Assignment 1

Data Exploration and Classification

Semester 1 2024

Student Name: Nikisha Chhima

Student ID: 22168889

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.**

Data Exploration and Classification:

Full Name: Nikisha Chhima
Student ID: 22168889

Table of Contents

1. Introduction 4

2. Data Exploration 4

2.1. How many features (attributes) and instance exist, and what data types are these? 4

2.2. Provide summary statistics 5

2.3. Perform an initial exploration 5

2.4. Illustrate the features of your dataset 6

2.5. Explain what you have learnt from the data exploration and visualizations 7

3. Classification Models.. 8

3.1. Preprocessing steps 8

3.2. Create a model using the Decision Tree algorithm and adjust two suitable parameters (one at a time) to reduce the tree's size 8

3.3. Describe the role of the two parameters in the model building 12

3.4. Feature Importance based on the final classification model 13

3.5. Generate the Confusion Matrix, provide the model summary report and discuss the metrics (accuracy, precision, recall, and F1-score) 13

4. Results and Discussion 14

List of Figures/Tables:

1. Tables:... 4

1.1. Features, Instances and Data Types 4

1.2. Summary Statistics.. 5

1.3. Initial Exploration.. 5

1.4. Feature Importance 13

2. Figures:.. 6

2.1. Feature Visualizations 6

2.2. Before optimizing Decision Tree ('max_depth'=8) 8

2.3. 10-Fold Cross Validation Scores for 'max_depth' 9

2.4. After optimizing Decision Tree ('max_depth'=6)... 10

2.5. Before optimizing 'max_leaf_nodes' Decision Tree ('max_depth'=6) 10

2.6. 10-Fold Cross Validation Scores for ‘max_leaf_nodes’ 11
2.7. Final Optimized Decision Tree 12
2.8. Confusion Matrix and Model Summary Report.. 13

1. Introduction:

The dataset I have chosen for this assignment is the “Maternal Health Risk” dataset. The pregnancy cycle has numerous aspects to investigate, though this chosen dataset specifically focuses on 7 features. By investigating these features, we will be able to determine possible risk factors and relationships that may influence maternal health risks.

I have 2 research questions that I want to answer throughout this report:

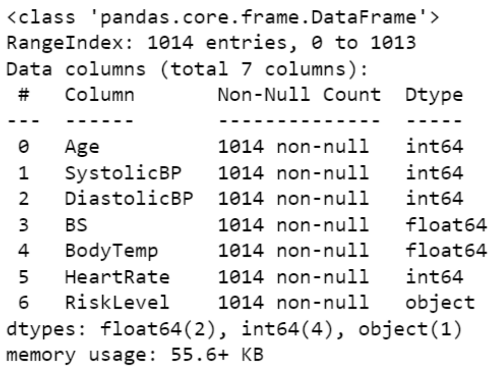
1. What are the most influential features that impact the target variable?
2. By how much does the accuracy results increase when you adjust two parameters in your Decision Tree?

The assumption for this report is that the data collected is accurate and precise. I am also assuming the information collected is from a diverse range of pregnant women at random.

2. Data Exploration: 580 words

2.1. How many features (attributes) and instance exist, and what data types are these?

1.1: Features, Instances and Data Types:

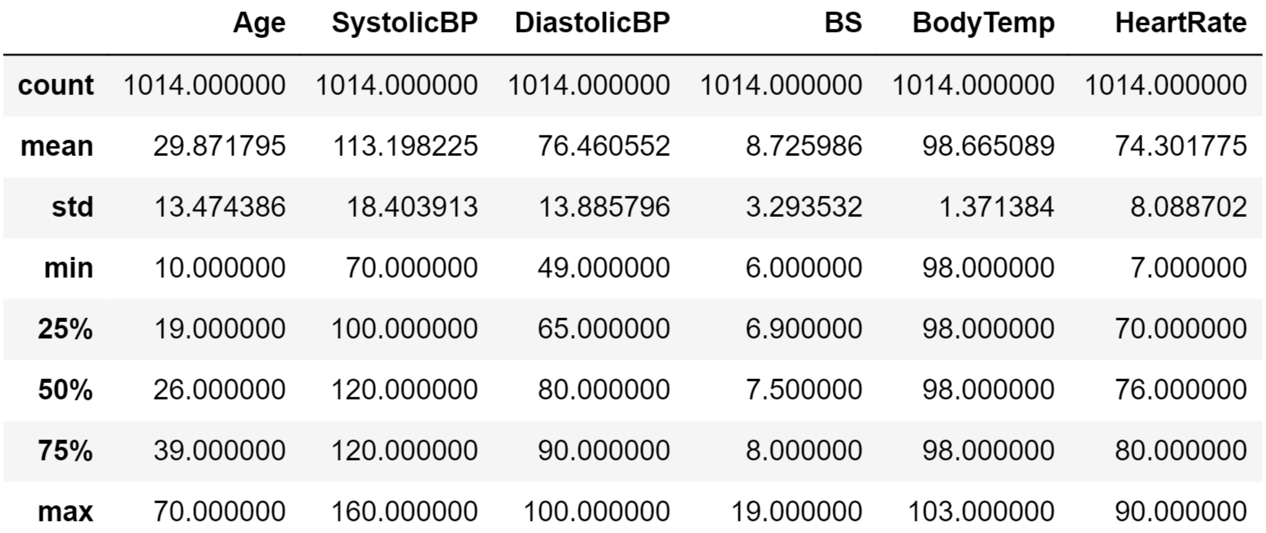


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1014 entries, 0 to 1013
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          1014 non-null   int64
 1   SystolicBP   1014 non-null   int64
 2   DiastolicBP  1014 non-null   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
dtypes: float64(2), int64(4), object(1)
memory usage: 55.6+ KB
```

In the Maternal Health Risk dataset there are 7 features which are named “Age”, “SystolicBP”, “DiastolicBP”, “BS”, “BodyTemp”, “HeartRate”, and “RiskLevel”. The data types for features “Age”, “SystolicBP”, “DiastolicBP”, and “HeartRate” are int64 (discrete data type). For features “BS” and “BodyTemp” they are float64 (continuous data type) and for “RiskLevel” it is categorical. The number of instances in this dataset is 1014 recorded observations.

2.2. Provide summary statistics:

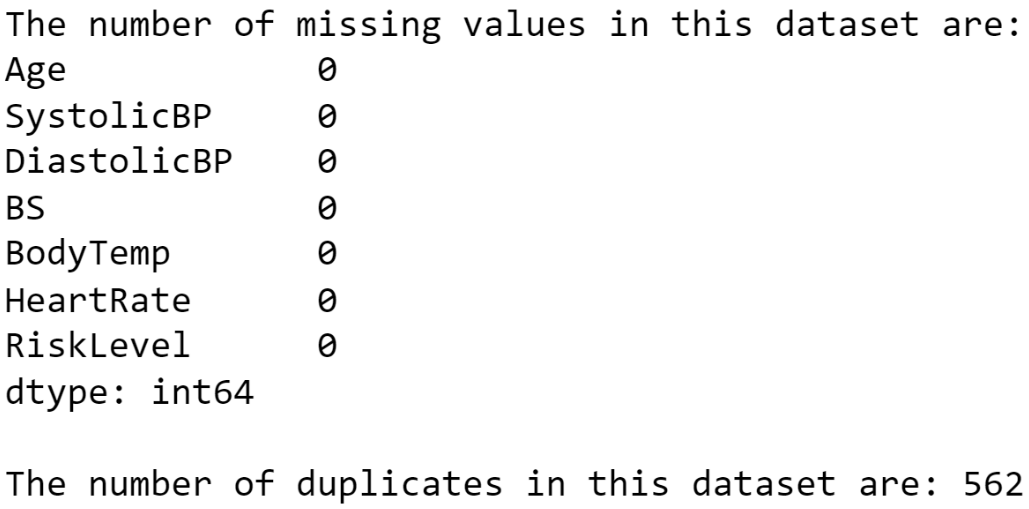
1.2: Summary Statistics:



	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

2.3. Perform an initial exploration:

1.3: Initial Exploration:

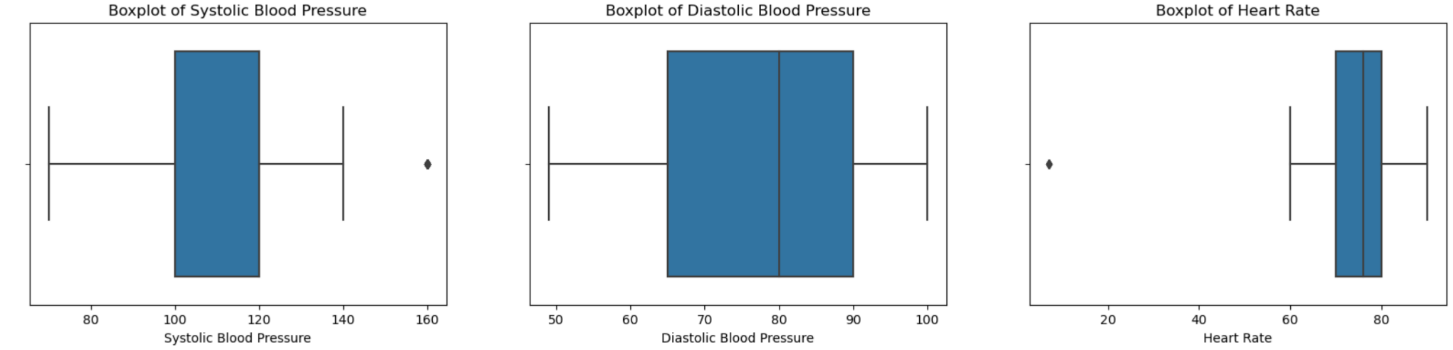


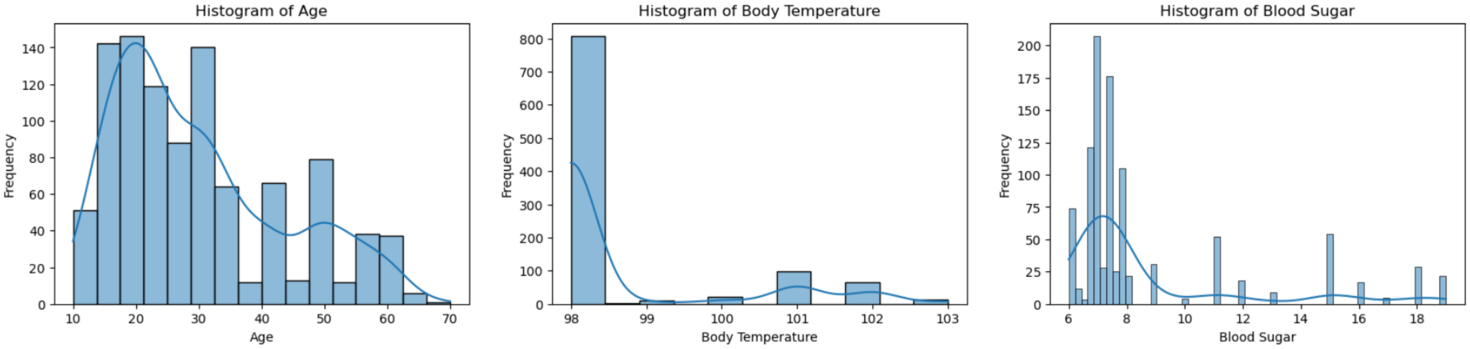
```
The number of missing values in this dataset are:
Age            0
SystolicBP     0
DiastolicBP    0
BS             0
BodyTemp       0
HeartRate      0
RiskLevel      0
dtype: int64

