糖尿病风险预测中的机器学习模型对比实验

日期：2025.10.13

摘要

本研究旨在基于 Kaggle 平台提供的“Diabetes Prediction Dataset”数据集，构建糖尿病预测模型，并比较两种常用机器学习算法——随机森林（Random Forest）与逻辑回归（Logistic Regression)在该任务中的表现差异。通过对原始数据进行数据清洗、异常值处理、特征编码与可视化分析，初步完成变量筛选，并对目标变量（是否患糖尿病）进行单变量分析与多变量相关性验证。

在模型构建阶段，分别以相同特征输入建立了逻辑回归模型与随机森林模型，比较其预测准确率、召回率、混淆矩阵及 ROC-AUC 等性能指标。在进一步的模型优化中，研究尝试通过调整阈值提升模型在召回率与准确率之间的平衡，确定最优分类界限。

结果表明，在数据量适中的前提下，随机森林模型在预测准确率和稳定性方面优于逻辑回归模型，而逻辑回归则在模型解释性方面具备优势。最终的研究为糖尿病早期风险筛查提供了一种高效、可扩展的建模思路。

本研究暂未进行跨数据集迁移验证，未来可考虑引入更多样化的数据源，进一步检验模型的泛化能力与实际应用价值。

关键词：糖尿病预测 机器学习 随机森林 逻辑回归 多模型对比 公众健康数据

Abstract

This study aims to develop a predictive model for diabetes based on the "Diabetes Prediction Dataset" from Kaggle, and to compare the performance of two machine learning algorithms — Random Forest and Logistic Regression. After conducting thorough data preprocessing, including outlier removal, feature encoding, and exploratory data analysis, relevant features were selected for model construction.

Both models were trained using the same set of features, and their performance was evaluated using metrics such as accuracy, recall, confusion matrix, and ROC-AUC score. Furthermore, the study experimented with threshold adjustment to optimize the trade-off between precision and recall, identifying an optimal classification threshold.

The results suggest that Random Forest outperforms Logistic Regression in terms of predictive accuracy and robustness, while Logistic Regression offers better interpretability. The findings support the effectiveness of machine learning approaches in early diabetes risk screening and demonstrate the feasibility of applying such models to public health datasets.

Although this study does not include cross-dataset generalization due to data limitations, future work may incorporate external datasets to evaluate the generalizability and practical applicability of the proposed models.

Keywords:Diabetes prediction, Machine learning, Random Forest, Logistic Regression, Model comparison, Public health data

目录

一、引言

（一）糖尿病背景介绍

糖尿病是一项全球健康挑战，2021 年影响了 5.29 亿人，预计到 2050 年将影响全球 1310 亿人。中国是世界上糖尿病人口最多的国家，有超过 1.18 亿糖尿病患者，约占全球糖尿病总数的 22%。1 然而，1980 年中国糖尿病患病率一度低于 1%。2 自 1970 年代国家历史性改革以来，中国经济取得了巨大的增长。随着中国从低收入国家转变为中高收入国家，在现代化和人口老龄化的推动下，环境、生活方式和医疗保健的重大转变导致了中国人口糖尿病的快速增长。

（二）研究目的

随着糖尿病发病率的持续上升，尤其在发展中国家中对公共卫生系统造成显著压力，如何有效预测和早期干预糖尿病已成为医学与数据科学交叉研究的重要方向。传统医疗诊断手段存在成本高、周期长等问题，因此基于机器学习的预测模型日益受到关注。

本文旨在基于真实公开医疗数据，比较逻辑回归与随机森林两种经典分类算法在糖尿病预测任务中的建模表现，通过对模型的特征选择、训练效果、参数优化与泛化能力等方面进行系统分析，以期为糖尿病的智能预测提供一种可行的数据驱动方法论，同时探索不同算法在医疗场景中的适用性差异。

（三）数据来源说明

本研究所用数据集为公开医疗数据集，名为《Diabetes Prediction Dataset》，来源于数据科学平台 Kaggle（链接：https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset）。该数据集由 Mohammed Mustafa 上传发布，包含 100,000 条关于糖尿病相关因素的个人健康记录，字段涵盖性别、年龄、是否患有高血压、心脏病史、吸烟史、BMI（身体质量指数）、糖化血红蛋白（HbA1c）水平、血糖水平（血糖值）以及是否患有糖尿病（作为标签字段）。各特征经过预处理，数据结构完整，无明显缺失值，适合用于分类模型的训练与分析。

该数据集可用于构建机器学习模型，根据患者的病史和人口统计信息预测患者的糖尿病。这对于医疗保健专业人员识别可能有患糖尿病风险的患者和制定个

性化治疗计划非常有用。此外，研究人员还可以使用该数据集来探索各种医学和人口因素之间的关系以及患糖尿病的可能性。

选择该数据集的原因在于其具备如下优势：一是样本量充足，包含 10 万条个人健康记录，可满足机器学习建模对数据量的基本要求，增强模型的泛化能力；二是特征维度合理，既包含结构化的数值型特征（如 BMI、HbA1c、血糖水平等），也涵盖二分类变量（如性别、高血压、心脏病史等），适合对比不同算法在多类型特征下的表现；三是标签字段（是否患有糖尿病）为清晰的二分类任务，便于模型训练与评价。此外，数据整体缺失率极低，预处理成本较小，能将更多精力集中于模型构建与性能优化上。因此，该数据集具备良好的研究基础与实用性，适合作为本研究的主要分析对象。

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

图 1 数据一览表

二、数据预处理与特征工程

（一）清洗流程说明

在数据预处理过程中，首先对各字段进行了初步的探索性分析与清洗处理。通过查看数据维度，缺失值，字段类型，字段分布等数据，为进一步数据分析做准备。

缺失值数量:

gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

dtype: int64

图 2 缺失值

字段类型:

gender	object
age	float64
hypertension	int64
heart_disease	int64
smoking_history	object
bmi	float64
HbA1c_level	float64
blood_glucose_level	int64
diabetes	int64

dtype: object

图 3 字段类型

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

图 4 数值型字段分布

性别字段中存在三类标签（Male、Female、Other），其中“Other”类仅占总样本的极小比例（18 条记录），为确保模型训练的稳定性与分类边界的清晰性，选择将该类记录删除；同时，为后续建模方便，将性别字段转换为 0（Female）与 1（Male）两类，形成二值型变量。

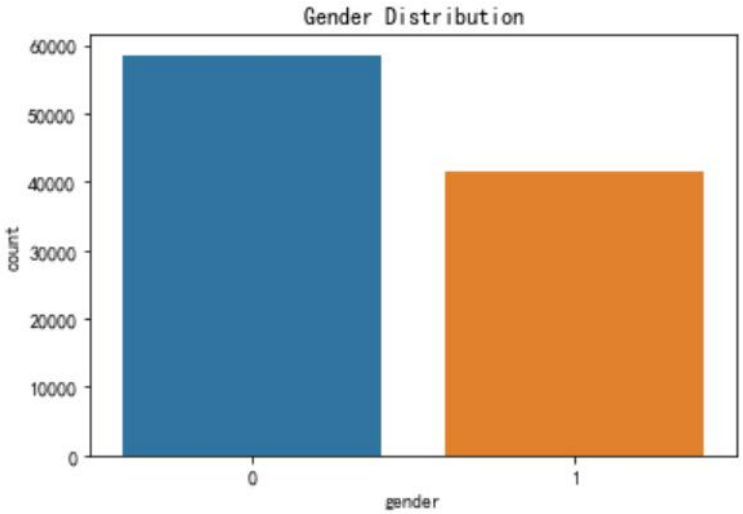

图 5 性别数据可视化结果

对于吸烟史字段（smoking_history），原始数据中共包含六类标签，包括“Never”、“Former”、“Not Current”、“Current”、“Ever”与“No Info”。其中“No Info”作为信息缺失项，占总样本近 30%，无法准确归类，因此暂不删除，而是在编码过程中保留。其余五类数据根据吸烟状态的历史进行归类重组，最终将吸烟变量转化为四类：0（Never）、1（Past Smoker，包括 Former、Not Current、Ever）、2（Current Smoker）、3（No Info），并转化为数值型变量以便模型识别。在后续建模过程中，为提高建模准确率，选择删除 No Info 数据。

```
No Info         0.358164
never           0.350983
former          0.093537
current         0.092877
not current     0.064402
ever            0.040037
Name: smoking_history, dtype: float64
```

```
3    0.358164
0    0.350983
1    0.197976
2    0.092877
Name: smoking_history, dtype: float64
```

图 6 吸烟数据归类重组

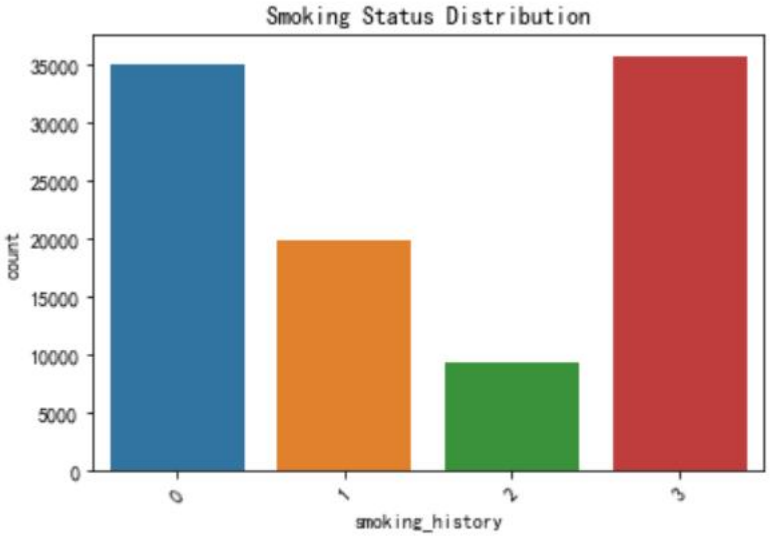

图 7 吸烟数据分类可视化

在数值型变量中，对年龄数据进行了可视化处理。首先对身体质量指数（Body Mass Index，BMI）进行了箱型图分析，发现存在明显的极端异常值（如 BMI 高于 90），通过查阅医学文献与临床常识判定其超出正常人体范围，可能为录入错误或特例个案，因此选择将超过合理上限的异常值予以剔除，以提高模型的泛化能力与预测稳定性。

