DIC Phase 2
Report

Geetansh - 50607410
Sataakshi - 50607324
Rudraksh - 50604938

1. Geetansh Question 1 :-

Question recap :- 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?

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

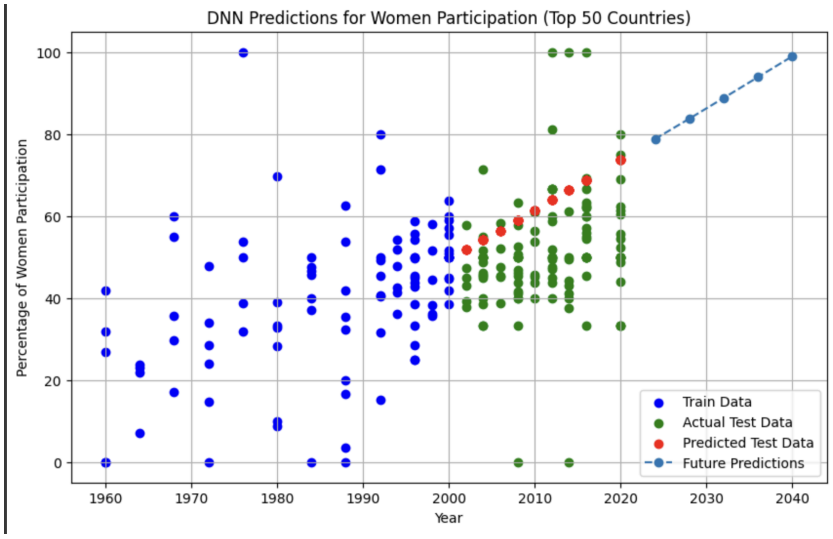
Data Pre-processing : - We have used the EDA steps for this in order to calculate the percentage woman by country for every Olympic year. Then, we have removed data points where the percentage was NA and taken the top 20 countries where this trend was significant in order to make our data more meaningful. We have also normalized the data so that it fits in the DNN model we have defined.

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.

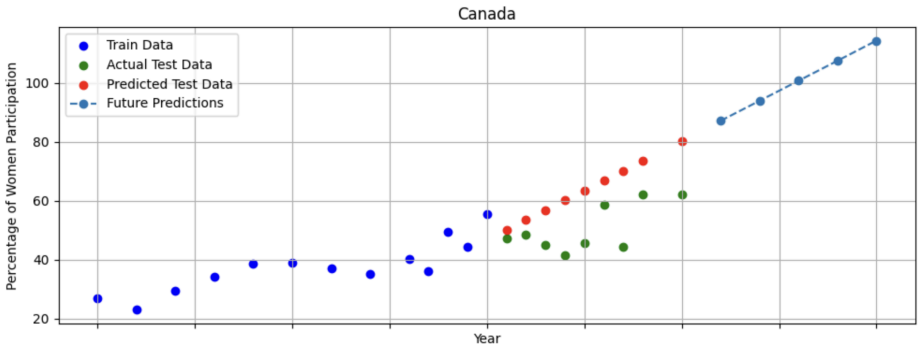


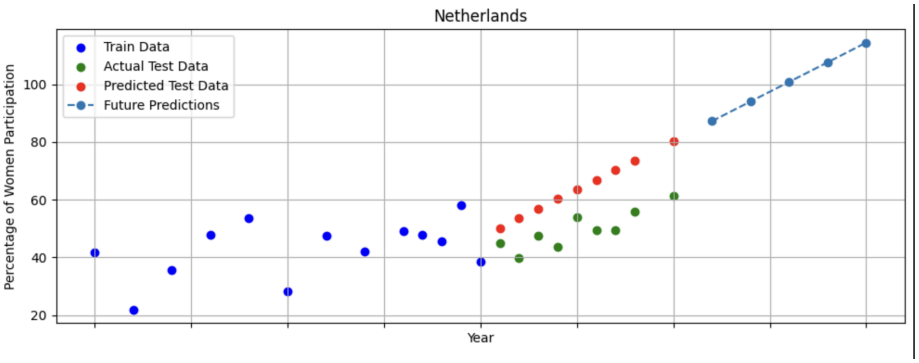
Evaluation Metrics : - For evaluating our results on the test data we have used r2 value and mean total loss.

