形状

中度可信度描述已自动生成

卡通人物画

中度可信度描述已自动生成

**机器学习**

题 目 导致肥胖的主要因素

学生姓名 刘自闲

专业班级 大数据管理与应用 21

学 号 2107120124

完成日期 20240711

目录

[第一章 数据准备 2](#_Toc171543436)

[第一节 数据获取 2](#_Toc171543437)

[第二节 数据描述 2](#_Toc171543438)

[第一章 数据清洗 4](#_Toc171543439)

[第一节 异常值处理 4](#_Toc171543440)

[第二节 标签重新分类 6](#_Toc171543441)

[第三节 缺失值处理 6](#_Toc171543442)

[第四节 重复值处理 7](#_Toc171543443)

[第五节 数据类型转换 7](#_Toc171543444)

[第三章 探索性数据分析 8](#_Toc171543445)

[第一节 描述性分析 8](#_Toc171543446)

[第二节 数据归一化 16](#_Toc171543447)

[第三节 直方图与Q-Q图分析 17](#_Toc171543448)

[第四节 KDE分布图 18](#_Toc171543449)

[第五节 双变量分析 19](#_Toc171543450)

[第六节 多重共线性检验 21](#_Toc171543451)

[第四章 特征工程 21](#_Toc171543452)

[第一节 特征删除 21](#_Toc171543453)

[第二节 平衡数据集 22](#_Toc171543454)

[第三节 嵌入法特征选择 23](#_Toc171543455)

[第五章 机器学习 24](#_Toc171543456)

[第一节 模型建立与评估 24](#_Toc171543457)

[第二节 五折交叉验证 32](#_Toc171543458)

[第三节 参数调优 34](#_Toc171543459)

[第六章 结果分析 35](#_Toc171543460)

# 第一章 数据准备

## 第一节 数据获取

数据来源于<https://www.kaggle.com>内的ObesityDataSet\_raw\_and\_data\_sinthetic数据集，其内涵盖了多种与肥胖相关的因素。Kaggle作为一个知名的数据科学和机器学习社区，汇聚了全球来自各个领域的数据科学家和研究人员。数据集具有高质量和可靠性，从而能够确保数据的准确性和完整性。

查看部分数据显示于下表1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | Height | Weight | family\_history\_with\_overweight | FAVC | FCVC | NCP |
| Female | 21.0 | 1.62 | 64.0 | yes | no | 2.0 | 3.0 |
| Female | 21.0 | 1.52 | 56.0 | yes | no | 3.0 | 3.0 |
| Male | 23.0 | 1.80 | 77.0 | yes | no | 2.0 | 3.0 |
| Male | 27.0 | 1.80 | 87.0 | no | no | 3.0 | 3.0 |
| Male | 22.0 | 1.78 | 89.8 | no | no | 2.0 | 1.0 |

续表1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| Sometimes | no | 2.0 | no | 0.0 | 1.0 | no | Public\_Transportation | Normal\_Weight |
| Sometimes | yes | 3.0 | yes | 3.0 | 0.0 | Sometimes | Public\_Transportation | Normal\_Weight |
| Sometimes | no | 2.0 | no | 2.0 | 1.0 | Frequently | Public\_Transportation | Normal\_Weight |
| Sometimes | no | 2.0 | no | 2.0 | 0.0 | Frequently | Walking | Overweight\_Level\_I |
| Sometimes | no | 2.0 | no | 0.0 | 0.0 | Sometimes | Public\_Transportation | Overweight\_Level\_II |

表1 部分原数据展示表

本报告旨在通过分析不同因素与肥胖之间的关系，帮助发掘导致肥胖的主要因素，并为肥胖人群提供科学的饮食建议和健康管理方法。

## 第二节 数据描述

数据集内总共包含2111条数据，其内含有性别、年龄、身高、体重等15个特征字段，1个标签字段，通过对数据进行分析以及通过pandas读取数据描述信息，所有数据无缺失值，并总结得出以下数据描述信息表1

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段名称 | 描述 | 字段类型 |
| 0 | Gender | 性别 | object |
| 1 | Age | 年龄（岁） | float64 |
| 2 | Height | 身高（厘米） | float64 |
| 3 | Weight | 体重（公斤） | float64 |
| 4 | family\_history\_with\_overweight | 家族中是否有超重史 | object |
| 5 | FAVC | 是否有食物作为主要选择的习惯(FAVC) | object |
| 6 | FCVC | 每日是否摄入蔬果（FCVC） | float64 |
| 7 | NCP | 每日进餐次数（NCP） | float64 |
| 8 | CAEC | 每周进食的碳水化合物频率(CAEC) | object |
| 9 | SMOKE | 是否吸烟（SMOKE） | object |
| 10 | CH2O | 每日饮用水量（CH2O） | float64 |
| 11 | SCC | 是否有慢性疾病（SCC） | object |
| 12 | FAF | 每周体力活动频率（FAF） | float64 |
| 13 | TUE | 每日看电视/使用电子设备时间（TUE） | float64 |
| 14 | CALC | 每周摄入钙频率（CALC） | object |
| 15 | MTRANS | 出行方式（MTRANS） | object |
| 16 | NObeyesdad | 肥胖类型分类(NObeyesdad) | object |

表1 数据描述表

对数据集内连续型变量进行进一步分析，计算其平均值、标准差、最小值、下四分位数、中位数、上四分位数、最大值等，对数据内连续性变量的分布情况和中心趋势进一步了解，其具体数据表x于下表2

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Height | Weight | FCVC | NCP | CH2O | FAF | TUE |
| count | 2111.00 | 2111.00 | 2111.00 | 2111.00 | 2111.00 | 2111.00 | 2111.00 | 2111.00 |
| mean | 24.312600 | 1.701677 | 86.586058 | 2.419043 | 2.685628 | 2.008011 | 1.010298 | 0.657866 |
| std | 6.345968 | 0.093305 | 26.191172 | 0.533927 | 0.778039 | 0.612953 | 0.850592 | 0.608927 |
| min | 14.000000 | 1.450000 | 39.00000 | 1.000000 | 1.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 19.947192 | 1.630000 | 65.473343 | 2.000000 | 2.658738 | 1.584812 | 0.124505 | 0.000000 |
| 50% | 22.777890 | 1.700499 | 83.000000 | 2.385502 | 3.000000 | 2.000000 | 1.000000 | 0.625350 |
| 75% | 26.000000 | 1.768464 | 107.430682 | 3.000000 | 3.000000 | 2.477420 | 1.666678 | 1.000000 |
| max | 61.000000 | 1.980000 | 173.0000 | 3.000000 | 4.000000 | 3.000000 | 3.000000 | 2.000000 |

表2 连续型变量数据分析表

从表内数据可以看出，数据集内年龄(Age)最小值为14，最大值为61，平均年龄为24.31，身高（Height）单位为米，最大值为1.98m，最小为1.45m。体重（Weight）平均值为86KG，最大值为173KG，最小值为34KG，每日摄入蔬果量（FCVC）平均为2.42，每日进餐次数(NCP)平均为2.69次每天，每日饮水量(CH2O)平均为2.01升，每周体力活动频率（FAF）平均为1.01，每日看电视/使用电子设备时间(TUE)平均为0.66小时。

# 第一章 数据清洗

## 第一节 异常值处理

图表, 箱线图

描述已自动生成异常值处理在数据分析和建模过程中至关重要，它涉及识别、理解和处理数据中的异常或离群值。异常值可能影响统计分析的结果和模型的准确性，因此对数据集内连续型数据绘制箱型图对其异常值进行查看，箱型图绘制于下图1

图1 连续型变量分布箱型图

根据上图可知，在数据样本中，Age和NCP两个特征可能存在异常值，对此进行进一步分析。箱型图显示Age特征上限为35，在35到82之间存在离群点，显示NCP特征上限为3.5，下限为2.25，在3.5到4.0之间存在离群点，在1.0到2.25中存在离群点。Age特征表示年龄，单位为岁，在数据样本中出现35到82岁的情况为正常现象，不计为异常值；NCP特征表示每日进餐次数，而数据样本中出现的一天1次到2.25次的进餐次数的情况以及一天3.5到4次的进餐情况，在健康和营养研究中，进餐次数的变化可能反映个体的饮食习惯和生活方式选择，因此这些离群点可能是数据样本的真实反映，同样不计为异常值。

图表, 瀑布图

描述已自动生成查看离散值特征分布情况于下柱状图2

图2 离散值柱状分布图

柱状图显示了数据集中不同特征变量的分布情况。在变量"Gender"上，男性和女性的分布相对均衡。然而，在"family\_history\_with\_overweight"（家族中是否有超重史）、"FAVC"（是否每日摄入蔬果）、"CAEC"（每周进食的碳水化合物频率）、"SMOKE"（是否吸烟）、"SCC"（是否有慢性疾病）、"CALC"（每周摄入钙频率）以及"MTRANS"（出行方式）等变量上存在明显的不均衡分布。

