Ian Sam
V00990790

STAT 469 Assignment #1

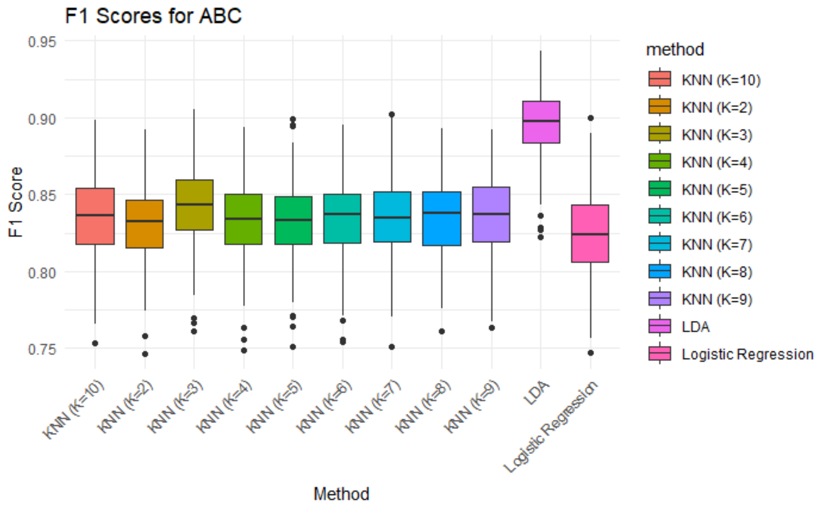
Introduction

This report analyzes the performance of three predictive models: Logistic Regression, Linear Discriminant Analysis (LDA), and k-Nearest Neighbors (KNN) in predicting binary resistance outcomes for five HIV drugs (ABC, 3TC, AZT, D4T, and DDI). The analysis employs 50 random splits with 5-fold stratified cross-validation, to maintain a balanced representation of resistant and non-resistant cases within each fold. The model's effectiveness is evaluated using F1-scores, recall, precision, and the misclassification rates. Furthermore, Wilcoxon tests are conducted to determine whether the observed differences in F1-scores are statistically significant, providing a comprehensive assessment of the comparative performance of the three models across all drugs.

Methods

For our 3 models Logistic Regression, LDA, and KNN a stratified 5-fold cross-validation was employed which was repeated across 50 random splits. A matrix was created to predefine fold indices. Each model was evaluated on the same training and testing splits. Logistic Regression was implemented using the glm() function. LDA was applied using the lda() function, while KNN was evaluated for various k values (from 2 to 10) using the knn() function. Performance metrics including misclassification rate, precision, recall, and F1-score were calculated using a function based on confusion matrix outputs. The results for each model and drug were stored in a structured list, and summary statistics such as median F1-scores were computed. Additionally, boxplots were generated for each drug to visualize F1-score distributions across methods. Finally, Wilcoxon tests were performed to assess the statistical significance of differences in F1-scores between Logistic Regression, LDA, and the best-performing KNN configuration.

