

引用: 窦智丽, 孙浩南, 刘一星, 张佳玮, 贾金柱, 韩东燃. 中医治疗 2 型糖尿病证候预测模型的建立[J]. 湖南中医杂志, 2023, 39(8): 135-140.

中医治疗 2 型糖尿病证候预测模型的建立

窦智丽¹, 孙浩南², 刘一星³, 张佳玮², 贾金柱^{1,4}, 韩东燃²

(1. 北京大学医学部公共卫生学院, 北京, 100191;

2. 北京中医药大学生命科学学院, 北京, 102488;

3. 北京中医药大学管理学院, 北京, 102488;

4. 北京大学统计科学中心, 北京, 100781)

[摘要] 目的: 构建 2 型糖尿病中医证候诊断预测模型, 探索中医治疗 2 型糖尿病特征对疾病预测的实际意义。方法: 本研究使用支持向量机、决策树、逻辑回归、K 近邻算法等, 首先构建标准化的症状、证型获取途径; 再分别融合中医治疗数据特征。结果: 本文通过对 321 篇随机对照试验(RCT)文章的诊疗信息统发现有 21 个证候类型、99 个主观症状、13 个中医脉象表现、23 个中医舌象、232 味中草药、20 个中医穴位。其中气阴两虚、口燥咽干、细脉、脾俞穴分别为高频; 决策树、支持向量机、逻辑回归算法的评估效果最佳。结论: 决策树是预测糖尿病中医证候分类的最佳算法。构建包含中医主观症状、舌象、脉象, 有助于根据早期症状表现划分中医证候类别。

[关键词] 2 型糖尿病; 中医证候; 预测模型

[中图分类号] R259.871 **[文献标识码]** A **DOI:** 10.16808/j.cnki.issn1003-7705.2023.08.035

机器学习方法作为数据挖掘和人工智能的重要方法, 已经被广泛应用于医疗领域, 在疾病的诊疗中发挥着重要的作用^[1]。2 型糖尿病作为中国的常见病之一, 也经常应用于预测模型的构建中, 常用的有 Cox 回归^[2-3]、多元 logistic 回归^[4]、最大似然估计^[5]、贝叶斯等^[6], 见表 1。而随着数据类型的扩大, 机器学习算法从多模态数据中自动提取特征, 使用有监督学习方法建立预测模型, 更有利于提高效率以及模型的准确度。

表 1 传统机器学习方法和深度学习算法的比较

项目	传统机器学习算法	深度学习算法
数据集规模	不限	海量数据
特征提取	相关专家进行特征设计和提取	原始数据自动提取
可解释性	具有良好的解释性	黑盒, 中间过程难以解释
优势	适用于对于特征提取容易的数据	适用于发病率评估、用药推荐

临床疗效预测模型的因变量常为某种临床结

局, 如疾病的发生、发展、预后、转归或者不良反应等。自变量则存在多种数据形式, 如性别、用药等分类变量, 以及生化指标中的连续性变量等。最终, 需要研究者根据不同的研究目的和数据形态, 选择恰当的疗效预测模型。

近 10 年来, 中国糖尿病患者急剧增加, 目前已经成为世界糖尿病第一大国。中国糖尿病的发病类型以 2 型糖尿病为主, 其发病率高、并发症广、致死致残率高, 除给患者自身造成巨大的痛苦外, 也给家庭及社会带来了极大的经济负担。因此, 积极预防早期治疗糖尿病至关重要。本研究拟采用症状-证候智能预测模型, 评价中医治疗 2 型糖尿病的临床效果, 从方法学角度对疗效评价模型这一关键技术充分应用。

1 研究方案与方法

1.1 数据来源 选用 1990—2020 年中国知网

基金项目: 科技部“十三五”国家重点研发计划项目(2019YFC1709801)

第一作者: 窦智丽, 女, 博士, 助理研究员, 研究方向: 中医药数据挖掘与医学统计学分析

通信作者: 韩东燃, 男, 博士, 教授, 博士研究生导师, 研究方向: 中医药大数据挖掘、信息化建设、智能中医健康管理, E-mail: handongr@gmail.com

(CNKI)、万方数据(WANFANG DATA)、维普中文科技期刊数据库(VIP)建立的中医优势病种的6万余篇数据库中有关中医诊疗糖尿病相关的临床研究文献3654篇,统计文献中所列的中药、证型、穴位等诊疗信息。最终筛选出661篇与2型糖尿病的中医治疗相关的研究。将获取的文献应用NoteExpress 3.2文献管理软件查找重复文献,由2名硕士研究生作为研究者分别对其进行筛选,第3名研究者将解决最初2名研究者之间的争议。最终涉及中药处方的有312篇,包含中医症状、舌象、脉象、证候类型,此部分文献用于建立中医证候预测模型分析数据标准。

