Predicting Diabetes with Accessible Parameters

STOR 520

Yun Ma, 730476815

Ziyin Zheng , 730703348

December 2, 2024

Introduction

Diabetes, particularly type 2 diabetes, is a major global health challenge that affects millions of people worldwide, placing a heavy burden on both individuals and healthcare systems. To tackle this challenge effectively, we need to ask innovative questions that go beyond the surface, delving into root causes and effective early intervention strategies. Using the CDC Diabetes Health Indicators dataset, we focus on two key questions aimed at deepening our understanding of type 2 diabetes and ultimately making a tangible impact on people's lives.

The first question we address is: What key features most influenced the detection of type 2 diabetes? Identifying the most important risk factors is crucial because it allows for more targeted prevention and intervention strategies. By analyzing features such as BMI, physical activity, diet, socioeconomic conditions, and access to healthcare, we can determine the most critical predictors. These insights can inform public health initiatives and empower individuals to make informed lifestyle changes to reduce their risk. For example, understanding the impact of access to healthy food and exercise opportunities can lead to community-based programs aimed at creating healthier environments and reducing diabetes rates at the local level.

The second question explores How to train and evaluate a model to effectively identify patients with early detection of type 2 diabetes for timely treatment? Early detection is critical because catching type 2 diabetes in its early stages allows for effective management or even potential reversal through lifestyle changes and prompt medical intervention. By employing predictive modeling, we can develop tools that identify at-risk individuals using key health indicators such as physical activity, and demographic data. The objective is not only to create accurate models but also to ensure that these tools are accessible, especially for populations most in need. This means integrating predictive models into community healthcare settings and ensuring equitable distribution of early intervention resources, particularly in underserved areas with limited healthcare access.

Together, by identifying the key predictors of diabetes and understanding how lifestyle and environmental factors contribute, we can empower individuals and communities to make healthier choices. Early detection models ensure that those at risk are identified and supported before complications develop. focus on both prevention and early detection, our study aims to lessen the impact of diabetes on individuals and alleviate the burden on healthcare systems, ultimately contributing meaningfully to public health.

Data

The dataset analyzed in this study is from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) conducted by the Centers for Disease Control and Prevention (CDC).The BRFSS is a large-scale public health survey program that collects data from all 50 U.S. states, Puerto Rico, the U.S. Virgin Islands, and Guam through monthly surveys. The 2015 dataset provides self-reported data, which are collected to monitor the health status and risk behaviors of the population. BRFSS data use rigorous data collection methods and provide a unique opportunity to investigate public health patterns. It is a trusted resource in the field of healthcare research and policy development.

Each observation in the dataset corresponds to an individual survey respondent, yielding a large number of health indicators. These include biometrics such as BMI (body mass index), lifestyle variables such as frequency of physical activity and dietary patterns, as well as key demographic details such as age, gender, ethnicity and socioeconomic status (Table 1).

```
Data columns (total 23 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  category
 1   HighBP                253680 non-null  category
 2   HighChol              253680 non-null  category
 3   CholCheck             253680 non-null  category
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  category
 6   Stroke                253680 non-null  category
 7   HeartDiseaseorAttack  253680 non-null  category
 8   PhysActivity          253680 non-null  category
 9   Fruits                253680 non-null  category
 10  Veggies               253680 non-null  category
 11  HvyAlcoholConsump     253680 non-null  category
 12  AnyHealthcare         253680 non-null  category
 13  NoDocbcCost           253680 non-null  category
 14  GenHlth               253680 non-null  category
 15  MentHlth              253680 non-null  category
 16  PhysHlth              253680 non-null  category
 17  DiffWalk              253680 non-null  category
 18  Sex                   253680 non-null  category
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  category
 21  Income                253680 non-null  category
 22  Diabetes_binary       253680 non-null  category
```

Table 1. Description of Features

The main variables of interest in this study include body mass index, physical activity levels, dietary patterns, and health care delivery indexes, as well as demographic variables such as age, gender, race, and socioeconomic status. The total sample size of the dataset is approximately 250,000 respondents, These variables are important for understanding the correlations between lifestyle choices, health outcomes, and demographic characteristics (Table 1).

A summary of key statistics from the dataset will include descriptive indicators such as mean, median, standard deviation, and range for variables such as body mass index, physical activity level and age. For example, the mean body mass index (BMI) in the dataset is

approximately 28.4 with a standard deviation of 5.6, indicating that there are differences in health indicators related to weight and height across the population (Figure 1). Similarly, the median age of respondents was 45 years, reflecting a broad age distribution (Figure 2). In addition, bar plots showing the presence or absence of diabetes in relation to age can reveal trends in diabetes risk in different age groups.

