Classification and Prediction of Diabetes

WQD7001: PRINCIPLES OF DATA SCIENCE (GROUP 9)

Group Members	Student ID	Role
Lai Yuen Seng	22076493	Leader
Justin Ee Qing Sem	s2111464	Oracle
Leong Lip San	17217825	Detective
Chen You Hui	22113675	Maker
Wang Zi	22105299	Secretary

University of Malaya

Faculty of Computer Science & Information Technology

January 16, 2024

Contents

Chapter 1

Introduction and Project Details

1.1 Project Background

Diabetes is a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both [1]. Diabetes is usually associated with long-term damage, dysfunction, and failure of different organs, such as the eyes, kidneys, nerves, heart, and blood vessels. Diabetes-related complications such as cardiovascular disease, kidney disease, nephropathy, blindness, and lower-extremity amputation are significant causes of increased morbidity and mortality among diabetic patients.

According to article [5], there were around 16.2 million diabetic individuals in the United States in 2005, and approximately 30% of the cases were undiagnosed. It is estimated that this number will grow to 48.3 million by 2050. These numbers will certainly create a substantial financial burden on the country's economy such as indirect costs of healthcare or unexpected disbursement due to reduced labour and productivity. Besides that, the framing of diabetes has moved from medical in 1993 to behavioural in 2001 then societal in 2013 which portrays modifiable risk factors for diabetes [8]. This has increased the importance and awareness of health monitoring nowadays.

Although medical science is rapidly growing, diabetes is still an incurable disease. Diabetes cannot be cured yet it is preventable based on the results of the controlled randomised trials stated in [9]. There is evidence which shows that intensive lifestyle interventions can effectively prevent or delay the onset of diabetes in high-risk individuals, by considering the main risk factors of diabetes.

In the early 2000s, there were several techniques for predicting diabetes, such as using statistical models and risk scores proposed in [15]. These models associate several attributes, such as age, BMI, waist circumference etc in scoring. Besides that, the emergence of Internet of Things (IoT) products such as wearable devices and sensors contributes valuable input of real-time health-related data. With relatively cheap computing power, we can adopt a machine learning approach to predict diabetes as an extension to the classical approaches [18].

As such, early diagnosis of diabetes using machine learning models can greatly benefit each individual such as early treatment to prevent complications like lower-extremity amputation, cardiovascular diseases etc and subsequently reduce the individual's financial burden and healthcare costs borne by the government.

1.2 Problem Statement

As mentioned in Section 1.1, prevention is better than cure [19]. However, the behaviour of diabetes factors changes over time and redefining the risk of factor requires to be updated time to time. Also, the rise of IoT technologies such as wearable health watch and body composition machine have provided convenience in measuring body and health data. We aim to identify the significant diabetes risk factors using statistical analysis for intensive lifestyle interventions, and develop machine learning models based on body data from wearable devices to predict whether an individual is at high risk for being diabetic.

1.3 Project Objectives

The objectives of the project were outlined as follows:

1. To identify the significant risk factors or features that results in diabetes using statistical techniques.

2. To construct an effective binary classification model to predict whether an individual is at high risk for being diabetic.
3. To develop an easily interpretable model that can explain the factors influencing diabetes to healthcare professionals and patients.

1.4 Project Scope/Domain

The project scope is outlined as follows:

1. Identify significant diabetes risk factors.
2. Patients are from adult group of age 18 and above.
3. Patients are from United States - Buckingham and Louisa .
4. To establish a machine learning model to predict the risk of getting diabetes, where input variables are quantities that can be measured conveniently using health devices.

Chapter 2

Literature Study

2.1 Introducing Diabetes and its Related Information

Diabetes is defined as chronically elevated blood sugar levels caused by problems with insulin secretion or action. It includes two main types: type 1 (autoimmune-related) and type 2 (insulin resistance and insufficient secretion). Diagnostic criteria include the hemoglobin (average 3-months blood sugar) A1c $\geq$ 6.5%, fasting plasma glucose $\geq$ 126 mg/dL, or 2-hour plasma glucose $\geq$ 200 mg/dL during an oral glucose tolerance test. Prediabetes indicates a risk for diabetes and includes impaired fasting glucose (100-125 mg/dL) and impaired glucose tolerance (2-hour values 140-199 mg/dL) [1].

Diabetes has a wide range of causes. The four main risk factors are age, gender, cholesterol, and glucose levels. Blood sugar levels are intimately associated with diabetes. Some glucose that is not used by cells enters the blood circulation and causes hyperglycemia [24]. Regular monitoring of diabetes is essential for early diagnosis, assessing patient compliance, and assessing whether antidiabetic medications are effective. Individuals who previously had poor control or who have persistent hyperglycemia should continue taking insulin and regularly check their blood glucose levels at home. This will help them determine when to reduce their insulin dosage and when to transition to non-insulin medications [1].

Research on cholesterol and diabetes indicates. Diabetes is thought to be a separate predictor of sudden cardiac arrest (SCA), and studies on the relationship between dyslipidemia and SCA risk have produced contradicting findings. Low-density lipoprotein (LDL) cholesterol is a marker of dyslipidemia, which is higher in people with diabetes and is recognized as a risk factor for cardiovascular disease [13]. Both low and high LDL cholesterol levels are linked to an increased risk of SCA in diabetic patients [13].

Recently published research on gender in diabetes shows low levels of reporting of gender and gender considerations in published diabetes studies [4].

Age and the length of diabetes seem to interact, according to a study on the condition. Due to the longer duration of the condition, patients younger than 60 years old have a higher risk of developing macrovascular and microvascular problems than do those older than 60 [26]. Only the duration of diabetes was found to be independently connected with microvascular problems, while the length of the disease was linked to macrovascular issues.

The relevant study shows that intensive management is important for all patients with type 2 diabetes, especially those with longer duration who are at a highest risk of complications. Age and age of diagnosis are also factors contributing to risk of macrovascular beyond duration alone. In multivariate analysis, older age, older age at diagnosis, and longer diabetes duration were independently associated with increased risk of macrovascular complications. After accounting for other risk factors, the only factor that was independently linked to microvascular problems was longer diabetes duration rather than age or age at diagnosis [20].

In recent years, although diabetes-related deaths have decreased, both type 1 and type 2 diabetes are on the rise across all age groups. This increased prevalence results from a combination of factors, including rising incidence among younger individuals, advancements in diagnosis and treatment, and better management of cardiovascular risk factors. Nevertheless, adults and adolescents with diabetes still face elevated mortality rates compared to the general population in [25]. The primary causes of death in this group include acute diabetic complications like hypoglycemia and ketoacidosis, cardiovascular diseases (CVDs), and renal failure.

2.2 Handling Missing Data

Even in a well-designed and controlled study, missing data occurs in almost all research. In many real-world data sets, missing values are common due to many reasons, including human errors, faulty sensors, or simply the absence of data for certain observations. Generally, the rows of a data matrix represent units, observations, cases or subjects depending on context. The column of a data matrix usually represents characteristics or variables measured for each observation [10]. Standard statistical methods have been introduced to analyze rectangular data matrices with no missing values. As such, missing data can reduce the statistical power of research and able to produce biased model estimates, leading to invalid conclusions [12].

There are several types of missing data as mentioned in [22].

