Exploring the Association Between Clinical Predictors and Diabetes Disease Progression: A Regression Analysis

Anteneh G. Yitayal 256983

2025-03-28

Introduction

Diabetes is a chronic metabolic disorder that poses significant health risks, including the progression of complications over time. Understanding the factors influencing disease progression is crucial for better patient management and prevention of severe outcomes. In this study, we investigate the relationship between disease progression after one year (measured as "progr") and various clinical predictors in a sample of 442 diabetic patients. The explanatory variables include age, sex, body mass index (BMI), average blood pressure (BP), total cholesterol (TC), low-density lipoproteins (LDL), high-density lipoproteins (HDL), the ratio between total cholesterol and HDL (TCH), triglyceride levels (TG), and blood glucose levels (GC). By analyzing these factors using regression techniques, the goal is to identify the key predictors of disease progression, thereby aiding in the development of more targeted treatment plans for diabetic patients.

Objective

The objective of this analysis is to explore how different clinical factors like age, BMI, blood pressure, cholesterol levels, and blood glucose impact the progression of diabetes over one year. By using models like regression trees, random forests, and boosting it is aimed to identify which factors are most strongly related to the worsening of the disease. The goal is to find key predictors of diabetes progression and evaluate which model best explains the changes in disease over time, helping to improve patient care and treatment strategies.

Data set

The dataset consists of 442 records with 11 features (10 predictors and 1 target variable). All variables are complete with no missing values. The target variable, disease progression ('progr'), shows considerable variation (25-346) with a mean of 152.1, suggesting diverse disease outcomes across patients. Key biomarkers show wide ranges: BMI (18.0-42.2), blood pressure (62-133 mmHg), and triglycerides (TG: 3.258-6.107). These ranges encompass both normal and significantly elevated values, providing a comprehensive dataset for modeling disease progression.

Exploratory Data Analysis

The correlation heatmap in figure @ref(fig:fig1) reveals strong positive relationships between progr (disease progression) and both TG (triglycerides) and BMI, with moderately strong correlations with BP (blood pressure) and GC (glucose). HDL shows a notable negative correlation with progression, confirming its protective effect. The scatter plot in figure @ref(fig:fig2) visually confirm these relationships, with clear positive trends between TG and disease progression, though with considerable variability this relationship

is consistent for BP. The plot comparing disease progression by sex shows only minimal differences between males and females, with females showing slightly higher median values and marginally wider distribution, but this difference appears clinically insignificant compared to metabolic factors.

