|  |  |  |
| --- | --- | --- |
| 选题 | **2024年第十四届APMCM**  **亚太地区大学生数学建模竞赛（中文赛项）** | 参赛编号 |
|  | apmcm\*\*\*\*\*\* |

标题（此处换成论文的标题）

摘要

针对问

1. 问题重述

1.1问题背景

ChatGPT吧！

1.2问题要求

附件1给出了各地季风强度、地形排水、河流管理等基本信息，以及该地发生洪水的概率。为了能更好地根据已知信息预测洪水的发生概率，最小化灾害损失，现需结合实际情况与所给信息建立数学模型，分析以下问题。

问题1：依据附件1中给出的基本信息，可视化处理相关数据，并单独分析洪水发生概率与每个基本信息的相关性，判断其关联程度大小，给出相关结论和建议。

问题2：

问题3：

问题4：

1. 问题分析

2.1问题1的分析

针对问题1，首先对train.csv中所给数据进行最大值、最小值、平均值、中位数等基本统计量计算、数据预处理等基础性工作,并记录洪水概率的最大、最小值，基于这些数据，采用概率密度直方图、核密度估计（Kernel Density Estimation, KDE）图可视化洪水概率的分布情况，同时用频数统计直方图可视化train.csv与test.csv中季风强度、地形排水等各个指标特征的分布情况，并计算上述所有指标分布的偏度与峰度系数用以量化各项评估指标与洪水概率的分布情况。随后，引入Spearman相关系数进相关性分析，并加以假设检验，用以量化各项指标与洪水概率之间的关系。最后，综合以上获得的相关性、数据分布情况等分析结果进行统计分析，指出对洪水概率影响较大的指标，并据此提出相应的预防措施。

2.2问题2的分析

针对问题2，首先标准化train.csv中洪水概率数据，并利用K-Means++聚类算法将其聚类为三簇，随后对聚类结果进行箱型图、频数分布直方图可视化，基于簇进行数据表分组，并计算每个分组的最大、最小值，用以评估聚类结果是是否存在交错，最后划分风险为高、中、低三类，将其作为分类标签，并对其分别进行统计分析可视化，用以分析不同风险对应指标特征。风险预警评价问题实质是基于连续数值到逻辑类别区间映射的回归+分类的合并问题，首先基于先前得出的三类分组数据中最大、最小值构造连续回归数值转离散类别数值的映射，并训练逻辑回归模型建立符合分布的洪水概率到风险等级的连续映射，实现有空缺映射区间的填充工作，以保证最终回归+分类模型评价能力的鲁棒性，随后，基于所有指标构建线性回归模型作为基准模型，实现较优结果，探索指标特征与洪水概率间潜在的线性相关性，并对所有指标之和与洪水概率进行可视化以确认其线性相关性，因此引入指标和作为关键指标，然后将季风强度、地形排水等指标在样本维度上从小到大排序，构造排序指标，从而消除指标维度上的数据差异性，最后，将排序指标与指标和作为最终指标训练CatBoost模型实现了优于基准线性回归模型的效果，并且其在三分类任务上同样表现优秀，同时，利用K折交叉验证法验证了模型的低灵敏度与高泛化性。

