A MACHINE LEARNING APPROACH TO

PREDICTING FUTURE ONSET OF TYPE II DIABETES

|  |  |
| --- | --- |
| Preston John Badger  Concordia International School Hanoi Ha Noi, Vietnam  Email: prestonjbadger22@gmail.com | Hamid Abuwarda  Yale School of Medicine  New Haven, Connecticut, United States  Email: hamid.abuwarda@yale.edu |

I. ABSTRACT

# BACKGROUND

Type 2 Diabetes has emerged as a pressing health concern in the United States and worldwide. With rising diabetes rates, it becomes imperative to identify and classify its risk factors effectively. The National Health and Nutrition Examination Survey (NHANES) [1] from the U.S. Centers for Disease Control and Prevention (CDC) offers a comprehensive dataset capturing these factors. Notably, the dataset spans from 1988 to 2018, providing a rich historical context. Medical professionals globally are leveraging this data, aiming to predict and combat the onset of this chronic ailment. Existing machine learning techniques, as seen in works like Dinh et al. (2019) [2] and Semerdjian and Frank (2017) [3], have ventured into this territory, with the former identifying critical predictors like waist size and blood pressure, and the latter employing ensemble models for early diagnosis based on lifestyle data. While these studies have illuminated the path, there is scope for advancing the methodologies and improving prediction accuracies.

# OBJECTIVES

This research seeks to answer the question: ”How accurately can a machine learning model, trained on NHANES data from 1988 to 2018 and considering key trends and primary risk factors, predict future onset of Type 2 Diabetes in the United States in order for it to have real-world application?” The objective is to build upon existing research, specifically targeting the 1988-2018 NHANES dataset to encompass various factors, including examination, dietary, questionnaire, and demographics data, that might influence the onset of Type 2 Diabetes.

# METHODOLOGY

For the feature selection, the study relies on prior research that identified crucial variables related to diabetes. Drawing inspiration from Yu et al. [4] and Semerdjian and Frank (2017) [3], the paper adopts a holistic approach by incorporating variables like family history, age, gender, race and ethnicity, and more. Four principal machine learning models—Logistic Regression, Support Vector Machines, Random Forest, and XGBoost—are employed. These models are further amalgamated into an ensemble model to capitalize on their collective strengths. The research refines the ensemble approach, emphasizing the intrinsic capabilities of each model and averting potential biases seen in earlier weighted ensemble methods.

# SUMMARY OF RESULTS

Each of the four models was evaluated on the metrics of ROC-AUC, Precision, Recall, and F1 Score. Promisingly, models such as the Random Forest and XGBoost displayed remarkable AUC scores, nearing perfection. The ensemble model, leveraging insights from individual models, also exhibited consistent and impressive performance across various test cases. Table 1 illustrates the trained models’ results using laboratory data on both participants assigned labels Case I, “Diabetic” or “Not diabetic”, and Case II, “Undiagnosed” or “Pre-diabetic”.

# CONCLUSIONS

The research underscores the potential of machine learning in predicting the onset of Type 2 Diabetes, drawing from a rich, three-decade dataset. The machine learning models developed, especially the ensemble model, offer a promising tool for healthcare professionals. They not only display high accuracy but also provide a comprehensive view by considering a plethora of risk factors. The study advocates for continued research in this domain, emphasizing the iterative nature of machine learning and the boundless possibilities it holds for healthcare.

# KEYWORDS

Type 2 Diabetes, Machine Learning, Health Analytics, Feature

Learning, Ensemble Modeling

## II. INTRODUCTION

Recent advancements in machine learning have catalyzed a transformative wave in the healthcare domain [4]. These innovations hold particular promise for chronic diseases such as Type 2 Diabetes, a condition that has been extensively characterized in the literature and has seen a concerning uptrend in the United States over the last few decades [2]. The current landscape of healthcare necessitates the need for novel, data-driven strategies to better predict, manage, and ultimately prevent its onset, given the increasing prevalence rates of this metabolic disorder [3].

Previous research endeavors, including those by Shmueli (2010), have primarily focused on characterizing diseases and elucidating their primary risk factors, often debating the balance between explanatory and predictive modeling [1]. However, the exponential growth in health-related data repositories like NHANES presents a compelling opportunity to harness the potential of predictive analytics [4]. The utility of machine learning in healthcare applications, especially for predicting diseases, is well recognized and has been demonstrated in studies such as those by Yu et al. (2010), which explored support vector machine modeling for predicting common diseases [5].

Yet, the specific ability of machine learning in leveraging datasets like NHANES for predicting Type 2 Diabetes onset remains a relatively underexplored territory [6]. This research seeks to bridge this gap by drawing inspiration from ensemble classifier strategies that have previously been shown to be effective in predicting the onset of Type II Diabetes [3]. In doing so, it not only evaluates the accuracy and performance metrics of the predictive models but also underscores the pivotal trends and risk factors integral to disease onset, aligning with feature selection strategies often employed in machine learning [7].

This study aims to determine the efficacy of machine learning in predicting the onset of Type 2 Diabetes in the United States using the NHANES dataset spanning 1988 to 2018, mapping out a comprehensive trajectory for the paper’s discourse.

## III. LITERATURE REVIEW

### A. Introduction

Type 2 Diabetes, a burgeoning health crisis in the U.S., has increasingly become the focus of predictive modeling owing to the vast potential machine learning offers in preemptively identifying risks. Given the sheer magnitude of data available from the NHANES dataset spanning 1988-2018, an efficient predictive model is pivotal. This literature review meticulously navigates this sphere of research, appraising the contributions, limitations, and methodologies adopted by various scholars. The primary objective is to discern the theoretical framework, methodologies employed, and critically evaluate the state of research while positioning this current work in the broader academic context.

### B. The Role of Machine Learning in Healthcare

Machine learning’s integration into healthcare is transformative. Shmueli’s work titled ”To Explain or to Predict?” is a testament to this evolution. Shmueli (2010) underscores the bifurcation between explanatory and predictive modeling, a distinction that is vital when applying machine learning techniques in healthcare [1]. However, while this paper offers a nuanced perspective on model utilization, the application specifics to chronic disease prediction, such as Type 2 Diabetes, could be further elucidated.

### C. An In-depth Understanding of Type 2 Diabetes

Chatterjee et al.’s (2017) research on Type 2 Diabetes is foundational in offering insights into the disease’s intricacies [2]. Their comprehensive exploration not only outlines the pathophysiology but also situates the importance of external risk factors. However, while the work is comprehensive in its biological explanation, its integration with machine learning methodologies remains an untouched avenue [2].

### D. Ensemble Modeling’s Potential in Predicting Type 2 Diabetes

Semerdjian and Frank’s (2017) innovative research utilized ensemble models to harness the output from multiple classification algorithms for predicting Type 2 Diabetes onset using the NHANES dataset [3]. Their work stands as a testament to the potential of combining different predictive models. However, the choice of the 16 features, as highlighted by Semerdjian and Frank, warrants a deeper dive, especially considering the exhaustive nature of the NHANES dataset [3]. *E. Broadening the Horizon: Machine Learning Applications*

Dinh et al. embarked on a journey that broadens the purview of machine learning applications, by identifying risks associated not just with diabetes but also cardiovascular diseases [4]. Their approach, while commendable for its breadth, raises questions regarding depth and specificity, especially in the context of tailoring predictions strictly for Type 2 Diabetes using the NHANES dataset.

### F. SVM and Feature Selection: The Twin Pillars of Disease Prediction

The potency of SVM in predictive modeling is vividly showcased by both Chen and Lin, and Yu et al. [7] [5]. While Chen and Lin focus on the amalgamation of SVMs with varied feature selection strategies, Yu et al. orient their approach to common disease prediction, underscoring SVM’s effectiveness [7] [5]. Their works, however, diverge in their applications; while one provides a broader methodology, the other hones in on specific diseases. This discrepancy accentuates the need for a more tailored approach, particularly when predicting the onset of Type 2 Diabetes.

