ORIE 5741

Final project (Data analysis)

Prediction of diabetes among Adults in the US: an analysis and application of nationally representative data from National Health and Nutrition Examination Survey

Naiwen Ji (nj294), Yuze Qin (yq276), Zhaohan Xing (zx333)

Content

Problem specification **2**

Description of Dataset **3**

Results **5**

Discussion **9**

Reference **11**

Problem overview:

The interested prediction question is:

With general information related to one's socio-demographic, lifestyle and non-invasive anthropometric characteristics, can we predict one's diabetes status?

In the US, about 38 million people have diabetes (1). Diabetes alone is the eighth leading cause of death in the US. However, according to a recent study, about 34.3 % of the diabetic population were unaware of their diabetes condition (2). In the past two decades, the prevalence of diabetes has increased dramatically in the US. In 2020, the estimated prevalence of prediabetes was 38.4 million people aged over 18 years had prediabetes which consisted of about 12% of the total US adult population (3). Diabetes is also an important risk factor for various medical conditions and chronic diseases including blindness, kidney failure, heart attack, stroke, and lower limb amputation. Studies also showed that diabetes was associated with an increased risk of most cancers and infectious diseases including COVID-19 (4).

The American Diabetes Association (ADA) published an economic cost of diabetes in the US in 2022 which indicated that the total annual cost of diabetes was 412.9 billion US dollars including 306.6 billion in direct medical costs and 106.3 billion in indirect costs. The indirect costs of diabetes were related to disability, presenteeism, and lost productivity due to premature death resulting in diabetes. Although diabetes as a chronic disease is manageable and has relatively slower progression compared to infectious disease, as shown in the report, it is causing additional loss of revenue and extra expenses for health insurance (5). The standards of care for people living with diabetes include diagnosis and classification of diabetes, glycemic monitoring and management, comprehensive medical evaluation and assessment of comorbidities, and facilitating positive health behaviors and well-being to improve health outcomes (6). The health insurance industry is highly involved in these processes of diabetes management. To better prevent diabetes-related economic loss and the decision-making process of health insurance, there is an urgent need to understand the dynamics of diabetes and develop a prediction model for diabetes.

Machine learning-induced prediction model is a rising method in the public health and nutrition fields. A good example of a successful implementation of the algorithm in the prediction of cardiovascular diseases (CVD) was the ASCVD risk estimator conducted and implemented by the American College of Cardiology. The prediction model was developed based on the generalized linear regression model and the model is widely used in research and clinical settings as an assessment and prevention tool (7–9). As for diabetes, there is no existing commonly used prediction model. However, in a 2019 literature review, the authors mentioned the potential opportunities of utilizing machine learning (ML) classifier algorithms to develop prediction models for diabetes (10). Studies have compared the performance of ML classification algorithms such as logistic regression, forests, support vector machine, artificial neural networks, and Naive Bayes to predict diabetes. Both of the studies achieved decent accuracy which indicates the feasibility of this proposed project (11,12). With access to the most updated nationally representative dataset in the US with high-quality data, we will be able to develop a prediction model for diabetes.

Description of dataset

Data utilized in this study from the National Health and Nutrition Examination Survey (NHANES). Survey design and details about data collection were available elsewhere (13). Briefly, NHANES is a publicly available, large-scale, and nationally representative health-related survey conducted in the US. In this study, we

will use data from NHANES cycle 2017-2018. Participants aged over 18 years diabetes measurements were included in the analyses.

From previous literature review, the potential correlates of diabetes including socio-demographic information including age, sex, body mass index (BMI), systolic and diastolic blood pressure, race/ethnicity (non-Hispanic White, non-Hispanic black, Hispanic, and other), education level (less than a high school diploma, high school graduate/GED, some college/AA degree, and college graduate or more)(14), and household income (≤130% (reference group), >130% to 350%, and >350% by the ratio of family income to poverty (FPL)) (15) were included in this study. Lifestyle factors including smoking (none, former, current smoker) (16-17), and body mass index (BMI) were included in the initial analyses as well.

Data preparation:

All aforementioned features underwent a data cleaning process according to the scientific recommended and commonly used cutoffs as described above. The continuous features such as BMI, household income, alcohol consumption and physical activity level were evaluated for outliers and assessed for biological plausibility. For ML approaches, the output was diabetes. The output was encoded as a binary dummy variable. Due to the nature of the algorithm suitable for classification, all categorical features will be encoded as dummy values, and the continuous variables were normalized before input in the ML models. The summary of the features included in the final prediction model was shown in table 1.

