Statistics and Application 统计学与应用, 2025, 14(6), 156-166
Published Online June 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.146155

Hans汉斯

基于数据挖掘的食品安全数据分析与应用

李志青

广西师范大学数学与统计学院，广西 桂林

收稿日期：2025年5月25日；录用日期：2025年6月17日；发布日期：2025年6月30日

摘　要

食品安全事关国家大计。近年来，相关法律法规不断出台，监管部门也在持续完善监管机制。然而，食品安全依然是一个需要高度重视和严格把控的重要问题。本研究基于机器学习方法，构建了多种分类模型，包括支持向量机(SVM)、随机森林(RF)、LightGBM和XGBoost，并对其性能进行了系统比较。结果表明，随机森林模型表现最佳，其次为支持向量机和XGBoost，而LightGBM的性能相对较差。为了进一步提升预测能力，研究进一步构建了Stacking融合模型，将随机森林、XGBoost和支持向量机的预测结果进行集成，并采用随机森林作为元学习器。实验结果显示，Stacking模型显著提高了整体预测性能。本研究结果为食品安全监管提供了重要参考，验证了数据挖掘技术在食品安全领域的有效性。相关技术可辅助监管部门实现食品安全问题的及时预测与处理，从而提升监管工作的效率与准确性。

关键词

食品安全，机器学习，Stacking融合模型

Data Mining-Based Food Safety Data Analysis and Application

Zhiqing Li

School of Mathematics and Statistics, Guangxi Normal University, Guilin Guangxi

Received: May 25th, 2025; accepted: Jun. 17th, 2025; published: Jun. 30th, 2025

Abstract

Food safety is a matter of national importance. In recent years, relevant laws and regulations have been continuously introduced, and regulatory authorities have been constantly improving the regulatory mechanisms. However, food safety remains a critical issue that requires high attention and strict control. This study, based on machine learning methods, constructed various classification models, including Support Vector Machine (SVM), Random Forest (RF), LightGBM, and XGBoost, and

文章引用: 李志青. 基于数据挖掘的食品安全数据分析与应用[J]. 统计学与应用, 2025, 14(6): 156-166.
DOI: 10.12677/sa.2025.146155

systematically compared their performance. The results showed that the Random Forest model performed the best, followed by Support Vector Machine and XGBoost, while LightGBM had relatively poorer performance. To further enhance the predictive capability, the study constructed a Stacking ensemble model, integrating the prediction results of Random Forest, XGBoost, and Support Vector Machine, with Random Forest serving as the meta-learner. Experimental results indicated that the Stacking model significantly improved overall predictive performance. The findings of this study provide important references for food safety regulation and validate the effectiveness of data mining techniques in the field of food safety. The related technologies can assist regulatory authorities in timely predicting and handling food safety issues, thereby improving the efficiency and accuracy of regulatory work.

Keywords

Food Safety, Machine Learning, Stacking Ensemble Model

1. 引言

(一) 问题描述

食品安全问题始终是我国重要的民生议题之一。食品安全直接关系到人民群众的身体健康，不合格食品可能引发多种健康风险，甚至诱发严重疾病。在食品安全保障方面，政府需确保食品供应来源的安全可靠，生产加工过程的合法合规，以确保公众所摄取食品的安全性和无害性。

自新中国成立以来，我国围绕食品安全问题相继制定并实施了一系列有针对性的政策与战略，以提升食品供应体系的规模化、规范化与安全化水平。早期，国家主要通过制定食品卫生相关规章与标准开展监管工作。从 1953 年卫生部颁布的《通知》和《暂行办法》，到 1964 年国务院转发的《食品卫生管理试行条例》，再到 1978 年改革开放后出台的各项政策措施，我国食品安全制度体系逐步建立，初步明确了食品违法行为的法律责任。然而，由于当时法律制度尚不健全，监管执行效果仍存在不足。

党的十八大以来，国家高度重视食品安全问题。2012 年，十八大报告首次提出建立统一的食品安全监管执法机制，以保障“舌尖上的安全”。2015 年，党的十八届五中全会通过“十三五”规划纲要，明确提出“实施食品安全战略”，并将其纳入“健康中国战略”之中。同期，《食品安全法》进行了重要修订，提出“预防为主、全过程控制、社会共治”的新理念，有效推动了食品安全治理体系和治理能力的现代化。

