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		CM*******

基于集成学习的大数据分析模型对母亲身心健康与婴儿成长关联的研究

摘要

母亲是孩子忠实的后盾，也是孩子一生的良师益友，母亲带来的影响会伴随孩子的一生。本文基于母亲的身心健康数据和婴儿的睡眠质量调查数据，综合运用机器学习模型融合算法、线性回归、运筹优化算法和 TOPSIS—熵权法等混合模型，结合实际情况，对数据进行建模和分析，揭示母亲身心健康与婴儿成长之间的关联。

针对问题一：本文首先对进行了**数据预处理**，处理了异常值和缺省值，将部分指标数值化。再将母亲的身体指标和心理指标归为一组“母亲指标”变量，婴儿的行为特征和睡眠质量归为一组“婴儿指标”变量，并假设两组典型变量服从正态分布。使用软件 Spss 对这两组典型变量进行**典型相关分析**，进行相关性检验，得出结论：在 99%的置信水平下认为两组变量存在典型相关关系。通过标准化典型相关系数，选择第一对典型变量建立典型相关模型，证明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存在相关关系。

针对问题二：本文首先根据机器学习的习惯与特点对数据进行预处理，包括部分指标标签编码、中英文本转换以及数据集划分。使用 **k 折交叉验证**方法来高效利用数据集，再使用基于 Python 的 scikit-learn 库选择了几种常见的机器学习算法：**随机森林、多层感知机、K 最近邻、决策树、XGBoost** 和**逻辑回归**进行模型训练，最后使用**硬投票法**作为**集成学习方法**，将这些模型的预测结果进行投票并融合，成功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，得到最终的行为特征预测值。

针对问题三：本文根据所给数据使用 Spss 建立了治疗费用与分数之间的**一元线性回归**模型，再使用枚举算法求出 238 编号婴儿的所有可能得分组合，使用训练好的**集成学习模型**预测行为特征，使用四分位数法、设置最小费用区间等方法，使用**运筹优化算法**找到使治疗费用最小的降低得分方案：从矛盾型转变为中等型、从矛盾型转变为安静型最少费用分别为 15022.67 元、30001.67 元。

针对问题四：本文使用 **TOPSIS—熵权法**，通过重新分配各指标权重计算出了婴儿综合睡眠质量得分并进行排序，再通过**模糊数学综合评价法**对睡眠质量进行分级，构建了睡眠质量评级模型。最后使用与问题二类似的模型融合算法，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，成功预测最后 20 组婴儿的综合睡眠质量评级。

针对问题五：本文使用问题四建立的睡眠质量评级模型，对 238 编号婴儿的睡眠质量进行评级，发现为良。对于问题三的两种治疗方案，发现并不能使睡眠质量评级变为优。因此治疗策略调整方向为：在成功使得睡眠质量等级转变为优的情况下最小化治疗费用。接着使用问题四建立的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按照问题三的方式预测该婴儿在不同得分组合下的睡眠质量评级，使用运筹优化算法找到使治疗费用最小的降低得分方案：从良变为优最少费用为 50001.33 元。

关键词：典型相关分析、随机森林、多层感知机、K 最近邻、决策树、XGBoost、逻辑回归、集成学习、线性回归、运筹优化算法、TOPSIS—熵权法、模糊数学综合评价

一、 问题重述

1.1 问题背景

高尔基曾经说过：“世界上一切光荣和骄傲，都来自母亲”。母亲对于婴儿的成长发育起着至关重要的作用。除了提供身体上的营养和保护外，母亲还需要通过情感的输出给婴儿提供情绪上的稳定性和安全感。因此，母亲的心理健康状况如果出现问题，如出现抑郁、焦虑和压力等不良情绪时，可能会对婴儿的认知、情感和社会行为等方面产生诸多负面影响。当母亲承受过大的压力时，甚至可能导致婴儿生理和心理无法正常发展，例如影响到他们的睡眠质量。

1.2 问题提出

为了更详细地了解母亲的身心健康对于婴儿的影响，本文将采用数学建模解决下列问题：

问题一：

根据母亲的身体指标和心理指标的评估结果，寻找它和婴儿行为特征和睡眠质量之间的规律，并预测它们之间的相关性。

问题二：

假设将婴儿的情绪和反应等行为特征分为三种类型：安静型、中等型、矛盾型。我们需要根据第一问的相关性结果，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，并且预测数据表最后 20 组婴儿的行为特征类型。

问题三：

研究表明，对母亲焦虑的干预能够显著提高母亲的心理健康水平，还可以改善母婴交互质量，促进婴儿的认知、情感和社交发展。经调研可得 CBTS、EPDS、HADS 的治疗费用相对于患病程度的变化率与治疗费用呈正比。我们需要根据不同分数对应的治疗费用，预测编号 238 的行为特征为矛盾型的婴儿最少需要花费多少费用，能够使得婴儿的行为特征从矛盾型变为中等型，并且寻找能够使其行为特征变为安静型的最佳治疗方案。

问题四：

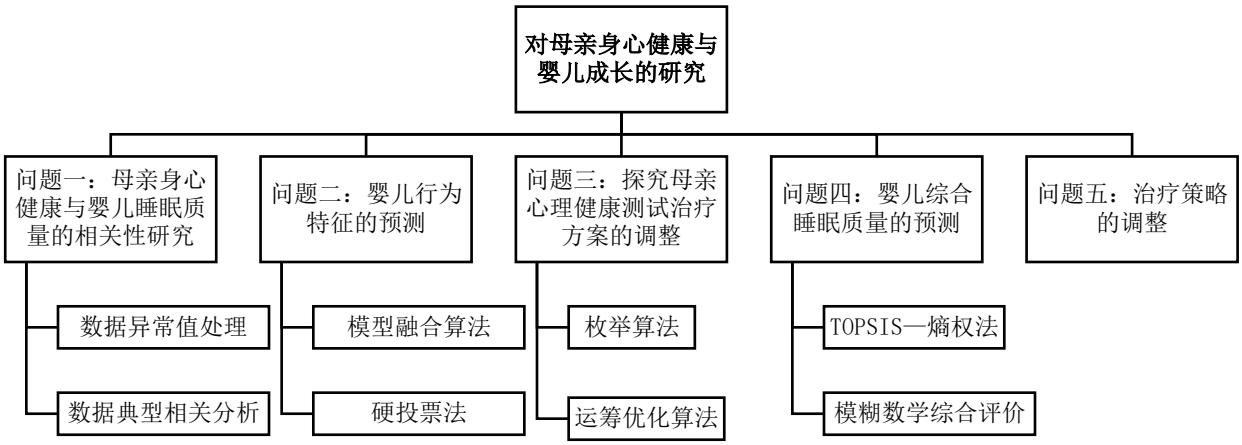
婴儿的睡眠质量指标包含整晚睡眠时间、睡醒次数、入睡方式。我们需要根据婴儿的睡眠质量指标对婴儿的睡眠质量进行优、良、中、差的评估，并且根据婴儿睡眠质量评估与母亲的身体指标、心理指标，建立关联模型，并预测最后 20 组婴儿的综合睡眠质量。

问题五：

在问题三的基础上，我们需要判断若让 238 号婴儿睡眠质量评级为优是否需要调整问题三的治疗策略，并给出调整方案、

二、 问题分析

文章框架图：



2.1 问题一的分析

问题一要求我们根据附件数据分析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是否有相关性并分析规律。可将母亲的身体指标和心理指标归为一组变量，称为“母亲指标”；婴儿的行为特征和睡眠质量归为一组变量，称为“婴儿指标”。假设两组典型变量服从正态分布，使用软件 Spss 进行分析，对两组典型变量的相关性进行检验，用这两组典型变量建立典型相关模型。

