Maternal health risk factors dataset: Clinical parameters and insights from rural Bangladesh

Arnav Tiwary

Rahul Kumar

Ranjit Kumar V.J.

Swarnav Kalita

Indian Statistical Institute

Bangalore center

April 2025

ABSTRACT

The aim of this report is to replicate and build upon the analyses presented in the paper 'Maternal health risk factors dataset: Clinical parameters and insights from rural Bangladesh' by Mayen Uddin Mojumdar, Dhiman Sarker, Md Assaduzzaman, Hasin Arman Shifa, Md. Anisul Haque Sajeeb, Oahidul Islam, Md Shadikul Bari, Mohammad Jahangir Alam and Narayan Ranjan Chakraborty.

In this report, we outline the methods used and the sources we consulted to carry out this project. By utilizing a combination of text-based analysis methods and models, we were able to expand on the original findings and provide new insights into the topic.

Table of Contents

Introduction to the Dataset

The dataset is a collection of clinical parameters of pregnant patients at General Hospital, Kurigram, Bangladesh. It provides an overview of each individual's health risk level and demographics. The key considered vital signs were

- Age
- BMI
- body temperature
- systolic and diastolic blood pressure
- blood sugar level
- pre-existing diabetes
- gestational diabetes
- previous complications
- mental health
- heart rate

Risk Level

Especially in rural and low-resource areas, pregnancy is a vital public health concern because of limited access to quality healthcare, inadequate nutrition, and insufficient prenatal and post-natal care. These areas face higher rates of maternal and infant mortality due to a shortage of skilled birth attendants and essential healthcare resources.

Diabetes, hypertension, and mental health problems are considered pregnancy complications that require close monitoring to avoid fatal outcomes for both the fetus and the mother. Unmanaged, these complications can lead to life-threatening conditions such as preeclampsia, eclampsia, and maternal hemorrhage, which are major causes of maternal and perinatal mortality.

The main objective of the research paper as well as this project is to understand and predict risk level from various factors.

Data Cleaning v/s Preprocessing

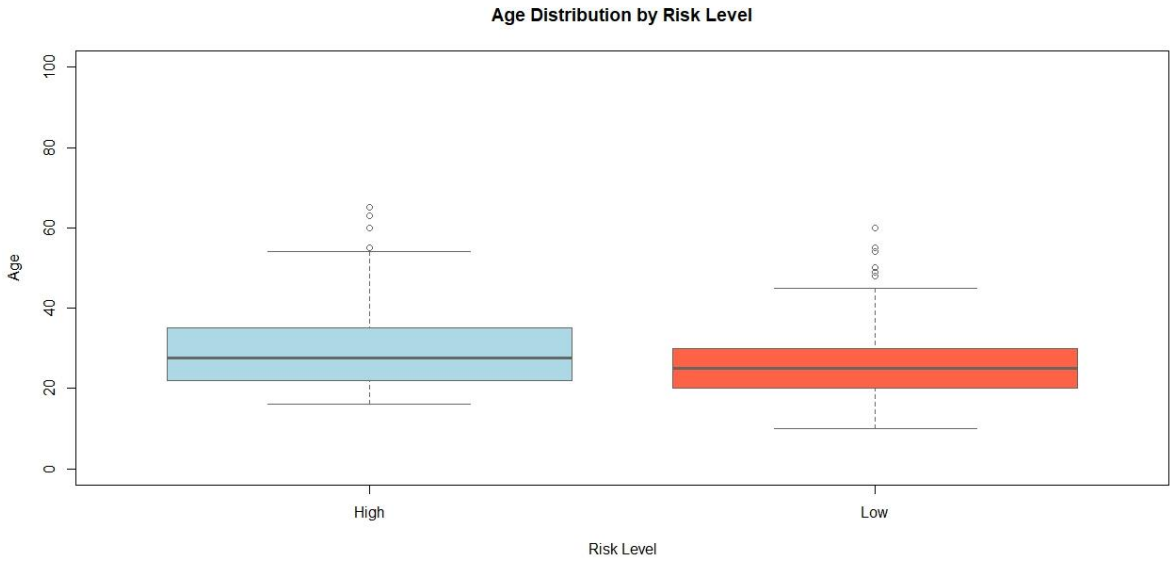
The given dataset comprised of 1205 entries, of which quite a few(39 entries) were missing values for certain variables like age or blood sugar level. There are two courses of actions we can take:

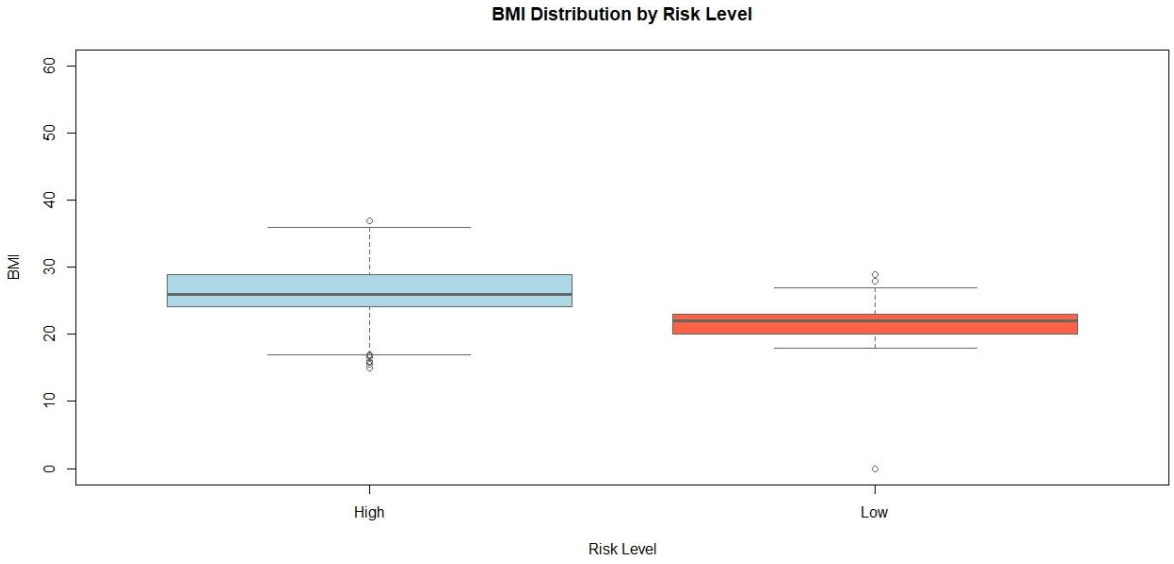
Preprocessing (Imputation) involves filling missing data points without affecting certain parameters, depending on the imputation performed. Mean imputation replaces missing values with the mean of the column, median imputation with the median and likewise for mode imputation.

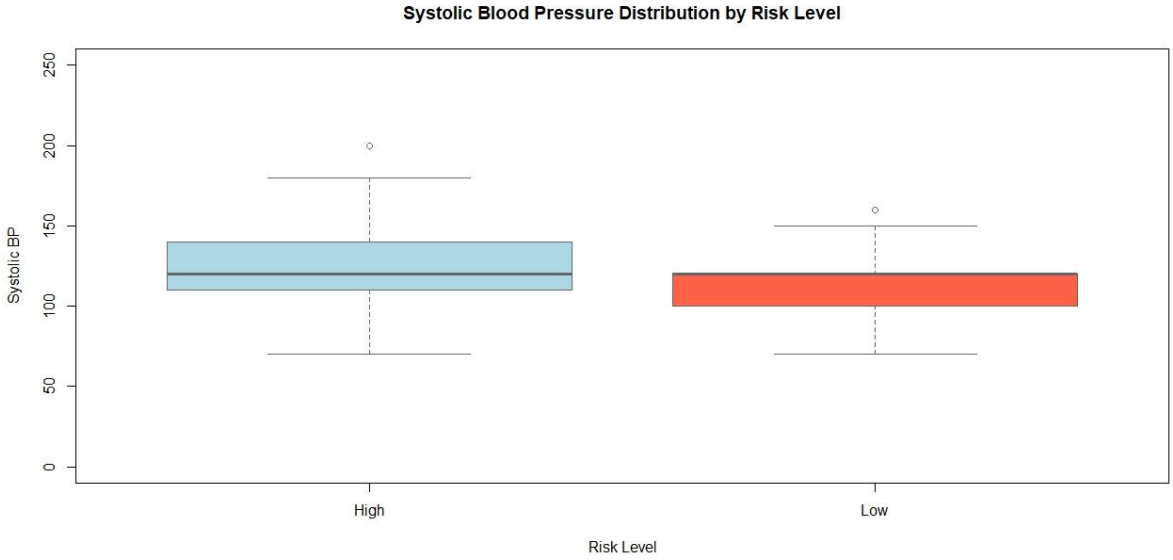
Data Cleaning involves complete deletion of the rows with any missing values.

