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：

出血性脑卒中具有起病急、进展快、预后较差的特点，且大多数患者发病后会有较严重的神经功能性遗留障碍。而造成出血性脑卒中预后较差的主要因素为血肿范围扩大和血肿周围水肿。因此，本文基于患者的医学影像数据，结合机器学习方法，对血肿扩张风险进行预测，推测水肿演进规律，并对临床预后进行预测。本文的具体作法如下：

针对问题 1，本文首先根据 48h 内血肿扩张的识别规则判断患者在发病 48h 内是否发生血肿扩张，并记录血肿扩张发生时间。接下来根据患者的个人史、疾病史、发病相关、影像检查数据，将是否发生血肿扩张作为响应变量。为了解决数据的维度较高而样本量较少的问题，本文使用 Spearman 相关性分析进行变量的特征选择，保留显著性小于 0.1 的变量共 7 个，作为自变量。以响应变量作为分类标签，分别建立了随机森林模型、bp 神经网络模型、支持向量机分类模型，并采用模拟退火进行寻优。比较发现，随机森林模型在测试集上的分类预测效果相对较好，精确率为 79.2%，F1 值为 0.775。

针对问题 2，本文首先对数据进行预处理后，绘制了水肿体积与发病时间的散点图，采用多项式方程进行拟合，并计算拟合残差。接下来，根据患者的个体差异指标，结合分类变量与连续变量，采用两步聚类法将患者聚为了 4 类，分别定义为亚组 1—亚组 4。对不同亚组中的水肿体积与发病时间数据，采用高斯拟合对不同亚组的散点图进行拟合；在分析治疗方法与水肿体积、血肿体积的关系过程中，进行数据的处理后，分别使用随机森林、决策树分类判别预测，来判别治疗手段对水肿体积、血肿体积的影响关系，再使用卡方检验出血肿体积和水肿体积在患者初始值大小以及变化量大小的一致性。

针对问题 3，根据所给的个人史、临床及治疗数据、影像相关数据来进行建模，对预

后 90 天 mRS 指标进行合理预测，以便对患者进行预后处理。我们通过变量间相关性，结合相关数据进行特征选择和关键因素探索，保留了 12 个指标，并用随机森林分类对 90 天 mRS 进行分类预测，测试集的精确率为 64.4%，F1 值为 0.617。再进一步采用岭回归来研究关键因素的影响效果，从而及时发现与其关联密切的临床特征及对其作用显著的对策、从而给出预后建议。

关键词：Spearman 相关系数、分类预测、特征选择、随机森林、两步聚类法、岭回归

目录

1. 问题背景及问题重述

1.1. 问题背景

由于非外伤性脑实质内血管破裂，从而引起的脑出血的这种现象，称作是出血性脑卒中，这占了全部脑卒中发病率的大约 10%。由于出血性脑卒中的病因复杂，而且常由于脑动脉瘤会发生破裂、异常等的因素，从而导致了血液从破裂的血管中涌入了脑组织，进而会造成脑部的机械性的损伤，进而引起一系列复杂的生理或病理性的反应。因为出血性脑卒中起病急、进展快，预后较差，急性期内病死率高，发掘出血性脑卒中的发病风险，整合影像学特征、患者临床信息及临床诊疗方案，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。

为了预防出血性脑卒中的血肿扩大带来的一系列危险后果，如导致脑组织受损，造成炎症反应等因素的不良状况，最终威胁患者的生命，因此，在本题中，也将监测并控制血肿的扩张，将这部分的影响进行分析建模，对识别和预测早期的患者血肿扩张和血肿周围水肿的发生及发展，具有极其重要的意义。

而现如今的医学影像技术的发展及进步，如何进行无创动态监测出血性脑卒中后脑组织损伤和演变，也是随着人工智能水平的发展及进步的又一关注问题。本文也将基于赛题提供的影像信息，结合患者的个人信息、治疗方案和预后等的数据，构建出一系列的智能诊疗模型，明确这些能够导致出血性脑卒中预后不良的危险因素，最终实现精准的、个性化的疗效评估和预后预测。

1.2. 问题重述

模型分为三个大的问题，第一个问题是针对血肿扩张风险相关因素进行探索的建模，分为两个小的问题，其中，（a）小问是需要对表 1，表 2 中患者等的信息，在 48 小时内是否会发生血肿扩张事件，而（b）小问则是在血肿扩张事件为目标变量的时候，判定预测所有患者的血肿扩张发生的概率；

第二个问题则是对血肿周围的水肿的发生及进展，对治疗方法进行分析，探索治疗干预对水肿进展的关联关系，其中，（a）小问是对全体患者的水肿体积进行的分析，探索的随时间变化的治疗手段如何影响水肿体积，并求出残差，（b）是需要自行将患者分为 3-5 亚组，然后来求出不同亚组下的患者真实值与预测值之间的残差，（c）小问则是需要分析不同治疗方法对水肿体积进展模式的影响，（d）是分析血肿体积、水肿体积与治疗方法之间的关系；

第三个问题则是需要探索出血性脑卒中患者在预后的预测以及影响预后的关键因素，（a）小问需要根据前 100 个患者的信息及影像构建预测模型来预测所有患者的 90 天 mRS 评分；（b）小问则是在（a）小问的基础信息上，增加后期随访的所有影像信息来预测所有患者在 90 天 mRS 鹅评分；而（c）是要分析出血性脑卒中患者的预后、个人史、疾病史、治疗方法及影像特征等的关联关系，进而需要对临床医学的决策提出建议。

2. 模型假设

1. 假设个人特征，病史及其影像数据的连续变量服从正态性，从而使用皮尔逊相关系数进行相关性分析，并假设相关系数低的变量相互独立。

2. 假设治疗手段是否有效根据首次影像及最后影像的变化对比来确定，血肿及水肿体积减小则判断为有效治疗手段，且不同治疗手段的作用相互独立。

3. 假设研究的血肿、水肿体积变化及 mRS 评分指标只受所给相关个人特征、临床信息、治疗手段和影像信息变量的影响，不受其他因素影响。

4. 由于临床数据和影像数据最早给到首次检查时，假设首次检查数据即为发病当时的原始数据。

3. 问题一：血肿扩张分类模型的建立与求解

3.1. 问题分析

问题一要求我们根据多次影像检查结果的血肿数据，结合发病时间和血肿扩张判断标准，判断 sub001-sub100 患者是否在发病 48 小时内发生血肿扩张，并针对发生血肿扩张的患者，记录其对应的血肿扩张时间距离发病时间的时间间隔。

接下来以发病后 48 小时内是否发生血肿扩张作为响应变量，结合文件中表 1 至表 3 的相关数据，使用 sub1-sub100 患者的数据构建二分类预测模型，并对 sub1-sub160 患者发病后发生血肿扩张的概率进行预测。

3.2. 患者血肿扩张判断分析

针对问题一 a)中所用到的数据，首先将表 1 的“入院首次影像检查流水号”、及“发病到首次影像检查时间间隔”数据、表 2 的各时间检查流水号及对应的 HM_volume 数据、附表 1 的检查时间点及流水号数据，按照 ID 列（患者编号列）进行合并，并根据流水号匹配对应的检查时间。

当检查时间满足：

$$t_i - t_0 + \Delta t \le 48h \tag{1-1}$$

如果检测到的血肿体积变化满足：

$$\begin{gathered} HM\_volume_i - HM\_volume_0 \ge 6ml \\ \text{or} \\ \frac{HM\_volume_i - HM\_volume_0}{HM\_volume_0} \ge 33\% \end{gathered} \tag{1-2}$$

可认为在 48 小时内发生了血肿扩张，并且将 $t_i - t_0 + \Delta t$ 记录为血肿扩张发生时间。其中，t_i 表示第 i 次随访的时间，t_0 表示首次入院影像检查的时间，Δt 表示发病至首次入院影像检查的时间间隔，HM_volume_i 表示第 i 次随访时的血肿体积，HM_volume_0 表示首次入院影像检查的体积。

在对三个表中的数据按照关键字合并时，发现 sub074 的首次检查流水号数据存在不一致的情况，具体如表 3-1 所示。

表 3-1 存在不一致情况的检查流水号表

表 1 中 sub074 患者的首次检查流水号	表 2 中 sub074 患者的首次检查流水号
20180719000630	20180719000020

结合附表 1 的流水号数据，将 sub074 患者的首次检查流水号定为 20180719000020，同时更改表 3 中对应的流水号。在对数据进行分析时，注意到，存下以下数据如表 3-2 所示。

表 3-2 发病到随访 1 时间间隔超过 48 小时的患者数据表

ID	发病到首次影像检查时间间隔	发病到随访 1 时间间隔	发病到随访 2 时间间隔	绝对体积增加 1	绝对体积增加 2	相对体积增加 1	相对体积增加 2
sub015	5.0	168.37	335.46	7.29	-23.98	21.90	-71.97
sub046	1.0	147.62	484.93	-13.07	-38.44	-17.62	-51.83
sub050	9.5	49.10	291.32	-0.83	-9.15	-9.05	-100.00
sub052	4.0	48.90	289.12	36.49	14.42	203.00	80.21

其随访 1 时间与发病的时间间隔超过 48 小时，无法按照上述公式进行计算，需要手动进行判别。判别及修正结果如表 3-3 所示。

表 3-3 需要手动判别的患者血肿数据

ID	是否发生血肿扩张	血肿扩张发生时间	修正后的血肿扩张发生时间
Sub015	0	—	—
Sub046	0	—	—
Sub050	0	—	—
Sub052	1	48.90	48.00

最终得到的 sub1-sub100 患者中，在发病 48 小时内是否发生血肿扩张及扩张时间如表 3-4 所示：

表 3-4 发生血肿扩张的患者数据表

ID	是否血肿扩张	血肿扩张时间	ID	是否血肿扩张	血肿扩张时间	ID	是否血肿扩张	血肿扩张时间
sub003	1	9.52	sub048	1	12.86	sub077	1	14.12
sub005	1	26.47	sub052	1	48	sub079	1	27.85
sub009	1	40.06	sub054	1	16.23	sub080	1	20.57
sub017	1	14.87	sub057	1	14.62	sub081	1	27.42
sub033	1	30.81	sub060	1	23.73	sub092	1	11.92
sub036	1	39.5	sub061	1	6.54	sub095	1	7.43
sub038	1	15.81	sub070	1	9.65	sub098	1	42.76
sub039	1	29.18	sub076	1	15.99	sub099	1	17.67

3.3. 血肿扩张发生情况的分类预测

根据题目要求，将上文得到的是否发生血肿扩张数据作为响应（目标）变量，将题目中要求的患者个人史等变量按照患者编号进行汇总，作为自变量，共得到 73 个初始变量特征。此题目即转化为有标签样本的分类预测问题，或称有监督机器学习。

题目中所给出的有标签样本为患者 sub001-sub100，样本容量 n=100，自变量维度 p=73。自变量维度过高而相应的样本容量过小，直接使用 73 个特征建立模型得到的效果并不好，首先需要通过特征选择对数据特征进行降维处理。

3.3.1. 特征指标降维

本题中的目标变量为是否发生血肿扩张，取值为：0（否）、1（是），为分类变量。本题考虑使用相关性分析的方法，保留与目标变量相关性较高的自变量，从而达到数据降维的目的。针对离散变量，绘制其频数分布条形图如图 3-1 所示。

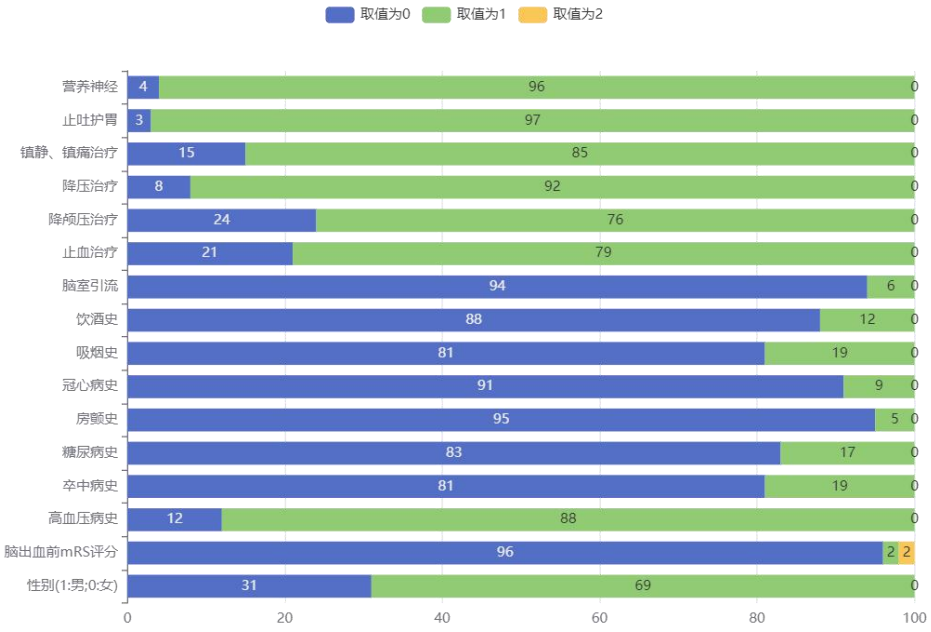


图 3-1 离散指标堆叠条形图

相关性分析中常用的是 Pearson 相关系数。Pearson 相关性分析可以对连续且服从正态分布的数据进行分析，但只能反映变量间有无线性关系。Pearson 相关系数的计算公式如下：

$$\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2\sum_{j=1}^{n}(Y_i-\mu_Y)^2}} \tag{1-3-1}$$

