| Problem Chosen | 2025 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2512735 |

# Olympic Medal Tables: Based on Heterogeneous Stacked Ensemble Learning Model


## Abstract

The number and distribution of Olympic medals are influenced by a variety of factors, including a country's geographical location, economic level, population size, and national physical fitness. By building a mathematical model to explore the patterns influencing Olympic medal distribution, this study not only helps countries scientifically improve their medal performance in future Olympics but also contributes to the global spread and development of the Olympic spirit.

**Question 1** focuses on predicting the 2028 Los Angeles Olympics. Due to the high uncertainty of the Olympic events , data preprocessing involved statistical analysis of medal counts by country and event in each Olympic Games, serving as influencing factors for training the model. Through multiple experiments, it was found that traditional time series forecasting models do not adapt well to this problem. Thus, **the data was stacked over time** giving more weight to years closer to the event. **LangChain technology** was used to collect background information on athletes' retirement times, quantifying the impact of their past performances on current outcomes. In terms of model selection, this study used **a heterogeneous ensemble stacking model**, combining base models including **Ridge Regression**, **Lasso Regression**, **Decision Trees**, **XGBoost**, **Neural Networks**, and **Multi-linear Regression** with strong anti-overfitting capabilities. To address **the data imbalance** caused by a small number of winning countries, **over-sampling techniques** were introduced into the ensemble model. The model's performance, stability, and generalization ability were quantitatively evaluated using metrics like $R^2$ and MAPE, and the confidence interval for total medal counts was calculated using MSE and accuracy. **For the first and second subquestions**, the study performed a set difference between the predicted data and historical medal data provided in the problem and selected countries with a probability **greater than 55%** as potential first-time medalists. The evaluation of each country's performance was based on **ranking changes** and **medal count changes**. Five countries were predicted to win medals for the first time, with an odds ratio of 31.75, and the five countries with the most significant improvements and declines were identified. **For the third subquestion**, the study **assigned different weights** to the gold, silver, and bronze medals of each country and event to measure the importance of each event to the country.

**Question 2** focuses on the "Great Coach Effect." To find evidence for this effect, the study began by examining changes in the number of medals, defining **a mutation effect function f(x)** that quantifies the extent of medal mutations in a given year compared to an initial reference year. The function f(x) considers both **the total number of medals** and **the time interval**, revealing the mutation in medal performance. The coefficient m controls the intensity of the mutation effect. The model suggests that when the value of f(x) is large, the mutation is significant, and the likelihood of the "Great Coach Effect" influencing athlete performance is higher. For each country, the study calculates the projects and years when f(x) reaches its maximum, checking whether there was a coaching change in that year to further verify the Great Coach Effect. Specifically, for **China in football**, **Kenya in athletics**, and **Japan in table tennis**, the coaching effect may warrant investment consideration.

**Question 3** focuses on the model's insights into the underlying factors affecting Olympic medal counts. It discusses three aspects: **historical status**, **national economic level**, and **the cultural identity and bias towards sports events**.



**Keywords** : Heterogeneous stacking ensemble learning ; Base learner; Meta learner ; Olympic prediction ; Great Coach Effect

# Contents

**1 Introduction** **2**
1.1 Background . . . 2
1.2 Restatement of the problem . . . 2
1.3 Analysis of the problem . . . 2

**2 Model construction** **3**
2.1 Data mining . . . 3
2.1.1 Reason . . . 3
2.1.2 Approach . . . 4
2.2 Preprocessing dataset format settings . . . 4
2.2.1 Format Introduction . . . 4
2.2.2 Reasons for this format . . . 5
2.3 Model Assumptions and Rationalization Verification . . . 6
2.4 Stacking modeling . . . 6
2.4.1 Introducing Submodels . . . 6
2.4.2 Introducing the processing of ensemble imbalanced datasets . . . 8

**3 Solution of the problem** **14**
3.1 Problem I . . . 14
3.1.1 Problem i . . . 15
3.1.2 Problem ii . . . 17
3.1.3 Problem iii . . . 17
3.2 Problem II . . . 19
3.2.1 Abrupt Change Effect Calculation . . . 20
3.3 Problem III . . . 22

**4 Evaluation of the model** **23**
4.1 Model Adaptability . . . 23
4.2 Robustness . . . 23
4.3 Ability to Handle Imbalanced Datasets . . . 23

**5 Conclusions** **24**

# 1 Introduction

## 1.1 Background

At the 2024 Paris Olympics, the U.S. led with 126 medals, tied with China for golds (40 each). France ranked fourth overall, despite being fifth in golds (16). Smaller nations like Albania, Cape Verde, Dominica, and Saint Lucia won their first medals, with the latter two securing one gold each. However, over 60 countries remain medal-less. Medal distribution is shaped by factors such as geography, economy, population, and fitness. A mathematical model exploring these variables could help countries boost medal counts and promote the Olympic spirit.

## 1.2 Restatement of the problem

By building a mathematical model, solve the following problems:

- Problem I: ***Medal Prediction Model***

  i Develop a medal prediction model for each country, assess performance, and analyze accuracy.

  ii Predict the 2028 Los Angeles Olympics medal table, including prediction intervals, and identify countries with potential performance shifts.

  iii Estimate the number of first-time medal-winning countries in 2028 and their probabilities. Analyze how event types affect medal counts and identify key events for countries.

- Problem II: ***"Great Coach" Effect***

  Assess the "Great Coach" effect on performance and estimate its impact on medal counts. Recommend sports for three countries based on coaching investment, estimating potential outcomes.

- Problem III: ***Insights on Medal Distribution***

  Apply data mining to uncover patterns in medal distribution and offer actionable insights for national Olympic committees.

## 1.3 Analysis of the problem

- Problem I: ***Medal Prediction Model***

  i Identify factors like economy, population, and history. Build model using machine learning, evaluate with MSE and MAE, and analyze accuracy with historical data.

  ii Update model for 2028 prediction, use bootstrapping for intervals, and analyze trends to find performance changes.

  iii Estimate first-time medal countries with Bayesian models. Analyze event-type influence via grouping and statistical tests.

- Problem II: ***"Great Coach" Effect***

  Collect pre- and post-coaching data, use t-tests or non-parametric tests for evidence, and apply regression for medal impact.

Analyze countries' strengths, talent, and growth, and estimate coaching impact through performance comparison.

- Problem III: ***Insights on Medal Distribution***

  Use clustering and association rule mining to provide insights on sport prioritization and collaboration for national Olympic committees.