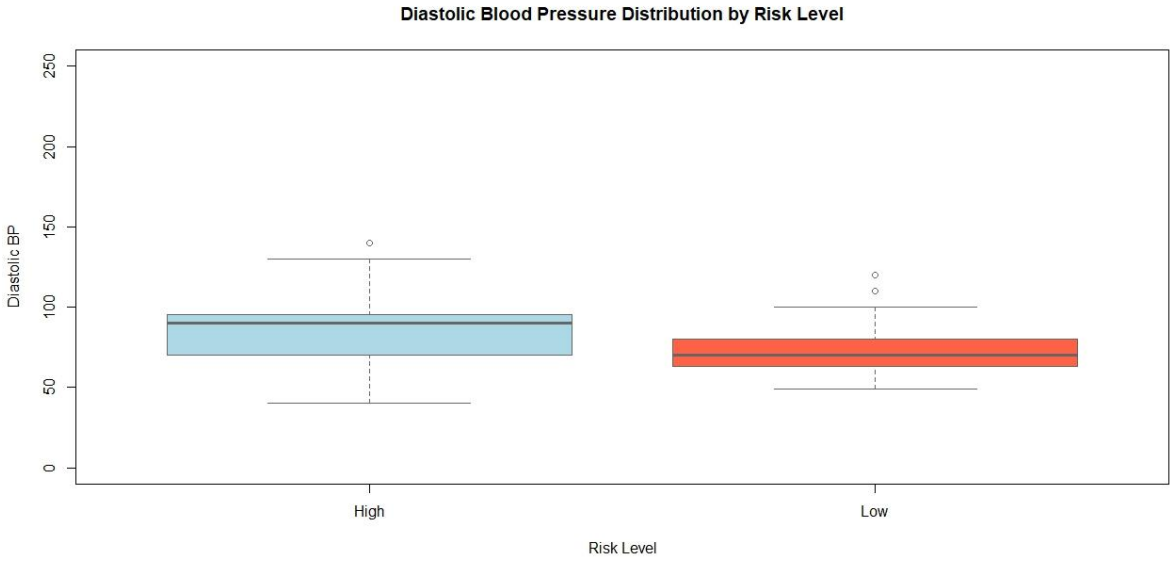
Imputation, say mean imputation, while preserving the mean, may introduce bias if the missing data wasn't random and may introduce values that do not accurately represent the true data. Therefore, we will be implementing data cleaning as our chosen method

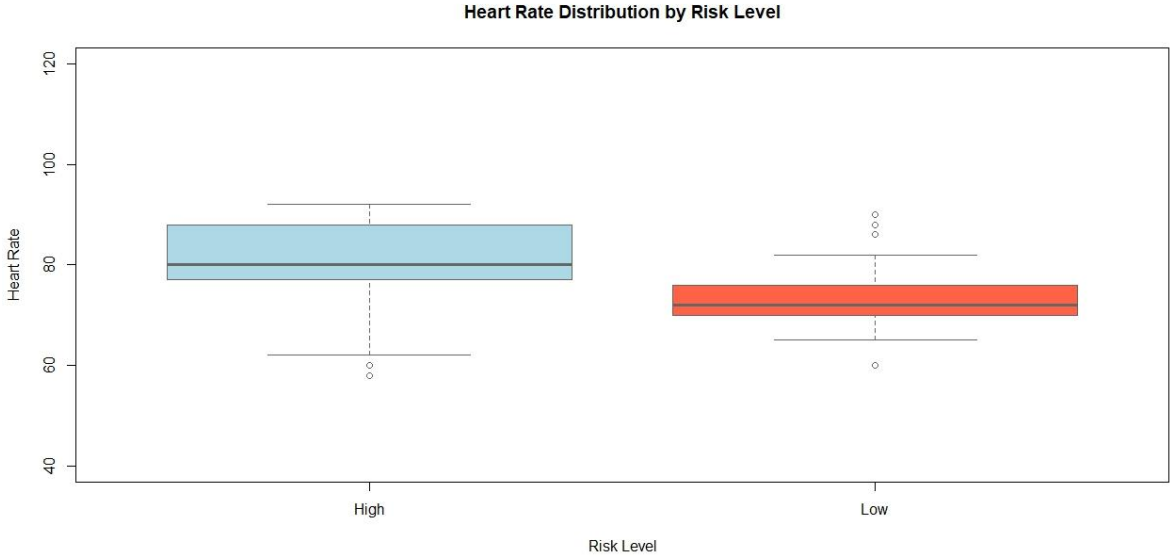
Boxplots of various variables against Risk Level






Systolic Blood Pressure Distribution by Risk Level
Systolic BP
0
50
100
150
200
250
High
Low
Risk Level


Diastolic Blood Pressure Distribution by Risk Level
Diastolic BP
0
50
100
150
200
250
High
Low
Risk Level


Heart Rate Distribution by Risk Level
Heart Rate
40
60
80
100
120
High
Low
Risk Level

Chi-Squared Test:

Chi-Squared Tests were used to determine whether two categorical variables have significant association between them.

It compares observed frequencies (f_o) with frequencies expected if the variables were independent (f_e).

Null Hypothesis ($\boldsymbol{H_0}$): The two categorical variables are independent.

Alternative Hypothesis ($\boldsymbol{H_a}$): The two variables are not independent.

Under the null hypothesis that the variables are independent, the test statistic

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Follows a chi-squared distribution with degrees of freedom, f, where f is found from the contingency table as

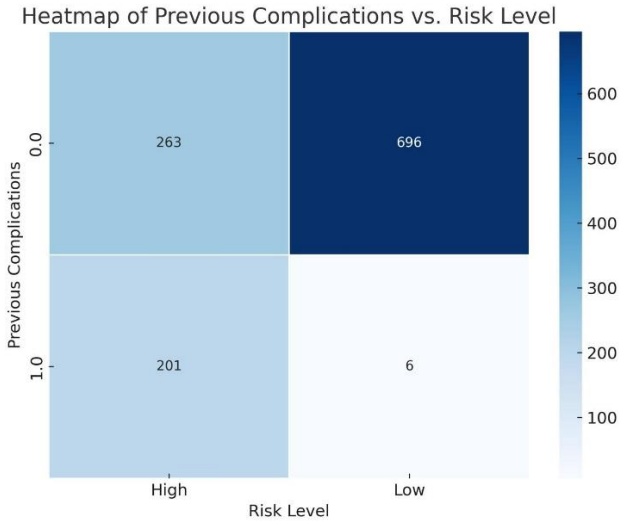
$$f = (No.\ of\ rows - 1)(No.\ of\ columns - 1)$$

Previous Complications-Risk Level:

Performing a chi-squared test on Previous Complications and Risk Level, the test statistic comes out as 342.09.

As the degrees of freedom is 1, the corresponding p-value comes out to be $<2.2 \times 10^{-16}$.

Hence, we reject the null hypothesis that previous complications and risk level are independent.

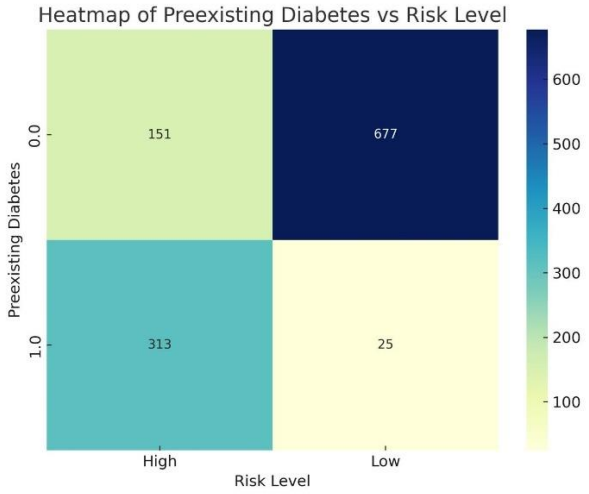


Pre-Existing Diabetes-Risk Level:

Performing a chi-squared test on Pre-Existing Diabetes and Risk Level, the test statistic comes out as 550.95.

As the degrees of freedom is 1, the corresponding p-value comes out to be $<2.2 \times 10^{-16}$.

Hence, we reject the null hypothesis that Pre-Existing Diabetes and risk level are independent.



One of the goals of the project is to understand relationships between various variables. For example, we may be interested in knowing whether pre-existing diabetes indicates higher BMI, or whether patients with mental health issues have higher blood pressure?

This helps in understanding how well we can predict one variable (say blood pressure) from another (say, mental health).

Also, when making models to fit the data (for example, through regression), it is important that the predictors we use should be independent, highly correlated predictors are redundant. So, this helps us know which predictors can be dropped from the model.

Hypothesis Testing:

Hypothesis testing was performed to examine differences in means/proportions of two populations. Recall:

- Under the null hypothesis that the means of two populations are equal, the test statistic is

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$\overline{X_1}$ and $\overline{X_2}$ are the means of the two samples from the two populations, where σ_1^2 and σ_2^2 are the variances of the two populations and n_1 and n_2 are the sizes of the two samples. Z follows a normal distribution with mean 0 and variance 1 for large samples. In case the population variances are unknown, we can use the sample variances (s_1^2 and s_2^2) as estimators of the population variances.

