南锐, 刘燕丽, 万祁阳, 等. 基于机器学习的糖尿病预测和因素分析[J]. 智能计算机与应用,2025,15(6):140-145. DOI: 10.20169/j.issn.2095-2163.250621

基于机器学习的糖尿病预测和因素分析

南 锐, 刘燕丽, 万祁阳, 涂博文
(武汉科技大学 理学院, 武汉 430065)

摘 要: 糖尿病是一种常见的代谢性疾病,其并发症会显著增加患者的死亡风险。因此,及早诊断和有效的预防对于降低患病风险至关重要。本文基于国家健康和营养调查平台的受试者记录,采用多种数据处理方法,以探究个人基本信息对糖尿病的预测效果。进一步研究了环境化学物质因素是否对糖尿病产生潜在的影响,并分析了特征的重要性。逻辑回归、随机森林和 XGBoost 三种模型的预测结果表明,引入环境特征有效地提高了模型的预测性能。与其它机器学习模型相比,XGBoost 的准确率、召回率和 *AUC* 值均提高了约 2%左右。特征的重要性分析显示年龄、身体健康指数等因素在糖尿病预测中具有重要的意义,为糖尿病的预防提供了参考依据。

关键词: 糖尿病; 机器学习; 环境化学物质; 因素分析
中图分类号: R587.1;TP181 **文献标志码**: A **文章编号**: 2095-2163(2025)06-0140-06

Machine learning based diabetes prediction and factor analysis

NAN Rui , LIU Yanli, WAN Qiyang, TU Bowen
(College of Science, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: Diabetes is a common metabolic disorder, and its complications significantly increase the risk of mortality among patients. Therefore, early diagnosis and effective prevention are of paramount importance in reducing the risk of developing the disease. This study is based on subject records from the National Health and Nutrition Examination Survey platform and employs various data processing methods to explore the predictive power of individual demographic information in relation to diabetes. Furthermore, the paper investigates whether environmental chemical factors may have a potential impact on diabetes and analyze the importance of various features. The predictive results from three models, namely Logistic Regression, Random Forest, and XGBoost, demonstrate that the incorporating environmental features effectively enhances the predictive performance of the models. Compared to other machine learning models, XGBoost exhibits approximately 2% increase in accuracy, recall, and AUC values. Feature importance analysis reveals that factors such as age and body mass index play a crucial role in diabetes prediction, providing valuable insights for diabetes prevention.

Key words: diabetes; machine learning; environmental chemical exposure; factor analysis

0 引 言

糖尿病是一种以血糖升高为特征的慢性系统性代谢疾病,具有病程长、致残率高、难以治愈的特点。根据国际糖尿病联盟统计的调查结果,在 2019 年全球共有 4.63 亿位糖尿病患者,预计到 2045 年,全世界将有 7 亿糖尿病患者,患病率的急剧攀升使得糖尿病成为威胁人们身体健康的重大问题[1-2]。现阶段,糖尿病的主要治疗方法是药物治疗和控制饮食,这也给患者造成了一定的经济压力和生活上的不便[3-5]。早预防早发现可以有效减少糖尿病引发的致病率和死亡率。因此,分析糖尿病患者的特征和影响因素具有重要的意义。

随着医疗数据规模的快速增长和机器学习技术的迅猛发展,机器学习方法应用于疾病的分析和诊断、为医生提供辅助诊断、提升人们对疾病的认识已

基金项目: 湖北省暨武汉市工业与应用数学学会开放基金(2022003); 大学生创新创业项目(202310488026X)。
作者简介: 南 锐(1999—),女,硕士研究生,主要研究方向:机器学习; 万祁阳(2002—),男,本科生,主要研究方向:组合优化; 涂博文(2004—),男,本科生,主要研究方向:组合优化。
通信作者: 刘燕丽(1980—),女,博士,副教授,主要研究方向:组合优化,机器学习。Email:yanlil2008@wust.edu.cn。
收稿日期: 2023-10-11

