# Early Diabetes Classification Using Bayesian Networks

Lucas Oliveira

# Abstract

This study explores the use of Bayesian networks for early diagnosis of diabetes based on a set of symptoms. The dataset was collected through surveys conducted on the phone using the Behavioral Risk Factor Surveillance System. The study compares the performance of three Bayesian network models with a Random Forest model, which is widely used as a control model. The Bayesian networks outperformed the Random Forest model when trained on the same data, suggesting that they are better suited for predicting diabetes based on the selected variables. The PC algorithm and HC algorithm were used for estimating the structure of Bayesian networks. The preferred model had 61% specificity and 81% F1 score, outperforming the other models and the Random Forest model. Overall, the study demonstrates the effectiveness of Bayesian networks in early diagnosis of diabetes and suggests further exploration of additional biomarkers to improve the accuracy of diagnosis.

# Introduction

Diabetes is a chronic metabolic disorder that affects millions of people worldwide. It is characterized by high blood sugar levels, which can lead to a number of health complications, including heart disease, stroke, blindness, and kidney failure. Early detection and diagnosis of diabetes are crucial for effective management of the disease. However, traditional methods for diagnosing diabetes, such as glucose tolerance tests and fasting plasma glucose tests, are often invasive and expensive. These tests are also often performed only after the onset of symptoms, which can delay diagnosis and treatment.

Bayesian networks (BNs) have emerged as a promising alternative for early detection of diabetes. BNs are probabilistic graphical models that can be used to predict the probability of a disease given a set of risk factors. These networks can help identify patients who are at high risk of developing diabetes, even before the onset of symptoms. They have been shown to be effective in a variety of applications, including medical diagnosis, fraud detection, and weather forecasting.

Several studies have shown that BNs can be effective in predicting the risk of diabetes. For example, a study by Alaa et al. (2018) used a BN to predict the risk of type 2 diabetes in a UK population. The study found that the BN significantly outperforms the traditional deployed risk scores, such as the Rothman index, MEWS, APACHE, and SOFA scores, in terms of timeliness, true positive rate, and positive predictive value.

One advantage of Bayesian networks over other machine learning techniques is their ability to incorporate domain knowledge into the model. In a study Border et al. (2018), developed a Bayesian network model to predict the likelihood of diabetic nephropathy, a common complication of diabetes that affects the kidneys. Through the use of BN, the researchers were able to determine the most important features for classification and improve the queries around those features. The resulting probabilistic queries give clinicians an initial estimate for the likelihood of stage membership.

BNs have also been used in other areas of healthcare. For example, BNs have been used to predict the risk of heart disease (Alaa et al., 2019) and breast cancer (Burnside et al., 2004). These studies found that BNs were able to accurately predict the risk of these diseases, allowing for early detection and intervention.

Recent research has also focused on the use of machine learning techniques, such as deep learning and artificial neural networks, for the early detection of diabetes. These techniques have been shown to be effective in predicting the risk of diabetes based on large datasets of patient information. For example, a study by Wang et al. (2019) used a deep learning algorithm to predict the risk of type 2 diabetes in a Chinese population. The study found that the deep learning algorithm was able to accurately predict the risk of diabetes, even in patients with no previous history of the disease. In this research a comparison is going to be conducted to further evaluate the results of the Bayesian network by comparing the result of the BN with the results of a Random Forest classifier.

BNs offer a potentially valuable tool for improving the management and outcomes of diabetes care, as well as other areas of healthcare. By identifying patients at high risk of developing diabetes, healthcare providers can implement early interventions and preventative measures. This can lead to better outcomes for patients and reduce the burden of diabetes on healthcare systems.

In this report, we will explore the use of Bayesian networks for early detection of diabetes. We will first review the literature on Bayesian networks and diabetes. We will then describe our proposed approach for using Bayesian networks to detect diabetes. Finally, we will present the results of our experiments.

