Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Implementation of Random Forest Algorithm Based Medal Count Prediction

Shuai Chen [1, *], Junjie Wang [2] and Yuhang Wu [1]

[1] School of Artificial Intelligence, North China University of Science and Technology, Tangshan, China

[2] College of Mechanical Engineering, North China University of Science and Technology, Tangshan, China

* Corresponding Author Email: chenshuai@stu.ncst.edu.cn

Abstract. This paper proposes a data trend prediction model based on random forest regression, focusing on the comprehensive use of machine learning and association rule mining in complex data prediction. Firstly, considering various factors, we constructed a medal prediction model based on random forest regression, assessed the accuracy of data, predicted the medal list of the 2028 Olympic Games and the first medal prediction results of the non-winning countries, and at the same time, we used the Apriori algorithm to analyze the relationship between the setting of the competition events and the distribution of the medals. Then, a regression model was constructed with "coaching effect" as the main variable, and the coefficients of "coaching effect" were judged according to the contribution degree of the coefficients. By applying the MK mutation test, we select and evaluate the changes in the number of medals of the three countries after investing in "great coaches". The model analyzes and validates the impact of great coaches on sports performance through a predictive model with excellent accuracy assessment obtained from Random Forest Regression.

Keywords: Random Forest Regression Algorithm; Apriori algorithm; Mann- Kendall method; predictive model.

1. Introduction

Focusing on data trend prediction with integrated learning models, this paper aims to explore the complex data mining prediction to find associations [1]. Firstly, a prediction model is built based on random forest regression to obtain detailed prediction results, and then the Apriori algorithm is introduced to mine the association between each item and medals in the centralized data [2]. Then, in order to get the coefficient of "coaching effect", the regression model is constructed with "coaching effect" as the main variable, and the statistical test is carried out by applying the method of MK mutation test to determine the size of its influence [3]. The experimental results show that the predictive model obtained by using the random forest regression algorithm with excellent accuracy assessment can analyze and verify the influence of excellent coaches on sports performance.

2. Modeling and Solving the Medal Count Prediction Model

2.1. Establishment of Prediction Model

In order to build a prediction model for the number of medals for each country, including three types of medals and the total number of medals. First, some factors are selected as predictor variables, including the proportion of winning athletes, the number of athletes participating in the Olympic Games, historical medal data, the number of sports events, and the host country variable host is added, with 1 marking that a country is the host country, otherwise 0.

By using the random forest algorithm to construct this multivariate prediction model, it can effectively deal with nonlinear feature interactions, be robust to noise, and also provide feature importance assessment for easy interpretation of the model.

The prediction formula for random forest is:

$$\hat{Y} = \frac{1}{N}\sum_{n=1}^{N} f_n(X) \tag{1}$$

Where:

N denotes the number of trees in the forest;

$f_n(X)$ is the predicted value of the nth tree.

Hyperparameter optimization optimizes the following parameters:

$N\_estimators$, Max_depth, $Min\_samples\_split$ and $Min\_samples\_leaf$.

The model equation for a random forest is:

The final model inputs include: $Year$, $Athletes$, $host$, $Goldlag$, $Totallag$, and the final output Gold is the prediction:

$$Gold_{i,t} = f(Year_{i,t}, Athletes_{i,t}, host_{i,t}, Gold_lag_{i,t}, Total_lag_{i,t}) \tag{2}$$

2.2. Predictive Model Training & Model Accuracy Analysis

2.2.1. Model training.

In order to evaluate the accuracy of the model, this paper divides the dataset into a training set (90%) and a test set (10%), and uses the optimized random forest model for training and prediction. The results of both training sessions showed high accuracy and good generalization ability in predicting the total number of gold medals and the total number of medals.

2.2.2. Model evaluation.

Comparison of prediction results for the training data of the model test set, as shown in Fig. 1, the $RMSE$ of the gold medal model and medal model test set are 1.2327 and 2.5939, respectively, and the smaller the data indicate that the prediction error of the model is smaller, and the model selection and optimization are reasonable. The R^2 of the gold medal model and medal model test set are 0.94969 and 0.96904 respectively, indicating that the model can explain about 95% of the changes in the number of gold medals and about 97% of the number of medals, and generates a small error in the results, and the model prediction performance is good [5].

