DATA 4419 Second Report

Carter Silos
Gabriel Aguirre
Marvin Bernardino

Working Title:

Predicting Diabetes Risk Using NHANES 2021-2023 Data and Machine Learning Models

Introduction:

With a focus on Diabetes and lifestyle risk factors, we turn our attention to Fasting Plasma Glucose (FPG). The FPG test checks fasting blood glucose levels where a value greater than or equal to 126 mg/dL indicates a diagnosis of Diabetes (ADA, 2025). Other tests include the A1C test which measures average blood glucose for the past 2-3 months, and Oral Glucose Tolerance Test (OGTT) which measures blood glucose levels before and two hours after drinking a sweet solution. OGTT will require fasting and waiting for 2-3 hours, while the A1C test is the most convenient as it does not require fasting (ADA, 2025). Fasting Plasma Glucose (FPG) is available in the NHANES 2021-2023 Data as GLU_L, with a range of values of 59 to 561, with a count of 3672 and 324 missing (National Center for Health Statistics (NCHS) ,2024). The A1C test is also available under Glycohemoglobin (GHB_L) dataset with a range of 3.2 to 17.1 with a count of 6715 and 484 missing (NCSHS, 2024).

According to the CDC (2025), Diabetes risk factors for Type 1 Diabetes include family history and age (where it usually develops in younger people), in the US, white people are more likely to develop type 1 diabetes. Risk factors for Type 2 Diabetes include obesity or being overweight, age of 45 or older, family history, low physical activity, having non-alcoholic fatty liver disease, giving birth to a baby weighing over 9 pounds. Being African American, Hispanic or Latino, American Indian or Alaska Native and some Pacific Islander and Asian American people also have higher risk.

Table 1B. **Standards of Care in Diabetes Criteria for Diagnosis and Classification of Diabetes (ADA, 2025)**

Test type	Criteria
A1C	>= 6.5% >= (>= 48 mmol/mol)
FPG	>= 126 m/dL (>= 7.0 mmol/L)
2-h PG (OGTT)	>= 200 mg/dl (>= 11.1 mmol/L)
Symptoms of hyperglycemia	Random PG >= 200 mg/dL

Dataset and Variable Selection

The NHANES August 2021-August 2023 dataset was collected after the suspension of field operations due to the COVID-19 pandemic. This dataset is the latest available data as part of the ongoing National Health and Nutrition Examination Survey. The data is separated into Demographics, Dietary, Examination, Laboratory, and Questionnaire Data. Part of the data is also in limited access which deals with youth alcohol, drug, reproductive health and sexual behavior

questions (NCHS, 2024). Variables that may or not be related to Diabetes were selected from this dataset. Several potential response variables were also included: A1C (LBXGH), and FPG (LBXGLU) laboratory data test result, and in the questionnaire data, the "Doctor told you have diabetes" response (DIQ010). Potential predictors include well-known diabetes risk factors like gender (RIAGENDR), age (RIDAGEYR), race/Hispanic origin (RIDRETH1), Body Mass Index (BMXBMI), physical activity (PAQ706). Notable factors also include poverty (INDFMPIR), sugar (DR1TSIGR) and fat (DR1TTFAT) intake, alcohol (DR1TALCO).

Data Preprocessing:

For this project, the data set provided by NHANES and the CDC has combined multiple datasets about patients with their demographic, the questionnaire, diet, laboratory data, and examination data. All data sets were combined to make one data set based off of the shared variable SEQN, the patient's unique identifier. Once this was done the data had many variables that had a significant amount of null observations. The threshold that was set was 30%. This threshold was set to avoid biased models and inaccurate results. This was also to make sure any important variables were not removed in the process. According to a study done by MIT, the removal of variables with a high amount of missing observations improve accuracy and overall performance of models which is the goal of this project as well [1]. Similar methods were performed in this project to improve the accuracy within the models used.
After the removal of these variables, there were still missing observations in some of the remaining variables. The missing values remaining were transformed using imputation. The missing values were replaced using the mode values for each variable's observations. This is because most of the variables that had missing values were categorical variables. This was to avoid losing data and improve model accuracy while keeping the bias as low as possible. From the same study from Bertsimas, et al (2018), it was proven that imputation improved overall model accuracy and performance of the models. Similar methods of imputation were used in this project to help in the accuracy of the models performed in a later part of this project.