## Problem and dataset

The problem that this research focuses on is the early classification of diabetes. To address this problem, a dataset was chosen from Kaggle containing relevant data for this research. The data originates from Behavioral Risk Factor Surveillance System and was obtained through a series of surveys though the phone across all states of United States of America. The data was preprocessed and transformed into the dataset being used in this research by Alex Teboul in this Kaggle notebook.

The dataset consists of twenty-two columns and a total of 253,680 records, with 217789 of the subjects answered they had never been diagnosed with diabetes and 35241 had been previously diagnosed with diabetes, resulting in a ratio of 1:6.

| Column | Type | Definition |
|---|---|---|
| **Diabetes_binary** | Discrete [0-1] | Have you ever been told that you have Diabetes? |
| **HighBP** | Discrete [0-1] | Have you ever been told that you have high blood pressure by a doctor, nurse, or other health professional? |
| **HighChol** | Discrete [0-1] | Have you ever been told that you have high blood cholesterol by a doctor, nurse, or other health professional? |
| **CholCheck** | Discrete [0-1] | Have you checked your cholesterol levels in the last 5 years? |
| **BMI** | Continuous [12-95] | Calculated variable by dividing the weight in kilograms by the squared height in meters |
| **Smoker** | Discrete [0-1] | Have you smoked at least 100 cigarettes in your entire life? |
| **Stroke** | Discrete [0-1] | Have you ever been told you had a stroke? |
| **HeartDiseaseorAttack** | Discrete [0-1] | Have you ever reported having coronary heart disease or myocardial infarction? |
| **PhysActivity** | Discrete [0-1] | Have you done any physical activity or exercise in the last 30 days other than your regular job? |
| **Fruits** | Discrete [0-1] | Do you consume fruits one or more times per day? |
| **Veggies** | Discrete [0-1] | Do you consume vegetables one or more times per day? |
| **HvyAlcoholConsump** | Discrete [0-1] | Are you a heavy drinker? (Adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) |
| **AnyHealthcare** | Discrete [0-1] | Do you have any kind of health care coverage? |
| **NoDocbcCost** | Discrete [0-1] | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? |

|  |  |  |
|---|---|---|
| **GenHlth** | Discrete [1–5] | Would you say that in general your health is? (1: Excellent, 2: Very Good, 3: Good, 4: Fair, 5: Poor) |
| **MentHlth** | Discrete [0–30] | For how many days during the past 30 days was your mental health not good? Includes stress, depression, and problems with emotions. |
| **PhysHlth** | Discrete [0–30] | For how many days during the past 30 days was your physical health not good? Includes physical illness and injury. |
| **DiffWalk** | Discrete [0–1] | Do you have serious difficulty walking or climbing stairs? |
| **Sex** | Discrete [0–1] | Indicate sex of respondent. |
| **Age** | Discrete [1–13] | Discrete representation of respondents' age. (18–24, 25–29, 30–34, 35–39, 40–44, …, 75–79, 80 or older) |
| **Education** | Discrete [1–6] | What is the highest grade or year of school you completed? (1: Never attended school, 2: Elementary, 3: Some high school, 4: High school graduate, 5: Some college or technical school, 6: College graduate) |
| **Income** | Discrete [1–8] | What is your annual household income in $ from all sources? (Less than 10k, 10–15k, 15–20k, 20–25k, 25–35k, 35–50k, 50–75k, 75k or more) |

*Table 1: Dataset columns and the respective descriptions*

# Methods

The problem of early diabetes classification can be approached using a variety of machine learning methods and techniques. In this section, we will describe the different methods and models that were employed in order to solve the problem at hand. Firstly, we will discuss the feature selection process, which involves identifying and selecting the most important features from the available data set. We will then proceed to describe the machine learning models that were implemented for classification, including the Bayesian networks. Finally, we will outline the model evaluation methods that were used to determine the effectiveness of the models.

## Feature Selection

Feature selection is an important step in machine learning that involves selecting a subset of relevant features from a larger set of available features to improve the performance of a model. The main motivation behind feature selection is to simplify the model by removing irrelevant or redundant features.