2.2 问题二的分析

问题二要求我们根据附件数据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，并以此预测数据表中最后 20 组的婴儿的行为特征。我们使用模型融合算法来建立关系模型。我们将采用随机森林、多层感知机（MLP）、K 最近邻（KNN）、决策树（Decision Tree）、XGBoost 和逻辑回归作为模型算法。这些算法在处理不同类型的数据和问题时都具有一定的优势。最后使用投票法作为集成学习方法，将这些模型的预测结果进行投票并融合，得到最终的预测值。通过模型融合，综合各个模型的优势，提高预测性能和准确度。

2.3 问题三的分析

问题三要求我们根据三个不同的心理测试 CBTS、EPDS、HADS 的治疗费用与患病程度变化率的正比关系来建立模型，分析编号为 238 的婴儿最少需要花费多少治疗费用能够使行为特征从矛盾型变为中等型；并进一步分析其行为特征变为安静型所需要花费的最少费用。我们利用提供的表格数据建立治疗费用与降低分数之间的一元线性回归模型，再使用第二问建立的模型融合分类模型预测该婴儿在不同 CBTS、EPDS 和 HADS 得分下的行为特征，使用运筹优化算法找到使治疗费用最小的降低得分方案。

2.4 问题四的分析

问题四要求我们根据附件数据对婴儿的睡眠质量进行分类评价，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并预测最后 20 组婴儿的综合睡眠质量。我们通过 TOPSIS—熵权法计算婴儿综合睡眠质量得分，通过模糊数学综合评价法对睡眠质量进行分级，构建睡眠质量评级模型。利用评级结果，使用与问题二类似的模型融合算法，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评级。

2.5 问题五的分析

问题五要求我们分析如何调整问题三的治疗策略使得编号为 238 的婴儿的睡眠质量评级变为优。问题四所建立的睡眠质量评级模型对该婴儿的睡眠质量评级为良。对于问题三的治疗方案，我们发现并不能使得睡眠质量评级变为优。因此我们的治疗策略调整方向为：在成功使得睡眠质量等级转变为优的情况下最小化治疗费用。我们使用问题四建立的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该婴儿在不同 CBTS、EPDS 和 HADS 得分下的睡眠质量评级，使用运筹优化算法找到使治疗费用最小的降低得分方案。

三、 模型假设

1) 假设在处理完异常值过后题目附件所给的研究调查数据真实有效；
2) 假设在不考虑其他因素下，题目所给出的母亲年龄、婚姻状况、教育程度、妊娠时间和分娩方式能够准确反映出产妇的身体指标，心理测试成绩 CBTS、EPDS 以及 HADS 能够准确反映出产妇的心理指标；
3) 假设在调查过程中所有参与人员都按真实情况填写；
4) 假设除本题设外的影响婴儿睡眠质量的因素都不对其产生影响，例如居住环境或家庭经济条件等。

四、 符号说明

符号	说明
u_k, v_k	第 k 对典型相关变量
$\rho_{X,Y}$	两个变量之间的总体的皮尔逊相关系数
ε_i	一元线性回归方程中的随机误差项
e_j	信息熵
ω_j	通过信息熵计算出的指标权重

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 典型相关分析模型准备

通过查阅文献[1]可知，典型相关分析用于研究两组变量（每组变量中都可能有多个指标）之间相关关系的一种多元统计方法。它采用类似于主成分分析的方法，在研究

两组变量间的相关关系时，首先在每组变量中找出变量的线性组合，代表了原始变量的大部分信息，并使得两组线性组合之间具有最大的相关系数，即相关程度最大。接着选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对，如此继续下去，直到两组变量之间的相关性被提取完毕为止。被选出的线性组合配对称为典型变量，它们的相关系数称为典型相关系数。典型相关系数度量了这两组变量之间联系的强度。

假设两组变量分别为：

$$X=(X_1, X_2, \dots, X_p), \quad Y=(Y_1, Y_2, \dots, Y_q)$$

分别在两组变量中选取具有代表性的综合变量 Ui、Vi，

将每一个综合变量看成是原变量的线性组合，即：

$$\begin{aligned} u &= a_1X_1 + a_2X_2 + \dots + a_pX_p \triangleq a'X \\ v &= b_1Y_1 + b_2Y_2 + \dots + b_qY_q \triangleq b'Y \end{aligned} \tag{1}$$

然后使用 Pearson 相关系数度量变量 u 与 v 的关系：

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{2}$$

我们期望得到一组最优的解 a 和 b，*使得* $\operatorname{corr}(u, v)$最大，这样所得到的 a、b 就是使得 u、v 具有最大关联的权重。

5.1.2 数据预处理

在对数据进行预处理之前，我们需要对调查数据进行清洗和筛选，剔除异常值或缺失值，防止其对最终结果产生误差。我们发现在婚姻状况一栏除了题设给出的 1（未婚）和 2（已婚）之外，还出现了若干个数字 3 和 6，这显然是异常值，在对数据进行处理之前，应该先清除掉这 6 条异常数据

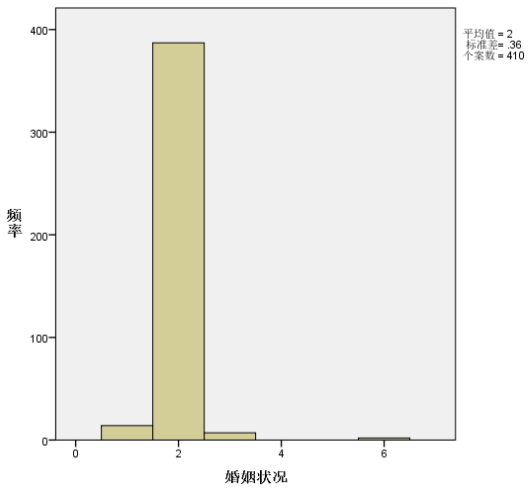


图 1 母亲婚姻状况

由题目所给的信息，母亲的身体指标与年龄、婚姻状况、教育程度、妊娠时间、分娩方式有关，心理指标与 CBTS、EPDS、HADS 有关，我们把与母亲相关的这八个指标归为一组变量，称为“母亲指标”；婴儿的行为特征分为：安静型、中等型、矛盾型，我们采取标签编码操作（Label Encoding）将婴儿行为特征字符串转换为数值型：“中等型”编码为 0、“安静型”编码为 1、“矛盾型”编码为 2。婴儿的睡眠质量与整

晚睡眠时间、睡醒次数和入睡方式有关，其中我们将整晚睡眠时间格式（时：分：秒）处理为小时制（例：10:30 改为 10.5）。我们把与婴儿相关的这四个指标归为一组变量，称为“婴儿指标”。用这两组指标作为典型相关变量，定义如表 1 所列：

表 1 典型相关变量组

典型相关变量	各变量指标
母亲指标	身体指标（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）
婴儿指标	行为特征睡眠质量（整晚睡眠时间、睡醒次数、入睡方式）

5.1.3 正态分布检验

典型相关性分析要求样本数据满足正态分布的要求，因此我们需要首先对样本数据的正态分布性进行检验。此处我们以参与实验的母亲年龄为例。

从图可以看出，样本数据与正态分布曲线拟合效果较好，于是我们用 SPSS 软件对样本数据进行了正态分布检验，发现 Shapiro-Wilk 检验的 p 值大于 0.05 水平，因此接受 H_0 假设，认为此样本符合正态分布的要求。

