***Olympic Data Analysis Using Data Science***

|  |  |
| --- | --- |
| Laxman Pradip Sawant  Computer Science and Engineering  (Data Science)  Vidyavardhini’s College of Engineering  laxman.213008101@vcet.edu.in | Harish Pralhad Sargar  Computer Science and Engineering  (Data Science)  Vidyavardhini’s College of Engineering harish.212998107@vcet.edu.in |
| Niharika Umachal Sahu  Computer Science and Engineering  (Data Science)  Vidyavardhini’s College of Engineering niharika.212988201@vcet.edu.in | Assistant Prof. Komal Champanerkar  Computer Science and Engineering  (Data Science)  Vidyavardhini’s College of Engineering  Champanerkar.komal@gmail.com |

*Abstract* - The Olympic Games, an iconic symbol of global unity and athletic excellence, bring together athletes from 205 nations every four years. Our research focuses on unraveling the layers of Olympic data, using data science to transform it into valuable insights. We have developed predictive algorithms considering factors like an athlete's physical attributes and past performance to estimate winning probabilities in specific events. This predictive element benefits athletes, coaches, and countries in optimizing their medal-winning strategies. Additionally, our research employs data visualization techniques to create an interactive world map, offering a comprehensive geographical perspective on countries' historical medal counts. This map deepens our understanding of the Olympics' global impact. In summary, our research provides a holistic journey through Olympic history, merging historical analysis, predictive modeling, and global visualization. It offers valuable insights for athletes, coaches, and policymakers as they navigate the evolving world of the Olympics and aim for future success in this renowned sporting event.

*Keywords - Exploratory Data Analysis, Prediction, Map, Visualization, Analysis, Olympics, Accuracy.*

# **Introduction**

The Olympic Games stand as a shining beacon of unity and athletic excellence, a global celebration 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.



**Fig. Olympic Games**

# **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.

.

# **Literature Survey**

**1] Olympic Data Analysis using Data Science**

## **Author**: Nishant Kulkarni, Pratik Patil, Rugved Pande, Dhiraj Patil, Pranav Nair, Parth Prabhu, Pratyush Doshi, Pranav Bhosale.

1. **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

## **Drawbacks**:

* 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**:

* 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

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

1. **Author**: Rahul Pradhan, Kartik Agrawal and Anubhav Nag
2. **Year**: 2021

## **Techniques** **Used:**

## The research paper employs Exploratory Data Analysis (EDA) as the primary technique to analyze the large dataset related to the Olympics.

## In this case, EDA is used to explore various factors contributing to the evolution of the Olympic Games over the years

## **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.

## **Future** **Scope**:

* Further Predictive Modelling: The authors can expand their analysis by developing predictive models to forecast future trends in Olympic games.
* Geospatial Analysis: Utilizing geographical data to analyze the geographical distribution of events.

**3] 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.

## **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.

## **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.

**4] Web Application of the Olympic data analysis.**

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

1. **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.

## **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.

## **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.

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

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

1. **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.

## **Drawbacks:**

## Data Collection and Preprocessing: Collected Olympic data from public sources and prepared it for analysis by cleaning and focusing on Summer Olympics data.

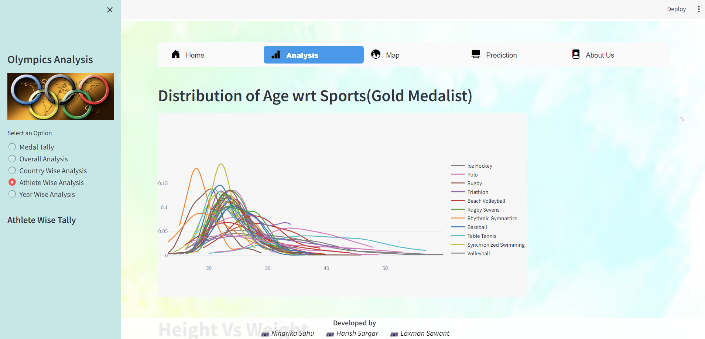
## Exploratory Data Analysis (EDA): Analyzed the data using Python libraries to gain insights into Olympic history, country performance, and athlete factors.

