A Deep Dive into Historical Olympic Performance Trends

Sataakshi Bhangelwal, Geetansh Kumar, Rudraksh Agarwal

***Abstract*—This project aims to analyze the historical trends of Olympic performances, focusing on medal tallies, country-wise achievements, sport-wise achievements, and participation in various sports. By examining these patterns, we can discover insights such as the performance of certain countries in specific sports, shifts in performance over time, and variations in participation based on factors like gender and athlete age, etc. The goal is to identify and explore significant trends of the Olympics over the years. In addition, we might take this analysis a step further by predicting the 2024 Olympic medal tally, with a plan to compare these predictions against the actual results.**

I. Introduction

The Olympic Games, a global stage for athletic excellence, have served as a mirror reflecting societal, cultural, and political trends for over a century. From showcasing the dominance of powerhouse nations to highlighting the rise of underdog stories, the Olympics offer a rich dataset that captures not only athletic performance but also global dynamics. This project delves into the historical trends of Olympic performances, aiming to uncover meaningful patterns that reveal the evolution of sports and their impact on participating nations and athletes.

By focusing on medal tallies, country-wise and sport-wise achievements, and the ever-changing dynamics of participation, this analysis seeks to highlight trends such as the consistency of dominant nations, emerging talents from new regions, and shifts in the popularity of sports. Gender and age-related participation patterns will also be examined, providing insights into inclusivity and representation over the years.

Furthermore, the project aspires to go beyond descriptive analysis. Using historical data and statistical modeling, we aim to forecast the medal tally for the 2024 Olympics. This predictive approach will allow for a unique comparison with the actual results, offering a practical evaluation of the model's accuracy and the reliability of data-driven sports analytics. This combination of retrospective and forward-looking analysis positions the project at the intersection of history, data science, and the future of sports performance analytics.

II. Motivation

The project "A Deep Dive into Historical Olympic Performance Trends" provides valuable insights into sports analytics by analyzing historical Olympic data. This contribution is crucial as it helps identify strengths and weaknesses in countries' performances. It will be a guide into how funds have been and will be allocated after carefully analyzing the country's performance. It will also be a way to encourage the participation of underrepresented groups. It will also promote inclusivity in sports. Additionally, predicting the 2024 Olympic medal tally using historical data fosters a data-driven approach to future preparations. Overall, this project significantly enhances the understanding of Olympic performance trends, ultimately helping athletes and sports committees improve performance and participation. We are also finding trends based on the ages of athletes to understand sports careers better. The motivation behind the project, **"A Deep Dive into Historical Olympic Performance Trends,"** lies in the transformative potential of sports analytics to drive meaningful change in the global sporting landscape. The Olympics, as a global stage for athletic performance, provide a unique opportunity to analyze not only athletic excellence but also the underlying factors that contribute to success and participation. By examining historical Olympic data, this project seeks to address key challenges and opportunities in sports management, inclusivity, and performance enhancement.

1) **Improving Resource Allocation:** Analyzing medal tallies and country-wise achievements highlights patterns in national performance and provides insights into how resources like funding, training infrastructure, and athlete development programs have impacted results. This analysis enables nations to make informed decisions about where to invest in the future to enhance competitiveness.
2) **Encouraging Underrepresented Groups:** Inclusivity is a cornerstone of modern sports. By examining participation trends across gender and age, the project aims to identify disparities in representation and performance. Insights derived can help design policies and programs that encourage the participation of underrepresented groups, ensuring a more diverse and equitable sporting community.
3) **Promoting Strategic Preparation:** The ability to predict future outcomes, such as the 2024 Olympic medal tally, introduces a data-driven approach to athletic preparation. These predictions can guide sports committees and athletes in setting realistic goals, tailoring training programs, and benchmarking performance against global standards.
4) **Understanding Athlete Careers:** By analyzing athlete age trends and their impact on performance, the project seeks to provide valuable insights into the lifespan of sports careers. This understanding is crucial for planning athlete development pathways, retirement transitions, and post-career opportunities.
5) **Identifying Strengths and Weaknesses:** Detailed anal-

yses of sport-wise performance trends allow countries to recognize areas of strength and pinpoint sports where they are underperforming. This helps refine focus and efforts on underdeveloped areas while sustaining success in established ones.

6) **Fostering Global Competitiveness:** By learning from the strategies and successes of leading nations, emerging countries can adopt best practices, fostering greater global competitiveness and raising the overall standard of sports.
7) **Driving Policy and Decision-Making:** The insights from this project can serve as a foundation for evidence-based policymaking in the realm of sports. From designing grassroots programs to redefining elite athlete training models, data-driven strategies will ensure a systematic approach to success.

In essence, this project is motivated by the aspiration to leverage data analytics for a more inclusive, equitable, and competitive sporting world. By diving deep into historical Olympic trends and combining descriptive and predictive analytics, this study not only provides a rich understanding of past performances but also contributes to shaping the future of global sports.

