Maternal Health Risk Analysis

Data Description

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

The dataset is made up of 1,013 observations collected from different hospitals, community clinics, and maternal health care in the rural areas of Bangladesh through the IoT-based risk monitoring system. The dataset aims to investigate maternal health risks by analyzing the six variables that are significant for maternal mortality. The IoT system used to collect the data provides real-time and reliable data, increasing the quality of data and decreasing human error. The dataset provides valuable data that is helpful to understand maternal health issues in a limited resources environment.

For our analysis, we used the following six variables from the original dataset:

1. **Age:**
   Any age in years when a woman during pregnant. It is a numerical variable ranging from 10 to 70 in the dataset.
2. **SystolicBP:**
   Upper value of Blood Pressure in mmHg, a significant attribute during pregnancy. This is a numerical variable ranging from 70–160 mmHg in the dataset.
3. **Diastolic BP:**
   Lower value of Blood Pressure in mmHg. This variable measures arterial pressure during the heart's resting phase. This is ranging from 49-100 mmHg in the dataset.
4. **BS (Blood Sugar):**
   Level of blood sugar during pregnancy. A high blood sugar level is a risk factor for diabetes.
5. **BodyTemp**:
   It is a body temperature ranging from 98.0–103.0°F in the dataset. It is normally stable, but deviation can cause serious infection.
6. **Heart Rate:**
   It is the heartbeat per minute. This is a numerical variable ranging from 7-90 bpm in the dataset.

Using these six variables, we conducted a regression analysis to find out how these variables affect the risk level. RiskLevel is a categorical variable that represents maternal health risks. 1 represents a high risk, 2 represents a low risk, and 3 represents a mid risk.

Hypotheses:

① Hypothesis: Blood Sugar (BS), Systolic Blood Pressure (SystolicBP), and Body Temperature are the most significant predictors of maternal health risks.

- BS: Blood Sugar is expected to have a strong positive relationship with maternal health risks.
- SystolicBP: Since high blood pressure is one of the main factors of complications during pregnancy, SystolicBP is expected to be an important predictor of maternal health risks.
- Body Temperature (BodyTemp): BodyTemp may not have as strong an influence as BS and SystolicBP have on maternal health risks, but it is still expected to be a significant predictor of maternal health risks because high body temperature can lead to infection.

② Hypothesis: Diastolic BP and Heart Rate are less significant predictors of maternal health risks compared to BS, SystolicBP, and BodyTemp.

- DiastolicBP: It is expected to have less influence on maternal health risks than SystolicBP. It is generally known that SystolicBP is a more direct indicator of any complications related to high blood pressure during pregnancy.
- Heart Rate: Heart Rate is expected to have the least predictive power among the included variables. This is because heart rate reflects general cardiovascular activities rather than maternal health risks.

Methodologies: 1. Linear Regression Analysis

To test the two hypotheses, we conducted a **linear regression analysis** with the following 7 steps.

Step 1. Data Preparation

We loaded the dataset into R using the read.csv() function. RiskLevel is a categorical variable that has three categories. To facilitate our analysis, we transformed this categorical variable into a numerical variable(risk_numeric) using as.factor() and as.numeric() function. Then, we created plots to analyze the relationship between risk_numeric and each independent variable.

Step 2. Model Specification

We developed a linear regression model to predict maternal health risks using the selected six variables. The following function was used to create the linear regression model: lm_model1 <- lm(risk_numeric ~ Age + BS + SystolicBP + DiastolicBP + BodyTemp + HeartRate, data = maternal_health_data)

summary(lm_model1)

We examined the p-values and coefficients of BS, Systolic BP, and Body Temperature to test the first hypothesis. The predictors were considered statistically significant if their p-values were less than 0.05.

We also investigated the p-values and coefficients of Diastolic BP and HeartRate to test the second hypothesis. Their higher p-values and coefficients would indicate their lower statistical significance, while smaller values indicate less predictive power compared to BS, SystolicBP, and BodyTemp.

Step 3. Model Diagnostics

We created diagnostic plots using this function:

```
par(mfrow = c(2, 2))

plot(lm_model1)
```

We obtained a Residuals vs. Fitted Plot and identified the linearity and constant variance (homoscedasticity) of residuals. We also obtained a Normal Q-Q Plot and found out whether residuals followed a normal distribution. By analyzing these plots, we could guarantee the reliability of the regression model and verify that the model's assumptions were not violated.

Step 4: Data Splitting

In order to evaluate how well the regression model generalized, we split the dataset into a training dataset (80%) and a test dataset (20%) using the following codes:

```
set.seed(123)

train.rows <- sample(nrow(maternal_health_data), round(0.8 * nrow(maternal_health_data)))

train <- maternal_health_data[train.rows, ]

test <- maternal_health_data[-train.rows, ]
```

The training set was used to fit the regression model, and the testing set was made to evaluate the out-of-sample performance.

Step 5: Model Selection

In order to identify the most significant predictors, we used the regsubsets() function to select the best subset.

```
regfit.full <- regsubsets(risk_numeric ~ ., data = train)

summary(regfit.full)
```

We used the evaluation metrics such as Adjusted R-squared, BIC, and CP to compare subsets.

Step 6: Final Model Construction

Based on these analyses, we constructed the final model using the following codes to evaluate the relative importance of BS, SystolicBP, and BodyTemp (Hypothesis 1) and compare them to DiastolicBP and Heart Rate (Hypothesis 2).

```
final_model <- lm(risk_numeric ~ BS + SystolicBP + DiastolicBP + HeartRate + BodyTemp, data = train)
summary(final_model)
```

Step 7: Performance Evaluation

The performance of the final model was evaluated on both the training and testing datasets to assess its in-sample and out-of-sample performance. We used the following codes to evaluate the in-sample and out-of-sample performance.

1. In-Sample Performance: mean((final_model$fitted.values - train$risk_numeric)^2)
2. Out-of-Sample Performance: test_predictions <- predict(final_model, newdata = test) mean((test_predictions - test$risk_numeric)^2)

A small difference between the in-sample and out-of-sample MSE values indicates good generalization, supporting Hypothesis 2.

Results and Conclusion: Linear Regression Model

Our linear regression analysis evidently showed the statistical significance of each predictor and validated the two hypotheses.

