International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I11P122

Volume 72 Issue 11, 192-219, November 2024

Original Article

A Comprehensive Analysis of Diabetes Risk Prediction Using Logistic Regression

Asish Pradhan

Technical Architect at Dell Technologies, USA.

Corresponding Author : Asish_Pradhan@Dell.com

Received: 08 October 2024 Revised: 09 November 2024 Accepted: 25 November 2024 Published: 30 November 2024

Abstract *- This study aimed to develop a predictive model for diabetes risk using a combination of demographic, examination, diet, and laboratory data. The dataset was processed through ETL (Extract, Transform, Load) and EDA (Exploratory Data Analysis) to identify potential correlations. A logistic regression model was built and evaluated using various metrics, achieving an accuracy of approximately 93%. The results indicate that the model can accurately predict diabetes risk, making it a valuable tool for healthcare professionals. The study demonstrates a comprehensive approach to building a predictive model for diabetes risk using a multidimensional dataset with potential applications in healthcare.*

Keywords *- Diabetes prediction, Exploratory data analysis, Healthcare data analysis, Logistic regression, Machine learning, Monte Carlo simulation.*

1. Introduction

The National Center for Health Statistics (NCHS) is a part of the Centers for Disease Control and Prevention (CDC) and responsible for producing vital health statistics for the United States. One of the NCHS's most significant programs is the National Health and Nutrition Examination Survey (NHANES), a program of studies designed to assess the health and nutritional status of adults and children in the United States. The NHANES program began in the early 1960s as a series of surveys focusing on different population groups or health topics. In 1999, the survey became a continuous program with a changing focus on various health and nutrition measurements to meet emerging needs. Since then, it annually examines a nationally representative sample of about 5,000 people nationwide. The NHANES is a unique survey combining interviews and physical examinations. The interviews include demographic, socioeconomic, dietary, and health-related questions.

The examination components consist of medical, dental, and physiological measurements and laboratory tests administered by highly trained medical personnel. Diabetes is a serious health condition that affects how the body metabolizes food, resulting in high blood sugar levels. According to the CDC, between 2013 and 2016, 12% of adults in the United States had diabetes, with the estimated percentage of adults diagnosed with diabetes being 9.4% and the estimated percentage of adults with undiagnosed diabetes being 2.6%. Unmanaged diabetes can lead to serious health problems like heart disease, vision loss, and kidney disease.

A predictive model for diabetes diagnosis is needed to help patients at risk of developing diabetes implement lifestyle changes that prevent diabetes or at least make living with diabetes easier, working against more harmful medical issues, like heart or kidney disease. Prior studies in this field focus on creating models with the given data without giving necessary attention to data sampling, yielding a model sensitive to the training data. A logistic regression model was developed to predict the risk of developing diabetes using demographic, examination, diet, and laboratory data from the NHANES program. With the introduction of bootstrap and Monte Carlo simulation methodologies, reliable prediction interval estimates and the stability of the predictions under diverse scenarios are ensured. This model was developed using a dataset of approximately 5,000 individuals and was found to have a high accuracy in predicting the risk of developing diabetes, with an accuracy of roughly 94%.

The model considers various demographic, examination, diet, and laboratory data, providing comprehensive information on the factors contributing to the risk of developing diabetes. Creating a successful predictive model for diabetes diagnosis will help patients at risk of developing diabetes implement lifestyle changes that prevent diabetes or at least make living with diabetes easier, working against more harmful medical issues, like heart or kidney disease. By developing a predictive model, healthcare providers can identify individuals at risk of developing diabetes and provide targeted interventions and preventative care, reducing the incidence and prevalence of the diabetic condition.

2. Literature Review

Diabetes is a complex and multifactorial health risk that affects millions of people worldwide. Predicting diabetes risk has been an active area of research in recent years. Diabetes prediction has been an active area of research in recent years, focusing on developing accurate and reliable early detection and prevention models. The literature on diabetes prediction is extensive and diverse, with various approaches and methods employed to identify individuals at risk of developing the disease.For diabetes risk predictions, traditional approaches mainly relied on statistical models, such as logistic regression, to identify risk factors and predict these risk outcomes due to its simplicity, interpretability, and ability to provide probabilistic outputs. For example, studies such as Sisodia and Sisodia [1] employed logistic regression on the Pima Indians Diabetes Dataset (PIDD) to predict diabetes risk, demonstrating its effectiveness in classifying outcomes based on clinical and demographic features.

Exploratory Data Analysis (EDA) and data preprocessing are vital steps in ensuring the quality and relevance of features used for modeling. The Extract, Transform, Load (ETL) process is often employed to handle complex datasets by integrating structured data (e.g., lab results) and unstructured data (e.g., physician notes) [2]. Prior studies have mostly worked on modeling, with a little focus on performing ETL and EDA on diverse data. Using diverse data allows for analyzing the data from a wider angle and increases the possibility of getting a more practical model. Both ETL and EDA work together to ensure that data is clean, consistent, and ready for meaningful insights by first exploring and understanding its patterns, then systematically integrating and transforming it from various sources into a usable format for analysis; essentially, EDA helps in discovering valuable information within the given data, while ETL prepares the data to be analyzed effectively, leading to more accurate and reliable results for decision-making. Prior studies have shown steps to perform EDAs on the type-2 diabetes dataset and with a smaller set of factors [21,22].

It is found that the bootstrap estimates and Monte Carlo simulations are advanced statistical methods that have been applied sparingly in diabetes prediction models. Bootstrap methods, as described by Efron and Tibshirani [10], involve resampling data with replacement to provide more robust confidence intervals and evaluate the variability of model parameters. Similarly, Monte Carlo simulations offer a systematic way to evaluate parameter distributions and assess model stability under different assumptions [3]. These techniques enhance model reliability, particularly in clinical contexts where decisions depend on robust statistical evidence. The concept of learning curves is commonly used in machine learning work to understand the relationship between model performance and training data. However, this concept is underutilized in diabetes prediction studies. Utilization of learning curves helps diagnose underfitting and overfitting issues in a model, providing critical insights into model generalizability and data sufficiency [11]. Their application to diabetes prediction offers an opportunity to optimize resource allocation for data collection and model training.

2.1. Novelty and Comparison with Existing Research

This paper advances the state of research in diabetes prediction using a few different methodologies. The prior studies focus more on the modeling without extensive use of the data sampling methods. The results, therefore, carry the risk of being sensitive to the trained dataset. While logistic regression has been extensively studied, this work extends its utility by incorporating bootstrap methods that help in estimating the variability and reliability of model coefficients, errors, etc., without relying on strong assumptions about the underlying data distribution that results in achieving more reliable interval estimates for predictions which then enhances the model's interpretability and trustworthiness in clinical applications. Furthermore, using Monte Carlo simulations adds a layer of robustness by simulating parameter variations to evaluate the stability of the predictions under diverse scenarios by generating multiple possible outcomes based on random sampling.

This technique ensures that the generated model is not sensitive to any specific data and, therefore, becomes dependable across different population subsets, a critical aspect of real-world healthcare systems [3]. The paper's ETL pipeline and detailed EDA on diverse data ensure the high quality and relevance of input features, laying a solid foundation for modeling. This study greatly uses the concept of learning curves by examining the influence of training data size on model performance. Gradual examination of the effects of the varying data size on model accuracy, this study provides valuable insights into the optimal data requirements needed to overcome underfitting and overfitting issues. Together, these elements contribute to developing a robust and reliable predictive model, advancing the applicability of machine learning and statistical methods in diabetes risk prediction.

3. Exploratory Data Analysis (EDA)

3.1. Dataset

The dataset obtained from Kaggle is the National Health and Nutrition Examination Survey (NHANES) dataset, a comprehensive program of studies designed to assess the health and nutritional status of adults and children in the United States. Conducted by the National Center for Health Statistics (NCHS), a part of the Centers for Disease Control and Prevention (CDC), NHANES combines interviews and physical examinations to provide a nationally representative sample of approximately 5,000 persons each year. The survey's continuous program, initiated in 1999, focuses on various health and nutrition measurements to meet emerging needs, examining a sample of individuals from 15 counties nationwide. The dataset includes demographic,

socioeconomic, dietary, and health-related questions, medical, dental, and physiological measurements, and laboratory tests trained medical personnel to administer. The 2013-2014 NHANES dataset contains the following components:

3.1.1. Demographics Dataset

This dataset contains columns such as age, sex, race, ethnicity, and socioeconomic status.

3.1.2. Examinations Dataset

The examination dataset contains variables related to physical examinations, including:

- Blood pressure
- Body measures (e.g. height, weight)
- Muscle strength (grip test)
- Oral health (dentition)
- Taste and smell
- More (see link for complete list)
- Dietary data: contains variables related to dietary intake, including:
- Total nutrient intake
- First-day dietary data
- Food security

3.1.3. Laboratory Dataset

This dataset contains variables related to laboratory tests, including:

- Albumin and creatinine levels (urine)
- Apolipoprotein B
- Blood lead, cadmium, total mercury, selenium, and manganese levels
- Blood mercury levels (inorganic, ethyl, and methyl)
- Cholesterol levels (HDL, LDL, triglycerides, total)

3.1.4. Medication Dataset

The medication dataset contains prescription medication data.

3.1.5. Questionnaire Dataset

This dataset contains variables related to health and lifestyle, including:

- Acculturation
- Alcohol use
- Blood pressure and cholesterol levels
- Cardiovascular health
- Consumer behavior
- Current health status
- Dermatology
- And for more information (see link for complete list),

3.2. Causal Loop Diagrams

Causal loop diagram is a valuable tool in data analysis, allowing researchers to visualize and analyze the complex relationships between various factors contributing to the development and progression of the diabetic condition. By mapping out the causal links between glucose levels, insulin levels, and insulin sensitivity, it helped to identify the underlying causes of diabetes and develop targeted interventions to address them. The benefits of causal loop diagrams in this study are improved understanding of the complex relationships between variables, better prediction of the likelihood of developing diabetes, identification of potential targets for treatment, evaluation of the effectiveness of interventions, and identification of potential confounding variables and mediators.

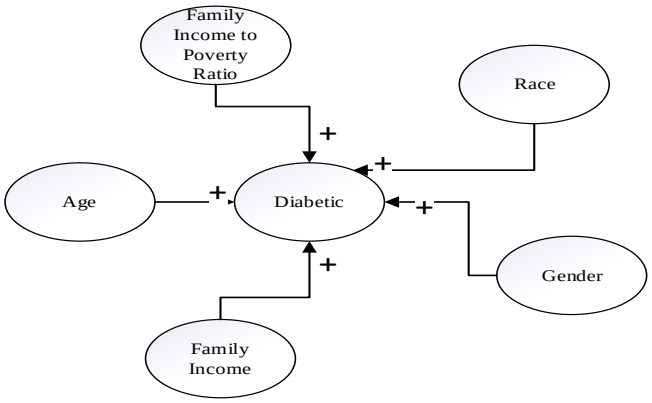


Fig. 1 Causal loop diagram from demographic data

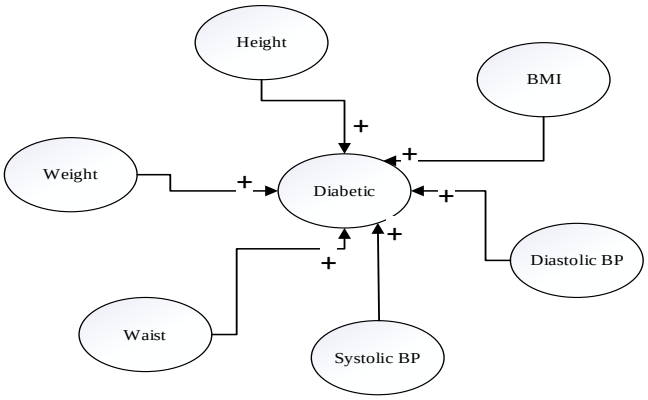


Fig. 2 Causal loop diagram from examination data

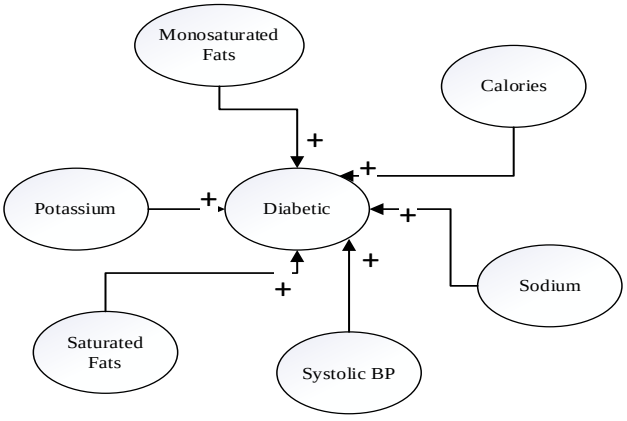


Fig. 3 Causal loop diagram from diet data

Fig. 4 Causal loop diagram from lab data

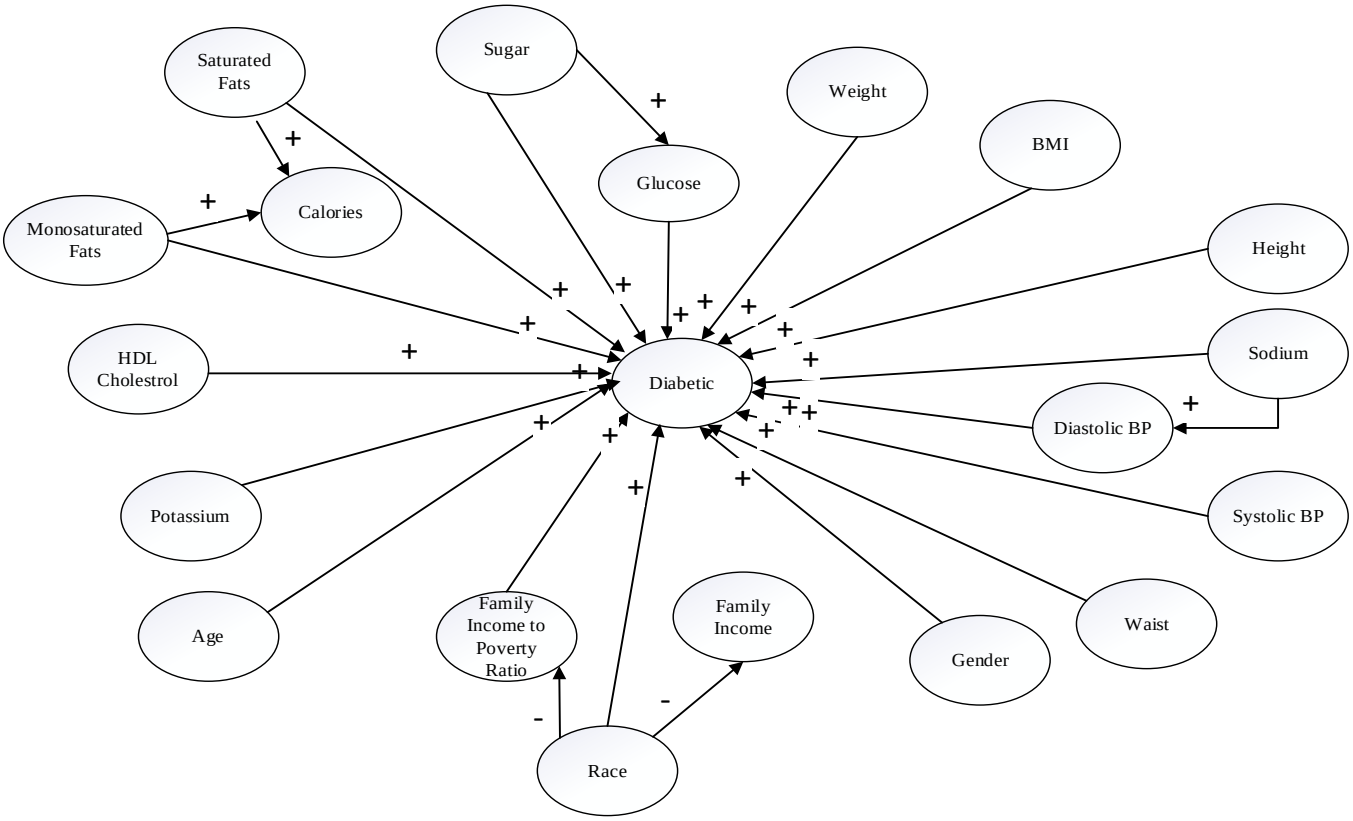


Fig. 5 Complete causal loop diagram using all variables

3.3. Extract, Transform and Load (ETL)

The original NHANES dataset consists of hundreds of variables related to health, lab results, demographics, and socioeconomic status. Since the focus is on studying diabetes, not all these variables were considered necessary. The most relevant variables were identified during the ETL process, and some demographic and socioeconomic attributes were included to explore potential relationships with a diabetic diagnosis. Variables with many missing responses, such as those related to pregnancy, were excluded.

The relevant data was extracted from CSV format and loaded into several tables in an SQLite database, establishing relationships between tables as needed.

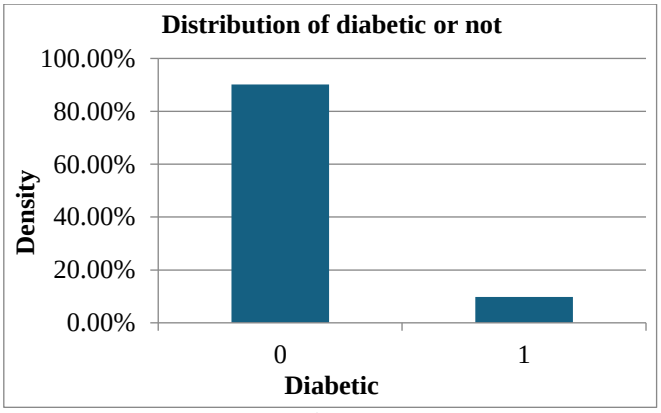


Fig. 6 Distribution of target variable diabetic

3.4. Descriptive Analysis – Single Variable EDA

3.4.1. Target Variable: Diabetic

Of the participants in the survey, 90.2% have not been diagnosed with diabetes, while 9.8% have been diagnosed with diabetes.

3.4.2. Calories

The minimum and maximum calorie intake values are 117 and 12108 kcal, respectively. The mean calorie intake is 2108 kcal, while the median is 1930 kcal. Notably, the mean is greater than the median, suggesting a right-skewed

distribution. Further examination of the data reveals that the value of 117 kcal is extremely low, which may indicate an input error. A visual examination of the box plot (Figure 1) reveals outliers above approximately 3980 kcal. The histogram above appears right-skewed and bell-shaped, with most data points falling between 1500 and 2000 kcal. Additionally, a small presence of calorie information beyond 4000 kcal is observed.

