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		CM2301248

母亲身心健康对婴儿成长的影响分析模型

摘要

本文研究如何根据母亲的身体指标和心理指标数据，结合婴儿的行为特征与睡眠质量两方面数据建立母亲身心健康对婴儿成长影响的分析模型，并利用此模型对婴儿行为特征进行预测以及相应治疗方案的制订。

针对问题一，为解决母亲身心指标数据对婴儿的行为特征和睡眠质量的关系，本文采用了**斯皮尔曼相关性分析**研究了数据之间的相关性，并通过**卡方检验**，寻找数据的差异性。接着，应用**灰色关联分析方法**，进一步探索相关指标数据与婴儿的行为特征和睡眠质量的关系。最终得到母亲的指标数据对婴儿行为特征与睡眠质量的特征重要性数据，证明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存在一定的影响。

针对问题二，我们先对附件中婴儿的行为特征已有数据进行数据编码，然后利用问题一得到的母亲身心指标对婴儿行为特征影响较为显著的变量，我们构建出可解释性强的**决策树分类模型**对婴儿的行为特征进行分类和预测，得到数据表中最后 20 组（编号 391-410 号）婴儿的行为特征信息。

婴儿编号	行为特征	婴儿编号	行为特征	婴儿编号	行为特征	婴儿编号	行为特征	婴儿编号	行为特征
391	**中等型**	392	**安静型**	393	**矛盾型**	394	**中等型**	395	**中等型**
396	**中等型**	397	**中等型**	398	**中等型**	399	**中等型**	400	**中等型**
401	**中等型**	402	**中等型**	403	**中等型**	404	**安静型**	405	**中等型**
406	**安静型**	407	**中等型**	408	**安静型**	409	**安静型**	410	**中等型**

针对问题三，我们根据题目所提供的治疗方案价格表，结合问题一中决策树所得到的对婴儿行为特征影响较大的几种变量，以及在题目要求所需治疗费用最小的约束条件下，构建线性规划模型，然后对该线性模型进行求解。最终得到将编号为 238 的行为特征**从矛盾型转变为中等型所需的最少治疗费用为 5252 元，从矛盾型转变为安静型所需的最少治疗费用为 7354 元。**

针对问题四，我们团队利用美国睡眠医学会（AASM）关于婴儿睡眠质量的评判标准对婴儿的睡眠质量进行按优、良、中、差四分类综合评判，并且将所得综合评判的结果进行数据编码后，运用鲁棒性强、可处理多分类问题的决策树分类模型对婴儿睡眠质量的综合评判结果进行分类，分析出婴儿综合睡眠质量与母亲的身体指标、心理指标之间的关联。最后，利用所构建的决策树分类模型，对附件数据表中最后 20 组（编号 391-410 号）婴儿的综合睡眠质量进行预测，得到相应的预测结果，具体结果已在原文展示。

针对问题五，我们运用问题四中所构建的决策树模型求得母亲身体指标和心理指标中各指标对婴儿睡眠质量综合评判的特征重要性强弱，并将其修正相对应的线性规划模型，最终得到在基础三的基础上，将 238 号婴儿的睡眠质量评级从良转变为优且转变为中等型所需的最少治疗费用为 8978 元，睡眠质量评级从良转变为优且转变为安静型所需的最少治疗费用为 12570 元。

关键词：**斯皮尔曼相关性分析 卡方检验 灰色关联分析 决策树分类模型**

一、问题重述

1.1 问题背景

母亲是把婴儿带到世上的使者，是婴儿生命的给予者。母亲不仅是婴儿的庇护所，为婴儿提供身体保护和营养物质，还是婴儿的情感启蒙的第一人，为婴儿提供情感支持和安全感。母亲心理的健康与否，可能会对婴儿的成长造成一定的影响。现有关于母亲与婴儿的一些已知相应指标与对应得分的数据，需要解决以下问题。

1.2 题目要求

问题一：根据附件中所提供的数据进行研究，分析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是否有关联。

问题二：在问题一分析的基础上，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，并根据所建立的模型对附件中最后 20 组缺失的婴儿的行为特征信息进行判断。

问题三：根据对母亲的 CBTS、EPDS、HADS 进行相关治疗的调研信息，并结合编号为 238 的行为特征为矛盾型的婴儿相应数据，建立相应的模型，分析最少需要花费多少治疗费用，能够使婴儿的行为特征从矛盾型变为中等型。

问题四：整晚睡眠时间、睡醒次数、入睡方式都是评价婴儿睡眠质量的指标，根据这些评价指标对婴儿的睡眠质量进行优、良、中、差四分类综合评判，并通过建立婴儿综合睡眠质量与母亲的身体指标、心理指标的关联模型，预测最后 20 组婴儿的综合睡眠质量。

问题五：在问题三将编号为 238 的婴儿行为特征从矛盾型变为中等型的基础上，若需让该婴儿的睡眠质量评级为优，问题三的治疗策略是否需要调整，若需要，又应该如何调整。

二、问题分析

2.1 问题一分析

问题一可分为两步处理。第一步，根据附件中所提供的数据，我们需要得到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关系。此关系包含相关性关系和差异性关系。考虑到母亲的婚姻状况、教育程度、分娩方式、婴儿的行为特征都为定类变量，不能采用分析连续型变量的皮尔逊相关分析和方差分析等方法。因此，在进行相关性分析时，可以采用针对定类变量的斯皮尔曼相关系数进行分析；在进行差异性分析时，则可以采用卡方检验和效应量化分析的方法。

第二步，根据第一步所得到的相关性大小，深入分析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之间的规律。

2.2 问题二分析

对于问题二，由第一问的分析结果可得到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之间的规律，基于此规律，可建立可解释性强、适用性广泛的决策树模型分析婴儿的行为特征与母亲的身体指标与心理指标的关系，并利用所建立的决策树模型对数据表中最后 20 组婴儿的行为特征信息进行预测。

2.3 问题三分析

对于问题三，我们可以先根据题目中的治疗方案的价格表，列出对应的微分方程。然后利用第二问中决策树所求得的母亲身心指标与心理指标对婴儿行为特征的影响特征重要性，由此构建关于治疗费用的线性规划模型，然后对模型进行求解，得到能够使婴儿的行为特征从矛盾型变为中等型的最少的治疗费用。

2.4 问题四分析

问题四需要根据婴儿的睡眠质量指标数据进行综合评判，并将其分成优、良、中、差四个类别，具体的分类标准可以参考美国睡眠医学会（AASM）的分类标准。在此基础上，可以通过应用可处理不相关特征、简单易于理解的决策树模型来建立婴儿综合睡眠质量与母亲的身体指标、心理指标的模型，并用于预测编号为 391-401 号婴儿的综合睡眠质量。

2.5 问题五分析

问题五需要求在问题三的基础上，编号为 238 号婴儿的睡眠质量评级为优所需花费费用最少的治疗方案。而这一求最小化结果属于线性规划问题，可以采用问题三的解决思路，利用问题四所构建的决策树模型可以得知母亲的身体指标与心理指标与婴儿的睡眠质量综合评判的特征重要性，并以此为基础修改原有的线性规划模型从而得到相应的调整后的方案。

三、模型假设

假设一：假设除母亲的身体指标和心理指标外，不存在其他物理和化学作用对婴儿的行为特征和睡眠质量有影响。

假设二：假设 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的关系不因治疗费用的不断增加而显著下降，即单位价格的治疗有效性不会下降。

假设三：假设婚姻状况处的大于 2 的情况表示该情况未知。

四、符号说明

符号	定义	单位
ρ	斯皮尔曼相关系数	/
x'	归一化后的数据	/
d_x	得分的变化率	/
Y	治疗费用总和	/
k	治疗费用相对于患病程度变化率的正比系数	/
σ	样本数据总和	/
μ	样本数据均值	/