1. ***Missing Completely at Random (MCAR):*** Data are MCAR when the missing data is independent of the observed and unobserved data. The advantages of data that are MCAR is that the estimated parameters will not be biased, though there might be a drop in statistical power, and often unrealistic. E.g., some participants could have missing values in their laboratory test due to faulty equipment.
2. ***Missing at Random (MAR):*** Data are MAR when the missing data is related to the observed data but not the unobserved data. As such, a complete statistical analysis may induce bias, and could be corrected if proper accounting for the known factors is done. E.g., a depression study might observe that the males tend to leave blanks in their questionnaire as compared to females, where the gender data is assumed to be complete.
3. ***Missing Not at Random (MNAR):*** Data are MNAR when the missing data is related to the observed data and the unobserved data. Similar to MAR, bias could be corrected, but the fact that the origin of missing data are themselves unmeasured means that this issue cannot be addressed in analysis and the conclusion might be biased. E.g., a depression study might observe MNAR data as heavily depressed participants may not complete the survey about depression severity.

Aside from leading to biased results, many machine learning models require the data set to be free of missing values. Therefore, many simple and sophisticated approaches have been developed to deal with missing values, and mostly based on the groundwork developed by [16, 22]. Here are some popular data deletion and imputation techniques from [6, 12].

1. ***Listwise or Case Deletion:*** This approach simply exclude those cases with missing data and analyze the remaining data. When the assumptions of MCAR are not satisfied, it may induce bias in parametric estimation. This strategy is reasonable when the MCAR assumptions are satisfied, power is not an issue and and the sample size is large enough
2. ***Pairwise Deletion:*** This approach only eliminates information when the particular data-point needed to test an assumption is missing. If there are missing values elsewhere, only the existing values will be used. However, it might result in different model statistics, such as sample size and degree of freedom. Though this approach is is less biased than the latter for MCAR and MAR data, too many missing values might result in deficient analysis.
3. ***Mean or Median Substitution:*** This approach involves replacing the missing values of a variable using the mean or median of that variable. When there are some extreme outliers, median imputation would be a more robust approach. However, with missing values that are not strictly random, it might lead to inconsistent bias. Also, there are no new information added using this approach, which might lead to an underestimate of errors.
4. ***Regression Imputation:*** In regression imputation, the existing variables are used to make a prediction, then the predicted value is substituted at the missing variable data. This approach works well when there exists some relationship between the missing variable and the other variables. Similarly, this approach might lead to an underestimate of errors.
5. ***Machine Learning-Based Imputation:*** These are sophisticated techniques that mostly involve a predictive algorithm to handle missing values using supervised or unsupervised learning. Common algorithms include

Decision Tree, k-Nearest Neighbour (k-NN) and Support Vector Machine (SVM). These models are able to capture complex relationships of the data, but are usually more computationally intensive.

To summarize, it is important to know the pros and cons of the imputation techniques in order to choose the appropriate method based on the nature of missing data, data set background and analysis goals. This is important as improper approaches may lead to improper inferences and conclusions.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a philosophy for data analysis that employs techniques either graphical or non-graphical [11] to:

1. Maximize insight and uncover underlying structure of a data set;
2. Extract important variables;
3. Detect outliers and anomalies;
4. Test underlying assumptions and develop parsimonious models.

Besides EDA, there are two other popular approaches, namely the Classical approach and Bayesian approach [11].

The Classical approach begins with data collection, followed by model imposition (normality, linearity, etc.) and the analysis, estimation and testing were based on the model parameters. This approach is more inclined towards confirming or rejecting predefined hypotheses, which limits our ability to reveal unexpected insights from the data.

The Bayesian approach incorporates prior domain/scientific knowledge by imposing a data-independent distribution on the parameters of the selected models (prior distribution). The analysis is then made by combining the prior and the data likelihood to derive the posterior distribution for inferences and/or test assumptions about the model parameters. While Bayesian analysis can be a powerful tool for drawing conclusions from data, it often requires a strong understanding of statistical theory and can be computationally intensive.

On the contrary, EDA works differently with the two approaches above, where data collection is not followed by a model imposition; rather it is followed by a straightforward analysis with a target of deciding what model would be appropriate. As compared to the classical approach, EDA is more open-ended as it does not involve an initial model imposition. When compared to the Bayesian approach, it is less technical and more accessible to broader users as it emphasizes on data exploration, visualization and summarization rather than formal statistical inference.

In summary, EDA is a crucial step in data analysis that differs from Classical and Bayesian data analysis in its primary purpose and approach. It is exploratory in nature, emphasizing on data visualization and understanding without preconceived models/hypotheses, making it a useful approach to reveal hidden insights and anomalies, discover underlying relationships among the variables and to summarize our data.

Here are some widely used graphical techniques used in EDA [2]:

1. ***Histogram:*** A histogram typically shows the location, spread, skewness, presence of outliers and presence of multiple modes in a data.
2. ***Boxplot:*** Box plot is an excellent tool for conveying location and variation information in data set, particularly for detecting and illustrating location and variation changes between different groups of data, and the presence of outliers.
3. ***Quantile-Quantile Plot:*** The quantile-quantile (QQ) plot is a graphical technique for determining if the univariate data comes from a specific distribution by plotting the empirical quantile function of the data against the quantile function of the theoretical distribution. If the data comes from the theoretical distribution, the points should fall approximately along the 45-degree reference line.
4. ***Scatter Plot:*** A scatter plot reveals relationships or association between two variables. Such relationships manifest themselves by any non-random patterns in the plot. Furthermore, it can be used to identify the plausible outliers in the data set.
5. ***Bar Charts:*** A chart that represents categorical data in rectangular bars, with height proportional to their values or observation counts. They are used to compare categorical variables.

Apart from graphical techniques, there are also some numerical techniques used in EDA [11].

1. ***Mean:*** The mean, also known as the arithmetic mean, is the sum of the data points divided by the size of data points. It is a measure of central tendency (location) and is generally a preferred measure when the distribution is symmetric.
2. ***Median:*** The median is defined as the value of the point which has half the data smaller than that point and half the data larger than that point. It is a measure of central tendency (location), and is more robust to outliers when compared to the mean, and is generally a preferred measure when the distribution is heavily skewed.
3. ***Variance & Standard Deviation:*** The variance is defined as the average squared distance from the mean, whereas the standard deviation is defined as the square-root of variance. Both techniques are measures of dispersion (scale).
4. ***Skewness (Fisher-Pearson coefficient of skewness):*** The skewness is a measure of symmetry. Non-zero skewness indicates that the distribution is asymmetric.

Chapter 3

Methodology and Results

3.1 Obtaining the Dataset

3.1.1 Data Set Sources and Reliability

The diabetes data set [23] are courtesy of Dr John Schorling, Department of Medicine, University of Virginia School of Medicine. The data consist of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans.

This data set was obtained from `http://hbiostat.org/data` courtesy of the Vanderbilt University Department of Biostatistics. As such, it is convincing to say that the data set is reliable for our project.

3.1.2 Types of Variables

We use the Python library-Pandas to explore the structure of the data set. Table 3.1 summarizes the variables present in the data frame, which specifies their units, interpretation, data type and the number of missing values.

