Machine Learning 2: Maternal Health Risk Classification

Aisha, Mufaddal, Raju Ahmed

2026-01-02

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**.

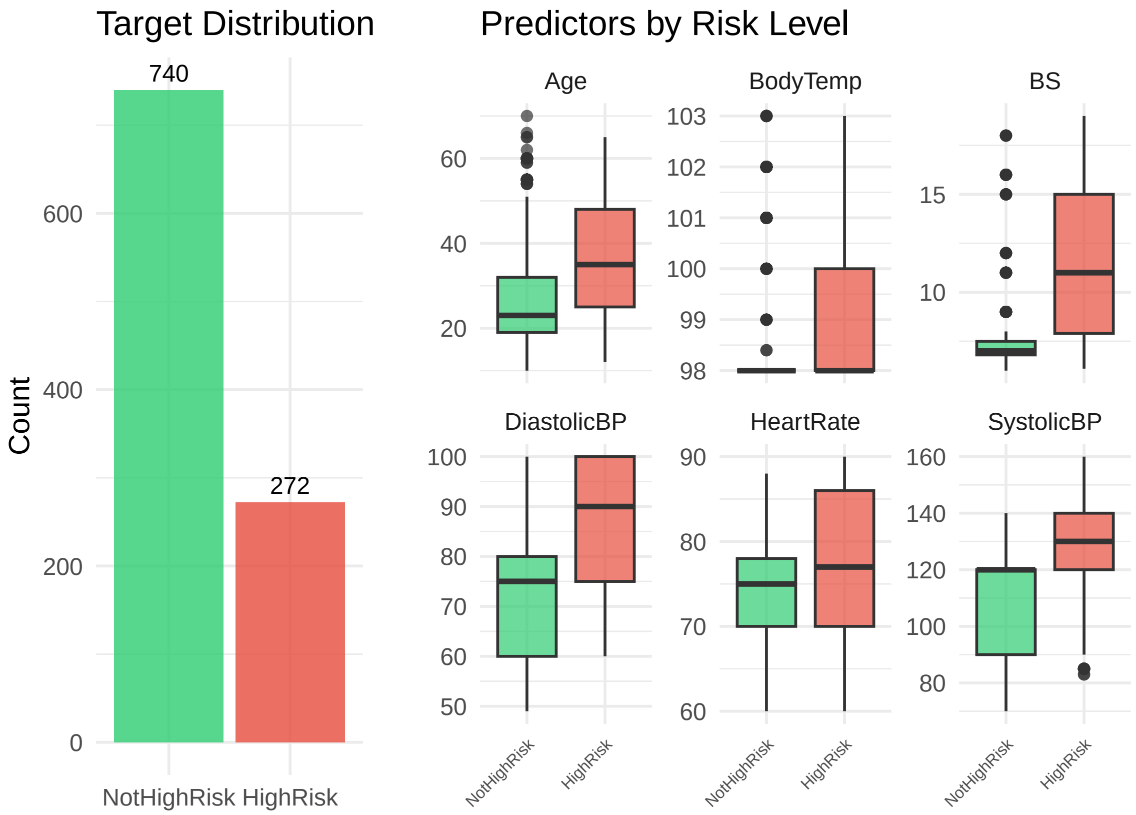

Figure 1: Target Distribution and Predictor Variables by Risk Level

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

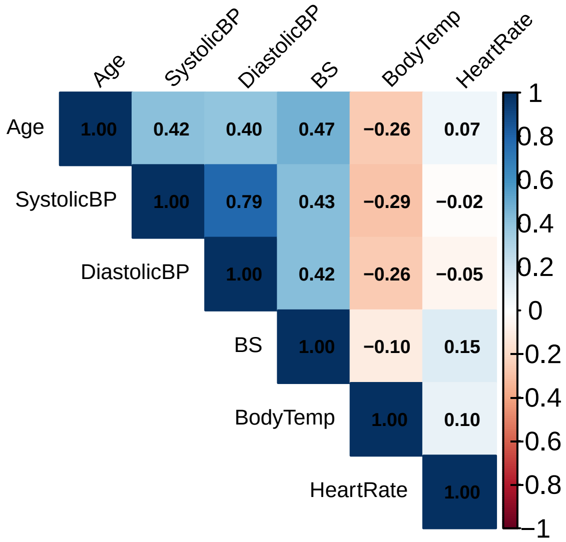

Figure 2: Correlation Matrix

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

2.2 Summary Statistics

Table 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 Rationale

Before diving into the mathematical foundations, we explain our choice of models for this classification task.

