心脏病预测：数据分析与建模报告

田宗源 2511110310

2026 年 1 月 2 日

1 数据读取

数据来自 `work.xls`。标签列为“结果”，其中 Healthy 编码为 0，Sick 编码为 1。

2 数据预处理

- 删除说明行（编号 =No）
- 数值变量：缺失值用中位数填补，并标准化
- 分类变量：缺失值用众数填补，并做 one-hot 编码

3 统计分析（显著性检验）

数值变量采用 Mann–Whitney U 检验；分类变量采用卡方检验（必要时 Fisher）。

3.1 数值变量检验结果

表 1: 数值变量在 Healthy 与 Sick 两组的差异（Mann–Whitney U）

变量	Healthy$_median$	Sick$_median$	p$_value$
最大心率	161	142	9.8e-14
峰值	0.2	1.4	2.41e-13
年龄	52	58	3.44e-05
静息血压	130	130	0.0347
血清胆固醇	234	249	0.0357

3.2 分类变量检验结果

表 2: 分类变量在 Healthy 与 Sick 两组的差异（卡方 / Fisher）

变量	方法	类别数	$\min_e xpected$	$\mathrm{p}_v alue$
缺损	chi-square	3	8.2	1.35e-19
胸痛类型	chi-square	4	10.5	1.33e-17
运动引起绞痛	chi-square	2	45.1	7.45e-14
斜度	chi-square	3	9.56	4.83e-11
性别	chi-square	2	43.7	1.88e-06
心电图	chi-square	3	1.82	0.00666
空腹血糖	chi-square	2	20.5	0.744

与 Healthy 组相比，Sick 组在年龄、最大心率、峰值等数值变量上存在显著差异（$p<0.05$）；在缺损、胸痛类型、运动引起绞痛等分类变量上也存在显著差异（$p<0.05$），空腹血糖差异不显著（$p \geq 0.05$）。