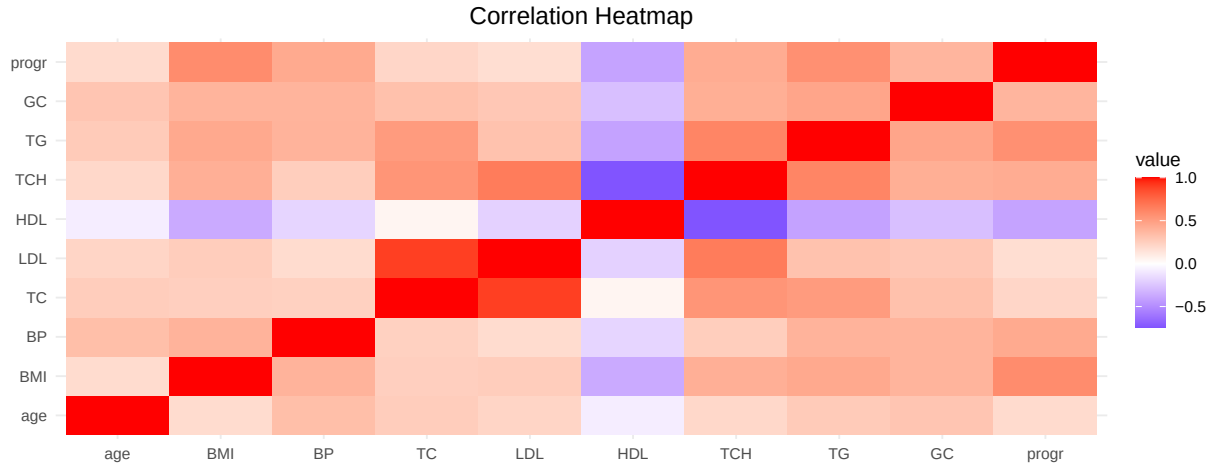

Figure 1: Pairwise correlation among the variables and the target

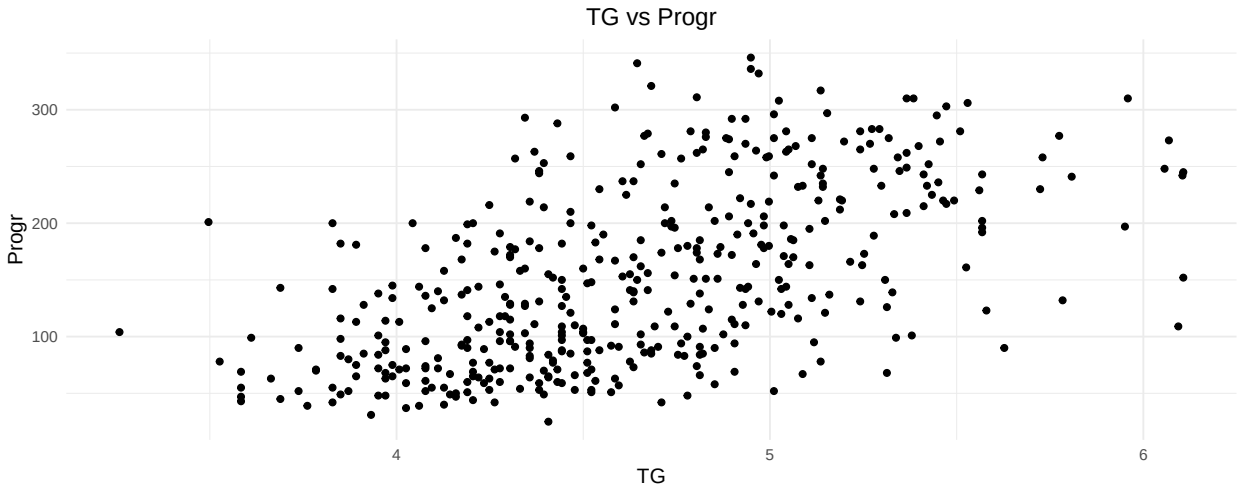

Figure 2: Distribution of Disease progression(Progr) with Triglycerides(TG)

Methods

In this analysis, there are three main regression models used to predict disease progression: a decision tree, a random forest, and a boosting model. Each model is evaluated based on its predictive performance, interpretability, and the importance of the variables used.