Table 1 summary of data types

Continuous	3 (Age, Body mass index, Diastolic blood pressure)
Categorical	1 (Race/ethnicity)
Ordinal	4 (Smoking status, Income to poverty ratio, Education level)

NHANES dataset used a complex surveying design that, in nature, captured representative samples across the US. However, due to the large scale of the survey and comprehensive information collected during the survey and lab visits, the missingness of the data is a serious concern. Even though most missingness among features in our dataset was not major, we decided to multiply imputation according to CDC recommendation before starting to apply other feature engineering techniques and fit them into prediction models. The distribution of each feature is described in Figure 1. We implemented a multivariate linear model with random forest as the approach to impute the missing values in the dataset using the relationship between features for categorical and ordinal features. As for continuous features, we used the median to impute the missingness in the dataset. To check the reasonability between and after the imputation, we used Chi-square Goodness-of-fit to test the distribution before and after the imputation, and Cramer's V test, which is a similar test for goodness-of-fit but excludes the effect of sample size. From the Chi-squared test, we found that the distribution of most of the variables did not change before and after the imputation. However, some variables failed to pass the test because of relatively large missingness (for example, "Alcohol" has over 50% missing rate). When we used Cramer's V, which is a statistic that can alleviate the effect of sample size, we found that all variables had Cramer's V below the general criteria. So, we can conclude that our imputation is reasonable from the distribution side. For further details, our code on GitHub explains and summarizes the results. Additionally, we used Pearson Correlation Spearman's Rank Correlation, and Covariance matrix to check whether the imputation captured the relationships among and between the features. Figure 2 visualizes the correlation between features. The correlation between features did not change significantly which indicated that our imputation captures the

correlations and covariances between all pairs of the variables. After imputation, the distribution of each feature did not change significantly with the data frame being more completed, potentially increasing our model accuracy and reducing bias.

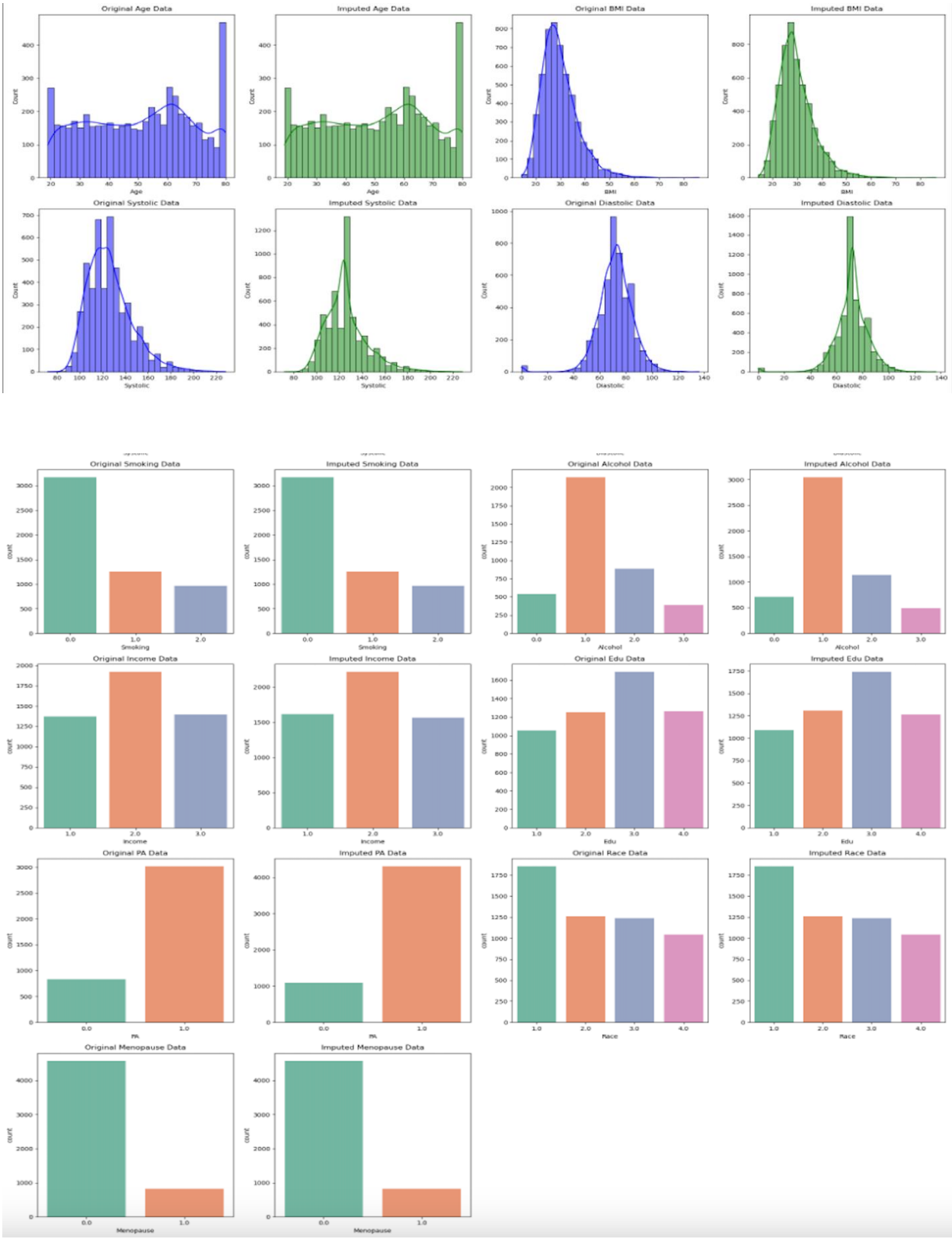

Figure 1 Distribution of included features before and after imputation.