Spearman 相关系数，又称秩相关系数，它适用范围相对较广，既可以对连续数据进行相关性分析，也可以针对离散数据进行相关性分析；而且它不仅仅只刻画变量间的线性相关关系，也可以刻画变量间的非线性相关关系。

Spearman 相关系数的基本步骤是将数据 $X_i(i=1,2,\ldots,n)$ 分别从小到大进行排序，得到秩数 $R_i(i=1,2,\ldots,n)$，并计算出秩平均数 $\overline{R}$。再分别用秩数 $R_i(i=1,2,\ldots,n)$ 替代上式中的 $X_i(i=1,2,\ldots,n)$，用秩平均数 $\overline{R}$ 替代 μ_X，依此类推，得到整理后的 Spearman 相关系数计算公式：

$$r_s=\frac{12\sum_{i=1}^{n}R_iQ_i-3n(n+1)^2}{n(n^2-1)} \tag{1-3-2))}$$

考虑到本题中的目标变量为分类变量，且有部分自变量也为分类变量，故对自变量与

目标变量使用 Spearman 相关性分析，保留的显著性 p<0.1 的变量。最终保留相关系数显著的 7 个变量，依次为：房颤史（p=0.002）、冠心病史（p=0.001）、饮酒史（p=0.038）、HM_ACA_R_Ratio（p=0.10）、ED_PCA_R_Ratio（p=0.09）、ED_Cerebellum_R_Ratio（p=0.1）、original_shape _Elongation（p=0.014）。

这 7 个变量与相应变量的 Spearman 相关系数热力图如图 3-2 所示。

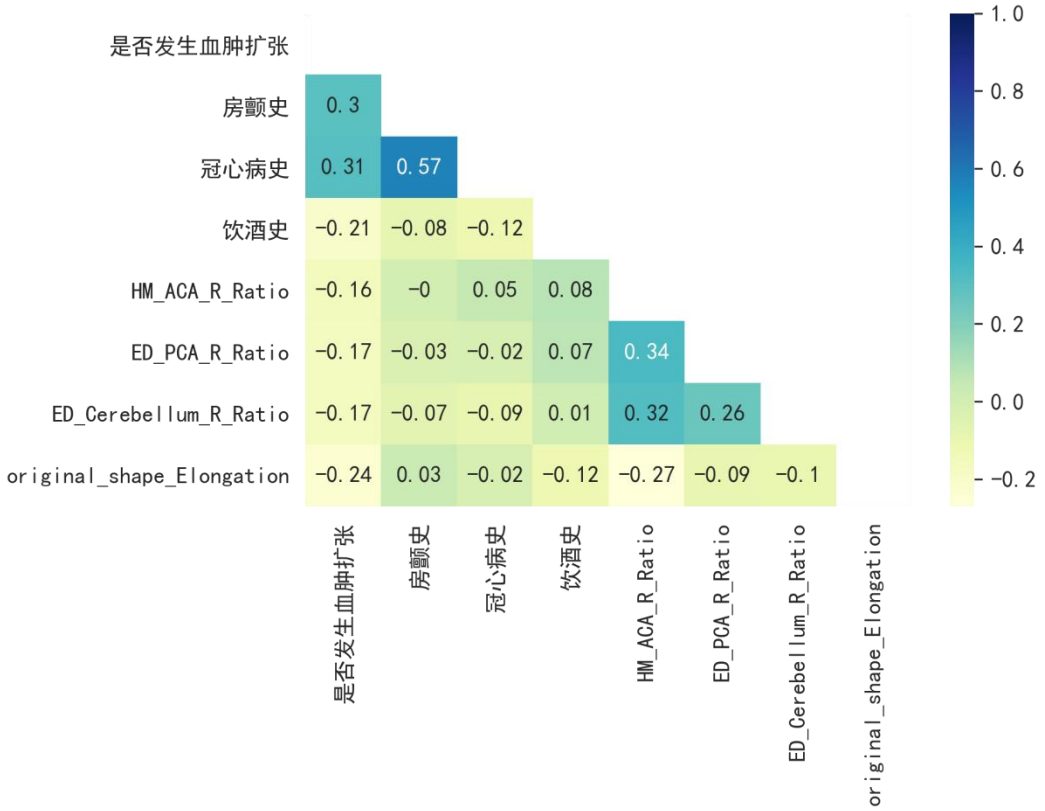


图 3-2 Spearman 相关系数热力图

从图中可以看出，与是否发生血肿扩张相关系数最高的是冠心病史（0.31）、其次是房颤史（0.3）。将是否发生血肿扩张作为响应变量，使用与响应变量相关系数较高的 7 个自变量建立分类预测模型

3.3.2. 血肿扩张的随机森林分类预测模型建立

使用 Spearman 相关性分析保留的 7 个相关性较高的变量，将 sub001-sub100 患者的样本数据，按照 7：3 的比例切分训练集和测试集，建立随机森林分类模型，并选用遗传算法进行启发式寻优，设置的初始种群个数为 50、最大迭代次数为 150、变异概率为 0.01、交叉概率为 0.5。随机森林得到的训练参数如表 3-5 所示。

表 3-5 随机森林训练参数表

参数名	参数值

训练用时	0.181s
数据洗牌	是
节点分裂评价准则	gini
决策树数量	100
有放回采样	true
划分时考虑的最大特征比例	auto
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阀值	0

从表 1-5 中可以看出，模型训练用时 0.181s，树的最大深度为 10，叶子节点的最大数量为 50。随机森林分类模型得到的特征重要性大小如图 3-3 所示。

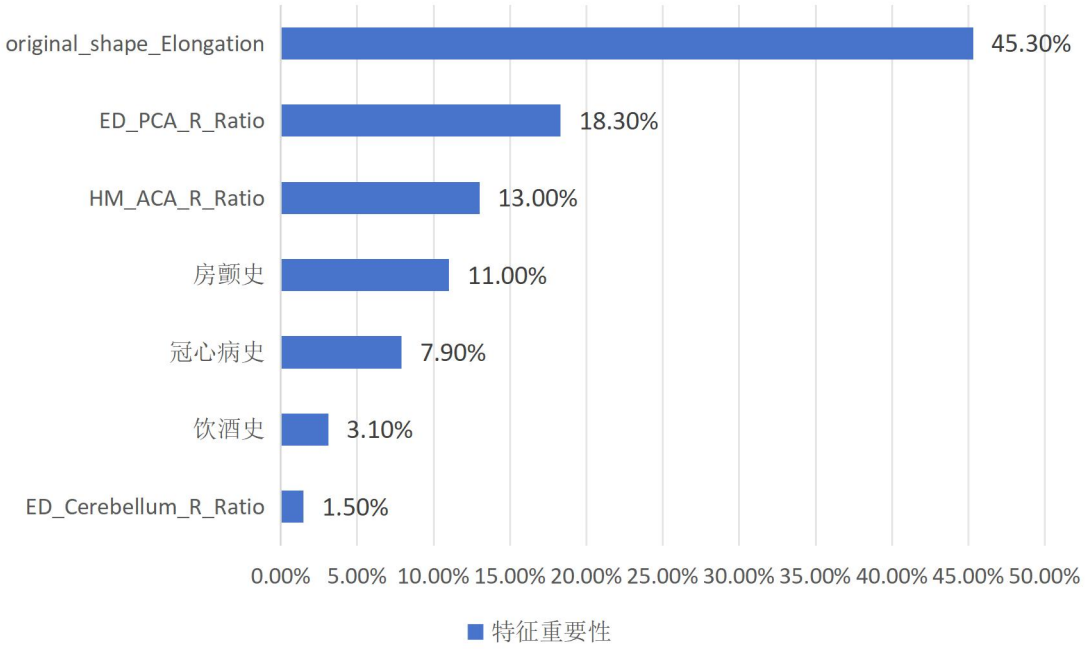


图 3-3 随机森林特征重要性图

从图 3-3 中可以看出，特征重要性从大到小排序依次为 original_shape_Elongation（45.30%）、ED_PCA_R_Ratio（18.30%）、HM_ACA_R_Ratio（13.00%）、房颤史（11.00%）、冠心病史（7.90%）、饮酒史（3.10%）、ED_Cerebellum_R_Ratio（1.50%）。模型在测试

集上的混淆矩阵如图 3-4 所示。

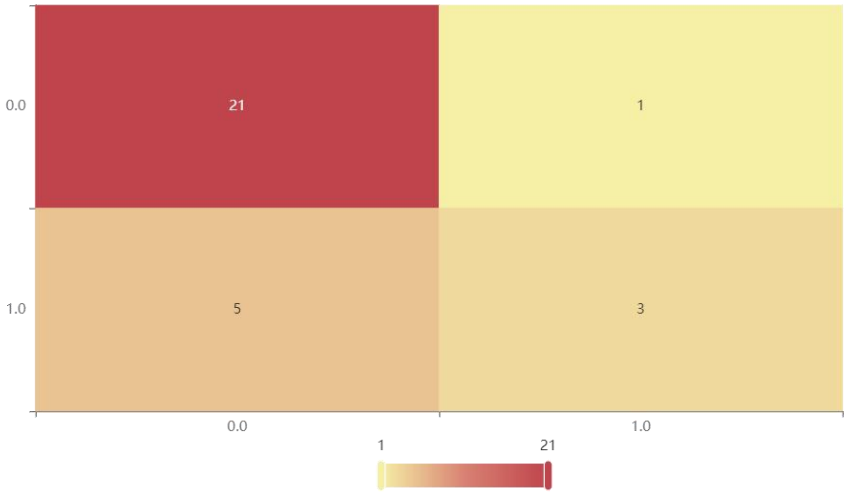


图 3-4 随机森林测试集的混淆矩阵图

从图中可以看出，模型在随机选取的 30 个测试集样本上，预测正确了 24 个，其中有 21 个属于未发生血肿扩张的样本，有 3 个为发生血肿扩张的样本；预测错误了 6 个，将 5 个原本发生血肿扩张的样本预测为了未发生血肿扩张，将 1 个原本未发生血肿扩张的样本预测为了血肿扩张。

3.3.3. 血肿扩张的 bp 神经网络分类预测模型建立

选用相关系数较高的 7 个特征作为自变量，将分类模型更改为 bp 神经网络模型，并选用遗传算法进行启发式寻优，设置的初始种群个数为 50、最大迭代次数为 150、变异概率为 0.01、交叉概率为 0.5。bp 神经网络得到的训练参数如表 3-6 所示。

表 3-6 bp 神经网络训练参数表

参数名	参数值
训练用时	0.05s
数据洗牌	是
激活函数	identity
求解器	lbfgs
学习率	0.1
L2 正则项	1
迭代次数	1000
隐藏第 1 层神经元数量	100

从表中可以看出，bp 神经网络模型训练用时 0.05s，与随机森林模型相比训练用时更短，模型采用的激活函数为 identity。bp 神经网络模型在测试集上的混淆矩阵如图 3-5 所示。

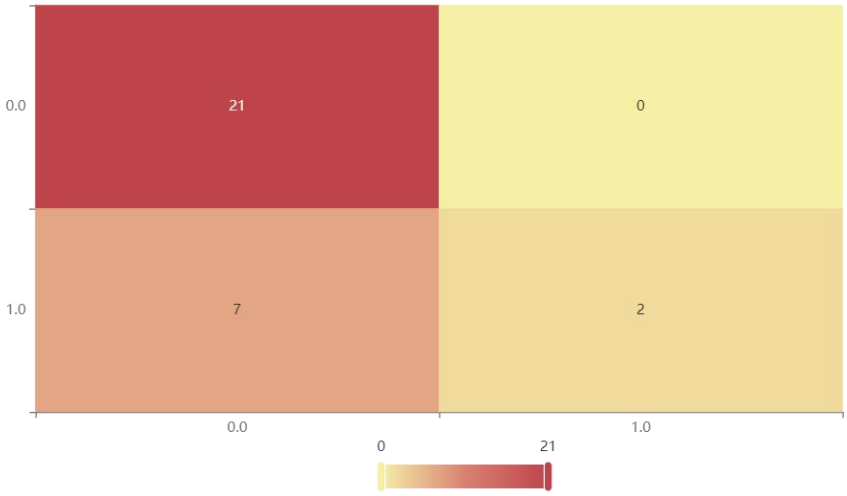


图 3-5 bp 神经网络模型测试集的混淆矩阵图

从图中可以看出，模型在随机选取的 30 个测试集样本上，预测正确了 23 个，其中有 21 个属于未发生血肿扩张的样本，有 2 个为发生血肿扩张的样本；预测错误了 7 个，将 7 个原本发生血肿扩张的样本预测为了未发生血肿扩张。

