Operations Research and Fuzziology 运筹与模糊学, 2023, 13(3), 2550-2564
Published Online June 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.133256

# 机器学习模型在白葡萄酒质量评价中的应用

**柴　桦**

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2023年5月23日；录用日期：2023年6月25日；发布日期：2023年6月30日

## 摘　要

**传统的葡萄酒质量检测由专业品酒师进行鉴评完成，存在检测成本高、周期长、主观臆断等缺点。建立一套客观、有效的葡萄酒质量评价体系，实现对葡萄酒质量的快速、批量检验是必要且重要的。本文对基于Logistic回归分析和随机森林两种白葡萄酒质量检测方法进行比较研究，选取了4898个白葡萄酒样本，通过混淆矩阵与十折交叉验证后，得出随机森林模型在测试集精确度及训练集精确度均优于Logistic模型，测试集精确度均值为88.48454%，相比于Logistic回归模型提高了8.36个百分点。本文使用机器学习模型为白葡萄酒的评价体系提供了一种快捷、准确且科学合理的方法。**

## 关键词

**随机森林，Logistic回归，十折交叉，混淆矩阵，二分类**

# Application of Machine Learning Model in Quality Evaluation of White Wine

**Hua Chai**

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: May 23rd, 2023; accepted: Jun. 25th, 2023; published: Jun. 30th, 2023

## Abstract

**Traditional wine quality testing is conducted by professional wine tasters, which has drawbacks such as high testing costs, long testing cycles, and subjective assumptions. It is necessary and important to establish an objective and effective wine quality evaluation system to achieve rapid and batch inspection of wine quality. This paper conducts a comparative study of two white wine quality detection methods based on Logistic regression analysis and random forest, selects 4898 white wine samples, and through the confusion matrix and ten fold cross validation, it is concluded that the precision of random forest model in the test set and training set is superior to the Logistic**

文章引用: 柴桦. 机器学习模型在白葡萄酒质量评价中的应用[J]. 运筹与模糊学, 2023, 13(3): 2550-2564.
DOI: 10.12677/orf.2023.133256

**model, and the average precision of the test set is 88.48454%, which is 8.36 percentage points higher than the Logistic regression model. This article provides a fast, accurate, and scientifically reasonable method for evaluating white wine using machine learning models.**

## Keywords

**Random Forest, Logistic Regression, Ten Fold Cross, Confusion Matrix, Binary Classification**



## 1. 研究背景

葡萄酒，即以葡萄为原材料经多种微生物发酵酿造而成的一款果酒[1]，葡萄酒中含有大量对调节人体健康有益的营养成分，适量引用葡萄酒对调节人体新陈代谢以及促进血液循环具有良好的作用[2]。近年来，随着民族、文化自信等因素助力国产品牌崛起，中国葡萄酒品质与世界接轨，屡次斩获国际性大奖，越来越多的消费者和经销商关注并选择国产葡萄酒。国产葡萄酒收割存量市场持续加速，“国产”替代“进口”的趋势明显。同时，伴随着进口葡萄酒的持续下滑，国产葡萄酒的替代效应越发明显。预计 2023 年，我国葡萄酒行业国产替代的速度将进一步加快。中国葡萄酒未来市场潜力大。一方面，中国葡萄酒市场消费呈年轻化趋势，消费能力快速提升，葡萄酒文化进一步普及，随着国民收入的增加，生活水平的提高，中国葡萄酒市场潜力巨大。另一方面，国家陆续出台了相关政策，鼓励发展葡萄酒行业，尤其是鼓励葡萄酒国产化，利好葡萄酒行业发展，预计中国葡萄酒行业市场规模也将稳步增长。目前国际上有二十多套评分体系，如美国葡萄酒协会(AWS)评分表、美国戴维斯(Davis)评分表、意大利评分表、OIV 评分表、法国酿酒师协会评分表等等。各个国家的饮食习惯和口味喜好不一，除了人工品尝的方式，葡萄酒的各项理化数据也应该是衡量质量的重要标准，因此一套科学合理的葡萄酒质量评估方法对于当前市场的重要性不言而喻。

## 2. 研究现状

对葡萄酒的评分，国内主要还是依靠品酒师的品尝感受来进行评价，再依靠各种质谱仪器来检测酒中各种化学成分含量来对葡萄酒质量进行综合评定。除此之外，部分质检部门也提出一套根据葡萄酒的特征指标，通过传统数理统计分析理论来进行更为科学的评判标准。近年来在实际应用中将层次分析法(Analytical Hierarchical Process，简称 AHP)应用于产品服务质量评定上，相似的也有不少学者提出应用于葡萄酒质量评估上。层次分析法被认为是将一个复杂的多目标决策问题作为一个系统，将目标分解为多个目标或准则，进而分解为多指标(或准则、约束)的若干层次，通过定性指标模糊量化方法算出层次单排序(权数)和总排序，以作为目标(多指标)、多方案优化决策的系统方法。此方法在一定程度上可以对葡萄酒产品做出综合的质量评估，也得到了众多专家的一致认可。但层次分析法[3]本身也存在局限性，主要表现为：1) 对决策目标的判定非常依赖评判专家的知识水平；2) 层次分析法[4]具有很高的主观性，因为它包含的虚拟评估取决于参数排名中的专家经验水平。基于以上原因，对质量评价结果有着较大误差[5]。

国外葡萄酒的生产和酿造技术已达到娴熟水平。在酒类鉴别方面，除了运用传统的品酒师品尝的方法外，早已使用现代科学仪器来分析酒类的各项化学成分，比如气相色谱仪、高液压相色谱仪等，以追求更科学更全面的对葡萄酒质量评价[6]。上世纪 80 年代，美国科研人员就提出将 PLS 应用于分析葡萄

酒各项化学成分和含量。国内方面，李志华设计了基于主要理化指标的分析软件来自动分析葡萄酒的质量，提升了葡萄酒产品质量[7]。张志然等通过对比不同葡萄酒产品感官评品数据与各类理化因子的关联性，利用因子贡献分析得到影响葡萄酒感官的主要理化因素[8]，唐文龙等通过分解葡萄酒的产品质量表达步骤，从目标受众、沟通媒介、沟通内容和沟通方法等方面分析了如何更好地构建葡萄酒产品质量的市场表达体系[9]，李记明等将统计学方法应用于葡萄酒质量分析与评价，可以更加清楚地了解葡萄酒成分与感官质量之间的关系[10]。也有更多的研究人员对人工神经网络关注越加密切，也有科研人员提出运用神经网络强大的逻辑分类能力用于鉴定葡萄酒分类上。但运用机器学习模型研究白葡萄酒质量评价还比较少，基于此，本文运用机器学习模型对白葡萄酒质量评价体系进行研究分析。

