北京大学
PEKING UNIVERSITY

健康数据科学的 Python 语言编程基础
课程报告

基于机器学习的儿童重症监护室患者住院死亡风险预测研究

学院 公共卫生学院

专业 XXXXXXXXXXXXXXX

学号 XXXXX

姓名 苏江慧

导师 高培 研究员

2026 年 1 月 18 日

摘　要

背景与目的：儿童重症监护室（Pediatric Intensive Care Unit，PICU）患者的病情变化快、死亡风险高，早期识别高危患者对临床资源分配及干预决策至关重要。本报告旨在利用 Python 数据科学工具链，基于公开的儿童重症监护数据集（Paediatric Intensive Care，PIC），挖掘常规人口学及实验室指标与患者住院死亡结局的关联，构建并评估死亡风险预测模型。

方法：本研究纳入 PIC 数据库中 13,258 名患者样本，通过 Pandas 进行数据清洗、缺失值中位数填充及 IQR 异常值处理。采用独立样本 t 检验与相关性热力图进行描述性统计分析。在建模阶段，构建了逻辑回归（Logistic Regression，LR）、随机森林（Random Forest，RF）及支持向量机（Support Vector Machine，SVM）三类预测模型，结合 SMOTE 策略与类别权重调整解决类别不平衡问题，并通过 GridSearchCV 进行超参数寻优。此外，引入 K-Means 聚类算法探索患者的风险亚群分布。

结果：数据分析显示，死亡组在月龄及多项实验室指标上与存活组存在显著差异（$p < 0.05$）。模型评估表明，随机森林模型在测试集上表现最优（AUC=0.735），优于 SVM（AUC=0.722）和逻辑回归（AUC=0.712）。SHAP 可解释性分析提示 `lab_5257_min` 和 `lab_5237_min` 是预测死亡风险的最关键特征。无监督聚类成功识别出死亡率高达 20.5% 的高风险亚群。

结论：基于机器学习的预测模型能有效利用常规临床指标评估 PICU 患者预后，随机森林模型结合 SHAP 分析为临床解释提供了透明度。本研究构建的数据分析流程与可视化网页展示了从数据挖掘到临床决策支持的完整转化路径，具有一定的应用参考价值。

关键词：PICU；死亡风险预测；机器学习；随机森林；SHAP 分析

目录

1 个人主页网址

作者个人主页公开访问地址为：https://jianghuisu.github.io/

2 数据分析报告

2.1 项目概述

儿童重症监护室（Pediatric Intensive Care Unit，PICU）不仅收治危重症患儿，其产生的大量连续监测数据也为临床研究提供了宝贵资源。本研究基于临床医学公开数据库儿童重症监护数据集（Paediatric Intensive Care, PIC），聚焦于“患者住院死亡结局预测”这一核心临床问题 [1]。

通过 Python 数据科学技术栈（Pandas, Scikit-learn, Matplotlib/Seaborn, SHAP），本研究实现了从数据读取、预处理、描述性统计分析、变量筛选、模型构建与评估的全流程分析。研究旨在：

1. 描述 PICU 患者群体的基本特征分布；
2. 筛选影响患者生存结局的关键实验室指标；
3. 建立并对比评估不同机器学习模型的预测性能；
4. 开发在线展示页面，提升研究成果的可视化与传播效率。

2.2 数据读取与基础概览

2.2.1 数据读取

本研究采用 Python 的 `pandas` 库读取 Excel 格式的原始临床数据。考虑到数据处理的复用性，封装了数据读取函数，核心代码如下：

Code Listing 1: 数据读取

```python
import pandas as pd
def load_and_overview(path):
    path = "data.xlsx"
    data = pd.read_excel(path)
    return data
```

2.2.2 数据探索

本次分析所用的 PIC 数据共包含 **13,258** 个患者样本，包含 7 个核心变量。特征变量涵盖人口学特征与临床实验室指标，结局变量为二分类变量。

1. 数据维度与变量含义

通过 `data.info()` 与 `data.columns` 命令探索数据，结果显示数据维度为 $(13258, 7)$。各变量的具体含义及类型界定如下：

- **人口学变量**（数值型）：`age_month`：患者进入 ICU 时的月龄，反映患者的生长发育阶段，是重要的基线特征。
- **临床实验室指标**（数值型）：`lab_5237_min`、`lab_5227_min`、`lab_5225_range`、`lab_5235_max`、`lab_5257_min`，分别对应某实验室指标的最小值、范围、最大值，反映患者生理状态。
- **结局变量**（分类变量）：`HOSPITAL_EXPIRE_FLAG`：0 代表存活，1 代表死亡。记录患者住院期间是否死亡，这是本研究的预测目标。

2. 缺失值具体分布

临床真实世界数据常伴随缺失现象。通过 `isnull().sum()` 统计发现，本数据集呈现“选择性缺失”特征：

- **完整变量**：`age_month` 与结局变量无缺失，说明基础登记信息完整。
- **缺失变量**：5 个实验室指标均存在缺失，`lab_5237_min`（4590 条）、`lab_5227_min`（4642 条）、`lab_5225_range`（4592 条）、`lab_5235_max`（4587 条）、`lab_5257_min`（4596 条），平均缺失率约为 35%。

这种缺失并非完全随机缺失（Missing Completely At Random，MCAR），更符合临床“按需采集”的逻辑，即病情较轻或非特定病种的患儿未进行相关检查。直接删除缺失样本会导致近三分之一的数据丢失，且可能引入选择偏倚，因此后续采用“中位数填充”补全缺失值。

3. 结局变量分布

对结局变量进行统计：

Code Listing 2: 统计结局变量分布

```
outcome_dist = data["HOSPITAL_EXPIRE_FLAG"].value_counts()
death_ratio = data["HOSPITAL_EXPIRE_FLAG"].mean()
```

结局变量呈典型的“类别不平衡”分布：存活样本数量为 12478 例，占比 94.1%，死亡样本数量为 780 例，占比 5.9%。如果直接建模，模型极易倾向于预测“存活”以获得高准确率，但会漏掉我们最关心的“死亡”案例。因此，后续在模型训练后续通过以下两种方式校正：（1）数据集划分时用 `stratify=y` 保持训练/测试集死亡占比一致；（2）模型训练时设置 `class_weight='balanced'`，提升少数类（死亡样本）的权重 [2]。

2.2.3 描述性统计分析

为深入理解数据分布特征，本研究通过 `exploratory_analysis` 函数对各变量进行了统计描述，包括基本信息统计、分布形态可视化及结局分组对比，为后续预处理与建模提供依据。

1. 集中趋势与离散程度

选取 6 个核心数值特征（1 个人口学变量 +5 个临床实验室指标），排除缺失值后计算其均值、中位数及方差，结果如表1所示。`age_month` 的中位数（8.0 个月）远小于均值（29.9 个月），提示数据呈严重的右偏分布，说明 PICU 收治的患儿以婴幼儿为主。实验室指标中，`lab_5257_min` 的方差高达 3052.25，提示个体间差异巨大，可能是区分病情严重程度的关键指标。

