



浙江农林大学
ZHEJIANG A&F UNIVERSITY

机器学习

学生姓名: 黄豪

学 号: 2024611031011


专业班级: 数计研 241 班

所在学院: 数学与计算机科学学院

浙江农林大学

硕士生课程论文（设计）诚信承诺书

我谨在此承诺：本人所写的课程论文（设计）《分类作业报告》均系本人独立完成，没有抄袭行为，凡涉及其他作者的观点和材料，均作了引用注释，如出现抄袭及侵犯他人知识产权的情况，后果由本人承担。

承诺人（签名）：

2024 年 12 月 21 日

分类作业报告

数学与计算机科学学院 数计研 241 黄豪 指导教师：夏凯

摘要：本研究基于葡萄酒质量数据集，探索多种机器学习模型在多分类任务中的性能表现。数据集包含红白葡萄酒的 12 项化学属性和质量评分，共计 6497 个样本，评分存在类别不平衡现象。实验选取逻辑斯蒂回归、朴素贝叶斯（伯努利、多项式、高斯）、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机及极限梯度提升等模型，采用交叉验证、准确率、交叉熵损失及容忍度准确率作为评价指标。实验结果显示，逻辑斯蒂回归模型准确率仅为 24.73%，容忍度准确率为 52.72%，性能较差；朴素贝叶斯模型中，多项式分类器表现最佳，准确率为 45.32%，容忍度准确率达 93.91%。深度随机森林准确率为 51.78%，容忍度准确率为 84.7%。极端随机树和随机森林表现出色，准确率分别达到 64.90%和 65.99%，容忍度准确率分别为 94.78%和 94.63%。极限梯度提升和梯度提升决策树的准确率分别为 61.28%和 56.27%，容忍度准确率分别为 94.42%和 93.47%。聚类实验中，KMeans 算法通过 PCA 降维后表现出较好的聚类效果，而 DBSCAN 算法因对密度的依赖导致部分类别混杂。研究表明集成学习模型在该数据集上的分类效果显著优于传统方法，特别是在处理类别不平衡问题时具有更高的鲁棒性。这为葡萄酒质量预测提供了数据支持，并为未来的模型改进提供了参考方向。

更多详细工作内容尽在：https://github.com/H-0526/Machine_Learning/tree/master

关键词：机器学习；集成学习；分类；随机森林

Classification Assignment Report

Abstract: This study explores the performance of various machine learning models in a multiclass classification task based on the Wine Quality dataset. The dataset includes 12 chemical attributes and quality ratings of red and white wines, totaling 6,497 samples, with significant class imbalance in the quality ratings. The experiments utilized models such as Logistic Regression, Naive Bayes (Bernoulli, Multinomial, Gaussian), Deep Random Forest, Extremely Randomized Trees, Gradient Boosting Decision Trees, LightGBM, Random Forest, Support Vector Machines, and Extreme Gradient Boosting. Evaluation metrics included cross-validation scores, accuracy, log-loss, and tolerance accuracy. The results indicate that Logistic Regression performed poorly, achieving an accuracy of only 24.73% and a tolerance accuracy of 52.72%. Among Naive Bayes models, the Multinomial classifier showed the best performance, with an accuracy of 45.32% and a tolerance accuracy of 93.91%. The Deep Random Forest achieved an accuracy of 51.78% and a tolerance accuracy of 84.7%. Extremely Randomized Trees and Random Forest demonstrated excellent performance, with accuracies of 64.90% and 65.99%, and tolerance accuracies of 94.78% and 94.63%, respectively. Extreme Gradient Boosting and Gradient Boosting Decision Trees achieved accuracies of 61.28% and 56.27%, and tolerance accuracies of 94.42% and 93.47%, respectively. In clustering experiments, the KMeans algorithm, after PCA dimensionality reduction, exhibited good clustering performance, while the DBSCAN algorithm showed mixed results due to its dependence on density, resulting in some class overlaps. The study demonstrates that ensemble learning models significantly outperform traditional methods on this dataset, particularly in addressing class imbalance, showcasing their robustness. These findings provide data-driven support for wine quality prediction and offer directions for future model improvements.

For more details, visit: https://github.com/H-0526/Machine_Learning/tree/master

Key words: Machine Learning, Ensemble Learning, Classification, Random Forest

目 录

| | |
|-----------------------|----|
| 摘要 | I |
| Abstract | II |
| 1 绪论 | 1 |
| 1.1 数据集介绍 | 1 |
| 1.2 数据集理解 | 1 |
| 1.3 研究目的与内容 | 2 |
| 1.3.1 研究目的 | 2 |
| 1.3.2 研究内容 | 2 |
| 1.3.3 技术路线 | 3 |
| 2 相关分类方法介绍 | 4 |
| 2.1 逻辑斯蒂回归 | 4 |
| 2.2 朴素贝叶斯 | 5 |
| 2.3 深度随机森林 | 5 |
| 2.4 极端随机树 | 6 |
| 2.5 梯度提升决策树 | 6 |
| 2.6 轻量级梯度提升机 | 7 |
| 2.7 随机森林 | 7 |
| 2.8 支持向量机 | 8 |
| 2.9 极限梯度提升 | 9 |
| 2.10 评价指标 | 9 |
| 3 基于机器学习算法的实验过程 | 11 |
| 3.1 数据探索 | 11 |
| 3.1.1 特征变量分析 | 11 |
| 3.1.2 目标变量分析 | 12 |
| 3.1.3 变量相关性分析 | 13 |
| 3.1.4 异常值检测 | 15 |
| 3.2 数据预处理 | 16 |

| | | |
|-------|-----------------|----|
| 3.2.1 | 数据清洗 | 16 |
| 3.2.2 | 特征工程 | 16 |
| 3.3 | 模型训练和调参 | 17 |
| 3.3.1 | 逻辑斯蒂回归 | 17 |
| 3.3.2 | 朴素贝叶斯 | 17 |
| 3.3.3 | 深度随机森林 | 17 |
| 3.3.4 | 极端随机树 | 18 |
| 3.3.5 | 梯度提升决策树 | 18 |
| 3.3.6 | 轻量级梯度提升机 | 19 |
| 3.3.7 | 随机森林 | 19 |
| 3.3.8 | 支持向量机 | 20 |
| 3.3.9 | 极限梯度提升 | 20 |
| 3.4 | 模型对比分析 | 21 |
| 3.5 | 聚类和降维 | 22 |
| 3.5.1 | KMeans 聚类 | 22 |
| 3.5.2 | DBSCAN 聚类 | 23 |
| 4 | 总结与展望 | 24 |
| 4.1 | 全文总结 | 24 |
| 4.2 | 展望 | 24 |
| 致 谢 | | 25 |

1 绪论

1.1 数据集介绍

葡萄酒质量数据集 (Wine Quality Data Set) 是一个经典的多分类任务数据集，广泛应用于机器学习和数据分析研究。该数据集来源于 UCI 机器学习数据集仓库，包含红葡萄酒和白葡萄酒的化学分析属性以及质量评分两部分，共计 6497 个样本（红葡萄酒 1599 个，白葡萄酒 4898 个）。数据集提供了 12 个物理化学属性，如固定酸度、挥发性酸度、柠檬酸含量、残糖、氯化物含量、自由二氧化硫、总二氧化硫、密度、pH 值、硫酸盐浓度和酒精浓度，以及范围为 0 到 10 的目标变量质量评分，这是由专业品酒师通过主观品评得出的等级。该数据集适合用于深入探索葡萄酒质量与化学特性的关系，涉及数据清洗、相关性分析、特征工程等步骤。分析中既可以采用监督学习模型进行分类任务，也可以借助无监督学习进行数据分组并与分类结果进行对比。此外该数据集对模型评估提出了较高要求，通过准确率、F1 分数、混淆矩阵等指标进行对比分析，并通过降维或可视化技术展示结果。这些分析为了解葡萄酒质量评分的决定因素提供了科学依据，也可为葡萄酒酿造工艺优化和市场策略制定提供参考。

1.2 数据集理解

多分类任务是机器学习中的一种问题类型，其目标是将数据样本分为多个离散类别，通常多于两类。与二分类任务不同，多分类任务需要预测样本属于某一类别的概率分布或直接输出类别标签，解决多分类任务需要克服类别不平衡、类别相关性以及噪声干扰等挑战，同时需要选择合适的模型和评价指标。经典模型如逻辑回归、支持向量机、随机森林以及梯度提升决策树等，均能有效应对多分类任务。

