Presence of women in sports - A step towards Equity

Tejaswini Satish Kumar
tsatishk@purdue.edu
Purdue University
West Lafayette, Indiana, U.S.A

Sahithi Kodali
kodali1@purdue.edu
Purdue University
West Lafayette, Indiana, U.S.A

Sanskriti Motwani
smotwani@purdue.edu
Purdue University
West Lafayette, Indiana, U.S.A

ABSTRACT

Female empowerment is a crucial aspect that needs to be addressed in various walks of life across the community. In this project, we analyze 120 years' worth of Olympics sports data to compare gender-based performance and participation in various sports categories. Further, we also predicted the sports suitable based on physical factors and predicted the chances of winning medals based on affecting factors in the dataset focusing on females. Later, a thorough analysis is conducted on the game of Tennis to understand the components that female athletes can improve upon to better their chances of winning.

1 PROJECT TOPIC AND BACKGROUND

The topic chosen for this project is 'Presence of women in sports - A step towards Equity', which involves gaining an insight into the participation and performance of women in sports. We assess their participation in Olympics based on physical attributes, nations, and sport type, along with analysing the skills required to excel in a particular sport (Tennis was chosen for this project scope). These tasks are performed by analyzing the historic data of 120 years of Olympic games [6], and the data showcasing skills required to excel in the Tennis game.

1.1 Importance and Motivation

Based on our common experiences and digging into the literature reviews, we realized that women are generally underrepresented in research and clinical studies. An eye-opening example was seen during the clinical studies related to covid-19 [2], where most studies failed to include gender as an analytical variable.

As a group of three women in the field of STEM, we found it necessary to contribute towards studies/research that is inclusive of the female gender. Despite having many powerhouses like Serena Williams, PV Sindhu, Simone Biles and Ronda Rousey in the sports field, the research in understanding the mechanics of women in sports is sparse [8].

Gaining an insight into what makes such athletic women great in their respective fields would definitely be beneficial for the women dreaming of thriving and mark their name in sports. Furthermore, predictive analysis to recognize women's performance in future Olympics or tournaments could be helpful for sports trainers/trainees while contributing to the community's growth.

2 ANTICIPATED PROJECT GOALS AND EVALUATION OUTCOMES

The anticipated goals of our project involve finding answers to the following questions.

- How might the participation of women in sports change in the upcoming years?
- Is there a relationship between the physical attributes (height, weight, age) and the performance of an athlete in a specific sport category?
- Are there any environmental factors that affect women's performance?
- How and when can women start holding more world records? (Currently, men hold most of the world records in sports)
- Which sport is most suitable for women and likely to attain more world records?

Our project goals are designed according to the outcomes we want to obtain. The success of these goals can be checked using the evaluation metrics and by reflecting on the following questions.

- Did we contribute something unique to the women in the sports field?
- Is there a way our research can be employed to further the careers of women in sports?
- Were we able to answer these questions with a certain degree of confidence?
- How can the outcomes be financially beneficial to women, governments and other organizations? (Cost vs gain analysis)

Further, we would also like to address how nutrition and required skills for a sport are related to women's performance, and what change in nutrition should be observed to achieve the best performance in a sport which is considered part of future work.

3 LITERATURE REVIEW

The research in the sports analytics field has gained a lot of traction over the years spawning many startups and independent contractors. Data has been collected in sports as "statistics" over centuries and has accumulated into an oracle that has helped to provide significant insights into players and team dynamics. A vast majority of papers focus on analyzing team sports such as basketball [7], baseball [3], football [5], and more importantly the analytics focus on improving male-dominated sports leagues. The National Basketball Association (NBA), National Football League(NFL), and Major League Baseball (MLB) have maximized the player performances resulting in numerous wins and as a result, far greater profits. Employing these existing studies as roadmaps, this paper probes the game of Tennis based on the singles games hosted by the Women's Tennis Association. By building a predictive model, the paper addresses a classification problem that anticipates a winner based on individual and head-to-head statistics. Utilizing the knowledge learned by the model, the most important statistics can be inferred. A player can then polish up skills that contribute to strengthening those statistics, thus improving their tennis game.

4 DATA AGGREGATION AND ANALYSIS

4.1 Description of datasets

- The dataset chosen for visualising, analysing and predicting the performance of female performance to males performance is "120 years of sports data in Olympics games" from Kaggle [1]. This data involves details of male and female athletes' physical aspects, game details and medals won. A snippet of the data is shown in figure 1 for reference.

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	0	China	0
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	0	China	0
2	3	Gunnar Nielsen Aaby	M	24.0	0.0	0.0	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	0	Denmark	0
3	4	Edgar Lindenau Aabye	M	34.0	0.0	0.0	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Denmark	0
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	0	Netherlands	0

Figure 1: Snippet of 120 years of Olympic games data

- The Women's Tennis Association has been collecting match statistics since the year 1920, which documents the tournament, year, type of court surface, players' physical attributes, nationalities, ages, and rankings. As part of statistics, WTA also recorded the number of aces, double faults, serve points, first serves made, first-serve points won, second-serve points won, serve games, break points saved and breakpoints faced for each player in a head-to-head match. For the purposes of this paper, the dataset is limited to the years ranging from 2000 to 2022 due to changes in the format of the games.[1]

4.2 Data Analysis and Observation

4.2.1 Olympics Data.

The goal of the project is to attain gender equity focused on female empowerment. To get to the goals we want to address, the data has been visualised on various metrics and categories. In the line graph showcasing the number of female to male participants ranging yearly as in figure 2, we can observe the female participants to be less than a fourth of male participants until the 1990s and a half or less than half after the 1990s. Though this indicates some improvement, we can clearly observe the need to encourage female athletes. The visualisation in figure 3 shows the participant's height and weight ratio based on gender, in which we can observe the weight-to-height ratio as linearly dependent except for a few outliers. From this observation, we can check if these outliers were in any way beneficial for the success of athletes in any sport category.

