated risk of heart disease, with estimates around 41%. Older patients also show a moderate risk of roughly 37%, reflecting age-related physiological and lifestyle factors. Among patients diagnosed with heart disease, high cholesterol remains the most common category (approximately 46–50%), followed by borderline cholesterol (33–38%) and desirable cholesterol (17–18%). Intercausal analysis further suggests that age and sex interact with cholesterol levels, with older male patients having a higher likelihood of elevated cholesterol. Overall, while the two models differ slightly in some conditional probabilities, both capture broadly consistent relationships between demographic and clinical variables, offering meaningful predictive and explanatory insights into heart-disease risk.

Conclusion

This project demonstrates the effectiveness of Bayesian networks in modelling heart-disease risk and providing an interpretable probabilistic understanding of how demographic and clinical factors interact. Key findings highlight the increased risk among older and male patients, as well as the strong influence of high cholesterol. Several limitations remain, including a reduced feature set, the exponential growth of conditional probability tables with additional parent states, and the static nature of the model, which does not account for changes in patient health over time. Future work could explore hybrid approaches that integrate expert-defined structures with data-driven learning, incorporate additional clinical features, and adopt more scalable or dynamic Bayesian models to further improve predictive accuracy and interpretability

Links to external resources

The dataset used in this project was obtained from Kaggle: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data

References

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[Ankan & Panda, 2015] Ankan, A., & Panda, A. (2015). pgmpy: Probabilistic Graphical Models using Python. *Proceedings of the 14th Python in Science Conference*.

[pgmpy Docs, 2025] pgmpy Developers. (2025). *Probabilistic Graphical Models using Python* [Documentation]. Available at: https://pgmpy.org/