所属类别	2023 年“华数杯”全国大学生数学建模竞赛	参赛编号
研究生组		CM2302809

母亲状态对婴儿行为与睡眠的影响

摘要

本文研究了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，并构建了相应的规划模型，以指导母亲心理问题的治疗策略。

对于问题一：要求探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，本文首先对母亲和婴儿数据进行数据预处理和异常值处理，后采用**卡方检验**和**斯皮尔曼相关系数法**，发现母亲的心理指标与婴儿的行为特征和睡眠质量存在显著关联，且**不同月龄**的婴儿对母亲的心理状况变化敏感性不同。这些研究结果为制定母亲心理问题治疗策略提供了重要依据。

对于问题二：需要根据母亲的心理和身体原因来预测婴儿的行为特征，本文先根据问题一中的相关性，来确认对于不同年龄的婴儿，母亲的哪些因素对其行为特征有着显著影响，以避免由于模型输入过多导致的欠拟合或过拟合问题。基于此，本文采用 **XGboost** 与**随机森林**两种模型以实现上述预测，对比发现随机森林模型更优，对于三种年龄的婴儿能达到 **64.29%**，**69.23%**,**72.00%**的分类准确率，并按此模型在后续问题预测被删除的婴儿行为特征。

对于问题三：要求能通过治疗母亲的三个不同的焦虑水平来使编号为 238 的婴儿行为特征由焦虑型转变为中等或安静型。基于上述要求，本文结合第二问中的模型与 **Minimize** 来寻找能达到要求的最低费用，使其经过治疗后模型能将其预测为中等型与安静型，所花费用分别为 **136263.82** 和 **339827.27**。

对于问题四：需要根据婴儿睡眠质量指标对婴儿睡眠质量进行评估，并建立母亲身体心理指标与婴儿入睡方式的关联模型。本文首先对婴儿睡眠质量指标进行了一致化与归一化处理，并采用熵值法与 Topsis 法相结合对睡眠质量进行综合评估，后将母亲特征作为输入，利用**梯度提升树**模型构建分类模型对睡眠质量进行预测。分类准确率 **41.03%**，后续可以考虑填充数据进行训练以优化模型分类效果。

对于问题五：需要根据在第三问的基础上继续完成对婴儿的治疗，新增约束条件不仅改善婴儿行为特征，还要改善婴儿睡眠特征。该问本质上属于**多约束单目标规划问题**，由于并不涉及多目标，本题模型仅需在问题三所建立的规划模型中加入**新的约束条件**即可。对于问题五的子问题 1，所花费用为 **451188.57**。而对于子问题 2，算法收敛失败，本文认为其原因在于原始数据集质量不高，导致误差积累。

关键词：睡眠质量；产后抑郁；梯度提升树；随机森林；卡方检验

一、 问题重述

母亲是婴儿生命中最重要的人之一，她不仅为婴儿提供营养物质和身体保护，还为婴儿提供情感支持和安全感。母亲心理健康状态的不良状况，如抑郁、焦虑、压力等，可能会对婴儿的认知、情感、社会行为等方面产生负面影响。压力过大的母亲可能会对婴儿的生理和心理发展产生负面影响，例如影响其睡眠等方面。

本文旨在对所给数据进行深度数据挖掘，探索母亲心理健康状态与婴儿认知、情感、社会行为，以及睡眠质量之间的关联。通过分析数据，我们将寻找可能存在的规律和模式，进一步了解母亲心理健康对婴儿发展的潜在影响。这样的研究有助于为提高婴儿健康发展的干预措施提供科学依据，从而促进婴儿全面成长，同时为母亲产后抑郁治疗方案提供参考。

二、 问题分析

2.1 问题一的分析

根据题意，需要对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进行探究。在开始实施研究之前，我们首先对题目给出的数据进行了初步可视化，并进行了正态分布检验，以确定后续使用何种相关性检验方法。在相关性分析过程中，我们发现婴儿特征与婴儿年龄之间具有较高的关联性。因此，接下来的分析中，我们选择将婴儿月龄进行分类讨论，并根据数据类型（离散或连续）选用适当的检验方法，例如卡方检验或斯皮尔曼相关系数检验。

2.2 问题二的分析

问题二需要建立起母亲身心状态到婴儿行为特征的关系模型。根据问题一中分析所得婴儿特征与婴儿年龄之间具有较高的关联性，将婴儿年龄分为 3 类，完成分类后，再将能显著影响不同年龄段婴儿行为特征的母亲身心数据作为分类的主要影响因素，最后基于机器学习方法如 XGboost、随机森林等模型实现对不同年龄段婴儿行为特征的预测分类。

2.3 问题三的分析

问题三的目标是找到使婴儿行为特征达到指定值时，需要对母亲实施治疗的最低费用。我们可以使用优化方法来实现这个目标。

首先，需要根据问题二中建立的关系模型，将婴儿行为特征与母亲的身体指标和心理指标联系起来。然后，本文引入一个变量，表示治疗费用，用来优化母亲焦虑的干预。接下来，我们可以建立一个优化模型，使用 Minimize 函数来最小化治疗费用，同时保证婴儿行为特征能在母亲经过治疗后达到指定值。

2.4 问题四的分析

问题四需要建立一个睡眠质量预测模型，根据婴儿的睡眠质量指标：整晚睡眠时间，睡醒次数与入睡方式，实现对睡眠质量的准确预测。本文拟采取以下两个个步骤解决问题：

1 综合评价模型的建立：通过数据转化将数据皆转变为极大型指标，采用熵增法确定各指标的权重，与 Topsis 方法结合对婴儿睡眠质量进行综合评价。

2 建立分类模型：基于选定的睡眠质量指标，建立睡眠质量预测模型。模型的建立需要考虑婴儿睡眠数据与睡眠质量之间的内在关系，可以采用机器学习算法，如回归模型或分类模型，来实现对睡眠质量的准确预测和判别。

2.5 问题五的分析

问题五的需求是在问题三的基础上，不仅需要考虑婴儿行为特征的治疗，还要确保婴儿的睡眠质量达到优。为了实现这个目标，本文引入新的分类模型约束，以确保治疗方案既能改善婴儿的行为特征，又能提高睡眠质量。

在问题三中，我们已经建立了母亲焦虑的治疗费用与得分之间的正比关系模型，这为我们提供了评估治疗费用的依据。需要根据问题五的要求，引入新的分类模型，该模型将婴儿的综合睡眠质量与母亲的身体指标和心理指标联系起来，求解得到可行的治疗方案。

三、 模型假设

（1） 假设参与问卷调查的母亲在填写问卷时提供了真实和准确的信息。

（2） 假设问卷设计和调查过程符合科学研究的原则，确保了问卷内容的合理性和有效性。

（3） 假设参与调查的母亲了解问题的含义，并能够正确理解和回答问卷中的问题。

（4） 假设在给定的数据集中，母亲的身体指标和心理指标与婴儿的行为特征和睡眠质量之间存在相关性。我们猜测母亲的身体健康和心理健康状况可能会影响婴儿的认知、情感、社会行为以及睡眠质量。

四、 符号说明

符号	符号含义
Var_p	变量 X 和 Y 之间的皮尔逊相关系数
P	置信度
X^2	卡方统计量
SSR	回归平方和
SST	总平方和

Accuracy	正确率
Pr*ecision*	精确率
Re*call*	召回率
F_1	F1 得分
u	样本均值
σ	样本方差

五、 问题一模型建立和求解

5.1 数据预处理

异常值和缺失值处理是数据预处理的重要步骤。异常值可能导致偏误和错误结论，而缺失值则减少样本量和影响结果准确性。处理这些问题有助于提高数据质量和分析结果的可信度。通过删除、修正或填充，我们可以使数据更完整、更可靠，为后续分析和建模奠定基础，确保准确有效地从数据中提取有价值的信息。

在对附件数据进行清理筛查后，发现数据不存在缺失值。接着对数据进行可视化观察，检查每一列数据是否符合该指标的取值要求。在发现异常值时，特别是母亲身体指标“婚姻状况”栏出现 3 和 6 这样不符合实际含义的数值，我们进行了修正，用 2（已婚）进行数据填补。另外，在婴儿睡眠质量“整晚睡眠时间”栏出现 99.99 小时的数据，显然有误，我们选择删除整行进行处理。异常数据的可视化结果如图 5-1 中“o”所示。

通过数据清理和异常值处理，我们提高了数据的准确性和可信度，确保了后续分析的准确有效性。

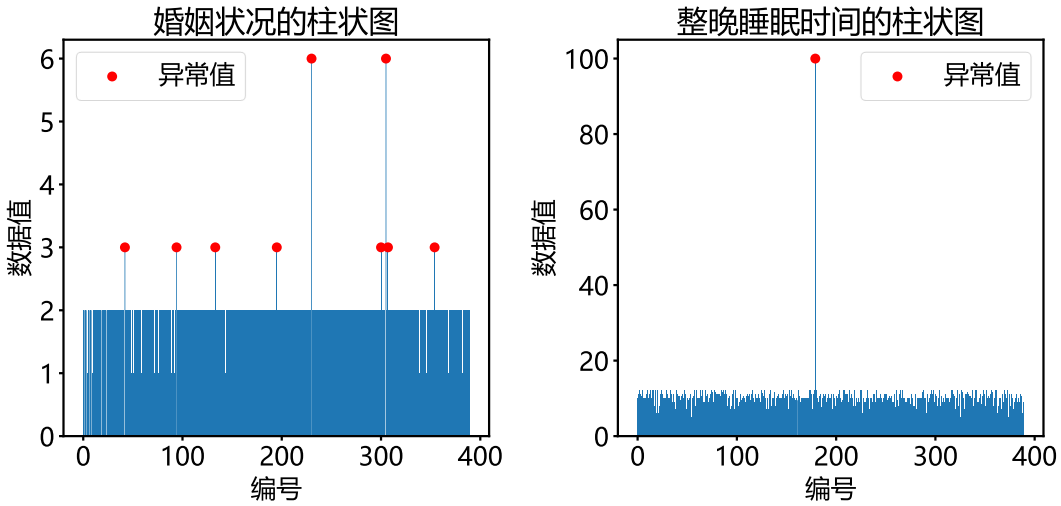


图 5-1 异常值处理图

5.2 数据检验

经典统计中的几乎所有模型的建立都依赖于原始数据或残差符合正态分布的假定，而服从正态分布的数据在回归任务中往往表现得比非正态的数据更加稳定。而且，符合正态分布的数据可以使用经典统计方法，而非正态的数据就需要数据变换或使用非参数统计的方法进行统计计算。因此，在建模之前，首先要对数据的正态性进行检验。本文采用 D'Agostino 检验，原假设是数据符合给定的分布。本文对附件中连续数据检验结果如下表所示：

