广州中医药大学医学信息工程学院

《医学统计学》
医学调查与数据分析

基于机器学习的肥胖水平估计研究

成员：刘思洋，杨猛，梁永健

目录

1. 研究目的

本研究旨在构建一个基于人口统计学特征、生活方式和饮食习惯等因素的肥胖风险水平分类预测模型。该模型能够准确预测个体未来患肥胖症的风险等级，并为个人提供个性化的健康建议，从而预防和控制肥胖症的发生和发展。

2. 研究意义

肥胖问题正日益受到医学、公共卫生、生物信息学和统计学等多个学科领域的关注。我们必须高度重视这个问题，努力降低肥胖率，并学习如何预防和合理控制体重。这就要求我们能够准确评估肥胖水平，以便制定针对性的科学预防和治疗方案。传统的肥胖评估方法，如身体质量指数（BMI），虽然简便易行，但存在明显缺陷，因为它无法区分不同组织的质量，也无法充分考虑个体差异和生活习惯等因素的影响，可能导致评估结果的误差。

此外，越来越多的研究表明，肥胖与生活方式、饮食习惯、基因、环境、社会和经济等多重因素密切相关。为了更深入地理解肥胖的成因和预测肥胖风险，研究人员需要从多角度对肥胖进行分析，并采用新技术方法以提高肥胖评估的精确度。当前，机器学习技术因其处理复杂数据集和提取有用信息的能力而备受瞩目，将其应用于肥胖研究具有显著优势。

通过机器学习算法，我们可以整合年龄、性别、生活方式、饮食习惯等多种因素来预测肥胖风险，并能在处理大规模、高维数据时提高预测的精确度和准确性。基于机器学习的肥胖水平评估方法不仅能够更准确地评估肥胖程度，还能综合考虑更多影响因素。这种方法通过大数据和机器学习算法，对个体的生物、社会和环境因素进行综合分析建模，从而预测个体的肥胖水平和风险，并制定相应的干预和治疗策略。这不仅有助于减少肥胖相关疾病的发生和发展，提升公众健康水平，还将在肥胖干预和防控领域带来显著的社会和经济效益。

因此，基于机器学习的肥胖水平评估研究在肥胖研究和预防领域具有重要的实际意义和应用价值。

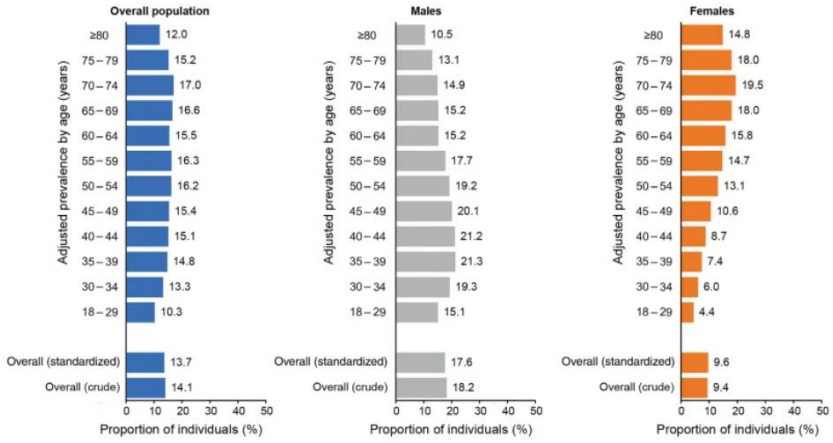

图 2.1 我国人群肥胖流行率

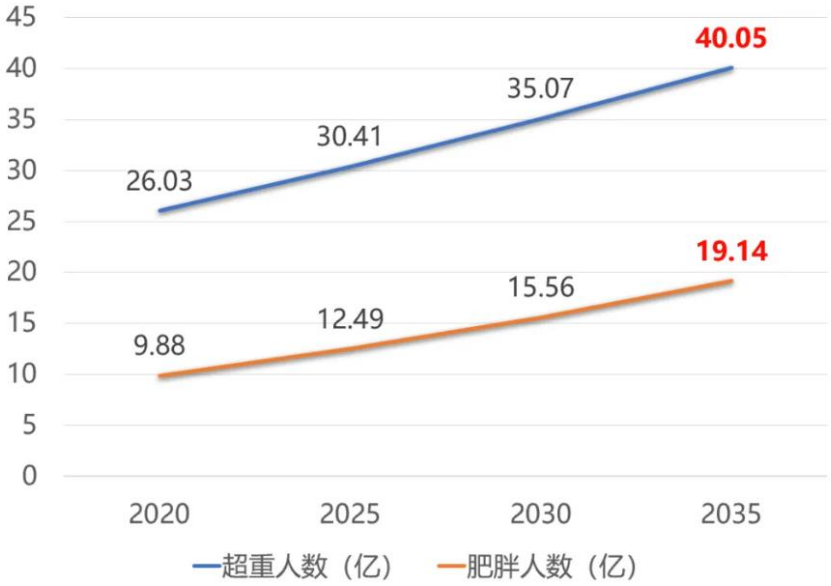

图 2.2 世界肥胖联盟估算从 2020 年到 2035 年全球的肥胖发展情况

3. 研究现状

当前，肥胖研究主要集中在几个关键领域。大部分研究关注的是肥胖的影响因素，其中饮食习惯和生活方式被认为是最重要的因素。此外，基因、环境、社会和经济因素也被纳入研究范畴。其次，机器学习方法在肥胖水平估计方面得到了广泛应用，包括朴素贝叶斯、支持向量机、随机森林和神经网络等算法。还有

一些研究探索了利用传感器数据和互联网技术来估计肥胖水平。最后，研究者们也在探索更有效、可持续的肥胖干预策略，如健康教育、运动干预和饮食干预，以帮助人们更好地控制体重，降低肥胖相关的健康风险。

尽管肥胖研究取得了一定的进展，但仍存在一些不足之处。对肥胖的发生发展机制的研究还不够深入，需要进一步探究。现有的肥胖预测模型的精度仍有待提高。此外，一些肥胖干预策略可能难以长期坚持，需要探索更加可持续的策略。现有的肥胖研究数据相对缺乏，需要进一步收集和积累数据，以便更好地进行肥胖研究。

未来的研究方向包括深入研究肥胖的发生发展机制，开发更加精准、高效的肥胖预测模型，探索更加有效、可持续的肥胖干预策略，以及加强肥胖研究数据的收集和积累。

4. 研究内容

4.1 数据集描述

表 4.1 数据变量描述

变量名	变量类型	描述
性别（Gender）	分类变量	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）	连续变量	受访者的身高，以厘米为单位
体重（Weight）	连续变量	受访者的体重，以千克为单位
年龄（Age）	连续变量	受访者的年龄，以年为单位
家族肥胖史（family_history_with_overweight）	分类变量	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）	分类变量	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）	分类变量	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）	分类变量	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）	分类变量	受访者在餐间摄入的食物，按 0 到 3 的等级来衡量
吸烟（SMOKE）	分类变量	如果受访者吸烟，则为 1；如果不吸烟，则为 0
饮水量（CH2O）	分类变量	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）	分类变量	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）	分类变量	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）	分类变量	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）	分类变量	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）	分类变量	表示受访者的主要交通方式
NObeeyesdad	分类变量	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）

4.2 研究流程

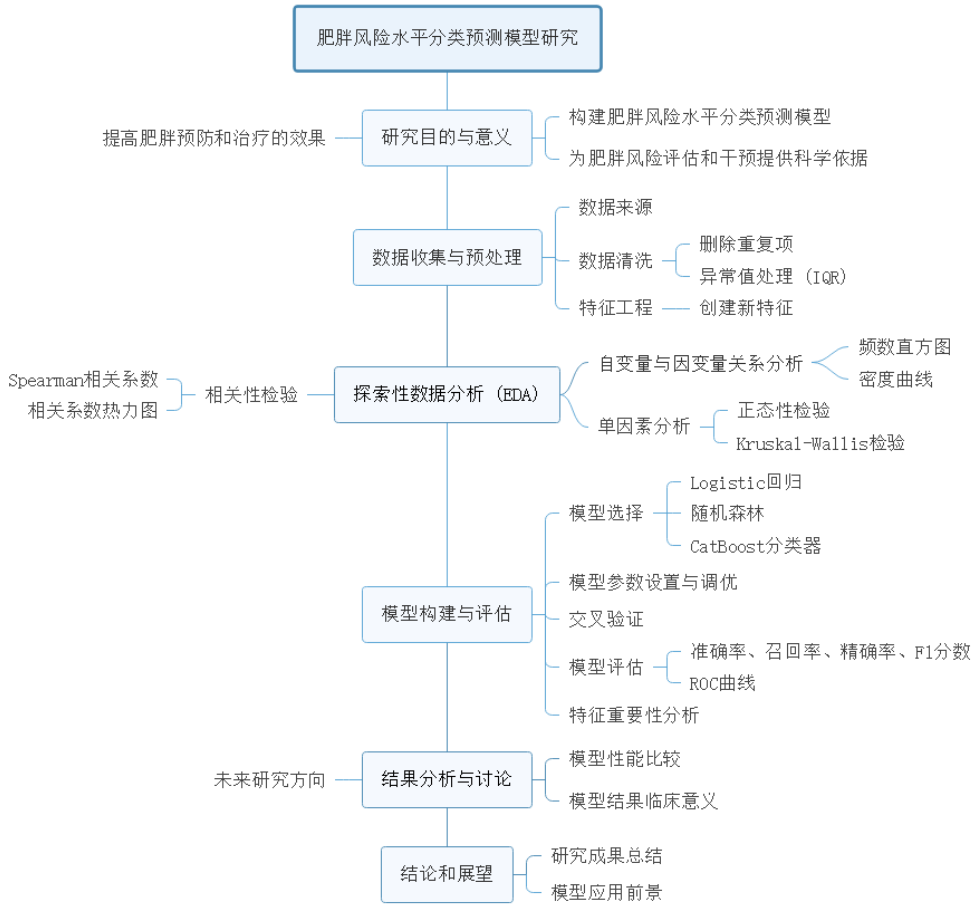

图 4.1 研究开展流程图

本研究针对肥胖问题，通过探索性数据分析（EDA）和相关性检验，构建了一个肥胖风险水平分类预测模型。研究背景和目的在于提供科学依据，为肥胖风险评估和干预提供支持。数据收集和预处理阶段，我们采用了 Spearman 相关系数和相关数热力图来筛选变量，并通过剔除重复项、清洗异常值以及创建新特征等步骤，对原始数据进行了全面的预处理。在 EDA 阶段，我们进行了自变量与因变量关系分析，包括频数直方图和密度曲线，以及单因素分析和正态性检验。这些分析帮助我们理解数据的基本分布和潜在模式。

模型选择与评估阶段，我们根据 EDA 结果选择了 Logistic 回归作为主要建模方法，并使用了随机森林和 CatBoost 分类器进行对比实验。为了评估模型的性能，我们进行了交叉验证，并计算了准确率、召回率、精确率和 F1 分数等指标。未来研究方向包括提高模型的泛化能力、引入更多有意义的相关变量以及探索更好的模型结构。本研究提出了一种基于 EDA 和统计检验的肥胖风险水平预测模型，通过深入的数据分析和模型选择，我们在一定程度上解决了肥胖风险评估

