Machine Learning-Based Prediction of Prediabetes Risk in Women Using Post-Pandemic NHANES Data (2021-2023)

By Saranna Lay
AOS C111 Machine Learning for Physical Sciences
Alexander Lozinski

1. Introduction

The growing prevalence of Type 2 Diabetes and its precursor, prediabetes, represents a significant and growing public health challenge in the United States that has touched many. Prediabetes is characterized by elevated blood glucose levels that do not yet meet the threshold for type 2 diabetes (CDC, 2024). It affects millions of individuals, some of whom don't even know of their prediabetes status, placing them at heightened risk for developing the full-blown disease and associated comorbidities. For individuals who plan to carry children, diabetes before and during pregnancy may present complications for both mother and baby during pregnancy and birth (*Pregnancy If You Have Diabetes - NIDDK*, n.d.). This makes early and accurate detection critical for timely lifestyle interventions, which align with current medical guidance to delay or prevent disease progression.

To address the need for improved diabetes screening tools, my project uses machine learning classification techniques to predict the risk of prediabetes or undiagnosed diabetes in individuals with childbearing potential. The analysis uses data from the National Health and Nutrition Examination Survey (NHANES), a program operated by the Centers for Disease Control and Prevention (CDC) that assesses the health and nutritional status of adults and children across the United States. This study uses publicly available NHANES Body Measures (BMX) data, which includes Body Mass Index (BMI) and Waist Circumference. Doctors use these as primary screening flags for diabetes risk and subsequent lab follow-up. A BMI of 25 kg/m^2 represents overweight, and equal to or greater than 30 kg/m^2 is considered obese, both of which are flagged as at risk for diabetes (CDC, 2024). Because BMI is not always reliable as a screening tool for populations like bodybuilders, whose low fat percentage and high muscle mass put them at a high BMI, doctors also consider waist circumference. Visceral fat is recognized as a stronger predictor of metabolic issues than BMI alone. Waist circumference criteria vary by race, so for simplicity, the project will focus on clinical guidelines established for female-assigned bodies. For women, a waist circumference of 88 cm (35 in) or greater is an indicator for diabetes risk (CDC, 2024). These body measurements, combined with Plasma Fasting Glucose (GLU) data, serve as the features used to train and test the created models.

The project is motivated by the potential effects of the COVID-19 pandemic on prediabetes risk factors in this specific cohort across races. Given that the pandemic period encouraged

widespread sedentary behaviors and altered dietary patterns due to fears and restrictions, this project focuses on the mid- to post-pandemic pandemic, from August 2021 to August 2023. By analyzing data, the project aims to identify and classify patterns associated with increased risk of prediabetes, thereby providing timely insights into the shifts in population health following a major global event.

A necessary limitation of this project is the definition of the childbearing cohort. Due to time and publicly available data constraints, this group is identified exclusively using the "Female" gender marker in the NHANES dataset. This approach is used to cover a statistically large majority of those with childbearing potential, but it is not intended to endorse exclusionary clinical or social definitions.

The resulting models will serve as tools for identifying individuals who may want further clinical evaluation to determine their diabetes status. The project intends to help all those classified by these metrics screen for diabetes, especially individuals who hope to bear children in the future.

2. Methods

Data Preprocessing

The variables of interest in the project come from two separate NHANES datasets during the August 2021 to August 2023 period: DEMO_L, BMX_L, and GLU_L.

1. Merge Datasets

The datasets were first matched based on the"SEQN" found in all datasets and then merged. DEMO_L had 11,933 observations. BMX_L originally had 8,860 observations. The resulting merged dataset matched 3,996 rows to the GLU_L dataset.

2. Feature Selection

The merged dataset originally contained 51 columns. A majority of columns were dropped due to irrelevance to the analysis or redundancy. This initial feature cleaning reduced the final dataset to include only gender, relevant body measurements, and fasting glucose columns. The selected columns are as follows.

Column Name	Description	Source Dataset
RIAGENDR	Gender	DEMO_L
LBXGLU	Fasting Glucose (mg/dL)	GLU_L
BMXBMI	Body Mass Index (kg/m^2)	BMX_L
BMXWAIST	Waist Circumference (cm)	BMX_L

3. Narrow to "Female" Population

Using the 'RIAGENDR' column, I dropped rows containing 1 (indicating Male) and NA values. After this process, the cleaned dataset contained 2196 observations (1,800 rows lost).

4. Handle Missing Values

Since body measurement and glucose data are central to defining the dependent variable (prediabetes/undiagnosed diabetes status), rows with missing values for these features were removed. After this process, the final cleaned dataset contained 1916 observations **(n = 1916)**.

