Machine Learning 2: Maternal Health Risk Classification

Aisha, Mufaddal, Raju Ahmed

2025-12-29

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.

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

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**.

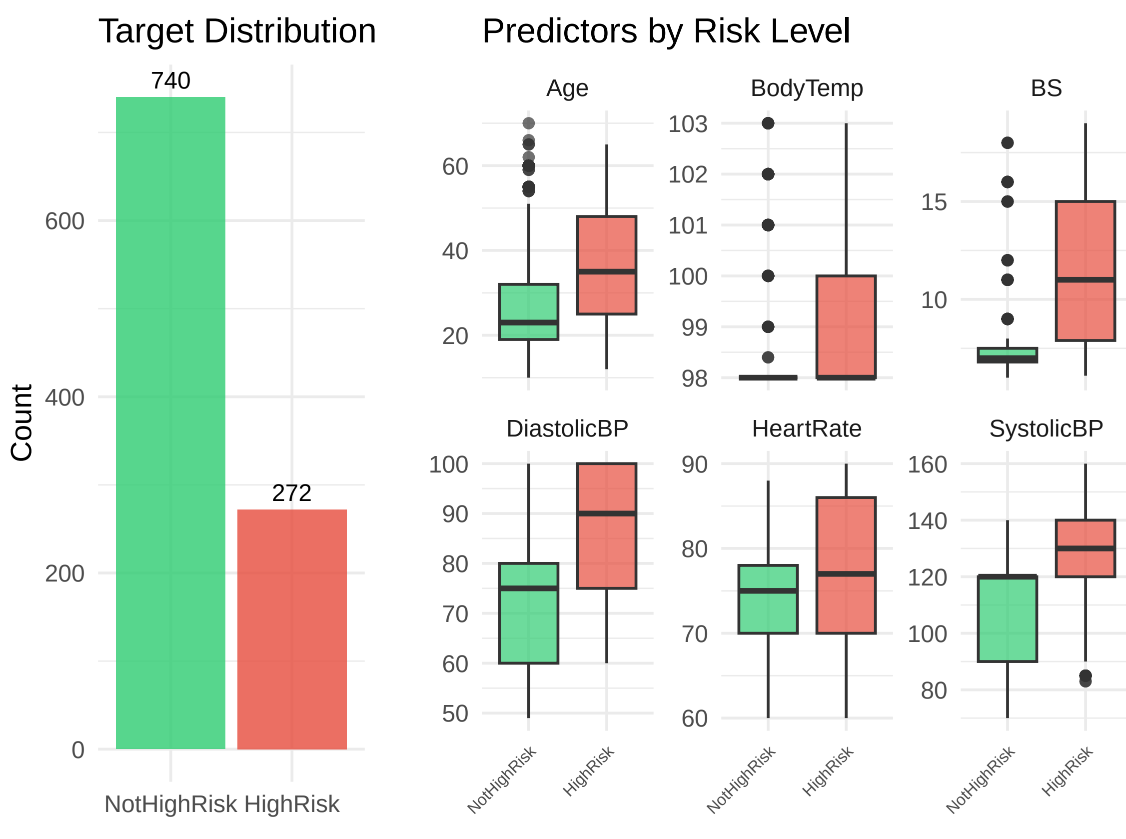

Figure 1: Target Distribution and Predictor Variables by Risk Level

Key Observations: Blood Sugar (BS) and Systolic BP are strong discriminators for high-risk cases. Class imbalance (~27% HighRisk) is addressed using stratified sampling.