3.3.4. 血肿扩张的支持向量机分类预测模型建立

选用相关系数较高的 7 个特征作为自变量，将分类模型更改为支持向量机分类模型，并选用遗传算法进行启发式寻优，设置的初始种群个数为 50、最大迭代次数为 150、变异概率为 0.01、交叉概率为 0.5。支持向量机得到的训练参数如表 3-7 所示。

表 3-7 支持向量机训练参数表

参数名	参数值
训练用时	0.026s
数据洗牌	是
惩罚系数	1
核函数	linear
核函数系数	scale
核函数常数	0
核函数最高项次数	3
误差收敛条件	0.001

参数名	参数值
最大迭代次数	1000
多分类融合策略	ovr

从表中可以看出，支持向量机的训练用时为 0.026s，与随机森林模型（0.181s）、bp 神经网络模型（0.05s）相比，训练速度更快支持向量机分类模型在测试集上的混淆矩阵如图 3-6 所示。

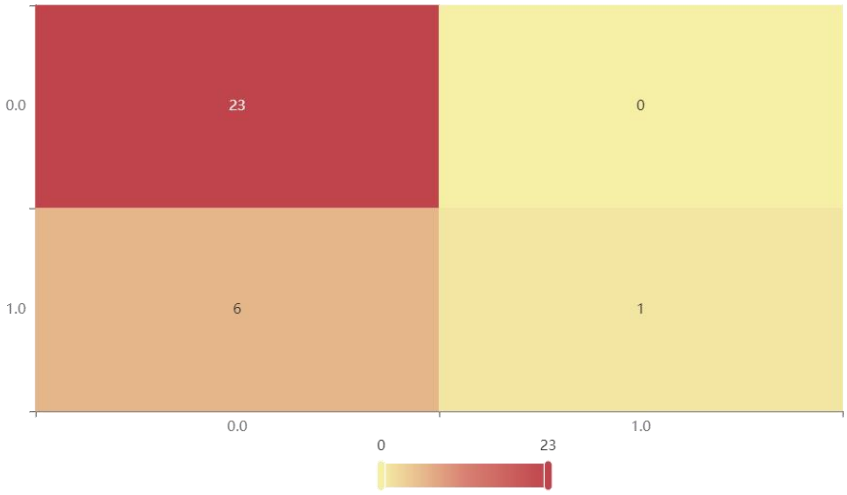


图 3-6 支持向量机测试集的混淆矩阵图

从图中可以看出，模型在随机选取的 30 个测试集样本上，预测正确了 24 个，其中有 23 个属于未发生血肿扩张的样本，有 1 个为发生血肿扩张的样本；预测错误了 6 个，将 6 个原本发生血肿扩张的样本预测为了未发生血肿扩张。

3.3.5. 不同模型分类效果的比较

常见的评价分类效果好坏的指标有：准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 值。

准确率的计算公式为：

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \tag{1-3-3}$$

精确率的计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{1-3-4}$$

召回率的计算公式为：

$$Recall = \frac{TP}{TP + TN} \tag{1-3-5}$$

F1 值的计算公式为：

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall} \tag{}$$

其中，*TP*、*FN*、*FP*、*TN* 的含义如表 3-8 所示。

表 3-8 判断混淆矩阵表

	实际为“1”	实际为“0”
预测为“1”	TP	FP
预测为“0”	FN	TN

针对随机森林模型、bp 神经网络模型、支持向量机分类模型，分别计算模型在训练集和测试集上的准确率、精确率、召回率、F1 值，用以比较模型分类预测的好坏，各模型的指标值如表 3-9 所示。

表 3-9 不同模型分类效果比较表

模型		准确率	召回率	精确率	F1
随机森林	训练集	1	1	1	1
	测试集	0.8	0.8	0.792	0.775
bp 神经网络	训练集	0.814	0.814	0.804	0.769
	测试集	0.767	0.767	0.825	0.709
支持向量机	训练集	0.786	0.786	0.779	0.731
	测试集	0.8	0.8	0.841	0.737

从上表中可以看出，在训练集上，随机森林的分类效果要优于 bp 神经网络和支持向量机。在测试集上，随机森林、支持向量机的准确率和召回率均为 0.8、0.8，优于 bp 神经网络的 0.767、0.767；从精确率来看，支持向量机的精确率最高。综合考虑，使用 F1 值作为评判标准，随机森林的 F1=0.775，优于 bp 神经网络的 F1 值（F1=0.709）和支持向量机的 F1 值（F1=0.737）。基于此，选择随机森林分类模型对患者是否发生血肿扩张进行分类预测应该能达到较好的分类效果。

3.4. 使用随机森林对患者是否血肿扩张进行分类预测

采用训练好的随机森林模型对 sub001-sub160 患者是否发生血肿扩张进行预测，得到的分类预测结果和预测概率如表 3-10 所示。

表 3-10 血肿扩张预测表

ID	预测结果	预测未发生血肿扩张的概率	预测发生血肿扩张的概率	是否发生血肿扩张（实际值）
sub001	0	0.96	0.04	0
sub002	0	0.98	0.02	0
sub003	0	1	0	1
sub004	0	0.797	0.21	0
sub005	0	0.98	0.02	1
sub006	0	0.95	0.05	0
sub007	0	0.98	0.02	0
sub008	0	0.93	0.08	0
sub009	1	0.16	0.84	1
sub010	0	0.61	0.39	0

4. 问题二：水肿进展预测及关联分析

4.1. 问题分析

本题的有四个小任务，分别需要处理的小问题是：

（1）要根据“表 2”中给出的前 100 个患者水肿体积和重复的检查时间点，来形成并构建出所有患者的水肿体积随着时间发展状况的一条曲线，这里要求我们计算出这部分患者的拟合值与真实值之间存在的残差情况。

（2）要求探索出患者的水肿体积随着时间发展过程中，患者之间存在的个体性差异，将这些患者划分为不同的组合类别，记录这些不同组合里的患者水肿体积，在随着时间进展过程时的图像曲线变化，与第一小问一样，我们需要记录并计算前一百名患者的真实值与曲线之间的残差。

（3）需要分析求得不同的治疗方法对水肿的体积进展模式的影响。

（4）继续分析：血肿的体积、水肿的体积以及治疗方法三者之间存在的关系。

问题二中主要存在的**难点**在于：

（1）本题中的患者本身身体素质和周遭环境等多方面的影响，会对随着时间增长而出现的部分数据产生偏差，因此我们需要合理的进行数据的分析和处理，才能得到效果较

好的模型结果和预测值。

（2）题目中要求我们来探索出患者的水肿体积随着时间变化时，存在的一些差异，由于专业限制，我们仅能通过有限的论文资料查询，相关的专业人士的文献研读，才能较为明确的选择并聚类到同一类型中的患者人群。

4.2. 血肿周遭水肿的发生及进展

本题中需要先将表格 2 中的测试集前一百个患者的重复检查时间点以及他们的水肿体积，分别作为坐标轴中的横坐标和纵坐标。因此，构造出问题 a 中的流程图如图 4.2 所示：

图 4.2 血肿周遭水肿扩张发生处理流程图

4.2.1. 数据的预处理

数据的预处理，通常就是要将数据进行分析，并在建模之前，就要将原始的数据就进行一些数据的清洗、筛选、转化等的过程，其目的也是为了提高数据的质量和模型建立的便易。数据预处理的得到，将会使得数据在后续的分析和建模中，得到更高效率的处理和更贴合需求的结果。

通常情况下，对数据进行预处理，会包括对数据进行清洗、转换、集成和规约四个方面，针对不同的数据特征，以及不同的任务目标的需求，我们可以对数据进行不同的处理和完善。

将数据进行处理，使得 x 轴为每次检查时的时间差值，y 轴为对应的水肿的体积大小，在数据处理过程中，完善不健康的数据情况，将明显不合理的数据指标进行异常值处理，将个体超出正常情况的数据也进行修正。

而在该小问中，我们对数据的预处理方面包括有：

（1）删除异常值以及替换异常值两部分。对于异常值数量比较少的，我们就考虑直接删除掉这些包含异常值的数据，再来分析数据，而对部分的异常值，我们也可以考虑替换掉这部分的异常数据，可通过该部分数据的中位数、平均数等来替换。

（2）处理掉一些重复的数据值，如果数据中重复的部分占比过大，可以考虑删除这部分的重复数据，若是重复数据占总的数据量较大的时候，我们对数据进行合并，可以考虑将两个相同的数据信息进行合并。

4.2.3 线性回归建立拟合曲线

根据整理的数据情况，进行数据的多项式拟合回归，进而来预测分析得到的拟合曲线与真实值之间的残差。建立模型：

$$f(x) = p1 * x^3 + p2 * x^2 + p3 * x + p4 \qquad (4\text{-}1)$$

其中，$p1 = -1.799 \times 10^{-4}; p2 = 2.169 \times 10^{-2}; p3 = -1.957 \times 10^{-1}; p4 = 2.566 \times 10^{-1}$。

计算出的拟合曲线的预测值与真实值之间的残差平方和为6.266×10^7，而计算出的RMSE为0.9255，如图4.2.1所示。

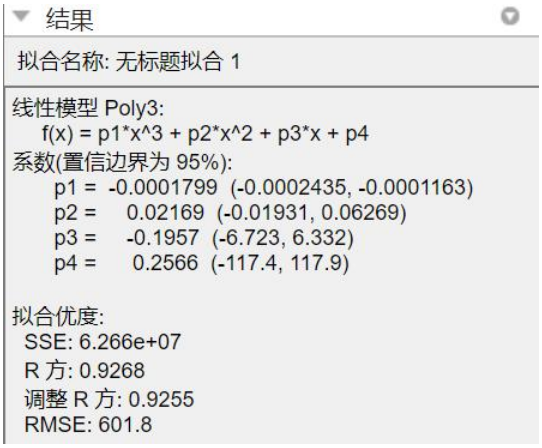


图 4.2.1 线性预测拟合方程结果图

其运行结果如下图所示，使用matlab的曲线拟合出的散点图（图4.2.2）及对应曲线，得出的模型，残差图如图4.2.3如下所示：

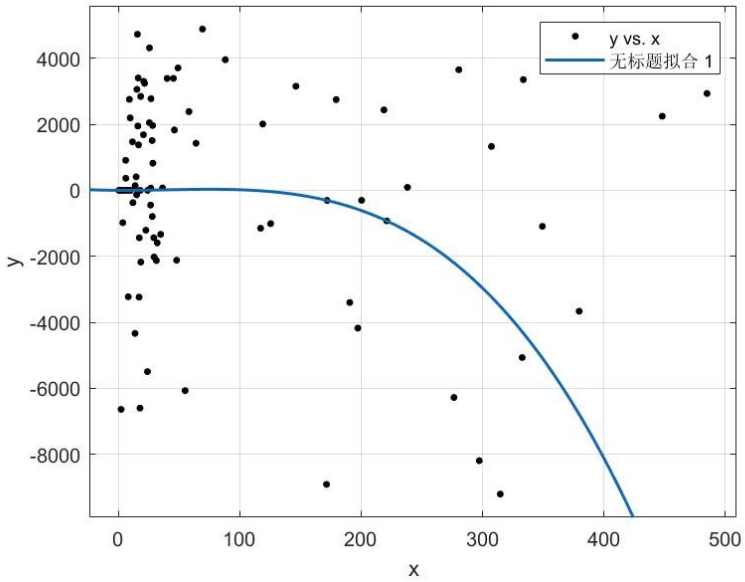


图 4.2.2 水肿体积随时间变化散点图及线性拟合曲线

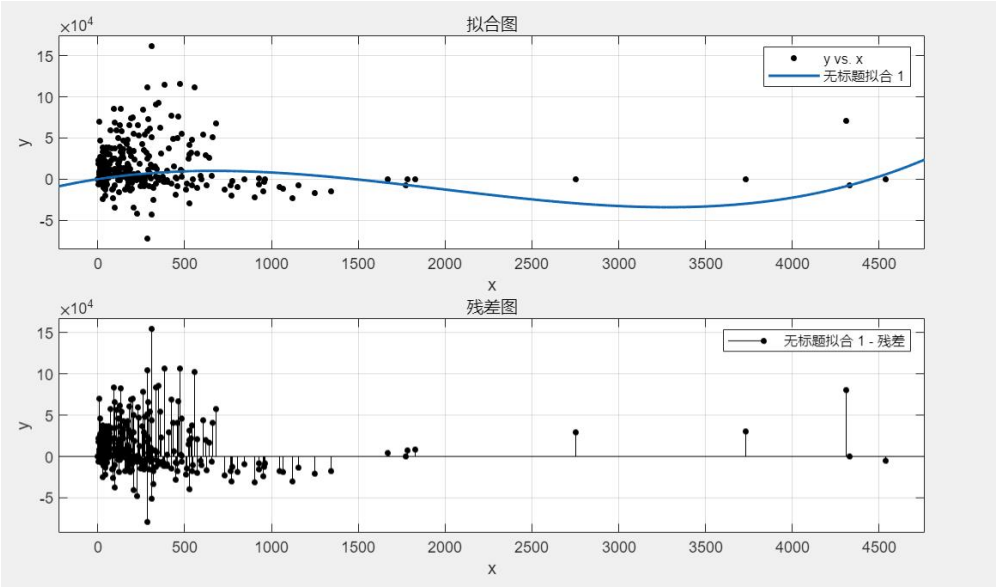


图 4.2.3 水肿体积随时间变化残差图对比

4.3 按患者信息分亚组来进行预测

4.3.1 按患者信息进行分组的思路

根据问题要求，首先提取出患者的个体指标（表 1 的 E-P 列），经过分析，“血压”指标和“发病到首次影像检查时间间隔”具有较大随机性，不属于个体差异指标范畴，将表 2.2.1 的 E-W 列确定为个体差异指标。

