

## ◆ 医疗场景 ◆

## 高血压患者管理多维度“健康指数”构建及应用研究\*

李 虎<sup>1,2</sup>, 甘昕艳<sup>1\*\*</sup>, 黄秀文<sup>2</sup>

(1. 广西中医药大学公共卫生与管理学院, 广西南宁 530200; 2. 广西壮族自治区人民医院信息网络管理中心, 广西南宁 530021)

**摘要:**本研究基于 LightGBM 算法, 构建一个基于大数据分析和人工智能技术的多维度“健康指数”评估模型, 通过综合分析个人体征、生化指标和高血压相关数据, 为高血压患者提供精准的健康管理服务。利用广西某大型三甲医院的医疗健康大数据, 设计并实现包含 23 项主要风险因素的算法模型。实证研究表明, 模型在提高高血压患者血压控制率和降低并发症风险方面展现出显著效果, 并有效提升了健康管理师的工作效率。本研究可为高血压患者管理提供一种新的量化、可视化工具, 并为其他慢性病的多维度健康管理研究提供方法论参考。

**关键词:** 高血压管理; 大数据; 机器学习; 健康指数; 健康画像

中图分类号: TP18, R544.1 文献标识码: A 文章编号: 1005-9164(2024)05-0854-10

DOI: 10.13656/j.cnki.gxkx.20250211.001

随着经济社会的发展和work节奏的加快, 全球疾病谱系和死因顺位发生了巨大变化, 高血压、糖尿病、冠心病、脑卒中等慢性病的患病率、死亡率逐年上升<sup>[1]</sup>。在我国, 高血压是患病率和致残率较高的慢性疾病, 是心血管疾病的主要代表, 给我国人民群众造成了严重的经济负担<sup>[1,2]</sup>。在高血压患者管理的场景中, 如何准确地预测患者高血压发展状态是医疗机构和研究人员十分关注的问题<sup>[3]</sup>。通过预测血压变化, 对可能发展为高血压的高风险个体采取早期干预措施, 可辅助医疗人员掌握高血压患者的疾病发展趋

势, 制定个性化治疗方案以阻止高血压疾病进一步发展, 从而降低高血压的发病率和健康风险。

传统高血压管理主要依赖于定期的血压监测、药物治疗和生活方式干预。患者通常需要定期到医院或诊所测量血压, 医生根据测量结果调整药物剂量。此外, 医生还会提供饮食、运动和戒烟等方面的建议, 以帮助患者控制血压<sup>[4]</sup>。这种基于医患交流互动的管理模式存在一些局限性, 如存在患者依从性不高、数据收集和分析效率低、血压监测缺乏实时性等问

收稿日期: 2024-05-15

修回日期: 2024-08-01

\* 广西壮族自治区健康治理智慧医疗工程研究中心建设和广西壮族自治区卫生健康委科研项目(Z-A20230148), 广西高等教育本科教学改革工程重点项目(2023JGZ138)资助。

## 【第一作者简介】

李 虎(1983—), 男, 高级工程师, 主要从事医疗信息化、健康大数据研究。

## 【\*\*通信作者简介】

甘昕艳(1972—), 女, 教授, 硕士研究生导师, 主要从事健康教育与健康促进、卫生信息技术研究, E-mail: 371590205@qq.com。

## 【引用本文】

李虎, 甘昕艳, 黄秀文. 高血压患者管理多维度“健康指数”构建及应用研究[J]. 广西科学, 2024, 31(5): 854-863.

LI H, GAN X Y, HUANG X W. Research on the Construction and Application of Multi-Dimensional “Health Index” in the Management of Hypertensive Patients [J]. Guangxi Sciences, 2024, 31(5): 854-863.

题。随着医疗信息化技术的发展,医疗健康数据量大幅增加。通过电子健康记录、移动健康应用以及远程监测系统,医疗机构能够收集到大量实时、连续的健康数据,不仅包括血压、心率等生理指标,还包括患者的日常行为、饮食习惯、运动频率等生活习惯数据<sup>[5]</sup>。在大数据时代,医疗健康领域产生了海量高容量和高维度的数据,采用机器学习、数据挖掘技术处理此类数据的优势日趋明显,随机森林、神经网络和支持向量机等机器学习方法在高血压风险预测领域的应用研究愈发广泛<sup>[6-9]</sup>。崔伟锋等<sup>[6]</sup>提出基于随机森林的疾病风险预后模型,对原发性高血压心血管风险的中西医预后危险因素进行筛选。Chen等<sup>[7]</sup>建立支持向量机回归模型和随机森林回归模型用于精确测量血压,发现后者在表现上优于前者,更接近水银血压计测量值。Shrivastava等<sup>[8]</sup>分别使用K最邻近算法(K-Nearest Neighbors, KNN)逻辑回归、决策树、随机森林预测收缩压和舒张压,结果表明随机森林表现出最高的预测准确率。Nusinovici等<sup>[9]</sup>在亚洲成年人队列中,对比分析5种机器学习模型对高血压和心血管疾病的预测效果,包括单隐藏层的神经网络、支持向量机、随机森林、梯度提升和KNN。这些研究主要采用单一的机器学习方法来开展研究工作。虽然构造的算法模型在实验中表现出较好的性能,但是单一的机器学习模型仍然存在准确率不高、泛化能力弱等不足,在高血压患者管理场景中缺少实践验证。

