Student ID Number: 18029208

Assignment 1

Data Exploration and Classification

Semester 1 2024

Student Name: William Bank Sukjaem
Student ID: 18029208
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**.

Contents

Table of Figures

Introduction to the dataset

The chosen data covers the topic of Maternal Health Risks in pregnant women with the data being collected from different maternal health care places in rural areas of developing countries, more specifically Bangladesh. The research relevant to the dataset analyses biometric data from wearable IoT devices from women during maternity with the goal being to identify, compare, and analyse relationships between the collected biometric data to mitigate and/or reduce maternal health risks. This analysis is under the assumption that the dataset used is accurate and that the women sampled are healthy.

1. Data Exploration

1.1 Data Types and Statistics Summary

The data set consists of 1014 instances and 7 attributes where 4 of the attributes are of type integer, 2 of type float, and 1 of type object. Six of the seven attributes are numerical being Age, SystolicBP, DiastolicBP, BloodSugar, BodyTemp, and HeartRate while the remaining is categorical being RiskLevel.

To briefly explain these attributes. The systolic blood pressure and diastolic blood pressure is the upper and lower values of blood pressure, respectively, which is measured in millimeter of mercury (mmHg). Blood sugar, which measures blood glucose levels (mmol/L), and body temperature, measured in Fahrenheit (°F).

	Age	SystolicBP	DiasolicBP	BloodSugar	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

Figure 1: Summary Statistics of continuous numerical features

The summary statistics in figure 1 provides key insight into biometric data of women in maternity and interesting statistical data. It is important to note that some of the data is not fully representative of the general population of women in maternity since the collection of data for the dataset was conducted within a single country.

In this dataset, the mean age of women in maternity is 29.8 years old with the lowest being 10 and maximum being 70. In this case, the minimum and maximum age are on the extreme ends of the spectrum which in reality will be quite rare when comparing amongst the general population indicating potential outliers. The mean age is normal when comparing to New Zealand's average age of mothers at the time of child birth (Statistica, 2023).

The mean blood sugar level is 8.7 mmol/L, with the median being 7.5 mmol/L. The minimum and maximum values are 6.0 mmol/L and 19.0 mmol/L, respectively. The blood sugar values have an extremely large range of 13 mmol/L.

According to the Mayo Clinic, an academic medical center, the baseline maximum for blood sugar level in pregnant women is around 7.8 mmol/L. This places the median within the range for normal blood sugar levels, while levels greater than 11.1 mmol/L show signs of diabetic blood sugar levels (Mayo Clinic, 2024). Although some values lie above 11.1mmol/L, some are well beyond the range of blood sugar levels of severe diabetic cases. The mean has a high value due to it being skewed by the high maximum value.

The mean heart rate is 74.3 bpm where the minimum is 7.0 bpm and maximum being 90 bpm. The minimum heart rate is abnormally low and is most likely an outlier. Heart rate is slowest when a person is sleeping which puts the heart rate at around 40 to 60 bpm. The range for heart rate is extremely large at 87 bpm.

	SystolicBP	DiasolicBP	RiskLevel
0	130	80	high risk
1	140	90	high risk
2	90	70	high risk
3	140	85	high risk
4	120	60	low risk
5	140	80	high risk
6	130	70	mid risk
7	85	60	high risk
8	120	90	mid risk
9	130	80	high risk

Figure 2: Table comparing Blood Pressure to Risk Level

For the blood pressure values, the mean SystolicBP is 113.20mmHg and DiasolicBP is 76.46mmHg. For the minimum and maximum, SystolicBP has values of 70mmHg and 160mmHg, and DiasolicBP has values of 49mmHg and 100mmHg. The mean values for blood pressure is normal and within the expected range healthy range of 120mmHg and 70mmHg or lower (Heart Foundation, n.d). However, when blood pressure exceeds 120mmHg, there is an overall higher risk of health complications shown when comparing RiskLevel to SystolicBP and DiasolicBP seen in figure 2 that shows first 10 rows of the dataset. High blood pressure is seen in women 20 weeks before or after who have Chronic Hypertension or Preeclampsia which is a serious disorder resulting from high blood pressure (ACOG, 2022). These abnormally high values, where SystolicBP is greater than 140mmHg, may be potential outliers in the dataset.

1.2 Dataset Cleanliness

```
Missing values:
 Age            0
SystolicBP      0
DiastolicBP     0
BloodSugar      0
BodyTemp        0
HeartRate       0
RiskLevel       0
dtype: int64
```

