2022 数维杯大学生数学建模竞赛论文

# 论文题目： 阿尔茨海默病识别优化模型

队伍编号：20190822

火灾等级评价与快速救援措施优化模型

**摘 要：**

本文讨论了火灾等级评价与快速救援措施优化问题,运用熵权法和 K-MEANS 算法,建立了熵权法火灾等级评价模型, 确定了所给火灾事件的危害等级和重点放火位置，并根据筛选出的重点放火位置给出了消防人员救援分配优化措施。

针对问题一, 本文利用熵权法建立了火灾事件危害性综合评级模型，用 R 软件编程求解得到了 2033 个火灾事件的危害性得分，并从中筛选出了 10 大火灾编号：880、1055、1610、1730、1480、1081、292、1631、689、683。之后运用上述熵权法得到的权重，通过加权欧式距离的 K-MEANS 聚类算法将所有火灾事件的危害等级分为 4 个等级，分别为特别重大火灾，重大火灾，较大火灾和一般火灾。

针对问题二, 本文在火灾发生位置和火灾事件危害等级的基础上综合考虑了各个省市的火灾发生概率和重大级以上火灾发生概率两个因素，利用熵权法给予各个指标以权重，对 每个位置发生火灾的风险性进行评分，得到了所有位置的火灾风险性得分，通过设置风险性 阈值筛选出 516 个重要放火点位置（详见材料中表三）。

针对问题三, 本文建立了以救援所造成经济损失最小为目标函数的线性规划模型. 利用R 软件求解, 得到了将救援人员合理分配至 516 个重点放火位置进行救援工作的人员分配方案（详见材料中表四）。

### **关键词：**火灾危害等级；熵权法；K-MEANS 算法；最优化问题

一、问题的重述

* 1. 研究背景及意义

阿尔茨海默病（AD），俗称老年痴呆症，它是一种起病隐匿的慢性神经退行性疾病，且患者大多数为60岁以上的老人。据统计，全国有700多万人患有AD,65岁以上老年人发病率在5.6%，而85岁以上的老人发病率高达20%。是世界上AD患病人口最多、增长速度最快的地区，给我国的患者、家庭、社会和医疗带来沉重负担。根据《神经病学》杂志中的数据，一年有超过500000例患者因阿尔茨海默病（AD）死亡，阿尔茨海默病所引起的脑细胞大量死亡是不可逆的，因此需要严密预防。

因此对大脑结构特征和认知行为特征进行评定，用于准确诊断阿尔茨海默病，一边进行辅助治疗等工作具有重要意义。

由于老年人的性别（男、女），年龄（50-90）、婚姻状况（是否离婚）呈现复杂状况。本论文依据大赛所给数据和其他相关数据，对附件中阿尔茨海默病类型诊断，并揭示不同类别疾病随时间的演变，旨在对患者的早期干预和诊断标准

### 问题的提出

本文将要解决以下几个问题：

(1)对所附数据的特征指标进行预处理，调查数据特征与阿尔茨海默病诊断之间的相关性。 (2)利用所附的大脑结构特征和认知行为特征来设计一种阿尔茨海默病的智能诊断 。

(3)首先，将CN、MCI和AD聚为三大类。然后，对于MCI中包含的三个子类（SMC、 EMCI和LMCI），聚类继续细化为三个子类。

(4)附件中相同的样本包含了在不同时间点收集的特征，请分析它们与时间点的关 系，以揭示不同类别疾病随时间的演变模式。

(5)请查阅相关文献，描述CN、SMC、EMCI、LMCI、AD五类患者的早期干预和诊断标准

二、问题的分析

对本文提出的 5 个问题，我们逐一做如下分析：

问题一的分析：问题一要求数据的特征指标进行预处理。剔除全为空的特征列， 此外还需要对表格所给的特征利用所给文档解释资料进行筛选，剔除冗余特征， 然后用皮尔逊相关系数进行相关性计算

问题二的分析： 利用第一题通过相关性分析得到的特征进行分类，通过训练集数据来建立XGBoost回归模型并计算特征重要性。将XGBoost回归模型应用到训练、测试数据，得到模型评估结果。由于XGBoost具有随机性，每次运算的结果不一样，若保存本次训练模型，后续可以直接上传数据代入到本次训练模型进行计算预测。

问题三的分析：影响消防员分配方案的因素包括初始火灾面积、火灾蔓延速度、预

期灾后经济损失状况、可调配的消防人员数量、灭火速度、灭火过程所耗费的时间等。结合 问题要求，如人员数量、人员分配合理等条件，可将该问题转化为一项人员分配的最优化问 题，建立线性规划模型进行分析。

## 三、模型假设

针对本文提出的问题，我们做了如下模型假设：

* + 1. 火势蔓延速度是连续的，且不受天气变化等的影响；
    2. 消防人员整体素质一致，灭火速度相对稳定；
    3. 附件 2 所给救援难度系数包含火灾点环境、救援设备调配、人员机动等对救援工作产生的影响；

针对问题三做如下假设：

1. 516 处火灾点同时发生火灾；
2. 救援人员均为同时调派，且调派时间与前往救援地点时间忽略不计；
3. 不考虑同一救援人员可依次被分到不同火灾点救援的可能（如某一救援人员完成某一火灾点救援任务后前往另一火灾点参加救援），则救援总人数不超过给定的 100000 人；
4. 不考虑初始灭火时间和完成灭火时间，只考虑救援过程中的消耗时间；

## 四、建模与求解

### 问题一的建模与求解

问题一要求对所附数据的特征指标进行预处理，并调查数据特征与阿尔茨海默病诊断之间的相关性。本文首先对 AGE、APOE4、CDRSB\_bl、ADAS11\_bl、ADASQ4\_bl、ADAS13\_bl这五个数据总体的各项统计指标进行整体描述分析，将所得到的结果进行正态分布校验，对数据进行Shapiro-Wilk检验，查看其显著性。接着皮尔逊相关系数假设检验， 最后用spss进行相关性分析。

#### 4.1.1数据的预处理

（1）数据处理

**A.**分析附件 中数据是否存在异常值、缺失值等影响模型建立结果的部分，发现数据未存在该类问题；

**B.**由于附件 2中“age”的数据缺失不便于填充，并且数据量庞大，所以对age缺

失的数据进行删除

**C.**对缺失值大于百分之五十的指标进行删除，并通过EM估计对缺失值小于百分之五十的指标进行填充。（详情见附录）

**D.**首先进行KMO和Bartlett的检验，判断是否可以进行主成分分析。通过主成分分析，选择出两组最主要的数据（d1,d2）。

1. 主成分分析

首先进行KMO和Bartlett的检验，判断是否可以进行主成分分析。

**A.**对于KMO值： 0.7-0.8之间一般适合，对于 Bartlett的检验，因为P小于0.05，拒绝原假设，则说明可以做主成分分析 。KMO检验的结果显示，KMO的值为0.771，同时，Bartlett球形检验的结果显示，显著性P值为0.000\*\*\*，水平上呈现显著性，拒绝原假设，各变量间具有相关性，主成分分析有效，程度为一般。

