Exploring the Impact of Race on Diabetes Risk Prediction Based on Lifestyle

Xizhe Jiang, Fude Cao, Ruobing Zhang

Abstract

Diabetes prevalence in the U.S. shows a concerning disparity across racial groups, with minority groups being disproportionately affected. This study utilized machine learning algorithms to examine lifestyle-related variables and their association with diabetes risk, focusing on these disparities.

Employing the NHANES dataset, we developed models to predict diabetes presence based on lifestyle data, with a specific emphasis on racial information. AdaBoost emerged as the superior model, demonstrating the highest f1-score and recall, particularly when examining data from single-race groups.

Our findings revealed a lower predictive performance for Asian and Hispanic datasets, indicating potential racial biases inherent in traditional predictive models. A deeper analysis into the AdaBoost models revealed that lifestyle habits such as physical activity and dietary choices have varied impacts on different racial groups. For Hispanics, vigorous activity and consumption of frozen meals showed a more correlation with diabetes prevalence, while for Asians, these factors were less impactful compared to the timing and intensity of physical activities. Our study underlines the critical role of tailoring public health initiatives to address cultural and lifestyle factors.

1 Introduction

1.1 Problem

Diabetes has emerged as a significant public health concern in the United States, with alarming statistics indicating its widespread prevalence and disproportionate impact on racial and ethnic minority groups. Literature review has revealed a concerning trend: members of racial and ethnic minority groups are disproportionately affected by diabetes compared to non-Hispanic Whites. Among individuals aged 20 years or older in the United States, the prevalence of diagnosed diabetes is alarmingly higher among Hispanics (11.7%), non-Hispanic Blacks (12.1%), and American Indians or Alaska Natives (14.5%) compared to non-Hispanic Whites (6.9%)(CDC, 2024). This significant disparity between racial and ethnic groups has motivated us to explore whether socioeconomic factors or differing lifestyle habits are the primary causes of this. And how could one know one's risk of having diabetes without doing a full examination in a hospital; and just by checking the lifestyle?

Percentage of US Adults 18 or Older With Diagnosed Diabetes, by Race and Ethnicity, 2019–2021

Race and Ethnicity	Percentage
American Indian or Alaska Native	14.5
Asian, non-Hispanic	9.1
Black, non-Hispanic	12.1
Hispanic, overall	11.7
White, non-Hispanic	6.9

2019–2021 National Health Interview Survey, except American Indian or Alaska Native data, which were from the Indian Health Service National Data Warehouse (2019 data only).

1.2 What others have done

There are plenty of research papers on the cause of diabetes, including biological, behavioral, social, environmental, and healthcare factors, and usually, those factors are highly correlated with each other. Most of those papers use statistics to explain the disparities among different groups. While extensive literature has explore the contributing factors to diabetes, what remains underexplored is the use of machine learning methods to further investigate whether models are equitable across different groups.

1.3 General ideas and Methods

Building on the extensive corpus of research addressing the multifaceted causes of diabetes, our work employed machine learning techniques to unravel the intricacies of diabetes prevalence among diverse racial groups. We used the NHANES dataset as the foundation. First, we designed several machine learning models that could predict whether the respondent has diabetes based on the

respondent's lifestyle with or without racial information. The candidate models included 5 separate models, which were trained on the data of a particular race, and a general model, which was trained on the whole dataset. We tested the common classification models, including KNN, Random Forest, AdaBoost, Logistic Regression, SVC RBF, SVC Poly, MLP, and Deep Neural Network, and selected the model with the highest f1-score. Then we analyzed the result of our candidate models to explore the impact of racial information on the predictions.