Table 3.1: Data Structures and Interpretation

Name	Labels	Units	Levels	Data Type	No. Missing Values
id	Subject ID			int64	0
chol	Total Cholesterol			float64	1
stab.glu	Stabilized Glucose			int64	0
hdl	High Density Lipoprotein			float64	1
ratio	Cholesterol/HDL Ratio			float64	1
glyhb	Glycosolated Hemoglobin			float64	13
location	(Buckingham or Louisa)		2	object	0
age		years		int64	0
gender	(male or female)		2	object	0
height		inches		float64	5
weight		pounds		float64	1
frame	Body Frame (small/medium/big)		3	object	12
bp.1s	First Systolic Blood Pressure			float64	5
bp.1d	First Diastolic Blood Pressure			float64	5
bp.2s	Second Systolic Blood Pressure			float64	262
bp.2d	Second Diastolic Blood Pressure			float64	262
waist		inches		float64	2
hip		inches		float64	2
time.ppn	Postprandial Time when Labs were Drawn	minutes		float64	3

3.2 Data Scrubbing

Data Scrubbing or cleansing is the crucial preparation process of detecting and correcting anomalies, errors or inaccurate information in our datasets. In our 'Scrub' stage, we used different techniques to strengthen the quality of the data as it is essential and has a significant impact on making informed decisions and drawing conclusions from our data-driven process.

From Table 3.1 shown in Section 3.1.2, the variables 'bp.2s' and 'bp.2d' have 262 null values out of 403 subjects. In addition, they are considered trivial variables, therefore we will be dropping these 2 columns from our dataset due to a large number of missing values and minor importance.

Due to the raw data obtained does not contain any response variable, hence we classified our subject's diabetic condition based on the 'glyhb' variable. According to [1], for an HbA1c test to be classified as normal or non-diabetic, the 'glyhb' value must be below 5.7%. Within the range of 5.7% to 6.4% is considered prediabetic and with 6.5% and above, diabetes can be diagnosed. Therefore, we will be classifying the diabetic condition as if the 'glyhb' level is above 6.5, it will be categorised as diabetic and vice-versa. The column of variable 'glyhb' will be deleted after the classification as it is not part of the significant risk factors or features that we trying to analyse to predict diabetes condition. Individuals without 'glyhb' will be deleted.

We will be applying the imputation technique to fill in the missing values for variables such as 'chol', 'hdl', 'ratio', 'height', 'weight', 'frame', 'bp.1s', 'bp.1d', 'waist', 'hip' and 'time.ppn'.

For 'chol', 'hdl', and 'ratio', the data is missing completely at random and there is only one missing value for each variable, hence we used the median of each column to fill in the missing values. The median is a better option here as it is less sensitive and more robust to the outlier, which increases the reliability and stability of the result of statistical analyses and modelling later.

We utilised the machine learning method for imputating 'weight'. To predict the value of weight, we perform multiple linear regression on 'hip' and 'waist' features. After we filled up the 'weight' column, we performed multiple linear regression as well to predict the missing values of 'hip' and 'waist'. This is because of the strong linear correlation between these 2 variables as shown in the matrix below from A.1 and scatter plots from A.2.

Table 3.2: Correlation Matrix between 'weight' with 'waist' and 'hip'

Attributes	waist	hip
weight	0.851561	0.829791

We imputed missing values of 'height' based on the median height grouped by gender, considering the approximate Normal distribution observed from the Quantile-Quantile plots in A.5.

In the following variables such as 'bp.1s', 'bp.1d' and 'time.ppn', we will be using the median imputation as well to eliminate the missing values. Although 'bp.1s' and 'bp.1d' are highly correlated, a regression-based imputation is not possible as the missing values come in pairs.

Lastly, we performed imputation for the variable 'frame' using a Decision Tree Classifier. Different variables were considered when imputing the variable 'frame' which we think has a strong correlation such as 'gender', 'height', 'weight', 'hip' and 'waist'.

3.3 Data Exploration

Our aim of this project is to identify the significant risk factors or features that result in diabetes, and utilise statistical techniques to construct an effective binary classification model to predict whether an individual is at high risk for being diabetic.

Hence, our EDA will mainly focus on exploring graphical and numerical summaries of diabetic and non-diabetic patients, along with outliers and distributional exploration.

3.3.1 On Categorical Variables

There are 4 categorical variables to be explored, namely 'location', 'frame', 'gender' and 'diabetic'. By considering the bar charts A.6 and proportions B.2 of the first 3 variables with respect to 'diabetic', we observe the following:

1. On 'location' wise, the *Buckingham* individuals seems to have a slightly higher proportion of diabetic patients as compared to *Louisa*.
2. On 'frame' wise, *large* frame individuals have the highest proportion of diabetic patients, followed by *medium* frame and *small* frame. The reason for this could be that, *large* frame individuals might tend to be obese, which is a contributing factor to diabetes [14].
3. On 'gender' wise, *male* individuals have a higher proportion of diabetic patients as compared to *female*. This is consistent with the results from [3].

Overall, we observe that the data set is dominated by non-diabetic patients, As such, when training for machine learning models, we have to address this issue by considering suitable techniques such as over-sampling and under-sampling as mentioned in [17].

3.3.2 On Central Tendency

By considering the mean and median B.3 of blood test results, we can observe that the mean and median levels of 'chol', 'stab.glu' and 'ratio' of diabetic patients are higher than non-diabetic patients, whereas the mean and median level of 'hdl' for diabetic patients are higher than non-diabetic patients.

Aside from that, we observe that the mean and median 'age' and 'bp.1s & bp.1d' of diabetic patients are higher than non-diabetic patients.

Lastly, on body measurements, it seems that diabetic patients have bigger body frames, as indicated by their higher mean and median 'height', 'weight', 'hip' and 'waist' measurements. This finding is consistent with Section 3.3.1.

Interpretation of box plots and histograms in A.4 and A.3 respectively yield similar observations.

3.3.3 On Dispersion and Outliers

Based on numerical measures of dispersion in B.4 and the box plots in A.4, the most notable variation is the 'stab.glu' level. We observe that the distribution of 'stab.glu' for diabetic patients is heavily dispersed as compared to non-diabetic patients. There are a couple of possible reasons for such results. Firstly, it can be due to the inconsistent variation in 'time.ppn' that results in the rapid fluctuations in 'stab.glu'. Besides, non-diabetic patients tend to stabilize their 'stab.glu' level rapidly, resulting in a lower variation of 'stab.glu' level.

Other notable observations include a lower 'hdl' level variation for diabetic patients as compared to non-diabetic patients. By considering the median values, we can say that diabetic patients seem to have generally lower 'hdl' ('good cholesterol') levels.

Furthermore, higher median and smaller dispersion of 'age' for diabetic patients suggest that most diabetic patients are older than non-diabetic patients.

Lastly, higher median and smaller dispersion of body measurements such as 'height', 'weight', 'waist' and 'hip' for diabetic patients suggest that most diabetic patients are generally bigger in body size.

There are two suspected outliers based on the box plots A.4, Namely the two extremes observed in 'stab.glu' and 'ratio'.

Careful inspection on the 'ratio' suggested that the observation might not be an outlier, as the ratio of 'chol' to 'hdl' is consistent with the observed 'ratio'.

However, for the extreme 'stab.glu' observation, further literature study is required to determine whether the observation should be excluded as the 'stab.glu' level is unusually high and is non-diabetic. It might be due to human error when inserting the data.

3.3.4 On Linear Relations and Distributions

When we examine the pairwise scatter plots A.2 and the linear correlation coefficients A.1, we can deduce some notable highly correlated pairs. They have been summarised in the table below.

Table 3.3: Notable Linearly Correlated Pairs

Variable Pairs	Pearson Correlation Coefficient	Strength
weight $\sim$ *waist*	0.853	Strong Positive Linear Relationship
hip $\sim$ *waist*	0.835	Strong Positive Linear Relationship
weight $\sim$ *hip*	0.831	Strong Positive Linear Relationship
ratio $\sim$ *hdl*	-0.682	Moderately Strong Negative Linear Relationship

With such relations, it helps us to determine redundant attributes when building a statistical model or machine learning models.

