2019

Olympic Athletes

USING DATA TO PREDICT OLYMPIC MEDALS

HEATHER EWTON

Table of Contents

Problem

One of the greatest honors for an athlete is to represent their country in the Olympic games. Held every four years, the summer Olympic games feature sports such as swimming, gymnastics, athletics (also called track and field), and equestrian events. During the Olympic Games, each country follows its athletes, carefully charting and tallying the number of medals garnered in friendly sport competition. At the close of each set of games, medal totals are calculated, and athletes return home to celebration, sponsorships, and occasionally monetary rewards.

Stakeholders

What if there was a way to predict which athletes were most likely to medal in their sport at the Olympic Games? If so, coaches and national Olympic committees around the world may want to use these guidelines when selecting athletes from each country to train for the Olympics. These predictions could significantly narrow the field from many to those most likely to earn a medal in their respective sport. This would free up coaches' time and allow funds spent on scouting to be redirected to other areas.

Hypothesis

- The null hypothesis for this analysis is that there is no relationship between an athlete's physical traits (age, weight, height) and an athlete's medal status (no medal, bronze, silver, or gold).

- The alternative hypothesis for this analysis is that there is a relationship between an athlete's physical traits (age, weight, height) and an athlete's medal status (no medal, bronze, silver, or gold).

Methods

- Dataset

The dataset used for this analysis and prediction has been sourced from Kaggle [https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete_events.csv] and contains 271,116 records of Olympians who have competed in summer and winter Olympic games. Of the data, 75% is male athletes due to restrictions on women competing in the Olympics. Age, height, and weight data are all included for most athletes in the dataset. Since 82% of the data set comes from the summer Olympic Games and the winter games are incomplete, the analysis will only consider the records for Summer Olympic Games, which consists of 222,552 records.

- Data Clean-up

The first step in working with this data was to eliminate the incomplete winter Olympics data through filtering. After completing this, it was discovered that there many of the athlete's physical information (age, height, weight) was not recorded for athletes that did not medal in the first few Olympic games. Due to this gap in values and the inability to locate that missing data, the missing values were dropped from the data set, leaving 166,706 records.

- Data Analysis

The first step in exploring the data was to look for outliers in the data set using the boxplots in Figure 1 (below). Looking at all the competitors' ages reveals a median age of 24 with outliers above 39 years of age. In the boxplot for height outliers, we see that the median for height is 175 cm, with outliers being shorter than 150 cm or taller than 205 cm. Lastly, from the Weight Outliers plot, we see that the median weight of athletes is right at 70kg, with outliers weighing less than 30 kg or more than 105 kg.

Fig. 1: Age, Height, and Weight of Olympic Athletes

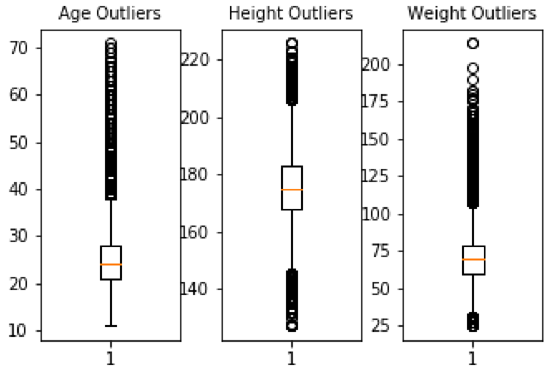
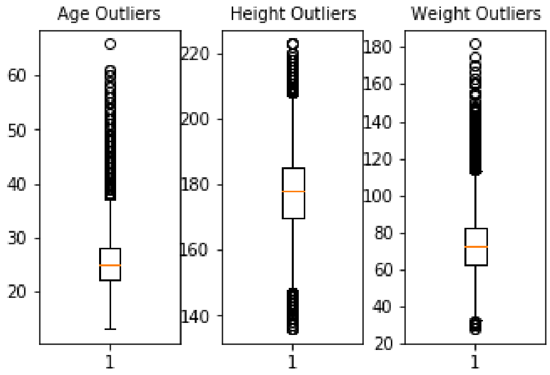