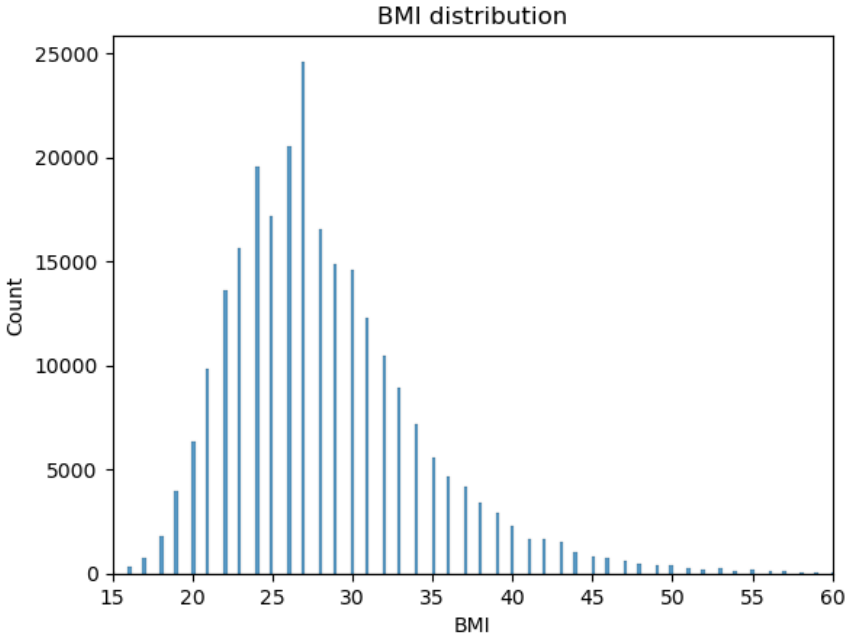

Figure 1. BMI Distribution

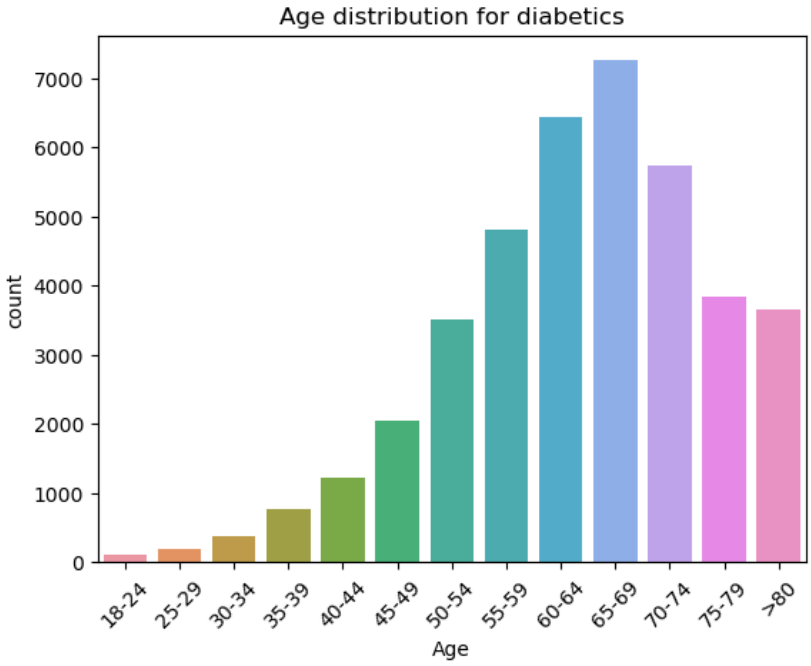

Figure 2. Age Distribution for Diabetics

Exploratory Data Analysis

Relationship Between Predictors and Response

The relationship between diabetes and income/education demonstrates a significant link between socioeconomic status and health outcomes (Figure 3). Individuals with lower income and education levels exhibit a noticeably higher prevalence of diabetes compared to those with higher income and education. For instance, individuals in the highest income brackets (e.g., levels 7 and 8) and with the highest education levels (e.g., levels 5 and 6) experience the lowest prevalence of diabetes, ranging from 0.08 to 0.15. In contrast, those in the lowest income and education levels face the highest prevalence, reaching approximately 0.40. These disparities highlight the critical role socioeconomic factors play in influencing diabetes risk.

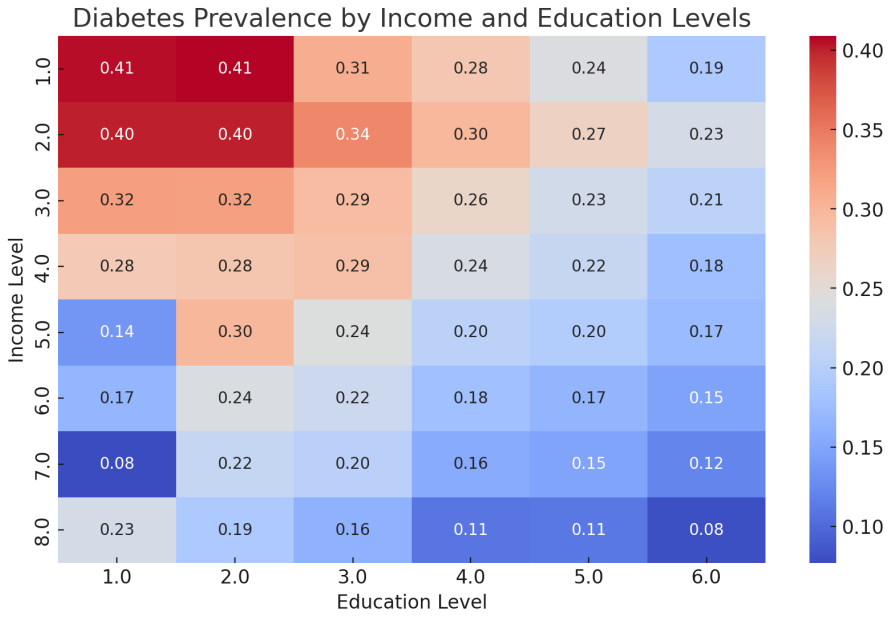

Figure 3. Diabetes Prevalence by Income and Education Levels

Lower-income and less-educated populations may encounter significant barriers to accessing healthcare resources, healthy food options, and opportunities for preventive care. These obstacles contribute to the elevated diabetes prevalence observed in these groups. On the other hand, individuals with higher income and education levels often benefit from greater access to health knowledge, resources, and preventive measures, which help reduce their diabetes risk. Addressing these disparities requires targeted interventions to improve health education and resource availability, particularly for vulnerable populations. This underscores the importance of addressing social determinants of health to mitigate diabetes risk.

BMI has a profound impact on diabetes prevalence, as evidenced by the dataset (Figure 3). Individuals with higher BMI levels are more likely to develop diabetes, with the prevalence steadily increasing as BMI moves into overweight and obese ranges. This strong correlation highlights excess body weight as a critical risk factor for diabetes. Furthermore, diabetics tend to have a higher median BMI compared to non-diabetics, reinforcing this relationship. These findings suggest that weight management and obesity prevention are essential strategies for reducing diabetes risk.

