ANALYSING THE FACTORS INFLUENCING DIABETES

Deborupa Pal

Aim:

This project aims to identify the most significant predictors of diabetes status using a dataset containing various patient laboratory results -related variables.

Research Question:

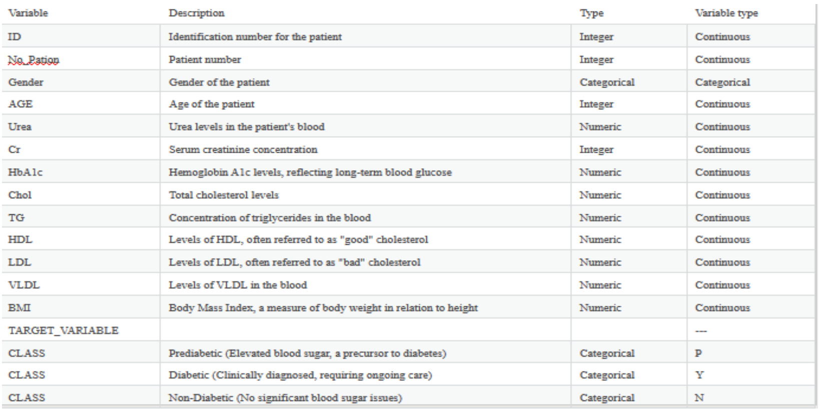
Which factors among Age, Urea, Creatinine (Cr), HbA1c, Cholesterol (Chol), Triglycerides (TG), High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Very Low-Density Lipoprotein (VLDL), and Body Mass Index (BMI) are the most influential in identifying diabetes status (prediabetic, diabetic, or non-diabetic) among patients.

Null and Alternate Hypothesis:

- Null Hypothesis (H_0): The combination of Age, HbA1c, Cholesterol, LDL, BMI, and Gender has no significant effect on predicting diabetes status.
- Alternative Hypothesis (H_a): The combination of Age, HbA1c, Cholesterol, LDL, BMI, and Gender is a significant predictor of diabetes status.

Description of the Dataset:

The Diabetes dataset represents a source of medical and health-related information sourced from 1000 patients in the Iraqi society. This dataset was compiled by gathering data from both the Medical City Hospital and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. The dataset is a comprehensive collection of patient files, consisting of various aspects of laboratory analysis.



Variable	Description	Type	Variable type
ID	Identification number for the patient	Integer	Continuous
No_Pation	Patient number	Integer	Continuous
Gender	Gender of the patient	Categorical	Categorical
AGE	Age of the patient	Integer	Continuous
Urea	Urea levels in the patient's blood	Numeric	Continuous
Cr	Serum creatinine concentration	Integer	Continuous
HbA1c	Hemoglobin A1c levels, reflecting long-term blood glucose	Numeric	Continuous
Chol	Total cholesterol levels	Numeric	Continuous
TG	Concentration of triglycerides in the blood	Numeric	Continuous
HDL	Levels of HDL, often referred to as "good" cholesterol	Numeric	Continuous
LDL	Levels of LDL, often referred to as "bad" cholesterol	Numeric	Continuous
VLDL	Levels of VLDL in the blood	Numeric	Continuous
BMI	Body Mass Index, a measure of body weight in relation to height	Numeric	Continuous
TARGET_VARIABLE			---
CLASS	Prediabetic (Elevated blood sugar, a precursor to diabetes)	Categorical	P
CLASS	Diabetic (Clinically diagnosed, requiring ongoing care)	Categorical	Y
CLASS	Non-Diabetic (No significant blood sugar issues)	Categorical	N

FIG: Description of the dataset

The dataset comprises various patient variables and their corresponding types, including Age (continuous), Urea (continuous), Serum Creatinine (continuous), Hemoglobin A1c (continuous), Total Cholesterol (continuous), Triglycerides (continuous), High-Density Lipoprotein (HDL - continuous), Low-Density Lipoprotein (LDL - continuous), Very Low-Density Lipoprotein (VLDL - continuous), Body Mass Index (BMI - continuous), and Gender (categorical). The response variable, "class," represents Prediabetic (elevated blood sugar), Diabetic (clinically diagnosed, requiring ongoing care), and Non-Diabetic (no significant blood sugar issues), all falling under the categorical type.