表 5-1 连续数据正态分布检验结果

特征	检验统计量	P 值
母亲年龄	7.7457	<0.0201
妊娠时间	173.3579	<0.0001
CBTS	42.9314	<0.0001
EPDS	29.4166	<0.0001
HADS	18.9448	<0.0001
整晚睡眠时间	56.7777	<0.0001

采用 Q-Q 图验证 D'Agostino 检验结果，如图 5-2 所示，数据点偏离直线，附件中连续数据不符合正态分布，在后续数据处理中需要考虑数据变换或使用非参数统计的方法进行统计计算。

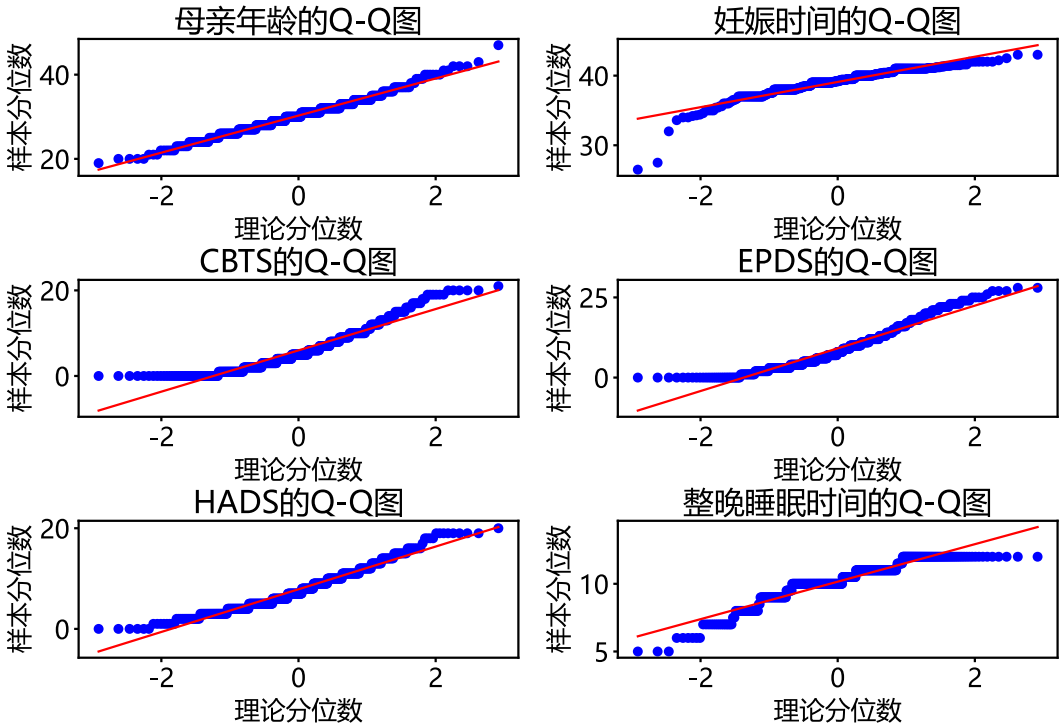


图 5-2 连续数据 Q-Q 图

问题一要求求出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，考虑到数据分布特点，本文首先采用斯皮尔曼相关系数法对各因素间相关性进行分析，并绘制热力图如图 5-3 所示。

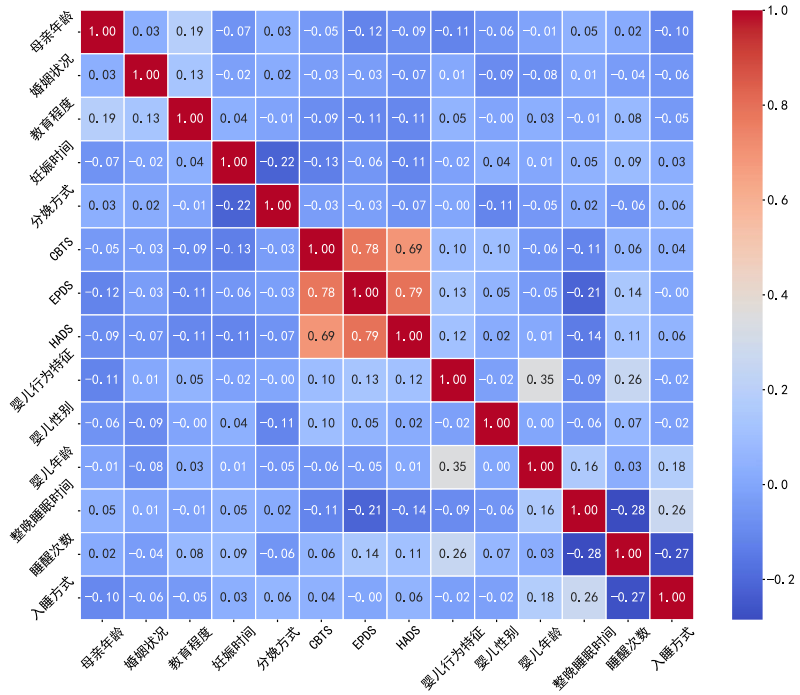


图 5-3 相关性热力图

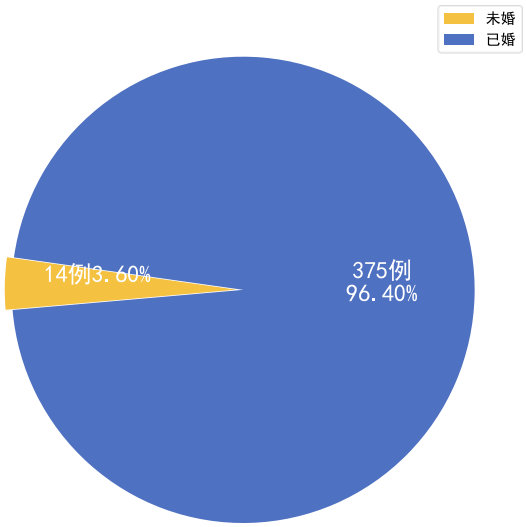
如图 5-3 所示，除了母亲心理指标 CBTS、EPDS 和 HADS 之间存在强相关性外，其他因素间的相关性较弱。这可能是因为婴儿行为特征和睡眠质量受多种因素影响，母亲的行为及状态可能只起到一部分影响作用。值得注意的是，婴儿年龄和婴儿行为特征之间存在 0.35 的相关性。因此，在后续相关性分析中，我们将根据婴儿年龄进行分组，以便更深入挖掘组内各因素之间的关系。

本文采用了多种统计方法和相关性分析来对数据进行全面的分析。对于离散数据，本文采用了例数描述，即对各个类别出现的频次进行统计，以直观地展示数据的分布情况。而在比较组间差异方面选择了卡方检验，这是一种适用于离散数据的非参数检验方法，用于判断不同组别之间是否存在显著差异。对于连续数据，本文采用了均值±标准差进行描述，这有助于理解数据的集中趋势和波动范围，并使用了斯皮尔曼相关性分析来研究数据间的关联性，这对于发现非线性相关或者排除异常值的影响非常有效。

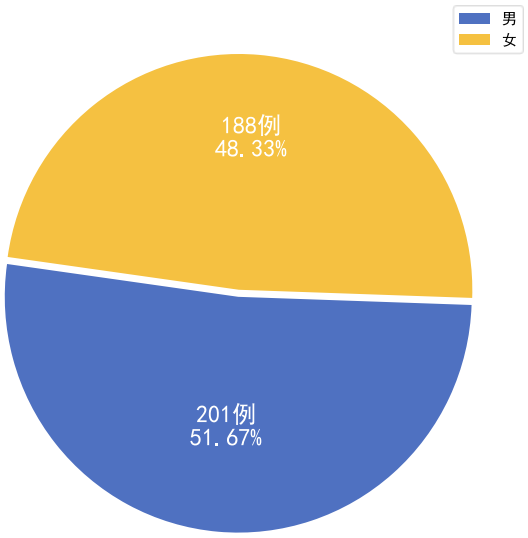
5.3 离散数据关联性分析

5.3.1 数据可视化

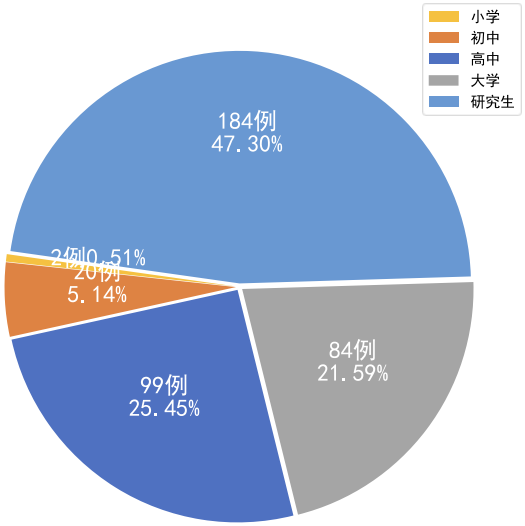
附件中离散数据包括：母亲婚姻状况、母亲教育程度、母亲分娩方式、婴儿性别、婴儿年龄、婴儿行为特征、婴儿睡醒次数以及婴儿入睡方式，具体分布情况如图 5-4 所示。