III. Data Retrieval

For our project, we leveraged several datasets available on Kaggle to analyze historical Olympic data and derive insights for predicting the outcomes of the Paris 2024 Olympics. These datasets were hosted on Google Drive, and we connected our notebook to Google Drive for seamless access and analysis. Below is a detailed description of the datasets used:

A. 126 Years of Historical Olympic Dataset

This comprehensive dataset provides detailed records of the Olympic Games from their inception in 1896 through to 2022. It encompasses various attributes such as athlete participation, medal tallies, country-specific achievements, and the progression of sports over time. By utilizing this dataset, we performed an in-depth analysis of historical trends, such as:

- The rise and fall of powerhouse nations in the medal standings.
- Changes in participation rates across genders and age groups.
- The introduction and evolution of new sports disciplines.

These insights enabled us to understand the evolution of the Olympics and provided a strong foundation for forecasting the outcomes of the Paris 2024 Olympics.

Source: Kaggle - 126 Years of Historical Olympic Dataset

B. World Population Dataset (1960–2018)

This dataset offers an extensive view of global population trends from 1960 to 2018, segmented by country. It includes valuable data points such as total population and growth rates. We integrated this dataset into our analysis to explore the relationship between population size and Olympic performance. By examining population trends, we aimed to:

- Understand how demographic factors might influence sports participation rates.
- Correlate population dynamics with medal tallies and participation rates, particularly for emerging countries.

This analysis provided valuable context for understanding the impact of demographic factors on a nation's performance in the Olympics.

Source: Kaggle - World Population Dataset

C. Integration of Datasets

By combining the insights from these two datasets, our analysis bridges historical Olympic performance with demographic trends. This holistic approach creates a comprehensive view of the factors influencing success at the Olympic Games. It not only enhances the descriptive analysis of past events but also strengthens the predictive model for Paris 2024.

IV. Data Cleaning

For the analysis of the Olympic dataset, several data cleaning techniques were employed to ensure the dataset's consistency and accuracy. These steps were carried out with the intention of refining the data for analysis and predictive modeling. Below is a detailed explanation of the key data cleaning processes applied:

1) **Removed Duplicates:** The dataset was first cleaned by removing any duplicate rows to ensure that each entry in the dataset was unique. This was achieved by using the `drop_duplicates()` function, which helps eliminate redundancy in the data.
2) **Removed Irrelevant Columns:** Unnecessary columns that did not contribute to the analysis, such as `sport_url`, `result_date`, `result_location`, `result_format`, `result_detail`, and `result_description`, were dropped. This was done using the `drop()` method, which helps in focusing on the relevant columns for analysis.
3) **Standardized Text Data:** Several text-based columns, including `event_title`, `edition`, `sport`, and `result_participants`, were stripped of any leading or trailing spaces and converted to lowercase for uniformity. This ensures consistency in text processing and comparison.
4) **Extracted and Cleaned Numerical Data:** For the column `result_participants`, which contained both the number of participants and the number of participant countries, a regular expression was used to extract these two values separately into new columns, `participants` and `participant_countries`. The extracted values were then converted into integers using `astype(int)` to facilitate numerical analysis.
5) **Dropped Unnecessary Columns:** After extracting the necessary information, the original column `result_participants` was dropped, as it had been split into the newly created columns.
6) **Extracted Year and Olympic Type:** From the `edition` column, the year and type of Olympic (i.e., Summer or Winter) were extracted using a regular

expression. The `str.extract()` method was used to split the `edition` column into two new columns: `year` and `olympic_type`. This made it easier to analyze trends over time and by Olympic type.

7) **Created Dummy Variables:** The `olympic_type` column, which contained categorical values ('summer', 'winter'), was one-hot encoded using the `get_dummies()` function. This generated new binary columns indicating the type of Olympic event (Summer or Winter) for each record. The resulting dummy variables were then added to the dataset.
8) **Data Imputation:** Missing values in certain columns were handled by imputing them based on contextual or statistical methods. In some cases, missing data was filled using the mean, median, or other appropriate values to ensure that the dataset remained complete for analysis.
9) **Normalization:** Large and small numerical values across the dataset were normalized to ensure consistency in scale. This helps in preventing any features with larger numerical values from dominating the analysis due to their scale.
10) **Ensuring Data Integrity:** Throughout the cleaning process, special attention was paid to maintaining the integrity of the dataset. This involved verifying data types, handling null values, and ensuring consistency across the various columns.

These data cleaning steps were essential in transforming the raw dataset into a format suitable for in-depth analysis. After these steps, the dataset was ready for the next phase of exploration, modeling, and prediction, which is crucial for identifying key trends and making predictions for future Olympic events like Paris 2024.

V. Problem Statements

A. Question 1- Geetansh

What is the general trend in women participation country wise over the years? What countries are doing well and how do they compare to the best performing countries?

This general question can lead to multiple hypotheses including the need of promoting women empowerment, understanding and highlighting gender disparity, urgent need to change in policies and awareness, etc. This is a significant question because in this day and age we should ideally have equal women participation in sports, especially in a major competition like the Olympics. This also helps us identify which countries need to focus on the gender disparity issue more

B. Question 2- Geetansh

Are there any sports which are on the decline and losing popularity among participants? Also, are there some sports which have gained popularity over the recent years?

This general question can again lead to multiple hypotheses including the need to spread awareness about some particular sports, predict which sport to remove from the Olympics, what sports have previously suffered this fate, etc. This is a significant question because we should be aware of which sports are losing popularity to save them from getting extinct. On the world stage we should be able to identify which sports are on the rise, it can be beneficial for marketing and branding (great opportunity for money making).

