基于机器学习的肥胖预测和相关性研究

摘要

肥胖是多种并发症的公认危险因素，包括2型糖尿病、血脂异常、高血压、心血管疾病、非酒精性脂肪性肝病、骨关节炎等。根据世界卫生组织WHO统计，超重和肥胖是全球引起死亡的第六大风险。每年，至少有340万成年人死于超重或肥胖。65%的全世界人口所居住的国家，其超重和肥胖死亡人数大于体重不足引起死亡的人数，其中包括了所有高收入和大部分中等收入国家。

为了研究这个问题并为预防肥胖症提出可行的建议，本文主要建立在机器学习的基础上，使用了XGBoost算法等，探究诸多因素与肥胖症的关联性，得出其中影响最为显著的一批因素，并将它们绘制成图表的形式呈现。

结果显示，男性更加容易超重和肥胖，此外，蔬菜的食用频率和两餐之间的食物摄取量和肥胖也有较大的关系，而遗传因素也有着不小的关联性。

关键词：机器学习，肥胖症，相关性研究

表格与插图清单

图1.各变量相关性的热力图

图2.各数值特征的分类直方图

图3.各数值特征的箱线图

图4.各分类特征的饼状图

图5.XGBoost参数最优值

图6.特征重要性

表1.七种模型及其简单介绍

表2.交叉验证的参数

表3.各模型得分

表4.XGBoost模型参数解释

表5.搜索参数空间范围

目录

[摘要 1](#_Toc166425149)

[表格与插图清单 2](#_Toc166425150)

[一、 引言 5](#_Toc166425151)

[二、基于机器学习的模型建立 5](#_Toc166425152)

[（一）数据的预处理 5](#_Toc166425153)

[（二）图形绘制及分析 5](#_Toc166425154)

[（三）模型选择 8](#_Toc166425155)

[（四）模型比较 9](#_Toc166425156)

[1.交叉验证的原理及以及在本文中的应用 9](#_Toc166425157)

[2.七种模型之间的比较 10](#_Toc166425158)

[3.XGBoost模型介绍 10](#_Toc166425159)

[4.XGBoost模型超参数调优 11](#_Toc166425160)

[（1）参数空间与参数解释 11](#_Toc166425161)

[（2）最优结果及对应参数 12](#_Toc166425162)

[三、结果分析 13](#_Toc166425163)

一、引言

肥胖症，或称肥胖，是由可能导致健康损害的脂肪过量堆积而定义的一种慢性复杂疾病。定义肥胖的体质指数分类会因人的年龄阶段和性别而异。一般而言，通过测量人的体重和身高，并计算体重指数BMI，就可做出超重和肥胖的诊断。体重指数BMI是肥胖的替代标志，腰围等其他测量值则可帮助做出肥胖诊断。

根据世界卫生组织（WHO）的数据显示，自1990年以来，全球成年人的肥胖率增加了一倍多，青少年的肥胖率则增加了三倍。截止到2022年，全世界有25亿（18岁及以上）的成年人，也就是43%的全球成年人超重。其中8.9亿人，也就是全球成年人口的16%罹患有肥胖症。超过3.9亿的5-19岁儿童和青少年超重，其中1.6亿患有肥胖症。

肥胖症是多种心血管疾病，糖尿病以及炎症等的并发危险因素。同时，在现代社会，由于科学技术的发展，人们在生活和工作中进行静态活动的时间明显增长，且肥胖症属于慢性疾病，在并发疾病产生时往往已经变得难以处理，使得它成为了现代人的一大健康杀手。

对儿童和青少年而言，肥胖症的影响和危害更大。研究表明，儿童的肥胖会导致骨龄发育提前，促使他们骨骼的干骺端提前闭合，也就使得身高发育不及预期。同时，肥胖会使得儿童的通气功能受到影响，让他们的肺活量和肺功能明显下降，最终可能导致儿童哮喘和睡眠障碍的产生。

因而，本文将利用机器学习对数据进行相关性分析。研究得出的结论，以对肥胖病的相关影响因素加深了解，依据这些结果为决策者提供合理建议和对策，以求有助于对我国日益严重的肥胖问题进行一定程度上的缓解。

二、基于机器学习的模型建立

（一）数据的预处理

原数据中存在一列与研究内容不相关的数据，故而将其去掉。此外，利用代码输出原数据中的缺失行以及重复行的数量，同时检验每一列中是否存在异常值。之后，依据列的数据不同，利用函数将原数据划分为训练集和测试集。将原数据中的最后一列作为目标变量，再对数据类型进行判别，找出原数据中的文本类型变量，将它们以及一些比较特殊的数值型变量，一起归为分类特征列。剩余的数值变量则归为数值特征列。

（二）图形绘制及分析

为预先了解各数值变量之间的相关性大小，先对其内部绘制热力图[图1]，根据结果显示，与体重较为相关的变量由高至低依次为，身高、年龄、每日水分摄取量以及蔬菜摄取量。图形用户界面