在党的十九大报告中，习近平总书记于 2017 年再次强调实施食品安全战略，进一步凸显保障食品安全在国家发展和民生福祉中的核心地位。自 2019 年以来，国家持续加强食品安全监管体系建设，进一步完善相关法律法规，加大监督执法力度，推动食品行业标准的优化与落地，强化生产、流通和销售各环节的质量安全责任追究机制，并对违规企业实施严厉惩处。

(二) 文献综述

食品安全问题一直以来都是全球关注的核心议题，它直接关系到民众的健康与生命安全，是社会稳定和经济发展的基础保障。随着市场经济的快速发展，特别是自 20 世纪 80 年代以来，诸如三聚氰胺奶粉、增白面粉、地沟油等食品安全事件频繁爆发，严重损害了消费者的信任，并暴露了现行食品安全监管体系中的诸多漏洞。这些事件的发生突显了食品安全监管体系的不足，亟需制定更为严格和有效的监管标准。

随着食品安全问题的日益严重，各国对食品安全问题的关注也日渐加深。欧美等发达国家，尤其是在疯牛病和禽流感等重大公共卫生事件发生后，已经开始加强食品安全监管。例如，欧盟于 20 世纪末成立了专门的食品安全监管机构——欧盟食品安全局(EFSA)，该机构负责执行全面的食品安全监管和风险评估。中国在此背景下，也不断完善食品安全监管体系，推动食品质量安全的现代化与信息化建设。近年来，随着监管系统的信息化发展，数据分析和预测技术逐渐成为食品安全管理的关键工具。

在食品安全监管过程中，每年都会产生大量的检测数据，这些数据具有信息量大、更新迅速、范围广泛以及突发性强等特点。如何从这些海量数据中有效提取有价值的信息，成为当前亟待解决的关键课题。数据挖掘技术，特别是机器学习方法，已被广泛应用于食品安全领域，显著提升了食品抽检和质量监控的智能化水平。例如，史运涛等[1]人提出了一种基于知识图谱的注意力网络模型，通过智能分析辅助食品安全风险评估，为监管部门提供决策支持。

在中国，数据挖掘技术逐步被应用于食品安全抽检工作，并帮助监管部门从海量数据中提取有价值的信息。尤其是在基于集成学习的机器学习方法方面，算法如随机森林(Random Forest)、XGBoost 和 LightGBM 已在食品安全数据分析中取得了显著成果。近年来，Stacking 融合模型也被提出并广泛应用，以进一步提高预测准确性。该模型通过结合多种基础学习模型的输出，能够在复杂的多分类任务中实现更优的预测效果。

2. 食品安全样本数据的收集处理与分析

(一) 数据来源

本文所使用的数据来源于广西壮族自治区市场监督管理局发布的食品安全抽检不合格信息。原始数据包含 12 项属性信息，包括：序号、标称生产企业名称、标称生产企业地址、被抽样单位名称、被抽样单位地址、食品名称、规格型号、标称商标、标称生产日期及批号、不合格项目、检验机构以及备注等。该数据为后续模型构建与分析提供了坚实的数据基础。

(二) 数据预处理

1) 数据清洗

本文对原始数据进行了预处理，首先剔除与研究无关的餐具类样本以及存在缺失值的记录，同时将不合格项目数量在两个及以上的样本进行拆分与补充处理。经过上述步骤，最终获得 579 条有效数据。根据数据缺失情况可视化图 1 可见，处理后的数据集已不存在缺失值，为后续建模分析提供了完整的数据基础。

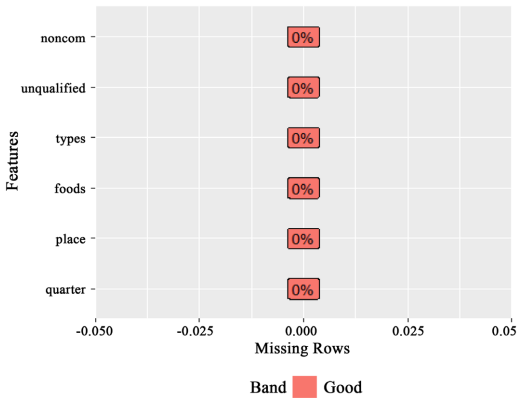

Figure 1. Chart of missing data
图 1. 数据缺失情况图

2) 数据规约

数据集中存在一些与研究目标不相关的指标。例如，生产季度、地址、食品名称、食品种类和不合格项目是与研究目标密切相关的关键指标；而生产企业名称、生产企业地址、被抽样单位名称、规格型号、商标、生产批号和备注等属性则不属于研究关注的内容。

