STAT 469/563 Assignment #3

Author: *Ana Compri Jacob*

Abstract

This report builds upon the findings of Assignment 2 by incorporating two additional classification methods: Random Forest and MTPS (residual stacking only). The study evaluates the performance of these methods alongside Linear Discriminant Analysis (LDA), and Elastic Net for classifying HIV drug resistance based on mutation data. The dataset was processed to convert continuous resistance levels into binary outcomes, and models were evaluated using 5-fold cross-validation across 10 iterations, a reduction from the 50 iterations in previous assignments. Performance was assessed using misclassification rate, precision, recall, and F1 score, and statistical comparisons were conducted using the Wilcoxon signed-rank test to evaluate the efficacy of the newly added methods against LDA. Results indicate that Elastic Net remains the most effective method, while Random Forest and LDA provide competitive performance.

1. Introduction

This study aims to evaluate different classification models for predicting HIV drug resistance by converting continuous resistance levels into a binary classification task. Building upon previous analyses, we extend our comparison to include Random Forest and MTPS (residual stacking only) alongside Linear Discriminant Analysis (LDA) and Elastic Net. The models are assessed using 5-fold cross-validation over 10 iterations, a reduction from the 50 iterations used in prior assignments, to ensure computational efficiency while maintaining robust performance evaluation.

Performance metrics include misclassification rate, precision, recall, and F1 score, and statistical significance is tested using the Wilcoxon signed-rank test, comparing LDA and Elastic Net with the newly introduced methods. The hypotheses remain as follows: H_0 (Null Hypothesis) states that there is no significant difference in classification performance among the models, while H_1 (Alternative Hypothesis) suggests that at least one model outperforms the others. The addition of ensemble-based (Random Forest) and stacking-based (MTPS) methods aims to explore whether these techniques enhance classification accuracy in drug resistance prediction.

2. Methodology

2.1 Data Preprocessing

To ensure a fair evaluation of classification models, stratified 5-fold cross-validation was employed. This technique ensures that each fold maintains approximately the same proportion of drug-resistant and non-resistant samples, preventing class imbalance from affecting model performance.

2.2 Classification Models

The models evaluated in this study are:

- **Linear Discriminant Analysis (LDA)** (assumes Gaussian-distributed data).
- **Elastic Net -** A regularized regression approach that combines L1 (Lasso) and L2 (Ridge) penalties, providing both feature selection and stability for correlated predictors.
    - The alpha value was set to 0.5 for simplicity.
    - Implemented using `glmnet` with cross-validation to determine the best lambda.
- **Random Forest -** An ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting.
    - Implemented using the `randomForest` package.
    - Cross-validation was used to evaluate performance across multiple runs.
- **MTPS (Residual Stacking Only) – Evaluated Separately**
    - A two-step model stacking approach that uses residual correction to refine predictions: 1 - A decision tree (rpart1) is used as a base model to approximate the decision boundary; 2 - A linear regression model (lm1) is applied to correct the residual errors, improving predictive performance.
    - Implemented with Pearson residuals and standardization for stability.
    - Unlike other models, MTPS was evaluated separately for each drug in a looped structure.
    - Results were aggregated across multiple repetitions per drug for consistency.

2.3 Performance Metrics and Analysis

The models were trained and tested, generating a Confusion Matrix (tn, fp, fn, tp). The Misclassification Rate (MCR), Precision, Recall, and F1-Scores were calculated. To determine whether differences in classification performance were statistically significant, Wilcoxon signed-rank tests were performed, comparing the F1 scores of Classification Tree and Elastic Net against LDA.

3. Results and Analysis

3.1 Statistical Significance (Wilcoxon Test Results)

The table below presents p-values from the Wilcoxon signed-rank test, comparing the classification performance (F1 scores) of LDA vs. MTPS, LDA vs. Random Forest, Elastic Net vs. MTPS, and Elastic Net vs. Random Forest across different drugs. The objective is to determine whether the differences in model performance are statistically significant.

