ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

A task-prediction model based on the random forest algorithm

Shuo Ma[a,*], Zhihao Zhao[b], Junfei Sun[c]

Inner Mongolia University of Science and Technology, Baotou, China
[a]13581102089@163.com, [b]1010204744@qq.com, [c]954673980@qq.com
[*]Corresponding author

Abstract: *The impact of maternal physical and mental health on infants is an important research field. We investigated the effects of maternal physical and psychological indicators on infant sleep quality through Spearman correlation analysis cluster analysis (K-Means) and random forest algorithm. The relative coefficient between each variable is calculated by organizing and optimizing the relevant data. At the same time, by optimizing the model to accurately predict and make optimal adjustments, the sleep quality of infants can be improved. By using a model to classify and predict training and testing data, and evaluating the accuracy, recall, accuracy, and F1 value of the model, it is concluded that there is a practical relationship between the mother's age, EPDS score, marital status, and other characteristics with the rating.*

Keywords: *Physical indicators, psychological indicators, sleep quality, random forest algorithm*

1. Introduction

The physical and mental health of mothers has a significant impact on the growth of infants. Firstly, the physical health of a mother is crucial for the development of a baby. A healthy mother can provide sufficient nutrition and energy to promote normal fetal development. The health status of the mother is also closely related to the sleep and health status of the baby after birth. Secondly, the mental health of mothers is crucial for the emotional and cognitive development of infants. A stable and positive mother can establish a safe parent-child relationship and provide ***sufficient*** care and support to the baby. On the contrary, mother's psychological problems such as anxiety and depression may affect emotional communication and parent-child interaction with the baby, causing negative effects on the baby's emotions and behavior. Emphasizing the physical and mental health of mothers is crucial for the growth of infants. By providing necessary support and care to ensure the physical and mental health of mothers, a favorable growth environment can be created for infants, promoting their comprehensive and healthy development. Through analysis of variance and Spearman correlation analysis, the differences and correlations between different clustering categories were revealed[1].

2. Model preparation

Assuming that the observation values between variables are independently sampled, there is no correlation between the physical and psychological indicators of each group of mothers and the data of infant sleep quality indicators.

2.1 Data preprocessing

Data processing mainly includes. Data cleaning: handling missing values, outliers, and duplicate values. This includes filling in missing values, deleting or correcting outliers, and detecting and processing duplicate data. The sleep time of sample 180 was 99.99, and there was an obvious error, so it was deleted. In the question, there are only two types of marital status: 1 and 2, while some samples have 3 and 6, which do not meet the conditions given in the question, so they will also be deleted. Feature selection: Select the most meaningful feature for the target task. By analyzing and evaluating the relationship, correlation, and importance between features, irrelevant or redundant features are excluded to improve the effectiveness of subsequent analysis. Feature transformation: Transforming raw data into a form suitable for model use. This may involve normalizing, standardizing, discretizing, or scaling the

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

data to eliminate differences in measurement units between different features or reduce the impact of noise. The training set is used for model training and parameter adjustment, the validation set is used for model selection and tuning, and the testing set is used to evaluate the performance of the model on new data. We convert the infant behavioral characteristics in the attachment, marking the quiet type as 0, the equal type as 1, and the contradictory type as 2. We convert the sleep time (hours: minutes: seconds) into hours. This involves operations such as merging, connecting, and stacking data for subsequent analysis and modeling. Data balancing: dealing with category imbalance issues, adjusting the sample ratio of different categories in the dataset through methods such as undersampling and oversampling to improve the performance of the classification model.

3. Problem analysis

Studying the impact of maternal physical and psychological indicators on infant behavioral characteristics and sleep quality. Firstly, it is necessary to establish a mathematical model, analyze the data in the attachment, and explore the correlation between maternal physical and psychological indicators on infant behavioral characteristics and sleep quality. Secondly, by establishing a relationship model between infant behavioral characteristics and maternal physical and psychological indicators, the behavioral characteristic types of the 20 deleted groups of infants were predicted. In addition, the impact of maternal anxiety interventions on mental health was studied, the relationship between treatment costs and mental health indicators was analyzed, and the conversion of contradictory behavioral characteristics into mild behavior was calculated[2-3].

