# 各课题环节日志记录清单

## 提出主题

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-05-01** | **研究** | **提出课题研究的AI及其它领域结合的方向。** | **1h** |
| **2023-05-02** | **研究** | **讨论一些热点领域，如果住房、医疗、教育，最终选择教育领域的问题。** | **2h** |
| **2023-06-24** | **研究** | **讨论教育领域突出问题，并选择的择校及内卷问题作为关注点，提出通过AI模型来帮助教育部门决策义务教育经费投入公平的问题。** | **2h** |
| **2023-06-24** | **文档** | **日志记录。** | **1h** |

## 制定方案

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-07-01** | **设计** | **思考课题范围及基本步骤。** | **2h** |
| **2023-07-04** | **研究** | **学习上海高中研究性课题报告的相关知识。** | **2h** |
| **2023-07-05** | **研究** | **学习上海高中研究性课题报告的相关知识。** | **2h** |
| **2023-07-08** | **设计** | **明确课题范围及基本步骤。** | **3h** |
| **2023-07-09** | **设计** | **选择合适的AI模型。** | **3h** |
| **2023-07-09** | **文档** | **日志记录。** | **1h** |

## 收集数据

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-07-10** | **研究** | **收集上海各区县历史人口、GDP数据。** | **3h** |
| **2023-07-12** | **研究** | **收集上海各区县历史人口、GDP数据。** | **2h** |
| **2023-07-14** | **研究** | **收集上海各区县历史教育投入数据。** | **2h** |
| **2023-07-15** | **研究** | **收集上海各区县历史教育投入数据。** | **3h** |
| **2023-07-15** | **文档** | **日志记录。** | **3h** |

## 数据分析

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-07-16** | **开发** | **搭建Python数据分析及开发环境。** | **3h** |
| **2023-07-17** | **开发** | **整理并分析收集到的上海各区县历史人口、GDP、教育投入数据。** | **3h** |
| **2023-07-18** | **开发** | **清洗数据，为模型训练做准备。** | **3h** |
| **2023-07-18** | **文档** | **日志记录。** | **1h** |

## 特征工程

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-07-22** | **开发** | **标签特征正态化。** | **3h** |
| **2023-07-22** | **开发** | **数据标准化。** | **3h** |
| **2023-07-22** | **文档** | **日志记录。** | **1h** |

## 训练模型

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **用时** |
| **2023-07-23** | **开发** | **构建训练、测试样本集。** | **3h** |
| **2023-07-23** | **开发** | **训练线性模型。** | **3h** |
| **2023-07-27** | **开发** | **训练非线性模型。** | **3h** |
| **2023-07-29** | **开发** | **集成线性及非线性模型。** | **3h** |
| **2023-07-29** | **文档** | **日志记录。** | **1h** |

## 预测及分析

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **字数** |
| **2023-07-30** | **开发** | **通过集成模型预测上海各区县相应年份义务教育经费。** | **3h** |
| **2023-07-30** | **开发** | **开发上海各区县义务教育经费多寡趋势图。** | **3h** |
| **2023-07-30** | **文档** | **日志记录。** | **1h** |

## 课题总结

|  |  |  |  |
| --- | --- | --- | --- |
| **时间** | **操作** | **内容** | **字数** |
| **2023-07-31** | **文档** | **撰写论文报告。** | **4h** |
| **2023-08-02** | **文档** | **整理研究日志。** | **4h** |
| **2023-08-05** | **文档** | **撰写论文报告。** | **4h** |
| **2023-08-06** | **文档** | **撰写论文报告。** | **4h** |

# 日志记录汇总统计

## 各课题环节日志记录清单

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **环节** | **研究** | **设计** | **开发** | **文档** |
| **提出问题** | 5 |  |  | 1 |
| **制定方案** | 4 | 8 |  |  |
| **收集数据** | 10 |  |  | 3 |
| **分析数据** |  |  | 9 | 1 |
| **特征工程** |  |  | 6 | 1 |
| **训练模型** |  |  | 12 | 1 |
| **预测及分析** |  |  | 6 | 1 |
| **课题总结** |  |  |  | 16 |

