Heart disease/failure prediction using Bayesian Networks

COMS4062A - Probabilistic Graphical Models

Luca von Mayer
2427051

I. Problem Statement and Domain Definition

Cardiovascular disease (CVD) is the leading cause of death globally, an estimated 17.9 million deaths per year, which accounts for 31% of all deaths worldwide. With one-third occurring prematurely in individuals under the age of 70. Heart failure is a common and potentially fatal outcome of CVD. Early detection and management of cardiovascular disease risk factors, such as hypertension, diabetes, hyperlipidemia, and existing conditions, are crucial for reducing mortality rates associated with these diseases. [fedesoriano, 2021]

The increasing availability of medical data and modelling methods present an opportunity to develop predictive models that can aid in the identification of heart disease. Graphical Models such as Bayesian networks, offer a promising approach to interpreting complex relationships between various related features and cardiovascular outcomes. By taking advantage these methods, we can expose patterns within the data allowing us to generate predictions that can aid in early diagnosis.

This study aims to develop a Bayesian network model capable of predicting the likelihood of heart disease/failure based on multiple attributes and risk factors from recorded data. This approach has the potential to contribute to the field of cardiovascular disease prevention and management by providing healthcare professionals with a valuable tool for risk assessment. Furthermore, the application of Bayesian networks in this context demonstrates the versatility and effectiveness of these techniques in addressing complex real-world problems within the medical domain.

II. Data Handling and Ethical Considerations

A. Data sourcing

The data sourced for this study is compiled by edesoriano [2021]. This dataset is the result of merging several previously separate datasets. It aggregates data from five distinct datasets, bringing together information on 11 shared features (and the target). As a result, it is the most extensive heart disease dataset currently available. The features covered in this dataset include:

1) **Age:** age of the patient [years]
2) **Sex:** sex of the patient [M: Male, F: Female]
3) **ChestPainType:** chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4) **RestingBP:** resting blood pressure [mm Hg]
5) **Cholesterol:** serum cholesterol [mm/dl]
6) **FastingBS:** fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
7) **RestingECG:** resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality, LVH: left ventricular hypertrophy by Estes' criteria]
8) **MaxHR:** maximum heart rate achieved [Numeric value between 60 and 202]
9) **ExerciseAngina:** exercise-induced angina [Y: Yes, N: No]
10) **Oldpeak:** oldpeak = ST [Numeric value measured in depression]
11) **ST_Slope:** the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12) **HeartDisease:** output class [1: heart disease, 0: Normal]

B. Data Preprocessing

The initial step involved loading the heart disease dataset into a pandas DataFrame, providing a structured format for data manipulation and analysis. To handle categorical variables, label encoding was applied to features like Sex, ChestPainType, RestingECG, ExerciseAngina, and ST-Slope. This assigns numerical values to categorical levels, facilitating their inclusion in the analysis. Numerical features, such as Age, RestingBP, Cholesterol, MaxHR, and Oldpeak, underwent MinMax scaling. This process ensures that features with different scales are treated equally during the modeling process, preventing any one feature from unduly influencing the results due to its magnitude. Following this, the scaled values were discretised and assigned integer values from 0-4.

C. Feature Engineering

Non-linear relationships within the data were captured through feature engineering, which involved binning numerical variables like RestingBP and MaxHR into distinct groups. While feature engineering can enhance model performance, the meaningfulness and interpretability of new features were carefully evaluated.

D. Feature Selection

Feature selection was performed using SelectKBest, to identify the most relevant features for predicting heart disease.

Employing chi-squared and ANOVA tests for categorical and numerical features, respectively. This reduced dimensionality and computational complexity while retaining the most informative features for predicting heart disease. However, after relevant testing it was found that the modelling and prediction was most effective using all provided features.

E. Ethical Considerations

Throughout the data handling process, ethical considerations were prioritized. Techniques like feature engineering and feature selection were applied reasonably, ensuring that the resulting model is not only accurate and interpretable but also free from any biases that could result in the editing of fundamentally sensitive data.

