文章编号:1008-1534(2025)02-0111-09

基于 Stacking 模型的早产预测方法

马金龙[1],史晓月[1],杜丽佳[2],王胜普[2],杨志芬[2]

(1. 河北科技大学信息科学与工程学院,河北石家庄 050018;2. 河北医科大学第四医院产科,河北石家庄 050035)

摘 要:为了解决传统机器学习模型在早产预测时综合性能不足的问题,提出一种基于 Stacking 模型的早产预测方法。首先,在数据预处理阶段,采用欠采样技术平衡正、负样本分布,并通过数据标准化消除变量间的数值差异;其次,通过分析特征之间的相关性和特征重要性分数,进行特征选择;然后,在 Stacking 模型构建时,通过分析机器学习算法预测结果间的皮尔逊相关系数,调整基分类器的类型和数量;最后,利用多种评价指标对基于 Stacking 模型的早产预测方法进行全面评估,并将其与现有方法对比分析,验证该方法的有效性。结果表明:所提方法在 ROC 曲线下面积(area under the curve, AUC)、准确率(Accuracy)、$F1$ 值和召回率(Recall)方面,分别达到了 0.921 9、0.922 9、0.916 4 和 0.858 5,均优于搭建 Stacking 模型所用的 11 个单一模型的最佳表现,且整体性能优于现有研究方法。所提方法能够高效识别孕早期的早产高风险人群,为早产的提前干预提供有力支持。

关键词:人工智能其他学科;机器学习;集成学习;神经网络模型;早产预测

中图分类号:TP181 **文献标识码**:A **DOI**: 10.7535/hbgykj.2025yx02002

Preterm birth prediction framework under Stacking model

MA Jinlong[1], SHI Xiaoyue[1], DU Lijia[2], WANG Shengpu[2], YANG Zhifen[2]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China; 2. Obstetrical Department, The Fourth Hospital of Hebei Medical University, Shijiazhuang, Hebei 050035, China)

Abstract: To address the issue of insufficient overall performance of traditional machine learning algorithms in preterm birth prediction, an innovative preterm birth prediction method based on Stacking model was proposed. Firstly, during the data preprocessing stage, an under-sampling technique was applied to balance the distribution of positive and negative samples, and numerical differences between variables were eliminated through data standardization. Secondly, feature selection was carried out by carefully analyzing the correlations between features and assessing their importance scores. Then, in the construction of the Stacking model, the Pearson correlation coefficient was calculated among the prediction results of different machine learning algorithms, and this analysis was used to adjust both the type and number of base classifiers. Finally, a comprehensive evaluation of the preterm birth prediction method based on the Stacking model was conducted using multiple evaluation

收稿日期:2024-03-14;修回日期:2024-10-25;责任编辑:丁军苗

基金项目:河北省自然科学基金(H2022206212,H2022206600);河北省医学科学研究课题计划 (20210715,20230775,20240817)

第一作者简介:马金龙(1981—),男,河北定州人,副教授,博士,主要从事生物信息学方面的研究。

通信作者:杨志芬副教授。E-mail:cxsxfy@126.com

马金龙,史晓月,杜丽佳,等. 基于 Stacking 模型的早产预测方法[J]. 河北工业科技,2025,42(2):111-119.

MA Jinlong, SHI Xiaoyue, DU Lijia, et al. Preterm birth prediction framework under Stacking model[J]. Hebei Journal of Industrial Science and Technology, 2025, 42(2): 111-119.

indicators, and compared and analyzed with existing methods to verify the effectiveness of the method. The results show that the proposed method achieves remarkable performance, with scores of 0.921 9 in AUC, 0.922 9 in Accuracy, 0.916 4 in $F1$ score, and 0.858 5 in Recall. These results significantly outperform the best performances of the 11 individual models used to build the Stacking model, and the overall performance is better than the existing research methods. The proposed method can effectively identify high-risk individuals for preterm birth in early pregnancy, providing strong support for early intervention in early pregnanly.

Keywords: other disciplines of artificial intelligence; machine learning; integrated learning; neural network model; preterm birth prediction

近年来,机器学习模型在数据分析领域的应用不断深化,为疾病预测研究提供了新的技术支持[1-2]。早产是导致新生儿死亡和发病的主要原因之一[3],其中约有 2/3 属于自发性早产,且通常发生突然[4],这使得孕妇难以获得及时转诊和高质量的围产期服务。传统的早产预测方法成本高昂且耗时长,近年来,研究人员广泛运用机器学习算法进行早产预测。

通过深入分析临床早产数据,机器学习模型能够有效识别与早产相关的风险因素并进行早产预测。MORKEN 等[5]使用逻辑回归模型(logistic regression,LR)进行早产预测,将数据划分为初产妇和经产妇 2 组,结果显示 ROC 曲线下面积(area under the curve, AUC)分别为 0.74 和 0.58。AHADI 等[6]收集了 600 名 1～13 周孕妇的数据,运用 LR 和支持向量机(support vector machine, SVM)[7]模型进行预测,SVM 模型的准确率(Accuracy)为 0.67,高于 LR 模型的 0.56。PREMA 等[8]提出了基于机器学习的早产危险因素识别方法,采用线性 SVM、非线性 SVM 和 LR 作为预测模型。这 3 种模型的 Accuracy 较高,分别为 0.861 1、0.861 1 和 0.872 3,但召回率(Recall)和 $F1$ 值较低。RAKESH 等[9]研究了居住环境因素对早产的影响,并比较了特征选择前后 LR 模型和决策树(decision tree,DT)[10]模型的精度,结果表明 LR 模型的表现更好。RAJA 等[11]通过机器学习模型 LR、DT 和 SVM 进行预测,发现 SVM 模型的 Accuracy 为 0.909 0,但 Recall 相对较低。尽管上述研究取得了一定成果,但单一模型容易存在偏差,并且不同模型在相同数据集上的表现具有较大差异。

