北京邮电大学实践课程实验报告

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程名称** | **创新创业实践** | **学院** | **计算机学院** | **指导老师** | **卢向群** |
| **知识模块** | **大数据** | **完成时间** | | **2024/11/17** | |
| **班级** | **序号** | **学号** | **学生姓名** | **成绩** | |
| **2023211321** | **22** | **2023212872** | **计子毅** |  | |

# 实验一 熟悉实验平台

## 实验目的

熟悉BDAP平台，工作流的基本操作和对应的组件0

## 实验内容

（这部分需要大家把进行实验的具体步骤分步骤描写，其中要附带组件的参数说明，即该组件的参数有什么意义，设置为多少等）

**第一步：**配置实验总体的工作流。包含组件：加载数据（加载选中的数据，并对其进行分割处理），列投影（选择数据中的部分属性列进行投影），表格（将传入的数据以表格的形式可视化出来），排序（将数据按一定规则进行排列），折线图（将传入的数据以折线图的方式可视化出来）。

**第二步：**配置工作流的各部分组件。

加载数据：文件路径是我们需要加载的数据的具体路径，参数为公共目录下的/Experiment\_1/Asia\_storm.csv；分隔符是将传入的内容以分隔符为界限进行切割，参数为：‘,‘是英文的逗号；是否带表头决定了数据分割后是否以原数据列的表头呈现，本实验参数为true；文件格式是文件路径中选择的文件的文件类型，若不一样则可能导致文件数据解析错误，参数为csv。

列投影：选择投影序列是将传入数据以选择的参数进行投影，投影出的结果将只有选择的参数列，参数为：ADV\_DATE,SPEED。

排序：sortKey是将数据以选择参数的数据列进行排序，参数为ADV\_DATE；sortOption是以升序或者降序的排序规则进行排列，参数为：up(升序排列)；sortType是选择的排序列中的数据项的数据类型，只有数据类型匹配才能保证排序操作合法，参数为：int

**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看折线图和表格的数据统计是否符合期望值。

## 实验步骤

图示

描述已自动生成

## 实验结果

表格

描述已自动生成

表明列投影成功投影到了ADV\_DATE和SPEED 两列上。

表格

描述已自动生成

表明排序将投影的两列数据列以台风发生的时间从早到晚进行了升序排列。

图表, 折线图

描述已自动生成

表明了可以看出台风风速随时间增长曲折变化；风速最高值和最低值交替出现。

## 实验总结

本次实验让我对BDAP平台及其工作流的基本操作和相关组件有了深入的了解和认识。通过实际操作和实验结果分析，我不仅掌握了数据处理工作流的基本构建和运行方法，还学会了如何通过参数设置和可视化工具对数据处理结果进行精确分析和评估。

# 实验二（1） 数据预处理（1）

## 实验目的

熟悉常见的数据预处理操作

## 实验内容

**第一步：**配置总工作流。包含组件：加载数据，排序，折线图，表格，groupby（将数据根据（By）若干个字段划分（group）为不同的群体）

**第二步：**配置工作流组件。

加载数据：文件路径参数选择/Experiment\_2/air\_quality.csv；分隔符参数为‘,‘英文的逗号；是否带表头为true；文件格式为csv。

Grouby：聚焦列是选择数据中某一列为聚焦字段；colName是计算数据指定列的指定值；option是选择需要计算出来的结果；newColName是计算出结果后的新列名；第一个grouby的聚焦字段为hour,colName为AQI，option为avg（平均值），newColName为avg\_h。第二个grouby的聚焦字段为date,colName为AQI，option为avg（平均值），newColName为avg\_d。

排序：对应第一个grouby的排序sortKey参数为hour（按时间点顺序），sortOption参数为up,sortType参数为int。对应第二个grouby的排序sortKey参数为date（按日期顺序），sortOption参数为up,sortType参数为int。

