Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Explainable Random Forest Model for Early Hypertension Risk Prediction: Integrating SHAP and Breakdown Analysis

Heshan Chandeepa Pathmakumara

bhcpathmakumara@students.nsbm.ac.lk

NSBM Green University https://orcid.org/0009-0000-5057-6102

Research Article

Keywords: Hypertension prediction, Random Forest, machine learning, explainable AI, SHAP values, risk stratification, clinical decision support, AUC ROC

Posted Date: May 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9585224/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Hypertension remains a leading global cause of cardiovascular mortality, but it often goes undiagnosed due to its silent progression. This study developed and evaluated an explanatory random forest model for the early prediction of hypertension risk using clinical and demographic data. A publicly available dataset of 4,240 medical records with 12 predictor variables (age, gender, blood pressure, body mass index, cholesterol, smoking status, glucose, heart rate, and others) was analyzed. After preprocessing the median imputation of missing values and removing predictors with near-zero variance, the data were divided into 70% training and 30% test sets. A random forest classifier was trained with 5-fold cross-validation to tune the mtry hyperparameter from 2 to 12, optimizing for the area under the receiver operating characteristic curve (AUC). The final model achieved an AUC value of 0.9365, an accuracy of 88.99% (95% CI: 87.13%-90.65%), a sensitivity of 79.45%, a specificity of 93.91%, and an F1 score of 0.8309, with the best mtry value being 2. Systolic and diastolic blood pressure were the most powerful predictors, followed by age, BMI, and heart rate. Model interpretability was improved using SHAP (SHapley Additive Explanations) values and DALEX breakdown plots, providing transparent, single-level explanations. While high specificity indicates excellent ability to exclude non-risk individuals, moderate sensitivity suggests room for improvement in identifying true hypertension cases. This study demonstrates that a tuned random forest model combined with explainable AI can serve as an effective, interpretable screening tool for hypertension risk stratification, data-driven clinical decisions, and early intervention strategies.

I. INTRODUCTION

Persistent high blood pressure, or hypertension, continues to be a leading cause of cardiovascular disease and death globally. The World Health Organization estimates that 1.28 billion people between the ages of 30 and 79 have high blood pressure, and about half of them are unaware that they have it [1]. Although a significant risk factor for stroke, heart attack, myocardial infarction, and chronic kidney disease, the disease often causes no symptoms, earning it the nickname “the silent killer” [2], [3]. Therefore, early detection and effective risk stratification are essential to reduce the growing burden of hypertension-related complications.

Traditional risk assessment relies on laboratory tests and clinical data (such as blood pressure, age, and body mass index), but these methods may not adequately account for complex, nonlinear interactions between risk variables such as comorbidities, genetics, and lifestyle [4]. By finding patterns in high-dimensional data and developing predictive models that aid in clinical decision-making, machine learning (ML) offers a viable alternative [5]. Random Forest (RF), a supervised learning algorithm, has attracted great interest in healthcare analytics due to its inherent feature importance ranking, resilience to overfitting, and ability to handle mixed data types (numeric and categorical) [6], [7].

Several studies have been conducted using RF for cardiovascular risk prediction. Using data from the Bangladesh Demographic and Health Survey, Islam et al. [8] used RF to predict hypertension with 93% accuracy, and found that overweight and older age were important predictors [8]. Explainable artificial intelligence (XAI) and machine learning models have recently been used to overcome the “black box” problem of complex algorithms. By using techniques such as breakdown profiles and SHAP (SHapley Additive Explanations), clinicians can increase confidence and facilitate adoption in real-world settings, helping them understand how individual features contribute to a prognosis [9], [10], [11].

Despite these developments, a large number of current predictive models are not sufficiently interpretable for clinical use or lack external validation. In addition, class imbalance in medical datasets can bias models toward the dominant class, reducing their sensitivity in detecting true cases of hypertension. Closing these gaps requires a methodological strategy that includes data preprocessing, fine-tuning, thorough performance evaluation, and clear justification of model choices.

The primary objective of this study is to develop and evaluate a random forest model for binary classification of hypertension risk using a publicly available dataset of 4,240 medical records. The specific objectives are: (1) pre-process the data by handling missing values and removing null variance predictors, (2) tune the RF hyperparameter mtry (the number of features sampled at each split) from 2 to 12 using 5-fold cross-validation, (3) assess model performance using accuracy, sensitivity, specificity, precision, F1 score, and AUC-ROC, and (4) improve model interpretation by applying

TABLE I: Description of Predictor Variables

Variable	Description	Type	Unit / Values
male	Gender	Binary	0 = female, 1 = male
age	Age	Continuous	Years
currentSmoker	Smoking status	Binary	0 = non-smoker, 1 = smoker
cigsPerDay	Cigarettes per day	Continuous	Number
BPMeds	Blood pressure medication	Binary	0 = no, 1 = yes
diabetes	Diabetes status	Binary	0 = no, 1 = yes
totChol	Total cholesterol	Continuous	mg/dL
sysBP	Systolic blood pressure	Continuous	mmHg
diaBP	Diastolic blood pressure	Continuous	mmHg
BMI	Body mass index	Continuous	kg/m²
heartRate	Heart rate	Continuous	beats/min
glucose	Glucose level	Continuous	mg/dL
Risk	Target variable	Binary	0 = no hypertension risk, 1 = at risk

SHAP and crash plots. We hypothesize that the random forest model will achieve high discriminant ability (AUC > 0.90) and that blood pressure-related features will be the most important predictors. The findings are expected to demonstrate the potential of ML-based tools for early hypertension screening and support data-driven intervention strategies in clinical practice.

II. RESEARCH MATERIALS AND METHODOLOGY

This section describes the dataset, preprocessing steps, model development, hyperparameter tuning, evaluation metrics and explanatory techniques used in this study. All analyses were performed using R version 4.5 with the packages tidyverse [12], caret [13], randomForest [14], pROC [15], DALEX [16], fastshap, corrplot. and ggplot2 [12].

A. Dataset

We used a publicly available hypertension risk dataset obtained from Kaggle [17]. The dataset includes 4,240 medical records with 13 variables: 12 predictor characteristics and a binary target variable (risk). Table I summarizes the predictor variables, their descriptions, data types and units of measurement.

The distribution of the target variable was unbalanced, with 1,317 individuals (31.1%) classified as “at risk” and 2,923 individuals (68.9%) classified as “not at risk.” Sensitivity and specificity were reported during evaluation to account for this discrepancy.

B. Data Preprocessing

Raw medical data often contains missing values and redundant features. We applied a structured preprocessing pipeline:

Missing values handling

There were incomplete records in the columns cigsPerDay (29 missing), BPMeds (53 missing), totChol (50 missing), BMI (19 missing), heart rate (1 missing), and glucose (388 missing). For continuous variables, missing values were imputed using the mean of the observed values, a robust method that preserves the central tendency from being affected by outliers [18]. For binary variables (male, current smoker, BPMeds, diabetes), we imputed using the mode (most frequent value). This approach resulted in a complete data set with no remaining missing entries.

Zero-variance predictor removal

Variables with very low variance contribute little to the prediction models and can increase computational cost. Using the nearZeroVar function from Carat, we identified predictors with frequency ratios > 30 and percent unique values < 1%. Two variables, BPMeds (frequency ratio 33.19) and Diabetes (frequency ratio 37.90), met these criteria and were removed. This step also reduces the risk of overfitting.

Encoding and class conversion

To facilitate classification, the target variable Risk was transformed into a factor with levels “No” and “Yes.” Every predictor variable continued to be numerical.