Figure 3: Missing Values

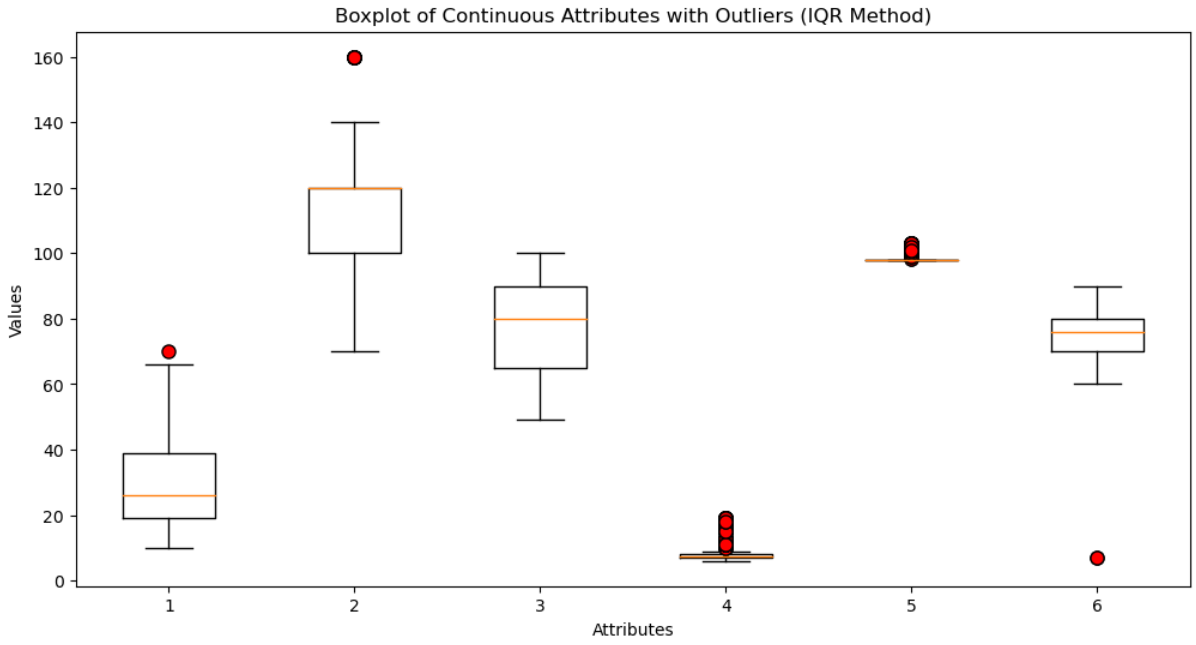

Figure 4: Boxplot of Continuous Attributes with identified outliers

In figures 3 and 4 helps in determining the cleanliness of the dataset. Figure 3 shows that there are no missing data in each attribute and in figure 4 shows the distribution of data points across each continuous attribute.

Upon looking at figure 4 and based on previous analysis on the statistics summary, there are certainly outliers that are not within reason of the general sample population of the dataset due to their extreme variance relative to the mean in their respective categories.

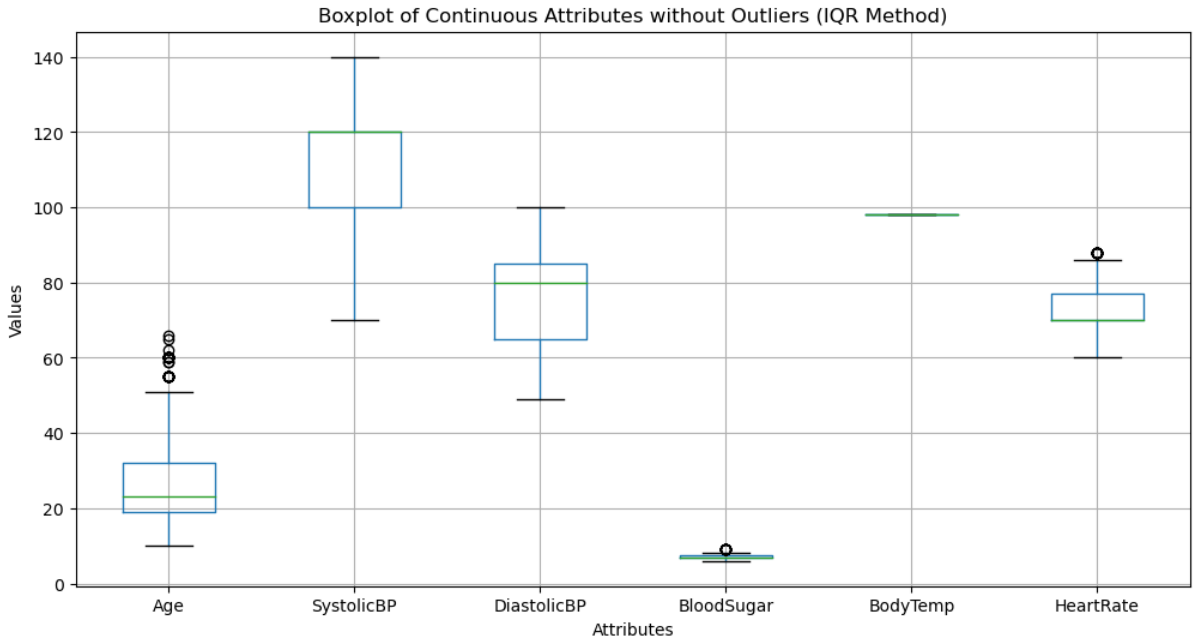

Figure 5: Boxplot of Continuous Attributes without Outliers

Since the box plots in figure 4 do not show signs of normality because of its asymmetry, the method of choice that is used to identify the outliers is the IQR method. The detected outliers are seen as red circles and are removed to have a cleaner dataset and better represent the general sample population as seen in figure 5.

The choice to remove these outliers are because these points lie greatly beyond what the human body would show in terms of biometrics. These cases are extremely unlikely to occur under the assumption that the women in the sample population are healthy. An example of this can be clearly seen through two categories: Heart Rate and Blood Sugar. The outliers for Blood Sugar go well beyond 11.1mmol/L which is extremely high even for diabetics and for heart rate, 7 bpm is extremely abnormal for high stress situations.

