文章编号：2095-2163(2023)01-0153-05 中图分类号：R587.1;TP181 文献标志码：A

基于机器学习的糖尿病预测及 SHAP 特征分析

李佳思
（上海工程技术大学 数理与统计学院，上海 201620）

摘 要：为了提升糖尿病诊断的准确率，找出糖尿病的影响因素，将机器学习算法应用于糖尿病诊断，建立糖尿病预测模型，为医生提供决策指导。实验使用 UCI 数据库中的皮马印第安人糖尿病数据集，首先对数据进行缺失值填充、异常值处理与标准化等预处理；在经过预处理后的数据集上建立单分类模型和集成学习模型，并通过 5 折交叉验证准确率和 AUC 值评估各模型的预测性能。结果表明，XGBoost 算法的预测效果最好，分类准确率达到 77.83%，AUC 值为 0.822。最后，引入 SHAP 模型增强模型的可解释性，归纳得出葡萄糖浓度、身体质量指数和年龄对糖尿病预测具有重要影响。

关键词：糖尿病；机器学习；疾病诊断；SHAP 特征分析

Diabetes prediction and SHAP feature analysis based on machine learning

LI Jiasi
（School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】 To improve the diagnostic accuracy of diabetes and identify the influencing factors of diabetes, the machine learning algorithm was applied to the diagnosis of diabetes, and a diabetes prediction model was established to provide decision guidance to physicians. The experiment uses the Pima Indian Diabetes Dataset in the UCI database. First, the data is preprocessed with missing value filling, abnormal value processing and standardization. Secondly, the single classification model and ensemble learning model are established on the preprocessed dataset, and the prediction performance of each model is evaluated by 5-fold cross-validation accuracy and AUC value. The results show that the prediction effect of XGBoost algorithm is the best, the classification accuracy is 77.83% and the AUC value is 0.822. Finally, the SHAP model is introduced to enhance the interpretability of the model. It is concluded that glucose, BMI and age have a significant impact on the prediction of diabetes.

【Key words】 Diabetes; machine learning; disease diagnosis; SHAP feature analysis

0 引 言

糖尿病是一种常见的慢性疾病，过去 30 年来，中国糖尿病的发病率急剧增加，国际糖尿病联盟协会(IDF)统计显示，2019 年中国糖尿病患者达到 1.16亿，位于世界第一位，成为威胁人们身体健康的一大问题[1-2]。糖尿病的典型症状为多食、多饮、多尿、体重减轻，根据病因不同可以分为 1 型糖尿病、2 型糖尿病、妊娠糖尿病和其它特殊类型糖尿病。其中，以 2 型糖尿病最为常见，约占糖尿病患者的 90%，主要由于遗传、环境等因素，使得胰岛素调节血糖能力下降[3]。糖尿病具有高达 100 多种并发症，如糖尿病心血管并发症、糖尿病足、神经等病变[4-5]。目前，糖尿病仍以饮食控制和药物治疗为主，给人们的生活带来极大不便，早发现早治疗可以减少由糖尿病并发症引起的死亡率。因此，寻找一种更加高效、准确的诊断方法具有重要意义。

随着互联网技术的快速发展与计算机性能的不断提升，机器学习和人工智能逐渐应用于各个领域，将机器学习方法与疾病诊断相结合，为医生提供辅助决策，建立智能化的疾病诊断系统成为热点研究方向。如今，机器学习已经形成成熟的理论体系，发展出许多算法模型，如：经典的决策树(Decision Tree)、逻辑回归(Logistic Regression)、支持向量机(Support Vector Machine)和 K-近邻(K-nearest Neighbor)等单学习器，以及随机森林(Random Forest)、AdaBoost 和 XGBoost 等集成学习算法，在模式识别、文本分类、医疗诊断等方面均得到了广泛应用。

