|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  |
| DETERMINING RISK FACTORS FOR DIABETES  MATH40031\_Statistical Data Analysis and Visualisation | | |
| N1086682 | | MOHAMED BASHARUL KHAN |

**Table of Contents**

**INTRODUCTION 3**

**MISSING VALUES3**

**N/A VALUES3**

**SUMMARY AND VISUALIZATION OF THE DIABETES DATASET4**

**STATISTICAL TESTS AND ITS VISUALIZATION5**

**CORRELATION COEFFICIENTS6**

**REGRESSION MODELLING7**

**PREDICTION OF GLUCOSE BASED ON AGE7**

**CONCLUSION8**

**BIBLIOGRAPHY9**

|  |  |
| --- | --- |
|  |  |
| Introduction The dataset provided pertains to the prevalence of diabetes among women of Pima Indian heritage, who are at least 21 years old. This report discusses about the missing entries in dataset, discarding variables, performing statistical analysis, discussing of insights and influence of diabetes in women, Visualization and Modelling techniques of the data will give a better understanding of relationship between clinical variables and diabetes of the provided dataset.  **Missing Values**  In this dataset, missing values are represented by 0s, this approach of representing missing data as 0s can be problematic, as it can distort the distribution and relationships of the variables.  For instance, consider the variable “BloodPressure”. It is unlikely that a person has a blood pressure as 0, So a 0 in this variable may indicate missing data. If missing values are not properly handled, it could lead to biased or inaccurate statistical analyses, as the missing data may have an impact on the relationships between variables.  Therefore, it is important to identify and handle missing data appropriately before conducting statistical analysis. This may involve imputing the missing values with estimates based on other variable data or removing the observations with missing data, depending on the extent and nature of missing values.  **N/A Values** | |
| After assuming that there is no missing data in “outcome” variable and discarded the “pregnancies” variable because it is not relatable with other variables in dataset and not required for further analysis.  Cleaning the data will make the missing data entries to “NAN” or “N/A” value and by proceeding for further analysis with that dataset will lead to biased or incorrect statistical analysis results.  Therefore, it is important to remove the column or replace missing values before performing statistical analysis. There are several ways to handle missing data, in that according to our situation Mean imputation method will resolve this issue. It replaces the missing values with the mean of the observed values for that variable. The variables such as Skin Thickness and Insulin have more missing values in their column and the imputation technique won’t be appropriate for those, so that we are removing the column, sometimes we need to take tough call like this otherwise the analysis won’t be accurate.  Although, variables like Glucose, Blood Pressure and BMI have fewer missing values and it is replaced by mean imputation technique.  Mean imputation implementation for Blood Pressure variable:  The below first line of code calculates the mean of the blood pressure variable by omitting “NA” values and then the second line of code imports the calculated mean value of blood pressure to the missing entries places of the blood pressure column in the diabetes dataset. This finally shows that there are no “NA” values, and the dataset is good to go for further analysis.  **R Code:**  mean\_glucose <- mean (Glucose, na.rm = TRUE)  diabetes\_data\_1[is.na(diabetes\_data\_1$Glucose), "Glucose"] <- mean\_glucose  **Summary and Visualization of the diabetes dataset**  In this dataset there are 768 health observation of women, in that 65.1% of women who do not have diabetes and 34.9% of women have diabetes as per data.  The mean glucose level for all women in the dataset is 121.6 mg/dL, with a standard deviation of 30.4. Women who have diabetes have a significantly higher mean glucose level (142.2 mg/dL) compared to women who do not have diabetes (110.7 mg/dL).  The mean BMI for all women in the dataset is 32.4 kg/m^2, with a standard deviation of 6.8. Women who have diabetes have a significantly higher mean BMI (35.3 kg/m^2) compared to women who do not have diabetes (30.8 kg/m^2).  The mean age for all women in the dataset is 33.2 years, with a standard deviation of 11.7. Women who have diabetes have a significantly higher mean age (37.0 years) compared to women who do not have diabetes (31.1 years).  The mean diabetes pedigree function (a measure of family history of diabetes) for all women in the dataset is 0.5, with a standard deviation of 0.3. Women who have diabetes have a slightly higher mean diabetes pedigree function (0.6) compared to women who do not have diabetes (0.4).  Also, various plots like histogram, scatterplot and bar plot are visualized and it shows significant distributions of each clinical variable.  When coming to the criteria of normality, we conducted Shapiro-Wilk Test, the p-values for all the variables is less than significant value of 0.05 and concluding that the whole data is not normally distributed.  [Fig 1. Visualization of clinical variables of diabetes dataset](https://drive.google.com/file/d/1jKL2x0weHPYJFXTihEgftTXCIt_zdxZS/view?usp=share_link) | |  |

