**研究目的：肥胖风险水平的分类预测模型（我的个人想法）**

**数据属性说明：**

* 性别（Gender）：如果受访者是男性，则为1；如果是女性，则为0。（1：男；0：女）
* 身高（Height）
* 体重（Weight）
* 年龄（Age）：受访者的年龄，以年为单位。
* 家族肥胖史（family\_history\_with\_overweight）：如果受访者有家庭成员现在或过去超重，则为1；如果没有，则为0。（0：无家族肥胖史；1：有）
* 经常食用高热量食物（FAVC）：如果受访者经常食用高热量食物，则为1；如果不是，则为0。（0：不经常食用高热量食物；1：经常食用）
* 通常食用蔬菜（FCVC）：如果受访者在他们的饮食中通常食用蔬菜，则为1；如果不是，则为0。（0：不经常食用蔬菜；1：经常食用）
* 每日主餐次数（NCP）：0表示1-2餐，1表示3餐，2表示超过3餐。
* 餐间食物摄入（CAEC）：受访者在餐间摄入的食物，按0到3的等级来衡量。
* 吸烟（SMOKE）：如果受访者吸烟，则为1；如果不吸烟，则为0。
* 饮水量（CH2O）：受访者每天饮水的量，按1到3的等级来衡量。
* 监测卡路里摄入（SCC）：如果受访者监测他们的卡路里摄入，则为1；如果不监测，则为0。
* 体力活动量（FAF）：受访者进行的体力活动量，按0到3的等级来衡量。
* 屏幕时间（TUE）：受访者每天花在看屏幕设备上的时间，按0到2的等级来衡量。
* 饮酒频率（CALC）：受访者饮酒的频率，按0到3的等级来衡量。
* 主要交通方式（MTRANS）：表示受访者的主要交通方式。
* NObeeyesdad：目标变量，肥胖等级分类：
  + 体重不足（Insufficient\_Weight）
  + 正常体重（Normal\_Weight）
  + 肥胖I型（Obesity\_Type\_I）
  + 肥胖II型（Obesity\_Type\_II）
  + 肥胖III型（Obesity\_Type\_III）
  + 超重I级（Overweight\_Level）
  + 超重II级（Overweight\_LevelI）

**研究流程：**

1. **数据预处理**

**1.1数据清洗：删除重复项数据、检查数据是否有缺失（分类变量可以采用众数填充）**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Column** | **Non-Null** | **Count** | **Dtype** |
| **0** | **Age** | **2111** | **non-null** | **float64** |
| **1** | **Gender** | **2111** | **non-null** | **Object** |
| **2** | **Height** | **2111** | **non-null** | **float64** |
| **3** | **Weight** | **2111** | **non-null** | **float64** |
| **4** | **CALC** | **2111** | **non-null** | **Object** |
| **5** | **FAVC** | **2111** | **non-null** | **Object** |
| **6** | **FCVC** | **2111** | **non-null** | **float64** |
| **7** | **NCP** | **2111** | **non-null** | **float64** |
| **8** | **SCC** | **2111** | **non-null** | **Object** |
| **9** | **SMOKE** | **2111** | **non-null** | **Object** |
| **10** | **CH2O** | **2111** | **non-null** | **float64** |
| **11** | **family\_history\_with\_overweight** | **2111** | **non-null** | **Object** |
| **12** | **FAF** | **2111** | **non-null** | **float64** |
| **13** | **TUE** | **2111** | **non-null** | **float64** |
| **14** | **CAEC** | **2111** | **non-null** | **Object** |
| **15** | **MTRANS** | **2111** | **non-null** | **Object** |
| **16** | **NObeyesdad** | **2111** | **non-null** | **Object** |

本研究数据总共有17个变量其中

数值变量有: ['Age', 'Height', 'Weight', 'FCVC', 'NCP', 'CH2O', 'FAF', 'TUE']

分类变量有：['Gender','family\_history\_with\_overweight', 'FAVC','CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS','NObeyesdad']

数据共有2111例，16个特征和最后的响应变量均不存在空值。

进一步探索发现，该数据存在24行重复数据，我们对重复数据项进行删除。

**利用四分位数法对数据中的异常值进行剔除。算法具体流程如下：**

四分位数法（Interquartile Range, IQR）是统计学中识别异常值的一种方法，主要适用于箱型图的绘制以及数据的初步异常检测。其算法流程如下：

1. 数据排序：首先将数据集按照大小顺序进行排序。
2. 计算四分位数：
3. 计算四分位距（IQR）：四分位距是上四分位数与下四分位数之差，即 IQR = Q3 - Q1。
4. 确定异常值范围：

