Diabetes Prediction Using Probabilistic Graphical Models

Rebecca Hadi
rhadi1@jh.edu
Probabilistic Graphical Models
625.692
Johns Hopkins Engineering for Professionals

Abstract

This paper implements different types of probabilistic graphical models on the Pima Indians Diabetes data set. The paper implements Naive Bayes, Bayesian Networks (Directed Models), Markov Networks (Undirected Networks), and uses Markov Chain Monte Carlo to estimate parameters of a Logistic Regression Model. The main goal of this paper is prediction accuracy, while inference methods are also explored. The primary model evaluation metric is ROC-AUC, and F1 Score is a secondary metric. Inference is explored by looking at the posterior likelihood of diabetes given varying sets of evidence. The best performing model was the Naive Bayes with ROC-AUC of 0.83 and 0.78 on the training and test sets, respectively. Ethical considerations are explored for any clinical implementation of models explored in this paper.

Keywords

Bayesian Networks, Markov Networks, Probabilistic Graphical Models, Markov Chain Monte Carlo, Diabetes, Medical Diagnosis

Introduction

According to the CDC, 34.1 million adults aged 18 years or older (13.0 percent of all US adults) have diabetes as of 2020[1]. The prevalence of diabetes increases with age, reaching a

[1] National diabetes report - centers for disease control and prevention. (n.d.). Retrieved April 15, 2022, from https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

prevalence of 26.8 percent in adults 65 years or older[2]. "In 2017, diabetes was the seventh leading cause of death in the United States[3]". Early detection of diabetes can be used to inform lifestyle changes for those at risk of developing Type 2 Diabetes. Risk factors for diabetes include smoking, being overweight/obese, physical inactivity, elevated A1C, high blood pressure, and high cholesterol[4]. Diabetes can also be a cause of significant healthcare spending, evidenced by 16 million emergency department visits (ED) in 2016[5], and 7.8 million hospital discharges related to diabetes in 2016[6]. Therefore, being able to predict diabetes or understand risk factors can help identify the need to implement prophylactic treatment. In this paper, we will implement different Bayesian Network structures as well as a Naive Bayes classifier to predict diabetes. We will evaluate the models using ROC-AUC as the primary evaluation metric, with F1 score (harmonic mean of precision and recall) as the secondary evaluation metric.

Literature Review and Context of Work

There have been several contributions to medical diagnosis using Bayesian networks. In "Comparison of Bayesian Networks for Diabetes Prediction"[7], the authors examine four different methods of creating Bayesian networks. There were hierarchical and non-hierarchical structures of networks compared. The accuracy metrics were ROC-AUC (77 percent) and F1 Score (24 percent), and were evaluated on training and test sets. This work heavily inspired my project and my goal is to produce similar or greater performance. My data set expands on this work by using a different data set, where this paper uses data obtained "from the Thai National Health Examination Survey"[8].

In "A Bayesian Network for Modelling Blood Glucose Concentration in Type 1 Diabetes"[9], the focus was on a temporal implementation of Bayesian networks to understand the "effects of everyday physical activity on blood glucose concentration in people with Type 1 diabetes"[10]. Markov Chain Monte Carlo methods were used to estimate model parameters. The time-series nature of this paper is different from my project (which is person-level with one record per person), it applies a Bayesian framework to a medical problem.

[2] Ibid
[3] Ibid
[4] Ibid
[5] Ibid
[6] Ibid
[7] Leerojanaprapa K., Sirikasemsuk K. (2019) Comparison of Bayesian Networks for Diabetes Prediction. In: Bhatia S., Tiwari S., Mishra K., Trivedi M. (eds) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, vol 924. Springer, Singapore. https://doi.org/10.1007/978-981-13-6861-5_37
[8] Ibid
[9] Ewings SM, Sahu SK, Valletta JJ, Byrne CD, Chipperfield AJ. A Bayesian network for modelling blood glucose concentration and exercise in type 1 diabetes. Stat Methods Med Res. 2015 Jun;24(3):342-72. doi: 10.1177/0962280214520732. Epub 2014 Feb 2. PMID: 24492795.
[10] Ibid

In "Prediction of diabetes using Bayesian Network[11], they use the WEKA[12] tool to predict diabetes. They report a 99 percent accuracy, which in this author's view, is a suspiciously high accuracy. They have three classes for diabetes (no, pre, and yes) as the outcome variable, whereas in my project I have a binary outcome for the presence of diabetes. The reported accuracy is 99.5 percent. In the "Preparing Dataset" section, I did not see any discussion of a training and test split for the model evaluation. This makes me think the model would not generalize well, as it may be severely over-fitting if it was not tested on data the model had not seen before. This underscores the importance of splitting into test/train sets in my project.

In the paper "Deep learning approach for diabetes prediction using Pima Indian dataset"[13], the authors compare four types of classifiers: "Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT), and Deep Learning (DL) with accuracy in the range of 90-98 percent". This paper used the same data set that I am using in my project, and can be a benchmark for accuracy on the same classification task. My project expands on this work by using Bayesian and Markov Networks, as well as taking an inference lens, in addition to prediction accuracy. I did not see discussion of a training/validation set in the paper design, so I am somewhat skeptical of the accuracy, but it's possible I did not have full insight into their method and evaluation. This paper discusses the potential application of the model to a "novel automatic prognosis tool", but it is my hope that they would collect more data that included males, as the Pima Indians data set includes only females.

Seven other papers were reviewed as part of the literature review of this project, and for brevity will not be described here, but are listed in the bibliography section. The overall theme is that there are several research papers showing the application of Bayesian Networks to the task of predicting diabetes, but adoption of these methods in practice is not widespread. There is also opportunity to partner with experts to incorporate domain expertise in addition to machine learning based features and model parameters. The context of my project is an application of Naive Bayes, Bayesian Networks, and Markov Networks to this task. My project's contribution is not necessarily novel, but it does provide a comparison of Bayesian and network based methods, as well as inferring posterior distributions giving evidence. My paper's main contribution is the ethical discussion of using Bayesian networks in an automated diagnosis setting, which is described at the end of this paper.