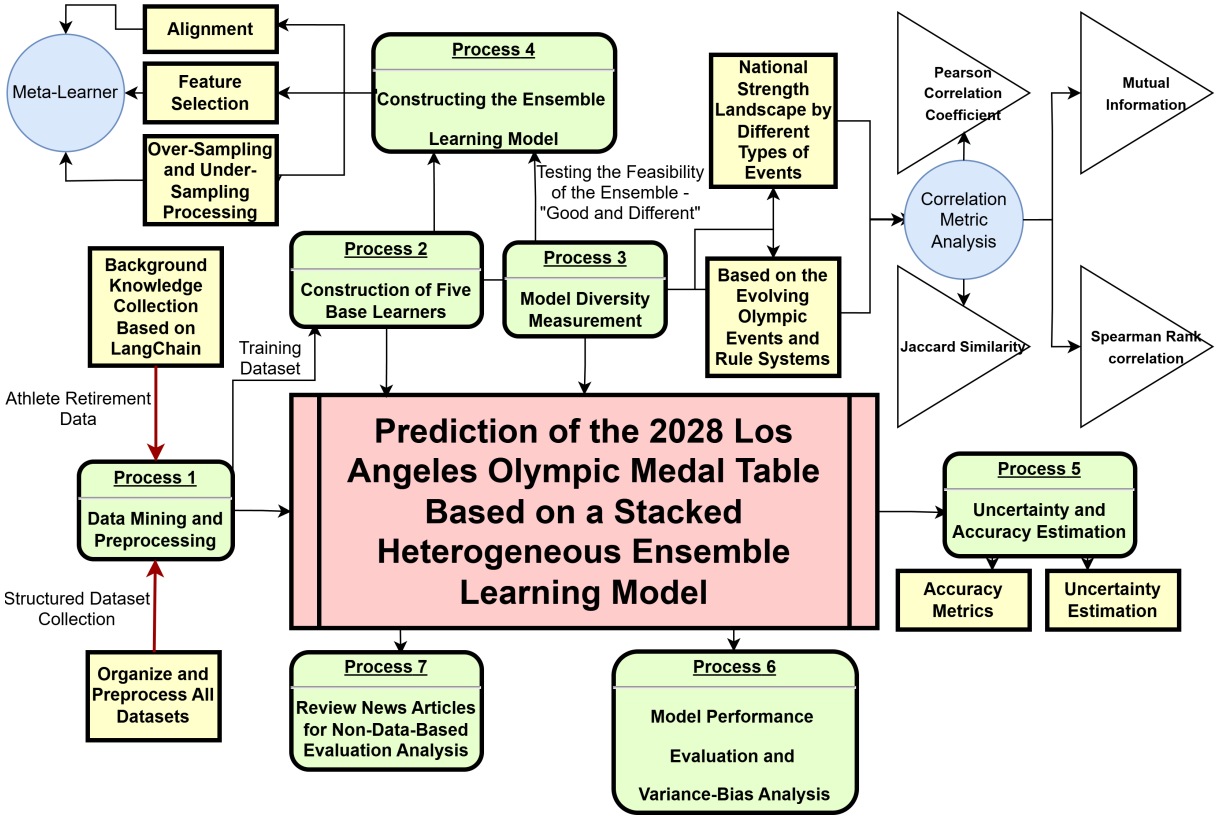



Figure 1: Flow chart

# 2 Model construction

## 2.1 Data mining

### 2.1.1 Reason

- Mining country-related data helps link economic factors and population size to sports investment and talent pool, both crucial for medal acquisition.
- Athlete and coach data can reveal the "great coach" effect, guiding resource allocation in coaching.
- Olympic project data identifies key events for medal success, helping countries select the most promising projects for competition.

### 2.1.2 Approach

**1. Langchain**
Create templated prompts and use a large language model to navigate the information of all athletes given in the data.
**2. Simplified Equivalence of the Olympic Games**
Traverse the table entries and consider athletes who have not appeared in the Olympics since then as retired
**3. Dataset equivalence processing**

In processing datasets, athletes who don't appear in consecutive Olympic Games are marked as retired, considering the limited duration of sports careers. If an athlete has no participation record in multiple Olympics, they are likely retired. The process involves checking each athlete's records. For every athlete, we review their participation years and verify if they appeared in any subsequent Olympics. If not, they are marked as retired.

## 2.2 Preprocessing dataset format settings

### 2.2.1 Format Introduction

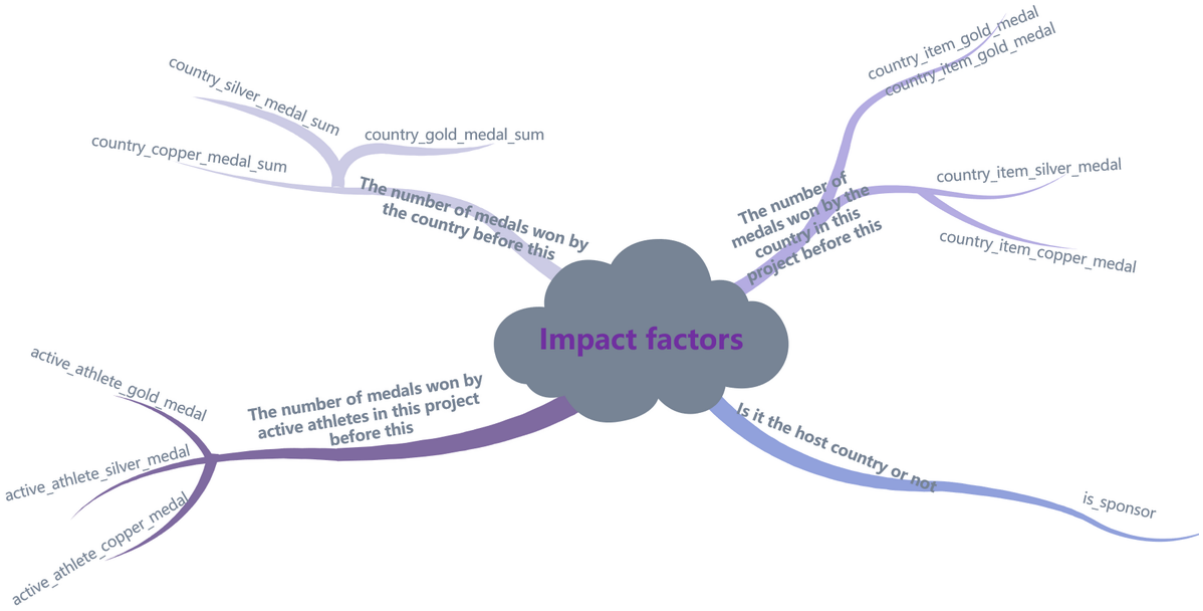



Figure 2: All influencing factors considered when setting up the data set

Considering that Germany-1 and Germany-2 both belong to Germany, firstly, we extract participating countries from the NOC column of ***summerOly_athletes.csv*** instead of the Team column by removing duplicates, and then extract all events from the Event column by removing duplicates. The country ID is represented by NOC, and the event ID defines a mapping relationship for a brief representation. Then, we make a Cartesian product of the country and the event to get the first column of the ID table. (As shown below) Through data processing, we set the basic unit as: ***country + specific***

***sub-event name***. This chart lists all the specific sub-events (Events) that China participated in under the major event (Sport) of Olympic diving.

### 2.2.2 Reasons for this format

**The actual meaning of each column in the dataset**

| ID | Description |
|---|---|
| CHN-DIV-M-3S | China Diving Men's Springboard |
| CHN-DIV-M-10P | China Diving Men's Platform |
| CHN-DIV-M-3SYNC | China Diving Men's Synchronized Platform |
| CHN-DIV-M-10SYNC | China Diving Men's Synchronized Platform |
| CHN-DIV-W-3S | China Diving Women's Springboard |
| CHN-DIV-W-10P | China Diving Women's Platform |
| CHN-DIV-W-3SYNC | China Diving Women's Synchronized Platform |
| CHN-DIV-W-10SYNC | China Diving Women's Synchronized Platform |

- ***Total medals:*** The country's overall sports strength and emphasis have some influence on the medals won in each event
- ***Medals won by active athletes in the country in this event:*** The strength of the national team in this event, with more detailed considerations and excluding the interference of retired athletes
- ***Medals won by the country in this event:*** Reflects the training environment of the country's athletes
- ***Whether the country is the host:*** Familiarity with the competition environment and the cost of adapting to local life

**Why use items?** Minor events provide detailed data, highlighting performance differences among countries in specific events. Focusing on these avoids generalizing strengths and weaknesses from major events, offering a clearer view of athletes' abilities. This method aligns with the official Olympic classification, reflecting the true structure of the Games.

**Why is the medal column presented in this way?**

1. This approach suits our regression, as we can rank countries based on previous data and predict medals accordingly. The design of the medal column aligns well with the solution model and fits the question.
2. No need to worry about predicted medals exceeding actual counts. Each file represents an event, and the medal column will only generate one gold, silver, and bronze per event, addressing the medal limit. This approach simplifies the program and improves efficiency.

## 2.3 Model Assumptions and Rationalization Verification

1. Data independence assumption: Samples in the dataset are independent, meaning medals won by different countries in the same year or by the same country in different years are independent. However, factors like changes in international sports policies or global event reforms may disrupt this independence.

2. Sample representativeness assumption: The training set is representative of the overall data, covering all relevant information of the Summer Olympics.

3. Error randomness assumption: The error between predicted and true values is due to uncontrollable random factors, not fixed model biases.

4. Threshold for a country winning two medals:

   - The dataset treats each country's small event as a single medal, missing cases where a country wins multiple medals in one event. Thresholds are designed to identify such cases (e.g., gold-silver, gold-bronze, or silver-bronze) based on performance gaps.
   - Threshold Design: A range of thresholds is set, prioritizing gold-silver combinations. If a country's score significantly exceeds second place and meets the threshold, two medals are awarded; otherwise, one is given. This accounts for data limitations and improves medal distribution accuracy.

5. Historical and future medals follow independent and identical distributions, meaning historical medals can predict future medal patterns.

6. Cases of shared medals are rare and can be ignored to avoid affecting the model.

## 2.4 Stacking modeling

"It uses a meta-learner and a base learner. The **base learner** is a **basic model** in an ensemble, while the **meta-learner** is the **second-layer model** that combines base learners' predictions, retraining them to produce the final result."

### 2.4.1 Introducing Submodels

**Selection of Base Learners**

**1.Lasso Regression**

**Principle**:In Lasso regression, the L1 regularization term $J(\theta) = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\theta_j|$ is introduced into the loss function for feature selection.