（注：未列出符号及重复符号以出现处为准）

五、问题一模型

问题一要求分析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之间是否存在影响。首先需要进行数据预处理，方便后续模型的建立。附件中所提供的数据既有定类数据，也有定量数据，因此我们使用斯皮尔曼相关性分析和卡方检验分析不同变量之间的关系，在此基础上，再运用灰色关联分析方法，分析具体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关联度，并且依据所得数据结果分析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之间是否存在影响。

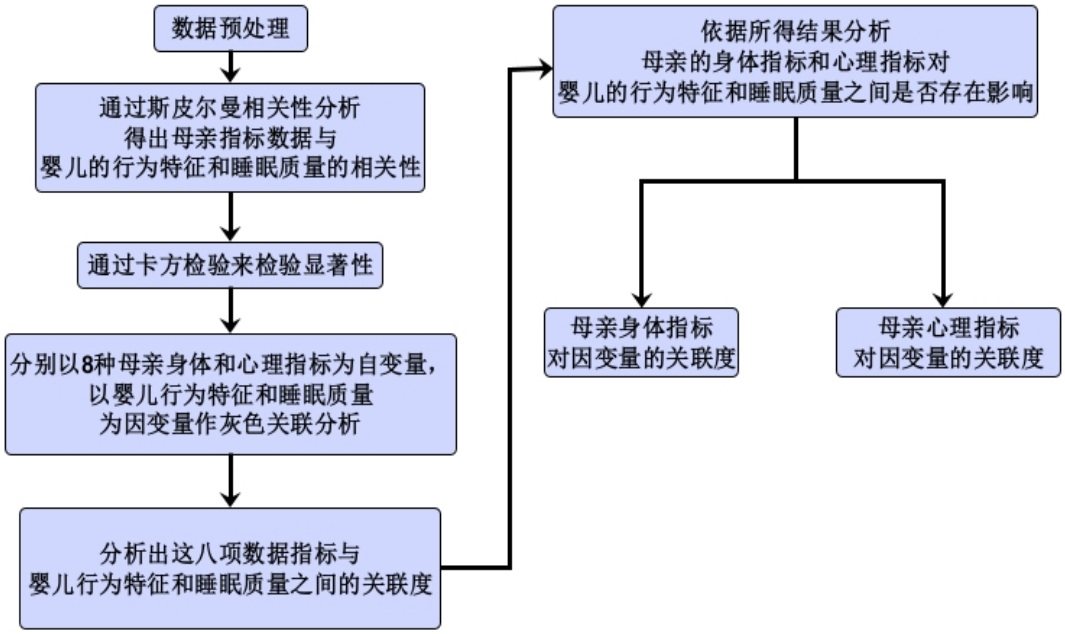

图 1 问题一流程图

5.1 数据预处理

通过对附件数据的分析，我们可以发现表单中部分母亲的身体指标数据存在异常值。为方便后续模型的建立，我们对附件数据进行预处理。

5.1.1 缺失值处理

由题目可知，附件的数据表中最后 20 组（编号为 391-410 号）婴儿的行为特征信息被删除，故为缺失值。另外，该 20 组的睡眠质量指标也是需要建立模型进行预测，没有提供相应的数据，故为缺失值。

5.1.2 异常值处理

通过对母亲的身体指标和心理指标进行分析，发现母亲的婚姻情况栏出现未标识的类别，由于此类数据本身是确定性答案，并且无法通过插值修改的方法进行修改，因此对于求解过程没有实际意义，我们将这一小部分数据进行剔除。另外，编号为 180 的数据中的睡眠时间显示为 99:99,显然是一个无效数据，故将其剔除。

5.1.3. 数据归一化

为了归纳统一样本的统计分布性，并消除奇异样本导致的不良影响，本文采用均值方差归一化的方法对数据进行预处理。归一化公式如下式所示：

$$x' = \frac{x_i - \mu}{\sigma} \tag{1}$$

其中，x'表示归一化后的数据，x_i为单个样本数据，μ为样本数据均值，σ为样本数据方差。

5.2 数据预分析

通过对附件中的数据进行分析，可以发现母亲的身体指标包括年龄、婚姻状况、教育程度、妊娠时间、 分娩方式，以及产妇心理指标 CBTS（分娩相关创伤后应激障碍问卷）、EPDS （爱丁堡产后抑郁量表）、HADS（医院焦虑抑郁量表）和婴儿睡眠质量指标包括整晚睡眠时间、睡醒次数和入睡方式。

在数据预处理之后，对附件中的数据进行分析，我们可以采用能够反应数据集中趋势的数据均值对不同婴儿的行为特征所对应的母亲的身体指标与心理指标进行分析，得到如表 1 的结果。

表 1 不同婴儿行为特征类型
对应的母亲身体指标与心理指标的均值情况

婴儿行为特征	母亲年龄	婚姻状况	教育程度	妊娠时间	分娩方式	CBTS	EPDS	HADS
中等型	30.050	1.959	4.131	39.077	1.018	6.217	9.267	8.045
安静型	30.905	1.966	4.026	39.222	1.009	5.353	8.190	7.259
矛盾型	29.409	1.977	4.114	39.116	1.000	6.591	11.023	8.750

将上表结果可视化，我们可以得到图 2 所示结果：

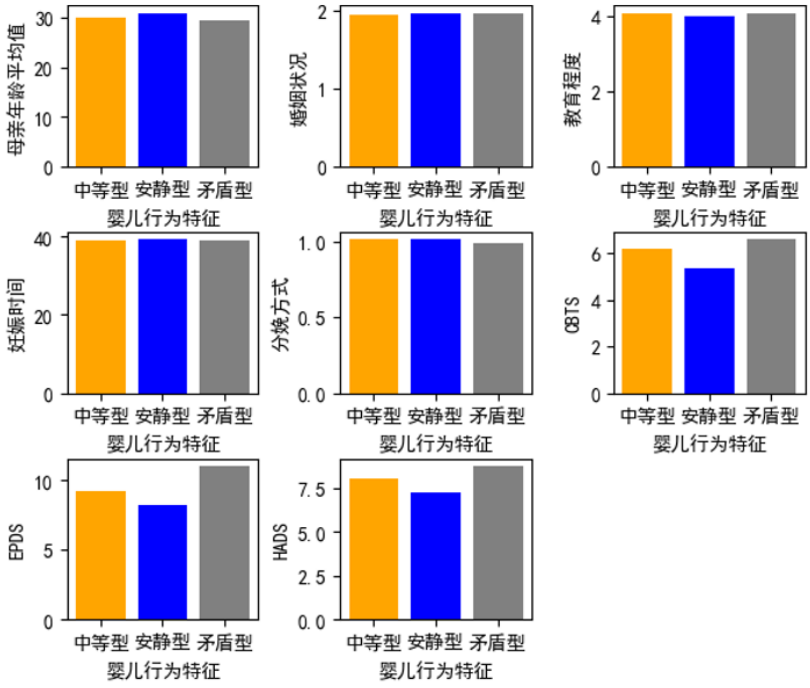

图 2 不同婴儿行为特征类型对应的
母亲身体指标与心理指标的均值分布情况

通过对表 1 的数据进行分析，我们可以发现，不同婴儿的行为特征对应母亲的大部分身体指标和心理指标的数据变化并不显著，但也有例外，如反映母亲心理指标数据的 CBTS、EPDS、HADS 指标与婴儿行为特征之间存在较大的关联，可直观地得出婴儿行为特征的规律：**婴儿行为特征对应的母亲心理指标沿矛盾型、中等型以及安静型的顺序递减。矛盾型的三项数据指标得分明显较高，而安静型的三项数据指标得分明显较低。**

5.3 基于斯皮尔曼相关系数与卡方检验的婴儿行为特征与睡眠质量分析模型

由于表单数据类型为定类型数据，需要分别衡量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，以及各项指标之间的关联程度，故本文采用斯皮尔曼相关性分析和卡方检验，分别得出不同影响因素之间的相关性和差异性。

