心力衰竭患者 28 天再入院的多阶段统计分析与机器学习探索

姓名：陈威
学号：2511210403
学院：北医三院
专业：临床研究方法学
日期：2026 年 1 月 15 日

方法学总结

本研究基于 PhysioNet 公共数据库中来自中国四川地区的回顾性心力衰竭住院队列，纳入 2016–2019 年间 2,008 例患者，共包含 168 个临床变量[1]。研究以 28 天内是否再入院（re.admission.within.28.days）作为唯一结局变量。基于明确的研究目标、临床相关性、数据可用性及方法学可控性，从原始变量中筛选出 10 个研究变量（1 个因变量、9 个自变量），构建统一的分析框架。

在数据预处理阶段，对分类与等级变量进行重编码，并依据变量分布特征完成缺失值填补；随后按照结局变量进行分层随机抽样，将数据划分为训练集（70%）与测试集（30%）。统计分析阶段使用未填补缺失值的数据集，采用描述性统计、成对删除法的组间比较以及单因素 Logistic 回归，探索变量与再入院之间的关联。

在监督学习阶段，基于填补后数据集构建 Logistic Regression、Random Forest 与 Support Vector Machine 模型，统一在训练集上建模、在测试集上评估，并通过 ROC 曲线与 AUC 比较模型性能。随后，对 AUC 表现最优的模型进一步开展混淆矩阵分析及模型可解释性分析；若最优模型为 Logistic Regression，则同时汇报其多变量回归系数、OR 及 95% 置信区间。

在无监督学习阶段，基于填补后数据集中连续变量进行标准化处理，采用 K-means 聚类开展探索性分析。聚类数通过肘部法则与轮廓系数综合确定，并利用主成分分析（PCA）对聚类结果进行二维可视化展示。

主要结果总结

描述性统计显示，研究人群 28 天再入院率为 6.97%，以高龄患者为主，整体心功能状态较差，合并症负担较为常见。组间比较结果表明，NYHA 心功能分级、收缩压及肌酐水平在是否再入院的两组之间存在统计学差异。

单因素 Logistic 回归分析进一步显示，NYHA 心功能分级、收缩压及肌酐水平与 28 天再入院风险存在显著关联。多变量 Logistic Regression 分析中，仅收缩压在调整其他变量后仍与再入院风险保持统计学显著关联，表现为收缩压越低，再入院风险越高，其余变量未达到统计学显著性。

在监督学习模型比较中，三种模型整体区分能力有限，其中 Logistic Regression 在测试集中 AUC 最高（0.58），因此被选为主要分析模型。混淆矩阵结果显示该模型特异度较高但灵敏度较低。SHAP 分析表明，模型预测主要依赖连续临床指标，其中肌酐与收缩压呈稳定的正、负向影响，而 B 型利钠肽在短期再入院预测情境下呈现出不同于传统临床预后的方向性。

无监督学习结果显示，基于连续变量的 K-means 聚类在 K=4 时可形成相对稳定的聚类结构。PCA 可视化提示研究人群中可能存在具有不同连续临床特征模式的潜在亚群，为进一步探索患者异质性提供了依据。

[1] Zhang, Zhongheng, et al. “Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data” (version 1.3). PhysioNet (2022). RRID: SCR_007345. https://doi.org/10.13026/5m60-vs44

阶段一：变量筛选

本研究原始数据库（Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data），该数据库为基于中国四川地区真实世界电子病历构建的回顾性心力衰竭数据集，收录了 2016–2019 年间 2,008 例住院心力衰竭患者的 168 个临床变量。该数据集弥补了中国人群心力衰竭相关高质量公共数据的不足，为开展流行病学研究及真实世界证据分析提供了重要数据基础。

为保证研究目标聚焦、分析流程可控并提升结果的临床可解释性，本研究未采用“全变量建模”，而是依据临床相关性与统计学可行性进行预先筛选，最终纳入 10 个变量（1 个因变量、9 个自变量）进入后续各阶段分析。

本研究因变量为：re.admission.within.28.days（28 天内是否再入院）。该指标直接反映心力衰竭患者的短期不良结局与医疗资源再利用情况，具有明确且重要的临床意义，因此被设定为唯一结局变量。

本研究自变量共 9 个，分别为：

- systolic.blood.pressure（收缩压）
- pulse（脉搏）
- albumin（白蛋白）
- brain.natriuretic.peptide（B 型利钠肽）
- creatinine.enzymatic.method（肌酐-酶法）
- gender（性别）
- ageCat（年龄分组）
- NYHA.cardiac.function.classification（NYHA 心功能分级）
- CCI.score（Charlson 合并症指数评分）

变量筛选主要基于以下依据：第一，优先选择与心力衰竭病情严重程度、器官功能状态及整体合并症负荷密切相关且临床意义明确的指标（如收缩压/脉搏反映血流动力学状态，白蛋白反映营养与炎症状态，利钠肽反映心衰负荷，肌酐反映肾功能，NYHA 与 CCI 反映功能分级与合并症负担），以确保模型输出具备临床解释价值。第二，从数据质量与分析可行性出发，纳入的变量类型覆盖连续（正态/右偏）、二分类与等级变量，便于在描述性统计、组间比较、单因素回归、监督学习与无监督学习等多阶段分析中保持一致的变量体系与可重复性。第三，控制自变量数量可降低高维变量带来的噪声与过拟合风险，从而提升统计检验与模型评估的稳定性。第四，考虑到变量缺失情况，本研究优先选择缺失比例相对较低、缺失模式较为稳定且适合进行合理填补处理的变量，以避免因大量缺失数据导致样本量显著下降或引入额外偏倚，从而保证后续统计分析与建模结果的可靠性与稳健性。

综上，本研究在 166 个候选变量中，围绕“28 天再入院”这一明确结局，基于临床相关性、数据可用性与方法学可控性，筛选并确定上述 10 个变量作为后续分析的统一研究对象。

阶段二：数据预处理

在数据预处理阶段，对变量进行了统一的重编码处理、缺失值填补以及数据集划分。

对部分分类与等级变量进行了重编码。年龄变量 ageCat 按原始分组区间由低到高依次编码为 1–8 级，具体对应为：(21,29] 记为 1，(29,39] 记为 2，(39,49] 记为 3，(49,59] 记为 4，(59,69] 记为 5，(69,79] 记为 6，(79,89] 记为 7，(89,110] 记为 8。NYHA.cardiac.function.classification 按心功能分级的临床严重程度递增进行编码，其中 I、II、III、IV 级分别记为 1、2、3、4。性别变量 gender 采用二分类编码，Male 记为 0，Female 记为 1。

