Student ID Number: 22180242

Assignment 1

Data Exploration and Classification

Semester 1 2024

Student Name: Juchang Kim
Student ID: 22180242
PAPER NAME: Foundations of Data Science

PAPER CODE: COMP615

Due Date: Sunday 14 April 2024 (midnight)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section.**

Task 1: Introduction (100-200 words) [10 marks]

Answer:

The dataset selected for this assignment is the Maternal Health Risk dataset. Maternal mortality is a significant concern globally, and identifying risk factors associated with it is crucial for improving maternal healthcare outcomes.

The aim of this work is to analyze the dataset to understand the relationship between various risk factors and the predicted risk intensity level during pregnancy. The research questions we attempt to answer include:

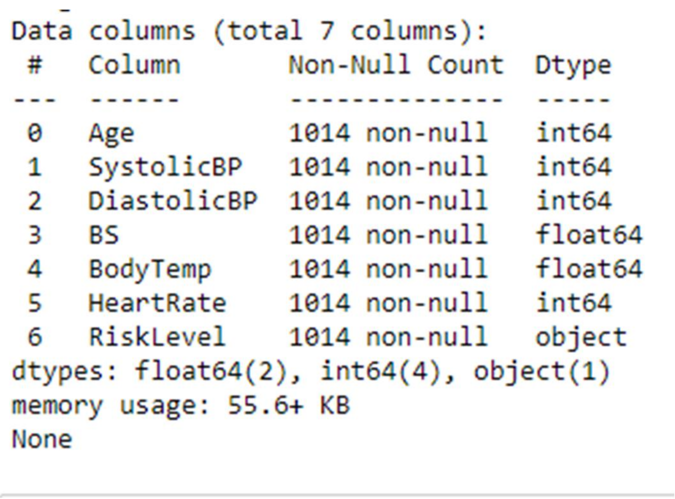
- What are the main risk factors associated with maternal health risk during pregnancy?
- How do these risk factors correlate with the predicted risk intensity level?
- Can we develop a classification model to predict the risk intensity level based on the given risk factors?

Assumptions:

- The dataset provides accurate and reliable measurements of the risk factors.
- The risk intensity level prediction is based on clinically validated criteria.

Task 2: Data Exploration (500-600 words) [20 marks]

Answer:

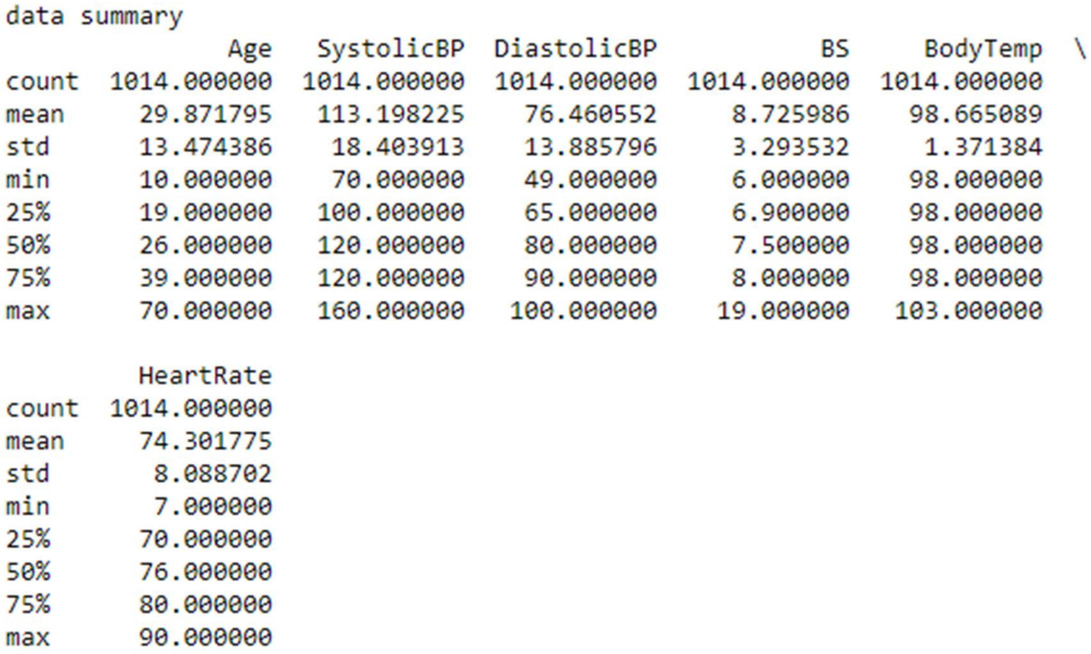


```
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Age          1014 non-null   int64  
 1   SystolicBP   1014 non-null   int64  
 2   DiastolicBP  1014 non-null   int64  
 3   BS           1014 non-null   float64
 4   BodyTemp     1014 non-null   float64
 5   HeartRate    1014 non-null   int64  
 6   RiskLevel    1014 non-null   object 
dtypes: float64(2), int64(4), object(1)
memory usage: 55.6+ KB
None
```

- Features and Instances:
  The features include Age, Systolic Blood Pressure (SystolicBP), Diastolic Blood Pressure (DiastolicBP), Blood Sugar (BS), Body Temperature (BodyTemp), HeartRate, and RiskLevel. Also, there are 1014 rows without null value rows.

- Data Types:
  The Age, SystolicBP, DiastolicBP, HeartRate are integer type, BS, BodyTemp are float type, the target variable RiskLevel is categorical. So, all attributes are numerical type without RiskLevel, while RiskLevel is a categorical attribute which is target value.

- Summary Statistics of Numerical Attributes

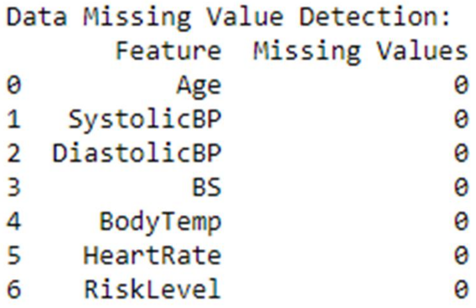


```
data summary
              Age   SystolicBP  DiastolicBP           BS     BodyTemp  \
count 1014.000000  1014.000000  1014.000000  1014.000000  1014.000000
mean    29.871795   113.198225    76.460552     8.725986    98.665089
std     13.474386    18.403913    13.885796     3.293532     1.371384
min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

        HeartRate
count 1014.000000
mean    74.301775
std      8.088702
min      7.000000
25%     70.000000
50%     76.000000
75%     80.000000
max     90.000000
```

It shows the summary of statistics of each numerical attributes there attributes' count, mean, standard deviation (std), minimum, 1st quartile (25%), media (50%), 3rd quartile (75%) and maximum. It can provide to understand each numerical attribute information overall.

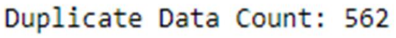
- Data Cleanliness Evaluation Step

First step is identifying missing data. Because those are cause building inaccurate models and should be removed. Then, it needs to detect duplicated data and decide how to deal with the duplicated data. As well as strategies for data cleaning may include outlier detection and removal, normalization or standardization of numerical features, and encoding categorical variables if necessary.



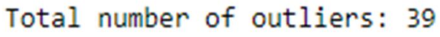
```
Data Missing Value Detection:
       Feature  Missing Values
0          Age               0
1   SystolicBP               0
2  DiastolicBP               0
3           BS               0
4     BodyTemp               0
5    HeartRate               0
6    RiskLevel               0
```

Handling missing value – There is no missing value. So, it does not need to have removing missing value process.