- Similarly, under the null hypothesis that the proportions of two populations are equal, the test statistic is

$$Z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$\widehat{p_1}$ and $\widehat{p_2}$ are the proportions in the two samples from the two populations, $\hat{p}$ is the population proportion, best estimated as $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$, the pooled proportion, n_1 and n_2 are the sizes of the two samples. Z follows a normal distribution with mean 0 and variance 1 for large samples.

▶ Correlation Test (Pearson's Test)

The Pearson correlation test is used to determine whether there is a significant linear relationship between two continuous variables, X and Y.

Under the null hypothesis that there is no linear correlation, given n paired observations, $(x_1, y_1), \ldots, (x_n, y_n)$, the sample correlation coefficient, r, is

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

And the test statistic

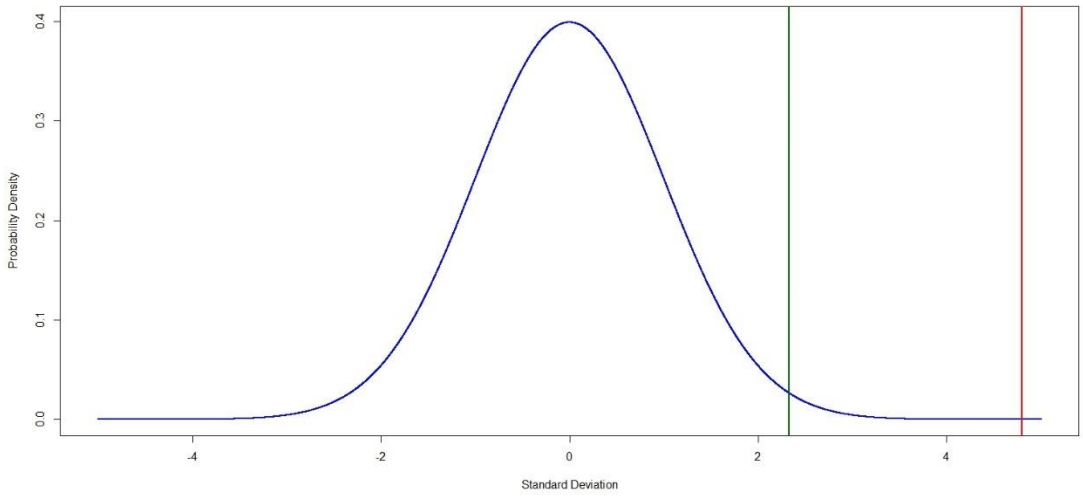
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Follows a t-distribution with $n - 2$ degrees of freedom.

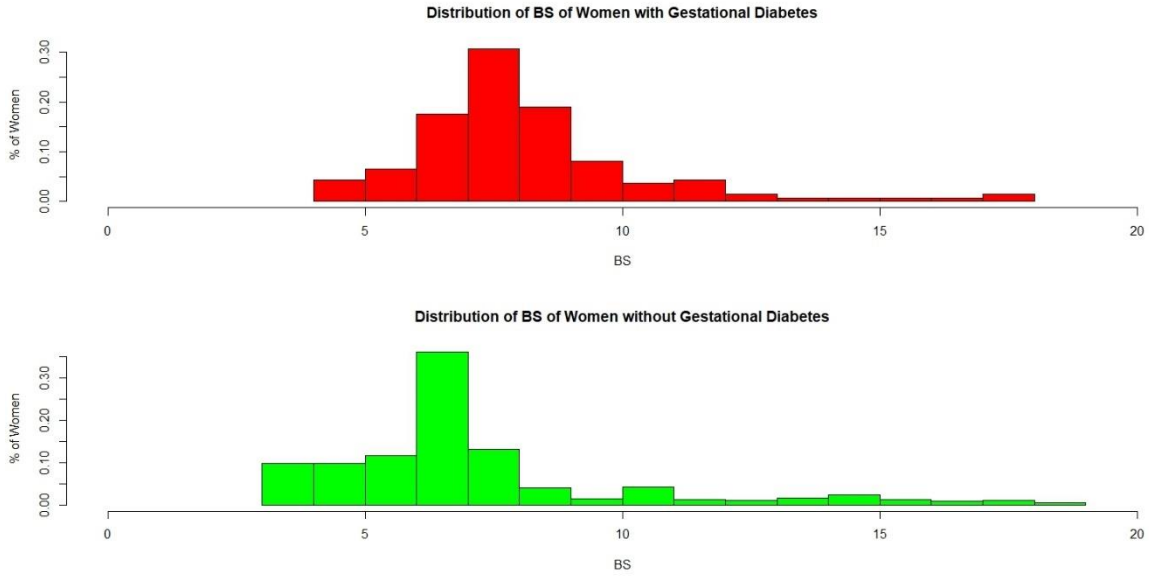
Blood Sugar-Gestational Diabetes:

Gestational diabetes mellitus (GDM) is a condition that affects pregnant women. It is crucial to understand how GDM influences blood sugar levels to assess potential health risks for both the mother and the baby. This study aims to statistically analyze whether there is a significant difference in blood sugar levels between pregnant women with and without gestational diabetes. In our data, Gestational Diabetes is a categorical variable, where 0 represents the absence of GD and 1 indicates its presence.

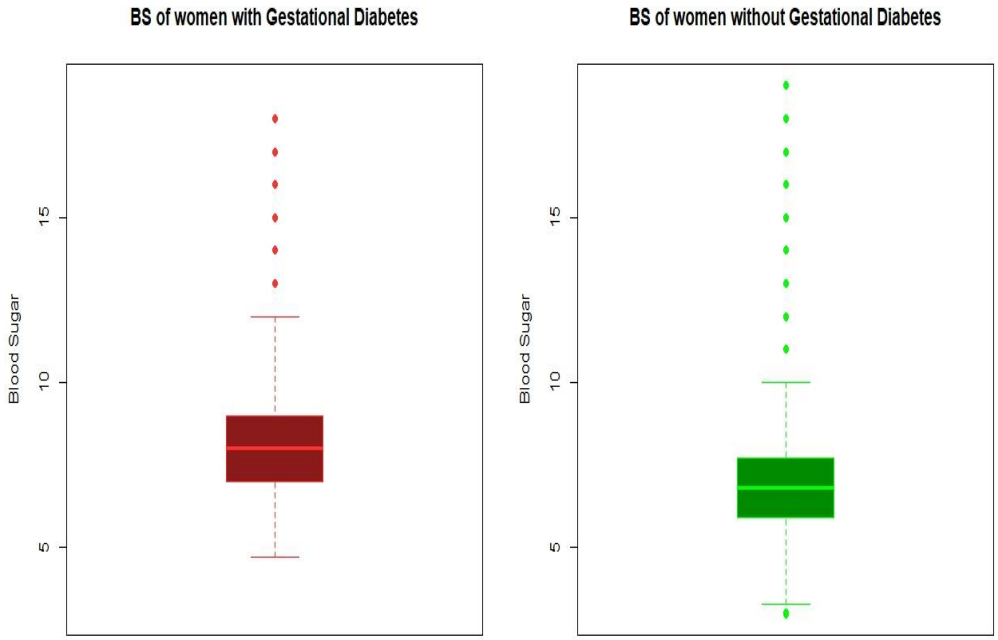
The hypothesis test results strongly indicate that gestational diabetes significantly impacts blood sugar levels, as the pnorm value is 0.9999992, meaning the difference in means is statistically significant. This suggests that blood sugar regulation is notably altered in pregnant women with GDM, reinforcing the need for medical monitoring and intervention.



