Assignment 1

Data Exploration and Classification

Semester 1 2024

Student Name: Julia Calma

Student ID:22167597

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**.

➢ Table of Contents: Include a table of contents to provide an overview of the report's structure.

Table of Contents

Task 1: Introduction (200-300 words) [10 marks]

Provide a statement of the problem, outlining the problem your chosen dataset addresses. The statement of the problem should briefly address the question: What is the problem that you will investigate in this assignment?

Your introduction must describe:

- The aim of your work, what are you trying to achieve, and research questions you attempted to answer.
- All assumptions that your data must meet.

In this assignment, I will be exploring and analysing the maternal health risk dataset. The data was obtained from various hospitals, community clinics, maternal health cares facilities in the rural areas of Bangladesh. The collection was done using an Internet of Things (IoT) based risk monitoring system. This dataset aims to solve the problem of creating a classification model that, given a pregnant woman's demographic and physiological information, can predict the level of risk intensity during her pregnancy. Classifying pregnant women into risk groups (e.g., low, medium, high) based on features including age, blood pressure, blood glucose levels, body temperature, and heart rate . My attention has been piqued by this topic since I have been curious about the factors that actually lead to the tragic occurrence of mothers or women who are about to

become mothers passing away when they are pregnant or giving birth. The formal name for this is maternal mortality, which refers to the death of pregnant women during the process of pregnancy or childbirth. This investigation will be conducted with the primary objective of identifying the factor(s) or feature(s) that are most responsible for the occurrence of maternal mortality. I will accomplish this by thoroughly exploring, cleaning each feature, and visualising its relationship between each feature and, of course, the target variable which in this case is the RIskLevel variable. For the purpose of this investigation, I am operating under the assumption that the data provided is both accurate and complete and all features are reliable, in order to maintain the validity and reliability of this investigation.

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

This section of your report must discuss the dataset and any features you consider relevant to the analysis and modelling task.

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

The maternal health risk dataset has 7 features and consists of 1014 instances. The included features are Age, SystolicBP, DiastolicBP, BS, BodyTemp, HeartRate , and Risk Level. All of their data types are numeric, with the exception of the RiskLevel variable, which is categorical.

```
Data Shape: (1014, 7)


Data Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
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
```

• Provide summary statistics of the continuous numerical features.

Summary Staistics:

	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.Describethe steps taken to address both data cleanliness evaluation and data cleaning strategies.

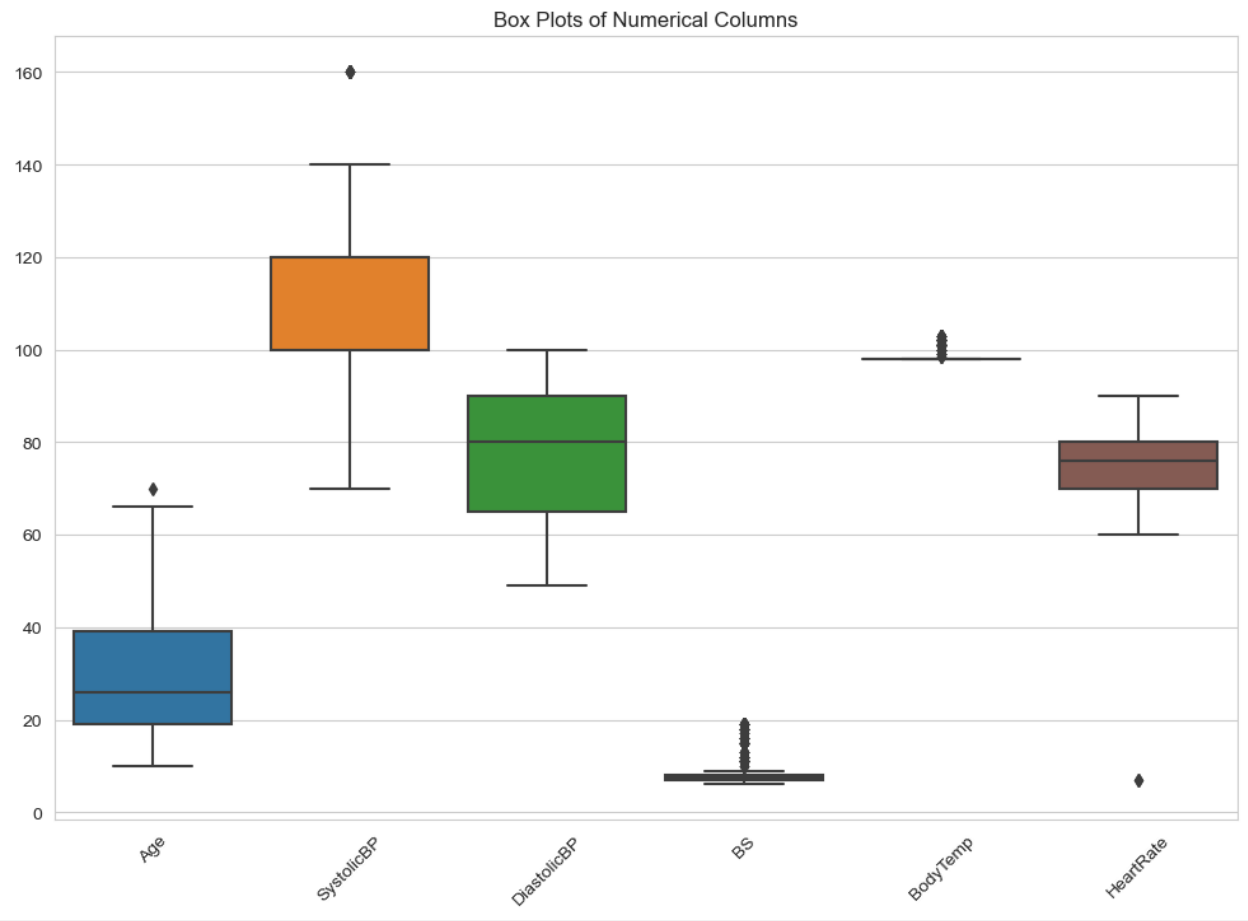
Missing values:

```
Missing values in the dataset:
Age             0
SystolicBP      0
DiastolicBP     0
BS              0
BodyTemp        0
HeartRate       0
RiskLevel       0
dtype: int64
```

Duplicates: Keep duplicates because all duplicates are greater than 50% which could lead to significant data loss and reduction in sample size

```
Number of duplicates: 866
```

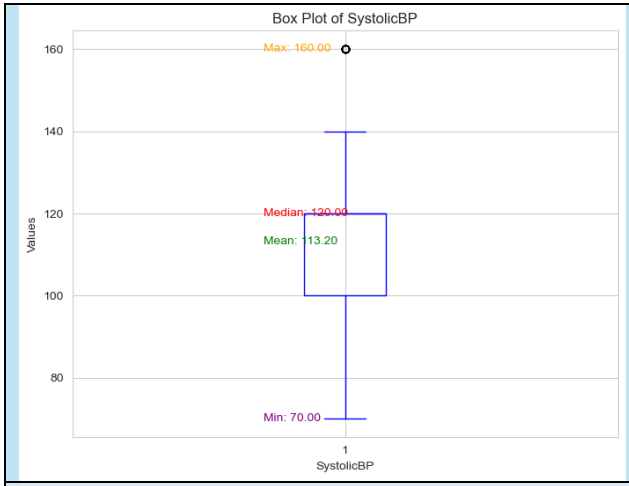
I will utilise this boxplot to visually detect potential outliers as this plot is effective in segregating outliers easily.

