Bayesian Inference for Early Stage Diabetes Risk Prediction

Marta Espejo, Josep Soler & Jordi Nadeu

Contents

1 Introduction

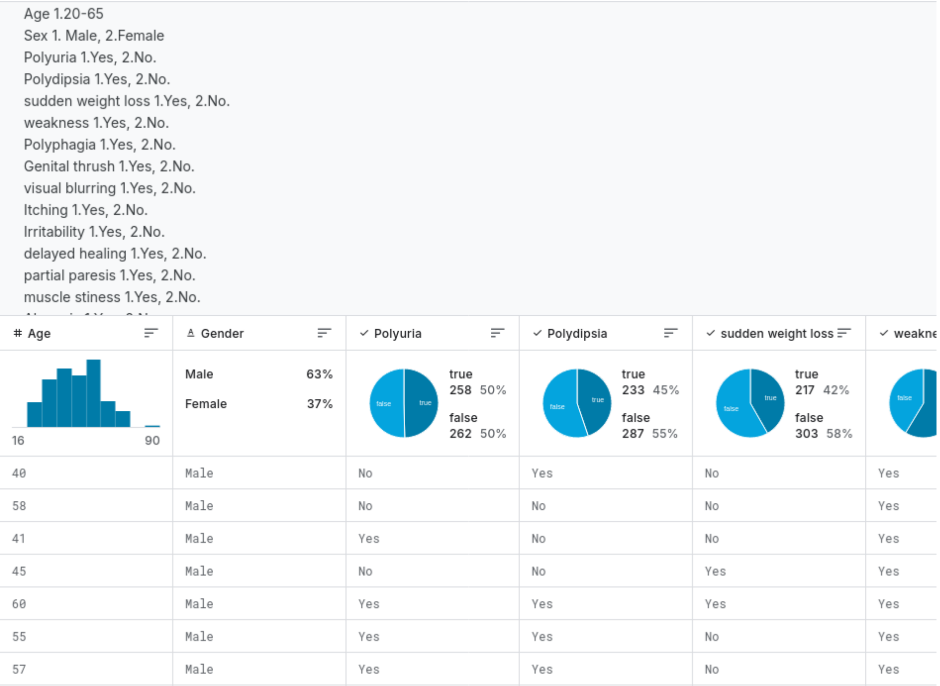
We can say an early diagnosis of diabetes is vital for effective treatment and management. This miniproject explores how Bayesian inference that is a probabilistic reasoning technique, can help model and predict the risk of diabetes based on patient symptoms and demographic information.

We will use the dataset "Early Stage Diabetes Risk Prediction Dataset"(1) from Kaggle platform, we aim to create a Bayesian Network to uncover causal and probabilistic relationships among symptoms and risk factors. Then we do a comparation with a non Bayesian baseline, in our case we choose a Gaussian (continuous) logistic regression baseline.

2 Dataset Description

The dataset, sourced from Kaggle (1), comprises 520 instances with 17 attributes, including symptoms like Age, Gender, polyuria (excessive urination), polydipsia (excessive thirst), sudden weight loss, visual blurring between others, and the outcome variable are the presence (Yes) or absence (No) of diabetes.

The data was collected through direct questionaires from patients at the Sylhet Diabetes Hospital in Bangladesh.



Age 1.20-65
Sex 1. Male, 2.Female
Polyuria 1.Yes, 2.No.
Polydipsia 1.Yes, 2.No.
sudden weight loss 1.Yes, 2.No.
weakness 1.Yes, 2.No.
Polyphagia 1.Yes, 2.No.
Genital thrush 1.Yes, 2.No.
visual blurring 1.Yes, 2.No.
Itching 1.Yes, 2.No.
Irritability 1.Yes, 2.No.
delayed healing 1.Yes, 2.No.
partial paresis 1.Yes, 2.No.
muscle stiness 1.Yes, 2.No.

