

**多元统计分析课程论文**

**糖尿病影响因素及血糖值预测探究**

**——基于电子健康记录系统数据**

**摘要**

本文使用来自和鲸网站的电子健康记录公开数据集，对各指标进行因子分析并对血糖值进行预测，从而希望能够分析各指标之间的联系，并探究公共因子对血糖值的影响程度，尝试建立糖尿病预测模型。本文首先介绍了糖尿病的历史背景以及机器学习在糖尿病领域上的应用与发展；其次对数据进行了一系列预处理和统计描述，展现了变量的分布特点；之后基于数据的特点，进行基于主成分和最大方差旋转的因子分析，找到各指标的内在联系；最终对提取出的公共因子和因变量血糖值分别用线性回归、随机森林和XGBoost进行回归分析，找到影响血糖值的关键因素。研究结果表明，血脂、肝功能和肾功能指标是诊断糖尿病最关键的三个因素，可以根据这些指标在糖尿病诊断上进一步提高效率。

**关键词**：糖尿病；因子分析；线性回归；随机森林；XGBoost;

目录

[1. 研究背景 2](#_Toc10942)

[2. 数据介绍 3](#_Toc29065)

[2.1. 数据来源 3](#_Toc7205)

[2.2. 数据清洗 5](#_Toc28039)

[2.2.1. 缺失数据处理 5](#_Toc23342)

[2.2.2. 异常值处理 6](#_Toc11397)

[3. 特征工程 7](#_Toc27277)

[3.1. 生成新变量 7](#_Toc22396)

[3.2. 变量相关性 7](#_Toc9224)

[3.3. 归一化处理 8](#_Toc11015)

[4. 因子分析 8](#_Toc7189)

[4.1. 因子分析简介 8](#_Toc4863)

[4.2. bartlett球形检验和KMO检验 9](#_Toc31603)

[4.3. 选择因子个数 10](#_Toc6188)

[4.3.1. 主成分法 10](#_Toc18099)

[4.3.2. 因子旋转 10](#_Toc22586)

[4.3.3. 碎石图 10](#_Toc2752)

[4.4. 因子解释 11](#_Toc17040)

[5. 回归分析 15](#_Toc24294)

[5.1. 模型选择 15](#_Toc26822)

[5.1.1. 多元线性回归 15](#_Toc28656)

[5.1.2. 随机森林 16](#_Toc26561)

[5.1.3. XGBoost 16](#_Toc9553)

[5.2. 实证分析 17](#_Toc22212)

[6. 总结 19](#_Toc14454)

# 研究背景

在医疗领域, 慢性病的及时诊断及慢性病患者的病情变化预测非常重要。在全世界大多数的国家里，糖尿病都是患者基础量巨大的一种疾病。20岁以上成人中，近10 %患有糖尿病，其中一半未得到诊断。每13人中就有1人出现[糖耐量异常](https://baike.baidu.com/item/%E7%B3%96%E8%80%90%E9%87%8F%E5%BC%82%E5%B8%B8/12692770" \t "https://www.heywhale.com/api/notebooks/60546bc24383a70015e97010/_blank)，每6个新生儿中，就有1个在孕期受到[高血糖症](https://baike.baidu.com/item/%E9%AB%98%E8%A1%80%E7%B3%96" \t "https://www.heywhale.com/api/notebooks/60546bc24383a70015e97010/_blank)的影响，每8秒就有1人死于糖尿病及其并发症。据估算，在中国的糖尿病患者已超过一亿人。一方面有许多患者是在疾病进入中后期才诊断出病情，另一方面有许多患者在经过医院治疗后, 血糖控制效果不佳, 而且随之产生的并发症进展尽速，产生更加严重疾病。这些都会给患者带来身心上的巨大打击，加重病人的负担，同时消耗紧张的医疗资源并对医疗保障体系产生巨大的压力。

随着电子健康记录系统在临床实践中的应用越发广泛，绝大多数医院都已在发展信息系统以更好的存储电子健康记录。电子信息系统和EHR的普及在一定程度上已经提高了医院和医生的工作效率，许多临床研究人员已经开始探索如何利用这一数据进行数据分析以支持医疗决策。在拥有了数据的基础上，把机器学习技术引入到海量医学数据的分析中，通常能取得较好的预测效果。这其中，应用最多的当属慢性病的风险与进程预测。举例来说，预测高风险人群未来发生糖尿病的风险指数是一个普遍的需求。在慢性病患者在疾病过程中往往会出现并发症，糖尿病患者往往具有较高的中风，心脏病，眼睛问题和肾衰竭的风险。慢性肾脏疾病则可导致贫血，心血管事件和死亡。预测此类慢性病的并发症也是一个非常普遍的需求。  
 综上，研究者在积累了大量数据，数据类型主要集中在慢性病患者的信息时，就形成了一个有潜力的应用场景。在此场景之上，可以探索使用或者提出一种机器学习预测模型，并在实际的医学疾病数据集上进行验证。对类似于并发症、疾病风险、用药副作用、存活率等诸多目标进行准确预测，可以使临床医生尽早采取预防性和个性化的干预措施，从而降低患者的风险并提高其生活质量。在这一应用场景下，合适的机器学习模型还将帮助医院节省诊断和治疗的巨大成本。

