# Olympic Data Analysis Using Machine Learning

Hause Manali Uddhav[1], Hule Pranjali Rajendra[2], KavatheVaishanvi Hanmant[3], Prof. S. G. Ekdante[4], Prof. J. M. Shaikh[5]

[1,2,3]*Dept. Computer Science and Engineering, Shri.Tulajabhavani college of Engineering, Tuljapur*

[4]*Guide, Dept. Computer Science and Engineering, Shri.Tulajabhavani college of Engineering, Tuljapur*

[5]*Co-Guide, Dept. Computer Science and Engineering, Shri.Tulajabhavani college of Engineering, Tuljapur*

**Abstract- The Olympic Games stand as a pinnacle of international competition and a source of immense pride for participating countries. Consequently, each nation endeavors to deliver its utmost performance during the event. Despite concerted efforts, numerous countries and athletes often fall short of securing medals, while others excel, amassing a substantial medal haul.**

**It's imperative for every country to conduct a meticulous analysis of past statistics to discern areas of improvement and rectify past mistakes. This introspection aids in future development and strategic planning, fostering enhanced performances in subsequent editions of the Games.**

**Keywords: Machine Learning, Kaggle, Notebook.**

## I. INTRODUCTION

The Olympic Games are the world's largest international sporting event, held every four years with participation from more than 200 nations. Over time, the Olympics have evolved in several dimensions, including the number of participating countries and athletes, changes in the total events, improvements in athlete performance, growth in women's participation, and variations in the male-to-female ratio. Additional factors such as hosting expenses, advancements in medical facilities, and global conditions like pandemics also influence Olympic outcomes. Analyzing these elements offers valuable insights into the historical evolution of the Games and assists in predicting future trends.[1]

## II. LITERATURE REVIEW

Previous research on Olympic data analysis has mainly focused on predicting medal outcomes, analyzing athlete performance, and studying country-wise trends using historical datasets.

Machine learning models such as Linear Regression, Logistic Regression, Decision Trees, and SVM have been used to forecast a nation's chances of winning medals. Studies also show that athlete attributes like age, height, weight, and past performance strongly influence medal probability. Exploratory Data Analysis (EDA) is widely used to visualize trends in participation, sports growth, gender ratios, and event evolution. While existing systems provide strong visual insights, many lack advanced prediction capabilities and deeper pattern identification. This highlights the need for an improved ML-based approach for Olympic performance analysis.[2]

## III. METHODOLOGY

The methodology adopted for Olympic data analysis using machine learning is structured into three major phases: Data Collection, Data Pre-processing, and Exploratory Data Analysis. Each phase contributes to preparing and understanding the data before applying advanced analytical techniques.

### A. Data Collection

This phase involves gathering datasets required to analyze historical Olympic trends.

- Multiple publicly available Olympic datasets were used to ensure completeness and reliability.
- Athlete dataset includes personal and performance attributes such as gender, age, height, weight, nationality, and medals won.
- Medal tally dataset provides country-wise medal counts across all Olympic years.
- Country code dataset helps uniquely identify participating nations.

- The combination of these datasets provides high-volume, diverse, and structured information essential for accurate analysis.

B. Data Pre-Processing

Raw data often contains inconsistencies and missing values, making preprocessing a crucial step.

- Cleaned and standardized all datasets to remove inconsistencies and incorrect entries.
- Missing numerical values were replaced using Mean/Median imputation for accuracy.
- Missing categorical values were filled using Hot Deck Imputation, ensuring similarity-based replacements.
- Duplicate records were removed to prevent bias in analysis.
- Final cleaned dataset ensures reliability and suitability for visualization and machine learning operations.

C. Exploratory Data Analysis (EDA)

EDA is conducted to extract initial insights, identify trends, and understand the data structure.

- Visual tools such as histograms, bar charts, line plots, scatter plots, and heatmaps were used.
- Trends analyzed include participation growth, medal distribution, event expansion, and gender representation over the years.
- Athlete characteristics (age, height, weight) were studied to observe performance-related patterns.
- EDA provides a clear understanding of Olympic evolution and forms the basis for future ML-based prediction.[3]

## IV. PROPOSED SYSTEM

The proposed system provides a structured analytical workflow to study the evolution and performance dynamics of the Olympic Games using machine learning and exploratory data analysis. The system is designed to ensure accuracy, scalability, and clear visualization of historical Olympic data.

A. Overview

The system aims to analyze Olympic datasets by following a systematic pipeline that transforms raw data into meaningful insights.

- Provides a standardized approach to handle multiple datasets.
- Ensures efficient preprocessing, analysis, and visualization of athlete and country performance.
- Helps identify long-term trends, participation patterns, and medal distributions.
- Serves as a foundation for advanced predictive modelling in future stages.

B. Proposed Workflow

The functioning of the system is structured into four major phases:

- Data Acquisition: Collects athlete, country, and medal datasets from reliable sources such as Kaggle.
- Data Pre-processing: Cleans, transforms, and integrates datasets by handling missing values and ensuring consistency.
- EDA and Visualization: Uses graphical techniques to observe patterns in participation, gender distribution, event growth, and medal performance.
- Machine Learning Integration: (Optional Future Scope) Applies basic ML models to identify key performance indicators and predict medal trends.

C. System Flow Diagram (Figure 4.1)

The system workflow is represented in a simple linear structure:

1. Input Data →
2. Data Cleaning & Pre-processing →
3. Exploratory Data Analysis →
4. Visualization & Interpretation

Each step processes the output of the previous one, ensuring smooth data transformation and consistent results.