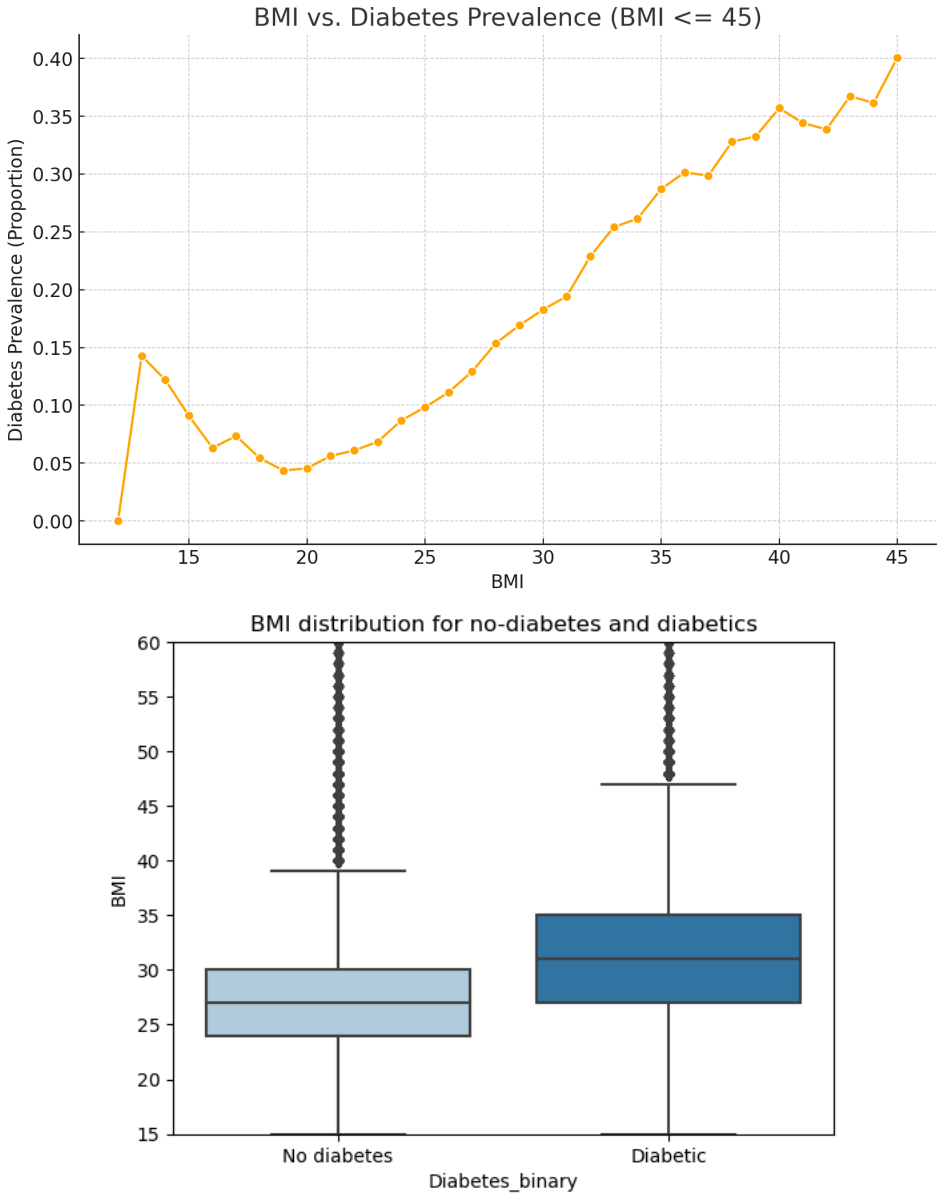

Figure 4. BMI and Diabetes Prevalence

Public health initiatives aimed at promoting healthy eating habits, regular physical activity, and awareness about the risks of high BMI could play a vital role in mitigating this growing concern. By addressing obesity as a central element in diabetes prevention programs, these efforts can help curb the rising prevalence of diabetes and improve overall health outcomes.

Lifestyle factors also show significant associations with diabetes prevalence (Figure 4). Individuals with high cholesterol (HighChol) and high blood pressure (HighBP) exhibit a notably higher prevalence of diabetes compared to those without these conditions. Similarly, smokers (Smoker) and heavy alcohol consumers (HvyAlcoholConsump) demonstrate higher rates of diabetes. Conversely, individuals engaging in regular physical activity (PhysActivity) have significantly lower rates of diabetes, while those reporting difficulty walking (DiffWalk) show a

higher prevalence. These patterns suggest that lifestyle choices may be closely linked to diabetes risk.

