**（所有的ID仅作为表示）**

**（样本是否均衡，）**

**开头要有一个帽子，数据有多少条，有哪些指标，这些指标的值域是多少，每个值的分布是多少，样本是否均衡，是不是有一些离群点？**

**（甚至分析一下，在你）**

**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。

（特征发生概率之间的关系）

异常点分析

特征的分布分析

**相关性检测：皮尔逊相关系数，斯皮尔曼等级相关系数，肯德尔等级相关系数，信息增益，协方差**

**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。

概率聚类，单纯的按照洪水发生概率，映射坐标后，进行分段函数（该部分如果没有明确的分界线，需要进行重叠部位的分析）

第一问问的的是特征和洪水发生概率之间相关性，第二问要主要指标与指标之间的关联

负相关的指标可以去掉，也可以尝试利用负相关的指标和其他指标构造正相关的指标

完全同质化的指标注意去重

**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？

注意你的数据集K-Flods,进行训练

8:2分割的方式去验证你模型的有效性

XgBoost模型就爸爸了在这里，SVM，线性拟合，

如果采用其他模型，可以采用多模型投票的方式

结合第二问，在第三问中，你可以把你最终的问题处理成为拟合洪水概率的问题，也可以增加一个小节按照高中低三种方式来进行处理

**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