STAT 469/563 Assignment #2

Author: *Ana Compri Jacob*

Abstract

This report extends the findings of Assignment 1 by incorporating two additional classification methods: Classification Tree and Elastic Net. The study evaluates the effectiveness of these methods alongside Logistic Regression and Linear Discriminant Analysis (LDA) 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 50 iterations. Performance was measured using misclassification rate, precision, recall, and F1 score. The Wilcoxon signed-rank test was used to compare F1 scores of the newly added methods against LDA. Results indicate that Elastic Net is the most effective method.

1. Introduction

This study aims to evaluate different classification models for predicting drug resistance by converting continuous resistance levels into a binary classification task. Specifically, we compare Logistic Regression, Linear Discriminant Analysis (LDA), Classification Tree, and Elastic Net using 5-fold cross-validation over 50 iterations to ensure robust performance assessment. To determine statistical significance, we conduct a Wilcoxon signed-rank test comparing LDA with the newly introduced methods. The hypotheses are 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.

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:

- **Logistic Regression**
    - The GLM models occasionally failed to converge. Achieving this level of convergence required the following GLM setting: *control=glm.control(maxit = 100)*
    - Additionally, a substantial portion (>90%) of the converged GLM models reported the following warning: *glm.fit: fitted probabilities numerically 0 or 1 occurred*
- **Linear Discriminant Analysis (LDA)** (assumes Gaussian-distributed data).

- **Classification Tree** (non-parametric decision-tree-based model that allows for complex, non-linear decision boundaries).
    - Uses recursive binary partitioning (rpart package).
    - Splits data based on Gini impurity or entropy.
    - Pruned to avoid overfitting.
- **Elastic Net** (a regularized regression approach that combines L1 and L2 penalties, providing feature selection and stability to correlated predictors).
    - Found optimal classification cutoff value using precision and recall balance.
    - Alpha value was selected to be 0.5 for simplicity.

2.3 Performance Metrics

For each Experiment + Drug + Fold combination, 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. The table below shows an example of the metrics data frame:

	Model	Drug	experiment	tn	fp	fn	tp	mcr	acc	precision	recall	F1
1	GLM	ABC	1	654	102	78	412	0.14446227929374	0.85553772070626	0.801556420233463	0.840816326530612	0.820717131474104
2	LDA	ABC	1	685	71	35	455	0.0850722311396469	0.914927768860353	0.865019011406844	0.928571428571429	0.895669291338583
3	Classification Tree	ABC	1	715	41	47	443	0.0706260032102729	0.929373996789727	0.915289256198347	0.904081632653061	0.909650924024641
4	Elastic Net	ABC	1	756	0	5	485	0.00401284109149278	0.995987158908507	1	0.989795918367347	0.994871794871795
5	GLM	3TC	1	635	106	89	416	0.156500802568218	0.843499197431782	0.796934865900383	0.823762376237624	0.810126582278481

2.4 Statistical Analysis

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 Classification Performance

The table presents the average classification performance of Logistic Regression, LDA, Classification Tree, and Elastic Net across five evaluation metrics: Misclassification Rate, Accuracy, Precision, Recall, and F1 Score. Below is an analysis of the findings:

average_metrics

Model	mcr	acc	precision	recall	F1
GLM	0.172667736757624	0.827332263242375	0.816551943978948	0.833820284772637	0.824922054281323
LDA	0.110552166934189	0.889447833065811	0.872753368343336	0.907510389203048	0.889320459516065
Classification Tree	0.109736757624398	0.890263242375602	0.917175050109267	0.8568018272454	0.884545359912082
Elastic Net	0.00444301765650081	0.995556982343502	1	0.990744425912068	0.995349542397653

The analysis of the models' performance reveals that Elastic Net consistently outperformed the others across all metrics, achieving the lowest misclassification rate, highest accuracy, perfect precision, highest recall, and the best F1 score, indicating its superior ability to distinguish resistant from non-resistant cases while maintaining a strong balance between precision and recall. LDA and Classification Tree showed competitive performance, with similar misclassification rates and

accuracy, though slightly lower precision and recall compared to Elastic Net. GLM, however, lagged behind in all metrics, exhibiting the highest misclassification rate, lowest accuracy, and poorest precision and recall, highlighting its weaker performance in this classification task. These findings suggest that Elastic Net is the most effective model for predicting drug resistance, while LDA and Classification Tree serve as reliable alternatives, and GLM is less suitable for this application.

