# 基于 RFECV-RF-Boosting 的烟叶感官质量预测研究

王龙鑫[1],冯文宁[2],崔扶芸[2],刘波[2],赵晖[2],申玉军[3],张渤海[1],来苗[1]

1. 河南农业大学 烟草学院,河南 郑州 450002;
2. 河北中烟工业有限责任公司 技术中心,河北 石家庄 050051;
3. 中国烟草总公司郑州烟草研究院,河南 郑州 450001

**摘要**:【目的】解决烟叶感官质量评价中存在的主观性强、数据获取困难等问题,实现基于数字化分析对烟叶感官质量的精准定量预测。【方法】以河南、湖南、云南和贵州 4 个典型风格产区的 264 份烟叶为研究对象,开展化学成分检测与感官质量评价,经化学指标间相关性分析剔除冗余后,采用基于随机森林的交叉验证递归特征消除方法(RFECV-RF)对各感官指标筛选最优特征子集,再分别采用极端梯度提升(XGBoost)、分类梯度提升(CatBoost)和轻量级梯度提升机(LightGBM)3 种经典梯度提升(Boosting)算法,经五折交叉验证优化超参数后建立 9 项感官指标预测模型。【结果】1)化学指标相关性分析剔除总糖、糖碱比、钾氯比和棕榈酸 4 项化学成分指标,保留总糖、还原糖、烟碱等 25 项化学成分指标用于后续建模。2)RFECV-RF 特征筛选选取各感官指标对应的最佳特征组,明确总氮、还原糖、钾和烟碱是影响烟叶感官质量的关键化学成分,其交叉验证的均方根误差(*RMSE*)均低于全特征集模型,有效降低模型复杂度,显著提升预测精度。3)最优算法下各感官指标决定系数($R^2$)为 0.711 3~0.894 0,*RMSE* 为 0.084 5~0.140 4,平均绝对百分比误差(*MAPE*)为 1.06%~1.70%,均取得了良好且稳定的预测效果。【结论】本文预测模型框架可实现烟叶感官质量高精度的量化预测,为卷烟产品数字化配方设计与品质控制提供了参考。

**关键词**:烟叶化学成分;感官质量;Boosting 算法;机器学习;特征选择

**中图分类号**:TS411 **文献标识码**:A

## 0 引言

烟叶感官质量是评价其品质及工业应用价值的关键指标。传统评价方法主要依靠经验丰富的评吸专家抽吸完成,高度依赖专家主观感受,易受其经验水平、生理及心理状态影响;当样品量过大时,人工感官评价不仅耗时费力,结果也难以保持一致[1]。烟叶的化学成分直接决定烟气组成,是影响烟叶感官质量的关键因素。因此,研究者长期致力于构建基于烟叶理化指标的质量评价方法,以量

**收稿日期**:2025-07-09;**修回日期**:2025-09-30

**基金项目**:河南省自然科学基金项目(232300421257);河北中烟工业有限责任公司重点科技项目(HBZY2024A047)

**作者简介**:王龙鑫(2000—),男,河南省郑州市人,河南农业大学硕士研究生,主要研究方向为烟叶质量评价。E-mail:wlx17838722310@163.com

**通信作者**:冯文宁(1978—),男,辽宁省锦州市人,河北中烟工业有限责任公司高级工程师,主要研究方向为卷烟产品开发及配方维护。E-mail:fengwn@126.com

化烟叶感官评分与客观理化指标的关联;传统数理统计方法包括相关分析、逐步回归、通径分析等,用以识别影响感官质量的主导因子[2-3]。此外,主成分分析、典型相关分析及灰色关联分析也被广泛应用,以揭示多指标之间的内在结构关联,进而辅助烟叶感官质量的客观评价[4-5]。然而,上述传统方法多侧重于线性关系的定性建模,难以捕捉化学成分与感官评分之间潜在的复杂非线性关联,预测能力有限。

随着化学数据挖掘与人工智能技术的快速发展,基于大数据驱动的机器学习方法已逐步被引入烟草品质分析领域。C. L. He 等[6]构建了基于反向传播神经网络(BPNN)的预测模型,将烟叶化学成分映射至香气质、香气量、透发性、杂气等关键感官特征,模型决定系数($R^2$)均高于 0.70,显示出较好的预测性能。黄建等[7]分别采用支持向量机(SVM)与随机森林(RF)构建感官质量档次预测模型探究不同档次烟叶化学成分差异和分布发现这 2 种模型的精确率、召回率及 F1 分数加权平均值均超过 84%,预测性能良好。侯冰清等[8]基于 BPNN 模型对雪茄原料常规化学成分与感官质量之间的复杂关联开展深入分析,发现该模型可实现感官质量的精准预测。张云伟等[9]提出的融合近红外光谱(NIR)与 Transformer 架构的烟叶感官质量预测方法,可实现烟叶风格特征、烟气特征与质量特征的精准预测,各感官指标的预测误差均不超过 0.56。别瑞等[10]将极端梯度提升(XGBoost)与遗传算法(GA)相结合,构建了基于化学成分的烟叶等级判别模型,并采用沙普利加和解释法(SHAP)分析各特征的重要贡献,发现该模型实现了对烟叶等级的高准确率判别,展现出良好的解释性。综上可知,机器学习技术虽在烟叶感官质量评价研究中取得显著进展,在提高评价效率与预测精度方面展现出巨大潜力。但现有研究在烟叶感官质量预测中普遍存在特征冗余、单一模型难以兼顾多维感官指标预测需求、关键化学成分作用规律挖掘不足等问题。

基于此,本研究构建了融合递归特征消除结合交叉验证(RFECV)的 RF 特征筛选方法与极端梯度提升(XGBoost)、分类梯度提升(CatBoost)和轻量级梯度提升机(LightGBM)这 3 种典型的梯度提升算法(Boosting)的建模框架,并探究还原糖、苹果酸、烟碱等化学成分对各感官指标的影响,以期为烟叶感官质量的客观评价提供高效可行的方法。

# 1 材料与方法

## 1.1 材料、试剂和仪器

主要材料:本研究所用实验材料选自 2022—2024 年典型浓香型、中间香型、清香型产区的代表性烟叶样品。涵盖河南、湖南、贵州和云南 4 个产地,均为当地主栽品种,包含上、中、下部位。具体样品信息如表 1 所示,样品总数为 264。每份样品切丝混匀后随机分为两组,一组在 40 ℃恒温条件下烘干、粉碎,过 80 目筛均质处理,用于后续化学成分测定;另一组卷制为标准烟支,平衡水分后用于感官质量评价。

主要试剂:无水 $CuSO_4$、无水 $Na_2SO_4$、NaOH、浓 $H_2SO_4$、二氯甲烷、乙酸、己二酸,均为分析纯,国药集团化学试剂有限公司;甲醇、$C_7$—$C_{40}$ 正构烷烃,均为色谱纯,德国默克有限公司;2-甲基-3-庚酮(色谱纯),上海麦克林生化科技有限公司;去离子水,由河南农业大学实验室自制。