5.3.1 斯皮尔曼相关性分析

斯皮尔曼相关系数是一个衡量两个变量之间依赖性的非参数指标，该系数对两个数据集是否属于同分布没有要求。斯皮尔曼相关系数分布范围从 −1 到 +1。当其取值为 0 时，表示两个参数之间没有相关性。当数据不重复，且两个变量完全单调相关时，斯皮 尔曼相关系数为 +1 或 −1。当自变量和因变量变化方向相同时，斯皮尔曼相关系数为 正，反之，当自变量和因变量变化方向不同时，斯皮尔曼相关系数为负。

斯皮尔曼相关系数定义如式 (2) 所示：

$$\rho = 1 - \frac{6d_i^2}{n * (n^2 - 1)} \tag{2}$$

其中d_i表示第i个数据对的位次值之差，n表示婴儿样本的总数目。

通过 SPSSPRO 软件计算斯皮尔曼相关系数，得到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之间的相关性大小，如图 3 所示：

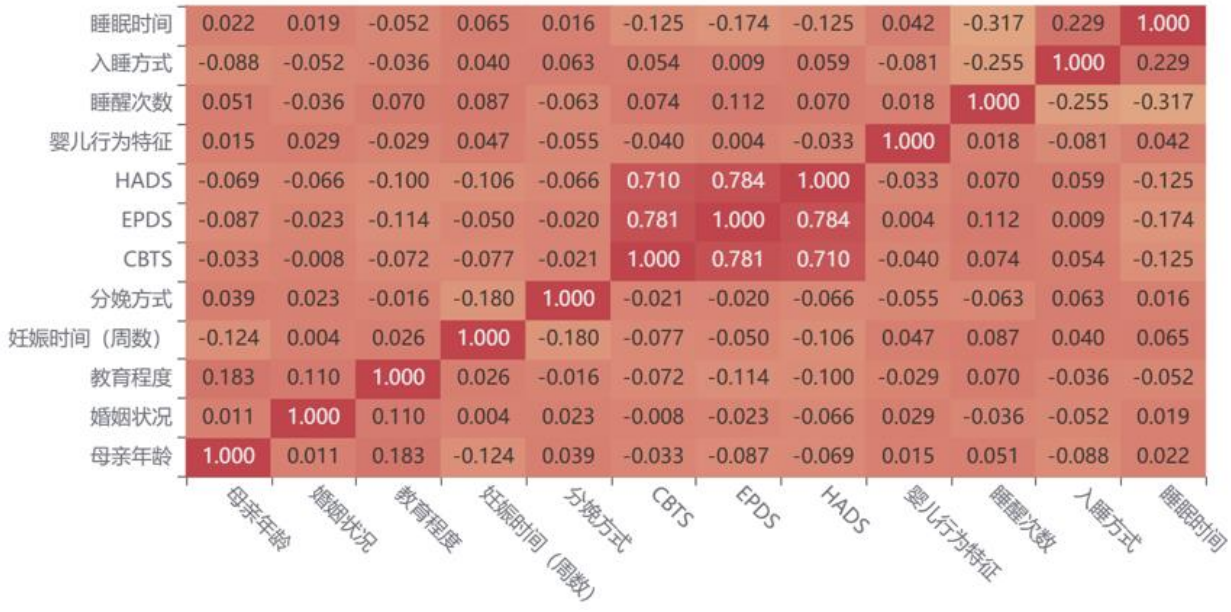

图 3 斯皮尔曼相关系数热力图

通过比较ρ值的大小，可以发现，与婴儿的行为特征相关性最大的是分娩方式，相关性系数为$\rho_{分娩方式}$=-0.05，其次是妊娠时间：$\rho_{妊娠时间} = 0.047$；最后是 EPDS 的数据，$\rho_{EPDS} = 0.004$。

5.3.2 卡方检验

为分析母亲的身体指标和心理指标之间的差异性，我们采用卡方检验进行分析。卡方值的大小由实际观测值与理论推断值之间的偏离程度决定，二者偏差程度越大，则卡方值越大；反之，卡方值越小。如两值完全相等，则卡方值为 0。

首先假设变量之间相互独立，并计算期望频次，代入卡方统计公式计算卡方值，公式如下式所示：

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{3}$$

其中x^2为卡方值，f_0为观测频率，f_e为期望频率。在此基础上，计算自由度。通过查表，得到并比较卡方值或P值。通过比较P值，可以得到母亲的身体指标和心理指标之间的关联程度。

通过 *SPSSPRO* 对附件数据进行处理，得到的卡方检验如表 2 所示：

表 2 卡方检验结果

题目	名称	婴儿行为特征			总计	X²	校正 X²	P
		1.0	2.0	3.0				
婚姻状况	1.0	9	4	1	14	0.368	0.368	0.832
	2.0	211	112	43	366			
教育程度	1.0	0	1	1	2	5.406	4.626	0.713
	2.0	12	7	1	20			
	3.0	54	31	11	96			
	4.0	47	26	10	83			
	5.0	107	51	21	179			
分娩方式	1.0	216	115	44	375	1.198	1.198	0.549
	2.0	4	1	0	5			
注：***、**、*分别代表 1%、5%、10%的显著性水平								

通过上述的卡方检验结果可以得知，对于母亲的婚姻状况与婴儿的行为特征，其显著性 P 值为 0.832，大于 0.05，接受原假设，不存在显著性差异。对于母亲的教育程度与婴儿的行为特征，其显著性为 0.713，大于 0.05，故接受原假设，不存在显著性差异。对于分娩方式与婴儿的行为特征，其显著性为 0.549，大于 0.05，故接受原假设，不存在显著性差异。

5.4 基于灰色关联分析的化学成分与风化分析模型

根据附件中表单的数据可知，婴儿的行为特征与睡眠质量受多种因素的影响，且每一影响因素的影响效果并不一致，任意数据的变动，都可能会引起婴儿的行为特征与睡眠质量。

为衡量附件所提供的数据变量对婴儿的行为特征与睡眠质量的影响程度的大小，本文采用对样本数量的多少和样本有无规律并无要求的灰色关联分析法，对表单中的数据进行处理，得到母亲的身体指标与心理指标与婴儿行为特征和睡眠质量影响程度排名。

5.4.1 灰色关联分析

灰色关联分析的步骤如下所示：

(1) 确定比较数列和参考数。其中反映系统行为特征的称为参考数列和影响系统行为的称为比较数列。在该模型中，玻璃是否风化得到参考数列，所有的化学成分比例得到比较数列。其中，比较数列如式（4）所示：

$$a_i = a_i(j) | j = 1,2,\dots,n, i = 1,2,\dots,m \tag{4}$$

其中，m 为评价对象的数目，题中评价的对象为是否风化。n 为评价指标的数目，题中 的评价指标为各化学成分的比例。

(2) 计算灰色关联系数。根据以上确定的数列，计算关联分析系数，该系数即决定了不同化学组成成分对于是否风化影响程度的大小。灰色关联分析计算公式如式（5）所示：

$$\xi(j) = \frac{\min\limits_{1\le s\le m}\min\limits_{1\le t\le n}|a_0(t)-a_s(t)| + \rho\max\limits_{1\le s\le m}\max\limits_{1\le t\le n}|a_0(t)-a_s(t)|}{\rho\max\limits_{1\le s\le m}\max\limits_{1\le t\le n}|a_0(t)-a_s(t)| + |a_0(j)-a_s(j)|}, \tag{5}$$
$$i = 1,\dots,m, j = 1,\dots,n.$$

$\xi(j)$为比较数列a_i对参考数列a_0在第j个指标上的关联系数，其中$\rho \in [0，1]$为分辨系数。其中，$\min\limits_{1\le s\le m}\min\limits_{1\le t\le n}|a_0(t)-a_s(t)|$、$\max\limits_{1\le s\le m}\max\limits_{1\le t\le n}|a_0(t)-a_s(t)|$分别称为两极最小差及两极最大差。一般来讲，分辨系数ρ越大，分辨率越大；ρ越小，分辨率越小。

