中国研究生创新实践系列大赛

“华为杯”第二十届中国研究生数学建模竞赛

学　　校	西安理工大学
参赛队号	23107000046
队员姓名	1. 刘进瑄
	2. 杨倩
	3. 郭永强

中国研究生创新实践系列大赛
“华为杯”第二十届中国研究生
数学建模竞赛

题　目：　出血性脑卒中临床智能诊疗建模

摘　要：

出血性脑卒中是中老年人的常见病、多发病，致残率和致死率都非常高，研究出血性脑卒中临床智能诊疗建模具有重大意义。本文采用信息增益、随机森林、神经网络、有序多分类 logistic 回归模型等数学算法，对数据进行了深度挖掘，发现了影响血肿及水肿体积变化的相关因素，探索出了治疗干预和患者预后预测的因素。具体做法如下：

针对问题一，本文采用数据整合和处理的方法来预测患者是否发生血肿扩张。首先，对表 1 和表 2 的数据进行整合，并计算出每次随访到发病的时间。然后，通过判断血肿体积的增加条件，确定患者是否发生血肿扩张，并记录血肿扩张发生的时间。为了降低数据维度，采用了信息增益方法对表 1 的数据进行降维，并采用 Lasso 回归模型和皮尔森相关系数法对表 2 的数据进行降维，选取了共 19 个特征，如年龄、血压、水肿体积、血肿体积等。最后，通过神经网络模型对所有患者的血肿扩张发生概率进行预测，并使用测试数据集评估模型的准确率，结果表明模型的准确率达到了 82%。

针对问题二，本文采用曲线拟合、聚类分析和关联分析等方法，探索了水肿体积的变化以及不同治疗方法对水肿体积的影响。在子问题 a 中，将表 2 中前 100 个患者的水肿体积和重复检查时间点数据进行整合，并通过启发式分析选择伽马分布作为合适的拟合模型，取绝对值后的残差和为 119864.72。在子问题 b 中，对表 2 的数据进行整合和归一化处理，使用 K-means 聚类算法将患者分为 4 类，并对每类患者绘制水肿体积随时间的进展散点图，并采用对勾函数、反比例函数、线性函数和多项式函数进行曲线拟合，取绝对值后的残差和依次为 41330.8、50970.3、 26450.80， 13650.1。在子问题 c 中，根据不同治疗方法对数据进行整合，使用皮尔逊相关系数分析各治疗方法对水肿体积进展的影响，并绘制折线图观察不同治疗方法对水肿体积进展模式的影响。在子问题 d 中，利用 Apriori 算法分析数据，探索血肿体积、水肿体积和治疗方法之间的关系，并分析哪些治疗方案能够降低血肿体积和水肿体积。其中的一条结果为【止血治疗；降颅压治疗；降压治疗；镇静、镇痛治疗；止吐护胃】，会使水肿、血肿体积减少。

针对问题三，本文通过不同的预测模型和相关性分析，研究了预后 mRS 评分与相关因素之间的关联性。在子问题 a 中，将表 1、表 2 和表 3 的相关数据整合，并使用随机森林、决策树和神经网络等预测模型对所有患者的 90 天 mRS 评分进行预测。结果表明，随机森林模型的预测准确率最高，达到了 88%。在子问题 b 中，对表 2 的数据进行降维处理，并与表 1 和表 3 的主要特征进行整合。使用随机森林预测模型对含影像检查的患者进行 90 天 mRS 评分的预测，准确率达到了 89%。在子问题 c 中，使用信息增益方法对表 1 的数据进行降维，并探究了年龄等特征与预后 mRS 评分的关联关系。同时使用有序多分类 logistic 回归模型

进行分析，发现年龄、预前 mRS 评分、水肿体积和血肿体积对预后 mRS 评分有显著影响，并提出了相应的对策和建议。
关键词：出血性脑卒中；信息增益；神经网络；随机森林；Kmeans 聚类；Apriori 算法；有序多分类 logistic 回归模型

目 录

一、问题重述

1.1 背景介绍

出血性脑卒中是中老年人的常见病，多发病，致残率和致死率都极高，即使进行手术治疗，预后的情况也依然较差。有数据表明，对病情发展到重型出血性脑卒中的 114 例病例，观察外科手术治疗的结果发现，痊愈 41 例，轻残 18 例，重残 9 例，植物生存 3 例，死亡 43 例[1] 。可以看出约 64%的出血性脑卒中患者会遗留比较严重的生理功能障碍，同时会给社会和家庭带来很严重的经济负担。因此，根据患者的发病风险，影像特征等相关信息，得出患者的治疗方案并及时优化治疗方案，从而预测患者预后的身体状况，根据此信息对优化临床决策具有非常重要的意义。

在我国[2]，近年来对出血性脑卒中的临床治疗方案和护理都有了明显的改善，但是患者的功能预后不良比例仍然较高，并且脑卒中病死率是最高的一类。影响出血性脑卒中发病的因素很多，其中包括年龄、血压、疾病史、用药情况、患者脑出血前 mRS 评分等。在出血性脑卒中发生短时间内，血肿范围会因为脑组织受损，病情反应等因素逐渐变大，导致颅内压增加和神经功能进一步恶化，甚至会影响患者的生命。此外，血肿周围的水肿扩张也会导致脑组织受压，加重患者的神经功能损伤。因此，对血肿扩张和血肿周围水肿的发生，进行早期诊断和及时治疗对提升患者的生活质量具有重要意义。

近年来，人工智能在计算机领域快速发展，医学成像过程中产生了许多复杂高维的图像信息，因此采用医学影像技术进行相关数据处理非常合适且准确率较高。在临床实践中，患者的影像数据在诊疗和随访过程中发挥非常重要的作用[3]。希望能够基于本赛题提供的影像信息，患者个人信息、治疗方案和预后等数据，构建智能诊疗模型，明确导致出血性脑卒中预后不良的危险因素，实现精准的疗效评估和预后预测。相信在不久的将来，相关研究成果及科学依据将能够进一步应用于临床实践，为改善出血性脑卒中患者临床和预后作出贡献。

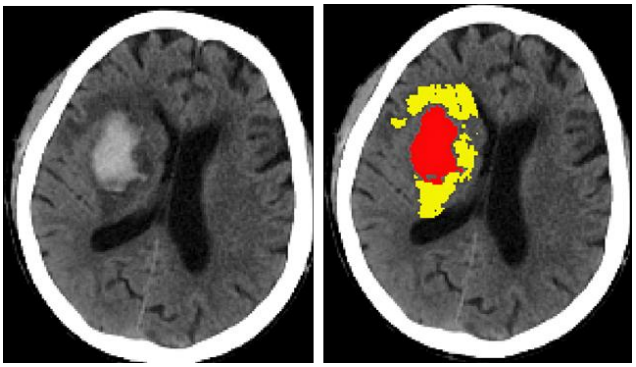

图 1. 左图脑出血患者 CT 平扫，右图红色为血肿，黄色为血肿周围水肿

1.2 问题重述

出血性脑卒中病因复杂，且会引发一系列复杂的生理病理反应。因此，发掘

出血性脑卒中的发病风险，整合影像学特征及临床诊疗方案，精准预测患者预后，并据此优化临床决策具有重要的临床意义。

问题一：血肿扩张风险相关因素探索建模。

1.根据“表 1”（字段：入院首次影像检查流水号，发病到首次影像检查时间间隔），“表 2”（字段：各时间点流水号及对应的 HM_volume），判断患者 sub001 至 sub100 发病后 48 小时内是否发生血肿扩张事件。如发生血肿扩张事件，请同时记录血肿扩张发生时间。

2.以是否发生血肿扩张事件为目标变量，基于“表 1” 前 100 例患者（sub001 至 sub100）的个人史，疾病史，发病相关、“表 2”中其影像检查结果及“表 3”其影像检查结果等变量，构建模型预测所有患者（sub001 至 sub160）发生血肿扩张的概率。

问题二：血肿周围水肿的发生及进展建模，并探索治疗干预和水肿进展的关联关系。

1.根据“表 2”前 100 个患者（sub001 至 sub100）的水肿体积和重复检查时间点，构建一条全体患者水肿体积随时间进展曲线（x 轴：发病至影像检查时间，y 轴：水肿体积，y=f(x)），计算前 100 个患者（sub001 至 sub100）真实值和所拟合曲线之间存在的残差。

2.探索患者水肿体积随时间进展模式的个体差异，构建不同人群（分亚组：3-5 个）的水肿体积随时间进展曲线，并计算前 100 个患者（sub001 至 sub100）真实值和曲线间的残差。

3.请分析不同治疗方法（“表 1”）对水肿体积进展模式的影响。

4.请分析血肿体积、水肿体积及治疗方法（“表 1”）三者之间的关系。

问题三：出血性脑卒中患者预后预测及关键因素探索。

1.请根据前 100 个患者（sub001 至 sub100）个人史、疾病史、发病相关（“表 1”字段 E 至 W）及首次影像结果（表 2，表 3 中相关字段）构建预测模型，预测患者（sub001 至 sub160）90 天 mRS 评分。

2.根据前 100 个患者（sub001 至 sub100）所有已知临床、治疗（“表 1”字段 E 到 W）、表 2 及表 3 的影像（首次+随访）结果，预测所有含随访影像检查的患者（sub001 至 sub100，sub131 至 sub160）90 天 mRS 评分。

3.请分析出血性脑卒中患者的预后（90 天 mRS）和个人史、疾病史、治疗方法及影像特征（包括血肿/水肿体积、血肿/水肿位置、信号强度特征、形状特征）等关联关系，为临床相关决策提出建议。

二、问题分析

2.1 对问题一的分析

（a）按照题目要求，对“表 1”、“表 2”进行数据整合，对后续随访的血肿体积与首次检查的血肿体积进行数据处理，并计算出每次随访到发病的时间，若在发病后 48 小时内后续检查比首次检查的绝对体积增加≥6 mL 或相对体积增加

≥33%，则认为该患者发生血肿扩张，反之，则不然。对于发生血肿扩张的患者，记录其扩张的时间。

（b）根据题意，需要将“表 1—表 3”所涉及的相关变量进行汇总，从而预测出所有患者发生血肿扩张的概率。因为表中变量数据复杂且繁多，所以在求解问题之前，需要对所有变量进行降维处理，找到影响血肿扩张的最主要的几个特征。为了提高选取特征的准确性，将使用不同的方法对不同的表进行降维处理，并提取每张表的主要特征，最后构建预测模型预测所有患者发生血肿扩张的概率。

2.2 对问题二的分析

（a）由题意，需要从表中筛选出前 100 名患者的水肿体积和重复检查时间点，并绘制出前 100 名患者真实值的散点图，根据观察散点图的特点，启发式的寻找函数进行曲线拟合，并计算前 100 名患者真实值和所拟合曲线之间存在的残差。

（b）根据题目要求，对“表 2”中患者、时间、水肿体积数据进行整合，并对数据采用归一化处理，使用 K-means 聚类方法对全体患者进行聚类，从而识别出具有相似水肿体积随时间进展模式的患者群体。绘制不同人群的水肿体积随时间进展的散点图，接着对每个亚组使用多项式回归拟合，拟合出类似于 $y = f(x)$ 的水肿体积随时间进展的曲线，并计算真实值和预测值之间的残差。

（c）根据题意，将表中的不同治疗方法和水肿体积数据进行汇总，使用皮尔逊相关系数法进行建模作图；后续将治疗方法相同的归为一类，计算随访 1 到随访 7，水肿体积的平均增长率，对不同治疗方法的水肿体积进展绘制折线图，观察不同治疗方法对水肿体积进展模式的影响。

（d）按照题目要求，将表中血肿体积、水肿体积及治疗方法的数据进行整合，探索其三者之间的关系，故可以采用 Apriori 算法对数据进行分析处理，Apriori 算法是一种经典的频繁项集挖掘算法，可以在大量数据中找到经常一起出现的数据集合，发现数据之间的内在联系，从而找到数据的关联规则。即可以探索出血肿体积、水肿体积及治疗方法三者之间的关系。

2.3 对问题三的分析

（a）题目要求根据前 100 名患者个人史、疾病史、发病相关及首次影像结果构建预测模型，数据处理过程与问题 1 的子问题 b)类似，对数据进行降维处理，找到影响 90 天 mRS 评分的最主要的几个特征。最后构建预测模型预测患者 90 天 mRS 评分，可以使用多种预测模型进行对比，选取最佳方案。

（b）题目要求根据前 100 名患者所有临床、治疗影像（首次+随访）结果，预测所有含随访影像检查的患者 90 天 mRS 评分。与（a）的数据相比多了后续随访的治疗影像，故先对“表 2”首次+随访的数据进行降维处理，选取“表 2”中的主要特征，与“表 1”，“表 3”的主要特征放在一起作为训练集进行预测。可以使用多种预测模型进行对比，选取最佳方案。

（c）题目要求分析出血性脑卒中患者的预后（90 天 mRS）和个人史、疾病史、治疗方法及影像特征等之间的关系，对表中相关数据进行整合。对“表 1”中的数据重新进行降维处理，选取“表 1”中的主要特征，与“表 2”，“表 3”的主要特征放在一起作为训练集。因为 mRS 的 6 个数据特征是离散的，对于留下的特征可以利用有序多分类 logistic 回归模型分析出血性脑卒中患者预后和个人史、疾病史、治疗方法及影像特征等关联关系，分析处理后的结果，为临床相关

