SCHOOL OF OPERATIONS RESEARCH
AND INFORMATION ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

HEALTHCARE ANALYTICS: PREDICTING DIABETES RISK FROM PUBLIC HEALTH DATA

FINAL REPORT

Chun Wang, Jiahao Li, Karen Kuang

ORIE 5741

Github link: https://github.com/karenSashimi/predicting-diabetes

2024 Spring

CONTENTS

1 INTRODUCTION

1.1 BACKGROUND INFORMATION

Diabetes is one of the widespread challenges that impacts millions of lives across the United States. As a potentially life threatening disease, diabetes affects how our bodies handle sugar, leading to a range of serious health issues, from heart disease to heart problems(1). Even more, CDC defined prediabetes as blood sugar level is higher than normal but not yet to be diagnosed as Type 2 diabetes, and it will raise the risk of patients having diabetes, hear attack and stroke. Alarmingly, according to CDC, more than 1 third of the population have prediabetes(1). Recognizing and addressing this disease in the early stage could significantly improve the quality of life for the patients, making this task more urgent and necessary.

1.2 PROBLEM STATEMENT

This project aims to identify individuals at high risk of developing diabetes based on a comprehensive analysis of health indicators, lifestyle factors, and the broader context of their living environments. Through the analysis and constructing predictive modeling, we seek to bring a personalized approach to diabetes prevention, and ensure early detection and awareness, making it possible to offer advice and interventions.

1.3 DATA SOURCE AND STRUCTURE

We obtained access to the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset(2), comprising 464,644 subjects and encompassing 279 variables. The BRFSS was established in 1984, initially involving 15 states conducting monthly telephone interviews to gather surveillance data on risk behaviors. BRFSS collects consistent, state-specific information on preventive health practices and behaviors associated with chronic diseases, injuries, and preventable infectious diseases affecting the adult population.

Within this dataset, 61,118 respondents were diagnosed with diabetes, 12,699 with prediabetes, and 390,827 had neither condition. The majority of variables were related to various chronic health conditions, such as cancer and asthma, aside from diabetes. Type 2 diabetes classification was based on age (>30 years), non-pregnancy status, and affirmative response to the survey questions.

The following Table 1 shows the detailed information about the dataset.

Feature Name	Description
Diabetes_012	0: No diabetes, 1: Prediabetes, 2: Diabetes.
HighBP	High Blood Pressure (1: Yes, 0: No).
HighChol	High Cholesterol (1: Yes, 0: No).
CholCheck	Cholesterol Checked within the last 5 years (1: Yes, 0: No).
BMI	Body Mass Index of the respondent.
Smoker	Smoking status (1: Yes, 0: No).
Stroke	Has had a Stroke (1: Yes, 0: No).
HeartDiseaseorAttack	Has had Heart Disease or a Heart Attack (1: Yes, 0: No).
PhysActivity	Exercised in the past 30 days (1: Yes, 0: No).
Fruits	Daily fruit consumption (1: Yes, 0: No).
Veggies	Daily vegetable consumption (1: Yes, 0: No).
HvyAlcoholConsump	Heavy alcohol consumption (1: Yes, 0: No).
AnyHealthcare	Access to healthcare (1: Yes, 0: No).
NoDocbcCost	Cost prevented seeing a doctor (1: Yes, 0: No).
GenHlth	General health status, rated from 1 (excellent) to 5 (poor).
MentHlth	Number of Days with poor mental health in the past 30 days.
PhysHlth	Number of Days with poor physical health in the past 30 days
DiffWalk	Difficulty walking or climbing stairs (1: Yes, 0: No).
Sex	0: Female, 1: Male.
Age	Encoded as an integer from 1 to 13.
Education	Education level encoded from 1 (no schooling) to 6 (college graduate).
Income	from 1 (lowest) to 8 (highest).

Table 1: Description of Data Features

2 DATA EXPLORING AND PREPROCESSING

The data size is 253,680 data points with 22 features where *diabetes_012* is the target variable. The dataset we used is a clean dataset of survey responds to the 2015 BRFSS. There are no missing values. 23899 records have the exactly same response, while it is reasonable to have individuals with the same background, and we decided to keep those data points.