目前，许多学者将机器学习算法应用于糖尿病的诊断。侯伟[6]等提出一种基于一维卷积神经网

基金项目：国家自然科学基金(62072296)。

作者简介：李佳思(1996-)，女，硕士研究生，主要研究方向：机器学习、生物统计。

收稿日期：2022-03-28

络的 DPN 糖尿病预测方法,使用天津医科大学代谢病医院的 898 个患者数据进行模型验证,分别建立了 1D-CNN 模型、支持向量机和 BP 神经网络预测模型。实验结果显示,1D-CNN 模型的效果最好,准确率达到 98.3%。乔瀚[7]等将基于多特征属性患者相似性的方法用于糖尿病诊断,通过聚类方法,分析了不同特征的相似性并进行分组,使用随机森林对分组结果进行拟合得到疾病预测结果。实验结果显示,所提方法相比其它方法更具有效性,预测准确率得到提升。章权[8]等将 Stacking 集成学习方法应用于糖尿病的诊断,在 UCI 数据库中的皮马印第安人糖尿病数据集中使用支持向量机、随机森林和人工神经网络作为基学习器进行 Stacking 集成。结果表明,融合后的模型比单个模型具有更好的分类效果,分类准确率为 92.2%。Sarwar[9]等将支持向量机、K-近邻、逻辑回归、随机森林等机器学习算法应用于糖尿病预测,实验使用 UCI 数据库中的皮马印第安人糖尿病数据集。结果显示,支持向量机和 K-近邻的分类准确率为 77%,优于其它分类算法,血糖浓度、体重指数和年龄是糖尿病预测的重要影响因素。于建宇[10]为了提升对妊娠期糖尿病的预测准确率,分别使用了 Xgboost、Lightgbm 和 Catboost 等集成学习算法,对天池比赛中的糖尿病数据进行预测,实验中使用交叉验证与网格搜索,确定模型的最佳参数。结果显示,Catboost 算法的预测准确率最高,达到了 76.5%。

机器学习算法在疾病预测方面的应用可以提升诊断效率和准确率,但是医疗数据通常具有大规模、高维度和异构性等特点,模型的预测准确率会受到多种因素的影响,同一模型在不同数据集中的表现也具有很大差异。虽然当前具有许多机器学习算法,但大多数算法都是黑箱模型,可解释性不强。针对于此,本文使用单分类模型(决策树、逻辑回归、支持向量机、K-近邻和朴素贝叶斯)和集成学习算法(随机森林、AdaBoost 和 XGBoost)用于糖尿病预测,对 UCI 数据库中的皮马印第安人糖尿病数据集进行建模,通过 5 折交叉验证和网格搜索获得模型的最佳参数,并使用 5 折交叉验证准确率和 AUC 值选出最佳预测模型,然后引入 SHAP 方法进行特征分析,找出影响糖尿病的主要因素。

1 方法研究

1.1 基于 SHAP 的糖尿病特征分析方法

由于大多数机器学习算法的可解释性较差,无法分析各个特征如何影响模型的预测结果,而这对于疾病诊断来说非常重要。因此,本文采用 SHAP 方法对模型中的糖尿病影响因素进行分析。

沙普利可加性模型解释方法(SHapley Additive exPlanation, SHAP)是 Lundberg 和 Lee[11]在 2017 年提出的,其基本思想是通过计算每个特征加入模型时的边际贡献达到解释模型的目的,可用于解释各种黑箱模型。该方法首先计算出每个特征的贡献值,可能为正向贡献,也可能为负向贡献,然后将所有特征的贡献值累加得到最终预测结果[12-14]。

假设第 i 个样本为 x_i, 第 i 个样本的第 j 个特征为 x_{ij}, 模型对该样本的预测值为 y_i, 整个模型的基线(即模型对所有样本预测值的均值)为 y_{base}, 则 SHAP 值服从以下等式:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) \quad (1)$$

式中,$f(x_{ij})$ 表示第 i 个样本中第 j 个特征对模型预测结果 y_i 的贡献值。当 $f(x_{ij}) > 0$ 时,表示该特征对模型预测结果具有正向影响,反之,该特征对模型预测结果具有负向影响[15-16]。SHAP 方法示意如图 1 所示。

