Summer Olympics 1976-2008

LINKEDIN

GITHUB

Presented by: Mohammad Amil Khan

Project Overview & Problem Statement

Problem Statement:

This project analyzes Summer Olympics data (1976–2008) to uncover trends in athlete participation, medal distribution, and country-wise performance. It also aims to predict medal types using machine learning models.

Objective:

Explore trends in athlete participation and gender distribution.

Identify top countries and sports by medal count.

Predict medal types using Logistic Regression and Random Forest.

Improve model performance with hyperparameter tuning.

Recommend future enhancements using advanced models and richer features.

Dataset Overview

Column Name	Description
City	Host city where the Olympic Games were held
Year	Year of the Olympic event
Sport	General category of the sport (e.g., Aquatics, Athletics)
Discipline	Specific discipline within the sport (e.g., Diving under Aquatics)
Event	Specific event name (e.g., 3m springboard, 100m freestyle)
Athlete	Full name of the athlete who participated
Gender	Gender of the athlete (Men or Women)
Country_Code	Country code abbreviation (e.g., USA, CHN)
Country	Full name of the athlete's country
Event_gender	Gender category of the event (M or W)
Medal	Medal won (Gold, Silver, Bronze, or missing if no medal)

Process

Data Prep: Cleaned dataset, encoded categories, extracted key fields

Participation Trend: Analyzed yearly athlete counts

Gender Analysis: Compared male vs female participation growth

Top Performers: Identified top countries and sports by medals

Medal Timeline: Tracked medal counts over years

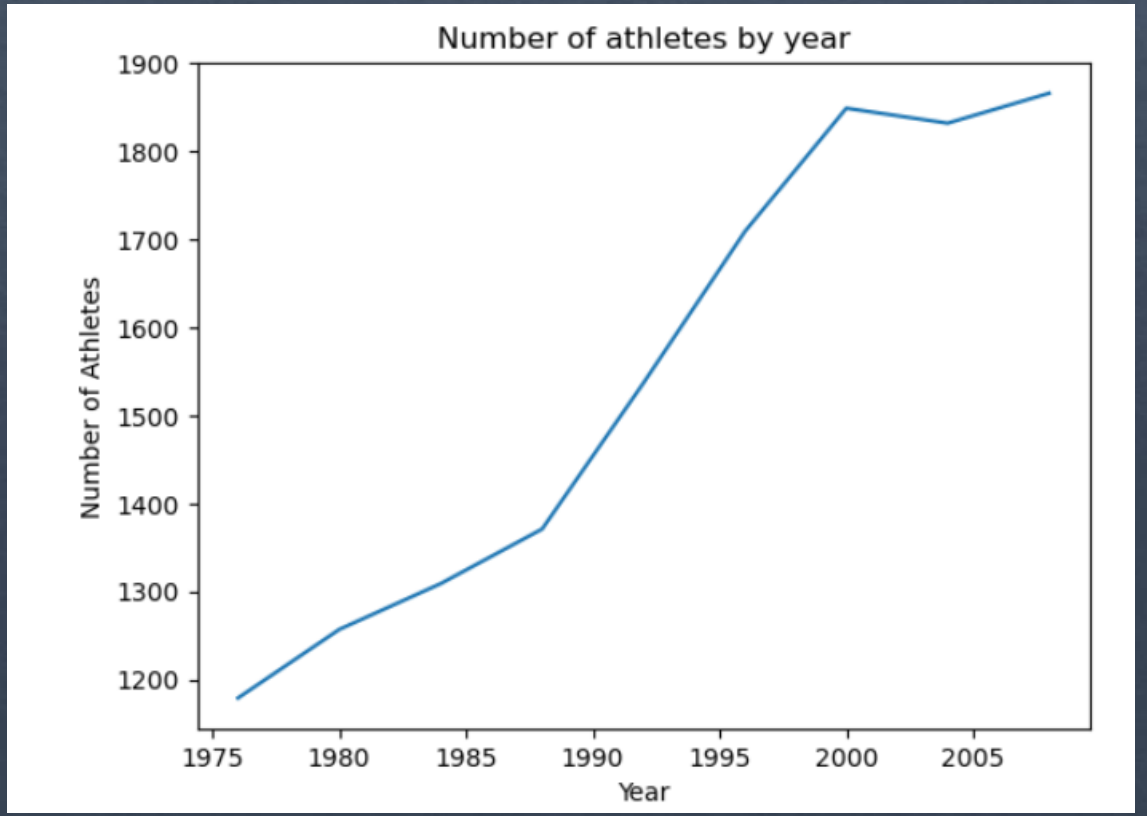
Modeling: Built Logistic Regression & Random Forest models