对缺失值进行了系统处理。由于二分类变量在原始数据中不存在缺失值，因此未进行额外处理。对于连续变量，根据其分布特征采用不同的填补策略：近似正态分布的连续变量采用均值填补，非正态分布的连续变量采用中位数填补。对于等级变量，同样采用中位数进行填补。

在完成变量重编码与缺失值填补后，对数据集进行了分层随机划分。依据因变量 re.admission.within.28.days 的类别分布进行分层，以确保再入院与未再入院患者在训练集与测试集中所占比例一致；随后在各分层内按比例随机抽样，将整体样本划分为训练集与测试集，其中训练集占 70%，测试集占 30%。

阶段三：统计分析

3.1 描述性统计

表 1: 研究对象的基线特征描述性统计

variable	descriptive statistics	missing values
gender		0 (0.00%)
Male	845 (42.08%)	
Female	1163 (57.92%)	
ageCat		0 (0.00%)
(21,29]	4 (0.20%)	
(29,39]	12 (0.60%)	
(39,49]	56 (2.79%)	
(49,59]	106 (5.28%)	
(59,69]	368 (18.33%)	
(69,79]	715 (35.61%)	
(79,89]	646 (32.17%)	
(89,110]	101 (5.03%)	
NYHA.cardiac.function.classification		0 (0.00%)
II	353 (17.58%)	
III	1039 (51.74%)	
IV	616 (30.68%)	
CCI.score		5 (0.25%)
0	56 (2.79%)	
1	770 (38.35%)	
2	699 (34.81%)	
3	368 (18.33%)	
4	94 (4.68%)	
5	15 (0.75%)	
6	1 (0.05%)	
systolic.blood.pressure	131.25 ± 24.23	3 (0.15%)
pulse	85.28 ± 21.46	1 (0.05%)
albumin	36.53 ± 4.98	102 (5.08%)
brain.natriuretic.peptide	753.03 (303.94, 1738.52)	35 (1.74%)
creatinine.enzymatic.method	87.10 (64.90, 122.70)	23 (1.15%)
re.admission.within.28.days		0 (0.00%)
0	1868 (93.03%)	
1	140 (6.97%)	

对因变量与各自变量按变量类型汇总：对于分类，报告各类别的频数及百分比 [n (%)]，并报告缺失值个数及其百分比 [n (%)]；对于近似正态分布的连续变量，报告均值 ± 标准差（mean±SD），并同

时报告缺失值个数及其百分比 [n (%)]；对于右偏分布的连续变量，报告中位数及四分位距 [median (IQR)]，并报告缺失值个数及其百分比 [n (%)]；

分类变量方面，研究总体样本量为 2008 例。性别分布为男性 845 例（42.08%）、女性 1163 例（57.92%），无缺失。年龄分组（ageCat）以（69,79] 与（79,89] 两组占比最高，分别为 715 例（35.61%）与 646 例（32.17%），其余年龄组占比较低，且 ageCat 无缺失。NYHA 心功能分级（NYHA.cardiac.function.classification）以 III 级占比最高（1039 例，51.74%），其次为 IV 级（616 例，30.68%）与 II 级（353 例，17.58%），该变量无缺失。Charlson 合并症指数（CCI.score）以 1 分（770 例，38.35%）与 2 分（699 例，34.81%）为主，其次为 3 分（368 例，18.33%），总体缺失 5 例（0.25%）。结局变量显示，28 天内再入院（re.admission.within.28.days=1）为 140 例（6.97%），未再入院为 1868 例（93.03%），结局变量无缺失。

连续变量方面，收缩压（systolic.blood.pressure）为 131.25±24.23 mmHg，缺失 3 例（0.15%）；脉搏（pulse）为 85.28±21.46 次/分，缺失 1 例（0.05%）；白蛋白（albumin）为 36.53±4.98 g/L，缺失 102 例（5.08%）。右偏连续变量方面，B 型利钠肽（brain.natriuretic.peptide）中位数为 753.03（IQR：303.94–1738.52），缺失 35 例（1.74%）；肌酐-酶法（creatinine.enzymatic.method）中位数为 87.10（IQR：64.90–122.70），缺失 23 例（1.15%）。总体来看，连续变量中缺失比例最高的是白蛋白（5.08%），其余连续变量缺失比例均 1.74%。

为直观展示各变量的分布特征，在描述性统计的基础上，本研究进一步进行了可视化分析。针对分类及等级变量，绘制了条形图，以展示各类别的样本数量分布情况；针对连续变量，绘制了直方图并叠加正态分布曲线，用以评估变量的分布形态及其与正态分布的吻合程度。

分类变量的条形图显示，研究人群以高龄患者为主，年龄主要集中在 69–79 岁和 79–89 岁区间；性别分布中女性略多于男性。NYHA 心功能分级以 III 级和 IV 级占多数，提示入组患者整体心功能状态较差。CCI 评分主要集中在 1–3 分区间，反映研究对象普遍合并一定程度的慢性疾病负担。结局变量方面，28 天内再入院事件所占比例较低，整体样本呈现明显的类别不平衡特征。

连续变量的分布图显示，白蛋白（albumin）、脉搏（pulse）及收缩压（systolic.blood.pressure）的直方图整体呈现近似对称分布，叠加的正态曲线与实际分布拟合较好，提示其可视为近似正态分布变量。相比之下，B 型利钠肽（brain.natriuretic.peptide）与肌酐（creatinine.enzymatic.method）的分布明显右偏，存在长尾现象，提示少数患者具有显著升高的指标水平，与其作为疾病严重程度和器官功能指标的临床特征相一致。

在直方图与条形图的基础上，本研究进一步为所有连续变量绘制了箱线图，以更直观地展示变量的集中趋势、离散程度及潜在异常值分布情况。

箱线图结果显示，收缩压（systolic.blood.pressure），脉搏（pulse）及白蛋白（albumin）的中位数位于箱体中央，四分位距相对集中，仅存在少量上下端离群点，整体分布较为稳定，与其近似正态分布特征一致。相比之下，B 型利钠肽（brain.natriuretic.peptide）及肌酐（creatinine.enzymatic.method）的箱线图呈现明显的右偏特征，上须较长且高值离群点数量较多，提示少数患者存在显著升高的指标水平，反映了疾病严重程度及肾功能受损的个体差异。