Fig. 2: Age, Height, and Weight of Olympic Medalists



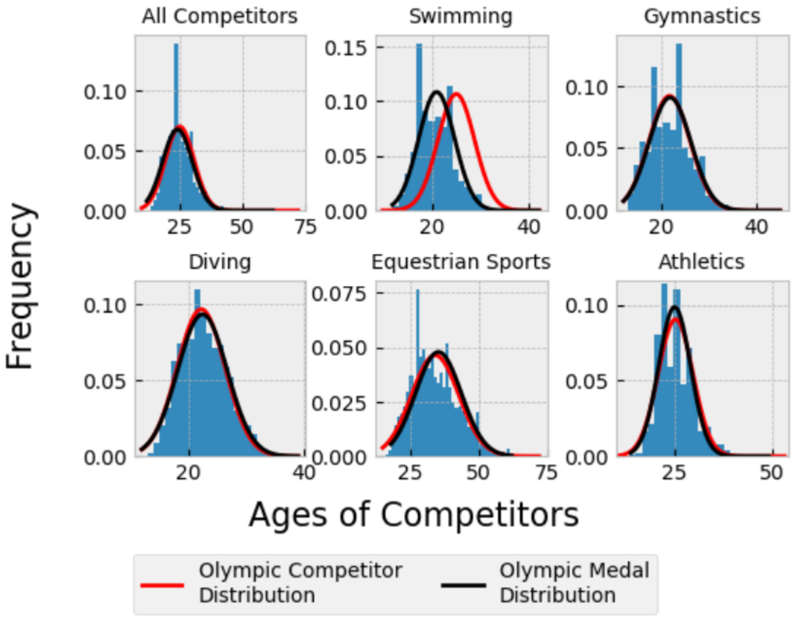
The next step in exploring the data was to determine if the same outliers hold true for medalists. After filtering the data to for medaling athletes, the data was plotted again to determine if the outliers remained the same for medalists. From Figure 2, we see that there is a difference in outliers for Olympic medalists compared to the outliers for all Olympians. The median age shifts slightly from 24 to 25 years of age, with outliers being anyone above the age of 36. The median height shifts slightly as well, from 175 cm to 178 cm, though outlier range for this group remains about the same. The median weight also shifts slightly higher, to 73 kg from 70 kg.

Using only the medalist data, the next step was to determine if there is a relationship between any of the variables measured. Several pair plots (see appendix) were constructed to determine if there was a relationship between any of the variables in the plots. Overall, the plots revealed some consistent trends. The first noticeable trend that appears is between age and sport; many physically demanding sports tend to have Olympians that are below 35 in age, while equestrian sports and art competitions tend to have Olympians that are above 45 in age. Another trend that becomes apparent is between weight/height and sport: there are several outliers that appear in the pairplot for height/weight but the sports tend to be clustered together for height/weight ranges. The last noticeable trend is that the age, height, and weight have all increased in range over the years but narrowed with the last summer Olympic competition (2016, Rio de Janeiro).

Although the pair plots revealed the above trends, the pair plots were jumbled and unclear due to the large number of sports held within the history of the Summer Olympics. To better analyze the data, the data was separated into six data sets: overall data (all Olympians in the data set), swimming events, gymnastics events, diving events, equestrian events, and athletics events. These sports were chosen as each sports category consists of several events that feature both genders and each category has a long history of inclusion in the Summer Olympic games, providing enough observations to make reasonable conclusions. Additionally, equestrian sports were included as the competitions tend to feature athletes who are considered outliers in age, height, and weight. These separated data sets were then compared to each other and the medalists for the respective sports.

In Figure 3 (below), the ages of the Olympic competitors were charted (blue histogram) and separated by their sport, then the normal curve was laid over each set of data. The black curve represents the Olympic distribution for the Olympic medalists in that sport while the red curve is the normal distribution line for all of that sport's competing athletes.