具体而言，"family\_history\_with\_overweight"中记录有超重史的人数显著高于无超重史的人数。在"FAVC"中，记录每日摄入蔬果量的人数明显多于不摄入蔬果的人数。"CAEC"中，记录在进食碳水化合物频率为有时的人数远远超过其他频率。"SMOKE"中，吸烟者的数量远多于非吸烟者。在"SCC"中，有慢性疾病的数据量明显高于无慢性疾病的数据量。"CALC"中，记录摄入钙频率为有时的人数明显超过其他频率。最后，在"MTRANS"中，使用公共交通方式出行的人数远多于其他出行方式的人数。

## 第二节 标签重新分类

对于标签数据（NObeyesdad），在各个肥胖类型分类上分布均匀，但由于研究的主要目标为“导致肥胖的主要因素”所以将目标标签进行重新分类，将目标标签分为肥胖（Obesity）和不肥胖（No\_Obesity）两类，标签数据重新分类后显示于下表3

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | Height | Weight | family\_history\_with\_overweight | FAVC | FCVC | NCP |
| Female | 21.0 | 1.62 | 64.0 | yes | no | 2.0 | 3.0 |
| Female | 21.0 | 1.52 | 56.0 | yes | no | 3.0 | 3.0 |
| Male | 23.0 | 1.80 | 77.0 | yes | no | 2.0 | 3.0 |
| Male | 27.0 | 1.80 | 87.0 | no | no | 3.0 | 3.0 |
| Male | 22.0 | 1.78 | 89.8 | no | no | 2.0 | 1.0 |

续表1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| Sometimes | no | 2.0 | no | 0.0 | 1.0 | no | Public\_Transportation | No\_Obesity |
| Sometimes | yes | 3.0 | yes | 3.0 | 0.0 | Sometimes | Public\_Transportation | No\_Obesity |
| Sometimes | no | 2.0 | no | 2.0 | 1.0 | Frequently | Public\_Transportation | No\_Obesity |
| Sometimes | no | 2.0 | no | 2.0 | 0.0 | Frequently | Walking | Obesity |
| Sometimes | no | 2.0 | no | 0.0 | 0.0 | Sometimes | Public\_Transportation | Obesity |

表3 标签重新分类后部分数据表

## 第三节 缺失值处理

查看数据缺失情况于下表4

|  |  |
| --- | --- |
| Gender | 0 |
| Age | 0 |
| Height | 0 |
| Weight | 0 |
| family\_history\_with\_overweight | 0 |
| FAVC | 0 |
| FCVC | 0 |
| NCP | 0 |
| CAEC | 0 |
| SMOKE | 0 |
| CH2O | 0 |
| SCC | 0 |
| FAF | 0 |
| TUE | 0 |
| CALC | 0 |
| MTRANS | 0 |
| NObeyesdad | 0 |

表4 缺失情况表

由表可知，数据集内不存在缺失值

## 第四节 重复值处理

查看数据重复情况，有24条重复数据，重复数据部分展示于下表5

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | Height | Weight | family\_history\_with\_overweight | FAVC | FCVC | NCP |
| Female | 21.0 | 1.52 | 42.0 | no | no | 3.0 | 1.0 |
| Female | 25.0 | 1.57 | 55.0 | no | yes | 2.0 | 1.0 |
| Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 |
| Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 |
| Male | 21.0 | 1.62 | 70.0 | no | yes | 2.0 | 1.0 |

续表1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| Frequently | no | 1.0 | no | 0.0 | 0.0 | Sometimes | Public\_Transportation | No\_Obesity |
| Sometimes | no | 2.0 | no | 2.0 | 0.0 | Sometimes | Public\_Transportation | No\_Obesity |
| no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public\_Transportation | Obesity |
| no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public\_Transportation | Obesity |
| no | no | 3.0 | no | 1.0 | 0.0 | Sometimes | Public\_Transportation | Obesity |

表5 部分重复数据表

由表可知，认为重复值由于巧合产生的可能性较小，认为重复值为记录异常值，对其进行删除处理

## 第五节 数据类型转换

数据集内存在object类型数据，为后续的分析工作对其进行数据类型转换。Gender、family\_history\_with\_overweight、FAVC、CAEC、CALC、MTRANS、NObeyesdad为object类型数据。其中CAEC、CALC有序，存在大小关系，对其进行手动赋值，按照大小关系依次从0进行赋值；NObeyesdad、family\_history\_with\_overweight、FAVC、SMOKE、SCC内为二元特征，对其进行手动赋值，将特征中yes字段赋值为1，no字段赋值为0；Gender、MTRANS为无序数据，调用LabelEncoder类进行赋值。

数据类型转换后部分数据显示于下表6

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | Height | Weight | family\_history\_with\_overweight | FAVC | FCVC | NCP |
| 0 | 21.0 | 1.62 | 64.0 | 1 | 0 | 2.0 | 3.0 |
| 0 | 21.0 | 1.52 | 56.0 | 1 | 0 | 3.0 | 3.0 |
| 1 | 23.0 | 1.80 | 77.0 | 1 | 0 | 2.0 | 3.0 |
| 1 | 27.0 | 1.80 | 87.0 | 0 | 0 | 3.0 | 3.0 |
| 1 | 22.0 | 1.78 | 89.8 | 0 | 0 | 2.0 | 1.0 |

续表1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| 1 | 0 | 2.0 | 0 | 0.0 | 1.0 | 0.0 | 3 | 0 |
| 1 | 1 | 3.0 | 1 | 3.0 | 0.0 | 1.0 | 3 | 0 |
| 1 | 0 | 2.0 | 0 | 2.0 | 1.0 | 2.0 | 3 | 0 |
| 1 | 0 | 2.0 | 0 | 2.0 | 0.0 | 2.0 | 4 | 1 |
| 1 | 0 | 2.0 | 0 | 0.0 | 0.0 | 1.0 | 3 | 1 |

表6 数据类型转换后部分数据表

# 第三章 探索性数据分析

## 第一节 描述性分析

在进行数据分析的过程中，对标签数据（目标变量）以及各个特征数据进行详细的可视化分析是至关重要的。通过数据分析揭示数据集的内在结构和特征之间的潜在关系，从而为后续建模和解释提供重要线索。通过可视化，我们能够直观地观察数据的分布情况。首先对数据集中的标签数据进行分析，探索其在不同特征条件下的表现和变化趋势，随后将深入研究各个特征的分布特征和可能的相关性，为全面理解数据集奠定基础。

图表, 饼图

描述已自动生成查看标签数据分布情况于下图3数据集内肥胖比例饼图

图3 肥胖比例饼图

据上图可得，数据集内肥胖占比为26.31%，非肥胖占比73.69%，数据集内肥胖与非肥胖的比例出现了不均衡情况，后续有必要进行平衡数据集操作

图表

描述已自动生成绘制年龄与是否肥胖的关系组合图于下图4

图4 年龄与是否肥胖的关系组合图

由图可得，在年龄中心点由14变化到21，肥胖率出现明显上升，在年龄中心点处于28至56时，肥胖率出现波动，在年龄中心点由56变化至63时，肥胖率出现明显下降。通过分析可得，在年龄中心点由56变化至63时出现的下降情况由数据分布导致，不做为分析依据，且由于在后续年龄增长时肥胖率在86%到100%之间波动，认为是否肥胖和年龄有较小的相关关系。

图表, 条形图

描述已自动生成绘制身高与是否肥胖的关系组合图5

图5 身高与是否肥胖的关系组合图

由图可得，随身高中心点由1.49变化至1.89，肥胖率未出现与身高上升相同或相反的变化趋势，整体变化趋势呈波动起伏情况，认为身高与是否肥胖无明显关系。

绘制体重与是否肥胖的关系组合图6

图表

描述已自动生成图6 体重与是否肥胖的关系组合图

由图可知，随体重中心点由40变化至160，肥胖率由0%上升至100%，存在明显正相关现象，但由于本文研究目的为”导致肥胖的主要因素”，而体重与上述的变量身高都是评价是否肥胖的主要依据，与肥胖的形成无关，与研究目的不相符合，后续需要考虑将这两个变量进行删除处理。