## 3. 数据来源与指标设计

本文所用的数据来自 UCL 机器学习数据库，数据集为白葡萄酒的质量特征，一共 4898 条数据，本文的因变量是白葡萄酒的质量状态，一共有两种可能的状态：优质和非优质。优质用“0”表示，非优质用“1”表示。数据变量说明见表 1。

**Table 1.** Data variable description table
**表 1.** 数据变量说明表

| 变量类型 | 变量名 | 详细说明 | 取值范围 |
|---|---|---|---|
| 因变量 | 是否优质 | 葡萄酒质量评分分数大于 7 的为优质红酒，低于 7 的为非优质红酒 | 0, 1 |
| 解释变量 | 非挥发性酸 Fixed.acidity | 大多数与葡萄酒有关的酸或固定或非挥发性(不易蒸发) | 3.8~14.2 |
| | 挥发性酸 Volatile.acidity | 葡萄酒中醋酸的含量过高会导致令人不愉快的醋味 | 0.08~1.1 |
| | 柠檬酸 Citric.acid | 柠檬酸可以增加葡萄酒的“新鲜度”和风味 | 0~1.66 |
| | 甜度 Residual.sugar | 发酵停止后剩余的糖量 | 0.6~65.8 |
| | 游离二氧化硫 Free.sulfur.dioxide | 游离形式的二氧化硫在分子 $SO_2$ (作为溶解气体)和亚硫酸氢根离子之间存在平衡 | 2~289 |
| | 氯化物 chlorides | 葡萄酒中盐的含量 | 0.009~0.346 |
| | 硫酸盐 sulphates | 一种葡萄酒添加剂，可以促进二氧化硫气体($SO_2$)水平，作为抗菌剂和抗氧化剂 | 0.22~1.08 |
| | 酒精 alcohol | 葡萄酒的酒精含量百分比 | 8~14.2 |
| | 密度 density | 酒精的密度接近水的密度，密度取决于酒精的百分比和糖含量 | 0.98711~1.03898 |
| | 氢离子浓度 pH | 描述葡萄酒的酸度或碱度大多数葡萄酒在 pH 值为 2.5~4.5 | 2.72~3.82 |
| | 总二氧化硫 Total.sulfur.dioxide | 游离和结合形式的 $SO_2$ 的量：在游离 $SO_2$ 浓度超过 50 ppm 时，$SO_2$ 在酒的鼻子和味道中变得明显 | 9~440 |

## 4. 模型简介

### 4.1. Logistic 回归模型

Logistic 回归又称定性定量回归，是一种广义的线性回归分析模型，其目标变量使用分类型字段而不是数值型，而自变量 $x_1,x_2,\cdots,x_n$ 可以是分类变量、连续变量或者两者的混合类型。Logistic 回归建立一组方程，把输出值域与输出字段每一类的概率联系起来。一旦生成模型，便可用于估计新的数据的概率。对于每一个记录，计算其从属于每种可能输出类的概率，概率最大的目标类被指定为该记录的预测输出值，类似于朴素贝叶斯分类方法。设目标变量 $y$ 是 0~1 型随机变量， $x_1,x_2,\cdots,x_k$ 是任意 $k$ 个变量。

$p=p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)$，那么变量 $y$ 关于变量 $x_1,x_2,\cdots,x_k$ 的 Logistic 回归模型是：

$$p=p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)=\frac{e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}} \tag{1}$$

或者：

$$\log\left(\frac{p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)}{1-p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)}\right)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k \tag{2}$$

有利于一个事件发生的机会比就是事件将要发生的概率与该事件将不发生的概率比：

$$odds=\frac{p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)}{1-p\left(y=1\middle|x_1,x_2,\cdots,x_k\right)}=e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k} \tag{3}$$

设目标变量 $Y$ 是 0~1 型随机变量， $x_1,x_2,\cdots,x_k$ 是对 $Y$ 的取值有影响的确定性变量。在 $x_{i1},x_{i2},\cdots,x_{ik}\left(i=1,2,\cdots,n\right)$ 处分别对 $y$ 进行了 $n$ 次独立观测，记第 $i$ 次观测值为 $y_i$。显然，$Y_i,i=1,2,\cdots,n$ 是相互独立的伯努利随机变量，其概率分布为：

$$P\left(Y_i=y_i\right)=p_i^{y_i}\left(1-p_i\right)^{1-y_i},y_i=0\text{ or }1 \tag{4}$$

于是 $y_1,y_2,\cdots,y_k$ 的似然函数为：

$$\ln L\left(Y,p\right)=\sum_{i=1}^{n}\left[y_i\ln p_i+\left(1-y_i\right)\ln\left(1-p_i\right)\right]=\sum_{i=1}^{n}\left[y_i\ln\frac{p_i}{1-p_i}+\ln\left(1-p_i\right)\right] \tag{5}$$

根据 Logistic 模型描述 $p_i$ 与 $x_{i1},x_{i2},\cdots,x_{ik}\left(i=1,2,\cdots,n\right)$ 之间的关系如下：

$$p_i=\frac{1}{1+e^{\beta_0+\beta_1x_1+\cdots+\beta_kx_k}} \tag{6}$$

其中 $\beta_0,\beta_1,\cdots,\beta_k$ 是待估参数，则：

$$\ln L\left(Y,p\right)=\sum_{i=1}^{n}\left[y_i\left(\beta_0+\sum_{j=1}^{k}\beta_{ij}\right)-\ln\left(1+\exp\left(\beta_0+\sum_{j=1}^{k}\beta_{ij}\right)\right)\right] \tag{7}$$

使得 $\ln L\left(Y,p\right)$ 达到最大值的 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$ 就是 $\beta_0,\beta_1,\cdots,\beta_k$ 的极大似然估计。

### 4.2. 随机森林模型

随机森林模型的基分类器为决策树(CART 算法形成的决策树)，而决策树形成的基本思想为“分而治之”，从上往下进行层层的细分。大致思想可以描述为，首先均匀地选择数据作为模型的训练集，然后对类型制定生成规则，其次根据特征的不同值往下建立分支，有的分支可能到停止，有的分支还可以继