表 1: 数值特征描述性统计（$n = 13258$）

变量名	均值	中位数	方差
age_month	29.989	8.0	1939.011
lab_5237_min	7.347	7.368	0.013
lab_5227_min	2.248	1.8	3.327
lab_5225_range	3.842	1.5	29.846
lab_5235_max	42.610	39.6	218.750
lab_5257_min	91.546	105.0	3052.255

2. 特征分布可视化

绘制 6 个变量的直方图以观察变量分布形态，如图1所示。`lab_5237_min` 呈尖峰分布，多数患者取值接近中位数，异常值极少。`lab_5257_min` 呈现出双峰分布特征（低值区 0-50 与高值区 100-150），暗示了潜在的两个不同患者群体，这为后续的聚类分析提供了依据。

3. 分组箱线图对比

以结局变量为分组依据，绘制箱线图探索组间差异，结果如图2所示，揭示了特征与预后的初步关联：age_month：死亡组的箱体位置明显更低，提示低月龄可能是死亡的危险因素。lab_5225_range 死亡组的中位数及离散度均显著高于存活组，说明生理指标的高波动性可能与不良预后相关。lab_5257_min 死亡组显著低于存活组，提示该指标的低水平可能是高危信号。

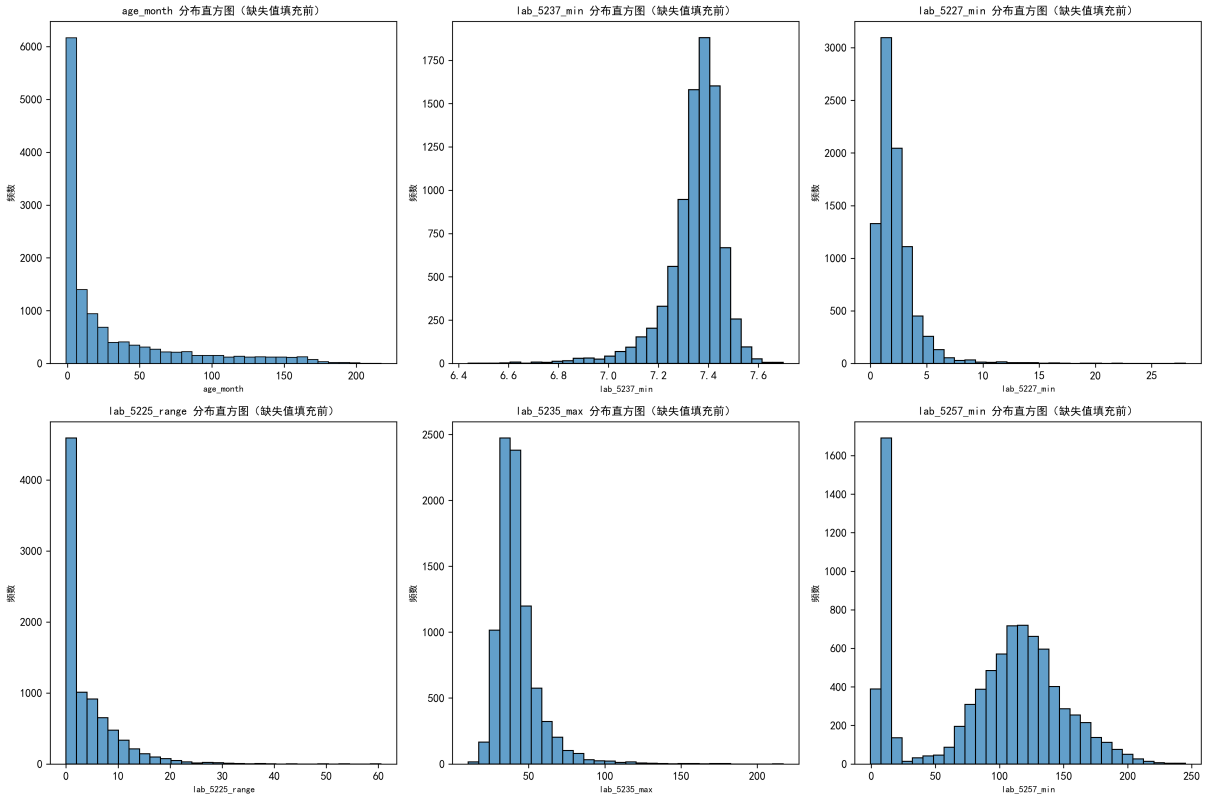


图 1: 数值特征分布直方图

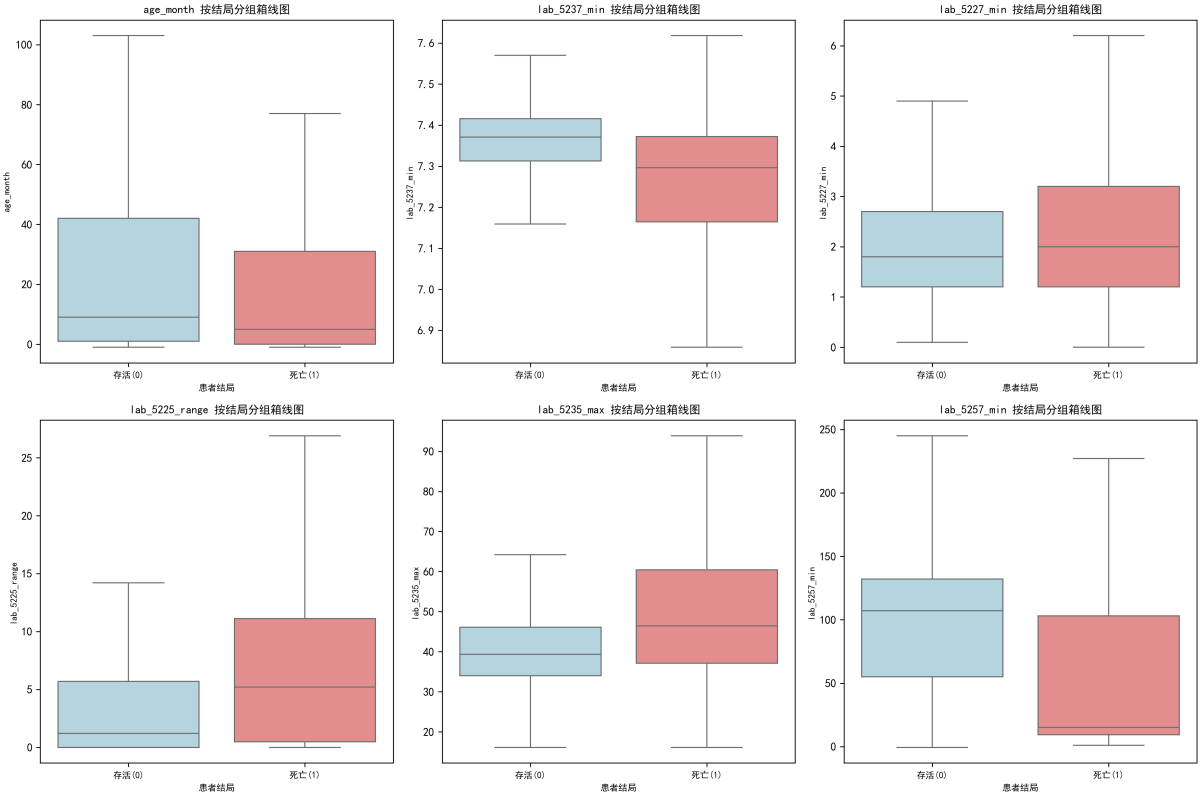


图 2: 按患者结局分组的特征箱线图

2.3 数据预处理

高质量的数据是建模的基础。针对前述发现的数据问题，本研究进行以下预处理。

2.3.1 缺失值处理

鉴于数据呈现偏态分布且含有离群点，平均值容易受极值影响，因此采用中位数填充法补全缺失值，以保持数据分布的稳健性。

Code Listing 3: 填充缺失值

```
# 使用中位数填充缺失值
data_no_na = data.fillna(data.median())
```

