

Machine Learning 2: Maternal Health Risk Classification

Aisha, Mufaddal, Raju Ahmed

2025-12-27

Contents

1	Introduction	1
1.1	Problem Statement	1
1.2	Dataset Description	2
2	Exploratory Data Analysis	2
2.1	Data Quality Check	2
2.2	Outlier Detection and Handling	2
2.3	Target Variable Distribution	3
2.4	Descriptive Statistics	4
2.5	Distribution of Predictor Variables	4
2.6	Correlation Analysis	5
3	Mathematical Overview of ML Methods	5
3.1	Random Forest	5
3.2	Decision Tree	6
3.3	Support Vector Machine (SVM)	6
4	Model Fitting and Comparison	6
4.1	Data Splitting	6
4.2	Cross-Validation Setup	6
4.3	Random Forest Model	6
4.4	Decision Tree Model	8
4.5	Support Vector Machine Model	9
4.6	Model Comparison on Test Data	11
5	Interpretable Machine Learning (XAI)	14
5.1	Feature Importance	14
5.2	Partial Dependence Plots	15
5.3	Local Explanations (LIME)	15
5.4	SHAP Values (Shapley)	17
6	Conclusions	17
7	References	18

1 Introduction

1.1 Problem Statement

Maternal mortality remains a critical global health challenge, particularly in developing regions. Early identification of high-risk pregnancies enables timely medical interventions that can save lives. This project develops and compares machine learning models to predict maternal health risk as a **binary classification task** (High Risk vs. Not High Risk) based on vital health indicators collected during pregnancy.

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 a single “Not High Risk” category. This binary framing directly addresses the clinical question: *“Is this pregnancy high-risk and requiring immediate attention?”*

1.2 Dataset Description

The Maternal Health Risk dataset was collected from hospitals and community clinics in rural Bangladesh through an IoT-based risk monitoring system (Ahmed and Kashem 2023). The dataset contains health records that can be used to predict pregnancy risk levels.

Dataset Dimensions: 1014 observations, 7 variables

Table 1: Variable Description

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

2 Exploratory Data Analysis

2.1 Data Quality Check

Table 2: Missing Values Summary

	Variable	Missing	Percentage
Age	Age	0	0
SystolicBP	SystolicBP	0	0
DiastolicBP	DiastolicBP	0	0
BS	BS	0	0
BodyTemp	BodyTemp	0	0
HeartRate	HeartRate	0	0
RiskLevel	RiskLevel	0	0

The dataset has **no missing values**, eliminating the need for imputation strategies.

2.2 Outlier Detection and Handling

HeartRate range: 7 - 90

Observations with HeartRate < 30: 2

Outlier row indices: 500 909

Outlier HeartRate values: 7 7

Issue Identified: Two observations have HeartRate = 7 bpm, which is physiologically impossible (normal resting heart rate is 60-100 bpm). These are clearly data entry errors.

Decision: Remove these 2 observations (~0.2% of data) as they represent erroneous values, not extreme but valid measurements.

Original observations: 1014

Table 4: Descriptive Statistics by Risk Level

RiskLevel	n	Age_mean	Age_sd	SystolicBP_mean	DiastolicBP_mean	BS_mean	BodyTemp_mean	HeartRate_mean
NotHighRisk	740	27.6	12.9	109.1	73.3	7.48	98.58	73.6
HighRisk	272	36.2	13.0	124.2	85.1	12.12	98.90	76.7

After outlier removal: 1012

Observations removed: 2

2.3 Target Variable Distribution

Table 3: Binary Target Variable Distribution

RiskLevel	Count	Percentage
NotHighRisk	740	73.1
HighRisk	272	26.9

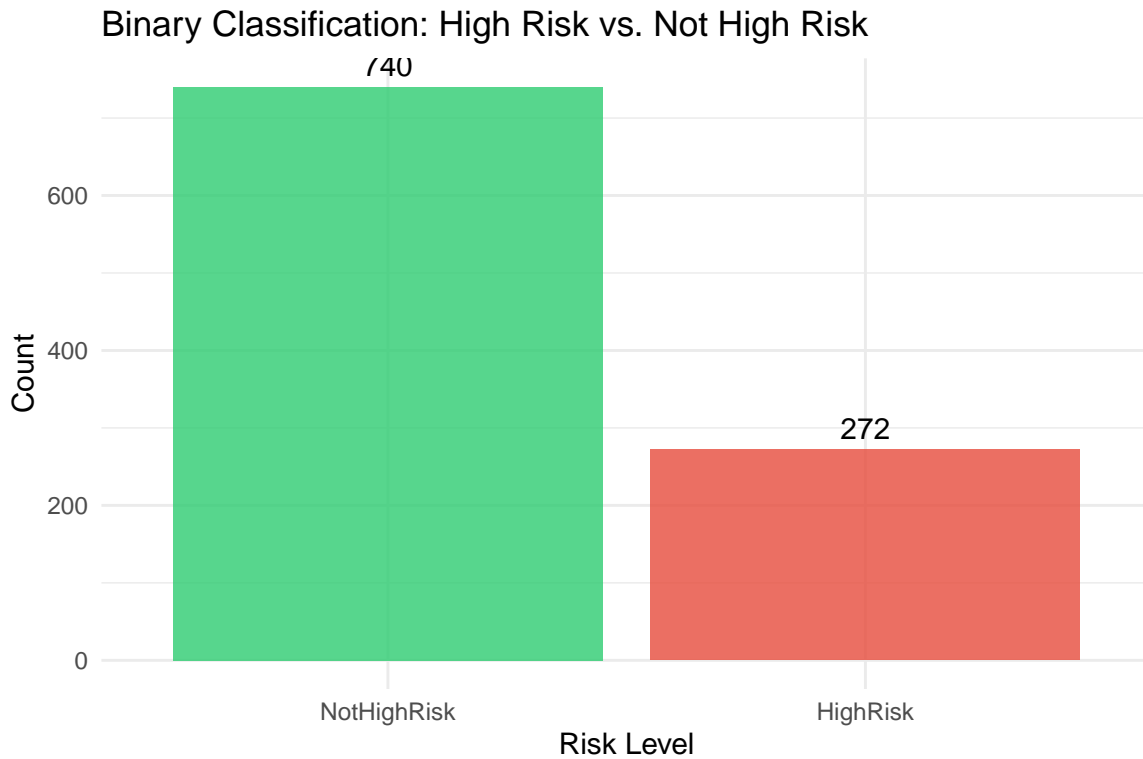


Figure 1: Distribution of Risk Levels (Binary)

The binary classification task has a class imbalance (~27% HighRisk, ~73% NotHighRisk), which we address using stratified sampling.