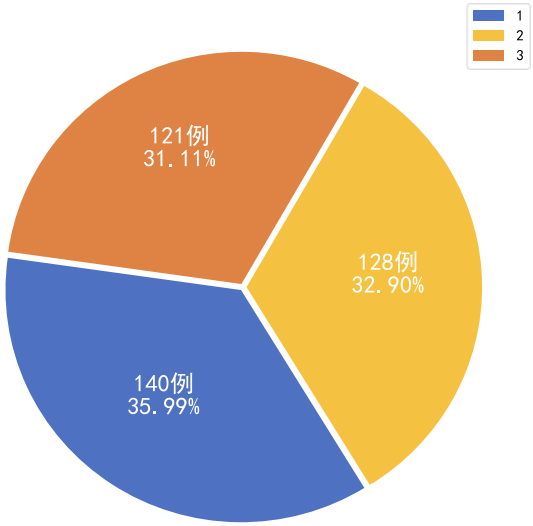
（a）母亲婚姻状况



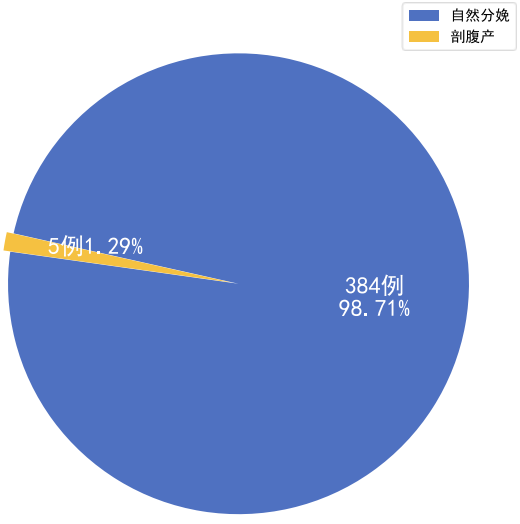
（d）婴儿性别



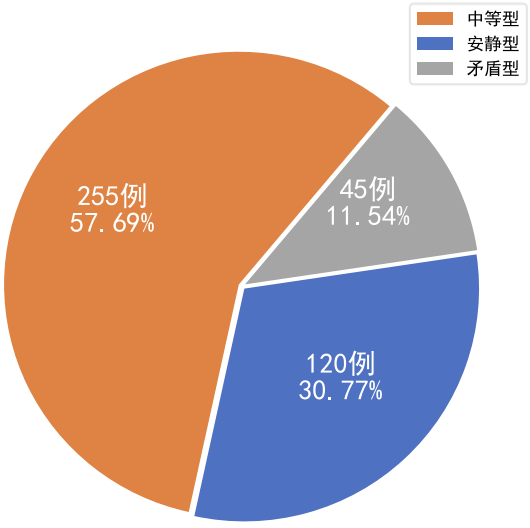
（b）母亲教育程度



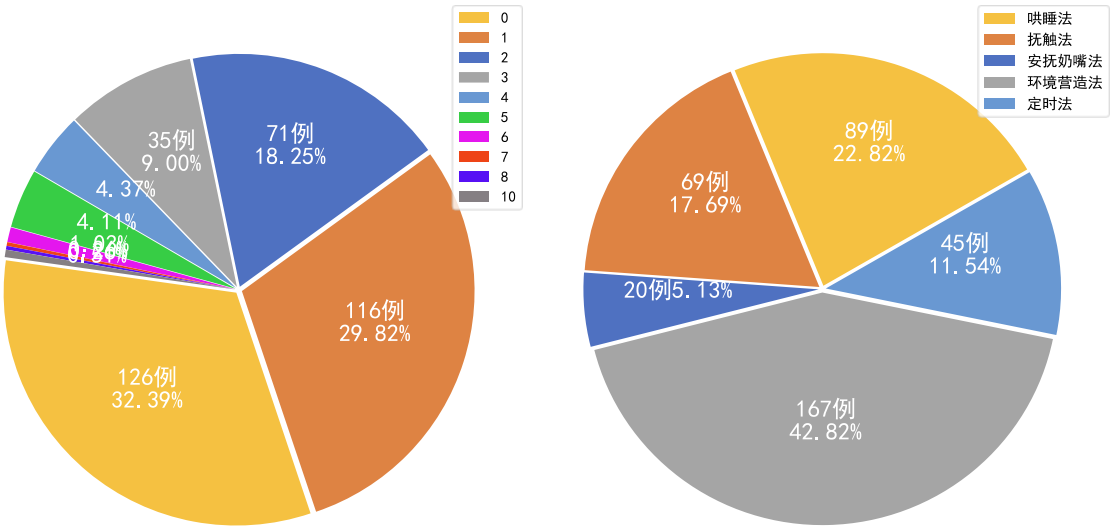
（e）婴儿年龄



（c）分娩方式



（f）婴儿行为特征



（g）婴儿睡醒次数　　（h）婴儿入睡方式

图 5-3 离散数据分布

5.3.2 卡方检验

卡方检验是一种用于比较两个或多个分类变量之间是否存在显著关联的统计方法。其数学原理涉及观察频数和期望频数之间的差异，通过计算卡方统计量来判断差异是否具有统计学意义。卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为 0，表明理论值完全符合。其步骤如下：

① 提出原假设，

原假设（H0）：两个变量之间没有关联，其独立性为真。

备择假设（H1）：两个变量之间存在关联，其独立性不成立。

② 将总体 X 的取值范围分为 k 个互不相交的小区间 $A1$，$A2$，$A3$,…，Ak，如可取 $A1=(a0,\ a1]$，$A2=(a1,\ a2]$,...，$Ak=(ak-1,ak)$。其中 $a0$ 可取 $-\infty$，ak 可取 $+\infty$，区间的划分视具体情况而定，但要使每个小区间所含的样本值个数不小于 5，而区间个数 k 不要太大也不要太小。

③ 把落入第 i 个小区间的 Ai 的样本值的个数记作 fi，成为组频数（真实值），所有组频数之和 $f1+f2+...+fk$ 等于样本容量 n。

④ 当 H0 为真时，根据所假设的总体理论分布，可算出总体 X 的值落入第 i 个小区间 Ai 的概率 pi，于是，npi 就是落入第 i 个小区间 Ai 的样本值的理论频数（理论值）。

⑤ 当 H0 为真时，n 次试验中样本值落入第 i 个小区间 Ai 的频率 fi/n 与概率 pi 应很接近，当 H0 不真时，则 fi/n 与 pi 相差很大。基于这种思想，皮尔逊引进如下检验统计量 $\chi^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$，在 0 假设成立的情况下服从自由度为 $k-1$ 的卡方分布。

卡方检验属于非参数检验，不存在具体参数，且不需要有总体服从正态分布的假设，是用途非常广泛的一种假设检验方法，可以用于本文数据分析。

5.3.3 检验结果

根据婴儿年龄将样本分为 3 类，对于婴儿行为特征，睡醒次数，入睡方式与母亲身体和心理特征关联性采用卡方检验分别进行检验，部分求解结果如表【】所示（此处只列出部分值得讨论和分析的结果，全部的结果请见附录。）：

① 一月龄婴儿

项目		安 静 型（61 例）	中等型（74 例）	矛盾型（5 例）	x^2	P
婚姻状况	未婚	2	1	0		
	已婚	59	73	5	0.706	0.703
教育程度	小学	0	0	0		
	初中	3	5	0		
	高中	20	16	2		
	大学	12	16	2		
	研究生	26	37	1	4.387	0.624
分娩方式	自然分娩	60	72	5		
	剖腹产	1	2	0	0.294	0.742

② 二月龄婴儿

项目		安 静 型（45 例）	中等型（74 例）	矛盾型（9 例）	x^2	P
婚姻状况	未婚	1	3	0		
	已婚	44	71	9	0.623	0.733
教育程度	小学	1	0	0		
	初中	4	2	1		
	高中	10	23	1		
	大学	12	12	1		
	研究生	18	37	6	9.301	0.318
分娩方式	自然分娩	45	73	9		
	剖腹产	0	1	0	0.736	0.692

③ 三月龄婴儿

项目		安 静 型（14 例）	中等型（76 例）	矛盾型（31 例）	x^2	P
婚姻状况	未婚	1	5	1		
	已婚	13	71	30	0.508	0.776
教育程度	小学	0	0	1		
	初中	0	5	0		
	高中	4	15	8		
	大学	2	20	7		
	研究生	8	36	15	7.485	0.485
分娩方式	自然分娩	14	75	31		
	剖腹产	0	1	0	0.597	0.712

在对婴儿身体特征进行单因素分析发现：对于每个年龄段的婴儿，不同的婚姻状况，教育程度及分娩方式下，婴儿行为特征分布基本相同，差异无统计学意

义。（P 值均大于 0.05）

在对婴儿每晚睡醒次数进行单因素分析时，考虑婚姻状况影响，对于一月龄小孩，母亲的婚姻状况对睡醒次数有影响，差异具有统计学意义（x2=25.588,p=0.0005）。对于其他年龄小孩则无差异，原因可能小孩刚出生时，睡眠质量受家庭因素影响，单母亲一人很难负担起孩子睡眠的任务，婚姻状况对婴儿睡眠质量影响很大。考虑教育程度影响，对于一月龄小孩，母亲的教育程度对睡醒次数有影响，差异具有微弱统计学意义（x2=28.618,p=0.12367）。考虑分娩方式影响，对于每个年龄段的婴儿，不同分娩方式下，婴儿睡醒次数情况基本相同，无统计学差异。

