Statistics and Application 统计学与应用, 2023, 12(4), 974-984
Published Online August 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.124101

Hans汉斯

基于机器学习方法的早期糖尿病风险预测

练春兰

云南财经大学统计与数学学院，云南 昆明

收稿日期：2023年7月16日；录用日期：2023年8月6日；发布日期：2023年8月18日

摘　要

糖尿病疾病是一个日益严重的医学问题，它是一种代谢疾病，身体内的葡萄糖长期处于一个高水平的状态，会产生尿频、口渴、饥饿程度加剧等症状，从而导致肾衰竭、中风、视力受损等并发症的产生。糖尿病的识别往往是病人询问医生或者是到诊断中心询问，会使诊断过程过于繁琐。但是逐步上升的机器学习方法解决了这一问题。本次研究的目的是采用机器学习方法，预测患者患糖尿病的可能性。因此采用四个机器学习分类算法，即朴素贝叶斯、决策树、随机森林及逻辑斯蒂回归，来检测早期糖尿病。实验采用的是UCI机器学习库中，从孟加拉国锡尔赫特的锡尔赫特医院患者那里收集的直接问卷。这四个算法的性能评估采用准确率来进行评估。实验显示随机森林的精度优于其他算法，达到了98.07%。

关键词

朴素贝叶斯，决策树，随机森林，逻辑斯蒂回归，R语言

Machine Learning-Based Approach to Early Diabetes Risk Prediction

Chunlan Lian

College of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Jul. 16th, 2023; accepted: Aug. 6th, 2023; published: Aug. 18th, 2023

Abstract

Diabetic disease is a growing medical problem. It is a metabolic disease in which glucose in the body remains at a high level for a long time, producing symptoms such as frequent urination, thirst and increased hunger levels, which can lead to complications such as kidney failure, stroke and impaired vision. Diabetes is often identified by the patient asking a doctor or visiting a diagnostic centre, which can make the diagnosis process too cumbersome. But progressively increasing machine learning methods solve this problem. The aim of this study was to use machine learning

文章引用: 练春兰. 基于机器学习方法的早期糖尿病风险预测[J]. 统计学与应用, 2023, 12(4): 974-984.
DOI: 10.12677/sa.2023.124101

methods to predict the likelihood of a patient developing diabetes. Four machine learning classification algorithms, namely, plain Bayesian, decision tree, random forest and logistic regression, were therefore used to detect early diabetes. The experiments were conducted using direct questionnaires collected from patients at Sylhet Hospital, Sylhet, Bangladesh, from the UCI Machine Learning Library. The performance of these four algorithms was evaluated using accuracy. The experiments showed that Random Forest outperformed the other algorithms with an accuracy of 98.07%.

Keywords

Naive Bayes, Decision Trees, Random Forest, Logistic Regression, R Language

1. 引言

2021 年 12 月 6 日，国际糖尿病联盟(IDF)发布了最新的全球糖尿病数据，据统计 2021 年全球约 5.37 亿成年人(20~79 岁)患有糖尿病(10 个人中就有 1 人为糖尿病患者)；预计到 2030 年，该数字将上升到 6.43 亿；到 2045 年将上升到 7.83 亿。在此期间，世界人口估计增长 20%，而糖尿病患者人数估计增加 46%。低收入和中等收入国家的患病率上升速度高于高收入国家。糖尿病是失明、肾衰竭、心脏病发作、中风和下肢截肢的主要病因。2000 年至 2019 年期间，糖尿病导致的死亡增加了 3%。2019 年，糖尿病以及糖尿病引起的肾脏疾病估计造成 200 万人死亡。可以看出糖尿病的经济成本似乎在全球范围内都有所增加。糖尿病是由于胰腺细胞产生的胰岛素不足或身体细胞对产生的胰岛素没有适当反应，导致碳水化合物、脂肪、蛋白质代谢紊乱，造成多种器官的慢性损伤、功能障碍甚至衰竭[1]。

