International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Player Winning Probability Model (PWPM): A Review

Shashank Gaur[1], Naveen Kumar Pandey[2], Archna Pandey[3], Mohammad Hasrat[4], Anuj Gupta[5]

[1]Assistant Professor, Computer Science and Engineering Department, Bansal Institute of Engineering and Technology, Lucknow
[2,3,4,5]Student, Computer Science and Engineering Department, Bansal Institute of Engineering and Technology, Lucknow

ABSTRACT
The Olympic Games are an international event and a source of pride for all countries. With over 200 countries participating in the Olympic Games, it is an international sporting event. The Olympic Games stand as a pinnacle of athletic achievement and global unity, showcasing the finest talents from diverse nations across numerous sporting disciplines. In this review we used machine Learning Algorithms to do prediction on Olympic medal and applying various tools and techniques for Data Analysis using historical data. The process involves Data Collection, Data Cleaning, Data Processing, Exploratory Data Analysis and Medal Prediction. The main objective of this work is to study the complete set of Olympic data to discover pattern and relationship between variables using analysis of data to assess how the Olympic Games changed over time and to do prediction on medals. In this work a PWP model has been proposed in which the Random Forest classifier algorithm has been selected for the classification of player winning probability with accuracy of 70.223325%.

Keyword: Player Winning Probability Model (PWPM), Random forest Classification, Adaboost Classification, Overfitting, Bagging Classifier.

INTRODUCTION
The Olympics, starting in 1896 and held every four years, are regarded as the premier global event where athletes from various nations showcase their skills. Originating from ancient Greece, sports have evolved significantly through the ages [1]. Recently, there has been an increased focus on using data and analytics to highlight the performance of athletes and countries in the Olympics. The purpose of this study is to analyze Olympic data from the last 120 years and identify patterns and trends. We collect data about the Medals, the performance of individual Athletes and their (Sex, Height, Weight), NOC, Region, Season etc. The Exploratory Data Analysis technique is employed to examine data and determine the number of reported cases (positive, deceased, discharged) both inside and outside China. [15] This paper utilizes data from various datasets and employs the Exploratory Data Analysis (EDA) technique to examine factors such as the number of recovered cases in January and February both inside and outside China, and the number of confirmed cases across different provinces in China and abroad up to February 16, 2020 [15]. We analyzed these data using data analysis methods and perform prediction on the data using Machine

Learning algorithms. Our research provides an in-depth analysis of Olympic data, illuminating the evolution of the original sports. When considering the changes in the Olympic Games over time, several scenarios emerge. These include an increase in the number of participating countries, more athletes competing, a greater variety of events, improved performances from certain countries, enhancements in individual athlete performances, changes in women's participation rates, shifts in male-to-female participant ratios, and predictions of medal outcomes versus actual results. Our analysis explores these aspects through various lenses, including medal tallies by country, a comprehensive overview of the Olympics, country-specific performance trends, individual athlete analyses, and medal predictions. This study aids in understanding which countries and athletes have been most successful in the Olympics.

For this Project, the datasets is collected from the Kaggle [7]. To offer a thorough analysis, we focused on various visualization, analysis, and machine learning techniques to present the data clearly and predict trends such as medal counts and other developments in the Olympics.. Key components of our analysis include a detailed review of medal tallies by country, providing a longitudinal perspective on the countries' performances across different Olympic editions. This is complemented by a thorough analysis of individual athletes over the years, highlighting standout performers and their contributions to their respective national tallies. Predictive analytics plays a crucial role here, enabling stakeholders to forecast future trends and prepare more effectively for upcoming competitions. By employing a range of visualization techniques and analytical tools, this research presents the data in an accessible format, making it possible not only to understand past and present trends but also to predict future developments in the Olympics. Before performing data analysis, the datasets are converted into Data Frames and undergo extensive data cleaning to remove any null values. Following this, data analysis is conducted using Python and libraries such as NumPy, Pandas, Sklearn, Plotly, Streamlit, and Matplotlib. Visualizations, including bar charts and pie charts, are used to display the results [20].

