**Title: Olympics Performance and Funding Analysis**

Data Mining and Business Intelligence

J Component

ITA5007

Review- I

Team Members:

1. Bharti Sharma - 22MCA0207
2. Ritika Gyanani- 22MCA0279

Faculty:

HARSHITA PATEL

**Olympics Performance and Funding Analysis**

Bharti Sharma 22MCA0207

Ritika Gyanani 22MCA0279

Faculty: Harshita Patel

School of Information Technology and Engineering

Vellore Institute of Technology

Vellore - 632014, Tamil Nadu, India

Abstract:

The Olympic Performance Profiler is a web app developed using Streamlit to explore, visualize, and analyze Olympic Games data. The app provides an interactive interface to perform various data analysis techniques, such as clustering analysis, association rule mining, and data preprocessing methods, on the Olympic Games dataset. The dataset contains information about the Olympic Games from 1896 to 2016, including athletes' information, medal tally, country-wise and athlete-wise analysis, and overall analysis.

The app's user interface includes different sections, such as data preprocessing, data analysis, and data visualization. In the data preprocessing section, users can handle missing values, outliers, and normalization in the dataset to improve data quality. The data analysis section provides clustering analysis and association rule mining techniques to discover patterns, similarities, and relationships between data attributes.

Regression analysis is a statistical technique used to investigate the relationship between a dependent variable (response) and one or more independent variables (predictors). In the case of the Olympic games dataset you provided, regression analysis could be used to explore the relationship between the cost of hosting the games (dependent variable) and various independent variables such as the year, country, type, number of events, and number of athletes.

Linear regression could be used to build a model to predict the cost of hosting the Olympics based on these independent variables. Additionally, other regression techniques such as polynomial regression, ridge regression, and lasso regression could be explored to improve the model's accuracy and reduce overfitting.

Data Collection:

The Kaggle 120 Years of Olympic History dataset is a comprehensive and detailed collection of data related to the history of the Olympic Games. The dataset includes information on every athlete, event, and medal awarded at the modern Olympic Games from Athens 1896 to Rio 2016. This dataset has been compiled by Kaggle, a subsidiary of Google LLC, and is available for public use.

The dataset contains two files one named athletes and event containing over 2,71,117 rows of data, including detailed information on each athlete's name, nationality, age, gender, and sport and the second file named NOC and regions containing around 230 rows having the data of various countries with their NOC.

The data is provided in a structured format that is easy to work with and analyze, making it an ideal resource for our data analysts in exploring trends and patterns in Olympic history.

Functional Requirements:

1. **Medal Tally Module:**

a. Display the total medal count of each country for a given Olympic event.

b. Display the gold, silver, and bronze medal count of each country for a given Olympic event.

c. Sort the medal tally based on the total medal count, gold medal count, silver medal count, or bronze medal count.

1. **Country Wise Analysis Module:**

a. Display the performance of a particular country over the years in the Olympics.

b. Display the top performing countries for a particular Olympic event.

c. Compare the performance of two or more countries for a given Olympic event.

1. **Athlete Wise Analysis Module:**

a. Display the performance of a particular athlete over the years in the Olympics.

b. Display the top performing athletes for a particular Olympic event.

c. Compare the performance of two or more athletes for a given Olympic event.

1. **Overall Analysis Module:**

a. Display the overall performance of all countries in the Olympics over the years.

b. Display the overall performance of all athletes in the Olympics over the years.

c. Perform data mining techniques to identify trends, patterns, and insights from the Olympic data.

Each of the above modules should have the following functionalities:

1. Accept user input to select the Olympic event, country, or athlete to perform the analysis.
2. Display the results of the analysis in a user-friendly format, such as tables or graphs.

# **Need for this Model:**

Clustering algorithms can group similar athletes based on their performance, personal brand, and other factors. This can help sponsors identify athletes who may be a good fit for their brand based on their shared characteristics.

Predicting the funding price of Olympics using regression analysis can be helpful in several ways. It can help in estimating the cost of organizing an Olympic event, which is important for budget planning and management. It can also provide insights into the factors that affect the cost of organizing an Olympics event, which can help in making decisions related to event management and planning. Moreover, it can help in predicting the funding required for future Olympic events, which can be useful for securing sponsorships and other sources of funding.

