Machine Learning 2: Maternal Health Risk Classification

Group I: Aisha, Mufaddal, Raju Ahmed

Abstract

Maternal health is a major challenge in Bangladesh, particularly in rural areas where healthcare is hard to access. Many pregnant women suffer from conditions like high blood pressure and infections that go unnoticed due to a lack of medical facilities and trained professionals. Early marriages, limited education, and poverty add to the problem. Women often cannot reach healthcare centers in time, leading to complications which increases the risk associated with the same. This project develops machine learning models to predict maternal health risk using vital health indicators, enabling early identification of high-risk pregnancies.

Contents

1 Introduction

1.1 Problem Statement

Maternal mortality remains a critical global health challenge. This project develops machine learning models to predict maternal health risk as a **binary classification** (High Risk vs. Not High Risk) based on vital health indicators.

Rationale for Binary Classification: The original dataset contains three ordinal risk levels (low, mid, high). Since ordinal relationships are not optimally captured by standard multi-class classifiers, we aggregate mid and low risk into “Not High Risk.” This directly addresses: *“Is this pregnancy high-risk?”*

1.2 Dataset Description

The Maternal Health Risk dataset was collected from hospitals in rural Bangladesh via an IoT-based monitoring system (Ahmed and Kashem 2023). It contains 1,014 observations with 6 predictor variables.

Dataset Source: UCI Machine Learning Repository - Maternal Health Risk

Attributes Description:

- **Age:** Age of the pregnant woman in years, ranging from teenagers to older mothers
- **SystolicBP / DiastolicBP:** Blood pressure measurements (mmHg), key indicators for hypertension-related complications like preeclampsia
- **BS (Blood Sugar):** Blood glucose level (mmol/L), important for detecting gestational diabetes
- **BodyTemp:** Body temperature (°F), elevated values may indicate infections
- **HeartRate:** Resting heart rate (bpm), abnormal values may signal cardiovascular stress
- **RiskLevel:** Target variable with three original categories (low, mid, high risk) converted to binary classification

Table 1: Sample Data: First 5 Observations

Variable	Description	Range
Age	Age of pregnant woman (years)	10-70
SystolicBP	Systolic blood pressure (mmHg)	70-160
DiastolicBP	Diastolic blood pressure (mmHg)	49-100
BS	Blood sugar level (mmol/L)	6.0-19.0
BodyTemp	Body temperature (°F)	98-103
HeartRate	Heart rate (bpm)	7-90
RiskLevel	Target: High Risk vs. Not High Risk	2 classes

Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
25	130	80	15.0	98	86	high risk
35	140	90	13.0	98	70	high risk
29	90	70	8.0	100	80	high risk
30	140	85	7.0	98	70	high risk
35	120	60	6.1	98	76	low risk

2 Exploratory Data Analysis

2.1 Data Quality and Preprocessing

The dataset has **no missing values**. Two observations with HeartRate = 7 bpm (physiologically impossible) were removed, leaving **1012 observations**.

