Circadian Rhythm Project - Molecules Toxicity

MRR course Project

Presented by:
Mohammed-Yassine Barnicha
Nezar Aberqi
Supervised by:
Ms. Juhyun Park

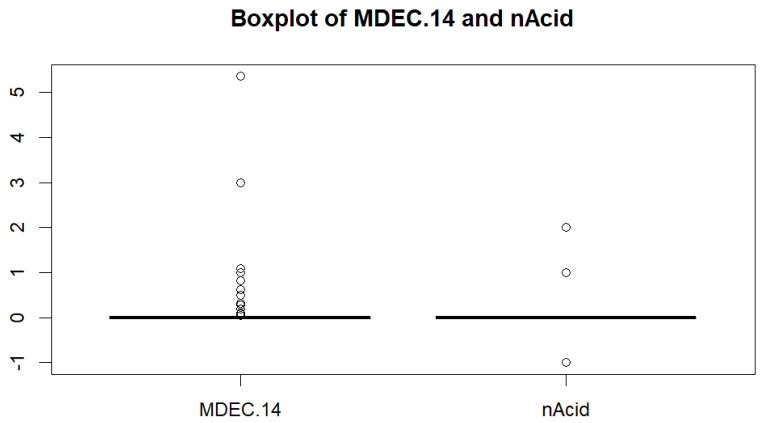
18/12/2023

General context

1. Study of circadian rhythms and molecular aspects in cells, focusing on toxicity classification and molecular interactions.
2. Purpose : Use of data science to analyze 178 cells for key variables influencing toxicity and developing a predictive toxicity model.

EDA - Dataset Overview and Feature Characteristics

- 171 observations and 1203 features.
- Two data types.
- Sparse feature Matrix.
- Different scales of statistical metrics.



Statistic	n6HeteroRing	MATS3m
Min.	0.000	-0.198700
1st Qu.	0.500	-0.052350
Median	1.000	-0.001600
Mean	1.216	0.003226
3rd Qu.	2.000	0.056550
Max.	4.000	0.168400

Exploratory Data Analysis

- Binary target variable.
- Few toxic molecules.

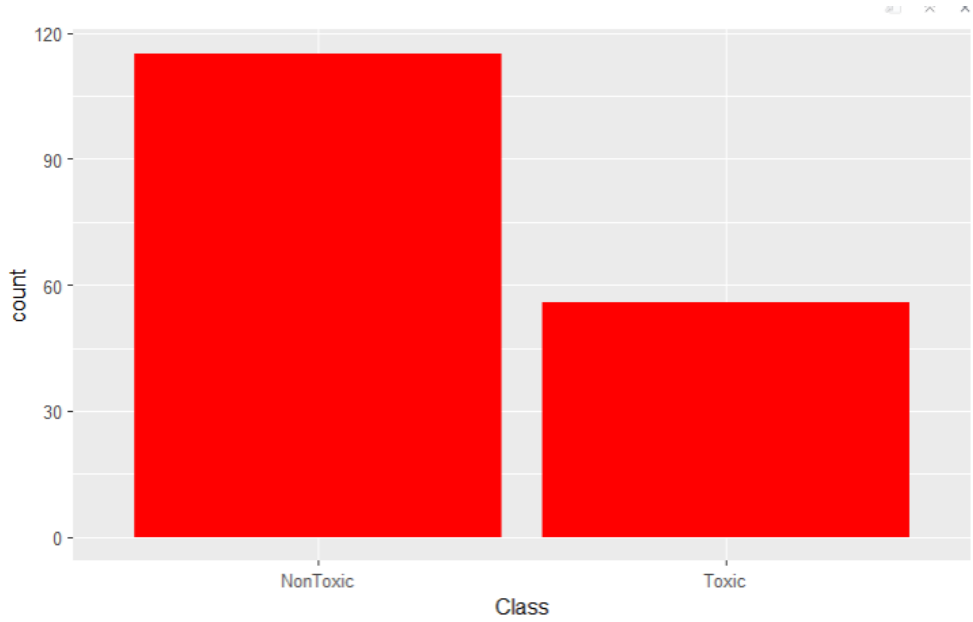


Figure: Train vs Test performance

Sensitivity Analysis with Student's t-test

For each feature X_i, we compare the means of two groups: $\mathrm{E}(X_i|_{y_i=1})$ and $\mathrm{E}(X_i|_{y_i=0})$, using the following t-statistic formula:

$$t = \frac{\bar{X}_0 - \bar{X}_1}{\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}}}$$