In figure 4, we can observe the type of medals secured by men and how they vary based on sport category. Similarly, figure 5 shows the same observation for female participants. We can make an interesting observation that a few sports were dominant in the female category for the number of participants, swimming, while some, like wrestling, have huge male domination, which can be due to physical factors and strength based on genders that are generally known to us. It would be interesting to analyse how we can improve females in sports with little success for participants. We can also observe a few sports that have similar success to the participants

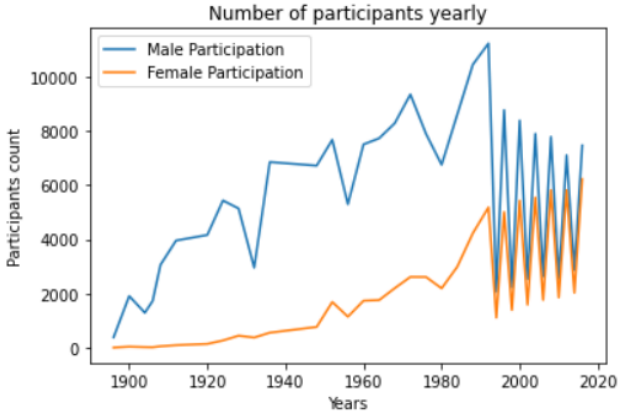

Figure 2: Participants vs year of participation based on gender

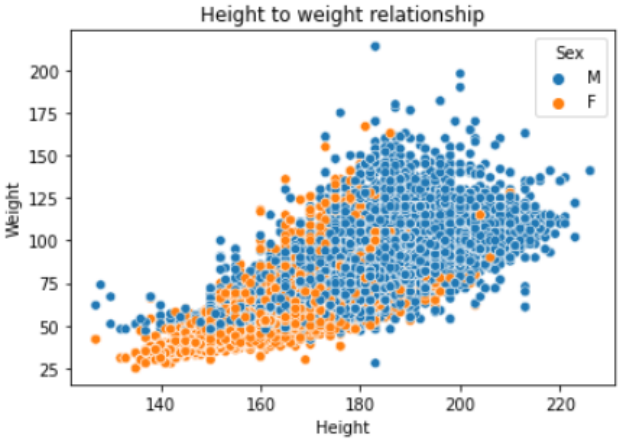

Figure 3: Height and weight relationship based on gender

in both genders, like Tennis, where we can analyse how to improve female success as we did in this project.

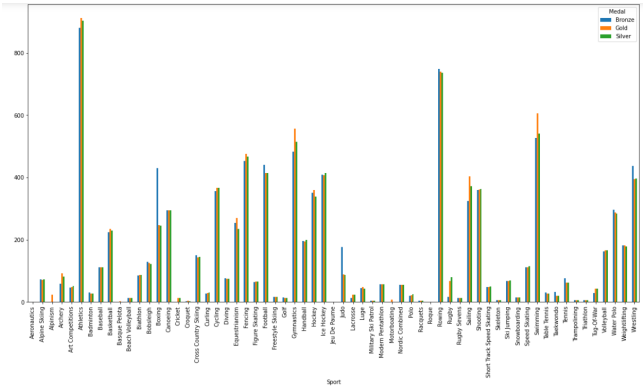

Figure 4: Medals won by men in each sport category

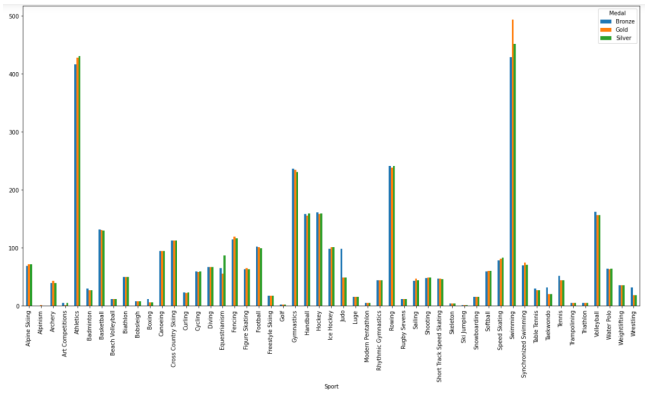

Figure 5: Medals won by women in each sport category

Further, we made a few visualisations focused on females and the factors affecting their number in sports. In figure 6, we can observe that there are very few sports like Athletics, swimming and gymnastics that have a good number of female participants. However, in figure 7, we can observe that only half of these participants are unique, i.e., not the same participant in multiple events in a sport each year, showing us that the number of unique females in the sports category is actually less.