Performance: Random Forest achieved ~53% accuracy

Goal: Uncover trends & predict medal outcomes for strategic insights

1. What has been the increase in number of athletes over time?

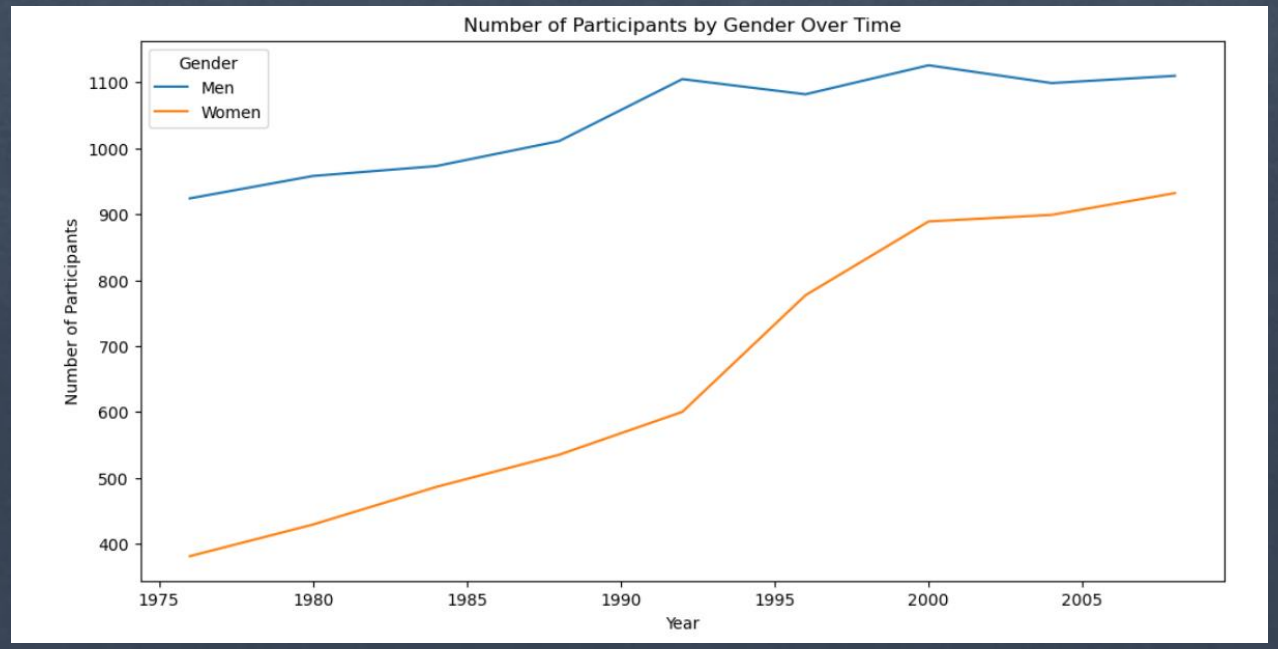
Insight: The number of Olympic athletes steadily increased from 1976 to 2008, highlighting growing global participation.



2. What has been the increase in participating atheletes over time by gender?

Answer: Consistent growth in both male and female Olympic participation from 1976 to 2008, with a sharper rise in female athletes.

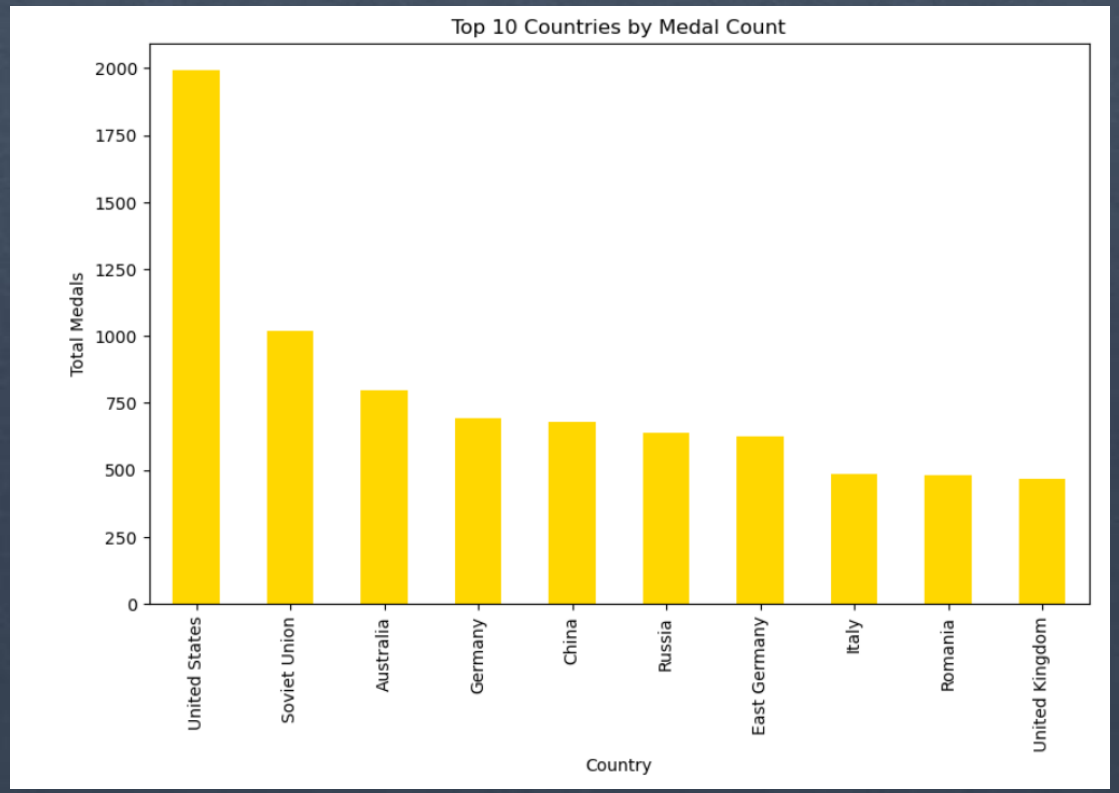
Insight: Female participation nearly doubled, reflecting increasing gender inclusivity. The narrowing gap highlights a positive shift toward gender balance in global sports.



