从体征到判别：心脏病数据的聚类、因子分析与预测模型对比

课程设计成果汇报

08023110陈奕诚，08023214张韫译萱，08023203周一鸣

2025.12.30

01/项目背景与分析思路

核心问题：

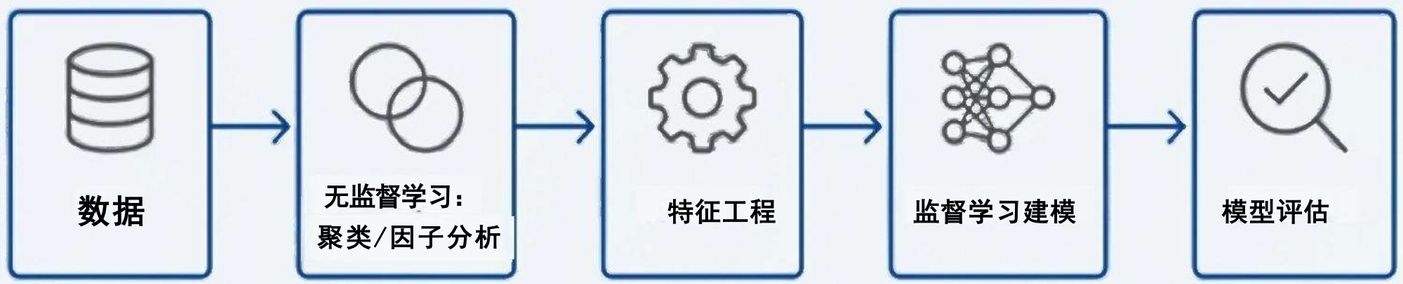
心脏病是全球主要健康威胁，早期筛查与精准预测至关重要。

数据驱动方法：

利用患者体征及检查指标，可通过数据分析构建有效的预测模型。

本项目分析思路：

本项目旨在探索一条完整的分析思路：从未经标记的数据中发掘结构(聚类、因子分析)，到利用这些结构构建并对比多种预测模型(Logistic 回归、贝叶斯判别)。



02 / 数据集与核心变量

目标变量（Target Variable）：
心脏病二分类结果（1=有，0=无）。

特征类型（Feature Types）： 数据集包含连续型与离散型两类特征。

初期焦点（Initial Focus for Tasks 1&2）：为进行聚类与因子分析，我们首先选取5个关键连续变量进行结构探索：

（1）age: 年龄

（2）trestbps: 静息血压

（3）chol: 血清胆固醇

（4）thalach: 最大心率

（5）oldpeak: 运动后ST段压低

年龄	性别	胸痛类型	静息血压	血清胆固醇	静息心电图结果	最大心率	运动诱发心绞痛	运动后ST段低压程度	ST段斜率类型	心脏病
70	1	4	130	322	2	109	0	2.4	2	1
57	1	2	124	261	0	141	0	0.3	1	1
64	1	4	128	263	0	105	1	0.2	2	0
74	0	2	120	269	2	121	1	0.2	1	0
65	1	4	120	177	0	140	0	0.4	1	0
56	1	3	130	256	2	142	1	0.6	2	1
59	1	4	110	239	2	142	1	1.2	2	1
60	1	4	140	293	2	170	0	1.2	2	1
63	0	4	150	407	2	154	0	4	2	1
59	1	4	135	234	0	161	0	0.5	2	0
53	1	4	142	226	2	111	1	0	1	0

03 / 数据预处理：Z-Score 标准化

挑战（The Challenge）：

5个连续变量的量纲和取值范围差异巨大，直接分析会使距离计算和相关性分析被个别指标主导。

解决方案（The Solution）：

采用Z-score标准化，消除量纲影响，使各变量具有可比性。

年龄	静息血压	血清胆固醇	最大心率	运动后ST段低压程度	心脏病
70	130	322	109	2.4	1
57	124	261	141	0.3	1
64	128	263	105	0.2	0
74	120	269	121	0.2	0
65	120	177	140	0.4	0
56	130	256	142	0.6	1
59	110	239	142	1.2	1

年龄	静息血压	血清胆固醇	最大心率	运动后ST段低压程度	心脏病
1.739810686	-0.049724911	1.623097397	-1.743495407	1.171516439	1
0.302919033	-0.393278841	0.278375499	-0.366685601	-0.655931433	1
1.076629923	-0.164242888	0.322464742	-1.915596633	-0.74295276	0
2.181931195	-0.622314795	0.454732469	-1.22719173	-0.74295276	0
1.18716005	-0.622314795	-1.573372687	-0.409710907	-0.568910106	0
0.192388906	-0.049724911	0.168152393	-0.323660294	-0.394867451	1
0.523979287	-1.194904678	-0.206606168	-0.323660294	0.127260512	1