[11] Prediction of diabetes using Bayesian network - IJCSIT. (n.d.). Retrieved March 4, 2022, from https://www.ijcsit.com/docs/Volume205/vol5issue04/ijcsit2014050477.pdf
[12] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
[13] Naz H, Ahuja S. Deep learning approach for diabetes prediction using Pima Indian dataset. J Diabetes Metab Disord. 2020;19(1):391-403. Published 2020 Apr 14. doi:10.1007/s40200-020-00520-5

Explanation of Data

The data set being used in this work is the Pima Indians data set from the National Institute of Diabetes and Digestive and Kidney Diseases, downloaded from Kaggle[14]. This data set is limited to females who are at least 21 years old of Pima Indian heritage.   There are 768 records which are further split into testing and training data sets.

The variable of interest is 'Outcome', which takes on value 1 for having Diabetes, and 0 for not having Diabetes.

The features included are: Age (Years), Number of Pregnancies, Skin Thickness (Triceps skin fold thickness mm) , BMI (Body Mass Index weight in kg/height in m^2), Glucose (Plasma glucose concentration 2 hours in an oral glucose tolerance test), Insulin (2-Hour serum insulin muU/ml) , Diabetes Pedigree Function (i.e. history of diabetes), Blood Pressure (Diastolic Blood Pressure mm Hg)[15].

Hypothesis

Null Hypothesis
μ_0: The ROC-AUC of the Bayesian Network will not be greater than 0.7

Alternative Hypothesis
μ_1: The ROC-AUC of the Bayesian Network will be greater than or equal to 0.7.

Experimental Design

To avoid over-fitting, the data were split into training and test data sets (70 percent and 30 percent, respectively). The models were trained on the training data set and evaluated on the test data set. This allows us to understand how the model would generalize in a practical application. Then, the ROC-AUC and F1 scores are calculated for each model, training, and test set. The results will then be compared to select a 'best model' optimizing for ROC-AUC.

[14] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press. https://www.kaggle.com/uciml/pima-indians-diabetes-database

[15] Ibid

Software

This project was implemented in Python, specifically using the libraries pandas[16], numpy[17], sklearn[18], imblearn[19], pgmpy[20], pymc3[21], and matplotlib[22].

Exploratory Data Analysis and Data Pre-Processing

Missing Data

The data set did not have any null records, but did have records with a value of "0" that did not make logical sense. For example, there were 227 records with a zero value for "SkinThickness", which is physically impossible. It is possible these values were caused by human-error (typed incorrectly), broken instruments, or the measurement wasn't able to be taken at the point the result of the data were collected. After inspecting each of the columns, we see that the number of zero records is not evenly distributed across columns, so we cannot assume the data are MCAR (missing completely at random), and therefore cannot ignore the data. Instead, we can assume that each of the columns were sampled from a Gaussian with mean and standard deviation based on the non-zero records, and draw from the respective distribution to impute the values. We see below the distributions for Blood Pressure and Skin Thickness.

[16] https://pandas.pydata.org

[17] https://numpy.org

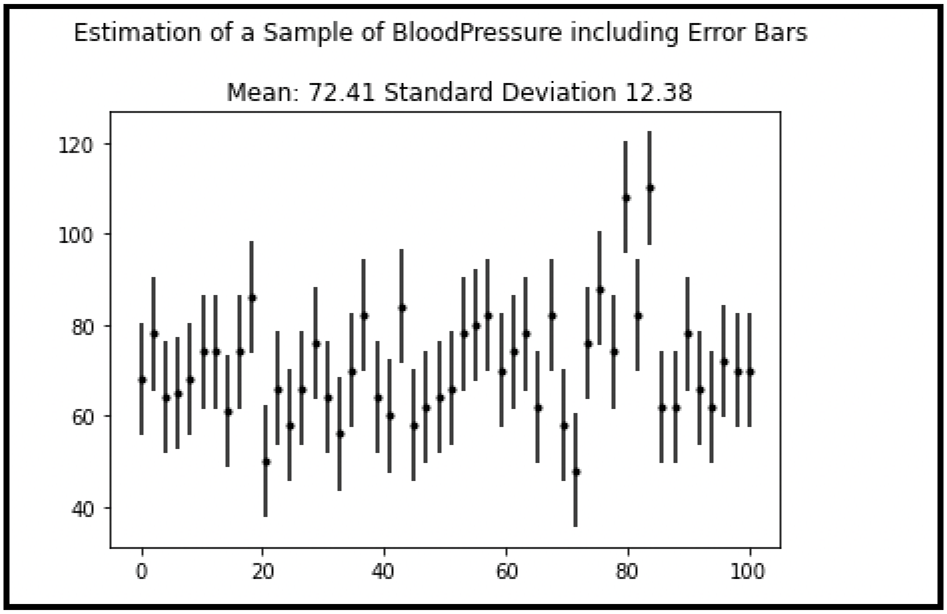
[18] https://scikit-learn.org/stable/about.html

[19] Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Guillaume Lematre and Fernando Nogueira and Christos K. Aridas, Journal of Machine Learning Researchhttp://jmlr.org/papers/v18/16-365.htm

