机器学习大作业——糖尿病数据的分类问题

PB23000187 商扬笛

2025 年 11 月 24 日

1 摘要

任务是用多种统计方法对数据进行处理、分析，构造不同的分类器，并加以比较，尽可能获得表现良好的分类器。为此首先进行数据处理，并将数据分割为训练集和测试集。本文训练并比较了KNN模型、LDA模型和逻辑回归模型三种分类器，并比较了模型的准确率、灵敏度、特异度、AUC、ROC曲线等性能，并以AUC为主要依据确定了最佳模型为逻辑回归模型，并根据模型分析得到了糖尿病预测因子重要性排序。最后对数据进行了PCA分析，确定了数据变异的主要因素。

2 背景

糖尿病是一种由胰岛素绝对或相对分泌不足以及利用障碍引发的，以高血糖为标志的慢性疾病。该疾病主要分为1型、2型和妊娠糖尿病三种类型。病因主要归结为遗传因素和环境因素的共同作用，包括胰岛细胞功能障碍导致的胰岛素分泌下降，或者机体对胰岛素作用不敏感或两者兼备，使得血液中的葡萄糖不能有效被利用和储存。一部分糖尿病患者和家族有疾病聚集现象。糖尿病的症状主要表现为“三多一少”，即多饮、多尿、多食和体重下降。此外，病程久的患者可能会引发眼、肾、神经、心脏、血管等组织器官的慢性进行性病变、功能减退甚至衰竭，并有可能引发急性严重代谢紊乱。数据显示，2023年我国年龄标准化糖尿病总体患病率为13.7%（包含儿童青少年，未区分糖尿病类型），患病人数达2.33亿，较2005年增长163%，且患者群体呈现年轻化趋势。若任由现状发展，患病率或将呈线性增长，2030年将达16.15%，2040年将达21.52%，2050年将达29.10%，其中天津和北京或超40%。作为慢性病，糖尿病早期往往没有表现，因此并不存在前兆。患者多通过体检、行血糖化验等发现患病，待出现“三多一少”等症状时，通常已经较为严重了，因此构建通过人体指标准确判断是否患糖尿病的分类器是必要的。

