#出血性脑卒中临床智能诊疗建模

##一、 背景介绍

出血性脑卒中指非外伤性脑实质内血管破裂引起的脑出血，占全部脑卒中发病率的10-15%。其病因复杂，通常因脑动脉瘤破裂、脑动脉异常等因素，导致血液从破裂的血管涌入脑组织，从而造成脑部机械性损伤，并引发一系列复杂的生理病理反应。出血性脑卒中起病急、进展快，预后较差，急性期内病死率高达45-50%，约80%的患者会遗留较严重的神经功能障碍，为社会及患者家庭带来沉重的健康和经济负担。因此，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。

出血性脑卒中后，血肿范围扩大是预后不良的重要危险因素之一。在出血发生后的短时间内，血肿范围可能因脑组织受损、炎症反应等因素逐渐扩大，导致颅内压迅速增加，从而引发神经功能进一步恶化，甚至危及患者生命。因此，监测和控制血肿的扩张是临床关注的重点之一。此外，血肿周围的水肿作为脑出血后继发性损伤的标志，在近年来引起了临床广泛关注。血肿周围的水肿可能导致脑组织受压，进而影响神经元功能，使脑组织进一步受损，进而加重患者神经功能损伤。综上所述，针对出血性脑卒中后的两个重要关键事件，即血肿扩张和血肿周围水肿的发生及发展，进行早期识别和预测对于改善患者预后、提升其生活质量具有重要意义。

医学影像技术的飞速进步，为无创动态监测出血性脑卒中后脑组织损伤和演变提供了有力手段。近年来，迅速发展并广泛应用于医学领域的人工智能技术，为海量影像数据的深度挖掘和智能分析带来了全新机遇。期望能够基于本赛题提供的影像信息，联合患者个人信息、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，实现精准个性化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

##二、 数据集介绍及建模目标

赛题提供了160例（100例训练数据集+60例独立测试数据集）出血性脑卒中患者的个人史、疾病史、发病及治疗相关信息、多次重复的影像学检查（CT平扫）结果及患者预后评估,该部分信息可在“表1-患者列表及临床信息”中查询。如图1为脑出血患者CT平扫，红色为血肿区域，黄色为水肿区域。赛题提供影像学检查数据，包括各个时间点血肿/水肿的体积、位置、形状特征及灰度分布等信息。体积及位置信息可在“表2-患者影像信息血肿及水肿的体积及位置”中查询。形状及灰度分布信息可在“表3-患者影像信息血肿及水肿的形状及灰度分布”中查询。

赛题目标：通过对真实临床数据的分析，研究出血性脑卒中患者血肿扩张风险、血肿周围水肿发生及演进规律，最终结合临床和影像信息，预测出血性脑卒中患者的临床预后。

目标变量：

- 发病48小时内是否发生血肿扩张：1是；0否。
- 发病后90天 mRS：0-6，有序等级变量。其中mRS是评估卒中后患者功能状态的重要工具，详见附件2相关概念。

临床信息：相关信息在“表1-患者列表及临床信息”中获取。

- ID：患者ID。
- 训练数据集：sub001至sub100，共计100例。包含：患者信息、首次及所有随访影像数据及90天mRS。
- 测试数据集1：sub101至sub130，共计30例。包含：患者信息、首次影像数据。不包含：随访影像数据及90天mRS。
- 测试数据集2：sub131至sub160，共计30例。包含：患者信息、首次及所有随访影像数据。不包含：90天mRS。
- 入院首次影像检查流水号：一个14位数字编码。前8位代表年月日，后6位为顺序编号（注意：不是时分秒）。流水号是影像检查的唯一编码，具体影像检查时间点可通过对应流水号在“附表1-检索表格-流水号vs时间”中检索。
- 年龄： 岁
- 性别：男/女
- 脑出血前mRS评分：0-6，有序等级变量
- 高血压病史：1是0否