为减少单一模型的偏差,部分研究人员采用混合模型进行早产预测。吴忆娜[12]构建了基于门控循环单元(gate recurrent unit, GRU)和梯度提升决策树(gradient boosting decision tree,GBDT)[13]的混合模型进行早产预测。研究结果表明,该混合模型的预测能力优于单一的 GUR 模型和 GBDT 模型,评价指标 Recall 为 0.77, AUC 值为 0.647。然而,混合模型仍然存在对特定数据集的依赖,且在提高准确性和泛化能力方面仍有不足。

为进一步提升早产预测模型的泛化能力和综合性能,本文提出了一种基于 Stacking 模型[14-15]的早产预测方法。通过欠采样解决数据不平衡的问题,并综合运用特征重要性和特征相关性 2 种特征选择方法进行特征筛选;同时,利用皮尔逊相关系数优化 Stacking 模型中基分类器的类型和数量,充分挖掘各基分类器的优势,从而更准确地识别孕早期高风险早产人群。

1 预测方法概述

基于 Stacking 模型的早产预测方法分为 3 部分:数据预处理、特征选择以及 Stacking 模型构建,如图 1 所示。

首先,进行数据预处理,运用 KNN Imputer 方法[16]对数据缺失值进行填补,使用欠采样方法解决数据集中正、负样本不平衡的问题,并通过数据标准化消除变量间的数值差异,使数据集服从正态分布。

其次,进行特征选择,计算 11 个机器学习模型的特征重要性分数,分析特征的相关系数,识别并剔除冗余特征,从而有效降低过拟合风险。

最后,构建 3 层架构的 Stacking 集成学习模型。在第 1 层,通过分析 11 个机器学习模型的预测结果间的相关性,筛选基分类器。筛选出的基分类器包括 LR、SVM、GBDT、Adaboost[17]以及多层感知器(multi-layerperceptron,MLP)。在第 2 层,对筛选出的基分类器进行五折交叉验证和超参数优化,以提高分类器的性能和鲁棒性。在第 3 层,将优化后的基分类器预测结果作为输入,评估 11 个模型分别作为元分类器的预测效果,最终选择 MLP 作为 Stacking 模型的元分类器。

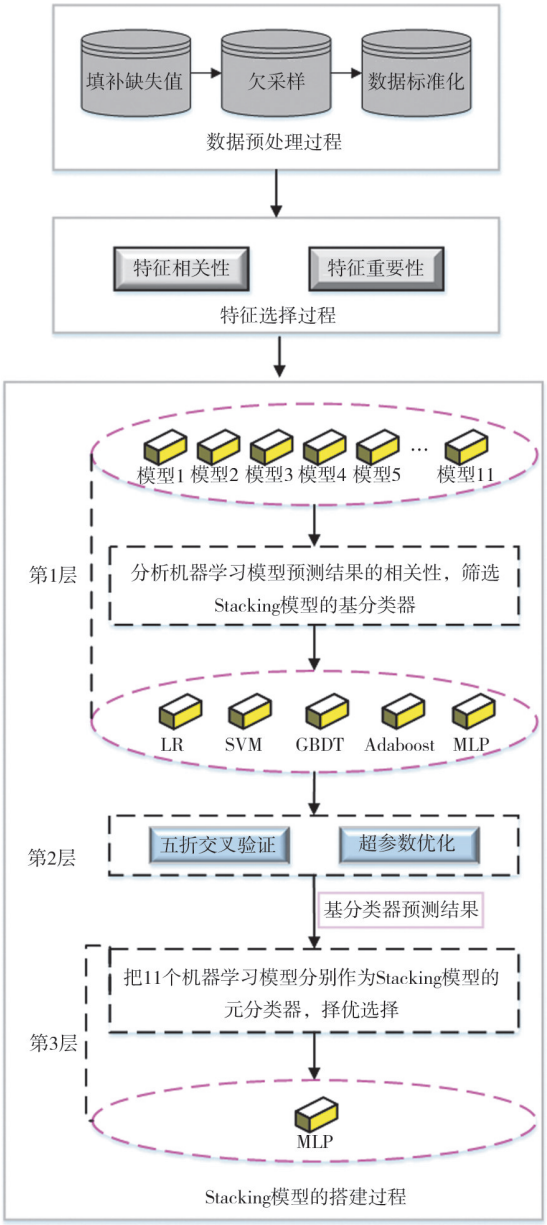


图 1　基于 Stacking 模型的早产预测方法框图

Fig. 1　Method block diagram of preterm birth prediction under Stacking model

2　研究方法

2.1　实验数据

数据集来自美国疾病控制预防中心(CDC)的综合数据库[18](http://wonder.cdc.gov)。该数据库包含 2021 年在美国各州登记的出生人口信息，覆盖了 99%以上的出生人口，共计 23 458 条记录，其中包括 2 692 条早产记录和 20 766 条非早产记录，每条记录包含 92 个特征。本文分析了 21 个特征，并以“是否早产”作为目标变量进行研究。正样本指发生早产的个体(目标变量为 1)，负样本指未发生早产的个体(目标变量为 0)。在研究数据中，正、负样本的比例为 8∶1。研究人群的特征如表 1 所示。