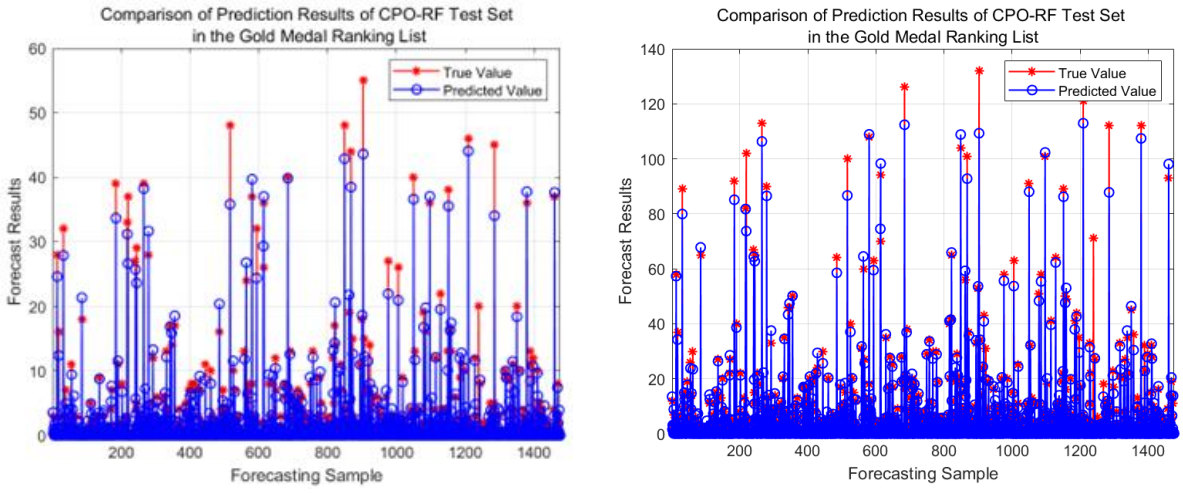

Figure 1. Comparison of predicted results for gold and medal table test sets

2.3. Medal Table Predictions and Country Performance Analysis for the 2028 Summer Olympics in Los Angeles

A trained random forest model was used to predict the medal distribution of the 2028 Los Angeles Olympic Games, with influencing factors including host country effect, proportion of winning athletes, number of athletes participating in the Olympics, historical medal data, and number of sporting events, and the predicted results are represented in Fig. 2:

NOC	Country	Gold	Silver	Bronze	Total
United States		39	33	38	110
China		37	30	24	91
Great Britain		17	5	39	61
France		16	3	35	54
Australia		18	17	16	51
Japan		20	7	21	48
Italy		13	17	8	38
Germany		13	2	19	34
Netherlands		14	8	12	34
Cannda		9	1	15	25
Brazil		4	4	13	20
New Zealand		9	7	4	20
Hungary		7	10	2	19
Spain		5	4	9	18
South Korea		5	4	2	11

Figure 2. 2028 Los Angeles Olympic medal table

Based on the results of the projected 2028 medal table compared to the final medal table of the last Olympics it can be seen that the United States and China are expected to continue to lead the medal table, with Great Britain and France likely to get a boost, and Japan, Australia, and the Netherlands likely to see their medal situation deteriorate.

Prediction intervals: In order to estimate the accuracy and uncertainty of the prediction results, this paper uses the Bootstrap method for the computation of prediction intervals in the following steps.

Perform putative sampling from the training set to generate multiple subsets, train the model on each subset, and make predictions on the 2028 data.

Calculate prediction intervals for each country. For example, a 95% confidence interval can be calculated using the following formula.

$$CI_{95\%} = [\hat{y} - 1.96 * SE, \hat{y} + 1.96 * SE] \tag{3}$$

Where $\hat{y}$ denotes the predicted value of the model;

CI denotes the confidence van of the estimates of the overall parameters;

SE denotes the standard deviation of the sample mean, which is used to estimate the precision of the overall mean.

Changes in the national gold medal table: The Fig. 3 shows the trend of changes through arrows and percentages, reflecting the dynamics of development and competition in the field of sports, and it can be clearly seen that compared with the data in 2024, according to the model solved in this paper, the

three countries of Cuba, Ukraine and Great Britain are the most likely to improve their performance, while China, Uzbekistan and South Korea may be relatively poor compared with the data in 2024. Uzbekistan and South Korea are the two countries that are most likely to improve, while China, Uzbekistan and South Korea are likely to do relatively poorly in comparison to the 2024 gold medal table data.

