浙江工商大學

《机器学习》课程论文

题目：基于逻辑回归的心脏病发作预测

学　　院：统计与数学学院

专　　业：应用统计

学　　号：**22020040148**

学生姓名：王锐

二〇二三 年 四 月

基于逻辑回归的心脏病发作预测

摘要

逻辑回归是机器学习中的一种分类算法，广泛应用于处理二分类问题，例如探索某疾病的危险因素，根据危险因素预测某疾病发生的概率等。本文通过 303 位心脏病患者的数据，使用逻辑回归算法对病人心脏病发作概率进行预测分类，最终预测精确度为 78.02%。本文主要简述逻辑回归的基本原理和一般步骤，同时对比了决策树和随机森林算法，分析比较三种算法之间的优劣差异。

关键词：逻辑回归；决策树；随机森林；预测分类

目录

一、 逻辑回归算法简述

1.1 基本概念

逻辑回归是一种广义的线性回归分析模型，属于机器学习中的监督学习，多用于处理二分类问题，从本质上看逻辑回归属于一种分类算法。逻辑回归的过程通常是面对一个回归或者分类问题，建立代价函数，通过优化方法迭代求解出最优的模型参数，然后测试验证我们这个求解的模型的好坏。

逻辑回归与多重线性回归较为相似，都可以归于广义线性模型。广义线性模型之间最大的差别在于因变量不同，将因变量分为连续的、二项分布、Poisson 分布、负二项分布，则分别对应多重线性回归、逻辑回归、Poisson 回归、负二项回归。

1.2 逻辑回归的优缺点

逻辑回归算法实现简单，应用广泛于二分类问题，并且在分类时计算量小，速度快，存储资源低；但是当特征空间很大时，逻辑回归的性能不是很好，容易出现欠拟合问题。

1.3 逻辑回归的一般问题

逻辑回归问题的一般步骤包括寻找假设函数、构造损失函数、求解损失函数的最小值并得出回归参数。求解最小值和回归参数时常用的方法是梯度下降法，对于逻辑回归的损失函数构成的模型，可能会由于权重的原因，导致过拟合问题，使得模型的复杂度提高，泛化能力较差。针对过拟合问题的常用解决方法是减少特征数量和正则化。

二、 其他算法

2.1 决策树算法

决策树是一种以树结构（包括二叉树和多叉树）形式来表达的预测分析模型，属于有监督学习算法。决策树由节点和分支组成，节点有根节点、内部节点、叶

节点三种类型。一般一棵决策树包含一个根节点，若干个内部节点和若干个叶节点，分支起到连接各个节点的作用。

决策树的构建是一个自根至叶的递归过程，需要在每个中间结点寻找一个“划分”属性。决策树算法的核心就在于如何选择最优划分属性。根据选取的划分属性不同，基于决策树衍生出一系列算法，例如以信息增益为准则来选择划分属性的 ID3 决策树学习算法、以增益率为准则来选择最优划分属性的 C4.5 决策树算法、以基尼指数来选择划分属性的 CART 决策树算法等。

2.2 随机森林算法

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树。它通过自助法（bootstrap）重采样技术，从原始训练样本集 N 中有放回地重复随机抽取 n 个样本生成新的训练样本集合训练决策树，然后生成 m 棵决策树组成随机森林。其实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于独立抽取的样本。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样本可以通过每一棵树的分类结果经统计后选择最可能的分类。

每棵树都选择部分样本及部分特征，使得随机森林算法能在一定程度避免过拟合问题，具有很好的抗噪能力，性能稳定。并且随机森林能处理很高维度的数据，不用做特征选择，但是随机森林对于小数据或者低维数据（特征较少的数据），可能无法产生很好的分类。

三、 模拟实验

3.1 描述性统计

数据集大小为 304 行，14 列。其中 14 个列变量分别为病人年龄、病人性别、胸部疼痛类型（0-无症状，1-非心绞痛，2-非典型心绞痛，3-典型心绞痛）、静息心率、胆固醇指数、空腹血糖是否超过 120mg/dl、静息心电图结果（0-正常，1-波正常，2-左心室肥厚）、达到的最大心率、有无运动型心绞痛、相对于休息来说运动引起的 ST 段抑制、运动高峰的心电图(1=上坡，2=平坦，3=下坡)、萤光显色的主要血管数目、铊压力测试结果、实际结果。