- 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.

References : -

https://medium.com/@zomev/deep-neural-network-dnn-explained-0f7311a0e869
https://www.tensorflow.org/tutorials/quickstart/beginner

2. Geetansh Question 2 :-

Question recap :- 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?

Model used :- 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.

Data pre-processing : - We first need to group the total participants for each sport. Now we need to define the slope for the trend, which will be one of the deciding factors while performing clustering. For calculating this, we have to use linregress to calculate the slope for the regression line between year and participants. This is the most appropriate way to calculate because slope is usually a linear line and we have only 2 columns to calculate slope. The slope will give us an idea for declining and rising trends. We have also normalized the input.

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 random state so that if we train our model again, we'll be able to produce the same results. We have decided on 2 labels called "Rising" and "Declining".

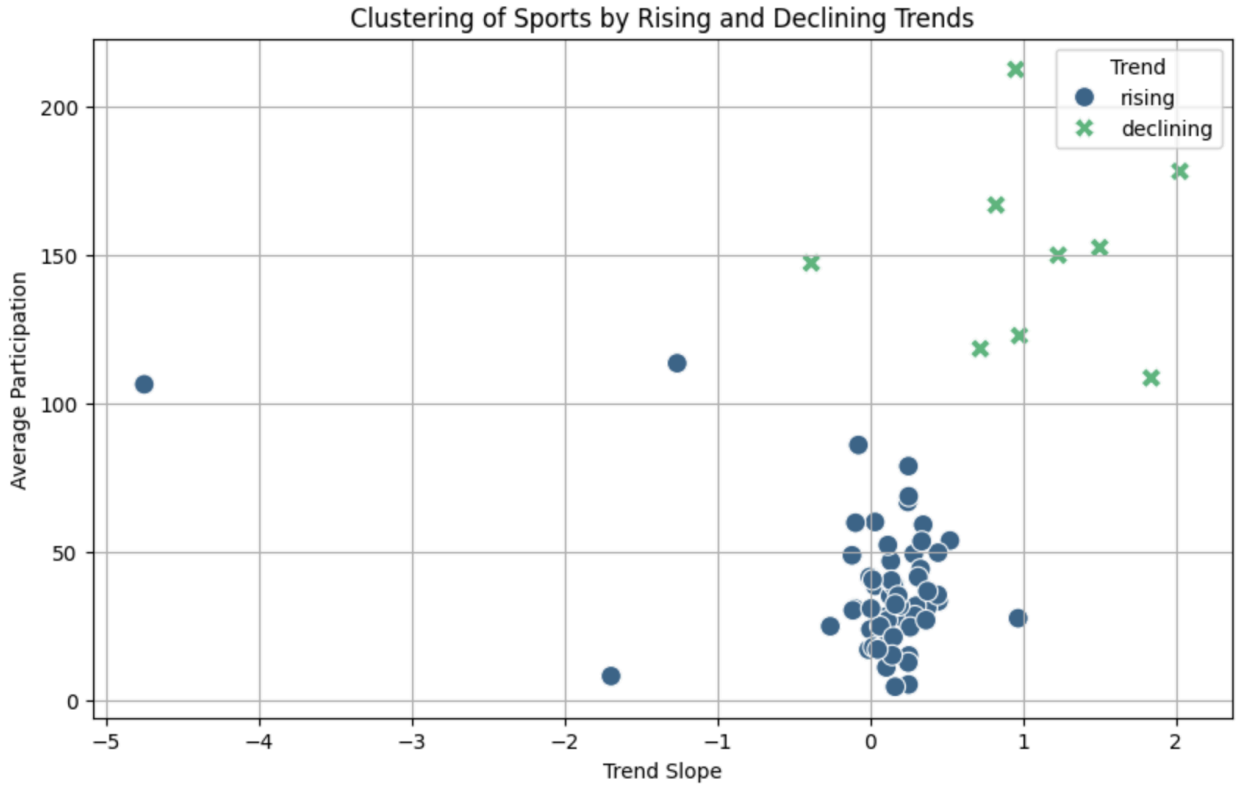
Evaluation Metrics : - For our evaluation, we have kept Silhouette Score, Davies-Bouldin Index, and Inertia.