2.3.2 异常值处理

为了防止极端异常值干扰模型的训练，使用四分位距（interquartile range,IQR）法对数据进行截断处理。

Code Listing 4: 异常值处理

```
def handle_outliers(df, features):
    # 将超过 Q3 + 1.5IQR 或低于 Q1 - 1.5IQR 的值截断至边界
    for col in features:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower = Q1 - 1.5 * IQR
        upper = Q3 + 1.5 * IQR
        df[col] = df[col].clip(lower, upper)
    return df
```

2.3.3 数据标准化

不同变量的量纲差异巨大（如 `age_month` 范围 0-200，而 `lab_5237` 仅在 7.0-7.6 之间）。为了消除量纲影响，加速梯度下降的收敛，并确保 SVM 等基于距离的算法有效，采用 `StandardScaler` 对变量进行标准化处理，将数据转换为均值为 0、方差为 1 的标准正态分布，以保证各变量的量纲一致。

2.3.4 数据集划分

随机划分可能导致测试集中没有足够的正样本。因此，采用分层抽样将数据按 8:2 比例划分训练集与测试集，训练集包含 10606 个样本, 测试集含 2652 个样本，训练集和测试集死亡占比均为 5.9%。

Code Listing 5: 数据集划分

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
    test_size=0.2, random_state=42, stratify=y)
```

2.4 统计分析

2.4.1 差异性检验

将变量按“存活/死亡”结局分组，对各自变量进行独立样本 t 检验。结果如（表2）显示，所有纳入的 6 个变量在存活组与死亡组之间均存在极显著差异（$p < 0.05$）。`lab_5237_min`（$t = 13.874$）和 `lab_5257_min`（$t = 15.284$）具有极大的 t 值，说明这两项指标在两组间区分度极高，提示它们可能在后续建模中通过提供高信息增益，可作为预后预警信号。死亡组月龄的均值和中位数均低于存活组，且 $p = 0.026 < 0.05$，说明低月龄可能额是患者住院死亡的潜在风险因素。实验室指标 `lab_5225_range` 的死亡组均值与中位数显著高于存活组，$p < 0.001$，提示该指标的“高波动”与病情危重强关联；

表 2: 存活组-死亡组特征的差异性分析（$n_{存活} = 12478, n_{死亡} = 780$）

变量名	存活组		死亡组		t 值	p 值
	均值	中位数	均值	中位数		
age_month	26.333	9.0	23.505	5.0	2.224	0.026
lab_5237_min	7.367	7.368	7.338	7.345	13.874	<0.001
lab_5227_min	1.851	1.8	1.940	1.8	-3.291	0.001
lab_5225_range	2.418	1.5	3.998	2.2	-12.331	<0.001
lab_5235_max	39.797	39.6	42.730	40.4	-11.614	<0.001
lab_5257_min	102.427	105.0	82.867	87.5	15.284	<0.001

注：$p < 0.05$ 表示组间差异具有统计学显著性

2.4.2 相关性分析

变量间的强相关性会导致模型参数估计不稳定。为了分析变量间的相关性，本研究绘制了变量相关性热力图，如图3所示。绝大多数变量间的相关性系数绝对值均小于 0.5，说明特征间不存在强线性关联，多重共线性风险较低，可直接用于后续建模。`lab_5237_min` 与 `lab_5235_max` 存在中等程度相关，相关性系数约为 0.6，提示二者可能反映相似的生理状态，但未达到共线性剔除阈值。因此，所有变量均保留进入建模阶段。

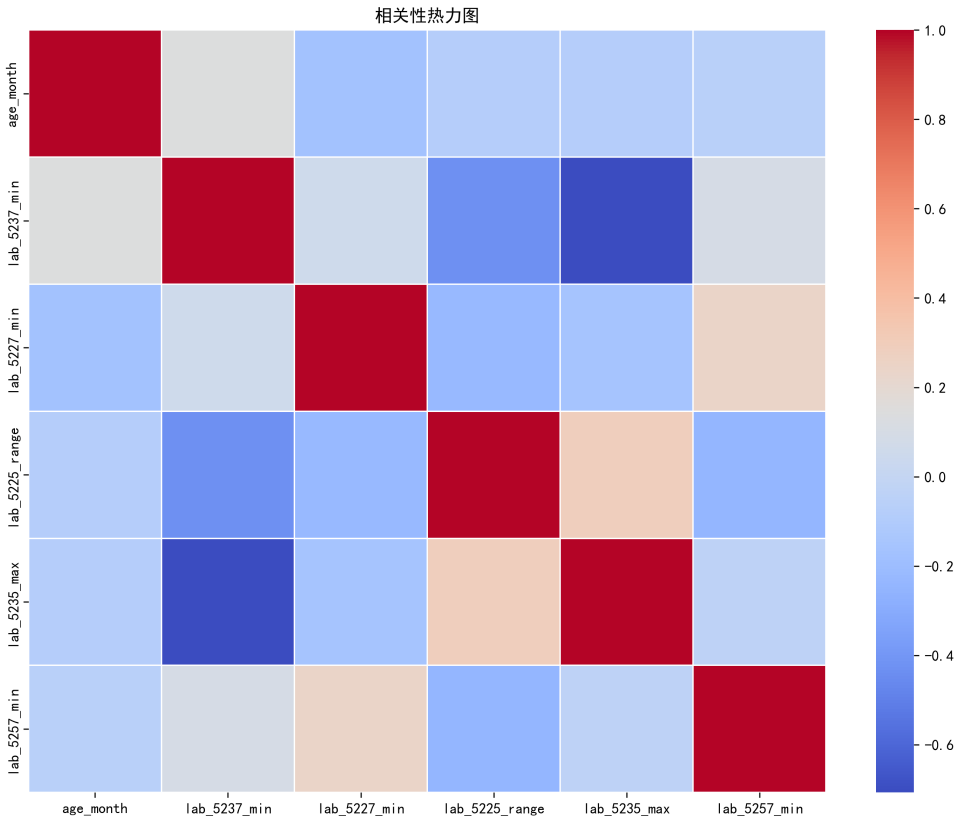


图 3: 变量相关性热力图

2.5 预测模型建立

本研究构建了三种不同机制的机器学习模型，包括逻辑回归（Logistic Regression，LR）、随机森林（Random Forest，RF）及支持向量机（Support Vector Machine，SVM），旨在寻找最适合该临床数据的算法。所有模型均采用 5 折交叉验证与 AUC 评分筛选最优模型，并统一设置 `class_weight='balanced'` 以校正类别不平衡 [3]。

2.5.1 逻辑回归模型

作为基准模型，逻辑回归具有可解释性强的优点。

- **调优参数**：正则化系数 $C \in [0.01, 100]$。
- **结果**：最优 $C = 0.01$，交叉验证 AUC $= 0.712$。线性模型表现尚可，但可能无法捕捉复杂的非线性生理机制。

Code Listing 6: 逻辑回归模型构建

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
lr_params = {'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l2'], 'solver': ['lbfgs']}
lr_grid = GridSearchCV(
    LogisticRegression(max_iter=1000, random_state=42, class_weight='balanced'),
    lr_params, cv=5, scoring='roc_auc', n_jobs=-1
)
lr_grid.fit(X_train, y_train)
best_lr = lr_grid.best_estimator_
```