C. Exploratory Data Analysis

Before modeling, exploratory data analysis (EDA) was performed to understand the data structure and relationships.

The distribution of the target variable is shown in Fig. 1. Approximately 69% of patients are classified as “no risk” and 31% as “yes risk,” and the bar graph validates the class heterogeneity. This visualization highlights the need for a thorough evaluation using class-sensitive measures.

Figure 2 presents the correlation matrix of all numeric predictors using a colour-coded upper triangular matrix.

Side-by-side box plots for each predictor stratified by risk status are shown in Fig. 3. Predictably, median values for continuous variables including systBP, diaBP, age and BMI were consistently higher in the “yes risk” group than in the “no risk” group. The clinical significance of the selected features was confirmed by these ocular examinations.

D. Data Partitioning

We used stratified random sampling to divide the data into training and test sets to assess the generalizability of the model. Thanks to the stratification, the percentage of “no” and “yes” cases in both subgroups was maintained. We assigned 70% of the data (n = 2,969) to the training set and the remaining 30% (n = 1,271) to the test set using the caret createDataPartition function. There were 2,047 “no” and “yes” cases in the training set and 876 “no” and “yes” cases in the test set. This division is consistent with standard machine learning procedures [19].

E. Random Forest Model

An ensemble learning technique called random forest (RF) builds multiple decision trees during training and uses majority voting for classification [6]. Each tree is built using a bootstrap sample of the data, and a random selection of features is taken into account for the split at each node. By pruning the trees, this unpredictability reduces variance and overfitting.

Three factors led us to use RF: (1) it handles nonlinear relationships and interactions without requiring significant feature engineering; (2) it has built-in feature importance measures; and (3) it is robust to noisy data and outliers, which are common in medical datasets [20].

F. Hyperparameter Tuning

The number of features randomly selected as candidates in each split, or mtry, is the primary hyperparameter of RF. Small values of mtry promote randomness and reduce bias, but can increase variance, while large values increase tree similarity and lead to overfitting [21]. As usual, we used 5-fold cross-validation on the training set to adjust mtry from 2 to 12.

The trainControl function in caret was used to conduct cross-validation (method = “cv,” number = 5). Four-fifths of the training data were used to train the model for each fold, and the remaining one-fifth was used for evaluation. The area under the ROC curve (AUC) served as the performance metric. The optimal mtry value was used to train the final model on the entire training set. The number of trees (ntree) was set at 500, which is adequate for the out-of-bag error to converge [22].

Figure 4 shows the cross-validated ROC (AUC) as a function of mtry. The highest AUC (0.9507) was obtained at mtry = 2, and the performance gradually decreased as mtry increased. Therefore, mtry = 2 was chosen as the optimal hyperparameter.

G. Model Evaluation Metrics

Using the following metrics, we evaluated the final model on the retained test set [23].

- **Accuracy**: (TP + TN) / (TP + TN + FP + FN)
- **Sensitivity (Recall)**: TP / (TP + FN) – ability to identify true hypertension cases
- **Specificity**: TN / (TN + FP) – ability to identify true non-risk cases
- **Precision (PPV)**: TP / (TP + FP) – proportion of predicted positives that are correct
- **Negative Predictive Value (NPV)**: TN / (TN + FN)
- **F1 Score**: 2 × (Precision × Sensitivity) / (Precision + Sensitivity) – harmonic mean of precision and sensitivity

- **AUC-ROC**: area under the receiver operating characteristic curve, measuring discriminative ability across all thresholds [24]

The binomial precision method was used to calculate the accuracy confidence intervals.

The confusion matrix is shown in Fig. 5. With 89 false positives and 51 false negatives, the model correctly identified 344 true positives and 787 true negatives. This corresponds to a specificity of 93.9% and a sensitivity of 79.5%.

The ROC curve is shown in Fig. 6. The excellent discrimination performance was demonstrated by an AUC of 0.9365 (95% CI not calculated). The ability of the model to distinguish between hypertensive and normal patients is confirmed by the sharp rise of the curve towards the upper left corner.

H. Out-of-Bag Error and Convergence

Out-of-bag (OOB) error is an internal error estimate provided by random forests. For each tree, about one-third of the training data is excluded (the "out-of-bag" sample). The predictions on these OOB samples are combined and the error rate is calculated, without the need for a separate validation set [25]. To evaluate convergence, we observed the OOB error as a function of the number of trees.

Figure 7 shows the OOB error curves for the total error, the no risk error, and the yes risk error. The total error decreased sharply from 0.35 to less than 0.12 in the first 50 trees and stabilized at around 0.10 after approximately 150 trees. The error in the yes risk class (green line) was higher (≈0.13–0.14) than in the no risk class (red line, ≈0.08), reflecting class imbalance. The convergence of 150–200 trees justifies our choice of 500 trees.

I. Feature Importance

The varImp function from Carrot was used to extract variable importance from the final random forest model. The average drop in fire filth across all trees is used to determine the importance score for binary classification [26]. Higher scores indicate greater contribution to the prediction.

The top ten features are listed in order of relevance in Fig. 8. The most significant predictor was systolic blood pressure (sysBP) (score 78.13), followed by diastolic blood pressure (diaBP, 52.66), age (13.16), BMI (9.65), and heart rate (6.21). These findings support the clinical understanding that hypertension is the major risk factor for hypertension.

J. Explainable Artificial Intelligence (XAI)

To enhance interpretability and trust in the model, we applied two complementary XAI methods.

Breakdown plots (DALEX): Model-agnostic breakdown profiles are implemented by the DALEX package [16]. Starting with the model's intercept (average prediction), a breakdown plot for a single observation illustrates how each attribute contributed to the final forecast. Negative contributions lower the projected risk, while positive contributions raise it.

A breakdown map for a typical high-risk patient (true risk = "yes") is shown in Fig. 9. The intercept (baseline probability) was 0.312. BMI contributed + 0.017, diastolic blood pressure (110 mmHg) contributed + 0.156, and systolic blood pressure (180 mmHg) contributed + 0.438. On the other hand, a glucose level of 99 mg/dL had a small negative effect (− 0.017). The patient was correctly classified as high risk based on a cumulative prediction of 0.94.

SHAP values (fastshap): SHApley additive explanations (SHAP) allocate the predicted variance to each feature on average based on cooperative game theory [9]. We calculated SHAP values using the fastshap package with 25 Monte Carlo simulations on a random sample of 200 test observations. The average absolute SHAP value per feature indicates global importance, while the bee plots visualize the direction and magnitude of the effects for individual observations.

The SHAP bar chart is presented in Fig. 10, showing the mean absolute SHAP values for each feature. The order largely confirms the fire-based importance: sysBP, diaBP, age, and BMI are the top contributors.

Figure 11 shows the SHAP heatmap of bees. Each point represents a single observation. The color indicates the trait value (blue = low, red = high). For sysBP and diaBP, higher values (red) are consistently associated with positive SHAP values (higher risk), while lower values (blue) push the predictions down. Age shows a more complex pattern: both very low and very high ages can increase risk. This visualization provides a nuanced understanding of how each trait affects predictions across the population.

III. CONCLUSIONS AND RECOMMENDATIONS

This study developed and evaluated a random forest model to predict hypertension risk using a publicly available dataset of 4,240 medical records. The model achieved strong discriminant ability with an AUC of 0.9365 and an overall accuracy of 88.99% (95% CI: 87.13%–90.65%). The optimal hyperparameter was mtry = 2, consistent with previous findings that a small number of randomly selected features in each split maximizes performance.