是智能医疗的一个重要研究方向。从单一预测模型到集成模型，机器学习应用于糖尿病的早期预测已取得可观的进步与成果。车前子等学者[6]对参与北京市房山区的 3 153 名居民进行分析，根据特征重要度排名筛选输入变量，构建了基于人工神经网络算法的 2 型糖尿病发病风险预测模型，准确率达到 94.0%。惠亚楠等学者[7]提出一种基于改进狮群算法优化神经网络的糖尿病风险预测模型，在狮群算法中引入非线性扰动因子，使得模型具有更好的预测性能。张春富等学者[8]提出基于遗传算法寻找最优模型参数的 GA-XGBoost 模型。该模型在天池竞赛平台提供的糖尿病临床数据中的预测性能优于其他模型。孙彤等学者[9]对多种基分类器组合利用加权求和的方法进行评估，提出了一种基于层次分析法的 Stacking 模型，提升了分类性能。

虽然机器学习算法已经在智能医疗上取得了不少的成果，但是仍然存在很多亟待解决的难点。比如，大多数机器学习算法的可解释性不强，同一模型在不同的数据集上的预测能力可能会有较大差异。数据规模大、高维度、特征繁杂等因素也直接影响模型的预测能力。本文针对国家健康与营养调查（National Health and Nutrition Examination Survey，NHANES）[10]的数据，采用逻辑回归、随机森林和 XGBoost 建立模型，分别研究了基于受试者基本信息的糖尿病的分类模型，以及添加环境化学物质暴露因素的预测模型。最后，通过分析特征重要性得出糖尿病的重要影响因素。

1　数据来源和预处理

国家健康和营养调查（NHANES）提供了受试者的基本信息和环境化学物质等 16 个数据文件[11]。每位受试者都拥有唯一的 ID，依据 ID 值聚合受试者的基本信息和环境信息，形成数据集。数据集去除信息缺失率大于 50%的记录，剩余 878 名受试者记录，其中包括 328 名糖尿病患者和 550 名未患糖尿病者。表 1 是特征集的描述。患者个人信息包括 6 个临床特征、6 个生活方式特征；环境化学物质类包括多环芳烃（PAHs）[12]、全氟烷基物质（PFAS）[13]、邻苯二甲酸酯（PAEs）[14]及金属与非金属元素[15]，共计 15 个特征。

1.1　数据预处理

数据预处理主要完成特征值替换、异常值处理、缺失值处理以及数据归一化，为模型训练提供较高质量的数据。

表 1　数据集的特征描述

Table 1　The feature description of the data set

特征类别	特征	特征个数
基本信息（12 个特征）	临床特征	6
	生活方式	6
环境化学物质（15 个特征）	多环芳烃（PAHs）	4
	全氟烷基物（PFAS）	4
	邻苯二甲酸酯（PAEs）	4
	金属与非金属元素	3

1.1.1　特征值替换

美国疾病控制与预防中心发布的关于人类接触环境化学品的国家报告提出，如果环境化学物质的特征值低于该特征的检测最小阈值（Limit of Detection，*LOD*），那么这些特征值需替换为 $LOD/\sqrt{2}$[16]。数据集中非金属元素砷存在低于其 *LOD* 的值。表 2 展示了砷的 *LOD* 以及特征值替换前后的结果。

表 2　砷的部分特征值替换前后的结果

Table 2　Results before and after replacement of partial values of AS

检测最小阈值	记录 ID	替换前	替换后
0.74	47	0.29	0.52
0.74	103	0.44	0.52
0.74	281	0.35	0.52

1.1.2　异常值处理

本文结合文献和箱线图，对特征进行了异常值检查。具体地，计算每个特征的第一分位数 Q_1、第三分位数 Q_3 以及四分位数极差 *IQR*，即小于 $Q_1 - 1.5IQR$ 或大于 $Q_3 + 1.5IQR$ 的值为异常值。检测结果表明年龄特征中 85 岁是异常值，但是联系资料和实际情况，该特征值仍然作为合法值处理。

1.1.3　缺失值的处理

部分特征存在数据缺失，但缺失比例均小于 30%。表 3 展示了这些特征的缺失率。

表 3　部分特征的缺失率

Table 3　The missing rate of partial features　%

收入水平	身体质量指数	摄入糖分	铅
1.21	0.48	2.29	15.41

为了避免数据缺失导致的模型性能下降，采用决策树模型预测的方法填充缺失值。具体地，将缺失值作为预测的目标变量，在不含缺失值的数据上训练决策树模型，利用训练好的模型对缺失值进行预测和填充。