2.5.2 随机森林模型

作为集成学习代表，随机森林通过构建多棵决策树来降低方差，适合处理非线性关系。

- **调优参数**：决策树数量 `n_estimators`，最大深度 `max_depth`，最小样本分裂数 `min_samples_split`。
- **结果**：最优参数为 $n_{\text{estimators}} = 100$，$max\_depth = 5$，$min\_samples\_split = 2$。交叉验证 AUC = **0.735**，优于逻辑回归模型，说明集成策略有效提升了泛化能力。

Code Listing 7: 随机森林模型构建

```
from sklearn.ensemble import RandomForestClassifier
rf_params = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 5, 10, 20],
    'min_samples_split': [2, 5, 10]
}
rf_grid = GridSearchCV(
    RandomForestClassifier(random_state=42, class_weight='balanced'),
    rf_params, cv=5, scoring='roc_auc', n_jobs=-1
)
rf_grid.fit(X_train, y_train)
best_rf = rf_grid.best_estimator_
```

2.5.3 支持向量机模型

构建支持向量机（SVM）模型，利用 RBF 核函数将数据映射到高维空间以实现线性可分。

- **调优参数**：正则化系数 `C`（0.1/1/10），核函数类型（`rbf/linear`）。
- **结果**：最优参数为 `C=0.1, kernel='rbf', gamma='scale'`。交叉验证 AUC = 0.722，优于逻辑回归（0.712），略低于随机森林（0.735），性能介于 LR 和 RF 之间。

Code Listing 8: 支持向量机模型构建

```
from sklearn.svm import SVC
svc_params = {
    'C': [0.1, 1, 10],
    'kernel': ['rbf', 'linear'],
    'gamma': ['scale']
}
svc_grid = GridSearchCV(
    SVC(probability=True, random_state=42, class_weight='balanced'),
    svc_params, cv=5, scoring='roc_auc', n_jobs=-1
)
svc_grid.fit(X_train, y_train)
best_svc = svc_grid.best_estimator_
```

2.6 无监督学习分析（KMeans 聚类）

为了探索除二分类外的患者内部结构，本研究基于 5 项实验室指标进行了 KMeans 聚类分析，旨在揭示患者风险亚群分布及与死亡结局的关联，为临床风险分层提供依据。

2.6.1 聚类数据预处理

选取 5 项实验室指标作为聚类特征，采用“均值填充缺失值 + 标准化”消除量纲影响，确保聚类结果不受特征量级干扰。

2.6.2 最佳簇数确定

利用肘部法则（Elbow Method）绘制簇内平方和（WCSS）曲线（图4），发现在 $K = 3$ 时曲线下降斜率发生明显转折，故选定将患者分为 3 个亚群。

Code Listing 9: 最佳簇数确定

```
# 肘部法则确定最佳K值
wcss = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state=42)
    kmeans.fit(data_cluster_scaled)
    wcss.append(kmeans.inertia_)  # 记录簇内平方和
```

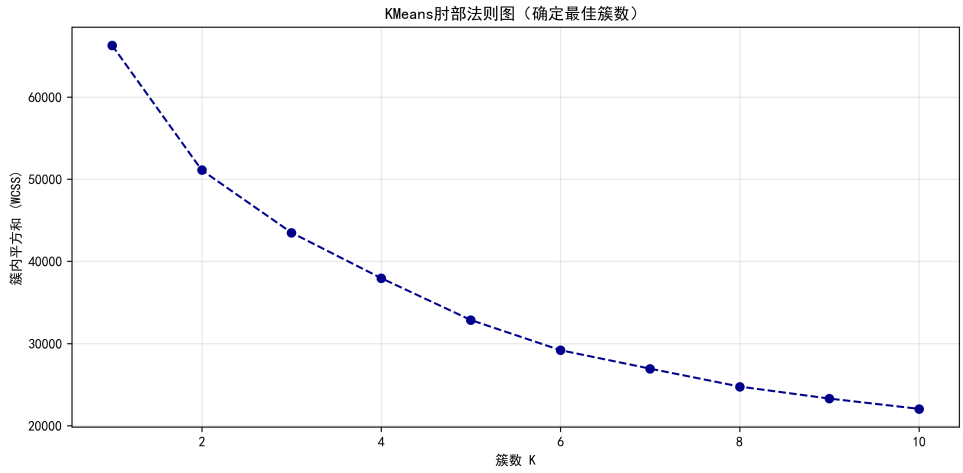


图 4: KMeans 肘部法则图

注：K=3 为最佳簇数，后续 WCSS 下降趋于平缓

2.6.3 风险亚群识别

采用最佳簇数 K=3 进行聚类训练，结合轮廓系数评估聚类质量。聚类后的临床结局分析（表 3、）揭示了极具临床意义的分层结果，如表 3所示：

- **聚类 0（低危组）**：占比 76.7%，死亡率仅 3.4%，这部分患儿生理指标最稳定。
- **聚类 1（高危组）**：占比 8.3%，但死亡率高达 **20.5%**，是临床需重点监护的对象。
- **聚类 2（中危组）**：占比 15.1%，死亡率 10.4%，属于中风险组。

表 3: 各聚类亚群的患者死亡率对比

聚类编号	样本数	死亡占比
0	10164	0.034
1	1095	0.205
2	1999	0.104

Code Listing 10: 聚类分析

```
# 最佳K值聚类训练
best_k = 3
kmeans = KMeans(n_clusters=best_k, init='k-means++', random_state=42)
cluster_labels = kmeans.fit_predict(data_cluster_scaled)
# 聚类质量评估（轮廓系数）
silhouette_avg = silhouette_score(data_cluster_scaled, cluster_labels
    )
# 聚类与临床结局关联
data_with_cluster = data.copy()
data_with_cluster['cluster'] = cluster_labels
cluster_outcome = data_with_cluster.groupby('cluster')['
    HOSPITAL_EXPIRE_FLAG'].agg(['count', 'mean']).round(3)
```