D. Key Advantages of the Proposed System

- Provides clear and structured insights into Olympic performance trends.
- Allows comparative analysis between athletes, countries, and sports categories.
- Enhances decision-making for researchers, sports analysts, and policymakers.
- Creates a foundation for future ML-based prediction systems.[4]

## V. SYSTEM ARCHITECTURE

The system architecture establishes the flow of Olympic data from input to analysis and visualization using a structured, modular approach.

A. Architecture Overview

- Provides a streamlined workflow for handling large Olympic datasets.
- Ensures smooth data transformation and supports future scalability.
- Organizes tasks into independent, efficient modules.

B. Main Components

1. Data Source Layer
- Collects athlete data, medal records, and country codes from reliable sources.

2. Data Pre-processing Layer
- Cleans raw data, removes duplicates, handles missing values, and standardizes formats.

3. EDA Layer
- Generates visual insights through histograms, bar charts, scatter plots, and heatmaps.

4. Machine Learning Layer (Future Scope)
- Applies models like Linear Regression, Logistic Regression, and Decision Trees for predictions.

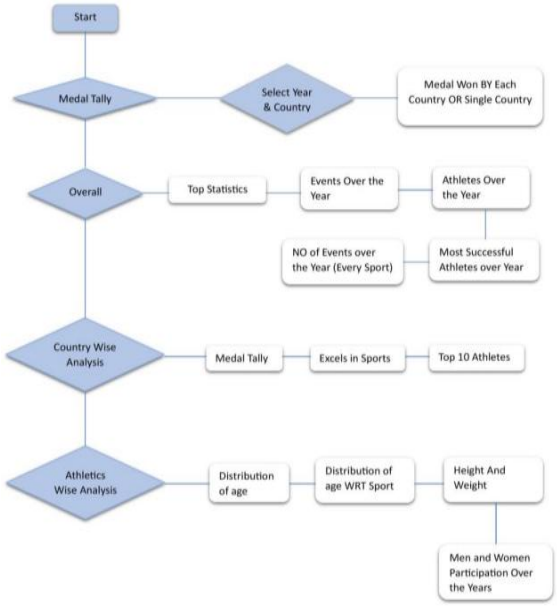
5. Visualization Layer
- Produces dashboards and charts showing medal trends, participation growth, and performance patterns.

C. System Workflow

1. Data Acquisition
2. Data Cleaning
3. EDA Insight Generation
4. Visualization
5. Optional ML Prediction

D. Benefits

- Efficient data handling and accurate analysis.
- Highly scalable and easy to expand with ML models.
- Produces clear visual outputs for better interpretation.[5]




## VI. RESULTS

Overall Medal Tally

- Shows long-term performance of each country from 1896–2016.
- USA consistently ranks the highest in total medals.
- Australia, Japan, France, and Hungary show varying contributions across years.

Top Statistics

- Olympic Games expanded in host nations, sports count, and event numbers.
- Shows continuous growth and diversification of Olympic competitions.

Heatmap of Events

- Athletics grew from 12 events (1896) to 47 events (2016).
- Wrestling and other sports also expanded significantly over time.

Age Distribution

- Displays athlete participation across different age groups.
- Helps identify preferred sports by age and recruitment opportunities.

Athletes Over the Years

- India improved from 0 medals (1992) to 6 medals (2016).

- USA shows fluctuating but overall dominant medal performance.
- Australia and Japan show sharp variations between Olympics.

Country-wise Medal Analysis
- USA leads with maximum gold, silver, and bronze medals.
- Australia and Japan show fewer golds but strong overall participation.
- France maintains balanced medal contributions.

Height vs Weight Analysis
- Most female medalists are 160–180 cm and 50–150 kg.
- Gold medal winners show height clustering around 175–180 cm.

Men vs Women Participation
- Male participation historically higher than female.
- Gender gap decreases over the years, showing improved inclusivity.

Country-wise Sports Performance
- USA excels in swimming; France performs strongly in wrestling.
- Country strengths vary based on sport and year.

Most Successful Athletes
- Michael Phelps leads with 28 Olympic medals.
- Higher athlete participation → higher total medal count for a country.[6]

## VII. CONCLUSION

- Study analyzes Olympic performance using ML and EDA.
- Shows how countries and athletes evolved over time.
- Highlights expansion of sports, participation, and competitiveness.
- Useful for researchers, sports authorities, and policymakers.
- Supports better planning for future Olympic performance strategies.[7]

## REFERENCES

[1] Kabita Paul, Elif Demir, Anjali Bapat: Olympic Data Analysis Project (May 2019)

[2] Sacha Schmidt, Limas, Wunderlich, Dominik Schreger (December 2020): Olympic Data Analysis Project

[3] Pradhan, Rahul & Agrawal, Kartik & Nag, Anubhav. (2021). Analyzing Evolution of the Olympics by Exploratory Data Analysis using R.

[4] Cutait, M.: Management performance of the Rio 2016 Summer Olympic Games. Research Paper submitted and approved to obtain the Master's degree in Sports Administration at AISTS in Lausanne, Switzerland.

[5] Moreno A, Moragas M and Paningua R 1999 The evolution of volunteers at the Olympic Games Proceedings of Symposium on Volunteers (Lausanne, Switzerland: Global Society and the Olympic Movement) pp 1–18

[6] Abeza G, Braunstein-Minkove J R, S´eguin B, O'Reilly N, Kim A and Abdourazakou Y 2020 Ambush marketing via social media: The case of the three most recent Olympic Games Int. J. Sport Communication 1–25

[7] Yamunathangam D, Kirthicka G and Shahanas P 2018 Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques International Journal of Recent Technology and Engineering (IJRTE) 7 251–3

[8] Wikipedia contributors: Exploratory data analysis, https://en.wikipedia.org/wiki/Exploratory data analysis, last accessed 2020/11/11.