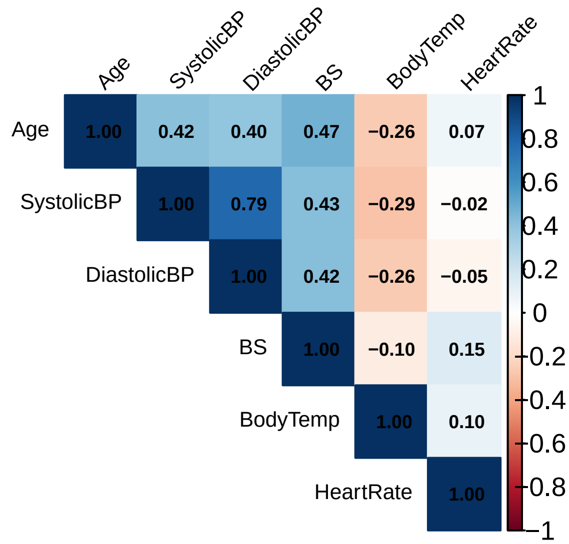

Figure 2: Correlation Matrix

No severe multicollinearity detected. Systolic and Diastolic BP show expected moderate correlation.

2.2 Summary Statistics

Table 2: 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 Mathematical Overview

3.1 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training (Breiman 2001). The algorithm works as follows:

1. **Bootstrap Sampling**: For each tree, draw a bootstrap sample (with replacement) from the training data
2. **Feature Randomization**: At each node split, only consider a random subset of $m = \sqrt{p}$ features
3. **Tree Construction**: Build each tree to maximum depth without pruning
4. **Aggregation**: Combine predictions via majority voting (classification)

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.

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` (number of trees), `mtry` (features per split), `nodesize` (minimum node size).

3.2 Support Vector Machine (SVM)

SVM finds the optimal separating hyperplane that maximizes the margin between classes (James et al. 2021). For non-linearly separable data, the soft-margin SVM solves:

Optimization Problem:

$$\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to: $y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i$ and $\xi_i \geq 0$

where C controls the trade-off between margin maximization and misclassification penalty, and ξ_i are slack variables.

RBF (Radial Basis Function) Kernel:

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

The RBF kernel maps data to infinite-dimensional space, enabling non-linear decision boundaries. Parameter γ controls the kernel width: larger γ means tighter fit (risk of overfitting).

Key Hyperparameters: C (cost/regularization), γ (kernel width, often denoted as `sigma` in R).

4 Model Fitting and Comparison

4.1 Data Splitting and Cross-Validation

Stratified 80/20 split: **810 training**, **202 test** observations. 10-fold CV used for hyperparameter tuning with AUC-ROC as the optimization metric.

4.2 Model Training

We trained three models (Random Forest, Decision Tree, SVM) and selected the **best 2 based on CV AUC-ROC** for final comparison.

```
## Model Selection (CV AUC-ROC):

## 1. Random Forest : 0.9802

## 2. SVM : 0.9531

## 3. Decision Tree : 0.9483

##
## Best 2 models selected: Random Forest and SVM
```

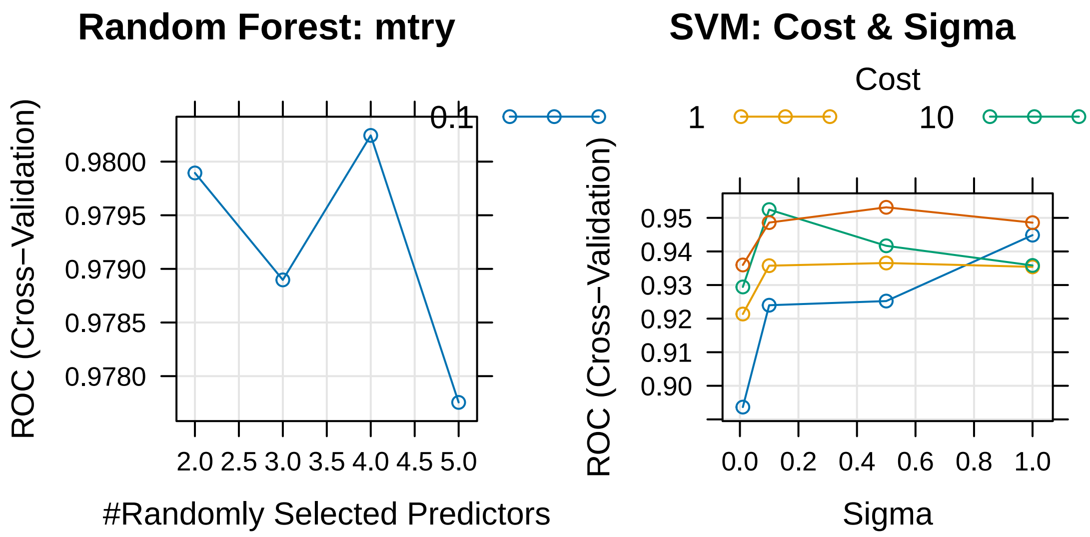

Figure 3: Hyperparameter Tuning Results (Best 2 Models)