① Results of linear regression analysis

We tested the first hypothesis with the results. **Blood Sugar** has turned out to be the most significant variable, showing small p-value (<2e-16) and a negative coefficient (-0.104844). These values indicate effective management of blood sugar can reduce maternal health risks. The low p-value (0.045064) of **SystolicBP** confirms it is a statistically important variable. Its positive coefficient (0.004240) shows higher SystolicBP increases maternal health risks. The p-value (0.045641) of **BodyTemp** is slightly lower than 0.05, thus it is a statistically significant variable. Its negative coefficient (-0.035960) indicates it can play a secondary role when predicting maternal health risks. Thus, these results prove that **Hypothesis 1 is true**.

We also tested the second hypothesis using the results. **DiastolicBP** demonstrated statistical significance (0.000149) but had a smaller regression coefficient (-0.010583) compared to Blood Sugar and SystolicBP. This result confirms that DiastolicBP has a moderate contribution to the model. The p-value of **HeartRate** (0.086684) indicates it is not statistically significant. The relatively high

p-value reflects its minimal role in predicting maternal health risks. Thus, the results show that **Hypothesis 2 is partially true** because it identified the minimal role of Heart Rate but underestimated the significance of Diastolic BP as a moderate predictor.

② Model Performance

The performance of the final model can also test the hypotheses. The model's in-sample Mean Squared Error (MSE) was 0.4408, and its out-of-sample MSE was 0.4613. The small difference between these two values proves that the model generalizes well to new data without overfitting. This result supports that BS, Systolic BP, and BodyTemp are significant predictors.

DiastolicBP's contribution in the model is confirmed by its small coefficient (-0.010583) and p-value(0.000149). This result indicates that DiastolicBP has a moderate, complementary role in predicting maternal health risks, improving the model's explanatory power when combined with primary predictors. On the other hand, HeartRate has a relatively high p-value(0.086684) and a small coefficient(-0.005256) in the final model. This indicates HeartRate has a much smaller impact on maternal health risks than other variables.

③ Unexpected Findings

Some unexpected results emerged. Although we first hypothesized that DiastolicBP would have limited significance, it showed moderate predictive power. This could be due to its role in systemic conditions such as hypertension which is known to influence maternal health. Conversely, HeartRate was not statistically significant. This may be attributed to its variability caused by temporary factors such as stress or physical activity, which might reduce its reliability as a consistent predictor of maternal health risks.

④ Conclusion

The linear regression analysis supports Hypothesis 1. BS is turned out to be the most significant predictor with a low p-value(<2e-16) and the largest coefficients (-0.104844). We also found that SystolicBP is a statistically significant variable(p=0.045064). Its positive coefficient (0.004240) shows that its increase leads to an increase in maternal health risks. BodyTemp has a relatively low p-value (0.045641), indicating its role as a marginally significant variable of maternal health risks. Thus, as predicted by Hypothesis 1, BS, SystolicBP, and BodyTemp are confirmed as the most significant predictors

Hypothesis 2 is partially validated. The high p-value of HeartRate indicates it is not a statistically significant variable and has limited predictive power. However, DiastolicBP was highly significant (p =0.000149) and demonstrated a moderate regression coefficient (-0.010583). This indicates although DiastolicBP is not a primary predictor it provides significant support in combination with key variables like BS and SystolicBP. Thus, while Hypothesis 2 assumes both DiastolicBP and HeartRate

have minimal significance, the analysis reveals that DiastolicBP plays a more substantial role than expected, whereas Heart Rate confirms its limited impact.

Methodologies: 2. Decision Tree

To test the two hypotheses, we also created a **decision tree model** with the following steps.

Step 1. Loading the Dataset and Data Splitting

We loaded the dataset into R using the read.csv() function. RiskLevel is a categorical variable with three categories: "low risk," "mid risk," and "high risk." To ensure correct handling during analysis, we transformed this variable into a factor using the as.factor() function. We used the summary() and names() functions to examine the structure of the dataset and confirmed the levels of RiskLevel using the levels() function.

To evaluate how well the decision tree model generalized to new data, we split the dataset into training and testing subsets.

- **Training Set (80%):** Used to build the decision tree model.
- **Testing Set (20%):** Reserved for performance evaluation.

We ensured reproducibility by setting a random seed using set.seed(380). The training set was created by sampling 80% of the data using the sample() function. The remaining 20% of the data was assigned to the testing set.

Step 2. Building the Decision Tree Model

We built the decision tree model using the tree library. The target variable, RiskLevel, was modeled as a function of all other variables in the dataset. The following function was used to create the model:
tree_model <- tree(RiskLevel ~ ., data = train)

To evaluate the tree, we summarized it using the summary() function. This provided details such as the variables used in the splits, the number of terminal nodes, residual mean deviance, and the misclassification error rate.

Step 3. Tree Visualization

To interpret the structure and decision rules of the tree, we visualized it using the plot() and text() functions. The plot showed the splitting variables and thresholds used for classification, helping to identify key predictors such as Blood Sugar (BS) and SystolicBP.

Step 4. Model Evaluation

To evaluate the performance of the decision tree on unseen data, we used the test dataset. Predictions were generated using the predict() function, specifying type = "class" to classify risk levels. The accuracy of the model was calculated by comparing the predicted and actual classifications using a confusion matrix (table()) and the formula:
accuracy <- mean(test_predictions == test$RiskLevel)

Step 5. Pruning the Decision Tree

To reduce overfitting, we simplified the tree by pruning it. Pruning removes less significant splits, resulting in a more interpretable and robust model. Using the prune.misclass() function, we created pruned versions of the tree with 3, 4, 5, and 6 terminal nodes. For each pruned tree, the following steps were performed:

Tree Summarization:
The pruned tree was summarized using the summary() function. This provided details about the variables used in the pruned model, the number of terminal nodes, residual deviance, and misclassification error rate.

Tree Visualization:
The structure of the pruned tree was visualized using the plot() and text() functions. This allowed us to interpret the decision-making process for each pruned version.

Predictions on Test Data:
The predict() function was used to generate predictions for the test dataset based on each pruned tree. The predictions classified patients into "low risk," "mid risk," and "high risk."