Fig. 3: Ages of Olympic Competitors by Sport

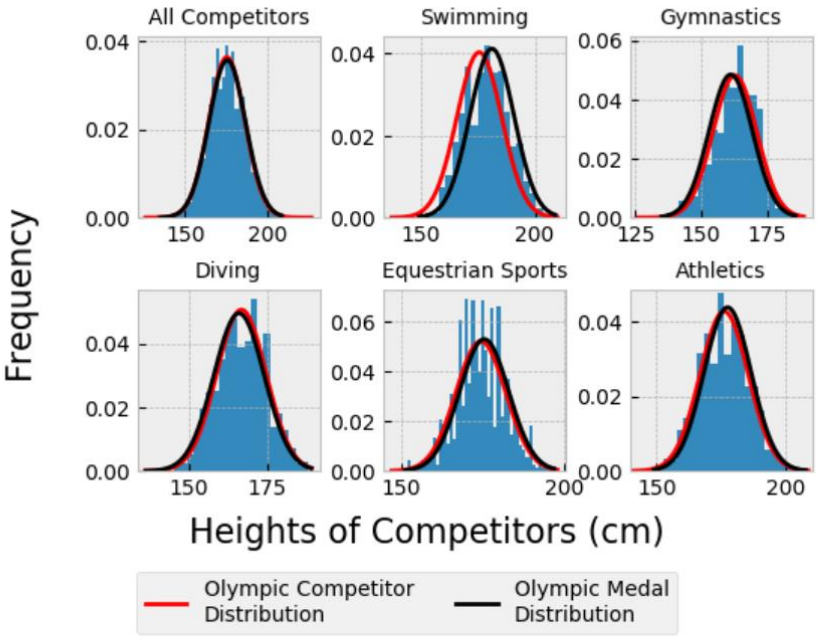


- **All Competitors**: The data from all competitors shows that the normal distribution for medalists is close to the normal distribution for all competing Olympians. The peak for the medalists is slightly shifted to the left when compared to the full data, but not significantly.
- **Swimming**: Interestingly, the distribution curve for swimming medalists is slightly left shifted, with the peak occurring at a mean age of 21 years of age. Although there are swimmers who compete into their thirties, the black medalist curve indicates that most medalists are located within the first half of the age group for competitive swimmers.
- **Gymnastics**: Age for gymnasts and Olympians medaling in gymnastics also peaks at the young age of 21.8 years of age. As these curves are nearly identical, it can be inferred that the athletes receiving medals mirror the larger competing population of Olympic gymnasts.
- **Diving**: The distribution of diving medalists is also closely aligned with the distribution of all Olympic divers. The medalists tend to predominantly be around age 22 and the medaling population closely mirrors the total Olympic diving population.
- **Equestrian sports**: Age for equestrian competitors varies widely, with peaks around 30, then in the upper thirties and then again at close to 50 years in age. There are several reasons why these peaks may exist including training the horse, continued improvement in ridership and as this sport is less harsh on the human body- an ability to continue competing past the years when one is in their peak physical condition. As a result of this variation within ages, the distribution for equestrian athletes is more widely spread out and right shifted than other sports. For medalists, the age is slightly older, with a mean age of 35.3 years of age and the medaling curve distribution

slightly shifted right of the distribution curve for all equestrian athletes, which may indicate that the medalists might be benefitting from practice and experience.
- **Athletics**: Peak ages for athletics competitors are at 22 and 24, with a right tail as the population of athletics competitors shrinks with age. Athletics athletes (both medaling and non-medaling) have a mean age of 25 years of age and both curves are narrow, suggesting that the peak Olympian for this sport will be in their twenties.

This same method of creating a histogram with overlaid distributions for medalists and all competitors was next repeated with Olympian heights. In Figure 4 (below), the heights of the Olympic competitors were charted (blue histogram), then the normal curve was laid over each set of data. The black curve represents the Olympic distribution for the Olympic medalists in that sport while the red curve is the normal distribution line for all of that sport's competing athletes.

Fig. 4: Heights of Olympic Competitors by Sport



- **All Competitors**: Looking at the medal data for all competitors (black curve), it is clear that athletes who medal are slightly taller than the average Olympic athlete who is 170 cm tall. Additionally, the narrower distribution of the curve suggests a narrower range of height, peaking at 175.75 cm.
- **Swimming**: Given that the best swimmers have often been tall, it is not surprising to find that the mean height of an Olympic medalist for swimming is 181 cm, or 11 cm above the median height for Olympians. What is surprising is that the medalist swimmers are taller than the average Olympic swimmers.
- **Gymnastics**: Gymnastics also offers some surprising results. The mean height for an Olympic medaling gymnast is 161.25 cm, compared to the slightly taller average of 162.85 cm. Shorter gymnasts seem to medal more often, thus the slight move of the curve to the left for medalists.
- **Diving**: The distribution for diving medalists is very closely aligned with the distribution curve of all Olympic divers; there is less than 1 cm of difference in height from the average medalist compared to the average Olympic diver.