针对单一机器学习方法存在的问题,集成学习方法通过组合多个弱学习器,优势互补形成强学习器<sup>[10-12]</sup>。LightGBM算法是一种典型的强学习器算法,它在梯度提升决策树(Gradient Boosting Decision Tree, GBDT)的基础上,引入了基于梯度的单边采样和互斥捆绑特征技术,旨在减少计算和内存消耗,是一种高效的梯度提升决策树算法<sup>[13]</sup>。目前LightGBM算法被应用于多个领域,尤其在经济与金融领域中具有较好的表现<sup>[14-18]</sup>。苗月等<sup>[14]</sup>使用LightGBM算法构建了信用风险评估模型,实验结果表明客户属性的分类精确度优于XGBoost(Extreme Gradient Boosting)集成学习算法,并且执行效率更高。付芷宁等<sup>[15]</sup>针对贷款市场复杂的个人信用风险问题,基于LightGBM算法构建了信用风险评估模

型,并且将其与逻辑回归、决策树等传统评估模型进行对比,实验证明LightGBM模型的效果最好。Zhao等<sup>[16]</sup>使用LightGBM算法预测股票价格,在所有测试数据上获得了最高的年回报率,优于XGBoost和决策树等算法。Bakhtiari等<sup>[17]</sup>考察LightGBM算法在信用卡欺诈检测中的应用,研究结果表明该算法通过集成学习防范和适当的组合策略,能够有效检测信用卡欺诈行为。在医疗领域,Han等<sup>[18]</sup>利用LightGBM分类器对阿尔茨海默病进行分类,并在真实数据集上进行评估,实验结果表明,LightGBM分类器的表现优于其他常规分类器。LightGBM算法能进一步提升机器学习模型的性能,但是目前在高血压患者管理方面研究较少。

高血压关联数据集通常包含大量患者的历史健康数据,包括血压读数、生活方式、遗传因素、医疗记录等,这些数据可能包含复杂的特征空间。另外,影响血压变化的风险因素众多,包括年龄、性别、体重、饮食习惯、遗传因素等。由于这些特性,应用于高血压患者管理的机器学习模型需要能够高效处理大规模、高维度的数据集。LightGBM算法在处理大规模和高维度数据集上有明显优势,能够在较少的计算资源和时间内,快速地训练出高性能模型,而且LightGBM算法的决策过程具有可解释性,可以帮助研究人员更好地理解模型的工作原理和预测结果<sup>[13-18]</sup>。因此,本研究提出一种基于LightGBM的“健康指数”算法模型,整合多种风险因素以较为全面地评估患者健康状态,并将模型部署到医院的主动健康管理应用中,帮助医护人员高效管理高血压患者。

## 1 模型与算法

### 1.1 风险因素与赋分规则

#### 1.1.1 筛选风险因素

根据高血压相关防治指南<sup>[19,20]</sup>,采用德尔菲法<sup>[21]</sup>收集广西某大型三甲医院(以下简称“三甲医院”)心血管内科专家团队的意见,筛选年龄、性别、体质指数(BMI)/腰围、血压、血糖、血脂、高尿酸血症、高同型半胱氨酸血症、家族史、生活习惯、心理疾病、靶器官损害、伴发临床疾病等风险因素作为高血压健康指数的评价指标(图1)。

