Maternal And Child Health Study Based on Decision Tree and Partial Least Squares Regression

Risheng Xu *

School of Automation and Electrical Engineering, Zhejiang University of Science &Technology, Hangzhou, China

* Corresponding Author Email: 3288142802@qq.com

Abstract. The physical and mental health of mothers plays a crucial role in the development of infants. In this paper, the influence of mothers' physical and mental indicators on infants' behavioral characteristics, the comprehensive sleep quality assessment system, and the treatment plan for mental illnesses are studied in depth. First, the decision tree model is used to explore the relationship between mothers' physical and psychological indicators on infants' behavioral characteristics. Second, a linear programming model was developed between the degree of illness and the scores of the three psychological indicators to derive the minimum cost of the mother's psychotherapeutic program that can change the infant's behavioral characteristics. Finally, the infant sleep quality indicators were graded, the weights of each sleep quality indicator were calculated using AHP hierarchical analysis, and the formula for calculating the sleep quality grade was derived using a partial least squares regression model. The study in this paper evaluates the physical and psychological indicators of mothers in a scientific way, thus ensuring the quality of mother-infant interaction.

Keywords: Decision tree; AHP; partial least squares regression.

1. Introduction

The physical and mental health of mothers is crucial to the growth of infants, especially the behavioral characteristics and sleep quality of infants are closely related to the physical and mental indicators of mothers [1], and this paper carries out an in-depth study on this issue. Firstly, we explore the relationship between mothers' physical and psychological indicators on infants' behavioral characteristics, calculate the corresponding configuration parameters and training time with the help of the decision tree model, calculate the scores of each indicator according to the correlation results, determine the intervals of the behavioral characteristics, and derive the behavioral characteristics of the last twenty infants. Then, the functional relationship between the degree of illness and the scores of the three psychological indicators was established, and the expression for the rate of change of the degree of illness was derived. Based on the fact that the rate of change in prevalence is positively related to the cost of treatment, a linear programming model between the cost of treatment and the change scores was developed to derive the minimum cost of a mother's psychotherapeutic program capable of changing the infant's behavioral characteristics. Finally, the infant sleep quality indicators were graded, the more favorable the infant sleep quality the closer to 5, the more unfavorable the closer to 1. Then the weight of each sleep quality indicator was calculated using AHP hierarchical analysis, and the model accuracy was verified with the help of consistency test to establish a comprehensive judgment system for infant sleep quality. Using the partial least squares regression model, the factor loading coefficients were calculated, and the formula for the sleep quality rating was derived. Thereby, the study in this paper provides help to improve the physical and mental health of mothers and the growth of infants.

2. Decision tree model

2.1. Mathematical principles of classification trees

Decision trees divide data recursively and can handle both categorical and numerical features, both uncorrelated data and continuous and discrete data, and uncorrelated features do not affect the decision tree [2].

The classification criterion used by ID3 is information gain, which indicates the degree to which the uncertainty of a sample set is reduced by learning about the information. Information entropy of the dataset:

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|} \tag{1}$$

Where C_k denotes the subset of samples belonging to the kth class of samples in the set D. For a certain feature A, the conditional entropy $H(D|A)$ for dataset D is:

$$\begin{aligned} H(D \mid A) &= \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) \\ &= -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \left(\sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} log_2 \frac{|D_{ik}|}{|D_i|} \right) \end{aligned} \tag{2}$$

Where D_i denotes the subset of samples in D for which feature A takes the ith value, and D_{ik} denotes the subset of samples in D_i that belong to the $k-th$ class.

Information gain = information entropy-conditional entropy: $Gain\ (D, A) = H(D) - H(D \mid A)$. A larger information gain indicates a larger purity gain obtained by using feature A for segmentation.

2.2. Behavioral Characteristics Results

The importance of the features is calculated through the created decision tree. The calculated importance ratios for each feature (dependent variable) are shown in Table 1 below:

Table 1. Importance of each characteristic

Function Name	Importance of characteristics
EPDS	29.00%
HADS	22.90%
Duration of pregnancy (weeks)	20.00%
Age of mother	12.00%
CBTS	9.50%
Educational attainment	5.20%
Marital status	1.50%
Delivery method	0.00%

Also, we plotted a heat map to show the confusion matrix, as shown in Fig. 1.