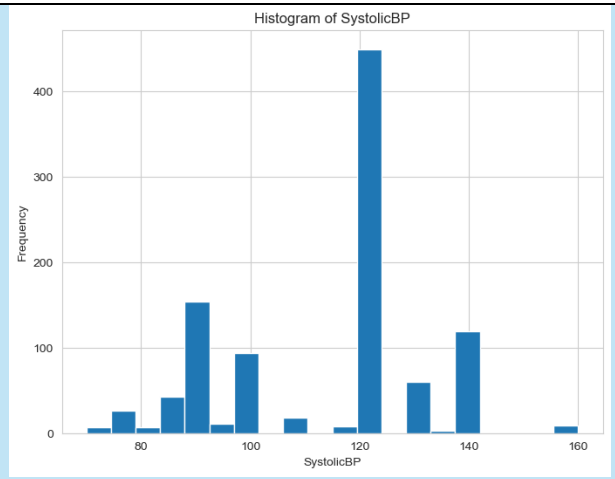


Outliers: Scanning through the boxplot I notice some outliers in Age, SystolicBP, BS BodyTemp, and HeartRate.

For the *Age* feature, I have decided to leave the outliers since it undeniably happens in real life where even ages 60 and above get pregnant.

Systolic Blood Pressure

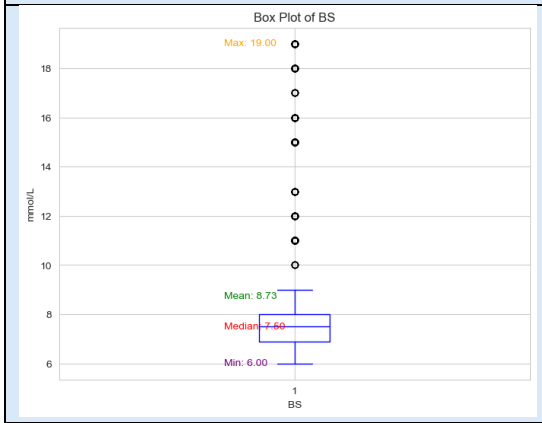


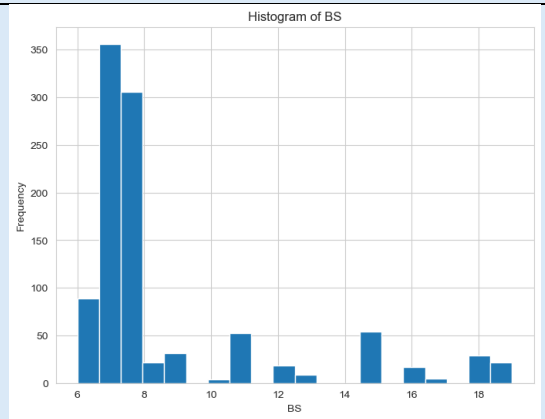


Normal distibuted graph: use z score

```
Indices of outliers detected using z-score:
Age            44
SystolicBP     10
DiastolicBP     0
BS             73
BodyTemp       79
HeartRate       0
dtype: int64
```

Blood Sugar



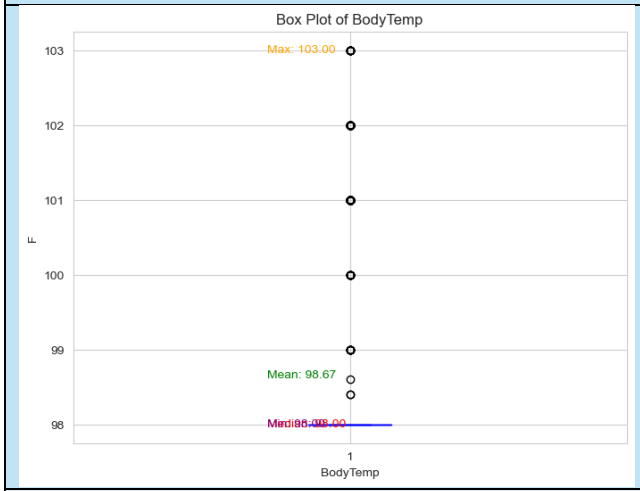


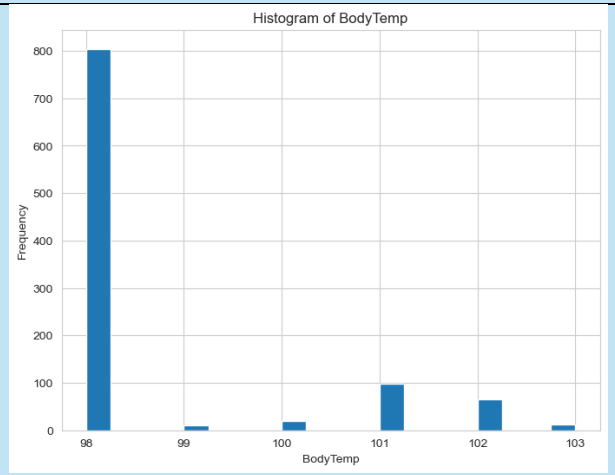
Right-skewed graph:use iqr

```
Number of outliers in each
numerical column using IQR method:


Age              1
SystolicBP      10
DiastolicBP      0
BS             210
BodyTemp       210
HeartRate        2
dtype: int64
```

Body Temperature



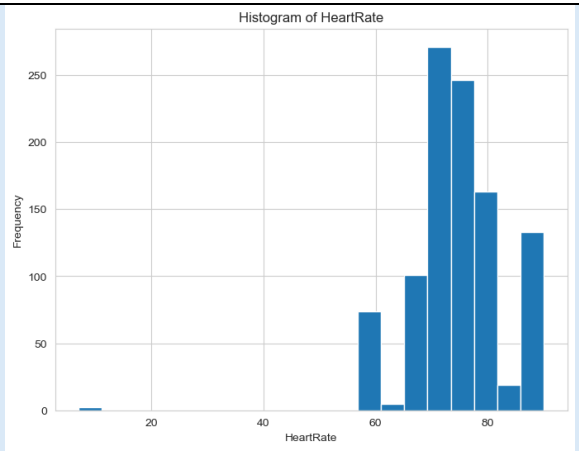


Right-skewed graph:use iqr

```
Number of outliers in each
numerical column using IQR method:


Age              1
SystolicBP      10
DiastolicBP      0
BS             210
BodyTemp       210
HeartRate        2
dtype: int64
```