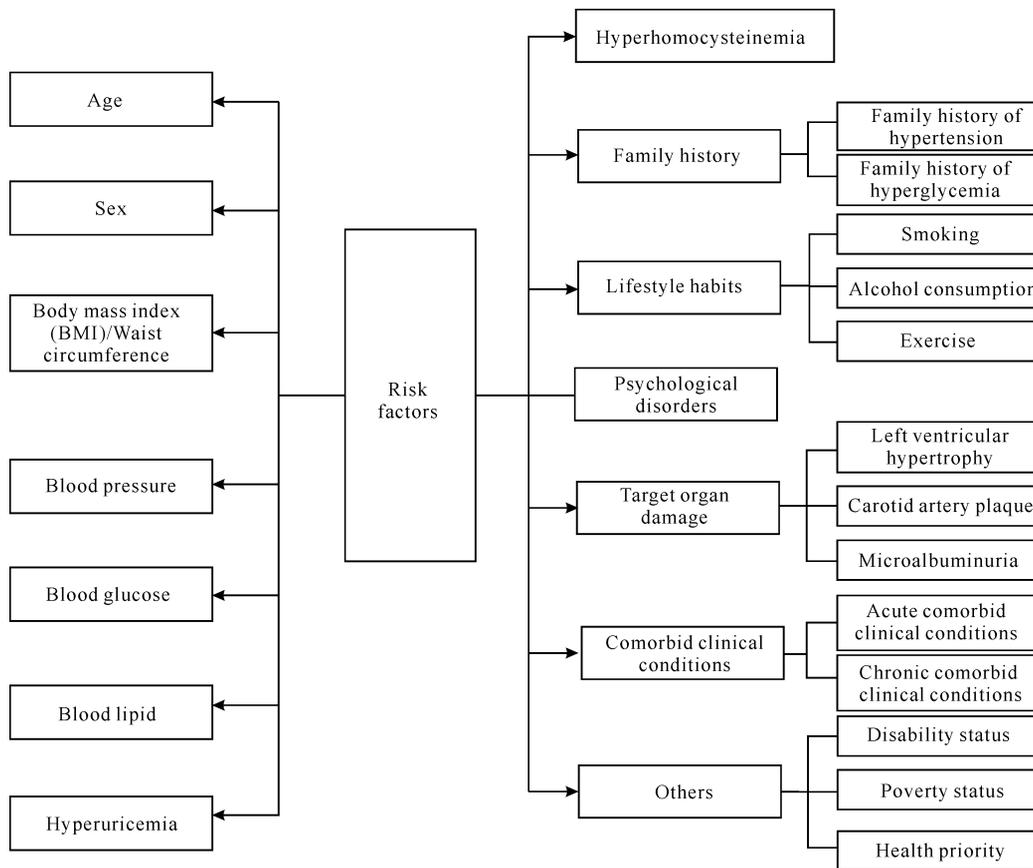


图1 风险因素指标

Fig. 1 Indicators of risk factors

### 1.1.2 制定赋分规则

如表1所示,每项指标按0—10分进行赋分,根据变量类型及其危害程度,采用分类分段、连续数值、等级数值等多种方法对各项因素赋初始值。例如,性别一般情况下只分为男女两类,是分类变量,采取分类赋值方法;血压、血糖和血脂的取值范围理论上可取任意实数,因此它们是连续变量,需采取连续数值进行赋值;对于吸烟、运动等具有程度差异的风险因素,可根据行为频率或强度进行划分,并采用等级赋值方法进行量化。

通过统计分析三甲医院的诊疗数据集,研究年龄、血压、血糖、BMI、血脂等数值变量与住院率的关系,然后进行初次赋分。以“年龄”风险因素为例,描述数值变量的初次赋分策略。本研究选取三甲医院主动健康大数据中心的1万名居民在2022年度产生的诊疗数据作为基础数据,通过统计分析年龄与住院率的关系,发现不同年龄阶段的居民高血压危险程度不同,总体呈现正相关趋势。<35岁患者的住院率最低,作为最小值,赋0分;≥85岁患者的住院率最高,作为最大值,赋10分;每5岁划分一个年龄段,在

最大住院率和最小住院率之间,不同年龄段按照住院率相对增量比进行赋分,公式为:某年龄段分值=(该年龄段患者住院率—最小住院率)×10/(最大住院率—最小住院率),由此可得各年龄段的分值,如表2所示。性别、家族史、生活习惯、心理疾病等分类变量参考《中国高血压防治指南(2018年修订版)》<sup>[19]</sup>和《中国2型糖尿病防治指南(2017年版)》<sup>[22]</sup>进行初次赋分;靶器官损害、伴发临床疾病等变量通过临床归类后,再根据急、慢性病特征进行分级分类赋分。所有变量初次赋分后,再根据德尔菲法<sup>[21]</sup>收集临床意见进行多轮修正。

表1 风险因素和变量分值

Table 1 Risk factors and variable scores

序号 Number	风险因素 Risk factor	变量分值/分 Range of variable score/points
1	Age	0—10
2	Sex	0—10
3	BMI/Waist circumference	0—10
4	Family history	0—10

续表

Continued table

序号 Number	风险因素 Risk factor	变量分值/分 Range of variable score/points
5	Blood pressure levels and control status	0-10
6	Blood pressure	0-10
7	Initial blood glucose level and control status	0-10
8	Blood glucose	0-10
9	Blood lipid	0-10
10	Hyperhomocysteinemia	0-10
11	Hyperuricemia	0-10
12	Psychological disorders	0-10
13	Smoking	0-10
14	Alcohol consumption	0-10
15	Exercise	0-10
16	Disability status	0-10
17	Poverty status	0-10
18	Health priority	0-10
19	Left ventricular hypertrophy	0-10
20	Carotid artery plaque	0-10
21	Microalbuminuria	0-10
22	Acute comorbid clinical conditions	0-10
23	Chronic comorbid clinical conditions	0-10

表 2 风险因素-年龄变量分值

Table 2 Risk factor-age variable scores

年龄/岁 Age/a	变量分值/分 Variable score/points	住院率/% Hospitalization rate/%
<35	0.0	5.65
35-39	1.6	9.13
40-44	1.7	9.32
45-49	1.8	9.62
50-54	2.0	9.87
55-59	2.4	10.76
60-64	3.2	12.55
65-69	3.8	13.94
70-74	5.4	17.32
75-79	7.4	21.72
80-84	9.3	25.73
≥85	10.0	27.28

## 1.2 数据预处理

### 1.2.1 缺失值处理

缺失值处理涉及识别、分析和处理数据集中的缺失值,以提高数据质量和后续分析的准确性。对于缺失值,具体处理如下:确定缺失值范围;删除缺失值超一半的无价值记录;填充缺失内容,其中连续特征通过计算均值进行填充,分类特征使用众数填充。例如,部分记录的血压有所缺失,进一步分析发现这些记录主要集中在年龄较大的患者中,这可能是老年患者血压测量频率偏低所致。考虑到血压是重要的健康指标,需对缺失值进行填充,通过计算该年龄段患者的平均血压值进行填充。

### 1.2.2 格式内容清洗

格式内容清洗是指对数据集中的内容进行标准化和纠正,以确保数据的一致性和准确性。首先,识别数据集中存在的不一致或错误的内容,包括拼写错误、格式不一致、单位不统一、数据类型错误等;根据数据特点和建模要求,制定拼写纠正、格式标准化、单位转换、数据类型转换等不同的清洗规则;通过手动修改、自动化处理脚本或数据清洗工具等进行清洗。最后检查数据集,保证所有记录的风险因素特征都符合统一的格式和标准。