在对婴儿入睡方式进行单因素分析时，考虑婚姻状况影响，对于每个年龄段小孩，不同婚姻状况下，哄婴儿入睡方式情况基本相同，无统计学差异。考虑教育程度影响，对于三月龄小孩，母亲的教育程度对睡醒次数有影响，差异具有微弱统计学意义（x2=22.992,p=0.11394）。考虑到分娩方式影响，对于二月婴儿，不同分娩方式下婴儿入睡方式不同，差异具有统计学意义（x2=10.720,p=0.02989）。但由于剖腹产母亲占总样本百分比较少，在后续数据处理中需要进一步分析。

5.4 连续数据关联性分析

5.4.1 数据可视化

附件中离散数据包括：母亲妊娠时间，母亲 CBTS，母亲 EPDS，母亲 HADS，婴儿整晚睡眠时间，具体分布情况如图 5-5 所示。

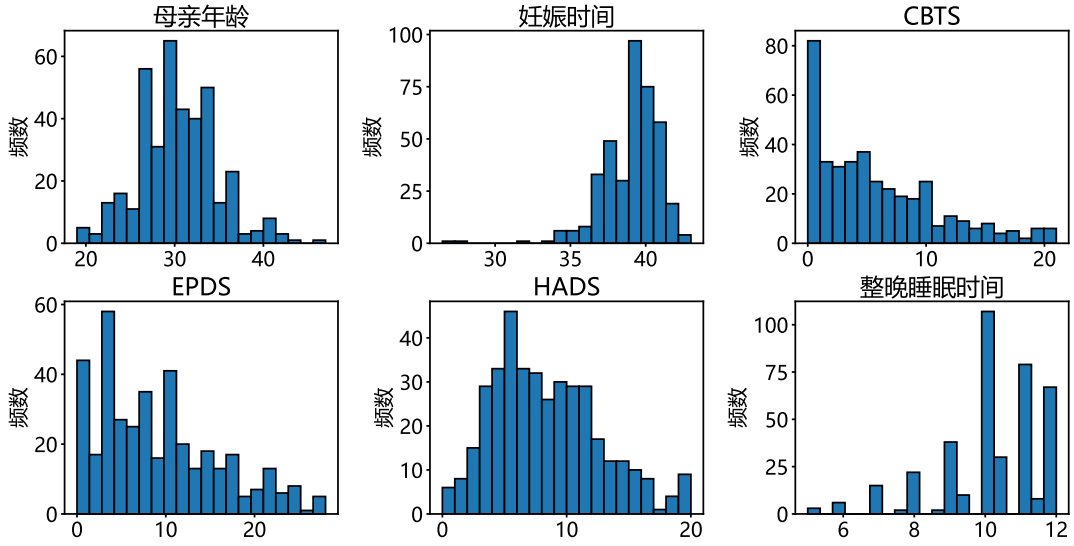


图 5-5 连续数据分布直方图

5.4.2 斯皮尔曼相关系数

斯皮尔曼相关系数法是一种无参数的检验方法， 用于度量变量之间联系的强弱。在没有重复数据的情况下，如果一个变量是另外一个变 量的严格单调函数，则斯皮尔曼相关系数为+1 或-1，称变量完全秩相关。使用该方法只 需要关

注个数值在变量内的排列顺序，如果两个变量的对应值在各组内的排序顺位是相同或类似的，则具有显著的相关性。

$$P_s=\frac{\sum_{i=1}^{N}(R_i-\tilde{R})(S_i-\tilde{S})}{\sqrt{\sum_{i=1}^{N}(R_i-\tilde{R})\sum_{i=1}^{N}(S_i-\tilde{S})^2}}=1-\frac{6\sum d_i^2}{N(N^2-1)}$$

其中，R_i 和 S_i 分别是观测值 i 的取值的等级，$\tilde{R}$ 和 $\tilde{S}$ 分别是变量 X 和变量 Y 的平均等级，N 是观测值的总数量，$d_i=R_i-S_i$ 表示二列成对变量的等级差数。

斯皮尔曼相关系数法适用于：只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级资料，不论两个变量的总体分布形态、样本容量 的大小如何，都可以用斯皮尔曼等级相关系数来进行研究。本题中各指标数据不满足正态分布，并且连续指标不属于分类变量，故选择斯皮尔曼相关系数法作为相关性分析方法使用相关系数用来衡量两个变量之间的线性相关程度。相关系数的值介于-1 和 1 之间，接近 1 表示正相关，接近-1 表示负相关，接近 0 表示无相关。P 值是用来判断相关系数是否具有统计学意义的指标，通常 P 值小于 0.05（通常以 0.05 作为显著性水平）时，认为相关系数是显著的，即具有统计学意义。

5.4.3 检验结果

与离散数据分析方法相同，首先将连续数据按年龄分为三组，对于婴儿整晚睡眠时间特征是连续数据，与母亲身体及心理特征的关联性使用斯皮尔曼相关系数法进行分析。求解结果如下表所示（此处只列出部分值得讨论和分析的结果，全部的结果请见附录。）：

① 一月龄婴儿

表 5-5 一月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	-0.047	0.579
婚姻状况	-0.046	0.583
教育程度	-0.039	0.647
妊娠时间	0.175	0.03**
分娩方式	0.073	0.392
CBTS	-0.015	0.862
EPDS	-0.132	0.117
HADS	-0.041	0.628

② 二月龄婴儿

表 5-6 二月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	0.104	0.242
婚姻状况	0.081	0.363
教育程度	-0.071	0.422
妊娠时间	0.049	0.586
分娩方式	-0.140	0.114

CBTS	-0.133	0.133
EPDS	-0.211	0.017**
HADS	-0.173	0.049**

③ 三月龄婴儿

表 5-7 三月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	<0.001	0.999
婚姻状况	0.047	0.605
教育程度	-0.106	0.245
妊娠时间	0.002	0.983
分娩方式	0.122	0.181
CBTS	-0.200	0.028**
EPDS	-0.127	0.166
HADS	-0.161	0.078*

通过婴儿整晚睡眠时间相关性分析可以看出，对于一月龄小孩，妊娠时间与整晚睡眠时间相关系数为 0.175，P 值为 0.03，这意味着妊娠时间与整晚睡眠时间存在一定程度正相关关系，并且这种相关关系在统计学上是显著的。其余因素与整晚睡眠时间无显著性相关关系。对于二月龄小孩，EPDS 和 HADS 和整晚睡眠时间相关系数为存在一定程度负相关关系，相关系数为-0.211 和-0.173，P 值为 0.017 和 0.049，并且这种相关关系在统计学上是显著的。对于三月龄小孩，CBTS 和 HADS 与婴儿整晚睡眠时间存在显著的负相关关系，相关系数为-0.200 和-0.161，p 值分别为 0.028 和 0.078。

由整晚相关系数与母亲身体特征和心理特征的分析可以看出，在婴儿一月龄大时，母亲的分娩情况是影响婴儿睡眠的主导因素，原因可能是母亲分娩情况直接与刚出生时的婴儿身体情况相关，影响直接体现到婴儿睡眠质量上。而在婴儿两月与三月大时，母亲的心理状况与婴儿睡眠时间有显著的负相关性，且侧重心理方面不同，伴有抑郁情绪母亲养育婴儿夜晚睡眠时间更短，与既往研究报道结果相似，另一方面，既往研究提示婴儿的睡眠问题也会增加母亲出现产后抑郁的症状，两者是可以相互印证的。

① 一月龄婴儿

表 5-8 三月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	-0.141	0.096*
妊娠时间	0.043	0.616
GBTS	0.172	0.042**
EPDS	0.206	0.015**
HADS	0.161	0.057*

② 二月龄婴儿

表 5-9 三月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	-0.075	0.398
妊娠时间	-0.038	0.672
GBTS	0.193	0.029**

EPDS	0.190	0.031**
HADS	0.082	0.357

③ 三月龄婴儿

表 5-10 三月龄婴儿整晚睡眠时间相关系数

项目	相关系数	P
母亲年龄	-0.048	0.599
妊娠时间	-0.064	0.484
GBTS	0.121	0.186
EPDS	0.103	0.031**
HADS	0.155	0.090**

如表 5-8～表 5-10 所示，针对婴儿行为特征的相关性分析发现，在不同月龄大的婴儿中，其行为特征都与母亲的心理状况呈正相关关系，且这种关系在统计学上是显著的，这表明在一定程度上，母亲的心理状况越差，婴儿的行为特征愈偏向矛盾型。对于婴儿夜晚睡醒次数的相关分析也证明了婴儿的睡眠质量与母亲的心理健康状况具有关联，且不同月龄的婴儿对于母亲的不同心理状况变化敏感性不同，而对于入睡方式的相关性分析都无法拒绝原假设，即母亲的身体指标和心理指标对婴儿入睡方式没有显著影响（篇幅有限，详情见附录）。

5.4.4 总结

母亲的身体指标与不同月龄的婴儿呈现不同程度的关联性，而母亲的心理指标对婴儿的身体行为特征和睡眠质量均有影响。此外，不同月龄的婴儿对于母亲的不同心理指标的敏感度也不相同。

与此同时，本文查阅文献得知，母亲的抑郁情绪对婴儿睡眠的影响可能与不良情绪影响亲子交流以及母亲对婴儿睡眠行为的反应有关，具有抑郁症状的母亲可能导致婴儿情感体验不良，进而影响婴儿睡眠行为的健康发展，且这种影响会随着月龄改变，与本文研究结果相同，在后续建模分析过程中将作为重点因素考虑。

六、 问题二模型建立和求解

6.1 数据分析

由问题一中的分析可知，对不同年龄的婴儿，母亲的身体与心理状况对婴儿行为特征有着不同的影响。其次，母亲的分娩方式与婚姻状况中，剖宫产与未婚占总体比率极低，数据量少，不具有分析价值，因此在后续的模型预测中将不再讨论。