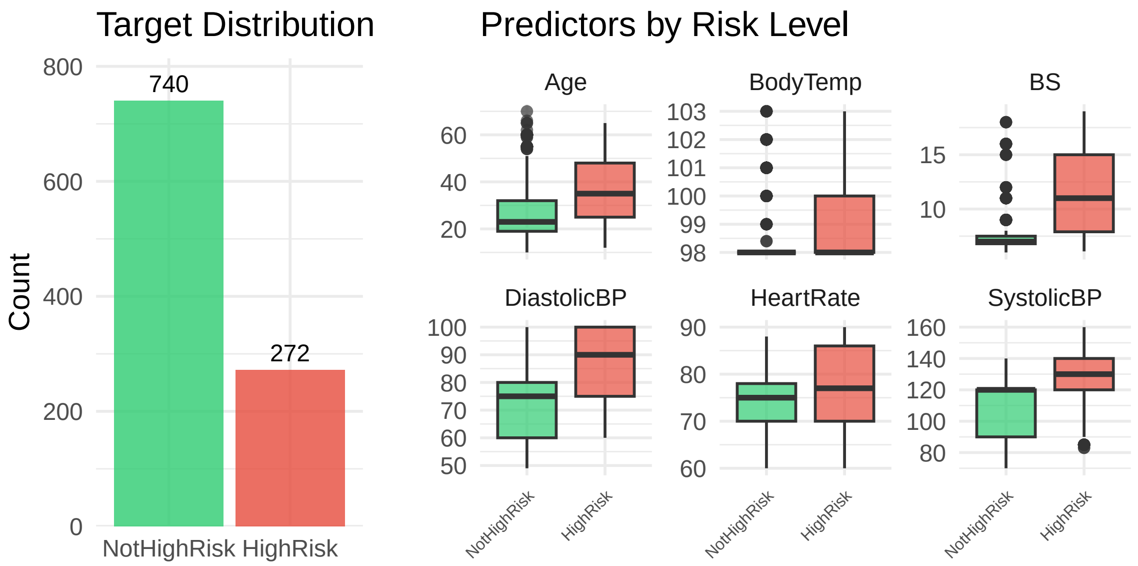

Figure 1: Target Distribution and Predictor Variables by Risk Level

Key Observations: From the boxplots, we can see that high-risk pregnancies tend to have elevated Blood Sugar (BS) and Systolic BP levels compared to non-high-risk cases. Since only about 27% of cases are high-risk, we use stratified sampling to ensure both classes are well-represented during training.

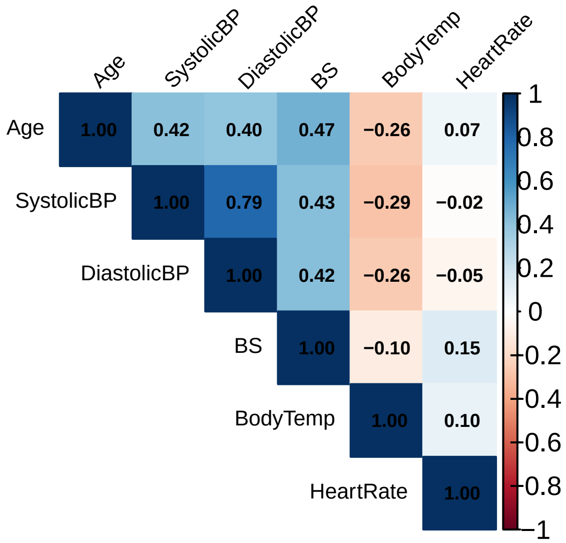

Figure 2: Correlation Matrix

The correlation matrix shows no severe multicollinearity among the predictors. Systolic and Diastolic blood pressure show a moderate positive correlation (r = 0.79), which is expected as both measure arterial pressure during different phases of the cardiac cycle (Franklin et al. 1999).

2.2 Summary Statistics

Table 3: Summary Statistics of Predictor Variables

Variable	Min	Mean	Max	SD
Age	10	29.9	70	13.5
SystolicBP	70	113.2	160	18.4
DiastolicBP	49	76.5	100	13.9
BS	6	8.7	19	3.3
BodyTemp	98	98.7	103	1.4
HeartRate	60	74.4	90	7.5

3 Model Selection and Mathematical Overview

By looking at the target variable, this is a binary classification problem. We decided to implement Random Forest and SVM models.

3.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training (Breiman 2001). For each tree in the forest, a bootstrap sample is drawn with replacement from the training data. During the construction of each tree, only a random subset of $m = \sqrt{p}$ features is considered at each node split, introducing additional randomness beyond the bootstrap sampling. Each tree is grown to maximum depth without pruning, allowing it to capture complex patterns in the data. Finally, predictions from all trees are combined through majority voting for classification tasks.

Prediction Formula:

$$\hat{f}(x) = \text{mode}\{h_1(x), h_2(x), ..., h_B(x)\}$$

where $h_b(x)$ is the prediction of tree b and B is the total number of trees (controlled by the `ntree` hyperparameter).

Split Criterion - Gini Impurity:

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the proportion of class k observations at node t. A split is chosen to maximize the reduction in impurity.

Key Hyperparameters:

- **`ntree`**: How many trees to grow. Adding more trees helps up to a point, but after around 500 the gains become negligible.
- **`mtry`**: At each split, only a random subset of features is considered. This keeps the trees different from each other. The default for classification is $\sqrt{p}$.
- **`nodesize`**: The smallest allowed leaf size. Letting leaves get very small (default is 1) can overfit the training data.

3.2 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised learning method primarily used for binary classification, though it can be extended to multiple classes. Mathematically, it works by identifying a hyperplane that serves as a boundary between different regions in a feature space. To find the "optimal" hyperplane, SVM seeks which is the boundary furthest away from the nearest data points. The distance between the boundary and the nearest points is called the margin.

