Classification Model

Based Inference of Diabetes Patient Features across Racial Dimensions

Questions:

Which factors play a significant role in predicting the onset of diabetes across various ethnic groups? Are there any differences in the importance of these factors among different racial populations, and if so, what are the key variations?

General Method Guideline:

We selected diabetes-related indicators from the 2017-2020 NHANES dataset and addressed data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE).

And we chose four classification models: Logistic Regression, Linear Discriminant Analysis, Random Forest, and Support Vector Machines, which offer complementary strengths and weaknesses.

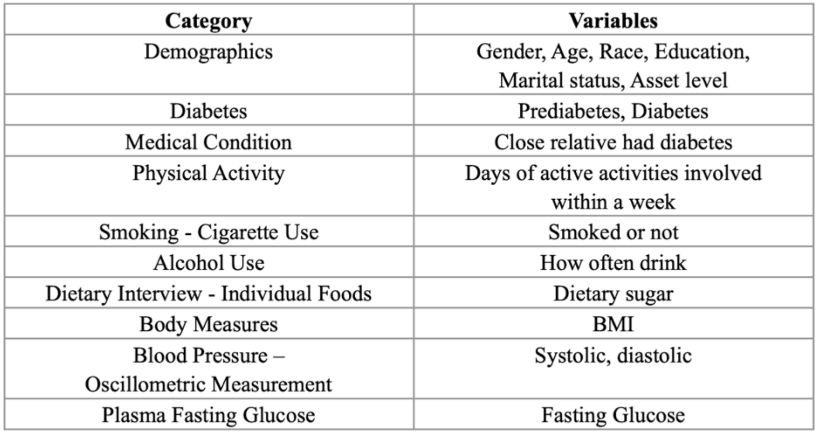
Considering the imbalanced data, we evaluated models based on recall and F1 scores. The Random Forest model, with the highest F1 score, was selected as the most suitable.

Finally, We divided the dataset into five racial subgroups and applied the Random Forest model separately to each group. By examining feature importance rankings, we identified key factors influencing diabetes risk across different racial groups.

Analysis:

1. Data Preparation

Based on the review of two research articles, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning" and "Predicting youth diabetes risk using NHANES data and machine learning," we identified the following variables for our analysis:

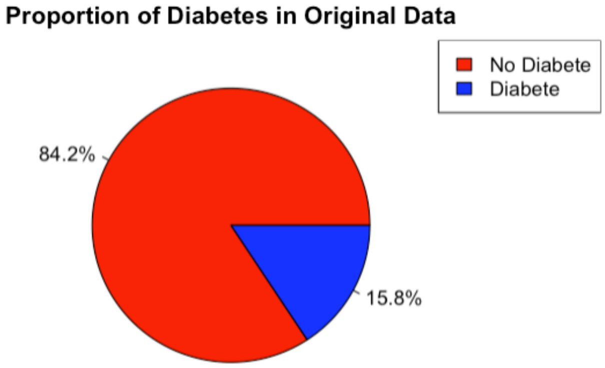


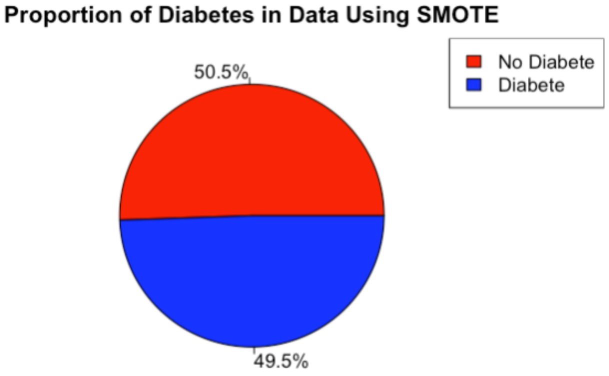
Category	Variables
Demographics	Gender, Age, Race, Education, Marital status, Asset level
Diabetes	Prediabetes, Diabetes
Medical Condition	Close relative had diabetes
Physical Activity	Days of active activities involved within a week
Smoking - Cigarette Use	Smoked or not
Alcohol Use	How often drink
Dietary Interview - Individual Foods	Dietary sugar
Body Measures	BMI
Blood Pressure – Oscillometric Measurement	Systolic, diastolic
Plasma Fasting Glucose	Fasting Glucose

We extracted the data for these variables from the 2017-2020 NHANES dataset. After extracting the data, we noticed that some variables had a significant number of missing values. We imputed the missing values based on the variable type, such as using the mode for categorical variables and the mean for continuous variables. For variables with only a few missing values, we opted to remove the records containing the missing values.

