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

Marvin Bernardino, Gabriel Aguirre, Carter Silos

Introduction:

Diabetes is a disease caused by the body's inability to manage the body's main source of energy, glucose. If a person has diabetes, the body does not make enough, or any insulin. Insulin helps glucose to be absorbed into the cells to be used for energy. Having diabetes risks damage to the eyes, kidneys, nerves, and heart (National Institute of Diabetes and Digestive and Kidney Diseases, 2023)[1].

In 2021, 38.4 million Americans or 11.6% of the population, had diabetes. In the same year, diabetes was also the eighth leading cause of death in the United States. Of the 38.4 million adults with diabetes, 29.7 million were diagnosed, and 8.7 million were undiagnosed (American Diabetes Association, 2023)[2]. The aim of this study is to explore indicators that would help in diagnosing diabetes. This project strives to perform statistical and machine learning techniques using the National Health and Nutrition Examination Survey (NHANES) dataset.

Literature Review:

Using the 2005-2016 NHANES data, Vangeepuram, et al. (2021)[13] analyzed 2970 participants aged 12-19 years to investigate youth diabetes risk through machine learning approaches. To decide which participants are at risk for diabetes, they used the American Diabetes Association (ADA) and the American Academy of Pediatrics (APA) guidelines. Some of these risk factors are being overweight, family history of type 2 diabetes in first or second-degree relatives, race/ethnicity, signs of insulin resistance or conditions associated with insulin resistance (hypertension, dyslipidemia). Incidentally, these risk factors appear in the complete NHANES data. Using the five variables listed above, the machine learning methods Vangeepuram, et al. (2021)[14] explored included ten established algorithms, and a five-fold cross-validation setup. The classification methods considered are: AdaBoost(M1), LogitBoost, Naive Bayes, Logistic(Regression). Support Vector Machine (SMO), Voted Perceptron, K-nearest neighbor (IBk), PART and J48 decision tree inference algorithms, and Random Forest. The study results show that the naive Bayes-based classifier performed better than the APA/ADA screening guideline (Vangeepuram, et al., 2021)[13].

Another study by Badger and Abuwarda (2024)[15], also utilized the NHANES data from 1988 to 2018 to predict the future onset of Type II Diabetes. The machine learning models used include Logistic Regression, Support Vector Machines(SVM), Random Forest, and XGBoost. They were able to achieve an Area Under-Receiver Operating Characteristics (AU-ROC) metric of 0.9772 for diabetic and non-diabetic patients, and 0.9806 for undiagnosed diabetic and pre-diabetic patients.

Dinh et al. (2019)[16] concludes in their study that "machine learning based on survey questionnaires can provide an automated identification mechanism for patients at risk of diabetes…" Using NHANES data from 1999-2014 consisting of 123-168 variables, Dinh et al. comes up with a top 24 features for diabetes classifiers. The most significant ones are: blood osmolality and sodium, followed by blood urea, nitrogen, triglyceride, and LDL cholesterol, for those that included lab results. For those instances without lab results, the most significant ones are waist, age, self-reported greatest weight, leg length, sodium intake. The machine learning

model used with the best AU-ROC was eXtreme Gradient Boost (XGBoost). Other machine learning models used were: Logistic Regression, Support Vector Machines (SVM), Random Forest Classifier (RFC), Gradient Boosted Trees (GBT), and Weighted Ensemble Model (WEM). Another study by Qin, et al. (2022)[17] seeks to predict diabetes by lifestyle type. The 1999-2020 NHANES dataset was used, this time, focused on the lifestyle questionnaire. The machine learning classifiers used were XGBoost, CATboost, SVM, Random Forest, and Logistic Regression. These five machine learning models identified dietary intake levels of energy, carbohydrate and fat contributed the most to the prediction of diabetes patients. The highest obtain AUC-ROC curve was 0.83 for CATBoost.

Data:

The CDC's National Center for Health Statistics (NCHS) conducts the NHANES. This is the only national health survey that includes health exams, laboratory tests, and dietary interviews for participants of all ages. Since 1999, the NCHS surveys each year, about 5000 adults and children in communities across the United States (NHANES, 2024)[3]. The NHANES uses a random sampling of households in the United States based on their similarity using US Census information. The survey involves the participants answering health questions, visiting a mobile exam center for health and lab tests, and answering questions about what they eat. The health exam may include height, weight, and other body measures, blood pressure reading, dental exam, vision and hearing tests, and laboratory tests for kidney and liver health. (NHANES, 2024)[4].

