Maternal Health Risk Data Exploration and Classification Report

Ishanika Singh 2023

Table of Contents:

1. Introduction 3

1. Statement of Problem…….
2. Aim of Work……...
3. Research Questions……
4. Assumptions……...

2. Data Exploration 4

1. Features and Instances……
2. Data Types……..……
3. Summary Statistics…….
4. Data Cleaning Evaluation……...
5. Data Cleaning Strategies…….
6. Visualizations and Insights…….

3. Classification Models 13

1. Preprocessing Steps……
2. Feature Importance…….
3. Decision Tree Model……..
4. Parameter Adjustment and Model Optimization……
5. Role of Parameters in Model Building……...
6. Confusion Matrix and Model Metrics……

4. Results and Discussions 18

1. Classifications and Results Analysis……..
2. Model Performance Comparison……
3. Evaluation Metrics……..
4. Conclusion and Recommendations…….

Table of Figures

Figure 1: Boxplot of Numeric Features 5
Figure 2: Scatter Plot for BS containing outliers 6
Figure 3: Scatter Plot for Body Temp containing outliers 6
Figure 4: Scatter Plot for Heart Rate containing outliers 7
Figure 5: Scatter Plot for HeartRate containing outliers 7
Figure 6: Scatter Plot for HeartRate with outliers removed 8
Figure 7: Histogram of numerical columns is maternal health risk dataset 10
Figure 8: Distribution of Risk Levels Bar Plot 11
Figure 9: New Boxplot of Dataset 11
Figure 10: Pairplot displaying individual distribution of features and the relationships between them. 12
Figure 11: Baseline Decision Tree 13
Figure 12: Plot for averaged CV scores for all max_depth tunings 14
Figure 13: Decision Tree max depth = 9 14
Figure 14: Tuned model by finding optimal max leaf nodes 15
Figure 15: Tuned decision tree 15
Figure 16: Confusion Matrix Plot 17

Table of Tables

Table 1: Summary Statistics of Continuous Numerical Features 4
Table 2: Using the Z Score to identify outliers 5
Table 3: Updated z-score table once all outliers are removed 8
Table 4: Checking for number of duplicate rows 8
Table 5: Checking for Data Types 9
Table 6: Feature Importance 17
Table 7: Classification Report 18

Task 1: Introduction *(100-200 words)*

Chosen Dataset: Maternal Health Risk

https://archive.ics.uci.edu/dataset/863/maternal+health+risk

Ahmed,Marzia. (2023). Maternal Health Risk. UCI Machine Learning Repository. https://doi.org/10.24432/C5DP5D.

This investigation involves studying data from medical facilities in Bangladesh to assess risks to maternal health. The data covers factors like age, blood pressure, blood sugar, body temperature, heart rate, and risk level, crucial for understanding maternal mortality risks.

The goal is to explore how these factors relate to maternal health risks, to enhance healthcare strategies. This analysis plans to investigate the distribution of age, variations in blood pressure levels, correlation with blood sugar levels, average body temperature and heart rate, and their impact on the risk level of maternal mortality. Assumptions include the dataset accurately representing the maternal population, the factors measured being reliable indicators of health risks, and no missing or significant data quality issues. The results could help reduce maternal mortality and support the different hospitals, communities' clinics and maternal health cares from the rural areas of Bangladesh.

Task 2: Data Exploration *(400-500 words)*

How many features (attributes) and instances exist, and what data types are these?

The dataset contains 1013 instances and 6 features, including age, systolic and diastolic blood pressure, blood sugar, body temperature, heart rate and risk level. These features are crucial for analyzing maternal health risks as they serve as significant indicators for maternal mortality. The dataset's multivariate nature allows for exploring relationships amount various variables. All features are of real or integer types, making them suitable for statistical analysis and modelling purposes. There are no missing values, ensuring the dataset's reliability for further analysis.

Provide summary statistics of the continuous numerical features.

Table 1: Summary Statistics of Continuous Numerical Features

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000	1014.000000
mean	29.871795	113.198225	76.460552	8.725986	98.665089	74.301775
std	13.474386	18.403913	13.885796	3.293532	1.371384	8.088702
min	10.000000	70.000000	49.000000	6.000000	98.000000	7.000000
25%	19.000000	100.000000	65.000000	6.900000	98.000000	70.000000
50%	26.000000	120.000000	80.000000	7.500000	98.000000	76.000000
75%	39.000000	120.000000	90.000000	8.000000	98.000000	80.000000
max	70.000000	160.000000	100.000000	19.000000	103.000000	90.000000

Perform an initial exploration of the provided dataset to assess its cleanliness. Describe the steps taken to address both data cleanliness evaluation and data cleaning strategies.

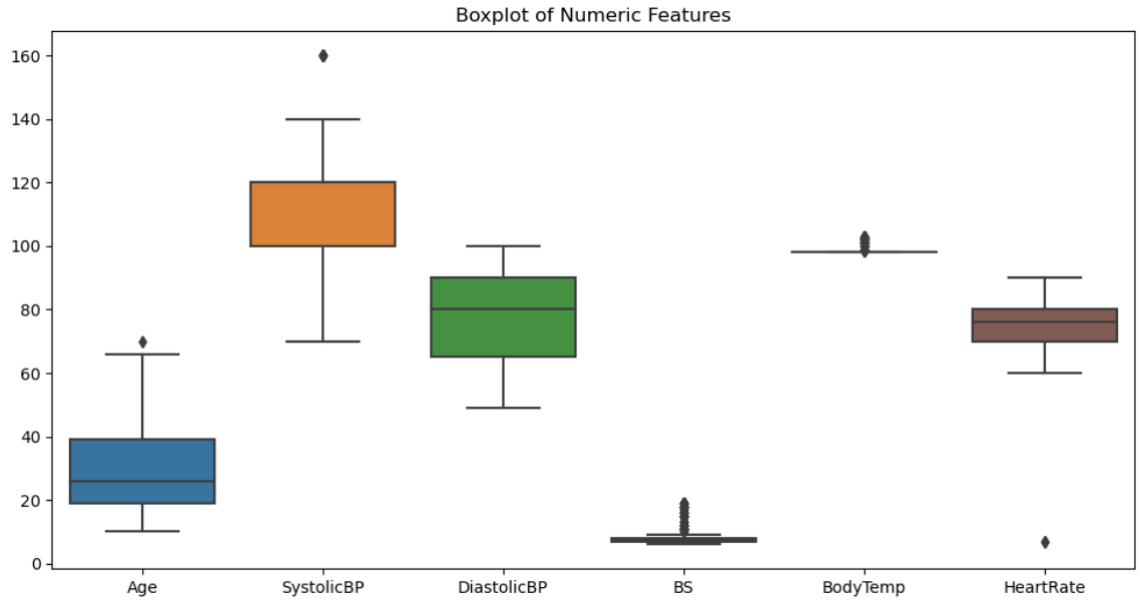


Figure 1: Boxplot of Numeric Features

Table 2: Using the Z Score to identify outliers

```
Outliers: 37
Outliers Status
Age            False
SystolicBP     False
DiastolicBP    False
BS              True
BodyTemp        True
HeartRate       True
dtype: bool
```

The images below show a scatter plot of the three columns containing outliers.