如图5所示，聚类分析成功识别出高风险亚群（聚类 1），其死亡风险是低风险亚群的 6 倍，可为临床优先干预提供依据。对聚类 1 患者加强监护频率、针对性治疗实验室指标异常，可降低死亡风险。

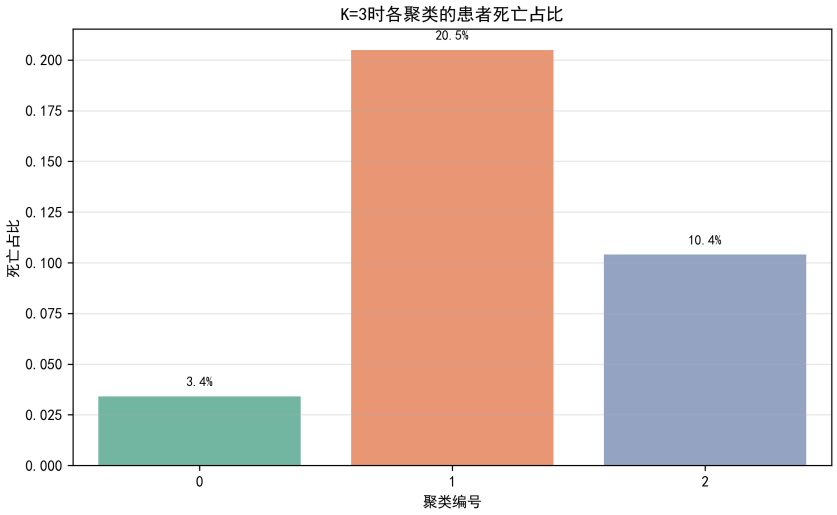


图 5: K=3 时各聚类的患者死亡占比

2.7 预测模型评估与可视化

2.7.1 核心评估指标对比

在独立测试集上对三个模型进行综合评估（表4）。由于数据不平衡，**准确率（Accuracy）**不再是核心指标，因为全猜存活也能达到 94% 准确率。我们需要关注 **AUC** 和**召回率**。随机森林模型展现了最佳的综合性能（AUC=0.7297，特异度 =0.8161），在保证较高查全率的同时，误报率较低。SVM 虽然召回率最高（0.5962），但其代价是极

低的精确率（0.1336），意味着存在大量假阳性，可能增加临床工作负担。而逻辑回归模型的各项指标均为三者最低，适配性最差，可能与临床数据多为非线性有关。

表 4: 模型在测试集上的性能评估

模型	准确率	召回率	精确率	特异度	F1 分数	AUC	Brier 分数
逻辑回归	0.7511	0.5321	0.1239	0.7648	0.2010	0.6792	0.2046
随机森林	**0.7994**	0.5321	0.1531	**0.8161**	**0.2378**	**0.7297**	0.1730
SVM	0.7489	**0.5962**	**0.1336**	0.2183	0.7584	0.7120	0.0525

2.7.2 ROC 曲线与混淆矩阵

ROC 曲线（图6）直观地展示了随机森林曲线（橙色）最靠近左上角，优于其他模型。SVM 曲线略低于随机森林（AUC=0.712），但优于逻辑回归（AUC=0.6792），三个模型的 AUC 均高于随机猜测，说明具备有效区分存活与死亡患者的能力。

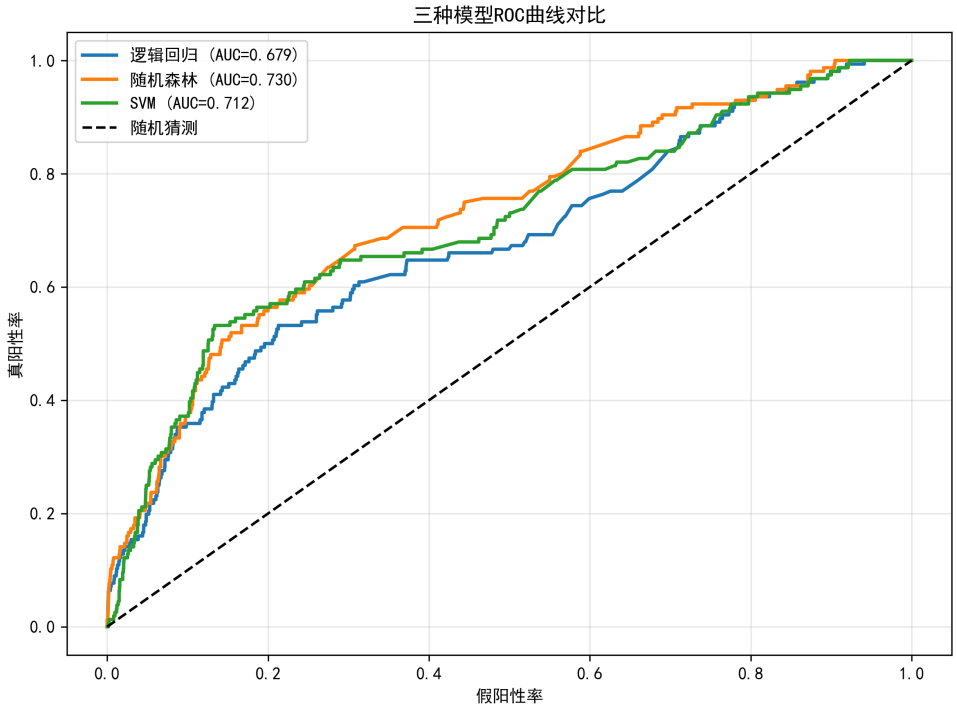


图 6: 三种模型 ROC 曲线对比

随机森林混淆矩阵（图7）进一步细化了预测结果，在随机森林模型中，成功识别了 83 例死亡患者（TP），但同时也漏报了 73 例（FN）。这提示模型虽有预警价值，但对死亡样本的识别能力有限，仍需结合临床医生判断以减少漏诊。