 卒中病史：1是0否
 糖尿病史：1是0否
 房颤史：1是0否
 冠心病史：1是0否
 吸烟史：1是0否
 饮酒史：1是0否
发病相关特征，共计2字段。
 血压：收缩压/舒张压。单位： 毫米汞柱
 发病到首次影像检查时间间隔：单位：小时
治疗相关特征，共计7字段。
 脑室引流：1是0否
 止血治疗：1是0否
 降颅压治疗：1是0否
 降压治疗：1是0否
 镇静、镇痛治疗：1是0否
 止吐护胃：1是0否
 营养神经：1是0否
影像相关特征，共计84字段/时间点。
 血肿及水肿的体积和位置信息在“表2-患者影像信息血肿及水肿的体积及位置”中获取，包含了：每个时间点血肿（Hemo）总体积及水肿（ED）总体积及不同位置的占比。体积占比定义：血肿/水肿在该位置的体积占总体积大小的比例，取值范围为：0-1。如：0代表该区域没有发生血肿/水肿，1则代表该患者所有血肿/水肿均发生在该区域，可通过占比换算出该位置绝对体积。本赛题采用通用模板，区分左右侧大脑前动脉（ACA_L，ACA_R），左右侧大脑中动脉（MCA_L，MCA_R），左右侧大脑后动脉（PCA_L，PCA_R），左右侧脑桥/延髓（Pons_Medulla_L，Pons_Medulla_R），左右侧小脑（Cerebellum_L，Cerebellum_R）共十个不同位置，具体位置和参考文献见附件2-相关概念。综上，总体积：2个字段（单位：10-3ml），位置：20个字段。在每个时间点，体积及位置特征共计22个字段。
 血肿及水肿的形状及灰度分布在“表3-患者影像信息血肿及水肿的形状及灰度分布”的两个不同标签页存放，可通过流水号检索对应数据。每个时间

点血肿及水肿的形状及灰度特征，反映目标区域内体素信号强度的分布（17个字段）及三维形状的描述（14个字段），因此，在每个时间点，血肿及水肿的形状+灰度分布特征共62字段。
注：重复影像数据根据临床真实情况提供，重复时间个体间可能存在差异。

##三、请建模回答如下问题
###1　血肿扩张风险相关因素探索建模。
a) 请根据“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。
结果填写规范：1是0否，填写位置：“表4”C字段（是否发生血肿扩张）。
如发生血肿扩张事件，请同时记录血肿扩张发生时间。
结果填写规范：如10.33小时，填写位置：“表4”D字段（血肿扩张时间）。
 是否发生血肿扩张可根据血肿体积前后变化，具体定义为：后续检查比首次检查绝对体积增加≥6 mL或相对体积增加≥33%。
注：可通过流水号至“附表1-检索表格-流水号vs时间”中查询相应影像检查时间点，结合发病到首次影像时间间隔和后续影像检查时间间隔，判断当前影像检查是否在发病48小时内。
b) 请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（字段E至W）、“表2”中其影像检查结果（字段C至X）及“表3”其影像检查结果（字段C至AG，注：只可包含对应患者首次影像检查记录）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。
注：该问只可纳入患者首次影像检查信息。
结果填写规范：记录预测事件发生概率（取值范围0-1，小数点后保留4位数）；填写位置：“表4”E字段（血肿扩张预测概率）。

###2　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。
a) 请根据“表2”前100个患者（sub001至sub100）的水肿体积（ED_volume）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。
结果填写规范：记录残差，填写位置“表4”F字段（残差（全体））。
b) 请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。
结果填写规范：记录残差，填写位置“表4”G字段（残差（亚组）），同时将所属亚组填写在H段（所属亚组）。
c) 请分析不同治疗方法（“表1”字段Q至W）对水肿体积进展模式的影响。
d) 请分析血肿体积、水肿体积及治疗方法（“表1”字段Q至W）三者之间的关系。

###3　出血性脑卒中患者预后预测及关键因素探索。
a) 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。
注：该问只可纳入患者首次影像检查信息。
结果填写规范：记录预测mRS结果，0-6，有序等级变量。填写位置“表4”I字段（预测mRS（基于首次影像））。
b) 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。
结果填写规范：记录预测mRS结果，0-6，有序等级变量。填写位置“表4”J字段（预测mRS）。
c) 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。

##1 血肿扩张风险相关因素探索建模

1.问题分析

出血性脑卒中是一个严重的医学问题，而现代的医学影像技术和人工智能为其提供了新的治疗和预测机会。赛题提供的数据集为我们提供了丰富的患者临床信息和影像学特征，目标是预测患者的临床预后。