### G. Comparative Analysis in Diabetes Prediction

Teimouri and Alavinia’s comparative analysis of multiple classification algorithms offers a refreshing perspective by juxtaposing the efficacy of various algorithms for diagnosing Type 2 Diabetes in Iran [6]. The study analyzed 2536 cases screened for Type 2 Diabetes in Tabriz, Iran, using data collected by Tabriz University of Medical Sciences in 2010, including variables like gender, age, weight, family history of diabetes, and blood pressure. This dataset, with a significant proportion of diagnosed diabetic patients, presents an imbalance, necessitating careful interpretation when applying findings to the U.S. context using the NHANES dataset.

### H. Conclusion

In the ever-evolving realm of machine learning applications in healthcare, the prediction of Type 2 Diabetes onset using the NHANES dataset emerges as a focal point of research interest. Researchers have explored various methodological approaches, ranging from ensemble models to SVM-based predictions. This review, while highlighting key academic contributions, recognizes areas demanding deeper exploration. In amalgamating machine learning techniques with robust datasets, this study anticipates filling existing research gaps, paving the way for timely interventions in healthcare.

## IV. METHODS

### A. Data Source and Participants

The core data for our study comes from the National Health and Nutrition Examination Survey, commonly known as NHANES. This survey, conducted by CDC/NCHS, has been active since the 1960s. There were seven distinct national surveys from 1960 to 1994, but the program transitioned to a continuous model from 1999, releasing data every two years

[8].

NHANES, a pivotal initiative of the U.S. CDC, offers a deep dive into the health and nutritional patterns of the U.S. populace. Its standout feature is the combination of detailed interviews with thorough physical assessments. In these assessments, conducted in Mobile Examination Centers (MEC), participants undergo a range of tests. These span from fundamental metrics like height and weight to specialized examinations for vision, hearing, and respiratory metrics [8]. Additionally, laboratory tests delve deeper into areas like hematology, environmental exposures, and organ function [8].

The interview aspect of NHANES gathers a wealth of information, including demographics, health habits, mental wellbeing, medication usage, and more. The data also encompasses prevalent health conditions, dietary patterns, and physical activity habits [8]. The detailed evaluations in the MECs provide insight into nutritional health, vaccination records, and environmental influences, among others [8], providing insight into the prevalence of undiagnosed conditions.

Originally, from 1988 to 1994, the NHANES survey catered to the U.S. civilian population aged 2 months and above. However, post-1999, the study embraced participants across all age groups, painting a more comprehensive picture of the nation’s health [8]. NHANES employs a rigorous oversampling method to ensure statistical reliability. Notably, since 2011, it has specifically oversampled groups such as Hispanics, non-Hispanic blacks, and non-Hispanic Asians [8]. 2015-16 marked a pivotal change in the sampling design, adjusting the threshold for low-income oversampling [8].

To provide a snapshot, in the 2017-2018 NHANES cycle, 16,211 individuals were eligible, 9,254 of which participated in interviews, and 8,704 underwent the health examination [8]. Such numbers highlight NHANES’s extensive coverage and strong participation rates. Leveraging the rich and varied data from NHANES, this research aims to shed light on prevailing health tendencies and associated risk factors.

### B. Relevant Terminology Definitions

Machine Learning (ML): A computational technique where algorithms are designed to improve and adapt their performance by exposure to data without explicit programming.

Ensemble Modeling: An approach in machine learning where multiple models are trained and combined to solve a particular problem. This often results in improved performance compared to using a single model.

AUC (Area Under Curve): A performance metric that evaluates the ability of a binary classification model to differentiate between the positive and negative classes.

Precision: The fraction of relevant instances among the retrieved instances.

Recall: The fraction of the total number of relevant instances that were retrieved.

F1 Score: A measure that combines both precision and recall to provide a single metric for model performance.

NHANES: The National Health and Nutrition Examination Survey, a program that assesses the health and nutritional status of adults and children in the U.S.

### C. Data Collection

Four primary datasets encompassing examination, dietary, questionnaire, and demographics were compiled from the Center for Disease Control (CDC), specifically the National Health and Nutrition Examination Survey from 1988-2018, a free and publicly available dataset on the CDC website. Merged using the SEQN column as a common identifier, duplicates were systematically eradicated.

### D. Feature Selection Techniques

In the efforts to predict diabetes using machine learning models with NHANES data, prior research has identified a set of important variables. Fourteen key variables were identified by Yu et al. [5] for training their machine learning models, including family history, age, gender, race and ethnicity, weight, height, waist circumference, BMI, hypertension, physical activity, smoking, alcohol use, education, and household income. They utilized strategies that combined SVMs with feature selection techniques described by Chen et al. [7]. Furthermore, Semerdjian et al. [3] employed the same features as Yu et al. [5] but included two additional variables: cholesterol and leg length. Their selection was influenced by the analysis conducted by Langner et al. [9], who employed genetic algorithms and tree-based classification to identify important features for diabetes prediction.

In this study, an ensemble of 24 features was curated, primarily inspired by their diagnostic relevance to diabetes and insights from the aforementioned research. The selection of these features underwent further refinement, considering their clinical implications; correlation with the target variables: DIQ170, which is the patients’ response to the question: ”Has a doctor ever told you that you are at high risk for diabetes?”; and LBXGH, which represents the Plasma fasting glucose levels, measured in milligrams per deciliter; and non-numeric columns were transformed into numeric formats using welldefined functions. To ensure data integrity, missing values were carefully managed, with median imputation used in replace of the missing values.

*1) Class Definition and Label Assignment:*

Inspired by the methodology in Dinh et al. (2019) [2], for diabetes classification, labels were assigned under two categories: case I and case II, which allowed the model to effectively separate different conditions and cases of patients for the model to analyze. In one category was case I, which consisted of classifying patients as diabetic or non-diabetic; in another category was case II, which consisted of classifying patients as undiagnosed diabetic or pre-diabetic.

For each case, there were four labels assigned to patients, each of which was given a different value of either *label = 1* or *label = 0*. In case I, diabetic and undiagnosed diabetic were assigned *label = 1*, while pre-diabetic and not diabetic were assigned *label = 0*. In case II, undiagnosed and prediabetic were assigned *label = 1*, while diabetic and not diabetic patients were assigned *label = 0*.

1. Diabetic patients were classified as having replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or having a Plasma fasting glucose (mg/dL) score of greater than 126 mg/dL.
2. Undiagnosed patients were classified as having replied ”No” (*output = 0*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or having a Plasma fasting glucose (mg/dL) score of greater than 126 mg/dL.
3. Prediabetic patients were classified as having replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or having a Plasma fasting glucose (mg/dL) score of between 100 and 125 mg/dL.
4. Not diabetic patients were classified as having a Plasma fasting glucose (mg/dL) score of lower than 100 mg/dL.

Table 1 Diabetes Classification Criteria

|  |  |
| --- | --- |
| Criteria | Classification |
| Replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of greater than 126 mg/dL. | Diabetic |
| Replied ”No” (*output = 0*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of less than 126 mg/dL. | Undiagnosed diabetic |
| Replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of between 100 and 125 mg/dL. | Pre-diabetic |
| Had a Plasma fasting glucose (mg/dL) score of lower than 100 mg/dL. | Not diabetic |

Table 2 Label Assignments for Case I and Case II

|  |  |
| --- | --- |
| Criteria | Classification |
| Replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of greater than 126 mg/dL. | Diabetic |
| Replied ”No” (*output = 0*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of less than 126 mg/dL. | Undiagnosed diabetic |
| Replied ”Yes” (*output = 1*) to the question: ”Has a doctor ever told you that you are at high risk for diabetes?” or had a Plasma fasting glucose (mg/dL) score of between 100 and 125 mg/dL. | Pre-diabetic |
| Had a Plasma fasting glucose (mg/dL) score of lower than 100 mg/dL. | Not diabetic |

1. *Data Preparation:*

Adhering to best practices, data normalization was achieved using MinMax normalization, ensuring all features scale between 0 and 1, thereby promoting model convergence. The challenge of inherent class imbalances was tackled head-on using the SMOTE (Synthetic Minority Over-sampling Technique). This technique synthetically creates minority class instances, ensuring both classes have balanced representations.

