基于三种预测模型构建医学生超重肥胖风险因素分析*

陆晓宇[1]　贾苑吏[2]　李萌萌[1]　赵泽坤[1]　曹肖肖[1]　樊梦婷[1]　夏　鑫[1]　成　丽[1]　薛　玲[1,3△]

【提　要】 **目的**　构建 logistic 回归、随机森林和 SVM 模型预测医学生超重肥胖发生的影响因素，并对模型性能参数进行评价和比较，以获得超重肥胖风险评估预测的最优模型。**方法**　参与者为 2020 年 5-12 月来自河北省某市 1866 名医学生，通过自测问卷收集筛查其超重肥胖相关数据；利用 Python 分别构建 logistic 回归、随机森林和 SVM 三种风险评估模型。**结果**　logistic 回归、随机森林和 SVM 模型准确度分别为 96.26%、98.66% 和 98.13%；特异度分别为 99.77%、100% 和 99.08%；F_1 值分别为 0.85、0.95 和 0.93，随机森林为最优预测模型。随机森林模型结果显示，主观幸福感、负性事件以及学生经济状况在模型中预测权重值均超过 10%。**结论**　主观幸福感水平、负性事件次数以及学生经济状况等为影响医学生超重肥胖发生率的主要因素；随机森林模型的预测效果较 logistic 回归和 SVM 更优。

【关键词】 医学生　超重肥胖　logistic 回归　随机森林　支持向量机

【中图分类号】 R195.1　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.01.006

Study on the Risk Prediction Models of Overweight and Obesity in Medical Students

Lu Xiaoyu, Jia Yuanli, Li Mengmeng, et al (*School of Public Health, North China University of Science and Technology* (063000), *Tangshan*)

【**Abstract**】 **Objective**　To construct logistic regression, random forest and SVM models to predict the influencing factors of overweight and obesity in medical students, and the prediction performance of the three models was compared, so as to obtain the optimal model for the risk assessment of overweight and obesity. **Methods**　Participants included 1 866 medical students from a city in Hebei Province from May to December 2020. The relevant data of overweight and obesity screening were collected through self-test questionnaire; three models of logistic regression, random forest and SVM are constructed by python. **Results**　The test set showed that the accuracy of logistic regression, random forest and SVM models were 96.26%, 98.66% and 98.13% respectively; the specificity were 99.77%, 100% and 99.00%, respectively; and the AUC were 0.88, 0.99 and 0.88 respectively. Random forest is the optimal prediction model; according to the random forest model results, subjective well-being, negative events and students' economic status are more than 10% of weight in the model. **Conclusion**　Subjective well-being, negative events and students' economic status are the main factors affecting the incidence of overweight and obesity in medical students; the prediction performance of random forest model was better than logistic regression model and SVM model.

【**Key words**】 Medical students; Overweight and obesity; Logistic regression; Random forest; Support vector machine

1997 年，世界卫生组织将肥胖定义为一种疾病，即指在遗传、环境因素交互作用下，因能量摄入超过能量消耗，导致体内脂肪积聚过多，从而危害健康的一类慢性代谢性疾病[1]。当前，全球超过 19 亿成年人（占世界人口总数的 39%）超重，其中超过 6.5 亿人肥胖[2]，作为一种能量代谢紊乱，肥胖对机体多个器官系统造成不良影响，且是多种疾病的关键危险因素，如心血管疾病、糖尿病、高血压和一些癌症。超重肥胖患病率的急剧增加及其对一些非传染性疾病负担的影响，使超重肥胖成为了一种具有严重后果的全球公共卫生问题[3]，明确影响超重肥胖发生的因素，制定有效的预防和控制措施，对于提高人口质量和群体健康水平具有十分重要的意义。

以往研究结果证实，成年群体早期生活条件及行为特点对其后期肥胖症的发生发挥重要作用[4-6]。因此，许多研究涉及青年群体超重肥胖相关的因素，但鲜有将医学生作为独立研究对象进行分析。医学生往往因课业繁多，实习工作强度高而面临巨大的身心压力，超重肥胖发生率也可能增加，这一群体的健康状况更应引起重视。另外，以往预测超重肥胖影响因素大多使用回归模型[7-8]，回归模型方法简便易行，可解释性强，但是对数据要求较严格[9]，自变量间若存在多重共线性则会降低模型效果；随机森林[10]是以多个决策树为基础构成的分类器，能显著降低计算复杂度，训练速度较快，部分特征缺失时模型仍可维持较高准确度；支持向量机模型（support vector machine，SVM）[11]则具有使人容易理解的分类思想和较强的抗变换性，可以更好地解决小样本、非线性和高维问题。

本文拟采用 logistic 回归、随机森林和 SVM 构建医学生超重肥胖风险评估预测模型，并对模型性能进行比较，以构建医学生超重肥胖最优风险评估预测模型，精准预测高风险因素，以遏制医学生群体超重肥胖的流行。