Systolic blood pressure (sysBP) and diastolic blood pressure (diaBP) emerged as the most important predictors, followed by age, body mass index, and heart rate. This is consistent with the clinical knowledge that hypertension is the primary modifiable risk factor for hypertension. The model showed high specificity (93.91%), with excellent ability to correctly identify non-at-risk individuals, which is clinically valuable in avoiding unnecessary follow-up testing and patient anxiety. However, sensitivity was low (79.45%), meaning that approximately one in five true hypertensive cases were missed. This trade-off reflects the class imbalance in the dataset (68.9% non-at-risk) and suggests that further optimization is needed to improve true positive detection.

Based on these findings, we recommend the use of random forest as an adjunct testing tool for hypertension risk stratification, especially in settings where disease exclusion is a priority. The improved interpretability of the model through SHAP and breakdown plots makes it suitable for integration into clinical decision support systems, allowing practitioners to understand why a patient has been flagged as high risk.

For future work, we recommend: (1) applying class balancing techniques such as SMOTE to improve sensitivity; (2) external validation on diverse, multi-center datasets to assess generalizability; (3) exploring other algorithms for comparison (e.g., XGBoost, neural networks); and (4) deploying the model as a web-based or mobile application for real-time risk assessment in primary care settings. Policymakers should support capacity-building initiatives in data science and artificial intelligence to enable the widespread use of evidence-based predictive analytics in public health.

IV. ETHICS STATEMENT

This study is a retrospective observational analysis using a publicly available anonymized dataset and does not constitute a clinical trial.

V. REFERENCES

1. Global report on hypertension the race against a silent killer. Accessed: Apr. 16, 2026. [Online]. Available: https://www.who.int/publications/i/item/9789240081062
2. Oparil S et al (2018) Hypertension, *Nat Rev Dis Primers*, vol. 4, no. 1, p. 18014, Mar. 10.1038/nrdp.2018.14
3. Iqbal AM, Jamal SF Essential Hypertension, in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2026. Accessed: Apr. 17, 2026. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK539859/
4. Charchar FJ et al (2024) Lifestyle management of hypertension: International Society of Hypertension position paper endorsed by the World Hypertension League and European Society of Hypertension, *J Hypertens*, vol. 42, no. 1, pp. 23–49, Jan. 10.1097/HJH.0000000000003563
5. Explainable Artificial Intelligence Models for Predicting Malaria Risk in Kenya Accessed: Apr. 17, 2026. [Online]. Available: https://www.researchgate.net/publication/390815411_Explainable_Artificial_Intelligence_Models_for_Predicting_Malaria_Risk_in_Kenya
6. Breiman L (2001) Random Forests, *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 10.1023/A:1010933404324
7. Yazici İ, Shayea I, Din J (Aug. 2023) A survey of applications of artificial intelligence and machine learning in future mobile networks-enabled systems. Eng Sci Technol Int J 44:101455. 10.1016/j.jestch.2023.101455
8. Ghosh PK et al (Jul. 2025) Identifying predictors and assessing causal effect on hypertension risk among adults using Double Machine Learning models: Insights from Bangladesh Demographic and Health Survey. PLoS Comput Biol 21(7):e1013211. 10.1371/journal.pcbi.1013211
9. Lundberg SM, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Accessed: Apr. 17, 2026. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
10. Biecek P, Burzykowski T (2021) Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. Chapman and Hall/CRC, New York. 10.1201/9780429027192
11. Muriithi DK, Lumumba VW, Awe OO, Muriithi DM (Feb. 2025) Explainable Artificial Intelligence Models for Predicting Malaria Risk in Kenya. Eur J Artif Intell Mach Learn 4(1):1–8. 10.24018/ejai.2025.4.1.47

12. Wickham H (2016) *ggplot2*. in Use R! Cham: Springer International Publishing. 10.1007/978-3-319-24277-4
13. Kuhn M (Nov. 2008) Building Predictive Models in R Using the caret Package. J Stat Softw 28:1–26. 10.18637/jss.v028.i05
14. Liaw A, Wiener M Classification and Regression by randomForest
15. Robin X et al (Mar. 2011) pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics 12:77. 10.1186/1471-2105-12-77
16. Biecek P (2018) DALEX: Explainers for Complex Predictive Models in R. J Mach Learn Res 19(84):1–5
17. Hypertension-risk-model-main Accessed: Apr. 17, 2026. [Online]. Available: https://www.kaggle.com/datasets/khan1803115/hypertension-risk-model-main
18. Zhang Z (Jan. 2016) Missing data imputation: focusing on single imputation. Ann Transl Med 4(1). 10.3978/j.issn.2305-5839.2015.12.38
19. Kuhn M, Johnson K Applied Predictive Modeling, ResearchGate. Accessed: Apr. 17, 2026. [Online]. Available: https://www.researchgate.net/publication/267989574_Applied_Predictive_Modeling
20. Couronné R, Probst P, Boulesteix A-L (Jul. 2018) Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19:270. 10.1186/s12859-018-2264-5
21. Probst P, Wright MN, Boulesteix A-L Hyperparameters and tuning strategies for random forest, 10.1002/widm.1301
22. Oshiro TM, Perez PS, Baranauskas JA (2012) How Many Trees in a Random Forest? In: Perner P (ed) in Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science, vol 7376. Springer Berlin Heidelberg, vol. 7376., Berlin, Heidelberg, pp 154–168. doi: 10.1007/978-3-642-31537-4_13.
23. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 10.1016/j.ipm.2009.03.002
24. Fawcett T (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 10.1016/j.patrec.2005.10.010
25. Liaw A, Wiener M (2002) Classification and regression by randomForest, *R News*, vol. 2, no. 3, pp. 18–22, Dec
26. Breiman L (2001) Random Forests, *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 10.1023/A:1010933404324