的问题。我们希望我们的研究成果能为公共卫生领域提供有益的参考，帮助人们更好地理解和预防肥胖。

5. 研究方案

5.1 数据预处理

5.1.1 数据清洗：删除重复项数据、检查数据是否有缺失

表 5.1 数据变量说明表

	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。

5.1.2 四分位数法处理异常值

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。

（5）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。

（6）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。

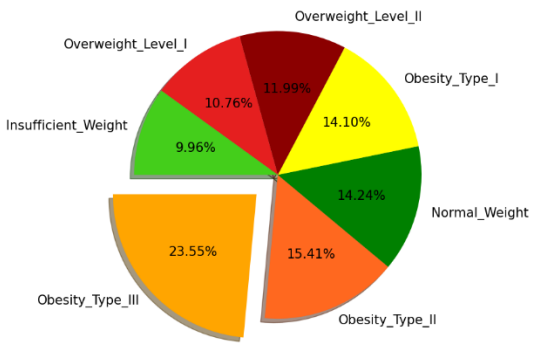

图 5.1 因变量分布情况

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%）。

5.1.3 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

本研究在数据预处理阶段，首先保留了原始的连续特征，如身高、体重、年龄等，以保留数据中的重要信息。接着，通过计算 BMI（体重指数）、每日用餐次数、总体育活动得分和每公斤体重的水分摄入量等新特征，进一步丰富了数据的信息维度。这些新特征的创建是基于对原始数据的深入分析，旨在捕捉数据中可能被忽视的模式。

在处理年龄特征时，本研究将年龄分为 Young、Adult 和 Elderly 三个类别，以简化模型并提高预测的准确性。此外，为了进一步简化模型并提高预测的准确性，本研究还将其他特征转化为分类变量，如将体重和身高换算为 BMI 指数并分等级，将 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）和 TUE（观看电子屏幕次数评分 0-2）等评分数据均转化为分类变量。其中，体重指数（BMI）等于 Weight/（Height）的平方，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

通过这种特征工程的方法，本研究在保留数据信息的同时，保持了模型的可解释性和稳健性。这为肥胖风险水平的分类预测模型提供了坚实的基础，有助于提高模型的预测能力和准确性。

5.2 探索性数据分析（EDA）（分析自变量和因变量之间的关系）

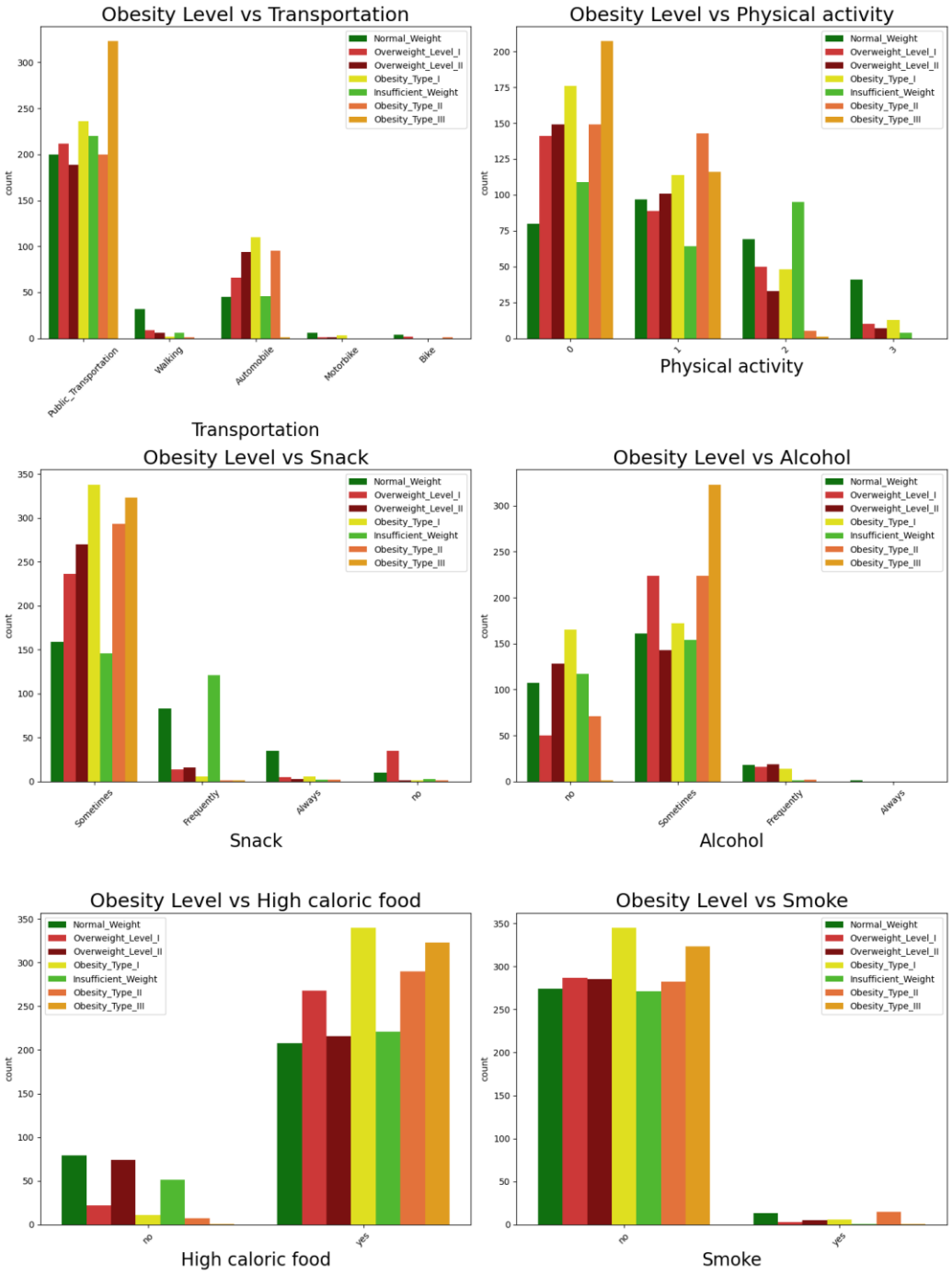

图 5.2 分类数据对应肥胖类型分布情况

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。

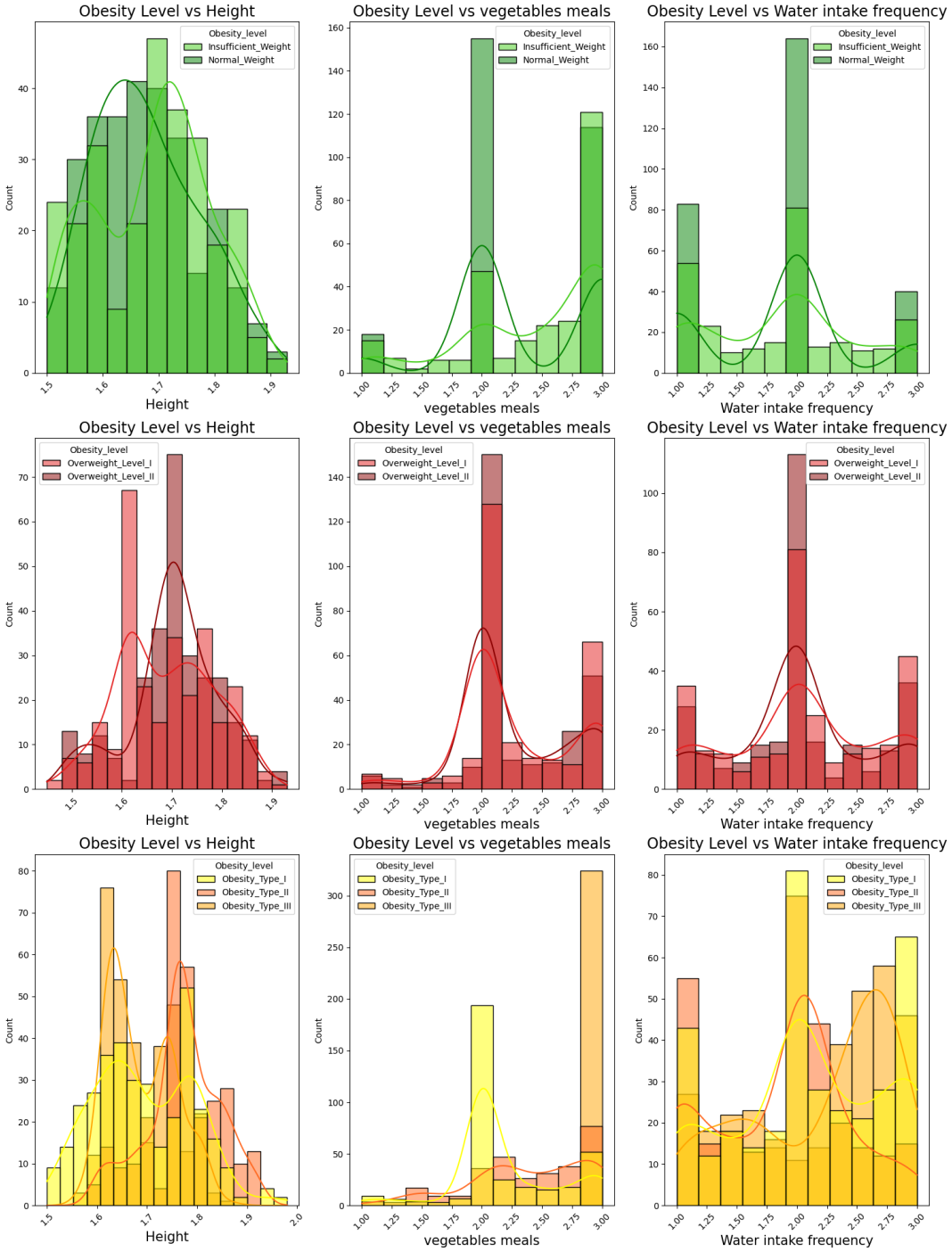

图 5.3 连续型数据对应肥胖类型分布情况

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。

5.3 单因素分析

5.3.1 假设检验

（1）定量数据与因变量 Y 体重分级（NObeyesdad）：多个独立样本 Kruskal-Wallis 检验

表 5.2 正态检验性结果

变量名	样本量	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
Water_Intake_Per_Kg	1376	0.024	0.01	1.265	2.501	0.919(0.000***)	0.123(0.000***)
Total_Activity_Score	1376	1.726	1.066	0.573	0.153	0.967(0.000***)	0.067(0.000***)
Meals_Per_Day	1376	5.429	0.568	-0.503	-0.666	0.861(0.000***)	0.211(0.000***)
Height	1376	1.714	0.087	-0.038	-0.367	0.996(0.001***)	0.043(0.012**)
BMI	1376	31.167	8.458	-0.014	-1.008	0.971(0.000***)	0.06(0.000***)
CH2O	1376	2.062	0.689	-0.082	-0.893	0.803(0.000***)	0.266(0.000***)
Weight	1376	92.136	27.435	0.033	-0.895	0.974(0.000***)	0.089(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