糖尿病主要有四种类型[2]，它们分别是：

1) I 型糖尿病：发病与 T 细胞介导的自身免疫导致胰岛 β 细胞的选择性破坏，胰岛素分泌减少和绝对缺乏有关，单用口服药无效，需要注射胰岛素来治疗[3]。

2) II 型糖尿病：发病由遗传易感性和现代生活方式(膳食、运动)造成的胰岛素分泌缺陷造成[3]。

3) 其他特殊类型：肝脏疾病、慢性肾功能不全、多种内分泌疾病、急性感染、创伤，外科手术都可能导致血糖一过性升高[3]。

4) 妊娠糖尿病：妊娠期间引发的糖尿病，产后需控制恢复，仍是危险人群。一般情况下在婴儿出生之后就会消退[3]。

机器学习的分类算法广泛应用与医学领域的数据分类。糖尿病受身高、体重、遗传和胰岛素功能等功能的影响，我们考虑的主要因素就是血糖浓度。早期识别是唯一远离并发症的补救方法[4]。许多研究者进行疾病诊断实验时，会使用各种分类的机器学习算法，例如：支持向量机(SVM) [5]、朴素贝叶斯[6]、决策树[7]、逻辑斯蒂回归[8]、神经网络[9]等等。数据挖掘[10]和机器学习方法对于来自不同数据源的数据的疾病诊断处理具有强大的能力[11]。在研究糖尿病，Nai-Arun 等人[12]提出了一种分类集成学习来研究糖尿病，利用增益比特征选择技术对数据进行分析。Orabi 等人[13]介绍了一种通过提高预防措施警报来帮助人们的方法。它是糖尿病疾病的预测系统，它将预测是否成为候选人以及在什么年龄。该系统基于机器学习概念，使用决策树技术，通过添加带有随机化代码的回归技术来预测年龄。Bamnote 等人[14]

提出了一种使用遗传编程(GP)检测糖尿病的分类器，使用分类表达式创建分类器。使用仅算术运算符的简化函数池，允许在交叉和突变期间进行较少的验证和宽大处理。Nai-Arun 等人[15]首先研究了四个众所周知的分类模型，即决策树、人工神经网络、逻辑回归和朴素贝叶斯。然后，研究了袋装和增压技术以提高此类模型的鲁棒性。诸如对糖尿病的研究还有很多很多，如，Vijiya Kumar 等人[16]提出的使用随机森林对糖尿病进行预测；Sisodia 等人[17]研究的是使用分类算法预测糖尿病等等。

本次的研究工作是关注早期糖尿病这一疾病。在这项工作中采取了朴素贝叶斯、决策树、随机森林和逻辑斯蒂回归，这四种机器学习方法来对早期糖尿病进行预测。在四种机器学习方法下，都取得了良好的精度。

其余的研究讨论组织结构如下：第二部分，介绍机器学习分类算法。第三部分，进行数据集的实证分析及评估结果。第四部分，进行研究总结。

2. 相关理论及方法

2.1. 朴素贝叶斯

朴素贝叶斯分类(Naive Bayes)是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 x 求出使得后验概率最大的输出 y [4]。它适用于数据不平衡及数据缺失，而且还适用于维度非常高的数据集。朴素贝叶斯分类广泛应用于文本分类、垃圾邮件过滤、情感分析等领域。

根据朴素贝叶斯算法可得：$P(Y|X)=P(Y)P(X|Y)/P(X)$，其中，$P(Y|X)$ 是目标类后验概率，$P(X|Y)$ 是预测类概率，$P(Y)$ 是 Y 概率是正确的，$P(X)$ 是预测的先验概率[6]。

2.2. 决策树

决策树(Decision Tree)是一个监督机器学习算法，主要用于研究分类问题。一个决策树学习算法需要包含特征选择、决策树生成和决策树剪枝过程。本文使用的是 ID3 算法，它的核心是在决策树各个节点上应用信息增益准则选择特征，递归的构建决策树。信息增益 $g(M,A)$ 也就是：$g(M,A)=H(M)-H(M|A)$，其中，$H(M)$ 是数据集 M 的经验熵，$H(M|A)$ 是数据集 M 的经验条件熵[7]。

