**Clinical Research**

**Risk Factor Analysis for Diabetes: Data Analysis Report**

**1. Introduction**

Diabetes is a major global public health issue, affecting millions of people worldwide. According to the International Diabetes Federation (IDF), approximately 537 million adults (20-79 years) were living with diabetes in 2021, with projections indicating this number could rise to 643 million by 2030 and 783 million by 2045. The disease is responsible for 6.7 million deaths annually, making it one of the leading causes of mortality.

**2. Study Objectives.**

To summarize key demographic and health related variables.

To assess the distribution of demographic and health-related characteristics among individuals with and without diabetes.

To explore associations between diabetes and key risk factors.

To develop a logistic regression model to assess the impact of various predictors on diabetes outcome.

To formulate conclusive insights and data driven recommendations for diabetes prevention based on the analysis

**3. Methodology**

**3.1 Data Collection and Preparation**

We utilized both SPSS and R Programming for the analysis, chosen for their strength in statistical analysis and modelling

The dataset used in this study consists of 718 observations and 9 variables, including demographic details, health conditions, and lifestyle factors. The data was obtained from a publicly available source (Kaggle).

We meticulously examined the data to identify and address any missing values, thereby enhancing the reliability of our analysis.

Duplicate records were identified and removed, ensuring that only unique study entries remained.

We identified categorical variables by systematically examining the unique values for each categorical variable

Categorical variables were converted into factors for statistical analysis.

Feature Engineering. We meticulously created new groups for all numeric variables levels to further our study.

**3.2 Statistical Analysis**

**3.2.1 Univariate Analysis**:

We conducted descriptive statistics for all numeric variables.

Frequency distributions were performed for all variables except id which acted as a unique identifier.

**3.2.2 Bivariate Analysis**:

Correlation Analysis was conducted to assess the relationship among numeric variables.

Chi-square tests and t-tests were conducted to examine associations between diabetes and other potential risk factors.

**3.3.3 Multivariate Analysis**:

Binary Logistic regression model was employed to identify significant predictors of diabetes.

The logistic regression model was assessed using the area under the curve (AUC) from ROC analysis.

Feature importance was evaluated to determine the most influential predictors of diabetes.

**3.3 Recommendations**

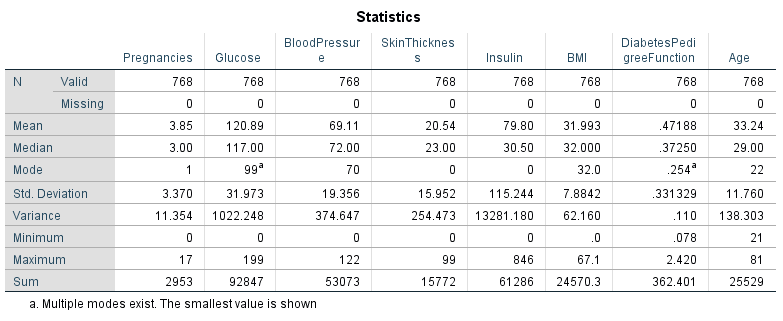
Based on our combined findings from SPSS and R, we provided recommendations for diabetes prevention to guide future actions and decision-making.

**4. Findings**

**4.1 Univariate Analysis**

Univariate analysis involves analyzing one variable at a time. It helps to understand the distribution, central tendency, and spread of individual variables.

**4.1.1 Descriptive Statistics**



The mean number of pregnancies was 3.85 indicating that majority of the participants had at least 4 pregnancies.

The mean blood glucose level was 120.89 indicating that majority of the participants had an abnormal blood glucose level.

The mean blood pressure was 69.11 indicating that majority of the participants had normal blood pressure.

The mean skin thickness was 20.54 while the median was 23.00.

The mean insulin level was 79.80 indicating that the majority of the participants had abnormal blood insulin levels.

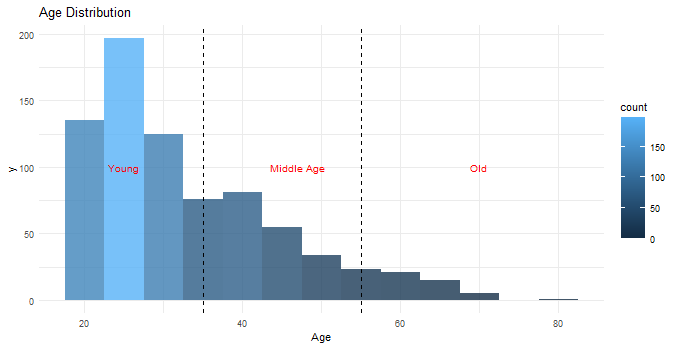
The mean bmi was 31.993 indicating that majority of the participants were obese.