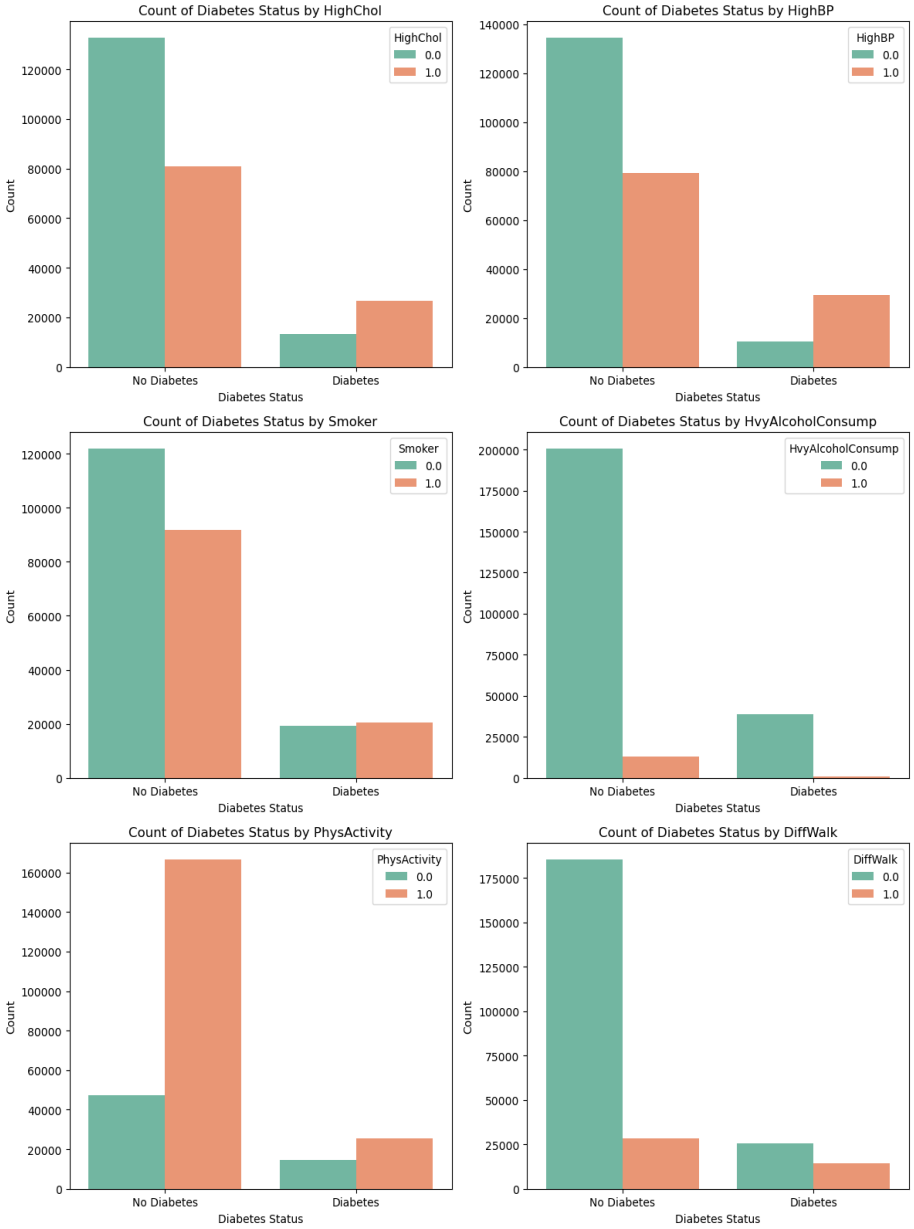

Figure 4. Various Lifestyle Parameters and Diabetes Prevalence

However, while these associations are clear, the data does not establish causation. High cholesterol and high blood pressure, for example, may either contribute to diabetes or result from it. Similarly, smoking and heavy alcohol consumption could exacerbate diabetes risk by impairing metabolic health, but other factors like diet, genetics, or socioeconomic status may also play a role. Regular physical activity, which correlates strongly with lower diabetes prevalence, underscores the importance of an active lifestyle. Yet, this relationship might also reflect overall better health among active individuals. The higher prevalence of diabetes in those with difficulty walking could be linked to reduced mobility or other co-occurring health issues.

These findings highlight the need for further research to investigate the causal mechanisms and interactions underlying these associations.

Understanding what key features most influenced the detection of type 2 diabetes is essential for better prevention and diagnosis strategies. By combining insights from demographic factors such as gender and age, physical attributes like BMI, and lifestyle factors including high blood pressure, high cholesterol, smoking, and physical activity, a more holistic view of diabetes risk can be achieved. These variables likely interact in complex ways, and analyzing them together can help uncover patterns and interactions that might not be apparent when studied individually. Machine learning models can be used to identify the relative importance of these features. Feature selection can quantify the contribution of each variable to the prediction of diabetes. This approach not only highlights which factors are most strongly associated with type 2 diabetes but also provides insights into how these factors interact. For instance, age and BMI may have a combined effect on diabetes risk that is different from their individual effects. Identifying these key features can inform targeted public health interventions and personalized treatment plans.

This analysis highlights the important association between diabetes and a variety of factors, including demographic, physical and lifestyle variables. The main patterns show that people with high cholesterol, high blood pressure and sedentary lifestyle are at greater risk of developing diabetes, while regular physical activity seems to reduce this risk. In addition, demographic factors such as age and gender, as well as physical characteristics such as body mass index, play an important role in understanding the prevalence of diabetes. However, while the relationship between these factors and diabetes is evident, the data alone cannot establish causality. A comprehensive approach to diabetes prevention and management is needed, combining lifestyle changes, regular health check-ups and interventions targeted at those at risk. By understanding and addressing the key drivers of type 2 diabetes, public health efforts can be more effective in reducing its prevalence and improving health outcomes.

PCA and Dimension Reduction

Principal Component Analysis (PCA) is a powerful technique for projecting high-dimensional data onto a lower-dimensional space, enabling initial examination of potential decision boundaries and data structure. Our PCA results indicate that the first three principal components (PCs) capture 90% of the variance (Figure 5), suggesting a strong underlying structure and potentially high redundancy in the data. However, visualizing the first two PCs reveals that the sample classes are extremely imbalanced, with no clear decision boundary in the two-dimensional space: the classes are highly embedded within each other (Figure 6). This observation suggests that the predictors in this dataset may lack robustness for accurately predicting diabetes.