Heart Rate



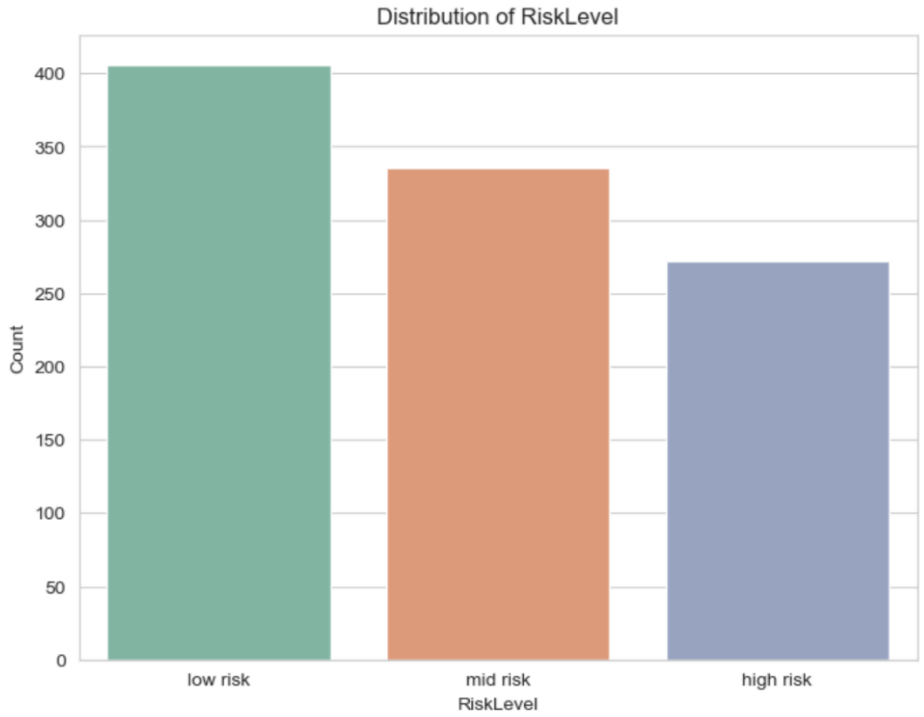
Left-skewed graph: use iqr

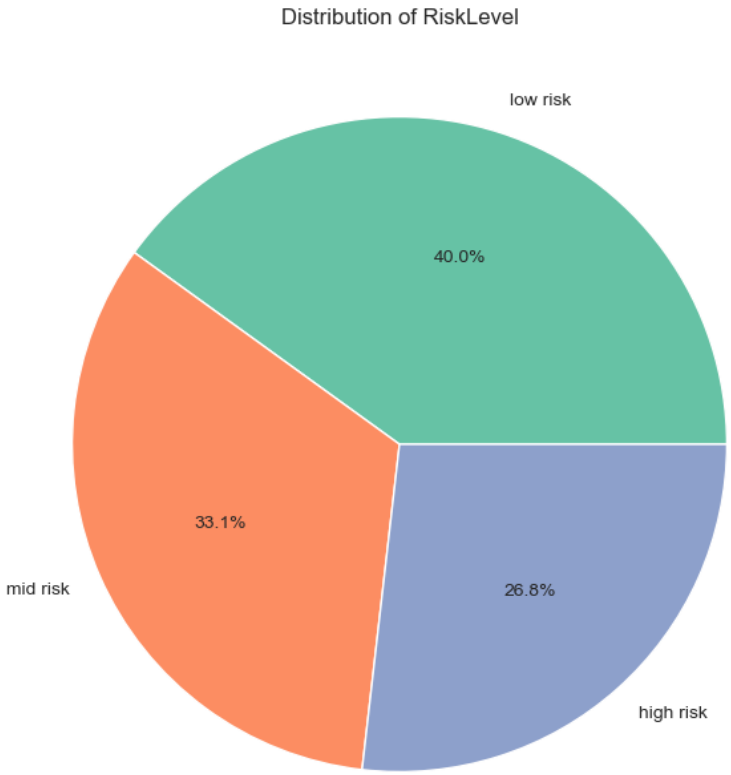
```
Number of outliers in each
numerical column using IQR method:


Age              1
SystolicBP      10
DiastolicBP      0
BS             210
BodyTemp       210
HeartRate        2
dtype: int64
```

• 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).

Distribution of Risk Level

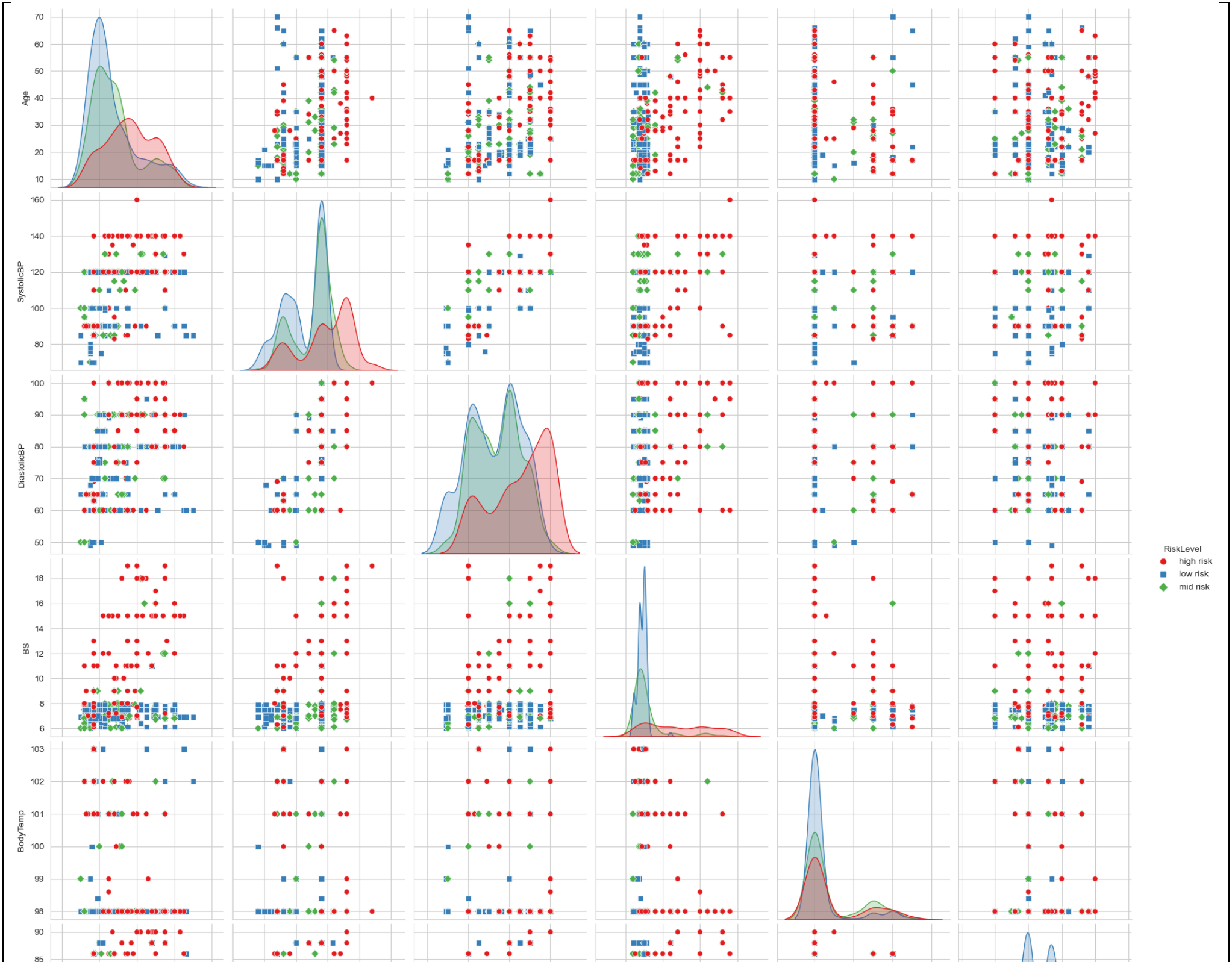





```
low risk     406
mid risk     336
high risk    272
```

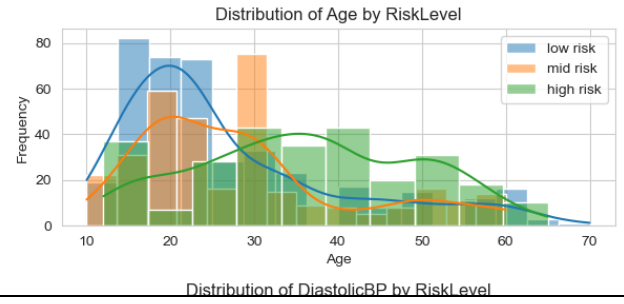
Explanation: We see that distribution is quite imbalanced. With the highest percentage, 40%, belonging to the low risk class, followed by the mid-risk class with 33.1%, and the high-risk class with the lowest percentage of 26.8%. However, the distribution is not highly imbalanced, indicating no significant differences between them. Consequently, the model would not exhibit a strong bias towards the majority feature and would not adversely impact the prediction performance on minority classes. Therefore, I conclude that there is no necessity for employing any handling techniques such as class resampling or weighting.

