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		CM2303252

母亲身心健康对婴儿睡眠质量的影响分析

摘要

母亲心理健康状态的不良状况会对婴儿的认知和情感等方面产生负面影响，因此母亲的身心健康对婴儿的健康成长至关重要。本文围绕母亲的身体和心理指标与婴儿的睡眠质量指标之间的关系进行分析和建模，该问题的研究可以有利于更好地提高母亲的心理健康水平和促进婴儿的生理和心理上的健康发展。

针对问题一，首先对数据进行清洗，将异常数据进行剔除。然后建立 **SEM 模型**，将母亲的身体指标和心理指标作为自变量，婴儿的行为特征和睡眠质量作为因变量，通过求解这些变量之间的路径系数，来研究母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量的影响规律。结果为母亲的心理问题症状越严重、年龄越大和妊娠时间越长，婴儿的睡眠质量越差，行为特征越偏向矛盾型。

针对问题二，为了建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，可以先使用**基于熵权法的 TOPSIS 模型**将身体指标和心理指标进行量化，计算其得分情况。再以 **Adaboost**，**GBDT** 和**随机森林**三种基分类器组合成更强的分类器，建立 **Soft Voting 集成学习算法**，对编号 391-410 号的婴儿的行为特征信息进行预测，该模型的准确率达到 85%以上，具有较高可靠性。最终选择预测概率最大值为婴儿的最终行为特征类型。

针对问题三，本文根据问题描述建立了治疗费用与患病得分之间的数学解析式，考虑到母亲的个体差异和特征，本文将母亲的特征纳入预测模型中，建立多输入多输出 **MIMO 神经网络模型**，输入母亲年龄、妊娠周数、教育程度以及特征类型得到对应的 CBTS、EPDS 和 HADS 的得分，模型的拟合优度达到 82.4%。再代入到数学解析式中求得 238 号婴儿行为特征从矛盾型变为中等型和安静型所需要的费用。

针对问题四，为了对婴儿的睡眠质量进行优良中差四分类综合评判，本文将其看作无监督学习问题，因此选择了 K-means 算法进行求解。通过建立 **K-means 模型**，将睡眠时间、睡醒次数和入睡方式作为特征进行聚类，对婴儿的睡眠质量进行优良中差评级。再使用**基于熵权法-Topsis 模型**和 **Soft Voting 模型**验证了 K-means 模型的分类结果的合理性，模型准确率达到了 75.4%。

针对问题五，本文基于问题四模型的预测得到 238 号婴儿的综合睡眠质量评级，再利用问题三建立的 **MIMO 神经网络模型**，当输入的婴儿综合睡眠质量为优时，得到对应的 CBTS、EPDS 和 HADS 得分，即可计算出所需要的费用。

关键词：SEM 模型；TOPSIS 模型；集成学习算法；MIMO 模型；K-means 模型

一、 问题重述

1.1 问题背景

处于婴儿期的婴儿发展潜能大，可塑性强，是促进体格、智力发育和良好行为习惯的关键时期[1]。而母亲是婴儿最亲密的人，需要承担儿童的喂养和教育等，既要为其提供营养和保护，又要给予其情感支持和安全感，是婴儿最重要的人之一。所以母亲的身心健康会对婴儿的成长有很大的影响，母亲抑郁、焦虑和压力等心理健康状态的不良状况会对婴儿的认知、情感、社会行为等方面产生负面影响，例如影响婴儿的睡眠质量等。因此，研究母亲身心健康对婴儿成长的影响有利于更好地提高母亲的心理健康水平和促进婴儿的生理和心理上的健康发展。

1.2 问题提出

基于问题背景、所给的数据以及查阅的相关文献，我们将建立数学模型解决以下问题：

(1) 研究附件中的数据，探究母亲的身体和心理指标是否对婴儿的行为特征和睡眠质量有一定的影响规律。
(2) 婴儿行为特征分为安静型、中等型和矛盾型三种，建立婴儿行为模型与心理指标的关系模型，并判断数据最后 20 组缺失的婴儿行为特征属于何种类型。
(3) 编号 238 的婴儿的行为特征为矛盾型，建立模型分析花费多少费用可以使其行为特征转变为中等型，如若转变为安静型，如何调整其治疗方案。
(4) 对婴儿的睡眠质量进行综合评判，分为优良中差四类，建立婴儿综合睡眠质量与母亲身体和心理指标的关联模型，并预测编号为 391-410 的婴儿的综合睡眠质量。
(5) 在问题 3 的基础上，如何调整问题 3 中的治疗策略从而使 238 号婴儿的睡眠质量评级为优。

二、 问题分析

2.1 问题一的分析

对于问题一，我们首先需要进行数据清洗，将异常数据进行剔除以减少对研究分析的影响。然后我们可以建立 SEM 模型，将母亲的身体指标和心理指标作为自变量，婴儿的行为特征和睡眠质量作为因变量，通过求解这些变量之间的路径系数，来研究母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量的影响规律。

2.2 问题二的分析

对于问题二，为了建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，我们可以先使用基于熵权法的 TOPSIS 模型计算身体和心理指标的得分，从而将其进行量化。然后再利用 Voting 投票模型，基于 Adaboost，GBDT 和随机森林三种基分类器组合成更强的分类器，对编号 391-410 号的婴儿的行为特征信息进行预测。

2.3 问题三的分析

对于问题三，我们先根据问题建立治疗费用的数学解析式，然后建立多输入多输出 MIMO 神经网络模型，输入母亲年龄、妊娠周数、教育程度以及特征类型得到对应的 CBTS、EPDS 和 HADS 的得分，再代入到解析式中既可以求得 238 号婴儿行为特

征从矛盾型变为中等型和安静型所需要的费用。

2.4 问题四的分析

对于问题四，为了对婴儿的睡眠质量进行优良中差四分类综合评判，我们将本问看作一个无监督学习问题，因此我们可以选择 K-means 算法进行求解，建立 K-means 模型，将睡眠时间、睡醒次数、入睡方式作为特征进行聚类，从而对婴儿的睡眠质量评级。最后，我们再使用基于熵权-Topsis 模型对 K-means 模型的分类结果进行验证。

2.5 问题五的分析

对于问题五，我们根据问题四的预测得到 238 号婴儿的综合睡眠质量评级，再利用问题三建立的 MIMO 神经网络模型，输入母亲年龄、妊娠周数、教育程度和睡眠质量评级为优，得到对应的 CBTS、EPDS 和 HADS 得分即治疗策略，并计算出所需要的费用。

三、 模型假设

1. 假设附件所给的数据真实有效。
2. 假设每个样本是独立观测的，不受其他样本的影响。
3. 假设附件所给数据能够准确地反映母亲的身体指标和心理指标，无其他因素影响。

四、 符号说明

符号	说明
X	外源指标向量
Y	内生指标向量
ξ	潜在变量
a_i	基分类器权重参数
w_i	基分类器的权重
p_i	基分类器的预测概率
$f(x)$	治疗费用

五、 问题一模型的建立与求解

5.1 数据预处理

我们首先发现附件中的数据存在一些错误，如异常数据、缺失值等，需要先进行预处理。经过对数据筛选，我们可以发现编号为 180 号的婴儿的睡眠时间为 99：99，属于异常值，故对其进行剔除。此外根据数据信息简介，我们可以得知母亲的婚姻状况只有未婚 1 和已婚 2 两种情况，但部分数据出现婚姻状况为 3 和 6，故也为异常值，需要进行剔除。