For the purposes of this study, the NHANES is a unique and useful data source as it combines health and interview questions along with actual laboratory tests. The NHANES data files and related documentation are also publicly available to download on their website. The research will be conducted using the latest NHANES data collected during August 2021 - August 2023.

The NHANES 8/2021 - 8/2023 data (NHANES, 2024)[5] is separated into 6 datasets:

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

The **NHANES Demographic Data** (NHANES, 2024)[6] contains basic demographic information including the participant's gender, age, race/Hispanic origin, country of birth, education level, pregnancy and marital status, and ratio of family income to the federal poverty line. More importantly, the NHANES Demographic Data also includes the respondent's unique identifier which can be used to link this dataset with other datasets included in the NHANES.

The **NHANES Dietary Data** (NHANES, 2024)[7] contains information on what the participant eats. This data includes grams, energy (in kcal), protein, carbohydrate, total sugars, total fat, cholesterol, and many others.

The **NHANES Examination Data** (NHANES, 2024)[8] contains information about balance tests, blood pressure readings, liver ultrasound tests, and more importantly, body measures. The body measures dataset contains height, and weight measurements, including finer details like arm, leg, hip, waist circumference measurements (NHANES, 2024)[9].

The **NHANES Laboratory Data** (NHANES, 2024)[10] contains information about routine blood laboratory work like Complete Blood Count (CBC), Comprehensive Metabolic Panel (CMP), Lipid Panel. Absent in this list is the A1C test which is one of the primary indicators for a diabetes diagnosis. Present in this dataset however is insulin, and plasma fasting glucose which will be more useful for signifying pre-diabetes.

The **NHANES Questionnaire Data** (NHANES, 2024)[11] contains information about the participant's lifestyle, it also contains questions about self-reported diabetes which will be useful for classification for later analysis. Other datasets include alcohol use, diet behavior and nutrition, early childhood weight, weight history, depression, physical activity, sleep disorders and smoking use.

The **NHANES Limited Access Data** (NHANES, 2024)[12] contains sensitive information that is not available to the public except through secure, on-site access. This includes youth alcohol use, drug use, reproductive health and sexual behavior questions for youth and adult respondents.

Table 1. Detailed Variable Description

DATA		
Laboratory Data	GHB_L	Glycohemoglobin (A1C)
Dietary Data	DR1TKCAL	Energy (kcal)
	DR1TPROT	Protein (gm)
	DR1TCARB	Carbohydrate (gm)
	DR1TSUGR	Total sugars (gm)
	DR1TFIBE	Dietary fiber (gm)
	DR1TTFAT	Total fat (gm)
	DR1TSFAT	Total saturated fatty acids (gm)
	DR1TMFAT	Total monounsaturated fatty acids (gm)
	DR1TPFAT	Total polyunsaturated fatty acids (gm

	DR1TCHOL	Cholesterol (mg)
	DR1TVB12	Vitamin B12 (mcg)
	DR1TVC	Vitamin C (mg)
	DR1TMAGN	Magnesium (mg)
	DR1TCAFF	Caffeine (mg)
	DR1TSODI	Sodium (mg)
	DR1TALCO	Alcohol (gm)
	DR1_320Z	Total plain water drank yesterday (gm)
Demographics Data	RIAGENDR	Gender
	RIDAGEYR	Age in years at screening
	RIDRETH1	Race/Hispanic origin
	DMDBORN4	Country of birth
	DMDYRUSR	Length of time in US
	DMDEDUC2	Education level - Adults 20+
	DMDMARTZ	Marital status
	RIDEXPRG	Pregnancy status at exam
	INDFMPIR	Ratio of family income to poverty
Examination Data	BPXOSY	Systolic - oscillometric reading (AVG)
	BPXODI	Diastolic - oscillometric reading (AVG)
	BMXBMI	Body Mass Index (kg/m**2)
	BMXWAIST	Waist Circumference (cm)
Questionnaire Data	DIQ010	Doctor told you have diabetes
	ALQ111	Ever had a drink of any kind of alcohol
	PAD790Q	Frequency of moderate LTPA
	PAD680	Minutes sedentary activity
	SMQ020	Smoked at least 100 cigarettes in life
	SLD012	Sleep hours - weekdays or workdays
	DPQ010	Have little interest in doing things
	DPQ040	Feeling tired or having little energy
	DPQ050	Poor appetite or overeating
	DPQ060	Feeling bad about yourself

Data Preprocessing:

Choosing a Response Variable

According to the American Diabetes Association Standards of Care in Diabetes, diabetes may be diagnosed by several criteria.

