|  |  |  |
| --- | --- | --- |
| 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 |
| B | apmcm24102596 |

标题（此处换成论文的标题）

摘要

(说明：以下开始写摘要，正文从下一页开始。摘要及正文格式基本要求是宋体，小四号，单倍行距，没有要求的地方就自行处理。看完后删除该说明)

目录

[1 引言 3](#_Toc171198412)

[1.1 背景 3](#_Toc171198413)

[1.2 问题重述 3](#_Toc171198414)

[1.3 工作概述 3](#_Toc171198415)

[1.3.1 问题一分析 3](#_Toc171198416)

[1.3.2 问题二分析 3](#_Toc171198417)

[1.3.3 问题三分析 3](#_Toc171198418)

[1.3.4 问题四分析 3](#_Toc171198419)

[2 一般假设和变量描述 3](#_Toc171198420)

[2.1 假设 3](#_Toc171198421)

[2.2 变量描述 3](#_Toc171198422)

[3 模型的建立与解决方案 4](#_Toc171198423)

[3.1 任务一 4](#_Toc171198424)

[3.1.1 数据预处理 4](#_Toc171198425)

[3.1.2 检验是否满足正态分布 4](#_Toc171198426)

[3.1.3 Spearman相关性分析 4](#_Toc171198427)

[3.2 任务二 4](#_Toc171198428)

[3.3 任务三 4](#_Toc171198429)

[3.4 任务四 4](#_Toc171198430)

[4 模型的评估 4](#_Toc171198431)

[4.1 优点 4](#_Toc171198432)

[4.2 缺点 4](#_Toc171198433)

正文内容(原则上不能超过20页)

# 引言

## 背景

洪水通常由暴雨、快速融雪、风暴潮等自然事件引发，导致河流、湖泊的水位急剧上升，是一种常见的自然灾害。洪水也被称为大水，指水体水位超标并威胁相关地区安全，有时甚至会导致灾害发生。随着人口增长，以及为了拓展农田和城市用地，人们对自然环境的破坏（如过度耕作、围湖造田、非法砍伐森林等）改变了地表条件和流水路线，加剧了洪水的严重性。在降水量大的年份，洪水灾害的发生和严重程度往往与人类活动密切相关，其中长期的森林破坏尤为关键。全球气候变化和城市化加速也增加了极端水文事件的频率，使得雨季洪水灾害更加频繁，严重威胁到人民的生命和财产安全[1]。

## 问题重述

1. 首先加载和初步处理train.csv文件，包括数据清洗和处理缺失值。使用皮尔逊相关系数等统计方法来分析20个指标与洪水发生概率之间的相关性，并可视化。根据指标的相关性强度，提出预防洪水的措施和建议。
2. 通过聚类分析，根据洪水发生的概率将数据分为高、中、低三个风险类别。选择合适的指标，并使用决策树、随机森林等方法分析指标重要性并计算权重，从而构建一个洪水风险预警评估模型。最后进行模型的灵敏度分析，评估不同指标权重变化对模型预测结果的影响。
3. 基于问题1的分析结果，选择合适的机器学习模型进行训练。考虑使用特征选择技术来优化模型，尝试只用5个关键指标来调整模型。使用交叉验证等技术验证模型的准确性。
4. 使用问题3中构建的模型，对test.csv中的数据进行洪水概率预测，并将预测结果填充到submit.csv文件中。绘制洪水概率的直方图和折线图，分析结果的分布特征是否正态分布。

## 工作概述

### 问题一分析

本问题主要目的是分析并可视化20个指标中哪些指标与洪水发生概率有密切的相关性。我们的方法首先使用箱形图对原数据进行预处理以排除异常值，接着分析预处理后的数据不满足正态分布，需要使用地理探测器对预处理后的数据进行相关性分析，得到相关性较高和相关性较低的变量，从而针对洪水进行提前预防。

### 问题二分析

本问题应该将train.csv中的洪水数据根据**洪水概率**进行**聚类**，人为地设定聚类的类别数，由每个聚类簇中簇中心的洪水概率值大小，将洪水数据划分为高中低风险三个等级。我们可以统计不同等级数据中**影响洪水因素的均值**，来分析不同等级洪水事件的指标特征。随后，针对指标的选取，考虑到选取的指标需要对洪水事件等级的不同存在一定敏感性，选取这些指标才能更加容易对不同事件进行区分，因而我们使用了**单因素方差分析**的方法，得到**所有的指标对不同事件的区分都是有益**的这一结论。根据这些指标，我们利用**熵权法**得到不同指标的权重，并在熵权法的基础上**结合层次分析法**形成一个**结合熵权法和层次分析法的洪水风险预警评价模型**。以聚类后的风险等级为依据，将评价模型得到的风险等级与之对比，从而评估评价模型的灵敏度。