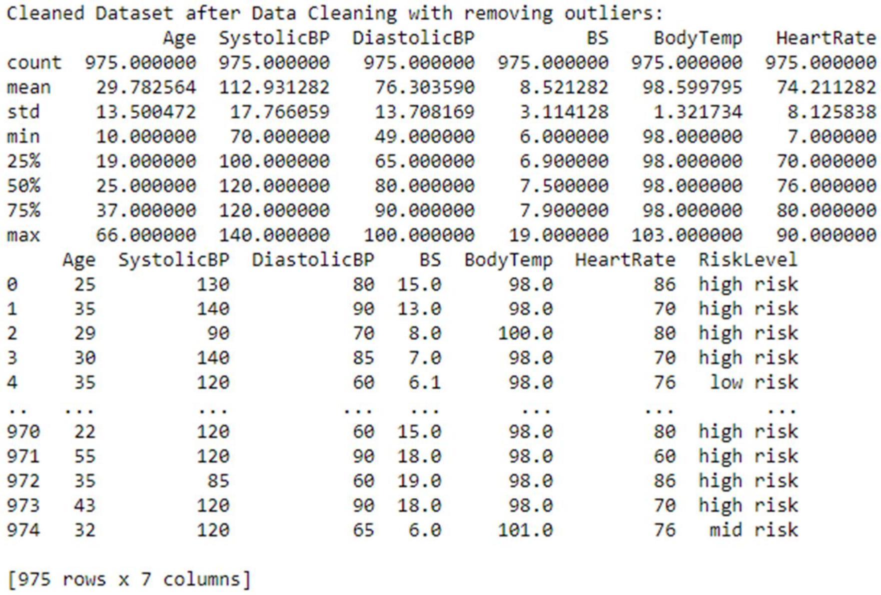
```
Duplicate Data Count: 562
```

Handling duplicated values – 562 data are duplicated. That amount is almost half of whole data amont. So, remaining the duplicated data should be better because if those are removed, there are lots of data loss and it makes inaccurate models.



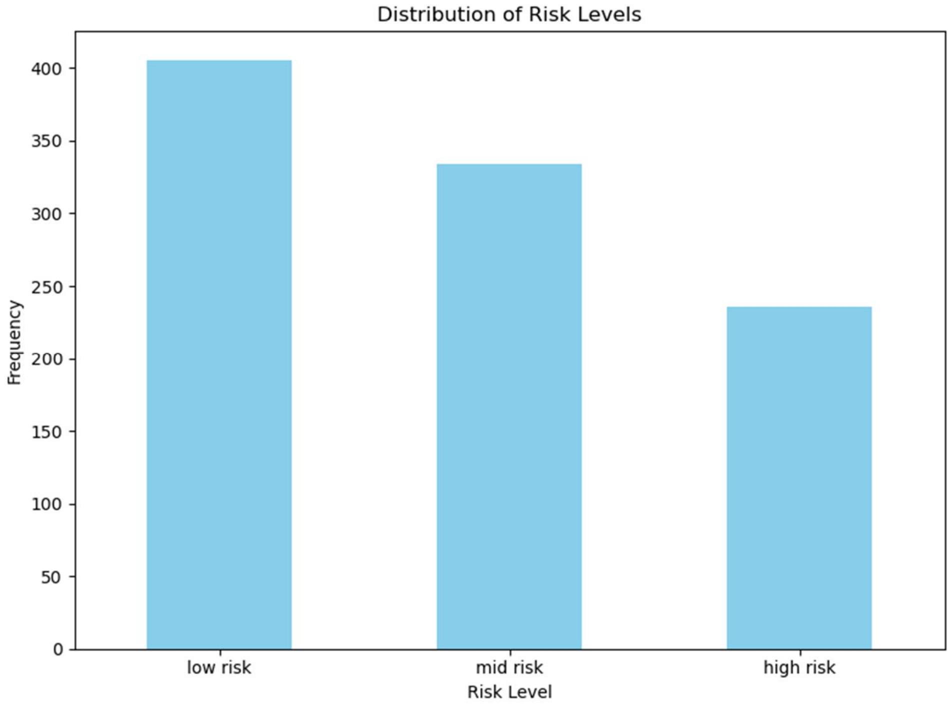
```
Total number of outliers: 39
```

Handling outliers – detecting outliers and remove to make clear dataset by using IQR method. So, the cleaned dataset has the 975 rows.



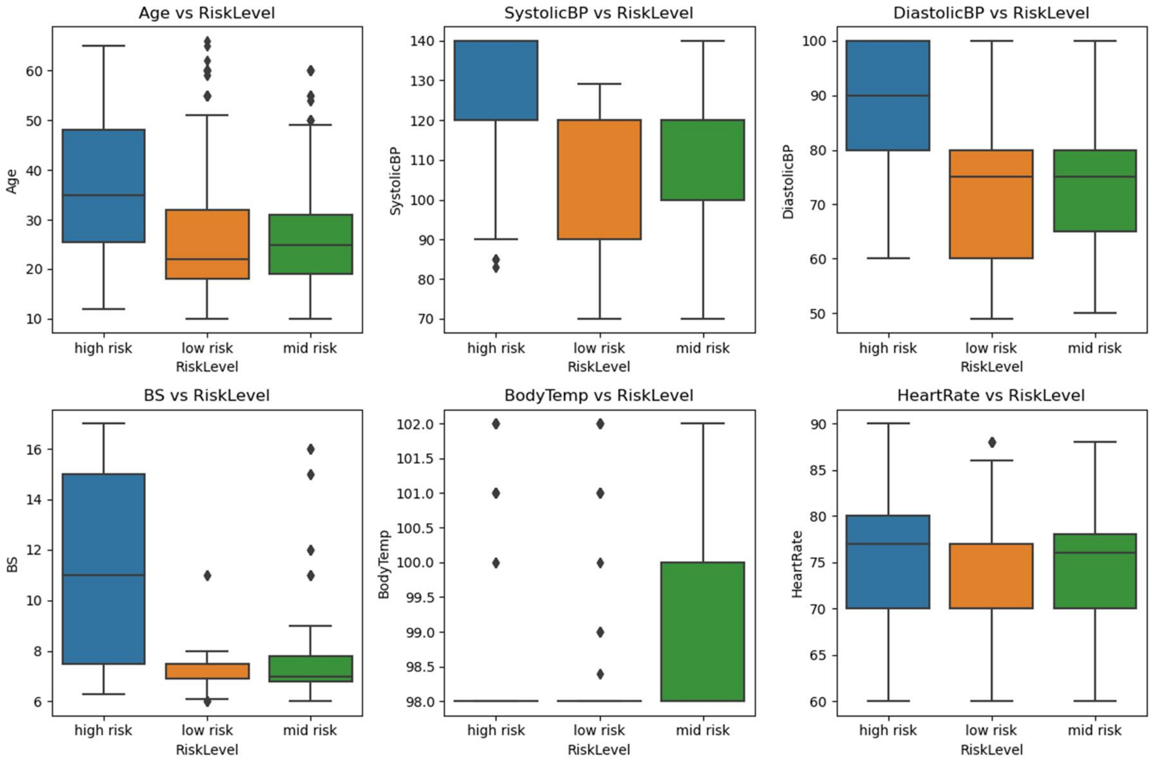
```
Cleaned Dataset after Data Cleaning with removing outliers:
              Age  SystolicBP  DiastolicBP          BS    BodyTemp   HeartRate
count  975.000000  975.000000   975.000000  975.000000  975.000000  975.000000
mean    29.782564  112.931282    76.303590    8.521282   98.599795   74.211282
std     13.500472   17.766059    13.708169    3.114128    1.321734    8.125838
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%     25.000000  120.000000    80.000000    7.500000   98.000000   76.000000
75%     37.000000  120.000000    90.000000    7.900000   98.000000   80.000000
max     66.000000  140.000000   100.000000   19.000000  103.000000   90.000000
     Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel
0     25         130           80  15.0      98.0         86  high risk
1     35         140           90  13.0      98.0         70  high risk
2     29          90           70   8.0     100.0         80  high risk
3     30         140           85   7.0      98.0         70  high risk
4     35         120           60   6.1      98.0         76   low risk
..   ...         ...          ...   ...       ...        ...        ...
970   22         120           60  15.0      98.0         80  high risk
971   55         120           90  18.0      98.0         60  high risk
972   35          85           60  19.0      98.0         86  high risk
973   43         120           90  18.0      98.0         70  high risk
974   32         120           65   6.0     101.0         76   mid risk

[975 rows x 7 columns]
```

Class Distribution (Risk Level):



- This is Distribution of Risk Level (class). In this dataset the low risk frequency is about 400, mid risk of frequency is around 325, high risk of frequency is about 250. So, low risk are most and follows mid risk, high risk.