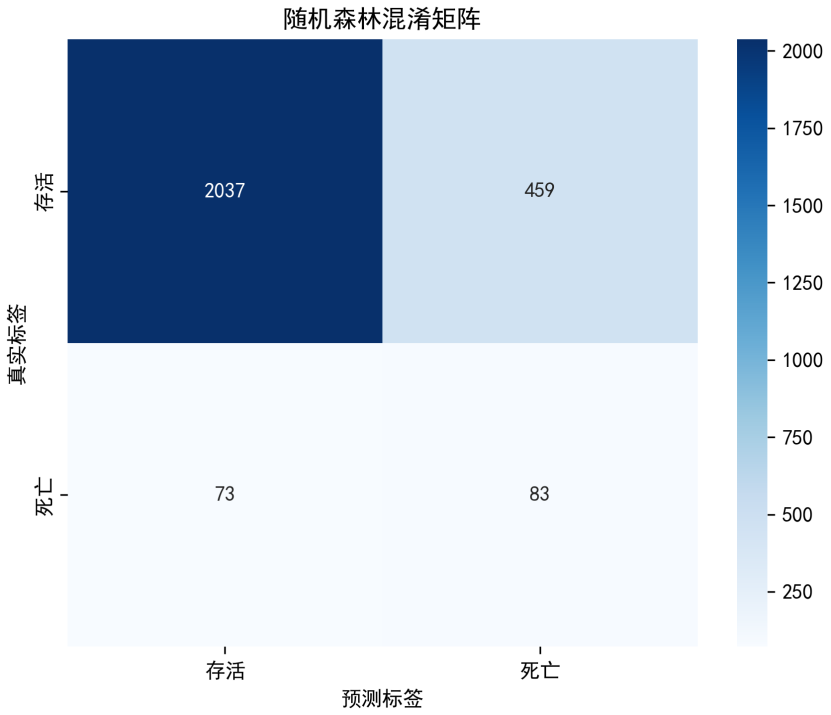


图 7: 随机森林混淆矩阵

Code Listing 11: 绘制 ROC 曲线

```
# 逻辑回归ROC
fpr_lr, tpr_lr, _ = roc_curve(y_test, lr_prob)
# 随机森林ROC
fpr_rf, tpr_rf, _ = roc_curve(y_test, rf_prob)
# SVM ROC
fpr_svc, tpr_svc, _ = roc_curve(y_test, svc_prob)
# 随机猜测线
plt.plot([0, 1], [0, 1], 'k--', label='随机猜测')
```

Code Listing 12: 绘制随机森林混淆矩阵

```
# 绘制随机森林混淆矩阵
y_pred_rf = best_rf.predict(X_test)
cm = confusion_matrix(y_test, y_pred_rf)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['存活
    ', '死亡'], yticklabels=['存活', '死亡'])
```

2.7.3 模型解释性分析

为了深入解模型判定患者高危程度的原因，本研究进行了特征重要性分析（图8）。结果显示 lab_5257_min 是最重要的预测因子。lab_5237_min 和 lab_5235_max 紧随

其后。本研究还对随机森林模型和逻辑回归模型进行了 SHAP 特征重要性分析，均呈现出相同的结果，如图 9所示 [4]。

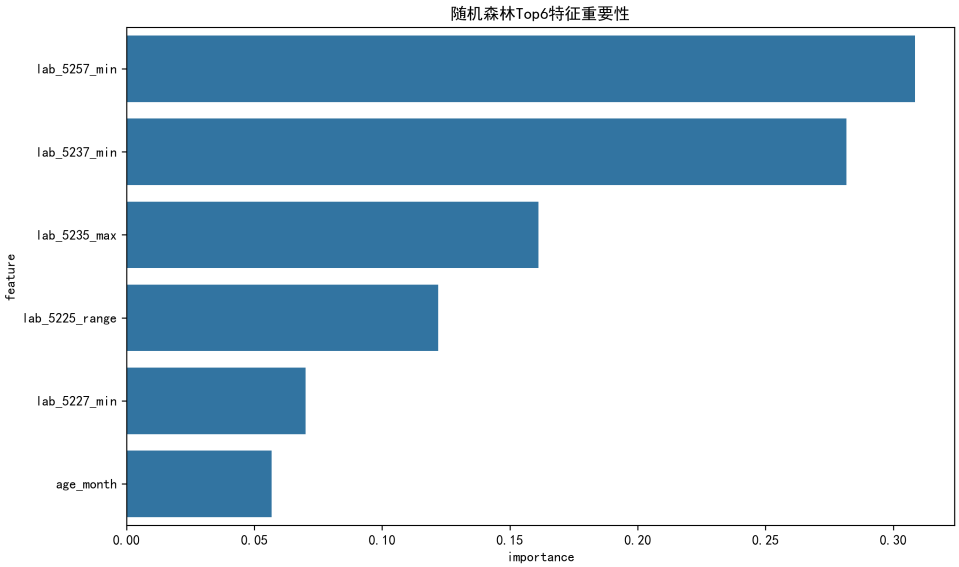


图 8: 随机森林特征重要性排序

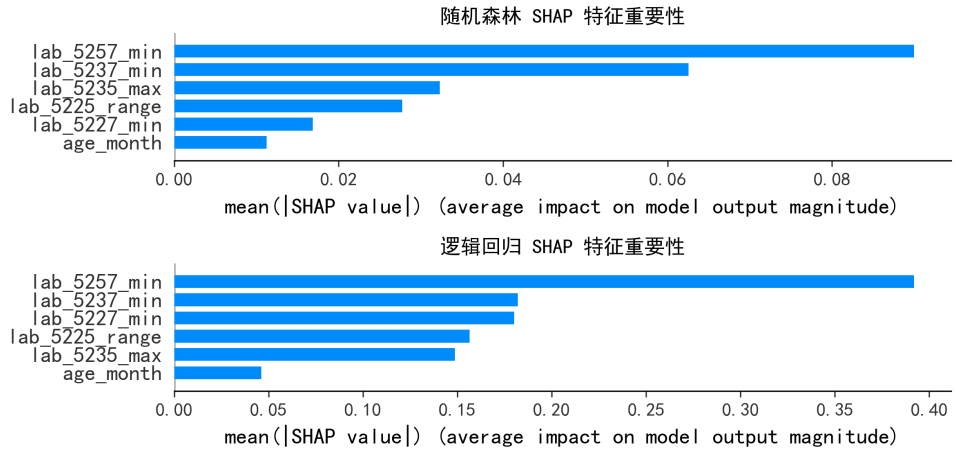


图 9: 随机森林与逻辑回归 SHAP 特征重要性分析

SHAP 依赖图（图10）进一步揭示了预测因子的方向性影响，`lab_5257_min` 的值越低，SHAP 值越高，呈显著负相关，提示 `lab_5257_min` 低水平显著增加死亡风险；SHAP 值随指标 `lab_5225_range` 的增大而上升，提示该指标波动增大会使死亡风险升高。`age_month` 在低月龄区间 SHAP 值较高，验证了“婴幼儿年龄越小风险更高”的临床常识。随机森林模型的 SHAP 依赖图如图10所示，逻辑回归模型的 SHAP 依赖图如图11所示。

