Olympic Data Analysis and Sports Prediction Using Machine Learning

Saifi Ali, Department of Machine learning, BMS College of Engineering  
Email: saifi.ai22@bmsce.ac.in

**Abstract—**

Sports performance is significantly influenced by an individual's physical attributes, such as age, height, weight, and sex. This project aims to develop a machine learning model that predicts the most suitable Olympic sport for a user based on these attributes. By analyzing historical Olympic athlete data, the model identifies patterns in body composition and assigns users to sports where they are most likely to excel.

Index Terms—Machine Learning, Sports Analytics, Olympics, Data Analysis, Prediction Models.

**I. INTRODUCTION**

Introduction

THE Olympic Games, known as Jeux Olympiques in French, represent premier international sporting events that showcase a variety of competitions in both summer and winter sports. These games bring together thousands of athletes from across the globe and are regarded as the pinnacle of global sports competitions, with participation from over 200 nations. Traditionally held every four years, the Olympics alternate between Summer and Winter Games every two years within this cycle. There are various games in Olympics like-

Athletics (Track & Field)

Boxing

Wrestling (Freestyle & Greco-Roman)

Judo

Taekwondo

Weightlifting

When examining the evolution of the Olympic Games over time, several key aspects come into focus. These include the growing number of participating nations and athletes, fluctuations in the number of events, and increasing expenditure associated with hosting the games. Additional factors to consider are the enhanced performance of specific countries and individual athletes, the rise in female participation, and the changing ratio of male to female participants. Other noteworthy trends include advancements in medical support for athletes, the impact of global events such as pandemics on player performance, and the growing emphasis on gender equality. Analyzing these scenarios provides valuable insights into how the Olympics have transformed over the years, offering a foundation for predicting future trends.

Their model was designed to provide users with an efficient, functional, and conventional platform, allowing them to benefit from visual representations. However, the proposed system was unable to predict winning probabilities or establish correlations between medal achievements and their influencing factors.

**II. LITERATURE REVIEW**

LITERATURE REVIEW

Predicting a nation's performance in the Olympics based on its historical achievements is a widely adopted approach. Utilizing previous data, such as top scores from earlier participations, helps estimate the probability of securing gold medals in future events, such as the 2016 Olympics. Similarly, an athlete's chances of earning a medal in upcoming games can be anticipated by analyzing their prior performance records. Machine learning techniques further enable heuristic modeling to predict a country's medal outcomes. Enhancing an athlete's training strategies by addressing their weaker areas can significantly improve their results. Analyzing efficiency and the societal importance of sports plays a vital role in assessing a country's Olympic achievements.

When examining sports categories, priority should be given to analyzing content that provides varied perspectives rather than focusing solely on spatial and temporal factors. Video content analysis offers richer insights compared to structured datasets. Additionally, exploratory data analysis employs visualization techniques to deliver detailed insights and statistical overviews, aiding in data comprehension and decision-making.

In May 2019, Kabita Paul, Elif Demir, and Anjali Bapot proposed a system utilizing data visualization to develop applications that incorporate concrete analysis examples.

**III. METHODOLOGY**

METHODOLOGY

A. Data Collection The initial step in any form of analysis, whether technical or non-technical, involves collecting data. To conduct an effective analysis on a particular problem, it is essential to gather a substantial amount of data. This data serves as the foundation for applying various techniques and algorithms to derive meaningful conclusions and achieve the desired outcomes. A larger volume of data typically leads to higher accuracy in results and increases confidence in decision-making based on these outcomes.

For our analysis of the Evolution of the Olympics over time, we utilized data from multiple sources to ensure a comprehensive and detailed study. We employed three datasets, each offering significant volume and variety for our analysis. The first dataset contains detailed information about athletes, including attributes such as gender, height, weight, country of representation, and medals won (gold, silver, and bronze), among other details. This dataset enables performance analysis for individual players and facilitates comparative studies between multiple athletes.

The second dataset includes information about participating countries and their total medal counts (gold, silver, and bronze). This dataset is valuable for conducting comparative analyses of country-level performances. Lastly, the third dataset lists the countries along with their respective country codes, which serve as unique identifiers. This dataset helps determine the total number of nations that have participated in the Olympics to date.

These datasets provide a diverse and extensive collection of accurate data, enabling the application of techniques such as Exploratory Data Analysis (EDA) to derive insights and reach meaningful conclusions.

Data Pre-Processing

After collecting data, the next crucial step is data processing. Raw data, which is the unprocessed information obtained directly from a data source like a dataset, cannot be used directly for applying techniques or machine learning algorithms such as Linear Regression, Decision Trees, or SVM. It needs to be refined and transformed into meaningful data.