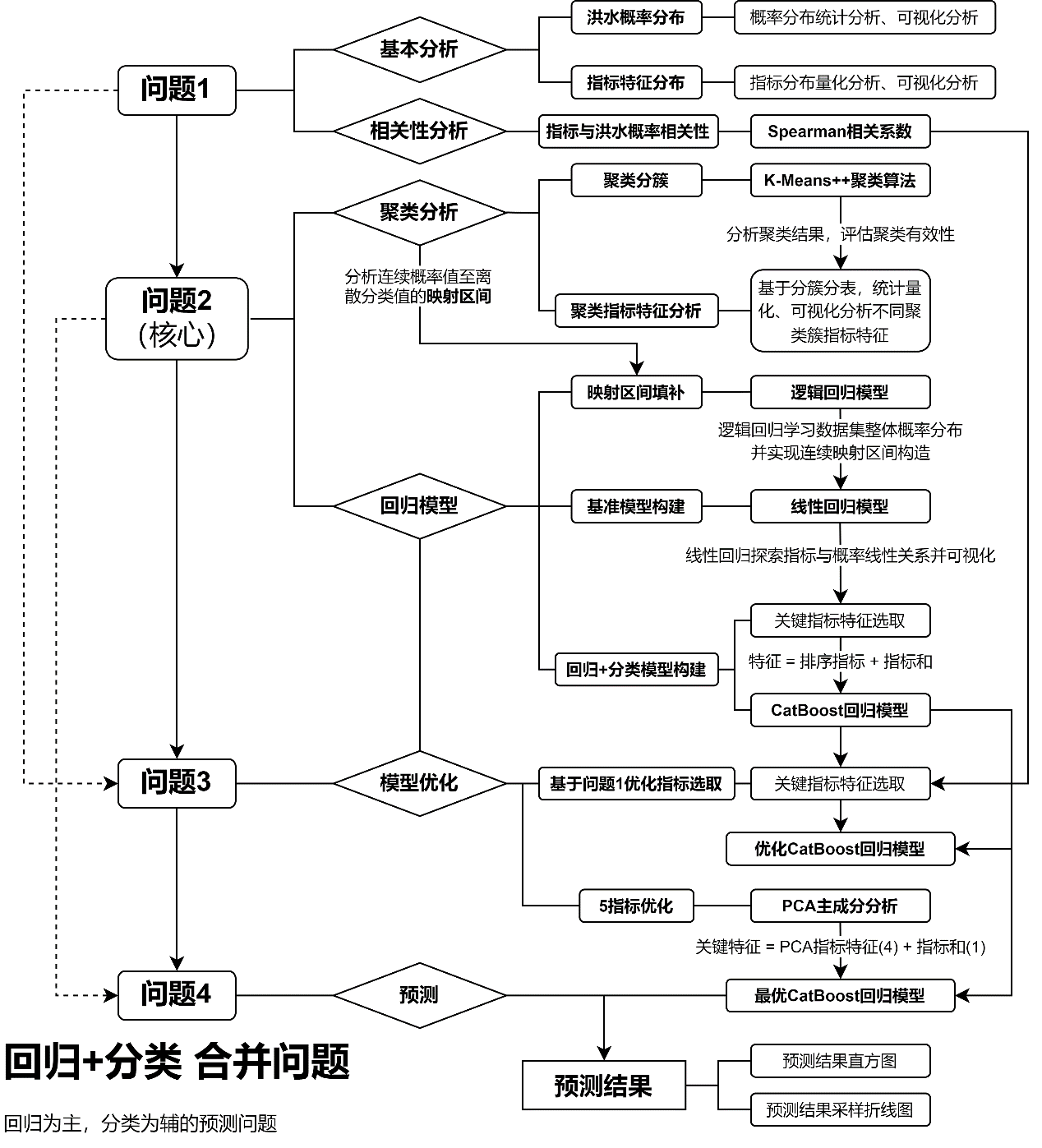
2.3问题3的分析

针对问题3，首先基于问题1分析出的各指标与洪水概率的Spearman相关系数选出相关性较高的指标特征作为基本指标，并加入指标和作为关键指标，训练CatBoost回归模型，得出较优回归预测模型。随后，使用主成分分析（Principle Component Analysis，PCA）进行数据降维，将20维指标特征降至4维作为关键指标，同时依旧加入指标和组成5个指标，用以训练CatBoost回归模型，得出最优回归预测模型。

2.4问题4的分析

针对问题4，直接选用问题3中得出的最优CatBoost回归模型进行test.csv的特征工程，才用最优5个关键指标进行结果预测，于submit.csv中填写预测结果，并对预测结果使用直方图、折线图（随机采样1000样本）进行可视化分析，计算其偏度峰度系数，分析其分布正态性，并与其他两个性能较低的CatBoost回归模型进行分布对比，验证其预测结果正态性与合理性。

**文章总体思路如图1所示：**



**图1总体思路图**

1. 模型假设

四、符号说明

五、数据侧写

1. 问题1模型的建立与求解

6.1基于数据的数字特征进行描述

题目要求对数据进行可视化处理，为了更好地描述洪水发生概率分布、基本信息特征的分布规律，本文引入峰度系数、偏度系数来描述统计数据。

1. 峰度系数

峰度系数（Kurtosis）用于衡量数据分布的峰度程度。例如时，数据的分布为正态分布；时，数据分布的峰度较小，数据更分散；时，数据分布的峰度较大，数据更集中。并且越大，数据分布的图像更加“陡峭”。峰度系数的计算公式如下：



1. 偏度系数

偏度系数（Skewness）用于衡量数据分布的偏斜程度。例如时，数据的分布为正态分布；时，数据呈左偏分布；时，数据呈右偏分布。并且越大，数据分布的偏斜越明显。峰度系数的计算公式如下：



