Investigating Genetic Algorithm as a feature Selection Method and Comparing with Logistic Regression and Elastic Net on HIV Drug Resistance Data

Shaghayegh AhooeiNejad[1], Farbod Esmaeili[2], Winnie Zheng[3]

[1] University of Victoria, V01017073

[2] University of Victoria, V01017091

[3] University of Victoria, V01006279

December 13, 2022

SUMMARY. The immune system of the body is attacked by the virus known as the Human immunodeficiency virus (HIV). There is no known cure for this condition, and people who contract HIV will live with the condition for the rest of their lives. However, HIV can be managed with the right medical attention. Effective HIV treatment enables people with HIV to enjoy long, healthy lives and safeguard their intimate partners. In this study, a machine learning-based strategy for identifying HIV medication resistance is presented. We compare logistic regression to genetic algorithms and elastic net regularization. Three experiments are presented to illustrate the comparability between these strategies. In the first experiment, we choose genetic algorithm parameters from a genetic perspective, but in the second experiment, logistic regression is utilised. In the third experiment, a novel strategy for training the elastic net regularisation model is proposed. Our model was evaluated to determine the ideal model for our response and to determine whether the medications utilising the predictor were resistant or non-resistant based on Stanford team-collected patient data. Our proposed elastic net regularization model attained a classification accuracy of 90.57% and a f1 score of 90.60%. Our algorithm demonstrates that elastic net regularization approaches enhanced by the suggested notion can bring a highly accurate approach to the problem of HIV medication resistance. Another result was that the Genetic Algorithm works almost nearly as Elastic Net and performs highly better than logistic regression.

Keywords: Genetic Algorithm, Logistic Regression, Elastic Net Regularization, Machine Learning, HIV

1 Introduction

Combination antiretroviral therapy (ART) is increasingly accessible to HIV patients. ART is typically efficient at suppressing the virus and maintaining the patient's health. However, treatment is only effective if the virus being treated is not resistant to the medications being employed. In the 1980s, when the first antiretroviral medications were launched, resistance was a given for all patients, and the duration of successful treatment was limited. Today, some patients receive treatment for years without encountering any resistance issues, whilst for others, resistance poses a grave threat to their health.

Therefore, in our research, we will therefore utilize HIV Drug Resistance data and three distinct Algorithms to determine drug resistance in five medicines, and we will detail these Algorithms and the Data Set we will employ.

1.1 HIV dataset

The team collected, annotated, and analyzed genotypic resistance data from previously untreated (ARV-nave) and previously treated (ARV-experienced) individuals, as well as systematically aggregated data from numerous published studies to analyze temporal global trends in resistance mutations leading to ADR and TDR, thereby creating a resource for clinical and molecular epidemiologists.

They are collaborating with researchers from all over the world to track global trends in ADR and TDR and to identify genetic mechanisms of HIVDR to ARVs that require additional data, including those belonging to the nucleoside RT inhibitor (NRTI), non-nucleoside RT inhibitor (NNRTI), protease inhibitor (PI), and integrase strand transfer inhibitor (INSTI) drug classes.

The sample size of the dataset was 1246, and the result was information regarding resistance to five NRTI-class medicines. IC50 (drug concentration in resistant strains) over IC50 (drug concentration in wild type) represents drug resistance. Biologists identified drug resistance (yes or no) based on their IC50 ratio (continuous value), which was calculated by five cutoff values. Each predictor isolate/virus possessed 229 mutations.

HIV drug resistance statistics are crucial for HIV drug resistance surveillance, ARV medication design, and the care of individuals infected with drug-resistant HIV. These data are most effectively recorded in a database that not only catalogs mutations linked with drug resistance but also connects the entire gene sequences to other types of data.

1.2 Algorithms

Genetic Algorithms (GA) are search approaches for machine learning that are inspired by Darwinian evolutionary models. GA has an advantage over factor analysis and other similar statistical models in that it can solve problems for which there is no human expertise or for which the sought-after answer is too complex for an approach based on human skill. Challenges that can be described as functional optimization issues are amenable to the application of GA. This

makes GA suitable for discrete combinatorial and mixed integer issues. Consequently, GA approaches are appropriate for finding solutions that need efficient search of a subset of characteristics to find near-optimal combinations that address high-dimensional classification problems, particularly when the search space is huge, complex, or poorly understood. Analyzing and predicting medication resistance based on numerous patient data is an example of such an issue.

Logistic Regression is a mathematical model that allows for the assessment of the likelihood of belonging to a certain class. In this study, the LR model is used to binary classification, but it is easily extensible to multi-label classification in other situations.

Lastly, the Elastic Net regression, it is a hybrid statistical approach used for regularizing and choosing essential predictor variables that have a substantial impact on the response variable and addressing the multicollinearity issue when it occurs between the predictor variables.

