Olympics Data Analysis

Project Title: Olympics Data Analysis

Domain: Data Analytics & Predictive Modeling

Tools & Technologies Used: Python, Pandas, Matplotlib, Seaborn, NumPy, Scikit-learn, Jupyter Notebook

Dataset: Summer Olympics Medals Dataset (1976 - 2008) - Medal winners from Montreal 1976 to Beijing 2008

Project Objective

- Explore and analyze historical data of Olympic medalists.
- Identify key trends across countries, sports, athletes, and gender.
- Use data-driven analysis to uncover hidden insights.
- Enhance decision-making using visualization.
- Build a model to predict whether an athlete is likely to win a medal using machine learning techniques.

Data Cleaning and Preprocessing

- Dropped unnecessary columns like Event_gender and Country_Code.
- Removed fully null rows (117 entries).
- Converted Year to integer data type.
- Checked and handled null values.

Exploratory Data Analysis (EDA)

- Analyzed Olympic hosting patterns and identified cities with the highest number of events.
- Found [top countries] to be the most consistent medal winners during 1976–2008.
- Observed gender participation trends, showing an increase in women's medal counts over time.
- Identified sport-wise country dominance (e.g., Korea in Archery, Australia in Swimming).
- Created visualizations to show medal trends, gender ratios, and top-performing athletes.

Predictive Analysis:

Model Used: Logistic Regression

Features Used:

- Country
- Sport
- Gender

Enhancements & Improvements:

- Dropped non-informative columns to improve model generalizability. - Label Encoded categorical features. - Combined 'Gold', 'Silver', 'Bronze' into a single binary target (1 = won a medal, 0 = no medal). - Trained and evaluated using train_test_split (70-30).

- Evaluation Metrics: - Accuracy Score - Confusion Matrix - Classification Report
- Model Outcome: - Reasonable accuracy in predicting medal winners based on limited features. - Can be further improved using athlete age, past records, country GDP, etc.

Insights & Conclusion

- USA and Russia were the most dominant countries overall.
- Certain sports/events are gender-exclusive or skewed.
- Michael Phelps stands out as the top-performing athlete.
- Visualizations helped uncover trends like city-wise hosting and sport-wise dominance.
- Predictive modeling demonstrated the possibility of anticipating medal wins, with scope for improvement.

Skills Demonstrated

- Data Cleaning & Transformation
- Exploratory Data Analysis (EDA)
- Data Visualization (Matplotlib, Seaborn)
- Machine Learning (Logistic Regression)
- Analytical Thinking & Interpretation

Challenges Faced & Overcome

- Incomplete data entries: resolved by dropping fully null rows.

- Non-uniform formatting: cleaned using pandas.
- Duplicated athlete names across events: noted and acknowledged.
- Modeling with limited features: simplified to demonstrate binary prediction.

Future Enhancements

- Integrate athlete physical metrics (age, height, weight).
- Incorporate country-wise sports infrastructure & funding.
- Build an interactive dashboard using Power BI or Tableau.