5.4.2 模型求解

通过 *SPSSPRO* 软件，以婴儿行为特征作为母序列，8 种母亲身心数据指标作为子序列，得到灰色关联度并排序，如表 3 所示 (完整结果参见支撑材料):

表 3 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灰色关联分析结果

	婴儿行为特征关联度结果		睡醒次数关联度结果		入睡方式关联度结果		睡眠时间关联度结果	
评价项	关联度	排名	关联度	排名	关联度	排名	关联度	排名
母亲年龄	0.796	1	0.815	3	0.782	5	0.902	4
教育程度	0.791	2	0.82	1	0.785	4	0.86	5
妊娠时间（周数）	0.79	3	0.815	2	0.788	1	0.925	2
婚姻状况	0.787	4	0.814	4	0.786	3	0.921	3
分娩方式	0.786	5	0.813	5	0.786	2	0.926	1
HADS	0.747	6	0.806	6	0.76	6	0.764	6
EPDS	0.715	7	0.804	7	0.714	7	0.709	7
CBTS	0.694	8	0.794	8	0.709	8	0.698	8

5.4.3 结果分析

根据表 3 所示结果，我们可以发现，利用灰色关联分析得到的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之间存在影响，并且有较高的关联度（处于 0.7 左右），有着较为明显的关系。同时也说明构建婴儿的行为特征与睡眠质量预测模型时，母亲的身体指标和心理指标都是相应需要考虑的变量因素。

六、问题二模型

6.1 问题分析

该问题由两部分组成，一是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，二是通过所建立的模型对数据附件种婴儿行为特征缺失的数据进行预测，判断其所属的类型。

关于建立婴儿行为特征与母亲的身体指标与心理指标的关系模型，首先需要分析出母亲各身体指标与心理指标与婴儿的行为特征的相关性强弱，而在第一问已经得到母亲各身体指标与心理指标与婴儿的行为特征相关性信息。可以在第一问的基础上，筛选出相关性较强的指标用于关系模型的构建。

关于对缺失的婴儿行为特征类型缺失数据的预测则可以通过建立具有良好的表达能力和泛化能力，广泛用于分类、回归等问题解决的 GBDT 预测模型对缺失的婴儿行为特征数据进行预测。

6.2 数据编码

为方便后续模型建立，现将婴儿行为特征进行数据编码，编码的标准如表 4 所示：

表 4 婴儿行为特征数据编码标准图

类型	编码
中等型	1
安静型	2
矛盾型	3

6.2 模型建立

GBDT（Gradient Boosting Decision Tree），全名叫梯度提升决策树，是一种迭代的决策树算法。其通过构造一组弱的学习器（树），并把多棵决策树的结果累加起来作为最终的预测输出。

GBDT 每次都以当前预测为基准，下一个弱分类器去拟合误差函数对预测值的残差（预测值与真实值之间的误差）。

GBDT 与负梯度近似残差

回归任务下，GBDT 在每一轮的迭代时对每个样本都会有一个预测值，此时的损失函数为均方差损失函数l：

$$l(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2 \tag{6}$$

损失函数的负梯度计算如下：

$$-\left[\frac{\delta l(y_i - \hat{y}_i)}{\delta \hat{y}_i}\right] = (y_i - \hat{y}_i) \tag{7}$$

梯度下降中，模型是以参数化形式表示，从而模型的更新等价于参数的更新。

$$\omega_t = \omega_{t-1} - \rho_t \nabla_\omega L|_{\omega=\omega_t-1} \tag{8}$$

$$L = \sum_i l\big(y_i, f_\omega(\omega_i)\big) \tag{9}$$

梯度提升中，模型并不需要进行参数化表示，而是直接定义在函数空间中，从而大大扩展了可以使用的模型种类。

$$F = F_{t-1} - \rho_t \nabla_F L|_{F=F_t-1} \tag{10}$$

$$L = \sum_i l\big(y_i, F(x_i)\big) \tag{11}$$

GBDT 模型的建立步骤：

（1）初始化模型：选择一个简单的基础模型作为起始模型，通常采用决策树。

（2）计算残差：使用当前模型对训练样本进行预测，计算预测值与实际值之间的残差。

（3）训练新模型：使用残差作为目标变量，训练一个新的基础模型，以拟合残差。

（4）更新模型：将新模型与之前的模型进行加权融合，并更新整体模型。

（5）重复步骤 2-4：重复上述步骤，不断迭代，每次迭代都将新模型的预测结果与之前模型的预测结果进行累加，逐步细化预测结果。

（6）构建最终模型：终止迭代后，将所有模型的预测结果加权合并，得到最终的预测结果。

GBDT 模型建立过程如图 4 所示：

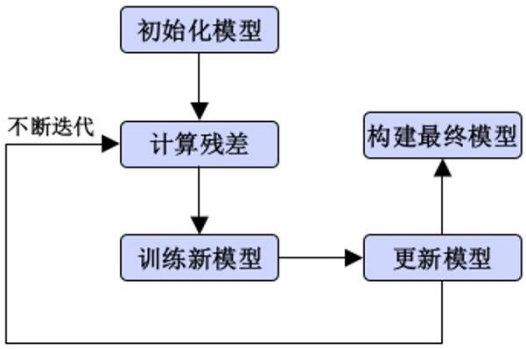

图 4 GBDT 模型建立过程

6.3 模型的求解

在进行 GBDT 模型求解前，可以先进行基本的决策树模型对预测模型进行可视化探索。

6.3.1 决策树构造

决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直接运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。

在数据清洗的基础上，我们采用随即洗牌的方法对数据进行随机排序，按照 8：2 的比例将数据分为训练集和验证集，训练得到所需决策树模型。

6.3.2 决策树运行结果

经训练得到的决策树的运行流程如图 5 所示，该决策树能够充分反映出以母亲的身体指标和心理指标为参考变量的基础上，婴儿行为特征的分类规律。

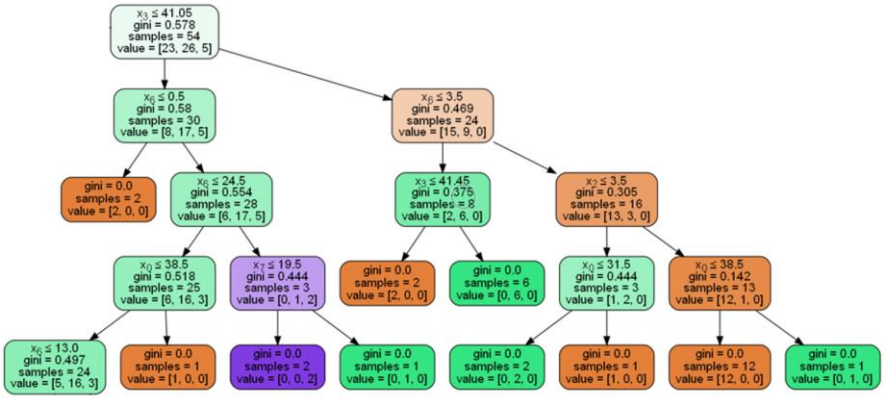

图 5 决策树结构图

各变量的特征重要性如图 6 所示：

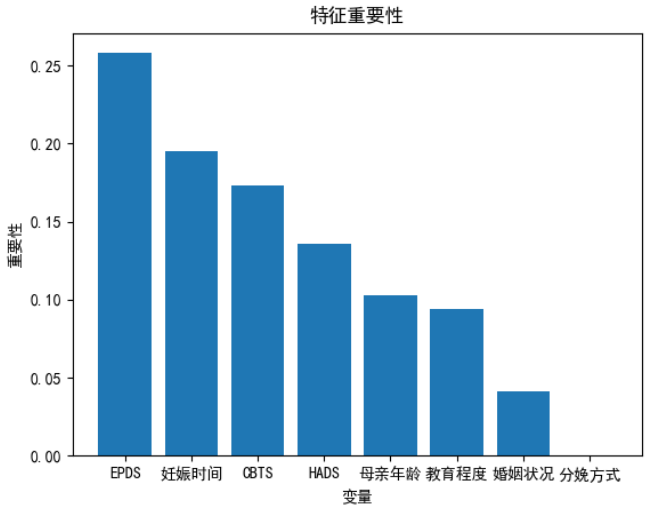

图 6 决策树模型特征重要性

对应的决策树的预测模型的准确性数据如表 5 所示：

表 5 决策树模型的准确率

	准确率	召回率	精确率	F1
训练集	0.857	0.857	0.867	0.847
测试集	0.774	0.774	0.767	0.768

6.3.3 采用 GBDT 模型进行求解

在数据清洗的基础上，我们采用随即洗牌的方法对数据进行随机排序，按照 8：2 的比例将数据分为训练集和验证集，训练得到所需 GBDT 模型的预测准确性结果如表 6 下：