### 1.2.3 数据分割

从三甲医院主动健康大数据中心选取原始健康数据作为样本,样本特征包括居民基本信息、慢性病专项信息、慢性病分级信息、随访记录、生活习惯、BMI、既往疾病史、家族疾病史等。经缺失值处理、格式内容清洗、结构化和标准化过程后,选取6万名用户的数据为源数据,其中5万份作为测试集,用于测试评分模型的预测准确性;根据赋分规则对剩余1万份源数据赋分,赋分结果为标签,作为模型的训练集。

## 1.3 “健康指数”模型

### 1.3.1 LightGBM 算法

LightGBM 是一种基于 GBDT 算法改进的算法<sup>[23-25]</sup>,适用于大规模数据集的训练和预测。GBDT 算法虽然具备较好的准确性,但在分裂决策树节点时,需对全局数据集的每一项特征变量进行遍历,以求出当前节点的最优分裂特征值,遍历过程十分耗时。为解决 GBDT 算法耗时长、执行效率低的问题,LightGBM 算法利用基于直方图的决策树优化策略,节省 GBDT 算法的执行时间,此外引入最大深度限制的叶子生长策略也使得算法的整体执行效率更好。

①LightGBM 算法采用直方图优化策略搜索最优分割点。直方图优化策略的主要原理:在训练前,对样本中每一维特征进行排序,然后对特征通过直方图进行划分,后续训练中算法使用直方图作为“特征”进行直方图构建,在连续遍历数据集后,根据直方图累积的统计量确定最优分割点。直方图优化策略将连续的特征值离散,使得直方图数量远少于数据量,因此,和预排序方法相比,采用直方图算法寻找最优分割点所需时间更少、占用内存更小。图2显示了直方图算法的核心思路。

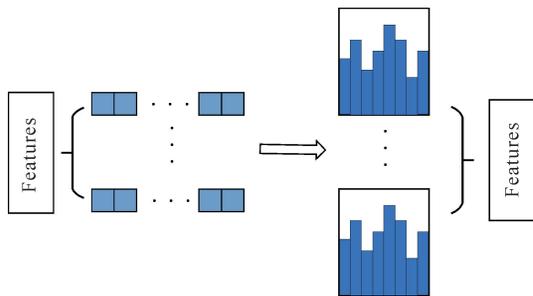


图2 直方图算法示意图

Fig. 2 Schematic diagram of the histogram algorithm

②在直方图算法之上,LightGBM 算法进一步选择了具有深度限制的按叶生长策略(图3)。LightGBM 算法每次从当前所有叶子中寻找分裂增益最大的叶子,然后分裂和循环。同按层生长策略(图4)相比,在分裂次数相同的情况下,按叶生长策略减少了训练误差,提高了算法的分类精度。但是,按叶生长策略的缺点是容易出现较深的决策树,产生过拟合,因此,LightGBM 算法在其之上增加了深度限制来预防过拟合问题。

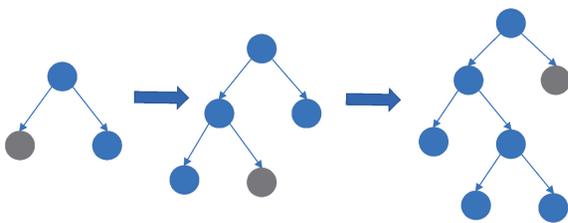


图3 按叶生长策略

Fig. 3 Leaf-wise growth strategy

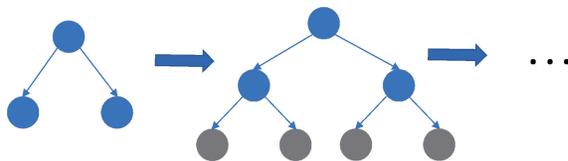


图4 按层生长策略

Fig. 4 Level-wise growth strategy

③LightGBM 算法在GBDT 决策树算法基础上引入梯度单边采样技术和独立特征合并技术。梯度

单边采样技术使得 LightGBM 算法在较传统 GBDT 算法少很多数据的情况下仍实现较高的准确率,独立特征合并技术能捆绑互斥特征,以减少计算冗余和提高内存利用率。LightGBM 模型在保持模型准确性的同时,显著减少了特征的数量和计算量,是学术界和工业界广泛使用的工具。本研究将居民的专家打分结果作为输入,对 LightGBM 模型进行训练,经过迭代修正后,模型输出即为健康指数分值。

### 1.3.2 “健康指数”模型算法

本研究的训练样本由风险因素和变量分值组成,训练样本集表示为  $\{X_m, Y\}$ ;  $X_m$  为居民高血压风险因素数据集,其中  $m$  表示数据集的样本数,  $n$  表示每个样本的特征数;  $Y$  为风险因素的变量分值向量,表示为  $Y = (y_1, y_2, \dots, y_m)^T$ 。算法流程如下。

