### A Bayesian Network Approach for Cardiovascular Risk Prediction: Insights and Applications

#### Abstract

Cardiovascular diseases (CVD) remain the leading cause of mortality worldwide, significantly impacting global health and economic resources. This paper presents the development and application of a Bayesian network model for predicting cardiovascular risk. By integrating extensive health data from annual assessments of Spanish workers with expert knowledge, the model highlights the complex relationships between non-modifiable, modifiable risk factors, and medical conditions. The accompanying software tool offers healthcare providers actionable insights to assess and mitigate cardiovascular risk.

#### 1. Introduction

Cardiovascular diseases (CVD) are responsible for millions of deaths annually, creating an urgent need for effective prevention strategies. Accurate risk prediction enables the identification of high-risk individuals, paving the way for targeted interventions and improved health outcomes. This study develops a Bayesian network to model the interdependencies among cardiovascular risk factors and to evaluate the impact of interventions on reducing disease risk.

The model integrates data from over 200,000 health assessments and expert input, resulting in a software tool that is accessible for research and clinical applications.

#### 2. Materials and Methods

##### 2.1 Data Collection

The dataset includes annual health assessments conducted between 2012 and 2016. The original data comprised over one million records, reduced to 205,087 after cleaning and preprocessing. The data encompasses a comprehensive range of variables:

* Non-modifiable risk factors such as age, sex, socioeconomic status, and education level were included to understand static determinants of cardiovascular health.
* Modifiable risk factors like BMI, physical activity, sleep duration, smoking habits, anxiety, and depression provided insights into areas where interventions could have significant effects.
* Prevalence of medical conditions such as hypertension, hypercholesterolemia, and diabetes were included to assess disease risk and progression.

##### 2.2 Data Preparation

The data underwent a meticulous cleaning process to ensure accuracy and relevance for the model:

* Records containing extreme outliers were removed by applying a three-standard-deviation rule for continuous variables, which eliminated 586 entries suspected of measurement or recording errors.
* Duplicated records and entries with incomplete information were excluded, accounting for 7,689 removals.
* To avoid duplication across time, only the latest assessment for each individual was retained, resulting in a final dataset of 205,087 records.

All variables were discretized for compatibility with Bayesian network modeling. Discretization simplifies the modeling process by transforming continuous data into categorical data, which is essential for Bayesian networks to handle probabilities efficiently. However, this process can also impact accuracy by reducing granularity, particularly for individuals whose values fall near category boundaries. Despite this, discretization enhances interpretability by aligning variables with clinically meaningful thresholds, such as WHO BMI classifications. For instance, BMI was categorized into underweight, normal weight, overweight, and obese using WHO standards, while age was grouped into ranges reflecting life stages.

##### 2.3 Bayesian Network Construction

The Bayesian network was developed in two stages:

1. **Structure Learning**: To identify relationships between variables, the Greedy Thick Thinning (GTT) algorithm was applied. This method utilizes conditional independence tests to uncover dependencies. It is well-suited for processing large datasets, as it systematically refines the network structure to align with observed data, ensuring that key dependencies are accurately represented in the model.
2. **Expert Refinement**: The initial structure was critically reviewed by domain experts who evaluated cause-effect relationships between variables. Over several iterations, they adjusted the structure by adding 15 new edges, reversing the direction of 7, and removing irrelevant connections to ensure the network accurately represented real-world phenomena.

The final network linked modifiable and non-modifiable risk factors to medical conditions, enabling a detailed analysis of their interactions.

##### 2.4 Probability Estimation

The network’s conditional probability tables (CPTs) were estimated using multinomial-Dirichlet models. To address uncertainty, uniform priors were used because they allow for an unbiased starting point when estimating probabilities in the absence of strong prior knowledge. This choice ensures that all possible outcomes are treated equally initially, while posterior distributions refine these probabilities based on the observed data, resulting in robust and data-driven predictions. This approach ensured that even with sparse data in certain nodes, the model remained robust.

##### 2.5 Model Validation

To assess the model’s predictive power, a 5-fold cross-validation method was used. The model was trained on 80% of the data and tested on the remaining 20%, iteratively for five cycles. The results demonstrated:

* A predictive accuracy of 91% for modifiable risk factors such as BMI and physical activity levels.
* A predictive accuracy of 93% for medical conditions like hypertension and diabetes.