- **Equestrian sports**: The height distribution for equestrian athletes and medaling equestrian athletes is approximately the same- there is approximately one-centimeter difference in height between medalists and the mean for all competing equestrian athletes. Equestrian athletes are slightly taller than the average Olympian, at 175 cm compared to 170 cm.
- **Athletics**: The mean height for medaling athletics competitors (177.6 cm) is also rather close to the mean height for all athletics competitors (176.27 cm). There is a very slight shift in the curve for medaling athletics Olympians, caused by this difference at all heights; for example, the shortest athletics medalist was 150 cm in height while the shortest competitor was 142 cm in height.

The last set of data to be charted in this same fashion was the weights of the Olympic competitors. In Figure 5 (below), the weights of the Olympic competitors were mapped (blue histogram), then the normal curve was laid over each set of data. The black curve again represents the Olympic distribution for the Olympic medalists in that sport while the red curve is the normal distribution line for all of that sport's competing athletes.

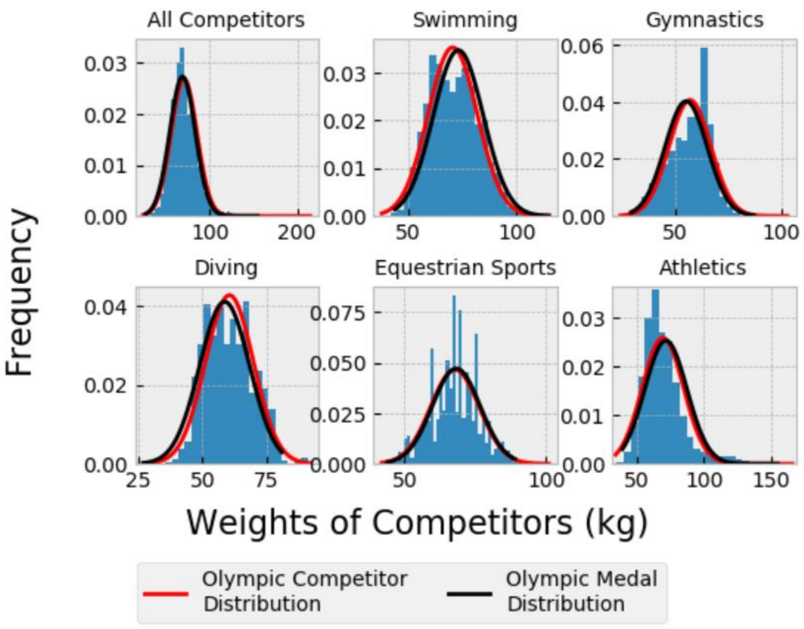


Fig. 5: Weights of Olympic Competitors by Sport

- **All Competitors**: The distribution for all Olympians is close to the distribution for all medaling Olympians for weight. The mean weight of a medaling Olympian is 68.9 kg while the mean weight of all Olympians is 70.67 kg. As you can see in Figure 5, the curves overlap except at the peak, where the means are slightly different.
- **Swimming**: As with height, the curve for weight of Olympic swimming medalists is slightly offset to the right from the distribution curve for the weight of all Olympic swimmers. The mean for the average Olympic swimmer is 70.59 kg while the mean weight for a medalist is 73.25 kg. This difference makes sense as the average medalist for swimming is taller than the average Olympic swimmer.
- **Gymnastics**: Similarly, the curve for gymnastics medalists is shifted to the left, with the average gymnastics' medalist weighing 55.07 kg, compared to the overall average of 58.89 kg. This is not

surprising as the height of gymnastics medalists was found to be slightly shorter than the average competing Olympic gymnast.
- **Diving**: There was a difference between the curves for diving, which was unexpected as the distribution curves for height were very similar. Diving medalists have a mean weight of 58.69 kg while the average competing Olympic diver has a higher weight average of 60.57 kg.
- **Equestrian sports**: The distribution for medaling equestrian athletes and all equestrian athletes is approximately the same, indicating that an athlete's weight may not have a significant impact on their ability to medal in the sport.
- **Athletics**: Since the height curve for athletics medalists was slightly shifted to the right from the average, it is not surprising that the weight curve for athletics medalists is also slightly shifted to the right. The average athletics competitor had a weight of 69.27 kg while the average athletics medalist had a weight of 71.51 kg.

After reviewing the histograms and pair plots, there seemed to be a link between weight and height for competitors. This was better examined by plotting all the data by sport in Figure 6 (below). Based on the scatter plots in Figure 6, there is indeed a positive correlation between weight and height of Olympians.

To determine if this has any impact on medalists, the medalists data was then laid on top of the previous chart in orange, creating Figure 7.