5.2 SEM 模型的建立

结构方程模型（SEM）是一种多变量之间内在结构关系的统计分析方法[2]，它结

合了因素分析、路径分析和回归分析等多种分析技术，旨在研究变量之间的因果关系和潜在结构。SEM 广泛应用于社会科学、教育研究、心理学等领域。

SEM 模型中分为 2 种变量：潜在变量、显性变量和残差变量，以及两个基本模型：测量模型和结构模型。其中测量模型用于描述观察变量和其背后的潜在变量之间的关系，其关系表达式为：

$$X = \Lambda_x \xi + \delta \tag{1}$$

$$Y = \Lambda_y \eta + \varepsilon \tag{2}$$

其中，X 为外源指标组成的向量；Λ_x 为 x 指标与潜在变量 ξ 的关系；δ 代表 x 测量上的误差；Y 代表内生指标组成的向量；Λ_y 代表 y 指标与潜在变量 η 的关系；ε 代表 y 测量上的误差。

结构模型用于描述潜在变量之间的因果关系和相关关系，其关系表达式为：

$$\eta = B\eta + \Gamma\xi + \zeta \tag{3}$$

其中，η 为内生潜在变量；ξ 为外源潜变量；B 为内生潜变量间的关系；Γ 为外源潜变量对内生潜变量的影响；ζ 为结构方程的残差项。

因此使用结构方程模型（SEM）来研究母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量的影响是合理的。它可以同时考虑观测变量和潜在变量，并通过路径分析来评估变量之间的直接和间接效应。我们利用 SEM 构建一个模型，将母亲的身体指标和心理指标作为自变量，婴儿的行为特征和睡眠质量作为因变量。通过附件 1 的数据并进行分析，通过可以评估这些变量之间的路径系数，从而确定身体指标和心理指标对婴儿行为特征和睡眠质量的影响程度。

5.3 SEM 模型的求解

我们利用 Matlab 对结构方程模型进行求解，并绘制出如下结构方程模型图，如图 1 所示。

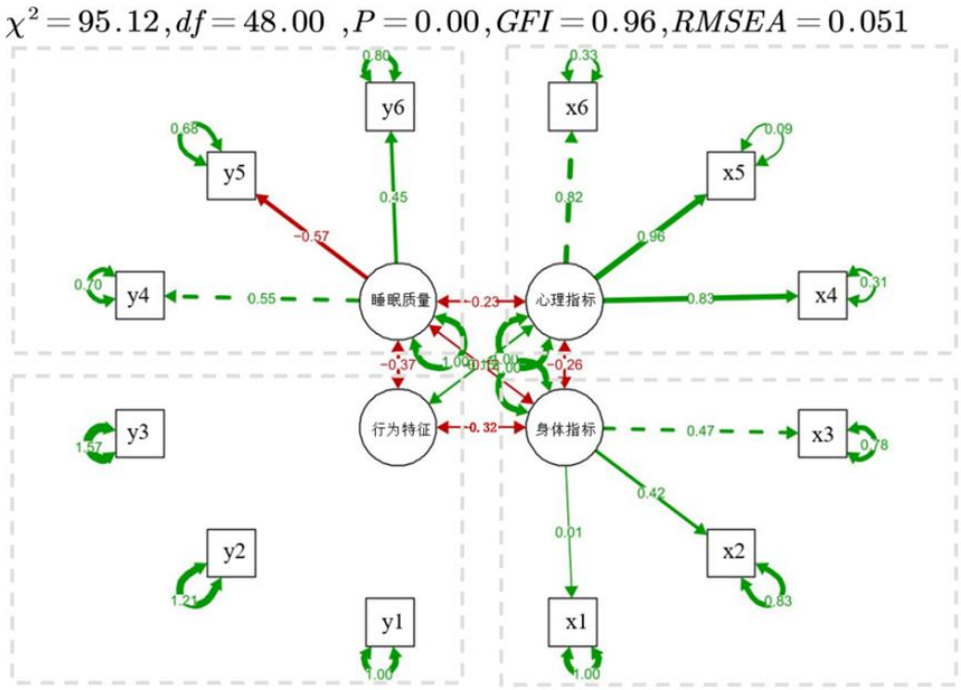

图 1 结构方程模型图

可以得出，CHISQ（卡方值）为 95.172，DF（自由度）为 48.000，PVALUE（P 值）为 0.000。这表明模型的卡方统计量较高，指示观察数据与模型之间存在显著的拟合差异。

GFI（拟合优度指数）为 0.960，CFI（比较拟合指数）为 0.952。这些值接近 1，表示模型的拟合优度较好，与基准模型相比有较好的改善。

RMSEA（近似均方根误差）为 0.051，表示模型与观察数据之间的拟合差异较小。

母亲的心理指标与婴儿的行为特征之间的路径系数为 0.32，而心理指标与睡眠质量之间的路径系数为-0.757。这表示心理指标与行为特征之间存在轻微负向关系，但与睡眠质量之间存在显著负向关系。对于它们之间的负向关系可以解释为：心理指标可以转化为问卷的得分，分数越高，表明心理问题症状越严重。因此可以得到结论：母亲的心理指标越健康，婴儿的行为特征越偏向于安静型，且婴儿的睡眠质量越好。由于数据处理的时候没有进行正向化，心理指标得分越高，婴儿的睡眠质量分数越低，婴儿的行为特征越高，即母亲的心理问题症状越严重，婴儿的睡眠质量越差，其行为特征越偏向于矛盾型。

母亲的身体指标与婴儿的行为特征之间的路径系数为-0.32，而身体指标与睡眠质量之间的路径系数为-0.56。表名身体指标的增加与婴儿行为特征和睡眠质量的减少之间存在一定程度的关联，具有中度显著影响。这意味着身体指标与行为特征和睡眠质量之间存在轻微负向关系。身体指标包括母亲年龄、妊娠周数和教育程度，因此可以得出结论：母亲的年龄越大、妊娠时间越长，婴儿的睡眠质量越差和行为特征越偏向矛盾型。

六、 问题二模型的建立与求解

6.1 基于熵权法的 TOPSIS 模型的建立

为了建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型，我们使用基于熵权法的 TOPSIS 模型计算指标得分，从而将其进行量化。

TOPSIS模型是一种常用的多属性决策分析方法[3]，它可以帮助决策者评估不同备选方案之间的综合表现，并确定最优方案。该模型基于对每个备选方案与理想解和负理想解之间的相似度进行排名，通过计算综合相似度来确定最优方案。该模型的建立过程如下：

(1) 指标的选取与正向化

首先我们需要对指标进行量化。可以发现，在母亲的身体指标中，婚姻状况为未婚所占比例仅为 14/380(3.6%)，分娩方式为剖宫产所占比例仅为 5/375(1.3%)，同时通过查阅资料以及搜索文献可以知道，母亲的婚姻状况以及分娩方式对于婴儿的行为特征影响较低，因此在建立关系模型前剔除这两项，从而使模型的准确性更加可靠。

剔除上述指标后，最后我们选取母亲的身体指标为母亲年龄、教育程度和妊娠时间，选取的心理指标为 CBTS、EPDS、HADS。

其中母亲年龄为区间型指标，经查阅相关资料发现，女性 25 岁-29 岁为最佳生育年龄；教育程度为极大型指标，分数越高越好；妊娠时间为中间型指标，正常妊娠时间为 40 周；根据附件中的描述，问卷得分越高，表明母亲的心理症状越严重，故心理指标为极小型指标。

我们需要对非极大型指标进行正向化，将其都转化为极大型指标。对于母亲年龄，我们可以采用如下公式进行转化：

$$x_{age}=1-\frac{|x_i-x_{best}|}{\max\{|x_i-x_{best}|\}} \tag{4}$$

其中x_i是一组中间型指标序列，x_{best}为其中最佳的数值。

对于妊娠时间，我们采用以下公式进行转化：

$$x_{time}=\begin{cases}1-\dfrac{x_i-25}{\max\{25-\min\{x_i\},\max\{x_i\}-29\}}, & x_i<25\\ 1, & 25\leqslant x_i\leqslant 29\\ 1-\dfrac{x_i-29}{\max\{25-\min\{x_i\},\max\{x_i\}-29\}}, & x_i>29\end{cases} \tag{5}$$