Methodology:

Data Preparation- In the initial phase, we focused on setting up our environment and loading essential libraries. The tidyverse library was loaded for data manipulation and visualization tools. Subsequently, we used the read.csv function to import our dataset into RStudio.

Data preprocessing – Our approach to data cleaning began with a thorough examination. Initially, we checked the dataset for missing values, ensuring that each variable had complete information. Following this, we identified outliers, although our dataset didn't exhibit any missing values and duplicate values, outliers were detected in all variables. We detected and removed outliers in continuous variables using Z-scores and the interquartile range (IQR) method, notably addressing outliers in Age (98 instances) and Creatinine (52 instances).

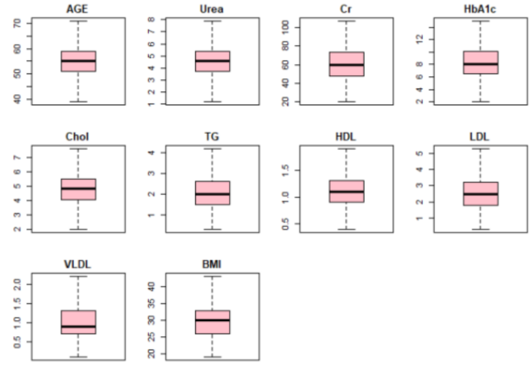


FIG: After outlier removal

Exploratory Data Analysis (EDA) -Before Outlier Removal the Summary statistics revealed diverse ranges in patient variables, such as "No_Pation," "Gender," and medical measurements like "Age," "Urea," "Cr," "HbA1c," "Chol," "TG," "HDL," "LDL," "VLDL," and "BMI." Patient ages, ranging from 20 to 79, showed a roughly symmetric distribution with a mean of 53.53. Post-outlier removal, the dataset exhibited refined central tendencies. Patient ages, with a mean of 53.53, displayed a symmetric distribution. Other variables, including Urea, HbA1c, Cholesterol, Triglycerides, and BMI, were within expected ranges. The "CLASS" variable lacked detailed distribution information.

Data Encoding- We encountered two categorical variables in our dataset: "Gender" and "Class." For conducting statistical analyses, we encoded these categorical values into numerical representations.

Data visualization:

Correlation matrix and pairplot of all the variables

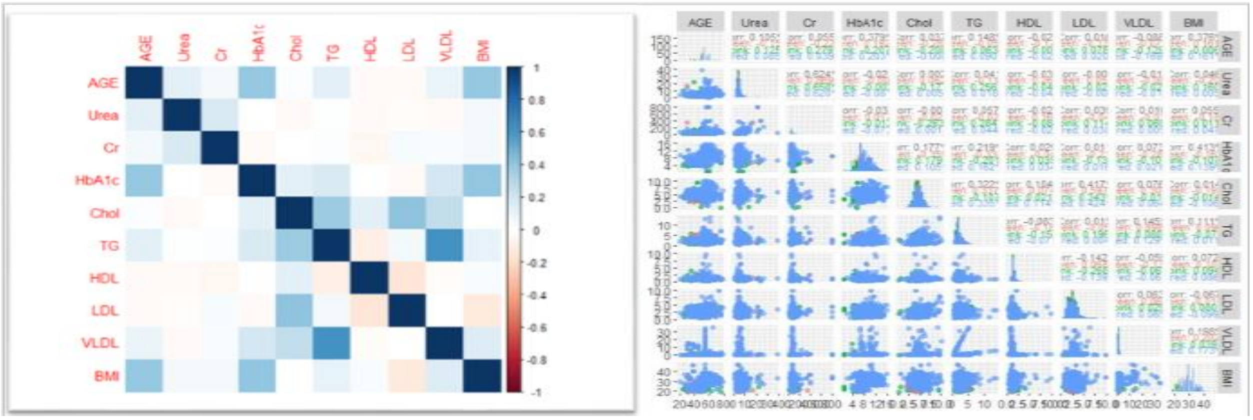


FIG: CORRELATION HEATMAP AND PAIRPLOT OF ALL THE VARIABLES

The heatmap and pairplot analyses were conducted to explore the relationships among each variable. The results indicate significant correlations as BMI positively correlates with AGE, HbA1c, TG, LDL, and VLDL, suggesting potential associations between body mass index and these factors. AGE and HbA1c exhibit a moderate positive relationship, indicating a correlation between age and glycated hemoglobin levels. Cholesterol (Chol) shows positive correlations with LDL and TG, while HDL negatively correlates with LDL and TG, revealing patterns in lipid profiles. Additionally, Urea and Cr display a moderate positive correlation, suggesting a connection between urea and creatinine levels.

Bivariate analysis KDE plots for all the variables:

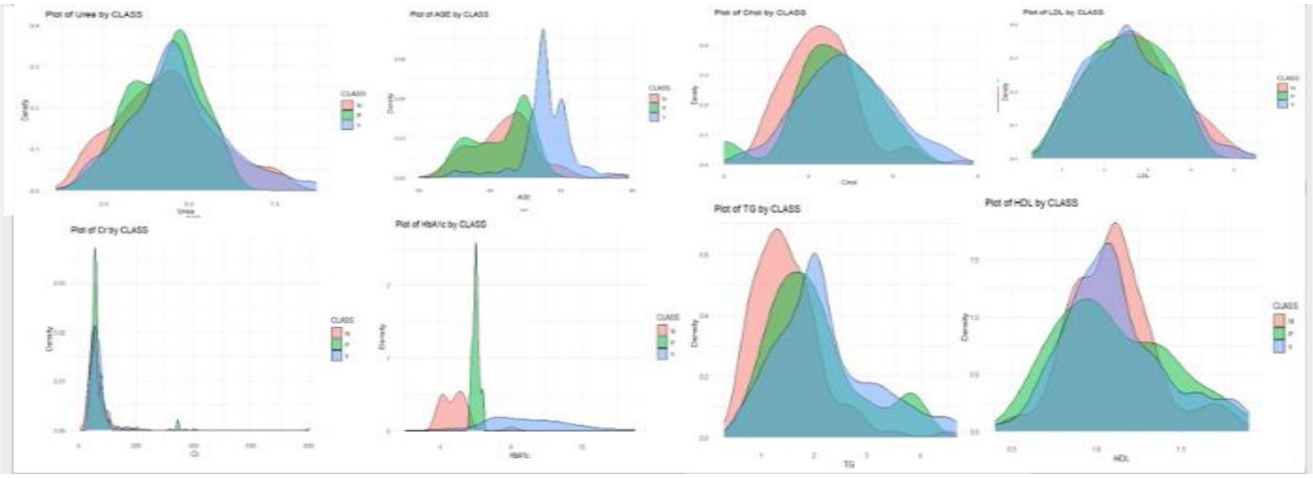


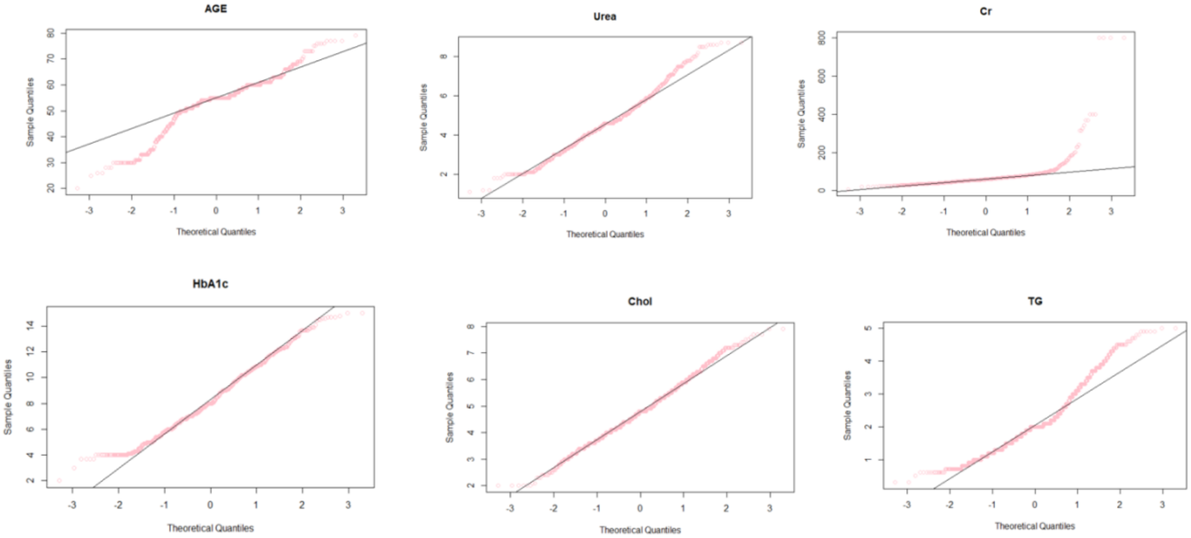
FIG: KDE PLOTS OF ALL VARIABLES

The bivariate analysis KDE plots indicate distinct patterns among variables for different classes. Individuals in CLASS=Y demonstrate higher Urea, Creatinine, cholesterol, triglyceride, and VLDL levels compared to CLASS=P and CLASS=N. While HDL and LDL levels show slight increases in CLASS=Y, individuals with this class also tend to have higher BMI and HbA1c levels. Moreover, AGE is higher in CLASS=Y compared to CLASS=P and CLASS=N

Statistical testing:

Shapiro test: For accessing the Normality: We conducted a normality test using the Shapiro-Wilk test for the variables AGE, Urea, Cr, HbA1c, Chol, TG, HDL, VLDL, and BMI. The results revealed significant deviations from normal distribution (p-values: AGE [1.44E-22], Urea [1.57E-09], Cr [7.92E-50], HbA1c [4.87E-08], Chol [0.00147283], TG [4.37E-19], HDL [1.49E-16], VLDL [1.50E-19], BMI [8.17E-09]).

QQ plots: Confirming these findings, QQ plots displayed discrepancies from the reference line, indicating non-normality. Due to the non-normal distribution and three variables in the target (CLASS N,Y,P), we opted for non-parametric testing and selected the Kruskal-Wallis test for further analysis.



