


人工智能（研讨）大作业

Assignment

Multi-Class Prediction of Obesity Risk

The goal is to use various factors to predict obesity risk in individuals, which is related to cardiovascular disease.

题目链接 

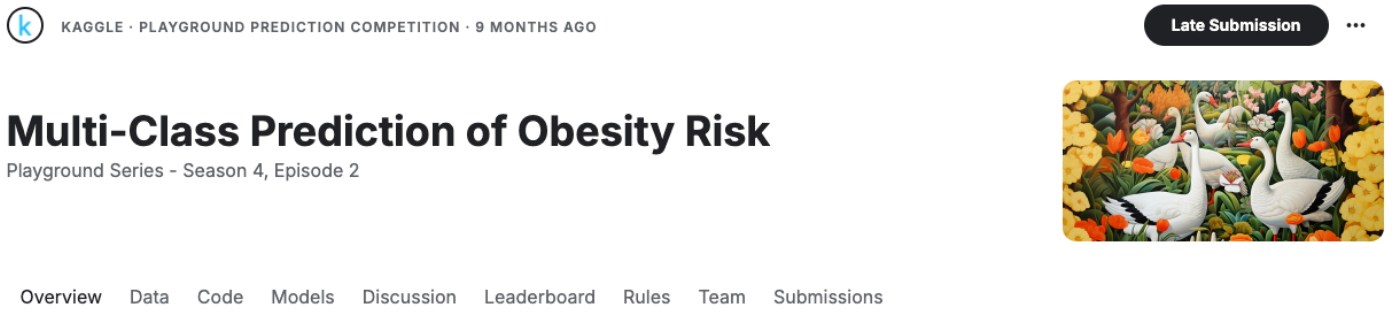
<https://www.kaggle.com/competitions/playground-series-s4e2/overview>

Keywords:

Tabular; Multiclass Classification; Accuracy Score

Kaggle

- Kaggle：一个进行数据挖掘和预测竞赛的在线平台



- 你需要做的：
 - ☐ 组队：不超过3个人
 - ☐ 仔细浏览题目描述，包括题目定义、数据描述、结果评估以及提交格式...
 - ☐ 下载题目提供的数据集（train/test，包括从训练集中划分出validate），并完成数据预处理，这将对后续预测的准确度起很大决定性。
 - ☐ 通过机器学习方法完成题目要求的预测，你可以使用Python（阅读学习sklearn库），Matlab, pytorch（深度学习）.....
 - ☐ 按提交格式要求生成结果（id, NObeyesdad），通常是.csv文件，上传到系统后查看accuracy score，并可以通过浏览Leaderboard定位你的大致rank。因为此项赛事已经结束，所以你的成绩不会再上传排名。

Tips

- 学习使用Jupyter Notebook/Colab（.ipynb文件），通过逐块执行脚本以及免费算力来提高生产效率
- This dataset is particularly well suited for visualizations, clustering, and general EDA. Show off your skills!
- Models提供了LLM的api也许可以尝试

Submission

- 本学期最后两周课，每队做一次**PPT presentation**，**不超过12分钟**，要求讲清楚如下过程：
 - 数据分析及预处理
 - 模型选择，至少选择2-3个模型，并讲清楚为何选择这些模型
 - 最终选择的模型的具体方法
 - 调参过程
 - 结果分析及展示
 - 队员具体贡献
 - 完成任务后，你需要提交一个**压缩包**，包括：
 - a. 一份**报告**，PDF格式，包括但不限于：
 - i. 问题背景描述
 - ii. 数据清洗及预处理
 - iii. 预测具体方法描述
 - iv. 实验结果分析
 - v. 系统给出的score证明
 - b. 你的完整**代码**:
 - .ipynb文件可能会是推荐做法
 - 给出适当的注释
 - c. 小组的PPT
- ☐ 以**小组为单位**，2-3个人一个小组
- ☐ 命名方式: **学号（8位数）-姓名-大作业**，打包上传，每个人都要提交，同组人的内容保持一致
- ☐ **提交地址**: <https://pan.seu.edu.cn:443/link/607B8138C761256941104B561B216087>
- ☐ **Deadline**: **考试周最后一周**

Grading Criteria

三个部分：PPT汇报演示、技术报告、代码运行结果

汇报展示人数不限，讲清楚分工

报告每个小组一份，组内成员分数的差异仅在于分工多少。

PPT汇报演示	15% (个人表现)	工作量在小组中的占比、是否参加答辩和回答问题等
	15% (小组表现)	内容是否完整：数据处理、方法选择、方法对比、实验结果分析和对比、结论 方法是否正确：有没有选择正确的预测方法、有没有正确定义loss等 答辩是否顺畅：有没有解释清楚、正确回答问题等
技术报告	30%	内容（包括数据分析、方法介绍、代码架构和实验结果分析）是否完整，结果是否正确、结果分析是否到位等
代码运行结果	40%	按照代码运行结果评分排名RANK打分

Q&A

有问题可以留言 - QQ: 1104471741（助教李东晓）



Good Luck!