* 基金项目：河北省民生科技专项项目（20377718D）

1.华北理工大学公共卫生学院（063000）
2.华北理工大学理学院
3.河北省煤矿卫生与安全重点实验室
△通信作者：薛玲，E-mail：xueling_heuu@163.com

对象和方法

1.研究对象

2020 年 5-12 月，采用多阶段抽样的方法对中国河北省某医学院校本科生进行横断面调查。首先从该校本科专业中随机选择六个专业，再在选定专业的 1~5 年级本科生中，每一年级随机选取 2~4 个班级，最后对选中班级内所有在册学生进行调查。

所有被调查者参与调查前均自愿签署知情同意书。本次调查共发放调查问卷 1880 份，回收有效问卷 1866 份，问卷有效回收率为 98.20%。

2.研究方法

(1)研究内容

采用自行设计的问卷进行调查，问卷由三部分组成：一般人口特征，包括年龄、性别、学生经济状况等；生活行为习惯，包括不良饮食习惯、吸烟、饮酒和上网时间等；家庭及学校环境，包括人际交往压力、学习压力、负性事件发生次数等。

主观幸福感测量采用清华大学李焰教授研发的大学生幸福感影响因素问卷对大学生的主观幸福感进行调查。该问卷涉及 40 个条目，共 7 维度，即包括自我意识、学校学习、人际交往、恋爱、工作与社会适应、家庭环境、休闲活动。评定与记分方法：基于问卷中列出的生活事件，根据被调查对象的感受，将其幸福感受按频率分为五个等级，“1”代表从来没有，“2”代表偶尔；“3”代表有时；“4”代表经常；“5”代表总能感受到幸福。被调查对象测试条目得分相加，分数越高，表明其幸福感越强。要求调查对象在 30 分钟内完成问卷涉及的全部内容。

(2)BMI 的测量标准

本研究以体重指数 BMI 作为超重肥胖的评估指标，BMI 与体内脂肪总量密切相关，该指标考虑了体重和身高两个因素。

通过测量标准对被调查对象进行体重和身高的测量，超重肥胖的界定标准参照《中国成人超重和肥胖症预防控制指南》[12]：成年人的 BMI 为 24.0~27.9 kg/m^2 被视为超重，28 kg/m^2 及以上被视为肥胖。

3.统计方法

(1)模型输入变量筛选

应用 PPCs-DEMATEL 法[13]，综合非条件 logistic 回归分析结果筛选模型的输入变量。

其中，皮尔逊相关系数(Pearson correlation coefficient，PPCs)可定量衡量变量间相关程度，标准值在-1 到 1 之间。决策试验评估实验室法[14](decision making trialand evaluation laboratory，DEMATEL)则是运用图论以及矩阵工具对系统要素进行定性分析，以判断要素间强弱的因果关系。PPCs-DEMATEL 法即将定量分析得到的相关系数和定性分析得到的中心度以 6∶4 的比例加和，并选择影响程度综合排名高于取值范围中位数的特征因素作为 PPCs-DEMATEL 法的模型输入变量筛选结果。

(2)模型构建与评价

通过 Python 分别构建 logistic 回归、随机森林和 SVM 模型预测医学生超重肥胖影响因素。模型开发过程中，采用 5 倍交叉验证，将样本数据分成 5 个大小相等的折叠，其中一个折叠用作测试集，其余作为训练集；再根据模型输入变量筛选结果选择相应变量，运用 sklearn 库中相应的函数构建模型，对训练集进行训练后引入测试集数据，最后使用 matplotlib 库分别对三种模型进行可视化，并通过混淆矩阵计算 AUC、准确性、敏感性和特异性等性能度量指标来验证三种模型性能。

(3)统计分析

采用 EpiData 3.0 创建数据库，采用 SPSS 25.0、R(3.6.3)和 Python 软件进行相应的作图和统计分析，用卡方检验和 t 检验/Wilcoxon 检验来比较组间分类变量和连续变量之间的差异。采用单因素和多因素 logistic 回归分析来筛选影响医学生超重肥胖的因素。P<0.05 被认为具有显著的统计学差异。

结　果

1.一般情况

1866 例调查对象中，超重肥胖发生率达 11.90%，其中男生 655 人，超重肥胖发生率为 18.63%；女生 1211 人，超重肥胖发生率 8.26%；不规律用餐、不运动、经常饮酒以及熬夜的医学生超重肥胖检出率较高(P≤0.05)。见表 1。

2.输入变量筛选

以是否超重肥胖为因变量(0=否，1=是)，将表 1 中单因素分析有意义的 15 个特征变量进行多因素 logistic 回归分析，结果显示，与医学生超重肥胖有关的因素有 12 个(P<0.05)，见表 3；PPCs-DEMATEL 法取值范围中位数为 1.273，即综合影响程度高于1.273 的 10 个特征因素可作为预测模型的输入变量，见表 4。