Moving on, we examine all continuous variables without considering the diabetic status. The Quantile-Quantile-plots in A.5 suggest that the standardised 'height' and 'bp.1d' seem to follow a standard normal distribution. This is further supported by their relatively small skewness values (0.040574 and 0.250472 respectively) as compared to the theoretical skewness of a normal distribution, which is 0.

Other standardised variables seem to not fit well with the standard normal distribution, as seen by the tail deviation along the theoretical Quantile-Quantile line.

Such information allows us to choose appropriate inferential tests for future work.

Chapter 4

Ethical Considerations

Our primary objective in conducting this study is to anticipate the likelihood of developing diabetes in order to raise awareness within society.

Throughout the study, we maintain a strong sense of morality and ethics. Our utmost priority is to minimize any potential harm that may arise, be it physical, social, or psychological. For instance, the data required for our analysis of diabetes classification and risk prediction is sourced from the Vanderbilt University Department of Biostatistics.

We have duly acknowledged and credited the relevant party for providing us with this data. It is important to emphasize that this data will only be used for the purpose of our present study on classification and risk prediction, and will not be shared, disclosed, or published without proper credit given to the related party. Additionally, since we have not conducted any primary data collection or surveys, consent is not required.

Furthermore, the data set does not contain any names, ensuring complete anonymity of the information. This guarantees that personal information cannot be traced or linked, and upholds a high level of confidentiality and ethical standards.

Chapter 5

Project Impact to the Society

With the rise of IoT technology, the use of data analytics and IoT technology to identify risk factors for diabetes has gradually increased. Smart devices will have the capacity to measure factors like a patient's blood glucose level and provide the user or doctor with real-time information about it through mobile or online applications. Our project will develop an easily interpretable model, this model will be based on the data from the wearable device model. This model will assist patients, potential risk groups, and healthcare professionals in understanding the factors that affect diabetes. Our project will predict whether a person is at high risk for being diabetic. We believe that our project will have a positive impact on our society and will be able to help the people associated with diabetes.

Our project will create an impact on the public and increase awareness of significant diabetes risk factors. In order to better understand which factors are more likely to lead to diabetes, our project will use statistical techniques to identify significant risk factors that are associated with the disease. People can regulate their dietary habits and increase the right amount of exercise to prevent the risk of diabetes based on the significant risk factors. The development of IoT technology has made it easier to track factors that may be linked to diabetes. People are beginning to realize how essential it is to change their dietary habits and to regard physical activity as crucial because they focus on health monitoring data and the results of our project.

Our project constructs an effective binary classification model that will provide society with a model for determining whether an individual is at high risk for diabetes. Individuals at high risk for diabetes have some common predisposing characteristics. Especially for those groups who retire early or exit from the labour market due to health issues, their health problems may increase the risk of diabetes [7]. These low-income or non-income groups will face an additional financial burden if they have the misfortune to get diabetes. Additionally, our project aims to make this group of people aware of how important diabetes prevention is. Furthermore, our project also aims to help the increasing proportion of elderly people in our society. Diabetes is a chronic disease with a high prevalence that has a detrimental effect on the quality of life of the elderly [21]. Early detection of diabetes can greatly reduce its serious impact on human life.

Our project aims to raise public and social awareness of the significance of diabetes prevention, the key factors influencing diabetes, and how to prevent diabetes triggers in everyday life. We also hope that this knowledge will be shared with those around them. Furthermore, our project constructs a model which will also help medical professionals to better categorize diabetes, helping them to predict which factors will lead to diabetes and distinguish whether an individual is at high risk for having diabetes. In a nutshell, this project aims to enhance the consciousness of society about diabetes prevention while also assisting in reducing the national financial burden, to create a healthier and euphoric community.