1.4 Results and Novelty statement

We found that Adaboost performed the best on the test data for both the separate models and the general model, achieving a f1-score of 35% and an accuracy of 65% on average. The separate models of Asian and Hispanic groups showed non-trivial performance differences from the general model. They had lower f1-scores than the general model. By analyzing the parameters of the two deviated models, we found that some lifestyles impose more or less effect on one racial group than on other racial groups. In this case, for example, minutes of vigorous recreational activities have more effect on the Hispanic group than on other groups. Although the models didn't achieve satisfying scores, we managed to show that the racial factors contribute to the diabetes prevalence difference across different racial groups and identify which lifestyle behaviors would affect some racial groups more or less, with our new machine learning method. Also, our predictive model should be able to predict whether one has diabetes given one's lifestyle with moderate accuracy.

The novelty of our study lies in the application of comprehensive machine learning algorithms to unravel the influences of lifestyle factors on diabetes prevalence across racial groups. Unlike previous research, our approach provides novel insights into the specific lifestyle variables that disproportionately affect Hispanic and Asian populations.

1.5 Organization of the paper

The paper starts with the Related Work section that surveys the landscape of diabetes prediction research. In the Data section, we detail the NHANES dataset and our preprocessing efforts. The Methods and Algorithms section outlines the selection and application of various machine learning models. Subsequently, we describe our approach to Model Evaluations, emphasizing the use of the f1 score as a primary metric. The process of Finding the Best Model, including hyperparameter tuning, is then explained, leading into a discussion of Training Small Models to detect potential biases in model performance. Results and Discussion synthesize the outcomes of our analyses, and we conclude by summarizing our findings and their implications for public health strategies.

2 Related Work

Machine learning (ML) algorithms are increasingly applied for disease prediction, with many studies focusing on diabetes. For example, Kandhasamy and Balamurali compared several classifiers, finding that J48 had the highest accuracy without data preprocessing, while KNN and Random Forest reached 100% accuracy post-preprocessing. Similarly, Yuvaraj and Sripreethaa used ML algorithms to predict diabetes on a pre-processed Pima Indian Diabetes dataset, achieving 94% accuracy with Random Forest. Tafa et al. developed an SVM and Naïve Bayes hybrid model that improved prediction accuracy to 97.6%. Deepti and Dilip evaluated Decision Tree, SVM, and Naïve Bayes on the Pima Indian dataset, with Naïve Bayes performing best with a 76.3% accuracy rate. Mercaldo et al. also used the Pima dataset, applying attribute selection algorithms to enhance classifier performance, with the Hoeffding Tree algorithm showing the best results(Larabi-Marie-Sainte et al., 2019).

Regarding the racial imbalance in diabetes, Spanakis and Golden discuss the possible cause in their paper "Race/Ethnic Difference in Diabetes and Diabetic Complications". They attribute the variance to biological factors and lifestyle, which are verified with scientific experiments. Also, social and environmental statuses play important roles. Minorities commonly live in inferior communities, which make them harder to get healthy food or regular exercise. Groups with less access to healthcare also tend to have a higher chance of having diabetes.

3 Data

In this study, we used a 2015-2020 data set from the National Health and Nutrition Examination Survey(NHSNES) to generate the models. NHANES is an ongoing, cross-sectional, probability sample survey of the U.S. population. It collects demographic, health history, and behavioral information from participants in home interviews.

We selected 10 lifestyle-related variables commonly associated with the risk for diabetes, including smoking, alcohol consumption, race, and physical activity, diet habit, sleeping and working. Features we select and descriptions are shown in Fig.2.

Feature description table

NHANES code	Description
RIDRETH3	race: Mexican American, Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian
DBD900	number of meals from fast food or pizza places in past 30 days
ALQ130	average number of alcoholic drinks per day - past 12 months
DBD910	number of frozen meals/pizza in past 30 days
SMD650	average number of cigarettes per day during past 30 days
PAD660	minutes of vigorous recreational activities
PAD675	minutes of moderate recreational activities
WHQ040	like to weigh more, less or the same
SLD012	sleep hours - weekdays or workdays
OCQ180	hours worked last week in total all jobs

