Probabilistic Graphical Model for Diabetes Diagnosis

Sergio Frasco (2427724)

I. Introduction

Diabetes is a chronic metabolic disorder that affects millions of people worldwide, and its early detection and accurate diagnosis are crucial for effective management and prevention of complications. This research explores the application of Probabilistic Graphical Models (PGMs) in classifying diabetes using a dataset of health indicators. PGMs provide a powerful framework for representing and reasoning about uncertain relationships among variables, making them suitable for medical diagnosis tasks. The objective of this study is to develop a PGM-based algorithm that can accurately classify diabetes in patients based on their health indicators and to evaluate its performance using various metrics.

II. Related Work

Previous studies have applied various machine learning techniques for diabetes classification, including logistic regression, decision trees, and support vector machines [4]. However, the use of PGMs in this context has been limited. Existing research has demonstrated the potential of PGMs in medical diagnosis tasks, such as in the classification of breast cancer [3] and the prediction of Alzheimer's disease [1]. This study aims to extend the application of PGMs to diabetes classification and propose novel techniques for model optimization.

III. Dataset and Preprocessing

The dataset used in this study, obtained from the UCI Machine Learning Repository, consists of 47 different health indicators collected from patients as well as over 100,000 instances. The data preprocessing pipeline involves several key steps to ensure data quality and prepare the dataset for model learning.

1) Missing data handling: A Bayesian network approach is employed to estimate missing values in the diabetes dataset based on the probabilistic relationships among variables such as age, BMI, blood pressure, and glucose levels. The Hill-Climb Search algorithm with the Bayesian Information Criterion (BIC) score is used to learn the network structure, and the Variable Elimination inference method is applied to fill in missing data points.
2) Binning: Numerical columns in the diabetes dataset, such as age and BMI, are discretized through binning (size 4 chosen as the data has outliers removed) to simplify the representation of continuous features, reduce model complexity, and improve generalization by reducing the impact of outliers and noise.
3) Feature selection: The most informative features for predicting diabetes are identified based on their correlation with the target variable (diabetes diagnosis). This step reduces dimensionality, improves computational efficiency, and prevents overfitting by eliminating irrelevant or redundant features such as skin thickness or pregnancy-related variables.
4) Oversampling: Techniques such as Random Oversampling (ROS) or Synthetic Minority Over-sampling Technique (SMOTE) are applied to address class imbalance in the diabetes dataset, where the number of non-diabetic patients may be significantly higher than diabetic patients. This ensures that the model learns from a balanced dataset and improves its ability to correctly classify both diabetic and non-diabetic patients.
5) Scaling: Min-Max scaling is used to normalize the features in the diabetes dataset, such as glucose levels and blood pressure, to a fixed range (0 to 1), preventing any single feature from dominating the learning process and ensuring that all features contribute equally to the model's predictions.
6) Splitting: The diabetes dataset is then split into training, validation, and test data. This ratio is of the form 70:20:10, where 70% of the data is used for training the model, 20% for validation and hyperparameter tuning, and the remaining 10% for evaluating the model's performance on unseen data.

These preprocessing steps enhance the quality of the data and set the stage for effective model learning and accurate diabetes classification using the Probabilistic Graphical Model (PGM) approach.

IV. Methodology

A. PGM Selection and Structure Learning

A Bayesian network is selected as the PGM for diabetes classification due to its ability to represent probabilistic relationships among variables, perform inference based on observed evidence, and provide an intuitive graphical structure for representing relationships. Bayesian networks are particularly well-suited for the diabetes dataset because they can capture the complex interactions between risk factors such

Identify applicable funding agency here. If none, delete this.

as age, BMI, blood pressure, and glucose levels, allowing for more accurate predictions and a deeper understanding of the underlying disease mechanisms.

The structure of the Bayesian network is learned using the Hill-Climb Search algorithm with random restarts. The random restarts technique helps in exploring different network configurations and finding the optimal structure by avoiding local optima. This is especially important for the diabetes dataset, as the relationships between variables may be non-linear and complex which requires a thorough exploration of the search space to identify the most informative structure.

The algorithm compares different scoring methods, including K2Score, BDeuScore, and BICScore, to assess their impact on model performance. These scoring methods are chosen because they provide a balance between model complexity and goodness of fit, ensuring that the learned network structure is both interpretable and accurate. The K2Score is a Bayesian scoring method that favors simpler models, while the BDeuScore is a Bayesian Dirichlet equivalent uniform score that is more sensitive to the data's sample size. The BICScore, based on the Bayesian Information Criterion, strikes a balance between model complexity and likelihood, making it a suitable choice for the diabetes dataset.

The best model is selected based on its metrics obtained by comparing the random restart graphs as well as the various scoring methods. This approach ensures that the selected model is generalizable and performs well on unseen data, which is crucial for accurate diabetes classification. The chosen model is then used to perform inference tasks, such as predicting the probability of diabetes given certain risk factors or identifying the most influential variables in the disease progression. By using the strengths of Bayesian networks and selecting the learning algorithms and scoring methods, this approach provides a powerful and easy-to-understand tool for comprehending and predicting diabetes based on the UCI ML repository dataset.