决策树一般适合处理离散型数据，计算复杂度不高，对中间值的缺失不敏感，可以处理不相关特征数据。但是决策树方法可能产生过度匹配问题，对连续性的字段比较难以预测。它通常适用于金融分析、医疗诊断、营销推荐、交通安全等[7]。

2.3. 随机森林

随机森林(Random Forest)是建立多个决策树并将他们融合起来得到一个更加准确和稳定的模型，是 bagging 思想和随机选择特征的结合。随机森林构造了多个决策树，当需要对某个样本进行预测时，统计森林中的每棵树对该样本的预测结果，然后通过投票法从这些预测结果中选出最后的结果[16]。

随机森林适用于高维数据，不容易产生过拟合。对于大部分数据遗失，仍然可以维持高准确度。对于数据集的适应能力强，既能处理离散型数据，也能处理连续性数据，数据集也无需规范化[16]。

2.4. 逻辑斯蒂回归

二项逻辑斯蒂回归模型(binomial logistic regression)是一种分类模型，它属于对数线性模型，原理是根据现有的数据对分类边界线建立回归公式，以此进行分类。它常用于数据挖掘、疾病自动诊断、经济预测等领域。它仅能适用于线性问题，容易欠拟合，导致分类精度不高[8]。

3. 实证分析

3.1. 数据集说明

本次使用的数据是早期糖尿病风险预测数据集，从孟加拉国锡尔赫特的锡尔赫特糖尿病医院的患者那里收集的直接问卷。问卷中包含了年龄、性别、多尿等等，如表 1，收集了 520 名患者数据。在实验中，随机选取 80%的数据用于训练过程，20%的数据用于测试过程。

Table 1. Variable description
表 1. 变量说明

变量名称	说明
年龄(Age)	定量变量，20~65
性别(Gender)	定性变量，男/女
多尿(Polyuria)	定性变量，Yes/No
烦渴(Polydipsia)	定性变量，Yes/No
突然减肥(Sudden weight loss)	定性变量，Yes/No
虚弱(Weakness)	定性变量，Yes/No
多食症(Polyphagia)	定性变量，Yes/No
生殖器鹅口疮(Genital thrush)	定性变量，Yes/No
视觉模糊(Visual blurring)	定性变量，Yes/No
瘙痒(Itching)	定性变量，Yes/No
易怒(Irritability)	定性变量，Yes/No
延迟愈合(Delayed healing)	定性变量，Yes/No
部分麻痹(Partial paresis)	定性变量，Yes/No
肌肉刺激(Muscle stiffness)	定性变量，Yes/No
脱发(Alopecia)	定性变量，Yes/No
肥胖(Obesity)	定性变量，Yes/No
等级(Class)	定性变量，Positive/Negative

3.2. 结果与分析

3.2.1. 朴素贝叶斯

利用朴素贝叶斯进行建模，绘制出 ROC 曲线，如图 1 所示，可以看出 AUC 的值达到 0.913，说明该分类器的性能比较好。然后使用混合矩阵，查看模型评估结果，如表 2 所示，计算出模型准确率为 90.38%。由于该组数据没有空值，则不需要拉普拉斯平滑处理。

