南非心脏病数据分析报告

黄脉 PB22151749

December 20, 2024

Abstract

本报告旨在通过分析南非心脏病相关数据集，运用统计方法探讨心脏病的发病率、影响因素及模型分类。首先，采用描述性统计方法对患者的基本特征进行分析，包括年龄、吸烟情况等变量。随后，运用多元回归模型深入分析各影响因素对心脏病发病率的具体影响程度。研究结果 tobacco，ldl ，typea，age，famhist 等因素与心脏病发病率显著相关。此外，SVM with RFE 在分类中的综合性能最好。

关键词：心脏病，发病率，统计分析，机器学习，可视化，致病因子

Contents

第 1 章　引言

1.1　引言

1.1.1　研究背景与目的

在南非，心脏病作为一种常见的慢性疾病，对公共卫生构成了重大挑战。了解心脏病的发病率及其影响因素对于制定有效的预防和治疗策略至关重要。

1.1.2　问题的陈述

本研究旨在识别和评估南非人群中心脏病的主要影响因素，并探究哪种模型预测效果最好。

1.1.3　研究的目的和意义

通过深入分析心脏病的影响因素，本研究旨在为公共卫生政策制定者提供科学依据，帮助他们设计更有效的干预措施。

1.2　方法

本研究使用了来自南非国家卫生数据库的心脏病患者数据。首先采用描述性统计方法对患者的基本特征进行分析，随后运用多元回归模型深入分析各影响因素对心脏病发病率的具体影响程度。

1.3　结果

研究结果显示，tobacco，ldl ，typea，age，famhist 等因素与心脏病发病率显著相关。此外，SVM with RFE 的综合性能最好。

第 2 章　探索性数据分析

2.1　数据集介绍与任务类型确定

本项目使用的数据集来自南非心脏病相关数据，共包含 462 个样本和 10 个特征。以下是对数据集的详细介绍及任务类型的确定。

2.1.1　数据预览与基本信息

首先，通过查看数据的前五行、数据总数以及各变量的数据类型，初步了解数据的结构和内容。

Table 2.1: 数据集前五行示例

sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
160	12.00	5.73	23.11	Present	49	25.30	97.20	52	1
144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	1
118	0.08	3.48	32.28	Present	52	29.14	3.81	46	0
170	7.50	6.41	38.03	Present	51	31.99	24.26	58	1
134	13.60	3.50	27.78	Present	60	25.99	57.34	49	1

2.1.2　数据特征概述

数据集中包含以下主要特征：

- **sbp**（收缩压）：连续变量，单位 mmHg。
- **tobacco**（烟草使用量）：连续变量，单位克。
- **ldl**（低密度脂蛋白）：连续变量，单位 mmol/L。
- **adiposity**（肥胖度）：连续变量，体脂百分比。
- **famhist**（家族史）：分类变量，是否有心脏病家族史（Present/Absent，已编码为 1/0）。

- **typea**（Type A 性格）：连续变量，评分。
- **obesity**（肥胖状态）：连续变量，体重指数（BMI）。
- **alcohol**（酒精消费量）：连续变量，单位克。
- **age**（年龄）：连续变量，单位岁。
- **chd**（冠心病）：目标变量，二分类（1= 有，0= 无）。

其中，‘famhist‘为分类变量，表示是否有心脏病家族史（Present/Absent），其余变量均为数值型变量。

2.1.3 数据编码与缺失值检查

为了便于后续分析，将 ‘famhist‘变量中的分类数据进行编码，将“Present”映射为 1，“Absent”映射为 0。

Table 2.2: 编码后的 ‘famhist‘列前五行

索引	famhist
0	1
1	0
2	1
3	1
4	1

编码后的 ‘famhist‘列唯一值为 1 和 0，分别表示有和无家族病史。
此外，数据集中不存在缺失值，各变量的缺失值数量均为 0，如下表所示：

Table 2.3: 缺失值检查结果

变量	**缺失值数量**
sbp	0
tobacco	0
ldl	0
adiposity	0
famhist	0
typea	0
obesity	0
alcohol	0
age	0
chd	0

2.1.4 极端值检测与处理

针对数值型变量（‘sbp‘, ‘tobacco‘, ‘ldl‘, ‘adiposity‘, ‘typea‘, ‘obesity‘, ‘alcohol‘, ‘age‘）进行了极端值的检测，发现共有 41 个极端值样本。为了保证数据质量，删除这些极端值后的数据集形状为 421 行 10 列。

2.1.5 数据标准化

为了消除不同变量量纲的影响，对数值型变量进行了标准化处理。标准化后的数据前五行及描述性统计如下：

Table 2.4: 标准化后的数据前五行

sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	famhist	chd
0.419544	-0.847256	-0.091844	0.465824	0.204035	0.785953	-0.656554	1.423183	0	1
-1.048592	-0.828294	-0.611869	0.940686	-0.112934	0.855226	-0.566158	0.268077	1	0
1.887680	1.181681	1.026489	1.684681	-0.218591	1.586442	0.490184	1.083446	1	1
-0.145124	2.834085	-0.600685	0.358430	0.732316	0.047041	2.198927	0.471920	1	1
-0.258057	0.829529	1.060039	1.449190	0.943629	1.273430	-0.032563	0.200130	1	0

Table 2.5: 标准化后数据的描述性统计

变量 chd	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	famhist
count 421	421	421	421	421	421	421	421	421	421
mean 0.3207	-7.09e-16	-8.44e-18	4.39e-16	2.30e-16	-2.53e-16	0.4133	-3.38e-17	0.3207	0.4133
std 0.4673	1.0012	1.0012	1.0012	1.0012	1.0012	0.4930	1.0012	0.4673	0.4930
min 0	-2.01	-0.85	-2.01	-2.36	-2.97	0.0000	-0.76	-1.84	0
25% 0	-0.7098	-0.8418	-0.7293	-0.7259	-0.6412	0.0000	-0.7366	-0.7511	0
50% 0	-0.1451	-0.3895	-0.1534	0.0971	-0.0073	0.0000	-0.4158	0.1322	0
75% 1	0.5325	0.4097	0.5679	0.7544	0.7323	1.0000	0.3528	0.8796	1
max 1	3.24	3.48	3.36	2.22	2.63	3.05	3.90	1.49	1

为了进一步评估数据质量，特别是检测是否存在过采样问题，我们通过方差分析来判断各特征的变异程度。标准化后的数据大部分数值型变量的标准差均接近 1，表明数据在各个特征上具有适当的变异性。这种均衡的方差分布有助于：

- **避免过拟合**: 合理的方差分布减少了某些特征因方差过大而在模型训练中占据主导地位，从而降低了模型过拟合的风险。

- **提高模型稳定性**: 均衡的方差分布使得模型在不同特征上的学习更加均衡，提升了模型的稳定性和泛化能力。
- **确认数据无过采样**: 方差接近 1 且均衡的分布表明数据在各个特征上具有足够的多样性，未出现因过采样导致的某些特征重复或冗余的情况。

此外，通过观察标准化后数据的最小值和最大值，可以确认数据未被极端值或重复样本所主导，进一步验证了数据集的多样性和代表性。

2.1.6 任务类型确定

根据数据集的特征和目标变量 ‘chd‘（冠心病发病情况），本项目的任务类型确定为 ** 二分类任务 **，旨在通过不同的机器学习模型预测个体患有冠心病的概率。

2.2 相关系数矩阵及可视化

在本节中，我们将展示并分析标准化后的数据的相关系数矩阵。相关系数矩阵用于衡量各变量之间的线性关系，帮助我们识别潜在的重要特征及其相互关系。以下表格展示了各变量之间的相关系数：

Table 2.6: 标准化后数据的相关系数矩阵

变量	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	famhist	chd
sbp	1.000000	0.176726	0.216120	0.365451	-0.072947	0.296126	0.148785	0.387176	0.118159	0.046773
tobacco	0.176726	1.000000	0.196769	0.308452	0.008411	0.164231	0.238313	0.457826	0.178355	0.265644
ldl	0.216120	0.196769	1.000000	0.462736	0.033600	0.376111	-0.050479	0.366162	0.199811	0.265644
adiposity	0.365451	0.308452	0.462736	1.000000	-0.061012	0.765086	0.133864	0.637513	0.199832	0.265644
typea	-0.072947	0.008411	0.033600	-0.061012	1.000000	0.052592	0.009810	-0.110953	0.034663	0.265644
obesity	0.296126	0.164231	0.376111	0.765086	0.052592	1.000000	0.121322	0.337564	0.135780	0.265644
alcohol	0.148785	0.238313	-0.050479	0.133864	0.009810	0.121322	1.000000	0.103655	0.046773	0.265644
age	0.387176	0.457826	0.366162	0.637513	-0.110953	0.337564	0.103655	1.000000	0.265644	0.265644
famhist	0.118159	0.178355	0.199811	0.199832	0.034663	0.135780	0.046773	0.265644	1.000000	0.265644
chd	0.046773	0.265644	0.265644	0.265644	0.265644	0.265644	0.265644	0.265644	0.265644	1.000000

2.2.1 相关系数矩阵分析

通过观察相关系数矩阵，我们可以识别出各变量之间的线性关系强度和方向。以下是一些关键发现：

- **收缩压（sbp）与年龄（age）**之间存在中等正相关（$r = 0.387$），表明随着年龄的增加，收缩压有上升的趋势。
- **收缩压（sbp）与肥胖指数（obesity）**之间存在中等正相关（$r = 0.296$），显示出肥胖可能与高血压相关。
- **吸烟量（tobacco）与年龄（age）**之间存在较强的正相关（$r = 0.458$），可能反映出随着年龄的增长，吸烟量的增加。
- **低密度脂蛋白（ldl）与脂肪含量（adiposity）**之间存在中等正相关（$r = 0.463$），表明高脂肪含量可能导致低密度脂蛋白水平的升高。
- **脂肪含量（adiposity）与肥胖指数（obesity）**之间存在强正相关（$r = 0.765$），这是预期中的结果，因为肥胖指数是衡量体脂百分比的指标之一。
- **年龄（age）与脂肪含量（adiposity）**之间存在较强的正相关（$r = 0.638$），表明随着年龄的增长，体脂百分比可能增加。
- **其他变量之间的相关性较弱**，如类型 A 性格（typea）与其他变量的相关系数均低于 0.2，表明类型 A 性格与心脏病相关性较低。

这些相关性分析为我们理解各变量之间的关系提供了重要的见解，帮助我们在后续的模型构建中选择和处理特征。

2.2.2 相关系数矩阵的可视化

为了更直观地展示各变量之间的相关关系，我们可以使用热图（Heatmap）进行可视化。热图通过颜色的深浅来表示相关系数的大小和方向，便于快速识别强相关和弱相关的变量对。

以下为相关系数矩阵的热图示例：

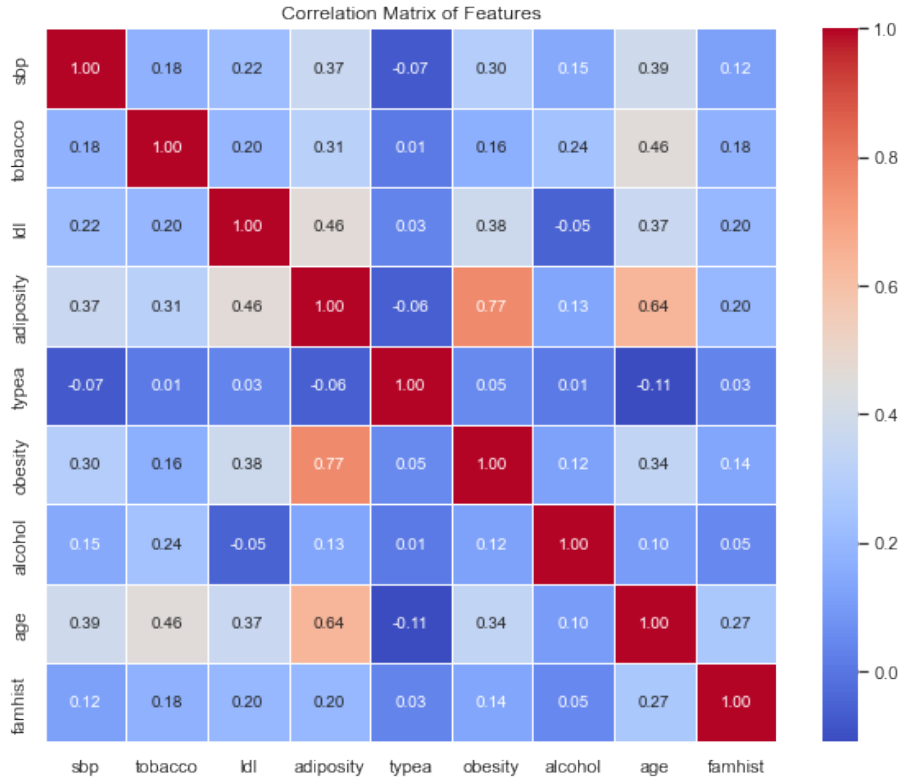


Figure 2.1: 标准化后数据的相关系数热图

图 2.2 展示了各变量之间的相关系数热图。通过颜色的深浅，可以直观地看到哪些变量之间存在强相关关系，哪些变量之间相关性较弱。特别是，脂肪含量（adiposity）与肥胖指数（obesity）的强正相关在图中表现得尤为明显。

2.2.3 结论

adiposity（肥胖度）、**obesity（肥胖状态）**的相关性最强，我们将结合接下来的 VIF 来做进一步的特征选取。其余变量由于相关性都没有这两个强，可以暂时不处理。在 **adiposity（肥胖度）**、**obesity（肥胖状态）**做完特征选取后，我们可以再次查看新的相关系数矩阵，判断是否进行下一步处理。

2.3 计算 VIF（方差膨胀因子）

在多元线性回归分析中，方差膨胀因子（Variance Inflation Factor, VIF）用于评估自变量之间的多重共线性程度。多重共线性指的是两个或多个自变量之间存在高度相关性，这可能导致回归系数的不稳定性和解释性的降低。通常，VIF 值超过 5 或 10 被认为存在严重的多重共线性问题，需要进一步处理。

以下表格展示了各自变量的 VIF 值：

Table 2.7: 各特征的方差膨胀因子（VIF）

特征	VIF
sbp	1.241939
tobacco	1.340871
ldl	1.339602
adiposity	4.128364
typea	1.045391
obesity	2.748950
alcohol	1.106764
age	2.265299
famhist	1.055453

以下是 VIF 的条形图示例：

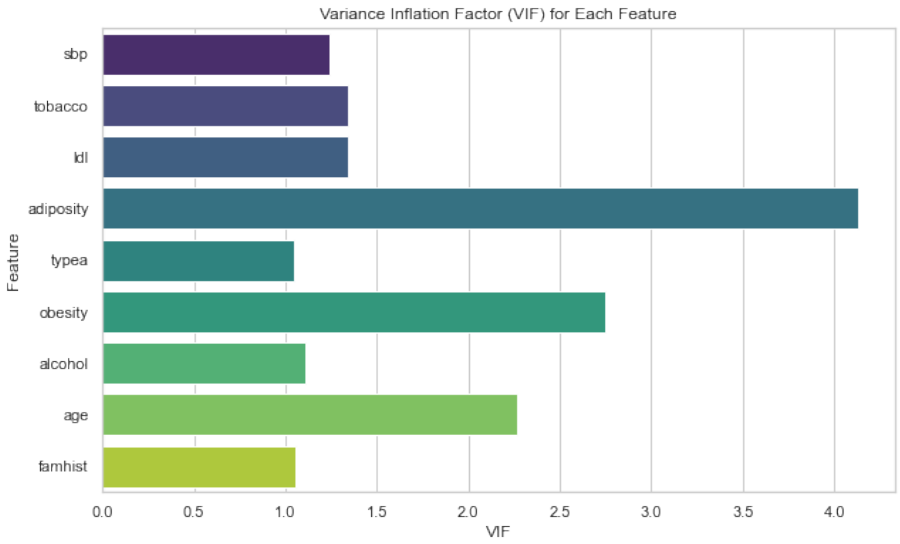


Figure 2.2: 数据集的 VIF 条形图

2.3.1 VIF 值分析

从表中可以看出，所有自变量的 VIF 值均低于 5，表明在本数据集中不存在严重的多重共线性问题。具体分析如下：

- **adiposity（肥胖度）**的 VIF 值为 4.13，接近 5，值得注意。这表明肥胖度与其他自变量之间存在一定程度的相关性，我们之后采取方法提取其特征。
- **obesity（肥胖状态）**和 **age（年龄）**的 VIF 值分别为 2.75 和 2.27，显示出中等程度的相关性，但依然在可接受范围内。