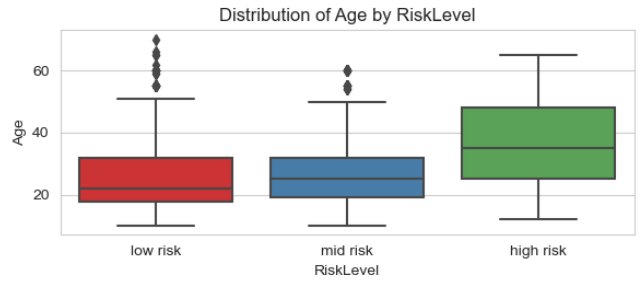
Pairplot-- pairwise relationship of measurements for each Risk Level

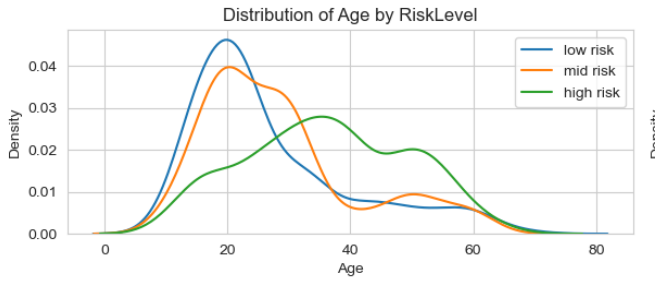

Age
SystolicBP
DiastolicBP
BS
BodyTemp
RiskLevel
high risk
low risk
mid risk

Explanation:
By the usage of pairplot, we can visualise pairwise relationships between our features (Age, Systolic BP, Diastolic BP, Blood Sugar (BS), Body Temperature, and Heart Rate) for each risk level (low risk, medium risk, and high risk). This helps us identify correlations or patterns within each risk level. In the pairplot, we observe that these relationships often show vertical or horizontal alignments rather than a gradient or slope. This is due to most of our features having points that are discrete numbers.
Overviewing the whole plot, we only see a few red points clustered with blue and green, and most of these red points appear scattered away from these clusters like outliers. This may imply that there are specific conditions or risk factors associated with the red points (representing high risk). These conditions could be unique or extreme compared to the majority of cases (blue and green points), suggesting distinct patterns or outlier behavior within the high-risk category. One notable linear pattern we observe is the relationship between Diastolic BP and Systolic BP, indicating a high correlation between these variables. We notice that the feature with the most correlation with all other features, as evidenced by the densest cluster of points, is Body Temperature. This suggests a strong relationship among the variables when Body Temperature is involved. Following this observation is Blood Sugar (BS), which also shows clustering of points. In this pairplot, we observe KDE plots diagonally, these plots show because it is the plot best in providing a smooth representation of a feature's own distribution.

Distribution of Age by RiskLevel

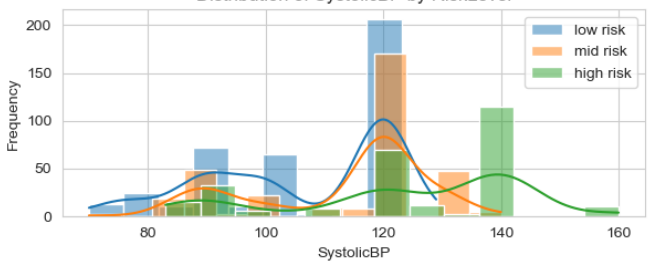


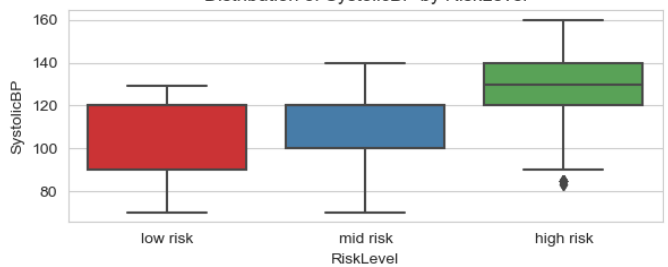


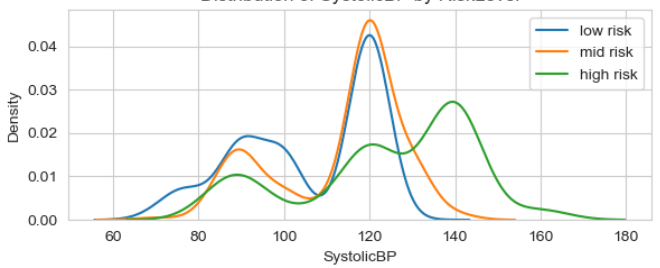


Explanation: These three plots show that as a pregnant woman's age increases (gets older), her maternal health risk level also increases (becomes high risk). Additionally, the KDE and histogram plots reveal that the majority of ages in this dataset center around 20 years old.

Distribution of SystolicBP by RiskLevel

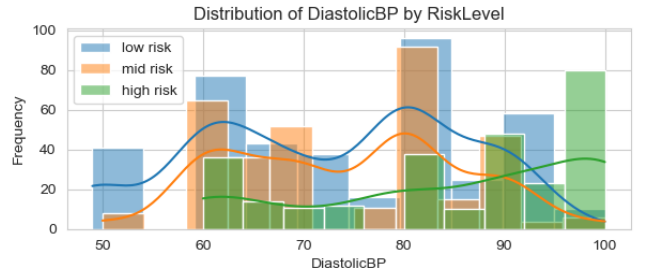


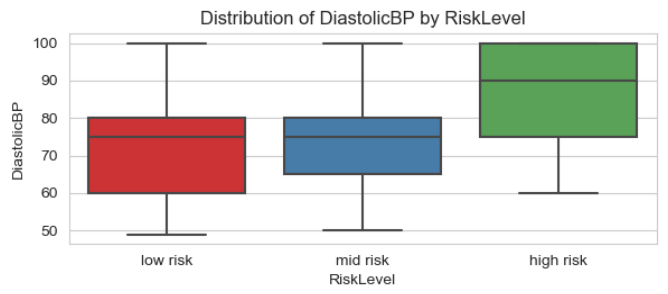


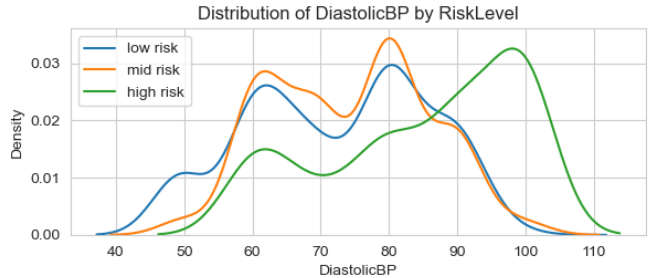


Explanation: The plots indicate that the largest proportion of pregnant women in this dataset classified as high risk occurs when their systolic blood pressure is 120 mmHg or higher. We also observe that a significant number of pregnant women have a systolic blood pressure of 120 mmHg, which indicates low and medium risk levels for maternal health.

Distribution of DiastolicBP by RiskLevel

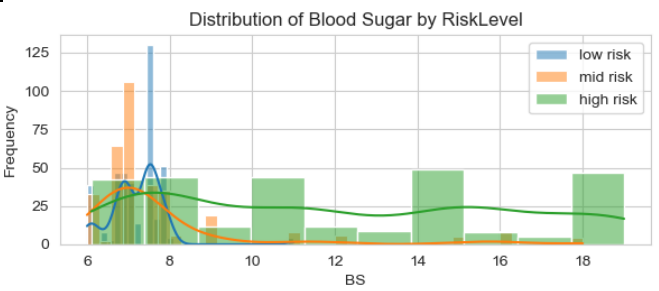


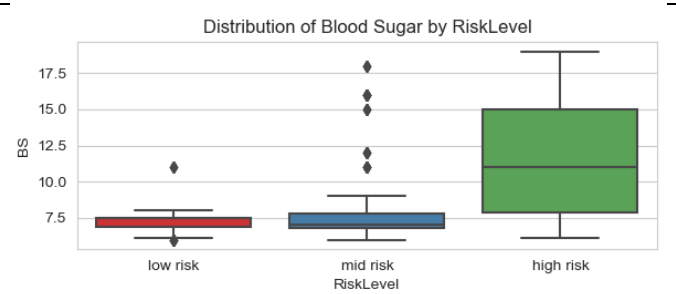


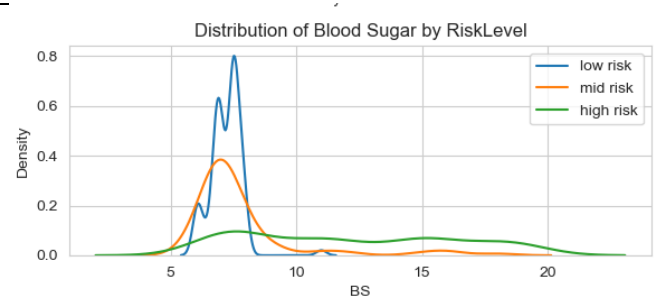


Explanation: These plots illustrate that as the diastolic blood pressure increases, the risk level also increases. We observe from the boxplot and KDE plot that a significant number of pregnant women have a high risk level for maternal health, basing off diastolic blood pressure.