LITERATURE SURVEY

The review and analysis of many research papers has enabled us to learn many new methods and procedures. [13] We explored heuristics, we utilized machine learning algorithms to estimate the number of Olympic medals won by athletes, and we have also learned that a country's level of success can be gauged through effective research and the significance of sports in society. A country's Olympic participation can be predicted on the basis of its historical achievements. The probability of winning a medal at the next Olympics is calculated based on whether an athlete has won a medal this year [13]. If they need improvement in a particular area, they can do so and participating in an appropriate learning program will have a significant impact on their results. The Olympic Games have been extensively analyzed through various methods, including statistical visualization, performance analysis of athletes, performance improvement across different countries, and more. The technique, specifically Exploratory Data Analysis, has been utilized to analyze the data. [17] This paper meticulously analyzed the Olympic Dataset to compare the overall performance of participating countries and each country's contribution to the Olympic Games. [17] The primary goal of this analysis was to track the performance growth of countries in the Olympics over the years. [17] Through such analysis, any player can review their progress record and also examine their opponent's progress. The analysis covered various aspects such as the total number of gold, silver, and bronze medals won by different countries, performance analysis of specific countries, and comparisons between various countries and participants [17]. Research data analysis is a popular and suitable method for studying the development of the Olympic Games. This type of analysis

involves examining large datasets and presenting their various characteristics primarily through visual formats such as graphs, charts, and more. The United States and China are expected to closely compete for the highest number of gold medals, with each country having equal chances of topping the leaderboard. The Unified Team of Germany and Britain will vie for the third position. [2]. EDA, or Exploratory Data Analysis, is an approach that offers a deeper understanding of the dataset. In addition to these approaches, a process known as survey data analysis has been used to statistically break down the data and provide a comprehensive understanding. Interpreting and analyzing data is one of the main challenges of big data analysis. The Olympics are the subject of a wide range of studies, including statistical visualizations, evaluating the performance of athletes and changes in the performance of different countries [9]. Another research paper illustrates the use of Exploratory Data Analysis (EDA) to investigate the origin and distribution of naturally occurring contaminants such as Fluorine, Barium, Manganese, Arsenic, and others in the groundwater of Southern Quebec, Canada. [16] For the study, researchers utilized a Groundwater Chemistry database combined with 16 regional projects. The final dataset included information about the supply framework, geological settings, hydrological conditions, and inorganic water chemistry. The results revealed that the sources of these contaminants are natural [16]. So, we concluded that research data analysis is a common and useful technique for assessing how the Olympics have changed over time.

PROPOSED SYSTEM

The proposed work for the Olympic data analysis and prediction system involve collecting and cleaning up large amounts of datasets taken from Kaggle. After the collection, and organized the data, we use research data analysis methods to gain meaningful insights into the performance of athletes, countries, and sports. We also use various visualization techniques to present the0 data in an attractive and user-friendly format that is easy to understand. These views will allow us to identify patterns, trends, and external factors that may not be apparent in the raw data. The aim is to provide a comprehensive understanding of Olympics data that can be used to inform decisions and strategies for participating countries, athletes and sports. The results of our analysis and prediction can be used to produce a report that encourages countries to improve their performance at prestigious sporting events. We use the various Python libraries and machine learning for the different analysis and prediction on the datasets. Our project provide the effective knowledge that will help countries improve their performance at the Olympic Games and other sporting events and make informed decisions about the future of sport.

For the prediction we tested some Machine Learning Algorithm to find the accuracy in which Bagging and boosting are powerful ensemble learning methods. For our dataset, problem statement and desired output, bagging is more suitable. It's great for reducing variance, especially with models like decision trees. Random Forest, a popular bagging technique, creates multiple decision trees from bootstrap samples of the data. Each tree is trained independently on a subset of features. The final prediction is made by averaging or voting. Random Forest is effective at reducing overfitting and is more robust to noisy data and outliers compared to boosting techniques like Gradient Boosting, AdaBoost etc. It often provides better generalization performance than Bagging Classifier and offers insights into feature importance.