2.建模思路

1) 数据预处理

- 对影像数据进行标准化处理，使其具有统一的尺度。
- 对分类特征进行独热编码。
- 对连续特征进行归一化处理。

2) 特征工程

- 从临床信息中提取关键特征，如疾病史、治疗方式等。
- 从影像数据中提取血肿和水肿的体积、位置、形状和灰度特征。

3) 模型构建

- 使用机器学习或深度学习模型进行训练。
- 使用交叉验证对模型进行评估，选择最佳模型。

4) 预测

- 使用模型预测发病48小时内是否发生血肿扩张。
- 使用模型预测发病后90天的mRS评分。

3.解决方案

1) 数据预处理

- 使用sklearn的preprocessing模块进行数据标准化和归一化。
- 使用pandas的get_dummies函数进行独热编码。

2) 特征工程

- 根据医学知识，选择与出血性脑卒中相关的关键特征。
- 使用pyradiomics库从CT影像中提取血肿和水肿的形态学和灰度特征。

3) 模型构建

- 尝试使用随机森林、支持向量机、神经网络等模型。
- 使用sklearn的GridSearchCV进行模型选择和参数调优。

4) 预测

- 对测试数据集进行预测。
- 使用sklearn的classification_report和confusion_matrix对预测结果进行评估。

4.潜在难点

- 数据不平衡：如果某一类的样本数量远少于其他类，可能需要进行过采样或欠采样。
- 特征选择：需要根据医学知识和数据特性选择最有意义的特征。
- 模型调优：可能需要多次尝试不同的模型和参数才能得到最佳效果。

5.结论

通过对出血性脑卒中的患者临床信息和影像数据进行深入分析和建模，我们可以更准确地预测患者的临床预后，从而为患者提供更好的治疗方案。这种方法结合了现代医学影像技术和人工智能技术，为出血性脑卒中的诊断和治疗提供了新的机会。

##1.a判断患者sub001至sub100发病后48小时内是否发生血肿扩张事件

##题目分析

1. 定义血肿扩张

首先，我们需要明确什么情况下被认为是血肿扩张。通常，如果血肿体积在一段时间内显著增加，我们可以认为发生了血肿扩张。具体地，我们可以定义：

$$
\text{血肿扩张} =
\begin{cases}
1 & \text{如果} \frac{HM{\text{volume后}} - HM{\text{volume前}}}{HM_{\text{volume前}}} > \theta \\
0 & \text{否则}
\end{cases}
$$

其中，θ 是一个阈值，例如0.1或10%，表示血肿体积增加的百分比。

2. 数据整合

- 从“表1”中，我们可以获取每个患者的首次影像检查的流水号和发病到首次影像检查的时间间隔。
- 从“表2”中，我们可以获取各个时间点的流水号和对应的血肿体积。

我们需要根据流水号将这两张表连接起来，以便我们可以针对每个患者和每个时间点比较血肿体积的变化。

3. 血肿扩张的判定

对于每个患者：

- 找到其首次影像检查的血肿体积。
- 查找在首次影像检查后48小时内的所有影像检查记录，并确定这段时

间内的血肿体积最大值。
- 使用上面的公式判断是否发生了血肿扩张。
- 如果发生了血肿扩张，记录扩张发生的时间为：扩张发生时的流水号时间减去首次影像检查的流水号时间。

##解题思路

1. 数据整合

我们的首要任务是对数据进行整合，确保我们可以轻松地跟踪每个患者在不同时间点的血肿体积。

公式1：数据整合
[
D{\text{merged}} = \text{合并}(D{\text{表1}}, D_{\text{表2}}, \text{基于流水号})
]

其中，(D{\text{merged}}) 是整合后的数据，(D{\text{表1}}) 和 (D_{\text{表2}}) 分别是表1和表2的数据。

2. 血肿体积变化的计算

我们需要计算每个患者在48小时内的血肿体积变化。这可以通过比较首次影像检查的血肿体积与48小时内的最大血肿体积来实现。

公式2：血肿体积变化
[
\Delta HM{\text{volume}} = HM{\text{volume, 48h max}} - HM_{\text{volume, 初始}}
]

3. 判断是否发生血肿扩张

对于血肿扩张的判定，我们可以设定一个阈值。如果血肿体积的变化超过这个阈值，我们就认为发生了血肿扩张。

公式3：血肿扩张判定

$$
\text{血肿扩张} =
\begin{cases}
1 & \text{如果} \frac{\Delta HM{\textit{volume}}}{HM{\text{volume, 初始}}} > \theta \\
0 & \text{否则}
\end{cases}
$$