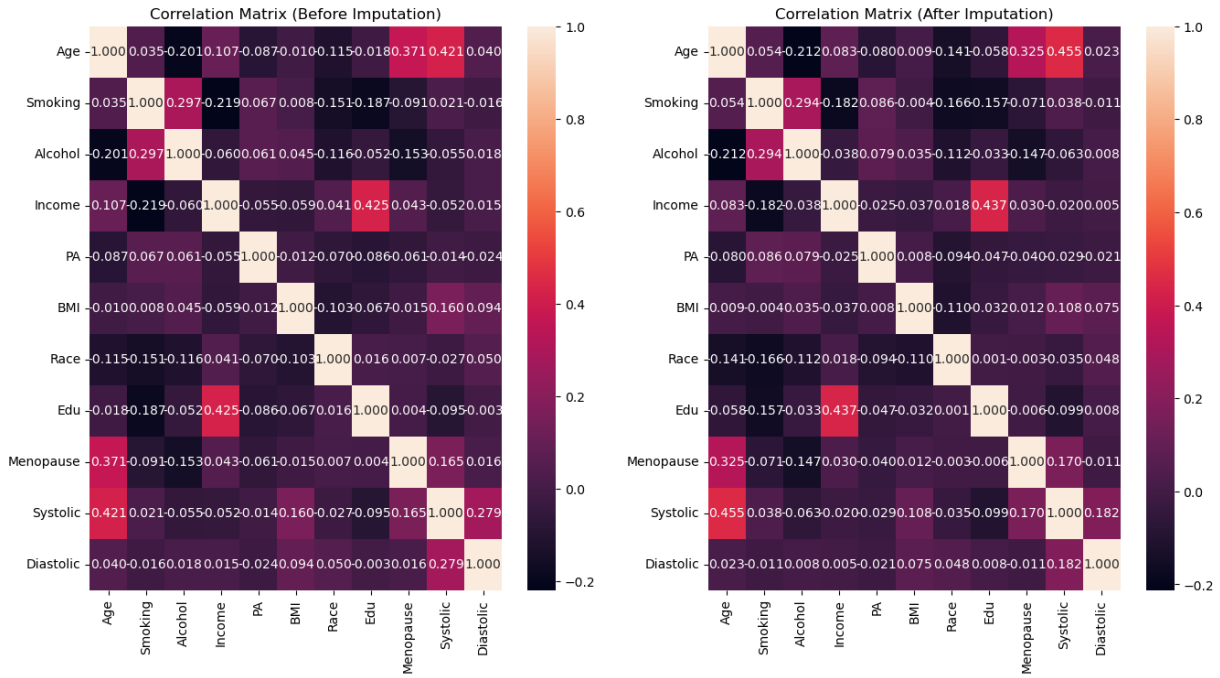

Figure 2 The heatmap that shows the covariances between features before and after the imputation.

Model Approaches

After data preparation, the dataset was composed of a sample size of 5712 adults living in the US aged over 18 years. To ensure the optimization of the performance of our ML models. We conducted feature engineering techniques such as one-hot encoding. In our dataset, we have physical activity level, race/ethnicity, and menopause status being categorical features. One-hot encoding could potentially improve model performance, and increase model flexibility while still preserving the information of the categorical features. We used different supervised ML classifiers to develop our prediction model for diabetes including support vector machine (SVM), random forest, logistic regression, and k-nearest neighbors (KNN).

All models mentioned above were tested for performance. The evaluation was based on sensitivity, specificity, accuracy, and likelihood ratio. The area under receiver operating characteristic curve (ROC) will be used to assess the performance of the models. The cross-validation procedure will be used to ensure the performance and to preserve generalizable model conclusions. A 10-fold cross-validation will be used in the training set. Then the best-fit model was used to test the prediction of diabetes in the testing set we previously defined. Since the tool we developed focused on the screening of diabetes, instead of diagnosing, it with non-invasive and easily obtained features, our main evaluation tool was accuracy. The screening of disease aims to achieve early detection, healthcare allocation, public health implications, and cost-effectiveness (18). Due to the goal of screening, we would like to develop a tool to maximize the benefits of using non-invasive and easily obtained features rather than having a tool that can predict diabetes with high accuracy but needs costly and complicated lab tests.

Results

Before starting the analyses, we visualized how the counts of diabetes are distributed among all these features, which is shown in Figure 3. We can see that some features have a very non-uniform distribution, which may

indicate that they are crucial for determining diabetes. In addition to the visualization, we also performed a preliminary analysis using logistic regression evaluate the feature significance in relation to DM because logistic regression is commonly used in the public health field to evaluate association. After the evaluation, we filtered out the feature, alcohol, physical activity, systolic blood pressure and menopause status due to the statistical insignificance ($p > 0.05$, data not shown).