### 问题三分析

### 问题四分析

# 一般假设和变量描述

## 假设

## 变量描述

# 模型的建立与解决方案

## 任务一

### 数据预处理

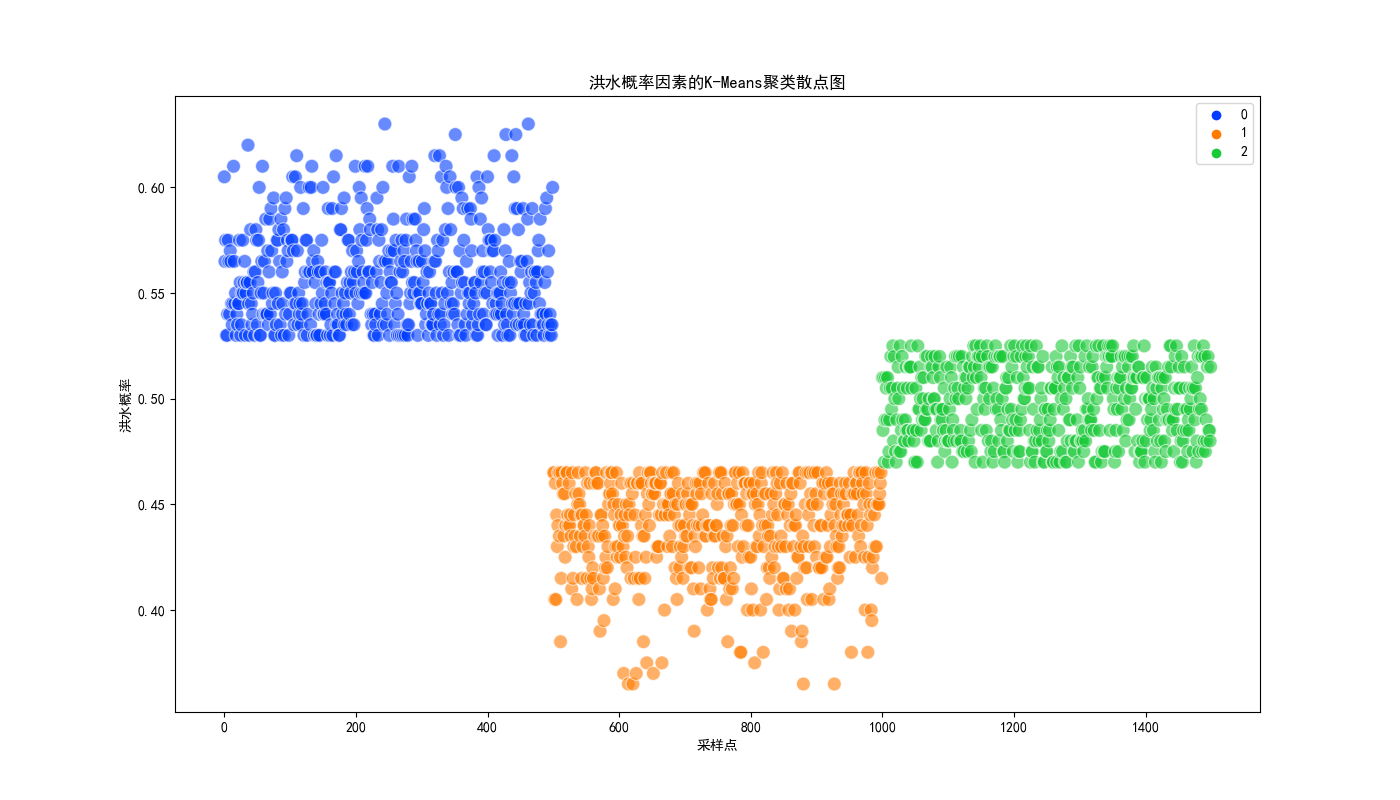