The number of duplicates in this dataset are: 562
```

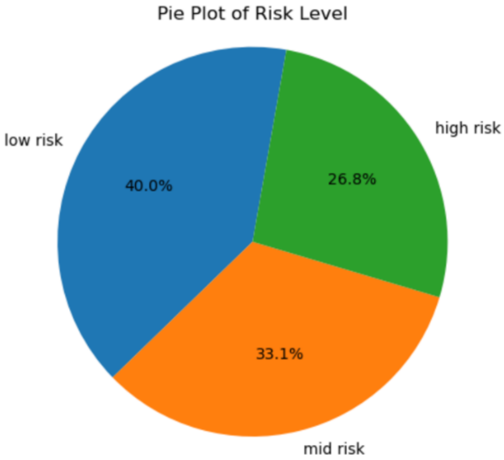
To assess the cleanliness of my dataset, I checked for missing values, duplications, and potential outliers. In this dataset there are no missing values, however, there are 562 duplicates. From the summary statistics, by investigating the minimum and maximum values I can identify potential outliers. For example, there may be an outlier in BodyTemp at 103 Fahrenheit or in SystolicBP at 160 as both these values are unusually high.

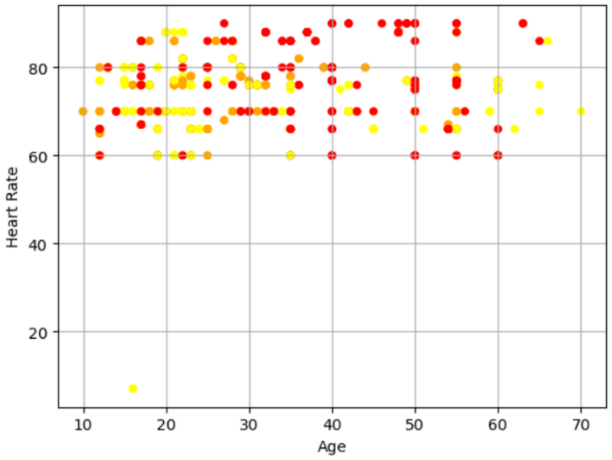
2.4. Illustrate the features of your dataset:

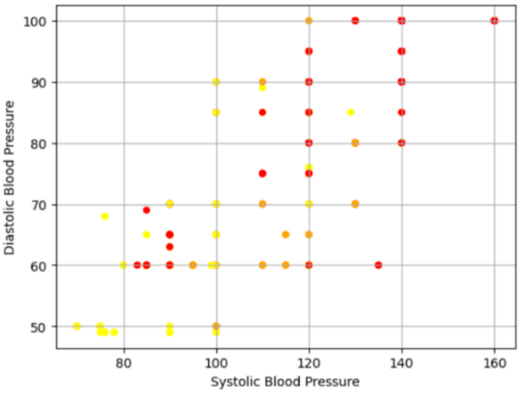
2.1: Feature Visualizations:











2.5. Explain what you have learnt from the data exploration and visualisations:

From the visualisations I can see the distribution for each feature and their relationships between one another in the grouped scatter plots. In the first 3 boxplots, I displayed Systolic Blood Pressure, Diastolic Blood Pressure, and Heart Rate. The Systolic Blood Pressure boxplot shows a range from 70 to 140 mmHg with an outlier at 160 mmHg. This could occur as an entry error considering its extreme value. The Diastolic Blood Pressure boxplot shows a range from 49 to 100 mmHg, having a large variety displaying no outliers. Lastly, the Heart Rate boxplot ranges from just below 60 to 90 bpm with an outlier at 7 bpm nearing flat line, this is very unlikely to occur, so it must be an entry error.

In the 3 histograms I displayed Age, Body Temperature and Blood Sugar. The Age histogram shows how most pregnant women are between 19 to 38 years old. This is accurate as womens' bodies are more capable to becoming pregnant during this period. After 35 years of age women can still become pregnant, however, it is less likely. Two outliers are shown at ages 10 and 70. The Body Temperature histogram shows majority of the temperatures are recorded around 98 to 98.5 Fahrenheit. Outliers in this graph are at 99, 100, 101, 102, and 103 Fahrenheit, this could be caused by potential health conditions. Lastly, the histogram of Blood Sugar shows a majority of the records from 6 to 9. It shows 9 outliers which lie at 10, 11, 12, 13, 15, 16, 17, 18, and 19.

The Risk Level feature is visualised in a pie chart as its data type is categorical. It shows that 40% of distribution is at low risk, 33.1% at mid risk and 26.8% at high risk.

For the grouped scatter plot of Age vs HeartRate grouped by RiskLevel. It shows low risk ranging from ages 10-25 with occasional high risks closer to the lower ages. From ages 26-60 there are both mid and high risk levels points. This could be because womens' bodies become weaker as they age, making it more difficult to carry a baby. For Heart Rate there are more low risks when it's around 70 bpm or below, when the heart rate reaches to 80 and over there are high risks.