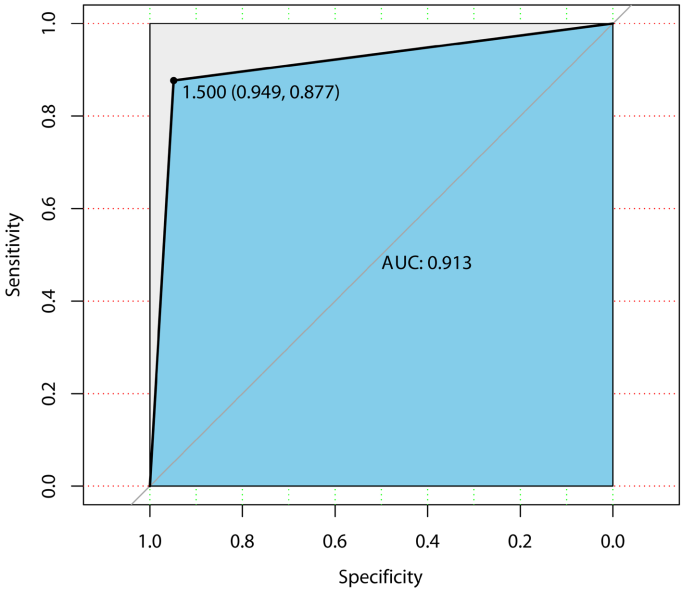

Figure 1. Naive Bayesian ROC curves
图 1. 朴素贝叶斯 ROC 曲线

Table 2. Confusion matrix of naive Bayesian
表 2. 朴素贝叶斯混合矩阵

	Negative	Positive
Negative	37	8
Positive	2	57

3.2.2. 决策树

利用决策树进行建模，对树进行可视化，如图 2 所示，通过检查底部节点，可以看到有多少分类是正确的。决策树对于检查特征的重要性、每个特征的预测能力也很有用，特征重要性按降序排序，如表 3 所示，可以看出对于早期糖尿病来说重要的因素是烦渴(Polydipsia)，多尿(Polyuria)，最不重要的因素是年龄(Age)，视觉模糊(visual blurring)。绘制出 ROC 曲线，如图 3 所示，可以看出 AUC 的值达到了 0.915，说明该分类器的性能比较好。然后使用混合矩阵，查看模型评估结果，如表 4 所示，计算出模型准确率为 91.35%。

Table 3. Importance of decision tree feature variables
表 3. 决策树特征变量重要性

Variable	Overall
Polydipsia	113.575726
Polyuria	85.630108
Gender	62.186220
Sudden weight loss	51.207733

Continued

Partial paresis	50.935121
Irritability	21.666840
Alopecia	12.636410
Muscle stiffness	7.827421
Genital thrush	3.917258
Delayed healing	3.816346
Polyphagia	3.500251
Itching	2.880000
Obesity	2.538690
Weakness	2.041063
Age	0.000000
Visual blurring	0.000000

