120 YEARS OF OLYMPIC DATA ANALYSIS

Best athletes and medal prediction in modern Olympic games.

Image 1

HDSC Spring '23 Premiere Project - Team Jupyter

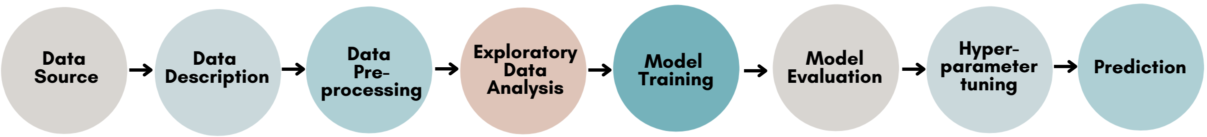
INTRODUCTION

Given a dataset containing 120 years of Olympic data, the problem at hand is to analyze Olympic athlete data and develop a predictive model to forecast the likelihood of winning a medal in future Olympic Games.

OBJECTIVES

The Olympic Athletes Analysis project aims to analyze the performance of athletes in modern Olympic games and offer a tool for predicting medal outcomes, thereby assisting stakeholders, including athletes, coaches, and sports enthusiasts, in making informed decisions and setting realistic expectations for future Olympic events.

FLOW PROCESS



Data Source → Data Description → Data Pre-processing → Exploratory Data Analysis → Model Training → Model Evaluation → Hyper-parameter tuning → Prediction

Image 2

DATA SOURCE - Kaggle

120 years of Olympic history: athletes and...

basic bio data on athletes and...

kaggle.com

DATASET DESCRIPTION

The datasets (df1 & df2) used in this project were sourced from Kaggle. It includes a comprehensive collection of data about athletes who participated in various Olympic Games. The datasets contain the following information:

- ID
- Name
- Sex
- Age
- Height
- Weight
- Team
- NOC — three-letter code for each country established by the National Olympics Committee
- Games
- Year
- Season
- City
- Sport
- Event
- Medal
- Region
- Notes

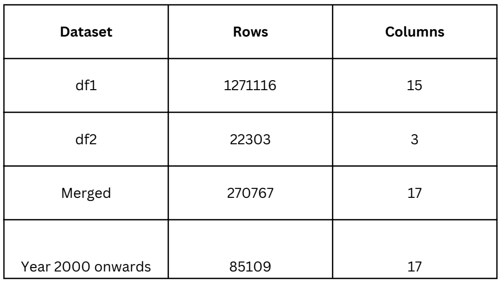
DATA PRE-PROCESSING

During the data loading and pre-processing stage, we aimed to maximize the analysis by merging the provided datasets, **df1** and **df2**, based on the common column '**NOC**'. By merging these datasets, we combined the relevant information from both sources, allowing us to leverage the shared knowledge and perform a comprehensive analysis using the merged dataset.

To ensure that our analysis reflects a more modern perspective, as indicated in the problem statement, we focused specifically on data and features from the year **2000** onwards. This means that we considered only the information that is relevant and available from the year 2000, excluding any earlier years.

By incorporating these steps into the data loading and processing stage, we set the foundation for a more focused and contemporary analysis (EDA), enabling us to gain valuable insights from the combined dataset.

Data size



Dataset	Rows	Columns
df1	1271116	15
df2	22303	3
Merged	270767	17
Year 2000 onwards	85109	17

Image 3

EXPLORATORY DATA ANALYSIS

During the exploratory data analysis (EDA) stage, we conducted various analysis and visualizations to gain insights into the Olympic athletes' data. Some of the key aspects we explored include:

1. **Top 10 countries with the most medals:** It is evident that the USA stands out as the dominant country in the analysis below.

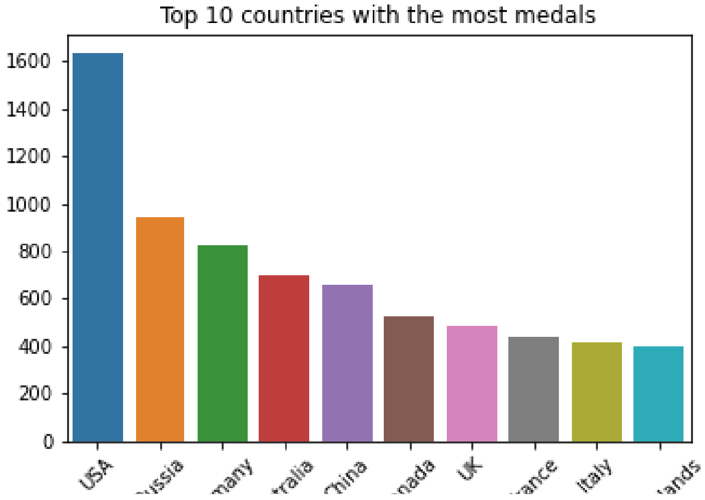


Image 4

2. Top 3 countries with the highest total medal in 16 years:

- This plot provides a **summary of the number of medals won by the top countries** ('USA', 'Russia', 'Germany') **over the years**.
- It allows for easy comparison of medal counts between these countries and can help identify trends or patterns in their performances over time.