2.4 Descriptive Statistics

2.5 Distribution of Predictor Variables

Predictors by Risk Level (Binary)

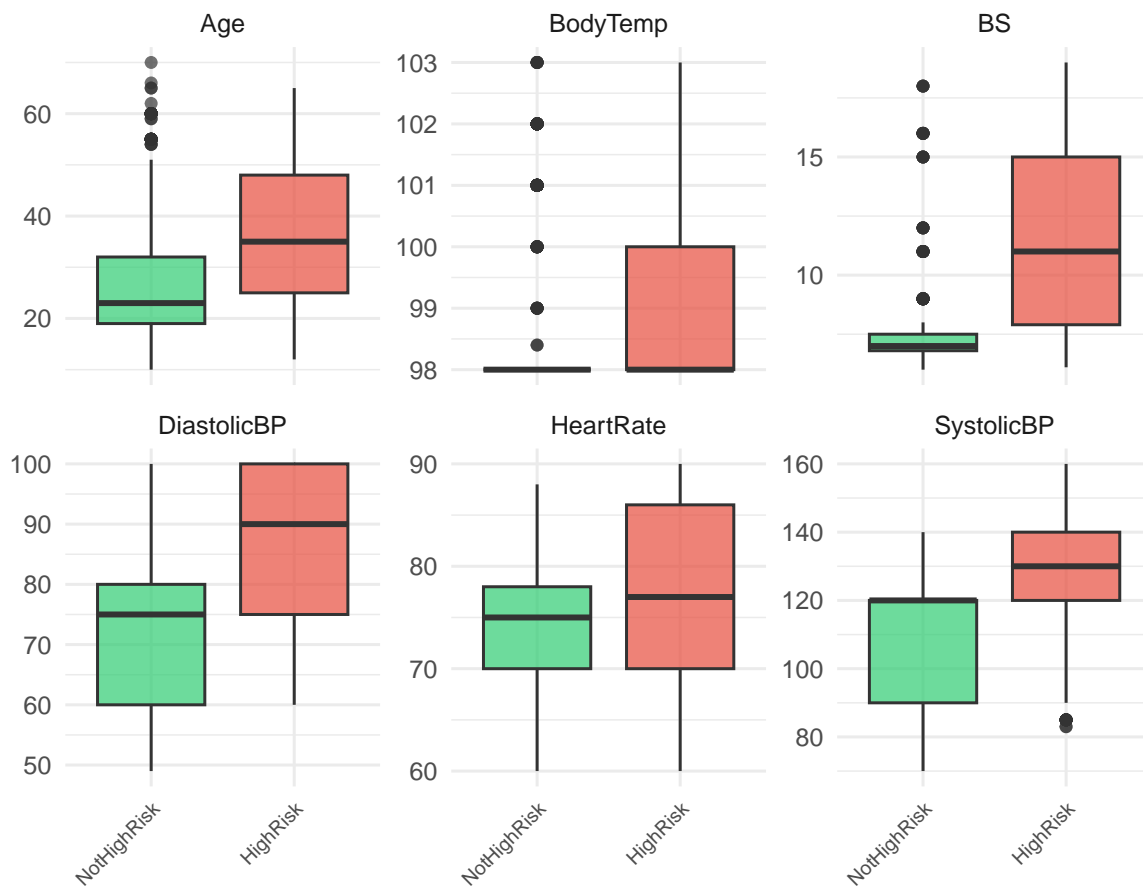


Figure 2: Box Plots of Predictor Variables by Risk Level

Key Observations:

- **Blood Sugar (BS):** Strong discriminator - high-risk patients have notably higher BS levels
- **Systolic BP:** High-risk patients tend to have elevated systolic blood pressure
- **Age:** Older patients show higher risk levels on average
- **Body Temperature:** Some high-risk cases show elevated temperature (fever)

2.6 Correlation Analysis

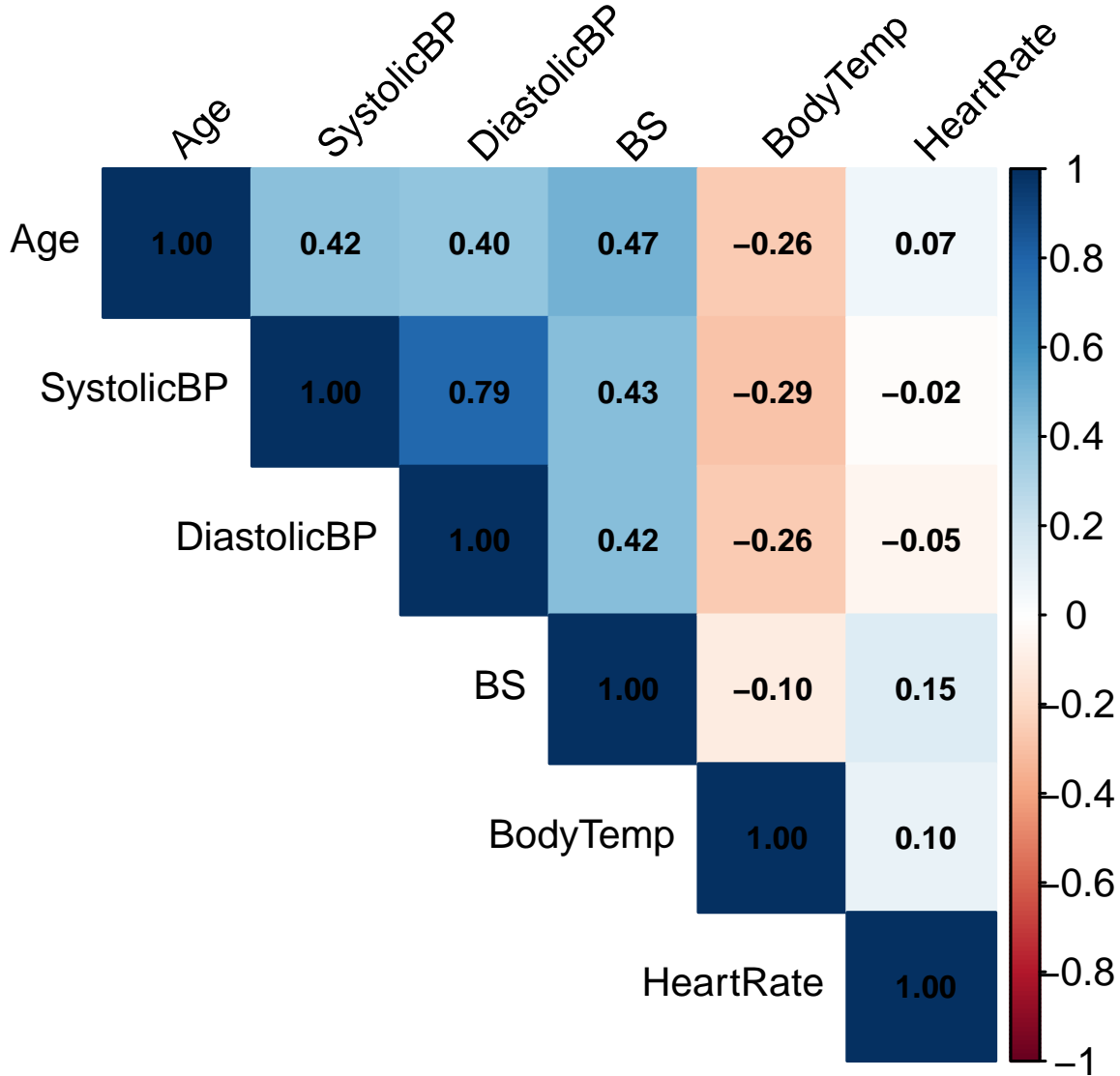


Figure 3: Correlation Matrix of Predictor Variables

Correlation Findings:

- Systolic and Diastolic BP show moderate positive correlation (expected physiologically)
- Most other predictors show weak correlations, suggesting independent information
- No severe multicollinearity issues detected

3 Mathematical Overview of ML Methods

3.1 Random Forest

Random Forest is an ensemble of decision trees using bootstrap aggregating (bagging) with feature randomization (Breiman 2001). Each tree is trained on a bootstrap sample, and at each split, only $m = \sqrt{p}$ random features are considered. Final predictions use majority voting: $\hat{y} = \text{mode}\{T_b(x)\}_{b=1}^B$.