Decision Tree

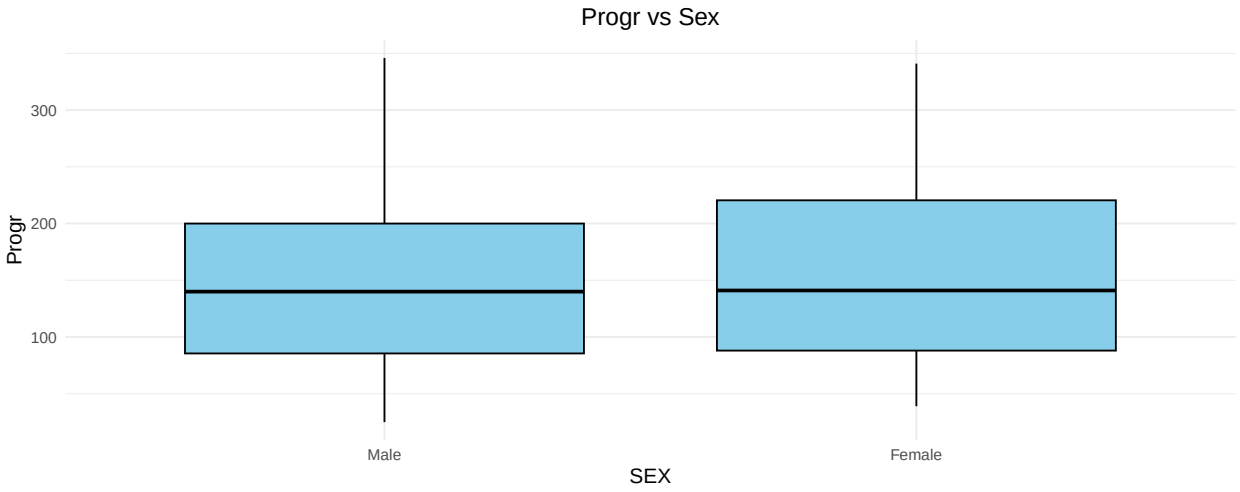

Figure 3: Distribution of Coronary Heart Disease(CHD) with Diastolic blood pressure.

```
# Fit the decision tree model
set.seed(1) # for reproducibility
dec_tree_model <- tree(progr ~ ., data = data_df)
summary(dec_tree_model)
```

```
##
## Regression tree:
## tree(formula = progr ~ ., data = data_df)
## Variables actually used in tree construction:
## [1] "TG"  "BMI" "HDL" "GC"  "BP"
## Number of terminal nodes:  12
## Residual mean deviance:  2674 = 1150000 / 430
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -140.900  -35.830   -4.805    0.000   33.540  154.100
```

The optimal tree size is computed using cross validation for the set of sizes, as in figure @ref(fig:fig4) captures the essential bias-variance tradeoff in decision tree modeling, showing how predictive performance relates to model complexity. Starting with high deviance (poor performance) at tree size 1, the error rapidly decreases until reaching an optimal minimum at approximately 5 terminal nodes, after which the deviance gradually increases again as the tree grows to 12 nodes. This U-shaped curve clearly identifies the optimal spot where the model has sufficient complexity to capture important patterns in the data without overfitting to noise, validating the decision to prune the original 12-node tree down to approximately 5 nodes for optimal generalization performance. The graph serves as empirical evidence for selecting the appropriate level of model simplification, balancing predictive accuracy against interpretability and preventing overfitting.

The full regression tree (12 nodes) shows a 20% lower training error (residual mean deviance 2674) using five predictors (TG, BMI, HDL, GC, BP) compared to the pruned tree's simpler structure (5 nodes, deviance 3215) using only TG and BMI. While this indicates better data fitting by the full model, these metrics reflect performance on the same data used for training rather than generalization ability. The pruned model likely offers better future prediction performance despite its higher apparent error, as it reduces the risk of overfitting to noise in the training data, which clearly confirmed with cross validation as in Figure @ref(fig:fig4).

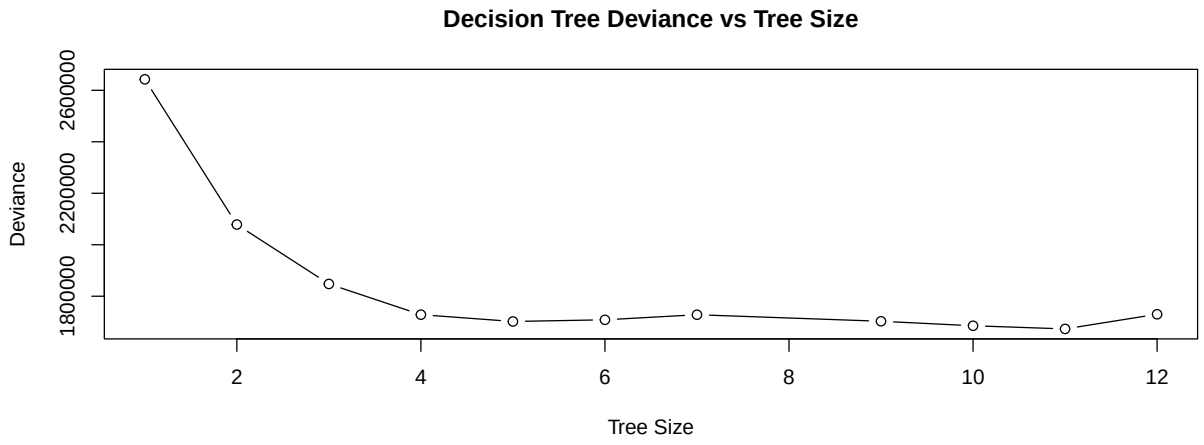

Figure 4: Decision tree size for set of tree sizes

```
set.seed(1)
# Prune the tree
pruned_tree <- prune.tree(dec_tree_model, best = 5)
summary(pruned_tree)
```

```
##
## Regression tree:
## snip.tree(tree = dec_tree_model, nodes = c(4L, 6L, 14L))
## Variables actually used in tree construction:
## [1] "TG"  "BMI"
## Number of terminal nodes:  5
## Residual mean deviance:  3215 = 1405000 / 437
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -156.60  -37.31   -7.31    0.00   37.40  156.70
```