描述已自动生成

图1 各变量相关性的热力图

为了直观了解各数值特征的分布情况，对其绘制分布直方图[图2]以及箱线图[图3]。为了了解各分类特征的分布情况，则对其绘制饼状图[图4]。由结果可知，原数据中男女性别对半，其中的大多数人生活习惯不算健康。这反映了现代社会中，人们的整体健康意识正在集体下滑。一部分人养成了许多不良习惯，进而促进了并发症的产生。

图表, 直方图

描述已自动生成

图2 各数值特征的分类直方图

图表, 箱线图

描述已自动生成

图3 各数值特征的箱线图

图表, 气泡图

描述已自动生成

图4 各分类特征的饼状图

（三）模型选择

基于本文的研究目的和所用数据的特征，本文选取了七个常见的可用于解决分类和回归问题机器学习模型。

七种模型及其简单介绍：

表1 七种模型及其简单介绍

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 用途 | 原理 | 特点 |
| Logistic Regression（逻辑回归） | 分类问题，尤其是二分类问题 | 通过线性函数，加上一个sigmoid函数来预测事件发生的概率 | 简单、计算效率高， 但不适合处理复杂的非线性关系 |
| Decision Tree（决策树） | 分类和回归问题 | 通过树状结构来模拟决策过程， 通过一系列规则来预测输出 | 直观、易于理解，但容易过拟合 |
| Extra Tree（额外树） | 分类和回归问题 | 一种随机化的决策树，通过随机选择特征来构建树的各个分支 | 通常比标准决策树更不易过拟合，但单个模型的性能可能较低 |
| Random Forest（随机森林） | 分类和回归问题 | 集成学习算法，由多个决策树组成，每个树在训练集的一个随机子集上训练 | 通常性能好于单个决策树，更稳定，但也更难以解释 |
| Ensemble Extra Trees（集成额外树） | 分类和回归问题 | 类似于随机森林，但是它使用的是额外树而不是标准的决策树 | 通常提供更好的性能和更高的稳定性 |
| XGBoost（极端梯度提升） | 分类和回归问题 | 是一种高效的梯度提升框架，使用树来构建模型 | 速度快，性能好，提供正则化项以防止过拟合 |
| SVM（支持向量机） | 分类和回归问题 | 找到最优的分割超平面，使得样本间的间隔最大化 | 对于小数据集表现良好，但计算复杂度较高，对于大规模数据集可能不够高效 |

（四）模型比较

因为本文的研究目的是找到对肥胖影响最大的特征，而不同模型得到的特征重要性得分会存在差异，所以本文将使用交叉验证法评估七种模型性能，选择得分最高的模型进行最后的特征重要性分析。

1. 交叉验证的原理及以及在本文中的应用

本文使用的是交叉验证中的分层K折交叉验证（StratifiedKFold），这种方法的原理是将原数据集等分为k组，一次实验取k-1组为训练集，剩下的一组为测试集，重复k次实验，取k次实验得分的均值为该模型在K折交叉验证中的最终得分。

分层K折交叉验证的参数及说明：

表2 交叉验证的参数

|  |  |  |
| --- | --- | --- |
| n\_splits | shuffle | shuffle |
| 5 | TRUE | 42 |

n\_splits=5：表示k为5，即进行5次实验。

shuffle=True：在每次选择不同的训练集和测试集之前，随机打乱数据集，有助于确保模型不会对数据集中的特定顺序产生依赖，从而提高了模型的泛化能力。

random\_state=42：这个参数用于保证实验的可重复性，去掉后将会生成随机数，使每次的结果都不同

1. 七种模型之间的比较

在比较不同模型对于本文所用数据的适用性时，分层K折交叉验证可以更加全面地评估模型在未见数据上的表现，并且能够克服单一测试集引起的评估结果的不稳定，从而更加科学地对各个模型的适用性做出判断。

各模型得分：

表3 各模型得分

|  |  |
| --- | --- |
| 模型名称 | 交叉验证F1得分 |
| Logistic Regression | 0.8224910204244449 |
| Decision Tree | 0.8264879049593091 |
| Extra Tree | 0.7151779084218829 |
| Random Forest | 0.8814805933207882 |
| Ensemble Extra Trees | 0.8520167091111424 |
| XGBoost | 0.8963440503131169 |
| SVM | 0.8374831108663361 |

上表表明XGBoost在本文数据中最为适用，故本文选定该模型进行特征重要性分析。

1. XGBoost模型介绍

XGBoost模型是Boosting模型中的一种。它通过加法模型和前向分步算法将多棵树模型串联起来形成一个适用性更强的新模型。因为在确定每一棵数的叶子节点的数目和值时使用的是前向分布算法，所以没有一个全过程的模型建立表达式。

