基于网格化搜索的 Adaboost 血糖值预测模型

摘要

通过对多种指标数据的综合分析，实现对血糖值的准确预测具有重要意义，这一方法有助于有效预防糖尿病的发生。本文通过建立基于 Adaboost 的血糖值预测模型和糖尿病风险关联模型，最终成功预测了附件二数据的血糖值。

对于问题一，对数据的预处理包括**基本处理**以及**异常值处理**和**缺失值处理**三个步骤，而统计分析则分为涵盖离散和连续量的**显著性分析**，以及所有指标的**描述性统计**，发现各特征指标单位不同，均值和方差相差甚远。

针对问题二，首先，将各项指标按照医学领域划分为五大类。然后，对各指标**正向化**和**标准化处理**。接着，计算每个分类中各指标间的**相关性**和**显著性 p 值**，选择彼此独立的指标来确定每个分类中有代表性的典型指标，利用**典型相关分析**以此对数据进行降维。最终得到了**九个**具有代表性，独立性和可解释合理性的指标，包括 ALT**(丙氨酸氨基转换酶)**、GGT**(谷氨酰基转换酶)**、GLB**（球蛋白）**、RBC**（红细胞计数）**、EBC**(白细胞计数)**、PLT**(血小板计数)**、LDL_C**(低密度脂蛋白胆固醇)**、HDL_C**(高密度脂蛋白胆固醇)**、USG**(尿素)**。

对于问题三，首先，用 **shuffle 法**将附件一中的数据分为训练集、交叉验证集和验证集三部分。然后，建立 **Adaboost 血糖值预测模型**，使用**网格化搜索**确定超参数的最优组合。接着，利用预处理后的特征数据带入其中得到血糖值预测数据。最终，最优模型的 MSE、RMSE、MAE 和 MAPE 四项指标分别为 **0.002**、**0.045**、**0.024**、**28.629**。

对于问题四，首先，运用问题三建立的模型对样本进行双重预测，并将平均值作为最终血糖值，从而构建**基于 Adaboost 的血糖预测模型**。接着，创建**基于 Adaboost 的糖尿病风险预测模型**，将血糖值分为正常、高血糖和异常三个区间，并赋予对应的 0、1、2 标签。因此，将预测得到的异常血糖区间概率作为衡量糖尿病风险的量化标准。最终，根据附件二的数据，发现 USG**（尿素）**是响应血糖值的**快速指标**，**肾功能指标**是衡量糖尿病患风险的**重要标准**。

最后，通过对问题三模型中的 ALT(丙氨酸氨基转换酶) 扰动 **1%**,**5%** 和 **10%** 时, 发现扰动百分比分别为 **3.74%**、**8.06%**、**9.21%**, 模型具有较好的鲁棒性。同时，也对模型的优缺点进行了评价。

关键字：网格化搜索　典型相关分析　基于 Adaboost 的血糖预测模型　基于 Adaboost 的糖尿病风险预测模型

一、 问题重述

1.1 背景资料

当人体胰岛素分泌不足或细胞对胰岛素反应异常时，就可能导致糖尿病。根据国际糖尿病联合会数据，全球成年糖尿病患者达 5.37 亿，我国糖尿病患者数量更是高达 1.4 亿，且有 51.7% 未被确诊。持续的高血糖环境可导致多系统器官损伤，包括心血管系统疾病、神经性病变以及肾功能障碍等复杂并发症，从而严重影响患者的生命质量和预期寿命。因此，通过检测数据信息来分析和预测血糖动态变化是至关重要的，可以为糖尿病的早期筛查和诊断提供重要依据。

1.2 需要解决的问题

（1）对附件一所给数据进行预处理后做相关的统计分析

（2）基于问题（1），为使 42 个检测指标更有代表性和独立性，对其降至 10 维以下，得出具有代表性和独立性的主要变量指标并说明变量筛选过程及合理性。

（3）基于问题（2）得到的主要变量指标，用数据挖掘技术建立血糖值预测模型，并对其验证。

（4）基于问题（3）的模型，预测附件二数据的血糖值，说明糖尿病的风险。

二、 问题假设

1. 所有检测项目均符合规范，记录的血糖数据为人体空腹情况下的数据。
2. 血糖数据可以正确反映人体的血糖含量情况。判断是否有高血糖的一般方法需要对人体血糖进行二次重复测量，而本文数据集只提供了一次的数据集未经过二次验证。
3. 假设样本中人群均未患有其他心血管疾病。

三、 符号说明

表 1　指标名称

符号	符号说明	符号	符号说明	符号	符号说明	符号	符号说明
GRA	中性粒细胞%	BAS	嗜碱细胞%	$MCHC$	红细胞平均血红蛋白浓度	MPV	血小板平均体积
KB_e	乙肝 e 抗体	EOS	嗜酸细胞%	EBC	白细胞计数	PDW	血小板体积分布宽度

$HBeAg$	乙肝 e 抗原	USG	尿素	ALB	白蛋白	CR	肌酐
K_HBc	乙肝核心抗体	UA	尿酸	RDW	红细胞体积分布宽度	PCT	血小板比积
K_HBs	乙肝表面抗体	Age	年龄	HCT	红细胞压积	PLT	血小板计数
$HBsAg$	乙肝表面抗原	Sex	性别	MCV	红细胞平均体积	LDL_C	低密度脂蛋白胆固醇
GLU	血糖	$TCHO$	总胆固醇	AG	白球比例	HGB	血红蛋白
TP	总蛋白数	LYM	淋巴细胞%	MCH	红细胞平均血红蛋白量	HDL_C	高密度脂蛋白胆固醇
MO	单核细胞%	TG	甘油三酯	RBC	红细胞计数	GLB	球蛋白

四、 问题一模型的建立与求解

4.1 问题分析

对附件一数据进行预处理[1] 包括三个主要步骤：基本处理（文本编码、数据标签化、冗余数据删除）、血糖值正常与否的判断以及缺失值处理。而统计分析分为显著性分析（离散量分析、连续量分析）和描述性统计（不显著相关性指标、显著相关性指标）。

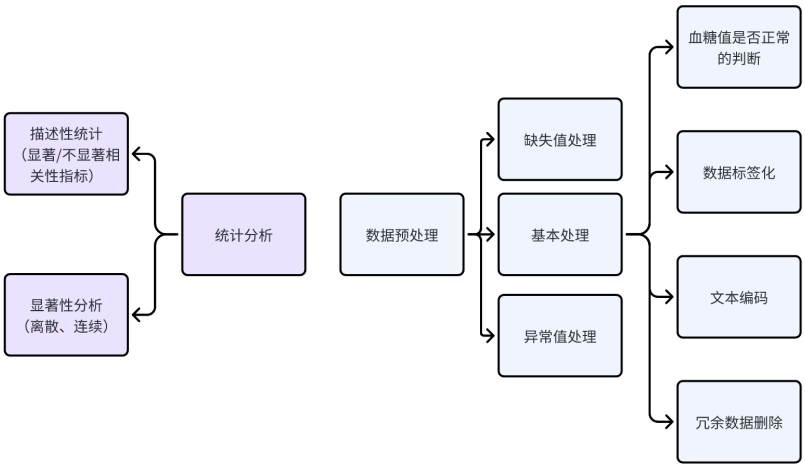

图 1　问题一流程图

4.2 数据预处理

1. 数据的基本处理

（1）文本指标编码处理：

为书写方便，将附件中出现的指标名字改成英文缩写（详细见表 1）。

（2）文本数据的标签化处理：

将男性标记为 1，女性标记为 2。

（3）冗余数据的处理：

鉴于检测时间仅分布在 2017 年 9 月 25 日至 2017 年 10 月 31 日之间，跨度较小，因此在后续分析中可将其忽略。而 id 编号没有实际意义，也忽略。

（4）血糖值正常与否的判断

根据题意，人在空腹时测得血糖值在 3.9~6.1 毫摩尔/升为正常（用 1 表示），所以可将空腹时所测血糖值在低于 3.9 毫摩尔/升记为低血糖不正常和高于 6.1 毫摩尔每升都记为异常（用 0 表示）。于是，可得出附件一中血糖值正常的有 4396 个，异常的有 969 个。

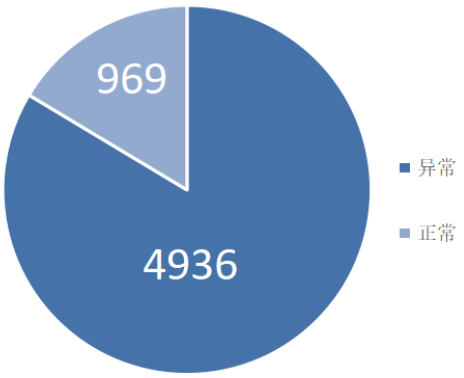

图 2　血糖值正常/异常数量

2. 数据的异常值处理

通过箱线法分析，可发现样本中存在血糖值含量为 38 的异常值。查阅文献显示，该值已远超人体所能承受的极限，不符合实际。通过剔除异常值样本，获得了最终的有效数据总数为 5904 个。