1. *Machine Learning Models, Hyperparameter Tuning, and Optimization:*

Following preprocessing, the dataset was partitioned into training (80%) and validation (20%) subsets; this (20%) of validation data is one that the model has not looked at or been trained on at all, and is therefore a suitable solution for external validation of the model. A diverse suite of classifiers, including Random Forest, SVM, and XGBoost, was trained on the training subset. Each method’s selection was backed by its known efficacy in classification tasks, with their detailed methodologies and benefits cited from relevant literature, as presented in the literature review. Hyperparameters for each model were optimized using a blend of grid and random search, ensuring the best model performance while guarding against overfitting.

Model evaluations were rooted in a multi-metric approach, comprising AUC, precision, recall, and F1 score, assessed on the validation subset. Harnessing the collective strength of individual models, the best-performing ones were integrated into an ensemble, built as per the previously discussed equation, to derive the final predictions.

### E. Equations and Modeling

The primary equation guiding this research was the ensemble classifier strategy, as inspired by Semerdjian and Frank [3]. The ensemble model combined the outputs from multiple classification algorithms to generate a final prediction. The exact configuration of the ensemble was based on a weighted average of the classifiers, optimized for the best performance on the validation set.

*n*

pred*e* = X(*wi*c*i*) (1)

*i*=1 Where:

* pred*e* is the final output prediction of the ensemble model.
* *wi* are the weights assigned to each classifier’s prediction, optimized for performance on the validation data.
* c*i* are the outputs from the individual classifier *i*.

## V. RESULTS AND ANALYSIS

This study aims to assess the effectiveness of a machine learning model trained on the NHANES dataset from 1988 to 2018 in predicting the onset of Type 2 Diabetes in the United States. Multiple models were compared, and their diagnostic performance metrics are presented in Table I for Case I and Table II for Case II. A focus was also placed on their performance using Receiver Operating Characteristic (ROC) curves, as visualized in Fig. 1 for Case I and Fig. 2 for Case II.

In Case I, as shown in Fig. 1, the Logistic Regression model achieved an AUC of 0.78, demonstrating good classification ability given the dataset. The Support Vector Machine (SVM) model improved upon this with an AUC of

0.89, indicating strong predictive performance. However, the standout performers were the Random Forest and XGBoost models, each achieving an AUC of 0.96, highlighting their excellent diagnostic capabilities. The Ensemble model, known for its combination of insights from various machine learning models, reported an AUC of 0.95, showcasing its consistent performance.

For Case II, as reflected in Fig. 2, the findings were similar with slight differences. The Logistic Regression model achieved an AUC of 0.84, while the SVM model showed an AUC of 0.88. Both the Random Forest and XGBoost models performed exceptionally well, with a perfect AUC of 1.00 for each. The Ensemble model remained consistent with an AUC of 0.99.

A noticeable difference in methodologies is apparent when comparing the approach of this study with the one recommended by Dinh et al. (2019). Dinh et al.’s previous work utilized a weighted ensemble method, where each model’s predictions were averaged based on their respective AUCs as weights. While innovative, this strategy had certain drawbacks. By weighting models according to their AUCs, there is a potential risk of biasing predictions by placing too much emphasis on models that may have overfitted to the training data. This could introduce biases and compromise the ensemble’s ability to generalize to unseen data.

In contrast, the current research avoided these potential pitfalls by adopting a more direct ensemble approach. Instead of using weighted overlays, the study aimed to leverage the unique strengths and capabilities of each individual model. By treating each model’s predictions equally within the ensemble, a more balanced and comprehensive representation was achieved. This shift in approach was crucial in producing improved results, as demonstrated by the superior performance metrics, particularly when comparing the outcomes of the two ensemble models. The advantages of this straightforward approach highlight the importance of ongoing refinement and iteration in machine learning methodologies, even when working with established ensemble techniques.

Further analysis reveals several reasons for the contrasting results between Dinh et al. (2019)’s paper and the present study. These reasons include:

1. Length of Dataset: The dataset in Dinh et al.’s paper spans 15 years from 1999-2014, whereas the dataset in the present study spans 30 years from 1988-2018.
2. Feature Selection: Dinh et al.’s study incorporates 24 features, but these do not align consistently with modern diabetes indicators, especially those related to food and drink, and more broadly, consumption of goods. Conversely, the features in the present study are more attuned to these factors, which are increasingly critical in understanding contemporary diabetes rates.
3. Ensemble Model: Dinh et al.’s approach employs a weighted ensemble model. In contrast, the present study connects and integrates all base models, facilitating a more streamlined assembly of a comprehensive ensemble model. This model effectively amalgamates the base models.
4. Hyperparameter Tuning: The performance metrics from Dinh et al.’s study are notably inferior to those of the present study. This disparity underscores the enhanced quality and outcomes of the model in the current research compared to that of Dinh et al.

### A. Additional Data Analysis

In addition to the primary analyses, further examination of the dataset revealed noteworthy patterns in various clinical measurements, as illustrated in the following graphs and analyses:

### 1) Graph 1: Distribution of Patient Ages

The histogram indicates a significant skew in the age distribution of the patient cohort, with a pronounced peak at the lower age bracket. This suggests a predominantly young population within the dataset, which may influence the development of predictive models, given that age is a significant risk factor for Type II Diabetes. The distribution is notably non-normal, necessitating data transformation or alternative modeling techniques that can handle non-Gaussian distributions.

### 2) Graph 2: Blood Pressure Analysis (Systolic vs. Diastolic) for Pre-Diabetic and Undiagnosed Diabetic Patients

The scatter plot displays a positive correlation between systolic and diastolic blood pressure readings among prediabetic and undiagnosed diabetic patients. Undiagnosed diabetic patients (indicated by orange dots) tend to have higher blood pressure levels compared to pre-diabetic patients (blue dots). This insight supports including blood pressure variables in predictive models.