5. Define the Target Variable

Based on established medical thresholds for fasting glucose and their corresponding classifications of diabetes status, the LBXGLU column was used to create a new binary variable, "glu_status", recorded as:

- 0: Normal range (LBXGLU < 100 mg/dL)

- 1: Risk of Prediabetes/Diabetes range (LBXGLU ≥ 100 mg/dL)

Another binary variable “waist_status” was created to indicate whether a participant's waist circumference (from the BMXWAIST column) exceeded the clinical threshold of 88 cm for women.

- 0: Normal (BMXWAIST < 88 cm)
- 1: Concern (BMXWAIST > 88 cm)

A third binary variable “bmi_status” was created to indicate whether a participant’s BMI is classified as Normal, Overweight, or Obese.

- 0: Underweight (BMI < 18.5 kg/m^2)
- 0: Normal (BMI 18.5 - 24.9 kg/m^2)
- 1: Overweight (BMI 25.0 - 29.9 kg/m^2)
- 2: Obese (BMI ≥ 30 kg/m^2)

The three variables are then combined to create the final analyzed diabetes risk score, “final_status”, which is the sum of “glu_status,” “waist_status,” and “bmi_status.” The following interpretations could be made:

- 0 - 1: Normal, low risk for diabetes based on analysis
- 2 - 3: Potential risk for diabetes
- 4: Highest risk for diabetes, clinical follow-up suggested

6. *Prepare for Model Training*

Prior to model training, the data were split into training (n = 1532) and testing (n = 384) sets. Numerical features were then scaled to prevent variables with larger magnitudes from influencing the learning process. The `StandardScalar` from the scikit-learn library was used. Z-score normalization here standardizes features by removing the mean and scaling to unit variance. The scaler fit was only on the training data. The resulting scaled parameters were then applied to the testing set to ensure that the test data remained unseen for the evaluation of the final model.

Exploratory Data Analysis

Figure 1: Distribution of Features ('BMXBMI', 'BMXWAIST', and 'LBXGLU')