1.3 Illustration and Explanation of Features

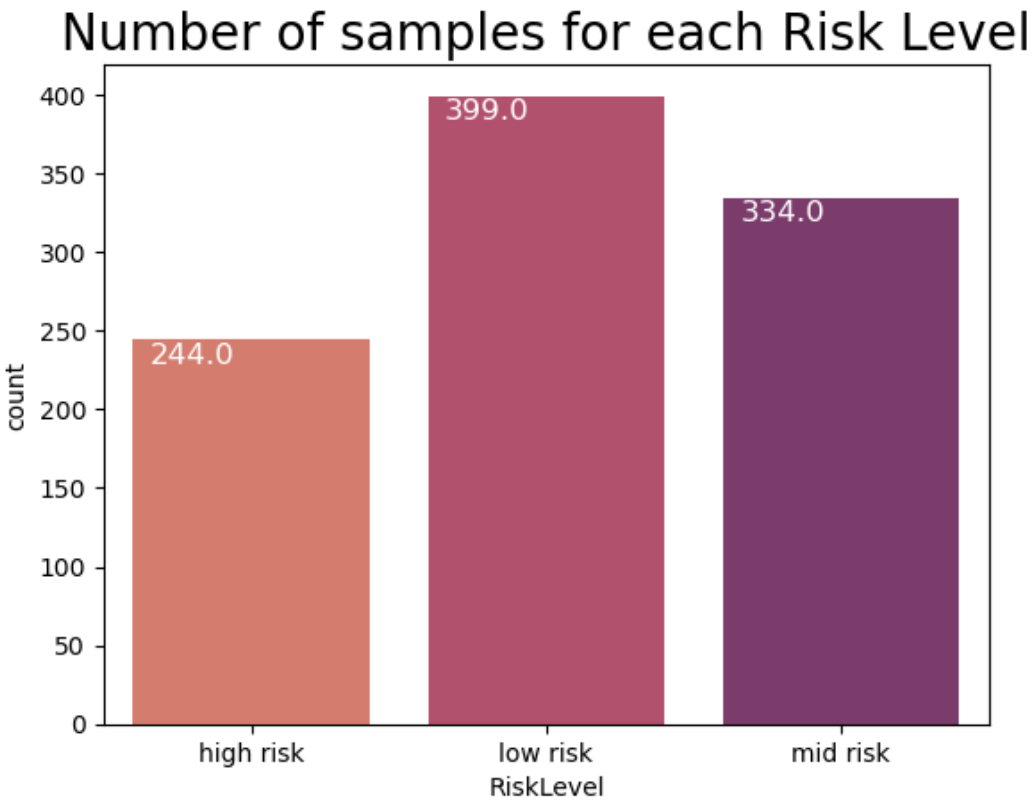

Figure 6: Count for each Risk Level

The risk levels in figure 7 show that most of the sample population of women are at a low or medium health risk when in maternity. There are still a significant number of women at a high risk but relative to the overall sample population, only 22.9% are at a high risk.