The pruned decision tree in figure @ref(fig:fig5) provides a streamlined model for predicting diabetes progression using only triglycerides (TG) and BMI. It first splits patients based on TG levels at 4.60015, with lower values indicating better outcomes. Patients with low TG are further divided by BMI at 26.95, resulting in either mild progression (96.31) or moderate progression (159.70). Those with high TG branch according to BMI thresholds (27.75 and 32.75), revealing a clear severity gradient from moderate (162.70) to severe (208.60) to very severe progression (268.90). This elegant five-terminal node structure efficiently captures the essence that patients with both elevated triglycerides and high BMI face substantially worse disease outcomes, with each factor intensifying the other's impact.

Random Forest

The random forest model was fitted to the data using 500 trees, with 10 variables randomly selected at each split. The mean squared residuals are 3384.245, which measures the average squared difference between observed and predicted values. The model explains about 42.86% of the variance in the response variable.

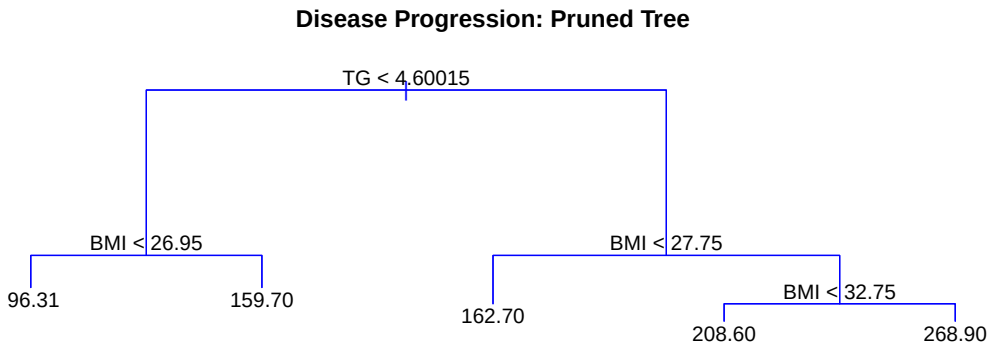

Figure 5: Pruned Decision Tree diagram

```
set.seed(1)
n_pred <- ncol(data_df) - 1
forest_tree_model <- randomForest(progr ~ ., data = data_df, mtry = n_pred, importance = TRUE)
forest_tree_model
```

```
##
## Call:
##  randomForest(formula = progr ~ ., data = data_df, mtry = n_pred,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 3384.245
##                     % Var explained: 42.93
```

The comparison between bagging model with all the features at each spiting and the random forst model reveals that the optimized model using only 2 variables per split (mtry=2) as shown in figure @ref(fig:fig6), which was selected using ten-fold cross validation, outperforms the model using all 10 predictors (mtry=10), achieving a lower mean squared error (3244.603 vs 3384.245) and explaining more variance in diabetes progression (45.28% vs 42.93%). This about 4.2% improvement in error metrics demonstrates the effectiveness of the random variable selection process in creating more diverse trees while reducing the impact of overfitting. The result challenges the conventional wisdom of using more information at each decision point, confirming that carefully tuned hyperparameters can lead to more accurate predictions even when limiting the information available at each split. This finding emphasizes the importance of proper cross-validation for hyperparameter tuning in ensemble models rather than relying on default settings or using all available predictors.

```
set.seed(1)
# Extract optimal mtry
optimal_mtry <- rf_tune$bestTune$mtry

# Fit the random forest model with optimal mtry
rf_model <- randomForest(progr ~ ., data = data_df, mtry = optimal_mtry, ntree = 500)
```

```r
# Print the model summary
print(rf_model)
```

```
##
## Call:
##  randomForest(formula = progr ~ ., data = data_df, mtry = optimal_mtry,      ntree = 500)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 3244.603
##                     % Var explained: 45.28
```