**KMO检验和Bartlett的检验**

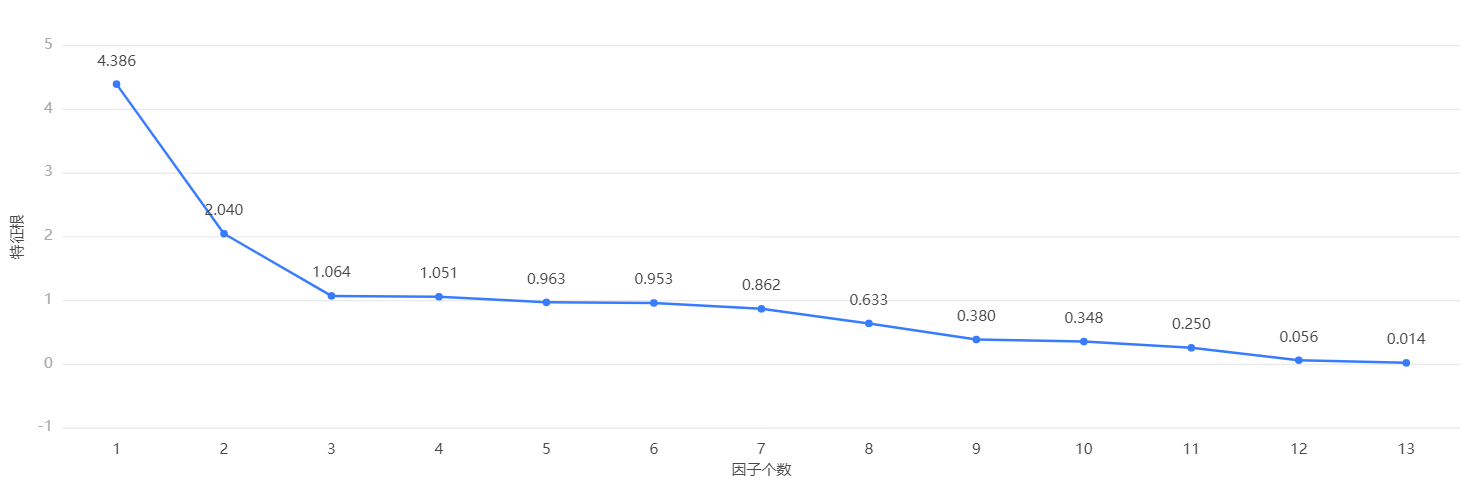
|  |  |
| --- | --- |
| KMO检验和Bartlett的检验 | |
| KMO值 | | 0.771 |
| Bartlett球形度检验 | 近似卡方 | 14905.513 |
| df | 78 |
| P | 0.000\*\*\* |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | |

**B.**通过分析方差解释表格和碎石图,主成分的数量方差解释表格主要是看主成分对于变量解释的贡献率。碎石图根据特征值下降的坡度来确认需要选择的主成分个数，这两者结合确认或调整主成分个数。根据分析可得在主成分5时，总方差解释的特征根低于1.0，变量解释的贡献率达到73.109 。

**方差解释表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 总方差解释 | | | |
| 成分 | 特征根 | | |
| 特征根 | 方差解释率(%) | 累积方差解释率(%) |
| 1 | 4.386 | 33.74 | 33.74 |
| 2 | 2.04 | 15.691 | 49.431 |
| 3 | 1.064 | 8.188 | 57.619 |
| 4 | 1.051 | 8.082 | 65.701 |
| 5 | 0.963 | 7.408 | 73.109 |
| 6 | 0.953 | 7.333 | 80.442 |
| 7 | 0.862 | 6.628 | 87.07 |
| 8 | 0.633 | 4.87 | 91.94 |
| 9 | 0.38 | 2.925 | 94.865 |
| 10 | 0.348 | 2.678 | 97.543 |
| 11 | 0.25 | 1.924 | 99.467 |
| 12 | 0.056 | 0.427 | 99.894 |
| 13 | 0.014 | 0.106 | 100 |

**碎石图**

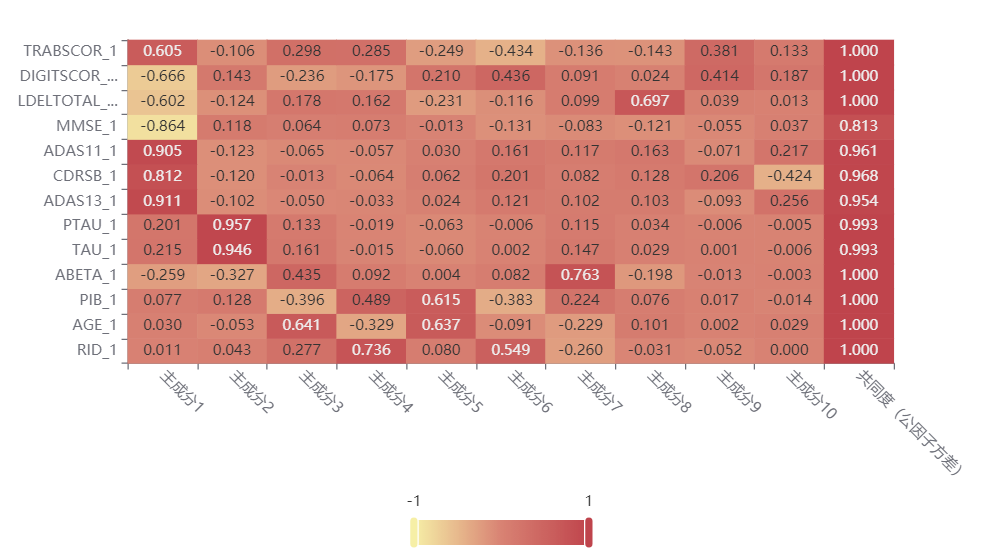


**C.**通过分析主成分载荷系数与热力图，可以分析到每个主成分中隐变量的重要性。可结合具体业务进行各主成分的隐变量分析。基于主成分载荷图通过将多主成分降维成双主成分或者三主成分。

**因子载荷系数表**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 因子载荷系数表 | | | | | | | | | | | |
|  | 因子载荷系数 | | | | | | | | | | 共同度（公因子方差） |
| 主成分1 | 主成分2 | 主成分3 | 主成分4 | 主成分5 | 主成分6 | 主成分7 | 主成分8 | 主成分9 | 主成分10 |
| RID\_1 | 0.011 | 0.043 | 0.277 | 0.736 | 0.08 | 0.549 | -0.26 | -0.031 | -0.052 | 0.000 | 1 |
| AGE\_1 | 0.03 | -0.053 | 0.641 | -0.329 | 0.637 | -0.091 | -0.229 | 0.101 | 0.002 | 0.029 | 1 |
| PIB\_1 | 0.077 | 0.128 | -0.396 | 0.489 | 0.615 | -0.383 | 0.224 | 0.076 | 0.017 | -0.014 | 1 |
| ABETA\_1 | -0.259 | -0.327 | 0.435 | 0.092 | 0.004 | 0.082 | 0.763 | -0.198 | -0.013 | -0.003 | 1 |
| TAU\_1 | 0.215 | 0.946 | 0.161 | -0.015 | -0.06 | 0.002 | 0.147 | 0.029 | 0.001 | -0.006 | 0.993 |
| PTAU\_1 | 0.201 | 0.957 | 0.133 | -0.019 | -0.063 | -0.006 | 0.115 | 0.034 | -0.006 | -0.005 | 0.993 |
| ADAS13\_1 | 0.911 | -0.102 | -0.05 | -0.033 | 0.024 | 0.121 | 0.102 | 0.103 | -0.093 | 0.256 | 0.954 |
| CDRSB\_1 | 0.812 | -0.12 | -0.013 | -0.064 | 0.062 | 0.201 | 0.082 | 0.128 | 0.206 | -0.424 | 0.968 |
| ADAS11\_1 | 0.905 | -0.123 | -0.065 | -0.057 | 0.03 | 0.161 | 0.117 | 0.163 | -0.071 | 0.217 | 0.961 |
| MMSE\_1 | -0.864 | 0.118 | 0.064 | 0.073 | -0.013 | -0.131 | -0.083 | -0.121 | -0.055 | 0.037 | 0.813 |
| LDELTOTAL\_1 | -0.602 | -0.124 | 0.178 | 0.162 | -0.231 | -0.116 | 0.099 | 0.697 | 0.039 | 0.013 | 1 |
| DIGITSCOR\_1 | -0.666 | 0.143 | -0.236 | -0.175 | 0.21 | 0.436 | 0.091 | 0.024 | 0.414 | 0.187 | 1 |
| TRABSCOR\_1 | 0.605 | -0.106 | 0.298 | 0.285 | -0.249 | -0.434 | -0.136 | -0.143 | 0.381 | 0.133 | 1 |