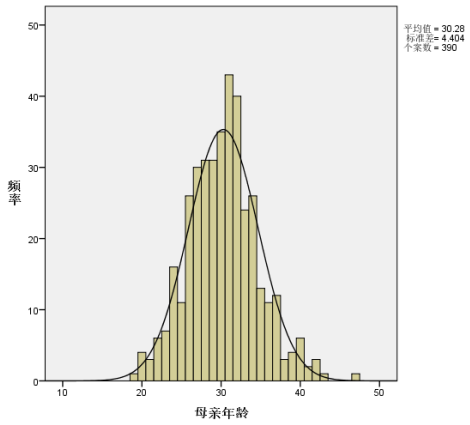


图 2：正态分布拟合

5.1.4 典型相关系数及其检验

我们通过 Spss 软件对母亲指标 u 和婴儿 v 进行典型相关分析，得出典型相关系数表，如表 2 所列：

表 2 典型相关系数

	相关系数	特征值	威尔克统计	F	分子自由度	分母自由度	p 值
1	.292	.093	.834	2.201	32.000	1391.902	.000
2	.226	.054	.911	1.698	21.000	1085.962	.026
3	.185	.035	.960	1.290	12.000	758.000	.219
4	.075	.006	.994	.430	5.000	380.000	.828

对得到的典型相关系数进行假设检验：原假设（H_0）是两组变量之间不存在典型相关关系，备择假设（H_1）是两组变量之间存在典型相关关系。置信水平有三个：

90%、95%、99%，其对应的显著性水平分别为 0.1、0.05、0.01。在 0.01 的显著性水平下，只有第一个典型相关系数的 p 值小于 0.01，即在 99%的置信水平下拒绝原假设，认为两组变量存在典型相关关系，表明能够用母亲指标变量组 u 来解释婴儿指标变量组 v。由于后三个典型相关系数未通过显著性检验，以下模型基于第一个典型相关系数建立。

5.1.5 建立典型相关模型及结论

由于数据的计量单位不一致，不宜进行直接比较，因此我们通过标准化典型相关系数来建立典型相关模型。通过 Spss 软件得出标准化典型相关变量对应的线性组合系数，如表 3、表 4 所列。由于只取第一个典型相关系数，其他数据可剔除。

表 3 母亲指标的标准化典型相关系数

变量	1	2	3	4
母亲年龄	-.304	.405	-.616	-.387
婚姻状况	.144	.603	.275	.301
教育程度	.286	-.095	-.222	.777
妊娠时间（周数）	-.063	-.410	-.536	.160
分娩方式	-.034	-.327	.237	-.173
CBTS	-.124	-.140	.161	.147
EPDS	-.205	.256	-.128	-.025
HADS	-.101	-.701	-.040	.106

表 4 婴儿指标的标准化典型相关系数

变量	1	2	3	4
婴儿行为特征	-.472	.091	-.489	-.777
整晚睡眠时间（时：分：秒）	.741	.226	.311	-.663
睡醒次数	-.256	.358	1.006	-.074
入睡方式	-.208	.959	-.116	.387

由表 3、表 4 可得标准化的第一对典型变量，即典型相关模型：

$$u = -0.304X_1+0.144X_2+0.286X_3-0.063X_4-0.034X_5-0.124X_6-0.205X_7-0.101X_8$$
$$v = -0.472Y_1+0.741Y_2-0.256Y_3-0.208Y_4 \quad (3)$$

结论：

在母亲的身体指标中，X_1（母亲年龄）的系数绝对值最大且为负号，反映身体指标的典型变量主要由母亲年龄决定，且为负相关，而妊娠时间和分娩方式虽然也为负相关关系，但是由于系数绝对值过小，因此对母亲指标的影响可以忽略不计；X_2（婚姻状况）和 X_3（教育程度）的系数为正，说明母亲的婚姻状况良好、教育水平优秀的条件下，对母亲的整体状况呈良好趋势。身体指标的系数与现实经验一致。在母亲的心理指标中，三种心理测试（X_6、X_7、X_8）的系数都为负，说明心理测试的分数越高，对母亲的心理健康负面形象越大，这与心理健康测试结果一致。

婴儿行为特征 Y_1、睡醒次数 Y_3、入睡方式 Y_4 的系数为负，说明当婴儿的入睡状态

越矛盾、睡醒次数越多、入睡方式越复杂，婴儿的睡眠质量越差；而整晚睡眠时间 Y2 的系数为正说明睡眠质量与整晚睡眠时间呈正相关。这都与现实中的经验一致。

即结论为：母亲年龄越小、婚姻状况越好、受教育程度越高，母亲的身体指标越大；三个心理测试的分数越低，母亲的心理指标越大，从而使得母亲指标更大，母亲的身心更加健康，对婴儿更能产生良性影响。婴儿行为特征数值越小（越趋于安静型）、睡醒次数越少、入睡方式数值越小（入睡趋于简单）、整晚睡眠时间越长，婴儿的睡眠质量越好。这和客观事实是相符的。

5.2 问题二模型的建立与求解

5.2.1 模型融合预测的准备

问题二要求我们预测表中后 20 组婴儿的行为特征。因此我们将附录中前 390 条信息作为训练集，选择随机森林（RF）、多层感知机（MLP）、K 最近邻（KNN）、决策树（Decision Tree）、XGBoost 和逻辑回归来建立婴儿的行为特征与母亲的身体指标和心理指标关系模型。

1. 随机森林（Random Forest）是一种集成学习算法，通过组合多个决策树来进行预测。它能够处理高维数据和非线性关系，并能够进行特征选择和处理缺失值。

2. 多层感知机（MLP）是一种人工神经网络模型，具有多个隐含层的结构。它适用于处理复杂的非线性关系，并能够学习到特征之间的复杂交互。

3. K 最近邻（KNN）是一种基于实例的学习算法，通过计算最近邻样本的类别来进行分类。它可以适应不同类型的数据，并能够处理非线性关系。

4. 决策树（Decision Tree）是一种基于树形结构的有监督学习算法，用于分类和回归问题。它通过对特征进行逐步的选择和划分，构建出一个树形结构来进行预测。决策树的优势在于生成的模型具有可解释性，也能够处理非线性关系和高维数据。

5. XGBoost（eXtreme Gradient Boosting）是一种集成学习算法，是一种梯度提升框架。XGBoost 通过训练多个弱学习器（通常是决策树），将它们组合成一个强学习器来进行预测和分类。它在处理高维数据和非线性关系方面表现出色，并具有出色的泛化能力。

6. 逻辑回归（Logistic Regression）是一种常用的分类算法，适用于处理二分类问题。它可以用于分析特征与目标变量之间的线性关系，并可以得到关于特征权重的解释。

在得出六种预测模型后，使用模型融合（Ensemble）算法进行融合。模型融合算法是一种将多个独立模型的预测结果结合起来以提高整体性能和准确性的方法，可以应用于各种机器学习任务，包括分类、回归和聚类等。对于问题二，选择硬投票法（Hard Voting Classifier）：根据少数服从多数的投票原则来确定最终的预测结果。

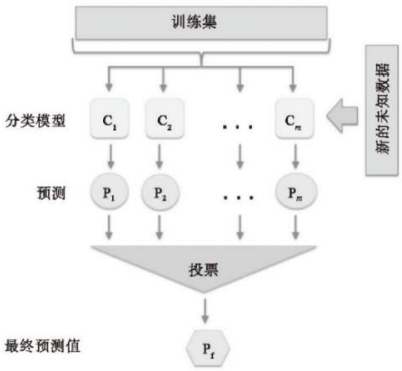


图 3：模型融合·硬投票法详解

5.2.2 数据预处理

我们使用了题目附录的数据作为数据集，并将在后续使用 k 折交叉验证（k-fold cross-validation）把数据集划分为训练集和测试集。训练集用于模型的训练和参数优化，测试集则用于评估模型在未见过的数据上的泛化能力。这样的数据集划分方式可以有效地评估模型的准确性、鲁棒性和泛化能力，并为进一步的模型优化和改进提供指导。

