{)I0UP_AGZAVGB$Y157CHTP 

**2022-2023学年第2学期**

**《数据挖掘项目》**

**课程报告**

学 院 理学院

专业班级 大数据（中法）201

学 号

姓 名

成 绩

目录

[1 项目介绍 2](#_Toc104921525)

[2 数据探索 2](#_Toc104921526)

[2.1 数据基础情况 2](#_Toc104921527)

[2.2 数据分布情况 2](#_Toc104921528)

[2.3 数据检验 2](#_Toc104921529)

[3 数据清洗 2](#_Toc104921530)

[3.1 数据筛选 2](#_Toc104921531)

[3.2 缺失值处理 2](#_Toc104921532)

[3.3 异常值处理 2](#_Toc104921533)

[4 特征提取 2](#_Toc104921534)

[4.1 特征选择 2](#_Toc104921535)

[4.2 降耦处理 3](#_Toc104921536)

[5 化合物生物活性预测模型 3](#_Toc104921537)

[5.1 模型介绍与参数设置 3](#_Toc104921538)

[5.2 训练集实验结果 3](#_Toc104921539)

[5.3 测试集实验结果 3](#_Toc104921540)

[6 化合物AMDET性质分类模型 3](#_Toc104921541)

[6.1 模型介绍与参数设置 3](#_Toc104921542)

[6.2 训练集实验结果 3](#_Toc104921543)

[6.3 测试集实验结果 3](#_Toc104921544)

[7 总结 3](#_Toc104921545)

1. **项目介绍**

本项目的目标为XXXXXXXX，主要完成以下三方面工作：

1）根据文件“Molecular\_Descriptor.xlsx”和“ERα\_activity.xlsx”提供的数据，针对1974个化合物的729个分子描述符进行变量选择，根据变量对生物活性影响的重要性进行排序，并给出前30个对生物活性最具有显著影响的分子描述符（即自变量），并请详细说明分子描述符筛选过程及其合理性

2）结合问题1，选择不超过30个分子描述符变量，构建化合物对ERα生物活性的定量预测模型，请叙述建模过程。然后使用构建的预测模型，对文件“ERα\_activity.xlsx”的test表中的50个化合物的pIC50值进行预测，并给出预测精度

3）利用文件“Molecular\_Descriptor.xlsx”提供的729个分子描述符，针对文件“ADMET.xlsx”中提供的1974个化合物的ADMET数据，分别构建化合物的Caco-2、CYP3A4、hERG、HOB、MN的分类预测模型，并简要叙述建模过程。然后使用所构建的5个分类预测模型，对文件“ADMET.xlsx”的test表中的50个化合物进行相应的预测，并给出预测结果的混淆矩阵。

1. **数据探索**

## 数据基础情况

对数据做基本展示，选择部分数据列，分析该列的值分布，检查是否具有缺失值、异常值、重复值

library(readxl)

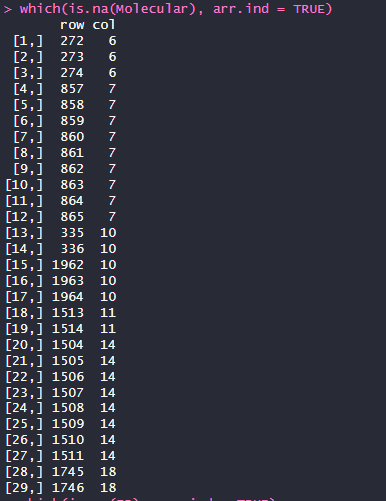
Molecular <- read\_excel("C:/Users/22867/Desktop/TEXTWORK/Licence3Suivant/CnProject/DataMing/Molecular.xlsx")

ER <- read\_excel("C:/Users/22867/Desktop/TEXTWORK/Licence3Suivant/CnProject/DataMing/ER.xlsx")

is.na(ER)

is.na(Molecular)

which(is.na(Molecular), arr.ind = TRUE)