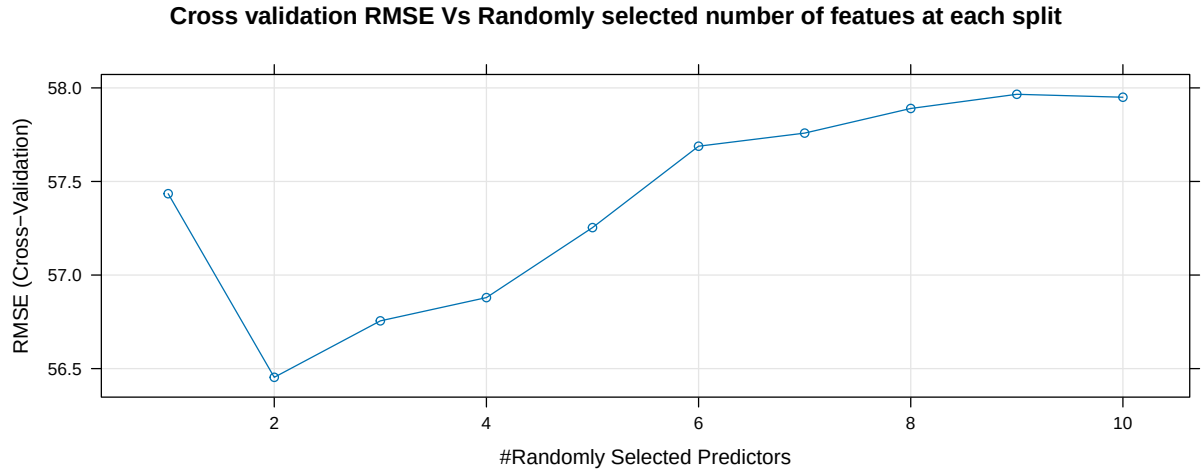

Figure 6: Random Forest RMSE with different number of randomly selected featues

This plot in figure @ref(fig:fig7) shows the relative importance of different predictors in explaining diabetes progression. Triglycerides (TG) and Body Mass Index (BMI) emerge as the most influential factors by a substantial margin, with TG slightly outperforming BMI. Blood pressure (BP) ranks as the third most important predictor, followed by HDL cholesterol and glucose control (GC) with moderate importance.

Boosting

A gradient boosted model with gaussian loss function model was fitted with 5000 trees, a shrinkage of 0.01 which is common, and a maximum interaction depth of 4. The optimal number of trees is selected using ten-fold cross validation for a series of number of tree of multiple of 100 between 100 and 2000 and the the minimum observations at the termial node set to 10 while keeping the other parameters the same, the RMSE is the lowest at the number of tress of 400 as shown in figure 8.

The optimized model with 400 trees reveals that diabetes progression is predominantly driven by just two metabolic factors (BMI and triglycerides), with a significant secondary contribution from blood pressure. The concentration of predictive power in these variables suggests that interventions targeting weight management and lipid control would likely have the greatest impact on diabetes outcomes.

The feature importance distribution in figure 9 reveals that just two metabolic factors (TG and BMI) account for over 66% of the predictive power, while the top three factors (including BP) represent more than 77% of the model's predictive capability. This strongly supports focusing clinical interventions on triglycerides, BMI, and blood pressure for managing diabetes progression.