3. 数据的缺失值处理

（1）所有指标缺失值占比排序

按顺序计算每个指标的缺失值占比如表 2。

表 2　每个指标的缺失值占比

指标名	缺失值占比	指标名	缺失值占比	指标名	缺失值占比	指标名	缺失值占比
*r-谷氨酰基转换酶	20.89754	乙肝表面抗体	76.08806	总胆固醇	20.77900	红细胞计数	0.33870
* 丙氨酸氨基转换酶	20.89754	乙肝表面抗原	76.08806	淋巴细胞%	0.33870	肌酐	23.48857
* 天门冬氨酸氨基转换酶	20.89754	低密度脂蛋白胆固醇	20.77900	甘油三酯	20.77900	血小板体积分布宽度	0.45724
* 总蛋白	20.89754	体检日期	0	白球比例	20.89754	血小板平均体积	0.45724
* 球蛋白	20.89754	单核细胞%	0.33870	白细胞计数	0.33870	血小板比积	0.45724
* 碱性磷酸酶	20.89754	嗜碱细胞%	0.33870	白蛋白	20.77900	血小板计数	0.33870
id	0	嗜酸细胞%	0.33870	红细胞体积分布宽度	0.33870	血糖	0.94835
中性粒细胞%	0.33870	尿素	23.48857	红细胞压积	0.33870	血红蛋白	0.33870
乙肝 e 抗体	76.08806	尿酸	23.48857	红细胞平均体积	0.33870	高密度脂蛋白胆固醇	20.77900
乙肝 e 抗原	76.08806	年龄	0	红细胞平均血红蛋白浓度	0.33870		
乙肝核心抗体	76.08806	性别	0.01693	红细胞平均血红蛋白量	0.33870		

由表 2，只有乙肝类指标数据存在显著的缺失现象（缺失值占比高达 50% 以上）。为确保数据分析的可靠性和准确性，后续的分析将数据划分为两组：一组为所有含有乙肝类指标的数据，另一组为剔除了乙肝类指标后的数据。

（2）性别缺失值的填充

附件一中的总样本数为 5904，其中包括 2996 名男性和 2907 名女性。值得注意的是，编号 573 的性别未知。考虑到数据分布，使用众数填充该缺失值，即假设其性别为男性。

4.3 统计分析

1. 数据的显著性分析

鉴于附件一中存在离散数据（年龄、性别）和连续数据（除年龄、性别外的指标），这里对其分开讨论。

（1）离散数据的正态分布检验

①正态分布的介绍

正态分布[2]，又称高斯分布。若随机变量 X 服从一个位置参数为 σ 尺度参数为 μ 的正态分布，记为：

$$X \sim N(\mu, \sigma^2) \tag{1}$$

则其概率密度函数为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

其中，正态分布的期望值 μ 等于位置参数，决定了分布的位置；其标准差 σ 等于尺度参数，决定了分布的幅度。

②离散数据的正态分布检验将附件一中每个人的性别、年龄和血糖值利用 spsspro 进行正态分布检验，结果如表 3。

表 3　正态性检验的结果

变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
血糖值是否异常	5904	0	0.164	0.37	1.814	1.292	0.446(0.000***)	0.507(0)
年龄	5904	45	45.691	12.996	0.412	-0.27	0.979(0.000***)	0.067(9.163530439287569e-24)

注：***、**、* 分别代表 1%、5%、10% 的显著性水平

根据表 2, 年龄的 K-S 检验得到的显著性 P 值为 0.000***，表明在统计学上呈现出显著差异。

年龄数据不服从正态分布。

鉴于此，需要对这两类数据选择多因素方差分析方法，以评估它们与血糖值之间的显著性关系。

（2）离散数据的多因素方差分析

①多因素方差分析的介绍

Step1：模型的构建

对于年龄、性别的 n=2 个因素，分别有 r1=78，r2=2 个水平，在水平组合下样本相互独立且满足式（3）：

$$X_{i_1,i_2\sim N(\mu_{i_1,i_2,i_3},\sigma^2)},(i_1=1,\ldots,r_1;i_2=1,\ldots,r_2;i_n=1,\ldots,r_n) \tag{3}$$

所以，因素 n 的第 i_n 个水平效应为 α_{ni_n}，第 i_n 个水平下各观察值的平均值为：

$$\mu_{**i_1}=\frac{1}{r_1r_2\ldots r_{n-1}}\sum_{i_1=1}^{r1}\sum_{i_2=1}^{r2}\cdots\sum_{i_{n-1}=1}^{r_{n-1}}\mu_{i_1,i_2,\ldots i_n} \tag{4}$$

若 $\mu_{i_1,i_2,\ldots i_n}=\mu+\alpha_{1i_1}+\alpha_{2i_2}+\ldots+\alpha_{ni_n}$，则模型为：

$$\begin{cases} X_{ijk}=\mu+\alpha_{1i_1}+\alpha_{2i_2}+\ldots+\alpha_{ni_n}+\varepsilon_{i_1i_2\ldots i_n} \\ \sum_{i_1=1}^{r_1}\alpha_{i_1}=0,\sum_{i_2=1}^{r_2}\alpha_{i_1}=0,\ldots,\sum_{i_n=1}^{r_n}\alpha_{i_1}=0 \\ \varepsilon_{i_1i_2\ldots i_n}\sim N(0,\sigma^2),\text{各个}\ \varepsilon_{i_1i_2\ldots i_n}\ \text{相互独立} \end{cases} \tag{5}$$

Step2：对 2 个因素提出假设

为了检验这 2 个因素的影响，需要对 n=2 个因素分别提出以下假设：

$$\begin{cases} H_{01}:\alpha_{11}=\alpha_{11}=\ldots=\alpha_{1r_1},\text{年龄对血糖值没有显著影响} \\ H_{02}:\alpha_{21}=\alpha_{22}=\ldots=\alpha_{2r_2},\text{性别对血糖值没有显著影响} \end{cases} \tag{6}$$

Step3：构建检验的统计量

为了检验假设是否成立，需要分别确定检验因素的统计量。则统计量:

$$\begin{cases} \bar{X}=\frac{1}{r_1r_2\ldots r_n}\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_n=1}^{r_n}X_{i_1i_2\ldots i_n} \\ \bar{X}_{i_1i_2\ldots i_n}=\dfrac{1}{r_2\ldots r_n}\displaystyle\sum_{i_2=1}^{r_2}\cdots\sum_{i_n=1}^{r_n}X_{i_1i_2\ldots i_n},\ldots\bar{X}_{i_1i_2\ldots i_n} \\ \qquad\qquad=\dfrac{1}{r_1r_2\ldots r_{n-1}}\displaystyle\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_{n-1}=1}^{r_{n-1}}X_{i_1i_2\ldots i_n} \end{cases} \tag{7}$$

引用式（8）：

$$S_T=\sum_{i_1=1}^{r_1}\sum_{i_2=1}^{r_2}\cdots\sum_{i_n=1}^{r_n}(X_{i_1i_2\ldots i_n}-\bar{X})^2=S_E+S_1+S_2+\ldots S_n \tag{8}$$

则年龄所产生的误差平方和为：

$$S_1 = r_2 \dots r_n \sum_{i_1=1}^{r_1} (\bar{X}_{i_n**} - \bar{X})^2 \tag{9}$$

性别所产生的误差平方和为:

$$S_2 = r_1 r_3 \dots r_n \sum_{i_2=1}^{r_2} (\bar{X}_{*i_2*} - \bar{X})^2 \tag{10}$$

因此，可以得到随机误差平方和:

$$S_E = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} (X_{i_1 i_2} - \bar{X}_{i_n**} - \bar{X}_{*i_2*} - \bar{X} + \bar{X})^2 = S_T - S_1 - S_2 \tag{11}$$

Step4：检验规则

采用 F 检验，若 $F_n > F_{n-\alpha}(r_n - 1, df_E$ 则拒绝原假设，表示因素 n 的个水平下的效应由显著差异。

②多因素方差分析的结果

表 4　多因素方差分析的结果

项	平方和	自由度	均方	F	P	R^2	调整 R^2
年龄	230.263	76	3.03	2.417	0.000***		
性别	22.265	1	22.265	17.765	0.000***	0.47	0.463
误差	7301.714	5826	1.253		NaN		

由表 4 可知，对于变量年龄，从 F 检验的结果分析可以得到，显著性 P 值为 0.000***，水平上呈现显著性，对 GLU(血糖值) 有显著性影响，存在主效应。而对于变量性别，从 F 检验的结果分析可以得到，显著性 P 值为 0.000***，水平上呈现显著性，对 GLU(血糖值) 有显著性影响，存在主效应。

（3）连续数据的相关性分析

①Spearman 秩相关系数的介绍

Spearman 秩相关系数被定义成等级变量之间的皮尔逊相关系数。对于样本容量为 n 的样本，n 个原始数据被转换成等级数据，相关系数 ρ 为

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - y)}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - y)^2}} \tag{12}$$

原始数据依据其在总体数据中平均的降序位置，被分配了一个相应的等级。如表 5 所示。

表 5　降序等级表

变量 Xi	降序位置	等级 xi	变量 Xi	降序位置	等级 xi
0.6	6	6	1.2	3	3
0.8	5	5	2.3	2	2
1.2	4	4	18	1	1

实际应用中，变量间的连结是无关紧要的，于是可以通过简单的步骤计算 ρ(被观测的两个变量的等级的差值) 为：

$$\rho = 1 - \frac{6\sum {d_i}^2}{n\left(n^2 - 1\right)} \tag{13}$$

②Spearman 显著性 p 值

对连续的指标变量和血糖值进行 Spearman 相关性分析，得到显著性 p 值如表 6 所示。

表 6　连续指标的显著性 p 值

指标	显著性 P 值	指标	显著性 P 值	指标	显著性 P 值	指标	显著性 P 值	指标	显著性 P 值
GRA	0.00042	RDW	0.00000	PLT	0.00000	$TCHO$	0.00000	TP	0.00078
MO	0.84146	HCT	0.00000	HGB	0.00000	TG	0.00000	GLB	0.00051
BAS	0.07598	MCV	0.78057	PDW	0.00126	HDL_C	0.00000	ALP	0.00000
EOS	0.85382	$MCHC$	0.00000	MPV	0.05085	GGT	0.00000	AG	0.32789
LYM	0.00008	MCH	0.00000	PCT	0.00000	ALT	0.00000	ALB	0.66377
EBC	0.00000	RBC	0.00000	LDL_C	0.00000	ACT	0.00000	USG	0.00000
UA	0.05708	CR	0.00000	KB_e	0.70600	$HBeAg$	0.38400	K_HBc	0.39100
K_HBs	0.430	$HBsAg$	0.948						