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

Table 2. **Standards of Care in Diabetes Criteria for Diagnosis and Classification of Diabetes (ADA, 2025)[29].**

FPG (Fasting Plasma Glucose) and the Glycohemoglobin Test (A1C) are both present in the NHANES data. The FPG test requires participants to fast for 8 hours, representing an extra hurdle in the collection of data. The A1C test however, does not require fasting and this is reflected in its abundance in the dataset (fewer missing data and double the number of tests). An A1C test result of 6.5 or greater is an indication of diabetes. Transforming this variable allows the use of this data for machine learning classification.

Transformation and Creation of Features

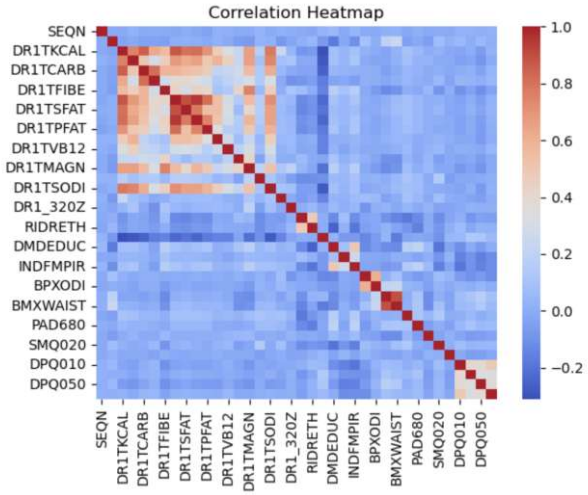
After cleaning null values, outliers and incomplete features, and creating several dummy variables for the various categories including those predictors for Education, Gender, Natural-born status, Marital status, and various depression-related questions, the final variable count is: 3520 observations and 36 features. For the response variable, A1C test, there are 398 positive cases versus 3122 negative cases (12% vs 88%). Upon application of SMOTE-NC, this is balanced to 3122 for positive cases, and 3122 negative cases for negative for an even split.

SMOTE-NC

To deal with the class imbalance, this study adopts SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) to deal with the unbalanced class data. SMOTE-NC is based on SMOTE which is a technique that calculates the Euclidean distance of data points in the feature space based on the nearest neighbor (Qin, et al., 2022) [17]. SMOTE is based on a method of over-sampling the minority class and under-sampling the majority class which improves classifier performance in the ROC space (Chawla, et al., 2002)[28]. Unlike SMOTE, SMOTE-NC is used for datasets containing numerical and categorical features.

Exploratory Data Analysis

A heat map was constructed to show the correlation between the variables. Values closer to 1 indicated a positive correlation between variables while values closer to –1 have a negative correlation. Values that are close to or are 0 have no correlation at all. The variables that have the most positive correlation are total monosaturated fats consumed and total fat consumed. The variables that have the most negative correlation are gender and calories consumed. The variables that have no correlation are total saturated fatty acids consumed and the patient's unique identifier.



Methods:

Logistic Regression:
Four machine learning classifiers will be utilized. Firstly, logistic regression. This method of machine learning takes variables and predicts a probability for a binary outcome of 0 (does not have diabetes) or 1 (has diabetes). Being the simplest, and easiest to perform, this will serve as the baseline for the other models.

Support Vector Machines
Another method is Support Vector Machines (SVM) which is a supervised learning algorithm that identifies the optimal hyperplane to serve as a decision boundary between different classes, making it suitable for binary classification problems such as diabetes prediction. The goal is to maximize the distance between the hyperplane and the closest data points (support vectors) which make up the margins to improve performance. SVM's usefulness comes from being able to handle data with varying degrees of complexity through the use of kernel functions. These kernel functions (linear, polynomial, radial basis function) are used to transform the feature space into a 3D space to provide better linear separation with a hyperplane (Geeks for Geeks, 2025)[19]. Outliers will also pose less of an issue because SVM relies on support vectors for the margins and hyperplane.

