A Probabilistic Approach to Diabetes Risk Assessment Using Bayesian Networks

Faezeh Sarlakifar
Master's Degree in Artificial Intelligence, University of Bologna
faezeh.sarlakifar@studio.unibo.it

February 12, 2025

Abstract

Diabetes is a chronic and widespread health condition that affects many individuals worldwide. Early prediction and risk assessment are essential for effective prevention and management. In this study, the application of Bayesian Networks (BNs) to model the probabilistic relationships between health indicators and diabetes risk is explored. Multiple Bayesian structures are developed using Naïve Bayes, Hill Climbing (with various scoring methods), Simulated Annealing, and domain knowledge-based techniques. To enhance model interpretability and predictive performance, a comprehensive feature selection approach is employed. The results demonstrate that Bayesian Networks offer an interpretable and robust framework for diabetes risk prediction, highlighting their potential for real-world healthcare applications.

Introduction

Domain

In the field of medical health assessment, the interpretability and reliability of Artificial Intelligence (AI) methods are crucial, especially when it comes to conditions like diabetes. Healthcare professionals must be able to understand and trust the outcomes of predictive models to make informed decisions about patient care. Bayesian Networks (BNs) are particularly suitable for this purpose in diabetes risk assessment, as they provide an interpretable framework that allows practitioners to understand the probabilistic relationships between various health indicators, such as age, BMI, and their impact on diabetes risk. This transparency is essential for building trust in AI models and ensuring they can be effectively integrated into clinical decision-making processes, ultimately improving patient outcomes.

This research is inspired by the study conducted by Kong et al. [2], whose findings have been utilized to design the domain knowledge-based Bayesian Network model.

Aim

The goal of this project is to explore the effectiveness of Bayesian Networks (BNs) in modeling the probabilistic relationships between health indicators and diabetes risk. By leveraging different structure-learning approaches, the study aims to develop a robust and interpretable framework for diabetes risk assessment.

Method

In this study, a Bayesian Network (BN) is built and evaluated using the "`pgmpy`" library. Several BN structures were explored, including Naïve Bayes, Hill Climbing (using various scoring functions: BIC, K2, and BDeu), Simulated Annealing, and a Domain Knowledge-based model.

To improve predictive performance and interpretability, a feature engineering step is applied to select the most important features for building the Bayesian Network. SHAP (SHapley Additive exPlanations) values are employed to identify features importance, implemented using the "`SHAP`" Python library. These values are computed by training an XGBoost classifier [1] and analyzing its explanations using TreeExplainer, a method specifically designed for tree-based models to capture complex nonlinear relationships. Since SHAP values provide a consistent measure of feature contribution in such models [3], they serve as a robust criterion for feature selection.

For parameter learning, the Conditional Probability Distributions (CPDs) are estimated using Maximum Likelihood Estimation (MLE) due to its efficiency in deterministic settings. The trained models are then evaluated using AUC-ROC scores to assess their classification performance. The top-performing Bayesian Networks are then employed for probabilistic reasoning, enabling predictions and risk assessments based on given evidence, with Variable Elimination used as the inference method.

Results

The experimental results show that the Hill Climbing method (with K2 scoring), with an AUC score of 81.30, and the Domain Knowledge-driven method, with an AUC score of 83.57, effectively capture the relationships between features for diabetes risk assessment. These results suggest that both data-driven and expert-informed Bayesian Network structures can model the probabilistic relationships among health indicators.

Model

In the proposed Bayesian Network, nodes represent key diabetes risk factors, while directed edges capture probabilistic dependencies based on expert knowledge and prior research.

Figure 1 presents the Domain Knowledge-Based Bayesian Network, illustrating the probabilistic dependencies between features.

To establish the network structure, two approaches are employed:

1. Data-Driven Learning: Bayesian structures are learned using algorithms including Hill Climbing (BIC, K2, BDeu) and Simulated Annealing, optimizing for the highest AUC-ROC score. These models establish probabilistic relationships purely from data. Using SHAP values, ten critical variables were identified: "HighBP[1]", "GenHlth[2]", "HighChol[3]", "Age", "Sex", "Income", "DiffWalk[4]", "BMI", "HeartDiseaseorAttack", and "Education". These variables serve as nodes in the AI-based Bayesian Networks.
2. Domain Knowledge-Based Network: A causally meaningful structure is designed based on established medical research, ensuring the relationships between variables align with known diabetes risk factors. The edges in this network reflect dependencies supported by scientific studies.