总体而言，箱线图进一步验证了连续变量在分布形态上的异质性，与前述直方图结果相互印证，为后续分析中连续变量的分布判定及统计方法选择提供了直观依据。

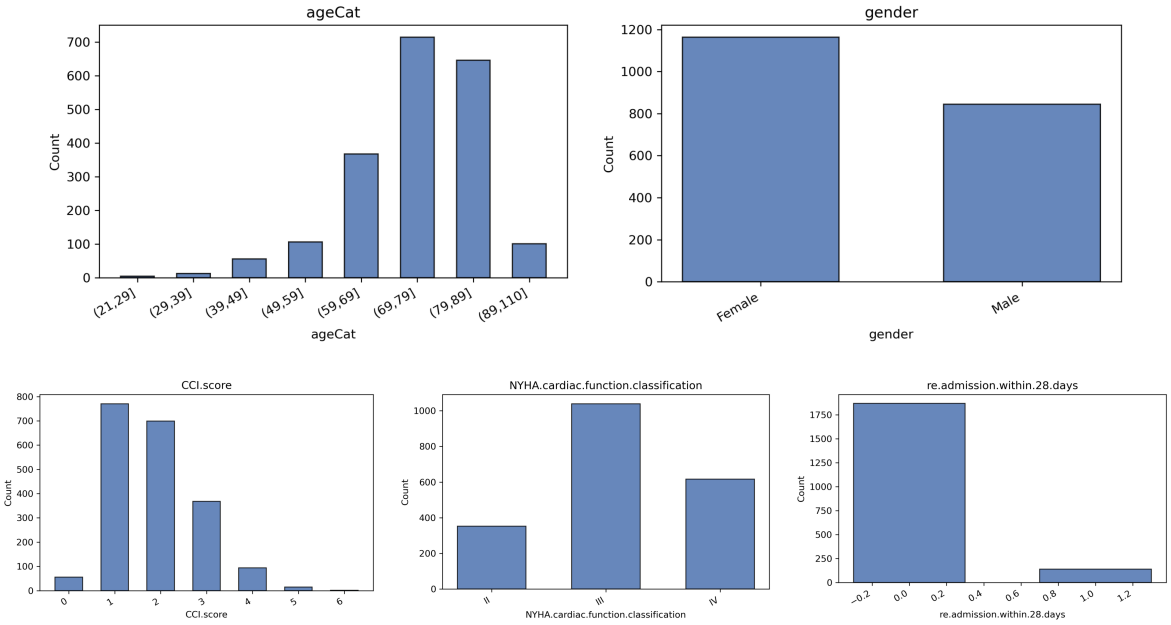

图 1: 分类变量条形图

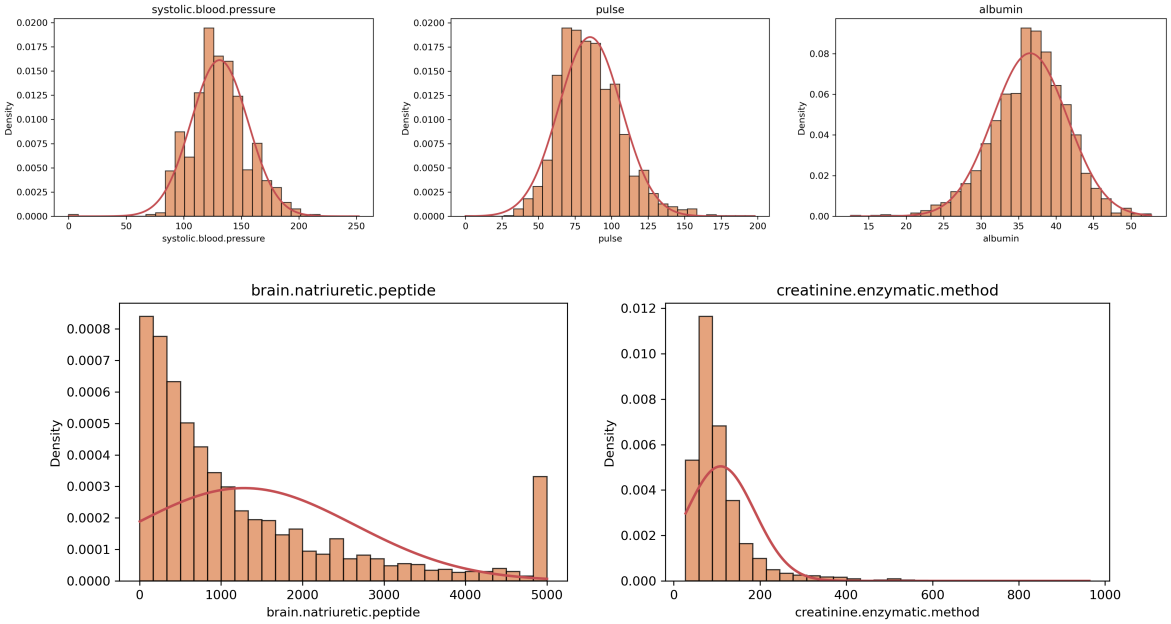

图 2: 连续变量的直方图叠加正态曲线图

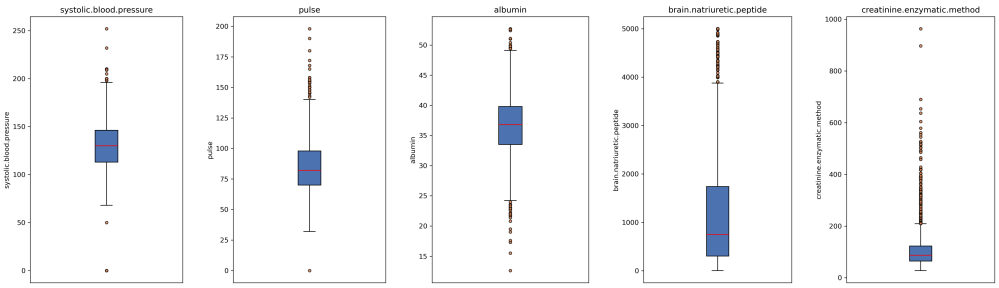

图 3: 连续变量的箱线图

3.2 分析性统计

在分析性统计阶段中，使用未进行缺失值填补的原始数据，并采用成对删除（pairwise deletion）策略处理缺失值。在每一项具体分析中，仅纳入该分析所涉及变量均为完整记录的样本。