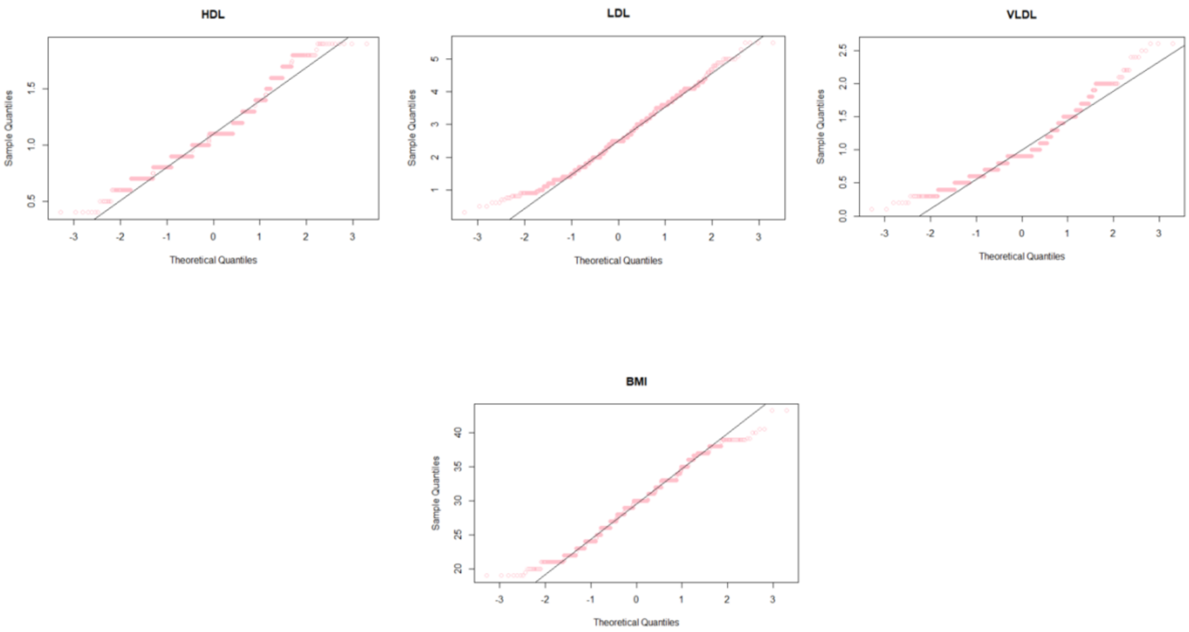


FIG:Q-Q PLOTS FOR ALL THE VARIABLES

Non parametric testing- Kruskal-walli's test: The Kruskal-Wallis test reveals significant differences in Age, HbA1c, and BMI across the three diabetes status groups. Additionally, Cholesterol and Triglycerides exhibit significant differences between prediabetic and non-diabetic groups (p-values: Age [all pairwise < 2.2e-16], HbA1c [all pairwise < 2.2e-16], BMI [all pairwise < 2.2e-16], Chol [prediabetic vs. non-diabetic p = 0.032, prediabetic vs. diabetic p = 0.439], TG [prediabetic vs. non-diabetic p = 0.0002, prediabetic vs. diabetic p = 0.3062]). Conversely, Urea, Creatinine, HDL, and LDL do not display significant differences across the groups (p-values: Urea [p = 0.07135], Cr [p = 0.3137], HDL [p = 0.9212], LDL [p = 0.7547]).

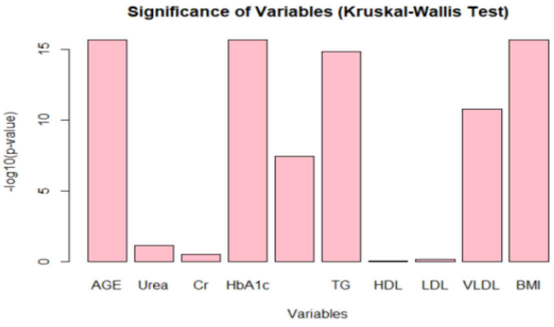


FIG: Bar graph showing significance of variables

In addition to the statistical analyses, visual exploration through bar plots was done. These visualizations supported the earlier statistical findings, emphasizing notable differences in age, HbA1c, BMI, cholesterol, triglycerides, and VLDL among different classes. Conversely, variables such as urea, Cr, HDL, and LDL did not display substantial differences visually.

Chi-square test –for categorical variable: The Pearson's Chi-squared test was conducted to assess the association between "Gender" and "CLASS" in the dataset. The results revealed a statistically significant association (X-squared = 18.202, df = 2, p-value = 0.0001116). Cramer's V, a measure of association strength, was calculated as 0.1349132, indicating a modest and relatively weak association between the two variables.

