Prepared by: Supreet Mutsuddi

Table of Contents

Pima Indians Diabetes Prediction Using Random Forest and XGBoost

1. Introduction

The goal of this analysis is to predict whether a person has diabetes using the Pima Indians Diabetes dataset. Two ensemble learning methods were employed: Random Forest and Extreme Gradient Boosting (XGBoost). These models are suitable for binary classification and capable of handling complex nonlinear interactions.

2. Data Overview

The dataset consists of 768 observations and 9 variables, where the target variable `Outcome` indicates diabetes presence (1) or absence (0). Summary statistics and structure were examined to understand distributions and detect any anomalies.

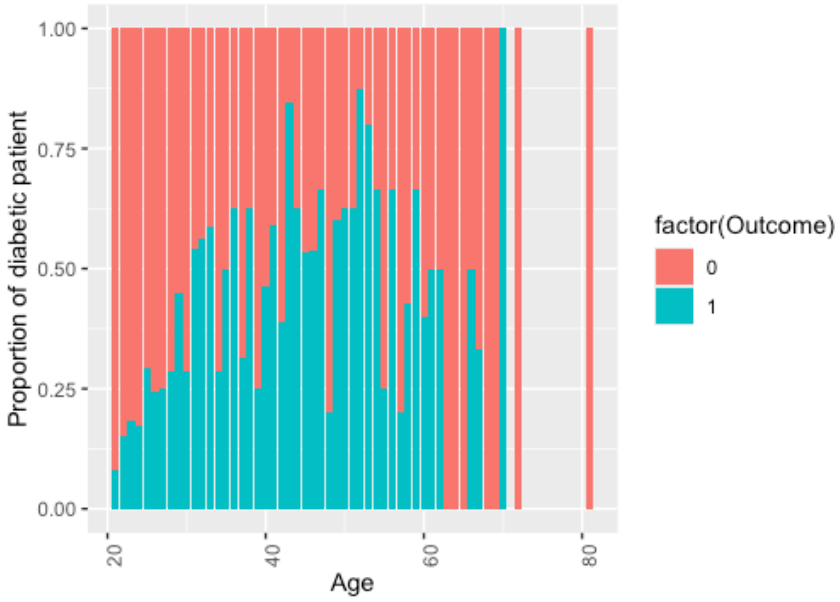
3. Data Preprocessing

Replaced `0` values in the `Insulin` variable with the median of non-zero insulin levels. This was done because a value of 0 for insulin is physiologically implausible and likely represents a missing measurement. Using the median (rather than the mean) helps reduce the influence of outliers and preserves the skewed nature of the distribution.

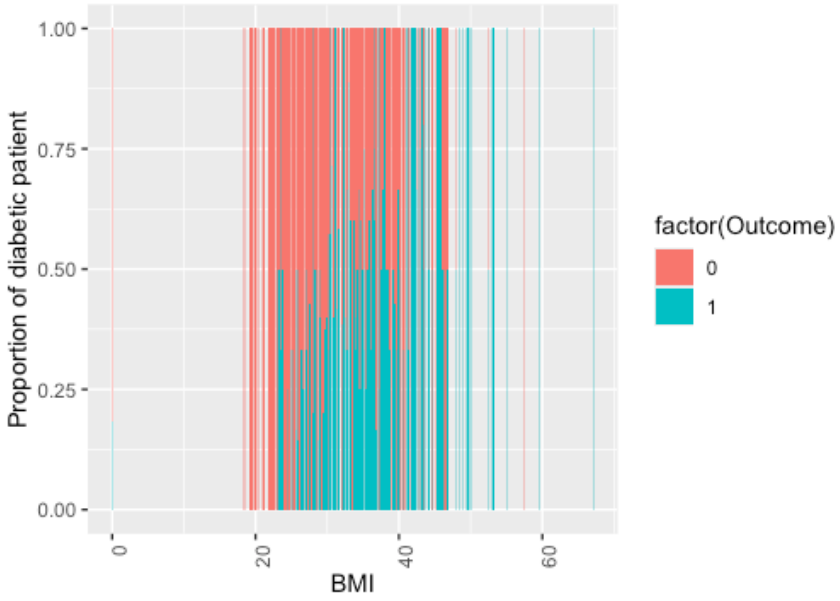
4. Exploratory Data Analysis

Bar plots were created for each predictor against the `Outcome`, showing the proportion of diabetic and non-diabetic cases. This revealed that predictors such as `Glucose`, `BMI`, and `Age` have strong discriminative power.