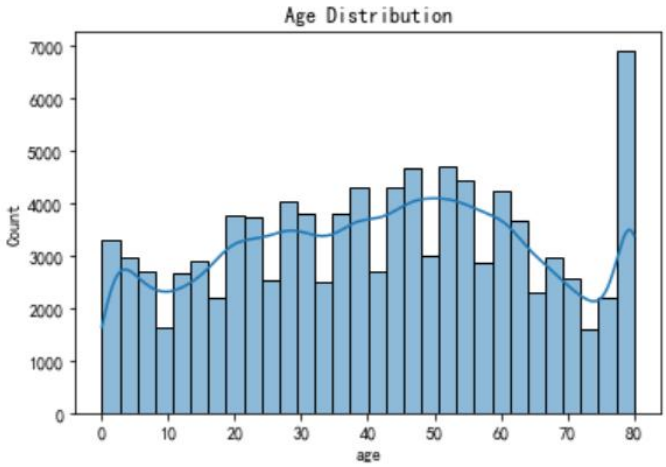

图 8 年龄数据可视化

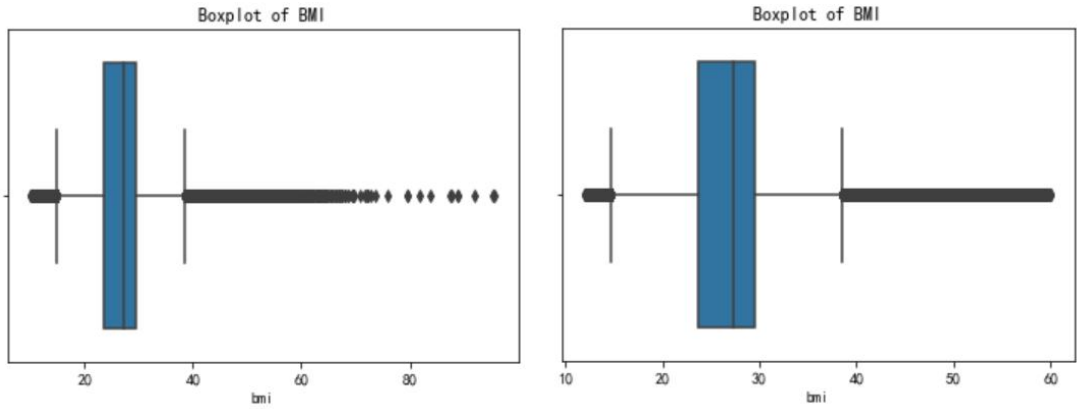

图 9 BMI 数据处理前后对比

糖化血红蛋白（HbA1c_level）与血糖（Blood_Glucose_Level）两个关键生理指标也进行了箱型图和分布分析。尽管存在少量极高或极低的数值，但数量极少，且通过对比医学标准后发现仍属于可能的临床边缘值，具有一定的医学合理性，因此保留全部记录，未作删除处理。

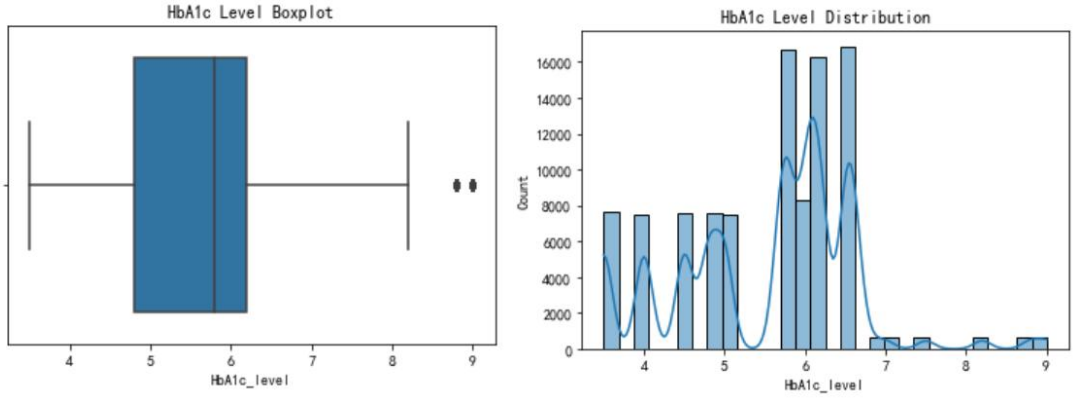

图 10　HbA1c level 数据可视化

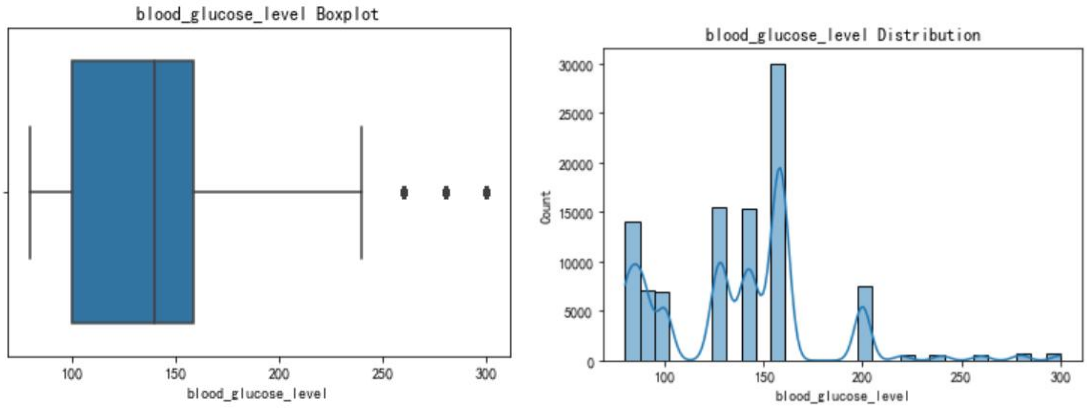

图 11 blood_glucose_level 数据可视化

（二）显著性分析

在完成初步清洗后，本研究对各变量与目标变量（是否患有糖尿病）之间的统计显著性进行了逐项分析，以判断其对建模预测的潜在贡献。

对于连续型变量，包括年龄（Age）、身体质量指数（BMI）、糖化血红蛋白（HbA1c_Level）与血糖水平（Blood_Glucose_Level），分别采用绘制箱线图，密度图，直方图等，分别进行 Mann-Whitney U 检验（即 Wilcoxon 秩和检验）对患病组与未患病组之间的分布差异进行非参数统计分析。结果显示，上述四个变量在两组间的差异均具有极强的统计显著性，p 值均趋近于 0（$p < 0.001$），表明这些连续变量在糖尿病判别中具有重要作用，后续建模阶段将予以保留。

