EconoBayes: Analyzing Income through Bayesian Networks

Abdul Moqeet
Master's Degree in Artificial Intelligence, University of Bologna
abdul.moqeet@studio.unibo.it

April 9, 2025

Abstract

This project employs Bayesian networks to assess income levels based on various socioeconomic factors. Initially, the domain-based model exhibited lower accuracy compared to the data-driven approach but demonstrated superior inference performance. By integrating significant data-driven insights into the domain-based model, we achieved enhanced accuracy upto 80 percent while maintaining robust inference capabilities. These findings underscore the effectiveness of a hybrid modeling approach that combines domain expertise with data-driven methodologies for income assessment.

Introduction

Domain

In this project, I built a Bayesian Network to predict income based on various socioeconomic factors and got inspiration from a paper (Ahmed Abd Elrahman 2024). Bayesian Networks are great for modeling complex relationships and understanding how different factors influence outcomes. This approach helps identify which variables have the most impact on income while also providing a clear, logical structure for analysis. By using this model, we can better understand the dependencies between different attributes and how they contribute to income variations.

Aim

The main aim of this project is to analyze whether a data-driven approach or a domain-knowledge-based approach works better for predicting income levels. By building a Bayesian Network, I aim to compare both methods in terms of accuracy and inference performance. Additionally, the project seeks to identify key factors that significantly influence income, especially those that might be overlooked in daily life. Understanding these factors can provide deeper insights into income distribution and socioeconomic patterns. Ultimately, the goal is to explore which approach pure data-driven modeling or expert-driven domain knowledge yields better results and whether a combination of both can enhance predictive performance.

Method

To build and test the Bayesian Network (BN) classifiers, I used the 'pgmpy' library for model creation and analysis. I applied a data-driven approach using hill climbing to discover the underlying connections between nodes. For domain knowledge, I incorporated scientific literature to establish edges within the network. Additionally, I utilized the NetworkX library to identify the most important nodes by examining centrality through edges. This helped improve the accuracy and performance of the models. I also visualized the networks to observe patterns and refine the inference process, leading to better results overall.

Results

I discovered that the structure of a Bayesian network plays a crucial role in preserving its semantic meaning. While data-driven models generally yield better accuracy, the variable elimination method benefits more from incorporating domain knowledge. Therefore, a hybrid approach, combining both data-driven and domain knowledge methods, is ideal for achieving the best results, as it leverages the strengths of both approaches.

Model

Before building the model, I thoroughly studied the dataset, which included both continuous values and categorical columns with many unique options. To handle the continuous values, I binned them after exploring the data, grouping them based on frequency. For columns with many categorical options, such as the ”education” node, which had over 15 possible values, I simplified the categories into broader groups like specialized courses, high school, and basic education. Next, I applied the hill-climbing method, a greedy data-driven approach, to maximize the model's performance, starting with an empty network. I used the maximum likelihood estimator since I lacked semantic meaning for the data. For the domain knowledge model, I used the Bayesian estimator, establishing causal relationships through logic and literature. I further improved the model by incorporating edge centrality and performed inference to evaluate the model's performance in both approaches.

Analysis

Experimental setup

At the beginning, I achieved around 80 percent accuracy with the data-driven approach and about 75 percent with the

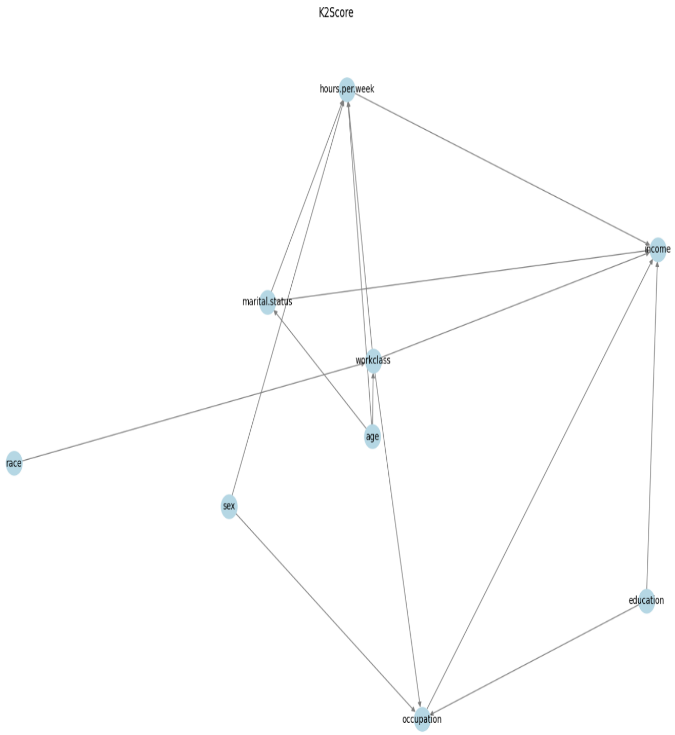

Figure 1: Bayesian network : Hybrid Approach

domain knowledge model. I expected the data-driven model to perform well in inference as well, but the results were the opposite. It turned out that some nodes were poorly connected to the income variable, which negatively impacted the data-driven approach. However, after incorporating hidden data patterns into the domain knowledge model, I was able to achieve 80 percent accuracy, while also seeing improved performance during inference.

Results

Bayesian Model	Accuracy
Data Driven	0.80
Domain Knowledge	0.76
Hybrid	0.806

Table 1: Results

Conclusion

I tested both approaches, but in my case, I found that a hybrid model worked best, as it combines both semantic meaning and underlying data connections. Using this approach, I achieved 80 percent accuracy along with better inference results. Since I worked with a dataset of around 30k records, I believe that for smaller datasets (e.g., around 1k records), the domain knowledge model might not require much assistance from the data-driven method. In such cases, domain knowledge alone could be sufficient.

Links to external resources

The code, dataset, and LaTeX file have been uploaded on Virtuale The notebook containing the project is available on GitHub. Refer (Bansal) for the dataset.

References

[Ahmed Abd Elrahman 2024] Ahmed Abd Elrahman, Marwan R Riad, M. M. 2024. Predicting adults' income using naive bayes classifier. 1–7.

[Bansal] Bansal, L. Adult census income. Retrieved: 10-03-2025.