### 检验是否满足正态分布

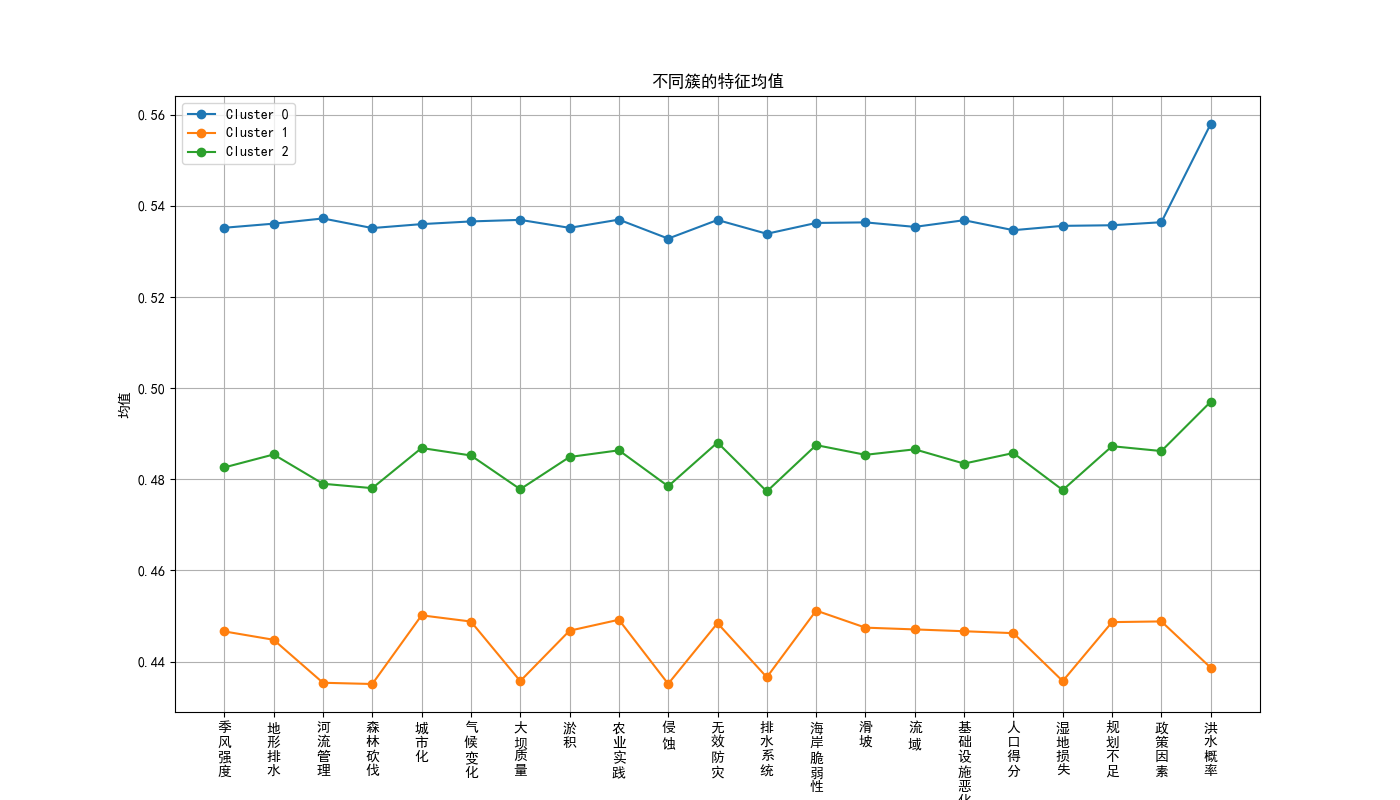
得到train.csv中的数据不符合正态分布，因此不能使用Pearson相关系数分析，数据不是二分类数据，因此不能使用Cochran's Q 检验。不是评级数据，所以不能使用Kappa一致性检验

