Beyond The Podium: Analysing Success and Fairness

Report submitted in partial fulfilment of the requirement for the degree of

B.Tech

in

Artificial Intelligence & Data Science

Submitted To:

Dr. Monika

HOD

(Artificial Intelligence And Data Science)

Submitted By:

Vridhi Aggarwal

Enrollment no.: 02320811922

Department of AI&DS

Bhagwan Parshuram Institute of Technology

PSP-4, Sec-17, Rohini, Delhi-89

DECLARATION

This is to certify that Report titled "Beyond The Podium: Analysing Success And Fairness", is submitted by me in partial fulfilment of the requirement for the award of degree of B.Tech in Artificial Intelligence & Data Science to BPIT Rohini Delhi affiliated to GGSIP University, Delhi. It comprises of my original work. The due acknowledgement has been made in the report for using other's work.

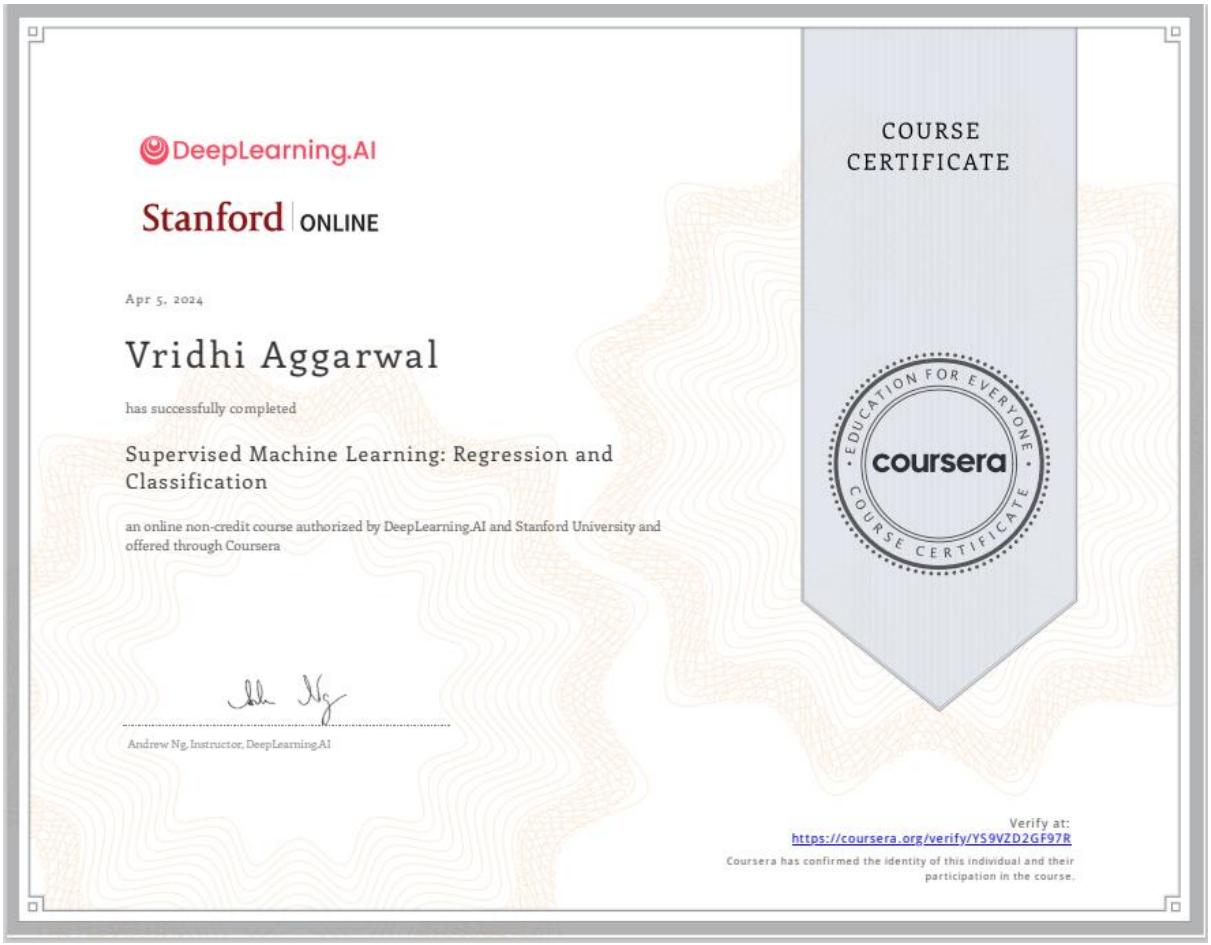
Date:

Name – Vridhi Aggarwal

Enrollment no. – 02320811922

Signature -

COMPANY CERTIFICATE



TRAINING COORDINATOR CERTIFICATE

This is to certify that Report titled "Beyond The Podium: Analysing Success And Fairness" is submitted by **Vridhi Aggarwal, Roll no. 23**, under the guidance of Dr. Monika in partial fulfilment of the requirement for the award of degree of B.Tech in Artificial Intelligence & Data Science to BPIT Rohini affiliated to GGSIP University, Delhi. The matter embodied in this Report is original and has been dully approved for the submission.

(signature)

Date:

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have supported and guided me throughout the completion of this project.

Firstly, I would like to extend my deepest thanks to Dr. Monika, whose invaluable expertise, insightful feedback, and continuous encouragement helped shape this report. Their guidance has been instrumental in my understanding and analysis of the data.

I am also deeply grateful to my professors, mentors, and colleagues for their constant support, valuable suggestions, and motivation throughout the duration of this project.

I am highly grateful and obliged to Mr. Andrew Ng, our instructor at Coursera, under whose tutorship I got to learn a lot many things and was able to apply the knowledge during the course of this project.

Finally, I would like to thank my family and friends for their encouragement, patience, and understanding, as they have been my source of strength throughout this journey.

(signature)

Date:

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

ABSTRACT

Doping has been a persistent issue in the Olympics, compromising the integrity of athletic competitions and posing significant health risks to athletes. This project leverages machine learning techniques to analyse historical doping data from past Olympic Games. Using Python as the core programming language, the project employs data preprocessing, exploratory data analysis (EDA), and advanced machine learning algorithms to detect patterns and anomalies related to doping. Key factors such as athlete performance, nationality, and sport are analysed to identify potential correlations with doping incidents.

Doping not only provides athletes with an unfair competitive advantage but also tarnishes the spirit of fair play and threatens the reputation of entire sports organizations. The long-term health risks for athletes who engage in doping can include cardiovascular diseases, hormonal imbalances, and even psychological disorders. Moreover, doping scandals can lead to disqualifications, stripping of medals, and legal consequences, severely impacting athletes' careers and their countries' standings. This project aims to better understand the underlying patterns of doping through data-driven insights, helping to address its damaging effects on both individual athletes and the broader sporting community.

The project integrates various algorithms, including logistic regression and random forests, to classify and predict doping occurrences. To enhance understanding, graphical and pictorial representations such as histograms, box plots, and heatmaps are used to visually communicate complex data insights. These visualizations help stakeholders, including sports analysts and policymakers, grasp trends and correlations effectively.

The results of this analysis aim to provide proactive insights into doping behaviors, potentially helping regulatory bodies like the World Anti-Doping Agency (WADA) in formulating more effective strategies for detection and prevention. Overall, the project underscores the potential of machine learning and data analysis in addressing one of the most pressing challenges in modern sports.

CHAPTER 1

INTRODUCTION

1.1. The Olympic Games, the pinnacle of global athletic achievement, symbolize human endurance, competitive spirit, and the pursuit of excellence. However, alongside these ideals, the issue of doping has cast a shadow over the games for decades. Doping refers to the use of banned substances or methods by athletes to enhance performance, which compromises fairness and threatens the integrity of sports. The World Anti-Doping Agency (WADA) and various sports governing bodies have implemented strict measures to detect and prevent doping, but the problem persists. In this context, technology, particularly machine learning (ML), has become a powerful tool for analysing patterns, detecting anomalies, and offering insights into doping practices in the Olympics.

1.2. This project focuses on leveraging machine learning techniques to analyse doping data from past Olympic events, with a special emphasis on identifying trends, anomalies, and correlations between various factors. Using Python as the programming language, the project employs advanced algorithms and visualizations to offer a deeper understanding of how doping has impacted the Olympics over the years. Additionally, the use of graphical and pictorial representations of data ensures that complex patterns and insights can be communicated effectively to a wide audience, including sports analysts, researchers, and policymakers.

1.3. Significance of Doping Analysis in the Olympics

1.3.1. Doping scandals have marred the history of the Olympic Games, with numerous high-profile cases leading to disqualifications, stripping of medals, and long-term damage to the reputation of athletes and their countries. Doping not only provides an unfair advantage but also poses severe health risks to athletes. Therefore, detecting and analysing doping patterns is crucial to maintaining the integrity of the sport.

1.4. Traditional methods of detecting doping, such as biological testing, are resource-intensive and often reactive, identifying doping incidents after they occur. Machine

learning, on the other hand, offers a proactive approach by analysing vast datasets, detecting hidden patterns, and predicting possible doping cases before they can disrupt the games. This shift from reactive to predictive analysis is a significant step forward in the fight against doping.

1.5. Objective of the Project

The primary objective of this project is to develop a machine learning model that can analyse historical doping data from the Olympics to:

1.5.1. Identify patterns and trends related to doping incidents.

1.5.2. Detect anomalies in athlete performance that may suggest potential doping.

1.5.3. Correlate factors such as country, sport, and athlete performance with doping cases.

1.5.4. Provide visual insights through graphs and pictorial representations to help stakeholders make informed decisions.