Key Hyperparameters: `ntree` (number of trees), `mtry` (features per split), `nodesize` (minimum leaf size).

Gini Impurity measures split quality: $G(t) = 1 - \sum_{k=1}^K p_k^2$

3.2 Decision Tree

Decision Trees recursively partition the feature space using binary splits (James et al. 2021). At each node, the algorithm selects the feature and threshold that best separates the classes. For classification, splits are evaluated using **Gini Impurity** or **Information Gain**.

Gini Impurity: $G(t) = 1 - \sum_{k=1}^K p_k^2$ where p_k is the proportion of class k at node t .

Key Hyperparameters: `cp` (complexity parameter for pruning), `maxdepth` (maximum tree depth), `minsplit` (minimum observations for split), `minbucket` (minimum observations in leaf).

Advantages: Highly interpretable, handles non-linear relationships, no feature scaling required.

3.3 Support Vector Machine (SVM)

SVM finds the optimal hyperplane maximizing the margin between classes (James et al. 2021). The soft-margin formulation with cost parameter C handles non-separable data:

$$\min_{\beta, \beta_0, \xi} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \xi_i \quad \text{s.t.} \quad y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i$$

Kernel Trick enables non-linear boundaries. We use the RBF kernel: $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$

Hyperparameters: C (cost/regularization), γ (kernel width). Binary classification uses direct optimization.

4 Model Fitting and Comparison

4.1 Data Splitting

```
## Training set: 810 observations
## Test set: 202 observations
##
## Class proportions in training data:
##
## NotHighRisk    HighRisk
##          73.1      26.9
##
## Class proportions in test data:
##
## NotHighRisk    HighRisk
##          73.3      26.7
```

4.2 Cross-Validation Setup

4.3 Random Forest Model

4.3.1 Hyperparameter Tuning

```
## Random Forest
##
## 810 samples
## 6 predictor
## 2 classes: 'NotHighRisk', 'HighRisk'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 729, 729, 729, 728, 729, 730, ...
```

```
## Resampling results across tuning parameters:
##
##   mtry  ROC      Sens      Spec
##   2     0.9798942 0.9712712 0.8623377
##   3     0.9788973 0.9729661 0.8854978
##   4     0.9802444 0.9746610 0.8945887
##   5     0.9777552 0.9746610 0.8991342
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 4.
```

Random Forest: mtry Tuning

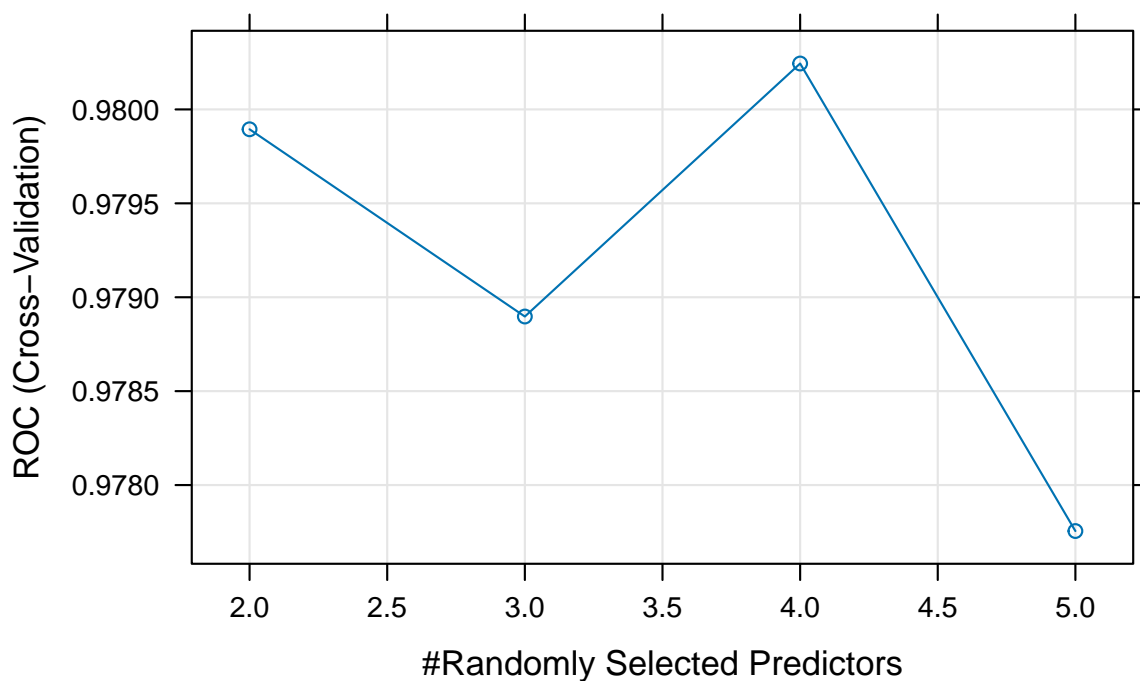


Figure 4: Random Forest Hyperparameter Tuning Results

4.3.2 Best Random Forest Model

```
## Best mtry: 4
## Best CV AUC-ROC: 0.9802
##
## Call:
## randomForest(x = x, y = y, ntree = 500, mtry = param$mtry, importance = TRUE)
##           Type of random forest: classification
##           Number of trees: 500
## No. of variables tried at each split: 4
##
##           OOB estimate of  error rate: 5.06%
## Confusion matrix:
##           NotHighRisk HighRisk class.error
## NotHighRisk      572      20 0.03378378
## HighRisk         21     197 0.09633028
```

4.4 Decision Tree Model

4.4.1 Hyperparameter Tuning

```
## CART
##
## 810 samples
## 6 predictor
## 2 classes: 'NotHighRisk', 'HighRisk'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 729, 729, 729, 728, 729, 730, ...
## Resampling results across tuning parameters:
##
##  cp      ROC      Sens      Spec
##  0.001  0.9493606  0.9459605  0.7937229
##  0.010  0.8978786  0.9426554  0.7800866
##  0.050  0.8725433  0.9156497  0.8209957
##  0.100  0.8725433  0.9156497  0.8209957
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.001.
```