Random Forest
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (developers, 2025) (scikit-learn.org) [26]. It has reduced risk of overfitting, and easy to determine feature importance (IBM)[27]. It does not perform well on unbalanced data. Random forest is one the best performing machine learning models that were used in the previous study by Qin, et al (2022)[17].

XGBoost
The final machine learning classifier is XGBoost (Extreme Gradient Boosting). XGBoost is an open-source software library based on gradient boosted trees, used for supervised learning

problems [22]. XGBoost utilizes decision trees and combines them sequentially which in turn enhances the performance. These trees are trained to correct errors previously made in other trees. This method is known as boosting. What makes this method “extreme” is the inclusion of regularization elements which are variables that correct overfitting[25]. XGBoost shows its utility with its accuracy of the predictions and capability of handling large data sets. This model also provides results and predictions quickly.[23] As shown in “Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction” by Kurshid et al. , the XGBoost model generally outperformed all other methods when predicting whether a person will develop diabetes, which makes this a reliable methodology to use in the research of this project.[24]

Results:

Logistic Regression

Logistic Regression from sci-kit learn was used with saga as solver and l1 as penalty for lasso regularization. Max iteration was changed to 10,000 as the model did not converge using the default of 1000.

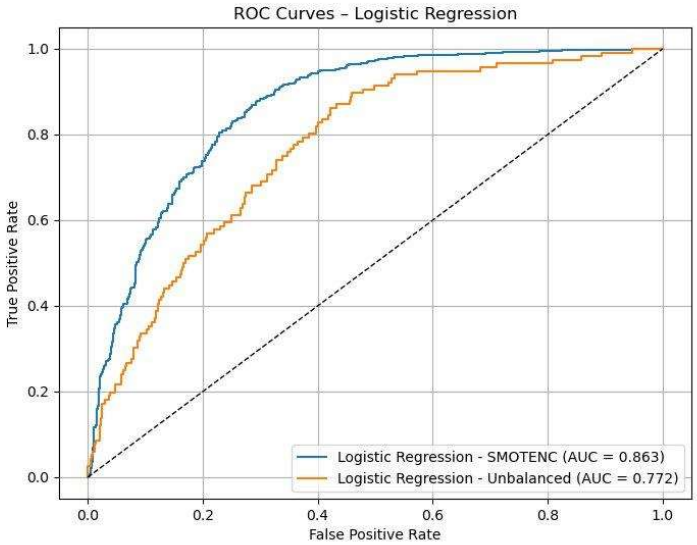


Figure 2. Logistic Regression for Balanced (using SMOTE-NC) and Unbalanced Data.

Using K=10 cross validation, an accuracy of 89% was obtained using unbalanced data. However, this is misleading as it has also produced a recall score of 9% and average precision of 51%. Using the data balanced by SMOTE-NC, the final K=10 cross validation accuracy score drops to 79% but with an increase in recall and average precision leading to a higher AUC-ROC curve score of 0.863. This was obtained from the test set.

Random Forest

Scikit-Learn Random Forest Classifier was used this time, along with K=10 cross validation. No tuning parameters were used, and as expected, an AUC-ROC curve of 1 was the result for training test. For the test set, however, we obtain 0.731. Using data balanced by SMOTE-NC, an AUC-ROC curve of 0.979 was obtained since it fixes the precision and recall obtained by the unbalanced data. The AUC and ROC was obtained from the test set.