决策提出建议。

三、模型假设

1.假设附件提供的所有数据均为真实数据；
2.假定数据样本的各个变量值都在规定的数据范围之内；
3.假设数据中个别缺失数据对结果不会产生重大影响；
4.假设除了附件中的变量，没有其他变量会对出血性脑卒中患者产生影响；

四、符号说明

符号的相关含义解释在文中均有注明，以下展示部分符号变量：

符号	意义
ε	残差
a	事件发生的次数
β	事件发生一次的概率
m	数据样本的个数
$\overline{X_i}$	样本水肿体积变化的平均值（i表示患者，$i=1,2,...,160$）
σ_1	样本标准差（i表示患者，$i=1,2,...,160$）
R^2	最小二乘法判定系数
Y	信息熵计算的集合
N	类别数目
p_i	类别i在子集的占比
p	概率
HMV_{ij}	第i号患者第j次检查后的血肿体积
$\Delta HMAV_{i(j-1)}$	血肿的绝对体积
$\Delta HMRV_{i(j-1)}$	血肿的相对体积
$j=1,\cdots,l_i$	第i号患者的总随访次数
Acc	模型的准确率
MSE	均方误差
$RMSE$	均方根误差
MAE	平均绝对误差

五、问题一模型建立与求解

5.1 子问题 a)的分析与求解

5.1.1 子问题 a)的分析

问题一的子问题 a)中，需要判断患者 sub001 至 sub100 发病后 48 小时内是否发生血肿扩张事件，若血肿扩张，则记录血肿扩张时间。

首先，从“表 1”中提取每个患者的入院首次影像检查流水号，根据流水号在“附表 1”中查找对应的患者数据，并计算每次随访到发病的时间，时间数据只保留发病后 48 小时内的信息；然后，在“表 2”中筛选出首次检查和每次随访的血肿体积，对后续随访的血肿体积与首次检查的血肿体积进行数据处理，并与“表 1”处理后的数据结合，若在发病后 48 小时内后续检查比首次检查的绝对体积增加≥6 mL 或相对体积增加≥33%，则认为该患者发生血肿扩张，思路流程图如图 2 所示。对于发生的血肿扩张事件，记录其扩张的时间。

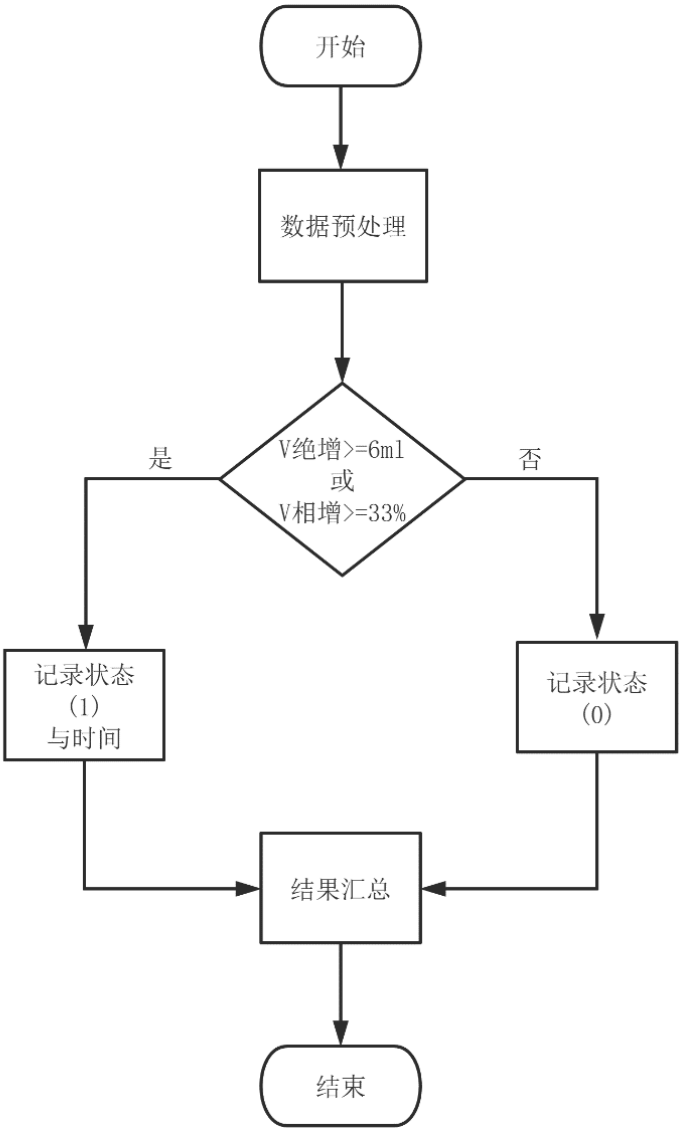

图 2 问题 1 子问题 a)的思路流程图

5.2.1 子问题 a)的模型建立与求解

$$HMV_{ij}, i=1,2,\cdots,100, j=0,\cdots,l_i$$

HMV_{ij} 表示第 i 号患者第 j 次检查后的血肿体积(HM_volume)，当 $j=0$ 时，表示患者进行首次检查。

HM of Absolutely Volume 即血肿的绝对体积:

$$\Delta HMAV_{i(j-1)} = \frac{HMV_{ij} - HMV_{i0}}{1000}, j=1,\cdots,l_i \tag{1}$$

HM of Relatively Volume 即血肿的相对体积:

$$\Delta HMRV_{i(j-1)} = \frac{\Delta HMAV_{i(j-1)}}{HMV_{i0}} \tag{2}$$

其中，$i=1,2,\cdots,100, j=1,\cdots,l_i$，$l_i$ 表示第 i 号患者的总随访次数。

发生血肿扩张，需满足的条件：

$$\Delta HMAV_{i(j-1)} \geq 6ml \text{ 或} \Delta HMRV_{i(j-1)} \geq 33\%, j=1,\cdots,l_i$$

结果表明，有 23 名患者会发生血肿扩张，77 名患者不会发生血肿扩张。部分结果如表 1 所示，全部结果见“表 4-答案文件”。

表 1 部分患者的数据结果展示

患者	是否发生血肿扩张	血肿扩张时间	患者	是否发生血肿扩张	血肿扩张时间
sub001	0	0.00	sub021	0	0.00
sub002	0	0.00	sub022	0	0.00
sub003	1	39.60	sub023	0	0.00
sub004	0	0.00	sub024	0	0.00
sub005	1	26.47	sub025	0	0.00
sub006	0	0.00	sub026	0	0.00
sub007	0	0.00	sub027	0	0.00
sub008	0	0.00	sub028	0	0.00
sub009	1	40.06	sub029	0	0.00
sub010	0	0.00	sub030	0	0.00
sub011	0	0.00	sub031	0	0.00
sub012	0	0.00	sub032	0	0.00
sub013	0	0.00	sub033	1	30.81
sub014	0	0.00	sub034	0	0.00
sub015	0	0.00	sub035	0	0.00
sub016	0	0.00	sub036	1	39.50
sub017	1	14.87	sub037	0	0.00
sub018	0	0.00	sub038	1	15.81
sub019	0	0.00	sub039	1	42.50
sub020	0	0.00	sub040	0	0.00

注：1：发生血肿扩张， 0：不发生血肿扩张；血肿扩张时间单位：小时.

5.2 子问题 b)的分析与求解

5.2.1 问题分析

在子问题 b）中，考虑到数据维数比较大，提出了三种特征选择的方法，对数据进行标准化，使用信息增益的方法对“表 1”进行降维；使用 Lasso 回归模型和皮尔森相关系数对“表 2”进行降维；求方差对“表 3”进行降维，找到影响血肿扩张最主要的几个特征，使用神经网络算法预测所有患者发生血肿扩张的概率。

5.2.2 数据预处理

特征选择的目的是过滤掉携带信息量较少的特征，只保留对分类贡献较大的特征。评价函数的好坏是影响特征选择的关键，常用的评价函数有信息增益、χ^2 统计、文档频率等[4]。

1、用信息增益的方法对“表 1”进行降维处理

（1）信息增益

信息增益表示得知某种特征 X 的信息使得目标信息 Y 的不确定性减少的程度，即目标信息 Y 的熵与属性 X 的条件信息熵的差值。

信息熵的计算公式如下:

$$\mathrm{Ent}(Y) = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{3}$$

其中，条件信息熵的计算公式如（4）所示：

$$\mathrm{Ent}(Y|X) = \sum_i p_i \,\mathrm{Ent}\ \ Y|X = X_i\ \ , p_i = p\ \ X = X_i\ \ , i = 1,2,3,\ldots,n \tag{4}$$

P 代表概率，Y 表示进行信息熵计算的集合，N 表示类别数目，p_i 表示类别 i 在子集的占比。

因此，可以根据信息增益性质得出公式（5）:

$$\mathrm{Gain}(Y, X) = \mathrm{Ent}(Y) - \mathrm{Ent}(Y|X) \tag{5}$$

即：信息增益=信息熵-条件信息熵。

（2）数据降维处理结果

处理数据时，记 x_1：年龄，x_2：性别，x_3：脑出血前 mRS 评分，x_4：高血压病史，x_5：卒中病史，x_6：糖尿病史，x_7：房颤史，x_8：冠心病史，x_9：吸烟史，x_{10}：饮酒史，x_{11}：发病到首次影像检查时间间隔，x_{12}：血压，x_{13}：脑室引流 ，x_{14}：止血治疗，x_{15}：降颅压治疗，x_{16}：降压治疗，x_{17}：镇静、镇痛治疗，x_{18}：止吐护胃，x_{19}：营养神经。

年龄记 1：0-20 岁,2：20-40 岁,3：40-60 岁,4：60-80 岁,5：80-100 岁。发病到首次影像检查时间间隔的均值为 3.28，向上取整为 4。

将血压划分为低、正常、高、严重高四个离散区间，具体划分方式如下：

1)低血压：收缩压<90 或舒张压<60；

2)正常血压：90≤收缩压<120 且 60≤舒张压<80；

3)高血压：120≤收缩压<140 且 80≤舒张压<90；

4)严重高血压：收缩压≥140 或舒张压≥90。

通过对“表 1”采用信息增益的方法降维后得出，影响血肿扩张的主要因素有年龄、脑出血前 mRS 评分、饮酒史、发病到首次影像检查时间间隔、血压、降颅压治疗、降压治疗，其中年龄、血压对血肿扩张的影响因素较大，具体结果如图 3 所示。

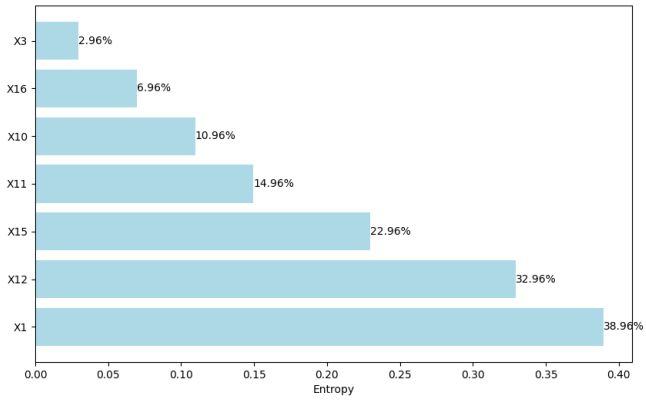

图 3 “表 1”中数据降维处理结果

注：X_1：年龄；X_3：脑出血前 mRS 评分；X_{10}：饮酒史；X_{11}：发病到首次影像检查时间间隔；X_{12}：血压；X_{15}：降颅压治疗；X_{16}：降压治疗.