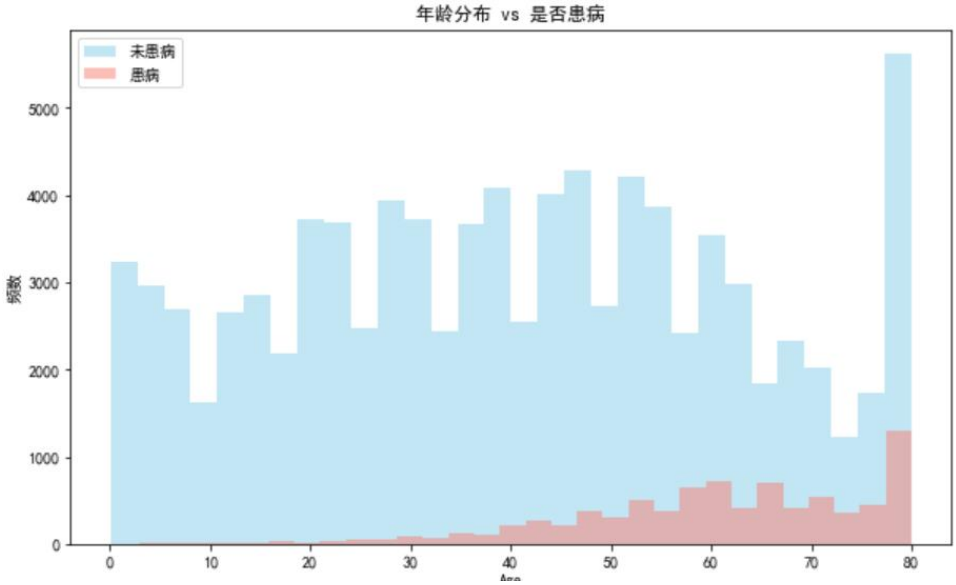

图 12 年龄分布与糖尿病患病情况对比条形图

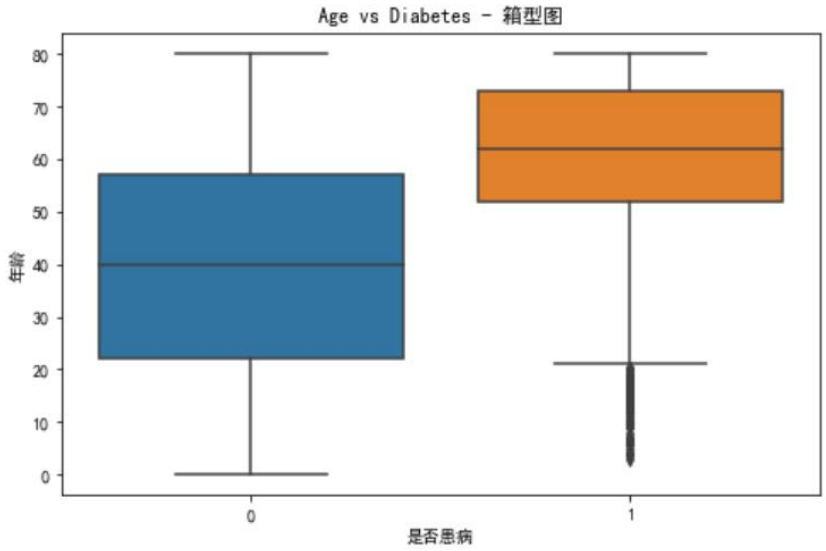

图 13 年龄分布与糖尿病患病情况对对比箱线图

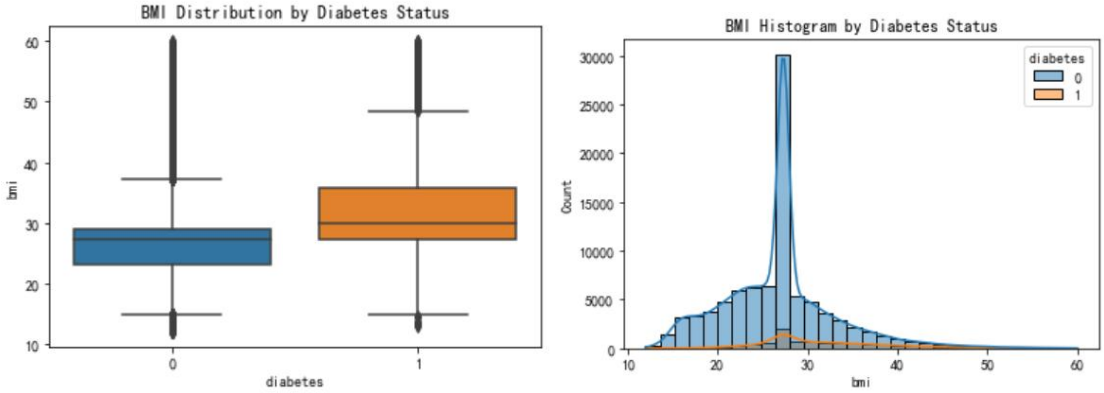

图 14 BMI 数据显著性分析可视化

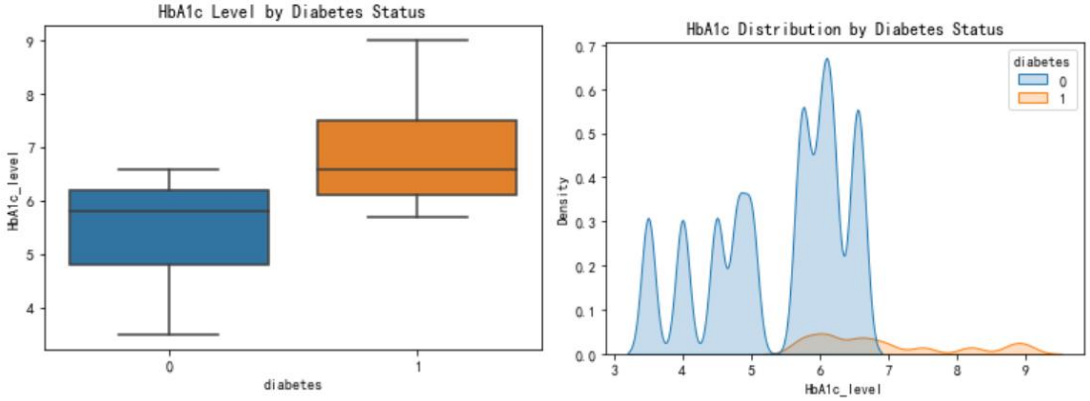

图 15 HbA1c level 数据显著性分析可视化

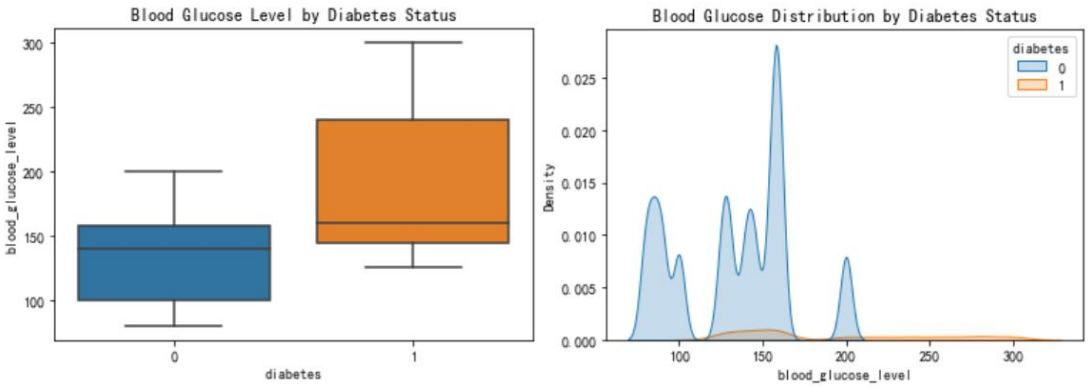

图 16 blood_glucose_level 数据显著性分析可视化

对于分类变量，包括性别（Gender）、是否患有高血压（Hypertension）、是否患有心脏病（Heart Disease）以及吸烟状态（Smoking_history），本研究首先绘制了频数直方图与交叉列联表，以观察不同变量类别在患病与未患病群体中的分布差异，随后进一步采用卡方检验（Chi-square Test）评估其与糖尿病发病情况之间的关联强度。分析结果显示，以上四项分类变量均在统计上呈现显著性差异，p 值均远小于显著性水平（$p < 0.001$），说明这些变量与糖尿病的发生具有密切关联，因而将在后续建模阶段中予以保留作为关键预测因子。