Performance Evaluation:
The test error rate for each pruned tree was calculated using the formula:
test_error_rate <- 1 - mean(pruned_predictions == test$RiskLevel)
This metric provided a quantitative assessment of how well the pruned tree performed compared to the unpruned tree.

Results and Conclusion: Decision Tree Model

① Results of Decision Tree Analysis
Upon a model performance-based reflection of all the decision trees, unpruned and pruned, we decided to test the first hypothesis using results from the 5-node decision tree. BS emerged as the most significant variable, consistently appearing as the root split with a threshold of **BS < 7.95**. This

confirms that BS plays a pivotal role in distinguishing high-risk patients. SystolicBP was the second most important variable, appearing in subsequent splits to classify low and mid-risk groups, validating its importance in maternal health risk prediction. BodyTemp also contributed to further splitting, supporting its role as a secondary predictor. These findings confirm that **Hypothesis 1 is true**: BS, SystolicBP, and BodyTemp are the most significant predictors of maternal health risks.

We also tested the second hypothesis using the 5-node tree. DiastolicBP did not appear in the decision tree, indicating it has minimal significance when combined with BS and SystolicBP. Similarly, HeartRate was excluded from the tree, reflecting its negligible predictive power. Thus, the results support **Hypothesis 2**, confirming that DiastolicBP and HeartRate have limited roles in predicting maternal health risks.

② Model Performance

The decision tree model's performance was evaluated on the test dataset:

- **Unpruned Tree:** Achieved a test error rate of **35.96%** with 8 terminal nodes.
- **3-Node Tree:** Simplest model using only BS and SystolicBP. High interpretability but high test error rate (39.9%), limiting reliability.
- **4-Node Pruned Tree:** Improved the test error rate to 38.92%. By adding Body Temperature as a predictor, this model balances simplicity with slightly better performance, making it moderately practical.
- **5-Node Pruned Tree:** Achieved a slightly higher test error rate of **36.45%**. However, this simpler model is more interpretable, retaining the key variables (BS, SystolicBP, and BodyTemp) while eliminating unnecessary complexity.
- **6-Node Pruned Tree:** Reduced the test error rate to 35.96%. Although it includes Age for additional depth, the marginal improvement in accuracy comes at the cost of increased complexity, limiting its practical value.

Why Choose the 5-Node Pruned Tree? Although the 5-node pruned tree has a marginally higher test error rate, its simplicity makes it better suited for real-world clinical use. With fewer terminal nodes:

- **Interpretability**: The tree provides clear and actionable thresholds for decision-making, making it easier for medical professionals to apply.
- **Avoiding Overfitting**: Pruning reduces the risk of overfitting, ensuring the tree generalizes better to unseen data.

③ Unexpected Findings

Some unexpected results emerged during the analysis:

- **DiastolicBP's Exclusion:** DiastolicBP did not appear in the tree, possibly be due to redundancy, as SystolicBP provides stronger predictive value for complications during pregnancy.
- **HeartRate's Negligible Role:** HeartRate was not included in the tree, likely because its variability (influenced by stress or activity) weakens its predictive power.
- **Mid-Risk Misclassification:** The model struggled with mid-risk classification, often misclassifying these patients as low risk. This highlights potential overlaps in features between the mid and low-risk groups.

④ Conclusion

The decision tree analysis supports **Hypothesis 1**, identifying BS, SystolicBP, and BodyTemp as the most significant predictors. BS consistently split the dataset at the root node, with SystolicBP and BodyTemp contributing to further refinement.

Hypothesis 2 is also validated. Both DiastolicBP and HeartRate were excluded from the tree, confirming their limited predictive value. However, the analysis revealed the challenge of accurately classifying mid-risk patients, suggesting the need for additional variables or advanced modeling techniques.

Final Model Recommendation

Based on the results of the linear regression model and the 5-node decision tree, we provide the following recommendations for assessing and predicting maternal health risks:

1. **Key Insights from the Models**:
    - **Linear Regression**:
        - Quantifies both the statistical significance and practical impact of predictors on maternal health risks.
        - Blood Sugar (BS) emerged as the most influential predictor, with a large negative coefficient, indicating that higher BS levels are strongly associated with reduced risk levels.
        - Systolic Blood Pressure (SystolicBP) and Body Temperature (BodyTemp) were statistically significant and had moderate practical impacts, while Diastolic Blood Pressure (DiastolicBP) had statistical significance but minimal practical effect. Heart Rate was not statistically significant.
        - Overall model performance (Adjusted R-squared = 0.2489) suggests the model explains 24.89% of the variance in maternal health risks, with a residual standard error of 0.6664.

- **Decision Tree**:
  - Provides a clear and interpretable framework for classifying patients into risk categories (low, mid, high).
  - Blood Sugar (BS) was the most impactful variable, forming the root split of the tree. SystolicBP and BodyTemp were also used in subsequent splits, refining risk classifications.
  - Diastolic Blood Pressure and Heart Rate were excluded, indicating they provide little practical benefit to classification accuracy.
  - The 5-node tree achieved a test error rate of 36.45%, making it reasonably reliable for real-world applications.

2. **Model Strength vs. Predictor Strength**:
   - The linear regression model excels at quantifying the strength of relationships between predictors and risk levels. For example, it highlights the statistical significance of BS and SystolicBP while showing that DiastolicBP, though statistically significant, has minimal practical impact.
   - The decision tree emphasizes practical impact by focusing on variables like BS, SystolicBP, and BodyTemp, which have the greatest influence on classification outcomes. Its simplicity and interpretability make it particularly useful for decision-making.
3. **Real-World Application**:
   - **Primary Recommendation**: The 5-node decision tree should be used as the primary tool for classifying maternal health risks into actionable categories (low, mid, high). Its interpretability and use of thresholds (e.g., BS $\geq$ 7.95) make it ideal for clinical settings.
   - **Supplementary Recommendation**: The linear regression model can complement the decision tree by providing insights into the relative importance and direction of each predictor. For example, BS's strong negative association with risk levels suggests prioritizing blood sugar management in interventions.
4. **Recommendations for Future Analysis**:
   - Explore alternative models, such as random forests or logistic regression, to improve accuracy while maintaining interpretability.
   - Consider validating the decision tree's thresholds and linear regression coefficients on a larger dataset to ensure robustness.
   - Investigate additional predictors or interactions (e.g., combined effects of BS and SystolicBP) to improve model performance.

