

**实验(实训)课程报告**

**课程 数据挖掘技术实践**

|  |  |  |  |
| --- | --- | --- | --- |
| 院　　别 | 信管学院 | 专 业 | 计算机科学与技术 |
| 届 别 | 2026届 | 班 级 | 计算机科学与技术（职高）一班 |
| 学生姓名 | 邱宇杰 | 学 号 | 202205650121 |
| 指导教师 | 钟敏娟 | 职 称 | 教授 |

湖南财政经济学院

2024年 8 月

**实验(实训)报告5**

实验(实训)时间: 2024年 12月 14日 指导教师评分：

|  |  |
| --- | --- |
| 实验项目 | 实验五：药物处方规律挖掘 |
| 组员名单 | 独立完成或姓名（学号）（操作性和验证性实验均独立完成，综合性实验3-5人一组） |
| 实验类型 | □基础性实验 □验证性实验 ☑综合性实验 |
| 实验(实训)时间 | 2024年 12月 14日 |

1. 实验目的和要求：

1.1 实验目的

案例数据是Modeler软件自带的一份关于药物研究的数据。大批患有同种疾病的不同病人，服用五种药物中的一种(Drug，分为Drug A，Drug B，Drug C，Drug X和Drug Y)之后都取得了同样的治疗效果。现需要利用数据挖掘发现以往药物处方适用的规律，给出不同的临床特征病人更适合服用哪种药物的建议，为未来医生开具处方提供参考。

案例数据是随同挑选的部分病人服务药物前的基本临床检查数据，包括：病人的年龄(Age)、性别(Sex)、血压(BP)、胆固醇(Cholesterol)、唾液中钠元素(Na)和钾元素(K)含量。

1.2 实验要求

请根据Drug.txt数据，完成下列任务：

(1) 读入数据集，并浏览数据内容；

(2) 观察各个变量的数据分布特征；

(3) 利用散点图、直方图和网状图等形式分析进行探索性分析，包括钠和钾含量与服用药物的关系、钠和钾的比值与服用药物的关系、不同血压与服用药物的关系以及其他影响因素与服用药物的关系；

(4) 构建模型与模型评价，建立决策树模型，对模型结果数据进行解读，给出不同临床特征病人更适合服用哪种药物的建议；

**注意：（1）任务不仅要有截图，更多的需要有文字说明，解释图例数据与内容。**

2．实验步骤（记录程序代码、分析实验结果）

请按照以下目录框架来撰写实验过程

**[2.1 认识并理解数据](#_Toc848_WPSOffice_Level1)**

[1) 读入数据源](#_Toc7246_WPSOffice_Level2)

[2) 浏览数据内容](#_Toc13533_WPSOffice_Level2)

[3) 类型设置](#_Toc18706_WPSOffice_Level2)

[4) 观察变量数据分布](#_Toc1341_WPSOffice_Level2)

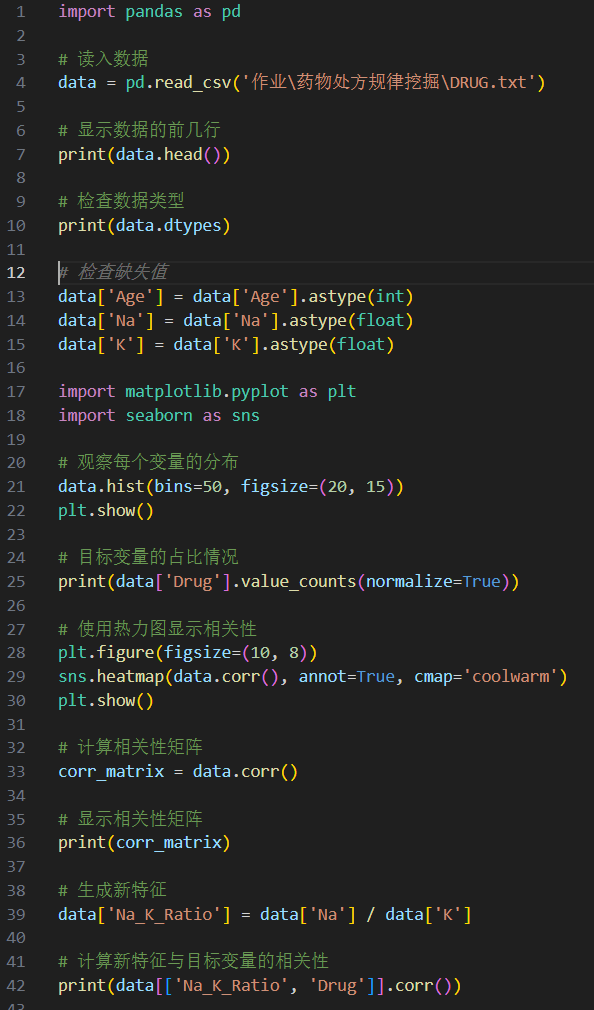
**[2.2 处理数据-试探分析](#_Toc14897_WPSOffice_Level1)**

