聚知刊出版社
JZK publishing

基于机器学习的 II 型糖尿病早期诊断模型构建

吉莉

吉林省松原市吉林油田总医院，吉林省松原市，138001；

摘要：针对 II 型糖尿病早期诊断难题，本研究构建了基于机器学习的诊断模型。采用 NHANES 及中国东北地区体检数据集，经 SMOTE 和 MICE 方法预处理后，训练评估 8 种机器学习算法。XGBoost 模型表现最佳，在 NHANES 数据集上准确率 0.87、AUC 值 0.90，在中国东北地区体检数据集上准确率 0.85、AUC 值 0.91，泛化能力强。该模型助力 II 型糖尿病早期干预治疗，改善患者预后，为医疗诊断模型选择提供参考。

关键词：糖尿病；机器学习；早期诊断；分类模型；ROC 曲线

DOI:10.69979/3029-2808.24.12.019

1 引言

1.1 研究背景

糖尿病已成为全球性的公共卫生挑战。2021 年，全球约有 5.37 亿成年人患有糖尿病，占全球成人的 10.5%。到 2022 年，这一数字已超过 5.5 亿，预计 2023 年将接近甚至超过 6 亿人。近年来，我国的糖尿病患病率也在显著上升。根据最新的流行病学调查，全国成人糖尿病患病率已超过 10%，也就是说每 10 个成年人中就有 1 人患有糖尿病。我国的糖尿病患者数量全球领先，估计超过 1.1 亿人。这庞大的患者基数给我国的医疗体系带来了巨大的压力。糖尿病会导致多种严重并发症，如心血管疾病、肾病、视网膜病变和神经病变等，这些并发症不仅严重影响患者的生活质量，还极大地增加了医疗资源的消耗。

1.2 糖尿病的类型与特点

糖尿病主要有两种类型：I 型糖尿病和 II 型糖尿病。I 型糖尿病由自身免疫反应引发，免疫系统攻击胰岛 β 细胞，导致胰岛素分泌不足，通常发生在儿童和青少年时期，发病迅速，症状包括极度口渴、多尿、疲劳和体重下降。而 II 型糖尿病则是胰岛素抵抗和胰岛 β 细胞功能不全共同作用的结果，通常与不良的生活方式、肥胖以及遗传因素相关。II 型糖尿病是最常见的糖尿病类型，占糖尿病病例的 90%以上。其症状发展较为缓慢，早期可能没有明显症状，或仅表现为轻度疲劳、口渴和视力模糊等。多数患者确诊时已处于中晚期，因此，II 型糖尿病的早期诊断至关重要。

1.3 现有诊断方法的局限性

目前，医院对 II 型糖尿病的诊断主要依赖于空腹血糖（FPG）、糖化血红蛋白（HbA1c）、口服葡萄糖耐量试验（OGTT）等指标，并结合年龄、BMI、家族史等风险因素。然而，这些方法在预测准确性和泛化能力方面仍存在不足，且多数患者是在出现明显症状后才就医，未能有效解决 II 型糖尿病的早期诊断和预防问题。

1.4 研究目的

鉴于现有诊断方法的局限性，本研究旨在构建一个基于机器学习的 II 型糖尿病早期诊断模型，以提高诊断的准确性和早期筛查能力，从而为 II 型糖尿病的早期干预和治疗提供有力支持。

2 研究方法

2.1 数据来源

本研究的数据集涵盖了基础人口学和生理指标以及临床生化指标，共包含 17 个变量，但不包含任何血糖相关变量。研究采用两个独立的数据集进行模型训练和外部验证，分别是美国国家健康与营养调查（NHANES）数据集和中国东北地区体检数据集。预测变量包括性别、年龄、BMI、腰围、收缩压、舒张压、吸烟状态、饮酒状态，以及临床生化指标，如 ALT、AST、GGT、血清肌酐、总胆固醇、甘油三酯、高密度脂蛋白胆固醇、低密度脂蛋白胆固醇和非酒精性脂肪肝。

NHANES 数据集来源于美国国家健康与营养调查，时间范围为 2011 年至 2018 年，共包含 15,528 名参与者，其中 2,426 名（15.62%）被诊断为 II 型糖尿病。该数据集具有代表性，广泛用于流行病学研究和疾病风险评估。中国东北地区体检数据集则来源于 2022 年 11 月至 2023 年 3 月的体检人群，共 4,952 名参与者，其中 447 名（9.03%）被诊断为 II 型糖尿病。该数据集反映了中国东北地区的健康状况和疾病分布特点，适合用于模型的外部验证。