III. PGM Selection and Application

A. Model Selection and Justification

Bayesian networks (BNs) offer a powerful and explainable framework for modeling complex problems involving uncertainty, making them well-suited for the task of predicting heart disease. The choice of BNs as the PGM is justified in their ability to represent and reason about the often convoluted relationships between various variables. BNs provide a intuitive graphical representation of the joint probability distribution, where nodes represent variables and edges capture the dependencies between them. This visual representation displays an effective depiction of the model's structure and the underlying relationships, making it easier to interpret and understand the model's outputs. [Arora et al., 2019]

Moreover, BNs excel at conducting individual-level prediction using Bayes' theorem, which is particularly valuable in the context of risk assessment and precision medicine. Additionally, BNs can be easily transformed into decision models, enabling the incorporation of decision nodes and utility functions, thereby supporting informed decision-making processes in healthcare. They are also generative, allowing new samples to be constructed, which is useful for data acquisition and demonstrating real-world possibilities of the data [Kim and Jung, 2003, Wu et al., 2001]

B. Management of Model Complexity

To address the inherent complexity of Bayesian network (BN) models, the study explores various variants and hyperparameters. This includes different structure learning algorithms, scoring methods, and inference techniques. By evaluating these factors, the study aims to develop a model that balances accuracy with interpretability and computational complexity.

C. Experiment Design and Model Evaluation

The initial step involves preprocessing the data, as outlined previously. The resulting features should exhibit positive values ranging from 0 to 1 for continuous features (MinMax scaling) and have been encoded with integer values for categorical variables, following that the continuous features are discretised and assigned integers ranging from 0-4. This preprocessing step ensures uniformity in the data representation, facilitating subsequent analysis.

The focus shifts to the structure learning phase. Despite the presence of ample expert knowledge in the domain, the adoption of algorithmic structure learning approaches has several advantages. This approach not only enhances robustness but also facilitates exploration of novel, unbiased avenues within the field. The selection of the specific learning technique and its corresponding scoring method will be guided by its performance. It will be chosen based on its ability to achieve optimal scores and other pertinent evaluation metrics.

For evaluating the model, several methods will be utilized :

Correlation Score: This function evaluates how well the model structure captures correlations in the data without parameterisation. It employs the use of the d-connection property of Bayesian Networks to compare variable correlations in the dataset. By testing pairs of variables for correlation and d-connection, it computes a score to assess model performance. This returns a value between 0-1 with higher scores being regarded as more strongly correlated.

Structure Score: Utilising standard model scoring methods, this score assigns a value to each structure. While its interpretation may not be straightforward, it serves to compare different models. A higher score indicates a better fit, and it only relies on the model structure, not requiring parameters.

The Log Likelihood Score will not be considered for evaluation due to its tendency to favor overfitting and more complex models. Additionally, it is computationally expensive, requiring pre-computed CPDs, rendering it infeasible for our purposes. [Ankan, 2023]

Subsequently, the dataset is partitioned into training and testing subsets. This partitioning enables the evaluation of model performance on unseen data, mitigating the risk of overfitting. In this case a 80/20 split of the 918 data points is used, which is deemed a robust partitioning strategy for ensuring adequate model training and evaluation.

Parameter estimation and inference will subsequently undergo evaluation on the test data, utilising a model learned on the training data, selected based on the aforementioned structure learning experiment. This process mirrors the preceding steps,

encompassing a variety of parameter estimation techniques coupled with several inference methodologies to identify the optimal combination yielding the highest scores. Performance assessment will entail computing accuracy and F1 scores, along with a confusion matrix and when applicable the runtime of the respectively algorithm, derived from the predictions made on the test data.

In terms of the inference technique employed for making predictions, the decision has been made to utilise exact inference. In the medical context, it is imperative to generate results with a high level of certainty, given the critical nature of applying such predictions. Estimation-based predictions carry inherent risks, as an erroneous prediction could potentially endanger lives. Within the domain of exact inference, the selection of the specific method will be guided by experimental validation, considering both the results obtained and their associated hyperparameters, and importantly the run time of the method.