其中，θ 是一个阈值，例如0.1或10%，表示血肿体积增加的百分比。

4. 计算血肿扩张发生的时间

如果在48小时内发生了血肿扩张，我们需要确定扩张发生的具体时间。

公式4：血肿扩张时间

$$
T{\textit{扩张}} = T{\text{发现最大体积}} - T_{\text{首次影像}}
$$

5. 结果记录

最后，我们需要将结果记录到“表4”中。

公式5：结果记录

$$
D_{\text{表4, C列}} =
\begin{cases}
1 & \text{如果发生血肿扩张} \\
0 & \text{否则}
\end{cases}
$$

$$
D_{\text{表4, D列}} =
\begin{cases}
T_{\text{扩张}} & \text{如果发生血肿扩张} \\
\text{N/A} & \text{否则}
\end{cases}
$$

结合这些公式和分析思路，我们可以编写相应的程序来处理数据，计算是否发生血肿扩张，以及扩张发生的时间，并将结果记录到“表4”中。

##编程思路

当然可以，我将按部就班地给出整体的编程思路，然后提供相应的程序示例。

编程思路：

1. **数据整合**:
    - 使用 pandas 库读取并加载所有的数据表。
    - 使用适当的键（例如“入院首次检查流水号”和“首次检查流水号”）将“表1”和“表2”进行合并。
    - 使用“附表1-检索表格-流水号vs时间”创建一个从流水号到具体时间的映射。

2. **计算时间间隔**:
    - 根据首次检查的时间，计算每次检查距离首次检查的小时数。
3. **血肿扩张的判断**:
    - 遍历每个患者，记录首次检查时的血肿体积。
    - 查找在首次检查后48小时内的最大血肿体积。
    - 使用公式(\frac{{\text{{max_hm_volume_48h}} - \text{{initial_hm_volume}}}}{{\text{{initial_hm_volume}}}} > \theta)来判断是否发生了血肿扩张，其中(\theta)是一个给定的阈值（如0.1，表示10%的增长）。
4. **记录结果**:
    - 在一个新的DataFrame中记录每个患者是否发生了血肿扩张以及发生的时间。

程序示例：

```
import pandas as pd

# 1. 数据整合
df1 = pd.read_excel('表1-患者列表及临床信息.xlsx')
df2 = pd.read_excel('表2-患者影像信息血肿及水肿的体积及位置.xlsx')
time_lookup = pd.read_excel('附表1-检索表格-流水号vs时间.xlsx')

# 根据流水号合并表1和表2
merged = pd.merge(df1, df2, left_on="入院首次影像检查流水号", right_on="首次检查流水号")

# 从时间查找表中创建流水号到时间的映射
time_dict = dict(zip(time_lookup['入院首次检查流水号'], time_lookup['入院首次检查时间点']))
```

```
13
14 # 2. 计算时间间隔
15 merged['时间'] = merged['首次检查流水号'].map(time_dict)
16 merged['hours_since_first'] = (merged['时间'] -
   merged.groupby('ID')['时
   间'].transform('first')).dt.total_seconds() / 3600
17
18 # 3. 血肿扩张的判断
19 theta = 0.1
20 result = pd.DataFrame()
21 result['ID'] = merged['ID'].unique()
22 result['是否发生血肿扩张'] = 0
23 result['血肿扩张时间'] = None
24
25 for patient in result['ID']:
26     initial_hm_volume = merged.loc[(merged['ID'] ==
   patient) & (merged['hours_since_first'] == 0),
   'HM_volume'].values[0]
27     max_hm_volume_48h = merged.loc[(merged['ID'] ==
   patient) & (merged['hours_since_first'] <= 48),
   'HM_volume'].max()
28
29     if (max_hm_volume_48h - initial_hm_volume) /
   initial_hm_volume > theta:
30         result.loc[result['ID'] == patient, '是否发生血肿
   扩张'] = 1
31         expansion_time = merged.loc[(merged['ID'] ==
   patient) & (merged['HM_volume'] == max_hm_volume_48h),
   'hours_since_first'].values[0]
32         result.loc[result['ID'] == patient, '血肿扩张时
   间'] = expansion_time
33
```

```
# 4. 记录结果
result.to_excel('表4-结果.xlsx', index=False)
```

