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** | **吴起凡** |
| **知识模块** | **大数据** | **完成时间** | | **2023年10月20日** |
| **班级** | **序号** | 学号 | **姓名** | **成绩** |
| **2023211307** | **31** | **2023211088** | **罗浩** | **100** |

报告内容包括：

实验目的

实验内容

实验步骤

实验结果

实验总结

实验一

实验目的：

由于最近内陆受到台风影响非常大，为了有效减轻台风带来的负面影响，本实验通过分析2008到2018年的台风数据集，统计全球台风风速随时间的变化趋势

实验内容于步骤：

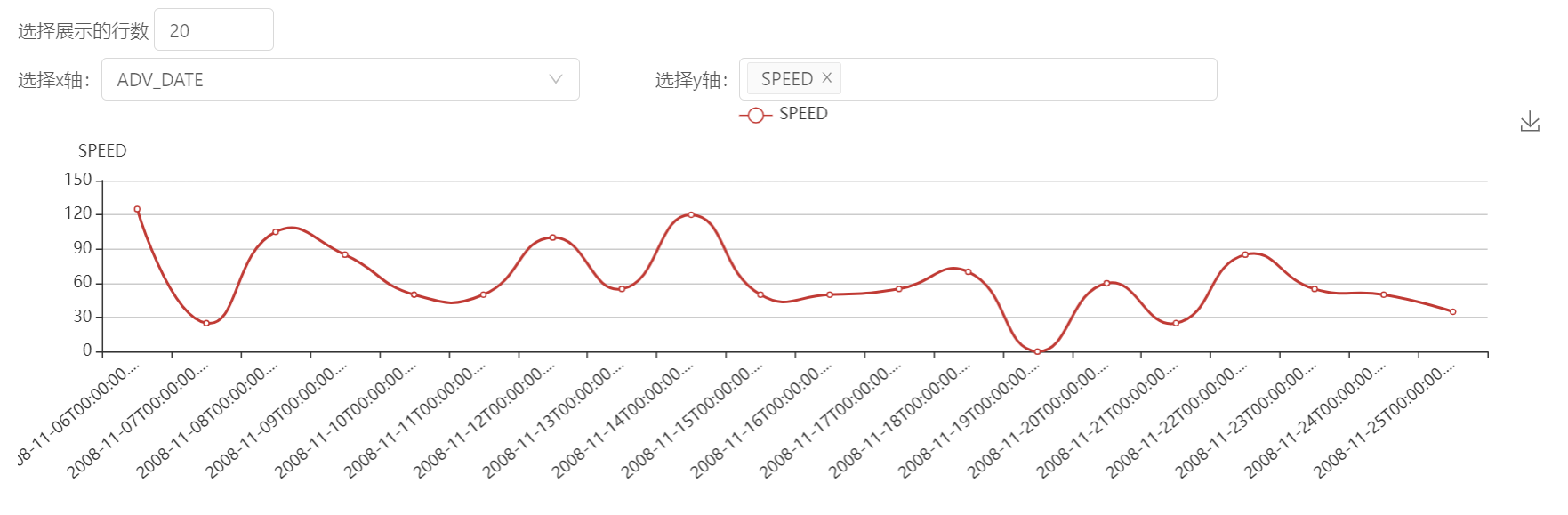
首先加载数据，选择ADV\_DATE和SPEED列进行投影，获取时间与风速的对应关系；