## **所有日志时间分布图**

# **日志详情**

## **提出主题**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 提出主题 | | |
| 建立日期 | 2023-05-01 | 最终日期 | 2023-06-24 | 版本次数 | 1 |
| 课题来源 | 教育事关国计民生，实现教育公平是教育工作的重点之一。在国家越来也重视素质教育的大背景下，实现教育资源的合理分配，将会有效解决长期困扰人们的择校问题，解决义务教育阶段唯分数是从的内卷现象。  在信息高度发达的当今社会，AI已被广泛接受。基于AI模型进行分析和预测，将会为教育工作提供决策支持的新视角。  人工综合多要素大数据，进行趋势预测分析，是比较困难的。 | | | | |
| 期望效果 | 基于AI模型分析上海各区在义务教育阶段的投入数据，结合人口、出生率、GDP等数据，建立上海各区的义务教育阶段投入的模型，预测各区的投入，进而与各区实际投入进行比较，从而得出各区教育经费投入多寡的趋势，为教育决策提供新视角。 | | | | |

## 制定方案

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 制定方案 | | |
| 建立日期 | **2023-07-01** | 最终日期 | **2023-07-09** | 版本次数 | 1 |
| 总统方案 | 参考文献进行数据搜集；利用Python进行数据分析；运用标准化和正态化进行特征工程；通过相关性分析选择AI模型；利用多模型集成方法来解决小数据集模型泛化能力弱的问题；通过交叉验证、网格搜索的方法来选择模型的超参；运用sklearn进行模型训练和预测；利用matplotlib绘制各区教育经费投入多寡的趋势图。 | | | | |
| 研究计划 | 1. **收集上海各区县历史人口、GDP数据** 2. **收集上海各区县历史教育投入数据。** 3. **搭建Python数据分析及开发环境。** 4. **分析、整理、清洗原始数据。** 5. **研究并选择合适的模型。** 6. 对原始数据进行特征工程。 7. 训练模型。 8. 利用模型进行预测上海各区县历史义务教育经费，并观察预测效果。 9. 绘制上海各区县历史义务教育经费多寡情况，并分析趋势。 10. 总结分析结论。 11. 完善课题报告文档。 | | | | |

## **收集数据**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 收集数据 | | |
| 建立日期 | **2023-07-10** | 最终日期 | **2023-07-15** | 版本次数 | 1 |
| 输入 | 上海各区常住人口、出生率、GDP、历年学生数据：  <https://population.gotohui.com/pdata-3>  上海各区教育投入：  <http://edu.sh.gov.cn/xxgk2_zdgz_cwgl_01/20221209/afcf2d9330fb4239b250e33aa36aedf5.html?eqid=f4303090008129ac0000000564311720>  <http://edu.sh.gov.cn/xxgk2_zdgz_cwgl_01/20211231/fa6ee6c94d794cb38ef8537cae3b7491.html>  <http://edu.sh.gov.cn/xxgk2_zdgz_cwgl_01/20201103/4692cd35e4d5448ba41533873e2e66da.html>  <http://edu.sh.gov.cn/shgy_jftrsy_cwgl_2/20211217/8bbf19e9b9d1486287c7069112536d2d.html>  <http://edu.sh.gov.cn/shgy_jftrsy_cwgl_2/20211222/v2-0015-gw_902992018003.html>  <https://edu.sh.gov.cn/xxgk2_zdgz_cwgl_01/20201015/v2-0015-gw_902992018001.html> | | | | |
| 输出 | 数据保存于“原始数据.xlsx”。  包含如下列：  区县，常住人口（万人，出生率(‰)，GDP（亿元），义务教育投入，年份，教育财政拨款情况（万元），高中在校学生人均经费情况（元），中等职业学校在校学生人均经费情况（元），初中在校学生人均经费情况（元），小学在校学生人均经费情况（元），幼儿园在园幼儿人均经费情况（元），大学生（万人），高中生（万人），初中生（万人），小学生（万人），在园幼儿（万人） | | | | |
| 难点 | 1. 免费网站数据不全，且不统一，难以符合要求。 2. 最近两年的数据缺失。 3. 中心城区数据较少。 | | | | |