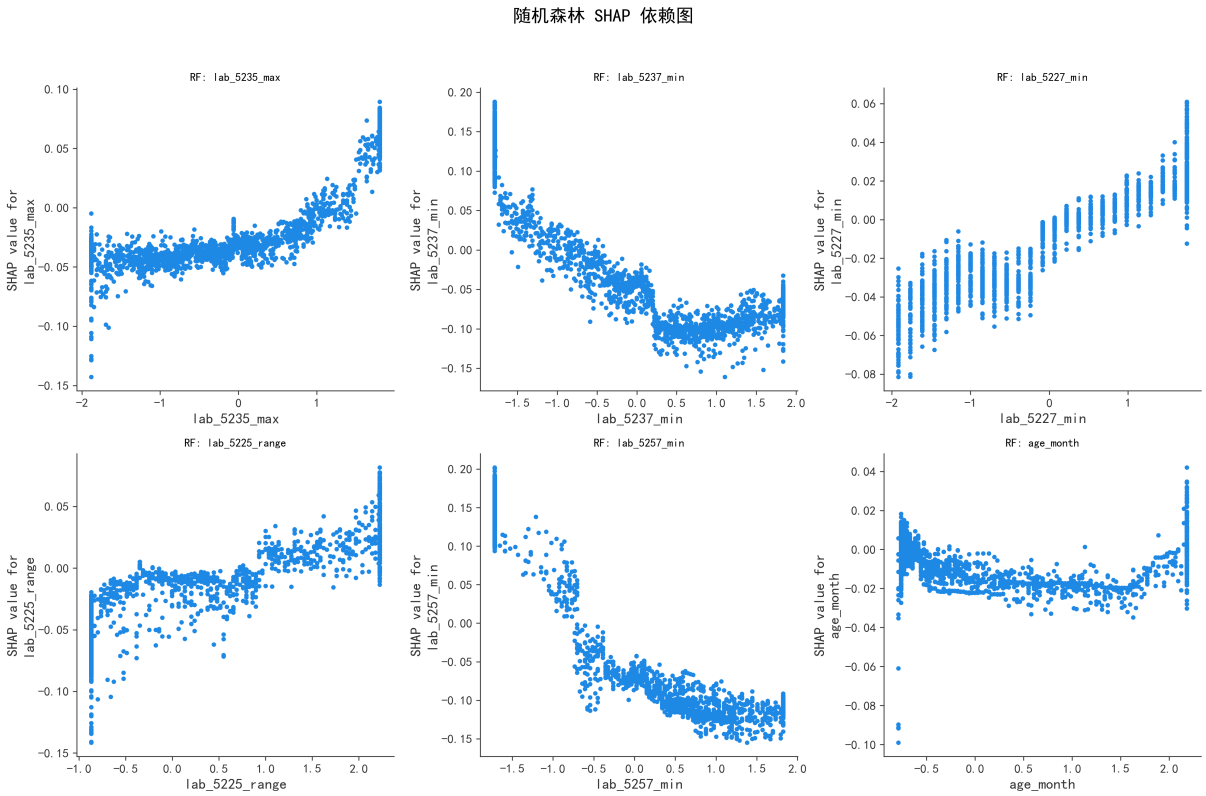


图 10: 随机森林模型的 SHAP 依赖图

Code Listing 13: 提取并可视化随机森林特征重要性

```
feature_importance = pd.DataFrame({
    'feature': selected_features, 'importance': best_rf.
        feature_importances_
}).sort_values('importance', ascending=False)
```

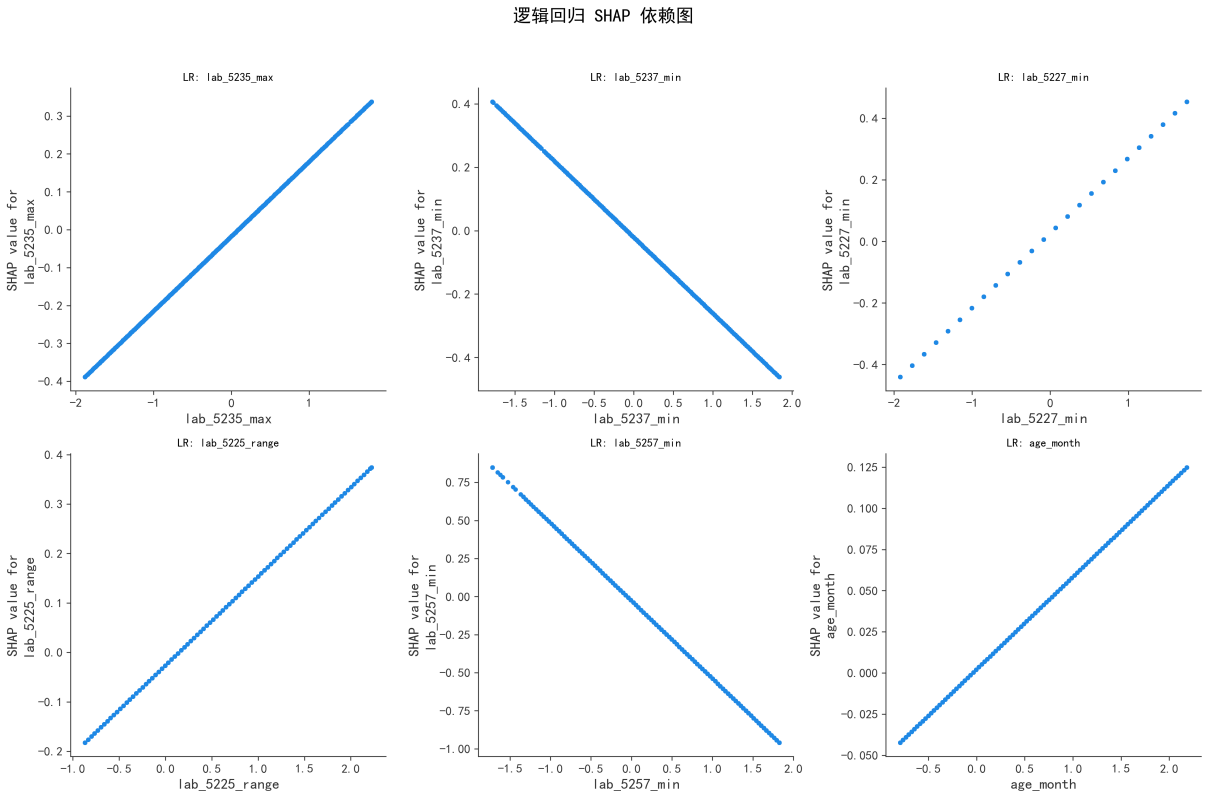


图 11: 逻辑回归模型的 SHAP 依赖图

Code Listing 14: 绘制 SHAP 特征重要性条形图

```
import shap
# 计算随机森林SHAP值
rf_explainer = shap.TreeExplainer(best_rf)
rf_shap_values = rf_explainer.shap_values(X_test_df)
rf_shap_values = rf_shap_values[1]  # 取死亡（正类）的SHAP值
# 计算逻辑回归SHAP值
lr_explainer = shap.LinearExplainer(best_lr, X_train_df)
lr_shap_values = lr_explainer.shap_values(X_test_df)
# 随机森林SHAP特征重要性条形图
shap.summary_plot(rf_shap_values, X_test_df, plot_type="bar", show=
    False)
# 逻辑回归SHAP特征重要性条形图
shap.summary_plot(lr_shap_values, X_test_df, plot_type="bar", show=
    False)
```