1.3 Steps

In this section, we would like to present the steps we took to obtain the final results.

- For starting and importing the data we install the "MTPS" package and load the HIV data for exploration
- Visualize the data
- Convert the ic50 ratio into binary drug resistance. Biologists decided 5 cut-off values to define drug resistance (Yes vs No) based on their IC50 ratios (continuous values)
- Visualize the binary data
- Apply prediction methods (Genetic Algorithm, Logistic Regression, and Elastic Net)
- Using 5-fold cross-validation to compare the performance and repeating it for 50 times to calculate uncertainty
- Compare the performance of these methods by their criteria (Accuracy, Precision, Recall, and F1-Scores)
- Using box plots to visualize comparison results and the Wilcoxon test to investigate whether the observed performance differences between methods are significant
- There are 5 different drugs, comparing methods on each drug, and comments on if the best method is consistent on all drugs

2 Methods

In this section, we will describe the data that we are trying to solve a multi-task prediction problem and build a model that can classify our data with a low rate of false negatives by using Genetic Algorithm, Logistic Regression, and Elastic Net.

The Data used is from "The HIV Drug Resistance Database" built by a Stanford team.

"HIV (human immunodeficiency virus) is a virus that attacks the body's immune system. If HIV is not treated, it can lead to AIDS (acquired immunodeficiency syndrome). HIV treatment (antiretroviral therapy or ART) involves taking medicine as prescribed by a health care provider. HIV treatment reduces the amount of HIV in your body and helps you stay healthy. There is no cure for HIV, but you can control it with HIV treatment."

The sample size is 1246. The outcome is Resistance information of 5 drugs in the NRTI class. NRTI is the abbreviation of Nucleoside Reverse Transcriptase Inhibitor which is an Antiretroviral (ARV) HIV drug class. Nucleoside reverse transcriptase inhibitors (NRTIs) block reverse transcriptase (an HIV enzyme).

HIV uses reverse transcriptase to convert its RNA into DNA (reverse transcription). Blocking reverse transcriptase and reverse transcription prevents HIV from replicating. This will be computed as follow:

$$Resistance = \frac{IC50(drug\ concentration\ for\ resistant\ strain)}{IC50(drug\ concentration\ for\ wild\ type)}$$

The wild type in the equation above is the naturally occurring, non-mutated strain of a virus. When exposed to antiretroviral (ARV) drugs, wild-type HIV can develop mutations that make the virus resistant to specific HIV drugs. In this data set, there are 228 mutations of each isolate/virus and five predictors that represent five different drugs. These drugs are Lamivudine (3TC), Abacavir (ABC), Zidovudine (AZT), Stavudine (D4T) and Didanosine (DDI).

These response variables have been changed to binary variables using specified cut-off values suggested by biologists. "HIV is so difficult to cure because the virus persists inside stable reservoirs that cannot be detected by the immune system." That is why we want to investigate the Genetic Algorithm for the data set above and find the best model for our response and see if the drugs using the predictors are resistant or non-resistant.

Moreover, we are using 5-fold cross-validation to investigate which model classifies the best with a low rate of False Negative and False Positive and repeating the process 50 times to learn the uncertainty of the comparison. We wanted to do the step-wise logistic regression and compare the results with the Genetic Algorithm to find the best algorithm or method of selecting the best subset of predictors. Unfortunately, the step-wise logistic regression failed to give any results due to the high volume of predictors and samples and neither time-wise nor memory-wise the computer was not able to compile the step-wise algorithm. Furthermore, in our previous studies, we already compared multiple models such as random forest, LDA, residual stacking, and elastic net.

As the result showed us elastic net was the best model and that is why we are comparing these two models, genetic

algorithm and logistic regression, with elastic net as well. We did Wilcoxon Test on F1-Scores to see if the differences observed in the box plots are significant or not.

We used Genetic Algorithm to find the best subsets of predictors with the best AIC and then use those proposed subsets to fit a logistic model. For the genetic algorithm, we set the size of each generation to 20 parents and the iteration number to 100. We chose one of the parents completely random and the other one base on their fitness. Then we used crossover and mutation at the rate of 0.01 to produce the next generation of parents. Each parent is a combination 0's and 1's representing if each corresponding predictor is included in the logistic regression or not.

For fitting elastic net, we used 5-fold cross-validation to decide on the best value of alpha and lambda based on their accuracy. The binomial family was chosen for the function.

2.1 Model Training and Testing

First, we stratified each drug into two groups of 0 and 1 and then we split each group into 5 folds and combined two groups of each fold to create the training data and the one group left is our test data. This procedure is repeated 50 times on each of the 5 drugs in order to compute the uncertainty of our results. After sampling is done, each drug outcome is modeled separately. All the models mentioned were fitted to our data sets and then their confusion matrices were stored. Finally, based on these confusion matrices we calculated and compared their run time, F1-Score, Accuracy, Precision, and Recall. The comparison was done by an intuition of box plots and also the Wilcoxson test. We also compared the models based on their running time and average overall running time for each model.