2.1 FEATURES' DISTRIBUTION

We first analyzed the distribution for the target variable as shown in Figure 1, where most of the data are classified as *No Diabetes*. With such unbalanced data, it is crucial to carefully tuning the models to correctly control type 1 error and type 2 error.

As the next step, we analyzed the distribution of all independent features shown in Figure 2. All the categorical features are binary categorized and therefore no need to transform. *Physlth* and *Menthlth* representing physical health and mental health show clearly skewed distributions where most of the participants respond with 0, meaning no health issue in the last 30 days for the corresponding field. For the binary features, some of them are not evenly distributed between two classes such as *Stroke*. Since this is a real life data, such imbalance is expected and we'll handle that in the feature engineering step.

2.2 Correlation Analysis

Correlation indicates the relationship between two features, helping the model to avoid overemphasis on a single area. Heat Map is constructed as shown in Figure 3 to visually understand the dataset. In summary, there's no serious correlation issue in the data, and no columns are dropped due to collinearity. General health have slight positive correlation with mental health and physical health, which is as expected. We do think separating general health into two categories could better help us to understand the importance between physical health and mental health. Another interesting finding is that General Health is slightly negatively correlated with income level, suggesting that higher earnings are associated with marginally poorer health. The validity of such finding could be a next step for this study.

As a first step of feature importance analysis, the correlation between dependent features and target variable is visualized in Figure 4. General Health shows the highest positive relation and Income shows the highest negative relation. Meanwhile, Healthcare conditions, sex, and eating habits are not the main influencers. This importance of features will be examined again by model selection.

2.3 Feature Engineering

The chi-squared (chi2) statistical test is employed to assess the relevance of features for classification. The chi-squared test statistics is calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

with O_i denoting the observed values (feature) and E_i denoting the expected values (diabetes stage). It measures the linear dependency between features and the target variable, facilitating the identification of irrelevant features. With the computed scores, the top 10 features (k=10) were selected for further analysis. Notably, among these features, variables such as blood pressure, high cholesterol levels, body mass index (BMI), history of heart disease, and self-reported general health emerged as the most relevant factors influencing the stage of diabetes. Conversely, features such as vegetable and fruit intake, gender, frequency of cholesterol checks, and engagement with healthcare services exhibited the lowest relevance. These findings align with existing medical literature, underscoring the significance of physiological indicators and health status in predicting health-related outcomes. By utilizing this method, we refine the feature set, enhancing the classification model's predictive accuracy and reliability.

In the next step, standardization was implemented on the feature columns to mitigate the risk of large-scale variables and the impact of outliers. Features were standardized by removing their mean and scaling them to unit variance. This normalization procedure enhances the comparability of feature magnitudes and thereby fostering the robustness of the classification model against potential biases and disproportional data .

Furthermore, to address the issue of class imbalance within the training dataset, the NearMiss algorithm was incorporated to dataset balancing. This imbalanced-learning library transforms the dataset and facilitates the undersampling of the majority class, thereby rebalancing the class distribution and mitigating the potential biases introduced by disproportionate representation. This strategic approach enhances the overall reliability and generalizability of the classification framework.

3 LINEAR CLASSIFIERS

Multiple linear classifiers were implemented to predict the stage of diabetes, aiming to capture the relationship between health-related and lifestyle variables and the development of diabetes. Linear classifiers operate under the assumption of linear separability, making them particularly adept at modeling linear relationships within the data. Leveraging this characteristic, we directly applied the preprocessed dataset to fit the linear classifier models, including Logistic Regression, Regularized Support Vector Machines (SVM), and Ridge Classifier. All models were trained using default hyperparameters to maintain consistency across the experimentation process. The model performances were assessed using 5-fold cross validation, and thereby using precision, recall, and f1-score as the performance indicators. The following table provides a comprehensive overview of the performance of the linear classifier models 2:

Model	**precision**	**recall**	**f1-score**
Ridge Classifier	0.82	0.8	0.8
Regularized SVM (SGD)	0.86	0.84	**0.84**
Logistic Regression	0.85	0.85	**0.84**