1.1.4　数据的归一化

特征之间的量纲差异大可能导致回归模型出现

较大的偏差,从而降低模型的训练效果。因此,为了确保逻辑回归(Logistic Regression,LR)模型的参数训练和预测能力,采用 Min-Max 归一化方法对连续型特征进行处理:

$$X_{nor}=\frac{X-X_{min}}{X_{max}-X_{min}} \tag{1}$$

其中,X_{nor} 表示归一化后的变量;X 表示原始变量;X_{min} 表示 X 中的最小值;X_{max} 表示 X 中的最大值。

1.2 特征选择

特征选择能够去除嘈杂、冗余或不相关的特征,达到降低训练难度、提高模型的准确度和泛化能力的目标,这是一种有效的降维策略。针对基本信息和环境化学物质两类特征,采用了以下 2 种特征选择方法。

(1)最大信息系数(Maximal Information Coefficient,*MIC*):*MIC* 是适用于连续型和离散型数据、鲁棒性强的相关性度量,用于衡量变量之间线性或非线性关联的强度,计算公式为:

$$MIC(X,Y)=\max_{g\in\Gamma(X,Y)}\frac{I(g(X),g(Y))}{\log(\min(|X|,|Y|))} \tag{2}$$

其中,$MIC(X,Y)$ 表示变量 X 和 Y 之间的最大信息系数;$I(g(X),g(Y))$ 表示在给定的离散化函数 g 下,X 和 Y 之间的互信息;$\Gamma(X,Y)$ 表示所有可能的离散化函数集合;$|X|$ 和 $|Y|$ 分别表示变量 X 和 Y 的不同取值的数量。

$I(g(X),g(Y))$ 的计算公式为:

$$I(g(X),g(Y))=\sum_{X,Y}P(g(X),g(Y))\log\frac{P(g(X),g(Y))}{P(g(X))P(g(Y))} \tag{3}$$

其中,$P(g(X),g(Y))$ 表示在给定的离散化函数 g 下变量 X 和 Y 之间的联合概率;$P(g(X))$ 和 $P(g(Y))$ 分别表示在函数 g 下变量 X、Y 的边际分布概率。

利用 *MIC* 对 12 个基本信息的特征进行选择,剔除 *MIC* 值小于 0.02 的 3 个特征,保留其余的 9 个特征。剔除的特征及其对应的 *MIC* 值见表 4。

表 4 剔除的特征及其 *MIC* 值

Table 4 Eliminated features and their *MIC* values

特征	*MIC*
性别	0.002 4
吸烟状况	0.002 1
种族	0.001 9

(2)基于决策树模型选择特征:决策树的生成过程是连续选择具有更确定信息的特征,以确定分类规则的过程。那些具有更大信息增益(信息增益比)的特征会被优先选择为决策树的分支点。因此,位于树根附近的特征对目标变量的影响更显著。

本文对环境化学物质特征建立了决策树模型,并选择了树结构中前 10 个出现的化学物质特征加入训练集,包括铅、全氟己烷磺酸、对羟基苯甲酸乙酯等。

2 模型构建

在文献[17-18]中,通常将回归模型作为糖尿病预测模型的对比基准。因此,本文采用逻辑回归、随机森林和 XGBoost 这 3 类具有不同工作原理的模型,对是否患有糖尿病进行预测,并对预测结果进行了比较。

2.1 逻辑回归

逻辑回归的本质是一种基于线性关系和概率的分类算法,具有简单、解释性强、计算效率高等优点。LR 通过最大似然估计来学习模型参数,建立一个线性模型,将输入特征的线性组合映射到概率值,最后利用阈值来进行分类预测。

LR 引入 Sigmoid 函数来处理二分类问题,计算公式为:

$$P(x)=\frac{1}{1+e^{-\theta^{T}\cdot x}} \tag{4}$$

其中,$x\in R^{d}$ 表示 d 维特征向量;$P(x)$ 表示样本 x 属于类别 1 的概率;θ 表示回归系数向量。

2.2 随机森林

随机森林(Random Forest,RF)是基于 Bagging 集成思想,将多棵决策树组合在一起的集成学习算法。RF 采用有放回的随机抽样方法,从训练样本集中抽取多个不同的样本子集,每个子集都可用于训练一棵独立的决策树。在训练每棵决策树时,RF 从特征集中选择最佳分裂特征。最后根据多数投票原则,集成每棵树的分类结果,以确定样本的最终类别。组合 N 棵决策树的随机森林如图 1 所示。