图表

描述已自动生成绘制家族中是否有超重史与是否肥胖的组合图7

图7 家族中是否有超重史与是否肥胖的组合图

由图所示，家族中有超重史的肥胖率为84%，家族中无超重史的肥胖率为25%，家族中有无超重史与是否肥胖存在相关性，家族中存在肥胖史的人更容易肥胖。

绘制是否有食物作为主要选择的习惯与是否肥胖的组合图8

图表

描述已自动生成图8 是否有食物作为主要选择的习惯与是否肥胖的组合图

由图所示，有食物作为主要选择的肥胖率为77%，无食物作为主要选择的肥胖率为47%，肥胖与有食物作为主要选择有一定的相关性，有食物作为主要选择的更有可能出现肥胖。

图表, 条形图

描述已自动生成绘制每日摄入蔬果与是否肥胖的关系组合图9

图9 每日摄入蔬果与是否肥胖的组合图

由图可得，每日摄入蔬果量中心点从1.2上升至1.6时，肥胖率出现上升，每日摄入蔬果量中心点从1.6上升至2时，肥胖率出现下降情况。随每日摄入蔬果量上升，肥胖率呈波动情况，每日摄入蔬果量与肥胖无较大相关性

图表

描述已自动生成绘制每日进餐次数与是否肥胖的关系组合图10

图10 每日进餐次数与是否肥胖的关系组合图

由图可知，在每日进餐次数中心点从1.2次变化至1.6次时，肥胖率出现较小的上升趋势，但当进餐次数中心点增加到2时，肥胖率出现明显下降趋势且后续无上升趋势。这种波动性表明进餐次数与肥胖率之间具有负相关关系。

图表

低可信度描述已自动生成绘制每周进食的碳水化合物量与是否肥胖的关系组合图11

图11 每周进食的碳水化合物量与是否肥胖的关系组合图

由图可得，由于进行了数据类型转换，0到3表示每周进食碳水化合物的频率逐渐提高，随着每周进食碳水化合物的频率提高，肥胖率出现先上升后下降在上升的情况，在每周进食碳水化合物频率为2（经常）时肥胖率出现最低值为16%，在每周进食碳水化合物频率为1（偶尔）时肥胖率出现最高值83%，肥胖与每周进食的碳水化合物频率无明显相关性。

