PGM Report

2446659 - Ru'aan Maharaj

I. Introduction

This paper presents a probabilistic graphical model for predicting heart disease severity using the Cleveland dataset. By leveraging Bayesian networks, the model not only estimates disease probability but provides transparent, interpretable inferences aligned with clinical reasoning.

II. Problem domain

Heart disease remains one of the leading causes of death globally, although it can be largely prevented through early detection and lifestyle interventions. The challenge lies in identifying at-risk individuals before symptoms escalate, using data that is both accessible and interpretable.

This project explores how probabilistic graphical models, specifically Bayesian Networks, can be used to model the conditional dependencies between medical risk factors (such as age, cholesterol, and blood pressure) and the presence of heart disease. Using the Cleveland Heart Disease dataset, the goal is to develop a system that not only predicts the severity of heart disease based on patient characteristics, but also provides interpretable reasoning behind these predictions using conditional probability distributions.

Unlike black-box classifiers, Bayesian networks offer a transparent and explainable AI approach, which is crucial in high-stakes domains such as health care. This project demonstrates how exact inference through Variable Elimination can be used to simulate real-world diagnostic queries, giving healthcare practitioners a way to reason about the likelihood of disease in the presence of partial or uncertain information.

By aligning domain knowledge from medicine with formal probabilistic reasoning, this work aims to show how Bayesian models can not only predict out comes but also expose hidden dependencies that can guide future research, prevention strategies, and personalized medicine.

III. Data Handling and Preprocessing

A. Data handling and Feature Engineering

The dataset used in this project is the Cleveland Heart Disease dataset, which contains 303 patient records and 14 attributes, including both continuous measurements and categorical indicators. Initial data analysis involved checking the shape and quality of the dataset, revealing the presence of missing values in the ca and thal columns, denoted by question marks. These were converted to proper null values and handled using row-wise deletion, a decision made after observing that only 6 rows were affected. With 297 remaining complete records, this trade-off maintained data integrity while avoiding the uncertainty of imputation.

Continuous features such as age, chol, trestbps, thalach, and oldpeak were discretized into categorical bins. This was done to accommodate the requirements of Bayesian networks, which operate more effectively on discrete variables, and to allow for interpretable conditional probability tables (CPDs). Discretization also reflects real-world clinical thinking, where risk factors are often assessed in categories (e.g., "high blood pressure" rather than a precise number).

Both original and discretized versions of the data were preserved to retain flexibility in future modelling or analysis. Feature engineering was performed with attention to domain knowledge, using medically accepted thresholds to bin variables such as cholesterol and blood pressure. This ensured that the preprocessing was aligned with clinical standards rather than arbitrary statistical boundaries.

For example, the age feature was discretized into three categories:

- young (≤ 40),
- middle (41–60),
- old (> 60)

Similarly, cholesterol and resting blood pressure were binned using medically accepted thresholds for normal, borderline, and high-risk values. This approach enabled the model to reason using categories that hold real-world significance rather than abstract numerical differences.

B. Ethical Considerations

In high-stakes domains like healthcare, ethical considerations must guide every stage of model development. A key motivation behind using Bayesian networks was their inherent transparency and explainability. Unlike black-box models such as deep neural networks, Bayesian networks produce interpretable CPDs that allow clinicians to understand why a certain prediction is made, a critical requirement in real-world diagnosis.

The choice to discretize features into clinically meaningful categories was driven not just by technical requirements, but by a commitment to ethical data representation. It ensures that outputs remain understandable to non-technical users and aligns with the kinds of decisions medical professionals actually make.

In handling missing data, a cautious approach was taken: instead of imputing with potentially misleading averages, rows were removed to preserve trust in the model's predictions. The dataset itself also contains sensitive health information, so care was taken to work with anonymized, public data that complies with ethical research standards.

IV. PGM Selection and Application

A. Model Selection and Justification

The core of this project relies on the use of Bayesian Networks, a class of Probabilistic Graphical Models (PGMs) particularly well-suited for representing causal relationships and conditional dependencies. Unlike black-box machine learning models, Bayesian Networks provide natural directionality between variables, which aligns perfectly with the problem of medical diagnosis, where understanding what causes what is crucial.