Appendix A

Appendix

A.1 Correlation Matrix

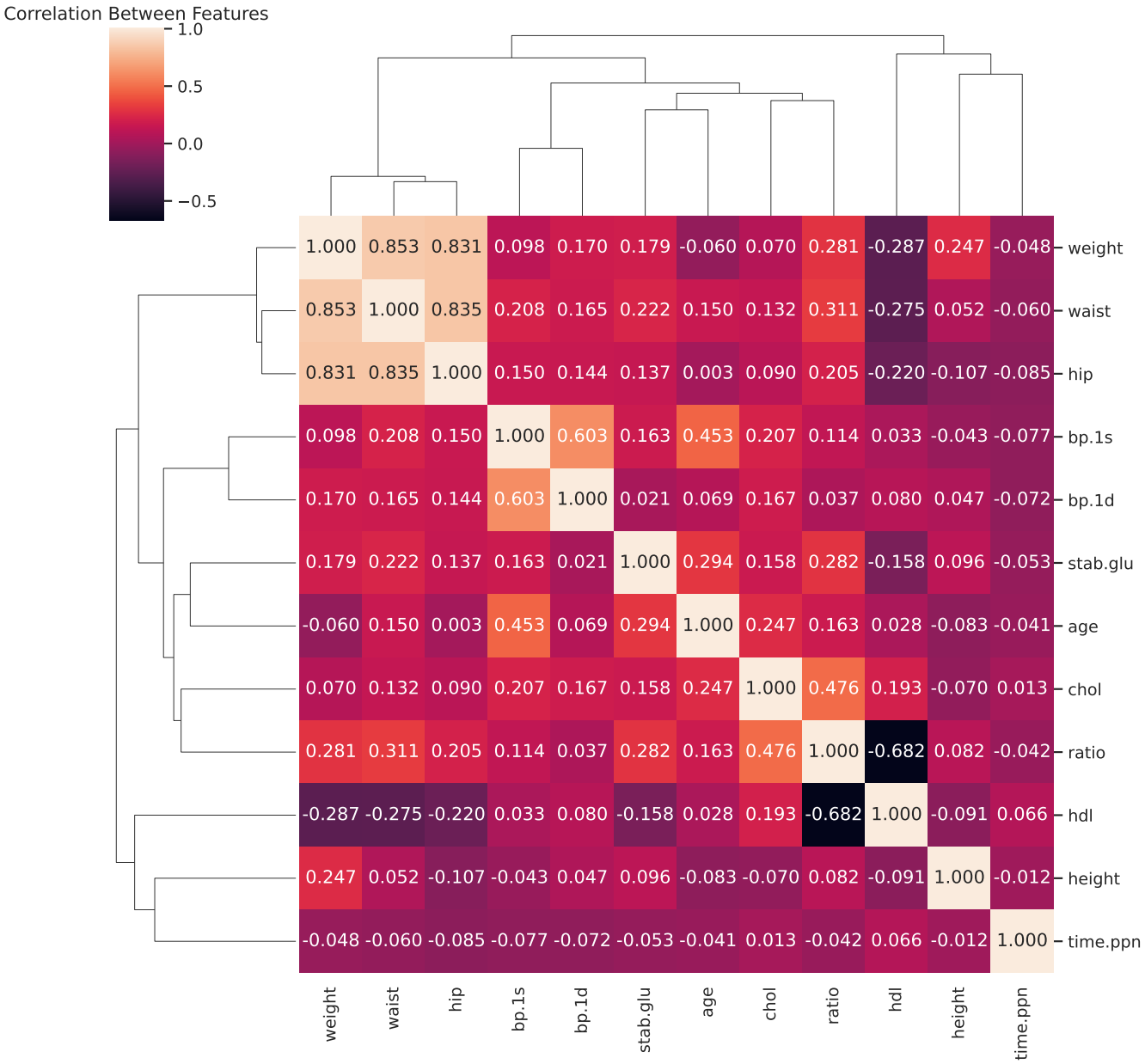

Figure A.1: Correlation Matrix

A.2 Pairwise Scatter Plots

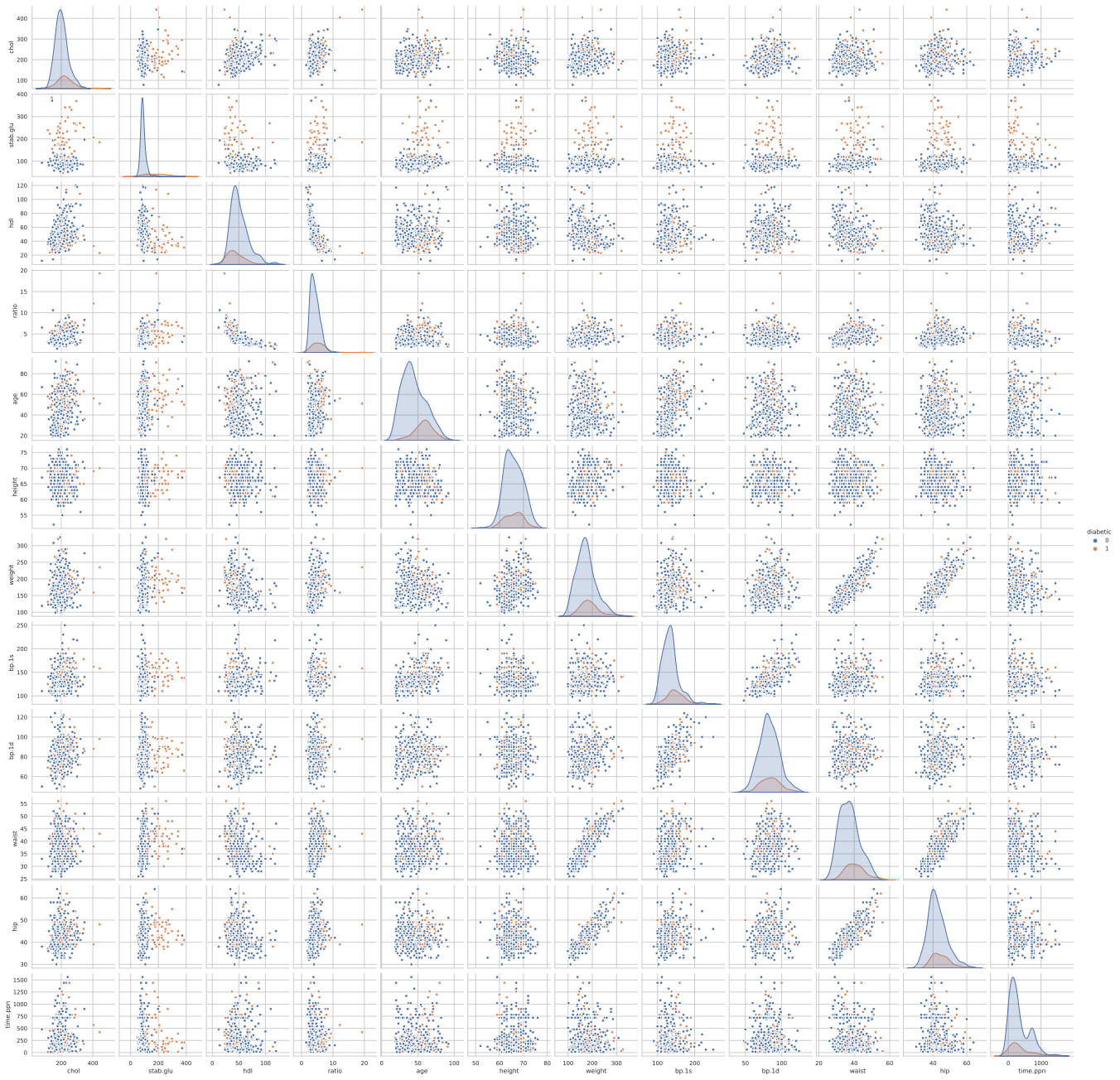

Figure A.2: Pairwise Scatter Plots

A.3 Histograms

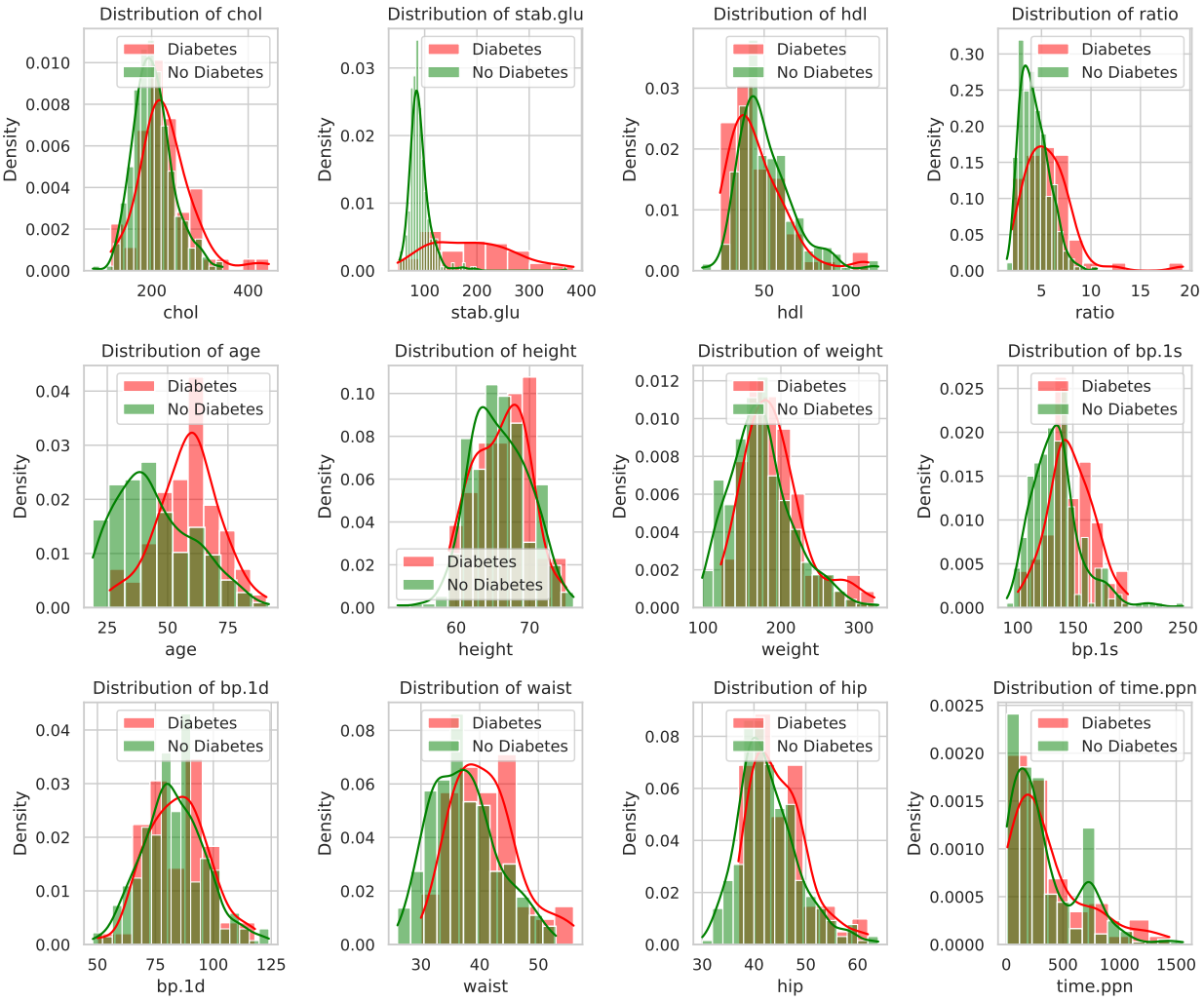

Figure A.3: Histograms of Continuous Variables

A.4 Boxplots

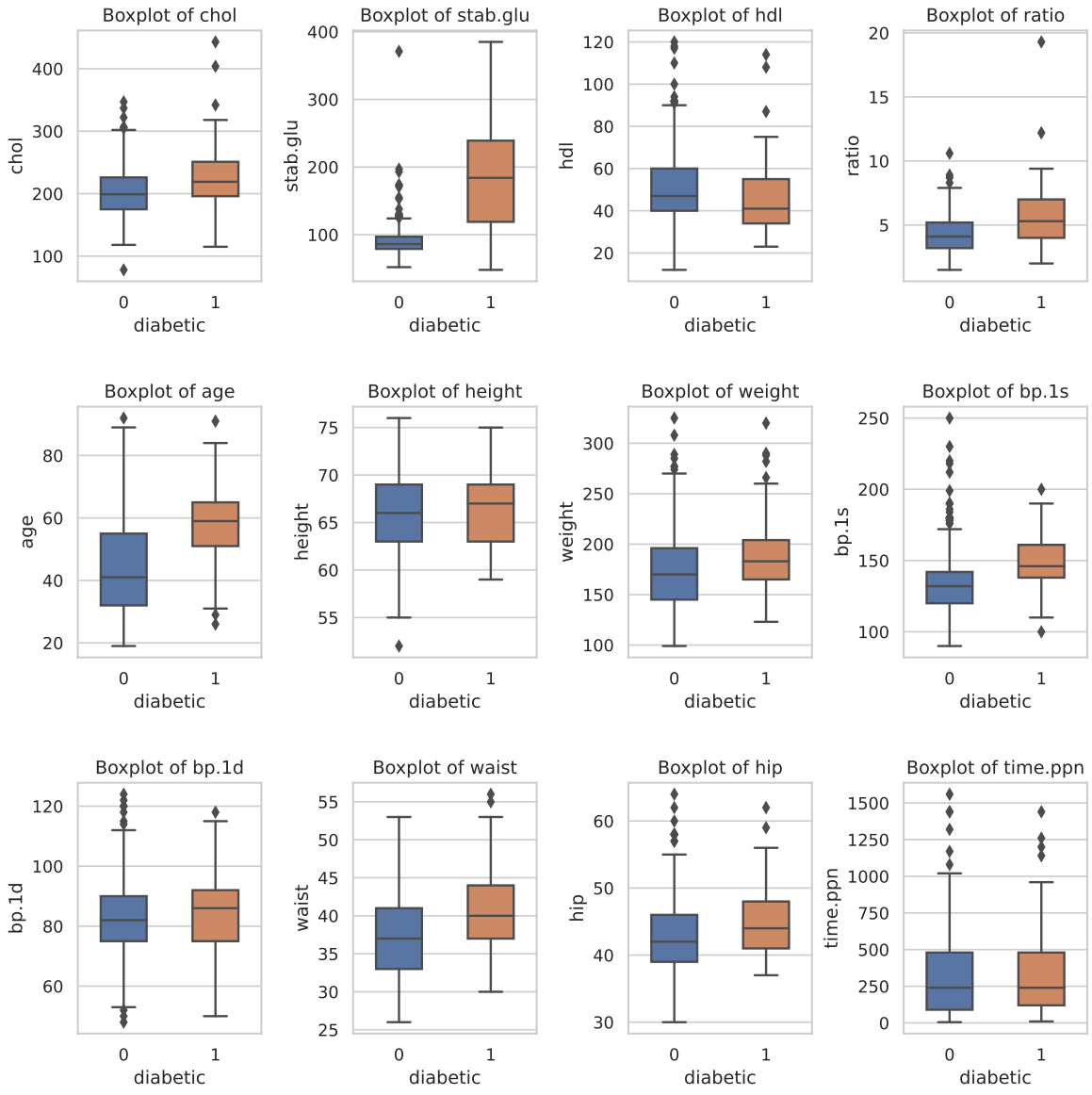

Figure A.4: Boxplots of Continuous Variables according to Diabetic Status

A.5 Quantile-Quantile Plots

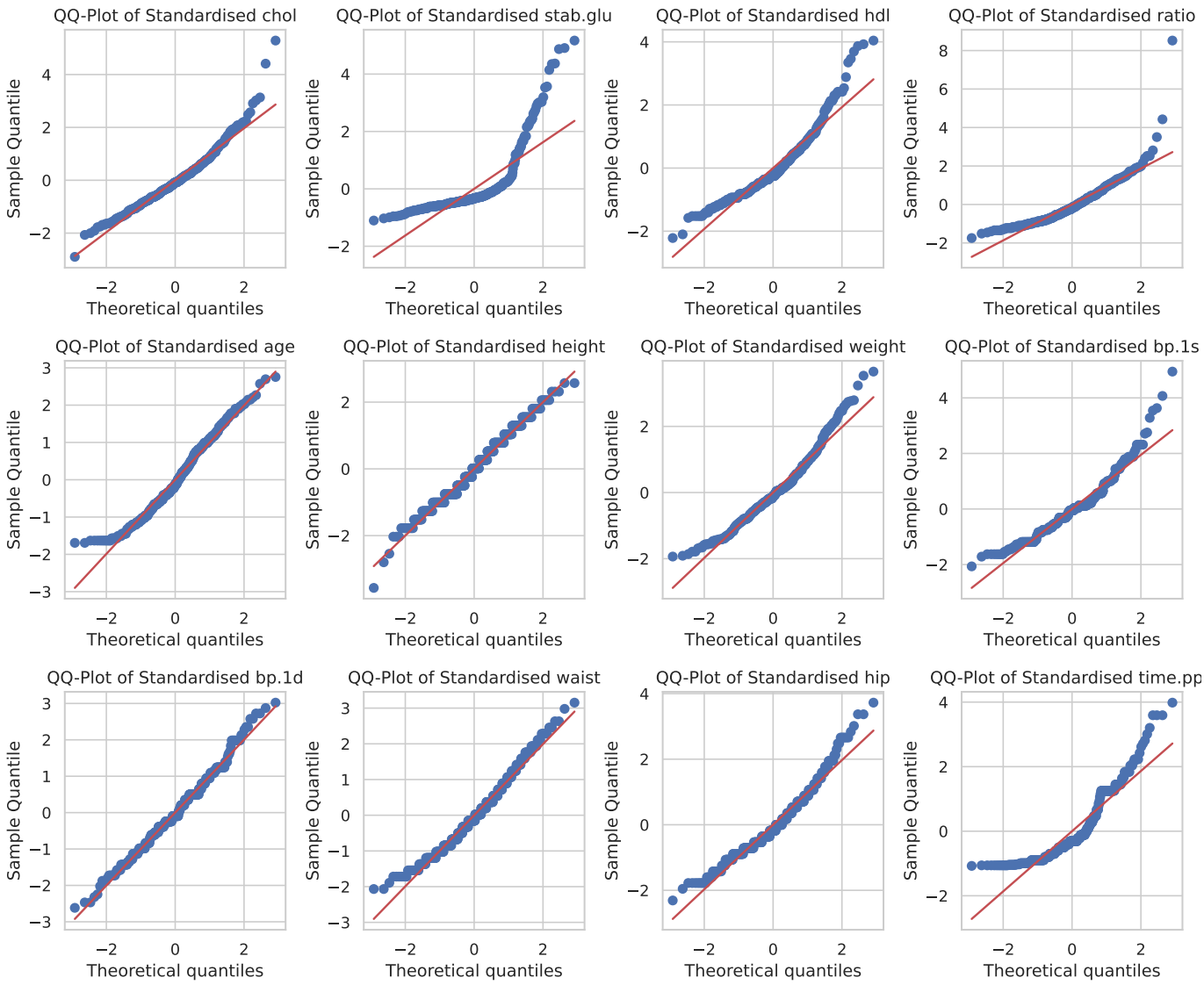

Figure A.5: Quantile-Quantile Plots of Standardized Continuous Variables with reference to $N(0,1)$ Distribution

A.6 Bar Charts

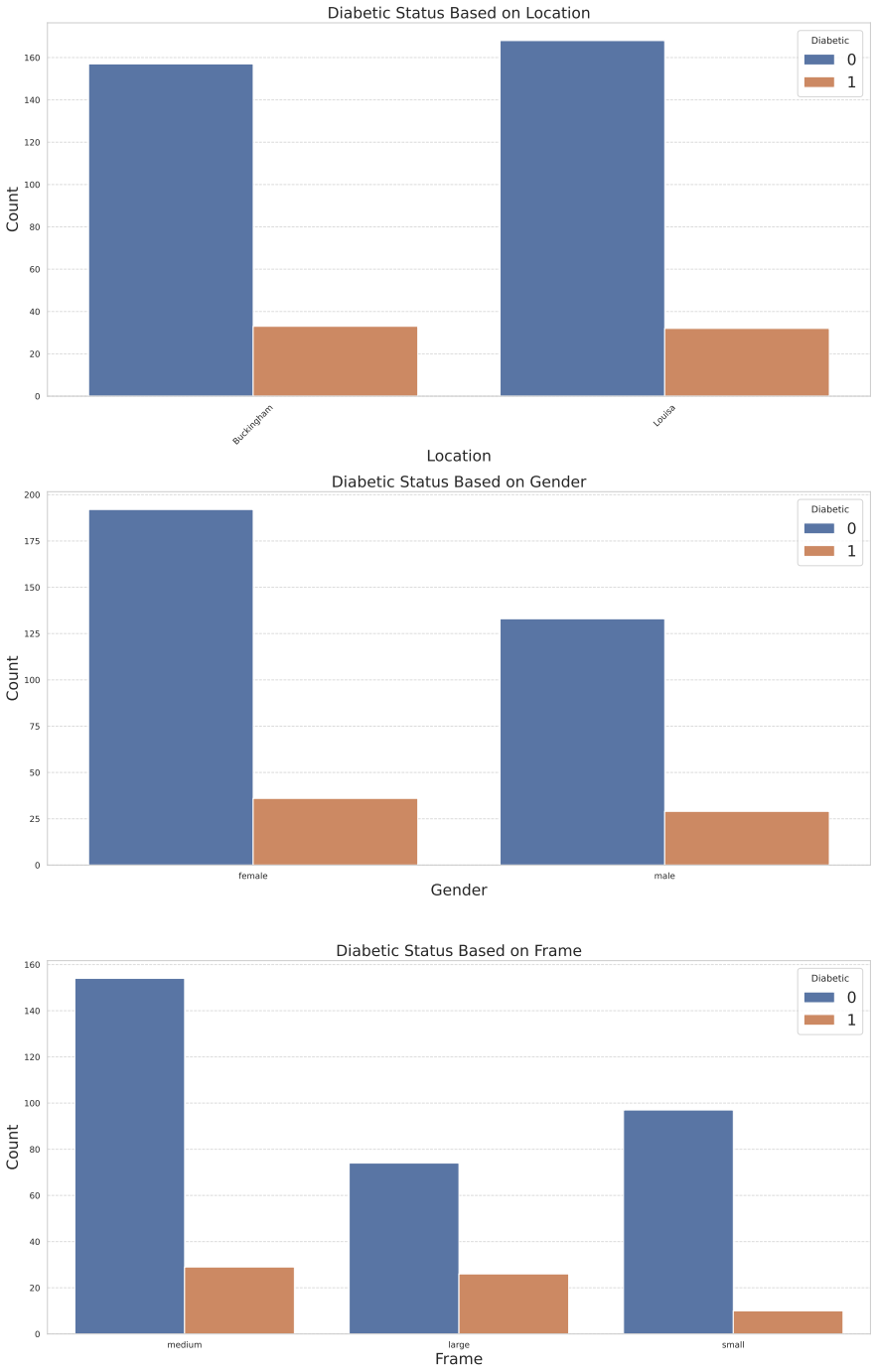