葡萄酒质量数据集是一个典型的多分类任务数据集，目标是预测葡萄酒的质量评分，评分范围从零到十，该数据集包含十二个物理化学属性，例如固定酸度、挥发性酸度和酒精浓度等，同时提供由专业品酒师主观评分的目标变量，该数据集存在显著的类别不平衡现象，例如评分为五和六的样本数量远多于评分为三或九的样本。此外评分之间还具有序列性和一定的主观性，这进一步增加了分类任务的复杂性。在分析过程中，需要通过数据清洗、特征工程以及相关性分析，识别对分类最重要的特征，同时可以结合逻辑回归、支持向量机、随机森林和梯度提升决策树等模型，通过调整

类别权重和使用过采样技术来缓解类别不平衡问题。

1.3 研究目的与内容

1.3.1 研究目的

本研究的目的是通过探索多种机器学习模型在葡萄酒质量数据集上的性能表现，揭示不同模型在处理多分类任务中的优势与不足，特别是在面对类别不平衡问题时的适应性和鲁棒性。研究旨在评估逻辑斯蒂回归、朴素贝叶斯、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升等模型在分类任务中的有效性，为葡萄酒质量预测提供可靠的技术支持，并为未来相关模型的改进提供参考方向，同时通过实验对比容忍度准确率、交叉验证、交叉熵损失等指标，进一步挖掘集成学习模型在处理复杂数据集时的潜力和实用价值。

1.3.2 研究内容

第一部分：绪论。绪论部分概述了数据集的来源并给与介绍，明确了葡萄酒质量多分类任务，同时本部分还阐述了研究的目的与内容，给出了研究的技术路线

第二部分：相关分类方法介绍。本部分系统地介绍了研究中采用的多种分类方法，包括逻辑斯蒂回归、朴素贝叶斯、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升等经典模型，此外还探讨了不同模型在多分类任务中的理论优势及其适用场景。

第三部分：基于机器学习算法的实验过程。本部分是研究的核心内容，详细描述了实验的设计与实施过程，包括数据预处理、特征选择、模型训练与优化、以及评价指标的选定（如交叉验证得分、交叉熵损失、准确率及容忍度准确率）。实验对比了多种机器学习模型在葡萄酒质量数据集上的性能表现，并结合分类与聚类实验结果，对模型适用性进行了深入分析。

第四部分：总结与展望。研究最后对实验结果进行了总结，指出了不同模型的优劣势及其对葡萄酒质量预测任务的启示。同时对本研究的不足之处进行反思，并展望了未来在分类方法改进、数据增强及应用领域扩展方面的潜在研究方向。

1.3.3 技术路线

本文的技术路线如图 1-1 所示。

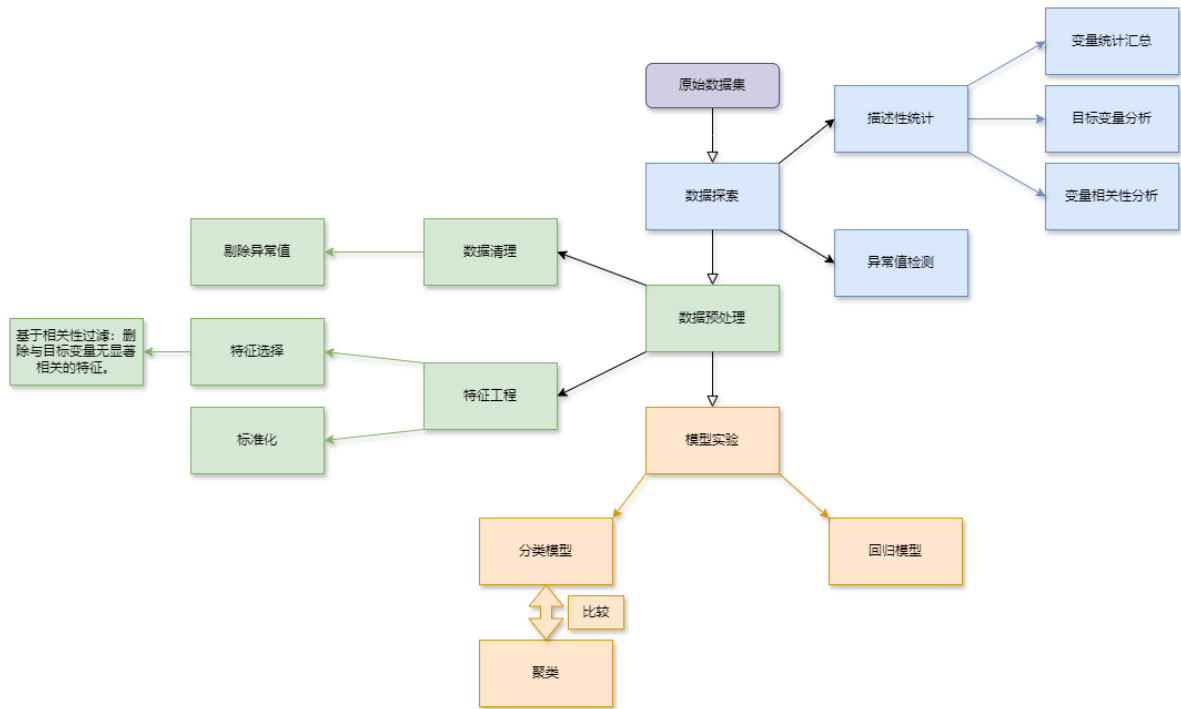


图 1-1 技术路线图

2 相关分类方法介绍

2.1 逻辑斯蒂回归

逻辑回归是用于分类的线性模型。虽然被称为“回归”，但它实际用于处理二分类问题，并使用 sigmoid 函数将特征映射为一个概率，表示数据为某类别的概率，因此对需要概率来帮助决策的任务很有帮助。

首先解释线性模型，即通过线性组合来进行预测的函数，向量形式如下

$$f(x) = \omega^T x + b \quad (2.1)$$

其中 x 为属性值的集合， w 是参数向量，表示属性权重， b 为偏置项， w 、 b 可通过模型学习来确定。

将线性模型推广可得到广义线性模型

$$f(x) = g^{-1}(\omega^T x + b) \quad (2.2)$$

其中 $g()$ 为联系函数。

逻辑回归就是 $g(\cdot)$ 为 sigmoid 函数时，

$$y = \frac{1}{1 + e^{-z}} \quad (2.3)$$

它将 z 值转化为一个接近 0 或 1 的值得到如下函数

$$y = \frac{1}{1 + e^{-(\omega^T x + b)}} \quad (2.4)$$

接着可将式 (2.4) 变化为

$$\ln \frac{y}{1-y} = \omega^T x + b \quad (2.5)$$

若将 y 视为 x 为正例的可能性，则 $1-y$ 使其为反例的可能性，两者比值

$$\frac{y}{1-y} \quad (2.6)$$

称为“几率”，反映 x 为正例的相对可能性，所以对几率取对数则得到“对数几率”

$$\ln \frac{y}{1-y} \quad (2.7)$$

由式 (2.4) 得知，实际上就是对数几率空间内的线性回归。

2.2 朴素贝叶斯

朴素贝叶斯法基于特征间独立和贝叶斯定理的假设。先计算目标的先验概率，然后使用贝叶斯定理来计算后验概率，并根据后验概率大小来决策。

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (2.8)$$

其中 X 为输入， Y 为后验概率最大输出。在算法中， $P(Y)$ 是事件发生的可能性，还是 Y 的先验概率， $P(X)$ 同理；条件概率为 $P(X|Y)$ ；后验概率为 $P(Y|X)$ ，表示某个因素引起事件发生的可能性。

朴素贝叶斯方法，计算简洁高效，计算量非常之小，训练和分类的速度很快，适用大规模数据集，对高维数据适用性好，还可以处理稀疏数据。但其存在一定缺陷，朴素贝叶斯分类器对数据特征要求相互独立，可现实中，这个情况是几乎不可能成立的。