更进一步，在获得了高准确率的预测表现后，一些新的研究方向则更加关注模型的可解释性。尤其是在医疗领域的应用中，更多的关注点是改进现有模型，使得模型的结果变得更具有解释性。这一方向正在过医疗专家与机器学习专家的共同努力下变得更有前景。

# 数据介绍

## 数据来源

本文数据来源于Heywhale公开数据集，共包含41个字段，5641条记录。数据的最后一个字段为标签列，即需要预测的目标血糖值。特征变量主要包括以下几个方面：

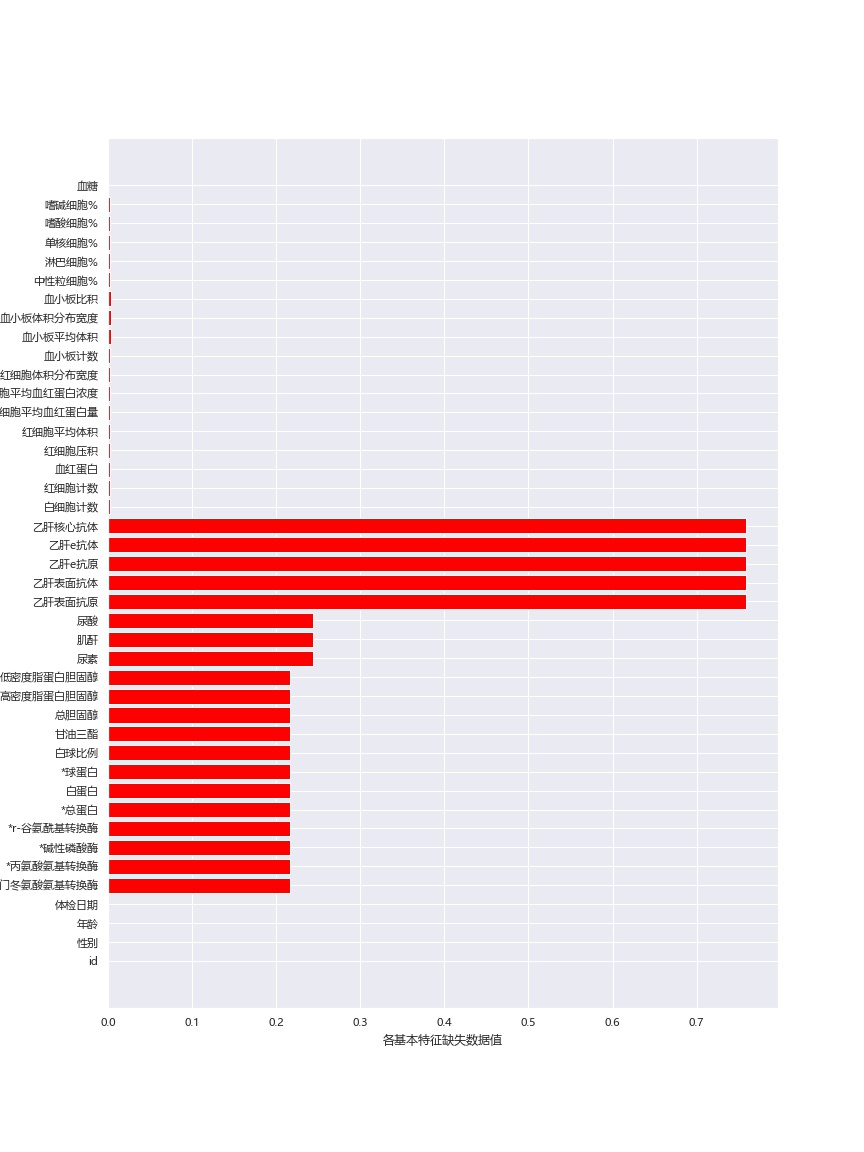
1. 个人信息，包括性别，年龄，体检日期等。
2. 各类酶的活性，包括天门东氨酸氨基转换酶、碱性磷酸酶等。
3. 蛋白含量，包括总蛋白、白蛋白、球蛋白等。
4. 脂质和其他有机物含量，包括甘油三酯、胆固醇、尿素、肌酐等。
5. 乙肝抗体，包括乙肝表面抗原、乙肝表面抗体等。
6. 红细胞和白细胞信息，包括计数、红细胞压积、红细胞平均体积等。
7. 血小板信息，包括血小板计数，血小板平均体积等。
8. 其他细胞含量，包括中性粒细胞、淋巴细胞、单核细胞等。