3.1 Why Random Forest and SVM?

For maternal health risk classification, we evaluated three candidate models: Random Forest, Decision Tree, and Support Vector Machine. Our final selection of **Random Forest and SVM** is based on the following considerations:

1. Performance-Based Selection: We trained all three models using 10-fold cross-validation and compared their AUC-ROC scores. Random Forest achieved the highest performance, followed by SVM, while Decision Tree showed lower performance due to its tendency to overfit without ensemble averaging.

2. Algorithmic Diversity: RF and SVM represent fundamentally different learning paradigms:

- **Random Forest** is an ensemble method combining multiple decision trees through bagging (bootstrap aggregating) and random feature selection. It naturally handles non-linear relationships and provides interpretable feature importance measures.
- **Support Vector Machine** is a geometric approach that finds the optimal hyperplane separating classes by maximizing the margin. Using the RBF (Radial Basis Function) kernel allows SVM to capture non-linear decision boundaries in the feature space.

3. Complementary Interpretability: Both models support model-agnostic explainability methods (feature importance, PDPs, LIME), allowing us to compare insights across different algorithmic perspectives. When both models agree on feature importance, we gain higher confidence in clinical interpretations.

4. Clinical Suitability: Both models output class probabilities, which is essential for medical risk stratification where clinicians need confidence scores rather than just binary predictions.

4 Mathematical Overview

4.1 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).

4.2 Support Vector Machine (SVM)

SVM finds the optimal separating hyperplane that maximizes the margin between classes (James et al. 2021). For the soft-margin SVM, the optimization problem is:

Optimization Problem:

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

subject to: $y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$

where:

- w is the weight vector in the transformed feature space
- b is the bias term
- C controls the trade-off between margin maximization and misclassification penalty
- ξ_i are slack variables allowing soft margin violations
- $\phi(x)$ is the feature transformation induced by the kernel

RBF (Radial Basis Function) Kernel:

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

The RBF kernel maps data into an infinite-dimensional space, enabling non-linear decision boundaries. It measures similarity between points based on their Euclidean distance, with γ controlling the influence radius of individual training examples.

Key Hyperparameters:

- C (cost): Controls regularization - higher values reduce misclassifications but may overfit
- γ (gamma): Controls kernel width - higher values create more complex boundaries

5 Model Fitting and Comparison

5.1 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.

5.2 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.

5.3 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 : 0.9856

## 2. SVM : 0.9676

## 3. Decision Tree : 0.9614

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

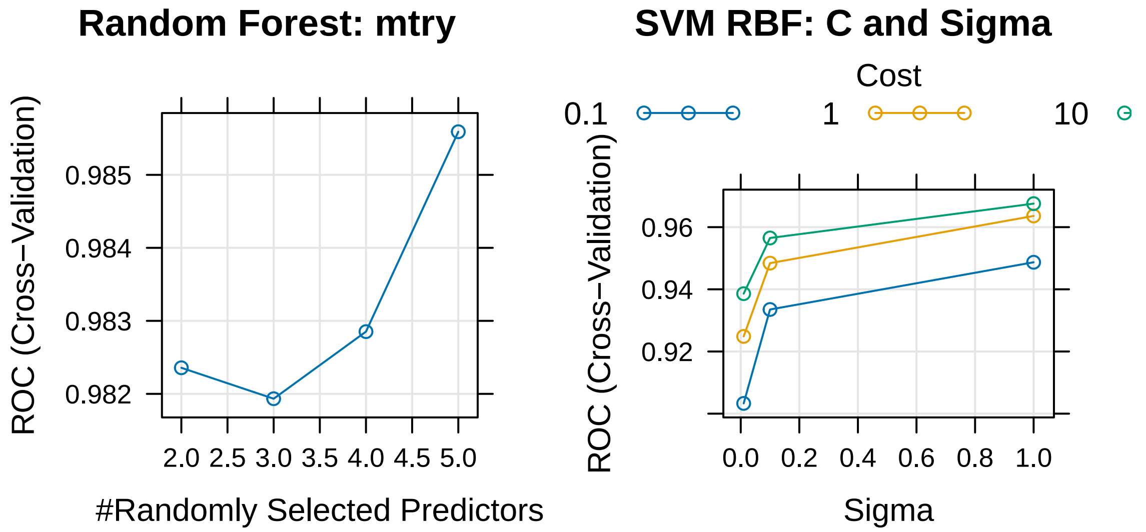

Figure 3: Hyperparameter Tuning Results (Best 2 Models)