表 1　研究人群的特征

Tab. 1　Characteristics of the study population

序号	特　征	特征说明
1	Age	年龄
2	Edu	受教育程度
3	Tbo	婴儿总的出生数目
4	Illb	自上次活产以来的间隔
5	Ilop	自上次其他妊娠以来的间隔
6	Ilp	自上次妊娠以来的间隔
7	Cig0	怀孕前吸烟数
8	Cig1	第 1 个妊娠期吸烟数
9	Cig2	第 2 个妊娠期吸烟数
10	Cig3	第 3 个妊娠期吸烟数
11	Height	身高
12	BMI	体重指数
13	Wtgain	怀孕增加的体重
14	Pdiab	孕前糖尿病
15	Gdiab	妊娠期糖尿病
16	Phype	孕前高血压
17	Ghype	妊娠期高血压
18	Ehype	高血压子痫
19	Ppterm	既往早产
20	Inftr	使用不孕症治疗
21	Cesarn	剖腹产次数

2.2　数据预处理

数据预处理流程包括填补数据集中的缺失值、进行欠采样处理和实施数据标准化操作。

1)运用 KNN Imputer 方法对数据缺失值进行填补。首先，利用欧氏距离矩阵计算出数据集中与缺失值相近的 k 个样本；然后，计算这 k 个样本内非空值的平均值，并用该平均值来填补数据中的缺失值。

2)在监督分类任务里，机器学习模型往往会过于侧重目标变量的多数类，这就引发了数据不平衡问题，进而对模型性能产生负面影响[19]。针对实验数据中目标变量正、负样本比例约为 8∶1 的情况，本文采用的数据不平衡处理方法是欠采样(NearMiss)[20]，其原理是通过去除部分多数样本使数据集达到平衡。首先，计算正、负样本之间的配对距离；然后，根据计算的距离，删除距离负样本较远的正样本实例。经 NearMiss 处理后的数据集正、负样本比例接近于 1∶1，能够有效提升模型预测的准确性、稳定性。

3)由于数据中的各个特征取值范围的差异较大，取值范围较大的特征可能在模型训练过程中占据主导地位，从而对模型权重产生较大的影响，而取

值范围较小但同样重要的特征就可能被忽视。为了解决这一问题，需要对数据进行标准化处理，消除变量间的数值差异，使所有特征在模型中具有相同的权重。标准化的核心是将数据转换为均值为 0、标准差为 1 的正态分布，其转化公式为

$$z=\frac{\chi-\mu}{\sigma}, \tag{1}$$

式中：χ 是特征的原始值；μ 是该特征的均值；σ 是其标准差。标准化不仅可以平衡特征对模型的影响，还能提高模型的训练效率和预测性能。

完成数据标准化后，采用分层抽样将数据集按 8∶2 的比例划分为训练集和测试集，并确保正、负样本比例在两者中保持一致，即训练集和测试集中的正、负样本比均为 1∶1。

2.3 特征选择

特征选择旨在筛选出对机器学习模型预测性能至关重要的特征[21]，以提高训练效率并降低过拟合风险，从而增强模型的整体性能[22]。原始数据集包含 21 个特征。在特征选择过程中，综合考虑 11 个机器学习模型计算的平均特征重要性分数和特征相关性，剔除相关性高且平均特征重要性较低的冗余特征，从而优化模型训练过程并提升预测性能。

首先，对特征相关性进行分析。在相关性热图中，颜色越深表示特征间相关性越高，颜色越浅表示特征间相关性越低。图 2 直观展示了数据集中 21 个特征的两两相关性系数。由图可见，2 个特征组的相关性较高：一组为第 1、2、3 个妊娠期吸烟数(Cig1、Cig2、Cig3)；另一组为自上次妊娠间隔(Ilp)和自上次活产间隔(Illb)。