In this data set there are various adjustments that need to be made for clarity, usability and to remove redundant observations. Some variables only have 1 unique value or are observation identifiers like 'SEQN', which was used to join the datasets, and will have effect on modeling and predictions. Some categorical variables have labels such as 'I don't know' and 'refused' that might add noise and uncertainty to the data. If the number of observations marked with these labels are small and can therefore be removed as outliers to improve data clarity without increasing bias. Reviewing the Valued counts of the categorical data we can find values with less than 5 occurrences in variables (DIQ010, DMDEDUC2, DMDMARTZ, ALQ111, SMQ020, DPQ040, DPQ050, DPQ060). These values reference questionnaire answers where the participant did not know the answer or simply refused to answer. The last main issues with the dataset was distinguishing between categorical and numerical values in the raw data since all variables were numeric. Reviewing the data documentation shows that all categorical variables that we are using have less that 10 values. Knowing this a function is built that will look at the value counts of each variable and

check if the value is less than 10. If it is the data type will be changed to an object. This will let us better handle the data down the line.

Data Visualization:

Choosing Response Variable for Diabetes Diagnosis

According DIQ010 (Diabetes Questionnaire), 1081 out of 11744 survey respondents claimed in the questionnaire data that they had a diabetes diagnosis. In another category for a diabetes diagnosis, A1C test, shows that 721 out of 5994 respondents are positive for diabetes. Using another category, PFG test, shows that 430 out of 3242 responds are positive for diabetes. Due to the logistics of taking a PFG test which requires fasting for 8 hours, the A1C test maybe a more reliable way for assigning a diabetes diagnosis, in accordance with the standards of care for diabetes (ADA, 2025).

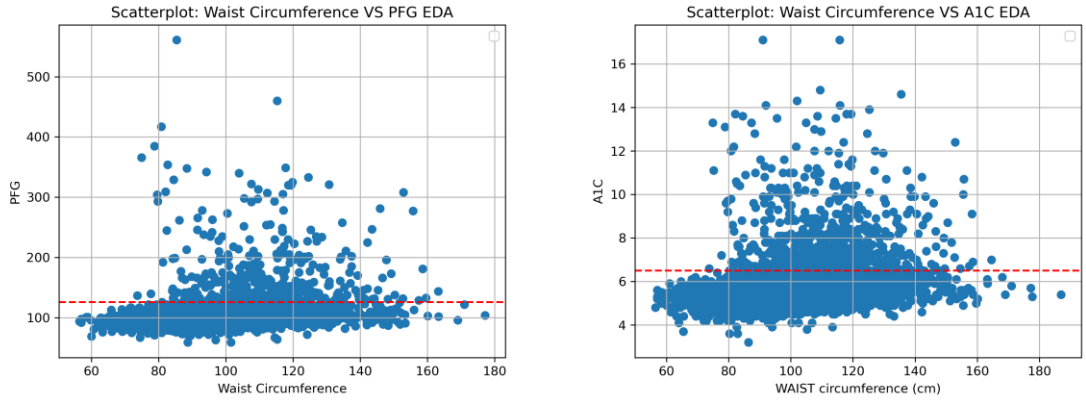
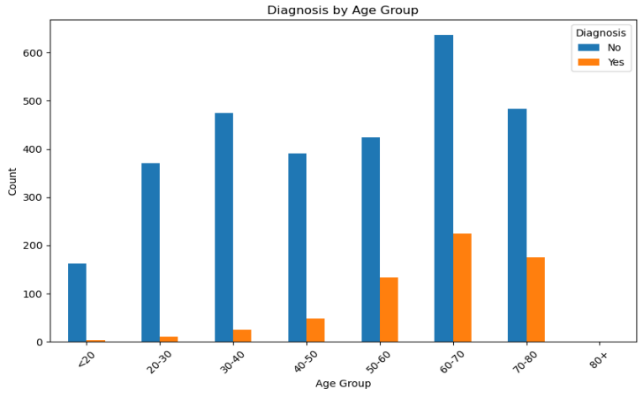