1.2 纳入标准 1)病种为2型糖尿病的研究;2)疾病证候诊断或疗效评价标准明确;3)干预措施为使用中药方剂;4)中药与其他治疗方法同时使用者,治疗组必须能体现出中医治疗的疗效;5)文献类型为随机对照试验(RCT)。

1.3 排除标准 1)重复发表;2)无法获取全文;3)综述、动物实验、个案报道、临床经验类;4)无法提取数据或资料不全;5)未明确诊断标准。

1.4 数据预处理 根据《中医诊断学》^[7]、《中医诊断学研究》^[8]、《中医内科学》^[9]、《证素辨证学》^[10]对文本提取的中医症状、中医证候信息进行标准化和归一化。包括21个证候类型、99个主观症状、13个中医脉象表现、23个中医舌象、232味中草药、20个中医穴位。

1.5 数据挖掘 选择Python语言编程,使用支持向量机^[11-12](SVM)、决策树^[13]、K近邻算法^[14]、逻辑回归^[15]、构建“症状-证候”预测模型,比较各个模型的准确度、精确度、F1、召回率、ROC曲线,选择最优模型作为预测模型。

1.6 模型评估 本研究中建模中随机将原始数据集分为80%训练集和20%测试集,将症状作为输入变量,将患者证候类型作为目标变量,使用训练集数据建立模型,再用测试集数据对该模型的性能进行评估。数据挖掘采用Python编辑器进行代码编辑,实现模型建立与评估。

模型评估参数^[16-17]选用准确度评估采用特征曲线(ROC)描述,曲线下面积(AUC)的大小代表灵敏度与特异度的高低。模型灵敏度和特异度采用ROC^[18]曲线描述,AUC ≥ 0.9 、但 <1.0 为优秀; $\geq$