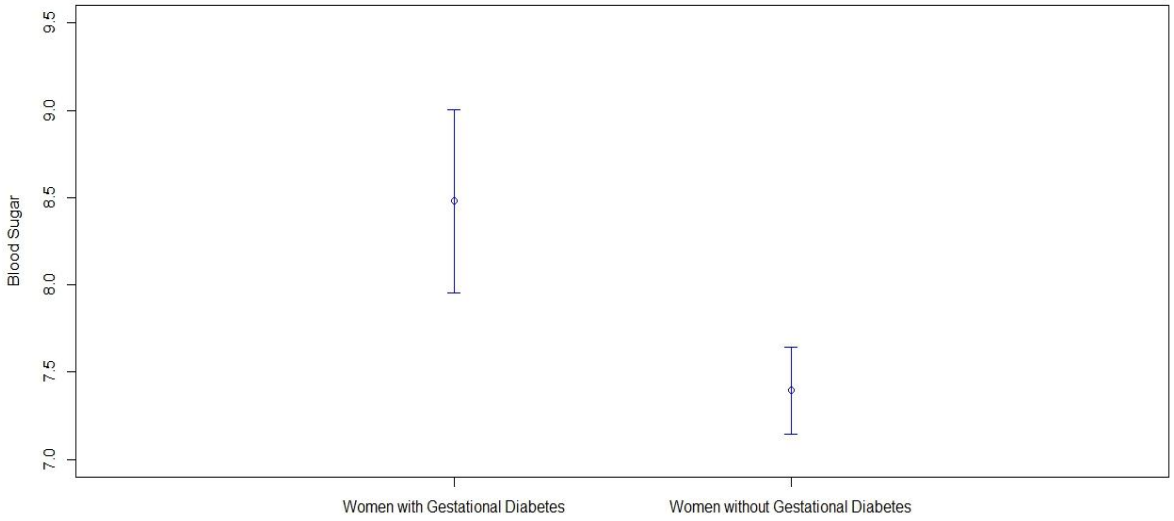
CONFIDENCE INTERVAL FOR NULL HYPOTHESIS – EQUALITY IN BLOOD SUGAR



Histograms visualizing the blood sugar level distributions in women with and without Gestational diabetes



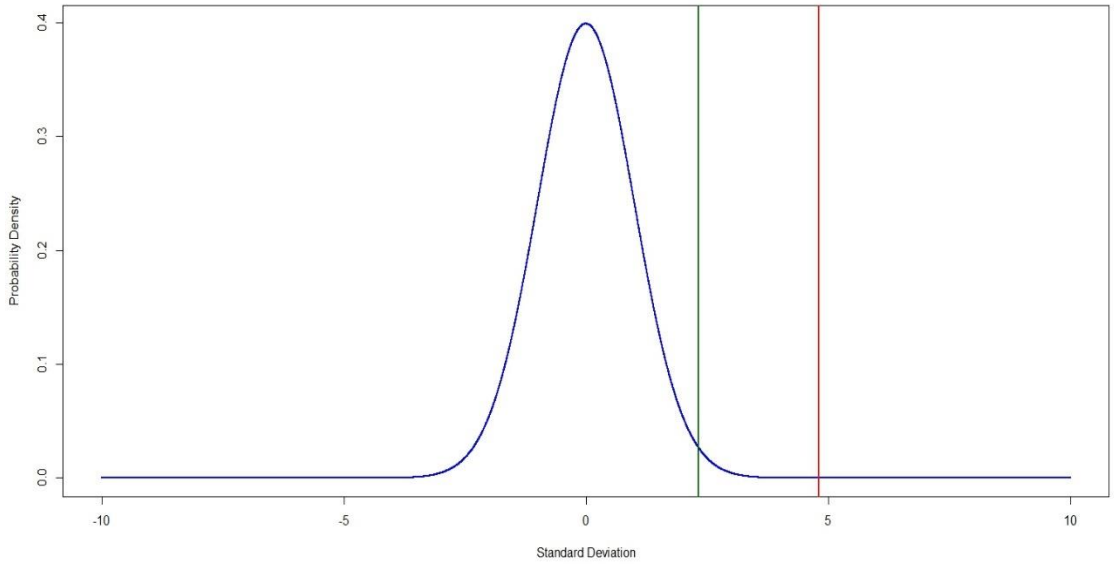
Boxplots visualizing the blood sugar level distributions in women with and without Gestational diabetes



interval estimates of the blood sugar level distributions in women with and without Gestational diabetes

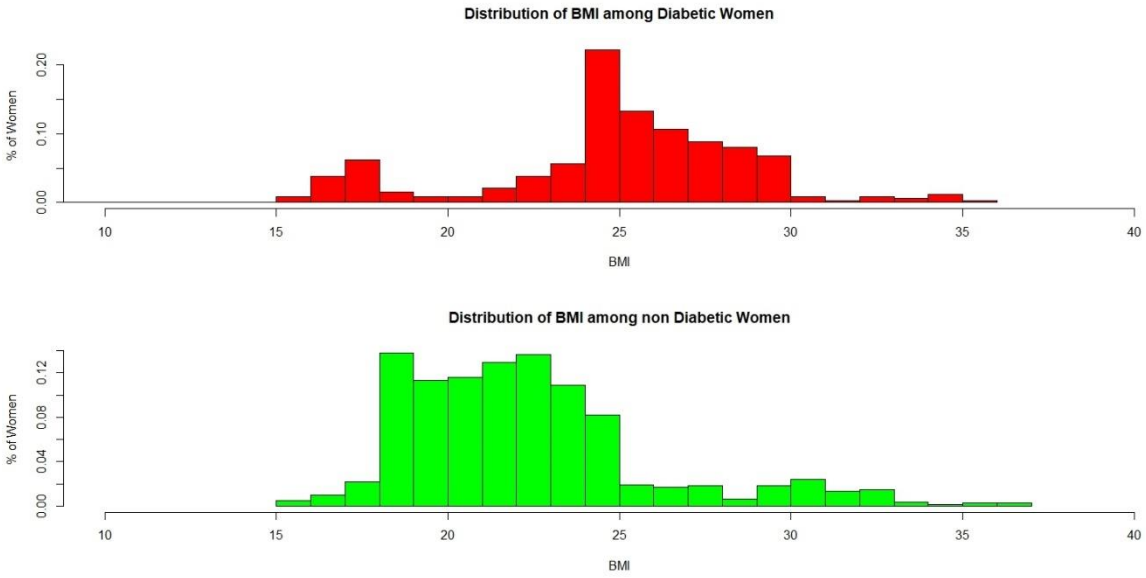
Pre-Existing Diabetes-BMI:

Diabetes is a widespread metabolic disorder that affects blood sugar regulation and is often linked to other health factors such as body weight. Understanding how diabetes impacts blood sugar levels and body mass index (BMI) can provide important insights into potential health risks for affected individuals. This study aims to statistically analyze whether there is a significant difference in BMI between pregnant women with and without pre-existing diabetes. Once again, Pre-Existing Diabetes is a categorical variable, where 0 represents the absence of diabetes and 1 indicates its presence.