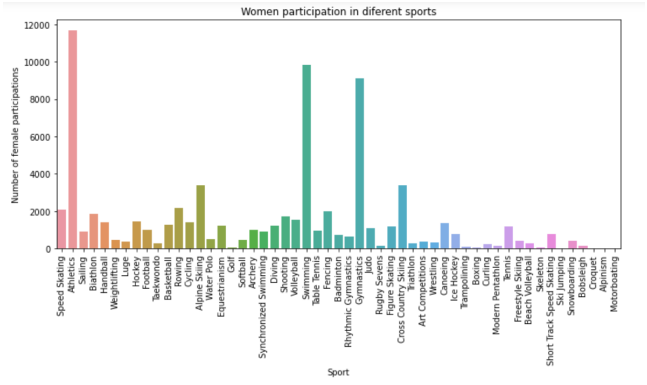

Figure 6: Female participants based on sport

Restricting these unique female participants, figure 8 shows that the age group of these participants primarily lies between 14-38.

Figure 9 shows the female count per year in all the sports, which is a positive indication of growth. Still, figure 10 shows the majority of sports events with female participants to be of less promising growth apart from two to three sports categories.

5 IMPLEMENTATION AND RESULTS

Observing the effects of various factors on the participation and performance of women in sports, we used existing models to predict the sport's suitability based on the physical factors and a prediction of winning medals based on physical, socio and game factors. Further, we picked the sport Tennis to understand and predict how to improve an athlete's game to win.

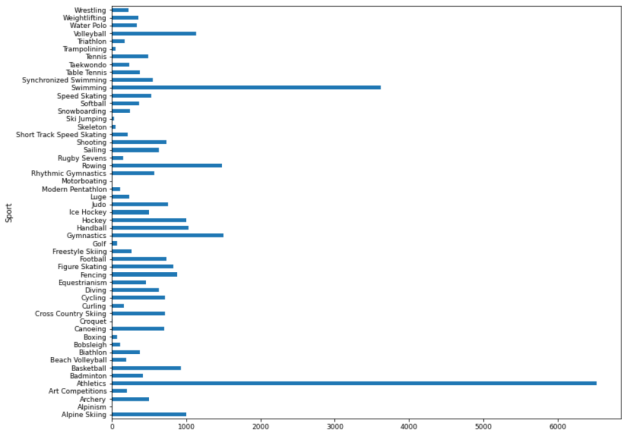

Figure 7: Unique female participants based on sport

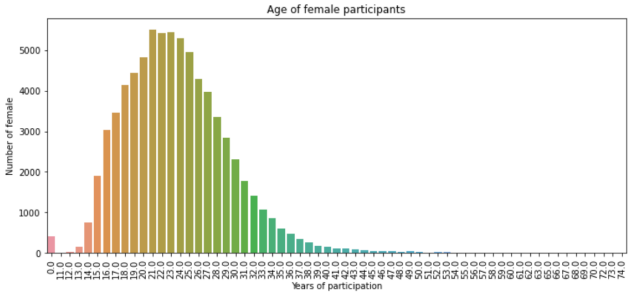

Figure 8: Age of the female participants

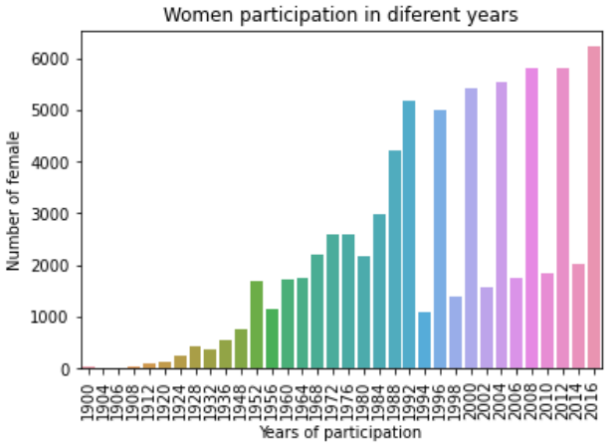

Figure 9: Female participants based on year

5.1 Olympic Dataset

5.1.1 Top 5 Sports Based on Physical Factors. To suggest sports that would be suitable based on age, weight, and height, we used a Decision tree model and a Neural Network based model separately to compare their performance.