先对 Water_Intake_Per_Kg（每公斤体重的摄水量）、Total_Activity_Score（总体育得分）、Meals_Per_Day（用餐次数）、Height（身高）、BMI（体重指数）、CH2O（饮水量）、Weight（体重）这七个连续型数据做了正态性检验，因为样本量为 1376 组属于大样本研究，因此正态性检验方法采用的是 S-W 检验，以上七个变量都不服从正态分布（显著性 P 值都>0.05），因此假设检验只能使用多个独立样本 Kruskal-Wallis 检验。

表 5.3 Kruskal-Wallis 检验分析结果表

分析项	分组变量	样	中位数	标准	统计量	P	Cohen'

		本量		差			s f 值
Water_Intake_Per_Kg	Normal_Weight	196	0.03	0.01	513.465	0.000***	0.05
	Overweight_Level_I	148	0.026	0.008			
	Obesity_Type_I	194	0.022	0.007			
	Overweight_Level_II	165	0.025	0.006			
	Obesity_Type_II	212	0.018	0.004			
	Insufficient_Weight	137	0.038	0.013			
	Obesity_Type_III	324	0.02	0.005			
	总计	1376	0.023	0.01			
Total_Activity_Score	Normal_Weight	196	2	1.236	111.453	0.000***	0.022
	Overweight_Level_I	148	1.596	0.91			
	Obesity_Type_I	194	1.625	1.265			
	Overweight_Level_II	165	1.962	1.049			
	Obesity_Type_II	212	1.518	0.617			
	Insufficient_Weight	137	2.068	1.028			
	Obesity_Type_III	324	1.001	0.939			
	总计	1376	1.675	1.066			
Meals_Per_Day	Normal_Weight	196	5	0.604	556.851	0.000***	0.036
	Overweight_Level_I	148	5.003	0.48			
	Obesity_Type_I	194	5	0.481			
	Overweight_Level_II	165	5	0.544			
	Obesity_Type_II	212	5.228	0.46			
	Insufficient_Weight	137	5.721	0.602			
	Obesity_Type_III	324	6	0			

	总计	1376	5.443	0.568			
	Normal_Weight	196	1.67	0.096			
	Overweight_Level_I	148	1.705	0.098			
	Obesity_Type_I	194	1.73	0.079			
	Overweight_Level_II	165	1.706	0.081	285.803	0.000***	0.03
	Obesity_Type_II	212	1.784	0.056			
	Insufficient_Weight	137	1.704	0.084			
	Obesity_Type_III	324	1.669	0.065			
	总计	1376	1.714	0.087			
	Normal_Weight	196	22.2	1.861			
	Overweight_Level_I	148	25.9	0.582			
	Obesity_Type_I	194	32.2	1.149			
	Overweight_Level_II	165	28.2	0.867	1334.252	0.000***	0.069
	Obesity_Type_II	212	36.3	1.194			
	Insufficient_Weight	137	17.5	0.881			
	Obesity_Type_III	324	41.9	2.58			
	总计	1376	31.25	8.458			
	Normal_Weight	196	2	0.633			
	Overweight_Level_I	148	2	0.661			
	Obesity_Type_I	194	2	0.762			
	Overweight_Level_II	165	2	0.6	69.287	0.000***	0.015
	Obesity_Type_II	212	2	0.548			
	Insufficient_Weight	137	2	0.654			
	Obesity_Type_III	324	2	0.759			
	总计	137	2	0.689			

		6					
	Normal_Weight	196	62	9.24			
	Overweight_Level_I	148	75	8.56			
	Obesity_Type_I	194	96.468	9.668			
	Overweight_Level_II	165	82.523	7.54			
Weight	Obesity_Type_II	212	119.066	6.095	1197.39	0.000***	0.067
	Insufficient_Weight	137	50.166	5.037			
	Obesity_Type_III	324	112.049	15.532			
	总计	1376	90.835	27.435			

注：***、**、*分别代表 1%、5%、10%的显著性水平

上表展示了 Kruskal-Wallis 检验的结果，包括中位数、统计量与效应量 Cohen's f 值。

- 分析每个分析项的 P 值是否显著(P<0.05)。
- 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据中位数±标准差的方式对差异进行分析，反之则表明数据不呈现差异性。
- Cohen's f 值： 表示效应量大小，效应量小、中、大的区分临界点分别是：0.1、0.25 和 0.40。

根据上述七个连续型自变量和因变量 NObeeyesdad（肥胖等级分类）之间的 Kruskal-Wallis 检验结果可知， Water_Intake_Per_Kg（每公斤体重的摄水量）、Total_Activity_Score（总体育得分）、Meals_Per_Day（用餐次数）、Height（身高）、BMI（体重指数）、CH2O（饮水量）、Weight（体重）都与自变量 NObeeyesdad 的的七个分级之间的差异都存在统计学意义（P<0.05），即可以认为这十个自变量都是 NObeeyesdad 分级的影响因素。

表 5.4 事后多重分析

两独立样本		样本量		中位数		统计量	P	中位数值差值	Cohen's d 值
分组项 A	分组项 B	分组	分组	分组项 A	分组项 B				

项	项	A	B						
Weight_Normal_Weight	Weight_Overweight_Level_I	1916	1418	62	75	4740.5	0.000***	13	1.397
Weight_Normal_Weight	Weight_Obesity_Type_I	1916	1919	62	96.468	110.5	0.000***	34.468	3.63
Weight_Normal_Weight	Weight_Overweight_Level_II	1916	1625	62	82.523	1852	0.000***	20.523	2.292
Weight_Normal_Weight	Weight_Obesity_Type_II	1916	1921	62	119.066	0	0.000***	57.066	7.154
Weight_Normal_Weight	Weight_Insufficient_Weight	1916	1317	62	50.166	23665	0.000***	11.834	1.612
Weight_Normal_Weight	Weight_Obesity_Type_III	1916	1932	62	112.049	0	0.000***	50.049	4.321
Weight_Overweight_Level_I	Weight_Obesity_Type_I	1418	1919	75	96.468	1310	0.000***	21.468	2.37
Weight_Overweight_Level_I	Weight_Overweight_Level_II	1418	1625	75	82.523	6829.5	0.000***	7.523	0.869
Weight_Overweight_Level_I	Weight_Obesity_Type_II	1418	1921	75	119.066	0	0.000***	44.066	5.972
Weight_Overweight_Level_I	Weight_Insufficient_Weight	1418	1317	75	50.166	20162	0.000***	24.834	3.536
Weight_Overweight_Level_I	Weight_Obesity_Type_III	1418	1932	75	112.049	0	0.000***	37.049	3.34
Weight_Obesity_Type_I	Weight_Overweight_Level_II	1916	1625	96.468	82.523	2823	0.000***	13.945	1.694
Weight_Obesity_Type_I	Weight_Obesity_Type_II	1916	1921	96.468	119.066	1553	0.000***	22.598	2.653
Weight_Obesity_Type_I	Weight_Insufficient_Weight	1916	1317	96.468	50.166	26578	0.000***	46.302	5.801
Weight_Obesity_Type_I	Weight_Obesity_Type_III	1916	1932	96.468	112.049	5554	0.000***	15.582	1.764
Weight_Overweight_Level_II	Weight_Obesity_Type_II	1625	1921	82.523	119.066	10	0.000***	36.543	5.331

Weight_Overweight_Level_II	Weight_Insufficient_Weight	1613	82.523	50.166	22600	0.000***	32.357	4.912
Weight_Overweight_Level_II	Weight_Obesity_Type_III	1632	82.523	112.049	0	0.000***	29.526	2.906
Weight_Obesity_Type_II	Weight_Insufficient_Weight	2113	119.066	50.166	29044	0.000***	68.943	11.99
Weight_Obesity_Type_II	Weight_Obesity_Type_III	2132	119.066	112.049	362555	0.551	7.017	0.223
Weight_Insufficient_Weight	Weight_Obesity_Type_III	1332	50.166	112.049	0	0.000***	61.884	5.328

注：***、**、*分别代表 1%、5%、10%的显著性水平

上表展示了对分组变量进行两两独立样本 MannWhitney U 检验的结果，包括样本数、中位数、统计量与效应量 Cohen's d 值。

- 分析每个分析项的 P 值是否显著(P<0.05)。
- 若呈显著性，拒绝原假设，说明两组数据之间存在显著性差异，可以根据中位数对差异进行分析，反之则表明数据不呈现差异性。

根据事后多重检验的结果可知，自变量 NObeyesdad 的 7 个分级，两两之间的差异都具有统计学意义（P<0.001），自变量 NObeyesdad 的 7 个分级两两之间都有不同。

（2）分类数据（年龄分级）与因变量 Y 体重分级（NObeyesdad）：多个独立样本 Kruskal-Wallis 检验

5.3.2 相关性分析（分析自变量与因变量之间的整体的相关性）

表 5.5 Kruskal-Wallis 检验分析结果表

分析项	分组变量	样本量	中位数	标准差	统计量	P	Cohen's f 值
CALC	Normal_Weight	196	1	0.586	222.888	0.000***	0.027
	Overweight_Level_I	148	1	0.476			
	Obesity_Type_I	194	0	0.521			

	Overweight_Level_II	165	1	0.617			
	Obesity_Type_II	212	1	0.267			
	Insufficient_Weight	137	1	0.5			
	Obesity_Type_III	324	1	0.056			
	总计	1376	1	0.475			
	Normal_Weight	196	1	0.29			
	Overweight_Level_I	148	1	0.181			
	Obesity_Type_I	194	1	0.199			
Age_Group	Overweight_Level_II	165	1	0.250	63.112	0.000***	0.016
	Obesity_Type_II	212	1				
	Insufficient_Weight	137	1	0.339			
	Obesity_Type_III	324	1	0			
	总计	1376	1	0.204			
	Normal_Weight	196	2	1.225			
	Overweight_Level_I	148	2	1.31			
MTRANS	Obesity_Type_I	194	2	1.082	83.852	0.000***	0.018
	Overweight_Level_II	165	2	1.197			
	Obesity_Type_II	212	2	1.385			

	Insufficient_Weight	137	2	0.883			
	Obesity_Type_III	324	2	0.167			
	总计	1376	2	1.095			
	Normal_Weight	196	1	0.732			
	Overweight_Level_I	148	1	0.383			
	Obesity_Type_I	194	1	0.291			
	Overweight_Level_II	165	1	0.323	269.611	0.000***	0.031
CAEC	Obesity_Type_II	212	1	0.097			
	Insufficient_Weight	137	1	0.567			
	Obesity_Type_III	324	1	0.056			
	总计	1376	1	0.426			
	Normal_Weight	196	0	0.31			
	Overweight_Level_I	148	0	0.35			
	Obesity_Type_I	194	0	0.101			
	Overweight_Level_II	165	0	0.154	96.099	0.000***	0.02
SCC	Obesity_Type_II	212	0	0.069			
	Insufficient_Weight	137	0	0.339			
	Obesity_Type_III	324	0	0			