Decision Tree: Complexity Parameter Tuning

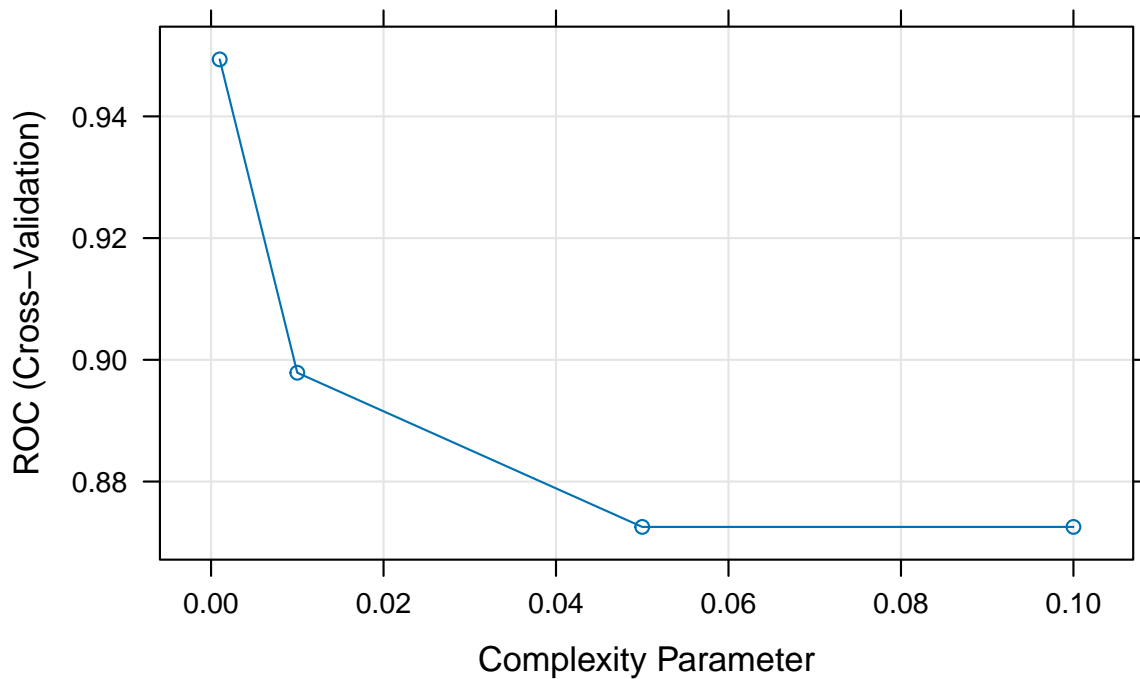


Figure 5: Decision Tree Hyperparameter Tuning Results

4.4.2 Best Decision Tree Model

```
## Best cp: 0.001
## Best CV AUC-ROC: 0.9494
```

4.4.3 Decision Tree Visualization

Decision Tree for Maternal Health Risk

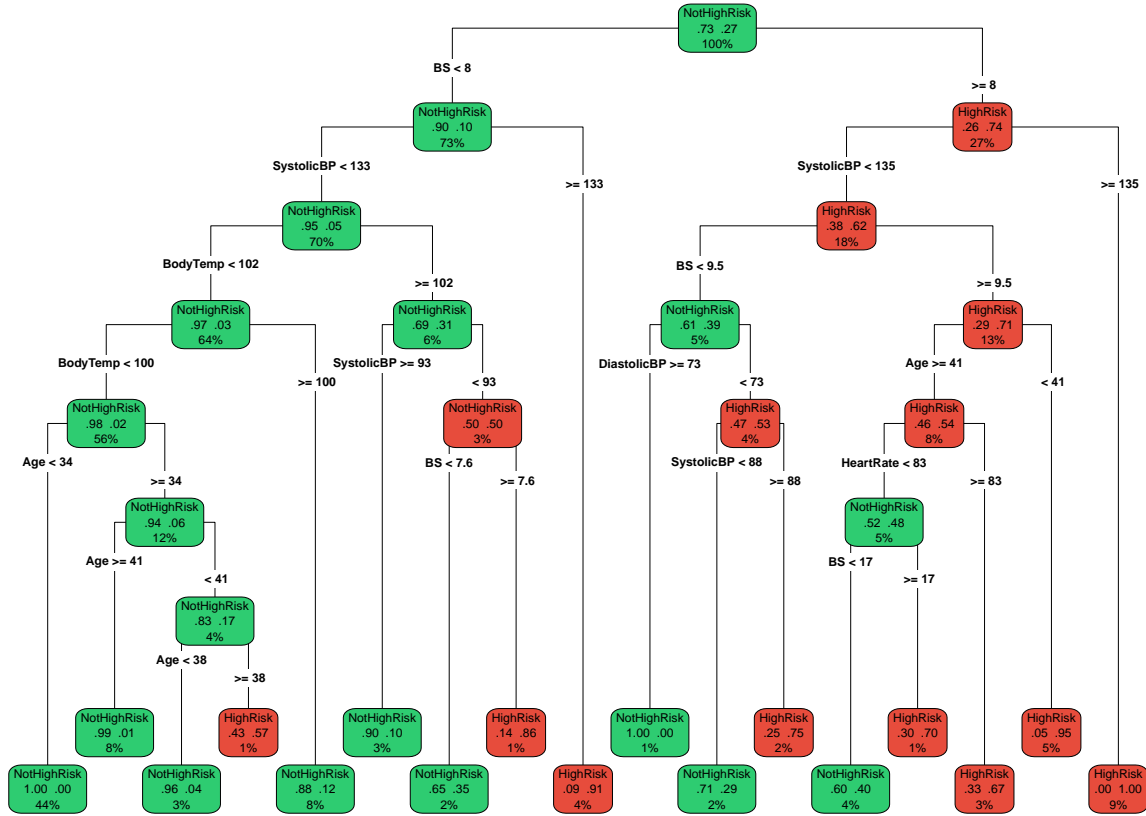


Figure 6: Decision Tree Structure

The decision tree provides a highly interpretable model. Each node shows the split condition, and the path from root to leaf represents the decision rules for classification.

4.5 Support Vector Machine Model

4.5.1 Feature Scaling

Scaling parameters (means):

##	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
##	29.98	112.76	76.23	8.76	98.68	74.77

##

Scaling parameters (std devs):

##	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
##	13.59	18.40	13.83	3.31	1.38	7.56

4.5.2 Kernel Comparison

Table 5: SVM Kernel Comparison (10-fold CV)

Kernel	Accuracy	Kappa
--------	----------	-------

linear	85.19	NA
radial	90.00	NA
polynomial	88.27	NA

```
##
## Best kernel: radial
```