METHODOLOGY

This methodology ensures a thorough understanding of the data. To evaluate how the Olympics Games changed over the time. The evolution of the Olympic Games is influenced by many variables. So we used

the following methodology on the Olympics dataset and create Olympics data analysis and prediction. In this work we have proposed an algorithm called Player Winning Probability Model (PWPM).
In this model a dataset has been used based on details of different players through which their winning percentage has been predicted by using PWPM which is based on Random Forest Classifier algorithm. In this model different steps has been followed they are:

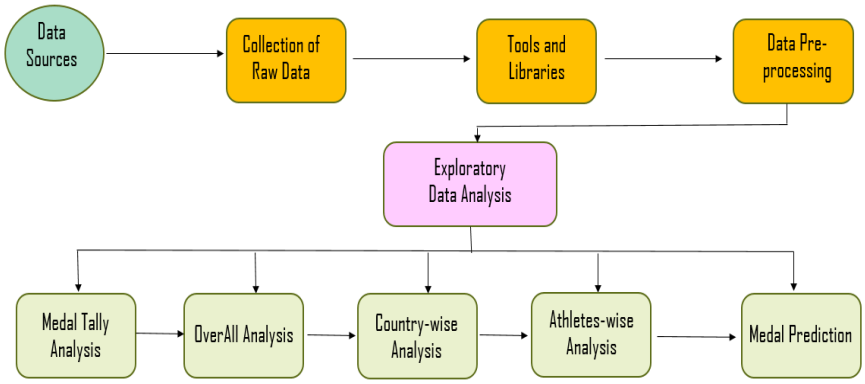

Fig 1: Steps in System Design

1. **Data Collection:** The initial step is to gather comprehensive datasets from Kaggle which includes the NOC, Athletes demographics (name, age, gender, nationality), Season, Sport, Event and medal outcome etc.
2. **Data Pre-processing:** After the collection of datasets the next step involves the data pre-processing of the datasets using various tools and python libraries to remove the incomplete or irrelevant entries, duplicates data, attributes having null values and standardizing data formats that can be used for further analysis and prediction.
3. **Exploratory Data Analysis (EDA):** Exploratory Data Analysis (EDA) involves examining and summarizing a dataset to uncover insights and understand the data's underlying patterns. This process may involve tasks like plotting and visualizing data, calculating summary statistics, and detecting patterns, outliers, and relationships within the data. EDA can also help identify potential data issues or inconsistencies that may need addressing during the pre-processing stage. Creating visual representation (graphs, scatter plots, heat maps) etc. This illustrate the relationship within data , identify trends and patterns and changes in the Olympics over the time Also various machine learning algorithm applies to predict the chances of winning medal (Gold, Silver and Bronze).
4. **Algorithm Selection:** Ensemble learning methods, particularly Bagging and boosting algorithms, were chosen for their ability to improve prediction accuracy and reduce variance.

These methods are well-suited for the project's problem statement, which involves predicting athletes' performance based on their physical attributes.

a. **Bagging (Random Forest):** Random Forest, a popular Bagging technique, was employed to create multiple decision trees from bootstrap samples of the data. Each tree was trained independently on a subset of features, and the final prediction was made by averaging or voting. Random Forest proved effective at reducing overfitting and was more robust to noisy data and outliers compared to boosting techniques like Gradient Descent and AdaBoost.
b. **HistGradient Boosting classifier :** HistGradient boosting classifier, a boosting algorithm, was tested for its ability to improve model performance by iteratively optimizing a loss function. However,

Gradient boosting yielded a lower accuracy of 62.28%, indicating limitations in handling noisy data and outliers compared to Random Forest.

c. **Bagging Classifier:** The Bagging Classifier algorithm was used to combine multiple base classifiers trained on different subsets of the training data. While Bagging Classifier improved upon Gradient Descent with an accuracy of 63.77%, it still fell short of Random Forest's performance.

d. **AdaBoost:** AdaBoost, another boosting algorithm, was applied to boost weak learners into a strong learner by assigning higher weights to misclassified instances. However, AdaBoost achieved the lowest accuracy of 52.36% among the tested algorithms, indicating challenges in handling the dataset's characteristics.

5. **Model Evaluation:** Random Forest emerged as the most suitable algorithm for the project, achieving an accuracy of 70.22%. The results demonstrate Random Forest's effectiveness in reducing overfitting and providing better generalization performance compared to Gradient Descent, Bagging Classifier, and AdaBoost.

RESULTS AND ANALYSIS

We are analyzing the Summer Olympics datasets and it includes the data collection from 1896 to 2016. Also the datasets contain the columns like (NOC, Athletes demographics (name, age, sex, region, height, weight), Season, Sport, Event and medal etc. and there is approximately 30,000 rows. With the help of data analysis we can identify patterns and create visual representation of datasets which contain (heat maps, scatter plots, graphs) and many more and for creating the visual representation we use the Data science, python and their libraries.