1.6. By achieving these objectives, the project aims to contribute to a better understanding of the dynamics of doping in the Olympics, offering valuable insights that can be used for policy formulation, athlete monitoring, and future research.

1.7. Methodology

1.8. The project follows a structured approach to analysing doping data, with a focus on both data-driven insights and the application of machine learning techniques. The methodology can be divided into several key phases:

1.9. Data Collection and Preprocessing:

1.9.1. The project begins with the collection of doping-related data from reliable sources such as WADA reports, athlete performance records, and publicly available Olympic datasets. This raw data is then cleaned, processed, and transformed into a format suitable for machine learning analysis. Various techniques such as normalization, handling missing values, and feature selection are applied to ensure the data is robust and ready for modelling.

1.10. Exploratory Data Analysis (EDA):

1.10.1. EDA plays a critical role in uncovering patterns and trends in the dataset. Visualizations such as histograms, box plots, scatter plots, and heatmaps are employed to examine the relationships between variables like athlete nationality, age, sport category, and performance metrics. These visual tools provide initial insights into doping trends, allowing for the identification of key features that might influence doping behaviour.

1.11. Machine Learning Model Development:

1.11.1. After the exploratory analysis, various machine learning algorithms are applied to the data to detect doping patterns. Algorithms such as logistic regression, decision trees, and random forests are tested and evaluated for their ability to classify and predict doping cases. The selection of the right model is based on factors such as accuracy, precision, recall, and F1 score. In this project, Python's extensive machine learning libraries such as scikit-learn, TensorFlow, and Keras are utilized to build and train the models.

1.12. Graphical and Pictorial Representation of Results:

1.12.1. One of the key aspects of this project is the visual representation of data and results. Graphs such as bar charts, line graphs, and pie charts are used to represent the frequency of doping incidents across different Olympic Games and sports categories. Moreover, heatmaps and correlation matrices are employed to show the relationship between various factors and doping incidents. These visualizations not only make the analysis easier to understand but also provide actionable insights for stakeholders.

1.13. Model Evaluation and Interpretation:

1.13.1. Once the machine learning models are trained, they are evaluated using test datasets to determine their effectiveness in detecting doping incidents. The results are interpreted in the context of the larger doping issue, with discussions on the accuracy of predictions and potential improvements. This evaluation phase is crucial in refining the model for future use.

1.14. Pictorial Representation of Doping Analysis

1.15. One of the standout features of this project is the use of pictorial representation to communicate complex data insights. By visualizing the results, the project ensures

that the findings are accessible to a wider audience, even those without technical expertise. The following visual tools are extensively used:

1.15.1. **Histograms:** To showcase the distribution of doping incidents across different Olympic Games.

1.15.2. **Box Plots:** To identify outliers in athlete performance, which might indicate potential doping.

1.15.3. **Correlation Heatmaps:** To display the relationships between variables such as athlete nationality, performance metrics, and doping cases.

1.15.4. **Scatter Plots:** To visualize the relationship between performance improvement and doping cases.

1.15.5. **Pie Charts and Bar Graphs:** To represent the percentage of doping cases by country, sport, and athlete age.

1.16. These visualizations allow for quick, intuitive understanding of the key trends and correlations within the dataset, making it easier for stakeholders to grasp the underlying issues in doping at the Olympics.

1.17. Conclusion

1.18. Doping in sports, especially in the Olympics, remains a significant challenge that requires innovative solutions for detection and prevention. This project demonstrates how machine learning and data visualization can be powerful tools in the fight against doping. By analysing historical data and using advanced machine learning techniques, this project provides actionable insights into doping trends and patterns, offering a proactive approach to combating the issue. The findings and visual representations serve as valuable resources for researchers, sports analysts, and regulatory bodies working to ensure the integrity of the Olympic Games remains intact for future generations.

CHAPTER 2

PROBLEM STATEMENT

2.1 Problem Statement:

2.2 Doping scandals have cast a long shadow over the Olympics, undermining the integrity and spirit of the Games. Athletes use performance-enhancing drugs (PEDs) to gain an unfair advantage, leading to numerous cases of cheating, disqualification, and the reassignment of medals. The problem has far-reaching consequences for athletes, nations, and the reputation of the Olympic movement. Analysing the Olympics dataset can help uncover patterns in doping violations and their impact on the Games, providing insight into the magnitude of the issue and potential preventive measures.

2.3 Key Aspects of the Problem:

2.3.1. Widespread Nature of Doping: Doping has been a recurring issue across multiple sports and countries. Certain nations and disciplines have been implicated more frequently, raising suspicions of systematic doping programs.

2.3.2. Impact on Results: Doped athletes often secure unfair victories, leading to distorted medal counts. Years later, retesting of samples sometimes reveals violations, leading to posthumous disqualifications and the reassignment of medals, which affects the historical records of the Games.

2.3.3. Detection and Challenges: Despite improved testing methods, many athletes evade detection during the events. Sophisticated doping techniques often outpace anti-doping agencies, leading to delayed discoveries of violations.

2.3.4. Global Governance: Agencies like the World Anti-Doping Agency (WADA) set anti-doping standards, but their enforcement is uneven across

different countries and sports. State-sponsored doping programs, like those revealed in the Russian doping scandal, showcase how entire nations can evade global detection.

2.3.5. **Public Trust:** Each doping scandal erodes public trust in the fairness of the Olympic Games. Fans question the legitimacy of the results, while clean athletes are overshadowed by those who cheat, affecting the reputation of the Olympics.

2.4Objectives Of the statement:

2.4.1. Understand the Scope of Doping in the Olympics:

2.4.1.1. Analyze the prevalence of doping violations over time to identify trends across different Olympic Games.

2.4.1.2. Determine the sports and countries most frequently implicated in doping scandals, uncovering any systematic patterns of cheating.

2.4.2. Assess the Impact of Doping on Athletic Performance and Medal Outcomes

2.4.2.1. Quantify the impact of performance-enhancing drugs (PEDs) on athletes' results by comparing doped athletes' performances to those who competed clean.

2.4.2.2. Investigate how doping influences medal standings and track how many medals have been reallocated due to doping-related disqualifications.

2.4.3. Examine the Effectiveness of Anti-Doping Efforts

2.4.3.1. Evaluate the effectiveness of current anti-doping measures by analyzing how frequently athletes are caught during events versus years later through retesting.

2.4.3.2. Study how advances in testing methods (e.g., more sensitive drug detection) have improved or failed to curb doping.

2.4.4. Develop Predictive Models for Future Doping Risks

2.4.4.1. Use historical data to build machine learning models that can predict the likelihood of doping incidents by analyzing factors such as country, sport, and performance metrics.

2.4.4.2. Identify athletes or countries that may be at higher risk for future doping based on past patterns of performance improvement and testing results.

2.4.5. Analyze the Socio-Political Factors in Doping Scandals

2.4.5.1. Explore the geopolitical dimensions of doping scandals, such as whether state-sponsored doping programs (e.g., the Russian scandal) are more prevalent in certain countries.

2.4.5.2. Investigate any potential bias in doping enforcement and testing across different nations and sports to see if certain athletes or regions are scrutinized more heavily.

2.4.6. Evaluate Public Sentiment and Media Impact

2.4.6.1. Analyze media coverage and social sentiment related to major doping scandals, assessing how the public's perception of athletes, sports, and countries shifts after each revelation.

2.4.6.2. Track changes in Olympic viewership and participation following significant doping incidents to measure the broader cultural impact of these scandals.

2.4.7. Support Fairer Competition and Future Policy Recommendations

2.4.7.1. Based on the data analysis, recommend improved anti-doping policies and testing methods to ensure fairer competition in future Olympic Games.

2.4.7.2. Propose adjustments to the rules governing doping detection and consequences (e.g., immediate medal reallocation, harsher penalties for countries involved in systematic doping) based on the analysis of historical patterns.

CHAPTER 3

SYSTEM ANALYSIS AND DESIGN

3.1 Software Requirement Specifications:

This SRS document provides detailed descriptions of the functional and non-functional requirements of the system, along with the design constraints, interface requirements, and user roles.

3.2.1 Product Perspective

The system will be a standalone web-based or cloud application that allows users to:

3.2.3.1 Upload and manage large-scale Olympic datasets.

3.2.3.2 Analyze doping cases and trends across multiple sports and countries.

3.2.3.3 Build predictive models for future doping risks.

3.2.3.4 Visualize results using dashboards, charts, and geographical maps.

3.2.2 Product Features

3.2.3.1 **Data Ingestion:** Support for uploading CSV, Excel, and API-based data sources.

3.2.3.2 **Data Cleaning:** Tools for cleaning and preparing datasets (e.g., removing duplicates, handling missing values).

3.2.3.3 **Trend Analysis**: Tools to analyze doping trends by country, sport, and year.

3.2.3.4 **Predictive Analytics**: Machine learning models to predict future doping violations based on historical patterns.

3.2.3.5 Performance Impact Analysis: Analyze how doping influences athletic performance and medal standings.

3.2.3.6 Visualization: Interactive charts, graphs, and geographical heat maps for visualizing doping patterns.

3.2.3 User Classes and Characteristics

3.2.3.1 Data Analysts: Experts who will analyze data for trends and insights.

3.2.3.2 Sports Authorities (e.g., IOC, WADA): Officials who will use the system to identify patterns and plan interventions.