|  |  |  |
| --- | --- | --- |
| 变量名称 | 变量类型 | 取值范围 |
| 性别 | 分类 | 0,1 |
| 年龄 | 数值 | 3-93 |
| 体检日期 | 字符串 | -- |
| 天门冬氨酸氨基转换酶 | 数值 | 10.04-434.95 |
| 丙氨酸氨基转换酶 | 数值 | 0.12-498.89 |
| 碱性磷酸酶 | 数值 | 22.98-70.35 |
| r-谷氨酰基转换酶 | 数值 | 6.36-736.99 |
| 总蛋白 | 数值 | 57.32-100.41 |
| 白蛋白 | 数值 | 29.54-54.08 |
| 球蛋白 | 数值 | 7.06-66.18 |
| 白球比例 | 数值 | 0.52-7.12 |
| 甘油三酯 | 数值 | 0.27-41.57 |
| 总胆固醇 | 数值 | 1.85-20.46 |
| 高密度脂蛋白胆固醇 | 数值 | 0.54-5.28 |
| 低密度脂蛋白胆固醇 | 数值 | 0.56-8.46 |
| 尿素 | 数值 | 1.5-13.39 |
| 肌酐 | 数值 | 39.43-177.42 |
| 尿酸 | 数值 | 118.67-774.06 |
| 乙肝表面抗原 | 数值 | 0.00-44.35 |
| 乙肝表面抗体 | 数值 | 0.00-42.49 |
| 乙肝e抗原 | 数值 | 0.01-17.52 |
| 乙肝e抗体 | 数值 | 0.00-7.17 |
| 乙肝核心抗体 | 数值 | 0.00-17.09 |
| 白细胞计数 | 数值 | 2.80-21.06 |
| 红细胞计数 | 数值 | 3.01-6.85 |
| 血红蛋白 | 数值 | 65-204 |
| 红细胞压积 | 数值 | 0.239-0.599 |
| 红细胞平均体积 | 数值 | 59-113 |
| 红细胞平均血红蛋白浓度 | 数值 | 16.00-44.70 |
| 红细胞体积分布宽度 | 数值 | 270-462 |
| 血小板计数 | 数值 | 37-745 |
| 血小板平均体积 | 数值 | 7.10-15.20 |
| 血小板体积分布宽度 | 数值 | 8.00-25.30 |
| 血小板比积 | 数值 | 0.042-0.71 |
| 中性粒细胞 | 数值 | 14.40-88.50 |
| 淋巴细胞 | 数值 | 7.50-76.30 |
| 单核细胞 | 数值 | 3.10-23.20 |
| 嗜酸细胞 | 数值 | 0.00-22.50 |
| 嗜碱细胞 | 数值 | 0.00-3.50 |
| 血糖 | 数值 | 3.07-38.43 |

## 数据清洗

### 缺失数据处理

对数据缺失情况进行统计，各变量缺失比例如下图。可以看出，除了患者的基本信息外，数据中绝大多数的列存在缺失值，其中乙肝相关检测值的缺失比例最高，分析可知，这是源于患者选择相关检测较少。



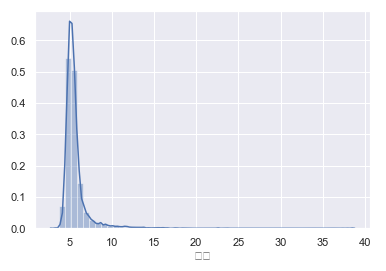
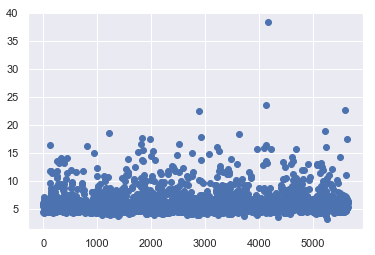
乙肝类检测值缺失比例超过了70%，故直接删除相关变量。此外，由于此次任务非时序数据预测任务，因此，单次‘体检日期’并不构成有效的时间特征，因此我们用相似的方法，将体检日期这一项特征也删除掉。

对于其他有缺失值的数据，由于缺失值所占比例较小，选择使用均值插补对其进行填充。

### 异常值处理

绘制出因变量血糖值的散点图和频数分布图如下，发现大多数血糖值集中在3~8之间，只有4个记录超过20，属于异常值，因此直接删除该记录所在的样本。

剔除之后，血糖值近似对称分布，可以认为此数据为平衡数据。



# 特征工程

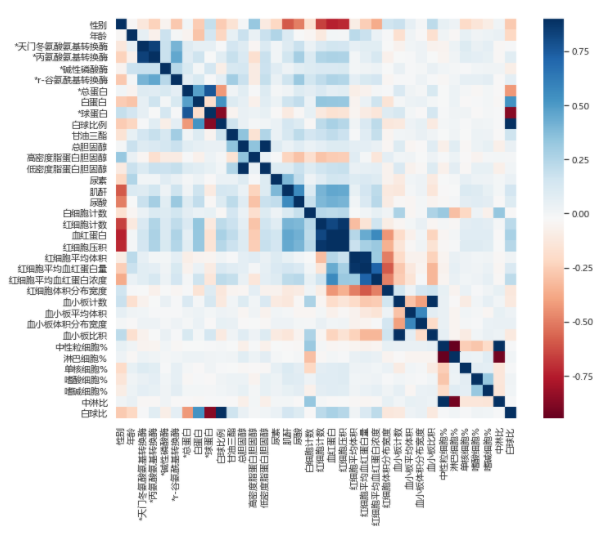
## 生成新变量

有文献证明中性粒细胞/淋巴细胞比值可用于评价多种疾病状态，因此我们可以根据已有的‘中性粒细胞’和‘淋巴细胞’属性生成新的特征‘中淋比’： '中淋比' = '中性粒细胞'/'淋巴细胞' 白球比是肝功能检查中的一项重要参考指标，反映的是肝脏的合成功能，在肝病的诊断上有重要意义。同样，根据已有的‘白蛋白’和‘球蛋白’属性生成新的特征‘白球比’： '白球比'= '白蛋白'/'球蛋白'。