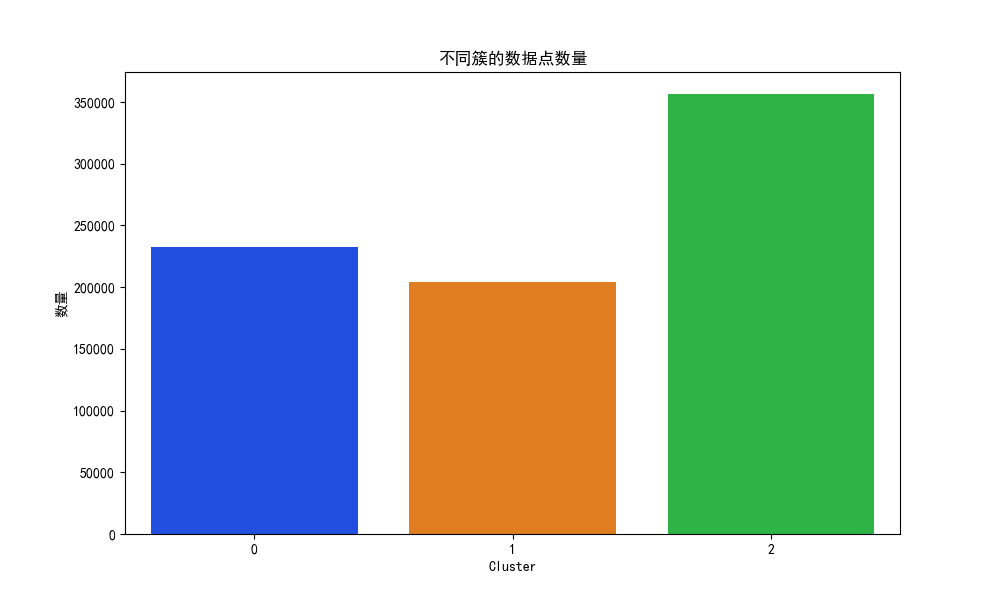
### Spearman相关性分析

## 任务二

### 针对洪水概率的聚类



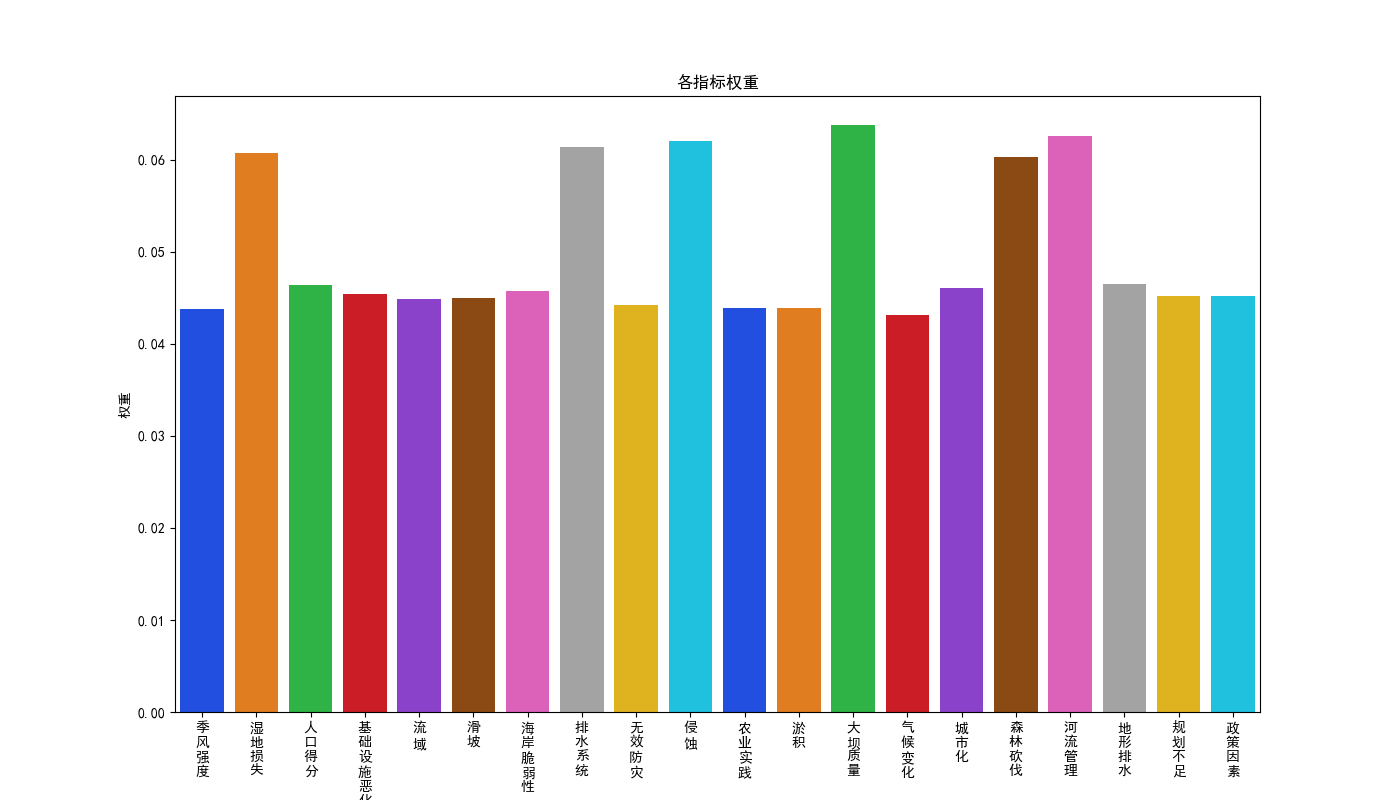


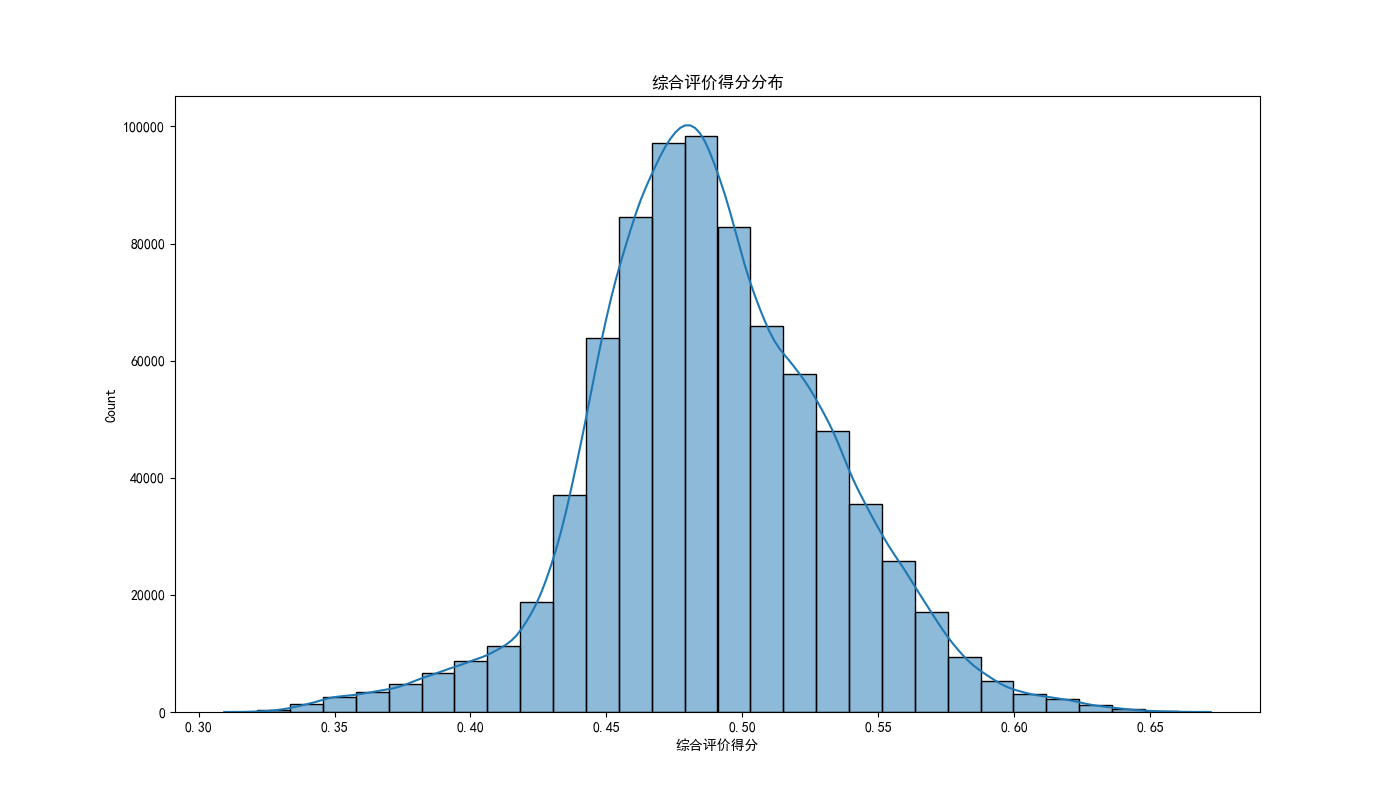