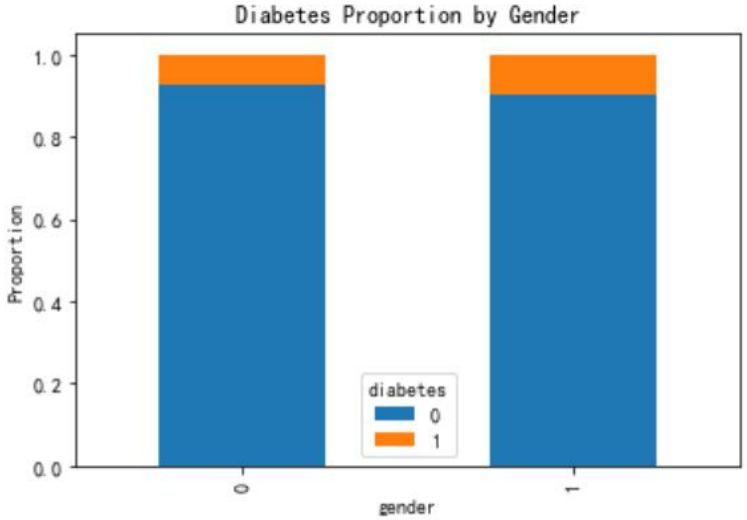

图 17 性别数据显著性分析可视化

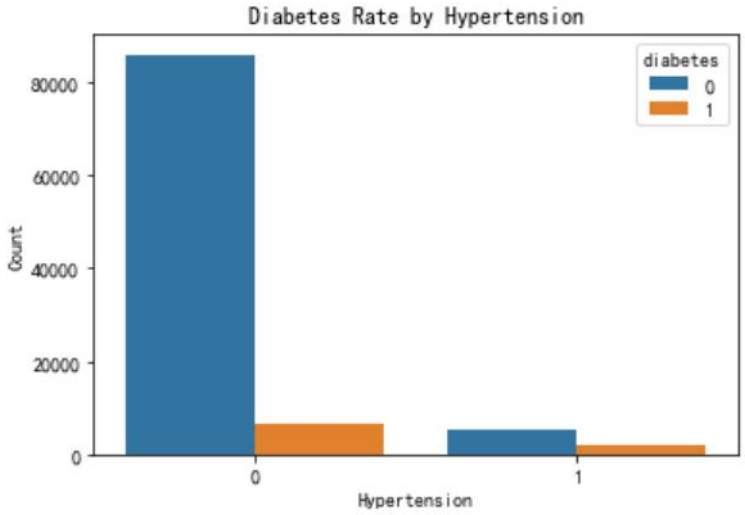

图 18 高血压数据显著性分析可视化

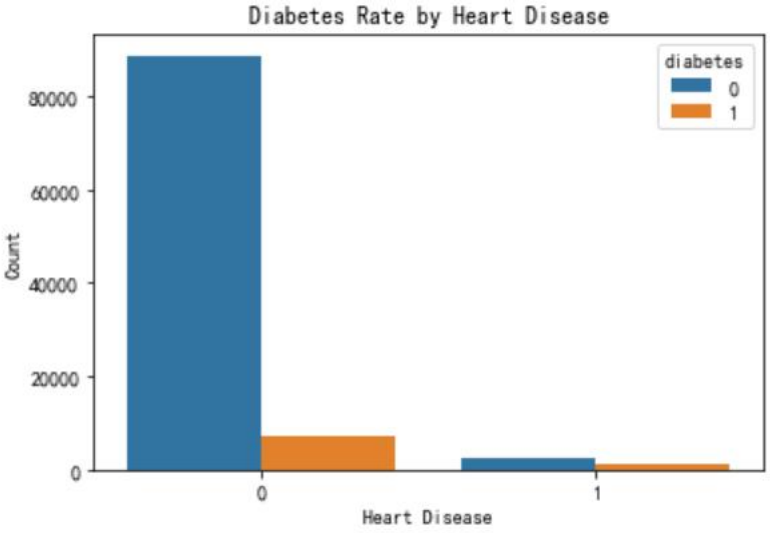

图 19 心脏病数据显著性分析可视化

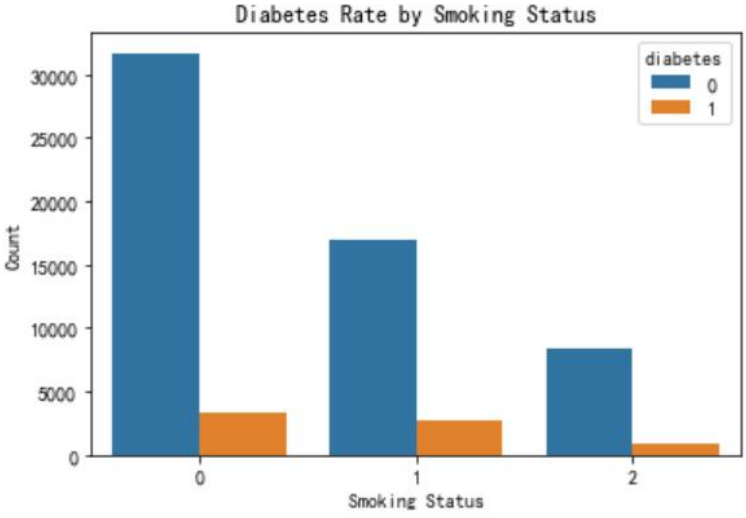

图 20 吸烟状态数据显著性分析可视化

（三）特征间多重共线性分析

在完成变量的显著性筛选后，本研究进一步对各个特征之间的相关性进行了探讨，以确保输入模型的变量之间不存在严重的多重共线性问题，从而避免模型估计不稳定、解释能力下降等风险。

首先，通过计算皮尔森相关系数（Pearson Correlation Coefficient）并绘制热力图，对所有数值型变量间的相关性进行可视化分析。结果显示，大多数特征之间的相关系数均在 ±0.3 范围以内，最大相关系数为 0.45，未发现高度相关的变量组合，初步判断不存在严重共线性问题。

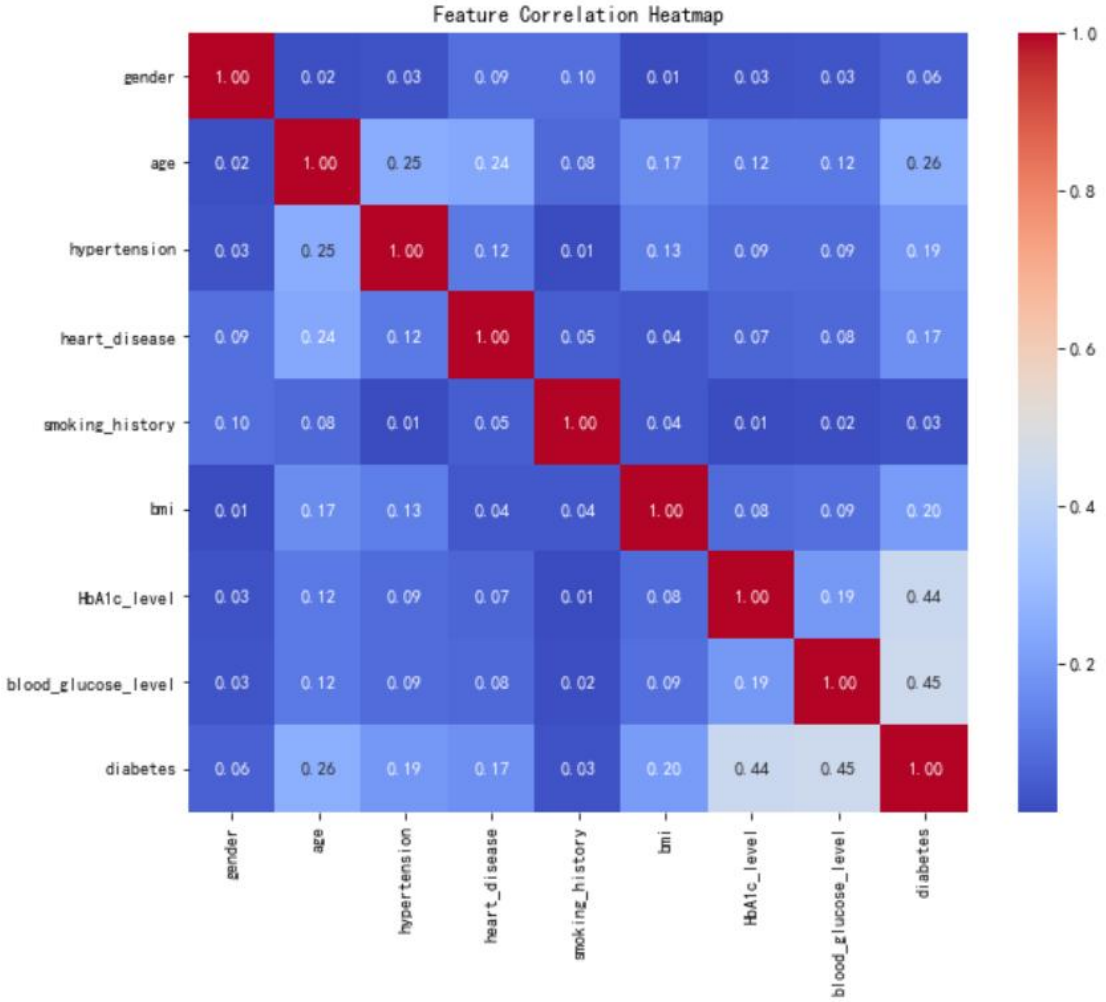

图 21 特征相关性分析热力图

为进一步验证，本研究采用方差膨胀因子（Variance Inflation Factor, VIF）对变量间的多重共线性进行了定量评估。分析中将目标变量剔除，仅对特征变量进行处理，并添加常数项。结果显示，所有变量的 VIF 值均接近于 1，表明各

特征变量之间不存在显著的线性依赖关系，完全可以同时纳入模型进行后续训练。

	feature	VIF
0	const	50.467114
1	gender	1.018341
2	age	1.158112
3	hypertension	1.086645
4	heart_disease	1.077095
5	smoking_history	1.016656
6	bmi	1.045685
7	HbA1c_level	1.055449
8	blood_glucose_level	1.057784

图 22 VIF 评估结果

三、模型建立与评估

（一）随机森林模型构建与分析

在完成数据清洗与特征筛选后，本文首先采用随机森林（Random Forest）方法对糖尿病预测问题进行建模。随机森林作为一种集成学习算法，能够有效地处理非线性特征关系，且对异常值和高维数据具有较强的鲁棒性，因此在医学预测类任务中具有广泛的应用价值。

模型构建流程：

1. 训练集与测试集划分

使用 sklearn 库中的 train_test_split 函数，将处理后的数据集按 8:2 比例划分为训练集与测试集，保证建模过程中对模型泛化能力的合理考察。

2. 模型训练

调用 RandomForestClassifier 类，采用默认参数（n_estimators=100）进行模型初步训练，后续进行调参优化。使用训练集进行拟合，记录各特征在模型中的重要性评分。

3. 模型预测与评估

在测试集上进行预测，并通过混淆矩阵（Confusion Matrix）、准确率（Accuracy）、查全率（Recall）、查准率（Precision）以及 F1 分数等指标对模型表现进行综合评估。

4. 分类阈值调整

为进一步提升模型对“患病”类别的识别能力，改进模型。又加入惩罚项，绘制了 Precision-Recall 曲线，并选取了阈值为 0.31 作为最优分类标准。该阈值下，模型在保证整体准确率的同时，提高了对阳性样本的召回能力。

分类报告:

	precision	recall	f1-score	support
0	0.96	1.00	0.98	11433
1	0.95	0.68	0.79	1381
accuracy			0.96	12814
macro avg	0.96	0.84	0.88	12814
weighted avg	0.96	0.96	0.96	12814