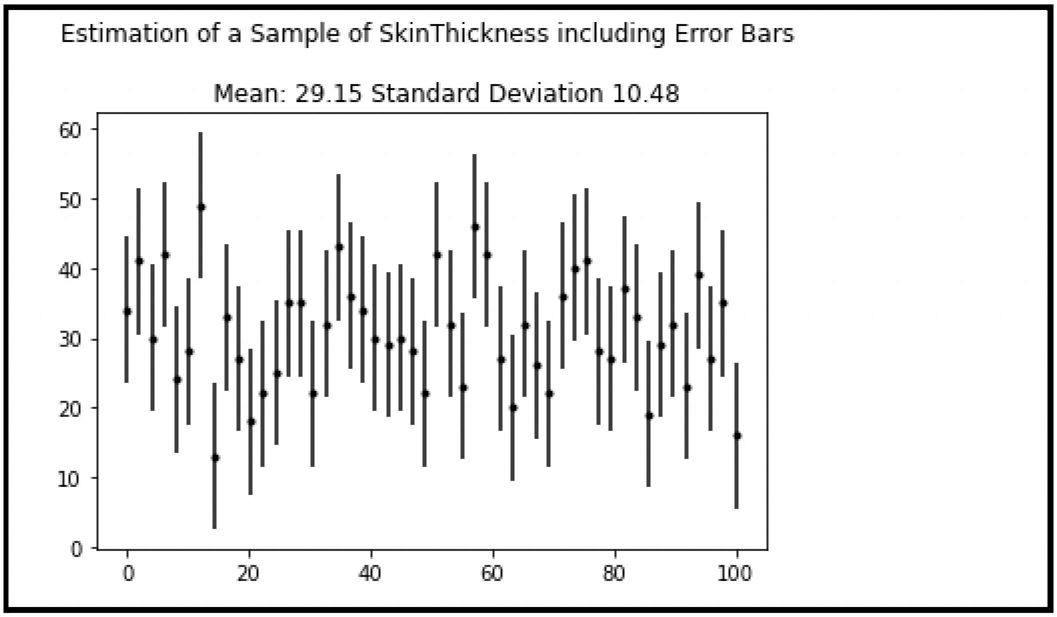
[20] https://pgmpy.org

[21] https://docs.pymc.io/en/v3/

[22] https://matplotlib.org

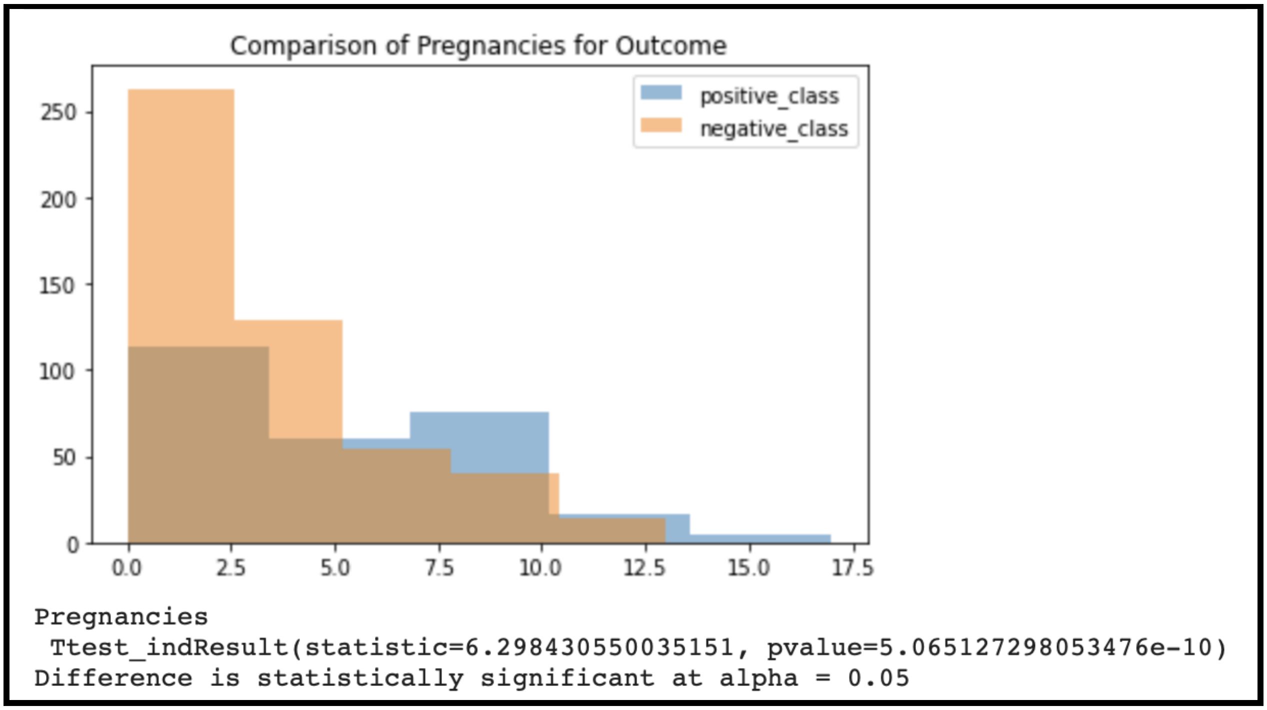


Estimating the Gaussian Distribution of BloodPressure Including Errors Bars



Estimating the Gaussian Distribution of SkinThickness Including Errors Bars

Before applying models, exploratory data analysis was conducted to get an understanding of feature importance before using t-tests. For each continuous variable, two arrays were generated for the positive class (having diabetes) and negative class (not having diabetes).



```
Pregnancies
 Ttest_indResult(statistic=6.298430550035151, pvalue=5.065127298053476e-10)
Difference is statistically significant at alpha = 0.05
```

Imbalanced Data

The proportion of diabetes in the training and test data sets was 35.2 percent and 34.6 percent. Ideally, in binary classification tasks the classes will be balanced at 50 percent each. The function `RandomOverSampler` from the Python library `imblearn`[23], was applied to the training data set to balance the classes. It was not applied to the test set because we want to test on data that resembles reality.

Feature Discretization

The continuous features were discretized using the qcut() function in Python to put into two or three discrete bins, depending on the uniqueness of the edges of the bins.

[23] Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Guillaume Lematre and Fernando Nogueira and Christos K. Aridas, Journal of Machine Learning Researchhttp://jmlr.org/papers/v18/16-365.htm

Naive Bayes

Naive Bayes is "a condition probability model"[24]. As stated in Koller & Friedman , "*Naive Bayes assumption* is that the features are conditionally independent given the instance's class"[25]. More formally,

$$(Xi \perp X_{-i} \mid C) \text{ for all } i \text{ where } X_{-i} = \{X1, \ldots, Xn\} - \{Xi\} \quad \text{(Koller 3.6)}$$

It can then be shown as a result of the independence assumption, the Naive Bayes model factorizes as:

$$P(C, X1, \ldots, Xn) = P(C) \prod_{i=1}^{n} P(Xi|C) \quad \text{(Koller 3.7)}$$

Bayesian Network Structure

When the independence assumption of the Naive Bayes model is relaxed, we can factorize networks in a similar way using the *chain rule for Bayesian networks*[26].

$$P(A, B, C, D) = P(A)P(B|A)P(C|B, A)P(D|C) \text{ (Koller 3.9)}$$

The Bayesian network structure we will use to model diabetes is defined as in Definition 3.1 from the Koller Friedman text.