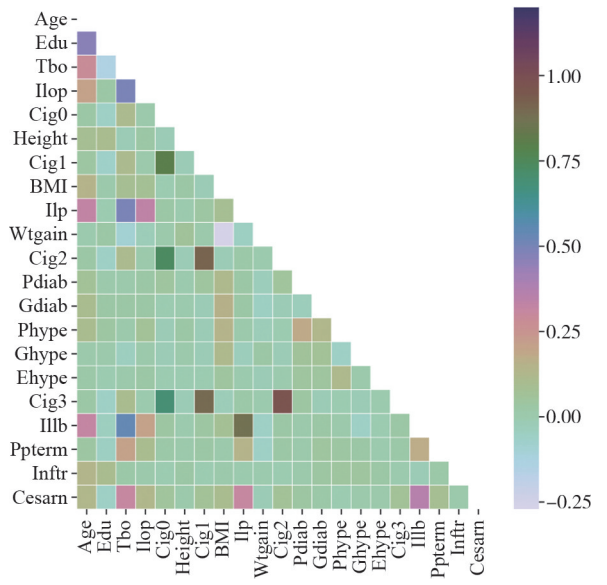


图 2 21 个特征的相关性热图

Fig. 2 Correlation heat map of 21 features

然后，通过 11 个机器学习模型分别计算 21 个特征的重要性分数，并对其取平均值，得到平均特征重要性，如图 3 所示。

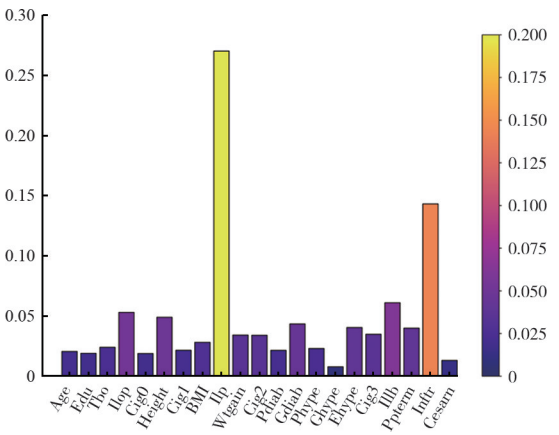


图 3 基于 11 个机器学习模型计算的平均特征重要性图

Fig. 3 Feature importance analysis based on 11 machine learning models

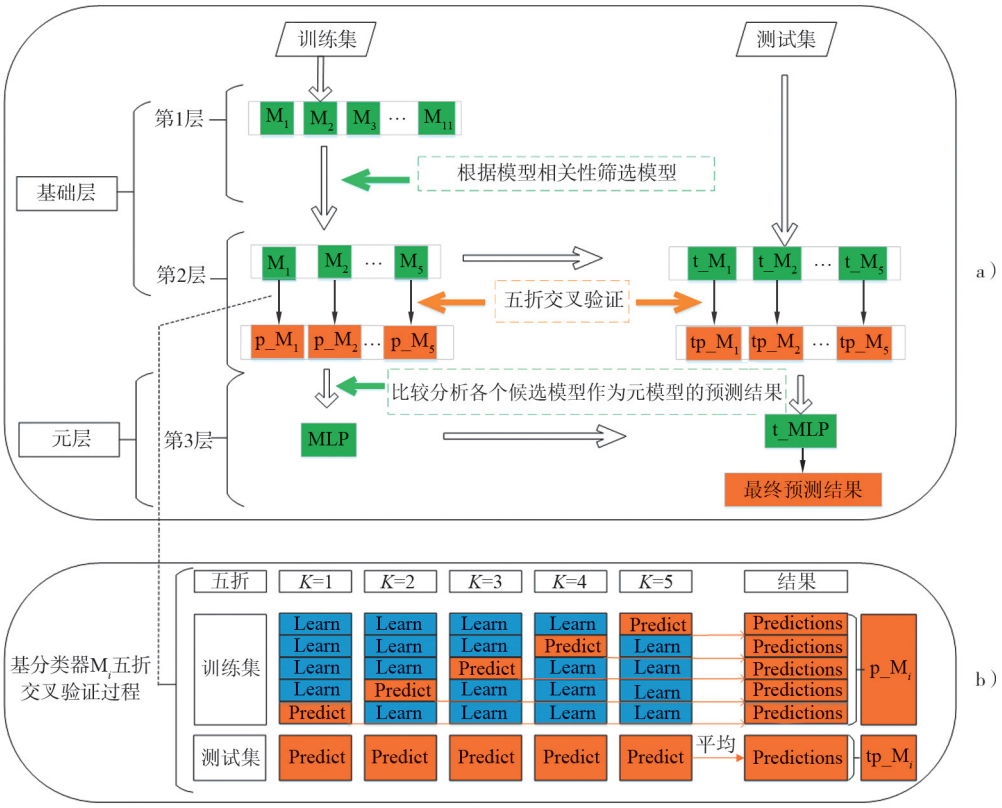
在特征选择过程中，过滤掉相关性较高且平均特征重要性较低的 2 个特征：Cig1 和 Illb。

2.4 Stacking 模型构建

所构建的 Stacking 模型分为 3 层，其中第 1 层和第 2 层是基础层，第 3 层是元层。在第 1 层设置了 11 个具有广泛代表性的机器学习模型，分别是 DT、SVM、LR、MLP、GBDT、LightGBM[23]、XGBoost[24]、CatBoost[25]、AdaBoost、随机森林(random forest，RF)[26]以及极端随机树分类器(extra trees classifier，ETC)[27]，根据机器学习模型预测结果间的皮尔逊相关系数分步筛选出 5 个基分类器。在第 2 层对筛选出的 5 个基分类器进行五折交叉验证以及超参数优化，以提升模型精度及泛化能力。在第 3 层将 11 个机器学习模型逐一设为 Stacking 模型的元分类器，统计分析 Stacking 模型的预测结果，选择效能最优的机器学习模型作为元分类器。Stacking 模型的分层集成框架如图 4 所示。

2.4.1 基分类器筛选

第 1 层是基分类器的筛选过程，当选择 11 个机器学习模型作为基分类器时，Stacking 模型的 AUC 值为 0.901 9。在图 4 a)中，M_1-M_{11} 表示 11 种机器学习模型。研究表明，Stacking 模型中基分类器的数量会影响模型整体性能[28]，且基分类器之间应保持独立，即预测结果间的皮尔逊相关系数要低。为此，对 11 种模型的预测结果进行了皮尔逊相关性分析，并逐步剔除相关系数最高的算法，最终确定 5 个基分类器。



注：a）Stacking模型框架；b）基分类器M_i的五折交叉验证过程。

图 4　Stacking 模型的分层集成框架

Fig. 4　Hierarchical ensemble framework of Stacking model

图 5 展示了 11 个机器学习模型预测结果的皮尔逊相关系数，其中 RF、XGBoost、LightGBM、CatBoost 以及 AdaBoost 之间存在较高的相关性。为降低 Stacking 模型基分类器的冗余性，进行第 1 次调整实验，评估相关性较高的 5 个机器学习模型对 Stacking 模型性能的影响。具体方法是进行 5 组实验，每组仅保留 RF、XGBoost、LightGBM、CatBoost 或 AdaBoost 之一，并与其余 6 个机器学习模型共同作为 Stacking 模型的基分类器进行训练，同时使用 11 个机器学习模型依次作为元分类器进行预测。第 1 次调整中 5 组实验的 AUC 结果对比如图 6 所示，其中 AdaBoost 组的 AUC 值最高（0.920 9），因此被保留。