Through EDA, we identified variables with a small number of extreme values and removed those outliers to improve the robustness of our analysis. After cleaning each dataset, we joined them using the unique identifier, SEQN, to create a comprehensive dataset containing all the selected variables. We identified the continuous variables in the dataset and standardized them to ensure that they are on the same scale, which is essential for the performance of some machine learning algorithms.

After preprocessing the data, we divided it into training (80%) and testing (20%) sets. As the problem at hand involves imbalanced data, with fewer instances of diabetic patients, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data to balance the number of cases for both classes. This approach improves the performance of our machine learning models by ensuring they are not biased towards the majority class.





2. Modeling Selection

Logistic Regression is a suitable choice for our dataset as it can handle both continuous and categorical features, such as sugar intake and marital status. Moreover, our research question aims to infer the profile of diabetic patients, so interpretability and ease of interpretation are important factors. Logistic Regression provides interpretable results, enabling us to understand the relationship between the predictor variables and the outcome.

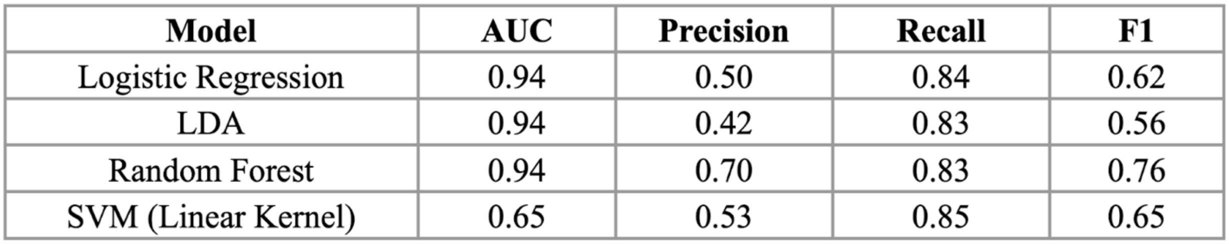
LDA is a robust method for our problem, as it is particularly effective in handling situations where predictor variables may exhibit internal correlations. Since our feature selection is based on previous literature, it is possible that some of our chosen variables are correlated. As a discriminant analysis method, LDA is more robust to multicollinearity than QDA, making it a strong candidate for our problem.

Random Forest is a good fit for our problem, as it can handle large datasets with high dimensionality, which is the case with our dataset containing numerous features and a large number of training examples. Similar to LDA, Random Forest is also robust to multicollinearity, as it is an ensemble learning method that builds multiple decision trees in parallel.

SVM with a linear kernel is an advantageous choice for our dataset. The linear kernel provides a balance between computational efficiency and model complexity, making it suitable for our dataset with a mix of continuous and categorical features. Furthermore, the linear kernel is particularly effective in high-dimensional spaces, allowing us to explore the relationships between our features and the outcome while maintaining a relatively simple model.

3. Modeling Evaluation

In this section, we discuss how we selected our final model. Given the nature of the imbalanced dataset, we are particularly concerned with the predictive ability of the models for the minority group, i.e., the diabetic population. We use the metrics from the classification report for the diabetic cases to determine the best model. First, we consider recall score to assess the models' ability to identify the minority group, and then we consider the F1 score to balance recall and precision performance. We select the model with the best overall performance in these aspects.