对于心理指标，我们采用以下公式进行转化：

$$x_{score}=\max\{x_i\}-x_i \tag{6}$$

(2) 标准化处理

为了消除指标不同量纲的影响，我们需要对已经正向化后的矩阵进行标准化处理，标准化的过程如下：

$$z_i=\frac{x_i}{\sqrt{\sum_i^n x_i}} \tag{7}$$

其中x_i是正向化后的指标序列，z_i为标准化后的的指标。

(3) 确定正负理想解

$$Z^+=\max_{1\leqslant i\leqslant n}\{Z_{ij}\}=(Z_1^+,Z_2^+,\cdots,Z_n^+) \tag{8}$$

$$Z^-=\max_{1\leqslant i\leqslant n}\{Z_{ij}\}=(Z_1^+,Z_2^+,\cdots,Z_n^+) \tag{9}$$

其中Z^+为正理想解，Z^-为负理想解。

(4) 加入熵权法计算的权重

利用Matlab进行熵权法权重的计算，具体结果如表1和表2所示：

表 1 身体指标权重

身体指标	信息熵值e	信息效用值d	权重(%)
母亲年龄	0.996	0.004	25.509
教育程度	0.99	0.01	63.652
妊娠时间	0.998	0.002	10.84

表 2 心理指标各权重

心理指标	信息熵值e	信息效用值d	权重(%)
CBTS	0.988	0.012	31.103
EPDS	0.987	0.013	35.232
HADS	0.987	0.013	33.664

(5) 计算距离

$$D_i^+ = \sqrt{\sum_{j=1}^{m}[w_i(Z_j^+ - Z_{ij})^2]} \tag{10}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m}[w_i(Z_j^- - Z_{ij})^2]} \tag{11}$$

其中D_i^+为评价指标与最大值的距离，D_i^-为评价指标与最小值的距离。

(6) 根据加权距离计算综合得分

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{12}$$

其中S_i为评价指标为归一化的得分。

6.2 TOPSIS 模型的求解

根据Matlab求解得到母亲身体指标和心理指标的综合得分，其中编号前10号的得分如表3所示。

表 3 综合评价得分

编号	身体指标得分	心理指标得分
1	0.85	0.62
2	0.88	0.91
3	0.79	0.67
4	0.92	0.49
5	0.82	0.88
6	0.91	0.90
7	0.80	0.16
8	0.86	0.65
9	0.91	0.87
10	0.58	0.69

通过计算我们得到了 380 名母亲的身体指标得分与心理指标得分，我们把该得分与婴儿性别、婴儿年龄，共计四个指标作为输入数据，婴儿行为特征作为输出结果，

建立机器学习模型。

6.3 Voting 模型的建立

投票法（voting）是集成学习里面针对分类问题的一种结合策略[4]。是一种遵循少数服从多数原则的集成学习模型，通过多个模型的集成降低方差，从而提高模型的鲁棒性。在理想情况下，投票法的预测效果应当优于任何一个基模型的预测效果。

投票主要有硬投票(Hard voting)和软投票(Soft voting)两种。硬投票是一种特殊的软投票，即各基分类器权重相同的投票，其原理为多数投票原则：如果基分类器的某一分类结果超过半数，则集成算法选择该结果；若无半数结果则无输出。软投票的原理也为多数投票，但是各基分类器投票所占的权重可自己定义。

Adaboost，GBDT 和随机森林都是基于 boosting 算法的分类器，分类结果较为理想，模型具有较强的泛化能力。Adaboost 是一种集成学习算法，它通过多次迭代训练多个基分类器，并将它们组合起来形成一个更强的分类器。每次迭代时，Adaboost 会调整样本的权重，更关注之前分类错误的样本。最终分类器是基分类器的线性组合，其中每个分类器的权重是根据其分类性能确定的。

GBDT 也是一种集成学习算法，它使用决策树作为基模型。GBDT 通过多次迭代，逐步提升模型性能。每次迭代时，GBDT 会拟合前一轮模型的残差，以捕捉前一个模型未能解释的部分。最终预测结果是所有基模型输出的加权和。

随机森林是一种基于决策树的集成学习算法，它使用多棵决策树来进行分类或回归。每棵决策树都是在样本和特征上进行随机采样的基础上训练得到的，这种随机性可以减少过拟合风险。最终预测结果是所有决策树的分类结果的多数表决或平均。

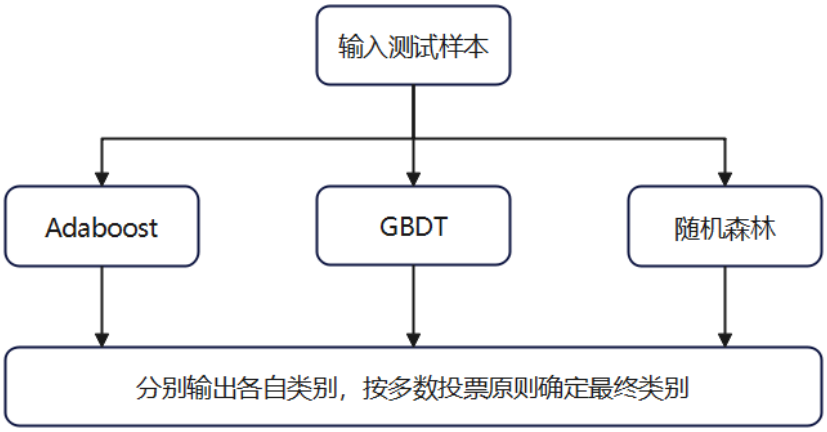

图 2 Voting 模型图解

6.4 Voting 模型的求解

对三种机器学习模型进行数据拟合和参数调优，三者在改数据集上的表现如表 4 所示，三个模型测试集的混淆矩阵如图 3 所示。

表 4 测试表现

基分类器	训练集	测试集	综合准确率
Adaboost	0.751	0.773	0.762
GBDT	0.693	0.709	0.701
随机森林	0.803	0.835	0.754

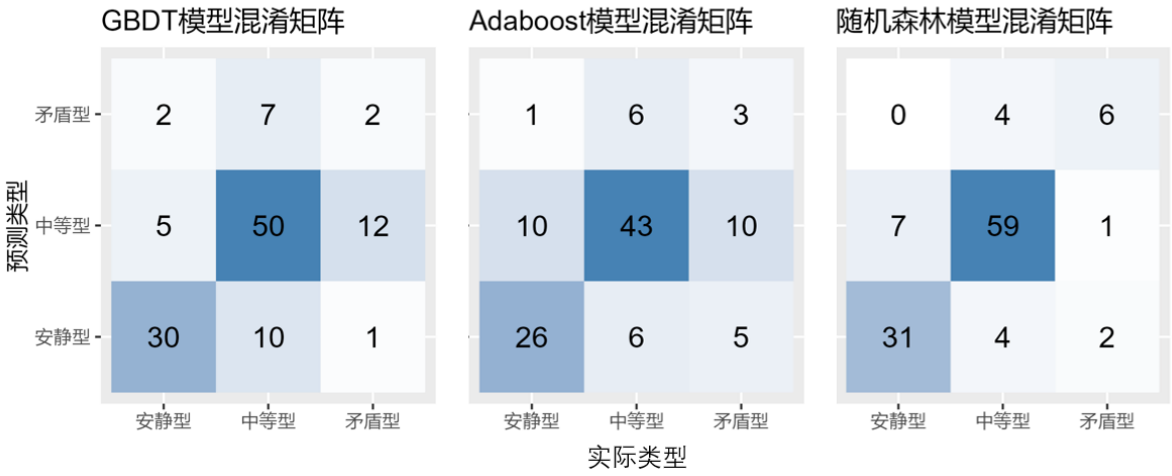

图 3 各模型混淆矩阵

从表中可以看出，训练集和测试集效果最接近的为随机森林，其余的分类器效果较为逊色。

因为各基分类器的分类效果不一样，所以本文选择软投票，软投票分类器采用加权融合方式。具体过程为：先确定好基分类器的权重w_i、基分类器的预测概率p_i，乘以基分类器的权重参数a_i再进行累加，最后选取累加最高的 P 为最终预测结果。

