OLYMPIC GAME ANALYSIS AND PREDICTION USING MACHINE LEARNING

Manjushree M P.E.S College of Engineering,Mandya

Kruthika Prasad P.E.S College of Engineering,Mandya

Kishore R E P.E.S College of Engineering,Mandya

Kashaf Khan K S P.E.S College of Engineering,Mandya

ABSTRACT:

This study presents an analysis of the Olympic Games using machine learning techniques. The Olympic Games represent a unique and extensive dataset of sports performance spanning over a century. In this research, we leverage machine learning algorithms to extract valuable insights from this data, such as predicting medal tally, country-wise analysis, athlete-wise analysis and overall analysis. The analysis begins with data pre-processing, including data collection, cleaning, and feature engineering. Various machine learning models are then applied to the dataset, including regression, classification, and clustering algorithms. These models are used to predict medal winners, classify athlete performance categories, and group similar sports disciplines based on historical data.

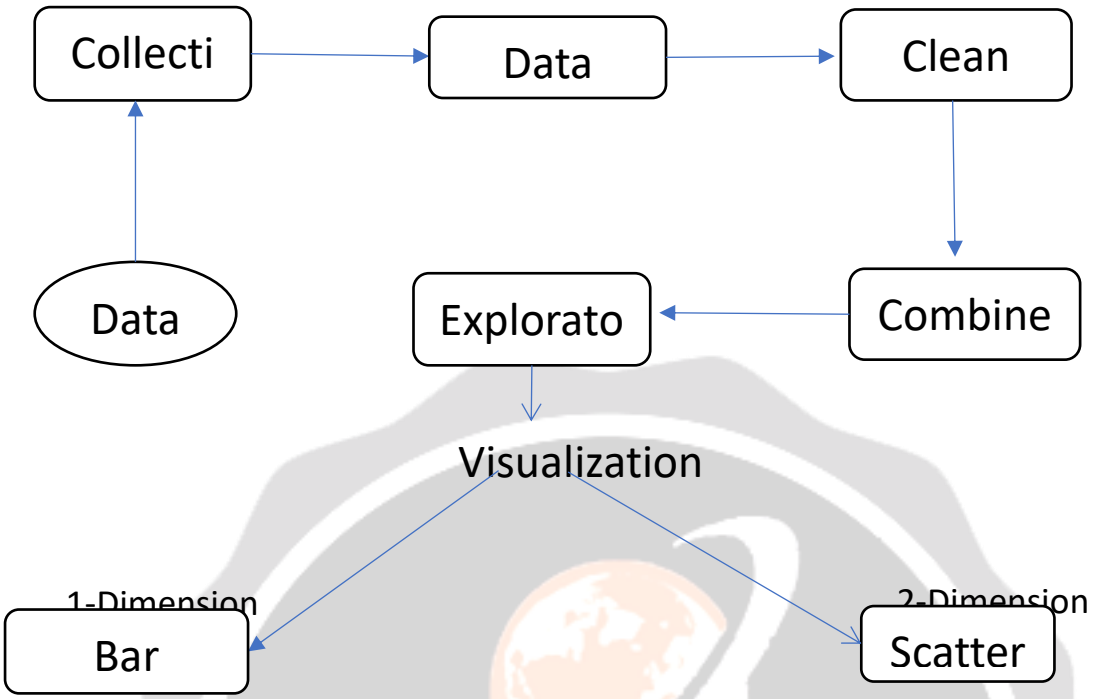
KEYWORDS:

Machine learning, Logistic Regression Random forest, Analysis, Prediction.

INTRODUCTION:

Machine Learning (ML) has emerged as a powerful tool for analyzing various aspects of sports, including the Olympic Games. ML algorithms can uncover hidden patterns and trends in large datasets, providing valuable insights into athlete performance, team strategies, and overall Olympic trends. Predicting medal counts accurately forecasting the number of medals a country is likely to win in future Olympic Games. Analyzing Athlete Performance using ML algorithms can analyze factors such as an athlete's age, height, weight, training records, and previous performance to identify patterns and predict their potential performance in future competitions. Understanding Olympic Trends using ML can be used to analyze historical data to identify trends in Olympic participation, medal distribution, and the evolution of sports and events over time. Uncovering Hidden Insights using ML algorithms can uncover hidden patterns and correlations in Olympic data that may not be apparent through traditional analysis methods, leading to new insights and understanding of the Olympic Games.

System Architecture:



A. METHODOLOGY:

1.Data Collection:

- Gather the dataset which has basic bio data on athletes and medal results from Athens 1869 to Rio 2016 from Kaggle.com.
(https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results)

2. Data Cleaning:

Missing Values: Identify and handle missing values in athlete information, past performance data, or economic indicators. Techniques include imputation (filling in missing values with estimates) or deletion depending on the data and its importance.