Figures

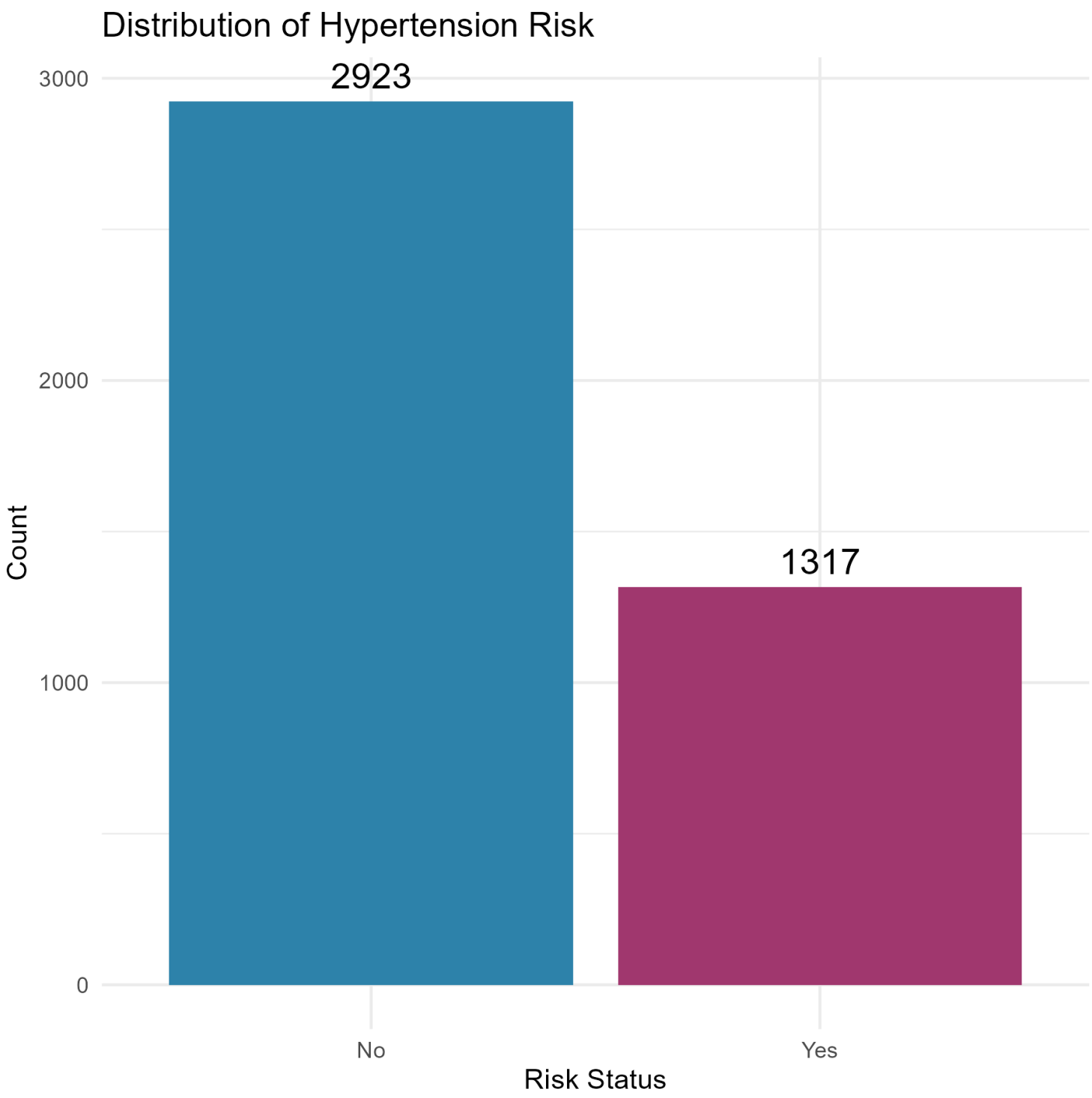


Figure 1

Distribution of Hypertension Risk

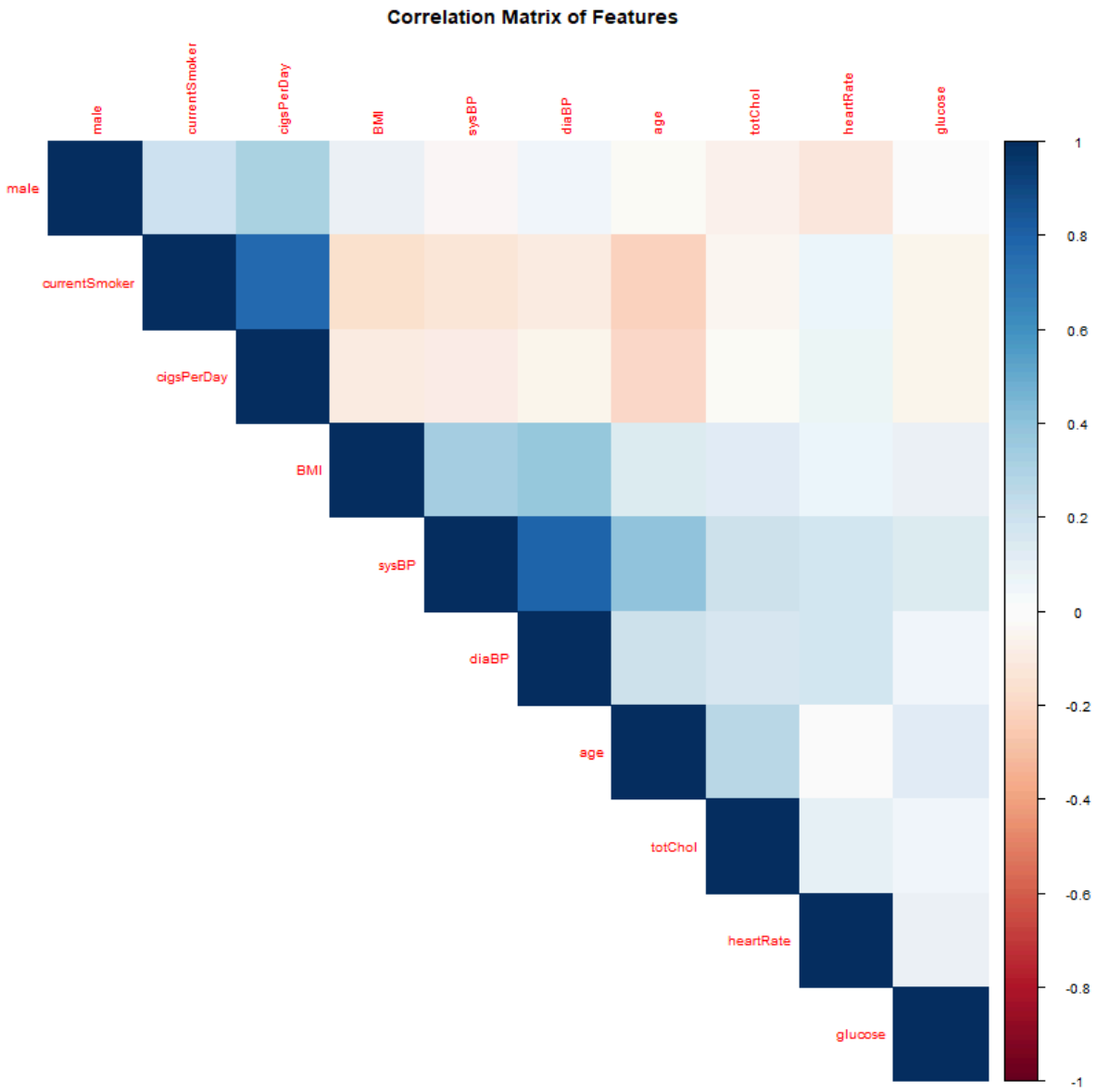


Figure 2

Correlation Matrix