Best Hyperparameters: RF: mtry = 4; SVM: C = 100, sigma = 0.5

4.3 Test Set Evaluation

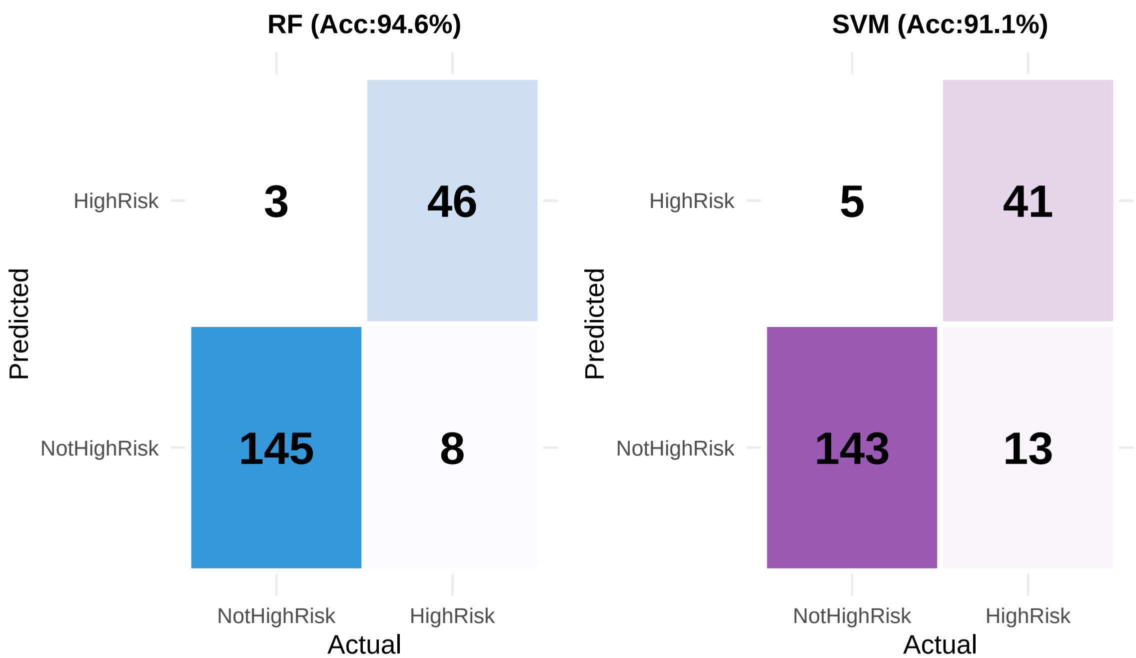

Figure 4: Confusion Matrices: Random Forest (left) and SVM (right)

Table 3: Test Set Performance Metrics (in percentage)

	Metric	RF	SVM
Accuracy	Accuracy	94.6	91.1
Sensitivity	Sensitivity	98.0	96.6
Specificity	Specificity	85.2	75.9
Pos Pred Value	Precision	94.8	91.7
F1	F1 Score	96.3	94.1