IV. Experimental Results Analysis

A. Structure learning

The first structure learning technique evaluated in our experiments was PC (Constraint-Based Estimator). The PC algorithm, a constraint-based method for estimating DAGs from data, operates by identifying conditional dependencies using statistical independence tests. It then constructs a DAG pattern that satisfies these dependencies, which can be further converted into a Bayesian Network. The algorithm employs a maximum of N independence checks, with a worst-case exponential complexity, but it performs efficiently for sparse graphs.

For finite samples, various methods such as significance tests, model selection criteria like chi-square, pearsonr, g-sq, log-likelihood, and cressie-read were evaluated. However, the algorithm may encounter challenges when the underlying probability distribution is unknown and estimated. To mitigate errors, it employs robust procedures, such as avoiding premature acceptance of conditional independences and ensuring consistency in separating sets, thus enhancing the interpretability and validity of the resulting DAG structures. [Lauritzen, 2011]

Following this, the Hill Climb Search method is evaluated. Hill climbing is a heuristic search algorithm for solving optimization problems, including learning network structures from data. The algorithm iteratively improves a solution by making small modifications based on a heuristic function that evaluates the quality of the solution. Starting from an initial state, the algorithm explores neighboring solutions and selects the one that leads to the best improvement. In our experiments, we utilized scoring functions such as BDeu, Bic, K2, and BDs to guide the search process. While hill climbing is simple to understand and implement, it may get stuck in local maxima, sometimes lacking full exploration of the search space. [GeeksforGeeks, 2023]

Lastly, the tree search algorithm was investigated, providing a powerful approach to learning graph structures, particularly for tree-related structures like Chow-Liu and Tree-augmented naive Bayes (TAN). Chow-Liu constructs a maximum-weight spanning tree using mutual information scores, while TAN extends the Naive Bayes classifier by incorporating a tree structure to capture feature interactions. Recent research in supervised learning has demonstrated the effectiveness of Bayesian classifiers, such as naive Bayes, despite their simplistic assumptions of feature independence. This prompted exploration into less restrictive classifiers, such as TAN, which combines the advantages of Bayesian networks with computational efficiency similar to naive Bayes. [Friedman et al., 1997]

These experiments culminated in the following results seen in Figure 1:

From the graph, we can observe that the Bic and K2 variants of the Hill Climb Search algorithm emerged as the best-performing structure learning methods, as evidenced by their higher correlation scores. These scores indicate that the learned structures effectively capture the underlying correlations present in the data. While most of the methods produced similar structure scores, the Hill Climb Search with the K2 scoring criterion exhibited a slight edge, prompting its selection as the model for subsequent experiments.

The learned structure of the Hill Climb Search model using the K2 score Figure 2 reveals interesting insights into the relationships among the various risk factors and the target variable. For instance, the graph demonstrates a direct influence of almost all the features on the likelihood of developing heart disease.

Furthermore, the model captures interesting dependencies between certain features, such as the relationship between sex and chest pain type / cholesterol, as well as the influence of Resting Blood Pressure on the resting electrocardiogram (ECG) results. These connections highlight the interconnected nature of the risk factors and their potential cascading effects on various physiological markers.

The structural representation provided by the Bayesian network can be invaluable for healthcare professionals in understanding the complex relations of different variables and their impact on heart disease. By identifying the most relevant risk factors and their relationships, clinicians can tailor preventive strategies to individual patient profiles.