下限：下限通常为 Q1 - 1.5 \* IQR（有时也会用3倍IQR作为更为严格的异常值界定）。

上限：上限为 Q3 + 1.5 \* IQR。

1. 识别异常值：任何小于下限或大于上限的数据点都会被视为异常值。
2. 处理异常值：由于数据量较大，异常值数目较少，且缺少相关专业知识，所以此处选择直接删除异常值。

异常值处理完成后，数据从2111例样本减少为1376例样本。

数据处理完成后，因变量分布情况如下：

图表, 饼图

描述已自动生成

根据分布饼图可知，因变量中Obesity\_Type\_III的占比最大（23.55%），Insufficient\_Weight占比最小（9.96%）。

**1.2特征工程：创建新的特征，BMI、每日用餐次数、总体育活动得分、年龄分级和每公斤体重的水分摄入量**

有几个问题：

创建新特征之后是否要去除原来的旧特征？

如果经过了运算（由两个或两个以上的变量运算出来的）的话就不用删除原数据吧，因为可以更好的捕捉模型的模式和性能。但是如果说是仅由一个变量转换而来的（比如说年龄等级）话就建模的时候就还是不计入建模吧

还有用餐次数，总体育活动得分，每公斤体重的水分摄入量怎么怎么计算？

用餐次数（Meals\_Per\_Day）等于FCVC+NCP

总体育得分（Total\_Activity\_Score）等于FAF+TUE

每公斤体重的摄水量（Water\_Intake\_Per\_Kg）等于CH2O / Weight

年龄按什么标准分为多少个等级？

可以根据年龄（0,18,60），分成三类（Young，Adult，Elderly）

另外，我在想需不需要把所有特征都化成分类变量：年龄分级，体重和身高换算为BMI指数并分等级，FCVC（是否经常食用蔬菜），'NCP'(每天吃主食的次数0-3), 'CH2O'（饮水量评分1-3）, 'FAF'（身体活动评分0-3）, 'TUE'（观看电子屏幕次数评分0-2）这些评分数据均可以看作是分类变量。最后得到的结果里面全为分类变量，可以考虑做logistic回归。

感觉不用全部化成分类变量吧，这样会使原数据丢失一些信息。

1. **探索性数据分析（EDA）（分析自变量和因变量之间的关系）**

**分类数据对应肥胖类型分布情况：**

手机屏幕的截图

低可信度描述已自动生成

根据定性的自变量与因变量之间的频数直方图可知：

主要交通方式为公共交通（Public\_Transportation）的样本数目最多，其次是自动单车（Automobile），主要交通工具样本数目最少的是自行车（Bike）；乘坐公共交通工具的数据样本里面肥胖III级（Obesity\_Type\_III）的占比最高，超重II级（Overweight\_LevelI）的样本占比最少，但是步行（Walking）、摩托车（Motorbike）、自行车（Bike）都是正常体重（Normal\_Weight）的占比最大，肥胖样本占比相对较小。因此可以初步判断主要交通方式与体重之间存在一定的关系。

体力活动量为0级的占比最多，随着级数的增加样本数目呈现递减的趋势；随着体力活动级数的增加，正常体重（Normal\_Weight）的占比逐渐增大，肥胖样本的占比逐渐减少。因此可以初步判断，体力活动量与体重之间存在一定的关系。

餐间食物摄入为有时（Sometimes）的样本数目最多，为不摄入（No）的样本数目最少；餐间食物摄入为有时（sometimes）中肥胖I级（Obesity\_Type\_I）的占比最大，偏瘦体重（Insufficient\_Weight）的占比最小；餐间不摄入（No）食物的样本中，体重超重I级（Overweight\_Level）的占比最大。因此暂不能看出餐间食物摄入与体重存在有明显的关系。

饮酒频率（CALC）为有时（Sometimes）的样本数目最多，其中占比最大的是肥胖III级（Obesity\_Type\_III），占比最小的是超重II级（Overweight\_LevelI）；饮酒频率为总是（Always）的样本数目最少，其中正常体重（Normal\_Weight）的占比最大；平时不饮酒（No）中肥胖I级（Obesity\_Type\_I）的占比最大，超重I级（Overweight\_Level）的占比最小；因此暂不能看出饮酒频率是否与体重之间存在明显的关系。

经常食用高热量食物（FAVC）为1的样本数目最多，其中肥胖I级（Obesity\_Type\_I）的占比最大，正常体重（Normal\_Weight）的占比最小；不经常食用高热量食物（0）的样本数目相对较少，其中占比最大的是正常体重（Normal\_Weight），占比最小的是肥胖III级（Obesity\_Type\_III）；因此可以初步认为是否经常食用高热量食物与体重之间存在一定的关系。

