计算机科学与技术学院大数据分析与实践课程实验报告

实验题目：数据采样方法实践		学号：202322130197
日期：	班级：数据	姓名：崔嘉铭
Email：cjm13969665900@gmail.com		

实验目的：

本实验旨在使用 Python 中的 Pandas 库，对给定数据集进行数据清洗、过滤和采样操作。通过实现随机采样、加权采样和分层采样等方法，加深对不同数据采样策略适用场景及其差异的理解，为后续数据分析与建模奠定基础。

实验软件和硬件环境：

操作系统：Windows 10

编程语言：Python 3.9

开发环境：Jupyter Notebook

第三方库：Pandas、NumPy

实验原理和方法：

数据采样是大数据分析中的重要步骤，合理的采样方法能够在减少计算成本的同时保持数据的代表性。本实验主要使用 Pandas 提供的数据处理接口，依次完成以下操作：

1. 使用 dropna 方法删除数据中的空行，保证数据完整性；
2. 通过条件筛选过滤无效数据；
3. 使用 sample 方法实现随机采样；
4. 通过设置权重参数实现加权采样；
5. 根据分类字段进行分层采样，保证各类别数据比例符合预期。

实验步骤：（不要求罗列完整源代码）

1. 使用 Pandas 读取原始数据文件；
2. 删除存在空值的无效记录；
3. 筛选出 traffic 不为 0 且 from_level 为“一般节点”的数据；
4. 在过滤后的数据集上分别进行随机采样、加权采样和分层采样；
5. 对不同采样结果进行输出与观察。

结论分析与体会：

通过本次实验，可以发现不同采样方法适用于不同的业务场景。随机采样实现简单，但在类别分布不均衡时可能导致样本偏差；加权采样能够提升重要类别被抽中的概率；分层采样则可以严格控制样本中各类别的比例，更适合对分类结果敏感的分析任务。本实验加深了我对数据采样方法及 Pandas 数据处理流程的理解。

就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟 1－3 道问答题：

问题 1：为什么需要先删除空行再进行采样？

答：空行会影响后续的过滤和采样操作，可能导致采样结果中包含无效数据，因此需要提前清洗。