MERR Project Report

Toxicity Analysis

SAES-VINCENSINI Mickaël, YANG Charles, YAO Luc

ENSIIE

2023-12-18

Nature of the problem

- The Research conducted focuses on the structure-based design and classification of small molecules that regulate the circadian rhythm period.
- The objective is to find out the molecular descriptors that helps to determine whether a molecule is toxic or not.

What is the dataset made of ?

- The toxicity data set is composed of 171 molecules with 1538 molecular descriptors. 334 features repeating the same value for all molecules were discarded. The remaining 1203 features were utilized to obtain the best feature to explain the toxicity of the molecules.
- Our dataset is composed of integers and floats variables
- We are trying to explain the toxicity of a given molecule, represented as a binary variable in our dataset.

Raw data graphics

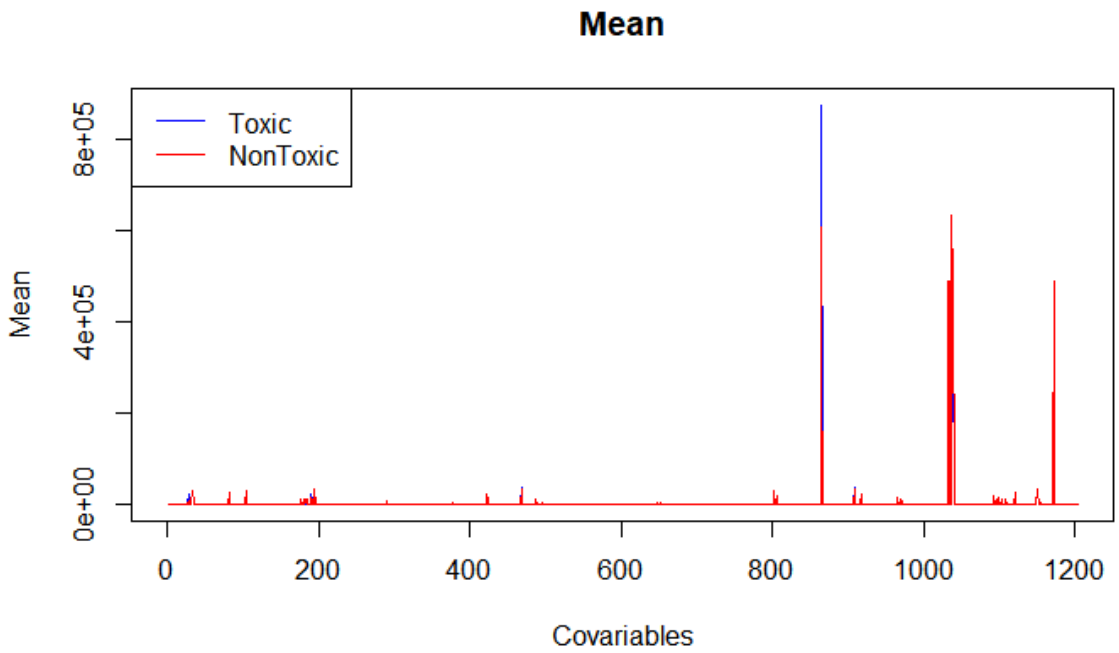

Figure 1 : Mean of raw covariables

How to normalize ?

Let D the dataset,
For any X $\in D$
We have $X = \begin{pmatrix} X_1 \\ ... \\ X_{1204} \end{pmatrix}$

Let X_{min} the minimal value of every X_i,
Let X_{max} the maximal value of every X_i
We apply the following formula for each $X_i \in X$ for i in [1, 1204]