## 变量相关性

对初步处理过后的36个指标进行相关系数的计算。pearson相关系数衡量的是线性相关关系。相关系数绝对值越接近于1，[相关度](https://baike.baidu.com/item/%E7%9B%B8%E5%85%B3%E5%BA%A6" \t "https://baike.baidu.com/item/Pearson%E7%9B%B8%E5%85%B3%E7%B3%BB%E6%95%B0/_blank)越强，相关系数越接近于0，相关度越弱。为了初步探究指标之间的相关性，用perason相关系数来了解概况，为后续的模型使用以及变量选择提供参照。

将相关系数矩阵绘图如下。其中，格子上的数值代表其对应指标的Kendall相关系数，格子越接近深色，表示问题结果相关性越高。从下图可以看出，大多数指标之间的相关系数在0-0.25之间，少数的问题之间相关系数达到0.7以上（）如红细胞计数、血红蛋白、红细胞压积）这些变量内部可能有潜在联系，因此可以考虑使用因子分析提取主要特征，对数据进行降维处理。



## 归一化处理

因为各变量的量纲不同，且取值范围相差很大，为了方便后续因子分析，对数据进行标准化并基于相关矩阵进行分析。

# 因子分析

## 因子分析简介

因子分析是指研究从变量群中提取[共性因子](https://baike.baidu.com/item/%E5%85%B1%E6%80%A7%E5%9B%A0%E5%AD%90/12716221" \t "https://baike.baidu.com/item/%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90/_blank)的[统计技术](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%8A%80%E6%9C%AF/7574831" \t "https://baike.baidu.com/item/%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90/_blank)。因子分析可在许多变量中找出隐藏的具有代表性的因子。将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设。

因子分析习惯性的做法是把因子的影响分解成公共因子和特殊因子，例如，反映维向量X的一般特征的为公共因子，仅反映X的个别成分的为特殊因子。即有惯用的因子分析模型：



式中，Q是一个维的矩阵，称为因子载荷矩阵；F是表示k个因子的k维向量；U是一个维的矩阵，称为特殊因子。随机变量F和U是不可观测、不相关的。

## bartlett球形检验和KMO检验

因子分析前，首先进行KMO检验和巴特利球体检验。KMO检验用于检查变量间的相关性和偏相关性，取值在0~1之间。如果原始数据中存在公共因子，则各变量之间的偏相关系数应该很小，这时，KMO的值接近于1，认为原数据适用于因子分析。实际分析中，KMO统计量在0.7以上时效果比较好；当KMO统计量在0.5以下，此时不适合应用[因子分析法](https://baike.baidu.com/item/%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90%E6%B3%95/6427615" \t "https://baike.baidu.com/item/Bartlett%27s%E7%90%83%E7%8A%B6%E6%A3%80%E9%AA%8C/_blank)，应考虑重新设计变量结构或者采用其他统计分析方法。计算公式如下：



如果变量间彼此独立，则无法从中提取公因子，也就无法应用因子分析法。Bartlett球形检验判断变量的相关阵是否为单位阵，原假设认为相关系数矩阵是单位矩阵。假设有r组变量，每组有同样多的数据，统计量近似服从r-1的卡方分布，p值<0.05时，拒绝原假设，认为数据呈球形分布，各个变量在一定程度上相互独立。计算公式如下：



对上述处理后的变量进行检验，检验结果如下：

|  |  |  |
| --- | --- | --- |
|  | | |
| KMO检验 | 0.5125 | |
| 巴特利球形度检验 | 近似卡方 | 显著性 |
| 518667.18 | 0.00 |

KMO检验的结果>0.5,且巴特利球形度检验的p值小于0.05,因此该数据适合做因子分析。

## 选择因子个数

因子分析中有很多确定因子变量的方法，如基于主成分模型的主成分分析和基于因子分析模型的主轴因子法、极大似然法、最小二乘法等，前者应用最为广泛。为了方便解释，本文基于主成分并采用方差最大旋转进行变量选择。

### 主成分法

主成分分析中一般取样本协方差矩阵S前k个最大的特征根，构造因子载荷矩阵Q的近似；在应用中，通常取特征值大于1对应的因子，或累计方差贡献率超过80%的前几个特征向量作为因子。

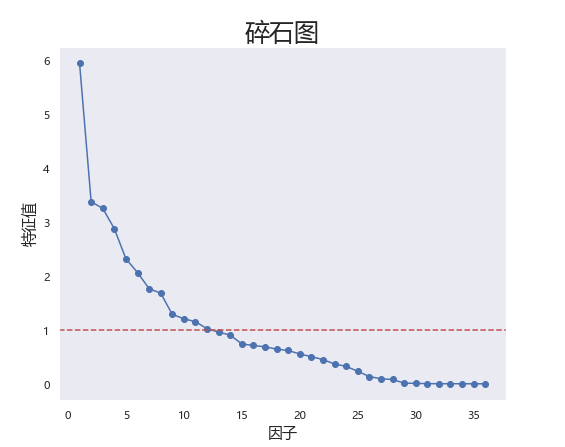
### 因子旋转

因为载荷矩阵和因子得分矩阵的不唯一性，可以通过正交变换使得载荷矩阵更具有解释性。一种常用的旋转方法是Kaiser提出的方差最大旋转法，即使得任何一个因子只在少数变量上有高载荷,而在其它变量上的载荷几乎为0。