吸烟（SMOKE）为No的样本数目最多，其中肥胖I级（Obesity\_Type\_I）的占比最大，偏瘦体重（Insufficient\_Weight）的占比最小；为Yes的样本数目相对较少，其中占比最大的是肥胖II级（Obesity\_Type\_II），占比最小的是偏瘦体重（Insufficient\_Weight），因此暂不能认为是否吸烟与体重之间存在明显的关系。

**连续型数据对应肥胖类型分布情况：**

图表, 直方图

描述已自动生成

根据连续的自变量与因变量之间的频数堆积条图和密度曲线可知：

身高（Height）的体重堆积条图显示，所有体重分级中身高总体都服从近似正态分布。肥胖等级与身高之间暂不能认为有直接的线性关系。

蔬菜食用（FCVC）的体重堆积条图显示，所有体重分级中蔬菜使用量总体都服从近似正态分布。肥胖等级与蔬菜食用量之间暂不能认为有直接的线性关系。

饮水量（CH2O）的体重堆积条图显示，所有体重分级中饮水量总体都服从近似正态分布。肥胖等级与饮水量之间暂不能认为有直接的线性关系。

1. **单因素分析（根据2、3步到时候建立模型时可不可以手动筛掉一些变量呢？）**

可以考虑筛选一部分变量，但是我觉得在这里不用筛掉，因为数据的变量的数量本身就不是很多，而且单因素没影响不一定放到模型中多个因素一起作用也没影响，可以考虑纳入多因素和单因素分析均有影响的变量。

**3.1假设检验（在分析定性自变量时可以分析到具体每一个参数与因变量的关系）**

**定量数据（身高、体重、BMI、每日用餐数、总体育活动得分、每公斤体重的水分摄入量）与因变量Y体重分级（NObeyesdad）：先进行正态性检验；再使用方差分析/多独立样本资料秩和检验**

由于定量数据几乎都不服从正态检验，所以采用了多独立样本的K-W检验。检验结果在对应的文档中。

**定性数据与因变量Y体重分级（NObeyesdad）：卡方检验**

**分类数据（年龄分级）与因变量Y体重分级（NObeyesdad）：两独立样本秩和检验**

**3.2相关性检验（分析自变量与因变量之间的整体的相关性）**

**spearman相关性分析**

相关分析的内容见相关分析文档

1. **特征分布分析**

**4.1创建箱式图来识别数据中的异常值**

这个在数据预处理阶段就可以做了，然后根据四分位数法，删除异常值

**4.2数值特征分布（用来确定数据是否需要标准化和归一化）**

**用柱状图+曲线图，绘制每一个变量的数值分布（分男女）**

**FAVC在男女性别中的分布（0：女，1：男）**

**图表, 箱线图

描述已自动生成**

**FCVC在男女性别中的分布：**

**图表, 箱线图

描述已自动生成**

**SMOKE在男女性中的分布：**

**图表, 箱线图

描述已自动生成**

**SCC在男女性别中的分布（0：女，1：男）**

**图表, 箱线图

描述已自动生成**

**图表, 箱线图

描述已自动生成**

family\_history\_with\_overweight在男女性别中的分布

**图表, 箱线图

描述已自动生成**

**男女性别中年龄段的分布：**

**图表, 箱线图

描述已自动生成**

**肥胖类型在性别中的分布：**

**图表, 条形图

描述已自动生成**

1. **数据标准化**

**使用StandardScaler、MinMaxScaler和QuantileTransformer对数据进行标准化处理**

由于使用的是集成学习模型，对数据的分布并不敏感，所以不需要进行标准化。

1. **模型构建**

**分类变量进行编码**

**模型一：采用Logistic逻辑回归，建立逻辑回归方程，算出模型拟合优度，算出OR值及其95%置信区间**

在逻辑回归文档中给出了回归系数，拟合优度，OR值等信息。

**模型二：使用LightGBM分类器构建模型，并使用Optuna进行参数调优。**

**模型三：使用CatBoost分类器构建另一个模型，并同样进行了参数调优。**

1. **模型评估**

**划分数据集（注意划分比例）**

对数据集进行了数据洗牌，并按照7：3的比例划分训练集和测试集，模型训练过程中进行五折交叉验证。

**算出准确率、召回率、精确率、F1分数以及接收者操作特征曲线下面积、混淆矩阵分析和绘制ROC曲线（表+图）**

对于最终结果，给出了准确度、召回率、精确度、F1分数及AUC值，对于集成学习模型还给出了对应模型的特征重要性。同时绘制了3种模型的ROC曲线。

1. **模型优化**

**调整模型参数，使用网格搜索、随机搜索、贝叶斯优化等方法寻找最优参数**

这个感觉可以直接合并到步骤6，在训练模型的过程中就已经对参数进行了调整。

1. **结果分析**