由表 6，可观察到在 p 值小于 0.01 的显著性水平下，以下指标显示出了显著相关性：GRA、LYM、EBC、RDW、HCT、$MCHC$、MCH、RBC、PLT、HGB、PDW、PCT、LDL_C、$TCHO$、TG、HDL_C、GGT、ALT、ACT、TP、GLB、ALP、USG 以及 CR。同时值得注意的是，乙肝类指标均不显著，所以后续分析中不需要单独对含有乙肝类指标的数据分组。

2. 描述性统计

(1) 不显著指标的描述统计

对不具有显著相关性的指标进行描述性统计，对于每个指标，计算数字频数、最小值、最大值、平均值、标准偏差和方差等统计量。如表 7 所示是部分指标计算的结果

（详细见附录四）。

表 7　不显著指标的描述统计

指标	样本量	最大值	最小值	平均值	标准差	中位数	方差
MO	5905	23.2	3.1	6.859	1.564	6.7	2.446
BAS	5905	3.5	0	0.603	0.291	0.6	0.085
EOS	5905	22.5	0	2.039	1.696	1.6	2.877
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
K_HBc	1412	17.09	0	1.876	1.556	1.645	2.42
K_HBs	1412	42.49	0	7.031	8.259	3.27	68.208
$HBsAg$	1412	44.35	0	0.919	5.602	0.05	31.382

总的来看可以发现，各特征字段数量不一、单位不同，均值和方差相差甚远。

（2）显著指标的描述统计

对具有显著相关性的指标进行了描述性统计分析。对于每个指标，计算数字频数、最小值、最大值、平均值、标准偏差和方差等统计量。如表 8 所示是部分指标计算的结果（详细见附录五）。

表 8　显著指标的描述统计

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差
LYM	5904	88.5	14.4	56.734	7.774	56.734	60.433
EBC	5904	76.3	7.5	33.766	7.234	33.6	52.337
RDW	5904	21.06	2.8	6.591	1.608	6.39	2.586
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
ALP	5904	374.32	22.98	87.562	22.717	87.562	516.047
USG	5904	13.39	1.5	4.989	1.138	4.989	1.294
CR	5904	177.42	39.43	78.397	12.106	78.397	146.565

可以发现，各特征字段数量不一、单位不同，均值和方差相差甚远。

五、 问题二模型的建立与求解

5.1 问题分析

为了降维的方法具有更好的指导意义，首先将各项指标按照医学领域进行明晰划分，主要分为肝功能、肾功能、血常规、血脂以及乙肝五大类。然后，将各指标正向化后再进行标准化处理。最后，计算每一个分类中各指标间的相关性，选取彼此独立的指标，进而确定每个分类中确立有代表性的典型指标，从而对数据集进行降维。最后，解释了所选变量指标的合理性。

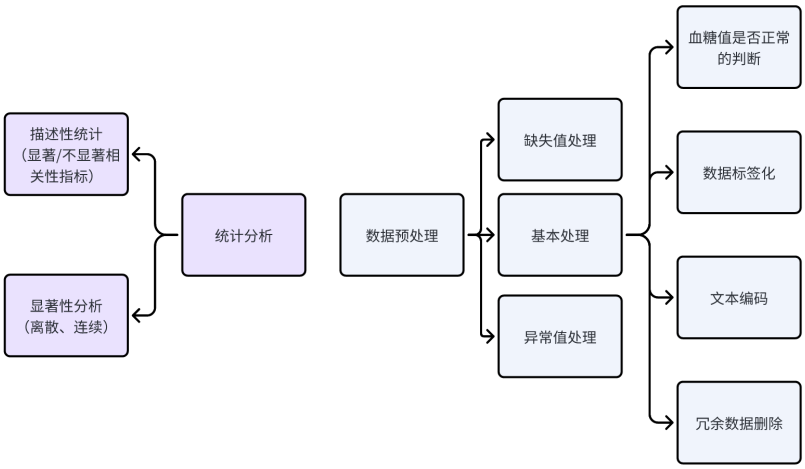

图 3　问题二流程图

5.2 数据预处理

1. 各指标的正向化和标准化处理

由于附件一中的医学指标有众多含义，首先需要依据其实际含义将高优指标和中间值指标正向化为低优指标。值得注意的是，尽管附件一展示的所有指标均不为百分比数据，但在分析其与血糖含量的相关性时，具体的数值并非主要关注点。因此对某一统计量 $X=(x_1,x_2,...,x_n)$，设该指标的正常值区间为 $[a,b]$ 直接采取如下方式正向化：

$$X=\begin{cases} b-X & \text{对于高优指标} \\ (b-a)-min\{0,|\frac{a+b}{2}-X|-\frac{b-a}{2}\} & \text{对于中间值指标} \end{cases} \tag{14}$$

然后，为了避免正向化后的指标间的数值差异对模型分析造成的影响，需要将这些数据置于相同尺度下，即进行标准化处理。

z-score 是以标准差为单位长度测量数据点与均值之间距离的值。在数据集经过 z-score 变换后其均值为 0，标准差为 1，并与原始数据集具有相同的形状属性。z-score 的计算方法为：

$$z\text{-}score=\frac{x-\mu}{\sigma} \tag{15}$$

其中 μ 为数据集的均值,σ 为数据集的标准差。重新缩放通过拉伸或压缩点来更改数据集中最小值和最大值之间的距离将数据集的值分布在 $[0,1]$ 之间。通过这样的缩放方式数据的 z-score 的特征会保留，其计算方式为:

$$x_{rescaled} = \frac{X - minX}{maxX - minX} \tag{16}$$

于是，可以得到各指标的正向化和标准化处理后的部分结果如表 9 所示（详细见附件 1）。

表 9　各指标的正向化和标准化结果

glucose	ALT	GGT	GLB	RBC	EBC	PLT	LDL_C	HDL_C	CR
0.08455	0.05924	0.01898	0.130270175	0.40700	0.86089	0.89546	0.29478	0.62272	0.74848
0.0656	0.09314	0.09942	0.35344	0.40700	0.73439	0.80551	0.29196	0.82272	0.67473
0.03478	0.02699	0.0189	0.26368	0.65498	0.64074	0.82576	0.63187	0.59545	0.86647
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.18297	0.05586	0.01422	0.53496	0.49865	0.70372	0.77066	0.46967	0.70454	0.86579
0.07579	0.07177	0.04527	0.36945	0.439	0.87732	0.77795	0.64033	0.40454	0.77329
0.06645	0.0330	0.01483	0.48222	0.55525	0.696	0.78930	0.25387	0.62272	0.8268

5.3 典型指标的选取

1. 检验指标的划分

查阅相关文献可知，题目中所给的医学指标可不重不漏地分类到如表 10 所示的六类检查项目中。

表 10　给定医学指标的检查项目分类

检查项目	检查内容	医学指标
肝功能检查	肝功能损伤	ALT,ACT,ALP,GGT,ALB,AG,GLB,TP
肾功能检查	肾功能损伤	USG,UA,CR
	红细胞项目	RBC,RDW,HGB,HCT,MCV,MCHC,MCH
血常规检查	白细胞项目	EBC,GRA,MO,BAS,EOS,LYM
	血小板项目	PDW,MPV,PCT,PLT
血脂检查	胆固醇含量	$LDL_C, HDL_C, TCHO, TG$

2. 各指标间的相关性

针对医学指标分类并选取与血糖浓度具有显著相关性的指标，进一步对每种检查项目中指标之间的相关性进行考察。为更直观地理解不同指标在检查项目中的关联情况，结果以热力图方式呈现如图 4。

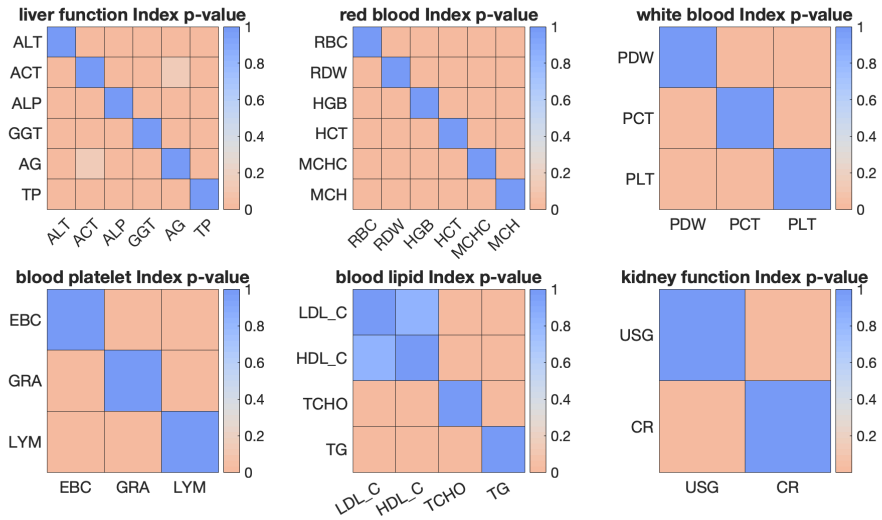

图 4　各变量指标的相关性 p 值热力矩阵图

3. 降维后得到的主要变量指标

基于相关性 p 值的分析，于每个检查项目中筛选具有较强独立性的指标。当两个指标之间的相关性 p 值较高（$p > 0.3$）时，视其为具有较强独立性。依此标准，共筛选出九个具有较强独立性的指标，选取结果如表 11 所示。