- 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.

References :

https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.KMeans.html

1. Sataakshi Question 1 :-

Question Recap:- 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?

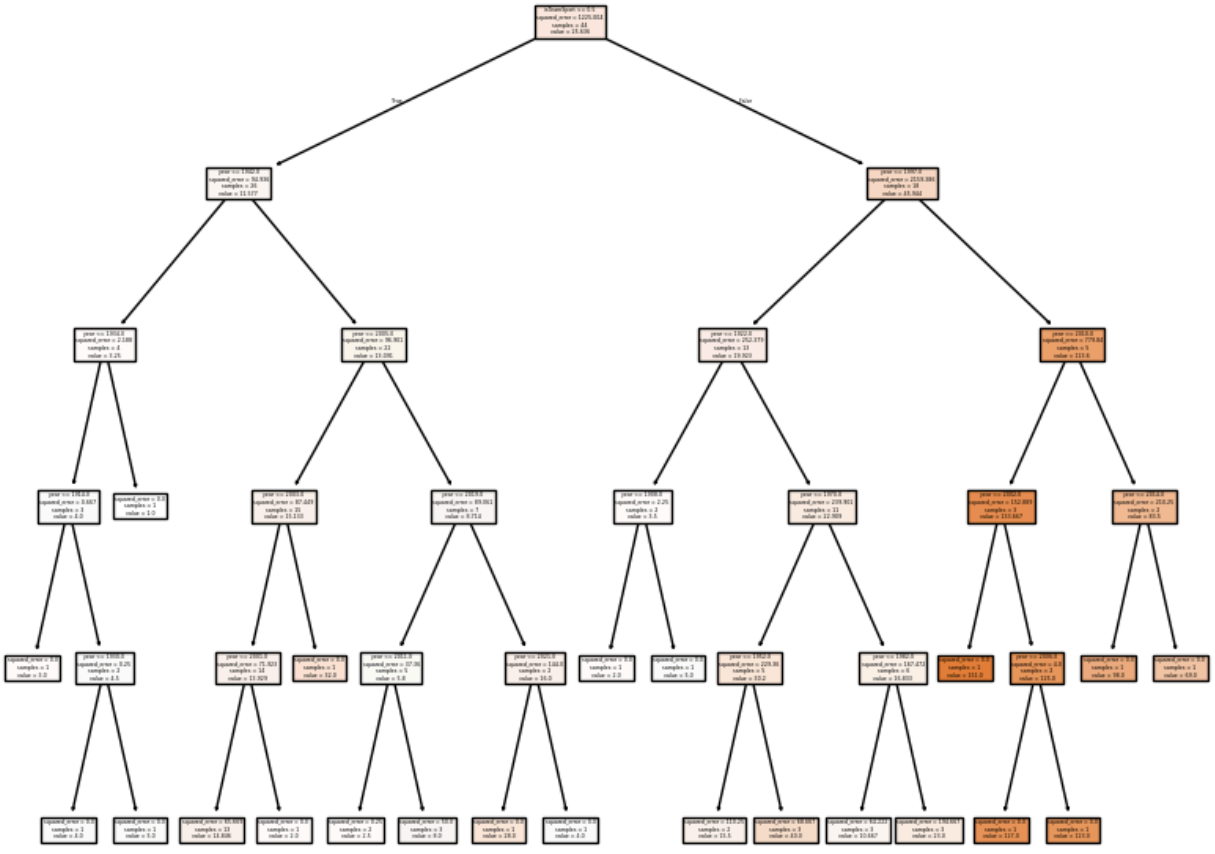
Model: - To solve this we are implementing Decision tree and Random forest model to evaluate the MSE and R^2

Data Pre-processing : - In the phase 1 of the project we cleaned the data by removing any null values. We performed one hot encoding encode the team vs individual sports using binary keys. Ensured that the features like 'year' and 'team sport' were encoded properly. We analysed how each country performs in individual vs in group sports.