6.2数据可视化

利用python作出洪水发生概率密度图像、基本信息的分布柱状图如下：

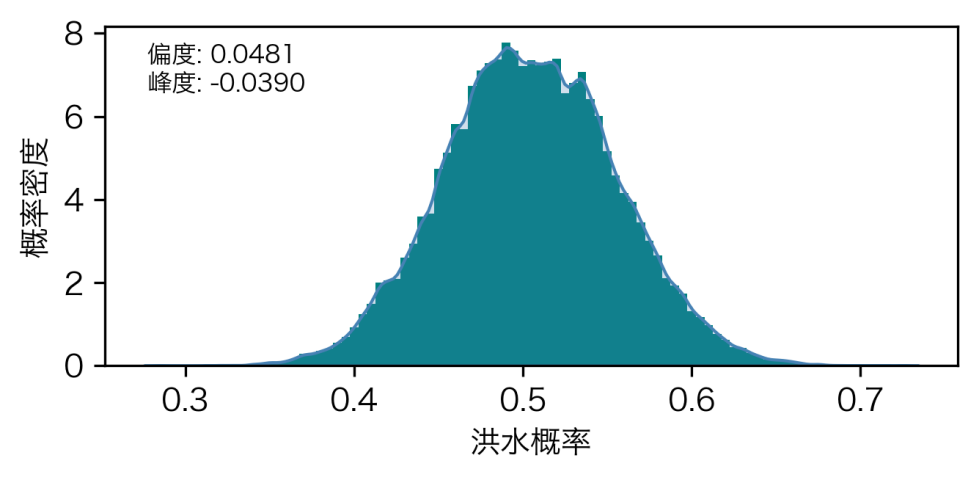
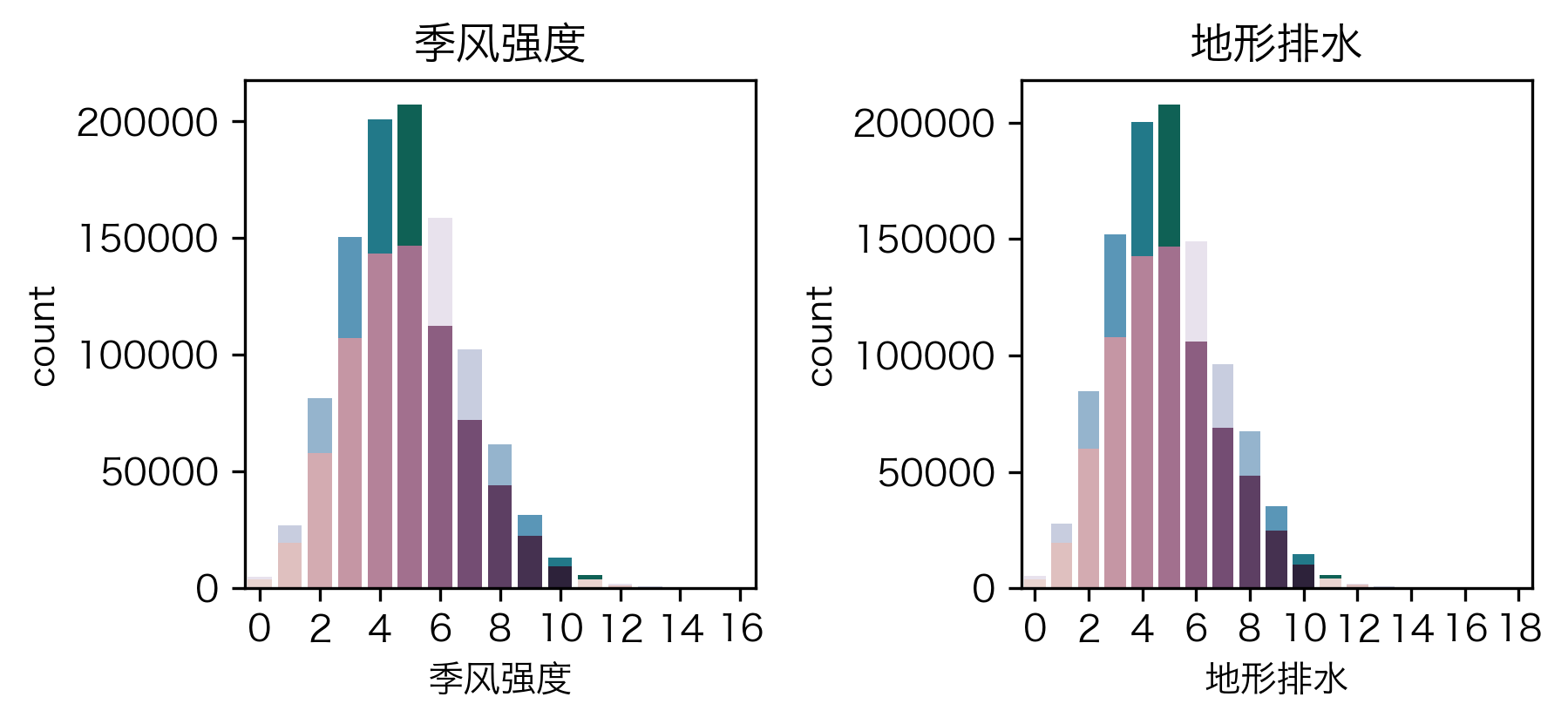


图2洪水概率密度的图像





七、问题2模型的建立与求解

八、问题3模型的建立与求解

九、问题4模型的建立与求解

十、模型的评价与推广

十一、参考文献

十二、附录

**参考文献 （可另起一页）**

参考文献的编号，如[1][3]等；引用书籍还必须指出页码。参考文献按正文中的引用次序列出，其中：**书籍的表述方式为**

[编号] 作者，书名，出版地：出版社，出版年。

**参考文献中期刊杂志论文的表述方式为**

[编号] 作者，论文名，杂志名，卷期号：起止页码，出版年。

**参考文献中网上资源的表述方式为**

[编号] 作者，资源标题，网址，访问时间（年月日）。

**附录（另起一页）**