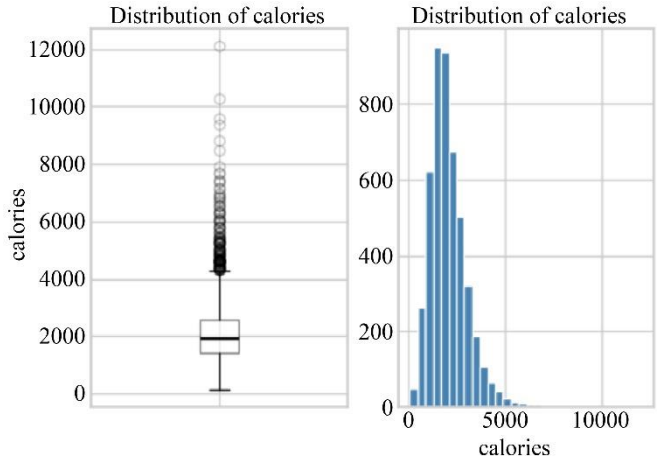


Fig. 7 Distribution of variable calories

3.4.3. Monounsaturated Fats

The monounsaturated fats attribute is a numerical variable. It represents the total monounsaturated fatty acids (gm) consumed the day before the interview. Monounsaturated fats are known as "good fats," such as those in olive oil and avocados. They are good for cardiac health. The recommended daily intake is 16-22 grams. The descriptive statistics for the monounsaturated fatty acid intake data indicate that the minimum and maximum values are 0 and 221 grams, respectively. The mean monounsaturated fatty acid intake is 28 grams, while the median is 24.6 grams. Notably, the mean exceeds the median, suggesting a potential right-skewed distribution. Further examination of the data reveals that the value of 0 grams may indicate an input error. The range is 221.67 grams, i.e. the difference between the max and min values. The interquartile range is 19.728 grams, and the coefficient of variance (standard deviation over mean in percentage) is 63.59%. A visual examination of the box plot reveals outliers above approximately 52 grams, which are not being ignored. The histogram appears right-skewed and bell-shaped, with most data points falling between 20 and 30 grams. Additionally, a small presence of monounsaturated fatty acid information beyond 70 grams is seen. Refer (Appendix 1) for plots.

3.4.4. Potassium

Potassium represents the amount of Potassium consumed in a day. Potassium is a numerical variable measured in mg. The recommended intake is 3,500–4,700 mg per day. Potassium is integral to a healthy diet because it helps normalize cell functionality. The descriptive statistics for the potassium data reveal that the minimum and maximum values are 110 and 15,876 milligrams, respectively. The mean potassium intake is 2,533 milligrams, while the median is 2,344 milligrams. A right-skewed distribution is expected because the mean is greater than the median. Additional descriptive statistics include a range of 15,766 milligrams, an interquartile range of 1,461 milligrams, and a coefficient of variance of 50.27%. The box plot and histogram (Appendix 1) reveal outliers above approximately 5,000 milligrams. The histogram is right-skewed and bell-shaped, with most data points falling between 1,500 and 2,000 milligrams. A small presence of potassium data points beyond 5,000 milligrams is also observed.

3.4.5. Saturated Fats

Attribute Saturated fats is a numerical variable. It represents total saturated fatty acids (gm). Unlike mono-unsaturated fats, saturated fats are associated with heart disease. Medical experts recommend limiting saturated fat intake. The saturated fatty acid data exhibits a range of 177.4 grams, with a minimum value of 0 grams and a maximum value of 177.4 grams. The mean intake is 26.09 grams, while the median is 22.84 grams. A right-skewed distribution is likely because the mean is greater than the median. Further descriptive statistics reveal an interquartile range of 19.027 grams and a coefficient of variance of 63.57%. The box plot and histogram (Appendix 1) illustrate outliers above approximately 52 grams. The histogram appears right-skewed and bell-shaped, with most data points falling between 15 and 25 grams. A small presence of saturated fatty acid data points beyond 60 grams is also observed.

3.4.6. Sugar

Sugar is a numerical variable. It represents the total grams of sugar consumed per day (both processed and non-processed). Processed sugars are hard for the body to regulate. Over consumption has long been correlated with health problems such as obesity and diabetes. The sugar data exhibits a range of 979.39 grams, with a minimum value of 0.13 grams and a maximum value of 979.39 grams. The mean sugar intake is 110.97 grams, while the median is 95.39 grams. A right-skewed distribution is expected if the mean is greater than the median. Additional descriptive statistics include an interquartile range of 84.05 grams and a coefficient of variance of 68.87%. The box plot and histogram (Appendix 1) reveal outliers above approximately 220 grams. The histogram is right-skewed and bell-shaped, with many data points falling between 50 and 60 grams. A small presence of sugar data points beyond 70 grams is also observed.

3.4.7. Sodium

Sodium is a numerical variable. Sodium in a participant's diet, measured in milligrams (mg). While sodium is important for maintaining fluid balance in the body, many Americans over consume sodium. The recommended intake is 2,300 mg

per day. Excess sodium can lead to high blood pressure and other health problems. The sodium data exhibits a mean of 3505.167 mg, which exceeds the recommended daily intake. The median value of 3169.00 mg also exceeds this recommendation. The 1st quartile is 2264.00 mg, while the 3rd quartile is 4355.00 mg. The minimum value is 29.00 mg, while the maximum value is 21,399.00 mg, resulting in a range of 21,370.00 mg. The interquartile range is 2,091.00 mg, and the coefficient of variance is 53.1%. The box plot and histogram (Appendix 1) reveal a few apparent outliers at approximately 20,000 mg. The histogram appears right-skewed and bell-shaped, with most data points clustering around 2,500 mg.

3.4.8. Diastolic Blood Pressure

Diastolic BP is a numerical variable for diastolic blood pressure (measured in mm Hg). Diastolic blood pressure is the pressure that is released when the heart fills with blood. A normal diastolic blood pressure for an adult is under 80 mmHg. High blood pressure is a cause for concern as it increases the risk of several medical emergencies, such as heart attacks and strokes. The diastolic blood pressure (BP) data exhibits a mean of 67.54 mm Hg and a median of 68.00 mm Hg. The first quartile is 60.00 mm Hg, while the 3rd quartile is 76.00 mm Hg. The minimum value is 0.00 mm Hg, while the maximum value is 120.00 mm Hg, resulting in a range of 120 mm Hg. The interquartile range is 16.00 mm Hg, and the coefficient of variance is 20.36%. The box plot and histogram (Appendix 1) reveal an even distribution of data points with no apparent outliers. The histogram appears to be bell-shaped, with a peak in data around 65 mm Hg. Notably, a single data point at 0.00 mm Hg appears anomalous for a diastolic BP value.

3.4.9. Systolic Blood Pressure

Systolic BP is a numerical variable for systolic blood pressure (measured in mm Hg). Systolic blood pressure is the pressure in the heart from contracting. The target systolic BP is 120 mmHg. High blood pressure is a cause for concern as it increases the risk of several medical emergencies, such as heart attacks and strokes. The systolic blood pressure (BP) data exhibits a mean of 119.81 mm Hg and a median of 116.00 mm Hg. The 1st quartile is 108.00 mm Hg, while the 3rd quartile is 128.00 mm Hg. The minimum value is 74.00 mm Hg, while the maximum value is 228.00 mm Hg, resulting in a range of 154.00 mm Hg. The interquartile range is 20.00 mm Hg, and the coefficient of variance is 14.33%. The box plot and histogram (Appendix 1) reveal a possible presence of a couple of outliers above 200 mm Hg. The box plot shows a wider distribution of points above the box, whereas the points below the box appear to be tighter and less dispersed. The histogram is bell-shaped, with a peak in data around 125 mm Hg.

3.4.10. Abdominal Diameter

Abdominal diameter is a numerical variable for the average sagittal abdominal diameter (measured in cm). Studies have investigated the potential of the sagittal abdominal diameter as a predictor of incident diabetes, suggesting a possible link between this measure and the development of the diabetic condition. [18] The abdominal diameter data exhibits a mean of 21.75 cm and a median of 21.30 cm. The 1st quartile is 18.20 cm, while the 3rd quartile is 24.70 cm. The minimum value is 11.90 cm, while the maximum value is 40.10 cm, resulting in a range of 28.20 cm. The interquartile range is 6.50 cm, and the coefficient of variance is 21.27%. The box plot and histogram (Appendix 1) reveal a possible presence of a couple of outliers at approximately 40 cm. The box plot shows a wider distribution of points above the box, whereas fewer points are below. The histogram appears to be somewhat bell-shaped, with a peak in data around 23 cm.

3.4.11. BMI

BMI is a numerical variable. BMI, or Body Mass Index, is one indicator of overall health. According to Wikipedia, it is calculated as $BMI = \frac{mass}{height}$. A BMI of 25 and over is categorized as overweight. Many health problems, including diabetes, have been linked to having a BMI that is considered overweight. Note that in recent years, BMI has been criticized as an indicator of health since it does not consider body type or composition. The body mass index (BMI) data exhibits a mean of 27.83 $kg/m^2$and a median of 26.90 kg/m^2. The 1st quartile is 22.90 kg/m^2, while the 3rd quartile is 31.40 kg/m^2. The minimum value is 13.40 kg/m^2, while the maximum value is 67.50 kg/m^2, resulting in a range of 54.10 kg/m^2. The interquartile range is 8.50 kg/m^2, and the coefficient of variance is 24.44%. The box plot and histogram (Appendix 1) reveal a distribution with all points above the box, with increasing spread as the BMI values increase. The histogram appears to be somewhat bell-shaped, with a peak in data around 30 kg/m^2.

3.4.12. Height

Height is a numerical variable for the participant's standing height (measured in cm). The height data exhibits a mean of 167.24 cm and a median of 166.80 cm. The 1st quartile is 160.10 cm, while the 3rd quartile is 174.30 cm. The minimum value is 136.30 cm, while the maximum value is 202.60 cm, resulting in a range of 66.30 cm. The interquartile range is 14.20 cm, and the coefficient of variance is 5.97%. The box plot and histogram (Appendix 1) reveal a distribution with most points below the box with a few above. The histogram is bell-shaped, consistent with the expected normal distribution of height data.

3.4.13. Weight

Weight is a numerical variable for the participant's weight (measured in kg). The weight data exhibits a mean of 78.23 kg and a median of 75.40 kg. The 1st quartile is 63.10 kg, while the 3rd quartile is 90.10 kg. The minimum value is 29.20 kg, while the maximum value is 195.40 kg, resulting in a range of

166.20 kg. The interquartile range is 27.00 kg, and the coefficient of variance is 27.62%. The box plot and histogram (Appendix 1) reveal a distribution with a few points below the box but most points above the box. The histogram appears bell-shaped, with spikes to the left of the curve that may indicate outliers.

3.4.14. High-Density Lipoprotein (HDL)

HDL, or high-density lipoprotein, is known as "good cholesterol". It is known to improve heart and liver functionality. It is measured in mg/dL. 60 mg/dL or higher is the desirable amount. High-Density Lipoprotein (HDL) cholesterol levels range from 10-173 mg. The mean HDL level is 50 mg, close to the median of 52.67 mg. The range of HDL levels is 163 mg, indicating significant variation in the data. The interquartile range (IQR) is 19 mg, which is smaller than the range, suggesting that the data is not excessively skewed. The coefficient of variance (COV) is 0.29, indicating that the variance is 29% of the mean. The distribution of HDL levels is roughly bell-shaped, with a long right skew and several high outliers.

3.4.15. Cholesterol

The cholesterol variable measures the total amount of cholesterol in the blood. It is measured in mmol/L. For adults, under 200 mmol/L is considered healthy. High cholesterol is associated with heart disease.The mean cholesterol level is 222 mmol/L, with a minimum value of 0 and a maximum of 3515 mmol/L. However, the minimum value of 0 is likely an error, as, practically, cholesterol levels cannot be zero. The range of the cholesterol data is 3515 mmol/L, including outliers. The interquartile range (IQR) is 269 mmol/L, and the coefficient of variance (COV) is 0.8592. The cholesterol data exhibits a heavy upward skew, with most data points falling on the lower side of the distribution. The box plot indicates that the observation of 3515 mmol/L is a highly high outlier, representing an error or an unusual case.

3.4.16. Glucose

Glucose, or blood sugar, is a vital energy source from human food. The glucose variable is measured in millimoles per liter (mmol/L). The mean glucose level is 5.59 mmol/L, with a minimum value of 2.72 mmol/L and a maximum value of 32.03 mmol/L. The standard deviation is 1.976 mmol/L. The range is 29.31 mmol/L, indicating significant variability in the data. The interquartile range (IQR) is 0.89 mmol/L, suggesting that most data points are clustered on the lower side of the distribution. The plots confirm that most of the data is on the low side, with a widespread in the upper quartile. Although there are some outliers, there are no glaringly obvious ones.

3.4.17. Age

The first demographic variable is age, representing the years a participant has lived. The participants in the study range in age from 12 to 80 years.The mean age is 41.39 years, with a median age of 40. The range of ages is 68 years, indicating a significant amount of variability in the data. The interquartile range (IQR) is 35 years, suggesting that most data points are clustered around the median. The coefficient of variation (COV) is 0.4876, indicating moderate variation in the data.

3.4.18. Family Income

Family income represents the annual household income, which was categorized into ranges during the data processing stage (ETL). The categories were mapped back from integer values to string category names. The family income measurements are binned at $5,000 and $10,000 intervals. Notably, a significant proportion of participants did not provide their actual income, and therefore, the income bin was estimated based on responses to various questions. This may complicate the analysis and regression modeling. Interestingly, the group with an income of $100,000 or more had the most significant observations.

3.4.19. Family Income Poverty Ratio

The family income poverty level is a ratio of family income to the poverty cut-off, with a range of 0-4.99. Observations with values greater than 4.99 were recoded as 5, resulting in a range of 0-5. The average family income poverty ratio is 3.94. The distribution of the variable is roughly uniform, with a notable peak at 5, likely due to the recoding of values greater than 5 to 5.

3.4.20. Gender

The gender variable captures the self-reported gender of the participant, with options being male or female. The survey did not include other gender identities, so any such responses would not be represented in the data. The gender distribution is roughly evenly split, with a slight majority of male participants, with 2390 females and 2423 males.

3.4.21. Race

The demographic variable is a categorical variable that captures the self-reported demographic background of the participant. Due to the limited options available, many demographic categories were aggregated into the "other" category. Most of the sample consists of individuals from diverse backgrounds, comprising approximately 42% of the participants. The second largest group is individuals from African American backgrounds, followed by other demographic categories.

3.5. Descriptive Analysis - Pairwise EDA

3.5.1. Diabetic Vs Calories

The first relationship examined is the association between diabetes and calorie consumption. Given the link between diabetes and unhealthy lifestyle habits, it was expected that individuals with diabetes would tend to consume more calories than those without diabetes. However, the data reveals a surprising finding: the mean calorie intake for individuals

with diabetes is 1871, while the mean for those without diabetes is 2134. Notably, both means exceed their respective medians, indicating right skewness in both distributions. Furthermore, the plots exhibit a normal distribution with right skewness, suggesting that the data is concentrated around the mean. This finding contradicts the initial hypothesis, highlighting the importance of exploring data-driven insights rather than relying solely on theoretical expectations.

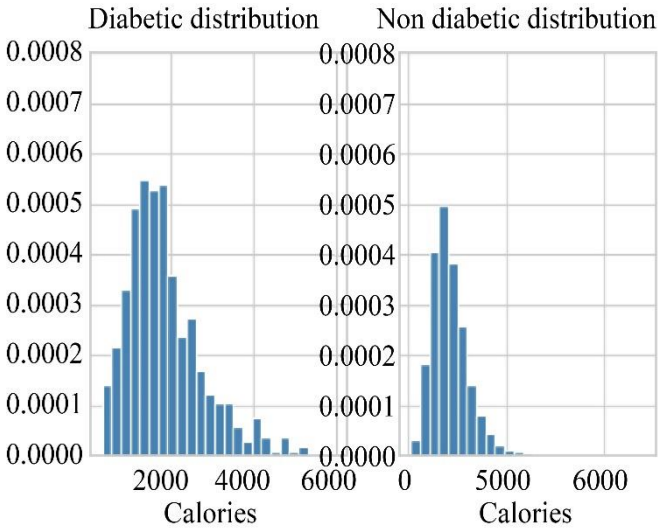


Fig. 8 Distribution of target variable diabetic against calories

3.5.2. Diabetic Vs Monounsaturated fats

Monounsaturated fats are considered "healthy fats" and are not expected to be associated with poor health outcomes. The mean monounsaturated fat intake for individuals with diabetes is 28.436 grams, while for those without diabetes, it is 26.48 grams. As in the previous cases, both means exceed their respective medians, indicating right skewness in both distributions. Furthermore, the plots exhibit a normal distribution with right skewness, suggesting that the data is concentrated around the mean. Consistent with expectations, the data does not show a strong correlation between monounsaturated fat intake and diabetes.

3.5.3. Diabetic Vs Potassium

The relationship between potassium levels and diabetes is of interest, as low levels of potassium have been linked to degraded cell performance. Therefore, it was hypothesized that individuals with diabetes would have lower potassium levels than those without diabetes. The means for potassium levels were 2426.9 mg for people with diabetes and 2544 mg for non-diabetics, with medians of 2349 mg and 2341 mg, respectively. However, a statistical analysis revealed no significant difference between the means. The plots for diabetics and non-diabetics exhibit a similar distribution, with samples concentrated around the mean and a bell-shaped distribution with right skewness. This suggests no significant difference in potassium levels between individuals with and without diabetes.

3.5.4. Diabetic Vs Saturated Fats

Saturated fats are known to affect health negatively, and it was expected that individuals with diabetes would have higher saturated fat intake. However, the data reveals a surprising finding: the mean consumption of saturated fats for the diabetic group was 23.27 grams, more than 3 grams less than the average for the non-diabetic group (26.4 grams), which contradicts the initial assumption. Notably, the saturated fat intake distribution for diabetics and non-diabetics appears to be normal with right skewness, with samples concentrated around the mean.

3.5.5. Diabetic Vs Sugar

Previous research has established a correlation between diets high in sugar and an increased risk of developing diabetes. Consequently, it was expected that the diabetic survey participants would have higher sugar intakes. However, the data reveals a surprising finding: the mean amount of sugar consumed per day among people with diabetes was 85.4 grams, which is lower than the mean of 113.74 grams among non-diabetic individuals. This may be because the diabetic group is defined as "has been diagnosed with diabetes", and therefore, individuals in this group may be actively managing their condition by reducing their sugar intake. The mean is greater than the median in both cases, indicating right skewness in both distributions. Furthermore, the plots are normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.6. Diabetic Vs Sodium

The relationship between sodium intake and diabetes has been explored. As diets high in sodium have been linked to blood pressure and heart disease, which are common comorbidities of diabetes, it was expected that the diabetic group would have higher sodium intake. However, the data reveals a surprising finding: the mean sodium intake for the diabetic group (3328.15 mg) is lower than that of the non-diabetic group (3524.37 mg). Furthermore, the median sodium intake for the diabetics is also lower than that of the non-diabetics. This goes against the initial prediction. The mean is greater than the median in both cases, indicating right skewness in both distributions. Additionally, the plots appear normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.7. Diabetic Vs Diastolic BP

The relationship between systolic blood pressure and diabetes has been explored. As heart disease is a common comorbidity of diabetes, it was expected that the diabetic group would have higher systolic blood pressure. However, the data reveals a surprising finding: the mean diastolic BP level for the diabetic group (68.87 mm Hg) is only slightly higher than that of the non-diabetic group (67.39 mm Hg). Notably, the mean is close to the median in both cases, indicating a symmetrical distribution with no skewness expected. Furthermore, the plots appear normal with right

skewness, suggesting that the data is concentrated around the mean. However, the data point of 0 mm Hg in both groups may be an outlier that warrants further investigation.