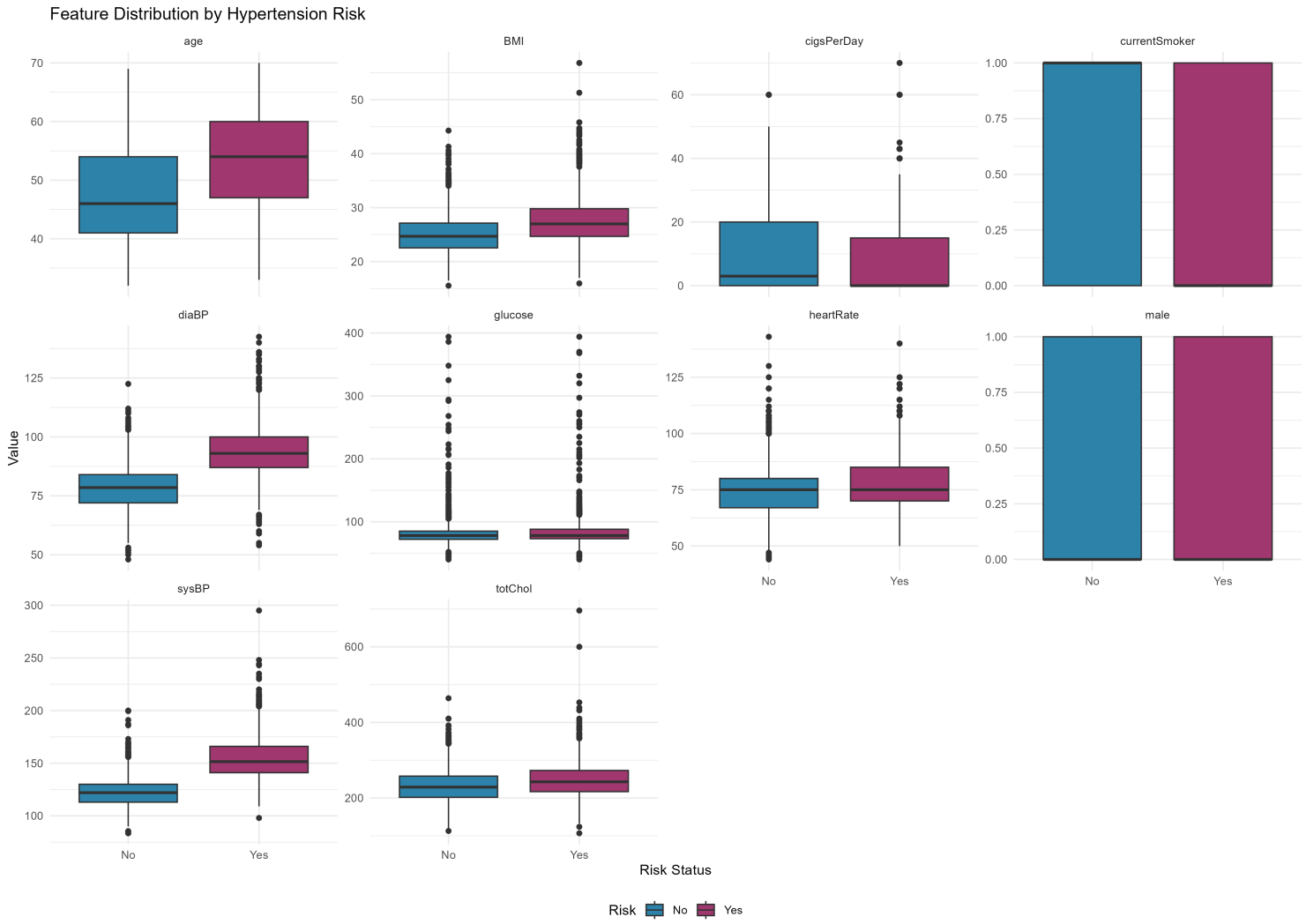


Figure 3

Feature Distribution Box plots

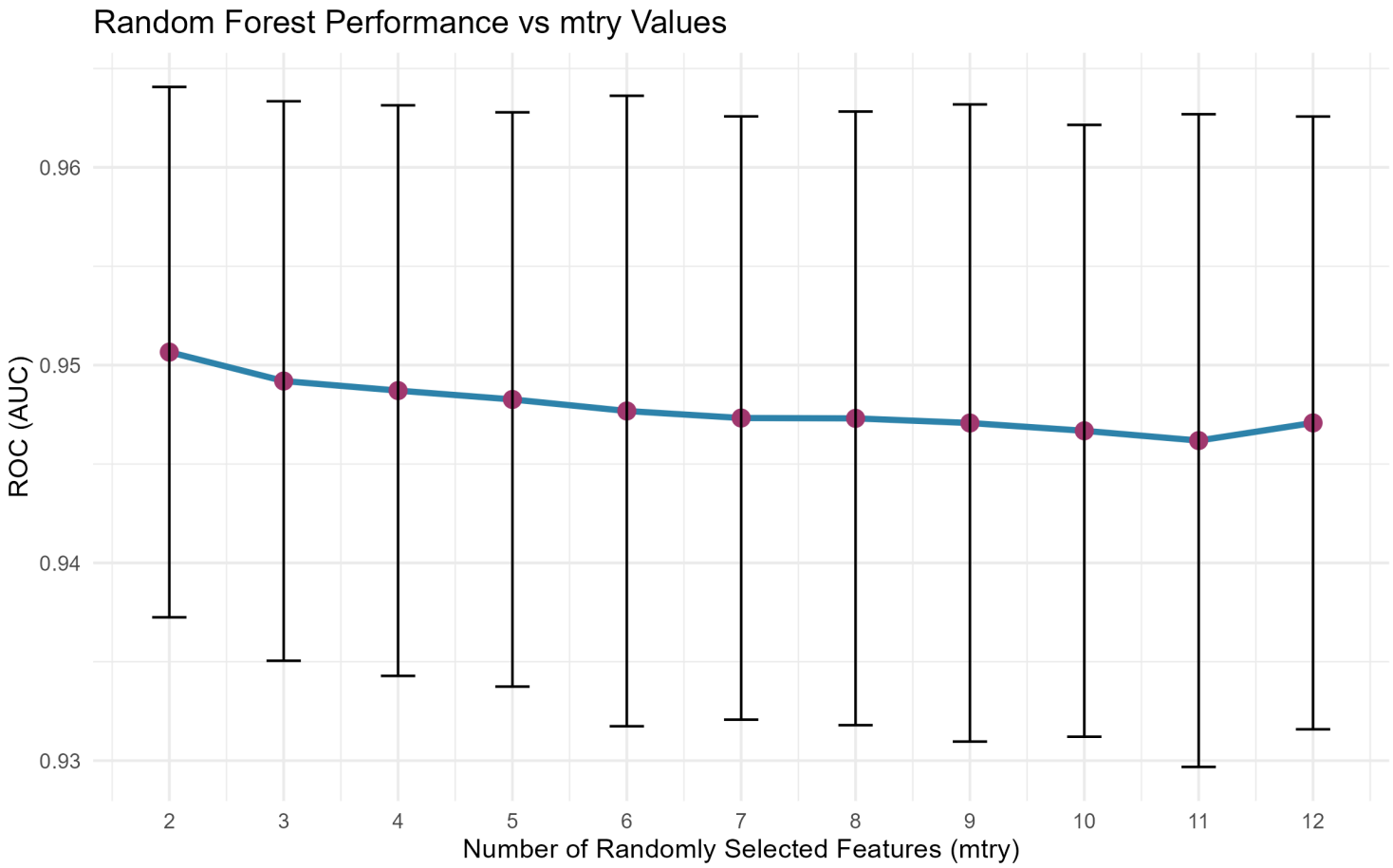


Figure 4