**Advantages**:It can effectively select key variables and simplify the model. Moreover, it creates a sparse model, reducing the computational complexity.

**Reasons**:The L1 regularization can shrink the coefficients of unimportant features to zero. It is suitable for feature selection from a medal dataset with 10 features, enhancing the accuracy and generalization ability of the model.

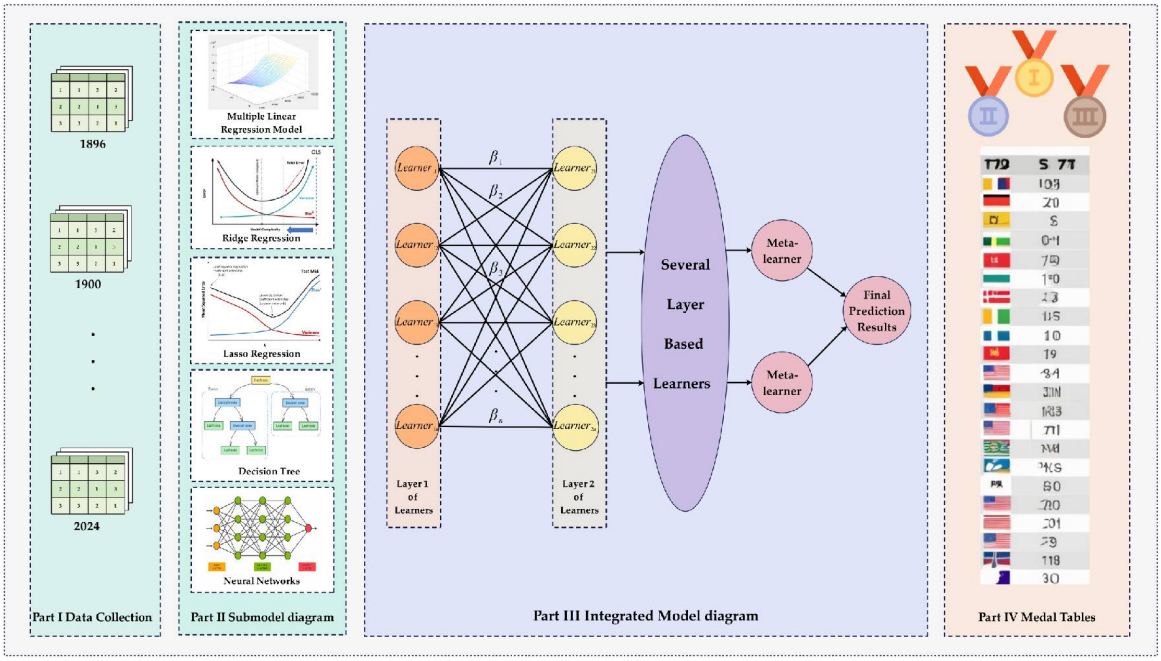



Figure 3: Integrated learning concept map

**2.Decision Tree Regression**

**Principle:** Decision tree regression selects the optimal feature and split point at each node to minimize mean-squared error. This process is recursively applied until stopping conditions are met. Predictions for leaf nodes are the average target values of training samples within them.

**Advantages:** It handles discrete data effectively, requires no assumptions about data distribution, and clearly shows the impact of features on medal outcomes.

**Reasons:** The medal dataset includes discrete features like medal counts and host country status, making decision trees well-suited for classification and providing strong interpretability.

**3.XGBoost Regression**

**Principle:** XGBoost is an ensemble algorithm based on gradient boosting that improves model performance by optimizing residuals of each tree. It adds regularization during training to prevent overfitting and enhances computational efficiency compared to GBDT.

**Advantages**:It can effectively handle discrete data, does not require assumptions about the data distribution, and has strong interpretability.

**Reasons**:It can effectively handle the discrete features in the medal dataset. The regularization can prevent overfitting, making it suitable for complex datasets.

**4.Neural Network Regression**

Principle The prediction is made through the formula:

$$\hat{y} = f_L\left(W_L \cdot f_{L-1}\left(W_{L-1} \cdot \ldots f_1\left(W_1 \cdot x + b_1\right) + b_2\right) + b_L\right)$$

. Advantages:It has strong fitting ability and robustness, can effectively handle large - scale datasets, and can integrate multi - dimensional information.

Reasons:The dataset contains rich information, and neural networks can fully utilize this information for prediction.

**Selection of Meta - Learners**

**Multiple Linear Regression** Principle The equation of the multiple linear regression model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$. The goal is to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing the mean - square error (MSE) between the predicted $\hat{y}$ and the actual value $y$.

Advantages:It can combine the prediction results of base learners to improve accuracy and stability. The results are interpretable and efficient.

Reasons:As a meta - learner in stacked regression, it can effectively integrate the prediction results of base learners, laying a strong foundation for further analysis.

**Heterogeneous Stacked Ensemble Learning**

**Introduction** By combining different types of base learners (Lasso regression, decision tree regression, XGBoost regression, neural network regression) and a multiple linear regression meta - learner, the advantages of each model are fully utilized to achieve more accurate and stable predictions.

**Optimization**

- GPU acceleration: It is used for neural networks to increase the number of layers and iterations.
- Parameter adjustment in XGBoost code:

  1. **scale_pos_weight**: Controls the ratio of weights for positive and negative samples. Increasing the weight of the positive class helps focus on the minority class, improving its classification.
  2. **max_depth**: Sets the maximum depth of the decision tree. Smaller values help prevent overfitting in imbalanced data, improving the model's generalization ability.
  3. **min_child_weight**: The minimum sum of sample weights in a child node. Larger values help avoid fitting noise or local patterns and stabilize learning in imbalanced data, especially for the minority class.
  4. **gamma**: Controls the minimum loss reduction required for a split. Larger gamma values make the model more conservative, reducing overfitting and effectively handling imbalanced data.

### 2.4.2 Introducing the processing of ensemble imbalanced datasets

**Definition of Imbalanced Data**

Imbalanced data occurs when there's a large disparity between class sizes, with the minority class being underrepresented. This affects algorithms' ability to detect minority class patterns, reducing

classification accuracy. For example, predicting 2028 Olympic medal outcomes requires handling imbalanced data, as winning medals is rare.

**Overview of Imbalanced Datasets**

In the context of medal achievements, most records correspond to athletes without medals, while only a small fraction achieve gold, silver, or bronze. This "few winners, many non-winners" pattern creates a typical case of data imbalance in the medal dataset.

**Solution**

***Oversampling Methods:*** Oversampling balances the dataset by increasing minority class samples through methods like random oversampling and SMOTE. SMOTE generates new samples by interpolating between k-nearest neighbors, boosting minority class instances and enhancing the model's generalization.

**Ensemble learning is to use multiple weak learners with large differences to integrate into a strong learner**

**Good but different** In ensemble learning, "good and different" refers to two key aspects: "good" means each base learner performs well, while "different" means base learners generate diverse predictions on the same data. Combining these improves the overall model's accuracy and robustness.

**Why choose stacking for heterogeneous ensemble learning?**

1. **Because ensemble learning is often more effective for real-life problems**

- ***Principle:***
  - Ensemble learning combines models (e.g., classifiers, regressors) to solve problems by aggregating their predictions, improving performance and reducing the risk of suboptimal models. Algorithms like Bagging, Boosting, and Stacked Generalization create diverse models and combine their strengths.
- ***Effectiveness:***
  - ***Improvement of Model Performance:***
    1. ***Variance and Bias Reduction:*** Ensemble learning reduces variance by bootstrap sampling and bias by correcting errors iteratively.
    2. ***Enhanced Generalization:*** By combining diverse base learners, ensemble learning adapts better to data patterns, improving generalization.
  - ***Addressing Data Challenges:***
    1. ***Handling Imbalanced Data:*** Ensemble methods address class imbalance with balanced subsampling or undersampling, focusing on minority classes.
    2. ***High-Dimensional Data:*** Random Forest reduces feature correlation in high-dimensional data, minimizing overfitting.

- ***Addressing Machine Learning Challenges:***
  1. ***Concept Drift:*** Ensemble methods use dynamic weighting and diversity-based approaches to handle concept drift in real-time applications.
  2. ***Expanding Search Space:*** By combining models, ensemble learning expands the search space for better data fitting, as seen in stock price prediction.