绘制是否吸烟与是否肥胖的关系组合图12

图表

描述已自动生成图12 是否吸烟与是否肥胖的关系组合图

由图可知，不吸烟人群的肥胖率为74%，吸烟人群的肥胖率68%，肥胖率并未因是否吸烟出现明显变化，肥胖与是否吸烟无明显的线性相关性。

图表, 条形图

描述已自动生成绘制每日饮用水量与是否肥胖的关系组合图13

图13 每日饮用水量与是否肥胖的关系组合图

由图可知，肥胖率随每日饮水量上升出现波动变化，在每日饮水量中心点为1时达最低为61%，在每日饮水量中心点为2.5时肥胖率达最高为90%，无明显简单的线性相关性。

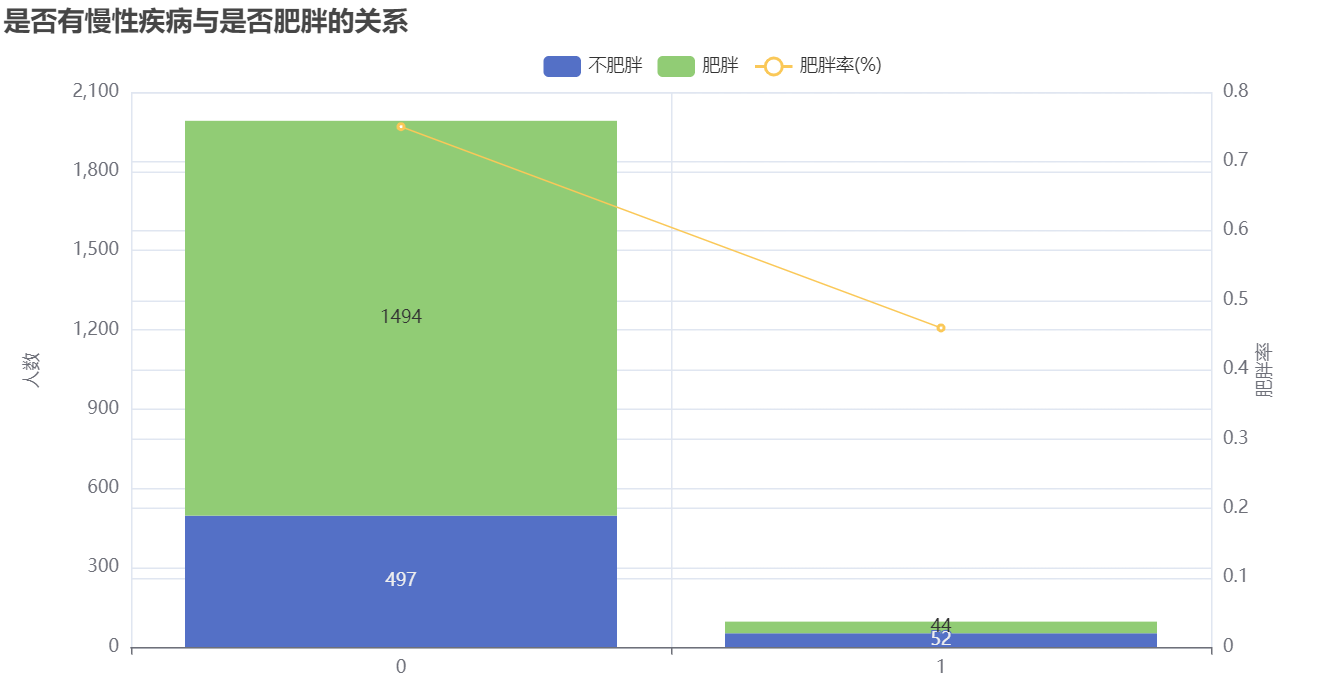
绘制是否有慢性疾病与是否肥胖的关系组合图14

图14 是否有慢性疾病与是否肥胖的关系组合图

由图可知，无慢性疾病的肥胖率为75%远高于有慢性疾病的肥胖率46%，是否有慢性疾病与是否肥胖存在一定相关性。

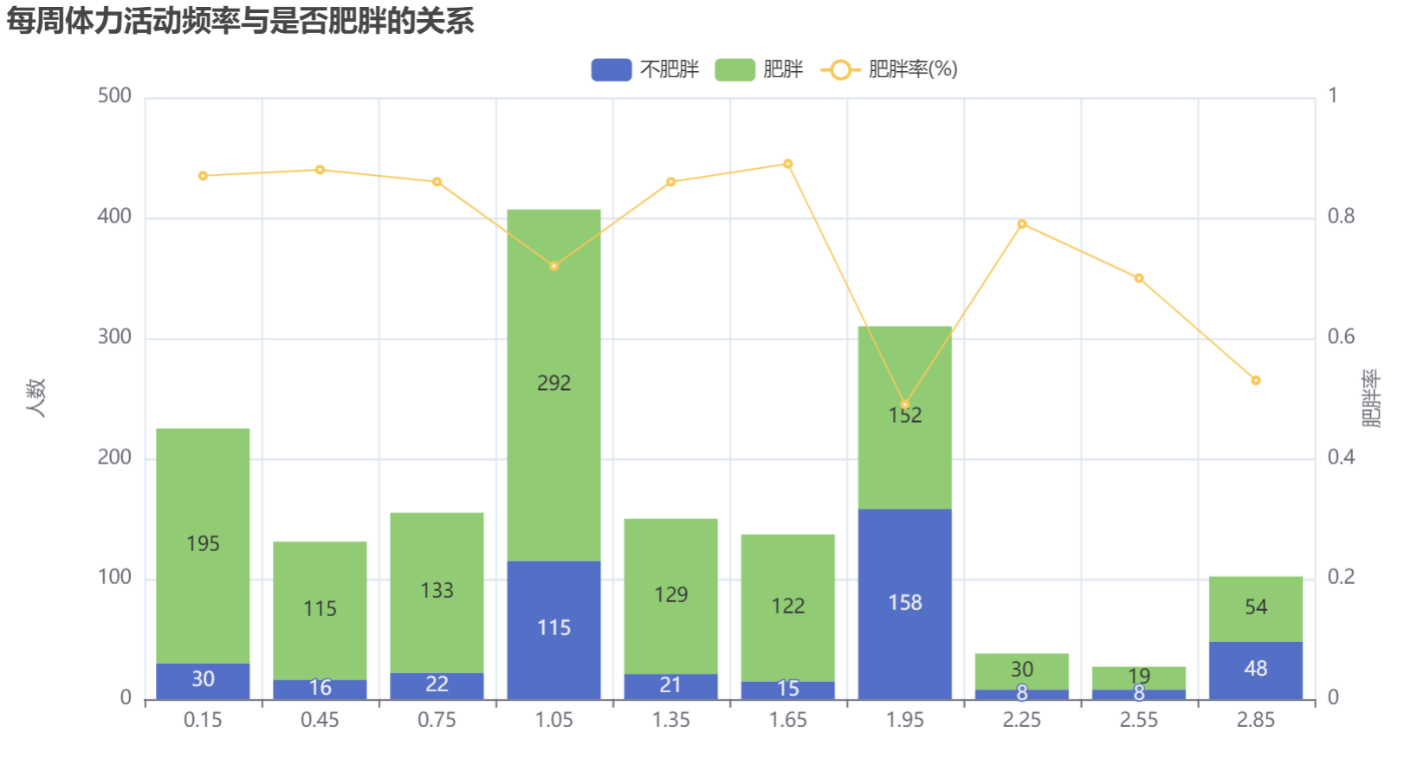
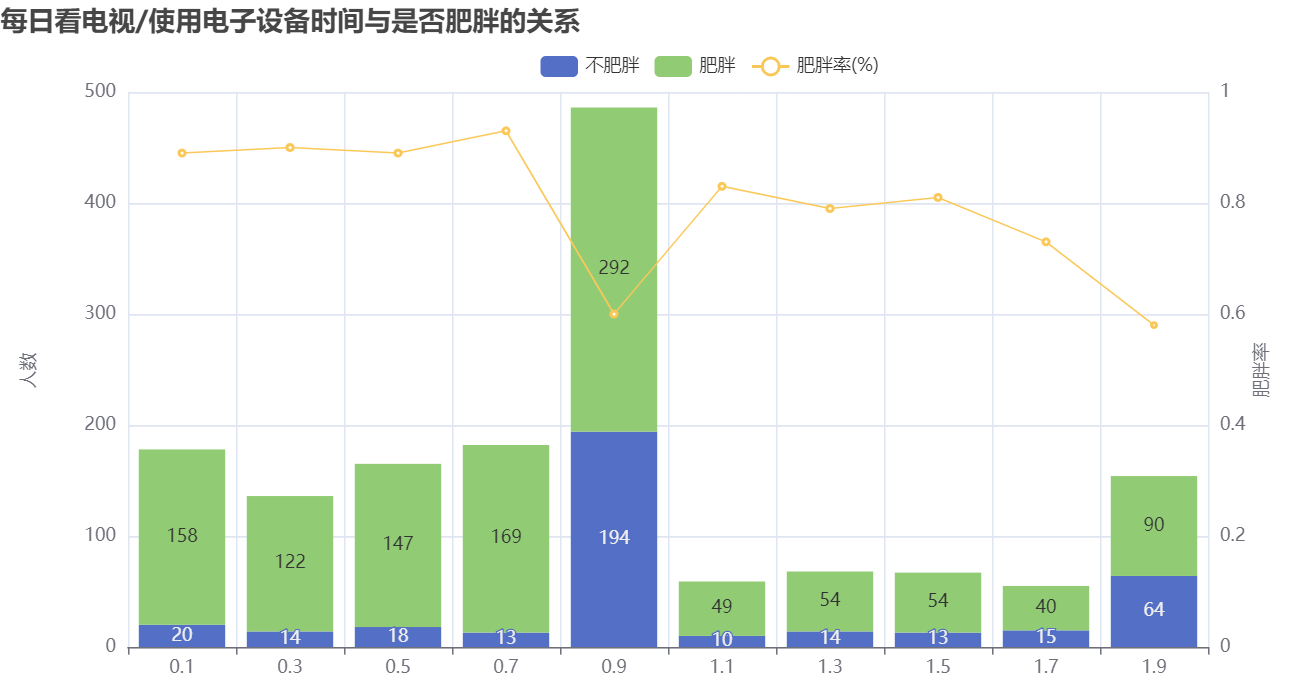
绘制每周体力活动频率与是否肥胖的关系组合图15

图15 每周体力活动频率与是否肥胖的关系组合图

由图可知，每周的活动频率由0.15增加到2.85时，肥胖率呈现波动变化，于每周活动频率为1.95时，达到最低为49%，在每周活动频率为1.65时达到最高为89%。是否肥胖与每周活动频率无明显线性相关性

绘制每日看电视或使用电子设备时间与是否肥胖的关系组合图16

图16 每日看电视或使用电子设备时间与是否肥胖的关系组合图

由图可知，每日看电视或使用电子设备时间由0.1增加到1.9时，肥胖率呈波动变化，在每日看电视或使用电子设备时间为0.9时，肥胖率达到最低为60%，每日看电视或使用电子设备时间为0.7时肥胖率达最高为93%。每日看电视或使用电子设备时间与是否肥胖无明显相关性

图表, 条形图

描述已自动生成绘制每周摄入钙频率与是否肥胖的关系组合图17

图17 每周摄入钙频率与是否肥胖的关系组合图

由图可知，每周摄入钙频率从0（从不）增加到3（总是）时，肥胖率呈现波动变化，在每周摄入钙频率为1（偶尔）时，肥胖率达最高为78%，在每周摄入钙频率为3（总是）时，肥胖率最低为0%。每周摄入钙频率与是否肥胖无明显线性相关性。

图表, 折线图

描述已自动生成绘制出行方式与是否肥胖的关系组合图18

图18 出行方式与是否肥胖的关系组合图

由图可知，肥胖率在出行方式为0（汽车）时最高为80%，在出行方式为1（自行车）时最低为43%。

## 第二节 数据归一化

由于数据未经敏感处理，每个特征的分布情况和数值可以进行描述性分析。为确保描述性分析数据的可用性，数据归一化在描述性分析之后进行.数据归一化在数据分析和机器学习中扮演着关键角色。它通过将不同特征的数值范围缩放到相似的区间，提升了算法的准确性和可靠性。

选择数值型数据进行数据归一化处理，数据归一化处理后部分数据展示于下表7

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | Height | Weight | family\_history\_with\_overweight | FAVC | FCVC | NCP |
| 0 | -0.526613 | -0.887408 | -0.872985 | 1 | 0 | -0.788364 | 0.390906 |
| 0 | -0.526613 | -1.960788 | -1.178508 | 1 | 0 | 1.082164 | 0.390906 |
| 1 | -0.212507 | 1.044677 | -0.376509 | 1 | 0 | -0.788364 | 0.390906 |
| 1 | 0.415705 | 1.044677 | 0.005395 | 0 | 0 | 1.082164 | 0.390906 |
| 1 | -0.369560 | 0.830001 | 0.112328 | 0 | 0 | -0.788364 | -2.225418 |

续表1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CAEC | SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| 1 | 0 | -0.007810 | 0 | -1.186977 | 0.554211 | 0.0 | 3 | 0 |
| 1 | 1 | 1.636552 | 1 | 2.328908 | -1.090505 | 1.0 | 3 | 0 |
| 1 | 0 | -0.007810 | 0 | 1.156947 | 0.554211 | 2.0 | 3 | 0 |
| 1 | 0 | -0.007810 | 0 | 1.156947 | -1.090505 | 2.0 | 4 | 1 |
| 1 | 0 | -0.007810 | 0 | -1.186977 | -1.090505 | 1.0 | 3 | 1 |