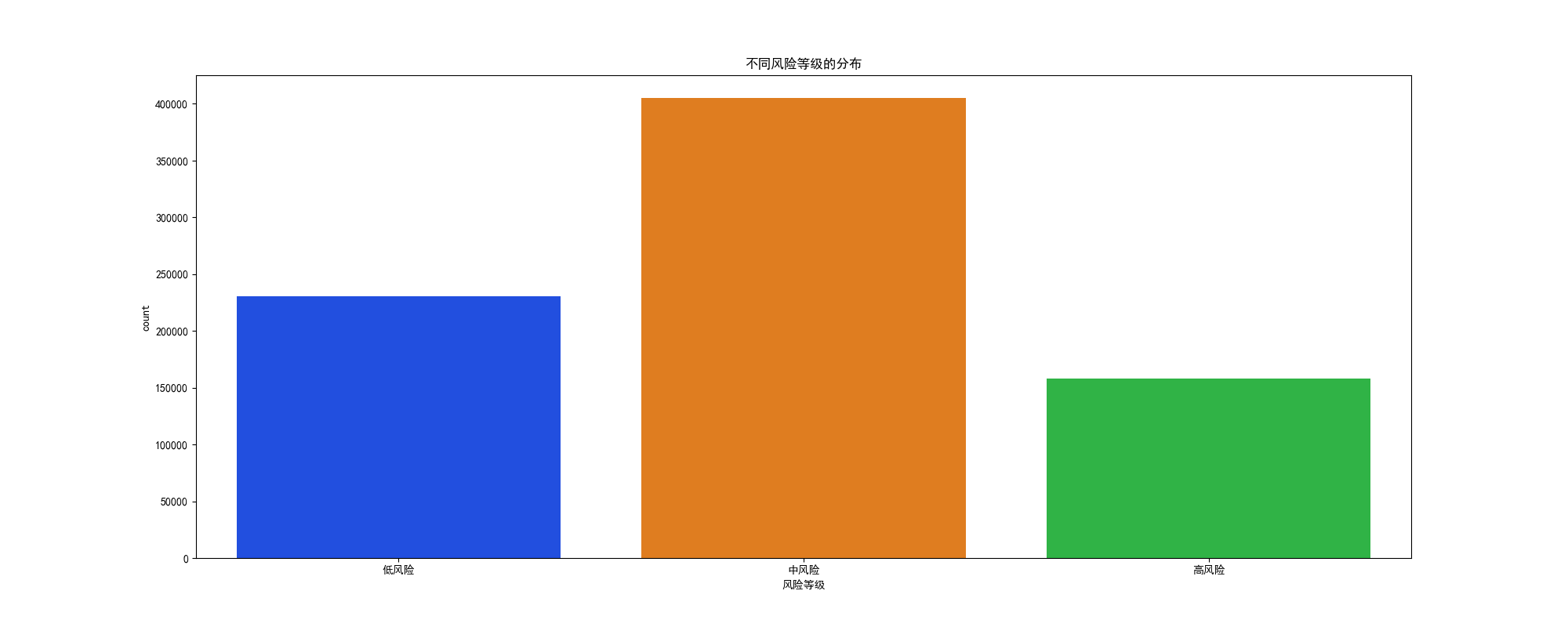
### 单因素方差分析

|  |  |
| --- | --- |
| 因素 | P值 |
| 季风强度 | 0.0 |
| 湿地损失 | 0.0 |
| 人口得分 | 0.0 |
| 基础设施恶化 | 0.0 |
| 流域 | 0.0 |
| 滑坡 | 0.0 |
| 海岸脆弱性 | 0.0 |
| 排水系统 | 0.0 |
| 无效防灾 | 0.0 |
| 侵蚀 | 0.0 |
| 农业实践 | 0.0 |
| 淤积 | 0.0 |
| 大坝质量 | 0.0 |
| 气候变化 | 0.0 |
| 城市化 | 0.0 |
| 森林砍伐 | 0.0 |
| 河流管理 | 0.0 |
| 地形排水 | 0.0 |
| 规划不足 | 0.0 |
| 政策因素 | 0.0 |