图 1 数据集前五行和后五行展示

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

数据集中病人平均年龄为 54.37 岁，最年轻的病人仅 27 岁，最年长的病人 77 岁，其中男性病人数量多于女性。超半数的病人患有心绞痛，并且超过四分之三的病人存在静息心率（超过 100 次/分钟）和胆固醇值过高（超过 120mg/dl）的问题。

图 2 描述性统计

	age	sex	cp	trtbps	chol	fbs	restecg
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

图 3 描述性统计

thalachh	exng	oldpeak	slp	caa	thall	output
303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

3.2 数据预处理

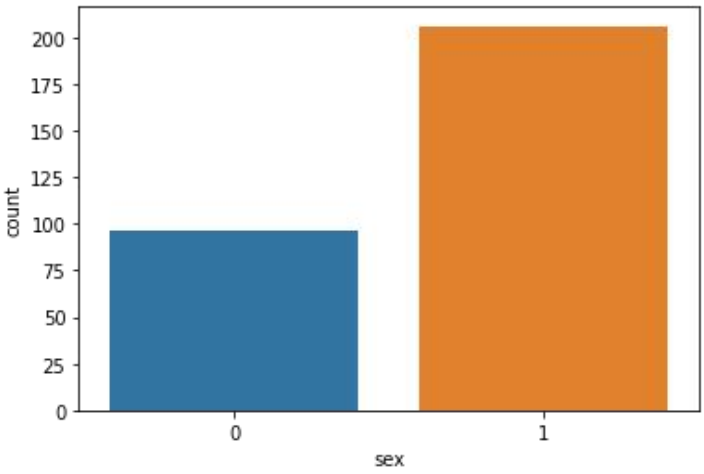
首先，查看数据集中是否存在缺失值，图 4 显示不存在缺失值。

图 4 查看缺失值

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

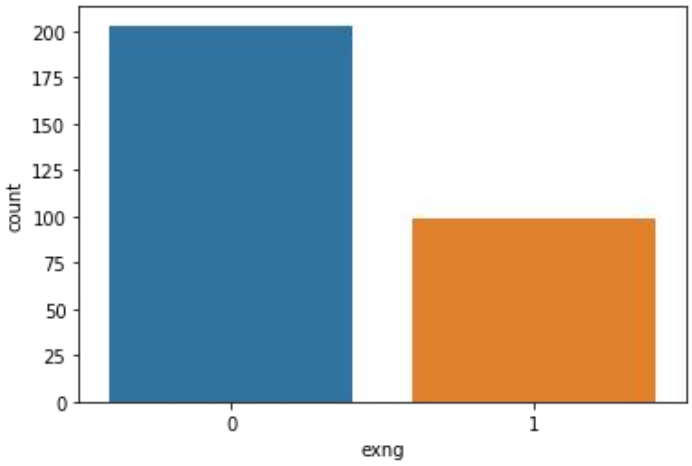
查看数据集中的男女人数，其中女性 96 人，男性 207 人。

图 5 男女病人数



查看数据集中患有运动型心绞痛的人数，其中 99 名病人具有运动型心绞痛。

图 6 运动型心绞痛人数



查看病人的年龄、静息心率、胆固醇、最大心率四个连续变量是否存在异常值。图 7 结果显示，病人年龄分布集中在 40-70 之间，年龄最小的病人不到 30 岁，年龄最大的超过了 70 岁。静息心率 60-100 次/分属于正常，根据图 8 可以看出绝大多数病人静息心率高于正常水平，最高值达 200 次/分。图 9 显示的病人的胆固醇分布大多集中在 200-300 之间，超过了正常的 0-200mg/dl，存在一位病人胆固醇值超 500mg/dl，可以认为该病人胆固醇值过高，而不作为异常值处理。图 10 显示的病人最大心率分布集中在 120-180 次/分，不存在异常值。

