Report On

**Olympics Data Analysis**

Submitted in partial fulfillment of the requirements of the Mini project in

Semester V of Third Year Computer Science and Engineering (Data Science)

by

Niharika Sahu (Roll No. 51)

Harish Sargar (Roll No. 52)

Laxman Sawant (Roll No. 53)

Mentor

Prof. Komal Champanerkar



**University of Mumbai**

**Vidyavardhini's College of Engineering & Technology**

**Department of Computer Science and Engineering (Data Science)**



**(A.Y. 2023-24)**

**Vidyavardhini's College of Engineering & Technology**

**Department of Computer Science and Engineering (Data Science)**

**CERTIFICATE**

This is to certify that the Mini Project entitled **“Olympics Data Analysis”** is a bonafide work of **Niharika Sahu (Roll No. 51) Harish Sargar (Roll No. 52) Laxman Sawant (Roll No. 53)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“Bachelor of Engineering”** in Semester V of Third Year **“Computer Science and Engineering [Data Science]” .**

|  |  |  |
| --- | --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Prof. Komal Champanerkar  Mentor | | |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Prof. Yogesh Pingle  Deputy-HOD CSE(DS) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Dr. Vikas Gupta  HOD CSE(DS) | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Dr. H.V. Vankudre  Principal |

**Vidyavardhini's College of Engineering & Technology**

**Department of Computer Science and Engineering (Data Science)**

# Mini Project Approval

This Mini Project entitled “**Olympics Data Analysis”** by **Niharika Sahu (Roll No. 51) Harish Sargar (Roll No. 52) Laxman Sawant (Roll No. 53)** is approved for the degree of **Bachelor of Engineering** in in Semester V of Third Year **Computer Science and Engineering [Data Science].**

**Examiners**

**1………………………………………**

(Internal Examiner Name & Sign)

## 2……………………………………….

(External Examiner name & Sign)

Date:

Place:

# Contents

# Abstract i

## Acknowledgments ii

## List of Figure iii

## Introduction 01

* 1. Problem Statement & Objectives 02
  2. Scope 02

1. **Literature Survey 04**
   1. Survey of Existing System 04
   2. Limitation Existing system or Research gap 07
   3. Mini Project Contribution 08

## Proposed System 09

* 1. Introduction 09
  2. Methodology 09
  3. Architecture and Block diagram 13
  4. Algorithm and Process Design 15
  5. Details of Hardware & Software 16
  6. Conclusion 17
  7. Future work. 17

## References 18

4 Annexure

**4.1 Published Paper /Camera Ready Paper/ Business pitch/proof of concept**

**Abstract**

The Olympic Games, an international sporting spectacle, hold a profound significance as a symbol of unity, global camaraderie, and the pursuit of athletic excellence. Athletes from approximately 205 nations converge every four years to partake in a diverse array of around 45 sports, competing with unwavering passion and determination to secure a place on the coveted Olympic podium. This grand event provides a unique opportunity to not only celebrate athletic achievements but also to delve into vast reservoir of data that underpins it.

Our research endeavors to unravel the multifaceted layers of the Olympics, emphasizing the imperative for countries and athletes to glean insights from past performances. By employing data science and analysis, we embark on a journey to transform raw data into valuable insights. We aim to provide a perspective on the Olympics, one that goes beyond the just winning medals and delves into intricacies of historical trends.

A key component of our research lies in the realm of machine learning. We've developed algorithms that predict the winning probabilities of athletes in specific events. These algorithms take into account various factors, including an athlete's physical attributes like height and weight, along with their past performance. By doing so, we aspire to offer a glimpse into the potential outcomes of future Olympic competitions. This predictive element is vital not only for athletes and their trainers but also for countries striving to optimize their medal-winning strategies.

In addition to predictive modeling, our research employs data visualization techniques to create an interactive world map that provides a comprehensive geographical perspective on the distribution of medals won by countries. By simply hovering over a nation, users can explore the historical medal count, fostering a deeper appreciation for the global reach and influence of the Olympics.

In essence, our research is a comprehensive journey through the Olympics' history and its future possibilities. By combining historical analysis, predictive modeling, and global visualization, we aim to create a research framework that will be invaluable to athletes, coaches, and policymakers alike. The Olympics have evolved over the years, and so should our understanding of them. Through this research, we seek to enhance the knowledge base, learn from the past, and chart a course for future achievements in the world's most celebrated sporting event.

**Acknowledgments**

We would like to share our heartfelt gratitude to our dedicated guide, Prof. Komal Champanerkar, for granting us the wonderful opportunity to work under her expert guidance. Prof. Komal's unwavering support, motivation, and inspiration played a pivotal role in shaping our Mini Project. We also wish to express our sincere appreciation to Dr. Vikas Gupta, the Head of the Computer Science and Engineering (Data Science) department. His continuous support and encouragement were instrumental in our project's success.

Our gratitude further extends to Prof. Yogesh Pingle, the Deputy Head of the Computer Science and Engineering (Data Science) Department. His valuable guidance throughout our Mini Project journey was invaluable. We are also indebted to the supportive staff of the Computer Science and Engineering (Data Science) department for providing us with all the necessary resources and facilities, which made our project achievable.

In addition, we would like to express our heartfelt thanks to Mrs. Maya Varghese, our Mini Project coordinator. Her assistance and guidance were greatly appreciated and contributed significantly to the project's success. Last but not least, we wish to express our profound gratitude to the almighty god and our parents for their blessings and unwavering support. Without their love and encouragement, this task would have been impossible to accomplish.

**List of Figures**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr No.** | **Figure No.** | **Name** | **Page No.** |
| 1 | Fig 1.2.1 | Scope for Olympic Data Analysis | 02 |
| 2 | Fig 3.2.1 | Top Statistics and Graphs | 10 |
| 3 | Fig 3.2.2 | Graph for Distribution of Age wrt Sports | 10 |
| 4 | Fig 3.2.3 | Visualization on World Map | 11 |
| 5 | Fig 3.2.4 | Predicting Medal Winners | 12 |
| 6 | Fig 3.3.1 | Flow Diagram | 13 |
| 7 | Fig 3.3.2 | System Architecture | 13 |

1. **Introduction**

The Olympic Games stand as a shining beacon of unity and athletic excellence, where athletes from across the world converge to showcase their remarkable talents. With approximately 45 diverse sports, these Games are a true spectacle of human prowess and dedication. Over 205 countries participate, each fervently vying to clinch the coveted Olympic medals and bring honor to their homeland. Held once every four years, the Olympics represent a unique opportunity to unravel a treasure trove of insights through data analysis. By delving into the vast repository of historical data, we can uncover valuable trends, athlete performance metrics, and factors that influence the ultimate outcome of these grand sporting events. This is where the world of data science and analysis steps in.