The mean of diabetes pedigree function was 0.47188 while its median was 0.37250.

The mean age was 33.24 indicating that majority of the participants were middle aged.

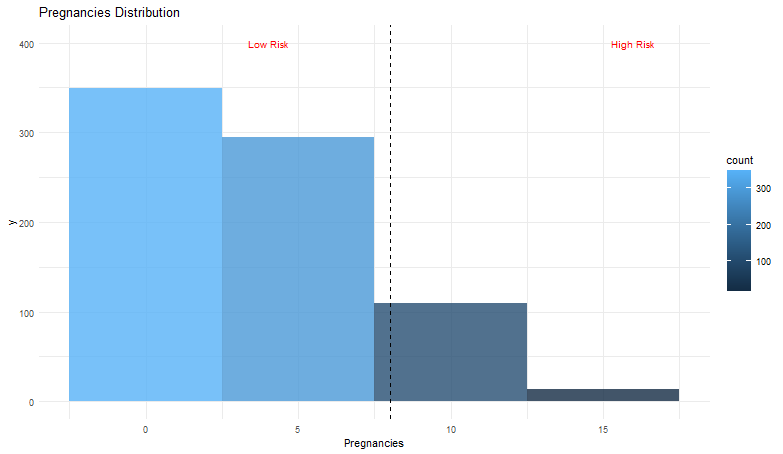
**4.1.2. Frequency Distributions and Visualizations of Key Variables in our study**

**4.1.2.1. Age**

****

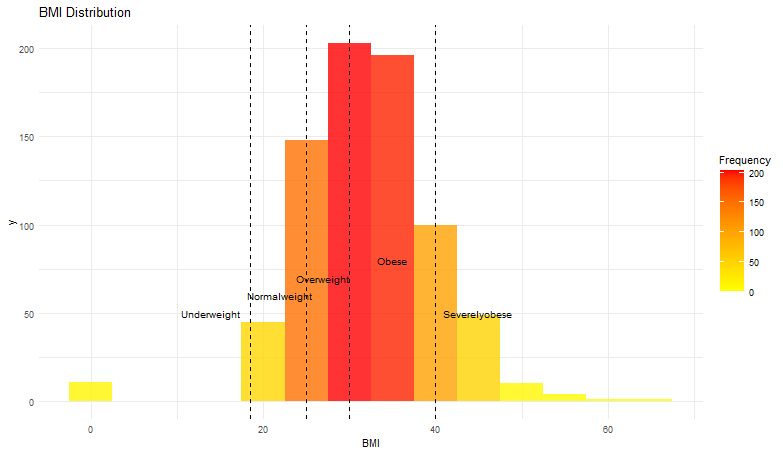
The distribution of age indicates a right scewed distribution.

**4.1.2.2. Number of pregnancies**

****

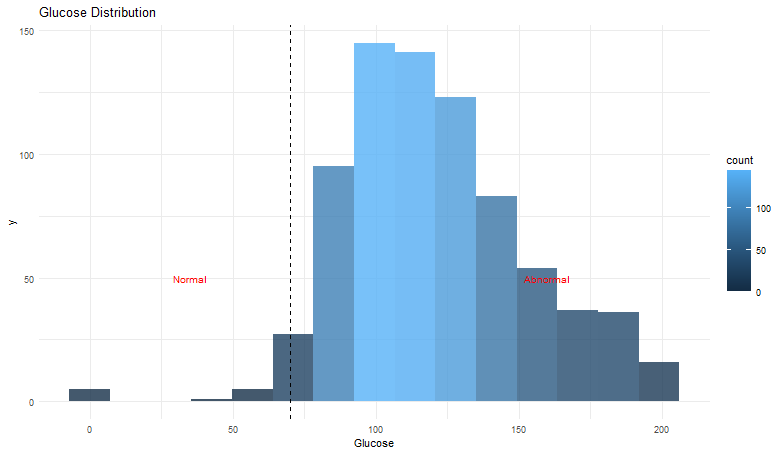
The graph indicates a right scewed distribution.