From the grouped scatter plot of Systolic Blood Pressure vs Diastolic Blood Pressure grouped by RiskLevel, it shows low risks from below 100 mmHg Systolic Blood Pressure to below 70 Diastolic Blood Pressure. I can see more mid to high risks when ranging from above 120 mmHg in terms of Systolic Blood Pressure and above 80 mmHg for Diastolic Blood Pressure.

3. Classification Models:

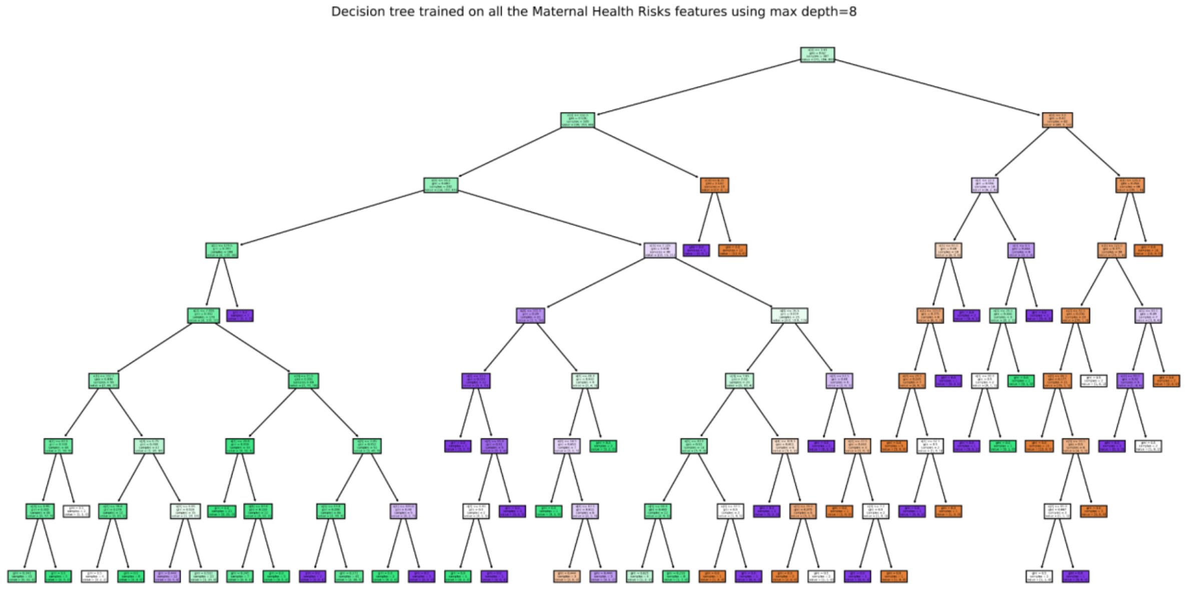
3.1. Preprocessing steps:

The outliers were reviewed and removed as they were interfering with the data quality rather than adding informative data. After removing them, there are now 538 duplicates rather than 562. Due to this large number, I analysed each duplicate closely and decided to remove them as well. I believe that keeping the duplicates would have resulted in lowered data integrity and weakened data quality. There were no missing values.

3.2. Create a model using the Decision Tree algorithm and adjust two suitable parameters (one at a time) to reduce the tree's size:

2.2: Before optimizing Decision Tree ('max_depth' = 8):

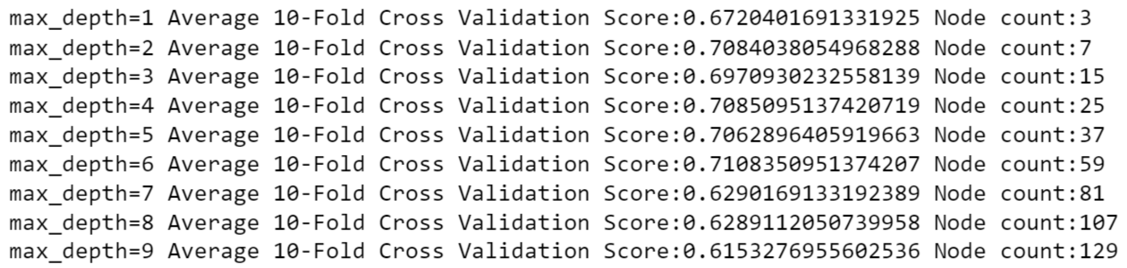
```
Number of tree nodes before optimizing max depth:  107
```


Decision tree trained on all the Maternal Health Risks features using max depth=8

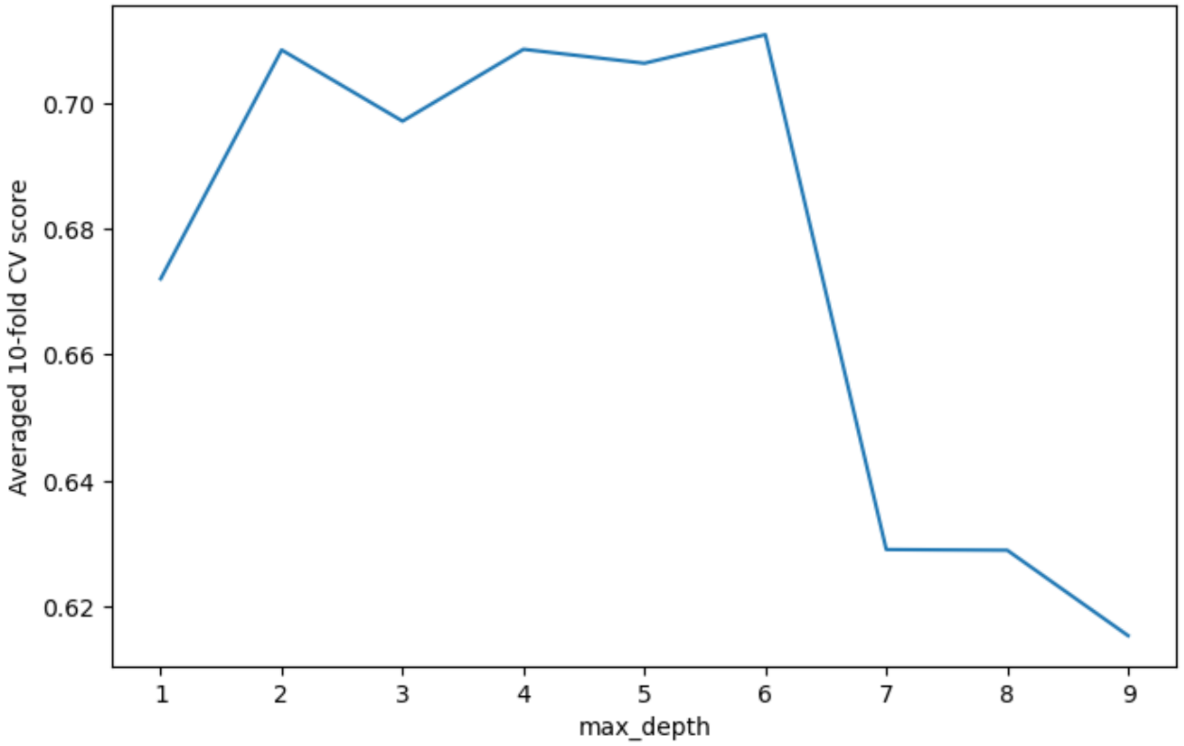
```
Accuracy score of our model with Decision Tree: 0.67
Precision score of our model with Decision Tree : 0.67
Recall score of our model with Decision Tree : 0.67
F1 score of our model with Decision Tree : 0.67
```

Before optimizing my model, I decided to experiment with a 'max_depth' value of 8, which gave an accuracy of 0.67. As the model increases in depth, the interpretability decreases. With 107 tree nodes, the interpretability decreases, however, it still indicates its complexity and capacity to capture patterns.

2.3: 10-Fold Cross Validation Scores for ‘max_depth’:



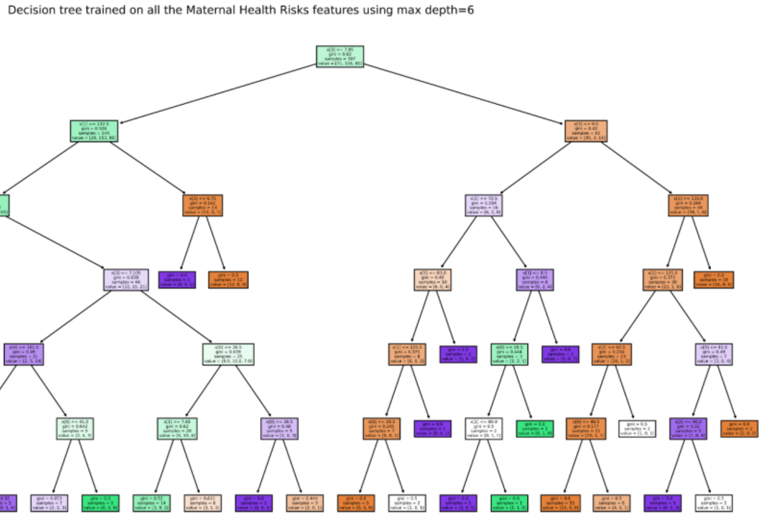
```
max_depth=1 Average 10-Fold Cross Validation Score:0.6720401691331925 Node count:3
max_depth=2 Average 10-Fold Cross Validation Score:0.7084038054968288 Node count:7
max_depth=3 Average 10-Fold Cross Validation Score:0.6970930232558139 Node count:15
max_depth=4 Average 10-Fold Cross Validation Score:0.7085095137420719 Node count:25
max_depth=5 Average 10-Fold Cross Validation Score:0.7062896405919663 Node count:37
max_depth=6 Average 10-Fold Cross Validation Score:0.7108350951374207 Node count:59
max_depth=7 Average 10-Fold Cross Validation Score:0.6290169133192389 Node count:81
max_depth=8 Average 10-Fold Cross Validation Score:0.6289112050739958 Node count:107
max_depth=9 Average 10-Fold Cross Validation Score:0.6153276955602536 Node count:129
```



The 10-Fold cross validation for ‘max_depth’ tests for its optimal value. The optimal value is 6 as it gives the highest score of 0.71.

2.4: After optimizing Decision Tree (‘max_depth’ = 6):