综合两种方法分析结果，最终选入 14 个特征因素作为构建三种预测模型的输入变量，分别是性别、年级、父母期望值、学生经济状况、熬夜、规律用餐、饮酒、运动频率、上网时间、主观幸福感、负性事件、学习压力、人际交往压力、心理学讲座。14 个构建模型所需输入变量的单因素与多因素的 logistic 分析结果，如表 5。

表 1　医学生特征因素超重肥胖检出情况及单因素分析结果

变量	正常及以下（n=1644）	超重肥胖（n=222）	χ^2/t 值	P
性别[n(%)]			43.597	<0.001
男	533(32.42)	122(54.95)		
女	1111(67.58)	100(45.05)		
民族[n(%)]			0.796	0.372
汉族	1616(98.30)	220(99.10)		
其他	28(1.70)	2(0.90)		
年级[n(%)]			21.607	<0.001
一	341(20.74)	21(9.46)		
二	307(18.67)	56(25.23)		
三	469(28.53)	80(36.04)		
四	333(20.26)	39(17.57)		
五	194(11.80)	26(11.70)		
父母离异[n(%)]			1.797	0.180
否	1523(92.64)	200(90.09)		
是	121(7.36)	22(9.91)		
父母期望值[n(%)]			25.195	<0.001
低	57(3.47)	20(9.00)		
一般	436(26.52)	34(15.32)		
高	1151(70.01)	168(75.68)		
学生经济状况(千元,$\bar{x}\pm s$)	5.0±1.18	4.5±1.33	5.757	<0.001
医疗负担[n(%)]			10.729	0.005
较轻	541(32.91)	50(22.52)		
一般	905(55.05)	136(61.26)		
较重	198(12.04)	36(16.22)		
规律用餐[n(%)]			57.271	<0.001
规律	523(31.82)	23(10.36)		
一般	402(24.45)	47(21.17)		
不规律	719(43.73)	152(68.47)		
不良饮食行为[n(%)]			0.270	0.604
否	1022(62.17)	142(63.96)		
是	622(37.83)	80(36.04)		
吸烟[n(%)]			3.624	0.163
从不	1272(77.37)	159(71.62)		
有时	276(16.79)	47(21.17)		
每天	96(5.84)	16(7.21)		
饮酒[n(%)]			112.290	<0.001
从不	804(48.91)	25(11.26)		
有时	745(45.32)	174(78.38)		
每天	95(5.77)	23(10.36)		
熬夜[n(%)]			16.104	<0.001
从不	368(22.38)	29(13.07)		
有时	692(42.09)	122(54.95)		
每天	584(35.53)	71(31.98)		
上网时间(时,$\bar{x}\pm s$)	20.83±20.25	16.40±13.25	3.170	0.002
运动频率[次,n(%)]			101.410	<0.001
0	282(17.15)	78(35.14)		
1~2	333(20.26)	82(36.94)		
3~4	669(40.69)	51(22.97)		
≥5	360(21.90)	11(4.95)		
主观幸福感(得分,$\bar{x}\pm s$)	125.38±21.34	126.68±13.88	−0.890	0.374
负性事件(次,$\bar{x}\pm s$)	0.72±1.09	1.35±1.10	−8.121	<0.001
学业压力(次,$\bar{x}\pm s$)	5.66±1.22	6.05±1.05	−4.548	<0.001
人际交往(次,$\bar{x}\pm s$)	6.28±1.20	6.54±1.62	−2.821	0.005
生活满意度[n(%)]			21.478	<0.001
不满意	20(1.22)	2(0.90)		
不太满意	711(43.25)	72(32.43)		
比较满意	854(51.95)	148(66.67)		
满意	59(3.59)	0		
心理学讲座[n(%)]			51.221	<0.001
从不	108(6.57)	42(18.92)		
偶尔	660(40.15)	53(23.87)		
经常	876(53.28)	127(57.21)		

表 2　多因素 logistic 回归分析变量赋值表

变量名	变量含义	赋值方式
Y	超重肥胖	1=正常及以下,2=超重肥胖
X_1	性别	1=男,2=女
X_2	年级	1=大一,2=大二,3=大三,4=大四,5=大五
X_3	父母期望值	1=低,2=一般,3=高
X_4	学生经济状况	实测值
X_5	医疗负担	1=较轻,2=一般,3=较重
X_6	规律用餐	1=规律,2=一般,3=不规律
X_7	饮酒	1=从不,2=有时,3=每天
X_8	熬夜	1=从不,2=有时,3=每天
X_9	上网时间	实测值
X_{10}	运动频率	1=从不,2=1~2 次,3=3~4 次,4=5 次及以上
X_{11}	负性事件	实测值
X_{12}	学业压力	实测值
X_{13}	人际交往	实测值
X_{14}	生活满意度	1=不满意,2=不太满意,3=比较满意和满意
X_{15}	心理学讲座	1=从不,2=偶尔,3=经常