图 7 年龄分布

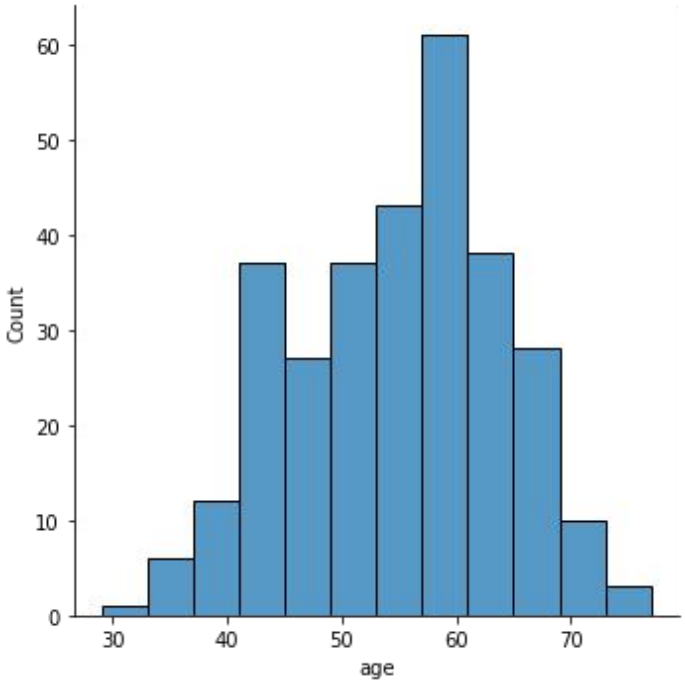


图 8 静息心率分布

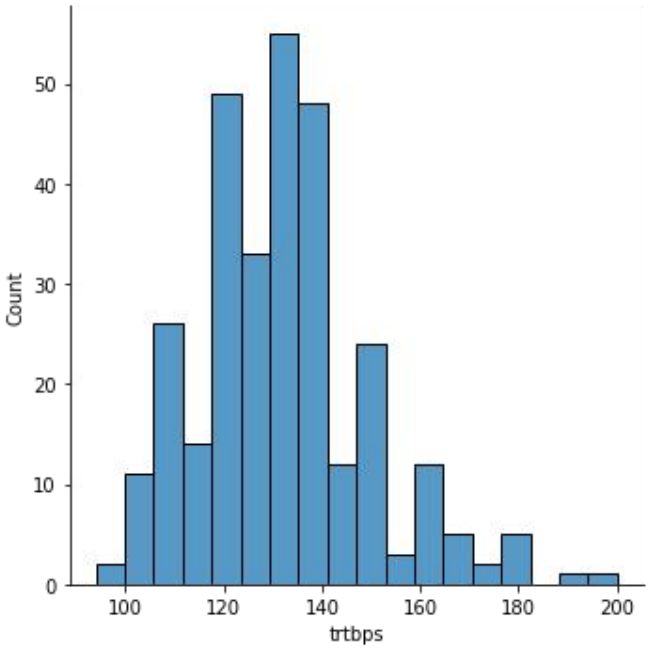


图 9 胆固醇分布