2、使用 Lasso 回归模型和皮尔逊相关系数对“表 2”进行降维处理

（1）皮尔逊相关系数

皮尔逊相关系数：用来反映两个变量 X 和 Y 的线性相关程度，$r\in[-1,1]$。计算公式如下：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \tag{6}$$

即可以看作是两个变量 X，Y 的协方差乘积和两者标准差乘积的比值。

当 Pearson 相关系数为正时，表示两个变量呈正相关关系；当 Pearson 相关系数为负时，表示两个变量呈负相关关系；当 Pearson 相关系数接近于 0 时，表示两个变量之间的线性相关性较弱或几乎没有线性关系。

（2）Lasso 回归模型

Lasso 的表达式

$$J(\theta)=\frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{k}|w_j|\right] \tag{7}$$

其中，m 为样本个数，k 为参数个数，$\lambda\sum_{j=1}^{k}|w_j|$ 为 L1 正则化，参数 λ 是正则化系数，$\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2$ 为均方误差。

（3）数据降维处理结果

对“表 2”数据进行降维处理，首先，通过观察“表 2”数据的特征，将特征中 0 超过 60%的数据删除，不进行考虑，因此，留下 10 个特征。使用皮尔逊相关系数方法计算 10 个特征的相关系数，留下相关性强的特征。如图 4 所示，

从图中可以看出 HM_MCA_R_Ratio、ED_MCA_R_Ratio、ED_volume、HM_volume、HM_PCA_R_Ratio、ED_PCA_R_Ratio 的相关性较强。

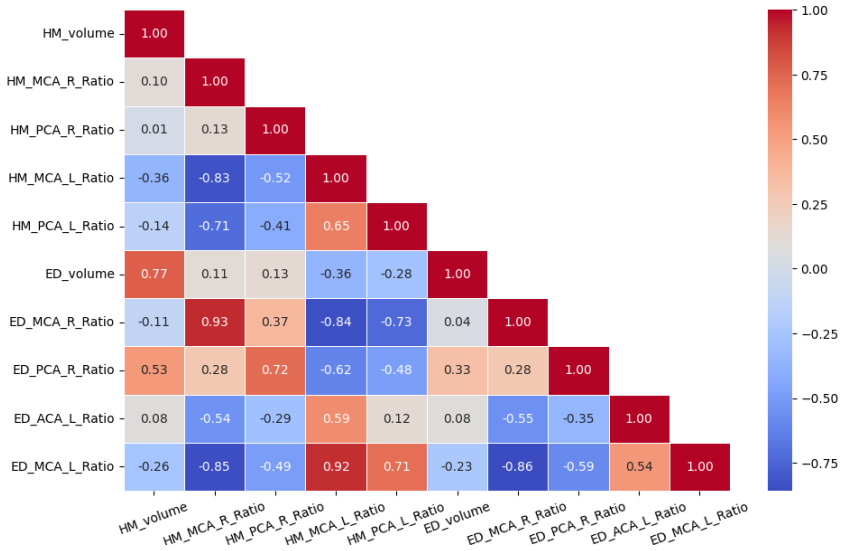

图 4 皮尔逊相关系数对数据的处理结果

为了使筛选的结果更可靠，同时使用 Lasso 回归模型和皮尔逊相关系数的结果进行对比，最终留下来的特征分别是 ED_volume、HM_volume。

3、使用方差对“表 3”中特征进行降维

因为“表 3”中的数据相互独立且数值较大，因此先对数据进行归一化，然后对每个特征求方差，方差越小说明数据较稳定，方差大则表明波动性比较强，因此，提取出方差较大的特征来作为主要特征。

对血肿、水肿数据分别进行处理后，各自筛选出 5 个方差较大的特征作为主要特征，具体结果如表 2、3 所示

表 2 血肿数据降维筛选后的主要特征

特征	方差
NCCT_original_firstorder_Entropy	0.061692
original_shape_Sphericity	0.058041
NCCT_original_firstorder_Uniformity	0.052042
NCCT_original_firstorder_Median	0.051612
original_shape_Elongation	0.048307

表 3 水肿数据降维筛选后的主要特征

特征	方差
NCCT_original_firstorder_InterquartileRange	0.064309
NCCT_original_firstorder_RobustMeanAbsoluteDeviation	0.062308
original_shape_Elongation	0.054433
original_shape_Flatness	0.054191
original_shape_Maximum2DDiameterSlice	0.047542

5.2.3 基于 BP 神经网络模型的建立与求解

（1）BP 神经网络模型

BP 神经网络[5]是按照误差逆向传播算法训练的多层前馈神经网络，其基本思想是梯度下降法。其主要特点是：信号将前向传播，误差会反向传播。基本的 BP 神经网络主要由三层组成，分别为输入层、隐含层和输出层。层与层之间采用的是全连接方式，同层的神经元之间没有任何连接。

每个隐含层和输出层的神经元，输出与输入的函数关系为：

$$I_j = \sum_i H_{ij} K_i \tag{8}$$

$$K_j = sigmod(I_s) = \frac{1}{1+e^{-I_s}} \tag{9}$$

其中，H_{ij} 表示神经元 i 与神经元 j 之间连接的权重，K_j 表示神经元 j 的输出，sigmod 是用于将任何实数映射到 $(0,1)$ 区间的函数。

计算输出层的误差：

$$E_j = sigmod'(K_j) * (Q_j - K_j) = K_j(1-K_j)(Q_j - K_j) \tag{10}$$

其中，E_j 表示神经元 j 的误差，K_j 表示神经元 j 的输出，Q_j 表示当前训练样本的参考输出，$sigmod'(x)$ 是一阶导数。

更新 H_{ij}：

$$H_{ij} = H_{ij} + \lambda E_j K_i \tag{11}$$

其中，λ 是学习率参数，一般 $\lambda \in (0, 0.1)$。

损失函数：

$$L = \sum_j (Q_j - K_j)^2 \tag{12}$$

其中，Q_j 表示当前训练样本的参考输出，K_j 表示神经元 j 的输出。

（2）模型求解

问题 1 子问题 b）的思路流程图如图 5 所示：

对“表 1，表 2，表 3”分别降维后，对选取的主要特征进行汇总。使用 BP 神经网络模型预测所有患者（sub001-sub160）发生血肿扩张的概率。其中训练集为患者 sub001-sub100 的主要特征及是否发生血肿扩张；测试集为患者 sub001-sub160 的主要特征。

具体步骤如下：

1）对数据进行归一化、数据标准化处理；

2）网络架构设计：设计神经网络的结构，其中包括输入层、隐藏层和输出层的数量和节点数；

3）初始化网络参数：随机初始化网络的权重和偏差；

4）前向传播：将输入数据通过神经网络进行前向传播，计算每个节点的输出；

5）计算损失：通过比较网络的输出和实际值，计算损失函数；

6）参数更新：使用优化算法更新网络的参数。通过迭代多次进行前向传播、计算损失和参数更新，逐步减小损失函数，使网络的预测结果更接近实际值；

7）模型评估：使用测试数据集评估训练好的神经网络模型的性能；

8）模型调优：根据模型评估结果，对网络架构和训练参数进行调整，以进

一步提高模型的性能；

9）预测：使用训练好的神经网络模型对新的输入数据进行预测。

使用神经网络预测模型，预测出所有患者发生血肿扩张的概率。部分结果如表 4 所示，全部结果见“表 4-答案文件”。

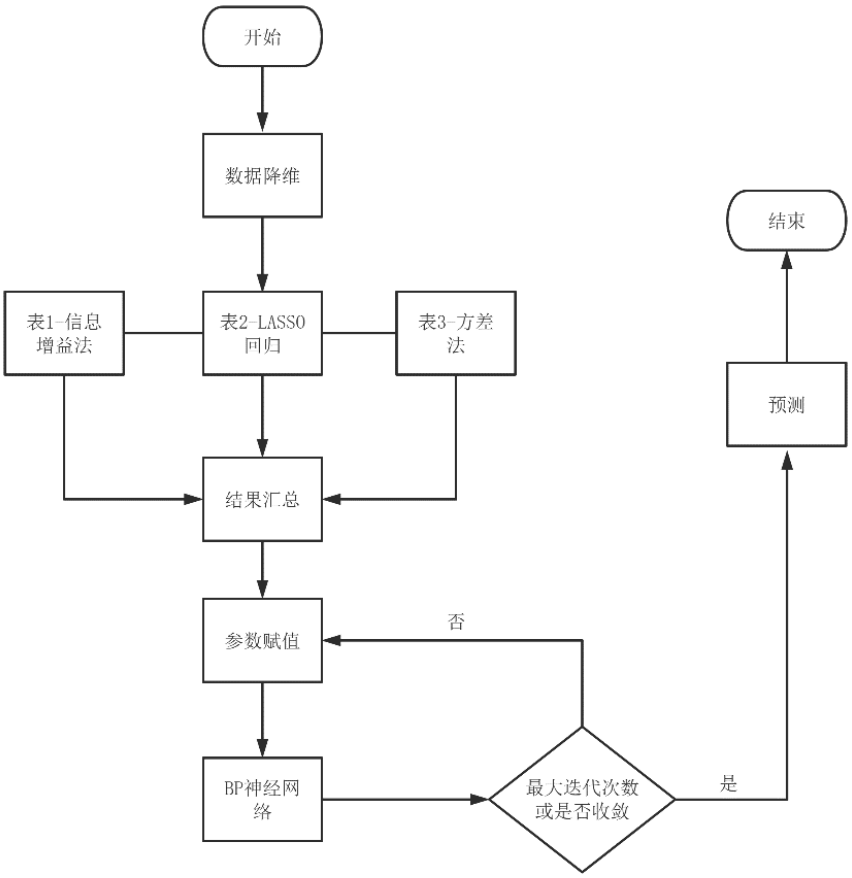

图 5 问题 1 子问题 b）的思路流程图

表 4 部分患者发生血肿扩张的概率

患者	发生血肿扩张的概率	患者	发生血肿扩张的概率
sub001	0.0921	sub021	0.1081
sub002	0.1468	sub022	0.0948
sub003	0.1658	sub023	0.4743
sub004	0.6131	sub024	0.1164
sub005	0.1150	sub025	0.1993
sub006	0.6118	sub026	0.6068
sub007	0.6746	sub027	0.1273
sub008	0.1154	sub028	0.1751
sub009	0.2505	sub029	0.2741
sub010	0.2350	sub030	0.1631
sub011	0.0890	sub031	0.0330
sub012	0.0929	sub032	0.4291
sub013	0.1992	sub033	0.0718

sub014	0.1192	sub034	0.6039
sub015	0.1026	sub035	0.0970
sub016	0.0647	sub036	0.4823
sub017	0.0867	sub037	0.0620
sub018	0.1561	sub038	0.4878
sub019	0.0644	sub039	0.0552
sub020	0.1031	sub040	0.0200

通过使用测试数据集评估训练好的神经网络模型的性能发现，模型评估的准确率达到 82%，所以认为 BP 神经网络通过处理所选出的主要特征，来预测所有患者发生血肿扩张的概率是可行的。

六、问题二模型建立与求解

6.1 子问题 a)的分析与求解

6.1.1 问题分析

将“表 2”中前 100 个患者的水肿体积和重复检查时间点数据 excel 进行整合，绘制出前 100 名患者真实值的散点图，启发式的寻找模型进行曲线拟合，根据观察散点图的特点，考虑到了卡方分布，高斯分布，多元回归等方法，最终确定使用伽马分布对散点图进行拟合，然后计算前 100 个患者真实值和所拟合曲线之间存在的残差，所得残差数据说明拟合的曲线效果较好，数据见“表 4”。

问题 2 子问题 a）的思路流程图如图 6 所示：

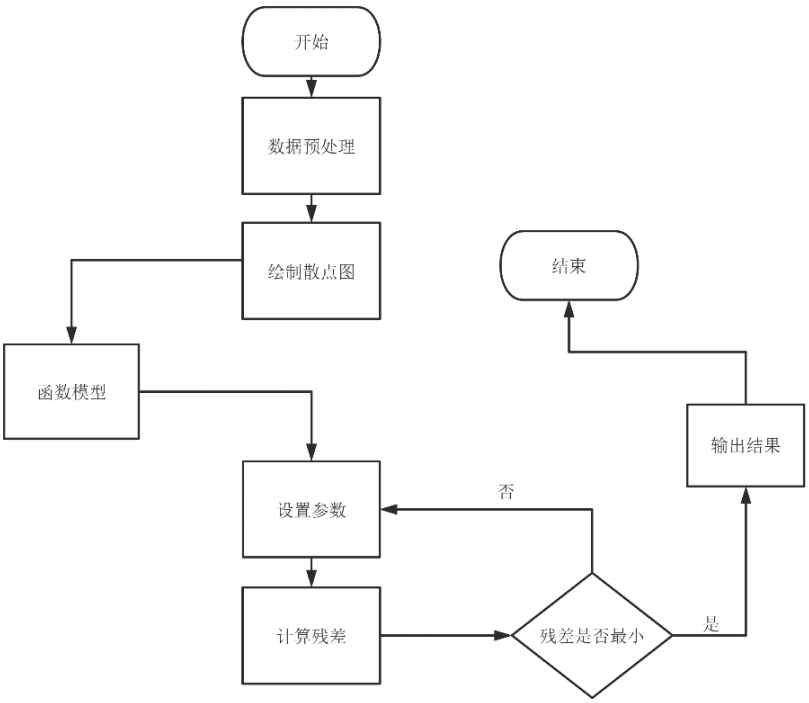

图 6 问题 2 子问题 a）的思路流程图

6.1.2 基于伽马分布的曲线拟合

（1）伽马分布：假设随机变量 x 为等到第 α 件事发生所需的等候时间，且每个事件之间的等待时间是互相独立的，α 为事件发生的次数，β 代表事件发生一次的概率，那么这 α 个事件的时间之和服从伽马分布，其概率密度函数为

$$f(x,\beta,\alpha)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x},x>0$$

（2）残差 (ε)=真实值-预测值

将“表 2”中前 100 个患者的水肿体积和重复检查时间点数据进行整合，绘制出患者真实值的散点图。使用伽马分布对散点图进行拟合，拟合的结果如图 7 所示。

拟合曲线为

$$\begin{aligned}y&=(a+(1+x)b)\cdot f(1-x,k,\theta)+c\\&=(a+(1+x)b)\cdot\frac{k^{\theta}}{\Gamma(\theta)}(x-1)^{\theta-1}e^{-k(x-1)}+c.\end{aligned}$$

其中 $a=22.700268652145077$，$b=-22.429994169505775$，
$c=0.0733000615020143$，$\theta=0.24886892329765503$，$k=9.225559098597493$.