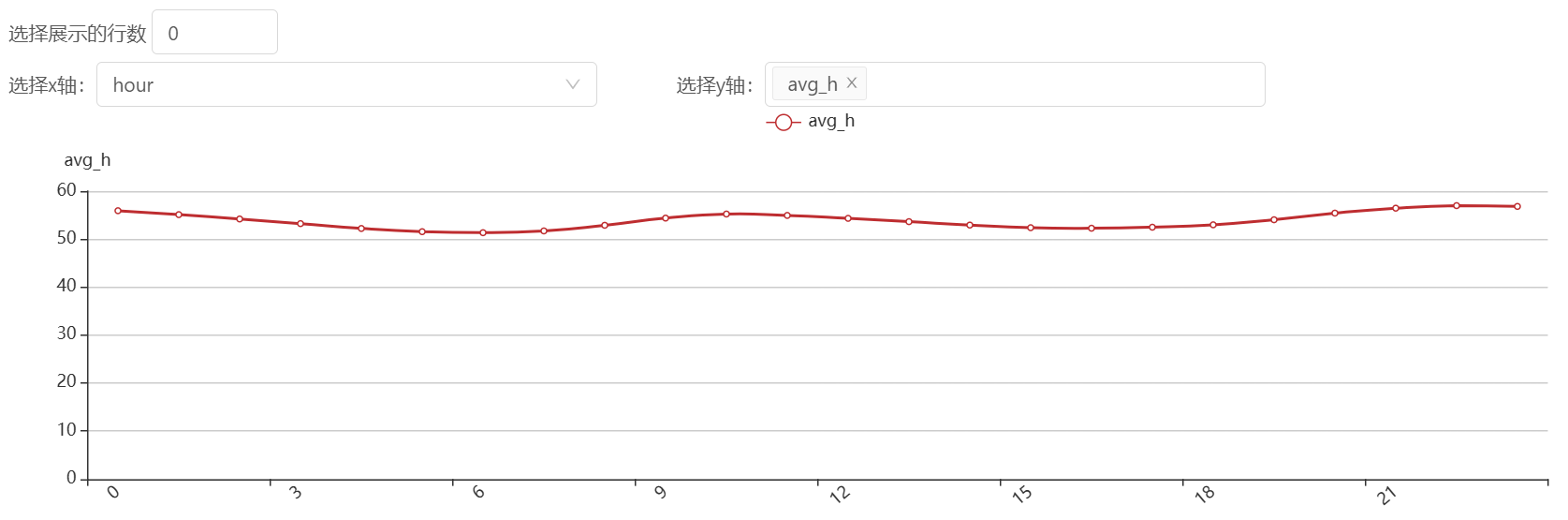
**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看折线图和表格的数据统计是否符合期望值。

## 实验步骤

图片包含 图示

描述已自动生成

## 实验结果

****

**图表

描述已自动生成**

通过对最终实验结果的分析，可以看出一天中每个时间点的AQI平均值都在50-60之间，变化不大；一年中每天的AQI平均值整体呈现出先减后增的趋势，其中中期变化较为平稳，前期和后期波动都比较大。

## 实验总结

通过groupby组件，我学会了如何根据指定字段将数据划分为不同的群体，并对每个群体进行特定的计算（如平均值）。这在实际应用中非常有用，可以帮助我们更好地理解数据的分布和特征。

在本次实验中，我分别对小时和日期进行了分组，并计算了AQI的平均值。这不仅让我掌握了groupby的基本用法，还加深了我对数据处理流程的理解

# 实验二（2） 数据预处理（2）

## 实验目的

熟悉常见的数据预处理操作。

## 实验内容

**第一步：**配置总工作流。包含组件：加载数据，排序，饼图（以饼图的形式可视化数据），表格，groupby（将数据根据（By）若干个字段划分（group）为不同的群体），计算生咸列（进行字符串操作）

**第二步：**配置工作流组件。

加载数据：文件路径参数选择/Experiment\_2/HZ\_game.csv；分隔符参数为‘,‘英文的逗号；是否带表头为true；文件格式为csv。

计算生成列：生成列名：对字符串操作结束后的结果形成新的列的列名； 选择函数：对字符串进行的操作；选择截取字段：指定一个数据列进行字符串操作；开始位置，结束位置：对字符串进行操作的下标。

Grouby：聚焦列是选择数据中某一列为聚焦字段；colName是计算数据指定列的指定值；option是选择需要计算出来的结果；newColName是计算出结果后的新列名；聚焦字段为famliyName, colName为famliyName，option为count（出现次数），newColName为famliycount。