**因子载荷矩阵热力图**



**D.**通过象限图的方式呈现主成分的空间分布。通过分析成分矩阵，得出主成分成分公式与权重。

**成分矩阵表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 成分矩阵表 | | | | | | | | | | |
| 名称 | 成分 | | | | | | | | | |
| 成分1 | 成分2 | 成分3 | 成分4 | 成分5 | 成分6 | 成分7 | 成分8 | 成分9 | 成分10 |
| RID\_1 | 0.002 | 0.021 | 0.26 | 0.701 | 0.083 | 0.576 | -0.302 | -0.049 | -0.138 | 0 |
| AGE\_1 | 0.007 | -0.026 | 0.602 | -0.313 | 0.661 | -0.095 | -0.266 | 0.16 | 0.006 | 0.084 |
| PIB\_1 | 0.018 | 0.063 | -0.372 | 0.466 | 0.638 | -0.402 | 0.26 | 0.12 | 0.044 | -0.039 |
| ABETA\_1 | -0.059 | -0.16 | 0.408 | 0.088 | 0.004 | 0.086 | 0.886 | -0.312 | -0.033 | -0.01 |
| TAU\_1 | 0.049 | 0.464 | 0.151 | -0.014 | -0.063 | 0.002 | 0.17 | 0.046 | 0.002 | -0.018 |
| PTAU\_1 | 0.046 | 0.469 | 0.125 | -0.019 | -0.065 | -0.006 | 0.134 | 0.054 | -0.016 | -0.015 |
| ADAS13\_1 | 0.208 | -0.05 | -0.047 | -0.031 | 0.025 | 0.126 | 0.119 | 0.162 | -0.245 | 0.736 |
| CDRSB\_1 | 0.185 | -0.059 | -0.012 | -0.061 | 0.064 | 0.211 | 0.095 | 0.202 | 0.542 | -1.218 |
| ADAS11\_1 | 0.206 | -0.06 | -0.061 | -0.054 | 0.031 | 0.168 | 0.136 | 0.258 | -0.186 | 0.624 |
| MMSE\_1 | -0.197 | 0.058 | 0.06 | 0.069 | -0.014 | -0.138 | -0.096 | -0.191 | -0.144 | 0.108 |
| LDELTOTAL\_1 | -0.137 | -0.061 | 0.167 | 0.154 | -0.24 | -0.122 | 0.115 | 1.101 | 0.102 | 0.038 |
| DIGITSCOR\_1 | -0.152 | 0.07 | -0.222 | -0.167 | 0.218 | 0.457 | 0.105 | 0.038 | 1.089 | 0.537 |
| TRABSCOR\_1 | 0.138 | -0.052 | 0.28 | 0.271 | -0.259 | -0.456 | -0.157 | -0.225 | 1.002 | 0.383 |

**因子权重分析**

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 方差解释率(%) | 累积方差解释率(%) | 权重(%) |
| 主成分1 | 33.74 | 33.74 | 34.59 |
| 主成分2 | 15.691 | 49.431 | 16.086 |
| 主成分3 | 8.188 | 57.619 | 8.394 |
| 主成分4 | 8.082 | 65.701 | 8.286 |
| 主成分5 | 7.408 | 73.109 | 7.595 |
| 主成分6 | 7.333 | 80.442 | 7.518 |
| 主成分7 | 6.628 | 87.07 | 6.795 |
| 主成分8 | 4.87 | 91.94 | 4.993 |
| 主成分9 | 2.925 | 94.865 | 2.999 |
| 主成分10 | 2.678 | 97.543 | 2.746 |

#### 4.1.2描述性统计数据

首先，对首先对AGE、APOE4、CDRSB\_bl、ADAS11\_bl、ADASQ4\_bl、ADAS13\_bl这五个数据的各项统计指标进行整体描述分析。其次，对异常或者表现得较为突出的指标进行分析，例如高方差、高平均值等等。

#### 4.1.3算法介绍

描述性统计，是指运用制表和分类，图形以及计算概括性数据来描述数据特征的各项活动。描述性统计分析要对调查总体所有变量的有关数据进行统计性描述，主要包括数据的频数分析、[集中趋势分析](https://baike.baidu.com/item/%E9%9B%86%E4%B8%AD%E8%B6%8B%E5%8A%BF%E5%88%86%E6%9E%90/22600393?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%8F%8F%E8%BF%B0%E6%80%A7%E7%BB%9F%E8%AE%A1/_blank)、离散程度分析、分布以及一些基本的统计图形。

**第一部分数据输出结果：**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 样本量 | 最大值 | 最小值 | 平均值 | 标准差 | 中位数 | 方差 | 峰度 | 偏度 | 变异系数（CV） |
| AGE | 15898 | 91.4 | 54.4 | 73.313 | 6.98 | 73.4 | 48.717 | -0.326 | -0.137 | 0.09520500540450336 |
| APOE4 | 15898 | 2 | 0 | 0.517 | 0.648 | 0 | 0.419 | -0.329 | 0.874 | 1.2522732623650317 |
| CDRSB\_bl | 15898 | 10 | 0 | 1.219 | 1.516 | 0.5 | 2.299 | 3.719 | 1.783 | 1.2440662869477734 |
| ADAS11\_bl | 15898 | 42.67 | 0 | 9.269 | 5.782 | 8 | 33.435 | 2.615 | 1.363 | 0.6238008568094767 |
| ADAS13\_bl | 15898 | 54.67 | 0 | 14.67 | 8.627 | 13 | 74.426 | 0.988 | 0.996 | 0.5880912940177317 |
| ADASQ4\_bl | 15898 | 10 | 0 | 4.694 | 2.843 | 4 | 8.083 | -0.924 | 0.334 | 0.6056979238743139 |

d1

**d1数据分析**

基于AGE，变异系数（CV）为0.095，小于0.15，当前数据中较小概率出现异常值，采用平均值进行描述分析。基于APOE4，变异系数（CV）为1.252，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于CDRSB\_bl，变异系数（CV）为1.244，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADAS11\_bl，变异系数（CV）为0.624，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADAS13\_bl，变异系数（CV）为0.588，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADASQ4\_bl，变异系数（CV）为0.606，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 样本量 | 最大值 | 最小值 | 平均值 | 标准差 | 中位数 | 方差 | 峰度 | 偏度 | 变异系数（CV） |
| APOE4 | 15898 | 2 | 0 | 0.517 | 0.648 | 0 | 0.419 | -0.329 | 0.874 | 1.2522732623650317 |
| ADAS13\_1 | 15898 | 85 | 0 | 16.751 | 9.556 | 16.735 | 91.309 | 5.132 | 1.643 | 0.5704426271081451 |
| ADASQ4\_bl\_1 | 15898 | 10 | 0 | 4.694 | 2.843 | 4 | 8.083 | -0.924 | 0.334 | 0.6057042836658916 |
| ADAS11\_1 | 15898 | 70 | 0 | 10.889 | 7.019 | 10.875 | 49.266 | 9.922 | 2.368 | 0.6445885891629438 |

d2

**d2数据与分析**

基于APOE4，变异系数（CV）为1.252，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADAS13\_1，变异系数（CV）为0.57，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADASQ4\_bl\_1，变异系数（CV）为0.606，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。基于ADAS11\_1，变异系数（CV）为0.645，大于0.15，当前数据中可能存在异常值，对异常的或者表现得较为突出的指标进行分析。