Definition 3.1[27]: ***A*** Bayesian network structure G ***is a directed acyclic graph whose nodes represent random variables*** X1, . . . , Xn***. Let*** PaGXi ***denote the parents of*** Xi ***in*** G***, and*** NonDescendantsXi ***denote the variables in the graph that are not descendants of*** Xi***. Then*** G ***encodes the following set of conditional independence assumptions, called the*** local independencies***, and denoted by*** Il(G)***:***

For each variable Xi***:*** (Xi ⊥ NonDescendantsXi | PaGXi)

We can express this structure G as a distribution P over the same space as[28]:

$$P(X1, \ldots, Xn) = \prod_{i=1}^{n} P(Xi|PaGXi) \quad \text{(Koller 3.17)}$$

[24] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

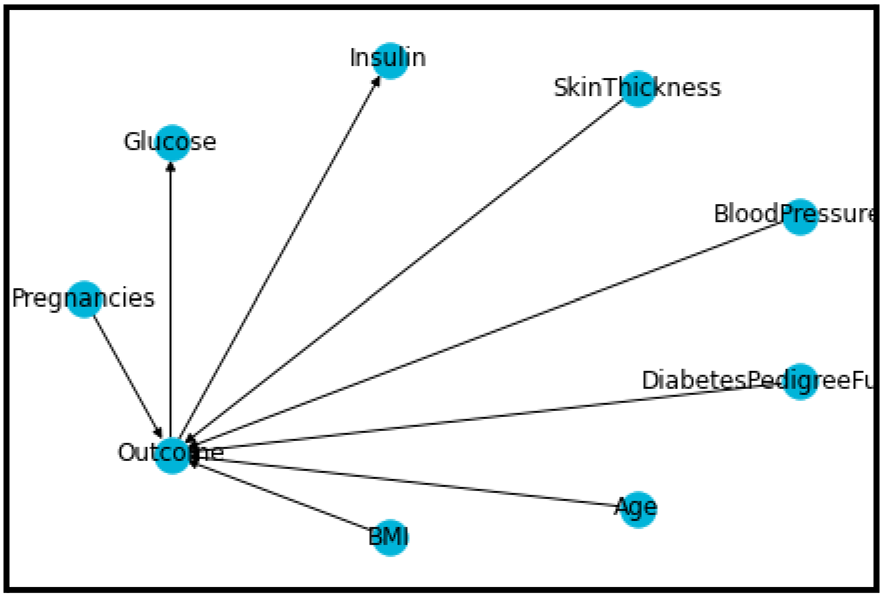
[25] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p50

[26] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p54

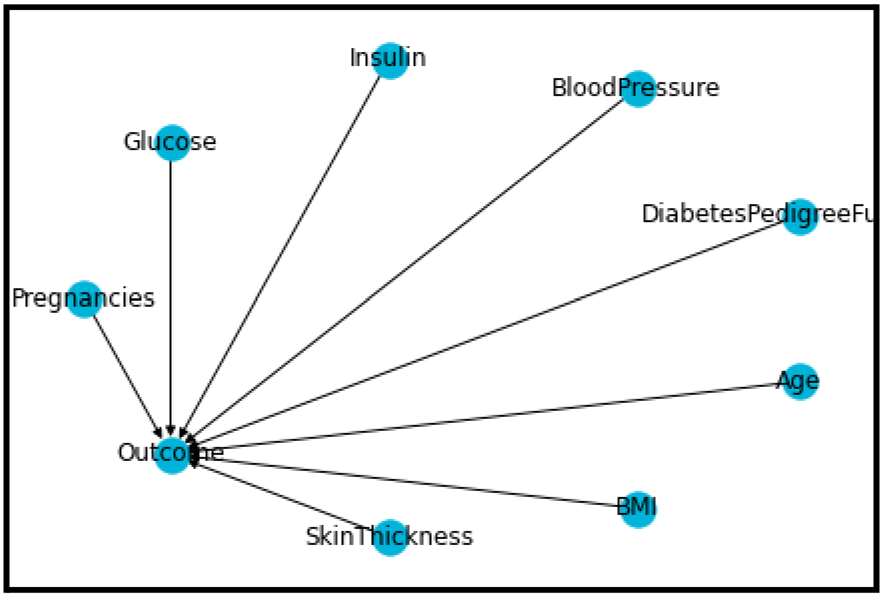
[27] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p57

[28] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p62

Four Bayesian network structures were implemented with varying levels of hierarchy. Two network images are reproduced below, with the other two available in the accompanying Jupyter Notebook PDF.



Bayesian Network with Outcome as the Parents of Insulin and Glucose



Bayesian Network with all Nodes as the Parent of Outcome

Expressing the first network using the chain rules for Bayesian networks, we can write:

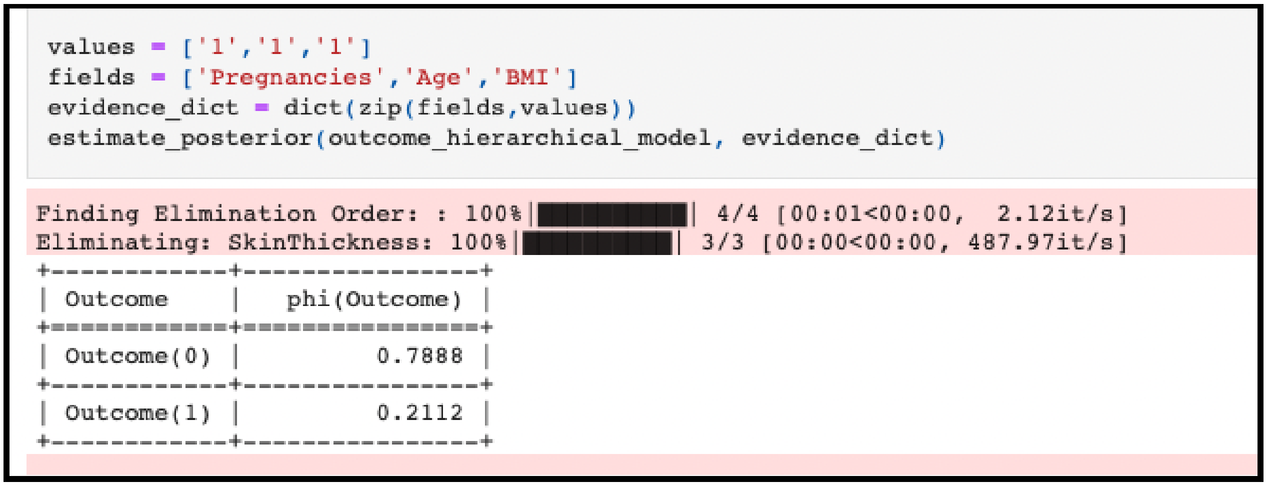
$$P(P, BP, DPF, B, ST, A, O, G, I) \; = \\ P(P)P(S)P(BP)P(DBF)P(A)P(B)P(ST)P(O|P, BP, DPF, B, ST, A)P(G|O)P(I|O)$$