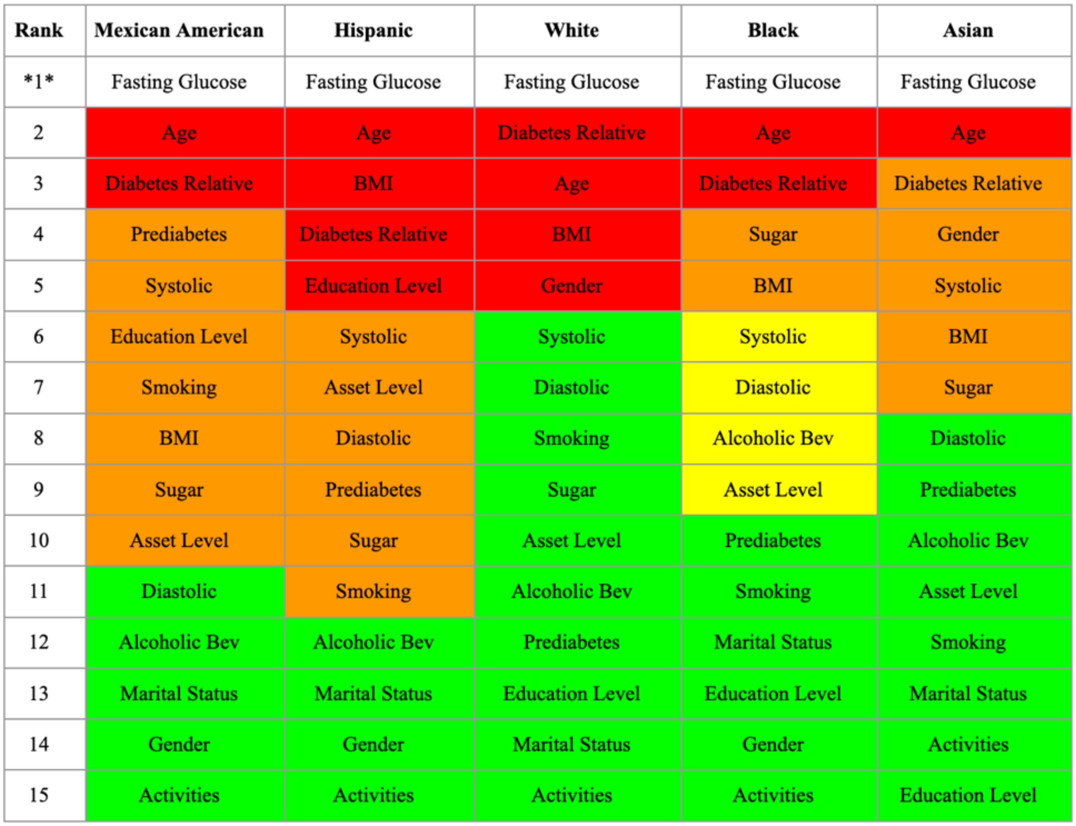


Model	AUC	Precision	Recall	F1
Logistic Regression	0.94	0.50	0.84	0.62
LDA	0.94	0.42	0.83	0.56
Random Forest	0.94	0.70	0.83	0.76
SVM (Linear Kernel)	0.65	0.53	0.85	0.65

Based on these results, We can see that the Random Forest model achieves a similar recall score 0.83 compared to the other models, while having the highest F1 score 0.76. This indicates that the Random Forest model strikes the best balance between precision and recall among the candidate models. Therefore, we select the Random Forest model as our final model due to its superior performance in handling the imbalanced dataset.

4. Race-specific Analysis

We aim to evaluate the importance of each feature for different racial groups using the Random Forest model. We calculate the feature importance based on the Mean Decrease in Gini Index, which represents the average decrease in the Gini Index when a particular feature is used as a splitting criterion in the decision trees. By ranking the features according to their Mean Decrease in Gini Index values, we can identify the most influential variables for predicting diabetes in different racial groups, providing valuable insights for further analysis and decision-making.



Rank	Mexican American	Hispanic	White	Black	Asian
1	Fasting Glucose	Fasting Glucose	Fasting Glucose	Fasting Glucose	Fasting Glucose
2	Age	Age	Diabetes Relative	Age	Age
3	Diabetes Relative	BMI	Age	Diabetes Relative	Diabetes Relative
4	Prediabetes	Diabetes Relative	BMI	Sugar	Gender
5	Systolic	Education Level	Gender	BMI	Systolic
6	Education Level	Systolic	Systolic	Systolic	BMI
7	Smoking	Asset Level	Diastolic	Diastolic	Sugar
8	BMI	Diastolic	Smoking	Alcoholic Bev	Diastolic
9	Sugar	Prediabetes	Sugar	Asset Level	Prediabetes
10	Asset Level	Sugar	Asset Level	Prediabetes	Alcoholic Bev
11	Diastolic	Smoking	Alcoholic Bev	Smoking	Asset Level
12	Alcoholic Bev	Alcoholic Bev	Prediabetes	Marital Status	Smoking
13	Marital Status	Marital Status	Education Level	Education Level	Marital Status
14	Gender	Gender	Marital Status	Gender	Activities
15	Activities	Activities	Activities	Activities	Education Level

Final Interpretation Based on the Feature EDA

Note: Some of the charts mentioned in this analysis are not included in the report. However, they can be found in the completed code output file accompanying this document.