表 6 GBDT 模型的预测准确率

	准确率	召回率	精确率	F1
训练集	0.932	0.932	0.94	0.931
测试集	0.841	0.824	0.832	0.839

由上表可知，利用 GBDT 模型对训练集进行训练后所得到的结果的准确率是比较高的，但是当其运用于测试集时，其准确率较低，为 0.841。为提高模型的预测精度，我们对模型采用能够在合理时间内找到接近最优解的解，且具有较好的收敛性和适应性的模拟退火算法对参数进行寻优处理。最终得到的模型的测试集预测准确率结果为：

表 7 进行调参后所得 GBDT 模型的预测准确率

	准确率	召回率	精确率	F1
训练集	0.932	0.932	0.94	0.931
测试集	0.863	0.841	0.812	0.810

6.3.4 GBDT 模型预测结果

运用所建立的 GBDT 预测模型，我们对数据表中最后 20 组（编号 391-410 号）婴儿的行为特征信息进行预测。最终得到的结果如表 8 下：

表 8 缺失的婴儿行为特征预测结果

婴儿编号	婴儿行为特征	婴儿编号	婴儿行为特征
391	中等型	392	安静型
393	矛盾型	394	中等型
395	中等型	396	中等型
397	中等型	398	中等型
399	中等型	400	中等型
401	中等型	402	中等型
403	中等型	404	安静型
405	中等型	406	安静型
407	中等型	408	安静型
409	安静型	410	中等型

七、问题三模型

问题三需要利用附件所提供的数据及题目中所提供的各心理指标治疗方案的价格建立相应的模型，并利用该模型求出能够使婴儿的行为特征从矛盾型转变为中等型所需花费的最少的费用。对此，我们可以先根据题目中的治疗方案的价格表，列出对应的微分方程。然后利用第二问中决策树所求得的母亲身心指标与心理指标对婴儿行为特征的影响特征重要性，由此构建关于治疗费用的线性规划模型，然后对模型进行求解，得到能够使婴儿的行为特征从矛盾型变为中等型的最少的治疗费用。

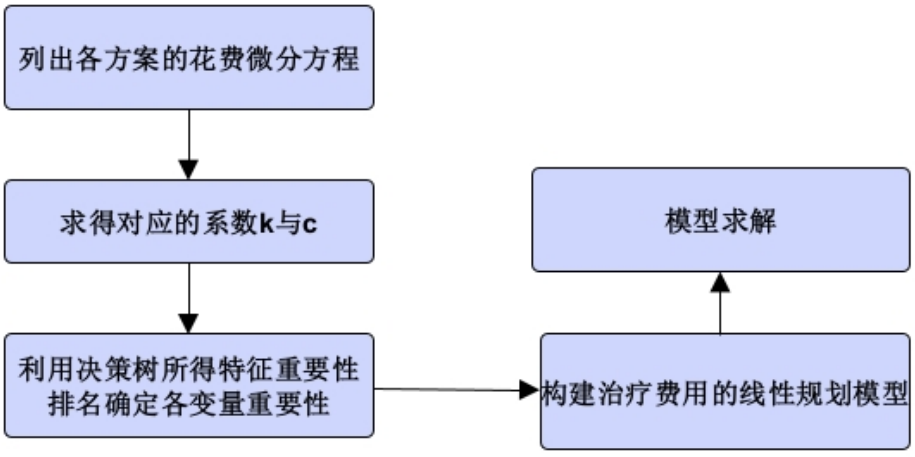

图 7 问题三流程图

7.1 治疗方案价格微分方程的建立

设单个治疗方案所需花费为y，相对应的得分为x，又因为 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。由此可得以下式子：

$$\frac{d_y}{d_x} = ky \tag{12}$$

求微分可得$\mathrm{y} = e^{kx+c}$,带入题目中所给的价格方案的数据，得到

$$y_1 = e^{0.525x+5.298}$$

$$y_2 = e^{0.654x+6.216},$$

$$y_3 = e^{0.654x+5.704},$$

由上式可以得到相应的治疗方案费用的式子为：

$$\begin{aligned}(min)\mathrm{Y} &= y_1 + y_2 + y_3 \\ &= e^{0.525x+5.298} + e^{0.654x+6.216} + e^{0.654x+5.704},\end{aligned} \tag{13}$$

7.2 分析婴儿特征类别影响因素

根据问题二，我们可以得到婴儿的行为特征所属类别与母亲的身体指标与心理指标都有一定的关系，并且通过问题二中所构建的决策树模型，我们可以得到母亲的身体指标和心理指标与婴儿的行为特征的特征重要性如图 8 所示：

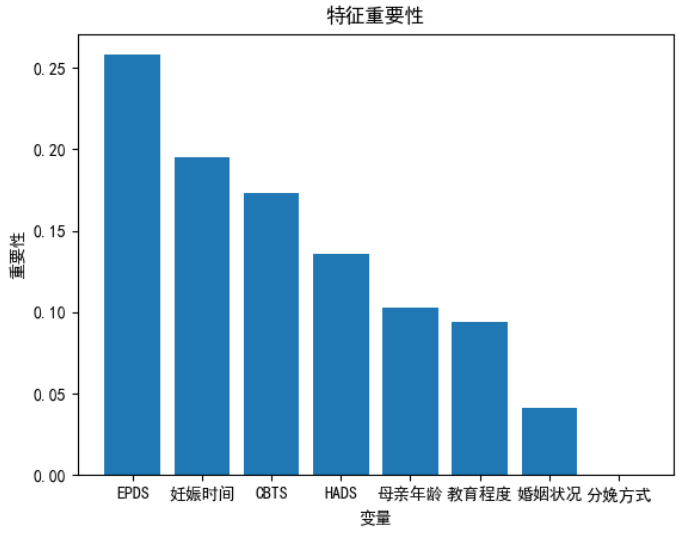

图 8 母亲身体指标和心理指标影响婴儿行为特征的重要性

综合上图中的数据，可以列出婴儿行为特征类别所属类型判别的变量影响式子为：

$$\mu = 0.258 * EPDS + 0.195 * \text{妊娠时间} + 0.173 * CBTS + 0.136 * HADS \tag{14}$$

$$+0.103 * \text{母亲年龄} + 0.094 * \text{教育程度} + 0.041 * \text{婚姻状况}$$

由上图可知，婴儿行为特征所属类别并不单纯只是受母亲的心理指标影响，而且还与母亲的身体指标密切相关。但是题目中只提供改变母亲的心理指标的方案，即说明题目需要通过单方面改变母亲的心理指标入手，但是在改变母亲的心理指标时，母亲的身体指标也是需要注意的，如母亲的妊娠时间对婴儿行为特征的特征重要性达 25.8%。

根据题目中所提供的价格表格，可以发现每一种治疗方案都存在一个基础费用（即无论是否有效）——CBTS 为 200 元，EPDS 为 500 元，HADS 为 300 元。而根据上述得到的微分方程，各方案中每降低 1 得分的花费比例为：**1.25 ： 1 ： 3.51。**

