Diabetes Risk Assessment and Patient Stratification: A Two-Phased Analytical Study

Natalie Chuang, Sara Pasquino

December 2023

Contents

1 Introduction

Type 2 diabetes is a widespread public health concern within the United States and globally, causing severe health complications and significant economic burden. According to the Centers for Disease Control and Prevention (CDC), about 38 million Americans have type 2 diabetes and approximately 20% of those individuals are undiagnosed. The key risk factors of type 2 diabetes include having prediabetes, being overweight, being physically active less than 3 times per week, being of certain race or ethnicity, and being 45 years or older, although increasingly younger adults and children are developing diabetes. Diagnosis not only results in significantly higher costs for healthcare providers, estimated to be $413 billion annually, but also for patients. Medical costs for diabetes patients are reported to be more than twice the costs of non-diabetes individuals [1]. Diabetes is therefore not only a public health crisis but an economic one as well.

2 Problem Summary and Data Source

This project aims to address diabetes risk by understanding the key drivers of the disease. To that end, we use classification and clustering methods to perform our analysis and leverage data from the Behavioral Risk Factor Surveillance System (BRFSS), an annual survey by the CDC. The goals of this work are twofold: first, to predict an individual patient's diabetes risk and second, to categorize patients into distinct, interpretable clusters. These results provide an understanding of patient risk profiles, allowing healthcare providers to tailor prevention and treatment strategies appropriately based on patient needs.

The Behavioral Risk Factor Surveillance System (BRFSS) is a comprehensive annual survey of U.S. residents in all 50 states as well as the District of Columbia and 3 U.S. territories which gathers data on the health and lifestyle behaviors of respondents. The BRFSS questionnaire which includes a core component and additional optional modules that individual states may opt into. The core module gathers demographic and health behavior data on its respondents. The optional modules provide additional information of topics including specific diseases and conditions, physical and mental health, and lifestyle behaviors. We utilized the 2015 results for this work, which contains 300+ features and 400,000+ responses [2].

3 Data Preprocessing

The BRFSS dataset included many additional calculated features in addition to the survey questions, so the first data preprocessing step involved removing highly correlated features as well as survey identifier features. This step reduced the initial 329 columns to 248. Because many of the BRFSS survey questions were part of optional modules and therefore not opted into, many of the features in the dataset had very few responses. We thus removed features missing more than 50% of values, which reduced the number of columns to just 78.

The remaining 78 columns still suffered from missing data. Dropping missing values was not appropriate since most rows contained some missing values. Instead, we utilized optimal k-NN imputation from the IAI OptImpute package [3]. Then we used the OptimalFeatureSelection method to further sparsify our features [4]. We observed no improvement in validation AUC using more than 9 features when testing sparsity of up to 15 features. These results can be seen in Figure 1. The 9 features chosen included BMI category (underweight, normal, overweight, and obese), whether patient has ever had high blood cholesterol levels, whether a patient has had a heart attack, time since a patient's last cholesterol check, estimated VO2 max, and general health and physical health related features. We also included features that represented the key risk factors of diabetes reported by the CDC (age, prediabetes or gestational diabetes, exercise frequency, and race) that were not already captured in the chosen features. After obtaining our final set of features, we performed one-hot encoding for correct interpretation of the categorical variables.