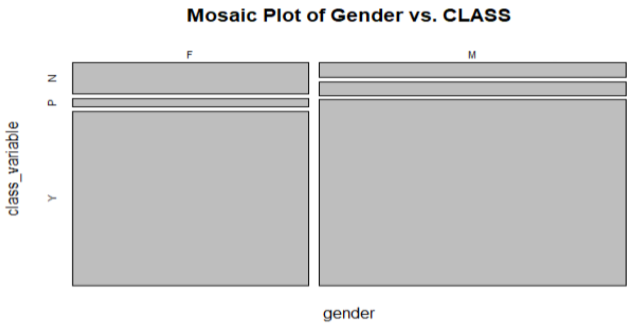


FIG: Mosaic plot of gender vs class

Multiple logistic regression: In the logistic regression results, Age ($p < 0.0001$), HbA1c ($p < 0.0001$), Cholesterol (p = 6.56e-04), LDL (p = 1.11e-08), and BMI (p = 9.56e-05) are highly significant predictors of diabetes status, whereas other variables do not exhibit statistically significant effects.

Limitations for the statistical methods used: While the statistical methods employed provided information into the dataset, it is essential to acknowledge certain limitations. The Shapiro-Wilk test and QQ plots revealed non-normal distributions for all the variables, prompting the use of non-parametric testing like the Kruskal-Wallis test. However, non-parametric tests have their limitations, such as reduced power compared to parametric tests. Additionally, while statistical tests identified significant differences in various variables across diabetes status groups, correlation does not imply causation. The findings should be interpreted cautiously, considering the observational nature of the study. Furthermore, the Chi-Square test highlighted a significant association between gender and diabetes status, but causal relationships cannot be established based solely on this statistical association. In logistic regression, it's crucial to note that the model assumes linearity, independence of errors, and absence of multicollinearity, assumptions that may impact the reliability of the results.

Summary of overall findings: In the logistic regression analysis, Age ($p < 0.0001$), HbA1c ($p < 0.0001$), Cholesterol (p = 6.56e-04), LDL (p = 1.11e-08), and BMI (p = 9.56e-05) emerged as highly significant predictors of diabetes status. These findings align with the non-parametric testing (Kruskal-Wallis), which identified significant differences in Age, HbA1c, and BMI across diabetes status groups. Additionally, visual exploration through bar plots supported these differences.

Moreover, the Chi-square test revealed a significant association (Cramer's V = 0.1349132) between "Gender" and "CLASS," although the association was relatively weak. The logistic regression further emphasized Creatinine (Cr), High-Density Lipoprotein (HDL), and BMI as highly significant predictors of diabetes status.

Conclusion:

In a thorough investigation utilizing non-parametric tests, chi-squared analyses, logistic regression, and visual scrutiny, a compelling argument surfaces for rejecting the null hypothesis. The interplay among Age, HbA1c, Cholesterol, LDL, BMI, and Gender emerges as a robust and statistically significant predictor of diabetes status within the examined population. Deviation from normal distribution in pivotal variables challenges parametric assumptions, necessitating the reliance on non-parametric methodologies such as the Kruskal-Wallis test, revealing distinct patterns in Age, HbA1c, and BMI across diabetes categories. The Chi-squared test unveils a significant association between "Gender" and "CLASS," supported by logistic regression findings that underscore Age, HbA1c, Cholesterol, LDL, and BMI as potent predictors. Visual examination through bar plots reinforces these distinctions. Furthermore, a secondary logistic regression broadens the predictive scope to encompass Creatinine (Cr), High-Density Lipoprotein (HDL), and BMI. This comprehensive analysis asserts that the amalgamation of these variables significantly shapes diabetes status, offering nuanced insights for personalized interventions and healthcare strategies. The study not only advances comprehension but also underscores practical implications, emphasizing the importance of a holistic approach to diabetes prediction and management.

Appendix:

References:

Official R documentation: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Detailed tutorial with code examples: GeeksforGeeks. "Kruskal-Wallis Test in R Programming." URL: https://www.geeksforgeeks.org/kruskal-wallis-test-in-r-programming/.

Package options for non-parametric tests: Data Analytics. "Non-parametric Tests using R." URL: https://www.dataanalytics.org.uk/non-parametric-tests-using-r/.