表 11　给定医学指标的检查项目分类

检查项目	检查内容	医学指标
肝功能检查	肝功能损伤	ALT　GGT　AG
	肾功能损伤	USG
血常规检查	红细胞项目	RBC
	白细胞项目	EBC
	血小板项目	PLT
血脂检查	胆固醇含量	LDL_C HDL_C

5.4 所选变量指标的合理性

使用典型相关分析最终确定属于五个医疗板块的九个医学指标，其合理性在于与因子分析相比，所选的变量指标更具有可解释性和独立性。

1. 因子分析的方差贡献率

使用因子分析法，通过碎石图辅助筛选变量因子确定了九个虚拟因子。其方差累计贡献率如表 12 所示。

表 12　累计方差贡献率

	旋转前方差解释率			旋转后方差解释率		
成分	特征根	方差解释率 (%)	累积方差解释率 (%)	特征根	方差解释率 (%)	累积方差解释率 (%)
1	4.497	18.739	18.739	300.156	12.506	12.506
⋮	⋮	⋮	⋮	⋮	⋮	⋮
8	1.123	4.681	73.618	143.431	5.976	72.408
9	1.056	4.4	78.018	134.629	5.61	78.018
10	0.886	3.692	81.71			
⋮	⋮	⋮	⋮			
24	0.001	0.003	100			

由表 12 可知，因子分析在保留多维数据的信息上较为优秀，在降维得到九维虚拟因子的情况下可以累计 78% 以上的方差解释率。

2. 虚拟因子的相关性矩阵

但是，这九维因子存在中每两个因子直接可能存在显著相关性，不符合题目中所述的需要具有独立性，其相关性热力图如图 5 所示。

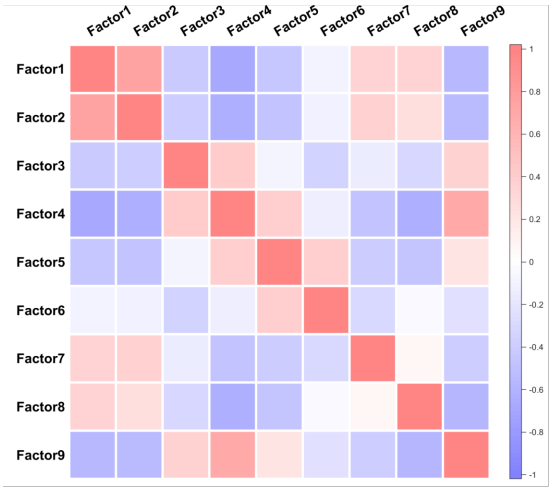

图 5　虚拟因子的相关性热力图

由图 5 可知，虚拟因子 1（Factor1）和虚拟因子 2 之间是存在较强相关性的。同时，也与虚拟因子 7 和虚拟因子 8 之间存在一定的相关性，因此因子分析法降维得到的九维虚拟因子正交性差，独立性差。

3. 典型分析筛选的指标拥有更好的正交性

为找到真正符合题目要求的，具有代表性、独立性、合理性的指标，我们对典型分析得到的九个指标绘制相关性热力图 6 所示。

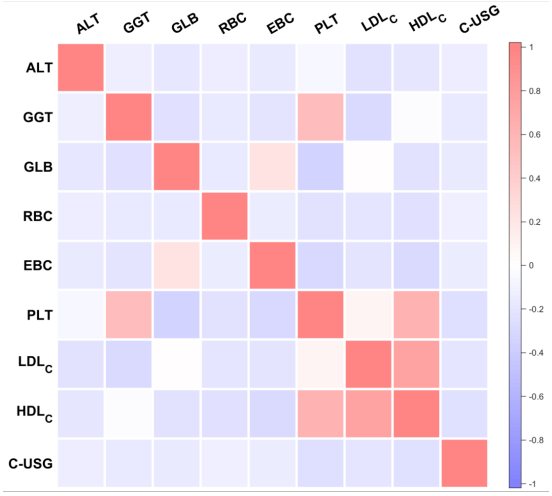

图 6　典型分析指标的相关性热力图

从图中可以很明确的看出，典型分析提取的九个主要指标之间颜色深的区块明显减少了，代表各个指标之间的正交性更优，更具有独立性。同时由于这九个指标都是拥有实际意义的医学指标，包括 ALT(丙氨酸氨基转换酶)、GGT(谷氨酰基转换酶)、GLB(球蛋白)、RBC（红细胞计数)、EBC(白细胞计数)、PLT(血小板计数)、LDL_c(低密度脂蛋白胆固醇)、HDL_c(高密度脂蛋白胆固醇)、USG(尿素) 等。因此，通过典型分析所降维的九个指标更具有代表性和可解释性，更为合理和独立。

六、 问题三模型的建立与求解

6.1 问题分析

用 shuffle 法将附件一中的数据划分为三部分，70% 的训练集、交叉验证集和 30% 的验证集。选择用 AdaBoost 模型建立血糖值预测模型，但在此之前需要对其超参数进行深入优化，得到最优超参数，利用预处理的特征数据带入其中，得到血糖值预测数据。最后通过 MSE、MAE、RMSE、MAPE 四项评价指标评估模型的性能。

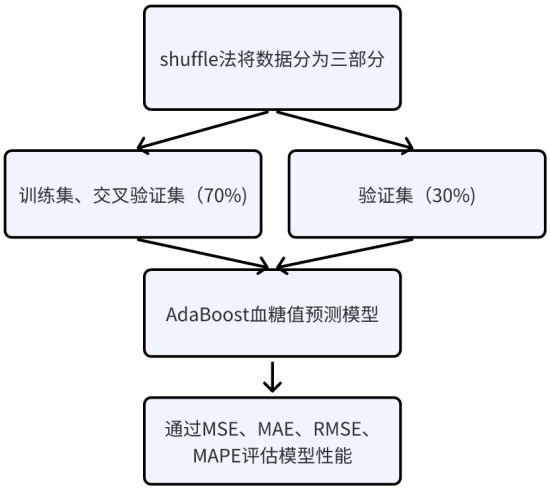

图 7　问题三流程图

6.2 Adaboost 自适应提升模型

6.2.1 模型介绍

AdaBoost（自适应提升）是一种机器学习中的集成学习算法[3]，用于提升弱分类器的性能。它是一种迭代算法，通过逐步改进分类器的准确性来构建一个强大的分类器。AdaBoost 被广泛应用于分类和回归问题，并且在实际应用中表现出色。

其具体步骤如下：

Algorithm 1 AdaBoost

procedure AdaBoost(X, T)　　▷ 输入：数据集 X，迭代次数 T

　初始化样本权重 $w_i = \frac{1}{N}$，$i = 1, 2, \ldots, N$

　初始化弱分类器列表 Classifiers

　for $t = 1$ to T **do**

　　训练一个弱分类器 h_t，使用权重 w_i 调整样本分布

　　计算弱分类器的分类误差 ε_t

　　计算弱分类器的权重 $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

　　更新样本权重 $w_i \leftarrow w_i \cdot \exp\left(-\alpha_t y_i h_t(x_i)\right)$，$i = 1, 2, \ldots, N$

　　归一化样本权重 $w_i \leftarrow \frac{w_i}{\sum_{j=1}^{N} w_j}$

　　将弱分类器 h_t 加入 Classifiers

　end for

　return Classifiers

end procedure

6.2.2 模型评估指标

通过模型评估指标选出最优回归预测模型，这里选用了 MSE，RMSE，MAE 和 MAPE 作为评估指标。

评价指标是针对相同数据，输入不同算法，或输入相同算法不同参数给出这个算法优劣的指标，我们将问题看成回归问题，以此来预测样本的血糖值。

上述 MSE，RMSE，MAE 和 MAPE 需要基于预测值和真实值的差异计算得出。

其中，均方误差（MSE，Mean Squared Error）是指对于每个样本，计算预测值与真实值之间的差异，将差异的平方累加，然后取平均值。这样可以得到平均平方误差。

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{17}$$

均方根误差（RMSE，Root Mean Squared Error）是指 MSW 的平方根，它将平均平方误差转化为与原始数据量纲相同的值。

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{18}$$

平均绝对误差（MAE，Mean Absolute Error）是指对于每个样本，计算预测值与真实值之间的绝对差异，将绝对差异累加，然后取平均值。这样可以得到平均绝对误差。

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{19}$$

平均绝对百分比误差（MAPE，Mean Absolute Percentage Error）是指对于每个样本，计算预测值与真实值之间的百分比差异，取绝对值后累加，然后取平均值。最终得到平均绝对百分比误差。

