

TEAM JUPYTER

# 120 YEARS OF OLYMPIC DATA ANALYSIS

Best athletes and medal prediction in modern Olympic games

24th of June 2023

# TEAM JUPYTER



**Emmanuel Omatozaye Aiyede**  
Presenter 1



**Jimoh Saheed Tunde**  
Presenter 2

**Project Lead**

Udofia Etietop

**Assistant Project Lead**

Emmanuel Omatozaye Aiyede

**Query Analyst**

Olamide Oladipo

**Other active members**

Shedrack David

Onuba Chibuike

Boudouin Alphonse Emmanuel Y

# PROBLEM STATEMENT

## 120 YEARS OF OLYMPIC DATA ANALYSIS

Best athletes and medal prediction in modern Olympic games.

- The problem is to analyze Olympic athlete data and develop a predictive model to forecast the likelihood of winning a medal in future Olympic Games.
- The goal is to understand the factors that contribute to an athlete's success and to build a model that can effectively identify potential medal winners based on their characteristics and performance..



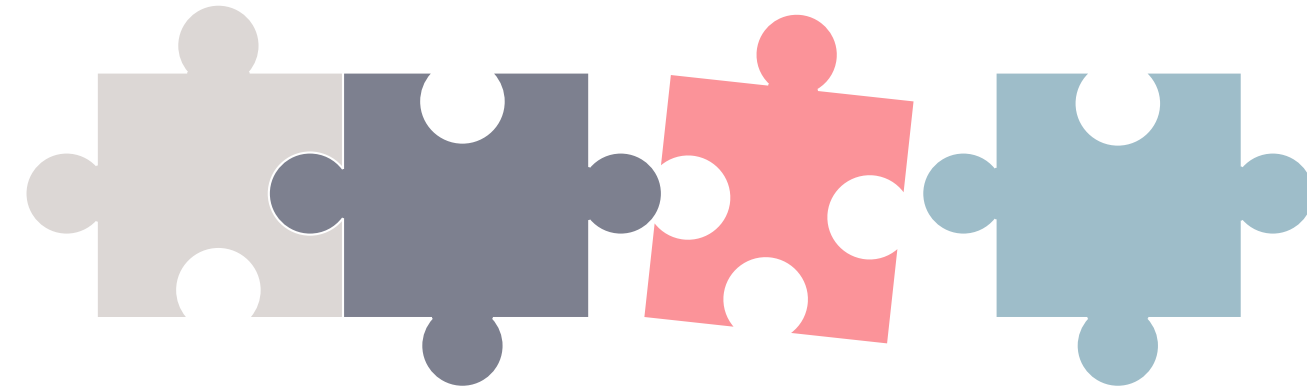
# EXISTING SOLUTION

- Researchers and data scientists have conducted numerous studies using these approaches to identify patterns and make predictions.
- Existing solutions for predicting Olympic medal winners and analyzing Olympic athlete data involve applying machine learning algorithms such as regression, classification, or clustering to historical Olympic data.
- These solutions consider factors such as athletes' past performance, demographics, and other relevant variables to predict medal winners.



# OUR APPROACH

- Data Source
- Data Description
- Data Pre-processing
- Exploratory Data Analysis
- Model Training
- Model Evaluation
- Hyper- parameter Tuning
- Prediction - Model Validation