针对一月婴儿，由第一问数据关联性分析可知‘母亲年龄’、‘CBTS’、‘EPDS’、‘HADS’对‘婴儿行为特征’有较强的相关性，而其母亲的其他状态对其影响不具有显著的相关性；同理，对于二月婴儿来说，‘CBTS’、‘EPDS’对‘婴儿行为特征’有较强的相关性，应该给予更大的关注；而对于三月婴儿，母亲的‘CBTS’、‘EPDS’、‘HADS’对‘婴儿行为特征’有较强的相关性。

综上，本文针对不同年龄段的婴儿，运用不同的母亲身体与心理因素，对‘婴儿行为特征’进行归类。

6.2 模型建立

6.2.1 XGboost 原理

XGBoost（eXtreme Gradient Boosting）是一种高效的机器学习算法，用于解决分类和回归问题。它基于决策树集成技术，并利用梯度提升框架来提高模型的性能。它结合了梯度提升和正则化技术，通过最小化损失函数的梯度来不断优化模型的预测能力，同时避免过拟合问题。其基本步骤如下：

Step1:初始化一个初始预测值，通常为目标变量的均值（对于回归问题）或各类别样本的概率（对于分类问题）。

Step2:计算当前模型对于每个样本的损失函数的梯度和二阶导数，这些信息将用于构建下一个弱学习器。

Step3:根据计算得到的梯度和二阶导数，构建一个决策树来拟合这些残差（负梯度），以逐步减小模型的预测误差。在构建决策树时，使用正则化技术来避免过拟合，如最大树深度限制、叶子节点权重的 L1 和 L2 正则化等。

Step4:将新构建的决策树与之前的模型进行结合，更新模型的预测值。

Step5:重复执行步骤二至四，直到达到预定的迭代次数或损失函数收敛。

Step5:得到最终的强大模型，它由多个弱学习器的加权组合构成，使得模型在训练数据上的预测效果达到最优。

XGBoost 还引入了特定的技巧来加快计算速度和提高模型的准确性，比如特征列排序、数据块并行处理、直方图近似算法等。这些优化使得 XGBoost 在许多机器学竞赛和实际应用中都表现出色。

6.2.2 随机森林原理

随机森林(Random forest,RF)是由美国科学家 Leo Breiman 在 1996 年提出的一种利 用多个树分类器进行分类和预测的方法，它是利用 bootstrap 重抽样方法从原始样本中抽取 多个样本，对每个 bootstrap 样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。

随机森林是 Bootstrap aggregating(Bagging)算法中的一种，Bagging 的算法流程图如图【】所示

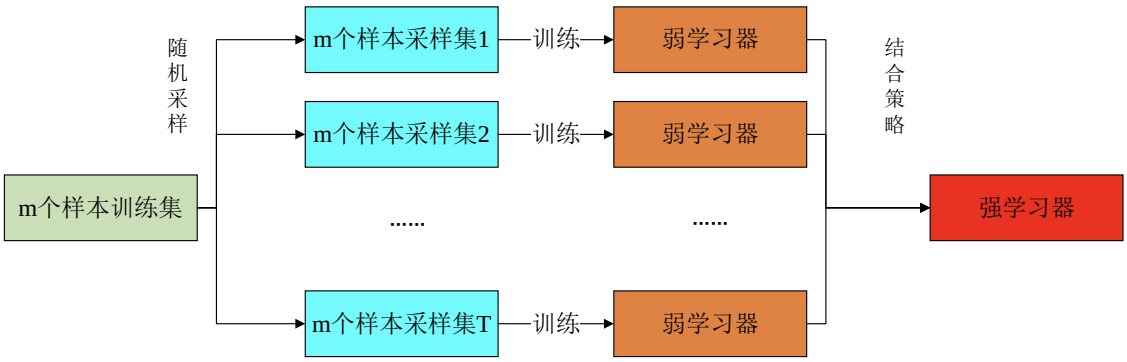


图 6-1 Bagging 方法原理图

Bagging 是一种集成学习方法，通常用于减少数据集中的方差。在 Bagging 学习方法中，训练集中的随机数据样本是通过替换选择的，意味着模型训练时可以多次选择单个数据点。在生成几个数据样本后，这些弱模型会被独立训练，并且根据任务的类型，例如回归或分类，获得更准确的估计值。

作为 Bagging 方法中的一种，大量的理论和实证研究都证明了随机森林方法具有很高的预测准确率，具有较好的鲁棒性，且不容易出现过拟合的情况。随机森林算法可以用于处理回归、分类、聚类以及生存分析等问题。

假设要我们要生成 T 个决策树，原始的训练集包含 m 个样本，特征个数为 n，那么随机森林算法的整个流程如下：

Step1:从原始的包含了 m 个样本的数据集中随机地有放回地采样 m 次，得到 m 个样本（会有重复样本）。

Step2:使用采样生成的数据集训练一个决策树。

Step3:重复步骤 1 和 2 共 T 次，得到 T 个训练好的决策树

Step4:采用投票法（分类树）或简单平均法（回归树）从 T 个决策树的预测结果中生成最终的结果。

可以看到，由于采用随机地有放回地采样得到训练集，因此不同的树用到的训练集将会有所差异；其次，每个树在结点分裂时并非是从所有的特征中选择最优特征和划分点，而是先随机地从所有特征 n 中选择一个包含了 k 个特征的特征子集，然后从特征子集中 10 选择最优特征和划分点，通过改变 k 的大小可以控制随机性的引入程度。随机森林中的“随机”含义指的就是前面说的这两个随机：数据随机和特征随机。

6.2.3 模型对比

根据本章中数据分析所述，将婴儿不同年龄下与婴儿行为特征相关性较强的

因素作为模型输入以预测‘婴儿行为特征’。表 6-1、表 6-2 展示了两种不同模型对‘婴儿行为特征’的分类准确度。

表 6-1 XGBT 预测结果

婴儿年龄	输入特征	准确率
一月	‘母亲年龄’、‘CBTS’、‘EPDS’、‘HADS’	60.71%
二月	‘CBTS’、‘EPDS’	61.54%
三月	‘CBTS’、‘EPDS’、‘HADS’	52.00%

表 6-2 随机森林预测结果

婴儿年龄	输入特征	准确率
一月	‘母亲年龄’、‘CBTS’、‘EPDS’、‘HADS’	64.29%
二月	‘CBTS’、‘EPDS’	69.23%
三月	‘CBTS’、‘EPDS’、‘HADS’	72.00%

6.3 总结

由于婴儿行为特征与母亲行为与心理的相关性均较弱而与婴儿自身的睡眠质量特征关系较强，导致以上模型并没有一个很好的预测效果，但是在对婴儿行为特征的分类中随机森林相较于 XGboost 还是有一个比较好的效果，因此，本文选用随机森林对后 20 组数据进行‘婴儿行为特征’的预测。结果如表 6-3 所示。

表 6-3 预测结果

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

七、 问题三模型建立与求解

7.1 子问题 1 规划模型建立与求解

① 确定目标函数

由问题三已知条件，CBTS、EPDS、HADS 的治疗费用相对于患病程度（得分）的变化率均与治疗费用呈正比，得微分方程如下式所示。

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

式中，$f(x)$ 表示治疗费用，$\frac{df}{dx}$ 表示治疗费用相对于得分的变化率，k 表示常数。

求解式得到治疗费用的函数如下式所示。

$$f(x) = ce^{kx}$$

式中，c 表示常数。

将问题三表 1 所给的 CBTS、EPDS、HADS 的相关数据分别代入上式，得到 CBTS、EPDS、HADS 的治疗费用如下式所示。

$$\begin{cases} f_1(x_1) = c_1 e^{k_1 x_1} = 200e^{0.8811x_1} \\ f_2(x_2) = c_2 e^{k_2 x_2} = 500e^{0.6649x_2} \\ f_3(x_3) = c_3 e^{k_3 x_3} = 300e^{0.7459x_3} \end{cases}$$

式中，$f_1(x)$ 代表 CBTS 的治疗费用函数，$f_2(x)$ 代表 EPDS 的治疗费用函数，$f_3(x)$ 代表 HADS 的治疗费用函数，x_1 代表 CBTS 的得分减小量，x_2 代表 EPDS 的得分减小量，x_3 代表 HADS 的得分减小量。

由此，得到目标函数如下式所示。

$$Min\ f(x_1, x_2, x_3) = f_1(x_1) + f_2(x_2) + f_3(x_3) = 200e^{0.8811x_1} + 500e^{0.6649x_2} + 300e^{0.7459x_3}$$

② 确定约束条件

根据编号为 237 的婴儿的 CBTS、EPDS、HADS 得分，及题目三要求使婴儿的行为特征从矛盾型变为中等型，得到约束条件如下式所示。

$$s.t. \begin{cases} 0 \le x_1 \le 15 \\ 0 \le x_2 \le 22 \\ 0 \le x_3 \le 18 \\ f(X) = 2 \end{cases}$$

式中，$f(X)$ 表示问题二的关系模型，2 表示矛盾型。

③ 规划求解

基于得到的目标函数和约束条件，建立规划模型如下式所示。

$$Min\ f(x_1,x_2,x_3)=f_1(x_1)+f_2(x_2)+f_3(x_3)=200e^{0.8811x_1}+500e^{0.6649x_2}+300e^{0.7459x_3}$$

$$s.t.\begin{cases}0\le x_1\le 15\\0\le x_2\le 22\\0\le x_3\le 18\\f(X)=2\end{cases}$$