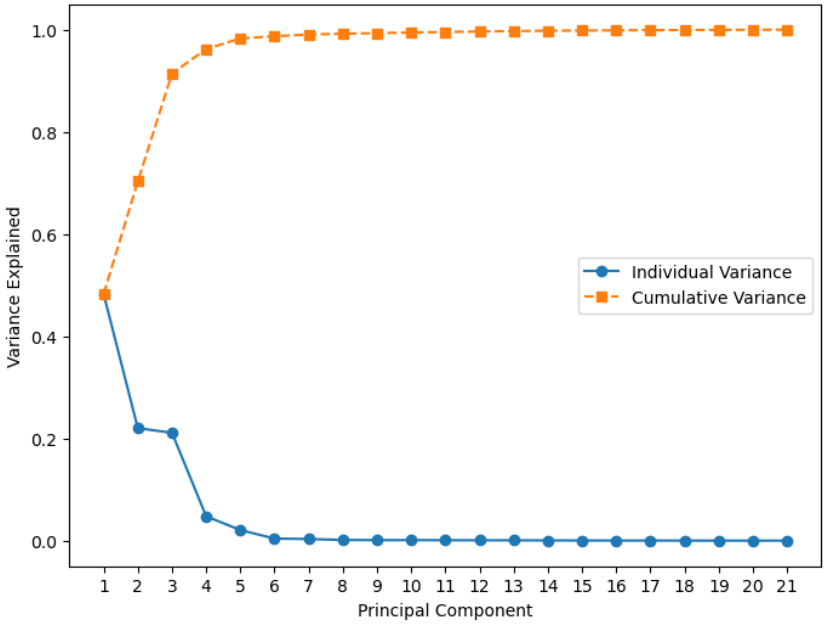

Figure 5. Elbow Plot

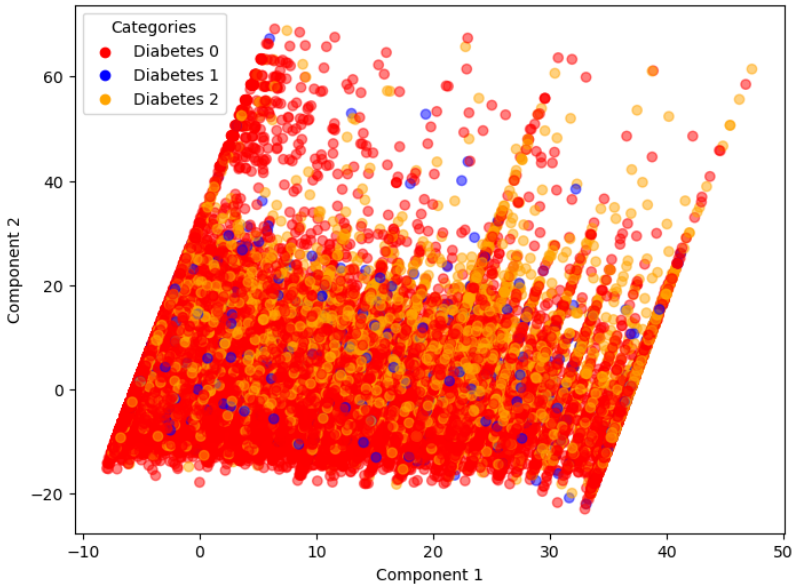

Figure 6. First 2 PCs of Multi-class Diabetes

Results

SMOTE and Class Imbalance

EDA revealed a significant class imbalance in our dataset, with Class 0 (healthy) comprising over 80% of the data (Figure 6, Table 2). This imbalance poses a serious concern: a model that classifies all patients as healthy achieves a baseline accuracy of 80%. While seemingly high, such a model is potentially harmful. Misclassifying diabetic patients as healthy during early diagnostics, such as in our model, cost them the optimal opportunity of receiving treatment, potentially worsening their outcomes. To address this issue, it is essential to correct the class imbalance before training any predictive model.

One effective approach is the Synthetic Minority Oversampling Technique (SMOTE). Unlike simple duplication of minority samples, SMOTE generates new synthetic samples by interpolating between an existing minority point and its k-nearest neighbors (Hoens, pp. 47). By applying SMOTE to our dataset, we successfully balanced the class sizes (Table 2), ensuring that the model is better equipped to identify diabetic patients without being biased toward the majority class.

	0	1	2
Original	213703	35346	4631
Train (No resample)	170908	28349	3687
Train (SMOTE)	170908	170908	170908
Test	42795	6997	944

Table 2. Sample Size of Each Class

Feature Selection

The detection of Type 2 diabetes requires careful consideration of multiple health indicators and demographic factors. Our feature selection process employed a comprehensive statistical framework to identify the most influential predictors while ensuring model stability and interpretability. Through systematic analysis combining multiple statistical approaches, we identified a set of key features that demonstrate significant predictive power for diabetes diagnosis.

Our initial step involved examining multicollinearity among numerical features using the Variance Inflation Factor (VIF) analysis. This critical assessment helped identify and eliminate redundant features that could potentially destabilize our predictive models. Features demonstrating VIF scores exceeding 10 were systematically removed from the dataset, ensuring that each retained feature contributed unique information to the diabetes detection process. This refinement process particularly impacted highly correlated health metrics, allowing us to maintain only the most informative indicators.

	Feature	VIF
19	Education	29.595848
2	CholCheck	23.188813
11	AnyHealthcare	20.846083
3	BMI	18.384195
20	Income	14.189437
13	GenHlth	10.849454
18	Age	9.908037
9	Veggies	5.828007
7	PhysActivity	4.647507
8	Fruits	3.032929
0	HighBP	2.329576
1	HighChol	2.047409
15	PhysHlth	1.999601
4	Smoker	1.933828
17	Sex	1.911345
16	DiffWalk	1.842262
14	MentHlth	1.463381
21	Diabetes_binary	1.421442
6	HeartDiseaseorAttack	1.294170
12	NoDocbcCost	1.216106
5	Stroke	1.127184
10	HvyAlcoholConsump	1.085138

Table 3. result of VIF analysis

Following the VIF analysis, we conducted Analysis of Variance (ANOVA) tests to evaluate the significance of mean differences across diabetes categories for our remaining features. The ANOVA results revealed substantial variations among diabetes groups, particularly in cardiovascular indicators and lifestyle factors. Features demonstrating strong F-statistics and low p-values were prioritized, as they showed clear differentiation capability between diabetes states.

	Feature	F-Statistic	P-Value
0	HighBP	1.014914e+04	0.000000e+00
1	HighChol	5.890843e+03	0.000000e+00
3	Stroke	1.475322e+03	0.000000e+00
4	HeartDiseaseorAttack	4.260879e+03	0.000000e+00
5	PhysActivity	1.923358e+03	0.000000e+00
11	PhysHlth	4.078700e+03	0.000000e+00
12	DiffWalk	6.727221e+03	0.000000e+00
14	Age	4.560441e+03	0.000000e+00
15	Diabetes_binary	inf	0.000000e+00
10	MentHlth	7.171174e+02	2.735179e-311
2	Smoker	5.072706e+02	1.363194e-220
7	Veggies	4.484959e+02	3.666426e-195
8	HvyAlcoholConsump	4.265869e+02	1.113041e-185
6	Fruits	2.275784e+02	1.788548e-99
9	NoDocbcCost	1.983483e+02	8.428089e-87
13	Sex	1.255480e+02	3.178224e-55

Table 4. result of ANOVA analysis

To complement our previous analyses, we implemented Chi-Square tests to evaluate relationships between categorical variables and diabetes status. After normalizing the data using MinMax scaling, we calculated chi-square statistics and corresponding p-values for each feature. This process revealed significant associations between diabetes status and various health

indicators, particularly in cardiovascular health markers and demographic factors. The results strengthened our feature selection by confirming strong statistical relationships between categorical variables and diabetes outcomes.