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

5.4 Test Set Evaluation

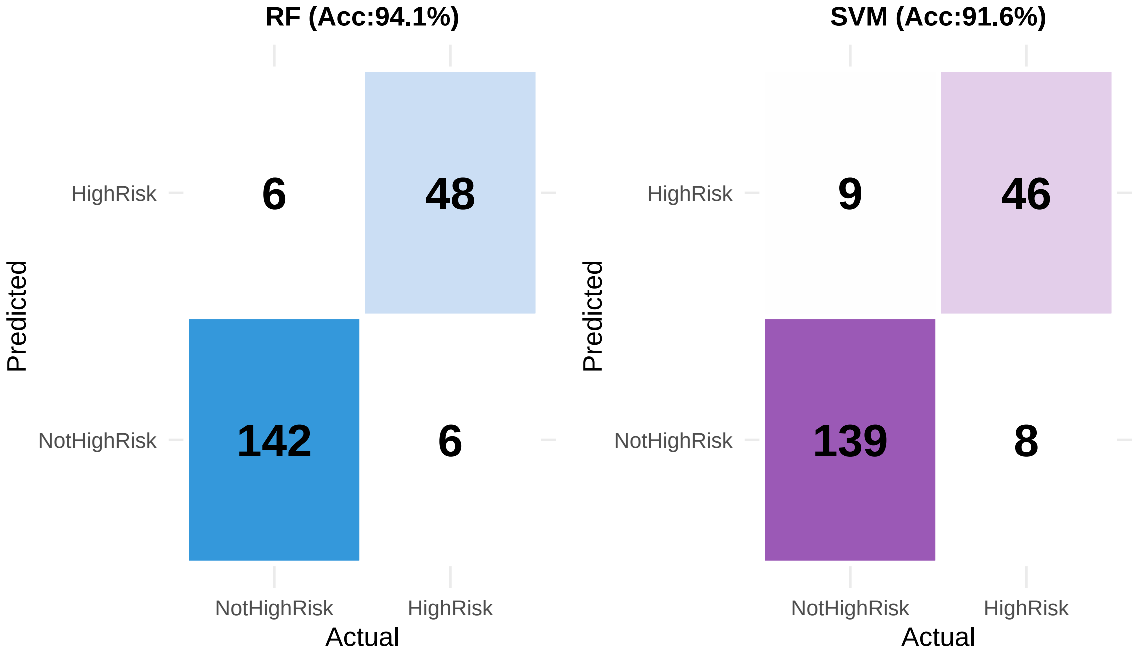

Figure 4: Confusion Matrices: Random Forest (left) and SVM (right)

Table 4: Test Set Performance Metrics (in percentage)

	Metric	RF	SVM
Accuracy	Accuracy	94.1	91.6
Sensitivity	Sensitivity	95.9	93.9
Specificity	Specificity	88.9	85.2
Pos Pred Value	Precision	95.9	94.6
F1	F1 Score	95.9	94.2

