Circadian Rhythm Project - Molecules Toxicity

Mohammed-Yassine Barnicha & *Nezar Aberqi*

I. Introduction :

In this MRR report, we are looking into how computer technology and medicine work together to find new drugs. Our focus is on using machine learning to predict if chemicals are toxic or safe, and their effects on the body's natural daily cycle. We studied a large dataset with 171 different cells that the researchers classified into toxic and non-toxic categories, each described by around 1500 characteristics. The key aim was to categorize these chemicals into harmful or safe groups and understand their impact on the body's daily cycle, especially looking at the CRY1 protein in certain cells. Our report will use the study's discoveries for our own analysis. The study identified essential characteristics of the chemicals that can predict their safety and their effect on the body's cycle. These insights are very useful for developing new medicines and show us more about how to use computer technology in medical research.

II. About our data:

A. Summary of the medical report :

1. Toxicity Classification :

A Decision Tree Classifier (DTC) identified a subset of 13 molecular descriptors that accurately predict toxicity with around 80% accuracy.(MDEC-23, MATS2v, ATSC8s, VE3_Dt, CrippenMR, SpMax7_Bhe, SpMin1_Bhs, C1SP2 GATS8e, GATS8s, SpMax5_Bhv, VE3_Dzi ,VPC-4)

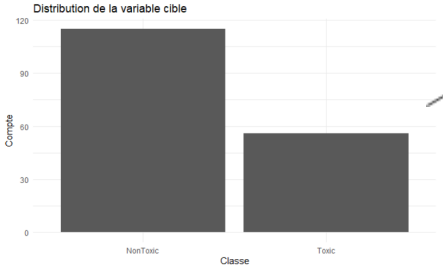
2. Period Lengthening Classification :

The study also investigated the impact of non-toxic molecules on the length of the circadian rhythm in U2OS cells.
An XGBoost Classifier (XGBC) identified 10 molecular descriptors that predict molecules capable of lengthening the circadian period with about 87% accuracy.(ATSC8c, MATS1e, minsCH3, MATS4e, MATS4s, ATSC7i, SpMin4_Bhp, MLFER_S, ATSC4p, SpMax2_Bhm)

B. Our Analysis :

1. Data Types:

```
     MATS3v              nHBint10
Min.   :-0.31150   Min.   :0.0000
1st Qu.:-0.06670   1st Qu.:0.0000
Median :-0.03250   Median :0.0000
Mean   :-0.03124   Mean   :0.3158
3rd Qu.: 0.00485   3rd Qu.:0.0000
Max.   : 0.14110   Max.   :4.0000
```



- Our feature are primarily of numeric, either continuous or discrete, with a large number of continuous variables, the wide range of means and standard deviations across the features indicates that the features are likely to have different scales, therefore, feature scaling might be necessary for models that are sensitive to the scale of the data.
- Our dataset contains no missing value, across all the 1203 features, no imputation is needed.
- The distribution of features is quite different, we notice that there are some that are highly skewed, and some others that are almost gaussian.
- The target variable is a binary response variable that is either equal to “Non Toxic” or “Toxic”, indicating whether the molecule is toxic or not, we notice that our dataset is imbalanced, there are twice as much ‘Non toxic’ (Majoritary Class) samples as ‘Toxic’ ones.

2. Feature Analysis and Selection :

Student t-Test :

As there is a very large number of variable in our dataset, we run the Student t-statistic in order to highlight the features in which the distribution depends on the whether the molecule is toxic or not, we stored the results of the test in a dataframe containing the different features and their correspondent **p-value** of the t-statistic.