Expressing the second network using the chain rules for Bayesian networks, we can write:

$$P(P, BP, DPF, B, ST, A, O, G, I) \ = \ P(P)P(S)P(BP)P(DBF)P(A)P(B)P(ST)P(G)P(I)P(O|P, BP, DPF, B, ST, A, G, I)$$

Bayesian Inference using Variable Elimination

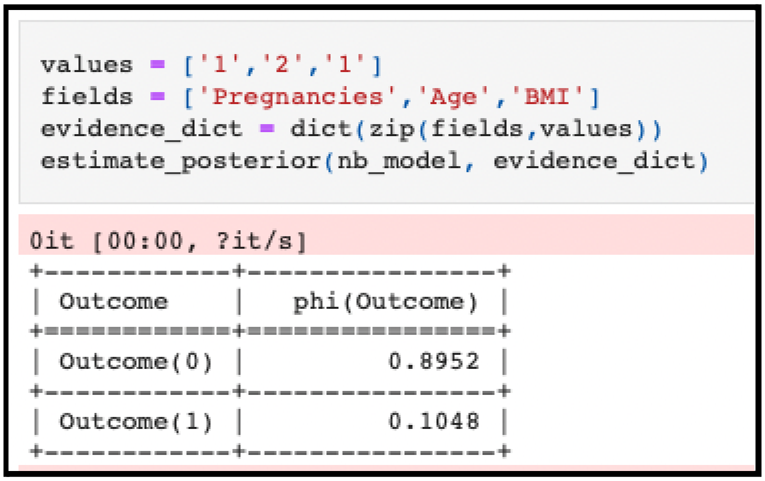
To conduct inference in the Bayesian models, we choose the best performing Bayesian model along with Naive Bayes (which was the best performing model overall). Through the process of using Variable Elimination, we can estimate the posterior distribution of Outcome given evidence.



```
values = ['1','1','1']
fields = ['Pregnancies','Age','BMI']
evidence_dict = dict(zip(fields,values))
estimate_posterior(outcome_hierarchical_model, evidence_dict)
```

```
Finding Elimination Order: : 100%|██████████| 4/4 [00:01<00:00,  2.12it/s]
Eliminating: SkinThickness: 100%|██████████| 3/3 [00:00<00:00, 487.97it/s]
+------------+----------------+
| Outcome    |   phi(Outcome) |
+============+================+
| Outcome(0) |         0.7888 |
+------------+----------------+
| Outcome(1) |         0.2112 |
+------------+----------------+
```

The first inference using the outcome hierarchical model and with evidence array [1,1,1] indicates that this person would be in the low range for pregnancies, age, and BMI. The posterior probability of diabetes is 78.9%, which tells us that it is not very likely this person has diabetes given the evidence.



```
values = ['1','2','1']
fields = ['Pregnancies','Age','BMI']
evidence_dict = dict(zip(fields,values))
estimate_posterior(nb_model, evidence_dict)
```

```
0it [00:00, ?it/s]
+------------+----------------+
| Outcome    |   phi(Outcome) |
+============+================+
| Outcome(0) |         0.8952 |
+------------+----------------+
| Outcome(1) |         0.1048 |
+------------+----------------+
```

The second inference using the Naive Bayes model with evidence array [1,2,1] represents low values for pregnancy and BMI, and medium value for age. The posterior probability of diabetes is 89.5%, which tells us it is not very likely this person has diabetes given the evidence.

Markov Network Structure

In a Markov network, we do not have directed edges like we do in a Bayesian network. The first picture below shows the structure of the Markov network, and it does not have arrows like in the figure of the Bayesian network because there are no directed edges. Mathematically, we say that[29]:

Definition 4.3: *A* distribution $P\Phi$ is a Gibbs distribution parameterized by a set of factors $\Phi = \{\varphi1(D1), . . . , \varphi K\ (DK\ )\}$ if it is defined as follows:

$$P\Phi(X1,...,Xn) = Z1\ \tilde{P}\Phi(X1,...,Xn),$$

where

$\tilde{P}\Phi\ (X\ 1\ , . . . , X\ n\ ) = \varphi1\ (\ D1\ ) \times \varphi2\ (\ D2\ ) \times \cdots \times \varphi m\ (Dm)$ *is an unnormalized measure and*

$Z= \sum_{X1,..,Xn} \tilde{\tilde{P}}\Phi(X1,...,Xn)\ X1\ ,...,Xn$ *is a normalizing constant called the* partition function*.*

We can then define a Markov Network as[30]:

Definition 4.4: We say that a distribution $P\Phi$ with $\Phi = \{\varphi1(D1), . . . , \varphi K\ (DK\ )\}$ factorizes over a Markov network H if each Dk (k = 1,...,K) is a complete subgraph of H.

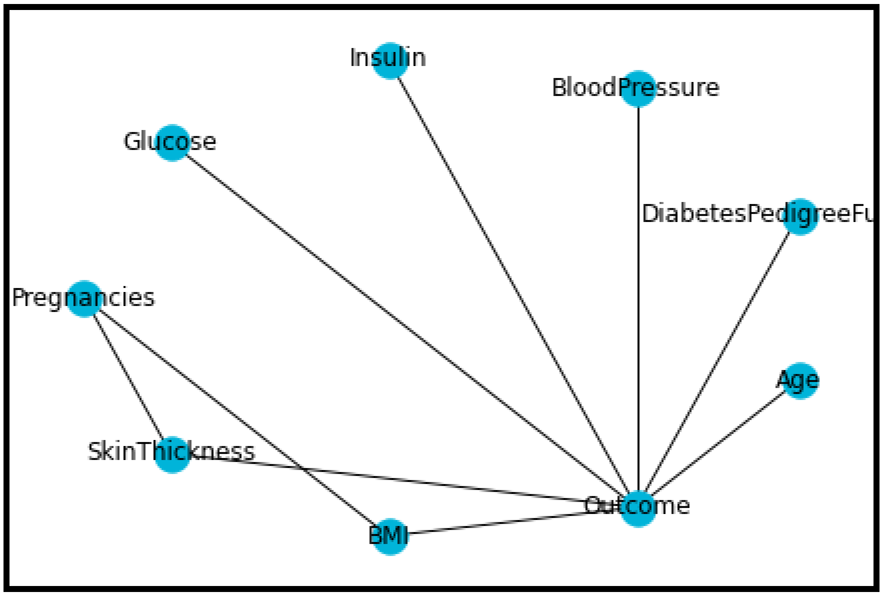
For inference using a Markov Network, we construct and visualize a Junction Tree which creates cliques based on the nodes in the original Markov Network. The Junction Tree is defined formally as[31]:

Definition 10.3 Let Φ be a set of factors over X . A cluster tree over Φ that satisfies the running intersection property is called a clique tree (sometimes also called a junction tree or a join tree). In the case of a clique tree, the clusters are also called cliques.