### 碎石图

碎石检验是根据碎石图来决定因素数的方法。R．B．卡特尔提出。在[主轴](https://baike.baidu.com/item/%E4%B8%BB%E8%BD%B4/3198308" \t "https://baike.baidu.com/item/%E7%A2%8E%E7%9F%B3%E6%A3%80%E9%AA%8C/_blank)因素法中，公共因素的方差贡献等于该因素对应的特征根的值，计算累积方差贡献率可用计算累积特征根百分比来代替，甚至可通过直接观察特征根的变化来决定因素数。当某个特征根较前一特征根的值出现较大的下降，而这个特征根较小，其后面的特征根变化不大，说明添加相应于该特征根的因素只能增加很少的信息，所以前几个特征根就是应抽取的公共因素数。

对数据基于主成分进行因子分析并进行最大方差旋转后，作出碎石图如下：



因此根据碎石图曲线的下降斜率和各特征值大小，选择因子个数为12。

## 因子解释

各因子对应的特征值和方差贡献表如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 因子 | 特征值 | 方差贡献 | 累计方差贡献 | 方差贡献占比 |
| 1 | 5.96 | 0.109 | 0.109 | 0.141 |
| 2 | 3.38 | 0.094 | 0.204 | 0.121 |
| 3 | 3.25 | 0.083 | 0.287 | 0.106 |
| 4 | 2.87 | 0.082 | 0.369 | 0.106 |
| 5 | 2.32 | 0.067 | 0.437 | 0.086 |
| 6 | 2.06 | 0.061 | 0.498 | 0.078 |
| 7 | 1.76 | 0.058 | 0.557 | 0.075 |
| 8 | 1.68 | 0.049 | 0.606 | 0.063 |
| 9 | 1.28 | 0.048 | 0.654 | 0.062 |
| 10 | 1.21 | 0.046 | 0.701 | 0.059 |
| 11 | 1.15 | 0.038 | 0.739 | 0.049 |
| 12 | 1.01 | 0.037 | 0.777 | 0.048 |