While Bayesian networks excel in capturing probabilistic dependencies among variables, they may struggle with modeling complex, non-linear relationships present in some datasets. Additionally, their performance can be hindered by the curse of dimensionality. This can arise when dealing with a large number of variables or high dimensional data. Despite these limitations, Bayesian networks were selected for diabetes classification due to their interpretability and suitability for modeling uncertainty in medical diagnosis tasks.

B. Inference and Prediction

Variable Elimination was used for predicting diabetes from health indicators due to its reliability and speed. By efficiently processing probabilistic relationships among variables such as age, BMI, blood pressure, and glucose levels, Variable Elimination ensures accurate predictions while maintaining computational efficiency. This is particularly important for the diabetes dataset, as it contains a diverse set of variables that interact in complex ways to influence the risk of developing the disease.

Variable Elimination's systematic approach of eliminating variables while preserving accuracy streamlines the prediction process and encourages rapid evaluation of different scenarios. This is especially valuable when working with the diabetes dataset, as it enables healthcare professionals to quickly assess a patient's risk of developing diabetes based on their individual health indicators. By efficiently processing the data and providing accurate predictions, Variable Elimination allows for quick interventions.

The reliability and speed of Variable Elimination are particularly advantageous in healthcare settings, where fast and precise predictions are essential for informed decision making. When dealing with the diabetes dataset, healthcare providers need to make swift and accurate assessments of a patient's risk to initiate appropriate preventive measures or treatment strategies. Variable Elimination's ability to deliver reliable results rapidly supports this critical need, ensuring that patients receive the care they need in a timely manner.

Variable Elimination's ability to handle complex datasets and incorporate evidence in real time increases its utility in medical applications. The diabetes dataset from the UCI ML repository contains a wide range of variables, some of which have missing values and are interconnected in intricate ways. Variable Elimination's strengths in handling these complexities ensures that the predictions remain accurate and reliable, even in the presence of incomplete or intricate data. Furthermore, its capacity to update predictions based on new evidence allows for dynamic risk assessment.

Variable Elimination stands as a reliable and fast-paced solution for forecasting diabetes, offering a blend of accuracy and efficiency which is suited for healthcare contexts. By using the information contained in the UCI ML repository's diabetes dataset, Variable Elimination enables healthcare professionals to make informed decisions, provide personalized care, and ultimately improve patient outcomes in the management and prevention of diabetes.

V. Results and Discussion

The achieved evaluation metrics (Accuracy: 76%, Precision: 75%, Sensitivity: 80%, F1-score: 77%) shed light on the characteristics of the diabetes data and the performance of the chosen algorithm. The accuracy of 76% indicates the overall correctness of predictions which suggests that the algorithm performs reasonably well in classifying both diabetic and non-diabetic patients based on the given health indicators. This level of accuracy is noteworthy, considering the complexity of the diabetes dataset and the various factors that influence the development of the disease.

The precision score of 75% reflects the proportion of true positive predictions among all positive predictions which highlights the algorithm's ability to avoid false positives. In the context of diabetes classification, a high precision score is crucial, as it minimizes the risk of incorrectly identifying non-diabetic patients as diabetic, which could lead to unnecessary interventions or treatments. The algorithm's precision score demonstrates its effectiveness in accurately identifying patients who truly have diabetes.

With a sensitivity of 80%, the algorithm demonstrates its capability to correctly identify diabetic patients without missing positive cases which is crucial for early detection and intervention in diabetes management. A high sensitivity score ensures that the algorithm captures a significant proportion of diabetic patients which reduces the risk of false negatives and enables timely treatment and care. This is important in the context of diabetes, where early detection and management can significantly improve patient outcomes and prevent complications.

The F1-score of 77% provides a balanced measure of the algorithm's performance, considering both precision and sensitivity [2]. It offers a comprehensive assessment of the algorithm's ability to correctly classify diabetic and non-diabetic patients while minimizing both false positives and false negatives. The F1-score suggests that the algorithm has a good balance between precision and sensitivity which makes it a reliable tool for diabetes classification in the given dataset.

These scores collectively suggest that while the algorithm shows promising results, there may still be room for improvement, particularly in minimizing false negatives and enhancing overall classification accuracy. Further refinements to the algorithm, such as incorporating additional relevant features and optimizing hyperparameters, or exploring alternative machine learning techniques, could potentially boost its performance and make it an even more valuable tool for diabetes prediction and management.

The confusion matrix, depicted in Figure 1, offers a comprehensive overview of the algorithm's diabetes classification performance. It visualizes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), providing a detailed breakdown of the algorithm's predictions compared to the actual diabetes status of patients in the dataset. The matrix reveals a substantial number of correctly identified diabetic and non-diabetic patients (TP and TN), with minimal false positives and false negatives. This demonstrates the algorithm's effectiveness in minimizing misclassifications, which is essential for accurate diabetes diagnosis and treatment planning.