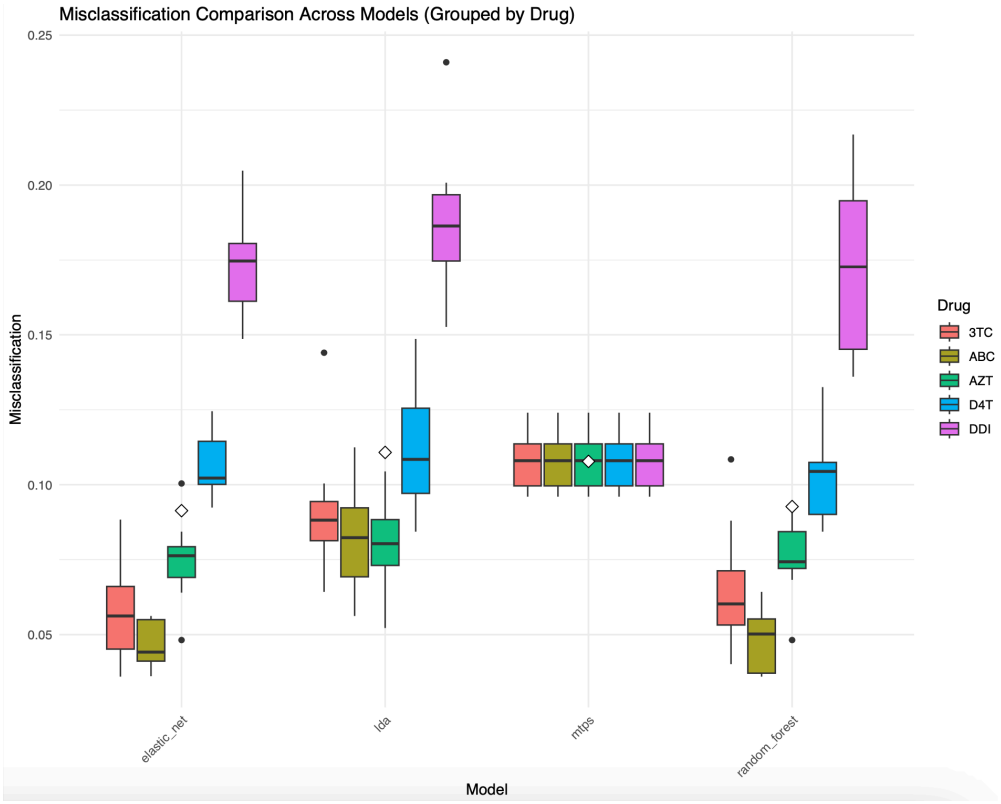
wilcox

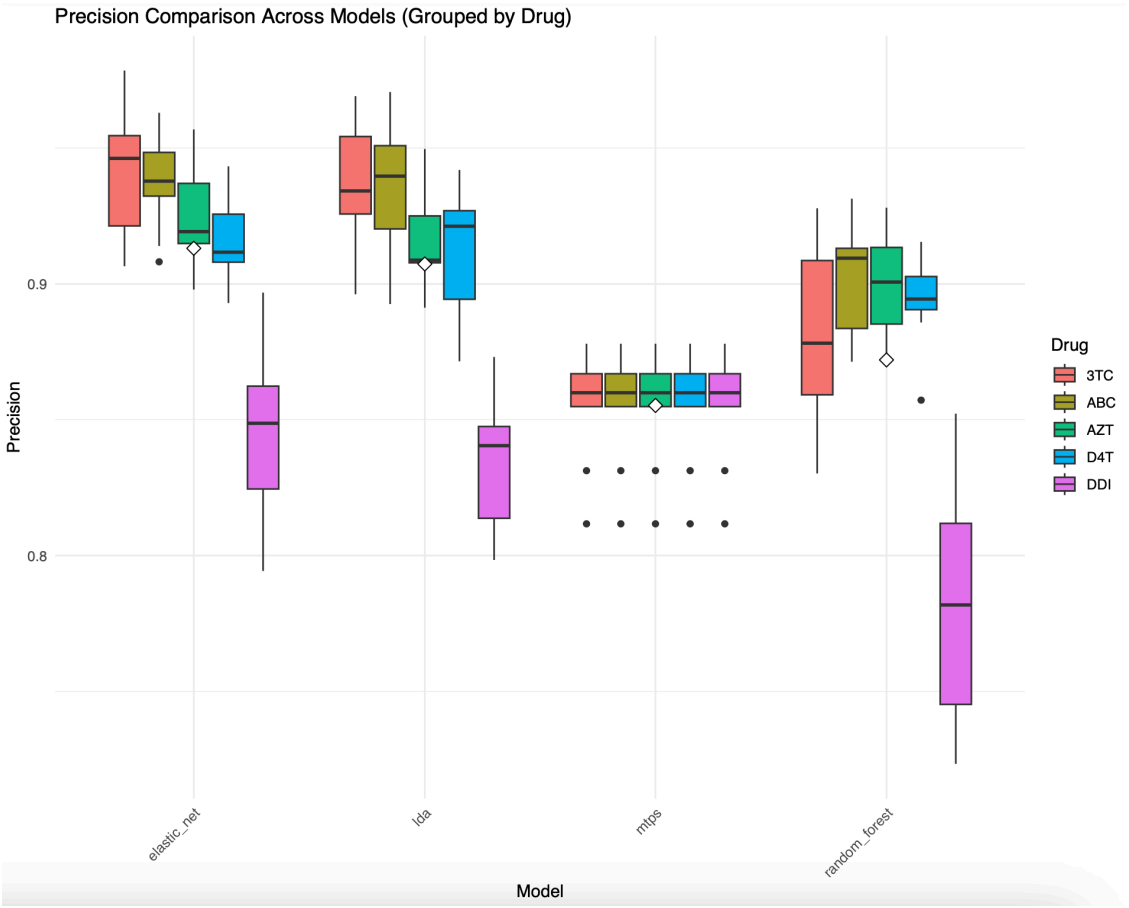
	LDA vs MTPS	LDA vs Random Forest	Elastic Net vs MTPS	Elastic Net vs Random Forest
ABC	0.16015625	0.00390625	0.001953125	0.322265625
3TC	0.322265625	0.037109375	0.001953125	0.009765625
AZT	0.001953125	0.625	0.001953125	0.375
D4T	0.037109375	0.375	0.00390625	0.625
DDI	0.001953125	0.625	0.001953125	0.232421875

