Predicting Maternal Mortality from Clinical Data

Antoinette Nguyen

Spring Quarter 2025

1 Introduction

1.1 Background

Maternal mortality refers to the death of a woman caused by complications related to pregnancy or childbirth. Maternal mortality has steadily declined throughout the twentieth and twenty-first centuries yet it remains a major health crisis, especially in less developed regions of the world. According to the World Health Organization, about 260,000 women have died from maternal causes in 2023 alone [1]. One way to address this health crisis is to explore how clinical information about pregnant women can be leveraged to predict Maternal Mortality Risk.

1.2 Objective

The objective of this study is to develop and evaluate predictive models that replicate physician-assigned maternal health risk categories based on clinical measurements. Given that the target variable was derived from clinical judgment of vital signs, this analysis aims to assess how well the models can approximate decision-making to support automated maternal risk assessment. Additionally, the analysis aims to identify the most significant predictors of maternal mortality risk.

2 Dataset

The dataset was sourced from the UCI Machine Learning Repository, a web-accessible repository containing a wide range of available datasets. The dataset consists of clinical records on pregnant women from various hospitals and clinics in rural areas in Bangladesh from 2020. Clinical information of these pregnant women included their Age, Systolic Blood Pressure, Diastolic Blood Pressure, Heart Rate, and Body Temperature measurements. To reduce multicollinearity before model development, only Systolic Blood Pressure will be used, excluding Diastolic Blood Pressure.

3 Methodology

For this multiclass analysis, a Multinomial Logistic Regression (MLR) Model and a Random Forest (RF) Model will be used to predict the risk level category among pregnant women. I will train and evaluate both models with a 5-fold cross-validation with 3 repetitions to ensure robustness across the estimates. An ROC Curve with a one vs. rest approach will be used to compare the classification performance of both models.

4 Exploratory Data Analysis Findings

Table 1: Summary Statistics of Maternal Health Dataset

Variable	**Min**	**Median**	**Mean**	**Max**
Age	10.00	26.00	29.87	70.00
Systolic BP	70.00	120.00	113.20	160.00
Diastolic BP	49.00	80.00	76.46	100.00
Blood Sugar (BS)	6.00	7.50	8.73	19.00
Body Temperature	98.00	98.00	98.67	103.00
Heart Rate	7.00	76.00	74.30	90.00
Risk Level (Categorical)				
Low Risk: 406	Mid Risk: 336	High Risk: 272		

4.1 Response Variable

The distribution of maternal health risk levels in the dataset is noticeably imbalanced. The majority of observations fell into the low-risk category, followed by mid-risk, then high-risk. A Chi-squared test for goodness-of-fit was conducted to evaluate whether the observed class proportions differ significantly from a uniform distribution. At a significance level of $\alpha = 0.05$, the test revealed a significant difference in the class proportions indicated by the low p-value (1.69×10^{-6}). The imbalance of the class distributions may influence the performance of the classification models that will be explored in this paper. Consequently, it is pertinent to take precautions when interpreting the model's evaluation measurements.

4.2 Numerical Variables

The distribution of Heart Rate seemed approximately normal, centered around 75 beats per minute. Age appeared slightly skewed to the right with wide distribution. Most women fall within the 20 to 30-year age group. The distribution of Blood Sugar appeared heavily right-skewed, with the majority of values clustering between 7 and 8 mmol/L, though a few extreme values suggest potential outliers. Systolic Blood Pressure (SystolicBP) exhibited a multimodal distribution with several minor peaks, but most readings are concentrated around 120 mmHg. Similarly, Diastolic Blood Pressure (DiastolicBP) also showed a multimodal pattern, though the distribution was more evenly spread and centered around 80 mmHg. Body Temperature remained relatively stable for most observations, clustering tightly around 98°F, though a few higher values around 101°F indicate the presence of outliers.