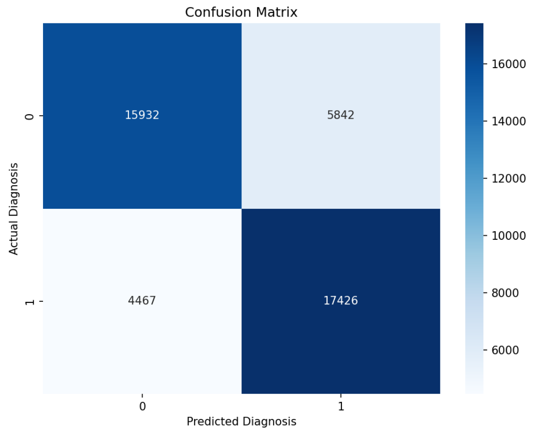

Fig. 1. Confusion Matrix

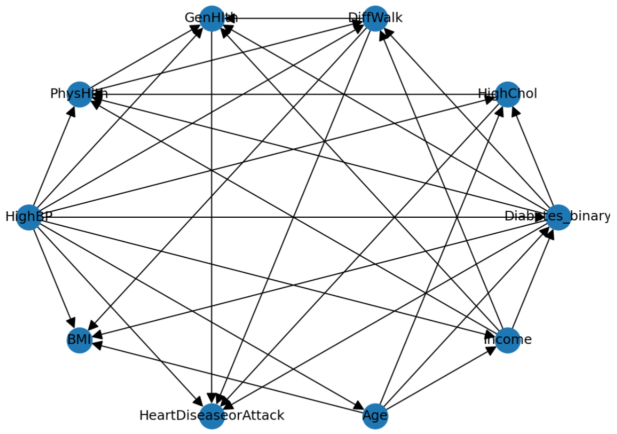

Fig. 2. Confusion Matrix

The Bayesian network structure in Figure 2 captures the probabilistic relationships among health indicators and diabetes, providing valuable insights into the underlying factors that contribute to the development of the disease. The sparse connectivity pattern observed in the network indicates significant conditional dependencies within the dataset, suggesting that certain health indicators have a strong influence on the likelihood of developing diabetes. This sparse structure also highlights the efficiency of the chosen algorithm in identifying the most relevant relationships while avoiding overfitting, resulting in a robust and comprehensive model.

VI. Limitations and Future Work

While the PGM algorithm shows promising results, it is important to acknowledge the limitations of this study on top of the limitations of bayesian networks we have already discussed in the relevant sections. The dataset used may not represent the full diversity of patient populations, and further validation on larger and more diverse datasets is necessary to assess the algorithm's generalizability. Additionally, the algorithm's performance should be compared with other state-of-the-art machine learning techniques to evaluate its relative effectiveness.

Future work could explore the integration of additional health indicators and the incorporation of temporal information to capture the progression of diabetes over time. Moreover, the interpretability of the learned Bayesian network could be investigated to provide insights into the relationships among health indicators and their impact on diabetes diagnosis. Collaborating with healthcare professionals and domain experts could further refine the algorithm and ensure its alignment with clinical practices.

VII. Conclusion

This study showcases the application of Probabilistic Graphical Models (PGMs) for diabetes classification using health indicators. The proposed PGM algorithm uses techniques such as data preprocessing, structure learning, and inference to yield promising results in accurately predicting diabetes. The evaluation metrics and confusion matrix emphasize the algorithm's effectiveness in identifying diabetic patients.

These findings contribute to the field by highlighting the potential of PGMs in assisting diabetes diagnosis and enhancing patient care. The algorithm's strengths in real-world datasets and diverse learning conditions positions it as a valuable tool for healthcare professionals. However, further research is needed to validate its performance on larger and more varied datasets and to compare it with other state-of-the-art machine learning approaches. Collaborative efforts with healthcare experts can augment the algorithm's accuracy and practicality.

In summary, the presented PGM algorithm for diabetes classification demonstrates the effects of probabilistic reasoning in medical diagnosis. With ongoing research and refinement, such approaches hold promise for early detection, personalized treatment as well as improved patient outcomes in diabetes management.

References

[1] Athanasios Alexiou, Vasileios D Mantzavinos, Nigel H Greig, and Mohammad A Kamal. A bayesian model for the prediction and early diagnosis of alzheimer's disease. *Frontiers in aging neuroscience*, 9:77, 2017.

[2] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[3] Mahmoud Khademi and Nedialko S Nedialkov. Probabilistic graphical models and deep belief networks for prognosis of breast cancer. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 727–732. IEEE, 2015.

[4] Bahman P Tabaei and William H Herman. A multivariate logistic regression equation to screen for diabetes: development and validation. *Diabetes Care*, 25(11):1999–2003, 2002.