**4.1.2.3. BMI**

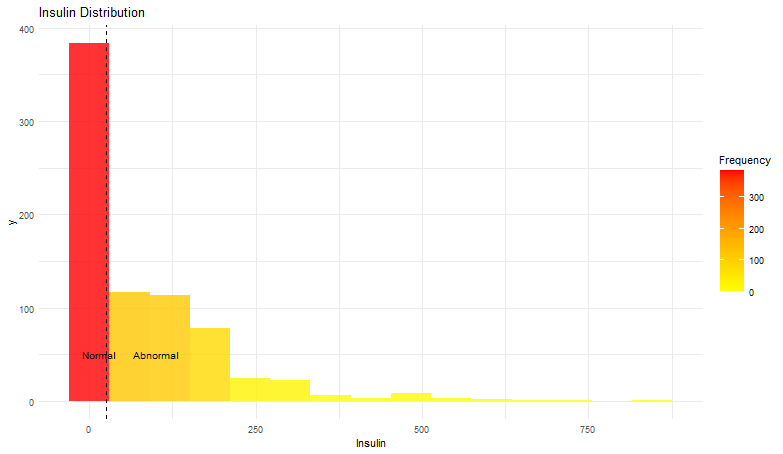
****

The graph indicates normal distribution of bmi.