总体来看，朴素贝叶斯是一个简洁高效的分类模型，但往往要根据数据集特征和分类目标做出改进。

2.3 深度随机森林

利用多层神经网络（通常是全连接层）对原始特征进行嵌入变换，生成高维、抽象的特征表征。

特征嵌入的函数形式如下：

$$h(x) = f(Wx + b) \quad (2.9)$$

其中 W 和 b 分别为神经网络的权重和偏置， $f(\cdot)$ 是激活函数。

之后是随机森林预测，公式如下：

$$y_i = \operatorname{argmax}_k \frac{1}{T} \sum_{t=1}^T I(f_t(h(x_i)) = k) \quad (2.10)$$

其中 T 是森林中树的数量， $f_t(\cdot)$ 表示第 t 棵树的预测结果， $I(\cdot)$ 为指示函数。

深度随机森林通过结合神经网络的特征提取能力和随机森林的高效分类能力，在多分类任务中表现优异。同时通过联合优化策略，深度随机森林能够在不显著增加模型复杂度的情况下提升分类性能，兼顾深度学习的表达能力和传统方法的高效性，是多分类任务中的一个重要方向。

2.4 极端随机树

Extra-Trees 属于基于树的集成学习方法，与随机森林类似，但在分裂点选择和样本使用上引入了更高的随机性，其目标是构建一组多样化的决策树，通过投票机制实现分类，设输入特征为 $x=[x_1, x_2, \dots, x_d]^T$ ，类别标签为 $y \in \{1, 2, \dots, K\}$ ，Extra-Trees 的目标是构建 T 棵极端随机树，通过投票机制统计每个类别样本的预测次数，具体公式如下。

$$P(y = k|x) = \frac{1}{T} \sum_{t=1}^T I(f_t(x) = k) \quad (2.11)$$

其中 $f_t(x)$ 是第 t 棵树对样本 x 的预测结果； $I(\cdot)$ 是指示函数，当 $f_t(x)=k$ 时， $I=1$ ，否则 $I=0$ ，然后将概率最大的类别作为最终预测结果。

$$\hat{y} = \operatorname{argmax}_k P(y = k|x) \quad (2.12)$$

2.5 梯度提升决策树

梯度提升决策树（GBDT）是提升方法（Boosting）的一种实现，通过迭代训练一组决策树，将每棵树的错误修正为下一棵树的目标，从而实现强大的预测能力。在多元分类任务中，GBDT 被扩展为适配多类别的预测，其核心思想是通过加权累加方式逐步逼近目标类别的真实分布，它通过多次迭代累加弱学习器的输出逼近目标函数。设输入特征为 x ，类别标签为 $y \in \{1, 2, \dots, K\}$ ，GBDT 的目标是预测每个样本属于 K 个类别之一的概率分布 $P(y=k|x)$ 。

GBDT 将模型表示为一组累加的决策树：

$$F_k(x) = \sum_{m=1}^M f_{k,m}(x) \quad (2.13)$$

其中 M 是决策树的数量， $f_{k,m}(x)$ 是第 m 棵树对类别 k 的预测， $F_k(x)$ 是模型对类别 k 的最终预测得分（logit 值），类别概率最终由 Softmax 函数计算得到：

$$P(y = k|x) = \frac{\exp(F_k(x))}{\sum_{j=1}^K \exp(F_j(x))} \quad (2.14)$$

然后将概率最大的类别作为最终预测结果

$$\hat{y} = \operatorname{argmax}_k P(y = k|x) \quad (2.15)$$

2.6 轻量级梯度提升机

轻量级梯度提升机 (LightGBM) 是一种基于梯度提升框架的高效算法, 专注于在大规模数据集和高维特征场景下提升训练速度与内存效率, 同时保留梯度提升决策树 (GBDT) 的强大性能。在多分类任务中, LightGBM 扩展了 GBDT 的框架, 通过优化数据结构和训练过程, 实现更快速的建模和高效的分类。

LightGBM 与 GBDT 的基本原理相同, 旨在通过迭代优化逐步逼近目标函数。在多分类任务中, 其目标是预测每个样本属于 K 个类别之一的概率分布 $P(y=k|x)$, 与传统 GBDT 的主要区别在于其对数据处理和树生长的优化。首先 LightGBM 引入了基于直方图的决策树构建方法, 通过将连续特征离散化为有限个区间, 大幅降低了分裂点搜索的复杂度, 先对每个特征 x_j , 将取值划分为 G 个区间。再在直方图上计算每个区间的梯度和损失, 选择最优分裂点, 其优化公式为:

$$Split\ Gain = \frac{(G_{left})^2}{H_{left} + \lambda} + \frac{(G_{right})^2}{H_{right} + \lambda} - \frac{(G_{total})^2}{H_{total} + \lambda} \quad (2.16)$$

其中 G 和 H 分别表示梯度和二阶梯度累积, λ 为正则化项。传统 GBDT 使用按层生长的树结构, 而 LightGBM 采用基于叶子的生长策略, 优先分裂能带来最大增益的叶子节点, 目标是最大化信息增益。

$$Leaf\ Gain = \frac{(G_{leaf})^2}{H_{leaf} + \lambda} \quad (2.17)$$

这一策略能更有效地适应数据分布, 提高分类任务的性能。

2.7 随机森林

随机森林模型是一种集成学习算法, 综合多个弱分类器的输出, 统筹做出预测, 一般以 决策树为弱分类器, 决策树的构造涉及特征选取、构造决策树、剪枝三步。随机森林在每个决策树的节点分裂时随机选择部分特征进行评估, 常用的特征选取方法包括 ID3、C4.5、CART。

信息增益 (ID3) 衡量特征对样本信息熵的降低程度, 信息增益更适合处理离散特征, 但对多值特征有偏好, 公式为:

$$g(D|A) = H(D) - H(D|A) \quad (2.18)$$

信息增益率 (C4.5) 修正信息增益对多值特征的偏好, 引入了特征值熵 $H(A)$, 其公式为:

$$g_r(D|A) = \frac{g(D|A)}{H(A)} \quad (2.19)$$

基尼指数（CART）：衡量样本集合的纯度，分裂节点时选择基尼指数最小的特征，其公式为：

$$Gini(D) = 1 - \sum_{k=1}^K p_k^2 \quad (2.20)$$

随机森林是集成学习 Bagging 的变体，训练过程中通过随机选择样本和属性，因此可以发现随机森林的弱分类器差异性及其来自样本和属性，可提升泛化能力。随机森林使用自助采样法，从包含 m 个样本的数据集 D 中进行有放回的抽样 m 次，形成新的数据集 D^* 。当样本数量趋于无穷，可计算出样本未被采集的概率：

$$\lim_{m \rightarrow \infty} \left(1 - \frac{1}{m}\right)^m = \frac{1}{e} \approx 0.368 \quad (2.21)$$

可看出元数据及中将有 36.8% 的样本未被采用，从未被用来训练，这将给机器学习带来较大差异，再结合少数服从多数来进行分类提高准确性。

2.8 支持向量机

支持向量机（SVM）是一种用于分类、回归和异常检测的监督学习模型，支持向量分类（SVC）是 SVM 在分类任务中的具体应用，旨在通过寻找最优的决策边界（超平面），将数据样本划分为不同类别。在多分类任务中，SVC 通过扩展一对多、一对一策略方法实现对多个类别的区分。

SVM 的目标是找到一个超平面 $w^T x + b = 0$ ，最大化边界间隔，同时确保样本被正确分类，其超平面公式如下：

$$w^T x + b = 0 \quad (2.22)$$

其中 w 是法向量，表示超平面的方向； b 是偏置，用来控制超平面与原点的距离，而为了处理非线性数据或允许分类错误，引入松弛变量 ξ_i ，优化目标变为：

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (2.23)$$

其中 C 是正则化参数，用于平衡边界间隔与分类错误的权重，而当数据无法通过线性超平面分割时，SVM 使用核函数将数据映射到高维特征空间，使其在高维空间中线性可分，常用的高斯核函数为：

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (2.24)$$

支持向量分类（SVC）通过最大化分类边界间隔实现分类，其核心特点是利用核方法解决非线性问题。在多分类任务中，SVC 通过一对多或一对一策略扩展为多分类模型，适用于高维和复杂分布数据，然而 SVC 的计算复杂度较高，且对超参数敏感，需要通过调优找到最佳设置。