In this research, Recursive Feature Elimination (RFE) was used to select the most relevant features for the models. By removing irrelevant features, we were able to simplify the models and improve their performance. The optimal number of features was determined using RFECV with five folds and a random forest as an estimator.

### Recursive Feature Elimination

Recursive feature elimination is a widely used feature selection method in machine learning. It works by recursively removing attributes and building a model on those attributes that remain. The goal is to identify a subset of input features that are most relevant to the target variable, which can lead to improved model performance and reduced overfitting.

In this study, we used RFE with a random forest classifier (RFC) as the estimator and the evaluation metric was balanced accuracy. This approach involves training the RFC model on the entire set of features, and then recursively eliminating the least important feature(s) based on the feature importance ranking provided by the model. The process is repeated until the desired number of features is reached or the performance metric no longer improves.

The use of balanced accuracy as the evaluation metric is particularly important in imbalanced datasets, where the class distribution is skewed towards one class. In our study, the original dataset had an imbalanced class distribution, with only 14% of samples belonging to the positive class (diabetic). Therefore, we used balanced accuracy, which takes into account both sensitivity (true positive rate) and specificity (true negative rate) to evaluate the model's performance on both classes.

## Bayesian Networks

Bayesian networks are probabilistic graphical models that represent the probabilistic relationships among a set of variables. Bayesian networks are based on Bayes' theorem, which states that the probability of an event occurring given some evidence is proportional to the prior probability of the event multiplied by the likelihood of the evidence given the event.

In the context of classification, Bayesian networks can be used to model the relationships among the input features and the target variable. This allows for the prediction of the target variable given some evidence about the input features.

## Estimating the Structure

The structure of a Bayesian network is crucial in determining the accuracy and effectiveness of the model. A poorly structured network can lead to incorrect predictions and conclusions.

Therefore, estimating the structure of Bayesian networks is an important task. There are various ways to estimate the structure of a Bayesian network, including constraint-based, score-based, and hybrid methods. Constraint-based methods use conditional independence tests to determine the graphical structure of the network. Score-based methods search for the optimal network structure that maximizes a given score function, such as Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC). Hybrid methods combine both constraint-based and score-based methods to estimate the network structure.

In addition to these methods, a domain knowledge approach can also be used to guide the structure learning process. Incorporating domain knowledge into the structure learning process can help to reduce the search space and improve the quality of the learned structure. For instance, domain knowledge can be used to identify variables that are known to be causally related.

### *PC Algorithm*

The PC algorithm is a commonly used method for estimating the structure of Bayesian networks. The algorithm uses conditional independence tests to iteratively build a network structure. It begins by assuming that all variables are marginally dependent, and then tests for conditional independence between pairs of variables, given all other variables. If the conditional independence test is satisfied, the algorithm adds an undirected edge between the two variables in the network. This process continues until no more edges can be added without violating the assumption of conditional independence. (Discovering Causal Structure from Observations)

### *HC Algorithm*

The Hill-Climbing (HC) algorithm is another popular method for estimating the structure of a Bayesian network. The HC algorithm starts with an initial graph and iteratively makes local changes to the structure until a good scoring DAG is obtained.

The algorithm begins by selecting an initial graph, often a fully connected graph, and evaluates its score. Then, it explores all possible single edge additions and deletions to the graph, evaluating the score of each new graph. If the score of a new graph is better than the current graph, the new graph becomes the current graph and the process continues. If no new graph improves the score, the algorithm terminates. (Adhitama & Saputo, 2022)

## Model Evaluation

### F1 Score

F1 score is a commonly used evaluation metric in classification tasks because it provides a more balanced measure of a model's performance on a binary classification task compared to other metrics such as classification.

The F1 score is the harmonic mean of the precision and recall.

$$\text{F1 score} = 2 * \frac{(precision * recall)}{(precision + recall)}$$

**Precision** measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

**Recall** measures the proportion of true positive predictions among actual positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

True positives (TP) are the samples that were correctly classified as positive by the model.

True negatives (TN) are the samples that were correctly classified as negative by the model.

False positives (FP) are the samples that were incorrectly classified as positive by the model.