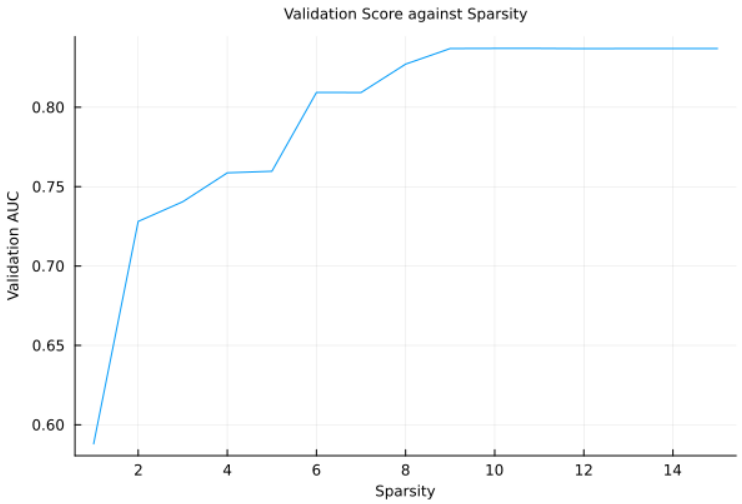

Figure 1: Plot of sparsity vs validation AUC for sparsity 1-15

Since our data is imbalanced with only 12% of observations having a positive outcome for diabetes, we chose to under-sample the negative outcomes in order to balance the two classes. After performing a stratified split of our data into train and test sets using a 70/30 split, we rebalanced the training set to have an equal number of positive and negative outcomes.

4 Methodology and Results

4.1 Classification with Tree-based Models

Firstly, we applied four different tree-based classification methods with grid search for hyperparameter tuning. Specifically, we employed CART (Figure 2), Random Forest, XGBoost, and Optimal Classification Tree (Figure 3).

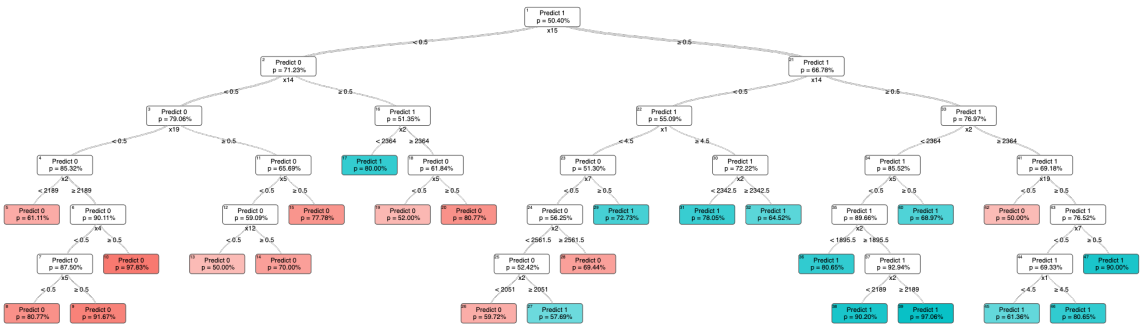

Figure 2: CART

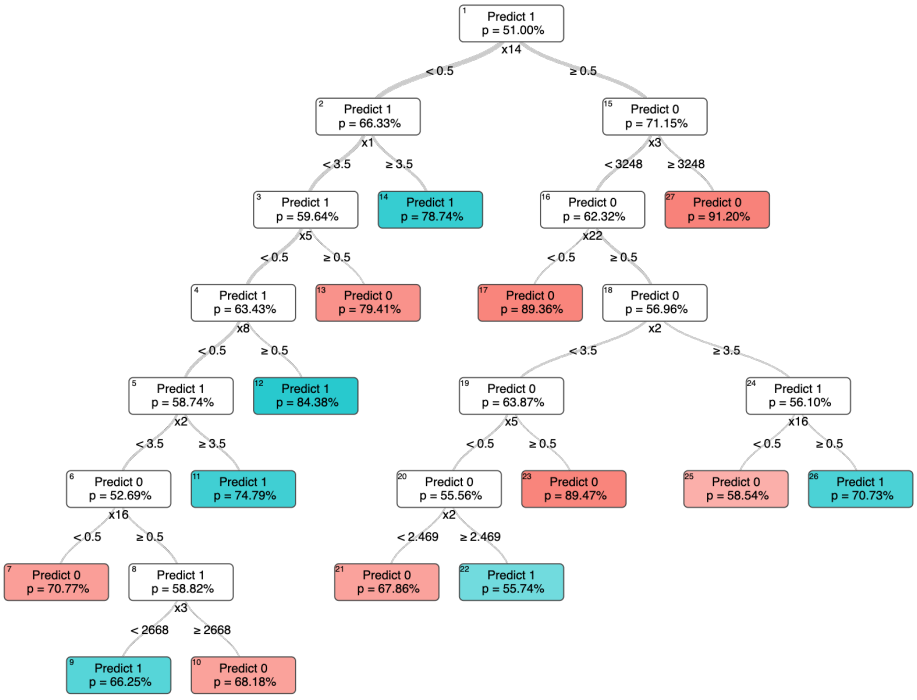

Figure 3: OCT

The performance of all four models was quite good and similar to one another (Table 1), with an average AUC above 0.8 for in-sample classifications and above 0.7 for out-of-sample classifications.