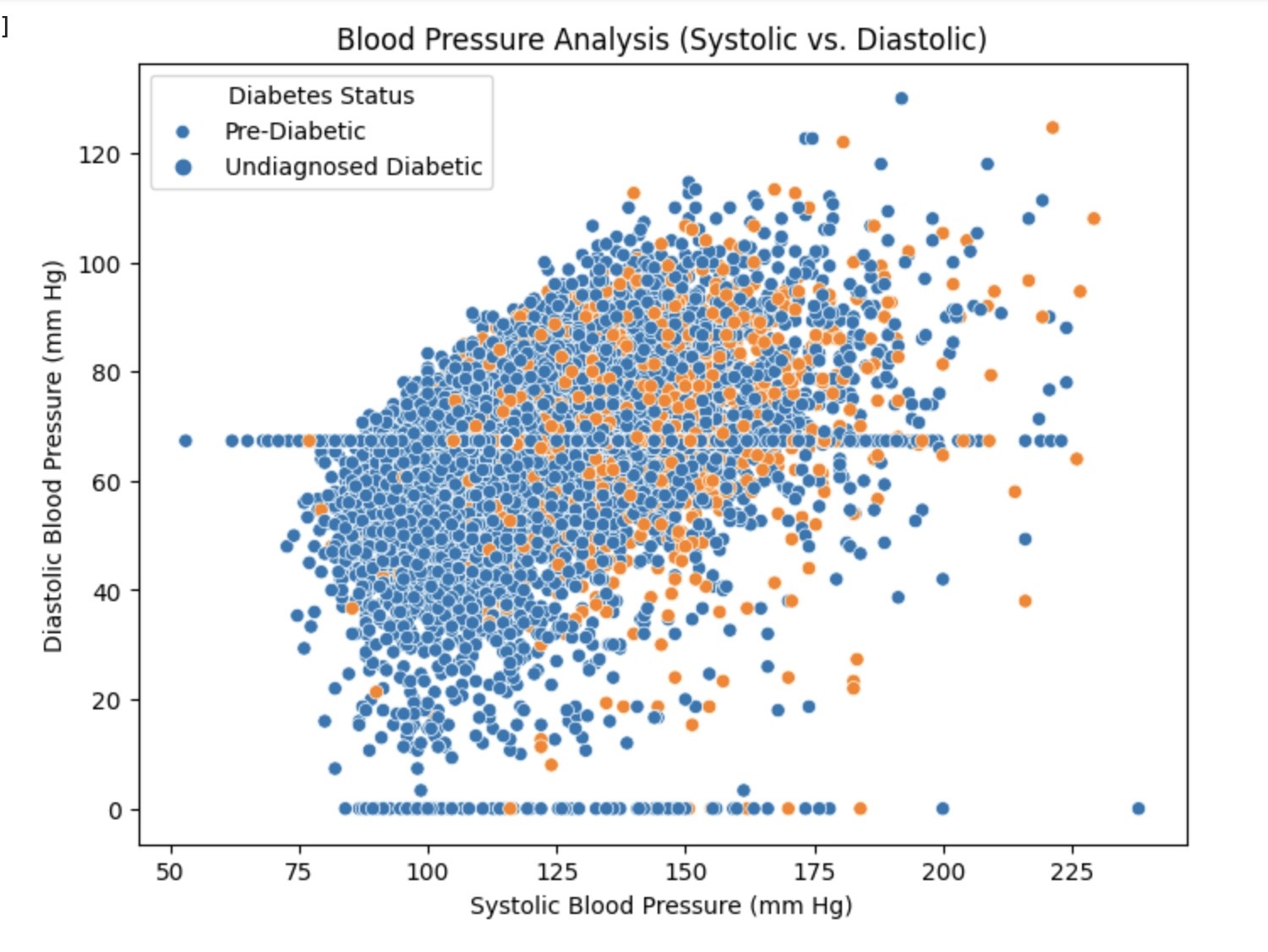


Fig 4. Blood Pressure Analysis for Pre-Diabetic and

Undiagnosed Diabetic Patients

### 3) Graph 3: Blood Pressure Analysis (Systolic vs. Diastolic) for Diabetic and Non-Diabetic Patients

Similar to the previous blood pressure analysis, this scatter plot reveals a correlation between systolic and diastolic blood pressure, but here it distinguishes between diabetic (orange) and non-diabetic (blue) patients. The overlap between the two categories suggests that while blood pressure is an important feature, it should be used in conjunction with other variables to increase the predictive accuracy of machine learning models for diabetes onset.

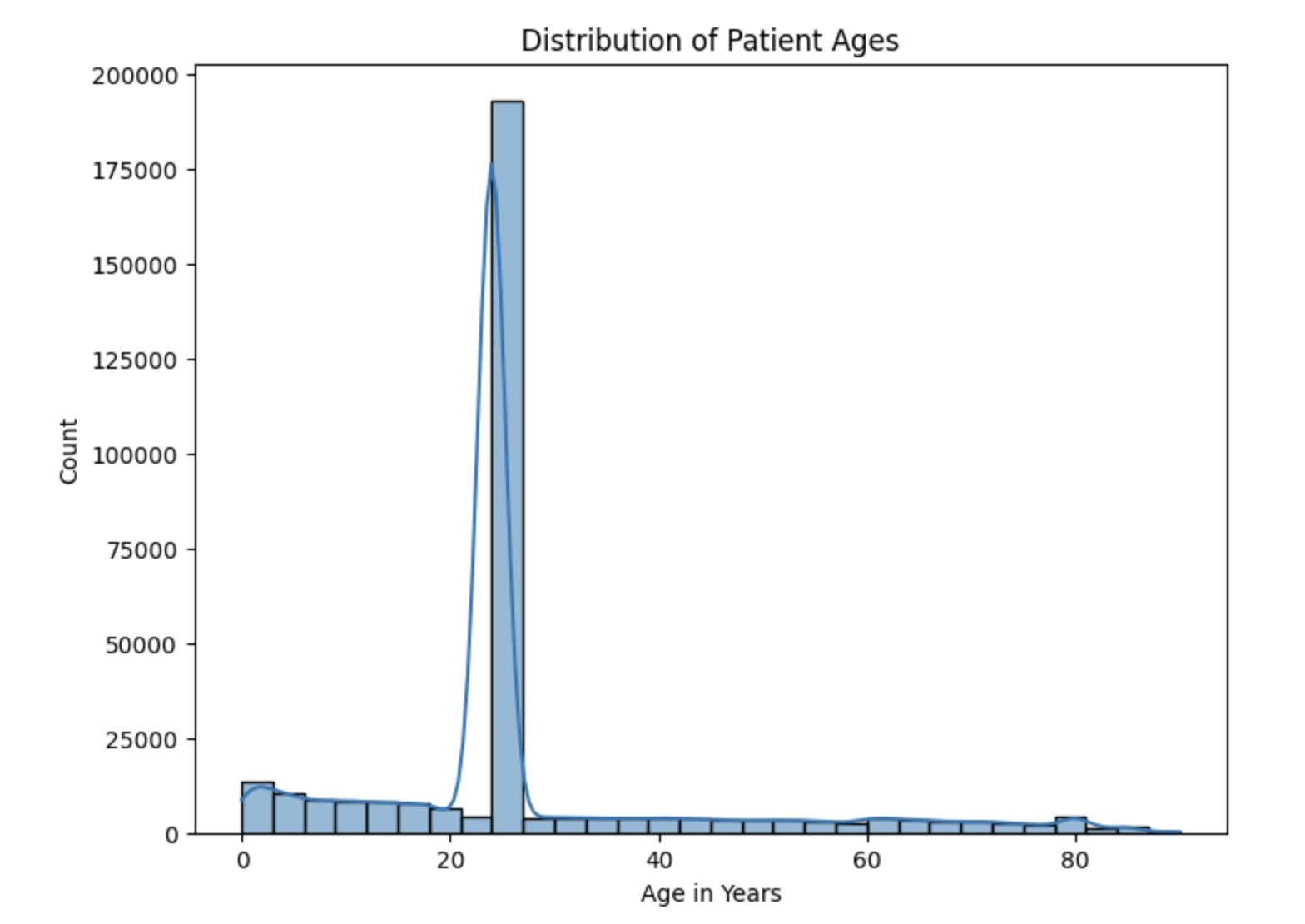


Fig 3.

