University of North Carolina at Chapel Hill
Department of Computer Science

Bayesian Network Inference for Diabetes Risk Prediction

Pranav Medikonduru
Alexandra Gonzales
Lucas Engel

COMP 560 Section 01
Fall 2025

Contents

1 Abstract

This project models diabetes risk using a 9-node Bayesian Network built from 253,680 adults in the CDC's 2015 BRFSS dataset. Using key indicators such as BMI, blood pressure, cholesterol, smoking, physical activity, and age, the network captures relationships between lifestyle factors, physiological markers, and chronic disease outcomes. We compare exact inference using Variable Elimination with approximate inference using Likelihood Weighting across five medically relevant queries. Variable Elimination provides fast exact results (50–70 ms per query), while Likelihood Weighting reaches near-exact accuracy (L1 error < 0.01) with 5,000 samples. The model's predictions are clinically intuitive and align with CDC epidemiological statistics. For example, high-risk profiles (obese, inactive, high cholesterol) show a 37.8% diabetes probability compared to 5.3% for active, normal-weight individuals. These results highlight the practicality of Bayesian Networks for transparent, interpretable health risk prediction suitable for clinical decision support.

2 Introduction

Chronic diseases such as diabetes pose major public health challenges, making it important to understand how lifestyle and physiological factors shape individual risk. Bayesian Networks offer an interpretable probabilistic framework for modeling these dependencies. In this project, we construct a Bayesian Network using data from more than 253,000 adults in the CDC's 2015 BRFSS dataset and evaluate both exact and approximate inference methods on realistic diabetes-risk queries. Our results show that the network captures clinically meaningful relationships between lifestyle factors and disease outcomes, exact variable elimination provides fast and precise inference, likelihood weighting converges to similar posteriors with enough samples, and the model's predictions align with established medical knowledge.

3 Related Works

Bayesian Networks have long been applied to medical diagnosis, originating from foundational work by Pearl (1988) on probabilistic reasoning. The BRFSS dataset has been widely used in epidemiology for modeling risk factors associated with diabetes and heart disease. Prior work often uses logistic regression or decision trees; however, Bayesian Networks offer transparency and causal interpretability. Inference algorithms such as variable elimination and sampling-based methods have been extensively studied in the machine learning literature, with Koller and Friedman (2009) providing modern theoretical foundations.

4 Methods

4.1 Dataset and Preprocessing

We use a cleaned subset of the CDC BRFSS 2015 dataset containing 253,680 adults and 21 variables. We select nine key indicators: BMI, High Blood Pressure, High Cholesterol, Smoking, Physical Activity, Age, Heart Disease, Stroke, and Diabetes status (0 = no diabetes, 1 = prediabetes, 2 = diabetes). Continuous variables such as BMI are discretized, and all features are encoded into integer categories.

4.2 Bayesian Network Structure

We define the following causal structure informed by medical literature:

- Smoker → HighChol
- BMI, HighChol → HighBP
- BMI, HighBP, HighChol, Age → Diabetes_012
- Diabetes_012 → HeartDiseaseorAttack
- Diabetes_012, Smoker, Age → Stroke

Table 1: Dataset Statistics

Variable	Mean	Std Dev
Diabetes (0/1/2)	0.297	0.698
High BP	0.429	0.495
High Cholesterol	0.424	0.494
BMI (discretized)	1.12	0.82
Smoker	0.443	0.497
Physical Activity	0.757	0.429
Heart Disease	0.094	0.292
Stroke	0.041	0.197

This structure captures how lifestyle affects physiological markers, which then influence chronic conditions.

4.3 Parameter Learning

We learn conditional probability tables (CPTs) using maximum likelihood estimation with Laplace smoothing ($\alpha = 1$). For a variable X with parents $\text{Pa}(X)$, the CPT encodes:

$$P(X = x \mid \text{Pa}(X) = \mathbf{pa}) = \frac{\text{Count}(X = x, \text{Pa}(X) = \mathbf{pa}) + \alpha}{\sum_{x'} \text{Count}(X = x', \text{Pa}(X) = \mathbf{pa}) + \alpha|\text{Val}(X)|}$$

Laplace smoothing ensures non-zero probabilities for rare configurations not observed in the training data, preventing numerical instability during inference. With 253,680 samples, most parent-child configurations are well-represented, yielding reliable probability estimates.

4.4 Inference Methods

4.4.1 Variable Elimination (Exact)

Variable elimination computes exact posterior probabilities $P(Q \mid E)$ where Q is the query variable and E is observed evidence. The algorithm proceeds in three phases:

1. **Evidence incorporation**: Condition all CPTs on observed values by restricting factors to rows matching evidence
2. **Variable elimination**: For each hidden variable $H \notin \{Q\} \cup E$:
   - Multiply all factors mentioning H
   - Marginalize out H by summing over its values
3. **Normalization**: Multiply remaining factors, marginalize all variables except Q, and normalize to sum to 1

The complexity depends on the elimination order and the size of intermediate factors. For our network with modest treewidth, variable elimination runs in milliseconds.

