Task – 3 Olympic Data Analysis

Unified Mentor PVT. LTD.

Project Name: E-commerce Furniture Dataset 2024 (Unified Mentor)

Tools Used ==>> Python (google colab).

About Dataset:

This dataset is a list of all the medal winners in the Summer Olympics from 1976 Montreal to 2008 Beijing. It includes each and every medal awarded within the period. This dataset is intended for beginners so that they can get a taste of advanced Excel functions which is perhaps one of the key skills required to be a great data scientist. I too got my hands dirty with the dataset and played with some advanced Excel functions. Further, this dataset can also be used for a predictive model as to which country is likely to fetch the highest number of gold in a particular sports category (just an example), etc.

Columns:

- **City: The city where the Olympics took place.**
- **Year: The year of the Olympics.**
- **Sport: The sport the event is categorized under.**
- **Discipline: A subcategory of the sport.**
- **Event: The specific event within a discipline.**
- **Athlete: The name of the athlete who participated.**
- **Gender: The gender of the athlete.**
- **Country_Code: The country code (abbreviation).**
- **Country: The full name of the country.**
- **Event_gender: The gender category of the event.**
- **Medal: The medal won (Gold, Silver, Bronze).**

Objective:

The primary goal is to:

1. Analyze the dataset to understand trends in medal distribution.

2. Identify the top-performing countries and athletes.

3. Study the gender distribution of events and medals.

4. Visualize the data using Python.

The goals were to:

• Study trends in medal distribution.

• Identify top-performing countries and athletes.

• Analyze gender distribution in sports and events.

• Build a predictive model to forecast medal outcomes.

I used Python with libraries like pandas, matplotlib, seaborn, and scikit-learn to perform the analysis.

Step Performed:

- Data Cleaning
- EDA (Exploratory Data Analysis)
- Predictive Analysis
- Conclusion

1. Data Cleaning

• Loaded the dataset using pandas and checked its structure (15,433 rows, 11 columns).

• Removed 117 duplicate rows to ensure accurate counts.

• Dropped 117 rows with missing values, as they were a small portion (0.7%) of the data.

• Converted the Year column from decimals (e.g., 1976.0) to integers (e.g., 1976) for clarity.

2. EDA (Exploratory Data Analysis)

• Generated summary statistics:–

- **Year ranged from 1976 to 2008 (mean: 1993.62)**
- **The USA and Aquatics (Swimming) appeared most frequently.**

• Plotted the number of medals per year to study trends.

• Identified the top 5 countries and athletes by medal count.

2.1Visualizations

• Created a bar plot of the top 5 and top 10 countries by medal count.
• Plotted the top 10 sports by medal count in 2008.
• Made a stacked bar plot to show medal distribution by sport and gender.

3. Predictive Analysis:

• Built a Random Forest model to predict if an athlete wins a Gold medal (1) or Sil ver/Bronze (0).
• Used features: Country, Sport, Discipline, Event, Gender.
• Converted categorical data to numbers using LabelEncoder.
• Split data into 80% training and 20% testing sets (3,063 test cases).

4. Results:

- Medal Trends: The line plot showed varying medal counts over years, likely due to changes in events or participating countries.

- Top Countries: The USA and Soviet Union led in medals, reflecting strong sports programs.

- Top Athletes: Athletes like Michael Phelps (if present) won multiple medals, espe cially in Aquatics.

- Gender Distribution: The stacked bar plot showed balanced gender participation in some sports (e.g., Gymnastics) but male dominance in others (e.g., Wrestling).

4.1Predictive Model Performance

- Accuracy:
71.2%, meaning the model correctly predicted Gold vs. Silver/Bronze 71.2% of the time.

- Precision and Recall:
- Silver/Bronze (0): 78% precision, 79% recall (2,065 cases).
– Gold (1): 56% precision, 55% recall (998 cases).

- The model performed better for Silver/Bronze due to more data in that category

4.2Key Learnings:

- Data Cleaning: I learned to handle duplicates and missing values to ensure accurate analysis.

- EDA:Summarizingdata andplotting trends helped me understand Olympic patterns, like which countries dominate.

• Visualizations: Creating plots with seaborn and matplotlib made complex data easy to interpret.

• Predictive Modeling: Building a Random Forest model taught me how to predict outcomes and evaluate performance using metrics like accuracy (71.2%) and preci sion/recall.

• Real-World Applications: The analysis can help Olympic committees decide which sports to fund or coaches choose athletes likely to win Gold.

5. Conclusion:

This project allowed me to apply data science skills to a real-world dataset. I cleaned the Olympic data, analyzed trends, visualized key insights, and built a predictive model with 71.2% accuracy. The analysis revealed the dominance of countries like the USA and sports like Aquatics, as well as gender participation patterns. The model, though not perfect, shows potential for predicting Gold medal outcomes.