

所属类别	2023 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科组		CM2305618

基于机器学习与集成学习模型探究母亲身心健康对婴儿成长的影响

摘要

针对问题一，我们使用 390 编号前 1 至 3 个月婴儿及其母亲的相关数据，探讨了母亲身体与心理指标对婴儿的行为特征与睡眠质量的影响。描述性统计直观地描述了母亲年龄、教育程度、孕期、夜间睡眠和模式的数据分布特征。通过 **ONE-HOT 编码** 系统地处理婴儿行为特征和入睡方式，使用**随机森林模型**评估母亲的身体指标和心理指标与婴儿的睡眠质量之间的关系。揭示了产妇心理健康与新生儿睡眠之间的相关性，尤其是**抑郁和焦虑症状对睡眠质量产生的消极影响**。产妇教育程度和婴儿睡眠模式也与睡眠质量有关，凸显产妇心理健康对婴儿发育的重要性。

针对问题二，整个过程包括数据预处理、复杂的特征工程以及细致的模型构建和评估。目的是根据给定特征预测目标变量。首先进行数据加载和验证，然后进行初步评估。预处理包括**分类变量编码、归一化、处理缺失数据和数字特征对齐**。特征工程转换数值属性，以便更好地对齐模型。模型构建涉及各种基本模型和高级模型。其中，为达最佳性能对随机森林超参数进行了调整，**逻辑回归**是最准确的模型，准确率约为 63.48%，另外图表等可视化助于深入了解数据和模型性能。

针对第三问，我们首先建立了关于 CBTS,EPDS,HADS 的变化与治疗费用的**线性回归模型**，划分测试集，代入模型求解将矛盾型转化为中等型治疗费用的最小值，求得为 **1.85 万元**，再从中等型转化为安静型，我们首选求得花费的最小值为 **1.94 万元**，此时对应的数值是 CBTS 得分为 17，HADS 得分调整为 15，EPDS 不变，对线性回归模型进行评价，求得模型的准确率，精确率，召回率分别为 **0.98, 0.976, 0.95**，模型效果较好，做出柱状图进行可视化。

针对问题四，我们对婴儿的睡眠质量进行综合评判，首先使用**熵权法**确定婴儿整晚睡眠时间、睡醒次数、入睡方式的权重，得到的权重分别为 19.43%、44.99%、35.57%，然后使用 BP 神经网络进行评判，将数据集划分训练数据 `train_data` 和相应的标签 `train_label`。然后使用构建 **BP 神经网络模型**，并通过 `train` 函数对模型进行训练。并定义一个待评判的测试样本 `test_data`，并使用训练好的神经网络模型对其进行评判，得个指标下婴儿的睡眠质量好坏。其次利用**逻辑回归模型**将婴儿综合睡眠质量与母亲的身体指标、心理指标进行关联，划分训练集与测试集，将前 390 行作为训练集，带入模型中，根据后 20 组的母亲身体特征与心理特征对婴儿综合睡眠质量进行预测，得到预测结果，利用混淆矩阵与 ROC 曲线对模型进行评价，发现模型效果良好。

针对问题 5，引入睡眠质量评级这一新约束条件，利用**阈值综合评价模型**，将睡眠时间、睡醒次数和行为特征得分作为因变量，婴儿的入睡方式作为分类变量，研究它们之间的关联性，得知问题三的治疗策略需要调整。睡眠质量评级表达式建立在逻辑回归模型的基础上，借鉴问题四中的熵权法计算权重，得出婴儿综合睡眠质量评级为优。最后通过理想点法等优化方法获得优化后的变量值调整。

关键词：随机森林；逻辑回归；BP 神经网络

一、 问题背景

分娩相关创伤后应激障碍（CBTS,Childbirth-related Post-traumatic Stress Disorder）和产后抑郁症（Edinburgh Postnatal Depression Scale, x）对婴儿睡眠质量和行为特征有较大影响，母亲的心理健康对婴儿的成长起着重要的影响，然而，目前关于这两种心理疾病与婴儿睡眠质量和行为特征之间的研究较少且结论不一致。

在本文中，我们将分析相关数据，主要以探究母亲心理指标、母亲生理指标、婴儿行为特征和婴儿睡眠质量的关系。

研究结果将为我们提供关于 CB-PTSD 和 EPDS 对婴儿睡眠质量和行为的影响的深入理解，为改善婴儿生活质量提供可能的途径和策略。这些研究成果将有助于提高对母亲心理健康问题的关注和干预措施的制定，从而促进婴儿的健康成长和发展。

二、 问题分析

2.1 问题一的分析

对于问题一，母亲的身体和心理指标如何影响婴儿的行为特征和睡眠质量。通过分析来自 390 名 3 至 12 个月婴儿及其母亲的数据，针对于揭示母亲的年龄、婚姻状况、教育程度、妊娠时间、分娩方式、心理健康状态等与婴儿的睡眠质量和行为特征之间的潜在关系，着重考虑以下几个方面：

- 1.我们需要了解数据的结构和内容，包括各指标意义与其可能性的关联，以确定分析的方法，由于数据包括数值和分类特征，需对如何处理该特征进行考量，例如使用 One-Hot 编码来转换分类特征。

- 2.在本案例中，由于随机森林模型可以处理复杂的非线性关系，并提供特征重要性的评估，故选择该模型。

- 3.解释模型的结果，理解母亲的身体和心理状态如何影响婴儿的行为和睡眠，有助于提供有针对性的干预或支持。

- 4.通过绘制描述性统计图表和模型结果图，提高分析的针对性与全局性。

2.2 问题二的分析

对于问题二，目标为预测婴儿的行为特征信息。我有一组特征，包括母亲年龄、婚姻状况、教育程度、妊娠时间(周数)、分娩方式及心理健康指标(CBTS、EPDS、HADS)。我们的任务是基于这些特征建立一个模型，以准确预测婴儿的行为类型。

2.3 问题三的分析

针对第三问，我们首先建立了关于 CBTS,EPDS, HADS 的变化与治疗费用的线性回归模型，划分测试集，代入模型求解将矛盾型转化为中等型治疗费用的最小值，求得为 1.85 万元，再从中等型转化为安静型，我们首选求得花费的最小值为 1.94 万元，此时对应的数值是 CBTS 得分为 17，HADS 得分调整为 15，EPDS 不变，对线性回归模型进行评价，求得模型的准确率，精确率，召回率分别为 0.98，0.976，0.95，模型效果较好，做出柱状图进行可视化。

2.4 问题四的分析

针对问题四，首先我们对婴儿的睡眠质量进行综合评判，首先使用熵权法确定婴儿整晚睡眠时间、睡醒次数、入睡方式的权重。然后使用 BP 神经网络进行评判，得到不同指标下婴儿的睡眠质量好坏，将结果填入数据集中。然后利用逻辑回归模型将婴儿综合睡眠质量与母亲的身体指标、心理指标进行关联，将前 390 行作为训练集，带入模型

中，根据后 20 组的母亲身体特征与心理特征对婴儿综合睡眠质量进行预测。

2.5 问题五的分析

对于问题五是在问题三的基础上，考虑婴儿综合睡眠质量与母亲身体指标和心理指标之间的关联，进行问题三中治疗干预方案在的再次优化。该问题可以看作是问题三的拓展，引入了睡眠质量评级这一新的约束条件进行求解，利用问题四建立的阈值综合评价模型，将整晚睡眠时间、睡醒次数、行为特征得分作为因变量，婴儿的入睡方式作为分类变量，探寻其之间的关联性。得出具体的睡眠质量评级表达式，用理想点法进行求解。