Figure 1

Only data pertaining to individuals over the age of 18 were selected. We preprocessed the data using the following rules (Fig.3).

```
'RIDRETH3': ([7, np.nan], [np.nan, np.nan]),
'DIQ010': ([2, 3, 7, 9, np.nan], [0.0, 1.0, np.nan, np.nan, np.nan]),
'ALQ130': ([777, 999, np.nan], [np.nan, np.nan, 0.0]),
'DBD900': ([5555,7777,9999,np.nan], [25.0, np.nan, np.nan,np.nan]),
'DBD910': ([6666, 7777, 9999, np.nan],[90.0, np.nan, np.nan, np.nan]),
'SMD650': ([777, 999, np.nan], [np.nan, np.nan, 0.0]),
'PAD660':([7777,9999,np.nan],[np.nan,np.nan,0.0]),
'PAD675':([7777,9999,np.nan],[np.nan,np.nan,0.0]),
'WHQ040':([7,9,np.nan],[np.nan,np.nan,np.nan]),
'SLD012':([np.nan],[np.nan]),
'OCQ180':([7777,9999,np.nan],[np.nan,np.nan,0.0]),
```

Figure 2

After carefully reviewing the questionnaire, we decided that we could replace some of the missing values with valid values because those missing values were skipped by the previous question instead of missed by the respondents. Then, we replaced invalid numbers and outliers with missing values and dropped all the rows with missing values. The remaining dataset has a mostly balanced number of respondents per race (about 2000 respondents per race), but it includes 20% of respondents with diabetes and 80% without diabetes. We then split the data into 80% training set and 20% test set. Because of the highly imbalanced labels in the training data, we used SMOTE to duplicate some data with diabetes labels to balance the training set to optimize the model. After preprocessing, we got a cleaned dataset with 10856 entries.

4 Methods and Algorithms

4.1 Model generation

In our algorithms, we utilized eight models to analyze the data for predicting diabetes based on lifestyle factors:

- **K-Nearest Neighbors (KNN)**
- **Random Forest**
- **AdaBoost**
- **Logistic Regression**
- **Support Vector Classifier with Radial Basis Function (SVC RBF)**
- **Support Vector Classifier with Polynomial Kernel (SVC Poly)**
- **Multilayer Perceptron (MLP)**
- **Deep Neural Network (DNN)**

We selected these models based on their prevalence in supervised learning tasks and their proven track record in various applications. The versatility of this ensemble ranges from simple, interpretable models like Logistic Regression and AdaBoost to complex, high-capacity models like Deep Neural Networks. This diversity ensures a comprehensive analysis of the data, allowing us to capture both linear and non-linear patterns.

Each model offers a unique approach to learning from data, which is crucial given the multifaceted nature of lifestyle-related factors in diabetes prediction. Models such as KNN and Random Forest can handle the intricacies of high-dimensional interactions without extensive feature engineering. Meanwhile, SVC with RBF and Polynomial kernels can model complex decision boundaries that might elude simpler linear models. Neural networks, including MLP and Deep Neural Networks, provide advanced feature learning capabilities, which can be particularly beneficial for uncovering subtle patterns within complex lifestyle data. Their depth and flexibility allow us to experiment with the architecture to best fit the dataset's nuances.

While there are indeed more sophisticated models available, they often come with increased computational costs and complexity, making them less practical for our scope. Additionally, complex models can pose challenges in fine-tuning and interpretability. Our chosen models strike a balance between performance, computational efficiency, and ease of interpretation. Therefore, we believe this set of models is well-suited for the objectives of our project.