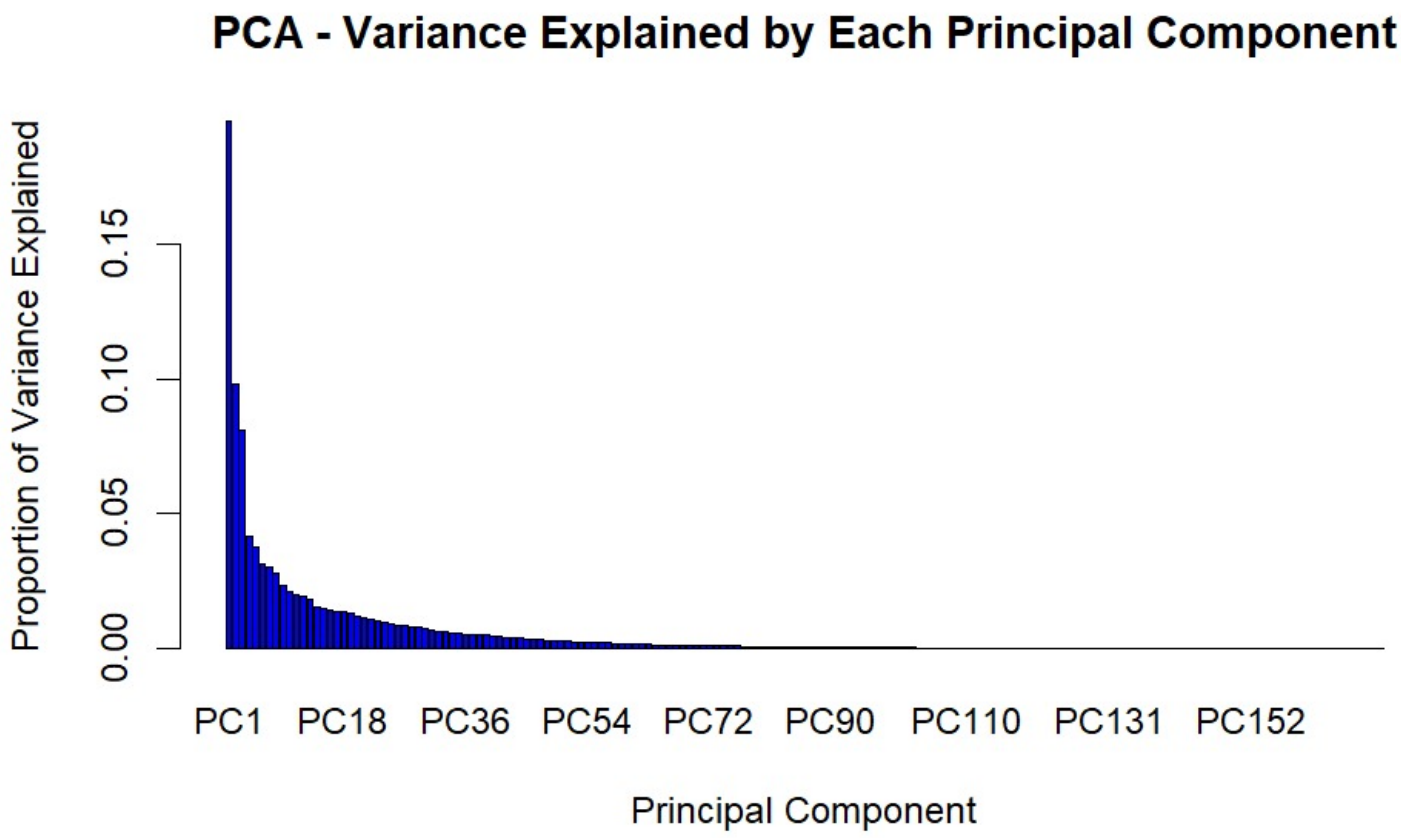
Setting our significance level at 0.05, the test indicated that only 34 variables are influential in determining molecular toxicity.