三、 模型假设

1. 假设每个样本（婴儿和母亲）是独立观测的，不受其他样本的影响。
2. 假设收数据是准确和完整的，没有系统偏见或误导性错误。
3. 假设样本具有代表性，可以推广到更广泛的母亲和婴儿人群。

四、 问题一的模型建立与求解

4.1 模型准备

4.1.1 数据预处理

数据集包括了母亲的年龄、婚姻状况、教育程度、妊娠时间、分娩方式、心理指标（CBTS、EPDS、HADS）以及婴儿的行为特征、性别、年龄、整晚睡眠时间、睡醒次数和入睡方式等信息。

由于数据量较小，直接观察可知本数据集无缺失值，整晚睡眠时间有一异常值 99:99，且该数据类型不方便进行分析，故转化为一小时为单位：

表 1 整晚睡眠时间数据转换

整晚睡眠时间(转变前)	整晚睡眠时间(转变后)
10: 00: 00	10
11: 00: 00	11
12: 00: 00	12
10: 30: 00	10.5

同样为便于后续进行分析，将字符行数据婴儿行为特征和入睡方式进行 One-Hot 编码。独热编码的主要优点是：1）能够有效解决混合的数据类型不易处理的问题；2）离散值进行独热编码后，特征之间的距离计算更加合理，离散型变量的特征连续化；3）数据集中的特征能够得到有效的扩充^[1]。

婴儿入睡方式被描述为五个不同的方法：哄睡法；抚触法；安抚奶嘴法；环境营造法；定时法，使用 One-Hot 编码，我们将这五个入睡方式转换为五个二进制特征。例如，如果某个婴儿使用“哄睡法”入睡，则“入睡方式_1”列将为 1，其他四列将为 0。对三个婴儿行为特征，如上述同样转换成三个二进制特征。

4.1.2 探索性数据分析

首先将提供数据集的基本统计描述，以了解各个特征的分布和概况。基本统计描述提供了每个特征的分布情况，例如均值、标准差、最小值和最大值等。如下表所示：

表 2 各特征分布情况

count	mean	std	min	25%	50%	75%	max
-------	------	-----	-----	-----	-----	-----	-----

母亲年龄	380	30.239474	4.412506	19	27	30	33	47
婚姻状况	380	1.963158	0.188622	1	2	2	2	2
教育程度	380	4.097368	0.98725	1	3	4	5	5
妊娠时间（周数）	380	39.120789	1.896564	26.5	38	39.2	40.4	43
分娩方式	380	1.013158	0.114101	1	1	1	1	2
CBTS	380	6.010526	5.029191	0	2	5	9	21
EPDS	380	9.157895	6.79487	0	4	8	13	28
HADS	380	7.902632	4.309099	0	5	7	11	20
婴儿性别	380	1.484211	0.500409	1	1	1	2	2
婴儿年龄（月）	380	1.944737	0.818389	1	1	2	3	3
整晚睡眠时间	380	10.171711	1.448189	5	10	10	11	12
睡醒次数	380	1.471053	1.621656	0	0	1	2	10
婴儿行为特征_1	380	0.305263	0.461126	0	0	0	1	1
婴儿行为特征_2	380	0.578947	0.494379	0	0	1	1	1
婴儿行为特征_3	380	0.115789	0.320394	0	0	0	0	1
入睡方式_1	380	0.221053	0.415503	0	0	0	0	1
入睡方式_2	380	0.178947	0.383814	0	0	0	0	1
入睡方式_3	380	0.052632	0.223591	0	0	0	0	1
入睡方式_4	380	0.428947	0.495578	0	0	0	1	1
入睡方式_5	380	0.118421	0.323532	0	0	0	0	1

为更直观、更深入地理解和解释数据集中的各种特征和关系。我们绘制如下的图像，让我们对数据有更直观的理解揭示潜在的趋势和模式，为进一步的分析、建模和解释提供基础，有助于形成如何改善婴儿的睡眠质量、母亲的心理健康等方面的策略和见解。

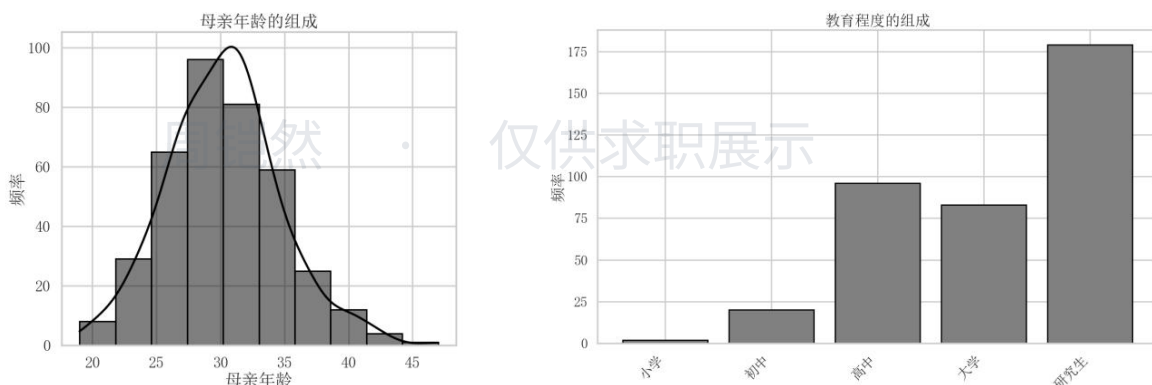


图1 母亲年龄和教育程度组成情况

由图1知，大部分母亲具有大学或高中教育背景，基本符合我国义务教育政策；母亲年龄高峰集中在25至35岁之间，符合生理适龄生育规律，如女性年龄与不孕及生育力减退一文所述，卵巢储备功能丧失女性的生育期约20年，其中25~35岁是女性的最佳生育时期，36~42岁的女性则面临生育力下降、中度的流产率和染色体异常率，43岁以上妇女则更严重，妊娠的概率极低^[2]。

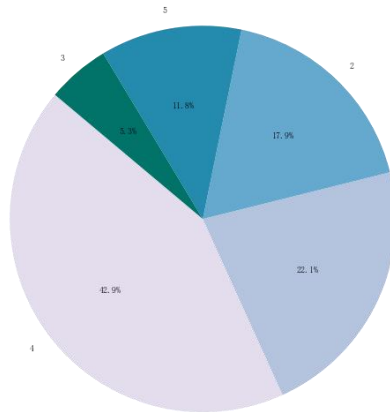


图2 入睡方式组成情况

由图2知, 1.环境营造法(Environmental Creation)占了最大比例, 约42.8%, 2.哄睡法(Coaxing Sleep)占了22.8%, 3.定时法(Timing Method)和抚触法(Touch Method)分别占了11.5%, 4.安抚奶法(Soothing Pacifier)占了较小的比例, 约5.1%, 上述符合北京市1~3月龄婴儿睡眠现状及影响因素研究文中所述, 婴儿睡眠昼夜节律紊乱受环境因素影响比例较大^[3]。