鉴于食品名称种类繁多，本文对食品名称进行了归约处理。采用层次规约方法，将较低层次的食品类别归约至更高层次的食品类别，如图 2 所示。具体而言，本文考虑的食品安全检测数据集中的食品类别主要包括蔬菜、水产品、饮料、糕点和粮食加工品，每一类食品下均包含一些更细分的低层次食品类别。例如，调味品类别可以细分为酱油、盐巴等低层次类别。食品名称归约后的数据集格式见表 1 所示。

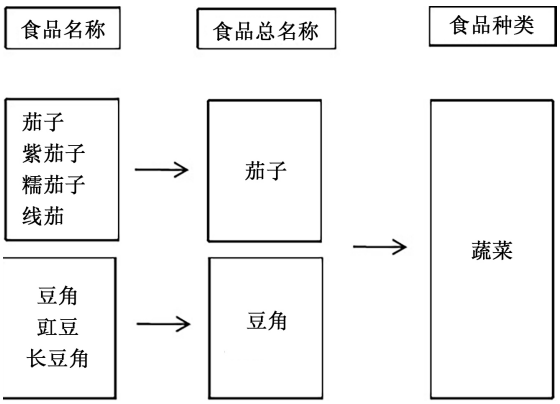

Figure 2. Example of food hierarchy reduction
图 2. 食品层次归约示例

Table 1. Partial dataset after food normalization
表 1. 食品规约后部分数据集

序号	生产时间	地址	食品名称	食品种类	不合格项目
1	2024.01.06	白色市	手磨浓香黑芝麻糊	方便食品	霉菌
2	2024.01.06	北海市	皮皮虾	水产品	镉
3	2024.01.06	钦州市	生姜	蔬菜	铅
4	2024.01.06	柳州市	大青椒	蔬菜	克百威
5	2024.01.06	桂林市	草鱼	水产品	孔雀石绿
6	2024.01.06	来宾市	木鸭(鸭肉)	家禽	呋喃唑酮代谢物
7	2024.01.06	梧州市	梧州白米醋	调味品	总酸
8	2024.01.06	梧州市	来利太平米饼特制紫薯饼	糕点	霉菌
9	2024.01.06	梧州市	来利太平米饼糯米肉饼	糕点	霉菌
10	2024.01.06	钦州市	皮皮虾	水产品	镉

将食品生产时间进行了归约处理，将其分为四个季度：第一季度(12 月至 2 月)、第二季度(3 月至 5 月)、第三季度(6 月至 8 月)和第四季度(9 月至 11 月)。归约后如表 2 所示。

Table 2. Production time normalization dataset
表 2. 生产时间规约数据集

序号	生产季度	地址	食品名称	食品种类	不合格项目
1	4	北海市	食用油	食用油	酸价
2	4	北海市	冰糖	食用糖	还原糖分

续表

3	4	河池市	食用油	食用油	苯并芘
4	4	南宁市	丝瓜	蔬菜	氯氟氰菊酯
5	4	梧州市	湿米粉	粮食加工品	脱氢乙酸
6	4	玉林市	咸水角	油炸食品	铝的残留量
7	4	南宁市	馒头	糕点	糖精钠
8	4	桂林市	油条	油炸食品	铝的残留量
9	4	桂林市	麻圆	油炸食品	铝的残留量
10	4	桂林市	油条	油炸食品	铝的残留量

本文将食品检测指标划分为九个大类，分别为：微生物、添加剂、理化、农药残留、兽药残留、金属元素、生物毒素、污染物和非法添加物。每个大类所包含的部分不合格指标详见表 3。基于表 3 中的分类结果，对原始数据进行规约处理，部分处理结果如表 4 所示。

Table 3. Partial classification table of nonconforming indicators
表 3. 不合格指标分类表部分