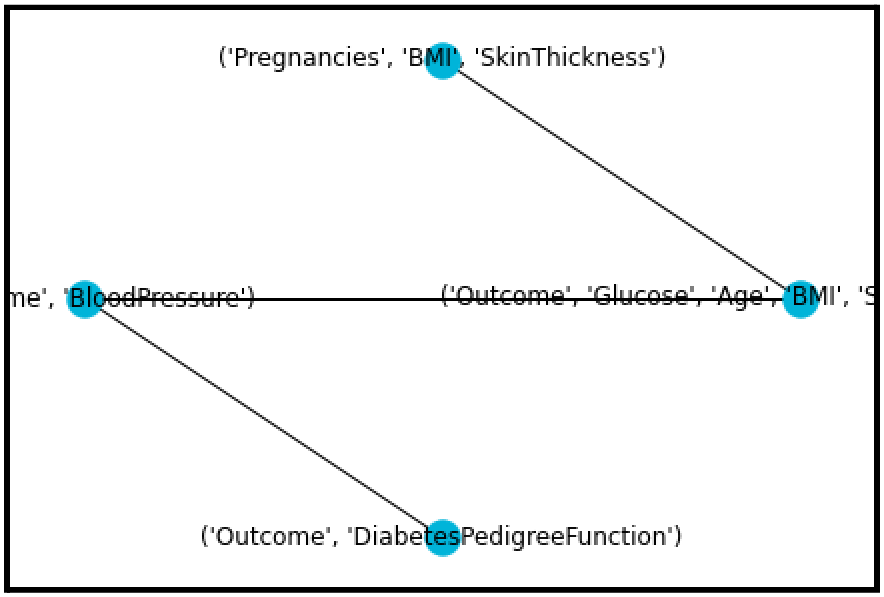
[29] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p108

[30] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. p109

[31] Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009. P 348



Markov Network with undirected edges

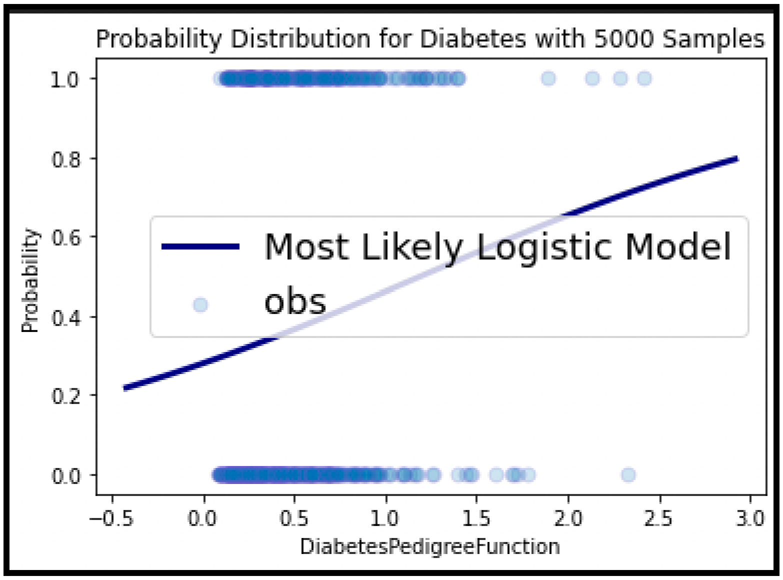

('Pregnancies', 'BMI', 'SkinThickness')
me', 'BloodPressure')
('Outcome', 'Glucose', 'Age', 'BMI', 'S
('Outcome', 'DiabetesPedigreeFunction')

Junction Tree representing the cliques of the Markov Network

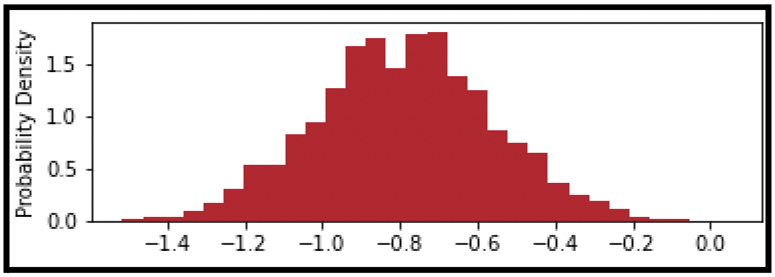
The visualization of the Junction Tree allows us to see the cliques of the Markov Network as (Pregnancy, BMI SkinThickness), (Outcome, Glucose, Age, SkinThickness), (Outcome, Blood Pressure), and (Outcome, DiabetesPedigreeFunction).

Markov Chain Monte Carlo

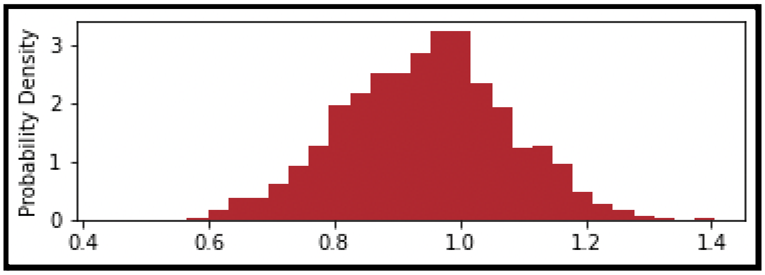
Markov Chain Monte Carlo (MCMC) is a method to estimate parameters. A logistic regression model was trained on the continuous data (before discretization) to generate prior beliefs for the parameters. The feature 'DiabetesPedigreeFunction' had the highest coefficient in the regression model output and thus was chosen (along with the intercept) for MCMC parameter estimation.



Logistic Regression model created from DiabetesPedigreeFunction and average value of alpha and beta from the MCMC simulation.