Data pre-processing involves converting raw data into a usable format by carefully identifying and addressing errors, as well as removing redundant, incomplete, or incorrect entries. For instance, datasets often include fields like Age and Gender that may contain null values. These null entries can cause errors, especially when visualizing data in graphical formats. To ensure accuracy, such null values must either be omitted or replaced with valid alternatives.

A method known as Deterministic Imputation is employed to handle this issue. This technique determines missing values (NA or NaN) based on other entries in the same column. Two common models for deterministic imputation are:

Basic Numeric Imputation Model: In this approach, null values are replaced with the mean or median of the non-missing values in the same column.

Hot Deck Imputation: This method substitutes missing values with a similar value from another record within the same column. Unlike the numeric imputation model, Hot Deck Imputation can be applied to both numerical and categorical values.

By implementing these techniques, errors are minimized, and the dataset is prepared for accurate analysis and visualization.

Exploratory Data Analysis

After completing data pre-processing, the next crucial step is data analysis. This involves utilizing various techniques, such as Text Analysis, Diagnostic Analysis, and Exploratory Data Analysis (EDA), alongside machine learning algorithms like Linear Regression, Logistic Regression, SVM, and Decision Trees, to derive meaningful conclusions.

Since our research focuses on visualization and comparative studies of factors contributing to the evolution of the Olympic Games over time, we employ EDA as a primary method for this analysis. EDA enables a comprehensive examination of data, summarizing its key features primarily through visual representations. By leveraging EDA, we can explore the structure and content of the dataset using various types of graphs and plots.

Some common visualization methods used in EDA include:

Histogram

Bar Graph

Box Plot

Scatter Plot

These visualizations allow us to interpret data effectively, make comparisons between different plots, and explain the insights derived from them. EDA plays a vital role in understanding the data and forming the basis for further analysis.

Methodology

An approach refers to a structured methodology designed to address problems and work toward solutions effectively. Both technical and non-technical challenges benefit from a systematic approach that outlines a clear pathway to achieve the intended objectives. This research paper aims to delve into the rich history of the Olympic Games and examine their progression over the years. Various factors contribute to this evolution, and a thorough analysis and comparison of these elements require a well-organized approach.

The proposed methodology, illustrated in Figure 1, serves as a strategic roadmap for tackling the research problem. Each phase of this approach is carefully outlined to ensure a detailed and comprehensive understanding of the subject matter.

**IV. ANALYSIS AND VISUALIZATION OF DATA**

ANALYSIS AND VISUALIZATION OF DATA  
A. Medal Tally from 1896 to 2016

Medal Count: Machine learning models can forecast the total number of medals a country might win by analyzing historical performance data, athlete demographics (such as age and experience), economic indicators (like GDP), and investments in sports infrastructure and training programs.

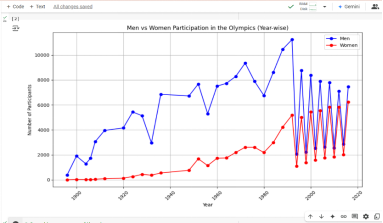
Medal Types: Algorithms can also estimate the probability of a country securing a specific type of medal—gold, silver, or bronze—by evaluating factors such as past achievements in particular events and the current rankings of athletes.

Individual Winners: Predicting the winners of individual events is more complex due to unpredictable factors like injuries or unforeseen circumstances. However, models can utilize data on athlete performance.

B. Men vs. Women Participation Over the Years

Separate Models for Men and Women: Develop distinct machine learning models for male and female athletes to account for biological and training differences between genders. This approach helps enhance the accuracy of performance predictions for both groups.

Incorporation of Gender-Specific Metrics: Include gender-specific factors in the models, such as body composition or hormonal influences, which can significantly impact performance in certain sports.

Ensuring Fairness and Reducing Bias: Apply strategies to minimize bias in the models that may disproportionately favor one gender. This can be achieved by employing techniques like data balancing to ensure equitable representation across genders.

C. Age Distribution

Identifying the Optimal Age for Performance: Analyze athlete performance data across various sports and age groups to determine the age range where peak performance is typically achieved in each sport.

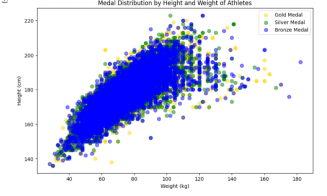
Influence of Training and Technology: Use machine learning models to evaluate how advancements in training techniques and sports science may affect the age range at which athletes reach peak performance, especially over time.

Forecasting Future Age Trends: Create models that predict potential changes in the age distribution of Olympic athletes. This can help anticipate shifts in training methods or adjustments in talent identification and development strategies.

Height vs. Weight

In the Olympic Games, there are no overall restrictions on the height or weight of athletes, although some sports may have specific weight classes or height criteria. While height and weight can influence performance in certain sports, using machine learning to predict Olympic outcomes requires a more nuanced perspective.