4.2 Model evaluations

We use f1 score when evaluating the performance of each model. The f1 offers a balanced measure of precision and recall. This balance is critical in diabetes prediction, where false negatives can have serious implications. The f1 score helps to reduce such errors without compromising overall model accuracy. Thus, the f1 score is the most suitable metric for our project.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (1)$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4.3 Finding the best model

To find the model that will achieve the highest f1 score on our dataset, we first fine-tuned each model. We utilized search algorithms specifically aimed at optimizing the f1 score to determine the optimal hyperparameters for every model. Grid search was applied to pre-built models like KNN, AdaBoost, Logistic Regression, and SVC RBF due to its thoroughness. However, for SVC Poly and MLP, we opted for random search given the excessive computational demands of grid search. For our deep neural network, we designed a flexible model that uses a list to store the hidden layers. The list and hidden layers will be initialized in accordance with the depth and width parameters in the constructor of our model. Again we use grid search to identify the best configuration for these parameters.

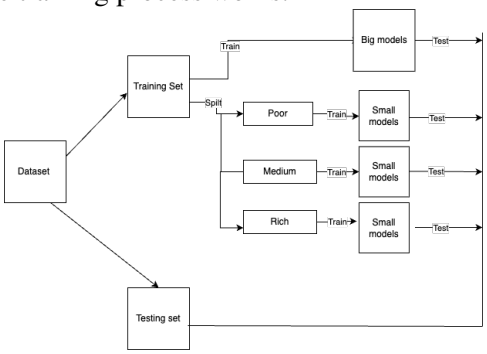
Upon identifying the optimal hyperparameters, we proceeded to train the models using the full dataset. For the deep neural network, training and testing loaders were established, and the model was trained over 1000 epochs. In contrast, for the remaining models, we utilized their respective built-in 'fit' functions for training. The performance outcomes of each model are detailed below.

model	recall	f1 score	accuracy
knn	52.63%	31.93%	62.71%
random forest	31.02%	30.21%	75.97%
adaboost	62.60%	35.87%	62.80%
logistic regressio	64.27%	35.10%	60.50%
svc	50.42%	35.79%	69.94%
mlp	52.63%	34.05%	66.11%
neural network	53.90%	30.40%	74.91%

In conclusion, the f1 scores of all these models are in the 30% - 36% range. Adaboost has the highest f1 score; however, its f1 score is not significantly higher than that of other models.

4.4 Training small models

To investigate potential biases within our dataset and their effects on model performance, we conducted a structured analysis. Initially, we segmented the dataset into distinct groups based on specific attributes, such as the respondents' wealth status and race group. We then trained our eight models -referred to as 'small models' - using data exclusively from each respective group. The efficacy of these small models was evaluated against a comprehensive test set containing respondents from all categories. This performance was also compared with 'big models', which were developed using the full dataset. If the small models perform less effectively than the big models, it could indicate that the data from the specific group, used to train the small model, has biases. To ensure a fair comparison between models in our study on data bias and its impact on performance, we maintained consistent hyperparameters across both the small and big models. This decision was made to attribute any observed performance disparities solely to data bias, rather than to differences in the models themselves. By keeping the hyperparameters uniform, we eliminated one variable that could otherwise skew the results, thus isolating the effect of the data's characteristics on model efficacy. Down below is a diagram that demonstrates how the training process works.

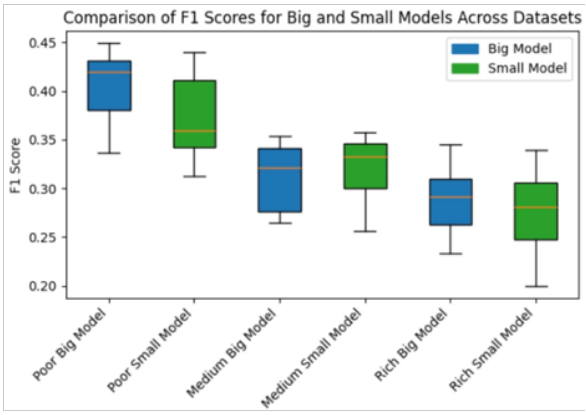


In addition to the initial approach, we employed a second method to construct small models. The training process was largely the same, except in this case, we trained the models on data from all groups with the exception of the target group.The purpose of this exclusion was to investigate whether the absence of a certain group would significantly improve the model's performance. Should we observe a marked enhancement in performance as a result

of this exclusion, it would suggest that the target group may contain biases that adversely affect the performance of the larger models. This could indicate the presence of underlying bias within the dataset pertaining to the target group.