	I						
	总计	1376	0	0.215			
	Normal_Weight	196	2	0.604			
	Overweight_Level_I	148	2	0.525			
	Obesity_Type_I	194	2	0.498			
	Overweight_Level_II	165	2	0.522	435.66	0.000***	0.033
FCVC	Obesity_Type_II	212	2	0.624			
	Insufficient_Weight	137	3	0.665			
	Obesity_Type_III	324	3	0			
	总计	1376	3	0.593			
	Normal_Weight	196	0	0.198			
	Overweight_Level_I	148	0	0.082			
	Obesity_Type_I	194	0	0.072			
	Overweight_Level_II	165	0	0.11	39.871	0.000***	0.012
SMOKE	Obesity_Type_II	212	0	0.257			
	Insufficient_Weight	137	0	0.085			
	Obesity_Type_III	324	0	0.056			
	总计	1376	0	0.14			

		6	4			
	Normal_Weight	196	1	0.498		
	Overweight_Level_I	148	1	0.5		
	Obesity_Type_I	194	1	0.439		
Gender	Overweight_Level_II	165	1	0.461	616.485	0.000***	0.042
	Obesity_Type_II	212	1	0		
	Insufficient_Weight	137	0	0.481		
	Obesity_Type_III	324	0	0.056		
	总计	1376	1	0.5		
	Normal_Weight	196	1	0.445		
	Overweight_Level_I	148	1	0.274		
	Obesity_Type_I	194	1	0.187		
FAVC	Overweight_Level_II	165	1	0.45	183.264	0.000***	0.026
	Obesity_Type_II	212	1	0.152		
	Insufficient_Weight	137	1	0.429		
	Obesity_Type_III	324	1	0.056		
	总计	1376	1	0.318		
family_history_with_overweight	Normal_Weight	196	1	0.499	417.856	0.000***	0.041

Overweight_Level_I	148	1	0.426
Obesity_Type_I	194	1	0.159
Overweight_Level_II	165	1	0.215
Obesity_Type_II	212	1	0
Insufficient_Weight	137	0	0.502
Obesity_Type_II	324	1	0
总计	1376	1	0.357

注：***、**、*分别代表 1%、5%、10%的显著性水平

根据上述十个自变量与因变量 NObeeyesdad 之间的 Kruskal-Wallis 检验结果可知，自变量 CALC（饮酒频率）、Age_Group（年龄分级）、MTRANS（主要交通方式）、CAEC（餐间食物摄入）、SCC（监测卡路里摄入）、FCVC（通常食用蔬菜）、SMOKE（吸烟）、Gender（性别）、FAVC（经常食用高热量食物）、family_history_with_overweight（家族肥胖史）都与因变量 NObeeyesdad 之间的差异都具有统计学意义（P<0.001），既可以认为这十个自变量都是因变量 NObeeyesdad 的影响因素。

表 5.6 事后多重分析

两独立样本							
分组项 A	分组项 B	样本量分组项 A	中位数分组项 B	统计量	P	中位数差值	Cohen's d 值
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Overweight_Level_I	119	411	1142.006	0.000***	0	0.453
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Obesity_Type_I	119	911	1096.610	0.000***	0	1.141

		6	4					
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Overweight_Level_II	196	165	11	9694	0.000***	0	1.013
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Obesity_Type_II	196	212	11	11448	0.000***	0	1.299
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Insufficient_Weight	196	137	10	14258	0.532	1	0.124
family_history_with_overweight_Normal_Weight	family_history_with_overweight_Obesity_Type_III	196	324	11	17496	0.000***	0	1.467
family_history_with_overweight_Overweight_Level_I	family_history_with_overweight_Obesity_Type_I	148	194	11	11331	0.000***	0	0.691
family_history_with_overweight_Overweight_Level_I	family_history_with_overweight_Overweight_Level_II	148	165	11	9914.5	0.000***	0	0.566
family_history_with_overweight_Overweight_Level_I	family_history_with_overweight_Obesity_Type_II	148	212	11	11978	0.000***	0	0.866
family_history_with_overweight_Overweight_Level_I	family_history_with_overweight_Insufficient_Weight	148	137	10	12920.5	0.000***	1	0.591
family_history_with_overweight_Overweight_Level_I	family_history_with_overweight_Obesity_Type_III	148	324	11	18306	0.000***	0	0.992
family_history_with_overweight_Obesity_Type_I	family_history_with_overweight_Overweight_Level_II	194	165	11	16368.5	0.503	0	0.121
family_history_with_overweight_Obesity_Type_I	family_history_with_overweight_Obesity_Type_II	194	212	11	20034	0.038**	0	0.235

family_history_with_overweight_Obesity_Type_I	family_history_with_overweight_Insufficient_Weight	194	137	10	19736.5	0.000***	1	1.407
family_history_with_overweight_Obesity_Type_I	family_history_with_overweight_Obesity_Type_III	194	324	11	30618	0.007***	0	0.265
family_history_with_overweight_Overweight_Level_II	family_history_with_overweight_Obesity_Type_II	165	212	11	16642	0.002***	0	0.34
family_history_with_overweight_Overweight_Level_II	family_history_with_overweight_Insufficient_Weight	165	137	10	16529.5	0.000***	1	1.238
family_history_with_overweight_Overweight_Level_II	family_history_with_overweight_Obesity_Type_III	165	324	11	25434	0.000***	0	0.388
family_history_with_overweight_Obesity_Type_II	family_history_with_overweight_Insufficient_Weight	212	137	10	21942	0.000***	1	1.627
family_history_with_overweight_Obesity_Type_II	family_history_with_overweight_Obesity_Type_III	212	324	11	34344	NaN	0	
family_history_with_overweight_Insufficient_Weight	family_history_with_overweight_Obesity_Type_III	137	324	01	10854	0.000***	1	1.871

注：***、**、*分别代表 1%、5%、10%的显著性水平

将 NObeeyesdad 作为分组依据，结合变量 family_history_with_overweight（家族肥胖史）为“1”的，进行事后多重检验，除了 family_history_with_overweight_Normal_Weight 和 family_history_with_overweight_Insufficient_Weight 之间（P=0.532>0.05）、family_history_with_overweight_Obesity_Type_I 和 family_history_with_overweight_Overweight_Level_II 之间（P=0.503>0.05）、family_history_with_overweight_Obesity_Type_II 和 family_history_with_overweight_Obesity_Type_III 之间（P 为 NaN）的差异

暂不存在统计学意义之外，其余变量两两之间的差异都具有统计学意义。

5.3.2 相关性检验（分析自变量与因变量之间的整体的相关性）

因为因变量为分类数据因此分析自变量和因变量之间的相关性的时候采用 spearman 相关性分析。

表 5.7 spearman 相关系数表

	Gender	Height	Weight	CALC	FAVC	FCVC	NCP	SCC	SMOKE	CH20	family_history_with_overweight	FAF	TUE	CAEC	MTRANS	BMI	Meals_Per_Day	Total_Activity_Score	Water_Intake_Per_Kg	Age	NObeyesdad
Gender	1(0.000***)	0.62(0.000***)	0.05(0.004**)	-0.058(0.003**)	0.032(0.005**)	0.418(0.000***)	-0.068(0.012**)	0.119(0.000***)	0.071(0.008***)	0.034(0.000***)	0.007(0.009***)	0.166(0.000***)	-0.018(0.517**)	-0.069(0.017**)	0.189(0.000***)	-0.181(0.000***)	-0.046(0.000***)	0.127(0.000***)	0.027(0.317)	0.029(0.286)	-0.373(0.000***)
Height	0.62(0.000***)	1(0.000***)	0.448(0.000***)	0.132(0.000***)	0.177(0.000***)	-0.135(0.000***)	0.016(0.565)	-0.162(0.000***)	0.083(0.002**)	0.141(0.000***)	0.212(0.000***)	0.277(0.000***)	-0.065(0.016**)	-0.074(0.006***)	0.117(0.000***)	-0.121(0.000***)	-0.125(0.000***)	0.252(0.000***)	-0.178(0.000***)	0.100(0.000***)	0.049(0.879)
Weight	0.05	0.44	1(0.000***)	0.26	0.31	0.22	0.02	-0.04	0.22	0.494(0.000***)	-0.01	-0.01	-0.01	0.01	0.93	0.25	-0.10	-0.50	-0.55		

	5(0.042**)	8(0.000***)	000***)	5(0.000***)	8(0.000***)	3(0.000***)	7(0.313)	242(0.000***)	7(0.079*)	9(0.000***)	0***)	059(0.029**)	058(0.032**)	342(0.000***)	7(0.537)	(0.000***)	8(0.000***)	05(0.000***)	85(0.000***)	193(0.000***)	7(0.000***)
CALC	-0.058(0.030**)	0.132(0.000***)	0.265(0.000***)	1(0.000***)	0.127(0.000***)	0.131(0.000***)	0.113(0.000***)	0.005(0.850)	0.1(0.000***)	0.065(0.016**)	0.022 (0.415)	-0.178(0.000***)	-0.015(0.591)	-0.106(0.000***)	0.004(0.892)	0.23(0.000***)	0.171(0.000***)	-0.164(0.000***)	-0.134(0.000***)	-0.103(0.000***)	0.205(0.000***)
FAVC	0.032(0.235)	0.177(0.000***)	0.318(0.000***)	0.127(0.000***)	1(0.000***)	0.007(0.798)	0.013(0.636)	-0.259(0.000***)	-0.027(0.318)	0.004(0.888)	0.253 (0.000***)	-0.155(0.000***)	0.044(0.101)	-0.122(0.000***)	0.07(0.009***)	0.294(0.000***)	0.019(0.492)	-0.108(0.000***)	-0.273(0.000***)	-0.035(0.191)	0.176(0.000***)
FCVC	-0.418(0.000***)	-0.135(0.000***)	0.223(0.000***)	0.131(0.000***)	0.007(0.798)	1(0.000***)	0.028(0.292)	0.092(0.001***)	-0.004(0.869)	0.142(0.000***)	0.011 (0.683)	0.000(0.992)	-0.052(0.054*)	0.016(0.558)	-0.088(0.001***)	0.315(0.000***)	0.9(0.000***)	-0.031(0.247)	-0.072(0.007***)	-0.049(0.069*)	0.454(0.000***)

NCP	－0.068(0.012**)	0.016(0.565)	0.027(0.313)	0.113(0.000***)	0.013(0.636)	0.028(0.292)	1(0.000***)	－0.048(0.078*)	0.024(0.373)	0.041(0.129)	－0.03(0.258)	－0.019(0.481)	－0.022(0.409)	0.039(0.146)	－0.014(0.600)	0.024(0.381)	0.173(0.000***)	－0.027(0.315)	0.026(0.338)	－0.054(0.044**)	0.062(0.022**)
SCC	－0.119(0.000***)	－0.162(0.000***)	－0.242(0.000***)	0.005(0.850)	－0.259(0.000***)	0.092(0.001***)	－0.048(0.078*)	1(0.000***)	0.037(0.166)	－0.01(0.714)	－0.274(0.000***)	－0.07(0.009***)	－0.078(0.004***)	0.141(0.000***)	－0.014(0.606)	－0.228(0.000***)	0.054(0.047**)	0.036(0.180)	0.176(0.000***)	0.034(0.203)	－0.142(0.000***)
SMOKE	0.071(0.008***)	0.083(0.002***)	0.047(0.079*)	0.1(0.000***)	－0.027(0.318)	－0.004(0.869)	0.024(0.373)	0.037(0.166)	1(0.000***)	－0.085(0.002***)	0.005(0.858)	0.026(0.336)	0.019(0.479)	0.028(0.296)	0.029(0.276)	0.002(0.937)	0.005(0.852)	0.046(0.086*)	－0.088(0.001***)	－0.031(0.245)	－0.037(0.168)
CH2O	0.034(0.208)	0.141(0.000***)	0.229(0.000***)	0.065(0.016**)	0.004(0.888)	0.142(0.000***)	0.041(0.129)	－0.01(0.714)	－0.085(0.002***)	1(0.000***)	0.18(0.000***)	0.054(0.044**)	－0.084(0.002***)	－0.116(0.000***)	－0.059(0.027**)	0.204(0.000***)	0.166(0.000***)	0.024(0.373)	0.568(0.000***)	－0.02(0.452)	0.171(0.000***)