这是一个高级程序示例，它提供了从数据加载到结果输出的完整流程。在实际应用中，可能需要进行更多的数据清洗和验证，以确保数据的准确性和完整性。

##1.b出血性脑卒中发病概率预测模型的建立
##问题分析

1. 问题定义

目标是构建一个模型，以预测出血性脑卒中患者在发病48小时内是否会发生血肿扩张。

2. 数据预处理

2.1 数据清洗

- 处理缺失值：使用均值、中位数或众数填充，或使用模型预测缺失值。
- 异常值处理：基于IQR或Z-score方法识别并处理。

2.2 数据转换

- 对于分类变量，例如性别、疾病史等，使用独热编码进行转换。
- 对于连续变量，例如年龄和血压，使用归一化或标准化。

[X' = \frac{X - \text{min}(X)}{\text{max}(X) - \text{min}(X)}]

其中，(X') 是归一化后的值，(X) 是原始值。

3. 特征工程

3.1 特征选择

基于医学知识，结合数据探索性分析的结果，选择与目标变量高度相关的特征。

3.2 特征构造

可以考虑从现有特征中构造新特征，例如：

- 结合年龄和其他健康指标。
- 从影像数据中提取的形状、位置和灰度特征的组合。

4. 模型选择与训练

4.1 模型选择

考虑使用以下模型：

- 逻辑回归
  $$ p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots)}} $$
- 支持向量机
- 随机森林
- 深度学习模型，如神经网络

4.2 模型训练

使用交叉验证选择最佳超参数。

5. 模型评估

5.1 性能指标

- 准确率

- 召回率
- F1分数
- ROC曲线与AUC

5.2 验证

使用独立的测试集评估模型性能。

6. 结果解释

使用特征重要性图或其他方法解释模型的预测结果，为临床决策提供洞察。

7. 结论

综合模型的预测性能、特征重要性和医学知识，提出关于预测出血性脑卒中患者是否会发生血肿扩张的结论和建议。

##解题思路
###1. 数据预处理

- 转换性别列为数值型（例如，男=1，女=0）。
- 选择一些关键特征，例如年龄、性别、高血压病史等。
- 分割数据集为训练集和验证集。

```python
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

# Convert gender to a numeric value
merged_df['性别'] = merged_df['性别'].map({'男': 1, '女': 0})

# Select features and target
features = ['年龄', '性别', '高血压病史', '卒中病史', '糖尿病史']
X = merged_df[features]
y = merged_df['是否发生血肿扩张']

# Split data into train and validation sets
X_train, X_valid, y_train, y_valid = train_test_split(X, y, train_size=0.8, test_size=0.2, random_state=0)
```

2. 模型构建与训练

- 使用逻辑回归进行模型训练。

```python
# Initialize and train a logistic regression model
model = LogisticRegression(max_iter=500)
model.fit(X_train, y_train)
```

3. 模型评估

- 使用验证集进行模型预测。
- 计算模型的准确度。

```
from sklearn.metrics import accuracy_score

# Predict on validation set
y_pred = model.predict(X_valid)

# Calculate accuracy
accuracy = accuracy_score(y_valid, y_pred)
```

通过这种方式，我们可以在给定的数据集上建立一个简单的逻辑回归模型。这只是一个起点，实际上，根据数据的复杂性和问题的特性，可能需要进行更多的数据工程、特征选择和模型调优。

2. 血肿周围水肿的发生及进展建模

##解题分析

a. 全体患者水肿体积随时间进展曲线

思路:

1. 使用患者的水肿体积和相对于首次检查的检查时间点来建立关系。
2. 使用回归分析来拟合数据并得到一个函数 (y = f(x))，其中 (y) 是水肿体积，(x) 是时间。
3. 使用实际的水肿体积值和拟合曲线计算残差。

公式:

- 拟合函数： (y = f(x))
- 残差计算： (\text{残差} = \text{实际值} - f(x))

b. 水肿体积随时间进展模式的个体差异

思路:

1. 使用聚类算法（如K-means）对患者进行分组，以找到不同的水肿体积随时间进展模式。
2. 对每个分组使用回归分析来拟合数据并得到一个函数。
3. 使用实际的水肿体积值和拟合曲线计算残差。

公式:

- 聚类算法： K-means
- 拟合函数： $y_i = f_i(x)$，其中 i 是分组标识
- 残差计算： $\text{残差}_i = \text{实际值} - f_i(x)$

2. 分析治疗方法对水肿体积进展模式的影响

思路:

1. 使用ANOVA或多重回归分析来研究不同治疗方法对水肿体积的影响。
2. 分析并解释统计结果。

3. 分析血肿体积、水肿体积及治疗方法之间的关系

思路:

1. 使用相关性分析来研究血肿体积和水肿体积之间的关系。
2. 使用多重回归分析来研究治疗方法、血肿体积和水肿体积之间的关系。

公式:

- 相关性分析： Pearson's r
- 多重回归分析： $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots$，其中 y 是水肿体积，$x_1, x_2, \ldots$ 是解释变量（如治疗方法、血肿体积等）

可视化图标建议：

1. **水肿体积随时间的变化**：可以使用散点图展示每个患者的水肿体积随时间的变化，同时添加拟合曲线展示总体趋势。
2. **不同治疗方法对水肿体积的影响**：可以使用箱线图或小提琴图展示不同治疗方法下患者的水肿体积分布。
3. **血肿体积与水肿体积的关系**：可以使用散点图，并添加回归线来展示两者之间的关系。
4. **多重回归结果**：可以使用柱状图展示每个解释变量的回归系数，以及其统计显著性。

这些图形将帮助我们更直观地理解血肿周围水肿的进展、治疗干预与水肿进展的关联关系，以及血肿体积、水肿体积及治疗方法之间的关系。

##建模思路

确定性地对血肿周围水肿的发生及进展进行建模是一个高度复杂的问题，涉及多种生物学、生理学和治疗干预因素

1. 血肿周围水肿的发生及进展建模

a. 全体患者水肿体积随时间进展曲线

思路:

1. **数据收集和整理**：首先，从“表2”中提取前100个患者的水肿体积和检查时间数据。
2. **模型选择**：由于数据点可能会有非线性趋势，可以考虑使用多项式回归、非线性回归或时间序列模型进行拟合。
3. **模型评估**：使用交叉验证和残差平方和 (RSS) 来评估模型的拟合质量。

公式:

- 多项式回归：[y = \beta_0 + \beta_1x + \beta_2x^2 + \ldots + \beta_px^p]
- 残差计算：[\text{残差} = y{*text{observed}} - y{\text{predicted}}]

b. 水肿体积随时间进展模式的个体差异

思路:

1. **数据分组**：利用聚类分析或主成分分析 (PCA) 对患者进行分组，以找到不同的水肿体积随时间进展模式。
2. **模型选择**：对每个分组进行多项式回归或非线性回归拟合。
3. **模型评估**：与上述类似，使用交叉验证和RSS来评估每个分组的模型拟合质量。

公式:

- 聚类算法：例如 K-means 或层次聚类
- 主成分分析 (PCA)

2. 分析治疗方法对水肿体积进展模式的影响

思路:

1. **数据整合**：从“表1”中提取治疗方法数据，并与“表2”中的水肿体积数据进行整合。
2. **模型选择**：考虑使用广义线性模型 (GLM) 或混合效应模型对数据进行拟合，以考虑到患者之间的差异。
3. **模型评估**：使用ANOVA或其他统计测试来评估模型中不同治疗方法的效果。

3. 分析血肿体积、水肿体积及治疗方法之间的关系

思路:

1. **数据整合**：将“表1”和“表2”中的数据整合到一个数据框中。
2. **模型选择**：使用多重回归分析来探索血肿体积、水肿体积和各种治疗方法之间的关系。
3. **模型评估**：使用F统计量、t统计量和R-squared值来评估模型的质量和单个系数的显著性。

公式:

- 多重回归分析：[y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_px_p]

这些建模和分析步骤提供了对血肿周围水肿发生、进展和治疗干预的深入理解。当然，在进行这些步骤时，可能还需要进一步的数据清洗、缺失值处理、异常值检测等预处理工作。

#3. 出血性脑卒中患者预后预测及关键因素探索

3.a. 预测90天mRS评分（基于首次影像）

解题思路:

1. **数据整合与预处理**:
   - 从“表1”中提取前100名患者的个人史、疾病史和发病相关数据（字段E至W）。
   - 从“表2”和“表3”中提取首次影像数据。
   - 对连续变量进行标准化或归一化处理。
   - 对分类变量进行独热编码。
2. **特征工程**:
   - 根据先验知识或特征重要性选择与mRS评分可能相关的特征。
   - 从影像数据中提取相关的信息，如血肿的大小、位置和形状。
3. **模型选择与建模**:

- 由于mRS评分是有序分类变量，考虑使用有序逻辑回归模型。
- 同时，可以尝试其他的机器学习模型如随机森林、支持向量机等，并对比其效果。

4. **模型评估与预测**:
   - 利用已知数据对模型进行训练，并进行交叉验证评估。
   - 利用模型预测sub001至sub160的90天mRS评分。
   - 使用混淆矩阵、精确度、召回率等指标进行模型评估。

公式与模型:

使用有序逻辑回归模型，其公式为：

$$
\log \left(\frac{P(Y \le k)}{1 - P(Y \le k)} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n
$$

其中，$P(Y \le k)$ 表示响应变量Y小于或等于k的概率。

3.b. 预测90天mRS评分（基于首次+随访影像）

解题思路:

1. **数据整合与特征工程**:
   - 与3.a类似，但此时需要整合随访影像数据，因此可能需要考虑时间序列的变化特点。
2. **模型选择、建模与预测**:
   - 考虑时间序列模型或递归神经网络（RNN）等模型来捕捉时间上的依赖关系。
   - 使用与3.a相同的方法进行模型评估与预测。

3.c. 预后与各因素的关联关系分析

解题思路:

1. **数据整合与特征工程**:
    - 同3.a.
2. **相关性与影响力分析**:
    - 使用多元线性回归或其他回归模型来探索mRS评分与其他连续变量之间的关系。
    - 对于分类变量，可以使用ANOVA或卡方检验来分析其与mRS评分的关系。
    - 对于影像特征，可以考虑使用深度学习模型如CNN来自动提取特征，并与mRS评分关联。
3. **建议与结论**:
    - 根据各因素与mRS评分的关联关系，为临床决策提供建议。
    - 分析各因素的影响力，并提出可能的干预措施。

公式与模型:

- 多元线性回归：

$$
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon
$$

- ANOVA:

  利用F统计量比较各组的均值差异。

- 深度学习模型（例如CNN）：

  使用卷积层、池化层和全连接层来自动提取影像特征。

建模思路

3.a. 预测90天mRS评分（基于首次影像）

解题思路:

1. **数据准备**:
   - 从“表1”中提取前100名患者的个人史、疾病史和发病相关数据（字段E至W）。
   - 从“表2”和“表3”中提取首次影像数据。
   - 对这些数据进行初步的统计描述，以理解数据的分布、缺失情况和可能的异常值。
2. **特征工程**:
   - 选择对mRS评分可能具有预测力的特征。
   - 考虑创建交互特征，如血肿大小与位置的组合，或血肿大小与某些临床特征的组合。
   - 对分类变量进行独热编码。
   - 对连续变量进行标准化或归一化处理。
3. **模型选择与建模**:
   - 由于mRS评分是有序分类变量，考虑使用有序逻辑回归模型。
   - 同时，可以尝试其他的机器学习模型如随机森林、支持向量机等，并对比其效果。
   - 使用交叉验证进行模型选择和参数优化。
4. **模型评估与预测**:
   - 使用已知数据对模型进行训练，并进行交叉验证评估。
   - 利用模型预测sub001至sub160的90天mRS评分。
   - 使用混淆矩阵、精确度、召回率等指标进行模型评估。

3.b. 预测90天mRS评分（基于首次+随访影像）

解题思路:

1. **数据整合与特征工程**:

- 与3.a类似，但此时需要整合随访影像数据，因此可能需要考虑时间序列的变化特点。
- 考虑时间窗口特征，例如首次到第二次、第二次到第三次的变化。

2. **模型选择、建模与预测**:
   - 考虑使用时间序列模型，如ARIMA或长短时记忆网络（LSTM）等模型，捕捉时间上的依赖关系。
   - 使用与3.a相同的方法进行模型评估与预测。

3.c. 预后与各因素的关联关系分析

解题思路:

1. **数据整合与特征工程**:
   - 同3.a.
2. **相关性与影响力分析**:
   - 使用多元线性回归或其他回归模型来探索mRS评分与其他连续变量之间的关系。
   - 对于分类变量，可以使用ANOVA或卡方检验来分析其与mRS评分的关系。
   - 对于影像特征，可以考虑使用深度学习模型如CNN来自动提取特征，并与mRS评分关联。
3. **建议与结论**:
   - 根据各因素与mRS评分的关联关系，为临床决策提供建议。
   - 分析各因素的影响力，并提出可能的干预措施。

综上，对于出血性脑卒中的预后预测及关键因素探索，首先需要对数据进行深入的理解和预处理，然后选择合适的模型进行建模和预测，最后分析各因素与预后之间的关联关系，并提出相应的建议。