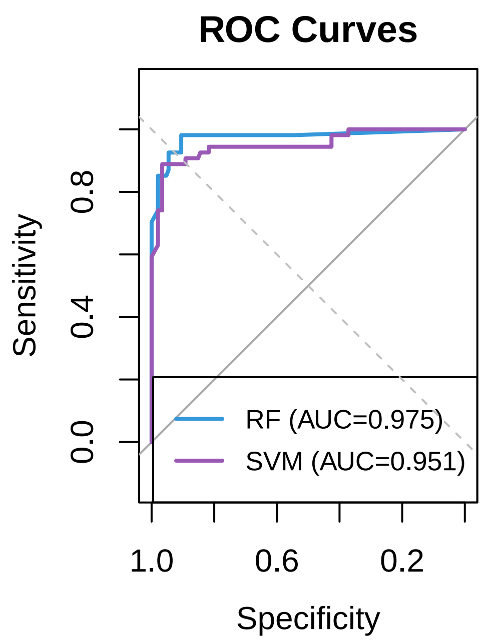

Figure 5: ROC Curves and Performance Comparison

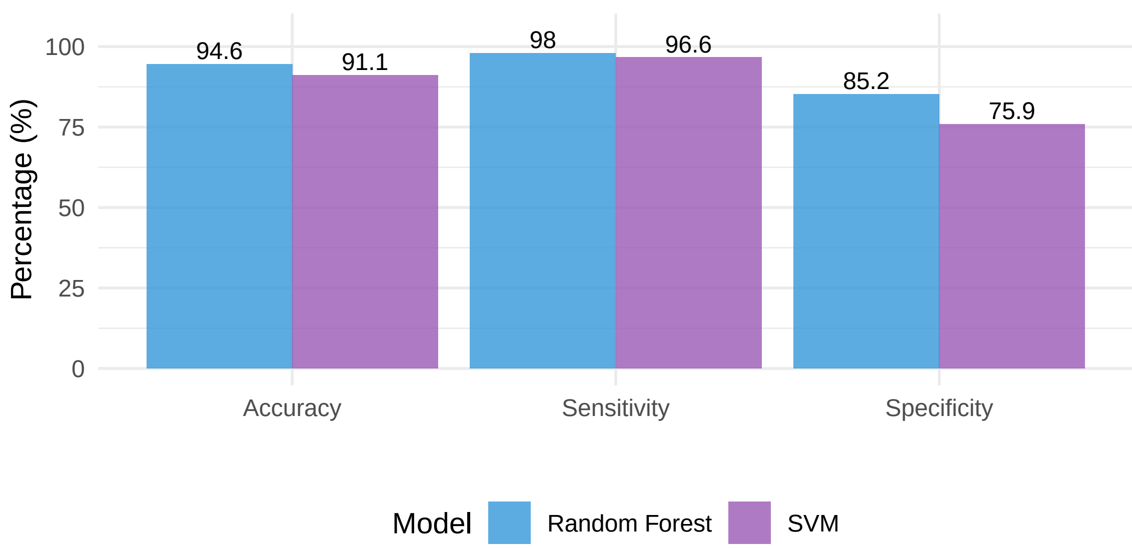

Figure 6: Model Performance Comparison

5 Interpretable Machine Learning (XAI)

5.1 Feature Importance

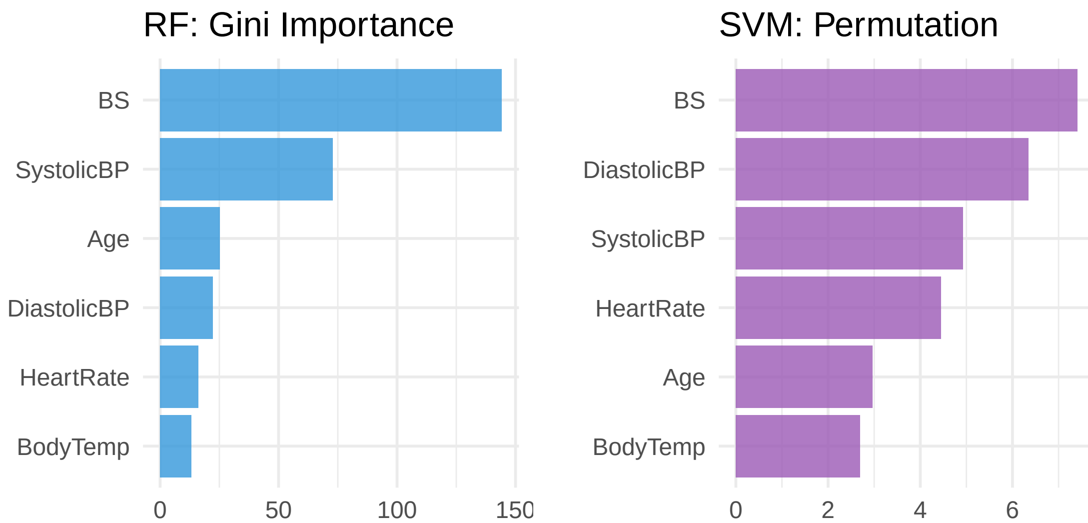

Figure 7: Feature Importance Comparison

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 marginal effect of a feature on the predicted outcome, averaging over all other features.