Our project embarks on a journey into the rich tapestry of Olympic data, using the versatile tool of Python. We transform raw data into insightful graphs and structured data frames, unveiling patterns and trends that may otherwise remain hidden. As we sift through the numbers and statistics, we gain a deeper understanding of the dynamics of the Olympics, both past and present. The intrigue doesn't stop there. With the aid of geographical data visualization, we take our analysis to the world stage. By plotting the medal counts on a world map, we offer a unique perspective on each country's performance in the Olympics. With a simple hover over a nation, you can see the medals it has garnered over the years. This adds a layer of comprehension and appreciation for the global scope of the Games.

But we don't stop at the past; we also look to the future. In our project, we've incorporated machine learning algorithms to predict the probability of an athlete winning a medal in a particular event. By considering variables like height, weight, and past performance, we aim to provide a glimpse into the potential outcomes of future Olympic competitions. In essence, our Olympic data analysis project is a multi-faceted exploration of one of the world's most significant sporting events. We dive into the data, offer a visual journey across the world, and even take a leap into the realm of predictive analytics. Through these endeavors, we seek to shed light on the remarkable stories that unfold at the Olympics and the incredible human achievements that transpire on the world's grandest stage.  
 In this era of data-driven insights, our Olympic data analysis project serves as a tribute to the spirit of human athleticism and the power of data science. By delving into the past, present, and future of the Olympic Games, we aim to celebrate not only the numbers but also the extraordinary stories and aspirations of athletes who inspire us all. Join us on this exhilarating journey as we unveil the hidden narratives and discoveries behind the world's greatest sporting spectacle.

**1.1 Problem Statement**

Analyzing the Olympic Games' vast dataset, which spans over a century, poses a significant challenge for individuals and organizations. This challenge arises from the sheer volume of data, including athlete details, event results, and historical records, encompassing hundreds of thousands of athletes, thousands of events across multiple Olympic Games, and gigabytes of data in diverse formats. The complexity of the data further compounds the challenge, as it involves multiple Olympic Games with distinct rules and formats, evolving event categories, and the influence of shifting geopolitical landscapes on athlete participation. Consequently, researchers and analysts confront information overload when attempting to extract meaningful insights from this wealth of data. Manual analysis, the typical approach, proves time-consuming and often inefficient, hindering the ability to draw timely conclusions. Addressing these challenges is vital to unlock the Olympic dataset's full potential, enabling a deeper understanding of both the historical and evolving aspects of this iconic global sporting event.

**1.2 Scope**



**Fig. 1.2.1 Scope for Olympic Data Analysis**

The scope of your Olympic data analysis project using Data Science is multifaceted and includes the following key components as follows:

**1] Historical Analysis:** This project delves into the historical data of the Olympic Games. It involves collecting, processing, and analyzing data from past Olympics to uncover trends, patterns, and insights. This analysis can include exploring changes in medal counts, athlete performance, and other historical data points.

**2] Predictive Modeling:** By incorporating machine learning algorithms, the project extends its scope to predicting athlete winning probabilities. This involves creating models that take into account various factors, such as an athlete's physical attributes and past performance, to forecast their likelihood of winning a medal in a particular event.

**3] Data Visualization:** The project includes data visualization techniques to present the findings in a visually appealing and informative manner. This can include creating graphs, charts, and interactive world maps to display historical medal counts by country and provide a geographical perspective on Olympic performance.

**4] Global Perspective:** The use of data visualization to map the distribution of medals won by countries across the world gives a global perspective on the Olympics. This adds a unique dimension to the project by allowing users to explore the historical success of countries in the Games on a world map.

**5] Learning from the Past:** One of the underlying themes of this project is the importance of countries and athletes learning from past performances. The scope includes providing insights and data-driven recommendations that can inform future strategies and decisions for athletes, coaches, and national Olympic committees.

**6] Comprehensive Insights:** The project aims to provide a holistic view of the Olympics' evolution, achievements, and trends. By combining historical analysis, predictive modeling, and data visualization, it seeks to offer a comprehensive understanding of the Olympic Games.

**7] Practical Applications:** The insights derived from this project have practical applications for athletes, coaches, and policymakers. They can use the findings to make informed decisions, optimize training, and develop strategies for future Olympic events.

**8] Educational Value:** The project has educational value. It can serve as a valuable resource for those interested in the Olympics, data analysis, and machine learning. It can be a learning tool for students and researchers in the field of sports analytics and data science.

In summary, the scope of this project encompasses historical analysis, predictive modeling, data visualization, and a global perspective on the Olympics. Its goal is to provide insights, inform decision-making, and contribute to a deeper understanding of the significance of the Olympic Games in the world of sports and data analysis.

**2. Literature Survey**

* **2.1 Survey of Existing System:**

**1. Olympic Data Analysis using Data Science**

**Author:** Nishant Kulkarni, Pratik Patil, Rugved Pande, Dhiraj Patil, Pranav Nair.

**Year**: 2022

**Techniques Used:**

* The primary technique used in this paper is EDA which is an approach of analyzing large datasets by summarizing data in graphical formats such as histograms, bar, box, & scatter plots.
* In this paper, EDA is used to visualize and analyze various factors related to the Olympics, including the performance of countries, the contribution of men and women, and trends over time

**Future Scope:**

* Geographical Visualization. The authors suggest that data could be visualized on a world map to provide a geographical perspective on countries performance in the Olympics. This could lead to more comprehensive insights.
* Machine Learning Predictions: As mentioned earlier, the paper could benefit from the application of machine learning algorithms for predictive analysis. Predictive models could be developed to forecast Olympic performance or identify factors contributing to success.

**2. Analyzing Evolution of the Olympics by Exploratory Data Analysis Using R**

**Author:** Rahul Pradhan, Kartik Agrawal and Anubhav Nag

**Year**: 2021

**Techniques Used:**

* The research paper employs EDA which is nothing but Exploratory Data Analysis as the primary original technique to analyze and understand the large dataset which is related to the most famous Olympics.
* EDA involves examining large datasets and exploring various characteristics through visual formats like graphs and charts.
* In this case, EDA is used to explore various factors contributing to the evolution of the Olympic Games over the years

**Future Scope:**

* Further Predictive Modelling: The authors can expand their analysis by developing predictive models to forecast future trends in Olympic Games based on historical data..
* Machine Learning Applications: Incorporating ML algorithms for more in-depth analysis, to gain a more comprehensive understanding of the factors affecting Olympic performance.
* Geospatial Analysis: Utilizing geographical data to analyze the geographic distribution of Olympic events and their impact on host cities and regions.
* Incorporating more recent data: As the paper was published in 2021, including data from subsequent Olympic Games to provide a more up-to-date analysis.