①初始化生成首棵决策树  $f_0(x)$  :

$$f_0(x) = \operatorname{argmin}_b \sum_{i=1}^m L(y_i, b), \quad (1)$$

式中,  $b$  为损失函数最小化的常数;  $L(y_i, b)$  为损失函数,用于真实值和预测值之差。

②进行  $K$  次迭代训练,在建立一系列 CART 回归树基础之上利用梯度提升技术分析残差来建立拟合模型。在第  $k$  ( $k=1, 2, \dots, K$ ) 次迭代中,对于任意样本  $x_{in}$ ,算法将样本的残差估计值规定为损失函数的负梯度值,因此样本  $x_{in}$  在第  $k$  次迭代中的残差估计值可确定为

$$r_{k,i} = - \left( \frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right), \quad (2)$$

式中,  $r_{k,i}$  为损失函数的负梯度值,  $f(x_i)$  是样本  $x_i$  在此次迭代训练中的拟合模型,  $L(y_i, f(x_i))$  是  $f(x_i)$  和  $x_i$  的真实变量分值  $y_i$  的损失函数。

③残差确定后,采用 CART 回归树进行拟合,此时设定回归树存在  $J$  个叶节点,叶节点域表示为  $C_{jk}$  ( $j=1, 2, \dots, J$ )。在拟合过程中,任务是确定每个叶节点损失函数最小化的最佳拟合值  $\alpha_{jk}$  :

$$\alpha_{jk} = \operatorname{argmin}_\alpha \sum_{x_{in} \in C_{jk}} L(y_i, f_{k-1}(x_{in}) + \alpha), \quad (3)$$

式中,  $f_{k-1}(x_{in})$  是第  $k-1$  次迭代的拟合模型,  $\alpha$  是使第  $k-1$  次回归树损失函数最小的值。在确定  $\alpha_{jk}$  值之后,第  $k$  次迭代对第  $k-1$  次形成的弱学习器进行强化,得到强学习器  $f_k(x_{in})$  :

$$f_k(x_{in}) = f_{k-1}(x_{in}) + \sum_{j=1}^J \alpha_{jk} \times \eta, \quad (4)$$

式中,  $\eta$  是学习速率。迭代结束后,最终得到目标函数  $F(x)$ ,可表示为

$$F(x) = f_0(x) + \sum_{k=1}^K \sum_{j=1}^J \alpha_{jk} \times \eta. \quad (5)$$

### 1.3.3 模型性能评价

$R^2$  称为决定系数(R-squared),用于回归模型的拟合度量,能够描述模型对观测值的解释程度。给定一个样本集  $y$ ,  $y_i$  表示样本集中第  $i$  个样本的真实值,该样本输入模型后得到预测值  $\hat{y}_i$ ,  $\bar{y}$  是样本集真实值的均值。 $R^2$  评价指标的计算方法如下:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}, \quad (6)$$

式中, $R^2$  取值范围为 $[0,1]$ ,若  $R^2$  为 0,表明模型拟合效果很差;若  $R^2$  为 1,说明模型预测值和真实值相同。通常, $R^2$  越大表示模型拟合效果越好。

### 1.3.4 “健康指数”模型训练

将 1 万名居民高血压的打分结果作为模型的训练集,未打分的居民健康数据作为测试集。训练集涵盖 1 万个样本,23 个特征变量,即 23 项高血压风险因素;1 个目标变量,即居民的健康指数分值。模型训练完成后,对测试集的 5 万个居民样本进行预测,最终得到健康状况评分。整合训练集和测试集的预测数据,形成包含 6 万份高血压特征样本及其健康指数分值的数据集,用于后续确定各项风险因素的权重因子。模型实现流程见图 5。

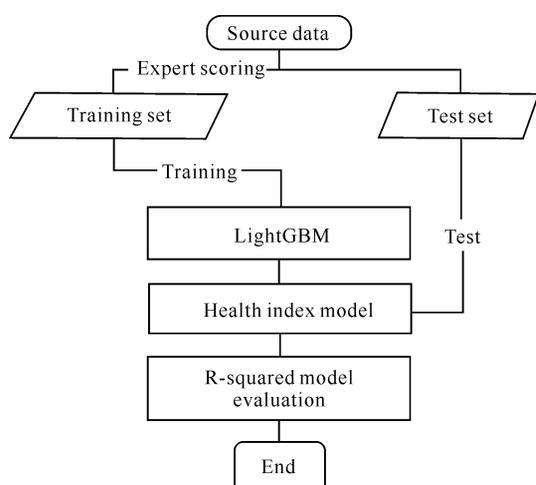


图 5 模型实现流程图

Fig. 5 Flowchart of model implementation

### 1.4 风险因素权重目标函数设定

前文构建了根据居民高血压风险因素自动打分的模型,但在实际推广应用,仍需设计一个简便易

行的指数计算方法,以能够便捷地计算出高血压健康指数。使用 SHAP 技术<sup>[26]</sup>分析风险因素对预测结果的非线性影响和因素间的相对重要度,得出结论:本研究的 23 项风险因素与高血压的危险程度呈现线性关系,即风险因素的水平越高,危险程度也相应越高,各项因素之间的关联性对预测结果的影响较弱,每项风险因素都是模型的有效特征。为此,本研究采用相加模型作为拟合的基础。该模型认为,当两个或两个以上的因素共同作用于某一事件时,如果它们之间不存在交互作用,那么它们的效应可以简单相加。在构造拟合函数之前,先定义目标函数  $f$ :

$$f = \sum_{i=1}^{23} a_i \times x_i + b, \quad (7)$$

式中,  $\{a_1, a_2, \dots, a_{23}\}$  是每个特征变量的权重,  $\{x_1, x_2, \dots, x_{23}\}$  为所有特征变量的集合,  $b$  是线性函数的偏移量。