主要仪器:AA3 型连续流动分析仪,德国水尔公司;LC-20AR 型高效液相色谱仪,日本岛津公司;TGL-16M 型台式高速冷冻离心机,上海卢湘仪离心机仪器有限公司;7890A 型气相色谱(配氢火焰离子检测器 FID)、5957C-7890A 型气相色谱-质谱联用(GC-MS)仪,美国安捷伦公司;FAS-3823 型多功能进样器,郑州安诺科学仪器有限公司。

表 1 样品清单

Table 1 Sample list

| 产区 | 品种 | 等级 | 单等级样本数 | 配方模块样本数 |
|---|---|---|---|---|
| 河南 | 中烟 100 | B2F、B3F、C2F、C3F、C3L、X2F | 35 | 27 |
| 湖南 | 云烟 87 | B2F、B3F、C2F、C3F、X2F | 28 | 24 |
| 贵州 | 云烟 87 | B2F、B3F、C2F、C3F、X2F | 37 | 32 |
| 云南 | 云烟 87、红花大金元 | B2F、B3F、C1F、C2F、C3F、C4F、X2F | 44 | 37 |

### 1.2 实验方法

**1.2.1 常规化学成分检测** 参照文献[11-15]检测样品中总糖、还原糖、总氮、烟碱、氯和钾含量,并计算钾氯比、糖碱比和氮碱比作为衍生指标。

**1.2.2 多酚检测** 参照文献[16],检测样品中新绿原酸、绿原酸、隐绿原酸、芸香苷和莨菪亭含量。

**1.2.3 有机酸检测** 参照文献[17]的方法,采用气相色谱-氢火焰离子化检测器检测样品中草酸、丙二酸、琥珀酸、苹果酸、柠檬酸、棕榈酸、十七酸、油酸、亚油酸、硬脂酸等非挥发性有机酸和高级脂肪酸含量。

**1.2.4 中性香气成分检测** 采用顶空-固相微萃取-气相色谱-质谱联用(HS-SPME-GC-MS)技术对样品中中性挥发性香气成分进行定性定量检测。

前处理方法:称取 0.5 g 充分均质的烟末于 20 mL 顶空瓶中,加入 10 μL 内标溶液(2-甲基-3-庚酮,1 mg/mL),于 70 ℃条件下孵育 10 min 实现组分平衡,在相同温度下萃取 40 min,进入 GC 进样口(250 ℃)解吸 5 min 后进行 GC-MS 检测。

GC 条件:HP-5 MS 石英毛细管柱(30 m×0.25 mm×0.25 μm);载气为高纯 He(1.0 mL/min,恒流);分流进样,分流比 2∶1,分流流量 1.6 mL/min;进样口温度 250 ℃;柱温程序为 50 ℃ 保持 2 min,以 2 ℃/min 的速率升温至 180 ℃并保持 2 min,再以 20 ℃/min 的速率升温至 250 ℃。

MS 条件:接口(传输线)温度 250 ℃,离子源温度 230 ℃,电子电离(EI, 70 eV),扫描范围($m/z$) 33~450 amu。

定性定量分析方法:基于 NIST 2023 质谱数据库(匹配度≥85%)进行化合物鉴定,并辅以保留指数(Retention Index,$RI$)比对,采用内标法定量,计算公式如下:

$$C_x = C_{is} \times \frac{A_x}{A_{is}}$$

式中,$C_x$ 为目标化合物含量,$C_{is}$ 为内标物已知浓度,$A_x$ 为目标化合物峰面积,$A_{is}$ 为内标物峰面积。

根据香味前体来源,将鉴定出的中性香气成分划分为类胡萝卜素降解产物、棕色化反应产物、苯丙氨酸降解产物、西柏烷类降解产物和叶绿素降解产物五大类。

**1.2.5 感官质量评价** 参照文献[18],采用 9 分制对样品感官质量进行评价。评价指标包括香气质、香气量、劲头、浓度、杂气、刺激性、余味、柔和细腻程度、回甜感 9 项,由郑州烟草研究院、河北中烟、湖南中烟、四川中烟等单位感官评价专家组成评价组,每位专家对样品各指标进行评价,具体评分标准如表 2 所示。

### 1.3 数据预处理

对样品化学成分检测与感官质量评价获得的原始数据集进行规范化预处理,保障后续模型训练的稳定性与预测结果的可靠性。

**1.3.1 异常值处理** 本研究采用箱线图(Box Plot)法进行异常值识别,将超出上下界的观测值标记为异常值。针对异常值,采用样本均值替代处理,在避免人为剔除样本基础上完成数据集平滑校正,为后续建模提供更为稳定的数据输入。

**1.3.2 数据标准化** 鉴于烟叶各项化学成分与感官指标在量纲与取值范围上的显著差异,为提升模型训练效率并确保各特征在机器学习过程中的等权参与,本研究对所有特征变量实施标准分数(Z-Score)标准化处理,将其统一映射至均值为 0、标准差为 1 的标准正态分布尺度。其计算公式如下:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

式中,$x_{ij}$ 为样本 $i$ 在指标 $j$ 上的原始测定值,$\mu_j$ 与 $\sigma_j$ 为该指标的均值和标准差,$z_{ij}$ 为经过标准化后的特征值。

表 2 感官质量评价指标及评分标准

Table 2 Sensory quality evaluation indicators and scoring criteria

| 评价指标 | 评分标准 | | |
|---|---|---|---|
| | 6.1~9.0 分 | 3.1~6.0 分 | <3.0 分 |
| 香气质 | 较好、很好 | 略差、稍好 | 差 |
| 香气量 | 较充足、很充足 | 略少、尚充足 | 少 |
| 杂气 | 微弱、无 | 略重、稍有 | 重 |
| 浓度 | 较浓、很浓 | 略淡、稍浓 | 淡 |
| 劲头 | 略小或略大、适中 | 较小或较大、很小 | 很大或很小 |
| 刺激性 | 微弱、无 | 略大、稍有 | 大 |
| 余味 | 较纯净、很纯净 | 略纯净、稍纯净 | 不纯净 |
| 柔和细腻程度 | 较细腻、很细腻 | 略粗糙、稍细腻 | 粗糙 |
| 回甜感 | 较强、很强 | 略弱、稍强 | 弱 |

### 1.4 特征筛选

**1.4.1 基于 Spearman 相关的冗余特征剔除** 本研究采用斯皮尔曼相关系数(Spearman Correlation Coefficient, $\rho$)开展变量间的相关性分析,首先对全部化学指标进行相关性分析,以识别高度相关的冗余变量。对强相关的特征对($|\rho|>0.85$),依据其与目标感官指标的相关强度,保留信息量更高者并剔除冗余变量,获得共线性低的特征用于后续递归特征筛选与模型构建。为确保相关性度量方法的适用性与分析结果的准确性,需先对全部化学成分做夏皮罗-威尔克(Shapiro-Wilk)检验,结果显示,除总糖外,其余大多数变量的 $\rho$ 值均显著小于 0.05,表明其不满足正态分布假设。$\rho$ 计算公式如下:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

式中,$d_i$ 为变量 $x$ 和 $y$ 对应观测值排序后的等级差,$n$ 为样本数量。$|\rho|$ 越接近 1,表明两变量的秩序关系越一致,接近 0 则说明无明显单调关系。