表7 数据归一化处理后部分数据表

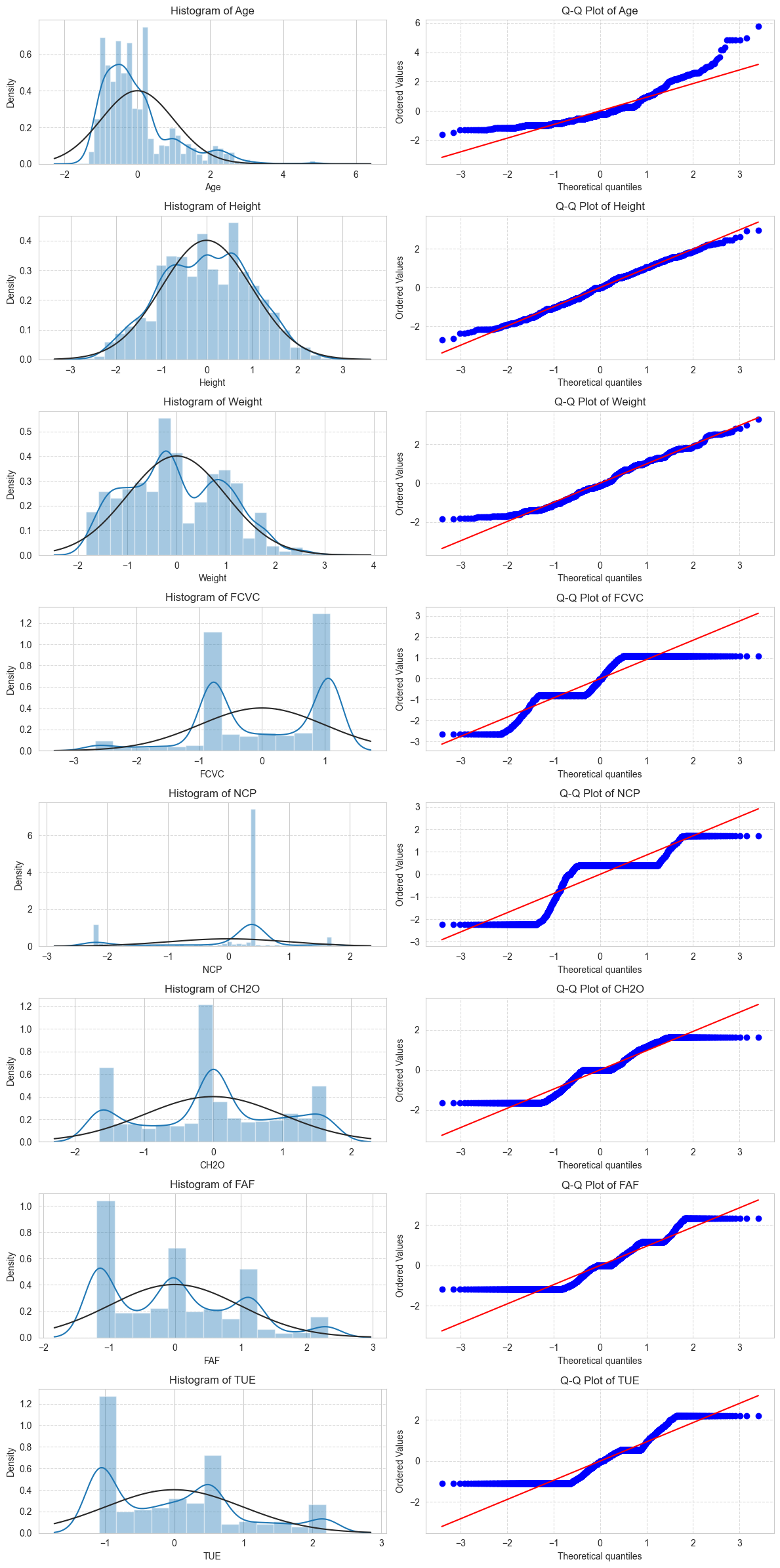
## 第三节 直方图与Q-Q图分析

绘制直方图和Q-Q图对数据分布进行分析，查看数据分布情况是否满足正态分布，为后续部分模型调优的特征处理提供依据。

图形用户界面

描述已自动生成绘制数值型特征的直方图和Q-Q图于下图19、20

图19 数值型特征的直方图与Q-Q图

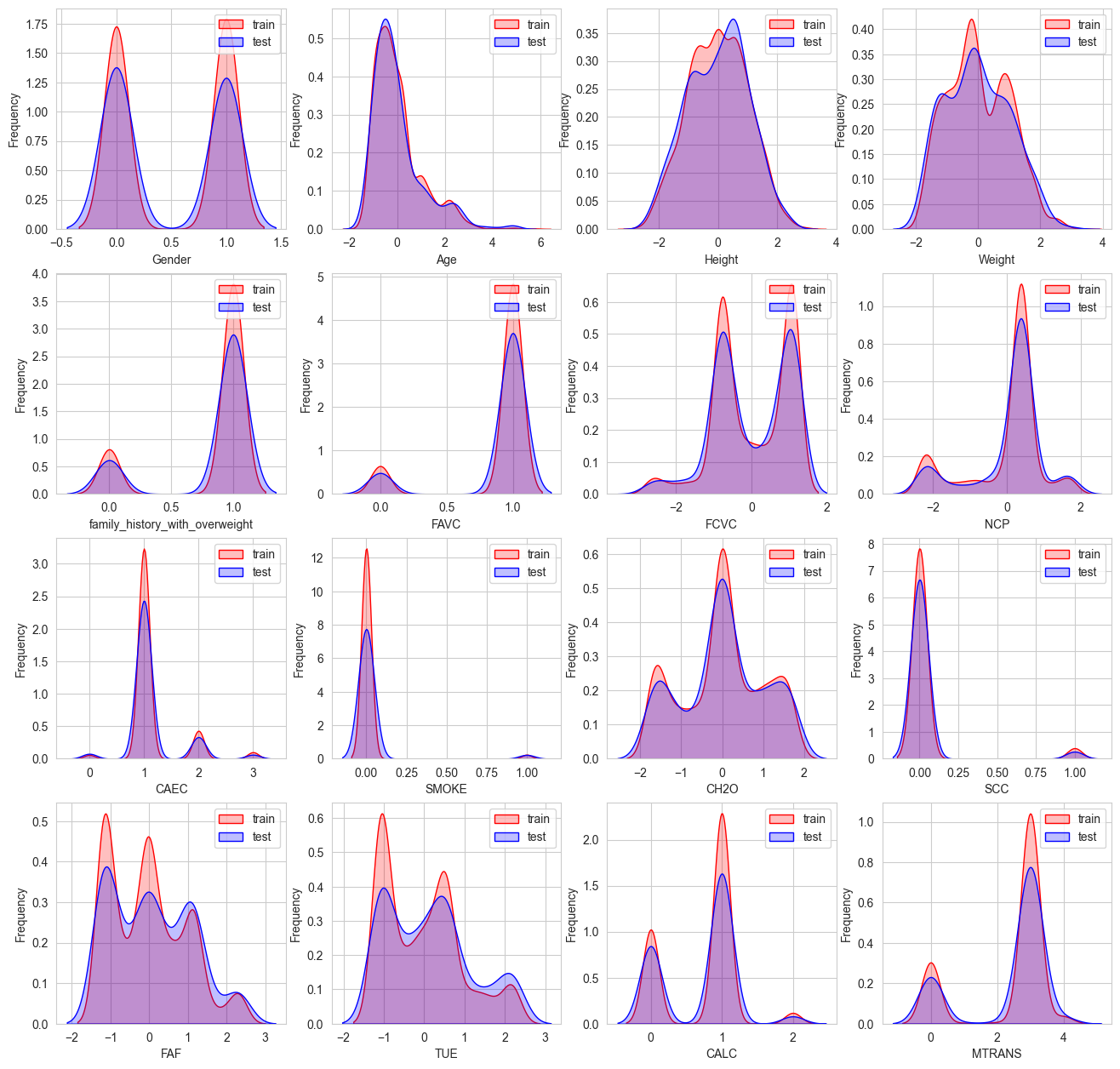
图20 数值型特征的直方图与Q-Q图

据上图可知，特征Age、Height、Weight较符合正态分布，而特征FCVC、NCP、FAF、TUE不符合正态分布，在后续可以考虑对特征进行合适的变化处理。

## 第四节 KDE分布图

通过绘制KDE分布图，查看对比训练集测试集中的特征变量的分布情况，数据在训练集测试集中分布不均可能会导致机器学习模型在实际应用中表现下降。这种情况会影响模型的泛化能力，使其在测试阶段无法有效预测新数据的特征分布，可能导致预测偏差和数据偏倚问题，最终影响模型的整体性能评估和应用效果。绘制KDE分布图对筛选分布不均特征，在后续特征工程中可以对其进行删除处理