2.2 Preliminary Visualization

The first visualization of our primary data before changing to a Binary response is as shown below:

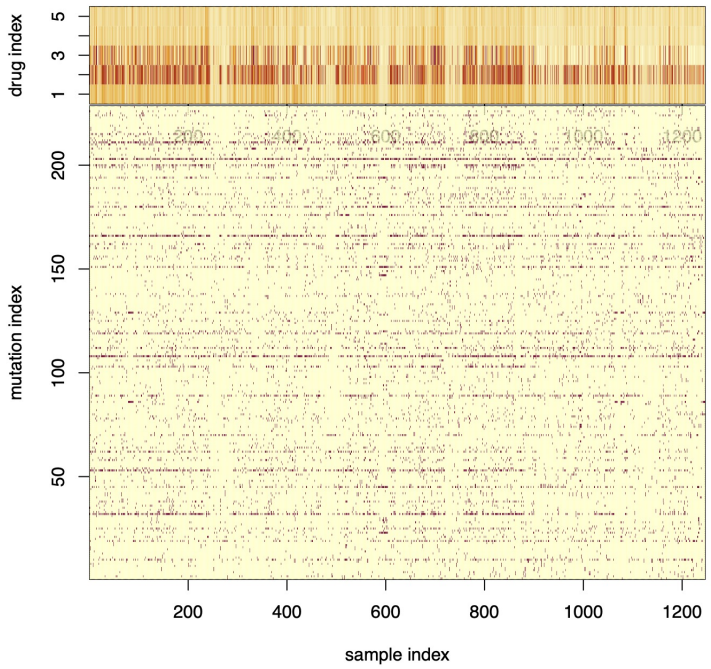

Figure 1: Primary Visualization of the date before changing the response in to a binary outcome

And after changing to a binary variable:

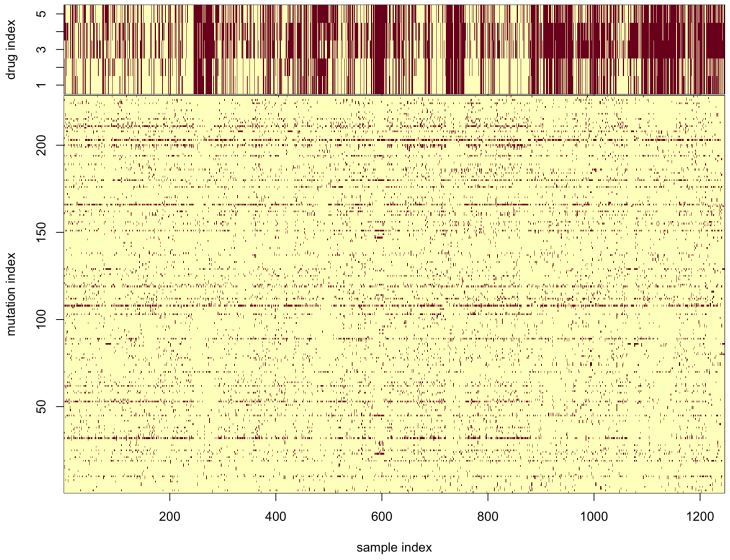

Figure 2: Secondary Visualization of the date after changing the response in to a binary outcome

3 Results

In this section, we will illustrate the outcome of the three proposed algorithms and derive our conclusion based on these outputs.

3.1 Comparison of Results: Combined Drug Box plots

For each Model, box plots show F1-score, Accuracy, Precision, and Recall values grouped by 5 different Drugs and in the 50 iterations. Higher values for all these measurements are preferred and indicate a good model is being fitted. According to the F1-Score plot, these conclusions are drawn:

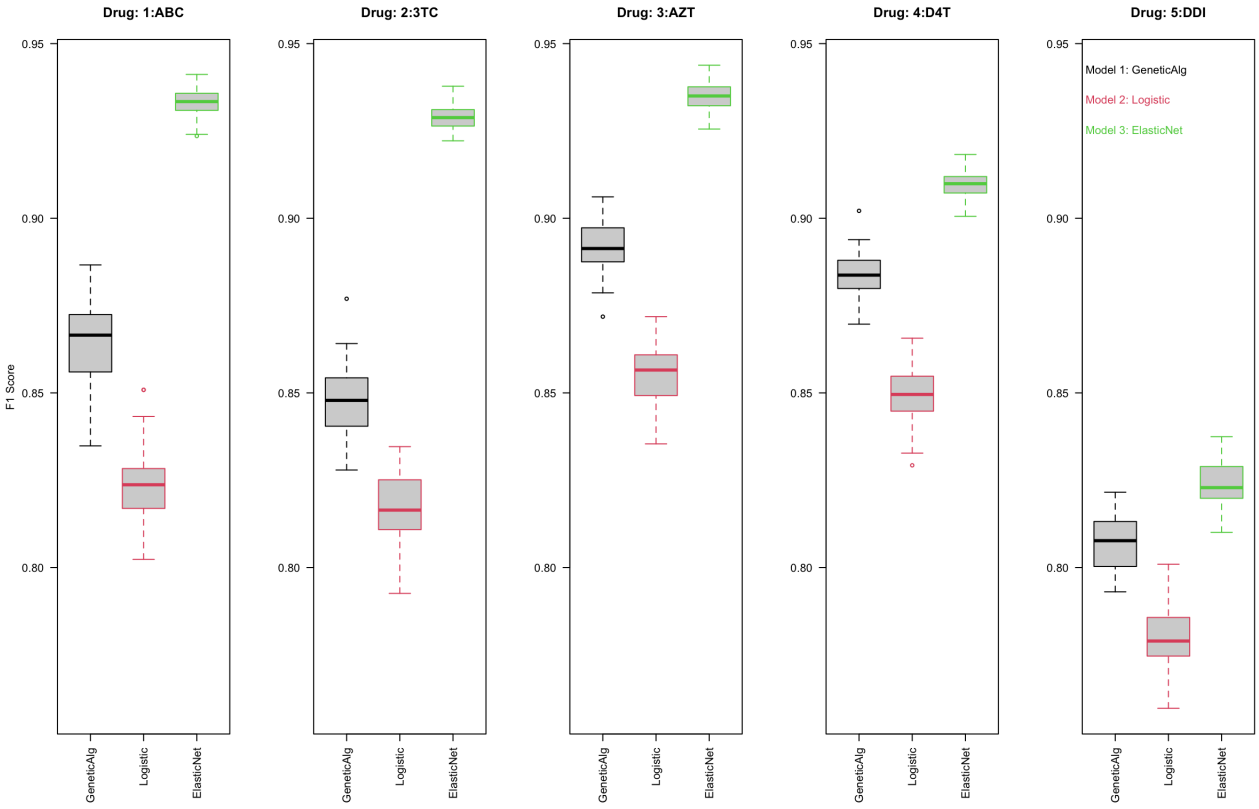

Figure 3: F1-score box plot for five different drugs. This plot is used to compare the F1-score of each model for each different drug

- Elastic Net outperforms other models with its high value for F1-Score.
- Drug 5 has a low F1-score in all the models compared with other drugs.
- Logistic Regression has the lowest F1 score of all drugs.
- Genetic Algorithm has a higher value than Logistic Regression and a lower value than Elastic Net.

According to the Accuracy plot, these conclusions are drawn:

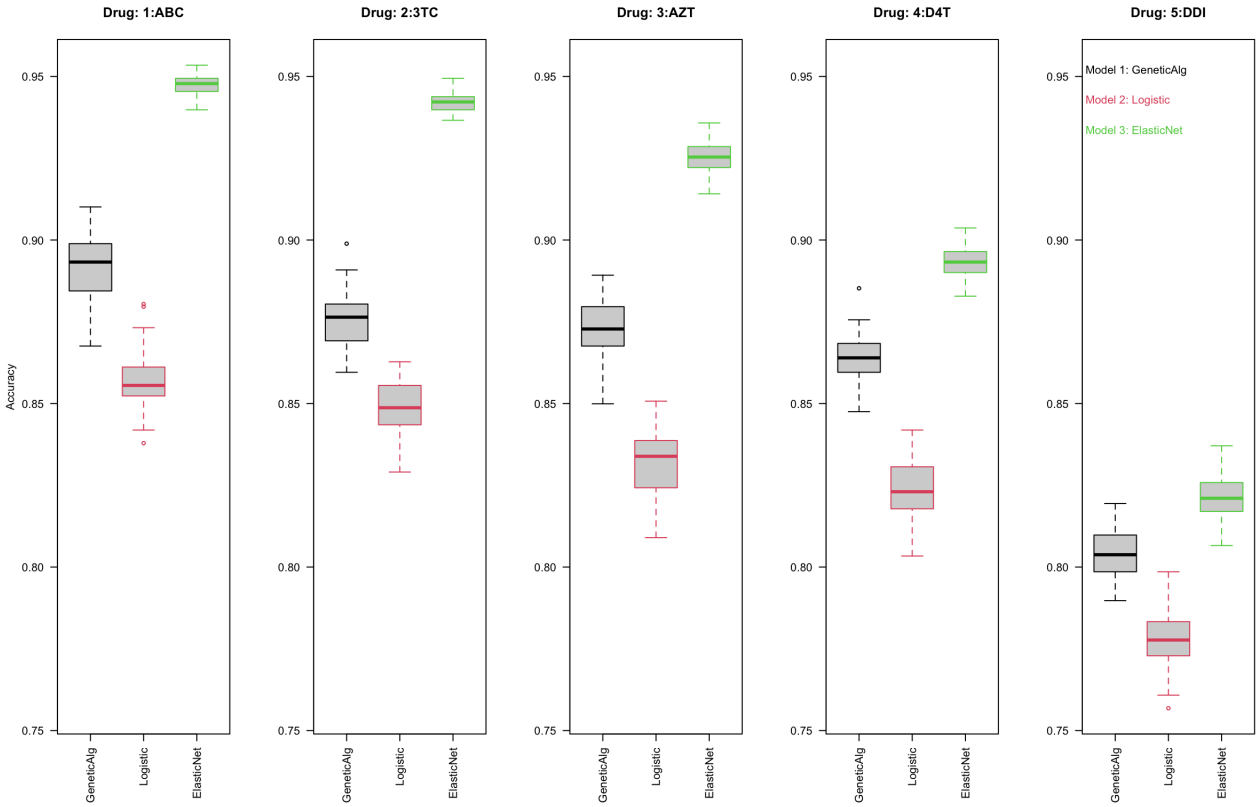

Figure 4: Accuracy box plot for five different drugs. This plot is used to compare the accuracy of each model for each different drug

- Elastic Net outperforms other models with its high value for accuracy.
- Drug 5 has low accuracy in all the models compared with other drugs.
- Logistic Regression has the lowest Accuracy of all drugs.
- Genetic Algorithm has an upper value than Logistic Regression and a lower value than Elastic Net.
- Excluding Drug5, the Genetic Algorithm shows a robust and stable Accuracy in all drugs compared to the other methods used.

According to the precision plot, these conclusions are drawn:

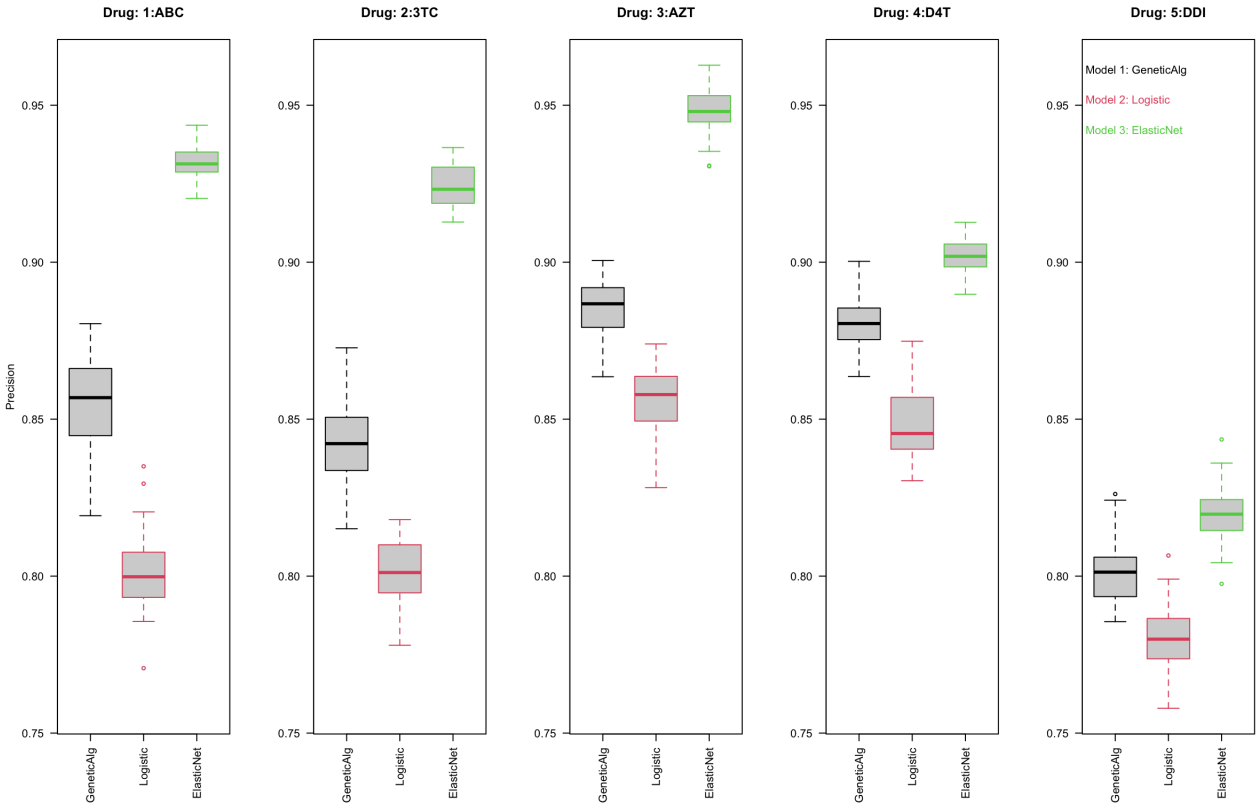

Figure 5: Precision box plot for five different drugs. This plot is used to compare the precision of each model for each different drug

- Elastic Net outperforms other models with its high value for Precision.
- Drug 5 has low precision in all the models compared with other drugs.
- Logistic Regression has the lowest precision of all drugs.
- Genetic Algorithm has an upper value than Logistic Regression and a lower value than Elastic Net.

