# Olympic History: Athletes and Results Data Analysis


Jake Haas
Computer Science
California State University, Sacramento
Sacramento, CA, USA
jakehaas@csus.edu

Manuel Herrera
Computer Science
California State University, Sacramento
Sacramento, CA, USA
mherrera@csus.edu



**ABSTRACT**

*The Olympics is an international sporting event. Participation in the event has expanded from 241 athletes to 11,500 since the last Olympics [1]. Given the historical data throughout the Olympics, the odds of winning a medal (gold, silver, or bronze) could perhaps be given based on a few biological attributes of the athletes.*

*Therefore, we decided to do exploratory data analysis so we may visualize patterns within the dataset. Furthermore, we wanted to predict if an athlete would win a medal based on those few attributes given. The dataset was provided by the Kaggle user 'rgriffin' under "120 years of Olympic History: Athletes and Results."*


## 1 Introduction

The Olympic Games have been expanding every year which can be seen by the records of the nations participating. The number has grown from 14 nations in 1896 in Athens to 207 nations in 2016 at the Rio Olympics [1]. This international sporting event where thousands of athletes from various countries compete in various sports every four years, has experienced enough growth in which we can begin to ask questions on the evolution of the Olympics based on gender participation or their performance and results based on basic biological information [2].

## 2 Design and Methodology

The dataset provided consists of 271,116 unique athletes with 15 attributes:

1. **ID** - Unique number for each athlete
2. **Name** - Athlete's name
3. **Sex** - M or F
4. **Age** - Integer
5. **Height** - In centimeters
6. **Weight** - In kilograms
7. **Team** - Team name
8. **NOC** - National Olympic Committee 3-letter code
9. **Games** - Year and season
10. **Year** - Integer
11. **Season** - Summer or Winter
12. **City** - Host city
13. **Sport** - Sport
14. **Event** - Event
15. **Medal** - Gold, Silver, Bronze, or NA

The dataset was obtained by the user 'rgriffin' by scraping and wrangling the data from a website dedicated to the collection of sport statistics. The collection includes all games from Athens 1896 to Rio 2016. Another file called "noc_regions.csv" was provided as well, however, we made the decision to drop the file. The National Olympic Committee (NOC) regions file simply defines which region each

NOC is associated with, however the original “athelete_events.csv” file already contains a column for the NOC that the athletes are associated with.

We did further preprocessing of the data by selecting attributes we deemed relevant such as: Sex, Age, Weight, Height, Sport, and Medal. We made the decision to remove ID, Name, Games, City, and Event. These were removed based on the idea that personal identifying information would not be useful in any predictions or data analysis. The Event column was removed because it splits the Sports column based on specific games based on the sport. For example, the swimming tag would be represented as Swimming Men’s 200 Meter Breaststroke, Swimming Men’s 400 Meter Breaststroke, and so forth in the Event column. Therefore, we made the decision to drop the column.

```
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   267538 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

The dataset came with null values that had to be resolved. We identified them by checking existing null values for each column within the dataset. Our results were 9,474 for Age, 60,171 for both Weight and Height, 3,578 for Sport and 231,333 for Medal. The reason the Medal column returned so many null values was because the dataset had the tags Gold, Silver, and Bronze for medalists and a null tag for non-medalists. The decision was made to give non-medalists the “NoMedal” string value to make further data analysis easier.

```
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport       3578
Event          0
Medal     231333
dtype: int64
```

At the end of our data preprocessing step, we came out with a total of 206,165 unique athletes with the attributes: Sex, Age, Weight, Height, NOC, Year, Season, Sport, and Medal.

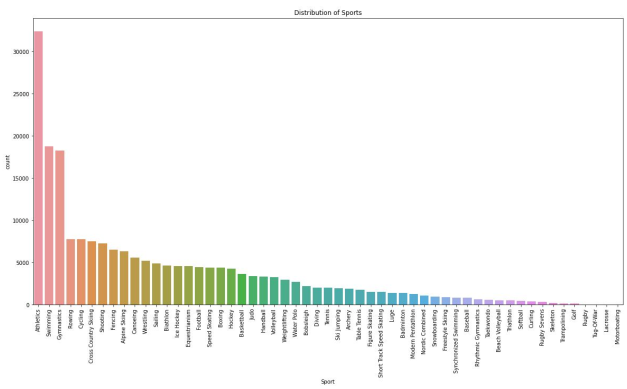
| | Sex | Age | Height | Weight | NOC | Year | Season | Sport | Medal |
|---|---|---|---|---|---|---|---|---|---|
| 0 | M | 23.0 | 154.0 | 45.0 | GER | 1896 | Summer | Athletics | NoMedal |
| 1 | M | 23.0 | 176.0 | 66.0 | USA | 1896 | Summer | Athletics | Gold |
| 2 | M | 23.0 | 176.0 | 66.0 | USA | 1896 | Summer | Athletics | NoMedal |
| 3 | M | 21.0 | 183.0 | 66.0 | USA | 1896 | Summer | Athletics | Gold |
| 4 | M | 21.0 | 183.0 | 66.0 | USA | 1896 | Summer | Athletics | Gold |

For predictions, we wanted to use different algorithms to plug our data into after a train/test split. We chose to use RandomForest, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN). Due to the range of medalists and non-medalist, we under-sampled to help with our predictions.

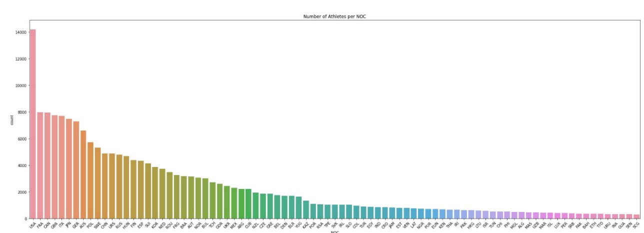
```
y.value_counts()