**3. Web Application of the Olympic data analysis.**

**Author:** Farkande Vaishnavi, Gurav Vaishnavi, Borse Tejas

**Year**: 2012

**Techniques Used:**

* The paper primarily employs Exploratory Data Analysis (EDA) to analyze Olympic data from 1896 to 2016.
* EDA involves examining large datasets and exploring various characteristics through visual formats like graphs and charts.
* The use of visual methods to provide a deep understanding and statistical summary of the data is highlighted.

**Future Scope:**

* Visualizing data in geographical formats, such as mapping countries' Olympic performance on a world map.
* Exploring the application of machine learning algorithms to the dataset to create predictive models for future Olympic Games.
* Conducting correlation analysis to analyze relationships between different variables in the dataset.
* These future scopes aim to enhance the analysis of the Olympics' evolution and provide more insights into the dataset.
* Visualizing data in geographical formats, such as mapping countries' Olympic performance on a world map.

**4. Data Analytics on Olympics Datasets.**

**Author:** Surya Sena Reddy, Suraj Kumar

**Year**: 2022

**Techniques Used:**

* The research paper employs data analytics techniques, particularly Descriptive Analytics and Predictive Analytics, to analyze data from the Olympics datasets.
* They utilize Python and various libraries, including NumPy, Pandas, Scikit-learn (Sklearn), Plotly, Streamlit, and Matplotlib for data analysis and visualization.
* **Future Scope:**
* The paper suggests areas for future research, including improving the accuracy of predictive models. For example, enhancing the weight prediction model based on height and exploring more advanced machine learning algorithms could be a potential future scope.
* Furthermore, expanding the analysis to include data beyond 2016 could provide insights into more recent Olympic events and trends.

**5. Web Application of Olympic Data Analysis**

**Author:** Shivam Bansal, Vaibhav Yadav, Sahil, Yashshavi Prajapti, Rashmi Tiwari

**Year**: 2023

**Techniques Used:**

* Data Collection and Preprocessing: Collected Olympic data from public sources and prepared it for analysis by cleaning and focusing on Summer Olympics data. Exploratory Data Analysis (EDA).
* Analyzed the data using Python libraries to gain insights into Olympic history, country performance, and athlete factors.

**Future Scope:**

* Implement predictive modeling using machine learning for Olympic predictions. Extend the application to cover major sporting events like the Commonwealth Games and Asian Games.
* Include real-time updates and live scores for ongoing Olympic events. Ongoing research for deeper insights into sports and countries' performance.
* **2.2 Limitation Existing System or Research Gap:**

**1. Olympic Data Analysis using Data Science**

**Drawbacks:**

* One potential drawback of this paper is that it does not mention the use of any specific machine learning algorithms for predictive analysis. While EDA is a powerful technique for data exploration and visualization, it is primarily descriptive in nature
* To enhance the paper's analytical capabilities, the authors could have explored the application of machine learning algorithms to predict future trends or outcomes in Olympic events.

**2. Analyzing Evolution of the Olympics by Exploratory Data Analysis Using R**

**Drawbacks:**

* Some potential limitations could include the reliance on historical data, the choice of analytical techniques, and the assumptions made during the analysis.
* Additionally, the paper primarily focuses on data visualization and analysis but may not delve into more advanced statistical or machine-learning methods for predictive modeling.

**3. Web Application of the Olympic data analysis.**

**Drawbacks:**

* The paper briefly mentions potential drawbacks but does not go into detail. These drawbacks include the possible variation in analysis due to historical or geographical changes.
* There is a focus on visualizing data in graphical formats, but potential limitations of this approach, such as data quality issues or biases in the data, are not thoroughly discussed.

**4. Data Analytics on Olympics Datasets**

**Drawbacks:**

* The research paper doesn't explicitly mention any drawbacks.
* However, based on the content provided, a potential limitation could be the assumption that linear regression is the best model for predicting athlete weight based on height, as the accuracy is around 63 percent. More sophisticated machine learning models might be explored for better predictions.

**5. Web Application of Olympic Data Analysis**

**Drawbacks:**

* Data Collection and Preprocessing: Collected Olympic data from public sources and prepared it for analysis by cleaning and focusing on Summer Olympics data.
* Exploratory Data Analysis (EDA): Analyzed the data using Python libraries to gain insights into Olympic history, country performance, and athlete factors.

**2.3 Mini Project Contribution:**

In our collaborative mini project, each team member played a distinct yet integral role, contributing their expertise to create a comprehensive and well-rounded final product. Laxman's accurate and complete work was pivotal in the analysis and prediction components. He handled the intricate data processing and cleaning, skillfully applied label encoding and decoding techniques, and perfectly trained the predictive models. Harish, on the other hand, brought a unique dimension to our project by focusing on geospatial mapping and analysis. His contribution allowed us to visualize and explore the data in a spatial context. Niharika, our third team member, excelled in creating the project's welcoming home page and informative about page, providing a user-friendly gateway to our work. Additionally, she contributed to the analysis component, enriching the project's overall depth. Together, our combined efforts resulted in a well-rounded and impactful mini project that successfully integrates various facets of data analysis, visualization, and user interaction.