Alpha parameter estimates from MCMC



Beta Parameter Estimates from MCMC

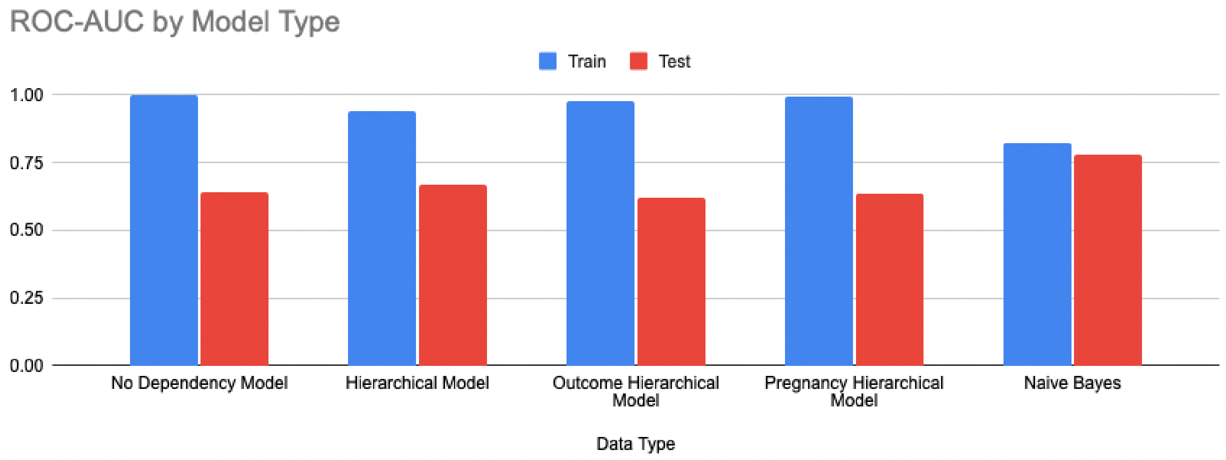
Results and Best Model

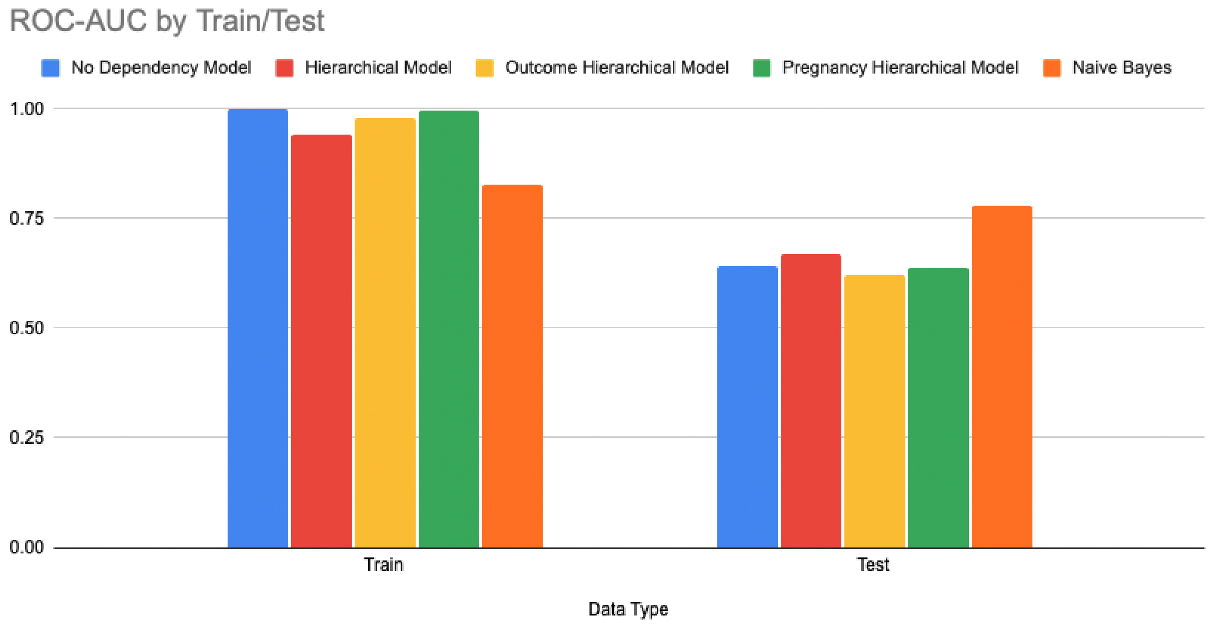
ROC-AUC

Data Type	No Dependency Model	Hierarchical Model	Outcome Hierarchical Model	Pregnancy Hierarchical Model	Naive Bayes
Train	1.00	0.94	0.98	1.00	0.83
Test	0.64	0.67	0.62	0.64	0.78
Train (Over-sampled)	1.00	0.95	0.98	0.99	0.80
Test (Over-sampled)	0.65	0.69	0.64	0.65	0.83

F1 Score

Data Type	No Dependency Model	Hierarchical Model	Outcome Hierarchical Model	Pregnancy Hierarchical Model	Naive Bayes
Train	0.97	0.80	0.86	0.96	0.76
Test	0.31	0.55	0.52	0.26	0.58
Train (Over-sampled)	0.99	0.88	0.92	0.92	0.62
Test (Over-sampled)	0.28	0.59	0.56	0.30	0.67





Discussion

Based on the results above, we see that the Bayesian networks perform very well on the training data sets, but have significantly lower ROC and F1 scores on the test data sets. Ideally, we want to see similar evaluation metrics on the training and test data sets. This tells us that the models are overfit and are "memorizing" the training data and therefore are not generalizing well. The Naive Bayes model has the lowest ROC score on the training data set and has a consistent ROC score on the test data set. Therefore, using ROC-AUC on the test data set as our primary evaluation metric, Naive Bayes is the best model. Of the Bayesian Networks, the 'Hierarchical Model' is the best performing model on the ROC-AUC score for the test data sets.

Conclusion

For the Bayesian networks, we fail to reject the null hypothesis that any model will have an ROC-AUC score above 0.7. The highest ROC-AUC on the test data set for the Bayesian Network models was 0.67, which is lower than 0.7 and not statistically significant on 231 observations at α=0.05 level of significance[32].

For the Naive Bayes network, we reject the null hypothesis that the model will not have an ROC-AUC score above 0.7. The ROC-AUC value for the Naive Bayes on the test set was 0.78, which is greater than 0.7 and statistically significant on 231 observations at the α=0.05 level of significance[33].

Therefore, the Naive Bayes model is the overall winner when it comes to the objective of maximizing the ROC-AUC score on the test data sets. The performance of the models generally improved when trained on the oversampled training sets.