Non-Linear SVM

When data cannot be separated by a straight line or plane, a linear boundary is insufficient. Non-linear SVMs address this by expanding the feature space.

The Kernel Trick: Rather than explicitly calculating the coordinates in a massive or infinite-dimensional space—which is computationally expensive—SVMs use kernel functions. This "trick" calculates the similarity between data points as if they were in a higher-dimensional space without actually performing the transformation.

Radial Basis Function (RBF) Kernel:

$$K(x_i, x_j) = \exp\left(-\gamma ||x_i - x_j||^2\right)$$

With respect to the dataset we have, it is not possible to separate it using a straight line, hence the Non-Linear SVM approach.

Hyperparameter Tuning:

- **C (Cost/Budget):** This is a general hyperparameter used in the Support Vector Classifier (soft margin) algorithm, which also applies to non-linear SVMs. It acts as a budget for the total amount of "slack" allowed for observations to fall on the wrong side of the margin or hyperplane.
- γ **(Gamma):** Defines the influence of a single training example. A low gamma makes the influence of each training point large, while a high gamma makes it small. Note: The `caret` package uses `sigma` where $\gamma = \frac{1}{2\sigma^2}$.

Model Evaluation: Confusion matrix and ROC-AUC curves can help us to evaluate classification problems.

4 Model Fitting and Comparison

4.1 Data Splitting and Cross-Validation

```
## Original Training Data Class Distribution:
```

```
##
## NotHighRisk    HighRisk
##         592         218
```

Data Split Strategy: Following professor's guidelines, since we use 10-fold cross-validation for hyperparameter tuning, we combine training and validation sets (60% + 20% = 80%). The test set (20%) is held out exclusively for final model comparison.

4.2 Handling Class Imbalance

The dataset shows class imbalance (~27% HighRisk vs ~73% NotHighRisk). To prevent bias toward the majority class, we apply **oversampling** to balance the training data.

```
## Balanced Training Data Class Distribution:
```

```
##
## NotHighRisk    HighRisk
##         592         592
```

After oversampling, the training data is balanced with equal representation of both classes. This ensures the models learn to identify high-risk cases effectively without being biased toward the majority class. The test set remains **unbalanced** to reflect real-world class distribution for fair evaluation.

4.3 Model Training

We trained Random Forest and SVM models using 10-fold cross-validation with AUC-ROC as the optimization metric.

For Random Forest, we set `ntree = 500` and tuned `mtry` over values 2, 3, 4, and 5 to find the best number of features considered at each split. We kept `nodesize` at its default value (1) since `mtry` has the most impact on performance. For SVM, we tuned `C` (cost) over 0.1, 1, and 10, and `sigma` over 0.01, 0.1, and 1 to control the margin flexibility and RBF kernel width.

```r
set.seed(123)
# Random Forest
rf_model <- train(RiskLevel ~ ., data = train_data, method = "rf", trControl =
 ↪  cv_control,
                  tuneGrid = expand.grid(mtry = c(2, 3, 4, 5)), ntree = 500,
                  importance = TRUE, metric = "ROC")

# SVM with RBF Kernel (requires scaling)
preProcValues <- preProcess(train_data[, 1:6], method = c("center", "scale"))
train_scaled <- predict(preProcValues, train_data)
test_scaled <- predict(preProcValues, test_data)

set.seed(123)
svm_model <- train(RiskLevel ~ ., data = train_scaled, method = "svmRadial", trControl =
 ↪  cv_control,
                   tuneGrid = expand.grid(C = c(0.1, 1, 10), sigma = c(0.01, 0.1, 1)),
                   metric = "ROC")
```

```
## Cross-Validation Results (AUC-ROC):

## Random Forest: 0.9856

## SVM: 0.9676
```