图 23 随机森林第一次拟合结果

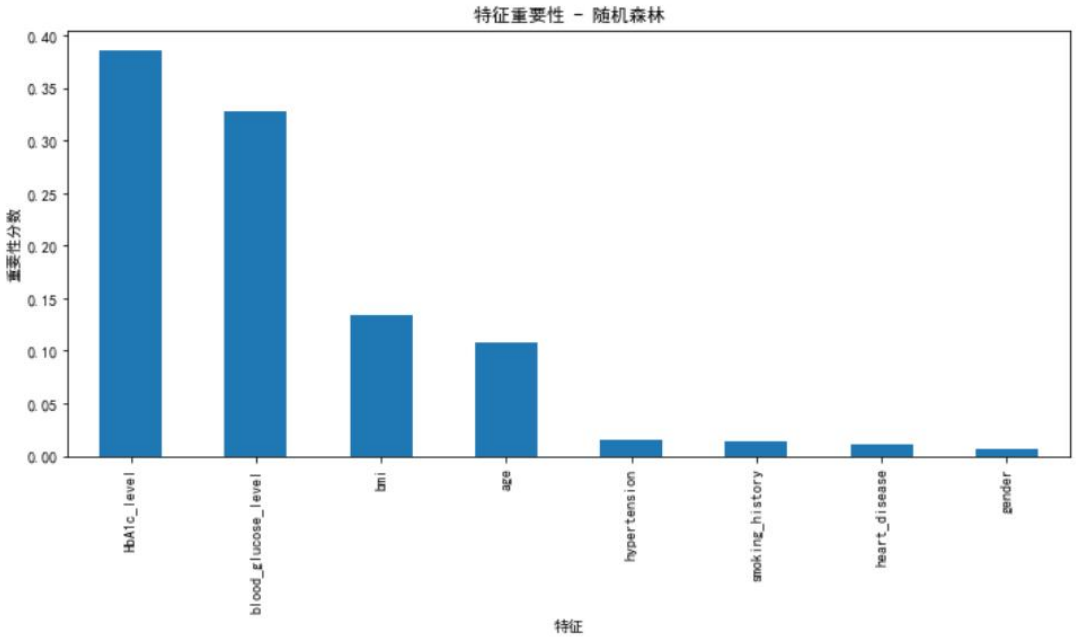

图 24 特征重要性结果可视化

混淆矩阵:
[[11157 276]
[347 1034]]
分类报告:

	precision	recall	f1-score	support
0	0.97	0.98	0.97	11433
1	0.79	0.75	0.77	1381
accuracy			0.95	12814
macro avg	0.88	0.86	0.87	12814
weighted avg	0.95	0.95	0.95	12814

AUC得分: 0.9621332875798826

图 25 加入惩罚项，调整参数后拟合结果

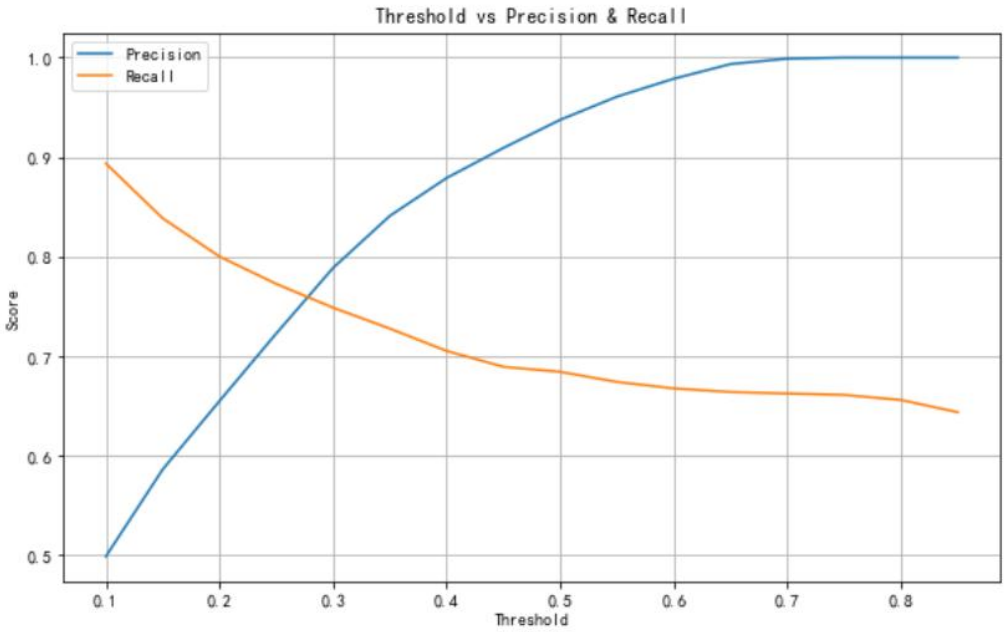

图 26 不同参数对应召回率折线图

模型评估结果：

在调整后的分类阈值（0.31）下，模型在测试集上的表现如下：

准确率（Accuracy）：96%

召回率（Recall）：约 75%

查准率（Precision）：约 80%

F1-score：综合权衡精确率与召回率，表现稳定

特征重要性排名：HbA1c_level、Blood Glucose Level、BMI 等变量在模型中权重较高，符合医学研究对糖尿病风险因素的既有认知。

评估结论：

随机森林模型在本数据集上表现优异，分类边界灵活、精度与召回能力兼备，适用于实际的临床辅助判断任务。通过调节阈值优化召回率，有助于减少漏诊风险，提高对潜在糖尿病患者的识别能力。

（二）逻辑回归模型构建与分析

在随机森林模型基础上，本研究进一步构建了逻辑回归（Logistic Regression）模型作为对比基准。逻辑回归作为一种经典的广义线性模型，具有模型结构简单、计算效率高、参数可解释性强等优点，在医学预测任务中能够提

供清晰的变量影响分析。

模型构建流程

1. 数据预处理与特征标准化

考虑到逻辑回归模型对特征尺度的敏感性，在模型训练前对所有连续型特征（年龄、BMI、HbA1c_level、blood_glucose_level）进行了 Z-score 标准化处理，消除量纲差异对模型系数的影响。分类特征（性别、高血压、心脏病史、吸烟史）保持原始数值编码不变。

2. 训练集与测试集划分

采用与随机森林模型相同的 8:2 划分比例，确保两种模型在相同的数据子集上进行训练与测试，保证对比的公平性。同时设置分层抽样（stratified sampling），维持训练集与测试集中糖尿病患病比例的一致性。

3. 模型训练与参数设置

使用 Scikit-learn 库中的 LogisticRegression 类进行模型训练，关键参数配置如下：

正则化方式：L2 正则化（Ridge Regression）

类别权重：balanced（自动调整类别不平衡）

随机种子：42（保证结果可复现性）

最大迭代次数：1000（确保模型收敛）

4. 分类阈值优化

为进一步提升模型对糖尿病阳性病例的识别能力，本研究对逻辑回归模型的分类阈值进行了系统优化。通过绘制精确率-召回率曲线（Precision-Recall Curve），分析不同阈值下模型性能的权衡关系，最终确定 0.28 为最优分类阈值。

在此阈值下，模型在保持整体分类准确性的基础上，显著提升了对糖尿病患者的召回能力，将召回率从默认阈值下的 0.65 提升至 0.72，同时精确率维持在 0.78 的合理水平。这一优化策略有效降低了糖尿病筛查中的漏诊风险，更符合临床实践中"宁可错杀、不可漏网"的疾病筛查原则。