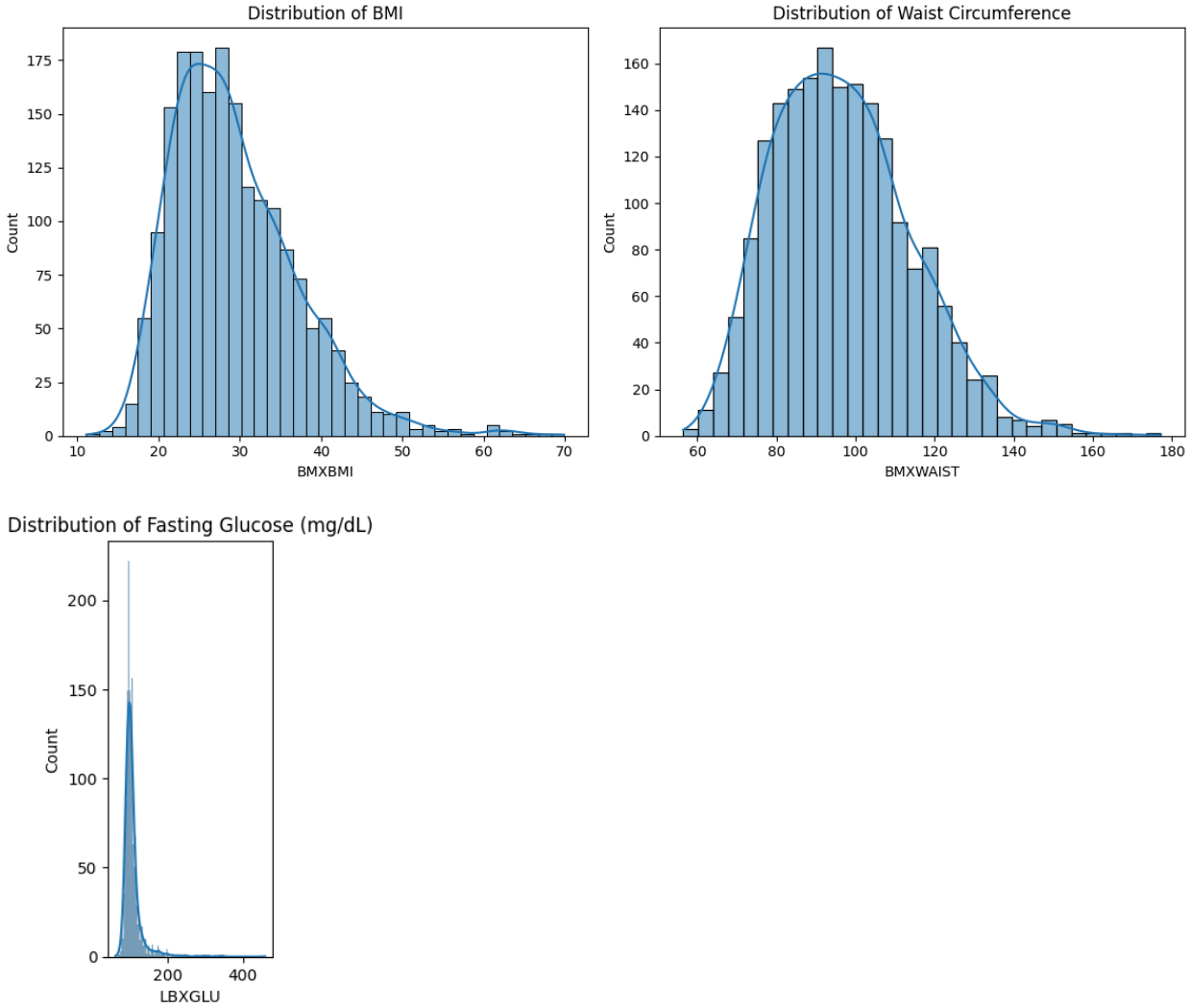

Figure 1 shows that BMI, Waist Circumference, and Fasting Glucose are generally normally distributed, though slightly right-skewed. This confirms that data transformation is not necessary for the data. Transformation may also decrease the interpretability of the final model.

Figure 2: Distribution of Target Variable, Diabetes Risk Status ('final_status')

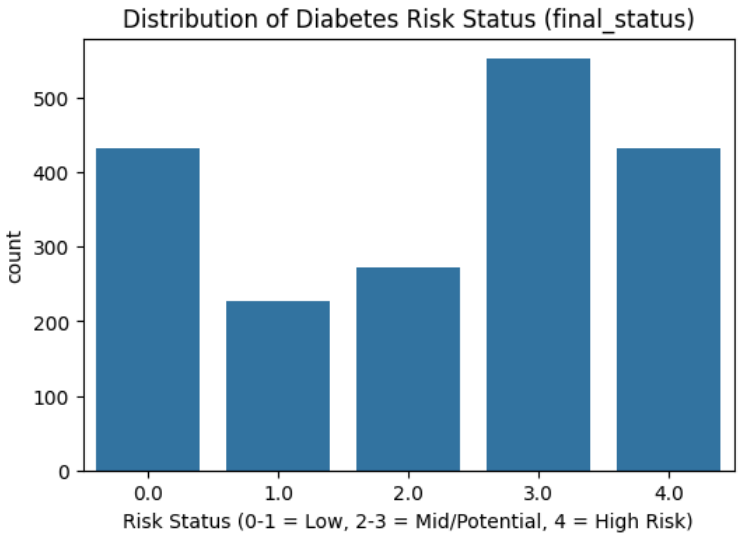

Figure 2 shows that a Risk Status of 0-1 (low risk) has the fewest observations, which may lead to training bias. This can lead the model to be overly cautious when predicting low risk, or it may struggle to distinguish between low and mid-level risk.