由于个体差异指标中既包含连续型变量，如年龄；又包含性别等分类变量，而 k-means

等常见的聚类方法针对的是连续型变量，无法使用于有离散变量的情形。两步聚类法用于聚类的变量既可以是连续变量，也可以是离散变量，故本文采用二步聚类法对sub001-sub100 患者的个体差异指标进行聚类分析。考虑到离散变量的存在，选取对数似然作为距离测量标准，设置最大聚类数目为 51，应用 BIC 准则评判聚类效果。

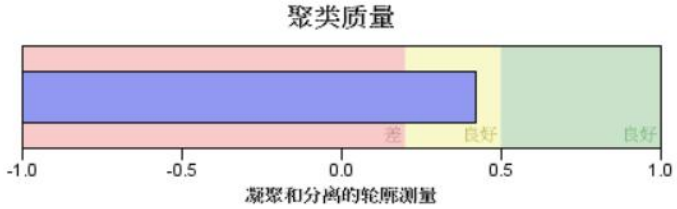


图 4.3.1 聚类质量图

从图 4.3.1 中可以看出，两步聚类的凝聚和分离的轮廓测量的分数在 0.4 左右，聚类质量较好。各聚类类别占比如图 4.3.2 所示。

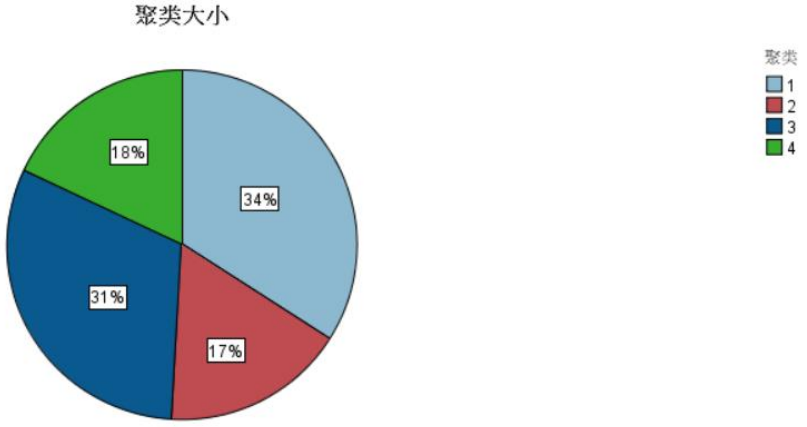
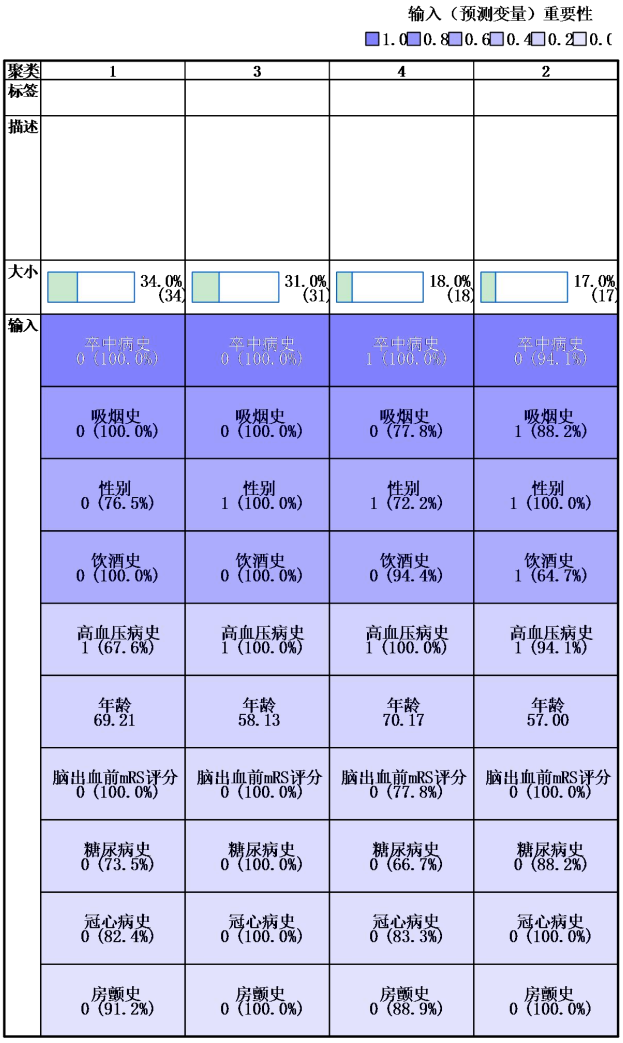


图 4.3.2 聚类大小图

从图 4.3.2 中可以看出，聚类类别为 1 的样本占比 34%，占比最多，有 34 个样本；聚类类别为 2 的样本占比 17%，有 17 个样本；聚类类别为 3 的样本占比 31%，有 31 个样本；聚类类别为 4 的样本占比 18%，有 18 个样本。最大聚类与最小聚类的比为 2。

得到的各聚类类别的变量信息如图 4.3.3 所示。



聚类	1	3	4	2
标签				
描述				
大小	34.0% (34)	31.0% (31)	18.0% (18)	17.0% (17)
输入	卒中病史 0 (100.0%)	卒中病史 0 (100.0%)	卒中病史 1 (100.0%)	卒中病史 0 (94.1%)
	吸烟史 0 (100.0%)	吸烟史 0 (100.0%)	吸烟史 0 (77.8%)	吸烟史 1 (88.2%)
	性别 0 (76.5%)	性别 1 (100.0%)	性别 1 (72.2%)	性别 1 (100.0%)
	饮酒史 0 (100.0%)	饮酒史 0 (100.0%)	饮酒史 0 (94.4%)	饮酒史 1 (64.7%)
	高血压病史 1 (67.6%)	高血压病史 1 (100.0%)	高血压病史 1 (100.0%)	高血压病史 1 (94.1%)
	年龄 69.21	年龄 58.13	年龄 70.17	年龄 57.00
	脑出血前mRS评分 0 (100.0%)	脑出血前mRS评分 0 (100.0%)	脑出血前mRS评分 0 (77.8%)	脑出血前mRS评分 0 (100.0%)
	糖尿病史 0 (73.5%)	糖尿病史 0 (100.0%)	糖尿病史 0 (66.7%)	糖尿病史 0 (88.2%)
	冠心病史 0 (82.4%)	冠心病史 0 (100.0%)	冠心病史 0 (83.3%)	冠心病史 0 (100.0%)
	房颤史 0 (91.2%)	房颤史 0 (100.0%)	房颤史 0 (88.9%)	房颤史 0 (100.0%)

图 4.3.3 聚类类别变量信息图

两步聚类的各变量的重要性如图 4.3.4 所示。

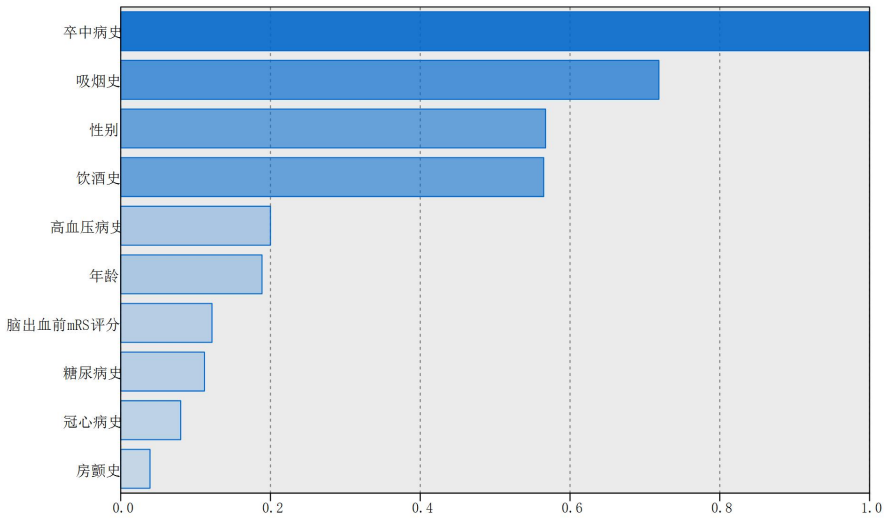


图 4.3.4 两步聚类变量重要性图

变量重要性从大到小一次为：卒中病史、吸烟史、性别、饮酒史、高血压病史、年龄、脑出血前 mRS 评分、糖尿病史、冠心病史、房颤史。

根据两部聚类结果将患者数据聚类为 4 类，将聚类类别 1 对应于亚组 1，依此类推，从而构建出 4 组不同人群（分亚组）。接下来将聚类标签加入水肿体积数据中，按照不同人群绘制水肿体积散点图。

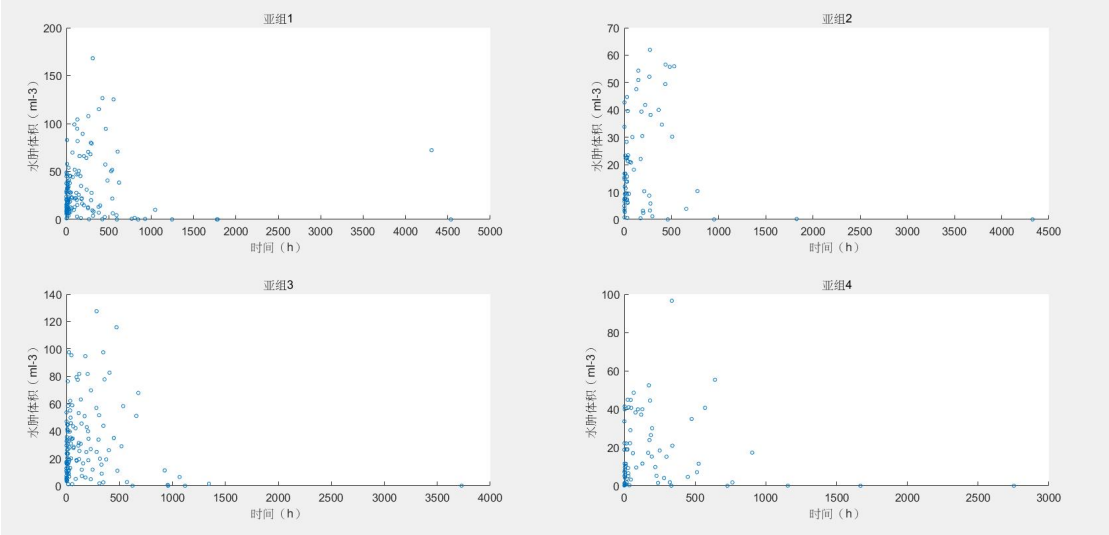


图 4.3.5 不同亚组的水肿体积散点图

针对不同亚组的水肿体积数据，使用 matlab 的 cftools 工具包进行拟合，综合比较线性拟合、多项式拟合、指数拟合、高斯拟合等多种拟合方式，发现高斯拟合得到的拟合效果相对较好。最终选择高斯拟合，得到 4 个亚组的得到的拟合效果如图 4.3.6-图 4.3.9：