RF performance VS mtry Values

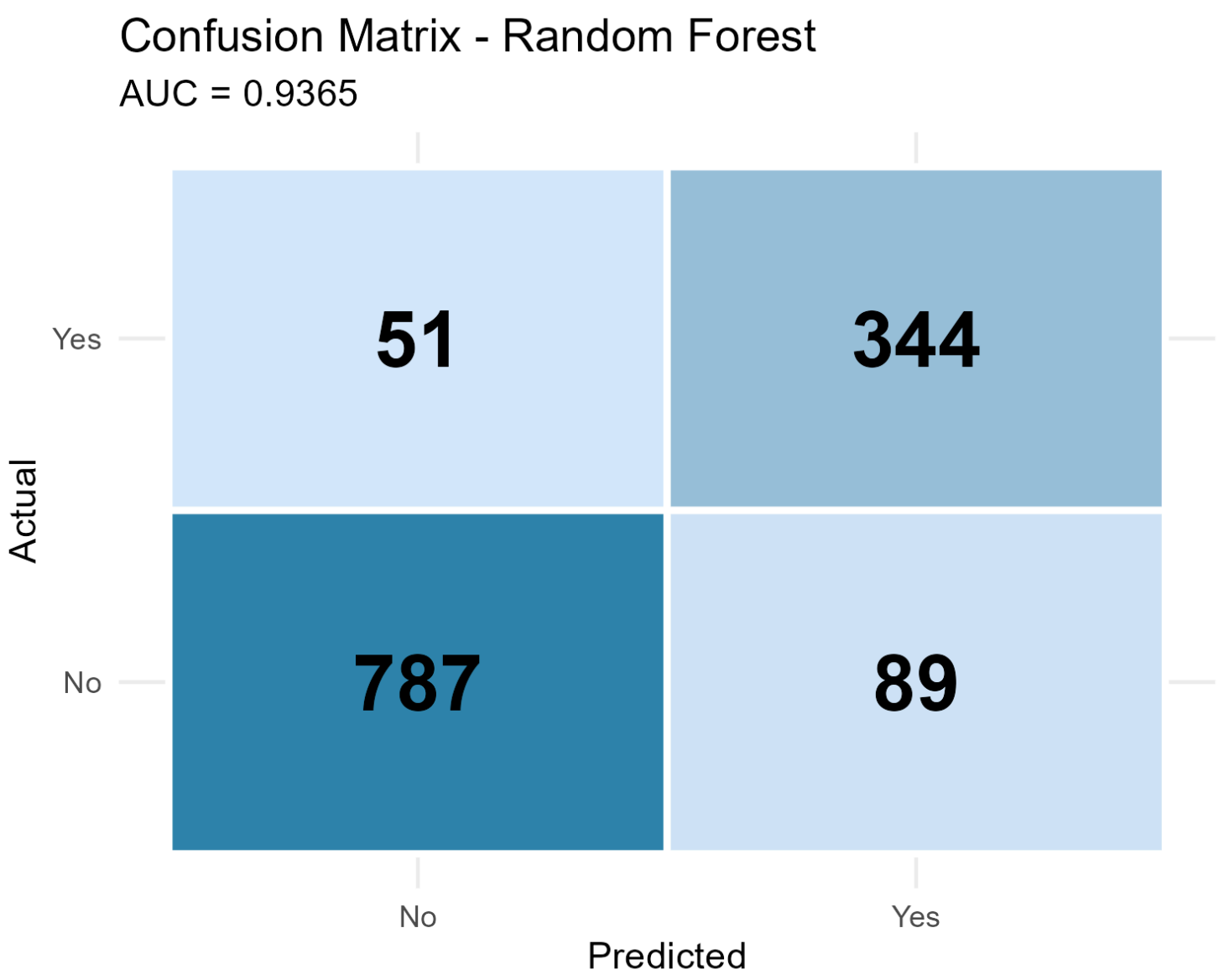


Figure 5

Confusion Matrix

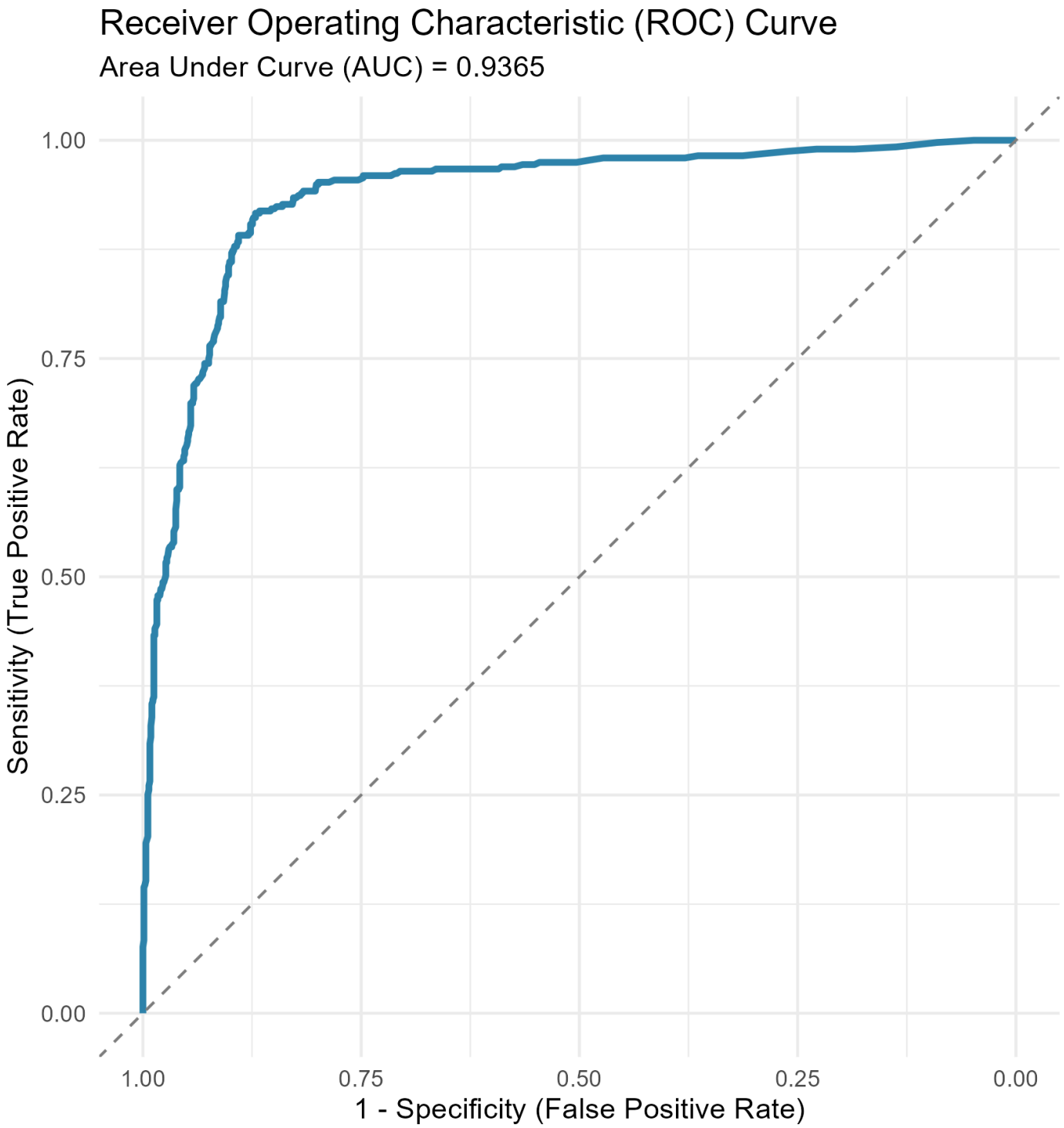


Figure 6

ROC Curve

