team3-Final

LD

Introduction

CRY1 cryptochrome circadian regulator 1 is a gene that helps regulate the circadian clock or the body's internal clock (NIH.gov). The gene is activated by proteins including CLOCK/ARNTL and is turned off in a feedback loop with proteins such as PER/CRY. Therefore, changes in the genes can affect sleep patterns.

The Dataset

The dataset includes records of 90 small molecules created for CRY1. Per dataset creators, 27 of these molecules make the circadian rhythm longer while the remaining 63 molecules have no affect. There are 1177 molecular discriptive features.

Objective

Perform EDA, feature selection and classification tasks. Use Gradient Boosting ML algorithms to identify the best set of molecular descriptors for the model. Finally, perform comprehensive predictive modeling techniques on the data.

EDA

Load Libraries

```r
library(knitr)
library(ggplot2)
library(plyr)
library(dplyr)
library(corrplot)
library(caret)
```

Read data

```
df_data <- read.csv("data.csv", stringsAsFactors = F)
df_period <- read.csv("Period-Changer-10F.csv", stringsAsFactors = F)
```

```
dim(df_data)
```

```
[1]   90 1178
```

```
sum(is.na.data.frame(df_data))
```

```
[1] 0
```

```
dim(df_period)
```

```
[1] 90 11
```

```
sum(is.na.data.frame(df_period))
```

```
[1] 0
```

```
str(df_data[,c(1:10)])
```

```
'data.frame':   90 obs. of  10 variables:
 $ MATS3v         : num  0.0908 0.0213 0.0018 -0.0251 -0.0094 -0.0619 -0.017 -0.0154 -0.0133
 $ nHBint10       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MATS3s         : num  0.0075 0.1144 -0.0156 -0.0064 0.0132 ...
 $ MATS3p         : num  0.0173 -0.041 -0.0765 -0.0894 -0.1035 ...
 $ nHBDon_Lipinski: int  0 0 2 3 0 1 1 0 0 0 ...
 $ minHBint8      : num  0 0 0 0 0 ...
 $ MATS3e         : num  -0.0436 0.1231 -0.1138 -0.0747 -0.0046 ...
 $ MATS3c         : num  0.0409 -0.0316 -0.1791 -0.1151 -0.087 ...
 $ minHBint2      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ MATS3m         : num  0.1368 0.1318 0.0615 0.0361 0.1063 ...
```

```
str(df_period)
```

```
'data.frame':   90 obs. of  11 variables:
 $ ATSC8c    : num  0.1053 -0.0037 0.565 -0.0643 0.026 ...
 $ MATS1e    : num  -0.0226 0.0504 -0.1202 -0.1401 -0.0104 ...
 $ minsCH3   : num  0 0 1.73 1.7 0 ...
 $ MATS4e    : num  -0.1538 0.162 0.1163 0.0831 -0.0318 ...
 $ MATS4s    : num  -0.0969 0.1151 0.0546 0.0213 0.0512 ...
 $ ATSC7i    : num  -1.61 -24.84 -47.08 -53.02 -7.41 ...
 $ SpMin4_Bhp: num  1.94 1.89 1.91 1.91 1.86 ...
 $ MLFER_S   : num  0 4.56 6.38 5.99 4.09 ...
 $ ATSC4p    : num  2.837 -0.603 -5.14 -5.761 -2.462 ...
 $ SpMax2_Bhm: num  4.19 4.07 4.18 4.18 4.04 ...
 $ Class     : chr  "NoChanger" "NoChanger" "NoChanger" "NoChanger" ...
```

```r
# Compute correlations and find correlated variables to remove
numeric_df <- df_data[sapply(df_data, is.numeric)]
correlations <- cor(numeric_df)
high_correlations <- findCorrelation(correlations, cutoff = .75)
length(high_correlations)
```

```
[1] 949
```

```r
# Check for degenerate distribution
degeneratecols <- nearZeroVar(df_data)

degenerate_table <- data.frame(
  Column = degeneratecols,
  Predictor = colnames(df_data)[degeneratecols],
  stringsAsFactors = FALSE
)
degenerate_table |> head() |> knitr::kable()
```

Column	Predictor
21	MDEC.14
36	mintN
37	nHsNH2
43	nF8Ring
51	StN
69	nsBr

```r
nrow(degenerate_table)
```

```
[1] 83
```

Resampling

Regularization