## 2 结果与分析

### 2.1 自动评分模型效果评价

在模型训练完成后,经计算,模型效果评价指标  $R^2$  为 0.831 8。通常, $R^2$  大于 0.75,表示模型拟合度较好,可解释程度较高;反之,模型拟合不成功,不宜进行回归分析。因此,自动评分模型的预测效果较好,预测值与真实值较接近。另外,对模型输出的评分结果进行抽样,并通过专家审核,结果表明评分结果具有较高的准确度。

### 2.2 风险因素权重因子

将 23 项风险因素作为特征变量,每一项风险因素的权重因子是目标变量,公式(2)的函数是目标函数,特征变量、目标变量、目标函数作为输入,使用 Python 依赖库中 `curve_fit` 函数进行拟合,最终获得 23 项风险因素的权重因子(表 3)。

表 3 各类风险因素的权重

Table 3 Weights of various risk factors

序号 Number	风险因素 Risk factor	权重/% Weight/%
1	Age	5.5
2	Sex	1.3
3	BMI/Waist circumference	2.2
4	Family history	1.5
5	Blood pressure levels and control status	7.9

续表

Continued table

序号 Number	风险因素 Risk factor	权重/% Weight/%
6	Blood pressure	1.4
7	Initial blood glucose level and control status	7.7
8	Blood glucose	1.3
9	Blood lipid	5.7
10	Hyperhomocysteinemia	3.0
11	Hyperuricemia	2.8
12	Psychological disorders	1.0
13	Smoking	1.3
14	Alcohol consumption	1.2
15	Exercise	1.2
16	Disability status	1.1
17	Poverty status	1.0
18	Health priority	0.9
19	Left ventricular hypertrophy	7.0
20	Carotid artery plaque	5.2
21	Microalbuminuria	5.3
22	Acute comorbid clinical conditions	17.3
23	Chronic comorbid clinical conditions	17.2
Total		100.0

表4 健康指数在不同高血压发展阶段的管理效果

Table 4 Management effectiveness of health index across different stages of hypertension development

	组别 Group	高血压用户 Users with hypertension		
		Stage 1	Stage 2	Stage 3
Health index group	Before management	100	100	100
	After management	47	25	11
	Decline rate/%	53	75	89
Non health index group	Before management	100	100	100
	After management	60	42	28
	Decline rate/%	40	58	72
Improvement rate of hypertension management effectiveness/%		13	17	17

经过1年的高血压管理,健康指数组各期高血压用户的血压值均有较为明显的下降,其中3期高血压用户人数下降最多。与非健康指数组相比,在引入健康指数后,高血压用户的血压下降效果更显著,高血压的管理效果有了明显提升。因此,健康指数能够有效帮助居民及时了解自身血压变化情况,尽早采取干预措施,防止血压升高;而对于高血压人群,健康指数

### 2.3 健康指数计算方法

为满足实际应用需求,高血压健康指数的基数分设定为1000分,高血压健康指数由基数分减去23项风险因素变量分值与其对应风险因素权重的乘积再乘以100,得分越低意味着风险越高。计算公式如下:

$$z = 1000 - \sum_{i=1}^{23} (x_i \times y_i) \times 100, \quad (8)$$

式中, $z$ 代表高血压健康指数, $x$ 代表各风险因素的变量分值, $y$ 代表对应的风险因素权重。

### 2.4 模型应用实践

为分析健康指数对高血压人群的实际管理效果,本研究根据3级高血压分期标准,分别在1期高血压(140/90 mmhg ≤ 血压 < 160/100 mmhg)、2期高血压(160/100 mmhg ≤ 血压 < 180/110 mmhg)、3期高血压(180/110 mmhg ≤ 血压 < 220/120 mmhg)用户中选取200名用户。在每个高血压级别中,用户被划分为健康指数组与非健康指数组,两组人数均为100人。健康指数组引入健康指数对高血压用户进行健康管理,非健康指数组则采用传统的健康管理方式。两组用户均在2022年1月开始采用对应的健康管理方案进行干预,2023年1月干预结束,统计用户干预结束后的血压变化情况(表4)。

能够反映降血压的效果,通过量化评估用户的高血压风险,准确分析用户的高血压发展阶段,对症下药采取更有针对性的干预方案,帮助用户及时控制血压,防止高血压继续发展进而产生并发症。

将高血压“健康指数”模型部署在三甲医院的主动健康管理系统上,在健康管理师分类管理和居民自我健康管理两种应用场景中进行应用,评估模型在实

际业务场景中的表现。

①健康管理师分类管理。图6的主动健康管理信息系统页面显示,健康管理师依据系统中的健康指数对高血压患者进行分类管理;普通管理或强化管理,确保重点人群得到充分关注,并据此制定个性化、精准化的健康管理方案。健康管理师通过健康指数的变化,把握高血压患者在一段时间内的健康动态,进而对管理方案进行及时、有效的调整。在日常随访和分层分级评估过程中,健康管理师可借助健康指数进行