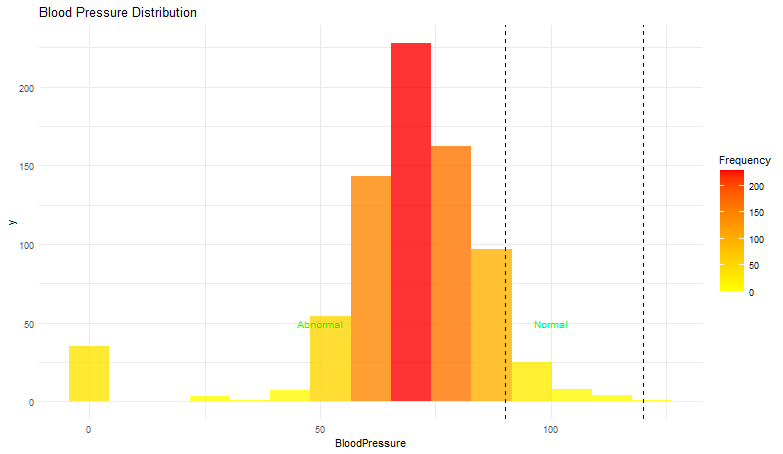
**4.1.2.4. Glucose levels**

****

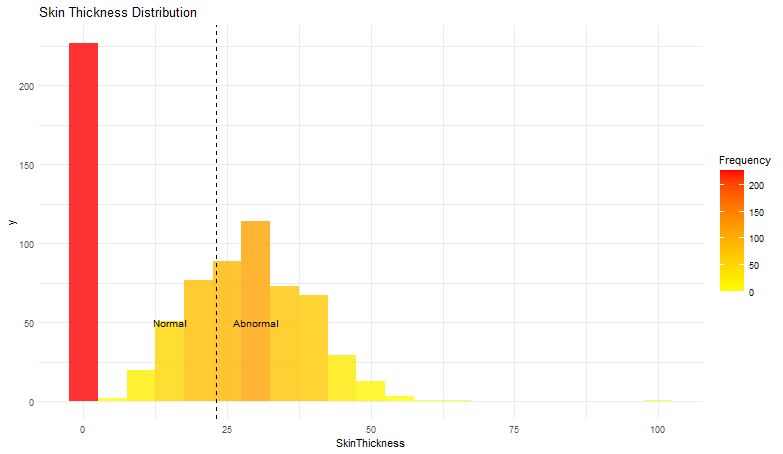
**4.1.2.5. Insulin**

****

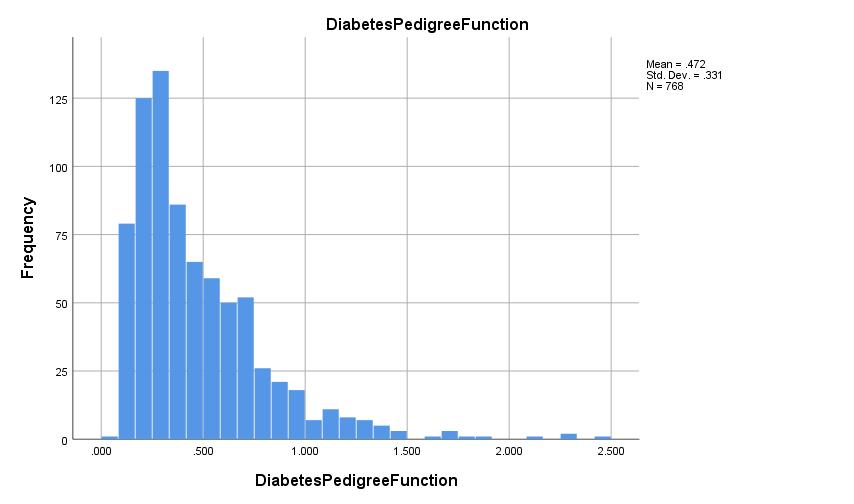
**4.1.2.6. Blood Pressure**

****

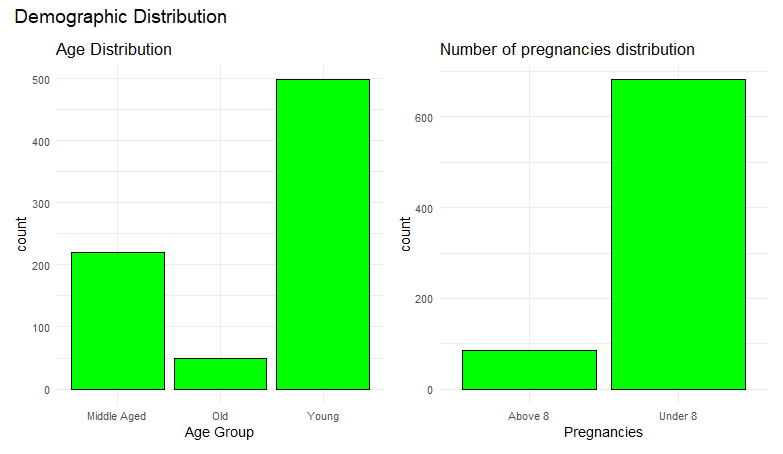
**4.1.2.7. Skin Thickness**

****

**4.1.2.8. Diabetes Pedigree Function**

****

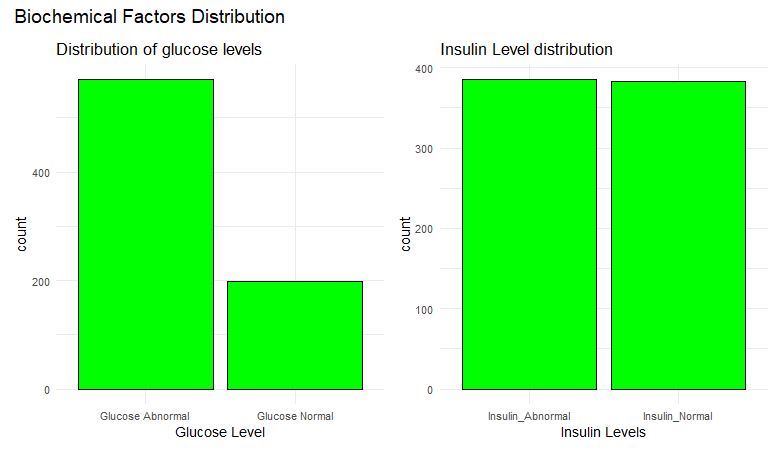
**4.1.2.9. Demographic Distribution**

****

The young participants were the highest followed by middle aged the old.

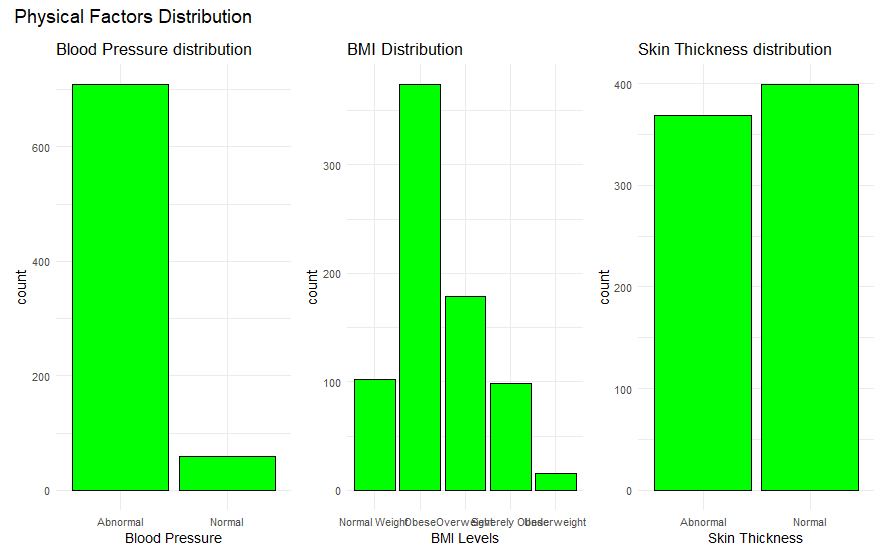
Majority of participants had less than 8 number of pregnancies.