# DATA DESCRIPTION

## Data Source

- 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

## Features

ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal, Region, Notes

## Data Wrangling/ Processing

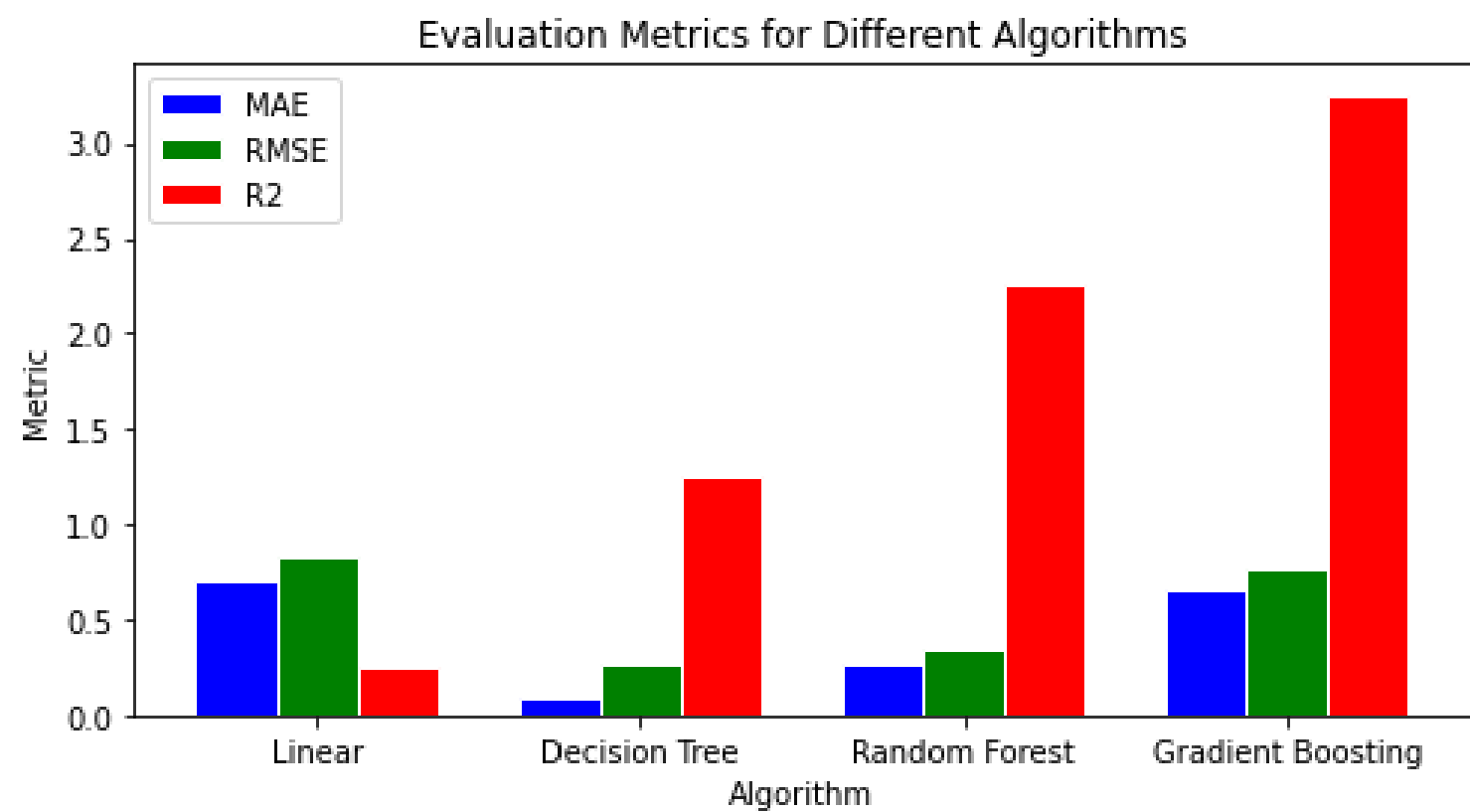
- Handling missing data (Age, Height, Weight, Notes)
- Transform dataset by merging df1 & df2
- Data Integration
- Feature Engineering(continent, age/height/weight category)
- Ensuring the quality, consistency, and reliability of the data before it can be used for further analysis

## Exploration

- Top 10 countries with the most medals (The USA stands out as the dominant country)
- Summary of the number of medals won by the top countries ('USA', 'Russia', 'Germany') over the years
- Top African countries with most medals
- Top 20 countries with the most GOLD medal
- Distribution of 'Age', 'Height', 'Weight' columns
- Distribution of 'Age-category', 'Height-category', 'Weight-category' of the top 10 countries

# MODEL

## Algorithms and Evaluation metrics



## Prediction/ Validation

Metrics	Actual Values	Predicted Values
Mean Absolute Error(MAE)	0.90263907	0.085612639
Root Mean Squared Error(RMSE)	0.262870101	0.252235941
R-squared (R2)	0.898210271	0.906699214

- This validation suggest that the model's predictions are quite accurate and have a strong correlation with the actual values.
- The low MAE and RMSE values and the high R-squared value indicate that the model performs well in predicting the target variable.

# SUMMARY

## Observation

- The analysis of Olympic athlete data provided valuable insights into various aspects such as athletes' demographics, sports performance, and medal achievements.
- Exploratory data analysis revealed interesting patterns and trends, such as the distribution of athletes' heights and weights, and the performance of different countries.

## Challenges

- Missing or incomplete data in certain columns, requiring data cleaning and imputation techniques.
- Limited availability of certain variables or features that could potentially provide more insights into the analysis.
- The complexity of predicting Olympic medal winners accurately based on the available data was also a challenge, as many factors contribute to an athlete's success.

## Recommendations

- Incorporating additional features or variables, such as athletes' training regimes, performance in qualifying events, and historical records, could enhance the accuracy of predictions.
- Exploring advanced machine learning techniques, such as deep learning, may provide improved predictive models.
- Investigating the impact of socioeconomic factors, cultural influences, and political dynamics on athletes' performance and medal achievements could provide a deeper understanding of the Olympic Games.