In the race-specific analysis, we employed the feature importance method to rank each variable based on their importance and output values. We categorized the variables into several levels according to their impact:

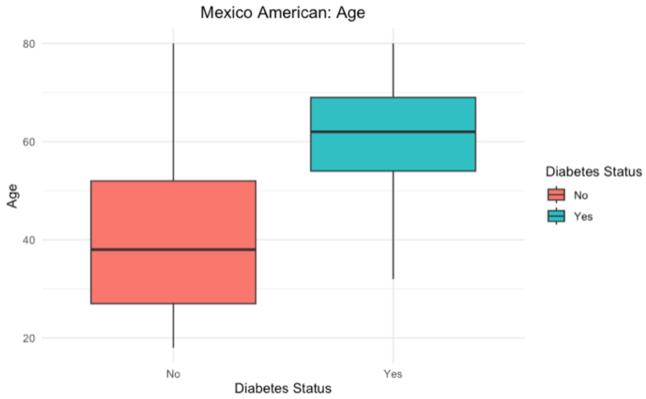
Most important (Dark Red), Moderately important (Rrange), Moderately impactful (Yellow), Very minor impact (Light green), No impact (Pure green).

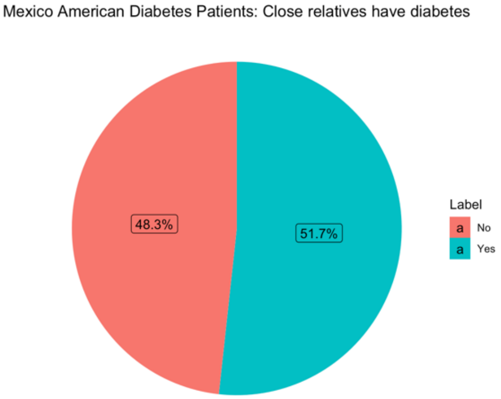
Notably, we excluded the top-ranked variable "Fasting Glucose" as it serves as a diagnostic indicator for diabetes, rendering its importance irrelevant for inferring group characteristics. We also removed the last-ranked variable "Race" because our models are focused on specific racial groups, making the race variable insignificant for reference.

With these considerations in mind, we proceeded with our race-specific analysis. By understanding the key variables that can help predict diabetes for different racial groups, we can gain valuable insights for further analysis and decision-making. Below is a summary of our findings for each racial group:

1. *Mexican Americans:*

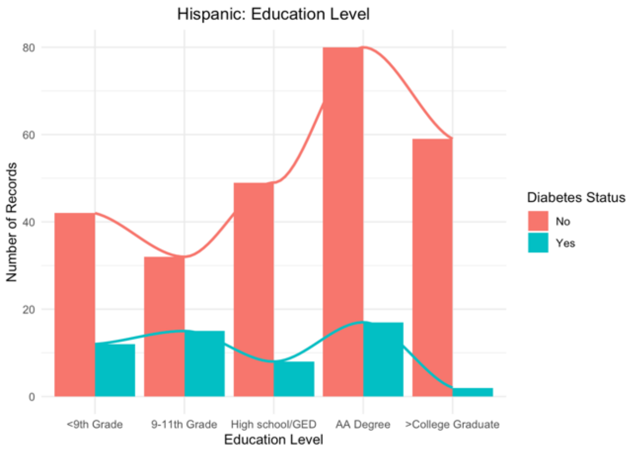
- Age is a significant factor in predicting diabetes, with a higher risk in the 55-70 range.
- Family history of diabetes is also important, but the pie chart reveals no clear trend.