## **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.

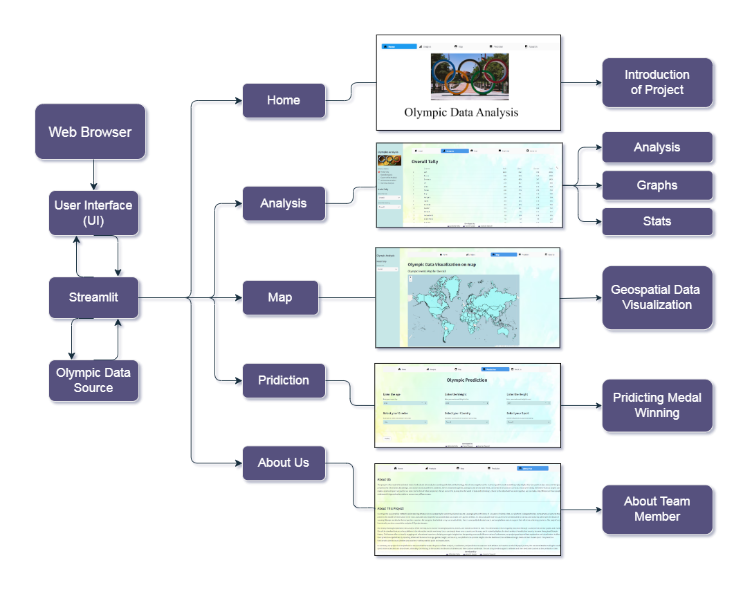
# **Proposed System**

## To address this challenge, we have data analytics techniques and technologies to efficiently process and analyze the Olympic dataset. Our goal is to extract valuable insights, uncover hidden patterns, and present the information in a way that is understandable and accessible to a wide audience. We want to develop a platform that will be visualizing data through graphs, structuring data into informative data frames, and applying statistical techniques to discover meaningful patterns within the data. Further we would develop a prediction model that is accurate which can effectively forecast future Olympic games medalists. We are also interested in developing a geospatial data visualization platform, where data is represented on a map allowing users to interactively explore and analyze stats related to each participating nation.



