Bayesian Network Modeling for Diabetes Risk Prediction

Pranav Medikonduru, Alexandra Gonzales, and Lucas Engel

Understanding Diabetes Risk With Bayesian Networks

Motivation

- Diabetes affects millions in the U.S. and is strongly linked to many lifestyle and metabolic factors
- We want to understand which factors matter most and how they interact to influence diabetes risk.

Dataset

- 253,680 adult survey responses from the CDC BRFSS (2015).
- Includes health/lifestyle variables (BMI, activity, smoking), metabolic factors (cholesterol, blood pressure), and outcomes (diabetes, heart disease, stroke).

Goal

- Built a Bayesian Network to quantify diabetes risk and evaluate approximate inference vs exact inference

Bayesian Network Structure + Inference Methods

- Selected 9 key variables from the dataset
- Exact Inference: Variable Elimination over CPT factors.
- Approximate Inference: Likelihood Weighting with N samples.
- Compared LW at N = 100, 500, 1k, 5k, 10k.

```python
# Bayesian network structure
nodes = [
    "Age",
    "Smoker",
    "PhysActivity",
    "BMI",
    "HighChol",
    "HighBP",
    "Diabetes_012",
    "HeartDiseaseorAttack",
    "Stroke",
]

parents = {
    "Age": [],
    "Smoker": [],
    "PhysActivity": [],
    "BMI": ["PhysActivity"],
    "HighChol": ["Smoker", "Age"],
    "HighBP": ["HighChol", "BMI", "Age"],
    "Diabetes_012": ["HighBP", "HighChol", "BMI", "Age"],
    "HeartDiseaseorAttack": ["Diabetes_012", "Age"],
    "Stroke": ["Diabetes_012", "Smoker", "Age"],
}

topo_order = [
    "Age",
    "Smoker",
    "PhysActivity",
    "BMI",
    "HighChol",
    "HighBP",
    "Diabetes_012",
    "HeartDiseaseorAttack",
    "Stroke",
]
```

Risk Factor Insights & Algorithm Comparison

Diabetes Risk Insights

- Obesity is the strongest predictor of diabetes.
- Inactivity amplifies obesity's effect.
- High cholesterol and blood pressure strongly increase risk.
- Smoking indirectly worsens risk through cholesterol and cardiovascular heal
- Age raises the baseline risk across all profiles.

Model Predictions (Exact VE)

- Healthy/active person: ~5% diabetes probability
- Obese + inactive + high cholesterol: ~38% diabetes probability
- Diabetes → 23% heart disease risk
- Diabetes → 10% stroke risk

Algorithm Comparison

- LW at N=100: high variance, unstable estimates
- LW at N=500–1000: large improvement
- LW at N=5000+: closely matches VE (L1 error ≤ 0.01)
- Runtime scales linearly (100 → ~2 sec; 10k → ~170 sec)
- VE remains extremely fast (~0.05 sec per query)