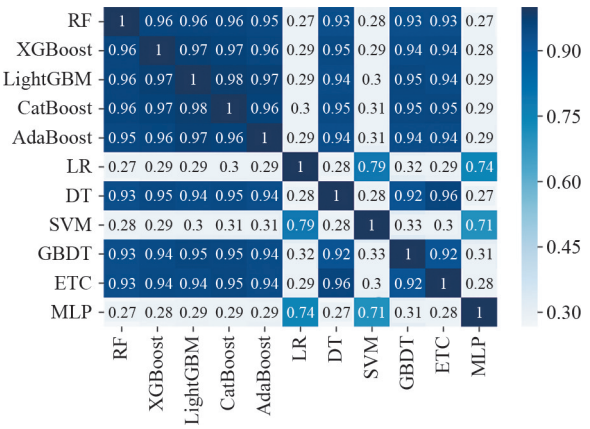


图 5　11 个机器学习模型的皮尔逊相关性分析图

Fig. 5　Pearson correlation analysis diagram of 11 machine learning model

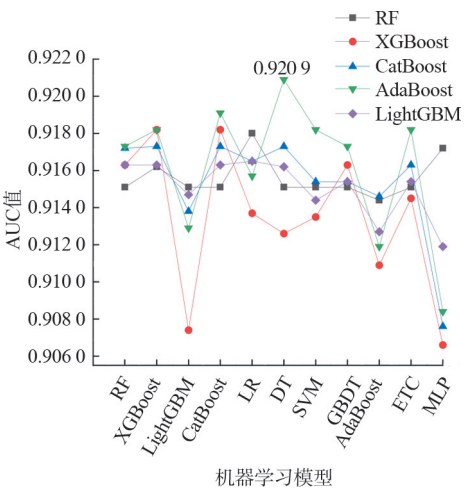


图 6　第 1 次调整中 5 组实验的 AUC 结果对比

Fig. 6　Comparison of AUC results for the five experimental groups in the first adjustment

在第 1 次调整后，DT、ETC 和 GBDT 仍存在且均为结构相似的树模型，为进一步降低 Stacking 模

型的冗余并提升基分类器的多样性，第 2 次调整实验旨在仅保留其中 1 个。具体方法是进行 3 组实验每组仅保留 DT、ETC 或 GBDT 之一，并与其余 4 个机器学习模型共同作为 Stacking 模型的基分类器进行训练，同时使用 11 个机器学习模型依次作为元分类器进行预测。第 2 次调整中 3 组实验的 AUC 结果对比如图 7 所示，GBDT 组的 AUC 值最高(0.921 9)，因此被保留。最终，筛选出 5 个相关性较低且性能优越的基分类器：LR、SVM、GBDT、MLP 和 AdaBoost。

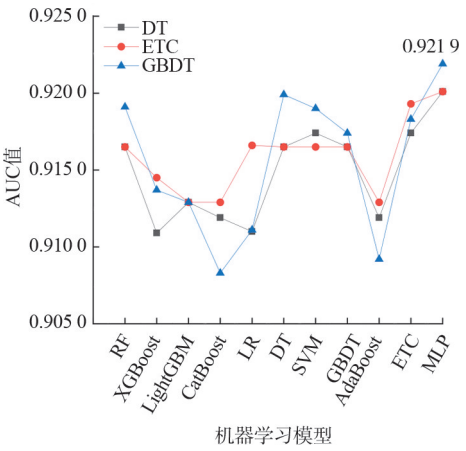


图 7 第 2 次调整中 3 个实验的 AUC 结果对比

Fig. 7 Comparison of the AUC values of the three experiments of validation in the second adjustment

2.4.2 五折交叉验证

在 Stacking 模型的第 2 层，采用五折交叉验证对基分类器进行训练，并通过超参数优化提升模型性能。在训练过程中，使用网格搜索调整各基分类器的关键参数，例如，SVM 可调节惩罚参数 C 和核函数参数 gamma。网格搜索在计算资源允许的范围内寻找最优超参数组合，以提升模型性能和泛化能力。各基分类器经过调整后的关键参数如表 2 所示。在进行超参数优化前，需要先将原数据集划分为训练集和测试集，其中训练集用于参数调优，测试集用于性能评估。为避免固定划分导致评估不稳定，采用交叉验证以提高评估的可靠性和稳健性。

表 2 Stacking 模型中各基分类器经过调整后的关键参数

Tab. 2 Optimized key parameters of base classifiers in the Stacking model

基分类器	关键参数
LR	penalty="l2", solver='newton-cg', C=0.6, max_iter=15, class_weight=None
SVM	C=1, kernel='rbf', gamma=3, gamma='auto', coef0=0.0
GBDT	learning_rate=0.05, n_estimators=418, max_depth=5, min_samples_split = 100, min_samples_leaf=50
MLP	solver='lbfgs', activation='logistic', hidden_layer_sizes=(100,100,10), random_state=20
AdaBoost	base_estimator=None, n_estimators=876, learning_rate=1.0, algorithm='SAMME.R'

在图 4 a)中，M_1—M_5 和 t_M_1—t_M_5 分别指代基分类器被训练集训练前后的状态，训练集对应的训练结果为 p_M_1—p_M_5。经五折交叉验证后，把测试集输入经过训练后的基分类器得出预测结果(tp_M_1—tp_M_5)。五折交叉验证通过多次随机划分数据并取平均值，有效降低模型评估的随机性，能够更全面地评估模型的性能。此外，在五折交叉验证过程中，Stacking 模型在不同数据集上反复训练和评估，能够全面验证其泛化能力。以基分类器 M_i 为例，其五折交叉验证过程如图 4 b)所示。如图所示，基分类器的预测组呈交错分布，经过交叉验证后可获得整个训练集的预测结果(p_M_i)。并且，在基分类器的每一轮迭代中，将五折交叉验证所产生的 5 组预测值进行算术平均运算，从而得出测试集的预测结果(tp_M_i)。