分类 1	分类 2
微生物	菌落总数、大肠菌群、铜绿假单胞菌、霉菌、商业无菌等
添加剂	苯甲酸、山梨酸、安赛蜜、亚硝酸钠、脱氢乙酸、三氯蔗糖、纳他霉素、柠檬黄等
理化	酒精度、氨基酸态氮、过氧化值、二氧化硫、还原糖分、溶剂残留、耗氧量等
农药残留	甲拌磷、多菌灵、毒死蜱、克百威、氯氰菊酯和高效氯氰菊酯、水胺硫磷等
兽药残留	氟苯尼考、硝基呋喃代谢物、氯霉素、克伦特罗、地西泮、恩诺沙星、孔雀石绿等
金属元素	铝、铅、镉、锌、无机砷、总砷、钠、铬、硒、锂、锶等
生物毒素	黄曲霉毒素、玉米赤霉烯酮、赭曲霉毒素、脱氧雪腐镰刀菌烯醇等
污染物	苯并芘、二甲基亚硝胺、溴酸盐等
非法添加物	罂粟碱、氯苯氧乙酸钠、苄基嘌呤、过氧化苯甲酰、罗丹明等

Table 4. Normalized data of nonconforming items
表 4. 不合格项目规约数据

生产季度	地址	食品名称	食品种类	不合格项目	不合格项目种类
4	南宁市	蔬菜	蔬菜	氯氟氰菊酯	农药残留
4	梧州市	湿米粉	粮食加工品	脱氢乙酸	添加剂
4	南宁市	馒头	糕点	糖精钠	添加剂
1	钦州市	面包	糕点	菌落总数	微生物
1	玉林市	湿米粉	粮食加工品	脱氢乙酸	添加剂
1	崇左市	蔬菜	蔬菜	二氧化硫残留量	理化
2	南宁市	粉	粮食加工品	菌落总数	微生物
2	南宁市	绿豆糕	糕点	过氧化值	理化
3	南宁市	黄骨鱼	水产品	孔雀石绿	兽药残留
3	南宁市	豆角	蔬菜	啶虫脒	农药残留
3	南宁市	饮用水	饮料	铜绿假单胞菌	微生物

3) 数据转换

鉴于 RStudio 对中文字符的兼容性较弱，为确保数据处理的稳定性与准确性，本文对数据中的部分中文属性值进行了编码转化处理。经过转换后，最终形成适用于数据挖掘分析的数据集。

3. 多分类模型在食品安全数据分析中的应用与比较

(一) 多分类模型理论

1) LightGBM 模型原理

LightGBM (Light Gradient Boosting Machine)是一种基于梯度提升框架的高效算法，广泛应用于处理大规模和高维数据(Ke *et al*., 2017) [2]。与传统的梯度提升决策树(GBDT)方法相比，LightGBM 在多个方面展现出显著优势。其核心创新包括基于直方图的特征分割方法，能够显著减少内存消耗并提高计算效率。此外，LightGBM 采用 leaf-wise 生长策略，而非传统的 level-wise 策略，这使得其能够在同等计算资源下生成更深的树，提高模型的准确性和拟合能力。在多分类任务中，LightGBM 通过构建多棵决策树，利用 Softmax 函数输出类别的概率分布，同时采用交叉熵作为损失函数，从而优化分类性能(Chen & Guestrin, 2016) [3]。由于这些技术创新，LightGBM 在训练速度、内存使用效率和模型的泛化能力方面，相较于传统 GBDT 方法具有显著的优势。