## **数据分析**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 数据分析 | | |
| 建立日期 | **2023-07-16** | 最终日期 | **2023-07-18** | 版本次数 | 1 |
| 输入 | 收集到的上海各区县历史人口、GDP、教育投入等数据。见附件“原始数据.xlsx”。 | | | | |
| 输出 | 1. 分析数据缺失情况。并选择合适的行和列作为初步特征数据集。 2. 分析数据标签特征列的分布情况。 3. 分析各特征列与标签列的相关性。 | | | | |
| 分析过程 | 1. 虽然有125行数据，但只有46行符合要求。 2. 我们可以用到“常住人口（万人） ”、“ 出生率(‰)”、“GDP （亿元） ”、“在园幼儿（万人）”、“小学生（万人） ”、“初中生（万人）”这六个特征，同时通过计算获得“义务教育经费（万元）”数据作为模型样本标签。 3. 计算获得的标签列“义务教育经费（万元）”柱状图偏左。后续需进行转换，以尽量服从正态分布。 4. 使用热力图查看字段间的相关性，显示“义务教育经费（万元） ”与其它六个特征都具有较强的相关性。 | | | | |

## **特征工程**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 特征工程 | | |
| 建立日期 | **2023-07-22** | 最终日期 | **2023-07-22** | 版本次数 | 1 |
| 输入 | 从“原始数据.xlsx”经过整理、清理后的初步特征数据集，包含如下列：  常住人口（万人），出生率(‰)，GDP（亿元），在园幼儿（万人），小学生（万人），初中生（万人），义务教育经费（万元）。 | | | | |
| 输出 | 标准数据集 | | | | |
| 过程记录 | 1. 对标签列“义务教育经费（万元）”进行转换 ，使之尽量呈正态分布。 2. 对非标签列进行标准化，使用z-score标准分数。。 | | | | |

## **训练模型**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 训练模型 | | |
| 建立日期 | **2023-07-23** | 最终日期 | **2023-07-23** | 版本次数 | 1 |
| 输入 | 经过特征工程处理后的标准数据集。 | | | | |
| 输出 | 训练好的模型 | | | | |
| 过程记录 | 1. 选择合适模型，目前仅选择岭回归和随机森林，作为线性回归和非线性回归的模型。将来，可以考虑神经网络作为非线性模型。 2. 由于样本数少，先将数据集分割为训练集和测试集，验证模型在测试集的效果。最终将所有样本作为训练集来训练模型。 3. 使用交叉验证及网格搜索的方法获得模型的最佳超参。 4. 使用求两个模型预测值的平均值的集成方法，最为最终的模型预测结果。 | | | | |

## **预测及分析**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 预测及分析 | | |
| 建立日期 | **2023-07-30** | 最终日期 | **2023-07-30** | 版本次数 | 1 |
| 输入 | 1. 训练好的岭回归和随机森林模型。 2. 求两个模型预测值的平均值的集成方法。 3. 处理过的上海各区县历史义务教育经费数据集。 | | | | |
| 输出 | 1. 上海各区县历史义务教育经费预测值。 2. 上海各区县历史义务教育经费多寡走势图。 3. 上海各区县历史义务教育经费多寡分析结论。 | | | | |
| 过程记录 | 1. 利用训练好的岭回归和随机森林模型，分别对处理过的上海各区县历史义务教育经费数据集预测义务教育经费。 2. 利用求两个模型预测值的平均值的集成方法，求出集成预测值。 3. 求出集成预测值与实际标签值的差值，并且百分化。 4. 使用matplotlib画出上海各区县历史义务教育经费多寡走势组合柱状图。 5. 分析上海各区县历史义务教育经费多寡走势情况。 | | | | |

## **课题总结**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 课题环节 | | | 课题总结 | | |
| 建立日期 | **2023-07-31** | 最终日期 | **2023-08-06** | 版本次数 | 1 |
| 输入 | 1. 收集到的上海各区县历史人口、GDP、教育投入等数据。见附件“原始数据.xlsx”。 2. 研究日志。 3. 程序代码。 4. 相关参考资料。 | | | | |
| 输出 | 1. 经过整理的研究日志。 2. 课题报告。 | | | | |
| 过程记录 | 1. 按照要求整理研究日志。其中报告各环节各操作的用时分布图。 2. 按照要求组织研究报告。 | | | | |

## 