**4.1.3.0. Biochemical Factors Distribution.**

****

Many participants had abnormal blood glucose level compared to those who had normal blood glucose level.

**4.1.3.1. Physical Factors Distribution**

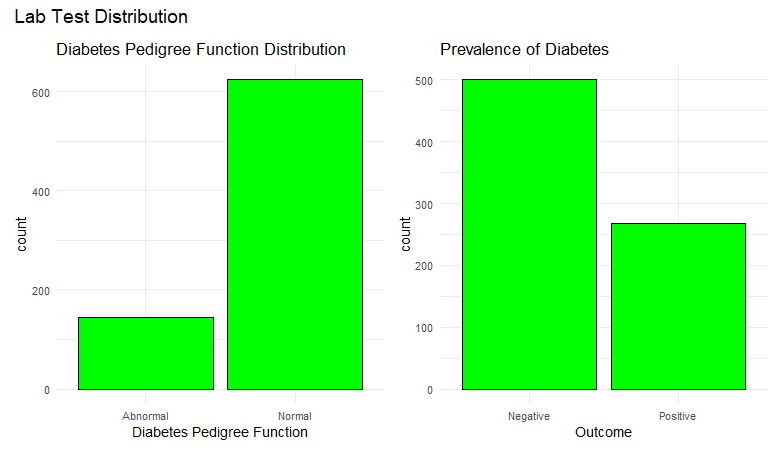
****

Many participants had abnormal blood pressure compared to those who had.

Many participants in the study were obese followed by overweight.

Almost half of the participants had abnormal skin thickness .

**4.1.3.2. Lab Test Distribution**

****

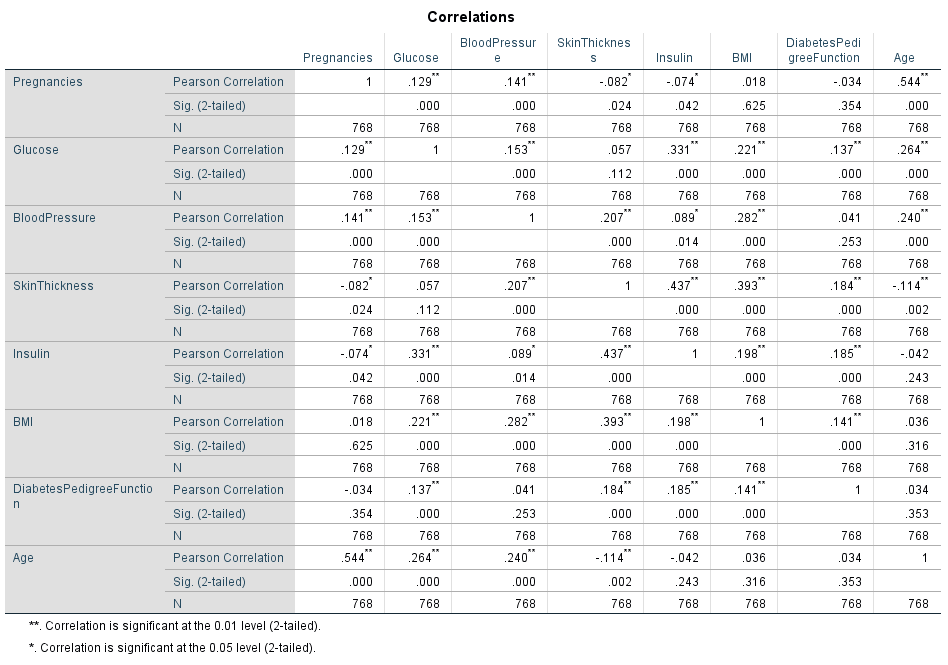
Majority of the participant had a normal diabetes pedigree function.

Negative participants to diabetes were higher compared to those who were positive.