2) XGboost 模型原理

XGBoost 是一种改进的梯度提升决策树(GBDT)算法。其训练过程中，每棵新树拟合前一轮模型的残差，从而不断优化预测结果。在多分类任务中，XGBoost 通过构建多棵决策树来估计各类别的预测概率，并采用带正则项的 Softmax 损失函数为 $\mathcal{L}(\phi)=-\sum_{i=1}^{N}\sum_{k=1}^{K}y_{i,k}\log\left(p_{i,k}\right)+\Phi(\phi)$。其中，$\mathrm{y_{i,k}}$ 为样本的真实标签，$p_{i,k}$ 为模型预测概率，$\Phi(\phi)$ 为正则项。第 t 轮的目标函数为 $Obj^{(t)}=\sum_{i=1}^{N}\mathcal{L}\left(y_i,\hat{y}_i^{(t-1)}+T\left(x_i;b\right)\right)+\Phi(\phi)$。在每轮迭代中，模型计算一阶梯度 $g_{i,k}$ 和二阶导数 $h_{i,k}$ 以进行最优划分 $\mathrm{g_{i,k}}=\frac{\partial\mathcal{L}}{\partial\hat{y}_{i,k}},\mathrm{h_{i,k}}=\frac{\partial^2\mathcal{L}}{\partial\hat{y}_{i,k}^2}$。最终，通过对所

有树的输出进行 Softmax 变换，获得每个样本的类别概率分布。

3) 随机森林模型原理

随机森林(Random Forest)是一种基于多棵决策树的集成学习方法，由 Leo Breiman 在 2001 年提出[4]。它通过构建多个决策树并结合它们的预测结果，来提升分类或回归任务的准确性和稳健性。随机森林能够处理分类和回归问题。此外，通过评估特征的重要性，随机森林可以帮助识别最具影响力的特征，辅助特征工程。其主要优点包括高准确性、抗过拟合能力强以及能够有效处理大规模的高维数据。

4) 支持向量机模型原理

支持向量机(Support Vector Machine, SVM)是一种由 Corinna Cortes 等人在 1995 年首次提出的算法，广泛应用于分类和回归问题[5]。支持向量机(SVM)是一种广泛应用于分类和回归任务的强大学习算法，尤其在分类任务中表现出色。对于线性可分的数据，SVM 通过最小化 $\frac{1}{2}|w|^2$ 并满足约束条件 $y_i\left(w\cdot x_i+b\right)\geq 1$ 来寻找最优超平面，其中 w 是超平面的法向量，b 是偏置项，y_i 和 x_i 分别是数据点的类别标签和特征向量。在处理非线性可分数据时，SVM 利用核函数将数据映射到高维空间以实现线性可分，常用的核函数有多项式核 $K(x,y)=(x\cdot y+c)^d$、径向基函数(RBF)核 $K(x,y)=\exp\left(-\gamma|x-y|^2\right)$ 和线性核 $K(x,y)=x\cdot y$。SVM 通过最大化间隔来增强泛化能力，减少过拟合风险，且在处理高维数据时表现出色，适用于如文本分类等任务。然而，SVM 对核函数及其参数的选择非常敏感，且在大规模数据集上的训练过程计算复杂度较高，因为需要求解二次规划问题。

5) Stacking 融合模型

Stacking (叠加泛化)是一种层次化的集成学习方法，旨在通过组合多个基础模型的预测结果来提高整

体性能[6]。该方法首先将多个第一层模型的输出作为输入特征，训练第二层的元模型(Meta-model)进行最终预测，从而充分发挥各基础模型的优势互补。与传统的单一模型相比，Stacking 能够显著增强模型的泛化能力，降低过拟合的风险，因此在处理复杂的机器学习任务时表现优越。Stacking 方法已被广泛应用于分类、回归以及其他各种机器学习问题中，成为集成学习中一种有效的策略[7]。

(二) 实证分析

1) 对训练集进行检验

由表 5 和图 3 训练集上模型对比可知，在训练集上，XGBoost 模型的分类准确率最高，其次为随机森林模型，支持向量机表现稍逊，LightGBM 模型的性能最弱。Kappa 值及其他评估指标的表现与准确率结果基本一致。因此，综合训练集上的各项性能指标，XGBoost 模型表现最佳，其次为随机森林和支持向量机，LightGBM 模型性能相对较差。

Table 5. Model comparison on the test set
表 5. 测试集上模型对比

	LightGBM	XGboost	随机森林	支持向量机
准确率	0.7917	0.9000	0.8924	0.8832
Kappa	0.7421	0.8759	0.8676	0.8561
平均 F1 分数	0.6994	0.8754	0.8700	0.8565
平均敏感性	0.7721	0.8651	0.8639	0.8471
平均特异性	0.9576	0.9792	0.9780	0.9759
平均正预测值	0.7587	0.8994	0.8850	0.8848
平均负预测值	0.9614	0.9811	0.9793	0.9778
平均精确率	0.7587	0.8994	0.8849	0.8848
平均召回率	0.7221	0.8650	0.8639	0.8471
平均检测率	0.1319	0.1500	0.1487	0.1472
平均平衡准确率	0.8398	0.9221	0.9210	0.9115
ROC 值	0.7460	0.9849	0.9728	0.9809