即12个因子能解释原始数据77%的方差，包含了较大部分的信息。

因子载荷如下：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | fac1 | fac2 | fac3 | fac4 | fac5 | fac6 | fac7 | fac8 | fac9 | fac10 | fac11 | fac12 |
| 性别 | -0.75 | -0.14 | -0.11 | -0.01 | -0.14 | 0.05 | 0.12 | -0.35 | -0.02 | 0.05 | -0.11 | -0.03 |
| 年龄 | -0.07 | -0.23 | 0.15 | 0.06 | -0.01 | 0.29 | -0.19 | 0.34 | -0.45 | 0.01 | 0.00 | -0.08 |
| 天门冬氨酸氨基转换酶 | 0.02 | -0.03 | 0.04 | -0.01 | 0.91 | 0.01 | -0.05 | -0.01 | 0.02 | 0.01 | 0.02 | -0.13 |
| 丙氨酸氨基转换酶 | 0.16 | 0.04 | 0.02 | -0.01 | 0.90 | 0.01 | 0.01 | 0.00 | 0.09 | 0.00 | 0.02 | -0.02 |
| 碱性磷酸酶 | 0.11 | -0.17 | 0.00 | 0.00 | 0.31 | 0.20 | 0.07 | 0.04 | -0.13 | 0.03 | -0.09 | 0.17 |
| r-谷氨酰基转换酶 | 0.14 | 0.03 | 0.06 | -0.01 | 0.64 | 0.13 | 0.01 | 0.10 | 0.03 | 0.00 | 0.07 | 0.20 |
| 总蛋白 | 0.07 | -0.67 | -0.01 | -0.02 | 0.07 | 0.09 | 0.05 | 0.05 | 0.70 | 0.02 | -0.03 | 0.01 |
| 白蛋白 | 0.22 | 0.28 | 0.05 | -0.01 | 0.06 | 0.06 | 0.01 | 0.09 | 0.87 | 0.04 | -0.03 | 0.02 |
| 球蛋白 | -0.08 | -0.96 | -0.05 | -0.01 | 0.03 | 0.06 | 0.04 | -0.01 | 0.15 | 0.00 | 0.00 | 0.00 |
| 白球比例 | 0.13 | 0.94 | 0.06 | -0.01 | 0.02 | 0.04 | -0.04 | 0.04 | 0.24 | 0.02 | 0.00 | 0.04 |
| 甘油三酯 | 0.08 | 0.11 | 0.00 | -0.04 | 0.24 | 0.34 | -0.06 | 0.10 | 0.02 | 0.01 | 0.06 | 0.59 |
| 总胆固醇 | 0.01 | 0.01 | 0.03 | -0.05 | 0.09 | 0.97 | 0.04 | 0.03 | 0.02 | -0.02 | 0.03 | 0.04 |
| 高密度脂蛋白胆固醇 | -0.34 | 0.05 | 0.06 | 0.04 | -0.13 | 0.42 | -0.01 | -0.16 | 0.08 | 0.03 | -0.02 | -0.50 |
| 低密度脂蛋白胆固醇 | 0.14 | -0.03 | 0.01 | -0.06 | 0.12 | 0.87 | 0.08 | 0.08 | -0.01 | -0.04 | 0.03 | 0.12 |
| 尿素 | 0.01 | 0.01 | 0.08 | 0.01 | -0.04 | 0.11 | -0.02 | 0.78 | -0.04 | 0.05 | -0.04 | -0.11 |
| 肌酐 | 0.41 | 0.07 | 0.03 | 0.00 | 0.07 | -0.03 | -0.06 | 0.72 | 0.04 | -0.02 | 0.06 | 0.06 |
| 尿酸 | 0.37 | 0.01 | 0.01 | -0.02 | 0.26 | -0.02 | 0.01 | 0.54 | 0.13 | -0.06 | 0.09 | 0.26 |
| 白细胞计数 | 0.27 | -0.08 | 0.05 | 0.37 | 0.08 | 0.00 | 0.42 | 0.03 | -0.06 | 0.04 | -0.02 | 0.39 |
| 红细胞计数 | 0.94 | 0.08 | -0.20 | 0.01 | 0.09 | 0.06 | 0.01 | 0.04 | 0.13 | 0.02 | 0.02 | 0.03 |
| 血红蛋白 | 0.88 | 0.10 | 0.37 | 0.00 | 0.12 | 0.07 | -0.09 | 0.09 | 0.10 | -0.02 | 0.04 | 0.05 |
| 红细胞压积 | 0.92 | 0.08 | 0.23 | 0.01 | 0.11 | 0.09 | 0.00 | 0.08 | 0.07 | 0.03 | 0.02 | -0.03 |
| 红细胞平均体积 | -0.09 | 0.00 | 0.88 | -0.01 | 0.04 | 0.04 | -0.03 | 0.08 | -0.14 | 0.02 | 0.01 | -0.12 |
| 红细胞平均血红蛋白量 | 0.05 | 0.05 | 0.95 | -0.02 | 0.06 | 0.03 | -0.17 | 0.10 | -0.04 | -0.06 | 0.05 | 0.05 |
| 红细胞平均血红蛋白浓度 | 0.25 | 0.11 | 0.62 | -0.02 | 0.07 | 0.00 | -0.29 | 0.06 | 0.13 | -0.14 | 0.07 | 0.27 |
| 红细胞体积分布宽度 | -0.22 | -0.05 | -0.74 | 0.05 | 0.02 | -0.01 | 0.00 | 0.05 | -0.09 | 0.03 | 0.06 | -0.01 |
| 血小板计数 | -0.09 | -0.06 | -0.16 | 0.01 | -0.01 | 0.07 | 0.88 | -0.06 | 0.06 | -0.33 | 0.05 | 0.04 |
| 血小板平均体积 | -0.03 | 0.01 | -0.05 | 0.02 | 0.00 | -0.03 | 0.03 | 0.02 | 0.01 | 0.91 | -0.04 | -0.05 |
| 血小板体积分布宽度 | 0.02 | 0.02 | -0.09 | 0.04 | 0.02 | -0.01 | -0.25 | -0.01 | 0.03 | 0.83 | 0.06 | 0.10 |
| 血小板比积 | -0.10 | -0.06 | -0.19 | 0.01 | -0.01 | 0.06 | 0.93 | -0.05 | 0.06 | 0.01 | 0.03 | 0.02 |
| 中性粒细胞 | -0.03 | 0.01 | -0.03 | 0.96 | -0.04 | -0.03 | 0.03 | -0.02 | 0.01 | 0.02 | -0.19 | 0.09 |
| 淋巴细胞 | -0.03 | 0.00 | 0.05 | -0.97 | 0.00 | 0.04 | -0.01 | -0.02 | 0.02 | -0.01 | -0.04 | 0.02 |
| 单核细胞 | 0.15 | -0.02 | -0.07 | -0.15 | 0.17 | -0.06 | -0.13 | 0.17 | -0.08 | -0.04 | 0.12 | **-**0.53 |
| 嗜酸细胞 | 0.08 | 0.01 | 0.00 | -0.09 | 0.02 | -0.01 | 0.03 | 0.03 | -0.05 | 0.01 | 0.79 | 0.00 |
| 嗜碱细胞 | 0.03 | 0.00 | 0.02 | -0.08 | 0.03 | 0.05 | 0.02 | 0.01 | 0.02 | 0.00 | 0.80 | -0.04 |
| 中淋比 | -0.01 | 0.00 | -0.02 | 0.96 | -0.02 | -0.03 | 0.01 | 0.01 | -0.02 | 0.02 | -0.07 | 0.02 |
| 白球比 | 0.14 | 0.94 | 0.06 | -0.01 | 0.01 | 0.04 | -0.05 | 0.04 | 0.24 | 0.02 | 0.00 | 0.03 |