3.2.3.3 Researchers/Academics: Users who will study the social and scientific aspects of doping scandals.

3.2.4 Operating Environment

3.2.3.1 Platform: Web-based application compatible with modern browsers (Chrome, Firefox, Edge).

3.2.3.2 Backend: Cloud-based processing for large datasets and machine learning.

3.2.3.3 Operating Systems: Compatible with Windows, Linux, and macOS.

3.2 Use Case Diagrams/DFD/ERD:

3.2.1 ER Diagram:

3.2.3.1 The **ER Diagram for the Olympics Dataset Analysis and Doping Analysis** combines key entities to model the relationships between athletes, events, doping tests, and countries. The central entity is the **Athlete**, which contains attributes such as Athlete ID, Name, Country, and Performance Metrics. Athletes participate in one or more **Events**, which are linked to specific **Sports** (e.g., athletics, swimming), represented by attributes like Sport Name and Category. Each event is defined by Event ID, Event Name, Sport, Year and Medal Won. An athlete's participation in events creates a many-to-many relationship between the **Athlete** and **Event** entities, allowing for detailed tracking of an athlete's performance in multiple events across different years. The ER diagram captures these relationships to support analyses on doping trends, athlete performances, and the effect of

doping violations on medals and national reputations, forming a complete framework for dataset exploration and predictive analytics.

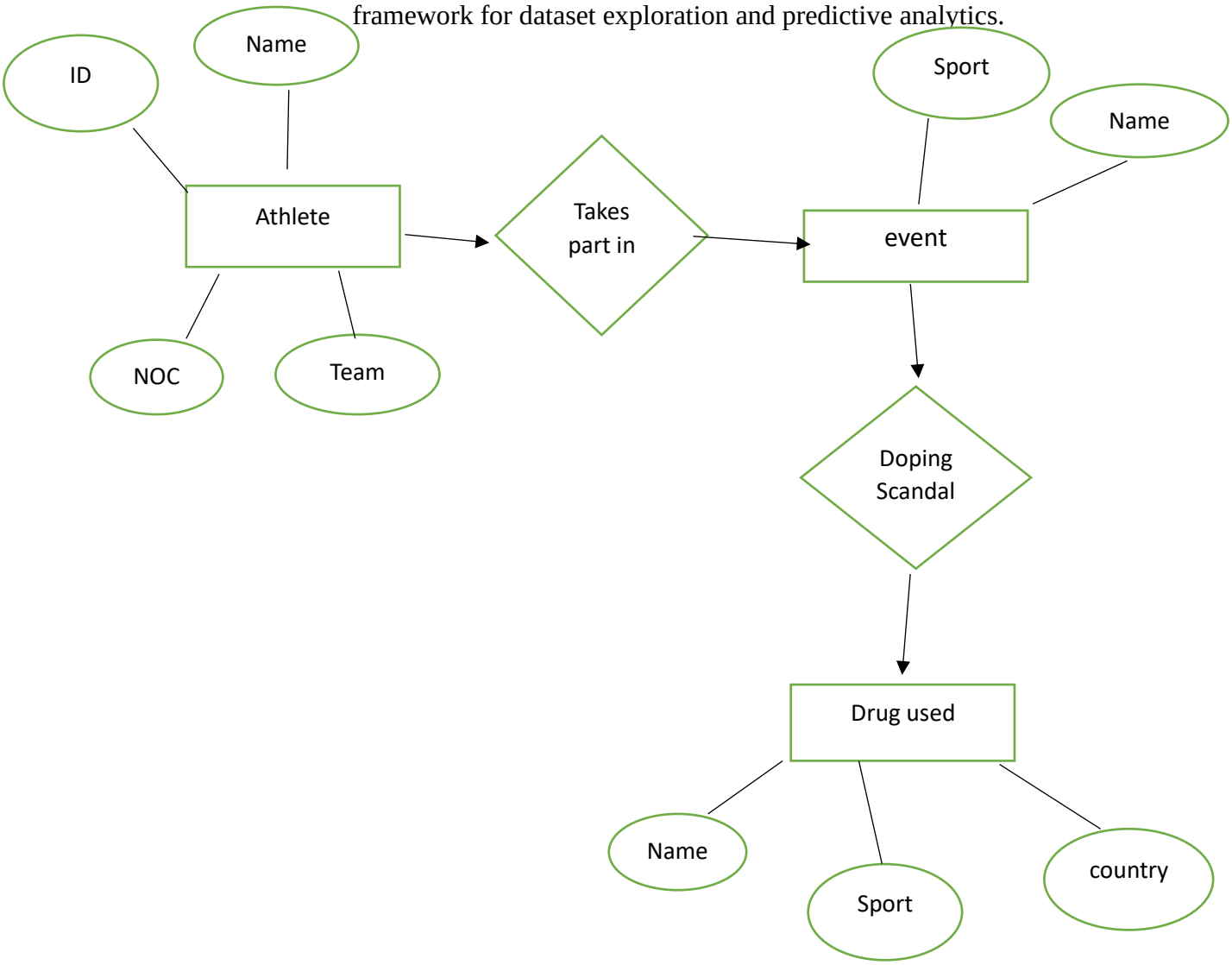


Figure 3.1: Entity Relational Diagram

3.2.2 Use Case Diagram:

3.2.3.1 The **use case diagram for the Olympics dataset and doping analysis system** involves multiple actors interacting with the system to perform specific tasks. The main actors are the **Data Analyst**, **Sports Authority** (such as IOC or WADA), the **General Public**, and the **Admin**. The **Data Analyst** plays a central role, responsible for uploading Olympic datasets, cleaning and preparing the data to remove inconsistencies, and analyzing doping trends across countries, sports, and years. They also use the system to generate predictive models to identify potential doping risks in the future. The Data Analyst can visualize data through charts and maps,

as well as generate detailed reports on doping trends and athlete performance. The **Sports Authority** interacts with the system primarily to monitor and track athletes, analyze doping trends, and access detailed reports. They use these insights to make policy decisions and improve doping regulations. They also receive real-time alerts if suspicious patterns, such as sudden performance improvements, are detected for any athlete.

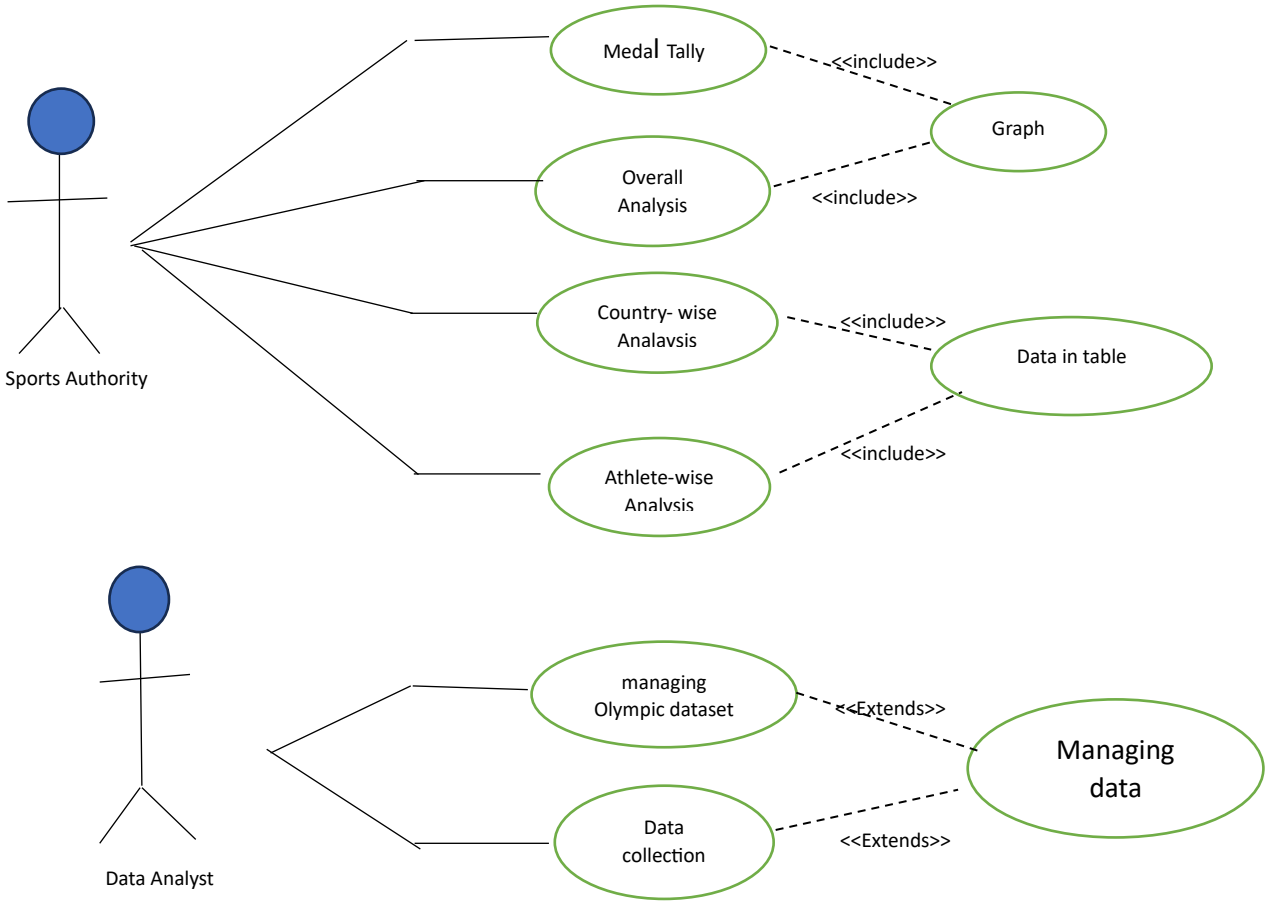


Figure 3.2: Use case diagram

3.2.3 Data flow diagrams

3.2.3.1 A Data Flow Diagram (DFD) for the Olympic dataset analysis and doping analysis would illustrate the flow of data between various processes, data stores, and external entities involved in the project. In the context of

the Olympics dataset, the DFD would depict the collection of data related to athletes, events, medals, and their performance metrics, which flows into data processing modules that analyze trends, correlations, and anomalies. For doping analysis, the DFD would show how data on athletes' test results and doping records are integrated and processed using machine learning models to detect patterns indicative of doping. The processed data outputs, including visualizations and statistical insights, would then be stored and displayed for decision-makers or analysts. External entities, such as data input sources (e.g., sports federations) and stakeholders (e.g., Olympic committees), interact with the system by supplying data and receiving analyzed reports.