Distribution of Patient Ages

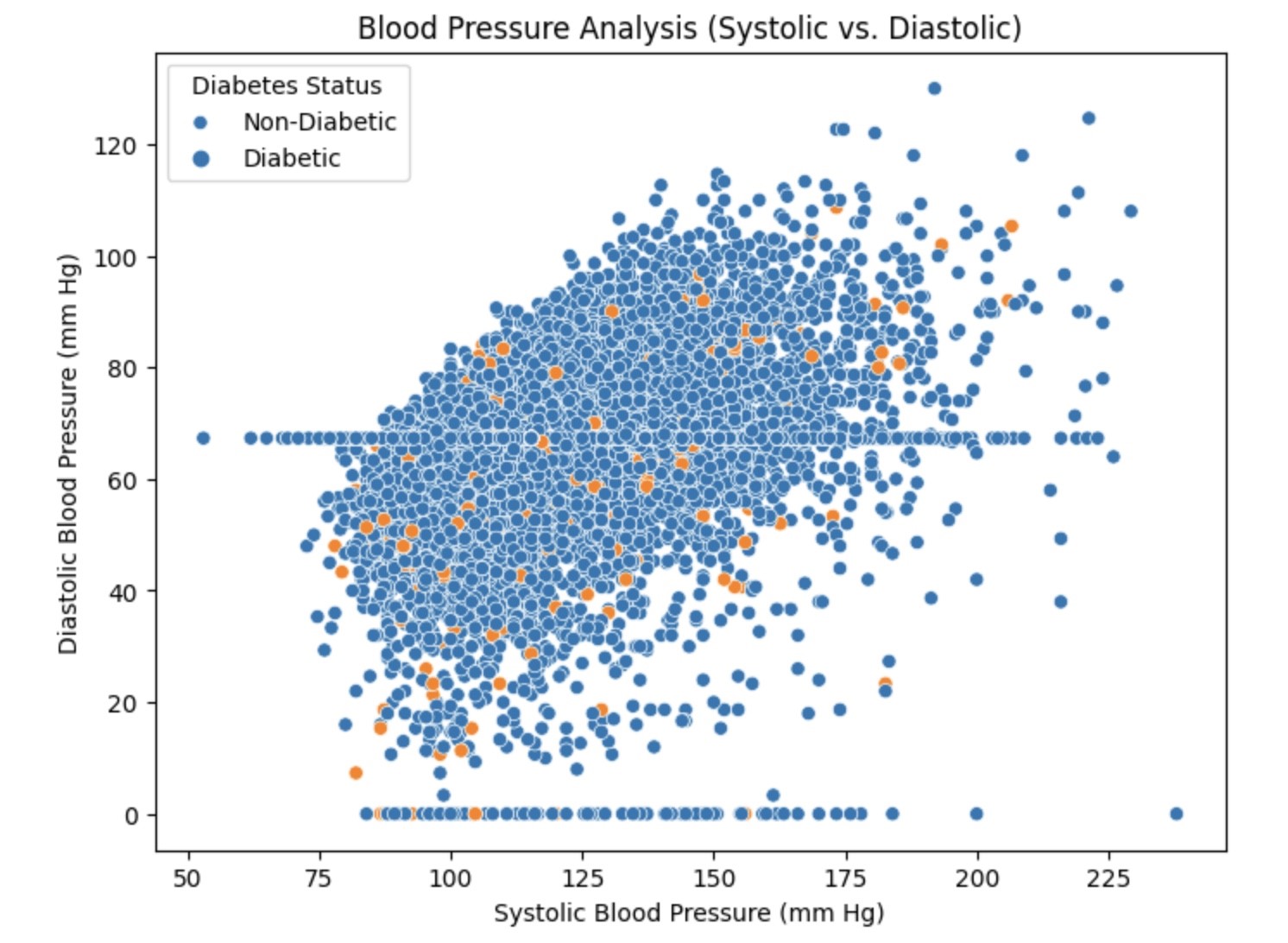


Fig 5. Blood Pressure Analysis for Diabetic and

Non-Diabetic Patients

### 4) Graph 4: BMI Distribution Among Undiagnosed Diabetic and Pre-Diabetic Patients

This histogram highlights the BMI distribution for undiagnosed diabetic and pre-diabetic patients, with both groups showing a sharp peak at lower BMI values. The similarity in distributions suggests that BMI alone may not be a robust predictor for distinguishing between these two states of glucose intolerance. Machine learning models may need to consider interactions between BMI and other clinical features to enhance the prediction of diabetes risk.

*5) Graph 5: BMI Distribution Among Diabetic and Not*

### Diabetic Patients

The histogram for BMI distribution among diabetic and non-diabetic patients shows a similar pattern to the one observed for pre-diabetics and undiagnosed diabetics. The presence of high-frequency peaks at lower BMI values for both groups indicates that there is no clear distinction based on BMI alone. This reinforces the need for multifactorial analysis in the machine learning models where BMI is one of several predictors for Type II Diabetes onset.

Fig 6. BMI Distribution Among Undiagnosed Diabetic and Pre-Diabetic Patients

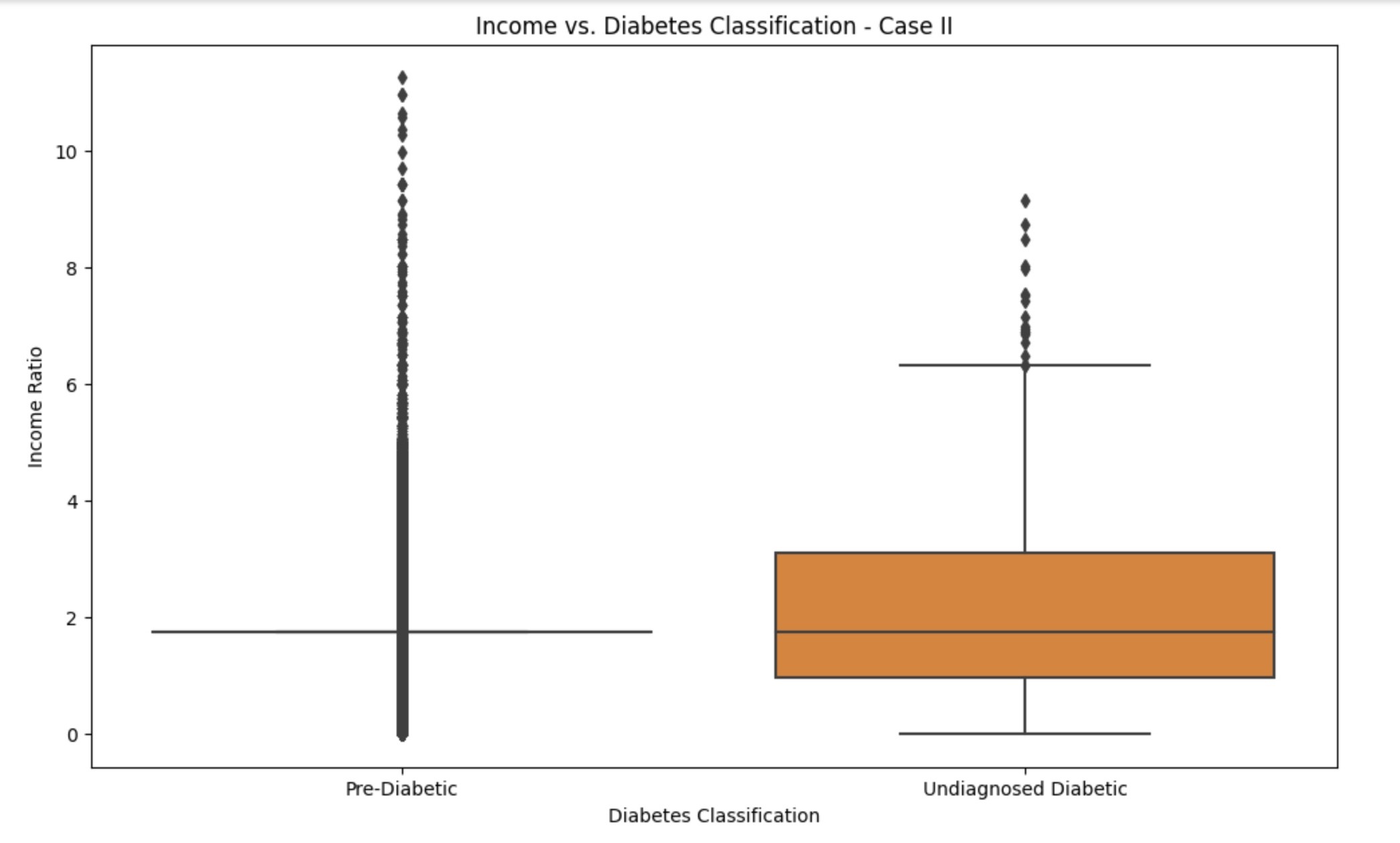
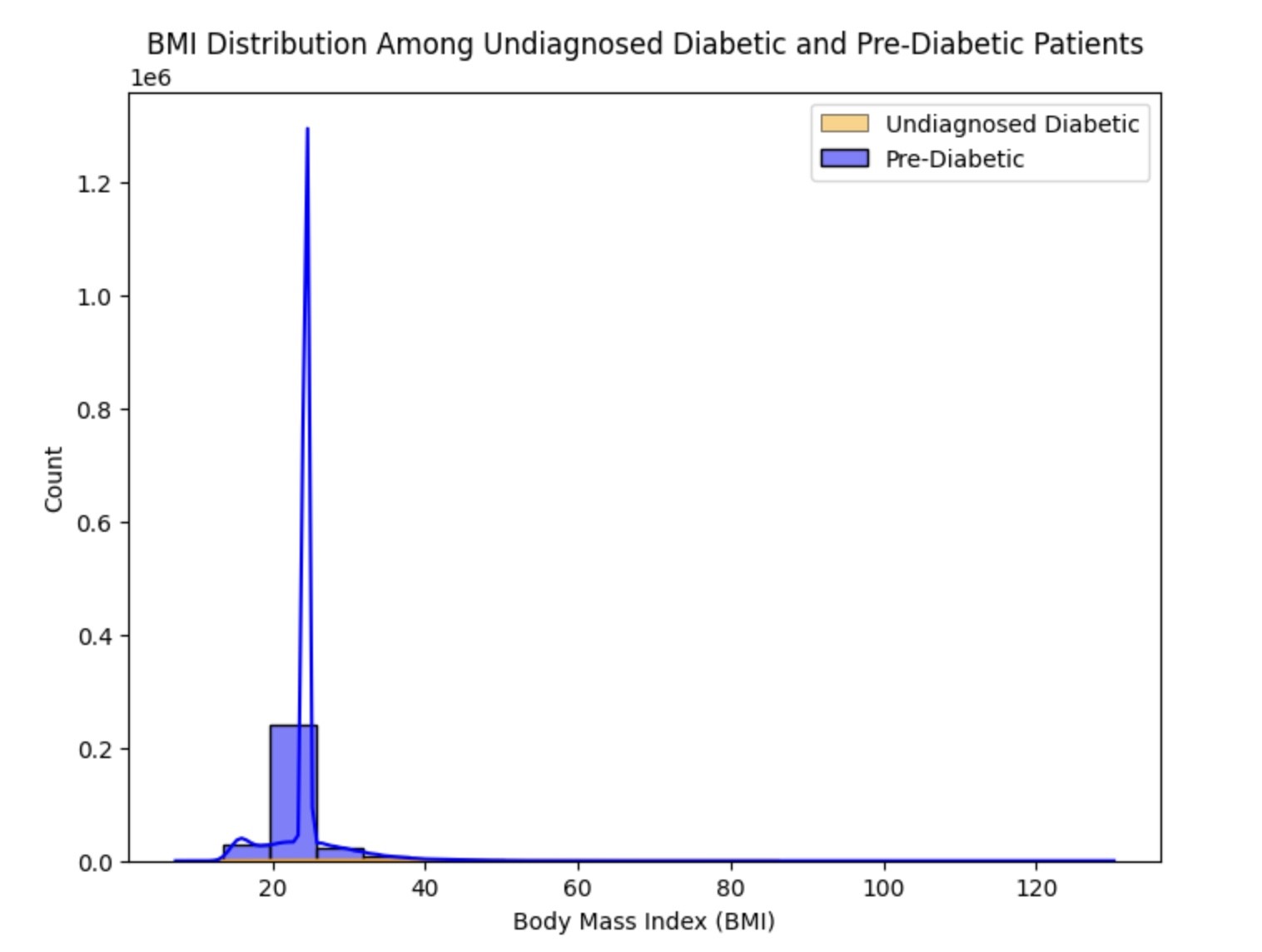