表 2: 不同 28 天再入院状态下各变量的组间比较

variable	group0 (n=1868)	group1 (n=140)	statistic	p
gender (male)	777 (42%)	68 (49%)	2.322	0.128
ageCat	6 (5, 7)	6 (6, 7)	122875.5	0.213
NYHA.cardiac.function.classification	3 (3, 4)	3 (3, 4)	110091.5	<0.001
CCI.score	2 (1, 2)	2 (1, 3)	119821	0.118
systolic.blood.pressure	131.62 ± 24.18	126.41 ± 24.47	2.432	0.016
pulse	85.34 ± 21.52	84.44 ± 20.66	0.498	0.619
albumin	36.54 ± 4.98	36.40 ± 4.99	0.303	0.762
brain.natriuretic.peptide	1264.90 ± 1339.29	1490.27 ± 1531.19	-1.67	0.097
creatinine.enzymatic.method	107.83 ± 78.44	123.29 ± 86.61	-2.051	0.042

以因变量 re.admission.within.28.days 为分组依据，将研究对象分为未再入院组（group 0）和 28 天内再入院组（group 1），对各自变量在两组间进行比较。根据变量类型选择相应的统计检验方法：连续变量采用 Welch t 检验；二分类变量采用卡方检验；等级变量采用 Mann–Whitney U 检验。所有检验均为双侧检验，结果以检验统计量及 P 值进行报告。

基于成对删除法的组间比较结果显示，在所分析的自变量中，部分变量在是否发生 28 天再入院的两组之间存在统计学显著差异。

等级变量方面，NYHA 心功能分级在两组之间差异显著（Mann–Whitney U 检验，$P < 0.001$），再入院组整体呈现更高的心功能分级，提示心功能状态较差的患者更易在短期内发生再入院。

连续变量方面，收缩压（systolic.blood.pressure）在两组之间存在显著差异（Welch t 检验，$P = 0.016$），再入院组的平均收缩压低于未再入院组；肌酐水平（creatinine.enzymatic.method）在两组间亦存在统计学差异（Welch t 检验，$P = 0.042$），再入院组肌酐水平相对较高，提示肾功能受损可能与短期再入院风险相关。

其余自变量在组间比较中未达到统计学显著性水平。总体而言，组间比较分析初步识别出 NYHA 心功能分级、收缩压及肌酐水平与 28 天再入院存在统计学关联。

表 3: 基于单因素 Logistic 回归的再入院风险分析

variable	OR	95% CI	p
systolic.blood.pressure	0.99	0.99 (0.98,1.00)	0.014
pulse	1	1.00 (0.99,1.01)	0.63
albumin	0.99	0.99 (0.96,1.03)	0.761
brain.natriuretic.peptide	1	1.00 (1.00,1.00)	0.062
creatinine.enzymatic.method	1	1.00 (1.00,1.00)	0.028
gender	0.75	0.75 (0.53,1.06)	0.108
ageCat	1.1	1.10 (0.94,1.29)	0.228
NYHA.cardiac.function.classification	1.58	1.58 (1.21,2.05)	<0.001
CCI.score	1.17	1.17 (0.98,1.39)	0.079

在单因素 logistic 回归分析中，9 个自变量分别与 28 天内是否再入院这一因变量进行建模分析。等级变量在本阶段作为连续变量处理，用于探索其与结局之间的线性关联趋势。结果以比值比（odds ratio，OR）、95% 置信区间（95% CI）及 P 值进行汇报。

分析结果显示，NYHA 心功能分级与 28 天再入院风险呈显著正相关（$OR = 1.58$，95% CI：1.21–2.05，$P < 0.001$），提示心功能分级每升高一级，再入院风险显著增加。收缩压（systolic.blood.pressure）与再入院风险呈显著负相关（$OR = 0.99$，95% CI：0.98–1.00，$P = 0.014$），表明较低的收缩压水平与更高的短期再入院风险相关。肌酐水平（creatinine.enzymatic.method）亦与再入院风险显著相关（$OR = 1.00$，95% CI：1.00–1.00，$P = 0.028$），提示肾功能受损可能增加短期再入院的可能性。

其余自变量在单因素 logistic 回归分析中未达到统计学显著性水平。总体而言，单因素回归分析进一步支持 NYHA 心功能分级、收缩压及肌酐水平与 28 天再入院之间的统计学关联。

阶段四：监督学习模型的建立与评估

在预测模型建立阶段，本研究基于已完成缺失值填补并完成分层随机划分的数据集开展分析。所有模型均在训练集上进行拟合与参数学习，并在独立的测试集上进行性能评估，以保证不同模型在相同样本划分和一致的数据基础上进行比较，从而避免因缺失处理或样本使用方式不同而引入偏倚。本阶段的研究目标是在统一的变量体系和处理规则下，构建并比较多种监督学习模型对 28 天再入院风险的预测能力及其泛化表现。

结局变量为 re.admission.within.28.days，为二分类变量，编码方式为 0 = 否，1 = 是。该变量仅作为模型的预测目标输入，不进行任何形式的标准化或变换。

自变量的处理遵循统一且可解释的规则，并根据变量类型及模型特性进行差异化处理。连续变量包括 systolic.blood.pressure、pulse、albumin、brain.natriuretic.peptide 及 creatinine.enzymatic.method。其中，brain.natriuretic.peptide 与 creatinine.enzymatic.method 在原始分布中呈明显右偏，因此在建模前进行对数转换（log transformation），以降低极端值对模型训练的影响。标准化处理根据模型类型进行：在 Logistic Regression 和 Support Vector Machine（SVM）模型中，对连续变量进行标准化；在 Random Forest 模型中，由于其基于树结构，对变量尺度不敏感，因此连续变量保留原始数值，不进行标准化。

二分类变量为 gender，采用 0/1 编码方式，不进行标准化处理，并直接作为数值变量输入所有模型。

等级变量包括 ageCat、NYHA.cardiac.function.classification 及 CCI.score。上述变量在原始数据中已按等级顺序完成编码，本阶段不再重新编码，也不进行标准化处理，而是在各模型中作为具有顺序意义的有序数值变量使用，以保留其潜在的梯度信息。

在具体建模过程中，不同模型采用如下最终输入策略：对于 Logistic Regression，连续变量经对数转换（如适用）后进行标准化，等级变量采用有序编码但不标准化，二分类变量以 0/1 形式输入；对于 Random Forest，连续变量使用原始数值（不标准化），等级变量采用有序编码，二分类变量以 0/1 形式输入；对于 SVM，连续变量在对数转换后进行标准化，等级变量保持有序编码且不标准化，二分类变量以 0/1 形式输入。