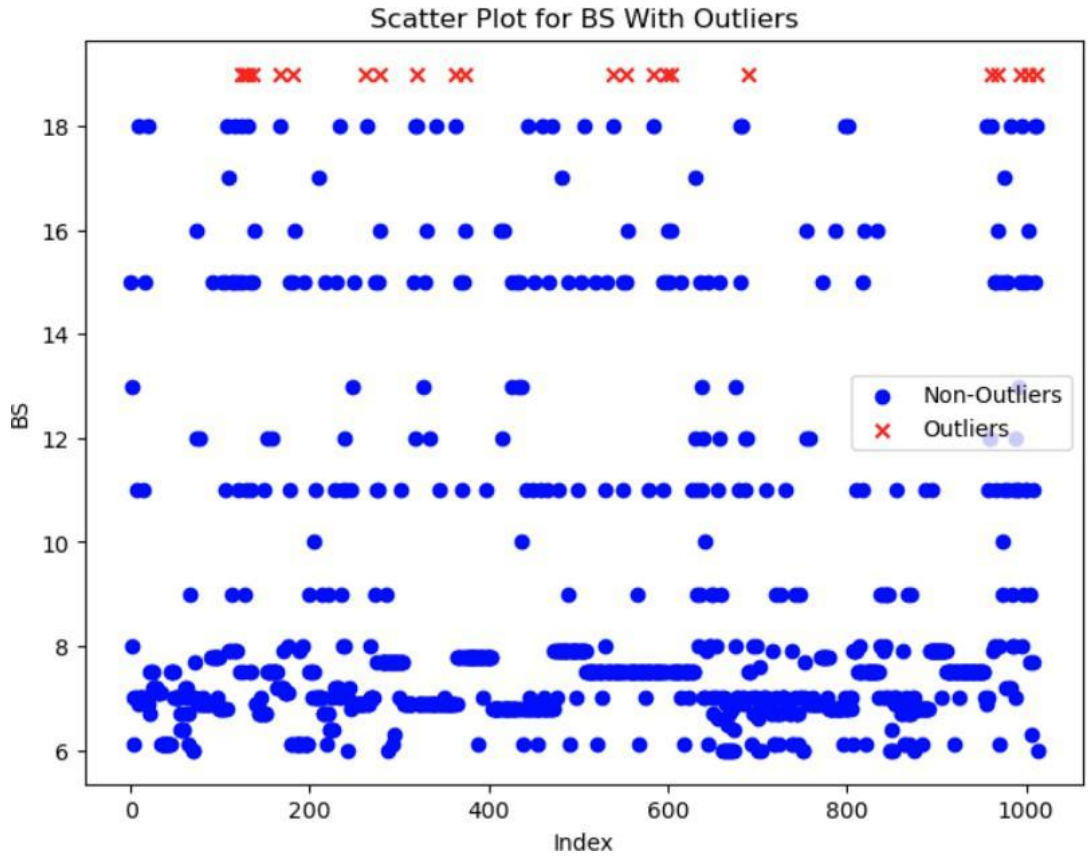


Figure 2: Scatter Plot for BS containing outliers

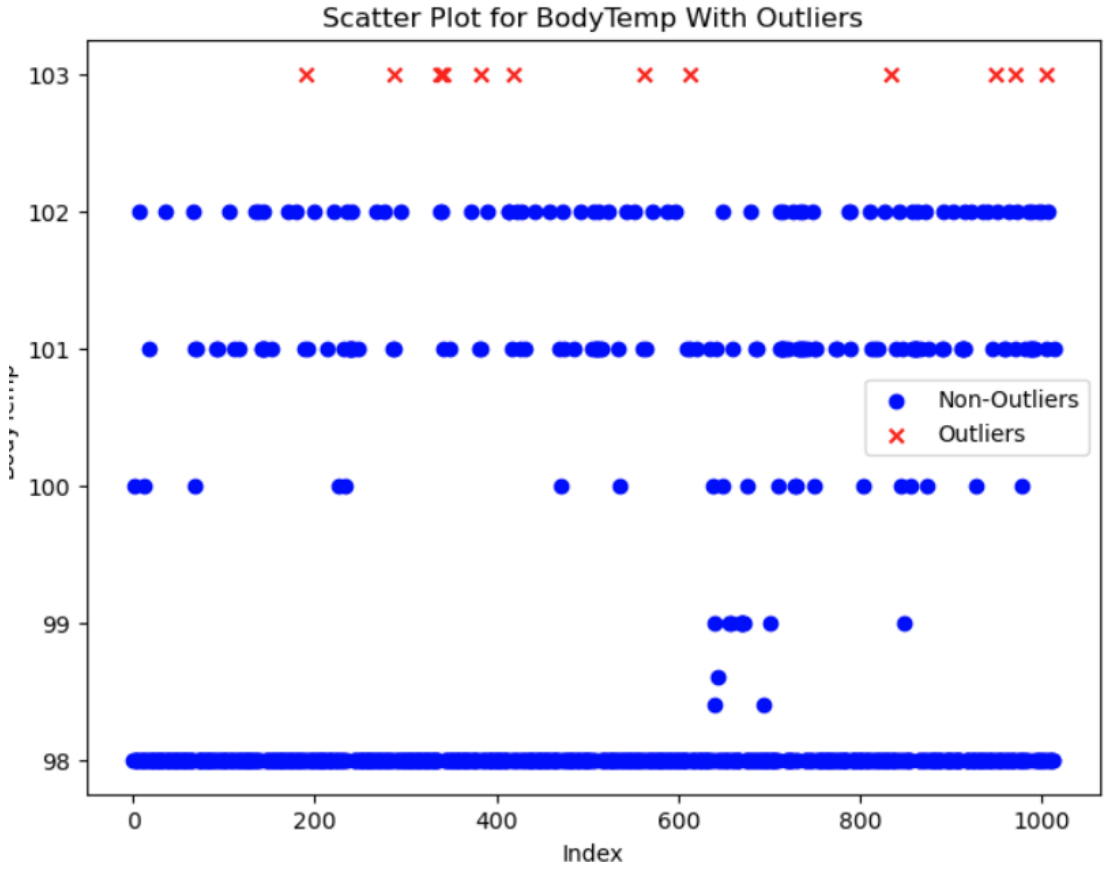


Figure 3: Scatter Plot for Body Temp containing outliers

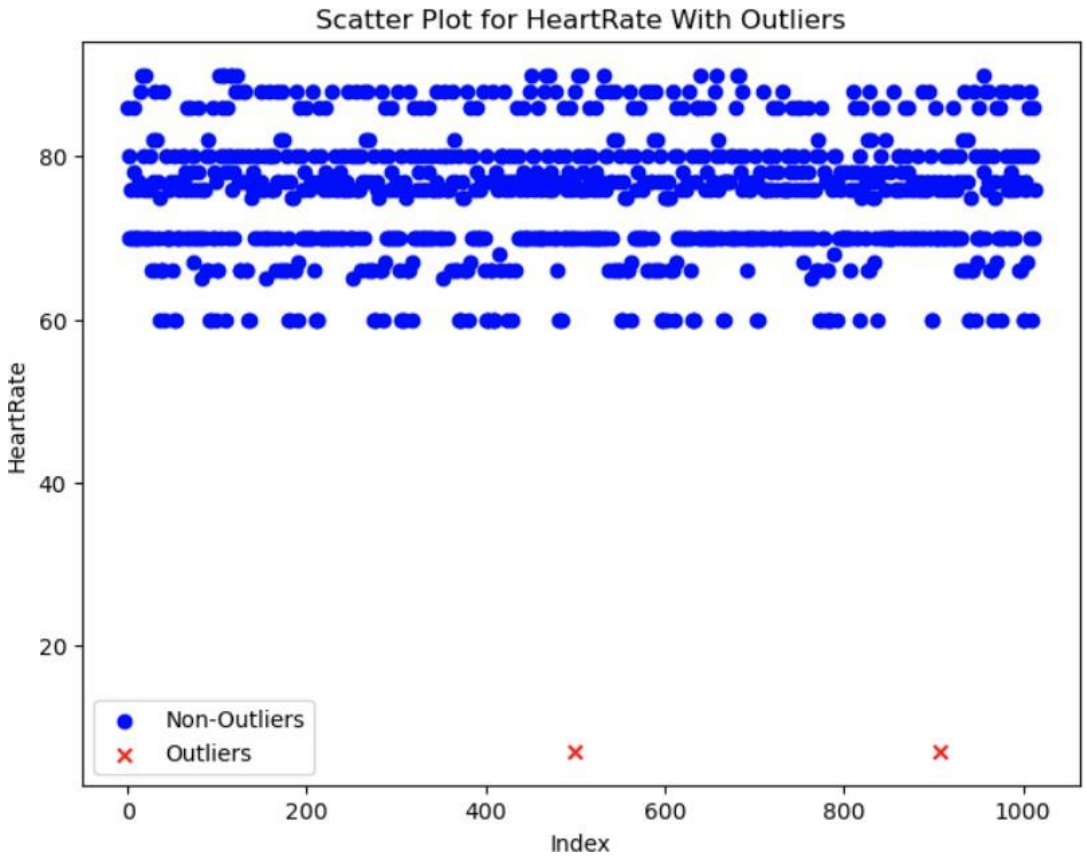


Figure 4: Scatter Plot for Heart Rate containing outliers

The scatterplots above display the columns of the dataset which contains outliers. BS, BodyTemp and Heart Rate. The only outliers which make logical sense to remove is the outliers in HeartRate as having a HeartRate significantly below 60 sugests that the individual would have a severe lack of oxygyn. It is also because there are only two outliers showing in the HeartRate column, which shouldn't make a significant impact to the data if they were to be removed. Therefore I have removed the rows with outliers in the HeartRate column. Below is a Figure displaying the scatterplot with the outliers removed.

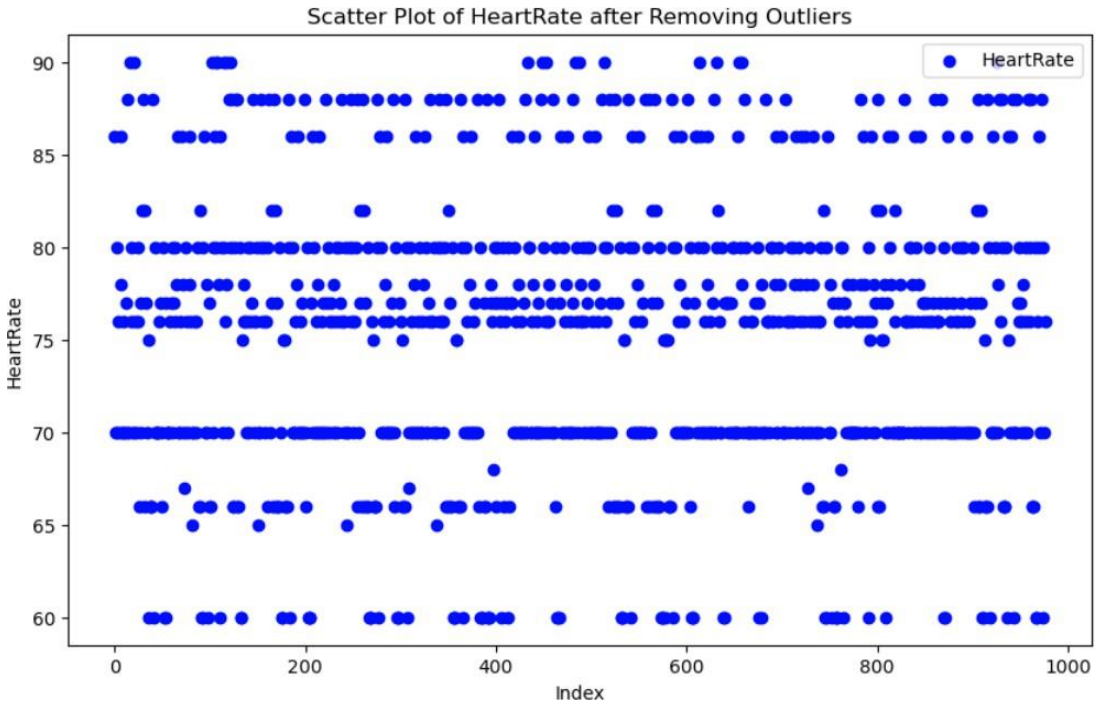


Figure 6: Scatter Plot for HeartRate with outliers removed

Table 3: Updated z-score table once alll outlieirs are removed

```
Outliers: 35
Outliers Status
Age            False
SystolicBP     False
DiastolicBP    False
BS              True
BodyTemp        True
HeartRate      False
dtype: bool
```

Table 4: Checking for number of duplicate rows

```
Number of duplicate rows: 561
Duplicate rows:
      Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel
67     19         120           80   7.0      98.0         70   mid risk
72     19         120           80   7.0      98.0         70   mid risk
97     19         120           80   7.0      98.0         70   mid risk