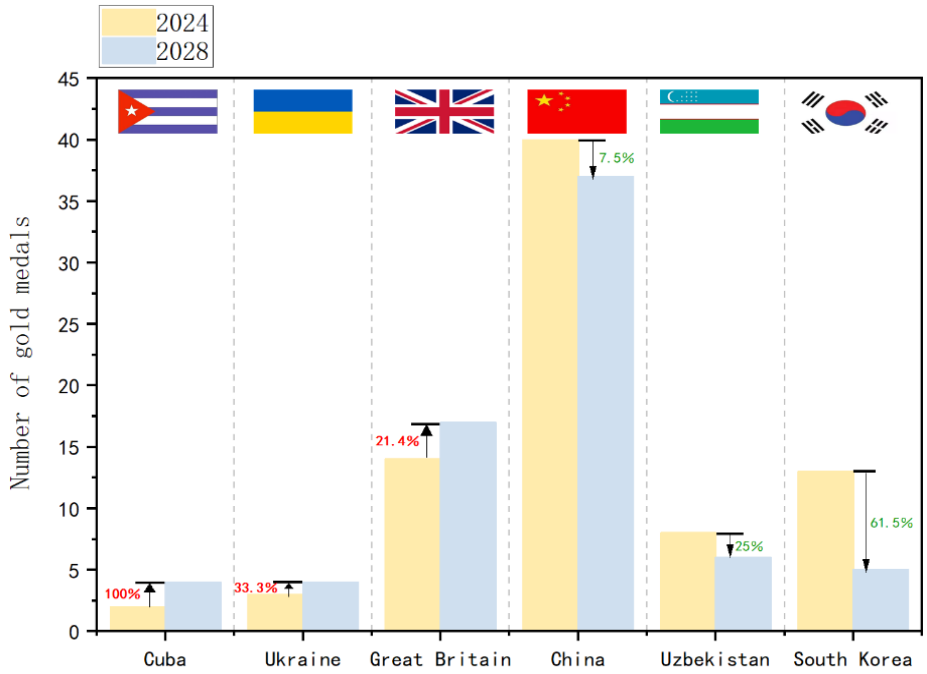

Figure 3. Changes in the national gold medal table where rankings may change

2.4. Prediction and Probability of First Olympic Medal for Country with Medal Deficit

The model also applies to countries that have not yet won a medal. According to the Fig. 4, taking into account the factor variables, it can be expected that at the next Olympic Games Kyrgyzstan, Malaysia, Albania and Cape Verde are most likely to win their first medal, with probabilities of about 80%, 70%, 30% and 20%, respectively.