Inconsistent Formats: Ensure consistency in units for measurements (e.g., centimeters vs. inches for height) and date formats across datasets from different sources.

Outliers: Detect and address outliers, which are data points significantly different from the rest. Consider removing them if they represent errors or investigating the cause if they're genuine.

3. Data Pre-processing:

- Conduct exploratory data analysis to understand the data's structure, distribution, and relationships between variables.
- Visualize the data using techniques like histograms, scatter plots, and box plots to identify trends, patterns, and outliers.
- Transform features as needed (e.g. encoding categorical variables) to make them suitable for machine learning algorithms.

4. Feature Engineering:

Create new features from existing ones to improve model performance. For example, calculate an "average medal count per Olympics" based on historical data.

Normalization and Scaling: Scale numerical features (like age, height, or GDP) to a common range to prevent certain features from dominating the model due to their scale.

Categorical Encoding: Convert categorical data (like sport or country) into numerical representations suitable for machine learning algorithms. This can involve techniques like one-hot encoding or label encoding.

5. Exploratory Data Analysis: Use data visualization techniques like histograms and scatter plots to understand data distribution, identify trends, and relationships between features. This helps determine the best preprocessing techniques.
Feature Importance Analysis: Analyze which features have the most significant influence on the target variable (e.g., medal count or winning probability) to guide feature selection and model building.

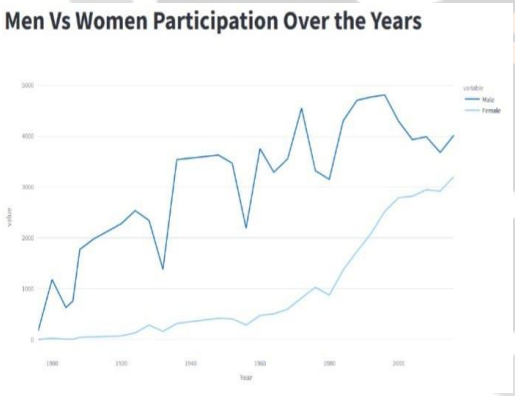
ANALYSIS AND VISUALIZATION

A. Medal Tally in last 1896 to 2016

Medal Count: Machine learning models can predict the total medal count for a country based on historical performance, athlete data (age, experience), economic factors (GDP), and investment in sports programs.
Medal Types: Algorithms can classify the likelihood of a country winning a specific medal type (gold, silver, bronze) for an event, considering factors like past performance in that event and current athlete rankings.
Individual Winners: Predicting individual winners in specific events is more challenging due to unforeseen circumstances like injuries. However, models can analyze athlete performance data, competition history, and recent form to suggest potential medal contenders.

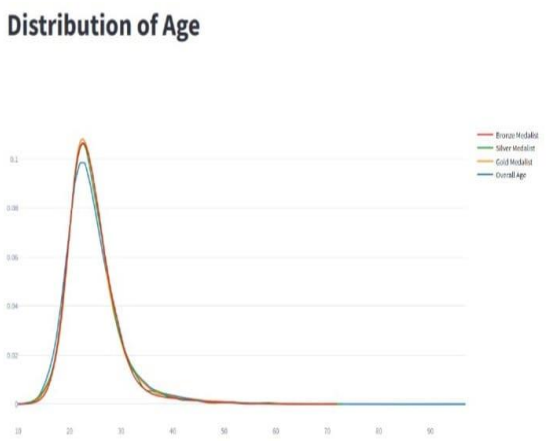
B. Men Vs. Women Participation over the years :

Separate Models for Men and Women. Train separate ML models for male and female athletes to account for potential biological and training differences between genders. This can lead to more accurate performance predictions for both.
Inclusion of Gender-Specific Metrics:. Integrate gender-specific metrics into the models, like body composition data or hormonal factors that might influence performance in certain sports.
Fairness and Bias Mitigation .Implement techniques to mitigate bias in the models that might favor one gender over the other. This might involve data balancing.



C. Age Distribution

Identifying Optimal Performance Age: Analyze athlete performance data across different sports and age groups. This can reveal the age range where athletes tend to peak in different sports.
Impact of Training and Technology: Use ML models to assess how advancements in training methods and sports science influence the optimal age range for peak performance over time.
Predicting Future Age Trends: Develop models to predict potential shifts in the age distribution of Olympic athletes. This can help anticipate changes in training strategies or talent development programs.