106    50         140           90  15.0      98.0         90  high risk
107    25         140          100   6.8      98.0         80  high risk
...   ...         ...          ...   ...       ...        ...        ...
1009   22         120           60  15.0      98.0         80  high risk
1010   55         120           90  18.0      98.0         60  high risk
1011   35          85           60  19.0      98.0         86  high risk
1012   43         120           90  18.0      98.0         70  high risk
1013   32         120           65   6.0     101.0         76   mid risk

[561 rows x 7 columns]
```

The table above showcases that the number of duplicate rows is 561. Since it is possible to have duplicated in a maternal health risk dataset as age for example can make other factors similar or the same, which is why there are duplicate rows. I have chosen not to remove duplicate rows for this reason. The dataset information stated that there are ***no missing values***.

Table 5: Checking for Data Types

```
Data Types:
Age              int64
SystolicBP       int64
DiastolicBP      int64
BS             float64
BodyTemp       float64
HeartRate        int64
RiskLevel       object
dtype: object
```

The table above showcases the data types. This tables helps with indicating whether normalization should occur. RiskLevel is identified as an object data type, as the data is a text (low risk, mid risk and high risk). I have decided not to normalise the data as the features, Age, SystolicBP, DiastolicBp, BodyTemp and Heart Rate are all on a similar scale with values that are within a reasonable range for each variable.