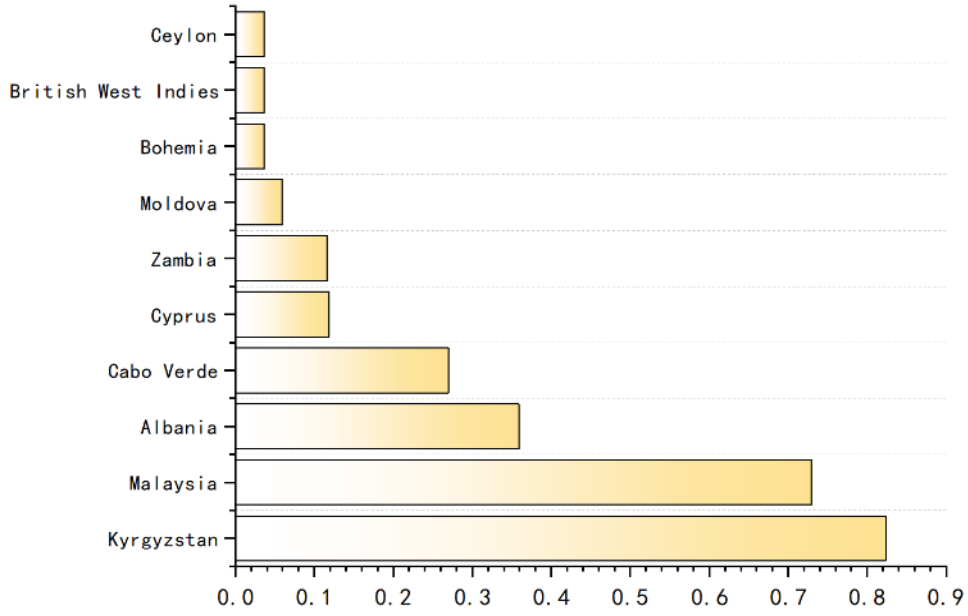

Figure 4. Predicted rate of the country's first gold medal

2.5. Modeling and Solving for the Relationship Between Events and Number of Medals

2.5.1. Model building.

The number of gold medals and the number of medals won by each country in each event at each Olympic Games, the total number of people in each event at each Olympic Games by country were extracted from the data provided.

Analyzing method:

1. Using Apriori method: the countries with the top medal counts are selected as an example and analyzed using the Apriori algorithm to generate frequent itemsets, discover frequent itemsets and association rules in the dataset, so as to find out the correlation between the items and the number of medals won by each country. The specific steps are as follows:

(1) Generate candidate item sets to determine the minimum support;

(2) Calculate the support and filter the frequent itemsets to calculate the candidate itemset support:

$$\text{support}(X) = \frac{\sigma}{\text{Total number of services}} \tag{4}$$

Where σ denotes Number of transactions containing itemset X.

(3) Generate association rules;

(4) Calculate the support and filter the frequent itemsets to calculate the candidate itemset support:

$$confidence(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)} \tag{5}$$

2. In order to measure the linear relationship between the different events and the number of medals won by countries. Pearson correlation coefficients were used.

2.5.2. Model solving and analysis.

Athletics, badminton, table tennis, and tennis were found to be highly correlated with the number of medals through the Apriori method, and the results were visualized in a heat matrix plot based on the correlation coefficients:

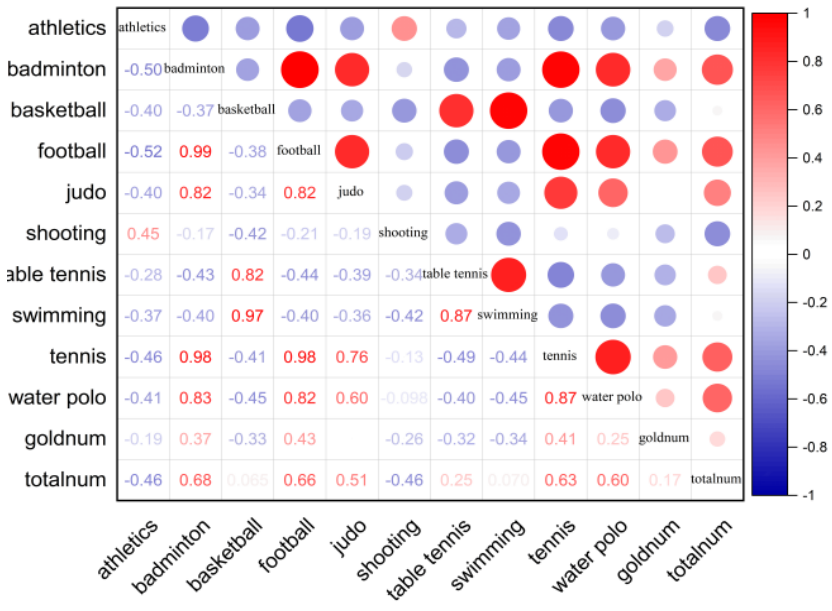

Figure 5. Heat map of Pearson correlation coefficient

From these Fig. 5, it can be inferred that sports with a higher correlation with the number of gold medals and the total number of medals are more important to individual countries.

When a country is the host country, it may invest more resources in the dominant sports and the athletes may be supported by the home crowd when competing at home, leading to better results in these sports.

3. Modeling and Solving for the "Great Coach" Effect

3.1. Analyzing the Effect of the "Good Coach" Effect on the Distribution of Medals

In order to analyze the effect of the "good coach" effect on the number of medals, a mathematical model using regression analysis is needed to quantify the effect of the coaching effect on the medals won by the athletes.

3.2. Selection of Data Samples and Control Variables

Dependent variable: the number of medals and the number of gold medals selected.