3.三种风险评估模型的建立及性能比较

采用 5 倍交叉法随机划分 1866 名调查对象的样本数据集,按照训练集和测试集 4∶1 的比例划分,训练集与测试集的样本例数分别为 1492 例与 374 例。根据表 6 结果显示,训练集和测试集的基线资料比较差异无统计学意义(P>0.05),即两组病例资料的基线特征一致,具有可比性。以超重肥胖为目标变量,14 个特征因素作为模型输入变量,分别构建 logistic 回归、随机森林和 SVM 三种模型,其中,随机森林模型准确率、特异度、F_1 分数以及 AUC 等多项参数值高于其他两种(表 7,图 1~2)。

4.重要变量列线图分析

选择预测能力最好的随机森林模型,筛选预测重要变量。对随机森林模型的输入变量按重要性由高到低排序,权重值排名前五的输入变量依次为:主观幸福感、负性事件、学生经济状况、上网时间、人际交往压力,详见图 3。

针对 5 个筛选变量绘制列线图,每个变量的对应线段刻度为该变量的取值范围,线段长度为该因素对医学生超重肥胖检出率的影响大小,图 4 结果显示,五个被选变量对医学生超重肥胖的影响程度从高到低分别为:负性事件、人际交往压力、每周上网时间、学生经济状况、主观幸福感。

讨　论

结果显示,本研究医学生超重肥胖发生率为 11.90%,超过以往调查研究中的普通大学生超重肥胖发生率(9.50%)[15],风险评估预测模型结果表明:人际交往压力、学生经济状况、每周上网时间、主观幸福感、负性事件发生次数、规律饮食以及熬夜等多个方面

表 3 医学生超重肥胖影响因素的多因素 logistic 回归分析

特征	β	$S.E$	Wald χ^2	P	OR	95% CI	
						下限	上限
性别	0.079	0.300	0.069	0.793	1.082	0.601	1.949
年级(一年级)			21.715	<0.001			
二年级	0.755	0.409	3.398	0.065	2.127	0.953	4.744
三年级	1.578	0.373	17.898	<0.001	4.847	2.333	10.070
四年级	0.435	0.460	0.895	0.344	1.545	0.627	3.806
五年级	0.758	0.383	3.913	0.048	2.133	1.007	4.520
父母期望值(低)			6.298	0.043			
中	−1.389	0.561	6.138	0.013	0.249	0.083	0.748
高	−1.147	0.508	5.097	0.024	0.318	0.117	0.860
学生经济状况	−0.847	0.137	38.408	<0.001	0.429	0.328	0.560
医疗负担(较轻)			0.549	0.760			
一般	−0.143	0.314	0.208	0.649	0.867	0.468	1.605
较重	−0.312	0.420	0.549	0.459	0.732	0.321	1.669
规律用餐(规律)			61.638	<0.001			
一般	0.935	0.427	4.786	0.029	2.547	1.102	5.885
不规律	2.942	0.412	51.012	<0.001	18.959	8.456	42.508
饮酒(从不)			70.670	<0.001			
有时	2.999	0.364	67.881	<0.001	20.070	9.833	40.964
每天	3.452	0.628	30.244	<0.001	31.555	9.222	107.978
熬夜(从不)			9.132	0.010			
有时	1.043	0.403	6.689	0.010	2.837	1.287	6.251
每天	1.389	0.460	9.108	0.003	4.010	1.627	9.884
运动频率(从不)			36.490	<0.001			
1~2 次/周	1.153	0.362	10.138	0.001	3.167	1.558	6.438
3~4 次/周	−0.228	0.333	0.469	0.493	0.796	0.414	1.530
5 次及以上/周	−1.483	0.544	7.424	0.006	0.227	0.078	0.660
每周上网时间	−0.016	0.007	5.634	0.018	0.984	0.971	0.997
负性事件	0.733	0.120	37.463	<0.001	2.082	1.646	2.632
学习压力	0.224	0.112	3.985	0.046	1.251	1.004	1.560
人际交往	0.518	0.097	28.337	<0.001	1.678	1.387	2.031
生活满意度(不满意)			3.479	0.176			
不太满意	−1.269	1.216	1.088	0.297	0.281	0.026	3.050
比较满意和满意	−0.709	1.229	0.333	0.564	0.492	0.044	5.471
心理学讲座(从不)			42.657	<0.001			
偶尔	−2.787	0.445	39.167	<0.001	0.062	0.026	0.147
经常	−1.649	0.420	15.376	<0.001	0.192	0.084	0.438