C. Question 3 - Sataakshi

How do the trends in medal counts for team sports compare to those for individual sports across different countries over the years, and what insights can be drawn from these comparisons regarding each country's performance in the Olympic Games?

This analysis compares the trends in medal counts for team sports versus individual sports across different countries over the years. Using queries to extract data I have analysed how the number of medals have been won in team events vs in individual events for different countries. This study provides an analysis of how countries perform in team versus individual competitions.

The question of how trends in medal counts for team sports compare to those for individual sports across different countries is significant for several reasons. It enhances our understanding of national strengths, allowing us to identify which countries excel in team versus individual sports and informing national sports policies. This analysis can also guide investment in sports programs. So, this question provides valuable insights that can shape the future of sports

D. Question 4 - Sataakshi

How has the participation of women athletes in various sports evolved, and what trends can be observed in terms of minimum and maximum participation levels across selected sports?

The analysis of women's participation in various sports over time reveals significant trends in their involvement. By filtering the data for women athletes, we observe how participation levels have increased in these years in all sports.

The analysis aims to highlight advancements in women's participation in sports, demonstrating progress in inclusion and representation. By examining trends over time, we can assess the effectiveness of initiatives aimed at increasing female participation and identify sports where participation still lags. This information informs sports organizations and advocates about the current situation of women's sports. This will help in gender equity in olympics.

E. Question 5 - Rudraksh

Table Tennis and Tennis are similar yet different sports. The players I have seen in both games seem to have different builds. The hypothesis is that we can build a model using Height, Weight, and athlete's country to predict which sport they belong to.

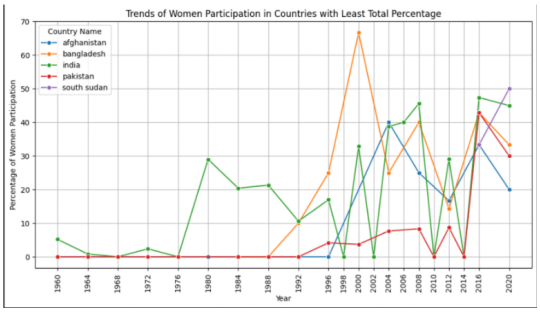
F. Question 6 - Rudraksh

In athletics, height, weight, age, and country are major indicators for success in Olympics. We have made to achieve the same.

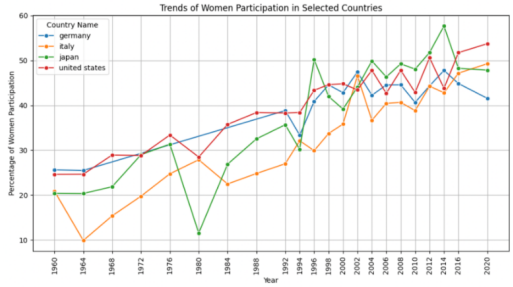
VI. John Tukey's Exploratory Data Analysis

A. EDA 1 (Geetansh) :- Percentage Women participation per country per olympic

Insights :- There are a lot of countries where women participation is significantly low as compared to men. These countries need to pay attention and target women participation in sports. Below are the trends for countries where women participation is the least



Now, let's compare this with leading countries in sports.

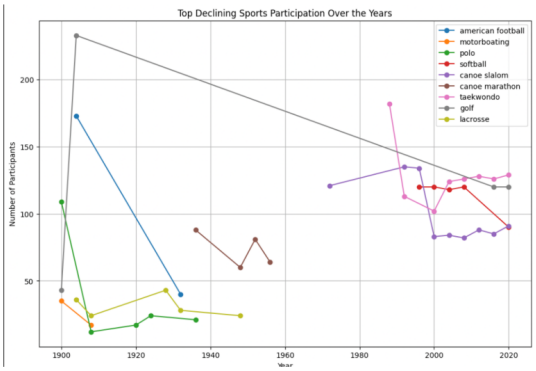


We clearly see the difference between the rise in women's participation in these two graphs.

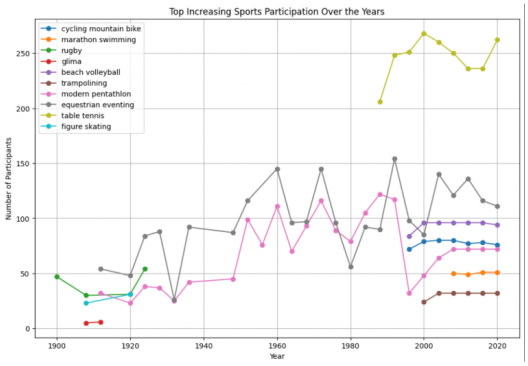
This data can be passed downstream and used for predicting the future participation of women in the Olympics. Also, this is a significant indicator of women's empowerment in each country and can be very useful to highlight these issues.

B. EDA 2 (Geetansh) :- Identifying Declining and On The Rise Sports

Insights:- There are a few sports whose trends are declining. In the past, the declining sports were removed from the Olympics altogether. The participation slowly decreases and then vanishes. Some sports have been gaining more popularity in the recent past. Below we see extinct sports in the past and sports whose popularity is decreasing.