Visualization (Box Plots):



- Age: The distribution of age appears similar across low risk level and mid risk level. The median age is around 25 years old in low and mid risk levels, while median age is more than 30 in the high-risk group. There are a few outliers on the younger side for all risk levels.

- Systolic Blood Pressure (SystolicBP): The median SystolicBP is highest in the high-risk group, followed by medium and then low risk. There are also more outliers in the high-risk group for SystolicBP.

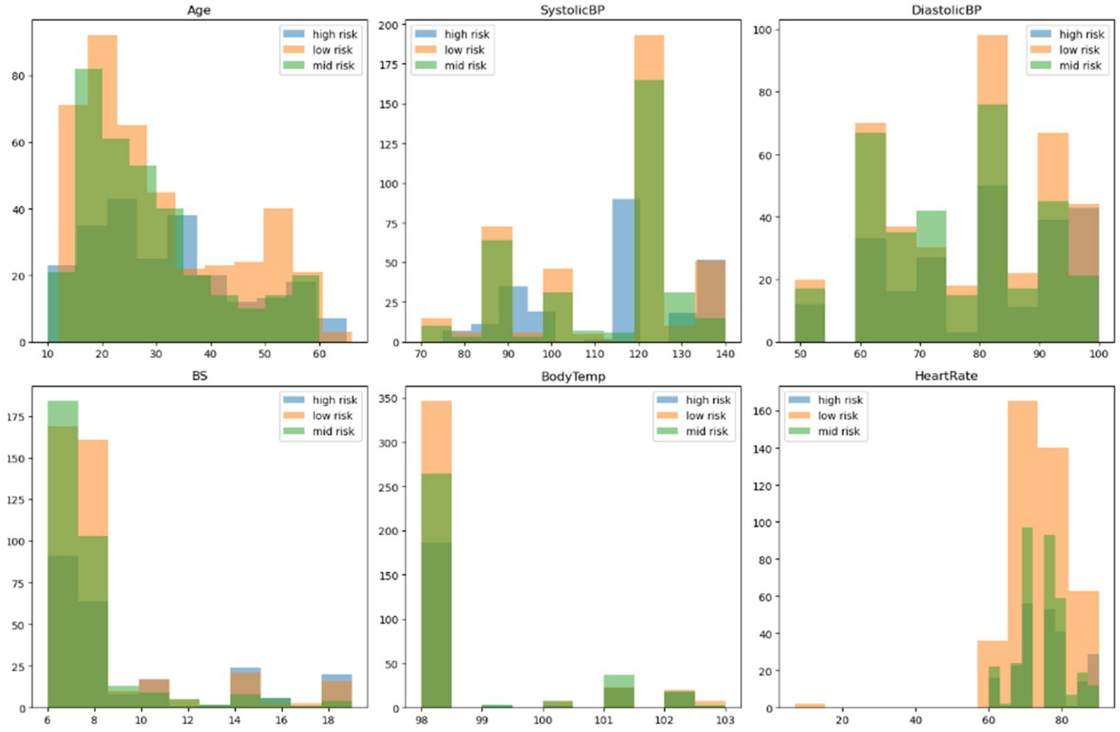
- Diastolic Blood Pressure (DiastolicBP): Similar to SystolicBP, the median DiastolicBP is highest in the high-risk group, followed by medium and then low risk. There are also more outliers in the high-risk group for DiastolicBP.

- Blood Sugar (BS): The distribution of BS appears similar across low and mid risk levels. The median BS is low in mid and low risk levels, However the median BS more than 10 in the high-risk group. There are a few outliers on both the low and mid risk group.

- Body Temperature (BodyTemp): It is hard to define the median across all risk levels. Mid risk group show higher than others high risk and low risk have some outliers.

- Heart Rate (HeartRate): The median HeartRate is highest in the high-risk group, followed by medium and then low risk.

  Visualization (Histograms):



- Age: The distribution of age appears similar across all three risk levels. There's a peak around 20 years old, with a slightly wider spread in the high-risk group compared to medium and low risk.

- Systolic Blood Pressure (SystolicBP): Generally, values are shifted to right side and low risk and mid risk group data amount are more than high risk group.

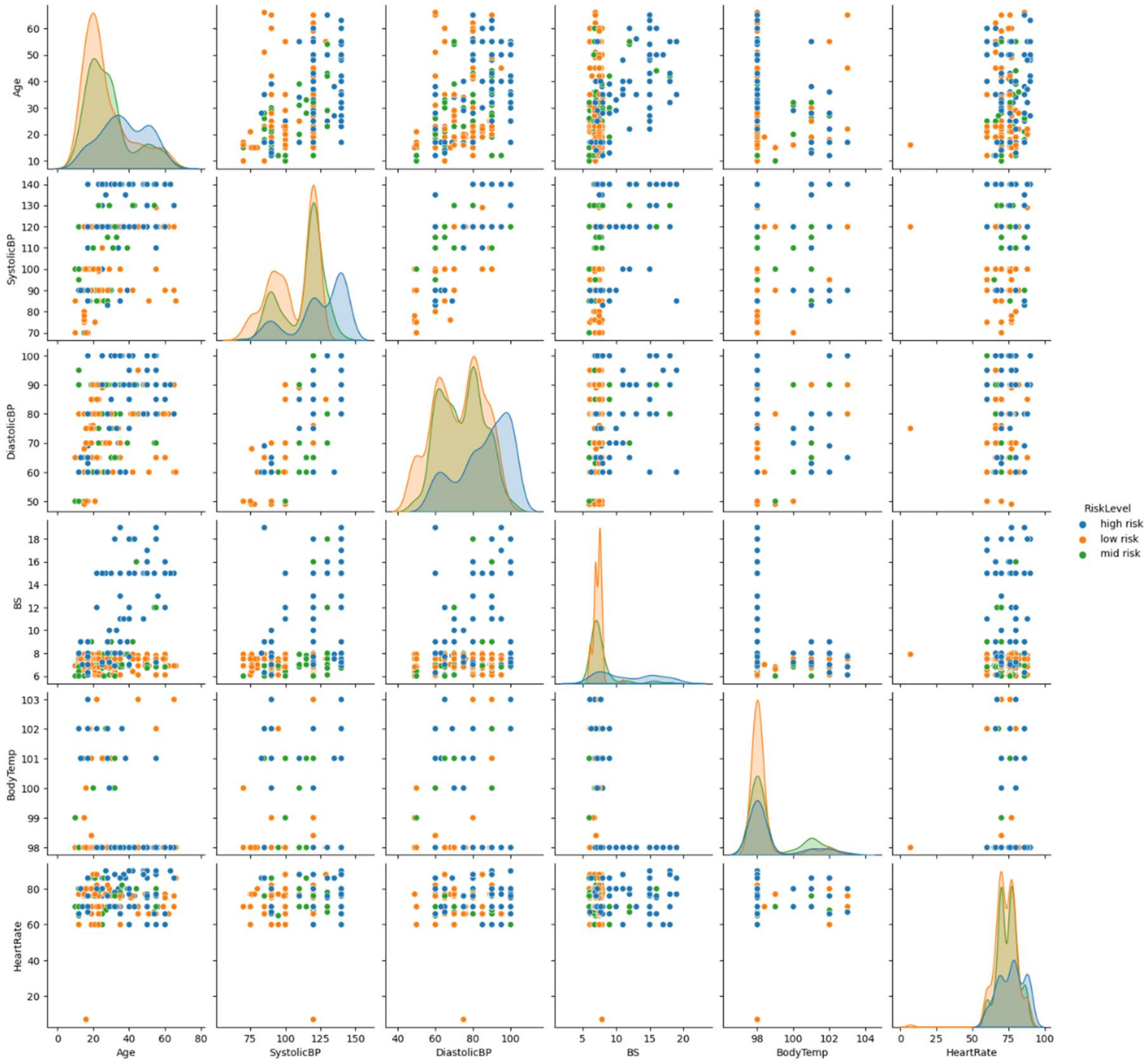
- Diastolic Blood Pressure (DiastolicBP): Similar to SystolicBP, the distribution of DiastolicBP is shifted to the right side and the high-risk group, indicating a tendency for higher diastolic blood pressure readings as well.