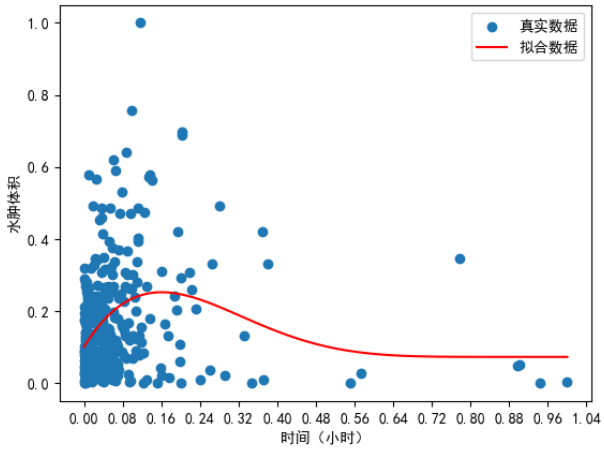

图 7 前 100 名患者真实值和拟合曲线图

拟合完之后，将前 100 名患者的真实值和拟合的数据整合到 excel 中，使用 excel 计算残差，观察残差值，发现曲线拟合的效果好，结果见“表 4-答案文件”。

6.2 子问题 b)的分析与求解

6.2.1 问题分析

在子问题 b 中，探索患者水肿体积随时间进展模式的个体差异，构建不同人群的水肿体积随时间进展曲线，首先对“表 2”中患者、时间、水肿体积数据进行整合，对数据归一化处理，使用 K-means 聚类对全体患者进行聚类，从而识别出具有相似水肿体积随时间进展模式的患者群体。

题中要求不同人群分亚组 3-5 个，本文将亚组分别聚成 3，4，5 类，用轮廓系数对 k 值的选取进行评估，其中 $k=4$ 时，轮廓系数最接近于 1，聚类效果最好。因此聚为 4 类，将聚类的结果分别记为 0,1,2,3。然后对 4 个类（亚组）绘制水肿

体积随时间的进展散点图，对散点图进行拟合，拟合出类似于 $y=f(x)$ 的水肿体积随时间进展的曲线，计算真实值和预测值之间的残差（见“表 4-答案文件”）。

观察 4 类患者的特征，观察 0 为患者有可能患有高血压病史，糖尿病史，冠心病史，均无卒中病史，吸烟史和饮酒史。1 为患者大概率可能患有高血压病史，吸烟史，均无卒中病史。2 为患者同时患有高血压病史和卒中病史，均无冠心病史，吸烟史，饮酒史，3 为均患有高血压病史和卒中病史，大概率还附带患有糖尿病史、房颤史，冠心病史别的病史。

问题 2 子问题 b）的思路流程图如图 8 所示：

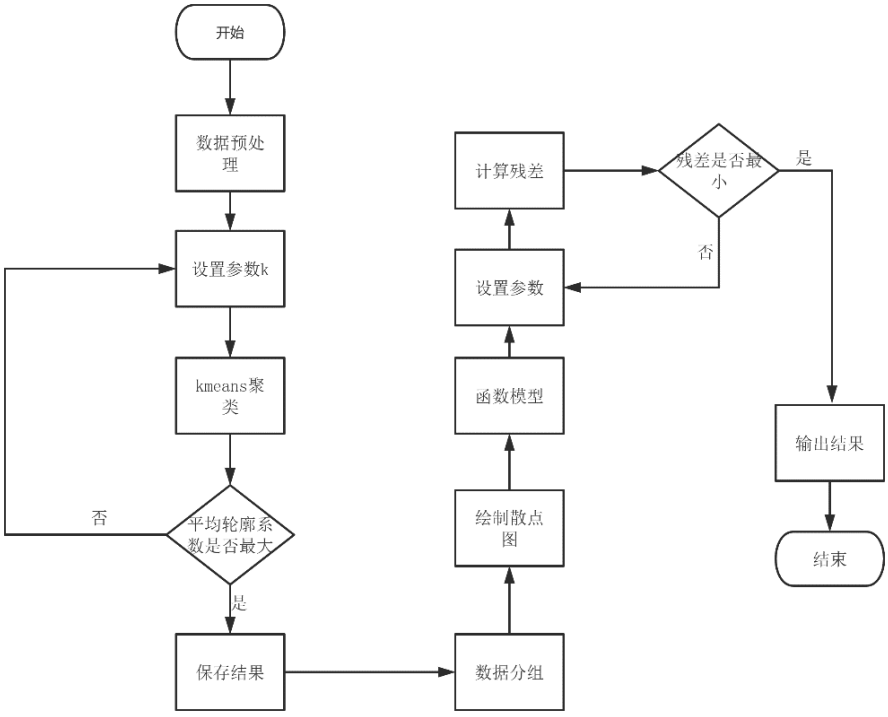

图 8 问题 2 子问题 b）的思路流程图

6.2.2 基于 K-means 聚类算法的模型建立与求解

（1）K-means 聚类算法[6]是按照某种特定的准则，将一个数据集合分为互不相交的多个簇，使得同一簇的数据对象尽可能地相似，不同簇的数据对象尽可能地有差异。

主要原理为：

1）首先随机选择 K 个样本点作为 K 个簇的初始簇中心；

2）计算每个样本点与这 K 个簇中心的相似度大小，并将该样本点划分到与之相似度最大的簇中心所对应的簇中;

3）根据现有的簇中样本，重新计算每个簇的簇中心；

4）循环迭代步骤 2），3），直到目标函数收敛，即簇中心不再发生变化。

（2）对勾函数

对勾函数是形如 $y=ax+\frac{b}{x}, ab>0$ 的一类函数。

（3）反比例函数

反比例函数的图像属于以原点为对称中心的中心对称的两条曲线，反比例函数图象中每一象限的每一条曲线会无限接近 x 轴和 y 轴。

形如 $y=k\frac{1}{x}$（k 为常数，$k\neq 0$）的形式，那么称 y 是 x 的反比例函数。

（4）线性函数

线性函数可以认为是一次曲线，形如 $y=ax+b$。

（5）多项式拟合

多项式拟合也是一个线性模型，其数学表达式为：

$$y(x,w)=\sum_{j=0}^{M}\omega_j x^j \tag{13}$$

其中 M 是多项式的最高次数，x^j 代表的是 x 的 j 次幂，w_j 是 x^j 的系数。

根据“表 2”整合的数据进行 K-means 聚类，依次聚类为 3,4,5，聚成 4 类时，轮廓系数最接近于 1，效果最好。然后对 4 个簇（亚组）绘制水肿体积随时间进展的散点图，观察散点图的趋势，对每个亚组进行拟合。

对记为 0 的亚组水肿体积和时间数据归一化处理，绘制散点图和拟合曲线，拟合出对勾函数 $f(x)=\frac{1}{x}+x(a=1,b=1)$，其中最小二乘法 $R^2=0.7128$。对记为 1 的亚组拟合出反比例函数 $f(x)=\frac{1}{x}+1(a=1,b=1)$，其中最小二乘法 $R^2=0.7543$。对记为 2 的亚组拟合出线性函数 $f(x)=4x$，其中最小二乘法 $R^2=0.8123$。对记为 3 的亚组拟合出多项式函数 $f(x)=2.8837x^3-3.6353x^2+1.2228x+0.1106$，其中最小二乘法 $R^2=0.8513$，拟合方法同记为 0 的亚组。根据最小二乘法的值来看，拟合的效果较好。

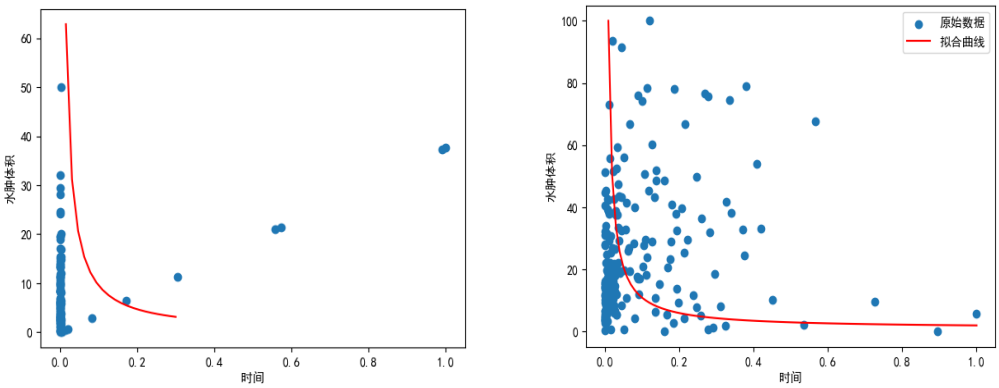

图 9 记为 0 和 1 的亚组水肿体积随时间的拟合曲线

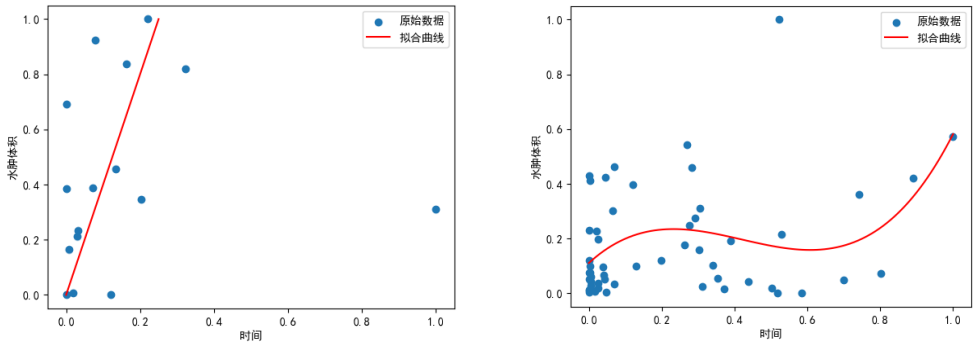

图 10 记为 2 和 3 的亚组水肿体积随时间的拟合曲线

对每个亚组拟合完后，将拟合的数据整合到 excel 中，使用 excel 计算残差，结果见“表 4-答案文件”。

从图 9 中可以看出，记亚组 0 和 1 的患者水肿体积开始时下降幅度较大，逐渐趋于平缓，患者在日后的生活中应注意保养。从图 10 中看，亚组 2 可以很明显看出同时患有高血压病史和卒中病史，均无冠心病史，吸烟史，饮酒史的患者随时间的变化，水肿体积在一直增大，在之后的生活中，患者若要避免脑卒中的发生，就要多关注和预防高血压和卒中病的发生。亚组 3 的患者开始时水肿体积相对平稳，但是随着时间的推移，水肿体积也在逐步增大，在日常生活中，患者应注意高血压和卒中病，以及附带的其他疾病的复发，引起水肿体积的增大，从而导致出血性脑卒中的发生。

注：亚组 0：患者有可能患有高血压病史，糖尿病史，冠心病史，均无卒中病史，吸烟史和饮酒史。亚组 1：患者大概率可能患有高血压病史，吸烟史，均无卒中病史。亚组 2：患者同时患有高血压病史和卒中病史，均无冠心病史，吸烟史，饮酒史，亚组 3：均患有高血压病史和卒中病史，大概率还附带患有糖尿病史、房颤史，冠心病史别的病史。

6.3 子问题 c)的分析与求解

6.3.1 问题分析

题中要求分析不同的治疗方法对水肿体积进展模式的影响，首先，用 excel 对“表 1”（sub001-sub100,sub131-sub160）的不同治疗方法和“表 2”水肿体积进行整合，由于患者随访 2 时间点水肿体积到随访 7 时间点水肿体积有缺失值，因此考虑最重要的两个时间点水肿体积，使用皮尔逊相关系数方法进行建模作图，结果见图 11。其中本文也考虑了对整合的数据，将治疗方法相同的归为一类，计算随访 1 到随访 7 时间点水肿体积平均增长率，对不同治疗方法的水肿体积进展绘制折线图，可以清晰地观察到不同治疗方法对水肿体积进展模式的影响。

6.3.2 相关系数求解

（1）皮尔逊相关系数：两个变量之间的皮尔逊相关系数定义为两个变量之间的协方差和标准差的商。

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}} \quad (14)$$

将“表 1”（sub001-sub100，sub131-sub160）的不同治疗方法和“表 2”水肿体积数据整合在一起，由于患者随访 2 时间点水肿体积到随访 7 时间点水肿体积有缺失值，因此考虑最重要的两个时间点水肿体积，使用皮尔逊相关系数方法建模作图。

初步观察图 11，可以看到脑室引流和降颅压治疗对患者的水肿体积有影响。

对于整合的数据，将 7 种治疗方法相同的归为同一种治疗方法，总共有 27 种治疗方法，如表 5 所示。计算随访 1 到随访 7，水肿体积的平均增长率。第二种治疗方法随访 2 到随访 3 时间点的水肿体积增加率为 52.77，数据异常大，剔除掉。对不同治疗方法的水肿体积进展绘制折线图，可以清晰地观察到不同治疗方法对水肿体积进展模式的影响，如图 12 所示。

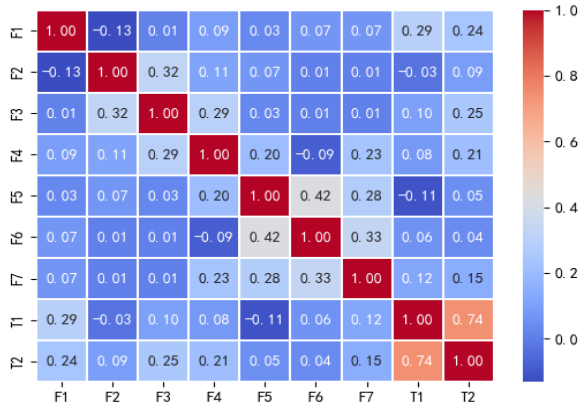

图 11 皮尔逊相关系数对数据的处理结果

注 F1：脑室引流，F2：止血治疗，F3：降颅压治疗，F4：降压治疗，F5：镇静、镇痛治疗，F6：止吐护胃，F7：营养神经，T1：患者首次水肿体积，T2：患者随访 1 时间的水肿体积。