依据分类器在整体数据上的综合准确率来判断各基分类器的权重，公式为：

$$w_i = \frac{Accuracy_i}{\sum_{i=1}^{3} Accuracy_i} \tag{13}$$

$$P = w_1 p_1 + w_2 p_2 + w_3 p_3 \tag{14}$$

最终的软投票结果如表 5 所示，软投票模型测试集的混淆矩阵如图 4 所示：

表 5 软投票结果

分类器	训练集	测试集	综合准确率
Soft Voting	0.831	0.851	0.841

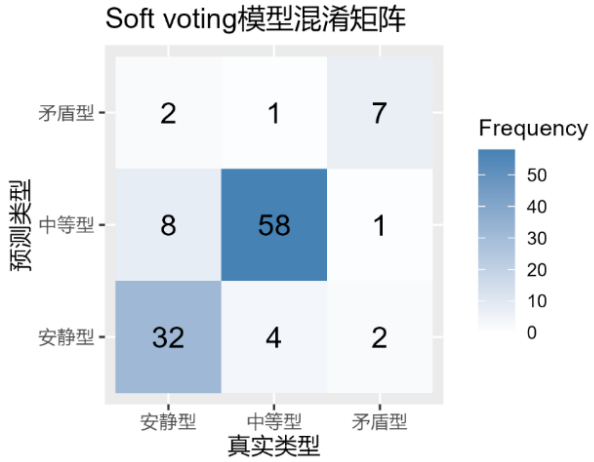

图 4 Soft Voting 模型测试集混淆矩阵

从表 5 和图 4 中可以看出，Soft Voting 模型在分类上的表现较好，高于之前的三个模型，准确率在 85%以上，说明集成学习算法在此数据集上有着较好的表现。

利用 Soft Voting 模型对编号为 391-410 号的婴儿的行为特征信息进行求解，得到三种情况的概率，选择概率最大值为婴儿的最终行为特征类型，结果表 6 所示。

表 6 婴儿行为特征预测

编号	预测安静型概率	预测中等型概率	预测矛盾型概率	最终类型
391	0.13	0.66	0.21	中等型
392	0.14	0.69	0.17	中等型
393	0.17	0.42	0.43	矛盾型
394	0.66	0.21	0.14	安静型
395	0.20	0.68	0.13	中等型
396	0.11	0.68	0.21	中等型
397	0.61	0.24	0.14	安静型
398	0.12	0.73	0.15	中等型
399	0.37	0.51	0.12	中等型
400	0.54	0.34	0.12	安静型
401	0.55	0.24	0.21	安静型
402	0.14	0.64	0.22	中等型
403	0.11	0.72	0.17	中等型
404	0.63	0.27	0.10	安静型
405	0.18	0.56	0.26	中等型
406	0.72	0.19	0.10	安静型
407	0.67	0.19	0.14	安静型
408	0.56	0.34	0.10	安静型
409	0.22	0.67	0.11	中等型
410	0.22	0.64	0.14	中等型

七、 问题三模型的建立与求解

7.1 建立模型前的数学分析

根据问题描述，我们假设 f(x)为 CBTS、 EPDS、 HADS 的治疗费用，x 为得分。根据 CBTS、EPDS、HADS 的治疗费用相对于患病程度的变化率均与治疗费用呈正比，又因为这里使用患病分数来衡量患病程度，即分数的变化与治疗费用也呈现正比关系。因此，我们可以得到如下关系式方程：

$$\frac{df(x)}{dx} = kf(x) \tag{15}$$

对方程两边同时进行积分，可以得到：

$$\ln f(x) = k^{*}x + C \tag{16}$$

$$f(x) = e^{kx+c} \tag{17}$$

$$f(x)=te^{kx} \tag{18}$$

根据问题 3 中已知的患病得分与治疗费用对应关系表，我们使 python 将给出的数据带入方程，得到对应的 t 值与 k 值。

$$cbts\ t=200.0\ k=0.88 \tag{19}$$

$$epds\ t=500\ k=0.66 \tag{20}$$

$$hads\ t=300\ k=0.75 \tag{21}$$

因此，我们将 t 和 k 值分别代入 f(x)，可以得到如下方程：

$$f(x_1)=te^{kx_1}=200e^{0.88x_1} \tag{22}$$

$$f(x_2)=te^{kx_2}=500e^{0.66x_2} \tag{23}$$

$$f(x_3)=te^{kx_3}=300e^{0.75x_3} \tag{24}$$

最终我们得到，当 238 编号得分分别为 15、22、18 时，治疗费用最小。所以可以得到治疗费用 w 的最终关系式方程：

$$w=200e^{0.88*15}-200e^{0.88x_1}+500e^{0.66*22}-500e^{0.66*x_2}+300e^{0.75*18}-300e^{0.75x_3} \tag{25}$$

因此问题 3 的问题可以转化为：如何找到对应的 x1、x2 和 x3，使得变为从矛盾型变为中等型，也即找到一个标准，即能满足为中等型，又能满足费用最小

7.2 MIMO 神经网络模型的建立

考虑到每位母亲的问卷得分与其身体指标有一定的关系，再计算可以改变婴儿特征的分数时，应考虑到母亲的身体特征进行预测，通过建立预测模型，可以更准确地预测在给定的身体指标下通过治疗改变婴儿特征后的问卷得分。如果我们能够准确地预测问卷得分，那么可以认为这种预测模型找到的方案就是使得费用最低的方案。

这种方法考虑到了母亲的个体差异和特征，通过将母亲的特征纳入预测模型中，我们可以更全面地了解母亲和婴儿之间的关系。相比于常规方法仅使用婴儿特征类型下各问卷的均值，这种方法更准确地考虑了母亲的影响因素，更加符合实际情况。

因此本文建立一个多输入多输出的 BP 神经网络模型来预测分数。

多输入多输出神经网络（Multi-Input Multi-Output Neural Network，简称 MIMO 神经网络）[5]是一种神经网络结构，它可以同时处理多个输入和多个输出。与传统的单输入单输出神经网络（SISO 神经网络）相比，MIMO 神经网络具有更强大的表达能力和灵活性。

在 MIMO 神经网络中，输入和输出可以是向量或矩阵。每个输入向量或矩阵对应一个输出向量或矩阵，因此可以有多个输入和多个输出之间的对应关系。神经网络的隐藏层可以具有不同数量的神经元，以适应不同输入和输出之间的复杂关系。其模型示意图如图 5 所示。