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{20}$$

6.3 基于 Adaboost 的血糖值预测模型建立

Step1. 特征数据预处理

基于问题二中的降维方法，得到了九个具有代表性、独立性以及医学解释合理性的变量指标。考虑到在问题二中已对这九个变量指标进行了正向化和标准化处理，因此可将经过处理的指标数据直接输入模型构建。

Step2. 血糖值预测模型的确立

鉴于通过弱分类器的集成，Adaboost 模型可显著提升预测性能，降低过拟合风险，同时又具备简单性与适应性，因此被选定作为血糖值预测模型。然而，在模型应用之前，有必要对其超参数进行深入优化，以进一步提升预测效果。

Step3.AdaBoost 超参数最优组合的确定

1. 网格化搜索最优超参数

Adaboost（自适应增强）模型的超参数优化过程，可通过网格化搜索方法来确定最佳组合：

将弱分类器数量设定于范围 100-200 之间，学习率设定于 0.1 到 1 之间进行探索。利用 Matlab 中的 $'OptimizeHyperparameters'$ 参数函数，成功确定了最优的弱分类器数量为 103，学习率为 0.12。

2. 最终模型的参数

由此，可确定最优化模型参数，如表 13 所示。

表 13　最优化模型参数表

参数名	参数值	参数名	参数值
训练用时	14.058s	L1 正则项	0
数据切分	0.7	L2 正则项	1
数据洗牌	是	样本征采样率	1
交叉验证	10	树特征采样率	1
基学习器	gbtree	节点特征采样率	1
基学习器数量	103	叶子节点中样本的最小权重	0
学习率	0.12	树的最大深度	10

步骤 4. 模型评估指标

因此，通过基于网格化搜索的最优 Adaboost 模型，可得到如表 14 所示的评估指标。

表 14　模型评估指标

	MSE	RMSE	MAE	MAPE
训练集	0	0.002	0.008	1.693
交叉验证集	0.002	0.042	0.024	21.258
测试集	0.002	0.045	0.024	28.629

其中，与优化前相比，主要是交叉验证集和测试集的 RMSE 从 0.049 优化到 0.045，优化了约 8.16%;，优化 MAPE 值从 33.907 到 28.629，优化了约 15.6%

6. 模型结果分析

由此可得附件 1 的血糖值预测数据，成功构建了基于 Adaboost 的血糖值预测模型，以首位六个数据为例，如表 15 所示，其余数据见附件 2。

表 15　测试集血糖值预测

预测测试集数据	GLU(血糖值)	因子 1	...	因子 9
5.214382271	6.33	209.5723939	...	73.63695689
5.022157178	6.43	197.673606	...	62.749536
5.671126366	4.41	247.009302	...	102.583342
⋮	⋮	⋮	⋮	⋮
5.371194147	4.92	193.898547	...	51.523017
5.768797722	5.8	228.3676489	...	82.11320189
5.107286555	4.58	191.5978324	...	62.56520887

通过 excel 可视化，可得前 38 个预测值和实际值的折线图，如图 8 所示。

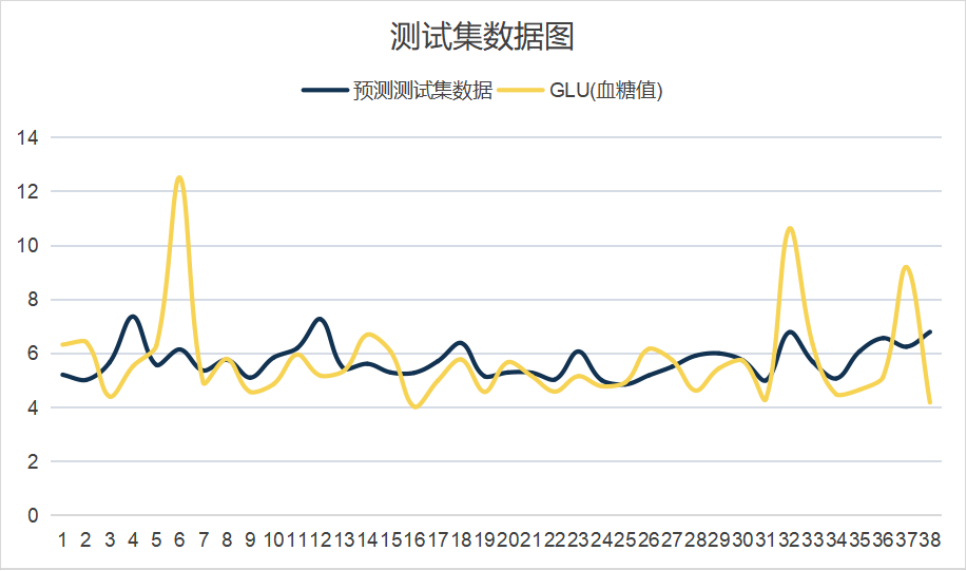

图 8　测试集数据图

七、 问题四模型的建立与求解

7.1 问题分析

首先，需要对问题三的血糖预测模型进行改进，以适用于附件二的预测，并对糖尿病风险进行解释。鉴于确诊糖尿病需要两次测量，所以，应对所有样本进行双重预测，并使用平均值构建基于 Adaboost 的二次血糖预测模型。同时，借助 Adaboost 模型，根据附件一的数据，构建糖尿病风险预测模型，将血糖值分为正常、偏高和异常三个区间，并使用 0、1、2 进行对应标记。在此基础上，将模型预测出的异常血糖区间概率视为糖尿病患病风险的量化标准。最后，将该方法应用于附件二的数据并解释糖尿病风险。

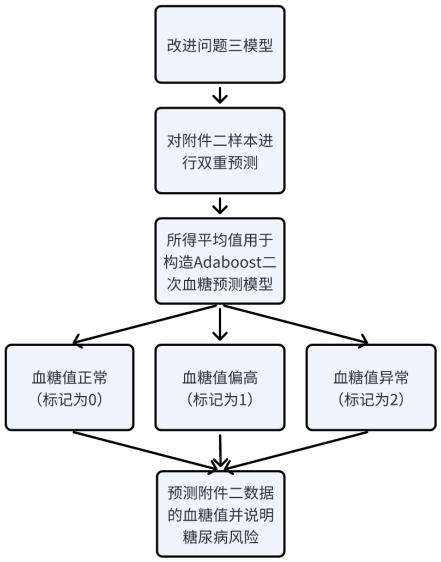

图 9　问题四流程图

7.2 基于 Adaboost 的二次血糖值预测模型的构建

Step1. 附件二数据预处理

提取附件二中所需的降维指标，进行正向化，标准化处理。

Step2. 模型超参数的确定

使用与问题三一样的超参数（见表 12）

Step3. 模型的训练

通过对附件一正向化标准化处理后的 5904 个有效指标对问题三所建立的模型进行训练，得到了训练好的基于 Adaboost 的二次血糖值预测模型。

Step4. 对附件 2 的检测数据的血糖值预测

根据题目所示，正常指标含量在 [3.9,6.1)，高血糖范围为 [6.1,6.7)。鉴于糖尿病的诊断需进行双次测量，因此针对同一样本进行了两次预测。取两次样本血糖预测值的平

均值作为最终血糖预测值，首尾共 6 个样本的预测结果如表 15 所示，其余预测样本见附件 3。

表 16　附件 1 血糖预测数据

样本序号	预测数据 1	预测数据 2	最终预测结果
6001	6.610011312	6.414875566	6.512443439
6002	6.00311086	5.840780543	5.921945701
6003	6.672794118	6.474264706	6.573529412
⋮	⋮	⋮	⋮
6138	7.562217195	7.31561086	7.438914027
6140	6.034502262	5.870475113	5.952488688
6141	6.00311086	5.840780543	5.921945701

步骤 5. 模型结果分析

由此，得到了所有样本的预测结果，图 10 呈现了前 50 个样本的最终预测曲线。

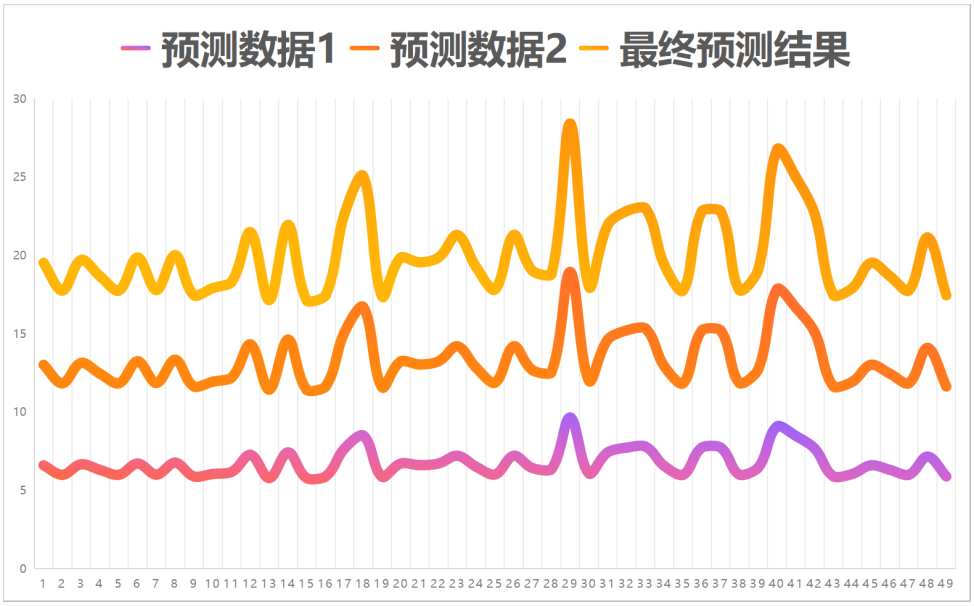

图 10　前 50 个样本血糖值预测曲线

总的来说，附件 2 的 141 个有效样本的分布情况如下：正常样本 62 个，高血糖样本 39 个，异常样本 40 个。

7.3 糖尿病风险关联模型的构建

步骤 1. 数据预处理

根据题目，将附件 1 中的血糖值分为 3 类，正常，高血糖，和异常血糖，对应 0，1，2 三类标签。由此可得 4985 个正常样本，585 个高血糖样本和 393 个异常样本。鉴于血糖是预测目标，故不将血糖值作为输入指标，只将降维后的 9 个指标作为特征指标输入。