minHBint4	ECCEN	MDEC.14	MDEC.23	SP.6	SP.5
SpAD_Dt	AATS8v	SpMax4_Bhm	ETA_Eta_F_L	SpDiam_Dt	nC
naAromAtom	SpMin3_Bhi	nHaaCH	ETA_Beta	nAcid	EE_Dt
nBondsD	ETA_Beta_ns	C2SP2	GATSV7v	SpMin4_Bhs	SpMin4_Bhi
SpMin4_Bhe	SpMax_Dt	MLogP	nWHBa	khs.aaCH	ZM1C1
C3SP2	naaCH	SpMAD_Dt	WTPT.1		

Table: Student T tests selected variables

Variables selection : PCA (Principal Component Analysis)

- Creating a synthetic dataset.
- Selection of 42 principal components.



Baseline Model

- The model didnt converge as we are facing a high dimensionnality problem, some coefficients couldnt be estimated.
- The model overfitted as it is too complex.

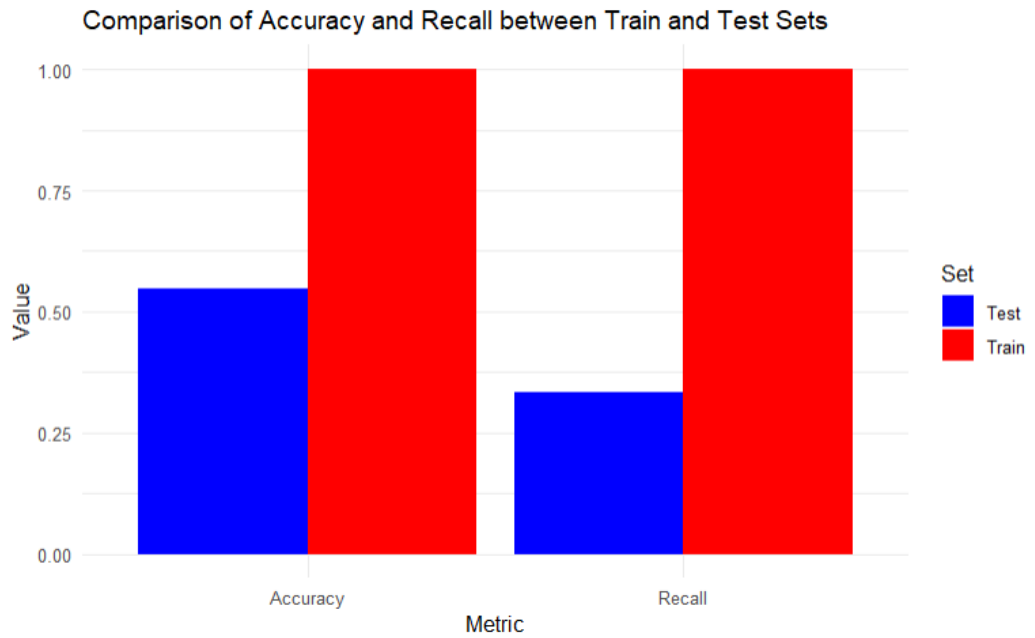


Figure: Baseline Model : Train vs Test performance

Baseline Model : using the t-test results

- The model has a better predictions performance due to the selected variables having a relatively high mutual information score with our target class.
- No overfitting was identified.
- However, many of these features are highly correlated.