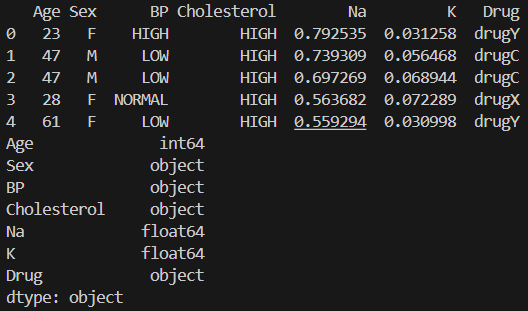
[1) 目标变量的占比情况](#_Toc28427_WPSOffice_Level2)

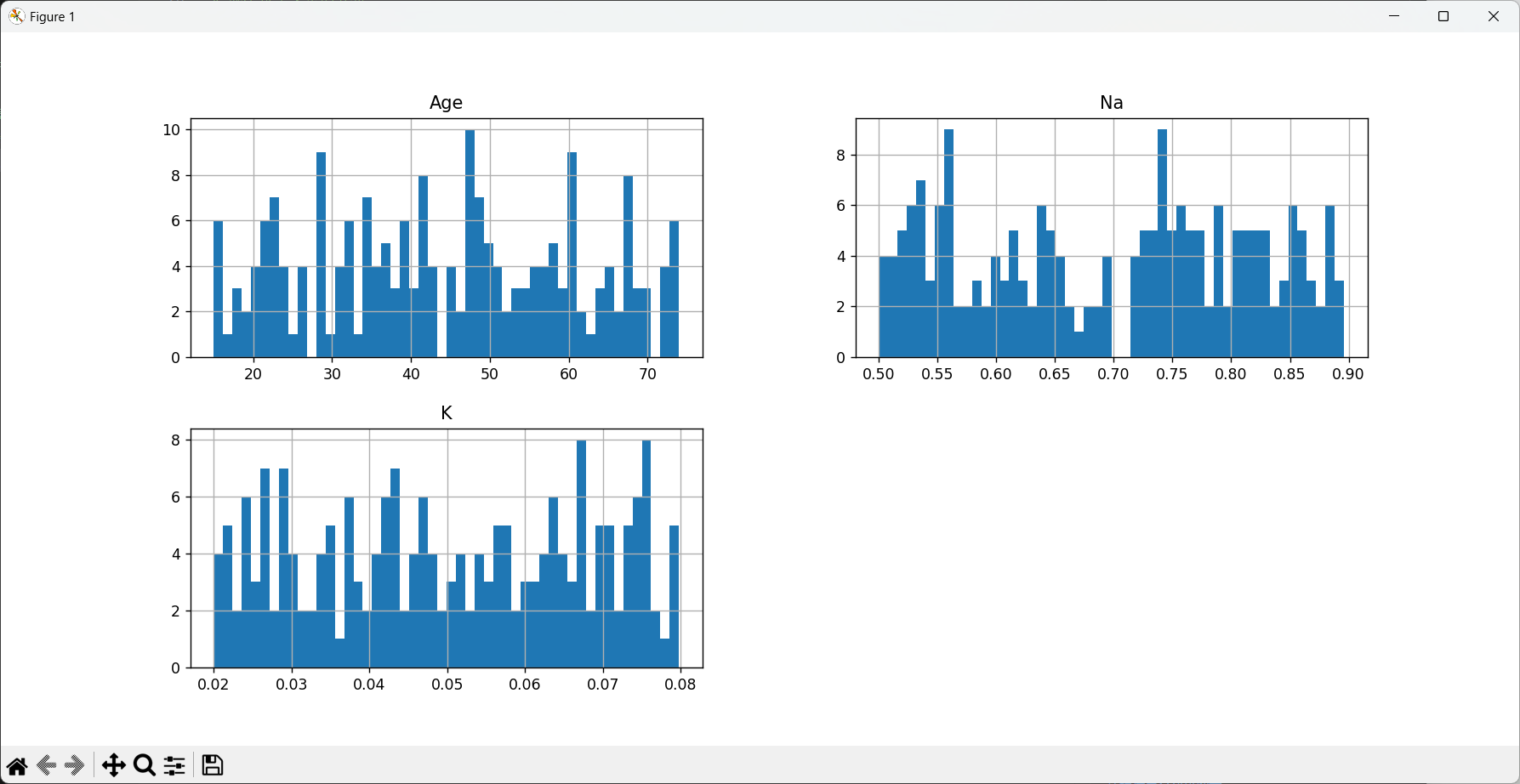
[2) 输入变量与目标变量之间的相关性分析](#_Toc31902_WPSOffice_Level2)