2.9 极限梯度提升

极限梯度提升（XGBoost）是一种基于梯度提升框架的改进算法，通过引入系统优化和正则化技术，提升了模型的训练速度、预测精度以及对过拟合的控制能力。在多分类任务中，XGBoost 能高效处理大规模数据，生成具有良好泛化能力的模型。

为了优化目标函数，XGBoost 使用二阶泰勒展开，将损失函数近似为：

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^N \left[g_i f_{k,t}(x_i) + \frac{1}{2} h_i f_{k,t}^2(x_i) \right] + \Omega(f_{k,t}) \quad (2.25)$$

其中 g_i 和 h_i 分别是一阶梯度和二阶梯度， $\Omega(f_{k,t})$ 是正则化项，用于控制模型复杂度，XGBoost 每次分裂选择增益最大的分裂点，分裂增益计算公式为：

$$Gain = \frac{1}{2} \left[\frac{(G_L)^2}{H_L + \lambda} + \frac{(G_R)^2}{H_R + \lambda} - \frac{(G_{total})^2}{H_{total} + \lambda} \right] - \gamma \quad (2.26)$$

其中 G_L , G_R 是左右子节点的一阶梯度和； H_L , H_R 是左右子节点的二阶梯度和； γ 和 λ 是正则化参数。

2.10 评价指标

各种评价指标对评估机器学习模型是至关重要的，一些评价指标如下：

表 2-1 分类混淆矩阵

| 真实情况 | 预测结果 | |
|------|------|----|
| | 正例 | 反例 |
| 正例 | TP | FN |
| 反例 | FP | TN |

(1) Acc (Accuracy, 准确率)：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.27)$$

(2) P (Precision, 精确率) :

$$P = \frac{TP}{TP + FP} \quad (2.28)$$

(3) R (Recall, 召回率) :

$$R = \frac{TP}{TP + FN} \quad (2.29)$$

(4) F1 分数 (F1 Score) 是精确率和召回率的调和平均数:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (2.30)$$

(5) 平均交叉验证得分为:

$$Mean\ CV\ Score = \frac{1}{K} \sum_{i=1}^K Accuracy_i \quad (2.31)$$

(6) 交叉熵损失 (Negative Log Loss) 为:

$$Log\ Loss = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log P(y = k|x_i) \quad (2.32)$$

(7) 宏平均 (Macro Average) 为:

$$Macro\ Avg = \frac{1}{K} \sum_{k=1}^K Metric_k \quad (2.33)$$

(8) 容忍度准确率 (Tolerance Accuracy) 为:

$$TA = \frac{N_T}{N} \quad (2.34)$$

其中 TA 容忍度准确率, NT 表示预测结果与真实标签的差值在容忍范围 T 内的样本数量 (即 $|\hat{y}_i - y_i| \leq T$ 的样本数), N 表示样本总数, 本研究设置的容忍度 T=1。

3 基于机器学习算法的实验过程

3.1 数据探索

3.1.1 特征变量分析

特征变量分析可以较全面了解数据的分布特性和基本结构，有助于发现数据质量问题，如缺失值、异常值或重复值，揭示变量间的相关性，为特征选择和特征工程提供依据，同时统计汇总能够评估模型对数据的适应性。通过统计汇总，还可以提高数据的可解释性，明确数据中潜在的规律，为后续的降维、聚类或建模奠定坚实基础，这一过程不仅是优化数据质量的必要环节，也是提升模型性能和决策效率的重要手段。