智能随访与评估。

②居民自我健康管理。医院的移动端健康管理应用也集成了高血压健康指数模型,应用界面如图7所示。居民通过健康指数分值的大小,直观地感知高血压风险程度的演变,通过量化指标清晰地把握自身的健康状况及发展趋势。应用程序中的智能推荐算法会根据健康指数的动态变化,精准推送健康科普知识,并根据个人健康状况推送健康管理建议和处方,让居民积极、主动地参与自身的健康管理。

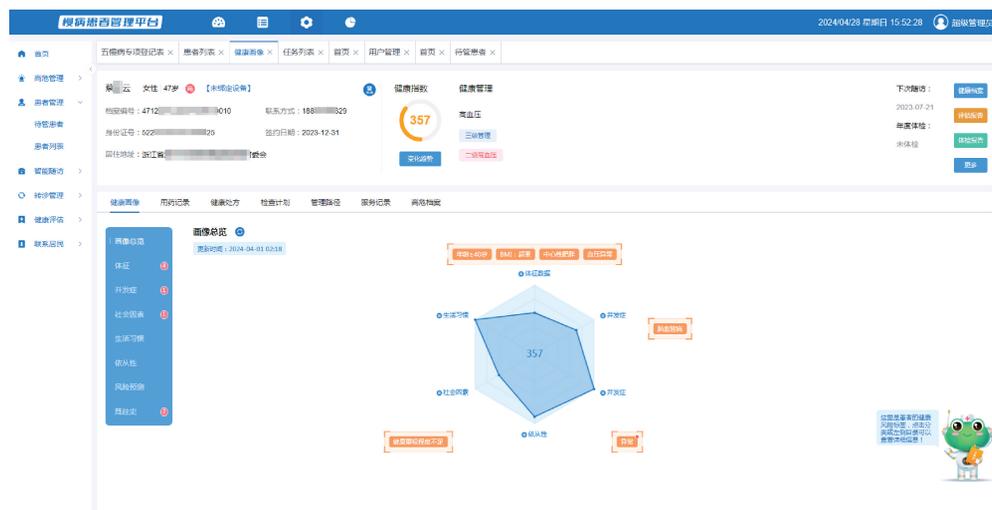


图6 主动健康管理信息系统页面

Fig. 6 Functional interface of the active health management system

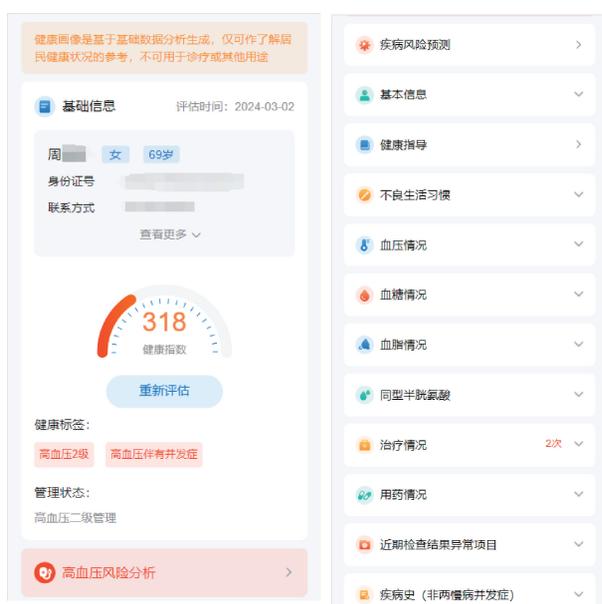


图7 移动端健康管理应用——“健康指数”页面

Fig. 7 “Health index” functional interface of the mobile health management application

### 3 结论

本研究根据国家高血压相关防治指南及专家意见,筛选年龄、性别、血压、高尿酸血症等23项作为风险因素,在居民电子健康档案收集数据并进行预处理,形成有效数据集,基于LightGBM算法框架构建多维度高血压“健康指数”评分模型,并将其部署在广西某大型三甲医院的主动健康管理系统上开展实证研究。实证结果表明,1期、2期、3期高血压用户的收缩压和舒张压都有较为明显的下降,高血压管理效果有所提升。多维度高血压“健康指数”模型能够帮助居民直观监测自身的高血压水平,为健康管理师提供了参考依据,为制定个性化健康干预方案提供了一种量化工具。然而模型也存在一些局限性,如医疗健康数据的完整性和有效性对模型结果的准确性有着直接影响。在未来,医疗健康数据的标准化和格式化是需关注的研究方向,另一个研究方向是调整风险因素和改进算法模型,推广应用到糖尿病、脑卒中等其他慢性病人群的健康管理中。