Why Decision tree and Random forest: - I tried linear regression first. This didnot seem to work well due to to lack of features. For building a better model we used decision tree and random forest. The Decision Tree algorithm is a choice when dealing with simple relationships between features. In this case, the primary features are 'year' and 'isTeamSport'. We get the knowledge of predicting '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, this will hence solve the problem better.

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 bad model. This has also helped in simplifying the tree. Additionally it has become 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 our results on the test data we have used R2 value and mean squared error.

- R2 value : 0.881
    - The R^2 score indicates that the model explains about 88.1% of the variance in the data. This suggests that the Decision Tree is fairly effective at capturing the underlying patterns.

- Mean Squared Error : 8.4
    - The MSE of 8.4 shows that there is some error in the model's predictions. But it is relatively low considering the complexity of the data. However, the Decision Tree's performance could be improvement with parameter tuning in futute.

Evaluation Metrics Random Forest:

- R2 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.

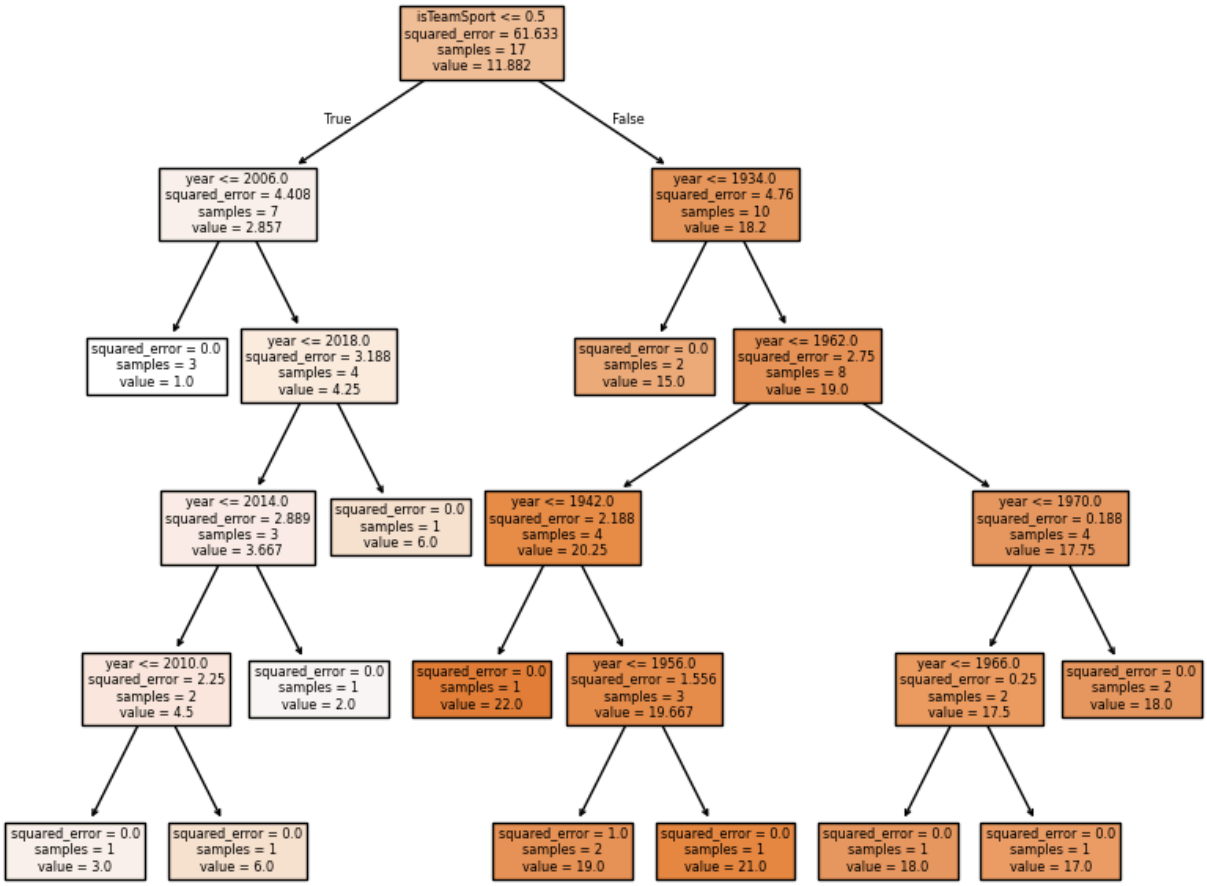
- Mean Squared Error : 5.72
    - The MSE of 5.72 is significantly lower than that of the Decision Tree, indicating more accurate predictions. The Random Forest model is better at generalizing to unseen data, reducing prediction errors compared to the single decision tree.