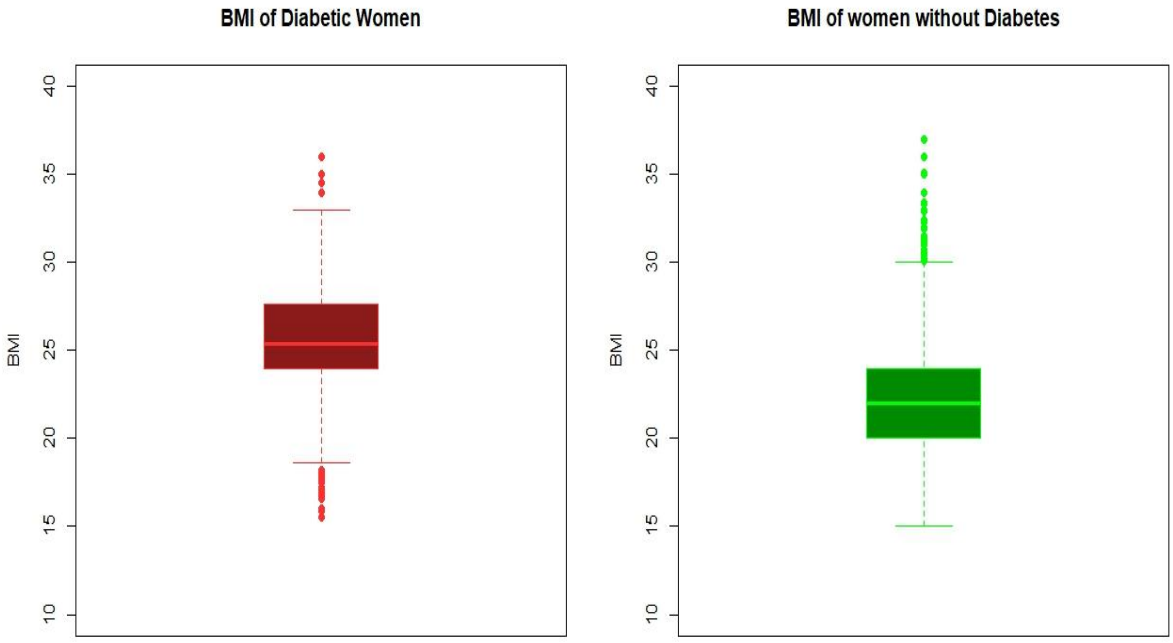


CONFIDENCE INTERVAL FOR NULL HYPOTHESIS – EQUALITY IN BMI

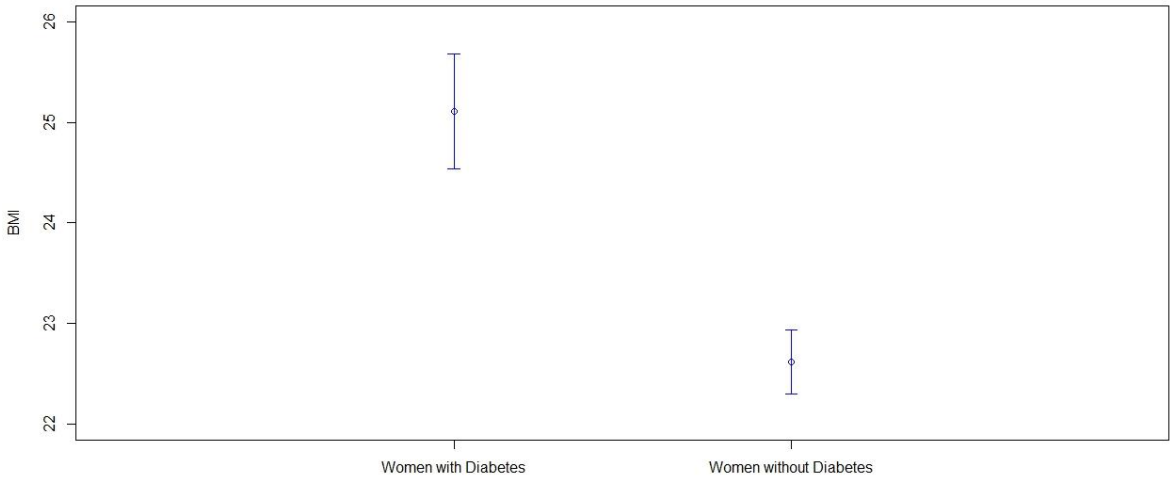
The test yielded a p-value of 3.75×10^{-23}, providing overwhelming evidence that BMI also significantly differs between these two groups. The histograms, and confidence intervals also support the result that BMI is significantly different in diabetic versus non-diabetic individuals.



Histograms visualizing the BMI distributions in both the groups.

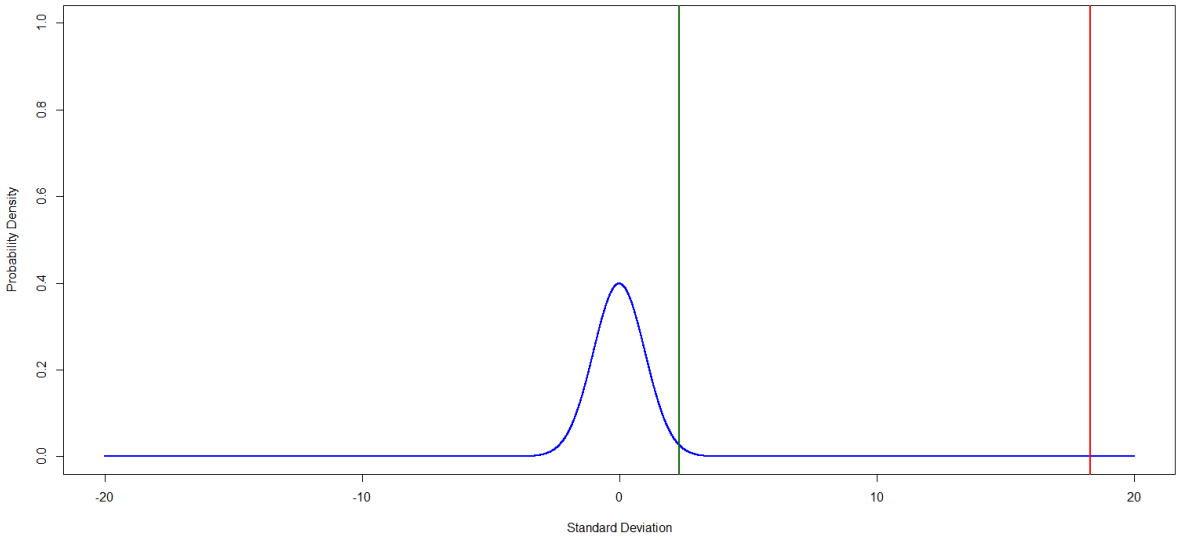


Boxplots visualizing the BMI in women with and without diabetes



interval estimates of the BMI in women with and without diabetes

BMI-Risk Level:

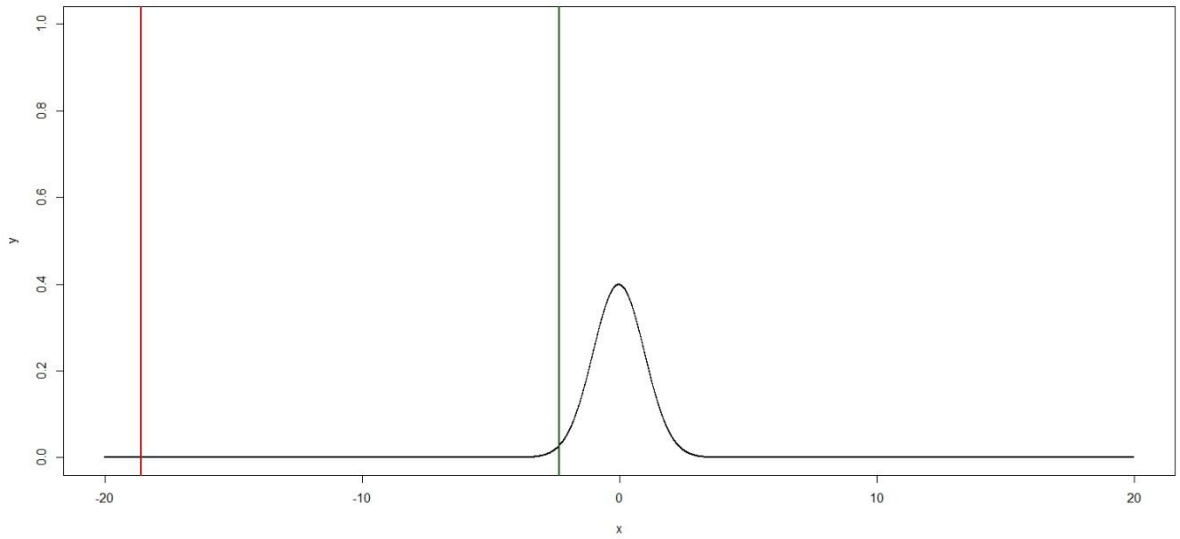


CONFIDENCE INTERVAL FOR NULL HYPOTHESIS – EQUALITY IN BMI

The hypothesis test results strongly indicate that risk level significantly impacts BMI, as the pnorm value is $1 - 3.72*10^{-75}$, meaning the difference in means is statistically significant. This suggests that BMI is notably different in pregnant women with high risk level, reinforcing the need for medical monitoring and intervention.

Previous Complications-Risk Level:

A history of complications in previous childbirth is often considered a risk factor for future pregnancies. To statistically assess this relationship, we conduct a hypothesis test to determine whether mothers with past childbirth complications have an increased risk in their current pregnancy. Both Previous Complications and Risk Level are binary variables, and we test for whether the proportion of women with high risk level is the same for the group with previous complications and for the group without them.

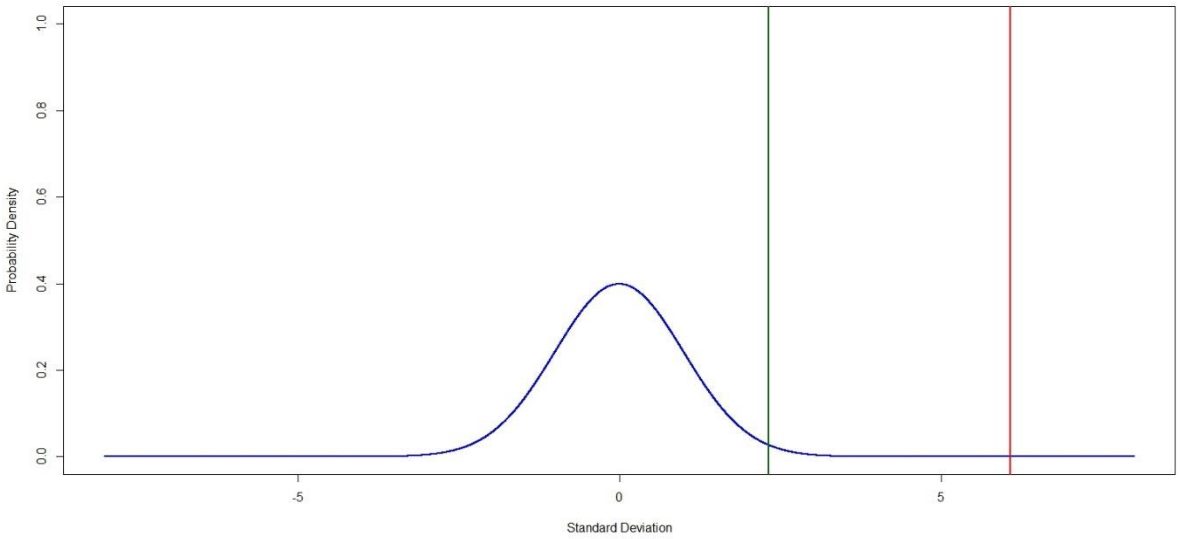


CONFIDENCE INTERVAL FOR NULL HYPOTHESIS – EQUALITY IN RISK LEVEL

The green lines demarcate 99% confidence region under the normal curve and the red line, the value of the test statistic. The corresponding p-norm value is 2.17 × 10^-47, indicating a statistically significant difference between the two groups. This suggests that previous complications are strongly associated with higher pregnancy risk.