根据ADV\_DATE对源数据进行排序；

根据排序后的数据绘制折线图，分析随着时间增长全球台风的风速变化趋势

实验结果与总结：

通过对最终实验结果的分析，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



实验二

实验目的：

熟悉常见的数据预处理操作

实验步骤与内容：

以“hour”为聚集字段，使用 groupBy 操作进行聚集，并计算同一个时间点指标AQI的平均值“avg\_h”；

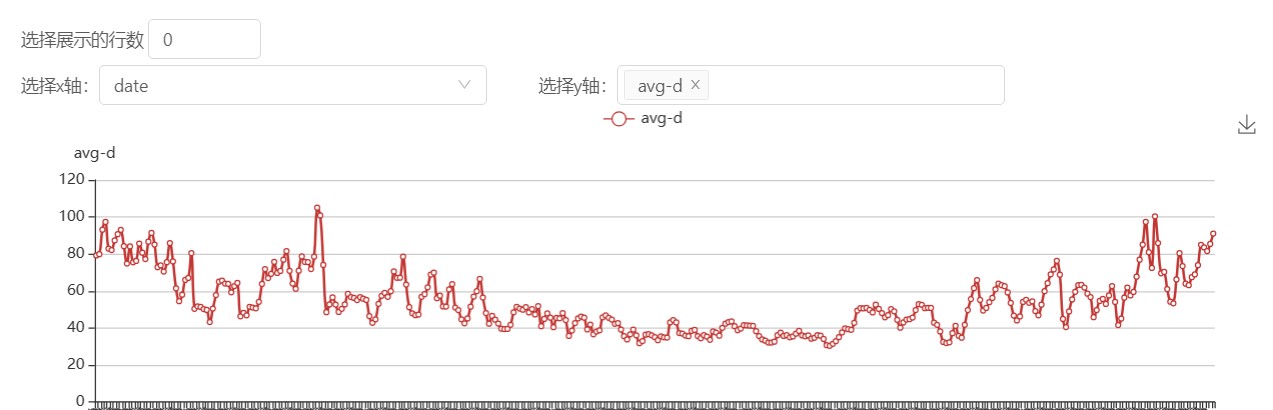
以“date”为聚集字段，使用 groupBy 操作进行聚集，并计算每天指标AQI的平均值“avg\_d”；