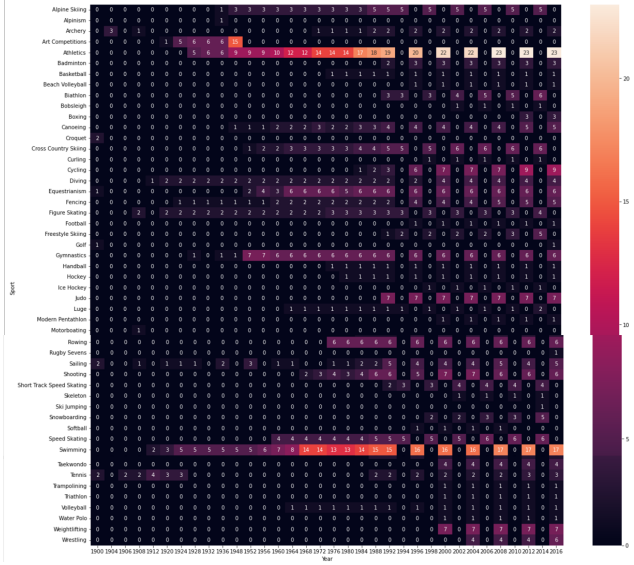

Figure 10: Number of events women participated in a sport per year

- Using Decision Tree:
  Considering the physical factors and the summer season games, a decision tree was used initially to predict the sport suitably. The decision tree makes an informed decision by analysing the odds of each output based on a feature and reading through a few articles similar to such predictions, and hence the existing model has been picked for implementation. The features have been cleaned and scaled for the sake of normalising, and the DecisonTreeClassifier of the sklearn library is used to train the model with physical features for a sport. As in figure 11, the training accuracy obtained is about 61% and can be visualised based on each of the 35 sports categories as labels. However, when the testing predictions are observed in figure 12, the model attained a 25% accuracy, only showing that the model overfits. This is due to the noise by the sports that did not have any players in a type of sport and also can be due to outliers.
- Using Neural Networks:
  For the same settings, a densely connected neural network with relu activation, except for the last output layer with softmax function to output the probabilities, is used. The model's architecture can be visualised in figure 13. The adam optimiser and the spare_categorical_crossentropy loss function are used in the model.
  The training and testing accuracy obtained by both models is 34%, as shown in figures 14 and 15, and we can see that the model learnt well with existing data. However, there are sports with no corresponding participants, which made the model prediction metrics less accurate than what is expected to be. When observing the individual sports categories, we can see that a few sports categories' predictions were accurate by more than 50%.
- Comparing the two models:

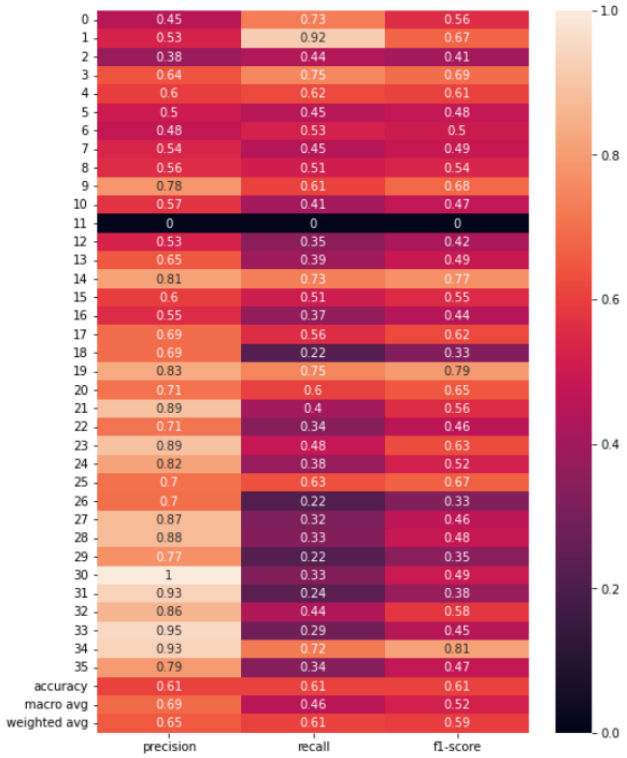

Figure 11: Training performance using Decision trees

Comparing the performance and predictions made by the two models, we can say that the neural network indeed performed better and can still be improved further. An own data sample data has been given to the model, and the prediction made by the models for given physical factors, and sports can be compared as shown in figures 16 and 17.

5.1.2 Predict the chances of winning a medal.

With the goal of predicting if a female athlete can win a medal or not, we utilized all the factors affecting the player's performance including physical, location and the game criteria to build models that can predict the chances of winning a medal.

- Using Random Forest
  The existing model of the Random Forest classifier from sklearn library is used to make the predictions on the scope of winning a medal. Random Forest though does not provide much interpretability about the data, since we already know what factors might affect athlete performance we proceeded to use this model for its beneficial outcomes based on a few papers we studied. This model obtained a very good accuracy of 93% in predicting the chances of winning a medal as shown in figure 18.

For the following algorithms, the dataset was modified as : Multiple columns had null values, and so removing features based on this would lead to skewed data. The dataset was modified such that null values in each column were replaced by the mode of that feature