ROC曲线是一种用于表示分类模型性能的图形工具。它通过将真阳性率（True Positive Rate，TPR）和假阳性率（False Positive Rate，FPR）作为横纵坐标来描绘分类器在不同阈值下的性能。真阳性率（True Positive Rate，TPR）通常也被称为敏感性（Sensitivity）或召回率（Recall）。它是指分类器正确识别正例的能力。真阳性率可以理解为所有阳性群体中被检测出来的比率(1-漏诊率)，因此TPR越接近1越好。假阳性率（False Positive Rate，FPR）是指在所有实际为负例的样本中，模型错误地预测为正例的样本比例。假阳性率可以理解为所有阴性群体中被检测出来阳性的比率(误诊率)，因此FPR越接近0越好。AUC（ROC曲线下面积）是ROC曲线下的面积，用于衡量分类器性能。AUC值越接近1，表示分类器性能越好；反之，AUC值越接

近0，表示分类器性能越差。在实际应用中，我们常常通过计算AUC值来评估分类器的性能。因此本文采用AUC作为判断分类器性能的主要依据。

3　数据处理

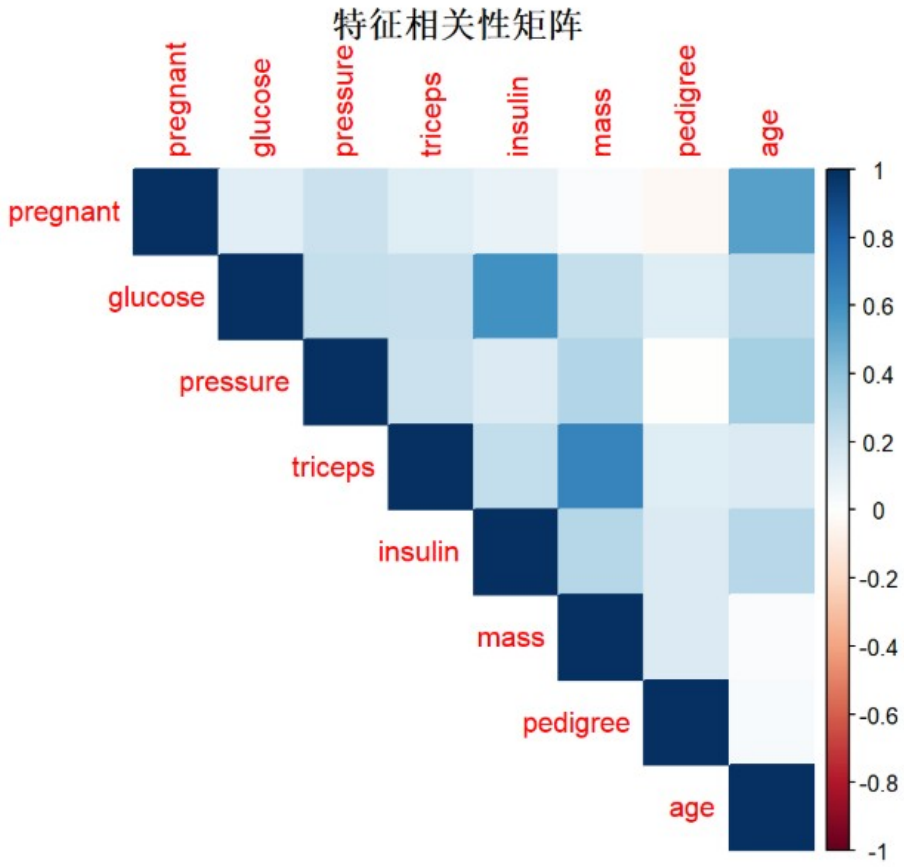
首先对数据进行预览

```
##     pregnant         glucose         pressure         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           mass          pedigree          age        diabetes
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   neg:500
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   pos:268
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

发现葡萄糖水平、血压、BMI、胰岛素水平、三头肌皮褶厚度这些不可能为零的人体指标数据为零，这些异常应该来源于数据的缺失，因此我们用”NA”来代替不合理的”0”，为了确保数据分布合理且与原始数据一致，使用KNN法插补缺失值，即基于欧氏距离计算k个和缺失值最近的观测值，然后使用这些观测值的加权平均值（距离越远权重越小）进行填补。