3. Top 10 Countries by Medal Count

Answer: The United States leads by a wide margin in total Olympic medals, followed by the Soviet Union and Australia.

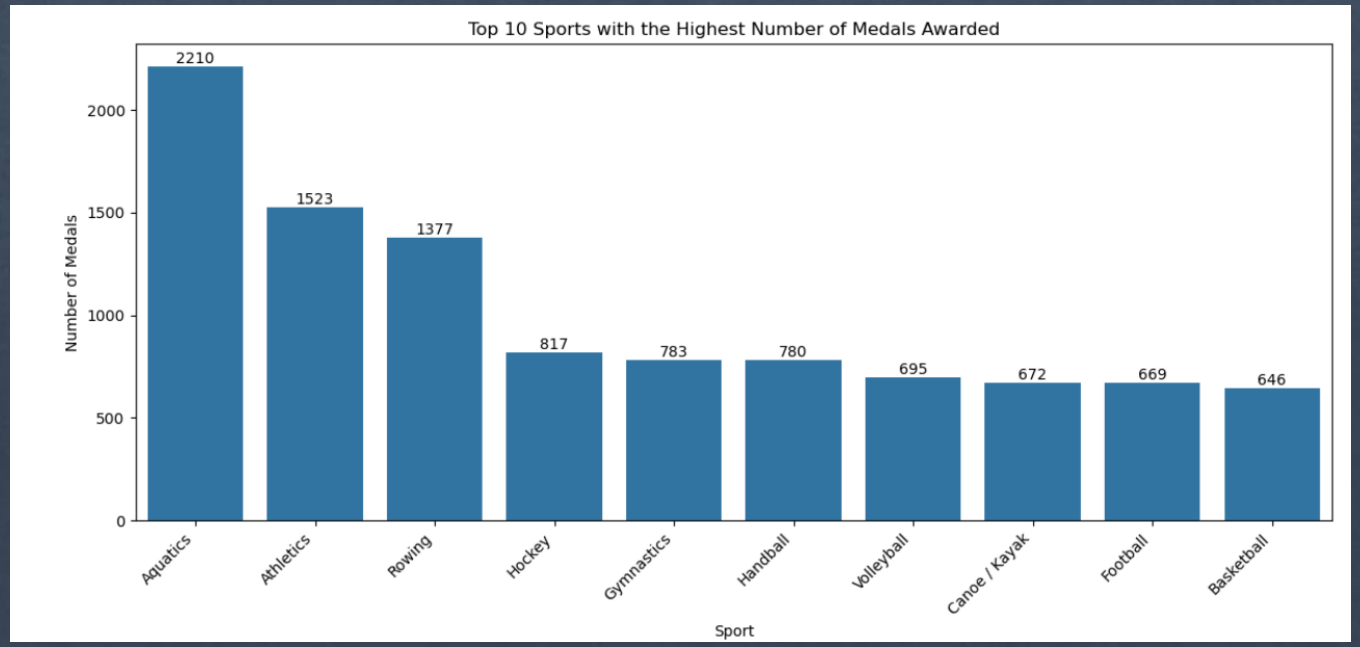
Insight: The U.S. dominance highlights sustained performance across decades. Other top countries show strong but narrower medal margins, reflecting regional athletic strengths.