Problem 1 analysis: Regarding the study of the impact of maternal physical and psychological indicators on infant behavioral characteristics and sleep quality, we describe it by obtaining correlation coefficients between various variables. Firstly, we preprocess the data in the attachment, conduct missing value testing and filling, and then convert other types of data into numerical types. Then we solved the Spearman coefficients between each variable, and used the obtained coefficients and significance P-values to analyze whether the mother's physical and psychological indicators had a significant impact on the baby's behavioral characteristics and sleep quality, and to analyze the degree and direction of the impact. The main task is to select random input variables from the processed data, further randomly combine input variables, and finally determine the random feature number.

Problem 3 analysis: Based on the data provided in Table 1, we can analyze the changes in infant behavioral characteristics and corresponding treatment costs based on the corresponding treatment costs for different scores. To change the infant's behavioral characteristics from contradictory to moderate, we need to calculate the changes in CBTS, EPDS, and HADS scores. Based on the data in the table, the treatment costs for CBTS, EPDS, and HADS were reduced, and statistical calculations were conducted to determine the behavioral characteristics of infants from contradictory to moderate, with the minimum required treatment costs. The behavioral characteristics of infants were adjusted according to the actual situation, and the corresponding treatment costs were calculated. Analysis of Question 4: For evaluating the comprehensive sleep quality of infants, indicators such as total sleep time, number of awakenings, and sleep patterns can be used for comprehensive evaluation. According to the specific situation, sleep quality can be divided into four categories: excellent, good, and medium to poor. A correlation model between the comprehensive sleep quality of infants and the physical and psychological indicators of mothers can be established. Cluster analysis (K-Means) and random forest algorithm are selected as a whole, and existing data are used to train the model, evaluate and optimize it. Finally, the trained model is used to predict the data of the last 20 groups of infants (numbered 391-410), Obtain the predicted comprehensive sleep quality classification results. Analysis of Question 5: Based on Question 3, the sleep quality of Baby 238 was rated as excellent. After adjusting the treatment strategy, a better optimization method was determined and specific adjustments were made.

4. Model establishment and solution

4.1 Solution to question two

1) Prediction result: 2 0 2 1 1 1 2 1 2 0 1 1 0 0 1 0 0 0 0 1

Chart description: The above table shows the preview results, only partial data is displayed. Please click the download button to export all the data. The above table shows the classification results of the random forest model on the test data, with the classification result values being the classification group

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

with the highest prediction probability. Summary of predictions: Based on the predicted results, the behavioral characteristics of each infant and the corresponding probability of predicted test results can be seen. The behavioral characteristics of infants are classified into quiet, moderate, and contradictory types, which predict the probability of test results_ 0.0. Probability of Predicting Test Results_ 1.0 and Probability of Predictive Test Results_ 2.0 represents the probability of belonging to quiet, moderate, and contradictory types, respectively. The predicted behavioral characteristics of infants numbered 391, 393, 397, and 399 are contradictory. The predicted behavioral characteristics of infants numbered 392, 400, 413, 414, and 416 were quiet.

The predicted behavioral characteristics of other infants are important codes for equality:

```
import seaborn as sns
from sklearn.metrics import confusion-matriximport matplotlib.pyplot as plt
sns.set()
f,ax = plt.subplots()
y_true = [,0,1,2,1,2,,2,2,,1,1]y_pred =[1,0,1,2,1,0,8,2,2,@,1,1]C2 =
confusion_matrix(y_true,y_pred,labels=[0,1,2])# C2 print(C2)
sns.heatmap(C2,annot=True,ax=ax) # ax.set_titile('confusion matrix') # ax.set_xlabel('predict') # ax.set_ylabel('true') #
```

4.2 Problem three solution

Based on the scores of CBTS, EPDS, HADS and corresponding treatment cost data provided by the title, we can adjust these scores to change the behavioral characteristics of infants. Based on the scores of CBTS, EPDS, and HADS given in the title and corresponding treatment cost data, we can adjust these scores to change the behavioral characteristics of infants. Based on the scores of CBTS, EPDS, and HADS given in the title and corresponding treatment cost data, we can adjust these scores to change the behavioral characteristics of infants. Further reduction in CBTS, EPDS, and HADS scores is necessary to change the behavioral characteristics of infants to a quiet type. The specific treatment plan and adjustments need to be developed based on the doctor's advice, and then the required treatment costs will be recalculated based on the data in Table 1.