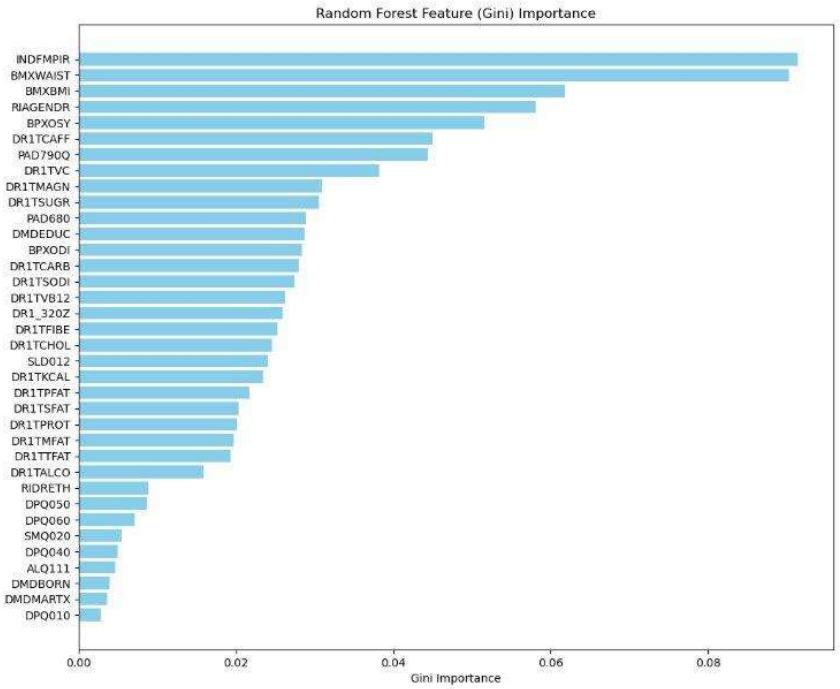


Figure. 3: Random Forest Feature Importance (Gini Index) using Balanced Data

Among the most important features for diabetes prediction are ratio of family income to poverty, waist circumference, body mass index, gender (being female), blood pressure, frequency of moderate LTPA (Leisure Time Physical Activity), sugar intake, and minutes Sedentary Activity which are well characterized in literature. Important features that are predicted by the model but are not classically linked to diabetes are: caffeine intake, vitamin C intake, magnesium intake, and education level.

XGBoost

Using the xgboost package in python, the XGBoost model was ran and optimized using hyperparameter tuning. The hyperparameters were tuned using Bayesian optimization. The first

iteration of the model was run without using cross validation or hyperparameter tuning which yielded a model AUC of 71%. Using 10-fold cross validation and hyperparameter tuning resulted in a higher model AUC of 78.6%. The model AUC increased slightly, but due to the imbalance of the data, it was not able to optimize as well as it could have. The AUC and ROC were obtained from the test set.

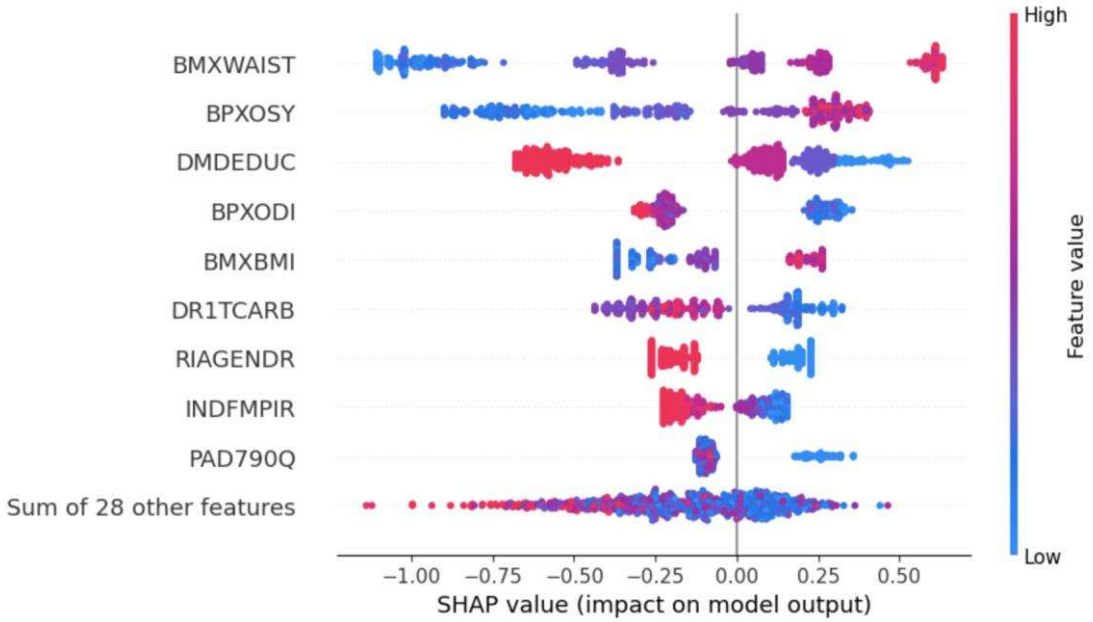


Figure 4: XGBoost variable importance

The figure above is a SHAP (SHapley Additive Plot) plot. This plot shows the variable importance in order, along with what values of each variable are connected with a diagnosis for diabetes. If a value for a variable falls on the right side of the line at 0, that value has an increased likelihood for a Diabetes diagnosis. For example, waist circumference (BMXWAIST). Higher values in waist circumference have a correlation with diagnosis while lower values have a decreased risk of diabetes.