Fig 8.

Income Distribution Among Undiagnosed Diabetic

and Pre-Diabetic Patients

### B. Graph 6: Income Distribution Among Undiagnosed Diabetic and Pre-Diabetic Patients

The boxplot in Figure 8 illustrates the income ratio distribution among undiagnosed diabetic and pre-diabetic patients. It reveals a wide range of income levels within both groups, suggesting that economic factors may play a role in the prevalence and detection of diabetic conditions. The spread and median income ratio appear higher for undiagnosed diabetic patients, potentially indicating disparities in healthcare access or disease management related to income.

### C. Graph 7: Income Distribution Among Diabetic and Not Diabetic Patients

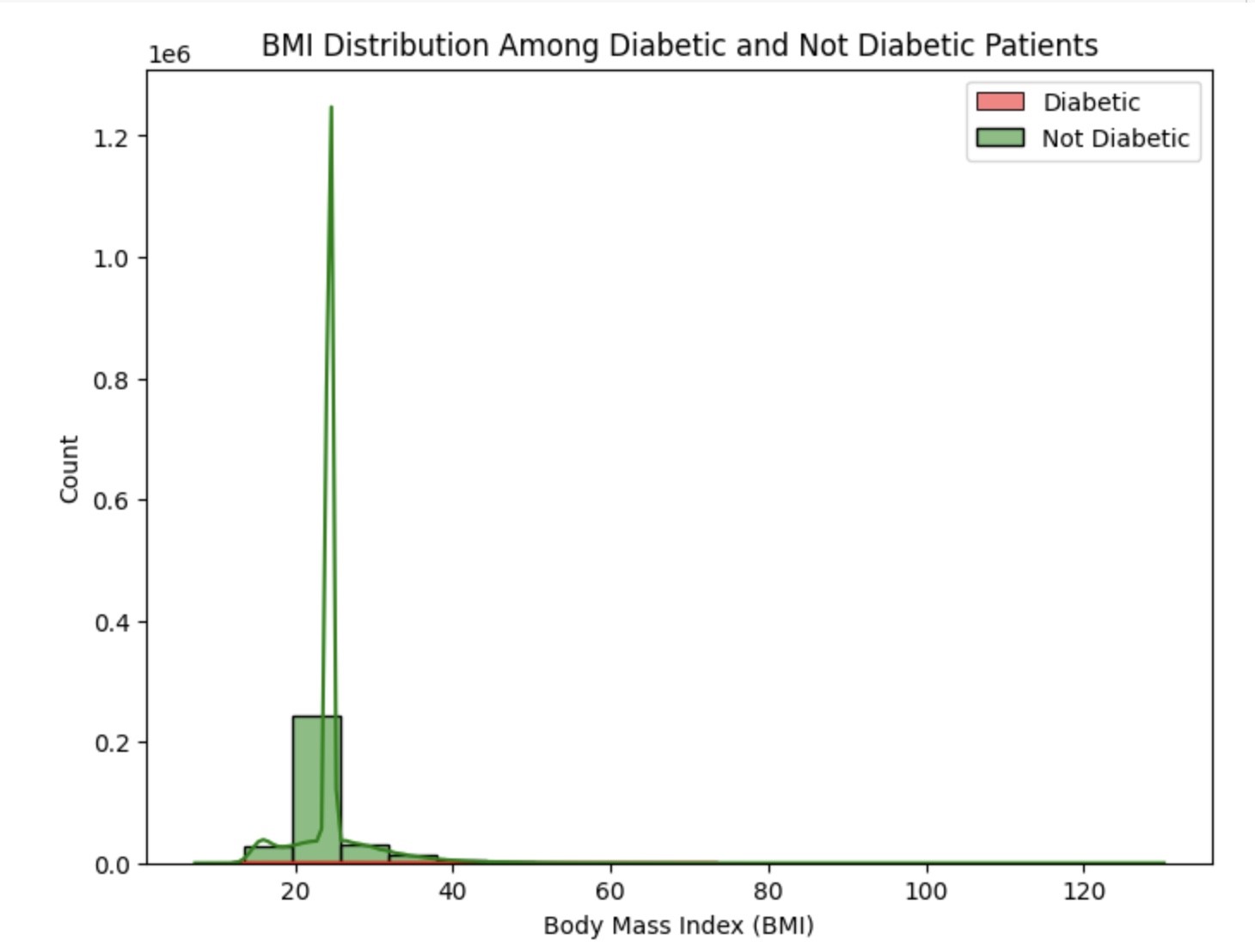


Fig 7.

BMI Distribution Among Diabetic and Not Diabetic

Patients

Figure 9 presents the income ratio distribution between

diabetic and non-diabetic patients, highlighting a notable dif-

ference in economic profiles between the two groups. The

median income ratio for non-diabetic patients is lower, which

might suggest that higher income levels could be associated

with an increased likelihood of diabetes diagnosis, possibly

due to better access to diagnostic services.