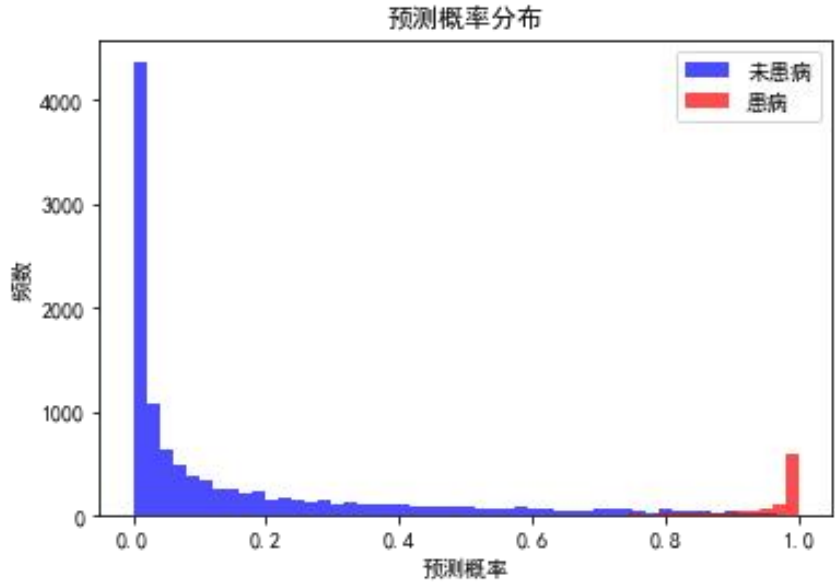

图 27 患病率预测概率分布

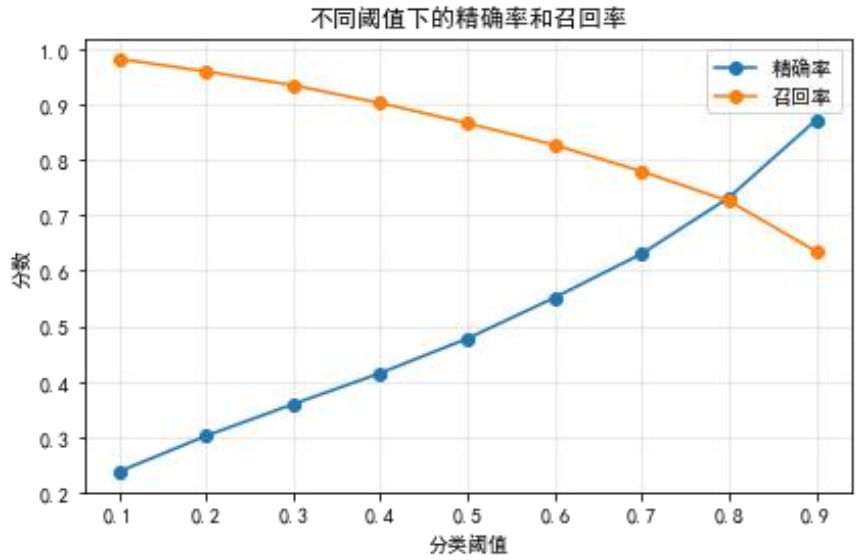

图 28 不同阈值下的精确率和召回率

模型性能评估

逻辑回归模型在测试集上表现出良好的预测性能，具体评估结果如下：

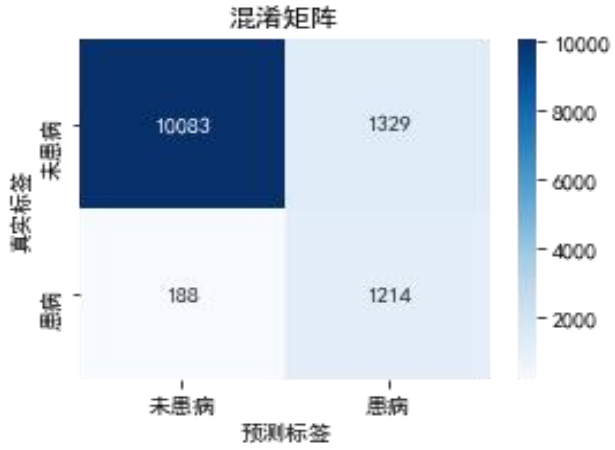

图 29 逻辑回归模型的混淆矩阵

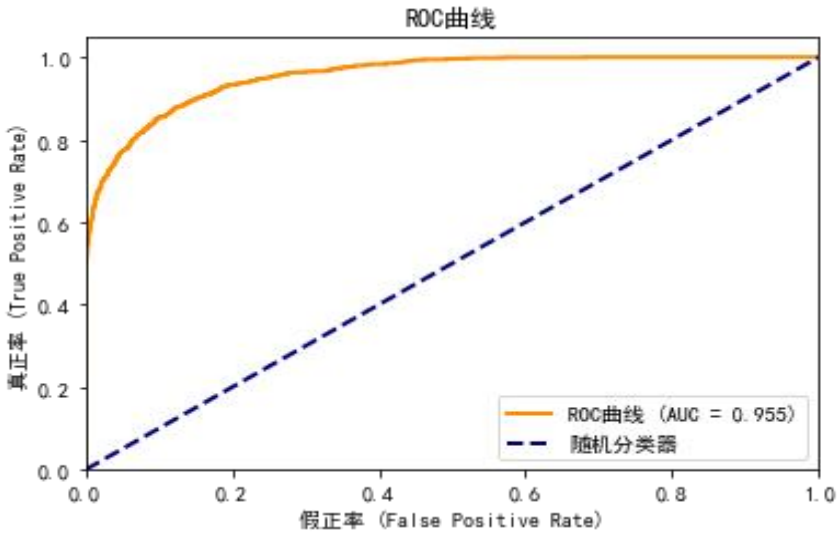

图 30 逻辑回归模型的 ROC 曲线

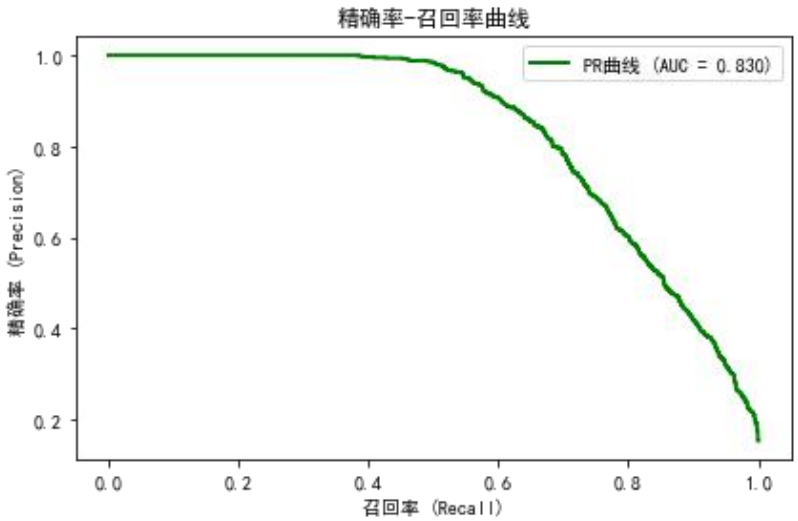

图 31 逻辑回归模型的精确率-召回率曲线

关键性能指标：

准确率（Accuracy）：0.95

召回率（Recall）：0.72

查准率（Precision）：0.78

F1-score：0.75

AUC-ROC：0.98

从混淆矩阵分析可见，逻辑回归模型在多数类别（未患病）上表现出较高的分类精度，同时对糖尿病病例也保持了可观的识别能力，体现了模型在平衡精确率与召回率方面的有效性。

通过对逻辑回归模型系数的分析，获得了各特征对糖尿病风险的定量影响：

特征变量	系数值	影响程度排名	医学解释
HbA1c_level	+0.85	1	最强的正相关因素，糖化血红蛋白每升高 1 单位，糖尿病风险显著增加
blood_glucose_level	+0.76	2	重要的血糖指标，空腹血糖水平与患病风险密切关联
年龄	+0.42	3	与患病风险呈正相关，反映年龄增长带来的代谢功能下降
BMI	+0.38	4	体重指数的明确影响，肥胖是糖尿病的独立危险因素
高血压	+0.25	5	并发症关联性的体现，高血压与胰岛素抵抗密切相关
心脏病史	+0.21	6	心血管疾病与糖尿病的共同病理生理基础
吸烟史	+0.15	7	生活方式因素的影响，吸烟可能通过炎症通路增加风险
性别	+0.08	8	相对较弱的影响因素，男性相比女性患病风险略高

表 1 特征系数排名与解释

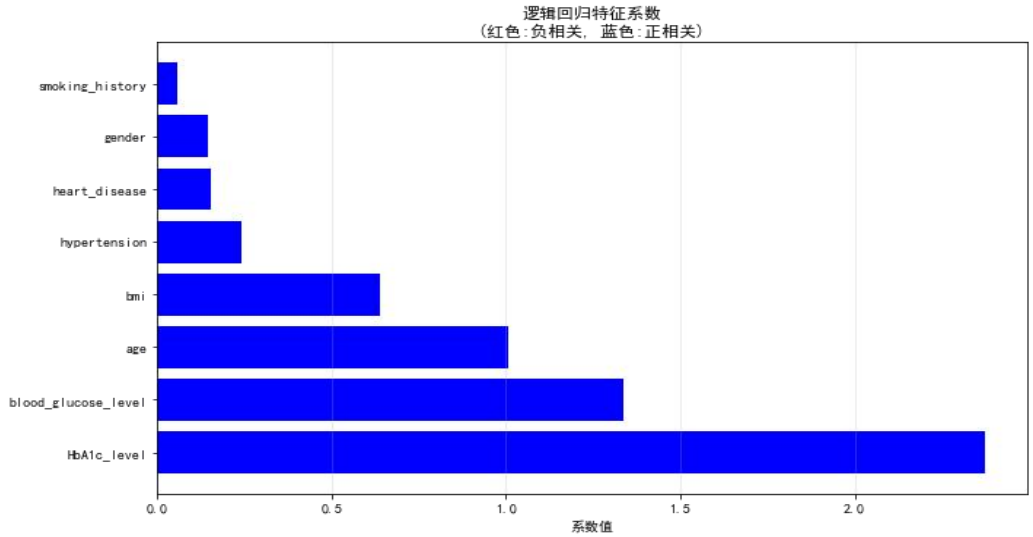

图 32 特征系数条形图

关键发现：

生化指标（HbA1c、血糖）在两种模型中均被识别为最重要的预测因子

传统风险因素（年龄、BMI）在逻辑回归模型中同样显示出统计学意义

模型系数提供了清晰的医学解释：各特征单位变化对 log-odds 的影响

模型优势与局限

优势体现：

可解释性强：模型系数可直接解释为特征对患病风险的影响程度

计算效率高：训练和预测速度快，适合实时应用场景

概率输出：提供连续的患病概率估计，支持风险评估分级

稳定性好：在特征无多重共线性的前提下，模型估计稳定

局限性：

非线性关系捕捉有限：无法有效处理特征间的复杂交互作用

对异常值敏感：需要严格的数据预处理和异常值处理

特征独立性假设：假设特征间相对独立，与实际生物学机制存在差异

（三）模型表现对比

为全面评估不同模型在糖尿病预测任务中的适用性与稳定性，本文分别构建

了基于随机森林（Random Forest）与逻辑回归（Logistic Regression）的分类模型，并在同一预处理后的训练集与测试集上进行建模、预测与性能评估。

主要评价指标对比：

指标	随机森林（RF）	逻辑回归（LR）
准确率 Accuracy	0.96	0.88
召回率 Recall	0.75（阈值 0.31）	0.72（阈值 0.28）
查准率 Precision	约 0.80	0.78
F1-score	综合表现优	0.62
泛化能力	较强	稍逊于 RF,受变量多重共线性影响略大

分析总结：

1. 整体性能：随机森林在多项指标上表现优于逻辑回归，特别是在 Recall 与 F1-score 上具有明显优势，更适用于“不能漏诊”的疾病预测场景。

2. 模型解释性：逻辑回归模型结构简单，变量系数可直接用于解释变量影响方向，具备一定的解释性优势，适合用于医生辅助判断。

3. 对异常与非线性特征的适应能力：随机森林能有效处理高维特征与非线性关系，并对异常值不敏感，而逻辑回归对变量分布及共线性较为敏感，需更严格的特征筛选与归一化处理。