3. 出血性脑卒中患者预后预测及关键因素探索

3.a. 预测90天mRS评分（基于首次影像）

编程思路:

1. **数据准备**:
   - 加载“表1”、“表2”和“表3”。
   - 对数据进行清洗：处理缺失值、异常值、对数据进行标准化或归一化。
2. **特征工程**:
   - 使用 `pandas` 进行数据合并和转换。
   - 使用 `sklearn.preprocessing` 进行特征的独热编码和标准化。
3. **模型选择与建模**:
   - 划分数据为训练集和验证集。
   - 使用 `sklearn` 中的有序逻辑回归或其他适合预测有序分类变量的模型，如随机森林、支持向量机、神经网络。
4. **模型评估与预测**:
   - 使用交叉验证评估模型效果。
   - 使用评估指标如混淆矩阵、AUC、F1分数等评价模型性能。
   - 对sub001至sub160的数据进行预测。
5. **超参数调优**:
   - 使用网格搜索或随机搜索优化模型的超参数。

代码示例:

```
# 数据准备
import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
```

```
6
7  table1 = pd.read_excel("表1.xlsx")
8  table2 = pd.read_excel("表2.xlsx")
9  table3 = pd.read_excel("表3.xlsx")
10
11 # 合并数据
12 merged_data = table1.merge(table2,
   on="ID").merge(table3, on="ID")
13
14 # 特征工程
15 features = merged_data[["列名1", "列名2", ...]]  # 根据前
   面的分析选择相关列
16 X = pd.get_dummies(features)  # 独热编码
17 y = merged_data["mRS评分"]
18
19 scaler = StandardScaler()
20 X_scaled = scaler.fit_transform(X)
21
22 # 划分数据集
23 X_train, X_val, y_train, y_val =
   train_test_split(X_scaled, y, test_size=0.2,
   random_state=42)
24
25 # 模型选择与建模
26 clf = LogisticRegression(multi_class='multinomial',
   solver='lbfgs')
27 clf.fit(X_train, y_train)
28
29 # 预测验证集
30 predictions = clf.predict(X_val)
```

##编程思路

3.b. 预测90天mRS评分（基于首次+随访影像）

编程思路:

1. **数据整合**:
   - 加载数据，并合并首次和随访影像数据。
2. **特征工程**:
   - 使用时间窗口特征，如首次到第二次的变化。
3. **模型选择与建模**:
   - 考虑使用深度学习模型，如LSTM，处理时间序列数据。
4. **模型评估与预测**:
   - 使用交叉验证进行模型评估。
   - 进行预测。
5. **超参数调优**:
   - 使用网格搜索或随机搜索优化模型的超参数。

代码示例:

```
# 数据整合
merged_data = table1.merge(table2, on="ID").merge(table3, on="ID")

# 特征工程
features = merged_data[["列名1", "列名2", ...]]
X = pd.get_dummies(features)
y = merged_data["mRS评分"]

# LSTM模型建立与训练
from keras.models import Sequential
from keras.layers import LSTM, Dense

```

```
model = Sequential()
model.add(LSTM(50, input_shape=(X.shape[1], 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')
model.fit(X, y, epochs=50, batch_size=32)

# 预测
predictions = model.predict(X)
```

3.c. 关联关系分析

编程思路:

1. **数据准备**:
   - 加载数据，合并所需的特征。
2. **相关性分析**:
   - 使用 `statsmodels` 进行多元线性回归分析。
   - 使用 `scipy.stats` 进行ANOVA或卡方检验。
3. **结果解释与建议**:
   - 基于回归分析的结果，解释各因素与mRS评分的关系。
   - 提出可能的干预措施。
4. **可视化**:
   - 使用 `matplotlib` 或 `seaborn` 绘制相关图形，如散点图、箱线图、热图等。

代码示例:

```
import statsmodels.api as sm
import scipy.stats as stats
import seaborn as sns
```

```
import matplotlib.pyplot as plt

# 数据准备
merged_data = table1.merge(table2,
on="ID").merge(table3, on="ID")

# 相关性分析
X = merged_data[["列名1", "列名2", ...]]
X = sm.add_constant(X)  # 添加常数项
y = merged_data["mRS评分"]

model = sm.OLS(y, X).fit()

# 结果解释
print(model.summary())

# 可视化
sns.heatmap(merged_data.corr())
plt.show()
```