2.3 极端梯度提升

极端梯度提升(eXtreme Gradient Boosting,XGBoost)是一种基于 Boosting 的高效的梯度提升决策树算法。与 RF 类似,XGBoost 也是多棵决策树的集成模型。但是,与 RF 的区别则在于训练过程中的第 t 棵树的预测目标是因变量的实际值与前 $t-1$ 棵树的预测值之差。XGBoost 通过不断对残差的拟

合，以达到提升整个集成模型预测能力的目标。XGBoost 的目标函数为：

$$obj^{t} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} W(f_i) \tag{5}$$

其中，$l(y_i, \hat{y}_i)$ 表示损失函数；y_i 表示第 i 个样本的真实值；$\hat{y}_i$ 表示第 i 个样本的预测值；n 表示样本总数；t 表示迭代次数；$W(f_i)$ 为第 i 棵树 f_i 的正则化项，表示模型的复杂程度，$W(f_i)$ 的计算公式为：

$$W(f_i) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

其中，γ 表示收缩系数；T 表示叶子节点数；λ 表示正则化系数；w_j 表示叶子节点 j 的输出。

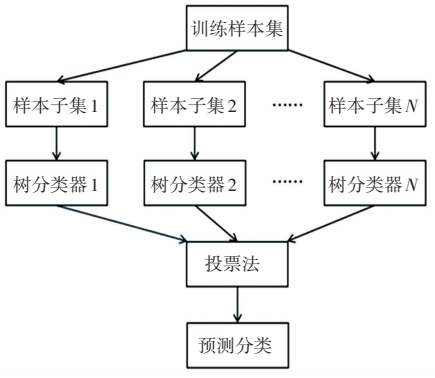

图 1　随机森林示意图

Fig. 1　The diagram of Random Forest

XGBoost 采用二阶泰勒展开来优化损失函数：

$$obj^{t} \approx \sum_{i=1}^{n} [\, l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \,] + \sum_{i=1}^{t} W(f_i) \tag{7}$$

其中，$f_t(x_i)$ 表示第 i 个样本在第 t 次迭代中的预测值；g_i、h_i 分别表示损失函数 $l(y_i, \hat{y}_i^{t-1})$ 的一阶导数和二阶导数。

根据二次矩阵最优化理论可以将式(6)化简为：

$$obj = \sum_{j=1}^{T} [\, G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \,] + \gamma T \tag{8}$$

其中，$G_j = \sum_{i \in I_j} g_i$；$H_j = \sum_{i \in I_j} h_i$；$I_j$ 表示第 j 个节点的样本集合。

可以求出最优权重为：

$$w^{*} = -\frac{G_j}{H_j + \lambda} \tag{9}$$

最终的目标函数为：

$$obj = \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{10}$$

2.4　参数优化

为了确定训练模型的最佳参数，本文采用了网络搜索方法。网格搜索[19]确定需要调整的超参数及每个超参数的可能取值范围，通过穷举搜索所有可能的超参数组合，寻找具有最佳性能的超参数组合。

选择 *AUC* 作为模型最优参数的评价指标。主要调整的模型参数和参数的调整范围见表 5。

表 5　主要调整参数和范围

Table 5　Main adjustment parameters and range

模型	调整参数	调整范围
随机森林	决策树个数	20~500
	决策树深度	1~10
XGBoost	决策树个数	20~500
	决策树深度	1~10
	学习率	[0.01,0.05,0.1]

3　实验与结果分析

3.1　实验操作和评价指标

为了研究受试者的个人基本信息和环境化学物质对糖尿病的影响，实验采用了两阶段的训练方式。具体地，首先，对仅包含个人基本信息的数据集进行模型训练；其次，基于个人基本信息和化学物质两类特征进行建模，并比较预测的结果。这种方法有助于深入探索研究不同特征对糖尿病的影响，以更全面地了解影响因素。

两阶段的模型建立均随机选取 70% 的数据作为训练集，30% 的数据作为测试集。为了提升模型的泛化能力，采用 5 折交叉验证，确保模型在不同数据子集上的表现稳定和可靠。同时，采用准确率、精确率和 *AUC* 作为模型性能的评价指标。