4.5.3 Hyperparameter Tuning for RBF Kernel

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 810 samples
## 6 predictor
## 2 classes: 'NotHighRisk', 'HighRisk'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 729, 729, 729, 728, 729, 730, ...
## Resampling results across tuning parameters:
##
##  C      sigma  ROC      Sens      Spec
##  0.1  0.01   0.8936576  0.9493785  0.6142857
##  0.1  0.10   0.9239888  0.9493785  0.6831169
##  0.1  0.50   0.9252279  0.9307910  0.7978355
##  0.1  1.00   0.9448563  0.9307627  0.8344156
##  1.0  0.01   0.9213656  0.9493785  0.5824675
##  1.0  0.10   0.9357619  0.9375424  0.7744589
##  1.0  0.50   0.9365550  0.9493503  0.7701299
##  1.0  1.00   0.9354076  0.9560734  0.7564935
## 10.0  0.01   0.9294244  0.9510734  0.7242424
## 10.0  0.10   0.9524572  0.9493785  0.8028139
## 10.0  0.50   0.9416746  0.9560734  0.8028139
## 10.0  1.00   0.9358371  0.9560734  0.7889610
## 100.0 0.01   0.9359744  0.9459887  0.7653680
## 100.0 0.10   0.9485955  0.9560734  0.8071429
## 100.0 0.50   0.9531418  0.9611299  0.8350649
## 100.0 1.00   0.9485764  0.9627966  0.8303030
##
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.5 and C = 100.
```

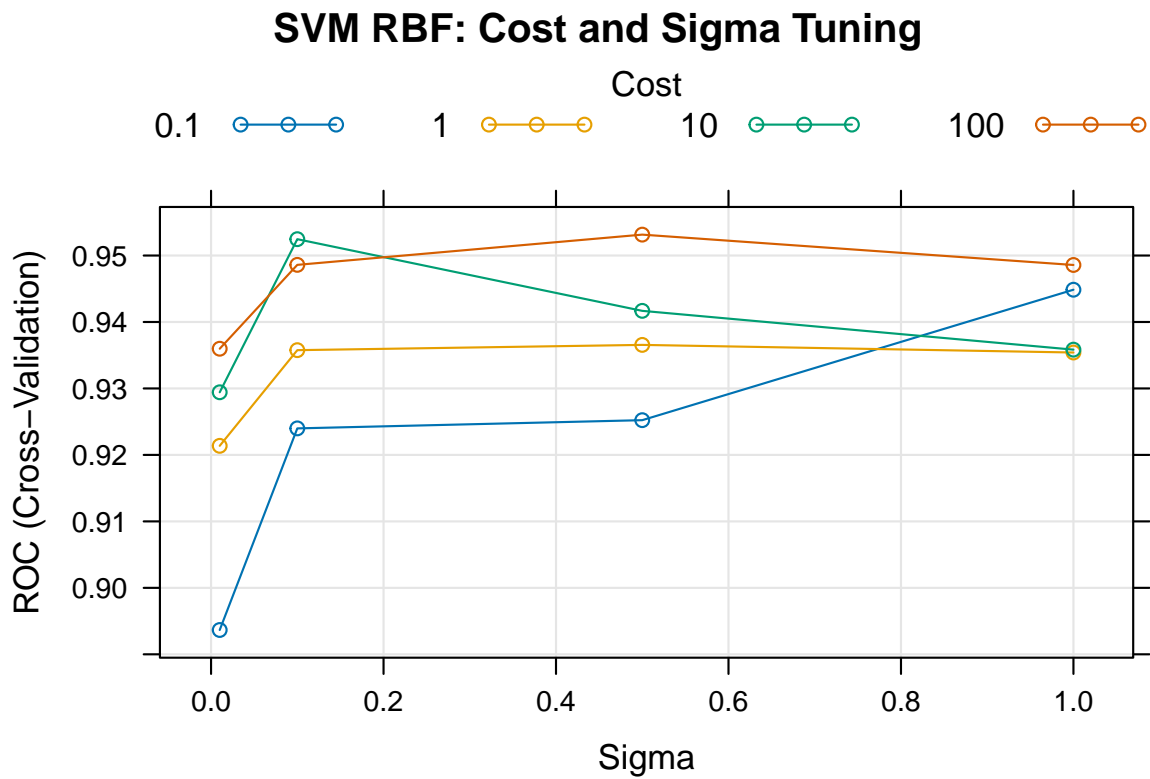