在该步骤中，本研究绘制了红白葡萄酒 11 个特征变量的直方图和箱型图。

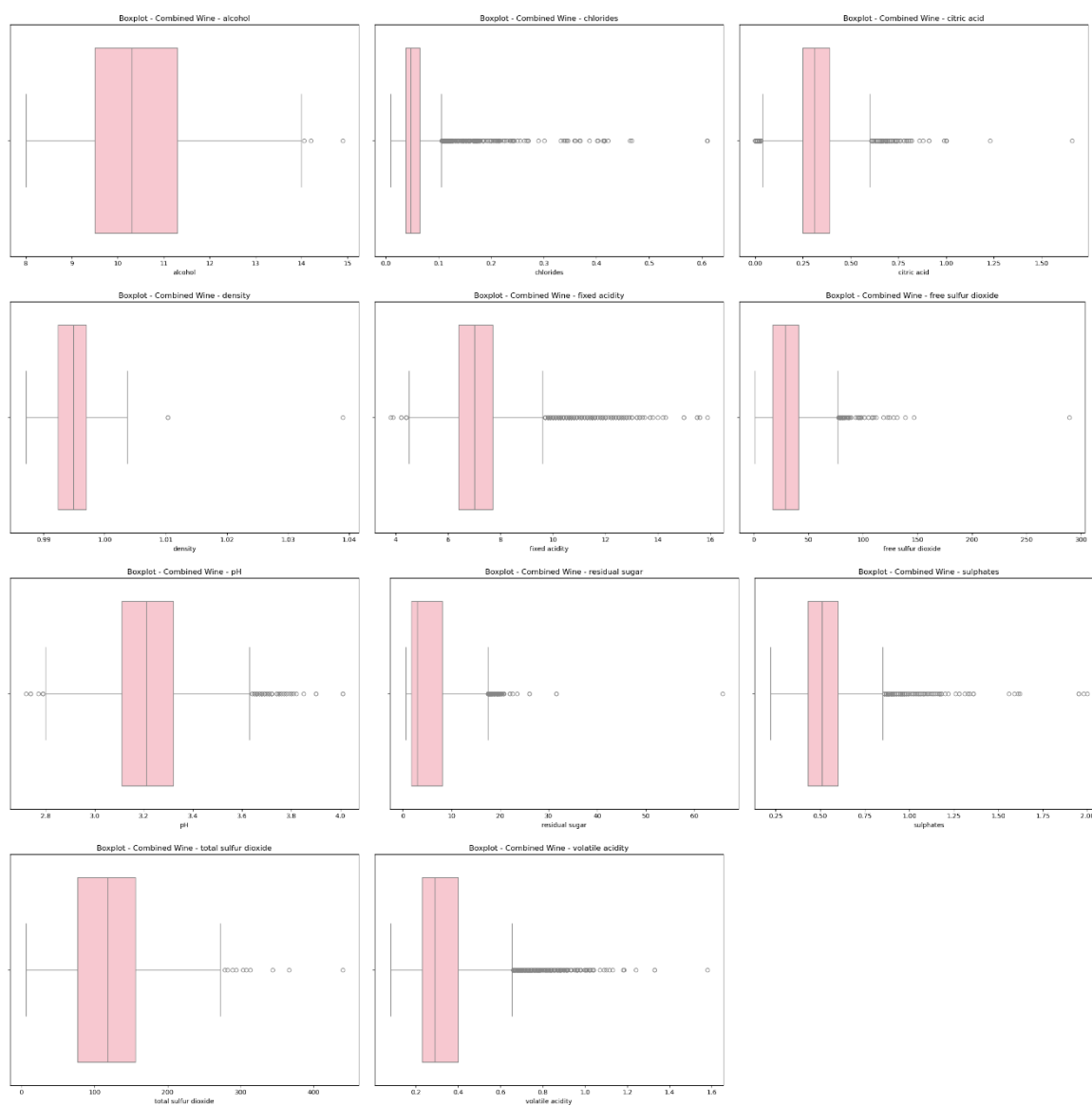


图 3-1 特征变量箱型图汇总

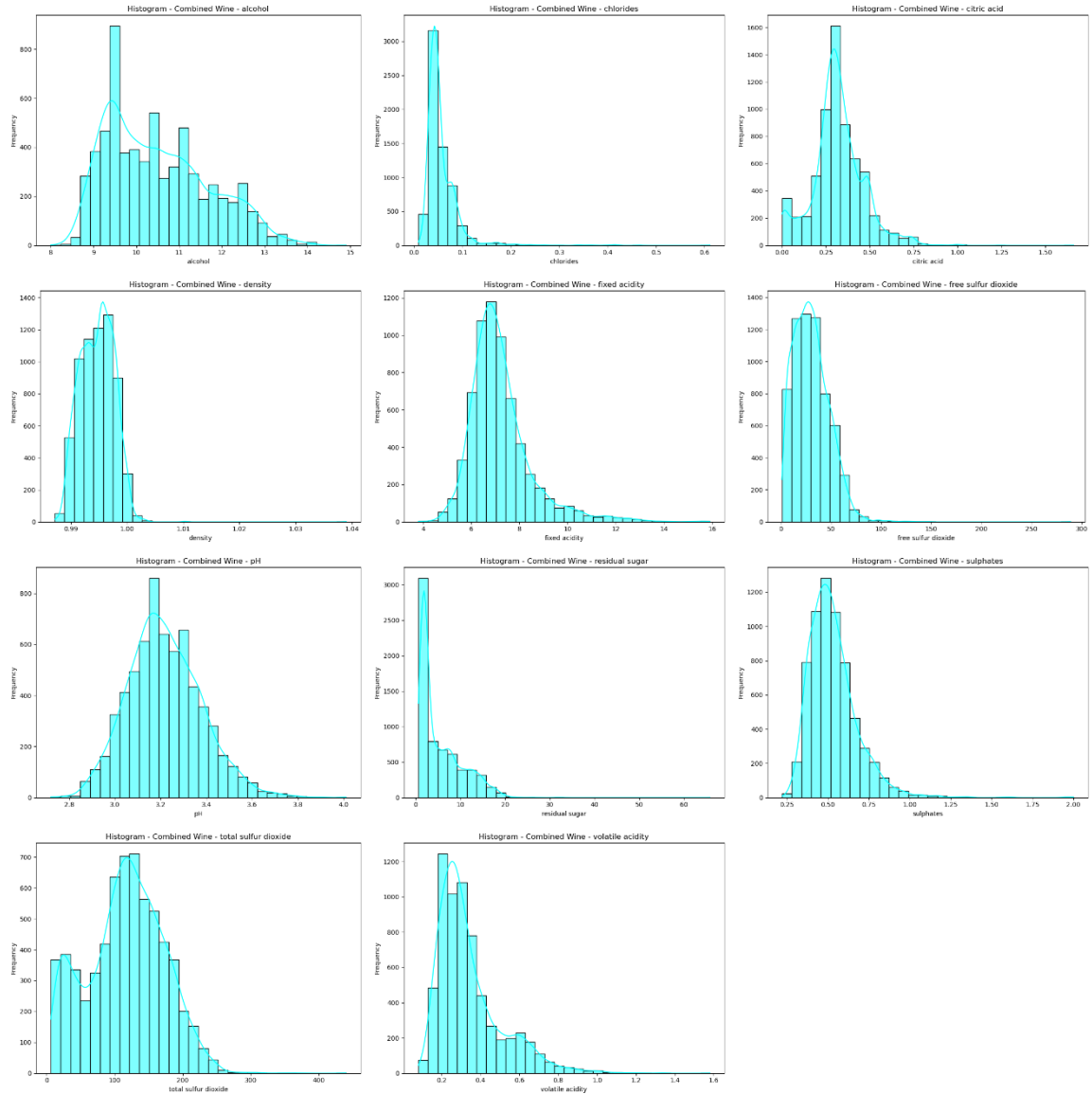


图 3-2 特征变量直方图汇总

3.1.2 目标变量分析

目标变量分析是数据探索中不可或缺的一部分，其核心在于全面理解预测目标的特性，为建模和数据处理提供坚实的基础。通过分析目标变量的分布形态，可以揭示集中趋势、离散程度以及潜在的异常值和缺失值，从而发现数据质量问题并采取相应的处理策略，例如分类任务中可以通过类别分布识别类别不平衡问题，而在回归任务中，偏态或多峰分布可能提示需要数据变换以提高模型效果，此外目标变量分析有助于评估任务复杂性，例如类别数量较多或类别边界模糊会增加分类难度，从而指导特定建模策略的选择。同时目标变量的特性还可以用于选择适当的模型评价指标，如类别不平衡时的宏平均 F1 分数或回归任务中的均方误差等。在特征工程方面，目标变

量与特征变量的相关性分析可以帮助筛选关键特征，为模型构建提供方向。通过目标变量分析，不仅可以优化建模策略，还能提升数据的可解释性和预测的科学性，为最终模型的准确性和鲁棒性奠定基础。

红白葡萄酒数据集的目标变量箱型图与直方图如图 3-3 所示。

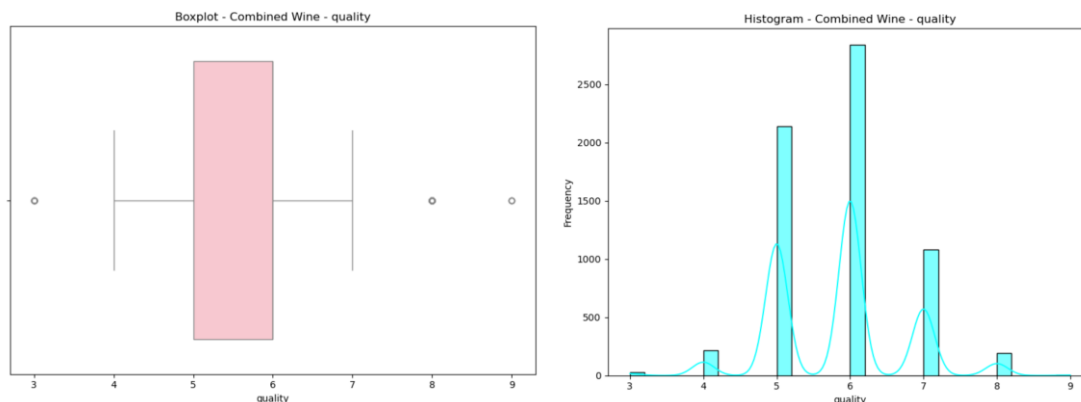


图 3-3 目标变量箱型图与直方图

目标变量的分布从箱型图和直方图中可以看出明显的集中趋势和类别不平衡特性。大多数样本的评分集中在 5 和 6，这两个类别的样本数量远远多于其他评分，如 3、8 和 9。箱型图也显示出离群点，这些极端评分虽然数量较少，但具有实际意义，如反映评分机制中的极端情况或特殊数据分布，因此在建模过程中本研究谨慎处理这些离群点，避免删除。

目标变量的类别分布还展示了一定的顺序性，例如从评分 3 到 9 的依次递增。这种特性提示我们可以在分类任务中引入容忍度，将模型预测值与真实值的偏差限制在一定范围内，提升分类任务的实际表现。此外轻微的偏态分布表明，虽然评分大致集中在中间部分（如 5 和 6），但高评分（如 8 和 9）数量较少且分布较远，这可能会对模型的训练和预测提出挑战。

总体来看目标变量的类别不平衡、离散性和顺序性对多分类任务提出了一定的难点，需要通过调整样本分布（如过采样或类别权重调整）、设计合理的损失函数，以及选择更合适的评价指标（如加权 F1 分数）来优化模型性能，同时考虑目标变量的顺序性和评分机制，可以尝试结合回归模型或容忍度策略进一步提升分类效果。

3.1.3 变量相关性分析

变量相关性分析是数据探索中的重要步骤，用于揭示特征变量之间以及特征变量与目标变量之间的关系，为特征选择和模型设计提供科学依据。通过相关性分析，可

以识别对目标变量具有显著预测力的关键特征，并剔除冗余或无关的特征，从而简化模型，提升训练效率和预测性能，例如计算皮尔逊相关系数或斯皮尔曼相关系数可以量化特征与目标变量的相关性，筛选出重要特征，同时通过分析特征间的强相关性，避免信息冗余和过拟合，此外相关性分析还能增强模型的解释性，帮助理解哪些特征对预测任务最重要，为业务优化或决策提供有力支持，因此相关性分析在提升模型性能和增强数据分析的科学性方面具有重要意义。