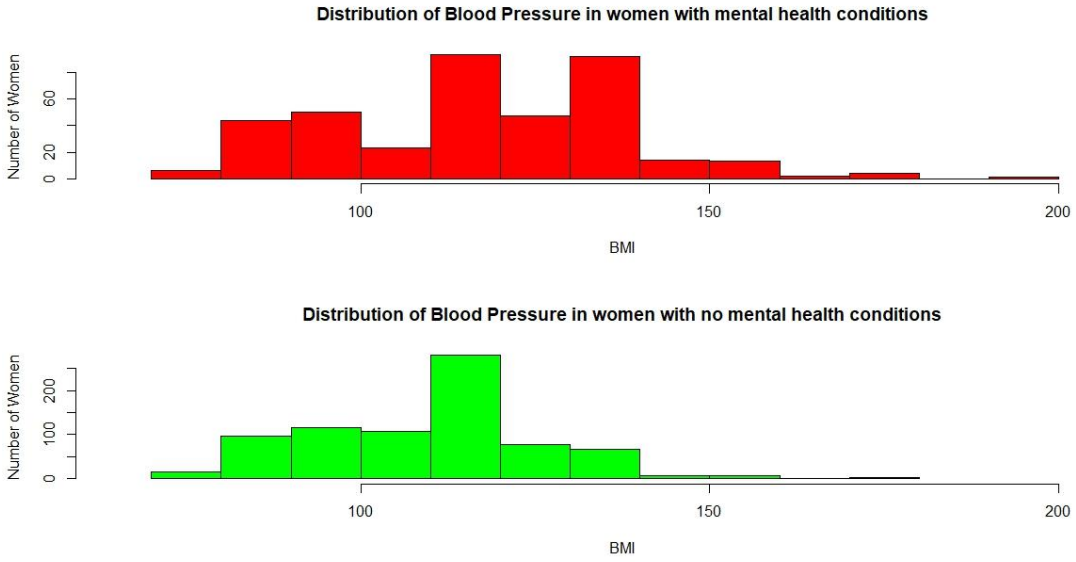
Mental Health-Blood Pressure:

Mental health is a crucial but often overlooked factor in pregnancy, with potential physiological effects on both the mother and the baby. One key aspect that may be influenced by mental health conditions is systolic blood pressure (SBP). Stress, anxiety, and other mental health issues can contribute to fluctuations in blood pressure, potentially increasing the risk of complications during pregnancy. This study aims to statistically assess whether mental health status affects systolic blood pressure in pregnant women. Mental Health is also a binary variable, with 0 and 1 denoting poor mental health and good mental health respectively.

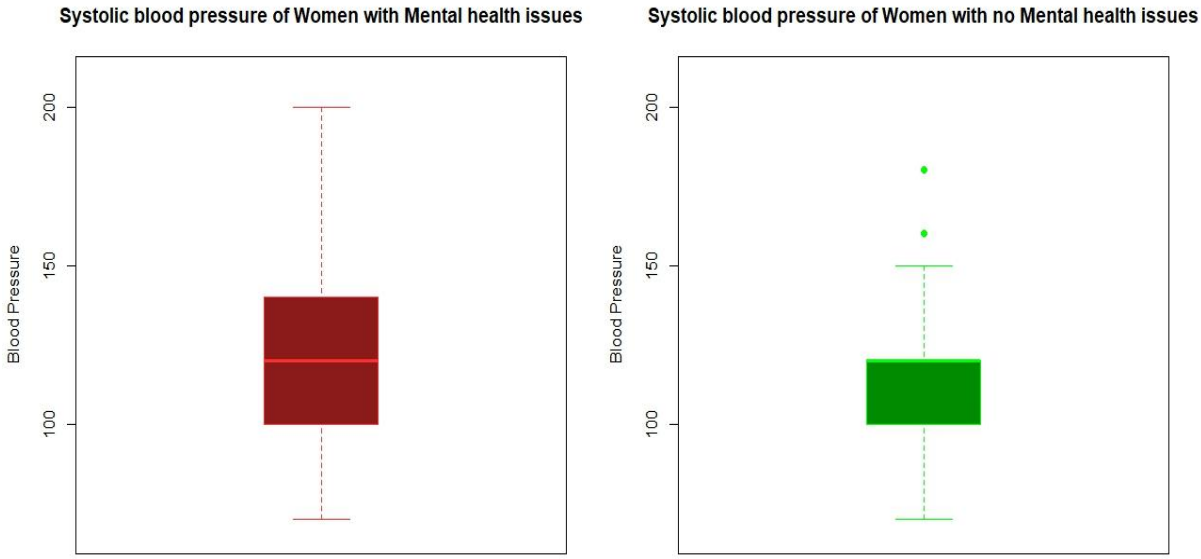


CONFIDENCE INTERVAL FOR NULL HYPOTHESIS – EQUALITY IN BLOOD PRESSURE

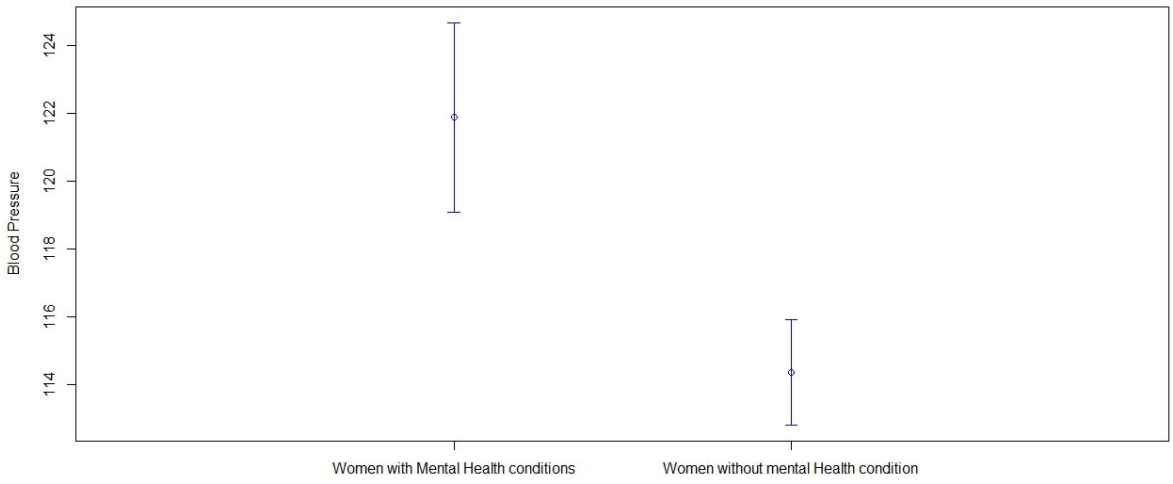
The test resulted in a p-norm value of 6.166968×10^{-10}, indicating a significant difference in blood pressure between the two groups. Interestingly, while the median systolic blood pressure was nearly the same in both groups, the mean blood pressure differed significantly. This is because most individuals with good mental health had a blood pressure close to 120 mmHg, whereas the group with mental health concerns showed a more scattered distribution, with a larger number of individuals having blood pressure levels exceeding 120 mmHg.



Histograms visualizing the blood pressure distributions in both the groups.



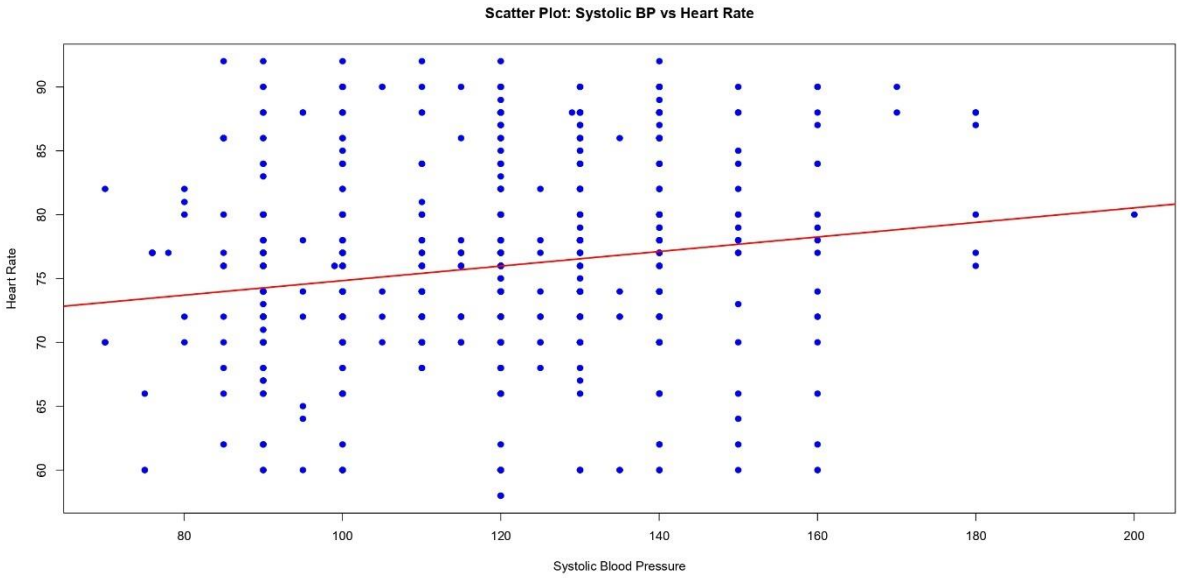
Boxplots visualizing the Blood Pressure in women with and without mental health issues