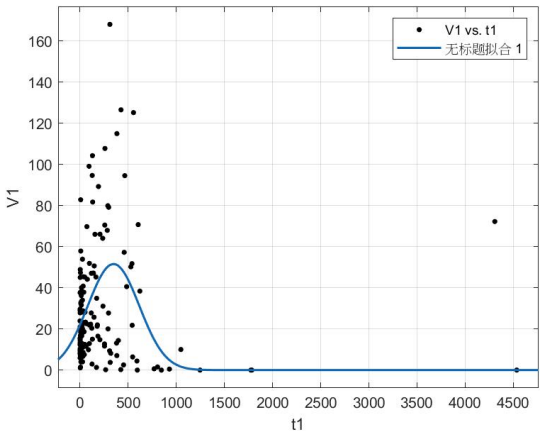


图 4.3.6 亚组 1 拟合曲线图

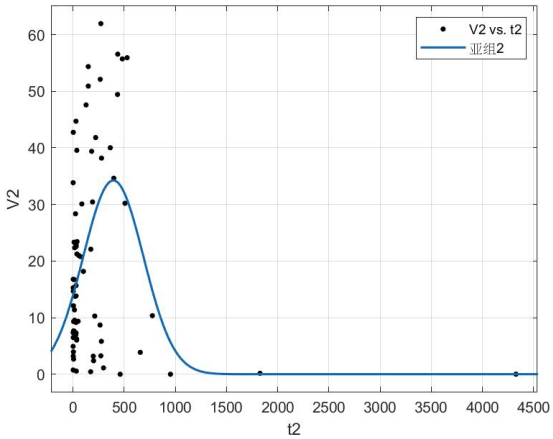


图 4.3.7 亚组 2 拟合曲线图

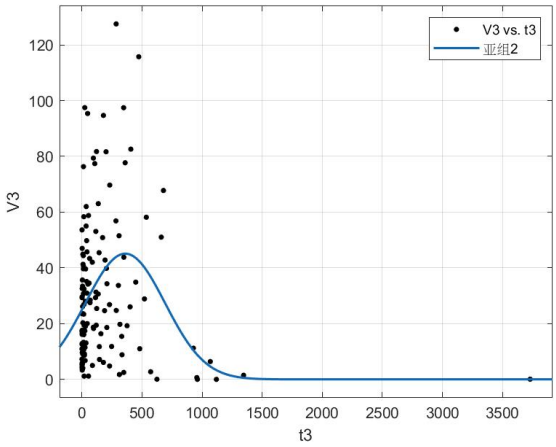


图 4.3.8 亚组 3 拟合曲线图

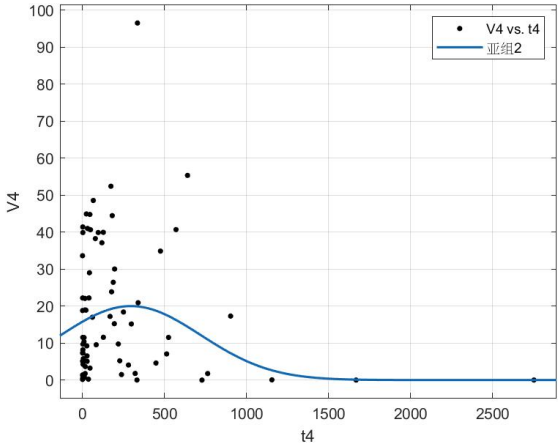


图 4.3.9 亚组 4 拟合曲线图

高斯拟合的通用方程为：

$$f(x) = a * \exp^{-\frac{(x-b)^2}{c}} \qquad (4\text{-}2\text{-}1)$$

其中，根据不同的数据，拟合得到的参数 a,b,c 的值也会不一样。针对不同亚组数据进行高斯拟合后，得到的模型系数如表 4.3.4 所示。

表 4.3.4 分亚组高斯拟合系数表

系数	亚组 1	亚组 2	亚组 3	亚组 4
a	51.58	34.23	45.1	19.99
	(41.08, 62.09)	(25.94, 42.52)	(35.73, 54.47)	(12.04, 27.94)
b	348.6	393.4	360.6	294.8
	(272.7, 424.5)	(269.8, 517.1)	(248.1, 473.1)	(49.72, 539.8)
c	377.9	416.3	468.9	604.6
	(263.9, 491.9)	(253.5, 579.1)	(290.4, 647.4)	(-4.995, 1214)

注：括号内相应地为该系数的 95%置信区间。

将表 4.3.4 中的数据带入公式 4-2-1 中，即可得到该亚组数据的高斯拟合方程。根据拟合值与数据真实值之间的差异，进一步可计算出拟合方程的评价指标如表 4.3.6 所示。

表 4.3.5 分亚组高斯拟合效果表

所属亚组	SSE	R 方	调整 R 方	RMSE
亚组 1	1.173*105	0.1407	0.129	28.25
亚组 2	1.74*104	0.221	0.199	15.65
亚组 3	8.289*104	0.1292	0.1163	24.78

亚组 4	2.036*104	0.1728	0.1495	16.93

根据不同亚组的高斯拟合方程和真实水肿体积值，计算出残差如表 4.3.6 所示。

表 4.3.6 分亚组拟合残差表

ID	残差（亚组）	所属亚组	ID	残差（亚组）	所属亚组
sub001	84546.8687	1	sub006	70009.60646	3
sub002	22014.01874	3	sub007	39120.56569	3
sub003	41200.29988	3	sub008	33480.68707	1
sub004	13955.39679	4	sub009	29826.38609	4
sub005	27797.09092	1	sub010	14249.07741	2

4.4 分析不同治疗方法对水肿体积进展模式的影响

4.4.1 使用不同治疗方法对水肿体积发展的思路

因为脑水肿是脑出血后必然出现的继发性脑损伤之一，因此分析不同治疗方法对水肿体积进展模式的时候，也是对于分析出反应脑出血血肿扩张的拓展。

在本小问中，因为表中未指出治疗方法的加入时间，因此我们不能明确知道源数据表 1 中的七种治疗方法：脑室引流、止血治疗、降颅压治疗、降压治疗、镇静及镇痛治疗、止吐护胃、营养神经时，率先考虑观察这部分数据的特征，发现除过脑室引流外，其他的六种治疗手段用的频率都在 70%以上，而水肿的体积大小，这里仅考虑患者在表中记录的最后一次复查中时的水肿体积与最初时刻检查的水肿体积的差值。

根据有限的时间数据信息，使用一定程度的数据预处理，将前一百名患者在医院的最后一次复查出的水肿体积大小与最初的那个检查出来的水肿体积大小做出差值，然后通过认定只要水肿体积差值为负值，则认为是该名患者的治疗方法对水肿的治疗起了作用。

通过分类判定，最后将得到一系列可用来做判别的数据（如下表 4.4.1 仅显示部分数据）：

表 4.4.1 脑水肿体积和治疗方法判别表

ID	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛	止吐护胃	营养神经	设定为判别值
sub001	0	1	1	1	1	1	1	0
sub002	0	1	1	1	0	1	1	0
sub003	0	1	1	1	1	1	1	1
sub004	0	1	1	1	0	0	0	1
sub005	0	1	1	0	0	1	1	1
sub006	1	0	0	1	0	1	1	0
sub007	0	0	1	1	0	1	1	0
sub008	0	1	0	1	0	0	1	1
sub009	0	1	1	1	0	1	1	1
sub010	0	1	0	1	1	1	1	0
sub011	0	0	0	1	1	1	1	1
sub012	0	1	1	1	0	1	1	0
sub013	0	1	0	0	1	1	0	0
sub014	0	1	1	1	0	0	1	1
sub015	0	0	1	1	0	1	0	1
sub016	0	1	1	1	1	1	1	1
sub017	0	1	0	1	1	1	1	1

4.4.2 使用随机森林分类判别预测

对该部分数据使用随机森林分类判别预测，进而显示出模型的特征值（治疗手段）的重要性程度（下图的柱形图 4.4.1 展示了各个特征或自变量的重要性，在整个治疗中占到的比例）如下：

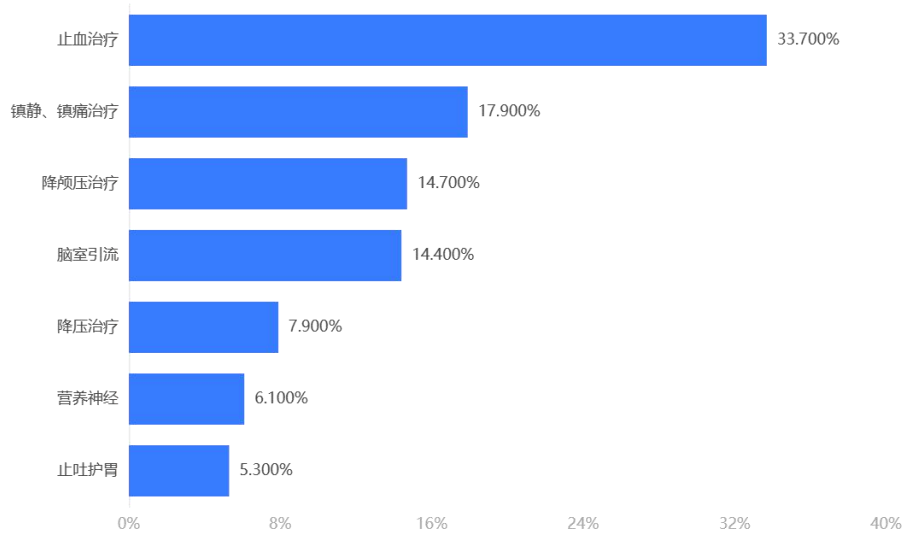


图 4.4.1 治疗手段的特征重要性程度

通过分析，进一步得到矩阵的热力图，可以用来可视化分类模型的性能表现，热力图能将分类模型预测结果与实际结果进行对比，并将结果用深浅不一的颜色的矩阵的形式呈现出来，并在最终以可以通过观察混淆矩阵热力图，得到模型的分类的情况，分类的数据以及会误分类的数量等指标，都可以在热力图中直观的观察到。该随机森林分类的热力图如下图 4.4.2 所示：

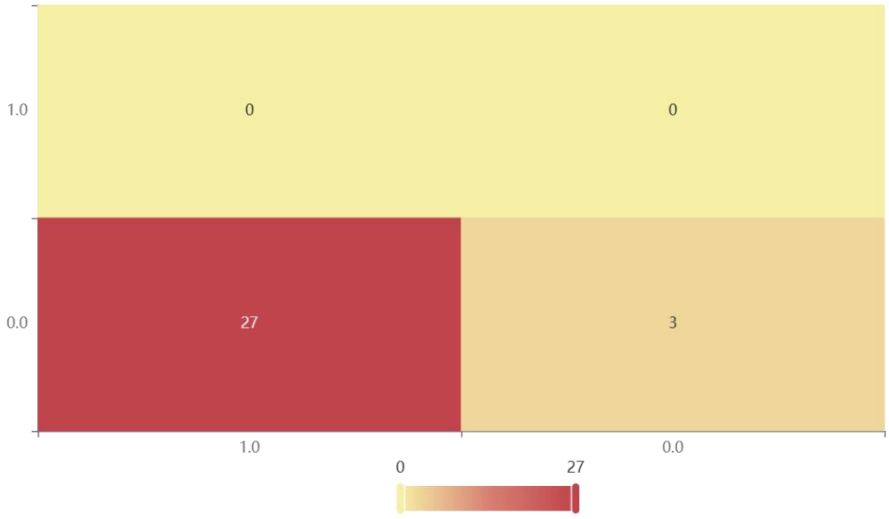


图 4.4.2 随机森林分类热力图

下面的表格表示，训练集和测试集的分类的评价的指标，通过对量化的指标的衡量，我们能得到对训练、测试的分类效果，而很明显在上图中，我们的精确率能达到 76.2%。准确程度较好，因此可以认为上述重要性排在前面的治疗手段对水肿体积的减少结果有着正向的帮助。

表 4.4.2 随机森林分类效果评价表

	准确率	召回率	精确率	F1
训练集	0.757	0.757	0.762	0.721
测试集	0.1	0.1	1	0.182

将部分的预测结果显示在下表 4.4.3 中：

表 4.4.3 随机森林预测判定值表

预测结果 Y	设定为判别值	预测测试结果概率_0.0	预测测试结果概率_1.0	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1
1.0	0.0	0.15	0.85	0	1	0	1	1	1	1
0.0	0.0	0.98	0.02	0	0	1	1	1	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1

