基于 PSO-XGBoost 的糖尿病预测模型

摘要

本文针对题目所提供体检数据集，研究并构建通过体检指标来预测血糖的模型、糖尿病风险的评估模型，期望对糖尿病进行科学有效的干预、预防和治疗，来降低发病率和提高患者的生活质量。

在数据的预处理方面，我们团队进行整体数据集的概述，对缺少值和异常值进行分析和选择，同时利用中间值填充法和高斯混合模型来进行数据的完善和舍去。

对于问题一，在进行数据预处理后，我们团队首先利用 Spearman 来分析缺失率过大的指标，排除乙肝指标的影响。然后，我们团队以 CART 回归树为基学习器的 XGBoost 集成学习算法，再依据 r^2-score 进行特征逐步消除。最后，选出了十项指标作为特征值：**'* 天门冬氨酸氨基转换酶', ' 尿酸', ' 年龄', ' 性别', ' 甘油三酯', ' 红细胞体积分布宽度', ' 红细胞平均体积',' 红细胞计数', ' 血小板平均体积', ' 血红蛋白'**。

对于问题二，通过对智能群优化算法的比较和选取，我们团队在问题一的模型基础上引入了**智能群优化算法**中的粒子群算法来针对数据集进行 **XGBoost 算**法参数的调整，并通过与其他预测模型（如'SVM'、'BP'、'KNN'、'Linear Regression'）进行比较，来验证我们团队模型的合理性。

对于问题三，我们团队根据聚类分析将血糖分成了四类，分别为患糖尿病低风险、患糖尿病中低风险、患糖尿病中高风险和患糖尿病高风险。我们采用了 K-means++ 算法对糖尿病评估模型进行训练，并分别绘制了 K-means++ 的肘部图和 K 值的轮廓分析图进行分析：**血糖 3.07-5.4 为患糖尿病低风险，5.04-7.04 为患糖尿病中低风险，7.05-10.45 为患糖尿病中高风险，10.45 以上为患糖尿病高风险**，与目前医学对糖尿病血糖的界定相符。最后，结合第二问的模型，具体说明体检数据对糖尿病风险评估的影响。

对于问题四，我们团队用血糖和体检数据的模型来对附件 2 中数据进行血糖预测，用血糖和糖尿病的模型做出评估。最后，计算出附件 2 的结果比例：**正常血糖为 42/141，具有患糖尿病风险为 74/141，糖尿病为 20/141，严重糖尿病为 5/141**。

我们团队将**群智能优化算法与梯度提升算法相结合**，克服了传统算法精度低、抗噪声能力弱和无法解决特征明显的结构化数据的局限性，提高了血糖预测的合理性和精度。同时，我们对聚类模型的类别数进行多方面检验，选取最合适的类别数来评估糖尿病风险，使结果符合医学标准。最后，我们建立了体检数据和糖尿病风险评估的关系，并利用血糖预测模型和评估模型进行双向构建，通过集成模型使结果更加符合医学实际。

关键字： XGBoost 算法　PSO 算法　智能群优化算法　Spearman　K-means++ 算法　CART 回归树　糖尿病预测　高斯混合模型

目录

一、 问题的重述和分析 1
1.1 问题重述 1
1.2 问题分析 1
二、 模型假设和符号说明 2
2.1 模型假设 2
2.2 符号说明 2
三、 模型的建立和求解 4
3.1 问题一的建模和求解 4
3.1.1 模型的建立 4
3.1.2 模型的求解 4
3.1.3 问题的分析 4
3.2 问题二的建模和求解 4
3.2.1 模型的建立 4
3.2.2 模型的求解 4
3.2.3 问题的分析 4
3.3 问题三的建模和求解 4
3.3.1 模型的建立 4
3.3.2 模型的求解 4
3.3.3 问题的分析 4
3.4 问题四的建模和求解 4
3.4.1 模型的建立 4
3.4.2 模型的求解 4
3.4.3 问题的分析 4
四、 模型的推广和评价 4
4.1 模型的优点 4
4.2 模型的缺点 4
4.3 模型的推广 4
五、 参考文献 5
..... 5
A 附录 附录标题 6