Appendix

Exhibit 1: Linear Regression - Data Summary

```
Rows: 1014 Columns: 7
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): RiskLevel
dbl (6): Age, SystolicBP, DiastolicBP, BS, BodyTemp, HeartRate

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> summary(maternal_health_data)
      Age          SystolicBP     DiastolicBP          BS           BodyTemp        HeartRate     RiskLevel
 Min.   :10.00   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000   Min.   : 98.00   Min.   : 7.0   Length:1014
 1st Qu.:19.00   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.900   1st Qu.: 98.00   1st Qu.:70.0   Class :character
 Median :26.00   Median :120.0   Median : 80.00   Median : 7.500   Median : 98.00   Median :76.0   Mode  :character
 Mean   :29.87   Mean   :113.2   Mean   : 76.46   Mean   : 8.726   Mean   : 98.67   Mean   :74.3
 3rd Qu.:39.00   3rd Qu.:120.0   3rd Qu.: 90.00   3rd Qu.: 8.000   3rd Qu.: 98.00   3rd Qu.:80.0
 Max.   :70.00   Max.   :160.0   Max.   :100.00   Max.   :19.000   Max.   :103.00   Max.   :90.0
> names(maternal_health_data)
[1] "Age"         "SystolicBP"  "DiastolicBP" "BS"          "BodyTemp"    "HeartRate"   "RiskLevel"
```

```
> levels(as.factor(maternal_health_data$RiskLevel))
[1] "high risk" "low risk"  "mid risk"
```

Exhibit 2: Linear Regression - Scatter Plots of Variables and Risk Levels

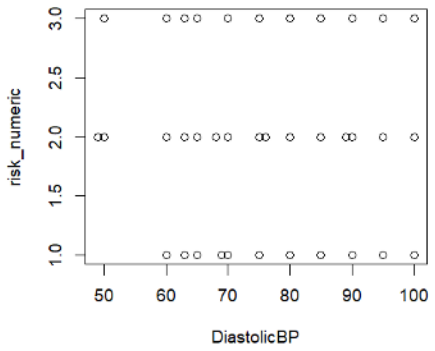

Figure 1: Relationship between Diastolic Blood Pressure and Risk Levels

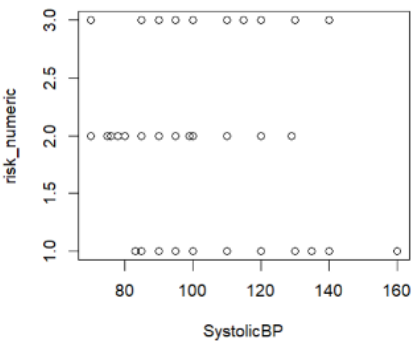

Figure 2: Relationship between Systolic Blood Pressure and Risk Levels

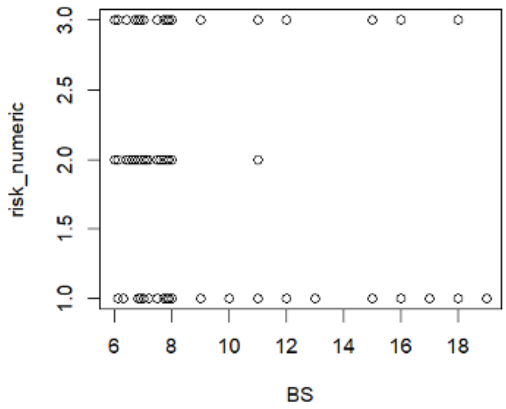

Figure 3: Relationship between Blood Sugar and Risk Levels

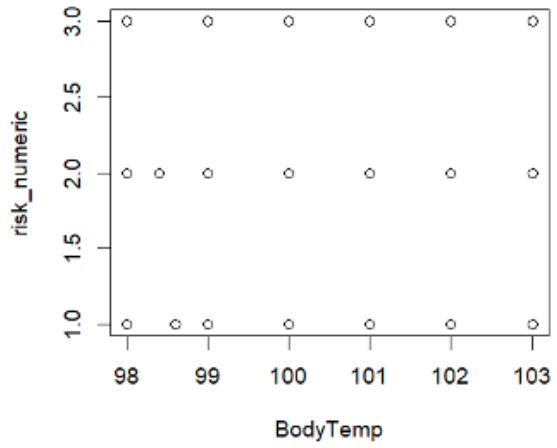

Figure 4: Relationship between Body Temperature and Risk Levels

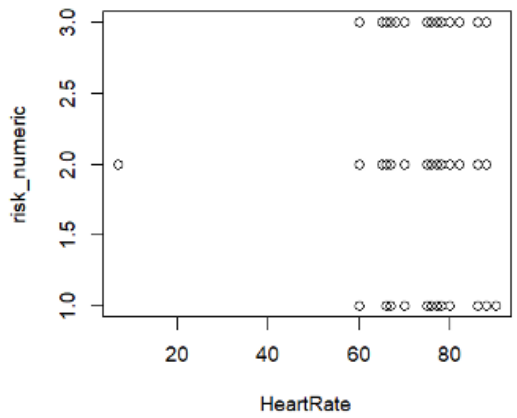

Figure 5: Relationship between Heart Rate and Risk Levels

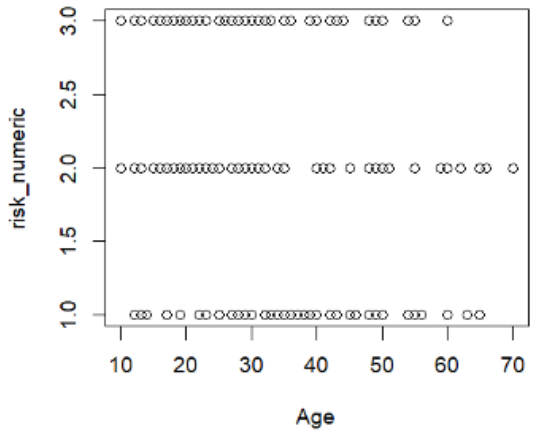

Figure 6: Relationship between Age and Risk Levels

Exhibit 3: Linear Regression - Summary

```
> summary(lm_model1)

Call:
lm(formula = risk_numeric ~ Age + BS + SystolicBP + DiastolicBP +
    BS + BodyTemp + HeartRate, data = maternal_health_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3809 -0.3634 -0.1773  0.6545  1.8593

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107843   1.649901   4.308 1.81e-05 ***
Age          0.001427   0.001875   0.761  0.44684
BS          -0.105818   0.007726 -13.696  < 2e-16 ***
SystolicBP   0.005448   0.001907   2.856  0.00437 **
DiastolicBP -0.012448   0.002501  -4.976 7.61e-07 ***
BodyTemp    -0.035191   0.016351  -2.152  0.03162 *
HeartRate   -0.004801   0.002660  -1.805  0.07139 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6689 on 1007 degrees of freedom
Multiple R-squared:  0.254,     Adjusted R-squared:  0.2496
F-statistic: 57.14 on 6 and 1007 DF,  p-value: < 2.2e-16
```