步骤 2. 模型的训练

模型的具体参数如表 16 所示。

表 17　模型具体参数

参数名	参数值	参数名	参数值
训练用时	6.069s	交叉验证	10
数据切分	0.7	基分类器数量	108
数据洗牌	是	学习率	0.21

步骤 3. 模型的评估

通过模型评估指标选出最优分类预测模型，这里选用了准确率，召回率，精确率和 F1 分数作为评估指标。上述准确率，召回率，精确率和 F1 分数需要基于混淆矩阵计算得出。混淆矩阵是衡量分类模型准确度中最基本和最直观的方法，但是混淆矩阵无论面对大量数据还是少量数据，它统计的只是个数，难以衡量模型的优劣。所以，我们在混淆矩阵统计的结果上进行计算准确率，召回率，精确率和 F1 分数。以此来得到稳定可靠的预测模型。

准确率表示被分类器正确分类的元组占总样本数的比例

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

召回率表示完全性的度量，也就是正例被标记为正的百分比

$$\text{召回率} = \frac{TP}{TP + FN} \tag{22}$$

精确率表示标记为正类的元组实际值为正类所占的比例

$$\text{精确率} = \frac{TP}{TP + FP} \tag{23}$$

F1 分数由精确率和召回率共同决定，可以看作二者的加权平均，取值范围为 [0,1]。

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{24}$$

模型性能的评估可依据准确率、召回率、精确率和 F1 指标完成如表 18 所示是血糖值预测模型的指标评估结果。

表 18　模型的评估指标

	准确率	召回率	精确率	F1
训练集	0.817	0.817	0.734	0.754
交叉验证集	0.816	0.816	0.724	0.752
测试集	0.835	0.835	0.749	0.785

其中，准确率超过 80%，说明模型具有较为广阔的实用价值。

步骤 4. 模型的预测

经过上述步骤，模型已训练完毕。对附件 2 数据进行预测后，获得了样本位于三个区间的概率，其中以 2 概率量化糖尿病风险。鉴于篇幅约束，仅呈现首尾 6 个样本的结果，如表 19 所示。

表 19　患病风险

编号	预测结果	患糖尿病风险	GLU(血糖值)
6001	1	0.565766677	6.512443439
6002	0	0.386830537	5.921945701
6003	1	0.566364839	6.573529412
⋮	⋮	⋮	⋮
6139	2	0.806750702	7.438914027
6140	0	0.416830537	5.952488688
6141	0	0.416530534	5.921945701

其中，预测为 0 的正常样本数为 62 个，正常样本数 39 个，异常样本数 40 个。

步骤 5. 模型结果分析

血糖值预测与风险关联模型为附件 2 中的数据提供了血糖预测值与患病风险概率。此外，该模型还揭示了各指标的特征重要性，如图 11 所示。

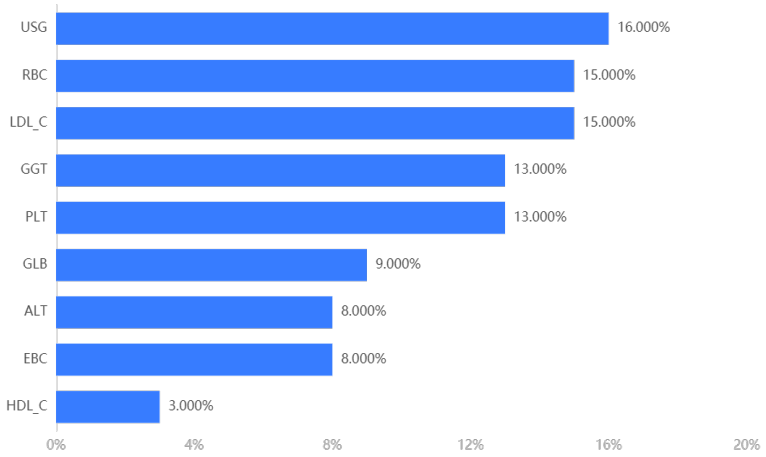

图 11　各指标的特征重要性

由图 11 可得，USG（尿素）是响应糖尿病的直接指标，肾功能指标是衡量糖尿病患病风险的重要指标。血糖值越高，其确诊为糖尿病，病情加重的风险越大。例如 6139 号样本的血糖值远超过 6.7，其被确诊为糖尿病的概率达到 80% 以上。而对于 6001 号样本，其血糖值接近 6.7，位于高血糖范畴，但仍有 56% 的潜在可能性进一步发展为糖尿病，风险较高。

八、 模型的稳定性分析

问题三中探究了各项指标对血糖值预测的数学模型。而问题三中所建立的模型是否稳定，不仅影响了不仅影响了问题三求解的精确性，也是问题四的求解的基础。

为深入分析各项指标对血糖值数学模型的稳定性，需要特别关注在扰动某一项指标情况下模型的稳定性。以扰动丙氨酸氨基转移酶（ALT）为例，模型稳定性如下：

(1) 在 ALT 扰动幅度为 1% 时，模型训练所需时间为 24.712 秒，扰动百分比为 3.74%。

(2) 在 ALT 扰动幅度为 5% 时，模型训练所需时间为 24.455 秒，扰动百分比为 8.06%。

(3) 在 ALT 扰动幅度为 10% 时，模型训练所需时间为 23.982 秒，扰动百分比为 9.21%。

观察到在扰动情况下，所得结果的误差有一定程度的增加，然而仍在可接受范围内。综合而言，在较低精度要求下，模型呈现较好的稳定性。

九、 模型的评价

9.1 模型的优点

1. 量化指标的选取简单易用且能较好地反映原数据集特征，有极高的实用价值。

2. 没有使用样本个体的年龄、性别等识别特征而具有较高的准确度，可以充分保护受试对象隐私。

3. 使用两次预测血糖，进一步提高了血糖含量预测的准确性，减少了误差率。

9.2 模型的缺点

1. 降维导致一些数据的信息缺失，使得模型预测的准确度有部分降低。

2. 对高血糖患者的潜在的糖尿病患病类型总结不足。尽管可以较为准确高效的预测血糖值含量及其患病风险，但是对于具体的糖尿病病症，例如 1 型和 2 型糖尿病不能很好的区分。

十、 模型的推广

该模型采取的数据降维思路可以应用于数据指标多、数据内部结构明显的分类和回归预测问题。如微生物群落生境的分类问题、河流湖泊演化反演、生化指标回归预测问题等。本文中使用的 Adaboost 方法也可以广泛应用于各类体检指标的判别分类，对世界范围内的医学、地学和生物学都有广阔而深远的应用前景。

十一、参考文献

[1] 佟志莹. 基于网格搜索算法的糖尿病风险预测模型应用研究. 硕士论文, 北京交通大学, 2022.

[2] 许涤龙, 陈春晖. 中国股市有效性分析中的正态性检验. 统计与信息论坛, 19(6):34–38, 1 2004.

[3] 李银燕. 基于 adaboost 方法的人脸检测. 硕士论文, 复旦大学, 2006.

附录清单

本文代码主要以 MATLAB，SPSSPRO 运行

软件运行环境: 操作系统：$windows11$

配置：$Intel\quad Cor\quad i9\quad 13900HX(2.2GHz/L3\quad 36M)$

显卡：$NVDIA\quad GeForce\quad RTX\quad 4090\quad Laptop\quad GPU\quad 24GB$

附录 一　数据预处理

1.1　数据导入

```
   clc;clear;close all;
opts = delimitedTextImportOptions("NumVariables", 42);
opts.DataLines = [2, Inf];
opts.Delimiter = ",";
opts.VariableNames = ["GGT", "ALT", "ACT", "TP", "GLB", "ALP", "ID", "GRA", "K_Be", "HBeAg",
    "K_HBc", "K_HBs", "HBsAg", "LDL_C", "Date", "MO", "BAS", "EOS", "USG", "UA", "Age", "Sex",
    "TCHO", "LYM", "TG", "AG", "EBC", "ALB", "RDW", "HCT", "MCV", "MCHC", "MCH", "RBC", "CR",
    "PDW", "MPV", "PCT", "PLT", "GLU", "HGB", "HDL_C"];
opts.VariableTypes = ["double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "datetime",
    "double", "double", "double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "double", "double"];
opts.ExtraColumnsRule = "ignore";
opts.EmptyLineRule = "read";
opts = setvaropts(opts, "Sex", "EmptyFieldRule", "auto");
opts = setvaropts(opts, "Date", "InputFormat", "dd/MM/yyyy");
rawCatedData = readtable("附件1—有血糖值的检测数据.csv", opts);
clear opts
save rawCatedData.mat
```

```
   clc;clear;close all;
opts = delimitedTextImportOptions("NumVariables", 41);
opts.DataLines = [2, Inf];
opts.Delimiter = ",";
opts.VariableNames = ["ID", "Sex", "Age", "Date", "ACT", "ALT", "ALP", "GGT", "TP", "ALB",
    "GLB", "AG", "TG", "TC", "HDL_C", "LDL_C", "USG", "CR", "UA", "HBsAg", "K_HBs", "HBeAg",
    "K_Be", "K_HBc", "EBC", "RBC", "HCT", "HGB", "MCV", "MCH", "MCHC", "RDW", "PLT", "MPV",
    "PDW", "PCT", "GRA", "LYM", "MO", "EOS", "BAS"];
opts.VariableTypes = ["double", "double", "double", "datetime", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double", "double", "double",
    "double", "double", "double", "double", "double", "double", "double"];
opts.ExtraColumnsRule = "ignore";
```

```
opts.EmptyLineRule = "read";
opts = setvaropts(opts, "Date", "InputFormat", "MM/dd/yyyy");
rawUncatedData = readtable("附件2—无血糖值的检测数据.csv", opts);
clear opts
save rawUncatedData.mat
```