(1)准确率 (*Accuracy*)。准确率是模型分类正确的样本数与总样本数之比。计算公式为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

(2)召回率 (*Recall*)。召回率是模型正确预测为正类的样本数与实际为正类的样本数之比。召回率衡量了模型对正类样本的覆盖程度。计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

(3) *AUC*。*AUC* 是 *ROC* 曲线下的面积，用于度量模型的分类性能。*AUC* 越接近 1，模型性能越好。*ROC* 曲线是以不同阈值下真正率(True Positive

Rate, *TPR*) 和假正率(False Positive Rate, *FPR*) 为横纵坐标绘制的曲线。*ROC* 曲线越接近左上角,模型性能越好。*TPR* 和 *FPR* 的计算公式为:

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$FPR = \frac{FP}{FP + TN} \tag{14}$$

其中, *TP* 表示预测为正类实际也为正类的样本数;*TN* 表示预测为负类实际也为负类的样本数;*FP* 表示预测为正类实际为负类的样本数;*FN* 表示预测为负类实际为正类的样本数。

3.2 预测结果

表 6 展示了基于基本信息训练的机器学习模型的预测能力。

表 6 基于基本信息的模型测试结果

Table 6 Test results of the models based on basic information

模型	*Accuracy*	*Recall*	*AUC*
LR	0.727 3	0.670 2	0.739 0
RF	0.723 5	0.691 5	0.739 4
XGBoost	**0.731 1**	**0.712 8**	**0.770 5**

表 6 的数据显示 LR 和 RF 的预测能力相近, *AUC* 的值分别达到了 0.739 0、0.739 4,具有不错的分类预测效果。XGBoost 的预测性能最佳,拥有最高的准确率、召回率和 *AUC* 值,其中 *AUC* 值达到了 0.770 5。对于糖尿病的预测诊断问题,基于传统信息建立的模型可以得到较高的预测准确率。

在传统基本信息的基础上,纳入环境化学物质等特征,进一步训练糖尿病预测模型。表 7 展示了加入化学物质后,分类指标的结果。XGBoost 同样具有最好的预测性能,准确率、召回率和 *AUC* 值分别为 0.738 6、0.776 6 和 0.772 1,均高于其它模型,能够得到更加可信的糖尿病预测结果。

表 7 加入环境特征的测试对比

Table 7 Comparison of tests incorporating environmental features

模型	*Accuracy*	*Recall*	*AUC*
LR	0.715 9	0.723 4	0.752 6
RF	0.731 1	0.702 1	0.752 3
XGBoost	**0.738 6**	**0.776 6**	**0.772 1**

对表 6 和表 7 的结果进行综合分析:引入环境特征后,仅有 LR 的准确率略微下降,而所有模型的准确率、召回率和 *AUC* 这 3 个评价指标都表现出改善。就主要评价指标 *AUC* 而言,LR、RF 和 XGBoost 分别提高了 0.013 6、0.012 9 和 0.001 6。因此,环境化学特征对糖尿病模型的预测是有意义的。

3.3 影响因素的分析

为了评估各特征对判断是否患有糖尿病的重要程度,本文利用纳入环境化学物质暴露因素的 XGBoost 模型,计算每个特征的重要性。图 2 展示了重要性排序前 8 的特征。

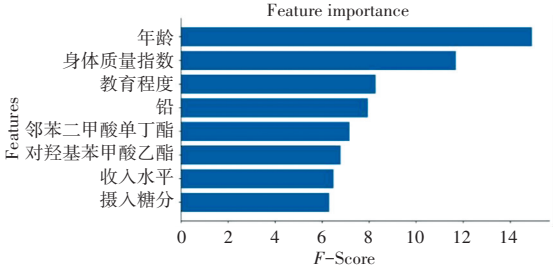

图 2 主要特征重要性分析

Fig. 2 Main feature importance analysis

图 2 显示年龄、身体质量指数、教育程度和铅是糖尿病的重要影响因素,其中年龄和身体质量指数的影响较为显著。

这 8 个重要的特征中,有 5 个基本信息特征和 3 个环境特征,这表明环境化学物质暴露因素对糖尿病具有预测价值,不应忽视环境对糖尿病的影响。

4 结束语

糖尿病的发生涉及多种因素,目前的主要治疗方法包括饮食控制、适度运动以及药物治疗。尽管已有这些治疗方法可供选用,糖尿病的临床致残率依然较高,且不少患者因糖尿病及其并发症而死亡。因此,进行早期糖尿病的风险评估和预防具有重要意义。本文采用机器学习方法,通过数据预处理和特征选择,构建了 2 种预测模型。一种是基于传统基本信息;另一种是在基本信息基础上加入环境特征。实验结果表明,XGBoost 模型具有较高的预测能力,而且环境化学物质在糖尿病的预测诊断中则有潜在的影响。

参考文献

[1] 刘洁. 糖尿病药物治疗的现状及进展研究[J]. 医学信息, 2022,35(9):69-72.