续进行细分，直至最后所有分支不可往下分类，此时决策树构建完成。每一个决策树都有根节点(Root nods)与叶节点(Internal nods)构成，其中叶节点又称作内部节点，而每一个结点都由一个父节点以及 2 个或多个子结点组成。

Breiman (1984)针对分类与回归问题，首先提出了分类树的思想，通过这种算法进行计算时，计算效率大幅度提高。随机森林(Random Forest，简称 RF)是 Bagging 的一个扩展变体。RF 在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了随机属性集合(假定有 $d$ 个属性)中选择一个最优属性；而在 RF 中，对基决策树的每个节点，首先从该结点的属性集合中随机选择一个包含 $k$ 个属性的子集，然后再从这个子集中选择一个最优属性用于划分。这里的参数 $k$ 控制了随机性的引入程度：若令 $k = d$，则基决策树的构建与传统决策树相同；若令 $k = 1$，则是随机选择一个属性用于划分。随机森林是基于 bagging 框架下的决策树模型。

## 5. 描述分析

本案例涉及 11 个解释变量：非挥发性酸、挥发性酸、柠檬酸、甜度、游离二氧化硫、氯化物、硫酸盐、酒精、密度、氢离子浓度。

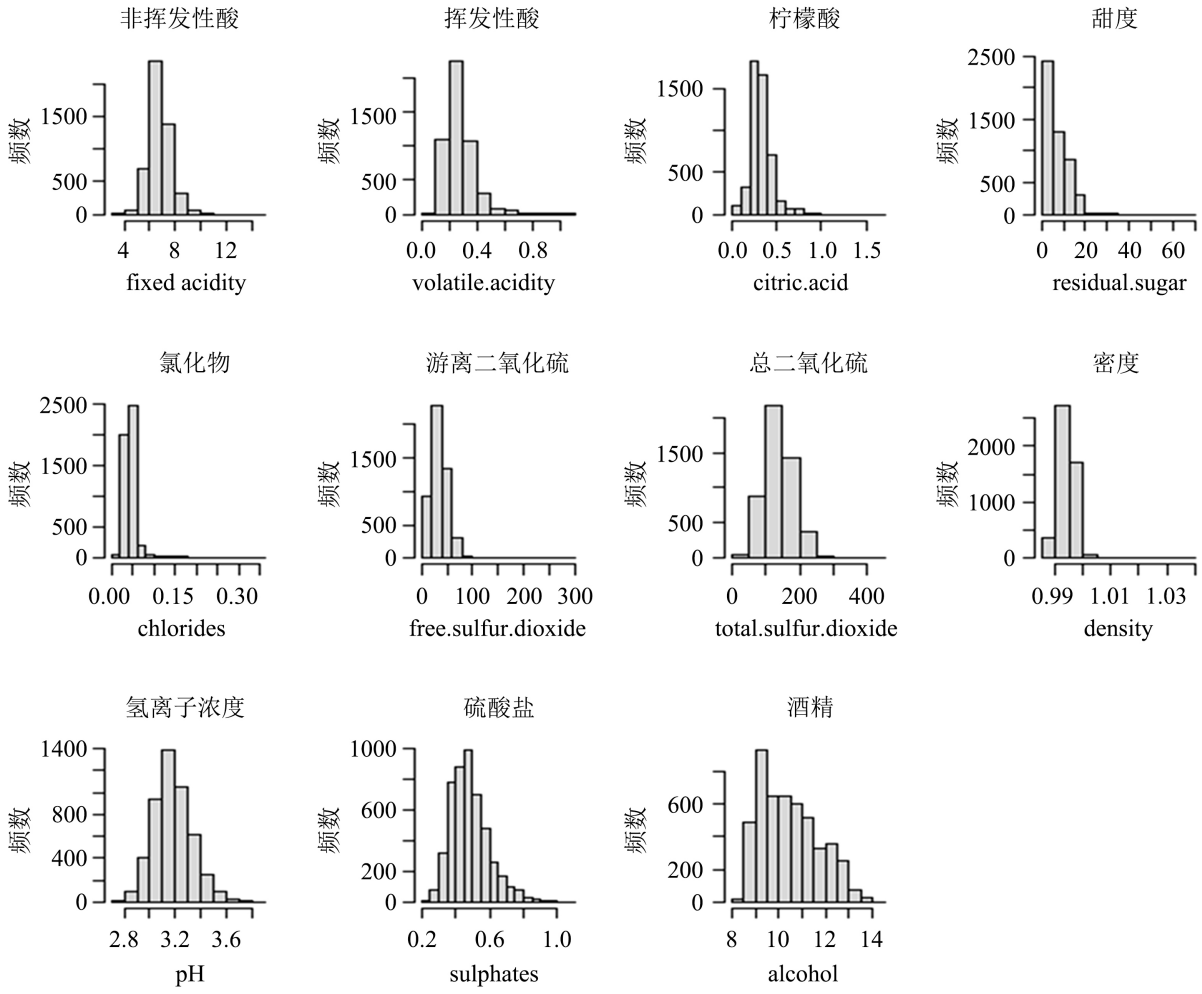


**Figure 1.** Histograms of 11 variables
**图 1.** 11 个变量的直方图

如图 1 所示，大多数白葡萄酒的非挥发性酸度在 6~8 g/dm$^3$，最小值是 3.8，最大值是 14.2。葡萄酒中醋酸的含量过高会导致令人不愉快的醋味。挥发性酸度分布偏向左边，大多数白葡萄酒的挥发性酸度低于 0.4 g/dm$^3$，我们可以猜测大于 0.4 g/dm$^3$ 含量的葡萄酒质量会比较差。柠檬酸可以增加葡萄酒的“新鲜度”和风味，柠檬酸分布偏向左边，大多数白葡萄酒的柠檬酸低于 0.5 g/dm$^3$。甜度低于 10 g/L 的白葡萄酒占了数据的绝大部分，甜度是发酵停止后剩余的糖量，很少能找到甜度含量低于 1 g/L 和超过 45 g/L 的葡萄酒。氯化物分布偏左，大多数白葡萄酒的氯化物含量小于 0.1 g/dm$^3$。游离二氧化硫可以防止微生物的生长和葡萄酒的氧化，可以保证葡萄酒不变质。但含量过高有可能会影响酒的口感。游离二氧化硫分布偏左，大多数白葡萄酒的游离二氧化硫含量小于 100 mg/dm$^3$。总二氧化硫呈正态分布，大多数总二氧化硫含量分布在 100~200 mg/dm$^3$。总二氧化硫是游离和结合形式的 $SO_2$ 的量，在低浓度下，$SO_2$ 在葡萄酒中几乎检测不到，但在游离 $SO_2$ 浓度超过 50 ppm 时，$SO_2$ 在酒中的味道变得明显。$SO_2$ 浓度过高时会影响葡萄酒的气味。密度分布偏左，大多数白葡萄酒的密度小于 1 g/cm$^3$。pH 呈正态分布，大多数 pH 分布在 3~3.3。硫酸盐呈正态分布，大多数硫酸盐含量分布在 0.4~0.6 g/dm$^3$。酒精分布偏右，大多数白葡萄酒的酒精含量小于 13%，白葡萄酒的 11 个物理和化学属性所有观测值近乎都可以呈正态分布，可能和白葡萄酒的质量存在很强的相关关系。

