

大学生学习生活数据集说明

1. 数据集简介

本数据集创建于 2023 年，包含江南大学君远学院选修机器学习与人工智能课程学生的 27 项学习生活相关数据，主要用于课程期末大作业。数据集未指明特征或标记，所有数据均假设为特征 feature，但可根据要研究的问题，选定若干项数据连接组合为特征 X，选定某项数据为标记 y，用于解决分类、回归两大类问题。

数据集历年采集情况如表 1 所示。

表 1 数据集历年采集情况

| 采集年份 | 专业 | 样本数 |
|------|--------|-----|
| 2023 | 机械电子工程 | 51 |
| 2024 | 机器人工程 | 63 |
| 样本合计 | | 114 |

2. 特征数据的组成

数据集包含的 27 项数据如表 2。

注意 1：所有数据均设置为 float 类型，在解决分类问题时，如有需要，可自行将类别型数据的 0、1（如性别、出生地等）转换成 int 类型。其中部分类别型数据也可视为数值，可作为特征参与模型学习，或作为标记用于解决回归问题；

注意 2：部分数据有偏斜，分类时要考虑混淆矩阵等指标；

注意 3：部分数据可能存在 outlier 点，需要考虑其影响；

注意 4：由于样本规模较小，最终结果不一定可信。但不妨碍提出要研究的问题、不妨碍运用机器学习方法、不妨碍得出最终结果；

注意 5：样本以提交调查表时间排序，可视为随机，但使用时仍可作 shuffle 操作。

表 2 特征数据组成

| 序号 | 数据项 | 字段名 | 类别型数据的分类值及说明 | 备注 |
|----|-----|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|
| 1 | 年份 | Year | | 目前均为 2023 |
| 2 | 性别 | Gender | [0]男 [1]女 | |
| 3 | 出生地 | Province | [0]北京市 [18]广东省 [1]天津市 [19]广西壮族自治区 [2]山西省 [20]海南省 [3]河北省 [21]四川省 [4]内蒙古自治区 [22]贵州省 [5]黑龙江省 [23]云南省 [6]吉林省 [24]重庆市 [7]辽宁省 [25]西藏自治区 [8]上海市 [26]陕西省 | 可自行组合，分类为南方、北方、西部、东部等。 |

| | | | | |
|----|----------------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|
| | | | [9]江苏省 [10]浙江省 [11]安徽省 [12]福建省 [13]江西省 [14]山东省 [15]河南省 [16]湖北省 [17]湖南省 [27]甘肃省 [28]青海省 [29]宁夏回族自治区 [30]新疆维吾尔自治区 [31]香港特别行政区 [32]澳门特别行政区 [33]台湾省 [34]海外 | |
| 4 | 身高 (cm) | Height | | 身高和体重也可组成 BMI 数据 |
| 5 | 体重 (kg) | Weight | | |
| 6 | 是否脱单 | Unavailable | [0]否 [1]是 | 如作为数值使用, 可计算脱单概率 |
| 7 | 近视度数 (两眼平均) | Shortsight | | |
| 8 | 平均每天吃饭花费 (元) | Dailymealexp | | |
| 9 | 平均每次吃多少米饭 (元) | Riceexp | | |
| 10 | 每月生活费 (元) | Monthlyexp | | |
| 11 | 每周点外卖次数 | Weeklytakeout | | |
| 12 | 每月网购次数 | Monthlyshopping | | |
| 13 | 平均每天睡眠时长 (小时) | Sleeptime | | |
| 14 | 平均每天入睡时间 | Timefallasleep | [0]22:00 以前 [1]22:00-22:30 [2]22:30-23:00 [3]23:00-23:30 [4]23:30-00:00 [5]00:00-00:30 [6]00:30-01:00 [7]01:00-01:30 [8]01:30-02:00 [9]02:00 以后 | 可作为数值使用, 比如 [0] 可视为 21:45, [1] 可视为 22:15 [1]反映了从 21:45 开始, 经过了 1 个 30 分钟, [2]则经过了 2 个 30 分钟 |
| 15 | 平均每天刷手机时长 (小时) | Dailyslidetime | | |
| 16 | 平均每天打游戏时长 (小时) | Dailygametime | | |
| 17 | 平均每天运动时长 (小时) | Dailysporttime | | |
| 18 | 每周和父母联系次数 | Weeklyfamilyfreq | | |
| 19 | 每周洗头次数 | Weeklyhairwash | | |
| 20 | 身体状况自我评价 | Healthstatus | [0]非常好 [1]比较好 [2]一般 [3]较差 [4]很差 | 可作为数值使用 |
| 21 | 脱发情况自我评价 | Hairloss | [0]不脱发 [1]微微脱发 [2]较严重脱发 | 可作为数值使用 |

| | | | | |
|----|----------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------|---------|
| | | | [3]严重脱发 | |
| 22 | 平均每天自习时长（小时） | Dailyselvestudy | | |
| 23 | 最常去的自习场所 | Selfstudypplace | [0]图书馆 [1]寝室 [2]教室 [3]其他 | |
| 24 | 上课时有效听课时间占比（%） | Lessoneffect | | |
| 25 | 平均每周完成作业用时（小时） | Weelyworktime | | |
| 26 | 焦虑状况自我评价 | Anxiousstatus | [0]不焦虑 [1]稍有焦虑 [2]较严重焦虑 [3]严重焦虑 | 可作为数值使用 |
| 27 | 班级成绩排名 | Rank | [0]前 10% [1]10%~20% [2]20%~30% [3]30%~40% [4]40%~50% [5]50%~60% [6]60%~70% [7]70%~80% [8]80%~90% [9]90%以后 | 可作为数值使用 |

3. 数据集的组成

- (1) 数据集依托 2 个.csv 文件和 1 个.py 文件生成：
- undergradute_dataset.csv： 包含所有样本的 27 项数据（见表 2）。
 - undergradute_dataset_classname.csv： 包含数据集基本信息，以及类别型数据的分类值和说明（详见表 2）。
 - dataset_setup.py： 包含了经.csv 文件生成数据集的函数

- (2) 数据集基本信息，见表 3。

表 3 数据集基本信息

| 序号 | 字段名 | 说明 |
|----|--------------------|-----------------------------------------|
| 1 | n_samples | 样本总数 |
| 2 | n_features | 数据特征数量（27 项） |
| 3 | classname_filepath | undergradute_dataset_classname.csv 文件路径 |
| 4 | dataset_filepath | undergradute_dataset.csv 文件路径 |
| 5 | all_feature_descr | 各数据项的中文解释 |
| 6 | all_feature_names | 各数据项的英文字段 |

- (3) 类别型数据的分类值和说明，字段见表 4，详见表 2。

表 4 类别型数据的分类值和说明