**1.4.2 基于 RFECV-RF 的特征优选** RFECV[19]是一种典型的包裹式(Wrapper)特征选择方法,其依据模型输出的特征重要性评分,逐步剔除权重最低的特征变量。为确定各感官指标最优特征组合,本研究针对香气质、香气量、刺激性等 9 项感官指标,分别构建基于 RFECV-RF 的特征筛选模型,并以五折交叉验证的平均 *RMSE* 作为评估指标,动态评估不同特征子集下的模型表现。具体流程如下:1)使用全部候选特征训练 RF,并通过五折交叉验证计算基准均方根误差(*RMSE*)。2)依据 RF 输出的特征重要度,定位权重最低的特征。3)移除该特征,重新训练模型并更新交叉验证 *RMSE*。4)重复步骤 2)和 3),直至 *RMSE* 不再下降或特征数量达到设定下限。5)记录交叉验证 *RMSE* 最小时对应的特征组合,即为该感官指标的最优子集。

### 1.5 Booting 模型构建

鉴于本研究烟叶化学指标数据维度高、稀疏性强且样本量相对有限,选取 XGBoost、CatBoost 和 LightGBM 作为后续建模算法,全面评估其在烟叶感官质量预测中的适用性。考虑到各感官指标分布的非均质性及模型泛化能力的验证需求,采用分层随机抽样(Stratified Sampling)方法,以各感官指标得分区间作为分层依据,将数据集按 8∶2 的比例划分为训练集和测试集训练集包含 211 个样本,用于模型构建及超参数调优;测试集包含 53 个样本,用于独立评估模型在未见数据上的预测性能。分别对 9 项感官指标构建独立的预测模型,并对各指标均单独实施分层抽样,以确保训练集与测试集在该指标得分分布保持一致并具有代表性。在模型训练阶段通过更改随机数种子进行 5 次独立重复训练,用以评估模型稳定性。

### 1.6 超参数优化

自动化超参数优化(Optuna)框架[20]是一款轻量级开源算法,因其高效性与灵活性在机器学习建模中被广泛应用。本研究的超参数优化采用 Optuna 框架实现,以五折交叉验证 *RMSE* 的均值最小化作为优化目标,优化过程在训练集内部进行,设置最大迭代次数为 200,超参数优化范围如表 3 所示,其余超参数为默认值。

### 1.7 模型评估

本研究选用 *RMSE*、平均绝对百分比误差(*MAPE*)和决定系数($R^2$)3 种常用回归任务评价指标对不同模型进行定量评估,综合评价各模型对不同感官指标的预测效果,结果以均值±标准差表示,具体计算公式如下:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

表 3 各个算法超参数优化范围

Table 3 Hyperparameter optimization ranges for each algorithm

| 超参数 | 指标名称 | XGBoost | CatBoost | LightGBM |
|---|---|---|---|---|
| 基学习器数量 | n_estimators | [50,300] | [50,300] | [50,300] |
| 树深 | max_depth | [2,10] | [2,10] | [2,10] |
| 学习率 | learning_rate | [0.01,0.3] | [0.01,0.3] | [0.01,0.3] |
| L1 正则化系数 | reg_alpha | (0,3] | — | (0,3] |
| L2 正则化系数 | reg_lambda | (0,3] | — | (0,3] |
| L2 叶子正则 | l2_leaf_reg | — | [1,10] | — |

注:—表示该参数未纳入本研究的化化范围。

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

式中,$n$ 为样本数量,$y_i$ 为第 $i$ 个样本真实值,$\hat{y}_i$ 为第 $i$ 个样本的预测值,$\bar{y}$ 为 $n$ 个样本的平均值。

## 2 结果与分析

### 2.1 化学指标相关性分析

根据 1.2 完成样品化学成分检测及感官质量评价获得数据集后,为降低多重共线性对建模效果的影响,本研究对总糖、还原糖等 29 个化学成分指标进行 Spearman 相关性分析,所得化学指标间相关性热图如图 1 所示,* 表示相关性显著($P<0.05$),** 表示相关性极显著($P<0.01$)。由图 1 可知,总糖与还原糖、烟碱与糖碱比、氯与钾氯比、棕榈酸与油酸及亚油酸之间均呈现出较强的相关性($|\rho|>0.85$),表明这些变量之间存在高度共线性。对这些变量剔除冗余变量,还原糖在烟支燃烧过程中可作为美拉德反应的关键物质,生成多种醛类、酮类、吡咯类、吡嗪类及呋喃类风味物质,对提升烟气的香气浓度与口感质量具有显著贡献;相比之下,总糖中所含的蔗糖等非还原糖在燃吸条件下不易生成挥发性风味物质,对感官质量的贡献相对有限,故本研究保留还原糖,剔除总糖[21-22]。在脂肪酸类化合物中,油酸和亚油酸为不饱和脂肪酸,在高温燃吸条件下通过热氧化产生醛类、酮类、呋喃类等香味物质,而棕榈酸作为饱和脂肪酸,在燃吸过程中不易转化为香气味物质,对感官质量的直接贡献较小,故保留油酸和亚油酸,剔除棕榈酸。烟碱和氯是影响烟叶感官质量的核心指标,前者主要决定吸食过程中的生理强度,后者影响烟支燃烧均匀性[23]。尽管衍生指标(如钾氯比及糖碱比)在一定程度上能够反映烟叶成分间的协调性,但与感官指标的直接相关性不强,且在机器学习模型中路径解释较为复杂,故本研究选择保留原始变量烟碱与氯,剔除其衍生指标,以增强模型的解释性与泛化能力。综上,本研究最终保留了还原糖等 25 项自相