Illustrate the features of your dataset using meaningful boxplots, histograms and grouped scatter plots (remember, these plots allow you to analyse the individual distribution of features and the relationship between them).

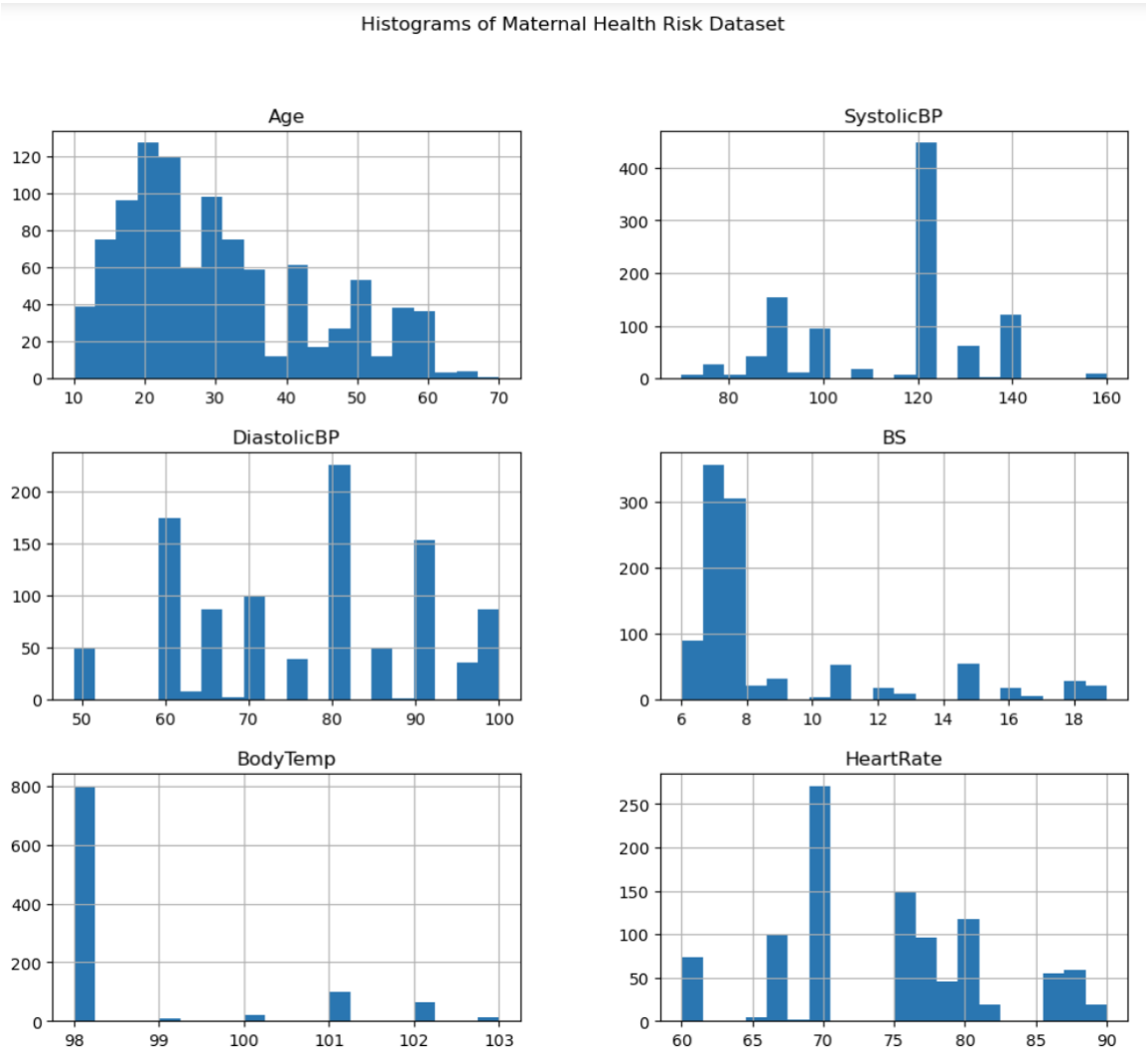


Figure 7: Histogram of numerical columns is maternal health risk dataset

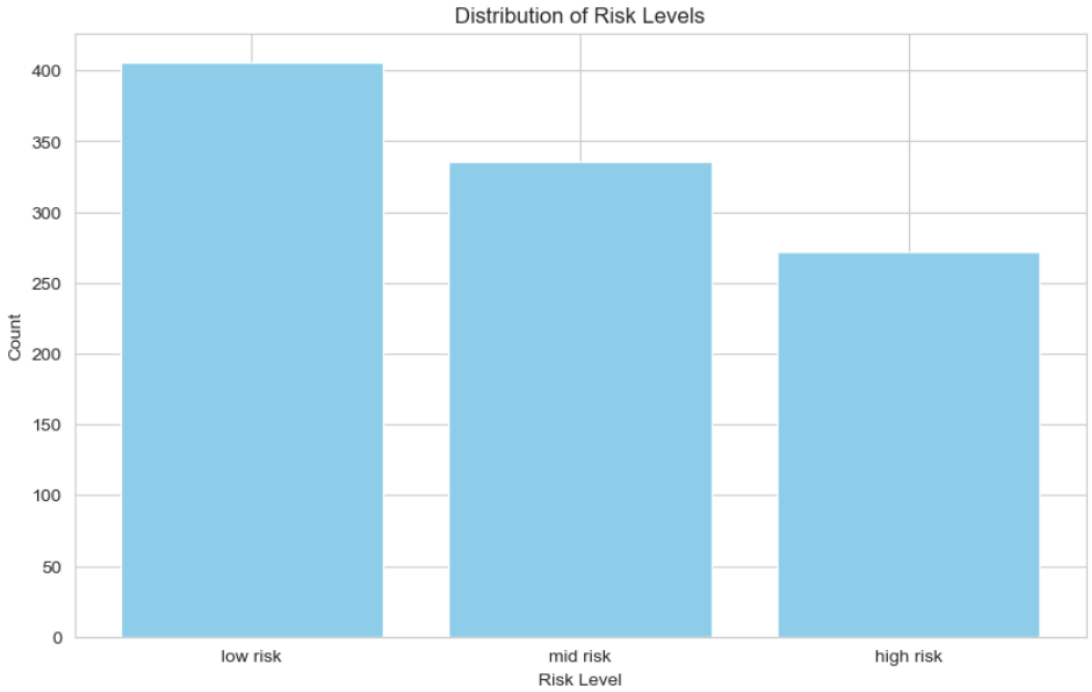


Figure 8: Distribution of Risk Levels Bar Plot

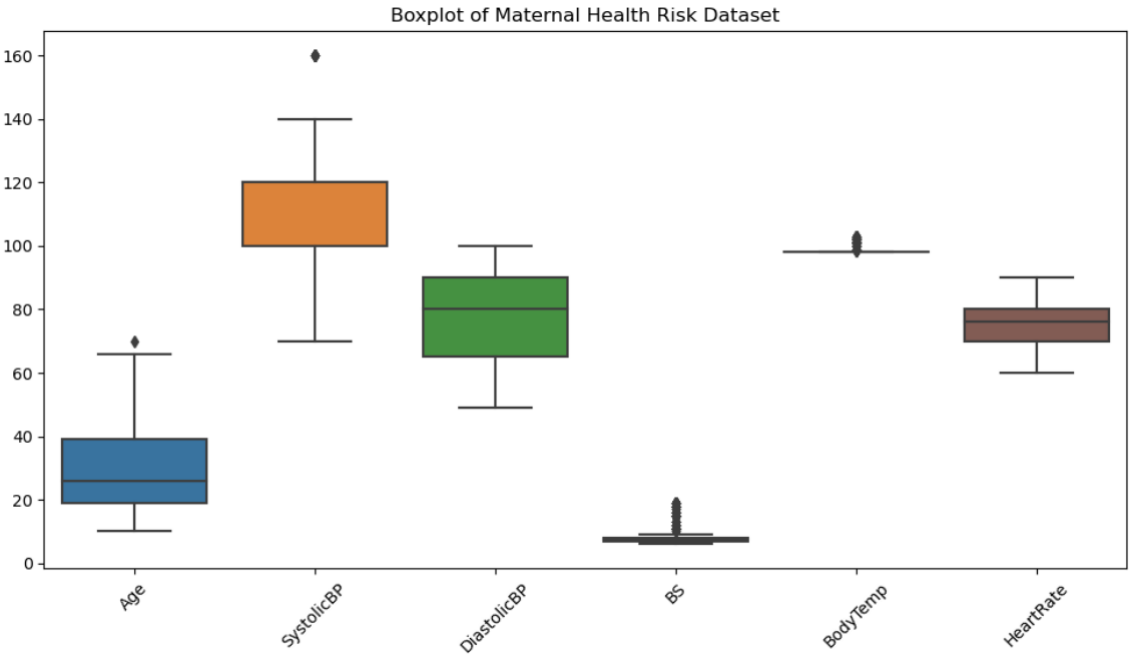


Figure 9: New Boxplot of Dataset

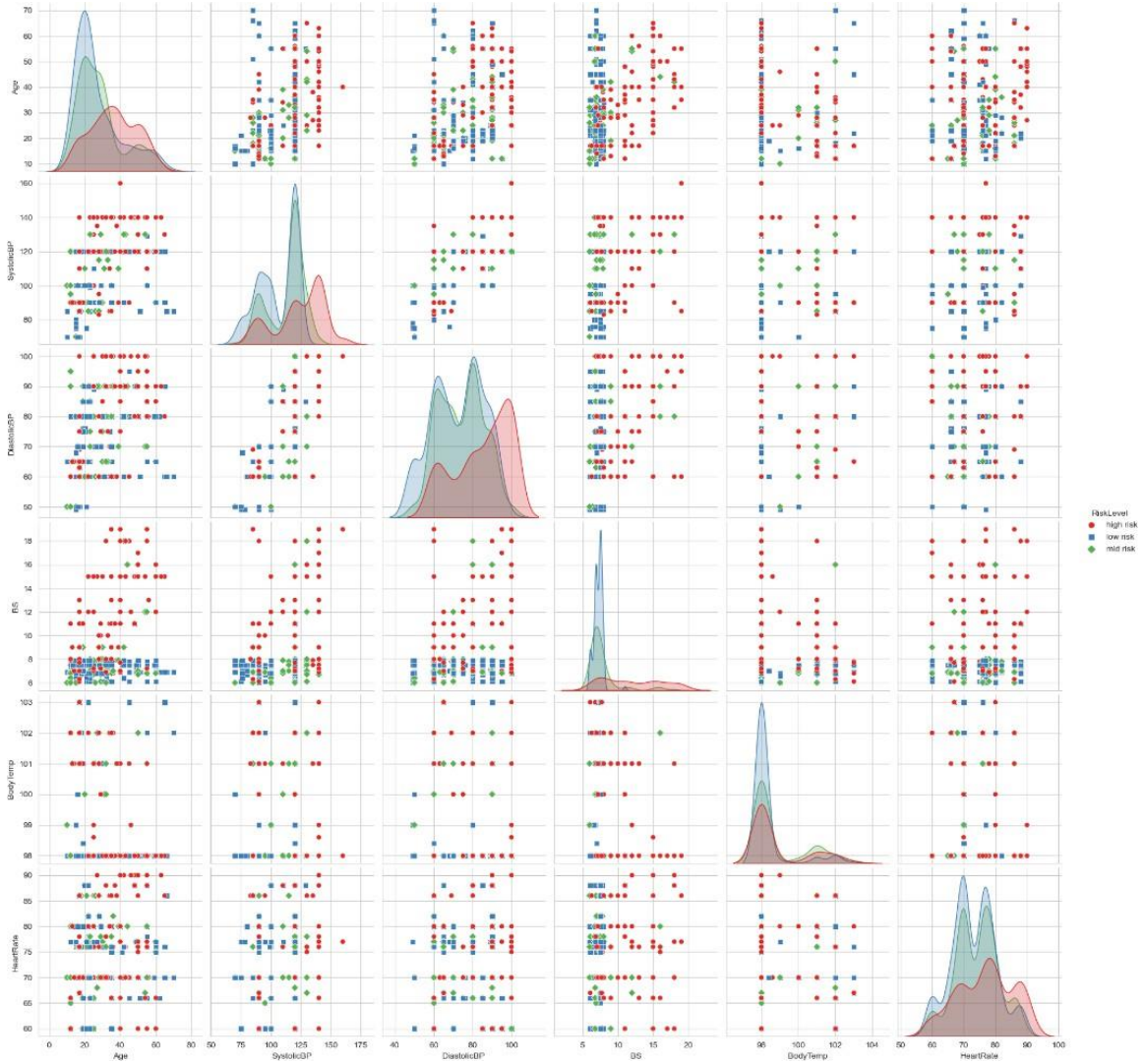


Figure 10: Pairplot displaying individual distribution of features and the relationships between them.

Explain what you can learn from your data exploration and visualisations provided.

From the data exploration and visualization provided, the histograms display the various health metrics. From this we can observe the range and frequency of values for each metric. This information can be helpful when identifying patterns, outliers and correlations between various metrics, which helps to understand and address maternal health risks. In the RiskLevel bar plot, it can be observed that the majority of cases fall under ‘low risk’, whilst there are smaller number of cases in mid risk and high risk risk levels. This plot can also help in identifying trends in the data, such as risk levels over time. From the exploration, I can learn that the dataset contains 561 duplicate rows, likely due to factors such as age causing similarities in other variables, an contained outliers, which was later processed and cleaned.