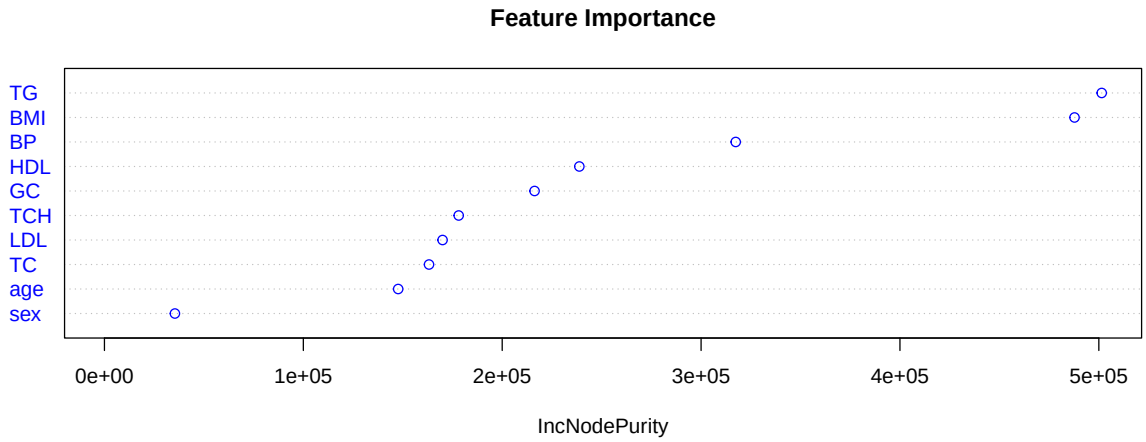

Figure 7: Random Feature importance

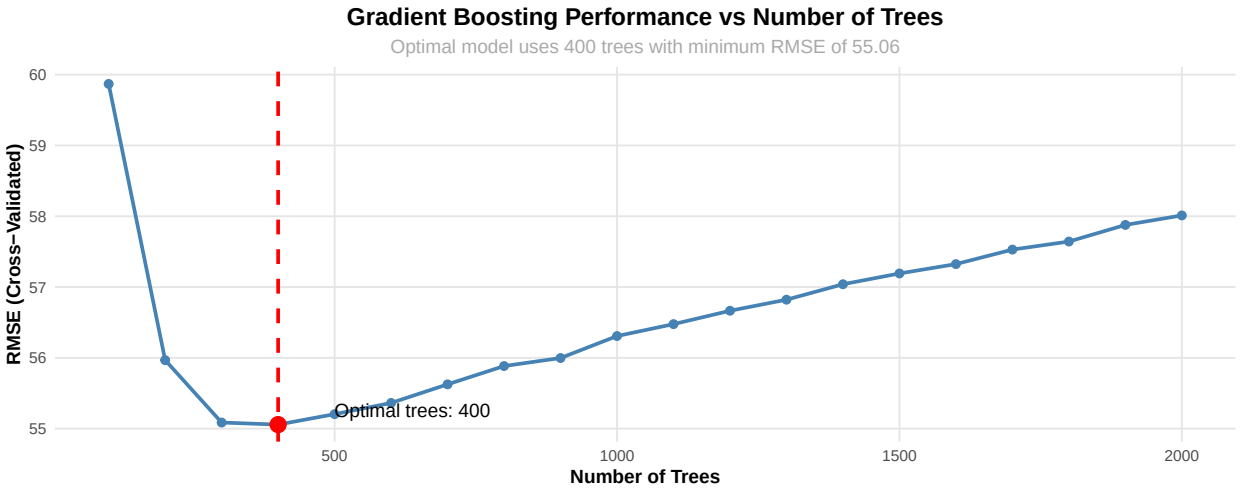

Figure 8: Gradient Boosting Performance by Number of Trees

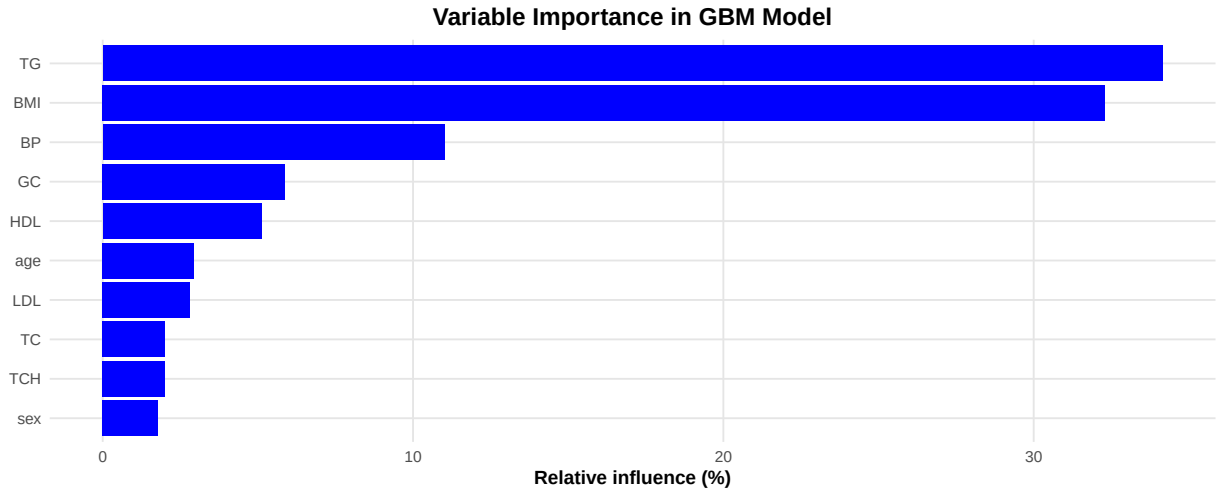

Figure 9: Gradient Boosting relative feature importance

Results and Discussion

A comprehensive 10-fold cross-validation framework was used to compare the three statistical models. For each fold, the data is split into training and testing sets, then each model is optimized and evaluated:

1. **Decision Trees**: First builds a full tree, then uses cross validation to determine the optimal complexity parameter, prunes accordingly, and calculates RMSE on the test fold.
2. **Random Forests**: Evaluates 10 different mtry values (variables considered at each split) using OOB error to select the optimal value, then builds a forest with 500 trees and this optimal mtry parameter.
3. **Gradient Boosting**: Uses 10-fold cross-validation to identify the best combination of hyperparameters from a grid of options: number of trees (100, 400, 500, 1000), interaction depth (1, 4, 5), and shrinkage (0.01, 0.1).

After collecting RMSE values for all 10 folds, the mean performance and standard deviations for each method was calculated, providing a robust comparison of predictive accuracy across the models. The boxplot in figure 10 compares the cross-validated