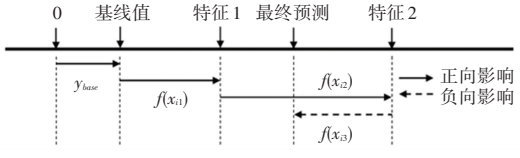

图 1 SHAP 示意图

Fig. 1 Schematic diagram of SHAP

1.2 模型评价方法

模型需要依靠合适的评价指标进行评估,对于二分类问题,常常使用混淆矩阵对模型的分类效果进行评价,混淆矩阵可以直观地反应出模型的分类结果,由混淆矩阵可以得到准确率(*Accuracy*)、精准率(*Precision*)、召回率(*Recall*) 和 F_1 值, 对模型性能进行全面评估。其中,准确率是分类正确的样本数量与总样本数量的比值;精准率表示预测为糖尿病的样本中实际也为糖尿病患者的比例;召回率表示真实为糖尿病的样本中,被预测为糖尿病的比例;*F*1 值为精准率与召回率的调和平均值。各评价指标的计算公式如下:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = \frac{2P \cdot R}{P + R} \tag{5}$$

其中，*TP* 表示真实为糖尿病同时也被预测为患糖尿病；*TN* 表示真实为未患糖尿病同时也被预测为未患病；*FN* 表示真实为患糖尿病而被预测为未患病；*FP* 表示真实未患病而被预测为患糖尿病。

受试者工作特征（Receiver Operating Characteristic，*ROC*）曲线，表示了模型在不同阈值下真正率（True Positive Rate，*TPR*）和假正率（False Positive Rate，*FPR*）之间的关系。*TPR* 和 *FPR* 的计算公式如下：

$$TPR = \frac{TP}{TP + FN} \tag{6}$$

$$FPR = \frac{FP}{TN + FP} \tag{7}$$

分别计算出不同阈值下的真正率和假正率，以假正率为横轴，真正率为纵轴，将各点在平面直角坐标系中绘制出来并进行连接得到 *ROC* 曲线（如图 2 所示），*ROC* 曲线越靠近左上角，表示模型效果越好。AUC 值是 *ROC* 曲线下的面积大小，其值越接近于 1，表示模型预测效果越好。本文通过 5 折交叉验证准确率和 AUC 值对各个模型的性能进行评价。

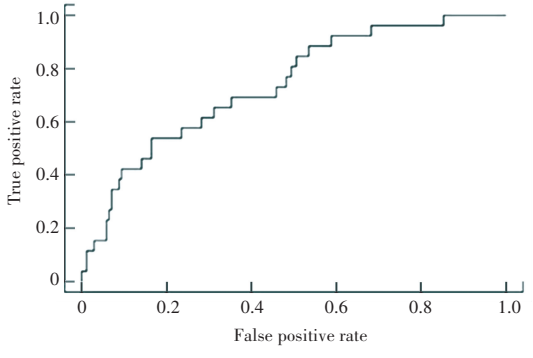

图 2　*ROC* 曲线示意图

Fig. 2　*ROC* curve diagram

2　实例验证

2.1　数据来源及预处理

本文实验过程主要包括数据预处理、建立预测模型、模型性能评估和影响因素分析 4 部分，实验流程如图 3 所示。实验使用 UCI 数据库中的皮马印第安人糖尿病数据集，该数据集共有 768 个样本，其中，500 例未患糖尿病，268 例患有糖尿病。数据集含有 8 个特征，各个特征的含义见表 1，其中，Outcome 为标签列，表示是否患有糖尿病。

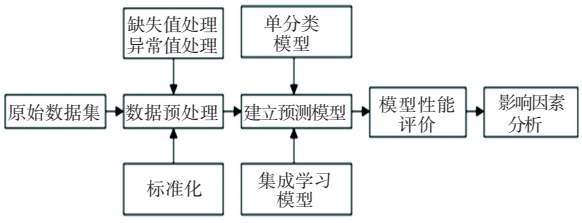

图 3　实验流程

Fig. 3　Experimental steps

表 1　糖尿病数据集特征描述

Tab. 1　Feature description of diabetes dataset

特征名称	含义
Pregnancies	怀孕次数
Glucose	葡萄糖浓度
BloodPressure	舒张压（mm Hg）
Skin Thickness	皮层厚度（mm）
Insulin	血清胰岛素（mu/ml）
BMI	身体质量指数
DiabetesPedigreeFunction	糖尿病家族影响指数
Age	年龄
Outcome	标签（0 表示未患病，1 表示患病）