因子1与红细胞计数和压积正相关，可以大体解释为“红细胞数量因子”；

因子2与球蛋白负相关，白球比正相关，可以大体解释为“白球比因子”；

因子3与红细胞平均血红蛋白量和平均体积正相关，可以大体解释为“血红蛋白因子”；

因子4与中性粒细胞正相关，和淋巴细胞负相关，可以大体解释为“中淋比因子”；

因子5与各类肝功能相关酶正相关，可以大体解释为“肝功能因子”；

因子6与各类胆固醇正相关，可以大体解释为“胆固醇因子”；

因子7与血小板计数和比积正相关，可以大体解释为“血小板数量因子”；

因子8与尿酸、尿素、肌酐正相关，可以大体解释为“肾功能指标”；

因子9与白蛋白和总蛋白正相关，可以大体解释为“白蛋白因子”；

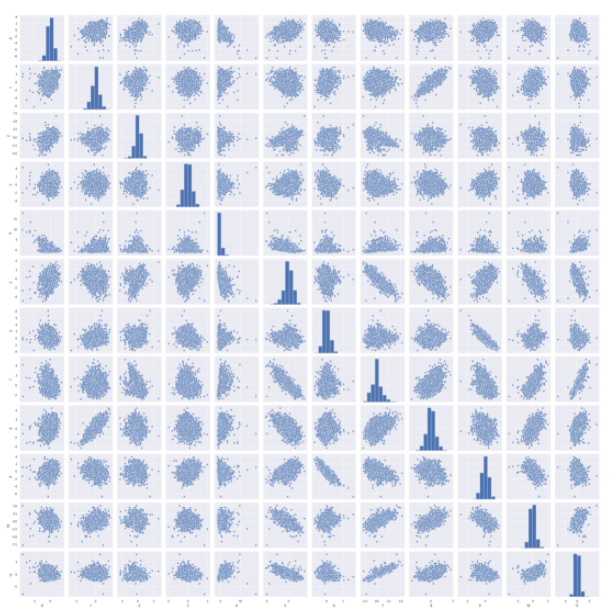
因子10与血小板平均体积和体积分布宽度正相关，可以大体解释为“血小板体积因子”；

因子11与嗜酸细胞和嗜碱细胞正相关，可以大体解释为“嗜酸嗜碱细胞因子”；

因子12各个变量上的载荷都很小，因此其解释较为困难，只能勉强解释为“单核细胞因子”。

总体来说，因子解释意义与医学常识相符，反映相同器官功能的指标往往可以用相同的公共因子解释。但仍有一些信息没有得到充分解释，比如患者的性别、年龄等。

降维后的因子两两特征图如下，显然因子是互不相关的，且大体分布均衡，方便后续的回归分析。



# 回归分析

## 模型选择

### 多元线性回归

我们把包括两个或两个以上自变量的回归称为多元线性回归。由于各个自变量的单位可能不一样，因此通常先将所有变量都标准化，再进行线性回归，此时得到的[回归系数](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E7%B3%BB%E6%95%B0" \t "https://baike.baidu.com/item/%E5%A4%9A%E5%85%83%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92/_blank)就能反映对应自变量的重要程度。这时的[回归方程](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E6%96%B9%E7%A8%8B" \t "https://baike.baidu.com/item/%E5%A4%9A%E5%85%83%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92/_blank)称为标准回归方程，回归系数称为[标准回归系数](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E5%9B%9E%E5%BD%92%E7%B3%BB%E6%95%B0" \t "https://baike.baidu.com/item/%E5%A4%9A%E5%85%83%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92/_blank)，表示如下：



多元线性回归与一元线性回归类似，可以用[最小二乘法](https://baike.baidu.com/item/%E6%9C%80%E5%B0%8F%E4%BA%8C%E4%B9%98%E6%B3%95" \t "https://baike.baidu.com/item/%E5%A4%9A%E5%85%83%E7%BA%BF%E6%80%A7%E5%9B%9E%E5%BD%92/_blank)估计模型参数，也需对模型及模型参数进行统计检验。

### 随机森林

随机森林由多棵决策树构成，且森林中的每一棵决策树之间没有关联，模型的最终输出由森林中的每一棵决策树共同决定。

**处理分类问题时**，对于测试样本，森林中每棵决策树会给出最终类别，最后综合考虑森林内每一棵决策树的输出类别，以投票方式来决定测试样本的类别；**处理回归问题时**，则以每棵决策树输出的均值为最终结果。

### XGBoost

XGBoost 是 “Extreme Gradient Boosting”的简称，其实现的是一种通用的Tree Boosting算法，此算法的一个代表为梯度提升决策树（Gradient Boosting Decision Tree, GBDT），又名MART(Multiple Additive Regression Tree)。

GBDT的原理是，首先使用训练集和样本真值（即标准答案）训练一棵树，然后使用这棵树预测训练集，得到每个样本的预测值，由于预测值与真值存在偏差，所以二者相减可以得到“残差”。接下来训练第二棵树，此时不再使用真值，而是使用残差作为标准答案。两棵树训练完成后，可以再次得到每个样本的残差，然后进一步训练第三棵树，以此类推。树的总棵数可以人为指定，也可以监控某些指标（例如验证集上的误差）来停止训练。在预测新样本时，每棵树都会有一个输出值，将这些输出值相加，即得到样本最终的预测值。