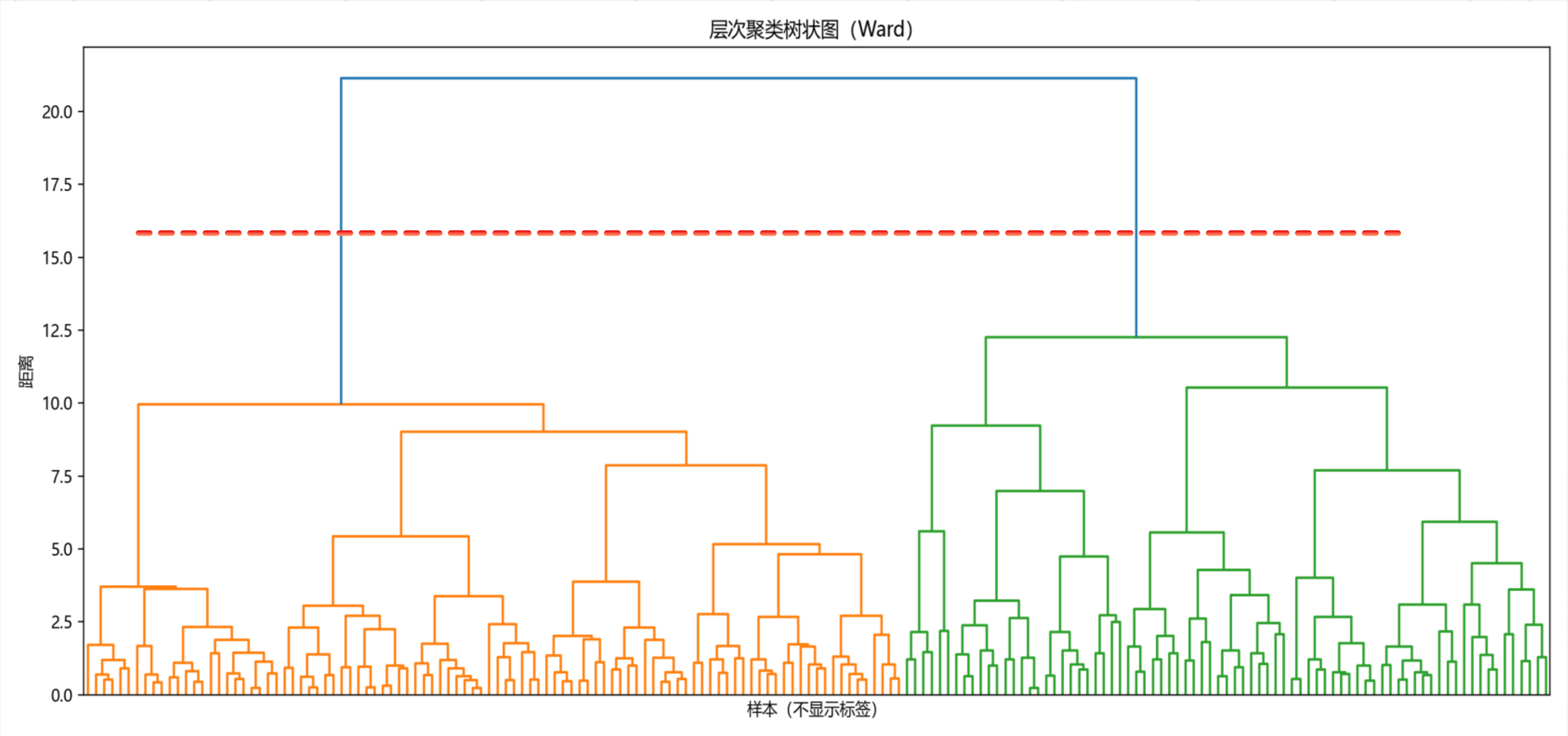
公式：

$$Z = \frac{X - \mu}{\sigma}$$

04/探索(一):层次聚类揭示潜在分组

方法:

采用Ward 法进行层次聚类，该方法旨在最小化合并后的类内方差。从树状图中可见，在较高的距离水平上存在一次明显的合并“跃迁”，表明数据自然地倾向于被划分为两类。



05/探索(二):相关性分析寻找变量间联系

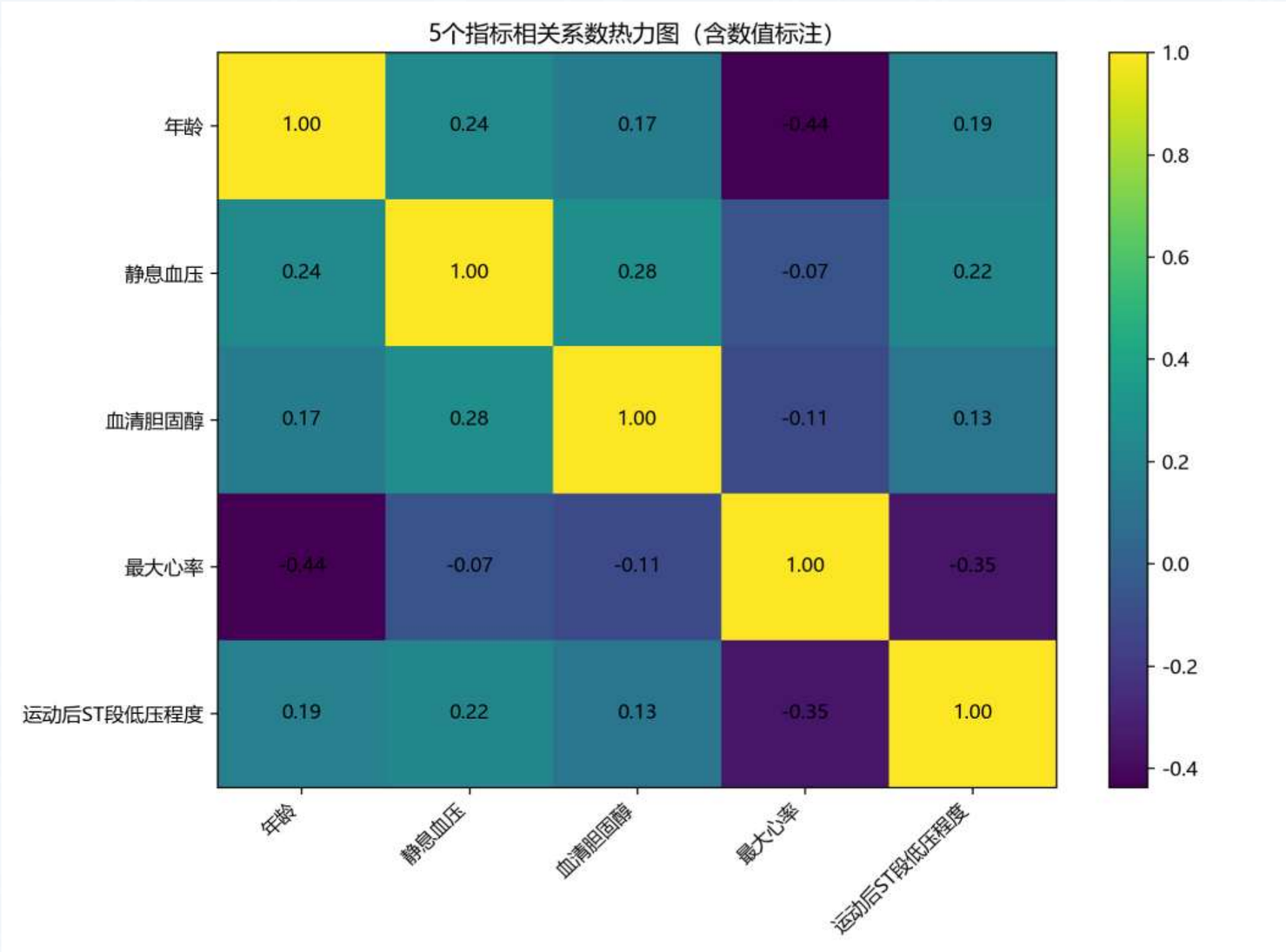
目标:
在因子分析前，检验5个变量间的相关性结构，为降维提供依据。

主要观察:

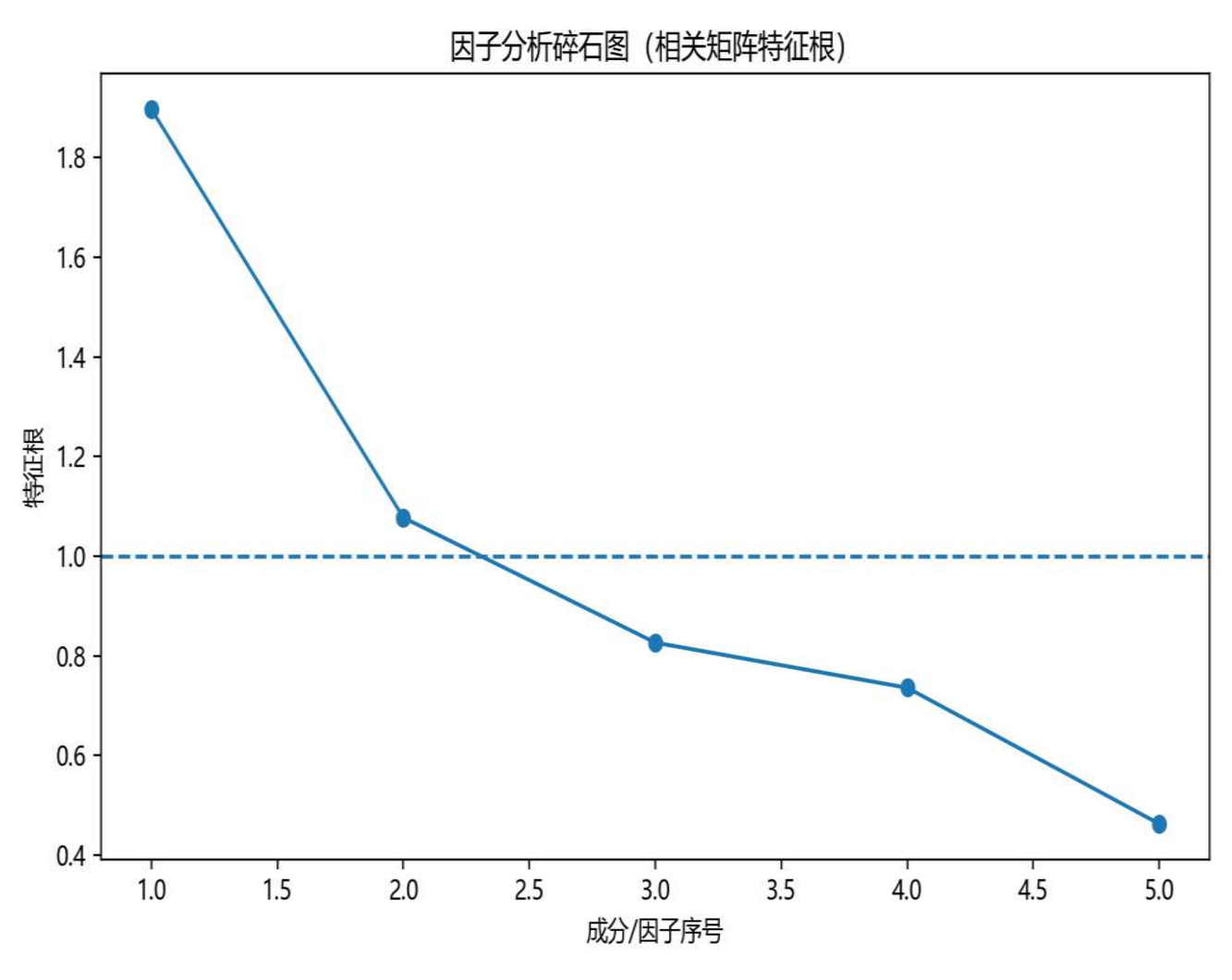
⑴ “年龄”与“最大心率”呈显著负相关。

⑵ “最大心率”与“运动后ST段压低”呈负相关。

（3）“静息血压”与“血清胆固醇”存在一定正相关。



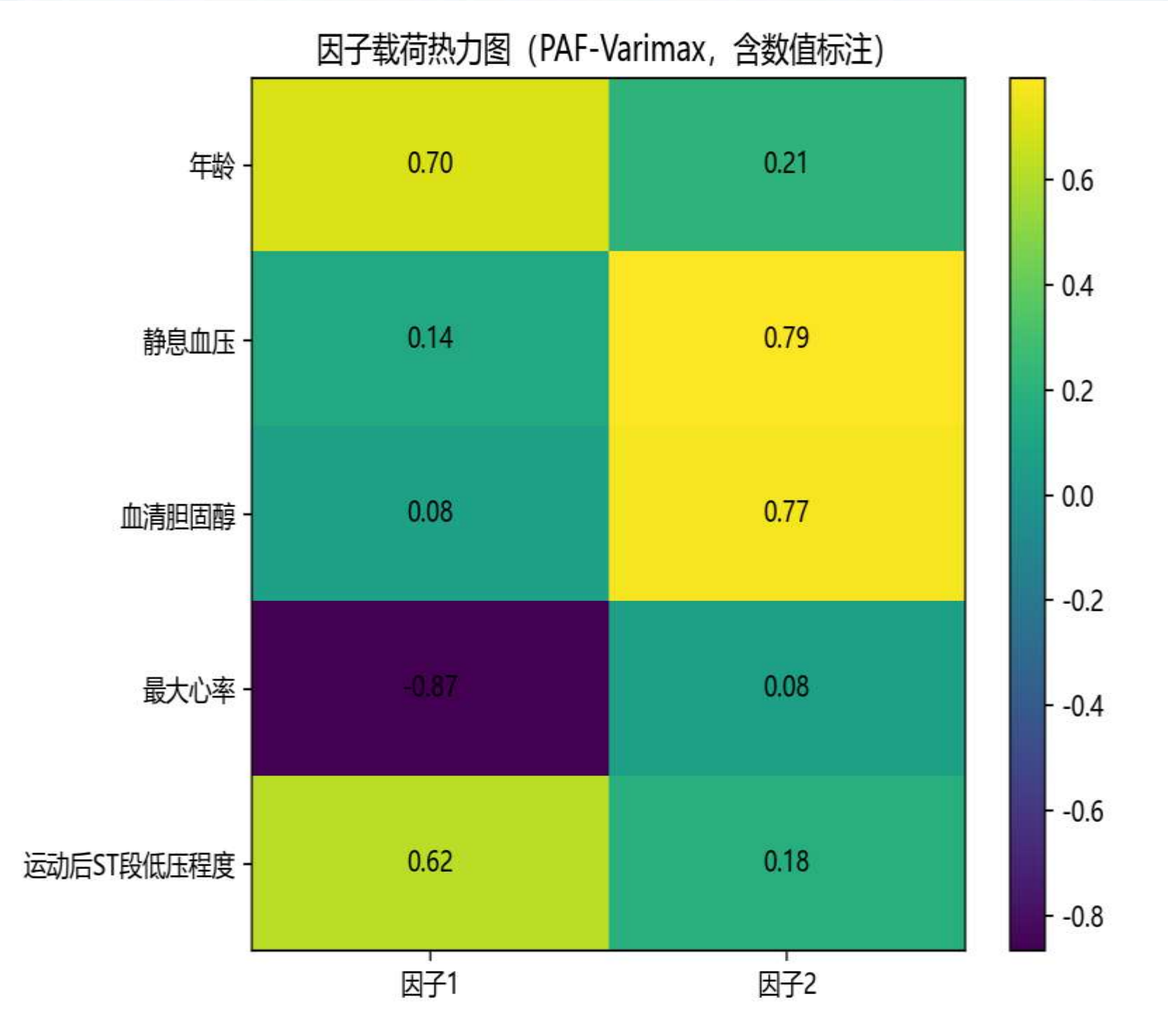
06 / 探索(三)：因子分析确定核心维度



- **目标：**将多个相关的变量浓缩为少数几个“因子”，以简化数据结构。
- **选择标准：**
  - **Kaiser准则：**保留特征根(Eigenvalue) > 1的因子。
  - **碎石图检验：**在图形“陡坡”与“缓坡”的交界处（拐点）确定因子数量。