In the context of heart disease prediction, it is not sufficient to simply classify an outcome, we need a system that can represent how and why risk factors contribute to disease. Bayesian Networks support this through their use of Conditional Probability Distributions (CPDs), which define the likelihood of a variable given its parents. In this project, we structured our model such that each of five medically relevant features (age group, cholesterol level, blood pressure, heart rate recovery, and ST depression) influenced the target variable (heart disease severity).

B. Managing Model Complexity

Model complexity was managed both at the data level and the structural level. The original dataset contained 14 variables, but after examining correlations and domain relevance, we reduced the network to five core features that are strongly associated with heart disease in medical literature.

This not only improved computational efficiency, by reducing the potential exponential blow-up from adding too many parent nodes to a single variable, but also ensured the model remained interpretable and aligned with real-world diagnostic processes. Each selected variable was also discretized, which reduced the state space of each node and made inference tractable using exact methods.

C. Experiment Design

To test the model, a series of inference queries were designed by comparing the profiles of seen vs unseen patients. For example, when given common evidence (such as middle-aged patients with normal blood pressure and high heart rate), the model returned confident predictions. However, when provided with rare combinations (e.g., old age, high cholesterol, low heart rate, and high ST depression), the model reverted to a uniform probability distribution, indicating that the evidence was not well represented in the training data.

This experiment illustrates how Bayesian Networks handle uncertainty gracefully, and provides insight into the model's ability to extrapolate or fail to, in cases of sparse data.

D. Model Evaluation

Evaluation was conducted through the inspection of learned CPDs, especially the distribution over the target given various evidence. The model outputs interpretable probabilities for each severity level (0 to 4) mild to severe, allowing us to understand the likelihood of each disease state under different input scenarios. This form of probabilistic output is more informative than simple binary classification, as it supports risk-based decision-making in healthcare contexts.

V. Inference and Results Analysis

Target	Probability
0	0.7263
1	0.1819
2	0.0004
3	0.0911
4	0.0004

TABLE I
SEEN EVIDENCE

Target	Probability
0	0.2000
1	0.2000
2	0.2000
3	0.2000
4	0.2000

TABLE II
UNSEEN EVIDENCE

Inference was conducted using Variable Elimination, an exact inference technique supported by the pgmpy library. The goal was to evaluate the model's ability to predict the probability distribution over heart disease severity (target) given different patient profiles (evidence). Two contrasting scenarios were tested to analyse the model's behaviour under both familiar and unfamiliar conditions.

In the first scenario (Table 1 - seen evidence), where the patient profile closely resembled the patterns present in the training data, the model returned a confident prediction, with a 72.63% probability of no heart disease (target = 0) and only minor likelihoods for other severity levels. This demonstrates the model's ability to leverage learned conditional probabilities effectively when evidence is well-represented in the dataset.

In the second scenario (Table 2 - unseen evidence), composed of a rare combination of high-risk features not well represented in the training data and the model returned a uniform probability distribution (20% across all severity levels). This indicates that, in the absence of sufficient prior data, the model refrains from making overconfident predictions, instead defaulting to a maximum entropy (uncertain) distribution. This behaviour reflects robust Bayesian reasoning, relying on observed evidence rather than assumptions.

These experiments also demonstrate an element of robustness and sensitivity: the model behaves predictably when confident, and cautiously when uncertain, a critical quality in high-stakes domains like medical diagnosis. It further highlights the importance of diverse training data, as Bayesian networks depend on prior observations to inform future inferences.

To evaluate the model's sensitivity, a follow-up experiment was performed by modifying a single variable in the original seen evidence. Specifically, blood pressure was changed from 'normal' to 'high', while keeping all other variables constant. The resulting inference is shown below:

Target Level	Probability
0	0.7472
1	0.2497
2	0.0010
3	0.0010
4	0.0010

TABLE III
INFERENCE RESULT AFTER INCREASING BLOOD PRESSURE FROM NORMAL TO HIGH.

