International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-09, Issue-01, Apr 2023

Olympics Data Analysis Web App Lication and Prediction Using Machine Learning

[1]Prof.Satish J.Manje, [2]Ms.Shruti Rakesh Dubey, [3]Ms.Sejal Sunil Parche, [4]Ms.Disha Laxman Bondre

[1]Asst.Prof.,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]satishmanje93@gmail.com, [2]dubeyshruti71@gmail.com, [3]sejalparche412@gmail.com, [4]dishabondre971@gmail.com,

***Abstract - The Olympics is a major international sporting events which has two aspects of competition namely summer and winter in which thousands of competitors from all over the globe compete in a variety of events*. Olympics are more than just a four-stroke multi-sport world championship. It is a lens which gives us an understanding of global history, including shifting geopolitical power dynamics, women's empowerment, and the evolution of society's values. The Olympic Games had been come to be regarded as world's leading sports competition, with more than 200 nations participating in it respectively. The total number of events in the Olympics is 46 but 2020 33 events took place. And there are winners in every event. Accordingly, various data are generated. The main goal is to shed light on major patterns in Olympic history. The NOC depends on where most athletes participate, the medal-winning probability, and the characteristics of the athletes (e.g., gender and physical size). Modules that will be implemented are Pandas for analyzing the data, NumPy for array processing, Matplotlib for mathematical extension, the Seaborn and Plotly libraries for plotting, and Random Forest regression for prediction.[1]**

Keywords- data analysis, prediction, machine learning, Random Forest.

I. INTRODUCTION

The process of extracting solutions to problems through data interrogation and interpretation is known as data analysis. Identifying problems, determining the availability of appropriate data, determining which method can aid in solving the interesting problem, and communicating the outcome are all part of the analysis process. For performing analysis, first the dataset should be divided into steps such as organizing, assembling, cleaning, re-analyzing, then applying models and algorithms, and evaluating the final results [1]. The Olympics are more than just a four-year multi-sport world championship. Visualization of data across multiple factors will provide us with a statistical view of the various factors that contribute to the evolution of the Olympic Games and improvements in the performance of various countries/players overtime. Various scenarios come to our mind when we look into the Evolution of the Olympic Games over the years [3]. The various scenarios mentioned above includes an increase in the number of participating nations, then an increase in the number of participating athletes, an increase or decrease in the number of events, an increase in the event's expenditure cost, is an improvement in the performance of a specific country, an improvement in the performance of a specific player, an increase in female participation, and a participation ratio of men to womenn in medication facilities during competition [1], [3].Each sport's performance can be compared to that of others. The sports that require more participation can be identified, and necessary action can be taken by players and nations to improve their future contributions to the Olympics. Using past sports performance data, one's future performance can be predicted. Their performance can also be improved based on prediction [2], [8].

II. AIMS AND OBJECTIVE

a) Aim

The goal is to the cleaning, transforming, and modelling processes to discover practical information that will aid in business decisions by providing beneficial and accurate insights and predictions based on athlete data [1],[6].

b) Objective

- To analyse all the factors that plays vital role in the evolution of Olympics game over the years.
- To visualize and explain change in trends of the various factors over the years which will help to predict the information of future Olympic games.

 DOI : 10.5281/zenodo.7863837

- Making Predictions after analyzing the data and getting the useful insights which will help the player know which country has more chance of winning the Olympics this year [2], [5].

III. LITERATURE SURVEY

Paper 1: Research on the Development Trend of Vault in Women's Competitive Gymnastics.

This paper has presented a Data Science approach in response to the national fitness plan and with the advent to of the digital expansion decisions in a timely manner and incorporating analytical results in terms of wave and information age, health big data and sports. This paper tries to adopt method of literature study, video observation, mathematical statistics and other research methods, to the Tokyo Olympic Games women's vaulting horse eight before the final technical action, the results of the competition, completion, action type as a research object [3].

Paper 2: Olympics Performance Evaluation and Competition Strategy based on Data Envelopment Analysis.

With increased focus on the Olympics, sports industry management has grown in importance in service industry management and industrial engineering. This system describes the index system for evaluating the efficiency of Olympic participants, presents a DEA model based on the best and worst practise frontiers for evaluating Olympic contestants' relative performance and examines competition strategies with the use of a competition DEA model [8].