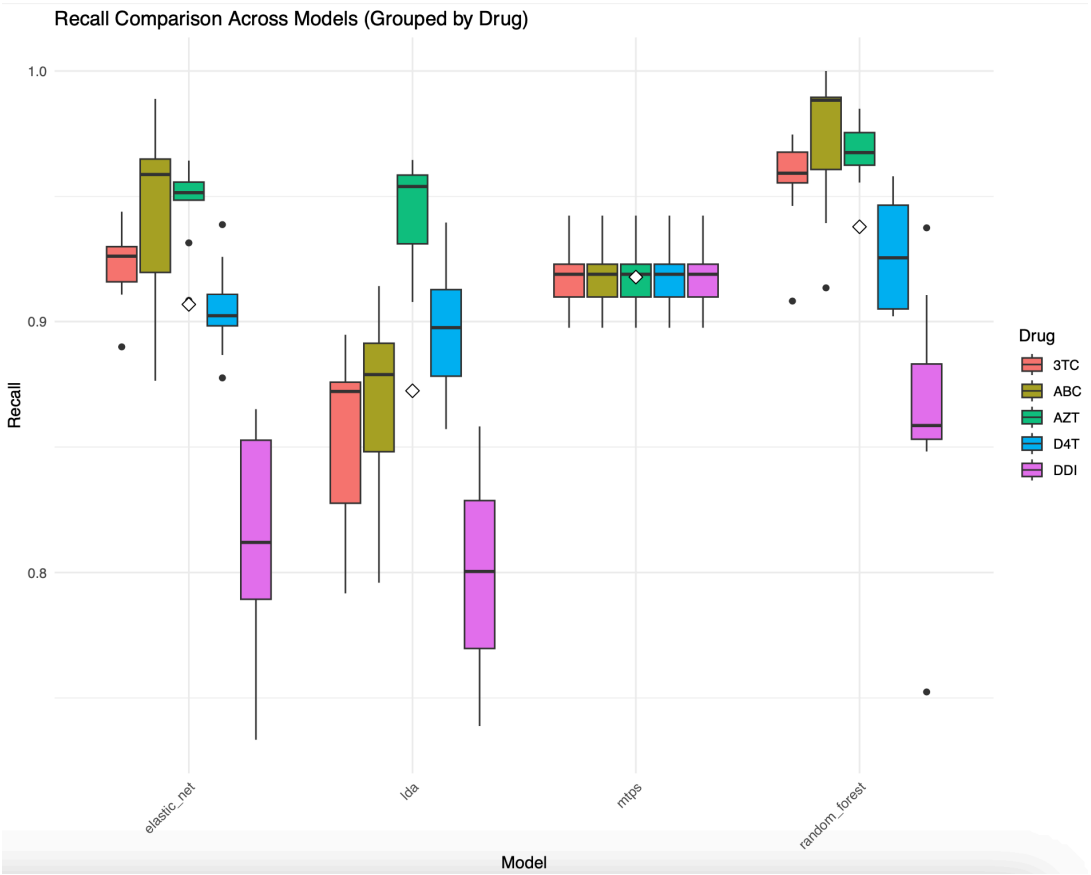
The results indicate that LDA performed similarly to MTPS for drugs ABC and 3TC, as the p-values exceed the 0.05 threshold, suggesting no significant difference in their classification performance for these drugs. Additionally, LDA performed similarly to Random Forest for AZT, D4T, and DDI, meaning that Random Forest did not provide a statistically significant improvement over LDA for these drugs. Similarly, Elastic Net performed comparably to Random Forest for ABC, D4T, AZT, and DDI, as indicated by higher p-values. This suggests that for these drugs, Elastic Net and Random Forest achieved similar classification performance and that neither model demonstrated a clear advantage. These findings highlight that while certain models outperform others in some cases, their effectiveness varies by drug. The lack of significant differences in some comparisons suggests that MTPS and LDA, as well as Elastic Net and Random Forest, may share similar decision boundaries or underlying classification mechanisms for specific drugs. On the other hand, the significant differences observed in other cases (e.g., Elastic Net vs. MTPS with consistently low p-values) suggest that Elastic Net may be a more robust model overall, particularly when MTPS struggles to achieve comparable performance.