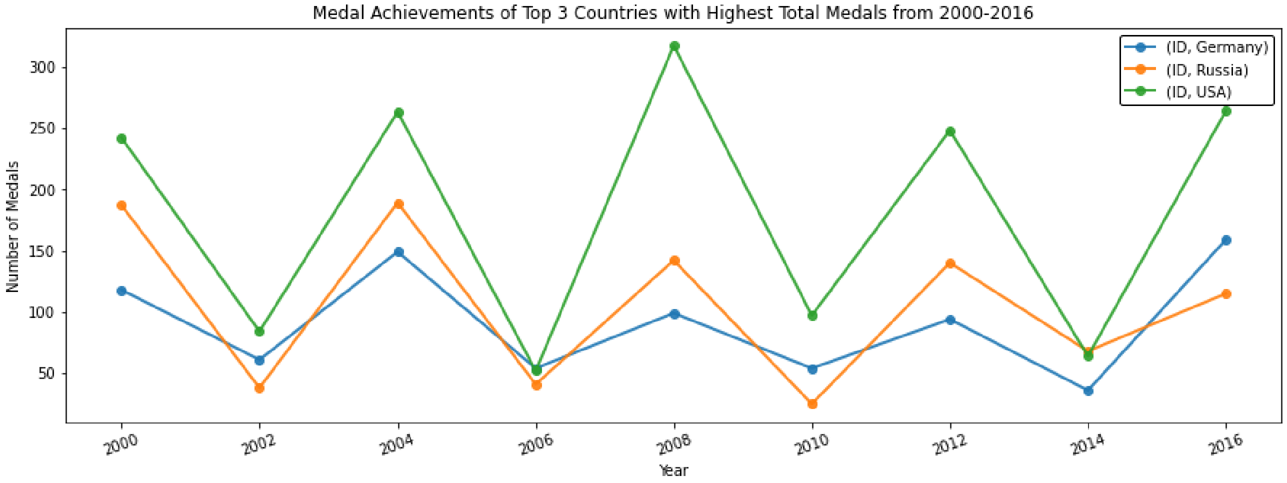


Image 5

3. Top African countries with most medals:

- The bar plot below displays the **top 5 African countries** with the **highest medal counts**.
- It provides a visual comparison of the medal counts among the selected African countries.

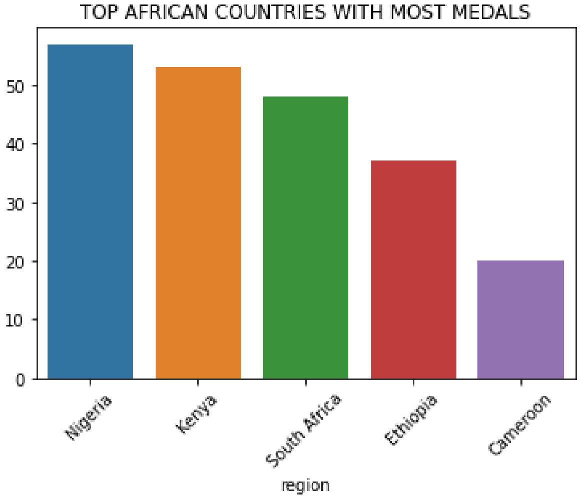


Image 6

4. Top 20 countries with the most GOLD medal:

- This bar plot visualizes the **top 20 countries** with the **most gold medals**.
- Each country is represented by a bar, and the height of the bar corresponds to the number of gold medals won by that country.

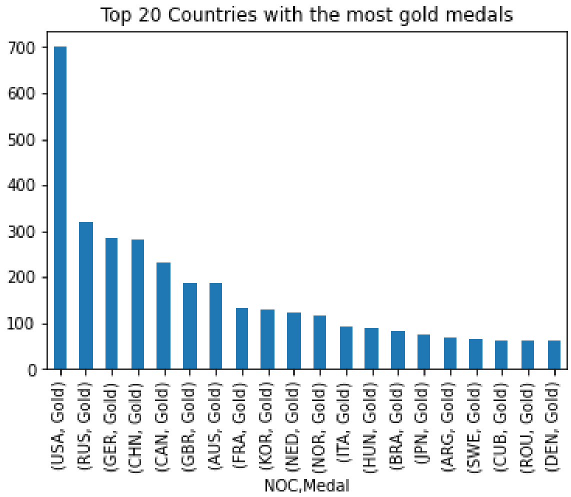


Image 7

5. Distribution of 'Age', 'Height', 'Weight' columns

According to these plots below:

- Athletes in their **mid-20s** tend to have a higher probability of achieving success and winning gold medals in the Olympic Games.
- The fact that the heights are mostly **over 180 cm** suggests that there is a prevalence of taller athletes among the top 20 countries.
- It appears that the weight distribution among athletes from the top 20 countries is concentrated around **70 kg**