Main independent variable: a binary variable indicating whether or not a country has a "great coach" in a given year. 1 means that there is a "great coach" and 0 means that there is not.

Control variables: including type of sport, number of athletes, host country effects, etc.

3.3. Regression Model Setup

$$Y_{t,j} = \beta_0 + \beta_1 X_{i,j} + \beta_2 Z_t + \varepsilon_{t,j} \tag{6}$$

Where:

$Y_{i,j}$: The number of medals won by the i country at the j Olympic Games.

$X_{i,j}$: A binary variable for whether the country was coached by a good coach, where 1 means coached and 0 means not coached.

$\mathrm{Z_i}$: Control variables, usually including the number of events, type of events, host country effect, etc.

β_0 : Constant term.

β_1 : Regression coefficient of coaching effect, indicating the contribution of "excellent coaching effect" to the number of medals.

$\varepsilon_{\mathrm{i,\ j}}$: Error term, which indicates the unexplained part of the model.

3.4. Model Solving

β is significant and positive, indicating that the coaching effect has a positive impact on the number of medals.

The results of the solution were visualized as shown in Fig. 6:

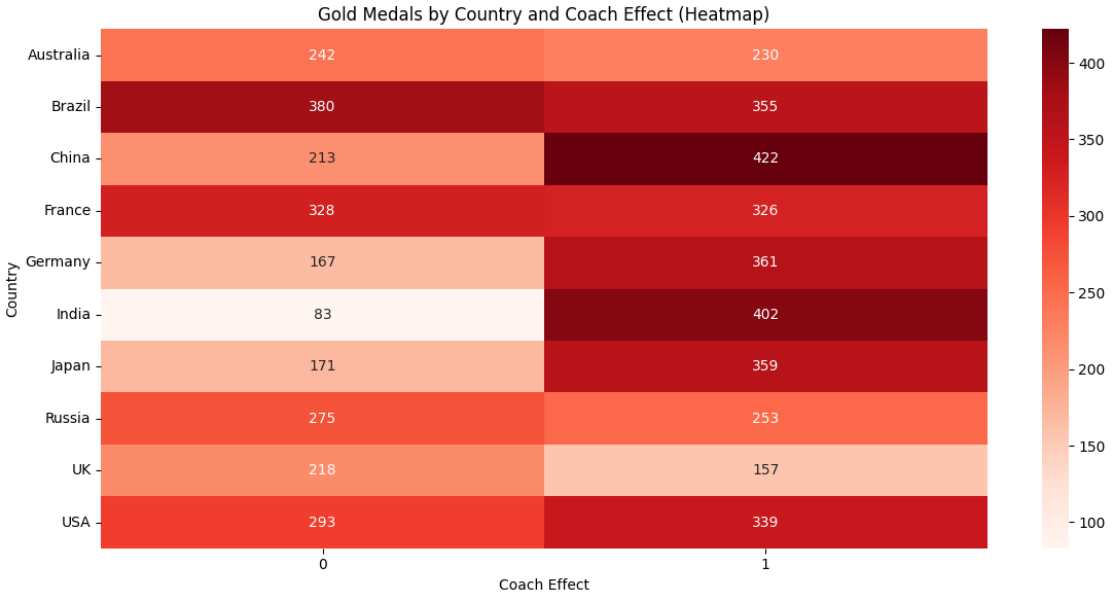

Figure 6. Heat map of the coaching effect

Fig. 6 demonstrates the relationship between the number of gold medals and coaching influence for ten countries. In terms of coaching influence, China and India are extremely affected by the positive influence of coaching, France is almost unchanged and the influence of coaching is not significant; the number of gold medals of Britain decreases when the influence of coaching is one. These differences may be due to the differences in training systems and athletes' basic qualities among countries.

3.5. Selection of Countries with Potential

This paper selects the United States China as well as Japan for analysis

1. Data extraction: This paper chooses volleyball, table tennis and swimming of three countries as examples, and selects the medal data of the last ten Olympic Games according to the coaching time period of the “great coaches”, and divides the time series into three phases: before coaching, during coaching and after coaching.

2. Conduct MK mutation test for these three items to find out the possible mutation points. If the mutation point appears during the coaching period of the “Great Coach”, it indicates that the “Great Coach” effect may exist.