A. Number of Events in the Olympics (1896 - 2016):

The increase in the number of Olympics events from 43 in 1896 to 306 in 2016 illustrates the significant expansion and diversification of the Olympic Games over a span of 120 years. The number of events also increased as the Olympics became more inclusive, particularly regarding female athletes. Women's events have been added progressively since 1900 when first participated in the Olympics. This has significantly increased the total number of events.

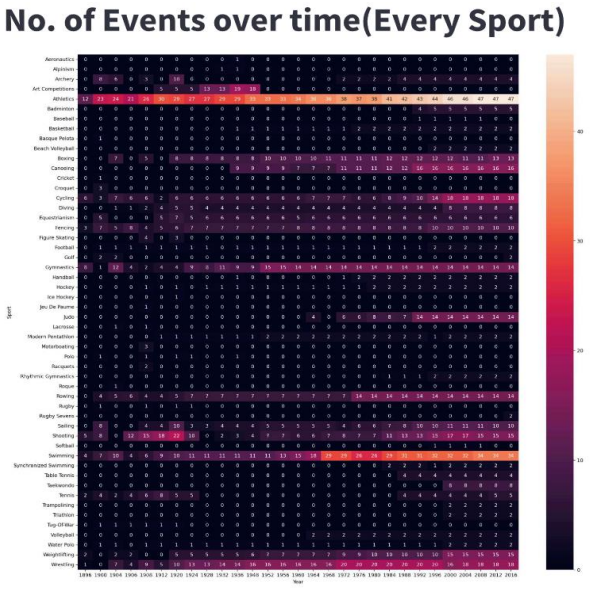

Fig 2: Heat map of no. of events over time (Every Sport)

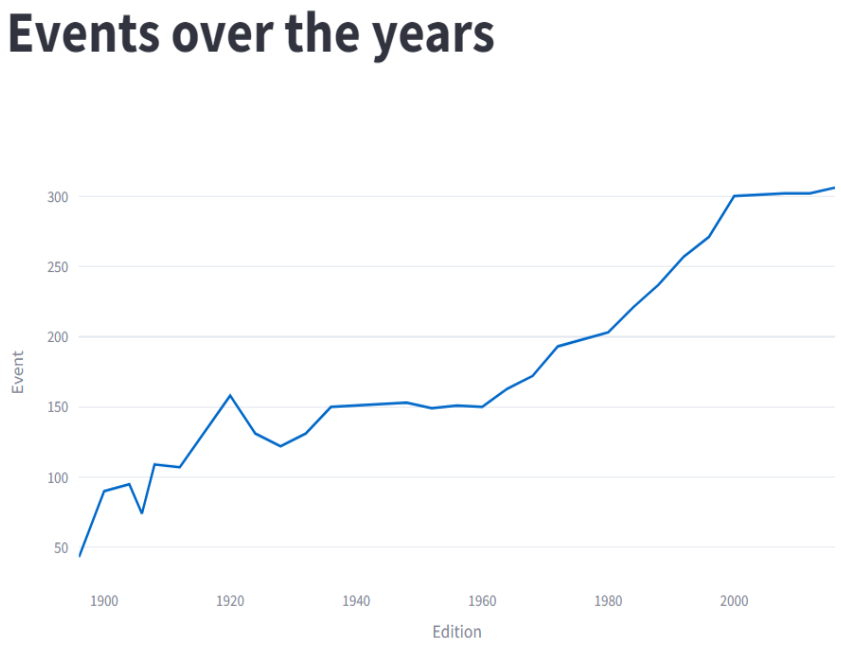

Fig 3: Events over the time

Figure 3 shows, the number of events was gradually increasing from 1896 – 1960 and then number of events increased tremendously to 2016.

B. Number of participating Nations in Olympics (1896 – 2016):

The increase in the number of participating nations at the Olympics from 12 in 1896 to 204 in 2016 illustrates the significant expansion and global reach of the Olympic Games over 120 years. Over the years, many new nations have been recognized internationally and have become members of the International Olympic Committee (IOC). This expansion includes countries that gained independence during the 20th century and chose participate in the Olympics as a mark of their new national identity.