预览处理后的数据

```
##     pregnant          glucose          pressure          triceps
##  Min.   :-1.1411   Min.   :-2.544134   Min.   :-3.909269   Min.   :-2.114485
##  1st Qu.:-0.8443   1st Qu.:-0.742960   1st Qu.:-0.678814   1st Qu.:-0.682775
##  Median :-0.2508   Median :-0.153485   Median :-0.032723   Median :-0.014643
##  Mean   : 0.0000   Mean   :-0.001016   Mean   :-0.005803   Mean   :-0.007485
##  3rd Qu.: 0.6395   3rd Qu.: 0.632482   3rd Qu.: 0.613368   3rd Qu.: 0.558040
##  Max.   : 3.9040   Max.   : 2.531902   Max.   : 4.005345   Max.   : 6.666670
##     insulin            mass           pedigree           age
##  Min.   :-1.19173   Min.   :-2.058843   Min.   :-1.1888   Min.   :-1.0409
##  1st Qu.:-0.55187   1st Qu.:-0.715880   1st Qu.:-0.6885   1st Qu.:-0.7858
##  Median :-0.21510   Median :-0.044399   Median :-0.2999   Median :-0.3606
##  Mean   :-0.01303   Mean   :-0.003466   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.31026   3rd Qu.: 0.598201   3rd Qu.: 0.4659   3rd Qu.: 0.6598
##  Max.   : 5.81306   Max.   : 5.002541   Max.   : 5.8797   Max.   : 4.0611
##  diabetes
##  neg:500
##  pos:268
```



本文采用7-3的比例分割训练集和测试集，为了确保随机性，不能简单的取前70%的数据为训练集，应对每个数据进行一次p=0.7的伯努利实验，若结果为1该数据分配至训练集，否则分配至测试集。

4 模型训练

本文训练三种模型，下面分别简介其原理

4.1 KNN模型

对于一个待预测的样本，在特征空间中，找到与它最邻近的K个已知标签的样本（即“近邻”），然后通过这K个邻居的标签来“投票”决定该样本的标签。对于分类问题采用“多数表决”原则。即K个邻居中，哪个类别的样本最多，就将待预测样本归为那个类别。K是一个用户定义的常数。K值过小（如K=1）：模型会变得非常复杂，容易受到噪声数据的干扰，即过拟合。模型的边界会变得崎岖不平。K值过大：模型会变得过于简单，可能会忽略数据中一些有用的模式，导致欠拟合。模型的边界会变得平滑。通常通过交叉验证来选择一个合适的K值，本文采用k=10，距离则采用欧氏距离。值得注意的是KNN方法对特征的量纲敏感，因此进行前需要进行数据标准化。

其优势主要为：1.直观易懂，原理简单。KNN的逻辑非常符合人类的直觉，不需要复杂的数学背景就能理解，非常适合作为机器学习的入门算法。2.无需训练/训练时间极短。KNN是一种“惰性学习”算法。它实际上并不从训练数据中学习一个明确的模型或规则，而只是把所有的训