1.0	0.0	0.15	0.85	0	1	0	1	1	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1
1.0	0.0	0.22	0.78	0	0	1	1	0	1	1
0.0	0.0	0.52	0.48	0	1	1	1	0	1	1
1.0	0.0	0.12	0.88	0	1	0	0	1	1	0
1.0	0.0	0.18	0.82	0	1	1	0	0	1	1
1.0	0.0	0.32	0.68	0	1	0	1	0	1	1
1.0	0.0	0.27	0.73	0	1	1	1	1	1	1

4.4.3 不同治疗方法对水肿体积的进展的影响分析

由上述分析可得出，在使用治疗方法是止血治疗，镇静镇痛治疗，以及降颅压治疗和脑室引流时，都会对结果产生较好的影响。且这四种治疗手段都会对水肿体积的减小产生较好的影响。而其他的治疗手段虽然有限，但也有一定的重要性程度和作用。

4.5 血肿体积、水肿体积和治疗方法的关系

4.5.1 三者关系的思路

在上一问中，我们已经得到了水肿体积与治疗方法的关系，因此，只要再将血肿的体积与治疗方法分析，将血肿体积和水肿体积的关系分析，就能够得到这三者之间的相互关系了。

与前文的思路一致，同样取每位患者在最后一次复查中的血肿体积大小与最初的血肿体积大小作差，得到了一个血肿的变化值，同样将为负的变化量认定为是治疗有效的部分，将正的变化量认为是治疗无效的部分，得到一系列的判别数据与治疗方法放一起。在这里我们运用决策树，得到数据将用来判别预测治疗方法对血肿体积的大小判定。

4.5.2 决策树来得出血肿体积和治疗方法的关系

因为决策树的分类也是基于准确率、召回率、准确率。因此通过建立决策树来得到不同的治疗手段在对降低血肿体积的重要性特征占比，如下表 4.5.2 所示：

4.5.2 决策树判别参数表

参数名	参数值

训练用时	0.007s
数据切分	0.7
数据洗牌	否
交叉验证	否
节点分裂评价准则	gini
特征划分点选择标准	best
划分时考虑的最大特征比例	None
内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
叶子节点的最大数量	50
树的最大深度	10
节点划分不纯度的阀值	0

从决策树结构的特点中能看出，内部的节点显示出了被分支特征的具体的切分情况，要根据这些特征中的部分某个分值，得到对水肿体积减小具有较好影响的治疗手段，进而将这些治疗手段结合成一种有效的治疗方法来参与水肿体积的判断中。如下图 2.4.2 所示，就是各个治疗手段的重要性的占比。

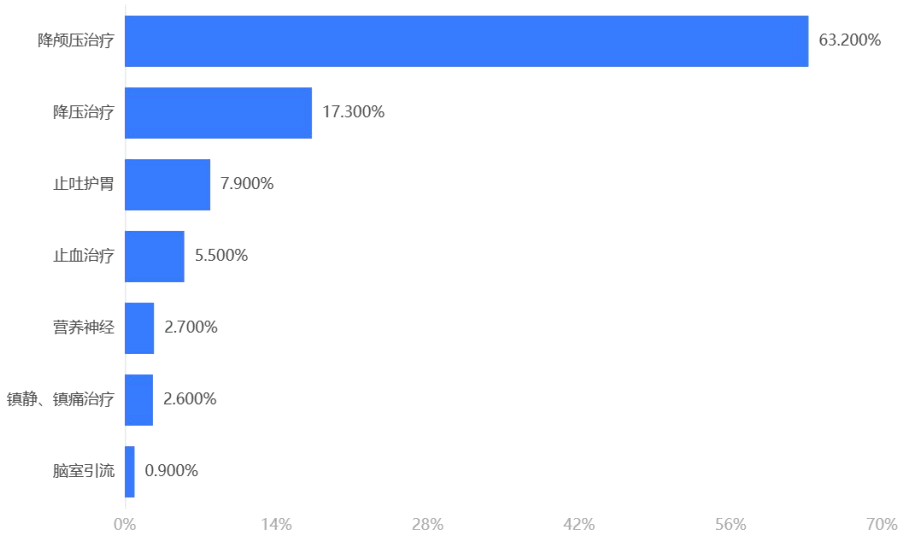


图 4.5.1 决策树判别治疗手段的占比

由决策树建立的模型预测出来的训练集和测试集的准确率如下表 4.5.2 所示：

表 4.5.2 决策树模型预测的准确率等的概率表

	准确率	召回率	精确率	F1
训练集	0.886	0.886	0.868	0.871
测试集	0.8	0.8	0.8	0.8

显示出来的准确率由 0.886，准确率较好，因此认为由较好的模型表现情况。

由该模型得到的数据预测情况，也部分展示如下表 4.5.3 所示：

表 4.5.3 决策树预测判别结果表

预测结果 Y	血肿判别	预测测试结果概率_0.0	预测测试结果概率_1.0	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
1.0	1.0	0	1	0	0	0	0	1	1	1
0.0	0.0	0.5	0.5	0	0	0	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0	1	0	0	0	0	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0	1	0	1	0	1	1	1	1
1.0	1.0	0	1	0	1	0	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	0.0	0	1	0	0	1	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0.125	0.875	0	1	1	1	1	1	1
1.0	1.0	0	1	1	1	1	1	1	1	1

表 2.4.3

由降颅压治疗为主要的治疗手段，添加降压治理和止吐护胃、止血治疗，认为这几个治疗手段对血肿体积的扩散具有较好的效果。

4.5.4 卡方检验血肿体积和水肿体积的关系

用上述的思路和预处理出来的数据进行分析，得到的是血肿体积变化量与水肿体积变化量，将这部分的数据量卡方检验，进而得到判定这两部分的结果如下表 4.5.1 所示：

表 4.5.1 卡方检验血肿、水肿体积差异性表

题目	名称	血肿判别		总计	检验方法	X^2	P
		0.0	1.0				
水肿判别	0.0	13	40	53	pearson 卡方检验	10.382	0.001***
	1.0	1	46	47			
合计		14	86	100			

注：***、**、*分别代表 1%、5%、10%的显著性水平

由此可知，血肿的变化量与水肿的变化量不存在一致性，而是存在一定的显著的差异。

将患者原始的血肿体积和水肿体积进行 Spearman 相关性检验，得到的数据检验效果结果如下表 4.5.2：

表 4.5.2 血肿、水肿体积相关性表

	HM_volume	ED_volume
HM_volume	1(0.000***)	0.626(0.000***)
ED_volume	0.626(0.000***)	1(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

因此由上表 4.5.2 可得出血肿体积与水肿体积存在着大约为 0.626 的相关性。

5. 问题三：患者预后预测及关键因素探索

5.1 问题分析

本题一共三小问，分别需要针对患者的不同特征情况、治疗手段及影像数据对患者进

行预后预测并且找出关键的影响因子。

a.需要结合患者的个人特征和首次影像结果，挑选自变量，对衡量患者之后功能状态的 mRS 指标构建拟合模型，在找出影响患者未来功能显著因素的同时，对 mRS 的值进行预测。

b.需要将随访的影像数据也纳入考虑范围，结合随时间推移的影像数据，进一步确认关键因素去逼近真实的 mRS 值。

c.分析 mRS 评分与个人特征、病历、治疗手段、脑部血肿水肿位置，及影像特征关联性，确定影响患者未来功能 mRS 评分的关键性因素并且找出合理的临床对策。

问题 3 的难点在于：

1）所给的 mRS 变量是等级分类数据，而所给特征、疾病症状及影像信息多为连续性的变量，需要采用机器学习的方法去对其拟合。

2）等级变量个数较多，在构建模型分析时，结果的容错率较低，泛化能力弱，很容易过拟合或者欠拟合，在测试集中展现较差的适应性。

3）自变量指标过多，且量纲差距较大，需要考虑进行归一或分类处理，对指标降维。从而提高模型参数的显著性，避免过拟合。

5.2 模型预测与关键因素探索

5.2.1 数据处理

对极端变量、缺失值、异常值进行处理，对表三的随访影像数据与表一二的基本信息及病情做流水号的匹配处理，由于后两次随访的影像数据缺失较多，仅保留前两次随访的数据来支持后续模型的构建。

5.2.2 自变量选择

（1）自变量独立性筛选

为避免多重共线性的影响，我们对自变量进行相关性检验。首先我们对于自变量进行独立性筛选，通过对表一年龄、表二血肿与水肿位置信息及表三影像信息的定量变量进行了正态性检验，QQ 图显示这些数据较好满足正态性假定，于是我们采用皮尔逊相关系数

对这些变量进行相关性检验，结果显示表二中水肿的位置与体积与其对应的血肿的体积有显著的相关性，其中左侧中动脉血肿率与水中率、左侧大脑后动脉血肿率和水肿率、右侧中动脉血肿率与水中率、右侧大脑后动脉血肿率和水肿率的皮尔逊相关系数分别为 0.921、0.934、0.863、0.8，并且通过 0.05 水平下的显著性置信水平。表三的图像数据也展现出很强的关联性，通过筛选和分析，我们保留了包括 Entropy、Rang 等 12 个水肿和血肿的影像指标。

	年龄	性别	高血压病史	卒中病史	糖尿病史	冠心病史	吸烟史	饮酒史	高压	低压	HM_volume	HM_ACA_R_Ratio	HM_MCA_R_Ratio	HM_PCA_R_Ratio	HM_Pons_Medulla_R_Ratio	HM_Cerebellum_R_Ra	HM_ACA_L_Ratio	HM_MCA_L_Ratio	HM_PCA_L_Ratio	HM_Pons_Medulla_L_Ratio	HM_Cerebellum_L_Rati
me	0.028(0.780)	0.118(0.243)	0.108(0.286)	0.209(0.037**)	81)	0.114(0.257)	0.183(0.068*)	0.178(0.076*)	47)	51)	00***)	00***)	93)	0.021(0.833)	0.029(0.776)	1)	48**)	69)	85)	0.026(0.794)	60)
ED_ACA_R_Ratio	-0.025(0.807)	0.01(0.918)	-0.157(0.120)	-0.019(0.852)	0.025(0.801)	-0.023(0.818)	0.079(0.436)	0.033(0.742)	-0.001(0.996)	0.003(0.972)	0.323(0.001***)	0.736(0.000***)	0.557(0.000***)	0.587(0.000***)	0.169(0.092*)	0.269(0.007***)	-0.053(0.603)	0.553(0.000***)	0.329(0.001***)	-0.005(0.959)	0.107(0.288)
ED_MCA_R_Ratio	-0.243(0.015**)	0.09(0.372)	0.087(0.391)	0.078(0.442)	0.129(0.201)	-0.048(0.637)	0.08(0.431)	0.051(0.616)	0.064(0.530)	0.086(0.393)	-0.021(0.837)	0.237(0.018**)	0.921(0.000***)	0.584(0.000***)	-0.004(0.972)	0.026(0.795)	-0.48(0.000***)	0.829(0.000***)	0.666(0.000***)	-0.217(0.030**)	-0.129(0.202)
ED_PCA_R_Ratio	-0.167(0.096*)	0.071(0.480)	0.015(0.880)	0.179(0.075*)	-0.068(0.503)	-0.02(0.840)	0.085(0.402)	0.069(0.496)	-0.024(0.814)	-0.016(0.872)	-0.077(0.444)	0.343(0.000***)	0.676(0.000***)	0.883(0.000***)	0.216(0.031**)	0.202(0.044**)	-0.29(0.003***)	0.799(0.000***)	0.461(0.000***)	0.001(0.996)	-0.06(0.555)
ED_Pons_Medulla_R_Rati	-0.317(0.001***)	0.145(0.149)	0.172(0.088*)	0.035(0.731)	0.014(0.893)	-0.054(0.595)	-0.002(0.983)	0.024(0.810)	-0.11(0.276)	0.021(0.834)	-0.153(0.129)	0.187(0.062*)	0.175(0.082*)	0.457(0.000***)	0.536(0.000***)	0.047(0.645)	-0.024(0.813)	-0.323(0.001***)	0.032(0.756)	0.404(0.000***)	0.158(0.116)
ED_Cerebellum_R_Ratio	-0.079(0.436)	0.033(0.744)	-0.01(0.920)	-0.048(0.636)	-0.04(0.691)	0.093(0.359)	-0.046(0.650)	0.008(0.938)	0.019(0.853)	-0.027(0.788)	-0.151(0.133)	0.316(0.001***)	0.078(0.440)	0.343(0.000***)	0.39(0.000***)	0.456(0.000***)	-0.076(0.452)	-0.135(0.180)	-0.018(0.863)	0.206(0.040**)	0.297(0.003***)
ED_ACA_L_Ratio	0.056(0.581)	-0.072(0.476)	0.002(0.987)	-0.138(0.171)	-0.007(0.943)	-0.065(0.523)	-0.11(0.277)	0.006(0.956)	0.058(0.566)	0.037(0.717)	0.276(0.005***)	-0.031(0.757)	-0.621(0.000***)	-0.51(0.000***)	0.027(0.790)	-0.043(0.669)	0.737(0.000***)	0.617(0.000***)	0.663(0.000***)	0.294(0.003***)	0.121(0.230)
ED_MCA_L_Ratio	0.168(0.095*)	-0.056(0.580)	0.012(0.905)	-0.113(0.264)	-0.073(0.468)	-0.029(0.773)	-0.039(0.704)	-0.014(0.888)	0.041(0.686)	-0.026(0.798)	0.045(0.657)	-0.316(0.001***)	0.782(0.000***)	0.732(0.000***)	-0.133(0.188)	-0.116(0.249)	0.294(0.003***)	0.934(0.000***)	0.569(0.000***)	0.061(0.544)	0.079(0.434)
ED_PCA_L_Ratio	0.135(0.180)	0.016(0.873)	0.047(0.645)	-0.034(0.740)	-0.084(0.406)	-0.09(0.372)	0.047(0.641)	0.002(0.983)	0.079(0.437)	0.091(0.369)	0.113(0.265)	-0.161(0.109)	0.772(0.000***)	-0.537(0.000***)	0.053(0.598)	-0.046(0.647)	0.452(0.000***)	0.738(0.000***)	0.8(0.000***)	0.376(0.000***)	0.245(0.014**)
ED_Pons_Medulla_L_Rati	-0.118(0.242)	-0.067(0.508)	0.009(0.929)	-0.065(0.524)	0.197(0.049**)	-0.121(0.232)	0.018(0.858)	0.024(0.809)	0.037(0.716)	0.1 (0.322)	0.027(0.788)	0.019(0.855)	-0.443(0.000***)	-0.142(0.159)	0.313(0.001***)	0.019(0.848)	0.385(0.000***)	0.274(0.006***)	0.618(0.000***)	0.589(0.000***)	0.275(0.006***)