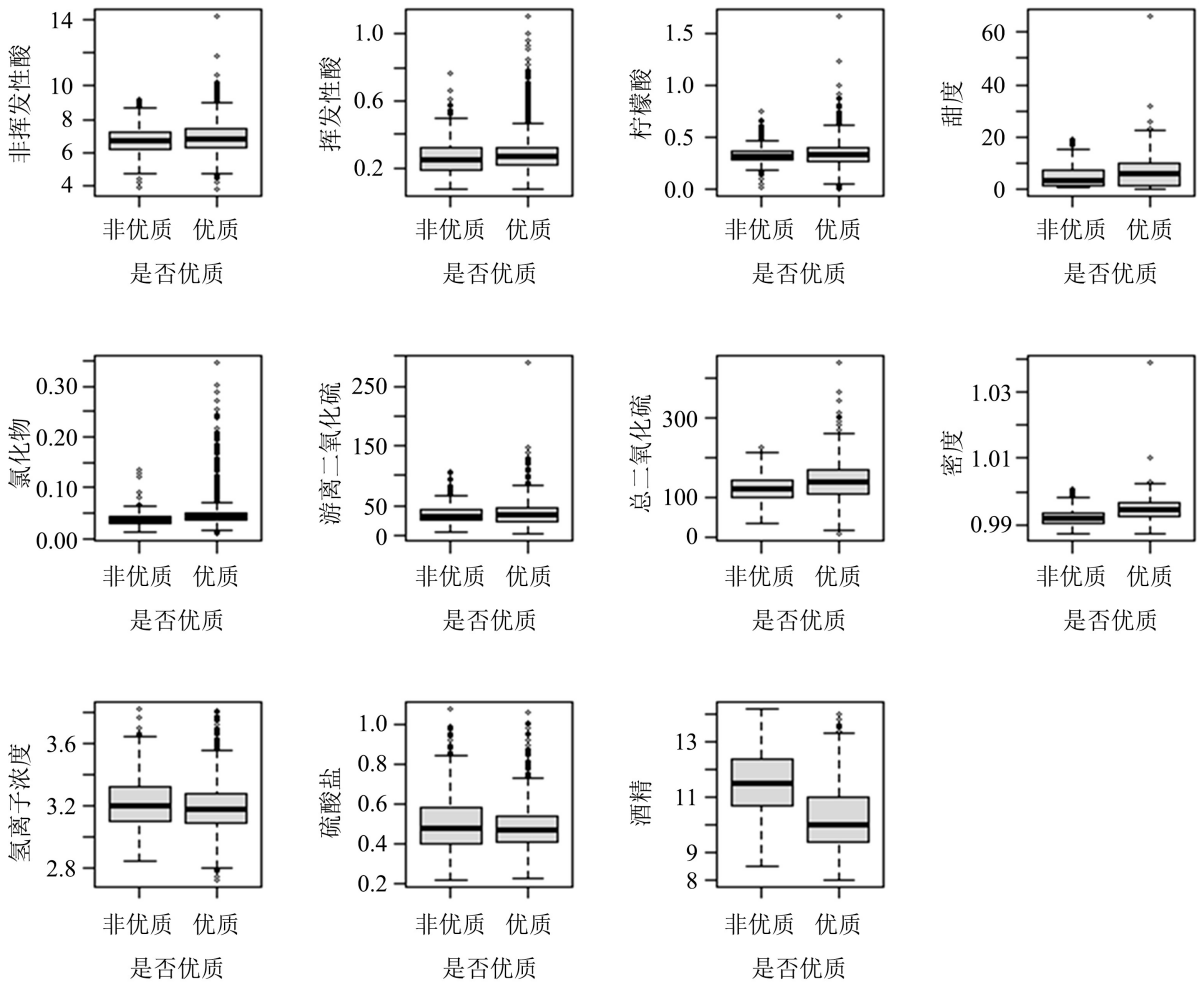


**Figure 2.** Box plot of 11 variables and white wine quality
**图 2.** 11 个变量与白葡萄酒质量的箱线图

接下来，对葡萄酒是否优质和各个解释变量的相关关系做简单描述。由于是否优质是一个 0~1 变量，它将整个数据天然地分为两类(优质 = 0 一类，非优质 = 1 一类)，因此可以对每一个解释变量用箱线图做对比分析。如图 2 所示，我们以酒精对白葡萄酒质量的影响情况来分析其箱线图，具体而言，对优质组(优质 = 0)，它的中位数要明显低于非优质组(非优质 = 1)。这说明酒精浓度过高的更容易被评为非优质白葡萄酒，同理对其他的变量也可以进行分析。这种分析结果是否具有统计学上的显著意义，需要后面的正式模型分析验证。