Figure 7: SVM Hyperparameter Tuning Results

4.5.4 Best SVM Model

Best C (Cost): 100

Best sigma: 0.5

Best CV AUC-ROC: 0.9531

4.6 Model Comparison on Test Data

4.6.1 Predictions

4.6.2 Confusion Matrices

=== RANDOM FOREST ===

		Reference	
## Prediction		NotHighRisk	HighRisk
## NotHighRisk		145	8
## HighRisk		3	46

##

=== DECISION TREE ===

		Reference	
## Prediction		NotHighRisk	HighRisk
## NotHighRisk		139	10
## HighRisk		9	44

##

=== SUPPORT VECTOR MACHINE ===

		Reference	
## Prediction		NotHighRisk	HighRisk

##	NotHighRisk	143	13
##	HighRisk	5	41

4.6.3 Performance Metrics Comparison

Table 6: Binary Classification Metrics on Test Data

	Metric	RandomForest	DecisionTree	SVM
Accuracy	Accuracy	94.550	90.590	91.090
Kappa	Kappa	0.857	0.758	0.761
Sensitivity	Sensitivity (Recall)	97.970	93.920	96.620
Specificity	Specificity	85.190	81.480	75.930
Pos Pred Value	Precision (PPV)	94.770	93.290	91.670
Neg Pred Value	NPV	93.880	83.020	89.130
F1	F1 Score	96.350	93.600	94.080

4.6.4 Visual Comparison

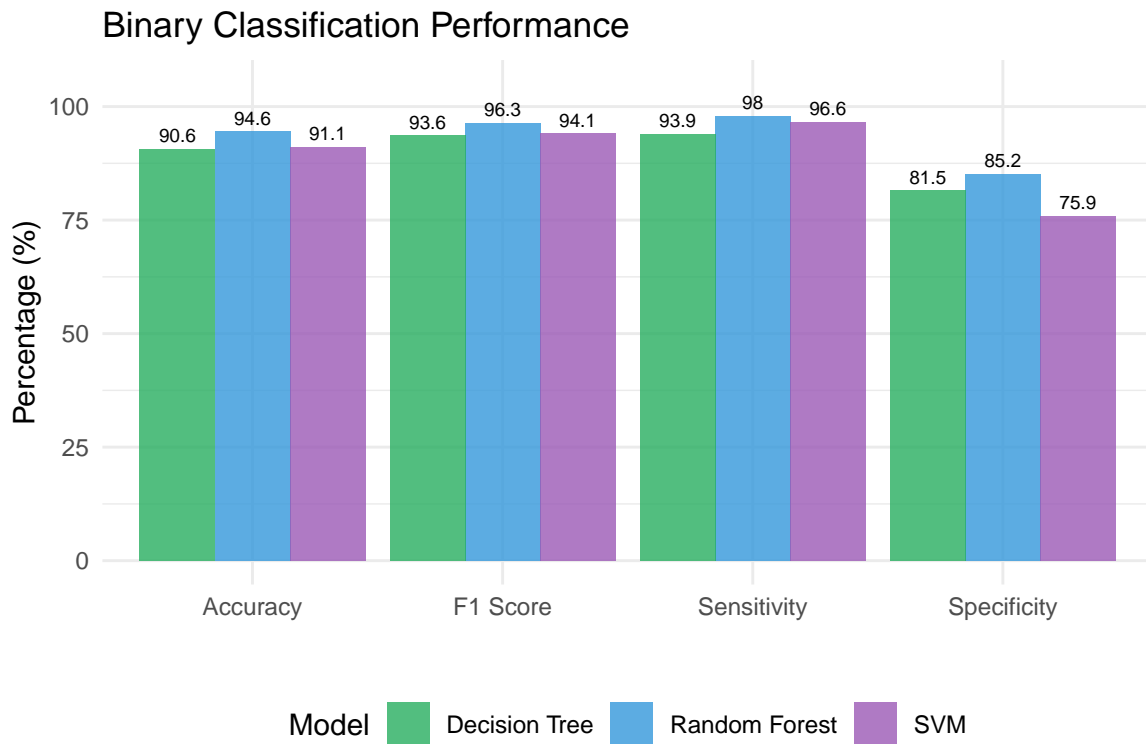


Figure 8: Model Performance Comparison

4.6.5 ROC Curve Analysis

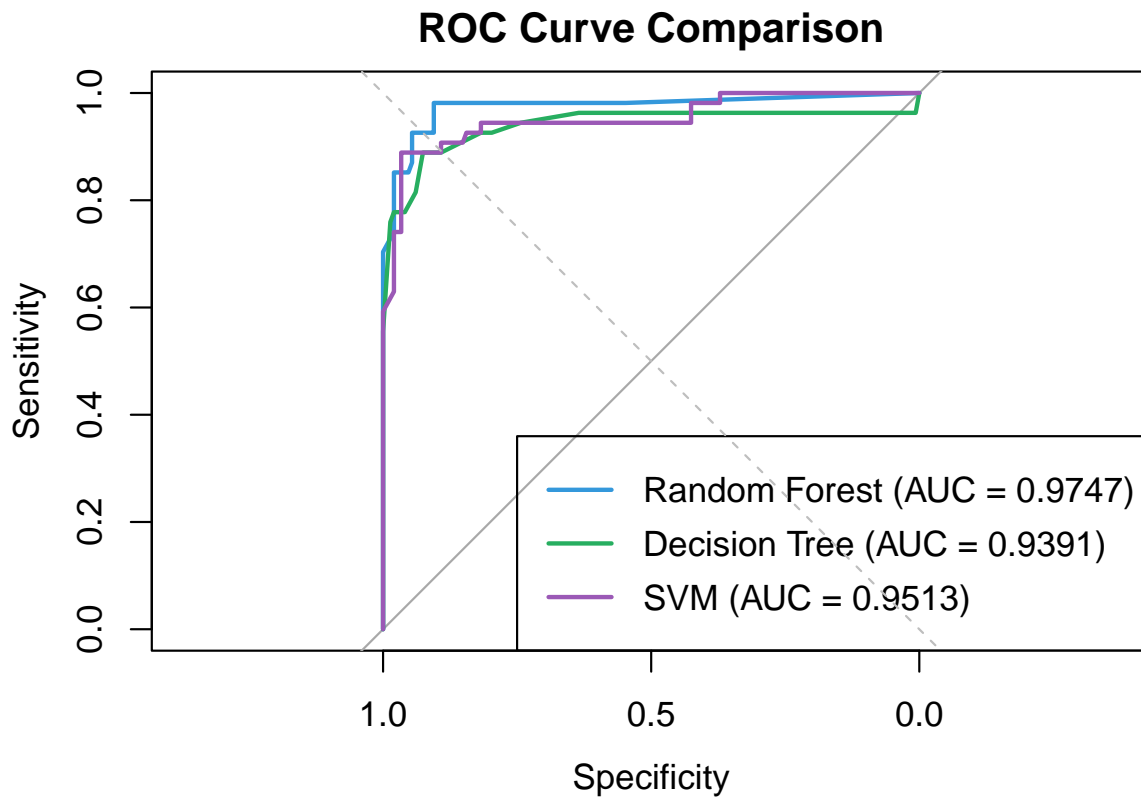


Figure 9: ROC Curves for Model Comparison

