Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 1, 2025

Machine Learning-Based Prediction of Olympic Medals and The Exploration of National Sports Expertise

Yuyang Gao *

International Business College, South China Normal University, Guangzhou, 528225, China

* **Corresponding author**: 20233638007@m.scnu.edu.cn

Abstract: This study focuses on research related to the Olympic Games, employing a combination of Analytical Hierarchy Process (AHP), Random Forest Regression, and K-means Clustering techniques to analyze athletes' abilities, medal outcomes of various countries, and their sporting specializations. First, an AHP-based athlete performance evaluation model is constructed, where athletes are scored based on total medal count, gold medal count, number of events participated in, and event diversity. Second, a Random Forest Regression prediction model is developed, which, using historical medal data of countries and athletes' scores, forecasts the medal outcomes for different countries in the 2028 Olympic Games. The model demonstrates strong predictive accuracy. Finally, K-means Clustering is applied to categorize countries, exploring the sports in which different clusters of countries excel. The findings of this research provide scientific evidence for formulating national sports development strategies and analyzing the relationship between sports and various factors, thus contributing to enhancing national sports competitiveness and advancing the development of sports.

Keywords: Olympic Medal Predictions; Random Forest Regression Model; K-means Clustering; Analytic Hierarchy Process.

1. Introduction

The Olympic Games, as the world's premier sporting event, have always attracted significant attention regarding the distribution of medals and the athletic specialties of various countries. Numerous scholars are dedicated to researching it from different perspectives to reveal the patterns of sports development and provide guidance for the sports undertakings of various nations. Wang Fang utilized neural network methods to explore the patterns presented by Olympic medals and the related experiences from the perspective of national GDP [1]; Xie Xiaopeng utilized a gray prediction model to conduct an in-depth analysis and processing of the results from previous Olympic Games, and made predictions regarding the number of medals for the 2020 Olympic Games [2].

The existing research still has room for improvement in the comprehensiveness of athlete capability assessment, the accuracy of medal predictions, and the depth of national sports expertise analysis.

In recent years, with the development of machine learning, the Analytic Hierarchy Process (AHP), Random Forest Regression, and K-means Clustering have been widely applied across multiple fields. AHP can effectively address complex decision-making problems involving multiple factors, providing a scientific framework for athlete capability assessment. Random Forest Regression excels in handling high-dimensional data and nonlinear relationships, demonstrating significant advantages in predictive tasks. K-means Clustering can uncover the inherent structure of data, enabling classification analysis of national sports expertise. Fu Chaoqun and others pointed out that the random forest regression algorithm exhibits good robustness against noisy data and missing outliers. Furthermore, the random forest regression algorithm reduces reliance on data preprocessing during the training phase, with a standardized implementation process and high learning efficiency [3]; Zhang Bowen pointed out that since its inception, the K-means clustering algorithm has rapidly become one of the most popular and widely used methods in clustering research due to its simplicity of thought, good clustering effect, ease of implementation, and strong interpretability. Its application areas cover business intelligence, pattern recognition, and natural language processing [4]; Zhang Dabin stated that the Analytic Hierarchy Process is a powerful decision-making tool that skillfully combines qualitative and quantitative analysis, effectively addressing complex and difficult-to-quantify evaluation issues [5]. In summary, there has been limited research applying machine learning to medal prediction for the Olympics and the analysis of sports expertise across different countries.

Therefore, this paper innovatively integrates the three machine learning methods to conduct a more in-depth analysis of issues related to Olympic predictions. By constructing models for athlete capability scoring, medal prediction, and national sports expertise classification, it aims to achieve an analysis of athlete abilities, medal predictions for various countries, and sports expertise analysis. This provides a scientific basis for countries to formulate scientifically sound sports development strategies while enriching the research outcomes related to the Olympics.

2. Materials and Methods

2.1. Data Acquisition and Preprocessing

The data used in this paper is sourced from an open-source website (https://www.comap.com/contests/mcm-icm). This paper collects the award-winning information of all countries participating in each Olympic Games from 1896 to 2024 across various sports, as well as the award-winning information of different athletes in their respective sports. Additionally, it includes economic indicator data from various countries. During the data collection phase, careful screening and organization of the data were conducted, eliminating outliers and supplementing missing values, ensuring the completeness and accuracy of the data, thereby laying a solid

foundation for subsequent data analysis. This paper decides to apply the method of maximum-minimum normalization to normalize the collected data, in order to facilitate subsequent calculations and the establishment of models. The specific formulas used are as follows:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

After performing max-min normalization, all data has been scaled proportionally to the range of [0, 1], eliminating the influence of dimensions and facilitating subsequent comparisons.