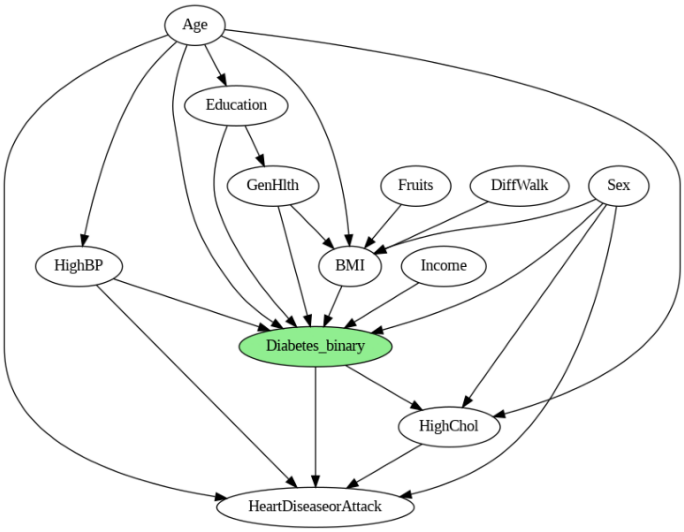

Figure 1: Proposed Bayesian Network.

For parameter learning, Maximum Likelihood Estimation (MLE) is applied to estimate the Conditional Probability Distributions (CPDs).

Analysis

Experimental setup

This study employs the Diabetes Health Indicators dataset from the Behavioral Risk Factor Surveillance System (BRFSS). This dataset consists of 21 health-related features, which include demographics, lifestyle factors, and medical conditions. This makes the dataset a valuable resource for assessing diabetes risk. The dataset is then divided into training and validation sets. The training set is used to fit models, while the validation set serves to evaluate their performance. Models are compared using the AUC-ROC evaluation metric.

[1]Indicates whether a person has high blood pressure.
[2]general health status.
[3]Presence of high cholesterol.
[4]Difficulty in walking.

It was hypothesized that the domain knowledge-based model would outperform the others, as it is inspired by insights from scientific literature. The results support this hypothesis, demonstrating the superiority of the domain knowledge-based approach.

Results

Table 1 presents a comparison of the various methods experimented for constructing the Bayesian Network. The results show that the Hill Climbing algorithm with the K2 scoring method outperforms other AI-based models, and the domain knowledge-based approach surpasses all. Although the simulated annealing model was anticipated to perform better than Hill Climbing, the Hill Climbing with K2 scoring method achieved higher AUC score for diabetes risk assessment, which is an intriguing finding.

Table 1: Comparison of Bayesian Network Models

Bayesian Network Model	AUC Score
Hill Climbing (BIC)	76.80
Hill Climbing (K2)	81.30
Hill Climbing (BDeu)	76.50
Naïve Bayes	72.40
Simulated Annealing	77.90
Proposed Model (domain knowledge-based)	**83.57**

Conclusion

This study validates the effectiveness of Bayesian Networks for diabetes risk assessment, achieving an AUC-ROC score of 83.57 through the integration of domain knowledge from scientific literature. Furthermore, with increased computational resources for Bayesian structure learning and parameter estimation, the models could potentially achieve higher predictive performance. Overall, the findings demonstrate that Bayesian Networks provide a valuable framework for diabetes risk prediction.

Links to external resources

- The source code of this project is available at: GiHub repository
- You can find the utilized dataset at: Kaggle diabetes health indicators dataset

References

[1] Tianqi Chen and Carlos Guestrin. "xgboost: A scalable tree boosting system". page 785–794, 2016. doi: 10.1145/2939672.2939785.

[2] Danli Kong, Rong Chen, Yongze Chen, Le Zhao, Ruixian Huang, Ling Luo, Fengxia Lai, Zihua Yang, Shuang Wang, Jingjing Zhang, Hao Chen, Zhenhua Mai, Haibing Yu, Keng Wu, and Yuanlin Ding. "bayesian network analysis of factors influencing type 2 diabetes, coronary heart disease, and their comorbidities". *BMC Public Health*, 24, 2024. doi: 10.1186/s12889-024-18737-x.

[3] Scott M. Lundberg and Su-In Lee. "a unified approach to interpreting model predictions". page 4768–4777, 2017.