期末考核题目及评分标准

1. 项目描述：

请同学们自行选择感兴趣的数据集，利用课程所学的相关数据挖掘模型和方法，展开分析和预测，以启发和指导商业决策。

2. 评分参考依据：

论文构成	内容和要求	得分
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	5 分
一、项目背景及意义	• 介绍研究问题的相关背景 • 阐述案例分析动机和意义	15 分
二、问题描述	• 清楚描述拟解决的数据挖掘问题 • 如有可能可对该问题进行建模 • 给出问题的数学定义	10 分
三、数据集介绍	• 数据集描述 • 数据集来源 • 数据表构成、字段描述等	5 分
四、数据的预处理与探索性分析	• 数据的预处理：如缺失值/异常值/特征变换等 • 数据探索分析：可视化、相关分析等	20 分
五、数据建模分析	• 所使用的数据挖掘技术 • 数据挖掘方法原理和实施过程 • 实验设置、参数调节、 • 模型效果评估：如多种方法比较 • 模型可解释性分析	25 分
六、数据分析的结论总结	• 数据挖掘分析的结论 • 模型使用过程中的不足 • 模型可改进的地方	15 分
七、实践价值与展望	• 就数据分析结论的实践价值展开论述	5 分

3. 项目辅助性链接（用于寻找数据集）：

- UCI 数据集： http://archive.ics.uci.edu/ml/
- 阿里天池数据集平台：
  https://tianchi.aliyun.com/datalab/index.htm?spm=5176.100065.1234.4.zDOlMf

- Kaggle 数据科学竞赛平台：https://www.kaggle.com/
- 优矿-通联量化实验室，大数据时代的金融量化平台：https://uqer.io/home/
- 斯坦福大型网络关系开放数据集：http://snap.stanford.edu/data/index.html
- 推荐系统与个性化相关数据集：https://cseweb.ucsd.edu/~jmcauley/datasets.html

4. 参考数据集

a) 健康与医疗：糖尿病预测&分类问题

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

数据背景：通过电话问卷采集，记录了个体生活习惯、身体指标、干预措施等特征属性和是否患糖尿病这一类别标签的联系。

研究问题：

- 结合多项特征进行糖尿病预测；
- 探究导致糖尿病的高危因素；
- 能否基于少量属性，提供糖尿病的精准预测，从而简化问卷设计。

难度系数：🌟🌟🌟

b) 课程与教学：评论文本&情感分析&评分预测

https://www.kaggle.com/datasets/anthonysusevski/course-reviews-university-of-waterloo

数据背景：来自 Waterloo 大学的课程评价，记录了课程名、评价文本、是否喜欢课程、难度评价等；

研究问题：

- 基于不同课程评价文本进行文本预处理&词云分析；
- 基于评价文本预测学生是否喜欢课程；
- 分析影响学生对课程喜欢程度的影响因素，是否存在课程差异。

难度系数：🌟🌟🌟🌟

5. 提交要求：

（1） 论文格式：正文 12 号字，1.5 倍行距，12-25 页；

（2） 电子版本统一通过 Canvas 平台进行提交；

（3） 期末报告展示：第 16 周课上展示。

6. 建议的项目报告格式：

《数据挖掘与商务分析》期末报告

-----项目/案例分析标题----

小组成员信息：

学号	姓名	工作分工	贡献比

完稿日期：_____

项目/案例分析标题

【摘要】

【用以对项目/案例分析进行全面的总结和归纳（300 字以内）】

一、 项目/案例分析的背景及意义

二、 项目/案例分析所涉及的数据分析问题/科学问题

三、 数据集介绍

四、 数据预处理与探索性分析

五、 数据建模分析

六、 数据分析的结论总结

七、 项目/案例的实践价值与展望

附录（参考文献、主要代码等）