Support Vector Machines

Support Vector Machines were tested using SMOTE-NC to address class imbalance and GridSearchCV to tune the soft margin parameter (C). Several kernel types were evaluated, and the linear kernel consistently provided the best balance between recall, precision, and AUC without overfitting the minority class. With 10-fold cross-validation, the final linear SVM achieved 78.9% accuracy, AUC of 0.86, precision of 0.75, and recall of 0.85, closely mirroring the performance of logistic regression

Figure 5: Decision boundary and Support Vectors

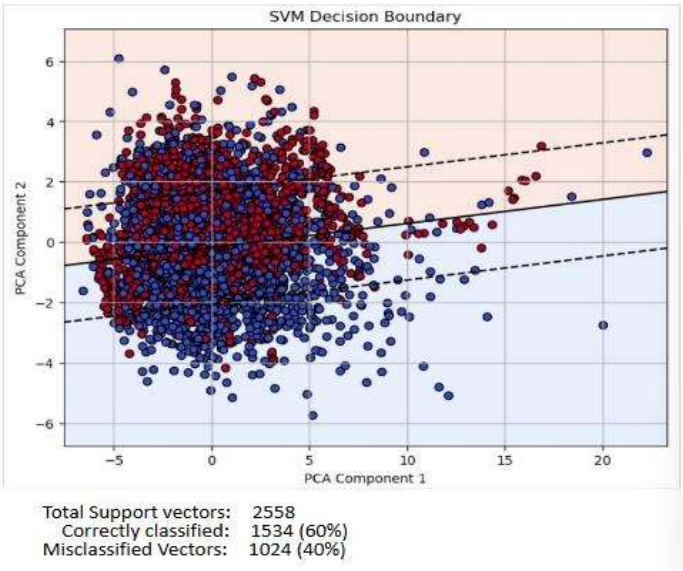


Figure 5 shows the linear decision boundary projected in PCA space. While the model handles most separation well, noise introduced by SMOTE-NC results in overlapping classes and a misclassification rate of roughly 40%. Still, the linear kernel offered a flexible, interpretable solution with strong generalization. Limitations remain in kernel selection and sensitivity to noisy synthetic data, but the SVM proved to be a reliable option when appropriately tuned.

Model Comparison:

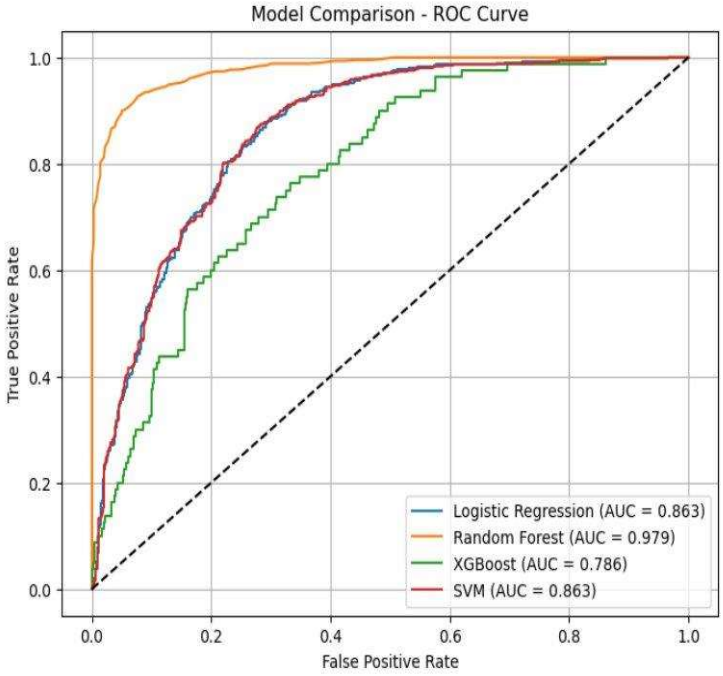
To assess model effectiveness in predicting diabetes, we evaluated their overall performance using accuracy, AUC-ROC, precision, recall, and mean squared error (MSE):