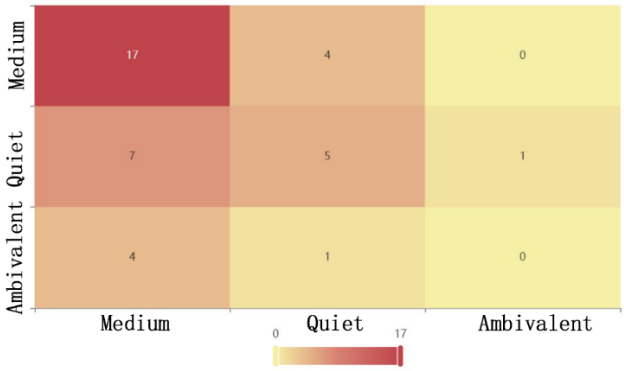

Fig. 1 Box plot of infant sleep quality and behavioral characteristics

Then, we apply the established decision tree classification model to the training data and test data to get the classification evaluation results of the model. Table 2 below lists the classification evaluation metrics for the training set and test set, and these quantitative metrics are used to measure the classification effect of the decision tree on the training data and test data.

Table 2. Results of model evaluation

Data set	Accuracy	Recall rate	Precision
Training set	0.798	0.798	0.799
Cross validation set	0.479	0.479	0.429
Test equipment	0.564	0.564	0.494

The data divided into test sets were substituted into the model for prediction and the results obtained are shown in Table 3.

Table 3. Predictions of infant behavioral characteristics

Outcome	Behavioral characteristics	Predicted Medium	Predicted Quiet	Predicted Conflictual	Mother's age	Marital status	Education	Duration of pregnancy		Mode of delivery	CBTS	EPDS
Medium	Medium	1	0	0	36	2	3	40	1	0	0	2
Medium	Medium	0.882	0.117	0	32	2	5	38	1	0	0	0
Medium	Quiet	0.615	0.230	0.153	26	2	3	40.1	1	0	4	4
Quiet	Medium	0.090	0.909	0	31	2	3	39	1	4	7	12
Medium	Medium	0.857	0.023	0.119	22	2	3	34.2	1	20	20	7
Medium	Medium	0.857	0.023	0.119	24	1	2	37.6	1	6	14	14
Medium	Quiet	1	0	0	36	2	4	38	1	10	11	12
Medium	Quiet	0.857	0.023	0.119	24	2	4	39.5	1	2	5	6
Quiet	Quiet	0	1	0	27	2	4	39	1	14	16	9
Medium	Medium	0.857	0.023	0.119	25	2	5	39	1	9	18	10
Medium	Medium	0.857	0.023	0.119	25	2	3	38	1	10	9	13
Medium	Medium	0.714	0	0.285	32	2	5	34.3	2	5	13	5
Quiet	Medium	0.2	0.8	0	28	2	4	41	1	5	9	9
Medium	Medium	0.714	0	0.285	37	2	5	41	1	8	12	8
Medium	Medium	0.878	0.090	0.030	35	2	5	39.3	1	5	6	10

From the table, it can be seen that the predicted results of infant behavioral characteristics can be used for prediction with a high degree of accuracy of 66.7% compared to the actual situation, which is a good fit.

3. Linear programming model

3.1. Model building

The behavioral characteristics of infants can be influenced by intervening on mothers' anxiety and improving their mental health in order to change the behavioral characteristics of infants. According to the analysis in Section 2, we know that the rate of change of treatment cost relative to the degree of prevalence of CBTS, EPDS, and HADS are all positively proportional to the cost of treatment, so we construct a linear function between the rate of change of the degree of prevalence of CBTS, EPDS, and HADS and the cost of treatment, respectively [3]. The full scores of CBTS, EPDS, and HADS are all 30, and the higher the score, the more serious the prevalence of the disease.

The degree of prevalence is denoted as $f(x)$, and the pre-treatment scores for CBTS, EPDS, and HADS are denoted as G, and the post-treatment scores are denoted as g. The prevalence formula is as follows:

$$F(X) = \frac{G_{max} - G}{G_{max} - G_{min}} \tag{3}$$

Where G_{max} = 30, and G_{min} = 0, so $F(X) \in [0,1]$, the closer $F(X)$ is to 0 the more severe the disease is and the closer it is to 1 the less severe the disease is. This gives the rate of change of the degree of illness $g(x)$ as:

$$d(x) = 1 - f(x) \tag{4}$$

From this we can construct a linear function $H(x)$ between the degree of illness and the cost of treatment as:

$$H(x) = Kd(x) + b \tag{5}$$

From equations (1) (2) (3),

$$H(x) = \frac{KG}{30-G} + b \tag{6}$$

Substituting the values of CBTS, EPDS, and HADS scores and treatment costs into the functional models developed above yielded a direct relationship between the prevalence of CBTS, EPDS, and HADS, respectively, and treatment costs:

$$H(x)_C = \frac{23508G}{30-G} + 200 \tag{7}$$

$$H(x)_E = \frac{19460G}{30-G} + 500 \tag{8}$$

$$H(x)_H = \frac{61000G}{30-G} + 300 \tag{1}$$

The mother's CBTS, EPDS, and HADS scores before treatment were recorded as $[G_1, G_2, G_3]$, the CBTS, EPDS, and HADS scores after treatment were recorded as $[g_1, g_2, g_3]$, and the change in the scores before and after treatment were recorded as $[a_1, a_2, a_3]$. The cost H of transforming the infant's behavioral characteristics through psychotherapy with the mother is:

$$H = H(a_1)_C + H(a_2)_E + H(a_3)_H \tag{2}$$

3.2. Establishing constraints

We used the minimum and average scores on each of the three psychological indicators for mothers of medium-sized infants as functional independent variables $[a_1, a_2, a_3]$ constraints. For making the behavioral characteristics of the infants to change from ambivalent type to medium type, it is necessary to satisfy:

$$\min_x \{H = H(a_1)_C + H(a_2)_E + H(a_3)_H\} \tag{3}$$

$$s.t. \begin{cases} 0 \le 15 - a_1 \le 6 \\ 0 \le 22 - a_2 \le 9 \\ 0 \le 18 - a_3 \le 8 \\ a_1 \in Z, i = 1,2,3 \end{cases} \tag{12}$$

The solution is obtained when a_1 = 9.a_2 =13.a_3 = 10, i.e., when the mother's CBTS, EPDS, and HADS scores are 6, 9, and 8, respectively, after treatment, the infant's behavioral profile shifts from ambivalent to moderate, and at this point the treatment costs the least, approximately $56,456.

Similarly, in order to make the infant's behavioral characteristics change from ambivalent to quiet, it is necessary to use the minimum and average scores of the three types of psychological indicators of the mothers of quiet infants as the constraints of the function independent variables $[a_1, a_2, a_3]$, respectively, the function constraints are:

$$s.t.\begin{cases}0 \le 15 - a_1 \le 5\\ 0 \le 22 - a_2 \le 8\\ 0 \le 18 - a_3 \le 8\\ a_1 \in Z, i = 1,2,3\end{cases} \tag{4}$$

Solving for a_1=10, a_2=14, a_3=10, when the mother's CBTS, EPDS, and HADS scores were 5, 8, and 8, respectively, after treatment, the infant's behavioral profile changed to quiet, at which point the cost of treatment was the least, approximately $59,282.

4. Comprehensive assessment system for infant sleep quality indicators

4.1. AHP Hierarchical Analysis

AHP hierarchical analysis divides complex problems into hierarchical structures based on dominance relationships [4], with each level consisting of interconnected and interacting elements. The relative importance of each element in the hierarchy is quantified by pairwise comparison and finally ranked. It divides the human thinking process into goal level, criterion level and program level, and analyzes them with the help of mathematical models, which is a practical decision analysis method that effectively combines the qualitative judgment of decision makers with quantitative calculation.

4.2. Comprehensive judgment on the classification of infant sleep quality

In order to establish a comprehensive judgment on the classification of infant sleep quality as excellent, good, moderate or poor, we decided to build an AHP model with the following steps:

First of all, we understood the division of advantages and disadvantages of infant sleep time, sleep-wake times, and sleep patterns by reviewing the literature and consulting experts. Finally, we came to the following conclusions: the more abundant the infant's nighttime sleep time is within the range of 12 hours, the better. The number of sleep-wake counts of 0-2 is the normal level, and in the case of greater than 2, the more wakes the more the quality of sleep is affected. The mode of falling asleep is to assist the infant to fall asleep on his/her own. The best way to fall asleep is to assist infants to fall asleep on their own, human intervention - coaxing/patting/stroking is second to pacifier, which will greatly reduce the proportion of children sleeping through the night.

Step 1. Develop scoring criteria for each indicator.

After inquiring relevant information, the three indicators of the infant's sleep time throughout the night, the number of waking times, and the way of falling asleep were categorized into five degree levels, and the indicator that had the greatest positive impact on the infant's sleep quality was recorded as 5, and the one that had the smallest positive impact on the infant's sleep quality was recorded as 1, and the results of the division are shown in Table 4.