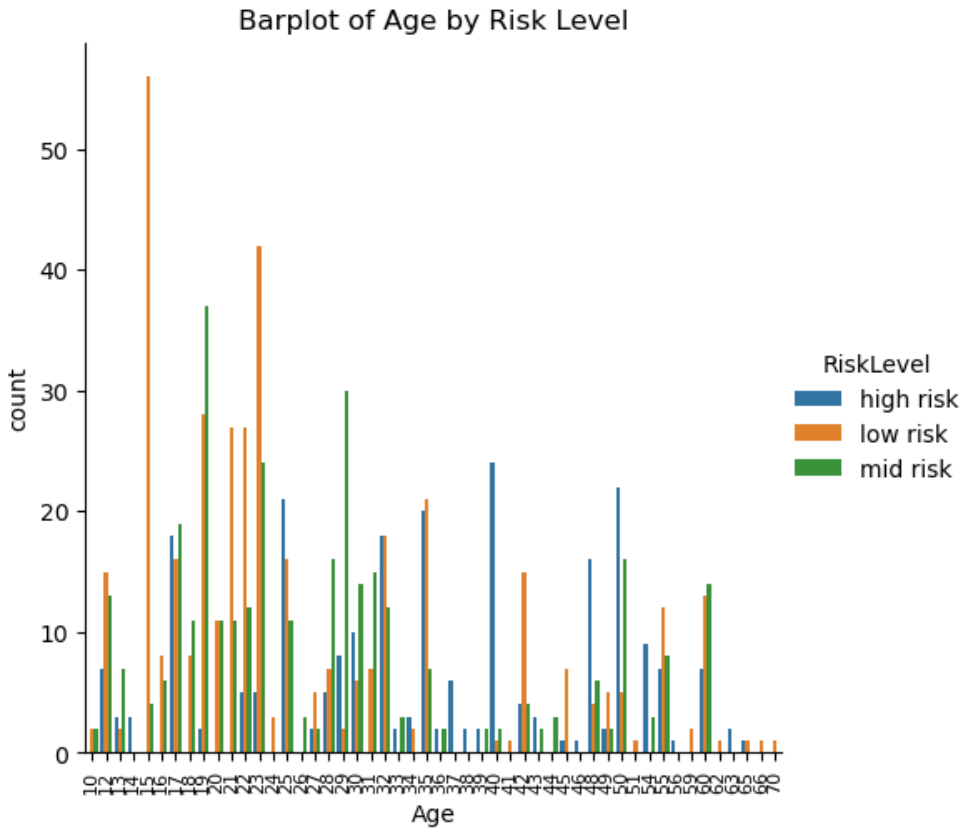

Figure 7: Age by Risk Level

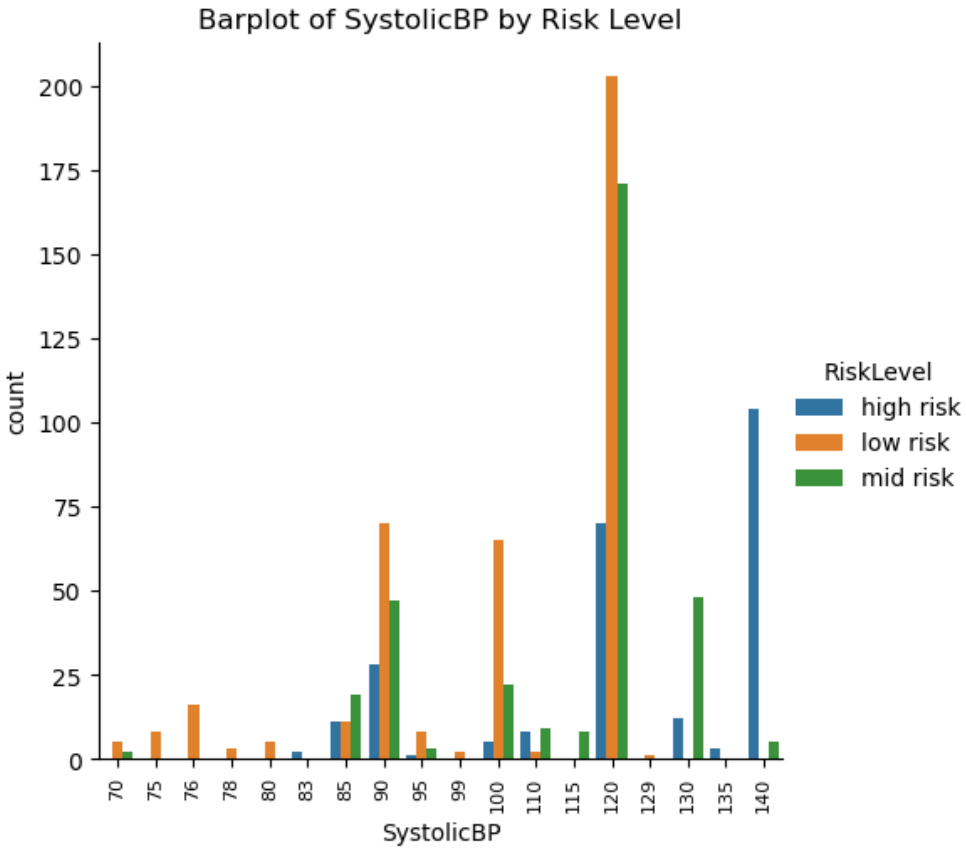

Figure 8: Systolic Blood Pressure by Risk Level

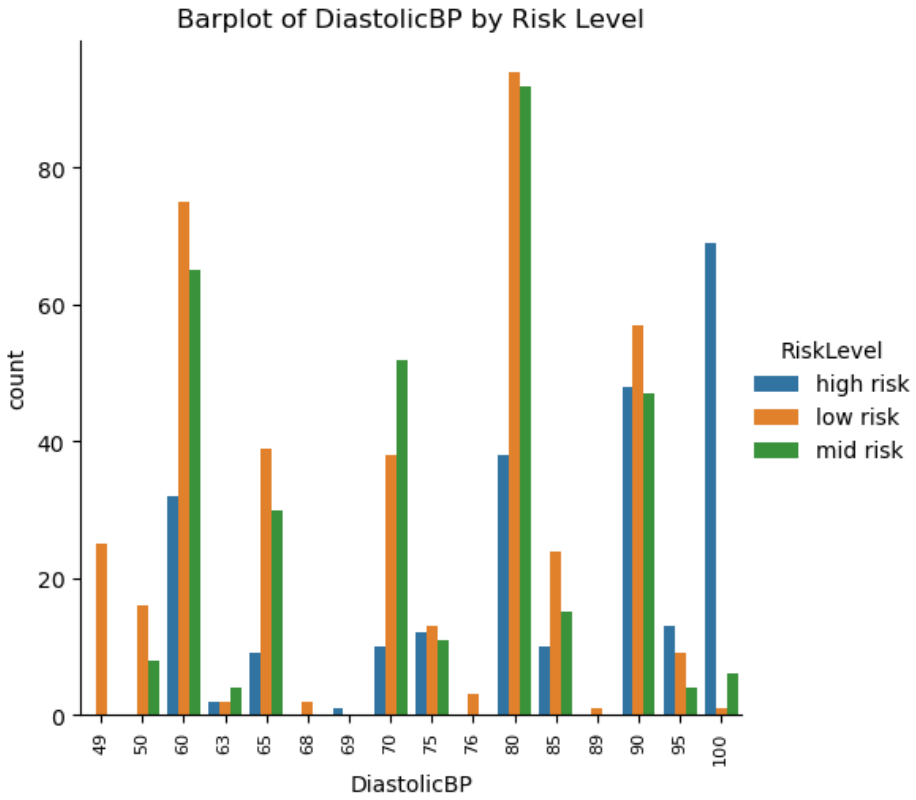

Figure 9: Diastolic Blood Pressure by Risk Level

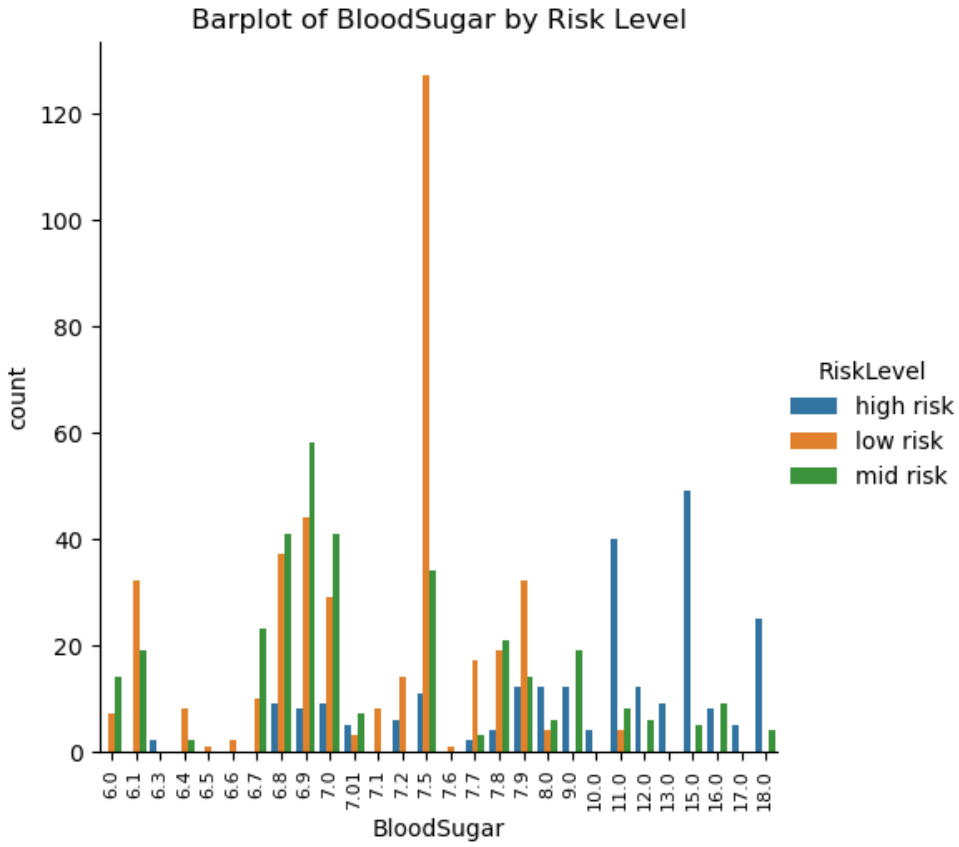