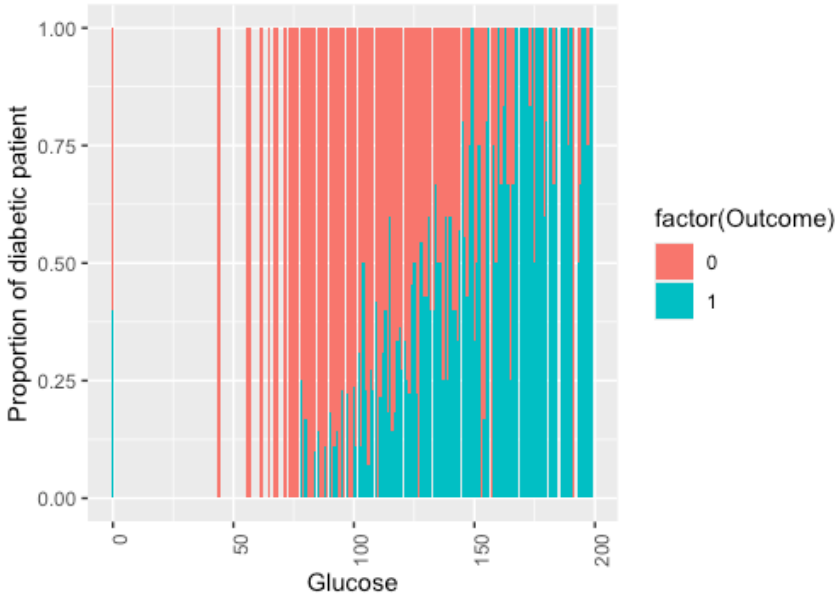
Proportion of Diabetic Patients by Age:



Proportion of Diabetic Patients by BMI:



Proportion of Diabetic Patients by Glucose:



5. Data Partitioning

The dataset was split into training (70%) and testing (30%) sets using stratified sampling to maintain class distribution. Outcome proportions remained consistent across splits.

6. Random Forest Modeling

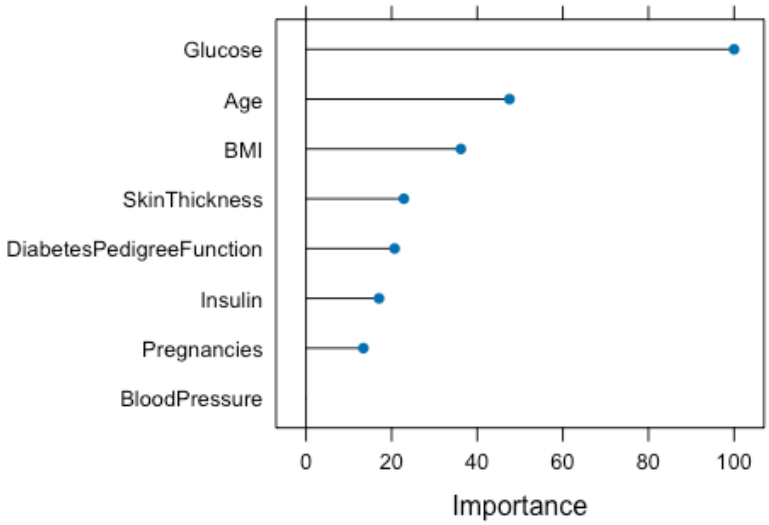
Three Random Forest models were built with `ntree` values of 100, 300, and 500. A repeated 5-fold cross-validation (3 repeats) with down-sampling was used to mitigate class imbalance.

Best model: `ntree = 300` achieved:

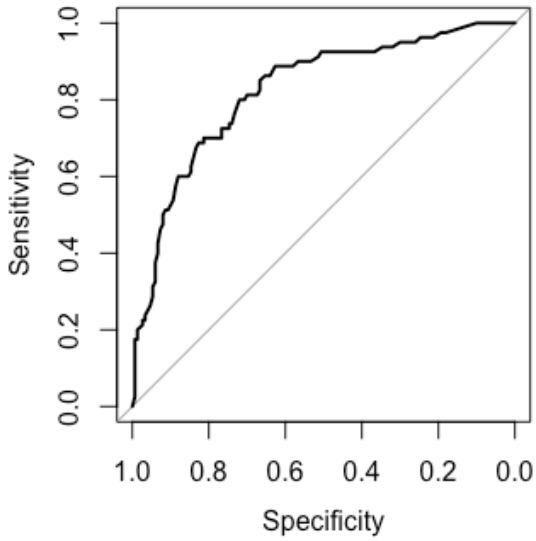
- Accuracy: ~0.7435

- AUC: 0.8364

Variable Importance (Random Forest):



ROC Curve (Random Forest):



7. XGBoost Modeling

A comprehensive grid search was conducted with parameters: `nrounds`, `eta`, `max_depth`, and `subsample`.

