Data mining - CDC Diabetes Health Indicators

Mohamed Ndiaye

Bill Muchero

Bhavya Shri Meda

Raghavendra Naidu Nayanuri

CIS 635, Knowledge Discovery & Data Mining, Grand Valley State University

Dr Yong Zhuang

PROJECT FINAL REPORT

INTRODUCTION

Our project examines the relationship between lifestyle factors and diabetes prevalence in the US, specifically focusing on less obvious determinants such as education level, mental health, and income. By analyzing these variables, we aim to uncover underlying issues contributing to diabetes beyond common factors like High Cholesterol, poor diet, and inactivity. Our team, with diverse backgrounds in business and health, brings a comprehensive perspective to this investigation. We chose this dataset because diabetes is a compelling topic with various associated factors. It is an excellent subject for a data mining project to explore how lifestyle elements, beyond medical conditions, contribute to diabetes risk. Give an overview of your approach and results.

RELATED WORK

1)*"Deep" Learning for Missing Value Imputation Tables with Non-numerical Data*

This article highlights the importance of data quality and completeness in data-processing applications, presenting a scalable imputation method for non-numerical values, including unstructured text. The proposed method outperforms traditional imputation techniques in both accuracy and scalability, with a median imputation F1 score of 0.93 across diverse datasets. This is particularly relevant to the limitations of our project, as it demonstrates advanced imputation techniques that could address the challenges we faced with non-numerical data and enhance the overall robustness of our data analysis.

2) *Error Consistency for Machine Learning Evaluation and Validation with Application to Biomedical Diagnostics*

This article discusses the significance of supervised machine learning classification in both industry and academic research, emphasizing the need for rigorous evaluation before real-world deployment. It introduces an enhanced technique for hold-out validation that not only tests the model on different samples but also assesses the consistency of the mistakes made by the learning algorithm, thereby improving the reliability and predictability of AI models. This technique is particularly relevant to our project's limitations, as it highlights the importance of validating models comprehensively, ensuring that our imputation methods and machine learning models are both accurate and consistent in their performance.

3) *Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction*

This article highlights the potential of automated machine learning (AutoML) tools to enhance the diagnosis and prediction of heart disease, a leading cause of mortality globally. The study evaluates PyCaret, AutoGluon, and AutoKeras, demonstrating that AutoML tools outperform traditional machine learning models, with AutoGluon achieving the highest accuracy rates. This is particularly relevant to our project, as it underscores the effectiveness of PyCaret and similar tools in generating robust models without requiring extensive expertise, thereby addressing our issue with the PyCaret installation and extending its application to non-numerical data for improved data analysis and model reliability.

METHODOLOGY

The diagram below represents the different steps of our methodology for this data mining project.

Data Collection

The data acquired was used to understand better the relationship between lifestyle and diabetes in the US. We collected 21 features, and the target variable was whether a patient had diabetes, was pre-diabetic or did not have diabetes (a total of 22 columns). The link below presents the summary and characteristics of the dataset:

https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

The dataset used as a reference is accessible at the link below:

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

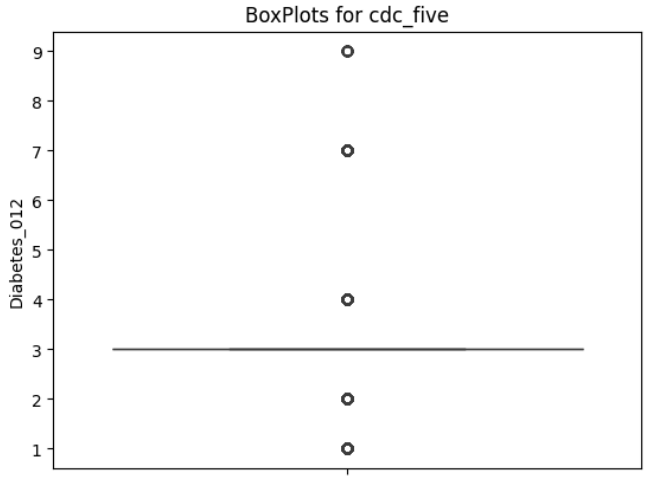
We noticed that the specified dataset had already been cleaned (BRFSS2015). We then acquired the original file from the CDC survey and performed the data preprocessing stage to master and better understand the techniques. The file we loaded had around 400,000 records and 330 features. To stay consistent with our study, we worked on getting the 22 features presented on BRFSS2015.