3.5.8. Diabetic Vs Systolic BP

The relationship between systolic blood pressure and diabetes has been explored. As heart disease is a common comorbidity of diabetes, it was expected that the diabetic group would have higher systolic blood pressure. The data reveals a finding consistent with this expectation: the mean systolic BP level for the diabetic group (129.54 mm Hg) is higher than that of the non-diabetic group (118.75 mm Hg).

Furthermore, the mean is greater than the median for the people with diabetes, indicating right skewness in the distribution. Conversely, the mean is less than the median for the non-diabetics, suggesting the opposite skewness. The center of the distribution for people with diabetes is indeed higher than that for non-diabetics, as expected. The plots appear normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.9. Diabetic Vs Abdominal Diameter

The relationship between abdominal diameter and diabetes has been previously studied, with higher abdominal diameter being identified as a potential risk factor for developing diabetes. Consistent with this finding, the data reveals a positive association between abdominal diameter and diabetes. The mean abdominal diameter for the diabetic group (25.78) is higher than that for the non-diabetic group (21.31). Moreover, the mean is slightly greater than the median for the people with diabetes, indicating a slight right skewness in the distribution. Conversely, the mean is less than the median for the non-diabetics, suggesting the opposite skewness. The center of the distribution for people with diabetes is indeed higher than that for non-diabetics, as expected. The plots appear normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.10. Diabetic Vs BMI

The relationship between body mass index (BMI) and diabetes has been well-established, with unhealthy lifestyles and obesity being common risk factors for developing the disease. Consistent with this expectation, the data reveals a positive association between BMI and diabetes. The mean BMI for the diabetic group (31.98 kg/m^2) is higher than that for the non-diabetic group (27.38 kg/m^2). Moreover, the mean is greater than the median for the people with diabetes, indicating a right skewness in the distribution. Conversely, the mean is less than the median for the non-diabetics, suggesting the opposite skewness. The center of the distribution for people with diabetes is indeed higher than that for non-diabetics, as expected. The plots appear normal with right skewness, suggesting that the data is concentrated around the mean. Notably, there appear to be some outliers in the diabetic distribution with BMIs greater than 60 kg/m^2.

3.5.11. Diabetic Vs Height

As previously observed, height follows a normal distribution. There is no established link between height and severe health problems. The data reveals that the mean height for people with diabetes (167.02 cm) is slightly lower than that for non-diabetics (167.27 cm). In the case of diabetics, the median and mean are nearly identical, indicating no skewness in the distribution. In contrast, the mean is greater than the median for non-diabetics, suggesting right skewness in the distribution. Notably, there is no significant difference in the distributions. The plots appear normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.12. Diabetic Vs Weight

While height is necessary in calculating BMI to determine whether someone is overweight, it is well-established that high weight is a strong indicator of being overweight. Therefore, it is reasonable to expect that the diabetic group would have a higher weight, given the association between weight and the risk of developing diabetes. The data supports this expectation, with the mean weight for people with diabetes (89.76 kg) being significantly higher than that for non-diabetics (76.98 kg). Notably, the weight distribution for people with diabetes exhibits left skewness, with the mean being less than the median. In contrast, the weight distribution for non-diabetics shows a slight right skewness, with the mean being slightly greater than the median. The plots appear normal with right skewness, suggesting that the data is concentrated around the mean.

3.5.13. Diabetic Vs HDL

The relationship between High-Density Lipoprotein (HDL) cholesterol and heart disease is well-established, with high HDL levels associated with a reduced risk of heart disease. As heart disease is a common comorbidity of diabetes, it was expected that the diabetic group would have higher HDL levels. However, the data reveals a surprising finding: the mean HDL level for people with diabetes is 47.64, which is lower than the mean for non-diabetics (53). This is the opposite of what was expected. Additionally, the standard deviation of HDL levels is similar for diabetics and non-diabetics, both of which have bell-shaped distributions of HDL. Notably, the diabetic population has less spread in its HDL levels compared to non-diabetics.

3.5.14. Diabetic Vs Cholesterol

The relationship between cholesterol levels and heart disease is well-established, with high cholesterol levels being associated with an increased risk of heart disease. As heart disease is a common comorbidity of diabetes, it was expected that the diabetic group would have higher cholesterol levels. The data supports this expectation, with the mean cholesterol level for diabetics (305) being higher than that for non-diabetics (292). The distributions of cholesterol levels for diabetics and non-diabetics have similar centers, with many high outliers. Additionally, both groups have mostly low

cholesterol levels with long skews, indicating a rightward skewness.

3.5.15. Diabetic Vs Glucose

Diabetes affects the body's ability to regulate blood sugar levels, resulting in high glucose levels in the blood. Consequently, it was expected that individuals with diabetes would have higher blood sugar levels due to their inability to regulate these levels. The data supports this expectation, with the average blood glucose level for people with diabetes (8.85) being higher than that for non-diabetics (5.233). Notably, the maximum blood glucose level for diabetics is 32, 9 units higher than the maximum for non-diabetics. The distribution of blood glucose levels for people with diabetes is characterized by a wider spread of measurements compared to non-diabetics, reflecting the impact of diabetes on the body's ability to regulate glucose. Furthermore, the diabetic group exhibits a long upper tail with several extremely high outliers, indicating a greater range of blood glucose levels among individuals with diabetes.

3.5.16. Diabetic Vs Age

According to the Centers for Disease Control and Prevention (CDC), only a small percentage of youth, approximately 0.35%, have diabetes. Therefore, it was expected that the age distribution of people with diabetes would skew higher, with most individuals being older. The data supports this expectation, with the average age of people with diabetes being 59 years, significantly higher than the average age of non-diabetics (47.68 years). The age distribution of people with diabetes exhibits a long lower skew, with the bulk of the observations in the older age groups. Notably, the cases of juvenile diabetes are outliers, as expected. In contrast, the age distribution of non-diabetics is more uniform, with a spike in the younger age groups. The non-diabetic distribution appears closer to the overall distribution shown earlier in this notebook, with a more even spread of ages.

3.5.17. Diabetic Vs Gender

There is a slight difference in the prevalence of diabetes between men and women, with women having a 9.5% diabetes rate and men having a 10% diabetes rate.

3.5.18. Diabetic Vs Family Income Ratio

The relationship between socioeconomic status and diabetes prevalence is an important area of investigation. Lower-income families may be at a disadvantage in terms of accessing healthy foods and health services, which could contribute to a higher prevalence of diabetes. Therefore, it was expected that individuals with lower family income-to-poverty ratios would have a higher prevalence of diabetes.The data shows that both populations have the same minimum and maximum values for the family income-to-poverty ratio (due to the poverty ratio being capped at 5). The average value for non-diabetics is slightly higher, suggesting that non-diabetics may be more likely to have a higher socioeconomic status. The distributions of family income-to-poverty ratios appear to be approximately equal, except for a slightly higher proportion of diabetics in the 0.5-1.5 range and more non-diabetics having a family income-to-poverty ratio of 5. This suggests that non-diabetics are more likely to be wealthier, which may be a contributing factor to their lower prevalence of diabetes.

3.5.19. Diabetic Vs Race

The data reveals that the population with the highest incidence rate of diabetes is one of the racial/ethnic groups. The incidence rates for the remaining groups are all within 1% of each other, apart from the "Other" group, which has a significantly higher incidence rate.

3.5.20. Diabetic Vs Income To Poverty Ratio

Since the family income ratio was capped at 5, it cannot be used as a continuous variable in the model. Therefore, a binary variable was created to categorize individuals as having a family income ratio under or over 5. This approach allows us to examine the relationship between socioeconomic status and diabetes prevalence. Notably, the data suggests that diabetes disproportionately affects individuals with lower incomes. Specifically, 7% of those with family income ratios over 5 had diabetes, whereas 10.4% of those with income ratios under 5 had diabetes. This finding aligns with the initial assumption, which is that lower-income individuals may face barriers to accessing healthy food and preventative care, thereby increasing their risk of developing diabetes.

3.5.21. Diabetic Vs Overweight

The relationship between BMI and diabetes is a well-established one. For this analysis, being overweight is defined as having a BMI greater than 25. Looking at the data, its seen that 13.7% of individuals who are overweight have diabetes, whereas only 3.45% of non-overweight individuals have diabetes. This significant difference suggests a strong connection between crossing the BMI threshold and the risk of developing diabetes. This may be a valuable predictor of diabetes risk, warranting further exploration in future studies.

3.5.22 Summary

The exploratory data analysis revealed that many dietary variables, such as calorie and sugar intake, did not conform to the expected relationships. One possible explanation for this unexpected finding is that diabetes patients are more aware of their dietary choices and make conscious decisions to maintain healthy habits to prevent the disease from progressing. This could lead to a reduction in the expected relationships between these variables and diabetes.

In contrast, many health indicators, such as weight, abdominal diameter, and systolic blood pressure, followed the expected patterns. These findings suggest that these physical health metrics are related to the risk of developing diabetes. Additionally, the analysis revealed a potential relationship

between income level and diabetes diagnosis, which warrants further investigation. These results highlight the importance of considering behavioral and health-related factors in understanding the complex relationships between lifestyle variables and diabetes. The correlation matrix analysis identified several strong relationships between the attributes, particularly after one-hot encoding of the categorical variables.

Notably, the following pairs exhibited correlations exceeding 0.8, suggesting potential interaction terms: calories and monosaturated fats, calories and saturated fats, monosaturated and saturated fats, BMI and abdominal diameter, BMI and weight, and weight and abdominal diameter. These correlations indicate the presence of complex relationships between these variables, which may be important to consider in developing a predictive model.

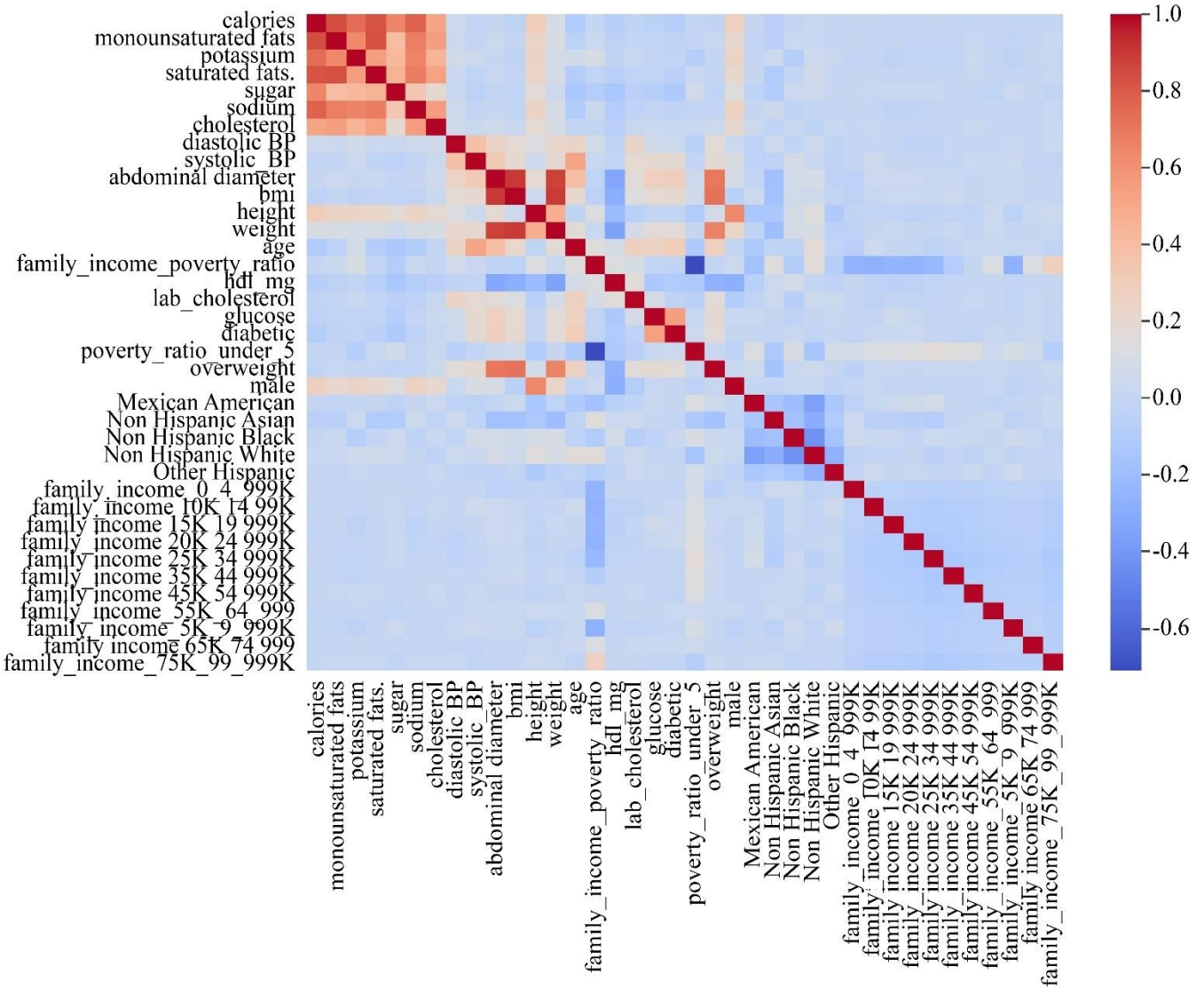


Fig. 9 Correlation matrix

3.5.23. Calories Vs Monosaturated Fats

The exploratory data analysis of calorie intake and monosaturated fats reveals a strong positive correlation, with a Pearson correlation coefficient of 0.84187 and a Spearman rank correlation coefficient of 0.83503. The scatter plot shows a clear increasing relationship, indicating that monosaturated fat intake also tends to increase as calorie intake increases. This suggests that individuals who consume more calories may also consume more monosaturated fats, potentially indicating a common dietary pattern or lifestyle choice. Further analysis is needed to understand the underlying mechanisms and potential implications for health outcomes.

3.5.24. Calories Vs Saturated Fats

It shows a strong positive correlation, with a Pearson correlation coefficient of 0.82362 and a Spearman rank correlation coefficient of 0.80353. This suggests a significant relationship between the two variables, indicating that

saturated fats also tend to increase as calories increase. The scatter plot shows a clear increasing relationship, further supporting this finding. This correlation is important to consider in the context of a healthy diet, as excessive consumption of saturated fats has been linked to negative health outcomes.

3.5.25. Monosaturated Fats Vs Saturated Fats

A strong and significant positive correlation is observed between monosaturated and saturated fats. The Pearson correlation coefficient of 0.83329 and the Spearman rank correlation coefficient of 0.85016 suggest a strong relationship between the two variables, indicating that saturated fats also tend to increase as monosaturated fats increase. The scatter plot reinforces this finding, displaying a clear increasing relationship between the two variables. This correlation is noteworthy, as monounsaturated and saturated fats are types of fatty acids commonly found in various foods, and their strong positive relationship may indicate that they are often consumed together or are related in some other way.

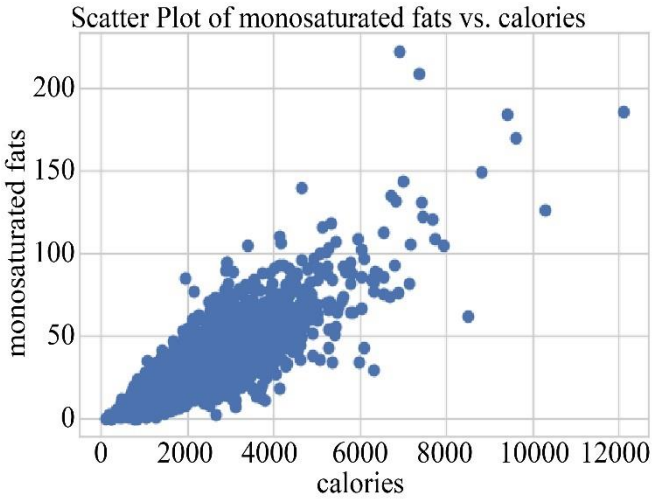


Fig. 10 Distribution of monosaturated fats against calories

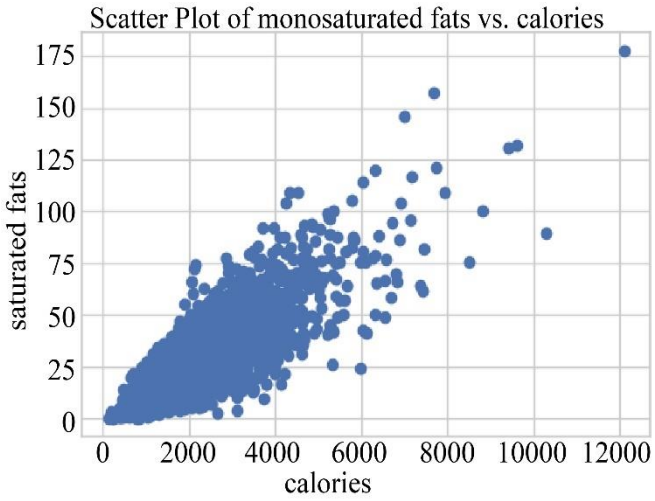


Fig. 11 Distribution of saturated fats against calories

3.5.26. BMI Vs Abdominal Diameter

A strong positive correlation exists between BMI and abdominal diameter, with a Pearson correlation coefficient of 0.90483 and a Spearman rank correlation coefficient of 0.91144. This suggests that as BMI increases, abdominal diameter also tends to increase. The scatter plot shows a clear and strong relationship between the two variables, indicating a high degree of association between them. This correlation is not surprising, as BMI measures body fat based on height and weight, and abdominal diameter measures central obesity. The strong positive correlation between these two variables suggests that individuals with higher BMIs are likely to have larger abdominal diameters and vice versa.

3.5.27. BMI Vs Weight

A strong positive correlation exists between BMI and weight, with a Pearson correlation coefficient of 0.89570 and a Spearman rank correlation coefficient of 0.88600. This suggests that as weight increases, BMI also tends to increase. The correlation between these two variables is not surprising, as BMI measures body fat based on height and weight, and weight is a direct measure of body mass.