# Age	Gender	✓ Polyuria	✓ Polydipsia	✓ sudden weight loss	✓ weakne
16 90	Male 63% Female 37%	true 258 50% false 262 50%	true 233 45% false 287 55%	true 217 42% false 303 58%	
40	Male	No	Yes	No	Yes
58	Male	No	No	No	Yes
41	Male	Yes	No	No	Yes
45	Male	No	No	Yes	Yes
60	Male	Yes	Yes	Yes	Yes
55	Male	Yes	Yes	No	Yes
57	Male	Yes	Yes	No	Yes

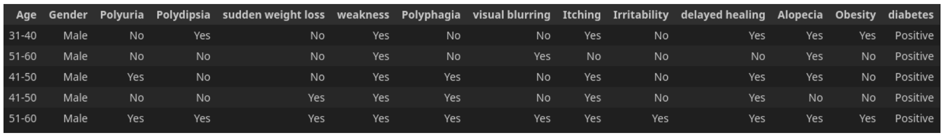
Figure 1: A sample of the data of the dataset

3 Methodology

3.1 Data Preprocessing

We do a preprocessing data process, starting with the rename of the columns like ”class” to ”diabetes” for clarity. Also categorical variables were encoded appropriately and missing values were handled. The Age attribute was discretized into bins to facilitate the construction of the Bayesian Network.

As shown below have an example of the discretization process of age attribute:



Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	visual blurring	Itching	Irritability	delayed healing	Alopecia	Obesity	diabetes
31-40	Male	No	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Positive
51-60	Male	No	No	No	Yes	No	Yes	No	No	No	Yes	No	Positive
41-50	Male	Yes	No	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Positive
41-50	Male	No	No	Yes	Yes	Yes	No	Yes	No	Yes	No	No	Positive
51-60	Male	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

Figure 2: Discretize with strictly increasing numeric bin edges

3.2 Bayesian Network Construction

Using the `pgmpy` (2) library in Python, we defined the structure of the Bayesian Network based on domain knowledge and the more technical knowledge on diabetes of a fifth year medicine student that we asked for help. The network includes nodes representing symptoms and the target variable ’diabetes’, with directed edges indicating conditional dependencies.

Then we do a Directed Acyclic Graph (DAG) manually constructed as shown below.

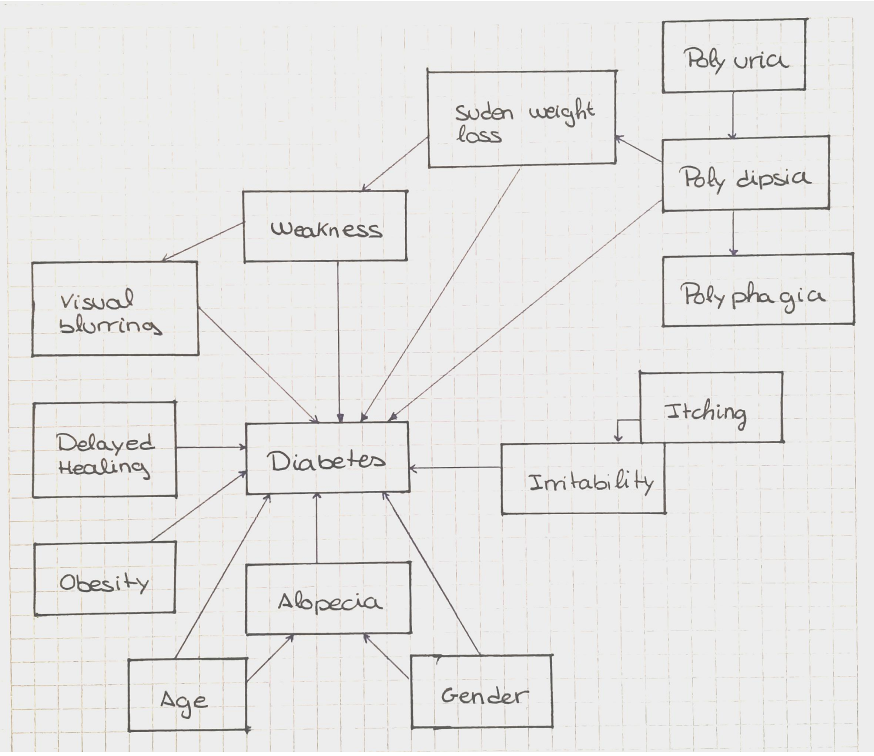
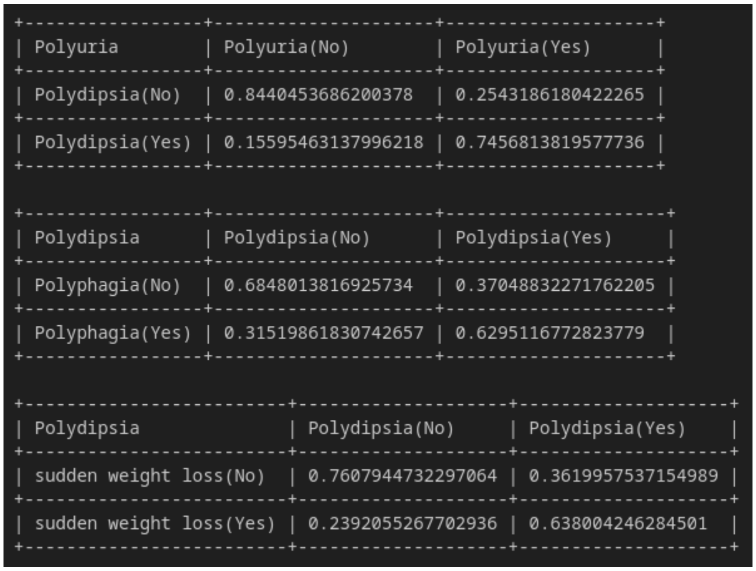