False negatives (FN) are the samples that were incorrectly classified as negative by the model.

# Experimental Setup

This section aims to outline the various methods utilized to prepare and transform the data for the models, including feature engineering, discretization, feature selection, and resampling.

## Feature Engineering

Feature engineering is the of selecting and transforming input features to enhance the performance of a machine learning model.

### Discretization

For the BMI, MentHlth, and PhysHlth columns, a discrete column was created from the continuous data using the "cut" method in R.

| ***Column*** | **Old Format** | **Discretized Format** |
|---|---|---|
| *BMI* | Continuous [12-95] | 0-18.5 "*Underweight*", 18.5-25 "*Healthy range*", 25-30 "*Overweight*", 30-40 "*Obesity*", 40-50 "*Severe obesity*", 50-55 "*Super Severe obesity*" |
| *MentHlth, PhysHlth* | Continuous [0-30] | *"1":* 0-3 days, *"2":* 3-7 days*, "3":* 7-20 days, *"4":* 20-30 days |

*Table 2: Intervals of discretized columns*

### Outlier Detection

This step is specific to the BMI column. As indicated in the table above, the BMI range in the original dataset spanned from 12 to 95, which is biologically implausible. For reference, a person with a height of 170 cm would need to weigh 320 kg to achieve a BMI of 50. Therefore, all samples with a BMI over 55 were removed from the dataset.

## Resampling

The initial dataset had a class imbalance issue, with a ratio of one diabetic person to six non-diabetic persons. To address this issue, a simple test was conducted by fitting the data to a Random Forest classifier and using balanced accuracy as the performance metric. The model achieved an accuracy of 60% with the original ratio. To improve the performance, the data was resampled using both Random Over Sampler and Random Under Sampler techniques. The resampling resulted in a ratio of one diabetic person to 2.5 non-diabetic persons, and the model accuracy improved to 75%. The combination of the two techniques was used to avoid excessive data duplication while maintaining as much of the original data as possible.

### *Random Over Sampler*

Random Over Sampler is a data pre-processing technique used to address the problem of class imbalance in binary classification tasks. It works by randomly duplicating instances of the minority class in the training dataset, such that the resulting dataset has a balanced number of instances from each class. (Imblearn documentation - Randomoversampler)

### *Random Under Sampler*

Random Under Sampler is a data pre-processing technique used to address the problem of class imbalance in binary classification tasks. It works by randomly selecting a subset of the majority class instances in the training dataset, such that the resulting dataset has a balanced number of instances from each class. (Imblearn documentation - Randomundersampler)

## Feature Selection

Feature selection is a crucial step in machine learning, as it aims to identify the most relevant input features that contribute to the performance of a model. In this section, we will present the methods used for feature selection in our study.

### *RFECV with Random Forest Estimator*

The main method used for feature selection was Recursive Feature Elimination with Cross-Validation (RFECV) with five folds and a Random Forest as estimator. The RFECV is an iterative process that removes the least important features from a model until the optimal number of features is reached. The RFECV approach was chosen because it provides an automatic and systematic way of identifying the optimal number of features for a model.

However, the outcome of the RFECV was not satisfactory as it showed that the optimal number of features was 21 (the total number of features in the dataset). The number of features had to be reduced for the sake of simplifying the Bayesian network model.

### *Analysis of Feature Importance*

After inspecting the coefficients of both the random forest classifier and logistic regression to determine the order of feature importance to the model, it was discovered that the model was assigning a higher importance to the variables that had a greater range, such as Income, which has a range of 1 to 8, compared to the binary variables. Even upon scaling the data using standard scaler, the behavior was still the same.

### *Domain Knowledge Approach*

Therefore, a different approach was used. First, a domain knowledge approach was used, and upon review, it was decided that Income and Education could be dropped as these are not indicators of diabetes. Then, the variable CholCheck was dropped as this is only a variable to complement the variable HighChol.