In summary, while these physical attributes may impact an athlete's performance in specific events, there are no general height or weight requirements for participation in the Olympics.



**V. APPLICATIONS OF MACHINE LEARNING ALGORITHMS FOR PREDICTION**

APPLICATIONS OF MACHINE LEARNING ALGORITHM FOR PREDICTION  
A. Logistic Regression

Logistic regression is a widely used technique for analyzing and predicting outcomes related to the Olympic Games. Its applications include the following:

Binary Outcomes:

Logistic regression is particularly effective for scenarios involving two possible outcomes, making it ideal for predictions such as:

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

Will a team win or lose a game?

Probability Estimates:

Instead of providing a definitive win or loss, logistic regression predicts the likelihood of an outcome. For example, it can estimate the probability of a particular country winning a medal in an event, offering a more detailed perspective.

Influencing Factors: The model allows the inclusion of various factors that might affect the results. Examples include:

Historical performance of teams or athletes.

World rankings.

Geographic or population size (helpful in forecasting medal counts).

Limitations: Despite its usefulness, logistic regression has certain constraints:

Data Dependence:

The model's accuracy relies heavily on the quality

Unpredictable Events:

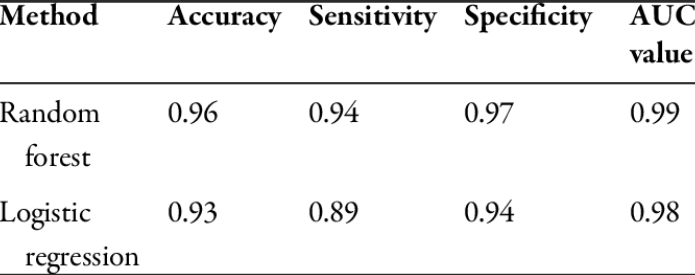
Unexpected factors like injuries,randomness.

B. Random Forest

Random Forest is a widely used machine learning algorithm introduced by Leo Breiman and Adele Cutler. Now the it works by aggregating the predictions of multiple decision trees to generate a unified outcome. Its simplicity, flexibility, and ability to handle both classification and regression tasks make it a popular choice.

Unlike relying on a single decision tree, Random Forest given uses predictions from all the trees in the ensemble, resulting in a more robust and accurate final prediction. This grouped collective approach significantly minimizes the risk of overfitting and leads enhances the model's performance.It works by randomly selecting subsets of data (bootstrapping) and features (feature bagging) to train each tree independently, reducing overfitting. The final prediction is obtained through majority voting (classification) or averaging (regression), making it highly effective for handling large datasets, missing values, and non-linear relationships.

Comparison of Logistic Regression and Random Forest



The Random Forest algorithm offers superior accuracy and resilience against outliers while reducing overfitting risks. However, its complexity and computational requirements can make it harder to interpret compared to Logistic Regression, which is simpler and faster but more susceptible to overfitting and outlier effects.

**VI. CONCLUSION**

CONCLUSION  
The analysis of Olympic data has provided valuable insights into historical patterns and the factors that influence performance. While historical data serves as a solid foundation for making predictions, the Olympic Games remain a dynamic event influenced by unpredictable elements such as injuries, rising stars, and the advantages associated with hosting.

Despite these uncertainties, data-driven analysis proves to be an essential resource for athletes, coaches, and sports organizations. By examining historical trends and pinpointing critical factors, stakeholders can design focused training programs and strategic plans to enhance performance.

In essence, the Olympic Games represent the peak of athletic excellence, uniting a global audience through extraordinary demonstrations of human capability and inspiring future generations.

**REFERENCES**

[1] Y. A. Ahmed, M. N. Ahmad, N. Ahmad, and N. H. Zakaria, "Social media for knowledge-sharing: a systematic literature review," Telematics and Informatics, vol. 37, pp. 72–112, 2019.

[2] A. Fink, Conducting Research Literature Reviews: From the Internet to Paper, 5th ed. Los Angeles, CA: Sage, 2019.

[3] B. Houlihan, B. Leopkey, and E. MacIntosh, "Introduction: youth, sport and the youth olympic games," Int. J. Sport Manage. Mark., vol. 4–6, pp. 237–241, 2017.

[4] S. Bellavia, M. Macconi, and B. Morini, "An aMne scaling trust-region approach to bound-constrained nonlinear systems," Applied Numerical Mathematics, vol. 44, pp. 257–280, 2003.

[5] Kaggle Dataset: 120 Years of Olympic History – Athletes and Results. [Online]. Available: https://www.kaggle.com/datasets/heesoo37/120-years-ofolympic-history-athletes-and-results