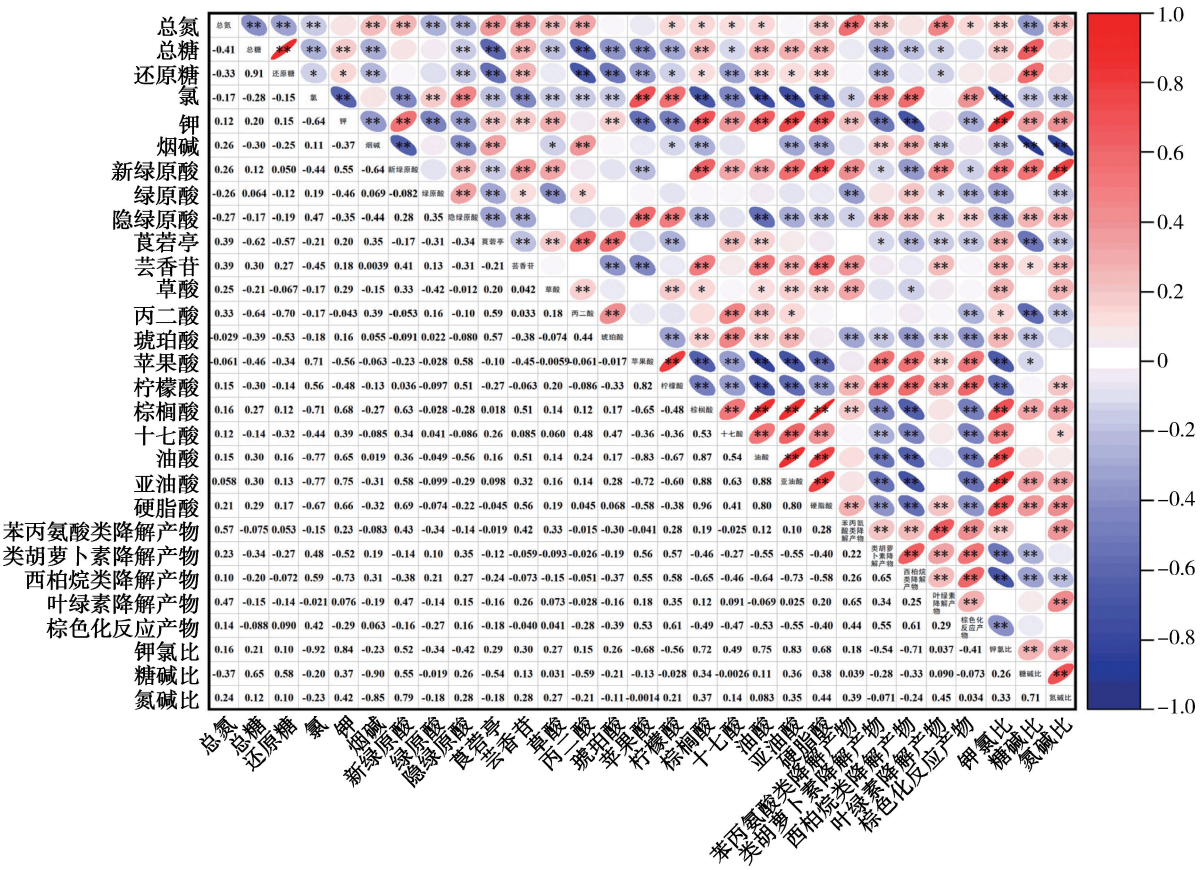


图 1 化学指标间相关性热图

Fig. 1 Correlation heatmap of chemical composition variables

关程度较低、信息代表性较强的化学成分指标，作为后续特征筛选与建模分析的基础变量集。

## 2.2 基于 RFECV-RF 的特征筛选

各感官指标在不同特征数量下的五折交叉验证平均 *RMSE* 变化趋势如图 2 所示。由图 2 可知，所有感官指标的五折交叉验证平均 *RMSE* 曲线在特征数量增加过程中变化趋势一致。当特征数由 1 增至 5 时，各感官指标的 *RMSE* 降幅显著，随后下降趋于平稳，表明早期引入的特征对模型预测性能提升具有较高的边际贡献；而当特征数量达到 10～25 范围时，*RMSE* 曲线斜率趋近于 0，新增变量对模型性能的增益趋于饱和，甚至可能引入冗余变量。以五折交叉验证平均 *RMSE* 最小值对应的特征数量为最优点，确定各感官指标的最佳特征子集规模，结果显示各感官指标最优特征数量在 15～25 之间，说明不同感官指标对化学成分的敏感性存在差异。进一步对比发现，除劲头外，其余感官指标的最优特征子集模型五折交叉验证的平均 *RMSE* 均低于全特征集模型，表明特征筛选有效降低了模型复杂度并显著提升了预测精度。

各感官指标对应的最佳特征组合见表 4。由表 4 可知，在本研究数据与模型条件下，总氮、还原糖、钾与烟碱是多个感官指标最优特征子集中的关键化学成分。在卷烟燃吸条件下，含氮化合物热裂解生成吡啶等碱性物质，从而提高了烟气 pH 值，烟气的刺激感与辛辣感随之增强；烟碱含量则直接影响