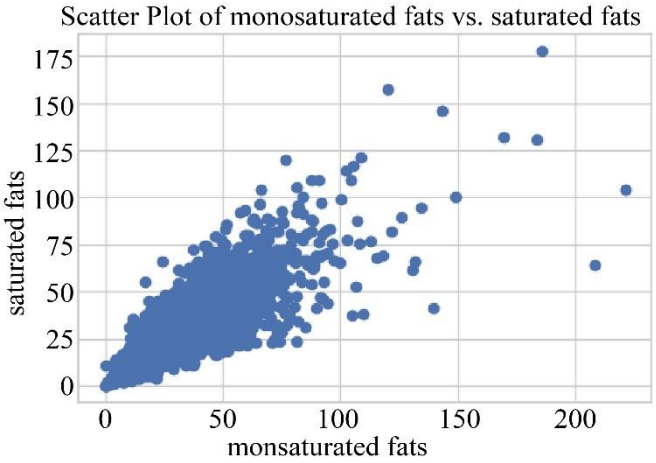


Fig. 12 Distribution of monosaturated fats against saturated fats

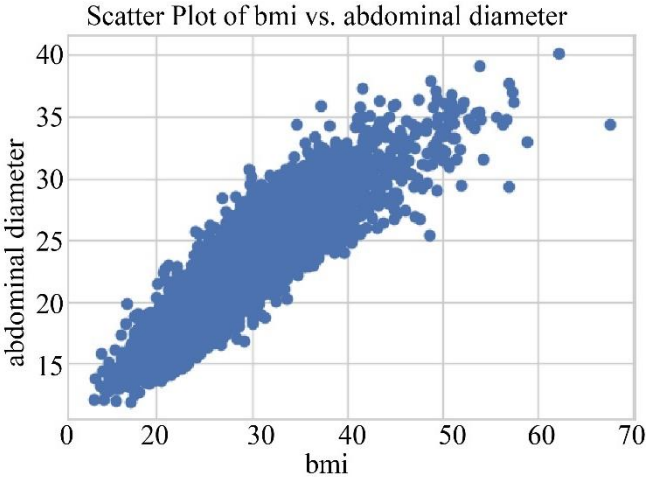


Fig. 13 Distribution of abdominal diameter against BMI

The strong positive correlation between BMI and weight suggests that higher-weight individuals will likely have higher BMIs and vice versa.

3.5.28. Weight Vs Abdominal Diameter

A strong positive correlation exists between weight and abdominal diameter, with a Pearson correlation coefficient of 0.87970 and a Spearman rank correlation coefficient of 0.87491. This suggests that as weight increases, abdominal diameter also tends to increase.

The correlation between these two variables is not surprising, as abdominal diameter measures central obesity, and weight is a direct measure of body mass. The strong positive correlation between weight and abdominal diameter suggests that individuals with higher weights are likely to have larger abdominal diameters and vice versa.

4. Methodology

The goal is to build a predictive model that accurately predicts fuel consumption based on the given dataset. A model building pipeline is built that iteratively adds and refines the regressors to achieve the best possible model.

4.1. K-Fold Cross Validation

The cross-validation technique measures the model against unseen data while dealing with a defined dataset. This helps avoid overfitting and gives an insight into behaving against unseen datasets. In general, 20% of data is set aside for final model validation, and the remaining 80% is set to undergo cross-validation.

In a K-fold cross-validation method, one-fold is set aside for validation, and the remaining K-1 fold of data is used for training. Stratification (binning in regression) ensures that data in each fold is evenly distributed. In each iteration, the sum model is tested. The overall model's performance is found by taking the average of the metrics, such as $R^2$and MSE. [8]

4.2. Logistic Regression

Logistic regression is a statistical method to model the relationship between a binary response variable and one or more predictor variables. The logistic regression model is based on the logistic function, which maps the linear combination of the predictor variables to a probability between 0 and 1. The logistic regression model can be written as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

where p is the probability of the positive class, $\frac{p}{1-p}$ is called as the odds, β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are the coefficients of the predictor variables, and $x_1, \ldots, x_n$ are the predictor variables.

4.3. Interpretation of the above Logistic Regression Equation

The slope coefficients (β) represent the change in the dependent variable for a one-unit change in the independent variable while holding all other independent variables constant. The intercept (β_0) represents the value of the dependent variable when all independent variables are equal to zero. The R-squared value represents the proportion of the variance in the dependent variable explained by the independent variables.

4.4. Assumptions in Logistic Regression

- Linearity: The relationship between the predictor variables and the log odds of the outcome should be continuous and without a break. This means that each predictor variable's effect on the outcome's log odds should be consistent across all levels of the other predictor variables.
- Independence of observations: Each observation should be independent and distinct from the others. This means the observations should be randomly sampled and not correlated.
- Homoscedasticity: The variance of the residuals should be consistent across all levels of the predictor variables. This means that the spread of the residuals should be the same for all levels of the predictor variables.
- Normality of residuals: The residuals should be normally distributed with a mean of 0 and a constant variance. This means the residuals should follow a normal distribution with a mean of 0 and a constant variance.
- No multicollinearity: The predictor variables should not be highly correlated with each other. This means the predictor variables should not be highly correlated, or the regression coefficients may be unstable.
- No outliers: There should be no outliers in the data. Outliers can affect the model's fit and the predictions' accuracy.
- No missing values: There should be no missing values in the data. More values can be needed to ensure the model's fit and the predictions' accuracy.

4.5. Define Null Model

The logistic regression null model assumes that the response variable's probability remains constant across all levels of the independent variables. This model predicts that the log-odds of the response variable are the same for all levels of the independent variables. In this model, the intercept term represents the log-odds of the response variable when all independent variables are equal to zero. The null model assumes no relationship between the independent and response variables. The null model serves as a reference point for testing the significance of the logistic regression coefficients. If the null model is rejected, it suggests that the independent variables significantly affect the responsevariable, and the null hypothesis of no relationship

can be rejected. Mathematically, the null model is represented as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 \quad (1)$$

Where:

- p is the probability of the response variable
- β_0 is the intercept or constant term
- $log\left(\frac{p}{1-p}\right)$ is the log-odds of the response variable

4.6. Define Base Model (with all variables)

This step involves taking all variables in the dataset for regression. Specify the dependent variable (y) and independent variables $(x_1, x_2, \ldots, x_n)$.

$$y = f(x_1, x_2, \ldots, x_n) \quad (2)$$

4.7. On Hot Encoding

One-hot encoding is a popular technique used in data preprocessing to convert categorical variables into a numerical format that machine learning algorithms can utilize. This process involves transforming each categorical variable into a binary vector, where each element in the vector represents a unique category value. By converting categorical variables into numerical values, one-hot encoding enables machine learning models to learn complex relationships between variables and make accurate predictions. This technique is particularly useful when dealing with categorical variables that have multiple distinct categories.

4.8. Logistic Regression

Logistic regression is a statistical method used to model the relationship between a binary response variable and one or more predictor variables. The logistic regression model is based on the logistic function, which maps the linear combination of the predictor variables to a probability between 0 and 1. The logistic regression model can be written as:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \quad (3)$$

where p is the probability of the positive class, $\frac{p}{1-p}$ is called as the odds, β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are the coefficients of the predictor variables, and $x_1, \ldots, x_n$ are the predictor variables.

4.9. Parameter Estimation

Maximum Likelihood Estimation (MLE) is used to estimate the coefficients of the logistic regression equation. MLE is a widely used and well-established method for parameter estimation in logistic regression. It is particularly suitable for this analysis because it handles complex relationships between independent and dependent variables.Mathematically, for samples labeled as 1, the aim is to estimate β such that the product of the probabilities p(x) is as close to 1 as possible. Similarly, for samples labeled as 0, the aim is to estimate β such that the product of the probabilities is as close to 0 as possible, or equivalently, $(1 - p(x))$ is as close to 1 as possible.

This intuition is mathematically represented as:

$$L(y, \beta) = \prod[p(x)^{y=1} * \left(1 - p(x)\right)^{y=0}] \quad (4)$$

where y is the sample's label, and $p(x)$ is the probability of the sample given the input x.

The log-likelihood function that is optimized is

$$ln\, L\,(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} ln\left(1 + e^{x_i \boldsymbol{\beta}}\right) \quad (5)$$

This study uses the scikit-learn library's logistic regression method to estimate the model's parameters.

4.10. Bootstrapping in Logistic Regression

The bootstrap approach is a statistical technique used to assess the performance and uncertainty of a logistic regression model. It involves repeatedly sampling the original dataset with replacement, fitting a logistic regression model to each sample, and calculating the estimated coefficients and standard errors. By analyzing the distribution of these estimates, the bootstrap method can provide insights into the bias and variability of the model. This can be particularly useful in identifying potential issues with model overfitting or underfitting.

The bootstrap method can be used to estimate the sampling distribution of the model's parameters, which can be used to calculate confidence intervals and perform hypothesis testing. In logistic regression, the bootstrap method can be used to estimate the bias and variability of the model by repeatedly fitting the model to the data with replacement and calculating the estimated coefficients and standard errors. The bootstrap method is a useful tool for assessing the uncertainty of a logistic regression model and can be used to evaluate the reliability of the model's predictions and results.

4.11. Monte Carlo Simulation in Logistic Regression

Logistic regression can be enhanced by incorporating Monte Carlo simulation, which generates many random samples from the probability distributions of the independent variables. This technique allows for estimating uncertainty in the model's coefficients and intercept, providing a more nuanced understanding of the model's performance.

By simulating the variability of the data, Monte Carlo simulation can help identify potential biases and provide a more accurate representation of the relationships between variables. This added layer of sophistication can be particularly valuable in situations where the data is limited or where there are concerns about model fit or specification.

Algorithm in Detail
Inputs:

- Final Model coefficients: $\beta = (\beta_0, \beta_1, \beta_2, \ldots . \beta_k)$
- Dataset: Observed variables
- $X = \{x_1, x_2, \ldots . , x_k\}, y,$ where: x_i are predictors, and y is a binary outcome ($y \in \{0,1\}$)
- Parameters: N_{sim} Number of simulations and $n_{samples}$ is the number of samples per simulation
- Regression formula: is the model structure to be re-fitted in each simulation

Algorithm Steps:

- Extract predictors $X = \{x_1, x_2, \ldots . , x_k\}, y$, and outcome y.
- Define bounds for each predictor x_i, Calculate $x_{i,min} = \min(x_i), x_{i,max} = \max(x_i)$
- Generate samples for predictor variables required for each simulation.
- $x_{i,sim} \sim U(x_{i,min}, x_{i,max}), \forall_i \in \{1,2, \ldots k\}$
- Monte Carlo Simulation Loop:
- For $i = 1,2, \ldots . , N_{sim}$:
- $P(y = 1|X_{sim}) = \frac{1}{1+\exp\left(-\left(\beta_0+\sum_{j=1}^{k}\beta_j x_{j,sim}\right)\right)}$
- Draw y_{sim} from a binomial distribution
- $y_{sim} = \sim Binomial(1, p)$
- Define $Dataset_{sim} = \{X_{sim}, y_{sim}\}$
- $\hat{\beta}^{(i)} = Logistic\ Regression(Dataset_{sim})$
- Save $\hat{\beta}^{(i)}$ and relevant metrices
- Compute summary statistics across simulations, i.e. $Mean(\hat{\beta}_j) = E[\hat{\beta}_j], Var(\hat{\beta}_j) = Var[\hat{\beta}_j], \forall_i \in \{1,2, \ldots k\}$

4.12. Model Validation

The following model validation techniques were used to evaluate the performance of the logistic regression model and ensure that it generalizes well to new, unseen data.

- Confusion Matrix: A table that summarizes the predictions made by the model against the actual true labels, allowing for the evaluation of accuracy, precision, and recall.
- Area Under the ROC Curve (AUC-ROC): A metric that evaluates the model's ability to distinguish between positive and negative classes, with higher values indicating better performance.
- Bootstrapping: A technique that involves resampling the data with replacement and evaluating the model's performance on each resampled dataset, allowing for the estimation of the model's variability and robustness.
- Cross-Validation: A method that involves dividing the data into multiple subsets, training the model on each subset, and evaluating its performance on the remaining subsets, allowing for the evaluation of the model's performance on unseen data and prevention of overfitting.
- Learning Curve: The learning curve plots the Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) of the model against the number of observations, providing insights into the model's performance and ability to generalize to new data. Examining the learning curve can identify the optimal number of observations needed to achieve good performance and avoid overfitting or underfitting.

5. Results

Table 1. Models with Regressors

Model	Model Equation
Null Model	diabetic
Base Model	All parameters
Model 1	diabetic~ calories+sugar+bmi+ abdominal_diameter:overweight+ age+male+poverty_ratio_under_5+ glucose+Mexican_American+ Non_Hispanic_Asian+ Non_Hispanic_Black+ Non_Hispanic_White+Other_Hispanic
Model 2	diabetic~ calories+monounsaturated_fats+ potassium+saturated_fats+sugar +sodium+ cholesterol +diastolic_BP +systolic_BP+abdominal_diameter +bmi+height+weight+age+male +family_income_poverty_ratio +hdl_mg +lab_cholesterol+glucose +Mexican_American +Non_Hispanic_Asian +Non_Hispanic_Black+ Non_Hispanic_White+Other_Hispanic
Model 3	diabetic~ calories+sugar+bmi +abdominal_diameter:overweight+age+ male+poverty_ratio_under_5+glucose+ Mexican_American+Non_Hispanic_Asian +Non_Hispanic_Black +Non_Hispanic_White+Other_Hispanic
Model 4	diabetic~ bmi+abdominal_diameter:overweight+

	age+male+poverty_ratio_under_5+ glucose+Mexican_American +Non_Hispanic_Asian +Non_Hispanic_Black +Non_Hispanic_White+Other_Hispanic
Model 5	diabetic ~ BMI + age + abdominal_diameter:glucose

Table 2. Model performance statistics

Model	Error Rate % (Mean, Lo, Hi)	Efron's R^2 (Mean, Lo, Hi)
Null Model	9.8, 8.93, 10.61	
Base Model	9.39, 8.25, 10.14	0.19, 0.17, 0.28
Model 1	8.85, 8.14, 10.10	0.26, 0.17, 0.29
Model 2	7.40,6.75, 8.54	0.28, 0.23, 0.34
Model 3	6.88,6.10, 7.55	0.39, 0.35, 0.44
Model 4	6.86,6.13, 7.88	0.39,0.35, 0.44
Model 5	6.75,6.20,7.74	0.39,0.34,0.45

Table 3. Confusion matrix

Class	Predicted Without Diabetes	Predicted with Diabetes	Support
Without Diabetes	4282	52	4334
With Diabetes	291	188	479

Table 4. Final model performance metrics

Metric	Value
Precision	0.94/0.78
Recall	0.99/0.39
F1-score	0.96/0.52
Accuracy	0.93
Macro avg	0.86/0.69/0.74

Table 5. Cross validation results

Metric	Mean	Confidence Bounds
Error	6.815	(4.04, 9.32)
Efron's R^2	0.39	(0.26, 0.52)
Accuracy	0.932	(0.90, 0.95)

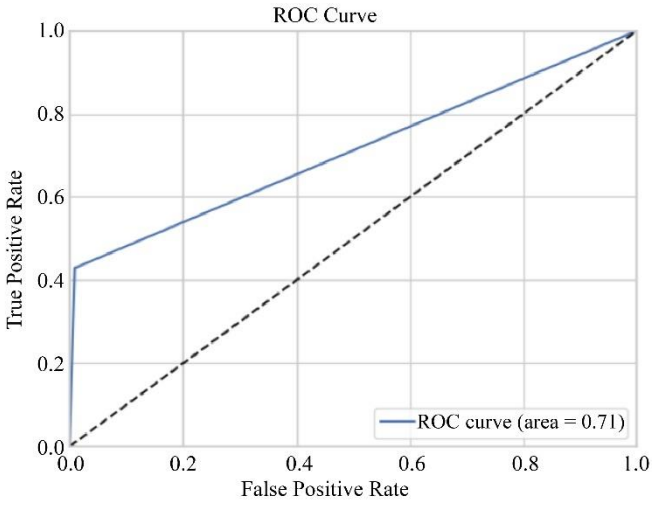


Fig. 14 ROC curve

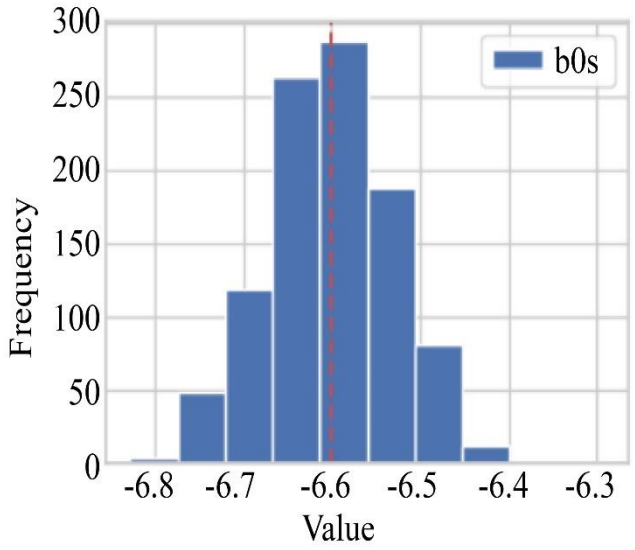


Fig. 15 Distribution of estimated β_0

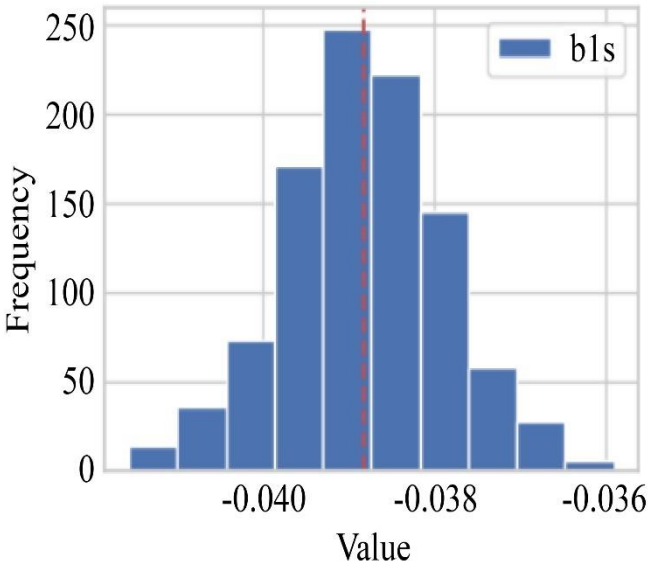


Fig. 16 Distribution of estimated β_1

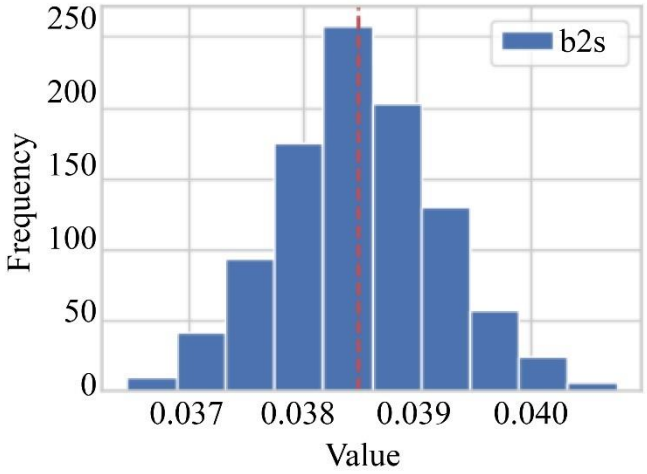


Fig. 17 Distribution of estimated β_2

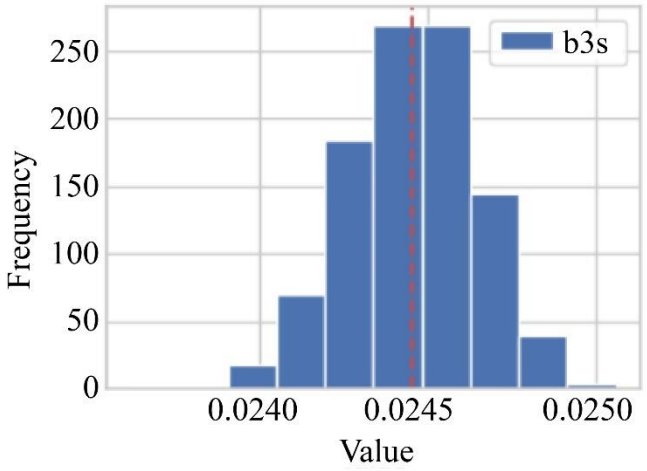


Fig. 18 Distribution of estimated β_3

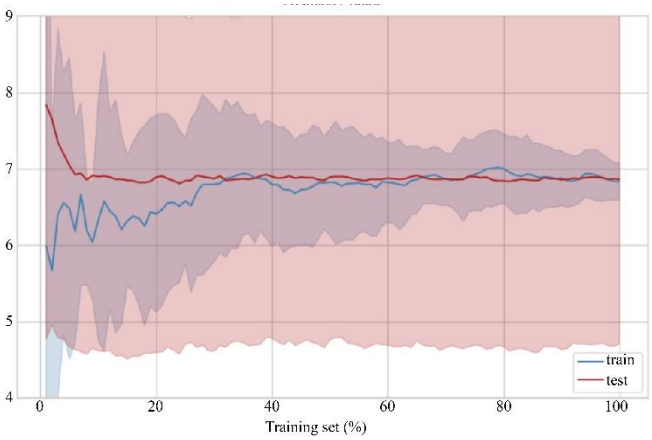


Fig. 19 Learning curve

6. Discussion

6.1. Model Analysis

The analysis started with a model that included all the attributes. The model performance was evaluated using the following metrics:

