Assignment #3

Belina Jang
Student Number: V00924202
Dr. Xuekui Zhang
Stat 469 A01
November 13, 2022

First, YY is converted into matrix then using the provided cut-off values for each drug to convert the matrix YY into a matrix of 1 and 0 by comparing the resistance with their corresponding cut-off values where 1 indicates the sample is not resistant to the corresponding column of drug since the resistance is smaller than the cut-off and 0 indicates the sample is resistant to the drug since the resistance is greater than the cut-off. Then the format of the matrix is changed into data frame.

Then the n-fold is set to be 5. Then in order to do the stratified random sampling with 5 folds, K means clustering method was used. It randomly generates a list of 1 to 5 for each sample within each cluster. Then the indices for 50 experiments are saved into a matrix to be used for each drug. So, all i-th experiment of different drug will use same sample data.

Then the format of the data is changed into dataframe so it's easier to work with. A matrix is created to store the confusion matrix values from different models for all drugs and experiments. Then a for loop is used for each experiment of MTPS model, then it creates a matrix to store confusion matrix values from each experiment, then there's another for loop for each fold. Within each fold, it fits a MTPS model and make predictions with the model then it evaluates the predictions and calculated values (tn, fp, fn, tp) for the confusion matrix. For same ith experiment it accumulates the confusion matrix values together. After all experiments, it passes those values into another matrix (metrics) to store good results. Then another for loop is used for each experiment of LDA, Elastic net, and random forest models. Then another for loop inside for each drug since those models doesn't predict for all drugs at once like MTPS. So for each i-th experiment (first loop), it will loop through all the drugs then move onto next (i+1)-th experiment. Then for each drug, it creates a matrix to store confusion matrix values from each experiment of each drug together. Then the predictor matrix XXdf is combined with a column of the drug from yBin. Then another loop for 5 folds is added. For each fold, it declares test and train data using the indices matrix created earlier to do stratified random sampling. I modelled and made predictions for LDA, Elastic net, and random forest models with the same train and test data. Then I evaluated the predictions and calculated values for the confusion matrix for each model. Repeat it 5 times for 5 folds then add them together for each model, drug and experiment. Then those values are saved into the matrix (metrics) used earlier to store results from MTPS. Then this repeats for each drug and each experiment.

To learn the uncertainties in my models, the experiment is repeated 10 times with different train and test data. Then I plotted different graphs to compare models in different drug. The resulting matrix (metrics) will have 200rows (10 experiments*5drugs*4models). Then it creates other columns namely mcr, acc, precision, recall, and F1 to calculate and save misclassification rate, accuracy, precision, recall, and F1 score respectively.

I used wilcox.test to see if the differences between those models are significant to justify a certain model is the best model for each drug and which drug is the best model in average for all drugs.

From figure 1-4 below, we can observe which model was the best model in terms of different criteria for each drug.

Figure 1 shows misclassification rate of different models for 5 different drugs. The lower the misclassification rate, the higher the accuracy of the model. Elastic net was the lowest in one, second lowest in one, a tie for the lowest in 2 drugs and third in DDI. LDA had the highest misclassification rate in 4 out of 5 drugs and second highest in one. MTPS was the highest in one drug, second highest in 2 drugs, had a better rate in D4T and lowest misclassification rate in DDI. Random forest was either first or second across all drugs. So, the models were not much consistent across different drugs but in overall random forest had the best performance across all the drugs in terms of misclassification rate.

Figure 2 shows precision of different models for 5 different drugs. The higher precision, the better the model performance is. Elastic net was second in 3TC and third in all other drugs. LDA was the lowest in all drugs. MTPS was highest in 3 drugs and second in ABC and third in 3TC. Random forest was second in 3 drugs and highest in 2 drugs. So, in general, the Random Forest and MTPS had the best performance in terms of precision.

Figure 3 shows recall of different models for 5 different drugs. The higher the recall, the better the model is. Elastic net was highest in 3TC and ABC, second in AZT and DDI, and was a tie in D4T. LDA was highest in AZT and DDI, second in 3TC and ABC, and was a tie in D4T. Both MTPS and Random Forest had low rate, but Random Forest was better than MTPS in all drugs except 3TC.