Fig. 6: Weight vs. Height for Olympic Competitors

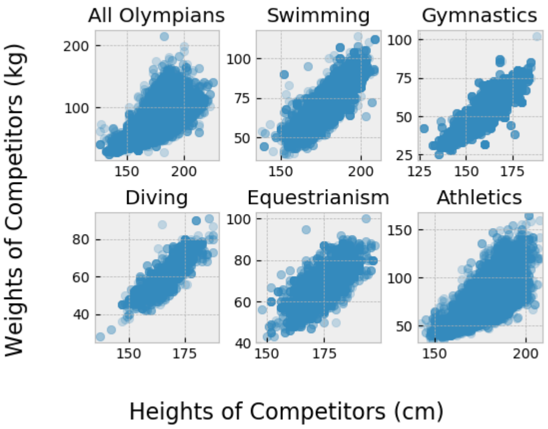
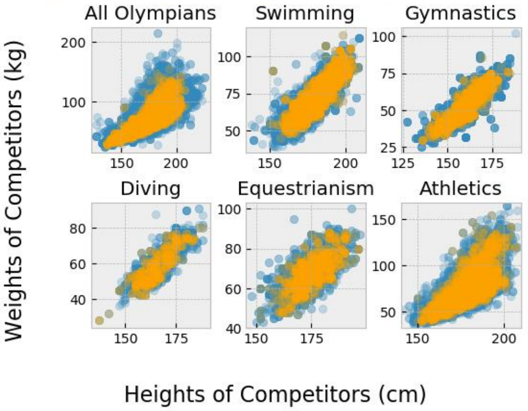


Fig. 7: Weight vs. Height for Olympic Competitors and Medalists



From Figure 7, we see that the weight/height correlation is even stronger for Olympic medalists, colored in orange. For example, the correlation coefficient between weight and height for all competing athletes was 0.79 and increased to 0.83 when only the medalists data was analyzed. An increase also occurred in the correlation coefficient for swimming, gymnastics, diving, and athletics. The only group to experience a decline in correlation between height and weight was the equestrian athletes. Although the decrease was small (0.002), it may suggest that height/weight is less important to a competitor's ability to medal in the sport.

- **Predictive Models**

  Once the data analysis was complete, the datasets were then analyzed using three types of predictive models: Random Forest Classification (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors Classifier (KNN). These classifiers were each run against a training set of data and a testing set of data. After running each classifier, a confusion matrix, classification report, and accuracy score were computed for each algorithm (located in Appendix); finally, the matrices, report, and accuracy scores were compared to the baseline and each of the other models to make a conclusion about the predictive abilities of each algorithm.

  In each of the data sets, the Random Forest Classifier was found to be the best predictor for determining an athlete's medal status. Although the SVC and KNN classifiers had a higher accuracy score in each data set, the increase in accuracy was obtained by the algorithm predicting a higher number of non-medalists and in some cases, predicting no or few medalists. The Random Forest Classifier likely performed better as it resamples the data repeatedly (1,000 to 5,000 times depending on the dataset as smaller datasets were resampled more), leading to an increased likelihood of each category being represented fairly. Given the resampling of the data and the prediction of each class, it is not surprising then, that the Random Forest algorithm had a slightly lower accuracy than the other classifiers but higher F-1 scores.

  Of each data set, the two with the most surprising results were gymnastics and equestrian events. The gymnastics data had a baseline accuracy score of 87.7%, which is unusually high compared to the other data sets analyzed (most were in high 70th to low 80th percentiles). This high baseline accuracy score indicates that not only can gymnastics medalists be predicted easily, but they are also primarily from a specific age group and country group. For the gymnastics medalists, the RF classifier returned an accuracy score of 93.3%, furthering the hypothesis that medalists can be predicted, at least for gymnastics.

  The other surprising result, equestrian events, stood out for the opposite reason: rather than high scores, each of the classifiers performed poorly on predicting medalists for this data set. The baseline accuracy for this set was 72.3%, indicating that medalists in this data set were likely harder to predict and less likely to follow a trend. We know from earlier analysis that this may be due to the large variation in age of medalists as well as height and weight. In this data set, the Random Forest Classifier was again the top performer, although it did not do well with the data, predicting zero silver medalists correctly.

Conclusion & Recommendations

Given the data analysis and the machine learning algorithms, I am comfortable rejecting the null hypothesis (there is no difference between medalists and non-medalists) and accepting the alternative hypothesis (there is difference between medalists and non-medalists). From the exploratory data analysis, there was a clear cluster of Olympic medalists when compared to all Olympic competitors.