2    175958
1     10166
0     10147
3      9866
Name: Medal, dtype: int64

np.unique(y_resampled, return_counts=True)

(array([0, 1, 2, 3]), array([9866, 9866, 9866, 9866]))
```

## 3 Results

Before we began the predictions, we did exploratory data analysis of the dataset. We wanted to view the distribution of different kinds of sport records across the whole dataset.



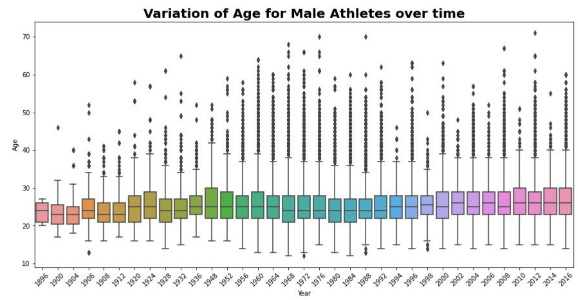

The distribution clearly shows many more records of 'athletic' events followed by 'swimming' and 'gymnastics, with such a smaller amount for the rest of the categories. Furthermore, we wanted to view the number of athletes per NOC tag.

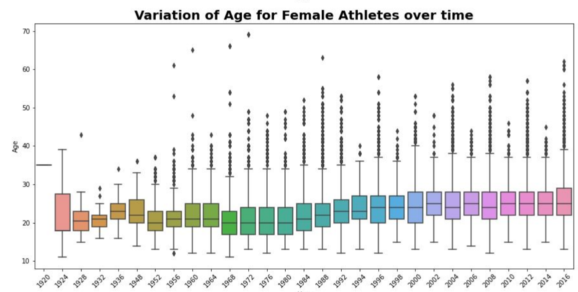



| | Age | Height | Weight |
|---|---|---|---|
| count | 206137.000000 | 206137.000000 | 206137.000000 |
| mean | 25.053445 | 175.372296 | 70.687805 |
| std | 5.478793 | 10.545929 | 14.340737 |
| min | 11.000000 | 127.000000 | 25.000000 |
| 25% | 21.000000 | 168.000000 | 60.000000 |
| 50% | 24.000000 | 175.000000 | 70.000000 |
| 75% | 28.000000 | 183.000000 | 79.000000 |
| max | 71.000000 | 226.000000 | 214.000000 |

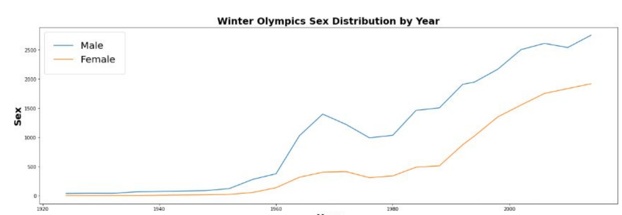
This distribution shows that the United States had more attending athletes than any other countries over time, but many of them are somewhat similar to one another. We also wanted to view a quick descriptive printout for the numerical attributes. The 25th and 50th percentile show that most athletes, at least overall, are around 21-24 years of age, 168-175 centimeters, and about 60-70 kilograms.

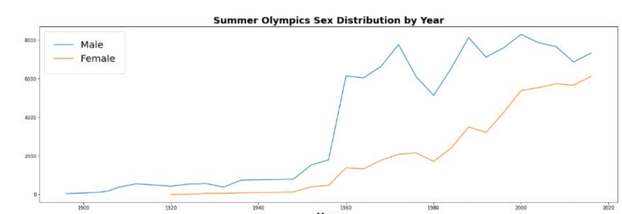
We also wanted to view the distribution through their quartiles with the ages between male and females over the years. As seen on the box and whisker plot, the average age is focused around the early 20's



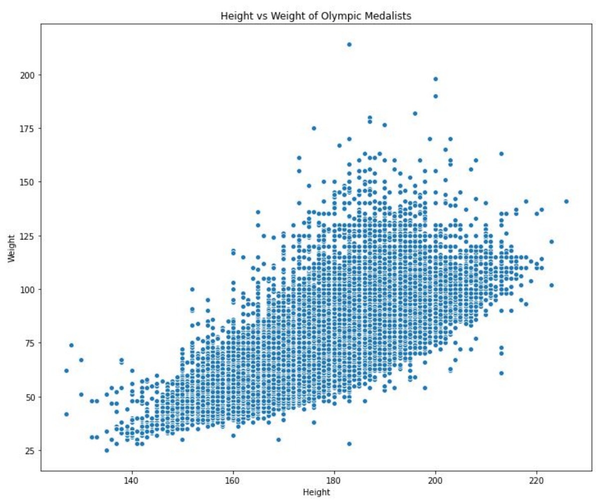




Furthermore, we wanted to view the attendance between males and females over the years. It's clear there is an imabl;ance in male to female attendance for the both summer and winter events. Therefore, we decided to split the data for the two types to see if it makes a difference in our predictions.



We believe Height and Weight played a vital role as well so we wanted to see if there was a trend that existed within our data. Our visualization showed there was a trend but it was not too extreme.

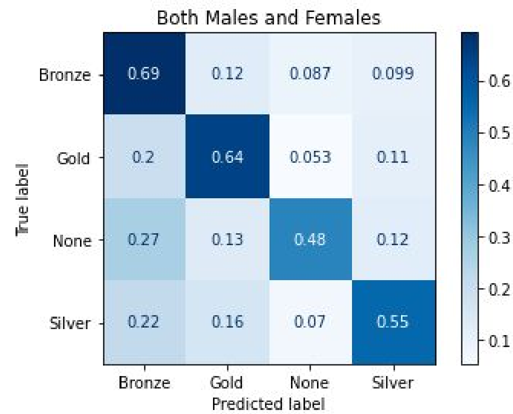



For our predictions, we wanted to compare the results of our predictions from the various algorithms we mentioned. The one below includes both males and females. We plugged the data into them after a train/test split. The scores were low except for KNN, which managed to get an average score of 59%. We included both the correlation matrix where age doesn't correlate much.