2.2. Introduction to Research Methods

(1) Rating of an athlete's ability: for the scoring of athletes' abilities, this paper has decided to use the Analytic Hierarchy Process. This paper combines existing theories and research to determine the corresponding characteristics and their relative importance ratios, and based on these ratios, through a series of calculations, derives the weights of each characteristic. Finally, using the normalized data and weights, the scores of each athlete are calculated through a formula.

(2) Prediction of medals: for the prediction of medals, this paper decides to adopt a random forest regression prediction model, using the total number of medals and the total counts of gold, silver, and bronze medals obtained by each country from 1896 to 2024, as well as the total and average scores of athletes from each country as input features, ultimately deriving the predicted values of total medals and total gold medals for each country in 2028.

(3) Categorical analysis of sports expertise in different countries: For the classification analysis of national sports expertise, this paper decides to adopt the K-means clustering method, using the total number of medals historically won in different sports by various countries as the input value for clustering, and setting the K value to 5, which means categorizing all countries into 5 groups, with different categories having different sports expertise.

3. Athlete Ability Scoring Model Based on Analytic Hierarchy Process

3.1. The Establishment of Analytic Hierarchy Process

This paper uses all the medals and gold medals won by an athlete in the history of participation, as well as the number of entries and the diversity of participation (the number of different events participated) as factors in judging the athlete's ability. The number of medals can intuitively reflect an athlete's ability, while the number of gold medals reflects whether the athlete is at the top of the event, and the number of entries is used as an indicator to judge the proficiency of athletes, and the diversity of participation is also one of the important indicators to judge the ability of athletes. The paper decided to build a hierarchical model, consisting of three layers: objective layer, Criterion layer and Alternative layer. Factors above are part of the criterion layer, the content of the Alternative layer consists of all athletes appearing in the attached data, and the objective layer is for rating the athletes' abilities.

The judgment matrix, in the analytic hierarchy process, is used to determine the relative importance of all factors through pairwise comparisons, with the importance represented by numbers from 1 to 9, where 1 indicates equal importance and 9 indicates that one factor is extremely important relative to another (the highest level of importance), as shown in Table 1.

Table 1. Judgment matrix

Scale	Meaning
1	Indicates that the two factors are equally important
3	Indicates that one factor is slightly more important than the other compared to two factors
5	Indicates that one factor is significantly more important than the other compared to two factors
7	Indicates that one factor is more strongly important than the other compared to two factors
9	Indicates that one factor is more important than the other compared to two factors
2, 4, 6, 8	The median of the above two adjacent judgments

3.2. The Solution of Athlete Ability Scoring Model

Based on Table 2, the paper use three formulas to calculate the results of the weights of each factors, which are:

Table 2. Importance numbers between different factors

	Medal	Gold	Count	Diversity
Medal	1	2	4	7
Gold	1/2	1	5	6
Count	1/4	1/5	1	3
Diversity	1/7	1/6	1/3	1

(1) This formulas is used to calculate the maximum eigenvalue $\lambda\max$, A is the judgment matrix (Table 2), λ is the root to be solved and I is the identity matrix.

(2) This formulas is used to calculate Consistency Index CI.

(3) This formulas is used to calculate Consistency Ratio CR, RI is Random Consistency Index.

$$\det(A - \lambda I) = 0 \tag{2}$$

$$CI = \frac{\lambda\max - n}{n - 1} \tag{3}$$

$$CR = \frac{CI}{RI} \tag{4}$$

Through a series of calculations, the Approximate result of CI is 0.048, CR is 0.043, Since CR < 0.10, the consistency of the judgment matrix A is acceptable. The paper uses the arithmetic average method, the geometric average method, and the eigenvalue method to obtain the weights respectively, and then evaluate the weights of these weights on average, and finally obtain the weights of different factors as: Medal=0.483, Gold=0.351, Count=0.113, Diversity=0.053. Finally, the paper work out the scores of nearly 130,000 athletes.

4. Medal Prediction Model Based on Random Forest Regression

4.1. Establishment and Solution of the Medal Prediction Model Based on Random Forest Regression

The features this paper choose to establish the model are the numbers of medals, gold medals, silver medals and bronze medals a country won in the Olympic game from 1896 to 2024, and the mean and sum scores of all athletes of a country (scores are derived from Athlete ability scoring model).