**d1d2图表说明：**

上表展示了描述性统计的结果，包括样本量、最大值、最小值等统计量，用于研究定量数据的整体情况。

1. 分析各项统计指标，对各项统计指标进行整体描述分析。

2. 对异常的或者表现得较为突出的指标进行分析，例如高方差，高平均值等等。

#### 4..1.4正态分布校验

对数据进行Shapiro-Wilk（检验，查看其显著性。若不呈现出显著性(P>0.05)，说明符合正态分布，反之说明不符合正态分布（PS：通常现实研究情况下很难满足检验，若其样本峰度绝对值小于10并且偏度绝对值小于结合正态分布直方图、PP图或者QQ图可以描述为基本符合正态分布）。

**算法介绍**

Kolmogorov-Smirnov是比较一个频率分布f(x)与理论分布g(x)或者两个观测值分布的检验方法。其原假设H0:两个数据分布一致或者数据符合理论分布。D=max| f(x)- g(x)|，当实际观测值D>D(n,α)则拒绝H0，否则则接受H0假设。

KS检验与t-检验之类的其他方法不同是KS检验不需要知道数据的分布情况，可以算是一种非参数检验方法。当然这样方便的代价就是当检验的数据分布符合特定的分布事，KS检验的灵敏度没有相应的检验来的高。在样本量比较小的时候，KS检验最为非参数检验在分析两组数据之间是否不同时相当常用。

**输出结果1：总体描述结果**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 样本量 | 中位数 | 平均值 | 标准差 | 偏度 | 峰度 | S-W检验 | K-S检验 |
| AGE | 15898 | 73.4 | 73.313 | 6.98 | -0.137 | -0.326 | 0.995(0.000\*\*\*) | 0.025(6.3497089141911326e-9) |
| APOE4 | 15898 | 0 | 0.517 | 0.648 | 0.874 | -0.329 | 0.718(0.000\*\*\*) | 0.355(0) |
| CDRSB\_bl | 15898 | 0.5 | 1.219 | 1.516 | 1.783 | 3.719 | 0.786(0.000\*\*\*) | 0.211(0) |
| ADASQ4\_bl | 15898 | 4 | 4.694 | 2.843 | 0.334 | -0.924 | 0.945(0.000\*\*\*) | 0.131(7.991062033103057e-239) |
| ADAS13\_bl | 15898 | 13 | 14.67 | 8.627 | 0.996 | 0.988 | 0.936(0.000\*\*\*) | 0.103(6.970484541777451e-148) |
| ADAS11\_bl | 15898 | 8 | 9.269 | 5.782 | 1.363 | 2.615 | 0.905(0.000\*\*\*) | 0.115(4.69749764359162e-182) |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | | |

**d1**

**d1图表说明：**

上表展示了AGE、APOE4、CDRSB\_bl、ADASQ4\_bl、ADAS13\_bl、ADAS11\_bl描述性统计和正态性检验的结果，包括中位数、平均值等，用于检验数据的正态性。

1. 通常正态分布的检验方法有两种，一种是Shapiro-Wilk检验，适用于小样本资料（样本量≤5000）；另一种是Kolmogorov–Smirnov检验，适用于大样本资料（样本量>5000）。

2. 若呈现显著性(P<0.05)，则说明拒绝原假设（数据符合正态分布），该数据不满足正态分布，反之则说明该数据满足正态分布。

**d1图表分析：**

AGE样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（-0.326）绝对值小于10并且偏度（-0.137）绝对值小于3)APOE4样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（-0.329）绝对值小于10并且偏度（0.874）绝对值小于3)CDRSB\_bl样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（3.719）绝对值小于10并且偏度（1.783）绝对值小于3)ADASQ4\_bl样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（-0.924）绝对值小于10并且偏度（0.334）绝对值小于3)

d2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 样本量 | 中位数 | 平均值 | 标准差 | 偏度 | 峰度 | S-W检验 | K-S检验 |
| APOE4 | 15898 | 0 | 0.517 | 0.648 | 0.874 | -0.329 | 0.718(0.000\*\*\*) | 0.355(0) |
| ADASQ4\_bl\_1 | 15898 | 4 | 4.694 | 2.843 | 0.334 | -0.924 | 0.945(0.000\*\*\*) | 0.131(8.530282626280515e-239) |
| RID | 15898 | 3750.796 | 3750.973 | 1024.927 | -0.47 | 4.167 | 0.846(0.000\*\*\*) | 0.174(0) |
| ADAS13\_1 | 15898 | 16.735 | 16.751 | 9.556 | 1.643 | 5.132 | 0.868(0.000\*\*\*) | 0.224(0) |
| ADAS11\_1 | 15898 | 10.875 | 10.889 | 7.019 | 2.368 | 9.922 | 0.802(0.000\*\*\*) | 0.247(0) |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | | | |

**d2图表说明：**

上表展示了APOE4、ADASQ4\_bl\_1、RID、ADAS13\_1、ADAS11\_1描述性统计和正态性检验的结果，包括中位数、平均值等，用于检验数据的正态性。

1. 通常正态分布的检验方法有两种，一种是Shapiro-Wilk检验，适用于小样本资料（样本量≤5000）；另一种是Kolmogorov–Smirnov检验，适用于大样本资料（样本量>5000）。

2. 若呈现显著性(P<0.05)，则说明拒绝原假设（数据符合正态分布），该数据不满足正态分布，反之则说明该数据满足正态分布。

APOE4样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（-0.329）绝对值小于10并且偏度（0.874）绝对值小于3)ADASQ4\_bl\_1样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（-0.924）绝对值小于10并且偏度（0.334）绝对值小于3)RID样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（4.167）绝对值小于10并且偏度（-0.47）绝对值小于3)ADAS13\_1样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（5.132）绝对值小于10并且偏度（1.643）绝对值小于3)

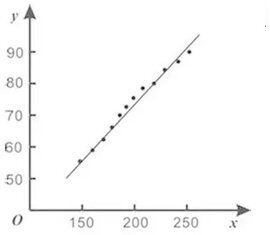
ADAS11\_1样本N >= 5000，采用K-S检验，显著性P值为0.000\*\*\*，水平呈现显著性，拒绝原假设，因此数据不满足正态分布。(其峰度（9.922）绝对值小于10并且偏度（2.368）绝对值小于3)