## 6. 模型建立

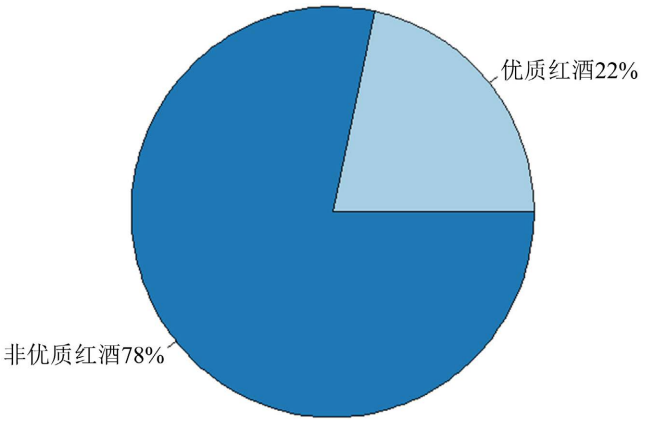


**Figure 3.** White wine quality pie chart
**图 3.** 白葡萄酒质量饼图

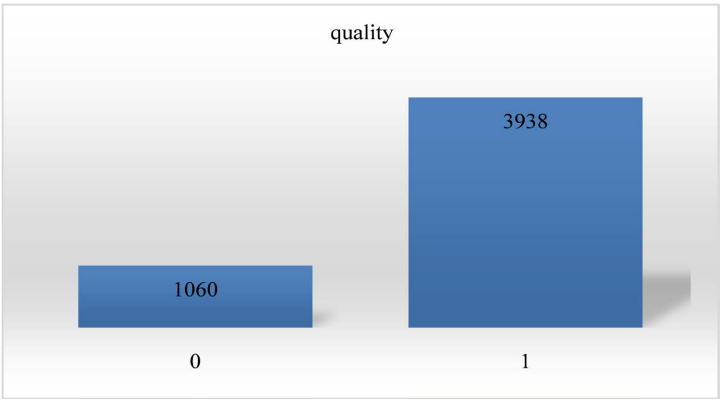


**Figure 4.** Histogram of frequency distribution of high-quality and non high-quality wines
**图 4.** 优质葡萄酒和非优质葡萄酒频数分布直方图

### 6.1. Logistic 回归模型

为了深入挖掘影响白葡萄酒质量是否优质的显著因素，本文将建立 0~1 回归模型(logistic 模型)。针对 4898 个白葡萄酒数据中的质量进行分类汇总，见图 3、图 4。从图 3 和图 4 可以看出，优质白葡萄酒占了总数的 22%，一共 1060 份，而非优质白葡萄酒占了 78%，一共 3938 份。用 70%的数据作为训练集；30%的数据作为测试集。使用前 11 个变量作为自变量，quality 作为因变量建立 logistic 回归模型，得到结果如表 2 所示：

**Table 2.** Logistic regression results
**表 2.** 逻辑回归结果

| 变量名称 | 回归系数 | 标准误 | P-值 |
|---|---|---|---|
| 截距项 | 1.67835 | 0.05718 | <2e−16 |
| Fixed.acidity | −0.43668 | 0.9006 | 1.24e−6 |
| Volatile.acidity | 0.38304 | 0.05939 | 1.12e−10 |
| Citric.acid | 0.03274 | 0.05613 | 0.55968 |
| Residual.sugar | −1.44753 | 0.21345 | 1.19e−11 |
| chlorides | 0.25053 | 0.09406 | 0.00773 |
| Free.sulfur.dioxide | −0.19164 | 0.06498 | 0.00319 |
| Total.sulfur.dioxide | 0.08785 | 0.07614 | 0.24863 |
| density | 1.89881 | 0.33690 | 1.74e−08 |
| pH | −0.45579 | 0.07584 | 1.85e−09 |
| sulphates | −0.24424 | 0.04747 | 2.67e−07 |
| alcohol | −0.20640 | 0.16607 | 0.21392 |
| 模型全局检验 | | P 值 < 0.001 | |

从结果可以看出，AIC 值为 2927.4，11 个变量中有柠檬酸、总二氧化硫、酒精的系数是不显著的，所以我们考虑利用逐步回归来删除数据中显著性表现不理想的变量。通过对训练集进行逐步回归，得出以下结果：

**Table 3.** Stepwise regression results
**表 3.** 逐步回归结果

| 变量名称 | 回归系数 | 标准误 | P-值 |
|---|---|---|---|
| 截距项 | 1.6780 | 0.05704 | <2e−16 |
| Fixed.acidity | −0.50777 | 0.06376 | 1.67e−15 |
| Volatile.acidity | 0.38015 | 0.05616 | 1.30e−11 |
| Residual.sugar | −1.68291 | 0.11565 | 2e−16 |
| chlorides | 0.24995 | 0.09377 | 0.00769 |
| Total.sulfur.dioxide | −0.14414 | 0.05239 | 0.00594 |
| density | 2.32106 | 0.12610 | <2e−16 |
| pH | −0.51304 | 0.05892 | <2e−16 |
| sulphates | −0.26006 | 0.04510 | 8.12e−09 |
| 模型全局检验 | | P 值 < 0.001 | |

我们由表 3 可知，逐步回归后，将模型中系数不显著的变量剔除，这时模型所有变量系数的显著性水平明显提升，且 AIC 值降到了 2924.1。用 $x_1, x_2, \cdots, x_8$ 分别表示非挥发性酸、挥发性酸、甜度、氯化物、游离二氧化硫、密度、氢离子浓度、硫酸盐这 8 个变量，则 Logistic 回归模型可以表示为：

$$p = p\left(y = 1 \middle| x_1, x_2, \cdots, x_8\right) = \frac{e^{1.678-0.50777x_1+0.38015x_2-1.68291x_3+\ldots-0.51304x_7-0.26006x_8}}{1+e^{1.678-0.50777x_1+0.38015x_2-1.68291x_3+\ldots-0.51304x_7-0.26006x_8}} \tag{8}$$

在其他解释变量保持不变的情况下，硫酸盐每增加一个单位，将会使优质葡萄酒发生的机会比减小原来的 $e^{-0.26006}$ 倍。

Logistic 回归参数的显著性检验的目的是逐个检验模型中的各个解释变量是否与 $\log\left(\frac{p}{1-p}\right)$ 有显著性关系，用卡方检验来检验回归系数是否显著。

**Table 4.** Chi square test results
**表 4.** 卡方检验结果

| 变量名称 | 偏差 | 残差 | P-值 |
|---|---|---|---|
| NULL | | 3592.2 | |
| Fixed.acidity | 19.026 | 3573.2 | 1.290e−05 |
| Volatile.acidity | 18.923 | 3544.3 | 1.361e−05 |
| Residual.sugar | 40.834 | 3513.4 | 1.657e−10 |
| chlorides | 199.184 | 3314.2 | <2.2e−16 |
| Free.sulfur.dioxide | 1.475 | 3312.8 | 0.2246 |
| density | 290.513 | 3022.2 | <2.2e−16 |
| pH | 82.883 | 2939.4 | <2.2e−16 |
| sulphates | 33.300 | 2906.1 | 7.898e−09 |
| 模型全局检验 | | P 值 < 0.001 | |