##### 2.6 Software Implementation

The Bayesian network’s accompanying software tool is a core component of its practical application. Built on the GeNIe platform, this tool enables clinicians and researchers to:

* Input patient-specific data and generate probabilistic predictions for medical conditions and risk factors.
* Simulate potential interventions by adjusting modifiable factors, such as BMI and physical activity, to assess their impact on outcomes.
* Explore the intricate interdependencies among risk factors in real-time, facilitating actionable insights for decision-making at both individual and population levels.

The software’s open-source availability on GitHub ensures accessibility for researchers to further customize and refine the tool according to specific needs and applications. This accessibility fosters collaborative advancements and broadens its utility across diverse healthcare environments. Future enhancements to the model will focus on expanding its applicability, improving its dynamic capabilities, and integrating real-time data sources. Future enhancements to the model will focus on expanding its applicability, improving its dynamic capabilities, and integrating real-time data sources.

#### 3. Results

##### 3.1 Individual Risk Prediction

The Bayesian network allows for personalized risk assessments by calculating the probability of developing medical conditions based on individual profiles. For example, a clinician might use the tool to assess the risk of diabetes in a middle-aged patient with a sedentary lifestyle and elevated BMI, guiding them to recommend targeted lifestyle modifications such as increased physical activity and dietary changes. Additionally, the network could help identify high-risk groups in a population, such as individuals with multiple overlapping risk factors, enabling proactive intervention strategies. For example, a 50-year-old male who is obese, physically inactive, and sleeps fewer than six hours per night is shown to have a significantly elevated risk of hypertension. Such detailed predictions enable tailored healthcare recommendations.

##### 3.2 Population-Level Insights

The model also identifies population-level trends. For example, individuals from lower socioeconomic groups were found to have higher incidences of hypertension and diabetes, emphasizing the role of social determinants in cardiovascular health. This insight supports targeted public health policies to reduce health disparities.

##### 3.3 Intervention Analysis

The Bayesian network is a valuable tool for assessing the impact of lifestyle changes on cardiovascular risk. For instance, increasing physical activity from insufficient to regular levels reduces the likelihood of hypertension by approximately 25%. This functionality allows clinicians to prioritize interventions based on their effectiveness.

#### 4. Discussion

##### 4.1 Advantages of the Bayesian Network Approach

1. **Comprehensive Analysis of Risk Factors**: The model integrates modifiable and non-modifiable factors, providing a holistic view of cardiovascular risk. This integration ensures that static attributes such as age and sex are analyzed alongside dynamic lifestyle factors like BMI and physical activity. By combining these elements, the model can identify patterns that may not be apparent when examining risk factors in isolation. For instance, it highlights how socioeconomic status interacts with lifestyle behaviors to influence hypertension risk, enabling clinicians to develop more nuanced and effective intervention strategies.
2. **Dynamic Predictions**: The Bayesian framework allows for probabilistic reasoning, meaning that predictions are updated as new evidence becomes available. For example, if a patient’s lifestyle factors, such as physical activity or smoking status, change over time, the model can dynamically adjust risk predictions to reflect these updates. This capability provides clinicians with a responsive tool for tailoring recommendations and monitoring patient progress.
3. **Causal Relationships**: By incorporating expert input, the model reflects causal pathways, enabling clinicians to identify direct influences between risk factors and outcomes. For instance, it elucidates how physical inactivity directly increases the likelihood of hypertension, rather than simply correlating the two. This level of understanding makes the model highly actionable for designing targeted and effective interventions, as it distinguishes between causal and coincidental relationships.
4. **Transparent and Interpretable**: The graphical representation of relationships simplifies complex interdependencies, making it easier for clinicians and researchers to understand and communicate findings. This clarity fosters confidence in the model’s predictions and supports its use in both academic and practical healthcare settings.
5. **Scalability**: The model effectively handles large datasets by utilizing advanced probabilistic algorithms that ensure computational efficiency. Specifically, the Greedy Thick Thinning (GTT) algorithm minimizes unnecessary computational overhead by focusing on the most relevant dependencies among variables. This optimization, combined with parallel processing capabilities and robust data handling techniques, allows the model to process extensive records while maintaining high accuracy. These features make it well-suited for analyzing large-scale population data, ensuring timely and reliable predictions.