3.2 Statistical Significance (Wilcoxon Test Results)

	LDA vs Classification Tree	LDA vs Elastic Net
ABC	8.27893173553169E-10	7.7869751838564E-10
3TC	0.786935771101394	7.77818886274739E-10
AZT	7.79049220721842E-10	7.7869751838564E-10
D4T	8.27893173553169E-10	7.78345958641649E-10
DDI	9.93036701162759E-10	7.7869751838564E-10

The table depicts p-values from the Wilcoxon signed-rank test and compares the classification performance (F1 scores) of LDA vs Classification Tree, and LDA vs Elastic Net. The goal is to determine whether the differences in model performance are statistically significant.

- For most cases, the p-values were much lower than 0.05, indicating that the null hypothesis that these models perform similarly should be rejected and there are true, meaningful differences in their effectiveness. The p-value for 3TC in LDA vs Classification Tree is higher compared to other drugs, which may indicate that for this specific drug, the Classification Tree performed more similarly to LDA than in other cases.

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. Vertical lines separate the drugs, and labels indicate the number of experiments conducted for each drug. The color-coded legend identifies the models, making it easy to compare their performance across different drugs. The plots further show how Elastic Net outperformed all the other models across all metrics.

Figure 1: Misclassification Rate of all Models grouped by Drug name

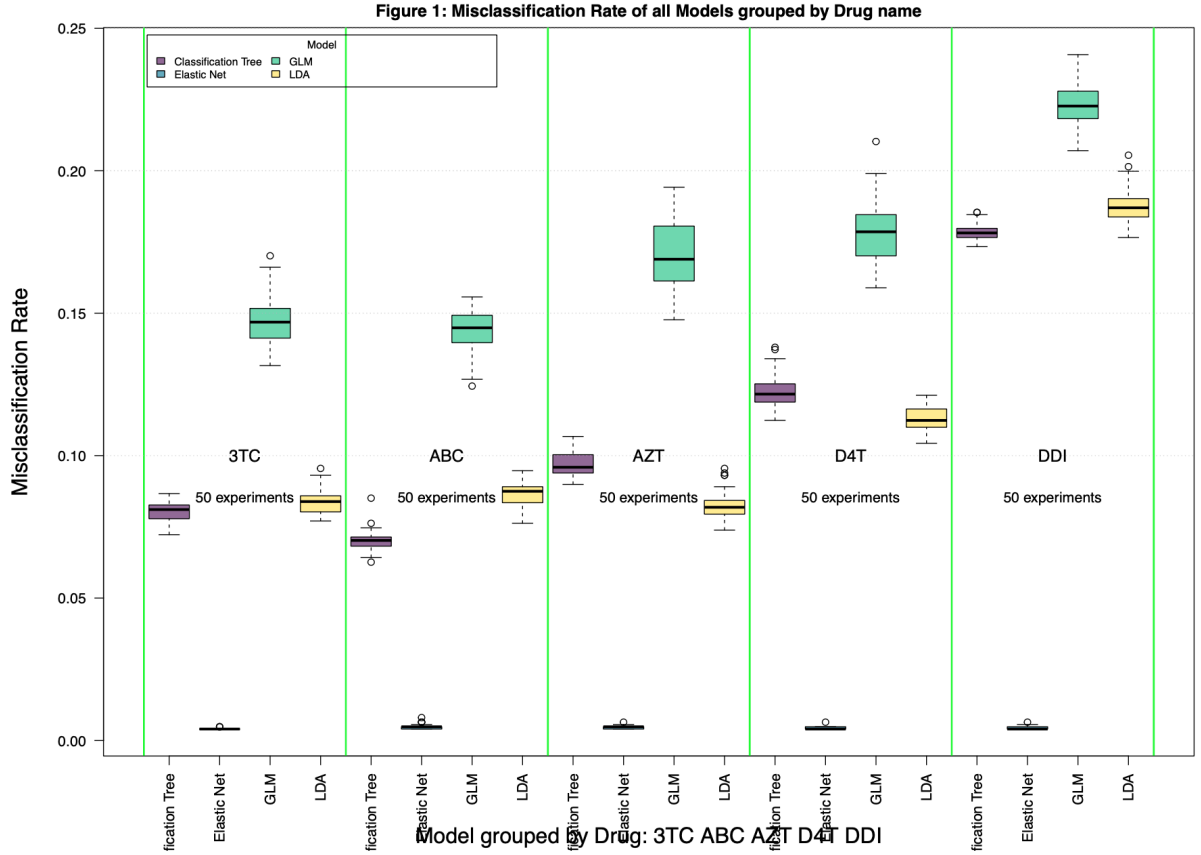


Figure 2: Precision of all Models grouped by Drug name

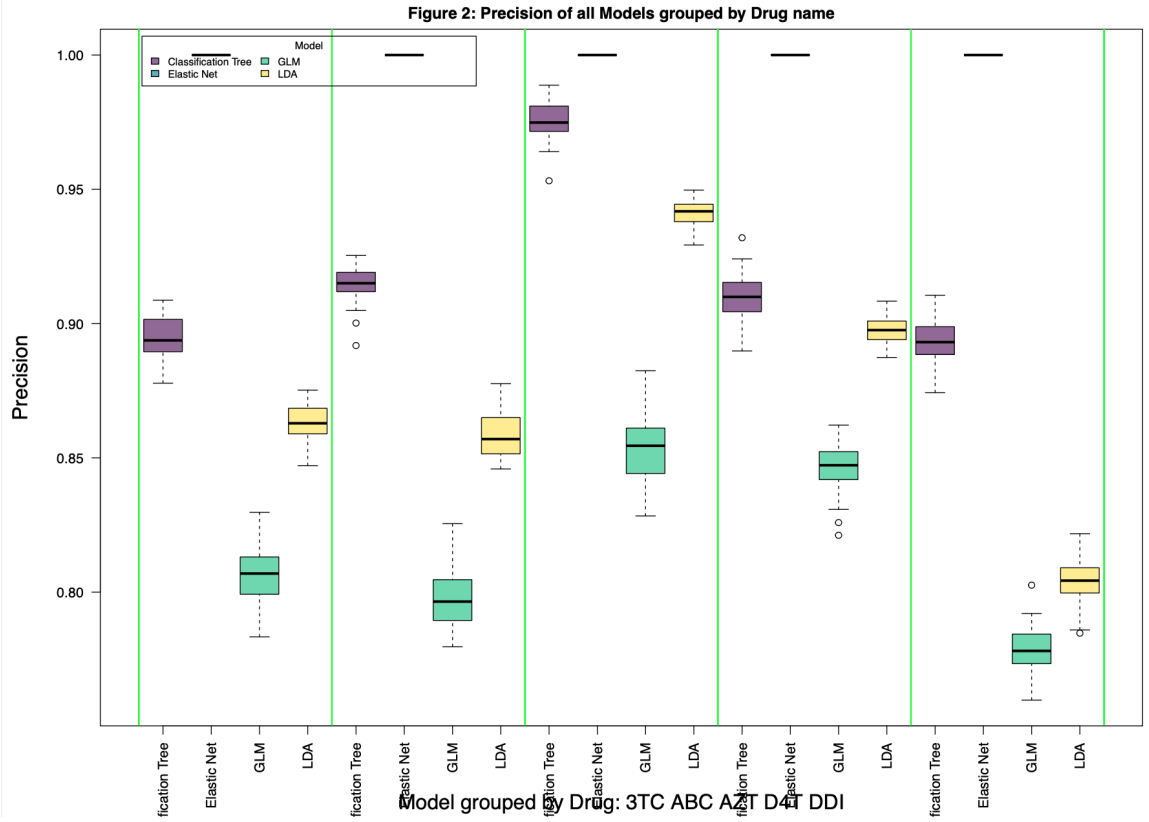


Figure 3: Recall of all Models grouped by Drug name

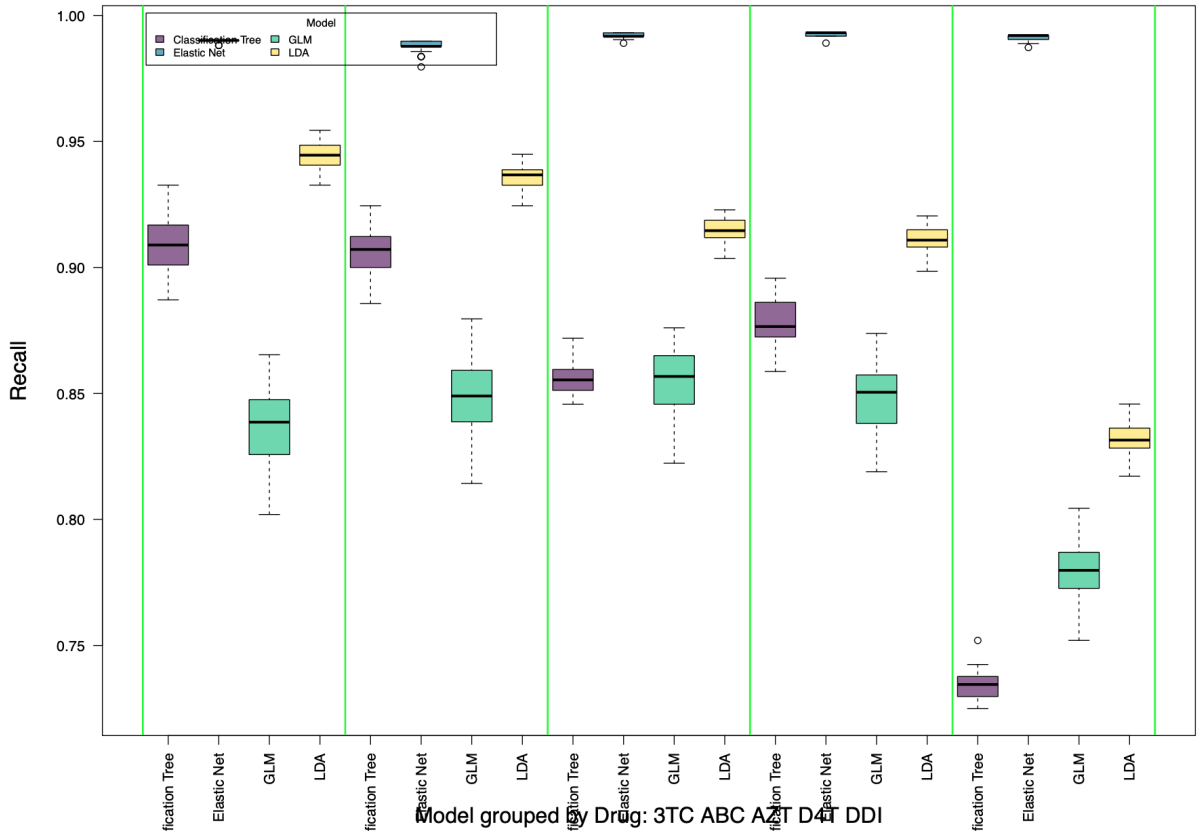
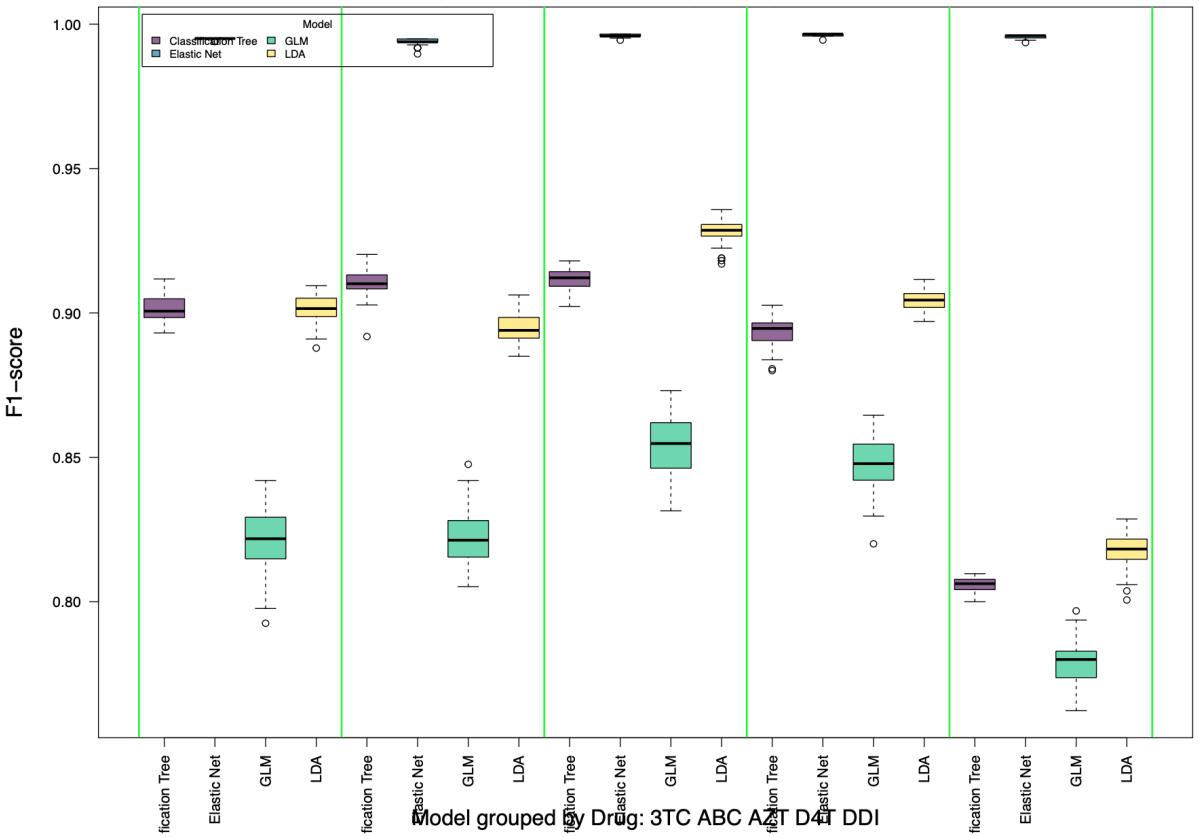