结论：两种方法均指向保留 **2个因子**。

07/因子解释：从数据到医学含义



方 法 ：

采用主轴因子法(PAF)提取，并进行方差最大化（Varimax）旋转以获得清晰的因子结构。

因子1：心功能与运动反应因子

主要负载于“年龄”(0.70)，“最大心率”(0.87)，“运动后ST段压低”(0.62)。可解释为“年龄越大、运动耐受越差、心肌缺血反应越明显，风险越高”。

因子2：基础心血管风险因子

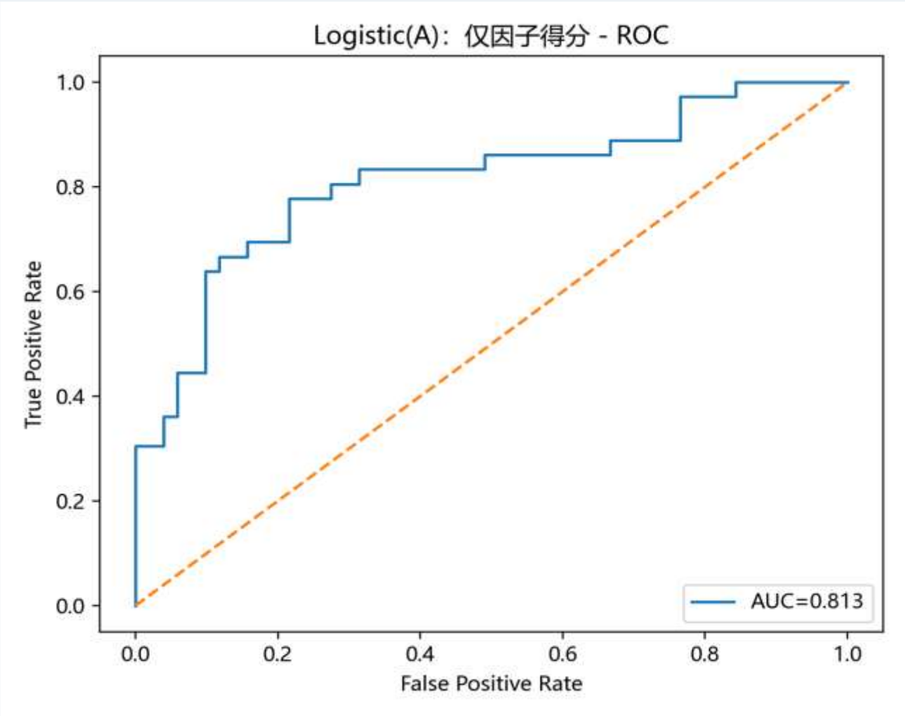
主要负载于“静息血压”(0.79)和“血清胆固醇”(0.77)代表了基础的血压与血脂水平风险。