模型性能评估以受试者工作特征曲线（ROC）及曲线下面积（AUC）为主要指标。在多种模型比较的基础上，选择 AUC 值最高的模型作为表现最优模型，并对其进行进一步的性能评估与解释性分析。

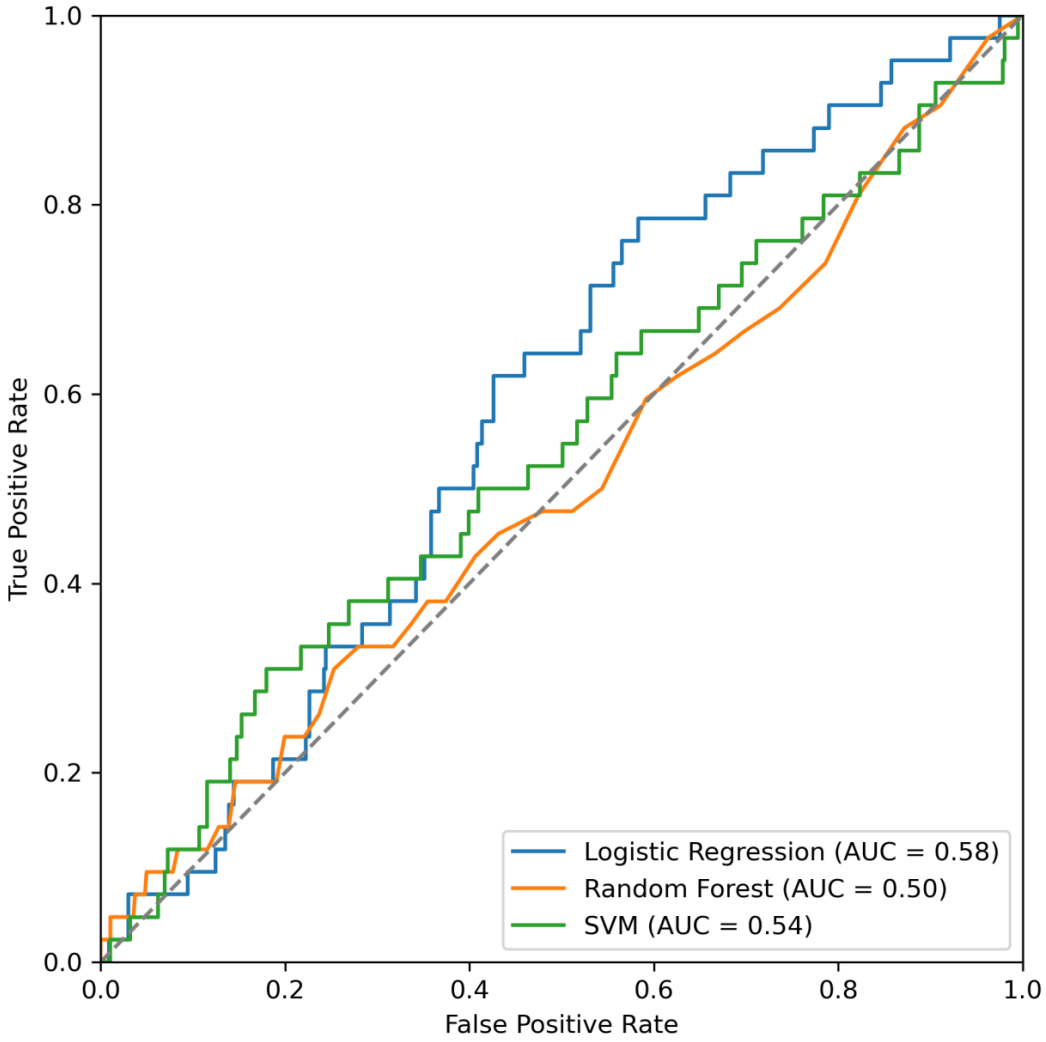

图 4: 不同预测模型对 28 天再入院的 ROC 曲线比较

ROC 曲线比较结果显示，在测试集中三种监督学习模型对 28 天再入院的预测能力整体有限，但模型之间仍存在一定差异。Logistic Regression 的 AUC 为 0.58，高于 SVM（AUC = 0.54）和 Random Forest（AUC = 0.50），表现出相对更好的区分能力。相比之下，Random Forest 的 ROC 曲线基本贴近对角线，其预测性能接近随机分类。

总体来看，三种模型的 AUC 均未达到较高水平，提示仅基于当前纳入的临床变量，对 28 天再入院进行预测的难度较大，单一模型的判别能力有限。在模型间比较中，Logistic Regression 在区分再入院与未再入院患者方面表现相对最优，因此被选定为后续进一步分析与解释的模型。