**4.2. BIVARIATE ANALYSIS**

Bivariate analysis is a fundamental statistical technique used to explore and establish relationships between two variables. By examining the associations between different pairs of variables, we gain valuable insights into the interplay and dependencies within our dataset. In this report, we embark on a comprehensive bivariate analysis, focusing on several key pairs of variables crucial to our understanding of Diabetes. Through the application of contingency tables and Chi-square tests, correlations, T-Tests and ANOVA, we aim to unravel significant associations and patterns that can inform decision-making and guide further investigations

**4.2.1. Relationship between continuous variables**



There is a weak positive relationship between pregnancies and glucose levels. The relationship is statistically significant.

There is a weak positive relationship between pregnancies and blood pressure. The relationship is statistically significant.

There is a weak negative correlation between pregnancies and glucose levels. The relationship is statistically significant.

There is a weak negative relationship between pregnancies and insulin levels. The relationship is statistically significant.

There is no significant relationship between pregnancies and skin thickness.

There is a weak positive relationship between pregnancies and age. The relationship is statistically significant.

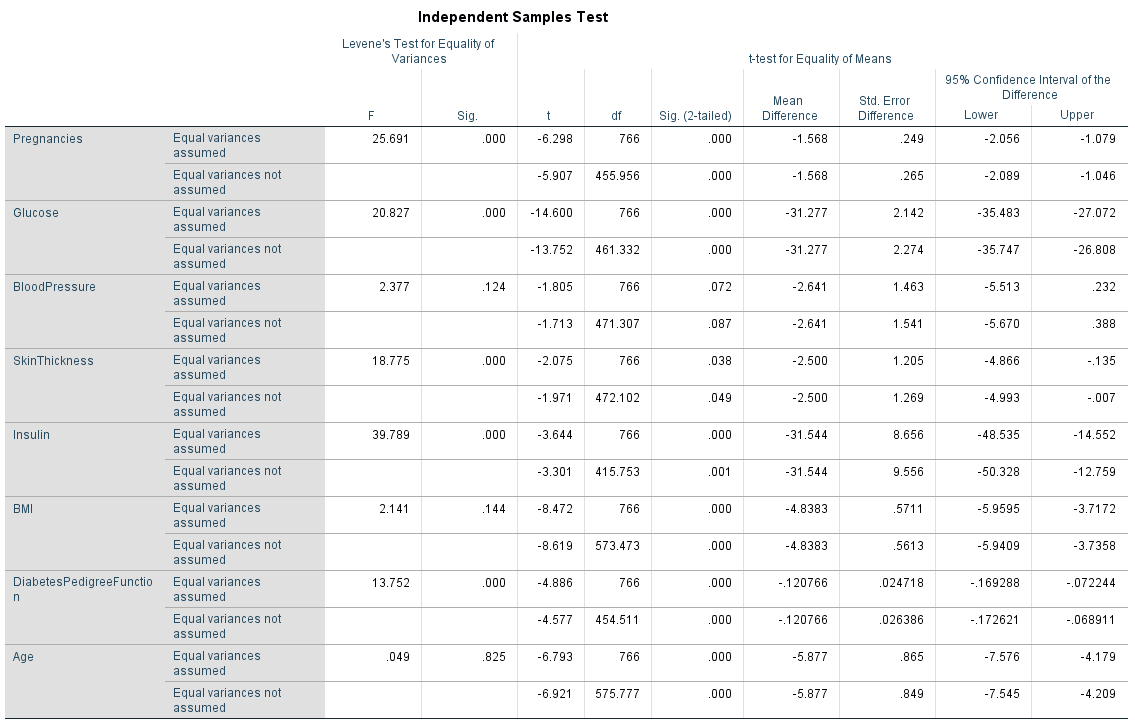
There is a weak positive relationship between blood pressure and glucose levels. The relationship is statistically significant.

There is a weak positive relationship between insulin and glucose levels. The relationship is statistically significant.

**T-Test**

T-Test helps determine whether there is a statistically significant difference in the means of the continuous variable across different levels of the categorical variable.

**4.2.2. Relationship between All continous variables and Diabetes Outcome.**



The p-value associated with pregnancies is very small than the significance level of 0.05. This indicates a strong statistical significance. This suggests that there is a statistically significant difference in pregnancies mean across different diabetes. Pregnancies has a significant impact on diabetes. Thus, the specific diabetes outcomes are likely to have different pregnancy means.

The p-value associated with glucose is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in average glucose level between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different glucose levels.