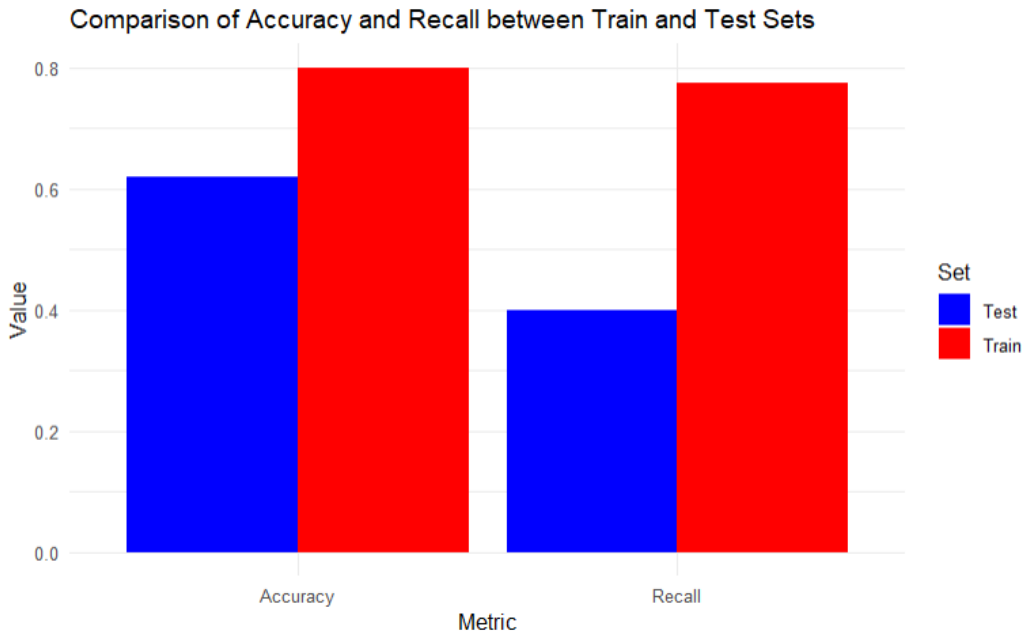
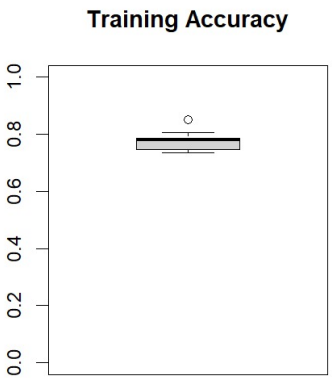
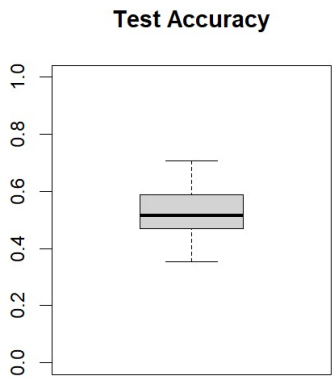


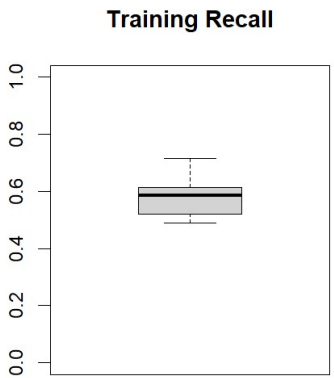
Figure: Train vs Test performance

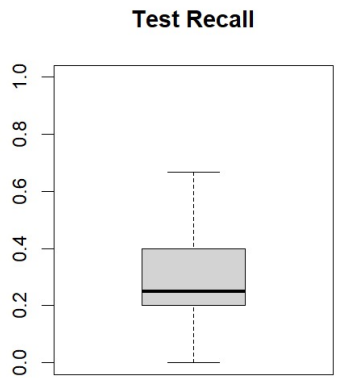
Baseline Model : using PCA results

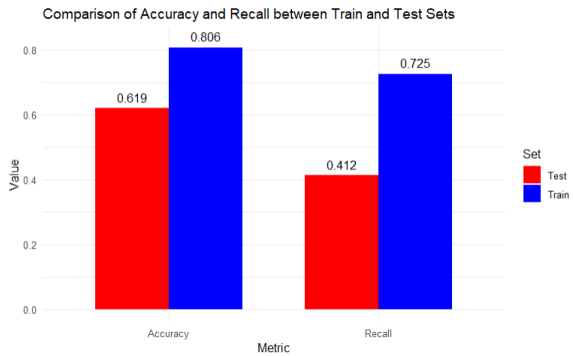
- Chosing PCs with a cumulative variance explained higher than 0.9.
- Good accuracy and a low residual Deviance .