**3. Proposed System**

**3.1 Introduction:**

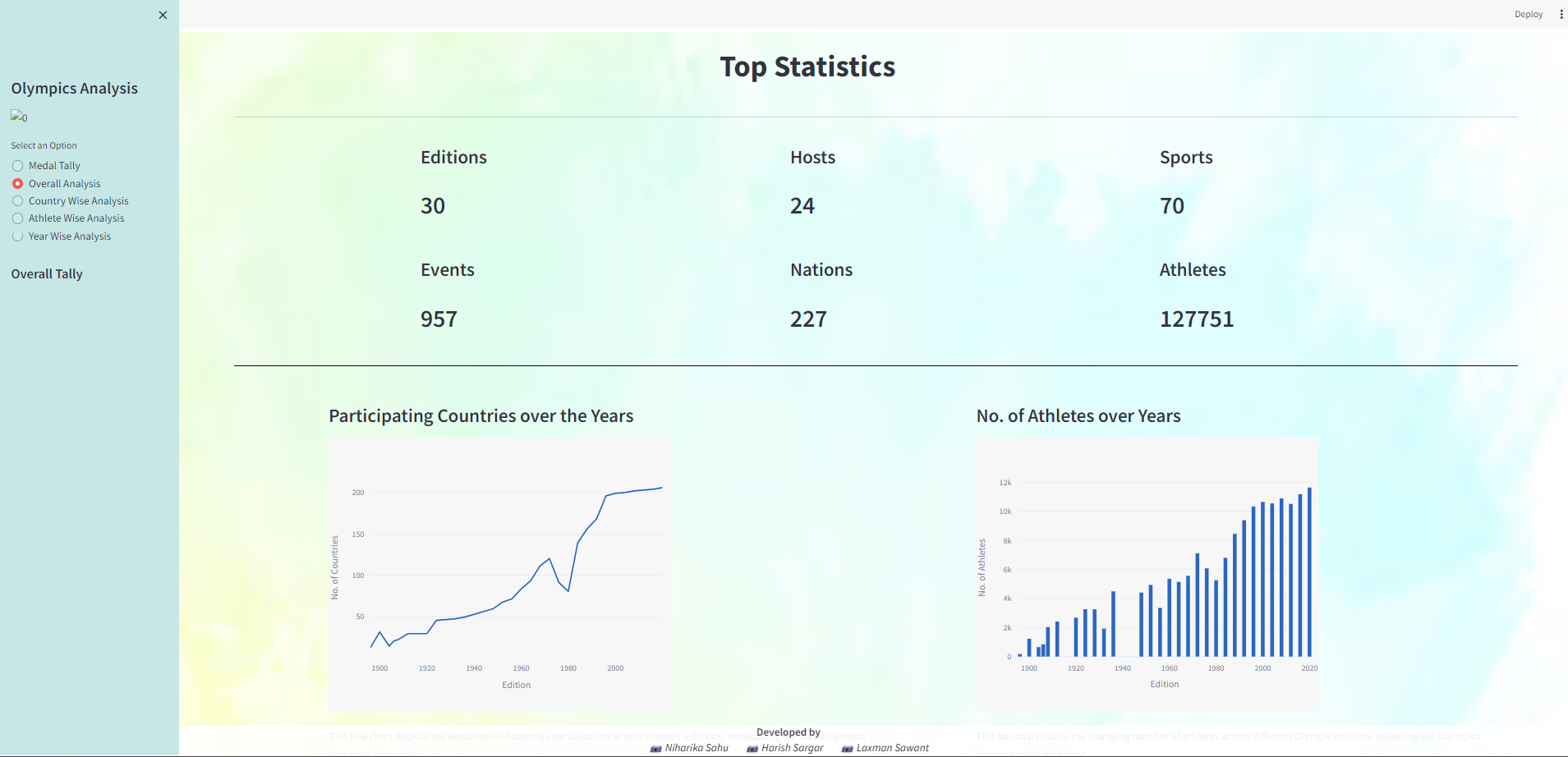
The proposed system represents a approach to address the crucial challenge of analyzing Olympic data comprehensively and effectively. Utilizing the power of data analytics techniques, our project aims to efficiently process and analyze the vast Olympic dataset. Our primary objective is to uncover valuable insights and hidden patterns, presenting this wealth of information in a manner that is not only understandable but also accessible to a wide-ranging audience. To achieve this, we intend to develop a versatile platform that employs graphical visualization, data structuring techniques, and statistical analysis to reveal meaningful trends within the data.

Furthermore, our project will go beyond conventional analysis by delving into the realm of predictive modeling, seeking to create an accurate model capable of forecasting future Olympic medalists. To enhance the user experience and interaction, we will also incorporate a geospatial data visualization component, allowing users to explore and analyze statistics related to each participating nation on an interactive map. This comprehensive system promises to be a valuable resource for both enthusiasts and stakeholders in the world of the Olympic Games.

