《大数据基础与实践》课程大作业要求

一、课程大作业要求

1. 大作业为个人作业，即每人独立完成；
2. 大作业立项由学生在第 5 次课后的作业中进行立项选题说明，包括项目名称、项目针对的问题、预期达到的效果、拟采用的数据集和数据收集方法等；此次立项后可以对立项内容继续进行完善和修改，但是必须此次立项时给出第一版的立项选题说明；
3. 大作业应该以 Notebook 形式完成，无论立项选题如何，都应该包含：问题描述、数据收集、数据预处理(数据清洗与转换)、数据建模分析(应包含统计分析和运用机器学习方法的分析两部分)、数据可视化等内容。Notebook 编写时注意以下问题
    1) 通过文本 Cell 添加必要的说明文字，不要只包含代码，或者在一个 Cell 中包含大量的代码；
    2) 代码 Cell 中应该添加适当的注释，以方便他人阅读理解代码；
    3) 使用函数将代码结构化，尽量复用代码，不要出现大量近似的代码；
    4) 在 Notebook 最前面用文字说明使用了哪些需要安装第三方工具库的包，以便助教在批改作业时安装配置环境；
    5) Notebook 使用的数据集要随 Notebook 一并压缩提交。
4. 大作业验收为答辩形式，先演示讲解 Notebook，然后回答问题。

二、大作业评分标准

总分 52 分

1. 数据获取 5 分
    1) 通过爬虫自己收集数据(Notebook 中包含爬虫代码)：5 分
    2) 通过网站下载或其他形式获取现成数据：2 分
2. 数据预处理 10 分
    1) 数据预处理方案合理，且符合大作业需求：10 分
    2) 数据预处理方案存在不合理之处：7 分
    3) 直接使用现成数据无需预处理：3 分
    4) 未进行数据预处理，导致数据集不符合大作业需求：0 分
3. 数据分析 15 分，包含下列 2 部分：
    1) 数据统计分析 6 分
    2) 基于机器学习的数据分析 9 分
        a) 只使用回归、聚类、决策树等简单机器学习方法进行数据分析：6 分
        b) 使用基于神经网络等进阶学习方法进行数据分析：9 分
4. 数据可视化 10 分，包含下列 3 部分：
    1) 绘制折线图、箱线图、条状图、热力图等简单图形 5 分
    2) 使用地图、词云等高级方式可视化数据 3 分
    3) 可视化方式合理、用户体验好：2 分
5. Notebook 编写规范 5 分，包含下列 2 部分：
    1) Notebook 中包含数量恰当的文字说明和代码注释：3 分
    2) 代码结构合理，通过函数提高代码复用性：2 分
6. 现场答辩 7 分，包含下列 2 部分：

1) 现场演示效果：3 分

   a) 演示结果正确，未出现异常：3 分

   b) 演示结果有误，出现异常：0 分

2) 回答问题正确：4 分

三、平时作业评分

每周布置 1 次作业，除答辩验收外，共 7 次作业 8 个 Notebook，每个 Notebook 中的任务 6 分，共 48 分。