The correlation matrix revealed several moderate relationships among the variables. Systolic Blood Pressure and Diastolic Blood Pressure exhibited a strong positive correlation ($r = 0.79$). Age showed moderate positive correlations with Systolic Blood Pressure, Diastolic Blood Pressure, and Blood Sugar, indicating that older women in the dataset tend to have higher blood pressure and blood sugar levels. Blood Sugar was moderately correlated with both types of blood pressure ($r = 0.42$), suggesting some co-occurrence of elevated glucose and hypertension. Body Temperature was negatively correlated with most other variables, especially with Age and Systolic Blood Pressure, though these relationships are relatively weak. Heart Rate showed very minimal correlation with the other features, indicating it may vary more independently in this population.

Next, I examined the relationship between the variables using a scatterplot with a Nadaraya-Watson estimated Kernel Regression overlaid to capture any non-linear patterns. Most of the relationships appeared ambiguous, except for the relationship between Diastolic Blood Pressure and Systolic Blood Pressure, which showed a moderately high positive association. Interestingly, the scatterplot of Age and Systolic Blood Pressure showed a downward-facing curved pattern in the relationship.

5 Modeling and Evaluation

5.1 Multinomial Logistic Regression

First, I fitted the data with a MLR in R. The low-risk level was used as the reference group for the model comparisons. The predictors selected in the model were Systolic Blood Pressure, Blood Sugar, Heart Rate, and Age. These variables were used in the final model of the MLR because they showed statistical significance when all variables, except for Diastolic Blood Pressure, were applied to the model.

The MLR revealed statistically significant associations with maternal risk levels for both sets of comparisons (Low to Mid and Low to High). The results showed that Blood Sugar, Systolic Blood Pressure, Age, and Heart Rate are significant predictors of the risk classification ($\alpha = 0.05$). Interestingly, Age appears to have a significant negative effect on a woman being predicted as high-risk in the Low to High comparison and as mid-risk in the Low to Mid comparison. This may be attributed to unique patterns in the dataset or the presence of potential confounding variables. One possible explanation is the correlation between Age and Income, where younger women may also have lower income, which in turn could influence maternal mortality risk.

Table 2: Multinomial Logistic Regression Results: Odds Ratios

Variable	**Mid Risk**			**High Risk**		
	OR	SE	p-value	OR	SE	p-value
Systolic BP	1.03***	0.006	<0.001	1.05***	0.008	<0.001
Blood Sugar (BS)	1.45***	0.080	<0.001	2.12***	0.086	<0.001
Age	0.98*	0.008	0.001	0.96***	0.013	0.001
Heart Rate	1.03*	0.012	<0.001	1.08***	0.017	<0.001

Notes: OR = Odds Ratio; SE = Standard Error; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

The class-level metrics showed varying degrees of performance across the risk levels: Sensitivity was moderately high for low-risk (78.51%) compared to the mid-risk level, which scored the lowest of the three (24.00%). Specificity also had varying degrees across the risk levels, ranging from 52.49% to 97.74%.

The McFadden pseudo R^2 value of 0.2181 indicates a moderate model fit, and the extremely low likelihood ratio test p-value (5.26×10^{-68}) confirms that the model significantly improves upon the null. Additionally, macro-averaged metrics: precision (0.604), recall (0.5557), and F1 score (0.5585) reflect moderate overall performance, driven largely by the model's success in classifying the high-risk group.

5.2 Random Forest Classification

Next, I fitted a RF classifier to the data and trained the model with the same cross-validation method: 5-fold cross-validation with 3 repeats. Using the RF, the model revealed substantial improvement in prediction accuracy compared to the MLR. The RF achieved an overall accuracy of 87.75%.

Class-level metrics showed strong performance of the model's distinguishing abilities across all risk levels: Sensitivity was high across the risk categories, with the lowest for mid-risk (84.00%) and the highest for the high risk (91.36%). Specificity was also high across the risk categories, exceeding 90% for each category. The model performed best for predicting the high-risk category, achieving the highest in Balanced Accuracy (94.77%) compared to the other categories. The macro-averaged precision, recall, and F1 scores all exceeded 87%, showing strong predictive performance across the classes.