使用梯度提升算法的好处是在提升树被创建后，可以相对直接地得到每个属性的重要性得分。一般来说，重要性分数，衡量了特征在模型中的提升决策树构建中价值。一个属性越多的被用来在模型中构建决策树，它的重要性就相对越高。属性重要性是通过对数据集中的每个属性进行计算，并进行排序得到。在单个决策树中通过每个属性分裂点改进性能度量的量来计算属性重要性，由节点负责加权和记录次数。也就说一个属性对分裂点改进性能度量越大（越靠近根节点），权值越大；被越多提升树所选择，属性越重要。性能度量可以是选择分裂节点的Gini纯度，也可以是其他度量函数。

最终将一个属性在所有提升树中的结果进行加权求和后然后平均，得到重要性得分。

## 实证分析

为了方便比较各因子的重要程度，本文将各公共因子标准化后分别使用多元线性回归、随机森林对数据进行实证分析。回顾模型中的因变量、自变量如下：

1. 因变量：患者的血糖值。
2. 自变量：红细胞数量因子，白球比因子，血红蛋白因子，中淋比因子，肝功能因子，胆固醇因子，血小板数量因子、肾功能因子、白蛋白因子、血小板体积因子、嗜酸嗜碱细胞因子、单核细胞因子。

多元线性回归得到的回归结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 自变量 | 回归系数 | p值 | 自变量 | 回归系数 | p值 |
| 红细胞数量因子 | 0.2176  (0.115) | 0.059  . | 血小板数量因子 | -0.3542  (0.267） | 0.185 |
| 白球比因子 | -0.0371  (0.238) | 0.876 | 肾功能因子 | 0.5381  （0.259） | 0.037  \*\* |
| 血红蛋白因子 | 0.0126  (0.117) | 0.914 | 白蛋白因子 | -0.0746  （0.273） | 0.785 |
| 中淋比因子 | 0.0283  (0.086) | 0.742 | 血小板体积因子 | -0.3484  （0.301） | 0.246 |
| 肝功能因子 | 0.4956  (0.133) | <0.001  \*\*\* | 嗜酸嗜碱细胞因子 | 0.1179  （0.142） | 0.406 |
| 胆固醇因子 | 0.8344  (0.214) | <0.001  \*\*\* | 单核细胞因子 | -0.1494  （0.206） | 0.468 |

注: 括号内表示回归系数估计值的标准误差，\*\*\*、\*\*、\*、.分别表示回归系数在0. 1%、1%、5%、10% 的显著性水平下显著。

|  |  |
| --- | --- |
| 模型拟合优度检验 | |
| 调整 | 0.005 |
| F统计量 | 3.465 |
| AIC | 35790 |
| BIC | 35870 |

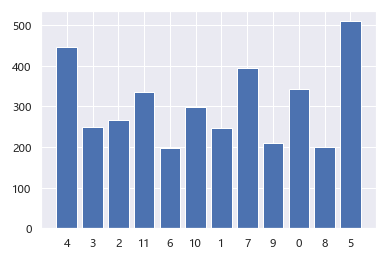
回归结果显示胆固醇指标对血糖影响最大，其次是肝功能指标，然后是肾功能指标。胆固醇主要影响血脂，查阅资料发现高血糖和高血脂没有直接关系，但很多的患者可能同时具有这两种疾病，因此胆固醇指标具有很大参考价值。肝功能也能影响血糖，肝脏内有合成糖类的功能，如果身体内的血糖量不充足时，肝脏就会分解原来形成的肝糖原对身体内的血糖进行补充；如果身体中血糖较多时，肝脏可以将血糖转化为肝糖原，从而对身体的血糖起到调节作用，使血糖在身体内保持稳定。对于肾功能，肾病患者更容易出现低血糖，肾功能不全的患者，肾脏对于胰岛素和一些口服降糖药的清除率会下降，导致患者发生低血糖的风险会增加。其他的指标如血小板指标、红细胞指标、白细胞指标等则不显著。

但总体来看，线性回归模型的拟合优度并不好，这说明血糖值与各指标之间可能不只是简单的线性关系。

为了尝试提高预测精度并进一步验证特征重要度，接下来使用集成方法随机森林和XGBoost在训练集上训练，并在测试集上检验精度，得到各模型的MSE如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | 多元线性回归 | 随机森林 | XGBoost |
| 训练集 | 1.792 | 0.273 | 2.102 |
| 测试集 | 3.778 | 3.141 | 2.653 |

可以看到，三个模型中XGBoost的精度最高，其得出的特征重要性如下图：



可以看到，重要度最高的前三个因子依次为5,4,7，即胆固醇因子、 肝功能因子和肾功能因子，这与前面线性回归得到的变量系数和显著程度相符。

# 总结

本文使用糖尿病公开数据集，对各指标进行因子分析并对血糖值进行预测，从而希望能够分析各指标之间的联系，并探究公共因子对血糖值的影响程度，尝试建立糖尿病预测模型。本文首先介绍了糖尿病的历史背景以及机器学习在慢性病初步诊断上的应用；其次对数据进行了一系列预处理和统计描述，展现了变量的分布特点；之后基于数据的特点，进行基于主成分和最大方差旋转的因子分析，找到各指标的内在联系；最终对提取出的公共因子和因变量血糖值分别用线性回归、随机森林和XGBoost进行回归分析，找到影响血糖值的关键因素。研究结果表明，血脂、肝功能和肾功能指标是诊断糖尿病最关键的三个因素，具有重要参考价值，可以根据这些指标在糖尿病诊断上进一步提高效率。