在数据预处理前通常需要对数据进行可视化分析，其有利于分析数据自身特性，帮助提升模型的分类效果。通过 Python 中 Pandas 库的 describe 函数，对原始数据进行描述性统计，可知各特征间的均值和方差存在较大差异。特征 Glucose、BloodPressure、SkinThickness、Insulin 和 BMI 5 个特征存在不同程度的缺失情况，其中 Insulin 的缺失比例达到48.7%。由于缺失数据较多，若直接删除含有缺失值的样本，将会导致模型拟合能力下降，因此，对于上述 5 个特征使用中位数进行缺失值填充。然后，使用数字异常值方法处理异常值，计算各列特征的四分位数间距 IQR，将低于 $Q_1 - 1.5IQR$ 的值或高于 $Q_3 + 1.5IQR$ 的值作为异常值。其中，Q_1 为下四分位数，Q_3 为上四分位数，IQR 为上四分位数与下四分位数之差。最后，将数据进行标准化处理以消除特征之间的量纲差异。

2.2　实验结果与分析

本文对 UCI 数据库中的皮马印第安人糖尿病数据进行缺失值、异常值处理并进行标准化后，分别使用决策树、逻辑回归、支持向量机、K-近邻、朴素贝叶斯 5 种单分类器，以及随机森林、AdaBoost 和 XGBoost3 种集成学习算法对预处理后的数据集进行拟合。实验过程中，使用 70% 的数据作为训练集，30% 的数据作为测试集，采用网格搜索和交叉验

证寻找各个模型的最优参数,将 5 折交叉验证准确率作为确定模型最优参数的评估指标。将建立好的模型在测试集中进行测试,以 5 折交叉验证准确率和 AUC 值作为模型优劣的评价标准,各模型的预测结果,见表 2。

表 2 不同模型性能比较

Tab. 2 Performance comparison of different models

类型	模型	5-折交叉验证准确率/%	AUC
单模型	决策树	70.70	0.706
	逻辑回归	77.35	0.820
	支持向量机	77.09	0.814
	K-近邻	73.70	0.769
	朴素贝叶斯	74.75	0.794
	随机森林	76.70	0.782
集成模型	AdaBoost	76.18	0.774
	XGBoost	77.83	0.822

由表 2 可知,在皮马印第安人糖尿病数据集中,决策树模型的预测效果最差,5 折交叉验证准确率为 70.70%,AUC 值为 0.706;XGBoost 算法的分类效果最好,5 折交叉验证准确率达到了 77.83%,AUC 值为 0.822,高于其它分类模型。

为了进一步研究糖尿病的主要影响因素,提升分类模型的可解释性,引入 SHAP 方法对糖尿病数据集进行特征分析。图 4 为糖尿病数据集的特征重要性分析图。由图中可知,葡萄糖浓度、身体质量指数和年龄是影响糖尿病的重要因素,这与临床经验基本一致。

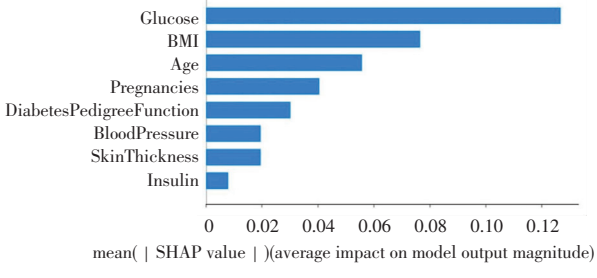

图 4 特征重要性分析

Fig. 4 Feature importance analysis

图 5 展示了各特征的 SHAP 值分布,从上到下按各个特征的重要性进行排序。横轴是模型的 SHAP 值,点的颜色表示特征值的大小。越接近红色,表示特征的值越大,越接近蓝色表示特征的值越小。SHAP 值为正,表示对模型预测为患糖尿病具有正向贡献;SHAP 值为负,表示对模型预测为患糖尿病具有负向贡献。由图 5 可知,Glucose 对模型的预测结果影响最大,且随着 Glucose 的值增大,会增加样本预测为患糖尿病的概率,即该特征对预测为患糖尿病具有正向影响。BMI 的趋势与 Glucose 类似,随着 BMI 值的增大,样本被判定为患糖尿病的概率增加,而 BMI 值的减小,则会增加模型判定样本不患糖尿病的概率,这与生活中肥胖患者更易患糖尿病相一致。