The p-value associated with blood pressure is very higher than the significance level of 0.05. This indicates a low statistical significance. This suggests that there is no statistically significant difference in blood pressure across different diabetes outcomes. Thus specific diabetes outcomes are likely to have similar blood pressure.

The p-value associated with skin thickness is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in skin thickness between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different skin thickness.

The p-value associated with insulin is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in average insulin levels between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different insulin levels.

The p-value associated with bmi is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in mean bmi between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different bmi levels.

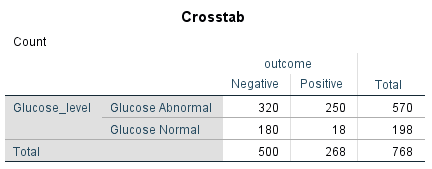
The p-value associated with diabetes pedigree function is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in mean diabetes pedigree function between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different diabetes pedigree function.

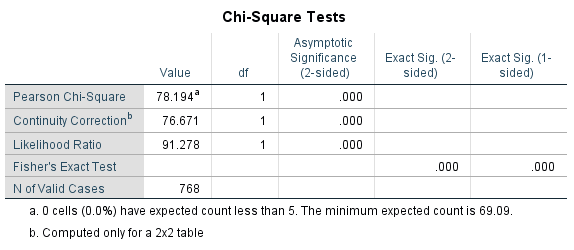
The p-value associated with age is very small indicating a strong statistical significance. This suggests that there is a statistically significant difference in mean age between negative and positive participants to diabetes. Different diabetes outcomes are therefore likely to have different age groups.

**Chi-Square Test**

The Chi-Square test is a statistical tool used to evaluate the association between categorical variables. It is particularly useful when we want to understand if there is a significant relationship or dependency between two or more categorical variables.

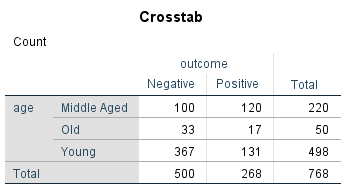
**4.2.3. Relationship between Glucose levels and Diabetes Outcome.**

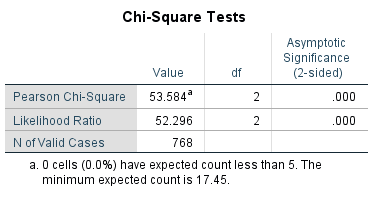
****



The Chi-Square test for independence was conducted to examine the relationship between glucose levels and diabetes. The test yielded a Chi-Squared statistic of 78.194 representing how much the observed data deviates from what would be expected if there were no relationship between the variables being studied. This high value suggests a strong association between the two variables. The extremely low p-value of (0.000), indicates a highly significant association between the two variables since the p-value is significantly smaller than the significance level of (0.05). This suggests that there is a substantial connection between glucose levels and diabetes.

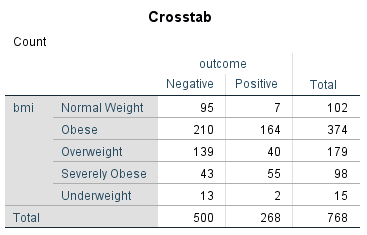
**4.2.4. Relationship between Age groups and Diabetes Outcome.**

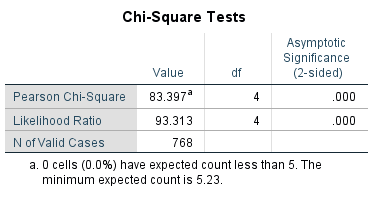
****



The Chi-Square test for independence was performed to investigate the relationship between age groups and diabetes. The analysis produced a Chi-Squared statistic of 53.584. The resulting p-value, which is less than the conventional significance level of 0.05, strongly indicates a significant association between the two variables. This suggests that age groups is closely linked with the diabetes outcome being studied.

**4.2.5. Relationship between BMI levels and Diabetes Outcome.**

****

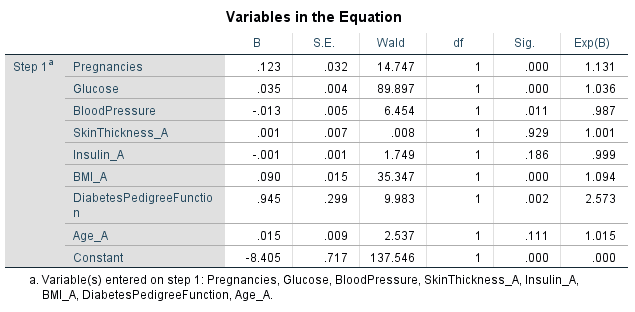


The Chi-Square test for independence was employed to explore the relationship between bmi levels and diabetes outcome. The analysis resulted in a Chi-Squared statistic of 83.387 with 4 degrees of freedom. The p-value obtained was less than the conventional significance level of 0.05, indicating a highly significant association between the two variables. This strongly suggests that bmi levels is substantially connected to diabetes outcome.