1.2 数据清洗与数据分类

```
    clc;clear;close all;
load rawCatedData.mat
rawCatedData_copy=rawCatedData;
rawCatedData_copy.ID=[];%脱敏数据去掉ID
%% 将日期信息转换为离散数值
date=rawCatedData_copy.Date;%datetime单独考虑
rawCatedData_copy.Date=[];
date=char(date);
month=str2num(date(:,4:5));
%因为所有样本均在9月与10月之间，所以先不考虑了
%% 检验年龄是否为正态分布
age=rawCatedData_copy.Age;
cdf=[age,normcdf(age,9.8225,5.0572)];
isAgeCdf=kstest(age,cdf);
%% 将乙肝患者的样本分离开
K_Be=rawCatedData_copy.K_Be;
HBVind=find(~isnan(K_Be));
HBVData=rawCatedData_copy(HBVind,:);
rawCatedData_copy=rawCatedData_copy(find(isnan(K_Be)),:);
save rawHBVCatedData.mat HBVData HBVind
%% 筛选缺失值
[len,rawVar]=size(rawCatedData_copy);
indValued=[];
countNan=zeros(1,rawVar);
for i=1:rawVar%统计每个指标的缺失值数量
    tmp=table2array(rawCatedData_copy(:,i));
    countNan(i)=sum(isnan(tmp));
    indNan=find(isnan(tmp));
    rawCatedData_copy(indNan,i)=array2table(mean(rmmissing(tmp)));
    if (countNan(i)<(len/2))
        indValued=[indValued,i];
    end
end
clear i indNan
catedData=rawCatedData_copy(:,indValued);%先过滤掉缺失值过大的指标
countNan=countNan(indValued);
clear indValued tmp
```

```
[countNan,ind]=sort(countNan,'ascend');%对剩下的指标依据缺失值排序
catedData=catedData(:,ind);%处理完成
clear ind rawVar
%% 数据正向化
catedData.AG=catedData.ALB./catedData.GLB;
catedData.ALB=[];catedData.GLB=[];
catedData.AG=abs(catedData.AG-2);
catedData.RBC=max(catedData.RBC)-catedData.RBC;
catedData.EBC=max(catedData.EBC)-catedData.EBC;
catedData.PLT=max(catedData.PLT)-catedData.PLT;
catedData.HDL_C=max(catedData.HDL_C)-catedData.HDL_C;
%% 计算与血糖相关性
glucose=rawCatedData_copy.GLU;%血糖
indSick=[find(glucose>6.1);find(glucose<3.1)];
cateSick=zeros(len,1);
cateSick(indSick)=1;%标记异常值
t=table(cateSick);
catedData=[t,catedData];%将非数值数据放回，这两个没有缺失值不参加排序放前面
clear t
varlen=size(catedData,2);
seq=table2array(catedData(:,5:varlen));%属于连续值的指标
seqnum=size(seq,2);
R=zeros(1,seqnum);P=zeros(1,seqnum);
for i=1:seqnum                  %计算每个连续值的指标与血糖间的相关性
    [tmpR,tmpP]=corrcoef(seq(:,i),glucose);
    R(i)=tmpR(1,2);P(i)=tmpP(1,2);
end
clear i tmpR tmpP
%% 切分显著指标与不显著指标
significantInd=find(P<=0.01);
significantSeq=seq(:,significantInd);
unSignificantInd=find(P>0.01);
unsignificantSeq=seq(:,unSignificantInd);
catedData=catedData(:,[1:4,significantInd+4]);%留在表中的都是显著的连续指标了

%% 输出标准化前结果
clear rawCatedData_copy
save catedData.mat catedData countNan isAgeCdf
save catedCorrcoef.mat seq R P significantSeq unsignificantSeq glucose
```

```
    clc;clear;close all;
load rawUncatedData.mat
rawUncatedData_copy=rawUncatedData;
rawUncatedData_copy.ID=[];
rawUncatedData_copy.Date=[];
%% 将乙肝患者的样本分离开
```

```
K_Be=rawUncatedData_copy.K_Be;
HBVind=find(~isnan(K_Be));
HBVData=rawUncatedData_copy(HBVind,:);
save HBVUncatedData.mat HBVData HBVind
rawUncatedData_copy=rawUncatedData_copy(find(isnan(K_Be)),:);
%%
[len,rawVar]=size(rawUncatedData_copy);
indValued=[];
countNan=zeros(1,rawVar);
for i=1:rawVar
    tmp=table2array(rawUncatedData_copy(:,i));
    countNan(i)=sum(isnan(tmp));
    indNan=find(isnan(tmp));
    rawUncatedData_copy(indNan,i)=array2table(mean(rmmissing(tmp)));
    if (countNan(i)<(len/2))
        indValued=[indValued,i];
    end
end
clear i indNan
uncatedData=rawUncatedData_copy(:,indValued);
countNan=countNan(indValued);
clear indValued tmp
[countNan,ind]=sort(countNan,'ascend');
uncatedData=uncatedData(:,ind);
clear ind rawVar
%%
save uncatedData.mat uncatedData countNan
```

```
    clc;clear;close all;
load rawHBVCatedData.mat
%% 填充缺失值
varnum=size(HBVData,2);
for i=1:varnum
    tmp=table2array(HBVData(:,i));
    indNan=find(isnan(tmp));
    HBVData(indNan,i)=array2table(mean(rmmissing(tmp)));
end
clear i tmp varnum indNan
%% 数据正向化
HBVData.AG=HBVData.ALB./HBVData.GLB;
HBVData.ALB=[];HBVData.GLB=[];
HBVData.AG=abs(HBVData.AG-2);
HBVData.RBC=max(HBVData.RBC)-HBVData.RBC;
HBVData.EBC=max(HBVData.EBC)-HBVData.EBC;
HBVData.PLT=max(HBVData.PLT)-HBVData.PLT;
HBVData.HDL_C=max(HBVData.HDL_C)-HBVData.HDL_C;
```

```
%% 计算与血糖相关性
GLU=HBVData.GLU;Sex=HBVData.Sex;Age=HBVData.Age;%血糖
HBVData.GLU=[];HBVData.Sex=[];HBVData.Age=[];
indSick=[find(GLU>6.1);find(GLU<3.1)];
cateSick=zeros(length(HBVind),1);
cateSick(indSick)=1;%标记异常值

varlen=size(HBVData,2);
seq=table2array(HBVData);%属于连续值的指标
seqnum=size(seq,2);
R=zeros(1,seqnum);P=zeros(1,seqnum);
for i=1:seqnum                  %计算每个连续值的指标与血糖间的相关性
    [tmpR,tmpP]=corrcoef(seq(:,i),GLU);
    R(i)=tmpR(1,2);P(i)=tmpP(1,2);
end
clear i tmpR tmpP varlen
%% 切分显著指标与不显著指标
significantInd=find(P<=0.01);
significantSeq=seq(:,significantInd);
unsignificantInd=find(P>0.01);
unsignificantSeq=seq(:,unsignificantInd);
HBVData=HBVData(:,unsignificantInd);
t=table(Sex,Age,cateSick);
HBVData=[t,HBVData];%将非数值数据放回
writetable(HBVData,'HBVData.xlsx');
clear t
%%

save HBVCatedData.mat
```

1.3 数据标准化

```
    clc;clear;close all;
load catedData.mat
load catedCorrcoef.mat
glucose=normalize(glucose,'range');
num=size(catedData,2);
seq=normalize(seq,'range');
SignificantSeq=normalize(significantSeq,'range');
UnsignificantSeq=normalize(unsignificantSeq,'range');
for i=5:num
    catedData(:,i)=normalize(catedData(:,i),'range');
end
catedData.GLU=normalize(catedData.GLU,'range');
clear i num
```

```
save normalizeCatedData.mat
```

```
data = xlsread('第四问无监督优化预测数据.xlsx');
data01=data;
min_values = min(data01);
max_values = max(data01);
normalized_data = (data01 - min_values) ./ (max_values - min_values)
```