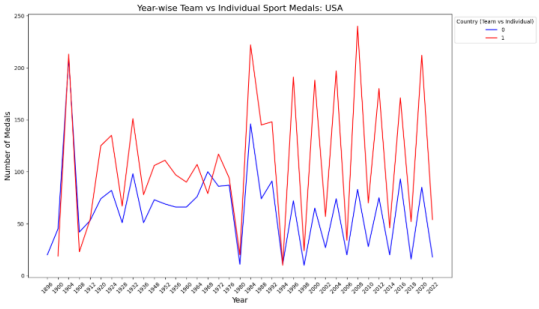


Below are the sports which are on the rise.



In the above two graphs, we see trends of increase and decline in sports. This data can be sent downstream, and we can predict which sport might get removed after how many years. We can also try to use this for marketing opportunities for the sports which are on the rise.

C. EDA 3 (Sataakshi) :- Country-wise Comparison of Team vs. Individual Sport Medals Over the Years



Inference: The graph depicting year-wise medal counts for team versus individual sports for the USA, we see:

The graph indicates that the USA has consistently won more medals in team sports (shown in red) compared to individual sports (shown in blue) across the years. There is a strong participation in team-oriented events reflecting the country's competitive advantage.

In summary, the graph highlights a strong national performance in team sports for the USA, while also providing valuable information about the potential for growth in individual sports through strategic investments and focused training programs.

D. EDA 4 (Sataakshi) :- Sport-wise Trend of Women's Participation Over the Years

	sport	year	women_participation_count
0	artistic gymnastics	1900	1
1	golf	1900	12
2	tennis	1900	15
3	archery	1904	17
4	figure skating	1908	6
..	...	...	...
585	speed skating	2022	184
586	ice hockey	2022	235
587	alpine skiing	2022	276
588	biathlon	2022	348
589	cross country skiing	2022	445

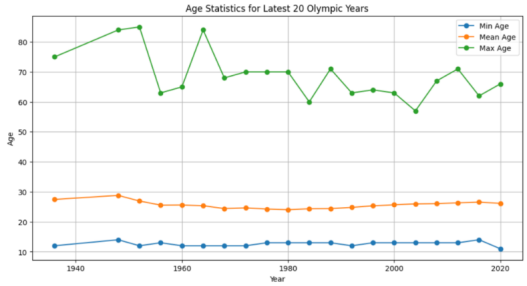
[590 rows x 3 columns]

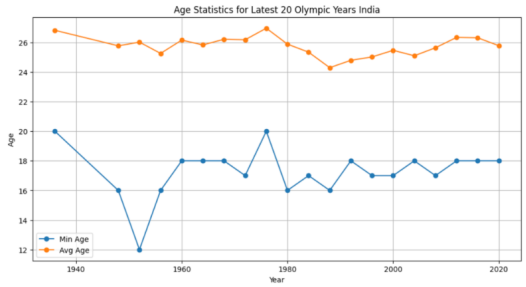
We start by filtering the data sports wise to aggregate the participation of women in different sports over the years followed by visualizing them.

We can infer from this graph that alpine skating has seen a significant increase in women's participation when compared from 1936 vs 2022. We can see that a sport that had less than 50 participants has over 300 participants at some point too. This clearly signs of the gender equality aspect. This number might be still very lease when compared to the male participants but it is a start. This EDA hence shows that all the sports will see even greater participation in the future too.

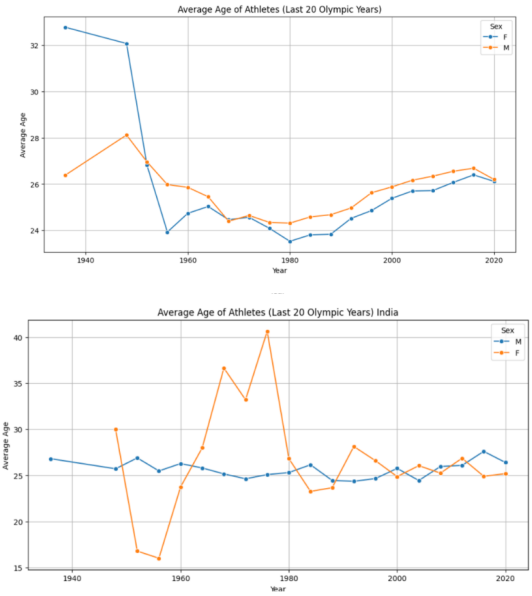
E. EDA 5 (Rudraksh)

Through the graphs made below, we can see that there is not much change in the average age of the athletes in the world. Even the minimum and maximum ages being observed are more or less in the same band of ages.



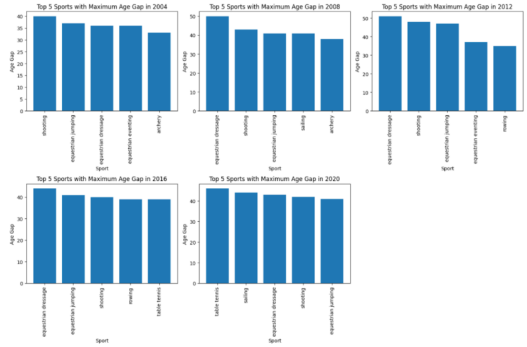


There is a clear difference in the average age of male and female athletes in the world. For males , it is consistently greater than that of females. For India, it looks different there is no clear conclusion that can be made.