- 其他变量如 **sbp（收缩压）**、**tobacco（烟草使用量）**、**ldl（低密度脂蛋白）**、**typea（Type A 性格）**、**alcohol（酒精消费量）**和 **famhist（家族史）**的 VIF 值均接近 1，表明它们之间的相关性非常低，几乎不存在多重共线性问题。

2.3.2 结论

根据 VIF，除了 **adiposity（肥胖度）**、**obesity（肥胖状态）**外，其他自变量之间的多重共线性问题较为轻微，整体数据质量良好。**结合相关系数**针对 **adiposity（肥胖度）**、**obesity（肥胖状态）**的高相关系数和高 VIF 值，我们将利用 PCA 提取其主要成分，进一步优化模型，确保其稳定性和预测准确性。低多重共线性的自变量不仅提升了模型的稳定性和解释性，还有助于提高预测准确性，因此，可以放心地使用这些自变量进行后续分析。

2.4 提取主成分，进行降维

在前面的分析中，我们发现 **adiposity（肥胖度）**与 **obesity（肥胖状态）**之间存在较高的相关性（VIF 值为 4.13），这可能导致多重共线性问题，影响模型的稳定性和解释性。为了解决这一问题，我们采用主成分分析（Principal Component Analysis, PCA）对这两个变量进行降维处理，提取出一个主要的主成分 **adiposity_obesity_pc**，以替代原有的两个变量。

主成分分析的步骤如下：

1. **标准化数据**: 对 **adiposity** 和 **obesity** 进行标准化处理，使其均值为 0，标准差为 1。
2. **计算协方差矩阵**: 计算标准化后的 **adiposity** 和 **obesity** 之间的协方差矩阵。
3. **特征值分解**: 对协方差矩阵进行特征值分解，提取出主要的主成分。
4. **构建主成分**: 根据特征值和特征向量，构建新的主成分 **adiposity_obesity_pc**。

通过上述步骤，我们成功提取了一个主成分 **adiposity_obesity_pc**，该主成分能够有效地代表原有的 **adiposity** 和 **obesity** 信息，减少了变量的维度，并降低了多重共线性的风险。

2.5 新数据集的相关系数矩阵及可视化

在提取主成分后，我们构建了一个新的数据集，该数据集将原有的 **adiposity** 和 **obesity** 替换为主成分 **adiposity_obesity_pc**。接下来，我们对新数据集进行相关系数矩阵的计算和可视化，以评估各变量之间的线性关系。

2.5.1 新数据集的相关系数矩阵

以下表格展示了新数据集中各变量之间的相关系数：

Table 2.8: 新数据集的相关系数矩阵

变量	sbp	tobacco	ldl	typea	alcohol	age	famhist	adiposity_obesity_pc
sbp	1.000	0.177	0.216	-0.073	0.149	0.387	0.118	0.352
tobacco	0.177	1.000	0.197	0.008	0.238	0.458	0.178	0.252
ldl	0.216	0.197	1.000	0.034	-0.050	0.366	0.200	0.446
typea	-0.073	0.008	0.034	1.000	0.010	-0.111	0.035	-0.004
alcohol	0.149	0.238	-0.050	0.010	1.000	0.104	0.047	0.136
age	0.387	0.458	0.366	-0.111	0.104	1.000	0.266	0.519
famhist	0.118	0.178	0.200	0.035	0.047	0.266	1.000	0.179
adiposity_obesity_pc	0.352	0.252	0.446	-0.004	0.136	0.519	0.179	1.000

2.5.2 相关系数矩阵的可视化

为了更直观地展示各变量之间的相关关系，我们可以使用热图（Heatmap）进行可视化。热图通过颜色的深浅来表示相关系数的大小和方向，便于快速识别强相关和弱相关的变量对。

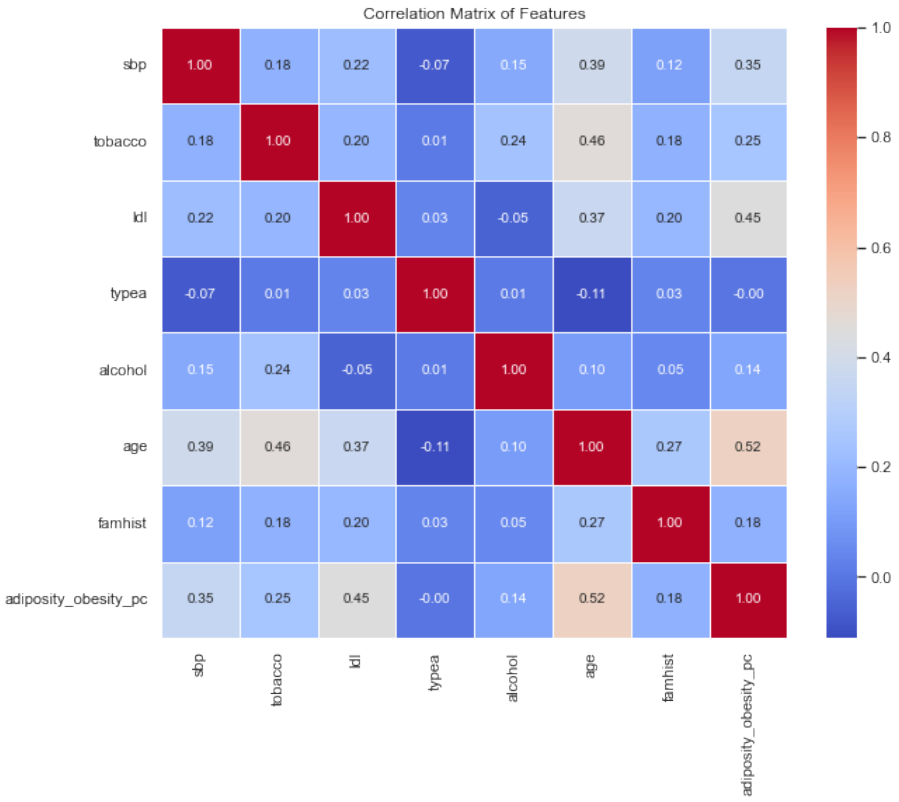


Figure 2.3: 新数据集的相关系数热图

图 2.4 展示了新数据集的相关系数热图。通过颜色的深浅，可以直观地看到哪些变量之间存在强相关关系，哪些变量之间相关性较弱。特别是，主成分（adiposity_obesity_pc）与收缩压（sbp）和低密度脂蛋白（ldl）的中等正相关在图中表现得尤为明显。

2.5.3 相关系数矩阵分析

通过观察新数据集的相关系数矩阵，我们可以识别出各变量之间的线性关系强度和方向。以下是一些关键发现：

- **收缩压（sbp）与年龄（age）**之间存在中等正相关（$r = 0.387$），表明随着年龄的增加，收缩压有上升的趋势。
- **收缩压(sbp)与主成分(adiposity_obesity_pc)**之间存在中等正相关($r = 0.352$)，说明肥胖相关因素对收缩压有一定影响。
- **吸烟量（tobacco）与年龄（age）**之间存在较强的正相关（$r = 0.458$），可能反映出随着年龄的增长，吸烟量的增加。
- **低密度脂蛋白（ldl）与主成分（adiposity_obesity_pc）**之间存在中等正相关（$r = 0.446$），表明高脂肪含量可能导致低密度脂蛋白水平的升高。
- **主成分(adiposity_obesity_pc)**与年龄(age)之间存在较强的正相关($r = 0.519$)，表明随着年龄的增长，体脂百分比可能增加。
- **其他变量之间的相关性较弱**，如类型 A 性格（typea）与其他变量的相关系数均低于 0.2，表明类型 A 性格与其他变量相关性较低。

这些相关性分析为我们理解各变量之间的关系提供了重要的见解，帮助我们在后续的模型构建中选择和处理特征。

2.6 计算新数据集的 VIF（方差膨胀因子）

在构建回归模型之前，评估自变量之间的多重共线性是一个重要步骤。多重共线性可能导致回归系数的不稳定性和模型解释性的降低。为此，我们计算了新数据集各自变量的方差膨胀因子（VIF），以评估自变量之间的多重共线性程度。

2.6.1 新数据集的 VIF 值

以下表格展示了新数据集中各自变量的 VIF 值：

Table 2.9: 新数据集的方差膨胀因子（VIF）

特征	VIF
sbp	1.24
tobacco	1.34
ldl	1.34
typea	1.05
alcohol	1.11
age	2.27
famhist	1.06
adiposity_obesity_pc	1.58

2.6.2 VIF 的可视化

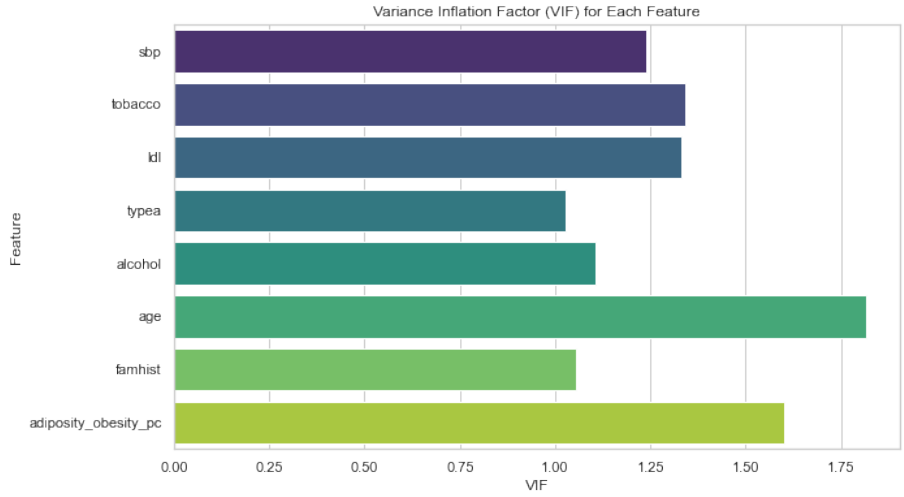


Figure 2.4: 新数据集的 VIF 条形图

2.6.3 VIF 值分析

从表 2.9 中可以看出，所有自变量的 VIF 值均低于 5，表明在新数据集中不存在严重的多重共线性问题。具体分析如下：

- **adiposity_obesity_pc（肥胖主成分）**的 VIF 值为 1.58，较之前的 **adiposity（肥胖度）**的 VIF 值（4.13）显著降低。这表明通过主成分分析，成功减少了原有变量之间的多重共线性问题。
- **age（年龄）**的 VIF 值为 2.27，显示出中等程度的相关性，但依然在可接受范围内。这意味着年龄与其他自变量之间存在一定的相关性，但不足以引起多重共线性的担忧。

- 其他变量如 **sbp（收缩压）**、**tobacco（烟草使用量）**、**ldl（低密度脂蛋白）**、**typea（Type A 性格）**、**alcohol（酒精消费量）**和 **famhist（家族史）**的 VIF 值均接近 1，表明它们之间的相关性非常低，几乎不存在多重共线性问题。

2.6.4 结论

通过方差膨胀因子分析，可以确认新数据集中各自变量之间的多重共线性程度较低。特别是，原有的 **adiposity（肥胖度）**通过主成分分析被替换为 **adiposity_obesity_pc（肥胖主成分）**后，其 VIF 值显著降低，表明多重共线性问题得到了有效缓解。这意味着各自变量在回归模型中相互独立，能够有效地反映各自对目标变量 **chd（冠心病）**的影响。

低多重共线性不仅提升了模型的稳定性和解释性，还有助于提高预测准确性。因此，在后续的回归分析和机器学习模型构建中，我们可以放心地使用这些自变量，而无需担心多重共线性带来的潜在问题。

第 3 章　　数据可视化分析

3.1　患病比例

患病比例为 32.07%。该比例表示在研究样本中，有 32.07% 的个体被诊断为患病。

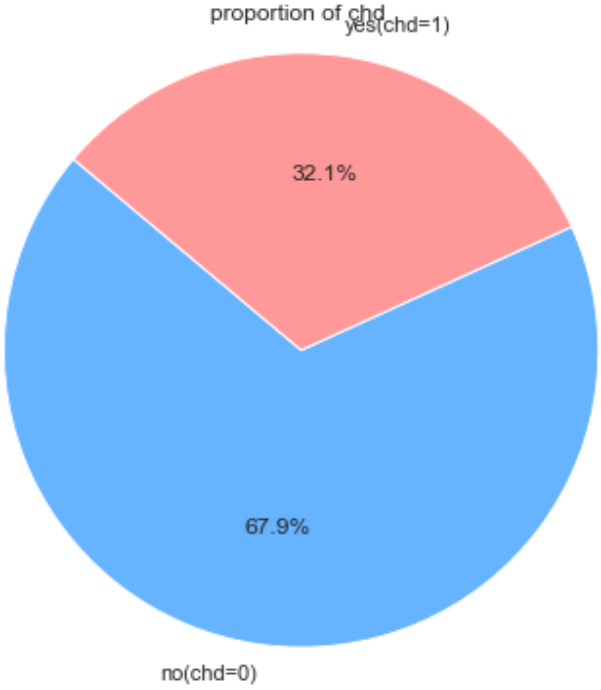


Figure 3.1: 患病比例饼图：直观显示样本中患病和未患病个体的比例。

3.2　对于每个连续变量，进行独立样本 t 检验

以下是对连续变量进行独立样本 t 检验的结果分析。独立样本 t 检验用于比较两个独立样本组在某个连续变量上是否存在显著差异。根据表格中的 p 值，我们可以判断各变量的统计显著性。

分析结果如下：

变量	t 值	p 值
sbp	-3.0087	0.0028
tobacco	-6.1480	0.0000
ldl	-6.4383	0.0000
typea	-2.5703	0.0105
alcohol	-0.7206	0.4716
age	-7.9122	0.0000
adiposity_obesity_pc	-3.9643	0.0001

Table 3.1: 连续变量的独立样本 t 检验结果

sbp（收缩压）：p 值小于 0.05，表示两组间在收缩压上存在显著差异。

tobacco（烟草使用）：p 值小于 0.001，表示两组间在烟草使用上存在极其显著的差异。

ldl（低密度脂蛋白）：p 值小于 0.001，表示两组间在低密度脂蛋白水平上存在极其显著的差异。

typea（A 型行为）：p 值小于 0.05，表示两组间在 A 型行为上存在显著差异。

alcohol（酒精摄入）：p 值大于 0.05，表示两组间在酒精摄入上没有显著差异。

age（年龄）：p 值小于 0.001，表示两组间在年龄上存在极其显著的差异。

adiposity_obesity_pc（肥胖百分比）：p 值小于 0.001，表示两组间在肥胖百分比上存在极其显著的差异。

综上所述，除了酒精摄入外，其他变量在两组间都显示出了统计学上的显著差异，其中烟草使用、低密度脂蛋白、年龄和肥胖百分比的差异尤为显著。这些结果可能表明这些变量与研究中的疾病或条件有关联

3.3 对于两个分类变量（famhist 和 chd），进行卡方检验

对于分类变量 `famhist` 和 `chd` 的卡方检验结果如下：

$$\chi^2 = 27.4438, \quad p\text{-value} = 0.0000$$

分析结果表明，卡方统计量为 27.4438，对应的 p 值小于 0.0001，远小于常用的显著性水平 0.05。这表明在统计上，我们可以拒绝零假设，即认为 `famhist` 和 `chd` 之间存在显著的关联。换句话说，家族病史 (`famhist`) 对心脏病 (`chd`) 的发生有显著影响。具体来说，这个极低的 p 值意味着观察到的 `famhist` 和 `chd` 之间的关联不太可能是由随机因素造成的，从而支持了家族病史是心脏病发生的一个重要因素的结论。