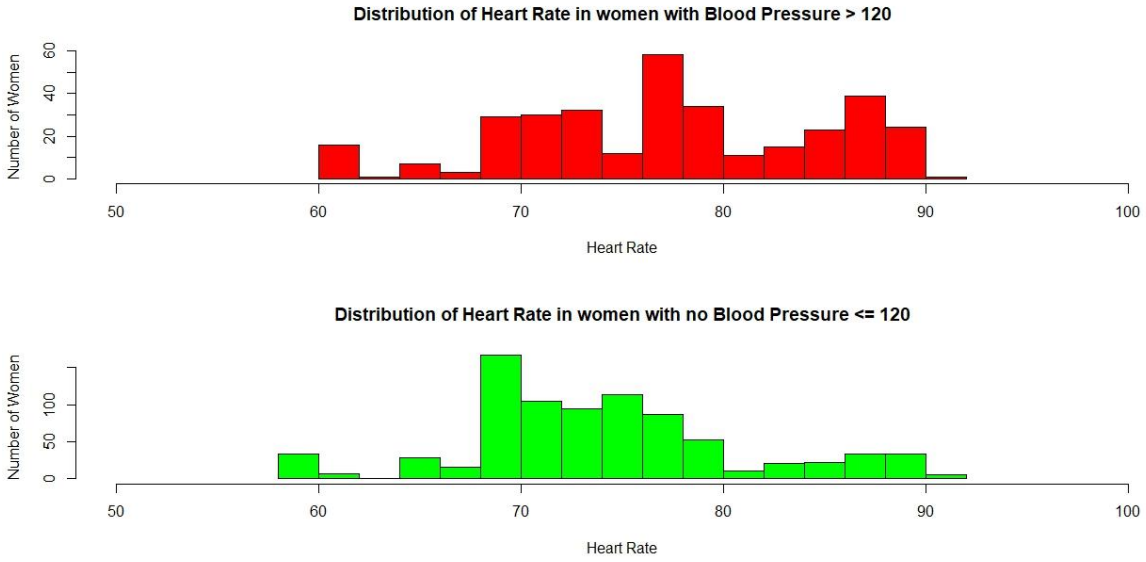
Blood Pressure-Heart Rate:

Blood pressure plays a crucial role in cardiovascular health, and its impact on heart rate is particularly important during pregnancy. Pregnant women experience physiological changes that affect circulation, and high or fluctuating blood pressure levels may put additional strain on the heart. Understanding whether blood pressure affects heart rate can help in assessing potential risks and ensuring better maternal health management. First, we divide people as those with high blood pressure (> 120 mm

of Hg) and those with low blood pressure (≤ 120 mm of Hg). Now, let us hypothesize on the mean heart rate of these two groups.

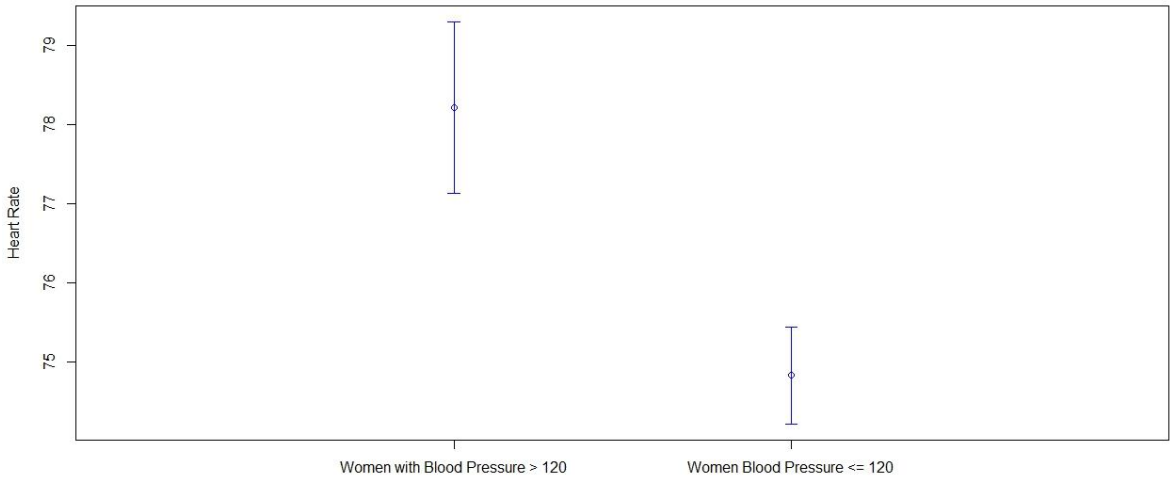


The R test demonstrates that there is indeed an increasing trend in Heart Rate with an increase in systolic Blood Pressure.



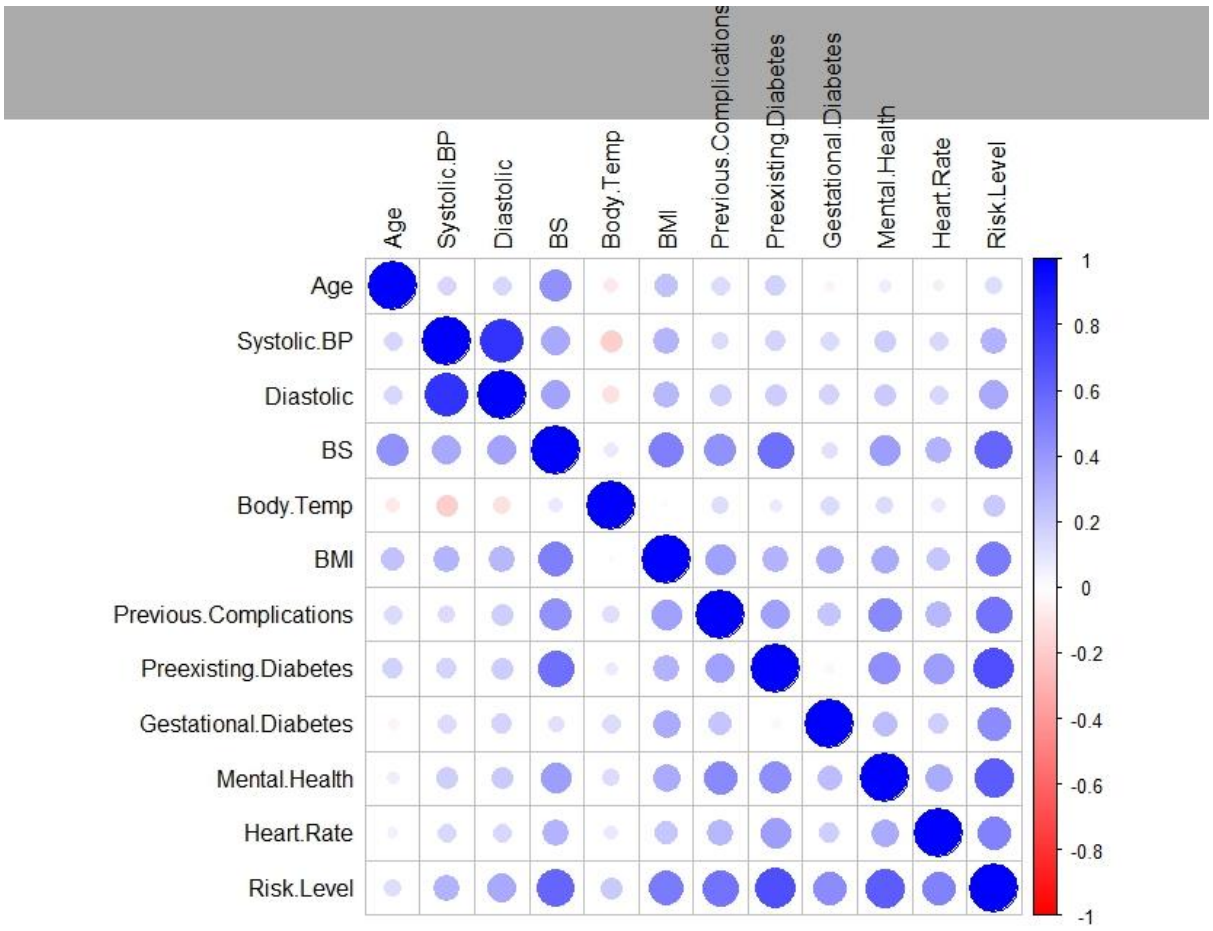
Histograms visualizing the heart rate distributions in both the groups.