Exhibit 4: Linear Regression - Diagnostic Plots for Regression Assumptions

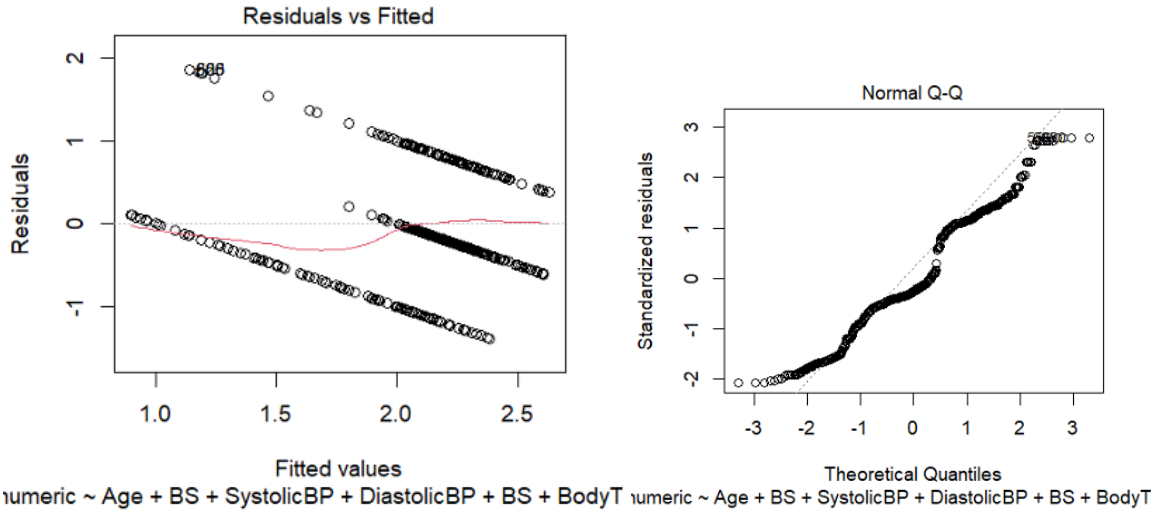

Exhibit 5: Linear Regression - Best Subset Selection

```
> summary(regfit.full)$adjr2
[1] 0.2334061 0.2395559 0.2442727 0.2470668 0.2488766 0.2480741
> plot(regfit.full, scale = "r2")
> summary(regfit.full)$bic
[1] -203.1675 -204.0047 -203.3568 -200.6680 -196.9284 -190.3722
> summary(regfit.full)$cp
[1] 17.781383 12.153397  8.079814  6.079783  5.140838  7.000000
```

```
> regfit.full1 <- regsubsets(risk_numeric ~ ., data = train, nvmax = ncol(train) - 1)
> summary(regfit.full1)
Subset selection object
Call: regsubsets.formula(risk_numeric ~ ., data = train, nvmax = ncol(train) -
    1)
6 Variables  (and intercept)
           Forced in Forced out
Age            FALSE      FALSE
SystolicBP     FALSE      FALSE
DiastolicBP    FALSE      FALSE
BS             FALSE      FALSE
BodyTemp       FALSE      FALSE
HeartRate      FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         Age SystolicBP DiastolicBP BS  BodyTemp HeartRate
1  ( 1 ) " " " "        " "         "*" " "      " "
2  ( 1 ) " " "*"        " "         "*" " "      " "
3  ( 1 ) " " "*"        "*"         "*" "*"      " "
4  ( 1 ) " " "*"        "*"         "*" "*"      "*"
5  ( 1 ) " " "*"        "*"         "*" "*"      "*"
6  ( 1 ) "*" "*"        "*"         "*" "*"      "*"
```

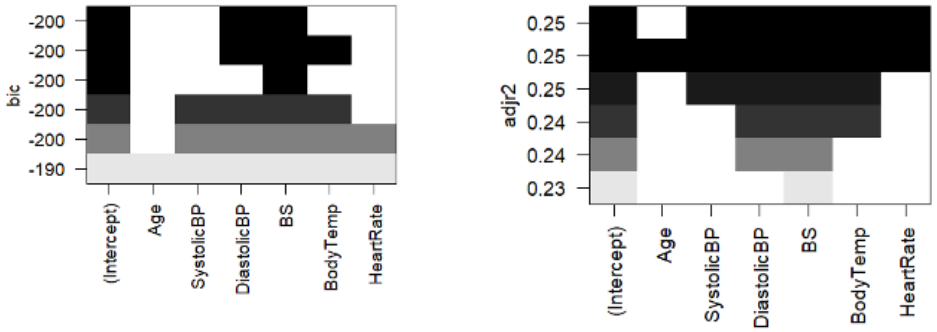

Exhibit 6: Linear Regression - Final Model Summary

```
> #Final Model Selection
> final_model <- lm(risk_numeric ~ BS + SystolicBP + DiastolicBP + HeartRate + BodyTemp, data = train)
> summary(final_model)

Call:
lm(formula = risk_numeric ~ BS + SystolicBP + DiastolicBP + HeartRate +
    BodyTemp, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3633 -0.3639 -0.1852  0.6725  1.8596

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.243365   1.813631   3.994  7.1e-05 ***
BS          -0.104844   0.008190 -12.802  < 2e-16 ***
SystolicBP   0.004240   0.002112   2.007 0.045064 *
DiastolicBP -0.010583   0.002777  -3.811 0.000149 ***
HeartRate   -0.005256   0.003064  -1.715 0.086684 .
BodyTemp    -0.035960   0.017964  -2.002 0.045641 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6664 on 805 degrees of freedom
Multiple R-squared:  0.2535,    Adjusted R-squared:  0.2489
F-statistic: 54.68 on 5 and 805 DF,  p-value: < 2.2e-16
```