**4.3. MULTIVARIATE ANALYSIS**

In this section, we delved into a multivariate analysis using a binary logistic regression model. The aim is to gain insights into how Age, BMI, Average glucose levels, blood pressure, diabetes pedigree function, skin thickness, insulin and number of pregnancies collectively influence diabetes outcome. This approach allows us to assess the combined impact of these important variables and understand their causal relationship with diabetes outcome.

**4.3.1. The effect of each predictor on the diabetes outcome.**



Pregnancies is statistically significant in determining diabetes outcome evidenced by its p-value less than 0.05.

Glucose is statistically significant in determining diabetes outcome indicated by its p-value less than 0.05.

Blood pressure is statistically significant in determining diabetes outcome evidenced by its p-value less than 0.05.

Skin thickness is not statistically significant in determining diabetes outcome evidenced by its p-value above 0.05.

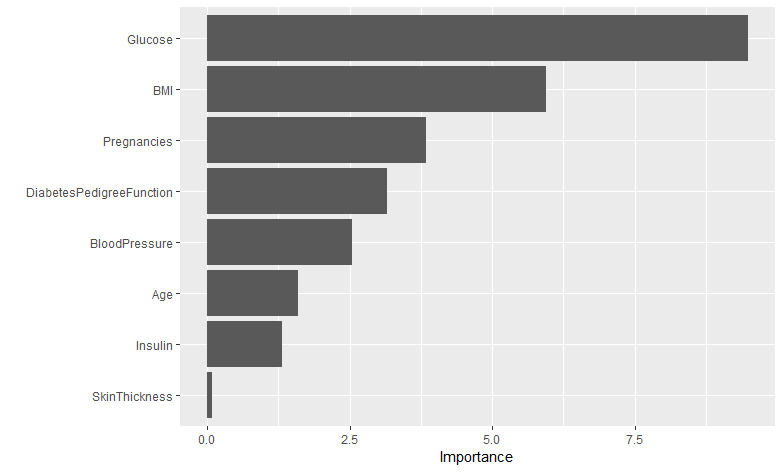
Insulin is not statistically significant in determining diabetes outcome indicated by its p-value above 0.05.

BMI is statistically significant in determining diabetes outcome evidenced by its p-value less than 0.05.

Diabetes pedigree function is statistically significant in determining diabetes outcome evidenced by its p-value less than 0.05.

Age is statistically significant in determining diabetes outcome evidenced by its p-value less than 0.1.

**4.3.2. The importance of each predictor in determining diabetes outcome.**

****

Glucose emerged as the top predictor of diabetes while skin thickness was the least predictor of diabetes.

**4.3.3. The performance of the model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **RMSE** | **AUC0** | **AIC** |
| **Value** | **0.391** | **0.839** | **718** |

The model achieved an auc of 0.839 making it an excellent tool in making predictions.

**5. Recommendations**

Based on the findings, the following recommendations are proposed: Routine Health Screenings: Regular medical check-ups for blood pressure, insulin levels, glucose levels, and BMI should be encouraged, particularly for individuals over 40.

Public Health Campaigns: Awareness programs should focus on the impact of obesity, hypertension, pregnancy, and insulin levels on diabetes risk.

Targeted Lifestyle Interventions: Weight management programs, and dietary guidance should be promoted to mitigate diabetes risks.

Early Detection Programs: Government and private health organizations should establish community-based early detection screenings for diabetes risk factors.

Workplace Health Initiatives: Employers should promote stress reduction programs, provide health check-ups, and encourage physical activity.

Enhancing Access to Healthcare: Healthcare providers should develop specialized clinics focusing on diabetes prevention and chronic disease management.

Further Research: More in-depth studies should be conducted with larger and more diverse populations to refine predictive models and uncover additional risk factors.

**6. Conclusion**

This study identifies age, blood pressure, insulin, bmi, number of pregnancies, diabetes pedigree function and glucose levels as major risk factors for diabetes. Addressing these factors through lifestyle modifications, early medical intervention, and public health policies can significantly reduce diabetes incidence and improve overall health outcom**es.**