[2] SANYAOLU A, MARINKOVIC A, PRAKASH S, et al. Diabetes mellitus: An overview of the types, prevalence, comorbidity, complication, genetics, economic implication, and treatment[J]. World Journal of Meta-Analysis, 2023, 11(5): 134-143.

[3] 张丽雯,阮梅花,刘加兰,等. 糖尿病领域研发态势分析[J]. 遗传,2022,44(10):824- 839.

[4] GOYAL S, RANI J, BHAT M A, et al. Genetics of diabetes[J]. World Journal of Diabetes, 2023, 14(6): 656-679.

[5] KHOR X Y, PAPPACHAN J M, JEEYAVUDEEN M S, et al. Individualized diabetes care: Lessons from the real - world

experience[J]. World Journal of Clinical Cases, 2023, 11(13): 2890-2902.
[6] 车前子,郑启文,陈思,等. 基于人工神经网络算法的 2 型糖尿病发病风险预测模型的构建[J]. 中国慢性病预防与控制, 2020,28(4):274-279.
[7] 惠亚楠,冯慧芳. 基于改进狮群算法优化神经网络的糖尿病风险预测[J]. 软件工程与应用,2023,12(3):474-484.
[8] 张春富,王松,吴亚东,等. 基于 GA-Xgboost 模型的糖尿病风险预测[J]. 计算机工程,2020,46(3):315-320.
[9] 孙彤,陈砚桥. 基于 AHP 的 Stacking 算法基分类器选择[J]. 兵工自动化,2022,41(1):39-42.
[10] BACHORIK P. NHANES - National Health and Nutrition Examination Survey Homepage (cdc. gov)[EB/OL]. (2025-04-01). https://www. niehs. nih. gov/research/atniehs/labs/crb/studies/nhanes.
[11] WEI Hongcheng, SUNC J, SHAN Wenqi, et al. Environmental chemical exposure dynamics and machine learning - based prediction of diabetes mellitus [J]. Science of the Total Environment,2022,806(2):150674.
[12] STALLINGS-SMITH S, MEASE A, JOHNSON T M, et al. Exploring the association between polycyclic aromatic hydrocarbons and diabetes among adults in the United States[J]. Environmental Research,2018,166:588-594.
[13] HE Xiaowei, LIU Yuanxin, XU Bo, et al. PFOA is associated with diabetes and metabolic alteration in US men: National Health and Nutrition Examination Survey 2003-2012 [J]. Science of the Total Environment,2018,625:566-574.
[14] SARGIS R M, SIMMONS R A. Environmental neglect: endocrine disruptors as underappreciated but potentially modifiable diabetes risk factors[J]. Diabetologia,2019,62(10):1811-1822.
[15] EICK S M, FERRECCIO C, ACEVEDO J, et al. Socioeconomic status and the association between arsenic exposure and type 2 diabetes[J]. Environmental Research,2019,172:578-585.
[16] ZHANG Yuqing, DONG Tianyu, HU Weiyue, et al. Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: Comparison of three statistical models[J]. Environment International,2019,123:325-336.
[17] CHU Sijia, JIANG Aijun, CHEN Lyuzhou, et al. Machine learning algorithms for predicting the risk of fracture in atients with diabetes in China[J]. Heliyon,2023,9(7):e18186.
[18] ABEGAZ T M, BALJOON A, KILANKO O, et al. Machine learning algorithms to predict major adverse cardiovascular events in patients with diabetes[J]. Computers in Biology and Medicine, 2023,164:107289.
[19] 刘海钰,曲海成. 基于数据挖掘的肝癌早期复发预测与阈值研究[J]. 智能计算机与应用,2021,11(8):35-41.