When predicting the total number of medals, the medal count is used as the target variable, while other variables (excluding the medal count) are treated as feature variables. Additionally, categorical variables (countries) are one-hot encoded, converting them into numerical features suitable for machine learning models. The data is then split into a training set and a testing set, with 80% allocated for training and 20% for testing. This ensures the independence of model training and evaluation. The number of decision trees is set to 100, which strikes a good balance between computational efficiency and model stability, and the random seed is fixed at 42. The model is fitted on the training set, leveraging the ensemble of multiple decision trees to capture the complex relationships between features and the target variable. The final model achieves a mean squared error (MSE) of 2.81 and an R^2 score of 0.994, indicating that the Random Forest regression model demonstrates excellent predictive performance on the data. The top 10 countries in terms of predicted medal counts for the 2028 Olympics are shown in the Figure 1.

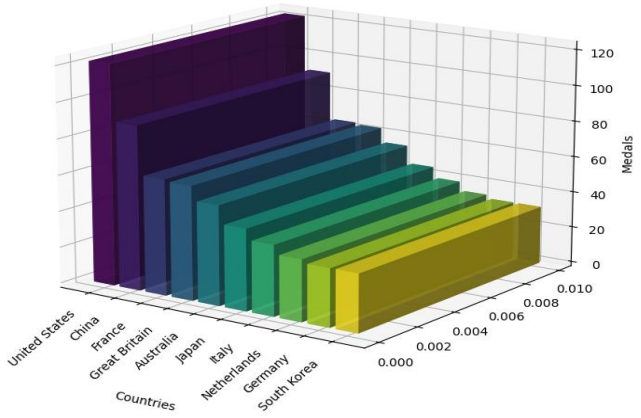

Figure 1. Top 10 Countries medal predictions

The remaining approximately 70 countries are not displayed due to the large volume of data.

In this prediction, some countries will show an improvement in their medal count, such as Uzbekistan, Kenya, Ukraine, United States, Great Britain, Germany, Italy, and Algeria. At the same time, some countries will experience a decline in their medal count, such as Albania, Cabo Verde, Saint Lucia, Singapore, India, and Australia.

The paper also made predictions for countries that have never won a medal in history, exploring whether any country could win its first medal in the 2028 Olympics. The paper first organized the data and then applied the model for prediction, and the results are shown in the Table 3.

Table 3. Prediciton of countries with its first medal in history

Country	Medal
Angola	1
El Salvador	1
Honduras	1
Laos	1
Malta	1
Monaco	1
Palestine	1
Trinidad & Tobago	1
U.S. Virgin Islands	1

In addition, the paper also made predictions for the number of gold medals each country will win in 2028, and the results are shown in the Figure 2.

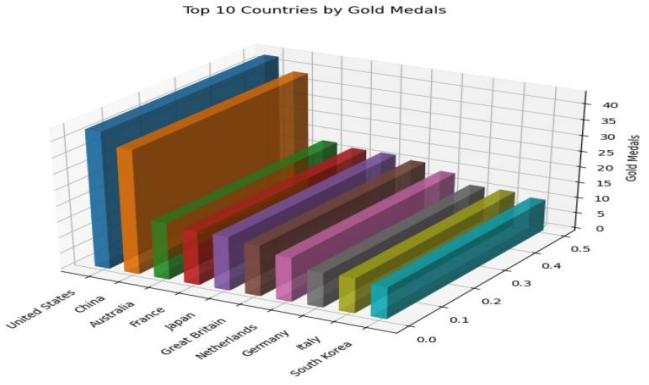

Figure 2. Gold medals prediction

4.2. Model Performance Evaluation and Prediction Accuracy Assessment

First, the paper decided to plot the learning curve of the model to assess its generalization ability, and the learning curve of the model is shown in the Figure 3.

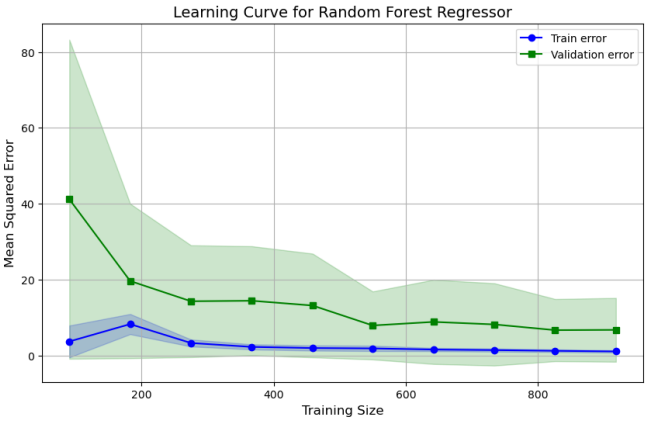

Figure 3. Learning curve

Figure 3 indicates that the model performs well on the training data, with the validation error gradually decreasing, suggesting that the random forest model is suitable for the current task and has good generalization capabilities.