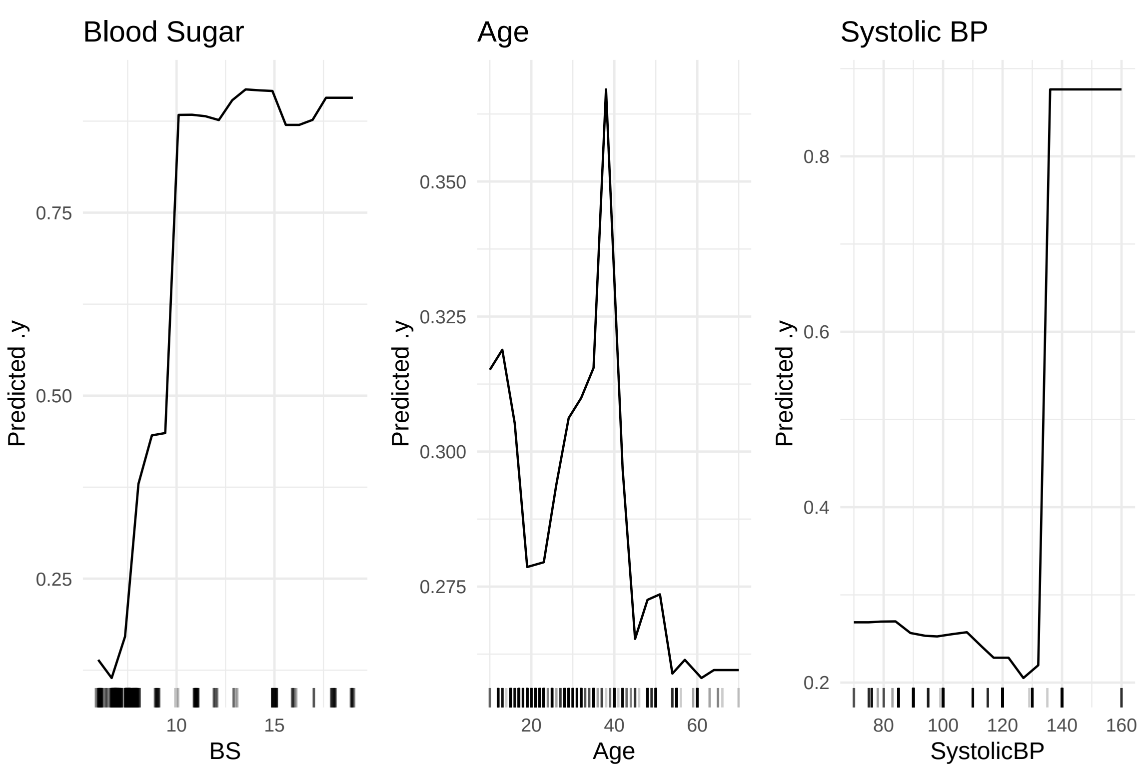

Figure 8: Partial Dependence Plots: Effect of Top 3 Features on High Risk Probability

Interpretation:

- **Blood Sugar (BS):** Strong positive relationship - risk increases sharply above 8 mmol/L, indicating a clinical threshold
- **Age:** Moderate positive effect - older maternal age associated with higher risk
- **Systolic BP:** Non-linear relationship - risk increases substantially above 130 mmHg (hypertension threshold)