| Males and Females Only | |
|---|---|
| **Algorithm** | **Score** |
| RandomForest | 0.3934 |
| LogisticRegression | 0.3095 |
| Support Vector Machine | 0.24 |
| K-Nearest Neighbor | 0.59 |

| | Age | Height | Weight |
|---|---|---|---|
| Age | 1.000000 | 0.103132 | 0.152635 |
| Height | 0.103132 | 1.000000 | 0.652414 |
| Weight | 0.152635 | 0.652414 | 1.000000 |

The confusion matrix did turn out better and showed the ability to predict being different per medal, or lack thereof.

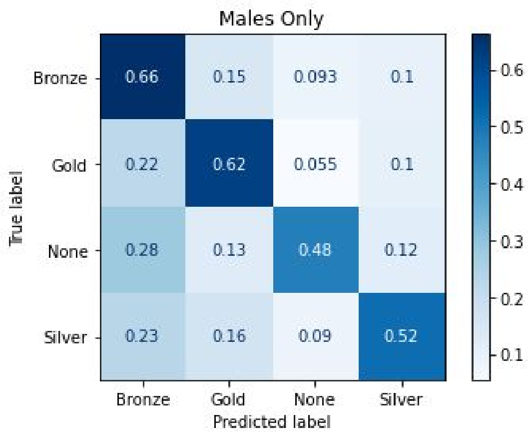



Next we wanted to analyze the predictions of only males. The results show the prediction results a slightly lower average score for KNN, buyt the results are fairly close to being the same as before. The correlation of the variables Height and Weight also remain very similar to the last attempt being at about 0.66 compared to 0.65.

| Males Only | |
|---|---|
| **Algorithm** | **Score** |
| RandomForest | 0.3724 |
| LogisticRegression | 0.2904 |
| Support Vector Machine | 0.231 |
| K-Nearest Neighbor | 0.5694 |

| | Age | Height | Weight |
|---|---|---|---|
| Age | 1.000000 | 0.097792 | 0.142562 |
| Height | 0.097792 | 1.000000 | 0.662574 |
| Weight | 0.142562 | 0.662574 | 1.000000 |

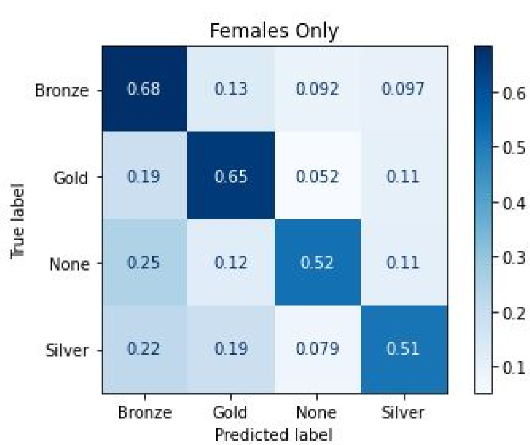
Furthermore, the results are reasonably similar for the confusion matrix with “non-medalists” still being under 50%




Next, the female only results proved to be the best but only by a margin.

| Females Only | |
|---|---|
| **Algorithm** | **Score** |
| RandomForest | 0.4056 |
| LogisticRegression | 0.316 |
| Support Vector Machine | 0.2456 |
| K-Nearest Neighbor | 0.5938 |

There's a slightly higher average score for KNN and the confusion matrix results were all above 50% for each category.

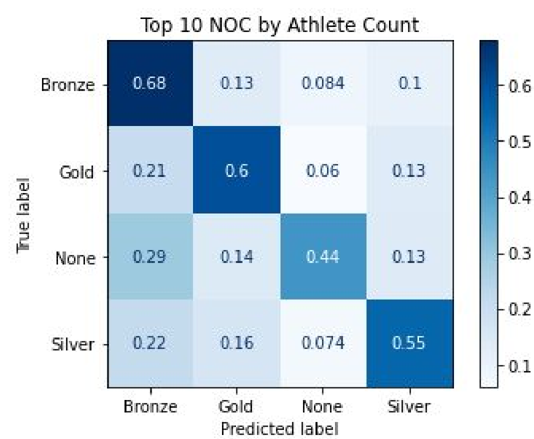



Finally, the top NOC by athlete count had lower results. It is apparent that there was not a lot of variety with the results so far, so our attempts to spit the data to get different results didn't help as much. In this section we wanted to see the top 10 NOC by total number of athletes over the years, which are the most active committees. The print out of the results were as follow:

| Top 10 NOC by Athlete Count | |
|---|---|
| **Algorithm** | **Score** |
| RandomForest | 0.4236 |
| LogisticRegression | 0.3287 |
| Support Vector Machine | 0.2502 |
| K-Nearest Neighbor | 0.5665 |