图 5.2.1 自变量相关性图

（2）因变量显著性筛选

为更有效对自变量进行降维，我们通过 spearman 相关系数探索第一步所筛选的自变量是否与因变量显著相关，保留显著水平 1%以上的显著因素。最终得到了包括年龄、糖尿病史、血肿体积、大脑左侧前动脉血肿率、血肿极差等 12 个的相关变量。

表 5.2.1 Spearman 相关性表

年龄	糖尿病史	冠心病史	饮酒史	HM_volume	HM_ACA_Ratio	HM_ACA_L_Ratio	edNCCT_original_firstorder_Entropy	edNCCT_original_firstorder_Range	edNCCT_original_firstorder_Uniformity	hmNCCT_original_firstorder_Entropy	hmNCCT_original_firstorder_Uniformity

9	0.2	0.2	0.2	-0.	0.3	0.18	0.2	0.224(0.	-0.203(0	-0.226(0.	0.211(0.	-0.206(0.
0	35(	99(	92(	268	66(	2(0.	64(	025**)	.042**)	024**)	035**)	039**)
天	0.0	0.0	0.0	(0.	0.0	069	0.0					
m	19*	02*	03*	007	00*	*)	08**					
R	*)	**)	**)	***)	**)		*)					
S												

5.2.3 模型探索与预测

由于因变量是等级序列变量，本文尝试了神经网络、支持向量机及随机森林的模型回归及分类。并且利用包括模拟退火和遗传算法的启发式算法对寻求模型的更优解，结果显示随机森林分类模型结合遗传算法能在更好地拟合模型。在训练集和测试集均有较好的表现。

刚开始我们由于分类等级是有序列，并且代表递增地代表严重程度，故而我们先用神经网络、支持向量机及随机森林三个方法对其进行回归分析，但是测试集的表现并不好，产生的估计值是连续的，需要采用四舍五入等方法对其进行再分类，误差较大。

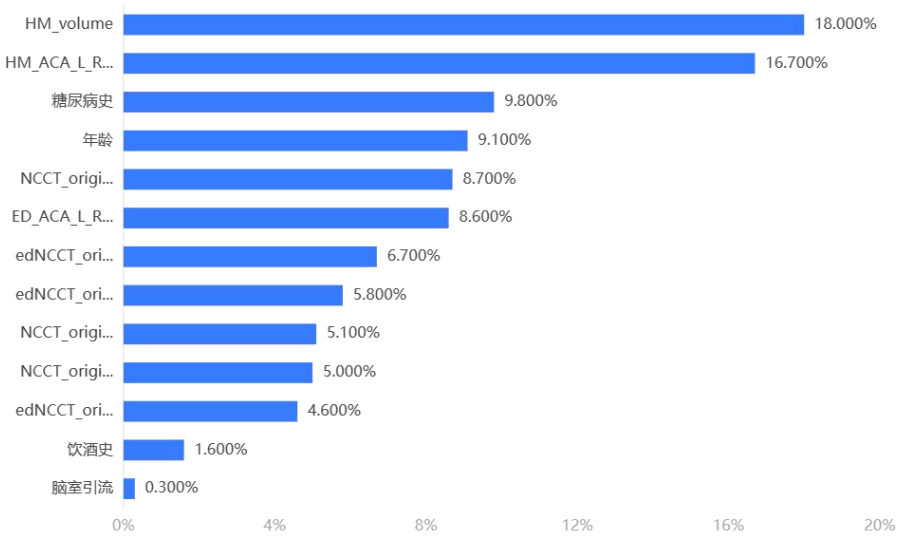


图 5.2.2 筛选后指标占比

表 5.2.2 测试集的误差表

	MSE	RMSE	MAE	MAPE	R^2

	MSE	RMSE	MAE	MAPE	R^2
训练集	0.415	0.644	0.552	28.986	0.856
测试集	1.836	1.355	1.053	44.69	0.31

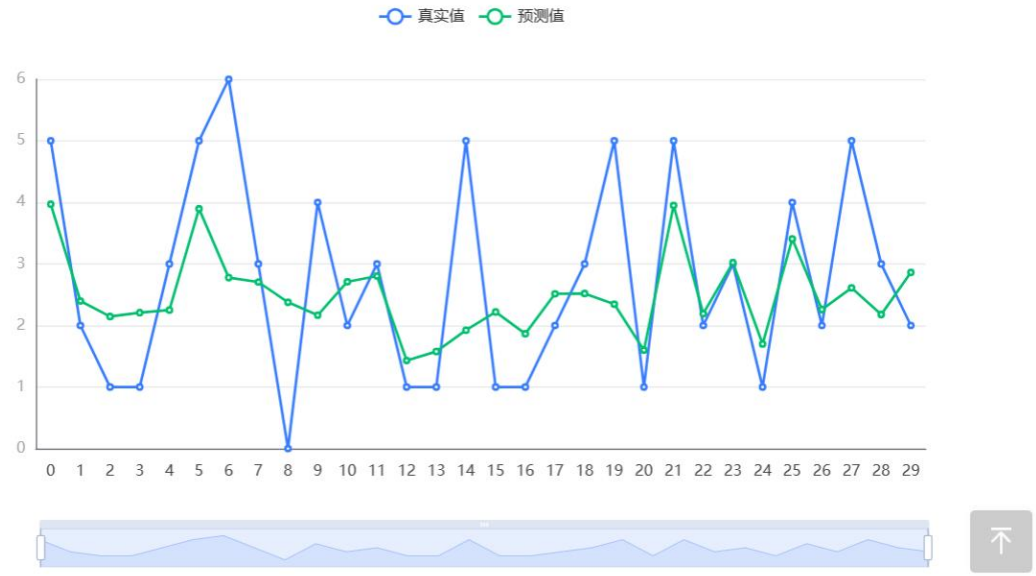


图 5.2.3 预测曲线与真实曲线的折线图

于是我们还是采用分类的思想使用上述三个模型对其进行分类估计，并且采用了遗传算法，结果表明测试集和训练集有较好的表现，实现了有效预测。

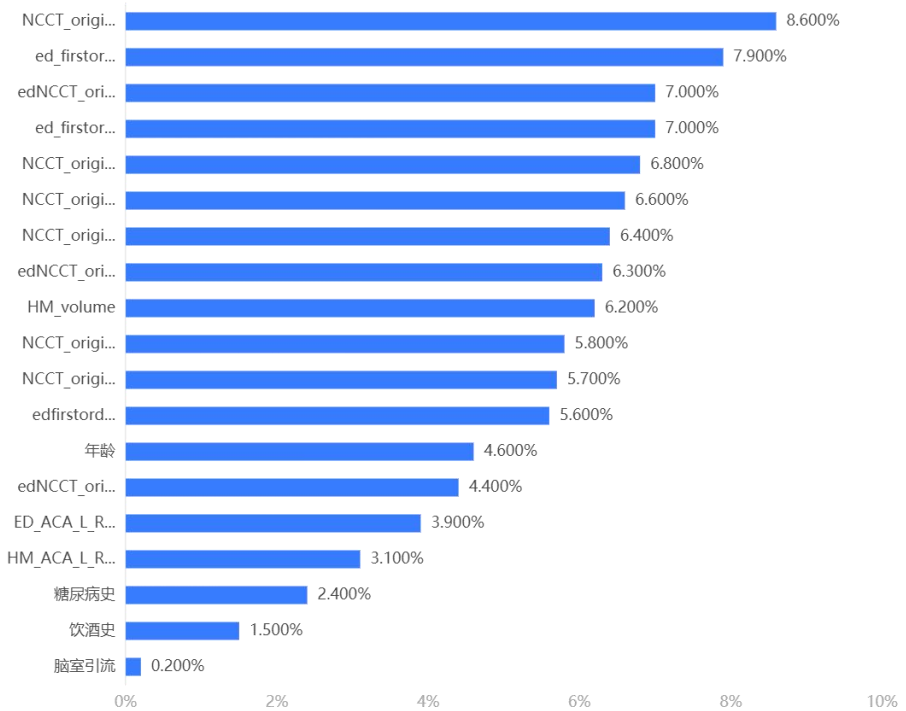


图 5.2.4 不同指标的预测占比

通过结果我们可以看到所筛选变量与随访影响信息均与最终评分关系显著，这也可以说明变量选择的有效性。

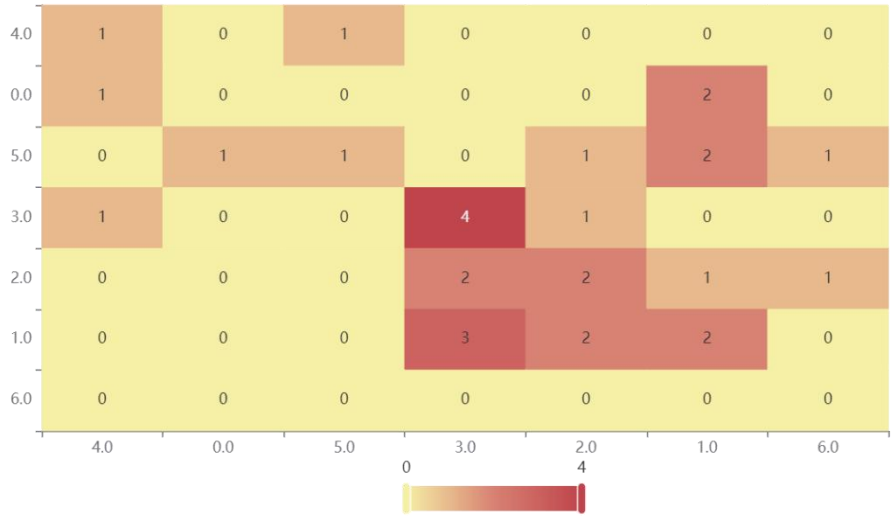


图 5.2.5 筛选后指标之间的相关性热力图

热力图表明虽然存在误差但总体的准确性还是可以保证的，通过下面的图我们也可以看到，拟合的结果还是较好的。

表 5.2.3 热力图反应的准确概率表

	准确率	召回率	精确率	F1
训练集	1	1	1	1
测试集	0.633	0.633	0.644	0.617

5.3 分析出血性脑卒中患者预后与其他因素的关联

由上述等的分析过程中，经过对数据的预处理等的操作后，整理出患者的预后与个人史、疾病史、治疗方法以及影像的特征，患者的影像特征中，包含着患者的血肿、水肿的体积、位置、信号强度特征、形状特征等的内容，因此在后续的分析过程里，将会对患者的各个指标进行相关性分析（Pearson 相关分析、灰色关联分析），将相关性高的指标进行预处理，然后将相关性高的重复指标进行剔除，剔除的指标应是相比较对目标影响较小

