DIABETES PREDICTION

Non parametric analysis using R

I. **Introduction:**

In the realm of health challenges, diabetes emerges as a pervasive and intricate concern affecting diverse populations globally. This project focuses on the diagnostic prediction of diabetes, utilizing a dataset originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset, carefully selected with specific constraints from a larger database, comprises exclusively of female individuals aged at least 21 years and of Pima Indian heritage. The file format of the dataset encapsulates numerous variables, including independent factors such as various medical predictor variables, and a singular target variable denoted as 'Outcome,' indicating the presence or absence of diabetes. In this project, I will concentrate on using non-parametric statistic to find the best model to predict the "Outcome".

II. .**Dataset overview:**

1. ***Data description:***

Diabetes dataset contains 768 rows and 9 columns. Here is the variable description:

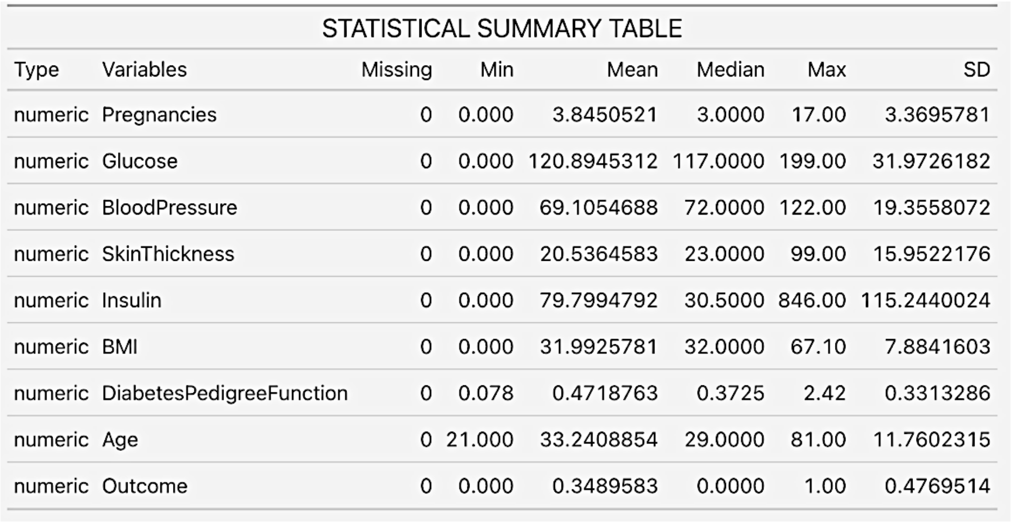
Variable name	Description
Pregnancies	To express the Number of pregnancies
Glucose	To express the Glucose level in blood
BloodPressure	To express the Blood pressure measurement
SkinThickness	To express the thickness of the skin
Insulin	To express the Insulin level in blood
BMI	To express the Body mass index
DiabetesPedigreeFunction	To express the Diabetes percentage
Age	To express the age
Outcome	To express the final result 1 is Yes and 0 is No

Table 1 Variables description

In this analysis, the "Outcome" is used as a response variable, and the other are used as predictor variables.

2. *Exploratory Data Analysis (EDA):*

a. Summary Statistic:



STATISTICAL SUMMARY TABLE

Type	Variables	Missing	Min	Mean	Median	Max	SD
numeric	Pregnancies	0	0.000	3.8450521	3.0000	17.00	3.3695781
numeric	Glucose	0	0.000	120.8945312	117.0000	199.00	31.9726182
numeric	BloodPressure	0	0.000	69.1054688	72.0000	122.00	19.3558072
numeric	SkinThickness	0	0.000	20.5364583	23.0000	99.00	15.9522176
numeric	Insulin	0	0.000	79.7994792	30.5000	846.00	115.2440024
numeric	BMI	0	0.000	31.9925781	32.0000	67.10	7.8841603
numeric	DiabetesPedigreeFunction	0	0.078	0.4718763	0.3725	2.42	0.3313286
numeric	Age	0	21.000	33.2408854	29.0000	81.00	11.7602315
numeric	Outcome	0	0.000	0.3489583	0.0000	1.00	0.4769514

Table 2 Summary Statistic

According to the statistical summary table, there is no missing data in this dataset, and all the numeric variables are positive number.

b. Outcome distribution:

The histogram provides a visual representation of the response variable in the dataset, revealing an apparent imbalance in the prevalence of diabetes. Among the 768 patients included in the study, approximately 268 patients exhibit diabetes, while the remaining 500 patients are non-diabetic. The non-diabetic patients are almost twice as the diabetic one.