练数据存储起来。所谓的“训练”过程只是把数据记下来，因此速度非常快。3.对数据分布没有假设。不同于很多算法（如线性回归、朴素贝叶斯）需要对数据的分布做出假设，KNN是一种非参数方法，它不预设任何数据分布形式，因此可以用于解决非常复杂的问题。4.在多分类问题上表现良好。无论是二分类还是多分类问题，KNN都能直接应用，无需像SVM等算法那样进行改造。

4.2 LDA模型

LDA的核心思想非常直观：“投影与类间分离”。它试图找到一个或多个线性组合（即新的坐标轴），将高维空间中的数据点投影到一条（或几条）低维的直线上。这个投影的目标是，使得不同类别的数据点在投影后的空间里，它们的中心（均值）尽可能互相远离；同时，同一类别的数据点在投影后的空间里尽可能聚集。简单来说，LDA就像在找一个最佳的“观察角度”，从这个角度看过去，不同类别的数据“团”被分得最开，而每个类别内部的数据点则抱得最紧。首先我们通过最大化瑞利商，即

$$J(w) = \frac{w^T * S_B * w}{w^T * S_W * w}$$

其中w是我们要求的最佳投影方向，$S_W S_B$分别代表类内散度矩阵（对于每个类别，计算其所有样本点投影后的值与该类中心投影后的值的差异的平方和）和类间散度矩阵（两个类别中心在投影后距离的平方）。接下来将待分类的样本x 投影到我们找到的判别方向上，得到它在低维空间（称为“LDA空间”）中的坐标。然后在低维空间中，使用一个简单的分类器（如最近中心分类器）进行分类。如计算每个类别的投影中心，并将待分类样本分配给投影后距离最近的那个类别的中心。

其优势主要为1.原理简单，计算高效。2.考虑了类别信息，降维后的特征通常对分类任务更有效。3.对线性可分或近似线性可分的数据效果很好。4.由于假设各类具有相同的协方差矩阵，能减少过拟合风险。

4.3 逻辑回归模型

它首先用一个线性函数（类似于线性回归$z = w^T x + b\Psi$来组合输入特征。然后，将这个线性函数的输出通过一个特殊的“Sigmoid函数”，将值压缩并映射到(0,1)区间内，这个值就被解释为“属于正类的概率”。我们通常会设定一个阈值（默认为0.5）。如果P(y=1| x) 大于等于0.5，则预测为类别1。如果P(y=1 | x) 小于0.5，则预测为类别0。