```
##  
## AUC-ROC Comparison:  
## Random Forest AUC: 0.9747  
## Decision Tree AUC: 0.9391  
## SVM AUC: 0.9513
```

5 Interpretable Machine Learning (XAI)

5.1 Feature Importance

5.1.1 Random Forest Feature Importance

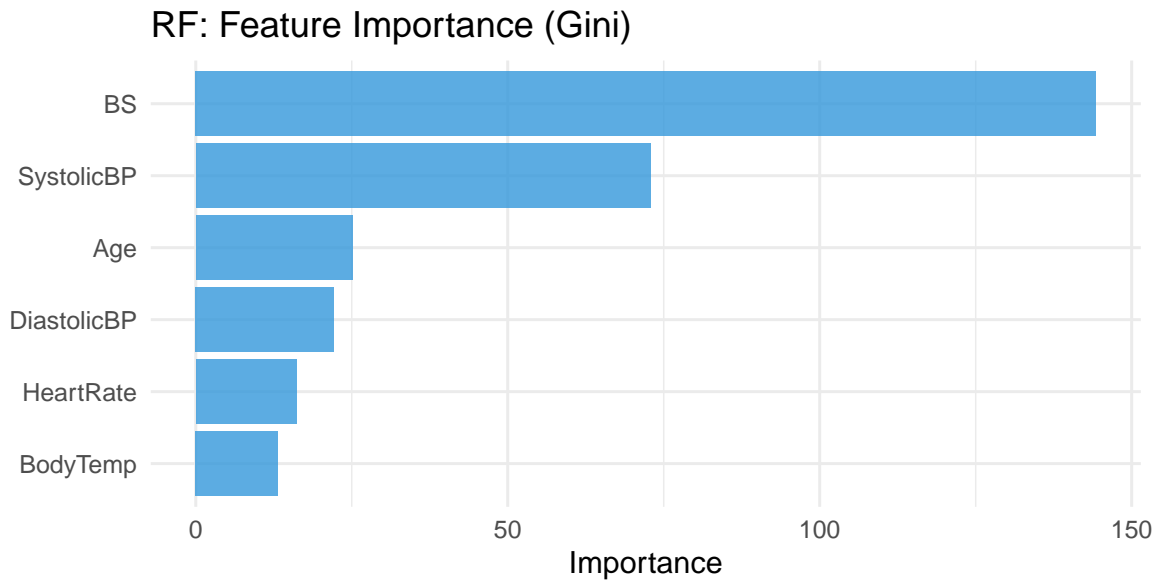


Figure 10: Random Forest Feature Importance

5.1.2 SVM Feature Importance

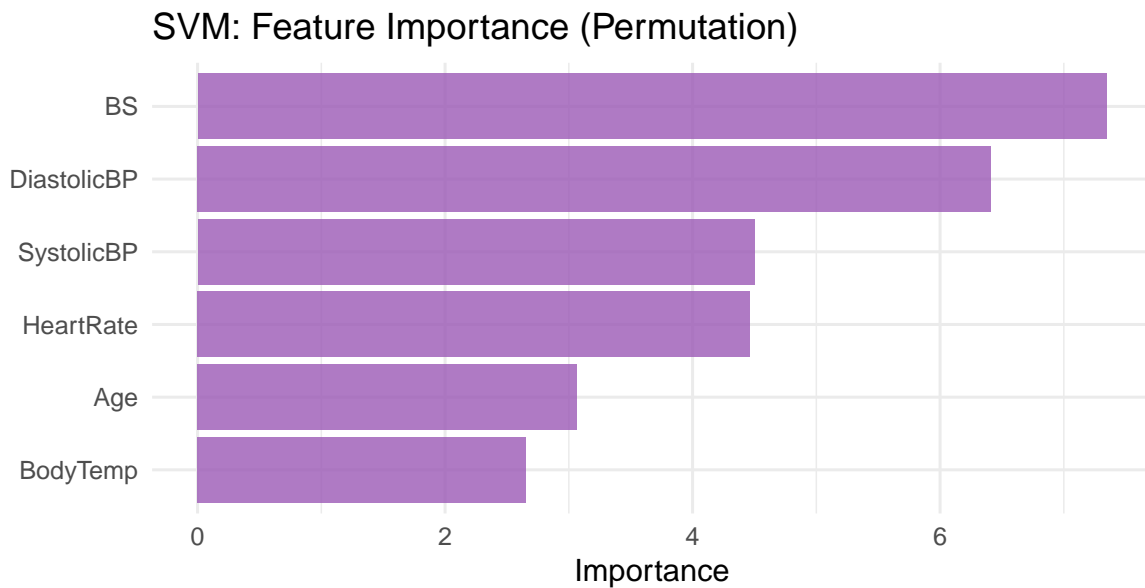


Figure 11: SVM Permutation Feature Importance

Both models rank **Blood Sugar (BS)** as the most important feature, followed by **SystolicBP** and **Age**. This consistency across different model types strengthens our confidence in these findings.

5.2 Partial Dependence Plots

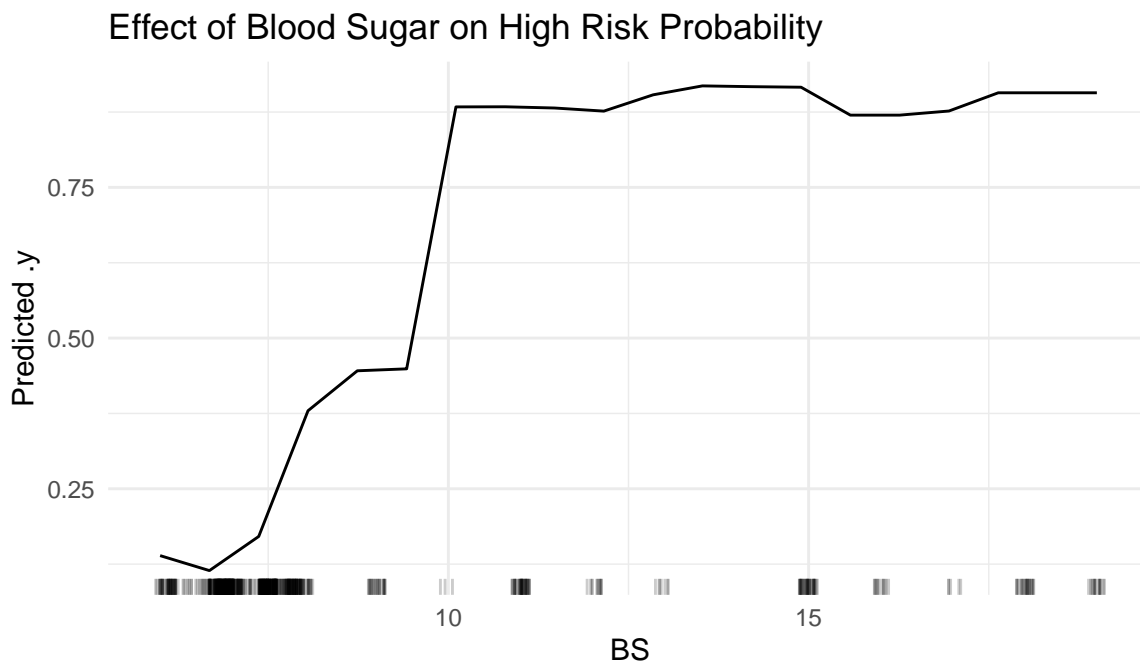


Figure 12: PDP: Blood Sugar Effect on High Risk Probability

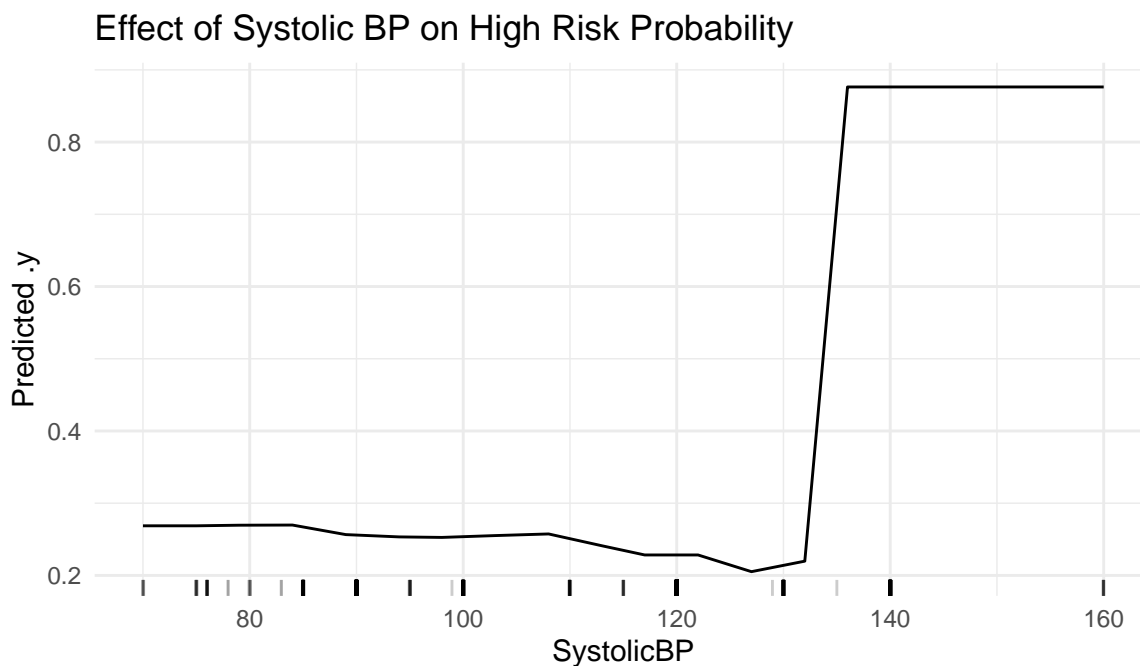


Figure 13: PDP: Systolic BP Effect on High Risk Probability

5.3 Local Explanations (LIME)

5.3.1 Setup LIME Explainer

5.3.2 Select Cases for Explanation

Selected test case indices: 1 23 2 30