F. EDA 6 (Rudraksh)

As can be seen, there are sports like shooting, equestrian jumping/dressage/eventing, archery, and table tennis, which have the longest careers spanning over 40 years. (12 Olympics)



As can be seen below graph, there are sports like boxing, triathlon, cycling, and gymnastics in which the career lasts for around 15 years (3-4 Olympics).

VII. Machine Learning and Statistical Modeling Algorithms

A. Model 1 (Geetansh)

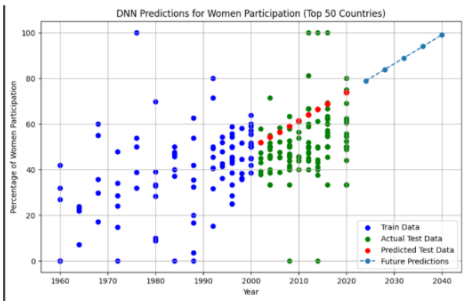
For this prediction problem we have used a custom designed Deep Neural Network (DNN). More details will be shared below.

Why DNN : - We tried linear regression first but that was giving very bad r2 value and wasn't giving good predictions. Then we went with DNN due to the flexibility it provides and its ability to learn on non-linear dataset. It also generalizes results better. After some hit trials we have also included some hidden layers with ReLU as the activation function and final layer with linear activation. Below is the DNN architecture we have created :-

- 3 hidden layers with 64, 32 and 16 nodes. ReLU was the activation function.
- Final layer with one output and linear activation function.
- Optimizer was Adam.
- Mean squared error as our loss function.

Training and Tuning : - For training we have taken data from year 1960 to year 2000 so that we have enough data for our model to calculate weights. Testing is done on data from year 2000 to 2020 in order to test the efficiency of our model. Furthermore, predictions are done for future Olympics as well

We have trained our DNN for 150 Epochs with a batch size of 4. By balancing the Epochs, loss value and batch size we have prevented overfitting on the training data. Lower batch size helped us gain more deeper insights into the training data.

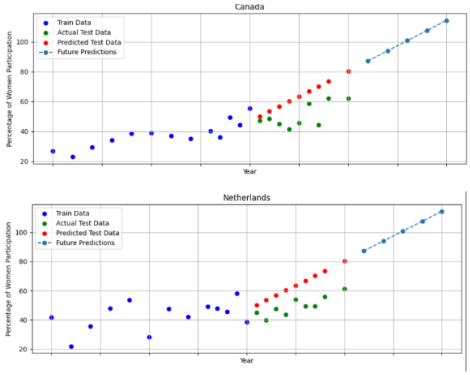


Evaluation Metrics : - For evaluating our results on the test data we have used r2 value and mean total loss.

R2 : - r2 value measures how close our predictions are from the actual data. It basically measures the distance of our prediction from the average of the actual data. R2 value is considered to be good if it is greater than 0.5 and closer to 1. For this model, our r2 value is 0.850, which is considered a good r2 value for any model.

Mean Squared Loss : - As the name suggests, this is commonly used to calculate the total loss for our dataset. A low loss suggests that the accuracy is good/high. We have tried to minimize this as much as possible and in our Python Notebook we can see this value reducing with every epoch. The final mean squared loss was 15.393.

Both our metrics are giving good values and together suggest that our DNN has performed well.



Intelligence Gained : - After testing from 2000 to 2020, we have also tried predicting the future which was our goal from the very start. We have predicted percentage women participation in general and also tried to plot the graphs for some specific countries with significant trends. Our analysis says that the women participation percentage is bound to increase in the future, which is a great sign for world sports.

The predicted increase helps us and all stakeholders understand the future progress of women participation in sports and olympics. This will help with better policies and awareness in the future. An increase in gender equality can be concluded from our future prediction analysis and intelligence gained.

B. Model 2 (Geetansh)

KMeans clustering for clustering on unlabelled data to divide sports into 2 clusters i.e. Rising and Declining. For slope, additionally we have used linear regression to calculate slope of the regression line. More details will be shared below.

Why KMeans : - We have used the KMeans algorithm here because we already knew how many clusters we wanted and the K in this algorithm signifies exactly that. KMeans clustering is very good for data with multi dimensional features and capturing patterns. We also considered DBScan but for this dataset KMeans was a perfect fit and more simpler to understand.

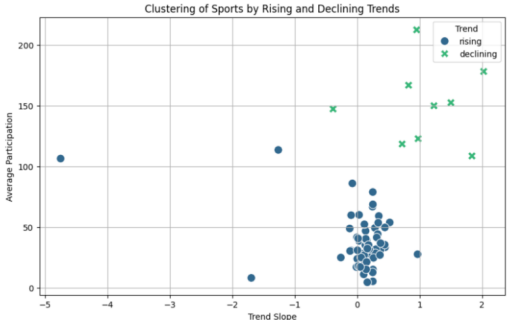
Training and Tuning : - For this we have firstly kept $n_clusters as 2, since we want to divide in 2 clusters. We have also set a rand$

Evaluation Metrics : - For our evaluation, we have kept Silhouette Score, Davies-Bouldin Index, and Inertia.