performance of the models show that Random Forest demonstrates the best overall performance with the lowest median RMSE and narrowest distribution, indicating consistent predictive ability across different data subsets. Gradient Boosting shows competitive performance with slightly higher variability but maintains strong predictive power. Decision Trees exhibit the highest median RMSE and widest spread, confirming they are the least accurate and most variable of the three approaches. The Random Forest's superior performance likely stems from its ability to capture complex relationships between key metabolic factors (TG, BMI, BP) while reducing overfitting through ensemble averaging.

Conclusion

This analysis demonstrates that ensemble methods provide superior predictive performance for diabetes progression while maintaining valuable clinical interpretability. The findings support focusing clinical attention on metabolic factors, particularly triglycerides and BMI, when developing interventions to slow disease progression. Future work could explore additional biomarkers, temporal dynamics of progression, and patient-specific response patterns.

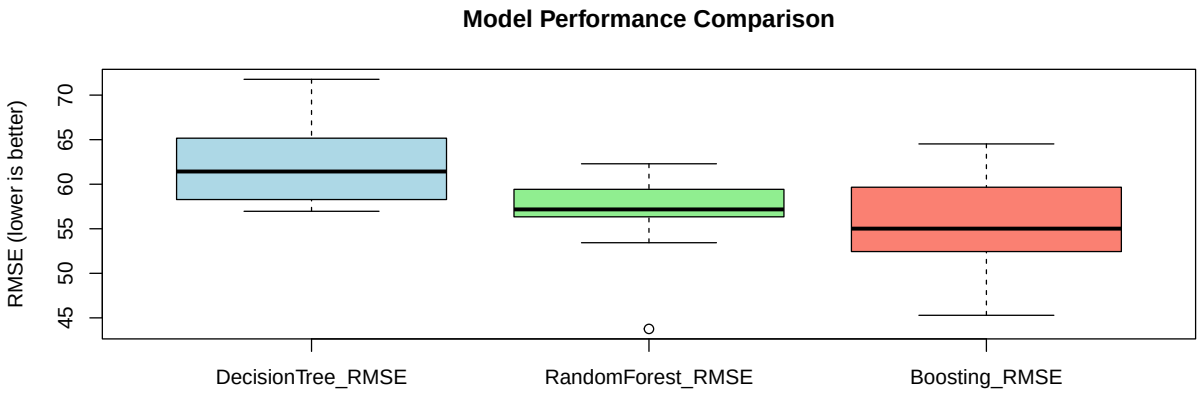

Figure 10: Performance comparision among the models

References

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

*Source code: https://github.com/abiget/*diabetes-progression