Paper 3: Web Outlier Prediction Using Random Forest Classifier.

Predicting outliers using Random Forest classifier is an ensemble learning method. An outlier is an uncertain and an inconsistent data point. Random forest classifier works by building numerous decision trees with training data and uses the majority of the class as output. The original datasets are divided into training and test sets in this model. Samples are drawn at random from the training data with replacement. These sample subsets are known as bootstrap samples [2].

IV. EXISTING SYSTEM

The current system is made up of datasets spanning more than 120 years of Olympic records for country-specific and individual sports events, which are being analyzed using data analysis and visualization, data cleaning, and data modelling processes. The current system includes methods such as plots, histograms, heat maps, charts, and so on for depicting Medal tally, Country-wise analysis, Overall Analysis, Athlete-wise analysis [1],[7].

V. COMPARATIVE STUDY

Sr. No	Paper Name	Author/ Publication	Technology	Advantage	Disadvantage
1.	Research on the Development Trend of Vault in Women's Competitive Gymnastics.	Jiao Jiao Dai, Jin Lin Liu.	Python, Data Analysis.	A multi-method approach, combining literature study, video observation, mathematical statistics, and other research methods to analyze the Olympic Games.	Lacks a clear research question or hypothesis, which may make it difficult to understand the overall purpose or significance of the study.
2	Olympics Performance Evaluation and Competition Strategy based on Data Envelopment Analysis.	Yang, F., Ling, L., Gou, Q., & Wu, H.	Python, Data Analysis.	Provides insights into the application of DEA in the sports industry and Olympic management.	Does not provide detailed information on the data used for the analysis, which makes it difficult to assess.
3.	Web Outlier Prediction Using Random Forest Classifier.	Mohandoss, D. P., Shi, Y., & Suo, K.	Python, Machine Learning.	The use of Random Forest classifier is a popular and effective method for outlier detection.	Does not discuss the potential limitations or challenges of using Random Forest classifier in outlier detection.

Table 1. Comparative Analysis of the System

VI. PROBLEM STATEMENT

The system analyses the Olympic data set and create an Olympic Data Analysis Web Application for analyzing data over time, which will assist athletes in broadening their horizons for winning a medal and be useful for future predictions of medals won by individual athletes, nations, and counties throughout various sports. This paper shows the

 DOI : 10.5281/zenodo.7863837

improvement in various design and implementation issues. Due to advancement in various python libraries visualization can be performed with accurate results. With the use of various advance tools, the computational complexity is reduced and the overall interface has been made user-friendly. With the introduction of the tool streamlit graphical interface deployment is enhanced [1], [3], [5].

VII. PROPOSED SYSTEM

This system shows Olympics Data Analysis Web Application using stream-lit and prediction using machine learning. Streamlit helps us create web apps for data science and machine learning. It is mainly compatible with most important Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc [1]. Then we perform Prediction on the analyzed data where our system will predict which country will win the medal in the current year using the Random Forest Algorithm as our is a type of classification problem [2]. Using different Python like -Numpy, Pandas, Matplotlib, Seaborn we will perform our Data Analysis and for deployment we will use Heroku. The system is going to be trained on huge Olympic data set and after cleaning and modeling the data set prediction process takes place where user enter some inputs in the input field to see the prediction output [2],[5].

VIII. ALGORITHM

The Algorithm for Predicting Olympic Medal Winner

Step.1: Start

Step.2: Importing the required libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

Step.3:

Importing the dataset Reading the csv file and putting it into 'df' object

df=pd.read_csv('athlete_events.csv')

Step.4:

Putting Feature Variable to X and Target variable to y

dataset = dataset.drop(['ID', 'Name', 'Team', 'Games', 'Year', 'City', 'Sport', 'Event'], axis = 1)

Step.5:

Train-Test-Split is performed

now splitting of the data into train and test is done

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.25,random_state=8)

Step.6:

Fitting of the Model

classifier.fit(X_train, y_train)

Step.7:

Making the Prediction

y_predictions=random_forest.predict(X_test

Step.8:

Evaluating the results

print("Accuracy:", (confusion_matrix[1,1]+confusion_matrix[0,0])/confusion_matrix.sum())