|  |  |
| --- | --- |
|  |  |
| Statistical tests and its Visualization To assess the differences in central tendencies of predictor variables and diabetes outcome, two sample t-test is used to check whether there is any statistically significant difference between individuals with diabetes and those without diabetes.  For all performed test we have used 95% confidence interval, not paired, and assumed non – equal variances.   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Comparison | Null Hypothesis | t-test value | p-value | Result | | Glucose ~ Outcome | Ho: The mean glucose level of non-diabetes = The mean glucose level of diabetes | -13.752 | < 2.2e-16 | Reject Null Hypothesis – p-value is less than significant value 0.05 Alternative Hypothesis: There is a significant mean difference in the glucose levels between patients with and without diabetes. | | Age ~ Outcome | Ho: The mean age of Non-diabetes = The mean age of diabetes | -6.9207 | 1.202e-11 | Reject Null Hypothesis – p-value is less than significant value 0.05 Alternative Hypothesis: There is a significant mean difference in the Age between patients with and without diabetes. | | BMI ~ Outcome | Ho: The mean BMI of Non-diabetes = The mean BMI of diabetes | -8.6193 | < 2.2e-16 | Reject Null Hypothesis – p-value is less than significant value 0.05 Alternative Hypothesis: There is a significant mean difference in the BMI between patients with and without diabetes. | | DiabetesPedigreeFunction ~ Outcome | Ho: The mean DiabetesPedigreeFunction of Non-diabetes = The mean DiabetesPedigreeFunction of diabetes | -4.5768 | 6.1e-06 | Reject Null Hypothesis – p-value is less than significant value 0.05 Alternative Hypothesis: There is a significant mean difference in the DiabetesPedigreeFunction between patients with and without diabetes. | | |

# Therefore, there are significant differences in central tendencies between clinical variables and diabetes outcome. To depict this visually refer figure 2 for the good understanding of differentiation within variables.

[Fig 2. Measure of Central Tendency within clinical variables and diabetes outcome.](https://drive.google.com/file/d/1UOKopW7m-lqTidl-iESvCZ8zHdi_rAyG/view?usp=share_link)