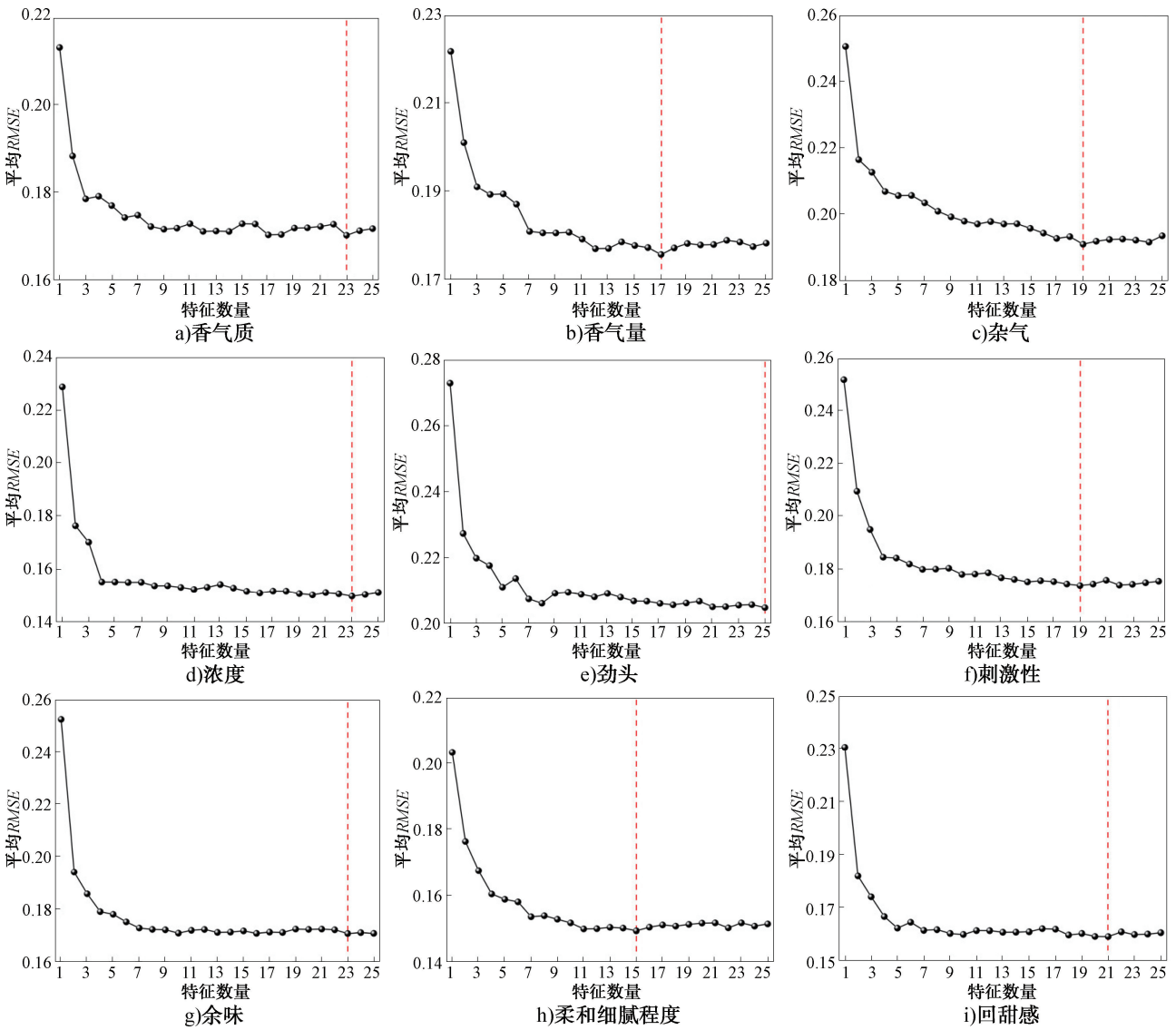


图 2 五折交叉验证平均 *RMSE* 变化趋势图

Fig. 2 Cross-validation curve of RFECV-RF

烟气劲头;还原糖与氨基酸通过美拉德反应可形成吡嗪、吡咯等致香物质,同时还原糖的酸性裂解产物能够中和碱性物质,降低刺激性,提升烟气甜润感与细腻度。钾通过改善烟支的燃烧均匀性,促进香气前体物质降解,减少不完全燃烧产生的杂气和粗糙感[24]。进一步研究发现,香气质、香气量、回甜感等关键指标主要受还原糖、有机酸(苹果酸)与多酚类物质(隐绿原酸)等影响,这些成分可能分别通过美拉德反应、酸碱平衡等路径协同发挥作用[25];浓度、劲头等指标则主要受烟碱、总氮、绿原酸、西柏烷类降解产物等影响。杂气、刺激性等负面感官指标与绿原酸、草酸、琥珀酸、硬脂酸等酸性物质密切相关,这些组分热裂解后产生刺鼻、酸涩的挥发物会加重烟气的刺激性与杂气感[26]。由上可知,模型筛选得到的关键变量与上述机理解释表现出较高的一致性,从理论层面验证了 RFECV-RF 特征筛选结果的准确性。

综上,RFECV-RF 筛选有效剔除了冗余变量,避免了共线性干扰,构建了更为精简、准确且具备机理解释的变量集合,为后续各感官指标的预测模型提供了稳固的数据基础。与传统“黑箱”模型直接输出结果不同,这种基于特征筛选的建模路径还清晰指出预测模型中起主导作用的关键化学成分,为感官机理解析与烟叶组分调控提供了明确指向。

### 2.3 各感官指标预测模型构建与性能评估

本研究进一步基于 XGBoost、CatBoost 和 LightGBM 这 3 种主流 Boosting 算法,分别构建 9 个感官指标的预测模型,全面评估不同集成学习方法在烟叶感官质量预测任务中的适用性与性能差异。各感官指标实测值与预测值对比结果如图 3 所示。由图 3 可知,3 种算法对各感官指标的预测曲线拟合效果良好,预测值与真实值变化趋势吻合度较高,表明基于特征筛选的化学指标有效地捕捉了烟叶化学成分与感官质量之间的非线性关系。3 种 Boosting 算法对各感官指标预测性能见表 5。由表 5 可知,3 种算法在各感官指标上均取得良好预测效果,且 5 次重复训练计算后各性能指标间波动较小,表明所构建模型具有较好的稳定性。然而,不同感官指标之间的预测性能差异显著,且 3 种算法对不同感官指标的侧重不同。XGBoost 算法对香气质、余味和杂气 3 个反映综合复杂风味的感观指标的预测表现最为突出,*MAPE* 分别为 1.06%、1.27%和 1.70%,拟合精度较高,表明 XGBoost 在建模复杂非线性关系及特征交互方面具备显著优势。与其余指标预测结果相比,尽管杂气的预测精度略低($R^2$=0.711 3),

表 4 RFECV-RF 筛选的最佳特征数量组合

Table 4 Optimal feature subset sizes selected by RFECV-RF

| 感官指标 | 最佳特征数量 | 最佳特征(按照特征重要性从大到小排序) |
|---|---|---|
| 香气质 | 23 | 还原糖、苹果酸、钾、隐绿原酸、草酸、氯、芸香苷、绿原酸、烟碱、莨菪亭、琥珀酸、类胡萝卜素降解产物、苯丙氨酸类降解产物、棕色化反应产物、油酸、总氮、柠檬酸、丙二酸、新绿原酸、叶绿素降解产物、硬脂酸、十七酸、氮碱比 |
| 香气量 | 17 | 钾、隐绿原酸、还原糖、芸香苷、苹果酸、莨菪亭、总氮、新绿原酸、十七酸、硬脂酸、烟碱、棕色化反应产物、氮碱比、琥珀酸、油酸、苯丙氨酸类降解产物、西柏烷类降解产物 |
| 杂气 | 19 | 绿原酸、草酸、苹果酸、钾、琥珀酸、还原糖、硬脂酸、柠檬酸、隐绿原酸、氯、棕色化反应产物、莨菪亭、油酸、西柏烷类降解产物、总氮、芸香苷、氮碱比、烟碱、十七酸 |
| 浓度 | 23 | 总氮、烟碱、钾、琥珀酸、绿原酸、亚油酸、西柏烷类降解产物、氮碱比、草酸、十七酸、隐绿原酸、芸香苷、叶绿素降解产物、苯丙氨酸类降解产物、苹果酸、新绿原酸、还原糖、柠檬酸、莨菪亭、丙二酸、棕色化反应产物、油酸、氯 |
| 劲头 | 25 | 烟碱、总氮、绿原酸、还原糖、西柏烷类降解产物、氮碱比、琥珀酸、草酸、隐绿原酸、类胡萝卜素降解产物、叶绿素降解产物、柠檬酸、莨菪亭、硬脂酸、棕色化反应产物、十七酸、苹果酸、芸香苷、油酸、苯丙氨酸类降解产物、钾、新绿原酸、丙二酸、氯、亚油酸 |
| 刺激性 | 19 | 绿原酸、莨菪亭、氯、还原糖、草酸、琥珀酸、总氮、硬脂酸、棕色化反应产物、新绿原酸、钾、十七酸、油酸、烟碱、丙二酸、隐绿原酸、西柏烷类降解产物、类胡萝卜素降解产物、苯丙氨酸类降解产物 |
| 余味 | 23 | 莨菪亭、芸香苷、还原糖、绿原酸、隐绿原酸、叶绿素降解产物、油酸、苹果酸、草酸、烟碱、柠檬酸、新绿原酸、总氮、钾、丙二酸、亚油酸、硬脂酸、氯、十七酸、氮碱比、棕色化反应产物、类胡萝卜素降解产物、西柏烷类降解产物 |
| 柔和细腻程度 | 15 | 总氮、还原糖、莨菪亭、绿原酸、草酸、苹果酸、棕色化反应产物、钾、西柏烷类降解产物、烟碱、硬脂酸、芸香苷、新绿原酸、类胡萝卜素降解产物、丙二酸 |
| 回甜感 | 21 | 芸香苷、还原糖、莨菪亭、总氮、钾、隐绿原酸、苹果酸、草酸、叶绿素降解产物、氯、烟碱、绿原酸、柠檬酸、苯丙氨酸类降解产物、十七酸、新绿原酸、类胡萝卜素降解产物、亚油酸、棕色化反应产物、氮碱比、丙二酸 |