5.3 Local Explanations (LIME)

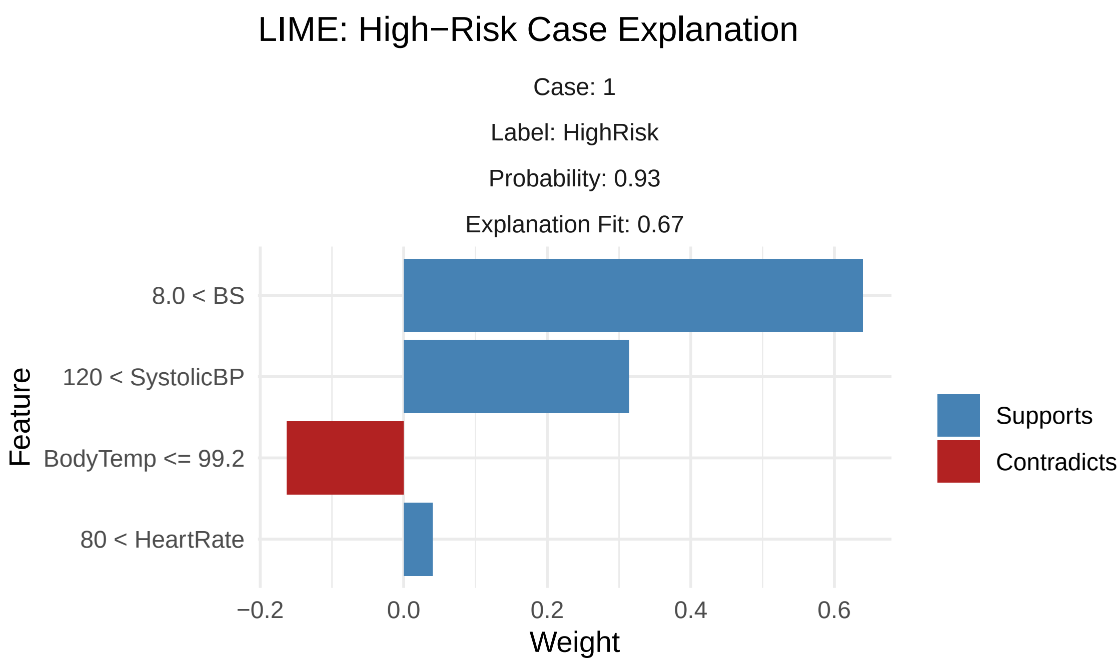

Figure 9: LIME Explanation for a High-Risk Case

LIME shows which features contributed most to the prediction for this specific case, providing interpretability for clinical review.

6 Conclusions

6.1 Summary

We trained three machine learning models (Random Forest, Decision Tree, SVM) and selected the **best 2 (Random Forest and SVM)** based on cross-validation AUC-ROC performance. Both selected models achieve strong performance in detecting high-risk pregnancies using binary classification.

6.2 Model Comparison Results

Table 4: Final Model Comparison Summary

Metric	RF	SVM
CV AUC-ROC	0.98	0.953
Test AUC-ROC	0.975	0.951
Accuracy	94.6%	91.1%
Sensitivity	98%	96.6%
Specificity	85.2%	75.9%
F1 Score	96.3%	94.1%

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 Clinical Recommendations

Based on our analysis, we recommend:

- **Primary Screening:** Use Random Forest model for initial risk assessment due to higher sensitivity
- **Key Indicators to Monitor:** Blood sugar levels should be closely monitored, especially values > 8 mmol/L
- **Blood Pressure Monitoring:** Systolic BP > 130 mmHg should trigger additional evaluation
- **Age Consideration:** Older maternal age warrants closer monitoring

6.5 Limitations

- Dataset size (~1,000 observations) may limit generalizability
- Geographic scope limited to rural Bangladesh
- Limited feature set (6 predictors) - additional clinical variables could improve predictions
- Binary classification loses granularity of 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.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in r*. 2nd ed. New York: Springer.