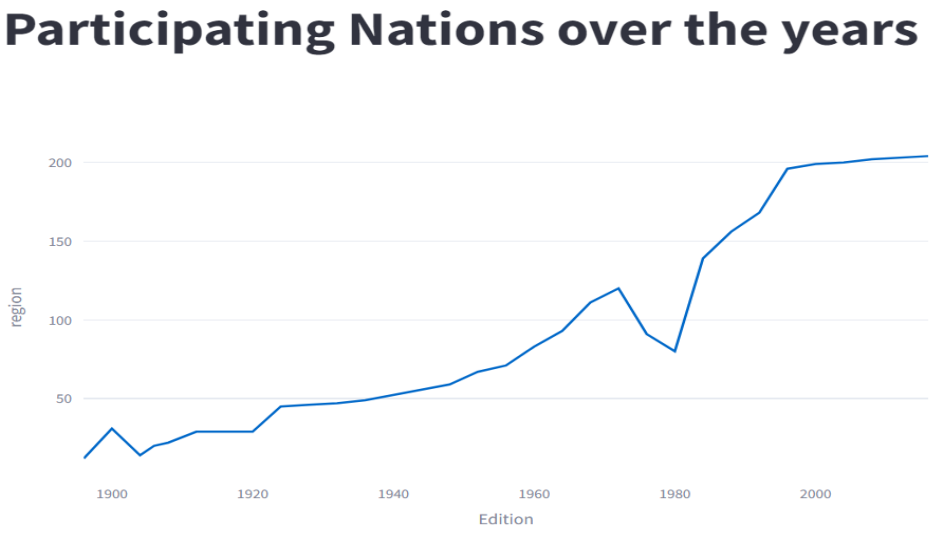

Fig 4: Participating Nations over the time

Figure 4 shows, the increase in the number of participating nations at the Olympics from 1896 to 2016 illustrates the significant expansion and global reach of the Olympic Games over 120 years.

C. Overall Most Successful athletes in Olympics (1896 – 2016):

Analyzing the top most athletes in the Olympics from 1896 – 2016. This athletes are analyzed on the basis of winning medals in Olympics till 2016. The athletes have more number of medals are the consider as the most successful athletes. These athletes not only achieved incredible individual success but also helped elevate their sports on the global stage, inspiring future generations of Olympians. Their records, stand as significant milestones in the history of the Olympic Games.

	Name	Medals	Sport	region
0	Michael Fred Phelps, II	28	Swimming	USA
30	Larysa Semenivna Latynina (Diriy-)	18	Gymnastics	Russia
49	Nikolay Yefimovich Andrianov	15	Gymnastics	Russia
73	Borys Anfiyanovych Shakhlin	13	Gymnastics	Russia
97	Takashi Ono	13	Gymnastics	Japan
130	Edoardo Mangiarotti	13	Fencing	Italy
144	Dara Grace Torres (-Hoffman, -Minas)	12	Swimming	USA
157	Aleksey Yuryevich Nemov	12	Gymnastics	Russia
178	Jennifer Elisabeth "Jenny" Thompson (-Cumpelik)	12	Swimming	USA
195	Birgit Fischer-Schmidt	12	Canoeing	Germany
208	Ryan Steven Lochte	12	Swimming	USA
222	Paavo Johannes Nurmi	12	Athletics	Finland
234	Sawao Kato	12	Gymnastics	Japan
258	Natalie Anne Coughlin (-Hall)	12	Swimming	USA

Fig 5: Overall most successful athletes

Figure 5 shows, the Top most successful athletes in the Olympics . Michael Fred Phelps, II (USA) is the most decorated Olympian of all the time, with a total of 28 medals in the Olympics. So these are the athletes achieved incredible individual success in the Olympics.

D. Analyzing the participation of Men and Women in Olympics (1896 – 2016):

Analyzing the participation of men and women in the Olympics from 1896 to 2016 provides a glimpse into how gender representation in global sports has evolved over more than a century.