绘制KDE分布图于下图21

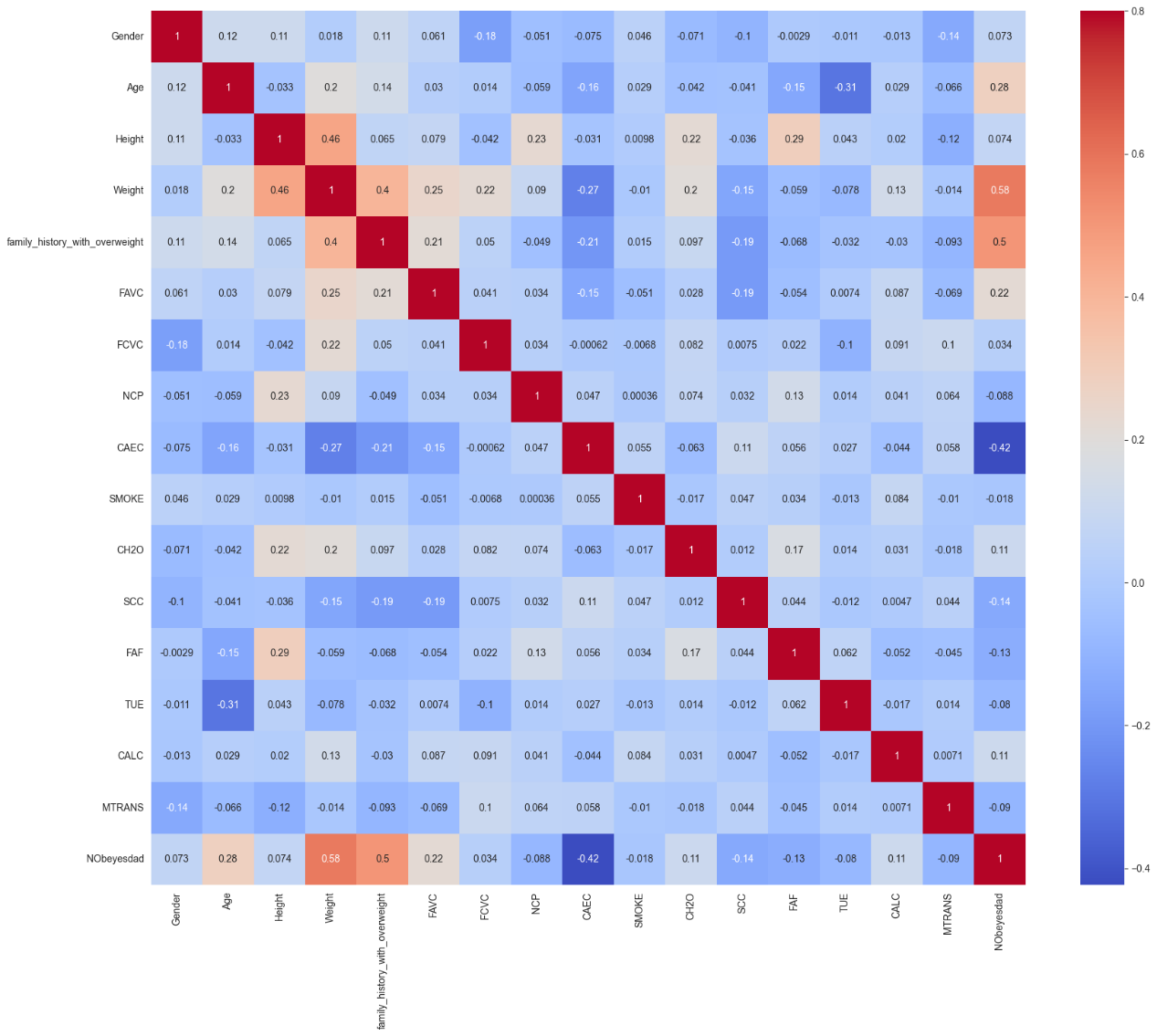
图21 KDE分布图

根据上图显示可得，各个特征在训练集和测试集上的分布较为均匀，不存在明显的分布不均问题。这种一致性表明，在数据的划分和处理过程中，特征的分布没有出现显著差异或偏离情况。这种均匀的分布有助于确保模型在训练和测试阶段能够有效地学习和泛化特征之间的关系，为模型的整体预测能力和稳定性起到一定的提高作用。

## 第五节 双变量分析

通过计算相关系数进行双变量分析的必要性在于评估两个变量之间的线性关系强度和方向。相关系数能够提供一个量化的指标，帮助确定变量之间是否存在显著的关联性。这种分析还可以为进一步的多变量分析和模型建立提供基础。通过相关系数，可以快速识别出对研究问题具有重要影响的变量，从而优化研究设计和分析策略，确保基于数据的科学决策和推断具有可靠性和有效性。

绘制各特征、标签相关系数热量图于下图22

图22 各特征、标签相关系数热力图

根据上图可得，除了各特征与标签的相关系数外，特征Weight与特征Height的相关系数较大，达到了0.46，存在中等程度的正相关关系，后续特征工程中可以考虑对其进行处理。

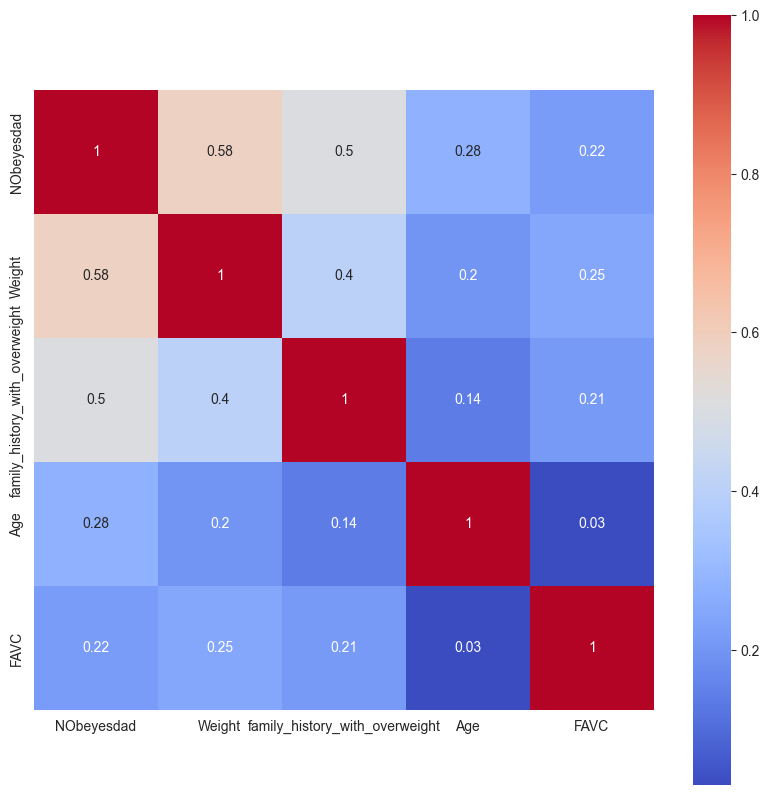
绘制与标签相关度最高的前五个特征的相关系数热力图23

图23 与标签相关度最高的前五个特征的相关系数热力图

根据上图可知，与标签相关系数最大的特征为Weight，相关系数为0.58；其次最相关的特征为family\_history\_with\_overweight，其和标签的相关系数为0.5。其余特征与标签特征的相关系数都较低，这种情况可能会导致后续在逻辑回归模型上预测的表现降低，在后续模型选择时可以选择对相关系数要求不高的分类模型如树模型等。

## 第六节 多重共线性检验

进一步进行多重共线性检验评估自变量之间的相关性程度，通过计算方差膨胀因子（VIF）来实现

各特征VIF值从小到大展示于下表

|  |  |
| --- | --- |
| SMOKE | 1.043511198697196 |
| NCP | 1.0802990666302068 |
| CH2O | 1.1061472943704815 |
| SCC | 1.1120033412563288 |
| TUE | 1.1195041615498433 |
| FCVC | 1.1498359817739066 |
| FAF | 1.2000763372165133 |
| Age | 1.209578790838902 |
| Height | 1.6760481229340467 |
| Weight | 1.9558870582209862 |
| Gender | 2.1673187614718175 |
| CALC | 2.9204822699252295 |
| MTRANS | 3.9756169185039387 |
| CAEC | 5.898354482807004 |
| family\_history\_with\_overweight | 5.940788142406638 |
| FAVC | 7.217314055764122 |

表8 各特征VIF值

据上表可得，各特征的VIF均未超过10，各特征之间未存在多重共线性。

# 第四章 特征工程

## 第一节 特征删除

基于研究的目的是探索“导致肥胖的主要因素”，身高和体重通常用于评估个体是否肥胖，而非直接导致肥胖的因素。因此，为了与研究的目标保持一致，将身高和体重这两个特征从分析中删除，有助于聚焦于那些被认为可能直接导致肥胖的潜在因素，而非仅仅是评估肥胖状态所需的身体测量。

删除后部分数据显示于下表9

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Gender | Age | family\_history\_with\_overweight | FAVC | FCVC | NCP | CAEC |
| 0 | -0.526613 | 1 | 0 | -0.788364 | 0.390906 | 1 |
| 0 | -0.526613 | 1 | 0 | 1.082164 | 0.390906 | 1 |
| 1 | -0.212507 | 1 | 0 | -0.788364 | 0.390906 | 1 |
| 1 | 0.415705 | 0 | 0 | 1.082164 | 0.390906 | 1 |
| 1 | -0.369560 | 0 | 0 | -0.788364 | -2.225418 | 1 |

续表1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SMOKE | CH2O | SCC | FAF | TUE | CALC | MTRANS | NObeyesdad |
| 0 | -0.007810 | 0 | -1.186977 | 0.554211 | 0.0 | 3 | 0 |
| 1 | 1.636552 | 1 | 2.328908 | -1.090505 | 1.0 | 3 | 0 |
| 0 | -0.007810 | 0 | 1.156947 | 0.554211 | 2.0 | 3 | 0 |
| 0 | -0.007810 | 0 | 1.156947 | -1.090505 | 2.0 | 4 | 1 |
| 0 | -0.007810 | 0 | -1.186977 | -1.090505 | 1.0 | 3 | 1 |

表9 删除特征后部分数据表

## 第二节 平衡数据集

由于标签变量NObeyesdad出现数据分布不平衡情况，不同类别的样本数量差异较大，可能会影响模型的训练效果和预测性能。需要对数据集进行类别平衡处理，以确保每个类别的样本数量相对均衡，从而提高模型的泛化能力和预测准确性。由于数据量较少，选择使用过采样对少样本数据进行增加。