Figure 10: Blood Sugar by Risk Level

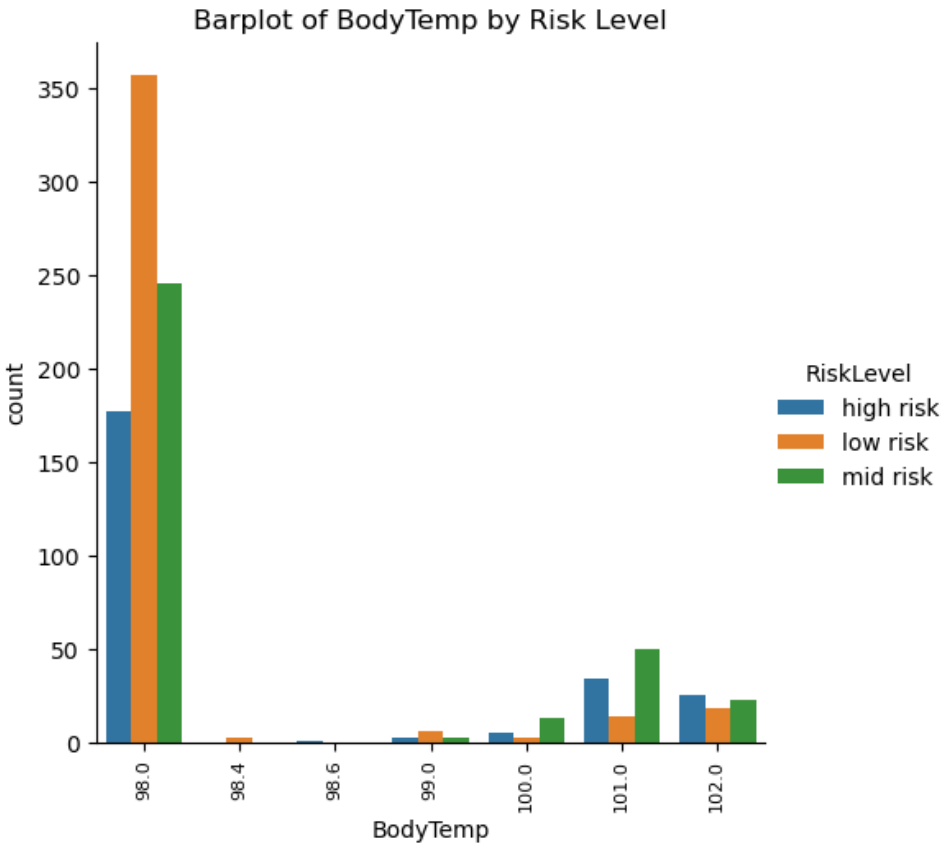

Figure 11: Body Temperature by Risk Level

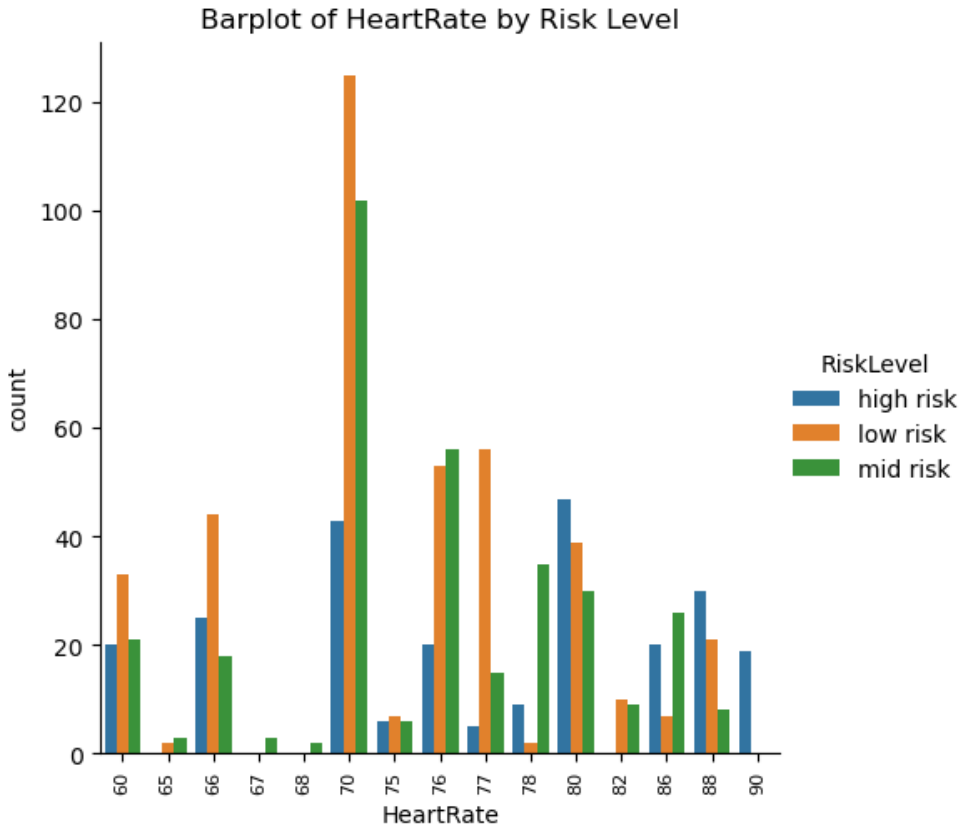

Figure 12: Heart Rate by Risk Level

Risk level in terms of age, there is a trend indicating increasing levels of health risks as you get older.

