数据统计分析
课程报告

心脏病数据的聚类、因子分析与预测模型对比

项目成员：陈奕诚、周一鸣、张韫译萱

专业：自动化

指导老师：薛磊

2026 年 1 月 3 日

目录

1 选题背景与数据预处理

1.1 选题背景

心脏病是全球的主要健康威胁之一，精准预测、早预防、早筛查是减弱这一威胁的重要手段。本次分析旨在通过无监督学习手段对数据集中的各特征进行聚类和因子分析，探索数据的内在结构，并利用分析得到的结构建立多种预测模型，对患者患病与否进行预测。

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

图 1: 心脏病预测数据集

1.2 数据预处理

在获取数据集后，我们首先进行了数据预处理。为了更好地进行聚类和因子分析，我们从数据集中选取了五个最为关键的连续变量进行分析，包括：年龄、静息血压、血清胆固醇、最大心率以及运动后 ST 段压低程度。

由于不同医学指标间的量纲不同，取值范围差异巨大（例如胆固醇数值可达几百，而 ST 段压低程度仅为个位数），若直接分析，会导致数值较大的特征主导分析结果。因此，我们首先对数据进行了标准化处理，即对每个特征减去样本均值并除以样本方差（Z-score 标准化），从而消除量纲影响，使得各指标在分析中具有同等的权重。

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

图 2: 使用 Z-score 法对数据进行标准化

2 数据分析

2.1 聚类分析

在完成标准化后，我们对选取的五个连续变量（年龄、静息血压、血清胆固醇、最大心率、运动后 ST 段压低程度）使用 Ward 最小方差法进行聚类分析。Ward 法的目标是使簇内方差最小化，能够生成结构紧凑的聚类结果。

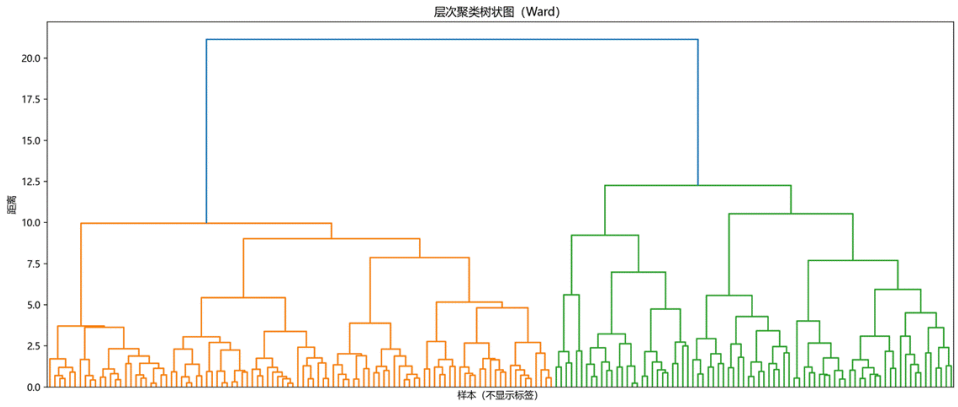

图 3: 层次聚类树状图

从生成的层次聚类树状图可以观察到，当合并距离较小时，样本主要在各自的小分支内逐步合并；而当合并距离上升到较高水平时，出现了明显的“距离突增”现象。这说明在该位置继续合并将显著降低类内相似性。我们发现样本在合并为 2 类时结构较为稳定，而在继续向上合并时，类间差异迅速减小。因此，在聚类类别数的选择上，我们最终将所有变量划分为两类。

聚类结果显示：第一类人群通常表现为：最大心率较低、运动后 ST 段压低程度较高、运动诱发心绞痛比例较高。这反映了心脏功能相对较弱或存在运动耐受性问题。第二类人群表现为：静息血压和血清胆固醇较高，提示存在较高的基础心血管风险。

2.2 因子分析

2.2.1 相关性检验

在进行因子分析之前，我们需要判断原始变量之间是否存在足够的相关性，以确定是否适合进行因子分析。我们计算了变量间的相关系数矩阵。

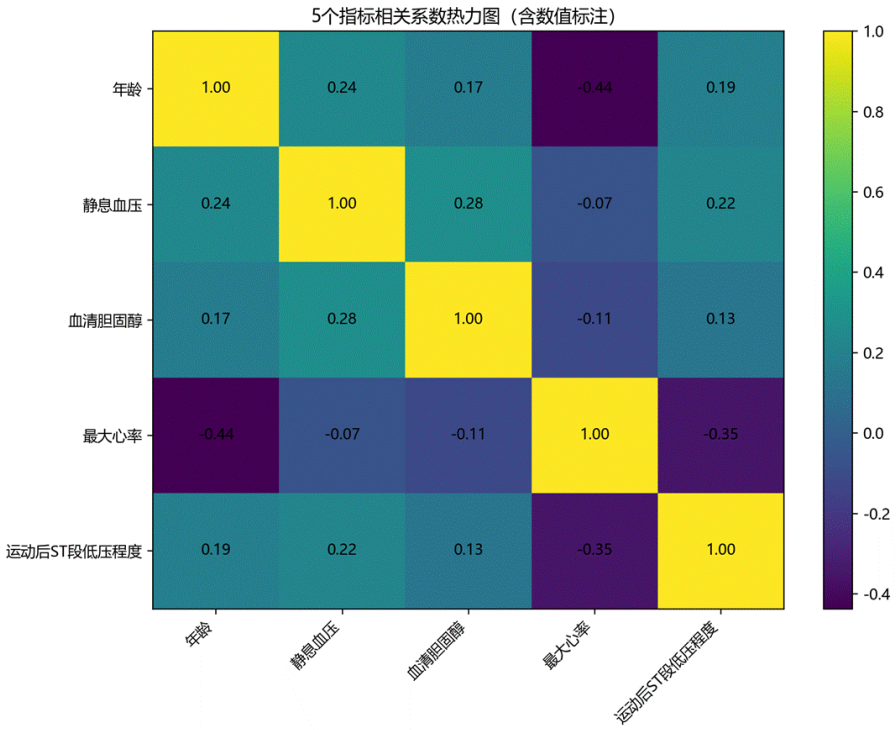

图 4: 变量相关性热力图

如图 2 所示，部分变量之间存在中等程度的相关性，例如“年龄”与“最大心率”呈负相关（-0.44），“最大心率”与“运动后 ST 段压低程度”呈负相关（-0.35）。虽然多数变量间的相关性较弱（这是因为医学指标受个体体质影响较大，而体质因素无法在数据中得到体现），但部分显著的相关性表明数据中存在潜在的公共因子结构，适合进行因子分析。

2.2.2 因子提取与旋转

我们采用主轴因子法（Principal Axis Factoring, PAF）提取公共因子，其目的是解析变量间的共同方差。为了提高因子载荷矩阵的可解释性，我们对提取的初始因子进行了方差最大化正交旋转（Varimax），使得每个变量在特定因子上的载荷尽量高，而在其他因子上的载荷尽量低。

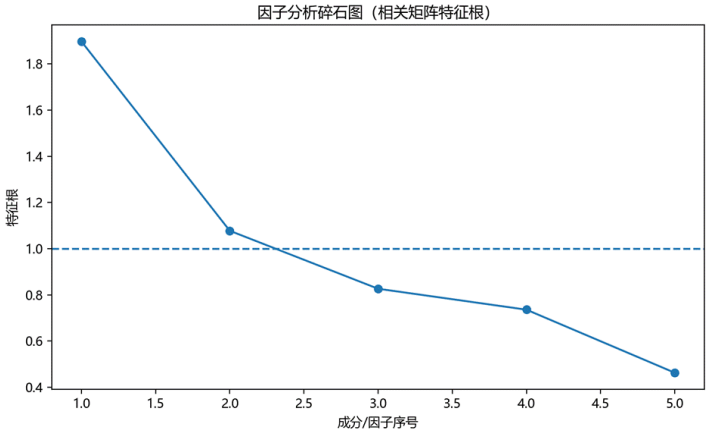

图 5: 碎石图

根据碎石图（图 3）观察，前两个因子的特征根均大于 1（Kaiser 判据），且从第三个因子开始，特征根曲线趋于平缓（“拐点”出现）。因此，我们最终决定保留 2 个公共因子。

2.2.3 因子解释

旋转后的因子载荷矩阵结构清晰：

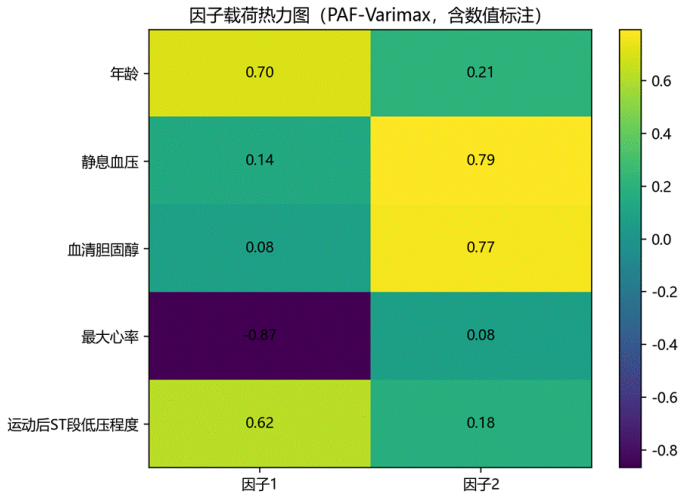

图 6: 旋转后的因子载荷矩阵

- **因子 1**：在“年龄”、“最大心率”（负向）以及“运动后 ST 段压低程度”上具有较高载荷。该因子主要反映个体在运动负荷下的心脏功能表现及年龄相关的生理衰退，我们将其概括为**“心功能/运动反应因子”**。
- **因子 2**：在“静息血压”和“血清胆固醇”上具有较高载荷。该因子更多体现个体的基础代谢和心血管健康状况，我们将其概括为**“基础心血管风险因子”**。

2.3 预测模型构建与对比

为了验证提取的因子对心脏病的预测能力，并探讨是否需要结合其他指标，我们设计了两套输入特征方案进行对比：

- **方案 A**：仅使用提取出的 2 个公共因子得分作为特征。
- **方案 B**：使用 2 个公共因子得分 + 其他离散变量（如性别、胸痛类型、空腹血糖等）作为特征。

我们分别使用逻辑回归（Logistic Regression）、线性判别分析（LDA）和高斯朴素贝叶斯（GNB）三种算法训练二分类模型，并使用 ROC 曲线下的面积（AUC）作为主要评价指标。

ROC 曲线与 AUC 指标说明：

- **ROC 曲线（Receiver Operating Characteristic Curve）**：ROC 曲线是反映二分类模型在不同阈值下分类性能的综合指标。其横坐标为假正率（False Positive Rate, FPR），纵坐标为真正率（True Positive Rate, TPR）。曲线越靠近左上角，说明模型的分类性能越好。
- **AUC（Area Under Curve）**：AUC 即 ROC 曲线下的面积，取值范围通常在 0.5 到 1 之间。AUC 值越大，表示模型区分正负样本的能力越强。一般认为，AUC 在 0.7-0.8 之间表示模型性能尚可，0.8-0.9 表示性能良好，0.9 以上表示性能优秀。

2.3.1 Logistic 回归模型

模型	AUC	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)	0.916	0.862	0.816	0.861	0.838

表 1: Logistic 回归模型性能对比

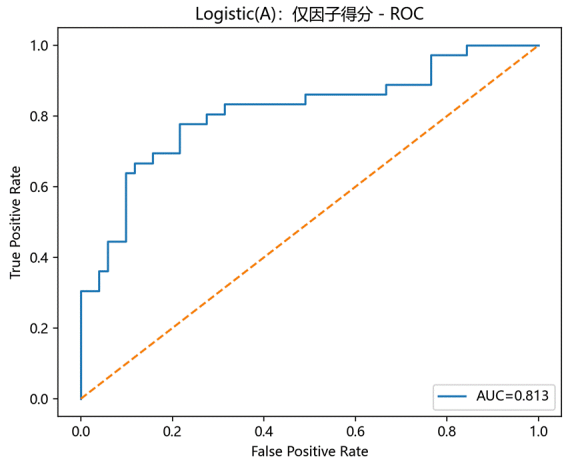

图 7: Logistic(A) ROC 曲线

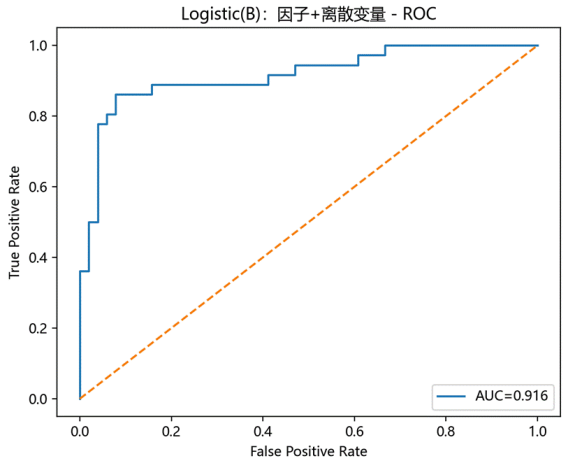

图 8: Logistic(B) ROC 曲线

结果显示，仅靠 2 个因子已有一定的区分能力（AUC>0.8），但加入离散变量后模型性能显著提升（AUC>0.9），说明其他的临床医学指标提供了重要的补充信息。

分析 Logistic 回归的系数发现：

- 因子 1（心功能因子）系数显著为正（0.9946），说明心功能越差（年龄大、最大心率低、ST 压低大），患病风险越高。
- 离散变量中，无症状胸痛、运动诱发心绞痛等变量系数较高，是显著的风险因素。

2.3.2 LDA 与 GNB 模型

我们同样对比了 LDA 和 GNB 在两种方案下的表现：

模型	AUC	ACC	Precision	Recall	F1
LDA_A (仅因子)	0.813	0.770	0.700	0.778	0.737
LDA_B (因子 + 离散)	0.912	0.885	0.842	0.889	0.865
GNB_A (仅因子)	0.813	0.736	0.659	0.750	0.701
GNB_B (因子 + 离散)	0.866	0.782	0.870	0.556	0.678

表 2: LDA 与 GNB 模型性能对比

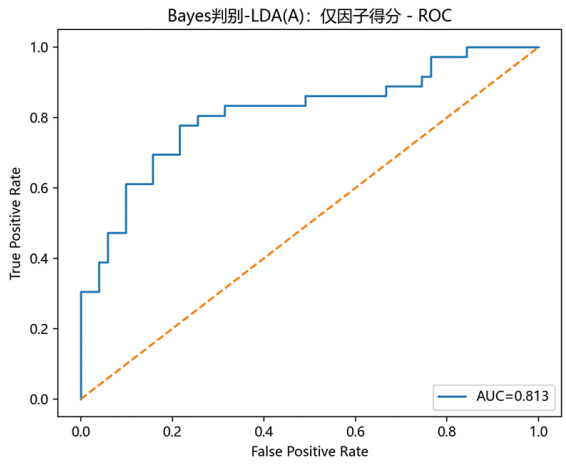

图 9: LDA(A) ROC 曲线

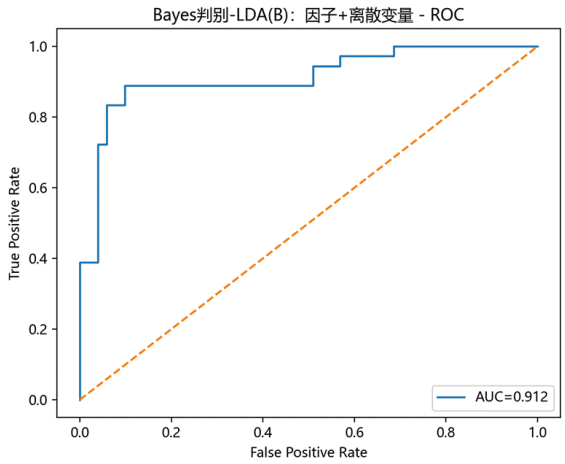

图 10: LDA(B) ROC 曲线

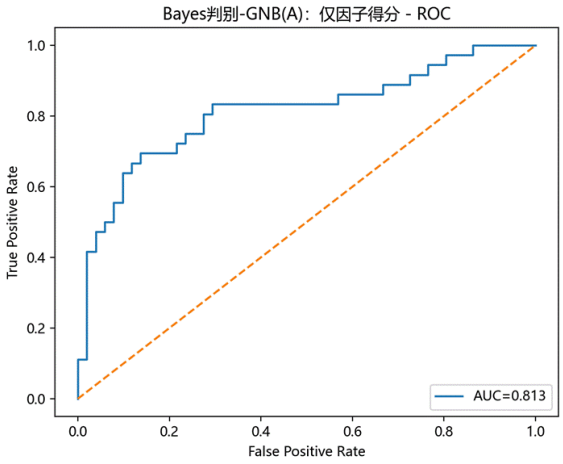

图 11: GNB(A) ROC 曲线

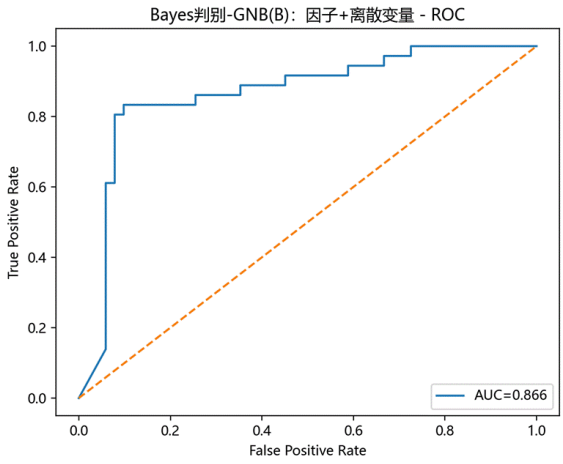

图 12: GNB(B) ROC 曲线

3 所得结论

通过本次对心脏病数据的深入分析，我们得出以下结论：

1. **模型性能对比：**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。其中，**Logistic 回归模型在方案 B 下的表现最佳（AUC=0.916）**，被确定为本次分析的最优预测模型。这表明心脏病的预测需要综合考虑连续的生理指标和离散的临床症状。

2. **医学相关结论：**

   (a) 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

   (b) **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。

   (c) **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。

   (d) **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”。

   (e) **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。

4 遇到的困难和解决方法

在本次实验过程中，我们主要遇到了以下困难并采取了相应的解决方法：

1. **变量选择的困境：**在进行因子分析初期，我们面临是否将所有变量（包含离散变量）都纳入分析的问题。**解决方法：**考虑到离散变量（如性别、胸痛类型）不符合因子分析对连续高斯分布的假设，我们最终决定仅选取 5 个关键的连续医学指标进行聚类和因子分析，保证了数学模型的严谨性和结果的可解释性。

2. **数据量纲差异：**原始数据中，胆固醇（200+）与 ST 段压低（0-6）等变量数值差异极大，直接聚类会导致结果被大数值变量主导。**解决方法：**我们在分析前对所有连续变量进行了 Z-score 标准化处理，消除了量纲影响，使得各指标能平等地参与到距离计算和方差解释中。

3. **模型解释性的提升：**初步提取的因子含义模糊，难以对应具体的医学概念。**解决方法：**我们引入了方差最大化旋转（Varimax），成功将因子分离为清晰的“运动反应”和“基础风险”两个维度，极大地提升了分析结果的医学价值。

5 参考文献及资料

5.1 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.

5.2 其他参考资料

1. 课程教材《数据分析》，范金城，梅长林，科学出版社

2. 课程 PPT