为了方便机器学习任务，我们将数据集的原始中文文本转换为英文，确保在使用基于 Python 的 scikit-learn 库进行模型训练时有更好的兼容性。

由于深度学习模型要求输入为数值型数据，而婴儿行为特征是以字符串形式表示的，因此我们采取标签编码操作（Label Encoding）将婴儿行为特征字符串转换为数值型：“中等型”编码为 0、“安静型”编码为 1、“矛盾型”编码为 2。

此为数据预处理后的部分数据：

表 5 数据预处理后的部分数据

number	mother_age	marriage	education	pregnant	birth_method
1	34	2	5	37	1
2	33	2	5	42	1
3	37	2	5	41	1
4	31	2	5	37.5	1
5	36	1	5	40	1
6	32	2	5	41	1
7	28	2	4	41	1

CBTS	EPDS	HADS	action_mode	mode
3	13	9	中等型	0
0	0	3	安静型	1
4	8	9	安静型	1
6	16	13	安静型	1
1	3	3	中等型	0
1	2	3	安静型	1
17	25	16	矛盾型	2

5.2.3 k 折交叉验证方法

预处理得到数据集后，我们使用基于 Python 的 scikit-learn 库进行模型训练，采用了几种常见的机器学习算法：随机森林（RF）、多层感知机（MLP）、K 最近邻（KNN）、决策树（Decision Tree）、XGBoost 和逻辑回归。不同的算法模型可以提供更多的多样性，每种模型都有其自身的优势和假设。通过比较不同模型的性能，我们可以更好地了解每种模型在给定数据集上的表现，得到更准确的集成预测。

我们采用 k 折交叉验证方法来有效地利用数据集：

首先，将原始数据集划分为 k 个相等大小的子集。然后，将数据集分割为 k 个互斥的子集，并依次选择其中一个子集作为验证集，而将其余 k-1 个子集作为训练集进行模型训练。重复进行 k 次训练-验证迭代，直到每个子集都充当了一次验证集。这样，我们获得了 k 个准确率值，每个准确率值代表了模型在相应验证集上的性能。

最后，我们计算这 k 个准确率值的平均值，并绘制准确率图像。图像的横坐标表示验证集迭代的次数，纵坐标表示对应验证集上的准确率。这个准确率图像可视化了模型在整个数据集上的性能，并帮助我们评估模型的泛化能力和稳定性。

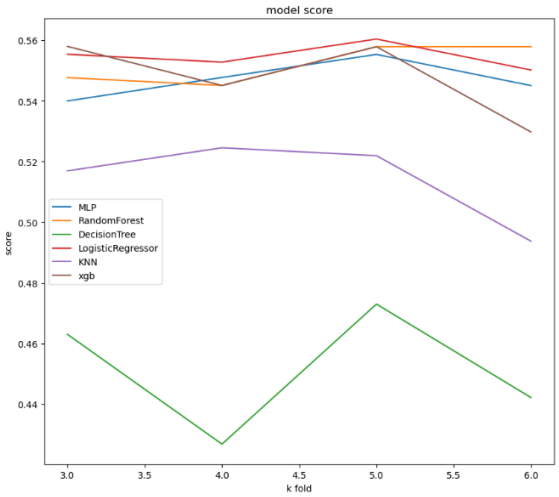


图 4 六种算法在不同训练集下的准确率折线图

每个模型都会独立地生成相应的预测结果，因此我们得到了六个模型通过训练集所预测的 20 组婴儿的行为特征的部分表格数据。（完整数据见附录）

表 6 六个模型的预测数据（部分）

	MLP	RF	DT	Logistic	KNN	XGB
391	中等型	中等型	中等型	中等型	中等型	中等型
392	中等型	中等型	矛盾型	中等型	中等型	中等型
⋮						⋮
409	中等型	中等型	中等型	中等型	中等型	安静型
410	中等型	中等型	中等型	中等型	中等型	中等型

5.2.4 投票分类器模型

为了提高预测性能和模型稳定性，并得出一个最终的预测值，我们采用了模型融合的技术。模型融合是一种将多个独立模型的预测结果进行综合的方法，通过整合不同模型的意见和决策，可以得到更加准确和可靠的最终预测结果。

根据问题二的分类算法性质，我们采用硬投票法（hard voting）。硬投票法是一种简单而有效的模型融合方法，它基于多数表决的原则。对于每个样本，五种模型会给出各自的预测结果，投票分类器通过计算预测结果的众数（即出现次数最多的预测类别）来确定最终的预测值。下图为投票分类器在不同训练集下的准确率折线图：

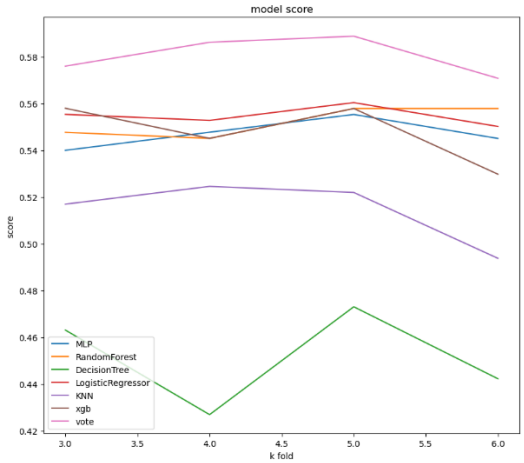


图 5 投票分类器与其它模型的正确率比较（折线图）

根据折线图，我们发现投票分类器得到的准确率远高于任何一个模型，可见模型融合算法能够将多个模型组合成一个更准确稳定的模型，使得预测准确率最高 。

最终，我们根据模型融合算法，得到最后 20 组（编号 391-410 号）婴儿的行为特征信息的预测值如下表：

表 7 婴儿行为特征最终预测表

编号	预测值	编号	预测值	编号	预测值	编号	预测值
391	中等型	396	中等型	401	中等型	406	安静型
392	中等型	397	安静型	402	中等型	407	中等型
393	中等型	398	中等型	403	中等型	408	中等型
394	中等型	399	中等型	404	安静型	409	安静型
395	中等型	400	中等型	405	中等型	410	中等型

5.3 问题三模型的建立与求解

5.3.1 一元线性回归模型准备

通过查阅文献[2]可知：一元线性回归分析主要研究两个变量之间的线性关系，回归模型为 $Y=\beta_0+\beta_1+\varepsilon$ ，其中 β_0，β_1 为待定系数。实际问题中，我们通过观测 n 组数据（X_i，Y_i）（$i=1,2,\cdots,n$），它们满足模型 $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$ （$i=1,2,\cdots,n$）并且通常假定 $E(\varepsilon_i)=0$，$Var(\varepsilon_i)=\sigma^2$ 各 ε_i 相互独立且服从正态分布。回归分析就是根据样本的

观察值寻找 β_0,β_1 的估计值$\hat{\beta}_0$, $\hat{\beta}_1$ ，对于给定的 x 值，取$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$，作为$E(Y) = \beta_0 + \beta_1 x$的估计，利用最小二乘法得到 β_0,β_1 的估计$\hat{\beta}_0$, $\hat{\beta}_1$，其中：

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 = \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) / \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \end{cases} \tag{4}$$

又因为在问题三中，根据题设已知“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”，因此在建立线性回归模型时无需额外进行线性假设检验。

5. 3. 2 建立题设一元线性回归模型

依题意，降低患病得分与所需要的治疗费用的关系如图所示：

表 8 患病得分与治疗费用

CBTS		EPDS		HADS	
得分	治疗费用（元）	得分	治疗费用（元）	得分	治疗费用（元）
0	200	0	500	0	300
3	2812	2	1890	5	12500

将数据输入 Spss 中进行运算，选择线性回归分析。可得回归模型系数表如下：

表 9 回归模型系数[a]（*a*.因变量：治疗费用（元））

模型		未标准化系数		标准化系数	t	显著性
		B	标准错误	Beta		
CBTS	(常量)	200	0		.	.
	得分	870.667	0	1	.	.
EPDS	(常量)	500	0		.	.
	得分	695	0	1	.	.
HADS	(常量)	300	0		.	.
	得分	2440	0	1	.	.