Table 4. Delineation results

numerical value	Sleeping time throughout the night	Number of wake-ups	How to fall asleep
1	5	10	Pacifier method
2	6-7	7--9	Sleep pattern
3	8	5--6	Touching
4	9-10	3--4	Matrix
5	11-12	0--2	Timing

Step 2: Establishment of AHP hierarchical analysis model and construction of a comprehensive evaluation system for infant sleep quality

Based on the query data, the extent to which the duration of sleep throughout the night, the number of awakenings, and the way of falling asleep influenced the infant's sleep quality was determined, and then a judgment matrix, as shown in Table 5 below, was created:

Table 5. Judgement matrix

Norm	Sleep duration	Number of wakings	How to fall asleep
Sleep duration	1	4	8
Number of wake-ups	0.25	1	3
How to fall asleep	0.125	0.333	1

Step3. Consistency check

In order to ensure whether the constructed pairwise judgment matrix satisfies the positive and negative matrices, a consistency test is performed on it. The consistency of judgment has the following theorem: an n-order positive and negative matrix A is a consistency matrix when and only when its largest characteristic root$\lambda_{max} = n$, and when the positive and negative matrices$\bar{A}$ is consistent, there must be $\lambda_{max} > n$ by calculating the consistency index CI to consistency test, $CI = 0$ is that completely consistent, the larger the CI is inconsistent. Consistency index CI is calculated as: $CI = \frac{\lambda_{max}-n}{n-1}$. By calculating, the judgment matrix has $CI = 0.009$, which is very strong consistency. Apply the arithmetic mean method to the judgment matrix to solve the weights, and the formula is as follows:

$$\alpha_i = \frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \tag{5}$$

The calculation results are shown in Table 6:

Table 6. Weighting table for sleep quality

Term	Eigenvector	Weighted value (%)
Sleep duration	2.144	71.465
Number of wakings	0.619	20.644
How to fall asleep	0.237	7.891

In summary, after the consistency test, we find that the matrix is built correctly, i.e., the weights are established reasonably.

Step4. Calculation of the integrated evaluation system

From this, a comprehensive evaluation system for infant sleep quality can be constructed with the evaluation formula:

$$S = 0.71465X + 0.20644Y + 0.07891Z \tag{15}$$

Subsequently, we screened for sleep duration, number of wakings, and mode of falling asleep, disaggregating the screening substitutions to ensure that the interval for the dataset of infant sleep quality indicator scores lies within the range of 1 to 5 (numerical values). The above formula was then substituted to produce infant sleep quality scores, which also lie within the range of 1 to 5 (numerical values). Finally, we comprehensively judged the infant's sleep quality in four categories: excellent, good, moderate, and poor, with the specific measures shown in Table 7 below.

Table 7. Judging criteria

Value of a score	Rating
1--2	Poor
2--3	Medium
3--4	Good
4--5	Excellent

So, we calculated and organized to get the infant comprehensive sleep quality rating scale, some data are shown in Table 8 below.

Table 8. Infant Comprehensive Sleep Quality Rating Scale

Serial number	Sleep Quality Score	Sleep Quality Rating
40	1.49179	Poor
41	1.64961	Poor
42	3.993065	Good
43	4.150885	Excellent
44	3.993065	Good
45	3.993065	Good
46	1.49179	Poor
47	3.84218	Good
48	1.64961	Poor
49	1.49179	Poor
50	2.20644	Medium

4.3. Mother-child association modeling and PLSR regression forecasting

PLSR, is a multivariate statistical method that can solve the problem of covariance [5], the simultaneous analysis of multiple dependent variables Y, as well as affecting the study of relationships when dealing with small samples. Given the limitations of the presence of multiple independent variables (physical and psychological indicators of mothers) and one variable (sleep quality rating of infants), the uncertainty of a linear relationship, and the small sample size of the observed data, we decided to solve the problem by using the PLS partial least squares regression model.

First, we dummy-variate the infant's sleep quality rating, which can also be interpreted as temporarily transforming the rating term into a dummy variable. At the same time, we downgraded each value of the mother's physical condition by PCA.

Secondly, we refer to the maximum number of principal components by Cumulative Projected Importance (VIP). After calculation, we get the independent variable VIP (Cumulative Projected Importance), as shown in Fig. 2. From the case of VIP shown in Fig. 2, it can be seen that it indicates the strength of the explanatory importance of X for Y when the number of components is different, and it can also be used to refer to the maximum number of principal components. In this case, for independent variables with large VIP (>1), it plays a relatively larger role in explaining the underlying factors.