Table 4. Confusion matrix of decision tree
表 4. 决策树混合矩阵

	Negative	Positive
Negative	36	6
Positive	3	59

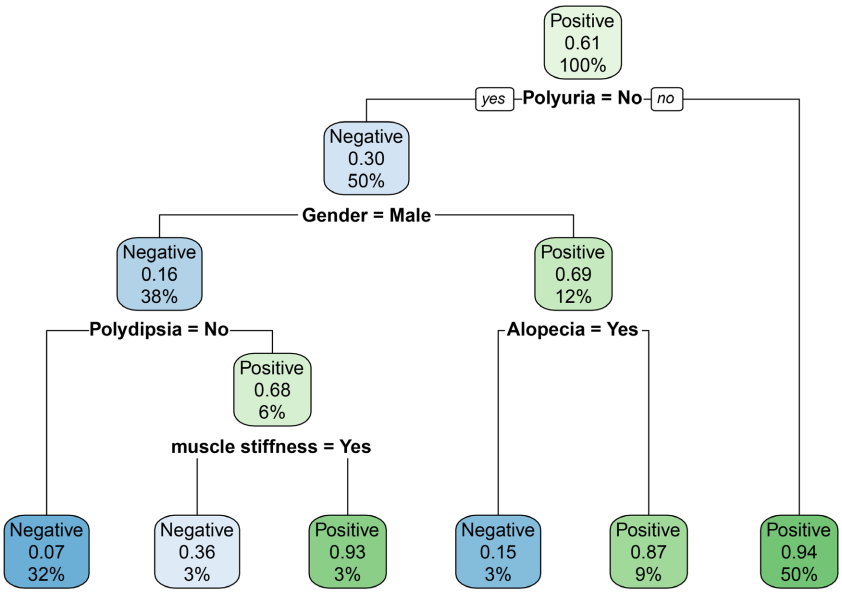

Figure 2. Tree visualization
图 2. 树可视化

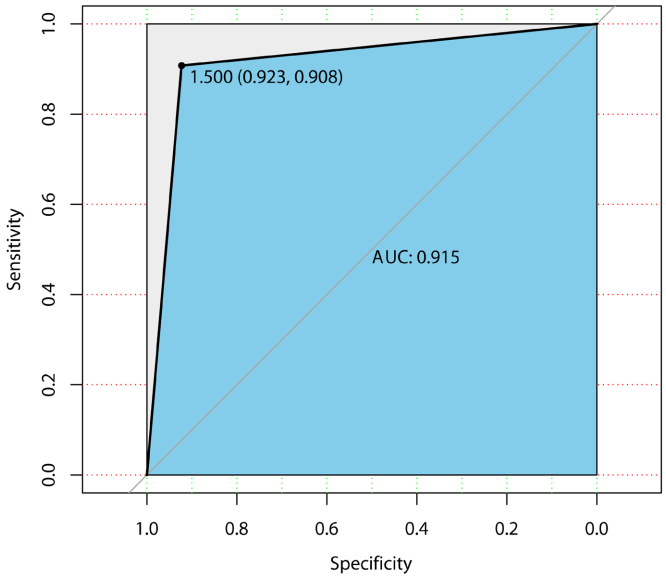

Figure 3. Decision tree ROC curve
图 3. 决策树 ROC 曲线

3.2.3. 随机森林

利用随机森林模型进行建模，该模型对于 mtry 和 ntree 两个参数的选取十分重要，由此在训练集中分出 80%的数据作为新的训练集，20%的数据作为验证集。在验证集上选取最佳的 mtry 为 3。如图 4 所示，当 ntree 取 100 时，模型内的误差就基本稳定了，出于更保险的考虑，我们确定 ntree 值为 100。

随机森林与决策树一样也可以检查对变量的重要性，如图 5 所示。在随机森林中变量的重要性计算时通过将相应变量替换成一列随机的数后，计算模型准确率或者 GINI 系数的降低。Mean Decrease Accuracy：表示变量替换后准确率的下降；Mean Decrease Gini：表示变量替换后 GINI 系数的降低。数值越大表示变量越重要。绘制出 ROC 曲线，如图 6 所示，可以看出 AUC 的值达到了 0.979，该值已经快要接近 1，说明该分类器的性能好。然后使用混合矩阵，查看模型评估结果，如表 5 所示，计算出模型准确率为 98.07%。