*:( )内为参照组。

表 4 PPCs-DEMATEL 法特征变量筛选结果排序

序号	指标	Pearson 系数	DEMATEL 中心度	影响程度
1	主观幸福感	0.866	2.211	1.404
2	熬夜	0.793	2.211	1.360
3	性别	0.746	2.209	1.331
4	学生经济状况	0.736	2.203	1.323
5	上网时间	0.705	2.198	1.302
6	学业压力	0.709	2.194	1.303
7	规律用餐	0.676	2.193	1.283
8	年级	0.671	2.191	1.279
9	运动频率	0.670	2.182	1.275
10	负性事件	0.666	2.177	1.271
11	人际交往压力	0.680	2.170	1.276
12	父母期望值	0.670	2.166	1.269
13	心理学讲座	0.649	2.164	1.255
14	饮酒	0.648	2.158	1.252
15	生活满意度	0.647	2.158	1.251
16	医疗负担	0.644	2.124	1.236
17	父母离异	0.639	2.088	1.219
18	民族	0.635	2.076	1.211
19	不良饮食行为	0.634	2.024	1.190
20	吸烟	0.633	1.948	1.159

表 5　模型输入变量的单因素及多因素 logistic 回归分析

特征	单因素分析		多因素分析	
	OR(95% *CI*)	*P*	*OR*(95% *CI*)	*P*
性别	0.393(0.296~0.522)	<0.001	1.020(0.575~1.807)	0.947
年级(一年级)	1.000		1.000	
二年级	0.460(0.252~0.838)	0.011	0.548(0.262~1.143)	0.109
三年级	1.361(0.827~2.241)	0.226	0.975(0.470~2.020)	0.945
四年级	1.273(0.793~2.043)	0.318	2.051(1.090~3.861)	0.026
五年级	0.874(0.516~1.480)	0.616	0.617(0.275~1.384)	0.242
父母期望(低)	1.000		1.000	
中	2.404(1.409~4.102)	0.001	3.705(1.385~9.916)	0.009
高	0.534(0.364~0.785)	0.001	0.623(0.320~1.210)	0.162
学生经济状况	0.729(0.654~0.814)	<0.001	0.463(0.370~0.579)	<0.001
熬夜(从不)	1.000		1.000	
有时	0.648(0.413~1.018)	0.06	0.174(0.083~0.365)	<0.001
每天	1.450(1.061~1.982)	0.02	0.700(0.429~1.144)	0.155
规律用餐	1.000		1.000	
偶尔规律用餐	2.659(1.588~4.451)	<0.001	0.057(0.027~0.119)	<0.001
不规律用餐	4.807(3.057~7.560)	<0.001	0.156(0.082~0.298)	<0.001
饮酒(从不)	1.000		1.000	
有时	0.128(0.070~0.235)	<0.001	0.045(0.015~0.137)	<0.001
每天	0.965(0.594~1.566)	0.884	0.913(0.336~2.481)	0.858
运动频率(从不)	1.000		1.000	
1~2 次/周	9.052(4.725~17.343)	<0.001	2.974(1.164~7.600)	0.023
3~4 次/周	8.059(4.221~15.387)	<0.001	10.509(4.288~25.753)	<0.001
5 次及以上/周	2.495(1.284~4.847)	0.007	2.647(1.094~6.408)	0.031
每周上网时间	0.986(0.977~0.995)	0.002	0.980(0.968~0.993)	0.003
主观幸福感	1.003(0.996~1.010)	0.373	1.016(1.004~1.029)	0.011
负性事件次数	1.530(1.372~1.705)	<0.001	2.081(1.654~2.619)	<0.001
学习压力	1.293(1.156~1.447)	<0.001	1.189(0.959~1.474)	0.114
人际交往	1.172(1.049~1.309)	0.005	1.709(1.423~2.052)	<0.001
心理学讲座(从不)	1.000		1.000	
偶尔	2.682(1.794~4.010)	<0.001	5.782(2.668~12.530)	<0.001
经常	0.554(0.396~0.775)	0.001	0.280(0.164~0.478)	<0.001

表 6　训练集与测试集基线描述差异比较

项目	训练集 (n=1492)	测试集 (n=374)	*P*	项目	训练集 (n=1492)	测试集 (n=374)	*P*
性别[n(%)]				熬夜[n(%)]			
男	517(34.7)	138(36.9)	0.451	从不	309(20.7)	88(23.5)	0.491
女	975(65.3)	236(63.1)		有时	655(43.9)	159(42.5)	
年级[n(%)]				每天	528(35.4)	127(34.0)	
一年级	294(19.7)	68(18.2)	0.174	运动频率[n(%)]			
二年级	280(18.8)	83(22.2)		从不	283(19.0)	77(20.5)	0.373
三年级	453(30.4)	96(25.7)		1~2 次/周	337(22.6)	78(20.9)	
四年级	287(19.2)	85(22.7)		3~4 次/周	585(39.2)	135(36.1)	
五年级	178(11.9)	42(11.2)		5 次及以上/周	287(19.2)	84(22.5)	
父母期望值[n(%)]				心理学讲座[n(%)]			
低	63(4.2)	14(3.7)	0.759	从不	127(8.5)	23(6.1)	0.076
一般	380(25.5)	90(24.1)		偶尔	575(38.5)	138(36.9)	
高	1049(70.3)	270(72.2)		经常	790(53.0)	213(57.0)	
规律用餐[n(%)]				学生经济状况[$M(P_{25},P_{75})$]	5(4, 6)	5(4, 6)	0.863
规律	454(30.4)	92(24.6)	0.072	每周上网时间[$M(P_{25},P_{75})$]	15(8, 25)	15(10, 27)	0.063
一般	349(23.4)	100(26.7)		负性事件次数($\bar{x}\pm s$)	0.79±1.11	0.82±1.08	0.603
不规律	689(46.2)	182(48.7)		学业压力[$M(P_{25},P_{75})$]	6(5, 7)	5(5, 6)	0.024
饮酒[n(%)]				人际交往压力[$M(P_{25},P_{75})$]	6(6, 7)	6(5, 7)	0.931
从不	663(44.4)	166(44.4)	0.522	主观幸福感[$M(P_{25},P_{75})$]	124(113, 140)	124(112, 137)	0.297
有时	730(49.0)	189(50.5)					
每天	99(6.6)	19(5.1)					