```
TopNOC.size().sort_values(ascending=False).index[:10]

Index(['USA', 'FRA', 'CAN', 'GBR', 'ITA', 'JPN', 'GER', 'AUS', 'POL', 'SWE'], dtype='object', name='NOC')
```

Overall, the results were lower than this than it was for the unmodified dataset and was even worse at identifying those who wouldn't get medal leading to a lot of false positive for the categories.




## 4 Related Works

The paper *Analyzing Sports Training Data with Machine Learning Techniques* by Purdue University students, dove deeper in the machine learning aspect to improve the training and coaching of their Women's Soccer Team [4]. Their data preprocessing included making the data anonymous and rearranging the data according to players vs according to training drills. There was player data corresponding to unique individuals and drill; data that was average out across all players. Furthermore, they normalized features into a common range.

## 5 Conclusion

Overall, it was a learning experience doing hands-on exploratory data analysis. Various useful data visualization and machine learning libraries were used. For instance, we had to learn more about a useful visualization named Seaborn to supplement our exploratory analysis section. There is no doubt the knowledge learned during this project will be incredibly useful later on. It should be noted that there are many different combinations of features to be used in this project that may give better predictions. For now, the experience attained during this project was a pleasant one.

## 6 References

[1] "Rio 2016." *International Olympic Committee*, 17 Apr. 2018, www.olympic.org/rio-2016.

[2] Rgriffin. "120 Years of Olympic History: Athletes and Results." *Kaggle*, 15 June 2018, www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results.

[3] VanderPlas, Jake. "Visualization with Seaborn." *Visualization with Seaborn | Python Data Science Handbook*, jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html.

[4] Mahfuz, Rehana, et al. "[PDF] Analyzing Sports Training Data with Machine Learning Techniques: Semantic Scholar." *Undefined*, 1 Jan. 1970, www.semanticscholar.org/paper/Analyzing-Sports-Training-Data-with-Machine-Mahfuz-Mourad/8a7a774e2aa3410575e0137071ed591fd65d1f78.

## 7 Appendix