**Fig. Distribution of Age wrt Sports.**

# **Architecture and Block Diagram**



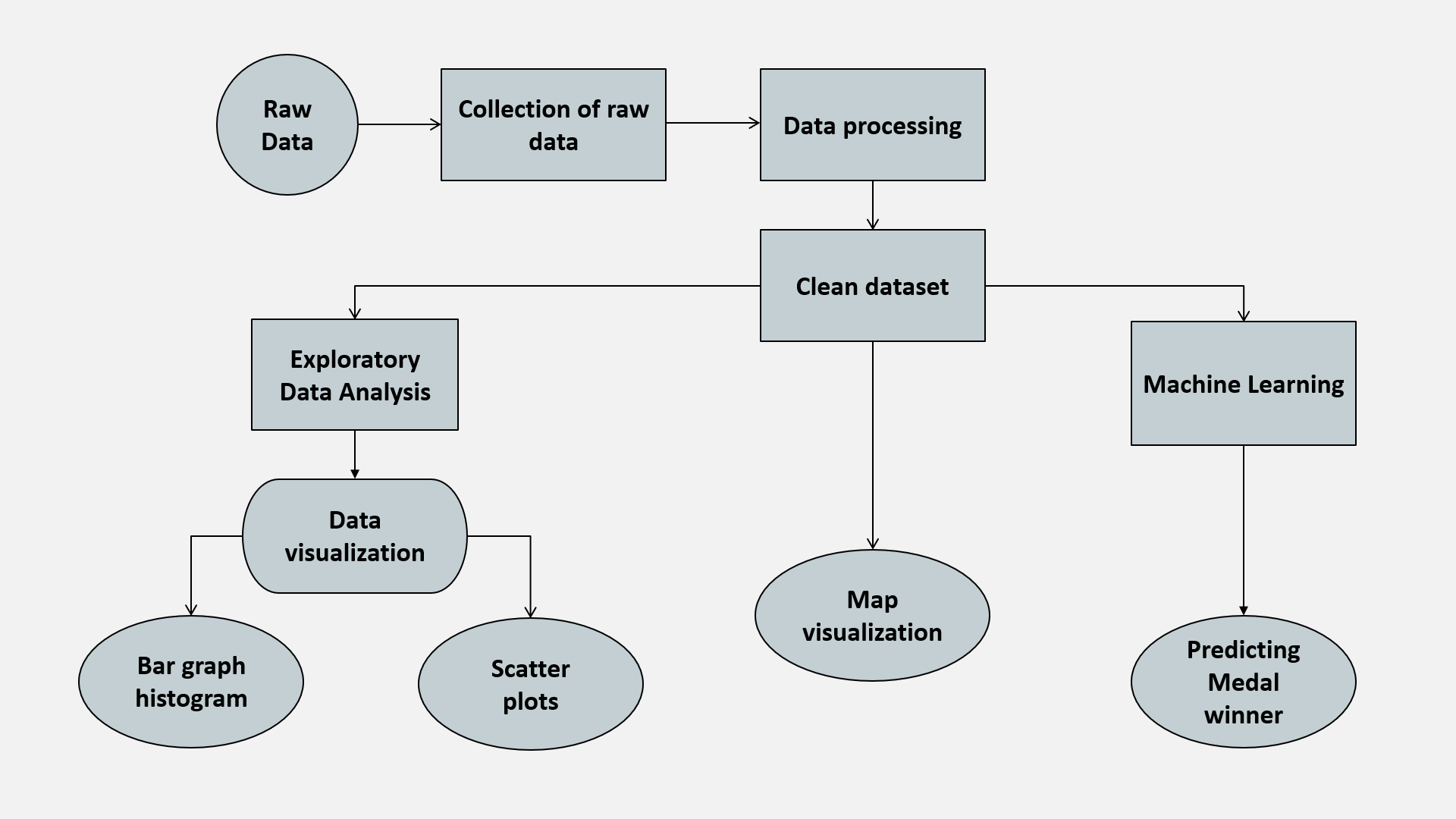
**Fig. Architecture**

**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 Olympic dataset, making it easily 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 beginning 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 analyse 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.



**Fig. Block Diagram**

# **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.

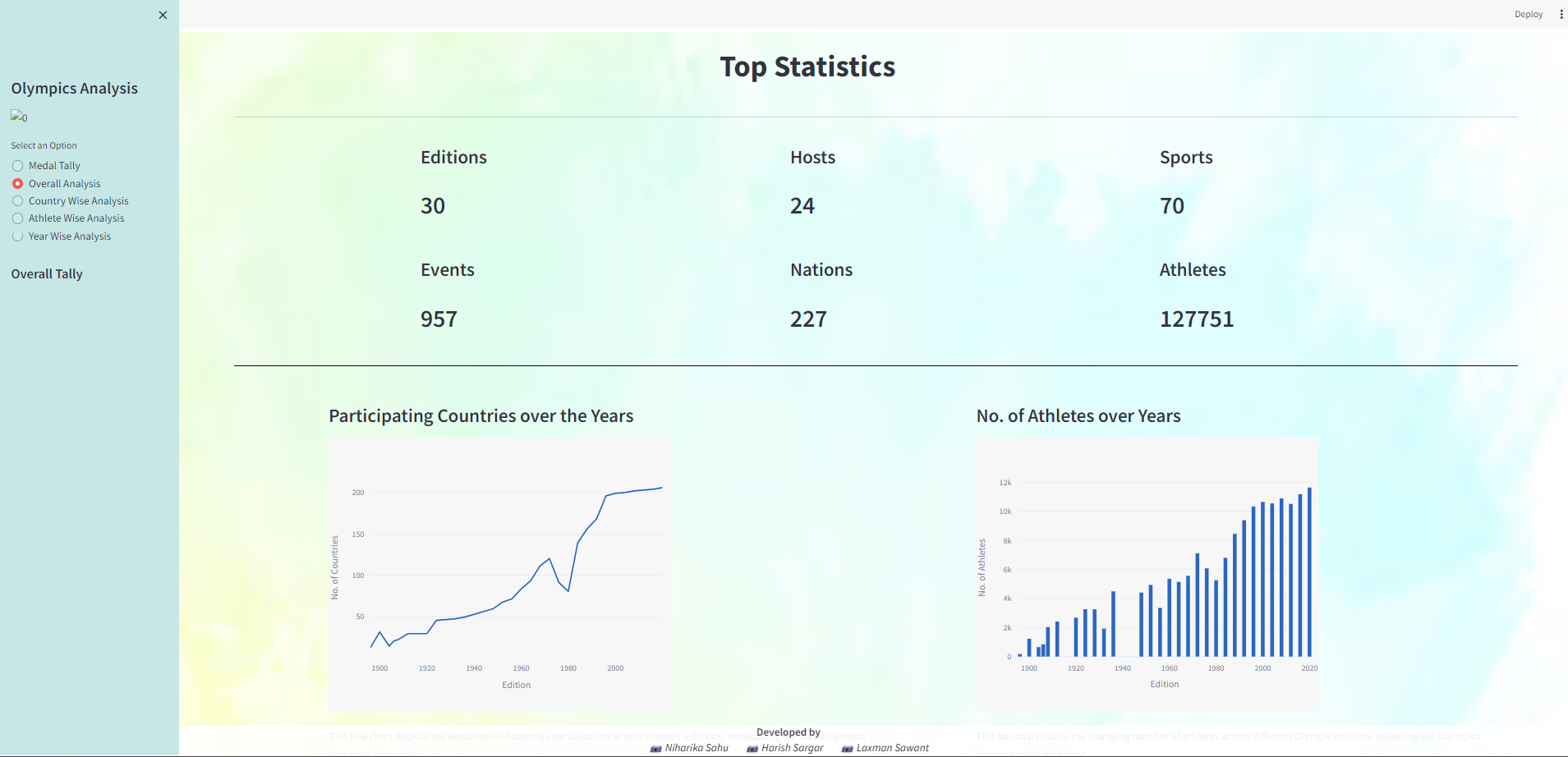
**Process Design - How is Logistic Regression used in predicting Medal Winners?**