Model	CART	Random Forest	XGBoost	OCT
In-sample AUC	0.8203	0.8222	0.8833	0.8213
Out-of-sample AUC	0.7635	0.7887	0.7837	0.7352

Table 1: In-sample and Out-of-sample AUC for various models

4.2 Ensembling with Optimal Policy Trees

To leverage the strengths of diverse predictive models and enhance prediction accuracy, we employed model ensembling using Optimal Policy Trees. This technique integrates insights from individual models, reducing overfitting. Our goal was to develop a robust, generalizable model that outperformed of any single initial model in predicting outcomes with greater confidence and precision. The ensemble was implemented using an Optimal Policy Tree (Figure 4) that minimized the logarithmic loss between the true and predicted values for each observation.

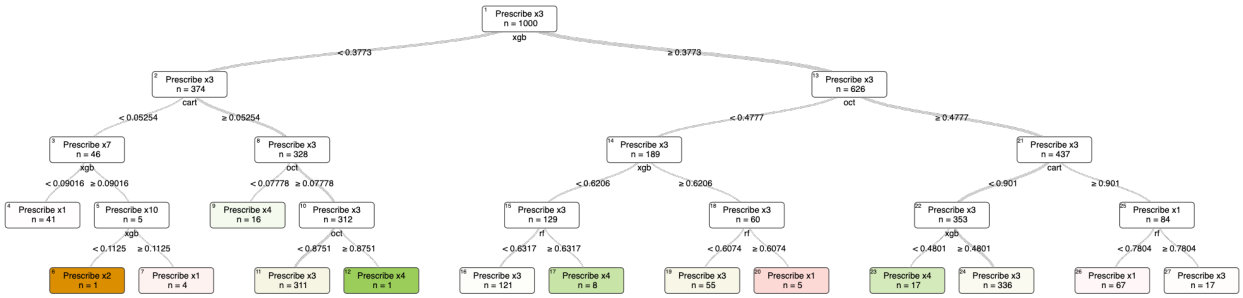

Figure 4: OPT

The resulting model achieved an out-of-sample AUC of 0.7892, improving over the performance of all other models individually.

4.3 Weighted Interpretable Clustering

This step was the core of our analysis: clustering individuals in our dataset into groups representing different risk exposures to developing type 2 diabetes, along with the characteristic features of each cluster.
The analysis proceeded in two steps:

- Weighted DBSCAN clustering of the complete dataset.
- Optimal Classification Tree analysis on the clusters to extract relevant features and provide interpretability.

In the first step, before standardizing the dataset, we assigned greater weight to the "probability of developing diabetes" to ensure the resulting clusters not only captured feature differences but also variations in disease risk probability. After several clustering methods including k-means and hierarchical clustering, we chose DBSCAN for its superior performance.

Then, we grouped the data by clusters and computed summary statistics for each cluster's diabetes risk probabilities, resulting in the BoxPlot shown in Figure 5.