Table 2: Performance of Linear Classifiers

Upon analysis, it was evident that Logistic Regression and Regularized Support Vector Machine (SVM) are as the top-performing classifiers. To delve deeper into their efficacy, we conducted a comprehensive investigation into feature importance within these models. For linear classifiers, feature importance is derived from the coefficients assigned to each feature. Features with larger coefficients exert a greater influence in determining the likelihood of an individual being at risk of diabetes. The results are depicted in Figure C. Notably, the feature importance plot elucidates that variables pertaining to physical and mental health conditions, income status, and body mass index (BMI) hold paramount significance in predicting diabetes risk. This underscores the critical role of holistic health indicators and socioeconomic factors in delineating susceptibility to diabetes.

4 Tree Classifiers

In this section, we extended our models to tree based methods which can capture nonlinear relationship between the features and the objective variable. As decision trees are not sensible to the scale of the variables, we directly used the preprocessed dataset to fit the models. We implemented Decision Tree, Random Forest, AdaBoost, Gradient Boosting Tree and XGBoost and chose accuracy as the performance metrics. All models are built with default hyperparameters. The validation and test accuracy is summarized the the table 3.

Model	**Val. Accuracy (%)**	**Test Accuracy (%)**
Decision Tree	76.53	77.12
Random Forest	83.91	84.42
AdaBoost	84.90	84.98
Gradient Boosting	84.87	**85.30**
XGBoost	84.79	**85.30**

Table 3: Performance of Tree Classifiers

We observed that Gradient Boosting and XGBoost achieved highest testing accuracy. We further investigated the feature importance by calculating the average impurity decrease for each split in these two models to understand which features are crucial to determine whether a person is at high risk of having the diabetes. The feature importance result is shown in Fig C. The feature importance plot indicates that blood pressure, general health condition, cholesterol, BMI and age are most important features to determine diabetes.

5 Conclusion

Our exploration of linear and tree-based classification models has provided valuable insights into the complex interplay between health and lifestyle indicators and the development of diabetes. Linear classifiers, predicated on the assumption of a linear relationship between features and the target, shed light on the importance of physical and mental health conditions, income status, and BMI as a significant predictors of diabetes risk. Conversely, tree-based models, adept at capturing nonlinear relationships, reveal a nuanced understanding of feature importance in diabetes prediction. While mental health and income emerges as significant feature in linear models, tree-based approaches recognized them as insignificant and unveils a broader spectrum of investigation features, including blood pressure, cholesterol levels and age. Lifestyle factors such as fruit and vegetable intake, as well as the frequency of health checkups, are deemed of lesser importance in hierarchy of predictors by both predictors. In essence, the divergent conclusions drawn by linear classifiers and tree classifiers regarding the feature importance underscore the necessity of employing complementary methodologies to glean a comprehensive understanding of the multifaceted factors influencing health outcomes.

The implications of these findings extend beyond theoretical understanding to practical applications in diabetes prevention and management. Addressing factors related to general health and body composition emerges as a pivotal intervention strategy, with targeted efforts towards managing blood pressure, cholesterol levels, and promoting overall well-being. While lifestyle factors such as fruit and vegetable intake and the frequency of health checkups are crucial to overall health, our analysis suggests that their direct impact on diabetes risk may be minimal in comparison to physiological indicators and body composition. By prioritizing interventions that address the fundamental determinants of diabetes, such as proactive control of health body composition, we can pave the way towards reducing the burden of diabetes on individuals and healthcare systems alike.

6 Fairness And WMD

Fairness in our study is crucial. Ensuring the predictive model does not increase current healthcare disparities is one of our goal. From the trained model, the most important features are mainly related to the health condition of an individual instead of personal background of the individual like sex. We remain vigilant in continuously validating the model across diverse populations to address any inadvertent biases and ensure the fairness of our insights.