Variable importance was assessed using the mean decrease in accuracy metric from the RF model. This method quantifies how much the model's average accuracy drops when the values of a given predictor are randomized, effectively breaking its association with the outcome. The results reveal that Blood Sugar was the most important predictor, followed by Systolic Blood Pressure, then Body Temperature. These variables contributed most significantly to the model's predictive performance and should be considered key indicators in assessing maternal health risk.

5.3 Model Comparison

Table 3: Model Performance Comparison

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Mean AUC
MLR	0.566	0.604	0.556	0.558	0.773
RF	0.877	0.884	0.879	0.881	0.950

The Random Forest (RF) model significantly outperformed the Multinomial Logistic Regression (MLR) model across all evaluated performance metrics. RF achieved an overall accuracy of 87.75%, in contrast to 56.62% for MLR. It also showed large gains in macro-averaged precision, recall, F1 score, and mean AUC, highlighting its strong classification performance and discriminative power across the risk categories. A comparison of the confusion matrices in Figures 6 and 7 further illustrates this improvement. The RF matrix displays a much more pronounced diagonal pattern, indicating more correct predictions, whereas the MLR matrix reveals greater deviance from the diagonal pattern, reflecting the misclassifications.

The MLR showed moderate predictive performance, with strong discrimination for the high-risk category, but performed poorly at classifying mid-risk. Future model improvements could focus on addressing the misclassifications of the mid-risk by implementing other predictive variables. Despite the differences in model performance, both models performed better than random guessing.

6 Conclusion

Both models successfully classified maternal risk levels, though with differing levels of accuracy. The RF, in particular, demonstrated strong predictive performance. While both models exhibited varying degrees of precision across the risk levels, they both suggest that Blood Sugar, Systolic Blood Pressure, and Heart Rate emerge as the most influential predictors of risk category. The MLR found that age seems to have a negative effect on a pregnant woman's predicted risk level, which may be attributed to possible confounding variables or the unique pattern of the dataset.

Since risk-level assignments were based on physician judgment using clinical measurements, it is expected that these variables will demonstrate strong predictive power. However, analyzing their relative importance provided valuable insights into how specific clinical factors influence maternal mortality risk. These findings can assist healthcare professionals, particularly in regions with limited access to care, by helping prioritize interventions targeting the most influential clinical indicators.

6.1 Limitations

The dataset was limited to clinical data of pregnant women residing in rural areas of Bangladesh. As a result, the findings of this analysis may not be generalizable to women in more urbanized or high-income regions, where healthcare infrastructure and access to medical services are typically more advanced. Disparities in mortality ratios across the globe are often driven by differences in healthcare accessibility, quality of care, and socioeconomic conditions.

6.2 Future Considerations

To improve the performance of the MLR model, it would be valuable to incorporate additional variables related to socioeconomic status, healthcare access, and individual health history. For instance, income level, educational attainment, and proximity to healthcare facilities may influence maternal risk outcomes through their impact on health literacy, nutrition, and timely care, respectively.

Another way to improve understanding of the data is to run separate pairwise binary logistic regression models for each class comparison (e.g., low vs. mid, low vs. high, mid vs. high). This approach provides more straightforward statistics for performing model diagnostics. Binary classification models have greater access to diagnostic tools such as the Hosmer-Lemeshow goodness-of-fit test, residual plots, and influence plots. This level of precision helps to assess the validity of each model and pinpoint areas for improvement.