Figure A.6: Bar Charts of Standardised Categorical Variables according to Diabetic Status

Appendix B

Appendix

B.1 Overall Continuous Numerical Summaries

Table B.1: Overall Continuous Numerical Summaries

	chol	*stab.glu*	*hdl*	*ratio*	*age*	*height*
count	390.000000	390.000000	390.000000	390.000000	390.000000	390.000000
mean	207.253846	107.338462	50.258974	4.525385	46.774359	65.946154
std	44.659404	53.798188	17.279856	1.736378	16.435911	3.914862
min	78.000000	48.000000	12.000000	1.500000	19.000000	52.000000
25%	179.000000	81.000000	38.000000	3.200000	34.000000	63.000000
50%	203.000000	90.000000	46.000000	4.200000	44.500000	66.000000
75%	229.000000	107.750000	59.000000	5.400000	60.000000	69.000000
max	443.000000	385.000000	120.000000	19.299999	92.000000	76.000000
skewness	0.960777	2.711121	1.230114	2.245135	0.332907	0.040574

	weight	*bp.1s*	*bp.1d*	*waist*	*hip*	*time.ppn*
count	390.000000	390.000000	390.000000	390.000000	390.000000	390.000000
mean	177.231599	137.133333	83.269231	37.876134	43.013662	335.384615
std	40.458726	22.849517	13.495570	5.755213	5.641488	307.825704
min	99.000000	90.000000	48.000000	26.000000	30.000000	5.000000
25%	150.000000	122.000000	75.000000	33.000000	39.000000	90.000000
50%	172.500000	136.000000	82.000000	37.000000	42.000000	240.000000
75%	199.750000	148.000000	90.000000	41.000000	46.000000	480.000000
max	325.000000	250.000000	124.000000	56.000000	64.000000	1560.000000
skewness	0.746538	1.100286	0.250472	0.475642	0.80637	1.283503