$$NewX_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

With this, each $NewX_i \in [0, 1]$

Normalized data graphics

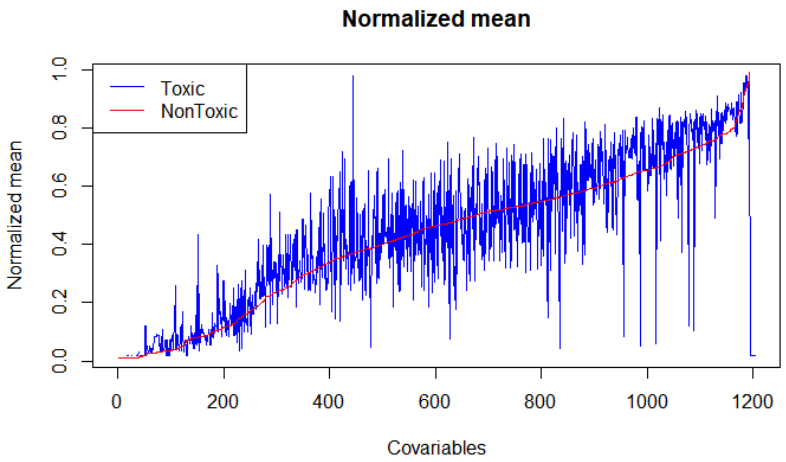

Figure 2 : Mean of normalized covariables

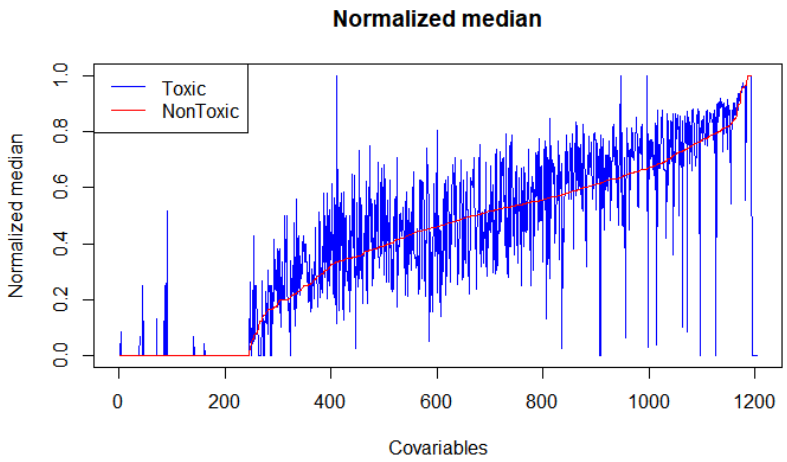

Figure 3 : Median of normalized covariables

We used the Student T-test to determine the variable that make a significant difference between the toxic and non toxic data.

Student test

Covariables

minHBint4	SpMin3_Bhm	SpMin4_Bhi
ECCEN	SpMin3_Bhi	SpMin4_Bhe
MDEC.14	nHaaCH	SpMax_Dt
MDEC.23	ETA_Beta	MLogP
SP.6	nAcid	nwHBa
SP.5	EE_Dt	khs.aaCH
SpAD_Dt	nBondsD	ZMIC1
AATS8v	ETA_Beta_ns	C3SP2
SpMax4_Bhm	C2SP2	naaCH
ETA_Eta_F_L	GATS7v	SpMAD_Dt
SpDiam_Dt	SpMin4_Bhs	WTPT
nC	SpMin3_Bhe	naAromAtom

Too much variable

Since we have n = 117 (observations) << p = 1203 (covariables)
We know that the matrix $X^T X$ is a non inversible matrix
Which implies that if we try to compute a lasso regression model, it will not converge.
We need to reduce the number of variables we are choosing to compute, by cleaning useless data.

Quantile

In order to reduce our number of variables, we are using the following method :

- We compute the absolute value of every value of a variable, and we compute the quantile of it, with a value of 70%
- Then we delete every variable which has a null quantile

We reduced the size of our dataset by 20% (1204 $\rightarrow$ 977 variables)

Logistic regression model

Variables :

- Y binary target variable ("Class") $\{0, 1\}$
- X covariates $X \in \mathbb{R}^n$

Sample of data i.i.d. :

- n = 171 i.i.d. observations
- $\mathcal{D}_n = \{(x_i, y_i), 1 \leq i \leq n, y_i \in \{0, 1\}\}$

For a given observation x_i, the model is :

- $\eta(x_i) = \mathbb{P}(Y = 1/X = x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$
- The β parameters are estimated by the maximum Likelihood.