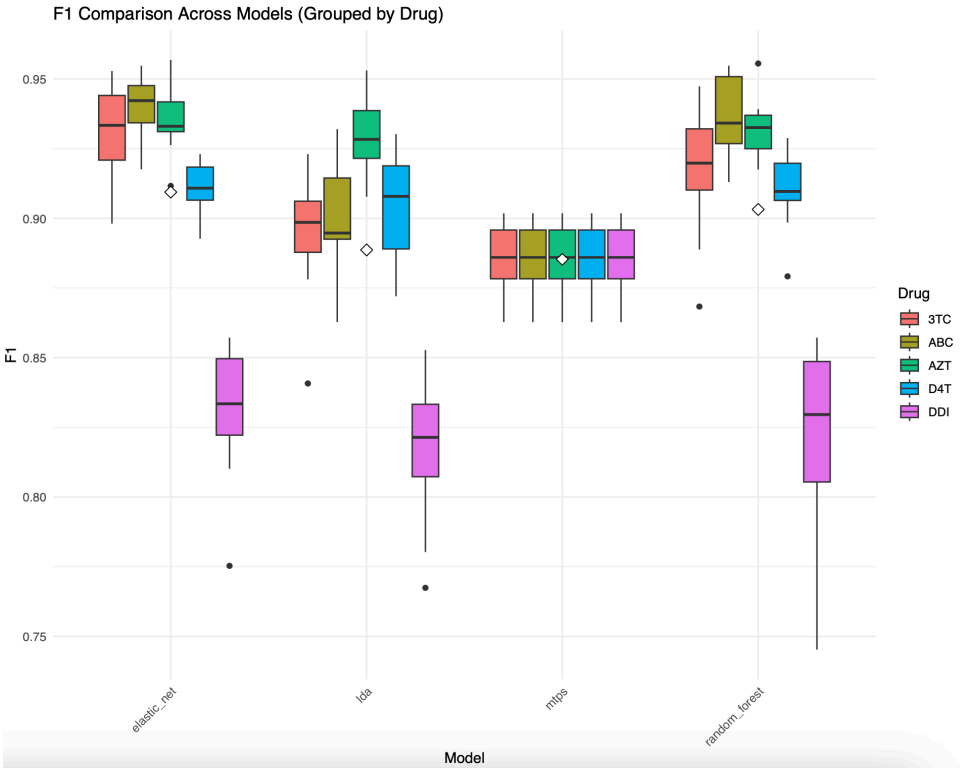
3.3 Visualization

The plots below visualize the performance metrics (Misclassification Rate, Precision, Recall, and F1-Score) of various models across the 5 drugs. Each boxplot groups the results by model and drug, allowing for a comparison of how each model performs in predicting drug resistance. The plots highlight potential correlations between drugs, such as 3TC and ABC, as well as AZT and D4T, based on similar medians in the metrics. The color-coded legend identifies the drug, making it easy to compare their performance across different models. The plots also demonstrate how MTPS had identical performance across all drugs indicating that it is a stable model. Additionally, they depict similarities across models like Random Forest and Elastic Net, LDA and Elastic Net (which was not predicted by the Wilcoxon Test).