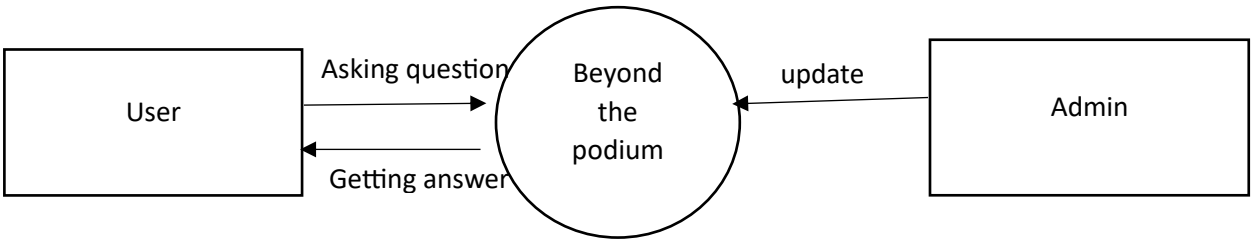


Figure 3.3: 0-level DFD

3.4.1 Flow Chart:

3.4.1.1. The flow chart for the Olympics Dataset and Doping Analysis outlines the step-by-step process involved in analyzing athlete performance and doping trends. The process begins with data collection, where Olympic performance data and doping test records are gathered from multiple sources. The next step is data preprocessing, where the datasets are cleaned, integrated, and organized using tools like Pandas and NumPy. Once the data is ready, Exploratory Data Analysis (EDA) is conducted using Matplotlib and Seaborn to visualize patterns and trends in the data, such as doping violations by country or sport.

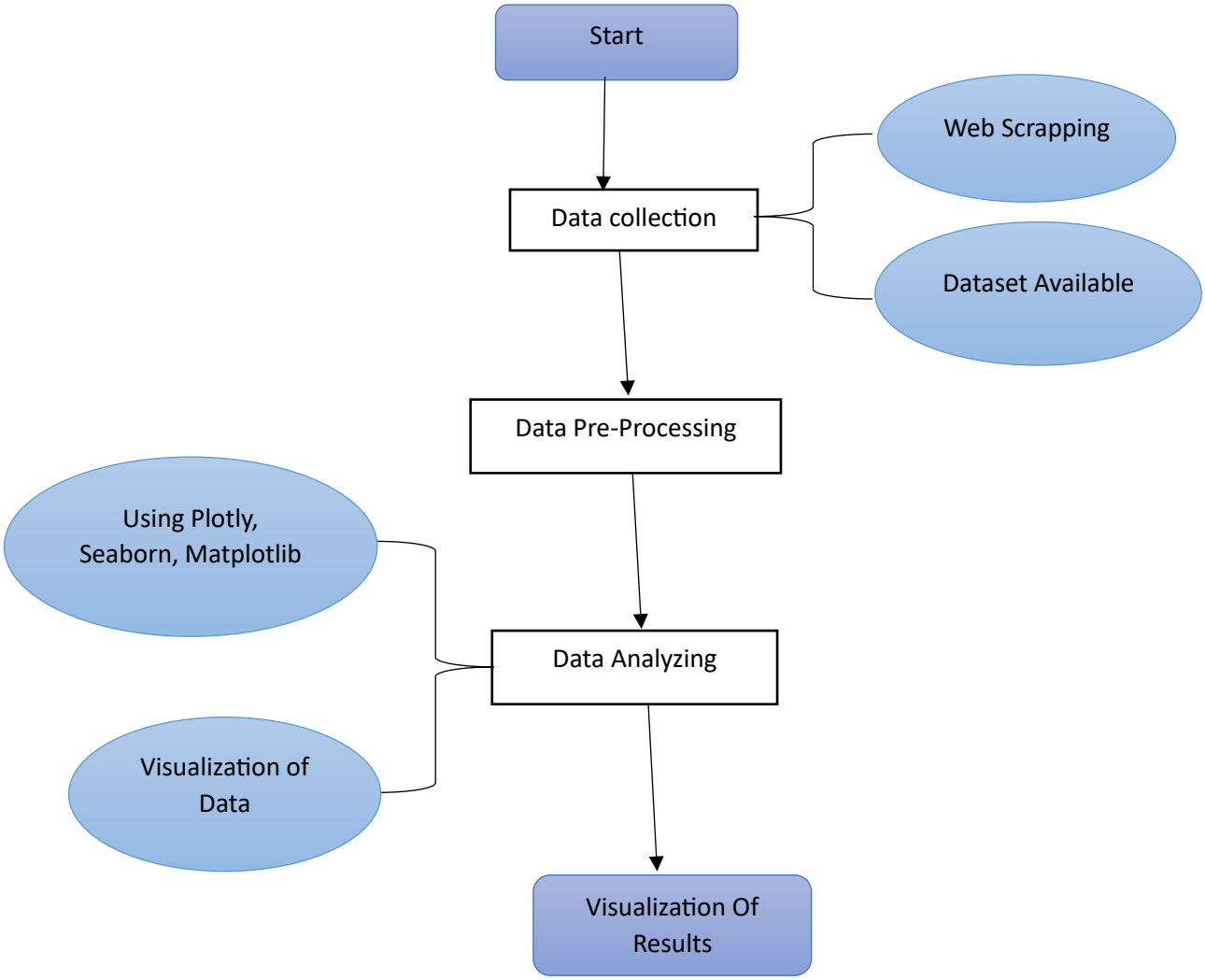


Figure 3.4: Flow Chart of Beyond The Podium

CHAPTER 4

METHODOLOGY

4.1 Process Selection:

The process selection for Beyond the Podium: Analysing success and failure focuses on choosing the most appropriate data analysis and machine learning techniques to uncover insights about athlete performance and doping trends. The key stages of this process are designed to ensure data integrity, analytical rigor, and actionable results.

4.2.1 Data Preprocessing and EDA:

4.2.1.1.Pandas and **NumPy** are used for cleaning, organizing, and integrating the datasets.

4.2.1.2. Exploratory Data Analysis (EDA) is performed with **Matplotlib**, **Seaborn**, and **Plotly** to visualize doping trends and athlete performance.

4.2.2 Predictive Modelling and Statistical Analysis:

4.2.2.1.Scikit-learn is used for machine learning models (e.g., logistic regression, decision trees) to predict doping risks.

4.2.2.2.SciPy is applied for statistical tests (t-tests, ANOVA) to assess the impact of doping on performance.

4.2.3 Visualization and Reporting:

4.2.3.1. Final results are visualized with **Plotly**, **Tableau**, or **Power BI**, ensuring interactive and intuitive reports for stakeholders.

4.2 Methodology used:

The methodology for **Olympics Dataset Analysis and Doping Analysis** involves several steps that aim to uncover trends, patterns, and insights related to athletic performance and doping violations. This methodology is designed to systematically process large-scale datasets, apply statistical and machine learning techniques, and generate actionable insights. The following points describe the key steps of the methodology:

4.2.1 Data Collection:

The first step involves collecting and gathering the relevant data required for analysis. This includes:

4.2.1.1.Olympic datasets: These datasets contain detailed information about athletes, countries, events, sports, and their respective performance data (e.g., times, scores, distances).

4.2.1.2.Doping data: This includes records of doping tests, results, substances detected, and athlete details (gathered from sources such as WADA and other sports authorities).

4.2.1.3.External data: Additional data, such as media reports and public sentiment from social media, is gathered to analyze the impact of doping scandals on public opinion.

4.2.2 Data Preprocessing:

Once the data is collected, it undergoes thorough preprocessing to clean, organize, and prepare it for analysis. This step involves:

4.2.2.1.Data Cleaning: Removing duplicates, handling missing or incomplete data, and correcting any inconsistencies or inaccuracies in the datasets.

4.2.2.2.Data Transformation: Formatting the data into a consistent structure, which might involve converting dates, standardizing units of measurement, and normalizing values across different datasets.

4.2.2.3.Data Integration: Combining datasets (e.g., linking athlete performance data with doping results) to create a unified dataset. The integration ensures that key relationships between entities such as athletes, countries, and doping tests are preserved and can be analyzed together.

4.2.3 Exploratory Data Analysis (EDA):

Exploratory Data Analysis is a critical step that involves analyzing the cleaned datasets to uncover initial patterns, trends, and insights. This phase includes:

4.2.3.1.Descriptive Statistics: Calculating summary statistics (e.g., mean, median, standard deviation) for athlete performances, doping tests, and other metrics to understand the central tendencies and variability in the data.