**2. The key to ensemble learning is "good but different"**

- ***The Importance of "Good and Different" in Ensemble Learning:***
  - Ensemble learning improves accuracy and generalization by combining diverse, well-performing base learners like ridge regression, Lasso, decision trees, and neural networks. This diversity avoids local optima and captures complex patterns, ensuring a more accurate analysis, e.g., in predicting Olympic medal counts.
- ***Why Heterogeneous Ensemble Learning More Easily Achieves "Good and Different":***
  1. ***Different Learning Algorithms:*** Heterogeneous ensemble learning combines algorithms like decision trees, neural networks, and SVM, each excelling in different tasks. This diversity ensures high performance and diversity by leveraging the unique strengths of each algorithm.
  2. ***Differences in Data Processing:*** Algorithms process data differently; neural networks need normalization, while decision trees don't. Some rely on dimensionality reduction, while others handle high-dimensional data directly, providing varied perspectives on the dataset.

**3. Stacking is extremely sensitive to data and considers minority labels, helping to address label imbalance.**

- ***Characteristics of the Stacking Method***
  1. ***Multilayer Structure:*** Combines multiple base learner predictions with a meta-learner to enhance generalization.
  2. ***Multiple Feature Consideration:*** Uses diverse algorithms and feature subsets to capture patterns, improving minority class recognition.
- ***Focus on Minority Class Labels***
  1. ***Data Fusion:*** Integrates predictions from base learners to offer a comprehensive view, improving minority class accuracy.
  2. ***Meta-Learner Optimization:*** Optimizes base learner combinations to focus on minority classes based on their distribution and characteristics.
- ***Empirical Evidence Support***
  1. ***Compared with Single Models:*** Balances learning between classes, reducing overfitting and enhancing generalization.

2. ***Compared with Other Ensemble Methods:*** Outperforms Bagging and Boosting in imbalanced data, improving performance without extra computational cost.

In summary, the Stacking method's multilayer structure and focus on data sensitivity make it effective for addressing imbalanced labels and improving predictions across diverse datasets.

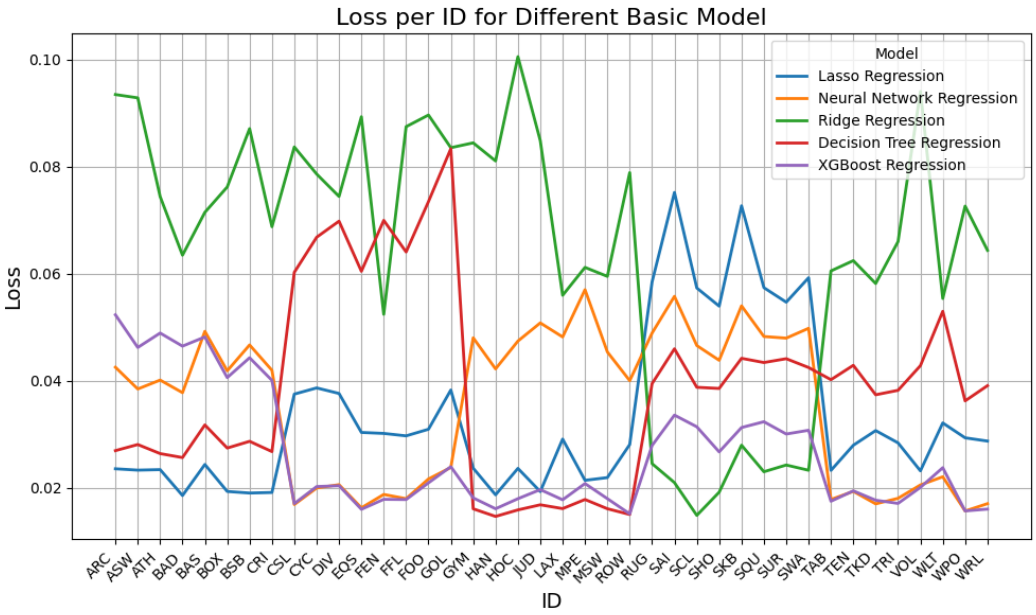



Figure 4: The performance of different models in different types of projects is different

**Segment interval accuracy comparison chart** There are differences in the performance of different models on various types of tasks, and multiple models can complement each other.

There are differences in the performance of different models on various types of years, and multiple models can complement each other.

**Diversity metrics** The four coefficients above show that the prediction results of each model are largely uncorrelated.

**Optimization of implementation and usage**

The model uses a multivariate linear regression meta-learner to combine predictions from base learners, forming a stacked regression model that improves accuracy on complex tasks. A two-layer structure is employed, where the first layer's predictions are input to the second layer, with the meta-learner making the final prediction. The model supports incremental learning, allowing updates with new data without retraining all models, and can incorporate new learners for added flexibility.

**Feature Selection**

Feature selection is a critical step in constructing ensemble heterogeneous learners. When using multiple different types of learning models in ensemble learning, feature selection helps to improve the efficiency and performance of the model.

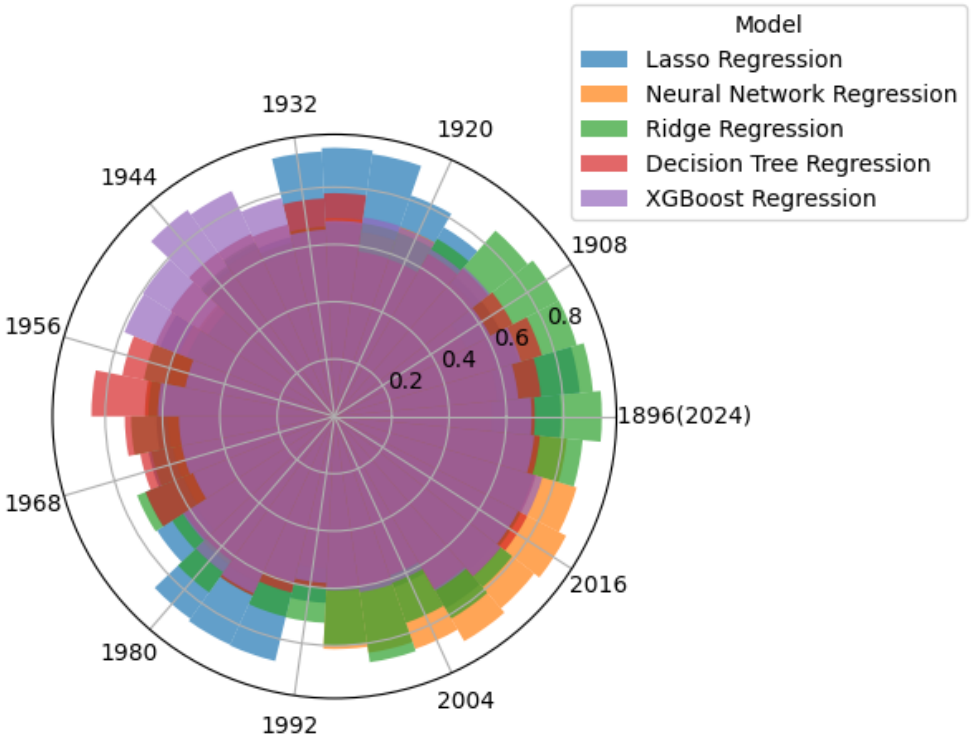



Figure 5: The performance of different models in different forecast years is different

### 1. Global Feature Selection

In some cases, a unified feature selection method can be applied to the entire dataset, and the selected features are then input into each base learner. This approach is suitable for cases where all models in the ensemble learning use the same features.

- **Common methods include:**
  Recursive Feature Elimination (RFE): RFE is a stepwise feature elimination process, where a model is trained, and each feature is evaluated based on the model's coefficients or importance. Features with smaller impact are gradually removed.

- **Tree-based Feature Selection:**
  This method uses tree-based algorithms (such as Random Forest or XGBoost) to compute the importance of features and selects those with higher importance.

### 2. Model-specific Feature Selection

For each learner, feature selection may require different strategies depending on the model. For example, decision trees and linear regression have different criteria for feature selection and may require distinct feature selection methods.