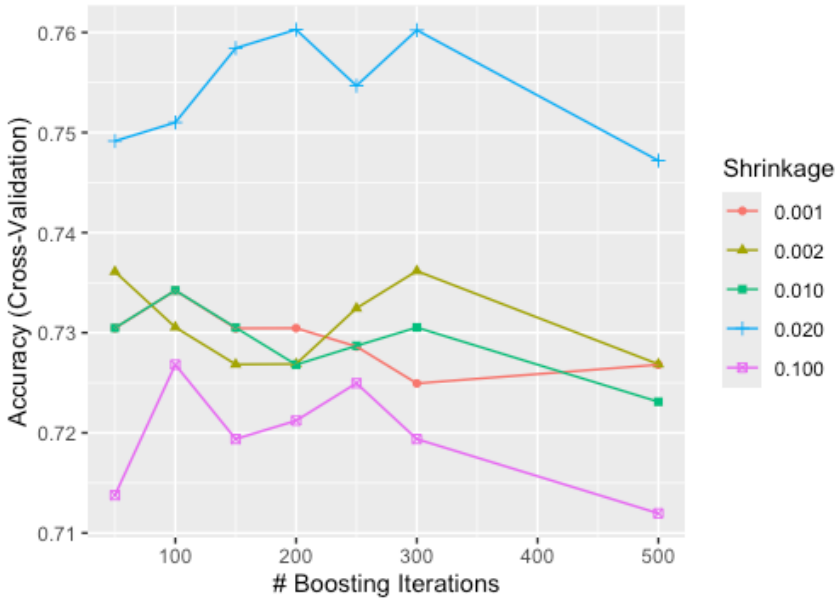
The final tuned XGBoost model used:

- `nrounds = 200`, `max_depth = 7`, `eta = 0.02`, `subsample = 0.6`

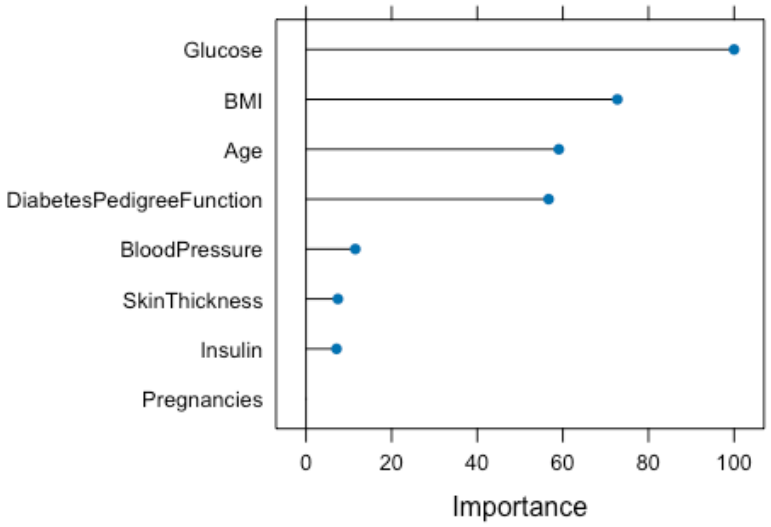
- Accuracy: ~0.7304

- AUC: 0.8153

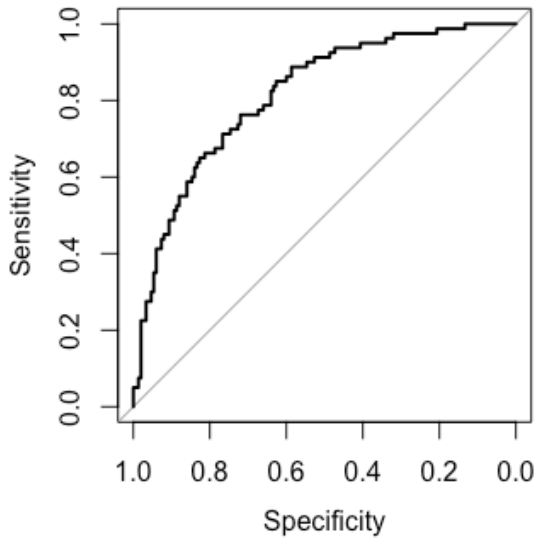
Tuning Performance Plot:



Variable Importance (XGBoost):



ROC Curve (XGBoost):



8. Model Comparison

Random Forest slightly outperformed XGBoost in both accuracy and AUC, indicating better generalization on this dataset.

Model	Test Accuracy	Test AUC
Random Forest	0.7435	0.8364
Boosted Tree	0.7304	0.8153

9. Conclusion

Both ensemble models performed well in predicting diabetes, with Random Forest offering a slight edge in test performance.

Key predictors consistently ranked across models include `Glucose`, `BMI`, and `Age`.

Interestingly, `BloodPressure` showed negligible importance in both models, which is unexpected given its known association with diabetes risk. This may be due to data quality issues (e.g., presence of zero values), lack of variability, or collinearity with other predictors. Further investigation into this feature's distribution and interaction effects is warranted.

Future work may explore regularization, interaction effects, or handling class imbalance with more advanced resampling.

10. Data Source

Kaggle: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database