Figure 3: Directed Acyclic Graph of our Bayesian Network

3.3 Parameter Learning

We employed the Bayesian Estimator with a BDeu prior to learn the Conditional Probability Distributions (CPDs) from the data. Also all variables were encoded as strings for compatibility. This approach allows for incorporating prior beliefs and handling data sparsity effectively.

Finally the inference queries do it after will be conducted using this model.



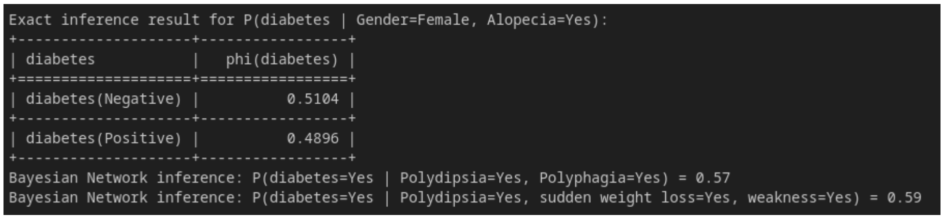
```
+-----------------+---------------------+--------------------+
| Polyuria        | Polyuria(No)        | Polyuria(Yes)      |
+-----------------+---------------------+--------------------+
| Polydipsia(No)  | 0.8440453686200378  | 0.2543186180422265 |
+-----------------+---------------------+--------------------+
| Polydipsia(Yes) | 0.15595463137996218 | 0.7456813819577736 |
+-----------------+---------------------+--------------------+

+-----------------+---------------------+---------------------+
| Polydipsia      | Polydipsia(No)      | Polydipsia(Yes)     |
+-----------------+---------------------+---------------------+
| Polyphagia(No)  | 0.6848013816925734  | 0.37048832271762205 |
+-----------------+---------------------+---------------------+
| Polyphagia(Yes) | 0.31519861830742657 | 0.6295116772823779  |
+-----------------+---------------------+---------------------+

+-------------------------+--------------------+--------------------+
| Polydipsia              | Polydipsia(No)     | Polydipsia(Yes)    |
+-------------------------+--------------------+--------------------+
| sudden weight loss(No)  | 0.7607944732297064 | 0.3619957537154989 |
+-------------------------+--------------------+--------------------+
| sudden weight loss(Yes) | 0.2392055267702936 | 0.638004246284501  |
+-------------------------+--------------------+--------------------+
```

Figure 4: A sample of the CPD's output

3.4 Inference

Variable elimination was used for exact inference, enabling the computation of posterior probabilities for the presence of diabetes given observed symptoms. (3)

We use three queries about how evidence changes the predicted probability of having diabetes as shown below:



```
Exact inference result for P(diabetes | Gender=Female, Alopecia=Yes):
+--------------------+-----------------+
| diabetes           |   phi(diabetes) |
+====================+=================+
| diabetes(Negative) |          0.5104 |
+--------------------+-----------------+
| diabetes(Positive) |          0.4896 |
+--------------------+-----------------+
Bayesian Network inference: P(diabetes=Yes | Polydipsia=Yes, Polyphagia=Yes) = 0.57
Bayesian Network inference: P(diabetes=Yes | Polydipsia=Yes, sudden weight loss=Yes, weakness=Yes) = 0.59
```

Figure 5: Output of the Inference of Bayasen Network

4 Conclusion

This miniproject demonstrates that Bayesian Networks are a powerful tool for healthcare applications, enabling interpretable and data-informed diagnosis support. Probabilistic reasoning can assist clinicians in estimating disease risk when complete patient data is unavailable, ultimately contributing to more personalized care.