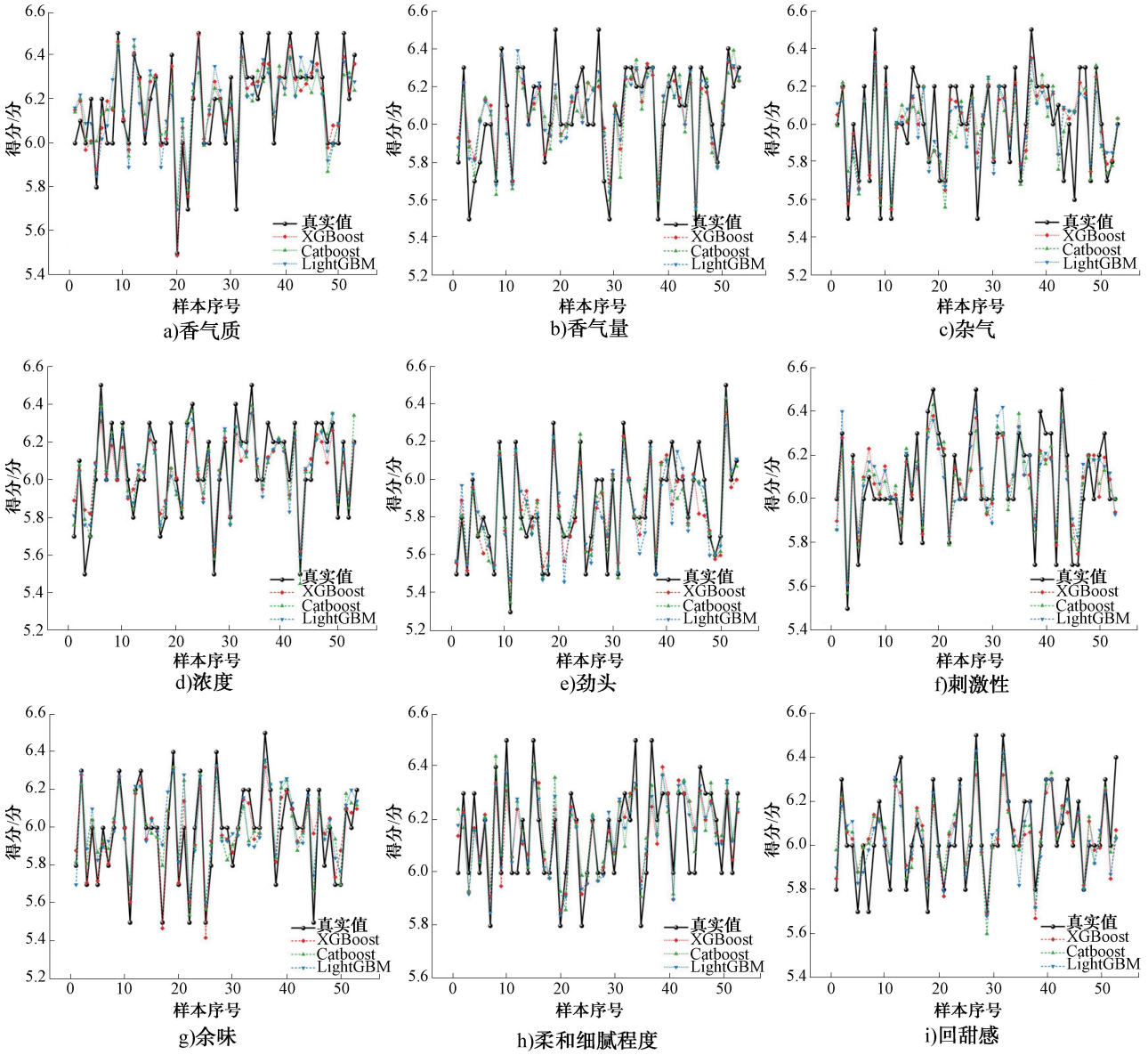


图 3 各感官指标实测值与预测值对比结果

Fig. 3 Comparison between measured and predicted values for each sensory indicator

表 5 Boosting 算法对各感官指标预测性能

Table 5 Predictive performance of boosting algorithms for sensory indicators

| 感官指标 | XGBoost | | | CatBoost | | | LightGBM | | |
|---|---|---|---|---|---|---|---|---|---|
| | *RMSE* | *MAPE*/% | $R^2$ | *RMSE* | *MAPE*/% | $R^2$ | *RMSE* | *MAPE*/% | $R^2$ |
| 香气质 | **0.085 6±0.002 0** | **1.06±0.04** | **0.853 3±0.007 0** | 0.117 7±0.001 7 | 1.63±0.03 | 0.722 4±0.008 1 | 0.115 6±0.001 5 | 1.62±0.03 | 0.732 3±0.007 2 |
| 香气量 | 0.136 0±0.002 2 | 1.79±0.05 | 0.724 7±0.009 1 | 0.149 7±0.000 9 | 2.01±0.03 | 0.666 7±0.003 9 | **0.123 0±0.001 5** | **1.62±0.03** | **0.775 0±0.005 6** |
| 杂气 | **0.140 4±0.000 6** | **1.70±0.03** | **0.711 3±0.002 5** | 0.169 4±0.001 5 | 2.17±0.03 | 0.579 7±0.007 5 | 0.148 9±0.002 4 | 2.06±0.04 | 0.675 3±0.010 6 |
| 浓度 | 0.109 4±0.002 2 | 1.53±0.03 | 0.801 6±0.008 1 | **0.086 8±0.002 8** | **1.12±0.05** | **0.875 0±0.008 1** | 0.098 9±0.001 9 | 1.37±0.04 | 0.837 8±0.006 4 |
| 劲头 | 0.115 2±0.002 1 | 1.52±0.05 | 0.803 1±0.007 3 | **0.084 5±0.001 8** | **1.13±0.03** | **0.894 0±0.004 4** | 0.124 1±0.001 6 | 1.63±0.03 | 0.771 5±0.005 9 |
| 刺激性 | 0.100 5±0.001 7 | 1.41±0.03 | 0.805 9±0.006 6 | **0.094 3±0.002 4** | **1.31±0.04** | **0.829 4±0.008 7** | 0.115 2±0.002 3 | 1.64±0.06 | 0.745 1±0.010 1 |
| 余味 | **0.102 4±0.002 4** | **1.27±0.04** | **0.828 6±0.007 9** | 0.118 0±0.002 2 | 1.59±0.04 | 0.772 2±0.008 3 | 0.134 2±0.002 5 | 1.82±0.05 | 0.705 4±0.011 0 |
| 柔和细腻程度 | 0.105 2±0.001 6 | 1.42±0.03 | 0.683 1±0.009 6 | 0.115 1±0.001 0 | 1.56±0.02 | 0.621 0±0.006 5 | **0.096 7±0.002 2** | **1.27±0.04** | **0.732 5±0.012 5** |
| 回甜感 | 0.112 1±0.002 2 | 1.44±0.05 | 0.704 2±0.011 7 | 0.106 1±0.002 0 | 1.45±0.04 | 0.735 1±0.010 1 | **0.101 0±0.001 4** | **1.27±0.04** | **0.760 1±0.006 7** |

**注:**加粗部分为该指标在 3 种算法下的最优结果。

但仍在可接受范围内,这可能归因于烟叶中杂气类型及成分构成复杂,导致建模难度增大。CatBoost 算法在浓度、劲头和刺激性 3 个感官指标上表现优异,*MAPE* 分别为 1.12%、1.13%和 1.31%。LightGBM 算法则在香气量、柔和细腻程度和回甜感 3 个感官指标的预测上表现更佳,这些指标往往体现出更加微妙和细致的感官特征变化,因而对模型的敏感性和精细化程度要求更高。LightGBM 基于 leaf-wise 的决策树生长策略构建其在处理微小变化时具备更高的灵敏度,有助于捕获复杂细节并提升预测性能[27]。