According to the Recall plot, these conclusions are drawn:

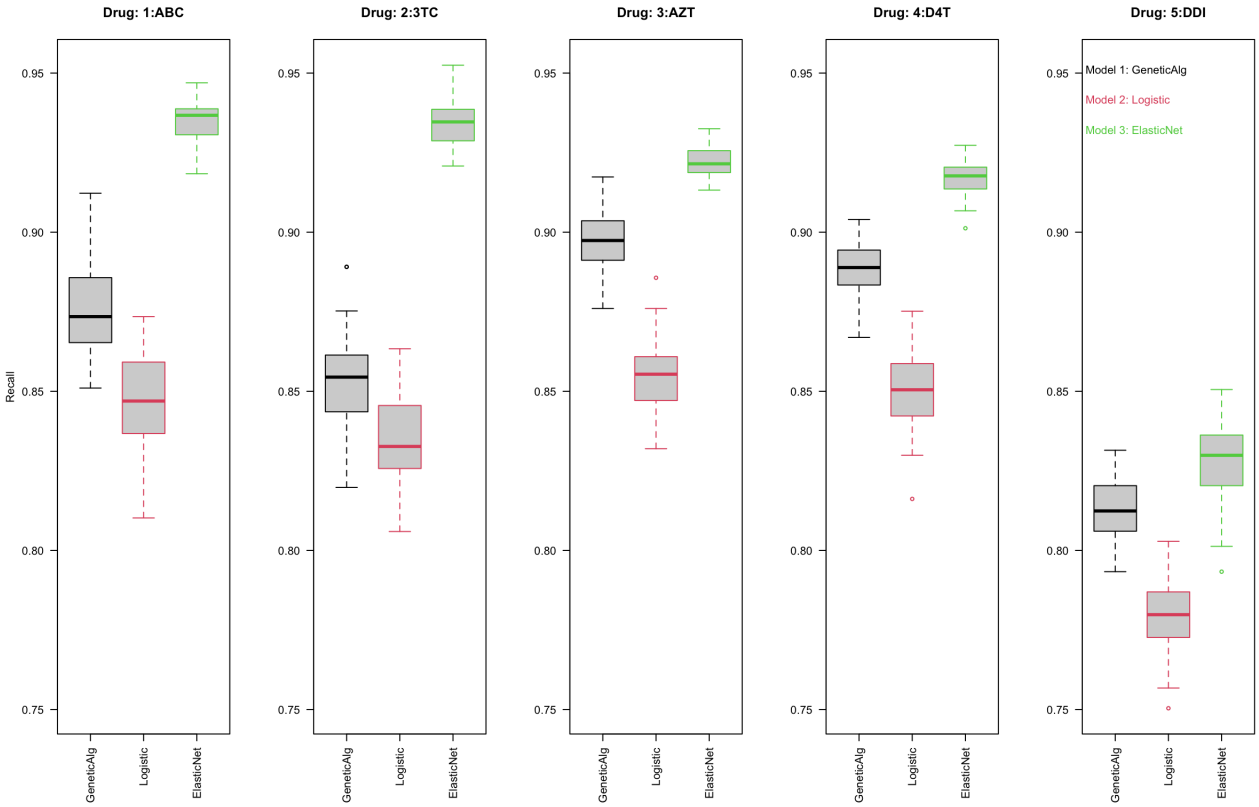

Figure 6: Recall boxplot for five different drugs. This plot is used to compare the recall of each model for each different drug

- Elastic Net outperforms other models with its high value for recall.
- Drug 5 has a low recall in all the models compared with other drugs.
- Logistic Regression has the lowest recall of all drugs.
- Genetic Algorithm has an upper value than Logistic Regression and a lower value than Elastic Net.

It is noteworthy that the GLM model was not the best model may be due to the warning of the code for convergence.

For a better understanding of these plots and to investigate more, we do the Wilcoxon Test P-Value Pairs of F1-Scores in the next part.

3.2 Wilcoxon Test P-Value Pairs of F1-Scores by Model and Drugs

Each table is related to one of the drugs containing the P-values of the Wilcoxon test for the same drug and between different models fitted (same models as written in the row and column). We can see that in the 95 confidence level, all differences are significant which means that all F1 scores of all methods and drugs performed significantly differently.