Figure 4 shows F1 score of different models for 5 different drugs. The higher the F1 score is, the better the model is. All models had much lower F1 score for DDI. For 3TC, Elastic net had a highest F1-score, and Random Forest had the second-best score. For ABC, Elastic net and Random Forest had the best performance in terms of F1-score. To see if either Elastic net or Random Forest was better, we will later check the p-value for the difference between two models. For AZT, MTPS had the lowest F1-score, and all other models were higher than MTPS but it's hard to tell which one is better than which. We will check if the difference between those three models is significant or not. For D4T, Elastic net and Random Forest had a better performance than other two models for D4T. But we also need to check if this is the case later. For DDI, overall F1-scores were all lower than other drugs. And MTPS had the best performance for DDI, and Random Forest and Elastic net was a tie for DDI. For F1-score we will look into the data with all drugs combined to see which model was the best model in terms of F1-score.

We will further investigate the observations made here later in wilcox test.

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

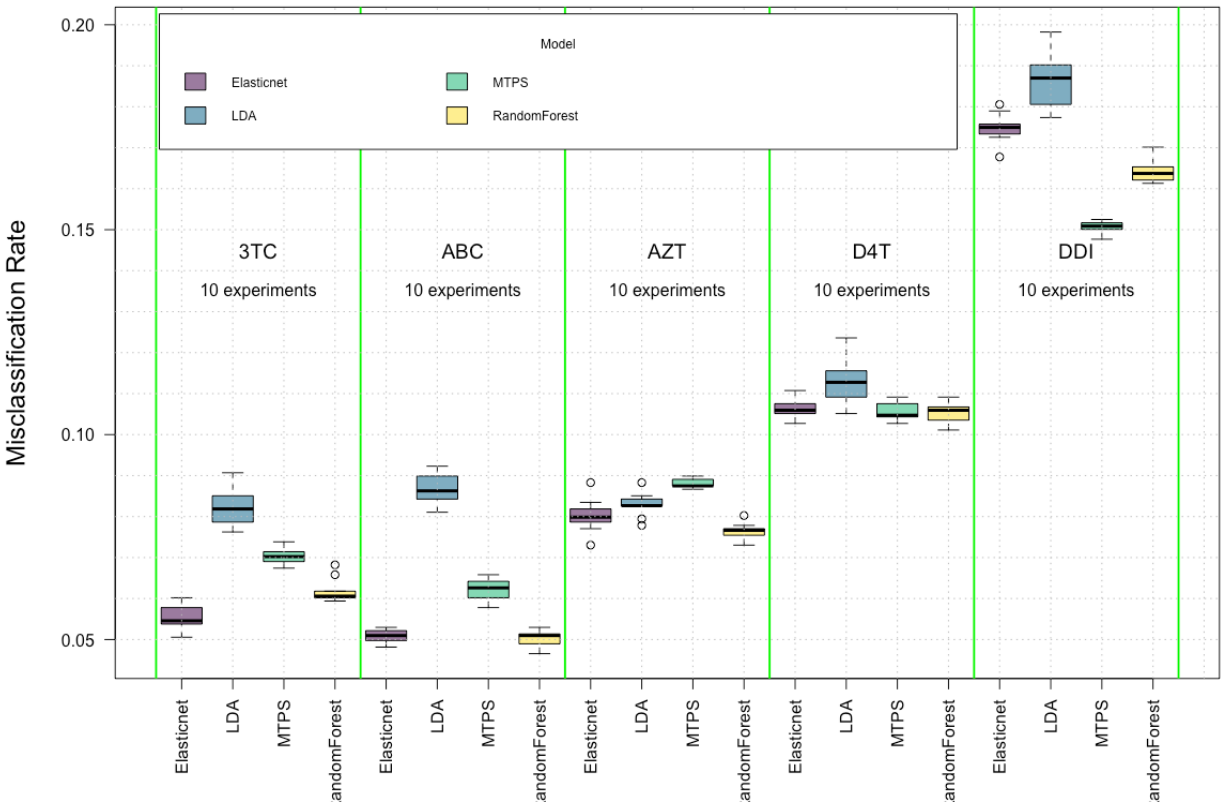


Figure 2: Precision of all Models grouped by Drug name

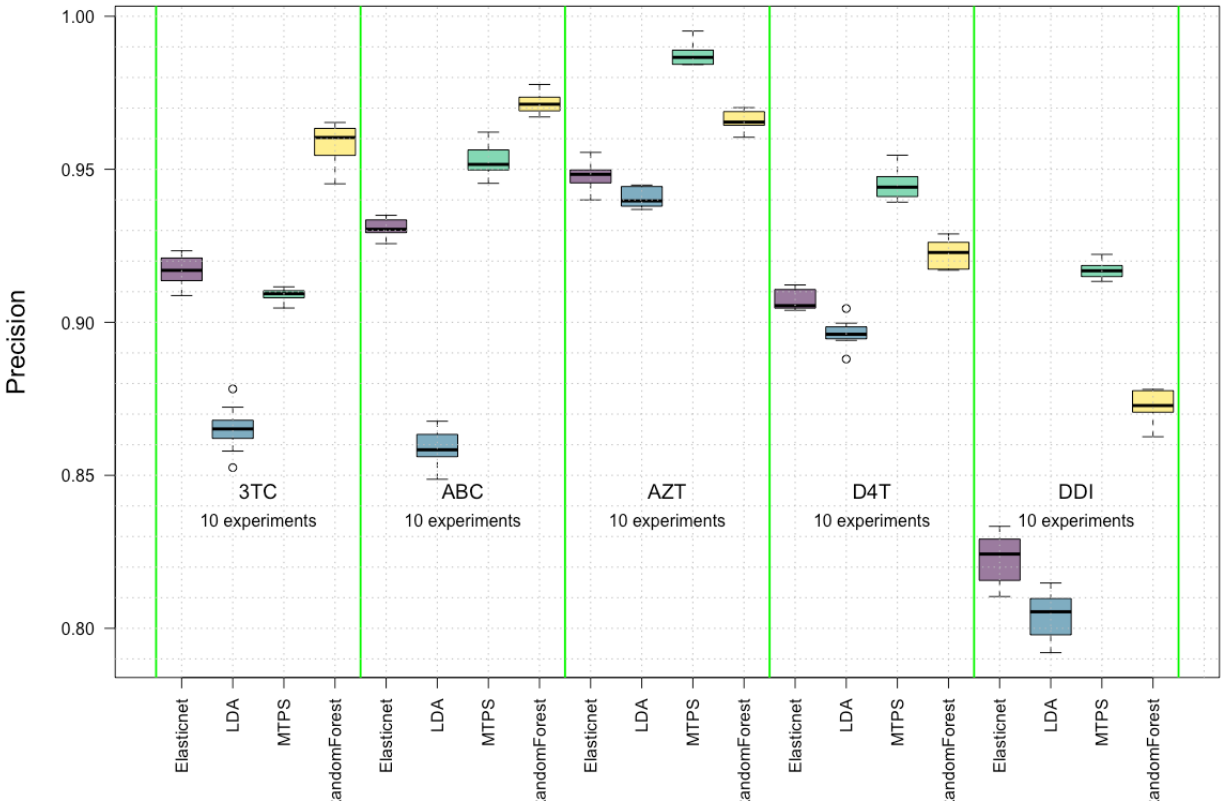


Figure 3: Recall of all Models grouped by Drug name

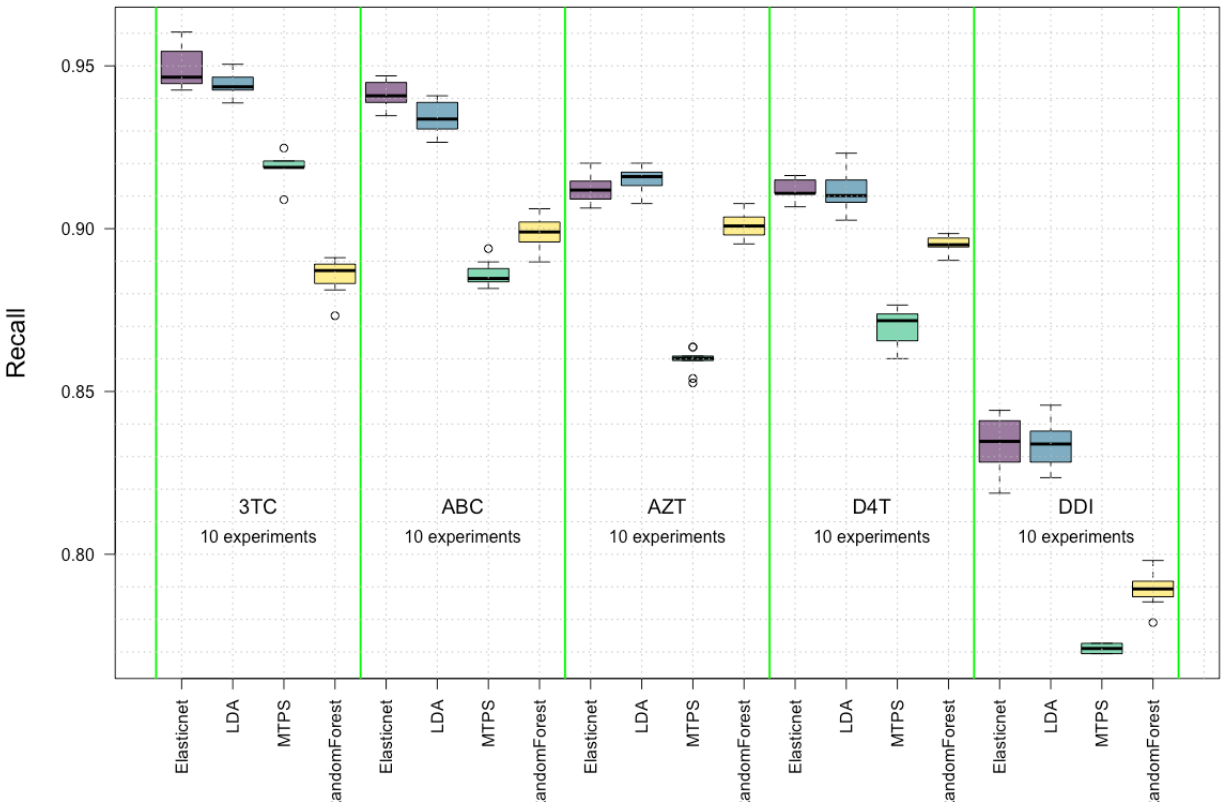
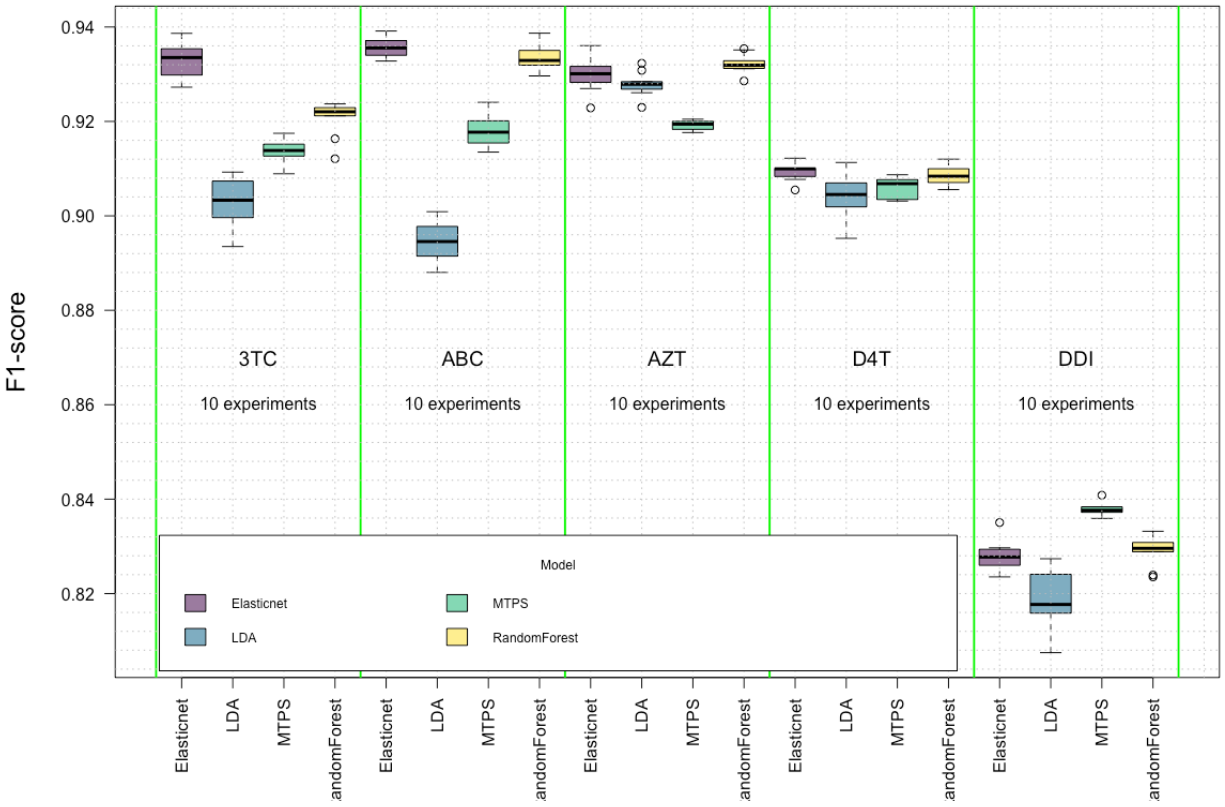


