数据挖掘 - 期末团队作业

姓名 + 学号

展示: 2025.12.24; 报告: 2026.1.11

1 期末团队作业要求

对于真实数据集，提出合适的数据挖掘任务，运用课程所学的数据挖掘技术予以解决。各团队既可以自行选择其他数据集，也可以选择整理好的 Yelp 数据集。对于 Yelp 数据集，可以从选题推荐中挑选感兴趣的数据挖掘任务进行解决，也可以自行构思其他任务。

期末团队作业成绩占总成绩的 30%。其中，团队展示占 5%，定于 2025 年 12 月 24 日，各小组会有 8-9 分钟的展示和 3-4 分钟的问答。最终报告占 25%，提交时间 2026 年 1 月 11 日。

团队展示和最终报告的内容包括：

- 标题和摘要：标题须具有针对性、摘要是对全文的总结和提炼。
- 选题背景及意义：介绍选题的相关背景、阐述选题动机和意义。
- 问题描述：清楚描述拟解决的数据挖掘问题。
- 数据集介绍：数据集来源、数据集构成、数据字段描述。
- 数据预处理与探索性分析。
- 数据挖掘建模分析：所使用的数据挖掘模型、模型简要原理及实施过程、模型效果评估及对比分析、模型结果可解释性分析。
- 数据挖掘的结论总结：数据挖掘的结论、模型使用过程中的不足、模型可改进的地方。
- 实践价值与展望：就数据挖掘结论的实践价值展开论述。

同时，请各团队在展示和报告中注明各成员分工。

2 自行选择数据集渠道推荐

- UCI 数据集：http://archive.ics.uci.edu/ml/。
- 阿里天池数据集平台：https://tianchi.aliyun.com/datalab/。
- Kaggle 数据科学竞赛平台：https://www.kaggle.com/

3 Yelp 数据集介绍及选题推荐

本次团队作业推荐使用的数据集整理自 Yelp 官方公开[1]的商户、点评和用户数据。整理好的 Yelp 数据集包含位于纳什维尔 (Nashville) 的所有商户信息 (bizs_nashville.txt)、所有商户在 2012 年 1 月至 2021 年 12 月的 10 年时间中累积的评论数据 (reviews_nashville.txt) 和所有相关用户个人信息 (users_nashville.txt)，各文件的数据字段名称详见 ReadMe.txt，数据字段含义详见 Yelp Dataset JSON.pdf 的高亮部分。

针对 Yelp 数据集，各团队可从以下推荐选题（排名不分先后）中挑选感兴趣的题目进行探索，也可以自行构思选题。要求选题可以运用数据挖掘技术进行解决。

- **分类分析**
  - 商户经营状况预测。
- **关联分析**
  - 业态关联分析。
- **聚类分析**
  - 餐馆聚类分析。
- ……

[1] https://www.yelp.com/dataset