DATA 4419 First Report

Carter Silos
Gabriel Aguirre
Marvin Bernardino

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

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.

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 very 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. These 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 a lot of very useful 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.

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
- 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.

Methods:

This project will focus on using three main methods for predicting diabetes. 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). The equation that represents the logistic regression model is equal to: $p = 1/1+e^{-(\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn)}$ [20].

Justification:

Logistic regression is useful due to its interpretability with odds ratios and p-values. In logistic regression, the odds ratios are represented by exponentiating β. If the odds ratio is less than 1 this means an outcome has lower odds of occurring. When the odds ratio is greater than 1, this means an outcome has higher odds of occurring. When the ratio is equal to 1, there is no effect on the outcome. Odds ratio is the magnitude of the association and also tells the direction that it is going in. When interpreting p-values the comparison is between the p-value and the significance level, $\boldsymbol{\alpha}$. If the p-value is lower than the significance level there is an association. If the p-value is higher than the significance level, there is no association.[20],[21]

Method:

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 hyperplane is typically defined by the equation: $w \cdot x + b = 0$

where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ represents input features, and $\boldsymbol{b}$ is the bias term. 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.

Justification:

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. Outliers will also pose less of an issue because SVM relies on support vectors for the margins and hyperplane.

Method:

The final model this project will look at is XGBoost (Extreme Gradient Boosting). XGBoost is an open-source software library based on gradient boosted trees, used for supervised learning problems. XGBoost is defined as $\hat{y}_i = \boldsymbol{\Sigma}\, f_k(x_i)$ where k is the number of trees and f is a function. 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.[22],[25]

Justification:

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. 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.[23],[24]

Tentative Schedule and Task Distribution:

- **Team Members / Chosen Method:**
    - Gabriel Aguirre: Support Vector Machines
    - Marvin Bernardino: Logistic Regression
    - Carter Silos: XGBoost
- **Data Processing (Feb 16 - Mar 2):**

  Task:
    - Explore NHANES dataset and check variable distributions
    - 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
- **Model Implementation and Evaluations(Mar 3 - Mar 16):**

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

  Deliverables:
    - Initial models implemented and evaluated
- **Progress Report and Peer Review (Mar 17 - Mar 23):**

  Task:
    - Conduct peer review for each model (all members)
    - Compile and organize findings into second report (all members)
    - Upload code documentation to GitHub

  Deliverables:
    - Second Report (Mar 23)
    - Updated GitHub repository
- **Model Optimization and Comparative Analysis (Mar 24 - Apr 12):**

  Task**:**
    - Optimization of models to improve accuracy (all members)
    - Perform Comparative Analysis to identify performance and key findings (all members)

  Deliverables
    - Optimized models
    - Comparative analysis report
- **Final Report and Presentation (Apr 13 - 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)

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 NHANE*S. 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). A*ugust 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 journal 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/