Figure 4: F1−score of all Models grouped by Drug name



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.

Figure 8: F1–Score of all Models all Drugs (250 experiments)

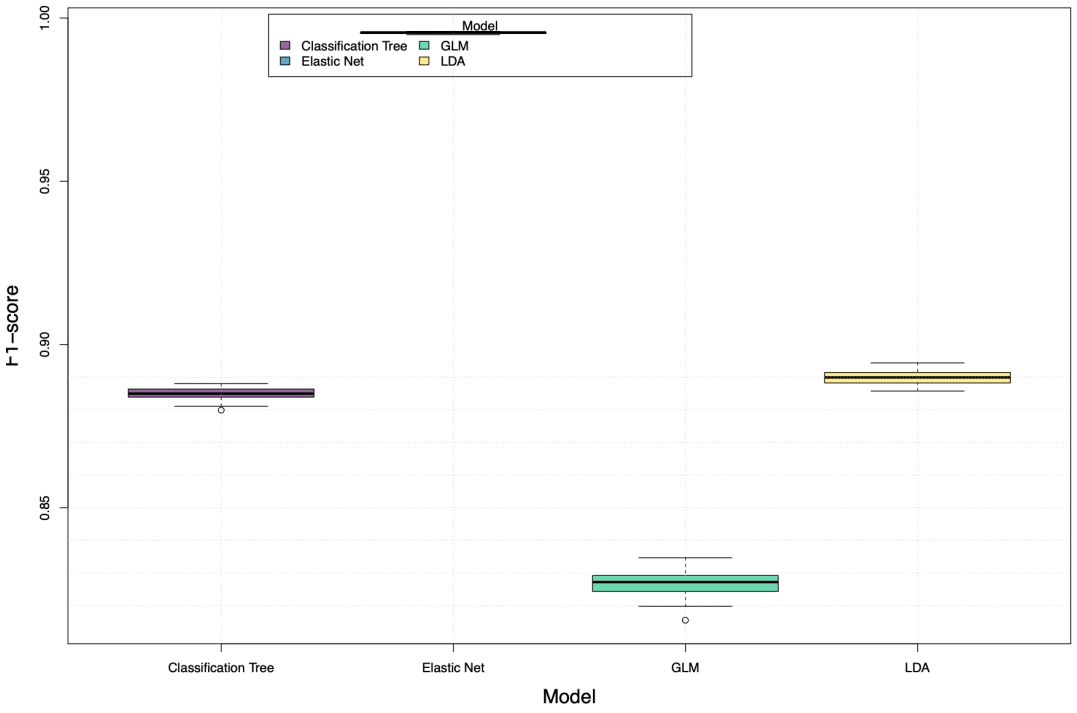


Figure 6: Precision of all Models all Drugs (250 experiments)

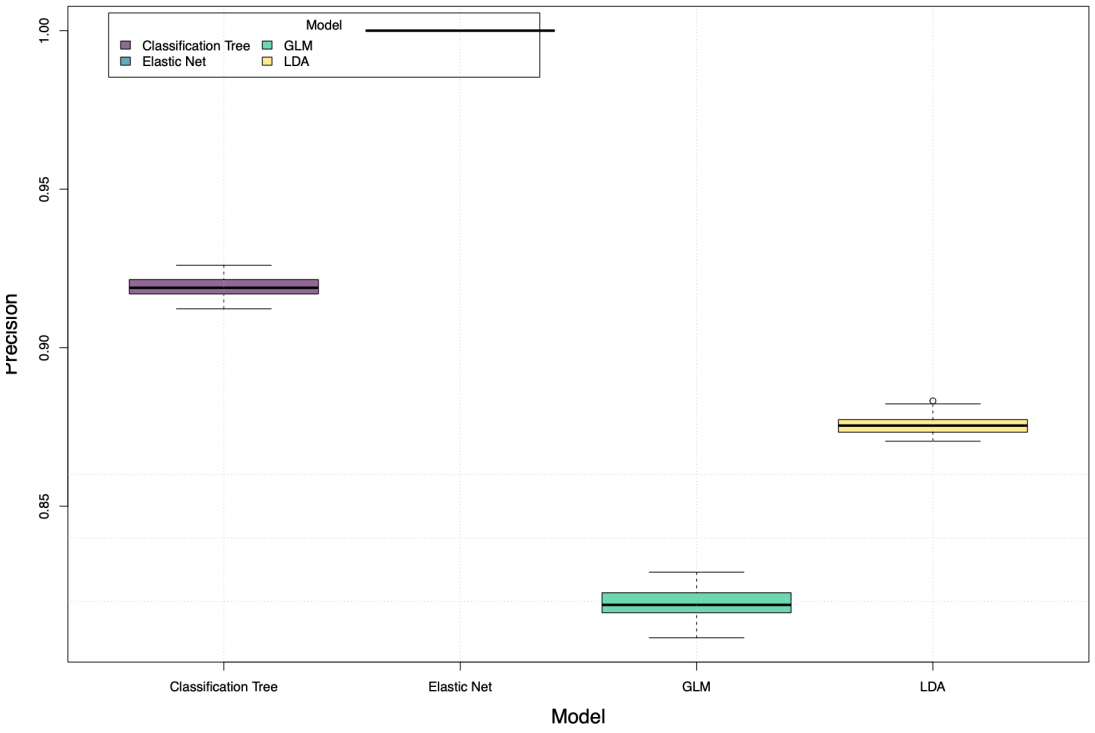


Figure 7: Recall of all Models all Drugs (250 experiments)