Important code

```
%%Unit cost k1=(2812-200)/3; k2=(1898-500)/2; k3=(12588-300)/5;
%% moderate type
c=[k1 k2 k3];
A=[1 0 0 ;0 1 0 ; 0 0 1;0.6 0.4 0.2];
1b=[9;13;10;50.45];
b=[15;22;18;51.7555];
[x,fval ]=intlinprog(c,[1 2 3],A,b,[ ],[ ],b); disp(x);
disp(c*x+1000);
%%quiet type c=[k1 k2 k3];
A=[1 0 0;0 1 0; 0 0 1;0.6 0.4 0.2]; 1b=[10;14;11;];
b=[15;22;18;50.45];
[x,fval]=intlinprog(c,[1 2 3],A,b,[],[],Ib); disp(x);
disp(c*x);
```

Output results: X1=9; X2=13; X3=10;

That is, min=k1x1 + k2x2 + k3x3 + 1000=41271, and the minimum cost of 41271 yuan of treatment, which can change the behavioral characteristics of infants from paradoxical to moderate.

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

4.3 Problem four analysis

Algorithm: Cluster Analysis (K-Means) Variable: {Total Night Sleep Time, Wake Up Times, Sleep Mode} Parameter: Number of Clusters: {4} Analysis results: Cluster analysis divides all samples into several categories based on data features: The clustering results are divided into four categories, with clustering categories_ The frequency of 1 is 116, accounting for 30.526%; Cluster Category_ The frequency of 2 is 35, accounting for 9.211%; Cluster Category_ The frequency of 3 is 96, accounting for 25.263%; Cluster Category_ The frequency of 4 is 133, accounting for 35.0%. Analysis steps: Category differential analysis is carried out on the basis of each field. The frequency of each cluster category is analyzed on the basis of the cluster summary. According to the clustering annotation of the dataset, it is possible to determine which category each sample data is classified into. The cluvster center coordinates can be used to analyze the distance between each sample and the center point[4].

Table 1: Output result 1: field difference analysis

Cluster category (mean ± SD)						
					F	P
	class 4(n=133)	class 1(n=116)	class 3(n=96)	class 2(n=35)		
Sleep time all night	11.269±0.716	9.698±1.323	9.661±1.139	9.0±2.0	61.463	0.000***
Wake up times	0.331±0.518	1.397±0.977	1.781±1.038	5.2±1.53	256.434	0.000***
The way to sleep	3.895±0.643	1.474±0.519	4.208±0.648	1.829±1.294	387.846	0.000***
Note: * * *, * * and * represent the significance levels of 1%, 5% and 10%, respectively						

TABLE 1. Show the results of the quantitative field difference analysis, including the results of the mean ± standard deviation, F-test results, and P-value of significance.

- The P-value for each analysis term was significant (P <0.05).

- If it is significant, the null hypothesis is rejected, indicating that there is a significant difference between the two groups of data. The difference can be analyzed according to the mean ± standard deviation, otherwise it indicates that the data does not show a difference.

Intelligent analysis: The results of the analysis of variance show that: For the variable overnight sleep time, the significance P-value is 0.000 * * *, showing significant significance at the level, rejecting the original hypothesis, indicating that there is a significant difference in the variable overnight sleep time among the categories classified by cluster analysis; For the variable wake-up frequency, the significance P-value is 0.000 * * *, showing significance at the horizontal level, rejecting the original hypothesis, indicating that there is a significant difference in wake-up frequency among the categories classified by cluster analysis; For the variable's sleep patterns, the significance P-value is 0.000 * * *, showing significant significance at the level, rejecting the original hypothesis, indicating that there is a significant difference in sleep patterns among the categories classified by cluster analysis;

Table 2: Output result 2: Cluster summary

Clustering categories	frequency	percentage%
Cluster category _1	116	30.526
Cluster category _2	35	9.211
Cluster category _3	96	25.263
Cluster category _4	133	35.0
total	380	100.0

TABLE 2 The above table shows the clustering results of the model, including frequency and percentage.