**4.1.5皮尔逊相关系数相关性分析**

先对XY之间是否存在统计上的显著关系(P<0.05)进行检验，分析相关系数为的正负向以及皮尔逊相关系数相关性程度，而后对分析结果进行总结。

**算法介绍**

皮尔逊积矩相关系数(又称作 PPMCC或PCCs)用于度量两个变量X和Y之间的相关(线性相关)，其值介于-1与1之间。该系数广泛用于度量两个变量之间的相关程度



1. **对第一部分数据输出结果1：**

**相关系数表**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | AGE | APOE4 | CDRSB\_bl | ADAS11\_bl | ADASQ4\_bl | ADAS13\_bl |
| AGE | 1(0.000\*\*\*) | -0.125(0.000\*\*\*) | 0.045(0.000\*\*\*) | 0.127(0.000\*\*\*) | 0.133(0.000\*\*\*) | 0.149(0.000\*\*\*) |
| APOE4 | -0.125(0.000\*\*\*) | 1(0.000\*\*\*) | 0.263(0.000\*\*\*) | 0.271(0.000\*\*\*) | 0.301(0.000\*\*\*) | 0.294(0.000\*\*\*) |
| CDRSB\_bl | 0.045(0.000\*\*\*) | 0.263(0.000\*\*\*) | 1(0.000\*\*\*) | 0.697(0.000\*\*\*) | 0.61(0.000\*\*\*) | 0.715(0.000\*\*\*) |
| ADAS11\_bl | 0.127(0.000\*\*\*) | 0.271(0.000\*\*\*) | 0.697(0.000\*\*\*) | 1(0.000\*\*\*) | 0.759(0.000\*\*\*) | 0.971(0.000\*\*\*) |
| ADASQ4\_bl | 0.133(0.000\*\*\*) | 0.301(0.000\*\*\*) | 0.61(0.000\*\*\*) | 0.759(0.000\*\*\*) | 1(0.000\*\*\*) | 0.878(0.000\*\*\*) |
| ADAS13\_bl | 0.149(0.000\*\*\*) | 0.294(0.000\*\*\*) | 0.715(0.000\*\*\*) | 0.971(0.000\*\*\*) | 0.878(0.000\*\*\*) | 1(0.000\*\*\*) |
| 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平 | | | | | | |
| d1 | | | | | | |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | APOE4 | ADAS11\_1 | ADAS13\_1 | ADASQ4\_bl\_1 | | APOE4 | 1(0.000\*\*\*) | 0.231(0.000\*\*\*) | 0.244(0.000\*\*\*) | 0.301(0.000\*\*\*) | | ADAS11\_1 | 0.231(0.000\*\*\*) | 1(0.000\*\*\*) | 0.957(0.000\*\*\*) | 0.551(0.000\*\*\*) | | ADAS13\_1 | 0.244(0.000\*\*\*) | 0.957(0.000\*\*\*) | 1(0.000\*\*\*) | 0.607(0.000\*\*\*) | | ADASQ4\_bl\_1 | 0.301(0.000\*\*\*) | 0.551(0.000\*\*\*) | 0.607(0.000\*\*\*) | 1(0.000\*\*\*) | | 注：\*\*\*、\*\*、\*分别代表1%、5%、10%的显著性水平  d2 | | | | | | | | | | | |

**图表说明：**

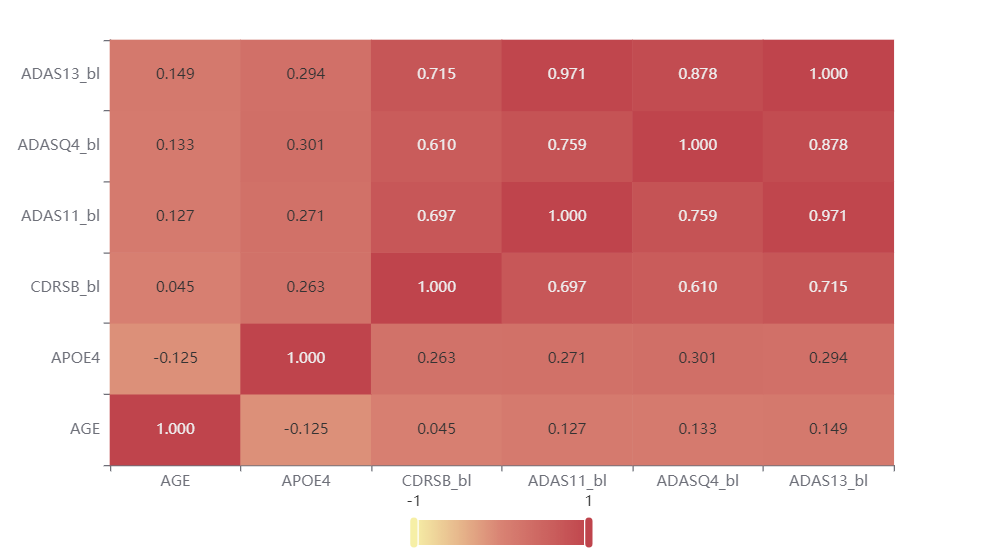
上表展示了模型检验的参数结果表，包括了相关系数、显著性P值。

1. 先对XY之间是否存在统计上的显著性关系进行检验，判断P值是否呈现显著性(P<0.05)。

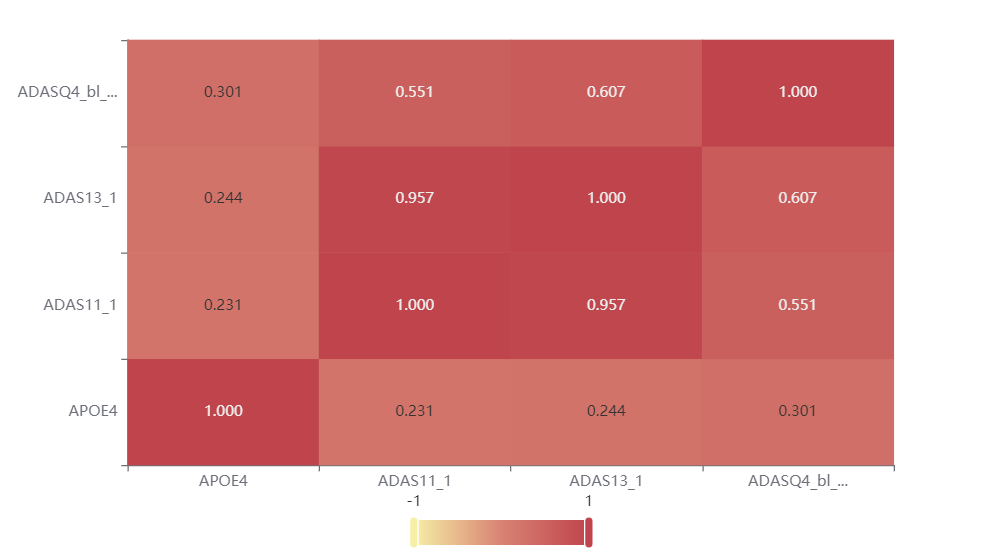
2. 若呈现显著性，则说明两变量之间存在相关性，反之，则两变量之间不存在相关性。

3. 分析相关系数的正负向以及相关性程度。

**输出结果2：相关系数热力图**



d1



d2

**图表说明：**

上图展示了热力图的形式展示了相关系数的值，主要通过颜色深浅去表示值的大小。

### 问题二的建模与求解

由所附的大脑结构特征和认知行为特征数据来设计阿尔茨海默病的智能诊断 。首先，利用第一题通过相关性分析得到的特征进行分类。通过训练集数据来建立XGBoost回归模型来计算特征重要性。将建立的XGBoost回归模型应用到训练、测试数据，得到模型评估结果。

#### 4.2.1通过训练集数据来建立XGBoost回归模型

#### 算法介绍

训练数据集和测试数据集，是机器学习领域，对于数据切分方式不同，而产生的2个概念。通常做法：将原始数据切分时，将原始数据的80%作为训练数据来训练模型，另外20%作为测试数据，通过测试数据直接判断模型的效果，在模型进入真实环境前不断改进模型。

**模型参数**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 1.62s |
| 数据切分 | 0.7 |
| 数据洗牌 | 是 |
| 交叉验证 | 5 |
| 基学习器 | gbtree |
| 基学习器数量 | 100 |
| 学习率 | 0.1 |
| L1正则项 | 0 |
| L2正则项 | 1 |
| 样本征采样率 | 1 |
| 树特征采样率 | 1 |
| 节点特征采样率 | 1 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |

**d1**

|  |  |
| --- | --- |
| 参数名 | 参数值 |
| 训练用时 | 1.62s |
| 数据切分 | 0.7 |
| 数据洗牌 | 是 |
| 交叉验证 | 5 |
| 基学习器 | gbtree |
| 基学习器数量 | 100 |
| 学习率 | 0.1 |
| L1正则项 | 0 |
| L2正则项 | 1 |
| 样本征采样率 | 1 |
| 树特征采样率 | 1 |
| 节点特征采样率 | 1 |
| 叶子节点中样本的最小权重 | 0 |
| 树的最大深度 | 10 |

**d2**

**图表说明：**

上表展示了模型各项参数配置以及模型训练时长。

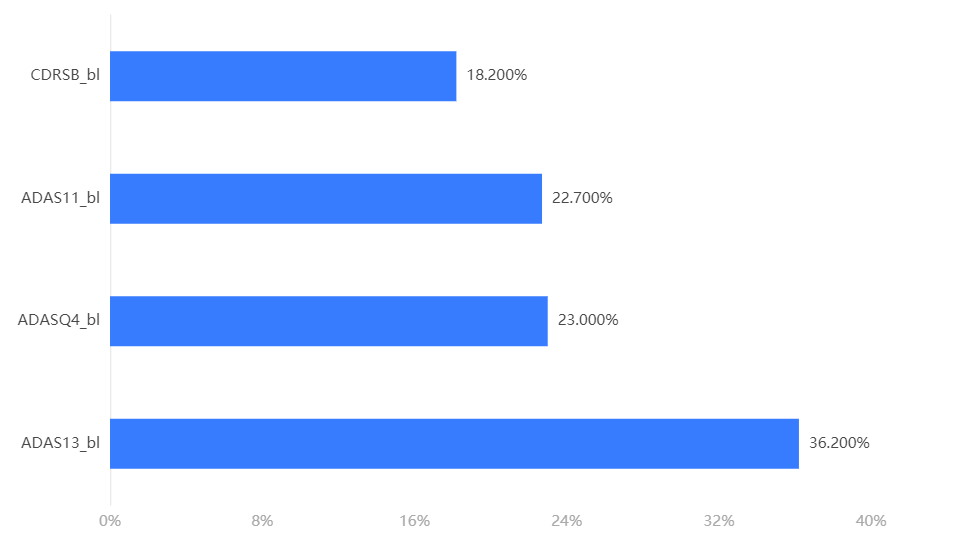
**4.2.2通过建立的XGBoost来计算特征重要性。**

**算法介绍**

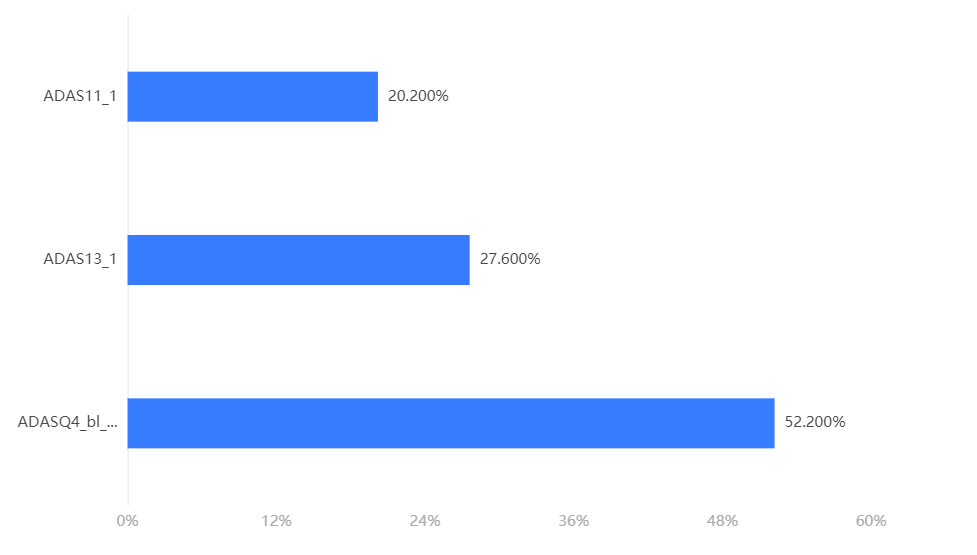
XGBoost是一个优化的分布式梯度增强库，旨在实现高效，灵活和便携。它在

Gradient Boosting 框架下实现机器学习算法。XGBoost提供并行树提升（也称为GBDT，GBM），XGBoost 是对梯度提升算法的改进，求解损失函数极值时使用了牛顿法，将损失函数泰勒展开到二阶，另外损失函数中加入了正则化项。训练时的目标函数由两部分构成，第一部分为梯度提升算法损失，第二部分为正则化项。

**特征重要性**



d1



d2

**图表说明：**

上柱形图或表格展示了各特征（自变量）的重要性比例。

**4.2.3建立的XGBoost回归模型，得到模型评估结果。**

输出结果：模型评估结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MSE | RMSE | MAE | MAPE | R² |
| 训练集 | 0.112 | 0.334 | 0.24 | 74.159 | 0.747 |
| 交叉验证集 | 0.56 | 0.748 | 0.599 | 473.881 | -0.267 |
| 测试集 | 0.528 | 0.727 | 0.579 | 252.859 | -0.253 |

**d1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MSE | RMSE | MAE | MAPE | R² |
| 训练集 | 0.27 | 0.519 | 0.423 | 105.588 | 0.334 |
| 交叉验证集 | 0.41 | 0.64 | 0.524 | 226.699 | -0.02 |
| 测试集 | 0.445 | 0.667 | 0.549 | 234.007 | 0.008 |

**d2**

**图表说明：**

上表中展示了交叉验证集、训练集和测试集的预测评价指标，通过量化指标来衡量XGBoost的预测效果。其中，通过交叉验证集的评价指标可以不断调整超参数，以得到可靠稳定的模型。

● MSE（均方误差）： 预测值与实际值之差平方的期望值。取值越小，模型准确度越高。

● RMSE（均方根误差）：为MSE的平方根，取值越小，模型准确度越高。

● MAE（平均绝对误差）： 绝对误差的平均值，能反映预测值误差的实际情况。取值越小，模型准确度越高。

● MAPE（平均绝对百分比误差）： 是 MAE 的变形，它是一个百分比值。取值越小，模型准确度越高。

● R²： 将预测值跟只使用均值的情况下相比，结果越靠近 1 模型准确度越高。

**4.2.4
由于XGBoost具有随机性，上传数据进行计算预测。**

测试数据预测结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 预测结果Y | APOE4 | CDRSB\_bl | ADAS11\_bl | ADASQ4\_bl | ADAS13\_bl |
| 0.865118145942688 | 1 | 6 | 33 | 10 | 45 |
| 0.29343461990356445 | 0 | 0 | 4 | 3 | 7 |
| 0.6995174288749695 | 0 | 0 | 5 | 5 | 11 |
| 0.08185886591672897 | 0 | 0.5 | 3.33 | 2 | 5.33 |
| 0.07391554117202759 | 1 | 1 | 4 | 5 | 9 |
| 0.985584020614624 | 1 | 0 | 10.67 | 3 | 14.67 |
| 0.24743741750717163 | 0 | 0 | 6.67 | 2 | 8.67 |
| 0.7881326675415039 | 1 | 3 | 20 | 10 | 31 |
| 1.104787826538086 | 0 | 2 | 13 | 5 | 18 |
| 0.7722784280776978 | 2 | 1.5 | 2.67 | 1 | 3.67 |
| 0.8947102427482605 | 0 | 0 | 5 | 5 | 10 |
| 0.866416335105896 | 0 | 1.5 | 12.33 | 5 | 20.33 |
| 0.12063263356685638 | 0 | 1 | 6 | 0 | 6 |
| 0.22346942126750946 | 0 | 0 | 6.33 | 5 | 11.33 |
| 0.8627852201461792 | 0 | 3 | 13 | 10 | 24 |