附录 二　分类预测方法

2.1　相关性 p 值计算

```
    function [coef,weight]=m_Factor(X,num)
    %因子分析Factor Analysis,X为自变量矩阵,Y为因变量矩阵,num为主因子个数
    Correspond=cov(X);
    [rate,latent,con]=pcacov(Correspond);
        %latent为Correspond的特征值，rate为各个主成分的贡献率
    f1=repmat(sign(sum(rate)),size(rate,1),1);       %构造与vec1同维数的元素为±1的矩阵用于求正分量
    rate=rate.*f1;
        %修改特征向量的正负号，使得每个特征向量的分量和为正，即为最终的特征向量
    f2=repmat(sqrt(latent)',size(rate,1),1);
    factormatOrigin=rate.*f2;                              %构造全部因子的载荷矩阵
    factormatMain=factormatOrigin(:,1:num);
    [factormat_rotate,t]=rotatefactors(factormatMain,'Method','varimax');
        %b为旋转后的矩阵,t为做变换的正交矩阵
    factmat=[factormat_rotate,factormatOrigin(:,num+1:end)];
    contribution=sum(factmat.^2);                          %计算各个因子的贡献
    rate=contribution(1:num)/sum(contribution);   %计算因子的贡献率
    coef=Correspond\factormat_rotate;           %计算得分函数的系数
    weight=rate/sum(rate);                                 %计算得分的权重
end
```

```
%GatherData
    clc;clear;close all;
load normalizeCatedData.mat
liverFunction=[catedData.ALT,catedData.ACT,catedData.ALP,catedData.GGT,catedData.AG,catedData.TP];
redBlood=[catedData.RBC,catedData.RDW,catedData.HGB,catedData.HCT,catedData.MCHC,catedData.MCH];
whiteBlood=[catedData.EBC,catedData.GRA,catedData.LYM];
plateletBlood=[catedData.PDW,catedData.PCT,catedData.PLT];
lipidBlood=[catedData.LDL_C,catedData.HDL_C,catedData.TCHO,catedData.TG];
kidneyFunction=[catedData.USG,catedData.CR];
[coef.Liver,p.Liver]=corrcoef(liverFunction);
[coef.Red,p.Red]=corrcoef(redBlood);
[coef.White,p.White]=corrcoef(whiteBlood);
```

```
[coef.Lipid,p.Lipid]=corrcoef(lipidBlood);
[coef.Kidney,p.Kidney]=corrcoef(kidneyFunction);
[coef.Platelet,p.Platelet]=corrcoef(plateletBlood);

%选取ALT,GGT,GLB,RBC,EBC,PLT,LDL_C,HDL_C,CR
regressData=[catedData.ALT,catedData.GGT,catedData.AG,catedData.RBC,catedData.EBC,catedData.PLT,catedData.LDL_C
[Coef,P]=corrcoef(regressData);
save orthCheck.mat Coef P
clear Coef P
regressData=[glucose,regressData];
save regressData.mat regressData coef
```

2.2 Adaboost 示例代码

```
import numpy
import pandas
from spsspro.algorithm import supervised_learning
data_x = pandas.DataFrame({
    "A": numpy.random.random(size=100),
    "B": numpy.random.random(size=100)
})
data_y = pandas.Series(data=numpy.random.random(size=100), name="C")
#adaboost回归，输入参数详细可以光标放置函数括号内按shift+tab查看，输出结果参考spsspro模板分析报告
result = supervised_learning.adaboost_regression(data_x=data_x, data_y=data_y)
print(result)
```

2.3 依据预测结果分类

```
%CateGLU
clc;clear;close all;
checkData = readtable("PredictResult.csv");
GLU=checkData.GLU;
len=length(GLU);
indNormal=sort(intersect(find(GLU>=3.9),find(GLU<6.1)),'ascend');
glucoseNormal=GLU(indNormal);
indRisky=sort(intersect(find(GLU>=6.1),find(GLU<6.7)),'ascend');
glucoseRisky=GLU(indRisky);
indOther=setdiff([1:len]',[indRisky;indNormal]);
glucoseOther=GLU(indOther);
clear len GLU
save checkData_cate.mat
```

附录 三　结论验证与呈现

3.1　灵敏度分析扰动代码

```
clc;close all;clear
data=xlsread('训练数据.xlsx');
needdata = data(:,3);
%%
% 设置扰动的百分比
perturbations = [1, 5, 10] / 100;

% 初始化存储扰动后数据的变量
data1 = cell(size(perturbations));

% 对每个扰动百分比进行循环
for i = 1:length(perturbations)
    perturbation = perturbations(i);

    % 对数据进行扰动
    perturbed_data = needdata * (1 + perturbation);

    % 存储扰动后的数据
    data1{i} = perturbed_data;
end

% 保存数据到 MATLAB 工作区
save('perturbed_data.mat', 'data1');
```

3.2　热力图矩阵绘制

```
    clc;clear;close all
load regressData.mat
map = [0.4698  0.6071   0.9678
    0.4876   0.6281   0.9765
    0.5083   0.6474   0.9825
    0.5290   0.6663   0.9882
    0.5529   0.6874   0.9922
    0.5712   0.7045   0.9961
    0.5923   0.7217   0.9961
    0.6133   0.7388   0.9961
    0.6363   0.7549   0.9961
    0.6554   0.7691   0.9922
    0.6725   0.7823   0.9882
    0.6964   0.7955   0.9843
    0.7146   0.8087   0.9765
```

```
    0.7357   0.8196   0.9663
    0.7527   0.8275   0.9608
    0.7687   0.8366   0.9477
    0.7871   0.8431   0.9345
    0.8081   0.8510   0.9213
    0.8214   0.8549   0.9081
    0.8385   0.8588   0.8948
    0.8556   0.8627   0.8738
    0.8710   0.8606   0.8545
    0.8860   0.8513   0.8320
    0.8992   0.8459   0.8145
    0.9124   0.8327   0.7909
    0.9255   0.8196   0.7646
    0.9349   0.8102   0.7420
    0.9442   0.7970   0.7225
    0.9496   0.7832   0.7003
    0.9569   0.7646   0.6725
    0.9608   0.7500   0.6485
    0.9647   0.7324   0.6265
];
map = flipud(map);
coefName={'Liver','Red','White','Platelet','Lipid','Kidney'};
titleName={'liver function','red blood','white blood','blood platelet','blood lipid','kidney
    function'};
stringName={{'ALT','ACT','ALP','GGT','AG','TP'},
           {'RBC','RDW','HGB','HCT','MCHC','MCH'},
           {'PDW','PCT','PLT'},
           {'EBC','GRA','LYM'},
           {'LDL_C','HDL_C','TCHO','TG'},
           {'USG','CR'}};
for i=1:6
    subplot(2,3,i);

    xvalues = stringName{i};
    yvalues = stringName{i};
    h = heatmap(xvalues,yvalues, getfield(p,coefName{i}));
    h.Title =strcat(titleName{i}, ' Index p-value');
    colormap(map);
    set(gcf,'Color',[1 1 1]);
    h.CellLabelColor = 'none';
end
clear i xvalues yvalues
```

附录 四　不显著指标的描述统计表

表 20　不显著指标的描述统计表

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差	峰度	偏度	变异系数（CV）
MO	5905	23.2	3.1	6.859	1.564	6.7	2.446	3.772	1.015	0.228
BAS	5905	3.5	0	0.603	0.291	0.6	0.085	5.998	1.466	0.483
EOS	5905	22.5	0	2.039	1.696	1.6	2.877	14.543	2.781	0.832
MCV	5905	113	59	89.083	4.456	89.3	19.854	4.823	-1.058	0.05
MPV	5905	15.2	7.1	10.656	0.983	10.6	0.967	0.678	0.076	0.092
AG	5905	7.12	0.52	1.501	0.195	1.501	0.038	117.203	4.366	0.13
ALB	5905	54.08	29.54	45.816	2.324	45.816	5.402	1.716	-0.162	0.051
UA	5905	776.59	118.67	355.441	84.211	355.441	7091.489	1.503	0.688	0.237
年龄	5905	93	3	45.691	12.996	45	168.907	-0.27	0.412	0.284

附录 五　显著指标的描述统计表

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差
RGB	5905	88.5	14.4	56.734	7.774	56.734	60.433
LYM	5905	76.3	7.5	33.766	7.234	33.6	52.337
EBC	5905	21.06	2.8	6.591	1.608	6.39	2.586
RDW	5905	23.8	10.9	12.739	1.016	12.6	1.033
HCT	5905	0.599	0.239	0.441	0.043	0.44	0.002
MCHC	5905	462	262	335.359	11.4	336	129.965
MCH	5905	44.7	16	29.89	1.995	30	3.981
RBC	5905	6.85	3.01	4.956	0.502	4.93	0.252
PLT	5905	1271	37	253.317	60.887	249	3707.261
HGB	5905	204	65	147.97	16.526	148	273.094

PDW	5905	25.3	8	13.311	2.17	13	4.709
PCT	5905	1.52	0.042	0.268	0.062	0.26	0.004
LDL_C	5905	8.46	0.56	3.367	0.766	3.367	0.587
TCHO	5905	20.46	1.85	5.234	0.914	5.234	0.835
TG	5905	41.57	0.27	1.843	1.577	1.76	2.487
HDL_C	5905	5.28	0.54	1.391	0.281	1.391	0.079
GGT	5905	736.99	6.36	38.822	35.982	33.44	1294.708
ALT	5905	498.89	0.12	27.703	20.047	26.26	401.886
ACT	5905	434.95	10.04	26.855	12.003	26.48	144.072
TP	5905	100.41	57.32	76.786	3.592	76.786	12.902
GLB	5905	66.18	7.06	30.97	3.182	30.97	10.128
ALP	5905	374.32	22.98	87.562	22.717	87.562	516.047
USG	5905	13.39	1.5	4.989	1.138	4.989	1.294
CR	5905	177.42	39.43	78.397	12.106	78.397	146.565