[2) 除目标变量外各变量之间的相关性](#_Toc16667_WPSOffice_Level2)

[3) 新特征生成](#_Toc369_WPSOffice_Level2)

[4) 生成的新属性与目标属性之间的相关性](#_Toc22203_WPSOffice_Level2)





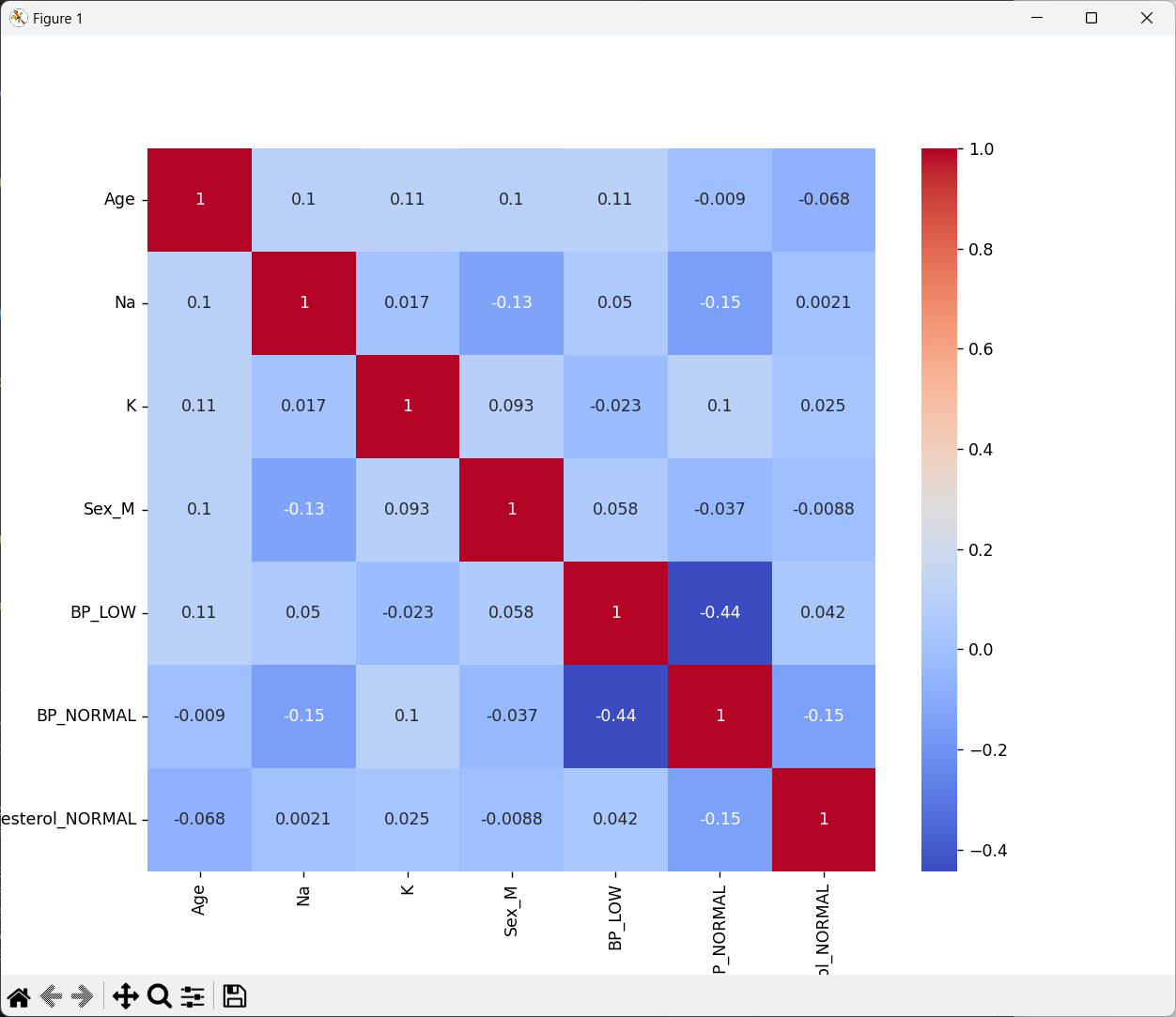
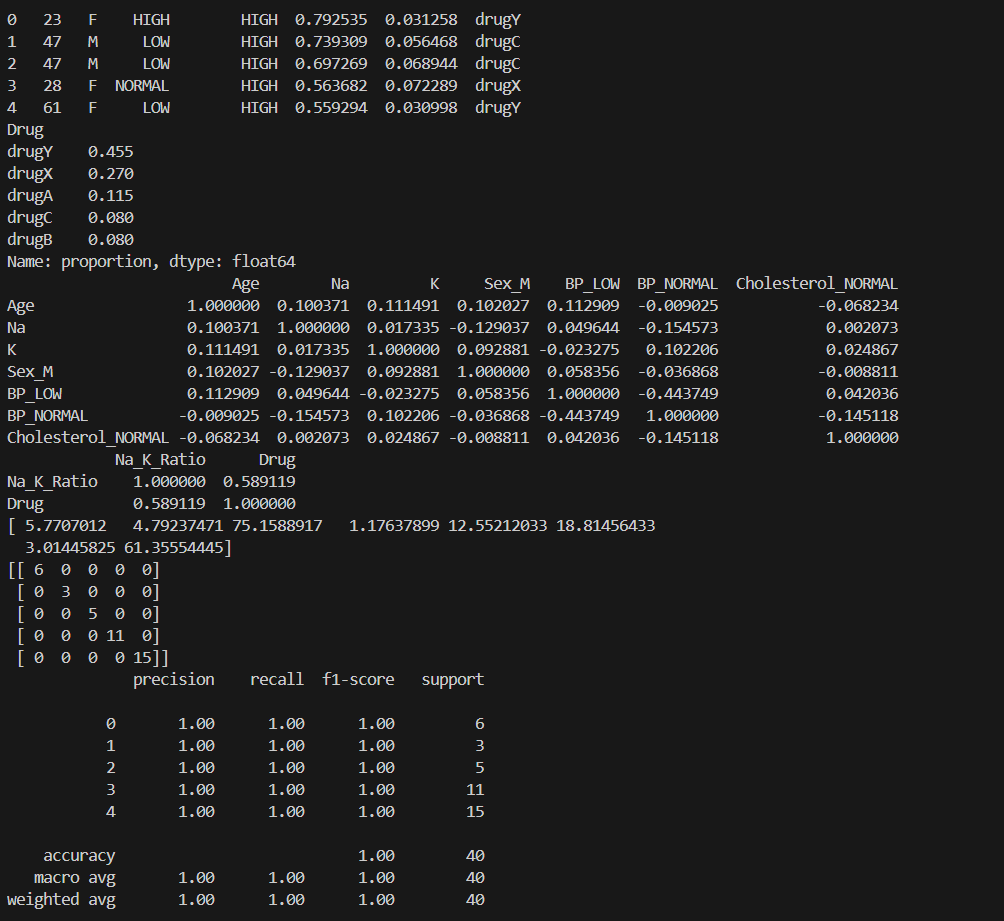
**[2.3 建模-深入分析](#_Toc26518_WPSOffice_Level1)**