表 5　不同治疗方案

治疗方法 / 方案种类	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇定治疗	止吐护胃	营养神经	患者数量
1	1	0	0	1	0	1	1	1
2	1	0	0	1	1	1	1	1
3	1	0	1	1	0	1	1	1
4	1	1	1	1	1	1	1	5
5	1	0	1	1	1	1	1	1
6	0	1	1	1	1	1	1	58
7	0	0	0	0	1	1	1	5
8	0	1	0	0	0	1	1	1
9	0	1	0	0	1	1	0	1
10	0	1	0	0	0	1	0	1
11	0	0	0	1	1	1	1	6
12	0	0	0	1	0	1	1	2
13	0	0	1	1	0	1	1	2
14	0	0	1	1	1	1	1	8
15	0	0	1	1	0	1	0	1
16	0	0	1	1	0	1	1	1
17	0	0	1	1	0	0	1	1
18	0	0	1	1	0	0	0	1
19	0	1	0	1	1	1	1	9
20	0	1	0	1	0	1	1	2
21	0	1	0	1	0	0	1	2
22	0	1	1	1	0	0	0	2
23	0	1	1	1	0	0	1	2
24	0	1	1	1	0	1	1	10
25	0	1	1	1	1	1	0	1
26	0	1	1	0	0	1	1	4
27	0	1	1	0	0	1	0	1

其中 1-27 代表 27 种不同的治疗方法，0 代表采取此治疗，1 代表未采取此治疗。

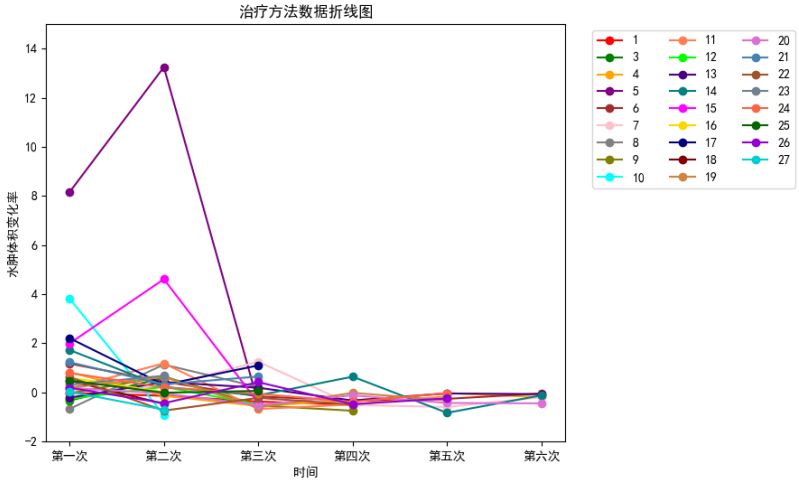

图 12 不同治疗方法对水肿体积变化率的影响

观察图 12 可以看出，患者进行治疗方法 5，第一次随访时间到第二次随访时间水肿体积增大，水肿体积的增大会导致血肿体积的增大，从而会引发脑卒中，在第二次随访时间到第三次随访时间水肿体积变化率迅速降低，同时，治疗方法只对一个患者进行了治疗，此组合方法对患者不稳妥，不建议使用。治疗方法 2，5，14，15，17，21，24 等的折线图可以反应出此些方法对患者降低水肿体积较为稳妥，在临床治疗方案上根据患者的年龄，水肿体积等因素建议采用。

6.4 子问题 d)的分析与求解

6.4.1 问题分析

将表中血肿体积、水肿体积及治疗方法的数据进行整合，探索其三者之间的关系，可以采用 Apriori 算法对数据进行分析处理，Apriori 算法可以在大量数据中找到经常一起出现的数据集合，发现数据之间的内在联系，从而找到数据的关联规则。即可以探索出血肿体积、水肿体积及治疗方法三者之间的关系。

问题 2 子问题 d）的思路流程图如图 13 所示：

6.4.2 基于 Apriori 算法的模型建立与求解

（1）Apriori 算法[7]

Apriori 算法基于 Apriori 原则，即如果一个项集是频繁的，那么它的所有子集也必须是频繁的。Apriori 算法通过依次迭代生成候选项集，计算候选项集的支持度，并剪枝来逐步获取频繁项集。

Apriori 算法的主要步骤如下：

1）初始化：扫描数据集，生成所有单个项的集合作为候选 1-项集。

2）迭代生成候选项集：根据 Apriori 原则，根据上一轮的频繁(k-1)-项集生成候选 k-项集。剪枝操作通过检查候选项集的所有(k-1)-子集是否都是频繁的来减少候选项集的数量。

3）计算支持度：扫描数据集，计算每个候选项集的支持度，即项集在数据集中出现的频率。

4）剪枝：根据最小支持度阈值，删除支持度小于阈值的项集，得到频繁项集。

5）迭代停止条件：如果没有频繁项集生成或者候选项集为空，则停止迭代。

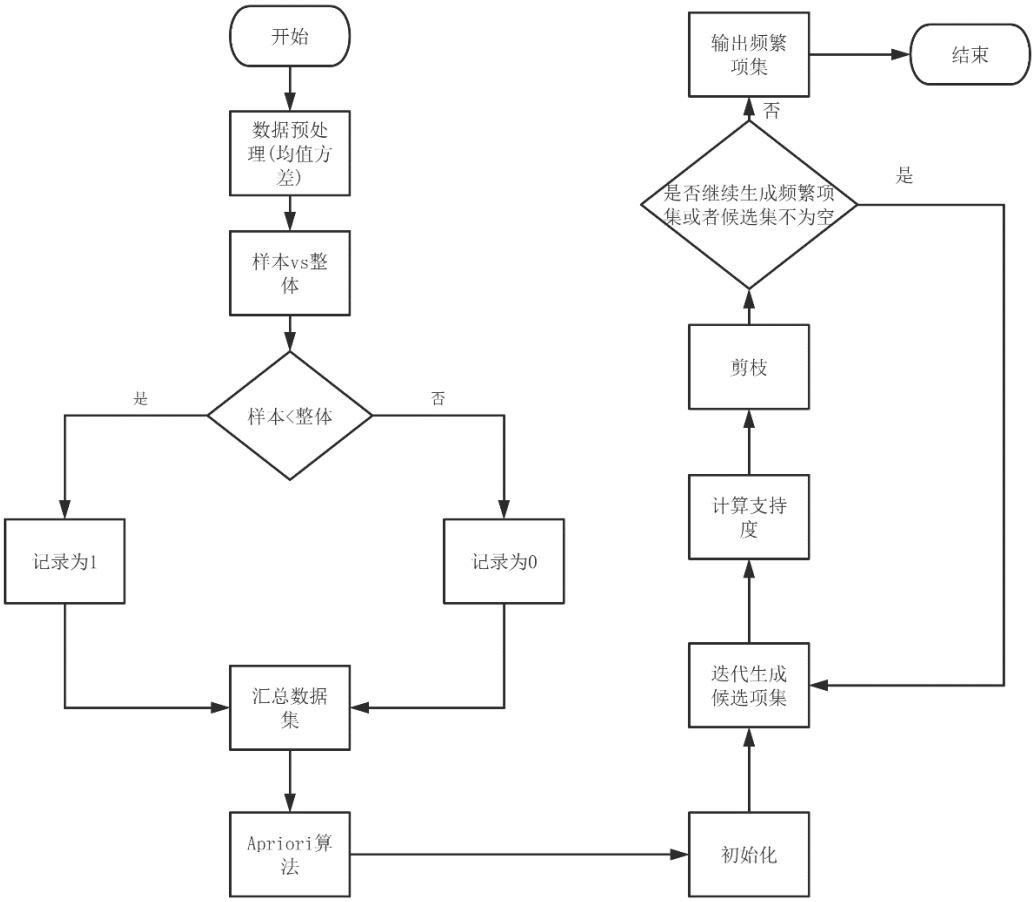

图 13 问题 2 子问题 d）的思路流程图

（2） 求解过程如下：

1）数据预处理，找到表中血肿体积、水肿体积及治疗方法的数据，使用 Excel 对其数据进行整合；

2）将实际问题转换为逻辑值，对数据进行离散化；

3）计算整体水肿体积变化的平均值 $\overline{X}$ 和标准差 σ；

4）计算每个样本水肿体积变化的平均值 $\overline{X_i}$ 和标准差 σ_i；

5）将 $\overline{X_i}$ 与 $\overline{X}$ 进行比较，如果 $\overline{X_i} < \overline{X}$，则记为 1，否则记为 0；

6）将 σ_i 与 σ 进行比较，如果 $\sigma_i < \sigma$，则记为 1，否则记为 0；

7）整合数据，用 Python 实现 Apriori 算法，得到最终结果，如表 6 所示。

通过 Apriori 算法得出以下结论：

$$\{3,4,5,6,7\} \Rightarrow \{8,9\}, \{2,4,5,6,7\} \Rightarrow \{8,9\}, \{2,3,5,6,7\} \Rightarrow \{8,9\},$$

$$\{2,3,4,6,7\} \Rightarrow \{8,9\}, \{2,3,4,5,7\} \Rightarrow \{8,9\}, \{2,3,4,5,6\} \Rightarrow \{8,9\}$$

即：当使用治疗方案

【降颅压治疗；降压治疗；镇静、镇痛治疗；止吐护胃；营养神经】、

【止血治疗；降压治疗；镇静、镇痛治疗；止吐护胃；营养神经】、

【止血治疗；降颅压治疗；镇静、镇痛治疗；止吐护胃；营养神经】、

【止血治疗；降颅压治疗；降压治疗；止吐护胃；营养神经】、

【止血治疗；降颅压治疗；降压治疗；镇静、镇痛治疗；营养神经】、
【止血治疗；降颅压治疗；降压治疗；镇静、镇痛治疗；止吐护胃】会使水肿、血肿体积减少。

表 6 通过 Apriori 算法获得的数据结果

治疗方案	体积减少
3，4，5，6，7	8，9
2，4，5，6，7	8，9
2，3，5，6，7	8，9
2，3，4，6，7	8，9
2，3，4，5，7	8，9
2，3，4，5，6	8，9

注：1 脑室引流； 2 止血治疗； 3 降颅压治疗；4 降压治疗；5 镇静、镇痛治疗；6 止吐护胃； 7 营养神经；8 水肿体积；9 血肿体积

七、问题三模型建立与求解

7.1 子问题 a)的分析与求解

7.1.1 问题分析

在子问题 a)中，要求根据前 100 个患者个人史、疾病史、发病相关及首次影像结果构建预测模型，首先，将“表 1”，“表 2”，“表 3”的相关数据使用 excel 进行整合，问题 1 的子问题 b)已经进行过降维，且降维效果很不错，因此直接在问题 1 降维的基础上进行预测，此问题使用了随机森林、决策树、神经网络 3 种预测模型预测患者（sub001 至 sub160）90 天 mRS 评分，其中使用随机森林预测模型预测患者的评分结果最好，预测准确率达到 88%，结果见“表 4-答案文件”。

问题 3 子问题 a）的思路流程图如图 14 所示。

7.1.2 基于随机森林、决策树、神经网络的模型求解

（1）随机森林

随机森林[8]是 bagging 方法的一种具体实现。它会训练多棵决策树，然后将这些结果融合在一起就是最终的结果。随机森林可以用于分类，也可以用于回归。主要在于决策树类型的选取，根据具体的任务选择具体类别的决策树。

对于分类问题，一个测试样本会送到每一颗决策树中进行预测，然后投票，得票最多的类为最终的分类结果；对于回归问题，随机森林的预测结果是所有决策树输出的均值。

随机森林[9]实现步骤：

1）使用装袋法在行列上进行随机抽样；

2）由很多决策数分类器组合而成；

3）单个的决策树分类器用随机方法构成。首先，学习集是从原训练集中通过有放回抽样得到的自助样本。其次，参与构建该决策树的变量也是随机抽样，参与变量数通常大大小于可用变量数；

4）单个决策树在产生学习集和确定参与变量后，使用 CART 算法计算，不剪枝(因为无需考虑过度拟合的问题)。

5）最后分类器结果取决于各个决策树分类器简单多数选举。

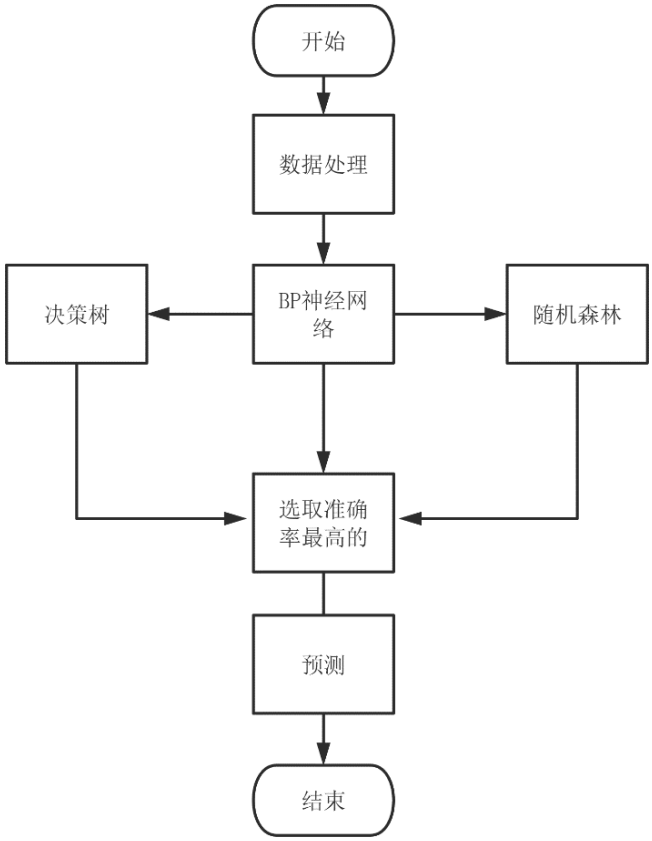

图 14 问题 3 子问题 a）的思路流程图

（2）决策树

决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。

一个决策树包含三种类型的节点：决策节点，机会节点，终结节点。

（3）神经网络

使用神经网络进行预测时，首先需要把数据归一化，把数据经过处理后使之限定在一定的范围内，然后进行网络设计。经典的 BP 神经网络通常由三层组成：输入层，隐含层，输出层。