本题使用 Python 中的 Minimize 函数，根据上述目标函数和约束条件进行规划的求解。

由于 Minimize 函数初值的设置对其收敛性有较大影响，本文在约束条件确定的范围内以固定间隔选取初值组合，并以相应的初值组合进行局部最优解的搜索。

其搜索过程如表 7-1 所示。

表 7-1 搜索算法流程

搜索算法
1、使用第一组初值组合进行规划，得到局部最优解。
2、判断是否有剩余组合，若有更换组合，转至步骤 1，若没有，转至步骤 3。
3、比较得到的所有局部最优解的大小，确定全局最优解。
4、结束。

执行程序后的得到的最优解如下表 7-2 所示，迭代过程如下图 7-1 所示。

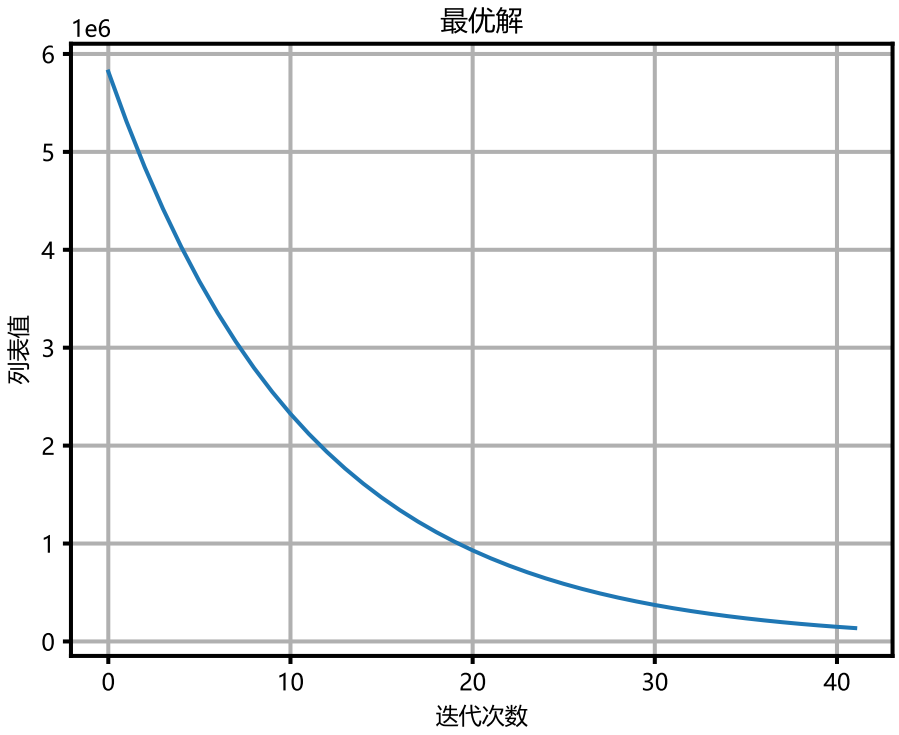


图 7-1 迭代过程

表 7-2 规划结果

x_1	x_2	x_3	$f(x_1,x_2,x_3)$
5.51	8.12	0	136263.8

7.2 子问题 2 规划模型建立与求解

由题意得，约束条件变更为使婴儿的行为特征从矛盾型变为安静型，新的规划模型如下式所示。

$$Min\ f(x_1,x_2,x_3)=f_1(x_1)+f_2(x_2)+f_3(x_3)=200e^{0.8811x_1}+500e^{0.6649x_2}+300e^{0.7459x_3}$$

$$s.t.\begin{cases}0\le x_1\le 15\\0\le x_2\le 22\\0\le x_3\le 18\\f(X)=1\end{cases}$$

执行程序后的得到的最优解如下表 7-3 所示，迭代过程如下图 7-2 所示。

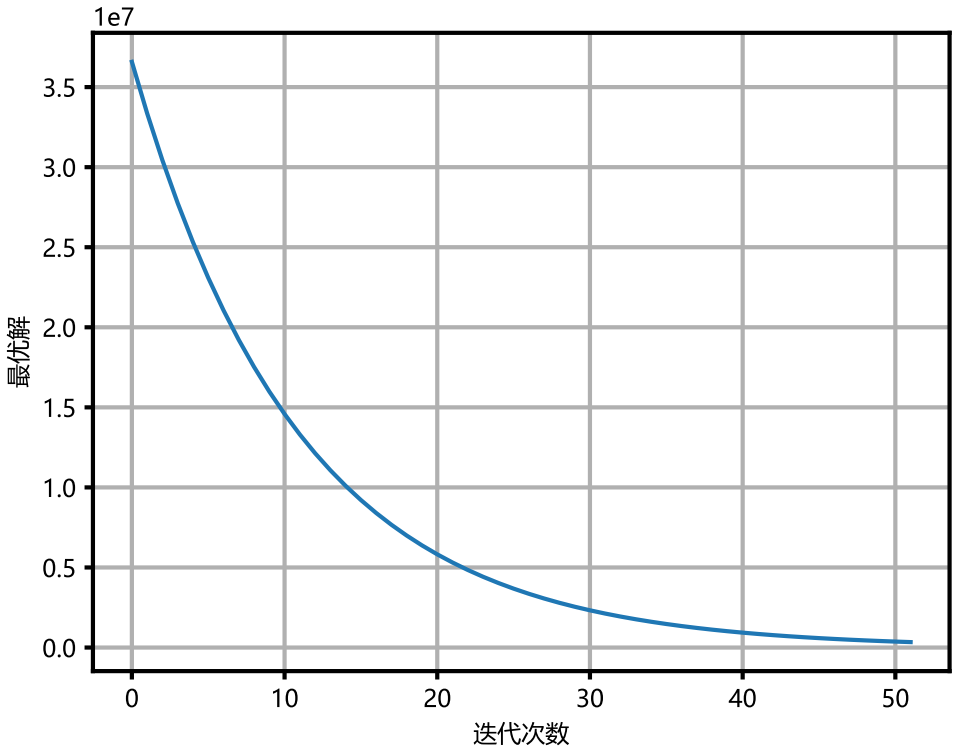


图 7-2 迭代过程

表 7-3 规划结果

x_1	x_2	x_3	$f(x_1,x_2,x_3)$
6.46	9.52	0	339827.28

八、 问题四模型建立与求解

8.1 打分模型的构建

建立婴儿综合睡眠质量与身体指标及心理指标关联模型前，需要先对婴儿睡眠质量进行综合评价。为了对整个系统进行综合评价，必须将各指标进行标准化处理，使定性指标科学地得以量化，将定量指标进行一致化处理和无量纲化处理。

8.1.1 综合评价数据处理

（1） 定量指标的一致化处理

一般来说，在评价指标体系中、可能会同时存在极大型指标、极小型指标、居中型指标和区间型指标。对于本题，整晚睡眠时间为极大型指标，睡醒次数为极小型指标，入睡方式未知，在评价前需要对其进行数据处理。

睡醒次数处理：对极小型指标 x_j，将其转化为极大型指标时，只需对指标 x_j 取倒数：

$$x'_j = \frac{1}{x_j}$$

入睡方式处理：由卡方检验可知，入睡方式与睡醒次数及婴儿行为特征存在较强相关性，且该相关性具有统计学意义。说明不同婴儿入睡方式对婴儿睡眠质量存在影响，婴儿对不同入睡方式存在偏好，因此可对不同婴儿入睡方式进行分类，按照睡醒次数的均值作为权重进行替代，求取均值结果如下表 8-1 所示：

表 8-1 入睡方式一致化处理结果

项目	哄睡法	抚触法	安抚奶嘴法	环境营造法	定时法
睡醒次数均值	2.426	1.391	1.700	0.861	1.800
评价权重	0.297	0.170	0.207	0.105	0.220

（2） 定量指标的归一化处理

无量纲化，也称为指标值的规范化，是通过数学变换来消除原始指标值的单位及其数值数量级影响的过程。本文对三个睡眠质量相关指标使用标准样本变换法进行归一化处理。

$$a^*_{ij} = \frac{a_{ij} - u_j}{s_j}$$

其中样本均值 $u_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$，样本标准差 $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - u_j)^2}$，$a_{ij}^*$ 成为观测标准值。

8.1.2 熵值法

熵值法是一种多准则决策分析方法，常用于对多个指标进行综合评价或权重分配。它通过计算指标的熵值来确定每个指标的权重，从而实现对指标的相对重要性排序。

熵值法的基本思想是：对于一个具有 n 个指标的决策问题，每个指标都有一个取值矩阵，其中每个元素代表决策方案在该指标上的得分。首先，将每个指标的得分标准化到[0, 1]范围内，然后计算每个指标的熵值。熵值反映了指标的信息量，即信息越多，熵值越大；信息越少，熵值越小。熵值法计算过程如下：

① 计算在第 j 项指标下第 i 个评价对象的特征比重。设第 i 个评价对象的第 j 项指标观测值的标准化数据 $b_{ij} > 0, i = 1, 2, \cdots, m$，则在第 j 项指标第 i 个评价对象的特征比重为:

$$p_{ij} = \frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, m$$

② 计算第 j 项指标的熵值为：

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} \ln p_{ij}$$

不难看出，第 j 项指标的观测值差异越大，熵值越小；反之，熵值越大。

③ 计算第 j 项指标的差异系数为

$$g_j = 1 - e_j$$

第 j 项指标的观测值差异越大，则差异系数 g_j 就越大，第 j 项指标就越重要。

④ 确定第 j 项指标的权重系数：

$$w_j = \frac{g_j}{\sum_{k=1}^{m} g_k}, j = 1, 2, \cdots m$$

⑤ 计算第 i 个评价对象的综合评价值：

$$f_i = \sum_{j=1}^{m} w_j p_{ij}$$

采用熵值法对处理后的数据进行权重评估，评估结果如表 8-2 所示，其中整晚睡眠时间的权重为 0.448，说明其在项目评估中占据了较大的比重，这可能是因为充足的睡眠时间对于婴儿身心健康有着重要的影响。：