本题中三种模型线性回归分析后的相关系数 R 均为一，说明患病得分与治疗费用之间存在完全正向的线性关系，这与题目中“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”一致。并且根据系数表，可以分别得到三个模型 CBTS（Y_1）、EPDS（Y_2）、HADS（Y_3）的线性回归方程和线性折线图：

$$Y_1 = 810.667x_1+200$$
$$Y_2 = 695x_2+500 \quad (5)$$
$$Y_3 = 2440x_3+300$$

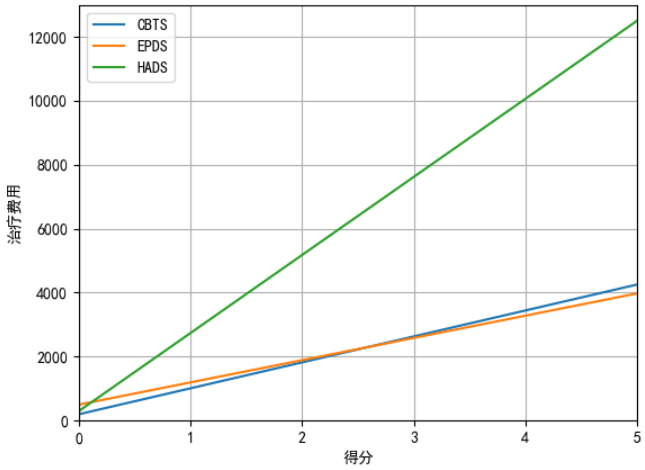


图 6 治疗费用和得分的线性回归图

5. 3. 3 最小治疗费用求解

对于编号为 238、行为特征为矛盾型的婴儿，他的母亲的 CBTS、EPDS、HADS 测试得分分别为 15、22、18。根据三条治疗费用与降低分数的线性模型，可计算出对于他的最大治疗费用为：15×810.667+200+695×22+500+2440×18+300=72370.005 元。这将作为治疗费用的上限参考值。

i. 从矛盾型变为中等型的最小治疗费用

为了最小化治疗费用，我们使用枚举算法：在 Python 中使用三个嵌套的 for 循环，对所有可能的 CBTS、EPDS 和 HADS 得分组合进行枚举。针对每种得分组合，进行以下步骤：

① 根据得分组合预测行为特征：使用第二问中训练好的融合模型，预测婴儿在不同得分组合下的行为特征，判断行为特征是否改变为中等型。

② 剔除未改变的行为特征：将行为特征未发生改变的结果剔除，只保留改变为中等型的结果。

③ 计算治疗费用：使用治疗费用与分数的线性模型计算从原先的得分组合降低到当前得分组合所需的治疗费用。

④ 在最大治疗费用的限制下，枚举循环完所有可能的得分组合后，我们得到了所有行为特征转变为中等型的治疗费用结果，进行升序排序：

表 10 不同得分组合下成功转变为中等型的治疗费用表

CBTS	EPDS	HADS	mode	price
14	21	18	0	2505.667
14	20	18	0	3200.667

13	21	18	0	3316.334
12	22	18	0	3432.001
14	19	18	0	3895.667
⋮				⋮

由表 10 可知，最小治疗费用约为 2505 元。

为了更好理解治疗费用的分布，我们计算了不同治疗费用区间的频数、累积频数与累计频率：

表 11 不同治疗费用区间的频数、累积频数与累计频率表

治疗费用	频数	累积频数	累积频率	治疗费用	频数	累积频数	累积频率
2000	1	1	0.1%	10000	181	277	24.4%
3000	4	5	0.4%	15000	213	490	43.2%
4000	7	12	1.1%	20000	221	711	62.8%
5000	9	21	1.9%	25000	208	919	81.1%
6000	10	31	2.7%	30000	150	1069	94.4%
7000	17	48	4.2%	35000	58	1127	99.5%
8000	25	73	6.4%	40000	6	1133	100.0%
9000	23	96	8.5%				

我们将数据输入到 Spss 中，更加直观地得出频率分布直方图：

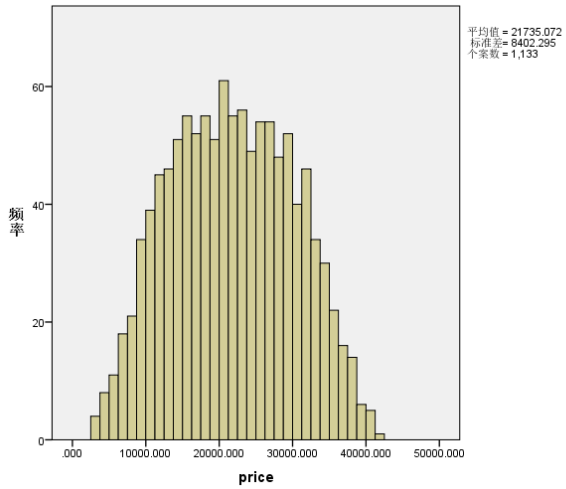


图 7 不同治疗费用区间的频率分布直方图

根据表 11、图 7，可知：随着费用增加，频数呈现逐渐递增的趋势。即费用较低的频数相对较少，而费用较高的频数逐渐增多，表明费用分布呈现右偏（正偏）的特征。根据频数图的分析，低于 2505 元的费用值的频数相对较少，而高于 2505 的费用值的频数逐渐增加。这意味着将 2505 元作为最低费用可能会导致遗漏较大比例的样本观测值，无法准确捕捉到费用分布的典型特征。

虽然机器学习模型可以提供费用预测，然而，由于模型的复杂性和对样本数据的拟合程度，模型预测可能存在一定误差和特例。根据频数图的分析，我们意识到直接选择融合模型预测的最低费用（2505 元）作为最低费用可能存在潜在风险，不具有代

表性。因此，我们决定选择累积频率作为阈值来确定最小费用区间，降低模型预测误差和特例的影响，以提高结果的准确性和可靠性。

我们使用四分位数法，将累积频率分为相等的四分值，其中 25%累积频率对应的价格即为 25%分位数，也被称为第一四分位数（Q_1）或下四分位数（Lower Quartile）。根据融合模型的特点与最小化治疗费用的需求，我们选择 25%的累积频率作为最小费用区间的阈值。

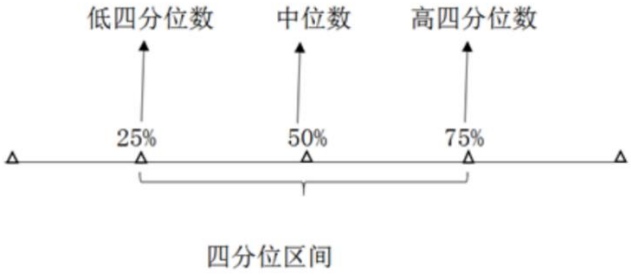


图 8 四分位数法示意图

由表 11 可知，24.4%的累积频率对应的价格为 10000 元。为了确定最小费用区间，我们需要找到累积频率大于 25%的最小费用作为起始点。当累积频率大于 25%时，最小费用区间的起始点价格为 15000。而最小费用是指最小费用区间内具有最低价格的价格值，根据表 12，可知最小费用为 15022.67 元。

表 12 不同得分组合下成功转变为中等型的治疗费用表（部分，见表 10）

CBTS	EPDS	HADS	mode	price
12	16	15	0	14922
11	10	17	0	15022.67
8	17	16	0	15029.67