Our model can be a potential WMD if not implemented with transparency and ethical oversight. The potential harm could be ranging from unnecessary anxiety to a lack of intervention for the disease. It is important that we can keep the data transparent while respecting participants' privacy. Since the model is focusing on the current condition of an individual and predicting the chance once, there's no feedback loops for our model.

In general, while the current model does not compromise fairness or qualify as a WMD, we acknowledge the potential for such issues. Monitoring the model and Making adjustments if any unfairness or negative impacts are identified is crucial for the future.

References

[1] Centers for Disease Control and Prevention, “What is diabetes?” https://www.cdc.gov/diabetes/basics/diabetes.html, 2023, accessed: 2024-05-08.

[2] ——, “2015 brfss annual data,” 2015, accessed: 2024-05-08. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2014.html

A Distribution for features

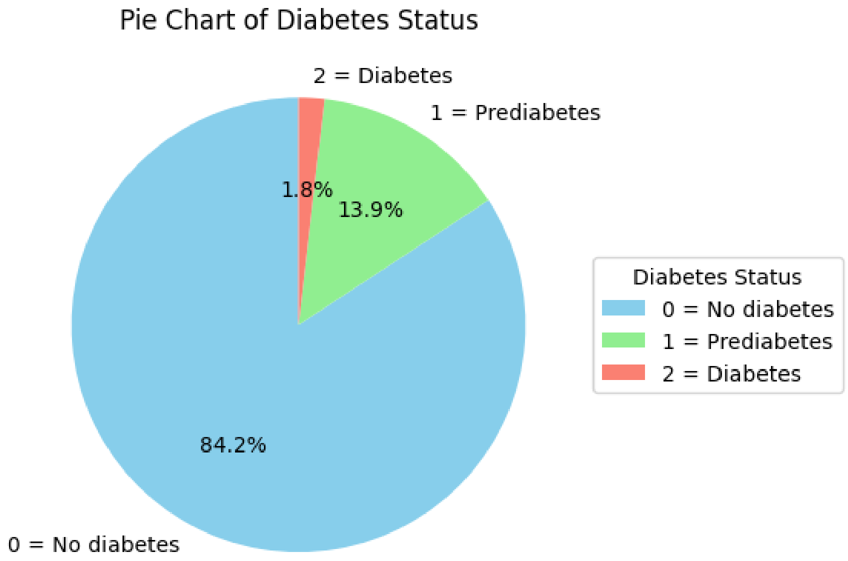

Figure 1: Distribution of target variable

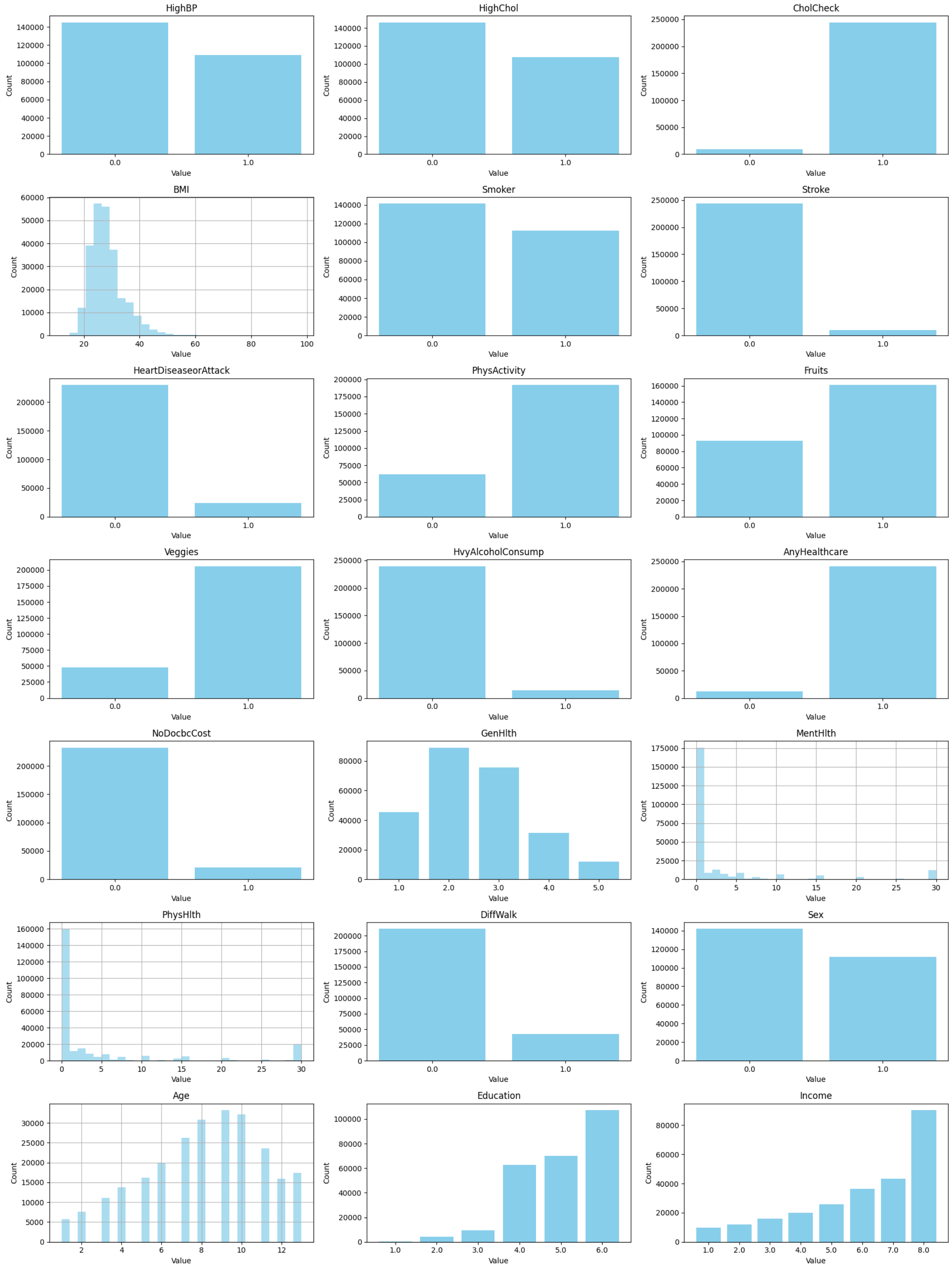

Figure 2: Distribution of All Features