- Blood Sugar (BS): The distributions of BS appear fairly similar across all three risk levels, but high-risk group are wider spreaded.

- Body Temperature (BodyTemp): The distributions of BodyTemp also appear similar across all three risk levels, with a peak around 98.5 degrees Fahrenheit.

- Heart Rate (HeartRate): The distribution of HeartRate is shifted to the right (higher values) for the high-risk group compared to medium and low risk. This suggests that pregnant women in the high-risk group tend to have higher heart rates.

Visualization (Grouped Scatter Plot):



- Age vs. Systolic Blood Pressure (SystolicBP): There appears to be a weak positive correlation, where systolic blood pressure tends to increase slightly with age. However, the spread in the data makes it difficult to see a clear pattern.

- Age vs. Diastolic Blood Pressure (DiastolicBP): Similar to SystolicBP, there's a possible weak positive correlation between age and diastolic blood pressure.

- SystolicBP vs. DiastolicBP: As expected, there's a positive correlation between systolic and diastolic blood pressure. This means that women with high systolic blood pressure also tend to have high diastolic blood pressure.

Summary and result of data visualization:

- Age and Risk Level:

The distribution of age appears similar across all three risk levels, with a peak around 20 years old. However, there's a slightly wider spread in the high-risk group compared to medium and low risk.

The median age is around 25 years old in the low and mid-risk groups, while it's more than 30 in the high-risk group. This suggests that older pregnant women may be at a higher risk of maternal health complications.

- Blood Pressure (SystolicBP and DiastolicBP) and Risk Level:

Both systolic and diastolic blood pressure show higher median values in the high-risk group compared to medium and low risk.

There are more outliers in the high-risk group for both systolic and diastolic blood pressure, indicating potential health concerns.

- Blood Sugar (BS) and Risk Level:

The distribution of blood sugar appears fairly similar across all three risk levels, but the high-risk group has a wider spread.

The median blood sugar is higher in the high-risk group compared to medium and low risk, suggesting potential metabolic issues in high-risk pregnancies.

- Body Temperature (BodyTemp) and Risk Level:

The distributions of body temperature are similar across all three risk levels, with a peak around 98.5 degrees Fahrenheit. However, the mid-risk group shows slightly higher values than others.

There are some outliers in both the high and low-risk groups, indicating potential variations in body temperature among pregnant women.

- Heart Rate (HeartRate) and Risk Level:

The median heart rate is highest in the high-risk group, followed by medium and then low risk.

Pregnant women in the high-risk group tend to have higher heart rates compared to medium and low risk, which may indicate increased cardiovascular stress.

- Relationships between Features:

Age and blood pressure (both systolic and diastolic) show weak positive correlations, suggesting that older pregnant women may tend to have higher blood pressure readings.

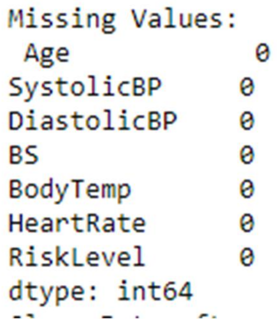
There's also a positive correlation between systolic and diastolic blood pressure, indicating that women with high systolic blood pressure tend to have high diastolic blood pressure as well.

Task 3: Classification Models [40 marks]

Answer:

a) The preprocess step.

1. Identifying Missing Data:



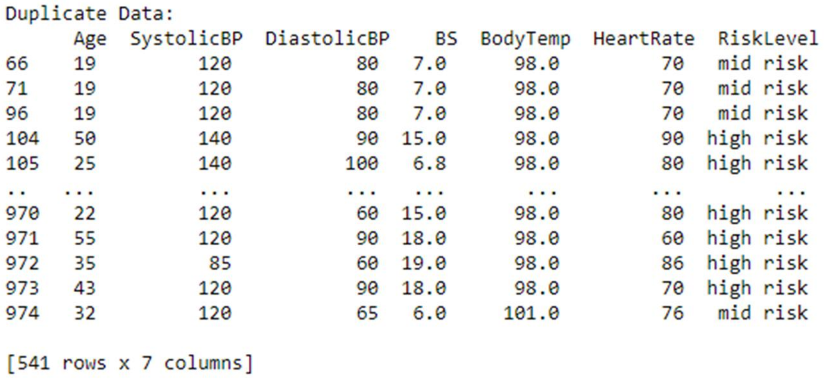
```
Missing Values:
 Age           0
SystolicBP    0
DiastolicBP   0
BS            0
BodyTemp      0
HeartRate     0
RiskLevel     0
dtype: int64
```

Check the dataset for any missing values in each attribute. Use methods like .isnull() or .info() to identify missing data.

2. Dealing with Missing Data:

If missing values are found, decide on an appropriate strategy to handle them. – There is no missing value, so it does not need to deal with it.

3. Identifying Duplicate Data:



```
Duplicate Data:
     Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel
66    19         120           80   7.0      98.0         70   mid risk
71    19         120           80   7.0      98.0         70   mid risk
96    19         120           80   7.0      98.0         70   mid risk
104   50         140           90  15.0      98.0         90  high risk
105   25         140          100   6.8      98.0         80  high risk
..   ...         ...          ...   ...       ...        ...        ...
970   22         120           60  15.0      98.0         80  high risk
971   55         120           90  18.0      98.0         60  high risk
972   35          85           60  19.0      98.0         86  high risk
973   43         120           90  18.0      98.0         70  high risk
974   32         120           65   6.0     101.0         76   mid risk

[541 rows x 7 columns]
```