由表 4 可以看出，经过逻辑回归后出现了 free.sulfur.dioxide 这个不显著的变量，为了后续研究，我们剔除这一变量。继续进行模型检验，得到如下结果如表 5 所示：

**Table 5.** Chi square test results
**表 5.** 卡方检验结果

| 变量名称 | 偏差 | 残差 | P-值 |
|---|---|---|---|
| NULL | | 3592.2 | |
| Fixed.acidity | 19.026 | 3573.2 | 1.290e−05 |
| Volatile.acidity | 18.923 | 3544.3 | 1.361e−05 |
| Residual.sugar | 40.834 | 3513.4 | 1.657e−10 |
| chlorides | 199.184 | 3314.2 | <2.2e−16 |

**Continued**

| | | | |
|---|---|---|---|
| density | 290.513 | 3022.2 | <2.2e−16 |
| pH | 82.883 | 2939.4 | <2.2e−16 |
| sulphates | 33.300 | 2906.1 | 7.898e−09 |
| 模型全局检验 | | P 值 < 0.001 | |

如结果所示，NULL 表示零模型，剩余偏差为：3592.2，接下来非挥发性酸进入模型，产生了 19.026 的模型偏差，p-value 非常小，显著性水平为 0.001 的情况下，所有系数的 P 值都小于 0.001，拒绝原假设，认为回归系数是显著的。

**Table 6.** H-L inspection results
**表 6.** H-L 检验结果

| df | P-value |
|---|---|
| 8 | 0.1269 |

Hosmer-Lemeshow 检验(HL 检验)为模型拟合指标，其原理在于判断预测值与真实值之间的 gap 情况，如果 p 值大于 0.05，则说明通过 HL 检验，即说明预测值与真实值之间并无非常明显的差异。反之如果 p 值小于 0.05，则说明没有通过 HL 检验，预测值与真实值之间有着明显的差异，即说明模型拟合度较差。由上述生成结果表 6 可知，p 值大于 0.05，说明通过了 HL 检验。

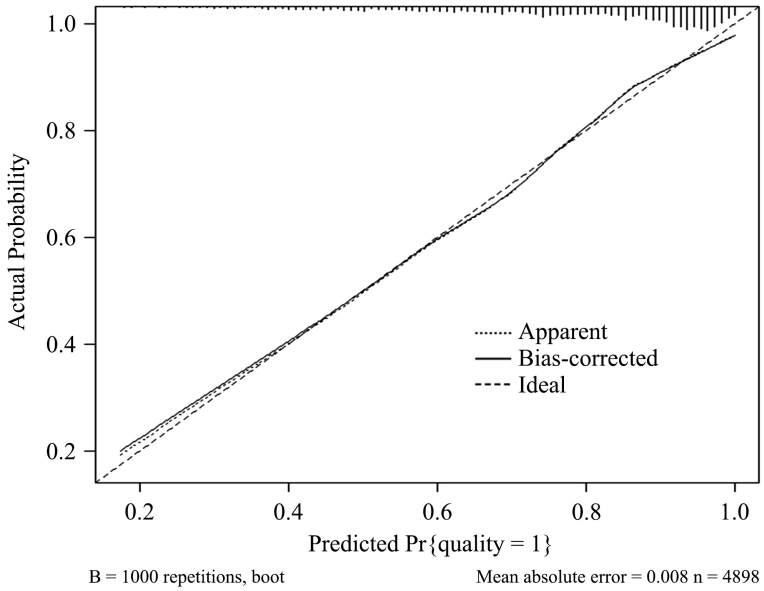


**Figure 5.** Calibration curve
**图 5.** 校正曲线

用 bootstrap 抽样方法重复 1000 次抽样对模型做偏引校正，以 P 值作为判断准则，对全模型的变量进行选择，结果与 AIC 的选择结果一致。得到如图 5 所示校正曲线，Apparent 是当前的模型，

Bias-corrected 是通过 bootstrap 检验偏引得到的模型，Ideal 是 45˚理想参考线，模型线都很接近参考线，所以模型拟合效果较好。从表 7 所示混淆矩阵来看，模型预测的准确率达到了 79.3%，优质白葡萄酒预测正确 96 例，非优质白葡萄酒预测正确 1070 例。灵敏度为 92.6%，F1-Score = 0.8759，说明该模型预测效果较好。

**Table 7.** Confusion matrix
**表 7.** 混淆矩阵

| true_value \ predict_value | 0 | 1 |
|---|---|---|
| 0 | 96 | 218 |
| 1 | 85 | 1070 |

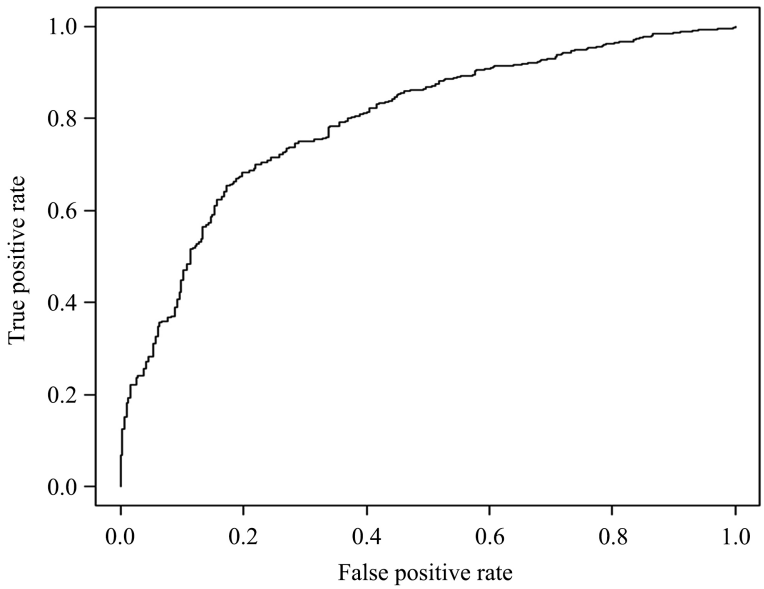


**Figure 6.** ROC curve
**图 6.** ROC 曲线图

在图 6 所示 ROC 曲线中，横轴为 FPR，纵轴为 TPR，理想目标是：TPR = 1，FPR = 0。该图的 AUC 值为 0.793，模型预测效果较好。

**Table 8.** Ten fold cross validation results
**表 8.** 十折交叉验证结果

| 组号 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 测试集精确度 | 0.793456 | 0.804816 | 0.7979592 | 0.8204082 | 0.8081633 |
| 训练集精确度 | 0.80313 | 0.8010436 | 0.8030853 | 0.798775 | 0.7996824 |
| 组号 | 6 | 7 | 8 | 9 | 10 |
| 测试集精确度 | 0.8020408 | 0.7959184 | 0.809816 | 0.7959184 | 0.7836735 |
| 训练集精确度 | 0.800363 | 0.8030853 | 0.8006351 | 0.8021779 | 0.8033122 |