In addition, the paper also plotted the relationship between mean squared error and the number of trees to assess the impact of the number of trees on model performance, as shown in Figure 4.

5. Country Expertise in Sports Classification Model Based on K-Means Clustering

5.1. Establishment of the Country Expertise in Sports Classification Model

As shown in Figure 4, the paper compiles and organizes data on the sports events in which different countries excel. Figure 4 only shows top 10 countries, the remaining approximately 70 countries are not displayed due to the large volume of data.

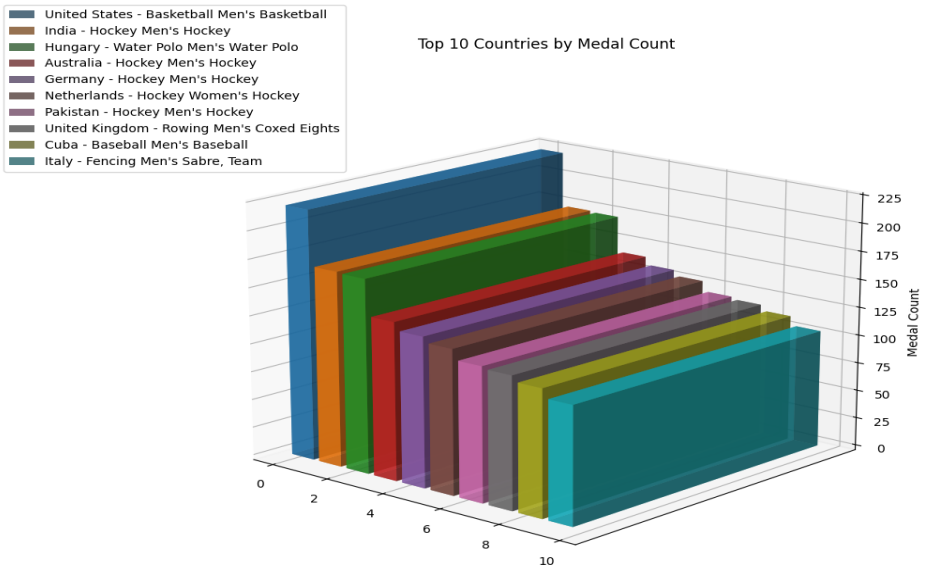

Figure 4. Countries and sport events rank

Then the paper decided to perform K-means clustering analysis on these countries to determine which Sports are important to which countries, because if we analyzed each country individually, the data volume would be too large and would severely slow down the computational efficiency. Therefore, we used K-means clustering to divide the countries into 5 categories and explored what Sports each of these five categories of countries excel in. This method can balance computational efficiency with the relationship between data complexity and data representativeness.

5.2. Solution of the Country Expertise in Sports Classification Model Based on K-Means Clustering

For this model, the first step is to select an appropriate K value for data classification, so the paper plotted the line graph of silhouette coefficients corresponding to different K values.

The closer the silhouette coefficient is to 1, the better the clustering effect. When K=2, the silhouette coefficient is the highest, indicating the best clustering effect. However, if K=2, the clustering becomes meaningless, as the number of categories is too small and each category lacks representativeness. Therefore, the paper decided to choose a K value of 5, dividing the countries into five categories to explore the sports that these categories of countries excel in. When K=5, the resulting clustering graph is shown in Figure 5.

After organizing the relevant data with code, the paper has concluded that the sports powerhouses represented by the United States (with an average cluster center value of 4.37 for all medals corresponding to all sports, the highest value) excel in 23 sports including wrestling, volleyball, table tennis, and others. The second-tier sports powers represented by the United Kingdom (with an average cluster center value of 3.18) do not have any particularly outstanding sports, but they have a high overall level. The third-tier sports powers represented by Germany, which consists of three countries (with an average cluster center value of 1.54), excel in Polo as a sport. The fourth-tier sports powers represented by China, which consists of ten countries (with an average cluster center value of 0.99), are proficient in 46 sports including swimming, football, gymnastics, and others. The fifth-tier sports powers represented by Afghanistan mainly excel in sports such as badminton and track and field. Each of these sports is important to their respective countries.