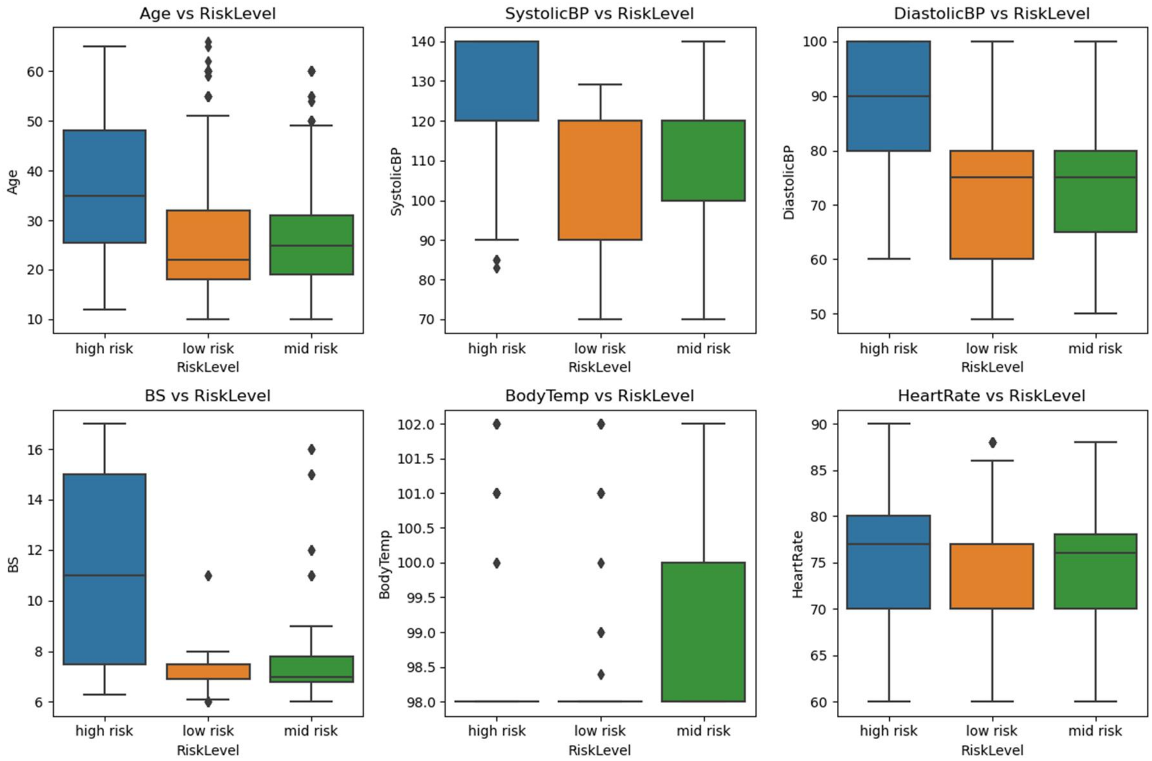
Check the dataset for any duplicate rows using methods like .duplicated().

4. Dealing with Duplicate Data:

If duplicate rows are found, decide whether to remove or keep them based on the significance of the duplicates to the analysis.

As I mentioned before, there are too many duplicated rows and if those are removed, there is big loss of data then, it occurs building inaccurate models. So, duplicated rows should be kept.

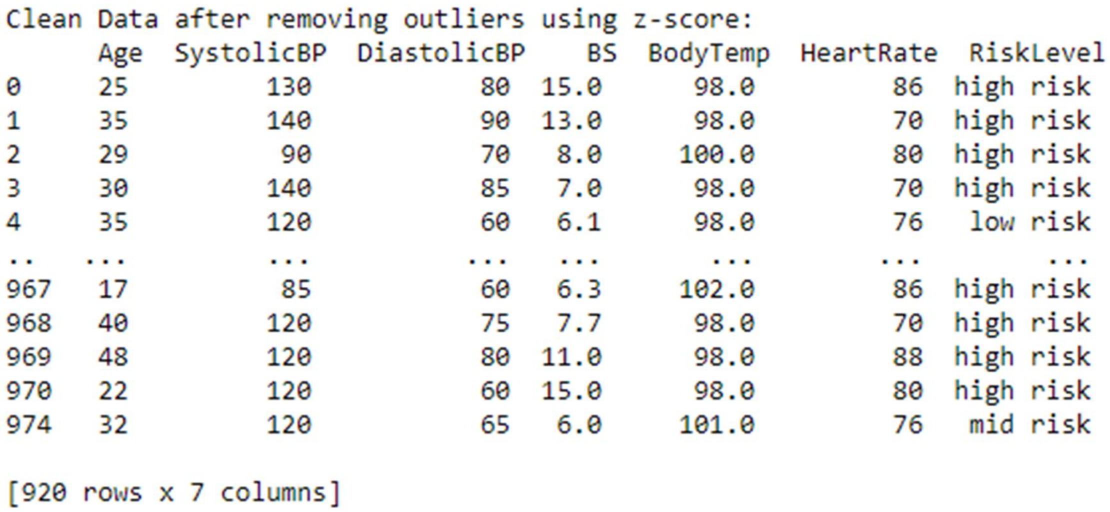
5. Identifying Outliers:



Examine the distribution of each numerical attribute using summary statistics and visualizations like box plots or histograms to detect outliers.

There are some outliers and those will be removed.

6. Dealing with Outliers:



```
Clean Data after removing outliers using z-score:
      Age  SystolicBP  DiastolicBP    BS  BodyTemp  HeartRate  RiskLevel
0      25         130           80  15.0      98.0         86  high risk
1      35         140           90  13.0      98.0         70  high risk
2      29          90           70   8.0     100.0         80  high risk
3      30         140           85   7.0      98.0         70  high risk
4      35         120           60   6.1      98.0         76   low risk
..    ...         ...          ...   ...       ...        ...        ...
967    17          85           60   6.3     102.0         86  high risk
968    40         120           75   7.7      98.0         70  high risk
969    48         120           80  11.0      98.0         88  high risk
970    22         120           60  15.0      98.0         80  high risk
974    32         120           65   6.0     101.0         76   mid risk

[920 rows x 7 columns]
```

Decide on an appropriate strategy to handle outliers based on their impact on the analysis and the specific requirements of the problem.

By using z-score, the outliers are removed and 920 rows are left.

b) Creating Decision Tree Model

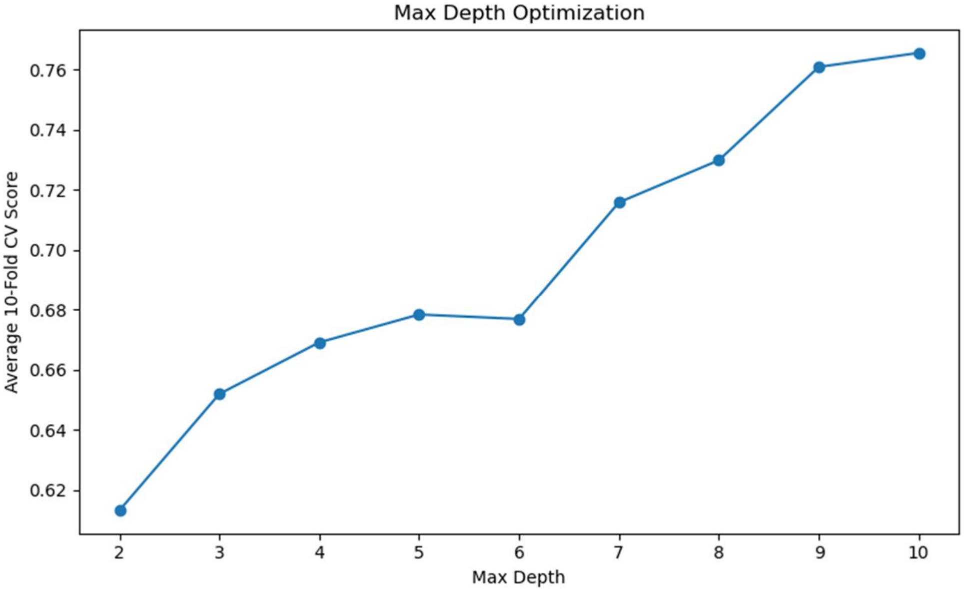
- Parameter Adjustment: Adjust two suitable parameters to reduce the size of the decision tree and improve its accuracy. Common parameters to adjust include max_depth, min_samples_split, min_samples_leaf, max_leaf_nodes, etc.
  For this task, max depth and max leaf nodes are selected to adjust decision tree and accuracy score.