Table 3: Summary of Model Classification Metrics

	Accuracy %	AUC-ROC	Precision	Recall	MSE
SVM	78.94	86.33	79.44	79.11	21.05
XGBoost	20.31	78.57	12.36	98.75	79.69
LR (SMOTE)	78.84	86.27	77.47	81.36	14.02
RF (SMOTE)	91.69	97.86	90.35	93.04	7.89

From the table it's clear that the random forest classifier outperforms other models, achieving the highest accuracy, AUC, and lowest MSE. The support vector Machines performed competitively, especially in terms of its AUC score and balanced precision/recall. The logistic regression model produced comparable results to them SVM model only having a slightly better recall. By contrast, XGBoost, despite its powerful boosting capabilities, performed poorly on all metrics besides recall, which suggest it over classified the minority class. This coupled with its low precision and high MSE suggest that it is struggling with the class imbalance, consequently overfitting the minority class.

Figure 5: SVM, RF, LR & XGB Comparison



Using the test set of the data we can visualize the performance of the models using ROC curves seen in figure 5, reinforcing the performance metrics discussed earlier. Random Forest stands out with the highest AUC, confirming its effectiveness for this classification task, though its complexity and lower interpretability make it more prone to overfitting. Support vector machines and logistic regressions have nearly identical performances, but with proper hyperparameter tuning and kernel choice, SVMs has the potential to improve. SVM is a reliable alternative to Random Forest if interpretability and margin control are priorities. While XGBoost underperforms in its current form, it demonstrates potential, particularly in recall. With further tuning and better handling of class imbalance, such as methods similar to SMOTE-NC, XGBoost could approach the performance of Random Forest.

Conclusions:

The Random Forest model demonstrated the strongest overall performance, achieving the highest accuracy and AUC. However Logistic Regression and Support Vector Machines offer a valuable balance between predictive power, model simplicity and interpretability, making then compelling alternatives for clinical settings. Some key features that consistently contributed to diabetes prediction across all models include blood pressure, waist circumference, gender and education level. This aligns with established clinical indicators in medical studies and show the value of socioeconomic and lifestyle factors. The main challenges faced were class imbalances and missing data, as mot all participants completed the same test and questionnaires leading to incomplete datasets. These limitations impacted model precision and recall, underscoring the need for robust handling of incomplete data. Future improvements should focus on exploring additional techniques to address class imbalance and refine models to find the right balance between accuracy and overfitting. Ultimately, these models have practical real-world potential to aid in early detection and risk assessment of diabetes supporting proactive healthcare interventions.

Comparison to Previous Studies

Classifier performance: AUC-ROC scores [0 – 1]				
	This Study	Qin et al (2022)[17]	Dinh et al (2019)[16]	Badger, Abuwarda (2024)[15]
Logistic Regression	0.8627	0.75	0.724	0.6626
Support Vector Machine	0.8633	0.74	0.887	0.7391
Random Forest	0.9786	0.81	0.937	0.8653
XGBoost	0.7857	0.77	0.957	0.8568

Table 4.AUC-ROC score comparison to previous studies

The results of this study are in line with previous studies, especially, when using unbalanced data. However, this study when compared with the work of Qin and others (2022)[17], did perform better using the balanced data. The study by Qin used data from 1999-2020 while this study was focused on the 2022-2024 NHANES dataset. It is unknown what the effect would be once the data from 1999 to 2024 is combined. Comparing the result of using SMOTENC on this new dataset versus without would be a worthy endeavor, but this would require another round of data processing and cleaning on an even larger dataset. This study is also unique as feature reduction was not done on the predictors, thus other non-traditional diabetes factors can be identified, including caffeine and carbohydrate intake, poverty ratio, and feelings of depression. These factors can be investigated further.