0.7、但 <0.9 为良好; ≥ 0.6 、但 <0.7 为一般; <0.6 为差。

2 实验设计逻辑图

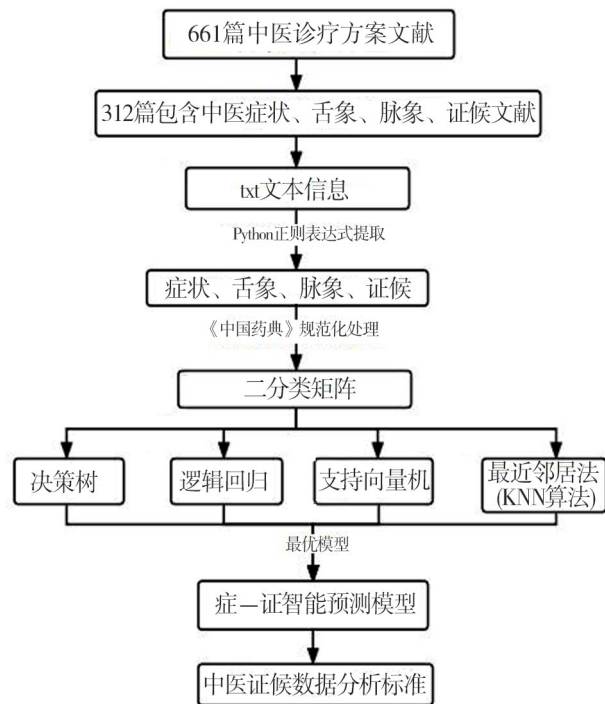


图1 中医症状-证候预测模型建立逻辑图

3 研究结果

3.1 临床资料统计结果 研究通过对321篇RCTs文章的诊疗信息统发现,有21个证候类型、99个主观症状、13个中医脉象表现、23个中医舌象、232味中草药、20个中医穴位,以下是对每个类别信息的分别统计。中医症状分布:口燥咽干(106)、口渴喜饮(101)、倦怠乏力(90)、神疲懒言(88)、大便干燥(84)、失眠(74)、消谷善饥(73)、心悸(68)、五心烦热(67)、心胸烦闷(56)、腰膝酸软(46)、自汗(43)、尿赤灼热(43)、盗汗(41)、小便频数(39)、头晕目眩(37)、消瘦(33)、肢体麻木(30)、小便色黄(29)、肥胖(28)、小便量多(27)、脘腹胀满(27)、食少纳呆(27)、烦躁易怒(24)、呕恶痰涎(20)、头重如裹(20)、胸胁胀闷(19)、耳鸣(18)、身体困重(15)、大便溏泻(15)、肢体水肿(14)、胸痛(14)、多饮(13)、肢体无力(12)、口苦(12)、视物模糊(12)、夜尿频数(11)、形寒肢冷(11)、健忘(11)、骨蒸潮热(11)、多尿(11)、肢体疼痛(10)、喜冷饮(9)、多梦(9)、面色少华(9)、口中黏腻(9)、口舌发紫(9)、小便短少(8)、肌肤甲错(8)、腰痛(7)、尿浊如脂膏(7)、手足

心热(7)、皮肤瘙痒(7)、大便溏结不调(7)、大便黏滞不爽(7)、尿糖增高(6)、面色晦暗(6)、早泄(5)、遗精(5)、疼痛部位固定(5)、面色黧黑(5)、口有秽臭(5)、口渴不欲饮(5)、易感冒(4)、阳痿(4)、胁痛(4)、喜叹息(4)、疼痛夜甚(4)、嗜睡(4)、皮肤干燥(4)、口淡无味(4)、恶心(4)、肢体酸痛(3)、肢体局部发暗(3)、腰膝酸冷(3)、背痛(3)、小便清长(3)、胃脘隐痛(3)。(见图 2-1)

中医证型分布:气阴两虚(125)、阴虚火旺(43)、脉络瘀阻(41)、阴阳两虚(22)、湿热困脾(17)、肺燥津伤(14)、痰瘀互结(13)、胃热炽盛(12)、痰湿内阻(12)、肾阴亏虚(11)、脾虚湿困(11)、脾虚痰湿(10)、肝肾阴虚(8)、脾肾阳虚(6)、肾虚血瘀(5)、气虚血瘀(5)、肝气郁结(4)、瘀热互结(3)、气虚(3)、脾肾阴虚(3)、脾肾气虚(3)、肝郁脾肾亏虚(3)、肝胃郁热(3)、肺胃热盛(3)、心脾两虚(2)、心肝热郁(2)、气虚痰浊(2)、脾阴虚(2)、肝脾阳虚(2)。(见图 2-2)

中医舌象分布:舌红少津(100)、舌苔薄(51)、少苔(44)、舌体胖大(38)、花剥苔(34)、苔薄黄

(31)、舌苔白(26)、舌质有瘀斑(26)、舌质紫暗(25)、舌苔厚腻(21)、边有齿痕(18)、舌质淡(18)、舌底脉络迂曲(16)、舌质红(14)、舌质晦暗(13)、苔黄腻(13)、舌淡红(12)。(见图 2-3)

细脉(111)、数脉(104)、弦脉(78)、脉无力(52)、滑脉(43)、沉脉(34)、涩脉(19)、弱脉(13)、虚脉(10)、濡脉(10)、有力(4)、洪脉(3)、缓脉(3)。(见图 2-4)

穴位及中药分布:脾俞穴(9)、肾俞穴(7)、胰俞穴(7)、足三里穴(7)、三阴交穴(6)、关元穴(5)、肺俞穴(4)、中脘穴(4)、气海穴(3)、神阙穴(3)、太溪穴(3)、胃俞穴(3)、丰隆穴(2)、肝俞穴(2)、膈俞穴(2)、内庭穴(2)、曲池穴(2)、三焦穴(2)、双侧内关穴(2)、太渊穴(2)。(见图 3-1)

黄芪(123)、生地黄(104)、麦冬(92)、葛根(88)、黄连(86)、丹参(79)、天花粉(79)、知母(72)、生山药(68)、茯苓(67)、枸杞子(47)、山茱萸(46)、玄参(46)、五味子(44)、生甘草(42)、生白术(41)、泽泻(39)、熟地黄(38)、党参(35)、太子参(34)、赤芍药(33)。(见图 3-2)