Step 9: exit

IX. MATHEMATICAL MODEL

The output will be closer to 1 then the model is more accurate.We can represent the dataset as $X = \{(x1, y1), (x2, y2), ..., (xN, yN)\}$, where $xi \in R^d$ is the ith observation and yi is its corresponding label.For each tree $t = 1, 2, ..., T$, we can denote the training dataset used to train the tree as Xt. The tree is trained to minimize the following objective function:

$$f_t(x) = argmax_k \sum_{(xi,yi) \in Xt : yi=k} I(yi=k)$$

$$\hat{y}(x) = argmax_k (1/T) \sum_{t=1}^{T} f_t(x)$$

where $\hat{y}(x)$ is the predicted label for observation x, and k is the set of possible labels. The entropy-based criterion is defined as:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} [(|D_j| / |D_p|) * I(D_j)] [7].$$

X. SYSTEM ARCHITECTURE

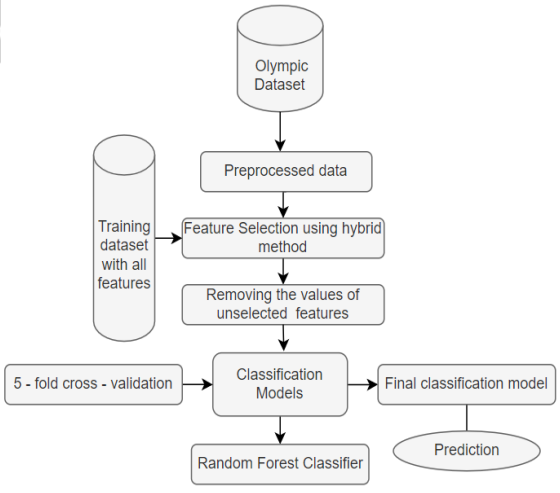

Fig.1: System Architecture

Description:

The system architecture for Olympic Medal Prediction using the Random Forest algorithm consists of a web application

 DOI : 10.5281/zenodo.7863837

where users can input various features related to Olympic events such as athlete's age, country, sport, previous Olympic experience, etc. This data is preprocessed and fed into the Random Forest algorithm, which is used to train a predictive model. The model is then used to make predictions about the number of medals that a given country is likely to win in the upcoming Olympic games. The predicted medal counts are then displayed to the user through the web application. The architecture also includes a database for storing user inputs and predicted results.

XI. ADVANTAGES

- Helps in analyzing over 120+ years of Olympics Historical Data.
- Calculates Overall Analysis of Olympic Games
- Predicts the future winning athlete which helps other country's athletes to improve their performance [2].
- By selecting any country all the country's history of winning the Olympics would display with medals (Gold, Silver, Bronze).

XII. DESIGN DETAILS

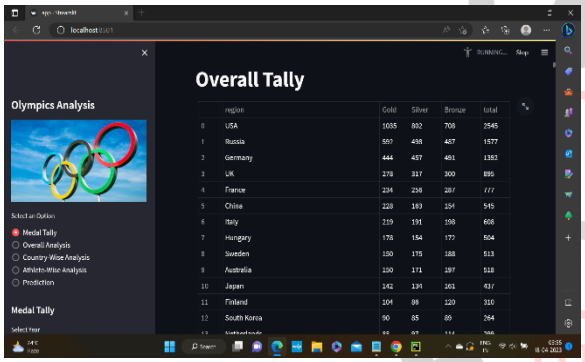

Fig.2: Medal Tally

Medal tally analysis is performed to study the distribution of medals across different countries. This analysis helps in identifying the countries that are performing well in different sports and events.

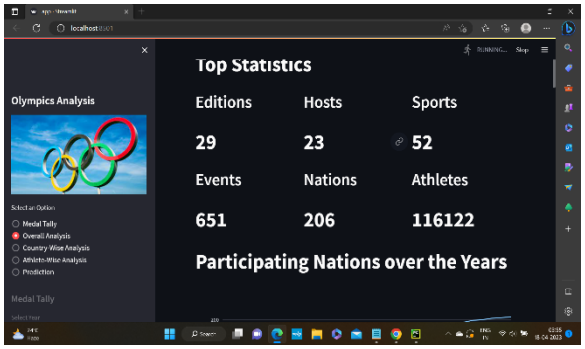

Fig 3: Overall Analysis

Overall analysis for Olympic analysis involves collecting and preprocessing a large dataset of historical Olympic games data and helps athletes, countries, and clubs to gain insights into the major sports world championships in Olympics, plan for future games, and make data-driven decisions.