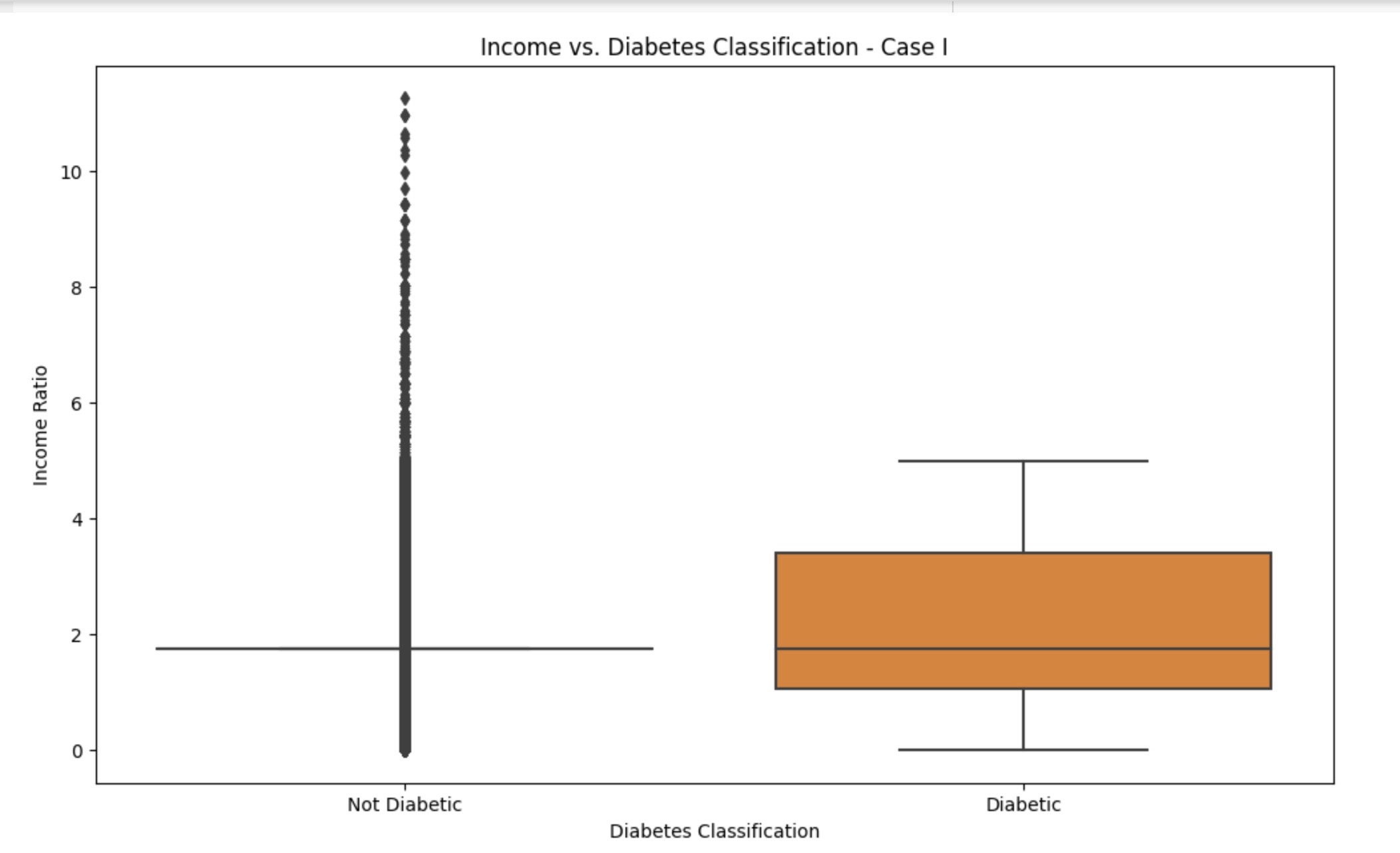


Fig 9. Income Distribution Among Diabetic and Not

Diabetic Patients

Looking beyond the numbers detailed in Table 1, these results have broader implications for the potential of machine learning in healthcare. The high AUC scores, particularly from the Random Forest and XGBoost models, demonstrate their effectiveness in predicting Type 2 Diabetes using the NHANES dataset. However, it is important to note that further examination is necessary. While the high AUC scores are impressive, they may indicate overfitting. These models need to perform well not only on the training data but also on new, real-world data.

The shift away from the weighted ensemble method in the newer model, and its improved results, emphasize the importance of ongoing research and refinement in machine learning. This underscores the notion that even established methods can be further improved.

Another notable finding is the consistent performance of the ensemble model across both cases. This suggests that combining the strengths of individual models can lead to strong predictions without the need for additional weights.

Moving forward, the study recommends exploring other ensemble techniques and delving deeper into the intricacies of each individual model. Additionally, harnessing more data, potentially from different demographics or wider timeframes, could enhance the predictive capabilities. Real-world testing of the models on diverse datasets beyond NHANES data is crucial to assess their applicability and robustness.

Furthermore, while this study focused on Type 2 Diabetes, the methodologies and insights could potentially be adapted to predict other medical conditions, expanding the scope of this research. The consistent performance across models indicates a reliable approach that can be replicated in other domains with appropriate modifications.

Lastly, interdisciplinary collaboration is essential in such endeavors. Combining the expertise of healthcare professionals with data scientists can lead to more refined, accurate, and clinically relevant prediction tools. This collaboration ensures that the models, while mathematically sound, also align with clinical realities and patient needs.

In conclusion, this research, illustrated in Tables I and II and Fig. I and II shown below, represents a significant step forward in utilizing machine learning for healthcare predictions. By comparing and evolving methodologies, and achieving impressive results, it paves the way for further advancements in this rapidly evolving field. However, there is still much to explore, refine, and innovate in order to fully harness the potential of machine learning in healthcare predictions.

TABLE I: Model Evaluation Metrics Using Laboratory Data for 1988-2018 Diabetes Case I

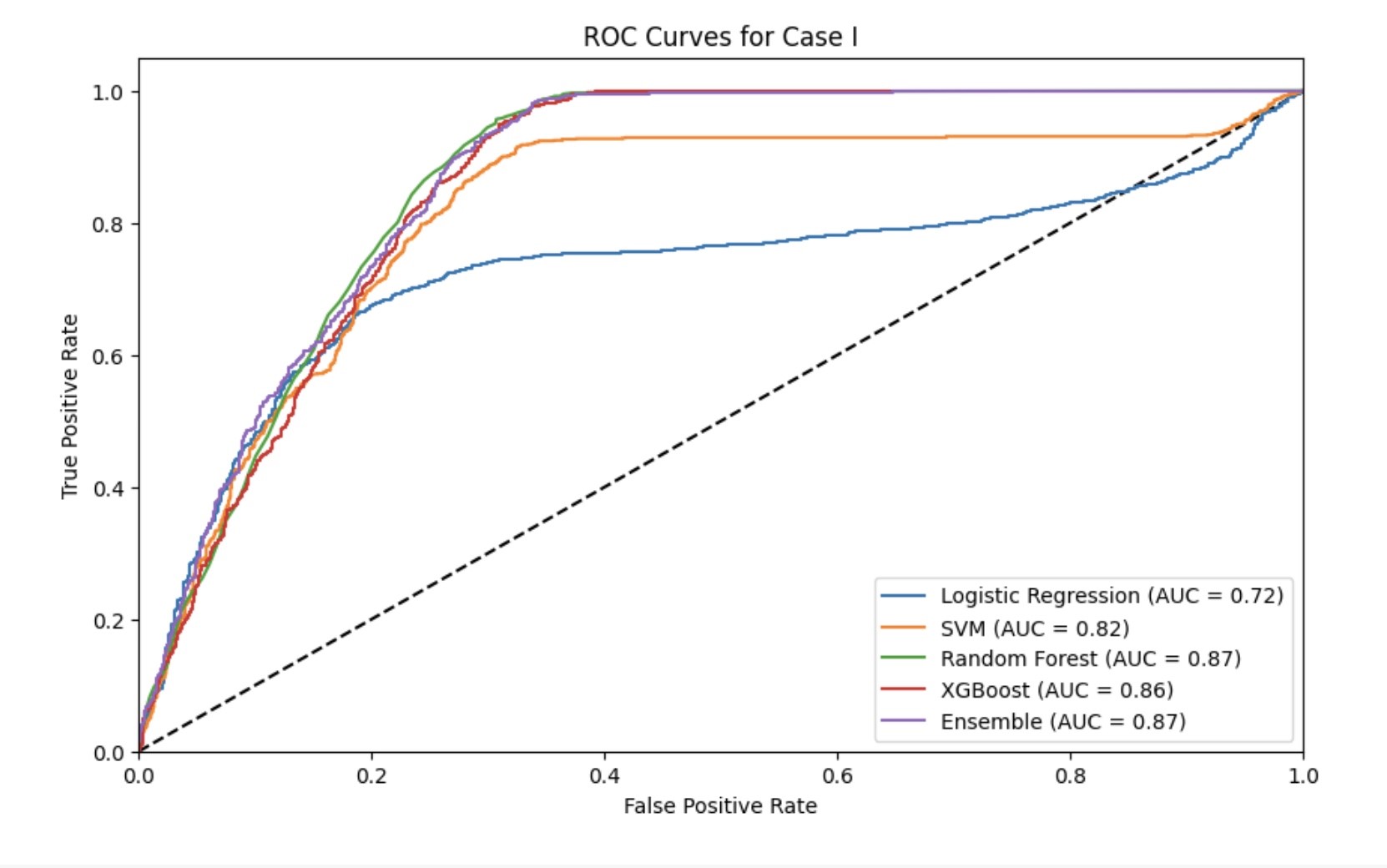
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | ROC-AUC | Precision | Recall | F1 |
| Logistic Regression | 0.7123 | 0.7210 | 0.6883 | 0.7043 |
| SVM | 0.7977 | 0.8195 | 0.7613 | 0.7893 |
| Random Forest | 0.8909 | 0.9064 | 0.8707 | 0.8882 |
| XGBoost | 0.8896 | 0.9076 | 0.8664 | 0.8865 |
| Ensemble | 0.8874 | 0.8747 | 0.9027 | 0.8885 |