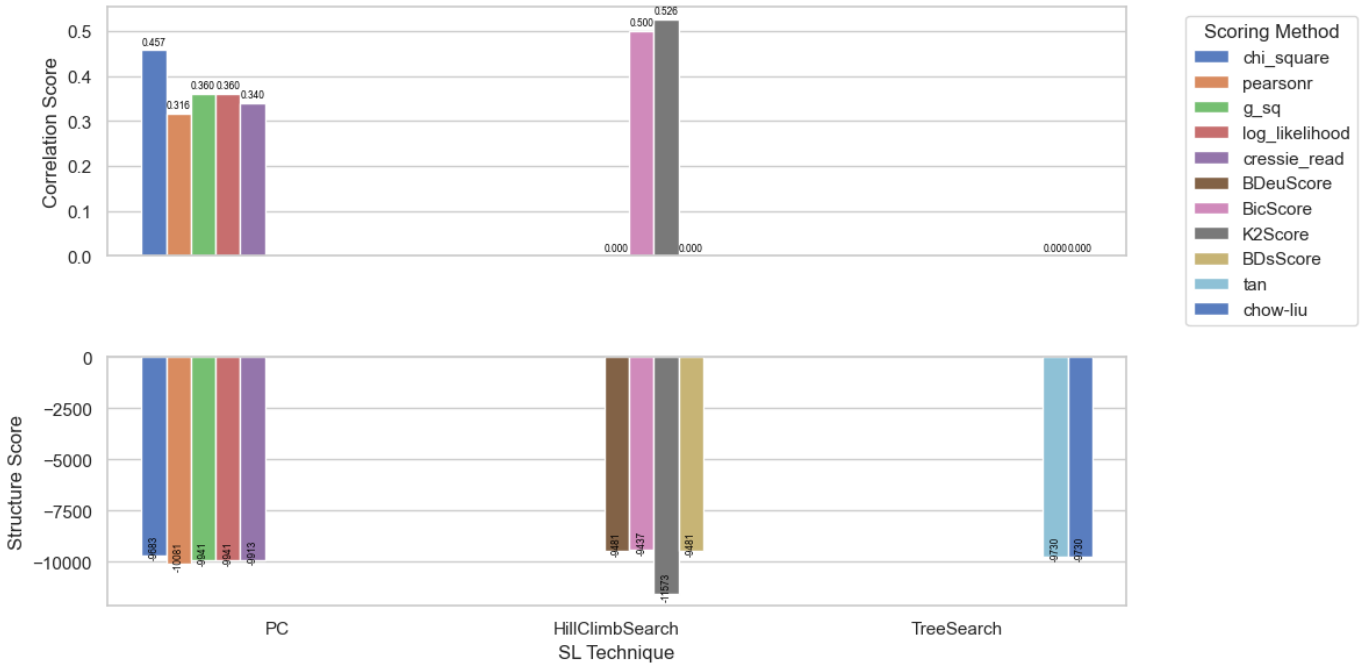

Fig. 1: Structure Learning Results

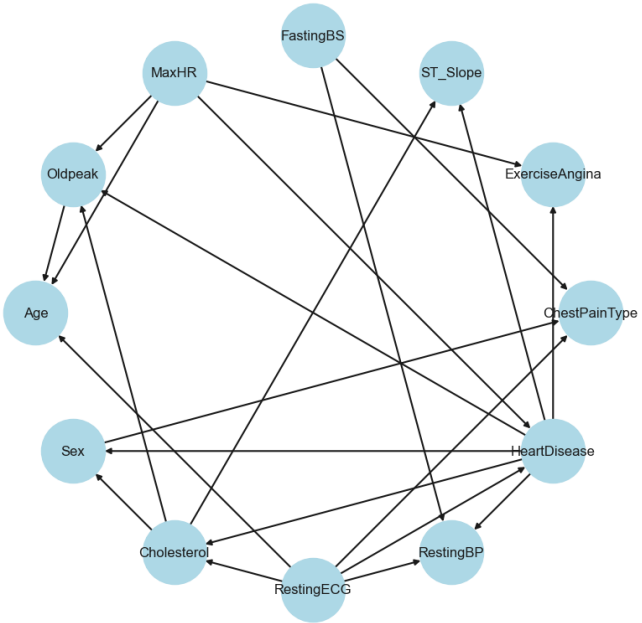

Fig. 2: Bayesian Network: Hill Climb Search using K2 score

B. Parameter Estimation

Maximum Likelihood Estimation (MLE) was employed to estimate the parameters of our distribution based on observed data. MLE seeks to maximize the likelihood function, determining the parameter values that are most likely to generate the observed data. It is known for its simplicity and consistency, ensuring that estimated parameters converge to true values as the sample size increases [Brilliant.org, 2024].