Logistic regression model

3 main steps

1. Estimation of the parameters of the model (β)
   - Estimation of $\hat{\beta}$ using the data
   - Variable selection using the Lasso penalization
2. Prediction (β)
   - for a new observation $x_{new} \notin \mathcal{D}_n$
   - Estimation of the probability : $\hat{\eta}(x_{new}, \hat{\beta}) = \frac{e^{\hat{\beta}^T x_{new}}}{1+e^{\hat{\beta}^T x_{new}}}$
3. Decision - Map (Maximum A Posteriori)
   - Given the value of a chosen threshold S $\in [0, 1]$
     - $\hat{Y} = 1$ if $\hat{\eta}(x_{new}, \hat{\beta}) > S$
     - $\hat{Y} = 0$ otherwise

Statisyical approach : forward selection

Step by step method :

- 1^{st} step : null model
- k^{th} step : the variable which reduced the AIC is added to the previous model
- Stop condition : model which we add no more variable has the lowest AIC

Machine Learning approach : penalization

Lasso penalization :

- minimise $\Phi(\beta) = (Y - X\beta)^T(Y_X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$
- Selection of important variables

Logistic regression model

1. Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
   - TP : True positive
   - TN : True negative
   - FP : False positive
   - FN : False negative
2. AIC $= -2\times$log-likelihood $+2\times$number of parameters
3. Residuals $=$ Observed Value Predicted Value

Logistics models

Since we can choose the λ of our models to select how much variable we choose to have in our model, we computed the performance of the model comparing the number of selected variables, for every criteria.