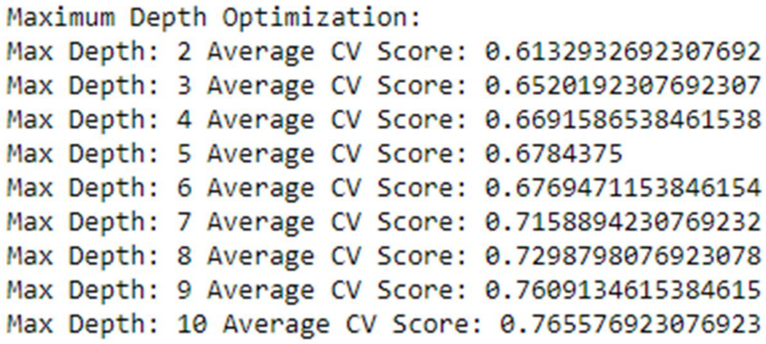
- Set range of max depth and max leaf nodes: Iterate over a range of values for each parameter and calculate the average cross-validation score for each value.

  max depth range are set by 2 to 10, max depth can be more and more max depth value show higher accuracy normally. However, more max depth value makes decision tree more complexicity and hard to visualize.

  Max leaf nodes range are set by 2 to 20, Also same reason with above, bigger max leaf nodes are better accuracy, but it makes big decision tree size and make overfitting model.

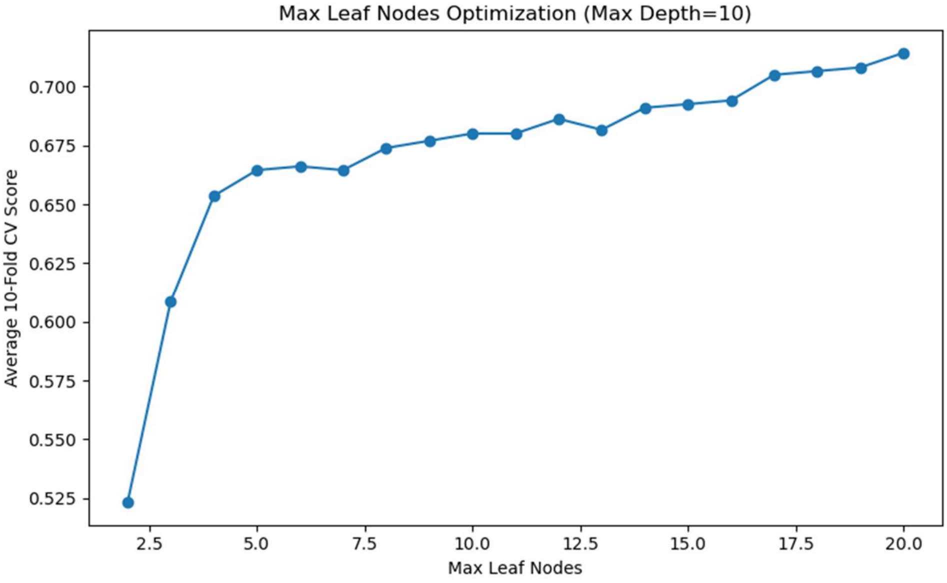
- Plotting and listing: Visualize the impact of parameter adjustments on the model's accuracy using plots. Plot the parameter values against the average cross-validation scores to understand how changing the parameters affects the model's performance.

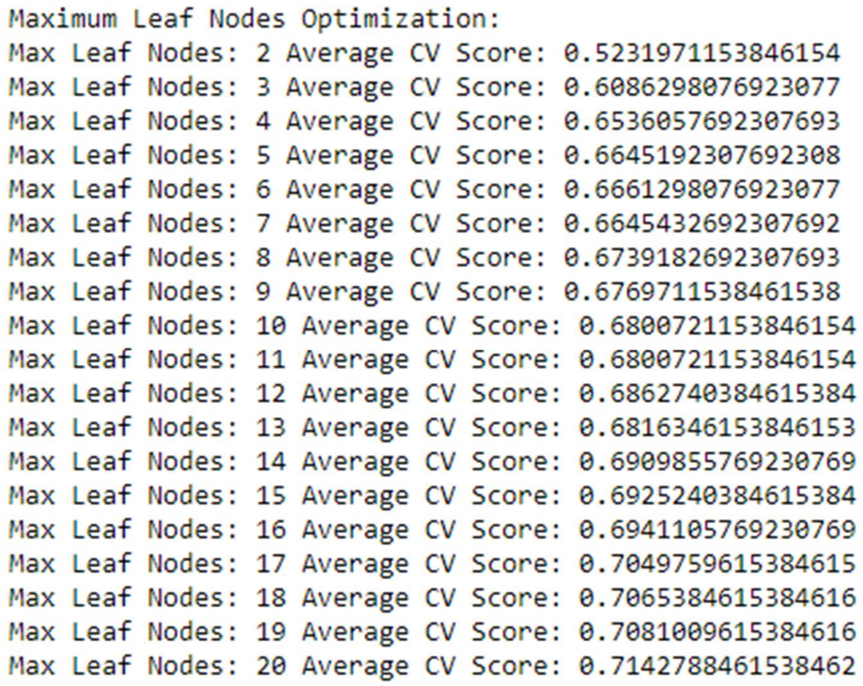




```
Maximum Depth Optimization:
Max Depth: 2 Average CV Score: 0.6132932692307692
Max Depth: 3 Average CV Score: 0.6520192307692307
Max Depth: 4 Average CV Score: 0.6691586538461538
Max Depth: 5 Average CV Score: 0.6784375
Max Depth: 6 Average CV Score: 0.6769471153846154
Max Depth: 7 Average CV Score: 0.7158894230769232
Max Depth: 8 Average CV Score: 0.7298798076923078
Max Depth: 9 Average CV Score: 0.7609134615384615
Max Depth: 10 Average CV Score: 0.765576923076923
```

In this plot and list, the 10 max depth is selected as adjusted one parameter because the average CV score is higher than the others. Then the 10 of max depth average CV score is 0.765576923076923.