08/模型(一)：Logistic 回归性能对比

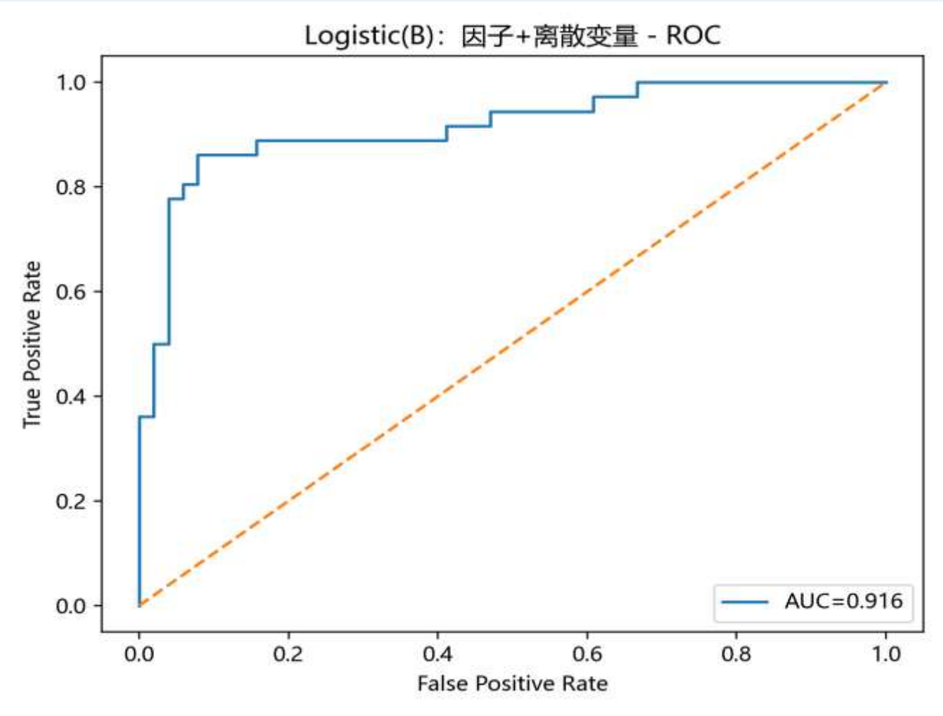
核心问题：仅用2个因子预测(方案A)，和“因子+离散变量”共同预测(方案B)，效果有多大差异?

结果：方案B（因子+离散变量)的**预测性能显著优于**方案A。

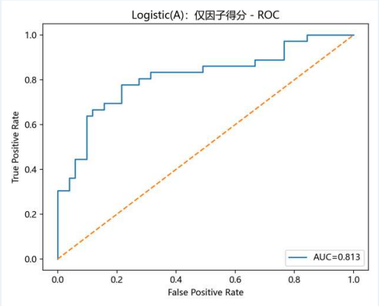
方案A: 仅2个因子

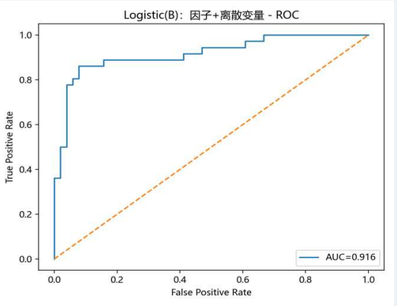


方案B: 因子+离散变量



08/模型(一)： Logistic 回归性能对比



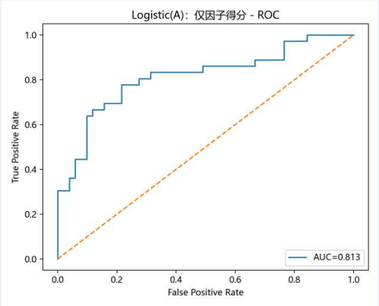


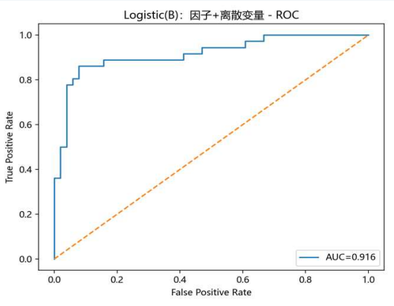
核心问题：仅用2个因子预测(方案A)， 和“因子+离散变量”共同预测(方案B)， 效果有多大差异?

结果：方案B (因子+离散变量)的**预测性能显著优于**方案A。

模型	Test_AUC	Test_ACC	Precision	Recall	F1
Logit_A(仅因子)	0.813	0.770	0.700	0.778	0.737
Logit_B(因子+离散,不含性别)	0.916	0.862	0.816	0.861	0.838

08/模型(一)：Logistic 回归性能对比





核心问题：仅用2个因子预测(方案A)，和“因子+离散变量”共同预测(方案B)，效果有多大差异?

结果：方案B（因子+离散变量)的**预测性能显著优于**方案A。

Logit_B（因子+离散）系数Top（展示前12项）：

变量	系数(β)
血管堵塞程度_0	-1.5465
胸痛类型_4	1.0293
血管堵塞程度_3	1.0110
铊显像结果_7	0.9922
铊显像结果_3	-0.9073
ST段斜率类型_2	0.6636
因子得分2	0.4545
胸痛类型_1	-0.4479
血管堵塞程度_2	0.4133
因子得分1	0.3749
胸痛类型_2	-0.3533
ST段斜率类型_3	-0.3455

Logit_A（仅因子）系数：

变量	系数(β)
因子得分1	0.9946
因子得分2	0.5390

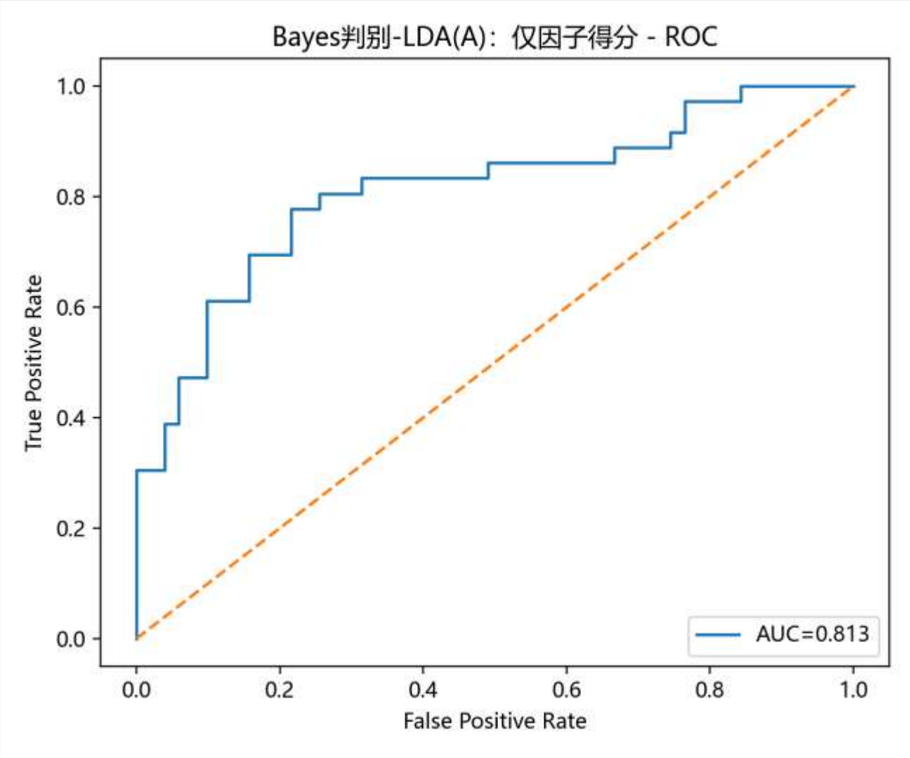
注：系数β的方向含义：β>0 表示该特征增大/该类别出现时，患病概率倾向上升；β<0 表示倾向下降。离散变量一般经过 One-Hot 编码，因此系数表示“相对于基准类别”的差异。

09/模型(二):线性判别分析(LDA) 结果

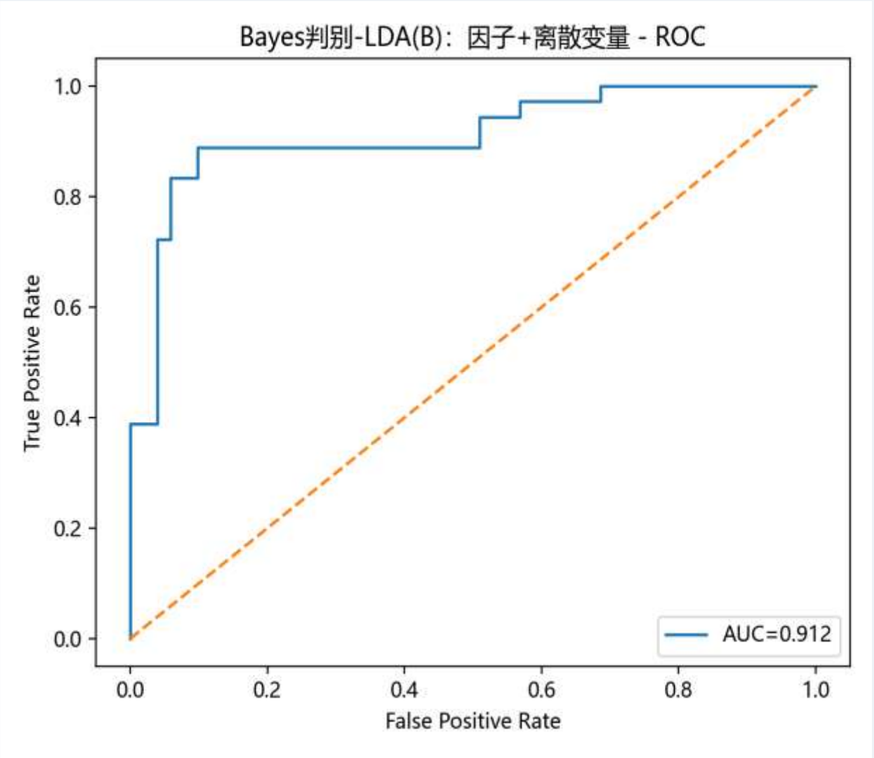
目标：使用另一种经典的线性分类器进行验证，以检验结论的稳健性。

结果：LDA的结果与Logistic 回归高度一致。方案B同样表现出色 (AUC≈0.912), 再次证明了增加离散变量的有效性。

方案A：仅2个因子



方案B：因子+离散变量

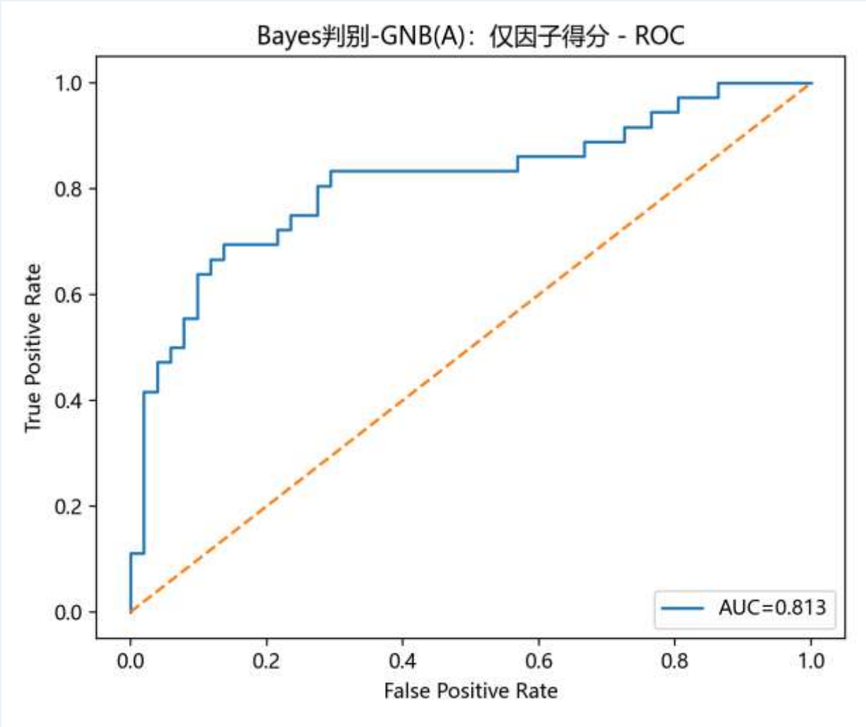


10/模型(三):高斯朴素贝叶斯(GNB)结果

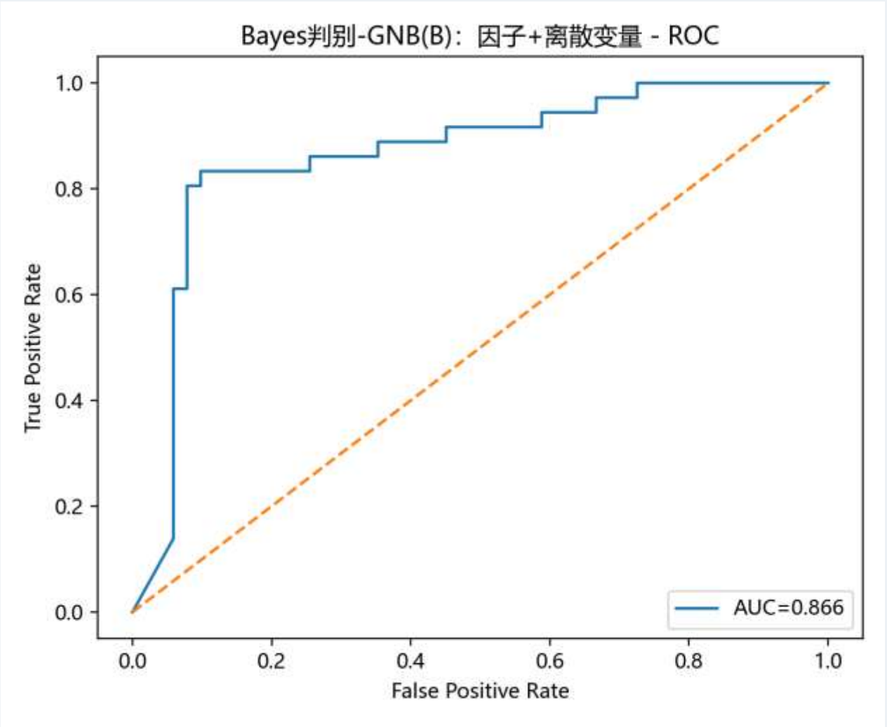
目标：测试基于概率和特征独立性假设的模型。

结果：方 案B 依然优于方案A，但整体性能（AUC≈0.866） 略低于Logistic和LDA。 这可能与其较强的“特征独立性”假设有关。

方案A：仅2个因子

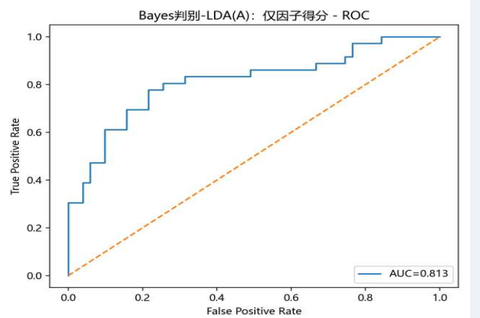


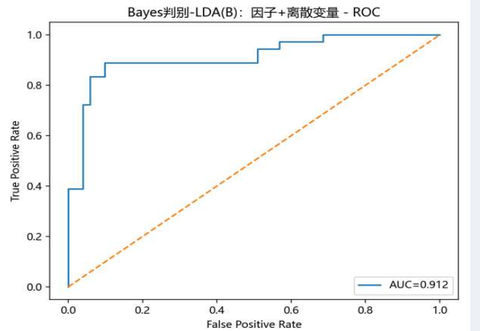
方案B：因子+离散变量

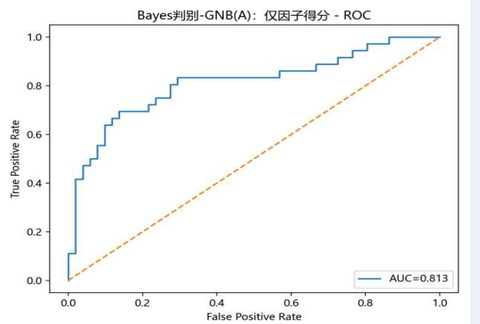


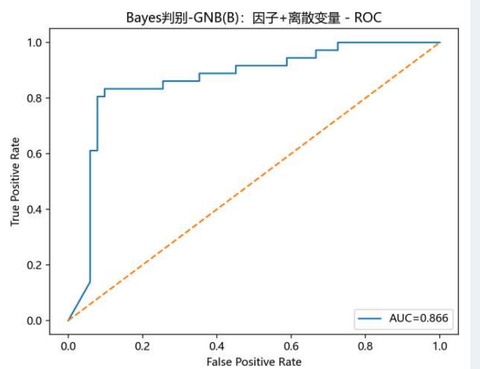
11/Bayes判别（LDA / GNB，方案A vs 方案B）

模型	Test_AUC	Test_ACC	Precision	Recall	F1
LDA_A(仅因子,按比例先验)	0.813	0.770	0.700	0.778	0.737
LDA_B(因子+离散,按比例先验)	0.912	0.885	0.842	0.889	0.865
GNB_A(仅因子,按比例先验)	0.813	0.736	0.659	0.750	0.701
GNB_B(因子+离散,按比例先验)	0.866	0.782	0.870	0.556	0.678









12/综合对比与核心结论

总体发现：在所有模型中，方案B(因子+离散变量)的性能均一致且显著地优于仅使用因子的方案A。

最终结论：Logistic回归模型在方案B下的表现最佳（AUC=0.916），是本次分析的最优预测模型。

模型	方案A:仅因子(AUC)	方案B:因子+离散(AUC)	性能提升(B-A)
Logistic回归	0.813	0.916	+0.103
LDA	0.813	0.912	+0.099
高斯朴素贝叶斯	0.813	0.866	+0.053