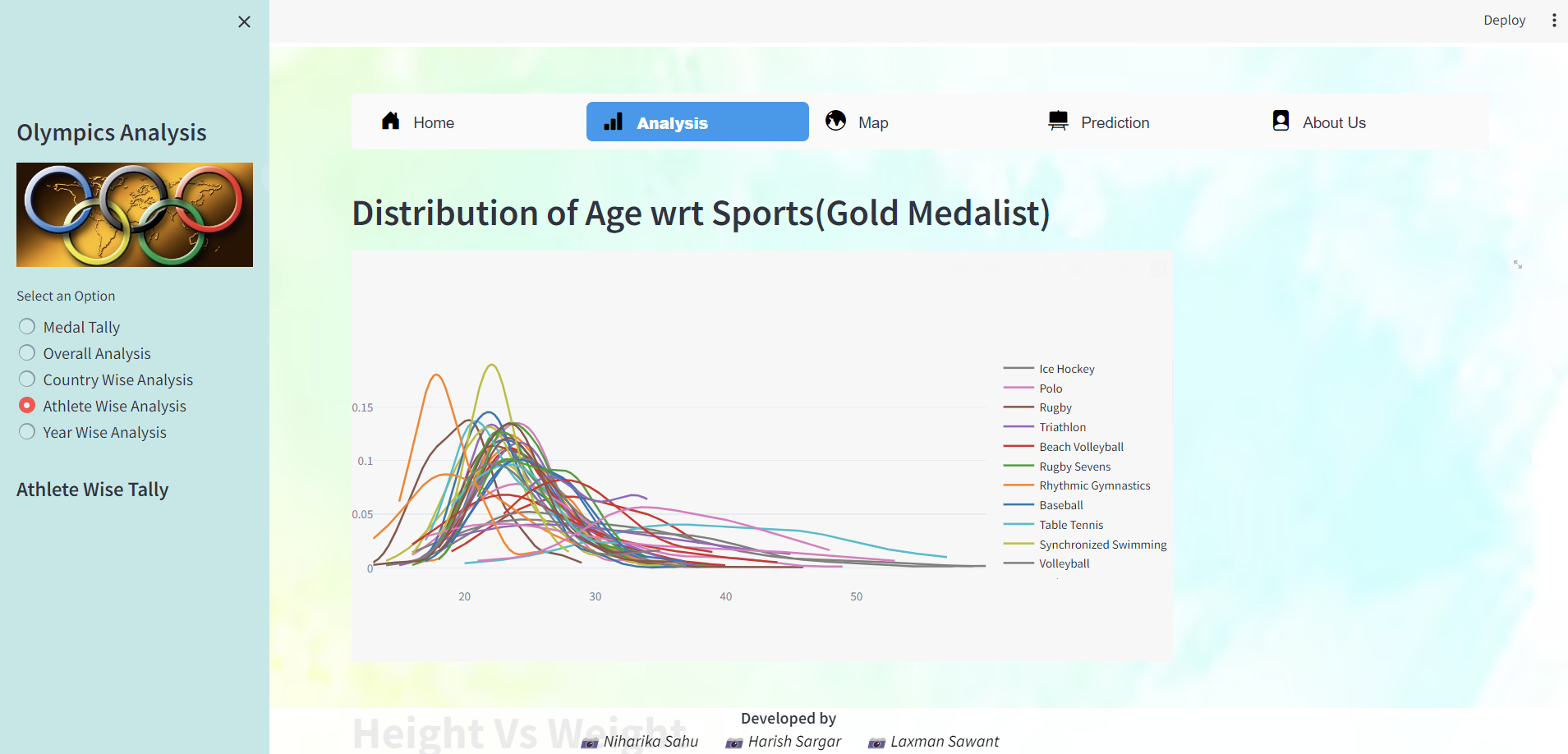
**3.2 Methodology:**

**1] Analysis:**

* **Define the Purpose and Scope:** We first clearly stated the objectives of the data analysis report and defined its scope.
* **Data Collection and Preparation:** We gathered and cleaned the data needed for analysis. We also handled missing values, outliers, and inconsistencies to ensure data integrity.
* **Data Exploration:** We performed exploratory data analysis (EDA) using Jupyter notebook IDE to understand the dataset's characteristics. We generated summary statistics, distributions, and basic visualizations.
* **Hypotheses and Questions:** We listed the hypotheses or questions we aim to answer through the analysis. This provides a clear focus for the report.
* **Findings and Insights:** We presented the results of the analysis, including any patterns, relationships, or anomalies discovered. We interpreted the data using various visualization tools and provided insights.
* **Data Visualizations:** We presented a variety of visualizations, including charts, graphs, and plots, to illustrate key trends. We used visualization tools like Matplotlib, Seaborn, etc.
* **Methodology and Techniques:** We described the analytical methods, techniques, and models used in the analysis. Explained why these methods were chosen.
* **Conclusion and Recommendations:** We summarized the main findings and provided conclusions based on the analysis. We also offered recommendations or actionable insights where relevant.
* **Limitations and Assumptions:** We clearly outlined the limitations of the data analysis, because of any assumptions we made during the process. We have acknowledged potential sources of bias or uncertainty that could affect the results.



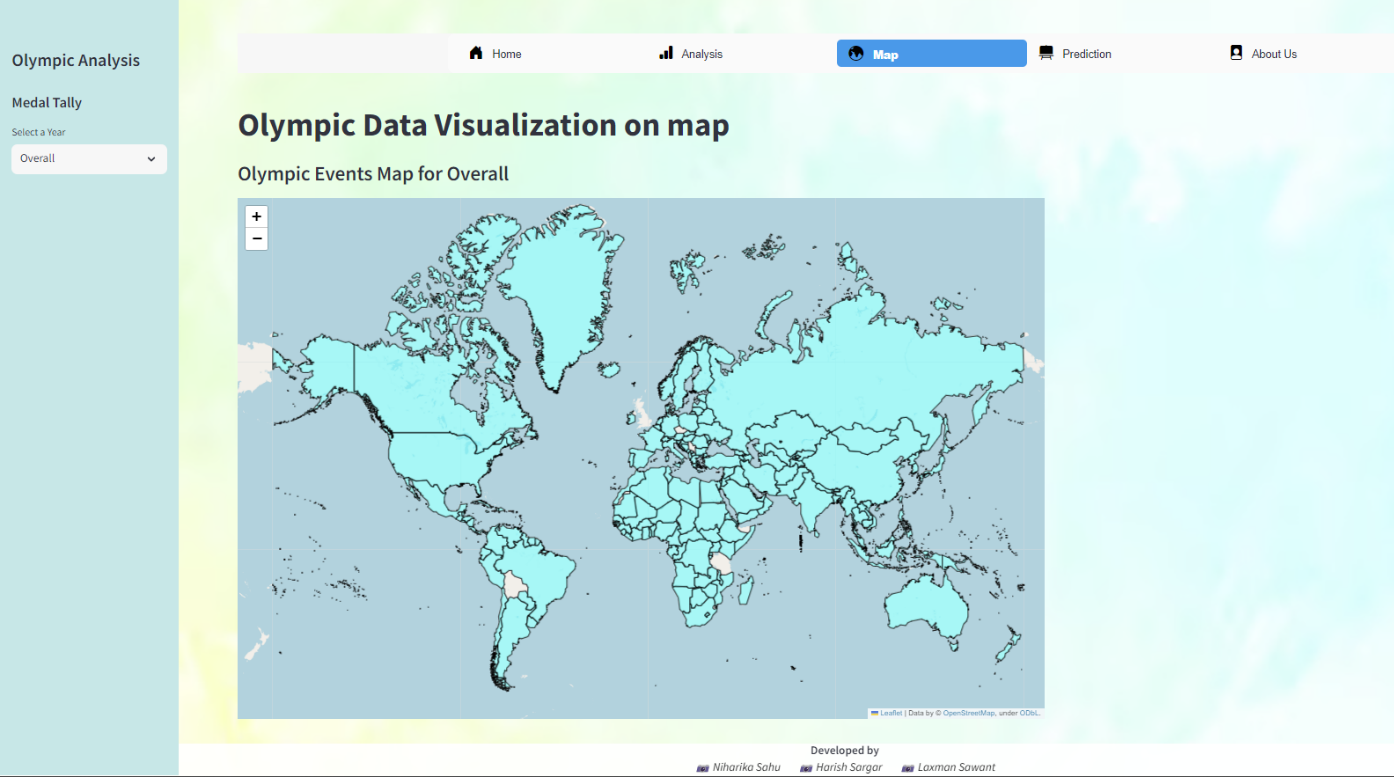
**Fig. 3.2.1 Top Statistics and Graphs**

****

**Fig. 3.2.2 Graph for Distribution of Age wrt Sports**

**2] Data Visualization on a World Map**

* **Data Collection:** We gathered historical Olympic data, including country names, years, and medal counts.
* **Data Preprocessing:** We cleaned and standardize the data, handling missing values.
* **Medal Tally Calculation:** With the preprocessed data, We calculate medal tallies for each country. This information forms the basis of our data visualization
* **Geospatial Data Integration:** We merge this data with geospatial information. This step establishes the link between a country's medal tally and its geographical location.
* **Map Center Determination:** Spain serves as the default center for map display. This is essential for the initial displaying of the map.
* **Color Scale Configuration:** We configured color scales to show countries medal counts on a world map when hovered on a particular place.
* **GeoJSON Layers:** Customized layers on the map visually represent medal counts.
* **Tooltips Customization:** We enhance user interaction by customizing tooltips with medal details. Also using Streamlit, we create an interactive web app for data exploration.
* **Final Output:** The app provides an informative platform for exploring Olympic data, offering insights into medal distribution across countries and years.