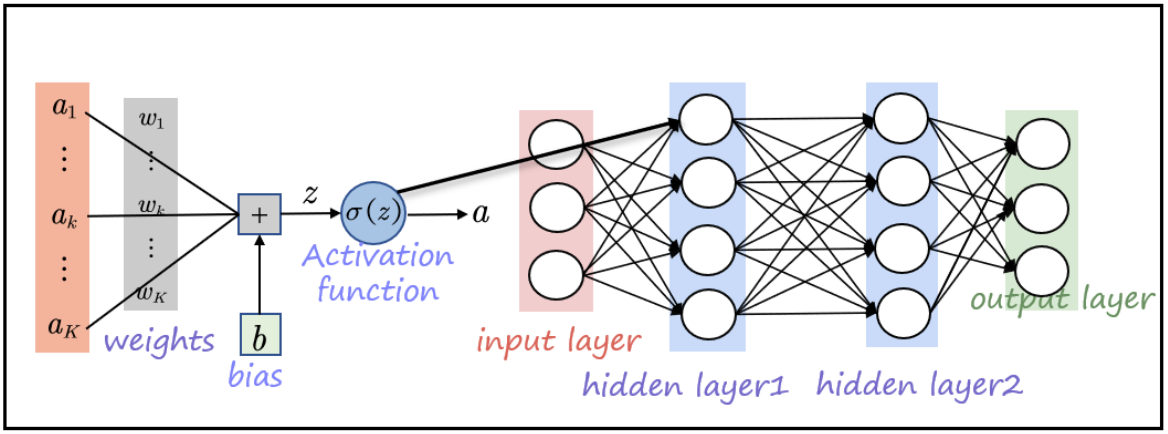

图 5 多输入多输出神经网络模型示意图

7.3 MIMO 神经网络模型的建立

我们使用 Matlab 建立 MIMO 神经网络模型并进行训练，可以发现进行到第 11 次时完成了迭代，此时拟合优度为 82.4%，可以认为模型拟合较好，结果如图 6 所示：

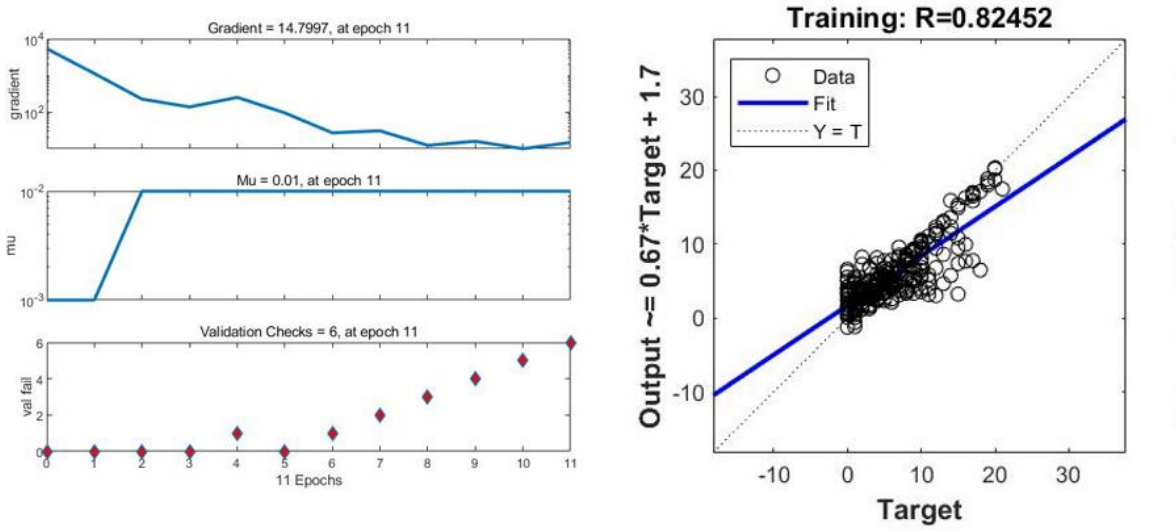

图 6 MIMO 神经网络模型运行结果

使用 matlab 对编号为 238 的婴儿，输入母亲年龄、妊娠周数、教育程度和特征类型(中等型)得到对应的 CBTS、EPDS 和 HADS 的得分分别为 11、19 和 18。带入到治疗费用公式得到，w=97664.67 元。

即最少需要花费 97664.67 元治疗费用，可以将 238 号婴儿的行为特征从矛盾型变为中等型。

同理若要使其行为特征变为安静型，输入母亲年龄、妊娠周数和教育程度，特征类型转为安静型时，得到对应的 CBTS、EPDS 和 HADS 得分分别为 5、12 和 16。带入到治疗费用公式得到，w=12881.34 元。

即最少需要花费 12881.34 元治疗费用，可以将 238 号婴儿的行为特征从矛盾型变为安静型。

八、 问题四模型的建立与求解

8.1 K-means 模型的建立

本问题需要以婴儿的整晚睡眠时间、睡眠次数和入睡方式为基础，将婴儿的睡眠质量进行优、良、中、差四分类综合评判。因此可以将本问题看作一个无监督学习问题，对于该问题的求解，我们可以选择 K-means 算法进行求解。

K 均值聚类（K-means clustering）是一种常用的无监督机器学习算法，用于将一组数据点划分为 k 个不同的簇（clusters）[6]。该算法通过最小化数据点与各自所属簇的簇中心之间的平方距离来进行聚类。K-means 最核心的部分就是先固定中心点，调整每个样本所属的类别来减少；再固定每个样本的类别，调整中心点继续减小。两个过程交替循环，单调递减直到最小值，从而使中心点和样本划分的类别同时收敛。其公式如下：

$$J(c,\mu)=\sum_{i=1}^{M}\|x_i-\mu_{ci}\|^2 \tag{26}$$

其中x_i代表第i个样本，c_i是x_i所属的簇，μ_{c_i}代表簇对应的中心点，M是样本总数。

K-means 的核心目标是将给定的数据集划分成 K 个簇，并给出每个样本数据对应的中心点，其具体步骤可分为以下 4 步，其模型步骤示意图如图 7 所示：

(1)数据预处理，标准化、异常值过滤

(2)随机选取 K 个中心，记为$\mu_1^{(0)},\mu_2^{(0)},\cdots,\mu_k^{(0)}$

(3)定义损失函数：$J(c,\mu)=\min\sum_{i=1}^{M}\|x_i-\mu_{ci}\|^2$

(4)令$t=0,1,2,\cdots$为迭代步数，重复如下过程直到J收敛：

a) 对于每一个样本x_i，将其分配到距离最近的中心，公式如下：

$$c_i^t\leftarrow arg\min_k\|x_i-u_k^t\|^2 \tag{27}$$

b) 对于每一个样本k，重新计算该类的中心，公式如下：

$$\mu_k^{(t+1)}\leftarrow arg\min_\mu\sum_{i-k}^{b}\|x_i-u\|^2 \tag{28}$$

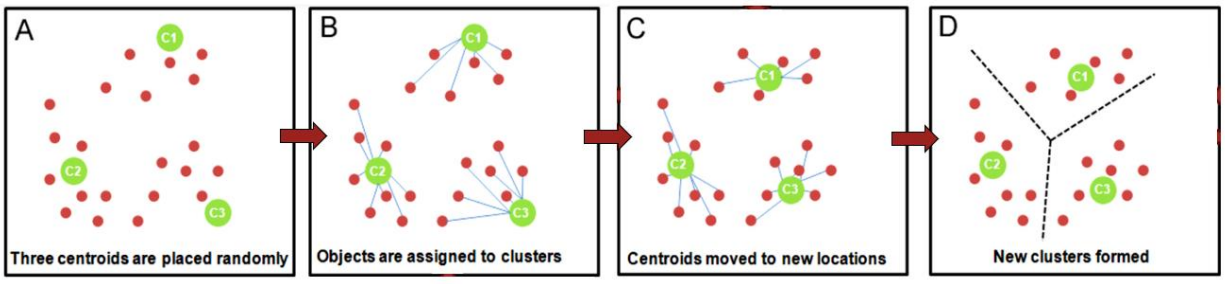

图 7 Kmeans 模型步骤示意图

8.2 K-means 模型的求解

我们通过 SPSS27 建立 K-means 模型，将睡眠时间、睡醒次数、入睡方式作为特征进行聚类，结果显示当迭代次数到达 17 次时，聚类中心中不存在变动，所有中心的最大绝对坐标变动为 0，因此认为此时模型到达收敛。此时的聚类中心如表 7 所示：