									)				)	)							
family_history_with_overweight	0.07(0.009***)	0.212(0.000***)	0.494(0.000***)	0.022(0.415)	0.253(0.000***)	0.011(0.683)	-0.03(0.258)	-0.274(0.000***)	0.005(0.858)	0.18(0.000***)	1(0.000***)	-0.117(0.000***)	0.042(0.116)	-0.171(0.000***)	0.032(0.232)	0.488(0.000***)	0.037(0.176)	-0.084(0.002***)	-0.242(0.000***)	-0.12(0.000***)	0.248(0.000***)
FAF	0.166(0.000***)	0.277(0.000***)	-0.059(0.029**)	-0.178(0.000***)	-0.155(0.000***)	0.000(0.992)	-0.019(0.481)	0.07(0.009***)	0.026(0.336)	0.054(0.044**)	-0.117(0.000***)	1(0.000***)	0.078(0.004***)	0.063(0.019**)	-0.03(0.268)	-0.18(0.000***)	-0.015(0.576)	0.8(0.000***)	0.145(0.000***)	0.089(0.001***)	-0.149(0.000***)
TUE	-0.018(0.513)	0.065(0.016**)	-0.058(0.032**)	-0.015(0.591)	0.044(0.101)	-0.052(0.054*)	-0.022(0.409)	-0.078(0.004***)	0.019(0.479)	-0.084(0.002***)	0.042(0.116)	0.078(0.004***)	1(0.000***)	0.032(0.231)	-0.079(0.003***)	-0.065(0.016**)	-0.054(0.043**)	0.566(0.000***)	0.022(0.419)	0.126(0.000***)	0.015(0.574)
CAEC	-0.065(0.01	-0.074(0.00	-0.342(0.00	-0.106(0.00	-0.122(0.00	0.016(0.558)	0.039(0.146)	0.141(0.000*	0.028(0.296)	-0.116(0.00	-0.171(0.000***)	0.063(0.019*	0.032(0.231)	1(0.000***)	-0.033(0.22	-0.358(0.00	0.014(0.609)	0.085(0.002***)	0.174(0.000***)	0.028(0.299)	-0.151(0.00

	7**)	6***)	0***)	0***)	0***)			**)		0***)		*)			5)	0***)					0***)
MTRANS	0.189(0.000***)	0.117(0.000***)	0.017(0.537)	0.004(0.892)	0.07(0.009***)	-0.088(0.001***)	-0.014(0.600)	-0.014(0.606)	0.029(0.276)	-0.059(0.027**)	0.032(0.232)	-0.03(0.268)	-0.079(0.003***)	-0.033(0.225)	1(0.000***)	-0.028(0.304)	-0.098(0.000***)	-0.053(0.049**)	-0.055(0.042**)	-0.086(0.001***)	-0.117(0.000***)
BMI	-0.181(0.000***)	0.12(0.000***)	0.93(0.000***)	0.23(0.000***)	0.294(0.000***)	0.315(0.000***)	0.024(0.381)	-0.228(0.000***)	0.002(0.937)	0.204(0.000***)	0.488(0.000***)	-0.18(0.000***)	-0.065(0.016**)	-0.358(0.000***)	-0.028(0.304)	1(0.000***)	0.357(0.000***)	-0.208(0.000***)	-0.578(0.000***)	-0.201(0.000***)	0.65(0.000***)
Meals_Per_Day	-0.46(0.000***)	-0.125(0.000***)	0.258(0.000***)	0.171(0.000***)	0.019(0.492)	0.9(0.000***)	0.173(0.000***)	0.054(0.047**)	0.005(0.852)	0.166(0.000***)	0.037(0.176)	-0.015(0.576)	-0.054(0.043**)	0.014(0.609)	-0.098(0.000***)	0.357(0.000***)	1(0.000***)	-0.043(0.112)	-0.078(0.004***)	-0.055(0.042**)	0.484(0.000***)
Total_Activity_Score	0.127(0.	0.252(0.	-0.105(	-0.164(	-0.108(	-0.031(	-0.027(	0.036(0.	0.046(0.	0.024(0.	-0.084(0.002***)	0.8(0.00	0.566(0.	0.085(0.	-0.053(	-0.208(	-0.043(	1(0.000***)	0.166(0.000*	0.14(0.0	-0.134(

	000***)	000***)	0.000***)	0.000***)	0.000***)	0.247)	0.315)	180)	086*)	373)		0***)	000***)	002***)	0.049**)	0.000***)	0.112)		**)	00***)	0.000***)
Water_Intake_Per_Kg	0.027(0.317)	-0.18(0.000***)	-0.585(0.000***)	-0.134(0.000***)	-0.273(0.000***)	-0.072(0.007***)	0.026(0.338)	0.176(0.000***)	-0.088(0.001***)	0.568(0.000***)	-0.242(0.000***)	0.145(0.000***)	0.022(0.419)	0.174(0.000***)	-0.055(0.042**)	-0.578(0.000***)	-0.078(0.004***)	0.166(0.000***)	1(0.000***)	0.14(0.000***)	-0.282(0.000***)
Age_Group	0.029(0.286)	0.000(0.990)	-0.193(0.000***)	-0.103(0.000***)	-0.035(0.191)	-0.049(0.069*)	-0.054(0.044**)	0.034(0.203)	-0.031(0.245)	-0.02(0.452)	-0.12(0.000***)	0.089(0.001***)	0.126(0.000***)	0.028(0.299)	-0.086(0.001***)	-0.201(0.000***)	-0.055(0.042**)	0.14(0.000***)	0.14(0.000***)	1(0.000***)	-0.092(0.001***)
NObeyesdad	-0.373(0.000***)	0.004(0.879)	0.557(0.000***)	0.205(0.000***)	0.176(0.000***)	0.454(0.000***)	0.062(0.022**)	-0.142(0.000***)	-0.037(0.168)	0.171(0.000***)	0.248(0.000***)	-0.149(0.000***)	0.015(0.574)	-0.151(0.000***)	-0.117(0.000***)	0.65(0.000***)	0.484(0.000***)	-0.134(0.000***)	-0.282(0.000***)	-0.092(0.001***)	1(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

上表展示了模型检验的参数结果表，包括了相关系数、显著性 P 值。

1. 先对 XY 之间是否存在统计上的显著性关系进行检验，判断 P 值是否呈现显著性(P<0.05)。

2. 若呈现显著性，则说明两变量之间存在相关性，反之，则两变量之间不存在相关性。

3. 分析相关系数的正负向以及相关性程度。

根据上述自变量和因变量之间的 spearman 检验结果可知，因变量 NObeyesdad 与自变量 Gender、Weight、CALC、FAVC、FCVC、NCP、SCC、CH2O、family_history_with_overweight、FAF、CAEC、MTRANS、BMI、Meals_Per_Day、Total_Activity_Score、Water_Intake_Per_Kg 、Age_Group 这 17 个自变量之间存在明显的相关性（P<0.05），但是与自变量 Height、SMOKE、TUE 这 3 个自变量之间咱没有明显的相关性(P>0.05)；其中，因变量 NObeyesdad 与 Weight、CALC、FAVC、FCVC、NCP、family_history_with_overweight、FAF、BMI、Meals_Per_Day 这九个自变量之间是正相关（相关系数 r>0），与 Gender、SCC、CH2O、CAEC、MTRANS、Total_Activity_Score、Water_Intake_Per_Kg 、Age_Group 这八个自变量之间是负相关（相关系数 r<0）。

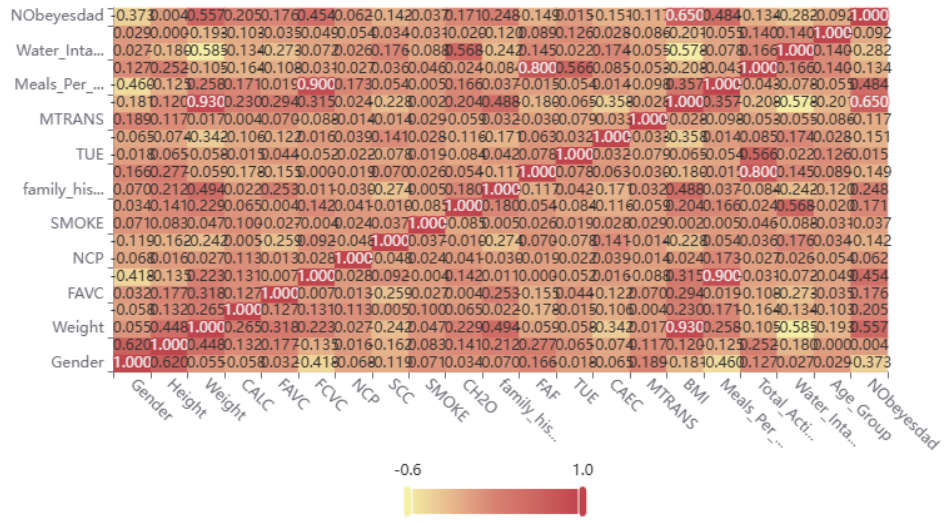

图 5.4 个变量直接相关系数热力图

根据各变量之间的相关系数热图可知，因变量 NObeyesdad 与自变量 Gender（R=-0.373）、Weight(R=0.557)、CALC(R=0.205)、FAVC(R=0.176)、FCVC(R=0.454)、NCP（R=0.062）、SCC（R=0.142）、CH2O（R=0.171）、family_history_with_overweight(R=0.248)、FAF(R=0.149)、CAEC(R=0.154)、

MTRANS （ R=0.117 ）、 BMI （ R=0.650 ）、 Meals_Per_Day （ R=0.484 ）、Total_Activity_Score(R=0.134)、Water_Intake_Per_Kg(R=0.282）、Age_Group（R=0.092）这 17 个自变量之间存在明显的相关性。