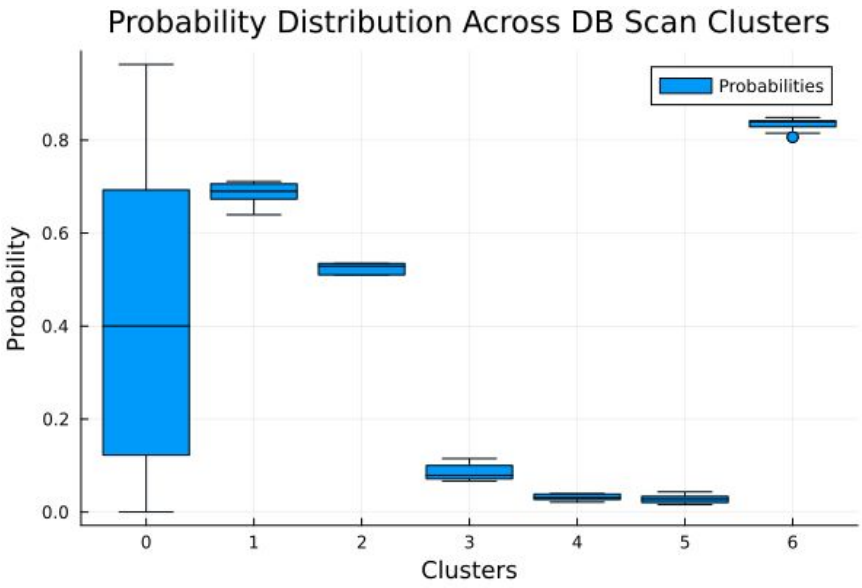

Figure 5: Probability by Cluster

Significant variability in diabetes risk probability between clusters was observed, indicating each cluster's distinct risk profile. Cluster 0 was the most diverse, encompassing a broad risk range, while other clusters exhibited lower variability within clusters and were positioned at distinct risk levels. The clusters were categorized as follows:

- **"Hard to Assess Risk"** (cluster 0): This cluster displays a wide range of diabetes risk (16% to 70%) with a median around 40%, making risk assessment challenging. These patients have varied health indicators, making them difficult to cluster with other patient groups.
- **"Higher Risk"** (cluster 1): Exhibits a consistently higher risk, around 70%, with minimal variability.
- **"Moderate - High Risk"** (cluster 2): Characterized by a moderately high probability of 50%, with low within-cluster deviation.
- **"Low Risk"** (cluster 3): Centered around a risk below 20%, indicating low risk with very low variability.
- **"Lowest Risk"** (clusters 4 and 5): Represent the lowest risk exposure, below 10%, with minimal deviation. These patients have excellent general health and engage in physical activity.
- **"Highest Risk"** (cluster 6): Individuals here are at serious risk, with probabilities tightly concentrated around 80%.

Upon clustering, we applied an Optimal Classification Tree to identify the distinguishing features of each cluster, as shown in Figure 6.

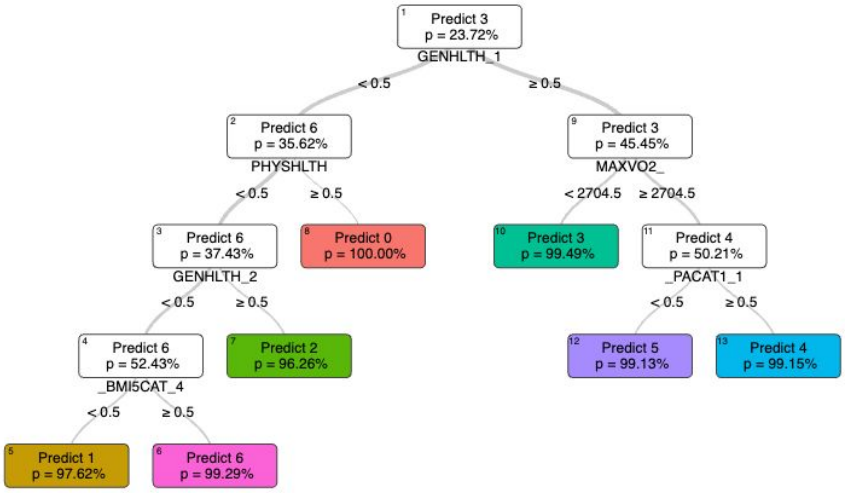

Figure 6: Cluster OCT

The tree revealed the following cluster characteristics:

- **"Higher Risk"** (cluster 1): Individuals with poor general and physical health, but not obese.
- **"Moderate - High Risk"** (cluster 2): People with fairly good general health, and poor physical health.
- **"Low Risk"** (cluster 3): Those with excellent general health but MAXVO2 (milliliters of oxygen consumed in a minute per kilogram of body weight) below 27.045. These are likely older adults.
- **"Lowest Risk"** (clusters 4 and 5): Individuals with excellent general health, MAXVO2 above 27.045, who engage in both 'high' and 'moderate' levels of physical activity. These are likely either younger or highly active adults.
- **"Highest Risk"** (cluster 6): People with poor general and physical health, suffering from obesity.