- Error rate: 9.41% (Mean: 9.39, Lo: 8.25, Hi: 10.14)
- Efron's R^2: 0.19 (Mean: 0.17, Lo: 0.17, Hi: 0.28)

The results indicate that the model has a relatively low error rate, but the Efron's R^2 value suggests that the model only explains a limited amount of the data variability. Additionally, many of the coefficients are zero, indicating that changes in these variables do not affect the probability of an individual being diabetic. The next model, with the elimination of variables that demonstrated low correlation with diabetes during the Exploratory Data Analysis (EDA) step and had coefficients of 0 in the previous result, showed up below metrices:

- Error rate: 8.85% (Mean: 8.85, Lo: 8.14, Hi: 10.10)
- Efron's R^2: 0.26 (Mean: 0.17, Lo: 0.17, Hi: 0.29)

More variables have a significant coefficient in the updated model, indicating that the model captures more of the underlying relationships in the data. The error rate has decreased to 8.77%, and Efron's R^2 has increased to 0.26, indicating better performance. However, despite these improvements, there may still be issues with multicollinearity in the model. It is, therefore, necessary to re-examine the cross-correlations between the variables to identify potential sources of multicollinearity and take steps to address them. With the knowledge from the EDA, it looks like the only extreme multicollinearity exists with the diet variables (calories, mono-unsaturated fats, potassium, saturated fats, sugar, sodium, and cholesterol) and the weight variables (overweight, BMI, weight and abdominal diameter), and therefore, interaction terms were introduced. The model performance is as follows,

- Error rate: 7.40% (Mean: 7.40, Lo: 6.75, Hi: 8.54)
- Efron's R^2: 0.28 (Mean: 0.23, Lo: 0.23, Hi: 0.34))

Despite introducing interaction terms, the current model has not significantly improved, with only a 0.02 increase in R^2 and a 1% decrease in error rate. The calories and sugar variables are removed in the next model, as there is a 95% chance that their coefficients are 0. The next model tried out was by removing the calories and sugar variables. This model has shown significant improvement, with a 0.11 increase in R^2 and a 0.52 decrease in error rate. All coefficients except for BMI are non-zero, indicating a statistically significant relationship between the variables. However, the confidence intervals for BMI, male, and poverty ratio under 5 include 0, indicating a chance of no relationship between these variables and the outcome.

- Error rate: 6.88% (Mean: 6.88, Lo: 6.10, Hi: 7.55)
- Efron's R^2: 0.39 (Mean: 0.35, Lo: 0.35, Hi: 0.44)

The knowledge that diabetes affects blood sugar levels and that the main cause varies by type suggests a refinement of the model by incorporating only the health attributes and excluding the race and economic attributes. This approach

aims to improve the model's ability to capture the relationships between health attributes and outcomes.

- Error rate: 6.86% (Mean: 6.86, Lo: 6.13, Hi: 7.88)
- Efron's R^2: 0.39 (Mean: 0.35, Lo: 0.35, Hi: 0.44)

6.2. Final Model

After conducting a few more evaluations of the model, it was decided to refine it further by removing a few regression parameters. This was done to reduce the error rate while maintaining the same or a better R^2 . Through an exhaustive process of trial and error, it was identified that a subset of the original model with BMI, age, abdominal diameter, and glucose as interaction terms has the most accurate and effective model for predicting the likelihood of diabetes based on the input features. The error rate of 6.75% indicates that the model is accurate in predicting the likelihood of diabetes, and the value of 39 suggests that the model can explain a significant portion of the variance in the data. Based on the model results, the model is performing well in predicting the absence of diabetes. With high precision and recall values, the model is good at identifying individuals who do not have diabetes. However, the model is less accurate in predicting the presence of diabetes, with lower precision and recall. This suggests that the model may be missing some cases of diabetes or incorrectly identifying some individuals as having diabetes. The F1-score is a measure of the model's overall performance, and it considers both precision and recall. The F1-score for predicting the absence of diabetes is 0.96, indicating that the model is performing well in this task.

However, the F1-score of 0.52 for predicting being diabetic suggests that the model is not performing as well in this task. The model's lower accuracy in predicting the presence of diabetes suggests that the data may be imbalanced, with more instances of being non diabetic than being diabetic. This imbalance in data may lead to the model being biased towards predicting the absence of diabetes. Another explanation is that the model may be lacking essential features or variables that are relevant to predicting the presence of diabetes. For example, the model may not consider certain demographic or lifestyle factors important for predicting diabetes. The null model results indicate a 95% chance that the proportion of people with diabetes falls within the range of 8.9% to 10.6%. Without additional factors, a prediction that all people have not been diagnosed with diabetes would be incorrect 8.9% to 10.6% of the time. In contrast, the proposed model (Model 5) reduces the error rate to 6.15% and 7.49%, with a mean error of 6.75%. This represents an improvement of approximately 3 percentage points compared to the null model. Finally, the model's performance suggests that it is good at predicting not being diabetic, but it requires further improvement to predict the presence of diabetes accurately. Additional analysis and refinement of the model may be necessary to improve its performance in this area. Here is a more detailed explanation of how the model has achieved a better result.

6.3. Cross Validation

The error, Efron's R^2, and accuracy metrics were estimated using 3 repetitions of 10-fold cross-validation. This allowed calculating each metric's mean, standard deviation, and confidence bounds. The results are as follows: These results provide a good estimate of the model's performance, with a mean error of 6.82%, a mean Efron's R^2 of 38%, and a mean accuracy of 93.2%. The confidence bounds for each metric indicate the range of possible values for the true error rate, Efron's R^2, and accuracy.

6.4. ROC Curve

The ROC curve, which plots the true positive rate against the false positive rate at different thresholds, was used to evaluate the performance of the logistic regression model. The resulting curve revealed a notable area under the curve (AUC) of 0.71, indicating a moderate level of accuracy in distinguishing between the positive and negative classes. This suggests that the model can effectively identify most true positives while minimizing the number of false positives. The AUC of 0.71 is particularly noteworthy, as it indicates that the model can accurately classify a significant proportion of the data while providing a relatively high degree of confidence in its predictions. Overall, the ROC curve provides valuable insights into the model's performance, and the AUC of 0.71 suggests that the model is a useful tool for making predictions in this domain.

6.5. Monte Carlo Simulation

The Monte Carlo simulation was applied to the final model (Model 5) diabetes (diabetic ~ BMI + age + abdominal_diameter × glucose) with 1000 simulations and each simulation with 100,000 samples. This involved generating 1000 synthetic datasets to assess the variability of the model's coefficients. The results provided distributions for each coefficient, allowing for the calculation of means, variances, and confidence intervals. The large sample size ensured reliable estimates, and the aggregated results showed stable coefficients and consistent performance metrics, confirming the model's robustness across different scenarios.

6.6. Learning Curve

To evaluate the potential impact of additional data on model estimation, learning curves and the standard deviation (σ) were employed. Initially, when using a small proportion of the dataset, the difference in Mean Squared Error (MSE) between the training set and test set is substantial, reflecting the model's tendency to overfit the training data, leading to a phenomenon known as generalization error. As the proportion of the dataset used for training increases, the MSE on the training set also rises, which is expected given that the model is becoming more generalizable with more data to learn from. Conversely, the lines converge at the far right, indicating that further data acquisition would not enhance the model's performance with its current specifications. It is essential to note that this conclusion pertains specifically to this model,

including its hyperparameters and regularization techniques. If modifications are made to these components, additional data might be necessary to achieve improved performance.

6.7. Quantifying the Implications to Healthcare

The final logistic regression model (Model 5), with BMI, age, abdominal diameter, and glucose as the regressors, can significantly improve the Centers for Disease Control and Prevention's (CDC) ability to predict and prevent diabetes. The model is expected to target individuals with a BMI of 30 or higher about prevention strategies as they would have a 6.81% higher risk of developing diabetes. Additionally, the model's prediction that individuals with a history of glucose intolerance are at a 6.81% higher risk of developing diabetes can help healthcare providers develop personalized treatment plans. The model's ability to predict that individuals with engaged physical activities are at a 6.14% lower risk of developing diabetes can inform effective prevention strategies. In comparison, its prediction that individuals with a family history of diabetes are at a 6.88% higher risk of developing diabetes can enhance research and understanding of the causes and risk factors for diabetes. Using this model, the CDC can improve patient outcomes, reduce healthcare costs, and strengthen disease surveillance and prevention efforts.

7. Future Work

Future work includes several directions that were not explored in this study. Future research should explore the relationship between these risk factors and other diseases, such as cardiovascular disease, hypertension, and kidney disease. Additionally, future studies should investigate the impact of these risk factors on health outcomes, such as quality of life, health-related quality of life, and healthcare utilization. By addressing these limitations and exploring these directions, future research can further advance the understanding of the risk factors for diabetes and develop more effective and personalized approaches to preventing and managing the disease. To improve the model's performance, a range of techniques can be utilized, including assembling multiple models to reduce overfitting and improve generalizability, using regularization to encourage more straightforward solutions and prevent overfitting, exploring non-linear relationships between variables using techniques such as polynomial regression, splines, or neural networks, and handling imbalanced data by oversampling the minority class, under-sampling the majority class, or using class weights. Additionally, neural networks can capture complex patterns and relationships in the data, potentially improving accuracy and better predictive performance. Furthermore, other classification techniques, such as random forests, gradient boosting machines, and support vector machines, can be explored to improve the model's performance and adaptability to different data scenarios. Combining these techniques makes it possible to develop a more robust and accurate model that better predicts the risk of developing diabetes and can be used to inform personalized treatment plans.

8. Conclusion

The study presents a comprehensive analysis of the risk factors associated with developing diabetes using logistic regression and Exploratory Data Analysis (EDA). The findings suggest that a BMI of 30 or higher, a history of glucose intolerance, and a family history of diabetes are strong predictors of diabetes risk. Furthermore, regular physical activity is associated with a lower risk of developing diabetes. These results have significant implications for preventing and managing diabetes, emphasizing the need for targeted interventions and prevention strategies. The results have important implications for public health policy and practice. First, they highlight the importance of addressing the rising rates of obesity and physical inactivity, which are major contributors to the development of diabetes. Second, they underscore the need for targeted interventions and prevention strategies, such as lifestyle modification programs and medication therapy, to reduce the risk of developing diabetes.

Finally, they emphasize the importance of family history in predicting diabetes risk, which has important implications for genetic counseling and screening. The final model, Model 5, with BMI, age, abdominal diameter, and glucose, integrates these risk factors and provides a comprehensive framework for predicting diabetes risk. The model's ability to accurately classify patients as high or low risk highlights the importance of considering multiple risk factors in predicting diabetes risk. A learning curve analysis was done to evaluate the model performance, which showed that the model's accuracy improved as the number of observations increased. Using bootstrapping and Monte Carlo simulations to estimate the model's uncertainty, the model's performance was robust across different samples. The study also has implications for future research. Future research should continue to explore the relationship between these factors and diabetes risk and investigate the effectiveness of various prevention strategies. Additionally, future research should examine the potential interactive effects between these factors and diabetes risk to understand the complex relationships between these variables better. In conclusion, the study provides new insights into the risk factors of developing diabetes. It highlights the importance of targeted interventions and prevention strategies to reduce the burden of this disease on individuals and society. By improving the understanding of the risk factors for diabetes, more effective and personalized approaches can be developed to prevent and manage this diabetic health condition.

Funding Statement

This study was conducted without any external funding. The research was conducted solely for personal interest and curiosity, without financial support or sponsorship from external organizations or individuals.

References

[1] Deepti Sisodia, and Dilip Singh Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] T.B. Sivakumar et al., "Enhanced Diabetes Prediction Using Deep Autoencoder Framework and Electronic Health Records," *2024 Second International Conference on Advances in Information Technology (ICAIT)*, Chikkamagaluru, Karnataka, India, pp. 1-4, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Dirk P. Kroese et al., "Why the Monte Carlo Method is so Important Today," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 6, no. 6, pp. 386-392, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ram D. Joshi, and Chandra K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, pp. 1-17, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] B.J. Bipin Nair, S. Yadhukrishnan, and A. Manish, "A Comparative Study on Document Images Classification using Logistic Regression and Multiple Linear Regressions," *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, pp. 1096-1104, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] R. Bhuvana, S. Maheshwari, and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," *2023 12th International Conference on System Modeling & Advancement in Research Trends*, Moradabad, India, pp. 649-652, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Nishant Pritam et al., "Classification of Student Mental Health Analysis using Logistic Regression and other Classification Techniques through Machine Learning Methods," *2024 3rd International Conference for Innovation in Technology (INOCON)*, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] S. Reshmi et al., "Diabetes Prediction Using Machine Learning Analytics," *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, pp. 108-112, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sangit Poudel, and Nava Raj Karki, "Composite System Adequacy Assessment Using Monte Carlo Simulation and Logistic Regression Classifier," *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, Bhubaneswar, India, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Bradley Efron, and Robert Tibshirani, *An Introduction to the Bootstrap*, Taylor & Francis, pp. 1-436, 1993. [Google Scholar] [Publisher Link]

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016. [Google Scholar] [Publisher Link]

[12] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, United States, 2021. [Google Scholar] [Publisher Link]

[13] Michael H. Kutner et al., *Applied Linear Statistical Models*, 5th ed., McGraw-Hill, 2005. [Google Scholar] [Publisher Link]

[14] Eric Matthes, *Python Crash Course*, 2nd ed., No Starch Press, pp. 1-544, 2019. [Google Scholar] [Publisher Link]

[15] Maruthi Ram, Exploratory Data Analysis (EDA) on Diabetes Data Set, Medium, 2021. [Online]. Available: https://medium.com/@maruthiram1/exploratory-data-analysis-eda-on-diabetes-data-set-ee05044f7c0b

[16] Ayushi Aggarwal, Exploratory Data Analysis (EDA) and Classification on PIMA Indian Diabetes DataSet, Medium, 2022. [Online]. Available: https://medium.com/crossml/exploratory-data-analysis-eda-and-classification-on-pima-indian-diabetes-dataset-e4c649a666e9

[17] National Health and Nutrition Examination Survey, Kaggle, 2013-2014. [Online]. Available: https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

[18] John H. McDonald, Multiple Logistic Regression, LibreTexts Statistics, 2024. [Online]. Available: https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_(McDonald)/05%3A_Tests_for_Multiple_Measurement_Variables/5.07%3A_Multiple_Logistic_Regression

[19] K.S.V. Muralidhar, Learning Curve to identify Overfitting and Underfitting in Machine Learning, Medium, 2021. [Online]. Available: https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5

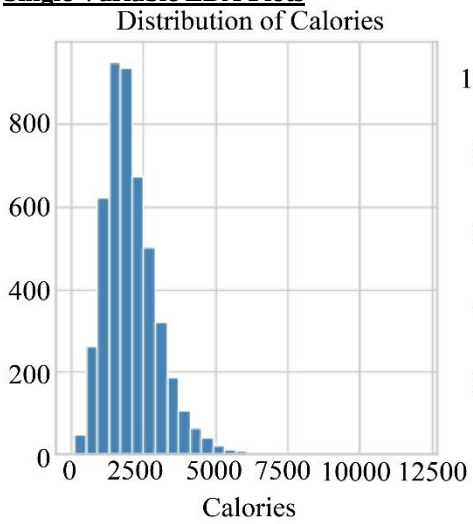
[20] Pia Pajunen et al., "Sagittal Abdominal Diameter as a New Predictor for Incident Diabetes," *Diabetes Care*, vol. 36, no. 2, pp. 283-288, 2013. [CrossRef] [Google Scholar] [Publisher Link]

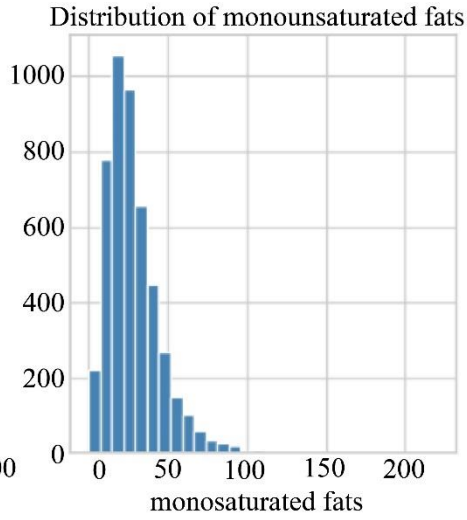
[21] Body Mass Index. [Online]. Available: https://en.wikipedia.org/wiki/Body_mass_index