Table 1: Wilcoxon Test P-Value Pairs of F1-Scores by Model for Drug 1

Drug 1	Genetic Algorithm	Logistic	Elastic Net
Genetic Algorithm	-	8.279e-10	7.79e-10
Logistic	8.279e-10	-	7.79e-10
Elastic Net	7.79e-10	7.79e-10	-

Table 2: Wilcoxon Test P-Value Pairs of F1-Scores by Model for Drug 2

Drug 2	Genetic Algorithm	Logistic	Elastic Net
Genetic Algorithm	-	7.79e-10	7.79e-10
Logistic	7.79e-10	-	7.79e-10
Elastic Net	7.79e-10	7.79e-10	-

Table 3: Wilcoxon Test P-Value Pairs of F1-Scores by Model for Drug 3

Drug 3	Genetic Algorithm	Logistic	Elastic Net
Genetic Algorithm	-	7.79e-10	7.79e-10
Logistic	7.79e-10	-	7.79e-10
Elastic Net	7.79e-10	7.79e-10	-

Table 4: Wilcoxon Test P-Value Pairs of F1-Scores by Model for Drug 4

Drug 4	Genetic Algorithm	Logistic	Elastic Net
Genetic Algorithm	-	7.79e-10	7.79e-10
Logistic	7.79e-10	-	7.79e-10
Elastic Net	7.79e-10	7.79e-10	-

Table 5: Wilcoxon Test P-Value Pairs of F1-Scores by Model for Drug 5

Drug 5	Genetic Algorithm	Logistic	Elastic Net
Genetic Algorithm	-	7.79e-10	1.513e-09
Logistic	7.79e-10	-	7.79e-10
Elastic Net	1.513e-09	7.79e-10	-

3.3 Run time by Model and Drugs

Below there is a graph illustrating the average run time by each model and in each drug.

Although logistic regression was the fastest, this method did not do any feature selection and it did not have a high value in any of the criteria discussed earlier.

Elastic net gradually increased from drug 1 to drug 5. Surprisingly, it passes the Genetic Algorithm.

Genetic Algorithm had the highest average run time in drug one to three. This algorithm takes over the second spot in drugs four and five as the Elastic net increases. The genetic Algorithm's average run time shows a steady decrease.

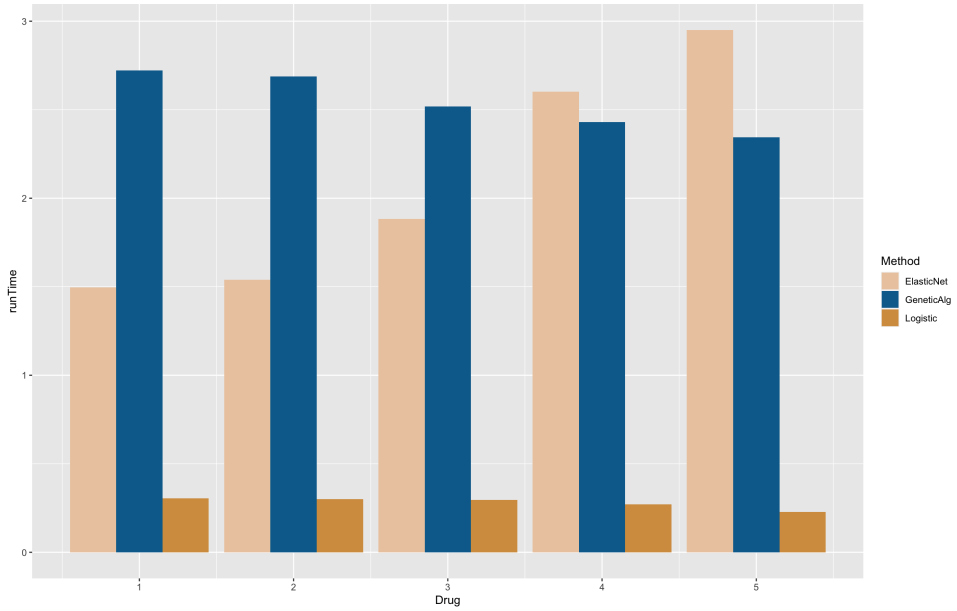

Figure 7: Average run time of three methods in all the five drugs

Logistic regression has the lowest run time in all the iterations of all drugs.

Elastic net rapidly rises from drug 1 to drug 5. Surprisingly, it passes the Genetic Algorithm, the variance of the run time increased over the drugs, and outliers appeared.

The genetic algorithm had the highest run time in all the iterations in drug one to three. In drugs four and five, this algorithm takes over the second spot as the Elastic net increases. The genetic Algorithm's run time shows a steady decrease and robustness.