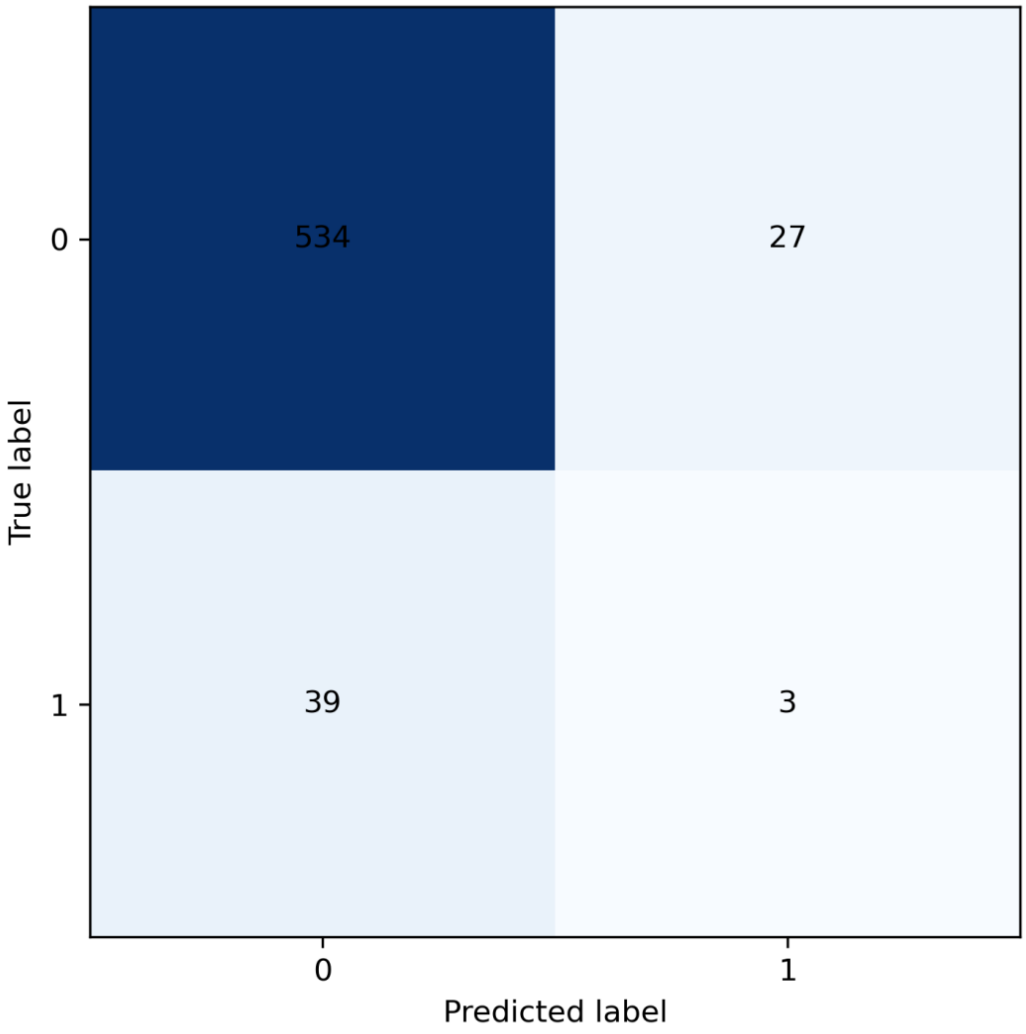

图 5: Logistic Regression 模型在测试集上的混淆矩阵

在混淆矩阵分析中，预测概率以 0.5 作为分类阈值，将模型输出转化为二分类结果。混淆矩阵结果显示，在测试集中，Logistic Regression 模型对未发生 28 天再入院（真实标签为 0）的识别能力较强，其中 534 例被正确预测为未再入院，27 例被误判为再入院。相比之下，该模型对再入院事件（真实标签为 1）的识别能力较弱，仅有 3 例被正确预测为再入院，而有 39 例再入院患者被误判为未再入院。整体来看，该模型具有较高的特异度，但灵敏度明显不足，提示其更倾向于预测患者为低风险人群。这一结果与样本中再入院事件比例较低的类别不平衡特征一致，也反映出在当前变量体系下，模型对短期再入院高风险患者的识别能力有限。

表 4: Logistic Regression 模型回归系数、比值比及其 95% 置信区间

variable	coefficient ( )	OR	95% CI lower	95% CI upper
ageCat	0.17	1.18	0.96	1.46
gender	-0.19	0.83	0.54	1.28
systolic.blood.pressure	-0.28	0.75	0.60	0.94
pulse	0.07	1.07	0.87	1.32
NYHA.cardiac.function.classification	0.29	1.34	0.97	1.84
brain.natriuretic.peptide	-0.02	0.98	0.78	1.22
creatinine.enzymatic.method	0.18	1.20	0.97	1.48
albumin	0.10	1.10	0.88	1.38
CCI.score	0.15	1.17	0.94	1.45

多变量 Logistic Regression 分析显示，在纳入 9 个自变量后，仅收缩压（systolic.blood.pressure）与 28 天再入院风险存在统计学显著关联（OR = 0.75，95%CI：0.60–0.94），方向性为负向，即收缩压越低，再入院风险越高。其余自变量均未达到统计学显著性（95%CI 均跨越 1）。其中，年龄分组（ageCat）、脉搏（pulse）、NYHA 心功能分级、肌酐（creatinine.enzymatic.method）、白蛋白（albumin）及 CCI 评分的 OR 均大于 1，呈正向趋势；性别（gender）及 B 型利钠肽（brain.natriuretic.peptide）的 OR 小于 1，呈负向趋势。上述方向性仅作趋势性描述，不作统计学结论。

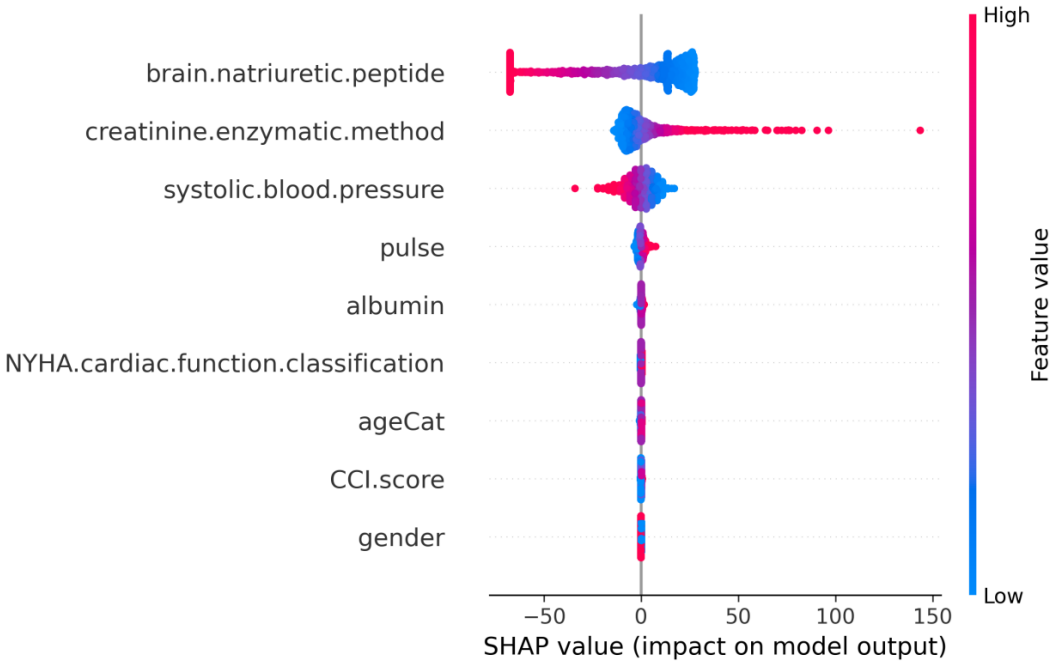

图 6: Logistic Regression 模型的 SHAP 汇总图（特征重要性与方向性）

SHAP 汇总图按平均绝对 SHAP 值由高到低展示了各变量对模型预测结果的相对重要性及其方向性。其中，brain.natriuretic.peptide、creatinine.enzymatic.method 和 systolic.blood.pressure 是对模型输出影响最大的变量。需要注意的是，brain.natriuretic.peptide 的高取值（红色）主要分布在负的 SHAP 值区域，而低取值（蓝色）更多分布在正的 SHAP 值区域，表明在该模型中，较高的