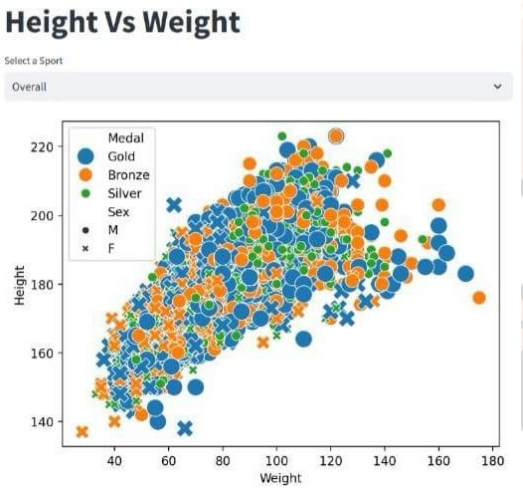


D. Height Vs. Weight

In the Olympics, there are no restrictions on the height and weight of athletes. However, certain sports have specific rules and regulations regarding weight classes or height requirements.

While height and weight can be factors in athletic performance, using machine learning for Olympic predictions requires a nuanced approach.

Overall, while height and weight can play a role in an athlete's success in certain sports, there are no universal restrictions or requirements in the Olympics.



APPLICATIONS OF MACHINE LEARNING ALGORITHM FOR PREDICTION

A. Logistic Regression

Logistic regression is a valuable tool used in analyzing and predicting aspects of the Olympic games. Here's how it's applied:

Binary Outcomes: Logistic regression excels at modeling events with two possible outcomes, which is perfect for situations like predicting:

Will a country win a medal (yes/no) in a specific event ?

Will a team win the game (win/loss) ?

Probability Predictions : Logistic regression doesn't predict a simple win or loss, but rather the probability of that outcome. This provides a more nuanced picture, like the likelihood of a specific country medaling in an event.

Factors Influencing Outcomes: The model can incorporate various factors that might influence the outcome. Examples include:

Past performance of athletes or teams

World ranking .
Geographic size (for medal count predictions).
Limitations: It's important to remember that logistic regression has limitations:
* Data Availability: The model's accuracy depends on the quality and comprehensiveness of historical data used.
* Unforeseen Events: Injuries, surprise performances, and other unpredictable events can impact outcomes.

B. Random Forest Random Forest

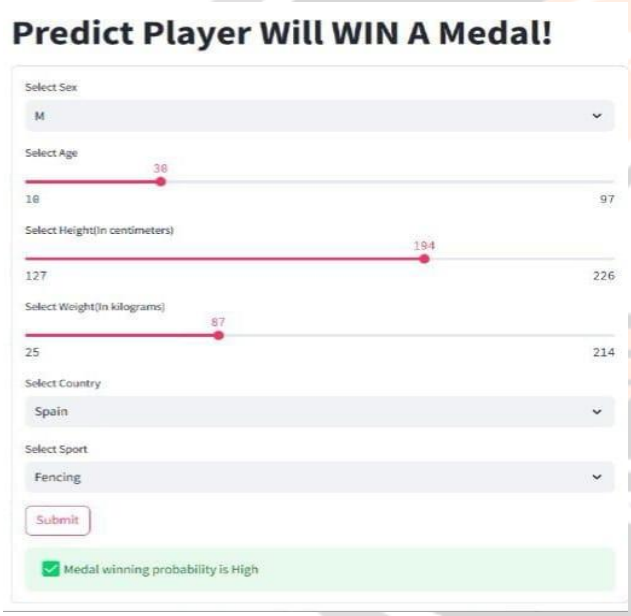
Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.**Scatter Plot and many more. Scatter Plot and many more.**

The random forest uses forecasts from all of the decision trees rather than just one, and bases its final prediction exclusively on user feedback. It reduces the problem of over fitting in the model.

COMPARISON OF LOGISTIC REGRESSION AND RANDOM FOREST

Parameters	**Logistic Regression**	**Random Forest**
Implementation	Easy to Interpret	Hard to Interpret
Accuracy	Accurate	Excellent Accuracy
Overfitting	Data likely to be over fit	Data unlikely to be over fit
Outliers	Highly susceptibe outliers	Resilient to outliers
Computations	Quickly construct	Slowly develop

RESULTS OF PREDICTION



CONCLUSION:

This analysis of Olympic data provided insights into historical trends and factors influencing performance. While past performance offers a foundation for predictions, the Olympics are a complex competition where unexpected factors can influence outcomes. Injuries, emerging talents, and even host country advantages can all play a role.

Despite these limitations, data analysis is a valuable tool for athletes, coaches, and sports organizations. By understanding historical trends and identifying key factors, stakeholders can develop targeted training programs and strategies to improve performance. Ultimately, the Olympic Games celebrate the pinnacle of athletic achievement, inspiring a global community through exceptional displays of human potential.