一、 问题的重述和分析

1.1 问题重述

糖尿病是一种代谢性疾病，其特征是患者的血糖长期高于标准值。胰腺无法产生足够的胰岛素或人体无法有效利用所产生的胰岛素时，就会出现糖尿病。糖尿病的临床表现包括频尿、口渴和饥饿感。同时伴随并发症如心血管疾病、中风、慢性肾脏病和足部溃疡等。根据 2021 年 IDF 发布的数据，全球成年糖尿病患者人数达到 5.37 亿（10.5%），约十分之一的成年人受到影响。在过去的 10 年间，中国糖尿病患者人数增加了 56%，其中约 7283 万名患者尚未被确诊，比例高达 51.7%。糖尿病种类主要分为 1 型糖尿病、2 型糖尿病、妊娠糖尿病和其他类型糖尿病。作为一种常见的慢性疾病，糖尿病目前无法根治，需要通过科学有效的干预、预防和治疗，来降低发病率和提高患者的生活质量。

附件 1 和 2 分别给出了有血糖值的检测数据和无血糖值的检测数据，包含年龄、性别、各项体检数据等 42 个监测指标，包含数值型、字符型、日期型等数据类型。题目需要我们解决以下四个问题:

问题 1：根据附件 1 的检测数据，从 42 个检测指标中筛选出主要变量指标，并说明筛选过程及其合理性。

问题 2：根据附件 1 的检测数据，建立血糖值的预测模型。

问题 3：根据附件 1 的检测数据，对糖尿病的风险进行评估。

问题 4：根据附件 2 的检测数据，对血糖值进行预测和评估糖尿病风险。

1.2 问题分析

首先，要对于数据集进行预处理，对于其中异常值、缺少值的数据进行删除或填充。然后，考虑不同生理指标和血糖值的相关性，并对其重要性进行排序。最后，筛选出相关性强的生理指标作为特征值。

在第一问得出的特征值，对数据预处理后的数据集来建立特征值和血糖的关系。通常可以采取：支持向量机回归、决策树回归、随机森林回归、逻辑回归等多种回归模型来拟合特征值和血糖的关系。

根据题目文本提供的信息，正常血糖值是指人在空腹状态下血糖值在 3.9～6.1 毫摩尔/升之间。然而，要判断是否存在高血糖，一般需要对人体进行两次重复测量，若两次测量的平均值大于 6.7 毫摩尔/升，则可以诊断为糖尿病。针对不同程度的血糖值，我们团队可以制定一个简单的分类标准，并利用聚类的思想，找出不同程度的血糖值对应患糖尿病的概率风险。此外，我们还可以通过分析血糖和体检数据之间的关系，评估体检数据对患糖尿病风险的影响。

在前三问的基础上，可建立出血糖的预测模型和血糖对应患糖尿病概率的评估模型，只需要将附件 2 的数据集利用两种模型来处理，即可得出结论。

二、 模型假设和符号说明

2.1 模型假设

图 **1**　**sym of whut**

(1) 这是第一点；
(2) 这是第二点；
(3) 这是第三点。

表 **1**　表格标题

1	2	3
4	5	6
7	8	9

2.2 符号说明

$$f(x) = \begin{cases} x, & x > 0, \\ -x, & x \leq 0. \end{cases}$$

$$\begin{aligned} a &= b + c \\ &= d + e \end{aligned}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

a+1
$\left(1+\frac{1}{n}\right)^n$
$(2+a)$
$a^{\frac{1}{n}}$
$\frac{a}{b}$

$$a^n$$

$$a_1$$

$$a^{n+1}$$

$$a_{n-1}$$

若 $a>0, b>0$, 则

$$a+b>0.$$

若 $a>0, b>0$, 则 $a+b>0$.

三、 模型的建立和求解

3.1 问题一的建模和求解

3.1.1 模型的建立

3.1.2 模型的求解

3.1.3 问题的分析

3.2 问题二的建模和求解

3.2.1 模型的建立

3.2.2 模型的求解

3.2.3 问题的分析

3.3 问题三的建模和求解

3.3.1 模型的建立

3.3.2 模型的求解

3.3.3 问题的分析

3.4 问题四的建模和求解

3.4.1 模型的建立

3.4.2 模型的求解

3.4.3 问题的分析

四、 模型的推广和评价

4.1 模型的优点

4.2 模型的缺点

4.3 模型的推广

五、 参考文献

[1] 作者. 文献[M]. 地点：出版社，年份。

[2] 作者. 文献[M]. 地点：出版社，年份。

附录 A　附录标题

这里是附录。