Using the data available (country, age, height, weight, sport, sex), the Random Forest algorithm performed the best in predicting medalist categories, although it was not a perfect predictor. To increase the accuracy of the algorithm, I would propose adding more biometrics and narrowing the analytics by country (ex: only including the United States athletes). Another possible solution would be to sort the competitors into two categories: medalists vs. non-medalists; this binary classification would perhaps provide a better basis for the machine learning algorithms while being less computationally expensive.

Going forward, this data should be used with some caution: it appears that medalists are different biologically than those who do not medal. However, without further data and analysis, it is not recommended that athletes are chosen only based on the categories analyzed here.

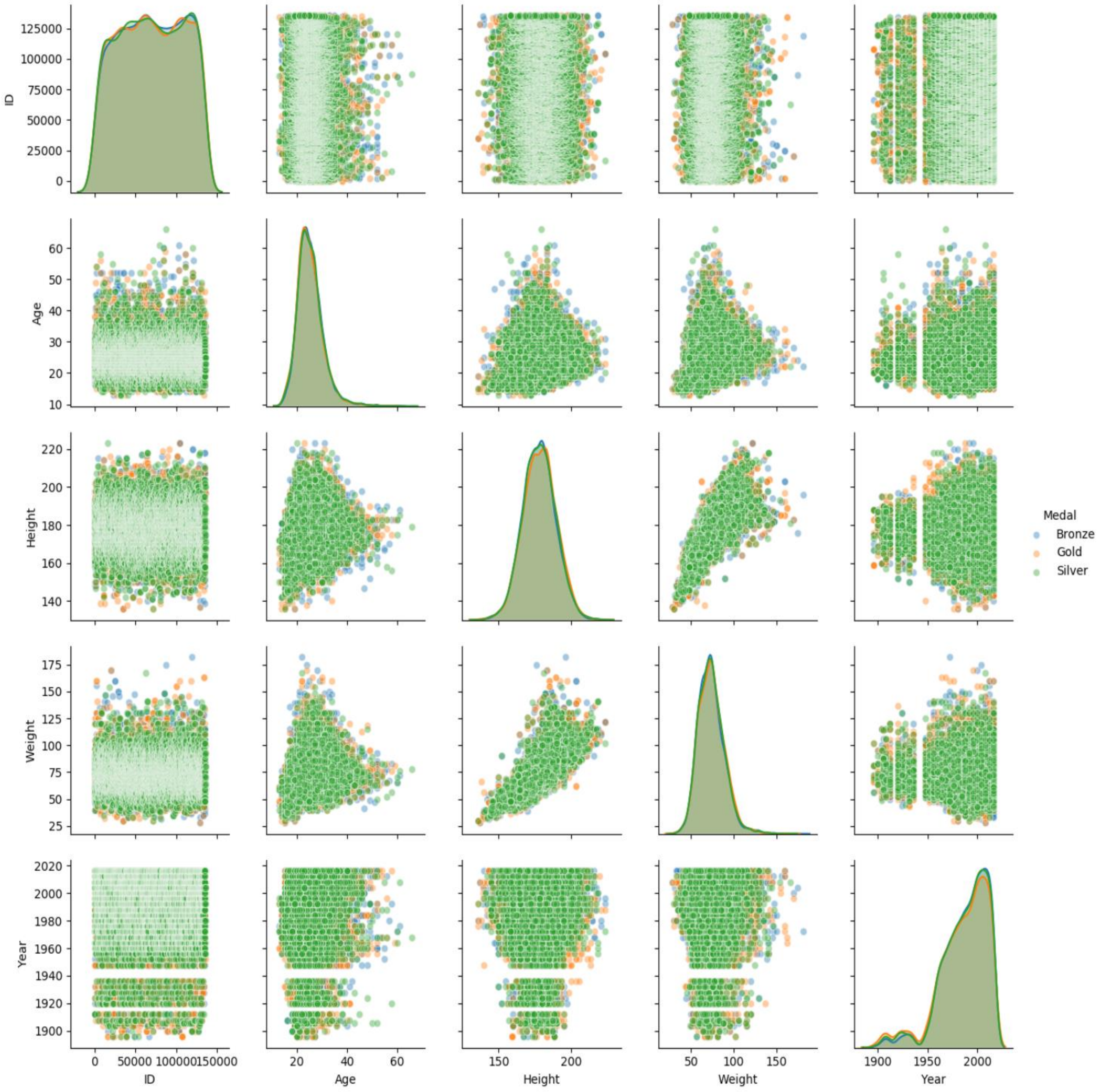
Acknowledgements

I would like to acknowledge and thank my mentor, Devin Cavagnaro, for his guidance in analysis and predictive modeling on this project.