5.4 模型构建

5.4.1 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

6	截距	12.749	3.083	17.094	1	<.001			
	SMOKE	-.827	.891	.862	1	.353	.437	.076	2.508
	family_history_with_overweight	2.852	.425	44.942	1	<.001	17.321	7.524	39.873
	FAVC	-.629	.308	4.172	1	.041	.533	.292	.975
	CAEC	-1.350	.276	23.871	1	<.001	.259	.151	.446
	TUE	.300	.191	2.472	1	.116	1.350	.929	1.963
	Gender	.193	.276	.492	1	.483	1.213	.707	2.082
	FAF	-.122	.148	.685	1	.408	.885	.662	1.182
	FCVC	.127	.244	.271	1	.603	1.135	.704	1.830
	MTRANS	.058	.103	.317	1	.573	1.060	.866	1.297
	CALC	.106	.234	.204	1	.651	1.112	.703	1.759
	SCC	-1.307	.669	3.812	1	.051	.271	.073	1.005
	Age_Group	.858	.474	3.272	1	.070	2.358	.931	5.972
	CH2O	.595	.220	7.342	1	.007	1.813	1.179	2.788
	NCP	-5.293	1.012	27.339	1	<.001	.005	.001	.037

图 5.5 Logistic 回归参数估计值

根据采用多元逐步向前逻辑回归回归，最终发现与因变量之间的差异具有统计学意义的变量有 family_history_with_overweight （ P<0.001 ）， FAVC（P=0.041<0,05)，CAEC（P<0.001)，CH2O（P=0.007<0.05)，NCP（P<0.001）这四个自变量与因变量之间存在相关性（P<0.05)，即这两个因素都是心肌炎的危险因素，Logistic 回归方程为：

Logit（P）=12.749-0.827 SMOKE-0.629 FAVC-1.35 CAEC+0.3 TUE+0.193 Gender-0.122 FAF+0.127 FCVC+0.058 MTRANS+0.106 CALC-1.307 SCC+0.858 Age_Group+0.595 CH2O-5.293 NCP

校正其他 7 个因素后，有家族肥胖史的人患肥胖的风险是没有家族肥胖史的人的 17.321 倍；经常食用高热量食物每增高一级患肥胖的分险就增加 0.533 倍；餐间食物摄入每增高一级患肥胖的分险就增加 0.259 倍；饮水量每增高一级患肥胖的分险就增加 1.813 倍；餐间食物摄入每增高一级患肥胖的分险就增加 0.259 倍；每日主餐次数每增多一次患肥胖的分险就增加 0.005 倍；餐间食物摄入每增高一级患肥胖的分险就增加 0.259 倍；吸烟的人患肥胖是不吸烟人的 0.437 倍；屏幕使用时间每增高一级患肥胖的分险就增加 1.35 倍；男性比女性的患肥胖的

概率会增加 1.213 倍；体力活动量每增高一级患肥胖的分险就增加 0.885 倍；通常食用蔬菜每增高一级患肥胖的分险就增加 1.135 倍；饮酒的人患肥胖是不饮酒人的 1.112 倍；监测卡路里的人患肥胖是不监测卡路里的 0.271 倍；年龄分级每增高一级患肥胖的分险就增加 0.005 倍。

模型拟合信息

模型	模型拟合条件 -2 对数似然	似然比检验 卡方	自由度	显著性
仅截距	5246.252			
最终	2784.991	2461.261	84	.000

拟合优度

	卡方	自由度	显著性
皮尔逊	29976.989	8148	.000
偏差	2784.991	8148	1.000

伪 R 方

考克斯-斯奈尔	.833
内戈尔科	.852
麦克法登	.469

图 5.6 模型拟合优度

根据多元 Logistic 回归模型拟合优度检验可知，该模型的-2 自然对数为 5246.252，拟合优度皮尔逊卡方值为 29976.989，显著性 P 值小于 0.001，因此该模型的拟合优度情况较好。

表 5.8 Logistic 逻辑回归模型参数

参数名	参数值
数据切分	0.7
交叉验证	5
正则化	none
设置常数项	true
误差收敛条件	0.001
最大迭代次数	1000

本研究 Logistic 逻辑回归模型的参数如上表所示，数据切分比例为 0.7，将数据切分为训练集和验证集，模型训练过程中进行五折交叉验证，误差收敛条件设置为 0.001，最大迭代次数设置为 1000，以调整到模型的最佳预测效率。

图 5.7 Logistic 逻辑回归混淆矩阵热力图

从混淆矩阵可以看出，该分类器在预测“正常体重”和“超重”方面表现良好，但在预测“肥胖_I 型”、“肥胖_II 型”和“体重偏瘦”方面的准确性较低。特别是对于“肥胖_I 型”，其假反例数高达 56 人，表明该类别存在严重的误判问题。这可能意味着需要进一步调整模型或收集更多相关数据来提高其性能。

表 5.9 Logistic 逻辑回归模型评估结果

	准确率	召回率	精确率	F1
训练集	0.95	0.95	0.951	0.95
交叉验证集	0.908	0.908	0.909	0.906
测试集	0.908	0.908	0.911	0.908

本研究的模型评估结果如上图所示，该模型的准确率、召回率、精确率、F1 分数都在 90%以上；训练集的准确率、召回率、F1 分数更是高达 95%，准确率达 95.1%。

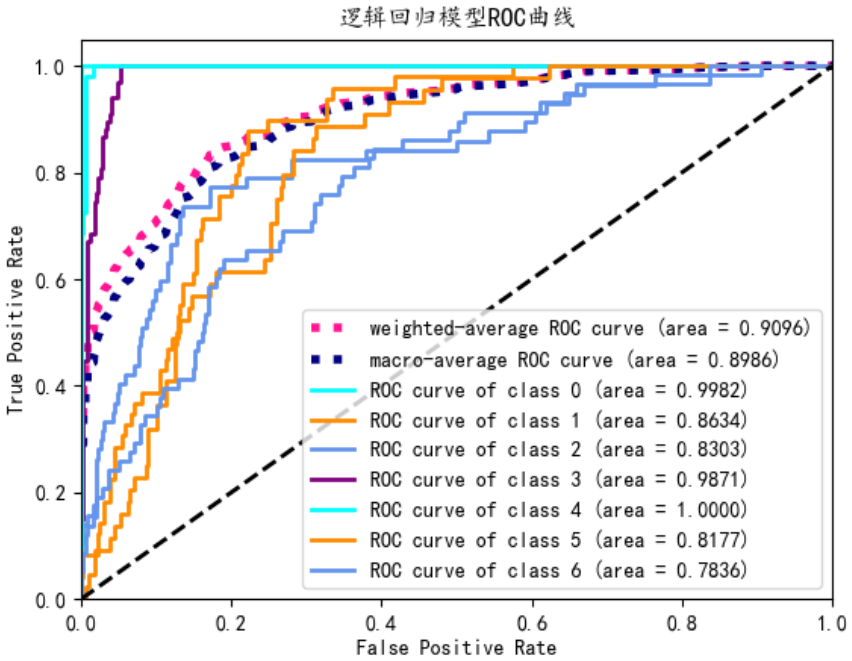

图 5.8 Logistic 逻辑回归模型的 ROC 曲线

根据绘制出的 ROC 曲线可知，体重不足（Insufficient_Weight）类别的曲线下面积达 0.9982，正常体重（Normal_Weight）类别的曲线下面积达 0.8634，肥胖 I 型（Obesity_Type_I）类别的曲线下面积达 0.8303，肥胖 II 型（Obesity_Type_II）类别的曲线下面积达 0.9871，肥胖 III 型（Obesity_Type_III）类别的曲线下面积达 1.000，超重 I 级（Overweight_Level）类别的曲线下面积达 0.8177，超重 II 级（Overweight_LevelI）类别的曲线下面积达 0.7836；各类别的预测准确率不一，肥胖 III 型的准确率最高达到了 100%，超重 II 级的准确率相对而言最低。

5.4.2 使用随机森林分类算法构建模型

表 5.10 随机森林模型参数

参数名	参数值
数据切分	0.7
交叉验证	5
节点分裂评价准则	gini
决策树数量	100
有放回采样	true
划分时考虑的最大特征比例	auto

内部节点分裂的最小样本数	2
叶子节点的最小样本数	1
叶子节点中样本的最小权重	0
树的最大深度	10
叶子节点的最大数量	50
节点划分不纯度的阀值	0

随机森林模型作为一种高效的机器学习算法，以其强大的预测能力和稳健性而著称。在本研究中，我们构建的随机森林模型采用了 0.7 的数据切分比例，即70%的数据用于模型训练，而剩余的 30%用于测试，确保了模型在未知数据上的泛化能力。此外，通过实施 5 次交叉验证，我们进一步增强了模型评估的准确性和可靠性，这种方法通过将数据集等分，轮流使用其中一份作为测试集，其余作为训练集，从而在不同数据子集上评估模型性能。

在节点分裂的评价标准上，我们选择了 Gini 指数作为衡量标准，这一指标有效量化了节点的不纯度，指导我们选择最优的分裂点，以期达到更准确的分类效果。模型中决策树的数量设定为 100，这一数量既能保证模型具有足够的多样性，又能在计算效率和预测精度之间取得平衡。有放回采样作为随机森林的核心机制，通过从原始数据集中重复抽取样本来构建每棵决策树，从而提升了模型的鲁棒性。

为了进一步优化模型，我们采用了自动调整的最大特征比例，使算法能够根据数据集的特征数量智能选择特征子集，这有助于在增加模型复杂度的同时避免过拟合。内部节点分裂的最小样本数设定为 2，这一设置有助于模型避免对个别样本的过度依赖，而叶子节点的最小样本数则设定为 1，保证了模型在细节上的区分能力。

在控制模型复杂度方面，我们将树的最大深度限制在 10，这有助于防止模型过度拟合到训练数据。同时，叶子节点的最大数量设定为 50，进一步限制了模型的复杂性，确保了模型在保持预测精度的同时，不会因过于复杂的结构而影响其泛化能力。

表 5.11 随机森林特征重要性

特征名称	特征重要性
Gender	5.70%
Height	6.10%
Weight	24.00%
CALC	1.80%

FAVC	1.00%
FCVC	1.80%
NCP	0.20%
SCC	0.30%
SMOKE	0.00%
CH2O	1.20%
family_history_with_overweight	1.90%
FAF	0.90%
TUE	1.00%
CAEC	1.40%
MTRANS	0.60%
BMI	35.40%
Meals_Per_Day	8.10%
Total_Activity_Score	2.50%
Water_Intake_Per_Kg	6.00%
Age_Group	0.20%

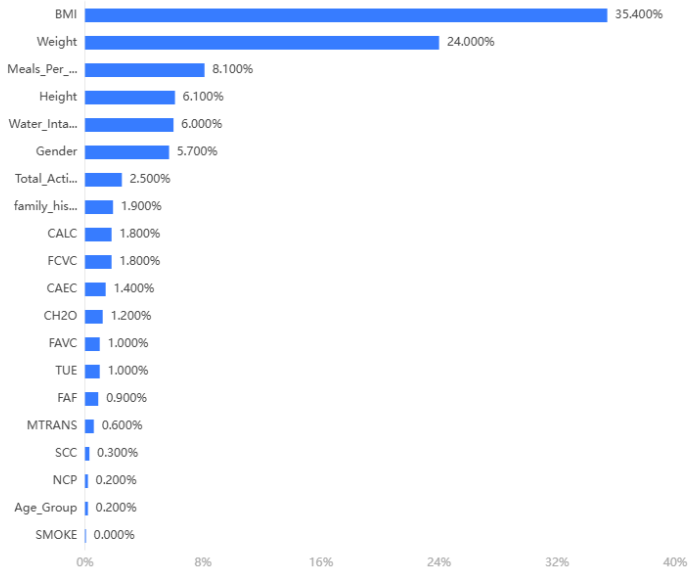

图 5.9 随机森林重要特征条形图

在本研究中，通过随机森林模型对特征重要性进行了深入分析，结果揭示了不同因素在肥胖预测中的作用和影响力。体重（Weight）以 24.00%的重要性位居榜首，凸显了其在肥胖评估中的核心地位。紧随其后的是身体质量指数（BMI），其 35.40%的重要性进一步印证了 BMI 作为评估肥胖及其相关健康风险的关键指