本研究中的红白葡萄酒数据集的变量热力图如图 3-4 所示。

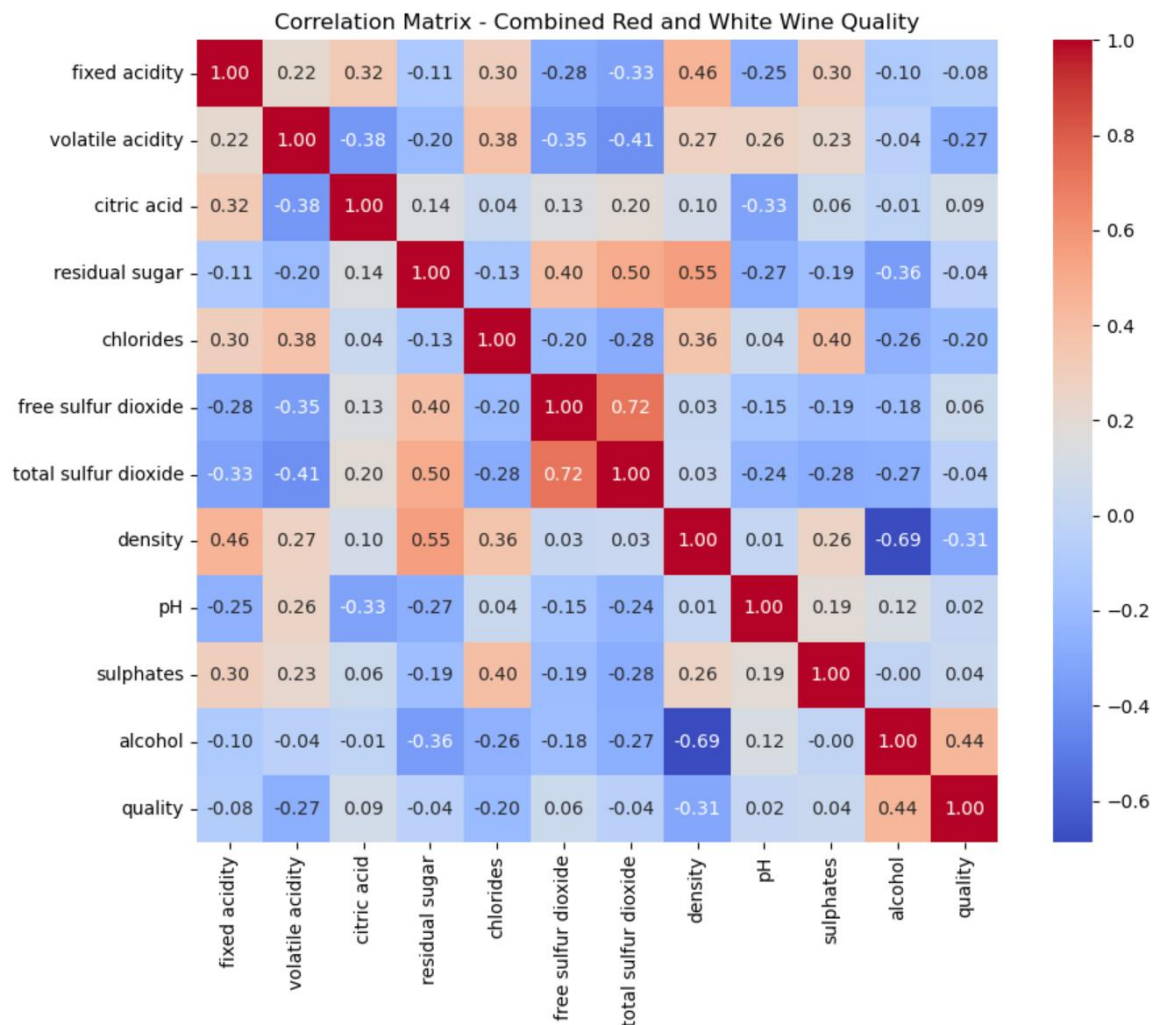


图 3-4 变量热力图

这张相关性矩阵图揭示了目标变量与特征变量之间的复杂关系，以及特征变量之间的交互作用。其中酒精含量与质量评分相关系数为 0.44，具有显著的正相关性，表明酒精是影响评分的重要因素，而挥发性酸度和密度则表现出负相关性，提示过高的酸度或密度可能降低葡萄酒质量。特征变量之间的高相关性，如自由二氧化硫和总二

氧化硫，相关系数为 0.72，反映了潜在的冗余信息，需在传统建模中进行降维或特征筛选，而深度学习模型则可以通过非线性变换和特征嵌入直接捕获其潜在关系。同时大多数特征与质量的相关性较弱，说明质量评分受多维因素影响，单一特征难以解释，需要更复杂的模型如深度学习，来挖掘特征间的隐含非线性关系。此外这张图对葡萄酒质量评估提供了业务启示，如酒精含量的正向作用和酸度平衡的负向影响，为生产改良和模型设计提供了有力支持。

3.1.4 异常值检测

本研究采用 IQR 方法进行异常值检测并移除，先计算四分位数：

$$Q1 = \text{quantile}(0.25) \quad (3.1)$$

$$Q2 = \text{quantile}(0.75) \quad (3.2)$$

由（3.1）和（3.2）式可得四分位距 IQR：

$$IQR = Q3 - Q1 \quad (3.3)$$

接着使用 1.5 倍的 IQR 来定义异常值的上下界，这个倍数是统计学中常用的经验值，然后过滤数据，令保留的数据满足以下（3.4）式子的条件，但由于本研究数据集的特殊性，于是保留了质量目标变量的异常值。

$$Q1 - 1.5 \times IQR \leq x \leq Q3 + 1.5 \times IQR \quad (3.4)$$

3.2 数据预处理

3.2.1 数据清洗

由于原数据集作者声明没有缺失值，且经检查无误，于是按上文提到的 IQR 方法进行异常值检测，并清理质量目标变量列以外的变量异常值，具体效果如表 3-1 所示。

表 3-1 移除特征变量异常值数目表

| 特征变量 | 移除异常值数目 |
|----------------------|---------|
| fixed acidity | 357 |
| volatile acidity | 374 |
| citric acid | 478 |
| residual sugar | 37 |
| chlorides | 393 |
| free sulfur dioxide | 54 |
| total sulfur dioxide | 7 |
| density | 0 |
| pH | 60 |
| sulphates | 141 |
| alcohol | 0 |

3.2.2 特征工程

特征工程是从原始数据中提取描述性强的特征以优化模型性能的过程。这一过程非常关键，因为优秀的特征可以显著降低模型复杂性和参数调整的需要，使得简单模型也能达到高效的性能，所以好的特征不仅提升模型的表现，还增强了模型的灵活性和维护的便捷性。

本研究的整个特征工程分为特征选择和标准化两个部分，根据特征热力图，选择跟目标标量相关性分数大于等于 0.05 的变量保留，并按照以下公式对红白葡萄酒数据集进行标准化。

$$z = \frac{x - \mu}{\sigma} \quad (3.5)$$

其中 μ 为均值， σ 为方差。

3.3 模型训练和调参

3.3.1 逻辑斯蒂回归

逻辑斯蒂回归模型在实验中的表现不佳，具体体现在平均交叉验证得分较低、交叉熵损失较高、准确率接近随机猜测水平，说明模型无法有效捕捉数据的特征模式。然而容忍度准确率相对较高，表明模型在预测的类别附近具备一定的判断能力，这反映了模型在容错范围内的潜在可用性。整体来看，逻辑斯蒂回归可能受限于数据特征与模型假设的线性关系较弱，以及类别不平衡等问题，实验结果如表 3-2 所示。

表 3-2 逻辑斯蒂回归模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|--------|--------------|--------|--------|--------|
| 逻辑斯蒂回归 | 0.2602 | 1.7219 | 24.73% | 52.72% |

3.3.2 朴素贝叶斯

朴素贝叶斯模型在实验中表现出一定的分类能力，尤其在容忍度范围内 (± 1) 的预测上表现优异，如伯努利分类器、多项式分类器和高斯分类器的容忍度准确率分别达到 92.24%、93.91%和 91.37%。然而整体分类准确率和平均交叉验证得分较低，分别为 46.92%、45.32%和 48.01%，交叉熵损失也相对较高，表明模型在特定类别上的区分能力有限。总体来看，这可能与模型假设（如特征独立性假设）与数据特征间的适配程度不足有关，实验结果如表 3-3 所示。

表 3-3 朴素贝叶斯模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|--------|--------------|--------|--------|--------|
| 伯努利分类器 | 0.4557 | 1.1892 | 46.92% | 92.24% |
| 多项式分类器 | 0.4569 | 1.2199 | 45.32% | 93.91% |
| 高斯分类器 | 0.4784 | 1.2991 | 48.01% | 91.37% |