##### 4.2 Limitations

1. **Static Nature**: The current model operates as a static snapshot, which does not account for temporal dynamics. This limitation restricts its ability to monitor changes in risk factors or outcomes over time, such as the progression of lifestyle improvements or the impact of long-term interventions. Developing a dynamic Bayesian network would address this by enabling longitudinal tracking and prediction of evolving health risks.
2. **Specific Population**: The dataset primarily represents Spanish workers, which introduces the possibility of limited generalizability to other populations with different demographic or health characteristics. Factors like cultural differences in diet, lifestyle, and healthcare access may influence the applicability of the findings. Therefore, the model should be recalibrated with data from diverse populations to ensure broader relevance and accuracy in varied healthcare contexts.
3. **Excluded Variables**: Factors like diet and alcohol consumption were not included, which could limit the comprehensiveness of the model in capturing all significant influences on cardiovascular health. For example, dietary habits play a critical role in modulating risk factors such as hypertension and diabetes. Similarly, alcohol consumption, depending on frequency and quantity, could interact with other variables to influence outcomes. Incorporating these variables in future iterations of the model would provide a more holistic perspective on cardiovascular risk.
4. **Computational Demands**: Training and inference for large and complex networks require advanced computational resources due to the high-dimensional data and probabilistic computations involved. However, implementing optimizations like parallel processing, efficient algorithms such as Greedy Thick Thinning (GTT), and streamlined memory management techniques can significantly reduce overhead, making large-scale applications more feasible.

##### 4.3 Future Directions

1. **Incorporating New Data**: Future iterations of the Bayesian network should integrate additional variables, including diet, alcohol consumption, and genetic information. These elements are critical in understanding cardiovascular risk holistically, as diet influences key factors like cholesterol and blood pressure, alcohol consumption can exacerbate or mitigate certain risks depending on usage patterns, and genetic predispositions provide foundational insights into non-modifiable risk factors. Incorporating such data would enhance the model’s accuracy and applicability in predicting and managing cardiovascular diseases.
2. **Dynamic Modeling**: Developing a dynamic Bayesian network would allow for longitudinal analyses, capturing the evolution of risk factors over time. For example, tracking changes in physical activity or BMI could enable personalized treatment adjustments by predicting how these evolving factors influence cardiovascular outcomes. Longitudinal modeling also facilitates early identification of trends, such as deteriorating health metrics, allowing timely interventions to mitigate potential risks.
3. **Expanding to Diverse Populations**: Applying the model to data from different age groups, ethnicities, and socioeconomic backgrounds would significantly enhance its generalizability. For instance, incorporating data from underrepresented groups could reveal unique risk patterns or interactions between factors, such as the interplay of cultural dietary habits and socioeconomic status in influencing cardiovascular health. This broader scope would allow the model to provide tailored predictions and interventions for diverse global populations.
4. **Real-Time Integration**: Connecting the model to wearable devices and electronic health records would enable real-time monitoring and prediction. For instance, wearable devices could provide continuous updates on physical activity, heart rate, or sleep patterns, allowing the model to dynamically adjust risk estimates. Similarly, integration with electronic health records could ensure that clinical data such as lab results and medication adherence are incorporated seamlessly, enhancing the model's predictive accuracy and utility in personalized healthcare.
5. **Cost-Effectiveness Analysis**: Incorporating decision and utility nodes into the Bayesian network would enable detailed evaluations of the financial implications of various interventions. For instance, the model could compare the cost of lifestyle modification programs (e.g., increasing physical activity or improving diet) with their projected benefits in reducing long-term healthcare expenses. This capability would support policymakers and clinicians in prioritizing cost-effective strategies while maximizing health outcomes.

#### 5. Conclusion

The Bayesian network developed in this study offers a robust framework for predicting cardiovascular risk and analyzing the interplay of various risk factors. By combining data-driven insights with expert knowledge, the model supports precision medicine and public health strategies. The accompanying software tool, implemented using the GeNIe platform, allows users to interact with the Bayesian network by inputting specific patient data to receive probabilistic predictions for various risk factors and medical conditions. Clinicians can easily adjust variables, such as BMI or physical activity, to simulate interventions and observe their effects. Additionally, the software supports real-time exploration of complex relationships between risk factors, offering actionable insights for both individual and population-level strategies. Its accessibility through an open-source GitHub repository ensures that researchers can further customize and adapt the tool for diverse applications.