Figure 3: Feature Distribution by Target Variable

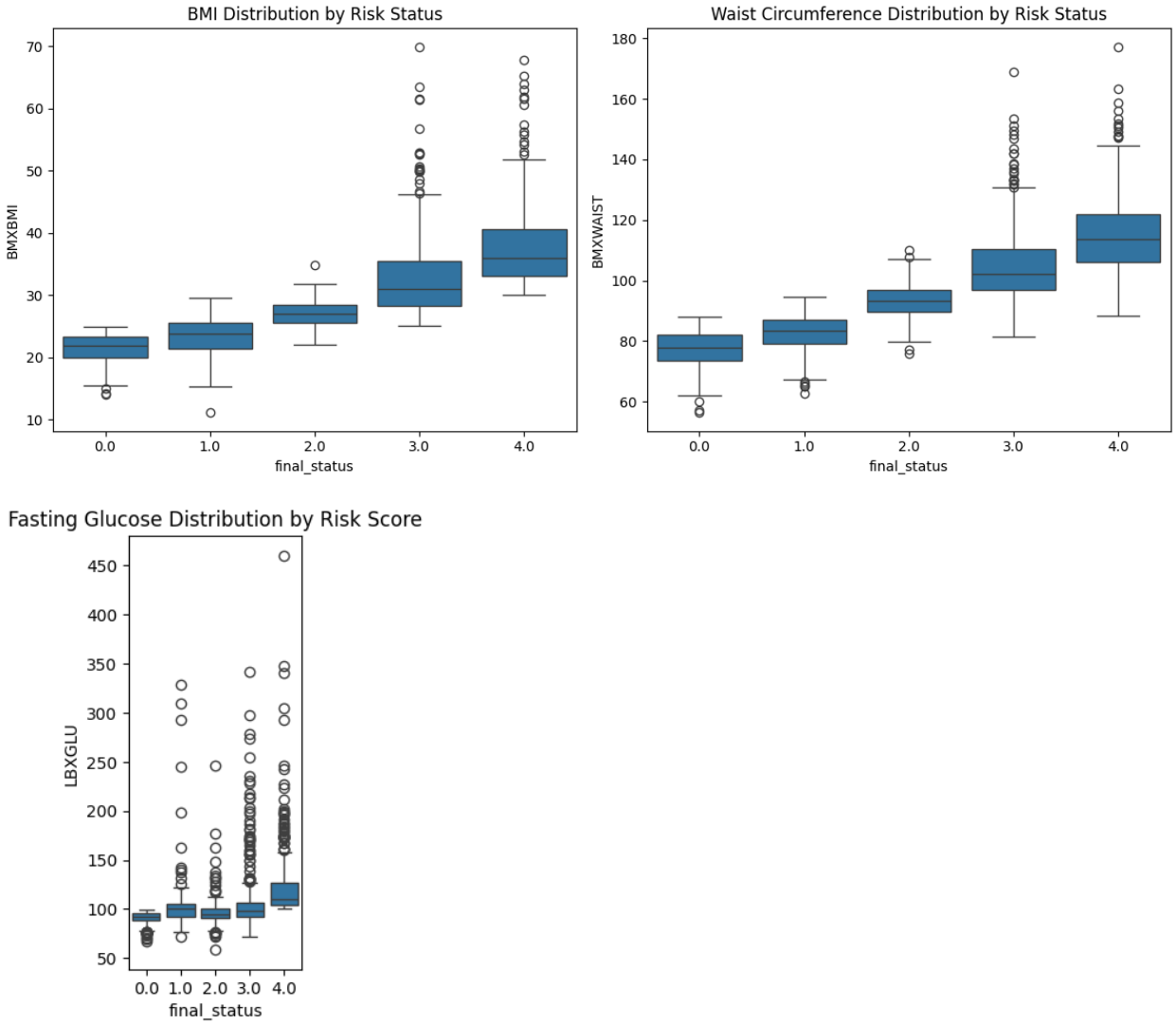

Figure 3 confirms a monotonic relationship. As risk status increases, the median BMI, Waist Circumference, and Fasting Glucose levels also increase. This is expected since risk status is the sum of the three features. This confirms that our data preprocessing steps were carried out as intended.

3. Model Training, Evaluation, and Interpretation

Multinomial Logistic Regression

A multinomial logistic regression model was trained to predict the three tiers of diabetes risk score in "final_status" (Risk Levels 0 to 4) using the scaled features BMXBMI, BMXWAIST, and LBXGLU as features and evaluated using the unseen test data. This type of model was used since the target variable has five categories. Logistic regression also quantifies predictors, allowing for easy interpretation, which is fitting for a screening tool.

The logistic regression model achieved an overall mediocre accuracy of 68% in predicting diabetes risk status on the testing data. Analysis of class-specific metrics showed varied performance across the risk categories. The model showed the best performance in identifying Low Risk individuals (Risk Level 0), achieving a high recall of 91%. This means the model is highly effective at correctly classifying those who are truly at low risk, minimizing the number of unnecessary false alarms. However, the model showed its weakest performance with Risk Level 1 having an F_1 score of only 0.49, suggesting difficulty in accurately distinguishing this level from neighboring classes. The model's performance on the Highest Risk group (Risk Level 4) was moderate, with a precision of 71% and a recall of 60%. While 71% of individuals flagged as Highest Risk were correctly classified, the low recall means that the model failed to identify 40% of the true high-risk individuals, resulting in a significant number of false negatives in the most critical category. Analysis of the confusion matrix shows that the model had a strong tendency to downgrade the risk level. This is seen in how the model classified 34 individuals truly in the Risk Level 4 category to Risk Level 3.