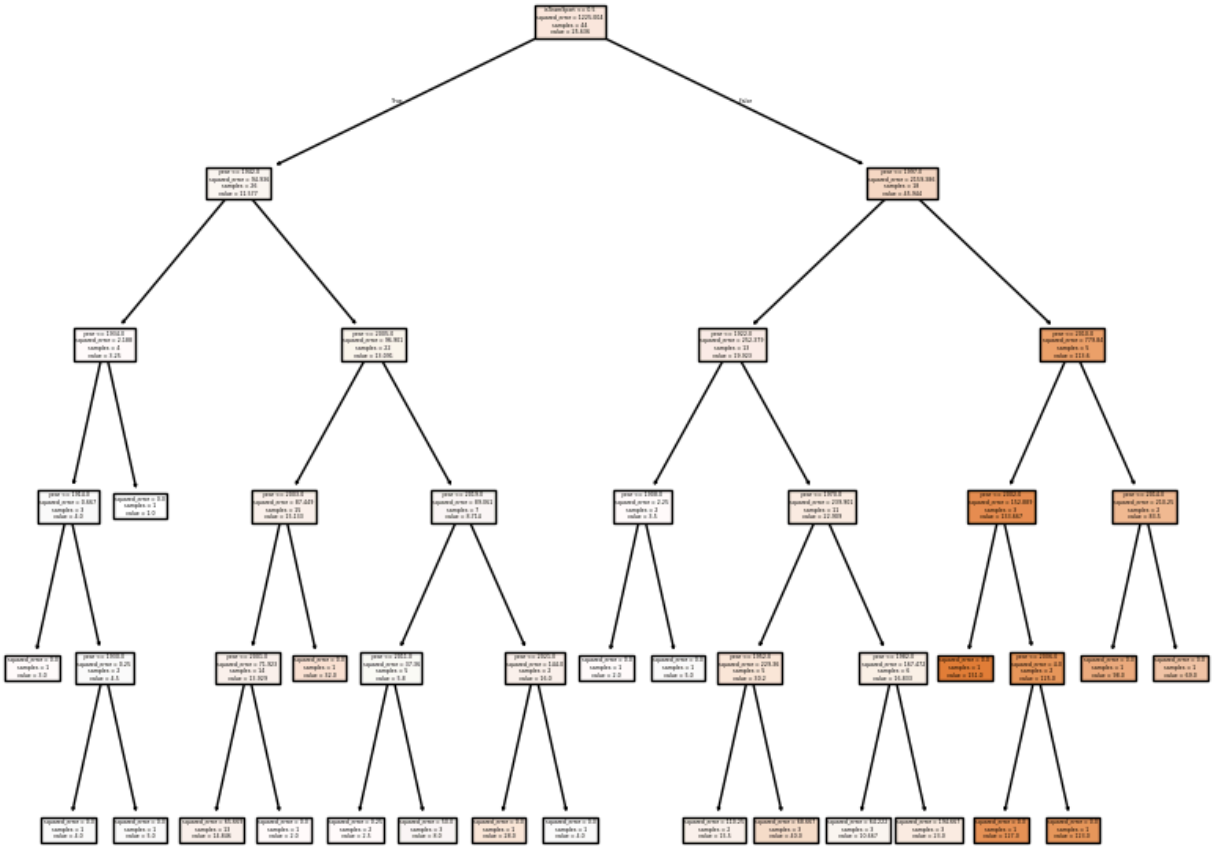
Similarly we have done analysis for Australia as well. This can be seen in the python notebook. In general R2 values should be as near to 1.