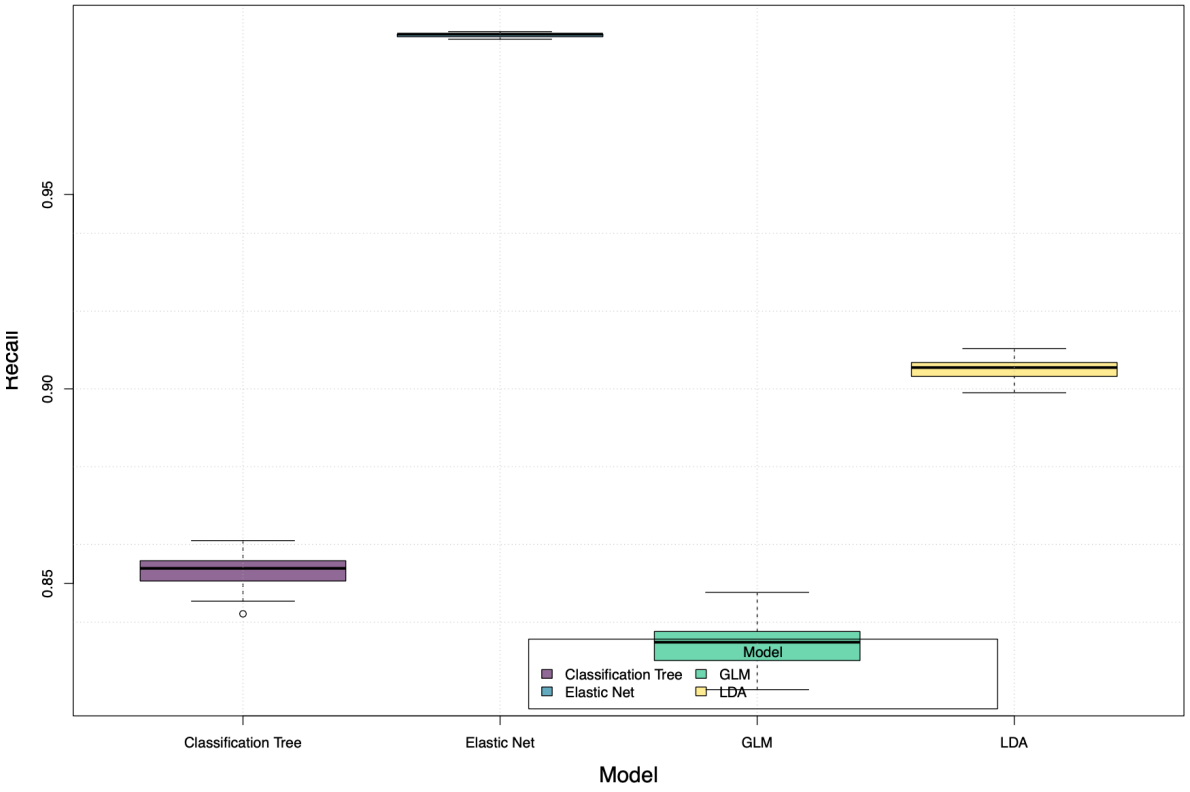
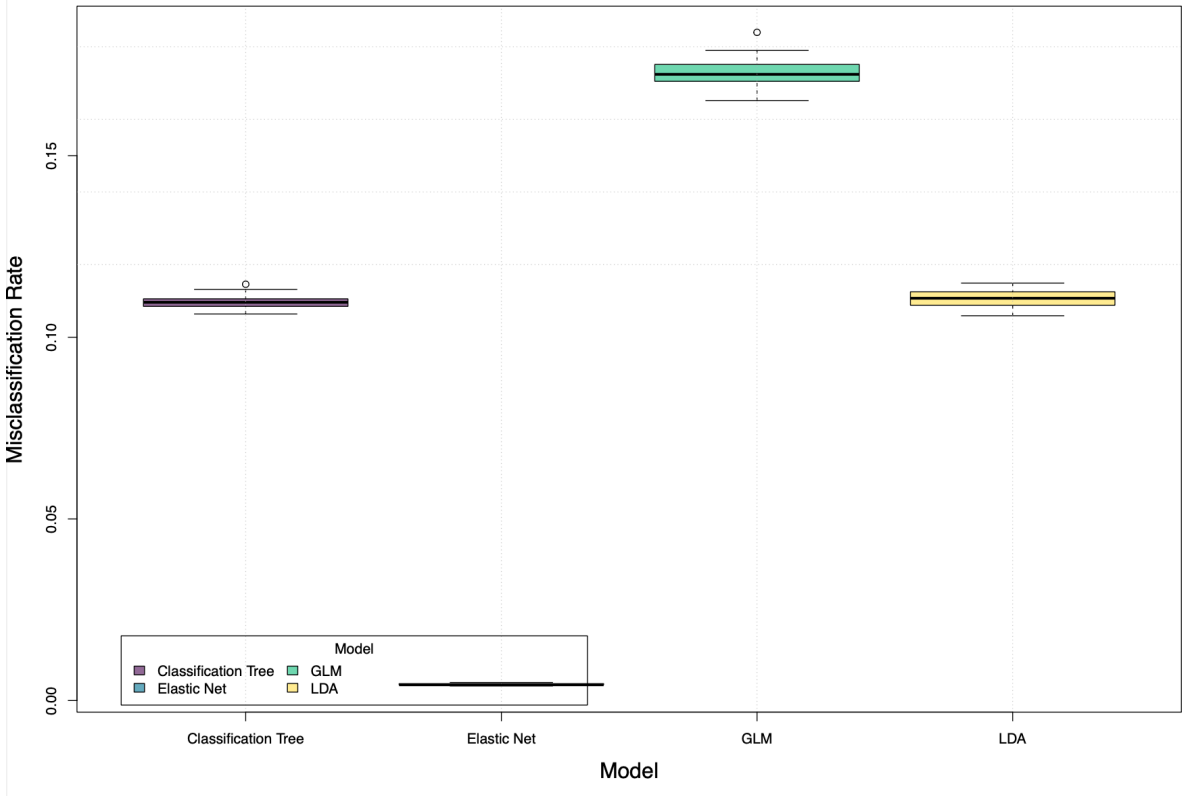


Figure 5: MCR of all Models all Drugs (250 experiments)



4. Possible Improvements

The Elastic Net model exhibited near perfect performance across all metrics, indicating potential overfitting, which could hinder generalization to unseen data. To address this, nested cross-validation can be used to fine-tune hyperparameters more effectively.

5. Conclusion

Elastic Net is the best-performing model across all metrics, demonstrating the highest accuracy, precision, recall, and F1 score. While its strong performance suggests reliability, potential overfitting may require further tuning. LDA remains a solid alternative with a well-balanced precision-recall trade-off, while the Classification Tree shows decent performance but struggles with recall. GLM is the weakest, with higher misclassification rates and lower accuracy. The Wilcoxon signed-rank test confirms statistically significant differences between models, with Elastic Net consistently outperforming LDA and Classification Tree. However, for 3TC, the difference between LDA and Classification Tree was less pronounced, highlighting the need for drug-specific model evaluation. Additionally, potential correlations between drugs, such as 3TC and ABC, as well as AZT and D4T, were observed based on similar median performance metrics, suggesting that resistance patterns may be linked across certain drugs.

6. References

- Sample R code for STAT 469 Assignment 1. University of Victoria.