7 Plots

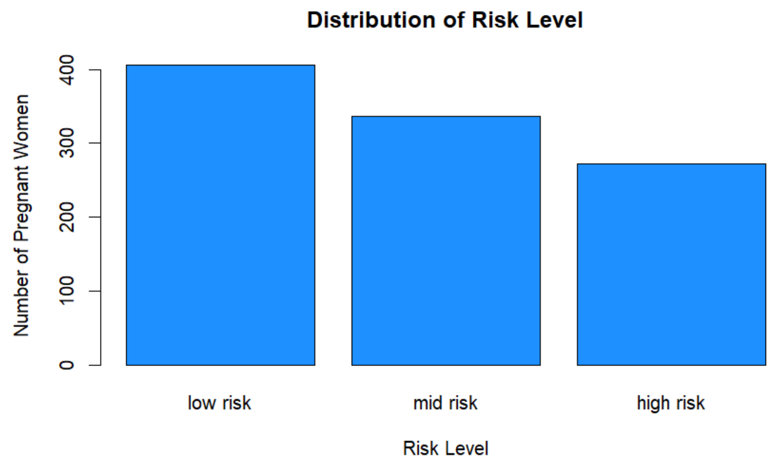

Figure 1: Distribution of Risk Level

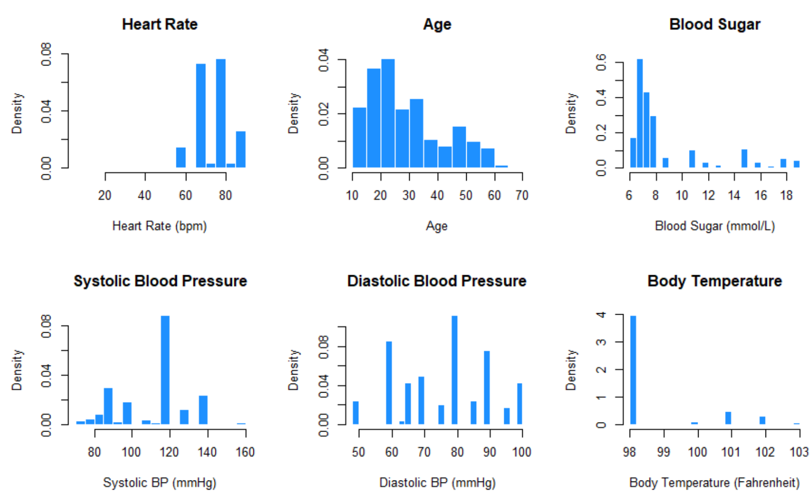

Figure 2: Prediction Distribution

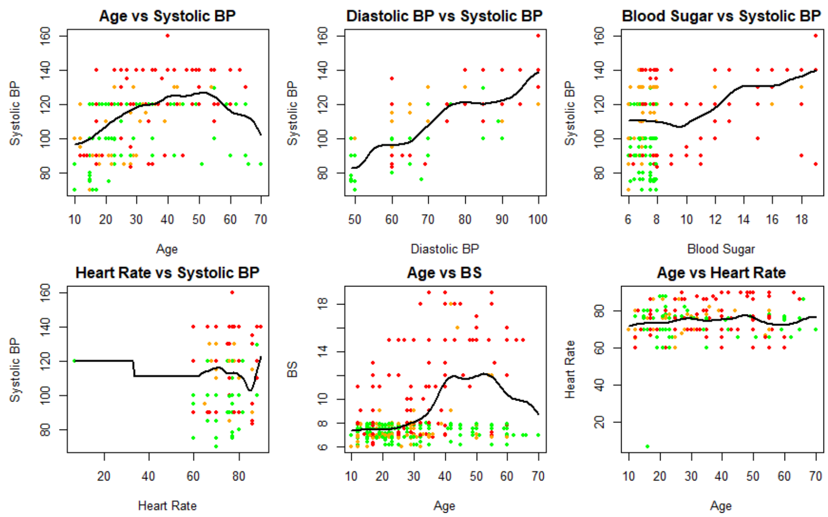

Figure 3: Relationship between Variables with Kernel Regression

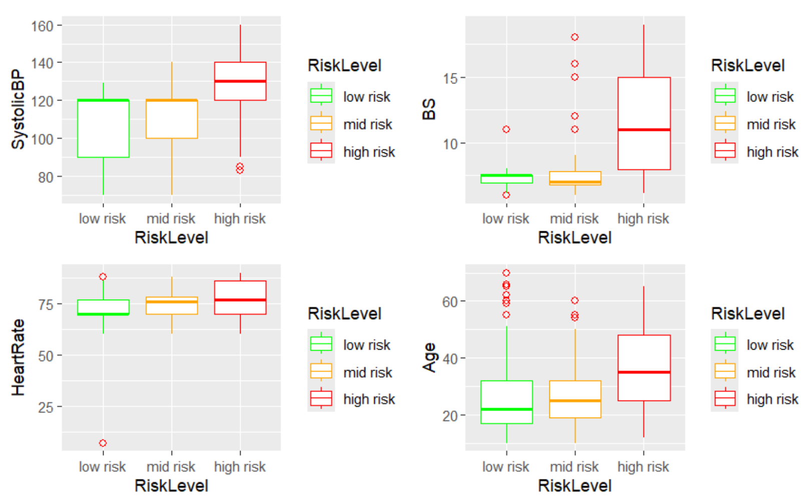