Drug ABC



A tibble: 11 × 5

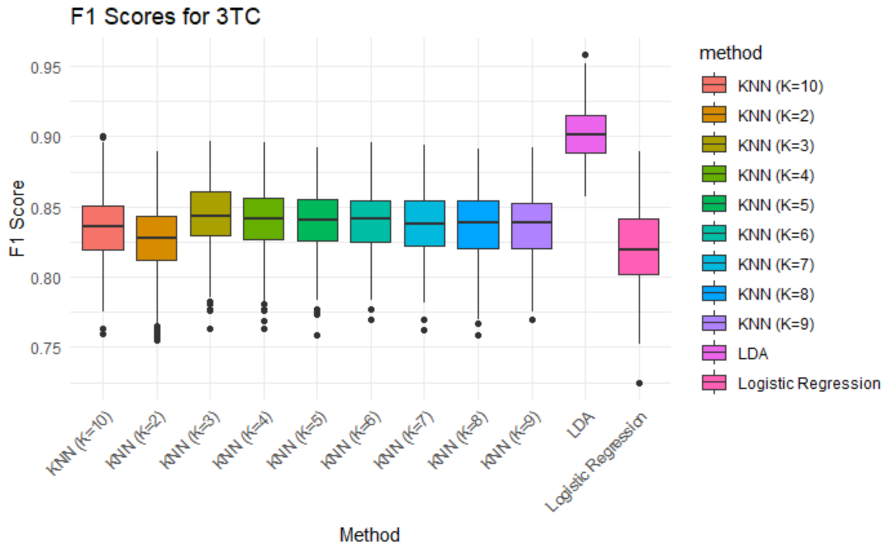
method <chr>	misclassification_rate <dbl>	precision <dbl>	recall <dbl>	f1_score <dbl>
KNN (K=10)	0.14112903	0.7674419	0.9263158	0.8362853
KNN (K=2)	0.14800000	0.7556404	0.9296465	0.8322188
KNN (K=3)	0.13600000	0.7673416	0.9393939	0.8430493
KNN (K=4)	0.14400000	0.7640036	0.9255319	0.8341232
KNN (K=5)	0.14428916	0.7644284	0.9207921	0.8329563
KNN (K=6)	0.14056225	0.7675493	0.9183673	0.8372093
KNN (K=7)	0.14112903	0.7692308	0.9170489	0.8350015
KNN (K=8)	0.14000000	0.7727335	0.9179466	0.8382738
KNN (K=9)	0.14056225	0.7723175	0.9200000	0.8372093
LDA	0.08433735	0.8611111	0.9393939	0.8975610
Logistic Regression	0.14112903	0.8019335	0.8465463	0.8241758

```
Drug: ABC
Logistic Regression vs LDA: p-value = 9.310216e-43
Logistic Regression vs KNN: p-value = 5.534785e-06
LDA vs KNN: p-value = 1.088691e-42
```

Results

For drug ABC, the results indicate that LDA is the best-performing method, achieving the highest median F1 score of (0.898) and the lowest misclassification rate of (0.084). Logistic Regression performed well, with a median F1 score of (0.824) and a misclassification rate of (0.141). While it exhibits low variability, it falls short of LDA in terms of precision (0.802 vs. 0.861) and recall (0.847 vs. 0.939). KNN, on the other hand, shows considerable variability depending on the value of k. The best KNN (k=3), achieved an F1 score of 0.843 and a misclassification rate of 0.136, which is similar with Logistic Regression but still inferior to LDA. The Wilcoxon tests confirm these findings with significant differences observed between Logistic Regression and LDA (p-value = 9.31e-43) and between LDA and KNN (p-value = 1.08e-42) indicating LDA is statistically better. Therefore, LDA is the most effective method for drug ABC, followed by Logistic Regression and KNN.

Drug 3TC



A tibble:11 × 5

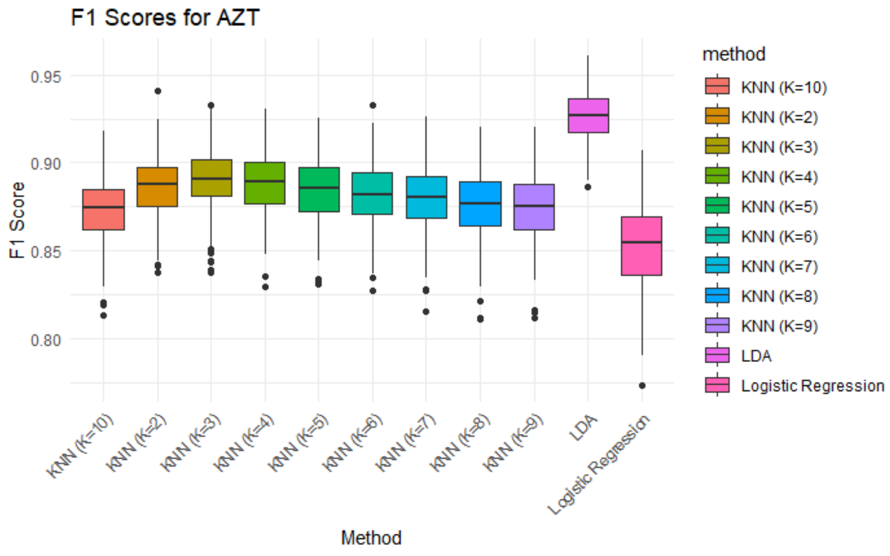
method <chr>	misclassification_rate <dbl>	precision <dbl>	recall <dbl>	f1_score <dbl>
KNN (K=10)	0.14800000	0.7615032	0.9292929	0.8356182
KNN (K=2)	0.15200000	0.7592593	0.9052632	0.8272727
KNN (K=3)	0.13709677	0.7796610	0.9215686	0.8433696
KNN (K=4)	0.14056225	0.7764298	0.9234432	0.8411215
KNN (K=5)	0.14400000	0.7721438	0.9227035	0.8403361
KNN (K=6)	0.14056225	0.7740533	0.9257289	0.8411215
KNN (K=7)	0.14400000	0.7714200	0.9278351	0.8377223
KNN (K=8)	0.14457831	0.7672414	0.9278351	0.8385688
KNN (K=9)	0.14457831	0.7654916	0.9292929	0.8384279
LDA	0.08433735	0.8644068	0.9484536	0.9015971
Logistic Regression	0.14859438	0.8049745	0.8398113	0.8196660

```
Drug: 3TC
Logistic Regression vs LDA: p-value = 9.3103e-43
Logistic Regression vs KNN: p-value = 1.210147e-15
LDA vs KNN: p-value = 1.025123e-42
```