Distribution of Blood Sugar by RiskLevel

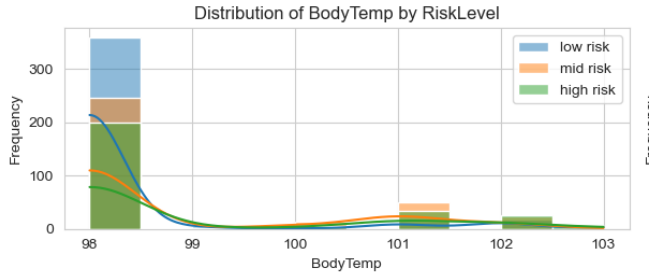


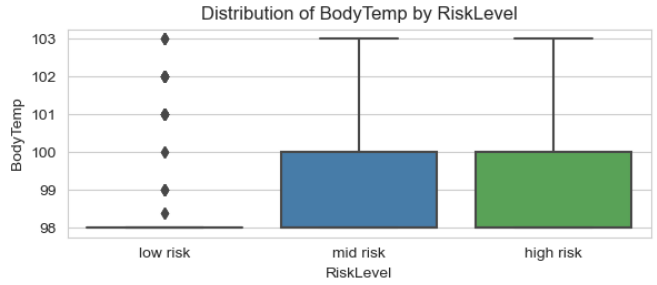


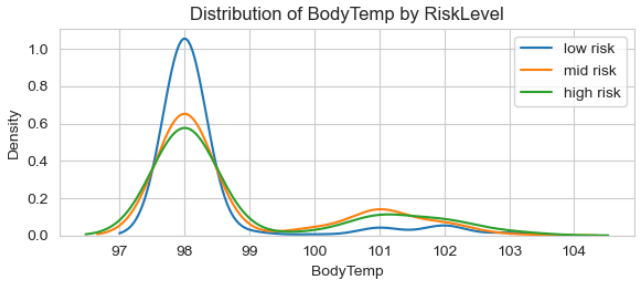


Explanation: We notice that, except for special cases (outliers in low risk and mid risk), blood sugar levels higher than 7.5 mmol/L pose a high risk for pregnant women. The difference between risk levels is evident in this distribution, which may imply strong predictive capability in predicting maternal health risk level. In the KDE and histogram, we observe that the most common blood sugar level for the low-risk category is approximately 7.5 mmol/L, as indicated by the peak of the low-risk level. This concentration of data suggests that a significant proportion of pregnant women in this dataset fall into this category, which is supported by the distribution of our target variable (Risk Level) shown in the pie chart previously.

Distribution of BodyTemp by RiskLevel

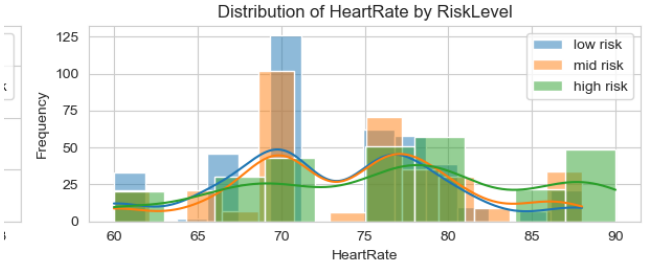


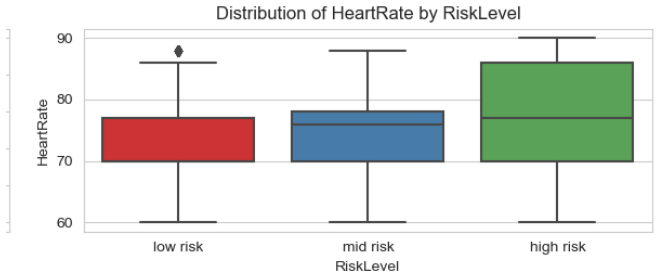


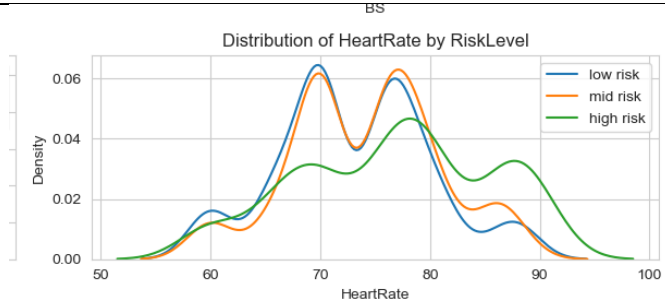


Explanation: In the boxplot, we identify that a pregnant woman can be at medium risk or high risk if the temperature is 98°F or above. This suggests that body temperature may not be the sole determining factor in maternal health risk level. Looking at the histogram and KDE plot, we observe that pregnant women at low risk have a body temperature right around 98°F. Additionally, we can see that the highest peak is at 98°F, indicating that most pregnant women in this dataset have a body temperature of 98°F and are at low risk as mentioned.

Distribution of HeartRate by RiskLevel







Explanation: In these plots, we notice that as heart rate increases, the risk level also increases. It is apparent that there is only a slight difference between low risk and medium risk in relation to heart rate. This could also signify that heart rate is not the best predictor for maternal health risk levels. On the other hand, viewing the histogram made it apparent that a great number of pregnant women in this dataset have a heart rate of 70 bpm, suggesting a low-risk maternal health level.

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

From conducting data exploration and visualization on this dataset, I learnt more about the relationship between each feature and the target variable. Using pie charts and bar graphs, I identified that most pregnant women in the maternal health risk dataset are categorized as low risk. This observation was further supported when examining the distribution of each feature by risk level using three types of plots (histogram, box plot, and KDE). These plots allowed me to assess the majority of risk levels by observing the height, width, and length of the box or the area. Outliers are also clearly displayed, particularly using box plots, which helped us understand the nature of the relationship between features and risk levels. What I concluded from this analysis is that out of all the features, the distribution of Blood Pressure (BP) against the risk level was the clearest. It is evident that the higher the blood sugar level, the higher the risk level. This insight provides valuable information for understanding maternal health risk factors and their impact on pregnant women in the dataset.

Task 3: Classification Models (500-600 words) [40 marks]

You need to create a model using the Decision Tree Classifier and answer the following questions based on the model built. In building the model, use the 10-fold cross-validation option for testing. Your answers need to be supported by suitable evidence, wherever appropriate. Some examples of suitable evidence are Confusion Matrices, Model Visualizations, and Model Summary Reports.

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]

Duplicates: I have decided to keep all duplicates because the number of duplicates is greater than 50% which could lead to significant data loss and reduction in sample size.

```
Number of duplicates: 866
```

Outliers: I have decided to retain all these outliers except the heart rate feature. I have conducted comprehensive research on all remaining features(Age, Systolic BP, Blood Sugar, and Body Temperature) and have conclusively identified that the values within them are indeed 'true outliers'. This decision is driven by the understanding that they represent valid and correct data points that can provide valuable insights and preserve crucial information within the dataset.

On the other hand, I have made the decision to remove the outliers for heart rate. This is because the normal range for heart rate is often between 60 and 100 beats per minute (bpm). Upon examining this particular feature, we have observed occasions where the heart rate was recorded as 7bpm,

which is significantly lower than the expected range. Therefore, we will eliminate these instances from our analysis. Additionally, there are only two occurrences of them, hence their impact on our analysis would be minimal.