以第t棵数为例，其建立过程的**目标函数（Objective Function）为：**

其中，N为样本数，表示损失函数，是真实值，是预测值，t表示第几棵树，是第t棵树的正则化项[[1]](#footnote-1)，公式如下

其中𝑇是第t棵树的叶子节点的数量，𝛾是每棵树的复杂度参数，𝑤𝑗 是树t中第𝑗个叶子节点的值，*λ*是 L2 正则化项（权重的平方和）。

1. XGBoost模型超参数调优

本文选择在超参数空间中进行搜索选择在分层K折交叉验证中获得F1最高得分的参数组合作为最优参数。

（1）参数空间及参数解释

表4 XGBoost模型参数解释

|  |  |
| --- | --- |
| 参数名 | 参数含义 |
| max\_depth | XGBoost 树的最大深度 |
| n\_estimators | 树的数目 |
| subsample | 训练时每棵树使用的数据样本比例 |
| colsample\_bytree | 训练时每棵树使用的feature比例 |
| reg\_alpha | L1 正则化项 |
| reg\_lambda | L2 正则化项 |
| max\_leaves | 树的最大叶子节点数 |
| learning\_rate | 学习率，也称为步长 |
| max\_bin | 用于数值特征的直方图的桶数量 |

表5 搜索参数空间范围

|  |  |  |  |
| --- | --- | --- | --- |
| 参数名 | 参数区间起始点 | 参数区间终点 | 参数区间步长 |
| max\_depth | 4 | 10 | 1 |
| n\_estimators | 300 | 600 | 100 |
| subsample | 0.5 | 1 | 0.1 |
| colsample\_bytree | 0.4 | 1 | 0.1 |
| reg\_alpha | 0 | 2\*e-1 | 0.025 |
| reg\_lambda | 0 | 2\*e-1 | 0.25 |
| max\_leaves | 8 | 30 | 2 |
| learning\_rate | 0.001 | 0.3 | 0.05 |
| max\_bin | 256 | 2096 | 16 |

（2）最优结果及对应参数

表6 XGBoost参数最优值及其得分

|  |  |
| --- | --- |
| 参数名及得分 | 参数空间中的最优取值 |
| max\_depth | 9 |
| n\_estimators | 600 |
| subsample | 0.7 |
| colsample\_bytree | 0.4 |
| reg\_alpha | 0.175 |
| reg\_lambda | 0.125 |
| max\_leaves | 24 |
| learning\_rate | 0.051 |
| max\_bin | 944 |
| F1得分 | 0.900726452796027 |

**三、 结果分析**

根据最终选定的XGBoost模型，本文得到了以下结果：

图片包含 日程表

描述已自动生成

图5 特征重要性

上图表明，对肥胖有着重要影响的特征有：

Gender\_Male, FCVC（蔬菜食用频率）, CAEC（两餐之间摄取食物的频率）, family\_history\_with\_overwieght（家庭肥胖史）, FAVC（经常食用高热量食物）, CALC（饮酒频率）, CH2O（饮水量），Age(年龄), TUE（使用电子设备的时间）等[[2]](#footnote-2)。

上述对肥胖有着重要影响的特征可以分为两类，其一为客观因素，其二为主观习惯。

从客观因素，如Gender\_Male可以看出，性别与肥胖的关系非常显著，这与我国的现状相符合[[3]](#footnote-3)，从科学角度来看，女性储存脂肪的能力实际上比男性更强，然而男性反而更容易发胖，这或许与现代社会中男性群体的客观处境以及由此衍生的主观习惯有关，例如男性饮酒频率远比女性更高[[4]](#footnote-4)。但无论如何，本文都建议男性群体应该更加注重自己的身体情况，防止肥胖的产生。

从主观习惯可以看出，一些通常人们认为会影响肥胖的因素确实会影响到肥胖和自身健康。所以本文建议人们养成规律良好的饮食安排习惯，少吃高热量食品，多吃蔬菜，将更有利于自身身体健康。

另外，从上表可见family\_history\_with\_overwieght（家庭肥胖史）这一特征确实对个体是否发展成肥胖有显著的影响。遗传因素在肥胖的发生中起着重要作用，因为肥胖具有遗传倾向性。家族中有肥胖成员的个体，其肥胖的风险会相应增加。此外，共享的家庭环境和生活习惯，如饮食模式和体力活动水平，也可能对个体的体重有重要影响。

因此，对于有家族肥胖史的个体，本文建议其更应该采取预防措施和健康的生活方式对减少肥胖风险。从本文的研究成果来看，这重点包括改善饮食习惯和使用电子设备的时间。

此外，利用得出的这些特征的相关性，研究者也可以将它们作为指标，利用预测模型对人群进行肥胖症风险评估。

1. 正则项的加入是为了防止每棵树模型出现过拟合现象。 [↑](#footnote-ref-1)
2. 注：以上部分特征变量并不是所选数据记录的原始特征而是经过独热编码后所得的新的特征。且因体重这一特征缺少单独研究的价值，所以不考虑在内。 [↑](#footnote-ref-2)
3. 根据《2023中国肥胖地图》的数据，超重的男性占比41.1%，而女性占比27.7%；肥胖的男性占比18.2%，女性占比9.4% [↑](#footnote-ref-3)
4. 2012年中国居民营养与健康状况监测结果显示，我国成年居民饮酒率为32.8%，其中男性的饮酒率为52.6%，而女性饮酒率远低于男性，仅为12.4% [↑](#footnote-ref-4)