B题 洪水灾害的数据分析与预测

题目就说了是分析和预测，真就只是分析和预测？吗？

第一段定义**洪水的危害**，是一种自然灾害。举出《尚书·尧典》的典故，和具体出现洪水的例子。

第二段分析洪水，洪水的频率和严重程度与**人口增长趋势**相当一致。

给出：人口增长->人为破坏增多（扩大耕地，围湖造田，乱砍滥伐）->改变汇流条件->加剧洪灾程度

- 原因：人为因素
- 影响：水土流失

第三、四段介绍三个数据集，简单来说，用test.csv训练，test.csv测试准确度，预测submit.csv的值。

附件 train.csv 中提供了超过 100 万的洪水数据，其中包含洪水事件的 id、季风强度、地形排水、河流管理、森林砍伐、城市化、气候变化、大坝质量、淤积、农业实践、侵蚀、无效防灾、排水系统、海岸脆弱性、滑坡、流域、基础设施恶化、人口得分、湿地损失、规划不足、政策因素和发生洪水的概率。

有洪水事件的 id和另外20个指标，其中，发生洪水的概率是**因变量**，其他因素均为**自变量候选变量。**

题目给的步骤非常明确，直接跟着题目思路走就行👇

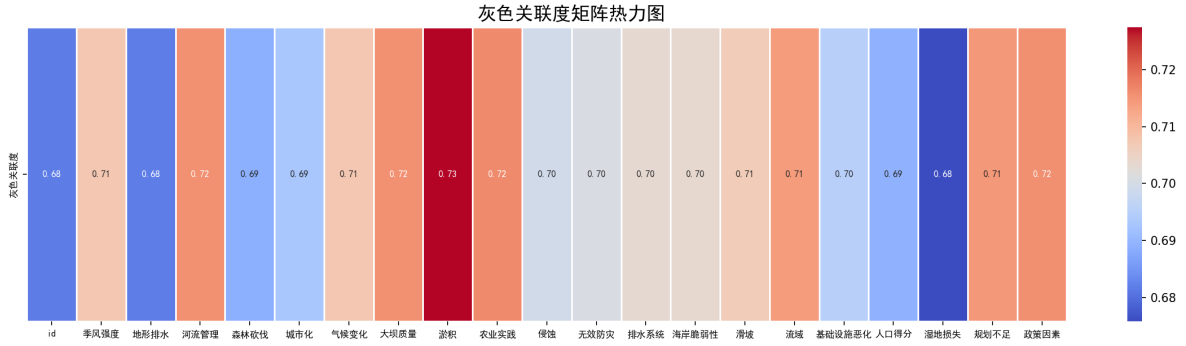
问题 1

请分析附件 train.csv 中的数据，分析并可视化上述 20 个指标中，哪些指标与洪水的发生有着密切的关联？哪些指标与洪水发生的相关性不大？并分析可能的原因，然后针对洪水的提前预防，提出你们合理的建议和措施。

洪水的发生就是发生洪水发生的概率，尝试了求皮尔森和斯皮尔曼相关系数之后，还是选择求**灰色关联度。**

灰色关联度系数如下：

id 0.681576 季风强度 0.707628 地形排水 0.681653 河流管理 0.715706 森林砍伐 0.688852 城市化 0.692787 气候变化 0.707646 大坝质量 0.715536 淤积 0.727431 农业实践 0.716449 侵蚀 0.699340 无效防灾 0.700754 排水系统 0.703328 海岸脆弱性 0.703549 滑坡 0.706461 流域 0.714254 基础设施恶化 0.695020 人口得分 0.689276 湿地损失 0.675788 规划不足 0.714684 政策因素 0.715782



根据提供的灰色关联度系数，可以将各指标与洪水发生的关联性进行分类分析，并针对洪水的提前预防提出合理的建议和措施。

得到第一问的结论：

高关联性指标（关联度 > 0.70）

1. 河流管理 (0.715706)
2. 大坝质量 (0.715536)
3. 淤积 (0.727431)
4. 农业实践 (0.716449)
5. 气候变化 (0.707646)
6. 季风强度 (0.707628)
7. 规划不足 (0.714684)
8. 政策因素 (0.715782)
9. 流域 (0.714254)

低关联性指标（关联度 < 0.70）

1. id (0.681576)
2. 地形排水 (0.681653)
3. 森林砍伐 (0.688852)
4. 城市化 (0.692787)
5. 侵蚀 (0.699340)
6. 无效防灾 (0.700754)
7. 排水系统 (0.703328)
8. 海岸脆弱性 (0.703549)
9. 滑坡 (0.706461)
10. 基础设施恶化 (0.695020)
11. 人口得分 (0.689276)
12. 湿地损失 (0.675788)

高关联性指标分析

1. **河流管理、大坝质量、淤积**：这些指标直接涉及河流系统和水利工程的管理和维护，若管理不善，会直接导致洪水风险增加。
2. **农业实践、气候变化、季风强度**：农业实践可能影响土壤吸水能力，气候变化和季风强度则直接影响降水量和水文条件，进而影响洪水发生概率。
3. **规划不足、政策因素、流域**：这些因素涉及政策和规划层面，合理的规划和有效的政策能显著减少洪水风险。