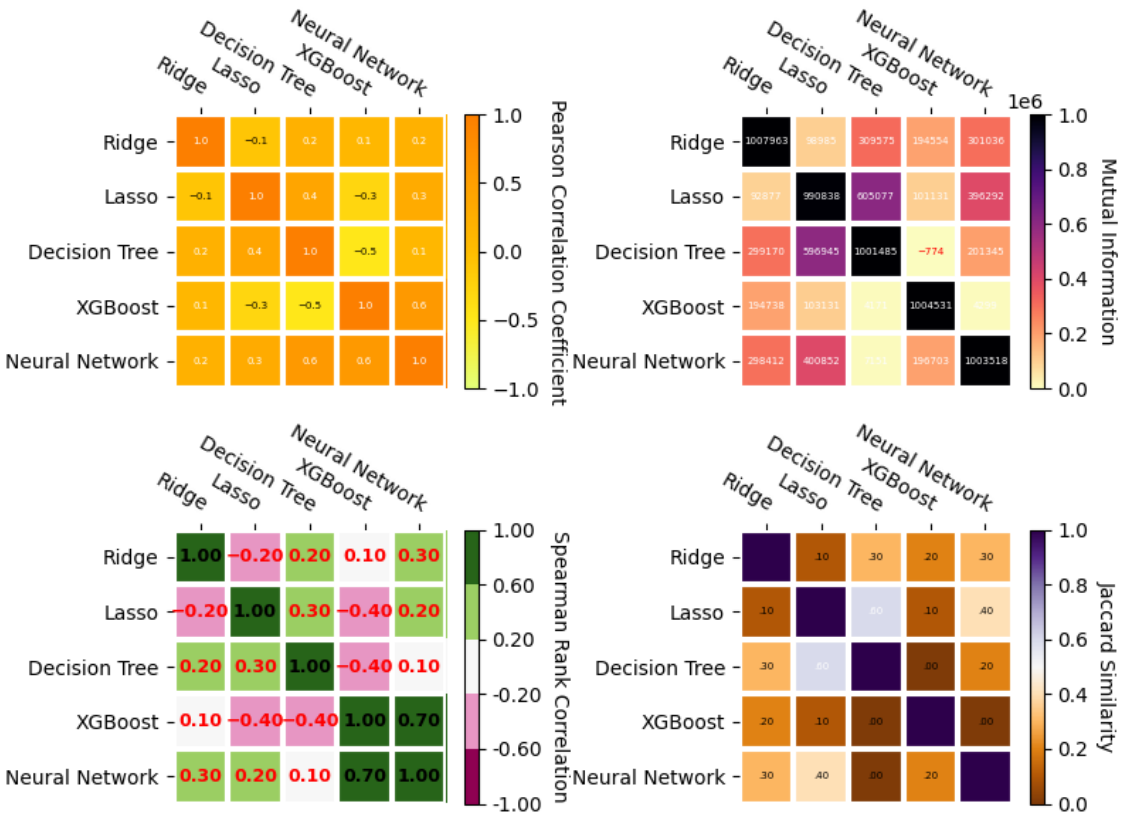



Figure 6: Each correlation index of the prediction results of different models measures the results

**Alignment** The alignment process is crucial for integrating different learners' advantages and enhancing learning performance. It has three key steps: using base learners, meta - learner secondary regression, and machine weighting.

1. **Using base learners**

- **Selection of Base Learners:** Select base learners like multivariate linear regression, ridge regression, Lasso regression, decision trees, XGBoost, and neural networks. Each has unique features based on different algorithms, useful for analyzing medal - related datasets.
- **Training of Base Learners:** Partition the dataset into training and test sets. Train each base learner independently on the training set to learn feature - target variable relationships and minimize training errors, providing diverse data for integration.

2. **Meta - learner quadratic regression**

- **Generation of Meta - Data**After training, base learners predict on the test set. Their predictions and true labels form meta - data, reflecting base learners' performance on the test set in the context of medal prediction.

- **Construction and Training of the Meta - Learner:** Choose a meta - learner (e.g., multivariate linear regression) and train it with meta - data. It learns the mapping between base learners' predictions and true labels to correct and optimize base learners' predictions.

**3. Machine Empowerment**

- **Weight Calculation Basis**Assign weights to base learners based on their performance during meta - learner training. Factors like prediction error on meta - data or correlation with true labels determine weights. Better - performing base learners get larger weights in medal - related predictions.
- **Generation of Final Prediction**Combine base learners' predictions through weighted aggregation (e.g., linear weighted sum). This leverages their advantages, improving the accuracy and stability of medal - ranking predictions.

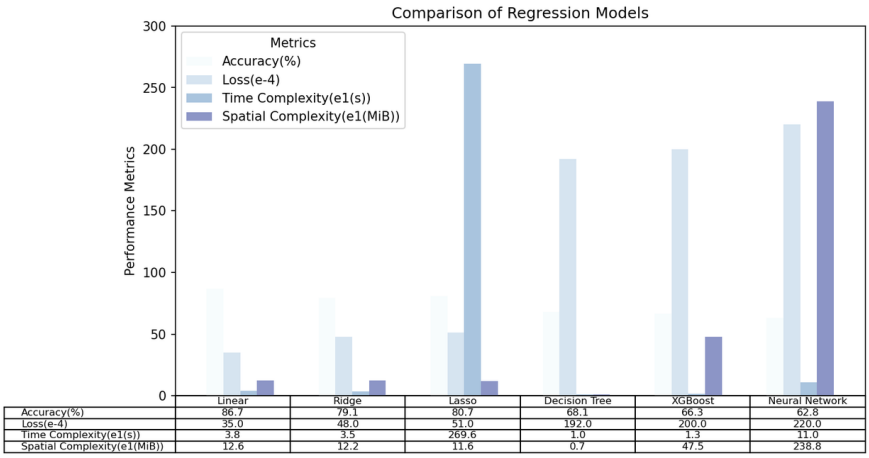
In summary, the heterogeneous learner registration process uses base learners for diverse predictions, a meta - learner for optimization, and machine - assigned weights for integration, leading to more accurate medal - related prediction results.

# 3 Solution of the problem

## 3.1 Problem I

In Figure 4-1, reflecting the differences between the sub-models in terms of accuracy, loss, and spatiotemporal complexity, while the ensemble model outperforms all the individual sub-models.

In Figure 4-2, aside from the aforementioned and the calculations provided, here is the error analysis of five base learners using RMSE, MAE, $R^2$, and MAPE.




| | Linear | Ridge | Lasso | Decision Tree | XGBoost | Neural Network |
|---|---|---|---|---|---|---|
| Accuracy(%) | 86.7 | 79.1 | 80.7 | 68.1 | 66.3 | 62.8 |
| Loss(e-4) | 35.0 | 48.0 | 51.0 | 192.0 | 200.0 | 220.0 |
| Time Complexity(e1(s)) | 3.8 | 3.5 | 269.6 | 1.0 | 1.3 | 11.0 |
| Spatial Complexity(e1(MiB)) | 12.6 | 12.2 | 11.6 | 0.7 | 47.5 | 238.8 |

Figure 7: Comparison chart of results evaluation of each model

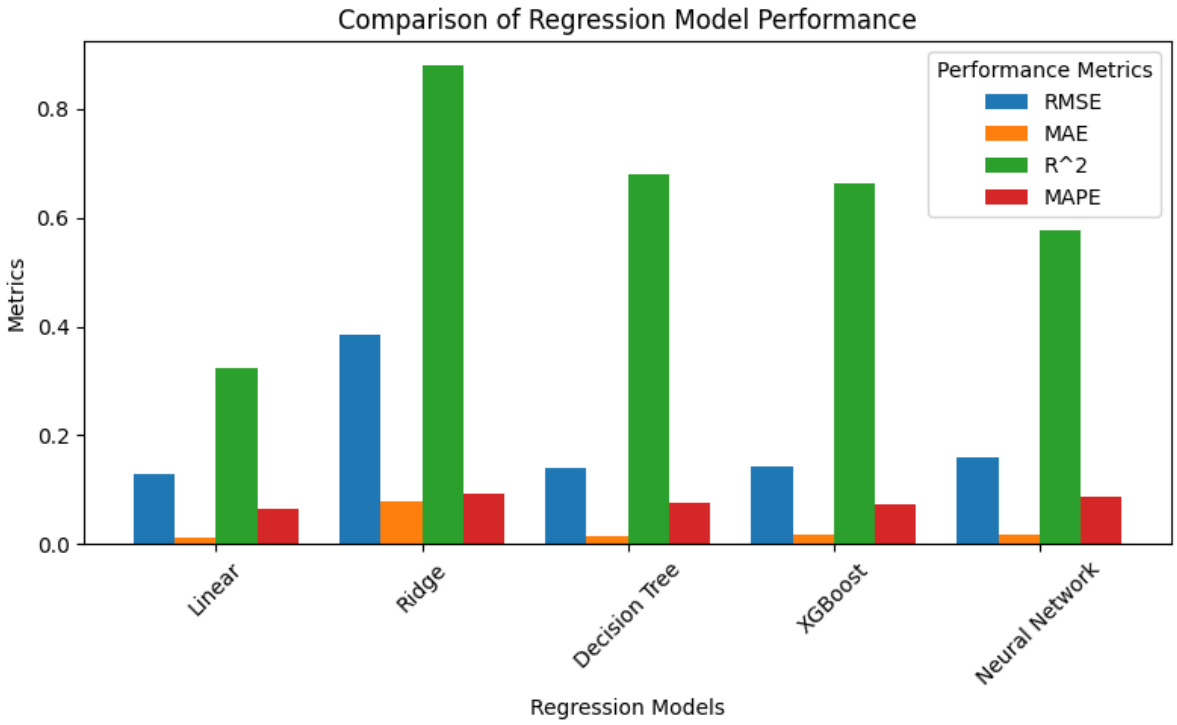



Figure 8: Each precision index of the prediction results of different models measures the comparison results