```
## Actual labels: HighRisk HighRisk NotHighRisk HighRisk
## Predicted labels: HighRisk HighRisk NotHighRisk NotHighRisk
```

5.3.3 LIME Explanations

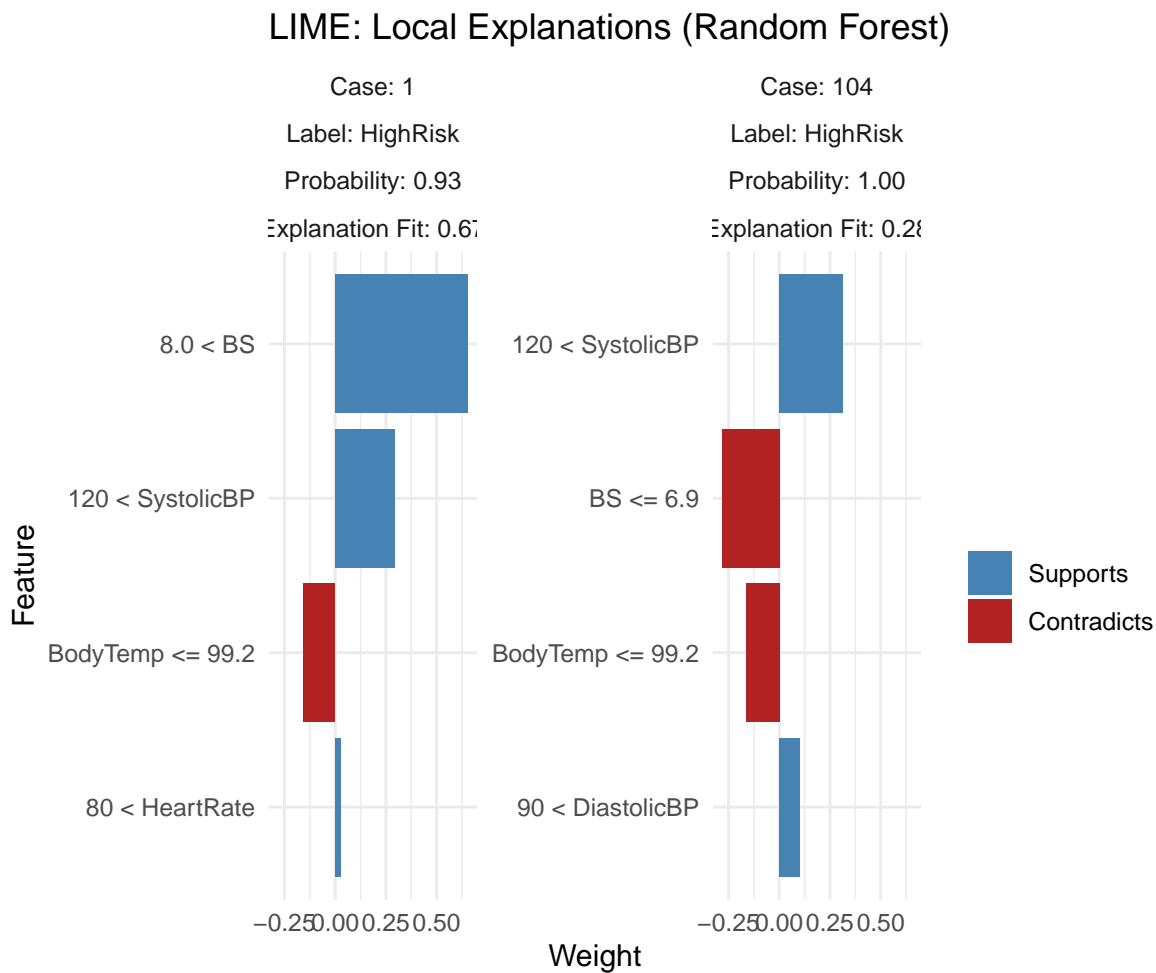


Figure 14: LIME Explanations for Selected Test Cases

5.4 SHAP Values (Shapley)

5.4.1 SHAP for Individual Predictions

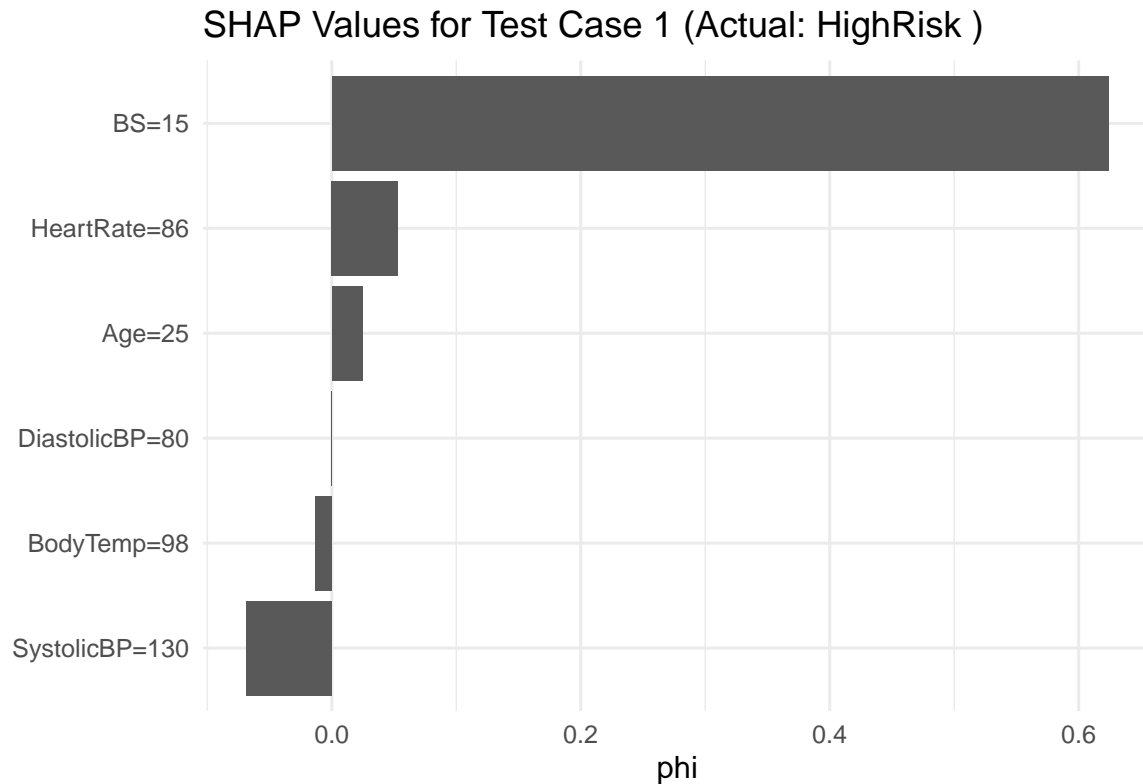


Figure 15: SHAP Values for a High-Risk Case

6 Conclusions

Summary: We framed maternal health risk prediction as a **binary classification task** (High Risk vs. Not High Risk), addressing the ordinal nature of the original 3-class labels. The dataset required minimal preprocessing (2 outliers removed). We compared three machine learning models: Random Forest, Decision Tree, and SVM.

Key Findings:

- All three models achieve strong performance in detecting high-risk pregnancies
- Blood sugar (BS) and systolic blood pressure are the strongest predictors across all models
- Decision Tree provides the most interpretable model with explicit decision rules
- Random Forest offers the best overall performance as an ensemble method
- SVM achieves competitive results with proper hyperparameter tuning

Key XAI Insights:

- Higher blood sugar strongly increases high-risk probability (PDP shows clear threshold)
- Systolic BP above 130-140 mmHg indicates elevated risk
- Decision Tree visualization provides clear clinical decision pathways
- LIME explanations provide case-by-case interpretability for clinical review

Clinical Recommendations:

- Use sensitivity as the primary metric (minimizing false negatives for patient safety)
- Blood sugar and blood pressure should be closely monitored during pregnancy
- Decision Tree can serve as an interpretable screening tool for non-technical staff
- Random Forest/SVM can be used for more accurate predictions when interpretability is less critical

Limitations: Dataset size (~1,000 observations), geographic scope (rural Bangladesh), limited feature set (6 predictors).

7 References

- Ahmed, Marzia, and Mohammad 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. Springer.