表 7 最终聚类中心

类别	睡眠时间	睡醒次数	入睡方式
1	8.00	5	2
2	9.66	3	4
3	9.68	1	1
4	11.26	0	4

在本题中，因为睡眠时间和睡醒次数可以作为评价睡眠质量好坏的主要指标，因此根据表中的聚类中心，可以根据其睡眠时间和睡醒次数去衡量睡眠质量的好坏，并进行评级如表 8 所示，婴儿的睡眠质量聚类分布如图 8 所示：

表 8 婴儿睡眠质量评级

类别	特征	评级
1	低睡眠时间~高睡醒次数	差
2	中睡眠时间~高睡醒次数	中
3	中睡眠时间~低睡醒次数	良
4	高睡眠时间~低睡醒次数	优

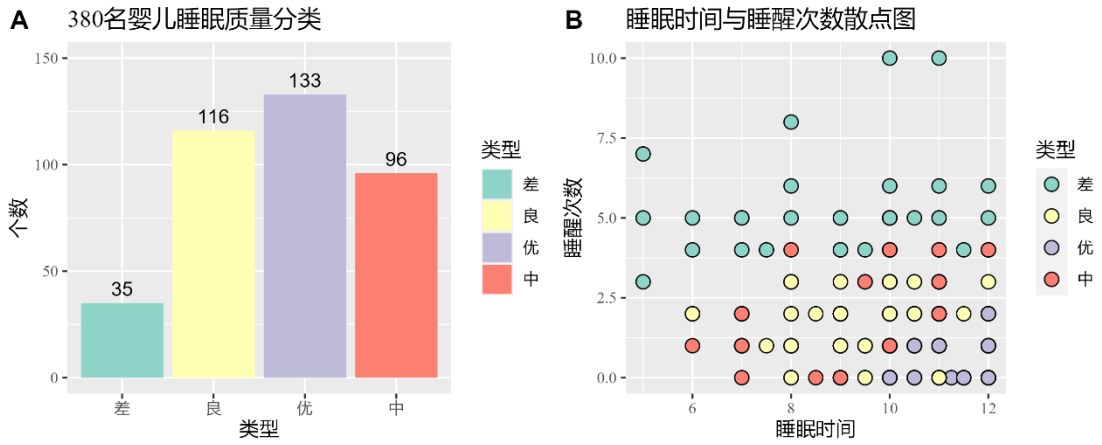

图 8 婴儿睡眠质量聚类分布

根据图 A，我们可以观察到不同睡眠质量级别的数量分布情况。图 B 则展示了一种趋势，即从左到右、从上到下的排列显示睡眠级别的增加。这种趋势验证了睡眠时间越长，睡醒次数越少，睡眠质量越高。

8.3 基于熵权-Topsis 模型验证 K-means 模型结果

为验证分类的合理性，我们使用问题二建立的熵权法-Topsis 模型计算睡眠指标，其中睡醒次数为极小型指标，使用公式转化为极大型指标并进行计算睡眠指标综合得分进行验证，计算出来的分数越高代表睡眠质量越好。

我们使用 MATLAB 求解，得到了睡眠指标的综合得分，其中编号前 10 号的得分如表 9 所示。

表 9 睡眠指标综合得分

编号	综合得分
1	0.71
2	0.88
3	0.95
4	0.84
5	0.82
6	1.00
7	0.76
8	0.76
9	0.76
10	0.88

然后我们使用 R 语言绘制得分散点图，以及各类型得分平均值如图 9 所示：

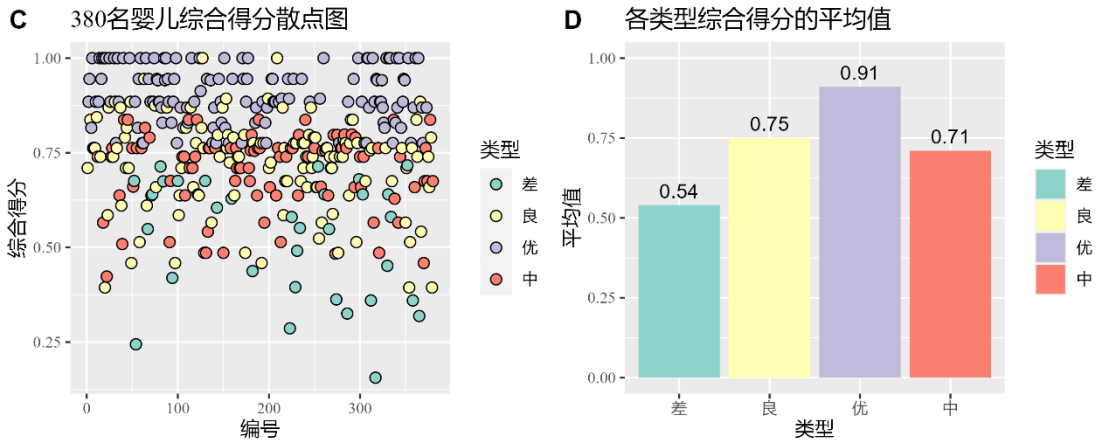

图 9 综合得分散点图及平均值

根据图 C，我们可以发现综合得分根据其类型（优、良、中、差）大致可以从上到下分为四类。此外，根据图 D 的结果，可以看出各类型的平均值呈现优>良>中>差的顺序。

这个观察结果与 K-means 分类的结果相一致，进一步验证了 K-means 分类的合理性。K-means 算法旨在将数据点划分为不同的簇，使得同一簇内的数据点相似度较高，而不同簇之间的数据点相似度较低。根据图 C 和图 D 的趋势，我们认为 K-means 分类的结果是有效、合理的。

最后我们通过 K-means 模型进行分类，我们得到了 380 名婴儿睡眠质量的评级，因此我们可以根据这些数据进行监督学习，使用问题二建立的 soft voting 模型进行训

练以及预测，模型混淆矩阵如图 10 所示，婴儿的综合睡眠质量结果如表 10 所示，模型准确率为 0.754。

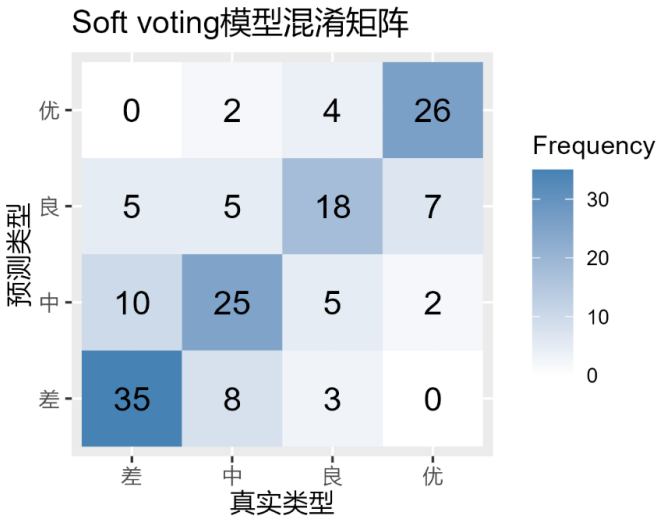

图 10 Soft voting 模型混淆矩阵

表 10 婴儿综合睡眠质量

编号	预测结果	编号	预测结果
391	中	401	中
392	良	402	差
393	差	403	中
394	优	404	良
395	优	405	优
396	优	406	优
397	中	407	中
398	良	408	差
399	差	409	良
400	优	410	优

九、 问题五模型的建立与求解

根据问题 4 求解的预测结果可以得知 238 号婴儿的综合睡眠质量评级为中。因此我们可以利用问题三建立的 MIMO 神经网络模型，将输入母亲年龄、妊娠周数、教育程度和睡眠质量评级作为输入，CBTS、EPDS 以及 HADS 得分作为输出，建立 MIMO 神经网络模型。

使用 matlab 建立模型进行训练，可以知道再进行到第 18 次时完成了迭代，此时拟合优度为 77.6%，认为模型拟合程度较好，模型的运行结果如图 11 所示。