另外，由于题目中要求相应的治疗方案的治疗费用最小化，故我们可以采用贪心算法进行结果寻优。根据决策树所提供的决策信息结合各方案的价格信息，我们可以得知各个方案的性价比排序依次为：EPDS > CBTS > HADS。故所制定的最终方案中，EPDS 治疗方案是比较重要的。

通过筛选出中等型与安静型婴儿行为特征的母亲身体与心理指标，并和编号为 238 的行为特征为矛盾型婴儿的母亲身体与心理指标进行比较，我们可以发现存在母亲的心理指标数据中的 CBTS 与 HADS 指标比编号 238 婴儿母亲的指标高的情况，又因为根据上述分析，EPDS 数据是性价比最高的治疗方案。另外，对编号 238 婴儿母亲的数据进行分析可知，该母亲的妊娠时间为 26.5 周，明显低于中等型母亲的妊娠时间，并且与接近 CBTS 与 HADS 得分数据的中等型母亲之间差了接近 10 周的时间，与安静型母亲之间相差了近 14 周的时间，属于早产儿。因此，要想该婴儿的行为特征由矛盾型转变为中等型则需要利用 EPDS 对行为特征类别的重要性来平衡该婴儿母亲妊娠时间所带来的影响。

综合上述分析，可以得到以下线性规划模型：

$$\text{s.t}\begin{cases} (min)\text{Y} = y_1 + y_2 + y_3, \\ \mu = \sum \alpha_i, \\ Y \geq 0, \\ EPDS \leq 30, CBTS \leq 30, HADS \leq 30 \end{cases} \tag{15}$$

根据上述所建立的线性规划模型及代入编号为 238 婴儿母亲的数据，可以得到让该婴儿的行为特征从矛盾型变为中等型最少需要花费 5252 元用于 EPDS，具体方案是通过 EPDS 治疗 10 分，从矛盾型转变为安静型最少需要花费 7354 元的 EPDS 治疗 14 分。

八、问题四模型

问题四需要根据婴儿的睡眠质量指标数据进行综合评判，并将其分成优、良、中、差四个类别，具体的分类标准可以参考美国睡眠医学会（AASM）的分类标准。在此基础上，可以通过应用可处理不相关特征、简单易于理解的决策树模型来建立婴儿综合睡眠质量与母亲的身体指标、心理指标的模型，并用于预测编号为 391-401 号婴儿的综合睡眠质量。

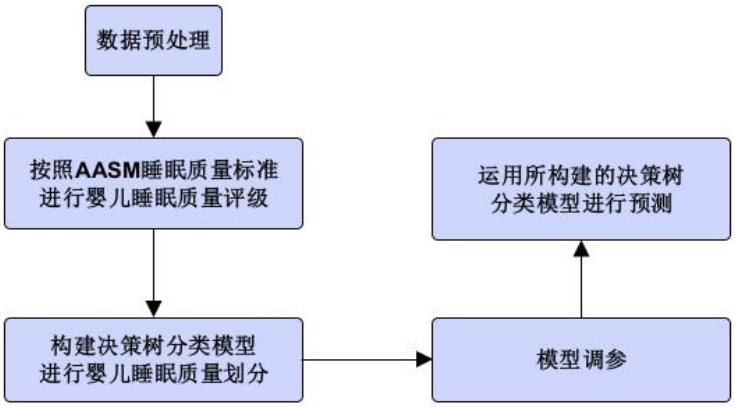

图 9 问题三流程图

8.1 数据预处理

1、数据编码

将用于评价婴儿睡眠质量的优、良、中、差的综合评判结果进行数据编码，具体的编码标准如下所示：

表 9 睡眠综合评判数据编码标准

睡眠综合评判	编码
优	1
良	2
中	3
差	4

2、综合评判

参考美国睡眠医学会（AASM）的分类标准，我们可以对婴儿的睡眠质量进行综合评判，具体的评判标准如表 10 所示：

表 10 婴儿睡眠质量评判标准

综合评判等级	评判标准
优	婴儿睡眠时长达到 11-14 小时， 或每天夜间醒来次数不超过 1 次， 或以环境营造法或定时法的方式入睡
良	婴儿睡眠时长达到 10-11 小时或 14-15 小时， 或每天夜间醒来次数在 2-3 次， 或以安抚奶嘴法入睡
中	婴儿睡眠时长达到 9-10 小时或 15-16 小时， 或每天夜间醒来次数在 4-5 次， 或以抚触方式入睡
差	婴儿睡眠时长低于 9 小时或超过 16 小时， 或每天夜间醒来次数超过 5 次， 或以哄睡法的方式入睡

8.2 基于决策树的婴儿睡眠综合质量划分模型

根据题目要求，我们需要以母亲的身体指标和心理指标作为婴儿睡眠质量综合评判划分的重要因素来构造预测模型。因此，我们选择了可处理不相关特征、简单易于理解的、准确率较高的决策树的婴儿睡眠综合质量划分模型。

8.2.1 决策树构造

结合题目的需求，我们选取母亲的身体指标和心理指标作为决策变量构造决策树。并且采用随机洗牌的方法对数据进行随机排序，按照 8：2 的比例将数据分为训练集和验证集，训练得到所需决策树模型。

8.2.2 决策树运行结果

经过训练得到的决策树运行流程图如图 10 所示，该决策树能够较为充分地反映出婴儿睡眠质量综合评判的分类规律。

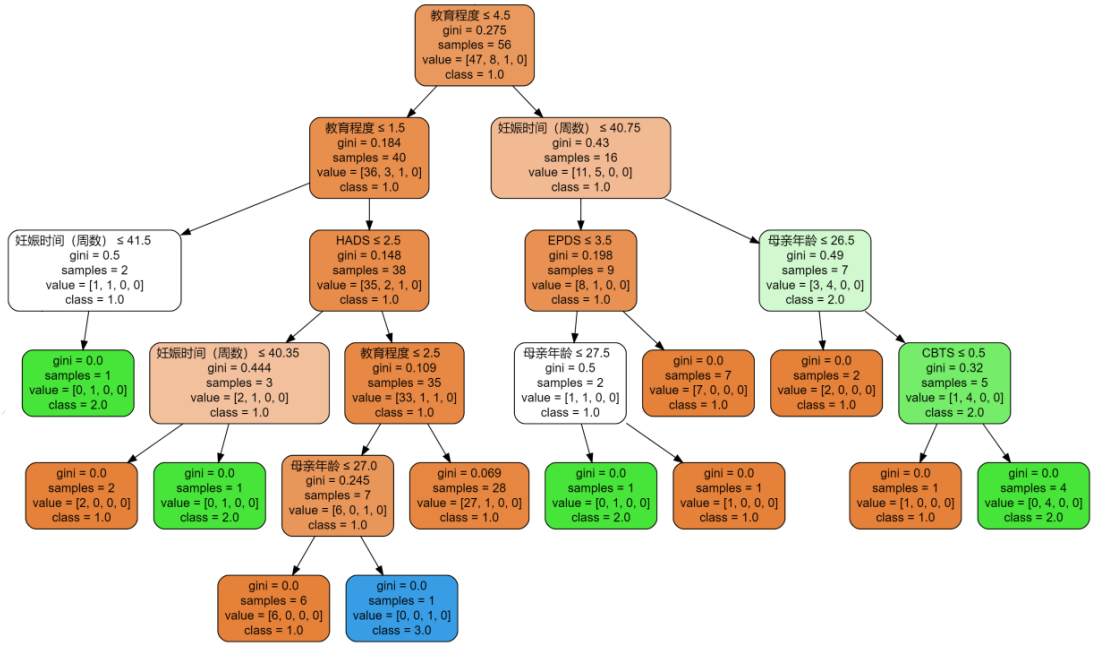

图 10 决策树分析过程

决策树的模型评价结果如表 11 所示：

表 11 决策树分析结果的准确率

	准确率	召回率	精确率	F1
训练集	0.974	0.974	0.974	0.973
测试集	0.821	0.810	0.836	0.855

8.2.3 模型求解

利用上述所建立的决策树模型对最后 20 组（编号 391-401 号）婴儿的综合睡眠评判结果进行预测，得到如表 12 的预测结果：

表 12 预测结果表

婴儿编号	综合睡眠质量	婴儿编号	综合睡眠质量
391	优	392	优
393	优	394	优
395	优	396	优
397	中	398	优
399	优	400	优
401	优	402	优
403	优	404	优
405	中	406	优
407	优	408	优
409	优	410	优