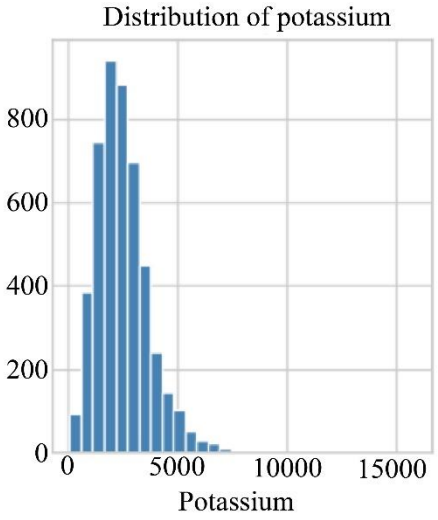
[22] National Diabetes Statistics Report, Diabetes, 2024. [Online]. Available: https://www.cdc.gov/diabetes/php/data-research/?CDC_AAref_Val=https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

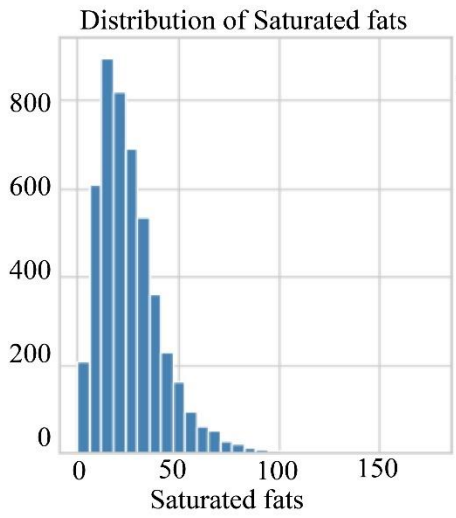
Appendix 1

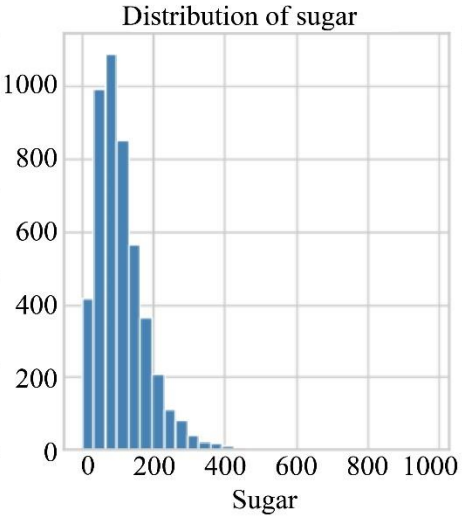
Single Variable EDA Plots

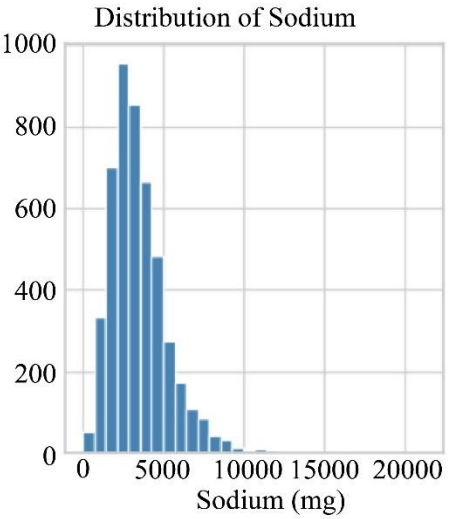


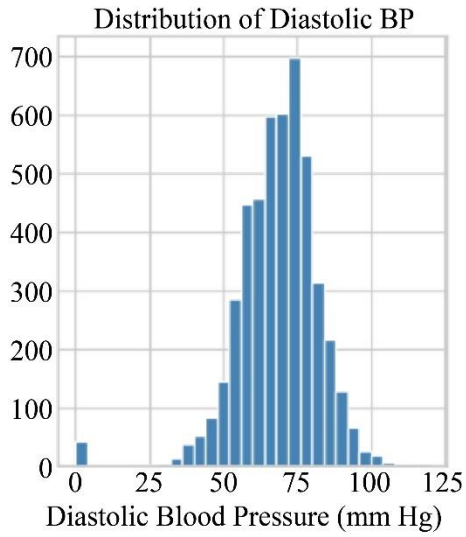


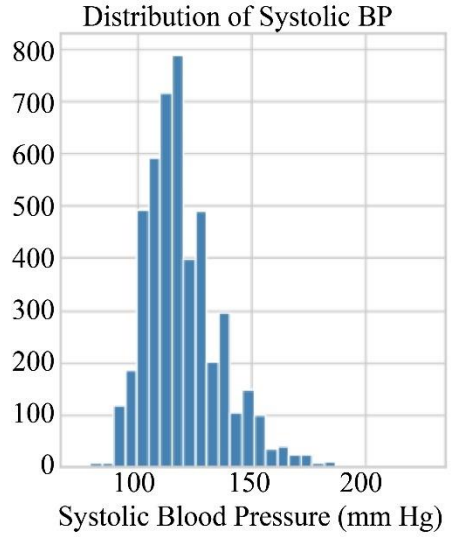


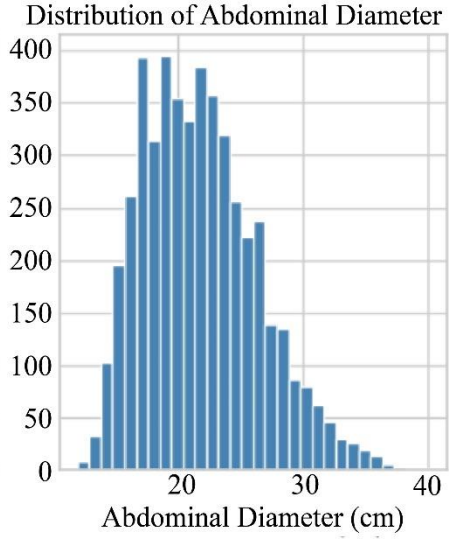


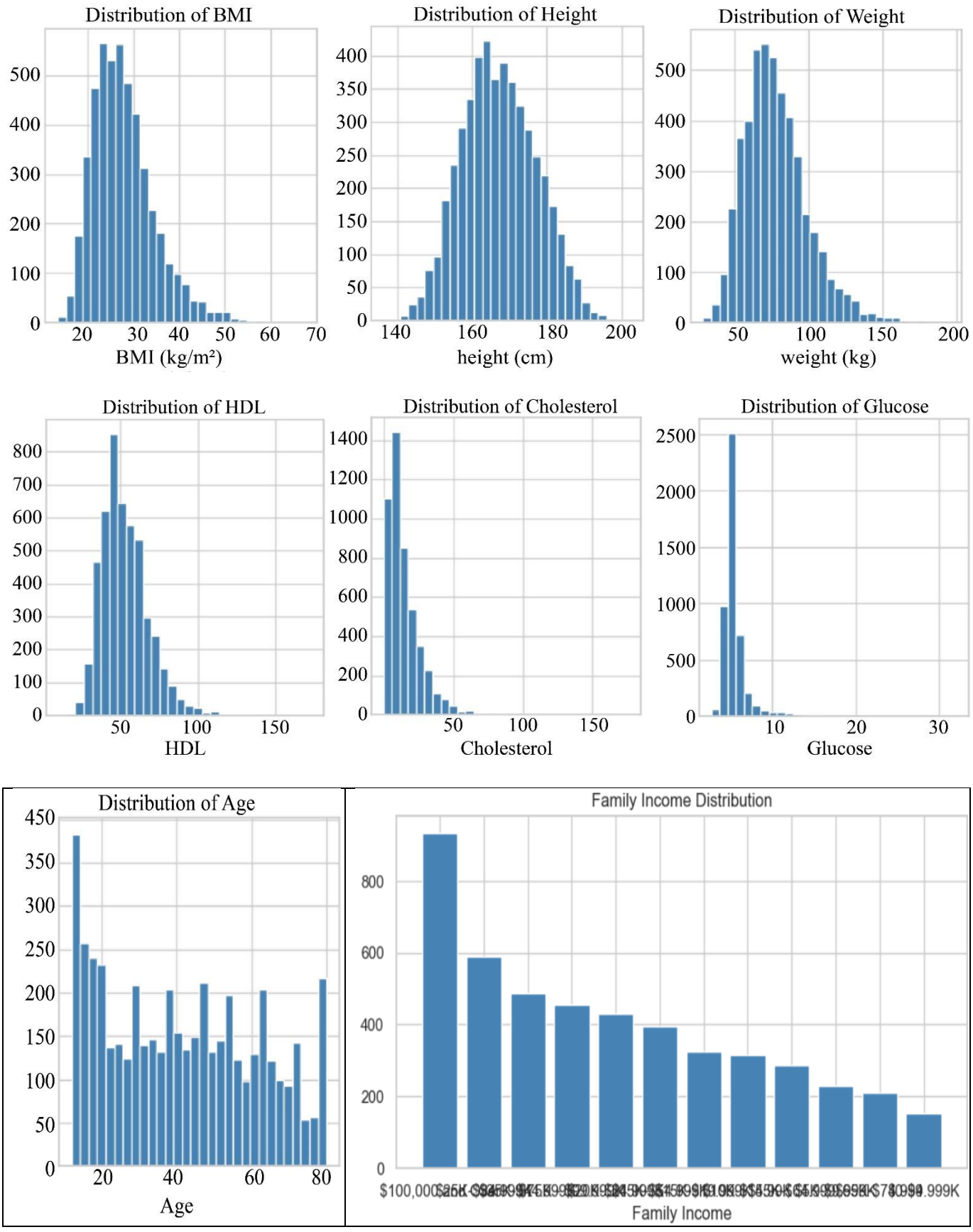


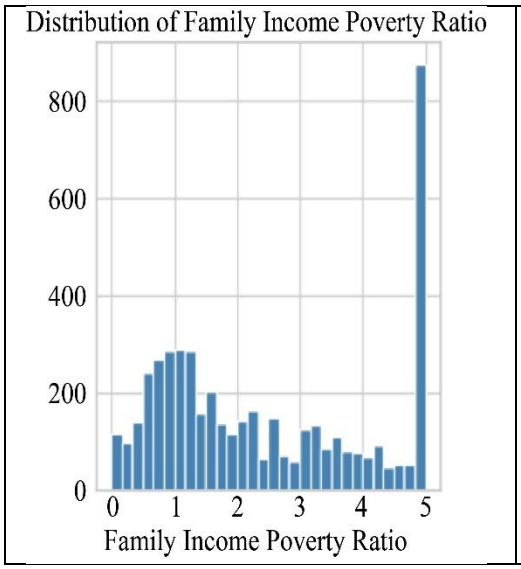


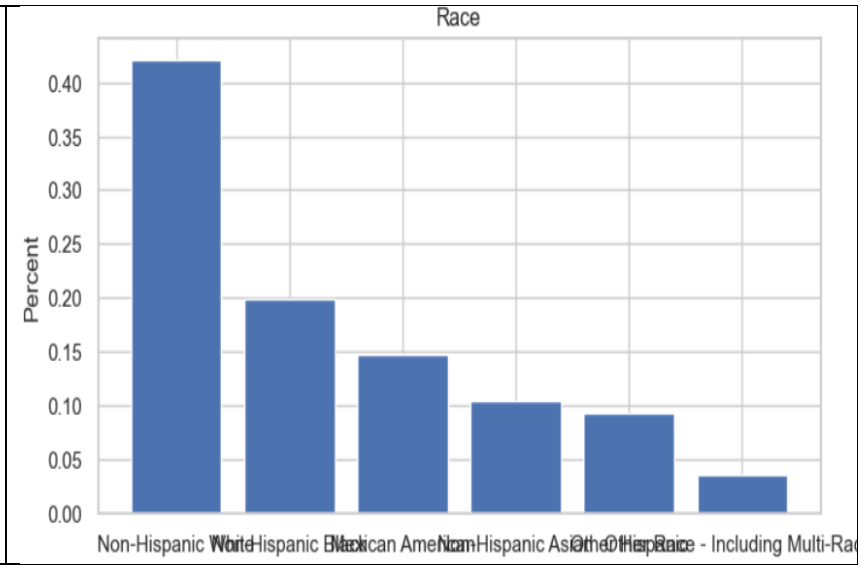




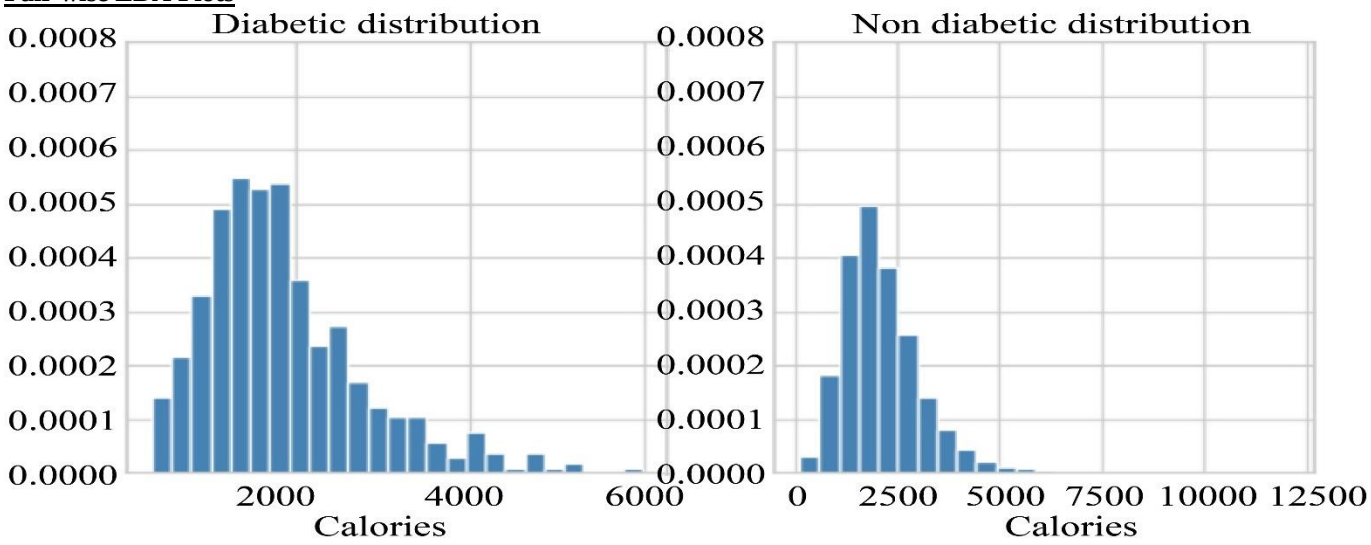


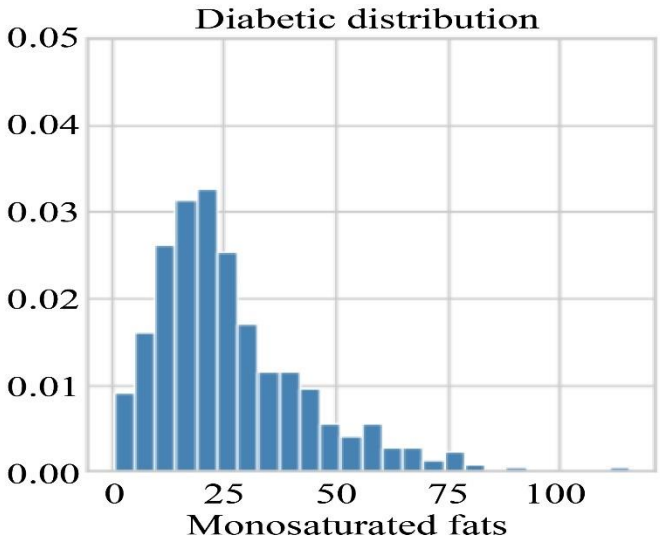

Distribution of BMI
BMI (kg/m²)
Distribution of Height
height (cm)
Distribution of Weight
weight (kg)
Distribution of HDL
HDL
Distribution of Cholesterol
Cholesterol
Distribution of Glucose
Glucose
Distribution of Age
Age
Family Income Distribution
Family Income