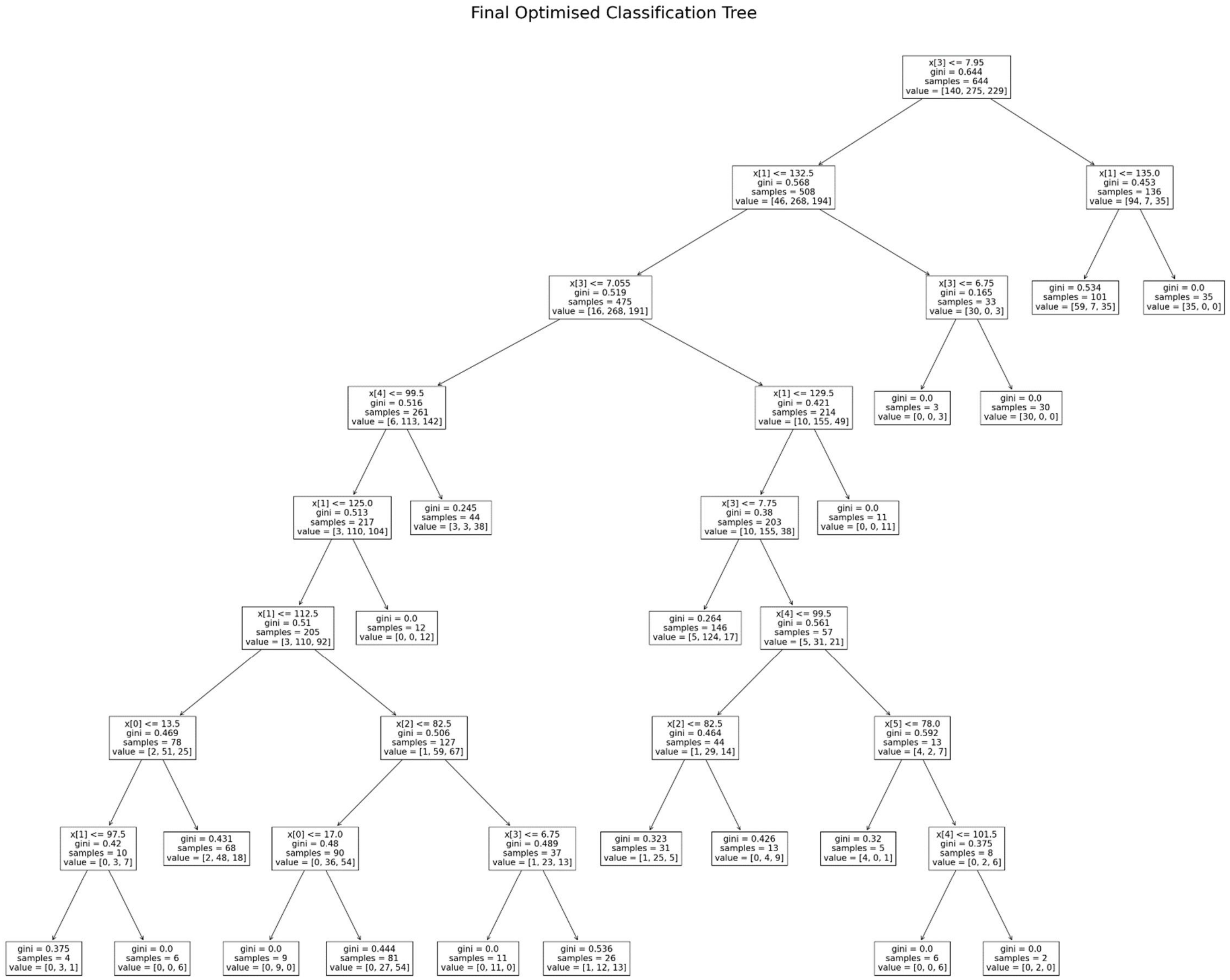




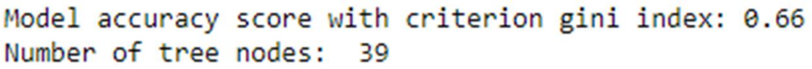
```
Maximum Leaf Nodes Optimization:
Max Leaf Nodes: 2 Average CV Score: 0.5231971153846154
Max Leaf Nodes: 3 Average CV Score: 0.6086298076923077
Max Leaf Nodes: 4 Average CV Score: 0.6536057692307693
Max Leaf Nodes: 5 Average CV Score: 0.6645192307692308
Max Leaf Nodes: 6 Average CV Score: 0.6661298076923077
Max Leaf Nodes: 7 Average CV Score: 0.6645432692307692
Max Leaf Nodes: 8 Average CV Score: 0.6739182692307693
Max Leaf Nodes: 9 Average CV Score: 0.6769711538461538
Max Leaf Nodes: 10 Average CV Score: 0.6800721153846154
Max Leaf Nodes: 11 Average CV Score: 0.6800721153846154
Max Leaf Nodes: 12 Average CV Score: 0.6862740384615384
Max Leaf Nodes: 13 Average CV Score: 0.6816346153846153
Max Leaf Nodes: 14 Average CV Score: 0.6909855769230769
Max Leaf Nodes: 15 Average CV Score: 0.6925240384615384
Max Leaf Nodes: 16 Average CV Score: 0.6941105769230769
Max Leaf Nodes: 17 Average CV Score: 0.7049759615384615
Max Leaf Nodes: 18 Average CV Score: 0.7065384615384616
Max Leaf Nodes: 19 Average CV Score: 0.7081009615384616
Max Leaf Nodes: 20 Average CV Score: 0.7142788461538462
```

A case of 10 of max depth, above plot and list show each max leaf nodes and each average CV score. 20 of max leaf nodes are selected and it shows the highest average CV score which is 0.7142788461538462.

- Final Optimized Model: Once you have identified the optimal parameter values, create the final optimized decision tree model using these values.



Final Optimised Classification Tree



```
Model accuracy score with criterion gini index: 0.66
Number of tree nodes:  39
```

- Describe the Tree Structure:
  The tree has a depth of 10, the total nodes are 39 and the accuracy score is 0.66.
  Structure Breakdown:
  The tree follows a binary structure, meaning each node splits the data into two branches based on a decision rule using a single feature. The decision rule involves a feature (e.g., x[3] <= 7.95) and a threshold value. If a data point's value for that feature meets the condition, it follows the left branch; otherwise, it goes to the right branch.

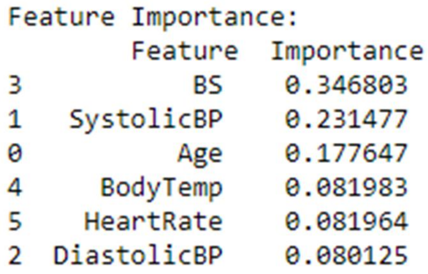
  Complexity:
  With a depth of 10 and 39 nodes, this tree has a moderate level of complexity. It's not overly simple, suggesting some depth for potentially

capturing complex relationships in the data. However, it's also not extremely complex, which helps maintain interpretability.
The text you provided shows lines with "gini" values followed by "samples" and "value" breakdowns:

Gini Value: This represents the Gini impurity at that node. A value closer to 0 indicates a purer node (better split), while a value closer to 0.5 signifies a more impure node (classes are mixed).
Samples: This indicates the number of data points that reach that particular node.
Value: This shows the distribution of class labels (e.g., [164, 279, 239] for "lowrisk", "midrisk" and "highrisk" classes) at that node.

c) Roles of the two parameters
In the model building process, two parameters were adjusted: max depth and max leaf nodes. Max depth controls the depth of the decision tree, balancing between complexity and generalization, while max leaf nodes limit the tree's size to prevent overfitting.
Whether the same parameter values will improve accuracy on other datasets depends on factors like dataset characteristics, size, and domain-specific considerations. While the obtained values can serve as a starting point, fine-tuning is necessary for optimal performance on different datasets.

d) Feature Importance



```
Feature Importance:
       Feature  Importance
3           BS    0.346803
1   SystolicBP    0.231477
0          Age    0.177647
4     BodyTemp    0.081983
5    HeartRate    0.081964
2  DiastolicBP    0.080125
```

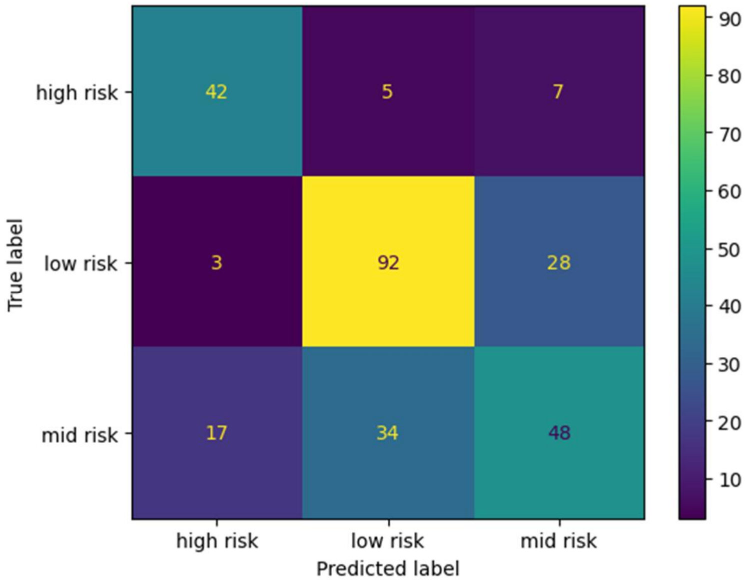
In the given feature importance output, each feature's importance score is provided based on the trained Decision Tree Classifier. Here's an explanation of the feature importance values:

- BS (Blood Sugar) 0.347 : The highest importance value suggests that blood sugar level (BS) is the most influential feature in predicting the target variable. This implies that variations in blood sugar levels have a strong impact on the outcome being predicted.
- Systolic Blood Pressure (Systolic) 0.231: The second-highest importance value indicates that age is the next most important feature. SystolicBP

often plays a crucial role in many predictive tasks, as certain conditions or risks may be more prevalent among specific SystolicBP groups.