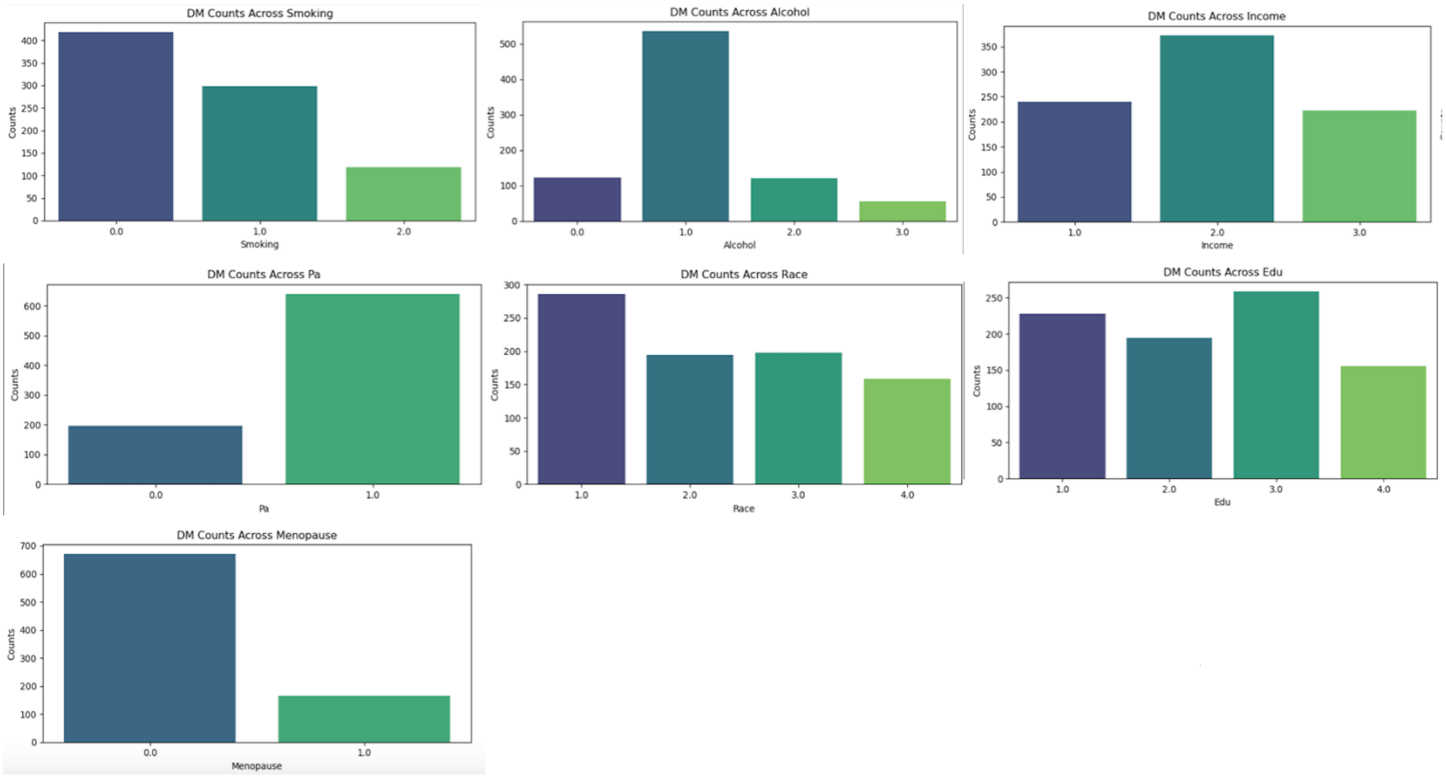

Figure 3 Cases of diabetes frequency by features.

Support vector machine: SVM is a powerful algorithm for classification tasks that can handle both linear and non-linear relationships between features and outcomes. This advantage makes it suitable for complex datasets.

After splitting the dataset into training and testing datasets by the ratio of 80/20. We first implemented SVM without tuning and kernel selection. In addition, the DM positive consisted of 15.47% of our total population which makes the classes unbalanced. The algorithm failed to predict the DM-positive cases due to the unbalanced class. We then decided to tune the hyperparameters and apply the "balance" function. Selecting a suitable Kernel transformed the feature when the dataset is not linearly separable. We chose the polynomial kernel as the kernel function. Moreover, we use the bagging classifier to improve model robustness and accuracy. We conducted a hyperparameter optimization process to determine the best configuration for our SVM model. We experimented with various kernel functions, regularization parameters, and other settings. By evaluating each model's performance on a validation set, we selected the SVM model configured with a polynomial kernel, a regularization parameter of 1. As shown in Figure 4, we found that the performance of SVM is good on response zero (no diabetes) and has an accuracy of 0.84 on testing data.