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 more robust and offers higher accuracy.

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.

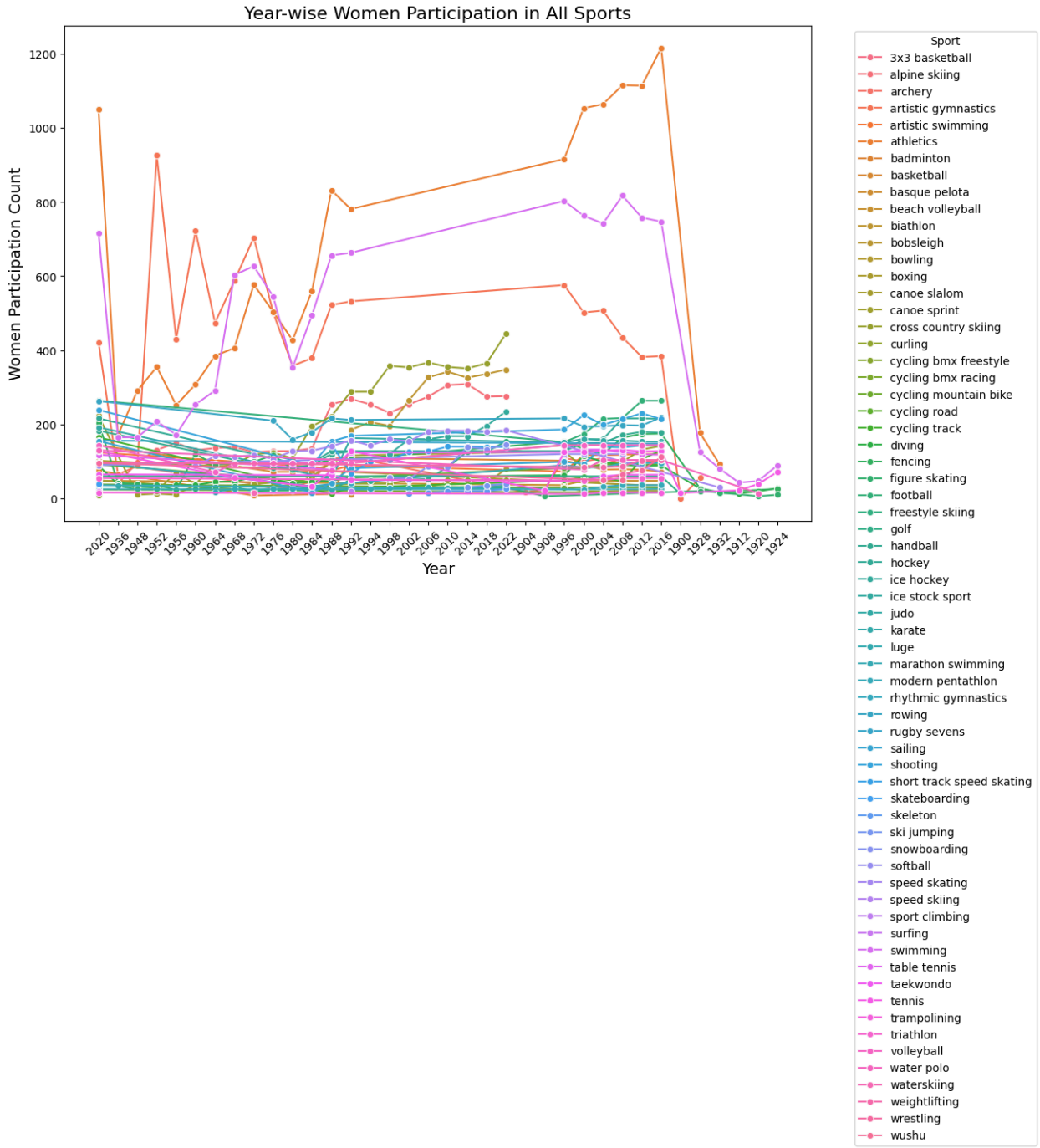
Intelligence Gained : I learnt that in case the paramenetrs are interpretability and simplicity. It is clear that decison treews are way better and useful. But if we want better accuracy Random Forest is the more effective choice.
I also understood that my Random Forest model could have been better improved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

2. Sataakshi Question 2 :-

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.