Although the highest probability remains at 'target = 0' (no heart disease), the model slightly reduces its certainty and increases the likelihood of 'target = 1', indicating mild disease. This demonstrates the model's ability to respond to changes in individual variables and reflects 'realistic medical reasoning', where elevated blood pressure is a known risk factor. The model's behaviour is both interpretable and sensitive, which are valuable traits in clinical decision support systems. Furthermore, the model assigns non-zero probabilities to higher severity levels (target = 2, 3, 4), albeit very small ($\tilde{0}.1\%$). This further highlights the model's nuanced reasoning and its tendency to reflect increased uncertainty when a known risk factor is introduced.

To further probe the model's reasoning, I crafted a profile with intentionally conflicting signals:

- 'age_cat': 'middle',
- 'bp_cat': 'high',
- 'chol_cat': 'low',
- 'thalach_cat': 'high',
- 'oldpeak_cat': 'moderate'

Surprisingly, the model predicted a 99.67% probability of target = 0 (no heart disease), despite the elevated blood pressure and moderate ST depression. This suggests that the model has learned from the dataset that the combination of low cholesterol and high heart rate recovery is a strong indicator of heart health, outweighing other risk factors. It reflects the data-driven weighting of features which is a core strength of Bayesian networks, where probabilistic reasoning overrides naive heuristics.

This reinforces the model's ability to distinguish noise from meaningful risk, and shows that even when evidence is mixed, it can lean confidently in one direction when supported by data.

To test the model's ability to detect high-risk cases, we constructed a synthetic patient with all known risk factors set to their highest levels. The resulting inference showed nearly 99.4% probability for target levels 3 and 4, which correspond to more severe forms of heart disease. This confirms that the model aligns with medical intuition and handles compound risk well. The input used is:

- 'age_cat': 'old',
- 'bp_cat': 'high',
- 'chol_cat': 'high',
- 'thalach_cat': 'high',
- 'oldpeak_cat': 'high'

As can be seen in the table below:

Target Level	Probability
0	0.0020
1	0.0020
2	0.0020
3	0.4969
4	0.4969

TABLE IV
INFERENCE RESULT AFTER INCREASING DATA TO THE MAXIMUM.

To demonstrate how the model reacts to balanced, mixed-risk profiles, we introduced a moderate-risk case. Unlike the previous confident or uncertain predictions, the model produced a nuanced probability distribution centred around target = 2, with notable weights on 0 and 3. This supports the model's ability to reflect gradual risk accumulation and handle probabilistic reasoning over binary decisions. The input used is:

- 'age_cat': 'middle',
- 'bp_cat': 'normal',
- 'chol_cat': 'high',
- 'thalach_cat': 'normal',
- 'oldpeak_cat': 'moderate'

And the results can be seen in the table below:

Target Level	Probability
0	0.2497
1	0.0010
2	0.4985
3	0.2497
4	0.0010

TABLE V
INFERENCE RESULT AFTER USING MODERATE DATA.

CONCLUSION

This project demonstrated the effectiveness of Bayesian Networks as a probabilistic tool for predicting heart disease severity based on clinical data. By using a well-structured model with carefully selected and discretized features, we were able to capture the conditional dependencies between risk factors and disease outcomes. The use of exact inference via Variable Elimination allowed for interpretable, scenario-specific reasoning, showcasing the model's capacity to handle uncertainty gracefully.

Inference results confirmed the model's ability to return confident predictions for well-represented evidence and fall back to uniform distributions when presented with rare combinations, reinforcing both its robustness and transparency. Sensitivity analysis further highlighted how changes in individual features affect outcomes, aligning with real-world medical intuition.

While the model was constrained by dataset size and variable selection, the project illustrates how probabilistic graphical models can serve as interpretable, trustworthy tools in medical decision support. Future work could involve structure learning, larger datasets, or integrating temporal models for patient monitoring. Ultimately, this approach reflects the potential of combining domain knowledge with probabilistic reasoning to improve outcomes in high-stakes fields such as healthcare.

REFERENCES

[1] D. Heckerman. "A tutorial on learning with Bayesian networks." In Innovations in Bayesian Networks, 2008.

[2] Cleveland Heart Disease Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/heart+disease