4.2.3.2.Visual Analysis: Using data visualization techniques such as histograms, scatter plots, bar charts, and heat maps to examine trends over time, identify patterns in doping violations, and highlight differences across countries or sports.

4.2.3.3.Correlation Analysis: Investigating relationships between variables (e.g., athlete performance and doping test outcomes) to understand how doping may have influenced performance metrics or medal outcomes.

4.2.4 Doping Trend Analysis:

4.2.4.1.Temporal Analysis: Examining doping trends over different Olympic cycles (e.g., every four years), identifying whether doping incidents increase or decrease with time.

4.2.4.2. Country and Sport Analysis: Analyzing which countries and sports have a higher prevalence of doping violations. This involves creating visualizations that show doping incidents by country or sport, allowing stakeholders to understand where doping risks are most prevalent.

4.2.4.3.Athlete-Level Analysis: Identifying patterns of doping across individual athletes, analyzing whether top-performing athletes are more likely to be involved in doping scandals.

4.2.5 Impact Analysis of Doping on Performance:

This phase involves understanding how doping has impacted athletes' performances and medal outcomes. This includes:

4.2.5.1.Comparative Analysis: Comparing the performance metrics (e.g., speed, endurance, strength) of athletes who were caught doping versus those who were clean. This helps in quantifying how much of a performance advantage doping provided.

4.2.5.2.Medal Reallocation: Analyzing cases where medals were stripped from athletes due to doping violations and redistributed to others. This analysis assesses how doping has altered the historical medal standings in the Olympics.

4.2.5.3.Statistical Testing: Applying statistical tests (e.g., t-tests, chi-square tests) to determine whether doping has had a significant impact on an athlete's likelihood of winning medals.

4.2.6 Visualization:

Once the analyses are complete, the results are visualized and communicated through interactive dashboards and reports. This step involves:

4.2.6.1.Dashboards: Creating interactive dashboards that allow users (e.g., sports authorities, analysts) to explore doping trends, athlete performances, and risk assessments. The dashboards can include graphs, geographical maps, and timelines to make the data easier to interpret.

4.2.6.2.Public Visualizations: For the general public, high-level visualizations and summaries of doping trends can be made available, focusing on transparency and informing public opinion about doping in sports.

4.3 Implementation:

4.2.1 Data Collection:

	player_id	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal
207154	252089	Patrick Vetterli	M	Switzerland	SUI	1984	Summer	Los Angeles	Athletics	Athletics Men's Decathlon	No medal
204974	249527	Dirk Tichelt	M	Belgium	BEL	2012	Summer	London	Judo	Judo Men's Lightweight	No medal
6231	7140	Thure Andersson	M	Sweden	SWE	1936	Summer	Berlin	Wrestling	Wrestling Men's Welterweight, Freestyle	Silver
201043	244878	Takale Tuna	M	Papua New Guinea	PNG	1988	Summer	Seoul	Athletics	Athletics Men's 400 metres	No medal
75649	91059	Tahesia Harrigan-scott	F	British Virgin Islands	IVB	2012	Summer	London	Athletics	Athletics Women's 100 metres	No medal

Figure 4.1: Historical data from previous Olympic Games.

name	country	year	banned substance	sport	season	city	medal
Hans-Gunnar Liljenwall	Sweden	1968	Ethanol	Modern pentathlon	summer	mexico	bronze
Bakaava Buidaa	Mongolia	1972	Dianabol	Judo	summer	munich	silver
Miguel Coll	Puerto Rico	1972	Amphetamine	Basketball	summer	munich	no medal
Rick DeMont	United States	1972	Ephedrine	Swimming	summer	munich	gold
Aad van den Hoek	Netherlands	1972	Coramine	Cycling	summer	munich	bronze
Jaime Huélamo	Spain	1972	Coramine	Cycling	summer	munich	bronze
Walter Legel	Austria	1972	Amphetamine	Weightlifting	summer	munich	no medal
Mohammad Reza Nasehi	Iran	1972	Ephedrine	Weightlifting	summer	munich	no medal
Blagoi Blagoev	Bulgaria	1976	Anabolic steroid	Weightlifting	summer	Montreal	silver
Mark Cameron	United States	1976	Anabolic steroid	Weightlifting	summer	Montreal	no medal
Paul Cerutti	Monaco	1976	Amphetamine	Shooting	summer	Montreal	no medal
Dragomir Cioroslan	Romania	1976	Fencamfamine	Weightlifting	summer	Montreal	no medal
Phil Grippaldi	United States	1976	Anabolic steroid	Weightlifting	summer	Montreal	no medal
Zbigniew Kaczmarek	Poland	1976	Anabolic steroid	Weightlifting	summer	Montreal	gold
Valentin Khristov	Bulgaria	1976	Anabolic steroid	Weightlifting	summer	Montreal	gold
Lorne Leibel	Canada	1976	Phenylpropanolamine	Sailing	summer	Montreal	no medal
Arne Norrback	Sweden	1976	Anabolic steroid	Weightlifting	summer	Montreal	no medal
Petr Pavlasek	Czechoslovakia	1976	Anabolic steroid	Weightlifting	summer	Montreal	no medal
Danuta Rosani	Poland	1976	Anabolic steroid	Athletics	summer	Montreal	no medal
Serafim Grammatikopou	Greece	1984	Nandrolone	Weightlifting	summer	Moscow	no medal
Vésteinn Hafsteinsson	Iceland	1984	Nandrolone	Athletics	summer	Moscow	no medal
Tomas Johansson	Sweden	1984	Methenolone	Wrestling	summer	Moscow	silver
Stefan Laggner	Austria	1984	Nandrolone	Weightlifting	summer	Moscow	no medal
Göran Pettersson	Sweden	1984	Nandrolone	Weightlifting	summer	Moscow	no medal
Eiji Shimomura	Japan	1984	Testosterone	Volleyball	summer	Moscow	no medal
Mikiyasu Tanaka	Japan	1984	Ephedrine	Volleyball	summer	Moscow	no medal

Figure 4.2: Historical data of doping from previous Olympic Games.

4.2.2 Data Pre-Processing:

	player_id	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal	Bronze	Gold	No medal	Silver
0	0	A Dijiang	M	China	CHN	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	No medal	0	0	1	0
1	1	A Lamusi	M	China	CHN	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	No medal	0	0	1	0
2	2	Gunnar Aaby	M	Denmark	DEN	1920	Summer	Antwerpen	Football	Football Men's Football	No medal	0	0	1	0
3	3	Edgar Aabye	M	Denmark/Sweden	DEN	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	0	1	0	0
4	26	Cornelia (-strannood)	F	Netherlands	NED	1932	Summer	Los Angeles	Athletics	Athletics Women's 100 metres	No medal	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
252545	4979556	Jasmine Jones	F	United States	USA	2024	Summer	Paris	Athletics	Women's 400m Hurdles	No medal	0	0	1	0
252555	4979790	Steven Insixiengmay	M	Lao PDR	LAO	2024	Summer	Paris	Swimming	Men's 100m Breaststroke	No medal	0	0	1	0
252558	4982762	Khrystyna Homan	F	Ukraine	UKR	2024	Summer	Paris	Judo	Women +78 kg	No medal	0	0	1	0

Figure 4.3: Processed data from previous Olympic Games.

4.2.3 Exploratory Data Analysis (EDA):

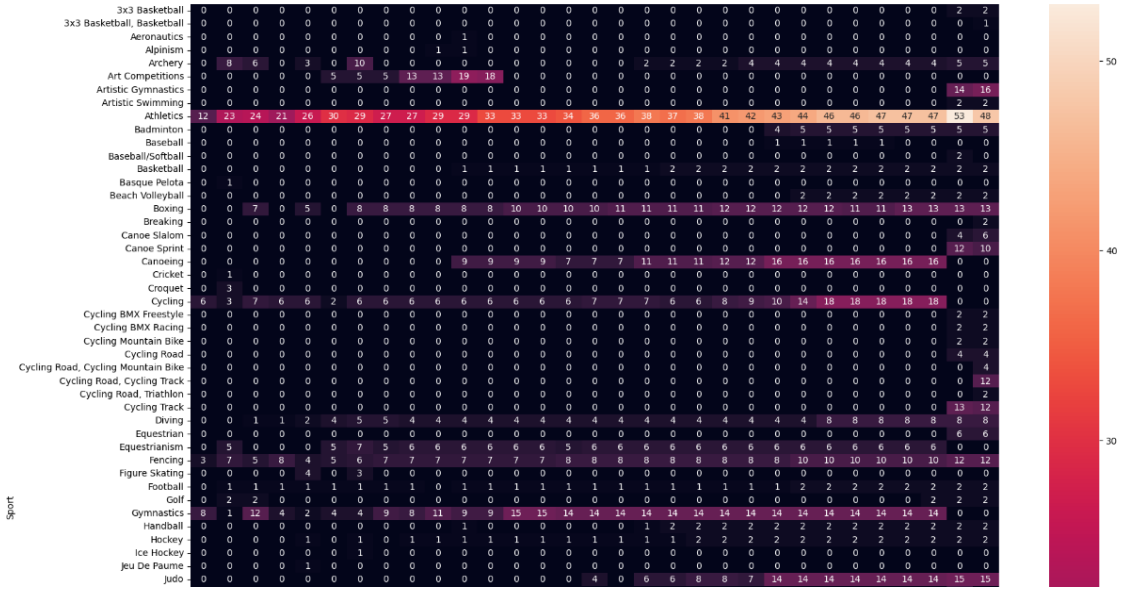


Figure 4.4: heatmap of the data from previous Olympic Games.