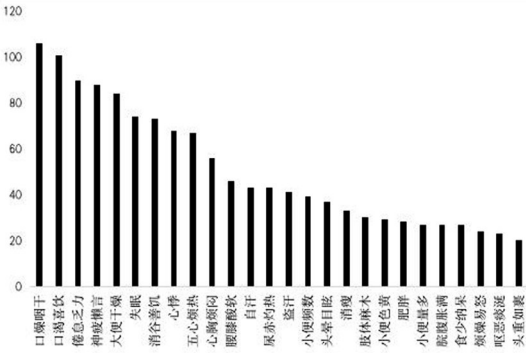


图 2-1 气阴两虚证候为最多见证候

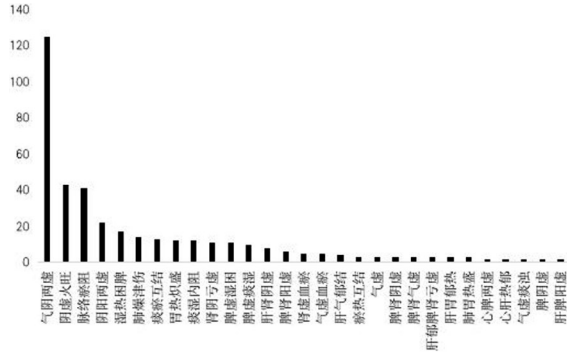


图 2-2 口燥咽干为最多见证状

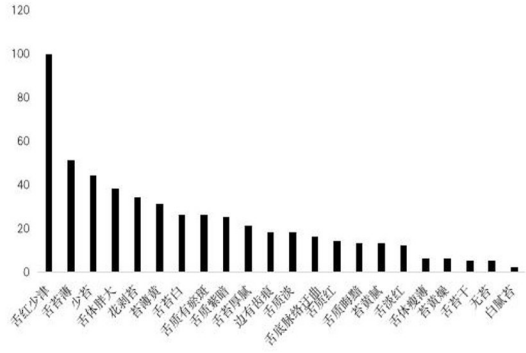


图 2-3 细脉为最多见证脉象

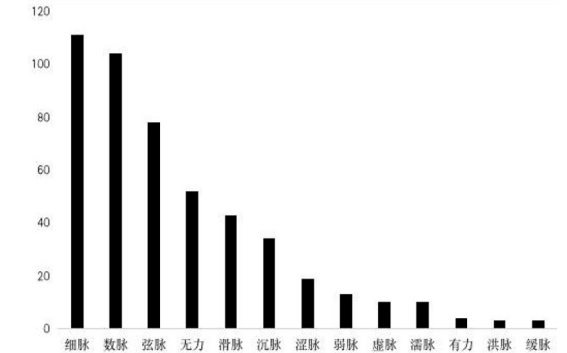


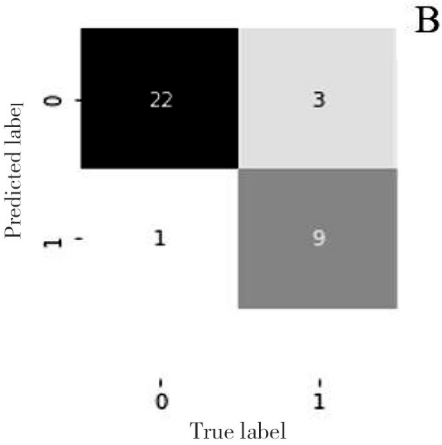
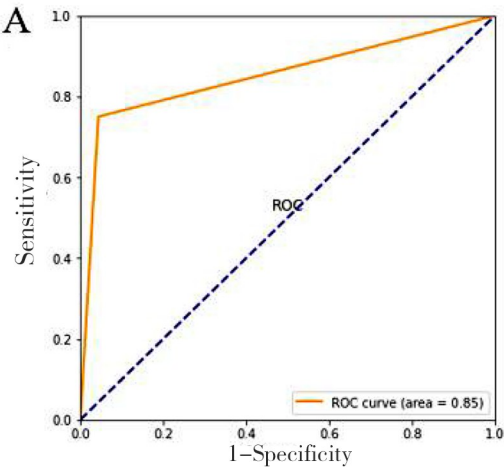
图 2-4 舌红少津为最多见舌象

图 2 中医症状证候信息分布

盛证与舌象: SVC: 0. 97; DT: 0. 88; KNN: 0. 97; LR: 0. 97。肾阴亏虚证与脉象: SVC: 0. 94; DT: 0. 91; KNN: 0. 94; LR: 0. 94。

气阴两虚证与脉象 SVC: 0. 78; DT: 0. 69; KNN: 0. 78; LR: 0. 69。阴虚火旺证与脉象模型的准确率: SVC: 0. 93; DT: 0. 93; KNN: 0. 90; LR: 0. 96。脉络瘀阻证与脉象: SVC: 0. 93; DT: 0. 93; KNN: 0. 90; LR:

0. 96。阴阳两虚证与脉象: SVC: 0. 93; DT: 0. 96; KNN: 0. 93; LR: 0. 93。湿热困脾证与脉象: SVC: 0. 93; DT: 0. 87; KNN: 0. 93; LR: 0. 87。肺燥津伤证与脉象: SVC: 0. 96; DT: 0. 93; KNN: 0. 96; LR: 0. 96。胃热炽盛证与脉象: SVC: 0. 96; DT: 0. 96; KNN: 0. 96; LR: 0. 96。肾阴亏虚证与脉象: SVC: 0. 93; DT: 0. 93; KNN: 0. 93; LR: 0. 93。



A:横轴 FPR: 1-TNR, 1-Specificity, FPR 越大, 预测正类中实际负类越多。纵轴 TPR: Sensitivity (正类覆盖率), TPR 越大, 预测正类中实际正类越多; Sensitivity、Specificity 越大效果越好; 本研究中 AUC 值等于 0. 85, 介于 0. 5 和 1 之间, 优于随机猜测分类, 有一定的预测价值。B: 横坐标为真实值标签, 纵坐标为预测值标签, 预测结果与真实值占比 30%。

图 4 ROC 曲线面积与混淆矩阵图

表 3 不同预测模型参数均值表

机器学习模型	准确率	精确度	F1	召回率	AUC 值
SVC	0. 94	0. 61	0. 92	0. 94	0. 72
DT	0. 90	0. 32	0. 91	0. 89	0. 76
LR	0. 92	0. 55	0. 90	0. 92	0. 63
KNN	0. 91	0. 56	0. 89	0. 91	0. 63

注: 支持向量机是预测糖尿病中医证候分类的最优算法。

4 讨 论

2 型糖尿病患者病程长, 与生活习惯和饮食习惯有密切关系^[19], 其并发症比原发病更严重, 如处理不当可危及生命。虽然临床上可以通过病史和临床表现诊断典型病例, 但是其中医证候类型较多, 临床 RCT 研究中中医证候类型判别依据多样, 缺乏统一标准, 常易无法科学解释中医理论。因此, 如何精准诊断甚至提前预测以及及时给予合理诊疗, 对控制患者血糖稳定和恢复胰岛器官功能十分必要^[20]。目前临床上仅依靠医师的主观经验评估患者的发病风险或提前诊断, 其精准度有限^[20]。因

此, 构建能够早期预测糖尿病患者发病风险的预测模型对于临床工作至关重要。本研究通过构建临床预测模型, 以期能够根据相关既往主观症状和四诊信息, 迅速判别人群类别, 进一步提高患病人群诊疗的准确性。

在大数据时代, 随着真实世界数据的增多, 运用机器学习对数据进行算法处理和开发程序化的预测模型成为临床提高诊疗效果的一种新方法^[20-22]。为了确保所建模型的有效性, 本文共建立了 4 个机器学习模型并进行了评估和比较, 结果显示决策树、支持向量机、逻辑回归算法的评估效果最佳, 这可能是因为所有证候预测模型都是通过相同的 99 个中医症状影响因素进行开发, 未能去除多余的数据, 从而使得这些预测模型保持较高的一致性。

综上所述, 相较于其他 3 个机器学习算法, 决策树是预测糖尿病中医证候分类的最佳算法。构建包含中医主观症状、舌象、脉象, 有助于根据早期症状表现划分中医证候类别。下一步研究需纳入

更大样本量进行临床验证试验,以期进一步改善并提高该模型的准确性。本文通过对文献数据建立预测模型,用准确率、精确度、召回率等参数评估模型的性能,一方面建立良好的2型糖尿病中医RCT文献数据的“症状-证候诊断模型”;另一方面,在提高临床电子病历数据二次挖掘的利用率的同时,为临床医生工作提供了理论支撑。决策树模型适用于分类、预测和规则提取,目前,决策树已经在中医病证诊断、辨证论治及预后等方面有了较好的应用^[23-24]。然而在中医优势病种高危因素预测、预后评估等方面应用不足,因此如何利用机器学习方法了解引发中医优势病种的高危因素,及时进行疾病风险评估以更好发挥中医“治未病”的优势是今后要努力的方向。同时,本文研究在证候模型建立过程中发现,在中医临床研究中,中医证候命名方法的多样导致了机器学习矩阵建立的困难,如多种证候类型名称在小样本模型矩阵分析中无法科学分类,这可能会影响分类模型的可靠性。因此,在下一步研究过程中证候类型采集时,我们将扩大样本量、注重证候类型名称的归一化与量化,提高预测准确性。