其优势主要为：1.输出有概率意义。不仅给出分类结果，还给出置信度，这在很多场景（如风险评估）中非常重要。2.计算效率高。训练和预测的速度都很快，适用于大规模数据集。3.可解释性强。权重w 的大小和正负可以直接解释为特征对结果的影响程度和方向。4.实现简单，易于部署。5.对线性可分或近似线性可分的问题效果很好。

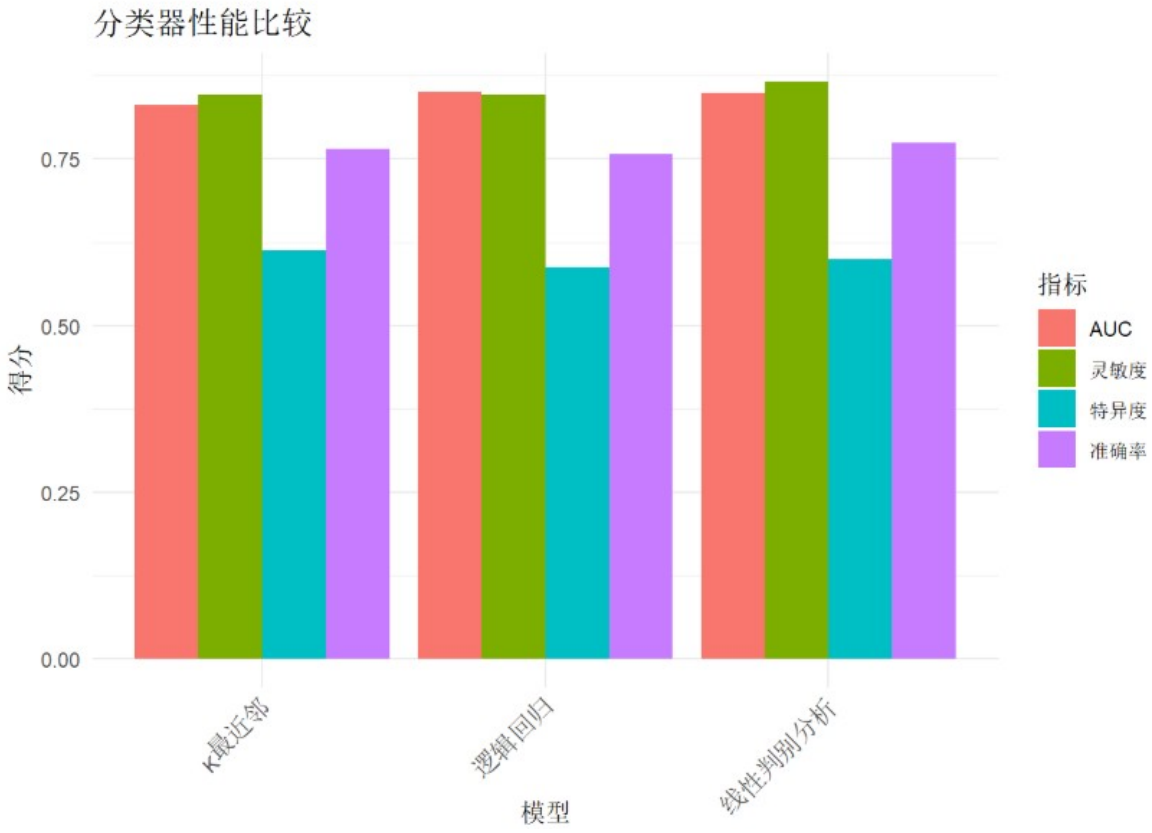
4.4 模型训练的控制参数

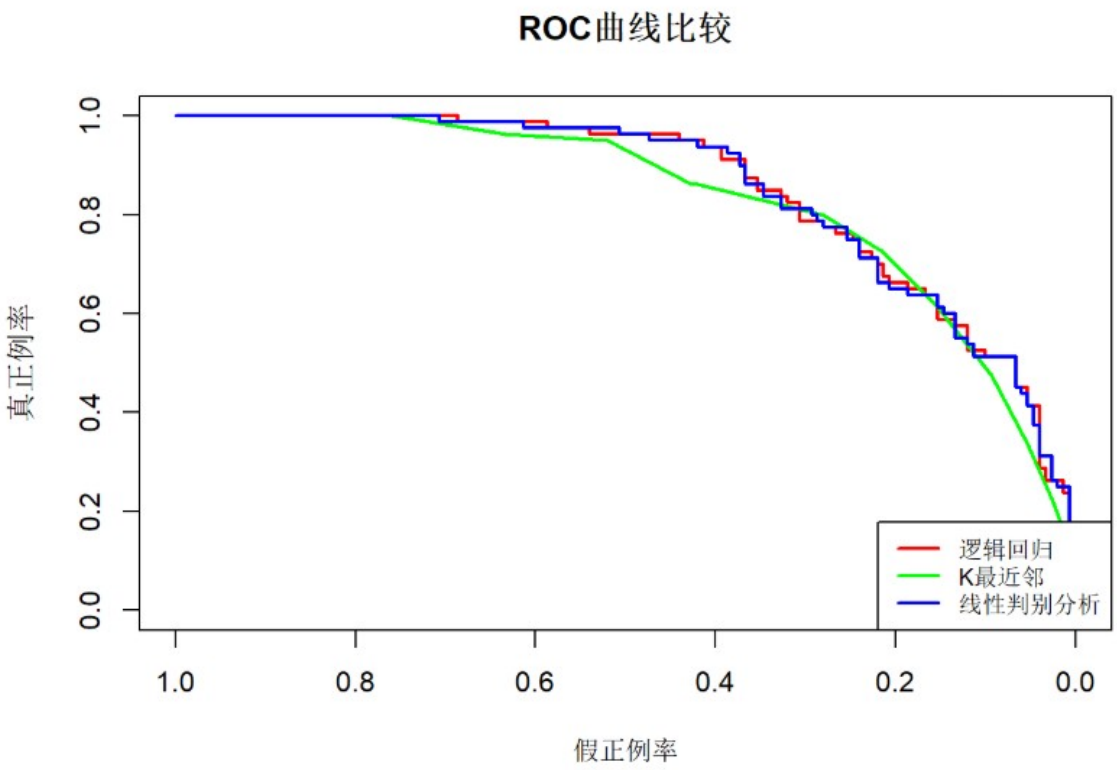
本文指定重采样方法为“交叉验证”（cross-validation）。交叉验证是一种常用的模型评估方法，它将训练数据分成多个部分，轮流将一部分作为验证集，其余作为训练集，多次训练和验证以评估模型性能。指定交叉验证的折数为10，即10折交叉验证。这意味着将训练数据随机分成10个子集，每次使用9个子集训练模型，1个子集验证模型，重复10次，每次使用不同的子集作为验证集。最终性能是10次验证结果的平均值。

在模型训练过程中计算并保存每个样本属于各个类别的概率。对于分类问题，设置为TRUE可以生成类别概率，这对于后续计算ROC曲线等指标是必要的。

5 性能比较

我们从准确率、灵敏度、特异度、AUC等多个维度比较分类器的性能





最终我们选择AUC最高的逻辑回归模型作为性能最优的分类器。

KNN表现不佳的可能原因

1.对高维数据效果不佳——维度灾难：在高维空间中，所有点之间的距离都会变得非常相似，导致“最近邻”的概念变得模糊，难以找到真正有意义的邻居，从而使得模型性能下降。本文采用的数据维度为8，维度较高。

2.对不平衡数据敏感：如果某个类别的样本数量远多于其他类别，那么在投票时，该类别就会有天然的优势，导致对新样本的预测更倾向于这个多数类，从而影响少数类的准确率。本文采用的数据中患病样本约为健康样本的两倍，可能导致对新样本的预测更倾向于患病。

LDA表现不佳的可能原因

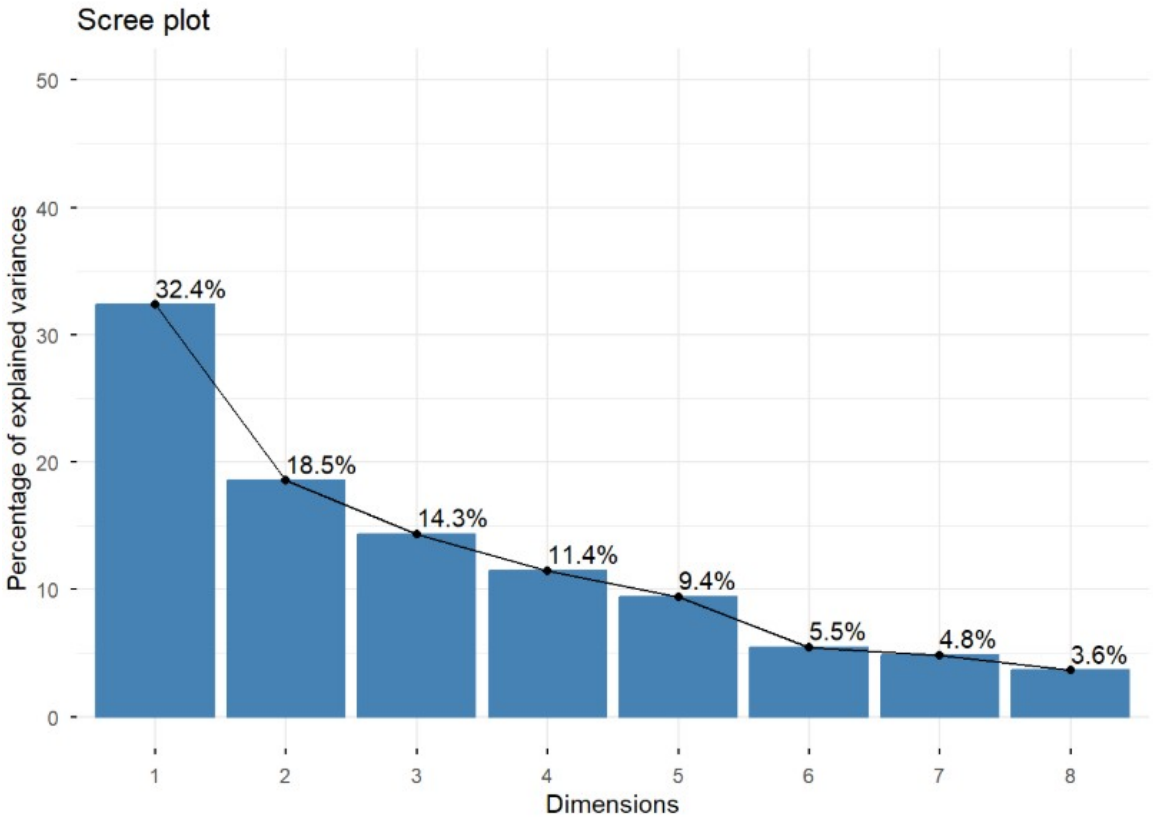