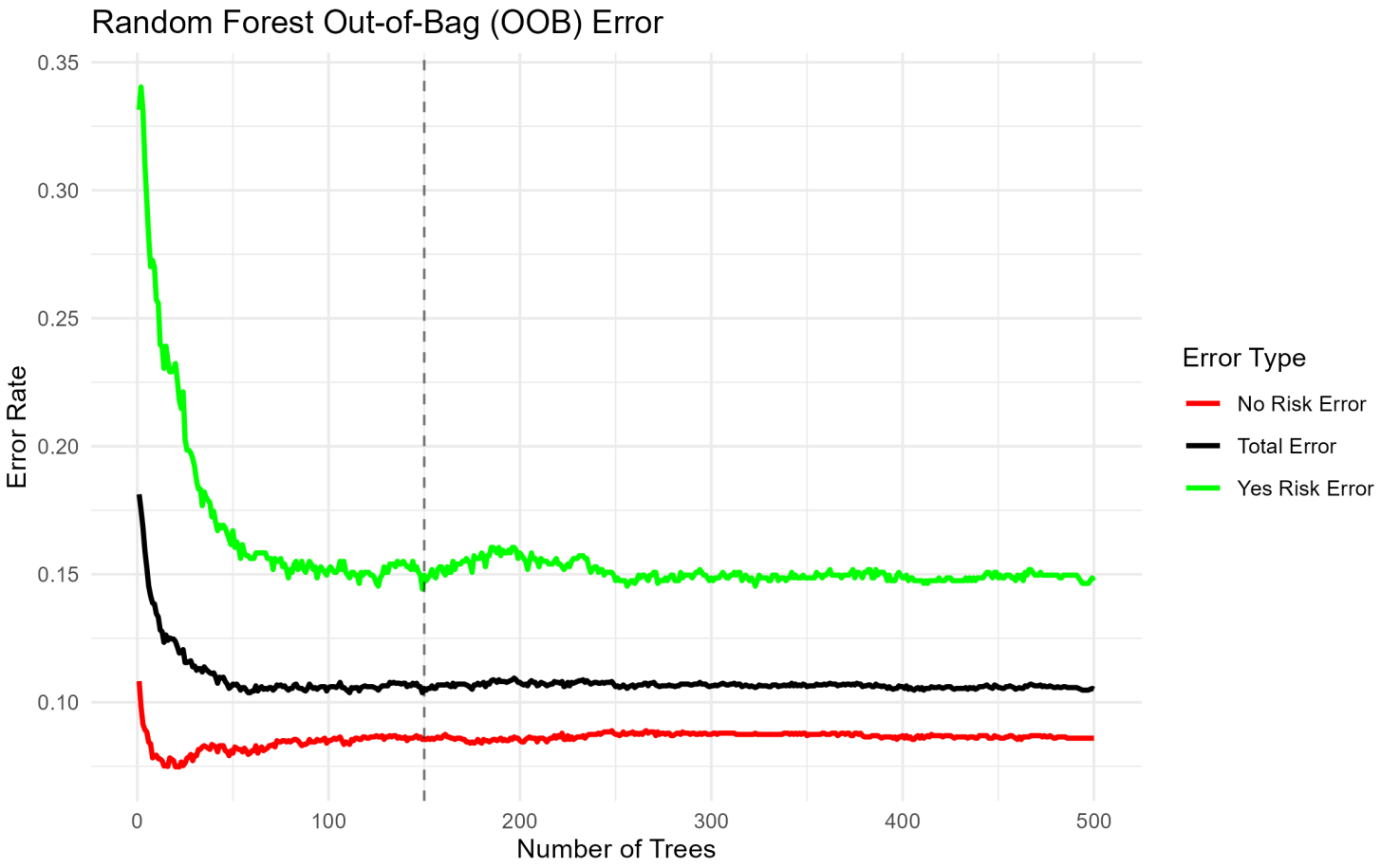


Figure 7

OBB Error Curves

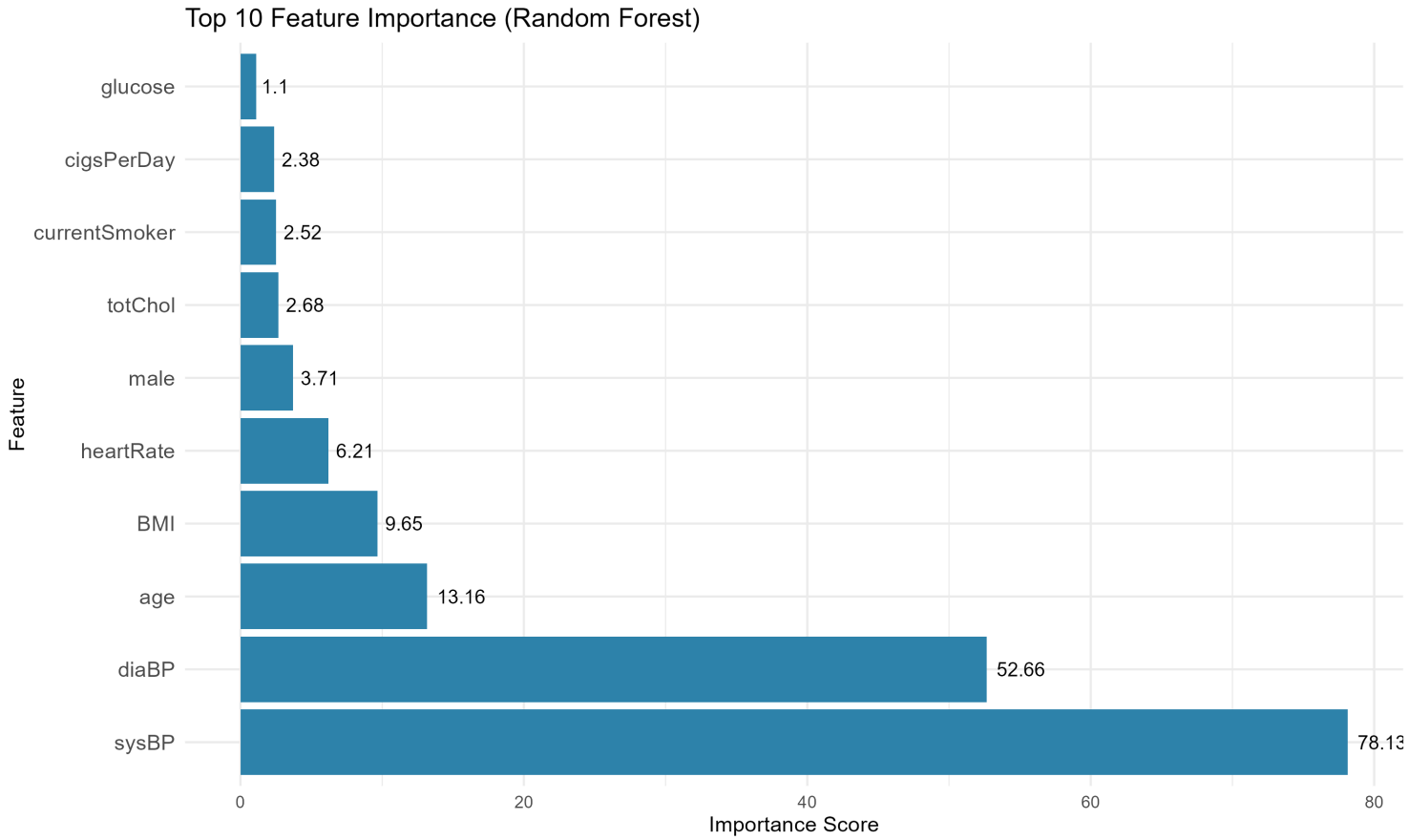


Figure 8

Top 10 feature importance