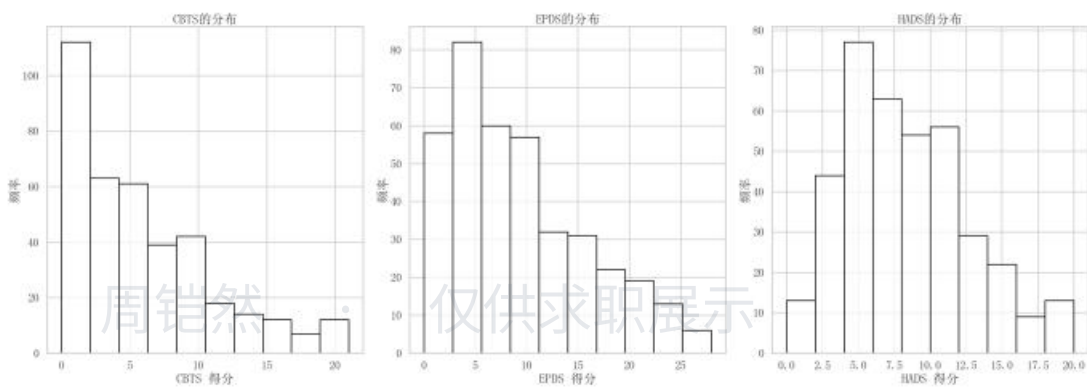


图3 母亲心理指标(CBTS、EPDS、HADS)的组成

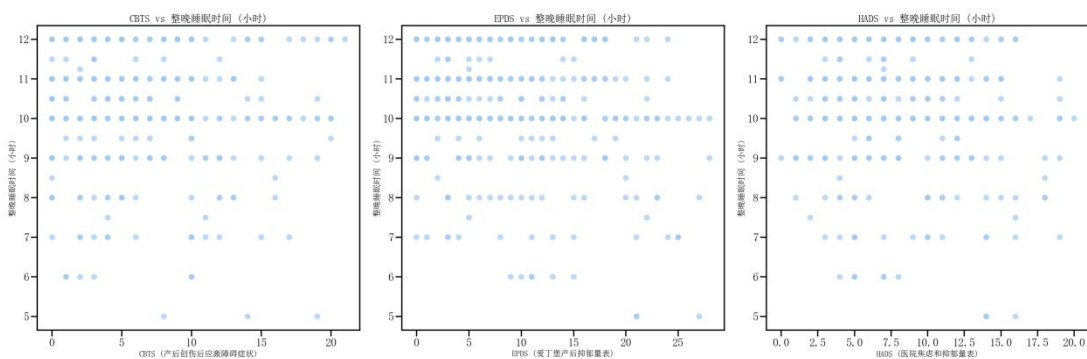


图4 母亲心理指标(CBTS、EPDS、HADS)与婴儿整晚睡眠时间的分布关系图

由图3-4知, 母亲心里指标与婴儿整晚睡眠时间关系较为模糊, 经阅读资料发现睡眠问题婴儿母亲抑郁情绪明显高于婴儿正常睡眠母亲^[4], 故需建立较复杂模型来捕捉潜在关系。

4.2 模型建立

我们目标为通过结合多个决策树的预测结果来提高模型的准确性, 从而减少分类错

误或回归预测误差。

决策变量为：

1. 训练集：原始数据集的子集，通过 Bootstrap 采样方法随机抽取形成。
2. 特征选择：在构建每个决策树时，随机选择部分特征用于节点分割。

目标函数为：

$$\frac{1}{T} \sum_{i=1}^T x_i \quad (1)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{\text{actual},i} - y_{\text{pred},i})^2 \quad (2)$$

其中 T 表示决策树的数量（随机森林中的树的个数）， x_i 表示第 i 棵决策树的预测结果。n 表示样本数量， $y_{\text{actual},i}$ 表示真实值， $y_{\text{pred},i}$ 表示模型预测值。

约束条件为：

特征从下述 15 个特征开始，母亲年龄；婚姻状况；教育程度；妊娠时间（周数）；分娩方式；CBTS；EPDS；HADS；婴儿性别；婴儿年龄（月）；整晚睡眠时间；睡醒次数；婴儿行为特征_1；婴儿行为特征_2；婴儿行为特征_3；入睡方式_1；入睡方式_2；入睡方式_3；入睡方式_4；入睡方式_5。

4.3 模型求解

同样，我们根据修改后建立的数学模型，利用 Python 编程对该问题进行了求解，得整晚睡眠时间和睡醒次数均方误差分别为 1.7285 和 2.23843，特征重要性分布如图 5-6。

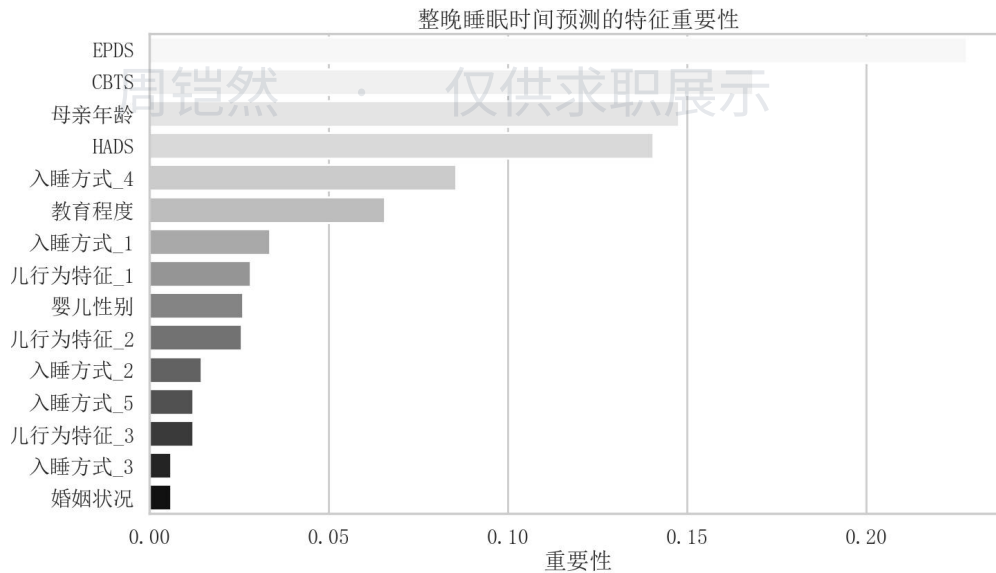


图 5 基于整晚睡眠时间预测的特征重要性

由图 5 知 EPDS 指标为最重要特征，重要性为 0.227；CBTS 指标重要性为 0.168；母亲年龄为 0.147；HADS 指标重要性为 0.140；入睡方式_4（环境营造法）为 0.085492；教育程度为 0.065534，其他特征重要性较低。

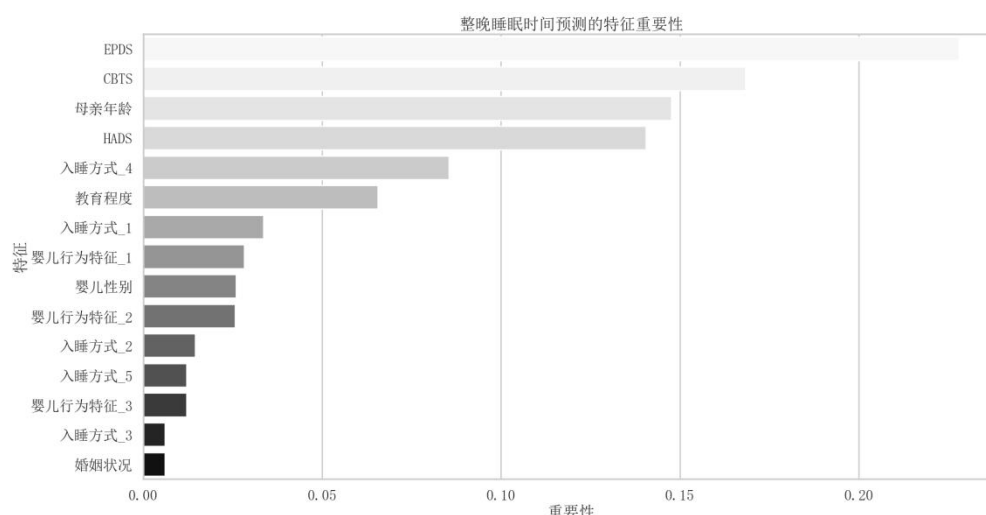


图 6 基于睡醒预测的特征重要性

由图 6 知 EPDS 指标为最重要特征，重要性为 0.2；母亲年龄重要性为 0.158；HADS 为 0.121；CBTS 为 0.117；入睡方式_4（环境营造法）为 0.085492；教育程度为 0.052，其他特征重要性较低。