- Age 0.178: Age ranks third in importance, suggesting that variations in age also contribute significantly to predicting the target variable. Older and younger age may be indicative of certain health conditions or risks.
- Body Temperature (BodyTemp) 0.082: Body Temperature follows age in importance. Body Temperature variations may provide valuable information for predicting health-related outcomes.
- Heart Rate (HeartRate) 0.082: Heart Rate has a relatively lower importance compared to the other features but still contributes significantly to the predictive performance of the model. Changes in Heart Rate can be indicative of various health conditions.
- Diastolic Blood Pressure (DiastolicBP) 0.080: Diastolic Blood Pressure has the lowest importance value among the features listed. However, it still plays a role in predicting the target variable, Heart Rate may provide additional predictive information.

e) Confusion Matrix



The confusion matrix is a valuable tool for evaluating the performance of a classification model by presenting a summary of the model's predictions versus the actual class labels across different categories.

In this particular confusion matrix:

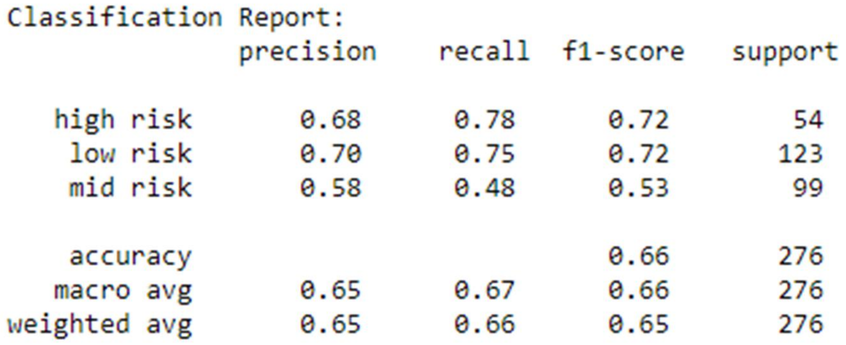
- The rows represent the actual classes or labels, while the columns represent the predicted classes by the model.
- Each cell in the matrix corresponds to the count of instances where the actual class matches the row label and the predicted class matches the column label.

Now, let's interpret the confusion matrix:

- High Risk Class (Row 1):
  - True Positives (TP): 42 instances were correctly predicted as high risk.
  - False Negatives (FN): 5 instances that were actually high risk were incorrectly classified as mid-risk.
  - False Positives (FP): 7 instances were incorrectly classified as high risk when they were actually mid-risk or low risk.
- Low Risk Class (Row 2):
  - True Positives (TP): 92 instances were correctly predicted as low risk.
  - False Negatives (FN): 3 instances that were actually low risk were incorrectly classified as high risk.
  - False Positives (FP): 28 instances were incorrectly classified as low risk when they were actually high risk or mid-risk.
- Mid Risk Class (Row 3):
  - True Positives (TP): 48 instances were correctly predicted as mid risk.
  - False Negatives (FN): 17 instances that were actually mid risk were incorrectly classified as low risk.
  - False Positives (FP): 34 instances were incorrectly classified as mid risk when they were actually high risk or low risk.

From the confusion matrix, several observations can be made:

1. The model generally performs well in correctly identifying instances of high and low risk, as evidenced by the high counts of true positives in these classes.
2. However, the model struggles more with accurately classifying instances of mid risk, as indicated by the lower count of true positives and higher counts of false positives and false negatives.
3. The false positives and false negatives in each class highlight areas where the model makes errors in classification, which can provide insights into potential areas for improvement or model refinement.



```
Classification Report:
              precision    recall  f1-score   support

   high risk       0.68      0.78      0.72        54
    low risk       0.70      0.75      0.72       123
    mid risk       0.58      0.48      0.53        99

    accuracy                           0.66       276
   macro avg       0.65      0.67      0.66       276
weighted avg       0.65      0.66      0.65       276
```

The model summary report includes various performance metrics such as accuracy, precision, recall, and F1-score.

- Accuracy**:** The overall accuracy of the model is 0.66, indicating that 66% of the predictions are correct.
- Precision: Precision measures the proportion of true positive predictions among all positive predictions.
    - Precision for high risk: 0.68
    - Precision for low risk: 0.70
    - Precision for mid risk: 0.58
- Recall**:** Recall measures the proportion of true positive predictions among all actual positive instances.
    - Recall for high risk: 0.78
    - Recall for low risk: 0.75
    - Recall for mid risk: 0.48
- F1-Score: The F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.
    - F1-score for high risk: 0.72
    - F1-score for low risk: 0.72
    - F1-score for mid risk: 0.53

Task 4: Results and Discussions (500-600 words) [20 marks]

Answer:

When evaluating the performance of the decision tree model, it's crucial to consider various metrics such as accuracy, precision, recall, and F1-score, particularly for each individual risk class. The decision tree model achieves an overall accuracy of 0.66, indicating that it correctly predicts the class label for approximately two-thirds of the instances. However, a deeper analysis reveals variations in performance across different risk classes within the model.

For the high-risk class, the model demonstrates a precision of 0.68 and a recall of 0.78, suggesting that it accurately identifies a significant portion of instances labelled as high risk while minimizing false positive predictions. Similarly, for the low-risk class, the model exhibits high precision (0.70) and recall (0.75), indicating effective identification of instances with low-risk attributes.

On the other hand, the model's performance declines when predicting instances belonging to the mid-risk class. Here, the precision drops to 0.58, indicating a higher rate of false positives, while the recall decreases to 0.48, suggesting that the model misses a considerable portion of actual mid-risk instances. This discrepancy highlights the challenge the model faces in accurately distinguishing between mid-risk and other risk categories.

Evaluation using confusion matrices provides further insights into the model's performance. The confusion matrix reveals the distribution of true positive, true negative, false positive, and false negative predictions across different classes. In the case of our decision tree model, the confusion matrix illustrates that while it accurately predicts a significant portion of high and low-risk instances, it tends to misclassify some mid-risk instances as either low or high risk. This pattern of misclassification highlights a potential area for improvement in the model's ability to differentiate between mid-risk and other risk categories.

In summary, while the decision tree model shows promise in predicting high and low-risk instances, there is clear room for improvement, particularly in accurately classifying mid-risk instances. Further refinement of the model, feature engineering, or exploring alternative algorithms may help address these challenges and enhance overall performance. Additionally, it's essential to consider the specific requirements and constraints of the application domain when evaluating and selecting the most suitable model. By iteratively refining the model and incorporating domain knowledge, we can strive to build more robust and reliable predictive models for risk classification tasks.