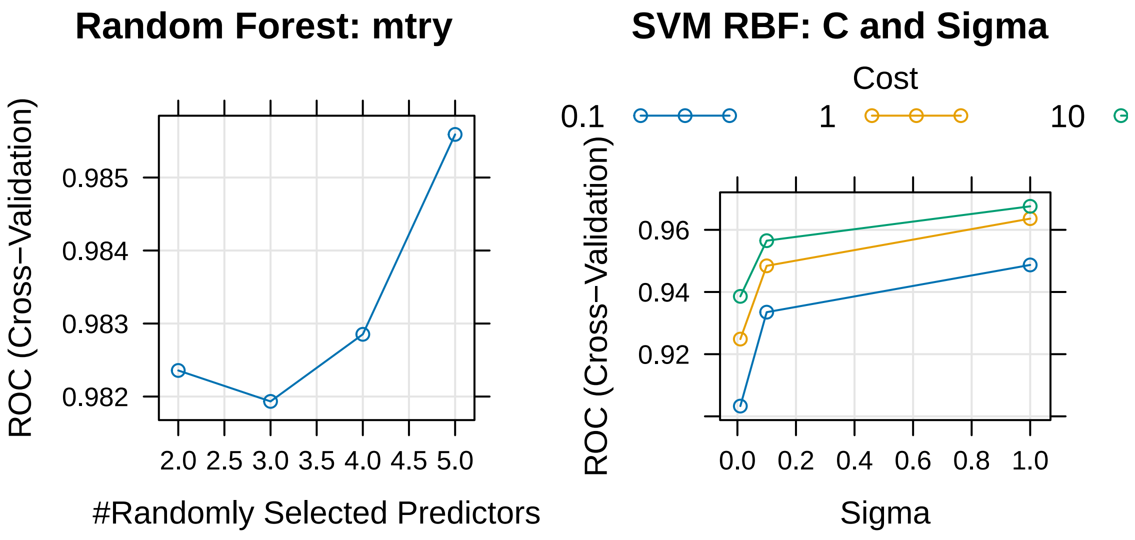

Figure 3: Hyperparameter Tuning Results

The left plot shows how cross-validated AUC changes with different `mtry` values for Random Forest. The right plot shows SVM performance across combinations of C and sigma — each line represents a different sigma value.

Best Hyperparameters: RF: mtry = 5; SVM: C = 10, sigma = 1

4.4 Test Set Evaluation

Both models are tested on unseen data to check how well they perform.

Table 4: Test Set Performance Metrics (in percentage)

	Metric	RF	SVM
Accuracy	Accuracy	94.1	91.6
Sensitivity	Sensitivity	88.9	85.2
Specificity	Specificity	95.9	93.9
Pos Pred Value	Precision	88.9	83.6
F1	F1 Score	88.9	84.4

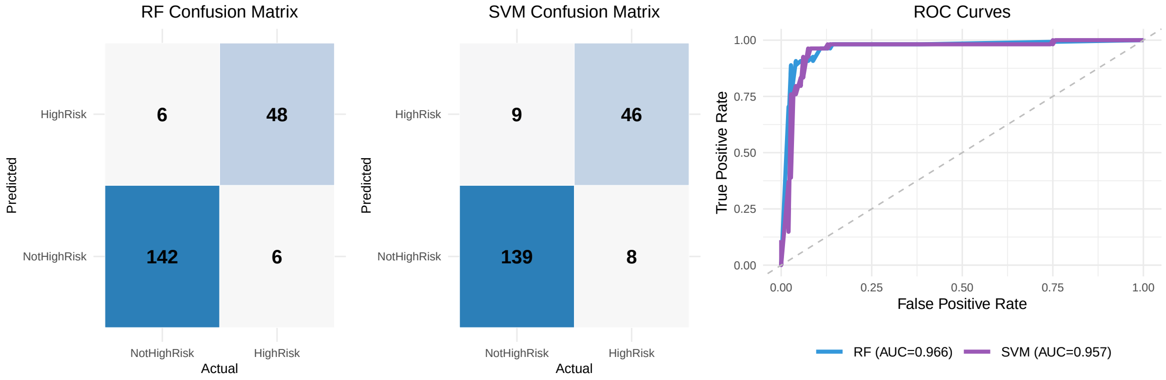

Figure 4: Confusion Matrices and ROC Curves

Overfitting Check: To ensure our models generalize well, we compare cross-validation AUC with test set AUC. Random Forest shows CV AUC of 0.986 and test AUC of 0.966. SVM shows CV AUC of 0.968 and test AUC of 0.957. The small gaps between CV and test performance confirm that neither model is overfitting.

Although both methods show good predictive results, machine learning models in the healthcare setting should be interpretable too. It is important to interpret the predictions to ensure trust and appropriateness. In the context of this study, interpretable machine learning methods will be used to interpret the global behavior and predictions of both models.