### *Correlation Analysis and Judgment*

Finally, an analysis of the correlation with the target variable and judgment led to dropping the columns Smoker, Fruits, Veggies, DiffWalk, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, and MentHlth because these had a very low correlation (between -0.1 and 0.1) with the target variable.

Additionally, the variable Sex was dropped upon analysis of the correlation among with the fact that both sexes had the same amount of diabetic people. This approach allowed us to reduce the number of features from 21 to 10, resulting in a more concise and interpretable Bayesian network model.

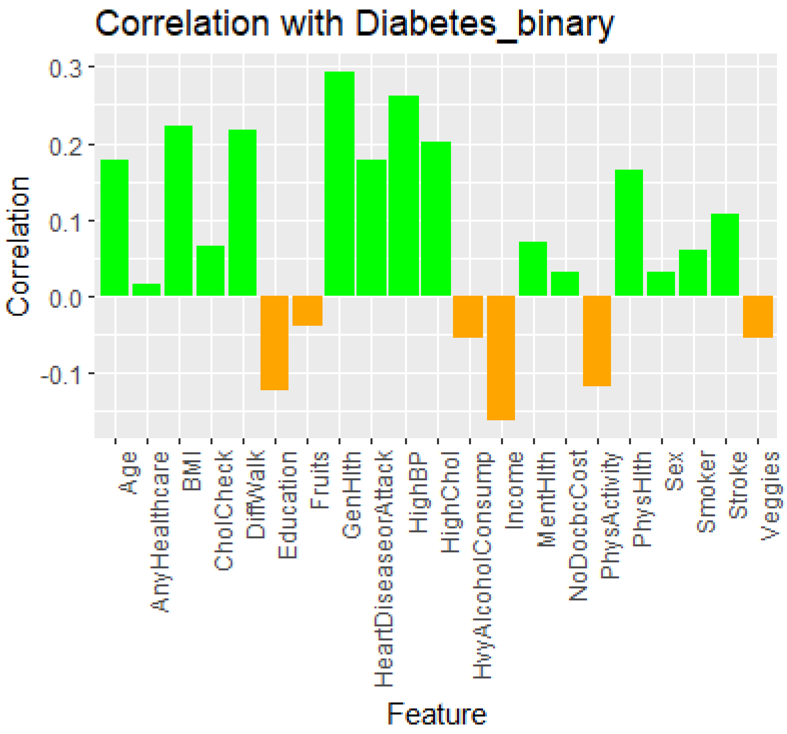

*Figure 1: Correlation of all the variables with the target variable*

# Results

This section aims to present the findings of the work, starting by the proposed network structure. The four figures below [2,3,4,5] show the Directed Acyclic Graphs (DAG) proposed.

The plots shown in figure 2 represent the network produced by the HC algorithm, this DAG was not considered appropriate because it maps causal relations that do not exist in real life, i.e., a causal relation from HighBP to PhysActivity or Age to BMI. The one in figure 3 which was produced by the PC algorithm was discarded because it contains edges which are not directed, together with the fact that it is overly complicated.

The DAG represented in figures 4 and 5 are the same except for the edges from Diabetes to HighBP, HighChol, Stroke, and HeartDisease. In figure 4 the edges are directed from these variables to Diabetes, and vice-versa. The cause for this change is because, according to several sources online Diabetes can cause or increases the risk of a person developing all of these symptoms. Therefor it was decided that the edges should be from diabetes to the symptoms to properly represent the relationship. Because of this a concern was raised that the model would not pick up the relation, due to the fact that the default predict method of the Bnlearn library only considers the parent nodes to calculate the probability of the target. The solution to this was that by changing the attribute "method" to "bayes-lw" the model now considers all the nodes available in the network not only the parents.