综上,3 种 Boosting 算法各有适用场景,当关注复杂多元的风味评价时,XGBoost 凭借强大的特征交互建模能力更具优势;对于劲头、浓度和刺激性,CatBoost 凭借在中小规模数据集上的稳健性与高精度更具优势;而针对香气量、柔和细腻度和回甜感这类细腻感官指标,LightGBM 凭借高灵敏度可实现更精准的预测。尽管不同算法在某些指标上表现有差异,但总体上差距有限,均能满足实际预测精度需求。因此,在具体应用中可依据需求灵活选择算法。若追求模型易用性和训练速度,可优先采用 LightGBM;若注重预测稳定性和精度,则 XGBoost 和 CatBoost 更为可靠。总体而言,3 种 Boosting 算法整体上为烟叶感官质量预测提供了有效工具,利用不同算法的融合和优化有望进一步提升预测性能。

## 3 结论

本研究以河南、湖南、云南、贵州 4 个典型产区的 264 份烟叶样品为研究对象,构建基于 RFECV-RF 特征筛选与 Boosting 算法的感官质量预测模型,在小样本、高维稀疏的数据场景下能够对烟叶香气质、香气量、劲头、浓度、杂气、刺激性、余味、柔和细腻度及回甜感 9 项感官指标实现高精度的量化预测,各指标的最优模型在独立测试集上均取得较高精度,$R^2$ 为 0.711 3~0.894 0,*RMSE* 为 0.084 5~0.140 4,*MAPE* 为 1.06%~1.70%。同时,本研究基于 RFECV-RF 特征筛选,精准识别出了影响各感官指标的关键化学成分,从数据驱动角度揭示了烟叶感官品质与理化指标之间的内在联系机制,所构建的感官质量预测框架兼具精度、解释性与工业适用性,为数字化分析烟叶感观质量提供参考。未来研究将进一步拓展样本覆盖范围,融合近红外光谱、高通量成分检测、生态种植因子等多源数据,构建更具普适性与鲁棒性的模型架构。

**参考文献:**

[1] 刘曙光,甘学文,王光耀,等. 基于主要化学成分的醇化片烟感官质量预测模型[J]. 西南农业学报,2020,33(7):1467-1473.
LIU S G,GAN X W,WANG G Y,et al. Construction of sensory quality of tobacco strips during aging based on main chemical constituents[J]. Southwest China Journal of Agricultural Sciences,2020,33(7):1467-1473.

[2] 王建伟,张艳玲,王桂瑶,等. 不同香型产区烤烟高可用性上部烟叶质量特征分析[J]. 烟草科技,2025,58(1):61-68.
WANG J W,ZHANG Y L,WANG G Y,et al. Quality characteristics of upper tobacco leaves with high usability and flavor types from different tobacco growing regions[J]. Tobacco Science & Technology,2025,58(1):61-68.

[3] 郜军艺,彭隆基,胡燕,等. 移栽期对云烟 87 烤后烟叶化学成分及感官质量的影响[J]. 西南农业学报,2024,37(9):2031-2041.
GAO J Y,PENG L J,HU Y,et al. Effect of transplanting period on chemical composition and sensory quality of cured Yunyan 87[J]. Southwest China Journal of Agricultural Sciences,2024,37(9):2031-2041.

[4] 李志伟,陈溪,王鹏泽,等. 基于因子、聚类及判别方法分析烟叶化学和感官质量[J]. 安徽农学通报,2023,29(13):144-149,161.
LI Z W,CHEN X,WANG P Z,et al. Analysis of chemical and sensory quality of tobacco leaves based on factor,clustering and discriminant methods[J]. Anhui Agricultural Science Bulletin,2023,29(13):144-149,161.

[5] 潘义宏,周芳芳,黄坤,等. 连作烤烟根系内次生代谢产物对烤烟品质因子的影响[J]. 西南农业学报,2023,36(10):2167-2174.
PAN Y H,ZHOU F F,HUANG K,et al. Effect of secondary metabolites in root of continuous cropping flue-cured tobacco on its quality factors[J]. Southwest China Journal of Agricultural Sciences,2023,36(10):2167-2174.

[6] HE C L,CHEN R X,REN K,et al. A predictive model for the sensory aroma characteristics of flue-cured tobacco based on a back-propagation neural network[J]. SN Applied Sciences,2020,2(11):1867.

[7] 黄建,杨新士,唐民,等. 江西省烟叶化学指标分析及

感官质量分类模型构建[J]. 湖北农业科学,2024,63(10):153-159.
HUANG J, YANG X S, TANG M, et al. Analysis of chemical indicators of tobacco leaves in Jiangxi province and construction of sensory quality classification model [J]. Hubei Agricultural Sciences, 2024, 63(10): 153-159.

[8] 侯冰清,王硕立,张友杰,等. 基于 BP 神经网络的雪茄原料感官质量预测模型构建[J]. 中国农学通报,2024,40(27):126-133.
HOU B Q, WANG S L, ZHANG Y J, et al. Prediction model of sensory quality of cigar raw materials based on BP neural network [J]. Chinese Agricultural Science Bulletin, 2024, 40(27): 126-133.

[9] 张云伟,张健涛,张海,等. 基于近红外光谱与 Transformer 的烟叶感官指标预测方法[J]. 农业机械学报,2026,57(1):386-396.
ZHANG Y W, ZHANG J T, ZHANG H, et al. Prediction method of tobacco sensory indicators based on near infrared spectroscopy and Transformer[J]. Transactions of the Chinese Society for Agricultural Machinery, 2026, 57(1): 386-396.

[10] 别瑞,周婷云,周显升,等. 基于 XGBoost 算法的山东烟叶质量预测模型初探[J]. 中国烟草科学,2022,43(5):80-86,93.
BIE R, ZHOU T Y, ZHOU X S, et al. Study on quality prediction model of Shandong tobacco based on XGBoost algorithm[J]. Chinese Tobacco Science, 2022, 43(5): 80-86, 93.

[11] 国家烟草专卖局. 烟草及烟草制品 水溶性糖的测定连续流动法:YC/T 159—2019[S]. 北京:中国标准出版社,2019.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of water soluble sugars—Continuous flow method: YC/T 159—2019[S]. Beijing: Standard Press of China, 2019.

[12] 国家烟草专卖局. 烟草及烟草制品 总氮的测定连续流动法:YC/T 161—2002[S]. 北京:中国标准出版社,2002.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of total nitrogen—Continuous flow method: YC/T 161—2002[S]. Beijing: Standard Press of China, 2002.

[13] 国家烟草专卖局. 烟草及烟草制品 总植物碱的测定连续流动(硫氰酸钾)法:YC/T 468—2021[S]. 北京:中国标准出版社,2021.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of total alkaloids—Continuous flow method (potassium thiocyanate): YC/T 468—2021[S]. Beijing: Standard Press of China, 2021.