Exhibit 7: Decision Tree - Dataset Summary

```
> summary(maternal_health_data)
      Age          SystolicBP     DiastolicBP           BS            BodyTemp        HeartRate     RiskLevel
 Min.   :10.00   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000   Min.   : 98.00   Min.   : 7.0   Length:1014
 1st Qu.:19.00   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.900   1st Qu.: 98.00   1st Qu.:70.0   Class :character
 Median :26.00   Median :120.0   Median : 80.00   Median : 7.500   Median : 98.00   Median :76.0   Mode  :character
 Mean   :29.87   Mean   :113.2   Mean   : 76.46   Mean   : 8.726   Mean   : 98.67   Mean   :74.3
 3rd Qu.:39.00   3rd Qu.:120.0   3rd Qu.: 90.00   3rd Qu.: 8.000   3rd Qu.: 98.00   3rd Qu.:80.0
 Max.   :70.00   Max.   :160.0   Max.   :100.00   Max.   :19.000   Max.   :103.00   Max.   :90.0
> names(maternal_health_data)
[1] "Age"         "SystolicBP"  "DiastolicBP" "BS"          "BodyTemp"    "HeartRate"   "RiskLevel"
```

Figure 6: Summary of the Maternal Health Dataset

Exhibit 8: Decision Tree - Data Splitting Dimensions

```
> dim(train) # Should be ~80% of rows
[1] 811   7
> dim(test)  # Should be ~20% of rows
[1] 203   7
```

Figure 8: Dimensions of Training and Testing Datasets

Exhibit 9: Unpruned Decision Tree

```
> summary(tree_model)

Classification tree:
tree(formula = RiskLevel ~ ., data = train)
Variables actually used in tree construction:
[1] "BS"         "SystolicBP" "BodyTemp"   "Age"
Number of terminal nodes:  8
Residual mean deviance:  1.231 = 988.7 / 803
Misclassification error rate: 0.2922 = 237 / 811
```

Figure 9: Summary of the Unpruned Decision Tree.

```
> print(conf_matrix)
           test_predictions
            high risk low risk mid risk
  high risk        39        1        6
  low risk          2       72       11
  mid risk         13       40       19
> # Calculate the accuracy
> accuracy <- mean(test_predictions == test$RiskLevel)
> print(paste("Accuracy:", round(accuracy * 100, 2), "%"))
[1] "Accuracy: 64.04 %"
```

Figure 10: Confusion Matrix and Accuracy of the Unpruned Tree

Exhibit 10: Pruned Decision Tree with 3 Terminal Nodes

```
> # Prune the tree with 3 terminal nodes
> pruned_tree_3 <- prune.misclass(tree_model, best = 3)
> summary(pruned_tree_3)

Classification tree:
snip.tree(tree = tree_model, nodes = c(3L, 5L, 4L))
Variables actually used in tree construction:
[1] "BS"         "SystolicBP"
Number of terminal nodes:  3
Residual mean deviance:  1.521 = 1229 / 808
Misclassification error rate: 0.3662 = 297 / 811
```

Figure 11: Summary of the Decision Tree with 3 Terminal Nodes.

```
> pruned_test_error_rate_3 <- 1 - mean(pruned_predictions_3 == test$RiskLevel)
> print(paste("Test Error Rate for Pruned Tree (3 Nodes):", round(pruned_test_error_rate_3 * 100, 2), "%"))
[1] "Test Error Rate for Pruned Tree (3 Nodes): 39.9 %"
```

Figure 12: Test Error Rate (39.9%) for Pruned Tree with 3 Nodes.

Exhibit 11: Summary of Pruned Tree with 4 Terminal Nodes

```
> # Prune the tree with 4 terminal nodes
> pruned_tree_4 <- prune.misclass(tree_model, best = 4)
> summary(pruned_tree_4)

Classification tree:
snip.tree(tree = tree_model, nodes = c(3L, 5L, 8L))
Variables actually used in tree construction:
[1] "BS"         "SystolicBP" "BodyTemp"
Number of terminal nodes:  4
Residual mean deviance:  1.44 = 1162 / 807
Misclassification error rate: 0.3292 = 267 / 811
```

Figure 13: Classification Tree with 4 Terminal Nodes

```
> # Predict on the test dataset with the pruned tree
> pruned_predictions_4 <- predict(pruned_tree_4, newdata = test, type = "class")
> pruned_test_error_rate_4 <- 1 - mean(pruned_predictions_4 == test$RiskLevel)
> print(paste("Test Error Rate for Pruned Tree (4 Nodes):", round(pruned_test_error_rate_4 * 100, 2), "%"))
[1] "Test Error Rate for Pruned Tree (4 Nodes): 38.92 %"
```

Figure 14: Test Error Rate for Pruned Tree with 4 Nodes

Exhibit 12: Summary of Pruned Tree with 5 Terminal Nodes

```
> # Prune the tree with 5 terminal nodes
> pruned_tree_5 <- prune.misclass(tree_model, best = 5)
> summary(pruned_tree_5)

Classification tree:
snip.tree(tree = tree_model, nodes = c(3L, 16L, 5L))
Variables actually used in tree construction:
[1] "BS"         "SystolicBP" "BodyTemp"
Number of terminal nodes:  5
Residual mean deviance:  1.364 = 1100 / 806
Misclassification error rate: 0.2947 = 239 / 811
```

Figure 14: Classification Tree with 5 Terminal Nodes

```
> pruned_predictions_5 <- predict(pruned_tree_5, newdata = test, type = "class")
> pruned_test_error_rate_5 <- 1 - mean(pruned_predictions_5 == test$RiskLevel)
> print(paste("Test Error Rate for Pruned Tree (5 Nodes):", round(pruned_test_error_rate_5 * 100, 2), "%"))
[1] "Test Error Rate for Pruned Tree (5 Nodes): 36.45 %"
```