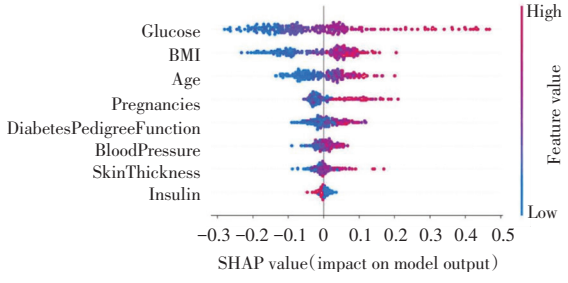

图 5 特征分析

Fig. 5 Feature analysis

SHAP 方法不仅可以在整体层面对预测模型的影响因素进行分析,还可以针对个体进行影响因素分析。分别选取一名被预测为糖尿病患者和一名被预测为非糖尿病患者进行个体影响因素分析。

图 6 为一名被预测为糖尿病患者的 SHAP 特征贡献图,红色部分表示对预测为糖尿病有正向影响,蓝色表示对预测为患糖尿病有负向影响。由图 6 可以解释这名患者被预测为患糖尿病的原因是葡萄糖浓度较高、身体质量指数较高、年龄较大等。

图 7 为一名被预测为非糖尿病患者的 SHAP 特征贡献图。由图 7 可以解释其被预测为非糖尿病的原因为葡萄糖浓度较低、年龄较小、身体质量指数较低等。

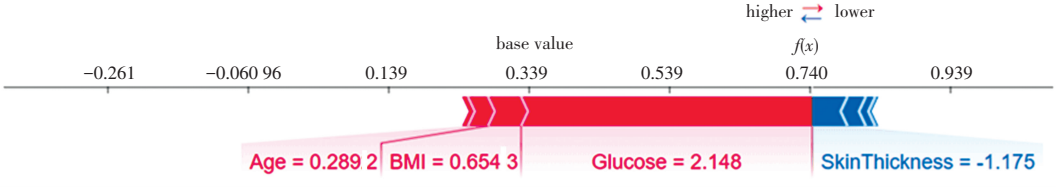

图 6 预测为糖尿病患者的 SHAP 解释示例

Fig. 6 Example of SHAP interpretation for predicting diabetes

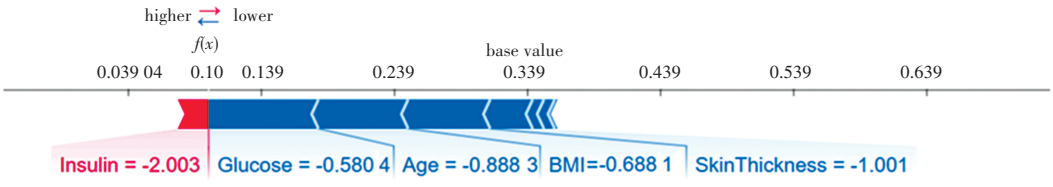

图 7　预测为非糖尿病患者的 SHAP 解释示例

Fig. 7　Example of SHAP interpretation for predicting Non-diabetes

3　结束语

糖尿病作为中国发病率较高的慢性疾病之一，给患者带来许多不便和负担。因此，将机器学习算法应用到糖尿病预测，对提升糖尿病诊断效率和准确率，了解糖尿病的发病机制具有重要意义。本文基于机器学习算法，使用 UCI 数据库中的皮马印第安人糖尿病数据集构建了糖尿病预测的单分类模型（决策树、逻辑回归、支持向量机、K-近邻和朴素贝叶斯）和集成学习模型（随机森林、AdaBoost 和 XGBoost），通过 5 折交叉验证准确率和 AUC 值对各个模型进行性能评估。实验结果表明，XGBoost 算法的预测效果最好，5 折交叉验证准确率为77.83%，AUC 值为 0.822。然后，引入了 SHAP 方法增强模型的可解释性，得到引起糖尿病的主要因素为葡萄糖浓度、身体质量指数、年龄、怀孕次数等，医生在进行决策时可以多关注这些特征。在未来，可以将更多与糖尿病相关的因素考虑到模型中，进一步提升糖尿病诊断准确率。