Basic Chi-square test: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Chi-square test for contingency tables: DataFlair. "Chi-square Test in R." URL: https://data-flair.training/blogs/chi-square-test-in-r/.

Packages for advanced chi-square tests: YouTube. "Advanced Chi-Square Tests in R." URL: https://m.youtube.com/watch?v=ln6WXYn4kw0.

Package for calculating Cramer's V: John Fox (2021). R Commander: A Basic-Statistics GUI for R. URL: https://www.john-fox.ca/RCommander/installation-notes.html.

Tutorial on Cramer's V and other effect size measures: YouTube. "Effect Size Measures in R." URL: https://m.youtube.com/watch?v=AxDpJ6qQVl8.

Basic logistic regression with glm: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

More advanced logistic regression with various packages: Medium. "Logistic Regression using RStudio." URL: https://medium.com/evidentebm/logistic-regression-using-rstudio-336a2b1af354.

Package options for model comparison and diagnostics: RPubs. "Logistic Regression Model Comparison." URL: https://rpubs.com/PandulaP/logisticregression_model_compare.

Base R plotting functions: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/.

Popular plotting packages: ggplot2 and lattice: Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org/; Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer, New York. URL: http://www.sthda.com/english/wiki/lattice-graphs.

Interactive visualization with plotly: Plotly. "Plotly for R." URL: https://plotly.com/r/.

General overview of statistical tests in R: YouTube. "Statistical Tests in R - Overview." URL: https://www.youtube.com/watch?v=_AhyWYz3DAo.

Packages for various statistical tests: John Fox (2021). R Commander: A Basic-Statistics GUI for R. URL: https://www.john-fox.ca/RCommander/installation-notes.html; R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://cran.r-project.org/package=Rcmdr.

Basic Shapiro-Wilk test: Statology. "Shapiro-Wilk Test in R." URL: https://www.statology.org/shapiro-wilk-test-r/.

Package options for various normality tests: R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org/; R-Forge. "R-Forge: R packages and projects." URL: https://r-forge.r-project.org/.