4 预测模型建立

使用 Logistic Regression 模型，采用训练集训练、测试集评估，并输出 ROC-AUC、分类报告。

5 模型评估与可视化

5.1 混淆矩阵

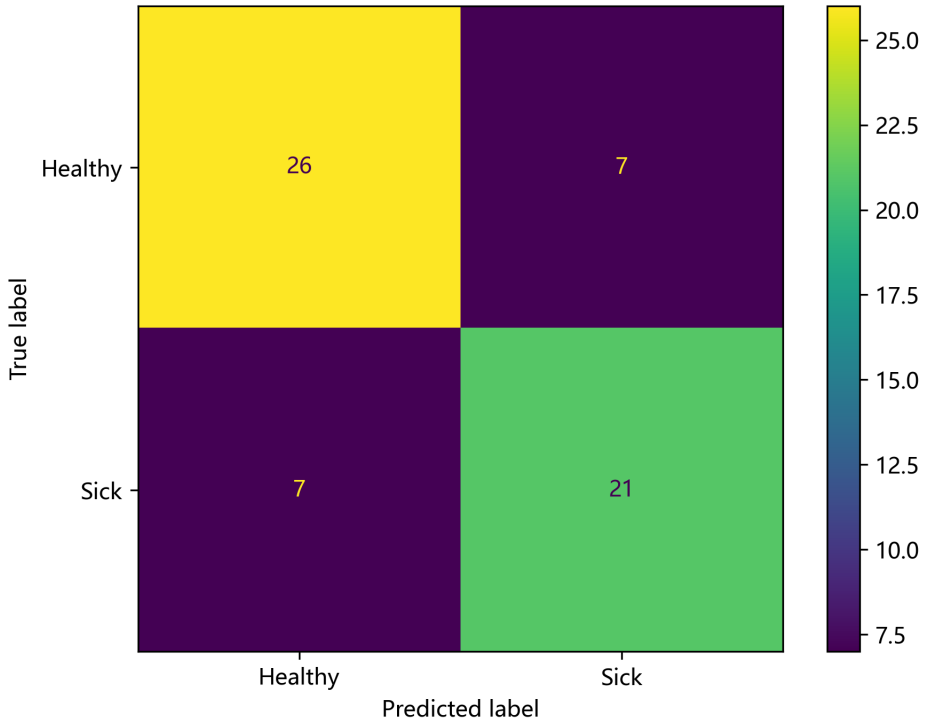

图 1: 混淆矩阵

5.2 ROC 曲线

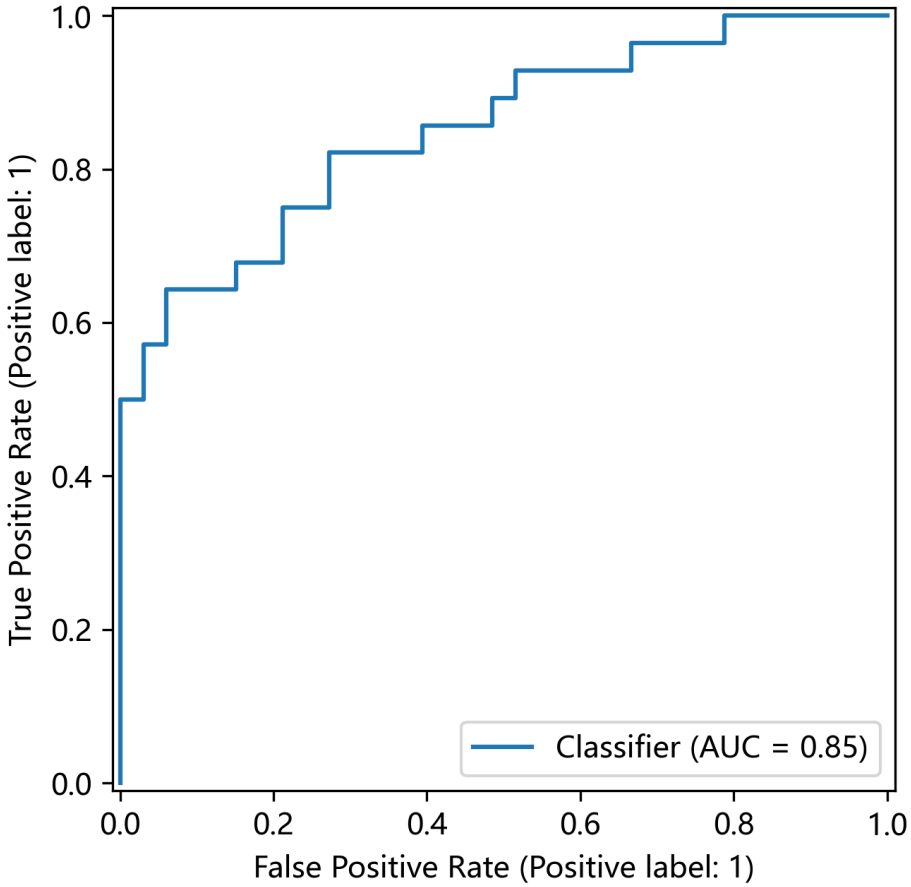

图 2: ROC 曲线

5.3 Precision-Recall 曲线

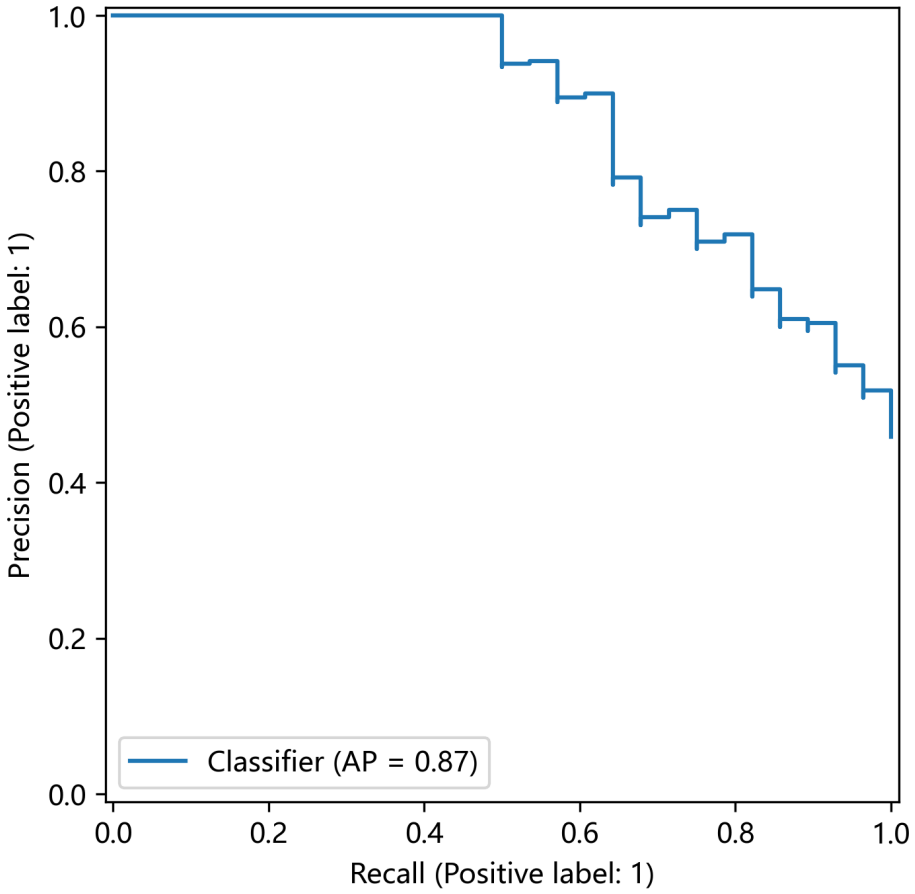

图 3: Precision-Recall 曲线