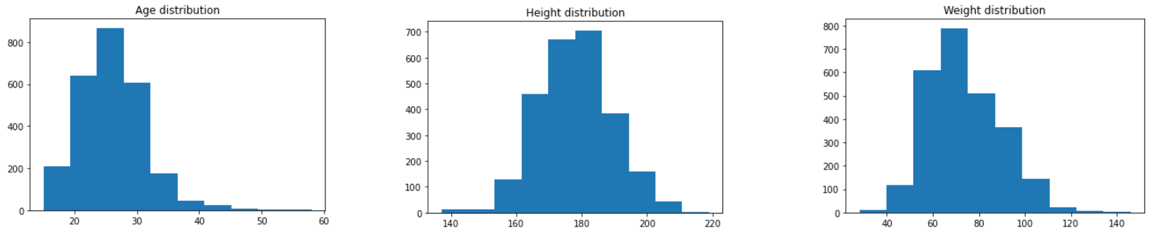


Image 8

6. **Distribution of 'Age-category', 'Height-category', 'Weight-category' of the top 10 countries:**

- The plot below shows the number of athletes in each Age category, height category and weight category for each country, with the countries differentiated by color.

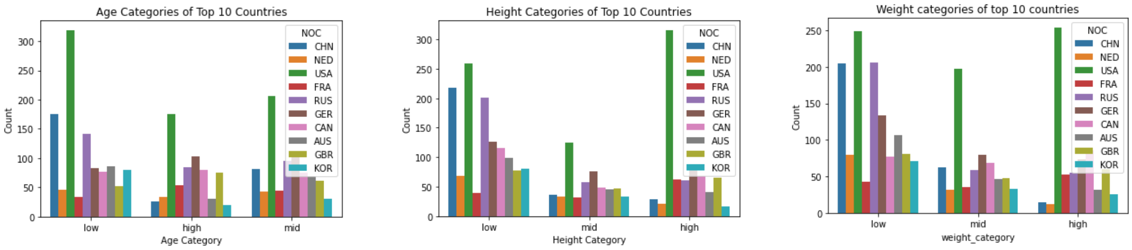


Image 9

MODEL TRAINING

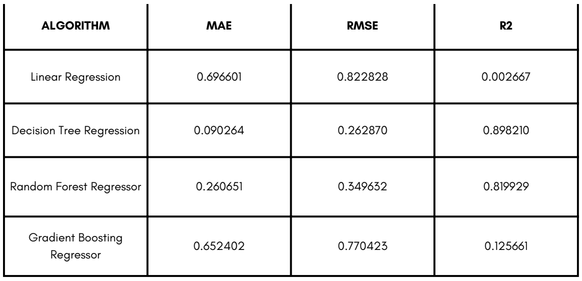
During the model training stage, we prepared the data for modeling and trained various machine learning algorithms to make predictions or gain deeper insights from the Olympic athletes' data. The key steps involved in this stage are as follows:

- **Data Preprocessing**: We handled missing values, applied data transformations such as encoding to prepare the data for modeling. This ensured that the input features were in a suitable format and range for the machine learning algorithms.
- **Feature Selection**: We selected the relevant features or variables that would be used as input for the models.
- **Splitting the Data:** We divided the dataset into training and testing sets. The training set was used to train the models, while the testing set was kept separate to evaluate the performance of the trained models on unseen data.
- **Model Selection:** We experimented with different machine learning algorithms such as linear regression, decision trees, random forests, and gradient boosting.
- **Model Training:** We trained the selected models using the training data. This involved fitting the models to the input features and their corresponding target variable.

MODEL EVALUATION

We assessed the performance of the trained models using appropriate evaluation metrics such as mean absolute error (**MAE**), root mean squared error (**RMSE**), and R-squared (**R2**) score. This allowed us to compare the performance of different models and identify the best-performing one(DTR-model).

Algorithms and Evaluation



ALGORITHM	MAE	RMSE	R2
Linear Regression	0.696601	0.822828	0.002667
Decision Tree Regression	0.090264	0.262870	0.898210
Random Forest Regressor	0.260651	0.349632	0.819929
Gradient Boosting Regressor	0.652402	0.770423	0.125661

Image 10

HYPER-PARAMETER TUNING

We performed hyperparameter tuning to find the optimal values for model parameters. This helped improve the model's performance and generalization ability.

PREDICTION (Model validation):

After training and tuning the models, we validated their performance using the testing set. This step provided an unbiased assessment of the models' predictive capabilities on unseen data.

We validated the trained model by comparing its **predicted values** to the **actual values**. This step helps us assess the accuracy and reliability of the model's predictions.

CONCLUSION & RECOMMENDATION

By following these steps, we trained and evaluated multiple models to identify the most accurate and reliable model for making predictions or gaining insights from the Olympic athletes' data.

Based on the insights gained from the analysis, we provide recommendations for various stakeholders. This may include suggestions for athlete training, team selection, resource allocation, or strategic decision-making. Recommendations should be practical and actionable, considering the context of the Olympic Games and the available data.

This stage wraps up the project by summarizing the main findings, highlighting the best-performing model, and providing actionable recommendations.

It ensures that the analysis has addressed the project objectives and has generated valuable insights that can inform decision-making in the context of Olympic sports.