九、问题五模型

问题五需要求在问题三的基础上，编号为 238 号婴儿的睡眠质量评级为优所需花费费用最少的治疗方案。而这一求最小化结果属于线性规划问题，可以采用问题三的解决思路，利用问题四所构建的决策树模型可以得知母亲的身体指标与心理指标与婴儿的睡眠质量综合评判的特征重要性，并以此为基础修改原有的线性规划模型从而得到相应的调整后的方案。

9.1 分析婴儿综合睡眠质量影响因素

根据第四问的分析，我们已知婴儿的睡眠质量评级所属级别与母亲的身体指标与心理指标都有一定的关系，并且通过问题二中所构建的决策树模型，我们可以得到母亲的身体指标和心理指标与婴儿综合睡眠质量评级的特征重要性如下：

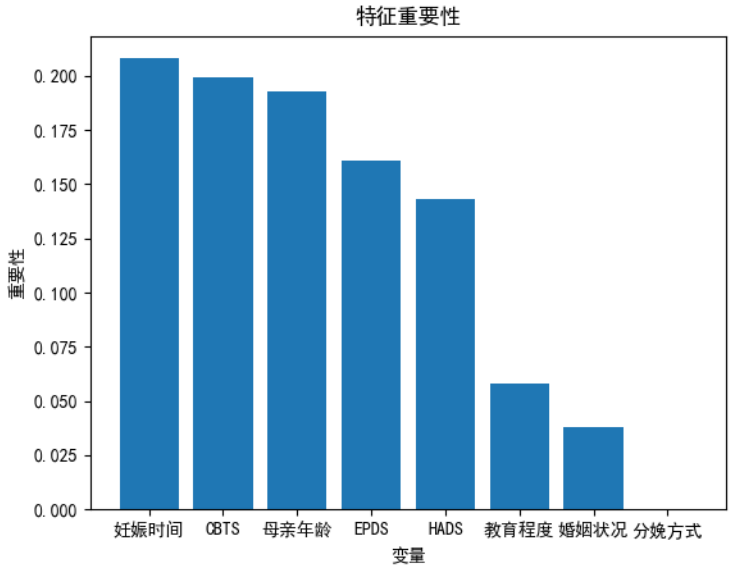

图 11 母亲身体指标和心理指标
对婴儿综合睡眠质量影响的特征重要性

综合上图中的数据，可以列出婴儿行为特征类别所属类型判别的变量影响式子调整为：

$$\mu = 0.208 * \text{妊娠时间} + 0.199 * \text{CBTS} + 0.193 * \text{母亲年龄} \tag{16}$$

$$+0.161 * \text{EPDS} + 0.143 * \text{HADS} + 0.058 * \text{教育程度} + 0.038 * \text{婚姻状况}$$

由上图可知，婴儿行为特征所属类别并不单纯只是受母亲的心理指标影响，而且还与母亲的身体指标密切相关。但是题目中只提供改变母亲的心理指标的方案，即说明题目需要通过单方面改变母亲的心理指标入手，但是在改变母亲的心理指标时，母亲的身体指标也是需要注意的，如母亲的妊娠时间对婴儿行为特征的特征重要性达 20.8%。

同样的，由于题目限定了从治疗心理指标入手，经过问题三的分析，我们已经知道选择 EPDS 治疗方案是比较明智的选择。

通过筛选出中等型与安静型婴儿行为特征并且睡眠质量为优的母亲身体与心理指标，并和编号为 238 的行为特征为矛盾型婴儿的母亲身体与心理指标进行比较，我们可以发现存在母亲的心理指标数据中的 CBTS 与 HADS 指标比编号 238 婴儿母亲的指标高的情况，又因为根据上述分析，EPDS 数据是性价比最高的治疗方案。另外，对编号 238 婴儿母亲的数据进行分析可知，该母亲的妊娠时间为 26.5 周，明显低于中等型母亲的妊娠时间，并且与接近 CBTS 与 HADS 得分数据的中等型母亲之间相差了接近 10 周的时间，与安静型母亲之间相差了近 14 周的时间，属于早产儿。另外，编号为 238 的婴儿的原始睡眠质量评级为良，与评级为优相差一个等级。因此，要想该婴儿的行为特征由矛盾型转变为中等型的基础上让其睡眠质量评级为优，则需要利用 EPDS 对行为特征类别与婴儿睡眠质量评级的重要性来平衡该婴儿母亲妊娠时间所带来的影响。

综合上述分析，可以得到以下线性规划模型：

$$\text{s.t}\begin{cases} (min)\text{Y} = y_1 + y_2 + y_3 \\ \mu = \sum \alpha_i \\ Y \geq 0 \end{cases} \tag{17}$$

根据上述所建立的线性规划模型及代入编号为 238 婴儿母亲的数据，可以得到让该婴儿的行为特征从矛盾型变为中等型最少需要花费 9517 元用于 EPDS 治疗，从矛盾型转变为安静型最少需要花费 12570 元的 EPDS 治疗。

十、模型评价

10.1 模型的优点

1、问题一中，利用斯皮尔曼相关性分析及灰色关联的方法，从母亲的身体指标与心理指标中筛选出具有重要特征信息的数据，便于下文构建相应的模型。

2、模型中所得到数据的结果经过多种方法的验证，如卡方检验等，确保数据的准确可信。

3、本模型考虑较为全面，模型简洁高效，便于理解和应用，得到的求解结果与实际情况接近，结果较为合理。

4、模型中的解答环环相扣、层次递进，没有顾此失彼，造成问题解答的割裂。

10.2 模型的缺点

本文的模型中，关于治疗方案的选择方面的解答还是不够全面，现实中的单位价格的医疗效果并不是一成不变的，还需考虑多种医疗方案有机结合下的医疗效果。

参考文献

[1]姜子艳. 基于计算机视觉的婴儿睡眠质量评估算法研究[D].电子科技大学,2023.DOI:10.27005/d.cnki.gdzku.2022.003669.

[1]孙佳慧,赵孟娇,梁熙. 母亲抑郁与家庭出生顺序对婴儿期睡眠的影响——母亲养育方式的多重中介作用[C]//中国心理学会.第二十三届全国心理学学术会议摘要集（下）.第二十三届全国心理学学术会议摘要集（下）,2021:62-63.DOI:10.26914/c.cnkihy.2021.039748.

[2]李天. 1-6 月龄婴儿睡眠现状及影响因素分析[D].兰州大学,2021.DOI:10.27204/d.cnki.glzhu.2021.002827.

[3]朱丽霞. 孕晚期母亲睡眠对子代生长及代谢影响的队列研究[D].上海交通大学,2018.DOI:10.27307/d.cnki.gsjtu.2018.002451.

[4]宋沅瑾. 睡眠对儿童饮食情况及身体活动水平影响的系列研究[D].上海交通大学,2015.

[5]孙箐爽. 父母情绪、睡眠与婴儿早期睡眠的相关性研究[D].兰州大学,2015.

[6]黄小娜. 基于生长发育轨迹婴儿期睡眠/觉醒模式及行为影响因素研究[D].中国疾病预防控制中心,2013.