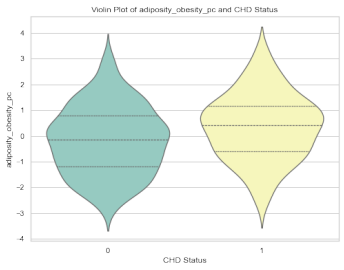
3.4 目标变量 chd 在不同连续变量下的分布情况

在之前的 t 检验中可以看到了除了 alcohol 之外的连续变量都和心脏病有显著关系。此外为了更直观地探究其中的关系，我们绘制小提琴图展示 chd 不同状态下各连续变量的分布情况。

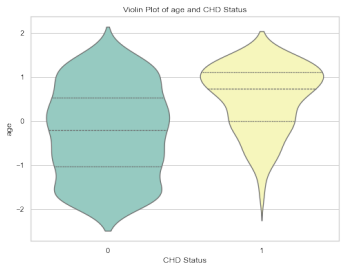
3.5 数据分析

3.5.1 小提琴图

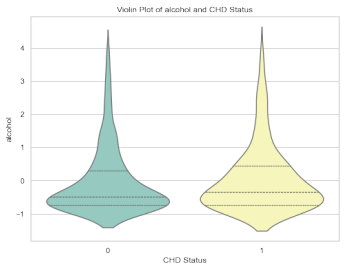
小提琴图展示了每个连续变量在 chd=0 和 chd=1 两种状态下的数据分布密度。以下是各个变量的小提琴图：



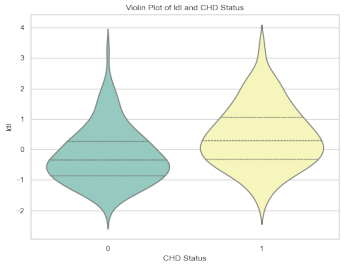
(a) adiposity_obesity_pc



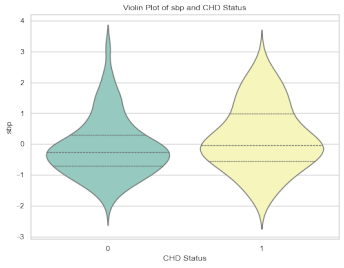
(b) age



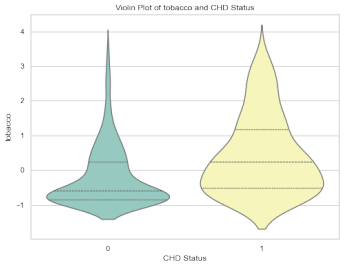
(c) alcohol



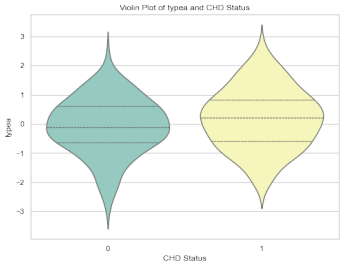
(d) ldl



(e) sbp



(f) tobacco



(g) typea

Figure 3.2: 连续变量的小提琴图

(a) adiposity_obesity_pc（肥胖百分比）：CHD 患者倾向于更高的肥胖百分比。

(b) age（年龄）：CHD 患者年龄分布偏向更年长，表明年龄是 CHD 的一个重要风险因素。

(c) alcohol（酒精摄入）：不同状态下的四分位数线，中位数线，四分之三位数线都相差不大，表明 alcohol 对心脏病可能不是简单的正相关或者负相关，它们之间的复杂关系值得我们在接下来研究。

(d) ldl（低密度脂蛋白）：CHD 患者的 LDL 水平分布偏向更高的值，是 CHD 的一个风险因素。

(e) sbp（收缩压）：CHD 患者的收缩压分布偏向更高的值，尤其是在高收缩压区域。

(f) tobacco（烟草使用）：CHD 患者中烟草使用者的分布更为明显，表明烟草使用是 CHD 的一个风险因素。

(g) typea（A 型行为）：CHD 患者和非患者在 A 型行为上的分布相似，但在高 A 型行为区域，CHD 患者的分布稍微更宽。

我们通过分组条形图和堆叠条形图展示了 `famhist` 与 `chd` 的关系。

3.6 目标变量 `chd` 在分类变量 `famhist` 下的分布情况

我们通过分组条形图和堆叠条形图展示了 `famhist` 与 `chd` 的关系。

3.6.1 分组条形图

分组条形图展示了不同 `famhist` 状态下 `chd` 的患病比例。从图中可以看出，有家族病史的个体中，患冠心病的比例显著高于没有家族病史的个体。这表明家族病史是冠心病的一个重要风险因素。

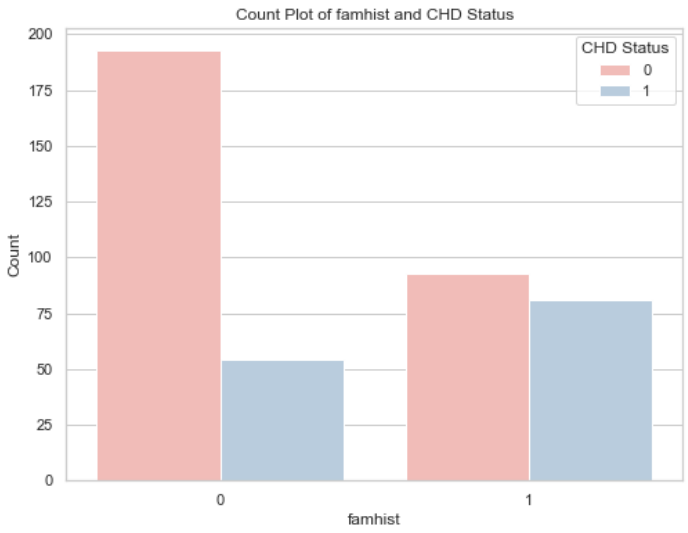


Figure 3.3: Count Plot of `famhist` and CHD Status

3.6.2 堆叠条形图

堆叠条形图紧凑地展示了不同 famhist 状态下 chd 的数量分布。图中显示，无论是否有家族病史，冠心病患者的数量都占据了一定比例，但有家族病史的群体中，冠心病患者的比例更为显著。这也表明家族病史是冠心病的一个重要风险因素。

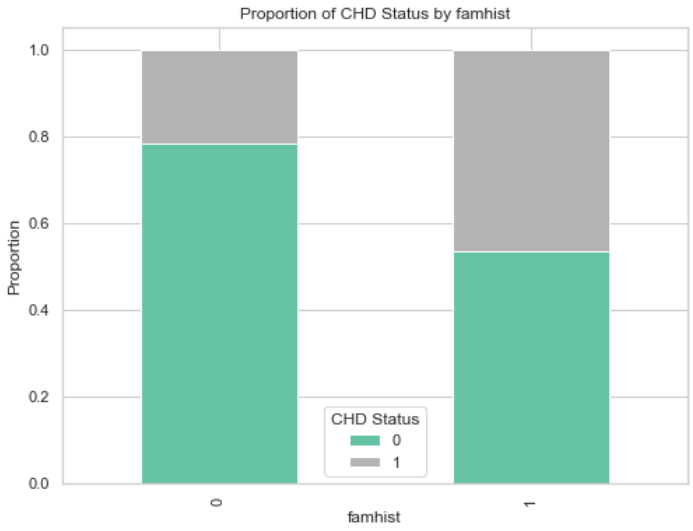


Figure 3.4: Proportion of CHD Status by famhist

3.7 单独研究酒精变量

3.7.1 分类统计

酒精变量的分类统计结果如下：

分类	样本数
Low	110
Medium	101
High	106
Very High	104

Table 3.2: 酒精变量的分类统计结果

3.7.2 酒精消费与 CHD 状态的关系

通过分组条形图、堆叠条形图和直方图，我们分析了酒精消费与冠心病（CHD）状态之间的关系。

分组条形图

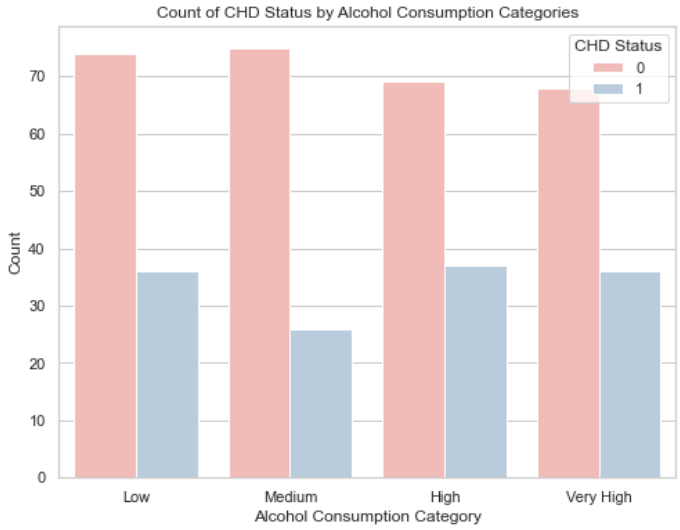


Figure 3.5: Count of CHD Status by Alcohol Consumption Categories

分组条形图显示，在所有酒精消费类别中，非 CHD 患者（用颜色表示为 0）的数量普遍高于 CHD 患者（用颜色表示为 1）。特别是在“Low”和“Medium”酒精消费类别中，非 CHD 患者的数量明显高于 CHD 患者。

堆叠条形图

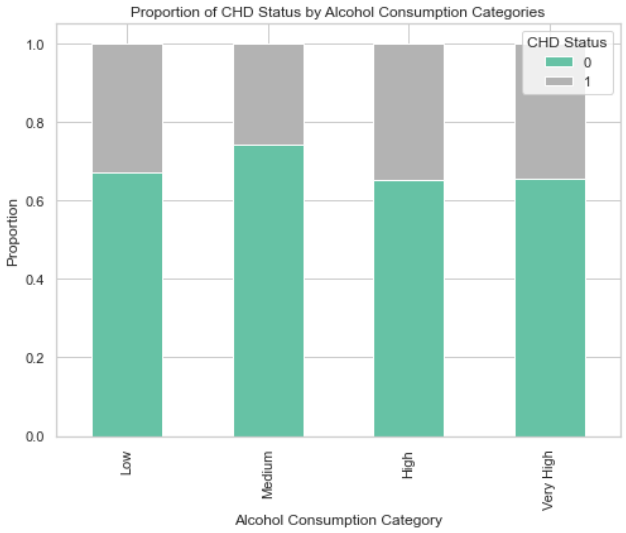


Figure 3.6: Proportion of CHD Status by Alcohol Consumption Categories

堆叠条形图进一步展示了不同酒精消费类别中 CHD 患者和非 CHD 患者的比例。在“Medium”类别中，CHD 患者的比例相对较低，而在“Very High”类别中，CHD 患者的比例略高，**这可能表明极端的酒精消费水平与 CHD 风险增加有关**。

直方图和核密度估计

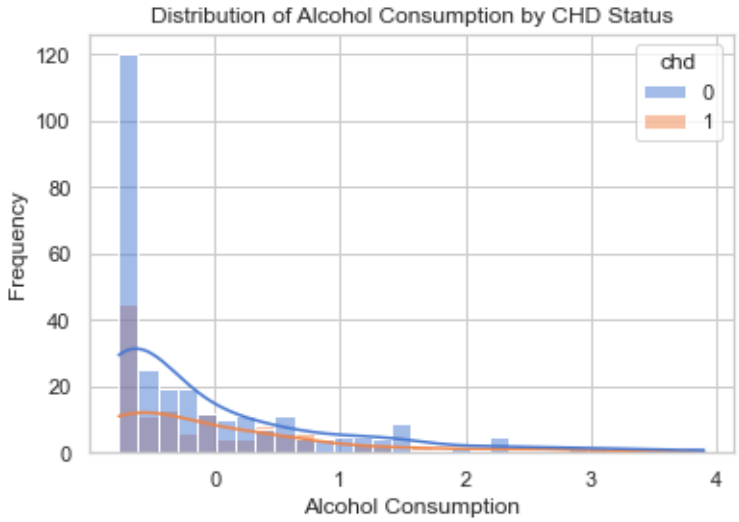


Figure 3.7: Distribution of Alcohol Consumption by CHD Status

直方图和核密度估计（KDE）展示了酒精消费的分布情况。CHD 患者（橙色）在低酒精消费区域的密度略高，而随着酒精消费的增加，非 CHD 患者（蓝色）的密度逐渐占据主导。**这表明适度的酒精消费可能与较低的 CHD 风险相关**。

第 4 章　机器学习

在本章中，我们将应用机器学习模型来预测心脏病（CHD）的发生。我们需要先划分数据集，然后将探索多种机器学习模型，包括但不限于逻辑回归、随机森林、XGBoost、AdaBoost、梯度提升等。每种模型都将在训练集上进行训练，并在测试集上进行评估，以确定其预测性能。

4.1　划分训练集和测试集

为了评估模型的性能，我们使用 `train_test_split` 函数将数据集划分为训练集和测试集。训练集和测试集的划分比例为 4 比 1，确保了模型在未见过的数据上具有泛化能力。我们设置了 `random_state` 为 42，以保证结果的可重复性。此外，通过 `stratify` 参数，我们保持了训练集和测试集中目标变量的分布与完整数据集中的分布相同，这对于处理不平衡数据尤为重要。

划分后的数据集大小如下：

- 训练集样本数: 336
- 测试集样本数: 85

4.2　逻辑回归

4.2.1　逻辑回归（直接运用模型）

在本节中，我们使用逻辑回归模型来预测心脏病（CHD）的发生。我们通过网格搜索（GridSearchCV）对模型的超参数进行了优化，以找到最佳的正则化强度（C）、正则化类型（penalty）和类别权重（class_weight）。

超参数优化过程中，我们考虑了 20 种不同的参数组合，并通过 5 折交叉验证进行了评估。最佳参数如下：

- 正则化强度（C）: 10
- 正则化类型（penalty）: 'l1'

- 类别权重（class_weight）：None

最佳交叉验证 AUC 得分为 0.7705，表明模型具有较好的区分能力。

在测试集上，模型的 AUC 得分为 0.8084，显示了模型良好的泛化能力。分类报告显示，模型在预测未患病（类别 0）时具有较高的精确度和召回率，而在预测患病（类别 1）时的性能相对较低。

```
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.90      0.82        58
           1       0.62      0.37      0.47        27

    accuracy                           0.73        85
   macro avg       0.69      0.63      0.64        85
weighted avg       0.71      0.73      0.71        85
```

混淆矩阵

混淆矩阵如下所示，其中 52 个未患病个体被正确识别，6 个未患病个体被错误分类为患病。同时，17 个患病个体被错误分类为未患病，10 个患病个体被正确识别。具体而言，模型的表现如下：

- 真正例（True Positives, TP）：10（患病个体正确识别）
- 假正例（False Positives, FP）：6（未患病个体错误分类为患病）
- 真负例（True Negatives, TN）：52（未患病个体正确识别）
- 假负例（False Negatives, FN）：17（患病个体错误分类为未患病）

从混淆矩阵中我们可以计算出以下性能指标：

- **准确率**（Accuracy）：$\frac{TP+TN}{TP+FP+TN+FN} = \frac{10+52}{10+6+52+17} = 0.79$
- **精确度**（Precision）：$\frac{TP}{TP+FP} = \frac{10}{10+6} = 0.625$
- **召回率**（Recall）：$\frac{TP}{TP+FN} = \frac{10}{10+17} = 0.370$
- **F1 值**（F1-Score）：$2 \times \frac{Precision \times Recall}{Precision+Recall} = 2 \times \frac{0.625 \times 0.370}{0.625+0.370} = 0.463$