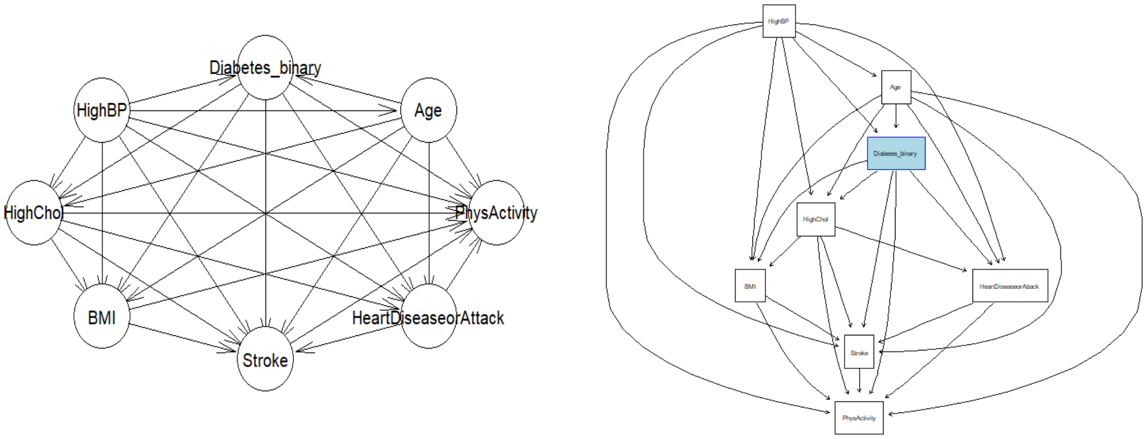

*Figure 2: Bayesian network after HC algorithm (dag 1)*

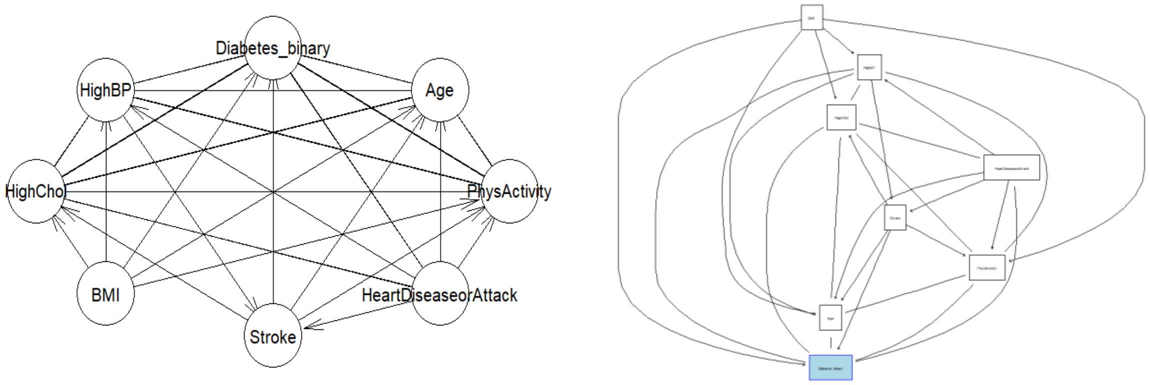

*Figure 3: Bayesian network after PC algorithm (dag2)*

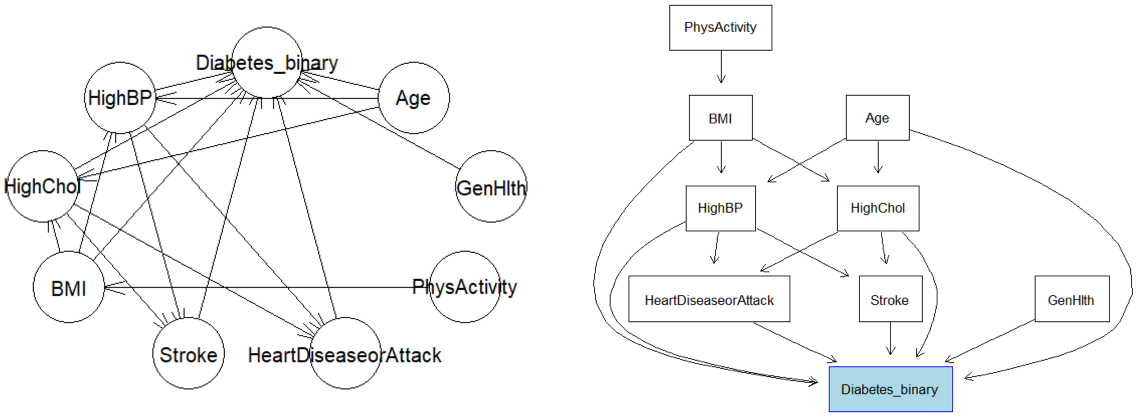

*Figure 4: Bayesian network with target variable in the bottom (dag 3)*

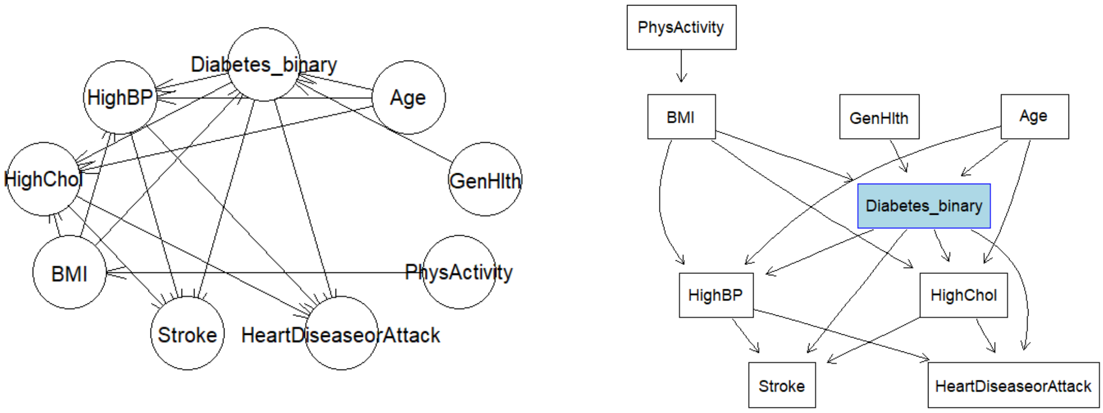

*Figure 5: Proposed final Bayesian network (dag 4)*

The results of the models are shown in the table below, including the results of the random forest classifier and the random forest classifier on the entire dataset.

| Metric/models | Dag 1 (HC) | Dag 3 | Dag 4 (Chosen BN) | Random Forest Classifier | Random Forest Classifier (All columns) |
|---|---|---|---|---|---|
| **Specificity** | 0.5867 | 0.5925 | 0.6130 | 0.5627 | 0.7732 |
| **Precision** | 0.8032 | 0.8060 | 0.8132 | 0.7969 | 0.8866 |
| **Recall** | 0.8433 | 0.8463 | 0.8258 | 0.8578 | 0.8869 |
| **F1 Score** | 0.8228 | 0.8256 | 0.8179 | 0.8262 | 0.8868 |

*Table 3: Evaluation of the models*

Out of the three Bayesian network models Dag 4, which was the preferred model, is better at predicting the positive class (Specificity) compared to the other two BN models. Compared to the Random Forest model the BN model outperformed the Random Forest which is widely used as a control model.

The performance of the Random Forest model trained on the entire dataset was notable as it outperformed the other models. However, upon a thorough analysis of the feature importance (see Appendix 1), the reliability of its results is questionable due to the model's tendency to assign higher importance to variables that are not reliable indicators of diabetes in the real world, such as income and education.

However, it's important to note that the Bayesian networks outperformed the random forest when trained on the same data. This suggests that the Bayesian networks are better suited to the specific task of predicting diabetes based on the variables selected. Additionally, using the entire dataset in the Bayesian networks makes the model overly complex and difficult to interpret.

Overall, the final preferred Bayesian network outperformed all other models in predicting the positive class, which is the most crucial aspect of the problem. It achieved a 61% specificity and 81% F1 score, slightly surpassing the other manually constructed BN with 59% specificity and 82% F1 score. Furthermore, the Bayesian network model outperformed the control Random Forest model, which had a 56% specificity and 82% F1 score.