Figure 4: Multinomial Logistic Regression Classification Report

	precision	recall	f1-score	support
0.0	0.81	0.91	0.85	87
1.0	0.59	0.42	0.49	45
2.0	0.62	0.75	0.68	55
3.0	0.63	0.65	0.64	111
4.0	0.71	0.60	0.65	86
accuracy			0.68	384
macro avg	0.67	0.67	0.66	384
weighted avg	0.68	0.68	0.68	384

Random Forest Model

A random forest model was trained to predict the five tiers of diabetes risk score as well. This model handles class imbalances well using the parameter `class_weight = 'balanced'`. The model also ranks the importance of features to aid in interpretation.

The Random Forest model achieved a great overall accuracy of 99% on the testing data, greatly outperforming the 68% accuracy of the Multinomial Logistic Regression model. The classification report showed a near-perfect performance across all five risk categories, with F_1 scores of 0.98 or higher in every class. This high level of performance is confirmed by the confusion matrix, which shows that very few misclassifications were present. The model achieved perfect Recall for the Highest Risk group (Risk Level 4), successfully identifying every high-risk individual.

The Feature Importance analysis strongly suggested that BMI is the dominant predictor, contributing approximately 39.4% of the model's decision power, which contradicts my initial hypothesis that BMI is too generalizing to various body types. The other two features, Fasting Glucose and Waist Circumference, contribute about the same at 30.4% and 30.2% respectively. While all metrics are strong predictors of diabetes, this finding shows the significance of general obesity in addition to visceral fat and fasting glucose levels in defining overall diabetes risk.

Figure 5: Random Forest Classification Report

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	87
1.0	0.96	1.00	0.98	45
2.0	0.98	0.98	0.98	55
3.0	1.00	0.99	1.00	111
4.0	1.00	1.00	1.00	86
accuracy			0.99	384
macro avg	0.99	0.99	0.99	384
weighted avg	0.99	0.99	0.99	384

4. Results and Conclusion

The final analyzed dataset contained 1,916 female respondents with complete records for the core risk features. The target variable ‘final_status’ (a cumulative risk score ranging from 0-4) was created to reflect five risk tiers.

Two classification models were tested to predict diabetes risk status: Multinomial Logistic Regression and Random Forest.

The Multinomial Logistic Regression model achieved an overall mediocre accuracy of 68% in predicting the risk status on the testing data. The Random Forest Classifier dramatically outperformed the linear model, achieving an overall accuracy of 99%. This model eliminated the weaknesses of the Logistic Regression model, achieving F_1 scores greater than or equal to 0.98. The Feature Importance analysis from the Random Forest model identified BMI as the dominant predictor of diabetes risk.

This project successfully developed and validated a classification model for diabetes risk status using three standard clinical measurements. The Random Forest Classifier ultimately proved to be the superior predictive model, achieving an accuracy of 99%. The high performance of the model suggests that the chosen features are highly effective at separating the five tiers of the risk levels in this dataset. The resulting model can be used as a screening tool for settings without easy access to clinicians, for example.

5. Limitations

The conclusions should be interpreted in light of several key limitations. The generalizability of the cohort is constrained since I was unable to exclude individuals with a prior clinical diagnosis of diabetes, which may have distorted the model's ability to identify patterns specific to undiagnosed risk. The analysis also did not account for pregnancy status, potentially introducing noise into the features and compromising the accuracy of the risk assessment.

Another necessary limitation of this project is the definition of the childbearing cohort. Due to time and publicly available data constraints, this group is identified exclusively using the "Female" gender marker in the NHANES dataset. This approach is used to cover a statistically large majority of those with childbearing potential, but it is not intended to endorse exclusionary clinical or social definitions.

Despite the model's high predictive power, it relies exclusively on three variables (BMXBMI, BMXWAIST, and LBXGLU). These simplifications neglect the complex nature of diabetes risk, which is also influenced by genetics, age, and lifestyle factors. Future projects would benefit from expanding the features to include other standard clinical risk factors, such as HbA1c or cholesterol levels, to provide a fuller picture of metabolic risk.

6. References

CDC. (2024a, June 5). *Healthy Weight*. Diabetes. https://www.cdc.gov/diabetes/living-with/healthy-weight.html

CDC. (2024b, December 6). *About Prediabetes and Type 2 Diabetes*. National Diabetes Prevention Program. https://www.cdc.gov/diabetes-prevention/about-prediabetes-type-2/index.html

Pregnancy if You Have Diabetes—NIDDK. (n.d.). National Institute of Diabetes and Digestive and Kidney Diseases. Retrieved December 5, 2025, from https://www.niddk.nih.gov/health-information/diabetes/diabetes-pregnancy