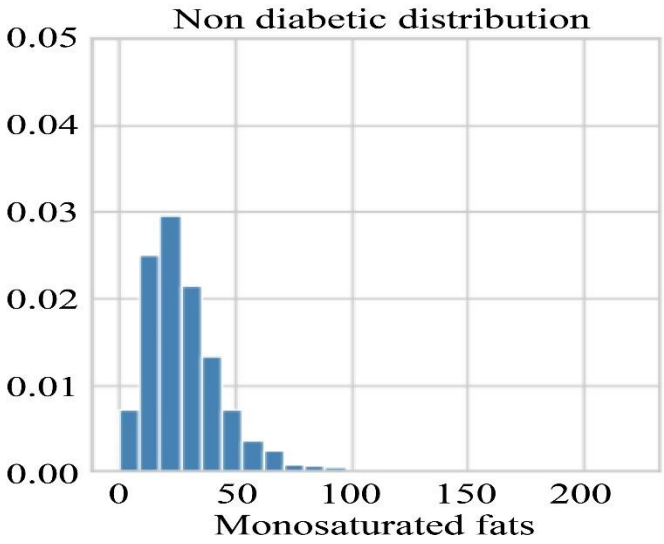


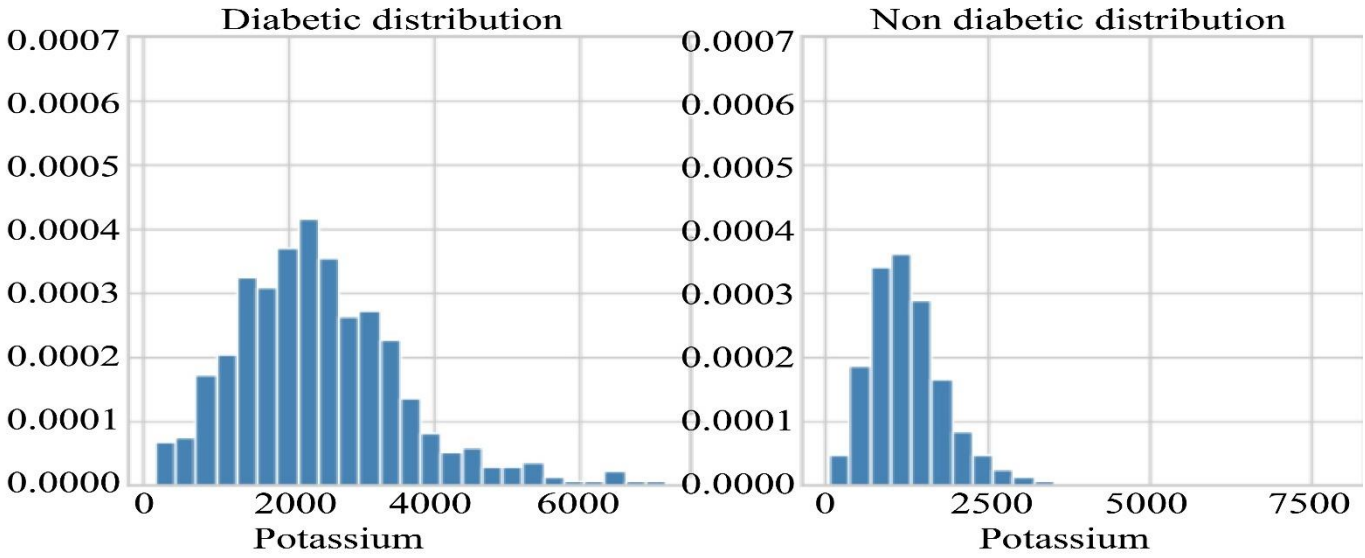


Pair wise EDA Plots

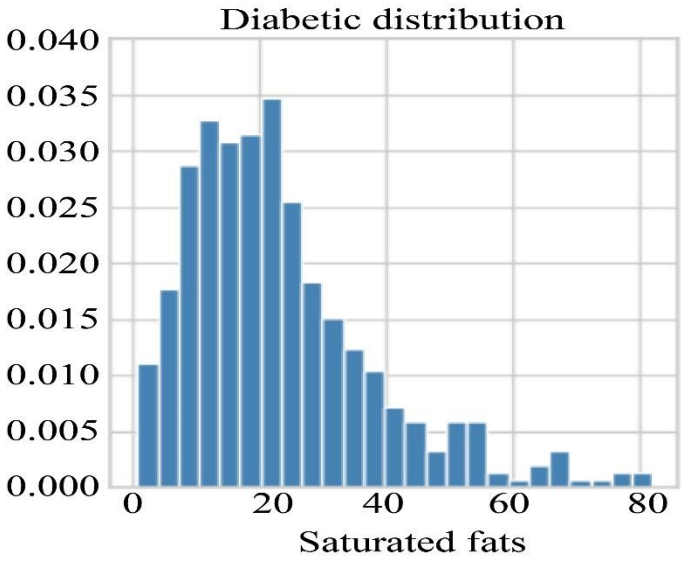


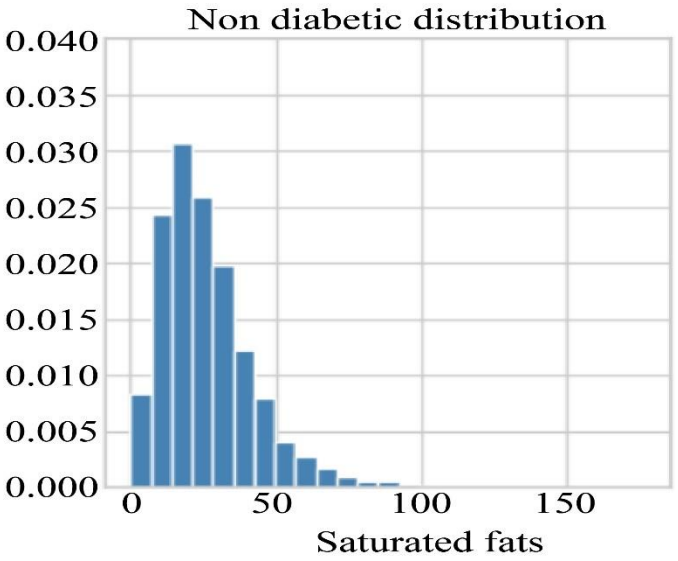


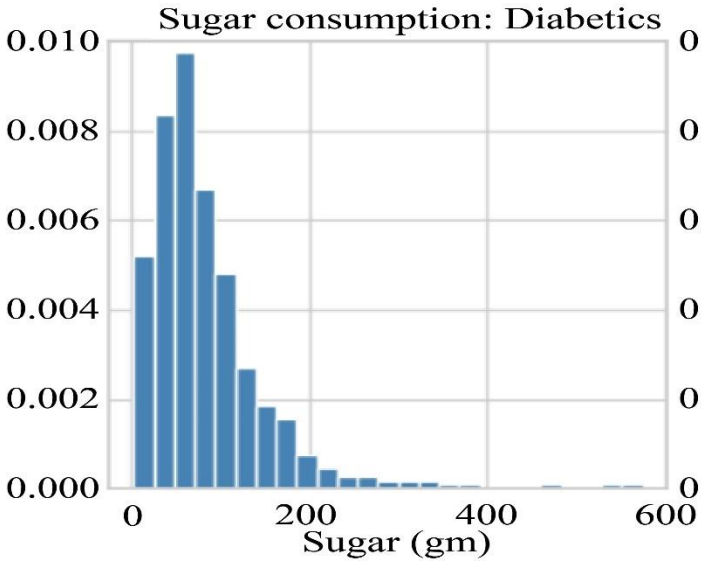


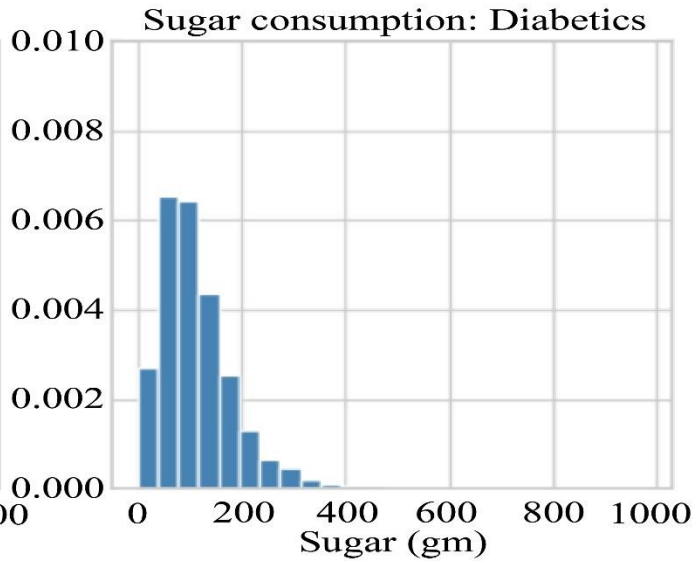

Diabetic distribution
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
2000
4000
6000
Potassium

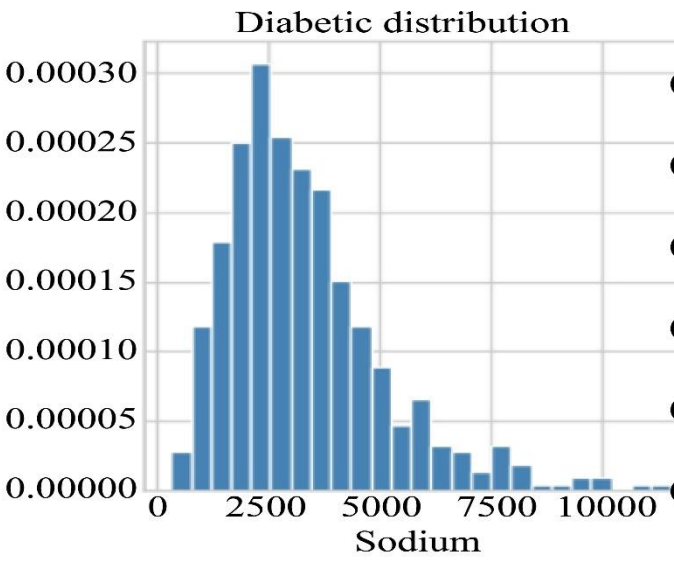
Non diabetic distribution
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
2500
5000
7500
Potassium

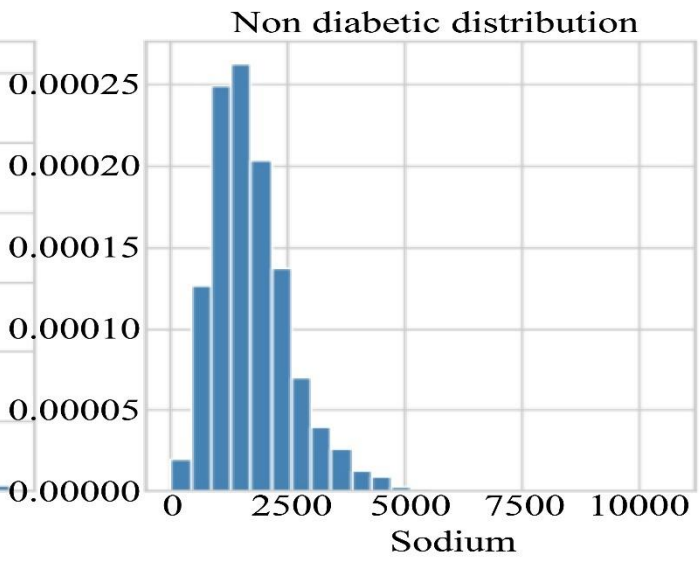

Diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
Saturated fats

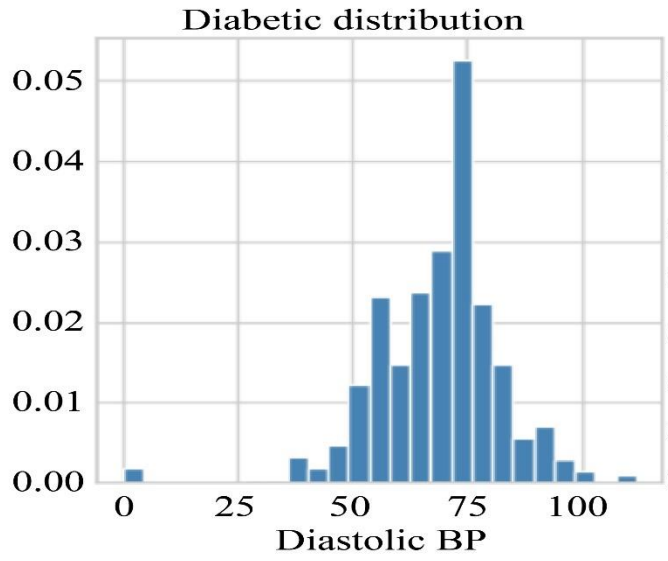

Non diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
Saturated fats

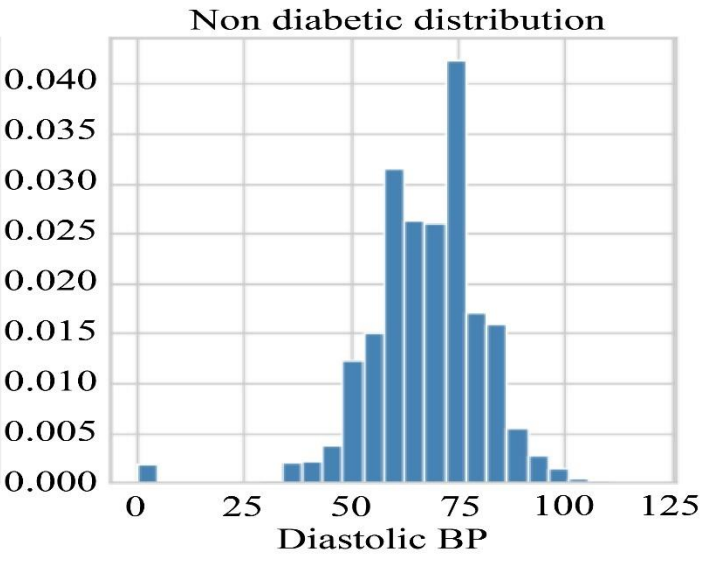

Sugar consumption: Diabetics
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
Sugar (gm)

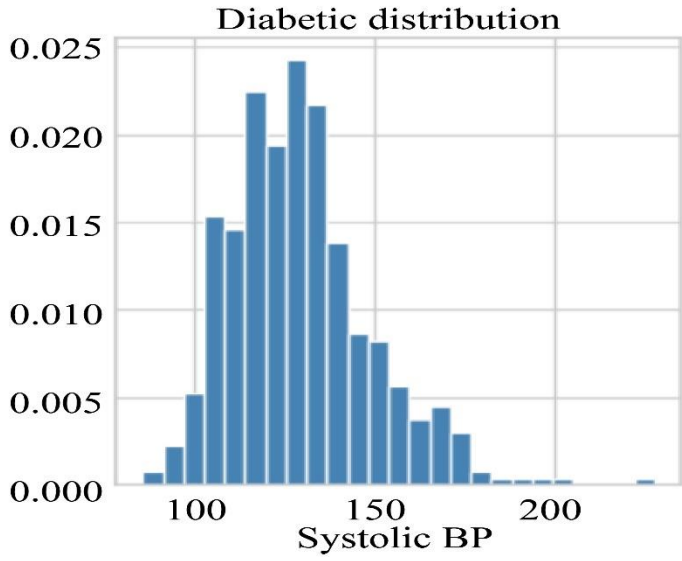

Sugar consumption: Diabetics
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
1000
Sugar (gm)

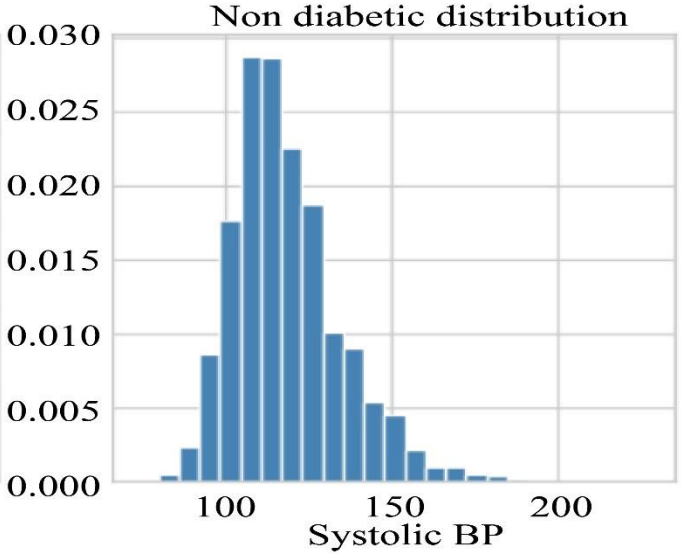

Diabetic distribution
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
Sodium

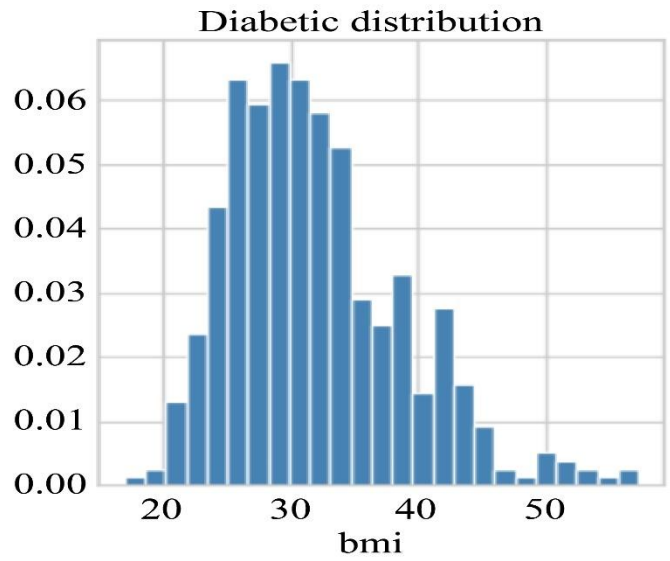

Non diabetic distribution
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2500
5000
7500
10000
Sodium

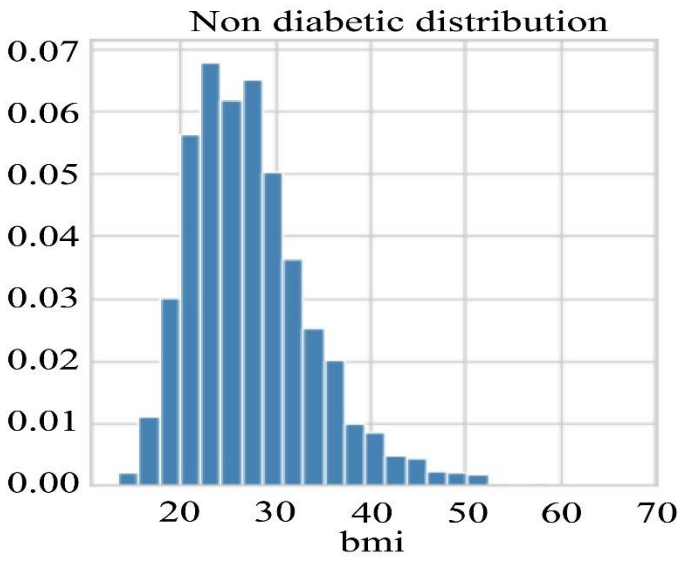

Diabetic distribution
0.05
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
Diastolic BP

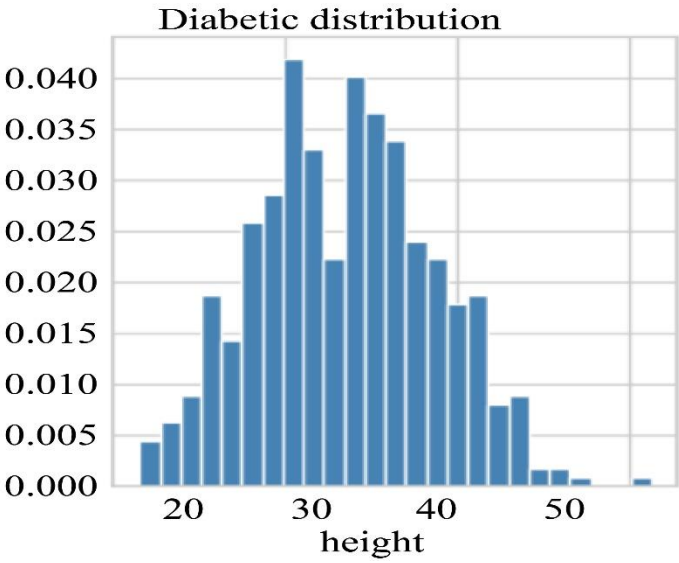

Non diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
125
Diastolic BP

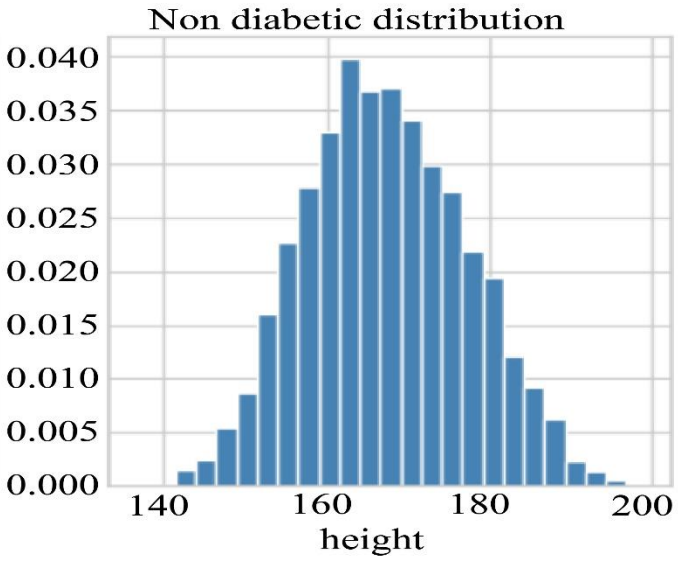

Diabetic distribution
0.025
0.020
0.015
0.010
0.005
0.000
100
150
200
Systolic BP