4.4 总结

通过分析和建模，我们发现母亲的心理健康指标，尤其 EPDS(爱丁堡产后抑郁量表)得分，与婴儿的整晚睡眠时间和睡醒次数有显著的关联性。除了 EPDS 得分外，其他重要的身体和心理因素，如妊娠时间、母亲年龄和 CBTS(分娩相关创伤后应激障碍问卷)得分。

这些因素可能与母亲在怀孕和分娩期间整体身心上的健康有关，从而影响婴儿的睡眠质量。从入睡方式对婴儿睡眠质量的影响上看，其各方式重要性不大相近，如哄睡法和环境营造法重要性较大，可能对某些婴儿的睡眠有积极影响。教育程度与母亲的心理健康和婴儿睡眠之间有潜在联系，虽然其影响可能不如上述因素显著，但可能与家庭环境和母亲的社交支持有关。

从实际角度出发，母亲的心理健康状况在一定程度上直接影响着婴儿的日常护理和互动方式。情感健康问题，可能直接影响母亲与婴儿之间的亲近程度；妊娠时间、分娩方式以及与分娩相关的创伤后应激状况，都可能与婴儿的早期发展产生关联，例如，分娩经历可能会影响母乳喂养的实施以及早期育儿实践的展开；教育程度和婚姻状况不仅反映了母亲自身的资源状况，还影响了她们面对压力时的应对能力，这些因素潜在影响着婴儿的睡眠质量。

五、 问题二的模型建立与求解

5.1 模型准备

5.1.1 数据预处理

由于定义的特征具有不同的量纲，所以使用了 Min-Max 缩放对其进行规范化以便所有特征量纲的影响，同时提高机器学习模型的训练和性能。

表 3 Min-Max 缩放

编号	母亲年龄	婚姻状况	教育程度	妊娠时间	分娩方式	CBTS	EPDS	HA DS
----	------	------	------	------	------	------	------	-------

0	0.535714	1	1	0.636363	0	0.142857	0.46428	0.45
1	0.500000	1	1	0.939393	0	0	0	0.15
2	0.642857	1	1	0.878787	0	0.19047	0.285714	0.45
3	0.428571	1	1	0.666666	0	0.285714	0.571428	0.65

规范化后的特征范围在 0 到 1 之间。但为了选择出最重要的特征，将进行特征工程。

首先使用热力图来可视化特征之间的相关性，但由绘制的热力图可以发现并没有明显高度相关的特征，如下图：

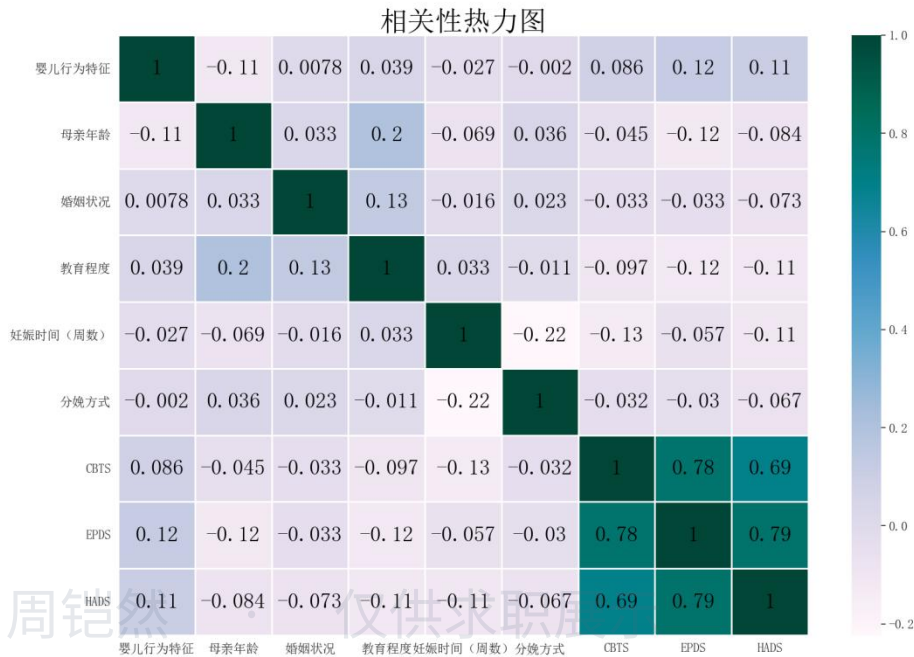


图 7 热力图

可以通过随机森林分类器的特征重要性来选择与目标变量最相关的特征来减少模型的复杂性和提高模型的准确性，如下图：

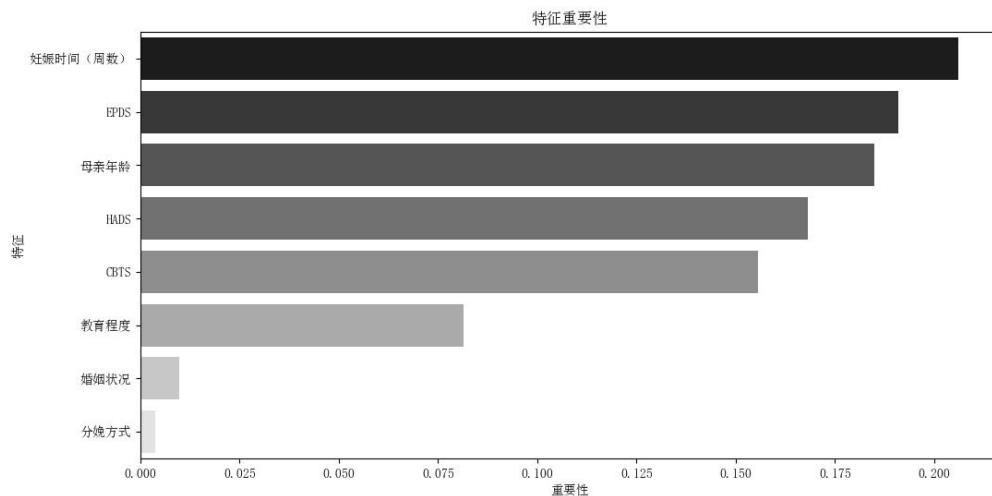


图 8 特征的重要程度图

5.2 模型建立

逻辑回归（Logistic regression）是机器学习领域最为常见的模型方法之一，常常用于作为处理各种任务的基准模型（baseline）。当然各类任务往往最后所用模型的性能远超过逻辑回归方法，但是其模型结构的复杂度和训练时间也往往超过逻辑回归方法。

逻辑回归 sigmoid 函数的数学形式为：

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

sigmoid 函数将属于原本属于实数范围的输入 z 映射到 $[0,1]$ 之间。并且对于 $Y \in \{0,1\}$ 的二分类问题而言，逻辑回归的条件概率分布 $P(Y|X)$ 可以写为：

$$P(Y = 1|X) = \frac{1}{1 + e^{-(w^T X + b)}} \quad (2)$$

$$p(Y = 0|X) = \frac{e^{-(w^T X + b)}}{1 + e^{-(w^T X + b)}} \quad (3)$$

$$P(Y|x) = P(Y = 1|x)^y P(Y = 0|x)^{1-y} \quad (4)$$

对给定训练集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，其似然函数写为：

$$\prod_{j=1}^N P(Y = 1|X_i)^{y_i} P(Y = 0|X_i)^{1-y_i} \quad (5)$$

其对数似然函数写为：

$$pL(w) = \sum_{i=1}^N [y_i \cdot (w^T x_i) - \log(1 + e^{w x_i})] \quad (6)$$

应用极大似然估计法需要对 $L(w)$ 求极大值，进而得到 w 的估计值。

$$\hat{w} = \sup_w -\frac{1}{N} \sum_{i=1}^N [y_i (w^T x_i) - \log(1 + e^{w x_i})] \quad (7)$$