We learned that the Bayesian models are prone to overfitting and that, in this case, the Naive Bayes generalized the best.

From an inference perspective, we learned that we can use the network structure to estimate the posterior probabilities for both the Bayesian and Naive Bayes models. We also constructed a Markov Network and tested converting it to a Bayesian Network, and understanding the cliques involved in the Markov Network.

Ethical Considerations

There may be a desire to use this model or a similar model in a clinical setting to help with clinical practice. That would not be ethical with this model due to the representation bias in the data itself. As discussed, the data set only contains females of Pima Indian heritage that are over 21 years old, and the model is trained on this subset of the entire true population. The model would not be applicable to males, or to any female that is under 21 or not of Pima Indian heritage. Even if this limitation were to be addressed, I believe it is not ethical to have an only automated diagnosis system, to have the algorithm replace the physician. In my view, the best application would be having the algorithm produce an initial result that was then

[32] https://www.statology.org/one-proportion-z-test-calculator/

[33] Ibid

reviewed by a physician, saving them time and allowing them to have an additional data point when making a diagnosis (e.g. "Human in the loop"[34]).

Future Work

Future work can improve on this project by collecting a sample that is more representative of the entire United States population (for a model applied in the US). Future work can also apply cross-validation in the model training process to understand a distribution of ROC-AUC scores across different validation sets. The Bayesian Network models were overfit, evidenced by the high ROC-AUC on the training set, and significantly lower ROC-AUC on the test set. Future work can improve on these methods by implementing methods to reduce overfitting on the model.

For the MCMC, future work can expand on this project by estimating parameters for the other features in the logistic regression model. In this project, only the feature DiabetesPedigreeFunction and intercept were used in the MCMC parameter estimation. Future work could also include implementation of a Hidden Markov Model (HMM) to include latent variables such as “Overweight” (for which BMI/SkinThickness are observed related observed variables).

Future work can also apply to the data collection and modeling approach. The current model is a fixed time and does not have temporal evidence. By collecting data on changes to diet (e.g. servings of vegetables per day, yes/no on plant based diet) and medications, these can be included as nodes in the model, and a state-space model framework could be implemented. This could look like implementing a time series and evaluating the effect of interventions (nutrition/medication) on the prevalence of diabetes (as a binary variable, or on the continuous variables of glucose/insulin). Hidden Markov Models could also be employed if the treatment variables were not observable, but could be observed through reduction in other nodes (e.g. latent variable: changing diet, observed variable: decrease in BMI).

Bibliography & References

Koller, D. and Friedman, N.. *Probabilistic Graphical Models: Principles and Techniques*. : MIT Press, 2009.

[34] Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., &amp; He, L. (2022, April 26). A survey of human-in-the-loop for machine learning. arXiv.org. Retrieved May 1, 2022, from https://arxiv.org/abs/2108.00941

National diabetes report - centers for disease control and prevention. (n.d.). Retrieved April 15, 2022, from https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

Leerojanaprapa K., Sirikasemsuk K. (2019) Comparison of Bayesian Networks for Diabetes Prediction. In: Bhatia S., Tiwari S., Mishra K., Trivedi M. (eds) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, vol 924. Springer, Singapore. https://doi.org/10.1007/978-981-13-6861-5_37

Ewings SM, Sahu SK, Valletta JJ, Byrne CD, Chipperfield AJ. A Bayesian network for modelling blood glucose concentration and exercise in type 1 diabetes. Stat Methods Med Res. 2015 Jun;24(3):342-72. doi: 10.1177/0962280214520732. Epub 2014 Feb 2. PMID: 24492795.

Prediction of diabetes using Bayesian network - IJCSIT. (n.d.). Retrieved March 4, 2022, from https://www.ijcsit.com/docs/Volume205/vol5issue04/ijcsit2014050477.pdf

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Naz H, Ahuja S. Deep learning approach for diabetes prediction using Pima Indian dataset. J Diabetes Metab Disord. 2020;19(1):391-403. Published 2020 Apr 14. doi:10.1007/s40200-020-00520-5

Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Guillaume Lematre and Fernando Nogueira and Christos K. Aridas, Journal of Machine Learning Researchhttp://jmlr.org/papers/v18/16-365.htm

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press. https://www.kaggle.com/uciml/pima-indians-diabetes-databas

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., &amp; He, L. (2022, April 26). A survey of human-in-the-loop for machine learning. arXiv.org. Retrieved May 1, 2022, from https://arxiv.org/abs/2108.00941

Sase Y, Kumagai D, Suzuki T, Yamashina H, Tani Y, Fujiwara K, Tanikawa T, Enomoto H, Aoyama T, Nagai W, Ogasawara K. Characteristics of Type-2 Diabetics Who are Prone to High-Cost Medical Care Expenses by Bayesian Network. Int J Environ Res Public Health. 2020 Jul 22;17(15):5271. doi: 10.3390/ijerph17155271. PMID: 32707809; PMCID: PMC7432350.

Guo, Y., Bai, G., & Hu, Y. (n.d.). Using Bayes Network for Prediction of Type-2 Diabetes. IEEE.

Danilo F. de Carvalho, Uzay Kaymak, Pieter Van Gorp, Natal van Riel,
A Markov model for inferring event types on diabetes patients data,
Healthcare Analytics, Volume 2, 2022, 100024, ISSN 2772-4425,
https://doi.org/10.1016/j.health.2022.100024.
(https://www.sciencedirect.com/science/article/pii/S2772442522000053)

Liu S, Zhang R, Shang X, Li W. Analysis for warning factors of type 2 diabetes mellitus complications with Markov blanket based on a Bayesian network model. Comput Methods Programs Biomed. 2020 May;188:105302. doi: 10.1016/j.cmpb.2019.105302. Epub 2020 Jan 2. PMID: 31923820.

Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. doi: 10.3389/fgene.2018.00515

https://github.com/WillKoehrsen/ai-projects/blob/master/markovchainmontecarlo/markovchainmontecarlo.ipynb

https://pgmpy.org/detailed_notebooks/2.20Bayesian20Networks.html
https://jakevdp.github.io/PythonDataScienceHandbook/04.03-errorbars.html
https://en.wikipedia.org/wiki/Naive_Bayes_classifier
https://pandas.pydata.org
https://numpy.org
https://scikit-learn.org/stable/about.html
https://pgmpy.org
https://matplotlib.org