[14] 国家烟草专卖局. 烟草及烟草制品 氯的测定 连续流动法:YC/T 162—2011[S]. 北京:中国标准出版社,2011.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of chloride—Continuous flow method: YC/T 162—2011[S]. Beijing: Standard Press of China, 2011.

[15] 国家烟草专卖局. 烟草及烟草制品 钾的测定 连续流动法:YC/T 217—2007[S]. 北京:中国标准出版社,2007.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of Potassium—Continuous flow method: YC/T 217—2007[S]. Beijing: Standard Press of China, 2007.

[16] 国家烟草专卖局. 烟草及烟草制品 多酚类化合物 绿原酸、莨菪亭和芸香苷的测定:YC/T 202—2006[S]. 北京:中国标准出版社,2006.
State Tobaao Monopoly Administration. Tobacco and tobacco products—Determination of polyphenols—Chlorogenic acid, scopletin and rutin: YC/T 202—2006[S]. Beijing: Standard Press of China, 2006.

[17] 刘瑞红,潘立宁,王晓瑜,等. 气相色谱测定烟草中非挥发有机酸方法改进[J]. 化学分析计量,2022,31(12):22-28.
LIU R H, PAN L N, WANG X Y, et al. Improvement of the analysis method for non-volatile organic acids in tobacco by gas chromatography [J]. Chemical Analysis and Meterage, 2022, 31(12): 22-28.

[18] 国家烟草专卖局. 烟草及烟草制品 感官评价方法:YC/T 138—1998[S]. 北京:中国标准出版社,1998.
State Tobaao Monopoly Administration. Tobacco and tobacco products—The sensory evaluation methods: YC/T 138—1998[S]. Beijing: Standard Press of China, 1998.

[19] EBRAHIMI WARKIANI M, MOATTAR M H. A comprehensive survey on recent feature selection methods for mixed data: Challenges, solutions and future directions [J]. Neurocomputing, 2025, 623: 129372.

[20] AKIBA T, SANO S, YANASE T, et al. Optuna: A next-generation hyperparameter optimization framework [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA. ACM, 2019: 2623-2631.

[21] BANOŽIĆ M, JOKIĆ S, AČKAR D, et al. Carbohydrates-key players in tobacco aroma formation and quality determination[J]. Molecules, 2020, 25(7): 1734.

[22] LIU T, NIU Y M, CHENG K X, et al. Exploring the formation pathway and antioxidant properties of the sugar-smoking pigment 5-GGMF [J]. Food Chemistry, 2024, 442: 138406.

[23] 谢恒多. 氮钾协同对烟碱合成的影响及 NtQPT2A 上游调控因子的筛选[D]. 成都:四川农业大学,2024.
XIE H D. The synergistic effect of nitrogen and potassium on nicotine synthesis and screening of upstream regulatory factors of NtQPT2A [D]. Chengdu: Sichuan Agricultural University, 2024.

[24] 曹景林,程君奇,李亚培,等. 烤烟常规化学成分与吸食品质关系的研究进展[J]. 湖北农业科学,2020,59(S1):253-258,262.
CAO J L,CHENG J Q,LI Y P,et al. Research progress on the relationship between routine chemical composition and smoking quality of flue-cured tobacco[J]. Hubei Agricultural Sciences,2020,59(S1):253-258,262.

[25] 刘天择,杨菁,汪旭,等. 不同部位烤烟化学成分及热解产物与加热卷烟感官质量的关系[J]. 中国烟草科学,2023,44(1):77-84.
LIU T Z,YANG J,WANG X,et al. Relationships between chemical components and pyrolytic products and sensory quality of heated tobacco of different position flue-cured tobacco leaves[J]. Chinese Tobacco Science,2023,44(1):77-84.

[26] 黄天雄,于洁,贾楠,等. 基于拉曼光谱法所建的多元校正模型预测烟草中绿原酸和芸香苷的含量[J]. 理化检验-化学分册,2022,58(2):210-215.
HUANG T X,YU J,JIA N,et al. Prediction of chlorogenic acid and rutin in tobacco by multivariate calibration model based on Raman spectroscopy[J]. Physical Testing and Chemical Analysis Part B (Chemical Analysis),2022,58(2):210-215.

[27] 朱晓晨,尹奇志,赵福芹,等. 基于 LightGBM 的船舶航速预测模型[J]. 大连海事大学学报,2023,49(1):56-65.
ZHU X C,YIN Q Z,ZHAO F Q,et al. Ship speed prediction model based on LightGBM[J]. Journal of Dalian Maritime University,2023,49(1):56-65.

# Sensory quality prediction of tobacco leaves based on RFECV-RF-Boosting

WANG Longxin[1],FENG Wenning[2],CUI Fuyun[2],LIU Bo[2],ZHAO Hui[2],
SHEN Yujun[3],ZHANG Bohai[1],LAI Miao[1]

1. *College of Tobacco Science,Henan Agricultural University,Zhengzhou* 450002,*China*;
2. *Technology Center,China Tobacco Hebei Industrial Co.,Ltd., Shijiazhuang* 050051,*China*;
3. *Zhengzhou Tobacco Research Institute of CNTC,Zhengzhou* 450001,*China*

**Abstract**:【**Objective**】To address the challenges of subjectivity and data acquisition difficulties in tobacco leaf sensory quality evaluation, and to achieve precise quantitative prediction of tobacco leaf sensory quality based on digital analysis.【**Methods**】A total of 264 tobacco leaf samples from four typical producing regions in China—Henan, Hunan, Yunnan, and Guizhou—were selected for chemical composition analysis and sensory quality evaluation. After removing redundant variables through correlation analysis of chemical indicators, RFECV-RF was used to select the optimal feature subsets for each sensory attribute. Three classic boosting algorithms—XGBoost, CatBoost, and LightGBM—were applied, with hyperparameters optimized using five-fold cross-validation within the Optuna framework to build prediction models for nine sensory attributes.【**Results**】1) Correlation analysis of chemical indices removed four chemical constituent indices, namely total sugar, sugar-to-nicotine ratio, potassium-to-chlorine ratio, and palmitic acid, and retained 25 chemical composition indices, including reducing sugar and nicotine, for subsequent modeling. 2) RFECV-RF feature selection identified the optimal feature subset for each sensory attribute, and further demonstrated that total nitrogen, reducing sugar, potassium, and nicotine were the key chemical constituents affecting tobacco leaf sensory quality. Except for "impact", the root mean square error (RMSE) obtained by cross-validation was lower than that of the full-feature model, indicating that feature selection effectively reduced model complexity and improved prediction accuracy. 3) Under the optimal algorithm, the coefficient of determination ($R^2$) for the sensory attributes ranged from 0.711 3 to 0.894 0, the RMSE ranged from 0.084 5 to 0.140 4, and the mean absolute percentage error (MAPE) ranged from 1.06% to 1.70%, indicating good and stable predictive performance.【**Conclusion**】The prediction model framework enables high-precision quantification of tobacco leaf sensory quality. The research result provide a reference for the digital formulation design and quality control of cigarette products.

**Key words**: chemical contents of tobacco leaves; sensory quality; boosting algorithms; machine learning; feature selection

[责任编辑:刘丹丹 贾学伟]