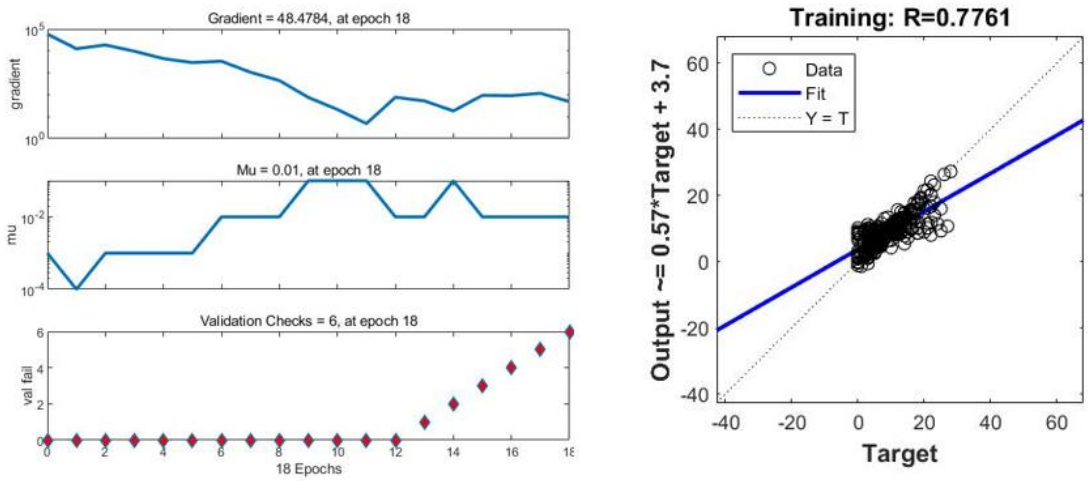

图 11 MIMO 神经网络模型运行结果

对于编号为 238 的婴儿，使用 Matlab 输入母亲年龄、妊娠周数、教育程度和睡眠质量评级为中，得到对应的 CBTS、EPDS 和 HADS 得分分别为 13、19 和 17。

同理，若要使其睡眠质量评级变为优，输入母亲年龄、妊娠周数、教育程度和睡眠质量评级为优，得到对应的 CBTS、EPDS和 HADS 得分分别为 5、12 和 16。带入到治疗费用公式得到，w=10767.28 元

即治疗策略调整为：CBTS、EPDS 和 HADS 得分从 13、19 和 17 调整为 5、12 和 16，最少需要花费的费用为 10767.28 元，可以将 238 号婴儿的睡眠质量评级从中变为优。

十、 模型的评价、改进与推广

10.1 模型的评价与改进

10.1.1 SEM 结构方程模型

优点：

(1)SEM 可以同时考虑观测变量之间的关系和潜在变量之间的关系，能够提供更全面的模型评估。

(2)可以用于验证理论模型，检验变量之间的因果关系。

(3)具备较好的可解释性，可以提供变量之间的直接和间接效应。

缺点：

(1)SEM 对于数据的要求较高，需要大样本量来保证模型的稳定性和准确性。

(2)SEM 的建模过程较为复杂，需要具备一定的统计和领域知识，在本文中只建立了简单的 SEM 模型。

10.1.2 Voting 投票模型

优点：

(1)Voting 模型可以结合多个基础模型的预测结果，从而提高整体的预测准确性和鲁棒性，在本文中 Voting 模型的准确率高于其余模型。

(2) 可以利用不同模型的优势，更好地应对不同类型的数据和问题。

缺点：

(1)Voting 模型可能受到弱分类器的影响，如果基础模型之间存在较大的差异，可能导致整体性能下降。

10.1.3 基于熵权法的 TOPSIS 模型

优点：

(1)TOPSIS 模型能够综合考虑多个指标的权重和绩效，提供全面的综合评价结果。

(2)熵权法能够考虑指标之间的相关性和权重分配的一致性，提高模型的可靠性。

缺点：

(1)TOPSIS 模型对于指标的权重敏感，权重的确定需要基于有效的权重分配方法。

10.1.4 K-means 聚类模型

优点：

(1)K-means 是一种简单易懂、易于实现的聚类算法，计算效率较高。

(2)可以自动发现数据中的类别和聚类中心，适用于无监督学习任务。

缺点：

(1)K-means 对初始聚类中心的选择敏感，初始值的选取可能导致不同的聚类结果。

10.1.5 MIMO 神经网络

优点：

(1) MIMO 神经网络能处理多输入和多输出之间的复杂关系，有较强的建模能力。

(2)可以同时处理多变量输入和输出问题，在许多应用领域中表现出很好的性能。

(3)可以通过大量的训练数据和网络设计来学习输入和输出之间的非线性映射。

缺点：

(1)MIMO 神经网络的训练和调优过程相对复杂，需要较长训练时间和计算资源。

(2)对于小样本数据集，可能出现过拟合，需要采取适当正则化和模型选择策略。

10.2 模型推广

10.2.1 Soft Voting 模型

使用多种机器学习算法以及软投票（soft voting）建立的模型推广到其他类似的研究领域。例如，可以将该模型应用于儿童或青少年的行为特征预测，以及与父母或监护人的身体指标和心理指标之间的关系模型建立。

10.2.2 MIMO 神经网络

本文使用 MIMO（Multiple-Input Multiple-Output）神经网络建立的模型用于预测母亲特征与问卷的关系，可以据此推广到其他相关研究领域。例如，可以将该模型应用于预测其他类型的问卷得分，如情绪评估、认知能力评估等。这可以帮助研究人员和临床医生更高效地评估个体的心理状态和认知能力水平。

10.2.3 K-means 聚类

本文使用 k-means 聚类算法对婴儿睡眠质量进行聚类并评级的模型，此模型适用性较广，可以将该模型应用于评估其他人群（如儿童、成人）的睡眠质量，并进行相应的评级。这对于睡眠研究、睡眠医学和睡眠障碍诊断等领域可能具有实际应用价值。

十一、 参考文献

[1]农雪艳,宋娟,朱锦渊等.母亲心理健康状况对婴儿体格和智能发育的影响[J].中国儿童保健杂志,2012,20(04):376-378.
[2]梁昌勇,代犟,朱龙.基于 SEM 的公共服务公众满意度测评模型研究[J].华东经济管理,2015,29(02):123-129.
[3]汪红,郭恒,武静云等.基于熵权-TOPSIS 法的发电企业低碳转型进程评价研究[J].热力发电,2023,52(07):26-32.DOI:10.19666/j.rlfd.202305057.
[4]张世文. 基于 Voting 策略的在线商品购买预测模型[D].湘潭大学,2022.DOI:10.27426/d.cnki.gxtdu.2021.001594.
[5]赵海阔. 基于深度学习的大规模 MIMO 检测算法研究[D].西安电子科技大学,2023.DOI:10.27389/d.cnki.gxadu.2022.000368.
[6]罗春芳,张国华,刘德华等.基于 Kmeans 聚类的 XGBoost 集成算法研究[J].计算机时代,2020(10):12-14.DOI:10.16644/j.cnki.cn33-1094/tp.2020.10.004.