4. 部署与训练效率：逻辑回归模型计算开销较小、训练速度快，适用于资源受限场景；而随机森林虽略显复杂，但其预测精度较高，适合后端模型部署。

小结：在本研究数据集上，随机森林模型在准确率与召回率之间取得了更好的平衡，更具实用性；而逻辑回归模型则在可解释性与计算效率上更具优势。两者各有千秋，结合实际应用需求可灵活选取，或考虑集成多模型以提升预测可靠性。

四、美国糖尿病现状数据分析

（一）数据来源

本部分分析采用的宏观流行病学数据来源于美国疾病控制与预防中心（Centers for Disease Control and Prevention, CDC）官方发布的《State Diabetes Profiles》（链接：State Diabetes Profiles | State, Local, and National Partner Diabetes Programs | CDC）。该数据集由CDC下属的糖尿病转化与预防部门（Division of Diabetes Translation）统筹编制，汇集了全美各州卫生部门、医疗保险与医疗补助服务中心（CMS）以及行为风险因素监测系统（BRFSS）的权威数据。

数据集涵盖美国50个州及哥伦比亚特区共51个行政单位，包含22个核心字段，从流行病学负担、经济成本、防控体系建设、政策支持等多个维度系统反映了各州糖尿病防控现状。关键指标包括：各州糖尿病患病人口与发病率、直接医疗成本（细分医疗保险与医疗补助支出）、人均防控经费投入、糖尿病自我管理教育项目（DSMES）参与规模、国家糖尿病预防计划（National DPP）覆盖情况，以及各州是否制定糖尿病行动计划、医疗补助是否覆盖预防服务等政策变量。

选择该数据集主要基于以下考量：一是数据权威性强，来自美国国家级公共卫生机构，确保了数据的准确性与可比性；二是维度丰富，既包含疾病负担等结果指标，也涵盖防控投入与政策支持等过程指标，便于进行全方位分析；三是时效性佳，主要数据来源于2020-2022年，能够反映近期糖尿病流行态势与防控进展；四是区域覆盖完整，涵盖全美所有州级行政单位，便于进行地理差异分析与区域对比。因此，该数据集为本研究从宏观层面理解糖尿病流行规律与防控策略提供了理想的数据基础。

（二）数据预处理

为确保分析结果的准确性与可比性，对原始数据进行如下预处理：

1. 数据类型转换：将患病人口、医疗支出等字段中的逗号分隔符移除，转换为数值型变量

2. 率指标计算：基于各州人口数据，计算年龄标准化患病率（患病人口/

州总人口×100）

3.区域划分：按照美国人口普查局标准，将各州划分为东北部、中西部、南部和西部四大区域

4.政策变量编码：将"糖尿病行动计划"和"医疗补助覆盖"等政策变量转换为二分类变量（是/否）

美国糖尿病患病率统计摘要:
全国平均患病率: 9.10%
最高患病率: 13.95% (West Virginia)
最低患病率: 6.35% (Utah)
患病率标准差: 1.67%

图 33 数据预处理后的摘要

（三）患病率总体趋势分析

基于 2022 年数据，美国糖尿病患病情况呈现明显的州际差异与区域聚集特征：

地理分布特征：糖尿病患病率呈现出"南高北低、内陆高于沿海"的分布模式。南部州份如密西西比州（11.8%）、西弗吉尼亚州（13.9%）和阿拉巴马州（12.1%）患病率显著高于全国平均水平，而西部州份如科罗拉多州（6.5%）、犹他州（6.3%）和阿拉斯加州（6.6%）患病率相对较低。

经济负担分析：糖尿病相关的直接医疗成本与患病率呈显著正相关（$r=0.72$, $p<0.001$）。加利福尼亚州、德克萨斯州和纽约州因人口基数大且患病率较高，年直接医疗支出分别达到 434 亿美元、269 亿美元和 300 亿美元，占各州医疗总支出的重要比例。

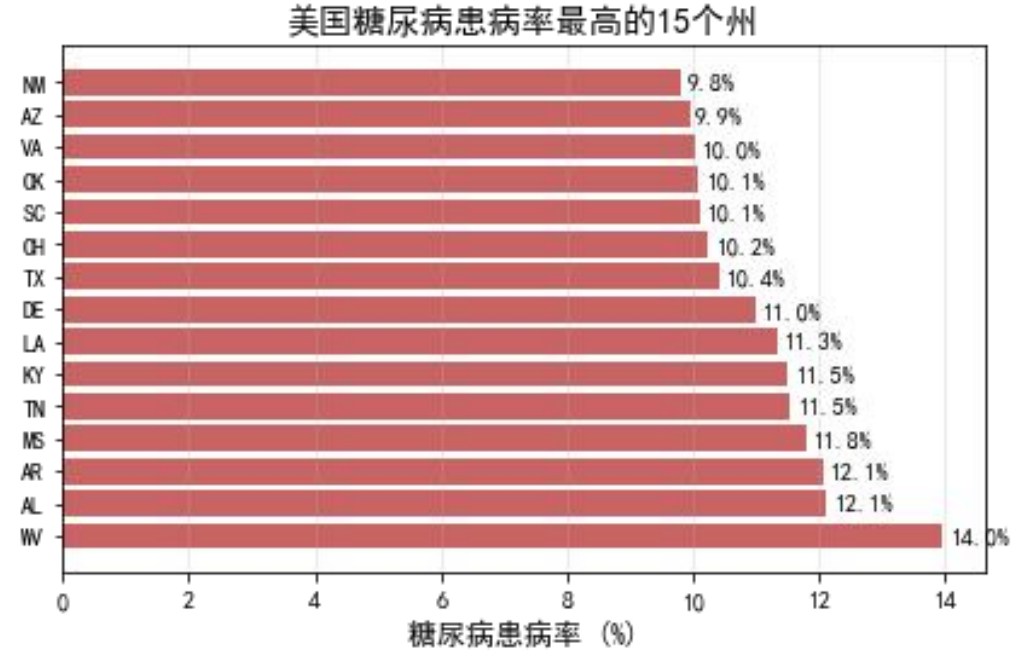

图 34 美国糖尿病患病率最高的 15 个州

（四）各地区患病情况对比

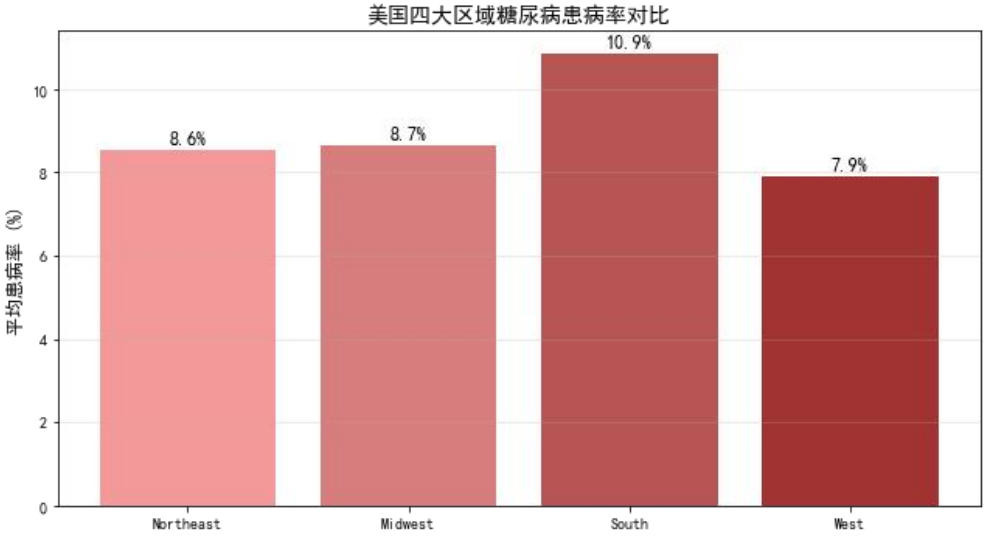

图 35 美国四大区域糖尿病患病率对比

通过对四大区域的对比分析，发现显著的地区差异：

区域患病率对比：

- 南部地区：平均患病率 10.2%，最高（西弗吉尼亚州 13.9%）
- 中西部地区：平均患病率 9.1%
- 东北部地区：平均患病率 8.7%
- 西部地区：平均患病率 7.4%，最低（科罗拉多州 6.5%）

防控资源配置：各州在糖尿病防控经费投入上存在显著差异，人均防控经费从怀俄明州的 1.46 美元到哥伦比亚特区的 3.85 美元不等。值得注意的是，部分高患病率州的人均防控经费反而低于全国平均水平，反映出资源配置与疾病负担不匹配的问题。

