B题 洪水灾害的数据分析与预测

洪水是暴雨、急剧融冰化雪、风暴潮等自然因素引起的江河湖泊水量迅速增加，或者水位迅猛上涨的一种自然现象，是自然灾害。洪水又称大水，是河流、海洋、湖泊等水体上涨超过一定水位，威胁有关地区的安全，甚至造成灾害的水流。洪水一词，在中国出自先秦《尚书·尧典》。从那时起，四千多年中有过很多次水灾记载，欧洲最早的洪水记载也远在公元前1450年。在西亚的底格里斯—幼发拉底河以及非洲的尼罗河关于洪水的记载，则可追溯到公元前40世纪。2023年6月24日8时至25日8时，中国15条河流发生超警洪水。2023年，全球洪水等造成了数十亿美元的经济损失。

洪水的频率和严重程度与人口增长趋势相当一致。迅猛的人口增长，扩大耕地，围湖造田，乱砍滥伐等人为破坏不断地改变着地表状态，改变了汇流条件，加剧了洪灾程度。在降水多的年份，洪水是否造成灾害，以及洪水灾害的大小，也离不开人为因素，长期以来人为的森林破坏是其重要原因。长江上游乱砍滥伐的恶果是惊人的水土流失。现已达35万平方千米，每年土壤浸融量达25亿吨。河流、湖泊、水库淤积的泥沙量达20亿吨。仅四川一省一年流入长江各支流的泥沙，如叠成宽高各1米的堤，可以围绕地球赤道16圈。我国第一大淡水湖洞庭湖每年沉积的泥沙达1亿多吨，有专家惊呼：“这样下去，要不了50年，洞庭湖将从地球上消失！”长江之险，险在荆江，由于泥沙俱下，如今荆江段河床比江外地面高出十多米，成了除黄河之外名副其实的地上河。对森林的肆意砍伐不仅危害自己，而且祸及子孙后代，世界上许多地方，如美索不达米亚、小亚细亚、阿尔卑斯山南坡等由于过度砍伐森林，最后都变成了不毛之地。

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。

附件test.csv中包含了超过70万的洪水数据，其中包含洪水事件的id和上述20个指标得分，缺少发生洪水的概率。附件submit.csv中包含test.csv中的洪

水事件的id，缺少发生洪水的概率。

请你们的团队通过数学建模和数据分析的方法，预测发生洪水灾害的概率，解决以下问题：

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。

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。

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？

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。

附件：

1. train.csv
2. test.csv
3. submit.csv