（4）对三种模型进行评估

评价指标：

1）准确率

$$Acc = \frac{TP}{N} \tag{15}$$

其中，TP (True Positive)：表示模型正确预测为正类的样本数，N：表示样本总数。

2）均方误差（MSE）：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{16}$$

3）均方根误差（RMSE）：

$$RMSE = \sqrt{MSE} \tag{17}$$

4）平均绝对误差（MAE）：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \tag{18}$$

其中，N 表示样本数量，y_i 是真实标签，$\hat{y}_i$ 是预测标签。

将三个表的相关数据进行整合后，使用了随机森林、决策树、神经网络预测模型预测患者（sub001 至 sub160）90 天 mRS 评分。首先，使用随机森林预测模型，对整合的数据进行归一化，划分训练集和测试集，进行交叉验证，得到的预测准确率为 88%，预测效果较为明显。分别使用神经网络，决策树的预测准确率为 82%，73%。

表 7 三种方法的评价指标

方法	Acc	MSE	RMSE	MAE
神经网络	82%	1.45	1.20	0.51
决策树	73%	1.91	1.38	0.63
随机森林	88%	0.64	0.8	0.28

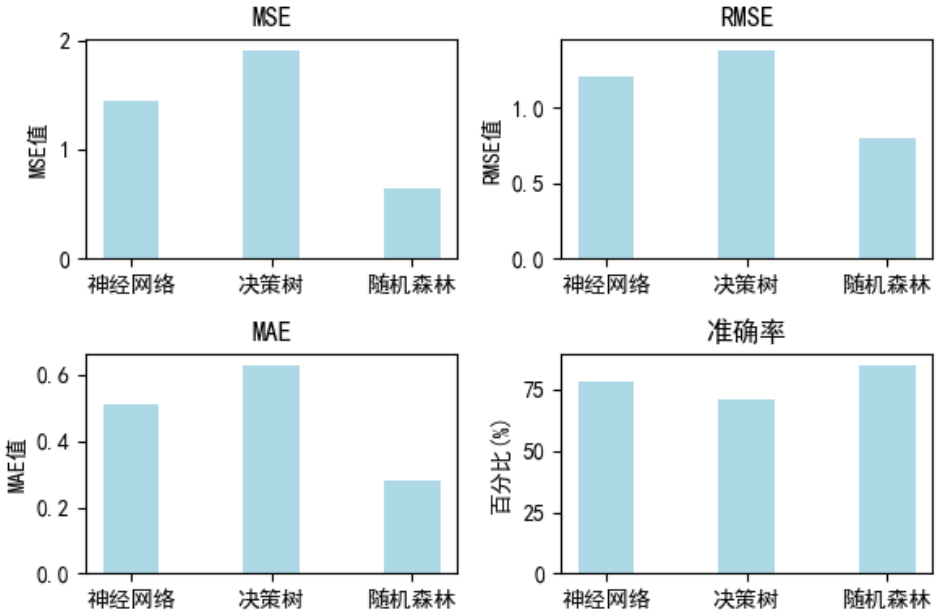

图 15 三种方法的评价指标

由表 7，图 15 所示，能明显的看出在预测患者（sub001 至 sub160）90 天 mRS 评分时，随机森林的方法最好，其次是神经网络和决策树。

7.2 子问题 b)的分析与求解

7.2.1 问题分析

在子问题 b)中，要求根据前 100 个患者所有临床、治疗影像（首次+随访）结果，对所有含有随访影像检查的患者 90 天的 mRS 评分进行预测，与 a）的数据相比多了后续随访的治疗影像，故先对“表 2” 首次+随访的数据进行降维处理，选取“表 2”中的主要特征，与“表 1”，“表 3”的主要特征放在一起作为训

练集进行预测。此问题使用了随机森林、决策树、神经网络 3 种预测模型预测所有含影像检查的患者（sub001 至 sub100）90 天 mRS 评分，其中使用随机森林预测模型预测患者的评分结果最好，准确率达到 89%，结果见“表 4-答案文件”。

问题 3 子问题 b）的思路流程图如图 16 所示：

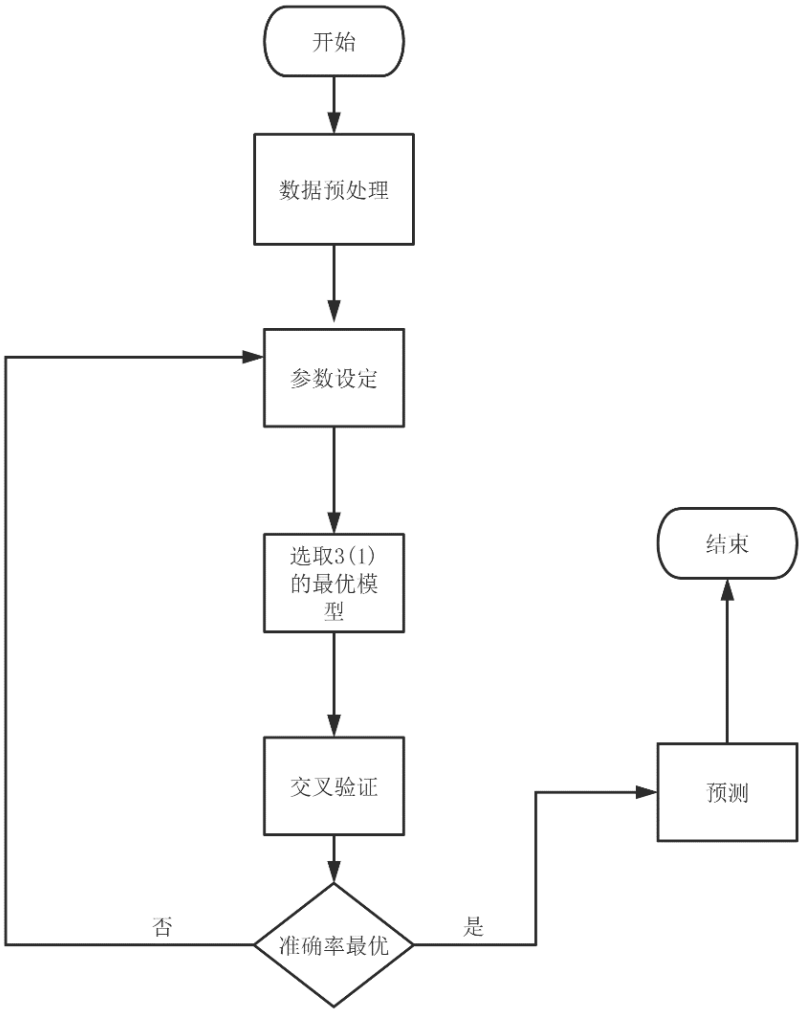

图 16 问题 3 子问题 b）的思路流程图

7.2.2 随机森林模型求解

此问题使用的方法同问题 3 中子问题 a)，使用了随机森林、决策树、神经网络 3 种预测模型预测所有含影像检查的患者（sub001 至 sub100）90 天 mRS 评分，其中使用随机森林预测模型预测患者的评分结果最好，准确率达到 89%，见表 8 所示，预测结果见“表 4-答案文件”。

表 8 随机森林的评价指标

方法	Acc	MSE	RMSE	MAE
随机森林	89%	0.73	0.85	0.31

注：Acc 为准确率，MSE 为均方误差，RMSE 为均方根误差，MAE 为平均绝对误差。

7.3 子问题 c)的分析与求解

7.3.1 问题分析

子问题 c)要求分析出血性脑卒中患者的预后（90 天 mRS）和个人史、疾病史、治疗方法及影像特征等之间的关系，对表中相关数据进行整合，对“表 1”中的数据使用信息增益方法重新进行降维处理，选取“表 1”中的主要特征，与“表 2”,“表 3”的主要特征放在一起作为训练集。通过计算信息熵的概率大小，留下了 6 个最重要的特征，因为 mRS 的 6 个数据特征是离散的，对于留下的特征利用有序多分类 logistic 回归模型分析出血性脑卒中患者预后和个人史、疾病史、治疗方法及影像特征等之间的特征关系，得出了模型拟合信息，平行线检验，参数估计值得出的结果如下。

问题 3 子问题 c）的思路流程图如图 17 所示：

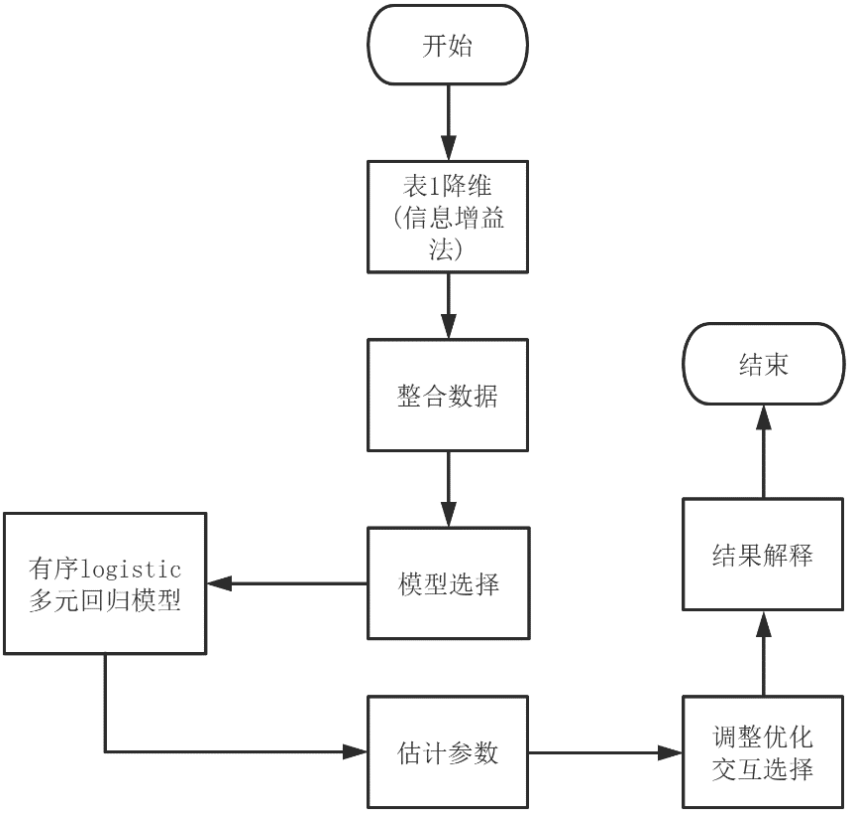

图 17 问题 3 子问题 c）的思路流程图

7.3.2 有序多分类 logistic 回归模型

（1）信息增益

信息熵：熵是信息论中一个很重要的概念，计算公式如下，

$$\mathrm{Ent}(Y) = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{19}$$

其中，$\mathrm{Ent}(Y)$越大，说明获得的信息量越大，同时说明 Y 更趋向于均匀分布，信息量大不大反应我们对事件发生预知的概率大不大。

（2）有序多分类 logistic 回归模型[10]

因变量为水平数大于 2 的有序多分类的资料，对这种资料可通过拟合因变量

水平数$n-1$个 logistic 回归模型，称为累计 logistic 模型。实质是依次将因变量按不同的取值水平分割成两个等级，对这两个等级建立因变量β_i为二分类的 logistic 回归模型，但模型中的各自变量系数β_i都保持不变，只改变常数项 (前提条件，需要验证)。以 4 个水平的因变量为例，其对应的概率为P_i，对 n 个自变量拟合 3 个模型（拟合累加模型），因变量有序取值水平的累计概率：

$$\log it\frac{P_1}{1-P_1}=logit\frac{P_1}{P_2+P_3+P_4}=-\alpha_1+\beta_1x_1+\ldots+\beta_nx_n$$
$$logit\frac{P_1+P_2}{1-(P_1+P_2)}=logit\frac{P_1+P_2}{P_3+P_4}=-\alpha_2+\beta_1x_1+\ldots+\beta_nx_n \quad (20)$$
$$logit\frac{P_1}{1-P_1-P_2-P_3}=logit\frac{P_1+P_2+P_3}{P_4}=-\alpha_3+\beta_1x_1+\ldots+\beta_nx_n$$

记β_i：自变量系数，P_i：概率.

（3）问题求解

1)首先，对题目中要求的相关数据用 excel 进行整合，使用信息增益方法对“表 1”进行降维，根据图 18 反应的不同特征对应的信息熵的概率，留下 6 个最重要的特征，分别是 Z1：年龄，Z2：性别，Z3：脑出血前 mRS 评分，Z8：HM_volume，Z11：original_shape_Elongation (ED_volume),
Z17：NCCT_original_firstorder_Median.