Figure 1.
A scatter plot of Waist Circumference versus PFG and A1C. Red dashed line shows the threshold where a diagnosis of diabetes would be warranted. (PFG ≥ 125 and A1C ≥ 6.5)

A quick scatter plot of waist circumference and PFG or A1C show that there is an imperceptibly better correlation when using the A1C test. Some respondents who were diagnosed before with diabetes but have now updated their lifestyle and have kept their glucose levels in check will show up as false positives in the DIQ010 questionnaire. Thus, owing to the logistics of fasting for a PFG test, and for the survey respondents who now have their diabetes under control, it is recommended that A1c test be used as the response variable.



This histogram visualizes the diagnosis of the patients compared to their age and the disparity of the data, which will have to be addressed in order to yield accurate results. This can be done by using resampling methods such as bootstrapping or random bagging. Figure 2

Class imbalance

Using A1C test, class balancing issues may arise due to 721 out of 5994 diabetes diagnosis. Qin et al (2022), explored the use of SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous). This will be explored as data cleaning is completed.

Table 2B. Detailed Variable Description

Dataset	Variable Name	Description	Variable type	Unique Values	Missing Values
Laboratory Data	**LBXGH**	**Glycohemoglobin (%) - A1C test**	Continuous	89	179
	LBXGLU	**Fasting Glucose (mg/dL) - FPG test**	Continuous	199	233
	LBXVIDMS	Vitamin D (25-hydroxyvitamin D2 +D3)	Continuous	888	282
Demographics Data	RIAGENDR	Gender	Categorical	2	0
	RIDAGEYR	Age in years at screening	Numerical	63	0
	RIDRETH3	Race/Hispanic origin	Categorical	6	0
	DMDBORN4	Country of birth	Categorical	2	0
	DMDYRUSR	Length of time in US	Categorical	6	2818
	DMDEDUC2	Education level - Adults 20+	Categorical	5	137
	DMDMARTZ	Marital status	Categorical	3	138
	RIDEXPRG	Pregnancy status at exam	Categorical		2921
	INDFMPIR	Ratio of family income to poverty	Continuous	416	479
Dietary Data	DR1TKCAL	Energy (kcal)	Continuous	1831	741
	DR1TPROT	Protein (gm)	Continuous	2657	741
	DR1TCARB	Carbohydrate (gm)	Continuous	2744	741
	DR1TSUGR	Total sugars (gm)	Continuous	2698	741
	DR1TFIBE	Dietary fiber (gm)	Continuous	739	741
	DR1TTFAT	Total fat (gm)	Continuous	2673	741
	DR1TSFAT	Total saturated fatty acids (gm)	Continuous	2722	741
	DR1TMFAT	Total monounsaturated fatty acids (gm)	Continuous	2768	741
	DR1TPFAT)	Total polyunsaturated fatty acids (gm	Continuous	2747	741
	DR1TCHOL	Cholesterol (mg)	Continuous	818	741
	DR1TVB12	Vitamin B12 (mcg)	Continuous	1266	741
	DR1TVC	Vitamin C (mg)	Continuous	1760	741

	DR1TMAGN	Magnesium (mg)	Continuous	607	741
	DR1TCAFF	Caffeine (mg)	Continuous	511	741
	DR1TSODI	Sodium (mg)	Continuous	2169	741
	DR1TALCO	Alcohol (gm)	Continuous	234	741
	DR1_320Z	Total plain water drank yesterday (gm)	Continuous	420	741
Examination Data	BPXOSY	Systolic - oscillometric reading (AVG)	Continuous	123	131
	BPXODI	Diastolic - oscillometric reading (AVG)	Continuous	83	134
	BMXBMI	Body Mass Index (kg/m**2)	Continuous	381	50
	BMXWAIST	Waist Circumference (cm)	Continuous	749	175
Questionnaire Data	**DIQ010**	**Doctor told you have diabetes**	**Categorical**	**2**	**0**
	ALQ111	Ever had a drink of any kind of alcohol	Categorical	2	513
	SMQ020	Smoked at least 100 cigarettes in life	Categorical	2	1
	SLD012	Sleep hours - weekdays or workdays	Numerical	23	31
	DPQ050	Poor appetite or overeating	Categorical	5	493
	DPQ060	Feeling bad about yourself	Categorical	5	494
	DPQ040	Feeling tired or having little energy	categorical	5	495