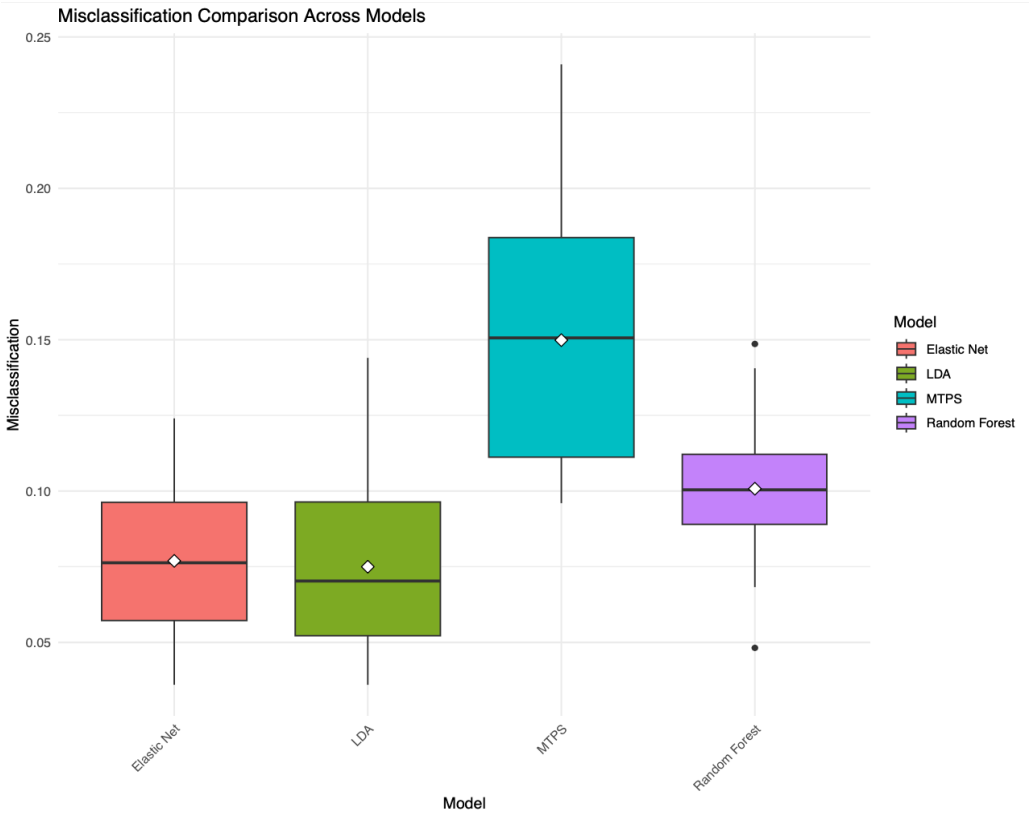


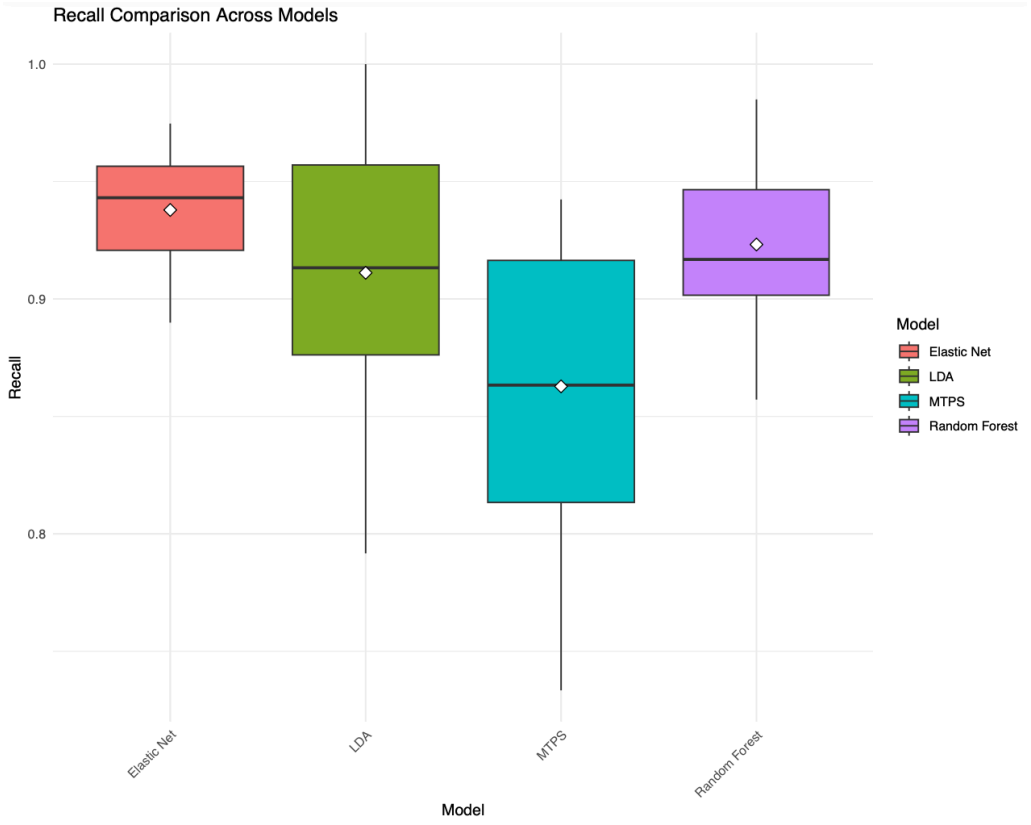

Precision Comparison Across Models (Grouped by Drug)
Precision
0.9
0.8
elastic_net
lda
mtps
random_forest
Model
Drug
3TC
ABC
AZT
D4T
DDI