2.4.3 元分类器筛选

第 3 层是筛选 Stacking 模型元分类器的过程。首先，将第 2 层中基分类器的预测结果作为新数据集；其次，依次使用 11 个机器学习模型作为元分类器进行训练；然后，通过多种评估指标计算每个模型的性能分数，如表 3 所示；最后，选择在各评价指标中表现最好的 MLP 作为最终的元分类器。

表 3 11 个候选模型作为元分类器的预测结果

Tab. 3 Predictions of 11 candidate models as meta-classifiers

模 型	AUC 值	Accuracy	Recall	*F*1 值
RF	0.919 1	0.920 1	0.850 9	0.913 0
XGBoost	0.913 7	0.914 6	0.858 5	0.908 2
LightGBM	0.913 0	0.913 6	0.871 7	0.908 6
Catboost	0.908 3	0.909 0	0.866 0	0.903 5
LR	0.911 1	0.911 8	0.867 9	0.906 4
DT	0.920 0	0.921 1	0.850 9	0.913 9
SVM	0.919 1	0.920 1	0.850 9	0.913 0
GBDT	0.917 4	0.918 3	0.862 3	0.912 2
Adaboost	0.909 2	0.909 9	0.867 9	0.904 6
ET	0.918 3	0.919 2	0.860 4	0.912 9
MLP	**0.921 9**	**0.922 9**	**0.858 5**	**0.916 4**

综上所述，Stacking 集成学习模型的基分类器由 5 种异构模型组成：LR、SVM、GBDT、AdaBoost 和 MLP，元分类器择优选用 MLP。

3　实验与结果分析

3.1　实验环境

实验在笔记本电脑上进行，该电脑预装 Windows 11 操作系统。实验过程中，使用 Python 语言进行编程，并借助 Scikit-learn 框架完成预测模型的训练与评估。Scikit-learn 框架提供丰富的机器学习算法和工具，支持数据预处理、特征工程、模型训练与优化及评估。笔记本电脑硬件配置中：CPU 为 Intel Core i5 1.80 GHz(英特尔公司提供)；内存为 8 GB RAM(三星电子(中国)公司提供)和 500 GB SSD(金士顿科技公司提供)。软件环境包括 Anaconda 2023.09、Python 3.9.13 和 Jupyter Notebook 6.5.4。

3.2　评价指标

衡量方法性能的指标包括 Accuracy、AUC 值、Recall、$F1$ 值[29]，这些常用的分类器评估指标可以全面地反映模型表现。

在 ROC 曲线中，横坐标为假阳性率(FPR)，纵坐标是真阳性率(TPR)。通常，ROC 曲线处于 $y=x$ 之上，并且越接近于左上角，表示模型灵敏度越高，误判率越低，整体性能越优。AUC 值用于量化 ROC 曲线下面积，以评估模型的判别能力。当 AUC 值为 1 时，表明该模型在早产预测中的表现极佳；而 AUC 值为 0.5，则意味着模型无法有效区分阳性和阴性样本，即不具备预测能力。

3.3　结果分析

3.3.1　特征选择前后的预测结果分析

特征选择过程中筛选掉的冗余特征为 Cig1 和 Illb。在使用 AUC 值、Accuracy、Recall 以及 $F1$ 值作为评价指标的情况下，对 11 个机器学习模型在特征选择前后的预测结果进行了对比分析，结果如表 4 所示。其中，CatBoost 模型在特征选择前后，4 个评价指标均未出现显著差异。而 SVM 模型的 AUC 值、Accuracy 和 $F1$ 值以及 AdaBoost 模型的 Recall 值虽然有所下降，但降幅在 0.005 7 以内，差异无统计学意义。除了上述评价指标外，其他模型的各项评价指标均呈上升趋势，表明所采用的特征选择方法能够有效提升模型的泛化能力。

表 4　11 个单一机器学习模型进行特征选择前后的结果对比

Tab.4　Comparison of the results of 11 single machine learning models before and after feature selection

模　型	AUC 值		Accuracy		Recall		$F1$ 值	
	之前	之后	之前	之后	之前	之后	之前	之后
RF	0.889 2	0.901 9	0.890 4	0.903 4	0.815 0	0.815 6	0.879 8	0.891 4
LR	0.644 5	0.646 6	0.646 2	0.648 0	0.537 7	0.554 7	0.599 3	0.608 0
DT	0.875 5	0.901 9	0.878 3	0.903 4	0.754 7	0.807 5	0.858 3	0.891 6
SVM	0.654 9	0.649 4	0.657 3	0.651 8	0.498 1	0.500 0	0.588 6	0.585 6
GBDT	0.891 7	0.899 3	0.893 2	0.900 6	0.800 0	0.818 8	0.880 5	0.890 2
ETC	0.866 6	0.874 1	0.868 1	0.875 5	0.769 8	0.784 9	0.851 7	0.861 2
MLP	0.643 8	0.663 3	0.645 3	0.664 8	0.549 0	0.567 9	0.603 7	0.625 1
XGBoost	0.896 4	0.901 3	0.897 8	0.902 5	0.803 7	0.826 4	0.885 6	0.892 9
LightGBM	0.894 7	0.899 4	0.891 3	0.900 6	0.791 3	0.824 5	0.882 6	0.890 9
Adaboost	0.899 3	0.901 9	0.900 6	0.903 4	0.818 8	0.813 2	0.890 2	0.891 6
Catboost	0.901 1	0.901 1	0.902 5	0.902 5	0.819 6	0.816 9	0.891 8	0.891 8