排序： sortKey参数为familycount，sortOption参数为down（降序）,sortType参数为int。

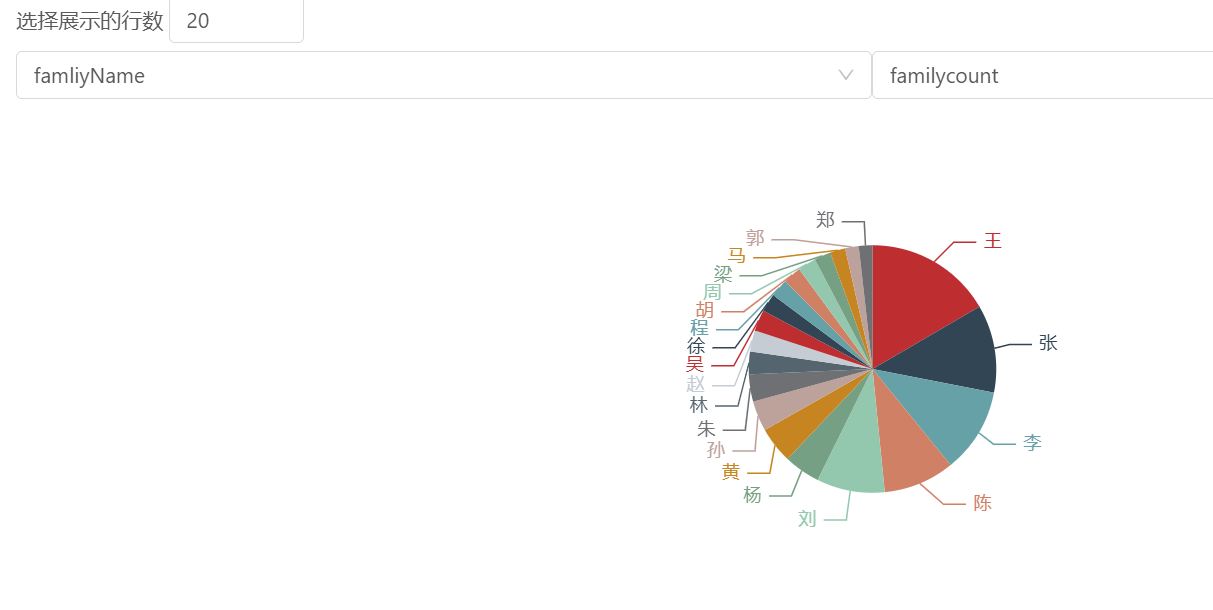
**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看饼图和表格的数据统计是否符合期望值。

## 实验步骤

图示

描述已自动生成

## 实验结果



由此可以看出各姓氏家庭的相对占比

## 实验总结

通过计算生成列组件，我掌握了如何在数据集中对字符串进行操作，并生成新的列。这包括选择操作函数、指定截取字段以及设置操作的下标等步骤。这为我处理包含文本数据的数据集提供了有力的工具。

# 实验三 决策树模型

## 实验目的

了解决策树的组成和构造算法，熟悉BDAP中机器学习相关的工作流和组件。

## 实验内容

**第一步：**配置总工作流。本实验总工作流分为两个部分，第一部分为决策树模型的训练流：包含加载数据，决策树（根据数据和参数进行决策树模型的训练）；第二部分为对决策树模型的结果进行预测，包含加载数据，已经训练的决策树模型，多酚类评估器（根据评价指标对预测结果进行评估），表格

**第二步：**配置工作流组件。

训练流加载数据：文件路径参数选择/Experiment\_3/iris.csv；分隔符参数为‘,‘英文的逗号；是否带表头为false；文件格式为csv。

决策树：模型路径：训练出的模型需要防止的地址，参数为自己建的私人文件夹；不纯度：用于信息增益计算的标准，参数为gini标准；每个节点的最大分支数，连续特征的最大划分数 范围 [分类数,100]，参数为100；最大深度，树的最大深度 非负，参数为10；最小信息增益 ，选择训练还是测试在树节点上考虑的拆分的最小信息增益，参数为0.1；节点的最少实例数，每个孩子在拆分后必须拥有的最小实例数，参数为1；随机数种子，参数为10；标签列，参数为\_c4；离散特征最大重复取值，对离散特征进行标号，提高决策树分类效果，参数为10;