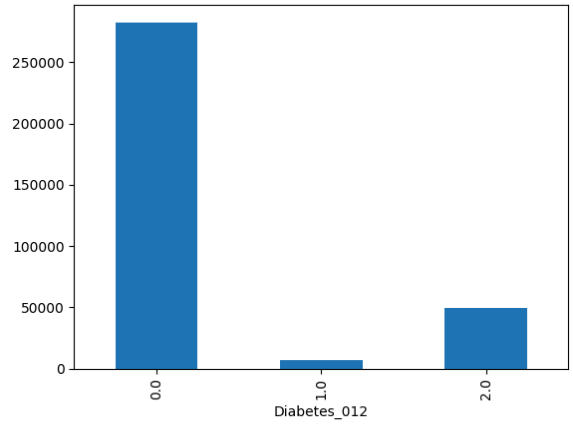
Data Cleaning

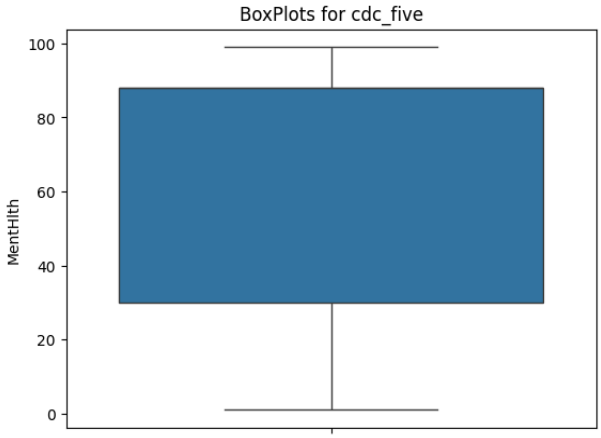
The following points represent the different activities performed to clean and prepare the data for the mining algorithm.

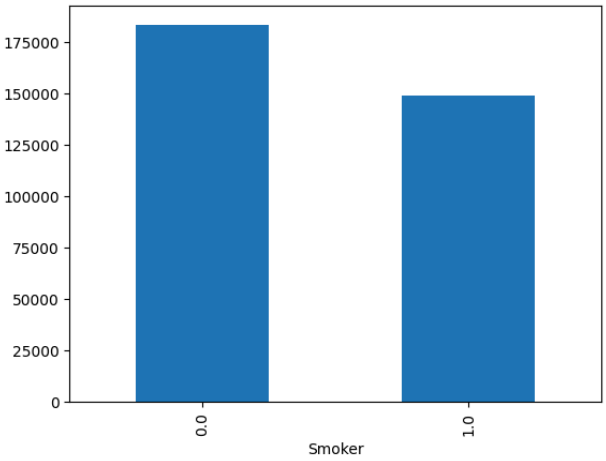
- Select and rename features based on the BRFSS 2015 dataset.
- Choose 5 key features: Diabetes_012, HighChol, Smoker, MentHlth, Education, Income.
- Study the target variable Diabetes_012, focusing on clear answers relevant to the study. Values such as “not sure” or “refused to answer” were excluded from the dataset for their lack of relevance.
- Based on the dataset analysis, clean and categorize Diabetes_012 into three groups: no diabetes, Prediabetes, and Diabetes.
- Remove outliers from the MentHlth, Smoker, HighCol, Education, and Income variables, ensuring that the values retained in the dataset are all relevant to the study.

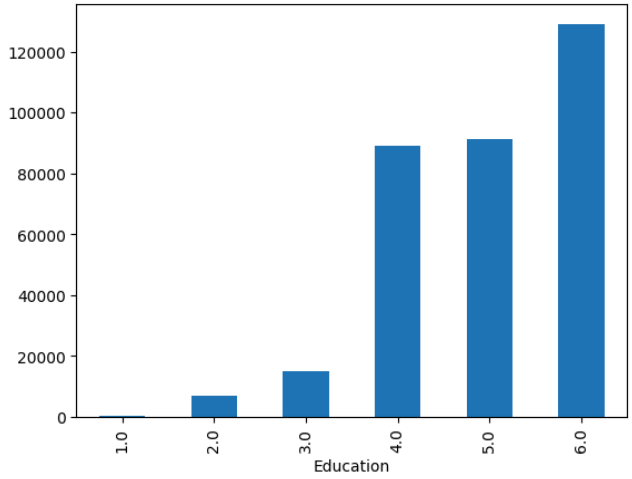
Graphs like boxplots helped identify the outliers in the feature variables, while bar charts helped measure the proportion of the values for a feature after cleaning.

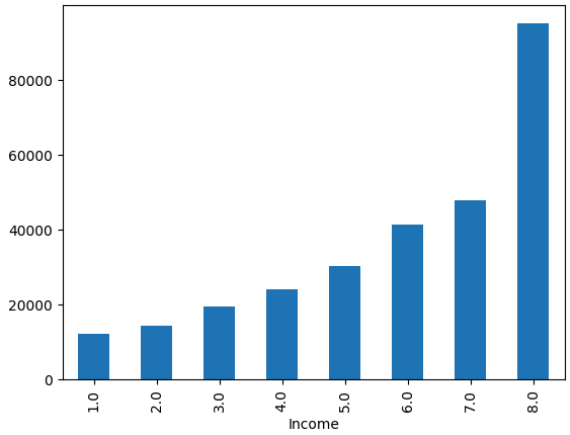












Selecting the model: Association rule Technique

We chose the association rule technique because it effectively identifies interesting relationships (associations) between variables in large datasets. For example, in retail, association rules can uncover patterns like "Customers who buy product A also tend to buy product B." In our study, we wanted to see whether there are relations between the 5 features selected and the likelihood of having diabetes (Diabetes_012 = 1).