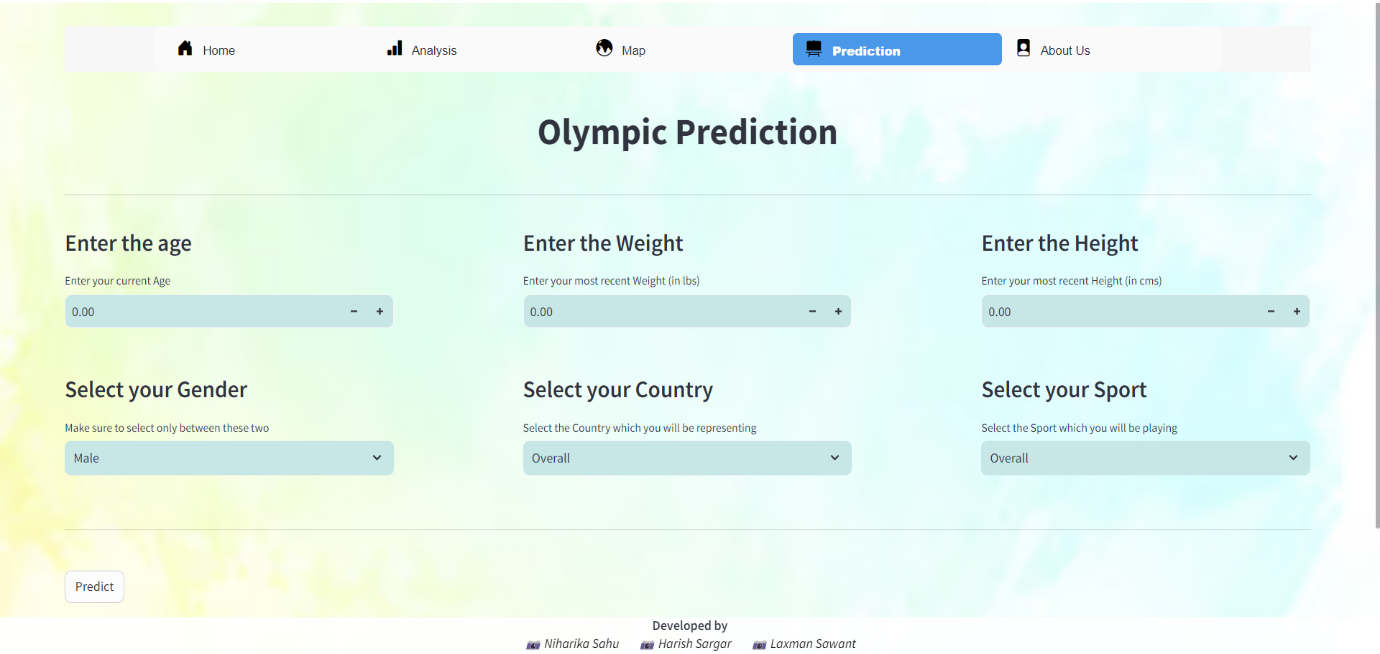


**Fig. 3.2.3 Visualization on World Map**

**3] Prediction**

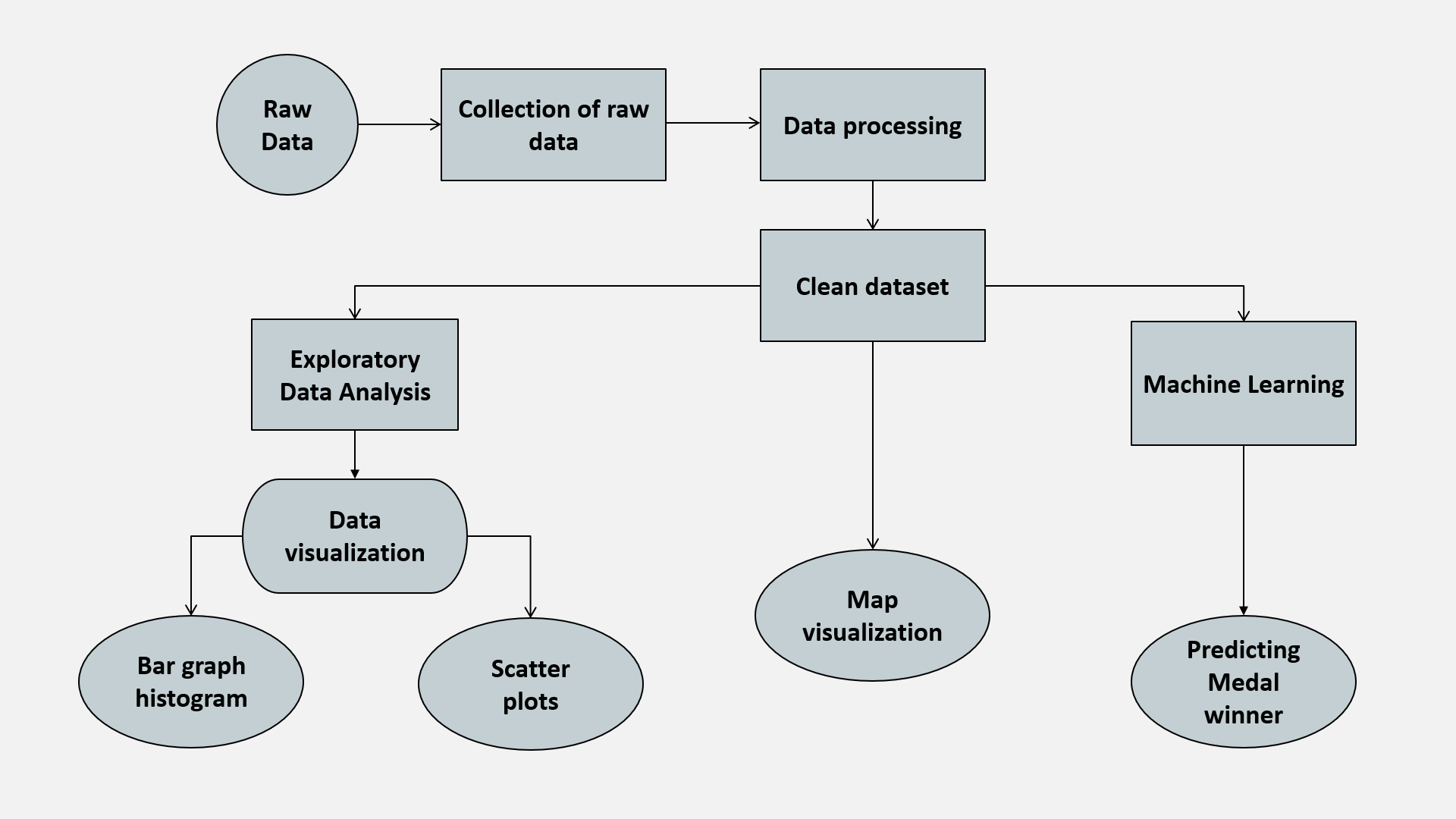
This methodology outlines the steps for building a Logistic Regression model to predict the winning probability of Olympic medals for athletes based on their height, weight, country, and sport. The aim is to develop a model that can provide insights into the likelihood of an athlete winning a medal in the Olympic Games.