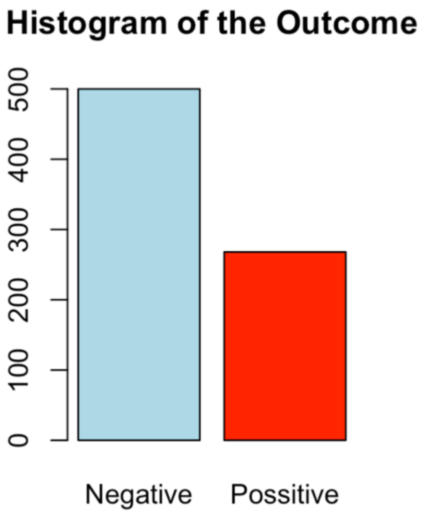
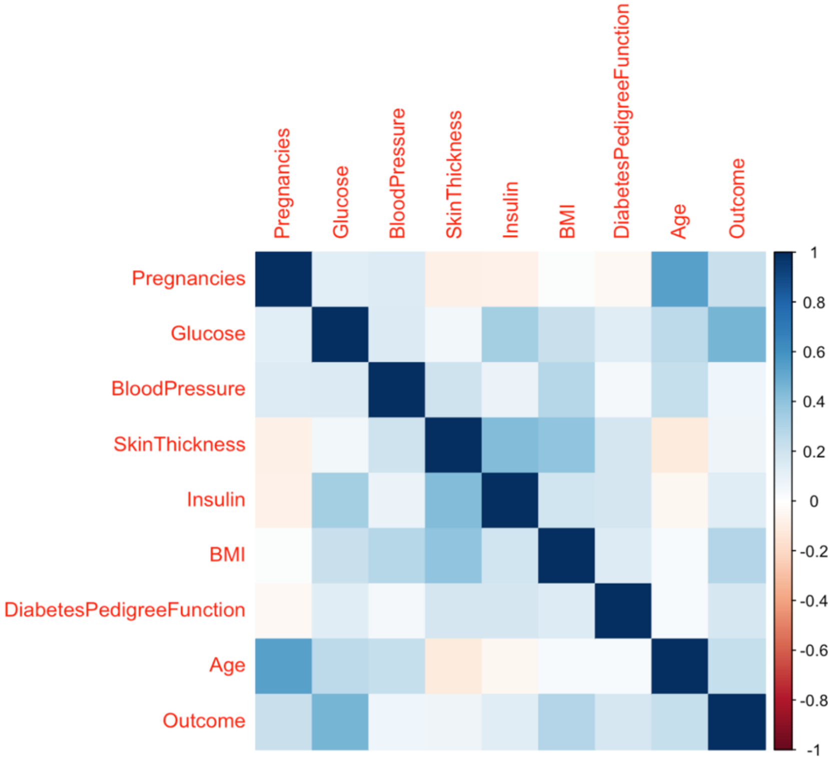


Figure 1 Outcome distribution

c. Correlation Analysis:



HEAT MAP OF CORRELATION

Figure 2 Correlation Heatmap

The heatmap, depicting the correlation among variables, highlights key insights into the factors influencing the outcome variable. Notably, the top five variables demonstrating the highest correlations are Glucose, BMI, Age, Pregnancies, and DiabetesPredigreeFunction (DPF). The pronounced correlations suggest a strong relationship between these factors and the outcome variable. This identification of influential variables provides valuable guidance for further exploration and emphasizes the potential significance of Glucose, BMI, Age, Pregnancies, and DPF in understanding and predicting the outcome variable in this analysis.

d. Variables distribution bases on “Outcome”:

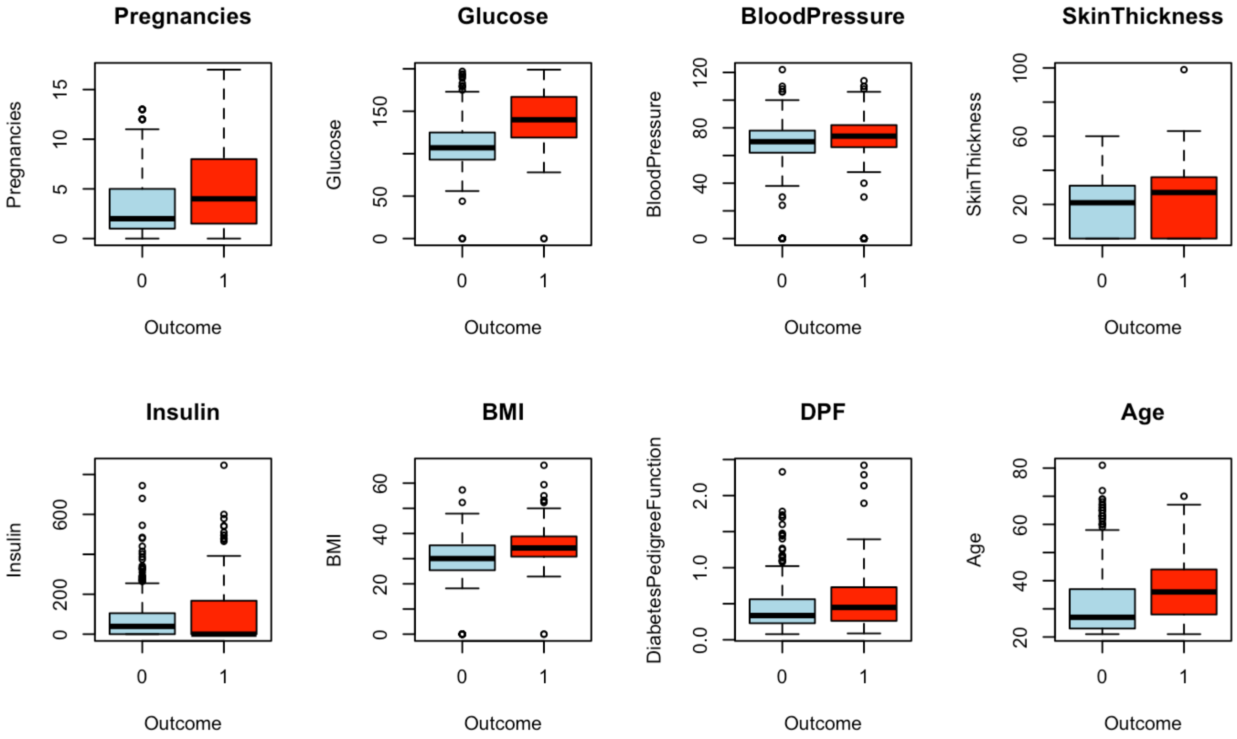


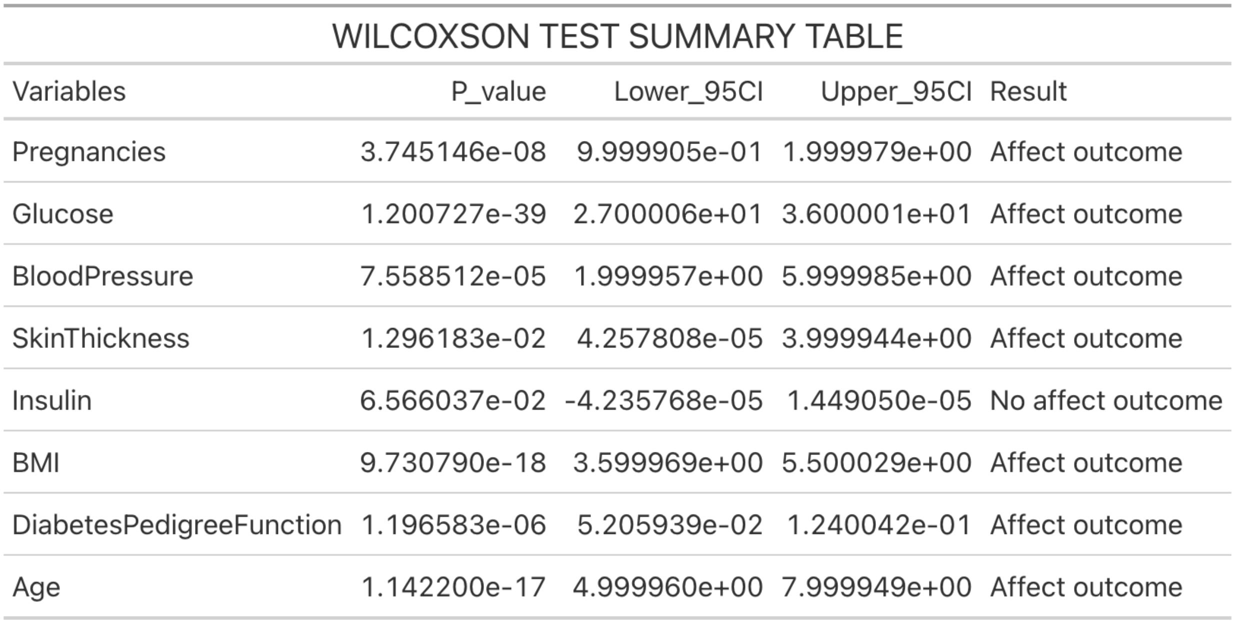
Figure 3 Multiple side-by-side boxplot by "Outcome"

The comprehensive set of box plots, portraying 8 distinct measurements by outcome, offers valuable insights into the distributional characteristics of diabetic and non-diabetic patients. Across most variables, diabetic individuals demonstrate noticeably higher median values, indicating a potential association between these measurements and diabetes status. However, an intriguing exception is observed in the case of Insulin, where the median values for both positive and negative outcomes exhibit a comparable pattern. This finding suggests that, unlike other measurements, Insulin may not be as differentiating a factor between diabetic and non-diabetic patients. These graphical representations enrich our understanding of the dataset, shedding light on the potential significance of each measurement in distinguishing diabetes outcomes.

III. Data Analysis:

1. Variable Selection:

a. Wilcoxon Hypothesis Tests



WILCOXSON TEST SUMMARY TABLE

Variables	P_value	Lower_95CI	Upper_95CI	Result
Pregnancies	3.745146e-08	9.999905e-01	1.999979e+00	Affect outcome
Glucose	1.200727e-39	2.700006e+01	3.600001e+01	Affect outcome
BloodPressure	7.558512e-05	1.999957e+00	5.999985e+00	Affect outcome
SkinThickness	1.296183e-02	4.257808e-05	3.999944e+00	Affect outcome