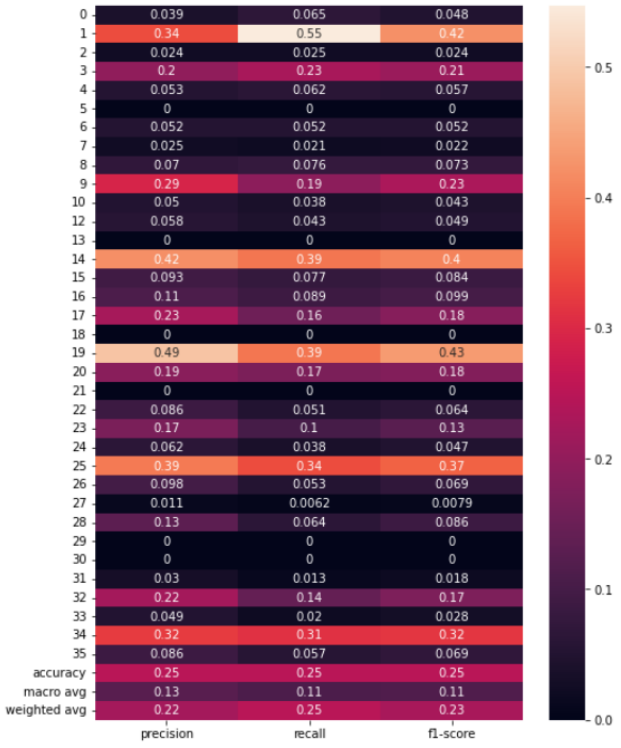

Figure 12: Training performance using Decision trees

```
model = tf.keras.models.Sequential([
    tf.keras.layers.Dense(4),
    tf.keras.layers.Dense(10, activation = tf.nn.relu),
    tf.keras.layers.Dense(20, activation = tf.nn.relu),
    tf.keras.layers.Dense(40, activation = tf.nn.relu),
    tf.keras.layers.Dense(len(unique_sports), activation = tf.nn.softmax)
])

model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

Figure 13: Model architecture

vector. Since the label here was the Medals column, removing rows was not ideal as some of the winners also had null values,

The next task was to remove features that were irrelevant for example ID. Some features were even duplicate of another vector, for country name and NOC relayed similar information but the NOC vector was clearer.

- KNN - was used as it can achieve high accuracy in a variety of prediction models. Since the label was medals, most rows in the training data had the label 0. Thus, using KNN was suitable as it doesn't rely too much on training data.
- Logistic Regression - is applied to predict the categorical dependent variable, i.e., it's used when the prediction is categorical, for example, yes or no, true or false, 0 or 1. The label that must be predicted here is winning a medal vs not winning a medal, thus Logistic Regression is another algorithm that is suitable.
- Naive Bayes - is good for making real-time predictions as it works fast. It also performs well in the case of categorical input variables compared to numerical variable(s).

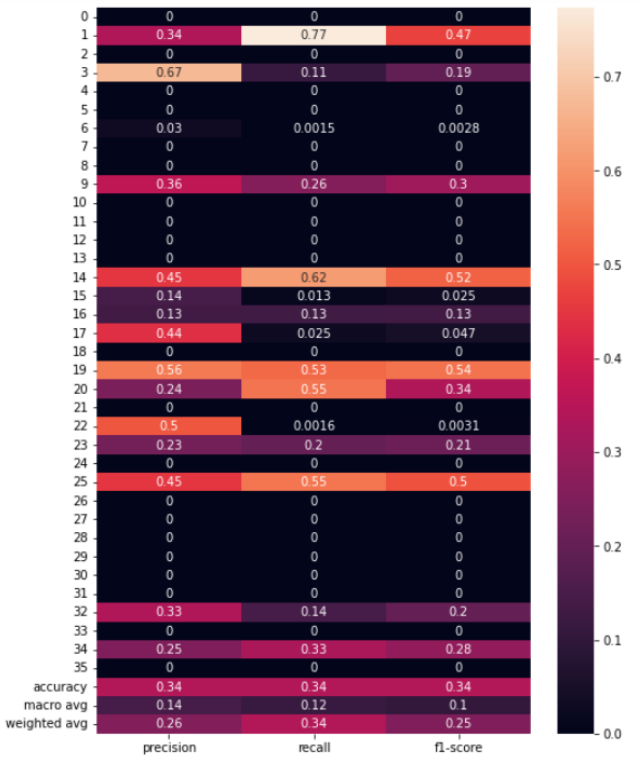

Figure 14: Training performance using Neural Network model

5.1.3 *Comparing the Outcomes.*

We can observe that the Random Forest predicted with great accuracy in predicting the chances of winning a medal compared to other classifiers we tried above, showing its efficiency in spite of not proviidng the interpretability capacity for the user.

5.1.4 *Understanding which Feature is Most Important.*

There are multiple factors that affect participation and performance From our dataset we can see that the feature vector "country" has the highest coefficiency.

From [4] it can be understood that Higher the women's empowerment in the public sphere relates to higher participation of countries in international women's athletics.

5.2 Tennis Dataset

The dataset had column labels which could lead to potential bias. Hence, the data was reorganized and relabelled to remove the usage of "winner" and "loser". The column headers were changed to address the players as "player1" and "player2". An extra column was added to indicate the winner of the match. The entire dataset was divided into 85% train and 15% test samples. An ordinal encoder is used to encode the categorical features as arrays of integers.

- The first model is a simple decision tree classifier without any depth constraints. This model was chosen for its speed and simplicity, allowing us to have a basic overview of the influence of different statistics. The resultant model has an

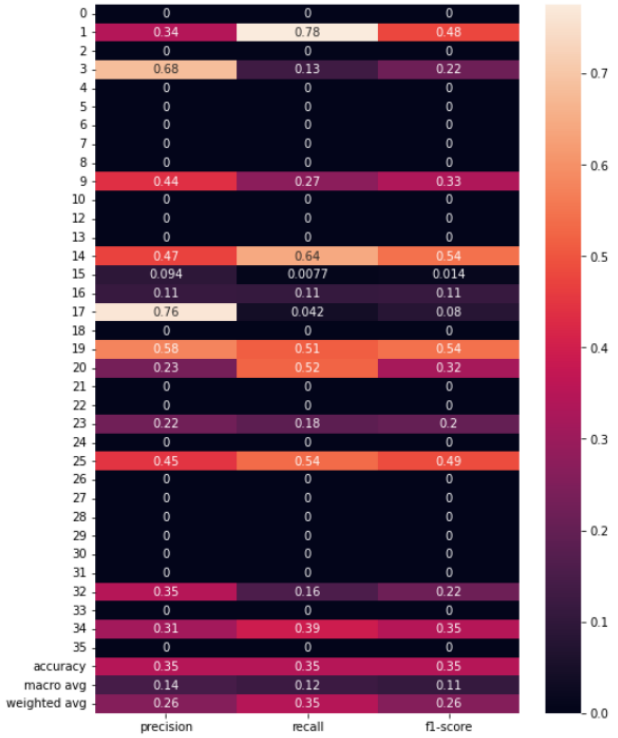

Figure 15: Testing performance using Neural Network model