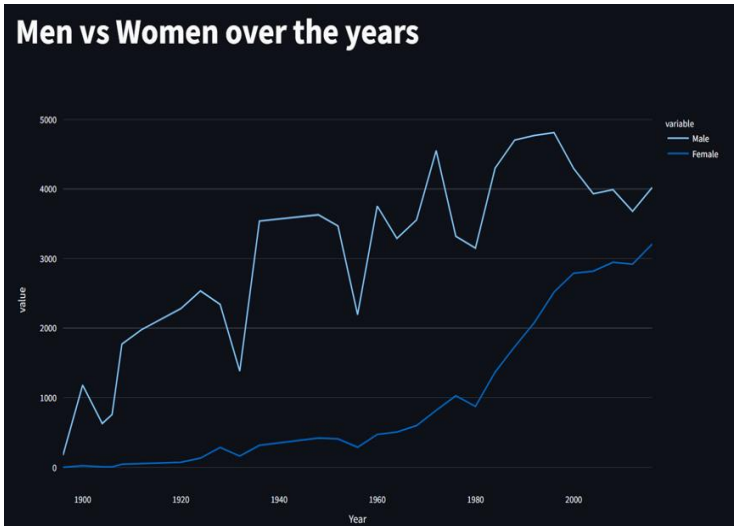

Fig 6: Men vs Women over the years

Figure 6 shows, that there is no women participated in the first modern Olympics Games, as they were initially intended only for male athletes. In the year 1900 women first participated in the Olympics and after that participation rates improves dramatically.

E. Analyzing the Height vs Weight

To analyze the chances of winning medal of male and female in the Olympics with respect to Height and Weight.

In which the male is represented with circle and female with cross.

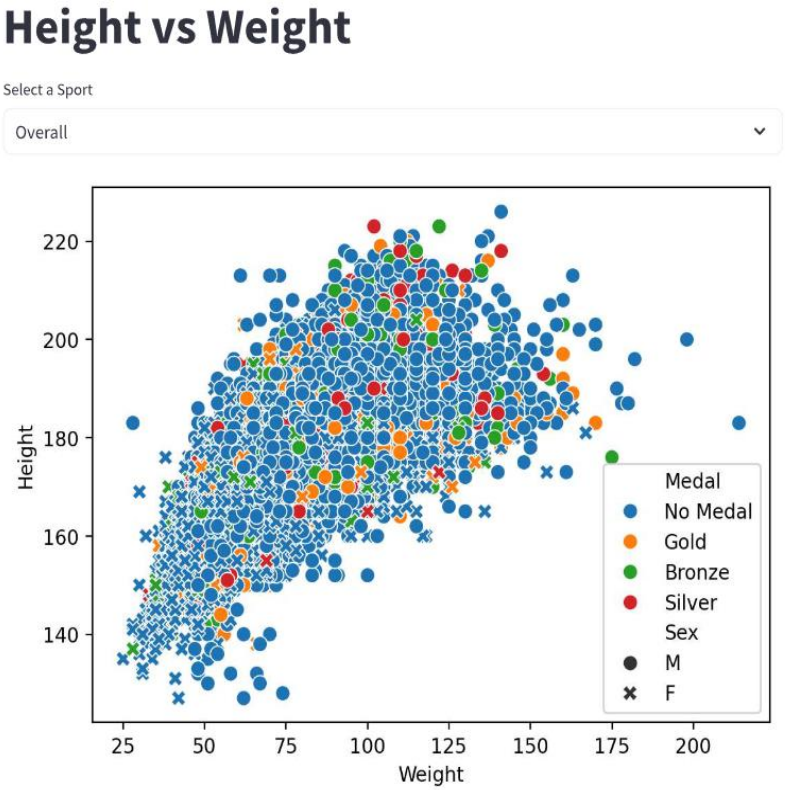

Fig 7: Height vs Weight

Figure 7 shows, according to the data female Olympic medal winners fall within a height range of 160 to 180cm and weight range from 50 to 150kg and men Olympic medal winners fall within a height range of 165 to 210cm and weight range from 55 to 170kg.

F. Player medal winning predictions

We tested the some machine learning algorithm for the prediction of winning medal of athletes. But we choose the Random Forest Classifier algorithm for the prediction of winning medal with accuracy 70.223325%.

S. No.	Algorithms	Accuracy
1.	HistGradient Boosting Classifier	62.28287%
2.	Random Forest Classifier	70.22332%
3.	AdaBoost Classifier	52.35732%
4.	Bagging Classifier	63.77171%

Table 8: Accuracy table of Predictive algorithm

	Actual_Value	Predicted_Value
166288	Silver	Bronze
71914	Bronze	Bronze
54639	Bronze	Silver
195047	Silver	Silver
58064	Gold	Gold
...	...	...
244871	Silver	Silver
153743	Silver	Bronze
94922	Bronze	Silver
204654	Gold	Gold
130870	Gold	Gold

403 rows × 2 columns

Fig 9: Prediction on winning medal

Figure 9 shows, the chances of winning medal through prediction using machine learning algorithm (Random Forest Classifier).