3.3.3 深度随机森林

深度随机森林模型在实验中的表现较为一般，平均交叉验证得分为 0.4967，准确率为 51.78%，略高于随机猜测水平，说明模型在捕捉数据特征模式上有一定能力但表现有限。交叉熵损失较高，为 6.2097，反映了模型在输出类别概率分布时存在较大偏

差。容忍度准确率达到 84.7%，表明模型在预测类别的容忍范围内 (± 1) 具备一定的判断能力。整体来看深度随机森林在容错范围内表现尚可，但准确率和概率预测质量有待进一步提升，未来可以考虑调整超参数或结合其他优化技术来改善模型性能，实验结果如表 3-4 所示。

表 3-4 深度随机森林模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|--------|--------------|--------|--------|--------|
| 深度随机森林 | 0.4967 | 6.2097 | 51.78% | 84.7% |

3.3.4 极端随机树

极端随机树模型在实验中的表现较为优异，平均交叉验证得分为 0.6432，准确率达到 64.9%，说明模型能够较好地捕捉数据特征模式，具备一定的分类能力。交叉熵损失为 0.9387，相对较低，表明模型在输出类别概率分布时具有较高的置信度和稳定性。此外容忍度准确率达到 94.78%，进一步体现了模型在容错范围内 (± 1) 具有较强的预测能力。总体来看，极端随机树在多分类任务中表现出色，适合处理特征较多且类别分布不平衡的问题，同时其预测结果在容忍范围内具备较高的可靠性，实验结果如表 3-5 所示。

表 3-5 极端随机树模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|-------|--------------|--------|-------|--------|
| 极端随机树 | 0.6432 | 0.9387 | 64.9% | 94.78% |

3.3.5 梯度提升决策树

梯度提升决策树模型在实验中的表现较为稳健，平均交叉验证得分为 0.5782，准确率为 56.27%，说明模型能够较好地学习数据特征并进行有效分类。交叉熵损失为 1.0660，表明模型在类别概率预测方面具有一定的置信度，但仍存在进一步优化的空间。容忍度准确率达到 93.47%，反映了模型在容错范围内 (± 1) 具备较强的预测能力。总体来看梯度提升决策树在多分类任务中表现较好，既能够有效捕捉复杂特征模式，又在容错范围内提供了较高的可靠性，但仍可通过优化参数进一步提升分类和概率预测性能，实验结果如表 3-6 所示。

表 3-6 梯度提升决策树模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|-------------|--------------|--------|--------|--------|
| 梯度提升决策 树 | 0.5782 | 1.0660 | 56.27% | 93.47% |

3.3.6 轻量级梯度提升机

轻量级梯度提升机在实验中展现了较为平衡的性能，其平均交叉验证得分为 0.5723，分类准确率为 54.68%，表明模型具备一定的预测能力，但在特征学习上尚有改进余地。交叉熵损失为 1.0376，体现了模型在预测类别概率时的合理表现，具有一定的置信水平。容忍度准确率达到 91.01%，说明在容忍范围（ ± 1 ）内，模型能较好地捕捉目标类别的分布特征。综合来看，轻量级梯度提升机在多分类任务中兼顾了效率与效果，适合大规模数据场景，但仍可通过优化参数或数据预处理提升整体性能，实验结果如表 3-7 所示。

表 3-7 轻量级梯度提升机模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|--------------|--------------|--------|--------|--------|
| 轻量级梯度提 升机 | 0.5723 | 1.0376 | 54.68% | 91.01% |

3.3.7 随机森林

随机森林模型在实验中表现良好，平均交叉验证得分为 0.6428，分类准确率为 65.99%，展现了其在多分类任务中的强大学习能力。交叉熵损失为 1.0268，说明模型在类别概率预测上具有较高的置信度和稳定性。容忍度准确率达到 94.63%，表明模型在容错范围（ ± 1 ）内对类别的预测能力相当出色。总体而言随机森林在捕捉复杂特征模式和处理类别不平衡方面具有显著优势，是一个可靠的多分类任务解决方案，但在提升效率或进一步降低损失方面仍有优化潜力，实验结果如表 3-8 所示。

表 3-8 随机森林模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|------|--------------|--------|--------|--------|
| 随机森林 | 0.6428 | 1.0268 | 65.99% | 94.63% |

3.3.8 支持向量机

支持向量机模型在实验中表现中规中矩，平均交叉验证得分为 0.5521，分类准确率为 54.53%，说明其在多分类任务中具有一定的预测能力，但未达到最佳状态。交叉熵损失为 1.0577，表明模型在概率预测方面的输出尚有改进空间。容忍度准确率为 93.47%，显示支持向量机在容错范围（ ± 1 ）内具备较高的预测可靠性。总体来看支持向量机在多分类任务中对边界样本的处理较为有效，但可能因类别分布不平衡或特征选择不足而导致性能受限，可以通过优化核函数或调整超参数进一步提升模型表现，本研究并未进行超参数优化，对此的实验有待补充，实验结果如表 3-9 所示。

表 3-9 支持向量机模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|-------|--------------|--------|--------|--------|
| 支持向量机 | 0.5521 | 1.0577 | 54.53% | 93.47% |

3.3.9 极限梯度提升

极限梯度提升模型在实验中表现出色，平均交叉验证得分为 0.6099，分类准确率达到了 61.28%，显示其在多分类任务中的可靠性和稳定性。交叉熵损失为 0.9752，表明模型在类别概率预测方面具有较高的精确度。容忍度准确率为 94.42%，进一步凸显了模型在容错范围（ ± 1 ）内的强大预测能力。总体来看极限梯度提升模型在捕捉复杂非线性特征和处理类别不平衡问题方面表现优异，是解决多分类问题的强有力工具，同时仍有优化超参数和提升效率的潜力，实验结果如表 3-10 所示。

表 3-10 极限梯度提升模型实验结果

| 模型 | 平均交叉验证 得分 | 交叉熵损失 | 准确率 | 容忍度准确率 |
|--------|--------------|--------|--------|--------|
| 极限梯度提升 | 0.6099 | 0.9752 | 0.6128 | 0.9442 |

3.4 模型对比分析

实验结果如表 3-11 所示。

表 3-11 模型容忍度准确率对比

| 模型 | 容忍度准确率 |
|-------------------------|--------|
| Logistic regression | 52.72% |
| Naive Bayes bernoulli | 92.24% |
| Naive Bayes multinomial | 93.91% |
| Naive Bayes gaussian | 91.37% |
| Deep RF | 84.7% |
| Etr | 94.78% |
| GBDT | 93.47% |
| LightGBM | 91.01% |
| RF | 94.63% |
| SVC | 93.47% |
| XGBoost | 94.42% |

从容忍度准确率的实验结果来看，不同模型在多分类任务中的表现差异显著。逻辑斯蒂回归的容忍度准确率最低，仅为 52.72%，表现明显逊色，这反映了其在处理类别间复杂关系和容错能力方面的局限性。在朴素贝叶斯模型中，多项式分类器的容忍度准确率最高，为 93.91%，其次是伯努利分类器和高斯分类器，分别为 92.24%和 91.37%，表明朴素贝叶斯模型在类别预测容错能力上具备较强的鲁棒性，尤其适用于特定特征分布的任务。集成学习方法中的极端随机树（94.78%）、随机森林（94.63%）以及极限梯度提升（94.42%）在容忍度准确率上表现最为优异，显示出它们在容错范围内的强大预测能力。梯度提升决策树（93.47%）、支持向量机（93.47%）以及轻量级梯度提升机（91.01%）也都展现了良好的容错能力。深度随机森林模型的容忍度准确率为 84.7%，虽然略低于其他集成方法，但仍表现出了较高的容错性。总体来看集成学习方法在容错能力上表现优异，尤其适用于多分类任务中的复杂数据分布。模型选择应结合具体任务需求，平衡容错能力和计算效率。

3.5 聚类和降维

3.5.1 KMeans 聚类

对实验数据进行 PCA 降维后在 KMeans 聚类中的分类结果如图 3-5 所示,通过主成分 1 和主成分 2 的二维投影直观呈现了三个聚类类别。聚类结果显示,紫色 Cluster 0 主要分布在左侧,青绿色 Cluster 1 集中于上方,而黄色 Cluster 2 位于右下方,各簇之间整体上具有一定的分离度,然而 Cluster 1 和 Cluster 2,以及 Cluster 0 和 Cluster 2 在部分边界区域存在重叠,表明类别之间的分离性尚不完全清晰,尽管 PCA 降维有效降低了数据维度并保留主要信息,但可能导致部分特征丢失,从而影响聚类效果。