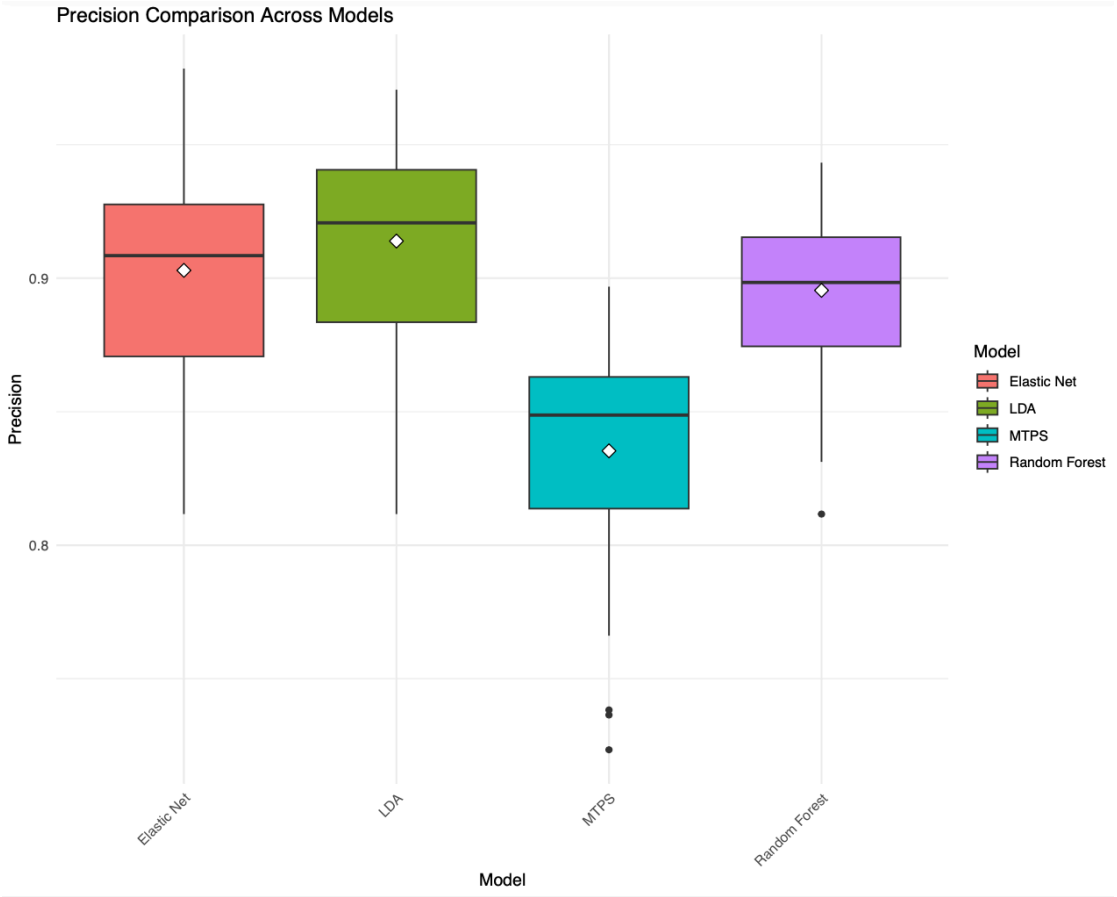

Recall Comparison Across Models (Grouped by Drug)
Recall
1.0
0.9
0.8
elastic_net
lda
mtps
random_forest
Model
Drug
3TC
ABC
AZT
D4T
DDI