Random forest: Random Forest is a versatile ensemble learning algorithm that offers valuable insights of feature importance while avoiding the disadvantage of decision tree, overfitting. Another advantage of random

forest is that the output is interpretable. By harnessing the power of the random forest, our analysis yields compelling insights, as illustrated in Figure 5, we found that age, BMI, and Diastolic play important roles in the random forest model. Also, from Figure 4, we can see that Random Forest is not better than SVM on testing data, with an accuracy of 0.76.

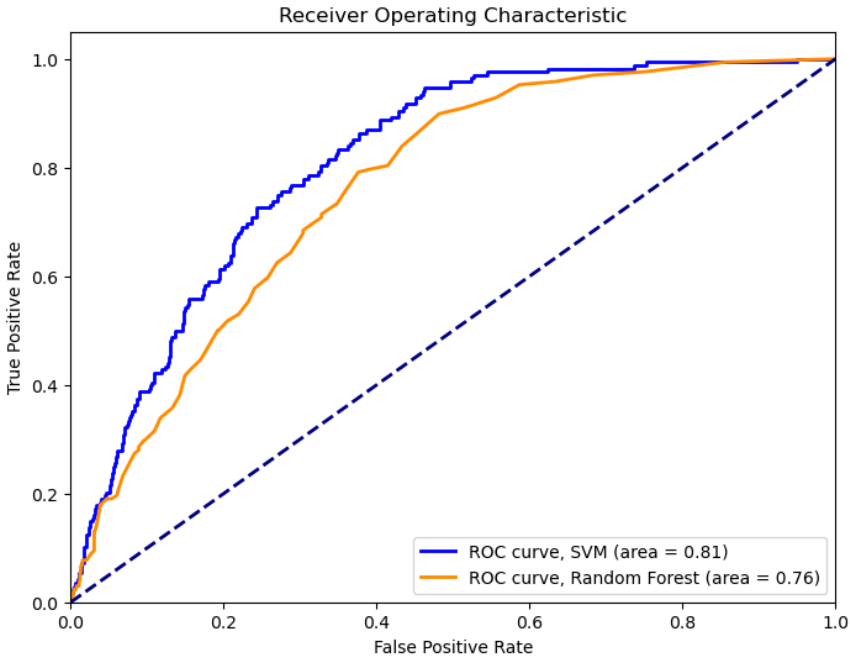

Figure 4 The ROC curve for the SVM and RF model. From the plot we see that SVM model has a larger AUC compared with RF model.

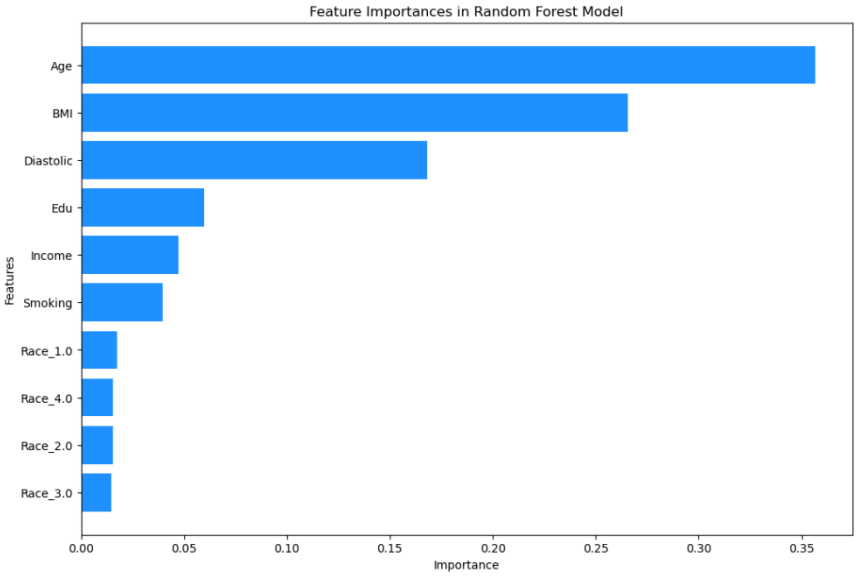

Figure 5 The importance rank of the features in the RF model. Age and BMI play important roles in this model. Note that we use One-hot encoding for Race, and it plays the lease importance in this model, which may indicate that Race is not crucial when determining diabetes.

K-Nearest Neighbors: KNN is an ML algorithm that uses the proximity to existing labeled data points. The challenge of implementing KNN is to find the "perfect" k. In order to find the k, we used the cross-validation to

perform hyperparameter tuning. The KNN model's performance (Figure 6) showcases a striking contrast between its ability to identify non-diabetic and diabetic individuals. Specifically, the model achieves perfect specificity and precision for non-diabetic predictions, indicated by a recall and precision of 1.00 for class 0, ensuring all non-diabetic cases are accurately identified without false positives. However, its effectiveness sharply drops when detecting diabetic cases, with a recall of only 0.01 for class 1 despite perfect precision. This means while every individual the model identifies as diabetic is correctly classified, it fails to identify 99% of the actual diabetic cases, highlighting a significant limitation in its utility for reliable diabetes screening.