### 3.1.1 Problem i

The ensemble model predicts medal outcomes for each country in every event, with results saved in a CSV file in the appendix. The country-NOC code mapping is sourced from the provided data, with NOC used as the country code, as Table 3. Historical countries, like the Soviet Union, are excluded from the final ranking. A multivariate linear regression model based on historical medal data generates the 2028 predicted ranking, as Table 2. The two results are correlated, and weights are adjusted through debugging to produce the final medal ranking.

| NOC | Gold | Silver | Bronze | Total |
|---|---|---|---|---|
| USA | 41 | 43 | 41 | 124 |
| CHN | 36 | 29 | 21 | 87 |
| GBR | 24 | 24 | 28 | 76 |
| ROC | 20 | 28 | 23 | 71 |
| ANZ | 32 | 2 | 33 | 67 |
| CRC | 0 | 9 | 53 | 63 |
| FRA | 15 | 23 | 19 | 57 |
| JPN | 23 | 13 | 17 | 53 |
| AUS | 17 | 15 | 18 | 51 |

Table 2: ”Predictions of the Medal Tables from Previous Editions”

| NOC | Gold | Silver | Bronze | Total |
|---|---|---|---|---|
| USA | 116 | 23 | 26 | 165 |
| CHN | 29 | 102 | 20 | 151 |
| FRA | 36 | 8 | 5 | 49 |
| GBR | 9 | 28 | 4 | 41 |
| AUS | 0 | 9 | 16 | 25 |
| COL | 0 | 4 | 12 | 16 |
| ARM | 1 | 2 | 11 | 14 |
| BRN | 0 | 1 | 9 | 10 |
| JPN | 0 | 0 | 6 | 6 |

Table 3: Medal counts for selected countries predicted by little programs

Each ID table predicts gold, silver, and bronze winners. Ensemble learning, combining base learners and a meta-learner, integrates predictions. The final 2028 results are output. The model's strong performance ensures stability, with accuracy as the metric to assess prediction quality, indicating the probability of correct predictions.

The steps to calculate the confidence interval for the predicted medal ranking using accuracy as the probability and MSE are as follows(known MSE is 0.018, and accuracy is 0.9)

**Calculating Standard Error (SE) and Confidence Interval (CI)** First, calculate the standard error (SE). The standard error is the square root of the mean squared error (MSE):

$$SE = \sqrt{MSE} \tag{1}$$

Given $MSE = 0.018$, we have:

$$SE = \sqrt{0.018} \approx 0.134$$

Next, determine the Z-value for the confidence interval. The Z-value depends on the chosen confidence level. For a 90% confidence interval, the Z-value is 1.645. For a 95% confidence interval, the Z-value is typically 1.96.

Finally, use the following formula to calculate the confidence interval (CI) for each predicted value:

$$CI = \hat{y} \pm Z \times SE \tag{2}$$

Where:

- $\hat{y}$ is the predicted value of the model
- $Z$ is the Z-value corresponding to the confidence interval
- $SE$ is the standard error calculated

| NOC | Total | CI_lower | CI_upper |
|---|---|---|---|
| USA | 141 | 140.50 | 141.54 |
| CHN | 108 | 107.48 | 108.52 |
| GBR | 58 | 57.50 | 58.54 |
| FRA | 47 | 46.53 | 47.57 |
| AUS | 44 | 43.47 | 44.51 |
| JPN | 37 | 36.46 | 37.52 |
| ITA | 22 | 21.51 | 22.57 |
| NED | 21 | 20.50 | 21.54 |
| GBR | 18 | 17.47 | 21.54 |
| CAN | 15 | 14.49 | 15.54 |

Figure 9: Confidence interval of the total for the top 10


Figure 10: 2028 Predicted Medal Standings

**Evaluation Criteria and Calculation of Scores** The evaluation criteria are: 1) change in ranking (a) and 2) change in the number of medals (b). The score is defined as:

$$\text{score} = f(\text{rank}) \times a + b \tag{3}$$

where $f(\text{rank})$ is a function of the 2024 ranking, and the higher the rank, the larger the value of $f(\text{rank})$. Define $f(\text{rank}) = \frac{m}{\text{rank}}$, where $m$ is a constant.

After calculation, the top five and bottom five $f(\text{rank})$ values are obtained, as shown in the Table 5.

| Rank | $f(\text{rank})$ Value |
|---|---|
| 1 | $\frac{m}{1}$ |
| 2 | $\frac{m}{2}$ |
| 3 | $\frac{m}{3}$ |
| 4 | $\frac{m}{4}$ |
| 5 | $\frac{m}{5}$ |
| ... | ... |

Table 4: Top 5 and Bottom 5 $f(\text{rank})$ Values

| Better | Worse |
|---|---|
| United States | Japan |
| China | Australia |
| Great Britain | South Korea |
| Brazil | Netherlands |
| France | Canada |

Table 5: Advancing and Declining Nations

### 3.1.2 Problem ii

By deduplicating the NOC column in the *summerOly_athletes.csv* file, we obtain all participating countries' NOCs, and by deduplicating the Team column in the *summerOly_medal_counts.csv* file, we get the NOCs of medal-winning countries. The difference between these sets reveals countries/regions that haven't won medals. Using the model, the 2028 medal ranking is generated, and the intersection with countries that haven't won medals identifies those likely to win their first medal.

**Prediction reasoning:** Monaco, which first participated in 1920 and missed only a few Olympics, has the most participation among countries without a medal, indicating strong Olympic commitment. By calculating the cumulative probability for each medal, countries with a probability above 55% are considered likely to win their first medal.

So we predict that 5 countries will win their first-ever medal , as Table 6.

**Odds Calculation:** The probability of guessing all five countries wrong is:

$$0.45 \times 0.40 \times 0.38 \times 0.43 \times 0.45 = 0.0316$$

The odds are the inverse of the probability, approximately:

$$\text{Odds} = \frac{1}{0.0316} \approx 31.75 \tag{4}$$

article graphicx

### 3.1.3 Problem iii

Explore the relationship between the events and how many medals countries earn. **What sports are most important for various countries? Why?**

| NOC | PR |
|---|---|
| Monaco(MON) | 55% |
| Myanmar(MYA) | 60% |
| United Arab Emirates(UAR) | 62% |
| American Samoa(ASA) | 57% |
| Mali(MLI) | 55% |

Table 6: The Five Countries Predicted to Win for the First Time.

**Defining the Score and Impact Factor** After processing the data, all sub-events, such as men's 400m, men's 800m, and women's 400m, are grouped under the category "ATH" (Athletics). This allows us to observe how different countries' medal counts in various events evolve over time.

We define the score as:

$$\text{score} = (\text{Gold Medals} \times 5) + (\text{Silver Medals} \times 3) + (\text{Bronze Medals} \times 1) \tag{5}$$

This score serves as an indicator of the importance of an event to a country. Below, we provide three 3D plots for the United States, China, and Australia. However, since each major event includes a different total number of medals, the score alone does not fully reflect the event's importance. Therefore, we calculate the impact factor by dividing the score by the total number of medals:

$$\text{Impact Factor} = \frac{\text{score}}{\text{Total Medals}} \tag{6}$$

This gives a normalized measure of the importance of the event to each country.