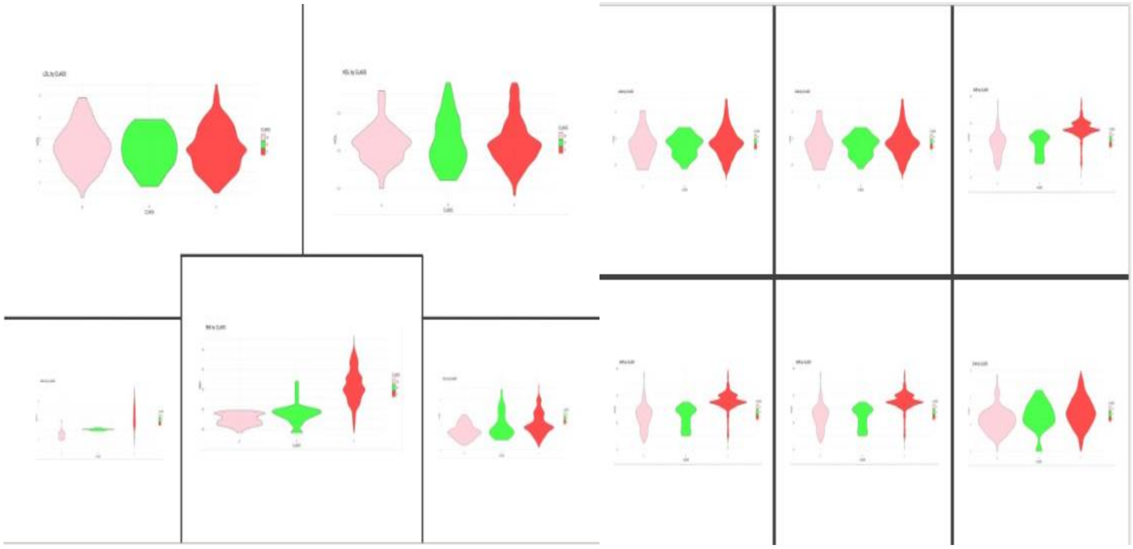


FIG: Violin plots for all the variables

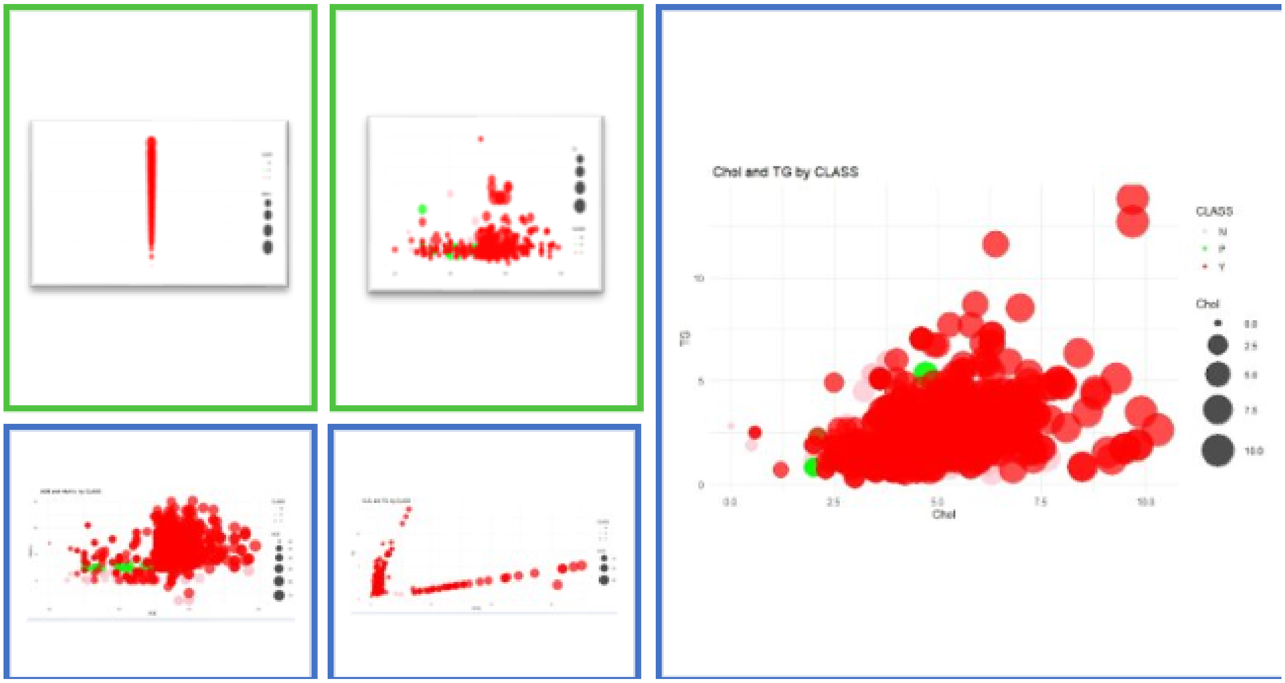


FIG: Bubble chart of HbA1C, AGE, urea and cr ,Chol and TG, VLDL and TG, Age and HbA1C by class

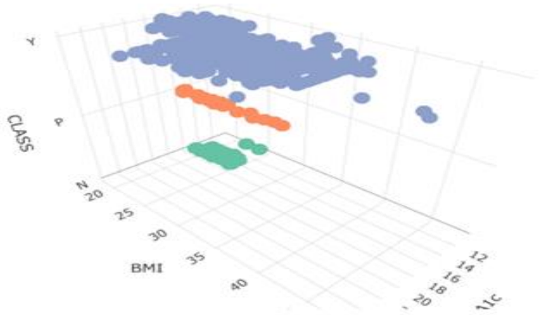


FIG:3D plot of HBA1C AND BMI with CLASS