**d1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 预测结果Y | APOE4 | ADAS11\_1 | ADAS13\_1 | ADASQ4\_bl\_1 |
| 0.30308660864830017 | 1 | 5 | 10 | 4 |
| 0.2373480200767517 | 1 | 5 | 9 | 4 |
| 0.36702221632003784 | 1 | 10.8754835581416 | 16.7348150124511 | 4 |
| 0.26444828510284424 | 1 | 6 | 12 | 4 |
| 0.3735928535461426 | 1 | 4 | 7 | 4 |
| 0.36702221632003784 | 1 | 10.8754835581416 | 16.7348150124511 | 4 |
| 0.31770047545433044 | 1 | 3 | 7 | 4 |
| 0.02031506411731243 | 1 | 1 | 1 | 2 |
| 0.2688102126121521 | 1 | 4 | 6 | 2 |
| 0.3052016794681549 | 1 | 10.8754835581416 | 16.7348150124511 | 2 |
| 0.22048260271549225 | 1 | 2 | 2 | 2 |
| 0.3938906788825989 | 1 | 4 | 4 | 2 |
| 0.3052016794681549 | 1 | 10.8754835581416 | 16.7348150124511 | 2 |
| 0.338725745677948 | 1 | 3 | 5 | 2 |
| 0.3052016794681549 | 1 | 10.8754835581416 | 16.7348150124511 | 2 |

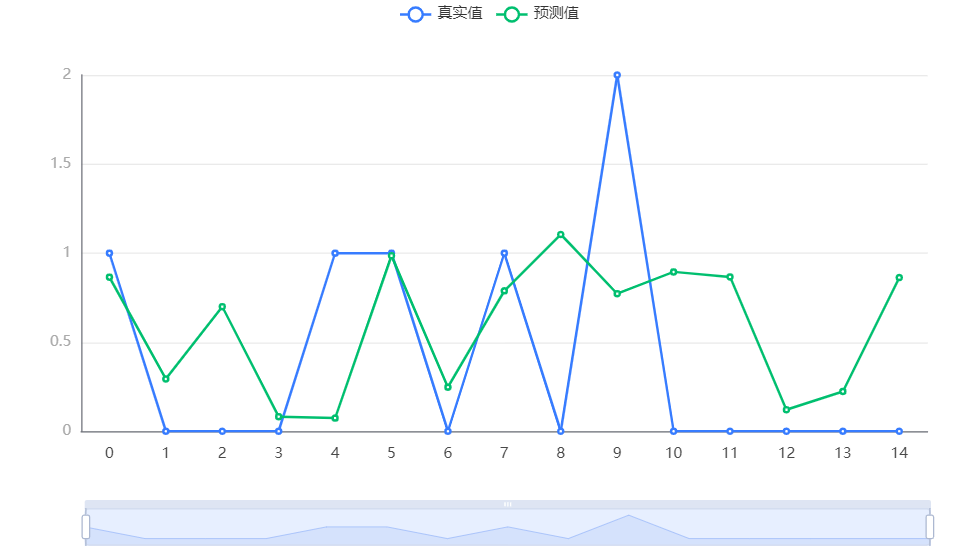
**d2**

**图表说明：**

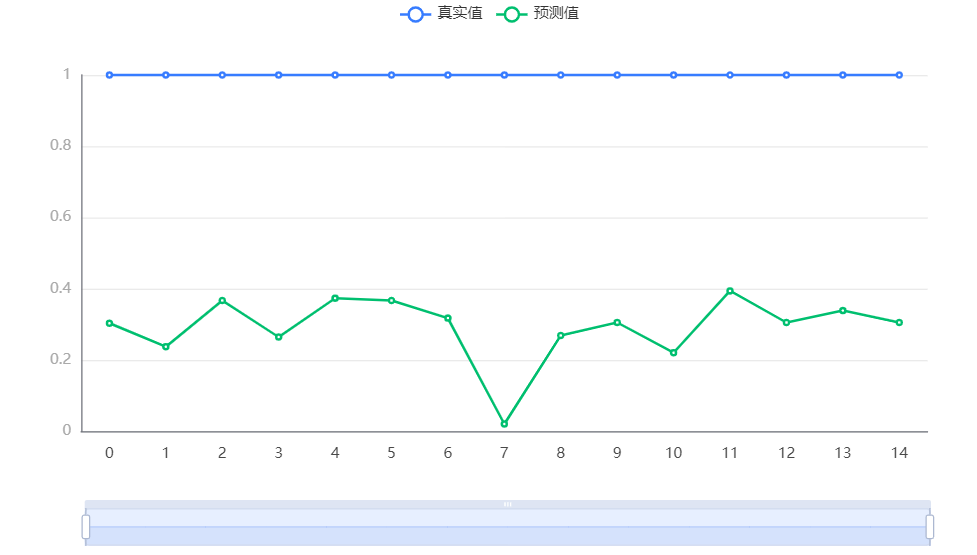
上表格为预览结果，只显示部分数据，全部数据请点击下载按钮导出。

上表展示了XGBoost对测试数据的预测情况。

测试数据预测图



d1



d2

**图表说明：**

图表说明：上图中展示了XGBoost对测试数据的预测情况。

## 问题三的建模与求解

首先对所有的数据进行量纲统一化处理，把量纲化后的数据进行聚类分析，在聚类算法中是要使用 K-MEANS 聚类算法，计算欧式距离，并根据最小距离重新对MCI中包含的三个子类（SMC、 EMCI和LMCI），相应对象细化为三个子类，重新计算每个聚类的中心对象，直到每个聚类不再发生变化为止。

*x*1*i*  *x*1 *j*

2

* *x*2*i*  *x*2 *j*  *x*3*i*  *x*3 *j*  *x*4*i*  *x*4 *j*

2

2

2

### 对所有数据进行量纲化处理

* + - 1. 算法介绍

量纲化的目的是将数据进行量纲单位统一化，有的量纲化具有实际意义，比如最小值化，最大值化，均值化，标准化等；分别代表数据除以平均值，数据除以第1个数，数据除以最小值，数据除以最大值，数据除以求和值，数据除以平方和值，标准化后数据的平均值为0标准差为1。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **描述统计** | | | | | |
|  | N | 最小值 | 最大值 | 均值 | 标准 偏差 |
| RID\_1 | 1284 | 7 | 7092 | 3982.20 | 2368.936 |
| AGE\_1 | 1284 | 54.4 | 90.1 | 73.928 | 7.3609 |
| ABETA\_1 | 1284 | 210.90 | 1681.00 | 754.5026 | 161.23296 |
| ADAS13\_1 | 1284 | .00 | 72.00 | 20.2839 | 9.03696 |
| CDRSB\_1 | 1284 | .0 | 16.0 | 2.307 | 1.9910 |
| ADAS11\_1 | 1284 | .00 | 57.00 | 12.7530 | 6.78729 |
| MMSE\_1 | 1284 | 8.0 | 30.0 | 26.032 | 3.4100 |
| LDELTOTAL\_1 | 1284 | .0 | 20.0 | 5.483 | 3.8563 |
| DIGITSCOR\_1 | 1284 | .0 | 70.0 | 36.374 | 11.8738 |
| TRABSCOR\_1 | 1284 | .0 | 300.0 | 135.034 | 78.3502 |
| 有效个案数（成列） | 1284 |  |  |  |  |