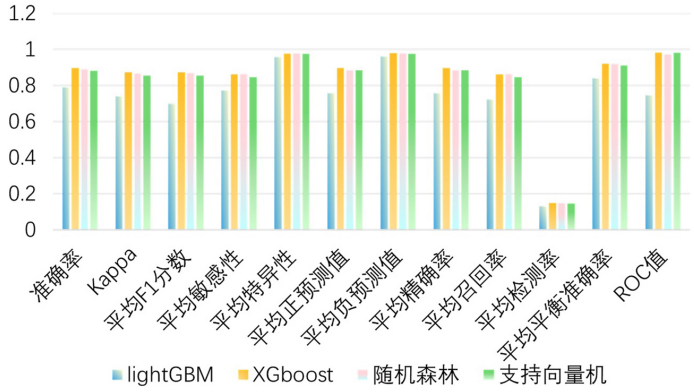

Figure 3. Model comparison on the training set
图 3. 训练集上模型对比

2) 对测试集进行检验

由表 6 和图 4 可见，在测试集上，随机森林模型的分类准确率最高，其次为支持向量机和 XGBoost 模型，LightGBM 模型表现最差。Kappa 值及其他评估指标的结果亦与准确率一致。因此，在测试集上，随机森林模型综合性能最优，具有较好的泛化能力。

Table 6. Model comparison on the test set
表 6. 测试集上模型对比

	LightGBM	XGboost	随机森林	支持向量机
准确率	0.7535	0.7746	0.8924	0.8239
Kappa	0.6953	0.7216	0.8676	0.7824
平均 F1 分数	0.6700	0.7209	0.8700	0.7844
平均敏感性	0.6765	0.7197	0.8639	0.7768
平均特异性	0.9506	0.9542	0.9780	0.9639
平均正预测值	0.7008	0.7334	0.8850	0.8125
平均负预测值	0.9534	0.9565	0.9793	0.9667
平均精确率	0.7008	0.7334	0.8849	0.8125
平均召回率	0.6765	0.7197	0.8639	0.7768
平均检测率	0.1255	0.1291	0.1487	0.1373
平均平衡准确率	0.8136	0.8369	0.9210	0.8704
ROC 值	0.6627	0.9325	0.9276	0.9451

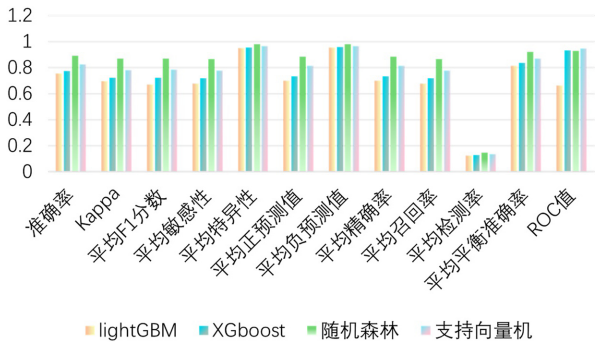

Figure 4. Comparison chart of single models on the test set
图 4. 单一模型测试集对比图

4. 融合模型及变量重要性分析

(一) 融合模型理论介绍

通过对上述单一模型的比较，我们发现各模型在性能上存在一定差异。为了进一步提升预测性能，本文尝试采用 Stacking 融合模型方法，将这些单一模型进行集成，以期获得更优的效果。在模型选择上，我们主要考虑其学习能力，并基于实验结果摒弃了表现较差的 LightGBM 模型。最终，选择随机森林、XGBoost 和支持向量机作为第一层基模型，并将随机森林作为第二层元模型。具体的融合过程如图 5 所示。