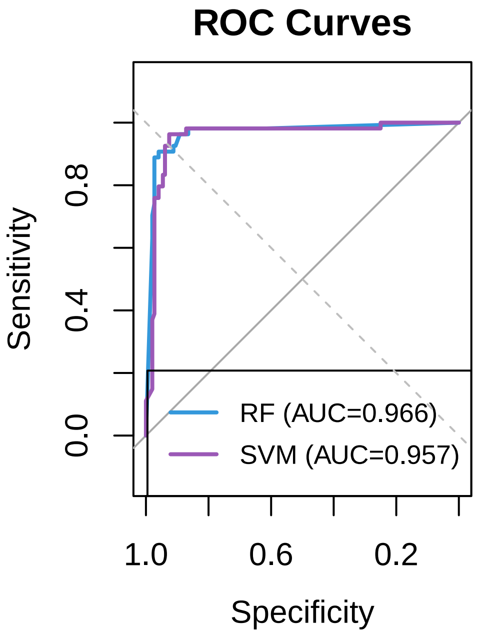

Figure 5: ROC Curves and Performance Comparison

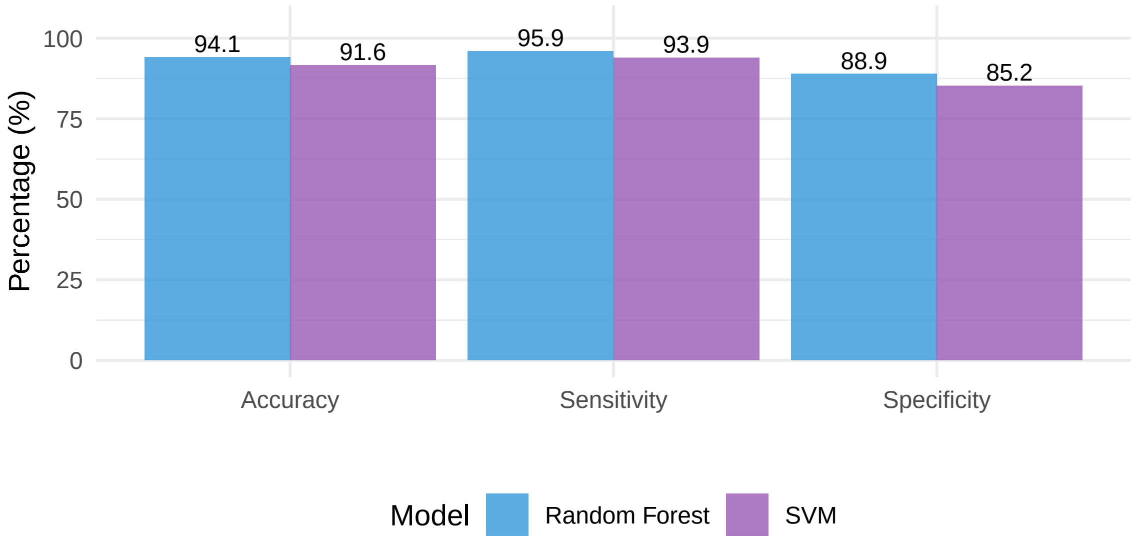

Figure 6: Model Performance Comparison

6 Interpretable Machine Learning (XAI)

6.1 Feature Importance

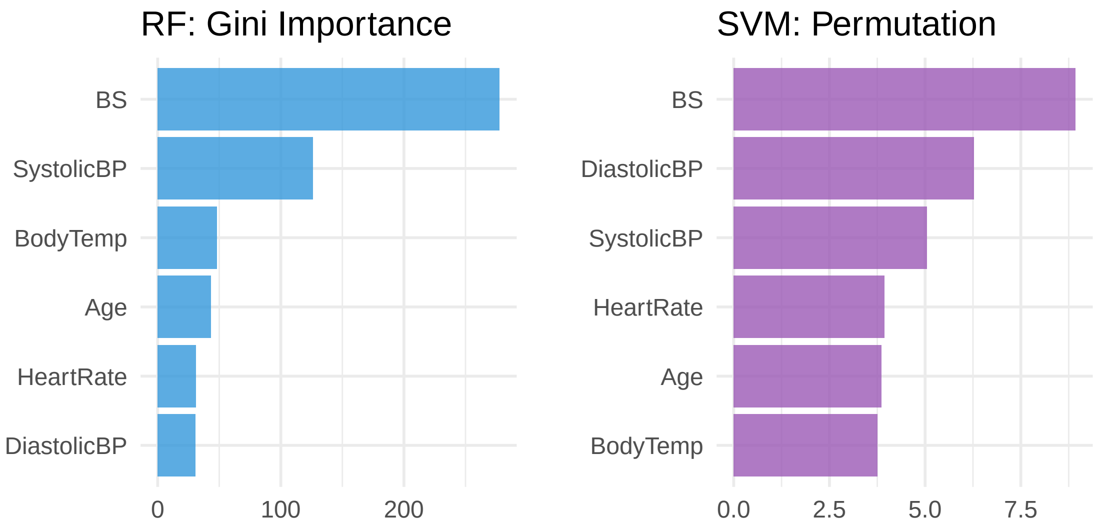

Figure 7: Feature Importance Comparison

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

6.2 Partial Dependence Plots

Partial Dependence Plots (PDPs) show the marginal effect of a feature on the predicted outcome, averaging over all other features. We compare PDPs for both Random Forest and SVM models.