Figure 4: F1-score of all Models grouped by Drug name

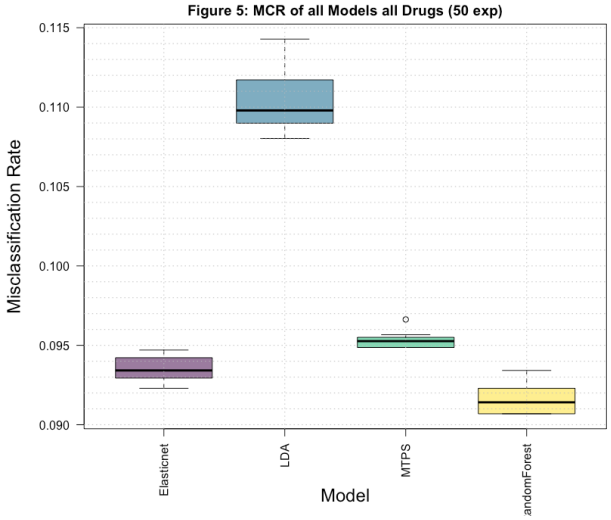


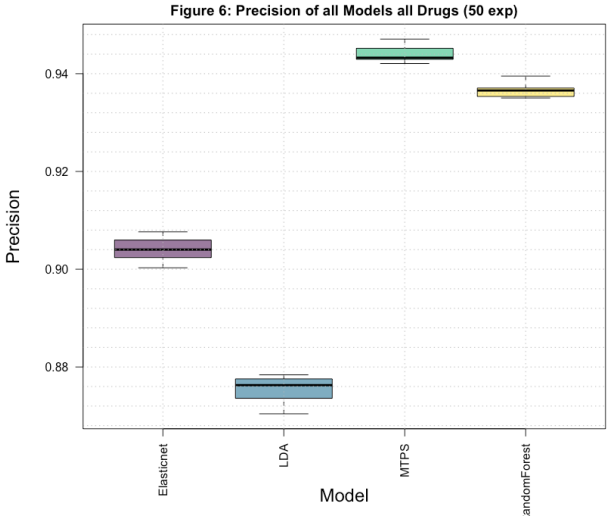
From figure 5-8 below, we can observe which model was the best model in terms of different criteria regardless of the drug type. Random forest was the best model in terms of misclassification rate. So, we can confirm the observation I made earlier in figure 1.

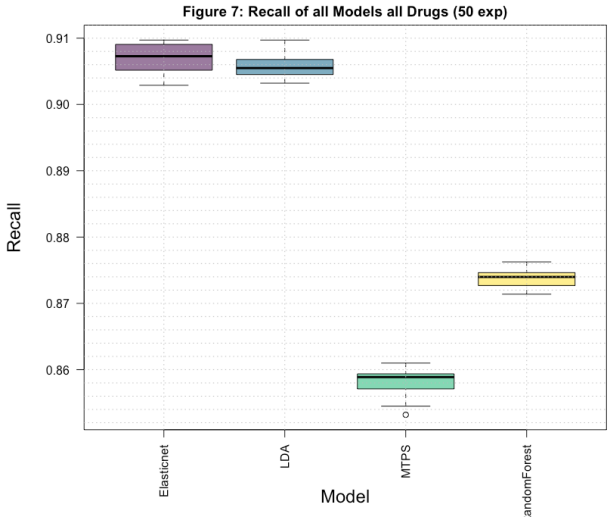
From figure 6, we can observe that MTPS was the best model in terms of precision and Random Forest was the second best which also confirms that observation from figure 2.

From figure 7, we can observe that the median of the elastic net is slightly higher than LDA and that overall box plot of elastic net locates higher than LDA. So, we can say that Elastic net was the best model in terms of recall in all drugs combined data.

From figure 8, we can observe that the elastic net was the best model in terms of F1-score and Random Forest was the second best. So, we can say Elastic net was the best model and Random Forest was the second in terms of F1-score.