B.2 Diabetic based Proportions on Categorical Variables

Table B.2: Proportions of Diabetic Patients based on Categorical Variables

location	*Diabetic*	
	0	**1**
Buckingham	0.826316	0.173684
Louisa	0.840000	0.160000

frame	*Diabetic*	
	0	**1**
large	0.740000	0.260000
medium	0.841530	0.158470
small	0.906542	0.093458

gender	*Diabetic*	
	0	**1**
female	0.842105	0.157895
male	0.820988	0.179012

B.3 Diabetic based Measures of Central Tendency

Table B.3: Mean and Median of Continuous Variables according to Diabetic Status

Variables	Mean	
diabetic	**0**	**1**
chol	202.941538	228.815385
stab.glu	90.889231	189.584615
hdl	51.212308	45.492308
ratio	4.303385	5.635385
age	44.443077	58.430769
height	65.916923	66.092308
weight	174.566154	190.558825
bp.1s	134.920000	148.200000
bp.1d	82.972308	84.753846
waist	37.297515	40.769231
hip	42.659471	44.784615
time.ppn	330.000000	362.307692

Variables	Median	
diabetic	**0**	**1**
chol	199.0	219.0
stab.glu	86.0	184.0
hdl	47.0	41.0
ratio	4.1	5.3
age	41.0	59.0
height	66.0	67.0
weight	170.0	183.0
bp.1s	132.0	146.0
bp.1d	82.0	86.0
waist	37.0	40.0
hip	42.0	44.0
time.ppn	240.0	240.0

B.4 Diabetic based Measures of Dispersion

Table B.4: Variance and Standard Deviation of Continuous Variables according to Diabetic Status

Variables	Variance	
diabetic	**0**	**1**
chol	1650.407066	3200.809135
stab.glu	615.246952	6232.746635
hdl	287.031947	334.097596
ratio	2.027427	6.560134
age	259.049991	164.905288
height	15.638139	13.960096
weight	1607.499468	1594.900231
bp.1s	515.882469	412.475000
bp.1d	183.866515	173.500962
waist	31.648899	30.899038
hip	31.922374	28.015385
time.ppn	91360.956790	112544.591346

Variables	Standard Deviation	
diabetic	**0**	**1**
chol	40.625202	56.575694
stab.glu	24.804172	78.947746
hdl	16.942017	18.278337
ratio	1.423877	2.561276
age	16.095030	12.841545
height	3.954509	3.736321
weight	40.093634	39.936202
bp.1s	22.713046	20.309481
bp.1d	13.559739	13.171976
waist	5.625735	5.558690
hip	5.649989	5.292956
time.ppn	302.259751	335.476663

Bibliography

[1] American Diabetes Association (2018). 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2018*. *Diabetes Care*, 41(Supplement 1):S13–S27.

[2] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (2018). *Graphical Methods for Data Analysis*. Chapman and Hall/CRC.

[3] Ciarambino, T., Crispino, P., Leto, G., Mastrolorenzo, E., Para, O., and Giordano, M. (2022). Influence of gender in diabetes mellitus and its complication. *Int. J. Mol. Sci.*, 23(16):8850.

[4] Day, S., Wu, W., Mason, R., and Rochon, P. A. (2019). Measuring the data gap: inclusion of sex and gender reporting in diabetes research. *Research Integrity and Peer Review*, 4(1).

[5] Deshpande, A. D., Harris-Hayes, M., and Schootman, M. (2008). Epidemiology of Diabetes and Diabetes-Related Complications. *Phys. Ther.*, 88(11):1254–1264.

[6] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021). A survey on missing data in machine learning. *J. Big Data*, 8(1):140.

[7] Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., and Moustakas, K. (2021). Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction. *IEEE Access*, PP:1–1.

[8] Foley, K., McNaughton, D., and Ward, P. (2020). Monitoring the 'diabetes epidemic': A framing analysis of United Kingdom print news 1993-2013. *PLoS One*, 15(1):e0225794.

[9] Galaviz, K. I., Narayan, K. M. V., Lobelo, F., and Weber, M. B. (2018). Lifestyle and the Prevention of Type 2 Diabetes: A Status Report. *Am. J. Lifestyle Med.*, 12(1):4–20.

[10] Graham, J. W. (2012). *Missing Data: Analysis and Design*. Statistics for Social and Behavioral Sciences. Springer Nature, New York, NY, 1 edition.

[11] Guthrie, W. F. (2020). NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151).

[12] Kang, H. (2013). The prevention and handling of the missing data. *Korean J. Anesthesiol.*, 64(5):402–406.

[13] Kim, Y. G., Jeong, J. H., Han, K.-D., Roh, S.-Y., Min, K., Lee, H. S., Choi, Y. Y., Shim, J., Choi, J.-I., and Kim, Y.-H. (2023). Association between low-density lipoprotein cholesterol and sudden cardiac arrest in people with diabetes mellitus. *Cardiovasc. Diabetol.*, 22(1):36.

[14] Klein, S., Gastaldelli, A., Yki-Järvinen, H., and Scherer, P. E. (2022). Why does obesity cause diabetes? *Cell Metabolism*, 34(1):11–20.

[15] Lindström, J. and Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes Care*, 26(3):725–731.

[16] Little, R. and Rubin, D. (1987). *Statistical Analysis With Missing Data*. Wiley Series in Probability and Statistics. Wiley.

[17] Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248.

[18] Mujumdar, A. and Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, 165:292–299. 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019.

[19] Naidoo, S. (2017). Prevention is better than cure! *Southern African Journal of Infectious Diseases*, 32:1.

[20] Nanayakkara, N., Ranasinha, S., Gadowski, A., Heritier, S., Flack, J. R., Wischer, N., Wong, J., and Zoungas, S. (2018). Age, age at diagnosis and diabetes duration are all associated with vascular complications in type 2 diabetes. *J. Diabetes Complications*, 32(3):279–290.

[21] Rosenzweig, J. L., Conlin, P. R., Gonzalvo, J. D., Kutler, S. B., Maruthur, N. M., Solis, P., Vijan, S., Wallia, A., and Wright, R. F. (2019). 2019 Endocrine Society Measures Set for Older Adults With Type 2 Diabetes Who Are at Risk for Hypoglycemia. *The Journal of Clinical Endocrinology & Metabolism*, 105(4):969–990.

[22] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

[23] Schorling, D. J. (1997). Diabetes Dataset. Data retrieved from Vanderbilt University Department of Biostatistics, `http://hbiostat.org/data`.

[24] Shettigar, L., Sivaraman, S., Rao, R., Arun, S. A., Chopra, A., Kamath, S. U., and Rana, R. (2023). Correlational analysis between salivary and blood glucose levels in individuals with and without diabetes mellitus: a cross-sectional study. *Acta Odontologica Scandinavica*, 0(0):1–11. PMID: 37823574.

[25] Svane, J., Lynge, T. H., Pedersen-Bjergaard, U., Jespersen, T., Gislason, G. H., Risgaard, B., Winkel, B. G., and Tfelt-Hansen, J. (2021). Cause-specific mortality in children and young adults with diabetes mellitus: A danish nationwide cohort study. *Eur. J. Prev. Cardiol.*, 28(2):159–165.

[26] Zoungas, S., Woodward, M., Li, Q., Cooper, M. E., Hamet, P., Harrap, S., Heller, S., Marre, M., Patel, A., Poulter, N., Williams, B., Chalmers, J., and ADVANCE Collaborative group (2014). Impact of age, age at diagnosis and duration of diabetes on the risk of macrovascular and microvascular complications and death in type 2 diabetes. *Diabetologia*, 57(12):2465–2474.