```
For a 18 year old female 150 cm and 110 kg:
[DT] Suggested sport is: Golf
1/1 [==============================] - 0s 19ms/step
[NN] Suggested sports are: Golf (90%), Swimming (4%), Rugby Sevens (2%), Weightlifting (1%), Triathlon (0%), Sailing (0%), Figure Skating (0%), Cycling (0%), Gymnastics (0%), Rhythmic Gymnastics (0%)


For a 18 year old female 175 cm and 50 kg:
[DT] Suggested sport is: Fencing
1/1 [==============================] - 0s 22ms/step
[NN] Suggested sports are: Fencing (76%), Weightlifting (10%), Canoeing (4%), Wrestling (4%), Diving (0%), Modern Pentathlon (0%), Boxing (0%), Water Polo (0%), Rugby Sevens (0%), Badminton (0%)


For a 18 year old female 175 cm and 80 kg:
[DT] Suggested sport is: Weightlifting
1/1 [==============================] - 0s 19ms/step
[NN] Suggested sports are: Weightlifting (24%), Rhythmic Gymnastics (14%), Swimming (9%), Wrestling (8%), Shooting (5%), Tennis (5%), Triathlon (5%), Rowing (3%), Sailing (3%), Badminton (2%)
```

Figure 16: Results obtained for the own sample data

```
For a 18 year old female 175 cm and 110 kg:
[DT] Suggested sport is: Swimming
1/1 [==============================] - 0s 19ms/step
[NN] Suggested sports are: Golf (36%), Weightlifting (29%), Swimming (28%), Triathlon (1%), Rugby Sevens (1%), Sailing (0%), Rhythmic Gymnastics (0%), Figure Skating (0%), Tennis (0%), Shooting (0%)


For a 18 year old female 200 cm and 50 kg:
[DT] Suggested sport is: Fencing
1/1 [==============================] - 0s 18ms/step
[NN] Suggested sports are: Fencing (46%), Weightlifting (44%), Wrestling (3%), Badminton (2%), Canoeing (0%), Athletics (0%), Modern Pentathlon (0%), Boxing (0%), Handball (0%), Rhythmic Gymnastics (0%)