5 Interpretable Machine Learning (XAI)

In this section, the post-hoc interpretability techniques are applied to the developed model of the Random Forest and the Support Vector Machine. Global feature importance and global feature effects are examined. Local interpretability is also explored.

5.1 Feature Importance

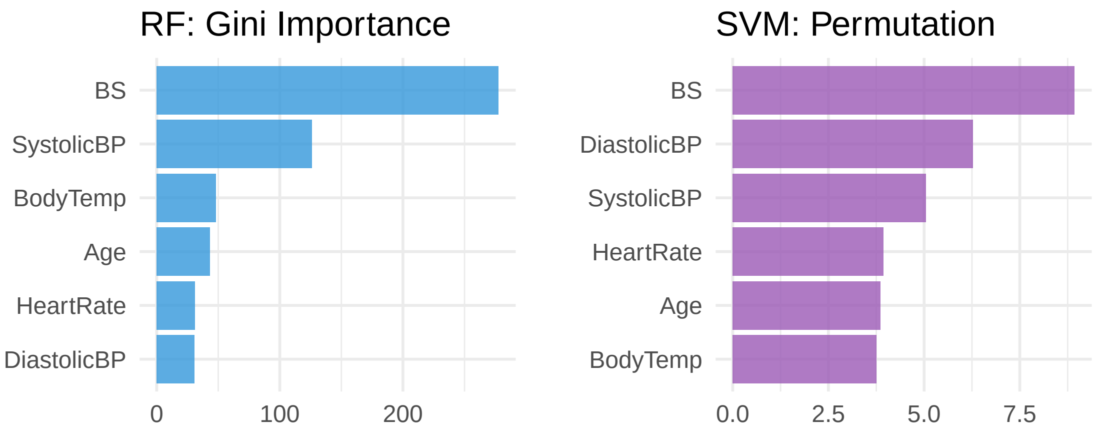

Figure 5: Feature Importance Comparison

Both models rank **Blood Sugar (BS)** as the most important feature, followed by **SystolicBP** and **Age**.

5.2 Partial Dependence Plots

Partial Dependence Plots (PDPs) show the average effect of a feature on the predicted outcome. The PDP value for a feature value s^* is:

$$\hat{f}_S(s^*) = \frac{1}{n}\sum_{i=1}^{n} f(S = s^*, C_i)$$

where C_i are the values of all other features for observation i.

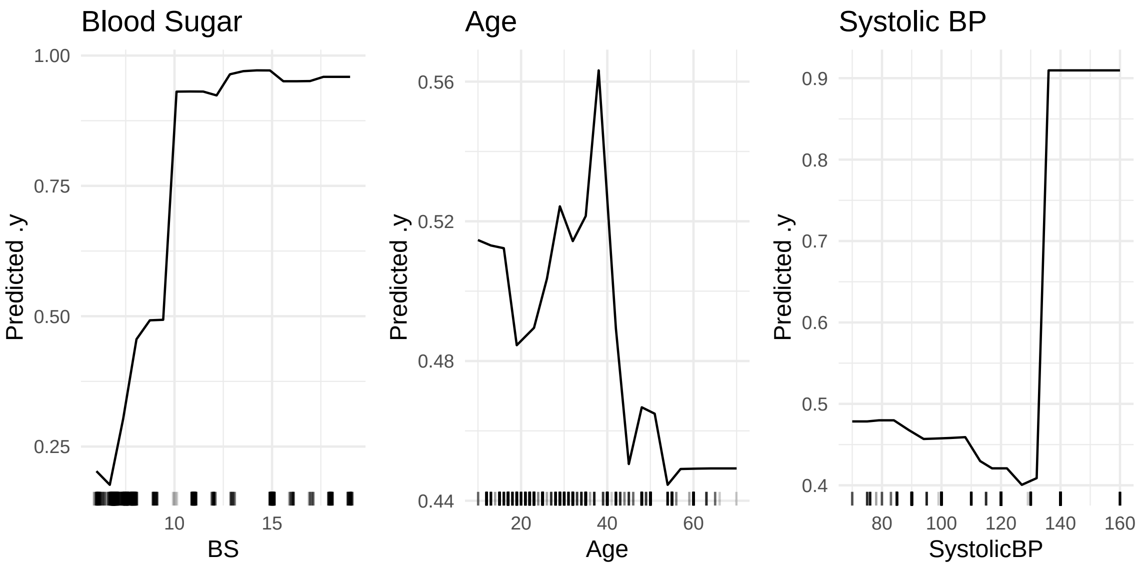

Figure 6: Partial Dependence Plots for Top 3 Features (Random Forest)

The PDPs show that higher Blood Sugar and Systolic BP values increase the probability of high-risk classification. Age shows a moderate positive effect, with risk increasing for older patients.