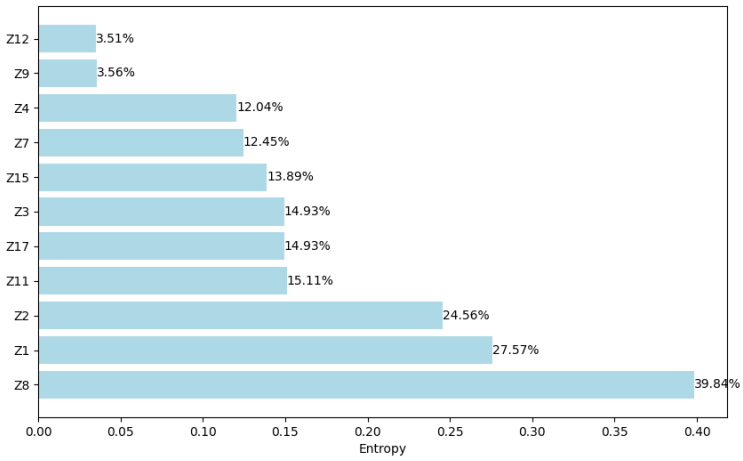

图 18　特征对应的信息熵值

注：　Z1：年龄 ；Z2：性别；Z3：脑出血前 mRS 评分；Z4：冠心病史 ；Z5：脑室引流；Z6：营养神经；Z7：ED_volume ；Z8：HM_volume；
Z9：NCCT_original_firstorder_InterquartileRange；Z15：original_shape_Sphericity；
Z10：NCCT_original_firstorder_RobustMeanAbsoluteDeviation；Z12：original_shape_Flatness；
Z11：original_shape_Elongation　(ED_volume)；Z14：NCCT_original_firstorder_Entropy；
Z13：original_shape_Maximum2DdiameterSlice；Z17：NCCT_original_firstorder_Median；
Z16：NCCT_original_firstorder_Uniformity；Z18：original_shape_Elongation　(HM_volume)；

接着分析降维后的主要特征和预后 mRS 的相关系数，结果如图 19 所示。

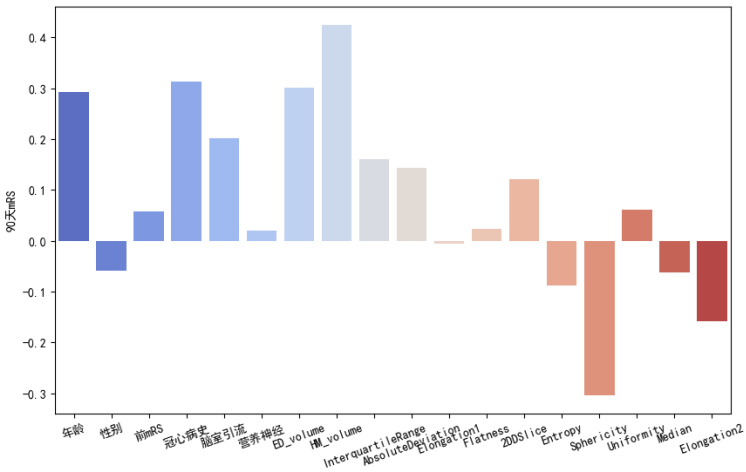

图 19 预后 mRS 与各变量的相关系数图

观察发现，年龄，冠心病史，ED_volume，HM_volume 的值越大，预后 mRS 的等级可能越高。下面进行更加细致的分析：

利用 SPSS 软件进行处理，使用了有序多分类 logistic 回归模型分析出血性脑卒中患者预后和个人史、疾病史、治疗方法及影像特征等之间的特征关系，得出了模型拟合信息，平行线检验，参数估计值的结果如表 9-11 所示。

表 9 模型拟合信息

模型	-2 对数似然	卡方	自由度	显著性
仅截距	371.465			
最终	310.070	61.395	19	.000

其中，假设检验 H_0：假设模型所有偏回归系数全部为 0，采用的是最大似然比检验，只含常数项模型与最终模型的负 2 倍最大似然比之差（-2 Log Likelihood 值之差）为卡方值，$\chi^2 = 371.465 - 310.070 = 61.395$，P<0.001，说明至少有一个自变量的偏回归系数不为 0。对模型的改善与否，主要看-2 对数似然是否有降低，其次，模型的显著性<0.05，认为模型中偏回归系数不全为 0，模型具有统计意义。

表 10 平行线检验 [a]

模型	-2 对数似然	卡方	自由度	显著性
原假设	310.070			
常规	193.927[b]	116.143[c]	95	.069

平行线检验是否满足风险比例假定，即检验自变量对因变量的影响在各个回归方程中是否相等（各自变量系数在各个模型中是否相等）。似然比 $\chi^2 = 116.143$，P=0.069＞0.05，满足平行性的假设。

将预后 mRS 评分作为因变量，将 $Z_1 - Z_{18}$ 作为自变量进行有序多元 logistic 回归分析,结果如表 11 所示。

表 11 参数估计值

		估算	标准错误	瓦尔德	自由度	显著性	95%置信区间下限
阈值	[90 天 mRS=0]	-20.136	30.209	.444	1	.505	-79.346
	[90 天 mRS=1]	-18.501	30.199	.375	1	.540	-77690
	[90 天 mRS=2]	-17.281	30.195	.328	1	.567	-76.462
	[90 天 mRS=3]	-16.003	30.193	.281	1	.596	-75.180
	[90 天 mRS=4]	-14.970	30.187	.246	1	.620	-74.135
	[90 天 mRS=5]	-12.536	30.169	.173	1	.678	-71.666
位置	年龄	**.036**	.018	4.283	1	**.038**	.002
	ED_volume	**6.029**	1.468E-5	16.856	1	**.000**	3.151E-5
	HM_volume	**2.828**	1.464E-5	3.735	1	**.045**	-4.021E-7
	InterquartileRange	.895	.753	1.411	1	.235	-.582
	AbsoluteDeviation	-1.620	1.849	.768	1	.381	-5.245
	Elongation1	.059	1.726	.001	1	.973	-3.323
	Flatness	-1.816	2.384	.580	1	.446	-6.489
	@2DDSlice	-.031	.012	6.445	1	.011	-.055
	Entropy	-1.936	6.505	.089	1	.766	-14.685
	Sphericity	-5.490	3.078	3.180	1	.075	-11.523
	Uniformity	35.225	82.616	.182	1	.670	-197.150
	Median	**-.022**	.011	3.792	1	**.052**	-.044
	Elongation2	-1.107	1.524	.528	1	.467	-4.093
	[性别=0]	-.517	.456	1.282	1	.257	-1.411
	[性别=1]	0[a]	.	.	0	.	.
	[前 mRS=0]	-1.996	1.576	1.603	1	.205	-5.086
	[前 mRS=1]	**3.069**	2.436	1.588	1	**.028**	-7.843
	[前 mRS=2]	0[a]	.	.	0	.	.
	[冠心病史=0]	-1.346	.785	2.939	1	.086	-2.885
	[冠心病史=1]	0[a]	.	.	0	.	.
	[脑室引流=0]	-.911	.861	1.117	1	.291	-2.599
	[脑室引流=1]	0[a]	.	.	0	.	.
	[营养神经=0]	-1.061	1.037	1.048	1	.306	-3.093
	[营养神经=1]	0[a]	.	.	0	.	.
	Elongation2	-1.107	1.524	.528	1	.467	-4.093

表 12 参数估计值得出影响 mRS 的特征

特征	对 mRS 的影响趋势
年龄	+
ED_volume	+
HM_volume	+
前 mRS	+

由表 12 可知，在显著性水平为 0.05 上共有 4 个因素进入回归方程，分别是年龄(x_1)，前 mRS(x_3)，ED_volume(x_7)，HM_volume(x_8)。通过对模型中的所有自变量的偏回归系数是否为 0 进行检验，$p<0.001$，故 logistic 回归方程具有统计学意义。即解释变量$x_1,\cdots,x_{18}$对解释预后 mRS 评分是有意义的。

由表 11 可知,年龄的偏回归系数$\beta_1=0.036>0$，这说明随着年龄的增长，一些人的身体状况不是很好，有的甚至只能坐在轮椅上，生活质量下降，幸福感较低，对这些人的关注，是促进人类进步的事情，有益无害。

前 mRS=1 的偏回归系数$\beta_3=3.069$，$OR3>1$ 。这说明随着前 mRS 的提高,预后 mRS 也上升，显然，这两个量之间的关系很密切，身体状况不佳的人，接受治疗后，能够治愈的很少，有的甚至变得更严重，这可能与患者本身的身体状况，经济情况，精神状态等有关。

水肿体积 ED_volume 的偏回归系数$\beta_7=6.209>0$，这说明随着水肿体积的增大，患者的情况越危险，大脑内部的结构组织复杂，患者的身体状况时刻都在发生变化，且水肿体积增大，会对周围的血管组织产生压迫，基于此，在遵循患者的意愿下，应立即采取措施，进行治疗。

血肿体积的偏回归系数$\beta_8=2.828>0$，这说明血肿体积越小，患者的身体状况越好，轻度患者和正常人无异，可以正常的生活，幸福感较高。相反，血肿体积越大，患者身处危险之中，后续的治疗，就可能得长期坚持下去。

综合分析后得出，这四种因素对预后的 mRS 解释性强，而前期的 mRS 更是有着指引性的作用，自己才是健康的第一负责人，要隔一段时间做相应的医疗检查，防患于未然。对于年龄较大的人，患脑卒中的风险更高，需要的社会倾注就越多。

（4）为临床相关决策提出建议

综合考虑之后，给出以下的对策和建议：

1）更加个性化的治疗方案：个性化治疗方案是根据脑卒中患者的病情、病因和个体特点制定的。例如，针对不同类型的脑卒中，医生会选择不同的药物治疗方案，以最大程度地减少再发和并发症的风险

2）心理疏导和支持：脑卒中患者常常面临心理压力和情绪变化，包括焦虑、抑郁和自卑等。专业的心理疏导和支持可以帮助他们应对这些困扰，并调整自我认知。在社会中，可以提供心理咨询和康复指导，使患者和家属更好地适应脑卒中后的生活变化，并提供他们所需的支持和安慰。

3）饮食和生活方式的改变：饮食和生活方式的改变对脑卒中的预防和康复至关重要。社会可以通过宣传教育和政策倡导，提倡健康饮食和生活方式。例如，推广蔬菜水果的摄入，减少高盐高脂食物的摄入，鼓励适量的有氧运动，如散步、游泳等。适当参与运动，可以降低患脑卒中的风险。

4）家庭护理和支持：家庭在脑卒中患者的护理和康复过程中起到至关重要的作用。社会可以提供家庭护理和支持的资源和服务。例如，社区护理机构可以派遣护士或康复治疗师上门为患者提供康复训练和护理指导，帮助患者恢复生活能力。

5）定期复查和随访：定期复查和随访是脑卒中患者康复中的重要环节。医疗机构可以建立定期复查的提醒系统，协助患者按时前往复诊，并提供相应的检查和评估。定期的复查和随访有助于及时调整治疗方案，提高康复效果和预后。

通过个性化治疗方案、心理疏导和支持、饮食和生活方式的改变、家庭护理和支持以及定期复查和随访等综合措施，社会为脑卒中患者提供了高尚的关爱和支持。这些措施的实施不仅帮助患者恢复生理功能，更是一种对人性关怀的体现。社会的参与和支持，让患者在脑卒中的挑战中感受到关怀与温暖，让他们在康复的道路上有更大的信心和动力。这种高大上的综合治疗和综合护理模式，不仅提高了患者的生活质量，也为构建健康社会提供了有益的借鉴和启示。以人为本、关爱生命，我们相信，通过社会的共同努力，脑卒中患者必将迎来更加美好的明天。

八、模型的评价与推广

8.1 模型的评价

8.1.1 模型的优点

1、本文在数据处理上，对“表 1、表 2、表 3”用了信息增益，Lasso 回归，方差最大的降维方法进行处理，最终选取的特征很有代表性，方便后续的分析和处理，在一定程度上可以提高求解的准确性。

2、本文在第三问的 a 中，通过多种预测模型进行求解，最终通过对比各模型的 Acc，MSE，RMSE，MAE，选出了随机森林模型。

3、本文在第三问的 c 中，利用有序多分类 logistic 回归模型分析了预后 90 天 mRS 与 18 个特征之间的因果关系，这里所选取的模型，适合处理因变量离散，自变量既有离散，也有连续的情况，最终得到的结果也符合人们的认知。

8.1.2 模型的缺点

1、由于数据集大小具有一定的局限性，在处理问题时可能会存在误差。

2、本文在第二问的 b 中，只选取了 kmeans 聚类去分析，虽然聚出的 4 个类别很有特点，但方法可能比较单一，也可以考虑别的聚类方法，比如：kNN，DBSCAN，FCM，LoDOG 等，进而去求解问题。

8.2 模型的推广与改进

1、本文提出的方法和模型可推广应用到其他临床智能诊疗的优化建模中。

2、模型可以在考虑变量耦合性的复杂问题上进行一些探索和优化，以便处理更复杂的数据挖掘问题。

3、可以考虑集成的模型，每一步都用很多模型去做，然后将每种模型的结果加权，作为下一次的输入，可能会提高结果的准确性。

参考文献

[1] 马东周,王红斌,翟风利,等. 重型出血性脑卒中[J]. 中国现代医学杂志,2002,12(10):98-99.

[2] 宋惠辉,罗欢,戴文卓. 血压与出血性脑卒中预后的关联研究[J]. 中国血液流变学杂志,2022,32(4):526-530,643.