BNP 水平倾向于降低模型预测为 28 天再入院的概率，而较低 BNP 水平则推动模型更倾向预测为再入院。creatinine.enzymatic.method 的高取值主要对应正的 SHAP 值，提示肌酐水平升高会推动模型预测为再入院；systolic.blood.pressure 则呈现相反方向，高收缩压对应负的 SHAP 值、低收缩压对应正的 SHAP 值，表明较低的收缩压水平会增加模型预测为再入院的倾向。其余变量（pulse、albumin、NYHA 心功能分级、ageCat、CCI.score 及 gender）的 SHAP 值整体集中在 0 附近，提示其对模型预测结果的边际影响较小。

需要指出的是，brain.natriuretic.peptide（BNP）在 SHAP 分析中呈现的方向性与一般医学常识中“BNP 升高提示病情加重、预后不良”的认识并不完全一致。在本模型中，较高的 BNP 水平倾向于降低模型预测为 28 天再入院的概率，而较低 BNP 水平反而推动模型更倾向预测为再入院。这一现象可能与本研究的结局定义及研究情境有关。本研究关注的是短期再入院而非死亡或长期不良预后，高 BNP 患者在住院期间可能接受了更充分的强化治疗、延长住院观察或出院后更密切的随访管理，从而在短期内降低了再入院的发生概率。此外，在多变量模型中，BNP 与心功能分级、肾功能等指标存在相关性，其独立信息在模型决策中可能被部分“吸收”或重新分配，导致其在预测短期再入院时呈现出不同于传统预后的方向性。因此，BNP 在本研究中的 SHAP 表现应被理解为模型在特定结局、特定人群及变量组合条件下的决策特征，而不应简单等同于其在传统临床预后评估中的意义。

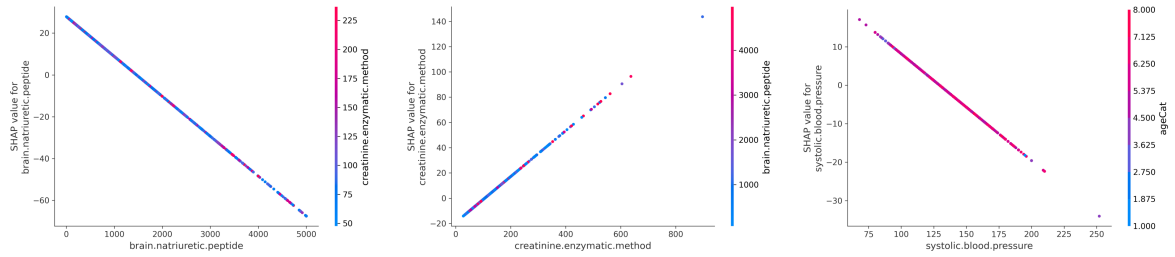

图 7: 主要连续变量的 SHAP 依赖图（Logistic Regression 模型）

SHAP 依赖图进一步展示了单个连续变量取值变化与其 SHAP 值之间的关系，从而刻画变量对模型预测结果的具体作用模式。

在 brain.natriuretic.peptide 的依赖图中，BNP 水平与其 SHAP 值呈现近似线性的负相关关系，即 BNP 值越高，其对应的 SHAP 值越低，表明在该模型中，较高的 BNP 水平持续推动模型预测为“未再入院”，而较低 BNP 水平则推动模型预测为“再入院”。这一关系在不同肌酐水平（颜色所示）下保持一致，未观察到明显的交互分层效应。

在 creatinine.enzymatic.method 的依赖图中，肌酐水平与 SHAP 值呈明显正相关关系，即随着肌酐升高，SHAP 值逐渐增大，提示肾功能受损程度越重，模型越倾向预测为 28 天再入院。不同 BNP 水平下，该正向关系总体保持稳定。

在 systolic.blood.pressure 的依赖图中，收缩压与 SHAP 值呈清晰的负相关关系，即收缩压越低，其 SHAP 值越高，推动模型更倾向预测为再入院；而较高的收缩压则对应负的 SHAP 值，提示其在模型中具有保护性方向。该关系在不同年龄分组（颜色所示）下整体一致。

总体而言，这三张 SHAP 依赖图表明，模型主要通过连续变量的单调关系进行决策：肌酐表现为稳定的正向风险因子，收缩压表现为稳定的负向风险因子，而 BNP 在本研究中呈现出与传统临床认知不同的负向预测方向，反映了其在短期再入院预测情境下的特定模型作用模式。

阶段五：无监督学习的探索性聚类分析

在无监督学习阶段，本研究基于已完成缺失值填补的数据集开展探索性聚类分析，旨在不依赖结局变量的前提下，识别心力衰竭患者是否可基于连续临床指标形成潜在的亚群结构。本阶段分析为纯探索性分析，不用于预测建模、模型性能比较或因果推断。本研究采用 K-means clustering（K 均值聚类）方法进行聚类分析。为保证距离计算的合理性，聚类分析仅使用连续变量，并在建模前对变量进行标准化处理。

聚类数（K）的选择采用肘部法则（elbow method）与轮廓系数（silhouette coefficient）相结合的方式综合确定。首先，通过绘制不同 K 值下的类内平方和变化曲线，观察误差下降趋势以初步识别可能的拐点；随后，结合不同 K 值对应的轮廓系数，对聚类结构的紧密性与分离度进行评估，最终在统计指标与临床可解释性之间取得平衡，确定最优聚类数。

在确定聚类结果后，采用主成分分析（principal component analysis，PCA）对高维连续变量进行降维，并将聚类标签映射至前两个主成分空间中进行可视化展示。PCA 可视化仅用于辅助理解聚类结构及样本分布特征，不参与聚类过程本身。