2. *Hispanics:*

- Age range with a higher risk of diabetes is 55-70 years, similar to Mexican Americans.
- Hispanics with a BMI of 30-35 are at a higher risk of diabetes.
- The pie chart shows no clear trend for family history of diabetes, akin to Mexican Americans.
- The histogram shows a more evenly distributed education background among diabetic Hispanics, thus education level may not be a viable predictor. However, it is noteworthy that non-diabetic Hispanics tend to have better educational background

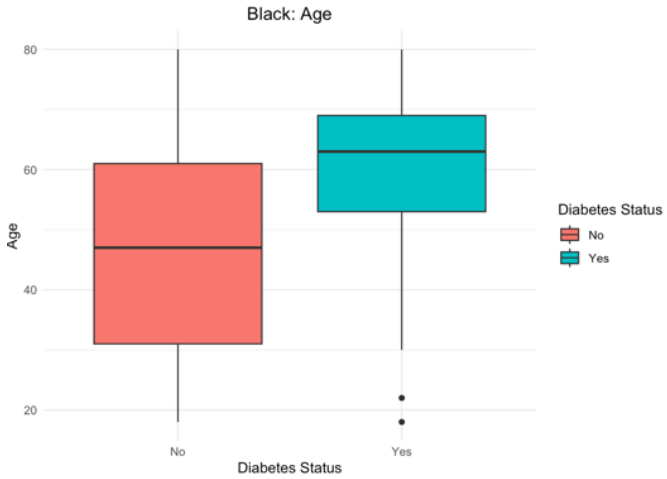


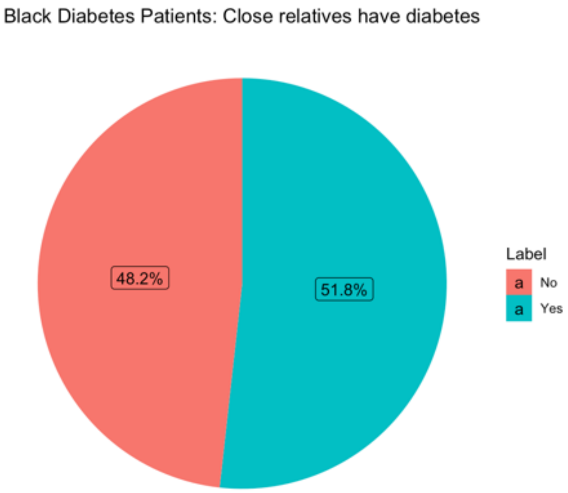
3. *White people:*

- Age plays a crucial role, with a higher risk in the 65-75 years age range.
- BMI also follows a similar pattern as in Hispanics and Mexican Americans, with a higher risk for individuals with a BMI of 30-35.
- Family history of diabetes shows no clear trend, consistent with the other groups.
- The gender distribution appears balanced.

4. *Black people:*

- Age and family history of diabetes are the most important predictors.
- The boxplot shows that the 60-70 age group has a stronger tendency to develop diabetes.
- There is no clear trend regarding family history in the pie chart.





5. *Asian people:*

- Age is the most crucial factor in predicting diabetes, with a significant number of cases occurring in the 55-62 age group. A box plot reveals that there is a lower incidence of diabetes below this age range, suggesting that the 55-62 age group should be the primary focus.

Conclusion:

For Mexican Americans, age (55-70 years) and family history of diabetes were identified as particularly important factors, although no clear trend was observed in the pie chart for family history. In the case of Hispanics, age (55-70 years), BMI (30-35), family history, and, to a lesser extent, education level were observed as significant factors, with non-diabetic Hispanics found to have better educational backgrounds. For White individuals, age (65-75 years), BMI (30-35), and family history were the main factors, with no clear gender trend. For Black individuals, a higher risk of diabetes in the age range of 60-70 years was observed, while family history did not exhibit a strong trend, consistent with the other groups. Finally, for the Asian population, a narrower high-risk age range (55-62 years) was identified, distinguishing it from the other racial groups.

Looking beyond race-specific trends, age, BMI, and family history of diabetes consistently emerged as significant factors across most of the racial groups. Also, some interesting insights emerged, such as the lower influence of BMI on the Asian population, which may be attributed to their typically lower body fat levels or less pronounced obesity issues.

Our race-specific analysis using the Random Forest model provided valuable insights into the varying patterns of feature importance in predicting diabetes across different racial groups.These findings can help inform targeted prevention and intervention strategies tailored to the unique characteristics of each racial group.