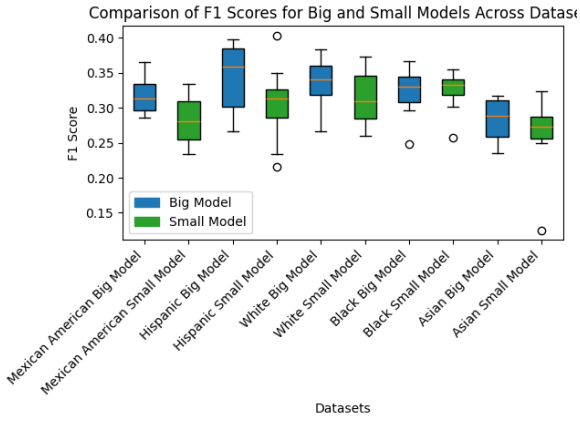
5 Results and discussion

We began by categorizing the dataset according to the economic status of the respondents, specifically into poor, medium, and rich groups. To illustrate the performance of the models, we utilized box plots to display the F1 scores achieved by the eight small models and the eight big models. The small models in this plot were trained by using the first method (training on models exclusively in the target group).



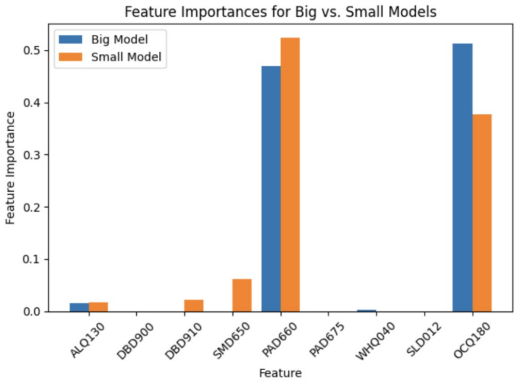
The box plot comparison revealed that there was no significant performance discrepancy between the big models and the small models across all wealth categories. We also used the second method(training the models on data from all groups with the exception of the target group), and the difference between performance is less obvious than using the first method. In line with our research focus, while we observed that the predictive performance of the models was somewhat lower for the less wealthy groups, our primary interest lay in examining the performance differential between the big and small models. Therefore, we chose not to investigate the broader performance trends in depth.

Next, we divided the group based on their racial identities and plotted the data in a box plot.



The disparity in performance between the big models and the small models is particularly pronounced within the Hispanic group, whereas for other racial groups, this difference is less distinct. We also use our second method to train small models, but the difference between performance is also less obvious.

To pinpoint the features contributing to this bias, we opted to utilize Adaboost. This model not only offers interpretability but also had the highest F1 score among our eight models when training on the entire dataset. We utilized the pre-built feature importance function to plot the following graph.



The results revealed that, in comparison to the big Adaboost model, the model trained on Hispanic data attributed greater importance to variables such as minutes of vigorous sports, number of frozen meals consumed, average number of cigarettes per day, and average number of alcoholic drinks per day. Conversely, it assigned lesser importance to the variable representing hours worked in the previous week.

We also did a similar analysis to the Asian group as we observed that there was a notable difference in big and small AdaBoost in the Asian group. Compared to the big model, the model specific to the Asian demographic assigned increased im-

portance to variables such as minutes of vigorous sports and weight loss aspirations, while it accorded reduced importance to the variable indicative of hours worked in the previous week.