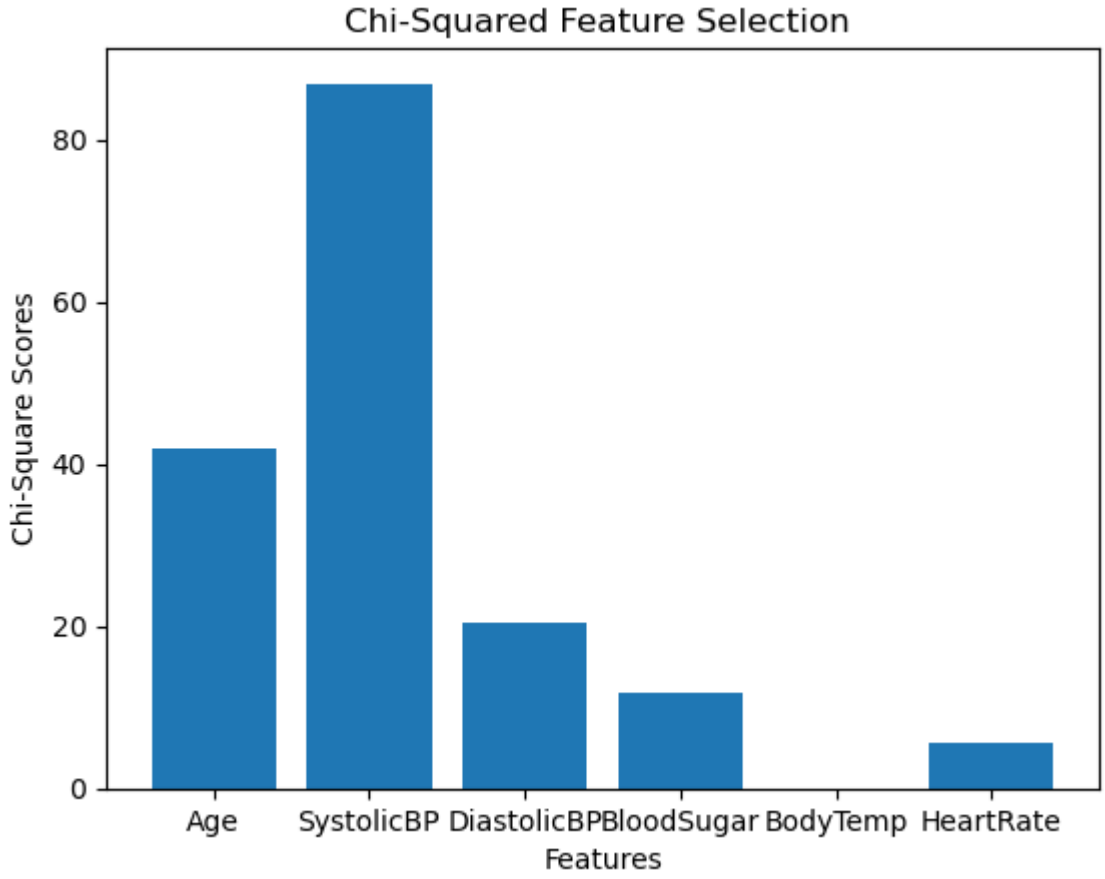

Figure 13: Chi-Squared Feature Selection Bar Plot

In figure 13, Attributes showing higher chi-square scores have more dependent information relating to the target variable being class. The attributes of age systolicBP, and distolicBP is largely influenced by class compared to the attributes with lower chi-scores.

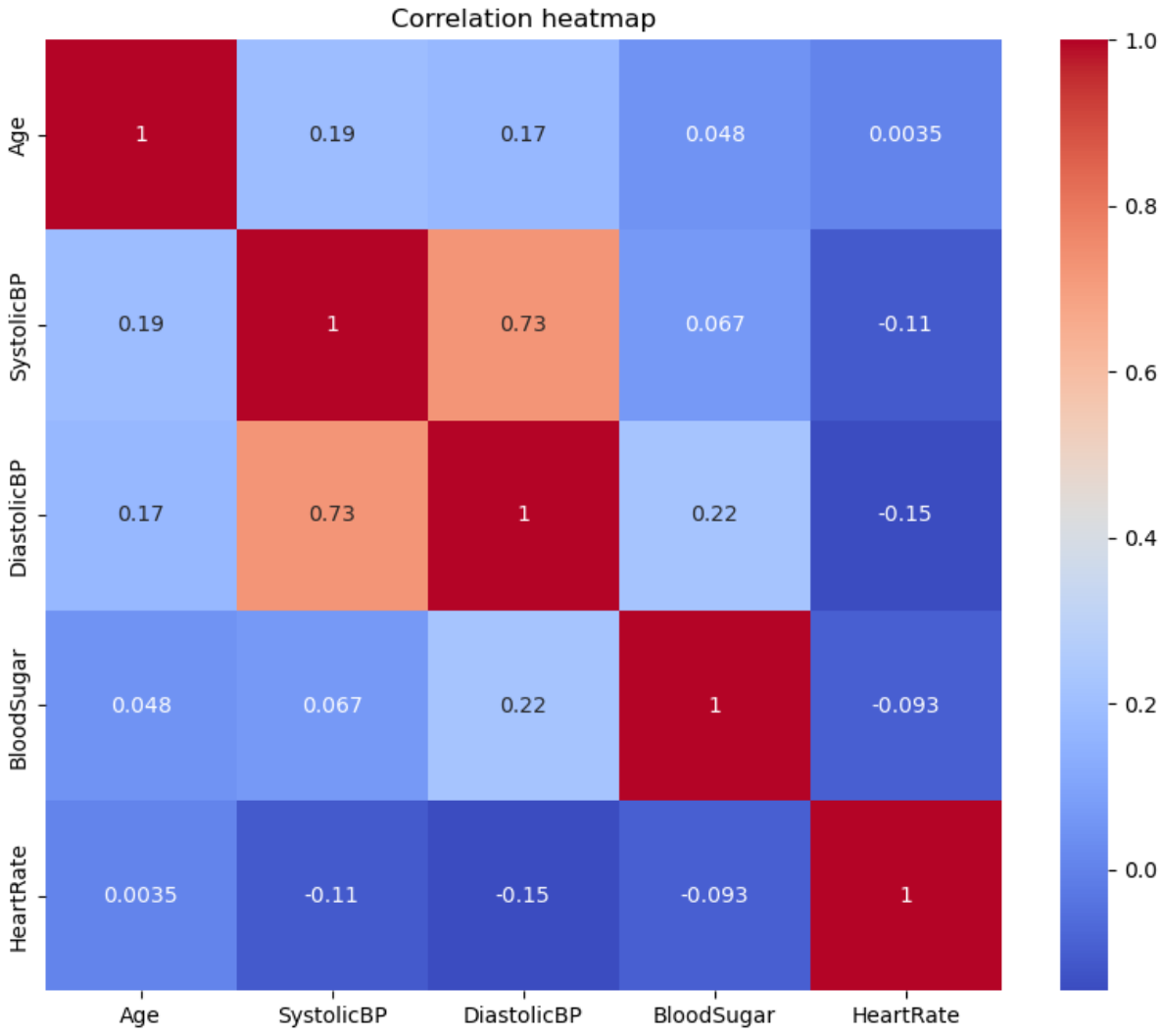

Figure 14: Correlation Heatmap

The correlation heatmap in figure 14 shows very few attributes correlation to each other. The biggest correlation is between systolicBP and diastolicBP due to their close relationship in both measuring Blood Pressure.

2.Data Classification Models

Cleaning the dataset involved identifying and removing the outliers. In the dataset, there is no duplicate or missing data found. The method used to remove the outliers is the IQR method because of the forementioned justification. The body temperature column was also dropped.