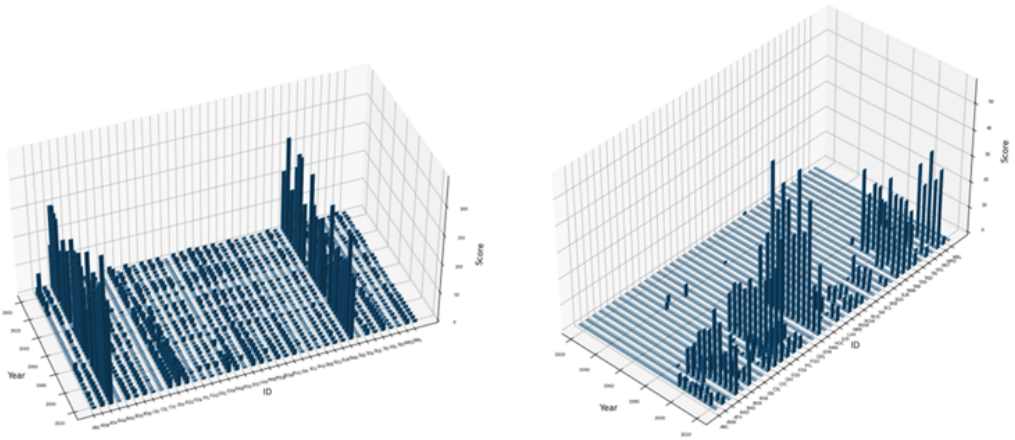



Figure 11: 2028 Predicted Medal Standings USA(left) CHN(right)

From the left chart in Figure 11, athletics and swimming are key for the U.S., with basketball, tennis, cycling, and gymnastics also important. The U.S. dominates athletics with 861 medals, especially in sprints, long-distance, and field events. Swimming has 614 medals, and gymnastics has 126, with women's gymnastics leading globally. The U.S. men's and women's basketball teams maintain an advantage, and cycling excels in men's sprint and women's road racing.

From the right chart in Figure 11, diving, gymnastics, and table tennis are key for China, with badminton, weightlifting, and shooting also important. Since the 1984 LA Olympics, China has

dominated diving, winning 162 medals. In table tennis, China has almost monopolized golds since its inclusion in 1988, with 56 medals. Badminton, especially women's, is another strong sport, while China excels in gymnastics, accumulating 96 medals since 1984. The weightlifting team maintains global leadership, and shooting, particularly in rifle and pistol events, remains a strength.

From the statistical data, it can be seen that gymnastics, fencing, and weightlifting are particularly important to Russia, while rowing, cycling, and equestrian are important to the United Kingdom. Judo and karate are particularly important to Japan, and fencing, cycling, and sailing are important to France. Detailed data for each country can be found in the "coach" folder (with subfolder names corresponding to each country's NOC, and the first column of the file representing the event ID).

**How do the events chosen by the home country impact results?** Visualize the total number of medals earned by all countries over the years Analysis: It can be observed that some countries have

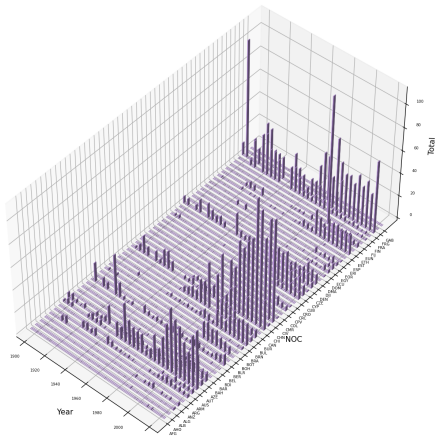



Figure 12: "Total number of medals won by each country per year"

consistently earned a high number of medals, while others have only become strong in recent years. The number of medals for each country has changed over time to varying degrees, with points of significant change referred to as mutation points. Visualize the 2D chart for a single country, as shown in the Figure 12, using China, the United States, and France as examples.

It is known by the given data that China hosted the Games in 2008, the United States in 1904, 1932, 1984, and 1996, and France in 1900, 1924, and 2024. It can be observed that the total number of medals for host countries significantly increased, which is closely related to the events chosen by the host country. Host countries tend to choose sports in which they have strong performances, which greatly impacts the final results.

## 3.2 Problem II

**Example: Changes in the United States Volleyball Team** Based on the known data, the medal history of the U.S. women's volleyball team is statistically analyzed. The score is calculated as:

$$\text{score} = (\text{gold medals} \times 5) + (\text{silver medals} \times 3) + (\text{bronze medals} \times 1) \tag{7}$$

It is known that Lang Ping became the coach of the U.S. women's volleyball team in 2005. From the data, we can see that in 2008, the U.S. women's volleyball team achieved a silver medal, which was

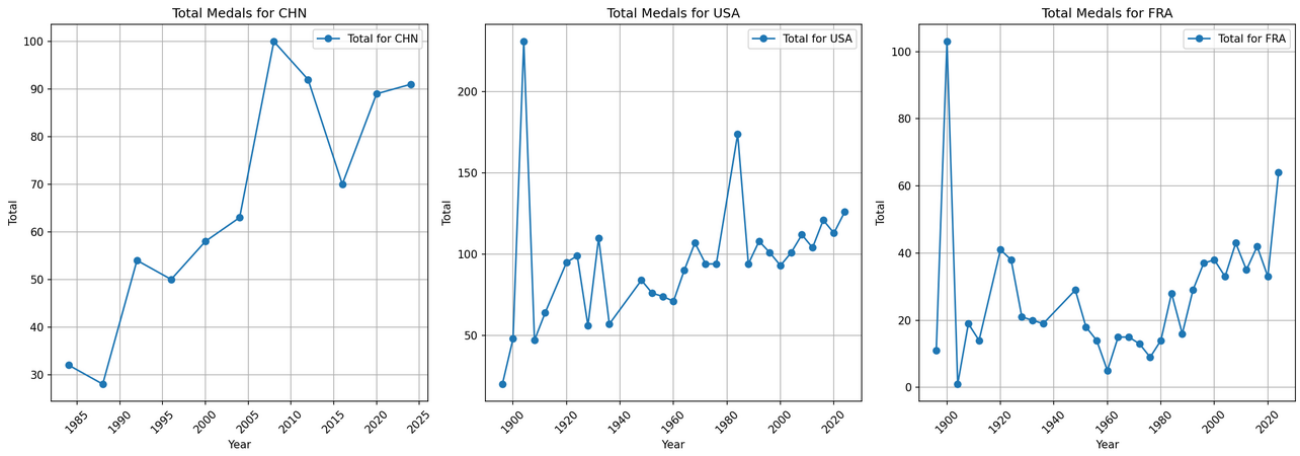



Figure 13: 2028 Predicted Medal Standings

their second silver medal since 24 years ago. In the subsequent Olympics, the U.S. women's volleyball team continued to perform well, proving the existence of the "great coach" effect, as Figure 14.

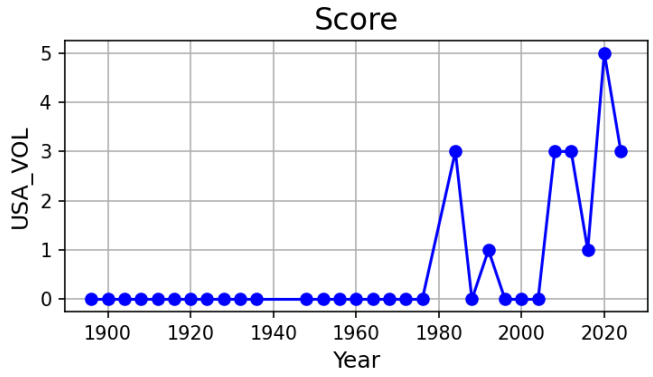



Figure 14: 2028 Predicted Medal Standings

As shown in Figure 10, after selecting a specific event for the y-axis, the score changes with the year, and it is often observed that there are abrupt changes. Of course, there are many factors that could influence these abrupt changes, such as historical and political reasons, but the possibility of the 'great coach' effect influencing the results also exists.

### 3.2.1 Abrupt Change Effect Calculation

Let the reference year be standard and the current year be now. Define the cumulative totals:

$$\text{sum\_G} = \sum_{y=\text{standard}}^{\text{now}} G(y), \quad \text{sum\_S} = \sum_{y=\text{standard}}^{\text{now}} S(y), \quad \text{sum\_B} = \sum_{y=\text{standard}}^{\text{now}} B(y)$$

Where $G(y)$, $S(y)$, and $B(y)$ represent the number of gold, silver, and bronze medals in year $y$ respectively.

For each year $y$ from standard to now:

If medals were won in year $y$, update:

$$\text{standard} = \text{year}, \quad x = \text{year} - \text{last\_award\_year}$$

Record the maximum gap $x$ and update the award totals:

$$\text{sum\_G} = \sum_{y=\text{standard}}^{\text{year}} G(y), \quad \text{sum\_S} = \sum_{y=\text{standard}}^{\text{year}} S(y), \quad \text{sum\_B} = \sum_{y=\text{standard}}^{\text{year}} B(y)$$

Finally, calculate the abrupt change effect:

$$f(x) = m \cdot x - 5 \cdot \text{sum\_G} - 3 \cdot \text{sum\_S} - 1 \cdot \text{sum\_B}$$

Where $m > 0$ is a constant.