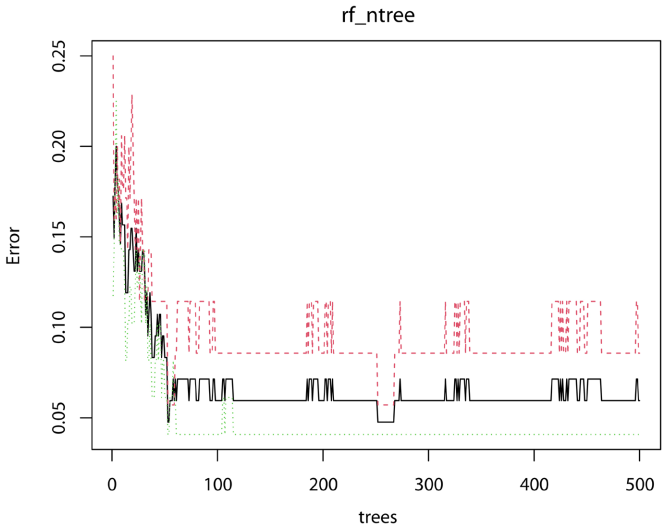

Figure 4. Ntree parameter selection
图 4. Ntree 参数选取

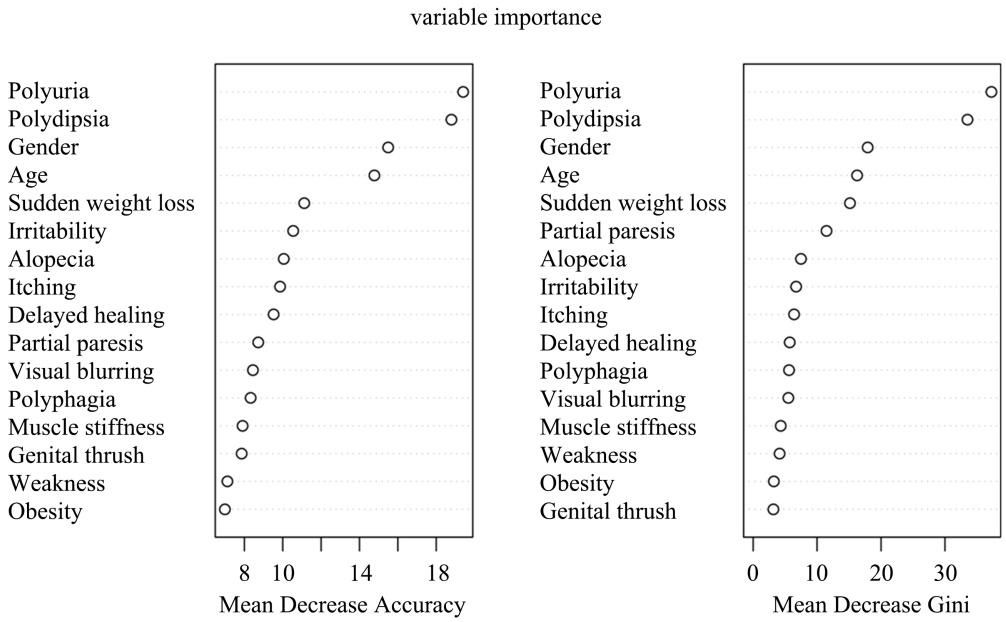

Figure 5. Importance of random forest variables
图 5. 随机森林变量重要性

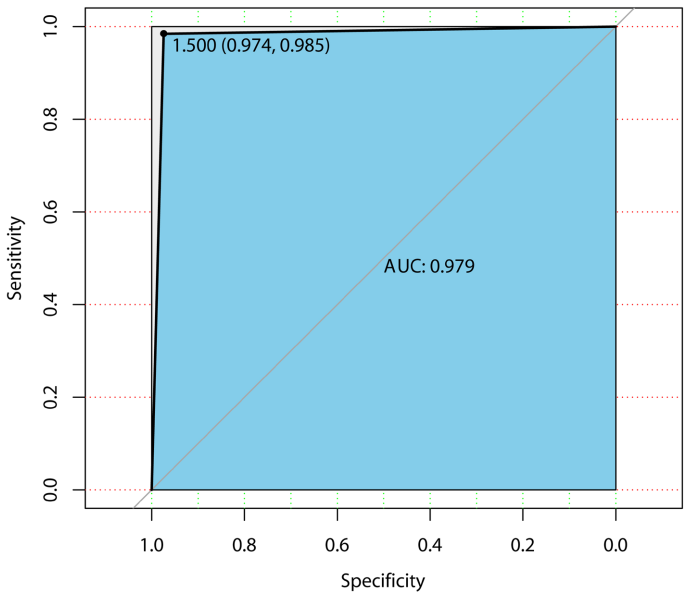

Figure 6. Random forest ROC curve
图 6. 随机森林 ROC 曲线

Table 5. Confusion matrix of random forest
表 5. 随机森林混合矩阵

	Negative	Positive
Negative	38	1
Positive	1	64

3.2.4. 逻辑斯蒂回归

利用逻辑斯蒂回归，创建模型，可以看到大部分变量都通过显著性检验，只有极少数的变量不显著，如表 6 所示。计算出在训练集上的$R2$ 为 0.7559，绘制出 ROC 曲线图，如图 7 所示，可以看出 AUC 的值达到了 0.944，说明该分类器的性能比较好。然后使用混合矩阵，查看模型评估结果，如表 7 所示，计算出模型准确率为 94.23%。