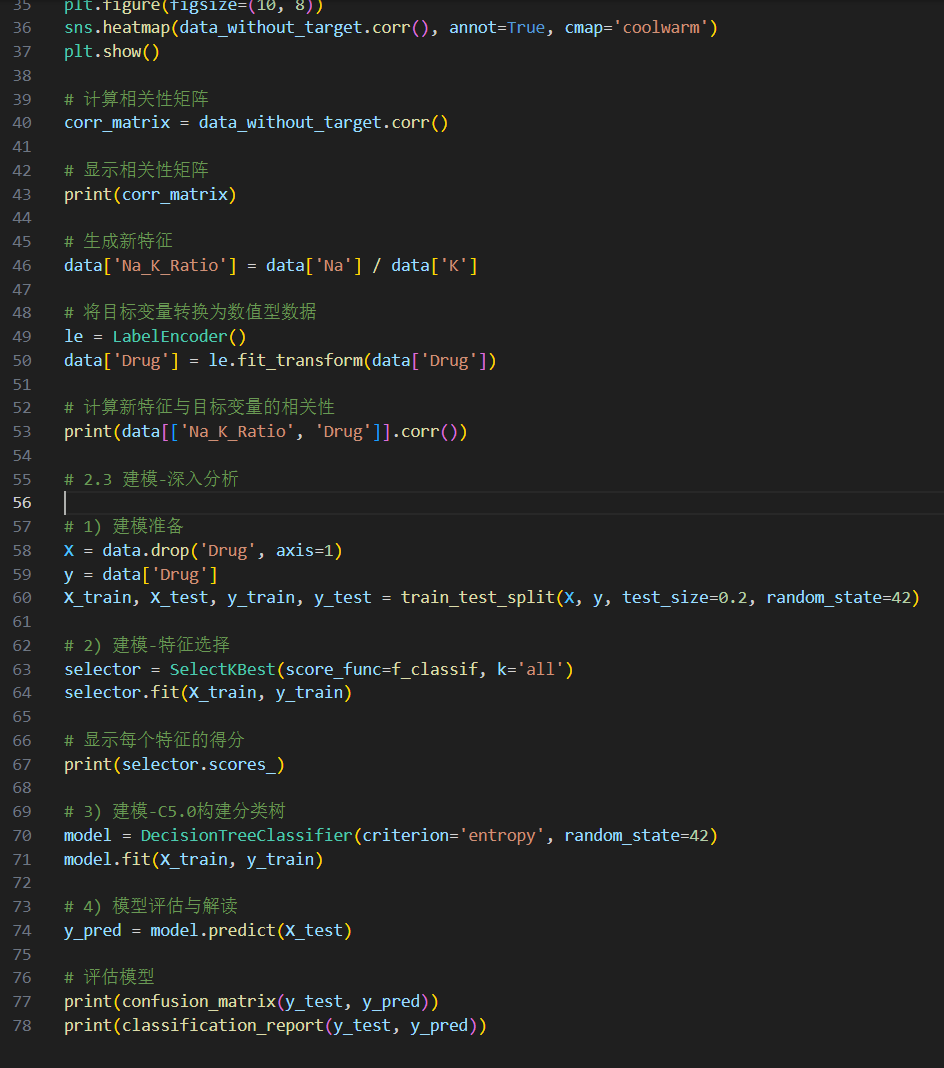
[1) 建模准备](#_Toc5354_WPSOffice_Level2)

[2) 建模-特征选择](#_Toc23909_WPSOffice_Level2)

[3) 建模-C5.0构建分类树](#_Toc848_WPSOffice_Level2)

[4) 模型评估与解读](#_Toc14897_WPSOffice_Level2)





目标变量 Drug 的占比情况如下：

drugY: 45.5%

drugX: 27.0%

drugA: 11.5%

drugC: 8.0%

drugB: 8.0%

变量之间的相关性

变量之间的相关性矩阵如下：

年龄与钠、钾、性别、血压和胆固醇之间的相关性较低。

钠和钾之间的相关性较低。

性别与其他变量之间的相关性较低。

血压（BP\_LOW 和 BP\_NORMAL）之间存在负相关。

胆固醇（Cholesterol\_NORMAL）与其他变量之间的相关性较低。

新特征与目标变量的相关性

新特征 Na\_K\_Ratio 与目标变量 Drug 之间的相关性为 0.589119，表明钠钾比值与药物类型之间存在一定的相关性。

特征选择

特征选择的得分如下：

Age: 5.77

Na: 4.79

K: 75.16

Sex\_M: 1.18

BP\_LOW: 12.55

BP\_NORMAL: 18.81

Cholesterol\_NORMAL: 3.01

Na\_K\_Ratio: 61.36

模型评估

使用决策树模型进行分类，模型的评估结果如下：

准确率（accuracy）为 1.00，表明模型在测试集上的表现非常好。

混淆矩阵显示所有类别的样本都被正确分类。

分类报告显示所有类别的精确率（precision）、召回率（recall）和 F1 分数（f1-score）均为 1.00。

1. 心得体会与建议

**从这次实验中我学到了很多分析数据的知识**