	Feature	Chi2 Statistic	P-Value
0	HighBP	10731.721009	0.000000e+00
16	DiffWalk	10627.556856	0.000000e+00
15	PhysHlth	4719.959441	0.000000e+00
13	GenHlth	4401.370425	0.000000e+00
6	HeartDiseaseorAttack	7468.339377	0.000000e+00
21	Diabetes_binary	213703.000000	0.000000e+00
1	HighChol	6483.776499	0.000000e+00
5	Stroke	2798.417025	0.000000e+00
18	Age	973.268766	4.543313e-212
7	PhysActivity	922.529401	4.734629e-201
20	Income	920.721528	1.169124e-200
14	MentHlth	820.248767	7.677916e-179
10	HvyAlcoholConsump	802.538572	5.382245e-175
4	Smoker	562.684715	6.524792e-123
3	BMI	398.378210	3.113658e-87
12	NoDocbcCost	362.740875	1.705374e-79
19	Education	211.763647	1.037787e-46
9	Veggies	168.560797	2.497400e-37
8	Fruits	166.174822	8.233717e-37
17	Sex	140.390490	3.270336e-31
2	CholCheck	43.816645	3.057289e-10
11	AnyHealthcare	3.381194	1.844094e-01

Table 5. result of Chi-Square tests

Through comprehensive statistical analyses integrating VIF, ANOVA, and Chi-Square tests, we identified twelve key features that significantly influence diabetes detection. These features span four critical categories: cardiovascular indicators (including High Blood Pressure, High Cholesterol, History of Stroke, and Heart Disease or Attack History), physical health metrics (encompassing Body Mass Index, Physical Activity Levels, and Difficulty Walking), general health status measures (comprising General Health Assessment, Mental Health Condition, and Physical Health Status), and demographic factors (specifically Age and Income Level).

Based on our comprehensive feature selection analysis, these twelve identified features will serve as the foundation for our predictive modeling approach. Moving forward, we will utilize these statistically validated indicators—ranging from cardiovascular metrics to demographic factors—to develop and evaluate various machine learning models for diabetes detection. Our modeling strategy will leverage the distinct predictive power of each feature while accounting for their interrelationships, aiming to create a robust and clinically applicable prediction system.

Classification Model

With the important features identified, we proceeded to train our classification models. To ensure both linear and non-linear decision boundaries were considered, we implemented three models: Random Forest, Naïve Bayes Classifier, and Logistic Regression.

The first model tested was Random Forest, this model has the advantage of handling non-linear decision boundaries. It achieved the highest accuracy among all models, with a testing accuracy of 78% (Figure 7). However, a closer examination of the confusion matrix revealed that

the model remained heavily biased toward healthy patients, even after addressing class imbalance in the training data using SMOTE. This bias indicates that the Random Forest model is not ideal, as it tends to misclassify diabetic or pre-diabetic patients as healthy. Such misdiagnoses are particularly concerning given the potential consequences of delayed treatment for these patients. Despite this limitation, the feature selection capabilities of the Random Forest model proved valuable. It allowed us to identify key features for further analysis, including High Blood Pressure, High Cholesterol, and General Health (Figure 8).