Intelligent analysis: The results of cluster analysis shown in Figure 1 that the clustering results are divided into four categories, Cluster Category_ The frequency of 1 is 116, accounting for 30.526%; Cluster Category_ The frequency of 2 is 35, accounting for 9.211%; Cluster Category_ The frequency of 3 is 96, accounting for 25.263%; Cluster Category_ The frequency of 4 is 133, accounting for 35.0%.

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

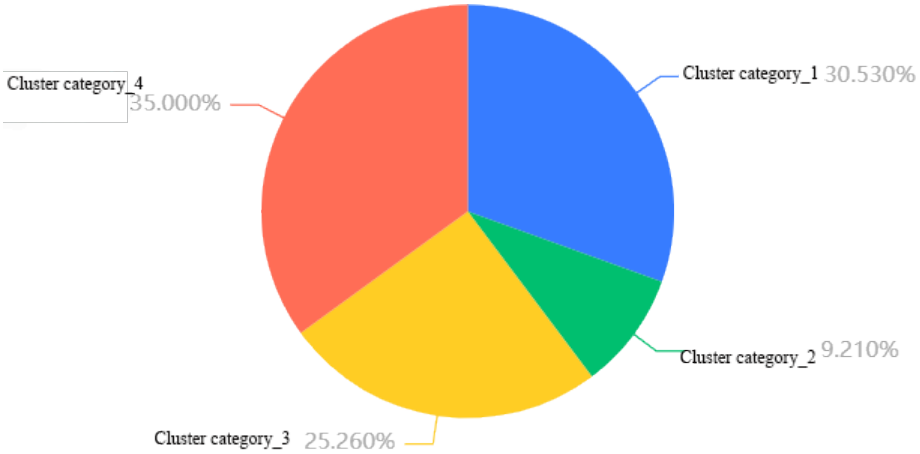

Figure 1: Output result 3: Cluster summary diagram

Figure 2 presents the clustering results of the model in a visual form, including frequency and percentage.

Table 3: Output result 4: data cluster annotation

Cluster species	Sleep time all night	Wake up times	The way to sleep
class 1	10	3	2
class 4	11	0	4
class 4	12	1	2
class 1	11	2	1
class 4	10.5	1	4
class 4	12	0	4
class 3	10	1	4
class 3	10	1	4
class 1	10	1	2
class 4	11	0	4
class 1	12	3	2
class 3	10	2	4
class 4	11	0	4
class 4	12	0	4
class 1	10	2	1

The TABLE 3. grid displays partial data clustering annotations of the model's clustering results, which are preview results and only display the top 15 comprehensively sorted items.

Table 4: Output result 5: Cluster center point coordinates

Cluster species	Center value _ All-night sleep time	Center value _ Wake times	Center value _ Sleep mode
1	9.698275862068966	1.396551724137931	1.4741379310344813
2	9	5.199999999999998	1.8285714285714292
3	9.661458333333334	1.7812500000000002	4.208333333333332
4	11.268796992481205	0.33082706766917425	3.8947368421052624

Table 4. displays some (or all) data from the clustering centers of the model, all of which can be downloaded from Excel by clicking on the top right corner.