```
Number of tree nodes after optimizing max depth:  59
```

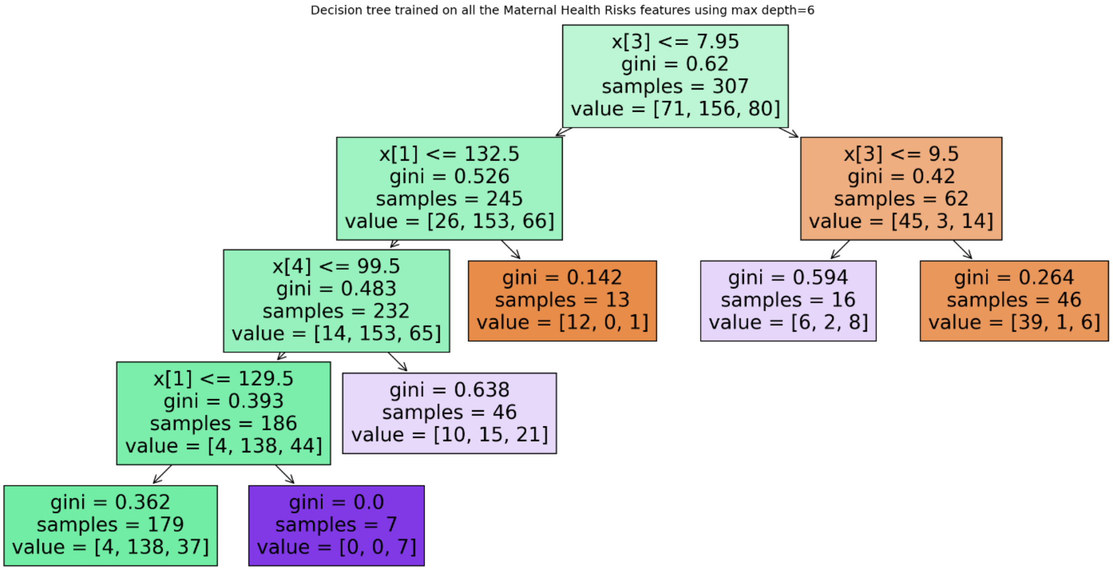

Decision tree trained on all the Maternal Health Risks features using max depth=6