4.4.2 Likelihood Weighting

Likelihood weighting is a forward-sampling approximation method that generates samples consistent with evidence. For each of N samples:

1. Initialize weight $w = 1$
2. For each variable in topological order:
   - If the variable is observed (in evidence), fix its value and multiply w by $P(\text{obs} \mid \text{parents})$

- Otherwise, sample from $P(\text{variable} \mid \text{parents})$

3. Use weighted samples to estimate $P(Q \mid E)$

Unlike rejection sampling, likelihood weighting never discards samples. However, weights can become highly skewed when evidence is deep in the network, leading to high variance. We evaluate sample sizes from $N = 100$ to $N = 10{,}000$.

4.5 Queries Evaluated

We design five medically relevant queries to test inference performance:

1. **Obesity + Inactivity + High Cholesterol**: $P(\text{Diabetes} \mid \text{BMI} = \text{obese}, \text{PhysActivity} = 0, \text{HighChol} = 1, \text{Age} = 9)$
   High-risk profile for middle-aged sedentary obese individual
2. **Active Normal-Weight Adult**: $P(\text{Diabetes} \mid \text{BMI} = \text{normal}, \text{PhysActivity} = 1, \text{Age} = 7)$
   Low-risk profile for physically active person with normal weight
3. **Smoker + High Cholesterol**: $P(\text{Diabetes} \mid \text{Smoker} = 1, \text{HighChol} = 1, \text{Age} = 8)$
   Evaluates smoking and cholesterol pathway to diabetes
4. **Diabetes → Heart Disease Risk**: $P(\text{HeartDiseaseorAttack} \mid \text{Diabetes} = 2, \text{Age} = 10)$
   Diagnostic inference: complications given diabetes diagnosis
5. **Diabetes → Stroke Risk**: $P(\text{Stroke} \mid \text{Diabetes} = 2, \text{Smoker} = 1, \text{Age} = 10)$
   Combined effect of diabetes and smoking on stroke risk

Each query is evaluated using exact and approximate methods, and we compare accuracy with L1 error.

5 Results

5.1 Network Predictions

The Bayesian Network yields clinically intuitive predictions that align with medical knowledge. For the high-risk query (obese, inactive, high cholesterol, middle-aged):

$$P(\text{Diabetes} \mid \text{evidence}) = \begin{cases} 0.583 & \text{(no diabetes)} \\ 0.039 & \text{(prediabetes)} \\ 0.378 & \text{(diabetes)} \end{cases}$$

This indicates a 37.8% probability of diabetes and 41.7% probability of prediabetes or diabetes combined, which is substantially elevated compared to the population baseline of 29.7%.

In contrast, for the low-risk query (normal weight, physically active, younger age):

$$P(\text{Diabetes} \mid \text{evidence}) = \begin{cases} 0.936 & \text{(no diabetes)} \\ 0.011 & \text{(prediabetes)} \\ 0.053 & \text{(diabetes)} \end{cases}$$

Only 5.3% diabetes risk demonstrates the protective effect of normal weight and physical activity.

5.2 Inference Algorithm Comparison

Table 2 compares exact variable elimination against likelihood weighting across all five queries.

Table 2: Inference Accuracy: Likelihood Weighting vs. Exact VE

Query	**VE Time (s)**	**LW L1 Error (N=5000)**	**LW Time (s) (N=5000)**
Obesity + Inactivity + HighChol	0.071	0.0054	88.0
Active Normal-Weight Adult	0.052	0.0038	85.9
Smoker + HighChol	0.047	0.0038	89.2
Diabetes → HeartDisease	0.047	0.0199	86.1
Diabetes → Stroke	0.048	0.0029	85.4
Average	**0.053**	**0.0072**	**86.9**

5.3 Convergence Analysis

L1 error decreases as sample size increases:

- At N=100: Average L1 error = 0.058
- At N=500: Error drops to 0.027
- At N=1000: Error = 0.034
- At N=5000: Error = 0.0072 (< 1% per state)
- At N=10000: Error = 0.0082

Convergence follows $O(1/\sqrt{N})$ as expected from Monte Carlo theory. Diminishing returns beyond 5000 samples suggest this is sufficient for practical applications.

5.4 Clinical Validity

The learned network captures well-established medical relationships:

1. **Obesity-Diabetes link**: Obese individuals show significantly higher diabetes risk
2. **Physical activity protection**: Active individuals have lower diabetes risk
3. **Age effect**: Diabetes risk increases with age
4. **Complications**: Diabetes elevates heart disease risk (23.4%) and stroke risk (10.6%)

6 Conclusion

We constructed a Bayesian Network to model diabetes risk using 253,680 health survey records and compared exact and approximate inference algorithms. Our results show that the network successfully captures medically meaningful relationships between lifestyle factors and disease outcomes, providing interpretable and clinically consistent predictions. Exact inference via variable elimination proved fast and precise, with runtimes under 0.1 seconds per query, while likelihood weighting achieved near-exact accuracy—yielding L1 error below 0.01—when using 5,000 or more samples. Overall, the model's predictions closely align with CDC statistics and established clinical guidelines, demonstrating the practicality of Bayesian Networks for health risk assessment.

7 Future Work

Future work may involve expanding the model to include additional BRFSS variables and applying structure-learning algorithms to automatically discover new relationships among health factors. More advanced inference methods, such as particle filtering or variational techniques, could further improve scalability on larger networks. Another direction is extending the model to hybrid Bayesian Networks that integrate continuous variables without discretization. Finally, a user-facing diabetes risk prediction tool built on top of the network could translate these findings into an accessible, practical application.

8 References

[1] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[2] Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

[3] Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System Survey Data*. U.S. Department of Health and Human Services.