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

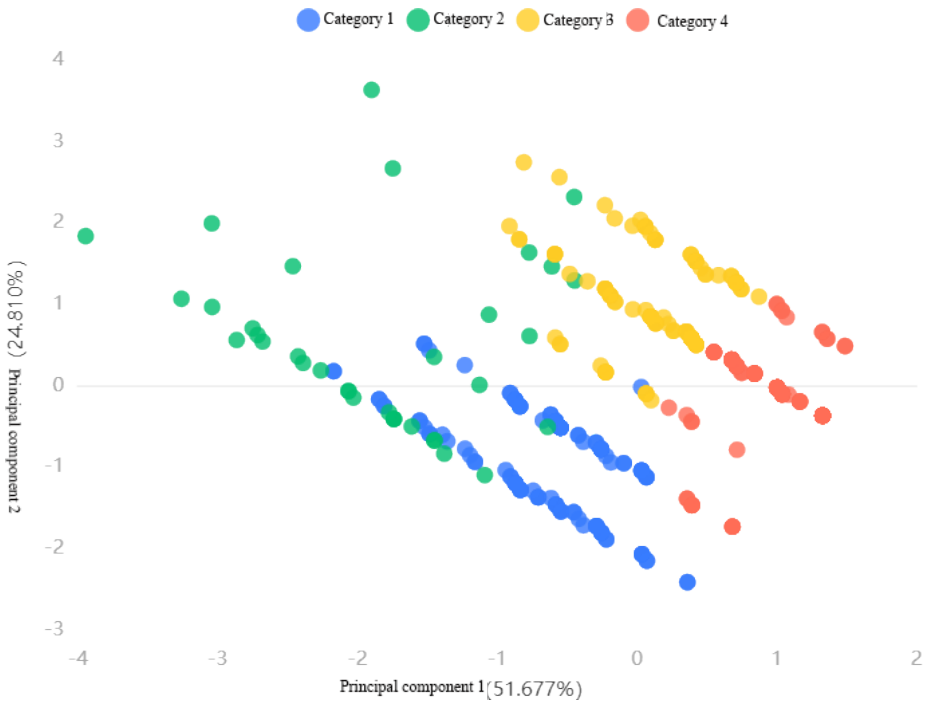

Figure 2: Output result 6: cluster scatter plot

Fig. 3 If the number of variables is equal to two, the above figure is a scatter plot based on the data of two variables; if the number of variables is more than two, the above figure is the first two principal components after the principal component analysis (PCA) to reduce the principal components, so as to view the clustering effect to some extent (if the variance interpretation rate of the first two principal components is low, the significance of the plot is not significant).

• The cluster scatter plot only shows the maximum sample size information of 1000. If the sample size is greater than 1000, the random sampling in the whole sample is carried out, and 1000 samples are selected for the scatter plot display.

Table 5: Output result 7: Evaluation index

The contour coefficient	DBI	CH
0.326	1.11	180.247

TABLE 5. Description:

• Profile coefficient: For a sample set, its profile coefficient is the average of all sample profile coefficients. The range of contour coefficient values is [-1,1]. The closer the distance between samples in the same category, the farther the distance between samples in different categories, and the higher the score, the better the clustering effect.

• DBI (Davies boldin): This indicator is used to measure the ratio of intra cluster distance to inter cluster distance between any two clusters. The smaller the indicator, the better the clustering effect.

• Calinski Haerbasz Score: measures the compactness within a class by calculating the sum of the squares of the distances between each point within the class and the center of the class (denominator), and measures the separability (numerator) of the dataset by calculating the sum of the squares of the distances between the center points between classes and the center points of the dataset. The CH index is obtained by the ratio of separability to compactness, and a larger CH indicates better clustering performance. Algorithm: Random forest classification. Variable: Variable X: {Mother's age, EPDS, marital status, HADS, education level, delivery method, gestational time (weeks), CBTS}; Variable Y: {Rating}

a) analytic result:

Random forest classification evaluates the model based on accuracy, recall, accuracy, and F1 indicators. Please refer to the detailed conclusions.

b) procedure of test

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

The established random forest classification model was applied to training and test data to obtain the classification evaluation results of the model. Due to the randomness in the random forest, the results of each operation are different. If the training model is saved, the data can be directly uploaded to the training model for calculation and classification in the future. Note: Random forests cannot obtain deterministic equations like traditional models, and models are usually evaluated by testing their data classification performance.

c) Detailed conclusions

Table 6: Output Result 1: Model Parameters

Parameter Name	Parameter value
Training time	0.126s
Data segmentation	0.9
data shuffle	deny
Cross validation	deny
Node splitting evaluation criteria	gini
Number of decision trees	100
There is a return sampling	true
Out of bag data testing	false
Maximum feature ratio considered during partitioning	auto
Minimum number of samples for internal node splitting	2
Minimum number of samples for leaf nodes	1
Minimum weight of samples in leaf nodes	0
The maximum depth of the tree	10
Maximum number of leaf nodes	50
Threshold for impure node partitioning	0

TABLE 6. The model parameter configuration and the model training time are shown.