- Silhouette Score : This measures how close one point is to the points in the other cluster. A value closer to 1 means it is well connected to its own cluster and far from neighboring clusters. For us this value was 0.703, which signifies that points in our clusters are well connected to their own clusters.
- Davies-Bouldin Index : This determines the average similarity for clusters, the lower the value the better. For this came at 0.534, which is a relatively low value.
- Inertia : This calculates how tightly the points are bound to their centroids (cluster center). For us this value is 77.11, which tells us that data points are closely connected to the centroids.

Combined, all these 3 metrics tell us that our clustering approach divides the data very well.

Intelligence Gained : We see in our output that the clusters have been defined pretty clearly with slope and average participation being the deciding factors and the data is labeled as well now. We have successfully labeled the input unlabeled data.



This visualization gives us a good idea on how to identify sports which are declining and need urgent attention in order to save them. Also, this gives us an idea on the rising sports which could be super beneficial for marketing, business and other opportunities. This proves our initial hypothesis correct.

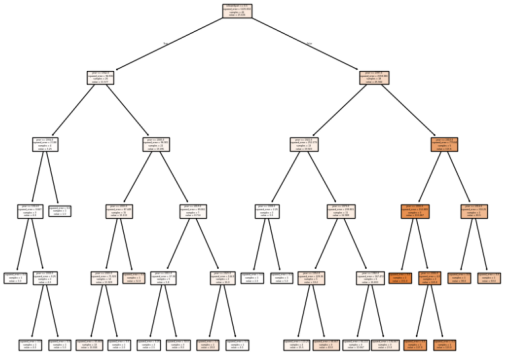
C. Model 3 (Sataakshi)

To solve this we are implementing Decision tree and Random forest model to evaluate the MSE and R^2

Why Decision Tree and Random Forest: - I tried linear regression first. This did not seem to work well due to the lack of features. For building a better model, we used Decision Tree and Random Forest. The Decision Tree algorithm is a good choice when dealing with simple relationships between features. In this case, the primary features are 'year' and 'isTeamSport'. We aim to predict 'medal_count'. Decision Trees are highly interpretable, allowing for easy visualization and understanding of decision-making processes. They are well-suited for this problem, where we want to identify trends in medal counts based on whether the sport is a team sport and the year of the event.

Random Forest is a similar method that builds multiple decision trees and aggregates their results. This technique improves on Decision Trees by reducing overfitting and increasing accuracy. Since the relationship between 'year', 'isTeamSport', and 'medal_count' might involve complex interactions, Random Forest can capture these better by averaging over many trees, thus providing a more robust solution to the problem.

Training and Tuning: - The Decision Tree was trained with a maximum depth of 5 to prevent overfitting. We know that deeper trees tend to memorize data instead of generalizing, which results in a poor model. This approach also helped in simplifying the tree, making it more interpretable. The model was fit using the training set X_{train} and y_{train}. Predictions were made on the test set X_{test}.



Evaluation Metrics Decision Tree: -For evaluating the performance of the Decision Tree model on the test dataset, we utilized two key metrics: the R^2 value and the Mean Squared Error (MSE). These metrics provide insight into how well the model fits the data and how accurate its predictions are.

- **R^2 Value: 0.881**
  - The R^2 score of 0.881 indicates that the model explains 88.1% of the variance in the data. This suggests that the Decision Tree is fairly effective at capturing the underlying patterns and relationships between the features and the target variable.
  - A high R^2 value like this typically indicates a good fit, where the model is able to predict the target variable with reasonable accuracy, though there might still be room for further improvement.
- **Mean Squared Error (MSE): 8.4**
  - The MSE value of 8.4 shows that the model's predictions have some degree of error, though the error is relatively low considering the complexity of the data and the nature of the Decision Tree model.
  - While the MSE is not negligible, it suggests that the model performs reasonably well in terms of minimizing the difference between predicted and actual values.
  - Further improvements can be achieved through techniques such as parameter tuning (e.g., adjusting tree depth, splitting criteria) or using ensemble methods like Random Forest to reduce overfitting and improve predictive accuracy.

In conclusion, while the Decision Tree model performs well with an R^2 value of 0.881, the MSE of 8.4 suggests that further improvements can be made to reduce prediction errors. Future work could focus on refining the model through parameter tuning and exploring more advanced techniques.

Evaluation Metrics Random Forest: -For evaluating the performance of the Decision Tree model on the test dataset, we utilized two key metrics: the R^2 value and the Mean Squared Error (MSE). These metrics provide insight into how well the model fits the data and how accurate its predictions are.

- **R^2 Value: 0.919**
  - The Random Forest achieved a higher R^2 score of 91.9%, indicating that it explains about 92% of the variance in the data. This is an improvement over the Decision Tree. Hence, Random Forest is better at capturing the complexity of the data.

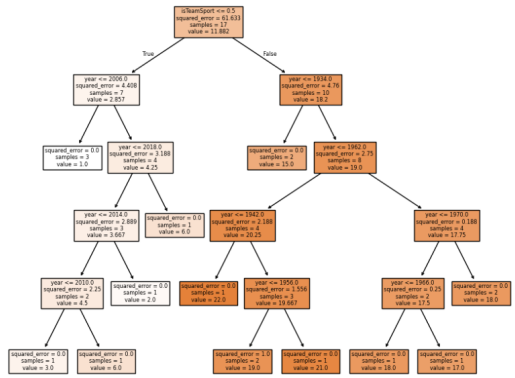
  - A high R^2 value like this typically indicates a good fit, where the model is able to predict the target variable with reasonable accuracy, though there might still be room for further improvement.