Results

For drug 3TC, the results indicate that LDA is the best-performing method with an F1 score of (0.902) and the lowest misclassification rate of (0.084). In comparison, Logistic Regression performs well but still loses to LDA, with an F1 score of (0.820), a higher misclassification rate (0.149), and slightly lower precision (0.805) and recall (0.839). Among the KNN configurations, KNN (k=3) achieves the highest F1 score (0.843) with a misclassification rate of 0.138, followed by KNN (k=4) and KNN (k=5). However, as k increases beyond 3, KNN shows diminishing returns, with only small improvements in precision and recall but no significant gain in overall performance. The Wilcoxon test results indicate significant differences in F1 scores between Logistic Regression and LDA (p-value = 9.31e-43) and between Logistic Regression and KNN (p-value = 1.21e-15). Similarly, LDA significantly outperforms KNN (p-value = 1.03e-42). In conclusion, LDA proves to be the best model for drug 3TC, followed by Logistic Regression then KNN.

Drug AZT



A tibble:11 × 5

method <chr>	misclassification_rate <dbl>	precision <dbl>	recall <dbl>	f1_score <dbl>
KNN (K=10)	0.1606426	0.8022617	0.9589041	0.8748089
KNN (K=2)	0.1400000	0.8393959	0.9415560	0.8877876
KNN (K=3)	0.1327973	0.8428953	0.9470199	0.8910256
KNN (K=4)	0.1368215	0.8374247	0.9490902	0.8888889
KNN (K=5)	0.1442892	0.8294118	0.9503546	0.8853503
KNN (K=6)	0.1480000	0.8208985	0.9522179	0.8819876
KNN (K=7)	0.1520000	0.8165119	0.9559286	0.8805031
KNN (K=8)	0.1560000	0.8075729	0.9566781	0.8767975
KNN (K=9)	0.1586290	0.8052962	0.9580420	0.8750000
LDA	0.0840000	0.9418174	0.9151933	0.9271599

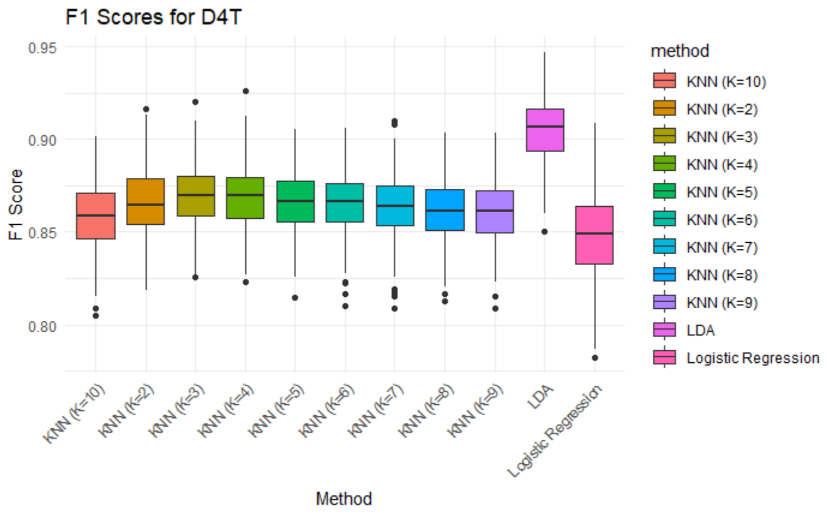
Logistic Regression	0.1686747	0.8541667	0.8536568	0.8542710

```
Drug: AZT
Logistic Regression vs LDA: p-value = 9.3103e-43
Logistic Regression vs KNN: p-value = 3.320514e-38
LDA vs KNN: p-value = 1.242652e-42
```