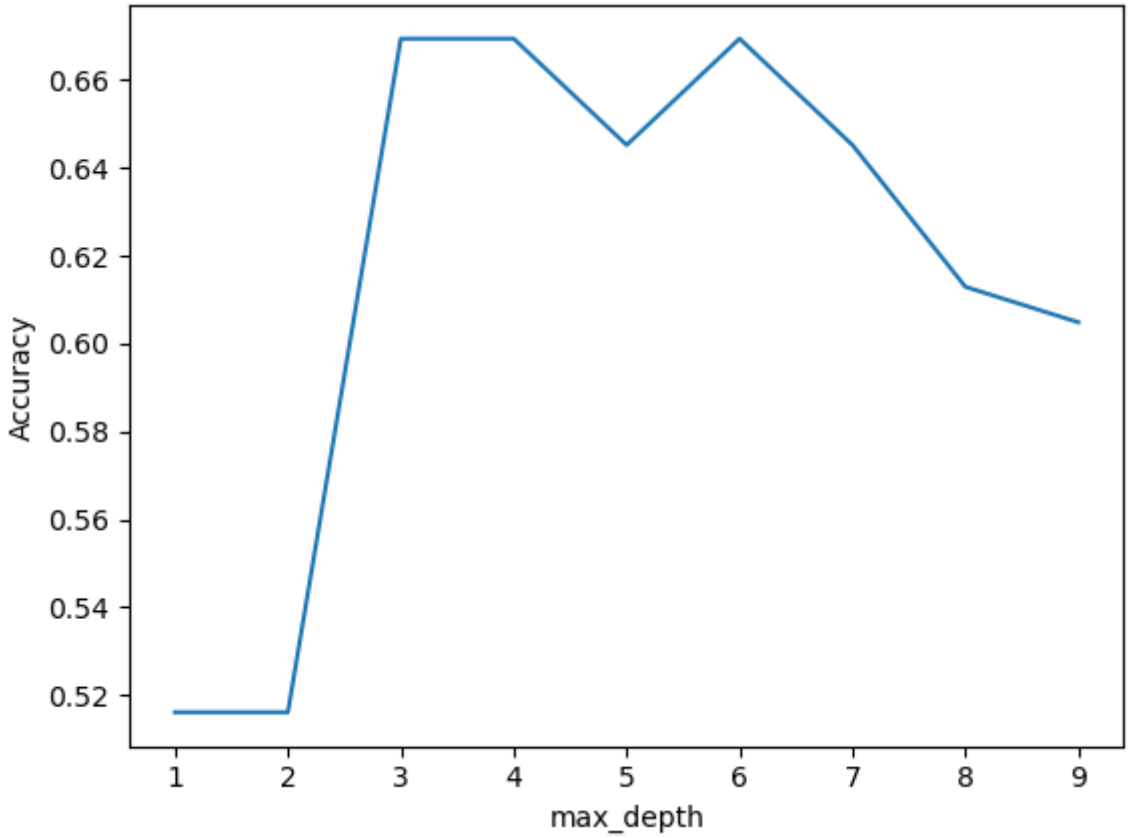

Figure 15: Decision Tree Accuracy (1 - 10)

If figure 15, the maximum depth of greater than 2 shows a higher accuracy and proceeds to fall when maximum depth goes beyond 6. This graph displays decision tree accuracy for when the maximum depth is between 1 and 10.

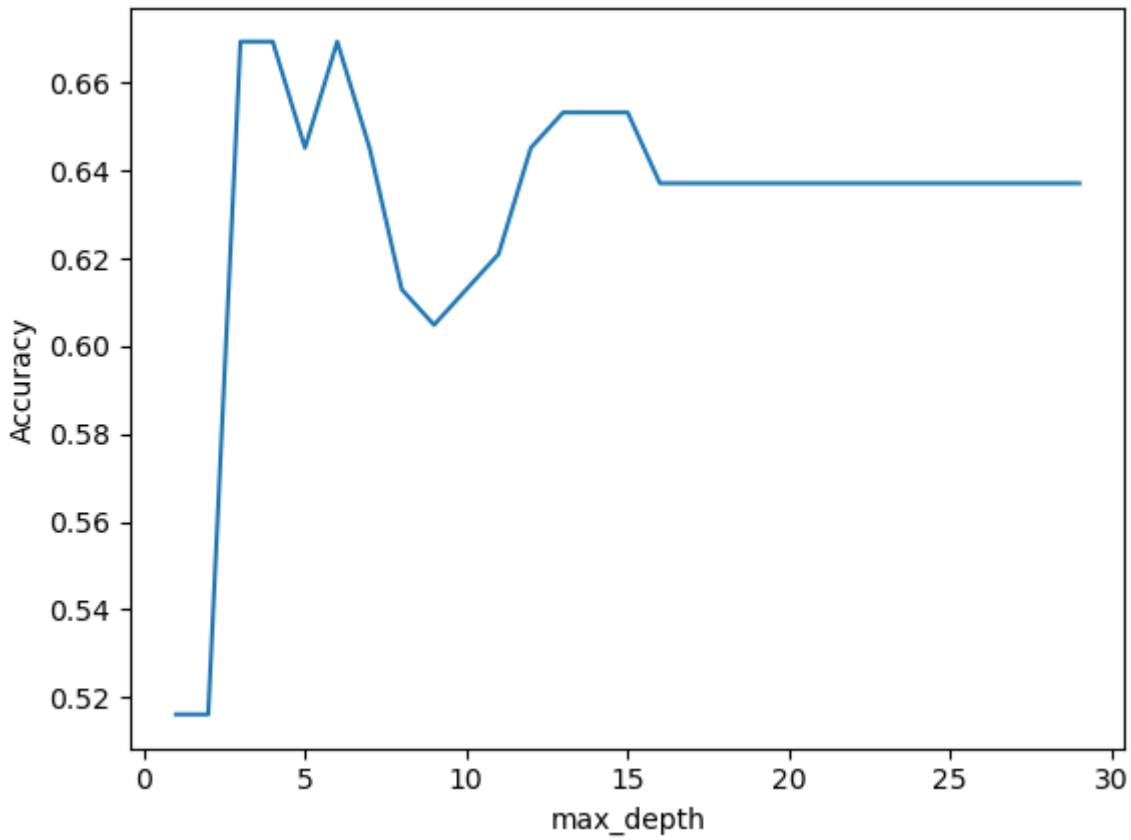

Figure 16: Decision Tree Accuracy (1 - 30)

Figure 16 is for when decision tree accuracy has a maximum depth range of 1 and 30. The accuracy increases again approximately between 6 and 14 and proceeds to stabilize beyond 15 although the highest accuracy still lies at around a maximum depth of 3. This graph will be more representative of due to the tree having a higher maximum depth making it more accurate, more complex, and less prone to overfitting.