Data Preparation for the Model
Before running the model, we copied the dataset containing only the rows where Diabetes_012 = 1. By targeting positive diabetes, we can see how the other features relate to the likelihood of having diabetes. With the Association rules technique, we have different parameters for evaluating the results of the model:
Support: Support refers to the frequency of an itemset's occurrence in a dataset. It measures how frequently an itemset appears in the dataset.
Confidence: Confidence measures the reliability or certainty of the association rule. It tells us, expressed as a percentage, how likely item Y will be purchased when item X is purchased.
- Lift: Lift measures how often the antecedent (X) and consequent (Y) of a rule occur together compared to what we would expect if they were statistically independent. Lift values greater than 1 indicate that the occurrence of X and Y together is more frequent than expected by chance, suggesting a stronger association between X and Y.

Running the Model
We used the Apriori Algorithm in Python to perform the association rules technique with a minimum support of 0.05.

RESULTS AND DISCUSSION

Following the completion of our model and subsequent analysis, our focus centered on Diabetes_012 as the primary outcome of interest. Our investigation delved into the antecedent relationships, revealing insightful support and lift values. Notably, we found that 62% of individuals diagnosed with diabetes also exhibited High Cholesterol, while 49% were identified as smokers. Remarkably, 32% of those diagnosed with diabetes presented both conditions concurrently.

These findings underscore the significant co-occurrence of High Cholesterol and smoking among diabetic individuals, highlighting their substantial roles as contributing factors to the likelihood of diabetes onset. Furthermore, the lift values substantiated the statistical relevance of these associations, further emphasizing their critical implications for understanding and potentially mitigating diabetes risk factors.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(HighChol)	(Diabetes_012)	0.620696	1.0	0.620696	1.0	1.0	0.0	inf	0.0
2	(Smoker)	(Diabetes_012)	0.492867	1.0	0.492867	1.0	1.0	0.0	inf	0.0
6	(Smoker, HighChol)	(Diabetes_012)	0.329465	1.0	0.329465	1.0	1.0	0.0	inf	0.0

After analyzing the model results, additional relevant points to consider include:

1. Association Strength: Besides support and lift values, examining confidence scores can provide insights into how strongly the antecedent variables (High Cholesterol, smoking) are associated with the consequence (Diabetes_012). Higher confidence scores indicate more reliable relationships between variables.
2. Variable Importance: Utilizing feature importance metrics such as feature weights or coefficients from the model can highlight which predictors have the most significant impact on predicting Diabetes_012. Understanding these influences can guide prioritization in interventions or further studies.
3. Interaction Effects: Exploring interactions between variables, especially those not directly mentioned in the support and lift values, can uncover nuanced

relationships. For instance, interactions between smoking and demographic factors like age or gender might reveal additional insights into diabetes risk.

These points deepen the analysis by providing a broader perspective on the relationships and factors influencing the likelihood of diabetes beyond the direct associations initially studied.

CONCLUSION

This project was beneficial in learning how to mind data and answer research questions. Our study was focused on analyzing the relationship of feature variables in the likelihood of having diabetes. The association rules technique was great for performing the analysis with the expected results. While working on our project, we encountered issues installing PyCaret on Python 3.10, as the **arules** module was unavailable. This version incompatibility required us to seek alternative packages or adjust our Python environment. Adjusting our setup was time-consuming and introduced additional inconsistencies. Ultimately, we had to manage our dependencies to maintain project stability. Future work could focus on ensuring data consistency by standardizing values across the dataset. Additionally, developing applications capable of analyzing data in various formats beyond just numerical would enhance the comprehensiveness and applicability of the analysis.

DATA AND SOFTWARE AVAILABILITY

Software
Google Colab,
Github
Links
GitHub Repository: https://github.com/GVSU-CIS635/term-project-CDC-Diabetes
Colab Files:
https://colab.research.google.com/drive/1da-f0f_HvIQ8zXDt4VofpXUoJgFkYfF8
Association
rules:https://colab.research.google.com/drive/1bb3GIjJxj1dkvtG7dsfF7XmMkjBkvSnt

REFERENCES

Alex Teboul, CDC
CDC Diabetes Health Indicators
https://doi.org/10.24432/C53919

1.*"Deep" Learning for Missing Value Imputation Tables with Non-numerical Data*

Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018, October). " Deep" Learning for Missing Value Imputation In Tables with Non-numerical Data. In Proceedings of the 27th ACM international conference on information and knowledge management (pp. 2017-2025).Link: https://ssc.io/pdf/p2017-biessmann.pdf

2.Error Consistency for Machine Learning Evaluation and Validation with Application to Biomedical Diagnostics

Levman J, Ewenson B, Apaloo J, Berger D, Tyrrell PN. Error Consistency for Machine Learning Evaluation and Validation with Application to Biomedical Diagnostics. Diagnostics. 2023; 13(7):1315.Link: https://www.mdpi.com/2075-4418/13/7/1315

3. *Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction*

Paladino LM, Hughes A, Perera A, Topsakal O, Akinci TC. Evaluating the Performance of Automated Machine Learning (AutoML) Tools for Heart Disease Diagnosis and Prediction. AI. 2023; 4(4):1036-1058. https://doi.org/10.3390/ai4040053