Table 6. Logistic regression
表 6. 逻辑斯蒂回归

Coefficients	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	2.53631	1.16473	2.18	0.0294 *
Age	−0.04206	0.02678	−1.57	0.1163
Gender: Male	−4.44419	0.67932	−6.54	6.1e−11 ***
Polyuria: Yes	5.11979	0.91574	5.59	2.3e−08***
Polydipsia: Yes	5.23044	0.96912	5.40	6.8e−08***
Sudden weight loss: Yes	−0.00139	0.64799	0.00	0.9983
Weakness: Yes	1.28695	0.63972	2.01	0.0442*
Polyphagia: Yes	1.37161	0.59031	2.32	0.0201*
Genital thrush: Yes	1.62509	0.63792	2.55	0.0109*
Visual blurring: Yes	0.98150	0.72366	1.36	0.1750
Itching: Yes	−2.82010	0.75721	−3.72	0.0002***
Irritability: Yes	2.06855	0.67013	3.09	0.0020**
Delayed healing: Yes	−0.64304	0.65075	−0.99	0.3231
Partial paresis: Yes	0.93269	0.59629	1.56	0.1178
Muscle stiffness: Yes	−1.28857	0.65397	−1.97	0.0488*
Alopecia: Yes	−0.29067	0.72604	−0.40	0.6889
Obesity: Yes	−0.35203	0.62357	−0.56	0.5724

Table 7. Confusion matrix of logistic regression
表 7. 逻辑斯蒂混合矩阵

	Negative	Positive
Negative	37	4
Positive	2	61

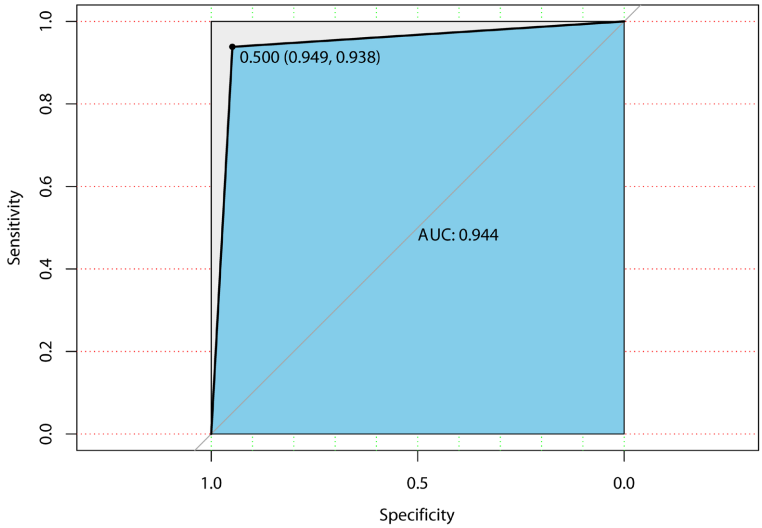

Figure 7. Logistic regression ROC curve
图 7. 逻辑斯蒂 ROC 曲线

3.2.5. 模型评价

本次分析采用了四种不同的模型，分别是：朴素贝叶斯，决策树，随机森林和逻辑斯蒂回归，获得了四种不同的计算结果。观察得到四种不同的结果准确率都高达 90%以上，如表 8 所示，由此可以得出这些模型对该数据集非常适用。本次模型训练集建议选取 80%，剩下 20%做测试集。此外通过决策树、随机森林及逻辑斯蒂回归分析糖尿病风险因素，发现多尿(polyuria)和烦渴(polydipsia)是导致糖尿病的主要因素。从结果分析中可以看出，随机森林算法获得了 98.07%的高精度，这比剩下的三种模型都要好得多。由此可以看出，利用随机森林模型可以很好的预测早期糖尿病这一类疾病。

Table 8. Models prediction accuracy
表 8. 模型预测精度

算法	准确度
朴素贝叶斯	90.38%
决策树	91.35%
随机森林	98.07%
逻辑斯蒂回归	94.23%

4. 总结

本文采用早期糖尿病风险预测数据集，通过使用朴素贝叶斯、决策树、随机森林及逻辑斯蒂回归模型，探讨了早期糖尿病风险因素的预测。风险因素预测在识别摆脱早期糖尿病疾病的风险方面起着重要的作用。从决策树、随机森林及逻辑斯蒂回归分析糖尿病风险因素，发现多尿(polyuria)和烦渴(polydipsia)是导致糖尿病的主要因素。