SUMMARY STATISTICS OF HEART (OUTLIER REMOVAL)

BEFORE

	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

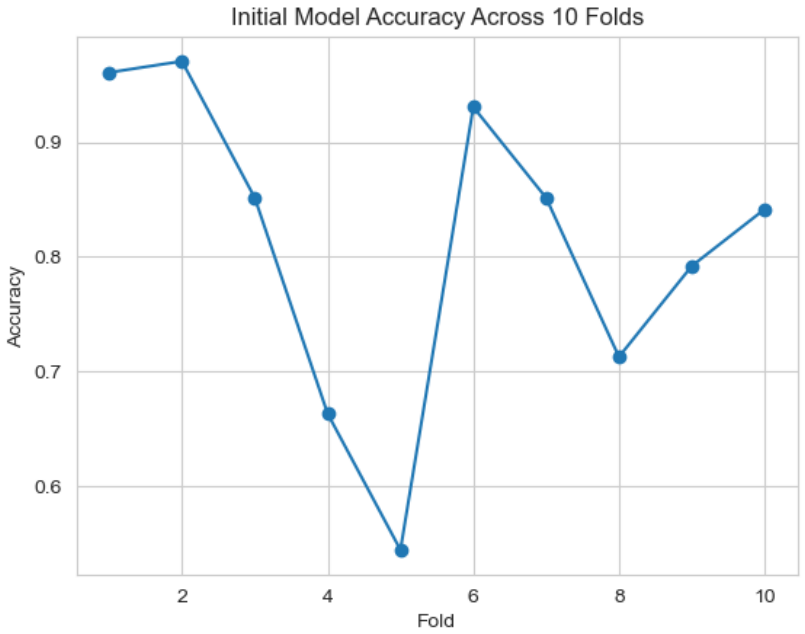
AFTER

```
Summary Statistics after Removing Outliers:
count    1012.000000
mean       74.434783
std         7.521857
min        60.000000
25%        70.000000
50%        76.000000
75%        80.000000
max        90.000000
```

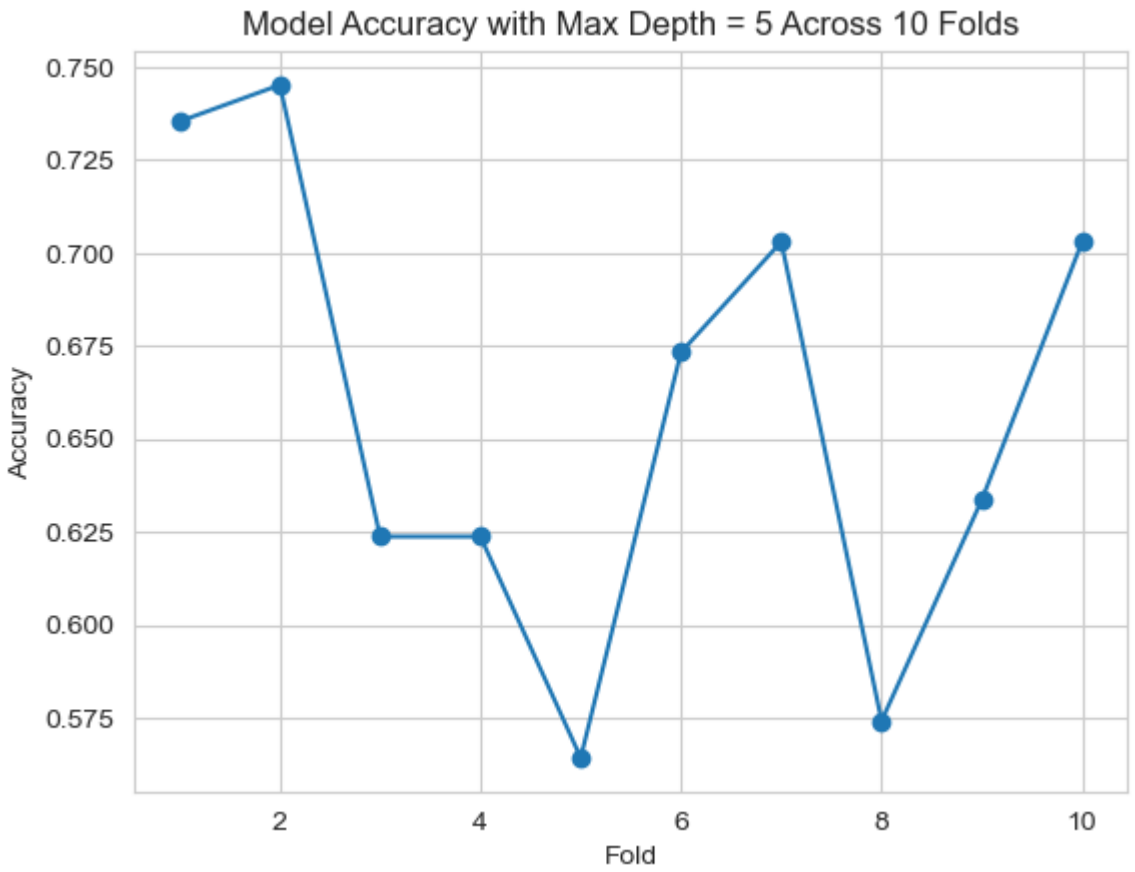
Missing data: As there were no missing data values, this simplifies the process of data cleaning or handling

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]

Initial Model



First parameter adjustment using ‘max_depth’



Accuracy scores with Max Depth = 5:

Fold 1: 0.7353

Fold 2: 0.7451

Fold 3: 0.6238

Fold 4: 0.6238

Fold 5: 0.5644

Fold 6: 0.6733

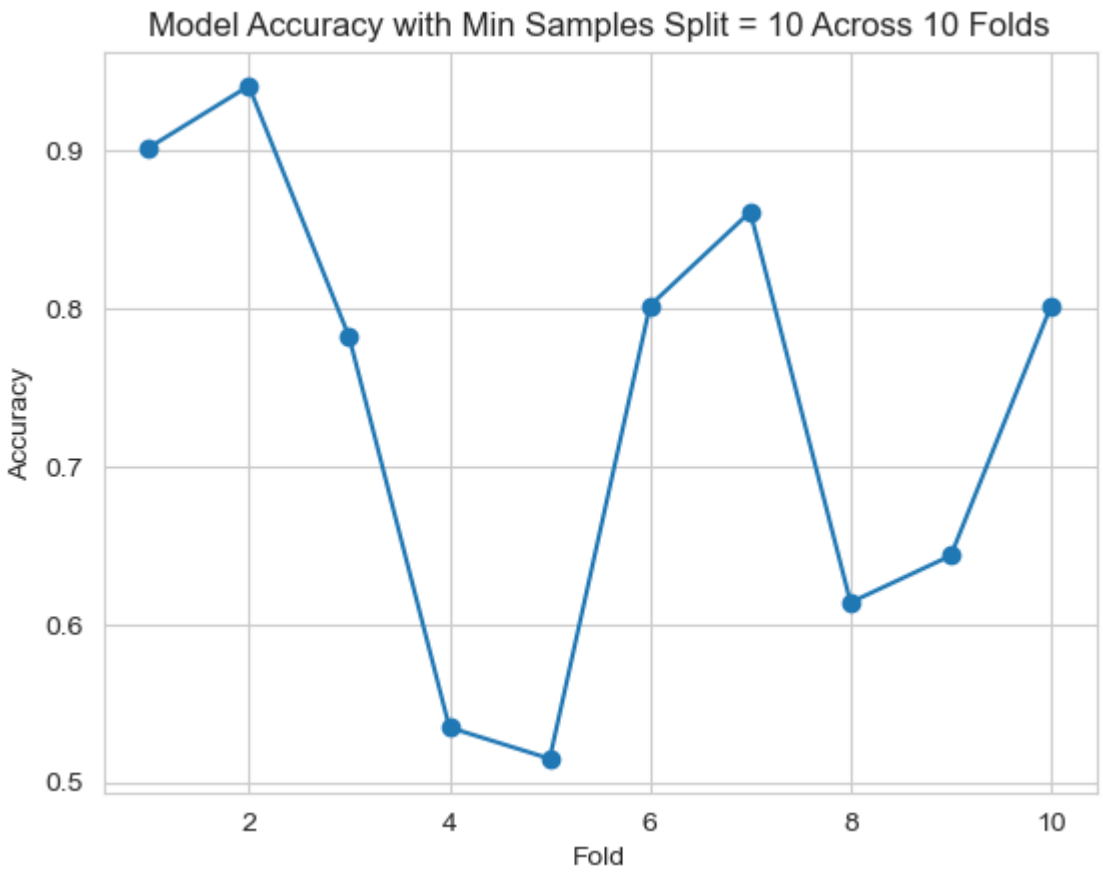
Fold 7: 0.7030

Fold 8: 0.5743

Fold 9: 0.6337

Fold 10: 0.7030

Second parameter adjustment using ‘min_samples_split’



Accuracy Scores with Min Samples Split = 10

Fold 1: 0.9020

Fold 2: 0.9412

Fold 3: 0.7822

Fold 4: 0.5347

Fold 5: 0.5149

Fold 6: 0.8020

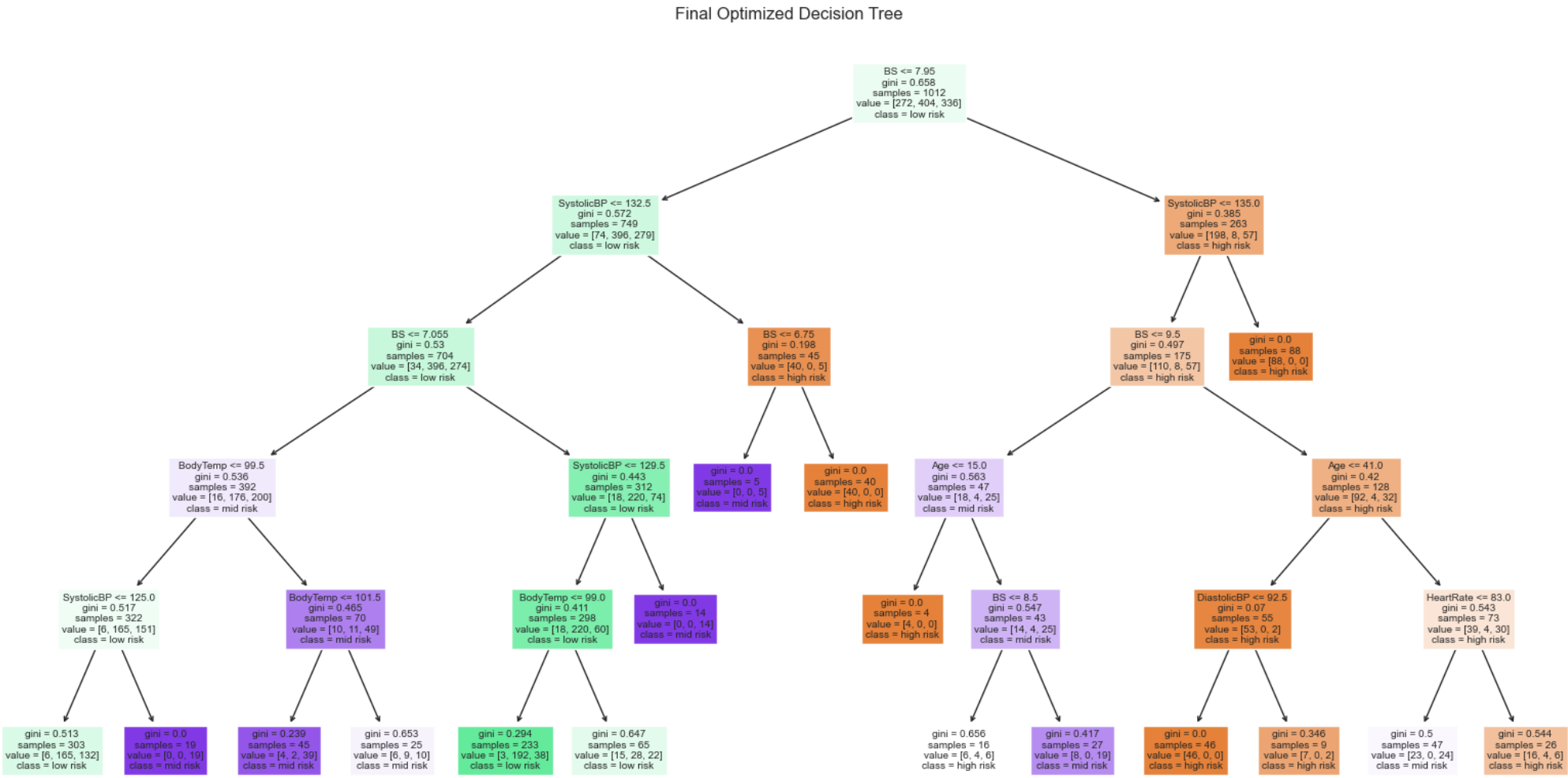
Fold 7: 0.8614

Fold 8: 0.6139

Fold 9: 0.6436

Fold 10: 0.8020

Final optimised classification tree



The tree has 33 nodes and a depth of 5. This means the decision tree has many specific decision rules because it has a lot of nodes. The tree depth of 5 tells us that it needed to make five levels of decisions to reach its final classification.

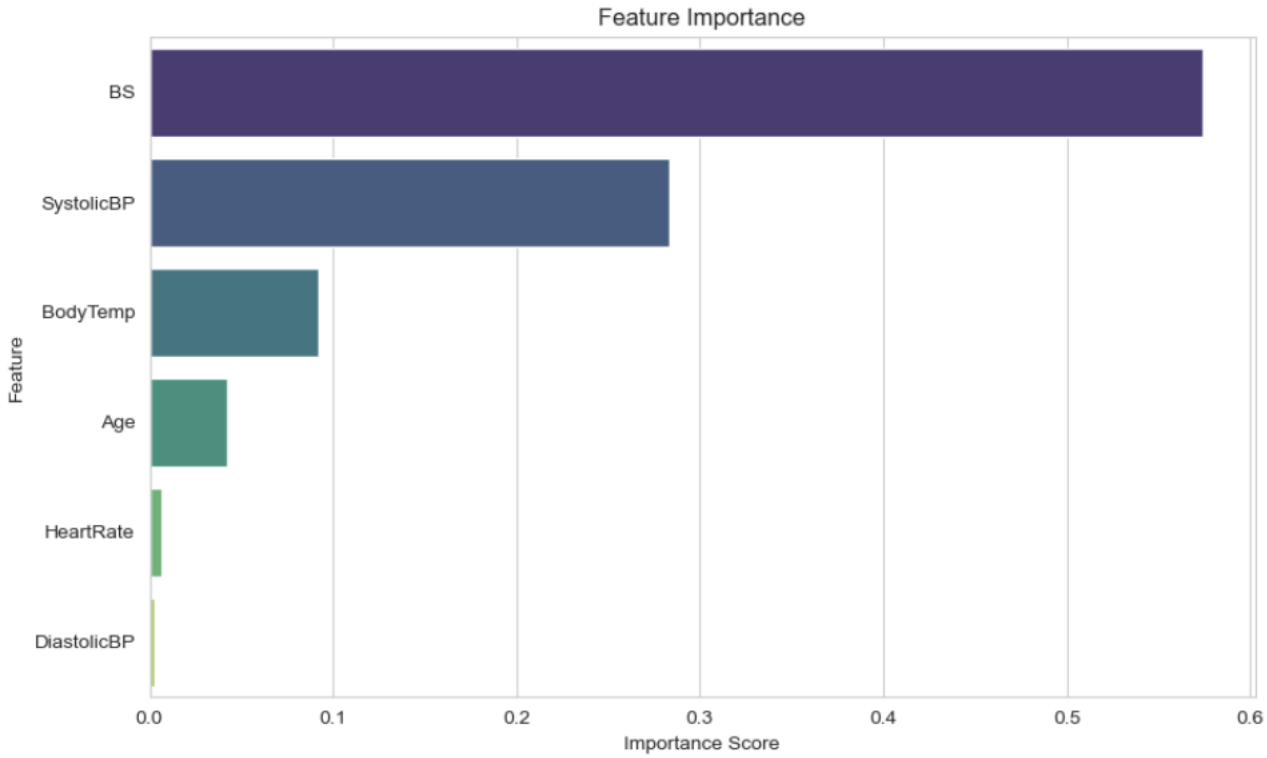
c) Describe the role of the two parameters in the model building you used in part b) above. Do you expect that using the same values obtained for this dataset will improve the accuracy of other datasets? Justify your answer. [8 marks]