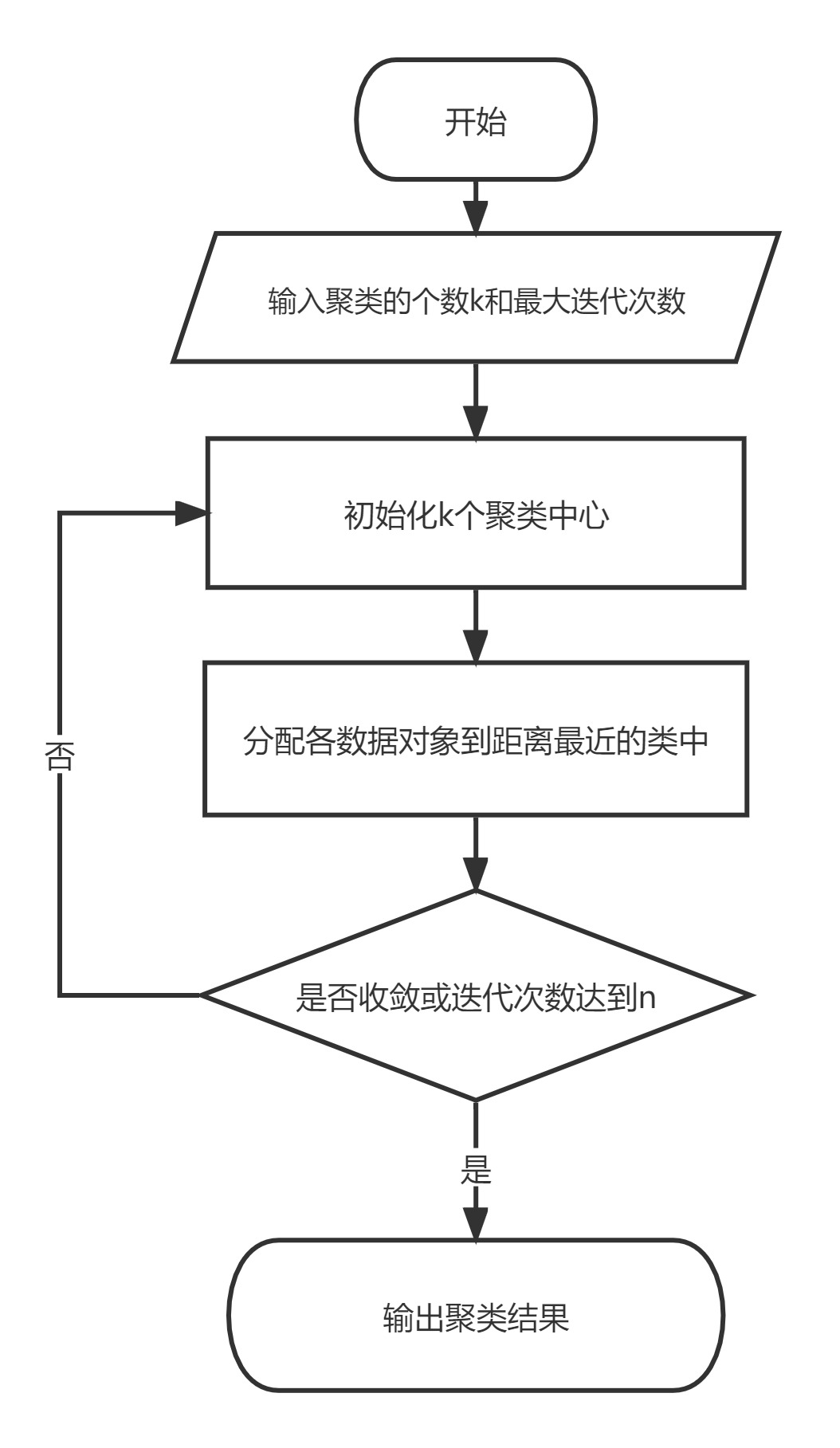
### 量纲化后的**K-means算法**聚类分析

1.随机数选取一个样本作为第一个初始点

2.计算每个样本与当前已有聚类中心的最短距离，值越大，表示被选取作为聚类中心的概率中心的概率越大，最后用轮盘法，选出下一个聚类中心

3.重复2，直到选出k个聚类中心，选出初始点，继续使用k-means算法。

4.随机地选择K个数据对象作为初始的聚类中心(不一定要是我们的样本点) ;、  
5.计算其余的各个数据对象到这K个初始聚类中心的距离，把数据对象划归到距离它最近的那个中心所处在的簇类中;  
6.调整新类并且重新计算出新类的中心;  
7.循环步骤三和四，看中心是否收敛(不变)，如果收敛或达到迭代次数则停止循环。



1. Means聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| **初始聚类中心** | | | |
|  | 聚类 | | |
| LMCI | SMC | EMCI |
| Zscore(RID\_1) | .43513 | .48452 | .94296 |
| Zscore(AGE\_1) | .77054 | .09128 | .39015 |
| Zscore(PIB\_1) | 11.47127 | -.03702 | -4.42108 |
| Zscore(ABETA\_1) | .07087 | -1.34776 | 2.75066 |
| Zscore(TAU\_1) | -.05611 | 8.18241 | -1.75524 |
| Zscore(PTAU\_1) | -.05809 | 9.68032 | -1.78830 |
| Zscore(ADAS13\_1) | 1.29646 | -.17859 | .33707 |
| Zscore(CDRSB\_1) | .09701 | -.65638 | .85039 |
| Zscore(ADAS11\_1) | .77306 | -.15956 | .08501 |
| Zscore(MMSE\_1) | -.30260 | -.30260 | .87042 |
| Zscore(LDELTOTAL\_1) | -1.42170 | -1.42170 | -.90307 |
| Zscore(DIGITSCOR\_1) | 1.65292 | .22119 | .47385 |
| Zscore(TRABSCOR\_1) | -.57478 | -.19189 | -.10254 |

|  |  |  |  |
| --- | --- | --- | --- |
| **最终聚类中心** | | | |
|  | 聚类 | | |
| LMCI | SMC | EMCI |
| RID\_1 | 1023 | 6553 | 4607 |
| AGE\_1 | 74.4 | 73.0 | 74.4 |
| PIB\_1 | 1.8423 | 1.8454 | 1.8526 |
| ABETA\_1 | 746.75 | 752.77 | 764.88 |
| TAU\_1 | 300.68 | 308.49 | 300.23 |
| PTAU\_1 | 29.58 | 30.36 | 29.53 |
| ADAS13\_1 | 19.85 | 21.20 | 19.82 |
| CDRSB\_1 | 2.1 | 2.5 | 2.3 |
| ADAS11\_1 | 12.52 | 13.28 | 12.47 |
| MMSE\_1 | 26.0 | 25.8 | 26.3 |
| LDELTOTAL\_1 | 5.4 | 5.5 | 5.6 |
| DIGITSCOR\_1 | 36.8 | 35.9 | 36.4 |
| TRABSCOR\_1 | 129.5 | 142.2 | 133.7 |

利用加权欧式距离的 K .均值聚类法针对附件中数据进行聚类，分为了LMCI，SMC，EMCI三类。表4 给出聚类结果。

|  |  |  |
| --- | --- | --- |
| **每个聚类中的个案数目** | | |
| 聚类 | LMCI | 453.000 |
| SMC | 422.000 |
| EMCI | 409.000 |
| 有效 | | 1284.000 |
| 缺失 | | .000 |