4. What sports have have the highest number of medals being awarded?

Answer: Aquatics, Athletics, and Rowing are the top three sports with the highest number of Olympic medals.

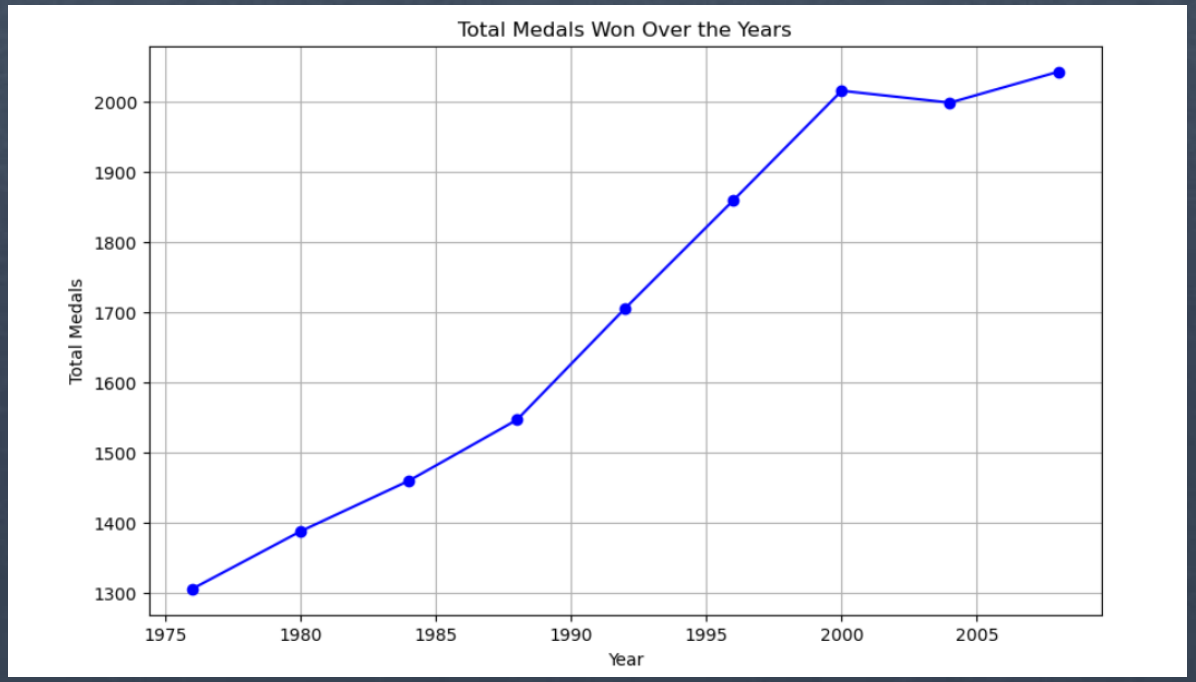
Insight: Aquatics leads with over 2,200 medals, indicating its broad event range and global competitiveness; other top sports show strong historical presence.



5. Total Medals Won Over the Years

Answer: Total Olympic medals increased steadily from 1976 to 2008, with minor dips after 2000.

Insight: The upward trend reflects rising participation and event expansion; the slight post-2000 dip may indicate competitive saturation or structural limits.



Best Model Random Forest

1. **Best Accuracy Achieved:** ~53.3% using tuned hyperparameters.
2. Best Parameters: max_depth=20, n_estimators=200, no bootstrap.
3. Class 1 (Gold) had the highest recall (0.60) and F1-score (0.56).
4. Class 2 (Silver) showed the weakest recall (0.47), indicating some underperformance.
5. Class 3 (Bronze) performed moderately with balanced precision/recall.
6. Macro & Weighted Averages: All at 0.53, suggesting uniform model behavior across classes.
7. Confusion Matrix shows good but not perfect separation—room for improvement.
8. Model is balanced, but additional features or class balancing may boost accuracy.

```
Best Parameters: {'bootstrap': False, 'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
Accuracy: 0.5331882480957563
Confusion Matrix:
 [[936 288 331]
 [424 715 372]
 [401 329 799]]
Classification Report:
              precision    recall  f1-score   support

           1       0.53      0.60      0.56      1555
           2       0.54      0.47      0.50      1511
           3       0.53      0.52      0.53      1529

    accuracy                           0.53      4595
   macro avg       0.53      0.53      0.53      4595
weighted avg       0.53      0.53      0.53      4595
```

Conclusion

The Olympics dataset highlights increasing global participation and growing gender diversity. The United States and Aquatics dominate in medal counts. For medal prediction, Random Forest outperforms Logistic Regression, but both are limited by basic features. Enhanced data and advanced models can significantly improve prediction and insight.