从表 8 结果可以看出，第四组的测试集精确度最大，到达了 82.04%，训练集精确度为 79.8%。

## 6.2. 随机森林模型

在训练随机森林模型前，需要寻找最优参数 mtry，即指定节点中用于二叉树的最佳变量个数。以及最佳参数 ntree，即指定随机森林所包含的最佳决策树数目。从图 7 中可以看出，当参数 mtry 为 8 时，模型误判率最低。从图 8 中可以看出，在树的数量大于 1200 的时候就基本稳定了，在保证效能的情况下减少决策树的数量，减少运行时间，因此参数 ntree 取 1200。

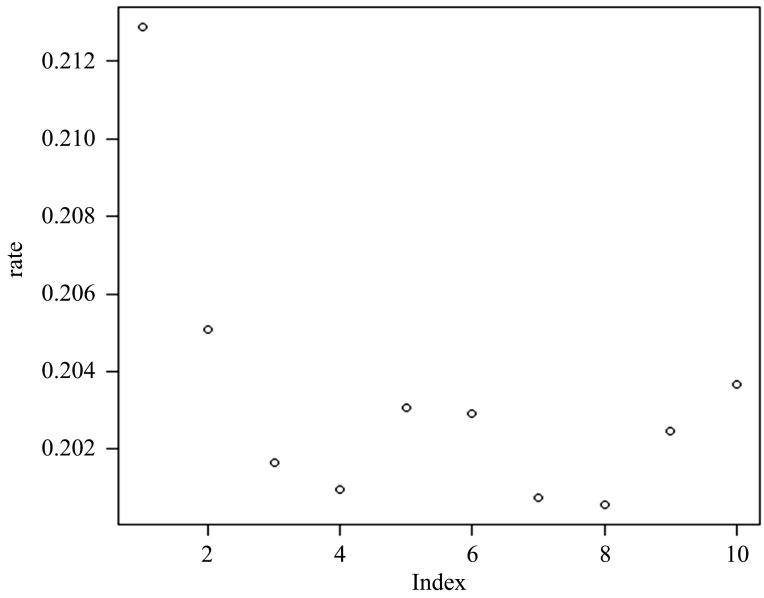


**Figure 7.** Misjudgment rate of different parameter models
**图 7.** 不同参数模型误判率

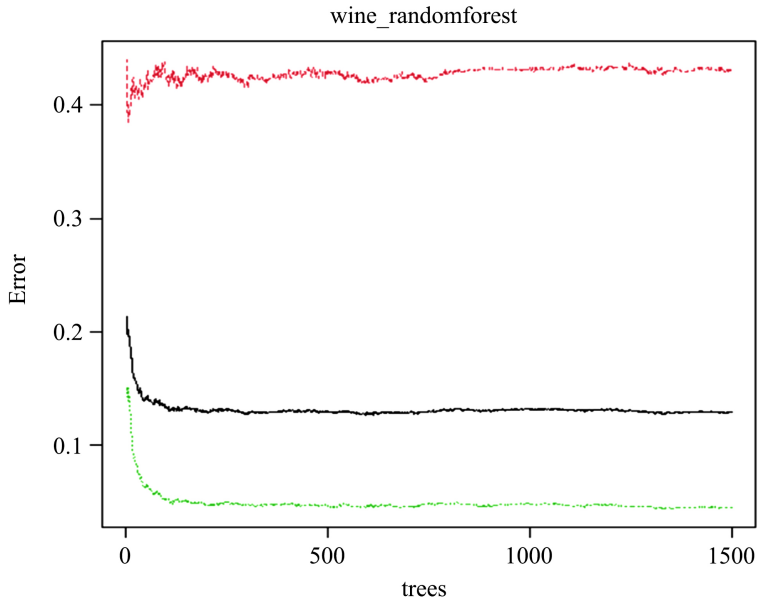


**Figure 8.** Misjudgment rate of different parameter models
**图 8.** 不同参数模型误判率

用 mtry = 8 和 ntree = 1200 来训练随机森林模型，可以得到如下图 9 所示结果：

```
Call:
 randomForest(formula = quality ~ ., data = data_train, mtry = i,      ntree = 1200)
               Type of random forest: classification
                     Number of trees: 1200
No. of variables tried at each split: 8

        OOB estimate of  error rate: 12.72%
Confusion matrix:
    0    1 class.error
0 431  315  0.42225201
1 121 2562  0.04509877
```

**Figure 9.** Random forest
**图 9.** 随机森林模型

用 mtry = 8 和 ntree = 1200 来训练随机森林模型，用测试集进行预测后，得到表 9 的混淆矩阵，可以看出测试集预测正确率达到 82.59075%，相比于 Logistic 回归模型的预测正确率更高。

**Table 9.** Confusion matrix
**表 9.** 混淆矩阵

| true_value \ predict_value | 0 | 1 |
|---|---|---|
| 0 | 69 | 34 |
| 1 | 7 | 380 |

为了探究随机森林模型预测准确率，利用十折交叉验证法进行训练，得到如表 10 所示的结果。从表中可以看出，第六组数据的效果最好，测试集预测精确度达到了 90.81633%，训练集预测精确度达到 99.93194%。

**Table 10.** Table of ten fold cross prediction results
**表 10.** 十折交叉预测结果表

| 组号 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 测试集精确度 | 0.8691207 | 0.8693878 | 0.8591837 | 0.9 | 0.8979592 |
| 训练集精确度 | 0.9997732 | 0.9995463 | 0.9997731 | 0.9993194 | 0.9990926 |
| 组号 | 6 | 7 | 8 | 9 | 10 |
| 测试集精确度 | 0.9081633 | 0.8836735 | 0.8732106 | 0.8938776 | 0.8938776 |
| 训练集精确度 | 0.9993194 | 0.9993194 | 0.9995464 | 0.9995463 | 0.9990926 |