附录

附录 1

R 语言 SEM 结构方程

```
library(tidyverse)
library(dplyr)
library(readxl)
library(writexl)
library(tidyr)
##读取数据
setwd("C:/desktop/2023 年 C 题")
data=read_xlsx("附件.xlsx")

#去除异常值
data=data[-180,1:15]
data=drop_na(data)
data=data[data$婚姻状况 <= 2, ]
#将睡眠时间转化为小时
sleep=data[,13]
datexpr2=as.data.frame(lapply(sleep,as.numeric))
data[,13]=datexpr2*24
#修改列名
colnames(data)[5]="妊娠时间"
colnames(data)[13]="睡眠时间"
#编码处理安静型为0，中等型为1，矛盾型为2

data = data %>%
  mutate(行为特征= case_when(
    婴儿行为特征 == "安静型"  ~ 0,
    婴儿行为特征 == "中等型" ~ 1,
    婴儿行为特征 == "矛盾型"  ~ 2
  ))


##创建 SEM 结构方程
install.packages("lavaan", dependencies = TRUE)  # 安装 lavaan 包
library(lavaan)   # 载入 lavaan 包


HS.model = ' 心理指标  =~ CBTS+EPDS+HADS
             身体指标  =~ Age+Edu+Ges
             行为特征 =~ 婴儿性别+婴儿年龄+特征类型
             睡眠质量 =~ 睡眠时间+睡醒次数+入睡方式
```

```
                行为特征~~心理指标
                睡眠质量~~心理指标
                行为特征~~身体指标
                睡眠质量~~身体指标'

#write_xlsx(cbind(mum1,mum2,baby1,baby2),"sem.xlsx")


fit1 <- sem(HS.model, data = cbind(mum1,mum2,baby1,baby2))
parameters <- parameterEstimates(fit1)
summary(fit1, fit.measure = TRUE)
##模型评估系数
fitMeasures(fit1,c("chisq", "df" ,
                   "pvalue" , "gfi", "cfi","rmr" , "srmr" , "rmsea")
  )

# 查看路径系数
path_coefficients <- parameters[parameters$op == "~", c("op", "lhs",
   "rhs", "estimate")]

library(semPlot)
## 加载 SEM 绘图包

pdf("plot2.pdf")
p2 = semPaths(fit, what = "std", layout = "circle", fade=F, nCharNod
  es = 0)
pdf("p2.pdf",width = 5,height = 6,family="GB1")
library(eoffice)
```

附录 2

Soft Voting　R 语言

```
library(adabag)
library(gbm)
library(randomForest)
# 训练 Adaboost 模型
adaboost_model <- boosting(formula, data = train_data, mfinal = 10)
# 训练 GBDT 模型
gbdt_model <- gbm.fit(x = train_data[, predictors], y = train_data$t
  arget, n.trees = 100, shrinkage = 0.1)
# 训练随机森林模型
rf_model <- randomForest(x = train_data[, predictors], y = train_dat
  a$target, ntree = 100)
```

```
# 定义 Soft Voting 模型
softvoting_model <- function(newdata) {
  adaboost_pred <- predict(adaboost_model, newdata, type = "response")
  gbdt_pred <- predict.gbm(gbdt_model, newdata, n.trees = 100, type = "response")
  rf_pred <- predict(rf_model, newdata, type = "response")
  # 对三个基分类器的预测结果进行投票
  voting_result <- ifelse(adaboost_pred + gbdt_pred + rf_pred >= 2, 1, 0) # 假设是二分类问题
  return(voting_result)
}
# 使用 Soft Voting 模型进行预测
predictions <- softvoting_model(test_data)
# 对预测结果进行评估
accuracy <- sum(predictions == test_data$target) / length(test_data$target)
```

附录 3

Matlab 神经网络

```
filename = 'C:/desktop/bp/bp 神经网络.xlsx';
data = xlsread(filename);
input_data = data(:, 1:4);  % 输入数据的列索引根据实际情况进行调整
output_data = data(:, 6);  % 输出数据的列索引根据实际情况进行调整
% 定义 BP 神经网络
hidden_size = 230;  % 隐藏层大小，根据需要进行调整
net = feedforwardnet(hidden_size);
% 设置训练参数
net.trainParam.epochs = 1000;  % 训练迭代次数，根据需要进行调整
net.trainParam.lr = 0.1;  % 学习率，根据需要进行调整
% 训练神经网络
net = train(net, input_data', output_data');
```

附录 4

Python 计算比例系数

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
# 给定的数据点
data = np.array([[0, 200], [3, 2812]])
```

```
#data = np.array([[0, 500], [2, 1890]])
#data = np.array([[0, 300], [5, 12500]])
# 提取数据点的 x 和 y 值
x = data[:, 0]
y = data[:, 1]
# 定义拟合函数
def func(x, t, k):
    return t * np.exp(k * x)
# 进行拟合
popt, pcov = curve_fit(func, x, y)
# 提取拟合参数的值
t = popt[0]
k = popt[1]
# 输出拟合参数的值
print("t =", t)
print("k =", k)
# 绘制原始数据和拟合曲线
plt.scatter(x, y, color='red', label='原始数据')
x_fit = np.linspace(min(x), max(x), 100)
y_fit = func(x_fit, t, k)
plt.plot(x_fit, y_fit, color='blue', linewidth=1.5, label='拟合曲线
')
plt.legend()
plt.xlabel('x')
plt.ylabel('f(x)')
plt.show()
```

附录2

R 语言第四问睡眠质量绘图

```
library(tidyverse)
library(dplyr)
library(readxl)
library(writexl)
library(tidyr)
##读取数据
setwd("C:/desktop/第四问")
data=read_xlsx("聚类后.xlsx")
data %>%
  count(分类)
data1 = data %>%
  mutate(类型= case_when(
    分类 == 1  ~ "差",
```

```
    分类 == 2  ~ "中",
    分类 == 3  ~ "良",
    分类 == 4  ~ "优",
  ))
##统计每一组的平均值
a=data1 %>%
  group_by(类型) %>%
  summarise(
    最小值 = min(综合得分),
    最大值 = max(综合得分),
    平均值 = mean(综合得分)
  )
##综合得分和类型
df=data1[,5:6]
library(ggplot2)
b=df %>%
  group_by(类型) %>%
  count
#380 名婴儿睡眠质量分
p1=ggplot(b, aes(x = 类型, y = n, fill = 类型)) +
  # 绘制柱状图
  geom_bar(stat = "identity") +
  geom_text(aes(label=n),vjust=-0.5)+
  labs(x="类型",y="个数",title="380 名婴儿睡眠质量分类")+
  scale_fill_brewer(palette = "Set3")+
  theme(text=element_text("serif"))+
  ylim(0,150)
##睡眠时间和睡眠次数
df1=data1[,c(1,2,6)]
p2=ggplot(df1, aes(x = 睡眠时间, y = 睡醒次数, fill = 类型)) +
  geom_point(shape = 21, size = 3.3) +
  labs(x = "睡眠时间") +
  labs(y = "睡醒次数") +
  labs(title = "睡眠时间与睡醒次数散点图") +
  scale_fill_brewer(palette = "Set3")+
  theme(text=element_text("serif"))
library(cowplot)
cowplot::plot_grid(
  p1,
  p2,
  labels = c("A", "B")
)
ggsave("分类统计.png",width=8.4,height = 3.4,dpi=300)
p3=ggplot(df,aes(x=1:380,y=综合得分,fill=类型))+
```

```
  geom_point(shape = 21, size = 2.5) +
  labs(x = "编号") +
  labs(y = "综合得分") +
  labs(title = "380 名婴儿综合得分散点图") +
  scale_fill_brewer(palette = "Set3")+
  theme(text=element_text("serif"))
a[, 4] <- round(a[, 4], 2)
a[2,4]=0.91
# 基于数据创建一个绘图对象
p4=ggplot(a, aes(x = 类型, y = 平均值,fill=类型)) +
  # 绘制柱状图
  geom_bar(stat = "identity") +
  geom_text(aes(label=平均值),vjust=-0.5)+
  xlab("类型") +
  ylab("平均值") +
  ggtitle("各类型综合得分的平均值") +
  scale_fill_brewer(palette = "Set3")+
  theme(text=element_text("serif"))+
  ylim(0,1)
cowplot::plot_grid(
  p3,
  p4,
  labels = c("C", "D")
)
ggsave("综合得分统计.png",width=8.4,height = 3.4,dpi=300)
```