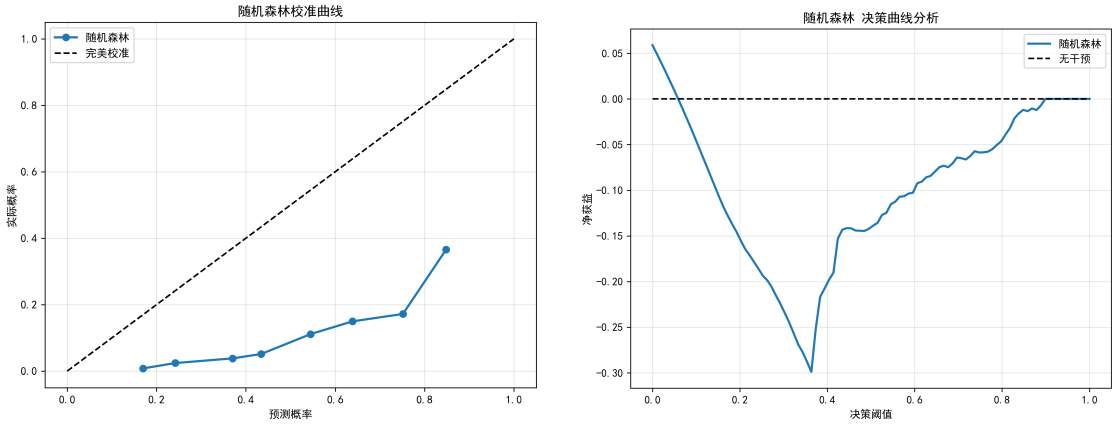
Code Listing 15: 绘制 SHAP 依赖图

```
# 随机森林模型SHAP依赖图
for i, feat in enumerate(dependence_features, 1):
    ax = plt.subplot(2, 3, i)
    shap.dependence_plot(feat, rf_shap_values, X_test_df,
                         interaction_index=None, show=False, ax=ax)
    ax.set_title(f"RF: {feat}", fontsize=10)
# 逻辑回归模型SHAP依赖图
for i, feat in enumerate(dependence_features, 1):
    ax = plt.subplot(2, 3, i)
    shap.dependence_plot(feat, lr_shap_values, X_test_df,
                         interaction_index=None, show=False, ax=ax)
    ax.set_title(f"LR: {feat}", fontsize=10)
```

2.7.4 模型预测可靠性与临床实用价值

本研究从模型预测可靠性与临床实用价值两个维度评估了随机森林模型的实际应用潜力，随机森林模型的校准曲线与决策曲线如图 12所示。随机森林的校准曲线（蓝色实线）整体沿理想校准线（黑色虚线）分布，仅在预测概率 0.5 ∼ 0.8 区间略有偏离，实际概率略低于预测概率，其余区间与理想线贴合度较高。这表明随机森林模型输出的死亡概率与患者真实死亡风险的一致性较好，预测结果具备一定的可靠性。

从随机森林模型的决策曲线可得，当决策阈值处于 0.05 ∼ 0.85 区间时，随机森林模型的净获益始终高于“无干预”基准，且在阈值 0.4 附近达到净获益峰值；仅当阈值 >0.8 时，模型净获益略低于无干预基准，此区间临床中极少采用，因为过高的阈值会遗漏大部分高风险患者。因此，在临床常用的决策阈值范围内，采用该随机森林模型指导风险分层决策，能够为临床带来正向的净获益，既可以有效识别高风险患者以实施针对性干预，又能避免对低风险患者的过度医疗，具备实际的临床应用价值。综上，本研究构建得随机森林模型不仅具备一定的概率预测可靠性，同时在多数临床决策场景下还能带来正向净获益。



(a) 校准曲线　　(b) 决策曲线

图 12: 随机森林模型校准与决策曲线

Code Listing 16: 绘制校准曲线和决策曲线

```
def plot_calibration_curve(model, X_test, y_test, model_name):
# 计算校准曲线
    prob_true, prob_pred = calibration_curve(y_test, model.
       predict_proba(X_test)[:,1], n_bins=10)
#决策曲线 (DCA)
def plot_dca(model, X_test, y_test):
    y_pred_prob = model.predict_proba(X_test)[:,1]
    thresholds = np.arange(0, 1.01, 0.01)
    net_benefit = []
    # 计算净获益
    for t in thresholds:
        tp = np.sum((y_pred_prob >= t) & (y_test == 1))
        fp = np.sum((y_pred_prob >= t) & (y_test == 0))
        n = len(y_test)
        nb = (tp / n) - (fp / n) * (t / (1 - t)) if t != 1 else 0
        net_benefit.append(nb)
# 绘制校准曲线和DCA曲线
plot_calibration_curve(best_rf, X_test, y_test, '随机森林')
plot_dca(best_rf, X_test, y_test)
```

2.8 项目展示网页开发

本项目的完整代码、分析报告及可视化结果已整理至个人 GitHub 展示页面，公开访问地址为：https://jianghuisu.github.io/portfolio/

3 结论与展望

本研究基于 PIC 数据集，利用 Python 进行了完整的数据分析和死亡风险预测模型构建与评估。

3.1 主要结论

1. 模型性能与优选

对比三种主流机器学习模型，随机森林模型展现了最佳的综合性能，明显优于逻辑回归和 SVM。这表明 PICU 临床数据中存在大量非线性关系和特征交互，集成学习算法比传统线性模型更适合此类场景。

2. 关键临床预测因子

特征重要性与 SHAP 分析一致发现，`lab_5257_min` 和 `lab_5237_min` 是影响患者生存的最核心指标。`lab_5257_min` 和 `lab_5237_min` 水平越低，患者死亡风险越高。另外，`lab_5225_range` 越大，死亡风险越高，说明这一生理指标的不稳定性是预后不良的危险因素。

3. 风险分层的价值

无监督 K-Means 聚类成功将患者分为三个风险层级。识别出的 **Cluster 1（高危组）**虽然仅占总人数的 8.3%，但其死亡率高达 20.5%，是低危组的 6 倍。这一发现可以为临床实现精准监护提供数据支持，识别出属于高危组特征的患儿，应立即启动更高级别的生命支持。

3.2 局限性与改进

尽管随机森林模型表现尚可，但本研究仍存在局限性：（1）数据不平衡：虽然使用了 Class Weighting，但精确率依然偏低，意味着模型存在较高的假阳性率，临床应用中可能引发 “报警疲劳”。（2）特征维度单一：目前仅依赖实验室指标和月龄，缺乏生命体征，如心率、血压等的时间序列数据。未来，可以通过引入 XGBoost 或 LightGBM 等梯度提升树模型，或尝试使用 SMOTE-ENN 等混合采样技术进一步解决类别不平衡问题。另外，PICU 数据本质是时间序列，可以引入长短期记忆网络（LSTM）或 Transformer 架构，利用患者生命体征的动态变化趋势来提升预测精度 [5]。结合非结构化数据，如医生的病程记录文本等，构建多模态预测模型，可以捕捉更多维度的病情信息，从而提升预测精度。

参考文献

[1] Xian Zeng, Gang Yu, Yang Lu, Linhua Tan, Xiujing Wu, Shanshan Shi, Huilong Duan, Qiang Shu, and Haomin Li. Pic, a paediatric-specific intensive care database. *Scientific Data*, 7(1):1–8, 2020.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[5] S. Sami et al. Predicting clinical outcomes using artificial intelligence and machine learning in neonatal intensive care units: a systematic review. *Journal of Perinatology*, 42(10):1273–1282, 2022.