表 8-2 熵值法权重计算结果

项目	整晚睡眠时间	睡醒次数	入睡方式
权重	0.448	0.275	0.278

8.1.3 Topsis 法

TOPSIS（Technique for Order of Preference by Similarity to Ideal Solution）法是一种多属性决策分析方法，用于帮助决策者在多个可选方案中选择最佳方案。该方法基于距离的概念，通过计算每个方案与理想解和负理想解之间的距离，来对方案进行排序，越接近理想解的方案排名越高。计算过程如下：

设综合评价问题含有 n 个评价对象 m 个指标，相应的指标观测值分别为：

$$a_{ij}, i=1,2,\cdots,n; j=1,2,\cdots m,$$

① 将评价指标进行预处理，并构造评价矩阵 $B=(b_{ij})_{n\times m}$

② 确定正理想解 C^+ 和负理想解 C^-。

设正理想解 C^+ 的第 j 个数学为 c^+_j，则 $C^+=[c^+_1,c^+_2,\cdots,c^+_m]$；负理想解 C 的第 j 个数学为 c^-_j，即 $C^-=[c^-_1,c^-_2,\cdots,c^-_m]$，则

$$c^+_j=\max_{1\le i\le n} b_{ij},\ j=1,2,\cdots m,$$

$$c^-_j=\min_{1\le i\le n} b_{ij},\ j=1,2,\cdots,m,$$

③ 计算各评价对象到正理想解及负理想解的距离。

各评价对象到正理想解的距离为

$$s^+_i=\sqrt{\sum_{j=1}^{m}(b_{ij}-c^+_j)^2},i=1,2,\cdots n,$$

各评价对象到负理想解的距离为

$$s^-_i=\sqrt{\sum_{j=1}^{m}(b_{ij}-c^-_j)^2},i=1,2,\cdots n,$$

④ 计算各评价对象对理想解的相对接近度

$$f_i=s^-_i/(s^-_i+s^+_i),i=1,2,\cdots n,$$

⑤ 按 f_i 由大到小排列各评价对象的优劣次序。

结合前节熵值法确实的权重结合 topsis 法对婴儿的睡眠质量进行打分，部分评估结果如下表 8-3：

表 8-3 Topsis 评分结果

项目	整晚睡眠时间	睡醒次数	入睡方式	Topsis Score	睡眠质量
128	12	0	0.296	1.000	优
212	12	0	0.296	1.000	优
63	12	1	0.296	0.943	优
……	……	……	……	……	……
373	5	3	0.170	0.386	差
54	5	5	0.170	0.317	差
319	7	5	0.105	0.303	差

8.2 分类模型的构建

8.2.1 学习器原理

GBDT 算法的全程是梯度提升树(Gradient Boosting Decision Tree)，其中 DT 指的是决策树(Decision Tree)；GB 指的是梯度提升(Gradient Boosting)，梯度提升是一种用于回归、分类和其他任务的方法。GBDT 的含义就是用 Gradient Boosting 的策略训练出来的 DT 模型。模型的结果是一组回归分类树组合(CART Tree Ensemble)：$T_1, T_1 .....T_K$,其中T_j学习的是之前 j-1 棵树预测结果的残差。

GBDT 算法的流程如下：

假设有训练集样本$T = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$作为输入，并且最大迭代次数设为 T，损失函数 L。输出是由多个基学习器训练出来的强学习器 $f(x)$，则使用 GBDT 算法用于回归预测的流程如下：

弱学习器的初始化得

$$f_0(x) = \arg\min \sum_{i=1}^{m} L(y_i, c)$$

对于迭代次数$t = 1, 2, ..., T$，首先对样本$i = 1, 2, ..., m$，计算负梯度得：

$$r_{ti} = \left[\frac{\partial L(y_i, f(x_i))}{\partial L f(x_i)} \right]_{f(x) = f_{t-1}(x)}$$

然后利用$(x_i, r_{tj})(i = 1, 2, ...m)$，拟合一颗 CART 回归树，得到第 t 颗回归树，其对应的叶子节点区域为R_{ti}， $j = 1, 2, ...J$ 。其中 J 为回归树 t 的叶子节点的个数。

再对叶子区域 $j = 1, 2, ...J$ ，计算最佳拟合值得：

$$C_{tj} = \arg\min \sum_{x \in R_{ij}} L(y_i, f_{t-1}(x_i) + c)$$

最后更新强学习器得：

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in R_{tj})$$

得到强学习器 $f(x)$ 的表达式

$$f(x) = f_T(x) = f_0(x) + \sum_{t=1}^{T} \sum_{j=1}^{J} c_{tj} I(x \in R)$$

GBDT 的原理流程图如图 8-1 所示：

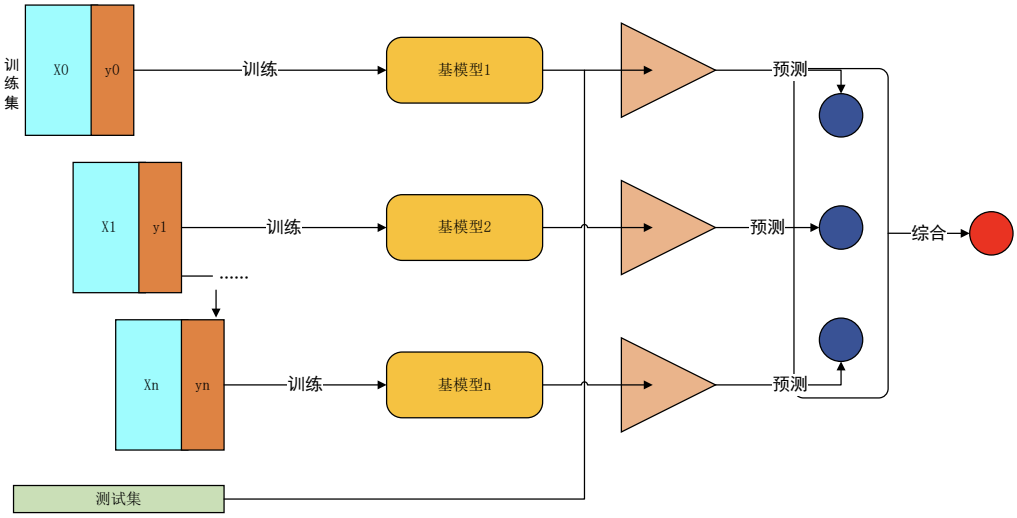


图 8-1 集成学习方法原理图

针对睡眠质量的预测，本文在上述问题的基础上进行进一步数据处理和特征提取，取前 8 个主要变量作为输入，本文利用 GBDT 集成学习算法构建了睡眠质量预测模型。

8.2.2 识别模型验证

根据前面的原理，我们对训练样本和测试样本进行了归一化处理。归一化公式如下：（这里填写归一化的具体公式）。该步骤的目的是加快模型训练速度，并消除不同样本之间量纲不同带来的影响。

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

在归一化之后对训练样本按比例进行了划分，将 80%的数据作为训练集，20%的数据作为测试集。采用混淆矩阵和热力图来可视化分类结果，其模型训练结果如下图 8-2 所示：

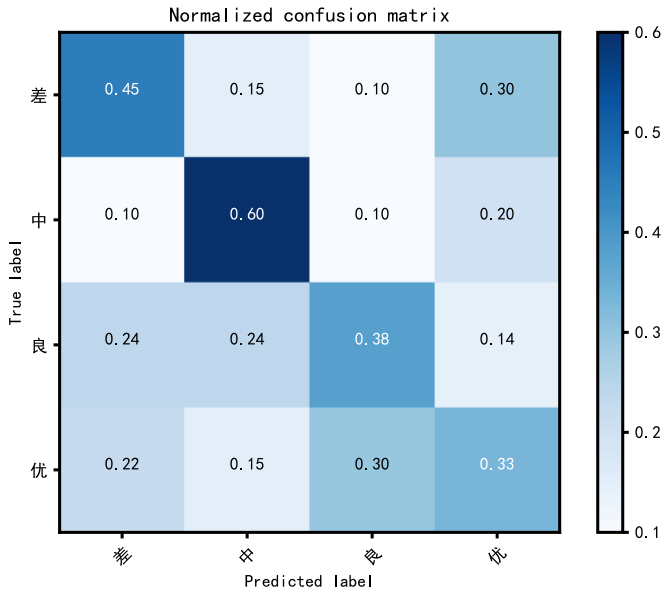


图 8-2 混淆矩阵热力图

根据图 8-2 的结果显示，我们成功地训练出了一个模型，通过母亲的身体和心理指标，对婴儿的睡眠质量进行基本的预测。这对于婴儿的健康和家庭的幸福都具有重要的意义。然而，我们也意识到数据质量对于模型的性能有着重要的影响，特别是在睡眠质量优良和良好之间的分类效果较差。这可能是由于数据集中缺乏足够的多样性和数量，导致模型对这两个类别分类效果较差，后续可以继续填充数据进行训练以优化模型分类效果。用该训练模型预测题所要求数据，结果如下：

表 8-4 预测结果

编号	婴儿行为特征预测
391	中
392	中
393	差
394	良
395	中
396	良
397	差
398	优
399	差
400	中
401	差
402	优
403	良
404	优
405	良
406	差
407	优
408	优
409	优
410	优

九、 问题五模型建立与求解

9.1 子问题 1 规划模型建立与求解

由问题五题意得，目标仍是治疗费用最低，但在问题三子问题 1 的基础上新增一个约束条件，即让 238 号婴儿的睡眠质量评级为优，新的规划模型如下式所示。

$$Min\ f(x_1,x_2,x_3)=f_1(x_1)+f_2(x_2)+f_3(x_3)=200e^{0.8811x_1}+500e^{0.6649x_2}+300e^{0.7459x_3}$$

$$s.t.\begin{cases}0\le x_1\le 15\\0\le x_2\le 22\\0\le x_3\le 18\\f(X)=2\\g(X)=4\end{cases}$$