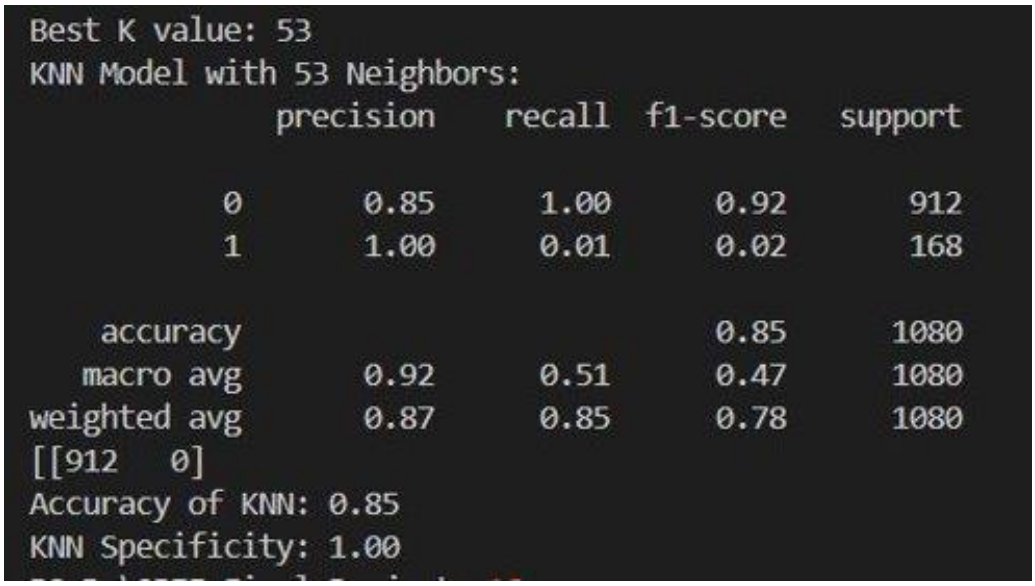

Figure 6 Model performance metric summary of KNN model when k = 53

Logistic regression: logistic regression, as the most commonly used algorithm in public health field, could provide not only classification but also provide the change of effect of certain feature. Comparing with most of ML algorithms, logistic regression offers more interpretable results and robust to noise while carries some cons such as intense assumption of linearity and limited complexity.

	coefficient	SE	P value
Age	0.067	0.012	0.001
Smoking	0.125	0.030	0.002
Income	-0.038	0.020	0.050
BMI	0.074	0.015	0.001
Race	-0.121	0.040	0.010
Education	-0.004	0.010	0.032
Diastolic	0.173	0.045	0.001

Table 2 The output of the LR model

Table 2 shows the output of the LR model. For interpretation, we converted coefficients into odds ratio in the following reporting. In summary, one year increase in age was associated with approximately a 7% increase in the odds of having diabetes, as the odds ratio (OR) for age is about 1.07, with a 95% confidence interval (CI) from 1.04 to 1.10. Being a current smoker was associated with 13% higher odds of having diabetes onsite (OR $\approx$ 1.13, CI: 1.07 to 1.20). Conversely, higher income levels are linked to a slight decrease in diabetes, reducing

the odds by 4% per unit increase (OR ≈ 0.96, CI: 0.93 to 1.00). A higher BMI significantly raises diabetes risk by 8% per unit increase (OR ≈ 1.08, CI: 1.05 to 1.10). The strong effects of age, BMI, and smoking highlight key areas for intervention to potentially mitigate the risk of developing diabetes.

Figure 6 shows the performance of the LR model. The LR model demonstrates robust performance with a recall of 0.77 for diabetic cases, indicating its effectiveness in identifying a high proportion of actual diabetes cases. Additionally, the model achieves a specificity of 0.69, affirming its ability to correctly identify a significant majority of non-diabetic individuals.