最终，我们得出结论：基于累积频率阈值为 25%，最小费用区间的起始点为 15000，对于编号为 238、行为特征为矛盾型的婴儿，最少需要花费 15022.67 元的治疗费用，使得 CBTS、EPDS、HADS 分别降为 11、10、17，能够使婴儿的行为特征变为中等型。

ii. 从矛盾型变为安静型的最小治疗费用

该问题与从矛盾型变为中等型的治疗费用最小化优化过程相同：枚举可能的得分组合，预测行为特征，剔除不符的结果，记录治疗费用与频数、累积频率：

表 13 不同得分组合下成功转变为安静型的治疗费用表

CBTS	EPDS	HADS	mode	price
14	21	18	1	1695
13	22	18	1	1810.667
14	20	18	1	2390
13	21	18	1	2505.667
12	22	18	1	2621.334
⋮				⋮

与中等型相同，我们选择 25%的累积频率作为最小费用区间的阈值。由表 13 可知，23.5%的累积频率对应的价格为 25000 元。为了确定最小费用区间，需要找到累积频率大于 25%的最小费用作为起始点。当累积频率大于 25%时，最小费用区间的起始点价格为 30000 元。而最小费用是指最小费用区间内具有最低价格的价格值，根据表 14，可知最小费用是 30001.67 元。

表 14 不同得分组合下成功转变为安静型的治疗费用表（部分，见表 13）

CBTS	EPDS	HADS	mode	price
7	6	13	1	29994.67
4	13	12	1	30001.67
10	6	12	1	30002.67

最终，我们得出结论：基于累积频率阈值为 25%，最小费用区间的起始点为 30000 元，对于编号为 238、行为特征为矛盾型的婴儿，需要修改治疗方案，继续加强治疗，最少需要花费 30001.67 元的治疗费用，使得 CBTS、EPDS、HADS 分别降为 4、13、12，才能使婴儿的行为特征变为安静型。

5.4 问题四模型的建立与求解

5.4.1 TOPSIS—熵权法模型准备

（一）TOPSIS—熵权法：

根据文献[3]，TOPSIS 是一种基于距离度量的综合评价办法，适用于有多个指标时的方案选择问题。它将每个指标的权重视为一个决策变量，并通过线性优化模型求解最优权重，TOPSIS 可以同时考虑指标的贡献度和互补性，更全面地评估方案的优劣。

熵权法是一种基于信息熵的权重确定方法，通过计算各个指标在不同方案中的信息熵，根据信息熵的大小分配权重，以确定各个指标的重要性。

熵权法和 TOPSIS 可以结合使用来进行多属性决策：

先使用熵权法：根据信息熵的定义计算熵值，确定各个指标的权重；再使用 TOPSIS：对指标进行归一化处理，与对应的权重相乘，得到加权后的矩阵，计算加权后矩阵的最优解和最劣解，计算每个评价对象与最优解和最劣解的距离，最后根据 TOPSIS 方法的公式---最劣解与评价对象的距离除以最劣解与最优解的距离之和，计算每个评价对象的得分。这种结合方法既考虑了指标的重要性和权重，又考虑了方案与理想解的接近程度，以提供更全面的决策分析。

（二）模糊综合评价法：

模糊综合评价法（Fuzzy Comprehensive Evaluation Method）是一种基于模糊数学理论的评价方法，考虑了多因素的影响，用于处理模糊、不确定或多指标的决策问题。该方法将模糊集合理论与数学模型相结合，通过量化和综合各种评价指标的模糊信息，得出最终的评价结果。

5.4.2 熵权法——基于信息熵对权重的重新分配

熵权法是基于信息熵对权重的重新分配。信息熵是偏于客观的确定权重的方法，它借用信息论中熵的概念，适用于多属性决策和评价。

Step1. 构造矩阵

以附录中 390 名婴儿的整晚睡眠时间、睡醒次数、入睡方式作为信息构建矩阵

X，每行代表一个婴儿对象，每列代表一个指标：（第 i 名婴儿的睡醒次数位为 a_i、第 i 名婴儿的入睡方式为 b_i、第 i 名婴儿的睡眠时间为 c_i）

$$X = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ \cdots & \cdots & \cdots \\ a_n & b_n & c_n \end{bmatrix} \tag{6}$$

Step2. 数据处理

为消除因量纲不同对评价结果的影响，需要对各指标进行正向化与标准化处理。

• 正向化处理：

矩阵第一列为睡醒次数，第二列为入睡方式，均为极小型指标，即期望值标值越小越好，因此需要正向化处理：

$$\mathrm{x}' = \mathrm{Max}\,(\mathrm{x}) - \mathrm{x} \tag{7}$$

矩阵第三列为睡眠时间，为极大型指标，即期望指标值越大越好，不做处理。

• 规范化处理：

每一列元素都除以当前列向量的范数（使用余弦距离度量）：

$$x_i = \frac{x_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}} \tag{8}$$

正向化与规范化后的矩阵（部分）：

$$X = \begin{bmatrix} 0.04086377 & 0.06267707 & 0.04941689 \\ 0.05837681 & 0.02089236 & 0.05435857 \\ 0.05253913 & 0.06267707 & 0.05930026 \\ \cdots & \cdots & \cdots \\ 0.04670145 & 0.08356943 & 0.05188773 \\ 0.04670145 & 0.08356943 & 0.02965013 \\ 0.05837681 & 0.02089236 & 0.04447522 \end{bmatrix} \tag{9}$$

Step3. 计算指标信息熵

对于第 j 个指标而言，其信息熵的计算公式为：

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij}\ln\left(p_{ij}\right)(j = 1,2,\cdots,m) \tag{10}$$

通过公式可得三个指标的信息熵为：

表 15 各指标的信息熵

信息熵	睡醒次数	入睡方式	睡眠时间
e	5.94163338	5.67154165	5.95277523

Step4. 计算指标权重

通过公式：

$$\omega_j = \frac{1 - E_j}{\sum_{k=1}^{10} (1 - E_k)} \tag{11}$$

计算出我们确定的各个指标新的权重为：

表 16 各指标新的权重

指标权重	睡醒次数	入睡方式	睡眠时间
ω	0.33925925	0.32071657	0.34002418

5.4.3 TOPSIS——计算综合得分

TOPSIS 是一种用于多指标方案选择问题的综合评价方法。它基于距离度量的原理，将每个指标的权重视为一个决策变量，并通过线性优化模型来确定最优权重。

Step1.计算加权矩阵

将指标值与对应的权重相乘，构造加权矩阵 X^*：

$$X^* = \begin{bmatrix} 0.01386341 & 0.02010158 & 0.01680294 \\ 0.01980487 & 0.00670053 & 0.01848323 \\ 0.01782438 & 0.02010158 & 0.02016352 \\ \cdots & \cdots & \cdots \\ 0.0158439 & 0.0268021 & 0.01764308 \\ 0.0158439 & 0.0268021 & 0.01008176 \\ 0.0198048 & 0.0067005 & 0.01512264 \end{bmatrix} \tag{12}$$

Step2. 计算最优解和最劣解的距离

计算最优解与最劣解：

$$\begin{cases} \mathrm{z}_{\mathrm{ij}}^{*+} = \underset{\mathrm{n,p}}{max}\left(\mathrm{z}_1^{*+}, \mathrm{z}_2^{*+}, \cdots, \mathrm{z}_{\mathrm{p}}^{*+}\right) \\ \mathrm{z}_{\mathrm{ij}}^{*-} = \underset{\mathrm{n,p}}{min}\left(\mathrm{z}_1^{*-}, \mathrm{z}_2^{*-}, \cdots, \mathrm{z}_{\mathrm{p}}^{*-}\right) \end{cases} \tag{13}$$