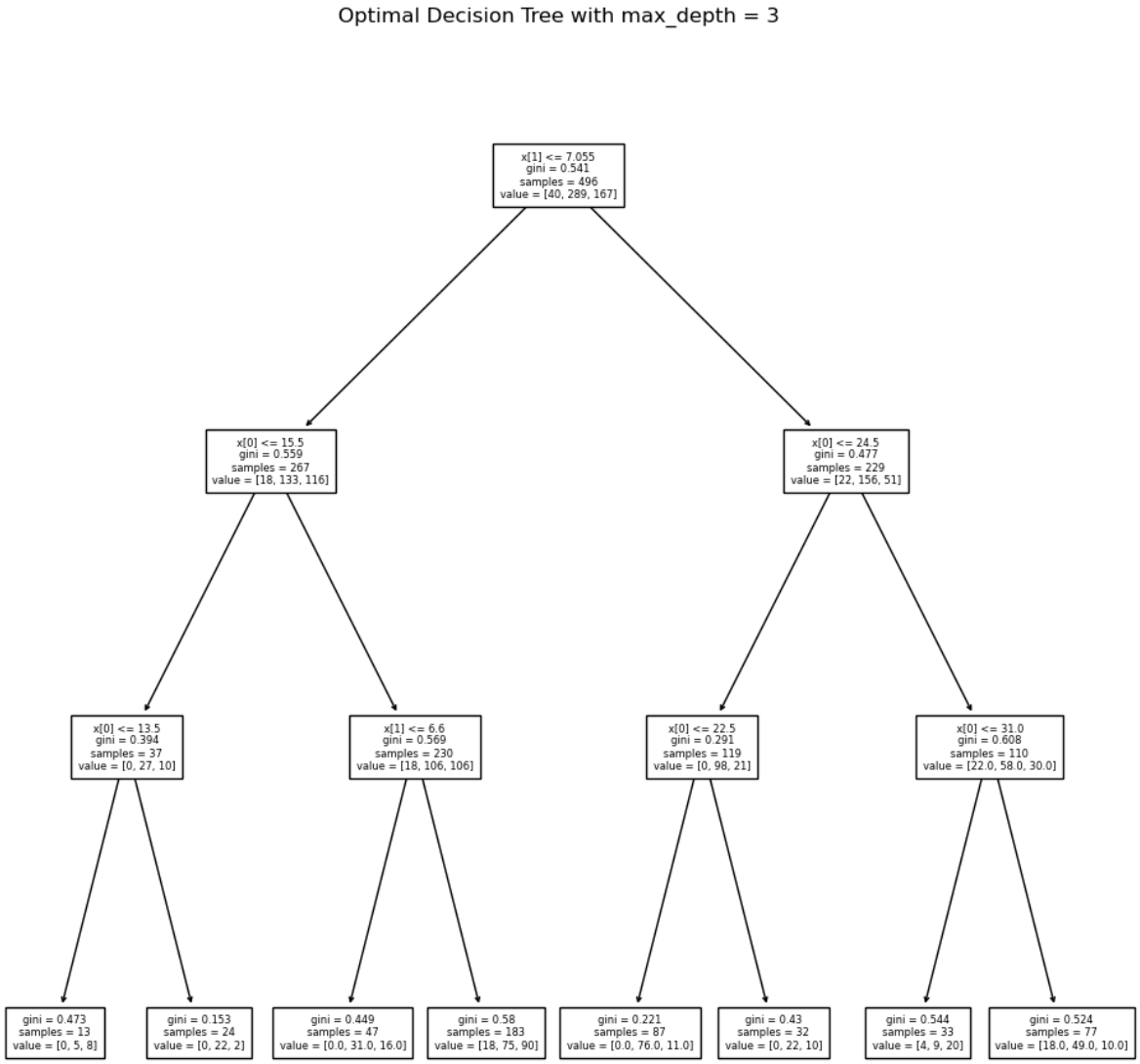

Figure 17: Decision Tree Based on Highest Accuracy

The tree in figure 17 is based on the highest accuracy seen in the line graphs. The split is relative to the features ‘Age’ and ‘Blood Sugar’. This split will show at the bottom the ‘risk level’ classification of each node.

3. Results and Discussion

This analysis goes into the dataset of maternal health in developing country settings, particularly in rural areas. The data was fed from wearable IoT devices to identify and classify maternal health risks in in pregnant women in Bangladesh.

The dataset used consists of 1014 data points across seven features related to health and risk levels. Through statistical analysis, it showed notable trends and outliers in certain biometric measures like Age, Blood Sugar, and Heart Rate, which were addressed using the IQR method to ensure data quality. These were cross referenced to ensure that the information represented medical accuracy and the human body's capabilities.

The distribution of risk levels showed that the majority of women classified as low or medium risk, indicating the minority is in the high risk group. Visualizations highlighted potential correlations between features and risk, such as the correlation between age and risk levels.

Data cleaning involved not only outlier removal but also feature selection, excluding Body Temperature due to the body's tight regulation therefore resulting in numbers being the same. The decision tree model focused on Age and Blood Sugar for risk classification, finding an optimal balance of accuracy and complexity at a max_depth of 14-15 although the accuracy still remains the highest at max_depth of around 3.

References

American Congress of Obstetricians and Gynecologists. (2020). Preeclampsia and High Blood Pressure During Pregnancy. https://www.acog.org/topics/hypertension-and-preeclampsia-in-pregnancy

American Heart Association. (2023). High Blood Pressure. https://www.heart.org/en/health-topics/high-blood-pressure

Diabète Québec. (n.d.). Diabetes in Pregnancy. https://guidelines.diabetes.ca/cpg/chapter36

Heart Foundation New Zealand. (n.d.). Managing High Blood Pressure. https://www.heartfoundation.org.nz/wellbeing/managing-risk/managing-high-blood-pressure
Mayo Clinic. (2023). Diabetes Diagnosis. https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/diagnosis/dxc-20371445

MedlinePlus. (2021). High Blood Pressure. https://medlineplus.gov/ency/anatomyvideos/000072.htm

National Institute of Diabetes and Digestive and Kidney Diseases. (2022). Diabetes. https://www.niddk.nih.gov/health-information/diabetes

Statista. (2023). New Zealand: Median age of mothers at childbirth from 1990 to 2023. https://www.statista.com/statistics/1064536/new-zealand-fertility-rate-by-age-group/