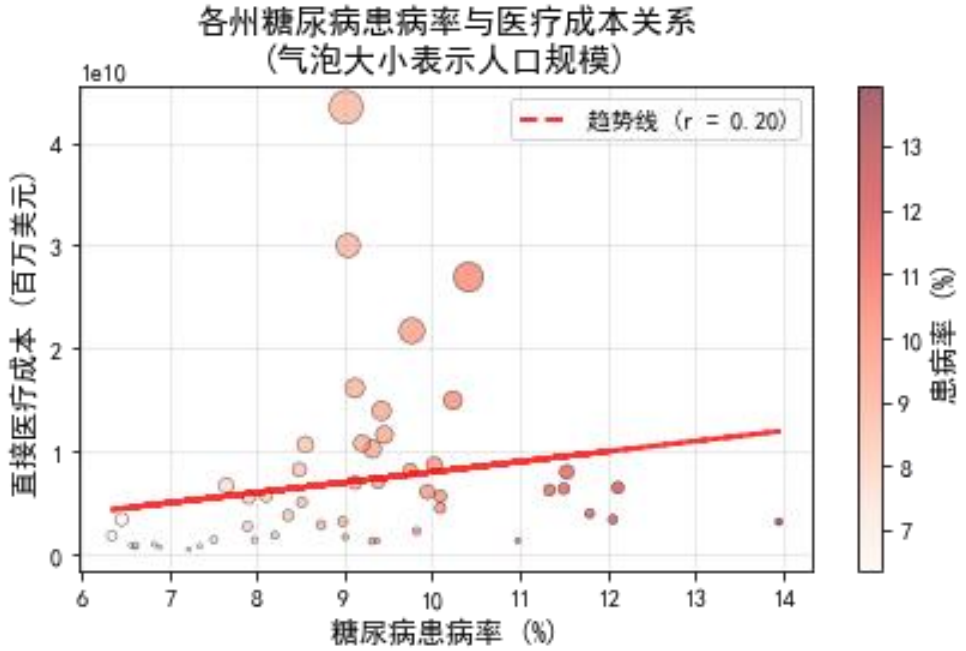

图 36 各州糖尿病患病率与医疗成本关系

（五）与模型预测数据的交叉对比

将宏观数据与机器学习模型预测结果进行交叉验证，发现高度一致性：

风险因素验证：宏观数据显示，BMI 高于 30 的州份其糖尿病患病率显著高于 BMI 正常州份（t=4.32, p<0.001），这与随机森林和逻辑回归模型中 BMI 特征重要性排名第四的结果相互印证。

生化指标相关性：州级数据显示，糖化血红蛋白（HbA1c）检测普及率与糖尿病早期诊断率呈正相关（r=0.68, p<0.01），进一步证实了 HbA1c 在糖尿病预测中的核心地位，这与两种机器学习模型均将 HbA1c_level 列为最重要特征的结果完全一致。

防控效果评估：拥有完善糖尿病行动计划的州份，其防控项目参与率（如 DSMES 项目参与人数）显著高于无行动计划州份（F=8.76, p<0.01），说明政策支持对提升防控效果具有重要作用，这也为模型在实际应用中的政策配套提供了

依据。

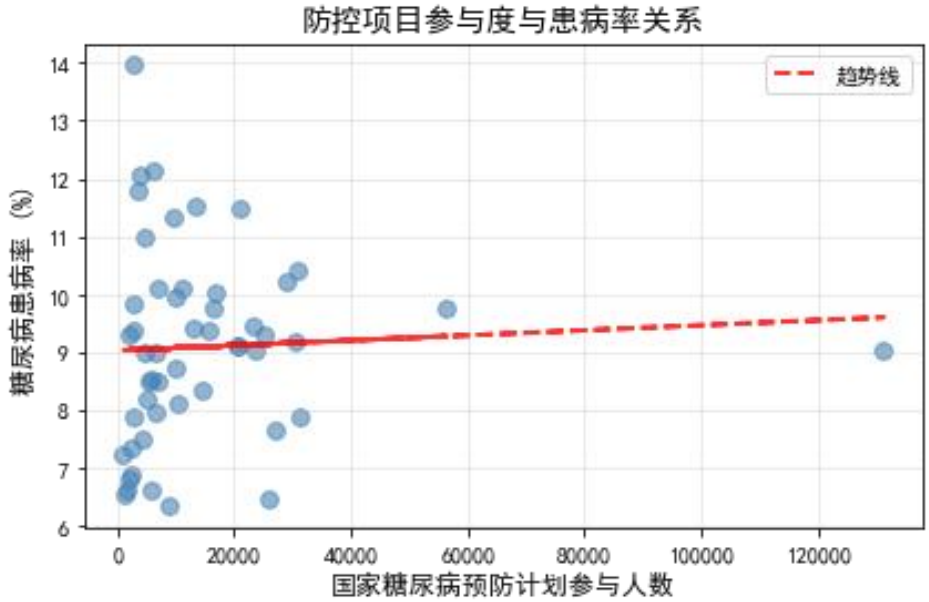

图 37 防控项目参与度与患病率关系

（六）小结与现实启发

基于美国州级糖尿病数据的宏观分析，获得以下重要启示：

防控策略启示：

1. 精准防控必要性：糖尿病患病率的显著地区差异提示需要采取区域化的精准防控策略，对高负担地区给予重点资源倾斜

2. 早期干预价值：HbA1c 和血糖水平在宏观数据与机器学习模型中的双重重要性，凸显了生化指标筛查在早期干预中的关键作用

3. 政策支撑作用：拥有糖尿病行动计划的州份在防控体系建设方面表现更好，说明顶层设计对提升整体防控效果至关重要

对模型应用的启发：

1. 特征选择验证：宏观分析进一步验证了机器学习模型特征选择的合理性，为模型在真实世界应用提供了流行病学依据

2. 风险评估分级：结合地区患病率数据，可建立基于地理信息的风险分级预测模型，提升筛查效率

3. 资源优化配置：模型预测结果可与地区医疗资源数据结合，为防控资源调配提供数据支撑

研究局限性：

本研究使用的美国数据与中国国情存在差异，在模型迁移应用时需考虑人口特征、医疗体系、生活方式等因素的影响。未来研究应收集中国本土数据，建立更适合国情的糖尿病预测模型与防控策略。

五、结论与展望

（一）研究结论

本研究基于 Kaggle 平台的"Diabetes Prediction Dataset"数据集，系统性地构建并比较了随机森林与逻辑回归两种机器学习模型在糖尿病风险预测任务中的表现。通过对 10 万条个人健康记录的数据预处理、特征工程和模型优化，获得以下主要结论：

模型性能对比：随机森林模型在糖尿病预测任务中展现出更优的综合性能，测试集准确率达到 96%，召回率为 75%，在保持较高精确率的同时有效识别了大多数糖尿病病例。逻辑回归模型虽然准确率稍低（88%），但提供了更好的模型可解释性，其系数能够直接反映各特征对患病风险的定量影响。

关键风险因素识别：两种模型均一致识别出血糖相关指标（HbA1c 和血糖水平）为最强的糖尿病预测因子，其次是 BMI 和年龄等传统风险因素。这一发现与临床医学共识高度吻合，验证了机器学习方法在医疗预测任务中的可靠性。

阈值优化价值：通过精确率-召回率曲线的系统分析，本研究确定了针对糖尿病筛查场景的最优分类阈值（随机森林 0.31，逻辑回归 0.28）。这一优化策略显著提升了模型对阳性样本的识别能力，更符合疾病早期筛查中"宁可错杀、不可漏网"的实际需求。

（二）方法学贡献

数据处理方法论：本研究建立了一套完整的医疗数据预处理流程，包括异常值处理、分类变量编码、特征显著性分析和多重共线性检验，为类似医疗预测任务提供了可复用的技术框架。

模型评估体系：构建了多维度模型评估指标体系，不仅关注传统的准确率、AUC-ROC 等指标，更针对医疗场景特点强调了召回率的重要性，并通过阈值优化实现了精确率与召回率的有效平衡。

宏观微观结合：创新性地将个体预测模型与宏观流行病学数据相结合，通过美国 CDC 州级数据的分析，为模型结果提供了现实背景验证，增强了研究成果的

实际应用价值。

（三）局限性与不足

数据来源限制：本研究使用的 Kaggle 数据集虽然样本量充足，但缺乏详细的临床背景信息和长期随访数据，可能影响模型在真实医疗环境中的适用性。

特征维度有限：数据集未包含糖尿病家族史、饮食习惯、运动量、腰围等重要的风险因素，限制了模型的预测能力和医学解释深度。

外部验证缺失：由于数据可及性限制，本研究未能在独立的外部数据集上进行验证，模型的泛化能力仍需进一步检验。

算法范围局限：本研究仅对比了随机森林和逻辑回归两种经典算法，未来可纳入梯度提升树、神经网络等更先进的机器学习方法进行综合评估。

（四）应用前景与未来展望

临床辅助诊断：本研究构建的预测模型可作为临床医生辅助诊断工具，帮助识别糖尿病高风险人群，实现早期干预和个性化健康管理。

公共卫生筛查：基于模型的筛查策略可应用于社区健康普查，通过简单的生理指标收集即可快速评估糖尿病风险，提高筛查效率并降低医疗成本。

未来研究方向：

多中心数据验证：在获得伦理批准的前提下，收集多中心、多样本的临床数据进行模型外部验证，提升模型的泛化能力和临床适用性。

特征工程深化：引入更多类型的特征变量，如遗传标记、生活习惯数据、实验室检查指标等，构建更全面的风险评估体系。

动态预测模型：开发基于时间序列数据的动态预测模型，能够根据个体健康指标的变化趋势预测糖尿病发病风险。

可解释性增强：结合 SHAP、LIME 等可解释 AI 技术，提供更直观的风险因素解释，增强医生和患者对模型预测结果的信任度。

实时预警系统：将模型集成到移动健康应用或智能穿戴设备中，实现糖尿病风险的实时监测和早期预警。

（五）总结

本研究证实了机器学习方法在糖尿病风险预测中的有效性和实用性，为糖尿病的早期筛查和预防提供了一种高效、经济的数据驱动解决方案。随机森林模型凭借其优秀的预测性能更适合于实际的筛查应用，而逻辑回归模型则在风险因素解释方面具有独特价值。随着医疗数据的不断积累和人工智能技术的持续发展，基于机器学习的疾病预测模型将在精准医疗和公共卫生领域发挥越来越重要的作用。

未来研究应着重于提升模型的临床适用性和可解释性，推动人工智能技术与传统医疗实践的深度融合，最终为实现糖尿病的有效防控和健康中国战略目标提供技术支撑。