Task 3: Classification Models *(500-600 words)*

a) You are required to report your preprocessing steps. The steps should include identifying any missing/duplicate data or outliers. Provide explanations of how you dealt with them. [5 marks]

The scatterplots showed outliers in the BS, BodyTemp and HeartRate columns. Removing outliers in HeartRate below 60 was logicial due to their potential health implications. With only two outliers in this column, their removal has minimal impact. The dataset contained 561 duplicate rows, likely due to similar factos such as age. These duplicated were retained. No missing values were found. Normalization was not implemented as all features were within a reasonable range.

b) Create a model using the Decision Tree algorithm. Adjust two suitable parameters (one at a time) to reduce the tree's size and improve your model's accuracy. Report the accuracy score for each parameter using the plots. Provide the final optimised classification tree and describe its structure. [12 marks]

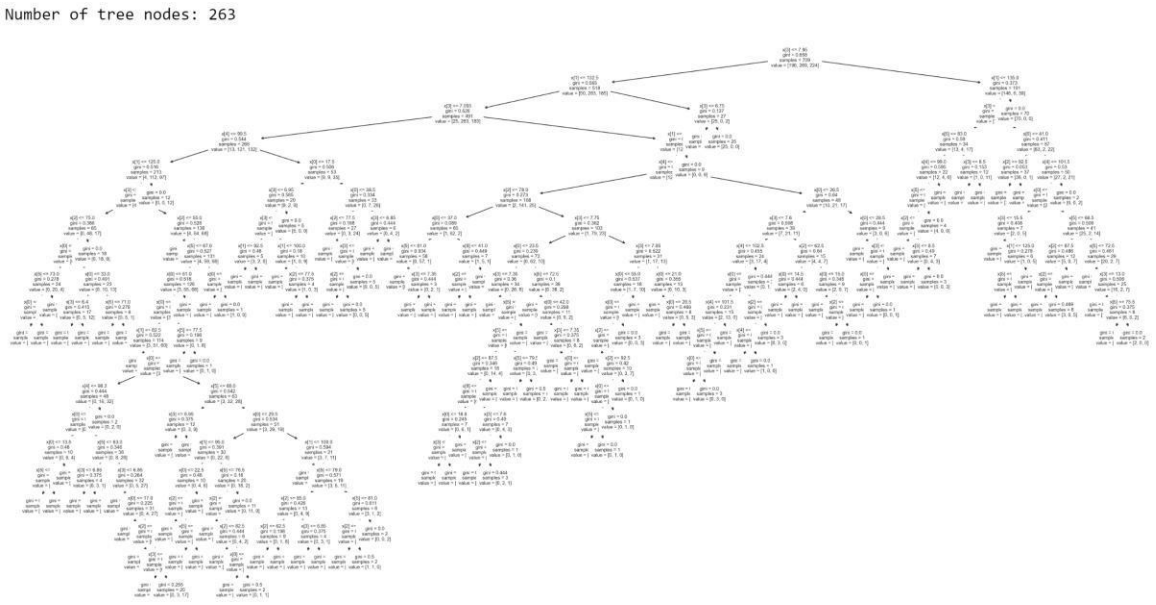


Figure 11: Baseline Decision Tree

```
max_depth=1 Average 10-Fold CV Score:0.5877305377596582 Node count:3
max_depth=2 Average 10-Fold CV Score:0.6271791885070859 Node count:7
max_depth=3 Average 10-Fold CV Score:0.6035818287711123 Node count:13
max_depth=4 Average 10-Fold CV Score:0.6478741991846243 Node count:21
max_depth=5 Average 10-Fold CV Score:0.6646476412347117 Node count:35
max_depth=6 Average 10-Fold CV Score:0.6290526111434673 Node count:49
max_depth=7 Average 10-Fold CV Score:0.6903708017860609 Node count:73
max_depth=8 Average 10-Fold CV Score:0.6932828576975344 Node count:105
max_depth=9 Average 10-Fold CV Score:0.7355562026790914 Node count:139
```

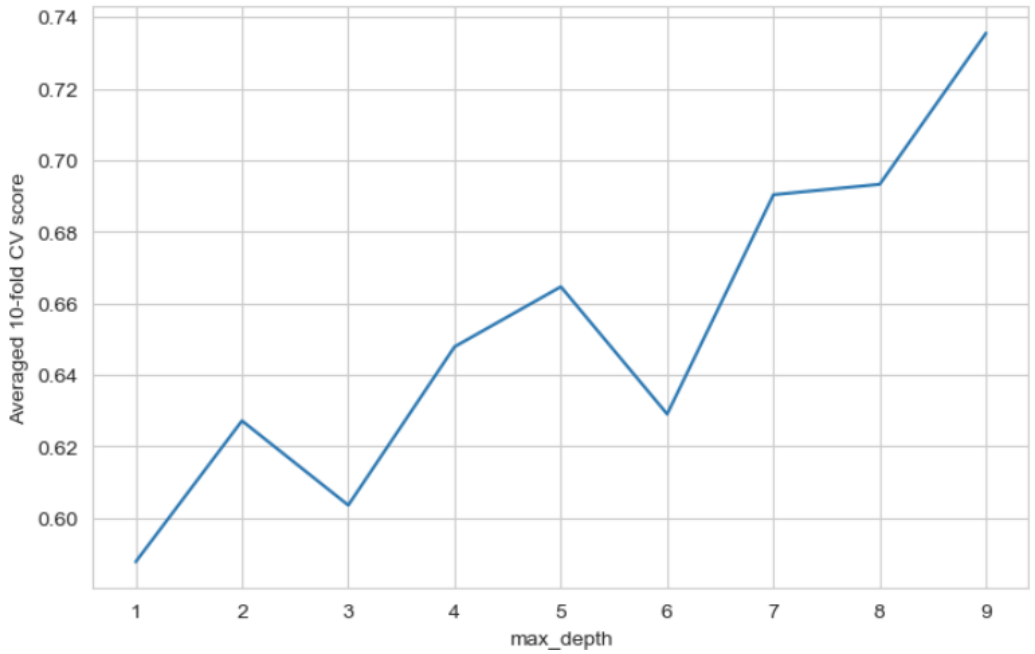
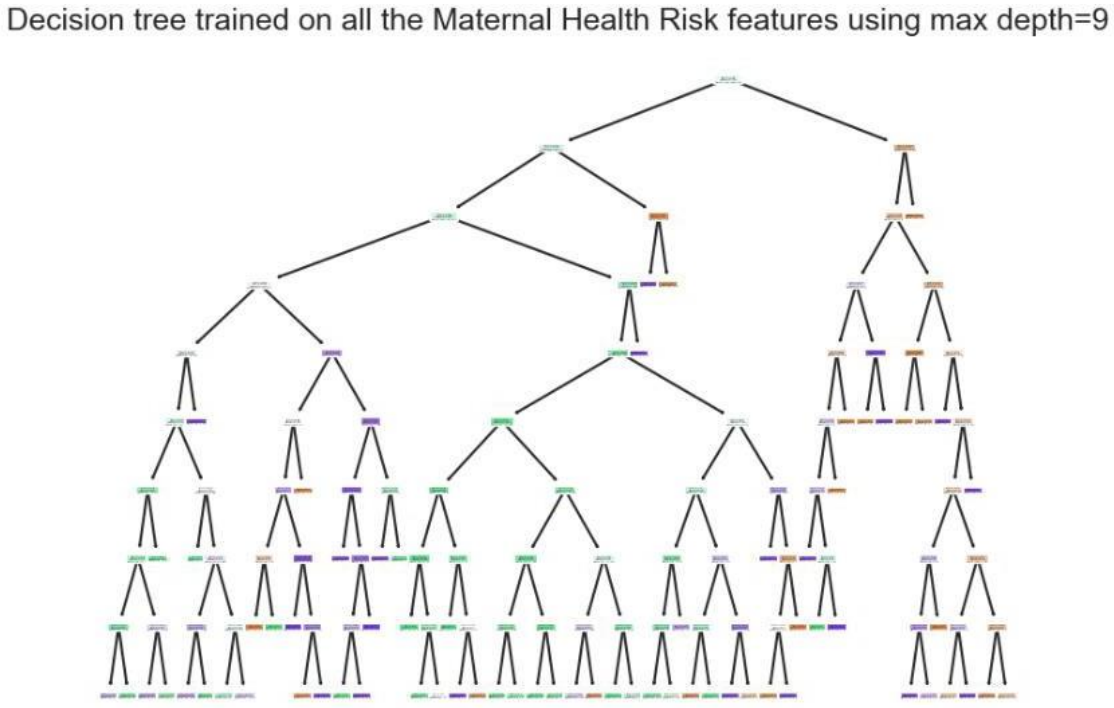
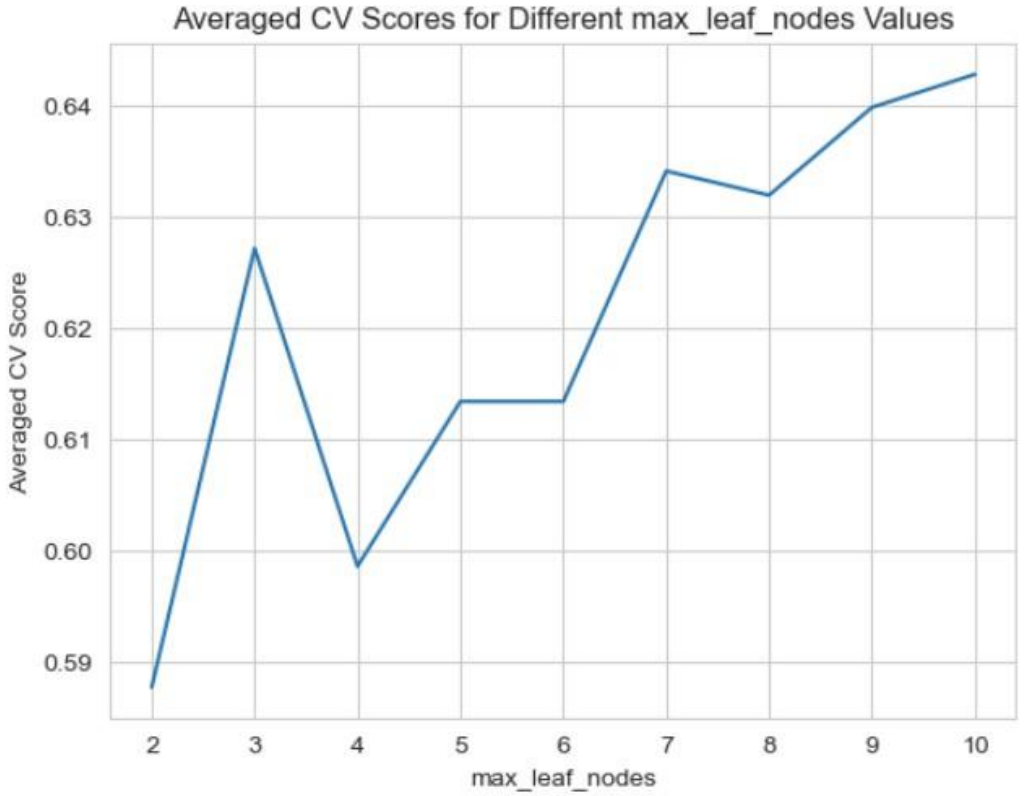