LITERATURE REVIEW

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Reference** | **Title** | **Techniques used** | **Dataset used** | **Conclusion** |
| [1.](https://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/download/1857/465) | Exploratory Data Analysis (EDA): A Study of Olympic Medal lists | The forms of analysis are presented in Univariate, Bivariate and Multivariate with  Data collection, Data Understanding, Data Pre-processing, Exploratory Data Analysis (EDA). | The data that is being used are data regarding the Olympic championship from 1896 – 2014 obtained from the public dataset. | The result of the univariate analysis is that most of the medalists from 1896-2014 is Male athletes. From bivariate analysis, there is no relationship between Year and Medal variables in the Olympics games. In other words, they are not independent of each other. |
| [2.](https://www.researchgate.net/publication/330847008_Performance_analysis_in_olympic_games_using_exploratory_data_analysis_techniques) | A longitudinal study on the Olympic Games | Panel data Analysis and Event Study | Official website of Olympic Movement | This study summarizes the effect of Olympics on tourism industry. |
| 3. | Mining Olympic Athletes Performance Using Data Mining Techniques | Association Rule Mining, Decision Tree, Naive Bayes, Normalization, Attribute Selection | Athlete Performance Dataset | Naive Bayes was the best classifier in predicting athlete success rate, the top 5 athletes were identified for each sport |
| 4. | Comparing the Well-Being of Para and Olympic Sport Athletes: A Systematic Review | Meta-Analyses, Quality Assessment | Eight electronic databases (MEDLINE, PsycINFO, Embase, AMED, CINAHL, SPORTDiscus, Scopus, and Web of Science) were searched from inception to September 2014. | Helped professionals and coaches aiming to facilitate the well-being needs of athletes. |
| 5. | Data Mining of Olympic Games: Prediction and Evaluation | Naive Bayes, K-NN, Decision Tree, Data Cleaning, Attribute Selection | Olympics Results and Demographic Data of Athletes | Naive Bayes outperformed other classifiers in predicting medal winners, age and past performance are important predictors |
| 6. | A Count Data Analysis Of National Olympic Performance | Poisson and negative binomial count data regression models and attaches statistical significance | Data from most recent 2004 Summer Games in Athens |  |
| 7. | A Data Mining-Based Method for Analyzing Olympic Games | Clustering, Decision Tree, SVM, Data Cleaning, Attribute Selection | Olympic Games Data | Decision Tree outperformed other classifiers in predicting medal winners, age and past performance are important predictors |
| 8. | Analysis of Olympic Heptathlon Data | Compact clustering, average Clustering, DySect | 1992 Women’s heptathlon data | Classifies the most critical events for g the better class of heptathlete. |
| [9](https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012058). | Analyzing the Evolution of the Olympics by Exploratory Data Analysis using R | Data collection, Data Pre-processing, Exploratory Data Analysis (EDA) | Olympic Games Data | It is true that the Olympic Games have evolved considerably over time from the 1896 Olympic Games till the 2016 Rio Olympics. |
| [10.](http://cs229.stanford.edu/proj2021spr/report2/81985704.pdf) | 2020 Summer Olympics Predictions using Machine Learning. | Linear Regression, Regularized Linear Regression, Poisson Regression, Random Forest Regressor, Logistic Regression, Random Forest Classifier, Gaussian Naive Bayes, Multilayer Perceptron, Support Vector Classification, and Support Vector Regression. | The dataset is collected from Kaggle containing Olympic medal results on an individual athlete basis. | Application of imputation technique may help in improving performance. |
| [11.](http://www.warse.org/IJETER/static/pdf/file/ijeter57852020.pdf) | Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms | Data Collection, Data Cleaning and Preprocessing, Missing Values, Normalization,  Model Selection, Decision Tree, KNN Algorithm, Linear Regression, Random Forest Algorithm, Bayesian Ridge Algorithm. | The dataset is collected from various sources on the web. The dataset contains 227 rows and 10 attributes. The attributes include countries, population, GDP, education expenditure, literacy rate, health expenditure, the prevalence of undernourishment, Gini index, number of games participated, and medals won. | The number of medals won is strongly positively correlated with the population and literacy rate; the Prevalence of undernourishment, GDP and Gini index is positively correlated with medals with some outliers; Education expenditure is in a weak positive correlation on the other hand health expenditure is not correlated; Health expenditure is in positive correlation with the prevalence of undernourishment. |
| [12.](https://www.ijraset.com/research-paper/olympic-data-analysis-using-data-science) | Olympic Data Analysis using Data Science | Data collection, Data Pre-processing, Exploratory Data Analysis (EDA) | 1st dataset consists of information about the players and their entire details like their Gender, Height, Weight, Country for which they play, Medals won (Gold, Silver and Bronze) and many more, 2nd dataset consists of the information of the countries which have participated in the Olympics , 3rd dataset consists of the list of countries along with their country code | Olympic Games have evolved considerably over time. There is an increase in the participation of females in both the Summer and Winter Olympics over time, the Total number of medals won by various participating countries over the years, the Average height and the weight of Players who contributes to the victory of the Games in the event |

Preprocessing Methods:

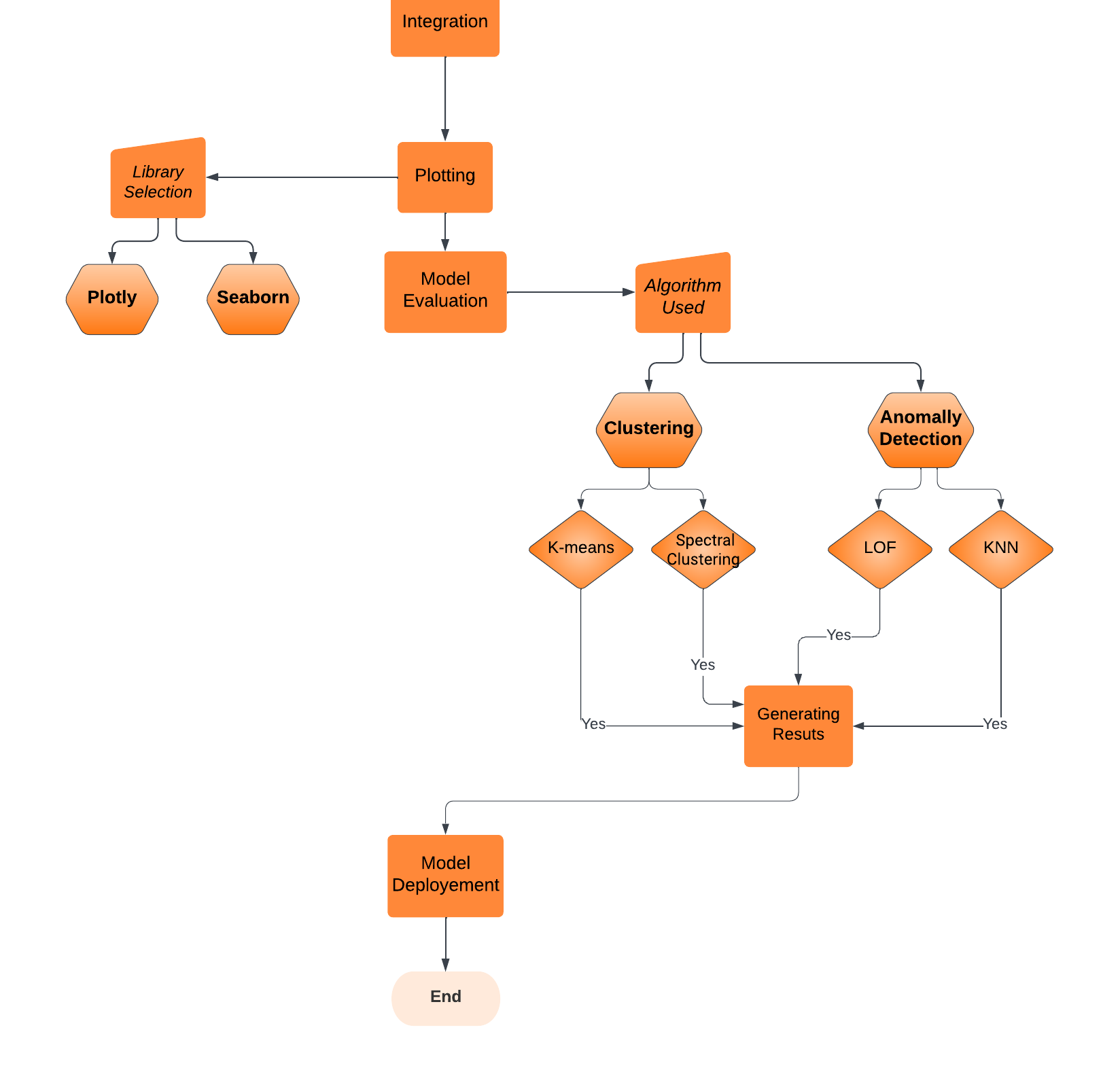
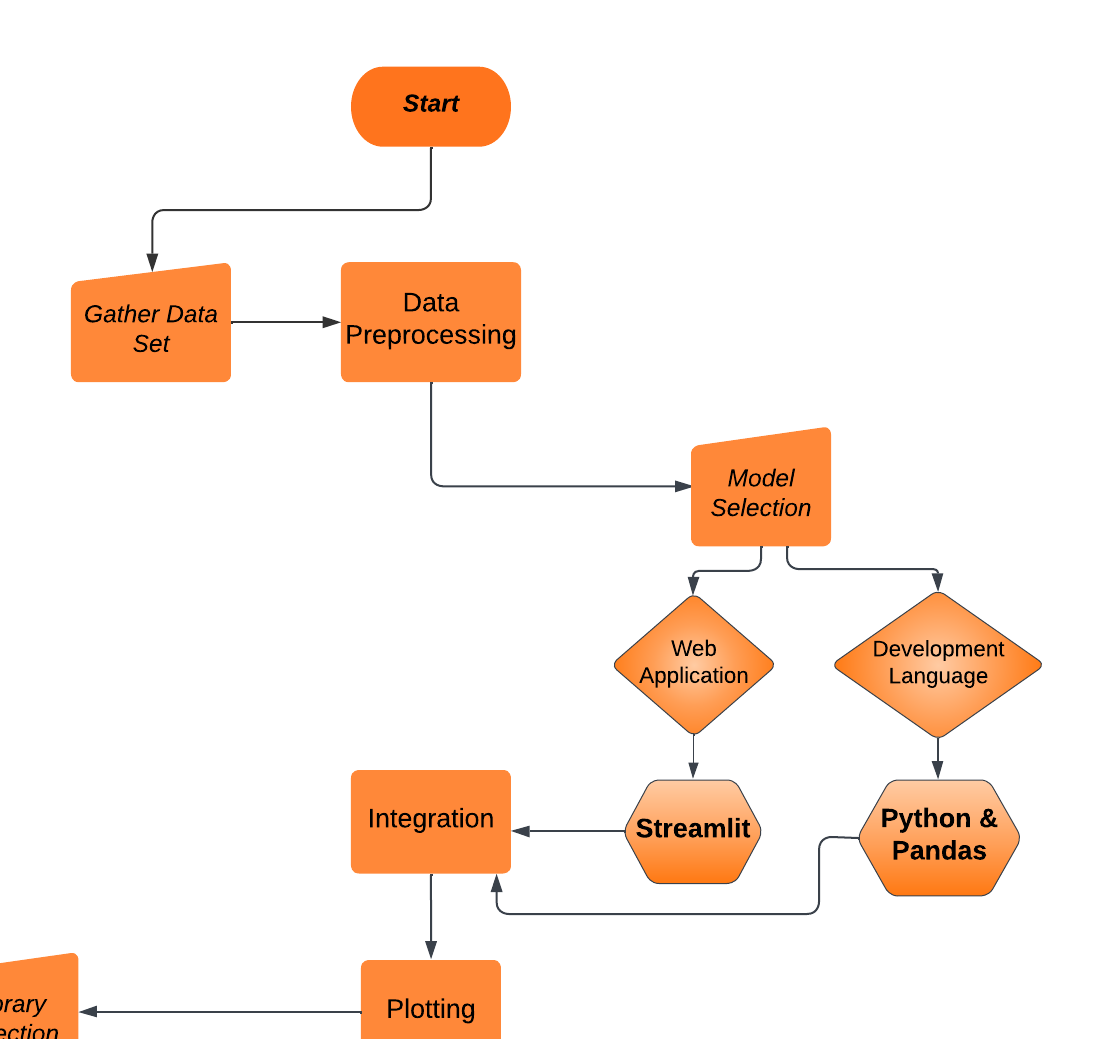
1. Handling missing values: The dataset contains NA values in height and weight. These missing values can be handled by imputing the mean or median of the respective column.
2. Filtering out unwanted data: Since the analysis is only focused on Summer Olympics data, the Winter Olympics data can be filtered out and removed from the dataset.
3. Feature Selection: Not all features are relevant for clustering or classification tasks. Feature selection can help to identify the most important features for the analysis and remove irrelevant features.
4. Outlier Detection: We will use statistical methods to detect outliers in our dataset and handle them accordingly.