5.3 Local Explanations (SHAP)

SHAP (SHapley Additive exPlanations) explains individual predictions using Shapley values from game theory. The Shapley value represents the average marginal contribution of a feature across all possible feature combinations.

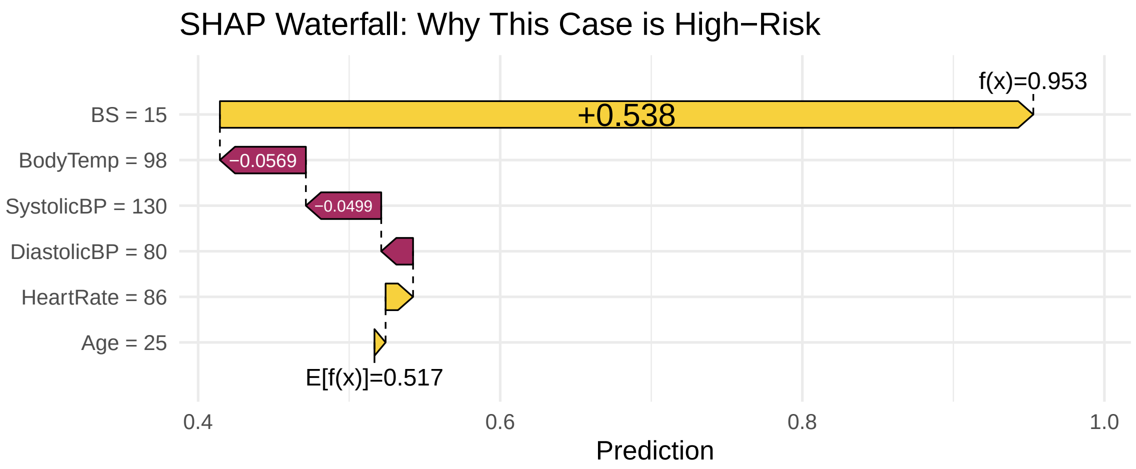

Figure 7: SHAP Waterfall Plots for a High-Risk Case

The waterfall plot shows how each feature contributes to pushing the prediction away from the baseline (average prediction). Features pushing right increase the high-risk probability, while those pushing left decrease it.

6 Conclusions

6.1 Summary

We trained Random Forest and SVM models for maternal health risk classification. Both models achieve strong performance in detecting high-risk pregnancies using binary classification.

6.2 Model Comparison Results

Table 5: Final Model Comparison Summary

Metric	RF	SVM
CV AUC-ROC	0.986	0.968
Test AUC-ROC	0.966	0.957
Accuracy	94.1%	91.6%
Sensitivity	88.9%	85.2%
Specificity	95.9%	93.9%
F1 Score	88.9%	84.4%

6.3 Key Findings

1. **Random Forest outperforms SVM** across most metrics, particularly in sensitivity which is critical for detecting high-risk cases
2. **Blood Sugar (BS) is the most important predictor** across both models, consistent with medical literature on gestational diabetes
3. **Systolic Blood Pressure and Age** are secondary important features, aligning with known risk factors for pregnancy complications
4. **Both models achieve excellent discrimination** with AUC-ROC > 0.90, indicating reliable separation between risk classes

6.4 Limitations

The dataset size (~1,000 observations) may limit generalizability, and the geographic scope is limited to rural Bangladesh. The feature set contains only 6 predictors, so additional clinical variables could improve predictions. Additionally, binary classification loses the granularity of the original ordinal risk levels.

7 References

Ahmed, Marzia, and Mohammod Abul Kashem. 2023. "Maternal Health Risk Data Set." UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/863/maternal+health+risk.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Franklin, Stanley S, Sarwat A Khan, Nathan D Wong, Martin G Larson, and Daniel Levy. 1999. "Is Pulse Pressure Useful in Predicting Risk for Coronary Heart Disease? The Framingham Heart Study." *Circulation* 100 (4): 354–60. https://doi.org/10.1161/01.CIR.100.4.354.