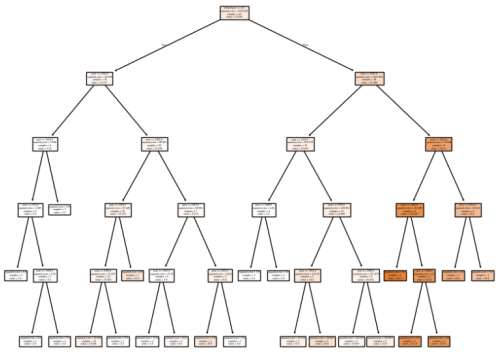
- **Mean Squared Error (MSE): 5.72**
  - The MSE of 5.72 is significantly lower than that of the Decision Tree, indicating more accurate predictions. The Random Forest model is better at generalizing to unseen data, reducing prediction errors compared to the single decision tree.
  - While the MSE is not negligible, it suggests that the model performs reasonably well in terms of minimizing the difference between predicted and actual values.
  - Further improvements can be achieved through techniques such as parameter tuning (e.g., adjusting tree depth, splitting criteria) or using ensemble methods like Random Forest to reduce overfitting and improve predictive accuracy.

Similarly we have done analysis for Australia as well. This can be seen in the python notebook. In general R2 values should be as near to 1.

Insights from Decision Tree:

The visualized decision tree provides insight into how the model splits the data based on the features. It shows that for certain years team sports have a higher medal count in comparison to the medal count from individual sports. I have done it for two countries here, India and australia. I have done two models, decison tree and random forest.





Insights from Random Forest: Random Forest provides a more reliable prediction model by combining the results of multiple trees. The model effectively handles variations in data, likely capturing the relationships between 'year', 'isTeamSport', and 'medal$_count'. While it is less interpretable than the Decision Tree, it is

In comparison, Decision Tree offers a simple and interpretable model, but it can be prone to overfitting and may miss out on more complex relationships. Its R^2 score of 0.881 and MSE of 8.4 reflects a decent performance. Random Forest provides a significant improvement over the Decision Tree, with an R^2 score of 0.919 and a much lower MSE of 5.72. It is more accurate and robust in predicting India's medal count in the Olympics, as it is better at handling complex interactions and generalizing to new data.

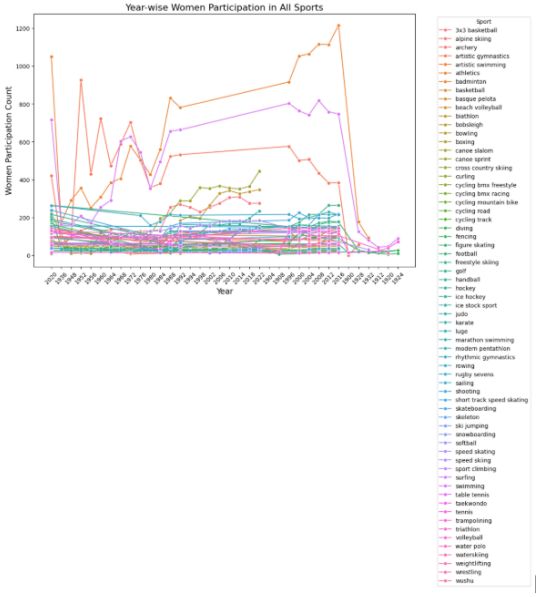
Intelligence Gained : I learnt that in case the paramenetrs are interpretability and simplicity. It is clear that decison treews are way better and useful. But if we want better accuracy Random Forest is the more effective choice. I also understood that my Random Forest model could have been better improved if trhere were more feautures. Hyperparameter tuning etc. can also yeld a better model.

D. Model 4 (Sataakshi)

Question recap :- How has the participation of women athletes in various sports evolved, and what trends can be observed in terms of minimum and maximum participation levels across selected sports?

Model used :- KNN model has been used.

How has the participation of women athletes in various sports evolved, and what trends can be observed in terms of minimum and maximum participation levels across selected sports?



Why KMeans: - The k-NN algorithm is a powerful non-parametric method that works well for regression tasks involving time-based data, such as tracking women's participation across various sports over the years. Since we're predicting participation based on a continuous feature like the year, k-NN is helpful because it makes predictions by averaging the

closest data points. This approach can capture local trends in participation growth or decline.

Training and Tuning: - Since k-NN is based on the distance between data points, feature scaling is crucial. I used the StandardScaler to standardize the 'year' feature, ensuring that the distances were calculated fairly.

Choosing k: The number of neighbors (k) was set to a minimum of 5 or less if there were fewer training samples. This ensures that the algorithm is not underfitting or overfitting.

Model Training: The data was split into training (70%) and testing (30%) sets. This allowed me to evaluate the model's ability to generalize.

Evaluation Metrics: -

Insights: The output indicates the performance of the k-NN regression model for predicting women's participation in four sports:

Results per Sport:

Swimming:

- **R^2:** 0.73 - The model explains 73% of the variance in swimming participation.
- **RMSE:** 125.06 - The model's predictions are off by an average of 125.06 units.

Diving:

- **R^2:** 0.77 - The model explains 77% of the variance in diving participation.
- **RMSE:** 10.88 - The predictions are off by an average of 10.88 units.