5 Challenges

Initially, we considered incorporating prescriptive analytics into our project in order to use obtain treatment strategies that are tailored to each patient cluster. However, we faced some challenges with this approach. First, we struggled to find a suitable dataset with diabetes treatment strategies that we could confidently map onto the BRFSS survey data which was used in our prediction and clustering models. Furthermore, the guidelines for treatment in preventative and early to moderate-stage cases are largely lifestyle interventions. Prescribing behavioral changes of this type, which include 'intentional caloric deprivation', 'structured exercise program', and 'good sleep hygiene' are not clearly defined and not easily enforced or adhered to [5]. This introduces a high level of uncertainty in the outcomes corresponding to these treatments, which makes trying to accurately assess outcomes for prescriptive models challenging. Despite not implementing prescriptive methods in our current work, this would be a natural extension of the project for future work.

6 Discussion and Conclusion

In conclusion, this two-phased analysis provides a comprehensive and methodical approach towards understanding and categorizing the risk of type 2 diabetes. We utilized an extensive data set from the Behavioral Risk Factor Surveillance System (BRFSS), applying data preprocessing techniques like optimal k-NN imputation and feature selection to refine their data.

The project's methodology entails a variety of tree-based models, including CART, Random Forest, XGBoost, and Optimal Classification Tree, to classify diabetes risk. The models demonstrated good performances, with an average AUC of around 0.8 for in-sample classifications and above 0.7 for out-of-sample classifications. This high level of accuracy underscores the effectiveness of their feature selection and data preprocessing methods. A pivotal aspect of the study is the use of model ensembling with Optimal Policy Trees. This technique not only addressed the limitations of individual models but also enhanced prediction accuracy, as evidenced by the improved out-of-sample AUC of 0.7892.

The second stage of the analysis – the weighted interpretable clustering – offers an insightful way to understand the diversity in diabetes risk. Through this process, we provided a holistic view of the population's risk profiles, as well as an interpretation of what are the factors underlying these diverse risk exposures.

In summary, our project provides significant insights into diabetes risk assessment, and paves the way for more personalized and effective healthcare interventions. Specifically, it provides a framework for earlier detection of diabetes and enables healthcare providers to focus their resources on education, intervention, and treatment of higher risk patient groups. In combination with potential extensions to prescriptive analytics, this work can help create specific and actionable treatment plans to meet the needs of each patient.

7 Contributions

While we both worked together throughout the project, the contributions were mostly split as follows: Natalie performed data preprocessing, including exploratory data analysis, missing data imputation, and feature selection. Sara implemented the CART, RF, XGBoost, and OCT classification models as well as the ensemble approach. We both contributed to the clustering analysis by assessing different clustering methods and constructing OCTs to interpret the clusters.

8 References

References

[1] *Diabetes*. URL: https://www.cdc.gov/diabetes/index.html.

[2] *Behavioral Risk Factor Surveillance System (BRFSS)*. URL: https://www.cdc.gov/brfss/index.html.

[3] Bertsimas, D., Pawlowski, C., Zhuo, Y. D. From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 2017.

[4] Bertsimas, D., Pauphilet, J., and Van Parys, B. Sparse classification and phase transitions: A discrete optimization perspective. *arXiv preprint arXiv:1710.01352*, 2017.

[5] Samson, S., et. al. American Association of Clinical Endocrinology Consensus Statement: Comprehensive Type 3 Diabetes Management Algorithm - 2023 Update. *Endocrine Practice*, 2023.