附录

由于本文模型用到的全部代码过多，而且部分代码存在复用性，所以附录只包括主要的源代码，其余源码可见于支撑材料。

支撑材料清单；

Python 程序文件	c1 问题一

1. 问题一代码

本部分代码是在 Pycharm 编写，在 Windows10 环境下编程，主要代码如下：

```
import matplotlib.pyplot as plt
import matplotlib

matplotlib.rcParams['font.sans-serif'] = ['SimHei']

# 创建一个包含 3 行 3 列的图形，并指定图形大小
fig, axs = plt.subplots(3, 3, figsize=(6, 6))

# 在每个子图中绘制柱形图
axs[0, 0].bar(['中等型', '安静型', '矛盾型'], [30.0, 30.9,
29.4],color=['orange', 'blue', 'gray']) # 第一行第一列的子图
axs[0, 0].set_xlabel('婴儿行为特征')  # 设置第一行第一列子图的横坐标标签
axs[0, 0].set_ylabel('母亲年龄平均值')  # 设置第一行第一列子图的纵坐标标签

axs[0, 1].bar(['中等型', '安静型', '矛盾型'], [1.95, 1.96,
1.97],color=['orange', 'blue', 'gray']) # 第一行第二列的子图
axs[0, 1].set_xlabel('婴儿行为特征')  # 设置第一行第二列子图的横坐标标签
axs[0, 1].set_ylabel('婚姻状况')  # 设置第一行第二列子图的纵坐标标签

axs[0, 2].bar(['中等型', '安静型', '矛盾型'], [4.1, 4.0, 4.1],color=['orange',
'blue', 'gray']) # 第一行第三列的子图
axs[0, 2].set_xlabel('婴儿行为特征')  # 设置第一行第三列子图的横坐标标签
axs[0, 2].set_ylabel('教育程度')  # 设置第一行第三列子图的纵坐标标签

# 在剩余的子图中绘制柱形图
axs[1, 0].bar(['中等型', '安静型', '矛盾型'], [39.1, 39.2,
39.1],color=['orange', 'blue', 'gray']) # 第二行第一列的子图
axs[1, 0].set_xlabel('婴儿行为特征')  # 设置第二行第一列子图的横坐标标签
axs[1, 0].set_ylabel('妊娠时间')  # 设置第二行第一列子图的纵坐标标签

axs[1, 1].bar(['中等型', '安静型', '矛盾型'], [1.01, 1.01,
0.99],color=['orange', 'blue', 'gray']) # 第二行第二列的子图
axs[1, 1].set_xlabel('婴儿行为特征')  # 设置第二行第二列子图的横坐标标签
axs[1, 1].set_ylabel('分娩方式')  # 设置第二行第二列子图的纵坐标标签
```

```
axs[1, 2].bar(['中等型', '安静型', '矛盾型'], [6.21, 5.35,
6.59],color=['orange', 'blue', 'gray']) # 第二行第三列的子图
axs[1, 2].set_xlabel('婴儿行为特征')  # 设置第二行第三列子图的横坐标标签
axs[1, 2].set_ylabel('CBTS')  # 设置第二行第三列子图的纵坐标标签

axs[2, 0].bar(['中等型', '安静型', '矛盾型'], [9.27, 8.19,
11.02],color=['orange', 'blue', 'gray']) # 第三行第一列的子图
axs[2, 0].set_xlabel('婴儿行为特征')  # 设置第三行第一列子图的横坐标标签
axs[2, 0].set_ylabel('EPDS')  # 设置第三行第一列子图的纵坐标标签

axs[2, 1].bar(['中等型', '安静型', '矛盾型'], [8.05, 7.26,
8.75],color=['orange', 'blue', 'gray']) # 第三行第二列的子图
axs[2, 1].set_xlabel('婴儿行为特征')  # 设置第三行第二列子图的横坐标标签
axs[2, 1].set_ylabel('HADS')  # 设置第三行第二列子图的纵坐标标签

# 隐藏第三行第三列的子图
axs[2, 2].remove()

# 调整子图之间的间距
plt.tight_layout()

# 显示图形
plt.show()


# 以下是随机森林模型进行预测
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib
from sklearn.metrics import accuracy_score
matplotlib.rcParams['font.sans-serif'] = ['SimHei']

data = pd.read_excel('华数杯.xlsx',usecols=range(15))
print(data)

le = LabelEncoder()
data['婴儿行为特征'] = le.fit_transform(data['婴儿行为特征'].astype(str))

train_data = data[data['编号']<390]
test_data = data[data['编号']>390]
```

```
train_features = train_data.drop(['编号', '婴儿行为特征', '整晚睡眠时间（时：分：秒）', '睡醒次数', '入睡方式'], axis=1)
train_target = train_data['婴儿行为特征']

test_features = test_data.drop(['编号', '婴儿行为特征', '整晚睡眠时间（时：分：秒）', '睡醒次数', '入睡方式'], axis=1)

rf = RandomForestClassifier(n_estimators=100, max_depth=3, class_weight='balanced', random_state=42)

rf.fit(train_features, train_target)

test_predictions = rf.predict(test_features)

test_predictions_labels = le.inverse_transform(test_predictions)

print("预测结果：")
print(test_predictions_labels)

accuracy = accuracy_score(test_data['婴儿行为特征'],test_predictions)
print("预测准确率：", accuracy)

cm = confusion_matrix(train_target, rf.predict(train_features))
sns.heatmap(cm,annot=True, fmt='d',
            xticklabels=le.classes_,
            yticklabels=le.classes_)



plt.ylabel('真实值')
plt.xlabel('预测值')
plt.savefig('D:/桌面/华数杯/第二问.png')
plt.show()
```

2. 问题二代码

本部分代码是在 Pycharm 编写，在 Windows11 环境下编程，主要代码如下：

```
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
# from sklearn.metrics import accuracy_score
import matplotlib
matplotlib.rcParams['font.sans-serif'] = ['SimHei']
# 定义数据
```

```
x = ['EPDS', '妊娠时间', 'CBTS', 'HADS', '母亲年龄', '教育程度', '婚姻状况', '分娩方式']
y = [0.258, 0.195, 0.173, 0.136, 0.103, 0.094, 0.041, 0]

# 绘制条形图
plt.bar(x, y)

# 添加标题和轴标签
plt.title('特征重要性')
plt.xlabel('变量')
plt.ylabel('重要性')



# 显示图形
plt.show()
```

3. 问题四代码

本部分代码是在 Pycharm 编写，在 Windows11 环境下编程，主要代码如下：

```
import pandas as pd
data = pd.read_excel('华数杯 1.xlsx',usecols=range(17))
result = []
for index, row in data.iterrows():
    if (row['睡眠时间'] >= 11 and row['睡眠时间']<=14) or row['睡醒次数']<=1 or row['入睡方式']>=4:
        result.append('1')
    elif (row['睡眠时间'] <11 and row['睡眠时间']>=10) or (row['睡眠时间']<=15 and row['睡眠时间']>14) or (row['睡醒次数']>1 and row['睡醒次数']<=3)or(row['入睡方式']==3):
        result.append('2')
    elif (row['睡眠时间'] <10 and row['睡眠时间']>=9) or (row['睡眠时间']<=16 and row['睡眠时间']>15) or (row['睡醒次数']>3 and row['睡醒次数']<=5)or(row['入睡方式']==2):
        result.append('3')
    else:
        result.append('4')

data.insert(loc=16, column='婴儿睡眠质量', value=result)


data.to_excel('华数杯 3.xlsx', index=False)
```

3. 问题五代码

本部分代码是在 Pycharm 编写，在 Windows11 环境下编程，主要代码如下：

```
import matplotlib.pyplot as plt
import matplotlib.ticker as mtick
# from sklearn.metrics import accuracy_score
import matplotlib
matplotlib.rcParams['font.sans-serif'] = ['SimHei']
# 定义数据
x = ['妊娠时间', 'CBTS', '母亲年龄', 'EPDS','HADS', '教育程度', '婚姻状况', '分娩方式']
y = [0.208, 0.199, 0.193, 0.161, 0.143, 0.058, 0.038, 0]

# 绘制条形图
plt.bar(x, y)

# 添加标题和轴标签
plt.title('特征重要性')
plt.xlabel('变量')
plt.ylabel('重要性')



# 显示图形
plt.show()
```