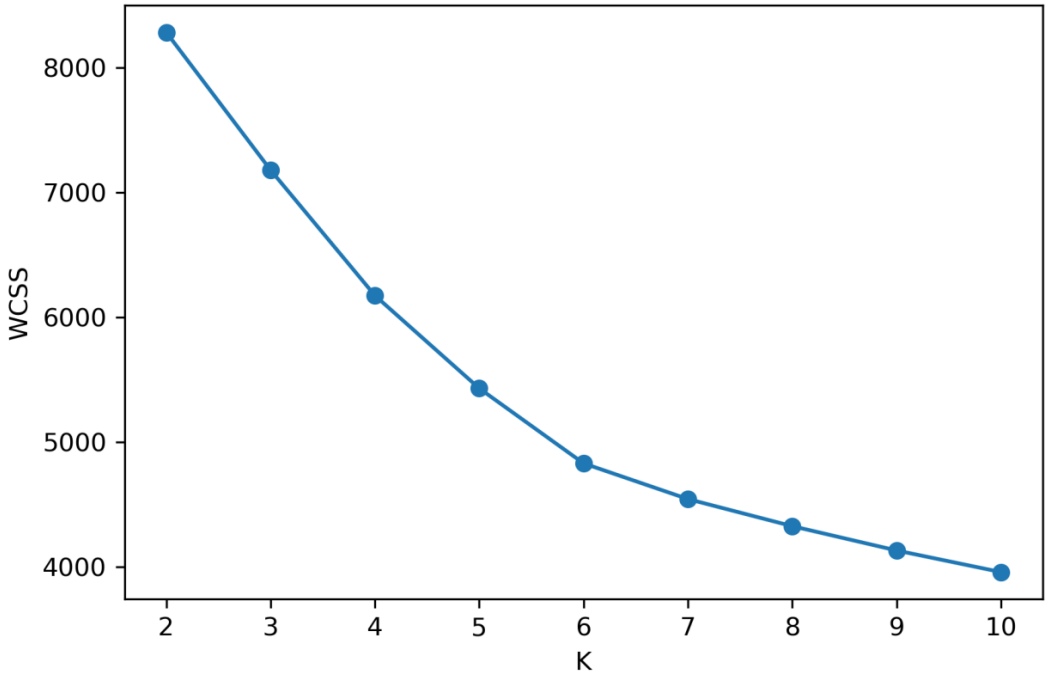

图 8: K-means 聚类的肘部法则图

肘部法则图显示，随着聚类数 K 的增加，类内平方和（within-cluster sum of squares，WCSS）持续下降，但下降幅度逐渐减小。当 K 从 2 增加至 4 时，WCSS 下降较为明显；而当 K 大于 4 后，曲线趋于平缓，进一步增加聚类数所带来的类内紧密度改善有限，提示在该区间内已接近“肘部”。结合轮廓系数结果，在 K=4 时模型仍能保持相对良好的类内一致性与类间分离度。综合考虑肘部法则、轮廓系数以及聚类结果的临床可解释性，本研究最终选择 K=4 作为 K-means 聚类的聚类数，用于后续聚类分析与可视化展示。

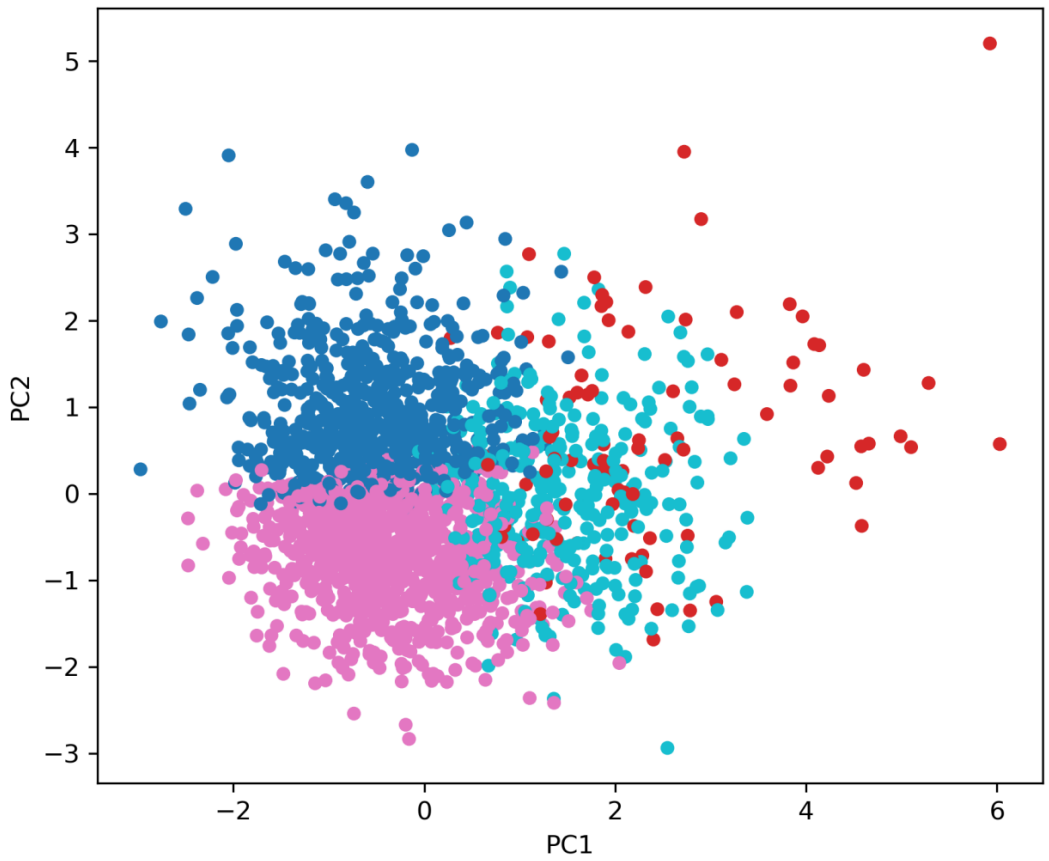

图 9: K-means 聚类结果的 PCA 二维可视化（K = 4）

基于 K-means 聚类（K=4）的结果，采用主成分分析（PCA）将高维连续变量降维至前两个主成分空间进行可视化展示。图中不同颜色的点代表不同的聚类亚群，横轴为第一主成分（PC1），纵轴为第二主成分（PC2）。从 PCA 可视化结果可以看出，四个聚类在二维空间中形成了相对清晰的分布结构，尽管部分区域存在重叠，但整体上仍表现出一定程度的分离趋势。其中，部分聚类主要沿 PC1 方向分布，提示其在主要变异方向上的特征与其他亚群存在差异；另一些聚类则更多体现在 PC2 方向上的区分，反映了次要变异维度上的异质性。需要指出的是，PCA 可视化仅用于辅助理解聚类结构，其二维投影无法完全呈现高维空间中的真实距离关系。尽管如此，该结果仍支持在连续临床指标基础上，研究人群中可能存在具有不同特征模式的潜在亚群，为后续对各聚类特征进行描述性比较提供了直观依据。