### AHP-TOPSIS综合评价模型





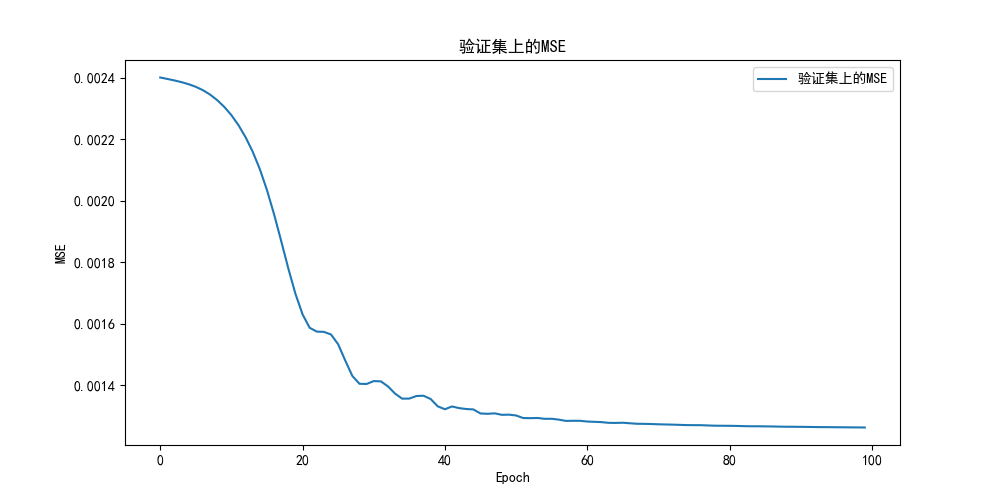
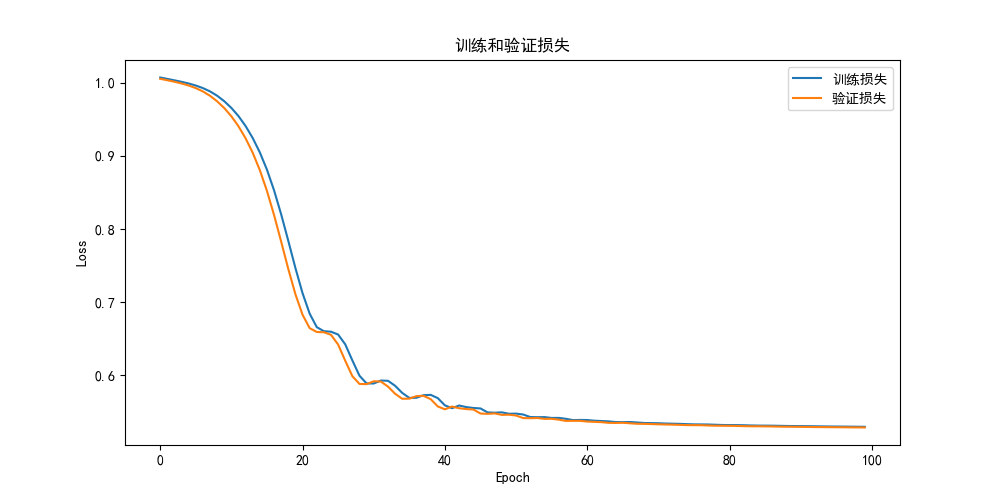


## 任务三

Validation R²: 0.4713341868839014

Validation Mean Squared Error: 0.001262481207959354

Validation Mean Absolute Error: 0.02876710332930088



## 任务四

# 模型的评估

## 优点

## 缺点

**参考文献**

1. 张建云,宋晓猛,王国庆等.变化环境下城市水文学的发展与挑战：I.城市水文效应[J].水科学进展,2014,25(4):594-605.
2. ZHANG Jianyun,SONG Xiaomeng,WANG Guoqing, et al.Development and challenges of urban hydrology in a changing environment:I: Hydrological response to urbanization[J]. Advances in Water Science, 2014,25(4):594-605.
3. AINUDDIN S, ROUTRAY J K, AINUDDIN S. People's risk perception in earthquake prone Quetta city of Baluchistan[J]. International Journal of Disaster Risk Reduction, 2014, 7: 165-175.
4. TABARI H. Climate change impact on flood and extreme precipitation increases with water availability[J]. Scientific Reports, 2020, 10(1): 13768.
5. **王劲峰,徐成东.地理探测器:原理与展望[J].地理学报,2017,72(01):116-134.**

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）**