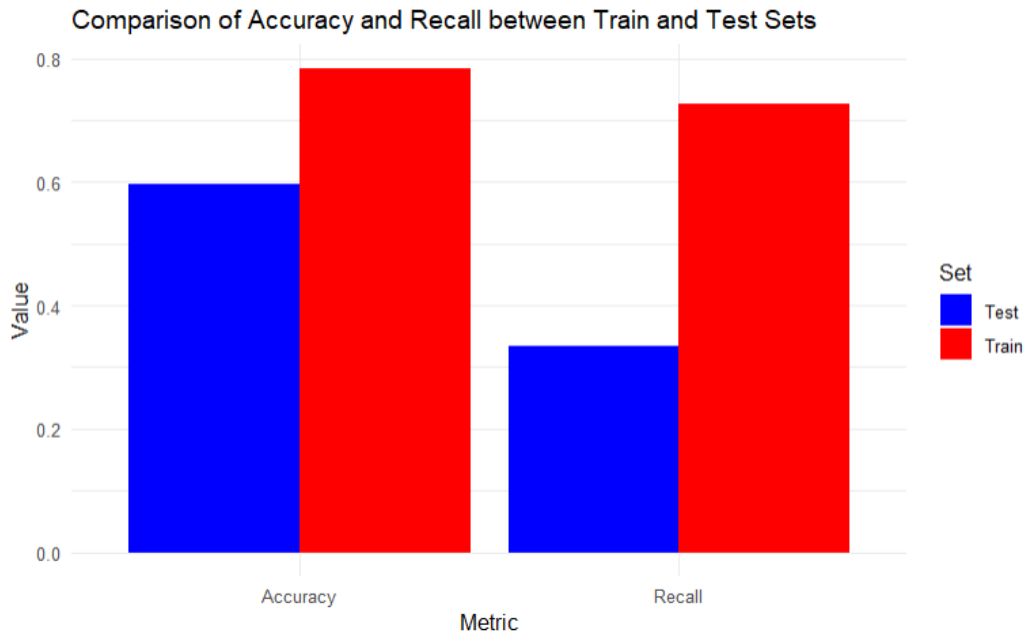




Forward Approach

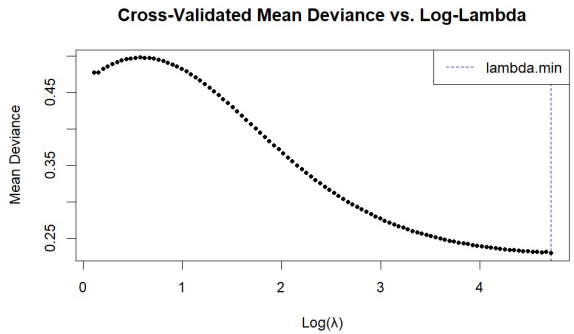
- Only 21 features were selected among which among 7 are significant according to the Ward test.
- Altough having a small AIC, it has a higher residual deviance compared to the baseline model.

