大数据分析课程项目

温州大学 | 数理学院 | 2023-2024 学年第二学期

大数据分析的课程项目要求每位同学基于在课程中学习到的数据获取、数据预处理、数据分析的方法，**独立完成**一项与数据分析相关的项目，使用 Jupyter Notebook，通过 Python 代码和文字说明结合的方式，阐述整个数据分析项目的流程。课程项目包括 6 个阶段，截止时间与评分标准如下：

内容要求	截止时间	评分占比
问题定义与数据获取	2024/5/10(周五)，21:00 前	10%
数据预处理与探索性数据分析	2024/5/24(周五)，21:00 前	20%
课程项目初稿	2024/5/31(周五)，21:00 前	20%
课程项目终稿	2024/6/11(周二)，21:00 前	30%
课程项目演示	2024/6/12,6/13,6/14	20%
项目互评	被评价者演示当天的 21:00 前	（+5%）

问题定义与数据获取(10%)

1. 陈述你想要通过数据分析项目**解答什么问题**：陈述问题的定义，定义需要具体且聚焦。
2. 陈述你提出**问题的意义与价值**：评估你问题的重要性、原创性、可行性。
3. 陈述你的**数据来源**，以及你是**如何获取数据**的。以下是推荐的一些公开的数据平台：
   a) Kaggle: https://www.kaggle.com/datasets
   b) 联合国数据库：https://data.un.org/
   c) UCI 机器学习资源库: https://archive.ics.uci.edu/ml/index.php
   d) 微软数据集: https://msropendata.com/
   e) Harvard Dataverse: https://dataverse.harvard.edu/
   f) 各平台的官方网站
4. **描述你获得的数据**，回答包括但不限于以下的问题：
   a) 你的数据中有哪些重要的变量？这些变量的数据类型是什么？
   b) 在你的数据中，什么是解释变量？什么是被解释变量？
5. 陈述你**初步计划**如何通过这个数据去回答你的问题，描述可能的途径和方法。
6. 提交材料：**数据文件**(尽量存储成.csv 或者.json 文件)以及**.ipynb 文件**和**.html 文件**（包含所有回答上述问题必要的代码和文本，并确保文档的格式清晰、逻辑通顺。

数据预处理与探索性数据分析(20%)

1. 陈述**原始数据中存在的问题**以及**数据预处理的过程**，回答包括但不限于以下的问题：
   a) 原始数据中存在什么问题？你是怎么发现这些问题的？

b) 你计划如何解决这些问题？为什么选择了特定的解决方式？

2. 陈述你是**如何进行探索性数据分析**，回答包括但不限于以下的问题：
   a) 你针对哪些数据进行了探索性数据分析？为什么？
   b) 你用什么样的方式进行了探索性数据分析？你为什么选择这些方式？
   c) 你在探索性数据分析中发现了什么？这些分析能如何帮助？
   d) 同时使用图表和可视化两种方式进行探索性数据分析。
3. 提交材料：**.ipynb 文件**和**.html 文件**（包含所有回答上述问题必要的代码和文本，并确保文档的格式清晰、逻辑通顺）。

课程项目初稿(20%)

1. 选择**至少两个常见的算法模型**进行你的数据分析，陈述你数据分析的过程，回答包括但不限于以下的问题：
   a) 你用了什么模型进行了数据分析？为什么选择这个模型？
   b) 每个模型有什么参数？你如何设置你模型的参数？为什么这么设置？
   c) 每个模型评价的相关数据有什么？你如何根据这些数据去评价你的数据？
   d) 你基于数据分析有什么发现？这些发现是如何帮助你回答你的问题的？
2. 提交材料：**.ipynb 文件**和**.html 文件**（包含所有回答上述问题必要的代码和文本，并确保文档的格式清晰、逻辑通顺）。

课程项目终稿(30%)

1. 基于课程项目初稿及其反馈，**补充和完善**包括但不限于以下的内容：
   a) 对初稿中的模型进行优化。
   b) 对比使用的模型的优劣，不同的模型在分析的过程中分别有什么优势与不足？
   c) 总结你在整个数据分析的过程中的发现和收获，并对你提出的问题提供一个阶段性的回答。
   d) 思考如果需要进一步优化和深入这方面的研究，你是具体如何打算的？
2. **检查整个课程项目流程的完整性和逻辑的连贯性**。
3. 提交材料：**.ipynb 文件**和**.html 文件**（包含所有回答上述问题必要的代码和文本，并确保文档的格式清晰、逻辑通顺）。

演示(20%)

1. 课程演示将于 6 月 12 日、6 月 13 日与 6 月 14 日上课期间进行。
2. 每个同学有 5 分钟的时间，基于 6 月 11 日提交的.ipynb 文件，简要介绍分享以下内容：
   a) 数据分析的目的以及意义
   b) 数据获取的步骤
   c) 数据预处理的步骤
   d) 数据分析所用到的模型
   e) 结论成果和收获
3. 将以**现场随机抽取**的方式抽取当天需要进行演示的同学。
4. 演示的评分将从**完整度、清晰流畅度、逻辑性**三个方面进行综合评估。

项目互评(+5%)

1. 每个同学将有机会评价 3 个随机分配的同学的课程项目，并提供 100 字左右的关于课程项目的点评和建议。评价内容本身将不影响被评价同学的最终得分，但评价者可以根据评价的完成度获得最高 5%的课程项目附加分。
2. 评价分配表将在课程项目演示前进行发放，项目互评的将通过问卷星的形式进行收集。

其他注意事项

1. 课程项目原则上不接受延迟提交，每一个过程模块都将进行阶段性评分。
2. 如果中途想要更改数据分析的主题与内容，请尽量第一时间与老师沟通，原则上每个过程模块的截止日期不会因为主题与内容的更改而延迟，请在一开始就慎重选择想要分析的主题。
3. 若出现其他突发状况，请第一时间与老师联系。