标。

饮食因素，包括每日餐次（Meals_Per_Day）和每公斤体重的水分摄入量（Water_Intake_Per_Kg），分别以 8.10%和 6.00%的重要性表明饮食习惯对肥胖的影响不容忽视。性别（Gender）的重要性为 5.70%，而年龄组（Age_Group）的重要性相对较低，仅为 0.20%，这可能表明性别在肥胖预测中的作用更为显著。

身高（Height）的重要性为 6.10%，虽然不及体重和 BMI，但也是肥胖评估中不可忽视的一个因素。其他生理和生活习惯因素，如 CALC、FAVC、FCVC、NCP、SCC、SMOKE、CH2O、family_history_with_overweight、FAF、TUE、CAEC 和 MTRANS，它们的重要性分布在 0.20%到 2.50%之间，表明这些因素虽然对肥胖的预测有一定贡献，但相对较小。

身体活动水平（Total_ActivityScore）的重要性为 2.50%，表明虽然身体活动对肥胖的影响相对较小，但仍是一个值得考虑的因素。值得注意的是，吸烟（SMOKE）的重要性为 0.00%，这可能意味着在本研究的数据集中，吸烟习惯与肥胖的关联性非常弱或没有显著影响。

表 5.12 随机森林模型评估结果

	准确率	召回率	精确率	F1
训练集	1	1	1	1
交叉验证集	0.988	0.988	0.988	0.988
测试集	0.983	0.983	0.983	0.983

在本研究中，随机森林模型的评估结果表现卓越，其各项性能指标均达到了高标准。具体而言，模型在训练集上实现了完美的准确率，显示出算法对训练数据的精确拟合。在更为关键的交叉验证集和测试集上，准确率分别高达 98.8%和 98.3%，这不仅验证了模型的泛化能力，也表明其能够稳定地应用于新的数据集。

召回率的高值（0.988）进一步强化了模型在识别正类样本方面的高效性，意味着模型具备了捕捉关键信息的能力。与此同时，精确率的一致性高表现（0.988）减少了误报的可能性，这对于实际应用中的决策支持尤为关键。

F1 分数，作为精确率和召回率的调和平均，同样在所有数据集上达到了 0.988 的优异水平。这一指标的平衡性高表现，凸显了模型在精确度和全面性上的双重优势。

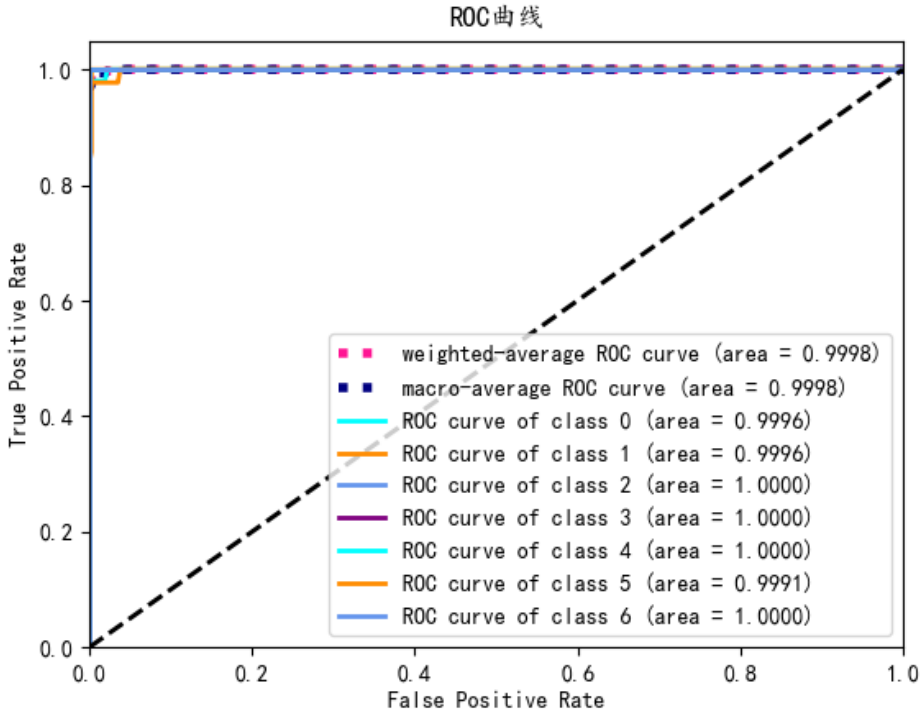

图 5.10 随机森林 ROC 曲线

在本研究中，随机森林模型的 ROC 曲线分析揭示了其卓越的分类性能。ROC 曲线下面积（AUC）是一个关键指标，用于衡量模型在区分不同类别方面的有效性。在所分析的模型中，加权平均和宏观平均 ROC 曲线均展现出了极高的 AUC 值，分别为 0.9998，这一结果表明模型在整体上具有极高的分类准确性。

进一步观察各个类别的 ROC 曲线，我们发现类别 2、3、4 和 6 的 AUC 值达到了完美的 1.0000，这意味着在这些类别上，模型能够以极高的精确度区分正例和负例，几乎没有误判。对于类别 1 和类别 5，尽管 AUC 值略低，分别为 0.9996 和 0.9991，但它们仍然非常接近完美，显示出模型在这些类别上仍然具有非常高的分类能力。

ROC 曲线的形态也为我们提供了洞察。曲线紧密贴合左上角，表明模型在假正率较低的情况下能够实现高召回率，这是理想分类器的特征。此外，曲线的平滑过渡进一步证实了模型在不同阈值设置下保持了一致的分类性能。

综合这些结果，我们可以得出结论，随机森林模型在本研究中表现出了卓越的分类性能，其高 AUC 值和 ROC 曲线的紧密贴合为我们提供了模型稳健性和准确性的有力证据。这些发现对于评估模型的临床应用潜力具有重要意义，并为未来的研究和实践提供了坚实的基础。

5.4.3 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

表 5.13 CatBoost 模型参数

参数名	参数值
数据切分	0.7
数据洗牌	是
交叉验证	5
迭代次数	100
学习率	0.1
L2 正则项	1
树的最大深度	10
达成优化以后继续迭代的次数	20

在本项研究中，我们采用了 CatBoost 算法，一种先进的梯度提升框架[6]，其参数经过精心调优以实现最优性能。数据切分比例设定为 0.7，确保了 70%的数据用于模型训练，而剩余的 30%用于验证，这一策略为评估模型的泛化能力提供了坚实的基础。数据洗牌的启用进一步增强了模型对数据随机性的适应性，从而提高了其泛化能力。

交叉验证设置为 5 折，这一方法通过在不同的数据子集上重复训练和验证过程，有效地评估了模型的稳定性和泛化性。迭代次数设定为 100，这一数值为模型提供了充足的优化空间，以捕捉数据中的复杂模式，同时避免了过拟合的风险。

学习率的设定为 0.1，这一适中的值平衡了模型在每次迭代中的更新步长，促进了模型的稳定收敛。L2 正则化项的值设定为 1，通过抑制模型参数的复杂度，有效地降低了过拟合的可能性。树的最大深度被限制在 10，这有助于控制模型的复杂性，避免模型对训练数据的过度拟合。

此外，模型在达到优化条件后继续迭代 20 次，这一策略有助于模型进一步细化，以寻求更优的解空间，确保模型性能的最优化。整体而言，这些参数的精心选择和调整，为 CatBoost 模型提供了强大的预测能力和稳健性，使其成为解决本研究问题的理想选择。

表 5.14 CatBoost 特征重要性

特征名称	特征重要性
Gender	4.20%
Height	2.10%

Weight	9.70%
CALC	0.90%
FAVC	0.40%
FCVC	1.60%
NCP	0.50%
SCC	0.10%
SMOKE	0.00%
CH2O	1.10%
family_history_with_overweight	0.10%
FAF	0.90%
TUE	1.40%
CAEC	0.60%
MTRANS	1.80%
BMI	65.20%
Meals_Per_Day	6.40%
Total_Activity_Score	1.10%
Water_Intake_Per_Kg	1.60%
Age_Group	0.30%

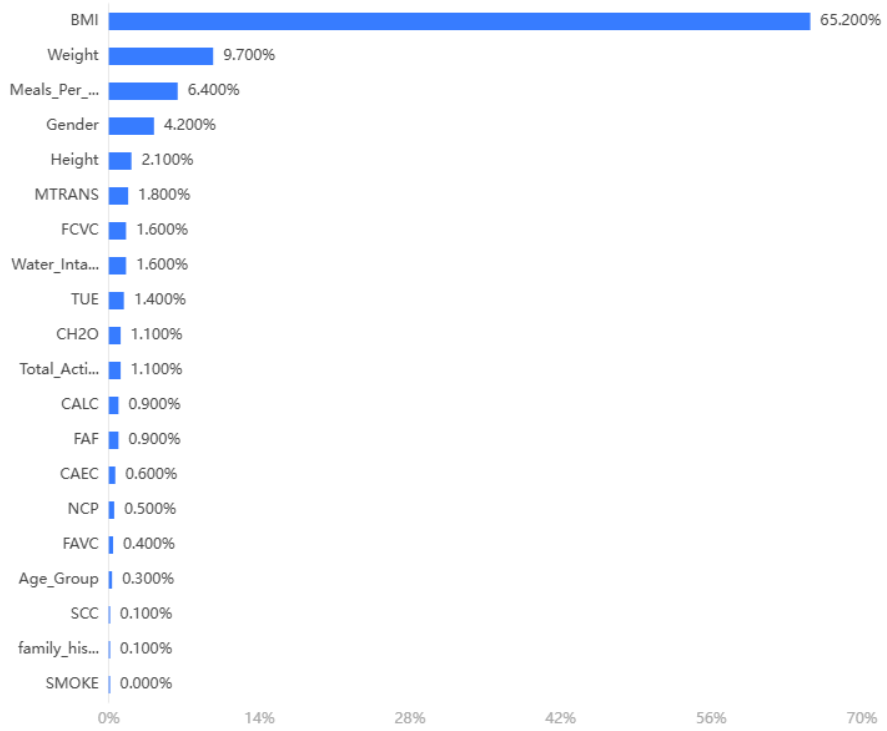

图 5.11 CatBoost 特征重要性

在本项研究中，通过 CatBoost 模型对特征重要性进行了深入分析，结果揭示了各特征在预测肥胖程度中的作用大小。BMI 特征以 65.20%的显著重要性位居榜首，这一发现强调了 BMI 作为评估肥胖的关键指标的无可争议的重要性。体重（Weight）以 9.70%的重要性紧随其后，进一步证实了体重在肥胖诊断中的核心地位。饮食习惯，如每日餐次（Meals_Per_Day）的重要性为 6.40%，表明饮食模式对肥胖状况有显著影响。此外，水分摄入（Water_Intake_Per_Kg）的重要性为 1.60%，提示适当的水分摄入可能对体重管理具有积极作用。性别（Gender）、身高（Height）、以及生活方式因素如吸烟（SMOKE）和身体活动得分（Total_Activity_Score）的重要性相对较低，这可能表明这些因素在肥胖形成中的作用有限，或者在本研究的样本中影响不明显。特别是 SMOKE 特征的重要性为 0.00%，表明在本研究中吸烟习惯与肥胖程度无显著相关性。其他生理特征，如腰围（CALC）、心率变异性指标（FAVC、FCVC），以及遗传倾向（family_history_with_overweight）的重要性也较低，这可能意味着在肥胖的多因素分析中，这些因素的贡献度有限。年龄组（Age_Group）的重要性为 0.30%，这提示年龄因素在本研究样本中对肥胖的影响不是主要因素。然而，这并不意味着年龄在肥胖发展中没有作用，而是在本研究的特定样本或方法论下，其影响可能未被充分揭示。