Feature	Occurence
SpMax4_Bhm	Student + Forward

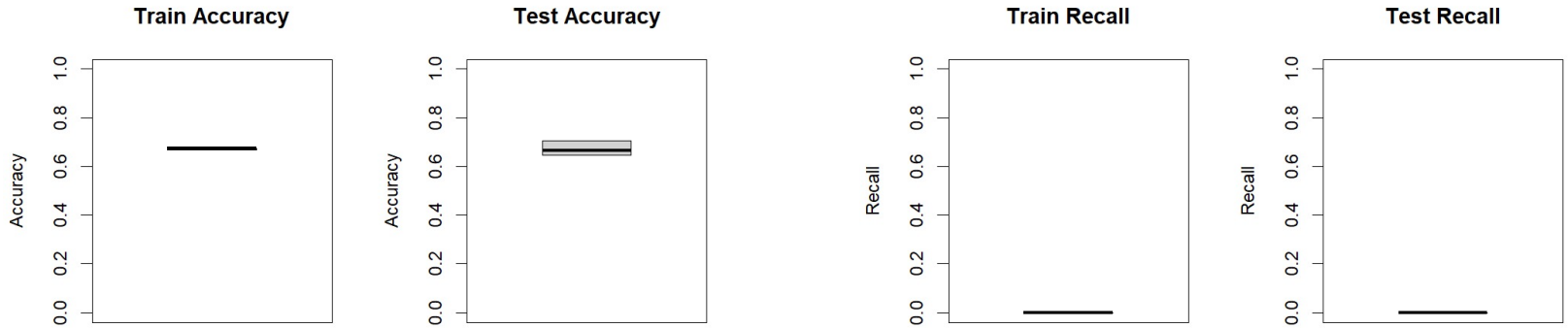


Logistic regression with Ridge penalization

- Optimal λ values are quite large.

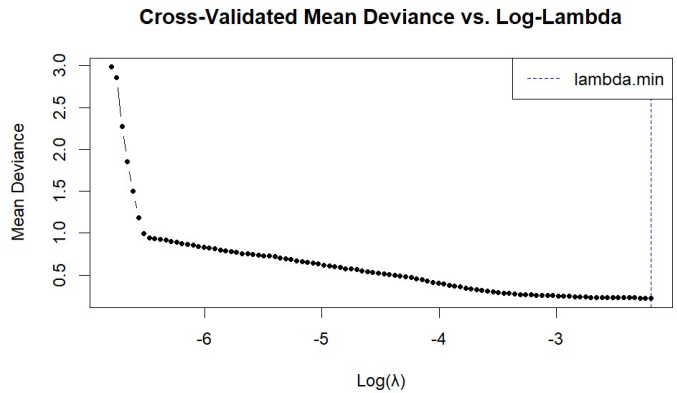


- Altough having a good accuray, the model has a very low recall on the positive class.

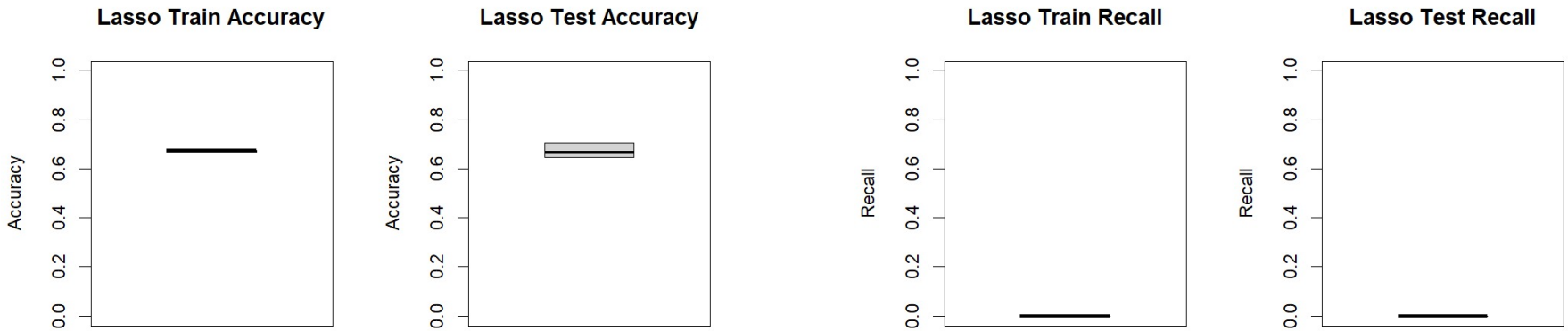


Logistic regression with Lasso penalization

- No features except the Intercept were selected for λ.min .
- 5 features were selected for λ.1se close to 0.