CONCLUSION

In conclusions from the detailed analysis of 120 years of Olympic data provide insightful observations into trends and patterns in Olympics Games. We utilized Exploratory Data Analysis to provide a detailed statistical and visual overview of national and individual performances from the 1896 to 2016 Olympics Games. Visualizations clarified the data, allowing us to draw significant conclusions. This analysis helps nations and athletes evaluate their performance, pinpoint areas for improvement, and make strategic decisions that enhance their prospects for future Olympic success and the prediction helps to know the probability of winning medal of for nations and athletes. Also the athletes aged 20 – 30 years being the most participative in the Olympics. The relationships between the numbers of athletes a country sends to the Olympics and its medal count is quite intuitive. Higher participation likely means more opportunities to win the Olympic medal. For the prediction we tested some Machine Learning Algorithm to find the accuracy in which Bagging and boosting are powerful ensemble learning methods. For our dataset, problem statement and desired output, bagging is more suitable as we are getting 70.223325% accuracy from Random Forest, 62.282878% from Gradient Descent, 63.771712% % from Bagging Classifier and 52.357320099% accuracy from AdaBoost method. It's great for reducing variance, especially with models like decision trees. Random Forest, a popular bagging technique, creates multiple decision trees from bootstrap samples of the data. Each tree is trained independently on a subset of features. The final

prediction is made by averaging or voting. Random Forest is effective at reducing overfitting and is more robust to noisy data and outliers compared to boosting techniques like Gradient Descent, AdaBoost etc. It often provides better generalization performance than Bagging Classifier and offers insights into feature importance.

REFERENCE

1. Wikipedia:https://en.wikipedia.org/wiki/Olympic_Games.
2. Antarlina Sen and Gaurang Margaj, “A prediction model for which country will win the highest
3. number of Gold“, 2016.
4. Leonardo De Marchi, “Data mining of Sports performance data”, 2011.
5. 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.
6. Chandra Segar Thirumalai and Monica Sankar, “Heuristic Prediction of Olympics using Machine Learning”, International Conference on Electronics, Communication and Aerospace Technology, April 2017.
7. Alexander Rathke and Ulrich Woitek, “Economics and Olympics: An Efficiency Analysis”, January 2007.
8. 120 years of Olympic Dataset, Available: https://www.kaggle.com/datasets/heesoo37/120-
9. years-of-olympic-history-athletes-and-results
10. Analyzing Evolution of the Olympics by Exploratory Data Analysis using R Rahul Pradhan1 , Kartik Agrawal1 and Anubhav Nag1.
11. DATA ANALYSIS AND VISUALIZATION OF OLYMPICS USING PYSPARK AND DASH-PLOTLY Harshal S.
12. Analysis of 120 Years of Olympic Results Kassidy Chaikin, Christopher Egan, Emily Lepore, William Wenze.
13. Web Application of Olympic data analysis Farkande Vaishnavi ,Gurav Vaishnavi, Borse Tejas
14. "The Modern Olympic Games" (PDF). The Olympic Museum. Archived from the original (PDF) on 6 September 2008. Retrieved 29 August 2008.
15. Olympics Data Analyzer with Prediction Hitanshi Shah, Jay Sheth, Hetvi Savla, Jyoti Bansode, Bijal Patel, Aruna Yewale
16. Analysing 120 Years Olympics Dataset Exploratory Data Analysis Prof. Lavina Jadhav, Aadarsh
17. Krishan Yadav
18. Dey S K, Rahman M M, Siddiqi U R and Howlader A 2020 Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach J. Med. Virol. 92 632–8
19. 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. 114 104500
20. 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) 7 251–3
21. 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

22. Olympics Data Analysis Web App Lication and Prediction Using Machine Learning Creators Prof.Satish J.Manje1 Ms.Shruti Rakesh Dubey2 Ms.Sejal Sunil Parche2 Ms.Disha Laxman Bondre2
23. Data Analytics on Olympics Datasets Surya Sena Reddy, Suraj Kumar