1.LDA理论基于数据服从正态分布的假设，在实际应用中可能不总是成立。

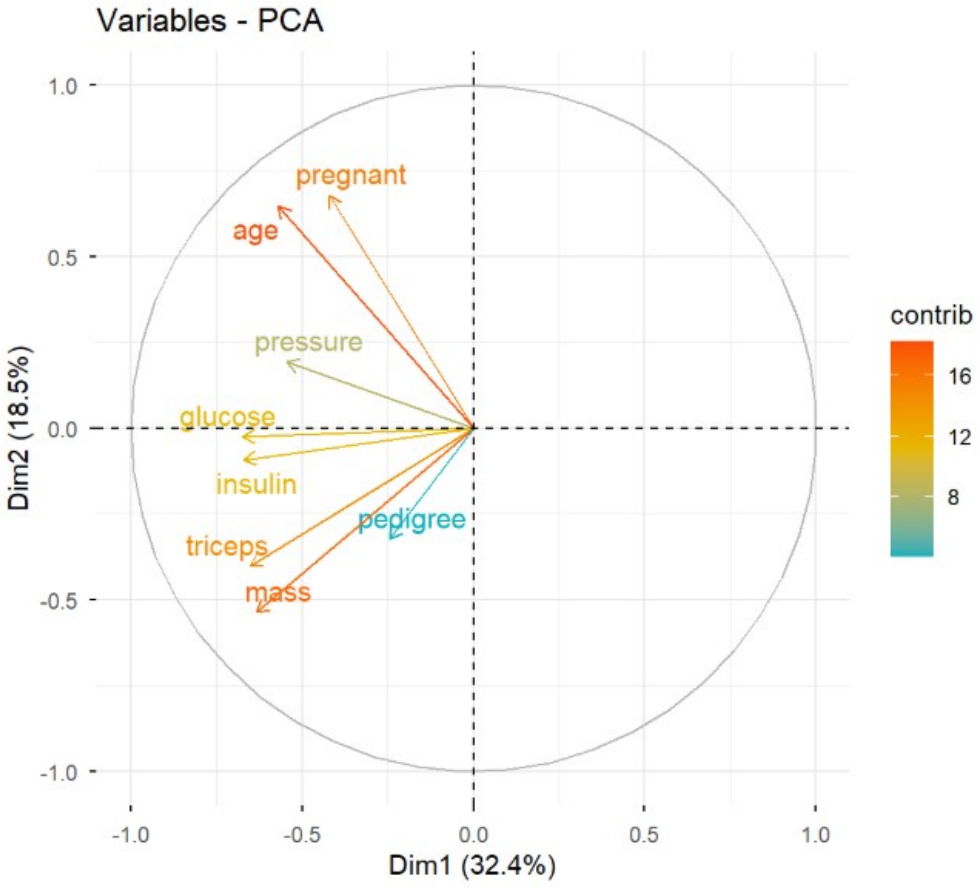
2.LDA理论要求所有类别的协方差矩阵大致相同。如果不同类别的数据分布形状差异很大，LDA的性能会下降。

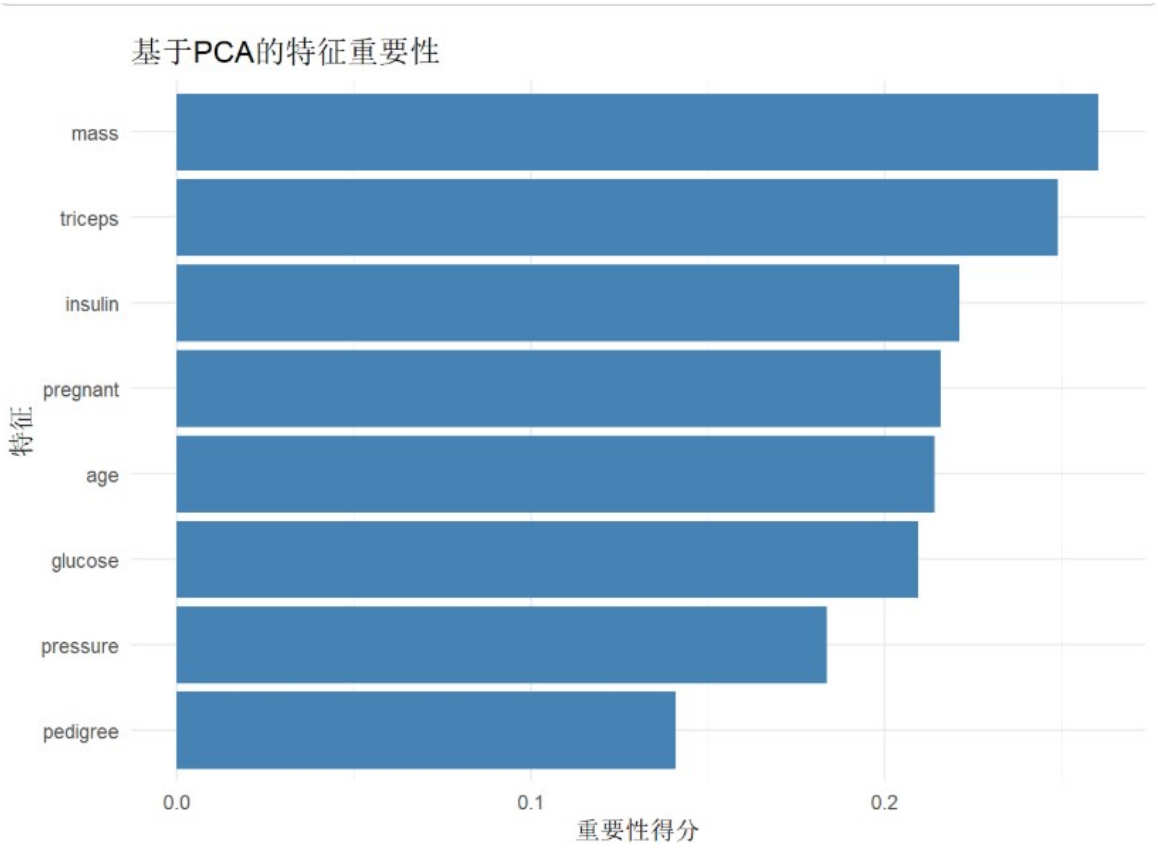
6　主成分分析

对数据进行主成分分析。PCA的核心思想为将原始数据投影到一组新的正交基上（称为主成分），使得投影后的数据方差最大化，从而用更少的特征保留尽可能多的信息。

步骤包括：1.中心化。将每个特征减去其均值，使数据均值为0。2.计算协方差矩阵以反映特征间的相关性。3.计算协方差矩阵的特征值和特征向量。4.将特征值从大到小排序，选择前k个最大的特征值对应的特征向量作为主成分。5.将原始数据投影到选定的主成分上，得到降维后的数据。


Scree plot
Percentage of explained variances
50
40
30
20
10
0
32.4%
18.5%
14.3%
11.4%
9.4%
5.5%
4.8%
3.6%
1
2
3
4
5
6
7
8
Dimensions

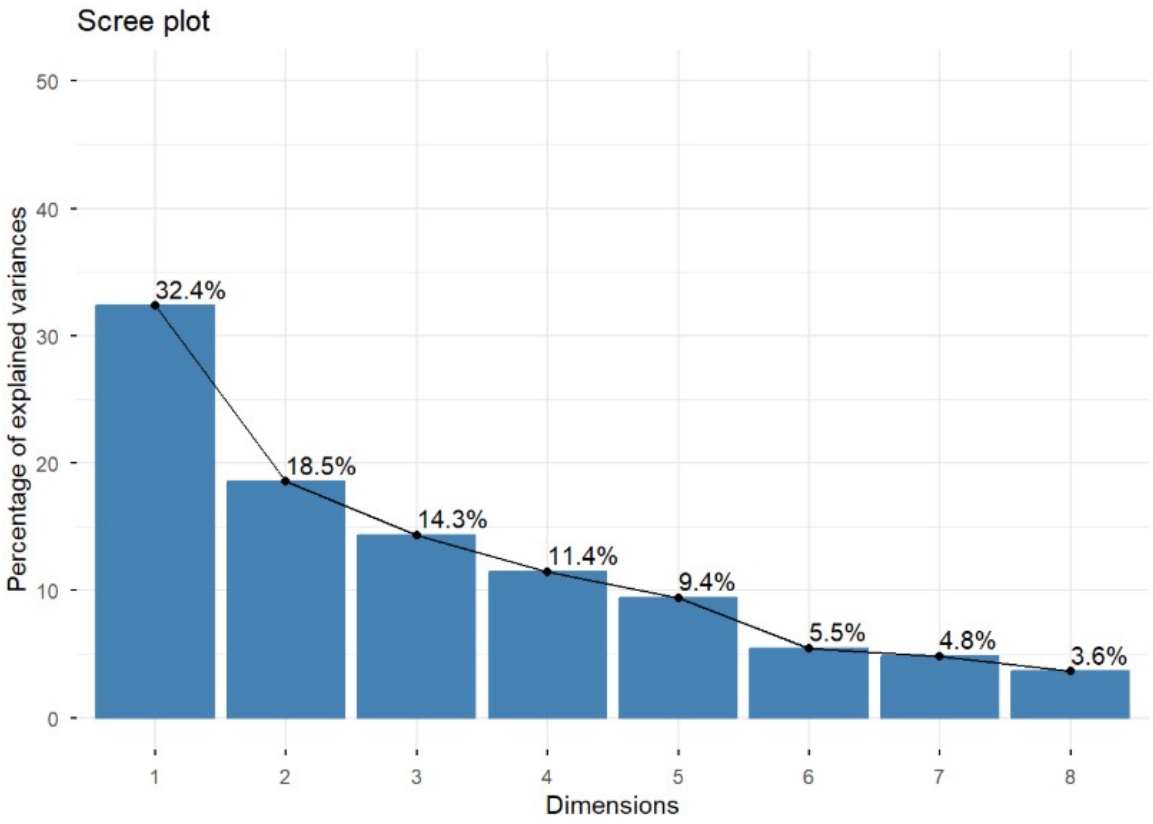

Variables - PCA
Dim2 (18.5%)
1.0
0.5
0.0
-0.5
-1.0
pregnant
age
pressure
glucose
insulin
pedigree
triceps
mass
-1.0
-0.5
0.0
0.5
1.0
Dim1 (32.4%)
contrib
16
12
8



可以看到重要特征排名：1．mass (重要性得分: 0.26)2．triceps (重要性得分: 0.249)3．insulin (重要性得分: 0.221)，即数据变异主要由身体成分（BMI、体脂）和代谢指标（胰岛素）驱动。值得注意的是从后面的分析中我们可以看到这并不是糖尿病预测中最重要的三个因素，这是因为PCA是无监督的降维，目的是最大化方差（数据变异性），并不能起到对标签预测重要性进行判断的功能。