2.2 数据预处理

在数据预处理阶段，我们针对两个数据集的样本不均衡问题和缺失值问题进行了系统化处理。由于数据集中 II 型糖尿病患者数量较少，模型容易偏向多数类样本，从而影响其泛化能力和准确性。为此，我们采用了 SMOTE（Synthetic Minority Over-sampling Technique）技术对少数类样本进行过采样，通过在少数类样本之间生成合成样本来平衡数据分布。这种方法不仅增加了少数类样本的数量，还帮助模型更好地学习到少数类的特征信息，从而提高模型的泛化能力和准确性。同时，针对数据中的缺失值问题，我们采用了 MICE（Multiple Imputation by Chained Equations）多重插补法进行处理。与传统的缺失值处理方法（如删除样本或采用均值/中值填充）相比，MICE 方法能够更好地保留数据中的信息，避免因缺失值处理而导致的信息丢失。此外，我们还通过 3σ 原则对数据中的异常值进行了筛选和处理，以确保数据的质量。

2.3 模型构建与训练

在模型构建与训练阶段，我们系统地采用了 8 种广泛使用的机器学习算法，以评估它们在 II 型糖尿病早期诊断中的有效性。所选算法包括传统的逻辑回归、决策树，以及更先进的集成学习方法如随机森林、梯度提升决策树（GBDT）、支持向量机（SVM）、K 近邻（KNN）、LightGBM 和 XGBoos。这些算法因其在处理复杂数据模式和高维特征空间方面的不同优势而被选中。为了全面评估这些模型的性能，我们选择了一组综合的评估指标，包括准确率、精确度、召回率和 F1-score。这些指标不仅能够反映模型的整体分类性能，还能特别揭示模型在识别正类（即 II 型糖尿病患者）方面的敏感性和特异性。在 NHANES 数据集上，我们实施了 5 折交叉验证，这是一种强大的评估方法，可以减少模型评估过程中的方差，提高模型性能估计的稳定性和可靠性。此外，我们特别比较了应用和未应用 SMOTE（合成少数类过采样技术）的情况下模型的性能。通过这一详细的模型训练和评估过程，我们旨在识别出哪些算法在处理不平衡数据集时能够提供更稳定、更准确的预测结果。这对于 II 型糖尿病的早期诊断尤为重要，因为早期诊断可以显著改善患者的治疗结果和生活质量。此外，通过比较不同算法的性能，我们可以为未来的研究和实际应用提供有价值的见解，帮助选择最合适的机器学习模型来解决特定的医疗诊断问题。

3 结果与分析

3.1 模型性能评估

从评估结果（表 1）来看，XGBoost 模型在所有指标上均展现出卓越的性能，具体表现为准确率为 0.87、精确度为 0.86、召回率为 0.87、F1 分数为 0.87 以及 AUC 为 0.90，这些结果均显著优于其他模型。XGBoost 的高准确率和高召回率表明该模型在正确识别糖尿病患者方面具有很高的能力，而其高精确度和高 F1 分数则显示了模型在平衡精确度和召回率方面的优异表现。此外，XGBoost 的 AUC 值最高，进一步证实了其在区分糖尿病患者和非糖尿病患者方面的优越能力。相比之下，其他模型虽然也展现出一定的预测能力，但在某些指标上存在不足。例如，逻辑回归模型的召回率相对较低（0.59），这可能意味着该模型在识别糖尿病患者方面存在一定的局限性。而 CART 模型虽然召回率较高（0.75），但其精确度（0.63）和 F1 分数（0.76）相对较低，表明该模型可能在识别非糖尿病患者时存在一定的误判。随机森林和 GBDT 模型的性能也较为出色，但在 AUC 值上略低于 XGBoost，显示出在整体预测能力上仍有提升空间。LightGBM 模型的表现也值得关注，其准确率、精确度、召回率和 F1 分数均与 XGBoost 相近，但在 AUC 值上略低，显示出在区分能力上稍有不足。SVM 和 KNN 模型的表现则相对较弱，特别是在精确度和 F1 分数上，这可能与这些模型在处理不平衡数据集时的局限性有关。综上所述，XGBoost 模型在本研究中展现出最佳的 II 型糖尿病早期诊断性能，其在准确率、精确度、召回率、F1 分数和 AUC 等关键指标上的优异表现，使其成为处理此类不平衡数据集的理想选择。这一发现不仅为 II 型糖尿病的早期诊断提供了有力的工具，也为未来在类似医疗诊断问题中选择和优化机器学习模型提供了宝贵的参考。

表 1　8 种机器学习模型在 NHANES 数据上的评估结果

Model	Accuracy	Precision	Recall	F1-score	AUC
LR	0.71	0.77	0.59	0.67	0.76
CART	0.70	0.63	0.75	0.76	0.76
RF	0.82	0.84	0.77	0.81	0.87
GBDT	0.83	0.81	0.82	0.84	0.88
SVM	0.73	0.73	0.73	0.74	0.77
KNN	0.65	0.72	0.60	0.58	0.73
LightGBM	0.83	0.85	0.82	0.81	0.88
XGBoost	0.87	0.86	0.87	0.87	0.90

