**数据说明**

本次比赛提供了15个班级从2023年8月31日到2024年1月25日共148天的学习行为模拟数据，包括学习者基本信息（共1364名）、题目基本信息（共44条）和学习者答题行为日志记录（共232818条），均以CSV格式给出。数据集中设置了缺失值、异常值或数据不一致等噪声信息（例如无效的班级、缺失的日志记录等），需要参赛者注意并处理。三个数据表具体字段说明如下：

学习者基本信息表

|  |  |  |
| --- | --- | --- |
| Data\_StudentInfo.csv | | |
| **字段名称** | **字段含义** | **相关说明** |
| index | 学习者索引 |  |
| student\_ID | 学习者编号 | 唯一编码 |
| sex | 性别 |  |
| age | 年龄 |  |
| major | 专业 |  |

题目基本信息表

|  |  |  |
| --- | --- | --- |
| Data\_TitleInfo.csv | | |
| **字段名称** | **字段含义** | **相关说明** |
| index | 题目索引 |  |
| title\_ID | 题目编号 | 唯一编码 |
| score | 题目分数 |  |
| knowledge | 知识点 | 题目可能考察多个知识点 |
| sub\_knowledge | 从属知识点 | 知识点可能存在多个从属知识点 |

学习者答题行为日志记录信息

Data\_SubmitRecord文件夹中共包含15个班级（Class1~Class15）的学习者答题行为日志记录数据文件，以SubmitRecord-Class1.csv为例。

|  |  |  |
| --- | --- | --- |
| SubmitRecord-Class1.csv | | |
| **字段名称** | **字段含义** | **相关说明** |
| index | 记录索引 |  |
| class | 班级 |  |
| time | 日志生成时间 | 时间戳，精确到秒 |
| state | 答题状态 | 例如完全正确、部分正确等，共12种情况 |
| score | 答题分数 | 通过测试用例获得的分数 |
| title\_ID | 题目编号 | 引用题目基本信息表中的title\_ID |
| method | 语言 | 学习者答题采用的编程语言 |
| memory | 内存 | 单位KB |
| timeconsume | 用时 | 单位毫秒 |
| student\_ID | 学习者编号 | 引用学习者基本信息表中的student\_ID |