如果我们将LDA(线性判别分析)与PCA(主成分分析)进行比较我们可以发现：前者的核心目标为有监督的降维，最大化类别可分性，后者则是无监督的降维，最大化方差；前者使用类别标签来寻找方向后者则不使用类别标签；前者的优化目标为最大化类间散度与类内散度的比值，后者则是最大化投影后的方差；前者的主要用途是分类前的特征降维和可视化，后者则是探索性数据分析、特征压缩、去噪。如果用一个形象的比喻概括则是：PCA是帮你找一个最好的角度来看清楚整个数据的整体结构和分布（哪个方向跨度最大）。LDA则是帮你找一个最好的角度来把不同颜色的弹珠分得最开（已知弹珠有不同颜色）。

7 预测因子重要性分析



因为LA模型中采用了一个线性函数（类似于线性回归$z = w^T x + b\Psi$来组合输入特征，因此该线性回归模型summary中的参数体现了预测因子的重要性：

1.系数估计（Estimate）。数值大小：系数的绝对值越大，表示该预测因子对响应变量的影响越大。符号方向：正号表示正向影响，负号表示负向影响。

2. 统计显著性（p-value）。p值小于0.001：高度显著，因子非常重要。小于0.01：很显著，因子比较重要。小于0.05：显著，因子有一定重要性。大于0.05：不显著，因子可能不重要。

3. t值（t value）t值解读：绝对值越大，说明该预测因子的效应越显著。经验法则：—t—大于2 通常表示统计显著。可用于比较不同预测因子的相对重要性。

可知葡萄糖水平是最重要的糖尿病预测因子，BMI和年龄也是关键风险因素，胰岛素水平和糖尿病血统函数提供了额外信息，血压和三头肌皮褶厚度的预测价值相对较低。

8 总结

本文在处理完数据后训练了KNN、LDA、LA三种分类器，通过对AUC的比较得到表现最好的分类器为LA分类器，并分析了其他两种分类器表现不佳的可能原因。然后进行PCA分析，得知数据变异主要由身体成分（BMI、体脂）和代谢指标（胰岛素）驱动。最后进行预测因子重要性分析，得知葡萄糖水平是最重要的糖尿病预测因子，BMI和年龄也是关键风险因素。