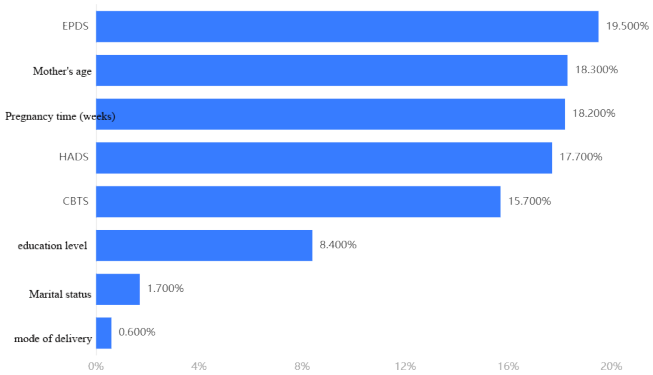

Figure 3: Output result 2: Characteristic importance

Fig. 3. Description: The upper bar chart or table shows the importance proportion of each feature (independent variable).

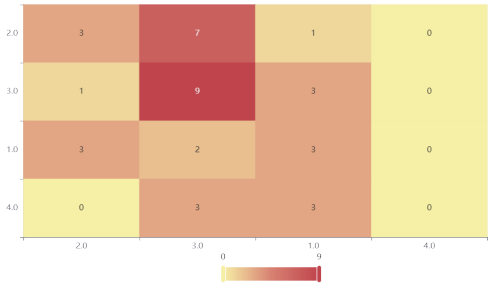

Figure 4: Output result 3: Heat map of the confusion matrix

Fig. 4. Description: The above table shows the confusion matrix in the form of a heat map.The above figure shows the classification results of the test set, presenting the confusion matrix in the form of a heatmap. The majority of samples are classified correctly, indicating that the model trained on the training set is effective and practical. From a data perspective, we generally use quantitative metrics such as accuracy, recall, precision, and F1 to measure the classification performance of random forests on training and test data:

TABLE 7 the classification evaluation indicators for training and testing sets are presented in, which

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

measure the classification effect of random forests on training and testing data through quantitative indicators.

Table 7: Output result 4: Model evaluation results

	Accuracy	recall	Accuracy rate	F1
training set	0.898	0.898	0.91	0.896
test set	0.395	0.395	0.334	0.348

• Accuracy: Predict the proportion of correct samples to the total sample, and the higher the accuracy, the better.

• Recall rate: In the actual positive sample results, the proportion of predicted positive samples is better with a larger recall rate.

• Accuracy: The proportion of predicted positive samples to actual positive samples, the higher the accuracy, the better.

• F1: The harmonic average of accuracy and recall, where accuracy and recall are interdependent. Although both are desirable, in reality, they often result in high accuracy and low recall, or low recall but high accuracy. If it is necessary to balance both, then the F1 indicator can be used.

• Oob_ Score: For classification issues, oob_ Score is the accuracy of data outside the bag. If we choose to have a return sampling during the tree building process, approximately one-third of the records have not been extracted. Unextracted data naturally forms a control dataset that can be used for model validation. So random forests do not need to reserve additional data for cross validation. Their own algorithm is similar to cross validation, and the out of pocket error is an unbiased estimate of the prediction error (only when the algorithm parameter selects "out of pocket test data" will the generalization ability of the model be tested through oob_score).