按照时间点的顺序(hour)，使用 sort 排序操作对整体数据进行排序；

按照日期的顺序(date)，使用 sort 排序操作对整体数据进行排序。

实验结果与总结：

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。



实验三

实验目的：

熟悉常见的数据预处理操作

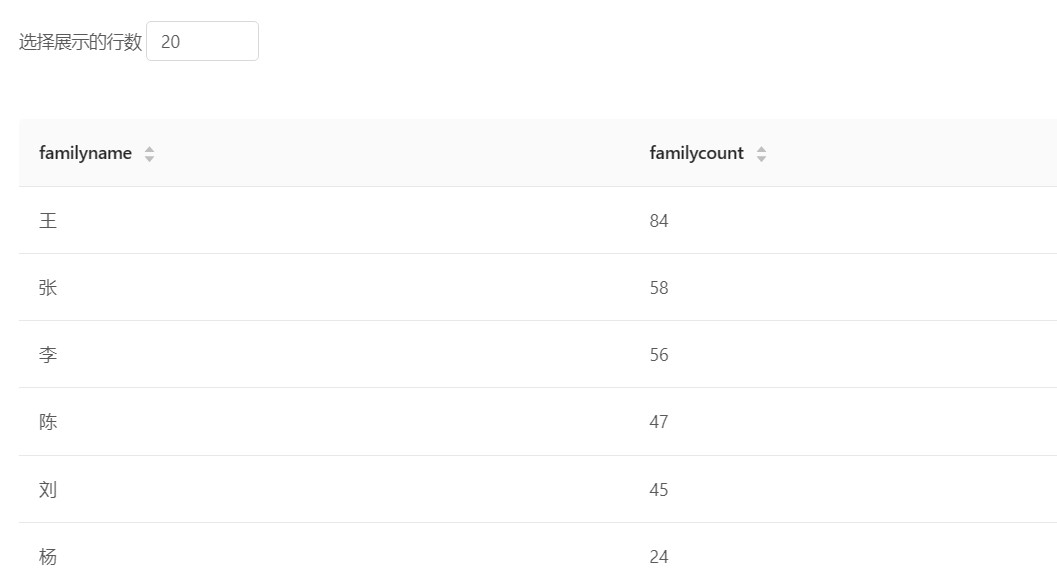
实验步骤与内容：

对字符串进行切割，获取“Name”的第一个字符，并命名为“familyName”。

以“familyName”为聚集字段，使用 groupBy 操作进行聚集，并统计每一个姓出现的次数“familyCount ；

按照出现次数“familyCount ，使用 sort 排序操作对整体数据进行排序。

实验结果与总结：



实验四

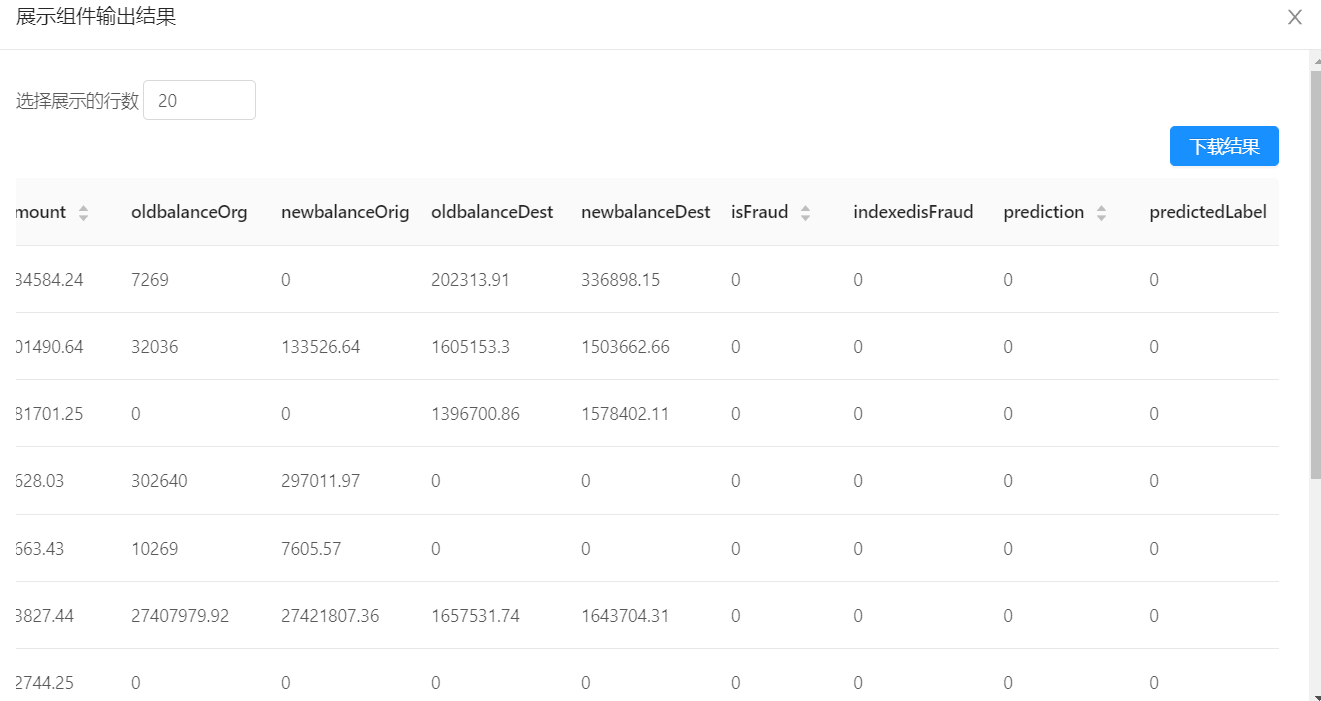
实验目的：

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。