```
Accuracy score of our model with Decision Tree: 0.73
Precision score of our model with Decision Tree : 0.73
Recall score of our model with Decision Tree : 0.73
F1 score of our model with Decision Tree : 0.73
```

The number of tree nodes has decreased by 48 by using the ‘max_depth’ optimal value. This increases interpretability and accuracy as this model provides a value of 0.73 in accuracy, which is a 0.06 increase.

2.5: Before optimizing ‘max_leaf_nodes’ Decision Tree (‘max_depth’ = 6):

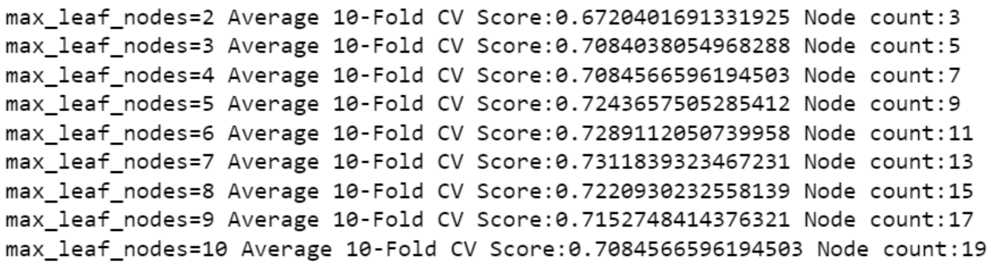
```
Number of tree nodes after optimizing max depth:  11
```



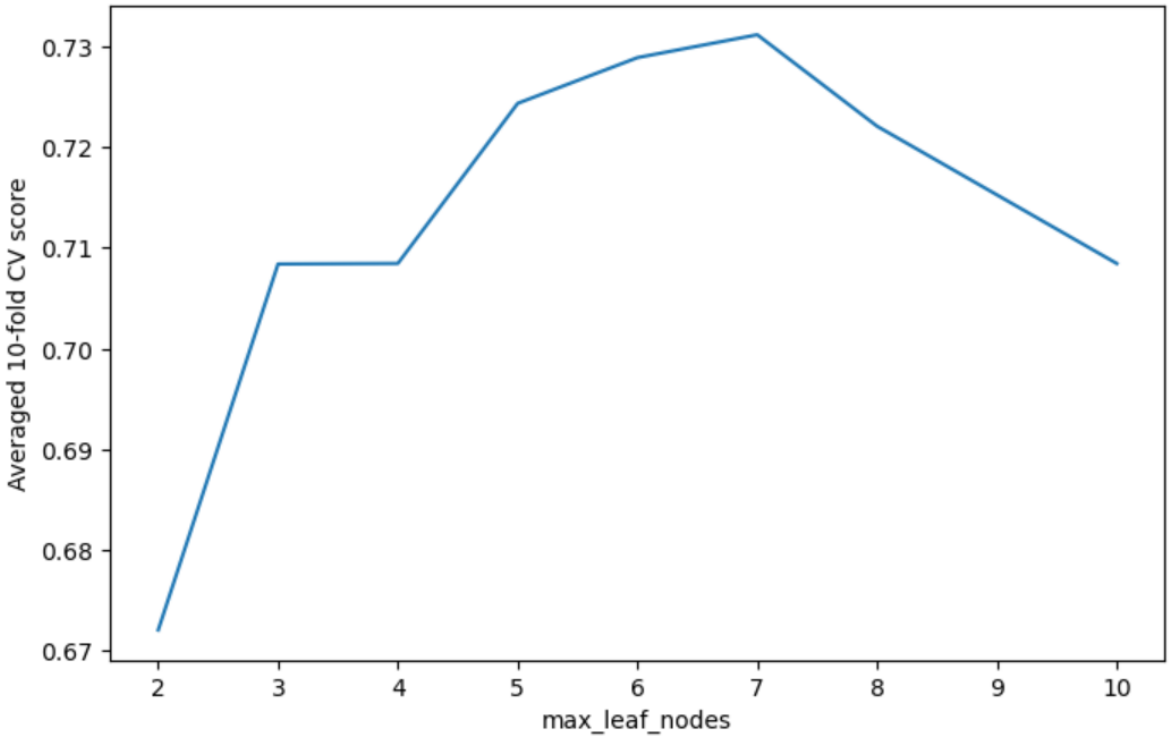
```
Accuracy score of our model with Decision Tree: 0.73
Precision score of our model with Decision Tree : 0.73
Recall score of our model with Decision Tree : 0.73
F1 score of our model with Decision Tree : 0.73
```

To optimize the model further, I experienced with the ‘max_leaf_nodes’ parameter. By setting the model to 6 ‘max_leaf_nodes’ it provided 11 tree nodes, keeping the accuracy at 0.73.

2.6: 10-Fold Cross Validation Scores for ‘max_leaf_nodes’:



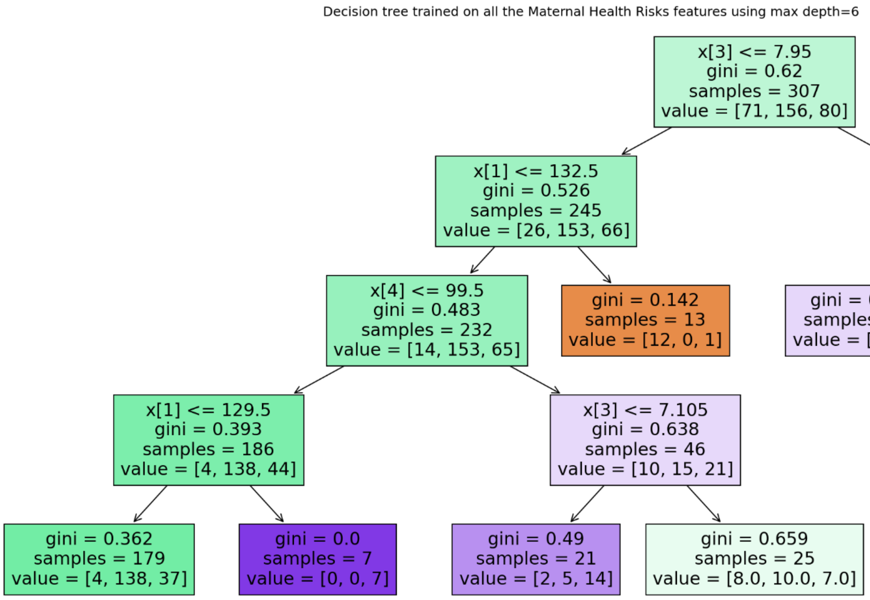
```
max_leaf_nodes=2 Average 10-Fold CV Score:0.6720401691331925 Node count:3
max_leaf_nodes=3 Average 10-Fold CV Score:0.7084038054968288 Node count:5
max_leaf_nodes=4 Average 10-Fold CV Score:0.7084566596194503 Node count:7
max_leaf_nodes=5 Average 10-Fold CV Score:0.7243657505285412 Node count:9
max_leaf_nodes=6 Average 10-Fold CV Score:0.7289112050739958 Node count:11
max_leaf_nodes=7 Average 10-Fold CV Score:0.7311839323467231 Node count:13
max_leaf_nodes=8 Average 10-Fold CV Score:0.722093023255813 Node count:15
max_leaf_nodes=9 Average 10-Fold CV Score:0.7152748414376321 Node count:17
max_leaf_nodes=10 Average 10-Fold CV Score:0.7084566596194503 Node count:19
```



The 10-Fold cross validation for ‘max_leaf_nodes’ tests for its optimal value. The highest score of 0.73 is achieved when ‘max_leaf_nodes’ is at 7.

2.7: Final Optimized Decision Tree:

```
Number of tree nodes after optimizing max depth:  13
```



```
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
F1 score of our model with Decision Tree : 0.74
```

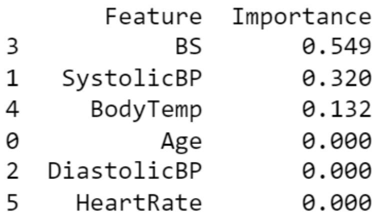
In the final Decision Tree model, we have optimized both parameters, ‘max_depth’ at 6 and ‘max_leaf_nodes’ at 7. This model gives better accuracy than the previous models with a value of 0.74. The number of tree nodes is 13 displaying complexity along with good interpretability.

3.3. Describe the role of the two parameters in the model building:

The two parameters that I adjusted was ‘max_depth’ and ‘max_leaf_nodes’. The ‘max_depth’ value controls the depth of the decision tree, the larger the ‘max_depth’, the more complex the tree will become. Before finding the optimal value, the tree was too large and complex to interpret, increasing the risk of overfitting. Inaccurate findings arise when overfitting takes place because the system learns noise from the training set rather than the underlying relationships. The ’max_leaf_nodes’ value controls the maximum number of node splits that the tree can make. This also relates to overfitting, by limiting the number of leaf nodes it prevents the tree from becoming overly complex, which overall reduces the risk of overfitting and increases the accuracy. For the ‘max_depth’ and ‘max_leaf_nodes’ values in this dataset, I do not expect these same values to improve the accuracy for other datasets. Due to the fact that datasets can have more complex data or more diverse information. The ‘max_depth’ and ‘max_leaf_nodes’ values are not the same for every dataset, so a different dataset would show different values for those parameters. The optimal values obtained are specific to this dataset, hence why it produces a more accurate model.

3.4. Feature importance based on the final classification model:

1.4: Feature Importance:

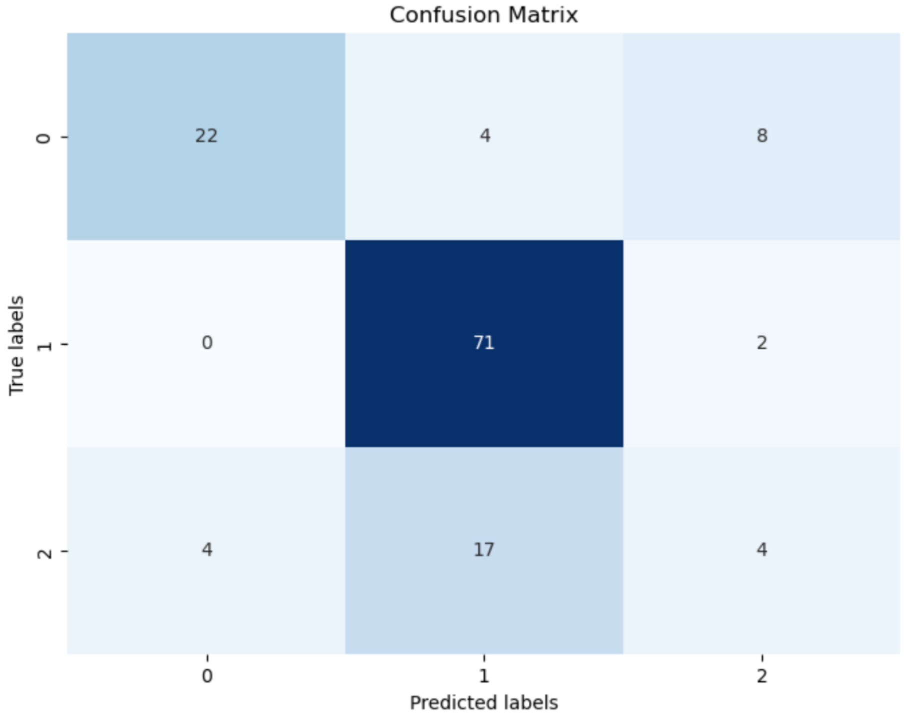


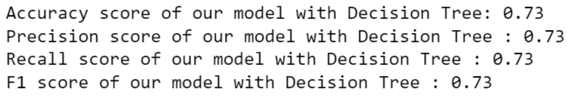
```
     Feature  Importance
3         BS       0.549
1  SystolicBP       0.320
4   BodyTemp       0.132
0        Age       0.000
2 DiastolicBP      0.000
5  HeartRate       0.000
```

Based on the final model I found the feature importance. Blood Sugar has the highest level of importance with a score at 0.549, this contributes moderately towards predictions of the RiskLevel variable overall.

3.5. Generate the Confusion Matrix, provide the model summary report and discuss the metrics (accuracy, precision, recall and F1-score):

2.8: Confusion Matrix and Model Summary Report:





```
Accuracy score of our model with Decision Tree: 0.73
Precision score of our model with Decision Tree : 0.73
Recall score of our model with Decision Tree : 0.73
F1 score of our model with Decision Tree : 0.73
```

The diagonal elements on the confusion matrix are correctly predicted samples, so a total of 97 samples were correctly predicted out of 132 samples. This model has an accuracy of 0.73, which indicates that is it roughly 73% accurate overall. With a precision score of 0.73, the prediction of the target class (RiskLevel) is 73% accurate. The recall score tells us how many times we made the right prediction, so 73% were correctly predicted. Our model's F1 score is 0.73, indicating that its precision and recall are weighted on average by 73%.

4. Results and Discussion:

The final decision tree model had ‘max_depth’ at 6 and ‘max_leaf_nodes’ at 7, overall creating a tree with 13 nodes. The root nodes initial split is based on Blood Sugar, whether it is less than or equal to 7.95. The Gini impurity value is 0.62 indicating distribution across three classes, in this case, across low risk, mid risk and high risk. This node has 307 samples, with 71 in low risk, 156 in mid risk and 80 in high risk. 245 samples meet this condition and now tests with the Systolic blood pressure at 132.5 or lower. The Gini impurity is 0.526, with 26 in low risk, 153 in mid risk and 66 in high risk. Only 13 samples didn’t meet this condition, this node has a low Gini impurity of 0.142. 12 out of the 13 samples have low risk, with 0 at mid risk and 1 in high risk. Another node with low Gini impurity of 0.264 appears when 46 samples have a Blood Sugar level over 7.95 and have Blood Sugar over 9.5. There are 39 in low risk, 1 in mid risk and 6 in high risk. There is only 1 node with a Gini impurity of 0.0. This occurs when 7 samples had Blood Sugar lower than or equal to 7.95, had Systolic blood pressure lower than or equal to 132.5, had Body Temperature lower than or equal to 99.5, and Systolic blood pressure higher than 129.5. These 7 samples with those conditions are within the high risk class. This division shows that the data within the model has been successfully segregated according to the chosen feature and threshold, producing a very uniform subset. These types of nodes are crucial as this is where the tree correctly divides the classes, demonstrating the decision tree’s success in classifying the data.

The first model had the ‘max_depth’ set at 8 with no optimal values. This is the most complex, which increases the risk of overfitting. Overfitting is when the system learns noise from the training set rather than the underlying relationships, therefore it provides invalid results. To reduce overfitting, 10-fold cross validation was performed to gain the ‘max_depth’ optimal value. Setting this to its optimal value of 6 allows the model to provide accurate results. Accuracy improves as the first model gives 107 tree nodes whereas with the optimal ‘max_depth’ set, the model only contains 59 tree nodes. ‘max_leaf_nodes’ was set to a random value of 6, this model’s accuracy remained the same but the tree nodes reduced to 11. To increase accuracy, I performed another 10-fold cross validation to find ‘max_leaf_nodes’ optimal value. 7 gave the highest score of 0.73, by using this value it effectively reduces the risk of overfitting and increases the accuracy. The next model made have both ‘max_depth’ and ‘max_leaf_nodes’ set to their optimal values. The number of tree nodes is now 13, with an accuracy value of 0.74, showing a total increase of 0.07 between the first and final model. This confirms the best performed model is when both parameters are at their optimal values.

The confusion matrix enhances our understanding by comparing actual against predicted outcomes. From 132 samples, 97 samples were correctly predicted, making the accuracy around 73% overall. The precision score of 73% (0.73) indicates the percentage of “positive” predictions that came true. Recall estimates the percentage of positive class samples that the model accurately recognized, so 73% of the samples were correctly predicted. The F1 score represents the mean of precision and recall, furthermore the model’s reliability is accurately indicated by the score of 0.73.