Additionally, we utilized Bayesian parameter estimation with both BDeu and K2 scores, which incorporates prior beliefs about model parameters with observed data to derive a posterior distribution. This approach offers flexibility by incorporating prior knowledge and allows for a more informed inference process. [Butler, 2021].

C. Inference

For exact inference in probabilistic graphical models, we utilized Variable Elimination, which decomposes complex problems into smaller ones, enabling efficient and complete computation of marginals and belief in Bayesian networks. This algorithmic framework unifies approaches from various fields, making it suitable for our diverse experimental needs.

Moreover, Belief Propagation was tested for inference in probabilistic graphical models by propagating evidence through iterative message passing. By creating a Junction Tree or Clique Tree from the input model and calibrating it using belief propagation, this method efficiently calculates marginal and conditional probabilities. Other techniques like Variable Elimination are effective for tree-structured graphs, Belief Propagation addresses issues in loopy graphs by constructing a clique tree, mitigating problems of message circulation and convergence [Stankiewicz and Domański, 2018].

Furthermore, we utilized causal inference to analyze complex relationships within our dataset. Despite challenges such as missing data and noncompliance, causal inference provided a flexible framework for handling these complexities effectively. This method involves formulating falsifiable hypotheses and employing statistical methods to test them, with a focus on discerning causal mechanisms rather than just correlation. [Rubin, 2010]

D. Parameter Estimation and Inference Results

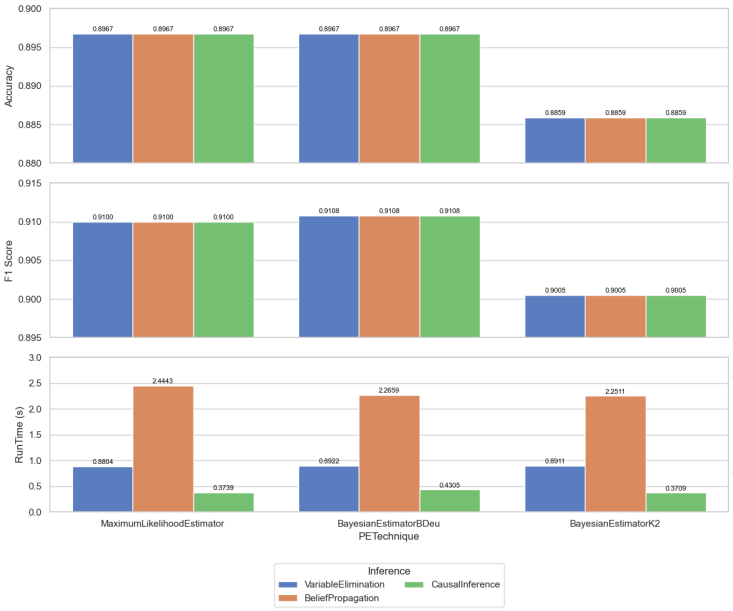

Fig. 3: Parameter Estimation and Inference Results

Upon examining the results presented in Figure 3, it is evident that all employed methods yielded high performance scores. As anticipated, Variable Elimination, Belief Propagation and Causal Inference demonstrated identical accuracy and F1 scores, affirming their status as instances of exact inference methods. The convergence of these techniques underscores the reliability and consistency of the inference process, ensuring that the obtained results are accurate and trustworthy.

Therefore, choosing a inference method for our model needs to be guided by the other factors, one that may account for future changes in our model structure and introduction of new data. In this case, causal inference is useful in that it accounts for missing data, currently our dataset does not have missing values but that could change as more data is introduced. Moreover, it ran the quickest out of the 3 methods, further emphasising its effectiveness in this context.