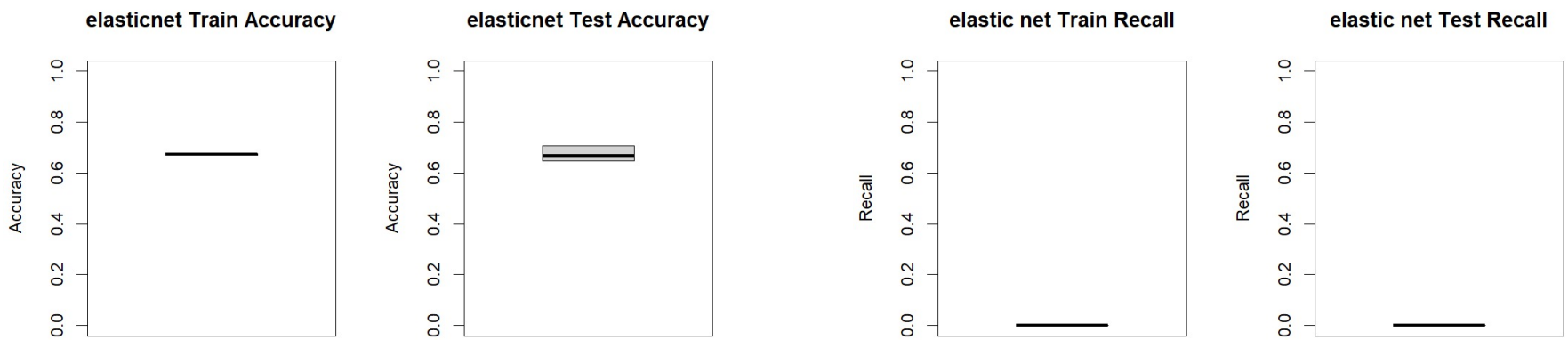


- Misleading high Accuracy but very low recall.



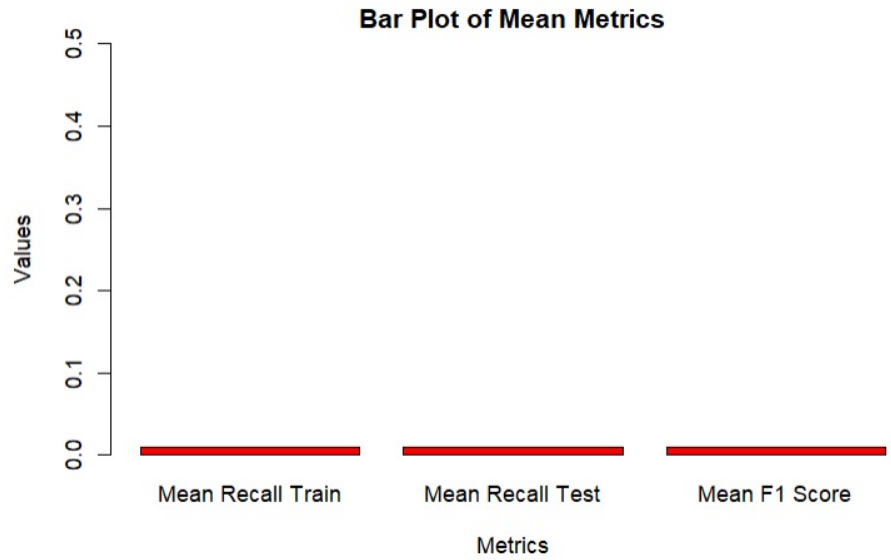
Elastic Net

- Equally weighted mixture of the two penalization $\alpha = 0.5$.