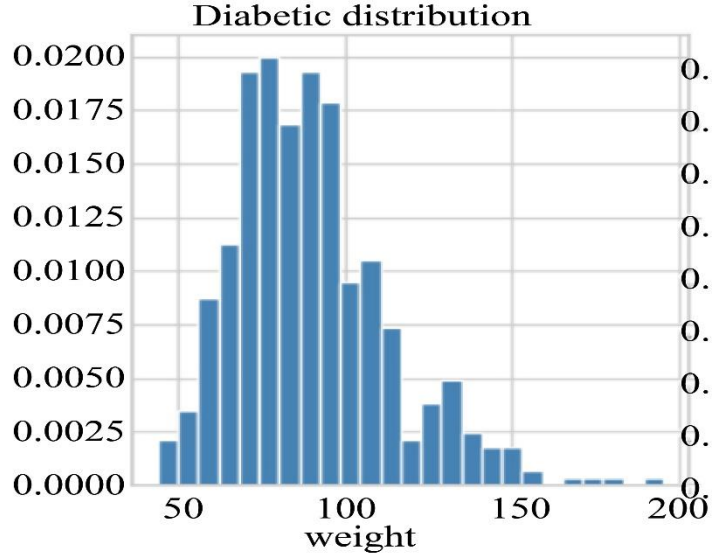

Non diabetic distribution
0.030
0.025
0.020
0.015
0.010
0.005
0.000
100
150
200
Systolic BP

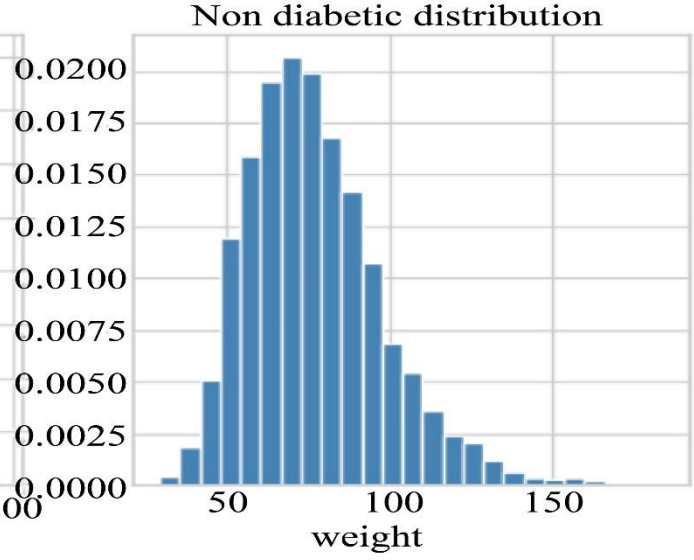

Diabetic distribution
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
bmi

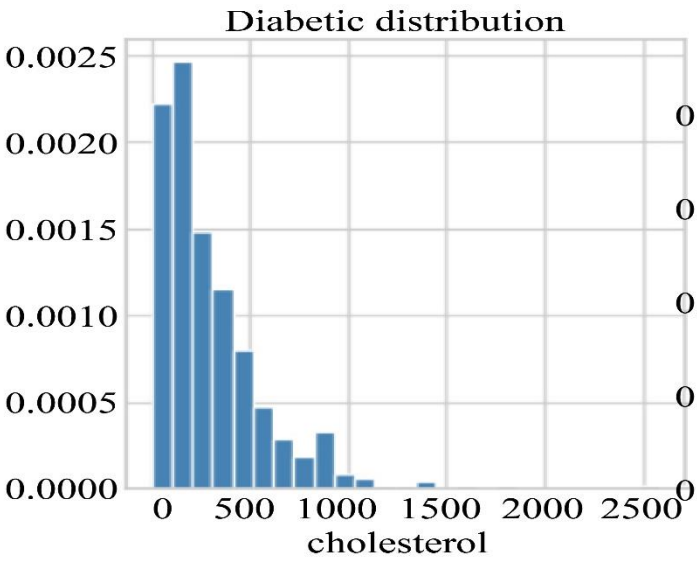

Non diabetic distribution
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
bmi

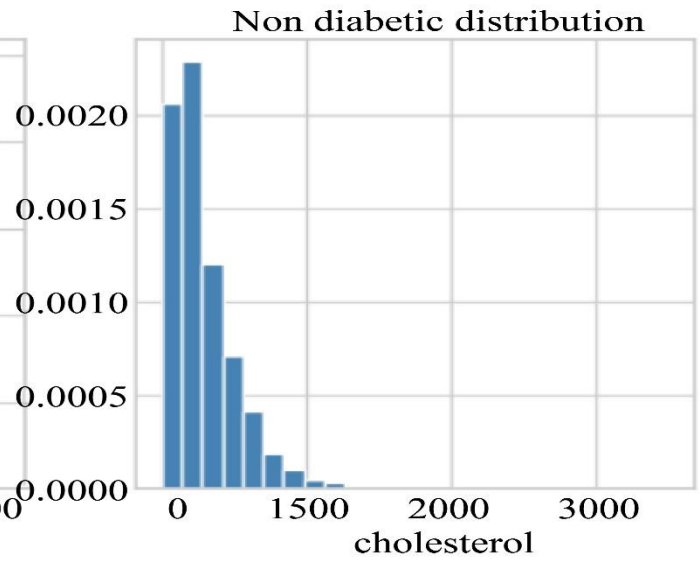

Diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
height

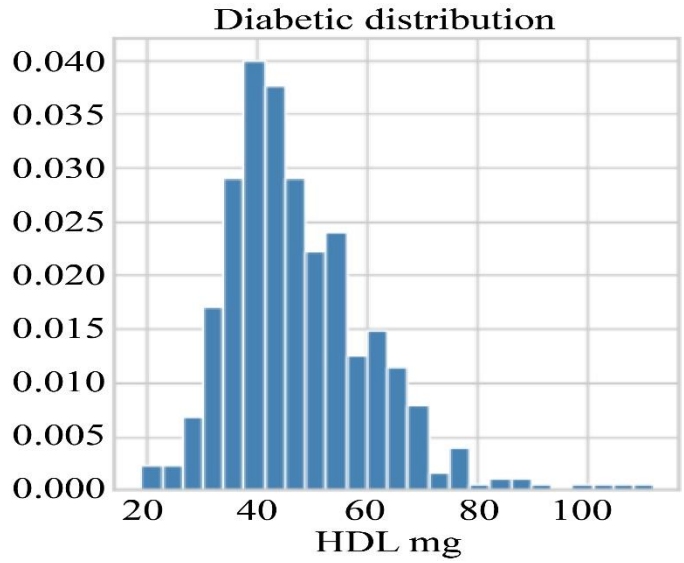

Non diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
140
160
180
200
height

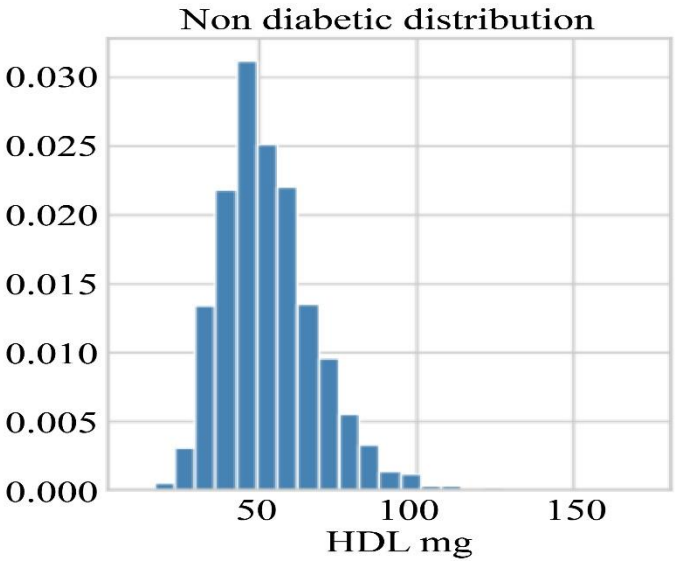

Diabetic distribution
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
200
weight

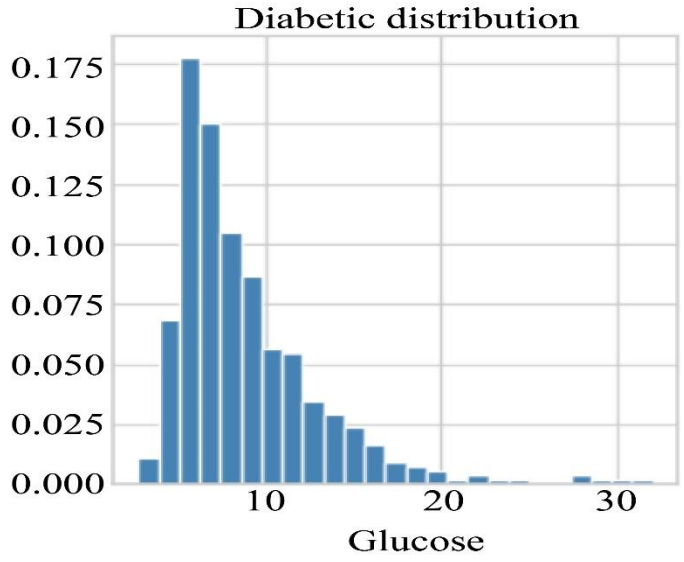

Non diabetic distribution
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
weight

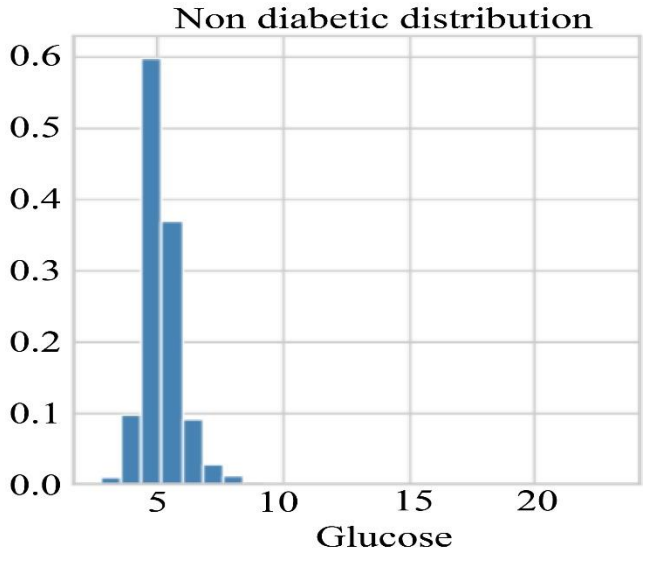

Diabetic distribution
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0 500 1000 1500 2000 2500
cholesterol

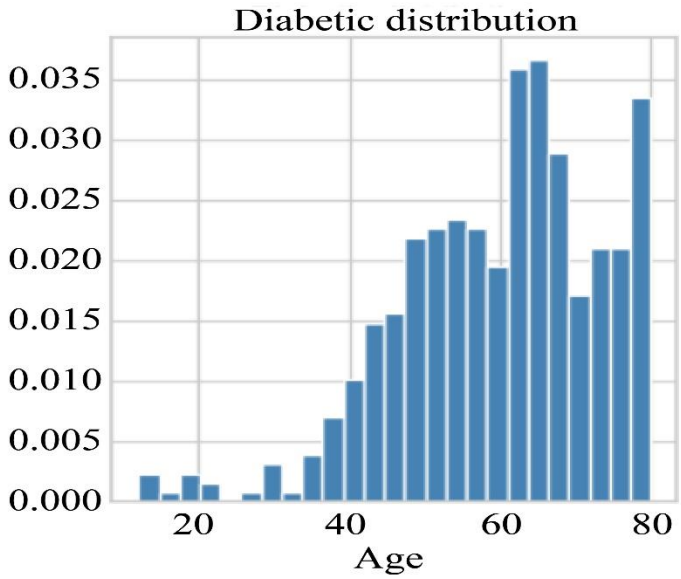

Non diabetic distribution
0.0020
0.0015
0.0010
0.0005
0.0000
0 1500 2000 3000
cholesterol

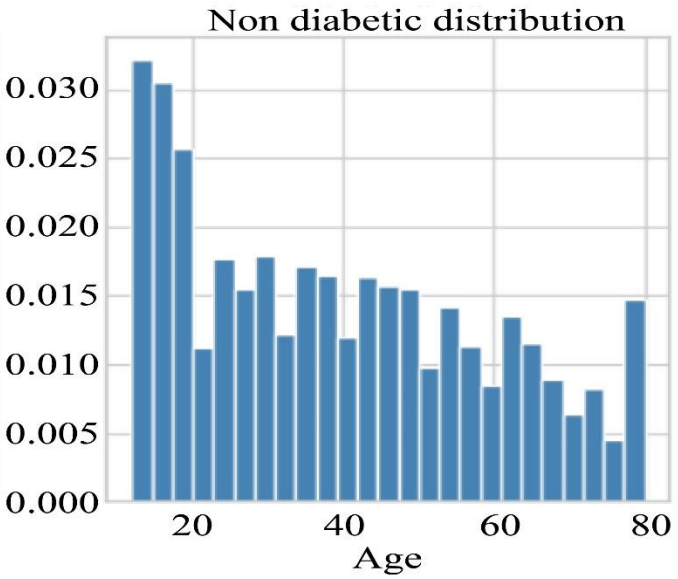

Diabetic distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20 40 60 80 100
HDL mg

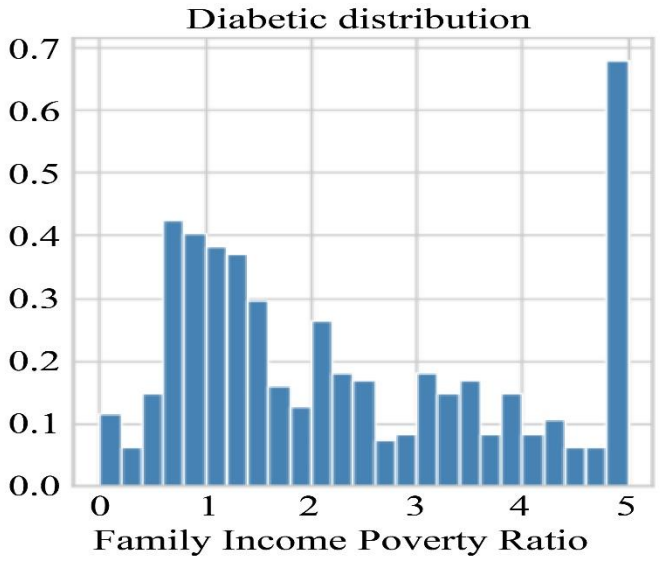

Non diabetic distribution
0.030
0.025
0.020
0.015
0.010
0.005
0.000
50 100 150
HDL mg

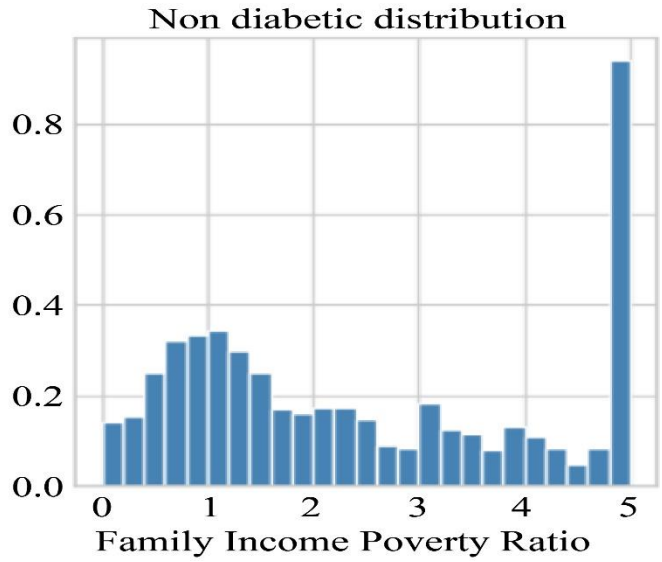

Diabetic distribution
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
10 20 30
Glucose

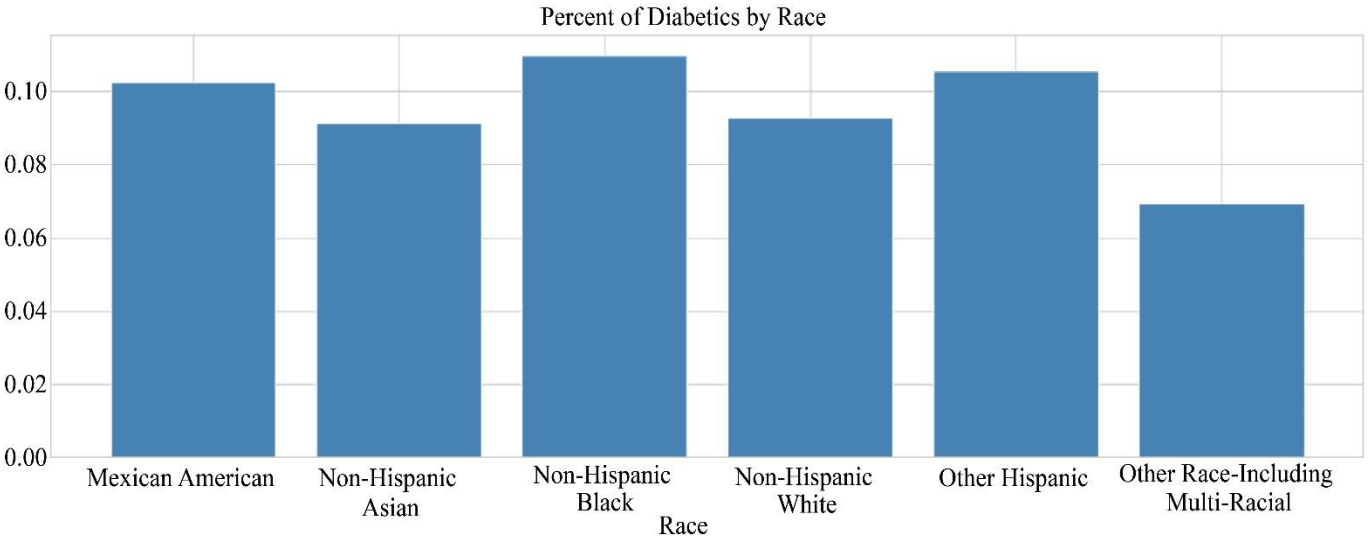

Non diabetic distribution
0.6
0.5
0.4
0.3
0.2
0.1
0.0
5 10 15 20
Glucose


Diabetic distribution
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
Age


Non diabetic distribution
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
Age


Diabetic distribution
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
Family Income Poverty Ratio


Non diabetic distribution
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
Family Income Poverty Ratio


Percent of Diabetics by Race
0.10
0.08
0.06
0.04
0.02
0.00
Mexican American
Non-Hispanic Asian
Non-Hispanic Black
Non-Hispanic White
Other Hispanic
Other Race-Including Multi-Racial
Race