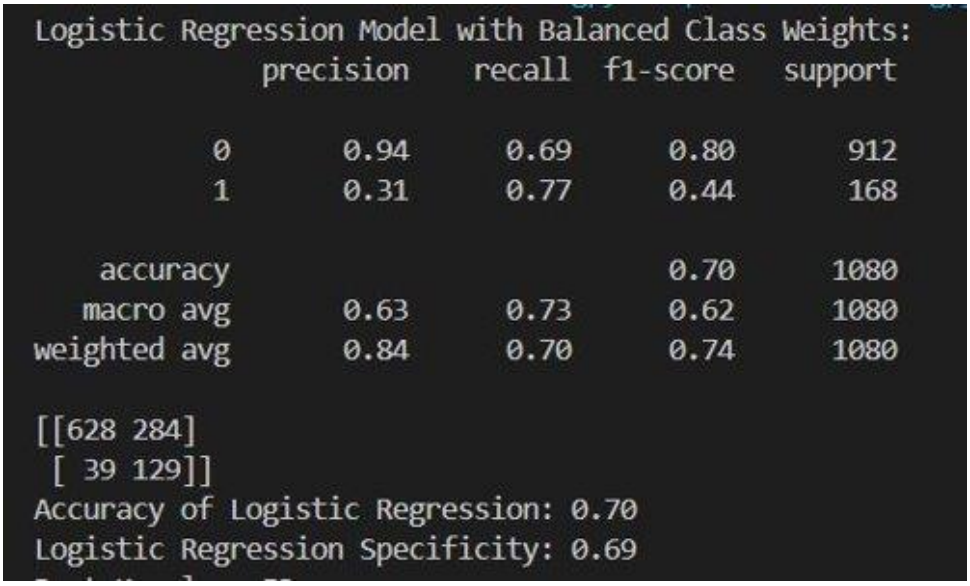

Figure 6 Model performance metric summary of LR

In our evaluation of different models for diabetes screening, LR clearly stands out due to its superior sensitivity, achieving a recall of 0.77. This is significantly higher compared to RF and SVM which only managed recalls of 0.08 and 0.07, respectively. KNN performed particularly poorly in this aspect, with a recall of just 0.01. While KNN demonstrated perfect specificity of 1.00, surpassing SVM (0.99), RF (0.98), and LR (0.69), its low sensitivity makes it less suitable for our primary goal of maximizing the detection of true positives without missing cases.

In our project aimed at developing a diabetes screening tool for use in public health and health insurance settings, the choice of the best model was critical. Our primary concern was ensuring a high detection rate of potential diabetes cases without allowing many cases to go unnoticed. This led us to prioritize models with high sensitivity, or the ability to identify most, if not all, true positive cases of diabetes.

Given these considerations, LR not only meets our need for high sensitivity but also maintains a workable level of specificity, creating an ideal balance for a screening tool. This model provides a robust framework for potentially improving early diabetes detection and management in public health settings, enhancing both individual patient outcomes and broader public health strategies.

By choosing LR, we are leveraging a model that effectively balances the trade-offs between identifying as many true cases as possible and managing false positives. This approach aligns with our overarching goal of

providing a reliable, efficient screening process that can be easily integrated into public health and insurance frameworks, ultimately facilitating better health management and resource allocation.

Discussion

Weapons of Math Destruction: Our study and developed models aim to offer a screening tool for diabetes in the public health prevention and insurance industry. The outcome, diabetes, is not hard to measure. The standardized test for diabetes is the fasting blood glucose level. However, in resource-limited settings or daily life, people will not go to their doctor specifically for a fasting blood glucose test. This explains the under-detection of diabetes in the general US population. From a public health perspective, the predictions will not harm anyone. The prediction model could serve as a screening tool for health professionals or even the general population to use and decide whether an individual needs to go to a doctor for a fasting blood glucose test. However, from an insurance industry perspective, this prediction model could be harmful when the insurance company decides to exclude somebody from being insured based on this prediction model. The prediction model could create a feedback loop. For example, in the public health setting, the individuals visit the setting and get their non-invasive information collected and may go to the doctor afterward for a blood test. Then the result of the diagnosis could feed back to the model and improve the model.

Fairness: From the perspective of public health screening purposes, we see our dataset, model, and results are fair. During the data collection process of NHANES, fairness was a consideration where they oversampled the population who were historically disadvantaged. With the discrepancies in healthcare access and vulnerability of different ethnic groups and income groups, our model did address the importance of treating them fairly and improving their awareness of possible disease onset. However, from the perspective of the health insurance industry, our model and results could be unfair considering the possible usage of the model to exclude disadvantaged individuals from getting appropriate health insurance.

Strengths and limitations: Our study used the most up-to-date cycle of NHANES data to conduct the prediction of diabetes. With the advantage of using nationally representative data, our models, in nature, are more generalizable to the general adult population in the US. In addition, to the best of our knowledge, we were the first study that used NHANES data and machine learning algorithms to predict diabetes. Diabetes was a major public health and medical issue in the US. Our model could help with both predicting diabetes for public policy uses and the health insurance industry to minimize additional costs. The features included in our models were non-invasive measurements which lower the risk and cost for the prediction model to be used in the real-world setting. However, our study also had some limitations. Some of the features included in our models, such as physical activity, smoking, and alcohol consumption, were based on self-reported survey data which could introduce recall bias. Even though the NHANES data is nationally representative data, it is not necessarily representative enough for some subpopulations in the US. For example, previous studies have shown that Black and Hispanic Americans were more vulnerable to diabetes (19, 20). However, we included race/ethnicity as one of our predictive features which could potentially minimize this limitation. Moreover, our feature selection was based on the literature review which could neglect some of the important features measured by NHANES.

Future improvement: Moving forward, it might be helpful to include more features and conduct feature selection based on unsupervised machine learning approaches, such as principal component analysis, to capture novel features to advance the performance of the prediction models. Even though the sample size of our study was relatively large, we can also pool data from other NHANES year cycles and further increase our sample

size. With the “Large Sample Theory”, we will be able to obtain a better performance model. Moreover, since diabetes is a global health problem, we would like to test the performance of our model in similar nationally representative datasets in other countries such as KNHANES in Korea and CHARLS in China. The model could be further updated and possibly more generalizable and widely used in screening scenarios.