Insulin	6.566037e-02	-4.235768e-05	1.449050e-05	No affect outcome
BMI	9.730790e-18	3.599969e+00	5.500029e+00	Affect outcome
DiabetesPedigreeFunction	1.196583e-06	5.205939e-02	1.240042e-01	Affect outcome
Age	1.142200e-17	4.999960e+00	7.999949e+00	Affect outcome

Table 3 Wilcoxson test summary

The results from the extensive multiple Wilcoxon rank sum tests across all variables suggest a notable distinction between diabetic and non-diabetic individuals, with nearly all independent variables displaying a significant difference. Notably, Insulin stands out as an exception, exhibiting no clear effect on diabetes outcomes. Consequently, based on this statistical evidence, Insulin has been purposefully excluded from the predictive model. This meticulous variable selection process refines the model, emphasizing the influential variables that robustly differentiate diabetes status in the dataset.

b. Ranking variables (Spearman)

In leveraging the Spearman rank correlation method, the variables in our project exhibit the following order of correlation coefficients in decreasing magnitude: Glucose, BMI, Age, SkinThickness, Pregnancies, DiabetesPedigreeFunction (DPF), BloodPressure, and Insulin. The selection of variables for the final model requires additional in-depth analysis. For a detailed overview of ranking, refer to the accompanying table.