Figure 12: Plot for averaged CV scores for all max_depth tunings




```
Accuracy score of our model with Decision Tree: 0.74
Precision score of our model with Decision Tree : 0.74
Recall score of our model with Decision Tree : 0.74
```

Figure 13: Decision Tree max depth = 9




```
Best max score: 0.6428072218986605
Best leaf node value: 10
```

Figure 14: Tuned model by finding optinal max leaf nodes

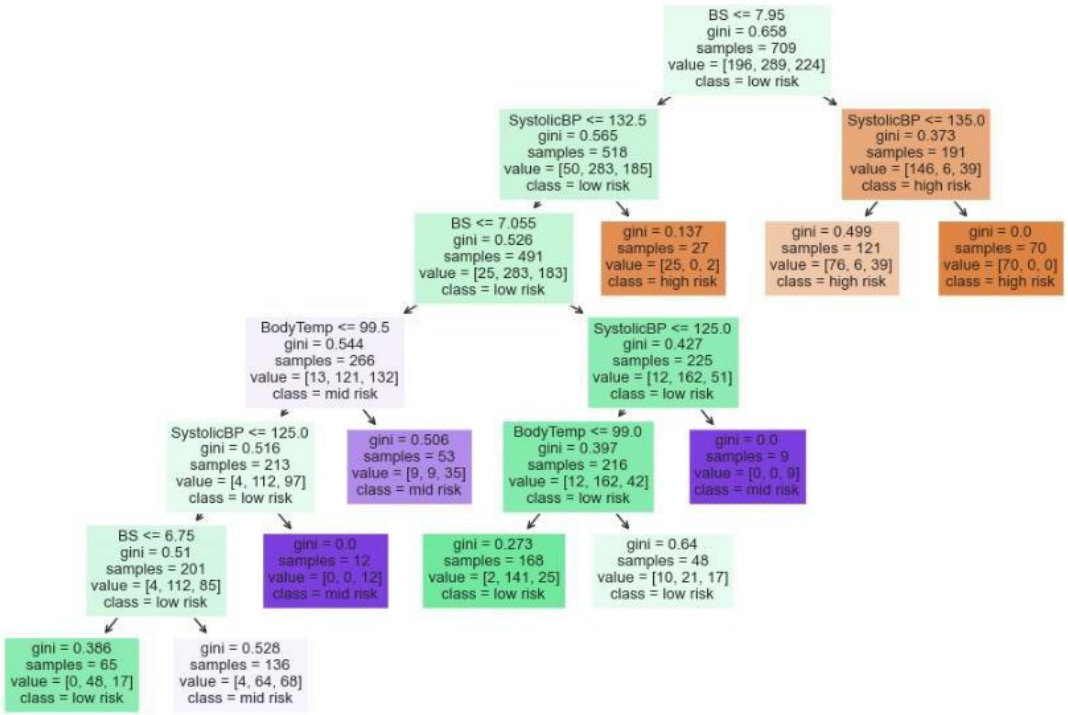


Figure 15: Tuned decision tree

```
Number of nodes in the tuned tree: 19
Accuracy score of our model with Decision Tree: 0.74
Precision score of our model with Decision Tree : 0.74
Recall score of our model with Decision Tree : 0.74
```

The decision tree structure above is used to predict a patients risk level based on their BodyTemp, SystolicBP, and BS.In this scenario, the target is the patients risk level, which is categorised as low-risk, mid-risk and high-risk.