低关联性指标分析

1. **地形排水、森林砍伐、城市化**：虽然这些因素对洪水有一定影响，但在具体情境中其影响力相对较小，可能是因为其他高关联性因素的影响更为显著。
2. **基础设施恶化、人口得分、湿地损失**：这些指标虽然也与洪水有关，但可能其直接影响力较弱，或是其影响是间接的，通过其他高关联性指标起作用。

建议和措施

1. **加强河流和水利工程管理**：定期检查和维护河流管理系统和大坝质量，及时疏通河道，防止淤积。
2. **优化农业实践**：推广节水农业和保护性耕作技术，减少土壤侵蚀，提高土壤吸水能力。
3. **应对气候变化**：加强气象预报能力，优化洪水预警系统，及时发布天气预警信息。
4. **合理的城市规划和政策制定**：制定科学的洪水防治规划，加强土地利用和水资源管理，确保规划和政策的有效执行。
5. **提升防灾减灾能力**：加强防洪基础设施建设，提升社区和公众的防洪意识和应急响应能力。
6. **保护和恢复自然生态系统**：保护湿地和自然蓄水区，恢复被破坏的生态系统，提高自然环境的洪水调蓄能力。

问题 2

将附件 train.csv 中洪水发生的概率聚类成不同类别，分析具有高、 中、低风险的洪水事件的指标特征。然后，选取合适的指标，计算不同指标的权重，建立发生洪水不同风险的预警评价模型，最后进行模型的灵敏度分析。

把洪水概率分成三类 高、 中、低风险

--->使用**K-Means算法**对洪水发生概率进行聚类

模型结果如下：

聚类中心： [[0.49744545]
[0.56086881]
[0.43834819]]

--->使用**互信息MI**对洪水发生概率进行聚类

聚类效果评价指标：MI, NMI, AMI（互信息，标准化互信息，调整互信息） 聚类评价指标条件互信息-CSDN博客

计算指标权重的方法是使用互信息（mutual information）。互信息是一种衡量两个变量之间的相互依赖程度的统计量。在特征选择中，互信息用于衡量每个特征（指标）与目标变量（在这里是洪水风险类别）之间的相关性。

各指标权重：
{
'id': 0.00141182,
'季风强度': 0.04919908,
'地形排水': 0.05264376,
'河流管理': 0.05228368,
'森林砍伐': 0.04956926,
'城市化': 0.04838909,
'气候变化': 0.05115447,
'大坝质量': 0.05148097,
'淤积': 0.05252767,
'农业实践': 0.04766383,
'侵蚀': 0.04790486,
'无效防灾': 0.04947149,
'排水系统': 0.05050857,
'海岸脆弱性': 0.0479157,
'滑坡': 0.05167836,
'流域': 0.04969479,
'基础设施恶化': 0.05090763,
'人口得分': 0.04874287,
'湿地损失': 0.04697528,
'规划不足': 0.05009433,
'政策因素': 0.0497825
}

选取指标：季风强度', '地形排水', '河流管理', '森林砍伐', '城市化', '气候变化', '大坝质量', '淤积', '滑坡'

（这里要补充可视化和中间量，待完善）

灵敏度分析:

灵敏度分析结果： [47594.71524907582, 67.29198830821967, 67.29157103953024, 67.28901003974, 67.2901429061167, 67.29008915827943, 67.2908827910544, 67.28893110083564, 67.29152304250229, 67.2900915118606, 67.28958835210439, 67.2899154973684, 67.28969125150985, 67.28906424307057, 67.29119017953643, 67.29136630751331, 67.29160369340391, 67.29145443999394, 67.28941056088411, 67.29028461401795, 67.29042880172499]

第一个id远高于其他，需要剔除

->改进：假设 `id` 只是一个唯一标识符，可以从特征中移除，并重新进行权重计算和灵敏度分析：

问题 3

基于问题 1 中指标分析的结果，请建立洪水发生概率的预测模型， 从 20 个指标中选取合适指标，预测洪水发生的概率，并验证你们预测模型的准确性。如果仅用 5 个关键指标，如何调整改进你们的洪水发生概率的预测模型?

选取合适指标： （按照第二问预警系统中选取的指标）

季风强度', '地形排水', '河流管理', '森林砍伐', '城市化', '气候变化', '大坝质量', '淤积', '滑坡'

5 个关键指标： （按照第一问求出的灰色关联度排序）

1. 淤积 0.727431
2. 农业实践 0.716449
3. 政策因素 0.715782
4. 河流管理 0.715706
5. 大坝质量 0.715536

在训练集上测试，寻找最好的模型

（模型对比）

基本思路

需要尝试:

线性回归 、决策树、随机森林、支持向量、梯度提升、多层感知器 、K 近邻

（参考一下宇哲和学长打的美赛校赛论文的结构）

问题 4

基于问题 2 中建立的洪水发生概率的预测模型，预测附件 test.csv 中 所有事件发生洪水的概率，并将预测结果填入附件 submit.csv 中。然后绘制这 74 多万件发生洪水的概率的直方图和折线图，分析此结果的分布是否服从正态分布。

得出答案，输出到submit.csv中，并进行正态分布检验，没试过，先得出结论丢给chat分析叭。