Variables	Correlation coefficients
Outcome	1.00000000
Glucose	0.47577631
BMI	0.30970674
Age	0.30904026
Pregnancies	0.19868875
DiabetesPedigreeFunction	0.17535347
BloodPressure	0.14292068
SkinThickness	0.08972776
Insulin	0.06647165

Table 4 Ranking correlation coefficient

c. Choosing variables:

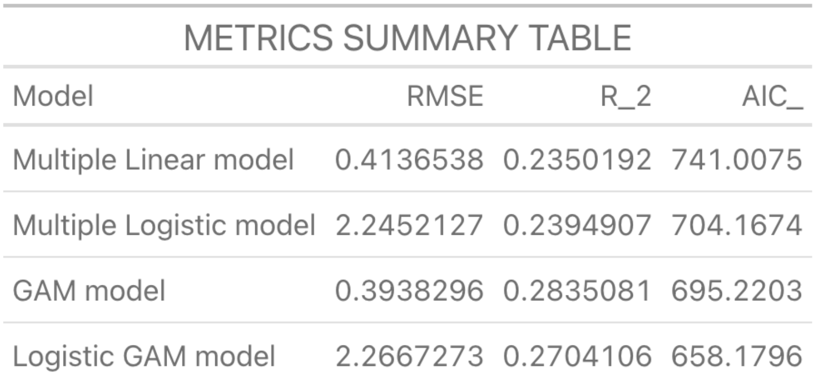
By referencing Table 4 for variable selection in the regression model (rfit), an insightful drop test analysis reveals that the most suitable variables for exclusion are Insulin and SkinThickness. Consequently, the refined set of predictor variables for forecasting the Outcome encompasses Pregnancies, Glucose, BloodPressure, BMI, DPF, and Age, emphasizing a meticulous approach to model refinement based on empirical evidence and statistical assessments.

Drop variables	p_value (from drop.test)
Insulin	0.11797
Insulin, SkinThickness	0.24789
Insulin, SkinThickness, BloodPressure	0.01320
Insulin, SkinThickness, BloodPressure, DPF	0.00017

Table 5 Drop test summary

2. Model selection:

In the process of model selection, I partitioned the dataset into a 95% training set and a 5% test set. Subsequently, I employed various modeling techniques, including multiple linear regression, multiple logistic regression, and Generalized Additive Models (GAM), as well as a Logistic GAM model. These models were trained on the training set and evaluated on the test set, with performance assessed using metrics such as Root Mean Squared Error (RMSE), R-squared (R^2), and the Akaike Information Criterion (AIC). The comprehensive metric summary table indicates that the Generalized Additive Model (GAM) stands out as the most suitable fit for the diabetes dataset.



METRICS SUMMARY TABLE

Model	RMSE	R_2	AIC_
Multiple Linear model	0.4136538	0.2350192	741.0075
Multiple Logistic model	2.2452127	0.2394907	704.1674
GAM model	0.3938296	0.2835081	695.2203
Logistic GAM model	2.2667273	0.2704106	658.1796

Table 6 Metrics summary

IV. Conclusion:

Through a systematic process, our diabetes prediction project began with 8 variables, which were refined using non-parametric methods like Exploratory Data Analysis (EDA), Wilcoxson test, and Spearman for variable selection. Following this process, we identified the optimal set of predictors: Pregnancies, Glucose, BloodPressure, BMI, DiabetesPedigreeFunction (DPF), and Age. Subsequently, various models were assessed for their predictive capability. The model selection journey culminated in the identification of the Logistic Generalized Additive Model (GAM) as the most effective since the outcome is considered as a categorical variable, emphasizing its prowess in capturing the intricate relationships within the diabetes dataset. This conclusion is rooted in a meticulous statistical approach that aligns with the project's overarching goal of accurate and robust diabetes prediction.

References:

Hollander, M., Wolfe, D. A., & Chicken, E. (Year of Publication). Nonparametric Statistical Methods (3rd ed.). Publisher.

Khare, A. D. (2022, October 6). *Diabetes dataset*. Kaggle. https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data