Data pre-processing : - We queried to get the participation of women in differemt sports over the olympic history.



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. Given that we're predicting participation based on a continuous feature like the year, k-NN is useful

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 verfitting;
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.

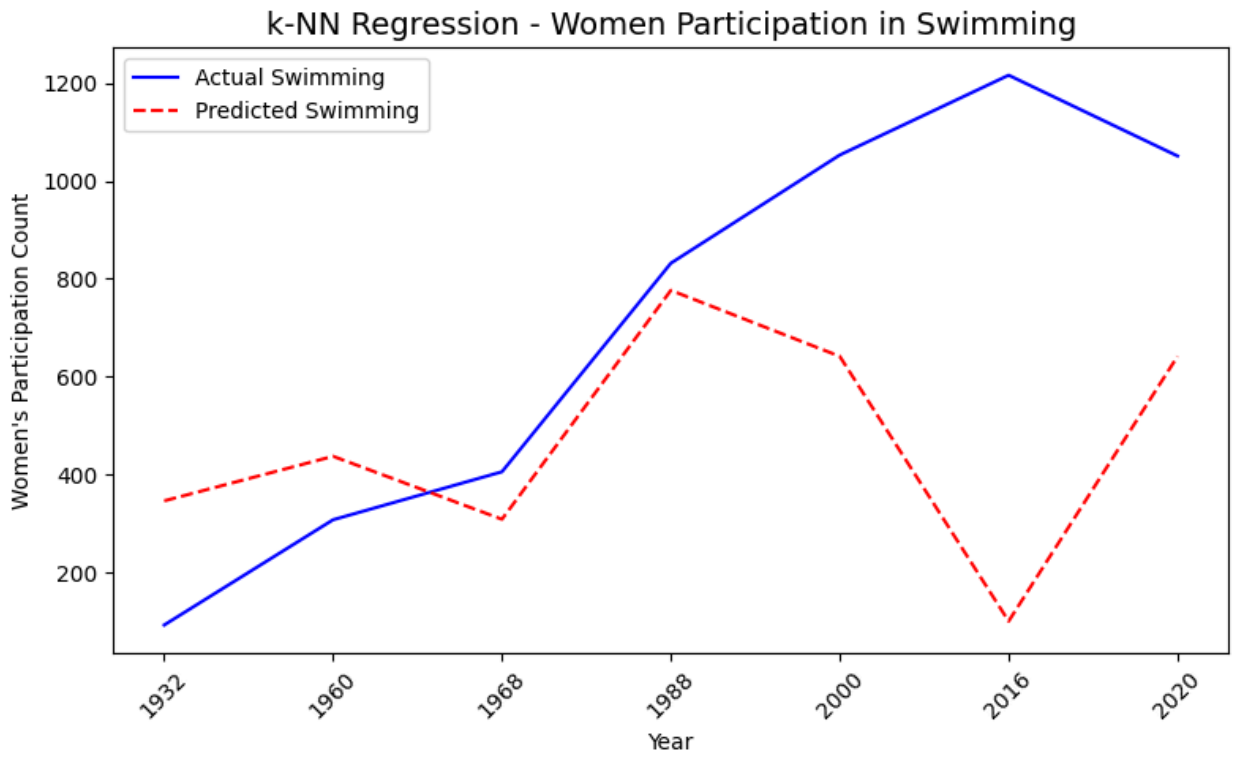
RMSE: 32.65 - The predictions are off by an average of 32.65 units.