的，最后对剩下的指标与患者的预后进行分析，得出最后的关系。

5.3.1 患者预后指标与其他因素的相关性

在将数据进行预处理的过程中发现，患者的信息与患者的影像特征指标信息有着部分的较高相关性的联系，因此在对数据进行一定程度的筛选和优化，将已经经过了上述的指标筛选后，进一步对数据进行灰色关联性的分析，得到了有关于数据的低相关性的指标量，将这部分的数据进行与出血性脑卒中的患者的预后指标 RMS 的指标一起相关性分析，得到了数据的相关性如下所示：

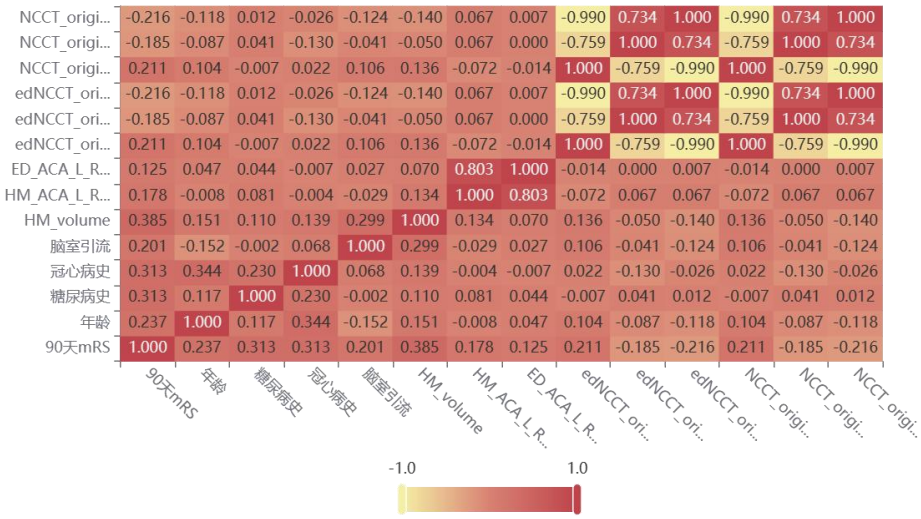


图 5.3.1 第一次筛选后的指标相关性热力图

由上述图可以知道，虽然指标间仍然存在着一定的相关性，但这些高相关性的指标与研究目标：预后 90 天 mRS 的相关性并不大，因此将这些指标进行岭回归，这些经过了进一步筛选后的指标的相关性已经极低，因此将这些低相关性的指标与预后指标进行岭回归分析，得到如下所示：

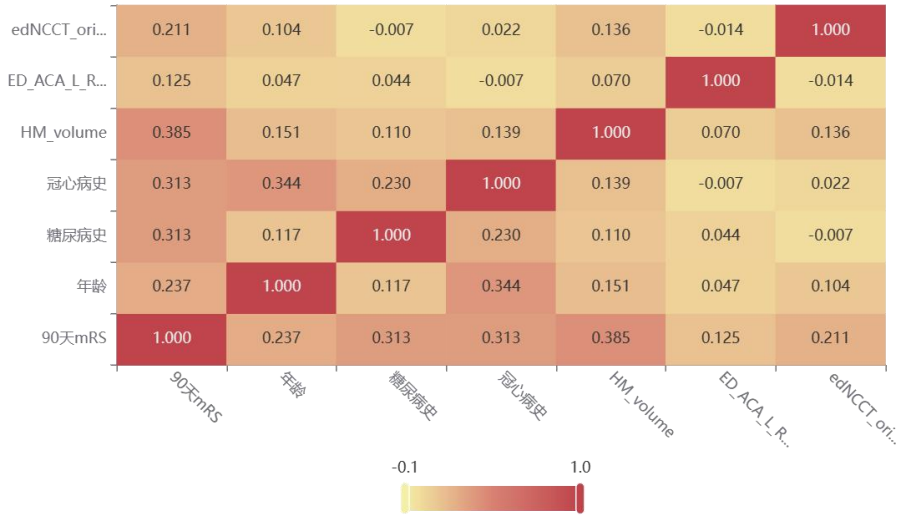


图 5.3.1 第二次筛选后的指标相关性热力图

5.3.2 患者预后指标与其他相关因素的回归分析

将这部分的指标对应的数值与其他因素回归处理，得到的模型涉及到的变量情况如下：

变量 X:{年龄，糖尿病史，冠心病史，脑室引流，HM_volume，HM_ACA_L_Ratio，ED_ACA_L_RatioedNCCT_original_firstorder_Entropy,edNCCT_original_firstorder_Range,edNCCT_original_firstorder_Uniformity，NCCT_original_firstorder_Entropy，NCCT_original_firstorder_Range，NCCT_original_firstorder_Uniformity}；

因变量 Y:{90 天 mRS}

将这部分的数据进行一定的建模，得到的模型公式为：

$$Y = 1.801 + 0.68X_1 + 0.922X_2 + 0.916X_3 + 2.502X_4 - 0.238X_5 + 0.173X_6 - 0.003X_7 - 2.918X_8 + 0.173X_9 - 0.003X_{10} - 2.918X_{11} + 0.818X_{12}$$

其中：

Y是90天mRS；X_1是年龄；X_2是糖尿病史;X_3是冠心病史;X_4是HM_ACA_L_Ratio;
X_5是 ED_ACA_L_Ratio;X_6是edNCCT_original_firstorder_Entropy
X_7是edNCCT_original_firstorder_Range;X_8是edNCCT_original_firstorder_Uniformity;
X_9是NCCT_original_firstorder_Entropy;X_{10}是NCCT_original_firstorder_Range;
X_{11}是 NCCT_original_firstorder_Uniformity;X_{12}是脑室引流

岭回归的做出的模型结果图如下 5.3.2 所示：

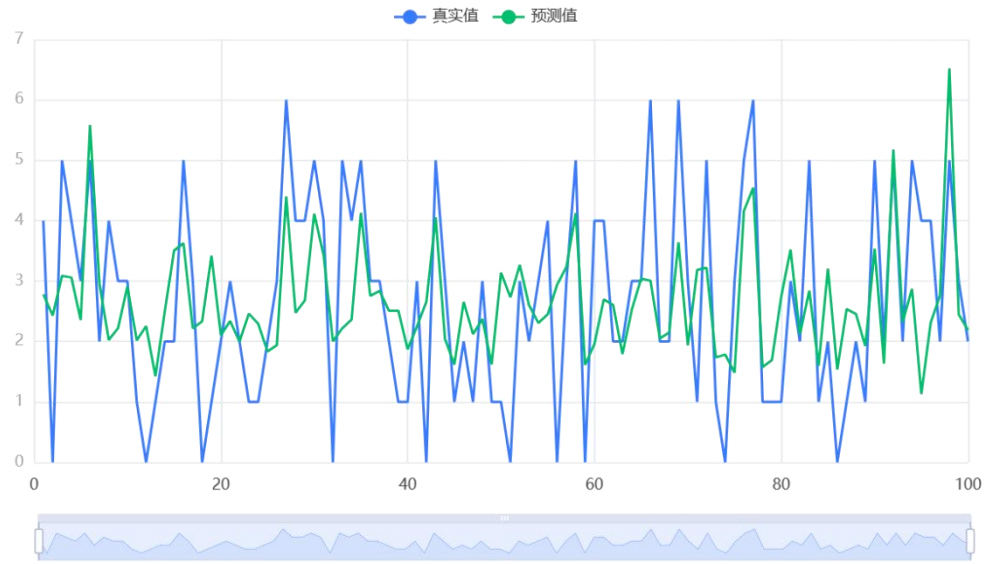


图 5.3.2 岭回归真实值与预测值的曲线拟合图

上图以可视化的形式展示出了本次用岭回归模型时的原始数据与模型拟合值所构成的图像。试用模型来进行结果的预测，得到的一个结果如下所示：

表 5.3.2 岭回归模型预测结果

变量	系数	测试值
常数	1.8006736164024921	1
年龄	0.11218358243247341	
糖尿病史	0.20606740944937188	
冠心病史	0.1560386138898756	
HM_volume	0.22314430611364428	
HM_ACA_L_Ratio	0.14645497177113176	
ED_ACA_L_Ratio	-0.020836105248052662	
edNCCT_original_firstorder_Entropy	0.020308206335156086	
edNCCT_original_firstorder_Range	-0.0359720966226415	
edNCCT_original_firstorder_Uniformity	-0.033980303894933304	
NCCT_original_firstorder_Entropy	0.020308206335156016	
NCCT_original_firstorder_Range	-0.03597209662264163	

NCCT_original_firstorder_Uniformity	-0.03398030389493329
脑室引流_1.0	0.11558652902842013
预测结果 90 天 mRS	1.801

表 5.3.2 是用来对岭回归模型进行预测的，数据检测值与数据测试值差距不大，在测试值设为 1 的时候，预测的结果 y 值为 1.801，与测试值结果较为贴合，因此认定这个模型预测效果较好。

5.3.3 患者预后指标与其他相关因素的关联分析

由本题的分析中得出，患者的预后指标与患者的个人史、疾病史、治疗方法及影像特征都存在着一定程度的相关性，但并没有决定性的关键因素，因此在本小问的求解中，我们使用了回归分析来建模关联到患者的预后指标。通过查阅一定的资料显示得出了结果。

5.3.4 模型结论及预后建议

通过相关性分析中血肿与水肿的数据相关性，我们可以得出血肿所引起的临床症状会对水肿有一定的影响，这说明血肿引起的渗透压变化会继而引发水肿，[1]，进而引发其他隐患，因此在临床上及时控制血肿至关重要。根据自变量和因变量的相关性检测及岭回归结果，我们可以看出年龄、糖尿病等疾病史都是对预后有较大影像的特征因素，故而优化习惯、合理饮食结构等都是对预后至关重要的。并且结果显示，这些患者最终的 90 天 mRS 预测值多集中在 2、3，说明病情得到了控制，功能有所恢复，但是很难得到完全的痊愈。故而脑卒中只要有过病史，在未来都应及时检查，积极配合治疗和习惯的保持，有感到异常及时就医，医生也应根据血肿体积、影像极差、一致性等影像数据及时给出包括脑室临流在内的有效治疗方案。

5.4 模型评价及改进

5.4.1 模型优点

模型充分考虑了变量之间的相关性，及其与所研究变量指标的关联性。针对所研究的是连续变量的情况采用多次回归的方法逐步拟合，并且及时剔除了极端值的影响。针对所研究情况是分类数据的情况，用多种机器学习方法进行探索，最终确定效果较少、出错率较低，且泛化能力较强的随机森林分类，并且使用遗传算法对其寻优改进。在分类变量和连续变量的聚类问题上也充分考虑了分类变量的影响，采用二分类 K-MEANS 法进行聚类，从而避免对分类数据聚类不当的问题。最后在决策上也采用对于数据拟合效果更好的岭回

[1]闻晓庆.血浆渗透压对脑出血后血肿周围水肿的影响及其预后的研究

归来进行预后探索，来获得更好的回归效果，从而给出有效临床决策建议。

5.4.2 改进意见

为防止相关数据剔除的误差，时间允许的情况下，可以再多用一些方法分析相关性，多采用几种方法剔除。并且在拟合时加入一些已经剔除的变量，从而避免剔除不当的问题。在对分类变量的拟合和预测上，为消除量纲及的差距及定量与定类拟合的差距，可以对一些定量因量纲展现的较大差距数据结合研究材料进行分级处理，缩小其因量纲展现的较大差距。并且可以跑更多次数据，来检验和确定自己所采用方法的效果是否最优。因为数据不平衡性剔除的末端回访特征影像也可以用于被检验的对象。

参考文献

[1] 黄晓宇. 基于 NCCT 影像组学预测原发性脑出血预后分层的研究[D].兰州大学,2023.DOI:10.27204/d.cnki.glzhu.2023.000177.

[2] 潘婷. 基于平扫 CT 影像组学模型预测自发性脑出血早期血肿扩张[D].吉林大学,2023.DOI:10.27162/d.cnki.gjlin.2022.006521.

[3] 张玉福,付茂盛,康杰等.“岛征”联合超早期血肿扩张速度在急性脑出血中预测 24 小时内血肿扩大的价值[J].医学影像学杂志,2021,31(02):187-189.

[4] 李青润. 基于 CT 影像组学模型对自发性脑出血血肿扩大的预测研究及与常规影像征象预测作用的价值比较[D].大连医科大学,2022.DOI:10.26994/d.cnki.gdlyu.2021.000198.

[5] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software].Retrieved from https://www.spsspro.com.

[6] 周志华.机器学习[M].清华大学出版社,2016.

[7] 闻晓庆.血浆渗透压对脑出血后血肿周围水肿的影响及其预后的研究