5.3 模型评估

5.3.1 随机森林分类器

随机森林分类器：在测试集上评估模型的性能，并使用混淆矩阵和分类报告来了解性能。表 4 提供了每个类别的精确度、召回率和 F1 分数。从这些结果中，我们可以看到模型在某些类别上的表现不佳，可能由于数据不平衡或特征不足以捕捉目标变量的复杂性所导致。

表 4 分类报告

	precision	recall	f1-score	support
中等型	0.62	0.82	0.70	49
安静型	0.33	0.17	0.22	24
矛盾型	0.00	0.00	0.00	5
accuracy			0.56	78
macro avg	0.32	0.32	0.31	78
weighted avg	0.49	0.56	0.51	78

5.3.2 ROC 曲线

ROC 曲线是反应敏感性和特异性连续变量的综合指标，ROC 曲线上每一个点表示在阈值下进行分类的真阳性率（True Positive Rate）和假阳性率值（False Positive Rate）。AUC 表示 ROC 曲线下的整个二维区域，其面积是衡量二分类模型优劣的一种评价指标。图 15 中调优后的决策树模型 AUC 值为 0.999。由此可见，该决策树优化模型具有较大的优势，灵敏性较优。在衡量学习器的泛化性能时，根据学习器的预测结果对样本排序，按此顺序逐个把样本作为正例进行输出，每次计算测试样本的真正率 TPR ，和假正率 FPR 并把这两项作为 ROC 的纵轴和横轴。其中真正率衡量实际值为正例的样本中被正确预测为正例的样本的比例，假正率表示实际值为负例的样本中被错误的预测为正例的样本的比例。

$$TruePositiveRate = \frac{TP}{TP + FN} \quad (1)$$

$$FalsePositiveRate = \frac{FP}{TN + FP} \quad (2)$$

具体来说，假设一逻辑回归分类器，根据其预测结果对样例排序，设初始阈值为最大，此时所有样本皆预测为反例，无样本被预测为正例即 $TP = 0, FP = 0, TPR = 0, FPR = 0$ ；之后我们不断调小阈值，在这个过程中被正确预测为正例的样本和被错误预测为正例的样本都在不断增多，同时被正确预测为反例的样本和被错误预测为反例的样本不断减小。因此 TPR 和 FPR 会不断变大。当阈值为最小时，各样本被预测为正，没有样本被预测为负即 $FN = 0, TN = 0, TPR = 1, FPR = 1$ ；根据这一系列阈值我们可以得到一系列一对应的真正率和假正率。

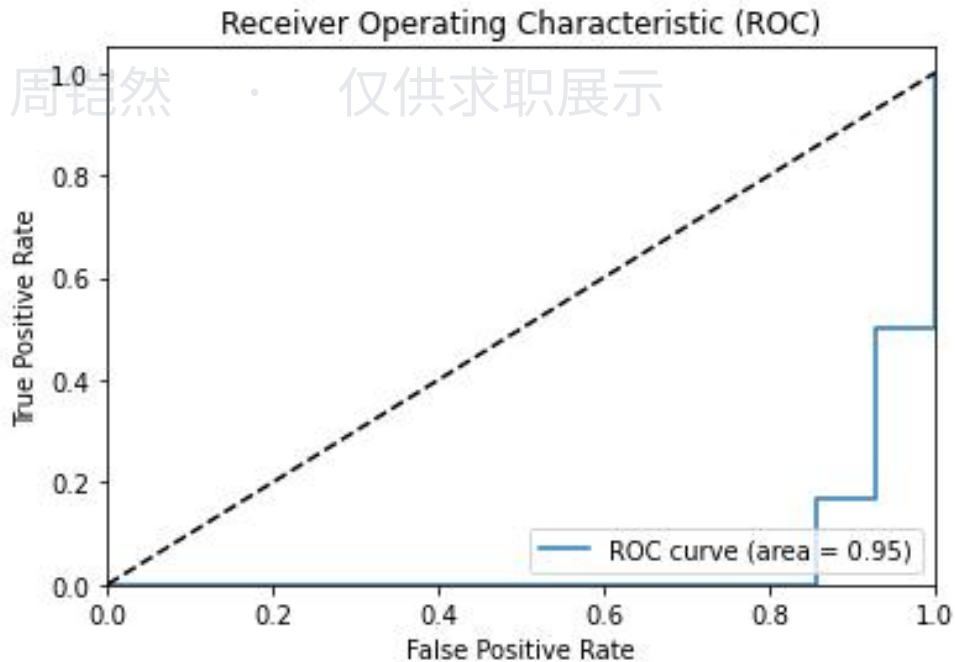


图 9 AUC 曲线

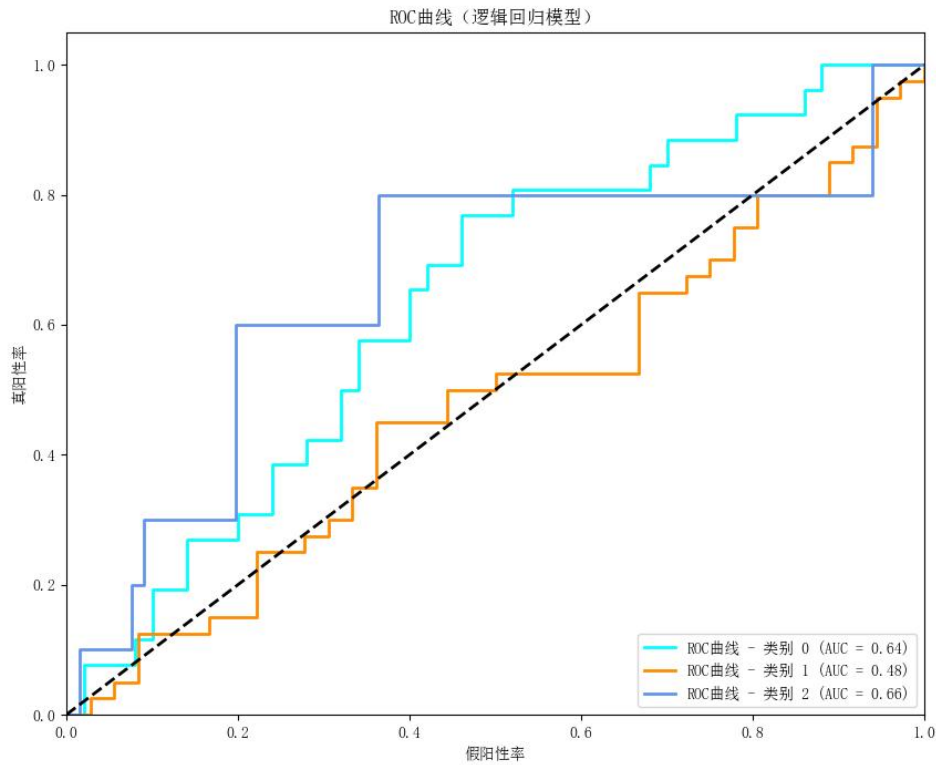


图 10 逻辑回归模型 ROC 曲线

根据表 5，故选择逻辑回归模型。

表 5 各模型 AUC 值

模型	类型 0 (安静型)	类型 1 (中等型)	类型 2 (矛盾型)
逻辑回归	0.64	0.48	0.66
支持向量机	0.54	0.36	0.37
K-近邻	0.44	0.47	0.53