参考文献

[1] 邵韦巧. 机器学习分类算法对糖尿病数据应用研究[D]. 兰州：兰州大学，2021.
[2] 张玉玺，贺松，尤思梦. 集成学习在糖尿病预测中的应用[J]. 智能计算机与应用，2019，9(5)：176-179.
[3] 曲凯扬. 基于机器学习的糖尿病预测模型[D]. 天津：天津大学，2019.
[4] 郑尔昌，邹金串，薛成斌，等. 糖尿病联合并发症发病风险计算与预测[J]. 华侨大学学报(自然科学版)，2022，43(4)：498-510.
[5] HASAN M K，ALAM M A，DAS D，et al. Diabetes prediction using ensembling of different machine learning classifiers[J]. IEEE Access，2020，8：76516-76531.
[6] 侯伟，赵耕，刘玉良，等.基于一维卷积神经网络的糖尿病周围神经病变预测模型研究[J]. 中国医学物理学杂志，2022，39(1)：127-132.
[7] 乔瀚，容芷君，许莹，等.基于多特征属性相似的糖尿病早期预测方法[J]. 科学技术与工程，2021，21(36)：15497-15502.
[8] 章权，周梁琦，邹琪，等.基于 Stacking 的糖尿病预测方法研究[J]. 智能计算机与应用，2020，10(2)：107-110.
[9] SARWAR M A，KAMAL N，HAMID W，et al. Prediction of diabetes using machine learning algorithms in healthcare[C]// 2018 24th international conference on automation and computing (ICAC). IEEE，2018：1-6.
[10] 于建宇.基于集成学习算法的妊娠期糖尿病预测模型研究[D]. 哈尔滨：哈尔滨工业大学，2019.
[11] LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions [C]// Advances in neural information processing systems，2017：4765-4774.
[12] 徐良辰，郭崇慧.基于集成学习的胃癌生存预测模型研究[J]. 数据分析与知识发现，2021，5(8)：86-99.
[13] 王鑫，廖彬，李敏，等. 融合 LightGBM 与 SHAP 的糖尿病预测及其特征分析方法[J]. 小型微型计算机系统，2022，43(9)：1877-1885.
[14] 陈小昆，左航旭，廖彬，等. 融合 XGBoost 与 SHAP 的冠心病预测及其特征分析模型[J]. 计算机应用研究，2022，39(6)：1796-1804.
[15] 李超，陈功，储文强，等. 基于改进 SHAP 的城市供水管网爆管主影响因素研究[J]. 科技通报，2021，37(1)：79-84.
[16] 张继婕，覃庆洪，刘雪萍，等.基于集成学习的乳腺癌生存预测研究[J]. 广西科技大学学报，2022，33(1)：101-109.

(上接第 152 页)

[10] Sonali Pradhan，Mitrabinda Ray. Transition coverage based test case generation from state chart diagram[J]. Journal of King Saud University-Computer and Information Sciences，2019：1.
[11] 冯杰. Web Service API 自动化测试与持续集成系统的设计与开发[D]. 上海：上海交通大学，2016.
[12] 刘明珠，丁亦楠，郑云非. XML 数据公交信息查询优化算法及实现[J]. 哈尔滨理工大学学报，2015，20(2)：85.
[13] Trinh Nguyen，Myungsik Yoo. A YAML-Based Profiler for Designing SFC Tests[J]. The Journal of Korean Institute of Communications and Information Sciences，2019，44(12)：2303.
[14] M. I. P. Salas，E. Martins. Security Testing Methodology for Vulnerabilities Detection of XSS in Web Services and WS-Security[J]. Electronic Notes in Theoretical Computer Science，2014，302：133.
[15] Raghu Ramakrishnana，Arvinder Kaur. Performance evaluation of web service response time probability distribution models for business process cycle time simulation[J]. The Journal of Systems and Software，2019，161：1.