实验步骤与内容：

训练出来的决策树模型，在预测集上的评估结果

实验结果与总结：



实验五

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。

实验步骤与内容：

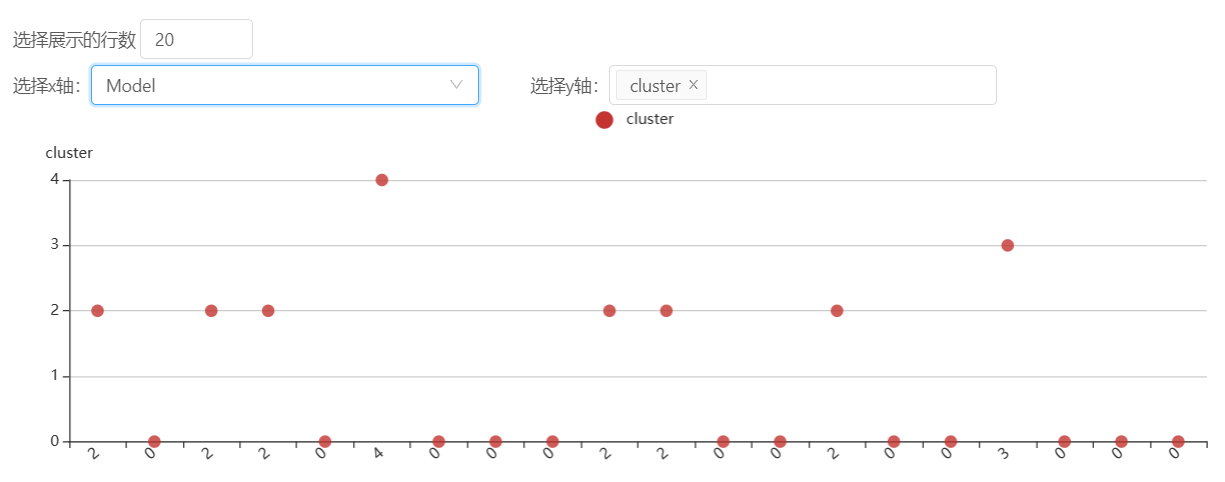
加载阿里云天池数据

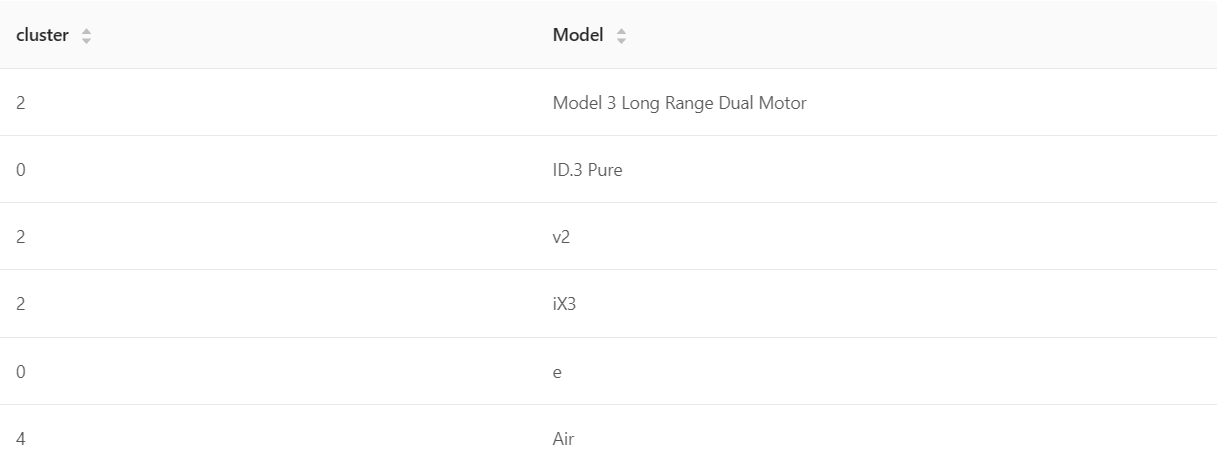
数据预处理

K-means聚类

可视化分析

实验结果与总给：





实验6

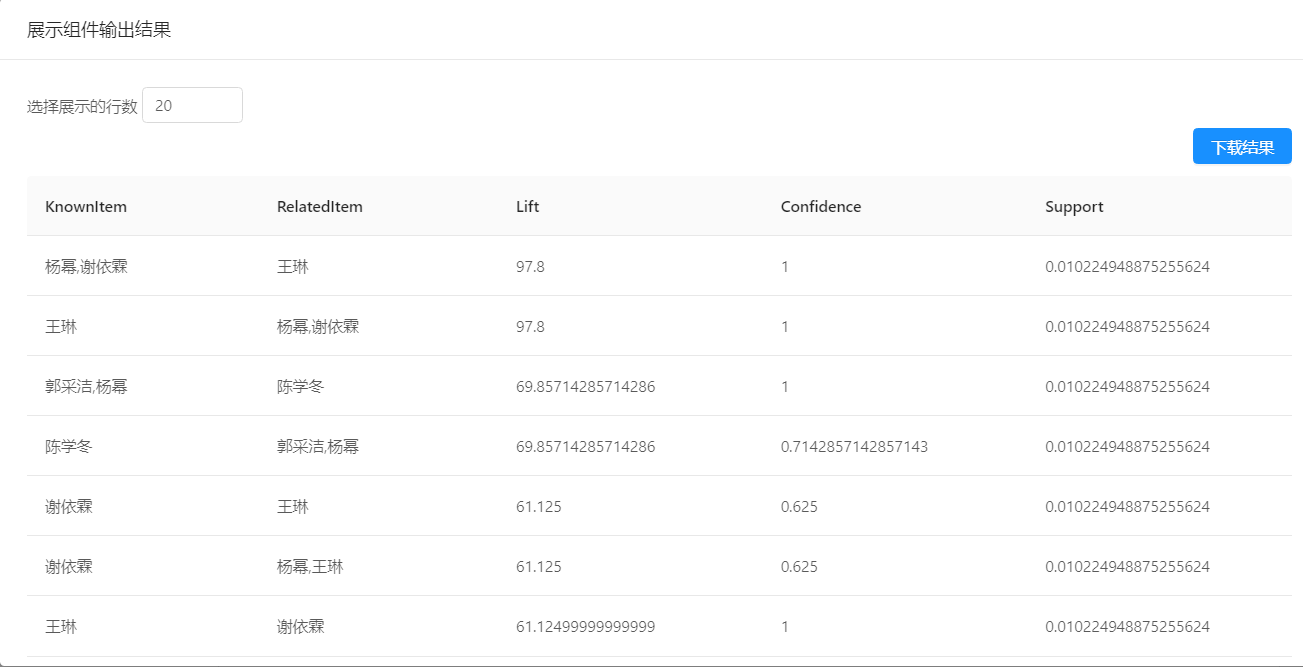
实验目的：

挖掘2015年以来最受欢迎电影中，演员之间的合作关系

实验步骤与内容：

利用Apriori组件进行数据分析

实验结果与总结：



实验7

实验目的：

SVM算法

实验步骤与内容：

下载数据集，转换数据集格式，数据集划分，模型训练，模型预测

实验结果与总结：

0.7075471698113207