Figure 15: Test Error Rate for Pruned Tree with 5 Nodes

Exhibit 13: Summary of Pruned Tree with 6 Terminal Nodes

```
> # Prune the tree with 6 terminal nodes
> pruned_tree_6 <- prune.misclass(tree_model, best = 6)
> summary(pruned_tree_6)

Classification tree:
snip.tree(tree = tree_model, nodes = c(3L, 16L))
Variables actually used in tree construction:
[1] "BS"         "SystolicBP" "BodyTemp"   "Age"
Number of terminal nodes:  6
Residual mean deviance:  1.343 = 1081 / 805
Misclassification error rate: 0.2922 = 237 / 811
```

Figure 16: Classification Tree with 6 Terminal Nodes

```
> pruned_predictions_6 <- predict(pruned_tree_6, newdata = test, type = "class")
> pruned_test_error_rate_6 <- 1 - mean(pruned_predictions_6 == test$RiskLevel)
> print(paste("Test Error Rate for Pruned Tree 65 Nodes):", round(pruned_test_error_rate_6 * 100, 2), "%"))
[1] "Test Error Rate for Pruned Tree 65 Nodes): 35.96 %"
```

Figure 17: Test Error Rate for Pruned Tree with 6 Nodes

Exhibit 14: Visualization of Decision Tree

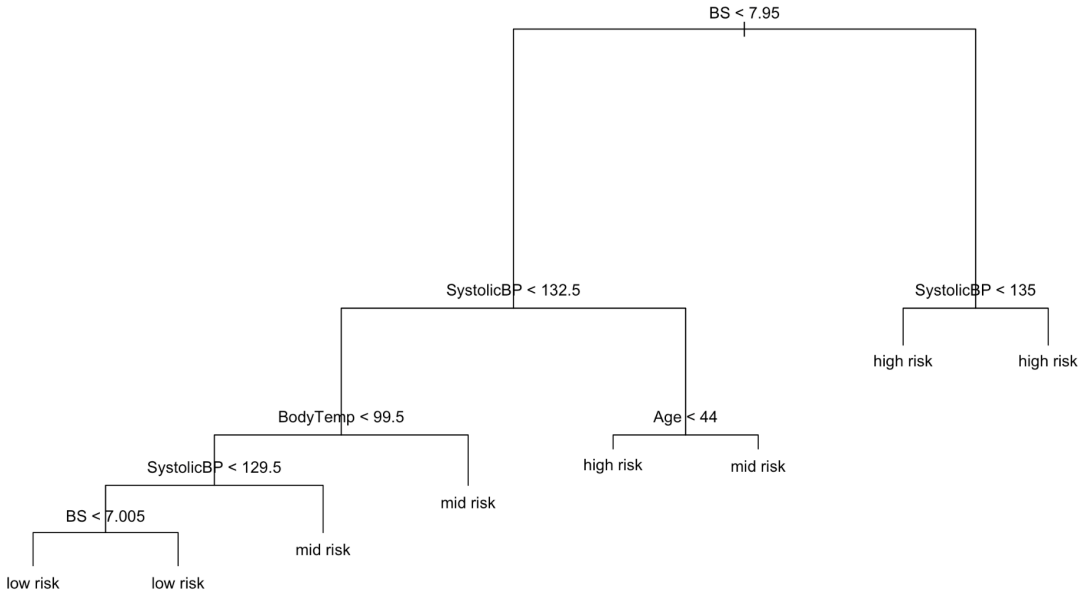

Figure 18: Visualization of the Unpruned Decision Tree

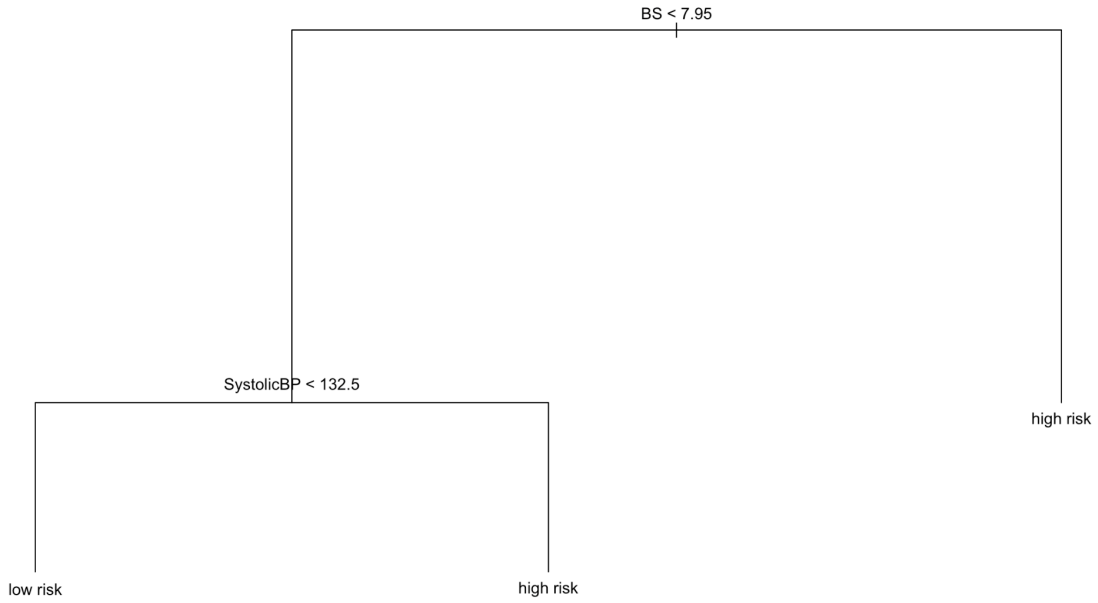

Figure 19: Visualization of the Pruned Decision Tree with 3 Terminal Nodes

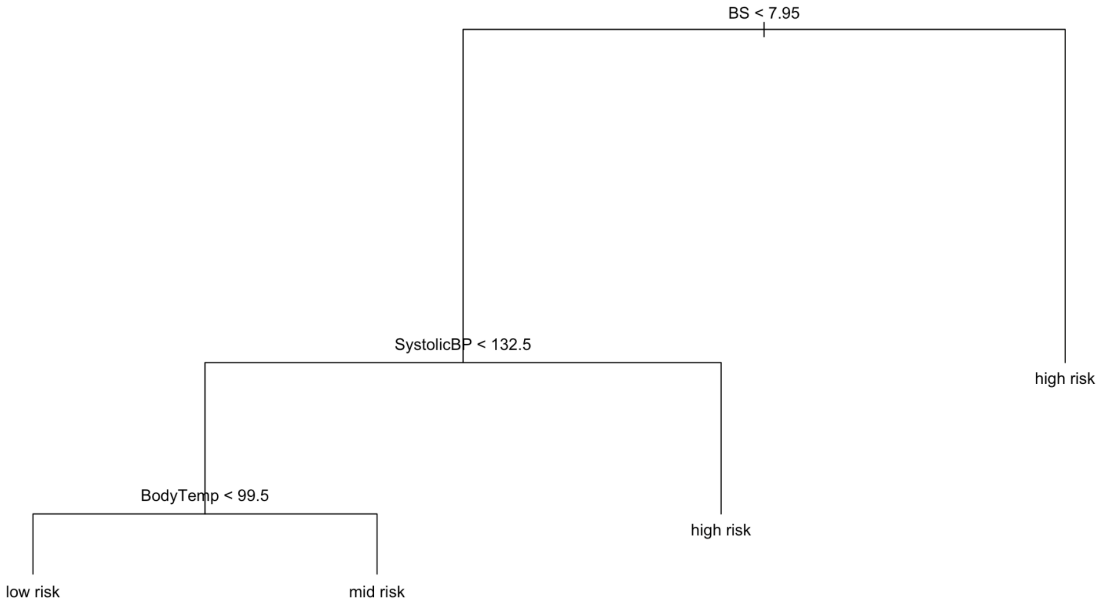

Figure 20: Visualization of the Pruned Decision Tree with 4 Terminal Nodes

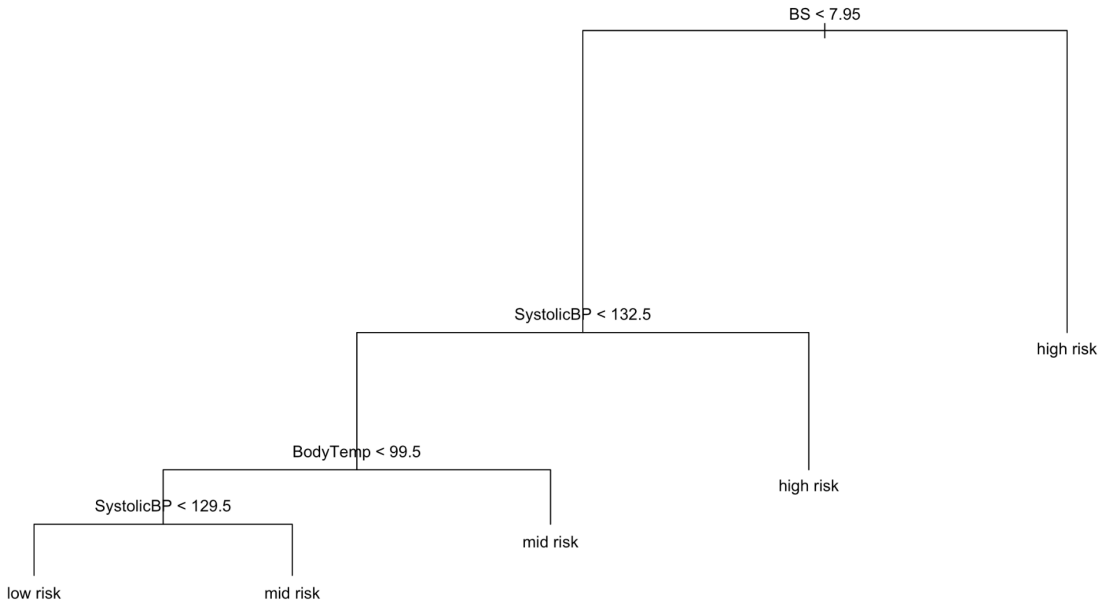

Figure 21: Visualization of the Pruned Decision Tree with 5 Terminal Nodes

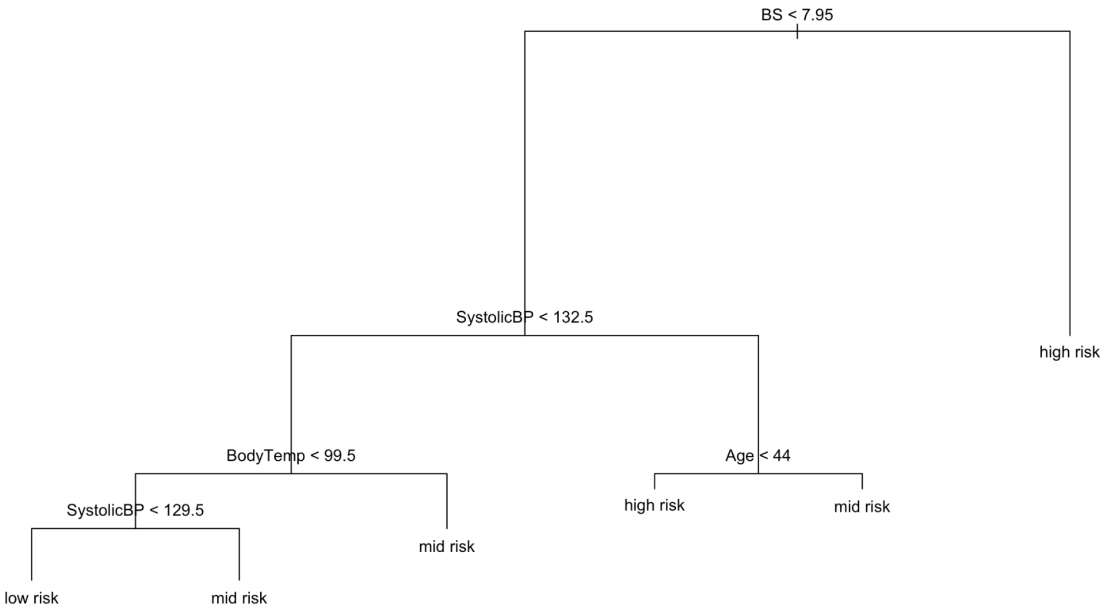

Figure 22: Visualization of the Pruned Decision Tree with 6 Terminal Nodes