Results

For drug AZT, LDA has the best performance, achieving the highest F1 score (0.928), precision (0.942), recall (0.916) and the lowest misclassification rate (0.084). Logistic Regression has an high F1 score of 0.854 but has a significantly higher misclassification rate (0.169). Furthermore, it has a lower precision (0.854) and recall (0.854) compared to LDA. Among the KNN configurations, KNN (k=3) delivers the best results, with an F1 score of 0.891 and a misclassification rate of (0.133), making it competitive but still inferior to LDA. Increasing k stabilizes performance, but no k-value achieves results comparable to LDA. KNN (k=4) and KNN (k=5) have reasonable F1 scores (0.889 and 0.885, respectively) but are still outperformed by both LDA and Logistic Regression. The Wilcoxon test confirms the superiority of LDA, with a highly significant difference in F1 scores compared to both Logistic Regression (p-value = 9.31e-43) and KNN (p-value = 1.24e-42). Overall, LDA is the best model for predicting resistance in drug AZT, followed by KNN (k=3) and Logistic Regression.

Drug D4T



A tibble:11 × 5

method <chr>	misclassification_rate <dbl>	precision <dbl>	recall <dbl>	f1_score <dbl>
KNN (K=10)	0.1880000	0.7717391	0.9720280	0.8588773

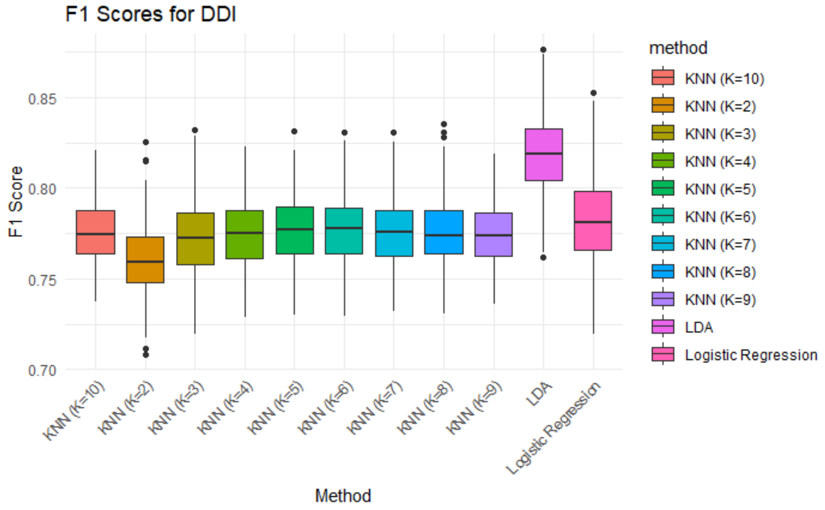
KNN (K=2)	0.1693548	0.8060606	0.9379310	0.8647757
KNN (K=3)	0.1680000	0.8057983	0.9469948	0.8699856
KNN (K=4)	0.1683373	0.8000000	0.9533333	0.8699768
KNN (K=5)	0.1726908	0.7908291	0.9581876	0.8665678
KNN (K=6)	0.1767068	0.7886957	0.9626783	0.8662562
KNN (K=7)	0.1800000	0.7811598	0.9652778	0.8637708
KNN (K=8)	0.1810873	0.7780781	0.9667770	0.8615868
KNN (K=9)	0.1840000	0.7737469	0.9712230	0.8615868
LDA	0.1120000	0.8989646	0.9128976	0.9064707
Logistic Regression	0.1774194	0.8480510	0.8482759	0.8486797

```
Drug: D4T
Logistic Regression vs LDA: p-value = 9.423058e-43
Logistic Regression vs KNN: p-value = 3.600221e-27
LDA vs KNN: p-value = 3.525982e-42
```