The function $f(x)$ represents the degree of abrupt change of the current year relative to the selected reference year. The larger the value of $f(x)$, the greater the degree of change, and the higher the likelihood of being influenced by the "great coach" effect.

Since the choice of the standard reference year affects the judgment of the abrupt change degree, through programming calculations combined with qualitative analysis, the following case was identified with a significant abrupt change:

In the 2020 Tokyo Olympics, the mixed doubles table tennis event in Japan, where $f(x)$ reached an astonishing 289 (with standard = 1990). This can be attributed not only to Japan being the host country, but also to Japan's increased investment and its proactive hiring of foreign coaches, especially Chinese coaches. This, to a certain extent, confirms the existence of the "great coach" effect. **Evidence 1:** Anastasia Ilyinichna Bliznyuk, a Russian national, coached the Chinese Rhythmic Gymnastics team, which won its first Olympic gold medal in the event at the 2024 Paris Olympics.

**Evidence 2:** In 2006, Liu Guoliang accepted an invitation from the Singapore Table Tennis Association and signed a three-year coaching contract. During Liu's tenure as coach, the team's performance significantly improved. At the 2008 Beijing Olympics, they successfully won the silver medal in the women's team event.

**How much do you estimate such an effect contributes to medal counts?**

- **Enhancing Strength through Personalized Training:** Excellent coaches create personalized training plans based on the athletes' characteristics, improving their skills, tactics, and physical fitness, thereby boosting their competitiveness and increasing the chances of winning more medals.

- **Resolving Conflicts to Strengthen Team Cohesion:** Coaches effectively manage conflicts, enhancing team cohesion, helping athletes maintain a good mental state, building confidence for competitions, and fostering collaboration to pursue Olympic medals.

- **Managing Risks to Enhance Competition Adaptability:** Coaches teach athletes to cope with risks, helping them adapt flexibly in complex competition environments, which is crucial for maintaining stable performance at the Olympics and improving the likelihood of winning medals.

**Three countries:**

**Scenario 1: Weak Strength**

**China Football:** Chinese football currently has relatively weak competitiveness both in Asia and globally. Despite significant investments and infrastructure development, its performance in international competitions has consistently fallen short of expectations. A great coach can not only improve the existing players' skills but also help establish a solid youth training system to nurture more young talent and lay a strong foundation for Chinese football. **Impact:** Help Chinese players improve tactical awareness and psychological resilience, leading to better results on the field.

**Scenario 2: Insufficient Resources**

**Kenya Athletics:** Kenyan athletics, especially long-distance running, has achieved remarkable international success in recent years. Kenyan athletes have frequently won world-class marathons, but there are still gaps and shortcomings in various track and field events. Although Kenyan runners have impressive endurance, their training methods and facilities are relatively basic. Introducing a great coach with modern training methods could help athletes improve the quality of their training and overcome the disadvantages brought about by resource limitations. **Impact:** Considering the athletes' endurance and Kenya's strength in long-distance running, a coach with cross-disciplinary management experience could help Kenya break through in other track and field events, expanding its competitive advantages.

**Scenario 3: Strong Strength**

**Japan Table Tennis:** While Japan's table tennis team has made breakthroughs in recent years, China remains the dominant power in the sport. Japan still faces a gap in overall competitiveness, especially when competing against strong teams like China and Germany. At crucial moments in major international tournaments, Japan is often limited by technical and tactical factors. A great coach can typically provide unique tactical guidance, helping athletes face tough opponents and improve their overall technical level. **Impact:** By bringing in coaches from table tennis powerhouses like China and Germany, Japan can learn their methods and techniques, complement its strengths, and improve its competitiveness.

## 3.3 Problem III

The Olympic medal count is not just a measure of a country's sports performance; it also reflects deeper aspects of history, politics, culture, and more. By analyzing the medal counts, we gain insight into a country's Olympic performance and the factors behind it.

**1. Olympic Medal Count Reflects a Country's Historical Status**

Data analysis shows that after the Soviet Union dissolved, Russia's medal count dropped significantly. During the Cold War, the Olympic rivalry between the U.S. and the Soviet Union symbolized their global competition. Post-dissolution, Russia's medals changed due to shifts in political and economic structures.

Similarly, China, initially weak, grew stronger over time as the country stabilized, invested more in the Olympics, and focused on youth development.

**2. Olympic Medal Count and Economic Level**

A strong economy allows countries like the U.S., Germany, and the UK to invest in sports infrastructure, athlete training, and event participation, boosting Olympic success. In contrast, low-income countries struggle to provide adequate resources, limiting their ability to perform well, especially in multi-sport events.

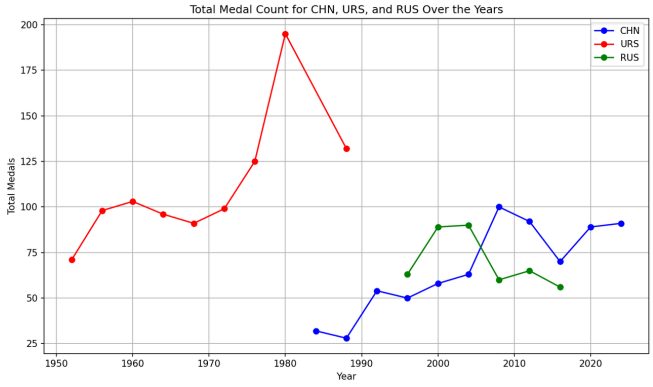



Figure 15: 2028 Predicted Medal Standings

**3. Olympic Medal Count and Cultural Influence**

Countries' sports cultures significantly impact their Olympic performance. For example, Brazil's football tradition has led to dominance in the sport but has also resulted in weaker performances in other events.

**Explain how these insight(s) can inform country Olympic committees**

For countries with strong economies and diverse sports, NOCs can create efficient resource allocation systems to ensure fair support across sports, boosting overall Olympic performance. For nations excelling in specific events, NOCs can concentrate resources to strengthen these areas. With increasing global competition, NOCs must assess the international landscape, focus on strong events, and invest in other areas to remain competitive.

# 4 Evaluation of the model

## 4.1 Model Adaptability

Linear regression suits linear data but underfits. Ridge regression works for correlated features but not nonlinear ones. Decision trees capture nonlinearity but may overfit. XGBoost adapts well to various data but needs tuning. Neural networks excel at complex problems but require extensive training. Stacked ensembles combine model strengths for better adaptability.

## 4.2 Robustness

Ridge regression is robust to multicollinearity. Decision trees handle noise but may overfit. XGBoost is strong against noise and imbalance but costly. Neural networks need regularization to avoid overfitting. Stacked ensembles improve robustness by combining models.

## 4.3 Ability to Handle Imbalanced Datasets

Linear and ridge regression struggle with imbalances. Decision trees need tuning for imbalanced data. XGBoost handles imbalances well with weighted loss. Neural networks work with class weighting or oversampling. Stacked ensembles combine models to handle imbalances.

# 5 Conclusions

Historical Olympic data is cleaned and standardized for consistency. Regression models and stacked ensemble learning (XGBoost, Random Forest) predict medal counts at various levels and identify strengths in specific sports for different countries. The analysis also predicts first-time medal-winning countries and investigates factors like policy changes, economic growth, and athlete training. The ”great coach” effect is explored, emphasizing the coach's role and suggesting improvements through high-level coach imports and better training systems. Additionally, medal data reveals insights into how economic, cultural, and social policies influence Olympic performance, supporting future sports policy and training strategy development.

# References

[1] Agarwal, Anurag, Davis, Jefferson T., Ward, T. Supporting ordinal four-state classification decisions using neural networks, Information Technology and Management, Vol. 2, 2001, pp. 5-26.

[2] Ho, Tin Kam. Random decision forests, In: Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995, Vol. 1, pp. 278-282.

[3] Jowett, Sophia. Coaching effectiveness: the coach–athlete relationship at its heart, Current Opinion in Psychology, 2017, Vol. 16, pp. 154-158.

# Appendix

## Appendix: List of Attached Codes and Data

| No. | File Name | Description and Data |
|---|---|---|
| 1 | final result.csv | Final predicted medal table |
| 2 | data.rar | Preprocessed dataset |
| 3 | medal_summary .csv | Sum of individual events' predicted medal tables . |
| 4 | predicted_2028.csv | Predicted medal table based on historical medal counts . |
| 5 | First-time winner .csv | List of countries winning medals for the first time |
| 6 | data_pre_process | Data preprocessing program |
| 7 | others | Other models and answers to some questions |