The role of the max depth parameter is to control the maximum depth to which the decision tree is capable of growing. Having a tree depth of 5 in our model prevents overfitting. Overfitting is when the model performs exceptional at training data but not good in testing data. On the other hand, the role of the “min_samples_split” is to set the minimum number of samples needed to split an internal node. While setting the max depth lower

prevents overfitting, in min sample split, having a higher number of splits prevents overfitting. The purpose of this is to make sure that nodes show more generalised patten and not just noise in the data. Using the same values obtained for this dataset may not improve the accuracy of other data set. This is because the nature of other data sets are unique,these include the relationships, size and etc. This may make the model performance poor.

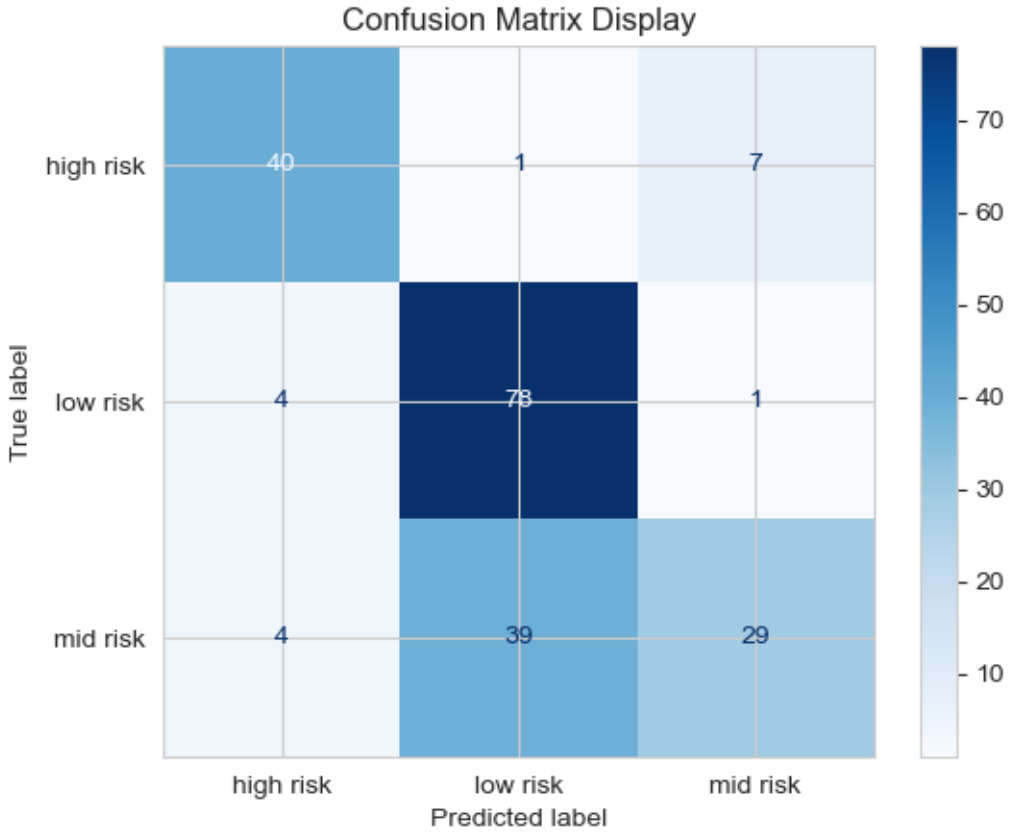
d) Find the feature importance based on the final classification model and explain your findings. [5 marks]

```
Feature Importance Scores:
       Feature  Importance
3           BS    0.573822
1   SystolicBP    0.282968
4     BodyTemp    0.091887
0          Age    0.042570
5    HeartRate    0.006358
2  DiastolicBP    0.002395
```



Based on the final classification model, the feature with the highest importance score of 0.574 is Blood Sugar (BS). This means that blood sugar has a significant effect in predicting the maternal health risk level of pregnant women. Following BS is the SystolicBP with an importance score of 0.283. This also signifies that systolic blood pressure also has a strong influence to predict health risk level. This is followed by body temperature, age, heart rate and diastolic blood pressure for the last one. Diastolic BP being the lowest importance score may suggest that it has a lowest impact out of all features in predicting health risk level.

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). [10 marks]



Interpretation:

High risk: 40 instances were accurately identified as “high risk”, 1 was misidentified as “low risk”, and 7 instances were misidentified as “mid risk”.

Low risk: 78 instances were correctly predicted as “low risk, 1 was incorrectly predicted as “mid risk”, and 4 were misidentified as “high risk”.

Mid risk: Only 29 instances of mid risk were accurately predicted as “mid risk”, 4 instances were misidentified as “high risk”, and 39 instances of “mid risk” were wrongly predicted as “low risk”.

Model Summary Report:

	precision	recall	f1-score	support
high risk	0.83	0.83	0.83	48
low risk	0.66	0.94	0.78	83
mid risk	0.78	0.40	0.53	72
accuracy			0.72	203
macro avg	0.76	0.73	0.71	203
weighted avg	0.75	0.72	0.70	203

Discussion :

The precision for high risk is 0.83 or 83%. This means that out of all instances, 83% identified were accurate predictions. For medium risk we have a percentage of 78% of the instances that were predicted correctly. On the other hand, the precision for predicting low risk is 66% . The recall tells if the model were able to correctly identify all the positive instances. The recall percentage for high risk is 0.83 too. This means that the model accurately identified 83% of the whole actual instances of high risk. For mid risk the recall is quite low which is at 40%. This may imply that our model had difficulty correctly recognising all mid risk instances. However, we have a high percentage of recall for low risk, at 94%, this means that 94% were identified correctly. The F1 score tells us about the balance involving precision and recall. In the high risk level we have a high percentage 83% which demonstrates excellent balance between precision and recall. This is important as it guarantees that the model’s prediction capability is both accurate and thorough. Second high F1 score is the low risk with a percentage of 78% which also suggests good balance. The lowest score would be the mid risk with an f1 score of 0.53. This indicates that the model had difficulty in minimizing inaccurate projections while accurately identifying every instance of "mid risk". The model’s overall accuracy is 72%, which represents the percentage of properly predicted instances throughout all classes.

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

Describe and analyse your classification results. Compare the performance of the models and explain which performed better and why. Evaluate the performance using confusion matrices, recall, precision, and accuracy metrics.

In conclusion, the classification results provided us with insights into how well the model performed. The decision tree model exhibited excellent overall performance, particularly in correctly predicting the "high risk" and "low risk" classifications. With the help of confusion matrix, recall, precision and accuracy metrics we were able to find out that the overall accuracy of 72%. While the model was giving accurate predictions for “high” and “low” risk instances, we found out that it was struggling to predict “mid risk” instances as we can see from the model’s recall.