B Correlation Analysis

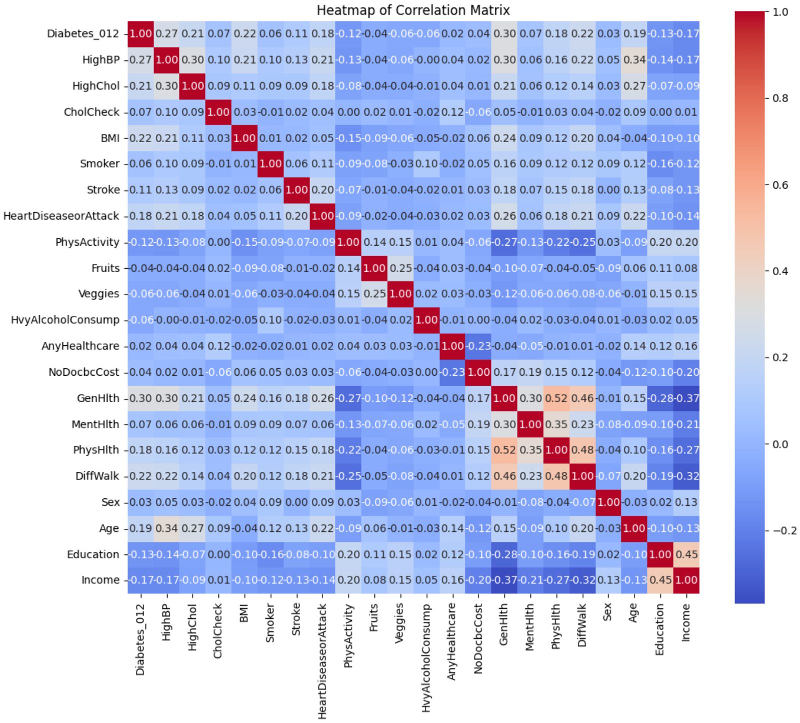

Figure 3: Correlation Matrix

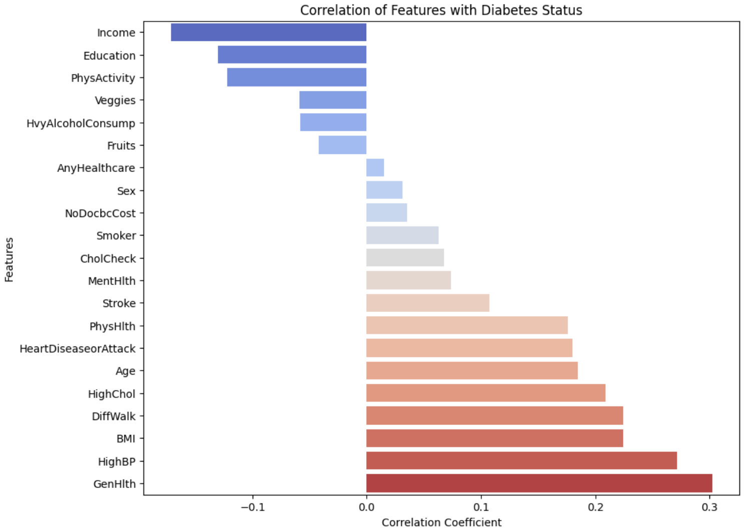

Figure 4: Correlation with the target variable

C Model Performance

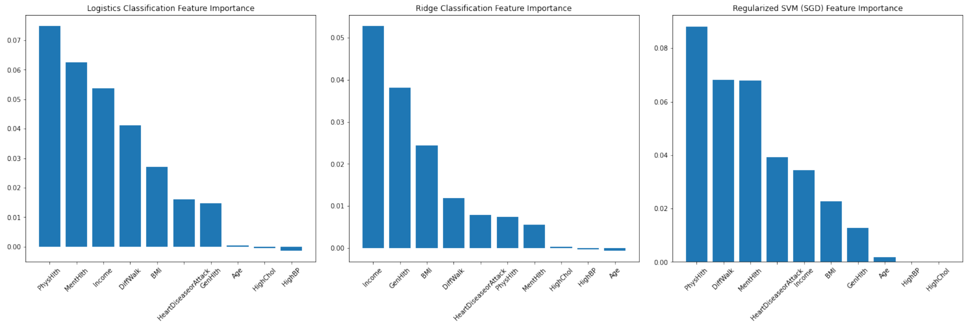

Figure 5: Feature importance for Logistics Classifier, Ridge Classifier and Regularized Support Vector Machine by ranking feature coefficients within the model.

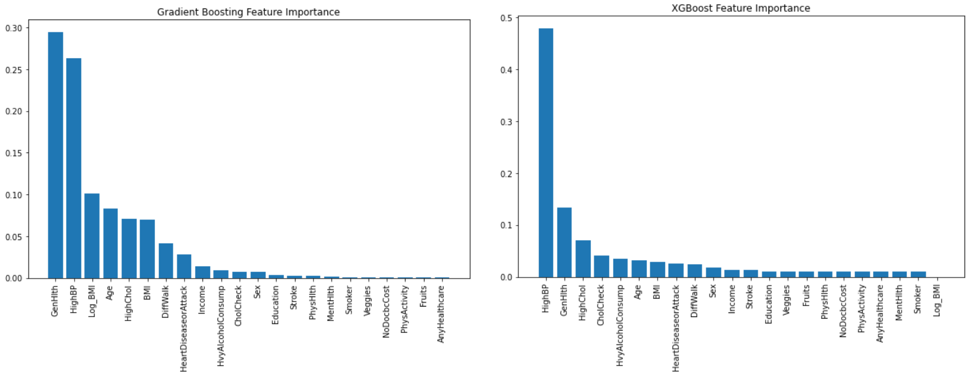

Figure 6: Feature importance for Gradient Boosting and XGBoost Model by calculating average impurity decrease in each split within the model.