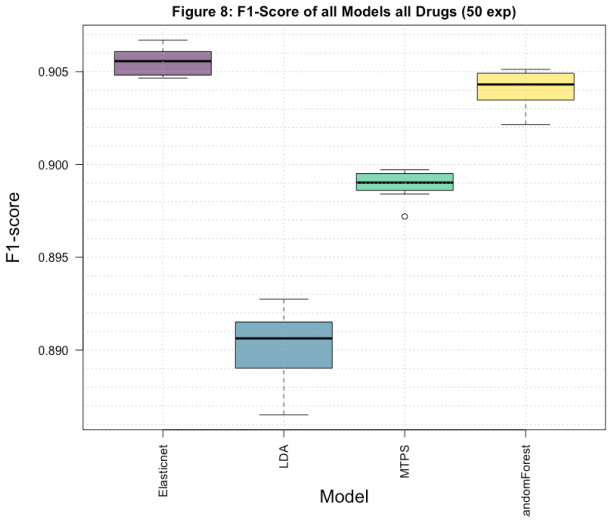


Table 3-7 shows p-value of wilcox pairs of F1-scores of different models. In most cases, p-values are very small except some cases.

In table 3, we can observe that the p-value for MTPS and Random Forest pair is between 0.01 and 0.05 which indicates that there's moderate significance for the difference between them for 3TC. This supports the observation made earlier about Elastic net being the best model and Random Forest being the second best for 3TC. All other p-values for other pairs are all between 0.001 and 0.01 so there is a significant difference between other models for 3TC.

In table 4, we can observe that p-value for Random Forest and Elastic net is big (> 0.1) for ABC, which indicates that the difference between Random Forest and Elastic net was insignificant. All other p-values are all between 0.001 and 0.01 so there is a significant difference between all other pairs for ABC. So, it does support the claim about Elastic net and Random Forest being the best model in ABC.

In table 5, we can observe that p-value for the pair Elastic net and LDA, and the pair Elastic net and Random Forest were big (> 0.1) for AZT, which shows that the difference between those pairs were insignificant. p-value for the pair LDA and Random Forest was between 0.01 and 0.05 which indicates that there's moderate significance for the difference between them for 3TC. P-value for all other pairs were all between 0.001 and 0.01 so there is a significant difference between those pairs. So, this supports the claim that MTPS was the worst model for AZT.

In table 6, we can observe that p-value for the pair Random Forest and Elastic net, and the pair LDA and MTPS were big (> 0.1) for D4T which shows that the difference between those pairs were insignificant. P-value of all other pairs were between 0.005 and 0.05 which indicates that there's moderate significance for the difference between them for D4T. So, it does support the claim that Elastic net and Random Forest was the best model for D4T.

In table 7, the p-value for the Random Forest and Elastic net was big (> 0.1) for DDI which indicates that the difference between Random Forest and Elastic net was insignificant for DDI. The p-value for all other pairs were between 0.001 and 0.01 which indicates that there is a significant difference between them for DDI. Therefore, it does support the claim that MTPS was the best model for DDI, and Random Forest and Elastic net was a tie for DDI.

Therefore, the p-values do support and explain the observations made earlier.

Table: 3: Wilcoxon Pairs of F1-Scores for Drug: 3TC (10 trials)

	Elasticnet	LDA	MTPS	RandomForest
Elasticnet	NA	0.0059215	0.0059215	0.0059215
LDA	0.0059215	NA	0.0059215	0.0059215
MTPS	0.0059215	0.0059215	NA	0.0108269
RandomForest	0.0059215	0.0059215	0.0108269	NA

Table: 4: Wilcoxon Pairs of F1-Scores for Drug: ABC (10 trials)

	Elasticnet	LDA	MTPS	RandomForest
Elasticnet	NA	0.0059215	0.0059215	0.1262789
LDA	0.0059215	NA	0.0059215	0.0059215
MTPS	0.0059215	0.0059215	NA	0.0059215
RandomForest	0.1262789	0.0059215	0.0059215	NA

Table: 5: Wilcoxon Pairs of F1-Scores for Drug: AZT (10 trials)

	Elasticnet	LDA	MTPS	RandomForest
Elasticnet	NA	0.1029175	0.0059215	0.1535764
LDA	0.1029175	NA	0.0059215	0.0108269
MTPS	0.0059215	0.0059215	NA	0.0059215
RandomForest	0.1535764	0.0108269	0.0059215	NA