Fig 4: Country Wise Analysis

Country-wise analysis in Olympic analysis involves analyzing the performance of different countries in the Olympics based on various parameters such as the number of medals won, the type of medals won, the number of athletes representing the country, the performance trend over the years, etc.This analysis is helpful for countries to identify the areas they need to focus on to improve their performance.

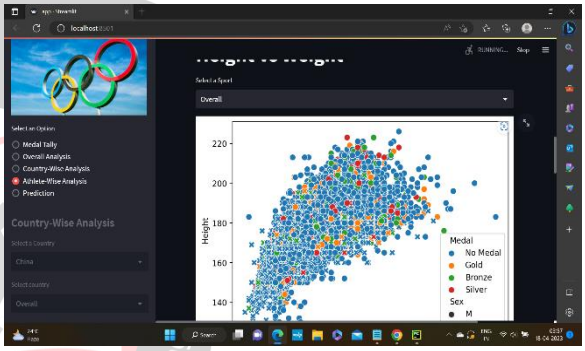

Fig 5: Athlete Wise Analysis

Athlete analysis involves analyzing the performance of individual athletes in various sports events in the Olympics. This analysis can provide insights into the strengths and weaknesses of individual athletes, as well as the factors that contribute to their success or failure.

XIII. CONCLUSION

Thus, we have tried to implement the paper "Jialu Wen, Xin Wang.","Study of the Visualization and Interaction of Data.", IEEE to analyze and visualize Olympic data using the random forest algorithm for prediction. The unique contributions of the paper include the development of a visual analytics tool for analyzing and predicting sports performance, as well as the use of the random forest algorithm for prediction. The theoretical and managerial implications of this project are significant, as the insights gained from the analysis and visualization of Olympic data can benefit athletes, countries, and sports clubs by informing decisions on training, participation, and investment. However, the limitations of the research include the use of a single algorithm for prediction and the potential for bias in the data. Future research directions could include the use of

 DOI : 10.5281/zenodo.7863837

additional algorithms for prediction, as well as the collection of more comprehensive and diverse data sets to reduce bias and improve accuracy. Overall, this project demonstrates the potential of visual analytics and machine learning in the field of sports data analysis.

REFERENCE

[1] J. Wen and X. Wang, "Study of the visualization and Interaction of data: Take the Historical Data of the Winter Olympics as an Example," 2020 International Conference on Innovation Design and Digital Technology (ICIDDT), Zhenjing, China, 2020, pp. 78-82, doi: 10.1109/ICIDDT52279.2020.00022.

[2] Mohandoss, D. P., Shi, Y., & Suo, K. (2021). Outlier Prediction Using Random Forest Classifier. 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC).

[3] J. J. Dai and J. L. Liu, "Research on the Development Trend of Vault in Women's Competitive Gymnastics: — Based on the Tokyo Olympics Games," 2021 International Conference on Health Big Data and Smart Sports (HBDSS), Guilin, China, 2021, pp. 248-251, doi: 10.1109/HBDSS54392.2021.00055.

[4] P. Rajesh, Bharadwaj, M. Alam and M. Tahernezhadi, "A Data Science Approach to Football Team Player Selection," 2020 IEEE International Conference on Electro Information Technology (EIT), Chicago, IL, USA, 2020, pp. 175-183, doi: 10.1109/EIT48999.2020.9208331.

[5] Rory P. Banker, Fadi Thabtah, "A machine learning framework for Sport result prediction", Applied Computing and Informatics (Volume 15, Issue 1, January 2019).

[6] Harsha Vardhan Goud, P. S., Mohana Roopa, Y., & Padmaja, B. (2019). Player Performance Analysis in Sports: with Fusion of Machine Learning and Wearable Technology. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). doi:10.1109/iccmc.2019.8819815

[7] Vishal R.Shinde, "Record linkage and data prediction" in IJREAM, ISSN: 2454-9150 Volume 04 Special Issue-iCreate-2019, APR 2019.

[8] F. Yang, L. Ling, Q. Gou and H. Wu, "Olympics Performance Evaluation and Competition Strategy Based on Data Envelopment Analysis," 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 2009, pp. 1-4, doi: 10.1109/CISE.2009.5367069

 DOI : 10.5281/zenodo.7863837