5.4 模型求解

将数据集划分好训练集与测试集后，运用训练好的逻辑回归模型对后 20 组婴儿行为特征进行预测，得到结果示例如表 6，其中婴儿行为特征安静型、中等型、矛盾型依次对应数值 1、2、3，完整结果见附件 2。

表 6 婴儿行为特征

编号	CBTS	EPDS	HADS	婴儿行为特征
391	7	15	12	2
392	9	14	12	2
393	7	12	7	2

六、 问题三的模型建立与求解

6.1 模型建立

在统计学中，线性回归方程利用最小二函数对一个或多个自变量之间关系进行建模

的一种回归分析。这种函数是一个或多个称为回归系数的模型参数的线性组合。只有一个自变量的情况成为简单回归，大于一个自变量的为多元回归。在线性回归中，数据使用线性预测函数来建模，并且未知的函数模型也是通过数据来估计，这种模型被称为线性模型。最常用的线性回归建模是给定的 x 条件下 y 条件分布的分位数作为 x 的线性函数表示。

当多个自变量与因变量之间是线性关系时，则多元线性回归模型为：

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k + e \quad (1)$$

其中， b_0 为常数项， $b_1, b_2, \dots, b_k$ 为回归系数， b_1 为 $x_2, x_3, \dots, x_k$ 固定时， x_1 每增加一个单位对 y 的效应，即 x_1 对 y 的偏回归系数；同理 b_2 为 x_1, x_k 固定时， x_2 每增加一个单位对 y 的效应，即， x_2 对 y 的偏回归系数等。如果多个自变量 x 对同一个因变量 y 呈线性相关时，可用多元回归模型的参数估计。在要求误差平方和 ($\sum e^2$) 为最小的前提下，用最小二乘法求参数。以线性回归模型[2]为例，求解回归参数标准方程组为。

6.2 模型求解

根据线性回归方程的建立易得：若要将编号 238 数据的婴儿行为特征从矛盾型转变为中等型需要花费至少 1.85 万；若要将变为安静型，花费至少需要 1.94 万元，

指标调整如表 7：

表 7 干预方案心理指标调整

	CBTS	EPDS	HADS
调整前	15	22	18
调整后	17	22	15

6.3 模型评估

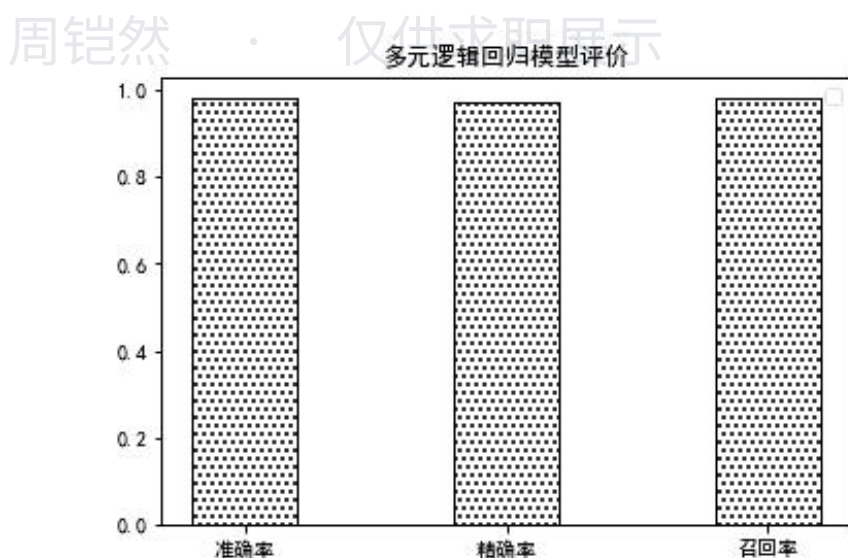


图 11 多元逻辑回归模型评估指标

由图 10 知，问题四模型准确率、精确率和召回率指标数值都接近于 1，说明模型表现良好。

七、 问题四的模型建立与求解

7.1 模型准备

熵权法 (the entropy weight method 简称 EWM) 是脱胎于信息论基本原理的解释, 信息是系统有序程度的一个度量, 熵是系统无序程度的一个度量; 如果指标的信息熵越小, 该指标提供的信息量越大, 在综合评价中所起作用理当越大, 权重就应该越高。熵权法是常用的一种求权重的方法。它是指一个随机变量与某一组随机变量间线性相依性的度量^[5]。

步骤:

1. 数据标准化

将各个指标的数据进行标准化处理。

假设给定了 k 个指标 $X_1, X_2, \dots, X_n$, 其中 $X_i = X_1, X_2, \dots, X_n$ 。假设对各指标数据标准化后的值为 $Y_1, Y_2, \dots, Y_k$, 那么

$$Y_{ij} = \frac{X_{ij} - \min(X_i)}{\max(x_i) - \min'(j)} \quad (1)$$

2. 求各指标的信息熵

根据信息论中信息熵的定义, 一组数据的信息熵。

$$E_j = -\ln(n)^{-1} \sum_{i=1}^n P_{ij} \ln p_{ij} \quad (2)$$

其中,

$$p_{ij} = \frac{Y_{ij}}{\sum_{i=1}^n Y_i} \quad (3)$$

如果, $p_{ij} \geq 0$, 则定义

$$\lim_{p_{ij} \rightarrow 0} p_{ij} \ln p_{ij} = 0 \quad (4)$$

3. 确定各指标权重

根据信息熵的计算公式, 计算出各个指标的信息熵为 $E_1, E_2, \dots, E_n$ 。通过信息熵计算指标的权重:

$$w_i = \frac{1 - E_i}{k - \sum E_i} (i = 1, 2, \dots, k) \quad (5)$$

4. 确定结果

通过对影响碳排放的参数进行熵权法确定权重, 得到的权重结果如下所示, 其中指标权重最大值为 建筑面积 (37.52%), 最小值为 建筑类型 (16.65%)。

表 8 熵权法权重确定

指标	权重
整晚睡眠时间	19.43%
睡醒次数	44.99%
入睡方法	35.57%

这里使用熵权法对神经网络算法的权值阈值进行优化，选择婴儿的睡眠质量指标作为输入参数，以提高稳定性和精度。为了提高使用算法计算时的收敛速度和寻找最优解的效率，将数据进行归一化处理，使输入和输出的数据【-1,1】在之间。

7.2 模型建立

7.2.1 Bp 神经网络的构建

Bp 神经网络这种由误差反向传播训练的多层前馈网络，由数个非线性变换单元组成，是当前应用较为广泛的神经网络模型。

Bp 神经网络由两部分组成，分别是信息流在网络进行正向传播以运算输出结果与此同时误差流不断向前反馈进行反向传播修正输出结果。Bp 神经网络的结构，如图所示。输入层在接收到归一化处理后的训练数据后，使用硬极限函数进行处理，信息由输入神经元经隐含层神经元层层处理后流向输出层。误差反向传播主要体现在训练过程中，当训练信号传到输出层进行输出后，估计结果与真实值对比然后计算误差，利用该误差来估计输出层前导层的误差，再利用前导层误差估计更前一层的误差，这样循环到获得所有其他各层的误差，最后使用所得误差对各层节点进行修正，来提高整个网络的精度^[6]。