Results

For drug D4T, LDA demonstrates the best performance among all methods, with the highest F1 score (0.907), precision (0.899), recall (0.913) and the lowest misclassification rate (0.112). Logistic Regression follows with a decent F1 score (0.849) and a misclassification rate (0.177), but its lower precision (0.849) and recall (0.849) indicate it is less effective then LDA. Among the KNN configurations, KNN (k=4) has the best scores with an F1 score (0.870) and a misclassification rate (0.168). As k increases beyond 4, KNN shows diminishing returns, with higher misclassification rates and decreasing precision. The Wilcoxon test results show significant differences in F1 scores between Logistic Regression and LDA (p-value = 9.42e-43), as well as between Logistic Regression and KNN (p-value = 3.60e-27). Similarly, LDA significantly outperforms KNN (p-value = 3.53e-42). Overall, LDA is the most effective method for drug D4T, followed by Logistic Regression, then KNN.

Drug DDI



A tibble:11 × 5

method	misclassification_rate	precision	recall	f1_score
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
KNN (K=10)	0.2680000	0.6745562	0.9126984	0.7743022
KNN (K=2)	0.2741935	0.6792453	0.8650794	0.7591000
KNN (K=3)	0.2640000	0.6851852	0.8880000	0.7724885
KNN (K=4)	0.2610442	0.6856244	0.8960000	0.7752090
KNN (K=5)	0.2570281	0.6876806	0.8968254	0.7772061
KNN (K=6)	0.2605221	0.6845238	0.8968254	0.7775932
KNN (K=7)	0.2640000	0.6807229	0.9047619	0.7758007
KNN (K=8)	0.2661290	0.6777988	0.9047619	0.7738476
KNN (K=9)	0.2661290	0.6749233	0.9126984	0.7738715
LDA	0.1867419	0.8049814	0.8333333	0.8187152
Logistic Regression	0.2200000	0.7829457	0.7857143	0.7810632

```
Drug: DDI
Logistic Regression vs LDA: p-value = 4.926246e-42
Logistic Regression vs KNN: p-value = 0.0004008831
LDA vs KNN: p-value = 1.56197e-41
```

Results

For drug DDI, LDA achieves the best overall performance, with the highest F1 score (0.818), precision (0.805), recall (0.833) and the lowest misclassification rate (0.186). Logistic Regression follows with a lower F1 score of (0.781) and a higher misclassification rate of (0.220). Its precision (0.782) and recall (0.785) falls behind LDA. For KNN, the performance across different k-values is relatively stable but falls short of both LDA and Logistic Regression. Among KNN configurations, KNN (k=5) achieves the highest F1 score (0.777) and a relatively low misclassification rate (0.257). However, the precision and recall for KNN are generally lower across all k-values compared to LDA. The Wilcoxon test results confirm significant differences in F1 scores between LDA and Logistic Regression (p-value = 4.93e-42) and between Logistic Regression and KNN (p-value = 0.0004). Similarly, LDA significantly outperforms KNN (p-value = 1.56e-41). In conclusion LDA is the best model for predicting resistance in drug DDI, with Logistic Regression coming second then KNN.

Conclusion

Overall, LDA consistently outperforms both Logistic Regression and KNN across all five drugs, achieving the highest F1-scores, highest precision, highest recall and lowest misclassification rates in most cases. Logistic Regression performed consistently well as a secondary model, delivering competitive F1-scores but generally lower than LDA, particularly for drugs AZT and DDI. KNN showed significant variability depending on the choice of k, with KNN (k=3 or k=4) often achieving competitive results. However, it failed to outperform LDA or Logistic Regression overall. Wilcoxon tests confirmed significant differences in F1-scores between LDA and the other methods across all drugs. These results indicate LDA's as the most effective classification method for predicting binary resistance outcomes with Logistic Regression as the next best and KNN's as the least effective model.