表 7　三种风险评估模型性能参数比较

指标	logistic 回归模型		随机森林模型		支持向量机模型		指标	logistic 回归模型		随机森林模型		支持向量机模型	
	训练集	测试集	训练集	测试集	训练集	测试集		训练集	测试集	训练集	测试集	训练集	测试集
分类正确数(例)	1417	360	1476	369	1466	367	特异度(%)	97.80	98.77	99.47	100.00	99.09	99.08
分类错误数(例)	75	14	16	5	26	7	约登指数	0.71	0.78	0.94	0.90	0.91	0.91
准确率(%)	94.97	96.26	98.93	98.66	98.26	98.13	F_1 分数	0.77	0.85	0.95	0.95	0.92	0.93
灵敏度(%)	73.41	79.60	94.80	89.80	91.91	91.84	AUC	0.8890	0.8775	0.9994	0.9991	0.8780	0.8733

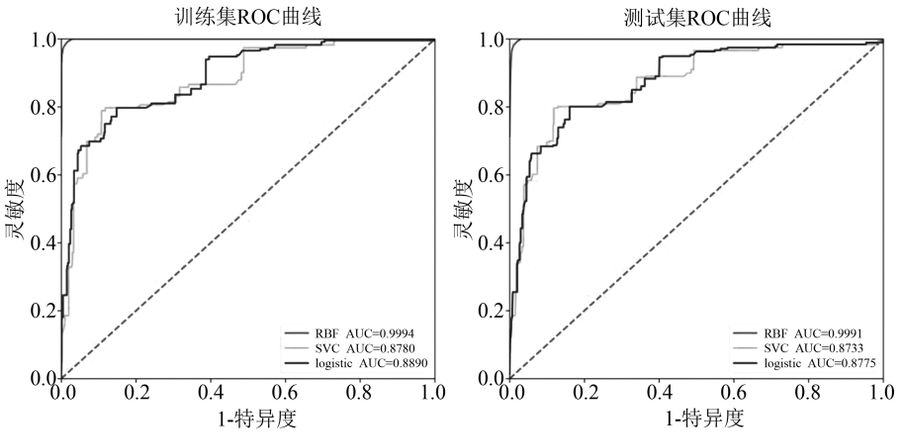

图 1 三种风险评估模型 ROC 曲线图

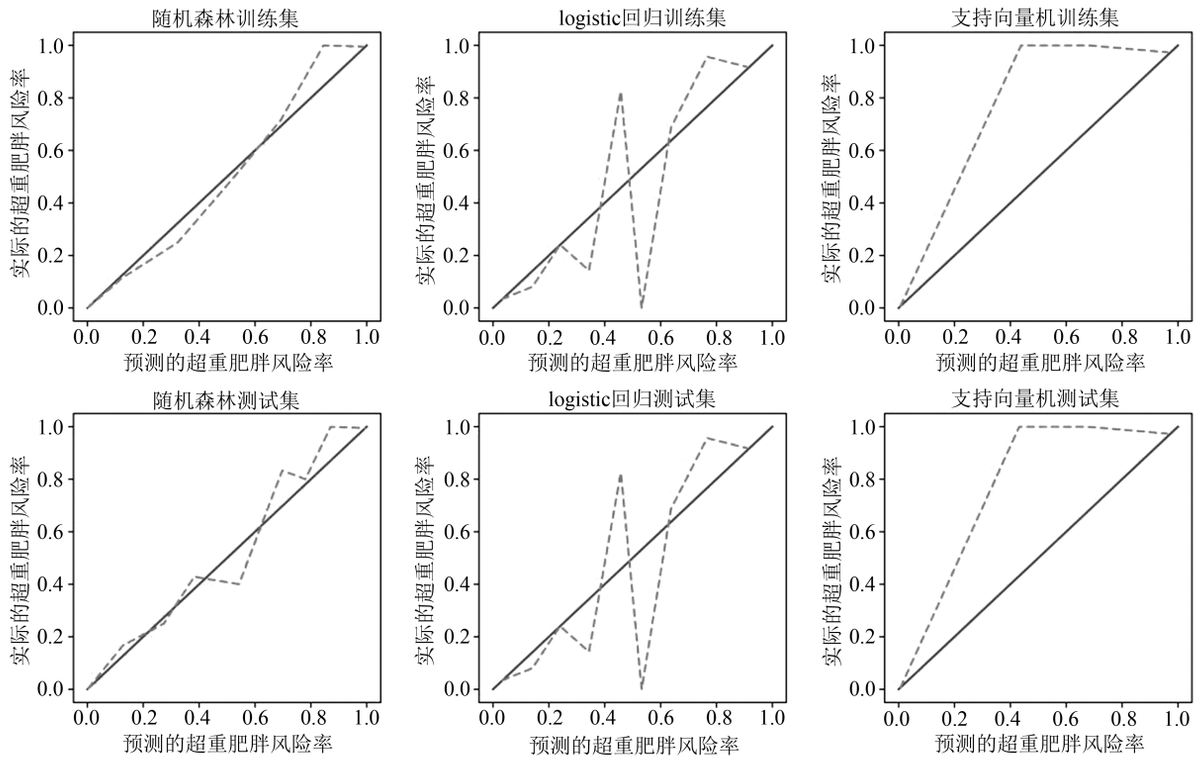

图 2 随机森林模型特征变量重要性排序

影响学生超重肥胖的发生情况，这与大多数研究结果一致[16-19]。本研究调查对象为医学生，鉴于这一特点，假设他们的健康知识优于一般人群，医学生群体可能不太容易发生超重肥胖。但是，本研究群体超重肥胖率仍高达11.90%。也许是因为医学生群体相较于其他同龄人，学习工作任务更重，尤其是高年级医学生，同时面临就业、学业、人际交往等各方面压力，繁琐的工作和学习任务增加医学生群体的久坐时间和熬夜机会、减少其规律用餐次数和时间。在一天紧张的高压下学习和工作的之后，因无法进食或没有足够时间进食，他们可能会长时间处于低能量状态，更需要补充热量，或者通过吸烟、饮酒来缓解身体饥饿感和心理压力，日积月累最终导致超重肥胖的发生。