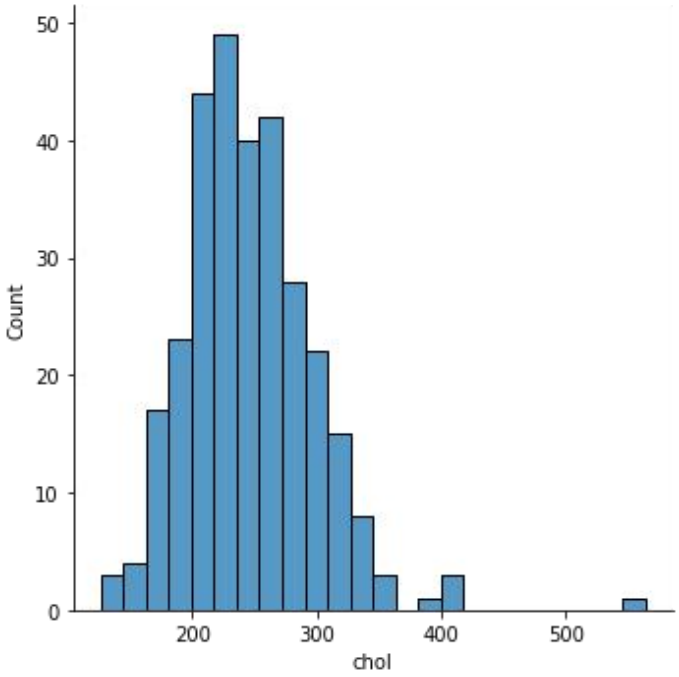


图 10 最大心率分布

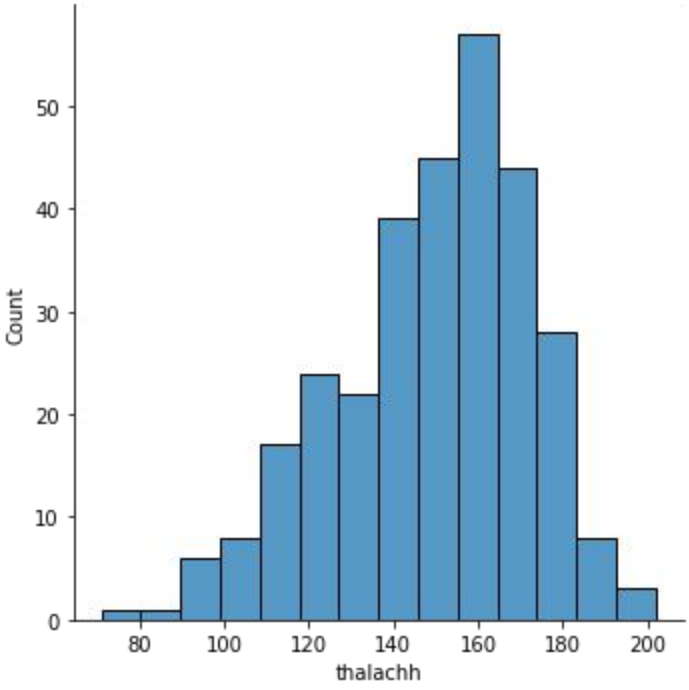
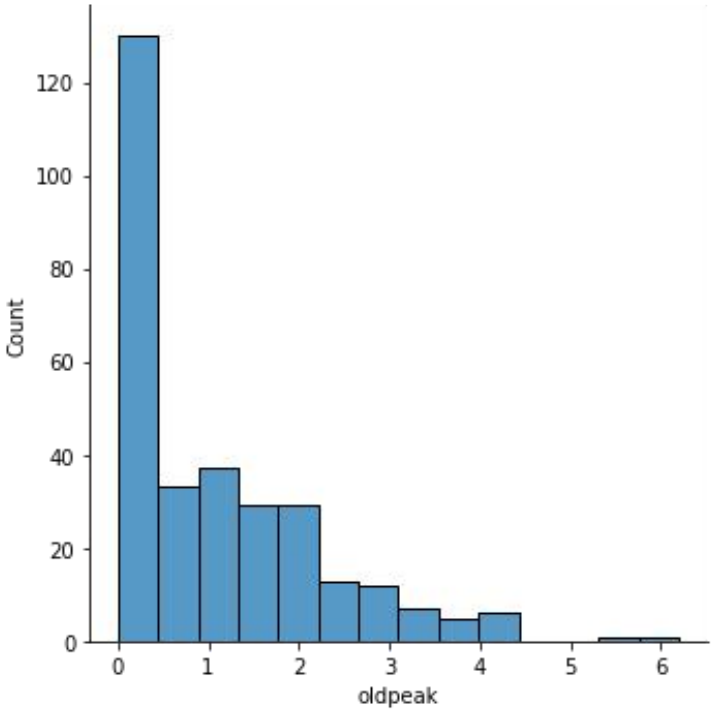