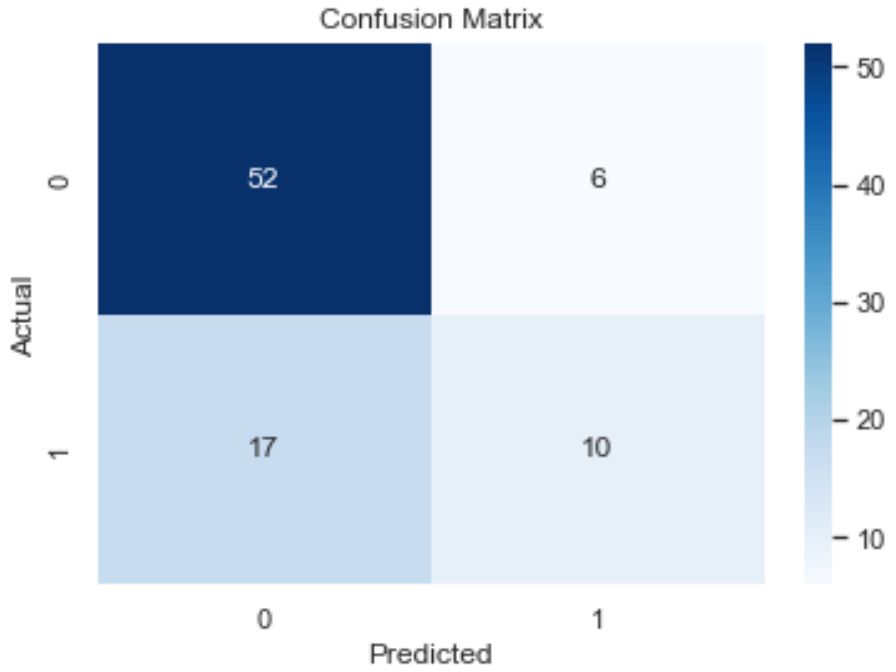


Figure 4.1: Confusion Matrix

ROC 曲线

接收者操作特征（ROC）曲线展示了模型在不同阈值下的性能。ROC 曲线的下面积（AUC）为 0.8084，表明模型的分类能力相对较强，具有良好的区分能力。AUC 值越接近 1，模型的预测性能越好。由于 AUC 为 0.8084，我们可以认为该模型的性能良好，能够有效地区分患病和未患病个体。

统计分析

使用 Statsmodels 进行的详细统计分析显示，以下变量是心脏病发生的显著预测因子：

- **年龄（age）**：系数为 0.7807，P 值为 0.000，表示年龄对患病的风险有显著影响。年龄越大，患病的风险越高。
- **烟草使用量（tobacco）**：系数为 0.3381，P 值为 0.013，表明吸烟量越大，心脏病的风险越高。
- **低密度脂蛋白（ldl）**：系数为 0.4530，P 值为 0.003，表示较高的低密度脂蛋白水平与心脏病发生风险增高相关。
- **家族病史（famhist）**：系数为 0.5351，P 值为 0.046，表示有家族病史的个体患心脏病的风险较高。

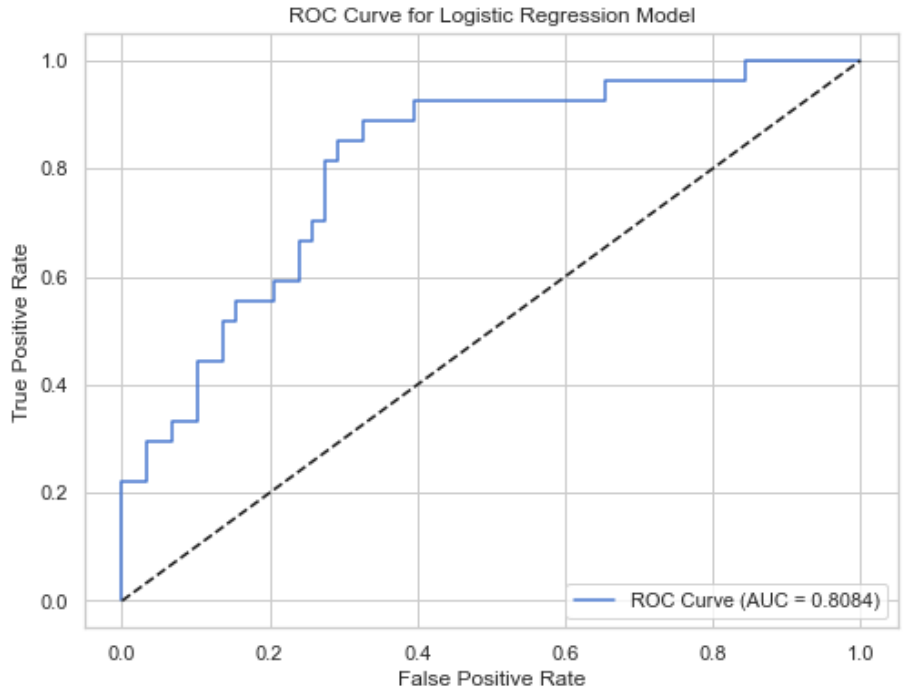


Figure 4.2: ROC Curve with AUC = 0.8084

其他变量如血压(sbp)、酒精消费量(alcohol)、肥胖指数 adiposity_obesity_pc 和性格类型（typea）在统计上未达到显著性水平，尤其是 P 值大于 0.05。

模型的伪 R^2 值为 0.1932，表明该模型能够解释约 19.32% 的因变量变异。尽管如此，这个值表明仍有很多未被模型解释的变异，可能需要进一步引入其他变量或使用更复杂的模型。

Logit Regression Results

Dep. Variable:	chd	No. Observations:	336
Model:	Logit	Df Residuals:	327
Method:	MLE	Df Model:	8
Date:	Thu, 19 Dec 2024	Pseudo R-squ.:	0.1932
Time:	00:02:55	Log-Likelihood:	-170.22
converged:	True	LL-Null:	-210.99
Covariance Type:	nonrobust	LLR p-value:	2.406e-14

	coef	std err	z	P>|z|	[0.025	0
const	-1.2614	0.193	-6.539	0.000	-1.639	-
sbp	0.0086	0.142	0.061	0.952	-0.269	
tobacco	0.3381	0.137	2.474	0.013	0.070	

ldl	0.4530	0.152	2.984	0.003	0.155
typea	0.5148	0.143	3.588	0.000	0.234
age	0.7807	0.196	3.986	0.000	0.397
alcohol	0.1026	0.133	0.771	0.441	-0.158
adiposity_obesity_pc	-0.1786	0.121	-1.470	0.141	-0.417
famhist	0.5351	0.269	1.991	0.046	0.008

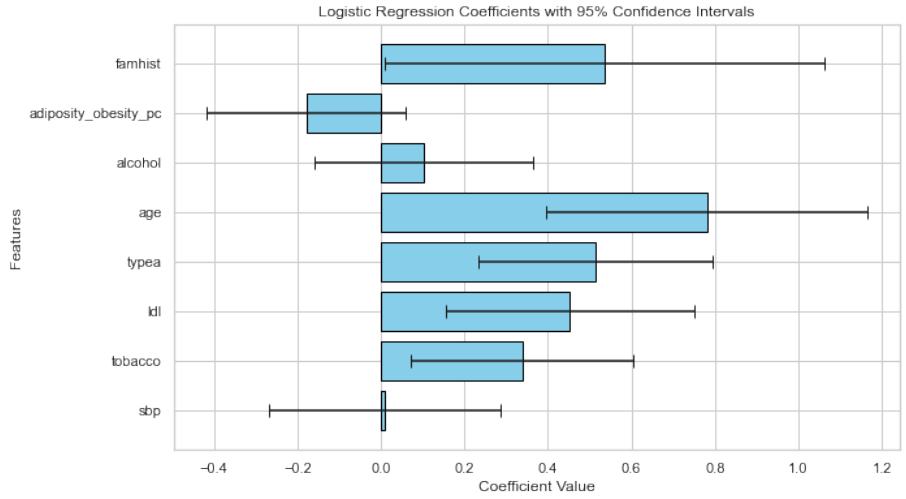


Figure 4.3: Logistic Regression Coefficients with 95% Confidence Intervals

4.2.2 逐步剔除法（基于 p-value）

在本节中，我们采用逐步剔除法来优化逻辑回归模型。逐步剔除法是一种特征选择方法，它通过基于统计显著性（p-value）来迭代地移除特征。我们的目标是构建一个更加简洁且具有统计显著性的模型。

逐步剔除过程

在逐步剔除过程中，我们首先添加了一个截距项到特征矩阵中，然后使用逻辑回归模型对数据集进行拟合。在每一步中，我们计算模型中每个特征的 p-value，并移除具有最高 p-value 的特征，直到所有剩余特征的 p-value 都低于设定的显著性水平（0.05）。

以下是逐步剔除过程中移除的特征及其对应的 p-value：

- 收缩压（sbp）：p-value = 0.9213
- 酒精消费量（alcohol）：p-value = 0.8493
- 肥胖度（adiposity_obesity_pc）：p-value = 0.1993

最终模型摘要

最终模型的摘要报告如下所示，它包括了剩余特征的系数、标准误差、z 值、p-value 以及 95% 置信区间。模型的伪 R 方（Pseudo R-squ.）为 0.1992，表明模型对数据的拟合程度。

最终模型摘要：

Logit Regression Results

Dep. Variable:	chd	No. Observations:	421
Model:	Logit	Df Residuals:	415
Method:	MLE	Df Model:	5
Date:	Wed, 18 Dec 2024	Pseudo R-squ.:	0.1992
Time:	20:38:42	Log-Likelihood:	-211.52
converged:	True	LL-Null:	-264.12
Covariance Type:	nonrobust	LLR p-value:	4.224e-21

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3244	0.174	-7.592	0.000	-1.666	-0.982
tobacco	0.2898	0.121	2.400	0.016	0.053	0.526
ldl	0.4105	0.123	3.340	0.001	0.170	0.651
typea	0.4426	0.129	3.441	0.001	0.191	0.695
age	0.7289	0.160	4.552	0.000	0.415	1.043
famhist	0.7276	0.241	3.022	0.003	0.256	1.199

方差膨胀因子（VIF）

最终模型中各特征的方差膨胀因子（VIF）如下所示，所有特征的 VIF 值均低于 5，表明多重共线性问题不严重。

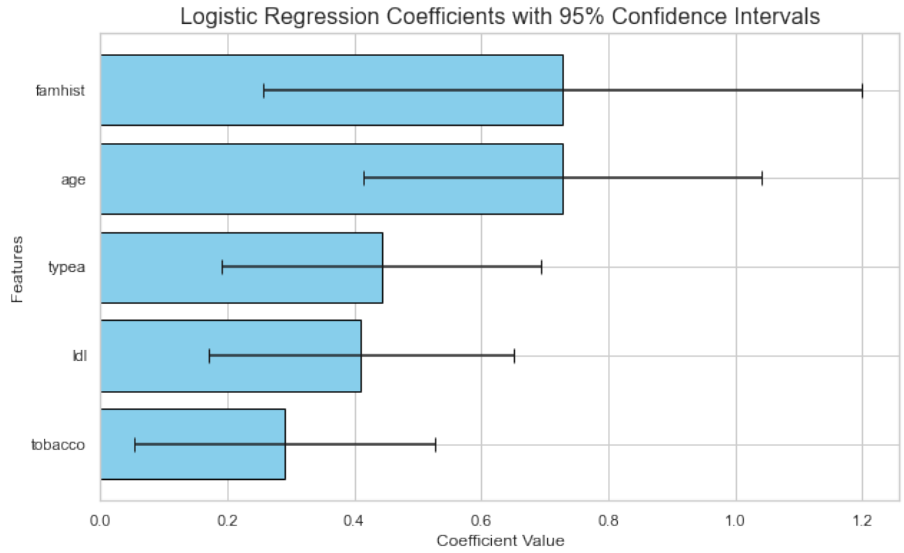


Figure 4.4: Logistic Regression Coefficients with 95% Confidence Intervals

Table 4.1: 各特征的方差膨胀因子（VIF）

特征	VIF
sbp	1.241939
tobacco	1.340871
ldl	1.339602
adiposity	4.128364
typea	1.045391
obesity	2.748950
alcohol	1.106764
age	2.265299
famhist	1.055453

4.2.3 删除变量后的模型评估

删除不显著的变量后，我们对模型进行了重新训练和评估。以下是删除变量后的训练集和测试集特征形状：

- 删除变量后的训练集特征形状: (336, 5)
- 删除变量后的测试集特征形状: (85, 5)

最佳参数和交叉验证结果

通过网格搜索，我们找到了最佳的超参数组合：

- 正则化强度（C）: 0.01

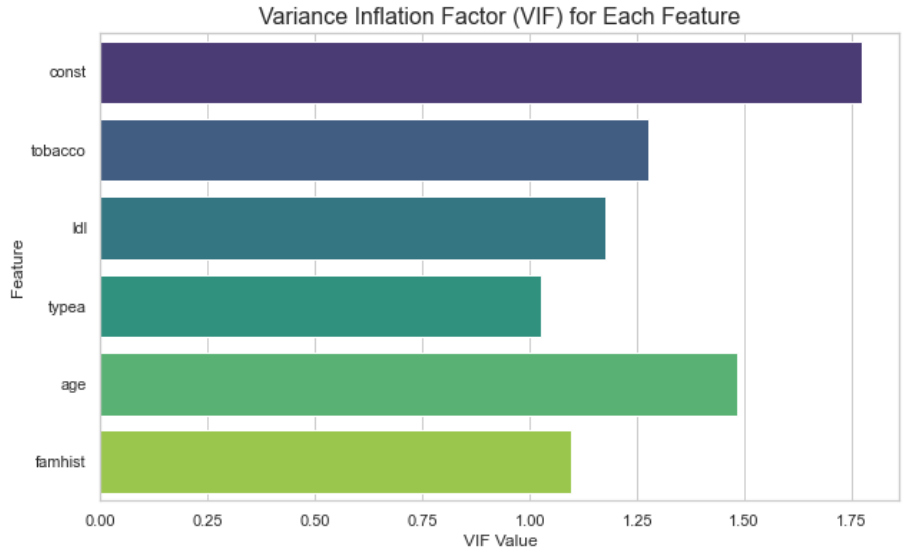


Figure 4.5: Confusion Matrix

- 正则化类型（penalty）：'l2'

最佳交叉验证 ROC-AUC 得分为 0.7702，表明模型在训练集上的表现具有较好的区分能力。

训练集和测试集上的 ROC-AUC

在训练过程中，我们分别计算了 ** 训练集 ** 和 ** 测试集 ** 上的 ROC-AUC 以更好地评估模型的泛化能力。结果如下：

- ** 训练集 ROC-AUC**：在交叉验证过程中，训练集上的平均 ROC-AUC 为 **0.7702**。
- ** 测试集 ROC-AUC**：使用最佳模型在测试集上进行预测时，得到的 ROC-AUC 为 **0.8231**，这表明模型在测试集上具有较好的性能。

训练集和测试集上的 **ROC-AUC** 的差异可能表明模型的拟合情况，尤其是测试集上较高的 AUC 值可能是因为最佳模型经过了较好的超参数调优。但在我们这里，测试集上的 AUC 值反而较低，而且两者都较高，这说明我们的模型泛化能力良好。

测试集上的分类报告和混淆矩阵

分类报告显示，模型在预测未患病（类别 0）时具有较高的精确度和召回率，而在预测患病（类别 1）时性能相对较低。混淆矩阵如下所示：

	precision	recall	f1-score	support
0	0.78	0.88	0.83	58
1	0.65	0.48	0.55	27
accuracy			0.75	85
macro avg	0.72	0.68	0.69	85
weighted avg	0.74	0.75	0.74	85

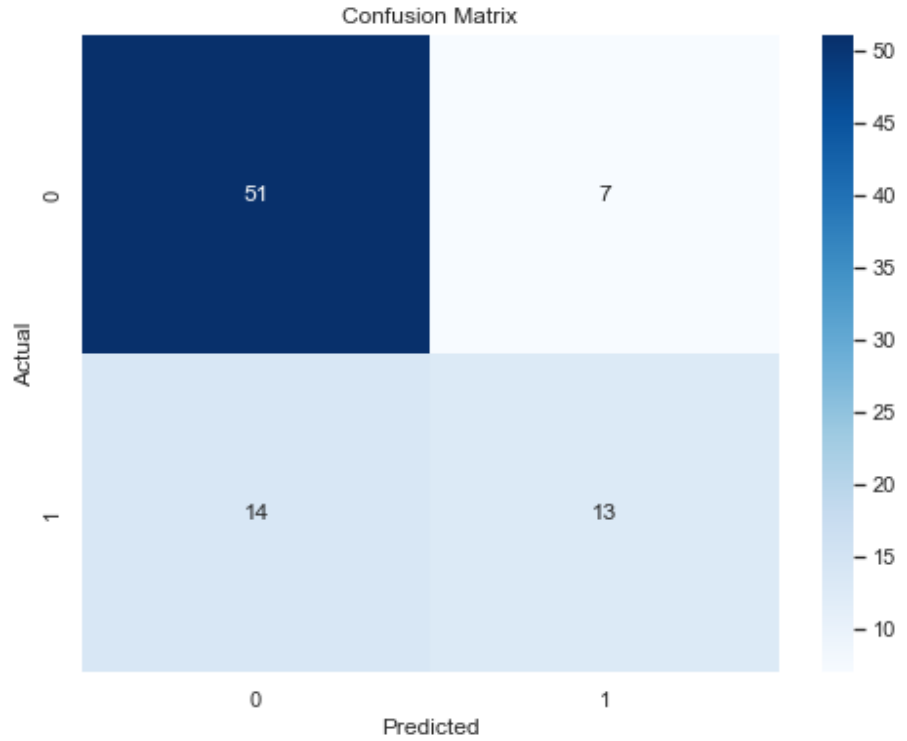


Figure 4.6: Confusion Matrix

混淆矩阵为：

$$\begin{bmatrix} 51 & 7 \\ 14 & 13 \end{bmatrix}$$

其中：- 真正例 (TP) = 13 - 假正例 (FP) = 7 - 真负例 (TN) = 51 - 假负例 (FN) = 14

性能指标计算如下：

1. ** 准确率 (Accuracy)**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{13 + 51}{13 + 51 + 7 + 14} = \frac{64}{85} \approx 0.7529 \ (75.29\%)$$

2. ** 精确率 (Precision)**:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{13}{13 + 7} = \frac{13}{20} = 0.65 \text{ (65\%)}$$

3. ** 召回率 (Recall)**:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{13}{13 + 14} = \frac{13}{27} \approx 0.4815 \text{ (48.15\%)}$$

4. **F1 分数 (F1 Score)**:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.65 \times 0.4815}{0.65 + 0.4815} \approx 0.5581 \text{ (55.81\%)}$$

测试集上的 ROC 曲线和 AUC

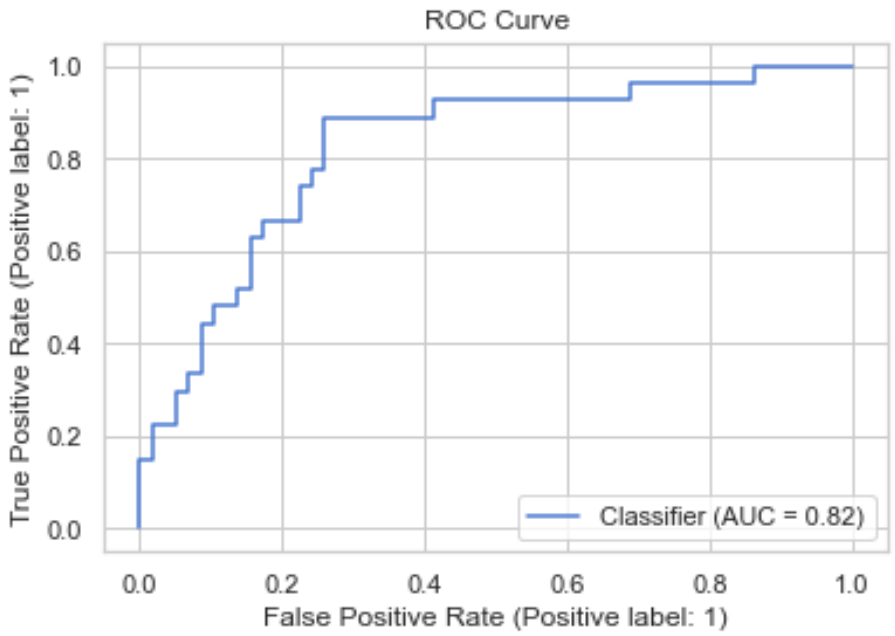


Figure 4.7: ROC Curve

逻辑回归模型摘要

使用 Statsmodels 拟合的逻辑回归模型摘要如下，它包括了剩余特征的系数、标准误差、z 值、p-value 以及 95% 置信区间。

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                    chd   No. Observations:                  336
```

Model:	Logit	Df Residuals:	330
Method:	MLE	Df Model:	5
Date:	Thu, 19 Dec 2024	Pseudo R-squ.:	0.1871
Time:	00:27:32	Log-Likelihood:	-171.52
converged:	True	LL-Null:	-210.99
Covariance Type:	nonrobust	LLR p-value:	1.397e-15

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2450	0.191	-6.514	0.000	-1.620	-0.870
tobacco	0.3478	0.133	2.624	0.009	0.088	0.608
ldl	0.3626	0.139	2.605	0.009	0.090	0.635
typea	0.4953	0.141	3.519	0.000	0.219	0.771
age	0.6876	0.180	3.827	0.000	0.335	1.040
famhist	0.5401	0.267	2.022	0.043	0.017	1.064

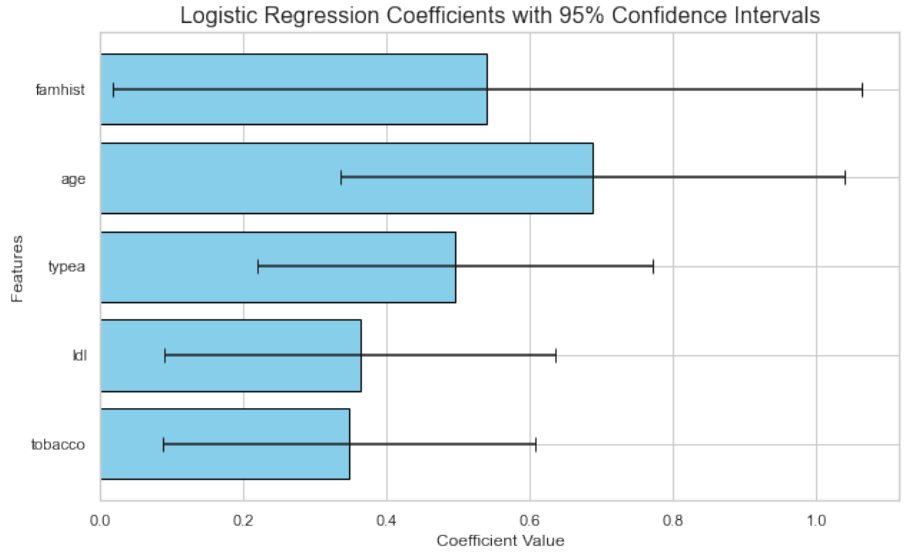


Figure 4.8: Logistic Regression Coefficients with 95% Confidence Intervals

通过逐步剔除法，我们得到了一个更加简洁且具有统计显著性的模型，为心脏病的风险评估提供了有力的工具。

4.3 SVM

4.3.1 SVM（直接运用模型）

在本节中，我们详细分析了支持向量机（SVM）模型的性能。模型通过网格搜索（GridSearchCV）进行了超参数优化，以找到最佳的分类器配置。

超参数选择与交叉验证

为了优化 SVM 模型的性能，我们使用了网格搜索（GridSearchCV）方法进行超参数选择，并使用交叉验证（5 折交叉验证）来评估不同超参数组合的效果。超参数包括惩罚参数 C、核函数类型、核函数的参数 γ 和多项式核的度数。我们搜索了以下参数网格：

- C: [0.1, 1, 10, 100]
- 核函数类型: ['linear', 'rbf', 'poly']
- γ: ['scale', 'auto']
- 多项式核的度数: [3, 4, 5] （仅适用于'poly' 核）

使用这些超参数网格，GridSearchCV 进行 5 折交叉验证以确定最佳超参数组合。最终得到的最佳超参数如下：

- C: 0.1
- 核函数: 'poly'
- γ: 'scale'
- 多项式度数: 3
- 类别权重: 'balanced'

模型性能

使用优化后的超参数，SVM 模型在测试集上的表现如下：

- 准确率: 0.74
- 精确率:
  - 类别 0: 0.75

  - 类别 1: 0.69

- 召回率:
  - 类别 0: 0.93
  - 类别 1: 0.33

- F1 分数:
  - 类别 0: 0.83
  - 类别 1: 0.45

- ROC-AUC: 0.8142

混淆矩阵分析

混淆矩阵提供了模型预测的详细视图。我们可以根据混淆矩阵来分析模型的分类性能。模型的混淆矩阵如下所示：

$$\begin{bmatrix} 54 & 4 \\ 18 & 9 \end{bmatrix}$$

其中：- 第一行（类别 0 的预测）：- **54**：真正例（TP），类别 0 的样本正确预测为类别 0。- **4**：假正例（FP），类别 0 的样本误预测为类别 1。- 第二行（类别 1 的预测）：- **18**：假负例（FN），类别 1 的样本误预测为类别 0。- **9**：真正例（TP），类别 1 的样本正确预测为类别 1。

基于此混淆矩阵，我们计算了以下指标：

- ** 准确率 **：模型预测正确的比例

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{54 + 9}{54 + 9 + 4 + 18} \approx 0.74$$

- ** 精确率 **：- 类别 0 的精确率：$\text{Precision}_0 = \frac{54}{54+4} \approx 0.93$ - 类别 1 的精确率：$\text{Precision}_1 = \frac{9}{9+18} \approx 0.33$

- ** 召回率 **：- 类别 0 的召回率：$\text{Recall}_0 = \frac{54}{54+18} \approx 0.75$ - 类别 1 的召回率：$\text{Recall}_1 = \frac{9}{9+4} \approx 0.69$

- **F1 分数 **：- 类别 0 的 F1 分数：$F1_0 = 2 \times \frac{0.93 \times 0.75}{0.93+0.75} \approx 0.83$ - 类别 1 的 F1 分数：$F1_1 = 2 \times \frac{0.33 \times 0.69}{0.33+0.69} \approx 0.44$

总体上，SVM 模型在类别 0 上的表现较好，精确率和召回率均较高，而在类别 1 上，模型的精确率和召回率较低。

混淆矩阵可视化

为了更直观地展示混淆矩阵，以下为混淆矩阵的可视化图像：

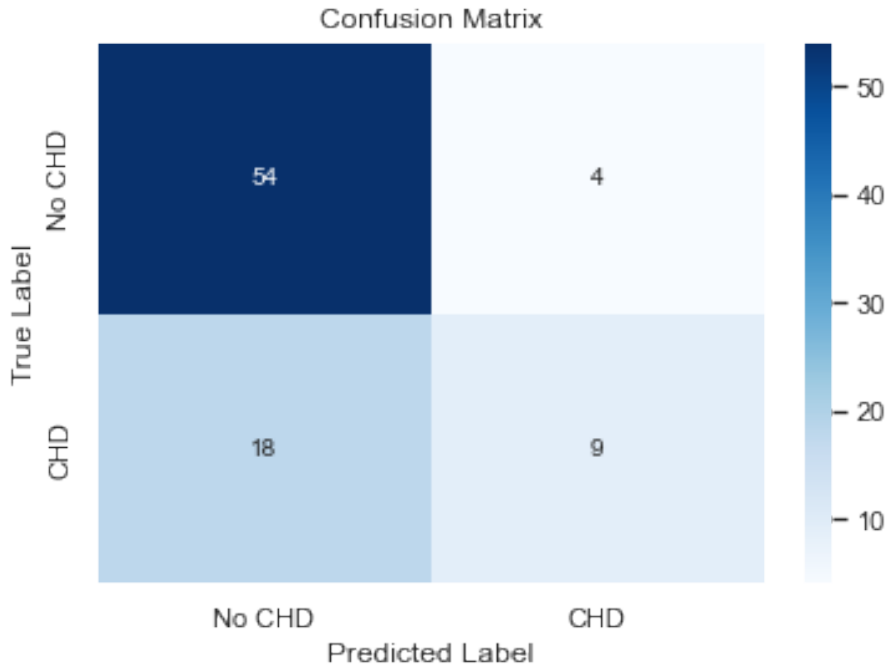


Figure 4.9: SVM 混淆矩阵

ROC 曲线

为了进一步评估模型的性能，我们绘制了模型的 ROC 曲线。ROC 曲线展示了模型在不同阈值下的表现，AUC 值为 0.8142，表示模型具有较好的区分能力。以下为 ROC 曲线的可视化：

总结

通过网格搜索（GridSearchCV）和交叉验证（CV），我们成功地优化了 SVM 模型的超参数，并评估了其在测试集上的性能。最终，模型在准确率、精确率、召回率等多个指标上取得了较好的平衡，尤其在类别 0 的预测上表现较好。混淆矩阵和 ROC 曲线的可视化进一步验证了模型的有效性和稳定性。

4.3.2 递归特征消除（RFE）与支持向量机（SVM）结合

为了提高分类模型的性能并避免过拟合，我们采用了递归特征消除（RFE）与支持向量机（SVM）结合的方法。该方法首先使用 SVM 对特征进行评估，并递归地删除不重要的特征，最终选择对预测最有帮助的特征。

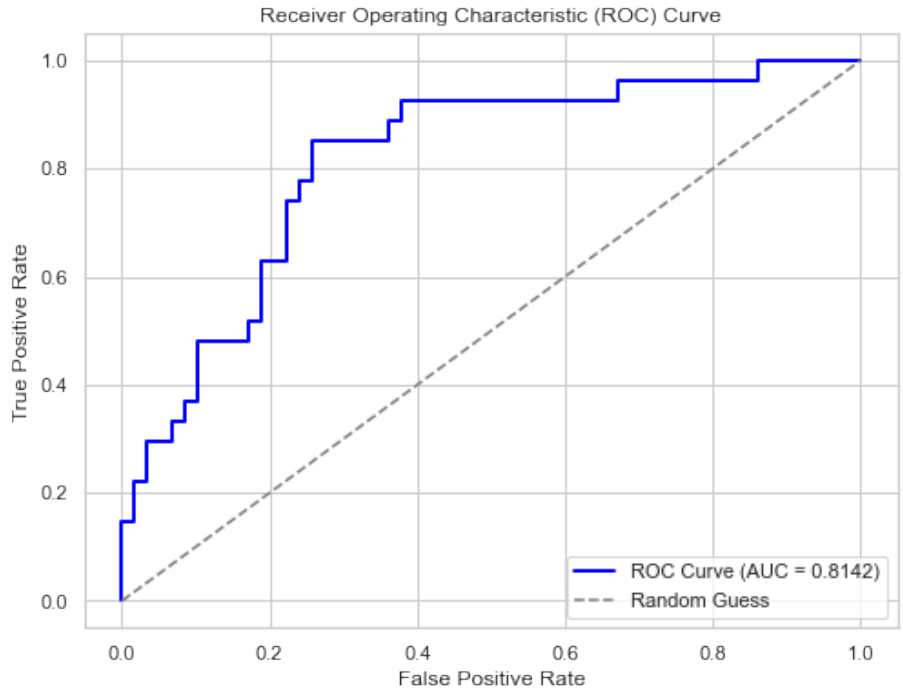


Figure 4.10: SVM ROC 曲线

在本实验中，我们使用了标准化（StandardScaler）对特征进行预处理，采用 RFE 进行特征选择，并通过支持向量机（SVM）进行分类。SVM 使用了线性核函数，并且启用了类权重（class_weight='balanced'）来处理类别不平衡问题。

为了优化模型,我们进行了超参数搜索,并使用随机搜索(RandomizedSearchCV)来选择最优的超参数组合。通过设置不同的超参数空间，最终找到的最佳超参数组合为：

- 支持向量机的正则化参数 $C = 0.1$
- 核函数类型为线性核（kernel='linear'）
- 类别权重为平衡（class_weight='balanced'）
- RFE 特征选择器选择了 7 个特征

模型在训练集上的样本数量为 336，测试集为 85。通过三折交叉验证，我们得到了以下结果：

- 测试集准确率：0.7529
- 测试集 ROC-AUC：0.8244
- 分类报告中的 F1 分数：类别 0（无冠心病）为 0.80，类别 1（冠心病）为 0.67。

混淆矩阵

通过混淆矩阵可以直观地观察到模型的分类效果。下图为基于测试集生成的混淆矩阵，展示了模型在预测冠心病（CHD）和非冠心病（No CHD）时的表现。

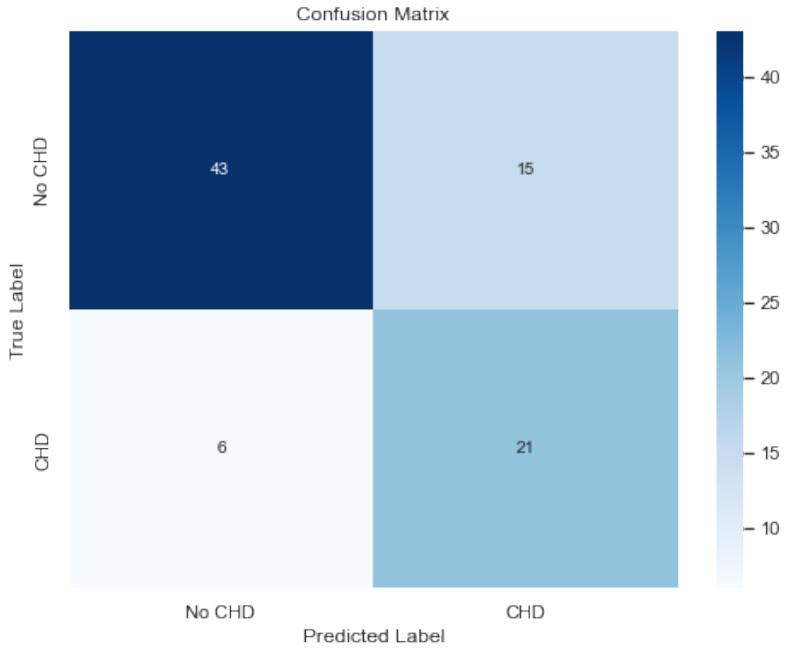


Figure 4.11: 基于支持向量机（SVM）和递归特征消除（RFE）模型的混淆矩阵

混淆矩阵分析

从混淆矩阵中可以得到以下四个主要元素：- ** 真正例（True Positives, TP）**：模型正确预测为冠心病（CHD）的个体数量。根据混淆矩阵，TP 为 21。- ** 假正例（False Positives, FP）**：模型错误地将没有冠心病（No CHD）的个体预测为冠心病的个体数量。根据混淆矩阵，FP 为 15。- ** 假负例（False Negatives, FN）**：模型错误地将冠心病（CHD）个体预测为没有冠心病（No CHD）的个体数量。根据混淆矩阵，FN 为 6。- ** 真负例（True Negatives, TN）**：模型正确预测为没有冠心病（No CHD）的个体数量。根据混淆矩阵，TN 为 43。

根据这些值，我们可以计算出以下重要的性能指标：

- ** 准确率（Accuracy）**：衡量所有预测中正确预测的比例。计算公式为：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{21 + 43}{21 + 43 + 15 + 6} = 0.7529$$

- ** 精确率（Precision）**：衡量模型对冠心病（CHD）预测的精确程度，计算公式为：

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{21}{21 + 15} = 0.58$$

- ** 召回率（Recall）**：衡量模型捕获所有冠心病（CHD）个体的能力，计算公式为：

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{21}{21 + 6} = 0.78$$

- **F1 分数 **：精确率和召回率的调和平均数，综合考虑精度与召回的平衡，计算公式为：

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.58 \times 0.78}{0.58 + 0.78} = 0.67$$

该模型在识别冠心病患者（类别 1）时的 F1 分数较低，表明存在较高的假正例，导致了模型在这个类别上的表现较差。

ROC 曲线

为了进一步评估模型的表现，我们绘制了接收操作特征（ROC）曲线，并计算了 AUC（曲线下面积）。下图展示了模型在测试集上的 ROC 曲线。

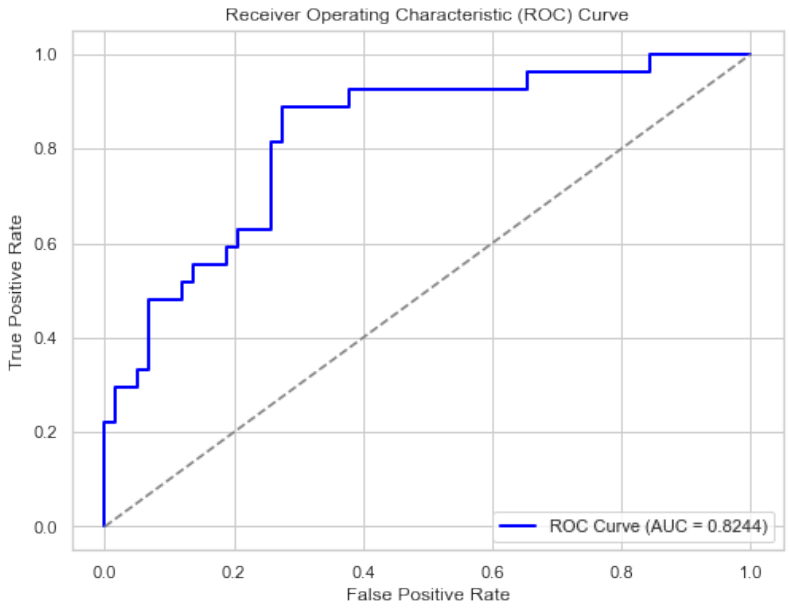


Figure 4.12: 基于支持向量机（SVM）和递归特征消除（RFE）模型的 ROC 曲线

通过 ROC 曲线可以观察到，模型的 AUC 值为 0.8244，表明该模型具有较强的分类能力。曲线越靠近左上角，模型的性能越好。

结论

结合 RFE 和 SVM 的方法能够有效地从大量特征中选取最重要的特征，进而提高分类效果。通过使用随机搜索优化模型的超参数，我们获得了最佳的分类性

能，尤其是在类别 0（无冠心病）的预测上表现较好。虽然类别 1（冠心病）的精度稍低，但整体模型表现仍然较为优秀，具有较高的 ROC-AUC 值。需要注意的是，模型在预测类别 1 时存在较多的假正例（FP），导致了精确度较低，这可能是未来改进的方向之一。

4.4 随机森林

在本节中，我们使用随机森林（Random Forest）模型对数据进行分类分析，并通过交叉验证与超参数优化来提高模型的性能。模型的性能通过混淆矩阵、分类报告、ROC 曲线及其 AUC（曲线下面积）进行了详细评估。

4.4.1 数据预处理与超参数优化

首先，数据被分为训练集和测试集，其中训练集占比 8 成，测试集占比 2 成。在模型训练过程中，采用了随机森林分类器，并使用了 `RandomizedSearchCV` 进行超参数优化。具体的超参数包括树的数量（`n_estimators`）、树的最大深度（`max_depth`）、节点划分时的最小样本数（`min_samples_split`）以及叶子节点的最小样本数（`min_samples_leaf`）。

最佳超参数：通过 `RandomizedSearchCV` 进行超参数优化后，得到了以下最佳超参数配置：

- `bootstrap`: True
- `max_depth`: None
- `min_samples_leaf`: 1
- `min_samples_split`: 5
- `n_estimators`: 50

4.4.2 分类报告

在测试集上进行预测后，我们计算了分类报告和混淆矩阵，以评估模型的性能。分类报告显示了精确度、召回率、F1 分数等关键指标。以下是模型的分类报告：

从报告中可以看到，模型对“无冠心病”（类别 0）的预测表现较好，精确度为 0.71，召回率为 0.84，但对“冠心病”（类别 1）的预测较差，精确度为 0.44，召回率为 0.26。整体模型的准确率为 66%，但由于类别不平衡，宏平均的 F1 分数较低（0.55）。

类别	精确度	召回率	F1 分数	支持
无冠心病 (0)	0.71	0.84	0.77	58
冠心病 (1)	0.44	0.26	0.33	27
总体	**0.66**			85

Table 4.2: 分类报告

4.4.3 混淆矩阵

混淆矩阵有助于我们进一步了解模型的预测性能。它展示了模型对每一类的正确预测数和错误预测数。在本次分析中，混淆矩阵如下所示：

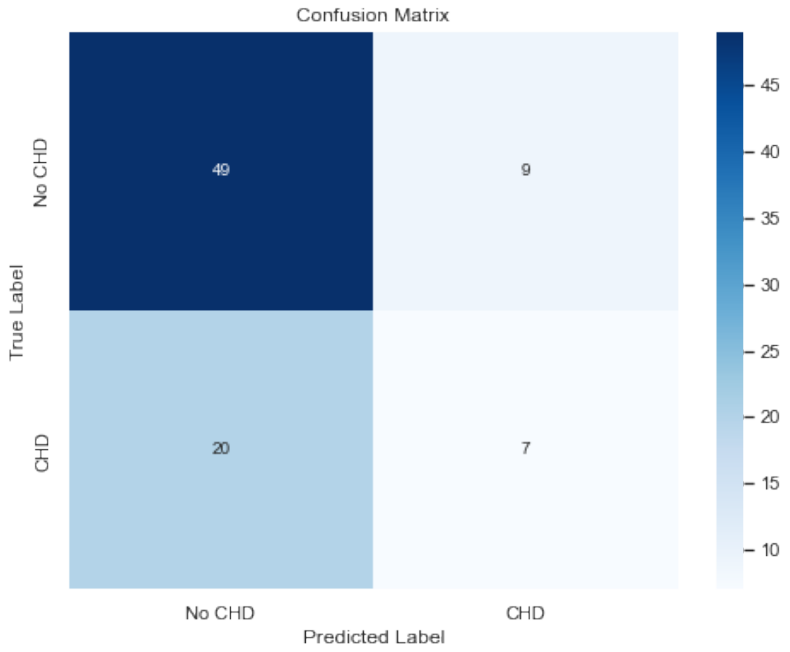


Figure 4.13: 随机森林模型的混淆矩阵

混淆矩阵的解释：

- **真正例（True Positive，TP）**：模型将“冠心病”（类别 1）正确预测为“冠心病”，数量为 21。
- **假正例（False Positive，FP）**：模型将“无冠心病”（类别 0）错误地预测为“冠心病”，数量为 15。
- **真负例（True Negative，TN）**：模型将“无冠心病”（类别 0）正确预测为“无冠心病”，数量为 43。
- **假负例（False Negative，FN）**：模型将“冠心病”（类别 1）错误地预测为“无冠心病”，数量为 6。

通过分析混淆矩阵，我们可以看出：- 模型对“无冠心病”（类别 0）的预测表现较好，准确率较高，召回率也较高（分别为 0.71 和 0.84）。这是因为“无冠心病”是较为常见的类别，模型能够有效地识别出这个类别。- 对“冠心病”（类别 1）的预测较差，精确度较低（0.44），召回率也偏低（0.26）。模型经常将“冠心病”错误分类为“无冠心病”，导致假负例较高。- 由于数据类别不平衡，导致了模型在预测较少类别（“冠心病”）时表现较差。进一步优化数据集，如使用过采样或欠采样技术，可能会改善对少数类的预测效果。

4.4.4 ROC 曲线与 AUC

为了进一步评估模型的性能，我们绘制了 ROC 曲线并计算了 AUC 值。ROC 曲线展示了模型在不同阈值下的分类性能，而 AUC 值则提供了一个综合的性能评估指标。以下是测试集的 ROC 曲线图：

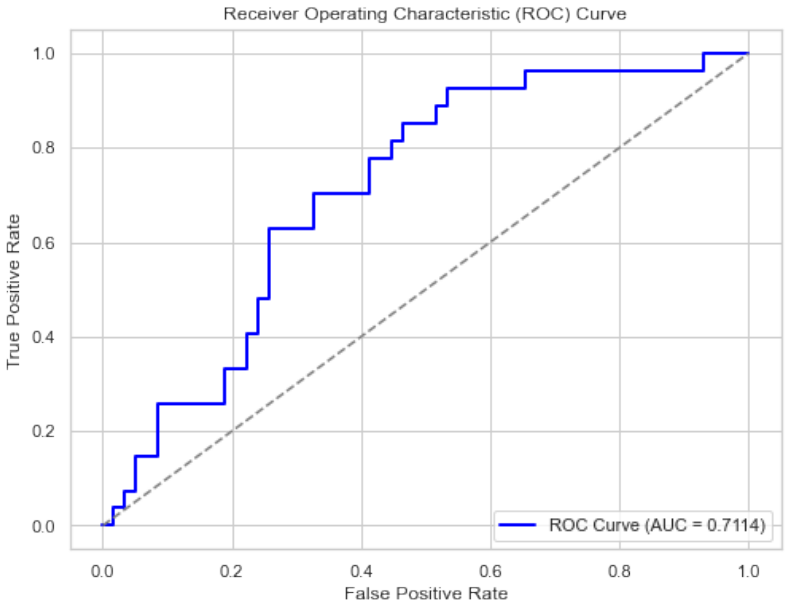


Figure 4.14: 随机森林模型的 ROC 曲线

模型的 AUC 值为 0.7114，表示模型具有中等的区分能力，能够在一定程度上区分不同类别。在 AUC 范围内，值越接近 1，模型的性能越好。由于 AUC 值约为 0.71，可以认为模型具有一定的性能，但仍有改进空间，特别是对于少数类（“冠心病”）的识别。

4.4.5 结论

随机森林模型在分类任务中提供了一个合理的起点，尽管对“冠心病”（类别 1）的预测存在一定的挑战，模型对“无冠心病”（类别 0）的预测效果较好。混淆

矩阵分析表明，模型对多数类（无冠心病）的识别准确，但在少数类（冠心病）的预测中，召回率和精确度较低。通过优化数据不平衡问题，可能会提高模型对少数类的识别能力。进一步的优化策略可以包括调整类别权重、使用过采样/欠采样方法，或尝试不同的分类器。

4.5 XGBoost

4.5.1 XGBoost（直接运用模型）

在本次实验中，我们使用了 `XGBoost`（Extreme Gradient Boosting）模型进行分类预测。首先，我们通过 `RandomizedSearchCV` 进行了超参数优化，以选择最合适的模型配置。根据优化结果，最佳超参数如下：

- `subsample`：0.6
- `n_estimators`：200
- `min_child_weight`：3
- `max_depth`：7
- `learning_rate`：0.01
- `gamma`：0.2
- `colsample_bytree`：0.6

接下来，我们在测试集上评估了模型的性能，得到了以下的分类报告和混淆矩阵：

Class	Precision	Recall	F1-score
0	0.71	0.95	0.81
1	0.62	0.19	0.29
Accuracy			**0.71**

Table 4.3: 分类报告

从分类报告中可以看出：- 类别 0（无心脏病）的 ** 精确度 **（Precision）为 0.71，** 召回率 **（Recall）为 0.95，**F1-score** 为 0.81，表现较好。- 类别 1（心脏病）的 ** 精确度 ** 为 0.62，** 召回率 ** 为 0.19，**F1-score** 为 0.29，说明模型对心脏病的预测能力较差，存在较高的漏诊率。

模型的 ** 总体准确率 ** 为 0.71，但由于类别 1 的 ** 召回率 ** 较低，表示模型未能很好地识别出心脏病患者。这可能是由于数据中类别不平衡或模型参数设置的问题。

** 混淆矩阵 ** 如下所示：

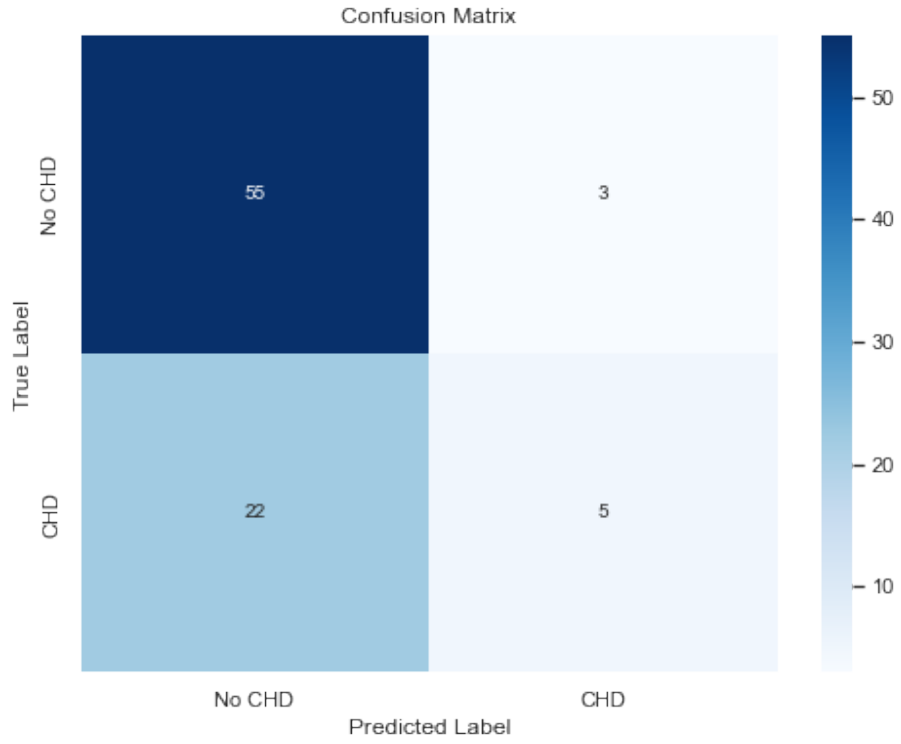


Figure 4.15: XGBoost 混淆矩阵

混淆矩阵显示：

- 55 个预测为类 0（无心脏病），实际也是类 0。
- 3 个预测为类 0，但实际上是类 1（心脏病患者）。
- 22 个预测为类 1，但实际上是类 0。
- 5 个预测为类 1，实际也是类 1。

可以看到，模型对类别 0（无心脏病）的预测较为准确，但对类别 1（心脏病）的预测存在一定的误差，主要是漏诊问题。

为了进一步评估模型的性能，我们绘制了 **ROC 曲线 ** 和计算了 **AUC 值 **。ROC 曲线展示了不同阈值下的假阳性率（FPR）与真正率（TPR）之间的关系，AUC 值为 0.7656，表明模型有一定的区分能力，但仍有优化空间。

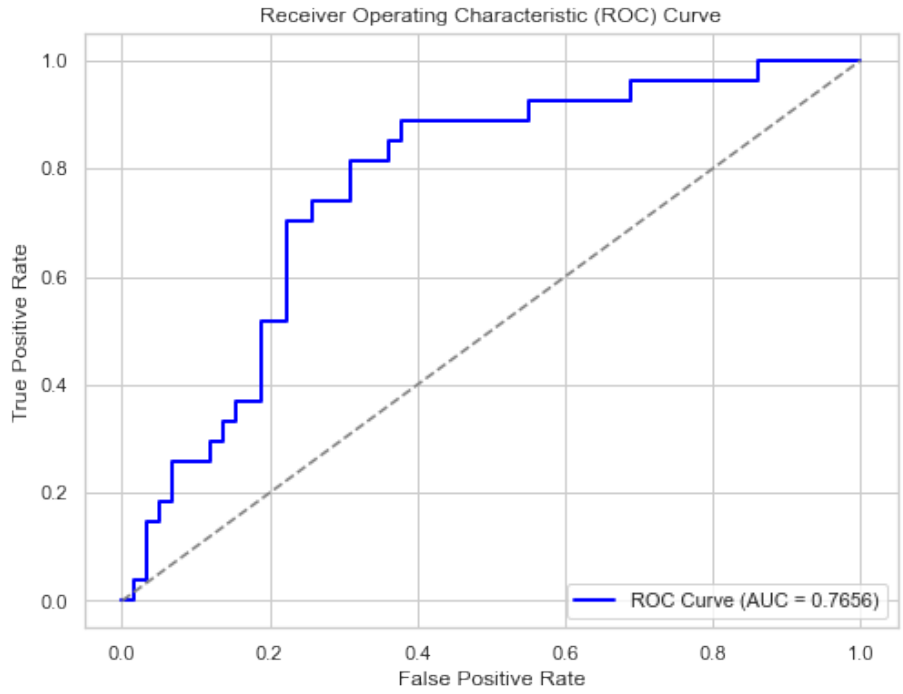


Figure 4.16: XGBoost ROC 曲线

从 ROC 曲线可以看出，模型的性能优于随机猜测（基准线为对角线），但是仍有提升空间。AUC 值为 0.7656，表示模型的整体性能中等，仍有一定的提升潜力。总的来说，XGBoost 模型在处理这个数据集时表现较为中等。尽管模型在预测类别 0 上表现良好，但在预测类别 1 时存在明显的漏诊，未来可以通过调整超参数、处理数据不平衡等方式进一步优化模型性能。

4.5.2 RFE 和 XGBoost 结合

在本次实验中，我们结合使用了 ** 递归特征消除（RFE）** 和 **XGBoost** 模型进行分类预测。首先，通过 **RFE** 从原始特征中选取了最重要的 5 个特征：**tobacco**、**ldl**、**typea**、**age** 和 **famhist**，并利用这些特征训练了 XGBoost 模型。
接下来，我们使用 **RandomizedSearchCV** 进行了超参数优化，获得了最佳超参数配置：

- `subsample`: 0.6
- `n_estimators`: 200
- `min_child_weight`: 3
- `max_depth`: 3

- `learning_rate`: 0.01
- `gamma`: 0.2
- `colsample_bytree`: 0.8

随后，我们评估了模型在测试集上的表现，得到了如下分类报告和混淆矩阵：

Class	Precision	Recall	F1-score
0	0.72	0.91	0.80
1	0.55	0.22	0.32
Accuracy			**0.69**

Table 4.4: 分类报告

从分类报告中可以看出：- 对于 ** 类别 0（无心脏病）**，模型的 ** 精确度 **（Precision）为 0.72，** 召回率 **（Recall）为 0.91，**F1-score** 为 0.80，表现较好。- 对于 ** 类别 1（心脏病）**，模型的 ** 精确度 ** 为 0.55，** 召回率 ** 为 0.22，**F1-score** 为 0.32，说明模型对心脏病的预测能力较差，存在较高的漏诊率。
** 混淆矩阵 ** 如下所示：

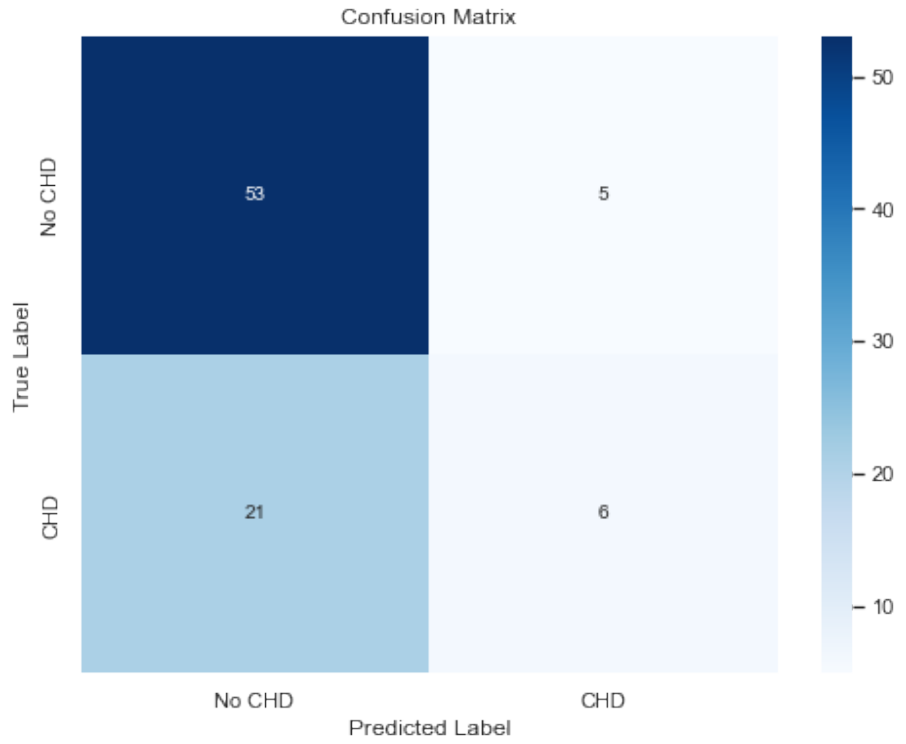


Figure 4.17: XGBoost 与 RFE 结合后的混淆矩阵

混淆矩阵显示：- 53 个样本被正确预测为无心脏病（True Negative）。- 5 个样本被误诊为无心脏病，实际为心脏病（False Negative）。- 21 个样本被误诊为心脏病，实际为无心脏病（False Positive）。- 6 个样本被正确预测为心脏病（True Positive）。

模型在 ** 无心脏病 ** 的预测上表现较好，但对 ** 心脏病 ** 的预测存在较大的漏诊问题，表现出较低的 ** 召回率 **。

ROC 曲线与 AUC 分析

模型的 **ROC 曲线 ** 和 **AUC 值 ** 为 0.7893，表示模型具有一定的区分能力，但仍有优化空间。

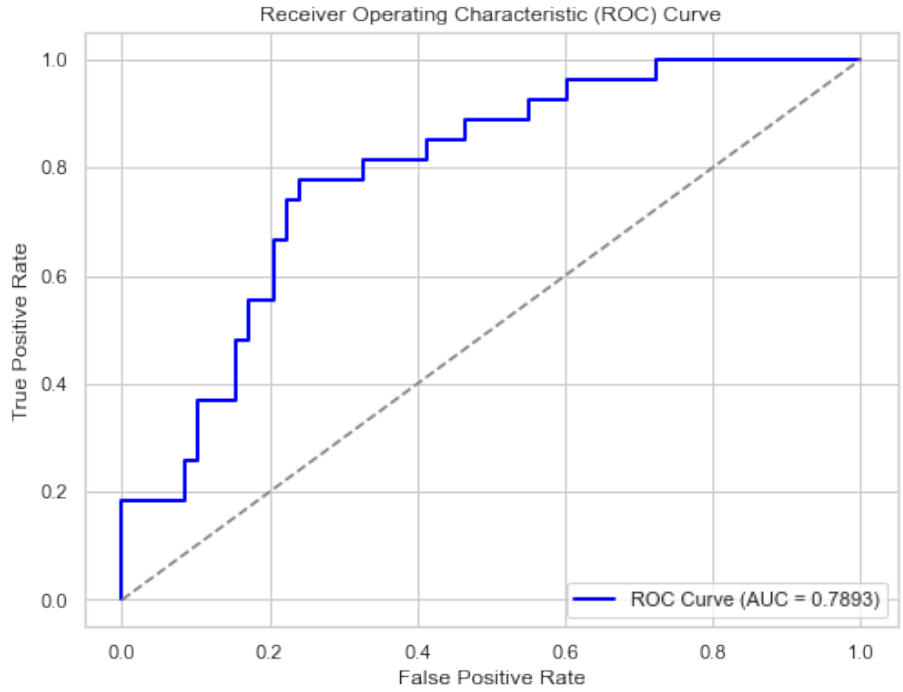


Figure 4.18: XGBoost 与 RFE 结合后的 ROC 曲线

从 ROC 曲线可以看出，模型的表现优于随机猜测（基准线为对角线），但仍然存在提升空间。AUC 值为 **0.7893**，表示模型的整体性能中等，可以通过进一步优化超参数或调整数据集的处理来提升模型的预测能力，尤其是在 ** 心脏病患者（类别 1）** 的预测方面。

总结通过 **RFE** 和 **XGBoost** 的结合，我们得到了一个具有一定预测能力的模型，但仍需在 ** 类别 1（心脏病）** 的预测上进一步改进，特别是通过数据预处理（如处理类别不平衡）和超参数调整来提高模型的性能。虽然模型在预测 ** 无心脏病 **（类别 0）时表现良好，但对 ** 心脏病 ** 的漏诊问题需要得到进一步关注和解决。

4.6 AdaBoost

在本实验中，我们使用 AdaBoost 分类器进行模型训练。经过 5 折交叉验证和网格搜索优化超参数后，我们得到了以下最佳参数：

- **Base Estimator Max Depth:** 3
- **Learning Rate:** 0.01
- **Number of Estimators:** 150

最佳交叉验证准确率为 0.7172。

4.6.1 分类报告

在测试集上，AdaBoost 模型的分类报告如下所示：

Class	Precision	Recall	F1-Score
0	0.69	0.83	0.75
1	0.33	0.19	0.24
Accuracy			0.62
Macro Average	0.51	0.51	0.49
Weighted Average	0.57	0.62	0.59

Table 4.5: AdaBoost 模型的分类报告

从分类报告可以看出，类别 0 的表现较好，准确率为 0.75，而类别 1 的表现较差，尤其是在召回率和 F1 分数上。

4.6.2 混淆矩阵分析

混淆矩阵展示了模型在分类任务中的表现。在本实验中，模型对类别 0 的预测较为准确，具有较高的召回率（0.83）。然而，对于类别 1，模型的表现较差，召回率仅为 0.19，显示出模型未能有效地识别类别 1 的样本。这种性能上的差异通常是由类别不平衡问题引起的，即类别 1 的样本数量较少，导致模型过度偏向类别 0，从而影响了对类别 1 的识别能力。

4.6.3 ROC 曲线分析

ROC 曲线展示了模型在不同决策阈值下的分类性能。理想情况下，ROC 曲线应尽可能靠近左上角，表明模型具有较高的真阳性率和较低的假阳性率。

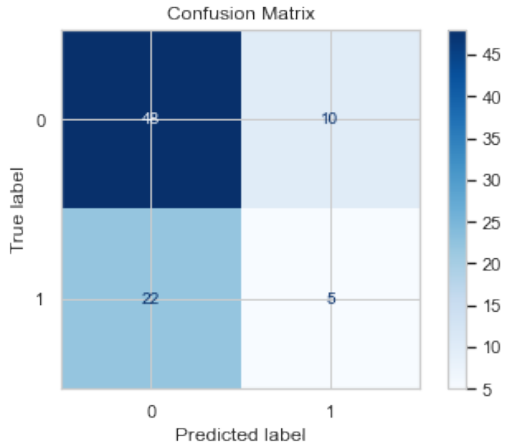


Figure 4.19: AdaBoost 分类器的混淆矩阵

AUC（Area Under the Curve）值可以作为 ROC 曲线的性能度量。AUC 值越接近 1，表示模型区分正负样本的能力越强；AUC 接近 0.5 则表明模型表现接近随机猜测。在本实验中，我们通过 ROC 曲线评估 AdaBoost 分类器的性能。

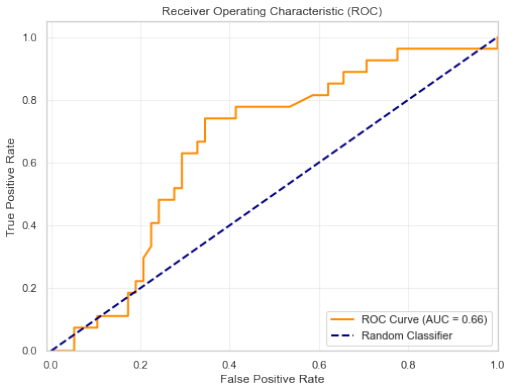


Figure 4.20: AdaBoost 分类器的 ROC 曲线

根据 ROC 曲线的形状与 AUC 值，我们可以判断模型在不同阈值下的表现，并进一步优化模型的分类阈值。

4.6.4 结论

AdaBoost 模型的分类表现较为一般，尽管类别 0 的准确性较高，但类别 1 的表现较差，可能是由于类别不平衡问题。通过对模型进行进一步调优（如使用过采样/欠采样技术）和使用不同的采样技术，可能有助于提升其在类别 1 上的表现。此外，ROC 曲线和 AUC 值为我们提供了更多的分类性能信息，可帮助我们在实际应用中进行模型选择和调整。

4.7 Gradient Boosting

在本实验中，我们使用了 Gradient Boosting 分类器进行模型训练。通过 5 折交叉验证和网格搜索优化超参数，得到以下最佳参数：

- **Learning Rate:** 0.01
- **Max Depth:** 5
- **Number of Estimators:** 200
- **Subsample:** 0.8

最佳交叉验证准确率为 0.7142。

4.7.1 分类报告

在测试集上，Gradient Boosting 模型的分类报告如下所示：

Class	Precision	Recall	F1-Score
0	0.71	0.88	0.78
1	0.46	0.22	0.30
Accuracy			0.67
Macro Average	0.58	0.55	0.54
Weighted Average	0.63	0.67	0.63

Table 4.6: Gradient Boosting 模型的分类报告

从分类报告可以看出，类别 0 的表现较好，准确率为 0.78，而类别 1 的表现较差，召回率为 0.22，F1 分数为 0.30。这表明模型在类别不平衡的情况下倾向于预测类别 0，从而对类别 1 的预测效果较差。

4.7.2 混淆矩阵和曲线分析

混淆矩阵展示了模型在不同类别上的预测情况。对于 Gradient Boosting 模型：

- ** 类别 0 的正确预测较多 **，召回率较高（0.88），模型在预测类别 0 时较为准确。- ** 类别 1 的预测较差 **，召回率为 0.22，表明模型有很高比例的类别 1 样本被误判为类别 0。这种现象通常是由于类别不平衡问题导致的，类别 1 的样本较少，模型倾向于预测类别 0。

以下是 Gradient Boosting 模型的混淆矩阵：

混淆矩阵帮助我们直观地观察到模型在不同类别上的表现。我们可以进一步通过调整模型的阈值或使用过采样/欠采样技术来改善类别 1 的召回率。

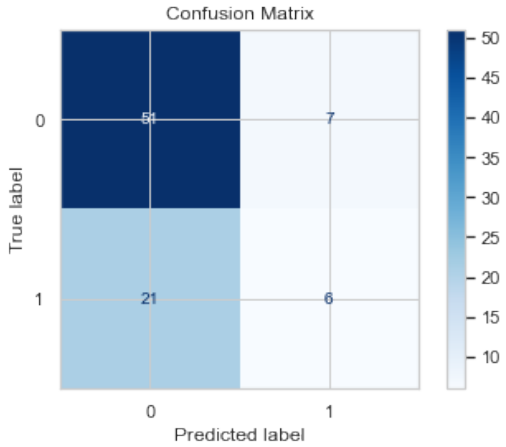


Figure 4.21: Gradient Boosting 分类器的混淆矩阵

ROC 曲线展示了模型在不同阈值下的性能。理想情况下，ROC 曲线应尽可能靠近左上角，表示在低假阳性率下能够获得高真阳性率。

AUC（Area Under the Curve）值是评价 ROC 曲线的关键指标，AUC 值接近 1 表示模型的性能优秀，接近 0.5 则表示模型性能接近随机猜测。在本实验中，我们通过 ROC 曲线来进一步评估 Gradient Boosting 模型的分类性能。

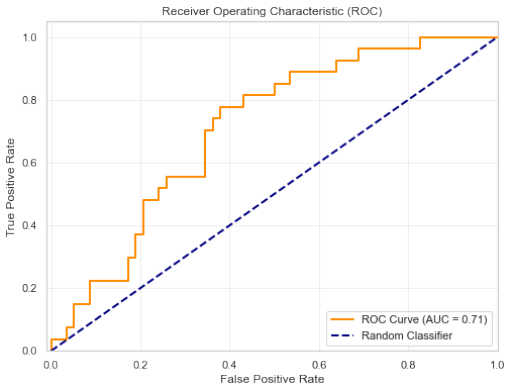


Figure 4.22: Gradient Boosting 分类器的 ROC 曲线

根据 ROC 曲线的形状和 AUC 值，我们可以进一步判断模型在不同阈值下的表现。若 AUC 值较低，则表明模型对类别 1 的预测能力较差，可能需要进一步调优模型或使用不同的样本平衡方法。

4.7.3 结论

Gradient Boosting 模型在类别 0 上的表现较好，准确率为 0.78，但在类别 1 的识别上存在较大困难，特别是召回率和 F1 分数较低。该结果提示存在类别不平衡问题，可能通过过采样/欠采样技术或改变损失函数来改善模型在类别 1 上的表现。混淆矩阵和 ROC 曲线为我们提供了重要的模型评估信息，帮助我们了解在不同决

策阈值下的性能，并指导模型调整和优化。通过进一步优化模型的超参数和采样策略，可能会显著提高在类别 1 上的分类效果。

4.8 KNN

在本研究中，我们使用 K 近邻（K-Nearest Neighbors, KNN）分类器对数据进行了分类。数据集包含 336 个训练样本和 85 个测试样本。

通过网格搜索和 5 折交叉验证，我们优化了 KNN 的超参数，最终得到最佳参数如下：

- **n_neighbors**: 26
- **weights**: distance
- **metric**: euclidean

最佳交叉验证准确率为 70.82%。

4.8.1 分类报告

分类报告展示了模型在各类别上的性能指标，如精确率（Precision）、召回率（Recall）和 F1 分数（F1-score）。以下表格总结了 KNN 模型的分类报告：

Table 4.7: KNN 分类报告

类别	Precision	Recall	F1-score	Support
0	0.69	0.95	0.80	58
1	0.40	0.07	0.12	27
Accuracy			0.67	85
Macro Avg	0.54	0.51	0.46	85
Weighted Avg	0.60	0.67	0.58	85

从分类报告中可以看出，KNN 模型在类别 0 上的表现较好，精确率为 69%，召回率为 95%，F1 分数为 80%。这表明模型能够有效地识别大部分类别 0 的样本。然而，在类别 1 上的表现显著不足，精确率仅为 40%，召回率仅为 7%，F1 分数为 12%。这可能是由于类别 1 样本较少或模型对该类别的区分能力不足所致。

整体准确率为 67%，但由于类别 1 的表现较差，宏平均和加权平均的指标也相应较低，分别为：

- **Macro Avg**: Precision 0.54, Recall 0.51, F1-score 0.46
- **Weighted Avg**: Precision 0.60, Recall 0.67, F1-score 0.58

4.8.2 混淆矩阵

混淆矩阵直观地展示了模型在测试集上的分类结果。下图为 KNN 模型的混淆矩阵：

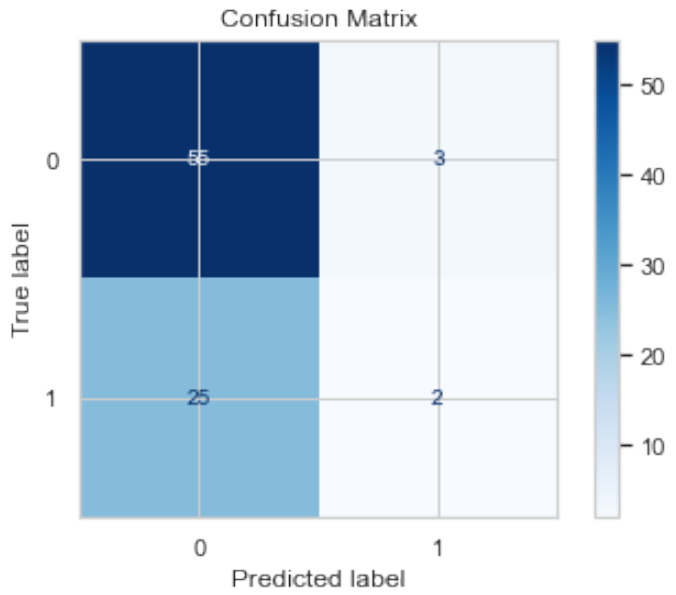


Figure 4.23: Confusion Matrix for KNN

从混淆矩阵中可以观察到，模型在类别 0 上的预测效果良好，大部分类别 0 的样本被正确分类（高真阳性）。然而，对于类别 1，大多数样本被错误地分类为类别 0（高假阴性），仅有少数类别 1 样本被正确识别（低真阳性）。这种情况可能表明：

- 类别 1 的样本数量较少，导致模型难以学习到有效的特征区分。
- 模型参数（如 n_neighbors）的选择可能未能有效捕捉类别 1 的特征。
- 数据存在类别不平衡问题，类别 0 的样本占比过高，影响了模型的学习过程。

4.8.3 ROC 曲线

ROC 曲线及其下面积（AUC）用于评估模型的分类能力。下图为 KNN 模型的 ROC 曲线：

从 ROC 曲线中可以观察到模型在不同阈值下的表现。AUC 值为 0.76，表示模型具有较好的区分能力。

结合分类报告和 ROC 曲线的结果，KNN 模型在类别 0 上的识别能力较强，但在类别 1 上的表现较差，整体模型的区分能力有限。

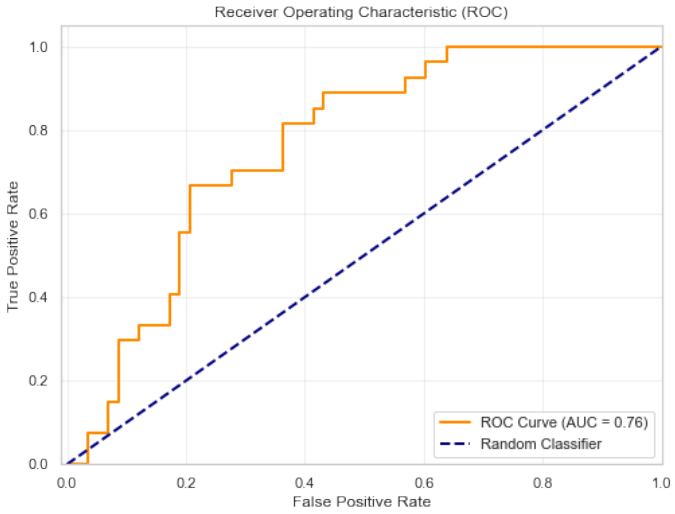


Figure 4.24: ROC Curve for KNN

4.8.4 小结

KNN 模型在本数据集上表现出中等的分类能力，尤其在识别类别 0 上效果显著，但在类别 1 上的表现较弱。这主要体现在高召回率和 F1 分数在类别 0 上的优秀表现，而类别 1 的召回率和 F1 分数极低，导致整体准确率受到影响。

未来的工作可以考虑以下几点以提升模型性能：

- **数据平衡**: 通过过采样、欠采样或使用合成少数类过采样技术（如 SMOTE）来平衡类别分布，提升模型对少数类的识别能力。
- **特征工程**: 增加更多的特征或进行特征选择，提取更具区分性的特征，以增强模型的分类能力。
- **模型参数优化**: 进一步细化超参数的搜索范围，例如扩大 n_neighbors 的范围，或尝试其他距离度量方法，以找到更优的参数组合。
- **尝试其他算法**: 由于 KNN 在本数据集上对类别 1 的识别效果较差，可以尝试其他分类算法，以比较不同模型的性能。
- **处理类别不平衡**: 在模型训练过程中，使用加权损失函数或调整分类阈值，以提高模型对少数类的敏感性。

通过上述方法，可以进一步提升 KNN 模型的性能，尤其是在类别 1 上的识别能力，从而实现更为准确和可靠的分类结果。

4.9 模型性能对比与总结

以下表格总结了不同模型的性能指标，包括准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 分数（F1-Score）以及 ROC-AUC 值。

Table 4.8: 不同模型的性能对比

模型	准确率	类别 0 精确率	类别 1 精确率	类别 0 召回率	类别 1 召回率	ROC-AUC
逻辑回归	0.73	0.75	0.62	0.90	0.37	0.8084
逻辑回归 + 逐步删除	0.75	0.78	0.65	0.88	0.48	0.8231
SVM	0.74	0.75	0.69	0.93	0.33	0.8142
SVM+RFE	0.75	0.88	0.58	0.74	0.78	0.8244
随机森林	0.66	0.71	0.44	0.84	0.26	0.7114
XGBoost	0.71	0.71	0.62	0.95	0.19	0.7656
XGBoost+RFE	0.69	0.72	0.55	0.91	0.22	0.7893
AdaBoost	0.62	0.69	0.33	0.83	0.19	0.66
Gradient Boosting	0.67	0.71	0.46	0.88	0.22	0.71
KNN	0.67	0.69	0.40	0.95	0.07	0.76

通过对不同模型的性能指标（准确率、精确率、召回率、F1 分数及 ROC-AUC 值）的对比分析，可以得出以下结论：

- ** 逻辑回归与改进 **：基础逻辑回归模型的准确率为 0.73，类别 0 召回率较高（0.90），但类别 1 召回率较低（0.37），说明模型对类别 0 的预测较优，对类别 1 的预测不足。经过逐步删除变量后的逻辑回归模型，其性能有明显提升，准确率提高至 0.75，类别 1 召回率上升至 0.48，且 ROC-AUC 值提高到 0.8231，表明模型更加均衡。

- **SVM 与变量选择 **：支持向量机（SVM）的基础模型准确率为 0.74，类别 1 召回率较低（0.33），但类别 0 召回率较高（0.93）。结合递归特征消除（RFE）后，SVM 模型的类别 1 召回率显著提高到 0.78，虽然类别 0 召回率有所下降，但总体性能更均衡，ROC-AUC 值上升至 0.8244，体现了特征选择的重要性。

- ** 随机森林 **：随机森林的基础模型表现较弱，准确率为 0.66，说明模型在预测时存在较大的误差，有很多样本被错误分类。类别 1 召回率仅为 0.26，这可能意味着模型对于正样本的识别能力较弱，导致大量正样本被误判为负样本。ROC-AUC 值为 0.7114，但仍然低于许多实际应用中期望的值（通常希望 ROC-AUC 值接近 1）。这表明模型在区分正负样本方面的能力有限。

- **Boosting 模型 **：XGBoost 模型的基础版本表现优于随机森林和 AdaBoost，准确率达到 0.71，ROC-AUC 值为 0.7656。然而，类别 1 召回率较低（0.19），表明对类别 1 的预测能力不足。结合 RFE 优化后，ROC-AUC 值上升至 0.7893，但其他指标提升有限。Gradient Boosting 和 AdaBoost 的性能表现整体较弱，准确率分别为 0.67 和 0.62，类别 1 的预测能力依旧不足。

- **KNN 模型 **：K 最近邻（KNN）模型的准确率为 0.67，类别 0 召回率较高（0.95），但类别 1 召回率仅为 0.07，说明该模型对类别 1 的预测几乎无效，整体性能较差。

综合来看，SVM 结合 RFE 优化后的模型在多个指标上表现最佳，尤其在 ROC-AUC 值上达到了 0.8244，显示了较好的分类能力。同时，逻辑回归经过逐步删除优化后也表现良好，适合在特征选择简单的情况下使用。而随机森林和 Boosting 模型尽管有一定的改进空间，但其性能相对欠佳，尤其是类别 1 召回率普遍较低。

** 推荐使用的模型 **：优先选择 SVM 结合 RFE 的模型，次选逻辑回归结合逐步删除的模型。