本研究以影响医学生超重肥胖的多种因素作为前提条件，借助各种分类算法和计算机软件建立相应模型[20-22]，预测其超重肥胖发生的主要影响因素，通过对超重肥胖风险评估预测模型性能参数分析比较发现，随机森林模型的校准曲线及大部分参数性能优于其他两种。在灵敏度、约登指数方面，随机森林模型略低于 SVM，高于 logistic 回归模型，原因可能是：模型

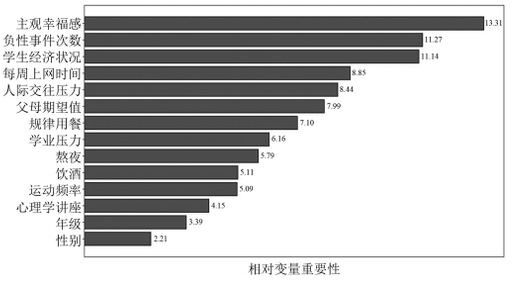

图 3 随机森林模型特征变量重要性排序

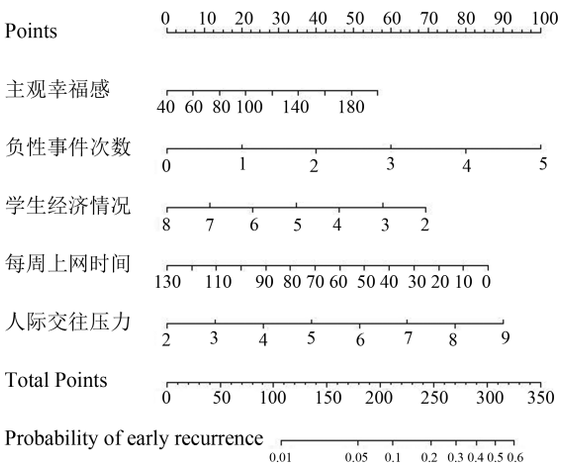

图 4　随机森林模型筛选的预测变量绘制的列线图

构建有一定的随机性，参数设置、训练次数及网络情况等都将影响模型表现。随机森林模型[23]内部结构较简单，作为近年来新兴的机器学习算法，是由很多个决策树组合成的森林，随机森林基于组合算法中的 Bagging 算法，对数据进行抽样，在生成决策树的时候引入随机变量，通过随机采样的方法，使其泛化能力增强，以解决决策树算法常出现的过拟合现象，具有高度灵活性，同时随机森林继承了决策树的大部分优点[24]，尤其对处理缺失值和噪声数据方面效果显著，在模式使用前不需做大量的数据预处理、每次训练完成后可以给出各个输入变量的预测权重，为疾病预防控制措施提供相应理论基础，更适用于医学研究。

参　考　文　献

[1] Woodhall-Melnik J, Misir V, Kaufman-Shriqui V, et al. The Impact of a 24 Month Housing First Intervention on Participants' Body Mass Index and Waist Circumference: Results from the At Home/Chez Soi Toronto Site Randomized Controlled Trial. PLoS One, 2015, 10(9):e0137069.