Conclusion

In conclusion, we consider our models and analyses can capture the relationship between the features and the disease outcome. Our model achieved a relatively good performance in predicting diabetes onsite in the aimed screening setting using non-invasive and easily obtained features. With the careful implementation of the model, we believe the model could be beneficial for diabetes prevention and early detection in the public health field. In the health insurance industry, our model could help determine the appropriate health insurance for individual best interests while reserving reasonable costs for the health insurance companies.

References:

1. Diabetes Quick Facts | Basics | Diabetes | CDC [Internet]. 2023 [cited 2024 Mar 12]. Available from: https://www.cdc.gov/diabetes/basics/quick-facts.html
2. Menke A, Casagrande S, Avilés-Santa ML, Cowie CC. Factors Associated With Being Unaware of Having Diabetes. Diabetes Care. 2017 Mar 13;40(5):e55–6.
3. National Diabetes Statistics Report | Diabetes | CDC [Internet]. 2023 [cited 2024 Mar 12]. Available from: https://www.cdc.gov/diabetes/data/statistics-report/index.html
4. NATIONAL CENTER FOR FARMWORKER HEALTH [Internet]. [cited 2024 Mar 12]. Diabetes Fact Sheet. Available from: https://www.ncfh.org/diabetes-fact-sheet.html
5. Parker ED, Lin J, Mahoney T, Ume N, Yang G, Gabbay RA, et al. Economic Costs of Diabetes in the U.S. in 2022. Diabetes Care. 2024 Jan 1;47(1):26–43.
6. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. Introduction and Methodology: Standards of Care in Diabetes—2023. Diabetes Care. 2022 Dec 12;46(Supplement_1):S1–4.
7. Ponce de León-Ballesteros G, Sánchez-Aguilar H, Aguilar-Salinas CA, Herrera MF. Improvement of the 10-Year Atherosclerotic Cardiovascular Disease (ASCVD) Risk Following Bariatric Surgery. Obes Surg. 2020 Oct;30(10):3997–4003.
8. Vassy JL, Posner DC, Ho YL, Gagnon DR, Galloway A, Tanukonda V, et al. Cardiovascular Disease Risk Assessment Using Traditional Risk Factors and Polygenic Risk Scores in the Million Veteran Program. JAMA Cardiol. 2023 Jun 1;8(6):564–74.
9. Ten-year risk assessment for cardiovascular diseases using ASCVD risk estimator plus: outcomes from hypertension and diabetes patients - PMC [Internet]. [cited 2024 Mar 12]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10604522/
10. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T. Current Techniques for Diabetes Prediction: Review and Case Study. Applied Sciences. 2019 Jan;9(21):4604.
11. Hasan MdK, Alam MdA, Das D, Hossain E, Hasan M. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access. 2020;8:76516–31.
12. Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Computer Science. 2018 Jan 1;132:1578–85.
13. NHANES - National Health and Nutrition Examination Survey Homepage [Internet]. 2023 [cited 2023 Nov 7]. Available from: https://www.cdc.gov/nchs/nhanes/index.htm
14. Scholes S, Bann D. Education-related disparities in reported physical activity during leisure-time, active transportation, and work among US adults: repeated cross-sectional analysis from the National Health and Nutrition Examination Surveys, 2007 to 2016. BMC Public Health. 2018 Jul 28;18:926.
15. Wang K, Zhao Y, Nie J, Xu H, Yu C, Wang S. Higher HEI-2015 Score Is Associated with Reduced Risk of Depression: Result from NHANES 2005–2016. Nutrients. 2021 Jan 25;13(2):348.
16. ALHarthi SSY, Natto ZS, Midle JB, Gyurko R, O'Neill R, Steffensen B. Association between time since quitting smoking and periodontitis in former smokers in the National Health and Nutrition Examination Surveys (NHANES) 2009 to 2012. J Periodontol. 2019 Jan;90(1):16–25.
17. Gay IC, Tran DT, Paquette DW. Alcohol intake and periodontitis in adults aged ≥30 years: NHANES 2009–2012. Journal of Periodontology. 2018;89(6):625–34.

18. Principles and practice of screening for disease. J R Coll Gen Pract. 1968 Oct;16(4):318. PMCID: PMC2236670.
19. Herman WH. The economic costs of diabetes: is it time for a new treatment paradigm? Diabetes Care. 2013 Apr;36(4):775-6. doi: 10.2337/dc13-0270. PMID: 23520368; PMCID: PMC3609514.
20. Cowie CC, Rust KF, Ford ES, et al. Full accounting of diabetes and pre-diabetes in the U.S. population in 1988-1994 and 2005-2006 [published correction appears in Diabetes Care. 2011 Oct;34(10):2338]. Diabetes Care. 2009;32(2):287-294. doi:10.2337/dc08-1296