* **Data Preparation:** We Collected dataset from Kaggle and cleaned it in Jupyter Notebook which included athlete info, country, sport, and medal data. We merged relevant data based on common keys.
* **Feature Selection:** We choose key features from the resultant dataset like height, weight, country, sport, etc.
* **Model Development:** We used Logistic Regression for Multi-class classification problem (Gold, Silver, Bronze or None).
* **Model Training:** Wetrained the model using the training dataset.
* **Predict Winning Probability:** We used the trained model to predict the winning probability of medals for new athletes or for athletes in historical data.
* **Model Evaluation:** We assessed model performance with metrics like accuracy and F1-score.

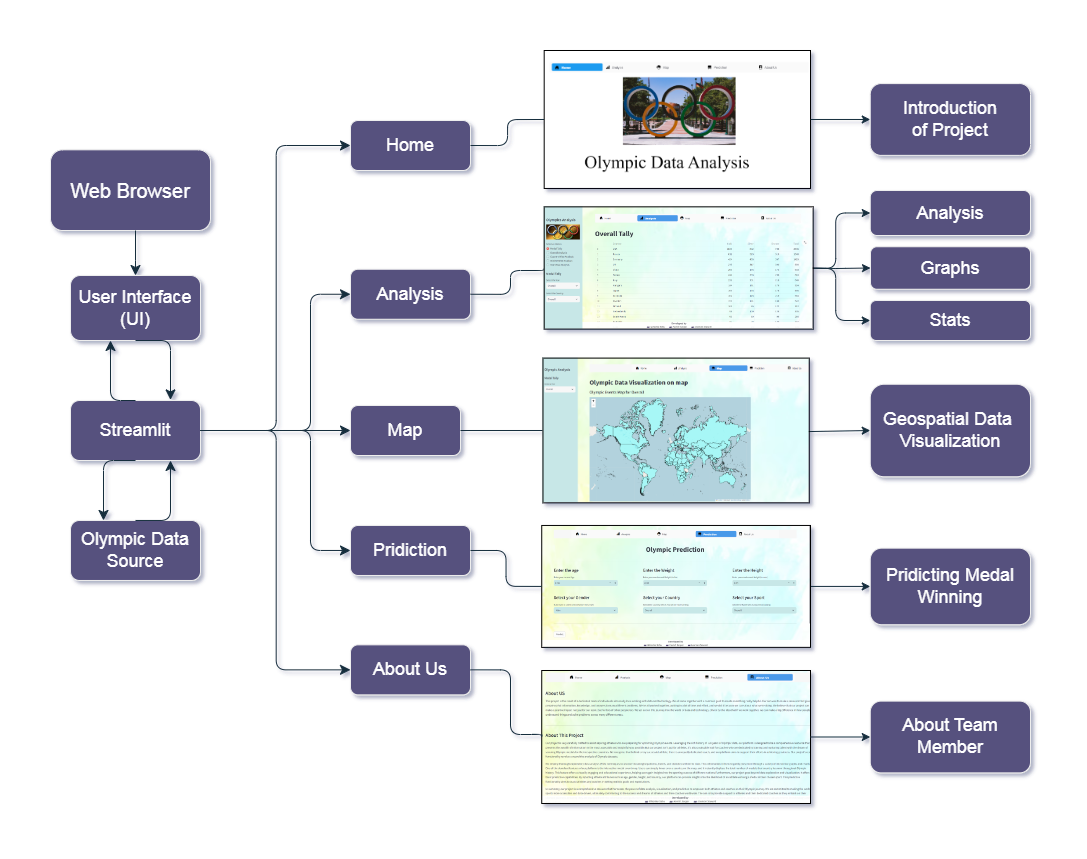


**Fig. 3.2.4 Predicting Medal Winners**

**3.3 Architecture and Block Diagram**



**Fig. 3.3.1 Flow Diagram**



**Fig. 3.3.2 System Architecture**

**Description:**

The architecture of our project is thoughtfully designed to provide a comprehensive and engaging user experience. At its core, we've employed Streamlit, a dynamic framework that seamlessly integrates with our “120 years of Olympic history” dataset, making it easily and efficiently accessible and interactive. Our project is set up like a well-organized book with different chapters. Each chapter covers a different aspect of the Olympics. Our architecture consists of several distinct branches, each serving a unique purpose:

* **Home**: The starting point of our project, the "Home" branch introduces users to the project's overview and goals. It serves as a welcoming entry point, providing an initial understanding of our work.
* **Analysis**: Under the "Analysis" branch, users can delve deeper into the data. It encompasses various sub-branches like "Analysis," where detailed insights into the dataset are offered, "Graphs" for visually representing data trends, and "Stats" to access statistical information, allowing users to explore and analyze Olympic data comprehensively.
* **Map**: The "Map" branch is dedicated to geospatial data visualization. Users can explore interactive maps to gain insights into the geographical distribution of Olympic data, offering a unique perspective on the information.
* **Prediction**: Here, the "Prediction" branch takes center stage. It utilizes data processing, cleaning, and predictive modeling to forecast potential medal winners, offering users an exciting glimpse into the future of Olympic events.
* **About Us**: The "About Us" branch introduces our team members and their respective contributions to the project. This section adds a personal touch, providing context and acknowledging the individuals behind this collaborative effort.

By structuring our project in this way, we aim to provide users with a seamless and informative experience, enabling them to navigate through various facts of the Olympic dataset and uncover valuable insights, predictions, and team background information. This architecture ensures that users can engage with the project in a user-friendly and intuitive manner, making the most of the Olympic data source. This design makes it easy for you to explore and understand Olympic data. It's like a friendly guide, helping you get the most out of the information. Whether you're a sports fan, a researcher, or just curious, this project is here to make the Olympic data fun and useful for you.Top of Form

**3.4 Algorithm and Process Design**

**Algorithm - Logistic Regression**

Logistic Regression is a statistical method used for predicting outcomes in a binary or multi-class classification problem. Unlike linear regression, which is used for predicting continuous values, logistic regression is specifically designed for predicting the probability of a binary outcome (0 or 1, Yes or No, True or False). In this case, we are using Logistic Regression to predict whether an athlete will win a gold, silver, or bronze medal in the Olympics, making it a multi-class classification problem.

