# 学前预警文档

运行环境

- python2 ,ipython notebook
- numpy,pandas,matplotlib,scikit-learn0.18

运行代码

- 学生就业数据特征提取：jiu_ye_feature.ipynb,运行后的结果保存在jy_data文件
- 贫困学生数据特征提取：poor_students_feature.ipynb,运行结果保存在poor_student_data文件夹里
- 心理预警学生数据特征提取：psy_warning_feature.ipynb,运行结果保存在psy_data文件里
- 入学学生数据特征提取：ru_xue_feature.ipynb,运行结果保存在rx_data文件
- 学业预警学生数据特征提取：students_warning_feature.ipynb,运行结果保存在student_warning_data文件

  以上文件运行无先后顺序，运行以上文件后再按顺序运行以下文件
- 把所有特征都连成一张表：merge_all_features.ipynb,运行结果保存在all_features文件
- 运行模型：model.ipynb,运行结果保存在output_result文件

额外的数据

- 保存在extra_data文件里，其包含有：
  1. 2012wlkfsx.csv——文理科分数线
  2. csdj.csv——城市等级
  3. csdm.csv——城市代码
  4. pai_ming.jpg——各省份基础教育排名图

所有特征的中文对照

| **sno** | **zf** | **tdf** | **female** | **male** | **ks_is_jf** |
|---|---|---|---|---|---|
| 学号 | 高考总分 | 加分后的总分 | 女 | 男 | 考生是否加分 |
| **ncyi** | **ncwj** | **czyj** | **czwj** | **jydf** | **csdm** |
| 农村应届 | 农村往届 | 城镇应届 | 城镇往届 | 省份基础教育得分 | 城市代码 |
| **kslb** | **kldm** | **csdj** | **is_qu** | **is_shi** | **is_xian** |
| 考试类别 | 科类代码 | 城市等级 | 家庭所在地是区 | 家庭所在地是市 | 家庭所在地是县 |
| **ybfsx_wk** | **ebfsx_wk** | **ybfsx_lk** | **ebfsx_lk** | **zfdj** | **zybm** |
| 文科一本线 | 文科二本线 | 理科一本线 | 理科二本线 | 高考分数等级 | 专业编码 |
| **xybm** | **poor_level** | **is_poor** | **jyzk** | **is_psy_warning** | **zy_pjf** |
| 学院编码 | 贫困等级 | 是否贫困 | 就业状况 | 是否心理预警 | 各专业平均分 |

| ssmzlqf_wk | ssmzlqf_lk |
| --- | --- |
| 民大少数民族文科最低录取分 | 民大少数民族理科最低录取分 |
| **hzlqf_wk** | **hzlqf_lk** |
| 民大汉族文科最低录取分 | 民大汉族理科最低录取分 |
| **zf_yb_d** | **zf_lqf** |
| 考生总分与一本分数线的差值 | 考生总分与民大最低录取分的差值 |
| **zf_pjf_d** | **zy_fs_pm** |
| 考生总分与本专业平均分的差值 | 考生的分数所在专业排名 |

模型

- 算法：`GradientBoostingClassifier`与`RandomForestClassifier`
- 标准化：`StandardScaler`
- 分层交叉验证：`StratifiedKFold`
- 划分测试集与训练集：`StratifiedShuffleSplit`
- 网格调参：`GridSearchCV`

详细用法请参见scikit-learn0.18官网文档 http://scikit-learn.org/stable/index.html

参考的网站

1. 使用sklearn进行集成学习 http://www.cnblogs.com/jasonfreak/p/5720137.html
2. Gradient Boosting tree 调参 http://blog.csdn.net/han_xiaoyang/article/details/52663170
3. 城市排名2017 http://www.mafengwo.cn/travel-news/542672.html
4. 中国各省份基础教育排名 http://www.360doc.com/content/15/1018/16/1353443_506512528.shtml
5. 处理数据类别不均衡的一些参考 http://www.tuicool.com/articles/EBJvEb3
6. 国外有名期刊上的一篇神文 http://jmlr.org/papers/v15/delgado14a.html ,摘要下有一个`[pdf]`按钮，点开就可以浏览