The decision tree contains 19 nodes connected by branches demonstrating its decisions. Each node signifies a decision based on a single feature, with branches portraying the potential outcomes. The leaves indicate final classifications. The decision tree initially divides the dataset using the BS feature, which employs a threshold of 7.95. If BS is less than or equal to 7.95, the tree then divides the data based on Systolic BP with a threshold of less than or equal to 132.5. If the Systolic BP is less than or equal to 132.5, the tree then splits again into BS, which then further splits into BodyTemp. If the BodyTem is less than or equal to 99.5, the tree further divides the data based on SystolicBP feature, with a threshold of 125.0. If the SystolicBp is less than or equal to 125.0, the tree divides the data based on the BS feature, with a threshold of 6.75. Finally, if the BS is less than or equal to 6.75, the tree categories the patient as low risk.

As shown above, the tree uses Gini impurity for splitting. With an accuracy, precision and recall score of 0.74, it provides an interpretable model but can overfit and be sensitive to feature order.

c) Describe the role of the two parameters in the model building you used in part c) above. Do you expect that using the same values obtained for this dataset will improve the accuracy of other datasets? Justify your answer. [8 marks]

The model building parameters include the maximum depth and minimum samples required to split an internal node. Max depth controls the tree's depth which influences its complexity. Higher values lead to more node, lower values to fewer nodes. The minimum samples parameter sets the threshold for further node splitting. Higher values form simpler trues, lower values, more complex. Optimal values. 9 for depth and 2 for samples, were discovers using cross-validation.

Using the same parameters on other datasets is unlikely to improve accuracy due to dataset variability in features, samples etc. Optimal values can vary between datasets. Tuning parameters individually for each dataset and problem is crucial. The techniques used in this

dataset can be applied to other datasets to find optimal valies for their specific characteristics.

d) Find the feature importance based on the final classification model and explain your findings. [5 marks]

Table 6: Feature Importance

	Feature	Importance
3	BS	0.470
1	SystolicBP	0.210
0	Age	0.124
4	BodyTemp	0.095
2	DiastolicBP	0.052
5	HeartRate	0.049

Based on the feature importance table, the most influential variables for the models predictions are SystolicBP, DiastolicBp and BS. SystolicBP is the most impactful, followed by DiastolicBP and then BS. BodyTemp ranks fourth, and HeartRate ranks fitfth. These findings highlight the significant of these variables in assessing heart risk failure (RiskLevel) and dividing preventative strategies.

e) Generate and carefully examine the Confusion Matrix and explain your findings. Provide the model summary report and discuss the metrics (accuracy, precision, recall, and F1- score).

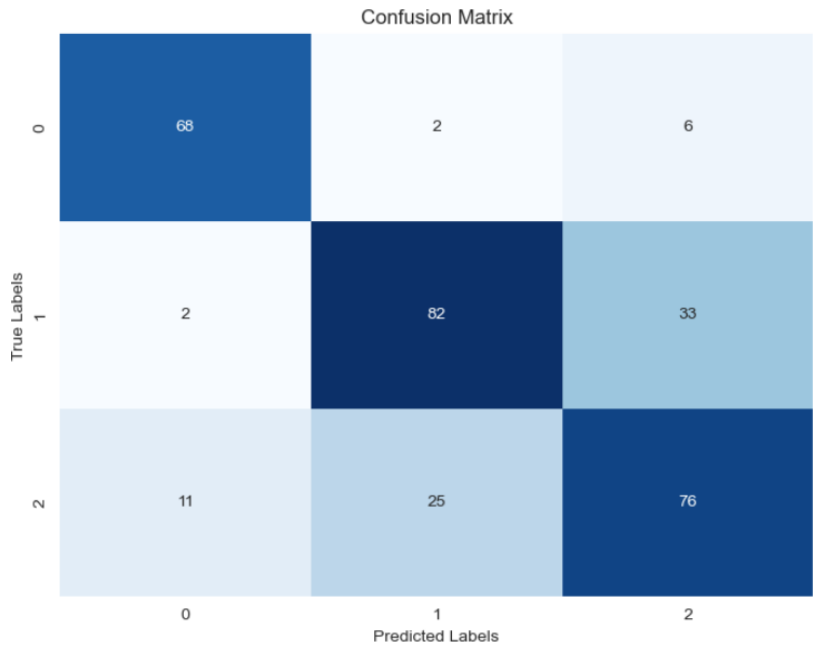


Figure 16: Confusion Matrix Plot

Table 7: Classification Report

```
Accuracy: 0.740983606557377
Precision: 0.7404531444721426
Recall: 0.740983606557377
F1-Score: 0.7401070310995791

Classification Report:
              precision    recall  f1-score   support

   high risk       0.84      0.89      0.87        76
    low risk       0.75      0.70      0.73       117
    mid risk       0.66      0.68      0.67       112

    accuracy                           0.74       305
   macro avg       0.75      0.76      0.75       305
weighted avg       0.74      0.74      0.74       305
```

Based on the confusion matrix, the classification algorithm appears to be performing well in predicting the positive class, with a high number of true positives (68) and a relatively low number of false negatives (11). The classification report is used to predicting risk levels. The overall accuracy of 0.741 indicated that it correctly predicted the risk level for 74.1% samples. The precision for the high risk was highest at 0.84, showcasing that when the model predicted a sample as high risk, it was correct 84% of the time. The recall for the high risk class was highest at 0.89, highlighting that the model identified 89% of the actual high risk instances. The F11 score for the high risk is 0.87, indicating a good balance between precision and recall. Overall the model performed reasonably well across all classes, with a slightly lower performance for the mid risk class.

Task 4: Results and Discussions *(300-400) words*

The analysis of the classification report provides valuable insights onto the performance of the classification algorithm in predicting maternal health risk levels. The high number of true positives (68), and the relatively low number of false negatives (11), suggest that the algorithm is performing well in identifying high risk cases, which is crucial for implementing targeted interventions to reduce maternal morality rates.

The overall accuracy of 0.741 indicates that the model correctly predicted the risk level for a significant portion of samples. While this is a decent result, it also suggests room to further improve, particularly in predicting mid risk cases. Further investigation into the factors

contributing to the misclassification of mid risk cases could have helped improve the models performance in this area.

The precision, recall and F1 scorer for the high risk class are particularly significant. The precision of 0.84 highlights that when the models predicted a sample as high risk, it was correct 84% of the time. The recall of 0.89 indicates that the model identified 89% of the actual high risk instances, the F1 score of 0.87 demonstrates a good balance between precision and recall for the high risk class, indicating that the models predictions for high risk cases are both accurate and comprehensive.

These results have a significant implication for healthcare strategies which are aimed at reducing maternal mortality rates in Bangladesh. The ability of the classification algorithm to accurately identify high risk cases suggest that it can be a valuable tool for targeting actions which can be taken such as early screening and monitoring, for women with a high risk for maternal death. By focusing on resources and attention on these high risk cases, healthcare workers can potentially reduce maternal death rates and improve on the maternal health outcomes.

In conclusion, whole the classification algorithm shows potential in predicting maternal health risk levels, there can still be improvements made, particularly in predicting mid risk cases. Further refinement, possibly through the addition of variables or more advanced techniques could improve its performance and be a valuable tool for assessing and addressing maternal health risks in Bangladesh.