4.2.4 Data Visualization:

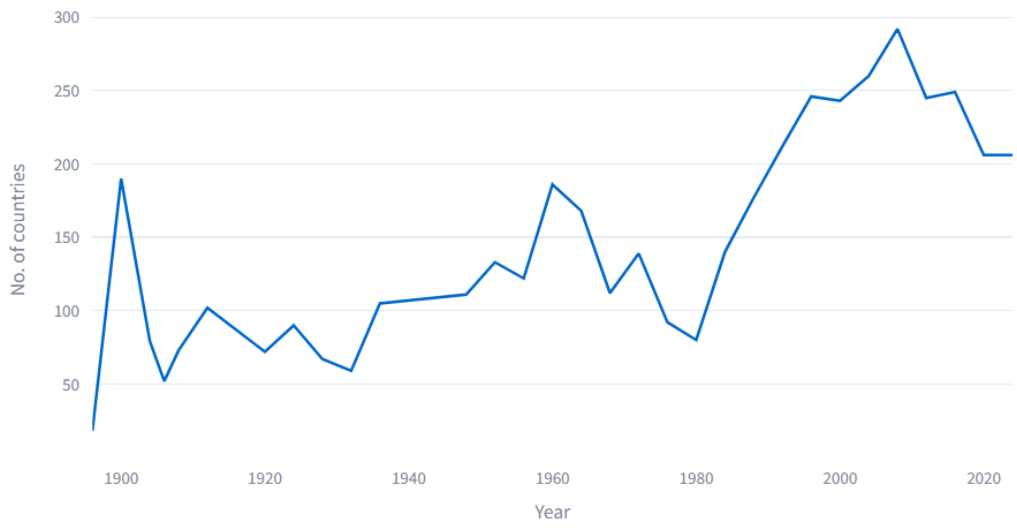


Figure 4.5: participating nations previous Olympic Games.

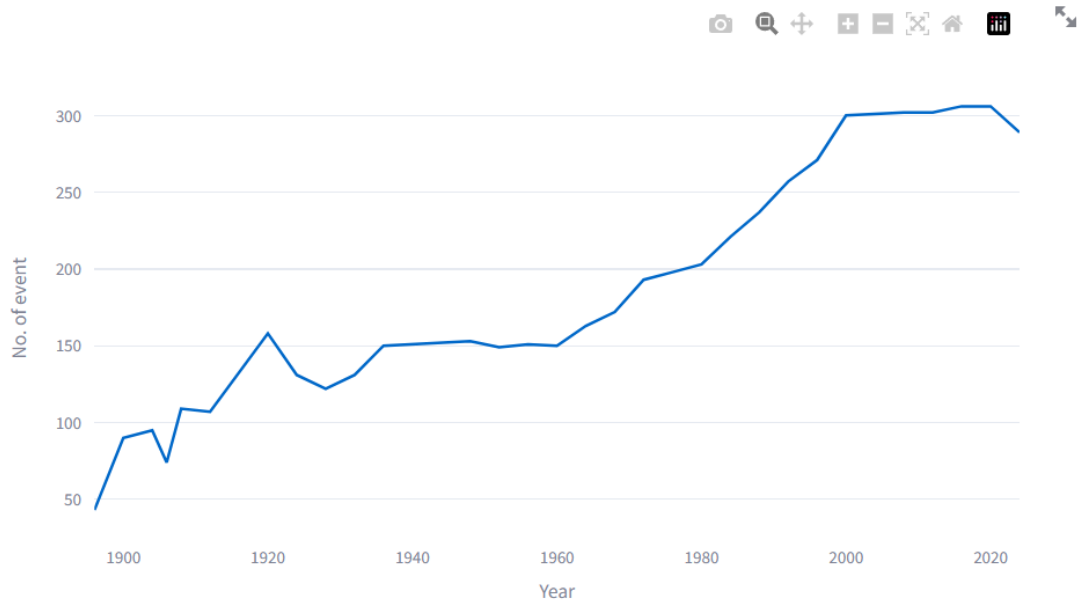


Figure 4.6: events held during Olympic Games.

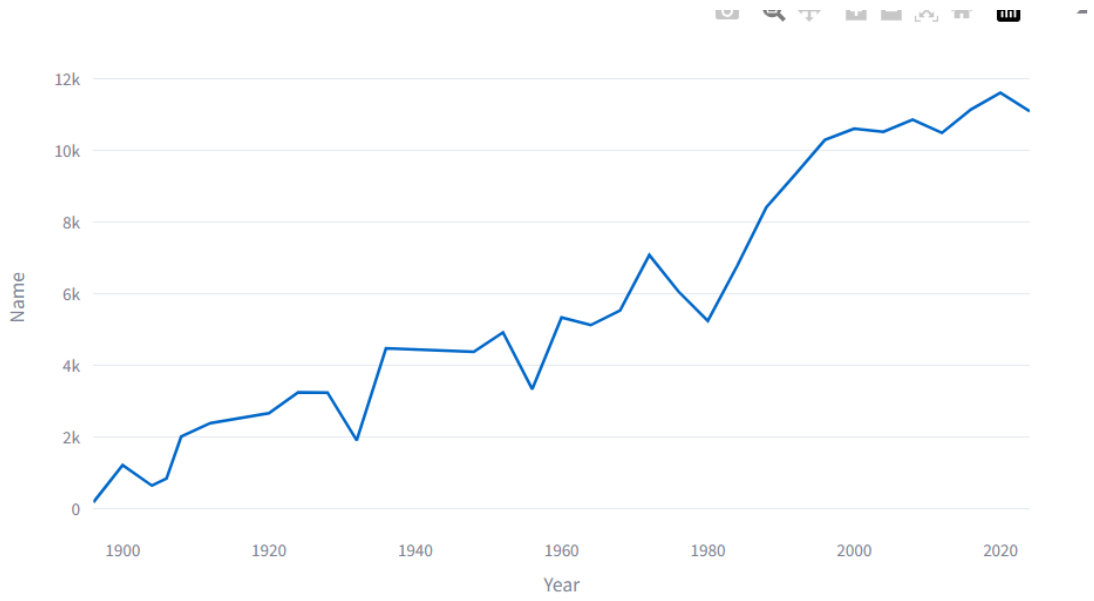
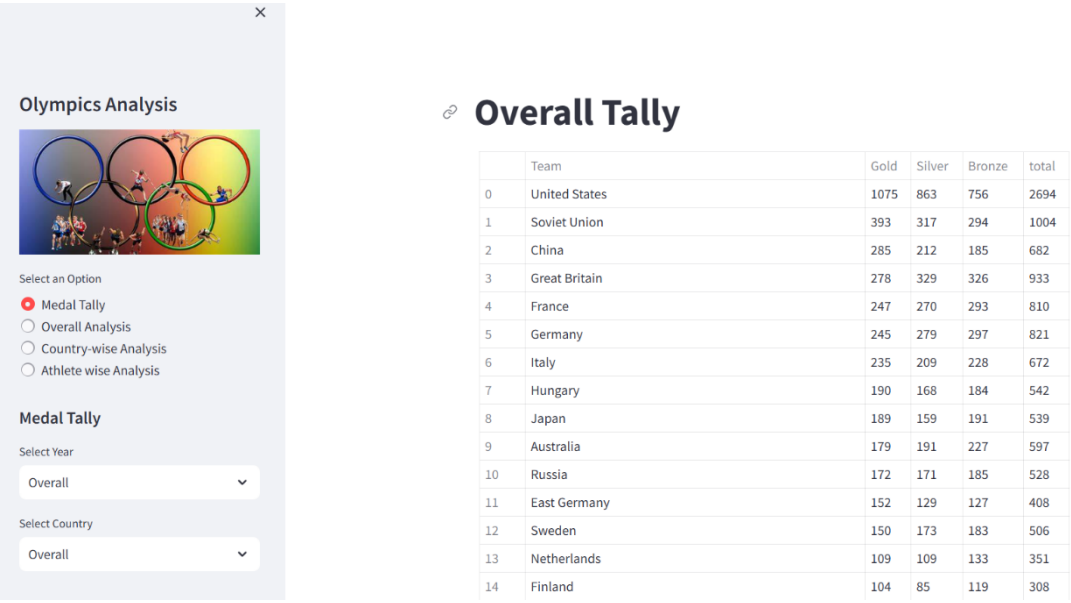


Figure 4.7: Participating Athletes during Olympic Games.

CHAPTER 5

RESULTS

The result of this project is a fully functional website that presents a comprehensive analysis of Olympic Games data over the years, focusing on athlete performances and doping trends. The website features interactive dashboards and visualizations, allowing users to explore various datasets related to Olympic history, medal counts, and country-specific performance. Additionally, it includes a detailed doping analysis section that highlights trends in doping violations across different sports and countries, with insights into how these violations have evolved over time. Predictive models are also integrated, offering forecasts on potential doping risks for future events. The user-friendly interface ensures that stakeholders, including sports authorities and the general public, can easily access and interpret the data. By combining historical data, advanced analytics, and visual storytelling, the website provides an engaging platform for exploring the integrity of sports and the impact of doping on athletic performance.