#查看Molecular数据集中的缺失值的索引

which(is.na(ER), arr.ind = TRUE)

#查看ER数据集中的缺失值的索引

duplicated(ER)

duplicated(Molecular)

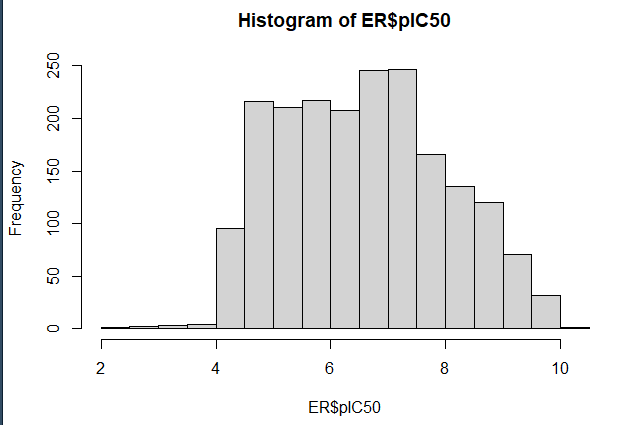
#检查重复值

ER\_sorted <- ER[order(-ER$pIC50),]

#对生物活性影响度坐降序排序

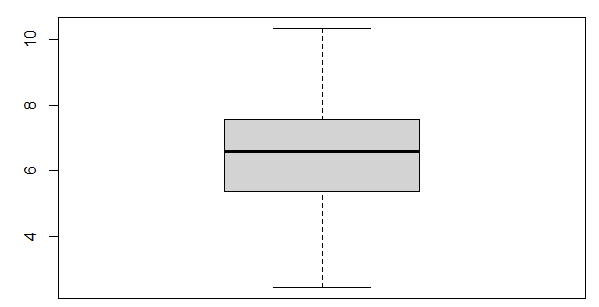
hist(ER$pIC50)

#查看不同的PIC50的分布频度图



boxplot(ER$pIC50)

#查看箱线图



## 数据分布情况

lapply(Molecular\_numeric, density)

#对Molecular数据集中的每列变量坐密度图

## 数据检验

完整性、一致性

1. **数据清洗**

## 数据筛选

## 缺失值处理

valeur\_manquer\_Molecular <- which(is.na(Molecular), arr.ind = TRUE)

# 将Molecular数据集中残缺值所在的坐标导入到valeur\_manquer\_Molecular数据集中

Molecular <- Molecular[-valeur\_manquer\_Molecular[,1],]

#使用了R语言的负数索引来删除数据集中的行。使用valeur\_manquer\_Molecular[,1]来获取缺失值所在的行号，并将其传递给负数索引，以删除这些行

ER <- ER[-valeur\_manquer\_Molecular[,1],]

同样的操作删除ER数据集中对应Molecular数据中的残缺数据行

## 异常值处理

Molecular <- na.omit(Molecular)

ER <- unique(ER)

并未在两个数据集中发现重复值和离群值

1. **特征提取**

## 特征选择

基于多种算法对特征进行选择

对Molecular数据集进行逐步回归法来筛选出最优自变量子集

在完成3.2中对数据集ER和数据集Molecular中的残缺值所在数据行的删除之后，

library(MASS)

model\_regression\_pas <- lm(pIC50 ~ ., data =Molecular)

step(model\_regression\_pas)

Molecular$pIC50 <- ER$pIC50

#在Molecular数据最后一列添加数据集ER中的生物活性值pIC50，作为筛选最优自变量子集的目标变量

Molecular\_regression\_pas <- Molecular[,2:ncol(Molecular)]

#防止Molecular数据集中第一列的化学结构式对逐步回归法产生干扰，并导致AIC负无限

model\_regression\_pas <- lm(pIC50 ~ . , data =Molecular\_regression\_pas)