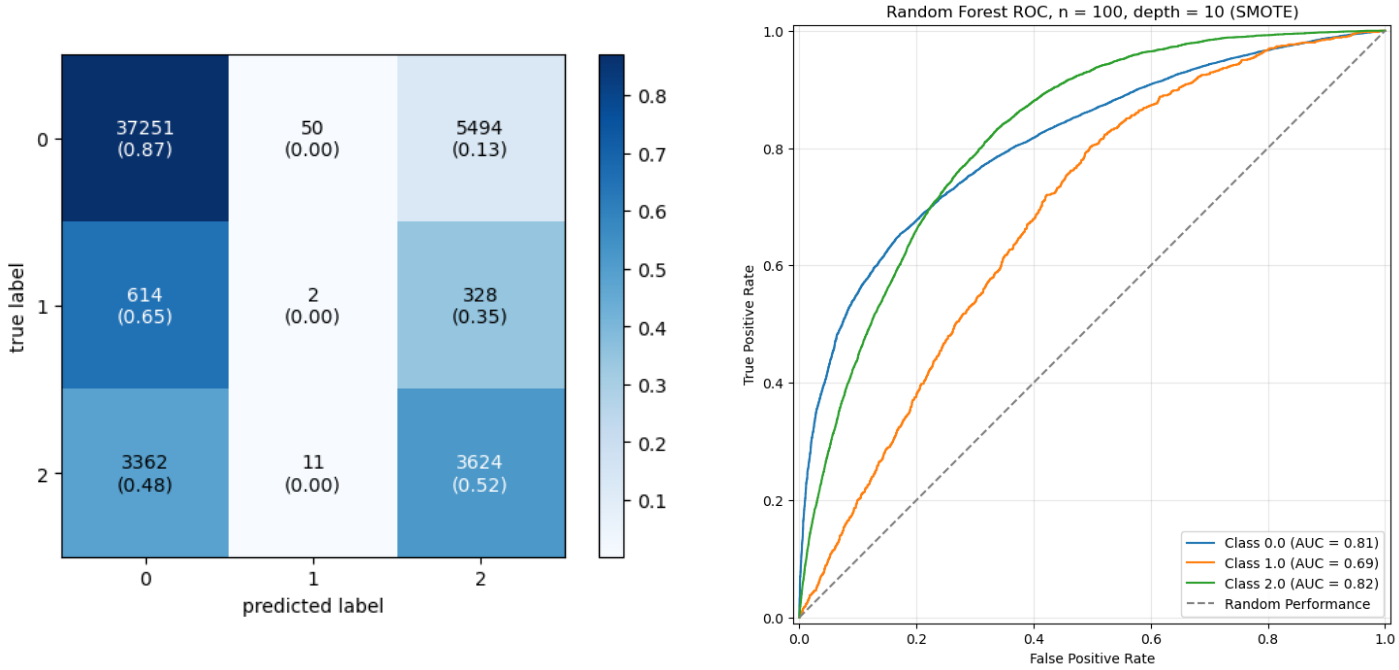

Figure 7. Test Performance of Random Forest Model

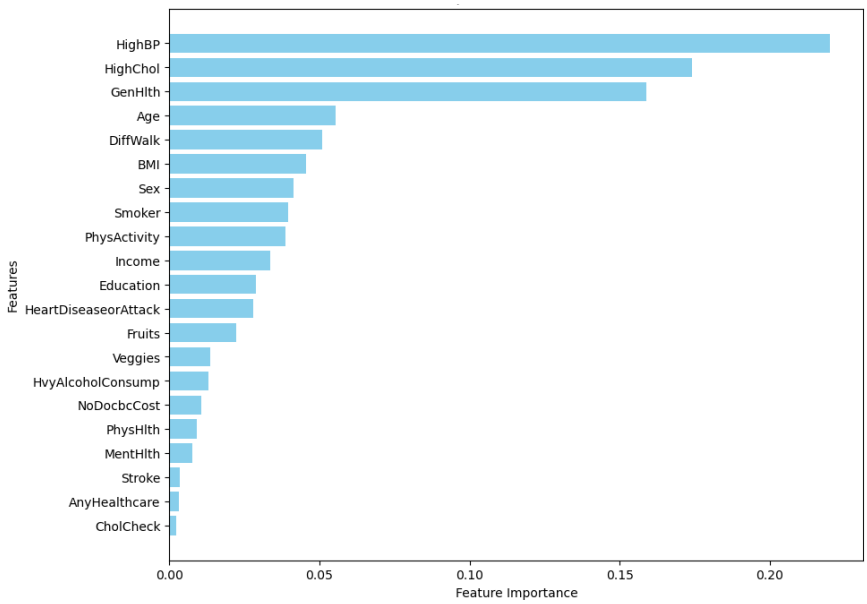

Figure 8. Feature importance in Random Forest Model

We next fitted a Naïve Bayes model, motivated by the observed low correlation between the previously identified important features (Figure 9). While low correlation does not confirm independence, it suggests the features may be sufficiently distinct for a model that assumes conditional independence, such as Naïve Bayes. This model is closest to the ideal Bayes classifier when the independence assumption is met and provides a distinct approach to classification compared to Random Forest. Given the absence of clear decision boundaries in our exploratory analysis (Figure 6), evaluating Naïve Bayes was a logical next step. Our Naïve Bayes model underperformed compared to Random Forest, achieving a test accuracy of 63% (Figure 10). However, analysis of the confusion matrix revealed a different type of bias: the model tended to classify pre-diabetic patients (Class 1) as diabetic (Class 2). While this misclassification is not ideal, it is less concerning than Random Forest's bias toward healthy classifications, as patients misdiagnosed with diabetes are more likely to seek further medical attention, leading to correct identification of their pre-diabetic status. Importantly, this bias does not pose an immediate threat to patient health.

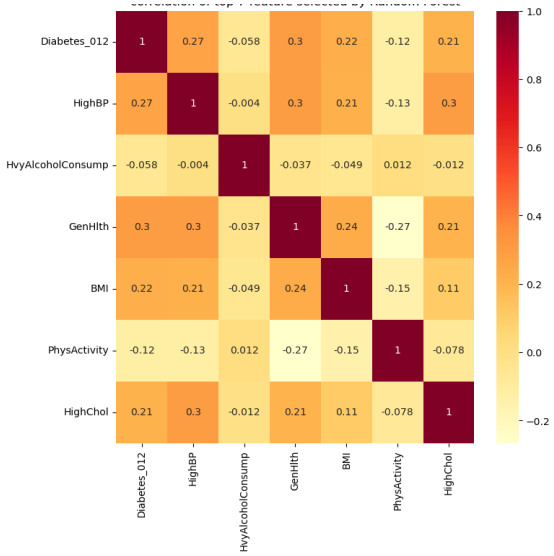

Figure 9. Correlation Among Naive Bayes Predictors

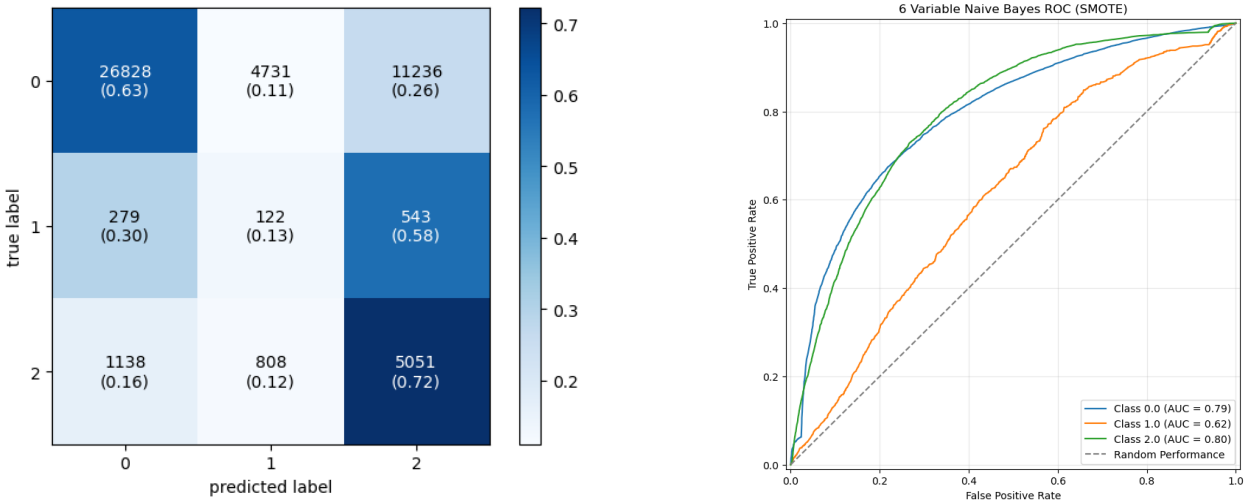

Figure 10. Test Performance of Naive Bayes Model

After evaluating the performance of two non-linear classifiers, we proceeded to assess a linear classifier by fitting a logistic regression model. The logistic regression model demonstrated similar overall performance to the Naïve Bayes classifier, achieving a test accuracy of 64%. However, unlike the non-linear models, it exhibited balanced prediction accuracy across all classes (Figure 11), which is a more favorable outcome.