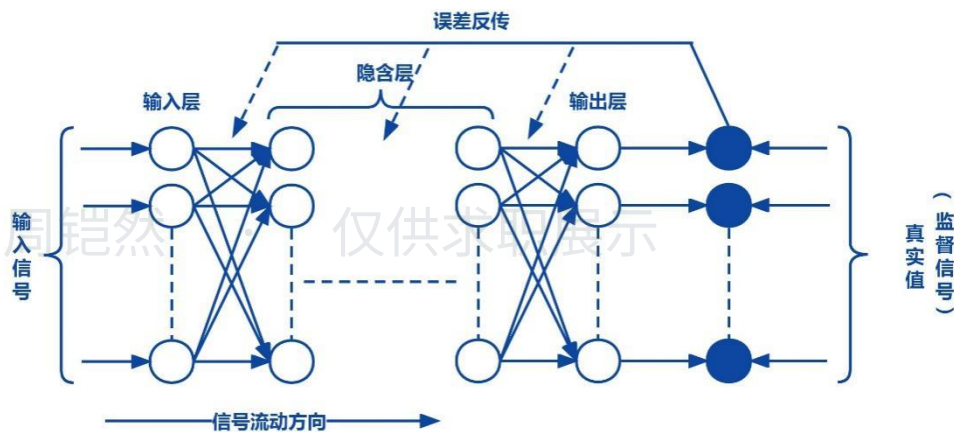


图 12 BP 神经网络示意图

在本模型中，输入为【客观变量】输出为当前状态下的【操作变量】。隐含层及输出各节点输出模型为：

$$S_j = f \left[\sum W_{ij} * X - b_j \right] \quad (1)$$

式中：f—传递函数； W_{ij} ——各神经元权值矩阵； b_j ——各神经元对应隐含层的阈值向量。为了让输出规整在（0，1）区间内，传递函数选择为双曲正切函数（1，2）。

$$f(x) = \tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2)$$

BP 算法的误差计算模型通常有 mas 和 mse 两种，这里选用均方误差 mse 模型，计算公式为：

$$E(k) = E[e^2(k)] \approx \frac{1}{nS^2} \sum_{j=1}^{s^2} \sum_{i=1}^{s^2} [t_i^2 - a_i^2(k)]^2 \quad (3)$$

式中：E(k) —第 k 次迭代的网络输出总误差；t—真实值，在训练过程中；t—当前实际的 SOC 值。

BP 网络的学习算法选用梯度最速下降法，这是一种有监督的学习方法。因 BP 学习算法与 LMS 算法基本相同，F(W) 来近似表述权值矩阵为 W 时网络的误差。F(W) 计算公式如下：

$$F(W) = e^T(k)e(k) \quad (4)$$

定义敏感度为：

$$s^m = \frac{\partial E}{\partial n^m} = \left(\frac{\partial E}{\partial n_1^m} \frac{\partial E}{\partial n_2^m} \cdots \frac{\partial E}{\partial n_s^m} \right) \quad (5)$$

使用链法则，通过网络将敏感性进行反向传播：

$$s^M = -2E(n^m)(t - y) \quad (6)$$

$$s^M = E^m(n^m)(W^{m+1})^T s^{m+1}; \quad m = M - 1, \dots, 2, 1 \quad (7)$$

式中：M—总隐含层数；m—当前隐含层数。

最后推导出权值和阈值的学习公式为：

$$W^m(k+1) = W^m(k) - \alpha s^m(y^{m-1})^T \quad (8)$$

式中：α—学习速率，这里选用 0.01。

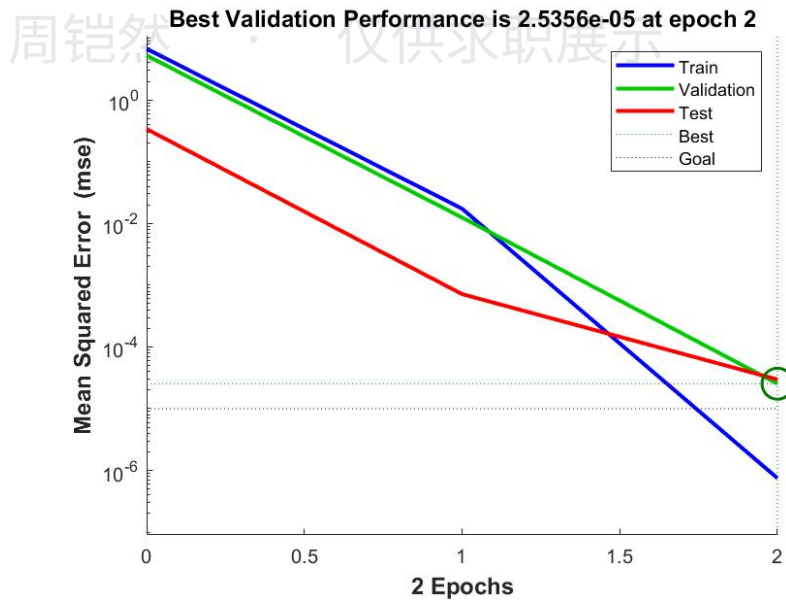


图 13 均方误差曲线图

图上三条有颜色的实线分别是：BP 训练过程的 MSE 指标在每一代中的表现，BP 交叉验证过程的 MSE 指标在每一代中的表现，BP 测试过程的 MSE 指标在每一代中的表现，从图中可以看出，当 BP 神经网络训练到第 2 代时的效果最好。

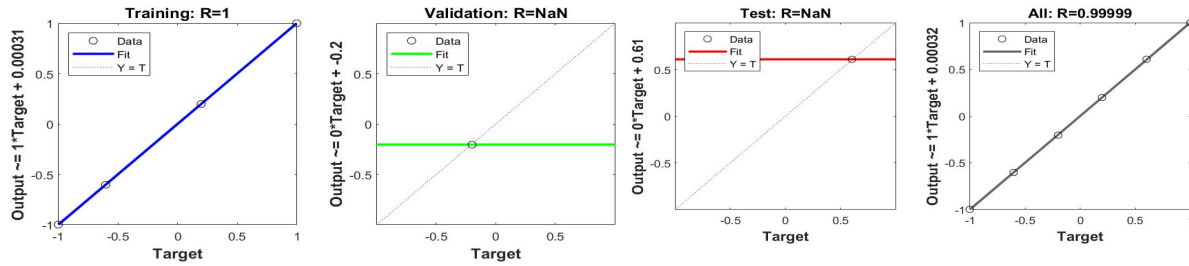


图 14 迭代效果检验图

在训练时，用 `training` 训练，每训练一次，系统自动会将 `validation set` 中的样本数据输入神经网络进行验证，在 `validation set` 输入后会得出一个误差（不是网络的训练误差，而是验证样本数据输入后得到的输出误差，可能是均方误差），这里对 `validation set` 会设置的步数为 `2echo`，并且在连续 2 次检验后不下降，说明这时候 `training set` 训练的误差已经不再减小，没有更好的效果了。

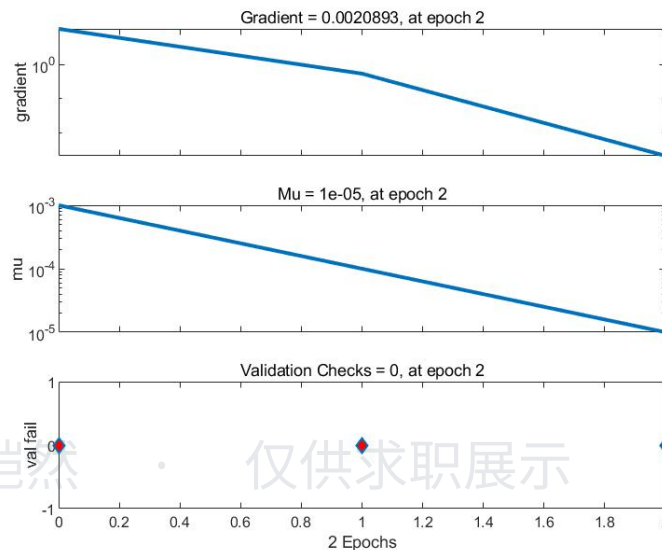


图 15 训练集、测试集与验证集适应曲线图

从上图可以看出，训练集、测试集与验证集的适应效果都非常好，说明我们模型训练的准确性。

7.2.2 逻辑回归模型的构建

与问题二使用相同模型，故见 5.2。

7.3 模型求解

7.3.1 Bp 神经网络的求解结果

将数据集划分训练数据 `train_data` 和相应的标签 `train_label`。然后使用构建 BP 神经网络模型，并通过 `train` 函数对模型进行训练。并定义一个待评判的测试样本 `test_data`，并使用训练好的神经网络模型对其进行预测，得到部分评价结果如下表，具体见附件。