从随机森林模型十折交叉验证的结果中，对预测效果最好的第六组的变量重要性进行分析，见图 10，可以发现重要性排序从大到小前七个指标分别是酒精、挥发性酸、游离二氧化硫、氢离子浓度、甜度、氯化物、硫酸盐，可以说明这七个指标对白葡萄酒品质预测的影响较大。

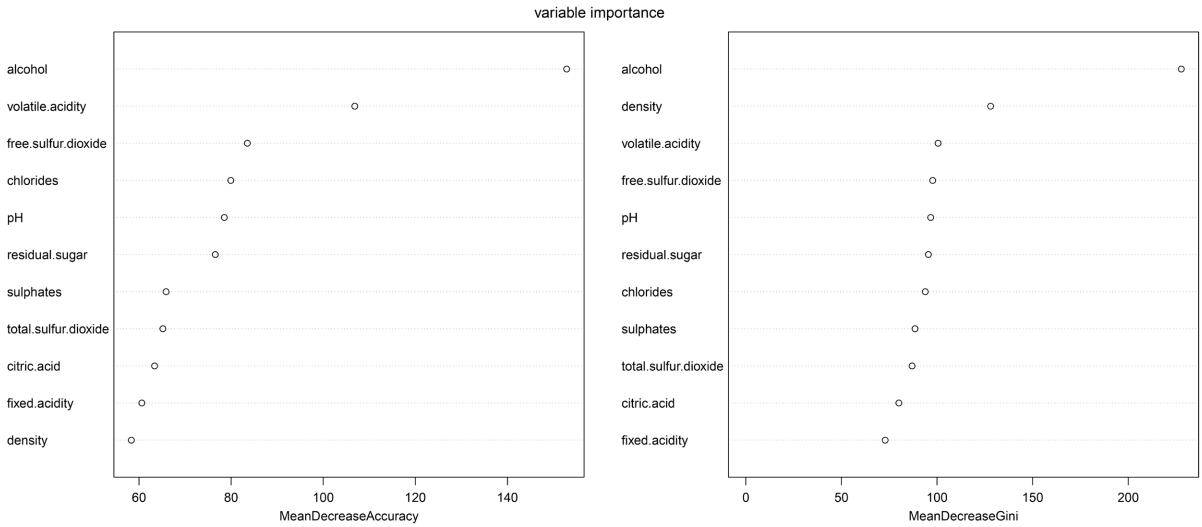


**Figure 10.** Sorting chart of important variables in the sixth group of data
**图 10.** 第六组数据重要变量排序图

## 7. 结论与展望

本文基于 UCL 机器学习数据库——白葡萄酒数据，以是否为优质白葡萄酒为因变量，以 11 个评价白葡萄酒质量的指标作为自变量，包括非挥发性酸、挥发性酸、甜度、氯化物、密度、氢离子浓度、硫酸盐，建立对白葡萄酒是否为优质具有一定预测能力的逻辑回归模型与随机森林模型。Logistic 回归模型选出 7 个显著变量非挥发性酸、挥发性酸、甜度、氯化物、密度、氢离子浓度、硫酸盐，也说明这七个变量对白葡萄酒品质有显著影响，ROC 曲线表明了该模型预测效果较好。而随机森林模型中变量重要性排序前七的指标分别是酒精、挥发性酸、游离二氧化硫、氢离子浓度、甜度、氯化物、硫酸盐，同时结合 Logistic 回归模型的结果可以发现，挥发性酸、氢离子浓度、甜度、氯化物、硫酸盐这五个指标对白葡萄酒品质预测的影响是最显著的，因此商家在生产葡萄酒时，可以在这五个指标上重点把控，生产出的优质葡萄酒的可能性更大。

Logistic 回归模型的测试集精确度均值为 80.12%，训练集精确度均值为 80.15%，再对比随机森林模型，通过十折交叉验证后，测试集精确度均值为 88.48%，训练集精确度均值为 99.94%，随机森林模型不论是测试集还是训练集都明显优于 Logistic 回归模型，尤其是测试集预测准确率相比于 Logistic 回归模型提高了 8.36 个百分点，明显提升了对葡萄酒特征的预测效果。随机森林模型是一种高准确度的分类器，具有快速简便的优点，可以在决定类别时，评估变量的重要性，对于不平衡的分类资料集来说，可以平衡误差。随机森林模型在酒类行业中，为葡萄酒的评价体系提供了一种快捷、准确且科学合理的方法。

## 参考文献

[1] 魏舜洋, 石国良. 基于模糊聚类模型的葡萄酒分类[J]. 中国传媒大学学报(自然科学版), 2015, 22(4): 49-53.

[2] 李璐, 李京鸿, 高影, 等. 葡萄酒对中枢神经系统损伤的保护作用研究进展[J]. 沈阳药科大学学报, 2021, 38(3): 314-320+327.

[3] 苏小菱, 洪昀. 基于层次分析评价模型的课程思政有效性评价探索[J]. 教育教学论坛, 2020(22): 150-152.

[4] 杨佳佳, 张铃丽, 路凯. 基于层次分析法的网络信息资源评价研究[J]. 许昌学院学报, 2020, 39(2): 129-133.

[5] 周灿, 廖振良, 孔令婷, 钱真. 基于熵权的模糊层次评价法在滴水湖水质评价中的应用[J]. 能源环境保护, 2020, 34(1): 82-87.

[6] 李伟康. 基于 GA-BP 神经网络对葡萄酒质量评估的研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2018.

[7] 李志华. 基于多元回归模型的葡萄酒“智”量分析[J]. 中国食品工业, 2023(3): 85-87+91.
[8] 张志然, 王恩辉, 李兴元, 等. 葡萄酒感官质量评价及其理化因子分析[J]. 现代食品, 2022, 28(7): 202-206.
[9] 唐文龙, 火兴三, 阮仕立, 等. 葡萄酒产品的质量维度与市场表达体系[J]. 中外葡萄与葡萄酒, 2022(6): 19-23.
[10] 李记明, 姜文广. 优质干红葡萄酒中主要质量指标的研究[J]. 中外葡萄与葡萄酒, 2018(6): 18-24.