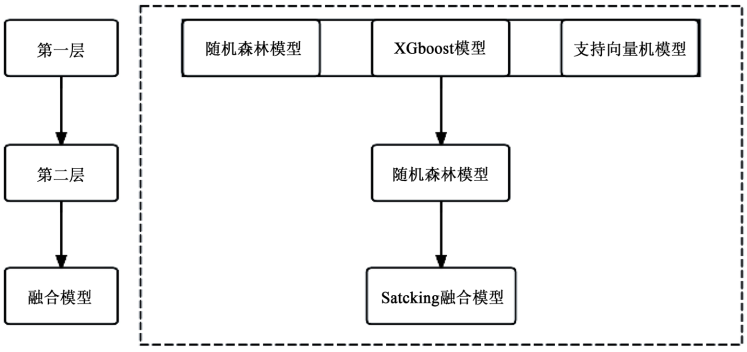

Figure 5. Flowchart of the Stacking ensemble model
图 5. Stacking 融合模型流程图

(二) 实证分析

1) 对训练集进行检验

由表 7 可知，Stacking 融合模型在训练集上的各项综合评价指标均表现良好，其中分类准确率为92.21%，Kappa 值为 0.9045，ROC 值达到 0.96，显示出该模型具有较高的准确性与稳定性。上述结果表明，Stacking 融合模型在食品安全预测任务中具备良好的应用潜力，可为食品安全监管提供可靠的决策支持工具。综合比较各模型性能，Stacking 融合模型整体表现优于其余三个单一模型。

Table 7. Comprehensive results of the Stacking ensemble model and single models on the training set
表 7. Stacking 融合模型及单一模型训练集综合结果

	LightGBM	XGboost	随机森林	支持向量机	Stacking
准确率	0.7917	0.9000	0.8924	0.8832	0.9221
Kappa	0.7421	0.8759	0.8676	0.8561	0.9045
平均 F1 分数	0.6994	0.8754	0.8700	0.8565	0.9081
平均敏感性	0.7721	0.8651	0.8639	0.8471	0.9044
平均特异性	0.9576	0.9792	0.9780	0.9759	0.9843
平均正预测值	0.7587	0.8994	0.8850	0.8848	0.9137
平均负预测值	0.9614	0.9811	0.9793	0.9778	0.9849
平均精确率	0.7587	0.8994	0.8849	0.8848	0.9137
平均召回率	0.7221	0.8650	0.8639	0.8471	0.9044
平均检测率	0.1319	0.1500	0.1487	0.1472	0.1537
平均平衡准确率	0.8398	0.9221	0.9210	0.9115	0.9444
ROC 值	0.7460	0.9849	0.9728	0.9809	0.9600

2) 对测试集进行检验

表 8 展示了测试集上各模型的预测效果综合结果。可以看到，Stacking 融合模型在测试集上的分类准确率为 86.62%，Kappa 值为 0.8352，ROC 值为 0.9195。与其他三个模型相比，Stacking 融合模型在测试集上的表现最为优异。综合来看，无论是在训练集还是测试集上，Stacking 融合模型均表现为最优模型。

Table 8. Comprehensive prediction results of the Stacking ensemble model on the test set
表 8. Stacking 融合模型测试集预测综合结果

	LightGBM	XGboost	随机森林	支持向量机	Stacking
准确率	0.7535	0.7746	0.8924	0.8239	0.8662
Kappa	0.6953	0.7216	0.8676	0.7824	0.8352
平均 F1 分数	0.6700	0.7209	0.8700	0.7844	0.8444
平均敏感性	0.6765	0.7197	0.8639	0.7768	0.8347
平均特异性	0.9506	0.9542	0.9780	0.9639	0.9726
平均正预测值	0.7008	0.7334	0.8850	0.8125	0.8614
平均负预测值	0.9534	0.9565	0.9793	0.9667	0.9742
平均精确率	0.7008	0.7334	0.8849	0.8125	0.8613
平均召回率	0.6765	0.7197	0.8639	0.7768	0.8347
平均检测率	0.1255	0.1291	0.1487	0.1373	0.1444
平均平衡准确率	0.8136	0.8369	0.9210	0.8704	0.9036
ROC 值	0.6627	0.9325	0.9276	0.9451	0.9195

(三) 变量重要性

1) 随机森林变量重要性分析

输入变量的重要性如表 9 所示。从表中可以看出，在金属元素、微生物以及其他类别的输入变量中，食品名称的重要性最高；而在添加剂、兽药残留和农药残留的输入变量中，食品种类的重要性最大。