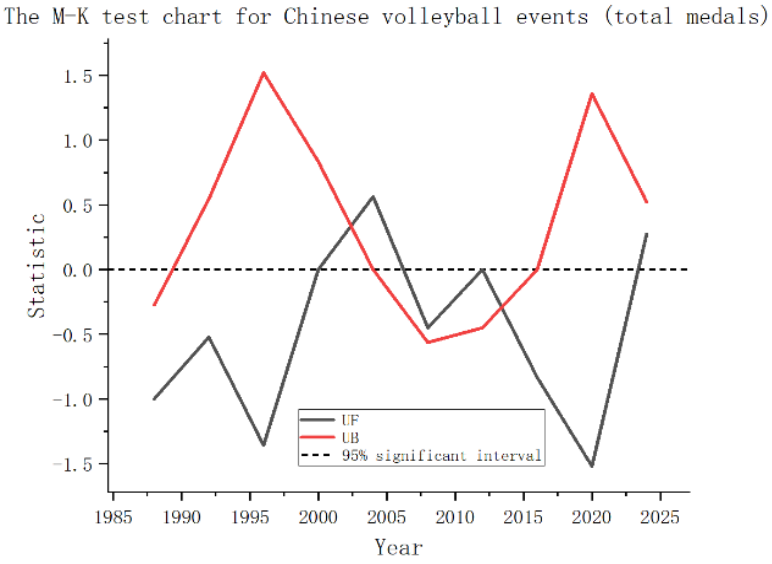

Figure 7. China Volleyball M-K Test Chart

Comprehensively analyzing the M-K test charts of each country's program, it is found that the UF and UB curves of each country's program show certain characteristics. Specifically, according to Fig. 7, the UF and UB curves of Chinese volleyball program fluctuate a lot and cross the 95% significant interval many times, so it is suitable for the introduction of “great coaches”. The UF curve of American table tennis program is above the 95% significant interval in some periods, which is also suitable for the introduction of “great coaches”. The UF curve of Japan's swimming program is also above the 95% significant interval in some periods, which is also suitable for the introduction of a “great coach”.

4. Conclusion

The Random Forest regression-based construction of data trend prediction model proposed in this paper plays a key role in complex data prediction. The experimental results show that by testing its gold medal model and medal model, it $RMSE$ is 1.2327 and 2.5939, and its R^2 is 0.94969 and 0.96904, indicating that the model prediction performs well. Firstly, based on the random forest regression of predicting the medal situation of each country, Apriori algorithm was applied to mine more factors affecting the performance. Then, in the face of the brand-new main variables to construct the regression model, the MK mutation test was applied to assess the influence size of the new main variables. In the future, the influence of more dynamic factors, such as the psychological state of athletes and the innovation of training methods, can be explored in depth. By continuously absorbing new knowledge and information, the model will be able to evolve and reach a more accurate and reasonable level of prediction [6].

References

[1] Nagpal, P., Gupta, K., Verma, Y., & Kirar, J. S. (2023, January). Paris Olympic (2024) Medal Tally Prediction. In International Conference on Data Management, Analytics & Innovation (pp. 249-267). Singapore: Springer Nature Singapore.

[2] Lina Lu, Yaping Chen, Heng-Yi Wei, Mai-Shun Yang. (2000). A study of Apriori algorithm in mining association rules. Small Microcomputer Systems (09), 940 - 943.

[3] Vagenas, G., & Vlach Kyriakou, E. (2012). Olympic medals and demo-economic factors: Novel predictors, the ex-host effect, the exact role of team size, and the “population-GDP” model revisited. Sport Management Review, 15 (2), 211 - 217.

[4] Yi-Sen Wang & Shu-Tao Xia. (2018). A review of random forest algorithms for integrated learning. Information and Communication Technology (01), 49 - 55.

[5] Eleftherios Kouloumpris & Ioannis Vlahavas. (2025). Markowitz random forest: Weighting classification and regression trees with modern portfolio theory. Neurocomputing129191 - 129191.

[6] Zhao, Xin, Xue, Ye & Niu, Chonghuai. (2013). Analysis of the correlation between the total number of Olympic medals and the total GDP of each country. Journal of Sports Culture (08), 1 - 4.