模型名称为决策树三分类-实验3。

预测流加载数据：文件路径参数选择/Experiment\_3/iris.csv；分隔符参数为‘,‘英文的逗号；是否带表头为false；文件格式为csv。

多分类评估器：评价指标，选择不同的评价指标：F1:统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率,F1分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是1，最小值是0. 准确率(Accuracy)=正确分类的样本/总样本. WeightedPrecision:加权后的准确率，对于每一个标签的准确率加权. 真正类率(TruePositiveRateByLabel):分类器所识别出的正实例占所有正实例的比例. LogLoss:反映样本的平均偏差. HammingLoss:表示所有label中错误样本的比例，该值越小则模型的分类能力越强.，参数为accuracy;标签列为indexed\_c4,预测标签为prediction。

**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看表格的数据统计是否符合期望值。

## 实验步骤

图示, 箱线图

描述已自动生成

图形用户界面

描述已自动生成

## 实验结果

**表格

描述已自动生成**

图形用户界面, 应用程序

中度可信度描述已自动生成

决策树模型预测结果拟合率达到96%。

## 实验总结

深入理解决策树模型：组成与原理：我清晰地了解了决策树的组成部分，包括根节点、内部节点（判断条件）、叶节点（分类结果）等。通过配置决策树的各项参数（如不纯度、最大深度、最小信息增益等），我进一步理解了这些参数对决策树模型性能和结构的影响。构造算法：通过实际操作，我掌握了决策树模型的训练过程，包括数据的加载、特征的选择、节点的拆分以及模型的生成。这些步骤让我对决策树的构造算法有了更加直观的认识。

模型评估与调优：评价指标：我学会了使用不同的评价指标来评估模型的性能，如准确率、F1分数、真正类率等。这些指标有助于我全面了解模型的分类能力和效果。模型调优：通过调整决策树的参数，我观察了模型性能的变化，并学会了如何根据实验结果进行模型调优。这种调优过程对于提高模型的准确性和泛化能力至关重要。

# 实验四 K-Means聚类

## 实验目的

分析不同的新能源汽车座椅数量和价格的聚类特征，尝试对于组内的相似性与组间的差异性进行解释。

## 实验内容

**第一步：**配置总工作流。包含组件：加载数据，填充/删除空白行（填充/删除模式下，空白区域将以对应列属性值的平均值填充/将空白区域删除），K-均值（计算K-Means均值），列投影，表格，散点图

**第二步：**配置工作流组件。

加载数据：文件路径参数选择/Experiment\_4/ElectricCarData\_Norm.csv；分隔符参数为‘,‘英文的逗号；是否带表头为true；文件格式为csv。

填充/删除空白行：模式选择填充模式。

K-均值：聚类簇数，聚类数量，参数为5；最大迭代数，参数为50；收敛阈值，迭代停止的阈值，参数为0.001；标签列为Model；训练字段为Seats，PriceEuro。