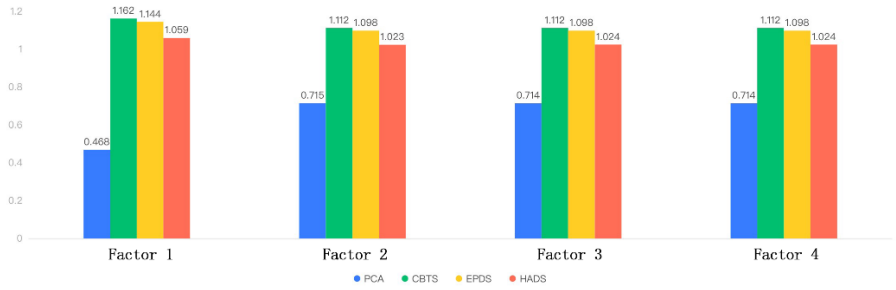

Fig. 2 VIP Histogram

Similarly, we calculated the required: component matrix table, factor loading coefficients table and model coefficient results table of the model in turn to show the results of this PLS model, which mainly includes the coefficients of the model for analyzing the relationship of the influence of the independent variable X on the dependent variable Y. The results are shown in Table 9, Table 10, and Table 11.

Table 9. Component matrix

Variant	Factor 1	Factor 2	Factor 3	Factor 4
PCA	0.234	1.007	0.052	0.06
CBTS	0.581	0.09	-0.473	-0.584
EPDS	0.572	-0.029	-0.329	0.794
HADS	0.529	-0.131	0.82	-0.238
Sleep Quality Rating	14.713	6.688	-2.038	0.543

Table 10. Table of factor loading coefficients

Variant	Factor 1	Factor 2	Factor 3	Factor 4
PCA	0.04	0.972	0.149	0.069
CBTS	0.587	-0.067	-0.583	-0.667
EPDS	0.601	-0.173	-0.202	0.736
HADS	0.577	-0.17	0.793	-0.094
Sleep Quality Rating	0.055	0.025	-0.008	0.002

Table 11. Table of model coefficient results

	Sleep Quality Rating	Sleep quality rating (standardized)
Constant	2.086	0
PCA	0.148	0.038
CBTS	0.073	0.037
EPDS	0.052	0.035
HADS	0.041	0.019

This completes the modeling of the mother-infant association. The standardized formula for the model is: sleep quality rating = $2.086 + 0.148 * PCA - 0.073 * CBTS + 0.052 * EPDS - 0.041 * HADS$.

Let sleep quality rating be the dependent variable y and PCA, CBTS, EPDS, and HADS be the independent variables x, which are set as x_1, x_2, x_3, x_4, then $y = 2.086 + 0.148x_1, -0.073x_2 + 0.052x_3 - 0.041x_4$.

5. Conclusion

In this paper, an in-depth study was conducted on the influence of mothers' physical and mental indicators on infants' behavioral characteristics, the comprehensive judgment system of sleep quality, and the treatment plan for mental illness. First, with the help of decision tree model to explore the relationship between mothers' physical and psychological indicators on infants' behavioral characteristics, the scores of each indicator were calculated to derive the latter twenty infants' behavioral characteristics. Then, a linear programming model was used to establish the functional relationship between the degree of illness and the scores of the three psychological indicators, and the minimum treatment costs were calculated for infants whose behavioral characteristics changed from ambivalent to moderate and quiet, respectively. Finally, the weights of each sleep quality indicator were calculated using AHP hierarchical analysis, and the accuracy of the model was verified with the help of consistency test, so as to establish a comprehensive judgment system for infant sleep quality. Using the partial least squares regression model, the factor loading coefficients were calculated, and the formula for the sleep quality rating was derived. Thus, it provides a scientific basis to help mothers rationalize their physical and psychological indicators and promote mother-infant interaction.

References

[1] Bayer, Jordana K., et al. Sleep problems in young infants and maternal mental and physical health. Journal of paediatrics and child health 43.1-2 (2007) 66-73.

[2] Kim, Kyoungok, and Jung-sik Hong. "A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis." Pattern Recognition Letters 98 (2017): 39-45.

[3] Lavery, Jessica A., et al. "Gestational diabetes in the United States: temporal changes in prevalence rates between 1979 and 2010." BJOG: An International Journal of Obstetrics & Gynaecology 124.5 (2017): 804-813.

[4] Darko, Amos, et al. "Review of application of analytic hierarchy process (AHP) in construction." International journal of construction management 19.5 (2019): 436-452.

[5] Abdi, Herve, and Lynne J. Williams. "Partial least squares methods: partial least squares correlation and partial least square regression." Computational Toxicology: Volume II (2013): 549-579.