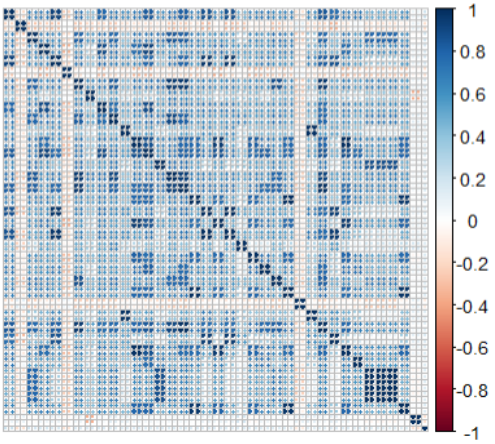
For a risk threshold of 0.05, we concluded that we reject the null hypothesis for only 36 variables, these variables have different behaviors depending on the molecule toxicity, therefore, we might consider these features to have a greater explanatory potential than the remaining features.

```
##  [1] "minHBint4"   "ECCEN"       "MDEC.14"     "MDEC.23"     "SP.6"
##  [6] "SP.5"        "SpAD_Dt"     "AATS8v"      "SpMax4_Bhm"  "ETA_Eta_F_L"
## [11] "SpDiam_Dt"   "nC"          "naAromAtom"  "SpMin3_Bhe"  "SpMin3_Bhm"
## [16] "SpMin3_Bhi"  "nHaaCH"      "ETA_Beta"    "nAcid"       "EE_Dt"
## [21] "nBondsD"     "ETA_Beta_ns" "C2SP2"       "GATS7v"      "SpMin4_Bhs"
## [26] "SpMin4_Bhi"  "SpMin4_Bhe"  "SpMax_Dt"    "MLogP"       "nwHBa"
## [31] "khs.aaCH"    "ZMIC1"       "C3SP2"       "naaCH"       "SpMAD_Dt"
## [36] "WTPT.1"
```

Selected Features from the Student t-test

Analysis of the Selected Features :

It is interesting to study the main properties of the obtained 36 features. We will try to see the relationship between these features that we consider to be important. We also included the target variable to see if it has any significant linear relationship with any of these features. The computation of the correlation matrix of the 36 features resulted in the following correlation heatmap :We notice that these features have a general tendency of being positively correlated, some of them are highly correlated.



Correlation Heatmap

```
##                        SP.6           SpAD_Dt       SpMin3_Bhe
## variable1              SP.6           SpAD_Dt       SpMin3_Bhe
## variable2              SP.5         SpDiam_Dt       SpMin3_Bhm
## correlation 0.98201409397246 0.964512826477179 0.97596044156147
##                   SpMin3_Bhe        SpMin3_Bhm      ETA_Eta_F_L
## variable1         SpMin3_Bhe        SpMin3_Bhm      ETA_Eta_F_L
## variable2         SpMin3_Bhi        SpMin3_Bhi         ETA_Beta
## correlation 0.998129018105592 0.967984699466117 0.98832321978461
##                    SpDiam_Dt       ETA_Eta_F_L          ETA_Beta
## variable1          SpDiam_Dt       ETA_Eta_F_L          ETA_Beta
## variable2              EE_Dt          nBondsD           nBondsD
## correlation 0.927245133214255 0.924820446898554 0.906504672527478
##                  ETA_Eta_F_L          ETA_Beta          nBondsD
## variable1        ETA_Eta_F_L          ETA_Beta          nBondsD
## variable2        ETA_Beta_ns       ETA_Beta_ns       ETA_Beta_ns
## correlation 0.951108380408378 0.950657146492826 0.94154453767645
##                   SpMin4_Bhs        SpMin4_Bhs        SpMin4_Bhi
## variable1         SpMin4_Bhs        SpMin4_Bhs        SpMin4_Bhi
## variable2         SpMin4_Bhi        SpMin4_Bhe        SpMin4_Bhe
## correlation 0.949862946826018 0.959707538731308 0.996166052123197
##                      SpAD_Dt         SpDiam_Dt                nC   nHaaCH
## variable1            SpAD_Dt         SpDiam_Dt                nC   nHaaCH
## variable2           SpMax_Dt          SpMax_Dt             MLogP khs.aaCH
## correlation 0.999937138032318 0.964797570539982 0.917244057857881        1
##             nHaaCH khs.aaCH          SpAD_Dt          SpMax_Dt
## variable1   nHaaCH khs.aaCH          SpAD_Dt          SpMax_Dt
## variable2    naaCH    naaCH          SpMAD_Dt          SpMAD_Dt
## correlation      1        1 0.940634682237468 0.938989820064587
```

After examining the correlation matrix, we've tried to output the highest correlated variables(with a 0.9 threshold) to have a better understanding of the relationship between these explanatory variables.

Highest correlated variables (>0.9 threshold)

- In order to study the impact of these 36 features, we created several KDE plots.

In order to see how the response variable changes with respect to these features, let's take **SpMin4_Bhi** as an example.

We can see that for this feature, the chances that the molecule is toxic get higher as the SpMin_Bhi increases.