3.2 模型的泛化能力

聚知刊出版社
JZK publishing

中国东北地区体检数据集作为外部验证的结果（表2）与 NHANES 数据集的结果相似，显示出模型在不同数据集上具有较好的一致性。在这些模型中，XGBoost 模型在测试数据上的表现依然最为出色，其准确率为 0.85，精确度为 0.87，召回率为 0.84，F1 分数为 0.85，AUC 为 0.91。这些结果表明 XGBoost 模型不仅在训练集上表现良好，而且在未见过的测试数据上也能保持高预测性能，显示出很强的泛化能力。其他模型如 LightGBM 和 GBDT 也展现出较好的性能，准确率分别为 0.82 和 0.84，AUC 值分别为 0.89 和 0.89，说明这些基于树的模型在处理复杂数据模式时也具有较强的泛化能力。随机森林（RF）模型同样表现良好，AUC 值为 0.86，准确率为 0.81，进一步证实了集成学习方法在处理此类数据时的有效性。SVM 和逻辑回归（LR）模型的表现中规中矩，AUC 值分别为 0.78 和 0.77，准确率分别为 0.71 和 0.72，这可能与 SVM 对参数选择敏感以及 LR 模型相对简单有关。CART 和 KNN 模型在测试集上的表现相对较弱，特别是在召回率和 F1 分数上，这可能是因为 CART 模型容易过拟合，而 KNN 模型对噪声敏感，导致它们在新数据上的泛化能力不如其他模型。从模型的泛化能力上分析，XGBoost 模型表现很好，可用于构建糖尿病早期诊断模型。

表 2　8 种机器学习模型在中国外部验证数据上的评估结果

Model	Accuracy	Precision	Recall	F1-score	AUC
LR	0.72	0.78	0.60	0.68	0.77
CART	0.68	0.65	0.77	0.70	0.74
RF	0.81	0.85	0.78	0.81	0.86
GBDT	0.84	0.82	0.83	0.82	0.89
SVM	0.71	0.74	0.70	0.72	0.78
KNN	0.66	0.73	0.62	0.67	0.74
LightGBM	0.82	0.86	0.83	0.84	0.89
XGBoost	0.85	0.87	0.84	0.85	0.91

4 讨论

本研究构建了基于机器学习的 II 型糖尿病早期诊断模型，对比 8 种算法后，XGBoost 在准确率、精确度、召回率、F1-score 和 AUC 等指标上表现最佳，且在两个数据集上均展现出良好泛化能力，为 II 型糖尿病早期诊断提供了高效工具，有助于改善患者预后。传统诊断方法存在一定局限性，多数患者在出现明显症状后才就医，错过最佳干预时机。而本研究的机器学习方法，尤其是 XGBoost 模型，能够综合考虑多种因素，挖掘数据中的复杂模式和潜在关联，从而更准确地识别早期糖尿病患者。例如，XGBoost 模型在 NHANES 数据集上的准确率达到 0.87，AUC 达到 0.90，在中国东北地区体检数据集上的准确率为 0.85，AUC 为 0.91，显著优于传统单一指标诊断方法。这表明机器学习模型在处理多维度数据和复杂疾病模式方面具有明显优势，能够为 II 型糖尿病的早期诊断提供更有力支持。

XGBoost 模型在处理不平衡数据集方面表现出色，通过 SMOTE 技术对少数类样本进行过采样，有效平衡数据分布，提高诊断准确性和召回率。同时，该模型具有强大的特征选择能力，能够自动筛选出对糖尿病诊断最具影响力的变量，如年龄、BMI、腰围、甘油三酯等，这些变量与已知糖尿病风险因素一致，验证了模型的合理性和可靠性。此外，XGBoost 模型在不同来源的数据集上均表现出良好的泛化能力，说明其具有较强的适应性和稳定性，适用于不同地区和人群的糖尿病早期诊断。尽管 XGBoost 模型表现优异，但仍存在局限性。研究数据集虽涵盖多个变量，但可能遗漏部分糖尿病相关因素，如饮食习惯、运动水平、心理压力等生活方式因素，这些因素对糖尿病的发生发展有重要影响。此外，机器学习模型的构建和应用需要一定技术和资源支持，性能依赖于数据质量和预处理过程，数据偏差或不完整性可能影响模型准确性和可靠性。

未来研究可从以下方面拓展：扩大样本量和数据来源，纳入更多地区、种族和年龄段的人群数据，提高模型代表性和普适性；结合更多维度数据，如基因数据、生活方式数据等，构建更全面精准的诊断模型；优化模型，尝试不同特征选择方法和参数调整策略，提高性能和效率；开展多中心临床验证研究，将模型与实际临床诊断流程结合，评估其在真实医疗环境中的应用效果和价值，为临床推广提供有力证据。

参考文献

[1]廖涌.中国糖尿病的流行病学现状及展望[J].重庆医科大学学报,2015,40(7):4.
[2]文世林.糖尿病流行的全球趋势和国内现状[J].河南诊断与治疗杂志,2002,16(1):3.
[3]刘佳璇,李代伟,任李娟,等.面向不平衡医疗数据的多阶段混合特征选择算法[J].计算机工程与应用,2025,61(2):158-169.

作者简介：吉莉（1972 年 9 月-），女，汉族，吉林省松原市，副主任中药师，本科，吉林省松原市吉林油田总医院，研究方向，中药的提取与应用