Table 8: Output result 5: Prediction result

Predicted results Y	grade	Predict the probability of the test results _1.0	Predict the probability of the test results _2.0	Predict the probability of the test results _3.0	Predict the probability of the test results _4.0	Maternal age	EPD S	Matrimoniares	HAD S	educational status	Mode of delivery	Pregnancy time (weeks)	CBT S
3.0	3.0	0.220354392218667 04	0.217230993539603 46	0.343731844914 109	0.218682769327620 37	26	12	2	7	3	1	40	5
2.0	1.0	0.31549076594334 5	0.446461131521415 37	0.141361004741021 37	0.096687097794218 36	32	16	2	13	3	1	37.4	5
3.0	3.0	0.174616396622565 42	0.338979623547325 14	0.416570498154272 7	0.069833481675836 71	35	3	2	10	5	1	39.7	3
3.0	2.0	0.236394576152253 95	0.292573370796310 35	0.352863115463955 9	0.118168937587479 72	32	13	2	5	5	2	34.3	5
3.0	3.0	0.264261384140634 3	0.286722006446638 53	0.370604149061871 5	0.078412460350855 33	29	3	2	4	4	1	38	2
3.0	3.0	0.34779922649053 3	0.220240377142179 17	0.353088569141402 6	0.078871827225885 26	26	11	2	12	5	1	41.2	7
3.0	3.0	0.118428106356889 16	0.200466615848386 82	0.510386857154486 2	0.170718420640237 8	23	2	2	3	3	1	37.5	0
1.0	2.0	0.483531945620049 03	0.170809425373618 43	0.188739837455776 2	0.156918791550556 13	29	9	2	11	5	1	41.2	6
3.0	2.0	0.263290909143552 36	0.251547790285284 9	0.360654713527008 4	0.124506586673354 29	27	0	2	9	5	1	40	6
3.0	4.0	0.247501813864518 22	0.250493347339444 9	0.450503044643769 83	0.051501794152267 12	26	17	2	11	5	1	40.4	10
3.0	3.0	0.234265337486873 25	0.221620944572584 58	0.468180864264454 5	0.075932853676087 44	34	0	2	4	3	1	38	0
3.0	4.0	0.228683230396558 66	0.208321242803720 02	0.435717468419365 43	0.127278058380355 94	31	11	2	4	3	1	40.2	3
1.0	4.0	0.351658366304843 6	0.337298778062657 3	0.244813093621075 6	0.066229762011423 61	25	11	2	15	5	1	38.5	5
2.0	3.0	0.265455267509838 44	0.537689315649093 8	0.077699866893615 77	0.119155549947452 03	32	24	2	19	3	1	35.4	19
3.0	2.0	0.318901256521611 5	0.252608191325861 15	0.368210393025208 64	0.060280159127318 544	31	8	2	8	5	1	39.5	11

TABLE 8 The grid is a preview result; only partial data will be displayed. Please click the download button to export all the data. The results of the classification of the random forest model are shown on the test data, classifying the outcome value as the group with the highest predictable probability.

ISSN 2616-5775 Vol. 6, Issue 11: 124-132, DOI: 10.25236/AJCIS.2023.061116

5. Conclusions

5.1 Merit

The use of cluster analysis to classify infant sleep quality and divide the samples into four categories provides an overall understanding of infant sleep quality. By frequency and percentage, the distribution of each category can be intuitively understood. Using Spearman correlation analysis, it was determined that maternal physical and psychological indicators have a certain impact on infant behavioral characteristics and sleep quality . Cluster summary analysis was conducted to provide the frequency and percentage of cluster categories, helping to better understand the cluster structure of the sample. By analyzing the distance between the sample and the center point through the coordinates of the cluster center point, and drawing a cluster scatter plot, the clustering effect is visually displayed, which helps to observe and understand the differences between different categories.

5.2 Weak point

In cluster analysis, the selection of the initial cluster center may have an impact on the results and requires multiple runs to avoid falling into local optima. The results of the random forest classification algorithm are based on data from training and testing sets, which have certain limitations and sample specificity. In practical applications, attention should be paid to the generalization performance of the model. The analysis results are only for reference, and the conclusion still needs to be judged based on comprehensive domain knowledge and actual situation.

References

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version1.0.11) [Online Application Software]. Retrieved from https://www.spsspro.com.
[2] Xu Weichao. Overview of Research on Correlation Coefficients [J]. Journal of Guangdong University of Technology, 2012, 29 (3): 12-17
[3] Zhou Zhihua. Machine Learning [M]. Tsinghua University Press, 2016
[4] Saroj, Kavita. Review: study on simple k mean and modified K mean clustering technique [J]. International Journal of Computer Science Engineering and Technology, 2016, 6(7):279- 281.