* **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.

# **Methodology**

**1] Analysis:**

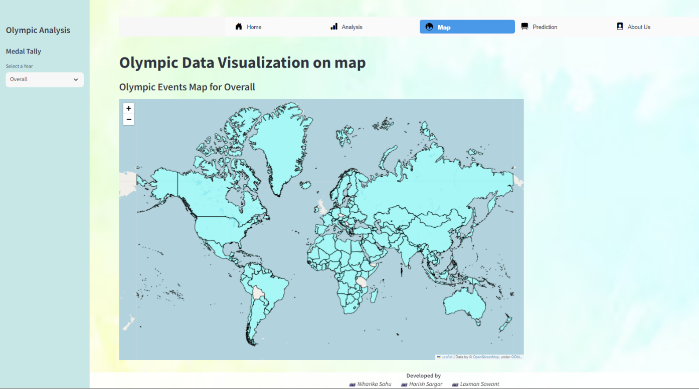
* **Define the Purpose and Scope:** We first clearly stated the objectives of the data analysis report and defined its scope.
* **Data Collection and Preparation:** We gathered and cleaned the data needed for analysis. We also handled missing values, outliers, and inconsistencies to ensure data integrity.
* **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.
* **Data Visualizations:** We presented a variety of visualizations, including charts, graphs, and plots, to illustrate key insights & trends. We used appropriate 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.
* **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.
* **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. Analysis and Visualization.**

**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 pre-processed 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 Centre Determination:** Spain serves as the default centre for map display. This is essential for the initial displaying of the map.
* **Colour Scale Configuration:** We configured colour 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.
* **Streamlit App Development:** 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.

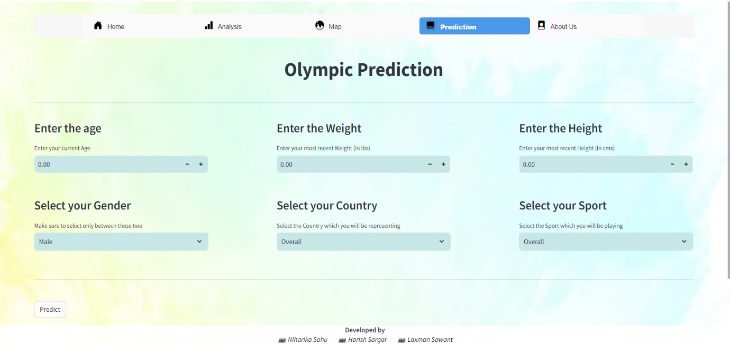


**Fig. Medal Count on World Map**

**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:** We trained 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. Predicting Medal Winners.**

# **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.

# **future scope**

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. Additionally, we are committed to expanding our project's user base by continually improving the user interface and introducing new features to make exploring Olympic data even more engaging and insightful. This web scraping integration will significantly enhance the system's accuracy and relevance while relieving the need for manual data updates, making it a more efficient and up-to-date tool for Olympic data analysis.

# **Acknowledgment**

##### We are deeply thankful to Prof. Komal Champanerkar for her expert guidance, Dr. Vikas Gupta for his continuous support, and Prof. Yogesh Pingle for his valuable guidance. We also appreciate the staff of the Computer Science and Engineering (Data Science) department for providing necessary resources. Mrs. Maya Varghese, our Mini Project coordinator, is acknowledged for her assistance. Lastly, we express profound gratitude to our parents and the almighty god for their unwavering support, without which the project would have been impossible to accomplish.

##### **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 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