Boxplots visualizing the Heart Rate in women with and without high blood Pressure



interval estimates of Heart in women with and without high blood pressure

Correlation Heatmap:



Regression Analysis:

What is Logistic Regression?

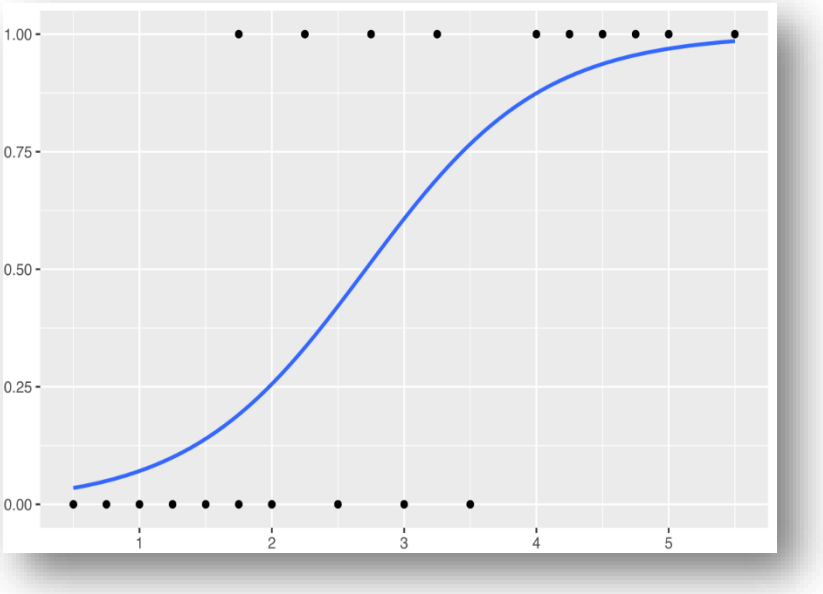
Logistic Regression is a statistical regression method applied when our dependent variable, that we wish to estimate with our predictors, is binary, i.e. it has two possible values (E.g., yes and no, 0 and 1).

Logistic Regression fits a Sigmoid curve into the data, giving a probability (of the binary dependent variable taking value say, 1) based on the predictors. Here, the risk level is a binary dependent variable, taking values 0 and 1.

The Sigmoid Curve:

$$f(X_1, \ \ldots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

In logistic regression, given values of the predictors, P($Y = 1$) is given as $f(X_1, \ \ldots, X_n)$, where f is the appropriate logistic curve, Y being the binary dependent variable, risk level in our cas



Sigmoid Curve for a single predictor, $y = \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}}$

Assumptions made during logistic regression:

$$Log - Odds = \log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right)$$

Taking Y to be the binary variable, the fraction, $\frac{P(Y=1)}{1-P(Y=1)}$, represents the "odds" of how much more likely the event, Y takes 1, is to the event that it doesn't.

Logistic Regression assumes a linear relationship between the predictors and the log-odds of the outcome, i.e.,

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

The predictors shouldn't have high degree of correlation amongst each other.

Independence of observations which implies there should be no autocorrelation.

Logistic Regression works better with larger samples, especially when dealing with rare events (high risk pregnancy).

Maximum Likelihood Estimation:

Logistic Regression uses Maximum Likelihood Estimation (MLE) to find the best regression coefficients, like how linear regression uses ordinary least squares.

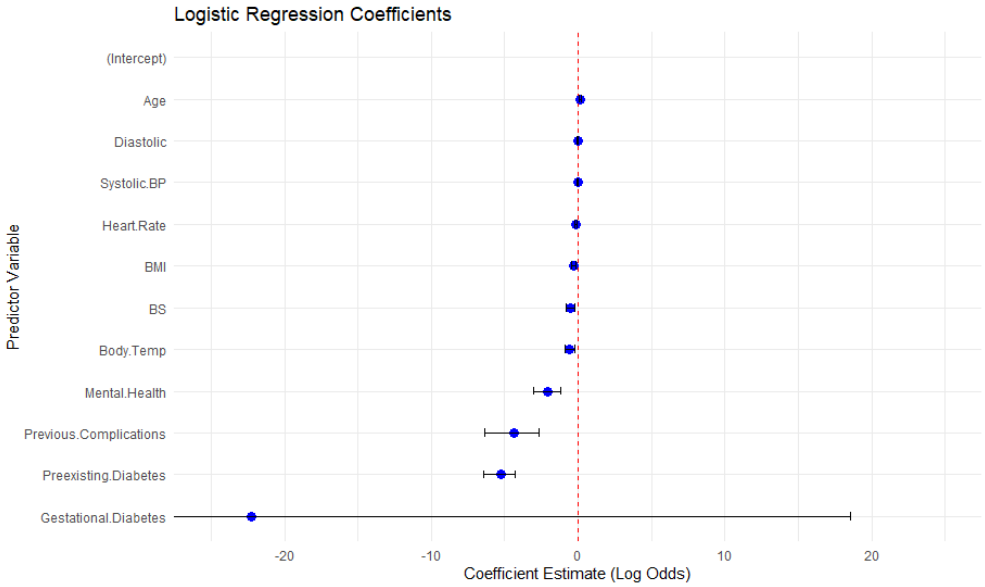
The following function is the log-likelihood function:

$$\mathcal{L}(\beta) = \sum_{i=1}^{N} [Y_i \log P_i + (1 - Y_i) \log(1 - P_i)]$$

where,

- N is the number of observations.
- $\beta = (\beta_0, \beta_1, \ldots, \beta_n)^T$ is the vector of regression coefficients.
- Y_i is the actual outcome (0 or 1).
- P_i is the predicted probability of the regression function for the i^{th} observation, taking β as the regression coefficients.

The Regression function that maximizes the probability of observing the given data is the β obtained by maximizing $\mathcal{L}(\beta)$. This can be done by various optimization methods like Newton-Raphson by a computer.



The predictors will largest coefficients strongly influence the dependent variable and those with small coefficients do not.

Strongest Predictors:

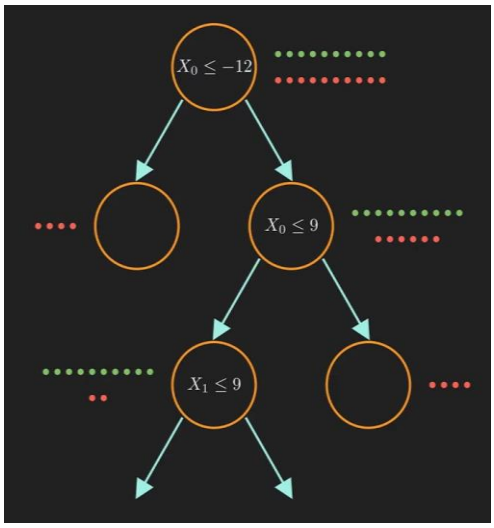
Gestational diabetes, pre-existing diabetes and previous complications are the strongest predictors of high-risk pregnancy and so, priority screening should be done for them.

Enhanced protocols should be followed for women above the age of 45 as they are at higher risk. Blood Pressure monitoring should be done checking for systolic hypertension.

Random Forest (Machine Learning model):

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that works by creating a multitude of decision trees during training.

The element of a random forest is a decision tree. A decision tree is composed of decision nodes and leaf nodes. A decision tree takes sample data and constructs conditions to divide data to (generally) two nodes. A leaf node is a binary quantity (0 or 1) and a decision node further imposes another condition.



Example of a decision tree

Decision Trees while working nicely on the data they are trained on, tend to fail for samples outside the data they were trained on. (Large Variance)

Random Forest works by creating a large number of decision trees to make it a better predictor. First, we select rows from the original dataset by sampling with replacement. Then we perform random sampling on this new dataset to select “some” rows or columns and on these smaller datasets, we train decision trees on these smaller datasets.

Reiterating these process to make a large number of decision trees is the Random Forest Algorithm. Finally, for a new sample, we run it through all the decision trees in our random forest, and the output (0 or 1) is decided by a majority vote.

Visit https://www.kaggle.com/code/ranjitkumarvj/notebook87c212fc33 to find our random forest model.

Conclusion

Healthcare infrastructure in developing countries often remains inadequate, making it crucial to understand and manage health risks effectively. Through this study, we aim to uncover significant relationships between various health parameters in pregnant women, contributing to better risk assessment and management to eradicate risks to the life and overall well being of both the mother and the infant.

References

- Maternal health risk factors dataset: Clinical parameters and insights from rural Bangladesh' by Mayen Uddin Mojumdar and others.
- All new tests introduced were taken from Wikipedia.
- The Elements of Statistical Learning by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie