Modeling and Simulation 建模与仿真, 2025, 14(7), 228-237
Published Online July 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.147531

Hans汉斯

基于Logistic回归模型的乳腺癌诊断特征筛选

常芳欣，杨双杨，冯爱芬*，曹君杰，蒋智涵，王世杰

河南科技大学数学与统计学院，河南 洛阳

收稿日期：2025年6月25日；录用日期：2025年7月18日；发布日期：2025年7月28日

摘　要

本文主要利用Logistic回归的数学建模方法对乳腺癌诊断问题进行研究，首先基于数据特征，对数据进行清洗、缺失值处理和标准化处理，采用点二列相关系数和基于Logistic回归的递归特征消除法(RFE)分别筛选出15个关键特征构建模型。通过对比两种方法所选特征构建的Logistic回归分类模型，发现基于RFE所选特征的模型准确率达到97.89%，优于基于点二列相关系数所选特征的模型准确率，表现出较高的预测性能。由此不仅验证了机器学习在乳腺癌预测中的有效性，还为未来模型优化和病因探索提供了重要参考，助力临床医生实现“三早”预防。

关键词

乳腺癌，Logistic回归，递归特征消除法，点二列相关系数

Feature Selection for Breast Cancer Diagnosis Based on the Logistic Regression Model

Fangxin Chang, Shuangyang Yang, Aifen Feng*, Junjie Cao, Zhihan Jiang, Shijie Wang

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang Henan

Received: Jun. 25th, 2025; accepted: Jul. 18th, 2025; published: Jul. 28th, 2025

Abstract

This paper mainly studies the problem of breast cancer diagnosis by using the mathematical modeling method of Logistic regression. Firstly, based on the data characteristics, the data is cleaned,

*通讯作者。

文章引用: 常芳欣, 杨双杨, 冯爱芬, 曹君杰, 蒋智涵, 王世杰. 基于 Logistic 回归模型的乳腺癌诊断特征筛选[J]. 建模与仿真, 2025, 14(7): 228-237. DOI: 10.12677/mos.2025.147531

missing values are processed, and standardized. The point-biserial correlation coefficient and the recursive feature elimination method based on Logistic regression (RFE) are used to select 15 key features to build the model. By comparing the Logistic regression classification models constructed with the features selected by the two methods, it is found that the accuracy rate of the model based on the features selected by RFE reaches 97.89%, which is better than the accuracy rate of the model based on the features selected by the point-biserial correlation coefficient, showing high predictive performance. This not only verifies the effectiveness of machine learning in breast cancer prediction, but also provides an important reference for future model optimization and etiological exploration, helping clinicians achieve "three early" prevention.

Keywords

Breast Cancer, Logistic Regression, Recursive Feature Elimination Method, Point-Biserial Correlation Coefficient

1. 引言

近年来，乳腺癌已成为全球女性发病率最高的恶性肿瘤[1]。2021 年世界卫生组织最新数据显示乳腺癌已经成为了全球新发病例最多的癌症，中国乳腺癌的发病数位居全球第四(WHO, 2020) [2]。目前，乳腺癌的诊断主要依赖于影像学检查、血清肿瘤标志物检测和病理学检查。传统的诊断方法依赖于病理学家的主观判断，这种方法耗时且易受人为因素影响。因此，寻找有效的生物标志物对提高乳腺癌早期诊断率具有重要意义。随着机器学习技术的不断发展与改进，利用机器学习技术对疾病进行预测不仅可以大大提高准确率，而且减少了人工成本，有望实现自动化诊断。例如，卢聪基于机器学习算法构建胰腺癌多器官转移病人的生存预测模型，使用单因素联合多因素 COX 回归分析筛选出影响 PCMOM 病人预后的独立危险因素，利用筛选的独立危险因素构建 COX 回归模型和随机生存森林模型，并对比随机森林模型机器学习模型，得出 COX 模型优于随机森林模型[3]。甄凯旋等学者采用逻辑回归、随机森林、支持向量机和 XGBoost 等机器学习方法建立了对儿童脓毒性休克临床预测模型的 Meta 分析，能够从大量的数据中学习模式，自动区分恶性肿瘤和良性肿瘤，从而提高诊断的准确性和效率，对未来工作提供了建设性建立[4]。王佩佩通过 LASSO 回归筛选出影响肾癌患者术后复发风险的因素，并根据这些因素建立了 logistic 模型、决策树模型、随机森林、贝叶斯模型等多种机器学习模型，并分析哪种模型最优，为临床决策提供了有效的参考意见[5]。时欣然学者运用 k 最近邻、支持向量机、决策树及随机森林构建女性压力性尿失禁发病的预测模型，并分析对比每种模型的效果，为 SUI 的早期诊断提供有力的参考[6]。由此可见，机器学习已经渗透到医学领域的方方面面。

本文以美国威斯康星州乳腺癌公开数据集[7]作为研究对象，选取相关性强的因素作为 Logistic 模型的自变量，对肿瘤是良性还是恶性进行预测，大大减少人工成本，并利用 K 折交叉验证对模型的准确率进行了评估。

2. 数据及可视化处理

2.1. 数据来源与数据预处理

本研究数据来源于美国威斯康星州乳腺癌数据集。该数据共有 569 个样本，其中良性肿瘤 357 个，

恶性肿瘤 212 个。每个样本包含了乳腺癌诊断的 30 个特征数据，描述了肿瘤的尺寸、纹理、周长、面积、平滑度、紧凑度、凹陷点数、对称性、分形维数等特性。这些特征被用来量化肿瘤的形态学特性，对于乳腺癌的诊断和分析至关重要，部分数据见表 1。

Table 1. Some breast cancer data from Wisconsin, USA
表 1. 部分美国威斯康星州乳腺癌数据

ID	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
842302	M	17.99	10.38	122.8	1001
842517	M	20.57	17.77	132.9	1326
84300903	M	19.69	21.25	130	1203
84348301	M	11.42	20.38	77.58	386.1
84358402	M	20.29	14.34	135.1	1297

数据预处理包括数据清洗和归一化。该数据表中没有缺失值，因此可以直接使用，再对特征数据进行 Z-score 标准化处理，确保数据的均值为 0，方差为 1。

标准化公式如下：

对于每个特征 X，计算其均值 μ 和标准差 δ，然后对每个样本 x_i 进行变换：

$$x_{s\tan dard} = \frac{x_i - u}{\delta}$$

在这个数据集中 diagnosis 状态为“M”时，表示为恶性，为“B”时表示为良性。为了方便后续对数据的梳理，我们采用独热编码的方式将恶性表示为 1，将良性表示为 0。

Table 2. The preprocessed partial breast cancer data
表 2. 经过预处理的部分乳腺癌数据

ID	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
842302	1	1.096099529	−2.071512302	1.268817263	0.98350952
842517	1	1.828211974	−0.353321523	1.684472552	1.907030269
84300903	1	1.578499202	0.455785908	1.565125984	1.557513185
84348301	1	−0.768233323	0.253509051	−0.592166123	−0.763791736
84358402	1	1.74875791	−1.150803847	1.775011328	1.824623802

2.2. 特征筛选

2.2.1. 基于点二列相关系数的特征筛选

基于特征之间相关性构建模型是一种常见的特征选择方法，尤其在处理高维数据集时。这种方法的核心是选择与目标变量相关性较强或者特征之间相互关联性较强的特征来构建模型。不仅具有降低维度、提高计算效率的优点，还具有减少多重共线性问题等优点。

为了更加直观地观测到特征之间的相关性如何，我们绘制了如下反映特征之间相关性的热力图。

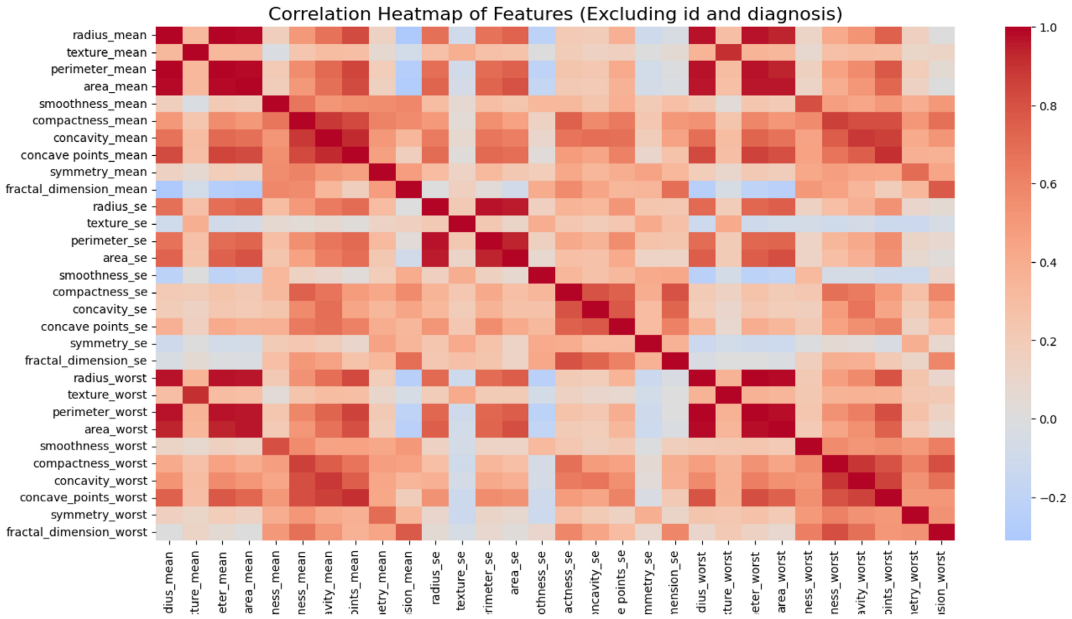

Figure 1. Correlation heatmap
图 1. 相关性热力图

为了进一步推断出哪些特征与目标变量“diagnosis”之间存在较强的关联性。本文采用点二列系数来量化这种关联性，根据和目标变量相关程度的大小，筛选出最有代表性的属性特征。

点二列相关系数是一种用于衡量二分类变量与连续变量之间的线性关系的统计量。其取值范围介于−1 和 1 之间，绝对值越接近 1 表示相关性越强。通过编程工具 Pycharm 的实现，我们得到了如下的十五个特征和目标变量 diagnosis 之间的相关系数(表 3)。

Table 3. Feature selection using point-biserial correlation
表 3. 基于点二列系数筛选的特征

特征英文名	特征中文名	与目标变量的相关性系数
concave points_mean	凹陷点数量的平均值	0.7766
perimeter_mean	周长的平均值	0.7426
radius_mean	半径的平均值	0.7300
concavity_mean	凹度的平均值	0.6964
compactness_mean	紧凑度的平均值	0.5965
area_mean	面积平均值	0.7090
concave_points_worst	凹陷点数量的最大值	0.7936
perimeter_worst	周长的最大值	0.7829
radius_worst	半径的最大值	0.7765
area_worst	面积的最大值	0.7338
concavity_worst	凹度的最大值	0.6596
compactness_worst	紧凑度的最大值	0.5910
radius_se	半径的标准误差	0.5671
perimeter_se	周长的标准误差	0.5561
area_se	面积的标准误差	0.5482

从表 3 可以看出，这十五个特征和目标变量 diagnosis 之间呈现了较强的相关性。

2.2.2. 基于 Logistic 回归的递归特征消除法

递归特征消除(RFE)是采用 Logistic 模型来评估特征的重要性[8]。在每次迭代中，REF 会训练一个 Logistic 模型，并根据模型的系数来评估每个特征的重要性。对于 Logistic 回归模型来说，特征的重要性通常通过模型的系数绝对值来衡量。系数绝对值越大，特征的重要性越高。之后在每次迭代过程中，REF 会去除重要性最低的特征，然后在剩余的特征上重新训练模型，直到达到所设置的特征数量阈值。最终，REF 会保留一组最重要的特征。

基于 Logistic 回归的递归特征筛选所得特征结果如表 4 所示。

Table 4. Logistic regression-based feature selection
表 4. 基于 Logistic 回归筛选的特征

特征英文名	特征中文名
area_mean	细胞核面积的平均值
compactness_mean	紧凑度的平均值
Concave points_mean	凹陷点数量的平均值
Radius_se	半径的标准误差
Perimeter_se	周长的标准误差
Area_se	面积的标准误差
Compactness_se	紧凑度的标准误差
Radius_worst	半径最大值
Texture_worst	纹理灰度标准差最大值
Perimeter_worst	周长最大值
Area_worst	面积最大值
Smoothness_worst	平滑度最大值
Concavity_worst	凹度最大值
Symmetry_worst	对称性最大异常值
Concave_points_worst	凹陷点数量最大值

由表 4 和表 3 不难看出，两种特征筛选方法最终所得到的结果差异较大。在基于 logistic 回归模型的递归特征消除(RFE)方法的过程中，所得到的特征集合与通过相关性分析方法筛选出的特征集合高度一致，仅在紧凑度的平均值(compactness_mean)这一特征上存在差异。具体而言，相关性筛选方法未将 compactness_mean 纳入最终特征集合，而 RFE 方法则将其排除，同时 RFE 方法额外选择了纹理的平均值(texture_mean)作为关键特征。

2.3. Logistic 回归方法

通过对数据进行初步探索(如图 2 所示)，我们发现该数据集呈现出典型的二分类特征，目标变量明确划分为两个类别(例如肿瘤的良性与恶性)。基于此，我们选择了 Logistic 回归作为建模方法。Logistic 回归作为一种广义线性模型，能够有效处理因变量为二分类的情形，其核心优势在于能够输出样本属于某一类别的条件概率。这种概率输出不仅为模型的分类结果提供了置信度评估，还为临床决策提供了量化依据，使得医学专家能够在参考模型预测概率的基础上，结合专业知识和经验，作出更加精准且具有依据的诊断判断。

Logistic 回归模型以其简单性和高效的训练速度而著称。它在实际医疗场景中能够实现快速部署，尤其适合需要快速响应的临床环境。这种模型的高效性不仅提高了诊断效率，还减少了患者的等待时间，为医疗决策提供了及时的支持。

从技术角度而言，Logistic 回归的核心思想是通过线性组合特征，并利用 Logistic 函数(即 Sigmoid 函

数)将输出映射到(0, 1)区间。这种映射方式使得模型能够预测目标变量属于某一类别的概率。其数学原理基于对数几率的线性组合和最大似然估计，这使得 Logistic 回归能够有效地处理二分类问题。此外，模型的系数可以直接解释为特征的重要性，为特征分析和模型解释提供了便利。

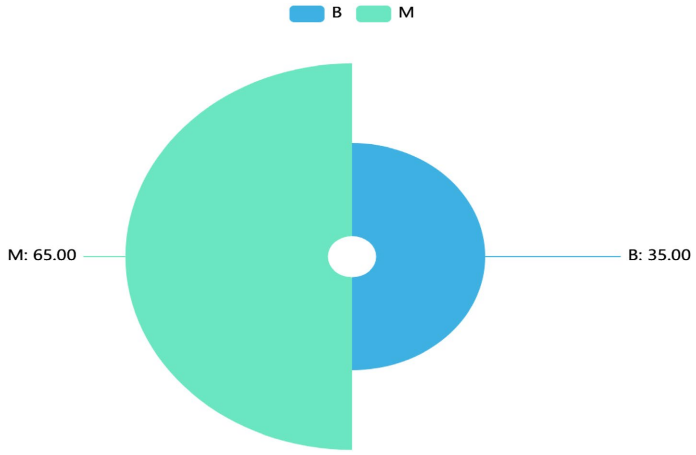

Figure 2. Distribution of diagnostic results for benign (0) and malignant (1) tumors
图 2. 良性肿瘤(0)和恶性肿瘤(1)诊断结果分布情况

Logistic 回归的数学原理

◆ 模型形式

Logistic 回归模型的目标是预测目标变量 y 属于某一类别的概率 p。模型的形式可以表示为[9]：

$$\text{logit}(p)=\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n$$

其中：logit (p)是对数几率(log-odds)，表示目标变量属于某一类别的对数几率。p/(1 − p)是几率(odds)，表示目标变量属于某一类别的概率与不属于该类别的概率之比。通过对数几率的线性组合，Logistic 回归将特征与目标变量的概率联系起来。

◆ 似然函数

Logistic 回归的训练过程是通过最大化似然函数来估计模型参数(系数 β)。似然函数表示模型在训练数据上的概率分布。对于二分类问题，似然函数为：

$$L(\beta)=\prod_{i=1}^{m}p_i^{y_i}(1-p_i)^{1-y_i}$$

其中：m 是训练样本的数量，y_i 是第 i 个样本的目标变量(0 或 1)，p_i 是模型预测第 i 个样本属于类别 1 的概率。

◆ 最大似然估计

Logistic 回归通过最大化似然函数来估计模型参数。最大化似然函数等价于最小化负对数似然函数(Negative Log-Likelihood, NLL)：

$$\text{NLL}(\beta)=\sum_{i=1}^{m}\left[-y_i\beta^{\mathrm{T}}\hat{x}_i+\ln\left(1+e^{\beta^T\hat{x}_i}\right)\right]$$

通过优化算法(如梯度下降)最小化负对数似然函数，从而估计出最优的模型参数 β。部分推导过程如下[10]：

① 设输入特征向量为 X∈R，模型参数为 θ∈R，建立线性关系：

$$z = w^T x + b \tag{2.1}$$

② 通过 sigmoid 函数将线性输出转换为概率：

$$y = \frac{1}{1+e^{-z}} \tag{2.2}$$

③ 并将其带入(2.1)式得到：

$$y = \frac{1}{1+e^{-\left(w^T x+b\right)}} \tag{2.3}$$

④可变化为：

$$\ln\frac{y}{1-y} = w^T + b \tag{2.4}$$

则得到对数概率(log odds，亦称 logit)：$\ln\frac{y}{1-y}$。接下来确定式(2.3)中的 w 和 b 将(2.3)式中的 y 视为类后验概率估计 $p(y=1|x)$，则(2.4)可重写为：

$$p(y=1|x) = \frac{e^{w^T x+b}}{1+e^{w^T x+b}} \tag{2.5}$$

$$p(y=0|x) = \frac{1}{1+e^{w^T x+b}} \tag{2.6}$$

于是通过极大似然法来估计 w 和 b.对数回归模型最大化“对数似然函数”为：

$$l(w,b) = \sum_{i=1}^{m} \ln p(y_i | x_i; w, b) \tag{2.7}$$

令 $\beta = (w;b)$，$\hat{x} = (x;1)$。再令 $p_1(\hat{x};\beta) = p(y=1|\hat{x};\beta), p_0(\hat{x};\beta) = p(y=0|\hat{x};\beta) = 1 - p_1(\hat{x};\beta)$，则(2.7)中的似然项可重写为：

$$p(y_i | x_i; w, b) = y_i p_1(\hat{x}_i;\beta) + (1-y_i) p_0(\hat{x}_i;\beta) \tag{2.8}$$

将(2.8)代入(2.7)，并根据(2.5)和(2.6)可知，最大化似然函数(2.7)等价于最小化

$$l(\beta) = \sum_{i=1}^{m}\left(-y_i \beta^T \hat{x}_i + \ln\left(1+e^{\beta^T \hat{x}_i}\right)\right) \tag{2.9}$$

3. 模型结果及评价方法

3.1. 基于两种特征筛选方法所得出来的模型结果

通过编程工具 Pycharm 所求的模型结果如表 5 所示。

Table 5. Model performance results
表 5. 模型结果

特征选择方法	模型准确率
基于点二列相关系数的特征筛选	0.9649
基于 Logistic 回归的递归特征筛选	0.9789

从结果可以看出，基于 Logistic 回归的递归特征选择方法所构建的模型准确率更高，这表明该方法所选择的十五个特征更为有效。因此，在后续进行模型性能评估时，我们只对基于 Logistic 回归的递归特征选择方法所建立的模型进行性能评估。

3.2. 模型评价方法

在机器学习中，评价指标用于评估模型的性能和预测能力。本文使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1-Score、K 折交叉验证作为评价指标[11]。

准确率(Accuracy)即模型预测正确样本占总样本数的比例。精确率(Precision)即模型预测为正例(Positive)的样本中，实际为正例的样本所占的比例。召回率(Recall)即实际为正例的样本中，被模型正确预测为正例的样本所占的比例。F1 得分(F1-Score)即精确率和召回率的加权平均值，用于综合评估模型的性能。设 TP 表示正确预测为恶性样本数量，FN 表示错误预测为恶性样本数量，FP 表示错误预测为良性的样本数量，TN 表示正确预测为良性的样本数量。它们的具体计算公式具体如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times P\mathrm{recision} \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$$

K 折交叉验证(K-Fold Cross Validation)：将数据集分为 K 个等大小的子集，每次选择一个子集作为验证集，其余 K-1 个子集作为训练集。重复 K 次，最终的评估结果是 K 次评估结果的平均值，用于综合评估模型的性能和鲁棒性。K 折交叉验证的取值范围没有固定的上下限，但通常 K 值越大，评估结果越稳定。

3.3. 对基于 Logistic 回归的递归特征筛选的模型性能进行评估

基于 Logistic 回归的递归特征筛选的模型性能评估的五折交叉验证的准确率见表 6。

Table 6. 5-fold cross-validation model accuracy
表 6. 五折交叉验证模型准确率

五折交叉验证	第 1 折	第 2 折	第 3 折	第 4 折	第 5 折
准确率	98.25%	94.74%	98.25%	99.12%	99.12%

由表 6 可以看出来，五折交叉验证的准确率在 94.74%到 99.12%之间，表明模型在不同的数据划分下表现相对稳定，但在某些划分下可能存在轻微的波动。

基于 Logistic 回归的递归特征筛选的模型性能评估的其他指标结果见表 7，ROC 曲线见图 3。

Table 7. Results of all metrics
表 7. 各项指标的结果

评价指标	精确率	召回率	F1-score
0	0.97	0.99	0.98
1	0.99	0.95	0.97

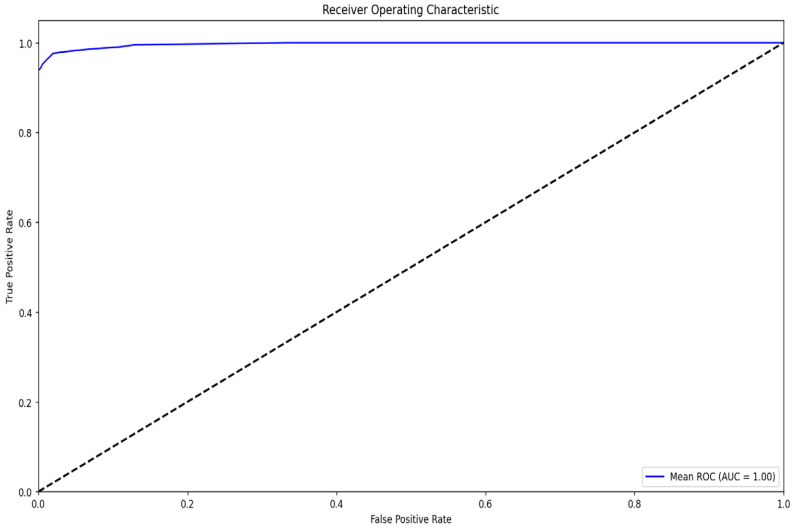

Figure 3. ROC curve
图 3. ROC 曲线

表 7 表示，分类报告中的精确率、召回率和 F1-score 均较高，表明模型在分类上任务表现良好，能够准确地预测正类和负类。图 3 也同样表示模型在区分正负类方面表现极佳。

4. 讨论

为了探究所选取的十五个特征是否真的最优，我们绘制了特征与诊断关系的箱线图(图 4)。由图 4 可以看出来，这些特征在两类诊断中的数据分布存在显著差异，特别是 radius_se，concave points_worst 等特征，这表明了这些特征在区分两类诊断上具有重要的价值。此外，大多数特征在两个诊断类别的中位数有所不同，这意味这些特征的平均水平在不同类别间具有显著变化。

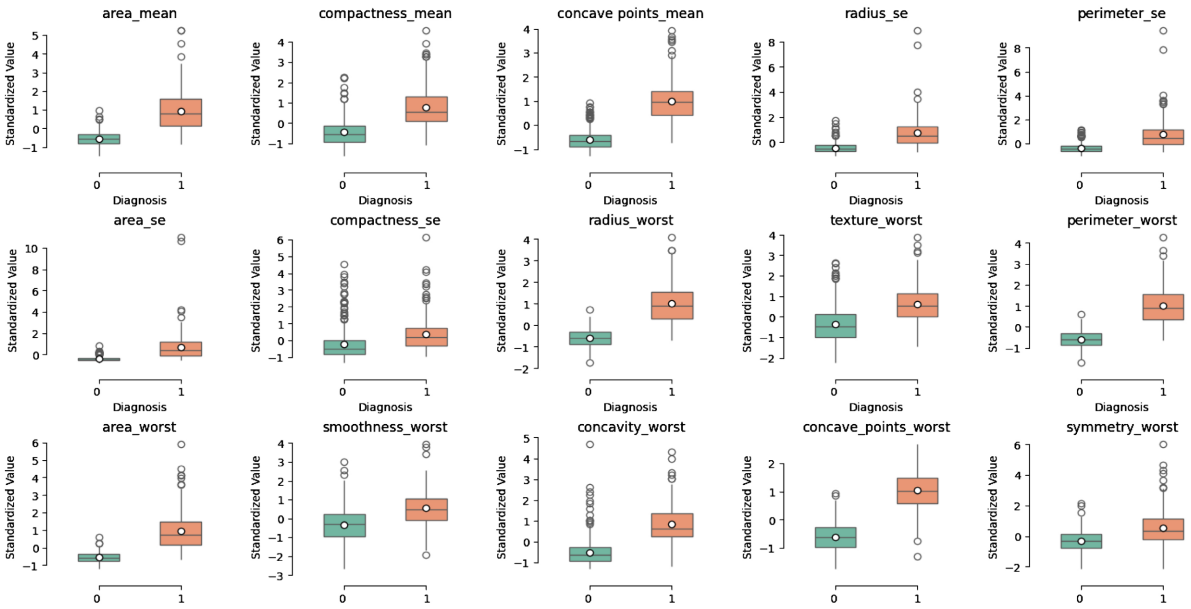

Figure 4. Boxplots of features versus diagnostic outcomes
图 4. 特征与诊断结果的关系箱线图

5. 结论

本研究基于美国威斯康星州乳腺癌数据集，通过数据预处理和两种特征筛选方法(点二列相关系数和递归特征消除法)构建 Logistic 回归模型。研究结果表明，基于递归特征消除法(RFE)筛选的特征所构建的模型准确率更高，达到 97.89%，优于点二列相关系数模型的 96.49%。通过五折交叉验证和各项评估指标(精确率、召回率、F1-score 及 ROC 曲线)的分析，证实该模型在区分良恶性肿瘤方面表现优异且稳定。特征可视化分析进一步验证了所选特征在诊断中的重要性。该研究不仅为乳腺癌早期诊断提供了高效、可靠的预测工具，还展示了机器学习在医学领域的应用潜力，对临床实现“三早”预防具有重要参考价值。

基金项目

河南省“专创融合”特色示范课程项目(2023)，河南科技大学教育教学改革研究与实践项目(2024BK089)，河南科技大学大学生创新创业计划项目(2024238)。

参考文献

[1] 张雅聪, 吕章艳, 宋方方, 等. 全球及我国乳腺癌发病和死亡变化趋势[J]. 肿瘤综合治疗电子杂志, 2021, 7(2): 14-20.

[2] 何梦. “丁香医生”微信公众号乳腺癌信息传播的框架研究[D]: [硕士学位论文]. 北京: 北京外国语大学, 2022.

[3] 鲁聪, 宋丹, 王伟, 等. 基于机器学习算法构建胰腺癌多器官转移病人的生存预测模型[J/OL]. 腹部外科, 1-15. http://kns.cnki.net/kcms/detail/42.1252.R.20250225.1326.002.html, 2025-07-23.

[4] 甄凯旋, 闫浩伦, 陈龙彪, 等. 基于机器学习儿童脓毒性休克临床预测模型的 Meta 分析[J]. 中国循证医学杂志, 2025, 25(2): 200-205.

[5] 王佩佩, 侯钊, 马慧, 等. 基于机器学习的肾癌患者术后复发风险预测模型的构建与评价[J]. 现代泌尿外科杂志, 2025, 30(3): 240-247.

[6] 时欣然, 庞震, 乔婷, 等. 基于机器学习的女性压力性尿失禁发病风险预测模型建立及效能评价[J]. 现代泌尿外科杂志, 2025, 30(3): 196-206.

[7] 阿里云. 威斯康星乳腺癌数据分析及自动诊断[EB/OL]. https://tianchi.aliyun.com/dataset/106831, 2021-07-21.

[8] 李华, 张伟, 王丽. 乳腺癌早期诊断中多模态影像特征与机器学习算法的结合应用[J]. 中国医学影像技术, 2023, 39(6): 854-859.

[9] 姜启源, 谢金星, 叶俊. 数学模型[M]. 第五版. 北京: 高等教育出版社, 2020.

[10] 周涛, 杨柳, 李娜. 乳腺癌风险预测模型的构建与验证[J]. 中国公共卫生, 2024, 35(4): 523-527.

[11] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.