Tentative Schedule and Task Distribution:

- **Team Members / Chosen Method:**
  - Gabriel Aguirre: Support Vector Machines
  - Marvin Bernardino: Random Forest AND Logistic Regression (Added Random Forest for a more complex approach)
  - Carter Silos: XGBoost
- **Data Processing and Exploring (Feb 16 - Mar 15): COMPLETED**

  Task:
  - Identify missing values, detect outliers, apply necessary data cleaning and transformations
  - Perform Data visualization and summary statistics

  Deliverables:
  - Cleaned data
  - Data visualizations and summary statistics

- **Report 2 and GitHub (Mar 18 - Mar 30): COMPLETED**

  Notes: deadline extended to the 30th

  Task:
  - Compile and organize findings into second report (all members)
  - Upload code documentation to GitHub

  Deliverables:
  - Second Report Due
  - GitHub repository created and updated

- **Model Implementation and Evaluations (Mar 31 - Apr 12):**

Notes: The original timeline was an optimistic timeframe. Changes in the schedule and delays put this on a more reasonable timeline

Task:

- Develop chosen model (all members)
- Conduct model training/testing and evaluations (all members)

Deliverables:

- Initial models implemented and evaluated

- **Model Optimization and Comparative Analysis (Apr 12 - Apr 19):**

  Task**:**

  - Optimization of models to improve accuracy (all members)
  - Perform Comparative Analysis (ROC Curves) to identify performance and key findings (all members)

  Deliverables

  - Optimized models
  - Comparative analysis report

- **Final Report and Presentation (Apr 19 - Apr 26):**

  Task:

  - Compile final report with detailed model methodology, analysis and conclusions (all members)
  - Fully updated GitHub repository (if not already done)
  - Develop presentation slideshow (all members)
  - Ensure all members understand all components of the project

  Deliverables:

  - Final Project report (Apr 26)
  - Finalized GitHub repository
  - Class Presentation (TBD)

Github Repository:

https://github.com/DATA-4419-Project/Machine-Learning-Diabetes-Modeling

References

1. American Diabetes Association (2025). *Diabetes Diagnosis & Tests*. Understanding Diabetes Diagnosis. https://diabetes.org/about-diabetes/diagnosis
2. National Center for Health Statistics (2024, September). U.S. Center for Disease Control and Prevention. *National Health and Nutrition Examination Survey: Plasma Fasting Glucose (GLU_L)* https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/GLU_L.htm
3. National Center for Health Statistics (2024, September). U.S. Center for Disease Control and Prevention. *National Health and Nutrition Examination Survey*. Glycohemoglobin (GHB_L) https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/GHB_L.htm
4. U.S. Center for Disease Control and Prevention (2024, May 15). *Diabetes Risk Factors*. https://www.cdc.gov/diabetes/risk-factors/index.html
5. Bertsimas, D., Pawlowski, C. Ying, D. (2018). Journal of Machine Learning Research. *From Predictive Methods to Missing Data Imputation: An Optimization Approach*. *https://* jmlr2020.csail.mit.edu/papers/volume18/17-073/17-073.pdf
6. American Diabetes Association. (2025, January). *Standards of Care in Diabetes - 2025*. Diabetes Care Volume 48, Supplement 1, January 2025. https://diabetesjournals.org/care/article/48/Supplement_1/S27/157566/2-Diagnosis-and-Classification-of-Diabetes

7. Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B. Yu, J., Li, C., Yu, F., Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. International journal of environmental research and public health, 19(22), 15027. https://pmc.ncbi.nlm.nih.gov/articles/PMC9690067/