表 17 各指标的最优解和最劣解

	睡醒次数	入睡方式	睡眠时间
最优解z^{*+}	0.01980487	0.0268021	0.02016352
最劣解z^{*-}	0	0	0.00840147

计算最优距离和最劣距离：

定义第 i 个（i=1,2,⋯,n）评价对象与最大值的距离：

$$D_i^+ = \sqrt{\sum_{j=1}^{m} W_j * \left(Z_j^+ - z_{ij}\right)^2} \tag{14}$$

定义第 i 个（i=1,2,⋯,n）评价对象与最小值的距离：

$$D_i^- = \sqrt{\sum_{j=1}^{m} W_j * \left(Z_j^- - z_{ij}\right)^2} \tag{15}$$

部分数据如下：

表 18 各对象的最优距离和最劣距离

	最优距离D^+	最劣距离D^-
婴儿对象 1	0.00956512	0.02582348
婴儿对象 2	0.02017168	0.02321146
婴儿对象 3	0.00698709	0.02932794
婴儿对象 4	0.00430264	0.03272652
婴儿对象 5	0.02035555	0.02116632

Step3.计算综合得分

根据 TOPSIS 方法的公式——最劣解与评价对象的距离除以最劣解与最优解的距离之和：

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{16}$$

计算每个评价对象的综合得分为：

表 19 评价对象的综合得分表

Number	Score	Number	Score
1	0.486124	6	0.669467
2	0.725397	7	0.535034
3	0.87087	8	0.712461
4	0.515278	9	0.457533
5	0.771514	10	0.466763

5. 4. 4 模糊数学综合评价——构建评级模型

通过 TOPSIS—熵权法，我们计算得到了婴儿睡眠儿综合睡眠质量得分。为了对婴儿的睡眠质量进行优、良、中、差四分类综合评判，我们对得分进行排序，利用模糊数学知识进行睡眠质量等级的划分：

表 20 得分排名模糊划分标准

模糊等级标准	得分排名
优（最高等级）	1-80
良	81-196
中	197-312
差（最低等级）	312-389

得到最终的排名和婴儿睡眠质量评价结果：

number	wake_times	sleep_way	sleep_time	Score	排名	等级
326	1	4	12	1	1	优
117	0	4	12	1	2	优
⋮						⋮
234	4	1	7	0.792745	79	优
179	1	4	12	0.789442	81	良
43	0	4	10.5	0.789442	82	良
⋮						⋮
71	0	4	11	0.544085	196	良
350	1	3	8	0.544085	197	中
207	1	1	12	0.544085	198	中
⋮						⋮
103	0	2	8	0.492372	312	中
1	3	2	10	0.486124	313	差
389	0	4	9	0.486124	314	差
⋮						⋮
133	1	4	7	0.200532	389	差

5.4.5 模型融合：建立关联模型并预测

在得到评级结果后，为了便于深度学习建模，我们采取标签编码操作（Label Encoding）将睡眠质量评级字符串转换为数值型：“优”编码为 0、“良”编码为 1“中”编码为 2，“差”编码为 3。

与问题二一致，在通过 TOPSIS—熵权法求得综合睡眠质量评分之后，我们需要将母亲的身体指标、心里指标和婴儿的综合睡眠质量建立模型。这与问题二“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“本质类似。因此，我们将采用问题二所使用的模型融合算法，使用随机森林、多层感知机、K 最近邻、决策树、XGBoost 和逻辑回归作为预测模型，最后使用投票法来融合得到最终的预测值。

最终，我们得到编号 391-410 号的婴儿的综合睡眠质量的预测值，如下表：

表 21 婴儿综合睡眠质量预测表

number	睡眠等级	number	睡眠等级	number	睡眠等级
391	中	398	中	405	良
392	良	399	差	406	优
393	良	400	良	407	优
394	优	401	中	408	优
395	良	402	中	409	中
396	良	403	中	410	中
397	优	404	良		

5.5 问题五模型的建立与求解

5.5.1 睡眠质量调整治疗方案

在我们建立的睡眠质量评级模型下，编号为 238 的婴儿的睡眠质量评级为良：

表 22 编号为 238 的婴儿的睡眠质量指标

编号	wake_times	sleep_way	sleep_time	score	睡眠等级
238	1	4	12	0.789442	良

在问题四中，我们训练得到了一个婴儿综合睡眠质量与母亲身体指标、心理指标的关联模型。

在问题三的两种治疗方案下，目标心理指标分别为 11、10、17 与 4、13、12。这两种目标结果都不能使得婴儿的睡眠等级从良转为优。因此，我们需要调整治疗方案，调整方向为：需要进一步加强治疗，使得心理指标继续降低；在成功使得睡眠质量等级转变为优的情况下最小化治疗费用。这与第三问的思路一致。

我们按照问题三的思路：枚举可能的得分组合，预测睡眠质量评级，剔除不符的结果，记录治疗费用与频数、累积频率，得到所有睡眠质量等级转变为优的治疗费用结果：

表 23 不同得分组合下成功转变为评级优的治疗费用表

CBTS	EPDS	HADS	price	睡眠评级
14	21	18	2505.667	0
14	20	18	3200.667	0
13	21	18	3316.334	0
14	19	18	3895.667	0
13	20	18	4011.334	0
⋮				⋮

表 24 不同治疗费用区间的频数、累积频数与累计频率表

治疗费用	频数	累积频数	累积频率	治疗费用	频数	累积频数	累积频率
2000	0	0	0.0%	15000	30	68	6.2%
3000	1	1	0.1%	20000	26	94	8.6%
4000	3	4	0.4%	25000	40	134	12.2%
5000	5	9	0.8%	30000	43	177	16.1%
6000	4	13	1.2%	35000	38	215	19.6%
7000	3	16	1.5%	40000	55	270	24.6%
9000	15	31	2.8%	50000	392	662	60.3%
10000	7	38	3.5%	70000	435	1097	100.0%

与问题三相近，为了降低模型预测误差和特例的影响，我们使用四分位数法，选择累积频率 25%作为阈值来确定最小费用区间，以提高结果的准确性和可靠性。由表 24 可知，24.6% 的累积频率对应的价格为 40000 元。为了确定最小费用区间，需要找到累积频率大于 25%的最小费用作为起始点。当累积频率大于 25%时，最小费用区间的起始点价格为 50000 元。而最小费用是指最小费用区间内具有最低价格的价格值，根据表 25，最小费用是 50001.33 元。

表 25 不同得分组合下成功转变为评级优的治疗费用表（部分，见表 23）

CBTS	EPDS	HADS	price	睡眠评级
10	10	3	49993.34	0
13	10	2	50001.33	0
13	17	0	50016.33	0

最终，我们得出结论：

若需要让 238 号婴儿的睡眠质量评级为优，问题三的两种治疗策略都无效，因此需要调整治疗方案：在成功使得睡眠质量等级转变为优的情况下，基于累积频率阈值为 25%，最小费用区间的起始点为 50000 元，对于编号为 238、睡眠评级为良的婴儿，最少需要花费 50001.33 元的治疗费用，使得 CBTS、EPDS、HADS 分别降为 13、10、2，才能使婴儿的睡眠质量评级变为优。

六、 模型的评价、改进与推广

6.1 模型的优点

（一）典型相关分析模型能够评估两组变量之间的相关性，可以直接计算多个变量之间的相关性从而反映两组指标之间整体的关联性，而不用逐一考虑各个变量之间的关系。这帮助了我们快速了解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的潜在联系。

（二）模型融合算法可以充分利用多个算法的优势，并综合它们的预测结果，从而提高预测性能和准确度。通过投票法，可以借助多个模型的意见来得出最终预测结果，提高预测正确率。