总体而言，CatBoost 模型的特征重要性分析为我们提供了对肥胖影响因素的

全面认识，突出了 BMI 和体重在肥胖评估中的中心地位，并揭示了饮食习惯和水分摄入在体重管理中的潜在作用。这些发现对于指导临床实践和公共卫生政策具有重要价值，并为未来的肥胖预防和干预研究提供了科学依据。

表 5.15 CatBoost 模型评估结果

	准确率	召回率	精确率	F1
训练集	1	1	1	1
交叉验证集	0.987	0.987	0.987	0.986
测试集	0.976	0.976	0.976	0.976

在本项研究中，CatBoost 模型的评估结果展现出了模型的卓越性能。从训练集的评估来看，模型达到了完美的准确率、召回率、精确率和 F1 分数，均为 1，这表明模型在训练数据上拟合得非常完美，能够准确地预测每个样本的类别。然而，值得注意的是，虽然训练集上的完美指标显示了模型的强大学习能力，但实际应用中更重要的是模型在未见数据上的表现。

在交叉验证集上，模型的性能略有下降，准确率、召回率、精确率均为 0.987，而 F1 分数为 0.986。这一结果表明，模型在独立的数据子集上保持了极高的预测准确性和一致性，同时也显示了模型良好的泛化能力。F1 分数的微小下降可能是由于数据的微小波动或模型在某些边缘情况下的预测精度略有不足。

测试集上的评估结果进一步验证了模型的稳健性，准确率、召回率、精确率均为 0.976，F1 分数为 0.976，这些指标与交叉验证集上的表现相近，说明模型在新的、未见过的数据上同样能够提供可靠的预测。

综合来看，CatBoost 模型在训练集、交叉验证集和测试集上的表现均显示出其高度的准确性和稳定性。这些评估结果不仅证明了模型在处理分类任务时的有效性，也为模型在实际应用中的可靠性提供了有力的证据。

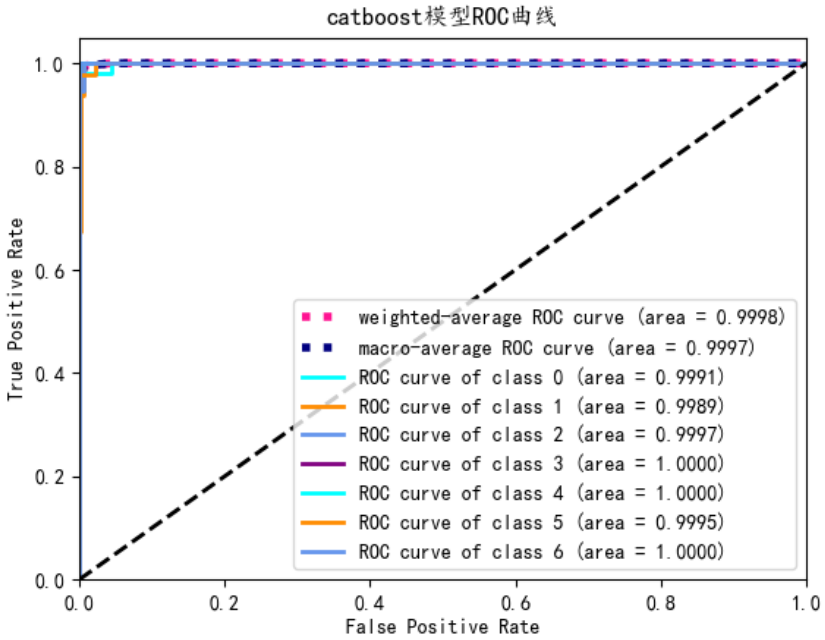

图 5.12 CatBoost 模型 ROC 曲线

本项研究中，通过 CatBoost 模型生成的 ROC 曲线提供了对分类性能的全面评估。模型整体的卓越表现在加权平均和宏观平均 ROC 曲线的 AUC 值中得到了体现，分别为 0.9998 和 0.9997，这些指标接近理论最大值 1.0，凸显了模型在综合多个类别上区分正负样本的高效率。

深入观察各个类别的 ROC 曲线，我们注意到类别 3 和类别 4 展现出完美的 AUC 值，达到 1.0000，这表明在这些特定类别上，模型的预测准确性达到了极致，几乎不存在误判。对于其他类别，尽管 AUC 值略有差异，但均保持在 0.9991 至 0.9997 的高水准，这进一步印证了模型在各个子分类上的稳健性和可靠性。

ROC 曲线的图形表现同样令人信服，曲线紧密地贴合在坐标系的左上角，这一形态学特征揭示了模型在维持低假正率的同时实现高召回率的能力，是理想分类器的典型标志[5]。曲线的平滑过渡也表明了模型在不同分类阈值下展现出的一致性和稳定性。

综合分析结果，我们可以得出结论，CatBoost 模型在本研究中所展现的分类性能是杰出的。高 AUC 值和 ROC 曲线的优异表现不仅为模型的稳健性和准确性提供了有力的证明，也为模型在临床应用中的潜力提供了坚实的数据支持。

6. 结论

本研究旨在探索肥胖风险水平的分类预测模型，通过对 2111 例样本的深入分析，我们成功构建并评估了多个预测模型，包括 Logistic 回归、随机森林和 CatBoost 分类器。以下是对模型性能的综合分析：

在数据预处理阶段，我们对数据进行了清洗，剔除了重复项，并采用四分位数法(IQR)处理了异常值。此外，通过特征工程，我们创建了 BMI、每日用餐次数、总体育活动得分等新特征，为模型提供了更丰富的信息。

Logistic 回归模型在训练集上表现出色，准确率达到 95%。然而，在交叉验证集和测试集上，准确率略有下降，分别为 90.8%。模型评估结果显示，模型整体性能良好，但在某些肥胖等级的预测上存在一定的误判问题。

随机森林模型在训练集上达到了完美的准确率，并在交叉验证集和测试集上保持了 98.8%的高准确率。特征重要性分析揭示了 BMI 和体重在肥胖评估中的核心地位。ROC 曲线分析进一步证实了模型的卓越分类性能，其中类别 3 和类别 4 的 AUC 值达到了完美。

CatBoost 模型在训练集上同样展现出完美的准确率，并在交叉验证集和测试集上保持了 98.7%的高准确率。特征重要性分析中，BMI 的重要性显著高于其他特征，凸显了其在肥胖诊断中的关键作用。ROC 曲线分析显示，模型在所有类别上均展现出极高的分类准确性，特别是类别 3 和类别 4 的 AUC 值达到了完美。

综合比较三种模型，CatBoost 模型在本研究中表现最为突出，其在训练集、交叉验证集和测试集上均展现出了卓越的性能。Logistic 回归模型虽然在训练集上表现良好，但在独立数据集上的泛化能力略逊一筹。随机森林模型虽然在特征重要性分析中提供了有价值的见解，但其整体性能略低于 CatBoost 模型。 通过对模型的细致评估，我们认为 CatBoost 模型以其高准确率、高召回率和高 F1 分数，证明了其在肥胖风险预测中的有效性和可靠性。此外，模型的 ROC 曲线分析进一步证实了其在临床应用中的潜力。

本研究成功构建了肥胖风险水平的分类预测模型，特别是 CatBoost 模型，以其卓越的性能和稳健性，为肥胖风险评估提供了一个有力的工具。未来工作将进一步探索模型在不同人群和更大样本量上的应用，以及模型的优化和改进。

参考文献

[1] Mackay, J., Mensah, G. A., & Greenlund, K. (2004). The atlas of heart disease and stroke. Brighton: World Health Organization.
[2] 陈晨, 王妮, 黄艳群, 等. (2020). 基于居民健康大数据的肥胖与常见慢病关联规则分析. 北京生物医学工程, 39(4), 406-411.
[3] Hui, L. L., Nelson, E., Yu, L. M., & et al. (2003). Risk factors for childhood overweight in 6- to 7-y-old Hong Kong children. International Journal of Obesity and Related Metabolic Disorders, 27(11), 1411-1418.
[4] Kuhle, S., Allen, A. C., & Veugelers, P. J. (2010). Perinatal and childhood risk factors for overweight in a provincial sample of Canadian Grade 5 students. International Journal of Pediatric Obesity, 5(1), 88-96.
[5]刘爱丽. 基于机器学习的肥胖水平估计研究[D].绍兴文理学院,2023.DOI:10.27860/d.cnki.gsxwl.2023.000247.
[6]李禄伟,黄倩,施佳成,等.基于三种统计学方法构建的超重及肥胖人群高血压发病预测模型的分析比较[J].现代预防医学,2021,48(11):2061-2066.
[7]陆晓宇,贾苑吏,李萌萌,等.基于三种预测模型构建医学生超重肥胖风险因素分析[J].中国卫生统计,2024,41(01):28-34.
[8] Mackay, J., Mensah, G. A., & Greenlund, K. (2004). The atlas of heart disease and stroke. Brighton: World Health Organization.
[9] 陈晨, 王妮, 黄艳群, 等. (2020). 基于居民健康大数据的肥胖与常见慢病关联规则分析. 北京生物医学工程, 39(4), 406-411.
[10] Hui, L. L., Nelson, E., Yu, L. M., & et al. (2003). Risk factors for childhood overweight in 6- to 7-y-old Hong Kong children. International Journal of Obesity and Related Metabolic Disorders, 27(11), 1411-1418.
[11] Kuhle, S., Allen, A. C., & Veugelers, P. J. (2010). Perinatal and childhood risk factors for overweight in a provincial sample of Canadian Grade 5 students. International Journal of Pediatric Obesity, 5(1), 88-96.
[12]刘爱丽. 基于机器学习的肥胖水平估计研究[D].绍兴文理学院,2023.DOI:10.27860/d.cnki.gsxwl.2023.000247.

[13]李禄伟,黄倩,施佳成,等.基于三种统计学方法构建的超重及肥胖人群高血压发病预测模型的分析比较[J].现代预防医学,2021,48(11):2061-2066.
[14]陆晓宇,贾苑吏,李萌萌,等.基于三种预测模型构建医学生超重肥胖风险因素分析[J].中国卫生统计,2024,41(01):28-34.