图 11oldpeak 分布



将病人心脏病是否发作的实际结果按照离散型变量进行分类，并绘制柱状图。图 12 显示，303 名病人中，有约 140 名未发作心脏病，约 160 名发作了心脏病。图 13 中，按性别划分的结果显示，女性病人中心脏病发作的比例高于男性。按是否患有心绞痛的分类结果显示，胸部无疼痛症状的病人心脏病发作的概率较低。同样，不患有运动型心绞痛的病人心脏病发作的人数远少于患有运动型心绞痛的病人。

图 12 心脏病发作结果分类

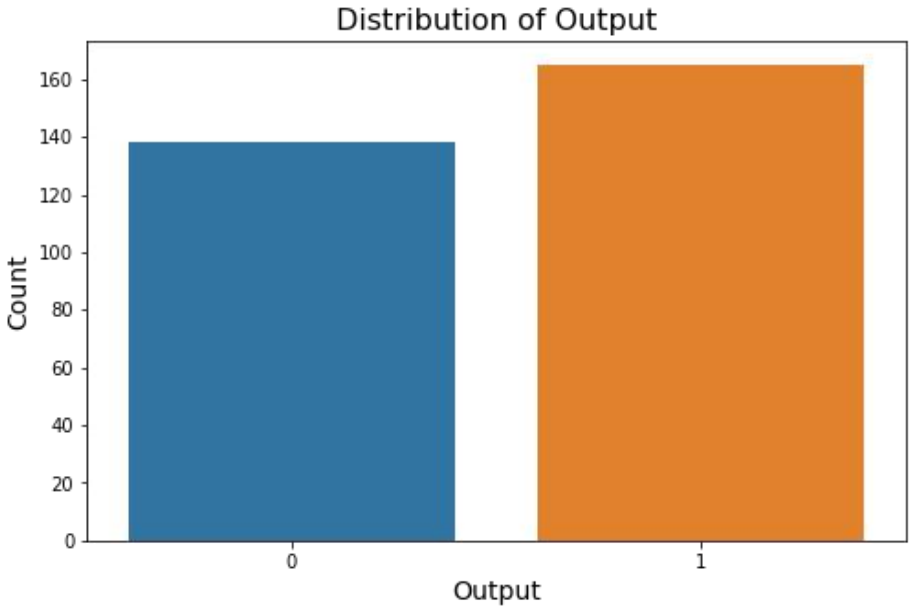
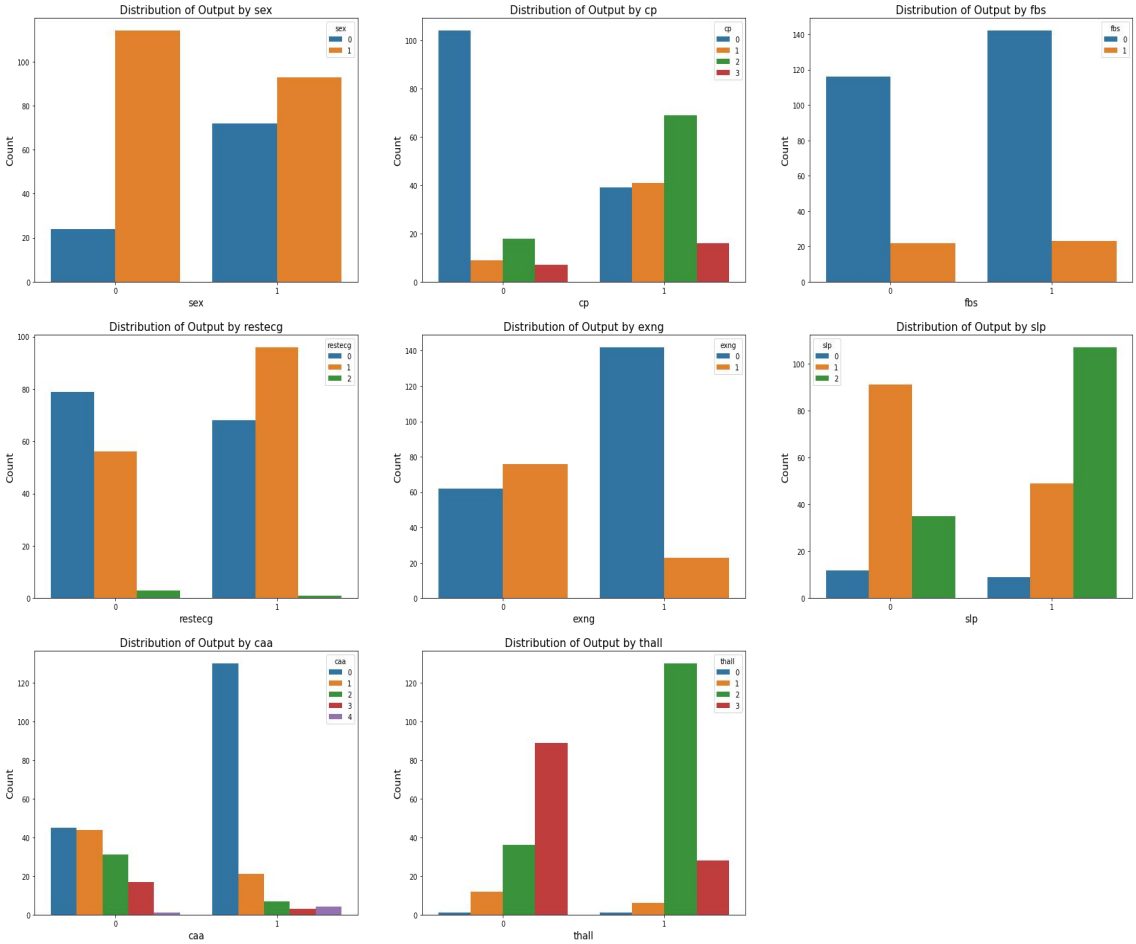
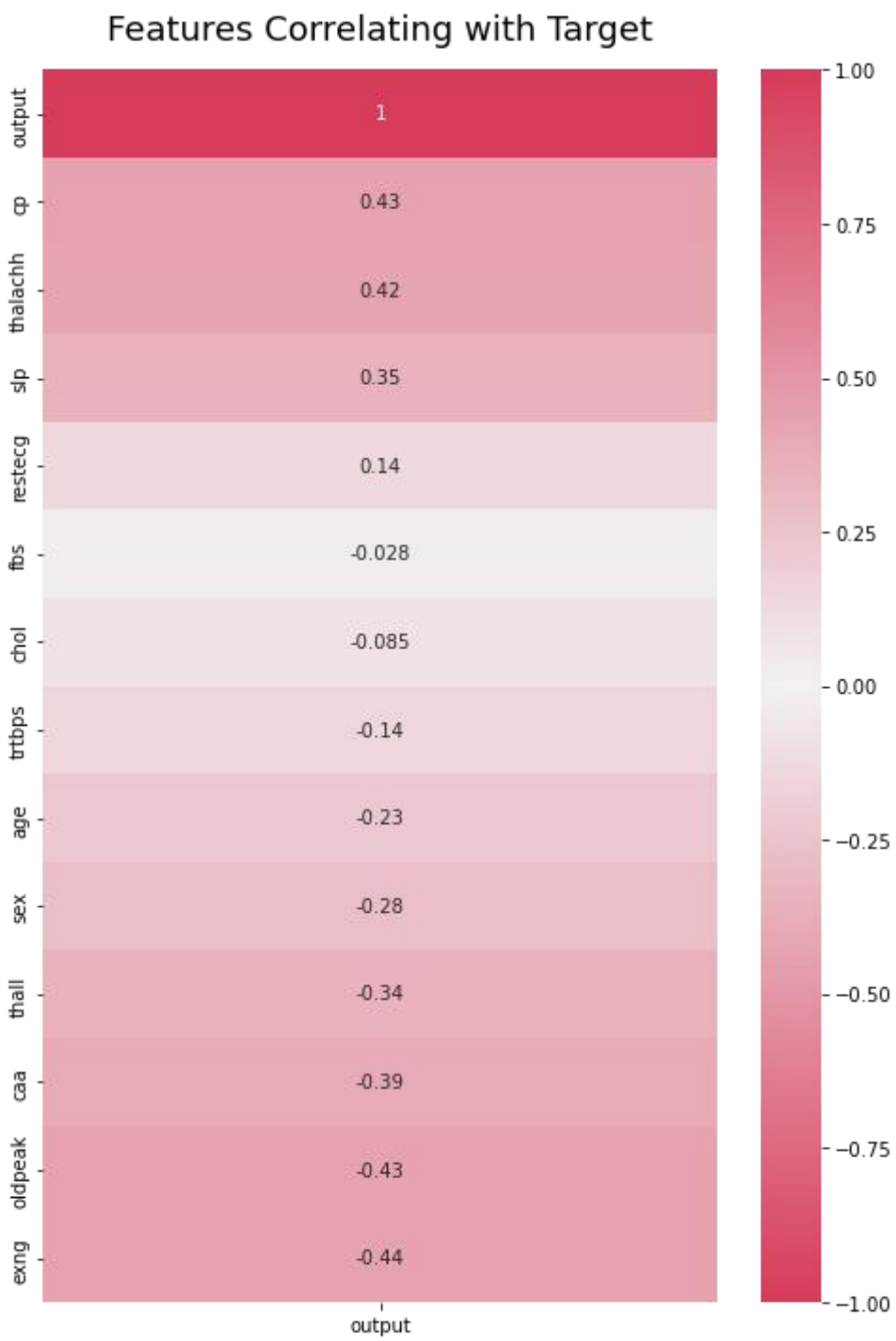