[3] 韩小伟,李茗,张冰. 人工智能在脑卒中神经影像中的应用[J]. 协和医学杂志,2021,12(5):749-754.

[4] 任克强,张国萍,赵光甫.基于相对文档频的平衡信息增益降维方法[J].江西理工大学学报,2008(05):68-71.
[5] 王辉. BFRC 力学性能试验研究及其本构模型的 BP 神经网络预测[D].安徽大学,2021.
[6] 孙林,刘梦含.基于自适应布谷鸟优化特征选择的 K-means 聚类[J/OL].计算机应用:1-13[2023-09-25].
[7] 杨晨. 基于 Apriori 关联算法的银行数据挖掘[D].云南财经大学,2023.
[8] 许芳芳,胡江,陈维仁,等. 基于随机森林预测国内外ICU患者的死亡风险比较研究[J].世界最新医学信息文摘（连续型电子期刊）,2020,20(8):15-16.
[9] 郭豪. 双权重随机森林预测算法及其并行化研究[D]. 黑龙江:哈尔滨工业大学,2016.
[10] 杜颖,龙婷.基于多分类有序回归模型的湖南省老年人中医养生认知及其影响因素实证研究[J].国外医学卫生经济分册,2017,34(04):157-16.

附录

支撑材料说明：

支撑材料中包含了表 4-答案文件.xlsx，是对所有问题的结果汇总。

除此之外，支撑材料中还包含了文件夹 1ab，2a，2b，2c，2d，3a，3b，3c，它们分别对应每一个小问，具体包含了

（1） 程序代码（txt 格式）

（2） 需要导入的数据集，以及结果数据集（Excel 文档）

（3） 图片结果（png 格式）

本文的所有程序代码基于 python，依赖的第三方库通过 import 导入，如果没有，则需要手动安装。

下面展示每个小问的核心代码：

```
核心代码：
1a:
result = []
formatOut = '%Y-%m-%d %H:%M:%S'
for i in range(100):
    Time1 = table1.iloc[i, 13]
    t1 = str(table3.loc[i, 'var1'])
    deltas = []
    flag = False
    for j in range(2, 10):
        if pd.isnull(table3.loc[i, 'var{}'.format(j)]):
            break
        t = str(table3.loc[i, 'var{}'.format(j)])
        delta = (datetime.strptime(t, formatOut) - datetime.strptime(t1, formatOut)).total_seconds() / 3600 / 24
        deltas.append(delta)
        if sum(deltas) + Time1 > 48:
            result.append([0, 0])
            flag = True
            break
        if j % 3 == 2:  # 如果是t2, t5, t8
            mn = table3.loc[i, 'var{}5'.format(j)] - table3.loc[i, 'var2']
            nm = mn / table3.loc[i, 'var2']
            if mn > 6000 or nm > 0.33 or nm == 0.33:
                result.append([1, sum(deltas)])
                flag = True
                break
    if not flag:
```

```
        result.append([0, 0])  # 没有满足条件的情况

  1b:
#计算信息增益
def calc_gains(data):
    data['T'] = data['T'].astype(int)
    gains = {}
    for f in data.columns[1:]:
        f_col = data[f]
        v_counts = f_col.value_counts()
        probs = v_counts / len(f_col)
        entr = -
probs.dot(probs.apply(lambda x: x * math.log2(x)))
        cond_entr = (
            f_col.groupby(data['T'])
            .apply(lambda x: -
(x.size / len(f_col)) * (x.value_counts() / x.size).dot((la
mbda y: y * math.log2(y))(y)))
            .sum()
        )
        gain = entr - cond_entr
        gains[f] = gain
    print("Gains:")
    print(gains)

#Lasso 回归
def lasso_selection(data):
    X = data.iloc[:, 1:]
    Y = data['T']
    lasso = Lasso(alpha=0.012)
    lasso.fit(X, Y)
    f_importance = lasso.coef_
    f_names = X.columns
    importance = pd.DataFrame({'F': f_names, 'Importance':
abs(f_importance)})
    sorted_importance = importance.sort_values('Importance'
, ascending=False)
    print("Lasso Feature Importance:")
    print(sorted_importance)

#加载数据
def load_data(filename):
    data = pd.read_excel(filename)
    return data
```

```
#数据分割
def prepare_data(data):
    new_columns = ['Y', 'Y1', 'Y2', 'Y3', 'Y4', 'Y5', 'Y6', 'Y7', 'Y8', 'Y9', 'Y10', 'Y11', 'Y12', 'Y13', 'Y14', 'Y15','Y16', 'Y17', 'Y18', 'Y19']
    data.columns = new_columns
    X = data.iloc[:, 1:]
    Y = data['Y']
    return X, Y
#数据最大最小值化
def normalize_data(X):
    scaler = MinMaxScaler()
    X_normalized = scaler.fit_transform(X)
    return scaler, X_normalized
#参数搜索
def parameter_search(X, Y):
    parameters = {'hidden_layer_sizes': [(10,), (20,), (30,)],
                  'learning_rate_init': [0.001, 0.01, 0.1],
                  'alpha': [0.0001, 0.001, 0.01]}
    mlp = MLPClassifier()
    grid_search = GridSearchCV(mlp, parameters, cv=5)
    grid_search.fit(X, Y)
    best_params = grid_search.best_params_
    return best_params
#训练模型
def build_and_train_model(X, Y, best_params):
    mlp_best = MLPClassifier(hidden_layer_sizes=best_params["hidden_layer_sizes"],
                             learning_rate_init=best_params["learning_rate_init"],
                             alpha=best_params["alpha"])
    mlp_best.fit(X, Y)
    return mlp_best
#交叉验证
def cross_validation(X, Y, model):
    skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=0)
    results = []
      for train_index, test_index in skf.split(X, Y):
          X_train, X_test = X[train_index], X[test_index]
          Y_train, Y_test = Y[train_index], Y[test_index]
          model.fit(X_train, Y_train)
          score = model.score(X_test, Y_test)
```

```
105.             results.append(score)
106.         return results
107.     #预测
108.     def predict_probabilities(data, model, scaler):
109.         X = data.iloc[:, 1:]
110.         X_normalized = scaler.transform(X)
111.         prediction_prob = model.predict_proba(X_normalized)
112.         return prediction_prob
113.
114.     2a:
115.     data = pd.read_excel('2a.xlsx')
116.     x1 = data.values[:, 0]
117.     x2 = data.values[:, 1]
118.     x1_minmax = (x1 - np.min(x1)) / (np.max(x1) - np.min(x1))
119.     x2_minmax = (x2 - np.min(x2)) / (np.max(x2) - np.min(x2))
120.     #伽马函数
121.     def func(x, k, theta, a, b, c):
122.         return (a + (1 - x) * b) * gamma.pdf(1 - x, k, scale=
    theta) + c
123.     #拟合
124.     fit_params, _ = curve_fit(func, x1_minmax, x2_minmax, p0=
    [2, 2, 1, 1, 1])
125.     fit_x = np.linspace(0, 1, 100)
126.     fit_y = func(fit_x, *fit_params)
127.     plt.scatter(x1_minmax, x2_minmax, label='真实数据')
128.
129.     2b:
130.     #kmeans 聚类
131.     kmeans = KMeans(n_clusters=k, random_state=0)
132.     data['cluster'] = kmeans.fit_predict(data[['年龄', '性别
    ', '脑出血前 mRS 评分', '高血压病史', '卒中病史', '糖尿病史', '
    房颤史', '冠心病史', '吸烟史', '饮酒史', '发病到首次影像检查时间
    间隔', '血压']])
133.     #分组拟合
134.     def fit_curve(data, group_num):
135.         scaler = MinMaxScaler()
136.         time_normalized = scaler.fit_transform(data["时间
    "].values.reshape(-1, 1))
137.         volume_normalized = scaler.fit_transform(data["体积
    "].values.reshape(-1, 1))
138.         if group_num == 1:
139.             func = lambda x, a, b: a * x + b
140.         elif group_num == 2:
141.             func = lambda x, a, b: a / x + b
```

```
142.     elif group_num == 3:
143.         func = lambda x: x * (1/0.25)
144.     elif group_num == 4:
145.         poly = PolynomialFeatures(degree=3)
146.         time_poly = poly.fit_transform(time_normalized)
147.         model = LinearRegression()
148.         model.fit(time_poly, volume_normalized)
149.         func = lambda x: model.predict(poly.transform(x.r
    eshape(-1, 1)))
150.     else:
151.         raise ValueError("Invalid group number")
152.     x = np.linspace(0, 1, 20)
153.     y = func(x)
154.     return x, y
155. 
156. 2c:
157. #相关系数图
158. sns.heatmap(pd.read_excel('数据
    c.xlsx').corr(), cmap='coolwarm', annot=True, fmt='.2f', li
    newidths=.05)
159. #水肿体积平均变化率
160. data = pd.read_excel('2c0.xlsx').replace(-1, pd.NA)
161. fig, ax = plt.subplots(figsize=(10, 6))
162. for _, row in data.iterrows():
163.     treatment = row['治疗方法']
164.     values = row.drop('治疗方法').dropna()
165.     ax.plot(values.index, values.values, marker='o', labe
    l=f'治疗方法 {treatment}')
166. 
167. 2d:
168. #计算整体均值，方差
169. def calculate_overall_mean_std(data):
170.     overall_mean = data.mean().mean()
171.     overall_std = data.std().mean()
172.     return overall_mean, overall_std
173. #样本均值，方差与整体比较
174. def compare_with_overall_mean_std(data, overall_mean, ove
    rall_std):
175.     result_comparison = (data[['水肿均值', '血肿均值
    ']] < overall_mean).astype(int)
176.     result_comparison = pd.concat([result_comparison, (da
    ta[['水肿方差', '血肿方差
    ']] < overall_std).astype(int)], axis=1)
177.     return result_comparison
```

```
178.  #整合数据
179.  def merge_data(A_data, result_comparison):
180.      merged_data = pd.concat([A_data.iloc[:, 16:23], resul
   t_comparison], axis=1)
181.      return merged_data
182.
183.  def transform_data(merged_data):
184.      te = TransactionEncoder()
185.      df = pd.DataFrame(te.fit_transform(merged_data.astype
   (str)), columns=te.columns_)
186.      return df
187.  #使用 apriori 算法挖掘关系
188.  def apply_apriori(df, min_support):
189.      frequent_itemsets = apriori(df, min_support=min_suppo
   rt, use_colnames=True)
190.      return frequent_itemsets
191.  3a:
192.  #数据列名的重新命名
193.  def load_data(filename):
194.      data = pd.read_excel(filename)
195.      new_columns = ['mRs', 'Y1', 'Y2', 'Y3', 'Y4', 'Y5', '
   Y6', 'Y7', 'Y8', 'Y9', 'Y10', 'Y11', 'Y12', 'Y13', 'Y14', '
   Y15', 'Y16', 'Y17', 'Y18', 'Y19']
196.      data.columns = new_columns
197.      return data
198.  #数据归一化
199.
200.  #对参数进行搜索
201.  def parameter_search(model, parameters, X, Y):
202.      grid_search = GridSearchCV(model, parameters, cv=10)
203.      grid_search.fit(X, Y)
204.      best_params = grid_search.best_params_
205.      return best_params
206.
207.  def main():
208.      # 加载数据
209.      data = load_data("三表整合 3a.xlsx")
210.      X = data.iloc[:100, 1:20]
211.      Y = data.iloc[:100, 0]
212.      # 归一化
213.      scaler, X_normalized = normalize_data(X)
214.      # 随机森林参数搜索
215.      rf_parameters = {'n_estimators': [10, 50, 100],
216.                       'max_depth': [4, 6, 8]}
```

```
    rf = RandomForestClassifier()
    rf_best_params = parameter_search(rf, rf_parameters, X_normalized, Y)
    # 决策树参数搜索
    dt_parameters = {'max_depth': [4, 6, 8, 10],
                     'min_samples_split': [2, 4, 6]}
    dt = DecisionTreeClassifier()
    dt_best_params = parameter_search(dt, dt_parameters, X_normalized, Y)
    # 神经网络参数搜索
    mlp_parameters = {'hidden_layer_sizes': [(10,), (20,), (30,)],
                      'learning_rate_init': [0.001, 0.01, 0.1],
                      'alpha': [0.0001, 0.001, 0.01]}
    mlp = MLPClassifier()
    mlp_best_params = parameter_search(mlp, mlp_parameters, X_normalized, Y)

3b:
#数据列名的重新命名
def load_data(filename):
    data = pd.read_excel(filename)
    data.columns = ['mRs', 'Y1', 'Y2', 'Y3', 'Y4', 'Y5', 'Y6', 'Y7', 'Y8', 'Y9', 'Y10', 'Y11', 'Y12', 'Y13', 'Y14', 'Y15', 'Y16', 'Y17', 'Y18', 'Y19']
    return data
#最大最小值归一化
#随机森林模型训练
def train_model(X_train, Y_train):
    param_grid = {'n_estimators': [10, 50, 100],
                  'max_depth': [4, 6, 8]}
    grid_search = GridSearchCV(RandomForestClassifier(), param_grid, cv=10)
    grid_search.fit(X_train, Y_train)
    best_params = grid_search.best_params_
    rf_model = RandomForestClassifier(n_estimators=best_params["n_estimators"], max_depth=best_params["max_depth"])
    rf_model.fit(X_train, Y_train)
    return rf_model
#预测结果
def predict_data(model, scaler, input_data):
    input_normalized = scaler.transform(input_data)
    prediction = model.predict(input_normalized)
```

```
252.      return prediction
```