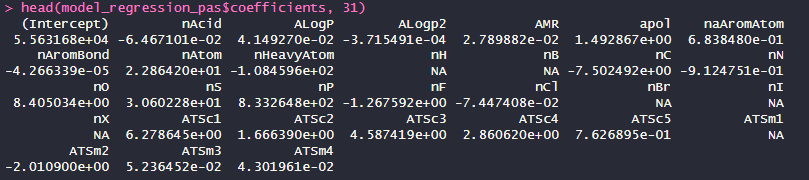
#建立线性回归模型

step(model\_regression\_pas)

#开始逐步回归

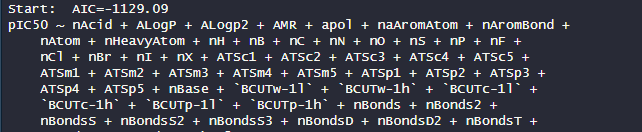
head(model\_regression\_pas$coefficients, 31)

#筛选出前30个对逐步回归法影响最强烈的决策变量，考虑到截距的存在，需要取前31个



summary(step(model\_regression\_pas))

# 评估模型



AIC = 1129.09

## 特征分析

1. **化合物生物活性预测模型**

## 模型介绍与参数设置

## 训练集实验结果

## 测试集实验结果

1. **化合物AMDET性质分类模型**

## 模型介绍与参数设置

## 训练集实验结果

## 测试集实验结果

1. **总结**

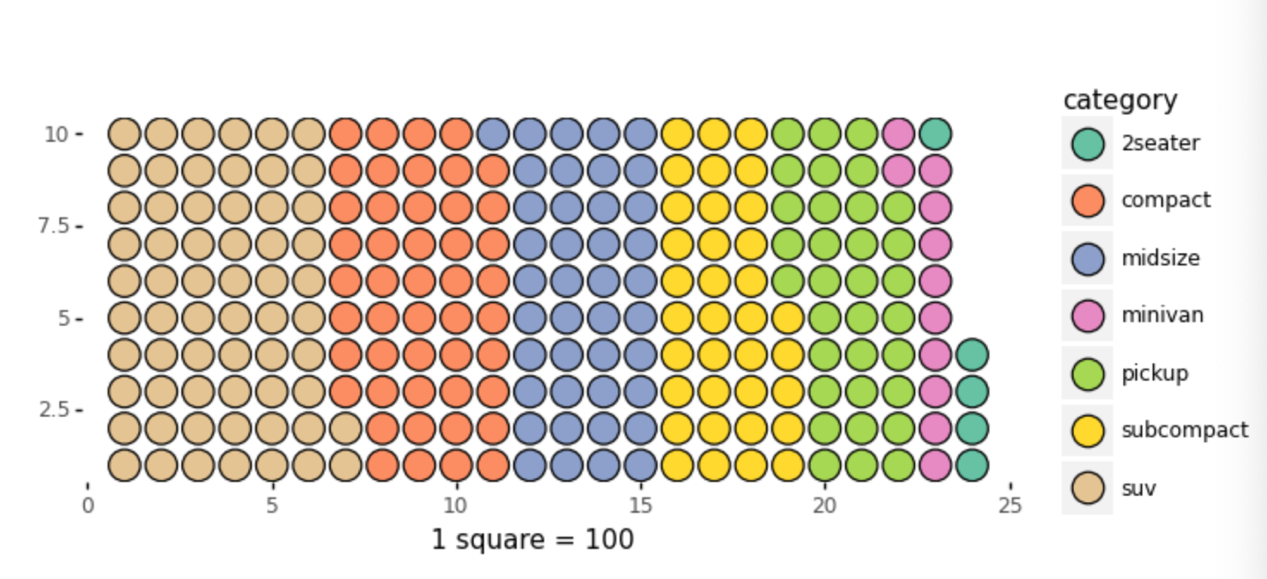


图1 示例图