## 参考文献

- [1] 中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告 2020 概要[J]. 中国循环杂志, 2021, 36(6): 521-545.
- [2] 杨玲娜, 彭思涵, 梁小利, 等. 中国社区老年高血压患者自我管理现状及影响因素[J]. 中国老年学杂志, 2018, 38(10): 2536-2538.
- [3] SILVA G F S, FAGUNDES T P, TEIXEIRA B C, et al. Machine learning for hypertension prediction: a systematic review [J]. *Current Hypertension Reports*, 2022, 24(11): 523-533.
- [4] CHOBANIAN A V, BAKRIS G L, BLACK H R, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report [J]. *The Journal of the American Medical Association*, 2003, 289(19): 2560-2572.
- [5] BEAM A L, KOHANE I S. Big data and machine learning in health care [J]. *The Journal of the American Medical Association*, 2018, 319(13): 1317-1318.
- [6] 崔伟锋, 刘萧萧, 韩静旖, 等. 基于随机森林的原发性高血压心血管风险预后模型[J]. 中国老年学杂志, 2020, 40(4): 814-816.
- [7] CHEN X, YU S, ZHANG Y, et al. Machine learning method for continuous noninvasive blood pressure detection based on random forest [J]. *IEEE Access*, 2021, 9: 34112-34118.
- [8] SHRIVASTAVA A, CHAKKARAVARTHY M, SHAH M A. A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics [J]. *Healthcare Analytics*, 2023, 4: 100219.
- [9] NUSINOVICI S, THAM Y C, CHAK YAN M Y, et al. Logistic regression was as good as machine learning for predicting major chronic diseases [J]. *Journal of Clinical Epidemiology*, 2020, 122: 56-69.
- [10] CHEN Z, DUAN J, YANG C, et al. SMLBoost-adopting a soft-margin like strategy in boosting [J]. *Knowledge-Based Systems*, 2020, 195: 105705.
- [11] 史东超. XGBoost 算法在河北省 GPM 卫星降水数据降尺度中的应用[J]. 水科学与工程技术, 2024(2): 35-38.
- [12] 王永亮, 李超, 许恩永, 等. 基于 XGBoost 算法的商用车驾驶风险辨识模型[J]. 机械设计与制造, 2024(8): 84-89, 95.
- [13] LIAO H, ZHANG X, ZHAO C, et al. LightGBM: an efficient and accurate method for predicting pregnancy diseases [J]. *Journal of Obstetrics and Gynaecology*, 2022, 42(4): 620-629.
- [14] 苗月, 吴陈. 基于 RF-FL-LightGBM 算法的信用风险评估模型研究[J]. 计算机与数字工程, 2024, 52(3): 808-813.
- [15] 付芷宁, 李慧敏, 徐亚田, 等. 基于自适应惯性权重 PSO-LightGBM 的信用风险评估研究[J]. 云南民族大学学报(自然科学版), 2024, 33(3): 345-350.
- [16] ZHAO X, LIU Y, ZHAO Q. Cost harmonization LightGBM-based stock market prediction [J]. *IEEE Access*, 2023, 11: 105009-105026.
- [17] BAKHTIARI S, NASIRI Z, VAHIDI J. Credit card fraud detection using ensemble data mining methods [J]. *Multimedia Tools and Applications*, 2023, 82(19): 29057-29075.
- [18] HAN L, YANG T, PU X, et al. Alzheimer's disease classification using LightGBM and euclidean distance map [C]//2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). Piscataway, NJ: IEEE, 2021, 5: 1540-1544.
- [19] 《临床医学研究与实践》编辑部. 中国高血压防治指南(2018年修订版)[J]. 临床医学研究与实践, 2019, 4(5): 201.
- [20] 国家基本公共卫生服务项目基层高血压管理办公室, 基层高血压管理专家委员会. 国家基层高血压防治管理指南[J]. 中国循环杂志, 2017, 32(11): 1041-1048.
- [21] 王少娜, 董瑞, 谢晖, 等. 德尔菲法及其构建指标体系的应用进展[J]. 蚌埠医学院学报, 2016, 41(5): 695-698.
- [22] 中华医学会糖尿病学分会. 中国 2 型糖尿病防治指南(2017年版)[J]. 中国实用内科杂志, 2018, 38(4): 292-344.
- [23] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine [J]. *The Annals of Statistics*, 2001, 29(5): 1189-1232.
- [24] FRIEDMAN J H. Stochastic gradient boosting [J]. *Computational Statistics & Data Analysis*, 2002, 38(4): 367-378.
- [25] KE G, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree [J]. *Neural Information Processing Systems*, 2017, 30: 52.
- [26] 金东镇, 郭城楠, 彭芳, 等. 从 SHAP 到概率: 可解释性机器学习在糖尿病视网膜病变靶向脂质组学研究中的应用[J]. 中国卫生统计, 2023, 40(4): 511-515.

# Research on the Construction and Application of Multi-Dimensional “Health Index” in the Management of Hypertensive Patients

LI Hu<sup>1,2</sup>, GAN Xinyan<sup>1\* \*</sup> , HUANG Xiuwen<sup>2</sup>

(1. School of Public Health and Management, Guangxi University of Chinese Medicine, Nanning, Guangxi, 530200, China;

2. Information Network Management Center of the People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530021, China)

**Abstract:** In this study, a multi-dimensional “health index” assessment model based on big data analysis and artificial intelligence technology was constructed using LightGBM algorithms. The model provides precise health management services for hypertensive patients by comprehensively analyzing individual signs, biochemical indicators, and hypertension-related data. The research designs and implements an algorithm model containing 23 major risk factors using the medical and health big data from a large tertiary hospital in Guangxi. Through empirical research, the model demonstrates significant effects in improving the blood pressure control rate and reducing the risk of complications in hypertensive patients, and effectively enhances the efficiency of health managers. This study provides a new quantitative and visual tool for hypertension patient management and serves as a methodological reference for multidimensional health management research in other chronic diseases.

**Key words:** hypertension management; big data; machine learning; health index; health profile

责任编辑:陆媛峰



微信公众号投稿更便捷

联系电话:0771-2503923

邮箱:gxkx@gxas.cn

投稿系统网址:<http://gxkx.ijournal.cn/gxkx/ch>