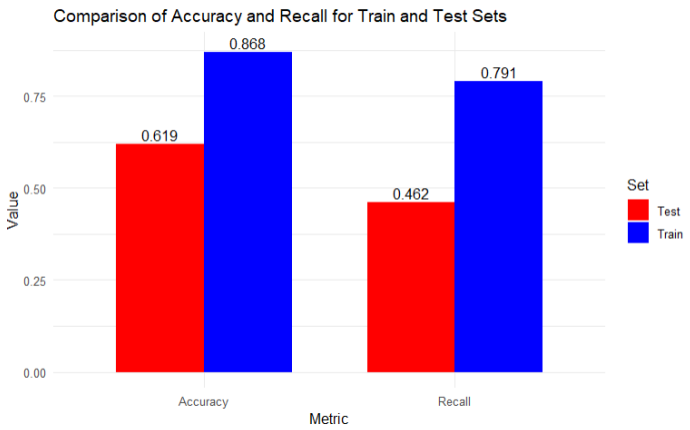
Comparison of the three penalizations

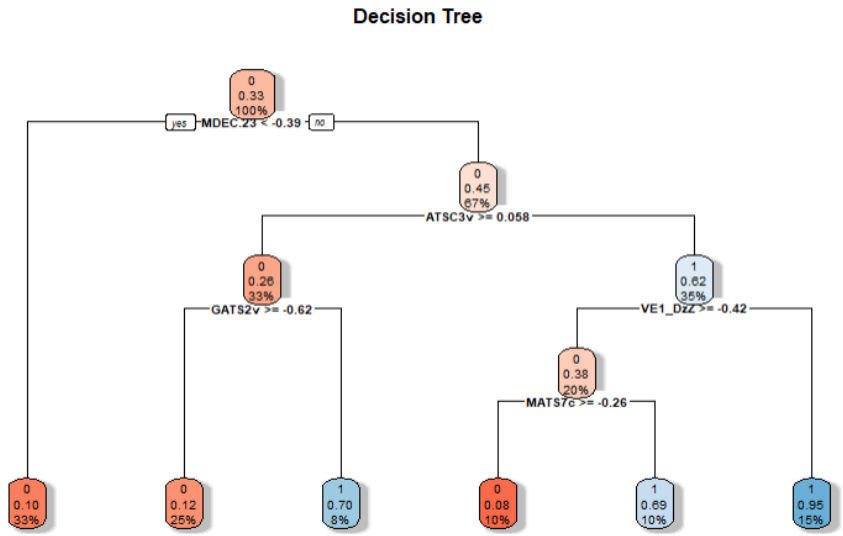
- The Mean_Deviance($\log(\lambda)$) is a decreasing function.
- Chosen λ min shrinks all coefficients towards zero, leaving only an intercept below 0.5 for Lasso.
- RESULT: all the regularization techniques overpredicted the majority class. The recall of Toxic classes is 0 .



Decision Tree classifier

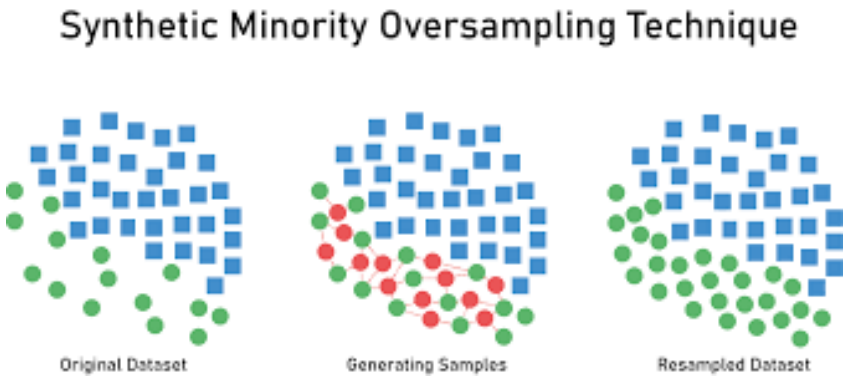
- High recall for toxic molecules, unlike regularized regression.
- 5 feature were selected after implementing pruning technique.
- Presence of some features quoted in the research paper.





Conclusion and Possible Prospects

- High dimensionality and data imbalance the dataset.
- Significance misinterpretation due to multicollinearity.
- Models couldnt outperform the baseline model.



+ How to improve the predictive performance?
- Trying weighted classes and oversampling techniques (SMOTE)
- Use other dimensionality reduction techniques.
- Use more robust classification models like XGBoost and Random Forest.