#### REFERENCES

1. G. Shmueli, “To explain or to predict?” *Statistical Science*, vol. 25, no. 3, 2010. [Online].

Available: https://projecteuclid.org/journals/statistical-science/volume-25/ issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full

1. S. Chatterjee, K. Khunti, and M. J. Davies. (2017) Type 2 diabetes. [Online]. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/ https://somepomed.org/pdf/abril/Type-2-Diabetes-2017.pdf

TABLE II: Model Evaluation Metrics Using Laboratory Data for 1988-2018 Diabetes Case II

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | ROC-AUC | Precision | Recall | F1 |
| Logistic Regression | 0.7557 | 0.7939 | 0.7482 | 0.7704 |
| SVM | 0.8020 | 0.8296 | 0.8058 | 0.8175 |
| Random Forest | 0.9920 | 0.9928 | 0.9928 | 0.9928 |
| XGBoost | 0.9928 | 1.00 | 0.9856 | 0.9928 |
| Ensemble | 0.9629 | 0.9578 | 0.9784 | 0.9680 |

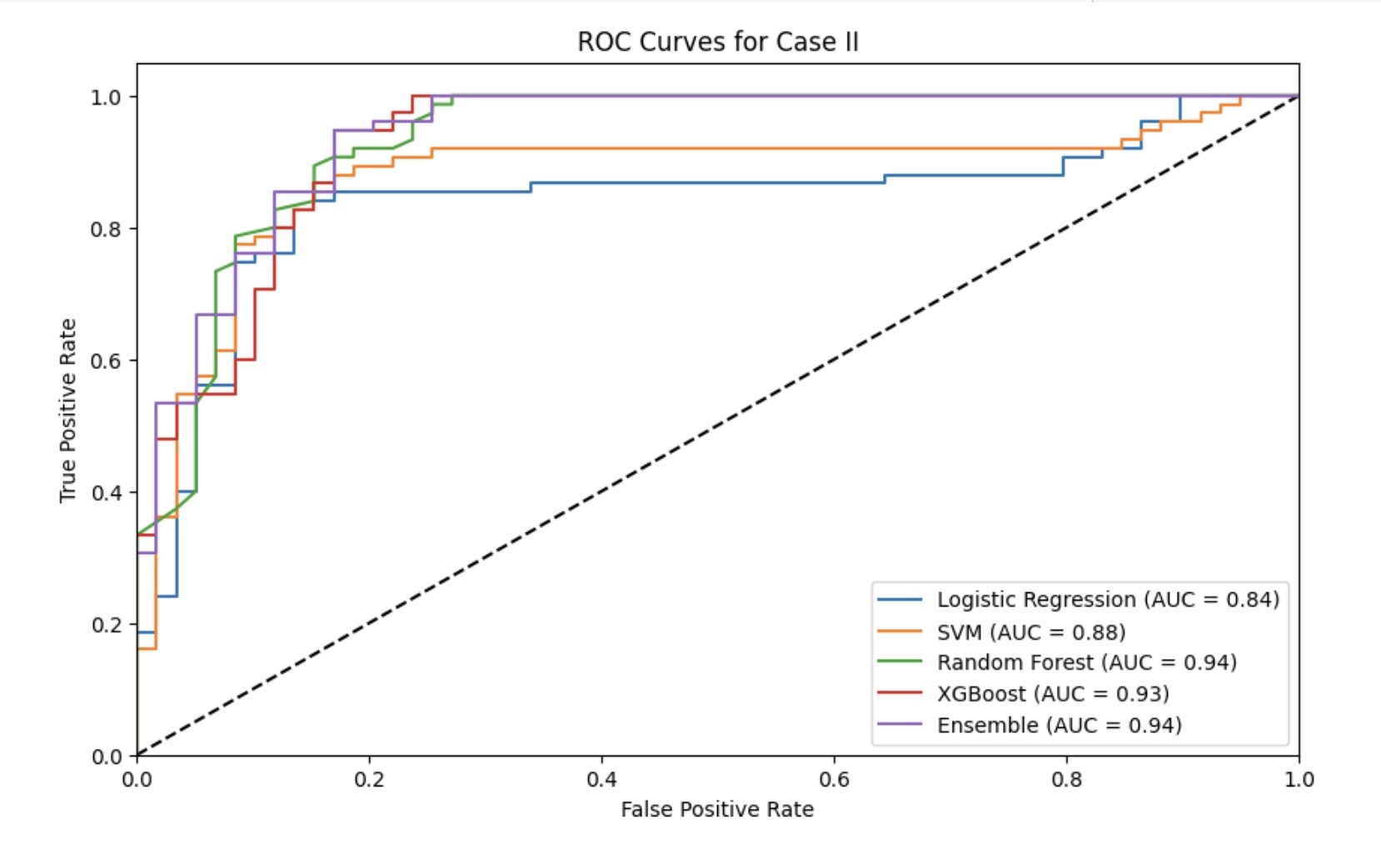
Fig 1. ROC Curves for Case I

Fig 2. ROC Curves for Case II

1. J. Semerdjian and S. Frank. (2017) An ensemble classifier for predicting the onset of type ii diabetes. [Online]. Available: https: //arxiv.org/abs/1708.07480
2. A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, “A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,” *BMC Medical Informatics and Decision Making*, 2019. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/ articles/10.1186/s12911-019-0918-5
3. W. Y. et al., “Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes,” *BMC Medical Informatics and Decision Making*, vol. 10, no. 16, 2010. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/ articles/10.1186/1472-6947-10-16
4. M. Teimouri and M. Alavinia. (2015) Comparison of various classification algorithms in the diagnosis of

type 2 diabetes in iran. [Online]. Available: https:

#### //www.researchgate.net/publication/277634174 Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran

1. Y.-W. Chen and C.-J. Lin, “Combining svms with various feature selection strategies,” in *Feature Extraction*. Springer, 2006. [Online]. Available:

https://link.springer.com/chapter/10.1007/978-3-540-35488-8 13

1. (2023) National health and nutrition examination survey (nhanes). Office of Disease Prevention and Health Promotion, U.S. Department of Health and Human Services. [Online]. Available: https://

health.gov/healthypeople/objectives-and-data/data-sources-and-methods/

data-sources/national-health-and-nutrition-examination-survey-nhanes

1. A. Heredia-Langner, K. H. Jarman, B. G. Amidan, and J. G. Pounds,

“Genetic algorithms and classification trees in feature discovery: diabetes and the nhanes database,” in *Proceedings of the International Conference on Data Mining (DMIN)*, 2013, p. 1.