3.3.2　基于 Stacking 模型的早产预测方法的结果分析

11 个单一机器学习模型在不同评价指标上各具优势，而所提出的 Stacking 集成学习模型在 Accuracy、Recall、$F1$ 值和 AUC 值等多项指标上均表现优异，具体结果见表 5。相比单一模型在各评价指标上的最佳表现，Stacking 模型进一步提升了预测性能，其中：AUC 值从 0.901 9 提升至 0.921 9，提升约 2.22%；Accuracy 从 0.903 4 提升至 0.922 9，提升约 2.16%；Recall 从 0.826 4 提升至 0.858 5，提升约 3.88%；$F1$ 值从 0.892 9 提升至 0.916 4，提升约 2.63%。此外，Stacking 模型的 ROC 曲线如图 8 所示。结果表明，所提出的模型能够有效融合单一机器学习模型的优势，显著提升整体性能。

表 5 Stacking 模型与 11 个单一模型的性能对比

Tab. 5 Performance comparison between Stacking model and 11 single models

模 型	AUC 值	Accuracy	Recall	*F*1 值
RF	0.901 9	0.903 4	0.815 6	0.891 4
LR	0.646 6	0.648 0	0.554 7	0.608 0
DT	0.901 9	0.903 4	0.807 5	0.891 6
SVM	0.649 4	0.651 8	0.500 0	0.585 6
GBDT	0.899 3	0.900 6	0.818 8	0.890 2
ETC	0.874 1	0.875 5	0.784 9	0.861 2
MLP	0.663 3	0.664 8	0.567 9	0.625 1
XGBoost	0.901 3	0.902 5	0.826 4	0.892 9
LightGBM	0.899 4	0.900 6	0.824 5	0.890 9
Adaboost	0.901 9	0.903 4	0.813 2	0.891 6
Catboost	0.901 1	0.902 5	0.816 9	0.891 8
Stacking	**0.921 9**	**0.922 9**	**0.858 5**	**0.916 4**

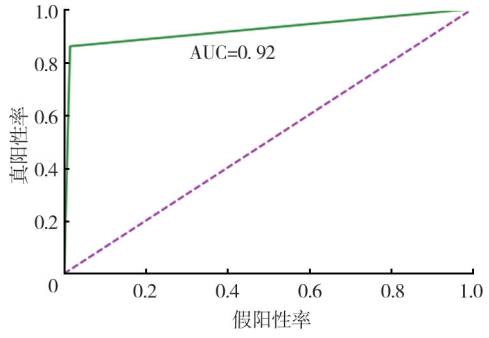


图 8 Stacking 模型的 ROC 曲线图

Fig. 8 ROC curve of the Stacking model

3.4 与现有早产预测方法的对比分析

本文提出的基于 Stacking 集成学习模型的早产预测方法通过计算各机器学习模型预测结果间的皮尔逊相关系数，优化基分类器的类型和数量，以提升预测性能。为验证该方法的优越性，将其与现有的早产预测方法进行对比，具体结果如表 6 所示。

PREMA 等[8]提出的基于机器学习的早产危险因素识别方法采用线性 SVM、非线性 SVM 以及 LR 作为预测模型，并运用合成少数类过采样技术(SMOTE)处理数据的不平衡问题。其中，在未进行数据平衡处理的模型分别标记为线性 SVM_1、非线性 SVM_1 和 LR_1，而经过 SMOTE 处理后的模型分别标记为线性 SVM_2、非线性 SVM_2 和 LR_2。在未进行数据平衡处理时，线性 SVM_1 与非线性 SVM_1 模型的 *F*1 值为 0，LR_1 的 Recall 和 *F*1 值较低，分别为 0.214 3 和 0.352 2。运用 SMOTE 平衡数据集后，虽然线性 SVM_2 与非线性 SVM_2 模型的评价指标 Accuracy 和 Recall 略有下降，但 *F*1 值显著提升，超过 0.74；LR_2 的 Accuracy 略微下降，但 Recall 和 *F*1 值显著上升，均超过 0.70。这表明数据平衡处理有效提升了模型的性能。RAJA 等[11]提出的基于机器学习的早产预测方法采用 DT、LR 和 SVM 作为预测模型。实验结果表明，SVM 分类器的 Accuracy 达到 0.909 0，相较于 DT 和 LR 模型表现更为优越。

表 6 本文所提方法与现有早产预测方法的对比

Tab. 6 Comparison between the proposed method in the paper and the existing preterm birth prediction methods

方 法	AUC 值	Accuracy	Recall	*F*1 值
PREMA(LR_1)		0.872 3	0.214 3	0.352 2
PREMA(线性 SVM_1)		0.861 1	0.861 1	0.000 0
PREMA(非线性 SVM_1)		0.861 1	0.861 1	0.000 0
PREMA(LR_2)		0.754 3	0.703 7	0.777 2
PREMA(线性 SVM_2)		0.763 4	0.838 7	0.803 6
PREMA(非线性 SVM_2)		0.731 2	0.704 5	0.747 5
RAJA(DT)		0.796 0	0.713 0	
RAJA(LR)		0.872 0	0.832 0	
RAJA(SVM)		0.909 0	0.891 0	
本文方法	**0.921 9**	**0.922 9**	**0.858 5**	**0.916 4**

在表 6 中可以看出，与以上研究相比，本文所提出的基于 Stacking 集成学习模型的早产预测方法在评价指标上表现更全面，而且此前 2 项研究缺少 AUC 值这一关键衡量指标。尽管基于 Stacking 集成学习模型的早产预测方法在 Recall 值上较低，但其 Accuracy 和 *F*1 值却有显著提高。*F*1 值作为精确率与召回率的调和平均数，有效平衡了 Stacking 集成学习的模型精确性与完整性，进一步证明了该方法在优化整体预测性能方面的优势。综上所述，该方法在多个评价指标上均表现优异，显著提高了早产预测模型的综合性能。

4 结 语

针对传统机器学习模型在早产预测中的局限性，本文提出了一种基于 Stacking 模型的预测方法，通过集成多个基分类器的预测结果，并利用元分类器对其进行学习和优化，提升了早产预测的整体性能。主要研究结论如下。