参考文献

- [1] HUANG JC, KO KM, SHU MH, et al. Application and comparison of several machine learning algorithms and their integration models in regression problems [J]. Neural Comput & Applic, 2020, 32: 5461-5469.
- [2] 宋超. 基于Cox回归模型和Logistic回归模型优选慢性心力衰竭中医临床疗效评价的替代指标[D]. 济南: 山东中医药大学, 2019.
- [3] 李思耐, 杨思雯, 夏非, 等. 基于中医证候要素的脓毒症预后因素多元Cox模型分析[J]. 中国中医急症, 2020, 29(7): 1135-1139.
- [4] 林于雄, 吴松鹰, 林菊珊. 卒中后抑郁中医易患因素的Logistic多元回归分析[J]. 福建中医药, 2017, 48(6): 40-41.
- [5] 赵小艳, 李继成, 段启宏. 最大似然估计法的教学设计与实践[J]. 大学数学, 2022, 38(4): 100-103.
- [6] 李纯净, 田闯, 李可可, 等. 现状数据下的贝叶斯比例风险模型的变量选择[J]. 数理统计与管理, 2022, 41(4): 679-688.
- [7] 袁肇凯, 王天芳. 中医诊断学[M]. 北京: 中国中医药出版社, 2007: 1-531.
- [8] 陈家旭. 中医诊断学研究[M]. 北京: 高等教育出版社, 2008: 1-302.
- [9] 张伯礼. 中医内科学[M]. 北京: 中国中医药出版社, 2016.
- [10] 朱文锋. 证素辨证学[M]. 北京: 人民卫生出版社, 2008.
- [11] 陈建新, 西广成, 王伟, 等. 数据挖掘分类算法在冠心病临床应用的比较[J]. 北京生物医学工程, 2008, 27(3): 249-252.
- [12] IMURA T, TODA H, IWAMOTO Y, et al. Comparison of supervised machine learning algorithms for classifying of home discharge possibility in convalescent stroke patients: A secondary analysis [J]. Stroke Cerebrovasc Dis, 2021, 30(10): 106011.
- [13] LEE H, LEE E J, HAM S, et al. Machine learning approach to identify stroke within 4.5 hours [J]. Stroke, 2020, 51(3): 860-866.
- [14] THAKKAR HK, LIAO WW, WU CY, et al. Predicting clinically significant motor function improvement after contemporary task-oriented interventions using machine learning approaches [J]. Neuroeng Rehabil, 2020, 17(1): 131.
- [15] WU Y, FANG Y. Stroke prediction with machine learning methods among older Chinese [J]. Int J Environ Res Public Health, 2020, 17(6): 1828.
- [16] CHUN M, CLARKER, CAIRNS BJ, et al. Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults [J]. J Am Med Inform Assoc, 2021, 28(8): 1719-1727.
- [17] YANG Y, ZHENG J, DU Z, et al. Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: Retrospective study [J]. JMIR Med Inform, 2021, 9(11): e30277.
- [18] MUSCHELLI J. ROC and AUC with a binary predictor: A potentially misleading metric [J]. J Classif, 2020, 37(3): 696-708.
- [19] 李想. 2型糖尿病患者的饮食护理研究进展[J]. 中国城乡企业卫生, 2022, 37(10): 19-21.
- [20] 冯强, 余永燕, 叶纪平. 基于机器学习法构建异位妊娠发生风险预测模型[J]. 广西医学, 2022, 44(1): 23-28.
- [21] 辜伟科, 王艳, 劳祥婷, 等. 基于机器学习的失眠症中医证素辨证模型的构建思路[J]. 电子技术与软件工程, 2020(21): 189-190.
- [22] 姚帅君, 闫敬来, 杜彩凤, 等. 基于机器学习的围绝经期综合征中医智能辨证模型构建研究[J]. 中国中医药信息杂志, 2022, 11(14): 1-8.
- [23] 樊佳赛, 杜艺菲, 许佳颖, 等. 基于中医证候和机器学习构建慢性心力衰竭中西医结合预后模型[J]. 基础医学与临床, 2022, 42(8): 1169-1175.
- [24] 马红丽, 徐长英, 杨新鸣. 决策树模型在中医药领域的应用现状[J]. 世界中医药, 2021, 16(17): 2648-2651, 2656.

(收稿日期: 2023-04-28)

[编辑: 韩晗]