Figure 4: Boxplot of Predictor Variables

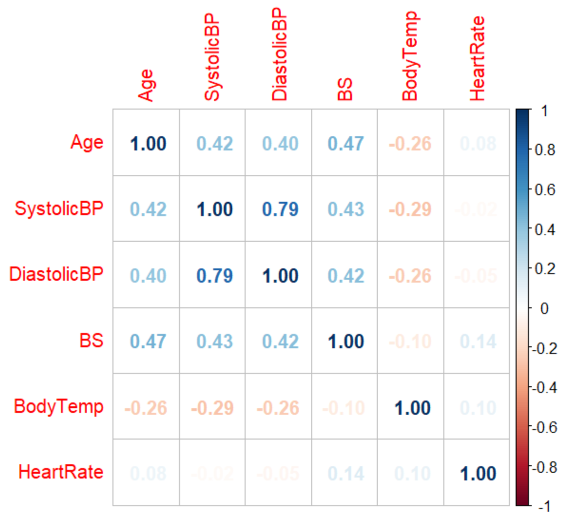

Figure 5: Correlation Matrix

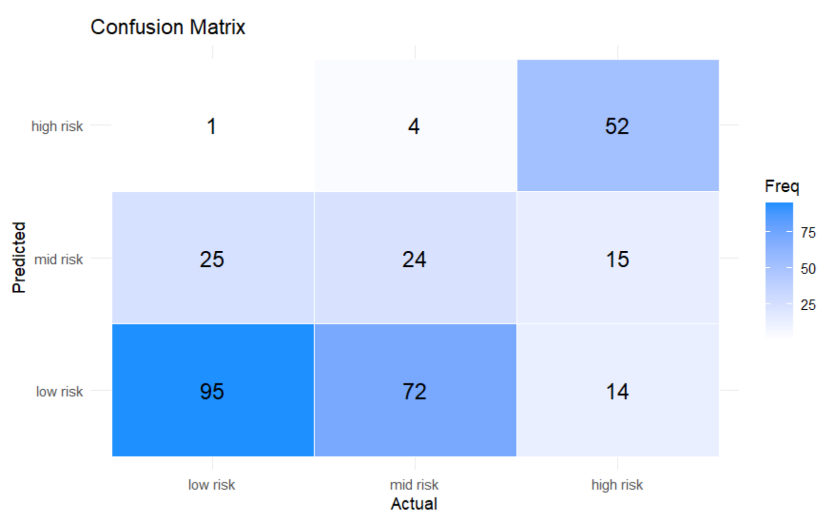

Figure 6: Confusion Matrix for Multinomial Logistic Regression

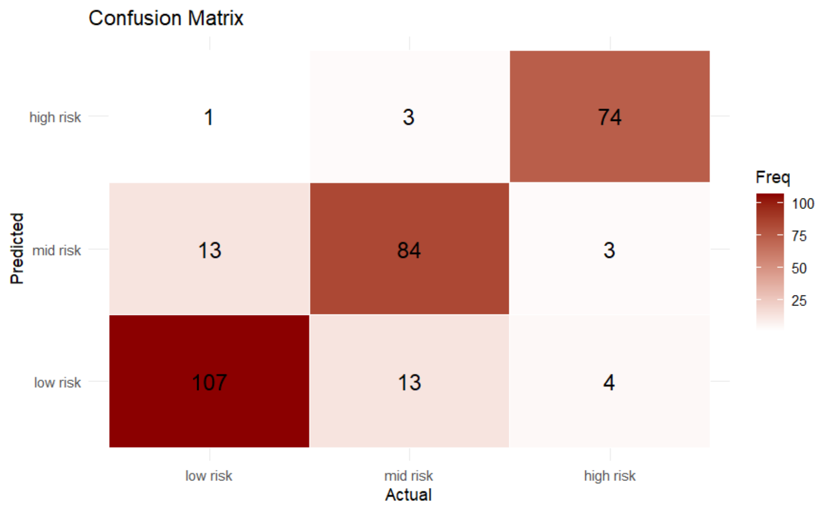

Figure 7: Confusion Matrix for Random Forest