**Process Design - How is Logistic Regression used in predicting Medal Winners?**

* **Data Preparation:** To use logistic regression for predicting medal winners, we used a dataset with historical information about athletes, such as their past performance, statistics, and other relevant features. Our dataset also has a target variable named Medal, which represents values as Gold, Silver, Bronze or None.
* **Data Cleaning**: Data cleaning was a pivotal phase in our project's data preparation. We meticulously addressed missing values, ensuring all data points were complete and reliable. Duplicates were removed to maintain dataset integrity. We also standardized formats, particularly for the Medal variable, streamlining it to represent Gold, Silver, Bronze, or None consistently. These steps were essential in creating a high-quality dataset for our logistic regression model, which relies on accurate and consistent data for predicting medal winners.
* **Feature Selection:** We selected a set of features from our dataset that we believed are relevant for predicting medal winners. These features can include factors like previous performance, rankings, age, training hours, or any other variables that may influence the outcome.
* **Label Encoding:** Since logistic regression deals with binary or multi-class classification, we need to encode the medal categories into numerical values. For example, we encoded gold as 1, silver as 2, bronze as 3 and none as 4.
* **Label Decoding:** Just as encoding is crucial for logistic regression, label decoding plays a vital role in making the output more interpretable and user-friendly. After running our model and obtaining numerical predictions, it's essential to decode these numerical values back into their original categories for clear and meaningful presentation. For example, we decoded 1 as 'Gold' 2 as 'Silver' and 3 as 'Bronze' allowing our users to easily understand the medal categories in the results displayed on the Streamlit interface. This label decoding step enhances the user experience and ensures that the output is easily understandable.
* **Model Training:** Logistic regression calculates the probability of an athlete winning each medal category. The logistic regression model learns to predict these probabilities based on the selected features from our dataset. The model is trained by optimizing the model parameters to fit the historical data.
* **Model Dumping:** In the model creation and dumping process, we constructed a predictive logistic regression model using the cleaned dataset. Once the model was trained and optimized, we serialized it using Python's Pickle library. This allowed us to save the model in a binary format for future use, making it readily available for predictions without the need to retrain it each time. Model dumping not only ensures the model's persistence but also offers efficient deployment for real-time or batch predictions, enhancing the project's usability.
* **Prediction:** Once the logistic regression model is trained, we used it to predict the probabilities of an athlete winning each medal category based on their input features.

**3.5 Details of Hardware & Software**

**Hardware Used:**

* Processor: Intel 5
* RAM: 8 GB
* Hard disk: 256 GB required.
* 256 GB Solid State Drive.

**Software Used:**

* Operating System: Windows 10.
* Programming Language: Python 3.6.
* Machine Learning Libraries: Logistic Regression.
* IDE: Jupyter notebook, PyCharm, Visual Studio Code.
* Library: Streamlit, Geopandas, Matplotlib, Plotly.

**3.6 Conclusion**

In summary, our project is a comprehensive resource that harnesses the power of data analysis, visualization, and prediction to empower both athletes and coaches on their Olympic journey. The aim is to provide support to athletes and their dedicated coaches as they embark on their Olympic journeys, ensuring they are equipped with the information and guidance needed to reach new heights of success. One of the standout features of our platform is the interactive medal count map. Users can simply hover over a country on the map, and it instantly displays the total number of medals that country has won throughout Olympic history. Furthermore, our project goes beyond data exploration and visualization. It offers basic predictive capabilities. By inputting athlete attributes such as age, gender, height, and country, our platform can provide insights into the likelihood of an athlete winning a medal in their chosen sport. This predictive functionality aims to assist athletes and coaches in setting realistic goals and expectations. In summary, our project is a valuable tool that uses data analysis, visual aids, and predictions to help athletes and coaches in their Olympic journeys. We want to provide support and information that can help them succeed. One cool feature is the interactive map that shows how many medals each country has won over the years. Additionally, our project can make predictions based on an athlete's age, gender, height, and country to give an idea of their chances of winning a medal. This prediction tool is meant to help athletes and coaches set realistic goals.

**3.7 Future Work**  
 The future scope of our project includes enhancing the accuracy and relevance of our Olympic predictions by incorporating data from the upcoming 2024 Olympics. We plan to create a new dataset with the same comprehensive features as our current dataset, leveraging web scraping techniques to extract real-time data from the official Olympics website. This endeavor will ensure that our platform remains up-to-date with the latest information, enabling more precise and timely predictions for our users. In the future, we aim to make our Olympic predictions even better by using data from the upcoming 2024 Olympics. We'll collect this data from the official Olympics website using web scraping, which means extracting information directly from the website. By doing this, we can keep our platform updated with the latest details, making our predictions more accurate and timelier for our users.

**References**

1. Nishant Kulkarni, Pratik Patil, Rugved Pande, Dhiraj Patil, Pranav Nair, Parth Prabhu, Pratyush Doshi, Pranav Bhosale. “Olympic Data Analysis using Data Science”, 2022.
2. Rahul Pradhan, Kartik Agrawal and Anubhav Nag. “Analyzing Evolution of the Olympics by Exploratory Data Analysis Using R”, 2021.
3. Surya Sena Reddy, Suraj Kumar. “Data Analytics on Olympics Datasets”, 2022.
4. Farkande Vaishnavi, Gurav Vaishnavi, Borse Tejas “Web Application of the Olympic data analysis”, 2012.
5. Shivam Bansal, Vaibhav Yadav, Sahil, Yashshavi Prajapti, Rashmi Tiwari. “Web Application of Olympic Data Analysis”, 2023.
6. Dey S K, Rahman M, Siddiqi U R and Howlader “Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach J. Med. Virol”, 2020.
7. Bondu R, Cloutier V, Rosa E and Roy M 2020 An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada) Appl. Geochem.
8. Cutait, M.: Management performance of the Rio 2016 Summer Olympic Games. Research Paper submitted and approved to obtain the Master’s degree in Sports Administration at AISTS in Lausanne, Switzerland.
9. Moreno A, Moragas M and Paningua R 1999 The evolution of volunteers at the Olympic Games Proceedings of Symposium on Volunteers (Lausanne, Switzerland: Global Society and the Olympic Movement)
10. Abeza G, Braunstein-Minkove J R, S´eguin B, O’Reilly N, Kim A and Abdourazakou Y 2020 Ambush marketing via social media: The case of the three most recent Olympic Games Int. J. Sport Communication
11. Yamunathangam D, Kirthicka G and Shahanas P 2018 Performance Analysis in Olympic Games using Exploratory Data Analysis Techniques International Journal of Recent Technology and Engineering (IJRTE)
12. Ramachandran K. M. and Tsokos C P 2020 Mathematical statistics with applications in R (Academic Press)
13. "The Modern Olympic Games" (PDF). The Olympic Museum. Archived from the original (PDF) on 6 September 2008. Retrieved 29 August 2008.
14. Antarlina Sen and Gaurang Margaj, “A prediction model for which country will win the highest number of Gold‟, 2016.
15. Leonardo De Marchi, “Data mining of Sports performance data”, 2011.
16. Huang-Chih Shih,” Survey on content-aware Video Analysis for Sports”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 99, No. 9, January 2017.
17. Chandra Segar Thirumalai and Monica Sankar, “Heuristic Prediction of Olympics using Machine Learning”, International Conference on Electronics, Communication and Aerospace Technology, April 2017.
18. Alexander Rathke and Ulrich Woitek, “Economics and Olympics: An Efficiency Analysis”, January 2007.
19. Leonardo De Marchi, “Data mining of Sports performance data”, 2011.
20. Sources: Athlete Events: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results