13/一些结论

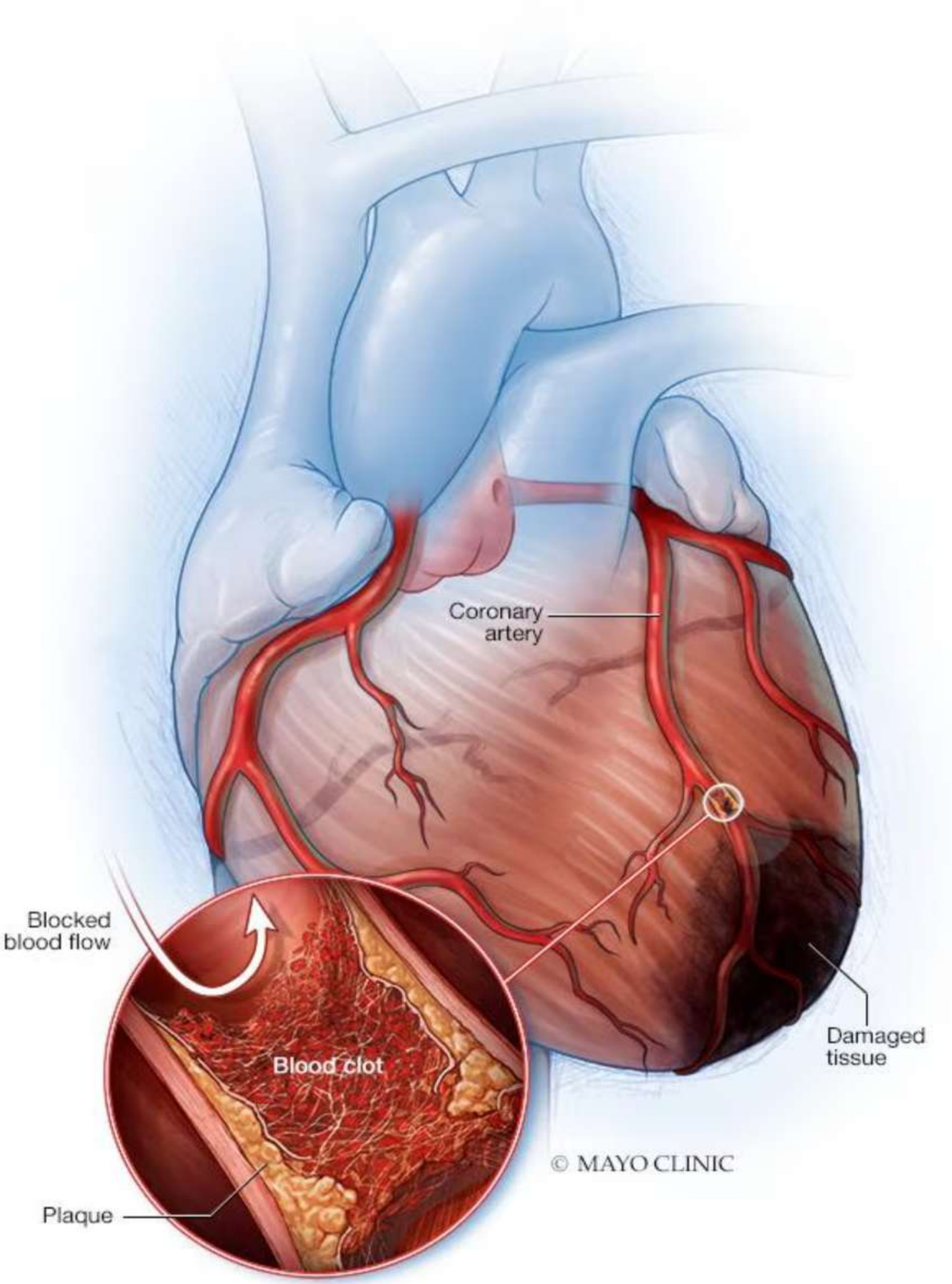
①年龄偏大的人运动时最大心率较低、运动后心电图 ST 段压低较明显，说明心脏运动耐受能力下降，患病风险更高

②静息血压偏高的人群，心脏病风险更大
在其他条件相近的情况下，静息状态下血压水平越高，心脏长期承受的负荷越大，从而：静息血压偏高往往伴随着更高的心脏病发生风险。

③运动诱发不适症状，提示心脏在负荷下存在隐患
如果个体在运动过程中出现心绞痛或明显不适，通常说明心脏在负荷条件下难以维持正常功能：运动诱发症状的出现，往往与更高的心脏病风险相关。

④多项轻度异常叠加，比单一严重异常更值得警惕
即使单个指标未达到明显异常水平，但多项指标同时处于不利区间时，整体心脏病风险仍会显著上升。这说明心脏病风险更多体现为多因素累积效应，而不是某一个指标“突然失控”。

⑤单一指标作用有限，多项指标联合判断更可靠
无论是连续生理指标，还是离散临床指标，单独使用某一个指标都难以准确判断是否患病。但当多类指标联合考虑时，判断效果显著提升。



14/遇到的问题与解决方案

问题1:如何处理混合类型数据进行探索性分析?

解决方案：采用分步策略。先仅对5个连续变量进行聚类和因子分析以挖掘核心结构；再将降维 后的因子与原始离散变量结合，构建更全面的预 测测模型(方案B)。

问题2:如何解决不同变量间的量纲差异?

对连续指标做 Z-score 标准化，保证聚类距离和因子分析的相关结构不会被量纲“绑架”。

参考文献：

[1] Dua D, Graff C. UCI Machine Learning Repository: Heart Disease Data Set[EB/OL]. Irvine, CA: University of California, School of Information and Computer Sciences, 2019. DOI:10.24432/C52P4X.

[2] Detrano R, Janosi A, Steinbrunn W, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease[J]. The American Journal of Cardiology, 1989, 64(5): 304-310. DOI:10.1016/0002-9149(89)90524-9.
PubMed

[3] Ward J H. Hierarchical grouping to optimize an objective function[J]. Journal of the American Statistical Association, 1963, 58(301): 236-244. DOI:10.1080/01621459.1963.10500845.

[4] Kaiser H F. The application of electronic computers to factor analysis[J]. Educational and Psychological Measurement, 1960, 20(1): 141-151. DOI:10.1177/001316446002000116.

[5] Cattell R B. The scree test for the number of factors[J]. Multivariate Behavioral Research, 1966, 1(2): 245-276. DOI:10.1207/s15327906mbr0102_10.

[6] Kaiser H F. The varimax criterion for analytic rotation in factor analysis[J]. Psychometrika, 1958, 23(3): 187-200. DOI:10.1007/BF02289233.

[7] Fisher R A. The use of multiple measurements in taxonomic problems[J]. Annals of Eugenics, 1936, 7(2): 179-188. DOI:10.1111/j.1469-1809.1936.tb02137.x.

[8] Hosmer D W, Lemeshow S, Sturdivant R X. Applied Logistic Regression[M]. 3rd ed. Hoboken, NJ: Wiley, 2013. DOI:10.1002/9781118548387.

感谢倾听！请批评指正！

08023110陈奕诚，08023214张韫译萱，08023203周一鸣

2025.12.30