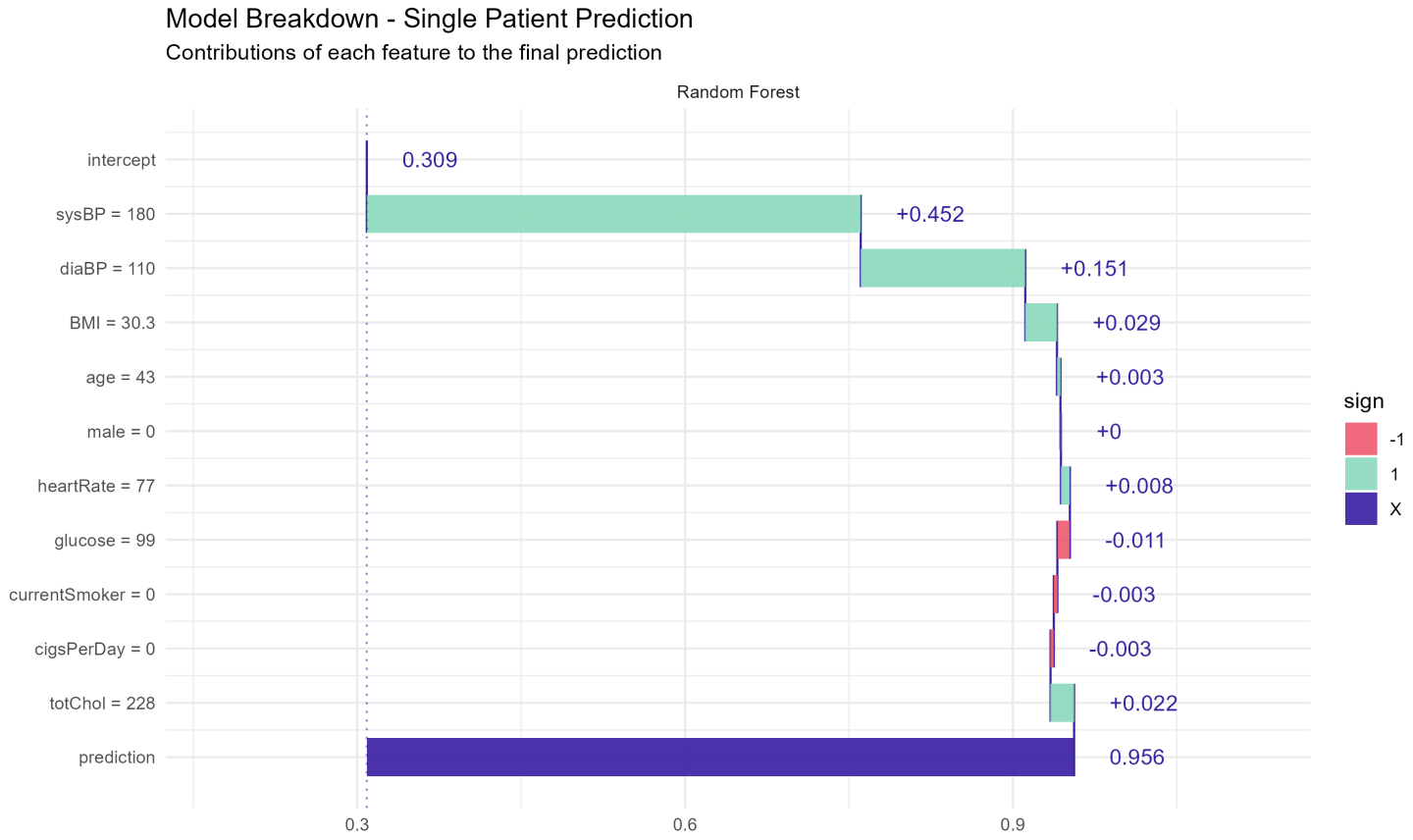


Figure 9

Breakdown map

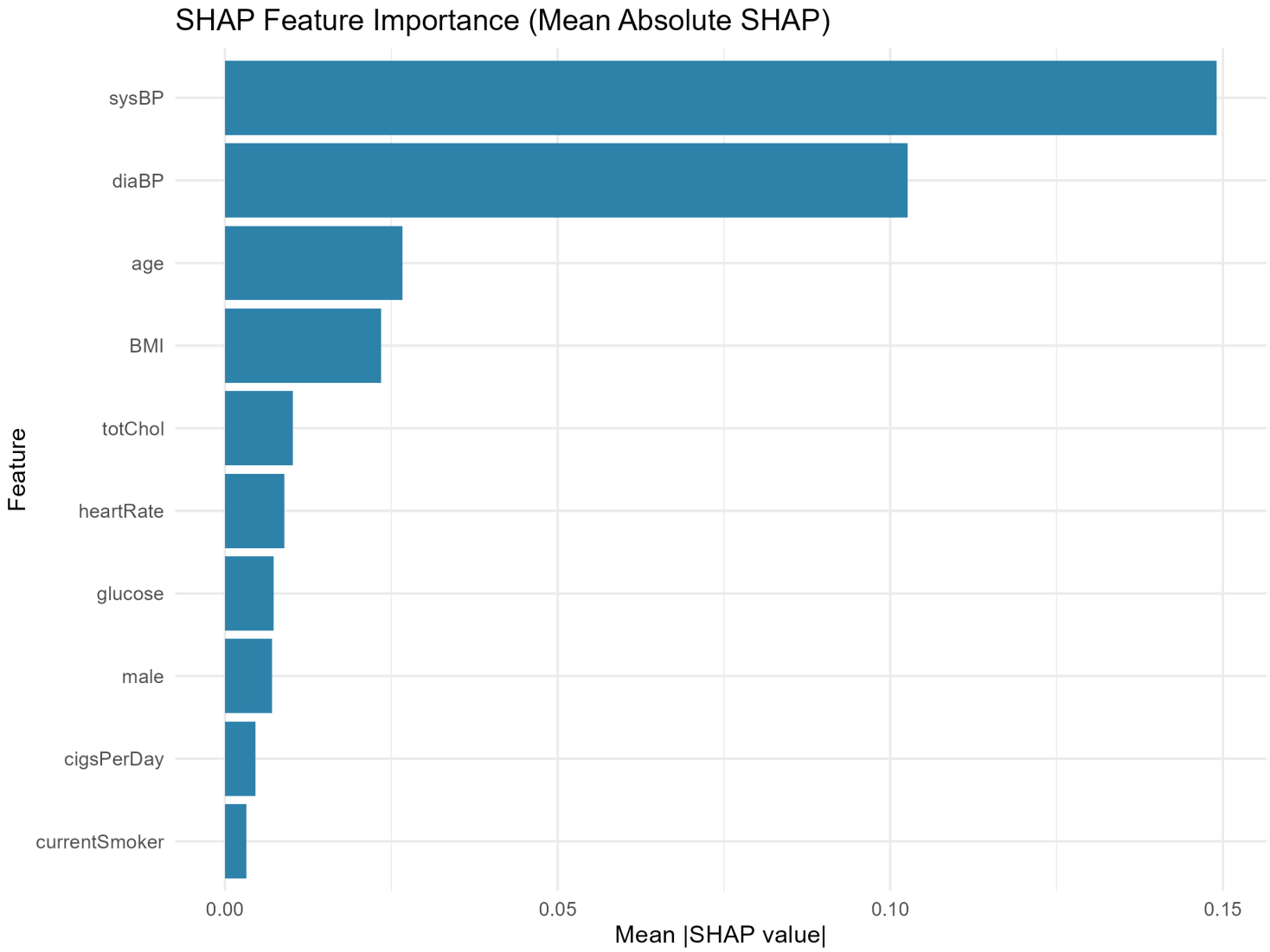


Figure 10

SHAP Feature Importance



Figure 11

SHAP Beeswarm Plot