Table Importance of features in Adaboost Models

Feature	Combined model	Hispanic	Asian
sleep hours	0	0	0
hours worked last week	0.48153501	0.34151889	0.45864596
minutes of vigorous sports	0.49095189	0.57692582	0.51319468
Minutes of moderate exercises	0	0	0
Like to weight more/less/same	0	0	0.00219183
number of meals from fast food	0	0	0
number of frozen meal	0	0.03666119	0
Avg # cigarettes/day	0	0.04724235	0
Avg # alcoholic drinks/day	0.0275131	0.04765176	0.02596753

6 Ethical implication and Solutions

6.1 Data Privacy and Anonymity

Although the National Health and Nutrition Examination Survey (NHANES) datasets are designed for public use, data privacy remains a paramount ethical concern. In our project, we only utilize the publicly accessible portions of the datasets and adhere strictly to the data use agreements stated on the NHANES website, ensuring that the identity of participants is not disclosed. We follow the NHANES Data Use Agreement, which mandates that users must not attempt to identify individuals. All data handling procedures are conducted in a secure environment.

6.2 Research Fairness

Another significant ethical aspect of our project involves the development of race-based predicting models. We are acutely aware of the potential to exacerbate existing social and health inequalities and are therefore committed to ensuring that our research methodologies are designed to benefit all groups equally. By actively seeking methods to reduce bias and inequality, our project aims to foster research fairness and contribute positively to the field, ensuring that the benefits derived from our findings are distributed fairly across all demographic groups.

6.3 Impact of Results

The potential misuse or misinterpretation of our research findings poses a considerable ethical challenge, particularly in how it could lead to prejudice or discrimination against certain groups. To mitigate this risk, we are dedicated to clearly communicating the limitations of our research. We will provide detailed explanations regarding how the results should be interpreted and applied when we publish and discuss our final outcomes. This approach is intended to prevent misunderstandings and misuse of the data, promoting a responsible dissemination of information that acknowledges the nuanced nature of our findings.

7 Conclusion

In this study, we evaluated eight classification models to detect diabetes cases within the U.S. population. The results indicated that the AdaBoost model outperformed all other models in terms of f1 score and recall. Subsequently, we explored whether the overall model exhibited any bias toward specific racial groups using this model. After controlling for socioeconomic differences, the analysis revealed that the model's predictive recall and f1 score were relatively lower for Asian and Hispanic datasets. This suggests the possibility of an unfavorable bias within the model against these two racial groups.

To further identify the root causes of this disparity, we analyzed the feature importance in AdaBoost models trained with data from separated racial groups.

We found that specific lifestyle habits significantly contributed to these disparities. For instance, for the Hispanic population, factors such as the duration of vigorous physical activity, frequency of consuming frozen meals, and smoking and drinking habits had a more pronounced impact on diabetes prevalence compared to other races. Conversely, for Asians, the influence of drinking habits and work hours on diabetes was less significant, while the timing and duration of vigorous physical activity played a more crucial role.

The Hispanic population's increased diabetes risk related to diet and exercise routines could reflect cultural food preferences and barriers to physical activity opportunities.

Tailored public health interventions and educational programs that respect and address these cultural nuances may prove to be effective in mitigating the diabetes risk in these communities.

8 References

- 1.Spanakis, E.K., Golden, S.H. Race/Ethnic Difference in Diabetes and Diabetic Complications. *Curr Diab Rep* **13**, 814–823 (2013). https://doi.org/10.1007/s11892-013-0421-9

- 2.National Center for Health Statistics. (2015-2020). National Health and Nutri-

tion Examination Survey. U.S. Department of Health & Human Services, Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/index.htm

- 3.Larabi-Marie-Sainte, Souad, Linah Aburahmah, Rana Almohaini, and Tanzila Saba. 2019. "Current Techniques for Diabetes Prediction: Review and Case Study" Applied Sciences 9, no. 21: 4604. https://doi.org/10.3390/app9214604

9 Appendix

All the data processing, model construction, and result analysis conducted for this project can be found at the following link https://github.com/RuobingZ0305/eecs448-project