Supplemental Materials:

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

References:

1. Bessesen, D. & Accili, D. (2023). National Institute of Diabetes and Digestive and Kidney Diseases. *What Is Diabetes?* National Institutes of Health. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes
2. American Diabetes Association (2023). *Statistics About Diabetes*. https://diabetes.org/about-diabetes/statistics/about-diabetes
3. National Health and Nutrition Examination Survey. (2024). *About NHANES*. National Center for Health and Statistics. https://www.cdc.gov/nchs/nhanes/about/index.html
4. National Health and Nutrition Examination Survey. (2024). *What to Expect*. National Center for Health and Statistics. https://www.cdc.gov/nchs/nhanes-participants/what-to-expect.html
5. National Health and Nutrition Examination Survey. (2024). *NHANES August 2021 - August 2023*. National Center for Health and Statistics. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023
6. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Data Documentation, Codebook, and Frequencies: Demographic Variables and Sample Weights (DEMO_L)*. National Center for Health and Statistics https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/DEMO_L.htm
7. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Dietary Data - Continuous NHANES*. National Center for Health and Statistics. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary&Cycle=2021-2023
8. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Examination Data - Continuous NHANES*. National Center for Health and Statistics https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&Cycle=2021-2023
9. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Data Documentation, Codebook, and Frequencies Body Measures (BMX_L)*. National Center for Health and Statistics. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2021/DataFiles/BMX_L.htm
10. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Laboratory Data - Continuous NHANES*. National Center for Health and Statistics.https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2021-2023
11. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Questionnaire Data - Continuous NHANES*. National Center for Health and Statistics. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2021-2023
12. National Health and Nutrition Examination Survey. (2024). *August 2021-August 2023 Limited Access Data - Continuous NHANES*. National Center for Health and Statistics. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Non-Public&Cycle=2021-2023
13. Vangeepuram, N., Liu, B., Chiu, P. H., Wang, L., & Pandey, G. (2021). Predicting youth diabetes risk using NHANES data and machine learning. *Scientific* reports, 11(1), 11212. https://doi.org/10.1038/s41598-021-90406-0
14. Vangeepuram, N., Liu, B., Chiu, P. H., Wang, L., & Pandey, G. (2021). Predicting youth diabetes risk using NHANES data and machine learning. *Supplemental Information*. https://pmc.ncbi.nlm.nih.gov/articles/instance/8160335/bin/41598_2021_90406_MOESM1_ESM.pdf
15. Badger, P. & Abuwarda, H. (2024). A Machine Learning Approach to Predicting Future Onset of Type II Diabetes. *Intersect: The Stanford Journal of Science, Technology, and Society*, 17(3). https://ojs.stanford.edu/ojs/index.php/intersect/article/view/3205
16. Dinh, A., Miertschin, S., Young, A. et al. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak* 19, 211 https://doi.org/10.1186/s12911-019-0918-5
17. 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://doi.org/10.3390/ijerph192215027
18. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With Applications in Python*.
19. GeeksforGeeks. (2025, January 27). *Support Vector Machine (SVM) algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/
20. Zabor, E. C., Reddy, C. A., Tendulkar, R. D., & Patil, S. (2022). Logistic Regression in Clinical Studies. International jour nal of radiation oncology, biology, physics, 112(2), 271–277. https://doi.org/10.1016/j.ijrobp.2021.08.007
21. Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease, 11(Suppl 4), S574–S584. https://doi.org/10.21037/jtd.2019.01.25
22. Introduction to boosted trees. Introduction to Boosted Trees - xgboost 2.1.3 documentation. (2022). https://xgboost.readthedocs.io/en/stable/tutorials/model.html
23. Simplilearn. (2022, November 22). What is XGBoost? An Introduction to XGBoost Algorithm in Machine Learning | Simplilearn. Simplilearn.com. https://www.simplilearn.com/what-is-xgboost-algorithm-in-machine-learning-article

24. Khurshid, M. R., Manzoor, S., Sadiq, T., Hussain, L., Khan, M. S., & Dutta, A. K. (2025). Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction. PloS one, 20(1), e0310218. https://doi.org/10.1371/journal.pone.0310218
25. GeeksforGeeks. (2021, September 18). XGBoost. GeeksforGeeks. https://www.geeksforgeeks.org/xgboost/
26. Scikit-Learn Developers. (2025). RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
27. IBM. What is random forest? https://www.ibm.com/think/topics/random-forest
28. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002, June 1). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.953
29. American Diabetes Association (2025). *Diabetes Diagnosis & Tests.* Understanding Diabetes Diagnosis. https://diabetes.org/about-diabetes/diagnosis