The discrete BN provides transparent, calibrated early stage risk predictions but at the cost of discretization and structural assumptions. The Gaussian logistic-regression offers simplicity and continuous modeling but lacks full uncertainty quantification. An hybrid approach with hierarchical Bayesian models with mixed discrete and continuous CPD's, could be interesnting and combine strengths of both.

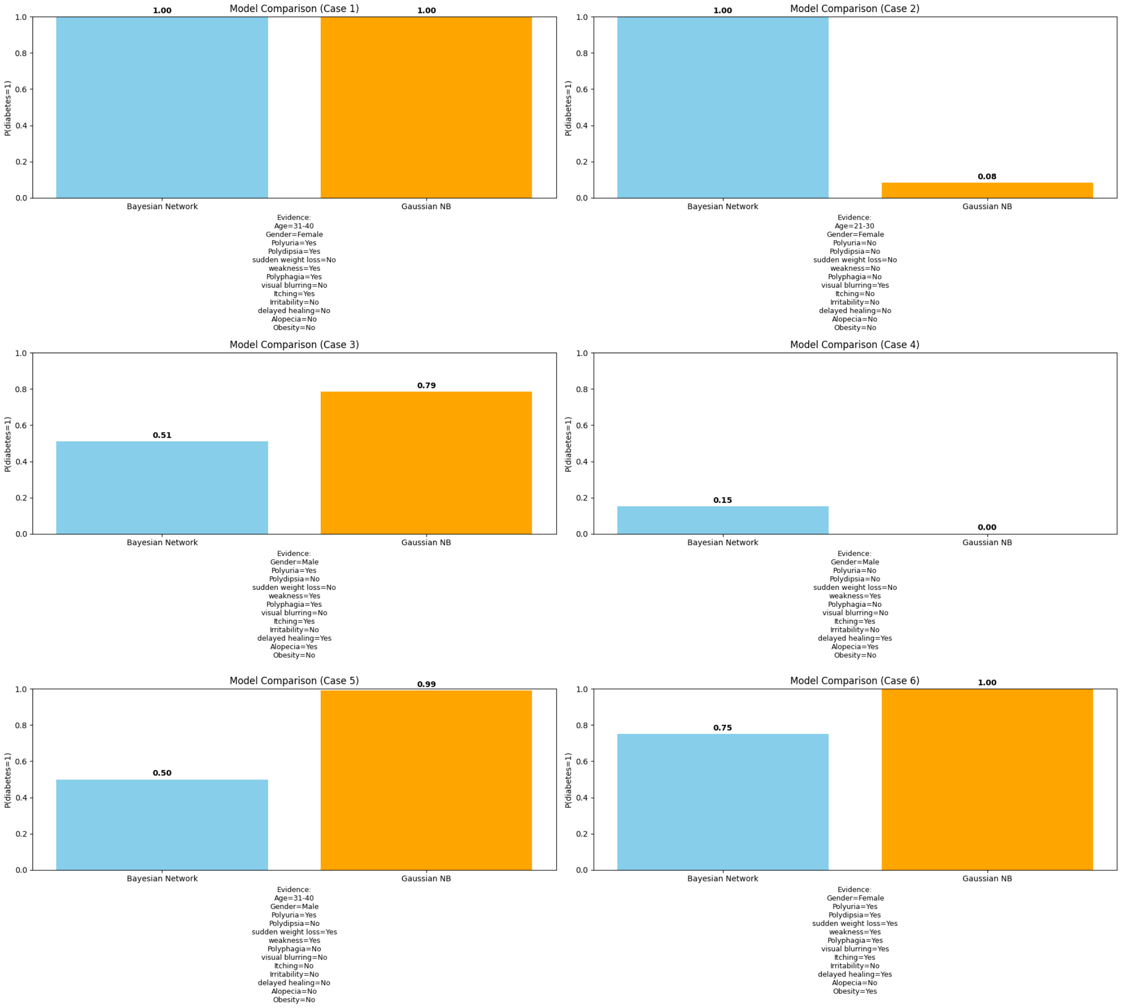


Figure 6: Results comparations between Bayesian Network & Gaussian NB

References

[1] Y. Hessein, "Early-Stage Diabetes Risk Prediction Dataset," Kaggle, 2020. https://www.kaggle.com/datasets/yasserhessein/early-stage-diabetes-risk-prediction-dataset

[2] A. Sharma et al., "pgmpy: Probabilistic Graphical Models using Python," *Journal of Machine Learning Research*, 2018.

[3] Bayesian inference. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Bayesian_inference