本题依旧采用 Python 中的 Minimize 函数，根据上述目标函数和约束条件进行规划的求解。

执行程序后的得到的最优解如下表 9-1 所示，迭代过程如下图 9-1 所示。

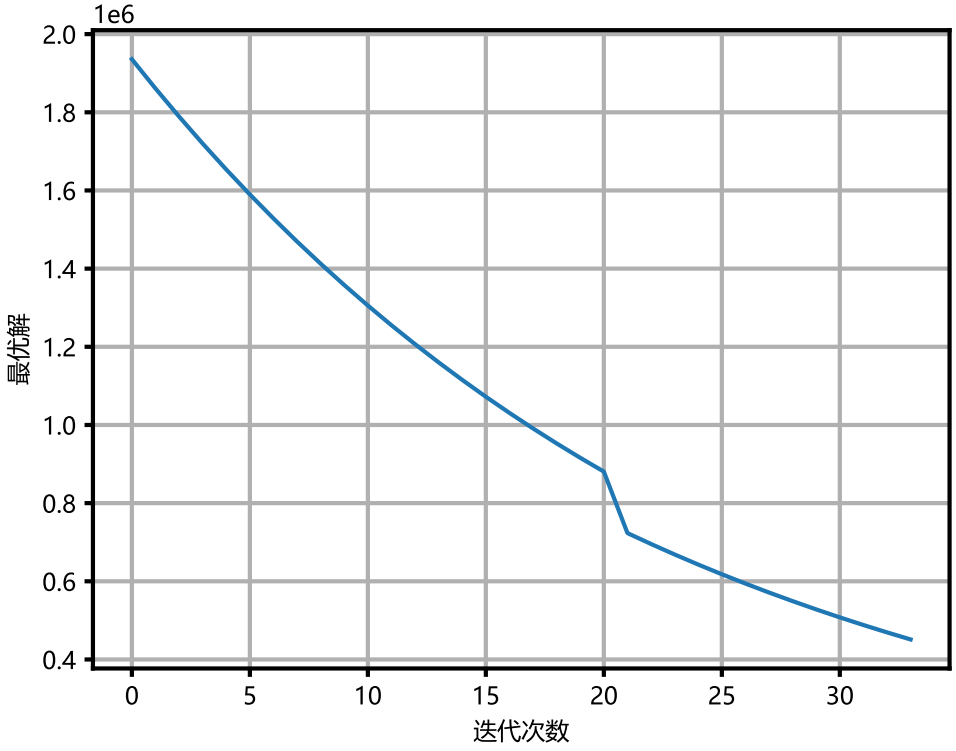


图 9-1 迭代过程

表 9-1 规划结果

x_1	x_2	x_3	$f(x_1, x_2, x_3)$
9	13.5	0	451188.57

9.1 子问题 2 规划模型建立与求解

由问题五题意得，目标仍是治疗费用最低，但在问题三子问题 2 的基础上新增一个约束条件，即让 238 号婴儿的睡眠质量评级为优，新的规划模型如下式所示。

$$Min\ f(x_1, x_2, x_3) = f_1(x_1) + f_2(x_2) + f_3(x_3) = 200e^{0.8811x_1} + 500e^{0.6649x_2} + 300e^{0.7459x_3}$$

$$s.t.\begin{cases} 0 \le x_1 \le 15 \\ 0 \le x_2 \le 22 \\ 0 \le x_3 \le 18 \\ f(X) = 1 \\ g(X) = 4 \end{cases}$$

执行程序后，发现无法同时满足安静型和优的条件，算法无法收敛。本文认为发生此现象的原因是原始数据集质量不高，造成问题二和问题四模型的准确度不够高，导致误差积累，进而使得算法无法收敛。

十、 模型的评价、改进

10.1 模型的评价

10.1.1 模型的优点

（1）数据处理方面：为了保证建模数据的高质量，建立模型之前，我们对数据进行了统一的处理并按类别分类进行相关性分析，保证了数据的完整性和准确性。
（2）准确性：为提高建模的准确性，考虑因素之间相关性影响，结多因素对婴儿身体状况的影响，取得了较不错的结果。
（3）多样性：在解决问题的过程中，采用了多种模型，包括 XGBT、随机森林，梯度提升树等等，通过多种方法进行对比，有 利于获得最佳的建模结果。

10.1.2 模型的缺点

（1）预测误差：即使使用最佳的模型与特征，对于婴儿行为指标及睡眠质量的预测仍存在一定误差，婴儿的成长的一个多因素作用的结果，且采用问卷对母亲状况的评估并非完全可靠。这些外在因素可能会导致模型预测结果存在一定的不准确性和误差。
（2）专业知识欠缺，如果对模型进行约束和先验信息的引入，可以提高模型对母亲及婴儿身体心理指标的理解和预测能力。

10.2 模型的改进

（1）增加更多特征：目前的数据只包括母亲笼统的特征描述，且多数为定性分析，考虑增加更多母亲相关信息，例如陪伴婴儿时间，性格评估等，这样可以更全面对母亲状况进行评估，从而提高预测模型的表现。
（2）当数据集存在一定的不平衡时，加入过采样/欠采样过程，使得模型可以处理不平衡数据集上的分类预测问题

参考文献

[1]黄巧瑜.儿童保健门诊 108 例婴儿睡眠现状分析[J].世界睡眠医学杂志,2022,9(10):1851-1853.

[2]陈孝颖. 某三甲医院 1～11 月龄婴儿睡眠现况及影响因素研究[D].南昌大学,2022.

[3]董响娇,李华珍.母亲孕期睡眠与婴儿早期夜间睡眠的相关性研究[J].全科护理,2022,20(03):404-406.

[4]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.

[5]闻芳,黄小娜,冯围围等.6 月龄婴儿夜晚睡眠-觉醒模式现状及影响因素分析[J].中国儿童保健杂志,2018,26(01):7-10+14.

[6]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.

[7]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.

附录（另起一页）

问题 1：

```
#############Spearman 等级相关系数###############
import pandas as pd
from scipy.stats import spearmanr
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号
# 读取 Excel 表格，假设数据在 Sheet1 中的 B 列和 J 列
data = pd.read_excel('附件.xlsx', sheet_name='Sheet1')
data = data.iloc[:, 1:15]
# 将婴儿行为进行转换为有序等级数据
def convert_behavior(behavior):
    if behavior == '安静型':
        return 1
    elif behavior == '中等型':
        return 2
    elif behavior == '矛盾型':
        return 3
category_mapping = {"安静型": 0, "中等型": 1, "矛盾型": 2}

# 使用 map 函数将"category"列转换为对应的数值
data["婴儿行为特征"] = data["婴儿行为特征"].map(category_mapping)
data1 = data[data.iloc[:, 10] == 1 ]
data2 = data[data.iloc[:, 10] == 2 ]
data3 = data[data.iloc[:, 10] == 3 ]




corr = data.corr()
ax = plt.subplots(figsize=(20, 16))#调整画布大小
ax = sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5,annot_kws={"fontsize": 20})
# 设置刻度字体大小
plt.xticks(fontsize=20)
plt.yticks(fontsize=20)
cbar = ax.collections[0].colorbar
cbar.ax.tick_params(labelsize=20)
ax.set_xticklabels(ax.get_xticklabels(), rotation=45)
ax.set_yticklabels(ax.get_yticklabels(), rotation=45)
# plt.savefig('热力图.svg',dpi=800)
```

```
##############卡方检验###############
import pandas as pd
from scipy.stats import chi2_contingency

# 读取 Excel 表格，假设数据在 Sheet1 中的两列分别是'婚姻状况'和'婴儿行为'
# data = pd.read_excel('垃圾.xlsx', sheet_name='Sheet1', usecols=['婚姻状况', '婴儿行为特征'])

# 创建一个交叉表，用于计算卡方检验
# 进行卡方分布检验
#需要进行卡方检验的数据
a = '睡醒次数'
b = '入睡方式'
cross_tab = pd.crosstab(data[a], data[b])
chi2, p_value, dof, expected = chi2_contingency(cross_tab)
print("卡方值:", chi2)
print("p 值:", p_value)

cross_tab = pd.crosstab(data1[a], data1[b])
chi2, p_value, dof, expected = chi2_contingency(cross_tab)
print('对于一月小孩')
print("卡方值:", chi2)
print("p 值:", p_value)

cross_tab = pd.crosstab(data2[a], data2[b])
chi2, p_value, dof, expected = chi2_contingency(cross_tab)
print('对于二月小孩')
print("卡方值:", chi2)
print("p 值:", p_value)

cross_tab = pd.crosstab(data3[a], data3[b])
chi2, p_value, dof, expected = chi2_contingency(cross_tab)
print('对于三月小孩')
print("卡方值:", chi2)
print("p 值:", p_value)
# 打印结果



##############相关性检验###############
import pandas as pd
```

```
from scipy.stats import chi2_contingency

# 读取 Excel 表格，假设数据在 Sheet1 中的两列分别是'婚姻状况'和'婴儿行为'
# data = pd.read_excel('垃圾.xlsx', sheet_name='Sheet1', usecols=['婚姻状况', '婴儿行为特征'])

# 创建一个交叉表，用于计算卡方检验
# 进行卡方分布检验
#需要进行卡方检验的数据
a = '睡醒次数'
b = '入睡方式'
correlation, p_value = spearmanr(data1[a], data1[b])

# 打印结果
print("Spearman 等级相关系数:", correlation)
print("p 值:", p_value)

correlation, p_value = spearmanr(data2[a], data2[b])

# 打印结果
print("Spearman 等级相关系数:", correlation)
print("p 值:", p_value)

correlation, p_value = spearmanr(data3[a], data3[b])

# 打印结果
print("Spearman 等级相关系数:", correlation)
print("p 值:", p_value)
```