Table 9. Importance of input variables
表 9. 输入变量重要性

	金属元素	微生物	添加剂	兽药残留	农药残留	其他
生产季节	1.46	20.45	22.74	2.56	13.28	6.46
地址	14.29	14.56	15.93	5.35	0.52	10.43
食品名称	41.06	58.04	49.99	25.56	34.05	22.37
食品种类	40.68	57.81	52.70	480.29	46.95	11.35

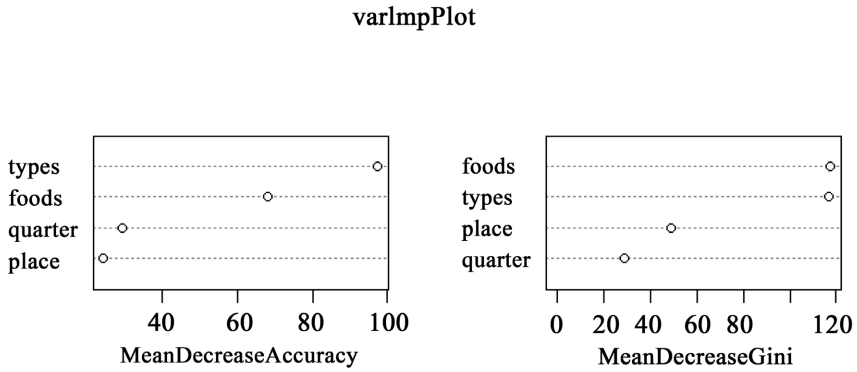

Figure 6. Measurement chart of input variable importance for random forest
图 6. 随机森林输入变量重要性测度图

见图 6，无论是对输出变量的预测精度，还是输出变量差异性下降的程度，食品种类和食品品种的影响均较为显著，显示出这两个变量在模型中的重要性。

2) XGboost 变量重要性分析

Gain 系数通常用于衡量各输入特征在模型中的相对重要性，数值越高，说明该特征对模型性能的贡献越大，具有更强的分类判别能力。表 10 列出了 Gain 系数排名前十的输入变量。从表中可以看出，食品种类中 Gain 值较高的包括蔬菜、饮料、糕点和粮食加工品，表明这些类别在食品安全问题中可能具有较高的风险，需予以重点关注；而地址属性中 Gain 值较大的为南宁和桂林，提示这两个地区可能存在较多食品安全隐患，有必要加强相关监管与风险防控措施。

Table 10. Ranking table of input variable Gain coefficients
表 10. 输入变量 Gain 系数排序表

变量	Gain 系数	变量	Gain 系数
蔬菜	0.154208416	第三季度	0.036471281
饮料	0.091160417	第一季度	0.031452919
糕点	0.084162458	桂林	0.031342253
粮食加工品	0.075453104	豆角	0.028963576
南宁	0.052946229	第四季度	0.028631852

5. 结论

本研究基于广西食品安全抽检数据，构建并比较了多种机器学习分类模型在食品不合格项目识别中的应用效果。结果表明，随机森林模型在测试集上表现最优，支持向量机和 XGBoost 次之，LightGBM 相对较弱。为进一步提升模型性能，研究引入 Stacking 融合模型，将随机森林、XGBoost 和支持向量机作为基模型，随机森林作为元模型进行融合。融合模型在训练集和测试集上均取得最优性能，准确率分别达 92.21%和 86.62%，显示出更强的泛化能力与稳定性。研究验证了机器学习，特别是集成学习方法在食品安全风险识别中的有效性，为监管部门提供了数据驱动的决策支持，有助于提升食品安全监管的科学性与精准性。未来研究可进一步探索深度学习与时空特征建模，以提升预测精度和动态响应能力。

参考文献

[1] 史运涛, 刘召, 李书钦, 等. 基于知识图谱注意力网络的食品安全风险评估模型[J]. 食品工业, 2021, 42(12): 471-475.

[2] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 3146-3154.

[3] Chen, T. and Guestrin, C. (2016) XGBoost. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[4] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[5] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1007/bf00994018

[6] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. https://doi.org/10.1016/s0893-6080(05)80023-1

[7] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/bf00058655