在进行过采样之前，数据集中正例样本的数量为1538个，而负例样本的数量仅为549个。通过过采样处理，正例样本数量增加为1538个，使整个数据集的样本总数扩展至3076条。

绘制平衡数据集后肥胖比例饼图于下图图表, 饼图

描述已自动生成24

图24 平衡数据集后肥胖比例饼图

平衡数据集后再次进行测试集训练集划分，完成划分后再次查看KDE分布图，观察训练集测试集分布情况。

绘制平衡数据集后KDE分布图于下图25

图图示

描述已自动生成25 平衡数据集后KDE分布图

## 第三节 嵌入法特征选择

嵌入法特征选择能够帮助提升机器学习模型的性能和泛化能力。在处理复杂的数据集时，选取合适的特征可以显著减少模型的复杂度，同时提高预测的准确性。通过嵌入法特征选择，从大量的特征中筛选出最具预测力和最相关的特征，从而有效地提升模型在训练和测试阶段的效率和精度。不仅能够加速模型的训练过程，还能够降低过拟合风险，提高模型的稳健性和可靠性。

对数据集进行建模，分别建立逻辑回归模型、k近邻模型、支持向量机、随机森林、决策树，后通过SelectFromModel基于不同的模型进行特征选择。由于k近邻是一种基于实例学习的方法，依赖于距离度量来找最近邻的样本，无系数或权重来衡量特征的重要性，所以不对KNN做特征选择。

首先将特征重要性选择的阈值设置为0.5，对不同的模型将特征选择器与训练数据相拟合，筛选出重要性大于0.5的特征，在逻辑回归模型中筛选特征数为5，支持向量积筛选特征数为2，随机森林与决策树筛选特征为0。所以进一步调整阈值使其筛选特征数达到期望值，经过多次尝试，设置决策树阈值为0.08，筛选特征数为6，设置随机森林阈值为0.08，筛选特征数为3。完成训练集特征筛选后使用在训练集上拟合的特征选择器对测试集进行转换，完成特征筛选过程

# 第五章 机器学习

## 第一节 模型建立与评估

对各模型计算准确率、精准率、召回率、F1分数绘制ROC曲线、AUC全面查看模型性能，由于研究的主题为“导致肥胖的主要因素”，且对标签数据进行了数据集平衡处理，在模型性能评估时主要关注的对象为准确率与ROC曲线的趋势和AUC的大小。

由于通过嵌入法对特征进行筛选删除具有一定的不可控性，特征的删除可能会导致模型的复杂度降低从而导致模型的性能较差，所以分别对进行特征筛选和未进行特性筛选的数据集进行模型训练，保证模型性能最佳。

计算逻辑回归在未进行特征筛选时的模型准确率、精准率、召回率、F1分数详细信息于下表10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.85 | 0.76 | 0.80 | 316 |
| 1 | 0.77 | 0.85 | 0.81 | 300 |
| accuracy |  |  | 0.81 | 616 |
| macro avg | 0.81 | 0.81 | 0.81 | 616 |
| weighted avg | 0.81 | 0.81 | 0.80 | 616 |

表10 未进行特征筛选1逻辑回归分类评估详细信息表

图表, 折线图

描述已自动生成绘制逻辑回归的ROC曲线图于下图26

图26 逻辑回归未进行特征筛选ROC曲线

计算逻辑回归在进行特征筛选后的模型准确率、精准率、召回率、F1分数详细信息于下表11

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.84 | 0.73 | 0.78 | 316 |
| 1 | 0.75 | 0.86 | 0.80 | 300 |
| accuracy |  |  | 0.79 | 616 |
| macro avg | 0.80 | 0.79 | 0.79 | 616 |
| weighted avg | 0.80 | 0.79 | 0.79 | 616 |

表11 逻辑回归特征筛选后分类评估详细信息表

图表, 折线图

描述已自动生成绘制逻辑回归的ROC曲线图于下图27

图27 逻辑回归特征筛选后ROC曲线

由上述数据可知，在进行特征筛选后模型的准确率出现小幅度下降，在未进行特征筛选前模型准确率为0.81，AUC为0.86，而在进行特征筛选后模型准确率为0.79，AUC为0.84。且未进行特征筛选的逻辑回归模型准确率宏平均值为0.81，加权平均为0.80均大于进行特征筛选的准确率各平均值。

计算k近邻的模型准确率、精准率、召回率、F1分数详细信息于下表12

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0. 88 | 0. 79 | 0. 83 | 316 |
| 1 | 0. 80 | 0. 88 | 0. 84 | 300 |
| accuracy |  |  | 0. 84 | 616 |
| macro avg | 0. 84 | 0. 84 | 0. 84 | 616 |
| weighted avg | 0. 84 | 0. 84 | 0. 84 | 616 |

表12 K近邻分类评估详细信息表

图表, 折线图

描述已自动生成绘制K近邻的ROC曲线图于下图28

图28 K近邻ROC曲线

由上图所示，K近邻的ROC曲线较陡，AUC面积为0.92，且准确率为0.84，模型预测性能较好。

计算支持向量积特征选择前的模型准确率、精准率、召回率、F1分数详细信息于下表13

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0. 90 | 0. 68 | 0. 78 | 316 |
| 1 | 0. 73 | 0. 92 | 0. 81 | 300 |
| accuracy |  |  | 0. 80 | 616 |
| macro avg | 0. 81 | 0. 80 | 0. 80 | 616 |
| weighted avg | 0. 82 | 0. 80 | 0. 79 | 616 |

表13 支持向量积特征选择前分类评估详细信息表

图表, 折线图

描述已自动生成绘制支持向量积特征选择前的ROC曲线图于下图29

图29支持向量积特征选择前ROC曲线

计算支持向量积特征选择后的模型准确率、精准率、召回率、F1分数详细信息于下表14

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0. 90 | 0. 68 | 0. 78 | 316 |
| 1 | 0. 73 | 0. 92 | 0. 81 | 300 |
| accuracy |  |  | 0. 80 | 616 |
| macro avg | 0. 81 | 0. 80 | 0. 80 | 616 |
| weighted avg | 0. 82 | 0. 80 | 0. 79 | 616 |

表14 支持向量积特征选择后分类评估详细信息表

图表, 折线图

描述已自动生成绘制支持向量积特征选择后的ROC曲线图于下图30

图30支持向量积特征选择后ROC曲线

由上述数据可知，在进行特征筛选后模型的AUC值出现小幅度下降，在未进行特征筛选前模型准确率为0.80，AUC为0.85，而在进行特征筛选后模型准确率仍为0.80，AUC下降至0.80。

计算随机森林特征选择前的模型准确率、精准率、召回率、F1分数详细信息于下表14

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0. 91 | 0. 98 | 0. 94 | 316 |
| 1 | 0. 97 | 0. 90 | 0. 93 | 300 |
| accuracy |  |  | 0. 94 | 616 |
| macro avg | 0. 94 | 0. 94 | 0. 94 | 616 |
| weighted avg | 0. 94 | 0. 94 | 0. 94 | 616 |

表14 随机森林特征选择前分类评估详细信息表

图表, 散点图

描述已自动生成绘制随机森林特征选择前的ROC曲线图于下图31

图31随机森林特征选择前ROC曲线

计算随机森林特征选择后的模型准确率、精准率、召回率、F1分数详细信息于下表15

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| 0 | 0. 91 | 0. 97 | 0. 94 | 316 |
| 1 | 0. 97 | 0. 89 | 0. 93 | 300 |
| accuracy |  |  | 0. 93 | 616 |
| macro avg | 0. 94 | 0. 93 | 0. 93 | 616 |
| weighted avg | 0. 94 | 0. 93 | 0. 93 | 616 |

表15 随机森林特征选择后分类评估详细信息表

图表, 散点图

描述已自动生成绘制随机森林特征选择后的ROC曲线图于下图32

图32随机森林特征选择后ROC曲线

由上述数据可知，在进行特征筛选后模型的准确率出现小幅度下降，在未进行特征筛选前模型准确率为0.94，AUC为0.99，而在进行特征筛选后模型准确率为0.93，AUC为0.97。且未进行特征筛选的逻辑回归模型准确率宏平均值为0.94，加权平均为0.94均大于进行特征筛选的准确率各平均值。

计算决策树特征选择前的模型准确率、精准率、召回率、F1分数详细信息于下表16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| 0 | 0. 81 | 0. 85 | 0. 83 | 316 |
| 1 | 0. 83 | 0. 80 | 0. 81 | 300 |
| accuracy |  |  | 0. 82 | 616 |
| macro avg | 0. 82 | 0. 82 | 0. 82 | 616 |
| weighted avg | 0. 82 | 0. 82 | 0. 82 | 616 |