1) 在特征选择前后，11 个机器学习模型中，AdaBoost、CatBoost 和 SVM 模型预测结果的变化不显著，而其他模型预测结果的评价指标普遍提升，

表明特征选择有效增强了大部分机器学习模型的性能和泛化能力，从而提高了预测的准确性和稳定性。

2）相比构建时所使用的单一模型，Stacking 模型在各评价指标上的提升幅度均超过 2%。具体而言，AUC 值从 0.901 9 提升至 0.921 9，提升约 2.22%；Accuracy 从 0.903 4 提升至 0.922 9，提升约 2.16%；Recall 从 0.826 4 提升至 0.858 5，提升约 3.88%；$F1$ 值从 0.892 9 提升至 0.916 4，提升约 2.63%。Stacking 模型显著提升了早产预测的整体性能。

3）与现有研究相比，Stacking 模型的 AUC 值为 0.921 9，展示了较强的分类能力；同时，Accuracy 为 0.922 9，优于 RAJA[11] 所提早产预测法，进一步提高了预测准确性。尽管 Recall 较低，但 $F1$ 值达到了 0.916 4，显著优于以往研究方法，证明了该方法在平衡精度与 Recall 方面的优势。所提方法较现有研究在整体性能上有了显著提升，在分类能力和整体性能上具有显著优势。

尽管所提出的方法在早产预测中表现出较高的性能，但目前仅用于预测是否早产，尚未考虑对早产风险进行分层预测。未来研究将结合统计分析方法，采用四分位间距法确定 Stacking 模型预测概率的截断值，将早产风险划分为 4 个层次：低风险、中风险、中高风险和高风险，从而实现风险分层预测。这一改进将进一步提高预测的精准度和可靠性，为早产防控措施的制定提供更为科学的依据。

参考文献/References：

[1] HOFFMAN M. Prediction and prevention of spontaneous preterm birth: ACOG practice bulletin, number 234[J]. Obstetrics and Gynecology, 2021, 138(6): 945-946.

[2] KOTELUK O, WARTECKI A, MAZUREK S, et al. How do machines learn? Artificial intelligence as a new era in medicine [J]. Journal of Personalized Medicine, 2021, 11(1): 32.

[3] LIU Li, OZA S, HOGAN D, et al. Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: An updated systematic analysis[J]. The Lancet, 2015, 385(9966): 430-440.

[4] MURRAY C J L, VOS T, LOZANO R, et al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010 [J]. The Lancet, 2012, 380 (9859): 2197-2223.

[5] MORKEN N H, KÄLLEN K, JACOBSSON B. Predicting risk of spontaneous preterm delivery in women with a singleton pregnancy[J]. Paediatric and Perinatal Epidemiology, 2014, 28 (1): 11-22.

[6] AHADI B, MAJD H, KHODAKARIM S, et al. Using support vector machines in predicting and classifying factors affecting preterm delivery[J]. Paramedical Sciences, 2016, 7(3): 37-42.

[7] DANENAS P, GARSVA G. Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach[J]. Procedia Computer Science, 2012, 9: 1324-1333.

[8] PREMA N S, PUSHPALATHA M P. Machine learning approach for preterm birth prediction based on maternal chronic conditions[C]//Emerging Research in Electronics, Computer Science and Technology. Singapore: Springer, 2019: 581-588.

[9] RAKESH R, INDRAJIT M, KANTI S B. A systematic review of healthcare big data[J]. Scientific Programming, 2020(1): 5471849.

[10] SONG Yanyan, LU Ying. Decision tree methods: Applications for classification and prediction[J]. Shanghai Archives of Psychiatry, 2015, 27(2): 130-135.

[11] RAJA R, MUKHERJEE I, SARKAR B K. A machine learning-based prediction model for preterm birth in rural India[J]. Journal of Healthcare Engineering, 2021, 2021: 6665573.

[12] 吴忆娜. 基于特征融合和深度学习的孕妇分娩时间预测模型的研究[D]. 杭州：杭州师范大学，2021.
WU Yina. Study on Prediction of Maternal Delivery Time Based on Feature Fusion and Deep Learning[D]. Hangzhou: Hangzhou Normal University, 2021.

[13] RÄTSCH G, ONODA T, MÜLLER K R. Softmargins for AdaBoost[J]. Machine Learning, 2001, 42(3): 287-320.

[14] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[15] 王鹏，曹丽惠，阮冬茹. 基于 Stacking 模型融合的店铺销量预测[J]. 河北工业科技，2022，39(3)：204-209.
WANG Peng, CAO Lihui, RUAN Dongru. Store sales forecast based on Stacking model fusion[J]. Hebei Journal of Industrial Scienceand Technology, 2022, 39(3): 204-209.

[16] JUNA A, UMER M, SADIQ S, et al. Water quality prediction using KNN imputer and multilayer perceptron [J]. Water, 2022, 14(17): 2592.

[17] YE J, CHOW J H, CHEN Jiang, et al. Stochastic gradient boosted distributed decision trees[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong: ACM, 2009: 2061-2064.

[18] OSTERMAN M J K, HAMILTON B E, MARTIN J A, et al. Births: Final data for 2021[J]. National Vital Statistics System, 2023, 72 (1): 1-53.

[19] LI D C, LIU C W, HU S C. A learning method for the class imbalance problem with medical data sets[J]. Computers in Biology and Medicine, 2010, 40(5): 509-518.

[20] LAURIKKALA J. Improving identification of difficult small classes by balancing class distribution[C]//Artificial Intelligence in Medicine. Berlin Heidelberg: Springer, 2001: 63-66.

（下转第 181 页）