其次从模型预测结果上来分析，四种模型的精度都达到 90%以上，说明这四种模型都能很好的预测早期糖尿病疾病。在所使用模型中随机森林模型的精度较高(98.07%)，所以预测早期糖尿病疾病来说采用随机森林效果更佳。

致　谢

在本论文的写作中，我也参照了大量的著作和文章，许多学者的科研成果及写作思路给我很大启发，在此向这些学者们表示由衷的感谢。感谢我的老师、家人、同学、朋友对我的大力支持，他们的无私奉献、关爱和支持使我能够继续去追求自己的人生理想和目标。感谢所有关心、帮助和支持我的人。

参考文献

[1] Kitabchi, A.E., Umpierrez, G.E., Miles, J.M. and Fisher, J.N. (2009) Hyperglycemic Crises in Adult Patients with Diabetes. *Diabetes Care*, **32**, 1335-1343. https://doi.org/10.2337/dc09-9032

[2] World Health Organization (2023) World Health Statistics 2023: Monitoring Health for the SDGs, Sustainable Development Goals. Geneva.

[3] Chandalia, H.B. and Tripathy, B.E. (2023) RSSDI Textbook of Diabetes Mellitus.

[4] Vijayan, V.V. and Anjali, C. (2016) Prediction and Diagnosis of Diabetes Mellitus—A Machine Learning Approach. 2015 *IEEE Recent Advances in Intelligent Computational Systems* (*RAICS*), Trivandrum, 10-12 December 2015, 122-127. https://doi.org/10.1109/RAICS.2015.7488400

[5] Huang, S., Cai, N., Pacheco, P.P., *et al*. (2018) Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, **15**, 41-51. https://doi.org/10.21873/cgp.20063

[6] Boyles, S., Fajardo, D. and Waller, S.T. (2007) Naive Bayesian Classifier for Incident Duration Prediction. *Transportation Research Board* 86*th Annual Meeting*, Washington DC, 21-25 January 2007.

[7] Song, Y.Y. and Lu, Y. (2015) Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, **27**, 130-135.

[8] La Valley, M.P. (2008) Logistic Regression. *Circulation*, **117**, 2395-2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[9] Bishop, C.M. (1994) Neural Networks and Their Applications. *Review of Scientific Instruments*, **65**, 1803-1832. https://doi.org/10.1063/1.1144830

[10] Bai, B.G.M., Nalini, B.M. and Majumdar, J. (2019) Analysis and Detection of Diabetes Using Data Mining Techniques—A Big Data Application in Health Care. *Conference Proceedings Emerging Research in Computing, Information, Communication and Applications*, Vol. 1, 443-455.

[11] Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, **9**, 1-16. https://doi.org/10.4236/jilsa.2017.91001

[12] Nai-Arun, N. and Sittidech, P. (2014) Ensemble Learning Model for Diabetes Classification. *Advanced Materials Research*, **931-932**, 1427-1431. https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427

[13] Ying, W.X., Hong, H.S., Quan, L.Z., *et al*. (2009) The Predictive Role of Diabetes Mellitus with Erectile Dysfunction on Cardiovascular Risk. *Chinese Circulation Journal*.

[14] Pradhan, M. and Bamnote, G.R. (2015) Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. *Advances in Intelligent Systems & Computing*, **327**, 763-770. https://doi.org/10.1007/978-3-319-11933-5_86

[15] Nai-Arun, N. and Moungmai, R. (2015) Comparison of Classifiers for the Risk of Diabetes Prediction. *Procedia Computer Science*, **69**, 132-142. https://doi.org/10.1016/j.procs.2015.10.014

[16] Vijiya Kumar, K., Lavanya, B., Nirmala, I. and Sofia Caroline, S. (2019) Random Forest Algorithm for the Prediction of Diabetes. 2019 *IEEE International Conference on System, Computation, Automation and Networking* (*ICSCAN*), Pondicherry, 29-30 March 2019, 1-5. https://doi.org/10.1109/ICSCAN.2019.8878802

[17] Sisodia, D. and Sisodia, D.S. (2018) Prediction of Diabetes Using Classification Algorithms. *Procedia Computer Science*, **132**, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122