For a 18 year old female 200 cm and 80 kg:
[DT] Suggested sport is: Athletics
1/1 [==============================] - 0s 19ms/step
[NN] Suggested sports are: Badminton (39%), Athletics (35%), Wrestling (10%), Rhythmic Gymnastics (8%), Weightlifting (2%), Tennis (1%), Modern Pentathlon (0%), Shooting (0%), Boxing (0%), Taekwondo (0%)
```

Figure 17: Results obtained for the own sample data (continued)

accuracy of 84.2%, with a maximum depth equal to 22. Since overfitting is a classic disadvantage of a decision tree, we decided to visualize the train and test AUC scores with the maximum depth ranging from 1 to 32 as seen in Figure 19. It is clear from Figure 19 that the tree is overfitting the data,

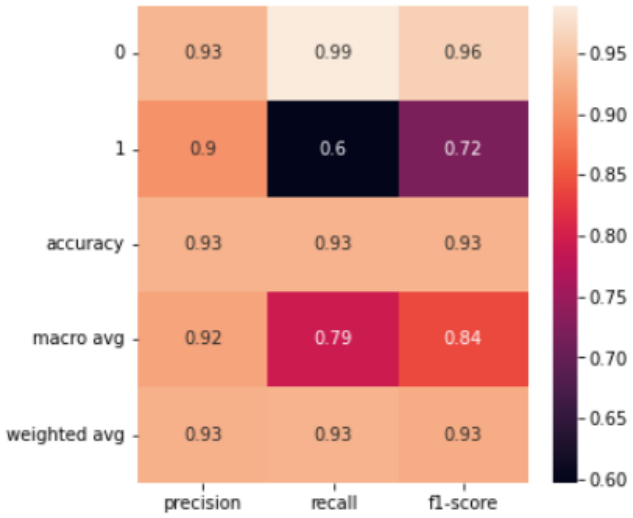

Figure 18: Accuracy attained using Random Forest

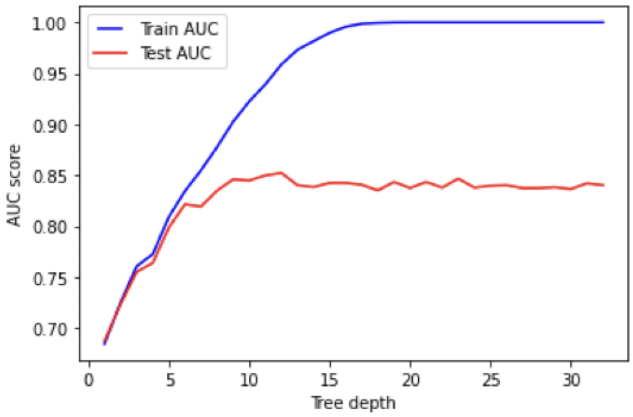

Figure 19: AUC Curve for Decision Tree Classifier

so the max depth of the tree was set to 12. The accuracy of the new tree is 84.8%, which is only a 0.6% increase from the previous tree. While there is still scope for improvement in accuracy, visualizing the tree and feature importance provided some insights into the decision-making. Figure **??** shows that the break points faced and the break points saved are the most influential features for the outcome of a game. To provide a better idea to those who are not Tennis enthusiasts, think of a breakpoint as when a player receiving a service game reaches the stage of being one point away from winning that game.[9] The non-statistical features such as player height, playing hand, game round, and surface do not play a big role in determining the outcome of a game. In future models, we focus on the statistical features to understand which of those a player can improve.

- As a step up to a decision tree classifier, a random forest classifier of 100 estimators was trained on the same input

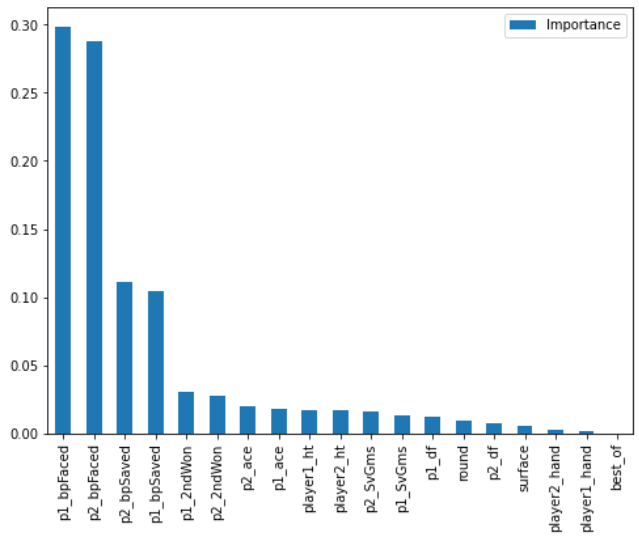

Figure 20: Feature Importances for Decision Tree Classifier

features. It produced an accuracy of 88.1% which is a significant increase from the previous model. By randomly shuffling the feature values, the change in the model's score is monitored to produce the permutation importances. Table 1 displays the importances as weights for each feature. The positive weights for break points faced, break points saved, ace percentage, and so on indicate that they are heavily responsible for swaying the output. The lower rows of the Table 1 contain negative weights for features indicating that those features do not have considerable influence on the model. Another interesting insight is that both the decision tree classifier and random forest classifier have similar feature importances with break points faced at the top. After visualizing the feature importances, another random forest classifier with the same number of estimators was built but the training set only included features whose weights were greater than 0.01. This provided an accuracy of 91.2% which is the best score seen so far.

- Lastly, an XGBoost classifier was built on the training set, which resulted in the best accuracy of the three models - 96%. To deduce the potential improvements that a player can make, further analysis is conducted by plotting partial dependence plots. Figures 21 and 22 show the partial dependence plots for the breakpoints faced and breakpoints saved respectively by player 1.
  The figures portray a sort of inverse relation between saving and facing break points. As the number of breakpoints saved increases up to approximately 7, the chances of winning also increase. However, beyond 9 saves, this feature loses importance in being a deciding factor towards the outcome. Similarly, the number of breakpoints a player faces affects a negative outcome or a loss until it reaches about 14. Beyond 14, it has very little effect on the outcome. Shapely values

Table 1: Permutation Importances for Random Forest Classifier

Weight	Feature
0.2525 ± 0.0128	p1_bpFaced
0.2523 ± 0.0159	p2_bpFaced
0.0641 ± 0.0018	p2_bpSaved
0.0552 ± 0.0102	p1_bpSaved
0.0149 ± 0.0085	p1_2ndWon
0.0108 ± 0.0067	p2_2ndWon
0.0073 ± 0.0064	p2_ace
0.0044 ± 0.0053	p1_ace
0.0039 ± 0.0035	p2_df
0.0038 ± 0.0055	p1_SvGms
0.0023 ± 0.0046	player1_ht
0.0012 ± 0.0046	p1_df