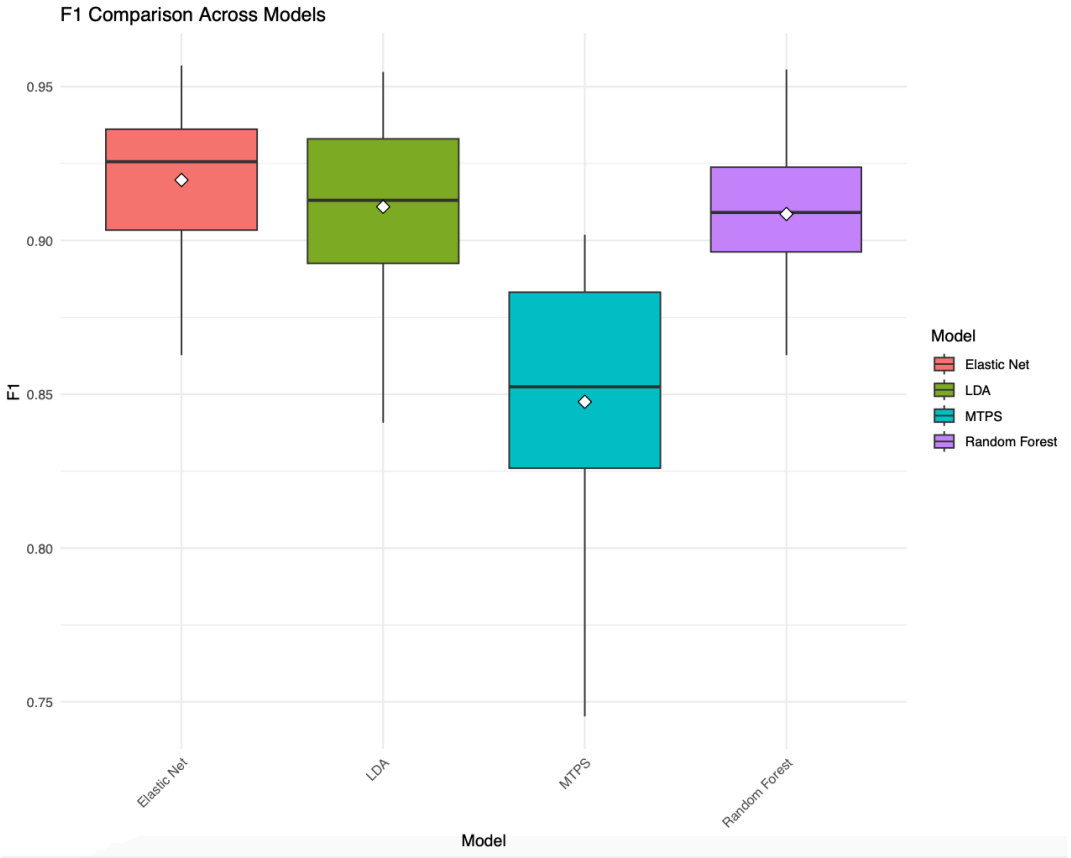


The plots below summarize the performance metrics (Misclassification Rate, Precision, Recall, and F1-Score) of all models across all drugs combined. Each plot groups the results by model, providing a comprehensive comparison of their overall performance in predicting drug resistance.




Recall Comparison Across Models
1.0
0.9
0.8
Recall
Elastic Net
LDA
MTPS
Random Forest
Model
Model
Elastic Net
LDA
MTPS
Random Forest


Precision Comparison Across Models
0.9
0.8
Precision
Elastic Net
LDA
MTPS
Random Forest
Model
Model
Elastic Net
LDA
MTPS
Random Forest



4. Possible Improvements

The performance of MTPS across all metrics was identical across all drugs, indicating a lack of variation in its classification results. This could be caused by a lack of drug-specific residual correction, or incorrect performance aggregation where results are averaged across drugs instead of being computed individually. Ensuring that MTPS is trained separately for each drug, applying drug-specific residual adjustments, and computing performance metrics at the drug level before aggregation will lead to more meaningful boxplots that reflect actual variations in model performance across different drugs

5. Conclusion

Elastic Net is the best-performing model across all metrics, demonstrating the highest F1 score and recall, making it the most effective at identifying resistant cases. LDA follows closely, achieving the lowest misclassification rate and highest precision. Random Forest performs well, but showed no significant advantage over Elastic Net for some drugs. MTPS, however, displayed identical performance across all drugs, indicating potential issues with drug-specific variation in its residual correction.

The Wilcoxon signed-rank test confirms statistically significant differences between models in most cases. However, LDA performed similarly to MTPS for ABC and 3TC, and to Random Forest for AZT, D4T, and DDI, suggesting that their classification effectiveness was comparable for these drugs. Additionally, Elastic Net and Random Forest showed no significant performance differences for ABC, D4T, AZT, and DDI, indicating they may capture similar resistance patterns.

Potential correlations between drugs were observed, such as ABC and 3TC, as well as AZT and D4T, where models exhibited similar median performance metrics. These findings suggest that certain resistance patterns may be linked across drugs, emphasizing the importance of drug-specific model evaluation and tailored classification approaches.

6. References

- Sample R code for STAT 469 Assignment 1. University of Victoria.
- Li, X., Lesperance, M. L., & Zhang, X. (2023). *Use the MTPS Package*. Retrieved from https://cran.r-project.org/web/packages/MTPS/vignettes/Guide.html