Among the parameter estimation techniques, Bayesian Estimation utilizing the BDeu scoring criterion marginally outperformed Maximum Likelihood Estimation (MLE). This observation aligns with the theoretical advantages of Bayesian estimation, which incorporates prior knowledge and beliefs about the model parameters in addition to the observed data. By combining these two sources of information, Bayesian estimation can potentially yield more accurate parameter estimates, especially in scenarios where the available data may be limited or subject to noise.

The superior performance of Bayesian estimation with the BDeu scoring criterion suggests that the chosen prior distributions and assumptions effectively captured the underlying characteristics of the heart disease domain. This result highlights the importance of carefully selecting priors and scoring functions in Bayesian parameter estimation, as they can significantly impact the quality of the estimated parameters and the accuracy of the overall model.

V. Conclusion

In conclusion, this study developed an effective Bayesian network model for predicting heart disease, utilising the Hill Climb method with the K2 score for structure learning, Bayesian estimation with the BDeu scoring for parameter estimation, and Causal Inference due to all their apparent benefits. This model configuration achieved high performance while maintaining interpretability and computational efficiency.

The application of Bayesian networks demonstrated their versatility in addressing complex healthcare problems, providing a valuable tool for heart disease risk assessment. The chosen approach balanced accuracy with reliability. Overall, this study showcased the promising potential of Bayesian networks in precision medicine and personalised healthcare.

References

[1] fedesoriano. Heart failure prediction dataset. https://www.kaggle.com/fedesoriano/heart-failure-prediction, September 2021.

[2] edesoriano. Heart failure prediction dataset. https://www.kaggle.com/fedesoriano/heart-failure-prediction, September 2021. Retrieved [Date Retrieved].

[3] Paul Arora, Devon Boyne, Justin J. Slater, Alind Gupta, Darren R. Brenner, and Marek J. Druzdzel. Bayesian networks for risk prediction using real-world data: A tool for precision medicine. *Value in Health*, 22(4):439–445, 2019. ISSN 1098-3015.

[4] In-Cheol Kim and Yong-Gyu Jung. Using bayesian networks to analyze medical data. In Petra Perner and Azriel Rosenfeld, editors, *Machine Learning and Data Mining in Pattern Recognition*, pages 317–327, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45065-8.

[5] Xiaofeng Wu, Peter Lucas, Susan Kerr, and Roelf Dijkhuizen. Learning bayesian-network topologies in realistic medical domains. In Jose Crespo, Victor Maojo, and Fernando Martin, editors, *Medical Data Analysis*, pages 302–307, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-45497-7.

[6] Ankur Ankan. Metrics, 2023. URL https://pgmpy.org/metrics/metrics.html. Copyright ©2023, Ankur Ankan.

[7] Steffen Lauritzen. Estimation of (causal?) structure. https://www.stats.ox.ac.uk/~steffen/teaching/gm11/structure.pdf, 2011. Lecture 14, Michaelmas Term 2011, University of Oxford; November 21, 2011.

[8] GeeksforGeeks. Introduction to hill climbing — artificial intelligence. https://www.geeksforgeeks.org/introduction-hill-climbing-artificial-intelligence/, 2023.

[9] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[10] Brilliant.org. Maximum likelihood estimation (mle). https://brilliant.org/wiki/maximum-likelihood-estimation-mle/, April 2024.

[11] Paul Butler. Intro to bayesian parameter estimation. https://medium.com/analytics-vidhya/intro-to-bayesian-parameter-estimation-f324498bb505, January 2021.

[12] Olgierd Stankiewicz and Marek Domański. Title of the chapter. In *Academic Press Library in Signal Processing*, volume 6. Academic Press, 2018.

[13] D. B. Rubin. Title of the chapter. In *International Encyclopedia of Education*. Third edition, 2010.