0.0009 ± 0.0011	player2_hand
0.0002 ± 0.0020	surface
0.0002 ± 0.0036	player2_ht
-0.0004 ± 0.0025	player1_hand
-0.0029 ± 0.0046	p2_SvGms

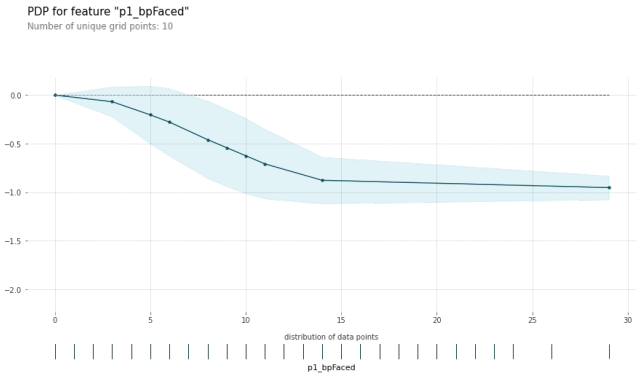

Figure 21: Feature Importances for Decision Tree Classifier

help in putting an individual instance under a microscope as seen in Figure 23.

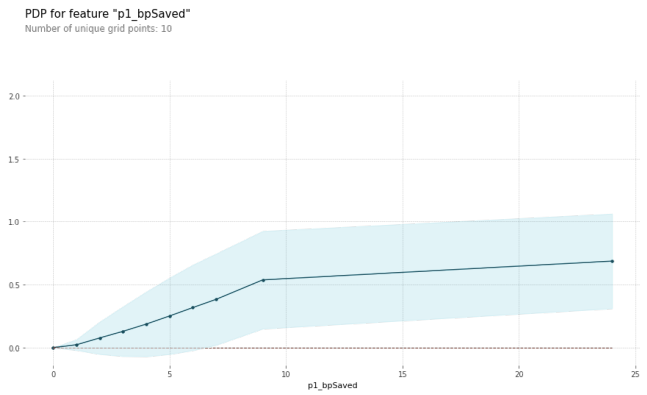

Figure 22: Feature Importances for Decision Tree Classifier

The separator indicating 0.95 refers to the model's score for a random observation from the test set which tends towards predicting a win for the player. The features marked in red, such as breakpoints faces, 2nd serve points won and double faults, push the model score higher. The features marked in blue, the number of serve games, and break points saved, pushed the model score lower. The biggest contributor to classifying this input as a win is the break point faced statistic, as confirmed by previous analyses.

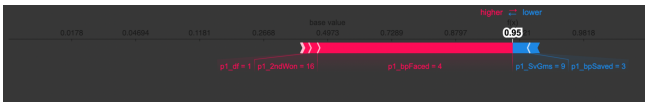

Figure 23: SHAP plot indicating feature influences for an instance from test set

6 EVALUATION

6.1 Olympic Dataset

We did reach a few of our goals in contributing to the female and sports coaches by showcasing the factors that affect a player's performance and the need to make sports more female-inclusive. Further, we also determined the chances of success in a sport based on physical and external factors, which would be a great starting point for a female to improve her chances of winning.

6.2 Tennis Dataset

From all the various investigations into the models, it can be concluded that women tennis players should focus on saving and converting their breakpoints. While they may not seem like the obvious statistic to improve compared to serve ratings or ace ratings, breakpoints can cause a huge change in the momentum of a game. This can be leveraged by the players to change the tide in their favor. Two other statistics that have room for improvement would be ensuring the win of second serve and increasing the ace percentage. Hence, our evaluations show two key features to enhance in both the rally and the serve sections of the game.

7 CONTRIBUTION OF TEAM MEMBERS

- Sahithi - Developed data visualization models described above using 120 years of Olympics data. Predicted top 5 sports using the models mentioned and chances of winning medals using Random Forest.
- Sanskriti - Predicted the chances of winning models using KNN, Logistic Regression and Naive Bayes classifier. Also determined important features for such predictions.
- Tejaswini - Thorough probe into the game of Tennis using 3 different models. Analysis of results to improve a player's game.

8 FUTURE WORK

- Include nutritional information based on skills required for a sport, similar to tennis game improvement.
- Perform similar evaluation and prediction for other genders to promote gender inclusivity in sports.
- Using the existing dataset, a head-to-head comparison model can be built versus the current generalized model. Personalized recommendations for a player could be the future of sports analytics in tennis, which has a history of being a conservative sport.

REFERENCES

[1] Amin Azad. Let data improve your tennis game, Jul 2020.
[2] Emer Brady, Mathias Wullum Nielsen, Jens Peter Andersen, and Sabine Oertelt-Prigione. Lack of consideration of sex and gender in covid-19 clinical studies, Jul 2021.
[3] Ramy Elitzur. Data analytics effects in major league baseball. *Omega*, 90, 11 2018.
[4] Henk Erik Meier, Mara Verena Konjer, and Jörg Krieger. Women in international elite athletics: Gender (in)equality and national participation, Aug 2021.
[5] P Rajesh, Bharadwaj, Mansoor Alam, and Mansour Tahernezhadi. A data science approach to football team player selection. In *2020 IEEE International Conference on Electro Information Technology (EIT)*, pages 175–183, 2020.
[6] Rgriffin. 120 years of olympic history: Athletes and results, Jun 2018.
[7] Vangelis Sarlis and Christos Tjortjis. Sports analytics — evaluation of basketball players and team performance. *Information Systems*, 93:101562, 2020.
[8] Joshua A. Senne. Examination of gender equity and female participation in sport, Feb 2016.
[9] TennisCompanion. Break point in tennis: Definition, examples, strategy, and questions, Apr 2022.