表 9 部分评价结果

编号	母亲年龄	婚姻状况	教育程度	妊娠时间（周数）
1	34	2	5	37
2	33	2	5	42
3	37	2	5	41

4	31	2	5	37.5
5	36	1	5	40
6	32	2	5	41
7	28	2	4	41
8	34	2	5	39
9	32	2	5	41.3
10	34	1	3	37.2
11	24	2	4	41
12	26	2	4	39.4
13	24	2	5	35
14	20	2	3	40.5
15	26	2	5	40

7.3.2 逻辑回归的求解结果

将综合评判结果导入数据处理后的数据集，带入逻辑回归模型中，划分训练集与测试集，这里我们将前 390 行作为训练集，带入模型中，根据后 20 组的母亲身体特征与心理特征对婴儿综合睡眠质量进行预测，得到的部分预测结果如下表所示，完整数据请见附件 2。

表 10 后 20 组婴儿睡眠质量预测结果

编号	母亲年龄	婚姻状况	教育程度	妊娠时间 (周数)	分娩方式	CBTS	EPDS	HADS	综合睡眠质量
391	29	2	4	40	1	7	15	12	2
392	29	2	3	42	1	9	14	12	2
393	23	2	2	38.5	1	7	12	7	1
394	27	2	3	36.3	1	8	4	5	2
395	36	2	4	39	1	6	6	8	2
396	30	2	5	41.2	1	5	8	5	2
397	28	2	2	40.6	1	8	11	9	1
398	32	2	5	37	1	3	6	7	2
399	28	2	5	38	1	7	11	5	2

通过上表我们可以清楚的了解，婴儿睡眠质量与母亲的身体与心理特征关系十分密切，这是我们值得注意的问题。

7.4 模型评估

混淆矩阵多用于判断分类器的优劣。混淆矩阵要表达的含义：一为，混淆矩阵的每一列代表了预测类别，每一列的总数表示预测为该类别的数据的数目；二为，每一行代表了数据的真实归属类别，每一行的数据总数表示该类别的数据实例的数目；每一列中的数值表示真实数据被预测为该类的数目。由上实验结果得到优化后的决策树模型为最佳模型，于是对该模型的混淆矩阵进行可视化操作。由图 可见，对角线数值均在 2500 以上且颜色较深，可证明该模型优化后，预测结果较准确，泛化性能较强，模型较好。

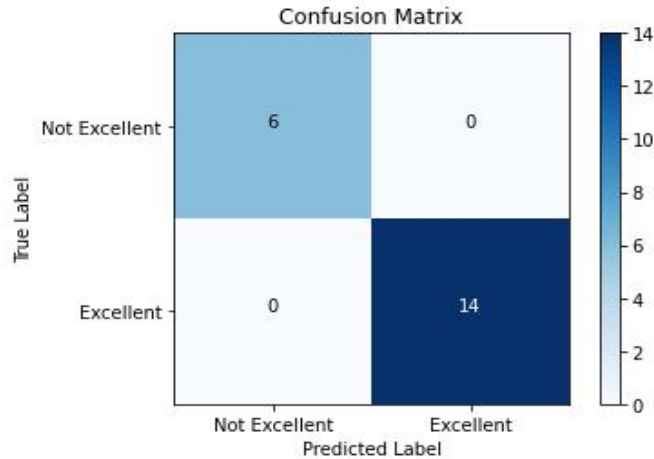


图 16 混淆矩阵

由图可见，对角线颜色较深，可证明该模型优化后，预测结果较准确，泛化性能较强，模型较好。

八、 问题五的模型建立与求解

睡眠质量评级表达式建立在原有的逻辑回归模型的基础上，将睡眠质量评级作为一个新的约束条件引入模型中。可以借鉴问题四中熵权法计算得到的权重，得编号 238 婴儿综合睡眠质量评级为优。

根据问题三建立的多元线性回归模型，加入新的睡眠质量评级约束条件进行优化，可以使用理想点法（Ideal Point Method）或其他合适的优化方法。

根据模型的求解结果，得到优化后的变量值调整。若将要婴儿从矛盾型变为中等型，根据模型可得需要花费 1.43 万元；若将要婴儿的睡眠质量变为安静型，可以使用理想点法求解出调整后的各指标得分，CBTS 得分为 12，EPDS 得分调整为 19，HADS 得分调整为 19，EPDS 不变，此时的费用为 1.95 万元。

九、 模型优缺点

模型优点：

- 1、随机森林它能够处理很高维度（feature 很多）的数据，并且不用做特征选择(因为特征子集是随机选择的)，在训练完后，它能够给出哪些 feature 比较重要
- 2、问题三提供了一种量化评估改善婴儿行为特征所需成本的方法，可以为实际干预提供参考。
- 3、在处理问题四时，创新性的分阶段进行数据降维，大大增强了模型的可解释性。
- 4、PR 曲线明显折点十分接近于 1，所以预测模型效果较好。
- 5、优化后模型的混淆矩阵对角线数值大、颜色深，即预测结果较准确，泛化性能较强，模型较好

模型缺点：

- 1、模型依赖性较强，需要保证数据本身的准确性
- 2、易出现过拟合情况
- 3、由于时间原因，未可以考虑统计模型中的线性模型、相关分析等方法。也未考虑可能存在的对数函数、指数函数关系

十、参考文献

- [1] RICE J. Mathematical Statistics and Data Analysis, International Edition(with CD Data Sets)J1. MathematicaGazette,2006,72(72):390-391.
- [2] 陈士岭,罗燕群,夏容等.女性年龄与不孕及生育力减退[J].国际生殖健康/计划生育杂志,2011,30(04):265-271.
- [3] 魏乾伟,游川,刘军等.北京市 1~3 月龄婴儿睡眠现状及影响因素研究[J].中国儿童保健杂志,2018,26(02):145-148.
- [4] 刘敏娜,肖琳,高雪婷等.西安市婴儿睡眠问题与母亲睡眠质量和情绪的相关性研究[J].中国儿童保健杂志,2012,20(09):790-793.
- [5] 何奕波,郭辉,张冰谦等.基于 GA-BP 神经网络的铁水硅含量预测方法[J/OL].钢铁研究学报:1-10[2023-08-06].DOI:10.13228/j.boyuan.issn1001-0963.20230158.
- [6] 杨孟娇,马赫,朱晓琳.基于博弈论组合赋权——TOPSIS 的微课评价模型[J]广播电视大学学报,2022(01):32-39.DOI:10.16161/j.issn.1008-0597.2022.01.005.

周铠然 · 仅供求职展示

附录

附录
介绍：支撑材料的文件列表
附件 1：编码后的数据集
附件 2：婴儿行为特征和综合睡眠质量预测结果
附件 3：问题一代码
附件 4：问题二代码
附件 5-7：问题四代码
附件 8-9：问题三代码