列投影：投影字段为Model，cluser，PriceEruo，Seats。

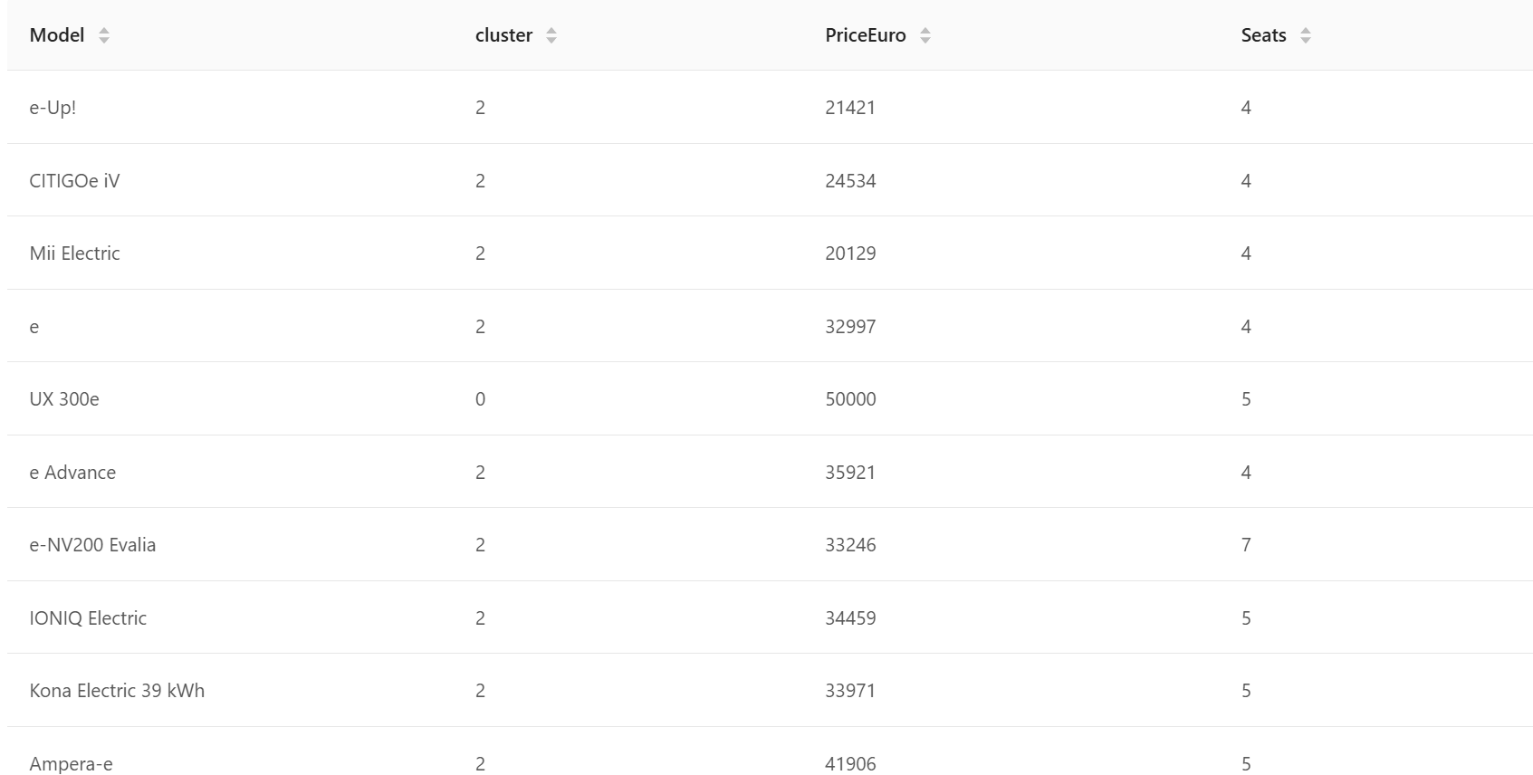
**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看散点图和表格的数据统计是否符合期望值。

## 实验步骤

图示, 工程绘图

描述已自动生成

## 实验结果



图表, 散点图

描述已自动生成

## 实验总结

K-Means聚类算法的应用：

在配置K-均值组件时，我深入了解了K-Means聚类算法的基本参数设置，包括聚类簇数、最大迭代数和收敛阈值。通过将这些参数设置为合理的值（如聚类簇数为5），我能够观察到数据被有效地分成了五个不同的组。

通过指定训练字段（Seats和PriceEuro），我学会了如何利用K-Means算法根据特定的特征对数据进行聚类。

聚类结果的分析：

在运行工作流后，我通过查看散点图和表格数据，对聚类结果进行了详细的分析。散点图直观地展示了不同聚类簇之间的分布和关系，而表格数据则提供了每个聚类簇中模型的详细信息。

通过分析，我尝试解释了组内相似性和组间差异性的原因。例如，某些聚类簇可能包含座椅数量较多且价格较高的高端新能源汽车，而另一些聚类簇则可能包含座椅数量较少且价格较低的经济型新能源汽车。