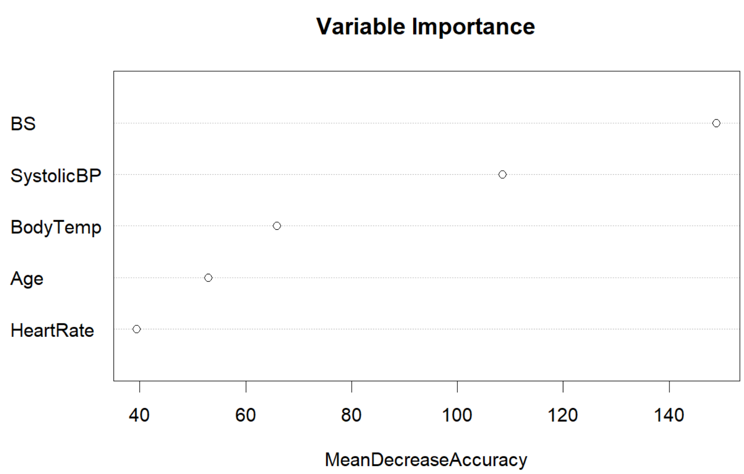

Figure 8: Random Forest Variable Importance

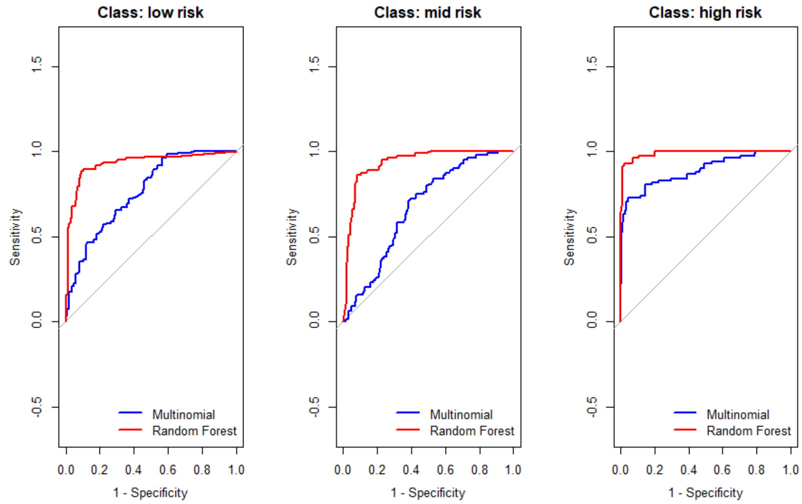

Figure 9: ROC Curves

References

[1] World Health Organization. Maternal mortality. `https://www.who.int/news-room/fact-sheets/detail/maternal-mortality`, 2023.