[2] Bessell E, Markovic TP, Fuller NR, et al. How to provide a structured clinical assessment of a patient with overweight or obesity. Diabetes Obes Metab, 2021, 23:36-49.

[3] López-Suárez A. Burden of cancer attributable to obesity, type 2 diabetes and associated risk factors. Metabolism, 2019, 92:136-146.

[4] Siddiqui MZ, Donato R. Overweight and obesity in India: policy issues from an exploratory multi-level analysis. Health Policy Plan, 2016, 31(5):582-591.

[5] Jiang SH, Peng SH, Yang TZ, et al. Overweight and Obesity Among Chinese College Students: An Exploration of Gender as Related to External Environmental Influences. Am J Mens Health, 2018, 12(4):926-934.

[6] Nau C, Schwartz BS, Bandeen-Roche K, et al. Community Socioeconomic Deprivation and Obesity Trajectories in Children Using Electronic Health Records. Obesity, 2015, 23(1):207-212.

[7] Al-Salameh A, Lanoix JP , Bennis Y, et al. The association between body mass index class and coronavirus disease 2019 outcomes. Int J Obes, 2021, 45(3):700-705.

[8] Bonanno L, Metro D, Papa M, et al. Assessment of sleep and obesity in adults and children Observational study. Medicine, 2019, 98(46):e17642.

[9] 张梅, 王丽敏, 陈志华, 等. 2013 年中国不同区域成人高胆固醇血症流行水平及相关因素分析.中华预防医学杂志, 2018, 52(2):151-157.

[10] Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. J Diabetes Sci Technol, 2018, 12(2):295-302.

[11] Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine. Front Microbiol, 2018, 9:476.

[12] 孟焕, 邱琳, 飒日娜, 等. 陕西省成人血脂异常流行现状及相关因素研究.中国慢性病预防与控制, 2021, 29(10):750-755.

[13] 崔惠敏, 薛惠锋, 王磊, 等. 基于 PCCs-DEMATEL 指标筛选的 BP 神经网络用水量预测.节水灌溉, 2019,(5):87-91+98.

[14] Kou G, Ergu D, Shang J. Enhancing data consistency in decision matrix: Adapting Hadamard model to mitigate judgment contradiction. Eur J Oper Res, 2014, 236(1):261-271.

[15] Yang TZ, Yu LW, Barnett R, et al. Contextual influences affecting patterns of overweight and obesity among university students: a 50universities population-based study in China. Int J Health Geogr, 2017, 16(1):18.

[16] Lampure A, Castetbon K, Hanafi M, et al. Relative Influence of Socioeconomic, Psychological and Sensory Characteristics, Physical Activity and Diet on 5-Year Weight Gain in French Adults. Nutrients, 2017. 9(11):1179.

[17] Tran DM, Dingley C, Arenas R. Perception and Beliefs Regarding Cardiovascular Risk Factors and Lifestyle Modifications Among High-Risk College Students. Can J Nurs Res, 2021, 53(2):94-106.

[18] 何春刚. 高校大学生体力活动与视屏时间交互作用对超重肥胖的影响.中国学校卫生, 2018, 39(12):1873-1876.

[19] 张玲玲, 熊家豪, 王纪川, 等.长沙市大学生外卖食品消费现状及其与超重肥胖的关联.中华疾病控制杂志, 2020, 24(9):1027-1031.

[20] 陈梦凡, 钱婷婷, 周梦林, 等. 基于人口学及临床特征的妊娠期糖尿病预测模型的研究.实用妇产科杂志, 2019, 35(2):117-122.

[21] 叶美华, 陈万远, 蔡博君,等. 基于卷积神经网络的甲状腺液基细胞学病理辅助诊断模型的研究.中华病理学杂志, 2021, 50(4):358-362.

[22] 张占林, 姚华, 孙勇, 等. 随机森林算法对体检人群糖尿病患病风险的预测价值研究.中国全科医学, 2019, 22(9):1021-1026.

[23] Dagliati A, Marini S, Sacchi L, et al. Machine learning methods to predict diabetes complications. J Diabetes Sci Technol, 2018, 12(2):295-302.

[24] Luo Y, Li Z, Guo H, et al. Predicting congenital heart defects: A comparison of three data mining methods. PLoS One,2017, 12(5):e0177811.

（责任编辑：张　悦）