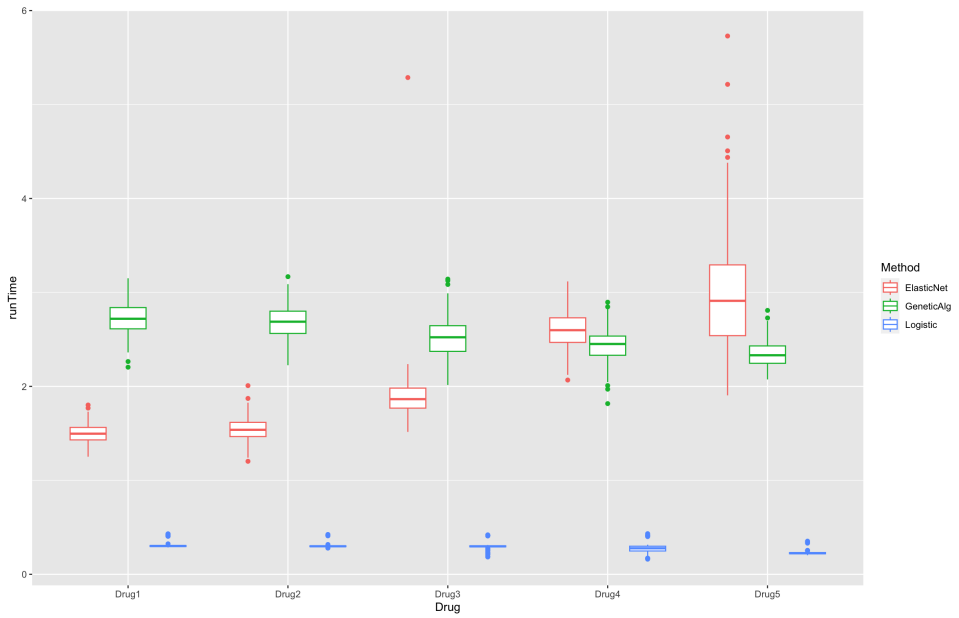

Figure 8: Run time of three methods in all the five drugs

4 Discussion

In this research, we focus on different methods to detect drug resistance to different drugs. We conducted three experiments using the same dataset but with different algorithms, showing the different performances of these algorithms.

Section 2 provides an overview of the experiments. The Stanford team's obtained dataset was used for the studies. In Section 1.1, a comprehensive description of the dataset is provided. In Section 2.1, the experiments are described in depth. The findings of this experiment are detailed in Section 3.1 and separated into four plots based on the objective function that was optimized (F1-score, accuracy, precision, or recall values).

The best model, Elastic Net, achieved a classification accuracy of 90.57% and an f1-score of 90.6%.

We then ran run time plots and found that Logistic Regression was the fastest, but this method did not do any feature selection and it did not have a high value in any of the criteria discussed earlier. Moreover, the genetic algorithm had the highest average running time.

This research compares three machine-learning techniques to determine the optimal model. Using various Algorithms, initially Logistic Regression, then Genetic Algorithm, and lastly Elastic Net, we produced progressively better outcomes with each successive experiment. The most accurate findings obtained had an f1 score of 90.6% and an accuracy of 90.57%. This demonstrates that Elastic Net is the optimal method for solving this kind of problem and this method is persistent on all the 5 drugs. Although Elastic Net outperforms the other method time-wise and by performance , our main goal was to show that Genetic Algorithm is a better substitute for step-wise logistic regression and logistic regression itself in a large data sets. In the future, we intend to apply ensemble methods and deep

science on larger data sets, integrating these methods with genetic/algorithmic approaches to research drug resistance identification.

5 Software and Codes

The R code was submitted along with this file. Also, the Source code for reproducibility can be found in the Supplementary Materials on Github(. Due to the high volume of the data and the 50 times to be repeated, it may take a while to run completely on a different computer. The outputs and confusion matrices were saved under the folder data as well.

6 Author Contributions

Shaghayegh AhooeiNejad and Farbod Esmaeili conceived the ideas and designed methodology and proposed the data; Farbod Esmaeili and Shaghayegh AhooeiNejad developed the R code; Winnie Zheng led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for submission

References

[1] Zhang, X. (2022, Dec 12). Machine learning lectures

[2] Stanford University (Ed.). (2022). HIV drug resistance database. HIV Drug Resistance Database. Retrieved November 27, 2022, from https://hivdb.stanford.edu/

[3] Centers for Disease Control and Prevention. (2022). HIV Basics. Living with the HIV. HIV Treatments. December 12,2022 from https://www.cdc.gov/hiv/basics/livingwithhiv/treatment.html

[4] Centers for Disease Control and Prevention. (2022). HIV Basics. About HIV. December 12,2022 from https://www.cdc.gov/hiv/basics/whatishiv.html

[5] Science of HIV. HIV Cure (2022). December 12,2022 from https://scienceofhiv.org/wp/cure/