Logistics models : Maximizing Accuracy

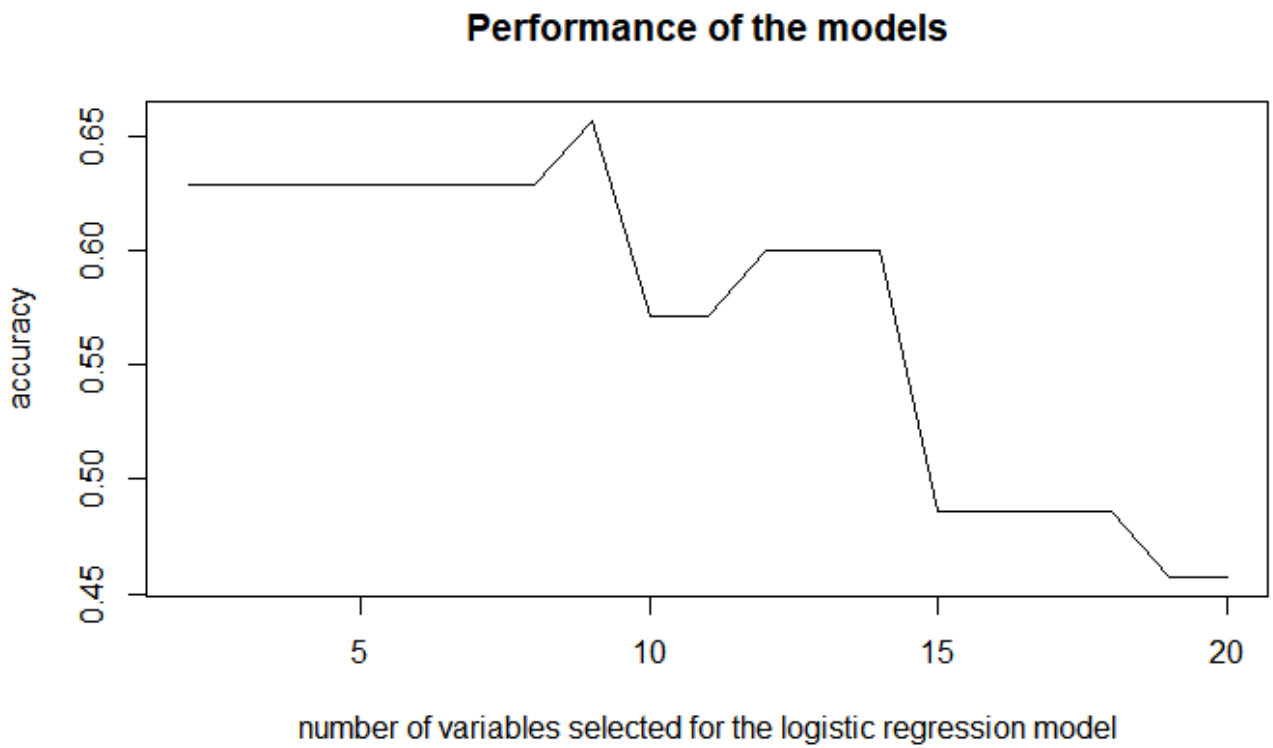

Figure 4 : Performance of the logistics model maximizing the accuracy

Logistics models : Minimizing AIC

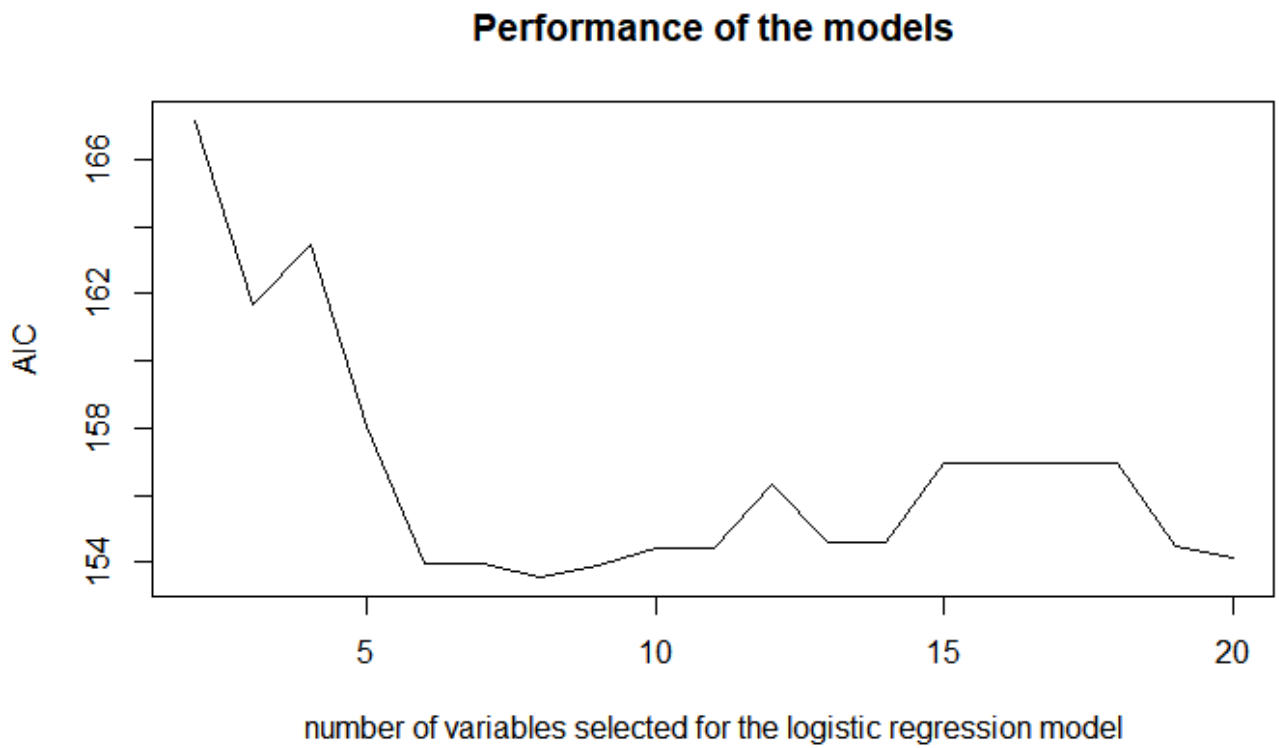

Figure 5 : Performance of the logistics model minimizing AIC

Logistics models : Minimizing Residuals

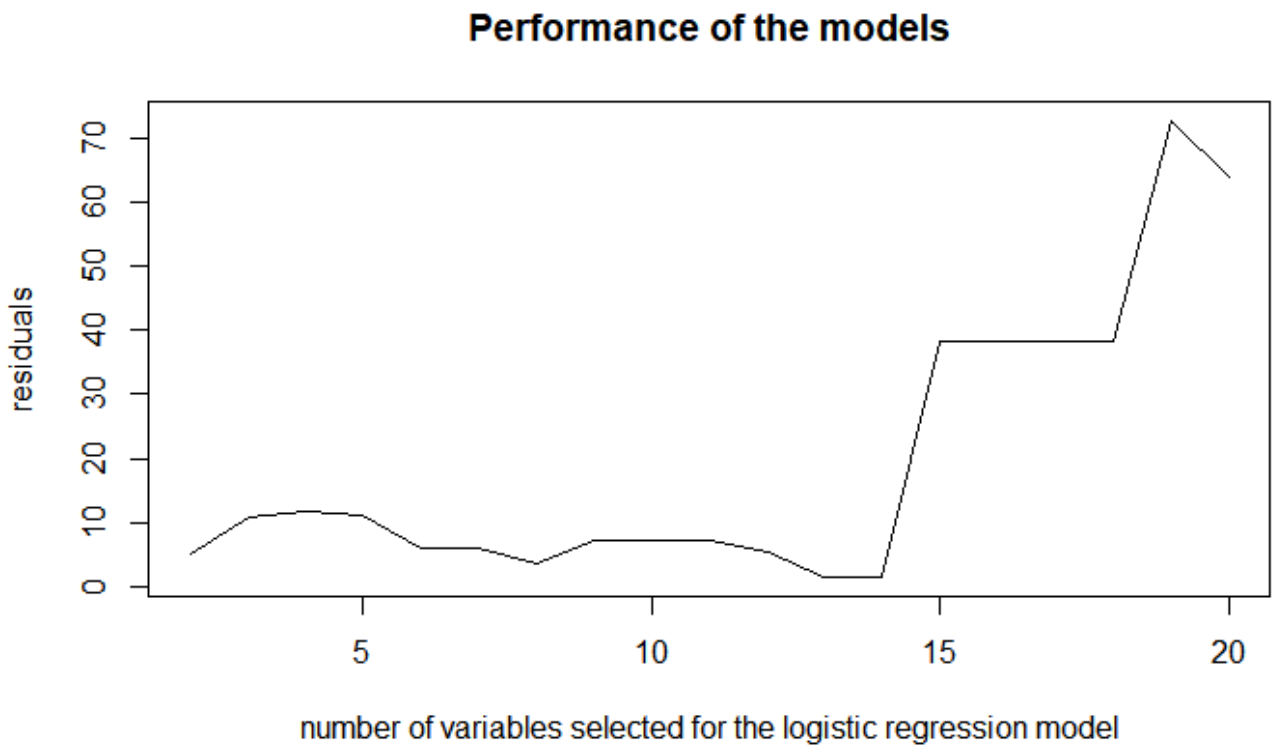

Figure 6 : Performance of the logistics model minimizing Residuals

Logistics models

It gives us 3 models, depending of the number of variables, which minizize or maximize our criterias.
We also have the forward model, and we will compare them

Summary

Logistic Model	Nb of variables	AIC	Residuals	Accuracy
AIC_{min}	8	153.6	3.61	63%
$Residuals_{min}$	13	154.6	1.42	60%
$Accuracy_{max}$	9	153.9	7.12	66%
Forward	Nb of variables	AIC	Residuals	Accuracy
	11	145.1	270	77%

Summary table

Selected covariables

covariable	coefficent	p-value	covariable	coefficent	p-value
SpMin3_Bhi	8.57	.	VR1_Dt	3.98e-07	.
AATSC6s	-4.71	**	AATSC2e	-1.06e+02	*
khs.aaN	-5.16e-01	*	SpMax4_Bhm	5.90	*
AATSC5m	1.96e-01	*	ATSC7p	1.40e-01	.

In red we highlighted the covariables that were already choosen by our first T-Test.