图 13 按离散变量的心脏病发作结果分类



考察各特征与心脏病是否发作之间的关联性，图 14 结果显示，与心脏病发作呈正相关的指标共有四个，分别为静息心电图结果、呼吸机参数、最大心率、胸部疼痛类型，其中胸部疼痛类型和最大心率的相关系数都超过 0.4。与心脏病发作呈负相关的指标共有 9 个，其中有无运动型心绞痛的负相关性最为显著，系数为-0.44。

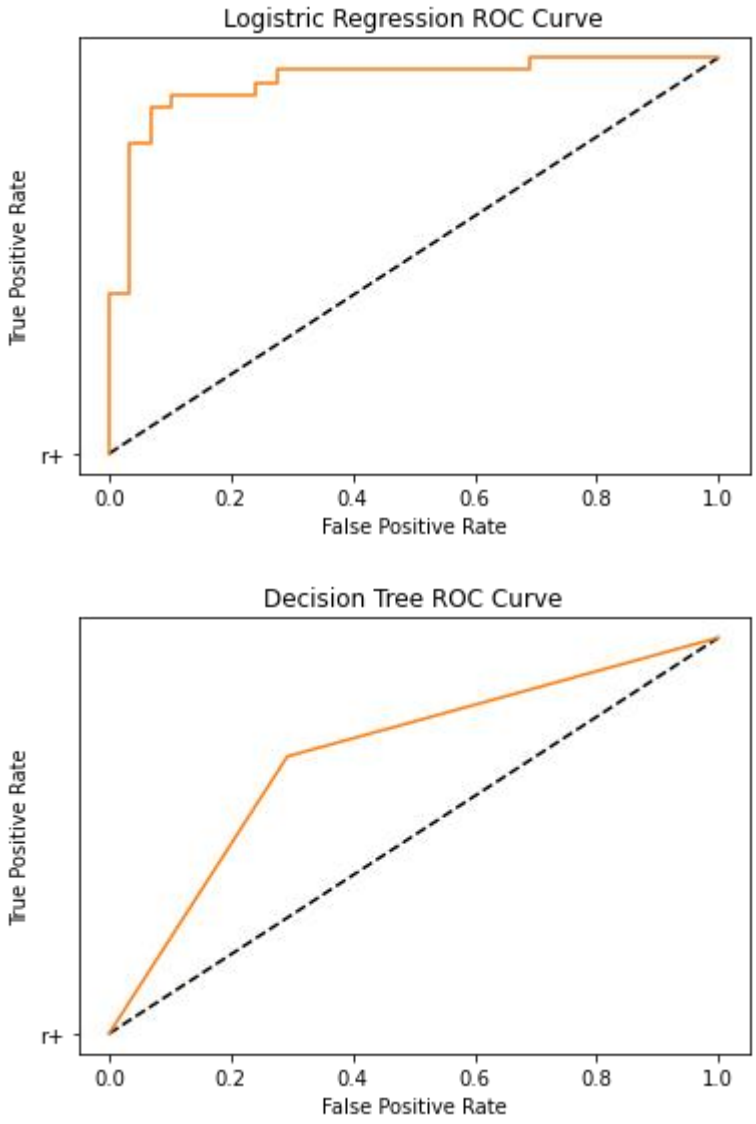
图 14 各变量与心脏病发作结果间的相关性

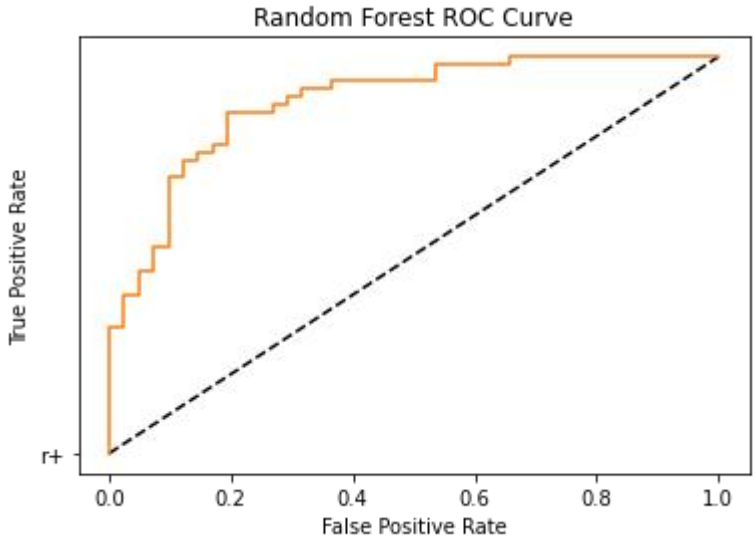


3.3 模型结果对比

将 303 个样本随机划分为 212 个训练集和 91 个测试集，分别使用逻辑回归、决策树、随机森林模型进行训练测试，得到的精确度分别为 90.16%、70.33%、75.82%，分别绘制 ROC 曲线图如下：

图 15 各模型 ROC 曲线图





四、 总结

本文决策树与随机森林的预测精确度都较低，分别为 70.33%、75.82%，随机森林算法在处理高纬数据时会更有优势，同时面对不平衡、部分特征缺失的数据时，随机森林都能更好地维持精确度，此外随机森林还具有较好的抗过拟合能力，但是面对低维数据时，随机森林往往达不到很好的分类效果。

本文选取的数据集数据量较少且数据较为完整，不存在缺失值、异常值，与预测目标关联度较高的变量同样较少，同时属于典型的二分类的问题，因此逻辑回归在训练和测试中呈现出的结果相对较好，精确度达 90.16%。

逻辑回归实现较为简单，被广泛应用于处理二分类的问题，在分类时计算量小、速度快、存储资源低，同时逻辑回归能够通过正则化方式很好地处理多重共线性问题。但是面对高维数据或者多分类问题时，逻辑回归往往效果一般，因此要充分考虑数据、问题的具体情况，选择合适的算法模型。

参考文献

[1] Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest[J]. Applied Soft Computing Journal,2022,118.

[2] Krzywinski, M., Altman, N. Classification and regression trees. Nat Methods 14, 757–758 (2017).

[3]Li Chaozhi. Predictors selection strategy based on stepwise random forests and logistic regression model[P]. Sichuan University (China),2023.

[4]Oh Gyeongseok,Song Juyoung,Park Hyoungah,Na Chongmin. Evaluation of Random Forest in Crime Prediction: Comparing Three-Layered Random Forest and Logistic Regression[J]. Deviant Behavior,2022,43(9).

[5]Tahani Daghistani,Riyad Alshammari. Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes[J]. Journal of Advances in Information Technology,2020,11(2).