|  |  |
| --- | --- |
|  |  |
| **Correlation Coefficients**  The strength of relationship between variables is called correlation, using “cor()” command we can compute the correlation coefficient, after trying different methods of correlation coefficient calculations like Pearson, Kendall and Spearman techniques, Pearson method is giving good results rather than the other techniques.  Using the “corrplot()” we can get a visualization of correlation matrix of diabetes data set, refer fig 3.  [Fig 3. Correlation Matrix of Diabetes Dataset](https://drive.google.com/file/d/1-9hcZ-_2mgqLkngZCbuQ3SZsrMCLpq4K/view?usp=share_link)  By observation of correlation matrix there are some high coefficient values between the clinical variables,   |  |  | | --- | --- | | **Observation** | **Correlation Coefficient** | | Glucose ~ Outcome | 0.49 | | BloodPressure ~ Age | 0.33 | | BMI ~ Outcome | 0.31 | | BMI ~ BloodPressure | 0.28 | | Glucose ~ Age | 0.27 |   The observation of Glucose and diabetes outcome shows moderate positive correlation, and it indicates that as glucose levels increases the likelihood of having diabetes also increases. Refer Fig 4 for the Plot.  [Fig 4. Glucose vs Outcome Plot](https://drive.google.com/file/d/102EWlHqyZxSnTjFvYFSGc7qH0EfUJQDd/view?usp=share_link)  The observation of Blood pressure and age shows moderate positive correlation, and it indicates that the aging of person does impact rise in blood pressure significantly. Refer Fig 5 for the plot.  [Fig 5. Bloodpressure vs Age Plot](https://drive.google.com/file/d/1lDGtaVN3rc6gdC4NiKhuiWFTosRZQWAf/view?usp=share_link)  The observation of body mass index and diabetes outcome shows that there is moderate positive correlation, and it indicates that if BMI increases, then there is a chance in presence of diabetes among patients. Refer Fig 6 for visualization.  [Fig 6. BMI vs Outcome Plot](https://drive.google.com/file/d/1JtYb3g3wYBRfWZGD6sqU81ozfpMoDiqN/view?usp=share_link) The observation of Body mass index and Blood Pressure shows a significant weak positive correlation, and this means that if body mass index is high, then then there is a chance of having high blood pressure. Refer Fig 7 for the visual plot. [Fig 7. BMI vs Bloodpressure Plot](https://drive.google.com/file/d/1vjZFvX1U6GeJiT_xK9P-k0mQxPssBbU9/view?usp=share_link) | |
| |  |  | | --- | --- | |  |  | | The observation of Glucose and Age shows a significant weak positive correlation, and this means that the aged person have a chance of high glucose level and normal glucose level for young person according to data. Refer Fig 8 for the visual plot. [Fig 8. Glucose vs Age Plot](https://drive.google.com/file/d/1ag1X6_Tkudax18KlIZuoDYy4RoTk9kkG/view?usp=share_link)  Therefore, Glucose and Outcome is near to the 0.5 value and this is considered as good predictors which are significantly correlated.  **Regression Modelling**  After analysis of clinical variables, there are some variables which may have an influence on diabetes. Those variables significantly help in regression modelling. The variables are Glucose, BMI, and Age, in that glucose plays a major part in deciding the diabetes outcome.  As Outcome variable is binary (diabetes or no diabetes), logistic regression may be an appropriate model to use rather than linear regression. The logistic regression model provides us with an estimated probability that a person has diabetes, based on the values of the predictor variables. The model assumes that the log odds of having diabetes is a linear function of the predictor variables.  First, we are splitting the dataset as training and testing dataset based on the diabetes outcome at a split ratio of 70%. Using Generalized linear model (glm) function in R, we can perform binomial function for the trained data set which will fit the logistic regression model. This will give us information about the coefficients, Standard errors, p – values and the confidence intervals for each predictor variable.  Second, we are using predict function to generate predicted probabilities for the testing data set and these predicted probabilities can be used to calculate predicted outcomes, which can then be compared to the actual outcomes in the testing set to assess the accuracy of the model.  At last, using ‘plot()’ and ‘lines()’ function, the final visualization is in Fig 9.  [Fig 9. Logistic Regression Model for Diabetes](https://drive.google.com/file/d/19b1Fp7xBMooRaPHW1bdb1ckdVqPg9z6Y/view?usp=share_link)  **Prediction of Glucose based on Age**  The linear regression model will best fit for this prediction. It is used to model the relationship between two continuous variables. This method helps to predict glucose levels for missing entries, by assuming glucose depends on only on age.  First, we need to subset the data into two parts as no missing entries in glucose **’data\_frame\_glucose’** and missing entries in glucose **‘data\_frame\_glucose\_missing’** from the master dataset. Then, we can fit a linear regression model to predict the glucose from age with the help of **’data\_frame\_glucose’**. This will give us information about the coefficients, Standard errors, p – values and the F- statistics. | | | |
| Second, by using the predicted model we need use ’predict()’ function for this ‘data\_frame\_glucose\_missing’ dataset, to get predicted values of glucose based on observation. Refer Fig 10 for the prediction results. [Fig 10. Prediction of Glucose values based on Age](https://drive.google.com/file/d/1XFR2c7oVnBGieN7mwzAmsPWq2wc8IjbY/view?usp=share_link) | |
| Conclusion  In conclusion, this report examines the diabetes dataset and here is our findings and computation.   1. Converted the missing values to ‘NaN’ and then the ‘NaN’ values is replaced by mean values using mean imputation technique. 2. Performed Statistical tests to assess the central tendencies of the predictor variables based on diabetes outcome. 3. Checked the correlation coefficients to test the assumption of independence among predictor variables. 4. Developed logistic regression model to describe the variables which may have an influence on diabetes outcome and plotted the logistic regression model to understand its performance and insights. 5. Predicted the Glucose level for missing entries in the dataset based on age using linear regression model.   Overall, this analysis provides useful framework for exploring and modelling diabetes datasets, with the potential to uncover important relationships between predictor variables and diabetes outcome. | |

|  |  |
| --- | --- |
|  |  |
| **Bibliography**  Dylan Z. Childs, B.J.H.and P.H.W. (2021) *APS 240: Data Analysis and statistics with R*, *B Writing a scientific report*. Available at: https://dzchilds.github.io/stats-for-bio/writing-a-scientific-report.html (Accessed: March 12, 2023).  Alam, M. (2021) *Reading and interpreting summary statistics*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/reading-and-interpreting-summary-statistics-df34f4e69ba6 (Accessed: March 12, 2023).  *Correlation plot in R ◤correlogram◢ [with examples]* (2020) *R CODER*. Available at: https://r-coder.com/correlation-plot-r/ (Accessed: March 12, 2023).  Dhandhania, K. (2018) *End-to-end data science example: Predicting diabetes with logistic regression*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16 (Accessed: March 12, 2023). | |