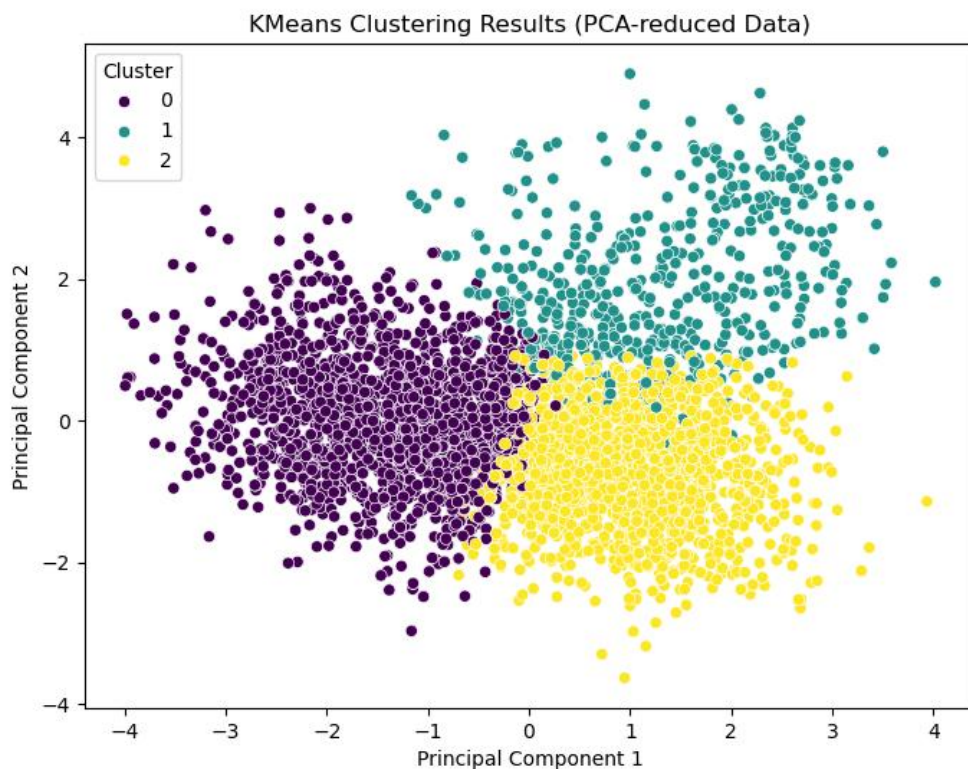


图 3-5 PCA 降维后的 Kmeans 聚类结果图

3.5.2 DBSCAN 聚类

DBSCAN（基于密度的聚类算法）通过指定半径参数 ϵ 和最小点数 minPts ，将数据点划分为密度可达的簇，同时将不满足条件的点视为噪声点。在本实验中，DBSCAN 对 PCA 降维后的数据进行了聚类，实验结果如图 3-6 所示，大多数数据点被划分为一个紫色大簇，表明这些点的密度较高，然而少量数据点被分为小簇或视为噪声点，反映出局部区域可能存在异常分布。整体结果显示 DBSCAN 能够有效识别密集簇，但由于参数选择可能导致对稀疏区域的划分不够细致。此外 PCA 降维可能压缩了高维数据的区分能力，从而影响了 DBSCAN 对数据内部结构的敏感性，未来可通过优化参数和降维方法进一步提升聚类效果。

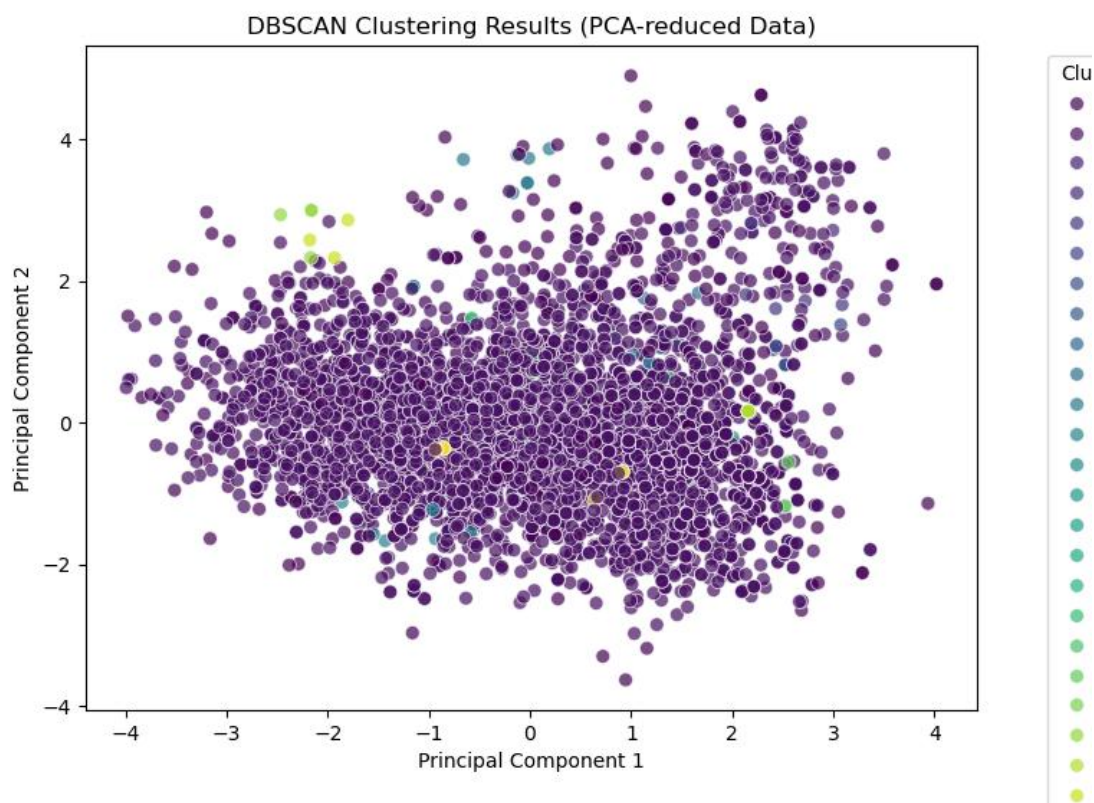


图 3-6 PCA 降维后的 DBSCAN 聚类结果图

4 总结与展望

4.1 全文总结

本研究以葡萄酒质量数据集为基础，评估多种机器学习模型在多分类任务中的性能。通过对逻辑斯蒂回归、朴素贝叶斯、深度随机森林、极端随机树、梯度提升决策树、轻量级梯度提升机、随机森林、支持向量机和极限梯度提升等模型的实验对比，结果显示集成学习模型（如随机森林、极端随机树、极限梯度提升）在准确率和容忍度准确率上表现突出，尤其在处理类别不平衡问题时展现出优越的鲁棒性。此外通过降维后聚类实验，KMeans 效果较好，而 DBSCAN 对密度参数较为敏感，本研究验证了集成学习模型在多分类任务中的优势，为葡萄酒质量预测提供了数据支持，并为未来优化模型和算法改进提供了参考方向。

4.2 展望

由于时间原因，本研究中大量模型没有进行细致的调参，没有将精度提升到这个模型应有的水平，另外本研究中还有大量的实验数据和评价指标做出了但没有进行可视化对比，以后时间充裕时，会在 github 仓库中进行可视化图表以及复现步骤的补充。

致 谢

在本次课程的学习过程中，我深刻体会到了机器学习的魅力与挑战。这门课程不仅拓宽了我的知识视野，还培养了我在数据处理与模型分析方面的实践能力。在此衷心感谢老师的悉心指导和无私分享，严谨的教学态度和丰富的专业知识让我受益匪浅。除此之外，也感谢课程为我提供了利用高质量教学资源 and 进行实验的机会，使我能够将理论知识有效应用于实际问题。这段学习经历将成为我未来研究与实践的宝贵财富，非常感谢！机器学习课程自开课以来，就需要一代代课程团队不断打磨，希望咱们学校的机器学习课程能够越办越好！