# Discussion and Conclusions

The Bayesian network presented in this report demonstrated its effectiveness as a valuable tool for diabetes diagnosis. As stated in the introduction, Bayesian networks have been widely used in medical diagnosis due to their ability to model the complex relationships between variables.

The results of the study were satisfactory, however, it's worth noting that the set of variables used might not be the optimal set for early diabetes classification. It is possible that including other variables such as insulin levels or other symptoms such as experiencing excessive thirst, frequent urination, among others, could improve the performance of the model. Therefore, future studies could focus on expanding the set of variables to include a broader range of features that could improve the accuracy of the model.

It is also important to highlight that the data used in this study was collected through phone surveys and has not been verified, meaning that it cannot be solely relied upon for medical diagnosis. Nonetheless, the results of the study provide valuable insights that could be used in conjunction with other medical tests and examinations to assist in early diabetes diagnosis.

In conclusion, the Bayesian network showed promising results in predicting diabetes based on the selected variables. However, further studies are needed to validate the model's accuracy using a larger and more diverse dataset with a wider range of features. The results obtained in this study could potentially serve as a useful tool for medical practitioners in supporting early diabetes diagnosis.

# References

Alaa, A. M., Yoon, J., Hu, S., & van der Schaar, M. (2018). Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes. IEEE Journal of Biomedical and Health Informatics, 23(6), 2432-2443.

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE, 14(5), e0213653.

Burnside, E. S., Rubin, D. L., & Shachter, R. D. (2004). Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography. Studies in health technology and informatics, 107(Pt 1), 13-7. PMID: 15360765.

Adhitama, R. P., &amp; Saputro, D. R. (2022). Hill Climbing Algorithm for Bayesian Network Structure. 4th international conference on frontiers of biological sciences and engineering (FBSE 2021). https://doi.org/10.1063/5.0099793

Carnegie Mellon University. (n.d.). Discovering Causal Structure from Observations. Carnegie Mellon University.

Border, S., Jen, K. Y., Dos-Santos, W. L., Tomaszewski, J., & Sarder, P. (2020). Probabilistic modeling of Diabetic Nephropathy progression. Proc SPIE Int Soc Opt Eng, 11320, 1132014.

Sklearn Metrics Precision_score Documentation. (2023). Retrieved March 29, 2023, from

https://scikitlearn.org/stable/modules/generated/sklearn.metrics.precision_score.html#sklearn.metrics.precision_score

Sklearn Metrics Recall_score Documentation. (2023). Retrieved March 29, 2023, from https://scikitlearn.org/stable/modules/generated/sklearn.metrics.recall_score.html#sklearn.metrics.recall_score

Sklearn Metrics F1_score Documentation. (2023). Retrieved March 29, 2023, from https://scikitlearn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score

Imblearn documentation - Randomoversampler. imblearn documentation . (n.d.). Retrieved April 19, 2023, from https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

Imblearn documentation - Randomundersampler. imblearn documentation . (n.d.). Retrieved April 19, 2023, from https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html

# Appendix

### Appendix 1– Feature importance of Random Forest model

| Variable | Importance |
|---:|:---:|
| HighBP | 4209.2991 |
| HighChol | 2191.5132 |
| CholCheck | 267.7504 |
| BMI | 3845.8972 |
| Smoker | 1239.2109 |
| Stroke | 572.6516 |
| HeartDiseaseorAttack | 1131.5474 |
| PhysActivity | 1097.2973 |
| Fruits | 1220.8298 |
| Veggies | 1048.0101 |
| HvyAlcoholConsump | 489.3520 |
| AnyHealthcare | 415.5601 |
| NoDocbcCost | 699.5942 |
| GenHlth | 5871.5614 |
| MentHlth | 1824.5363 |
| PhysHlth | 2240.5864 |
| DiffWalk | 1681.1019 |
| Sex | 1156.4070 |
| Age | 6405.8759 |
| Education | 2910.0573 |
| Income | 4668.4549 |

### Appendix 2- Link to the dataset

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?datasetId=1703281&sortBy=voteCount