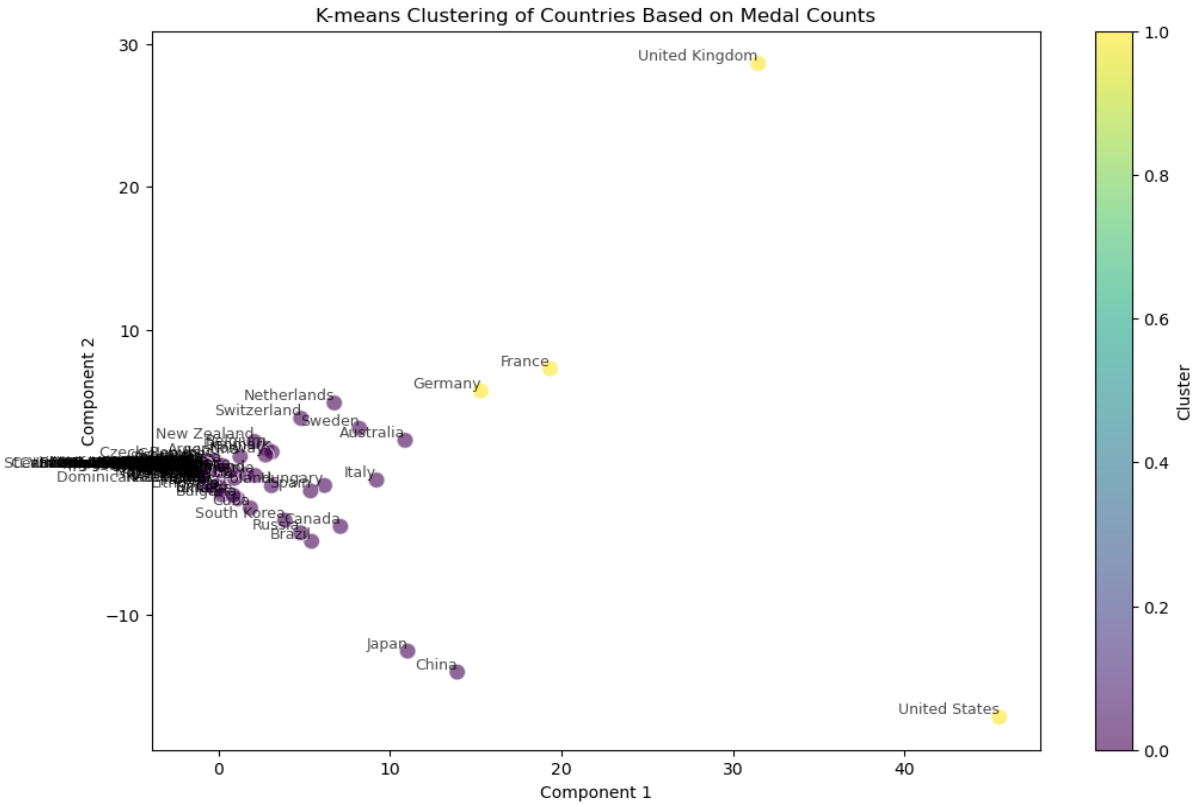

Figure 5. K-means clustering of countries Based on medal counts

6. Conclusion

The research establishes three core models: an athlete capability scoring model based on AHP, a medal prediction model using Random Forest Regression, and a national sports specialization classification model via K-means Clustering, providing actionable insights for nations to optimize resource allocation and tailor sports development strategies. While the paper offers valuable contributions, several avenues for improvement and extension remain. The future work will focus on expanding the dataset to include socioeconomic factors (e.g., GDP, population demographics, sports infrastructure investment) could enhance model robustness and uncover deeper correlations between external variables and athletic performance and exploring advanced machine learning techniques, such as deep learning or hybrid models, may further improve prediction accuracy, particularly for countries with limited historical data.

References

[1] Wang Fang. 2020 Olympic Games medal performance prediction based on neural network [J]. Statistics and Decision, 2019, 35(05): 89-91.

[2] Xie Xiaopeng. Research on prediction of Olympic results using grey prediction model [J]. Electronic World, 2018, (02): 48-49.

[3] Fu Qunchao, Zeng Hydro Detail, Chen Pei, et al. Drilling operation cycle prediction based on random forest regression model [J]. Well Logging Engineering, 2024, 35(04): 39-47.

[4] Zhang Bowen. Fast K-means clustering based on gridding and attribute weight matrix [D]. Sichuan Normal University, 2024.

[5] Yang Zhaosheng. Research on risk management of K domestic waste emergency landfill project based on analytic hierarchy process [D]. Guizhou University, 2024.