# 实验五 关联规则

## 实验目的

第一步：配置总工作流。包含组件：加载数据，表格，Apriori（运用Apriori算法对数据进行剪枝）

第二步：配置工作流组件。

加载数据：文件路径参数选择/Experiment\_5/casts.csv；分隔符参数为‘\t‘英文的逗号；是否带表头为true；文件格式为csv。

Apriori：数据列，算法输入的数据列名，参数为people;分隔符，每一项内的分隔符，参数为‘,‘；最小支持度，参数为0.01；最小置信度，参数为0.3

第三步：运行工作流，实验结果分析。工作流执行完毕后，查看表格的数据统计是否符合期望值。

## 实验内容

图形用户界面

低可信度描述已自动生成

## 实验步骤

（这一步需要同学们上传完整的代码截图，包含开头的导入数据及最后的结果打印代码）

## 实验结果

## 图形用户界面, 文本, 应用程序, 电子邮件 描述已自动生成

表格

描述已自动生成

## 实验总结

掌握了关联规则挖掘的基本流程和Apriori算法的工作原理。

学会了如何正确配置和运行数据分析工作流。

加深了对数据预处理和算法参数设置的理解。

提高了数据分析和解读能力。

获得了实际挖掘关联规则的经验和技巧

# 实验六 SVM拓展实验

## 实验目的

了解SVM算法，拉格朗日乘子法

## 实验内容

**第一步：**配置总工作流。本实验总工作流分为两个部分，第一部分为SVM模型的训练流：包含加载数据，多分类SVM（根据数据和参数进行SVM模型的训练）；第二部分为对SVM模型的结果进行预测，包含加载数据，已经训练的SVM模型，多分类评估器（根据评价指标对预测结果进行评估），表格

**第二步：**配置工作流组件。

训练流加载数据：文件路径参数选择/cad\_train.csv；分隔符参数为‘,‘英文的逗号；是否带表头为true；文件格式为csv。

多分类SVM：模型路径：训练出的模型需要防止的地址，参数为自己建的私人文件夹；模型名称为svm6；最大迭代次数为30；正则项参数，正则项在SVM的优化问题中充当了惩罚项，有助于解决分类问题并找到合适的超平面，控制模型的复杂性和训练误差之间的权衡，参数为2；标签列为Cath；训练字段为除Cath的剩下数据列。

预测流加载数据：文件路径参数选择/cad\_test.csv；分隔符参数为‘,‘英文的逗号；是否带表头为true；文件格式为csv。

多分类评估器：评价指标，选择不同的评价指标：F1:统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率,F1分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是1，最小值是0. 准确率(Accuracy)=正确分类的样本/总样本. WeightedPrecision:加权后的准确率，对于每一个标签的准确率加权. 真正类率(TruePositiveRateByLabel):分类器所识别出的正实例占所有正实例的比例. LogLoss:反映样本的平均偏差. HammingLoss:表示所有label中错误样本的比例，该值越小则模型的分类能力越强.，参数为accuracy;标签列为Cath\_indexed,预测标签为prediction。

**第三步：**运行工作流，实验结果分析。工作流执行完毕后，查看表格的数据统计是否符合期望值。

## 实验步骤

图形用户界面

描述已自动生成

## 实验结果

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

预测准确率达到%72.5

## 实验总结

在配置多分类SVM组件时，我深入了解了SVM算法的基本原理和参数设置。特别是正则项参数的设置，让我意识到在模型训练中，正则化对于防止过拟合和提高模型泛化能力的重要性。通过调整正则项参数，我学会了如何平衡模型的复杂性和训练误差，从而得到更好的模型性能。

# 实验七

（没有做选做实验的同学不用撰写实验七的报告）

## 实验目的

（同其他实验）

## 实验内容

（这个部分需要大家讲述代码编写的思路步骤，其中需要包含必要的变量名称及函数名称。如第一步：从xxx库导入数据；第二步：调用xxx库函数xxx分割数据集为xxx，第三步：…）

## 实验步骤

（这一步需要同学们上传完整的代码截图，包含开头的导入数据及最后的结果打印代码）

## 实验结果

（这一步打印混淆矩阵并结合实验内容作分析）

## 实验总结

（同其他实验）