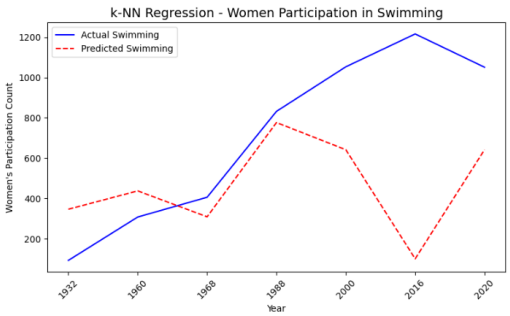
Fencing:

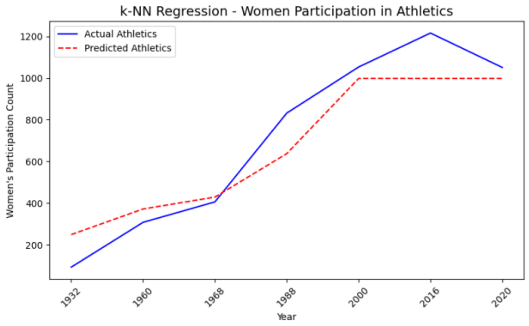
- **R^2:** 0.59 - The model explains 59% of the variance in fencing participation.
- **RMSE:** 32.65 - The predictions are off by an average of 32.65 units.

Athletics:

- **R^2:** 0.89 - The model explains 89% of the variance in athletics participation.
- **RMSE:** 131.02 - The predictions are off by an average of 131.02 units.

Intelligence Gained: We can see that women's participation in most sports is increasing based on our analysis. Our model correctly matches the actual data and follows a similar trend, as seen in one example given below. Using this analysis, specific sports can be targeted for women's participation awareness campaigns, and more attention can be given.





Conclusion k-NN was a strong model for predicting women's participation in sports with consistent trends over time, such as swimming and athletics.

E. Model 5 (Rudraksh)

Question recap :- Table Tennis and Tennis are similar yet different sports. The players I have seen in both games seem to have different builds. The hypothesis is that we can build a model using Height, Weight, and athlete's country to predict which sport they belong to.

Model: - I have used regression, ensemble, and clustering models to see which gives the best output for our hypothesis testing and model building.

Why these models: Ensemble-based models like Gradient Boosting and Xgboost are great for classification, and thus I have used them. I have used a lot of classification models to achieve the best results one can get.

Training and Tuning: - For training, we have taken all of the athletes who play Tennis and Table Tennis. Models are trained separately for men and women.

Evaluation Metrics: - For evaluating our results on the test data, we have used Accuracy, Precision, Recall, F1 score. We have also made Confusion metrics as well. F1 score achieved for Males: 95

```
--------------------Random Forest results----------------------------------
              precision    recall  f1-score   support

table tennis       0.95      0.96      0.95       208
      tennis       0.96      0.95      0.95       214

    accuracy                           0.95       422
   macro avg       0.95      0.95      0.95       422
weighted avg       0.95      0.95      0.95       422

Accuracy: 0.95260663507109
Confusion Matrix: [[199   9]
 [ 11 203]]
F1 Score: 0.9526087642965514
```

```
--------------------Random Forest results----------------------------------
              precision    recall  f1-score   support

table tennis       0.92      0.88      0.90       190
      tennis       0.90      0.94      0.92       223

    accuracy                           0.91       413
   macro avg       0.91      0.91      0.91       413
weighted avg       0.91      0.91      0.91       413

Accuracy: 0.9128329297820823
Confusion Matrix: [[168  22]
 [ 14 209]]
F1 Score: 0.9126636412652459
```

Intelligence Gained: it is indeed possible to see and make a model to predict which game an athlete plays based on their weight, height, and country, telling the hypothesis we started was right and it is possible to use it as a predictor for selecting the appropriate sport.

F. Model 6 (Rudraksh)

Question recap :- In athletics, height, weight, age, and country are major indicators of success in the Olympics. We have made to achieve the same.

Model: - I have used regression, ensemble, and clustering models to see which gives the best output for our hypothesis testing and model building.

Why these models: - Ensemble-based models like Gradient Boosting and XGBoost are excellent for classification tasks, which is why I chose them. These models combine the predictions of several base learners to improve accuracy and reduce overfitting, making them ideal for achieving high-performance results. I have experimented with various classification models to ensure that the best possible results are obtained.

Training and Tuning: - For training, we focused on all athletes who participate in Athletics, ensuring that the model was trained on relevant data for this particular sport.

Evaluation Metrics: - For evaluating the results on the test data, the following metrics were used:

- **Accuracy:** Measures the overall correctness of the model.
- **Precision:** Indicates how many of the predicted positive cases were actually positive.
- **Recall:** Shows how many actual positive cases were correctly identified by the model.
- **F1 Score:** A balanced metric that combines precision and recall into one score.

F1 Score achieved: 83

```
---Gradient Boosting---
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      6577
           1       1.00      0.00      0.00       834

    accuracy                           0.89      7411
   macro avg       0.94      0.50      0.47      7411
weighted avg       0.90      0.89      0.83      7411

Accuracy: 0.8875995142355957
F1 Score: 0.8348809051144537
```

Intelligence Gained: It is possible to build a model that tells athletes' success in athletics.