Table: 6: Wilcoxon Pairs of F1-Scores for Drug: D4T (10 trials)

	Elasticnet	LDA	MTPS	RandomForest
Elasticnet	NA	0.0108269	0.0108269	0.3589514
LDA	0.0108269	NA	0.4148231	0.0190589
MTPS	0.0108269	0.4148231	NA	0.0080452
RandomForest	0.3589514	0.0190589	0.0080452	NA

Table: 7: Wilcoxon Pairs of F1-Scores for Drug: DDI (10 trials)

	Elasticnet	LDA	MTPS	RandomForest
Elasticnet	NA	0.0080452	0.0059215	0.3589514
LDA	0.0080452	NA	0.0059215	0.0059215
MTPS	0.0059215	0.0059215	NA	0.0059215
RandomForest	0.3589514	0.0059215	0.0059215	NA

Table 8-10 shows the P-value of wilcoxon pairs of the F1-scores for each drug for different models. Most p-values are very small (<0.001) or small (<0.005) except few cases. For elastic net, p-values are all very small or small (< 0.005) except the pair 3TC and ABC and the pair 3TC and AZT. This is because Elastic net had similar F1-score for 3TC, ABC and AZT. So, observations made about elastic net is still valid. For LDA, pair D4T and 3TC had a big p-value (> 0.1) and it was also because LDA had very similar F1-score in D4T and 3TC. For the MTPS, ABC and AZT a big p-value (> 0.1) and it was also because MTPS had similar F1-score in ABC and AZT. So, these does not affect our observation on choosing the best model.

Table: 8: Wilcoxon Pairs of F1-Scores for Model: Elasticnet

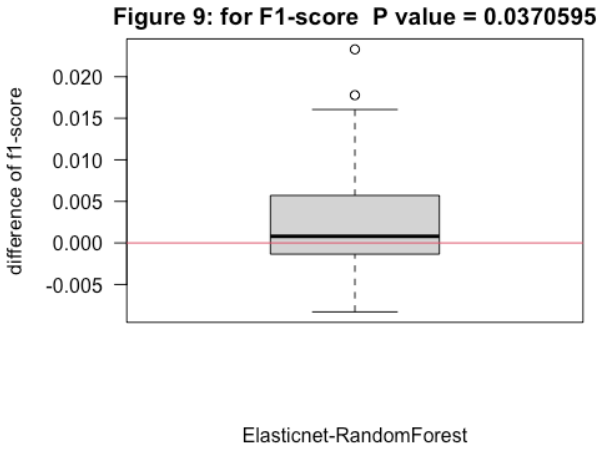
	3TC	ABC	AZT	D4T	DDI
3TC	NA	0.1212245	0.1041099	0.0001817	0.0001827
ABC	0.1212245	NA	0.0010080	0.0001817	0.0001827
AZT	0.1041099	0.0010080	NA	0.0001817	0.0001827
D4T	0.0001817	0.0001817	0.0001817	NA	0.0001817
DDI	0.0001827	0.0001827	0.0001827	0.0001817	NA

Table: 9: Wilcoxon Pairs of F1-Scores for Model: LDA

	3TC	ABC	AZT	D4T	DDI
3TC	NA	0.0028273	0.0001827	0.5707504	0.0001827
ABC	0.0028273	NA	0.0001827	0.0005828	0.0001827
AZT	0.0001827	0.0001827	NA	0.0001827	0.0001827
D4T	0.5707504	0.0005828	0.0001827	NA	0.0001827
DDI	0.0001827	0.0001827	0.0001827	0.0001827	NA

Table: 10: Wilcoxon Pairs of F1-Scores for Model: MTPS

	3TC	ABC	AZT	D4T	DDI
3TC	NA	0.0021685	0.0001786	0.0001786	0.0001766
ABC	0.0021685	NA	0.4270075	0.0001806	0.0001786
AZT	0.0001786	0.4270075	NA	0.0001806	0.0001786
D4T	0.0001786	0.0001806	0.0001806	NA	0.0001786
DDI	0.0001766	0.0001786	0.0001786	0.0001786	NA



From figure 9, the p-value for the difference between Elastic net and Random Forest was between 0.01 and 0.05 which indicates that there's moderate significance for the difference between them. Therefore, we can conclude that Elastic net was little bit better than Random Forest in terms of F1-score.

In conclusion, for all the drugs, the best model was Elastic net and Random Forest was the second-best model considering F1-score and other criteria.