	Team	Gold	Silver	Bronze	total
0	United States	1075	863	756	2694
1	Soviet Union	393	317	294	1004
2	China	285	212	185	682
3	Great Britain	278	329	326	933
4	France	247	270	293	810
5	Germany	245	279	297	821
6	Italy	235	209	228	672
7	Hungary	190	168	184	542
8	Japan	189	159	191	539
9	Australia	179	191	227	597
10	Russia	172	171	185	528
11	East Germany	152	129	127	408
12	Sweden	150	173	183	506
13	Netherlands	109	109	133	351
14	Finland	104	85	119	308

Figure 5.1 : Overall medal tally of Olympics

Figure 5.2: Statistics of the Olympics

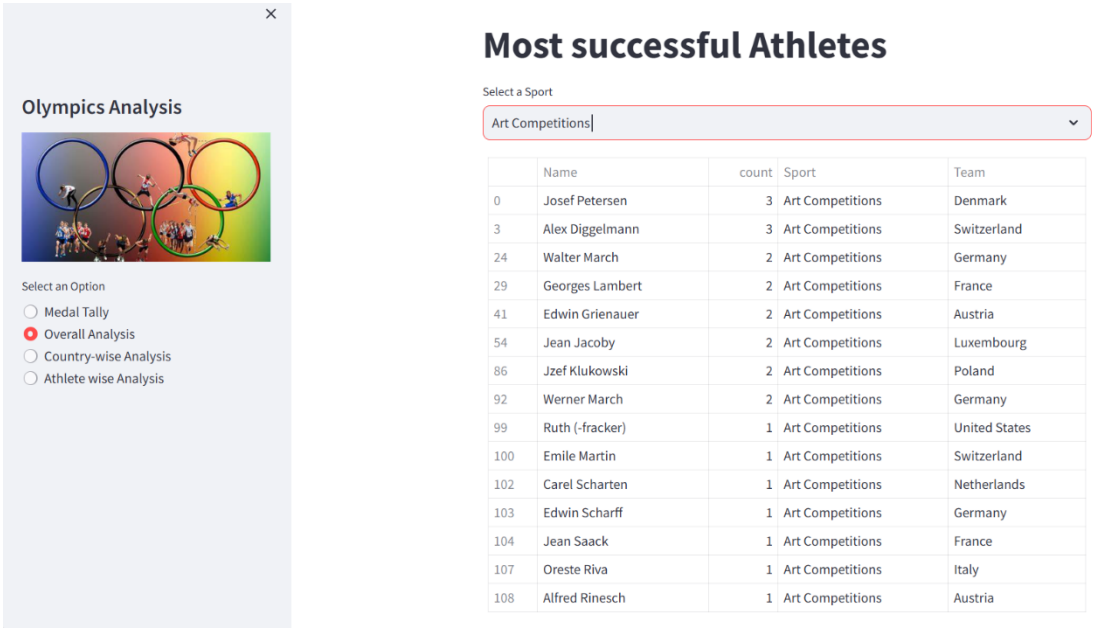


Figure 5.3: Most successful athlete of the particular game

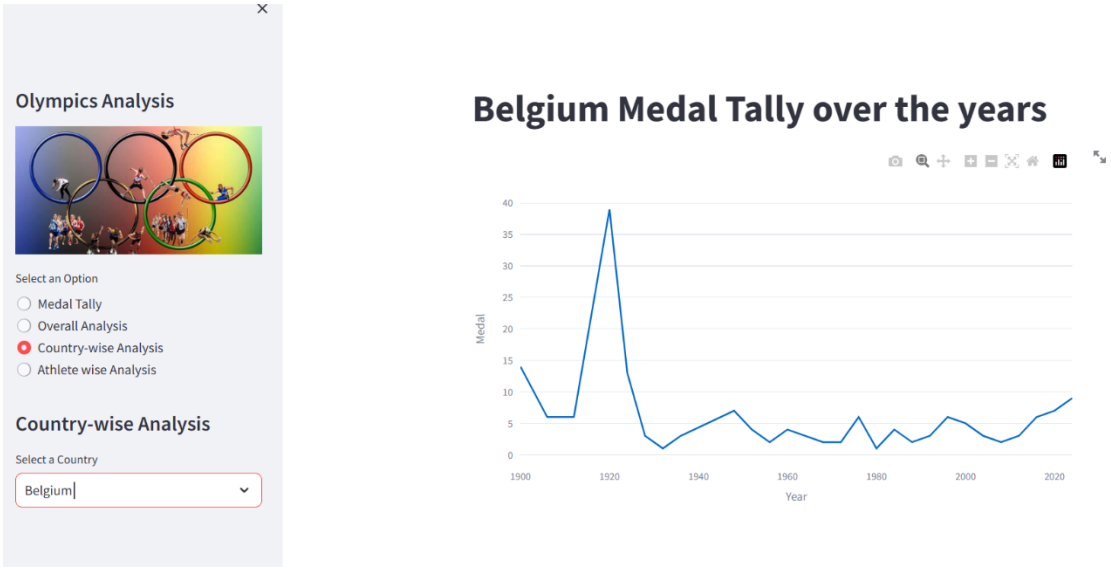


Figure 5.4: Performance Tally of a particular Country

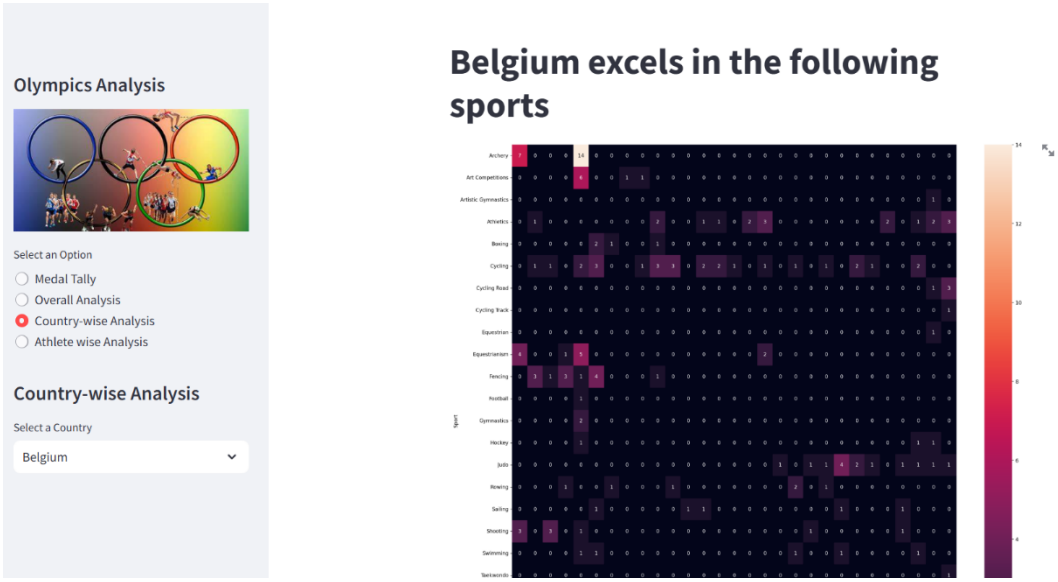


Figure 5.5: Heat map of a particular country

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion:

The development of a website that analyzes Olympic data and doping trends offers a powerful platform for understanding the historical and current landscape of athletic performance and integrity. By integrating various datasets and leveraging data visualization tools, the website provides clear insights into key trends, such as country-specific performance, doping violations, and their impact on the Olympic Games. The use of predictive modeling further enhances the site's value, allowing stakeholders to anticipate future doping risks. This project not only sheds light on the effects of doping but also emphasizes the importance of clean competition and ethical sportsmanship. The accessibility of this analysis makes it useful for sports authorities, analysts, and the general public, fostering transparency in the Olympic Games.

6.2 Future Work:

6.1.1 Real-Time Data Integration:

Incorporate real-time data from upcoming Olympic events and doping test results to provide dynamic, up-to-date analysis.

6.2.1 Sentiment Analysis:

Integrate social media and news sentiment analysis to track public perception surrounding doping scandals and athlete integrity.

6.3.1 Enhanced Predictive Modeling:

Utilize more advanced machine learning algorithms to improve the accuracy and reliability of doping risk predictions for future Olympic events.

6.4.1 Expansion to Other Sporting Events:

Extend the platform to include data and analysis from other global sporting events beyond the Olympics, offering a broader perspective on doping trends.

6.5.1 Personalized Insights and Custom Reports:

Implement user authentication to provide personalized insights, allowing users to generate custom reports tailored to their specific needs (e.g., country-specific performance, sport-specific trends).

6.6.1 Advanced Data Visualization:

Add more sophisticated visualizations, including network graphs and interactive maps, to explore relationships between athletes, countries, and doping cases in greater detail.

6.7.1 Collaborations with Sports Authorities:

Partner with sports authorities and anti-doping agencies to improve data accuracy and enhance the predictive modeling capabilities for real-world applications.

REFERENCES

Journal

[1] Forrest, D., McHale, I. G., Sanz, I., & Tena, J. D. (2017). An analysis of country medal shares in individual sports at the Olympics. *European Sport Management Quarterly*, *17*(2), 117-131.

[2] Tsivou, M., Kioukia-Fougia, N., Lyris, E., Aggelis, Y., Fragkaki, A., Kiousi, X., ... & Georgakopoulos, C. (2006). An overview of the doping control analysis during the Olympic Games of 2004 in Athens, Greece. *Analytica chimica acta*, *555*(1), 1-13.

[3] Kolliari-Turner, A., Lima, G., Hamilton, B., Pitsiladis, Y., & Guppy, F. M. (2021). Analysis of anti-doping rule violations that have impacted medal results at the summer olympic games 1968–2012. *Sports Medicine*, *51*, 2221-2229.