Proposed Methodology:

**Algorithms Used:**

1. **Clustering Analysis:** Clustering analysis of the Olympic dataset based on medal count can help identify groups of countries with similar performance patterns. This can provide insights into the factors that drive medal success, as well as inform training and development strategies.
2. **Association Rule Mining:** Association rule mining can be used on your Olympic dataset to discover interesting relationships between different attributes such as sport, event, country, and medal. This can provide insights into which combinations are most likely to occur together.

Flow Chart of Proposed System:



Performance Measures

1. **Data Cleaning:** This performance measure refers to the accuracy and completeness of the data that has been cleaned, preprocessed and formatted. It measures how well the data cleaning process has been performed.
2. **Data Visualization:** This performance measure refers to the quality of the visual representations of the data. It measures how well the data is being presented and how clear and informative the visualizations are.
3. **Data Modeling:** This performance measure refers to the quality of the models created for the data analysis. It measures how well the models fit the data and how accurate the predictions or insights are.
4. **Time and Resource Usage:** This performance measure includes the time and resources required to complete the project. It measures how efficiently the project has been executed in terms of time and resources.
5. **Interpretation of Results**: This performance measure refers to the ability to effectively communicate the results and insights derived from the data analysis. It measures how well the results have been interpreted and communicated to stakeholders.