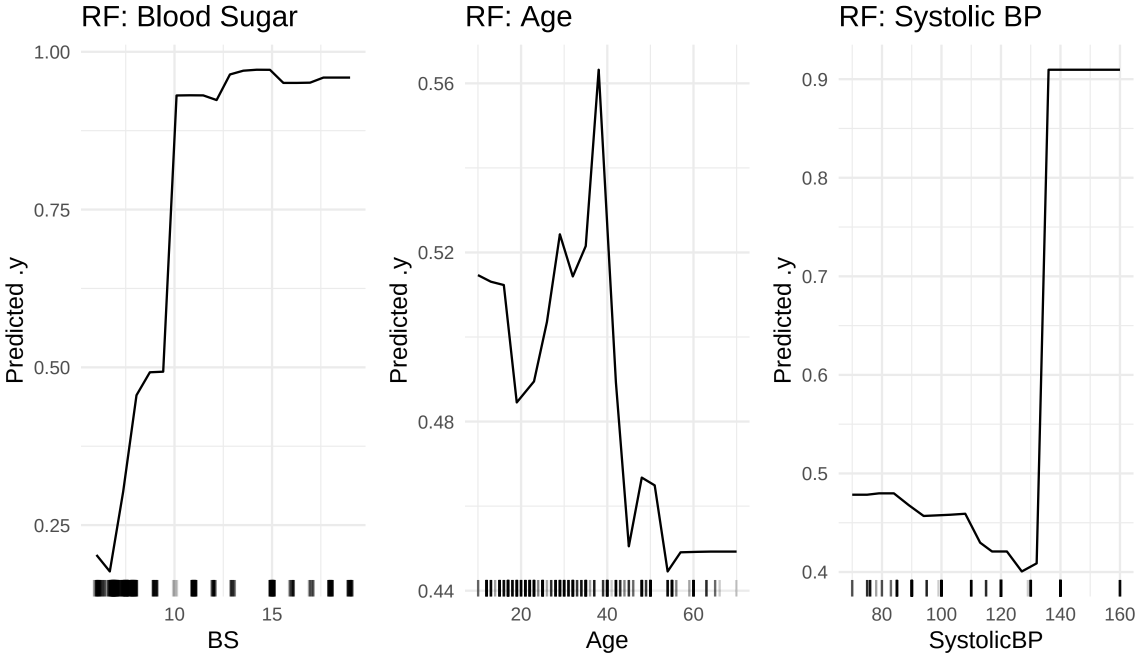

Figure 8: Random Forest: Partial Dependence Plots for Top 3 Features

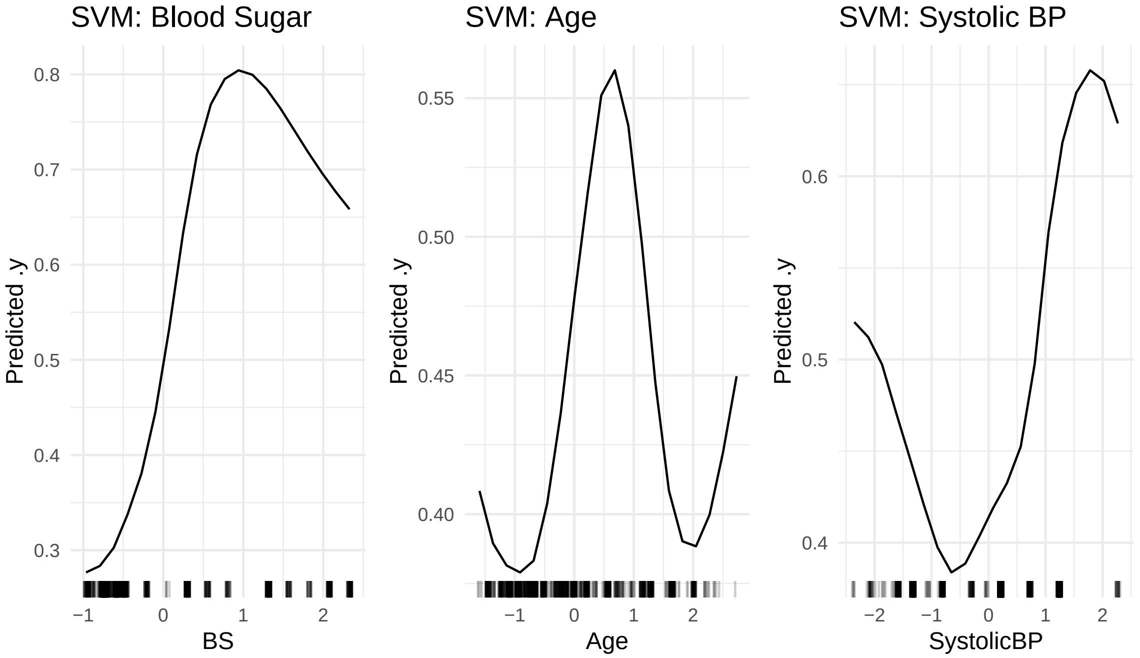

Figure 9: SVM: Partial Dependence Plots for Top 3 Features

PDP Comparison:

- **Blood Sugar (BS):** Both models show strong positive relationship. RF shows step-like pattern (tree-based), while RBF SVM shows smooth non-linear transitions due to its kernel-based decision boundary.
- **Age:** Both models show moderate positive effect, with RF displaying more abrupt changes and RBF SVM showing smoother curves.
- **Systolic BP:** Both identify ~130 mmHg as a risk threshold. Both models capture non-linear patterns, with RF showing discrete steps and SVM showing continuous transitions.

6.3 Local Explanations (LIME)

LIME (Local Interpretable Model-agnostic Explanations) provides instance-level explanations by fitting a simple interpretable model locally around a prediction. We compare LIME explanations for both models on the same high-risk case.

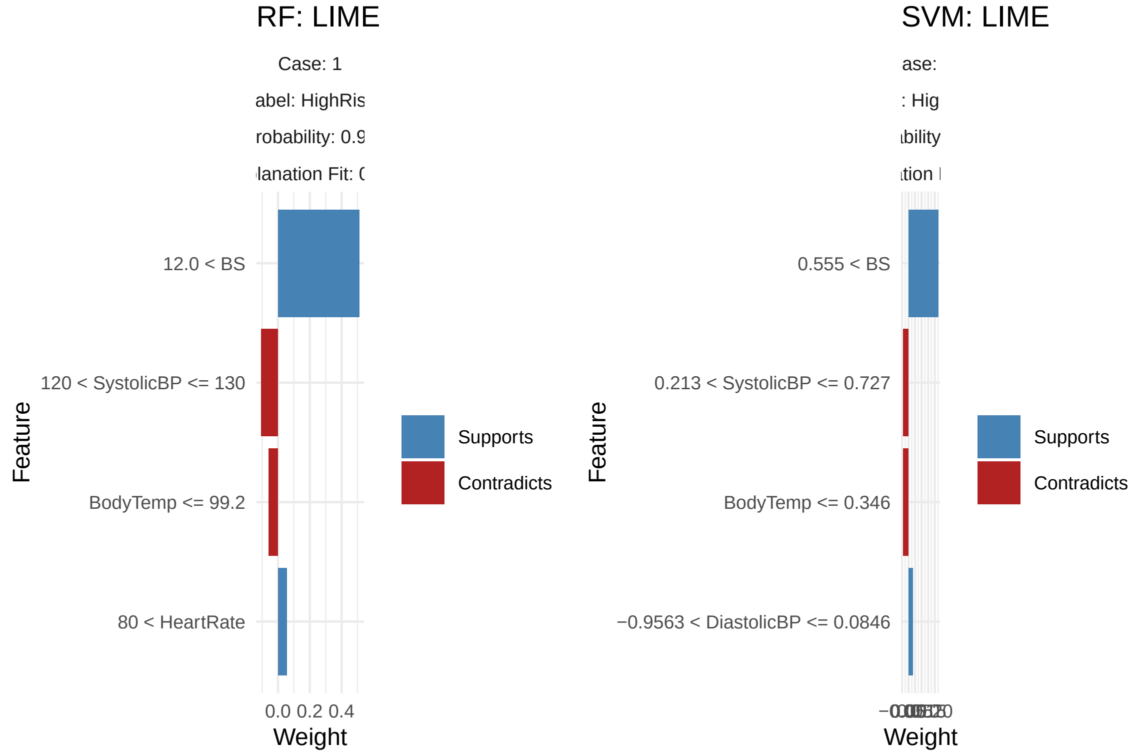

Figure 10: LIME Explanations: RF (left) vs SVM (right) for Same High-Risk Case

LIME Comparison: Both models identify similar key features for this high-risk case, but with different contribution magnitudes. This consistency across models strengthens confidence in the clinical interpretation.

7 Conclusions

7.1 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.

7.2 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	95.9%	93.9%
Specificity	88.9%	85.2%
F1 Score	95.9%	94.2%

7.3 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

7.4 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

7.5 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

8 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.