[4] Pereira, H. M. G., Sardela, V. F., Padilha, M. C., Mirotti, L., Casilli, A., de Oliveira, F. A., ... & de Aquino Neto, F. R. (2017). Doping control analysis at the Rio 2016 Olympic and Paralympic Games. *Drug testing and analysis*, *9*(11-12), 1658-1672.

[5] Sharma, R. (2016). Analytical study of doping cases of banned substances during Olympics games from 1968 to 2012. *Int. J. Phys. Educ. Sports Health*, *3*, 3-37.

[6] Spyridaki, M. H., Kiousi, P., Vonaparti, A., Valavani, P., Zonaras, V., Zahariou, M., ... & Georgakopoulos, C. (2006). Doping control analysis in human urine by liquid chromatography–electrospray ionization ion trap mass spectrometry for the Olympic Games Athens 2004: Determination of corticosteroids and quantification of ephedrines, salbutamol and morphine. *Analytica Chimica Acta*, *573*, 242-249.

Link

[1] Doping at the Olympic Games - Wikipedia

[2] app · Streamlit

APPENDIX

Coding

```
import streamlit as st
import pandas as pd
import preprocessor,helper
import plotly.express as px
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.figure_factory as ff

df = pd.read_csv('olympics_dataset.csv')
#region_df = pd.read_csv('noc_regions.csv')

#df = preprocessor.preprocess(df,region_df)
df = preprocessor.preprocess(df)

st.sidebar.title("Olympics Analysis")
st.sidebar.image('https://th.bing.com/th/id/OIP.bBI9ZHgiXFWupoayJgNySQHaD3?rs=1&pid=ImgDetMain')
user_menu = st.sidebar.radio(
    'Select an Option',
    ('Medal Tally','Overall Analysis','Country-wise Analysis','Athlete wise Analysis')
)

if user_menu == 'Medal Tally':
    st.sidebar.header("Medal Tally")
    years,country = helper.country_year_list(df)

    selected_year = st.sidebar.selectbox("Select Year",years)
    selected_country = st.sidebar.selectbox("Select Country", country)

    medal_tally = helper.fetch_medal_tally(df,selected_year,selected_country)
    if selected_year == 'Overall' and selected_country == 'Overall':
        st.title("Overall Tally")
    if selected_year != 'Overall' and selected_country == 'Overall':
        st.title("Medal Tally in " + str(selected_year) + " Olympics")
    if selected_year == 'Overall' and selected_country != 'Overall':
        st.title(selected_country + " overall performance")
    if selected_year != 'Overall' and selected_country != 'Overall':
        st.title(selected_country + " performance in " + str(selected_year) + " Olympics")
```

```
 st.table(medal_tally)

user_menu == 'Overall Analysis':
 editions = df['Year'].unique().shape[0] - 1
 cities = df['City'].unique().shape[0]
 sports = df['Sport'].unique().shape[0]
 events = df['Event'].unique().shape[0]
 athletes = df['Name'].unique().shape[0]
 #nations = df['region'].unique().shape[0]
 nations = df['Team'].unique().shape[0]


 st.title("Top Statistics")
 col1,col2,col3 = st.columns(3)
 with col1:
     st.header("Editions")
     st.title(editions)
 with col2:
     st.header("Hosts")
     st.title(cities)
 with col3:
     st.header("Sports")
     st.title(sports)

 col1, col2, col3 = st.columns(3)
 with col1:
     st.header("Events")
     st.title(events)
 with col2:
     st.header("Nations")
     st.title(nations)
 with col3:
     st.header("Athletes")
     st.title(athletes)

 nations_over_time = helper.data_over_time(df,'Team')
 fig = px.line(nations_over_time, x="Year", y="No. of countries")
```

```
    nations_over_time = helper.data_over_time(df,'Team')
    fig = px.line(nations_over_time, x="Year", y="No. of countries")
    st.title("Participating Nations over the years")
    st.plotly_chart(fig)

    events_over_time = helper.event_over_time(df, 'Event')
    fig = px.line(events_over_time, x="Year", y="No. of event")
    st.title("Events over the years")
    st.plotly_chart(fig)

    athlete_over_time = helper.athletes_over_time(df, 'Name')
    fig = px.line(athlete_over_time, x="Year", y="Name")
    st.title("Athletes over the years")
    st.plotly_chart(fig)

    st.title("No. of Events over time(Every Sport)")
    fig,ax = plt.subplots(figsize=(20,20))
    x = df.drop_duplicates(['Year', 'Sport', 'Event'])
    ax = sns.heatmap(x.pivot_table(index='Sport', columns='Year', values='Event', aggfunc='count').fillna(0).astype('int'),
                annot=True)
    st.pyplot(fig)

    st.title("Most successful Athletes")
    sport_list = df['Sport'].unique().tolist()
    sport_list.sort()
    sport_list.insert(0,'Overall')

    selected_sport = st.selectbox('Select a Sport',sport_list)
    x = helper.most_successful(df,selected_sport)
    st.table(x)

if user_menu == 'Country-wise Analysis':

    st.sidebar.title('Country-wise Analysis')

    country_list = df['Team'].unique().tolist()
```

```
user_menu == 'Country-wise Analysis':

 st.sidebar.title('Country-wise Analysis')

 country_list = df['Team'].unique().tolist()
 country_list.sort()

 selected_country = st.sidebar.selectbox('Select a Country',country_list)

 country_df = helper.yearwise_medal_tally(df,selected_country)
 fig = px.line(country_df, x="Year", y="Medal")
 st.title(selected_country + " Medal Tally over the years")
 st.plotly_chart(fig)

 st.title(selected_country + " excels in the following sports")
 pt = helper.country_event_heatmap(df,selected_country)
 fig, ax = plt.subplots(figsize=(20, 20))
 ax = sns.heatmap(pt,annot=True)
 st.pyplot(fig)

 st.title("Top 10 athletes of " + selected_country)
 top10_df = helper.most_successful_countrywise(df,selected_country)
 st.table(top10_df)

user_menu == 'Athlete wise Analysis':
 athlete_df = df.drop_duplicates(subset=['Name', 'Team'])

 x1 = athlete_df['Age'].dropna()
 x2 = athlete_df[athlete_df['Medal'] == 'Gold']['Age'].dropna()
 x3 = athlete_df[athlete_df['Medal'] == 'Silver']['Age'].dropna()
 x4 = athlete_df[athlete_df['Medal'] == 'Bronze']['Age'].dropna()

 fig = ff.create_distplot([x1, x2, x3, x4], ['Overall Age', 'Gold Medalist', 'Silver Medalist', 'Bronze Medalist'],show_hist=False,
 fig.update_layout(autosize=False,width=1000,height=600)
 st.title("Distribution of Age")
 st.plotly_chart(fig)
```

```python
def data_over_time(df,col):

    nations_over_time = df.drop_duplicates(['Year' , 'Team'])['Year'].value_counts().reset_index().sort_values('Year')
    nations_over_time.rename(columns = {'Year':'Year' , 'count' :'No. of countries' }, inplace = True)
    return nations_over_time

def event_over_time(df,col):

    event_over_time = df.drop_duplicates(['Year' , 'Event'])['Year'].value_counts().reset_index().sort_values('Year')

    event_over_time.rename(columns = {'Year':'Year' , 'count' :'No. of event' }, inplace = True)
    return event_over_time


def athletes_over_time(df,col):

    athletes_over_time = df.drop_duplicates(['Year' , 'Name'])['Year'].value_counts().reset_index().sort_values('Year')
    athletes_over_time.rename(columns = {'Year':'Year' , 'count' :'Name' }, inplace = True)
    return athletes_over_time


# def most_successful(df, sport):
#     temp_df = df.dropna(subset=['Medal'])

#     if sport != 'Overall':
#         temp_df = temp_df[temp_df['Sport'] == sport]

#     x = temp_df['Name'].value_counts().reset_index().head(15).merge(df, left_on='index', right_on='Name', how='left')[
#         ['index', 'Name_x', 'Sport', 'region']].drop_duplicates('index')
#     x.rename(columns={'index': 'Name', 'Name_x': 'Medals'}, inplace=True)
#     return x

def most_successful(df , sport):
    temp_df = df[df['Medal'].notna() & (df['Medal'] != 'No medal')]

    if sport != 'overall':
        temp_df = temp_df[temp_df['Sport'] == sport]
```

```python
        return x

    def yearwise_medal_tally(df,country):
        temp_df = df[df['Medal'] != 'No medal']
        #temp_df.drop_duplicates(subset=['Team', 'NOC', 'Games', 'Year', 'City', 'Sport', 'Event', 'Medal'], inplace=True)
        temp_df.drop_duplicates(subset=['Team', 'NOC', 'Year', 'City', 'Sport', 'Event', 'Medal'], inplace=True)


        new_df = temp_df[temp_df['Team'] == country]
        final_df = new_df.groupby('Year').count()['Medal'].reset_index()

        return final_df

    def country_event_heatmap(df,country):
Click to add a breakpoint f[df['Medal'] != 'No medal']
        temp_df.drop_duplicates(subset=['Team', 'NOC',  'Year', 'City', 'Sport', 'Event', 'Medal'], inplace=True)  #'Games',

        new_df = temp_df[temp_df['Team'] == country]

        pt = new_df.pivot_table(index='Sport', columns='Year', values='Medal', aggfunc='count').fillna(0)
        return pt
```

```python
#def preprocess(df,region_df):
def preprocess(df):
    # filtering for summer olympics
    df = df[df['Season'] == 'Summer']
    # merge with region_df
    #df = df.merge(region_df, on='NOC', how='left')
    # dropping duplicates
    df.drop_duplicates(inplace=True)
```