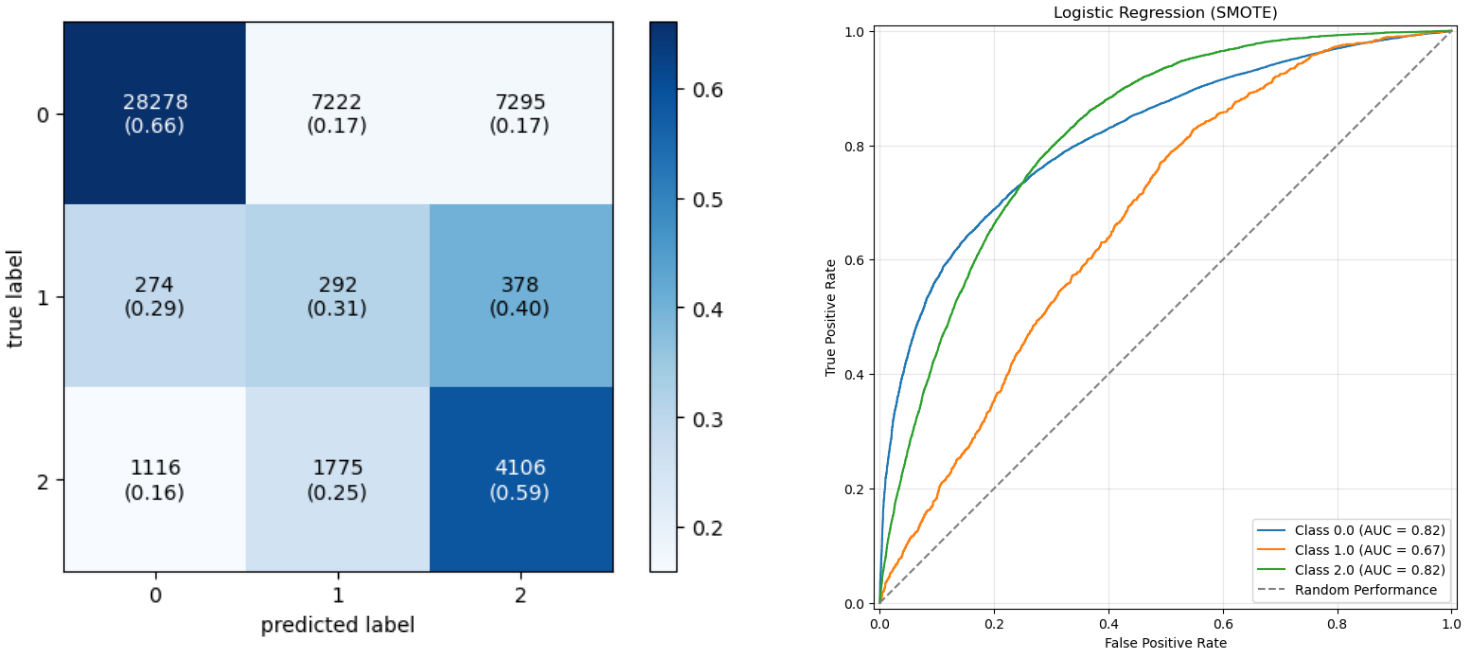

Figure 11. Test Performance of Logistic Regression Model

A comparison of the one-vs.-rest ROC curves for all three models revealed that all models struggled to accurately predict prediabetes (Class 1) (Figures 5, 8, and 9). Although combining prediabetes (Class 1) and diabetes (Class 2) into a single category significantly improved the accuracy score, we opted to retain the multi-class predictor due to its practical significance. Identifying the onset of diabetes, which is sometimes reversible, is critically important compared to classifying diabetes itself, which is generally irreversible. Developing a model that accurately detects diabetic onset remains a meaningful step toward preventing and treating diabetes.

Conclusion

Summary and Practical Application

In this study, we developed and evaluated three models—Random Forest, Naïve Bayes, and Logistic Regression—to predict diabetes (healthy, diabetic, and prediabetic) using 21 predictors across 253,680 samples. Our analysis identified variables such as Blood Pressure, Cholesterol, BMI, and General Health Index as significant contributors to diabetes diagnostics. These findings align with existing medical knowledge, underscoring the interpretability of our models. Further, they offer valuable insights for future medical studies to prioritize these factors, as our data suggest they are among the most relevant for diabetes diagnostics.

Among our models, Random Forest achieved the highest overall accuracy but was strongly biased toward predicting patients as healthy, even after addressing class imbalance through oversampling. This bias is particularly concerning, as it could lead to the misdiagnosis of diabetic patients as healthy, delaying their treatment and threatening their health outcomes. The Naïve Bayes model, while less accurate overall, tended to predict patients as diabetic. Although this bias is not ideal, it is less harmful, as it would likely prompt further medical attention and enable accurate diagnosis through additional testing. Logistic Regression, though slightly less accurate than Random Forest, achieved the most balanced prediction accuracy across all classes. Given this balance, Logistic Regression emerged as the most suitable model for this dataset, suggesting that the underlying structure of the data may be linear.

On the other hand, while combining the diabetes and prediabetes classes significantly improved prediction accuracy, we retained the multi-class model due to its greater clinical relevance. The inherent difficulty in distinguishing prediabetes from diabetes likely contributed to the lower prediction accuracy. Additionally, PCA analysis suggested that the predictors used may not be robust enough for precise diabetes classification. However, these predictors have the advantage of being easily accessible, as all 21 predictors can be self-reported without requiring advanced clinical tools. This makes our model a practical first-line predictor, helping individuals assess their risk and promoting universal, straightforward diabetes awareness and testing.

Future Directions

Future studies can address the class imbalance issue more effectively. Our current approach used oversampling of the minority class by interpolating between existing minority data points. However, since PCA revealed significant embedding of the classes, this method may also amplify background noise. An alternative approach could involve down-sampling the majority class, but this approach risks losing substantial amounts of data, illustrating the tradeoff between variance and bias. It would be worthwhile to compare the performance of these two methods in future analyses.

Additionally, fine-tuning model parameters through grid search with cross-validation could enhance performance by optimizing model complexity, potentially improving prediction accuracy and robustness. Finally, it would also be worthwhile to search for more accessible and robust indicators for diabetes to improve our model.

Citation

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: From sampling to classifiers. In *Imbalanced learning: Foundations, algorithms, and applications* (pp. 43–59). https://doi.org/10.1002/9781118646106.ch3