表16 决策树特征选择前分类评估详细信息表

图表, 折线图

描述已自动生成绘制决策树特征选择前的ROC曲线图于下图33

图33决策树特征选择前ROC曲线

计算决策树特征选择后的模型准确率、精准率、召回率、F1分数详细信息于下表17

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| 0 | 0. 90 | 0. 68 | 0. 78 | 316 |
| 1 | 0. 73 | 0. 92 | 0. 81 | 300 |
| accuracy |  |  | 0. 80 | 616 |
| macro avg | 0. 81 | 0. 80 | 0. 80 | 616 |
| weighted avg | 0. 82 | 0. 80 | 0. 79 | 616 |

表17 决策树特征选择后分类评估详细信息表

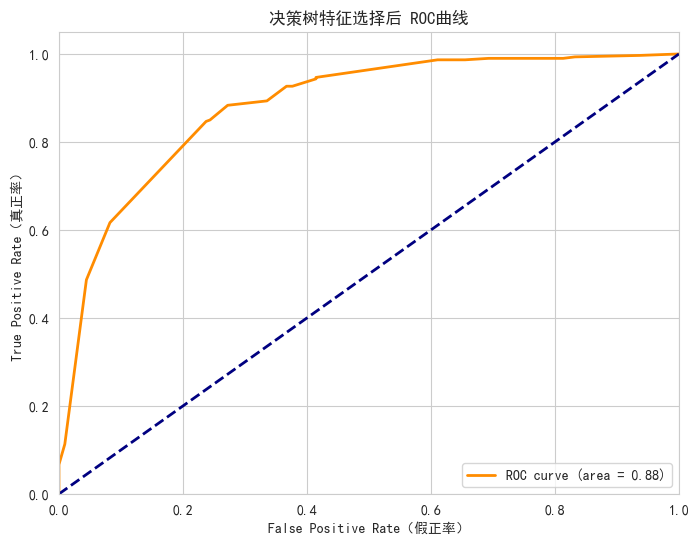
绘制决策树特征选择后的ROC曲线图于下图34

图34决策树特征选择后ROC曲线

由上述数据可知，在进行特征筛选后模型的准确率出现小幅度下降，在未进行特征筛选前模型准确率为0.82，AUC为0.89，而在进行特征筛选后模型准确率为0.80，AUC为0.88。且未进行特征筛选的逻辑回归模型准确率宏平均值为0.82，加权平均为0.82均大于进行特征筛选的准确率各平均值。

综上所述，进行了特征筛选后的模型准确率普遍比未进行特征筛选的模型低，在逻辑回归和支持向量机上表现较差，主要导致模型表现不佳原因可能是因为各特征分布不符合正态分布导致。在随机森林上表现最好，模型准确率达到0.94，AUC面积达到0.99。

## 第二节 五折交叉验证

准确率（accuracy）、精确率（precision）、召回率（recall）和F1分数（F1-score）是评估分类模型性能的常见指标。其通过不同的角度反映了模型在预测任务中的表现质量。然而，单次训练得到的评估结果可能受到数据分布的偶然性影响，特别是在数据量较少或者数据本身具有一定随机性时为了准确评估模型的泛化能力和稳定性。

所以需要进一步进行验证，通过交叉验证，能够综合考虑不同数据子集的特征分布和模型训练的随机性，从而更加客观地评估模型在各种数据情境下的表现。其有助于减少由于数据分布的偶然性引起的评估偏差，提高评估结果的可靠性和稳定性。同时能够检测模型是否存在过拟合或欠拟合问题。

绘制各模型进行五折交叉验证分数折线图于下图35

图图表, 折线图

描述已自动生成35 各模型交叉验证分数折线图

图表, 折线图

描述已自动生成据上图可得，在所有模型中随机森林仍表现最好，训练集平均准确率为1.0，测试集准确率平均值为0.95。未确保随机森林不存在过拟合现象，绘制随机森林学习曲线于下图36

图36 随机森林学习曲线

由图可得，随数据集逐渐增大，测试集的交叉验证分数平均值呈逐渐增加的趋势，这种情况说明随机森林未出现过拟合现象。综上所述，选择随机森林作为分类模型，选择未进行特征选择数据作为最终使用数据集。

## 第三节 参数调优

根据随机森林学习曲线来看，随机森林未出现过拟合问题，在训练样本增加到2000时，继续增加训练样本至2500，模型交叉验证分数并未出现明显提升，考虑是否存在欠拟合问题，对模型进行部分参数的调优，尝试进一步提高模型准确率。

为节省时间与计算资源，选择采用贝叶斯进行调优，通过利用先前的试验结果建立一个高斯过程模型，能够在每一步选择最有希望的参数来进行下一次尝试，从而找到最优解。

首先进行最优参数选择，制定参数选择空间展示于下表18

|  |  |  |
| --- | --- | --- |
| 参数 | 范围 | 步长 |
| n\_estimators | (100,500] | 10 |
| max\_depth | (10,20] | 1 |
| max\_features | (10,20] | 1 |
| min\_impurity\_decrease | (0,3] | 1 |

表18 随机森林调优参数空间

分别迭代50次、100次进行最优参数选择，分别选择参数于下表19

|  |  |  |
| --- | --- | --- |
| 参数 | 迭代50次 | 迭代100次 |
| n\_estimators | 210 | 140 |
| max\_depth | 15 | 15 |
| max\_features | 15 | 12 |
| min\_impurity\_decrease | 0 | 0 |

表19 最优参数选择结果

再次进行交叉验证查看优化效果于下图37

图图表, 折线图

描述已自动生成37 第一次调优交叉验证折线图

由图可得，两次调优的训练集准确率和调优前训练集准确率无较大提升，而训练集准确率在前三折较调优前有所提升，而在4、5时出现下降，计算交叉验证分数平均值，调优前与调优后在测试集上的分数无明显变化。

# 第六章 结果分析

最终选定构建弱分类器数为100，其他参数使用默认值的随机森林做为分类模型，训练模型后查看模型特征对于标签数据重要性，展示于下表20

|  |  |
| --- | --- |
| Feature | Importance |
| Age | 0.170938 |
| CAEC | 0.150723 |
| family\_history\_with\_overweight | 0.145747 |
| TUE | 0.095513 |
| NCP | 0.085893 |
| FAF | 0.085716 |
| CH2O | 0.077219 |
| FCVC | 0.051830 |
| MTRANS | 0.039330 |
| CALC | 0.033272 |
| Gender | 0.024502 |
| FAVC | 0.019996 |
| SCC | 0.013841 |
| SMOKE | 0.00547 |

表20 各特征对标签的重要性

根据上表可得，通过对肥胖标签重要性特征分析得出，导致肥胖最相关的几个特征为Age(年龄)、CAEC(每日碳水化合物频率)、family\_history\_with\_overweight(家族中是否有超重史)、TUE(每日看电子设备时间)、NCP(每日进餐次数)。

随年龄增长新陈代谢下降，肌肉组织下降，使得体脂的积累更为容易，以此导致肥胖风险增加；而每日碳水化物摄入的频率过高同样会导致高糖分、高脂肪的摄入，这些饮食习惯同样是肥胖的主要诱因之一。这些高量的碳水化合物不仅会提供过量能量，还可能导致血糖波动，增加人体脂肪存储；家庭遗传同样是导致肥胖的主要因素之一，遗传因素可能会导致人体新陈代谢速率、人体细胞对胰岛素的敏感性等生理特征，这些特征同样会导致肥胖的产生；同样久坐也会导致肥胖的产生，当每日看电子设备时间增加时同样会导致久坐时间的增加，降低人体能力消耗，导致人体的能量储备增加从而导致肥胖；饮食结构和进餐次数对于控制体重同样重要，频繁的小餐有助于维持血糖稳定和新陈代谢，从而减少过量进食的风险，降低肥胖的风险。

综上所述，年龄较大人群更应该重视肥胖问题，减少每日高碳水化合物食物的频率和摄入量，特别是高糖分和高脂肪的食物。选择复杂的碳水化合物，如全谷物类食物，有助于稳定血糖水平，减少脂肪储存的风险。同时，采用健康的饮食结构，例如适当分配主食、蛋白质和蔬菜的比例，并控制进餐次数。频繁而适量的小餐有助于维持血糖稳定和新陈代谢，减少过量进食的可能性也可以预防肥胖的发生。在进行电子设备使用时，避免久坐的产生，定期进行身体活动和运动，有利于维持健康的代谢状态和肌肉组织。在家族有超重史时，应该意识到遗传因素对肥胖的产生有一定的影响，家族中有超重史的人群更应该关注饮食和运动的健康习惯，以减少遗传风险带来的影响。