（三）TOPSIS—熵权法能够考虑指标间的相对性：TOPSIS 方法基于指标之间的相似度来评估方案的优劣程度。熵权法可以通过计算指标之间的信息熵来反映指标的相对性。结合这两种方法可以更好地衡量指标之间的关系，充分考虑到它们的相对贡献和相关性。

6.2 模型的缺点

（一）模型融合算法的解释性和可读性可能相对较差。由于预测结果来自多个模型的投票和融合，理解每个单独模型的贡献和决策过程可能变得困难。

（二）典型相关分析可能对数据分布的假设比较敏感，强制性地要求两组典型变量服从正态分布。如果数据违反了假设的分布条件，模型的结果可能会失真。

6.3 模型的改进

（一）我们可以进一步调整模型融合策略，例如权重调优或组合方法的改进。在投票法之外，还可以考虑其他集成学习方法，如加权投票、堆叠（stacking）或层次化投票，以进一步提高预测性能。

（二）鉴于典型相关模型对数据分布的敏感性，我们可以考虑使用非参数或鲁棒方法来处理数据，以减轻对正态分布要求的假设，可以更好地适应数据的实际情况。

6.4 模型的推广

本论文构造的集成学习融合模型，通过组合多个机器学习算法来提高性能和稳定性，解决了问题二、三、四、五。这表明该模型在一定程度上具有推广性，可以应用于其他类似的领域和实际问题，比如通过分析和预测与母婴关系相关的变量之间的关系，可以帮助研究人员更好地理解母婴关系的影响因素，并预测特定情境下的结果。

七、 参考文献

[1]费宇 《多元统计分析-基于 R》，人民大学出版社，2012.
[2]李博纳,赵新泉 概率论与数理统计[M]，北京：高等教育出版社，2009.
[3]张敬信 《数学建模：算法与编程实现》，机械工业出版社，2022

附录

Python 代码

问题二、问题三、问题五求解代码：集成学习模型.py

```
from sklearn.naive_bayes import GaussianNB, BernoulliNB
from sklearn.model_selection import cross_val_score,StratifiedKFold
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor, VotingClassifier
import pandas as pd

from sklearn.neighbors import KNeighborsRegressor
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPRegressor
from sklearn.neural_network import MLPClassifier

from sklearn.tree import DecisionTreeRegressor
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression

from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split

from xgboost import XGBClassifier
from xgboost import XGBRFClassifier

if __name__ == '__main__':
    # 母亲身体指标
    mother_physics = ['mother_age', 'marriage', 'education', 'pregnant', 'birth_method']
    # 母亲心理指标
    mother_mental = ['CBTS', 'EPDS', 'HADS']
    # 婴儿睡眠质量,
    sleep_quality = ['wake_times', 'sleep_way', 'sleep_time']
    #婴儿信息
    baby = ['baby_gender',    'baby_age']
    # 婴儿行为模式
    action_mode = ['mode']
```

```
    #导入训练集
    data = pd.read_excel('data.xlsx')
    #导入预测集
    target = pd.read_excel('predict.xlsx')
    X = data[mother_mental+mother_physics+baby]
    y = data[action_mode]
    x = target[mother_mental+mother_physics+baby]

    #切分训练集和验证集
    #X_train, X_test, y_train, y_test = train_test_split(X, y,train_size=0.75,shuffle=True)

     #更换各种模型进行预测
     MLP =MLPClassifier()

     rf = RandomForestClassifier()
     dt = DecisionTreeClassifier()
     logic = LogisticRegression()
     KNN = KNeighborsClassifier()
     xgb = XGBRFClassifier()


     #投票分类器
     vote =
VotingClassifier(estimators=[('MLP',MLP),('rf',rf),('xgb',xgb),('logic',logic),('KNN',KNN)],vo
ting='hard')
     model = xgb
     #model.fit(X,y)
     #print(model.predict(x))
     '''
     #分层 k 折交叉验证
     #更换 k 值以求得最佳比例
     stratifiedkf = StratifiedKFold(n_splits=5)
     #计算准确率
     score = cross_val_score(model,X,y,cv=stratifiedkf)
     print(score)
     score=pd.DataFrame(score)
     score.to_csv("result.csv",index=False,mode='a')
     #rf.fit(X_train,y_train)
'''
     #模型训练
     model.fit(X,y)
     #预测未知婴儿模式
     result = pd.DataFrame(model.predict(x))
     print(result)
```

```
    #result.to_csv('presume.csv',index=False,mode='a')
    #print(rf.score(X_test,y_test))
```

问题二代码：线性回归画图.py

```
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']  # 显示中文
plt.rcParams['axes.unicode_minus'] = False  # 正常显示负号

# 数据
cbts_scores = [0, 5]
cbts_costs = [200, 810.667*5+200]

epds_scores = [0, 5]
epds_costs = [500, 695*5+500]

hads_scores = [0, 5]
hads_costs = [300, 12500]

# 绘制图形
plt.plot(cbts_scores, cbts_costs, label='CBTS')
plt.plot(epds_scores, epds_costs, label='EPDS')
plt.plot(hads_scores, hads_costs, label='HADS')

# 添加标题和标签

plt.xlabel('得分')
plt.ylabel('治疗费用')

# 添加图例
plt.legend()

# 设置横坐标范围
plt.xlim(0, 5)

# 设置纵坐标范围
plt.ylim(0, 13000)

# 显示网格线
plt.grid(True)

# 显示图形
plt.show()
```

问题四代码：topsis 熵权法.py

```
import numpy as np
import pandas as pd
import openpyxl

# 指标属性同向化
def normalize_direction(matrix, reverse_cols):
    normalized_matrix = np.copy(matrix)
    for col in reverse_cols:
        normalized_matrix[:, col] = np.max(normalized_matrix[:, col]) -
normalized_matrix[:, col]
    return normalized_matrix

def entropy_weight(x):
    # 处理数组，将 0 替换为一个较小的非零值
    x_processed = np.where(x == 0, 1e-10, x)
    # 计算每个指标的熵值
    m, n = x.shape
    e = np.zeros((1, n))
    for j in range(n):
        p = x_processed[:, j] / x_processed[:, j].sum()
        e[0][j] = - (p * np.log(p)).sum()
    #print(e)
    # 计算每个指标的权重
    w = np.zeros((1, n))
    for j in range(n):
        w[0][j] = (1 - e[0][j]) / (np.sum(1 - e))
    return w

def topsis(x, w):
    # 将 x 归一化处理
    m, n = x.shape
    x_norm = np.zeros((m, n))
    for j in range(n):
        x_norm[:, j] = x[:, j] / np.sqrt((x[:, j]**2).sum())
    # 计算加权后的矩阵
    x_weighted = np.zeros((m, n))
    for j in range(n):
        x_weighted[:, j] = w[0][j] * x_norm[:, j]
    # 计算最优解和最劣解
    max_vec = x_weighted.max(axis=0)
    min_vec = x_weighted.min(axis=0)
```

```
        # 计算每个评价对象与最优解和最劣解的距离
        d_plus = np.sqrt(((x_weighted - max_vec)**2).sum(axis=1))
        d_minus = np.sqrt(((x_weighted - min_vec)**2).sum(axis=1))
        #print( d_minus)
        # 计算得分
        score = d_minus / (d_minus + d_plus)
        return score

x = np.array([[1,2,3],[4,5,6]]) #示例数据
reverse_cols = [0, 1]   # 第一列和第二列是负指标
x = normalize_direction(x, reverse_cols)

# 计算熵权法得到的权重
w = entropy_weight(x)

# 计算 TOPSIS 得分
score = topsis(x, w)
#print(w)

#df = pd.DataFrame(score, columns=['Score'])
# Export the DataFrame to an Excel file
#df.to_excel('score.xlsx', index=False)
```