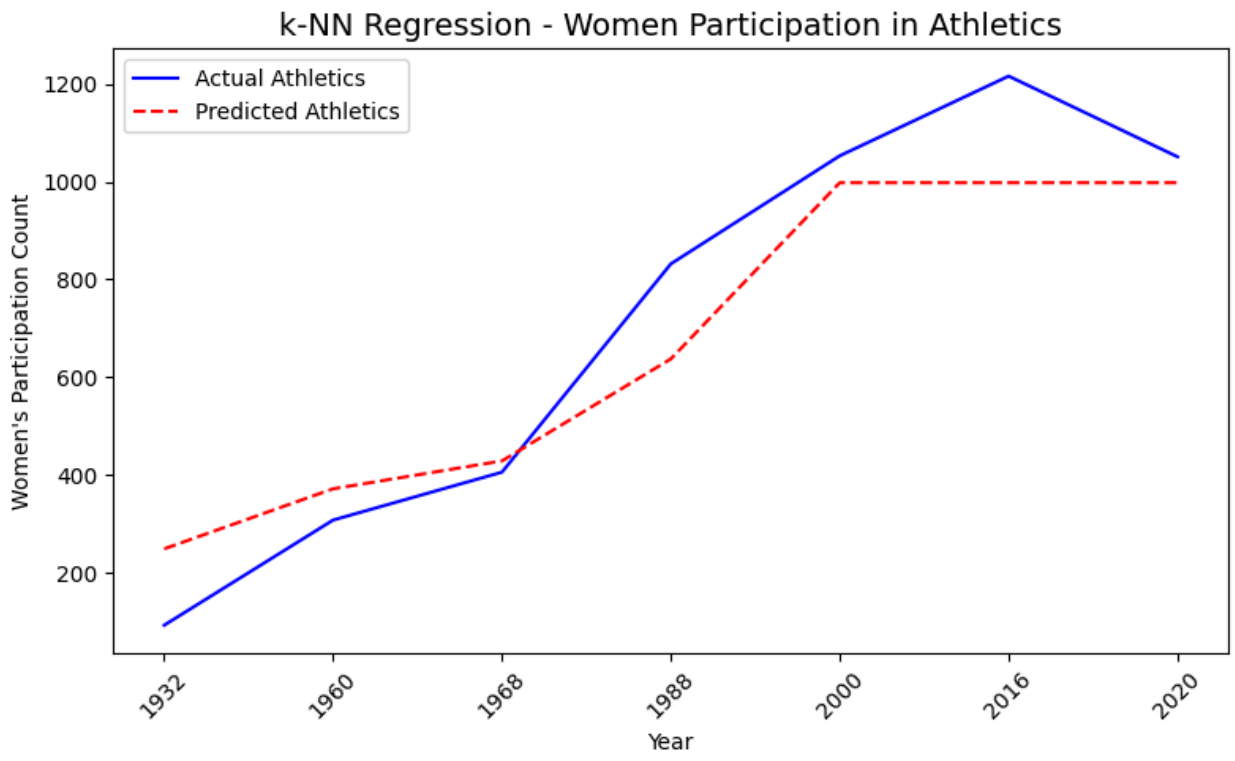
Athletics:-

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 the women participation in most sports is increasing based on our analysis. Our model correctly matches the actual data and follows a similar line trend as we can see in one example given below. Using this analysis, specific sports can be targeted for women 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.

Rudraksh Question 1:-

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.

Data Pre-processing: - We are using only Summer Olympics Data. Men, Women data are being handled separately.

Why these models: Ensemble-based models like Gradient Boosting, Xgboost are great when it comes to classification and thus I have used them. I have used a lot of classification models in order 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
```

F1 score achieved for Females: 91

```
--------------------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.

Rudraksh Question 2:-

Question recap:- In athletics, height, weight, age, and country are major indicators for success in 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.

Data Pre-processing: - We are using only Summer Olympics Data. The medal column is made into a medal or no medal.

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

Training and Tuning: - For training, we have taken all of the athletes who play Athletics.

Evaluation Metrics: - For evaluating our results on the test data, we have used Accuracy, Precision, Recall, F1 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 indeed possible to build a model which tells the success of athletes in athletics.