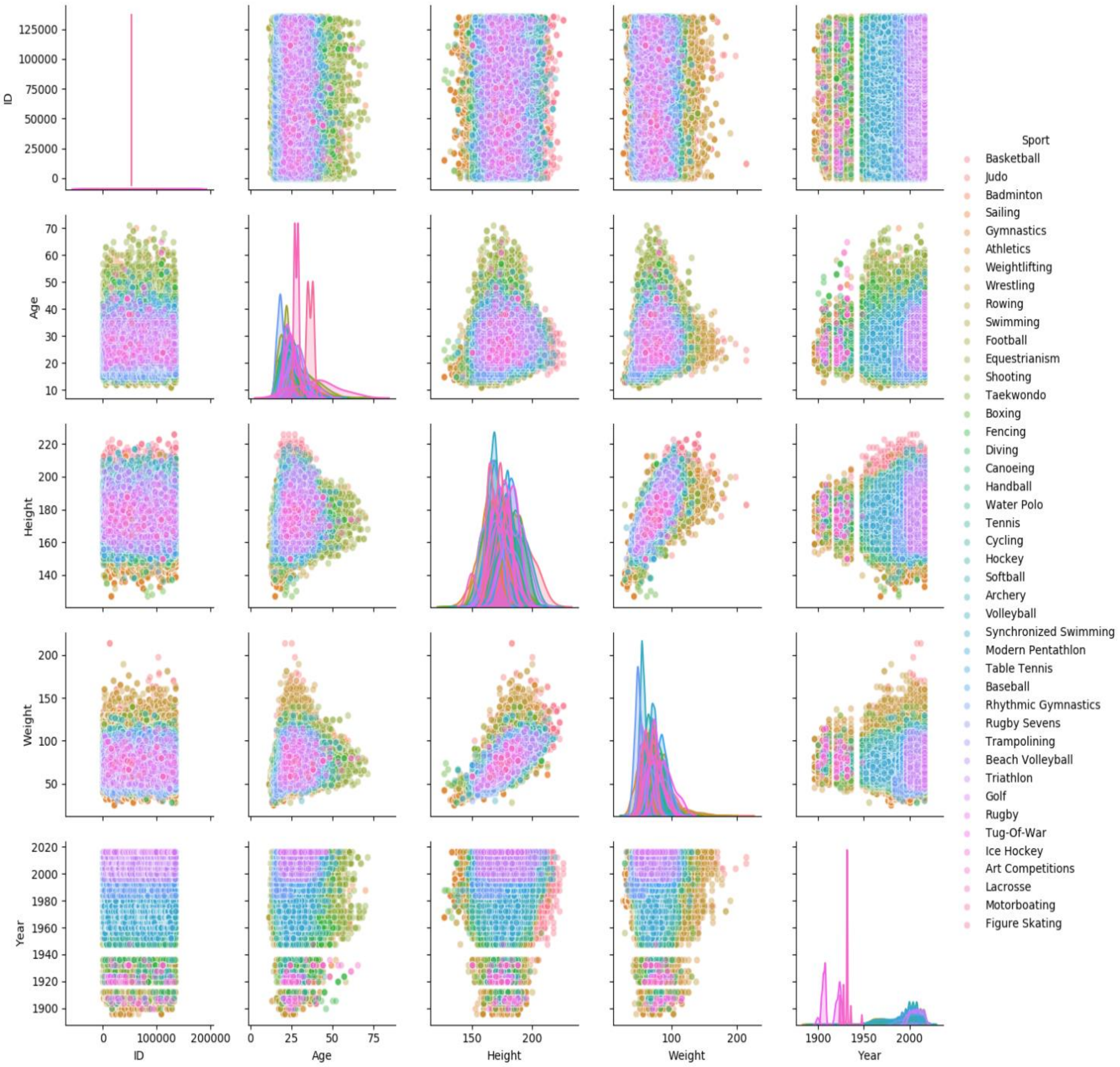
Appendix

Pair plots

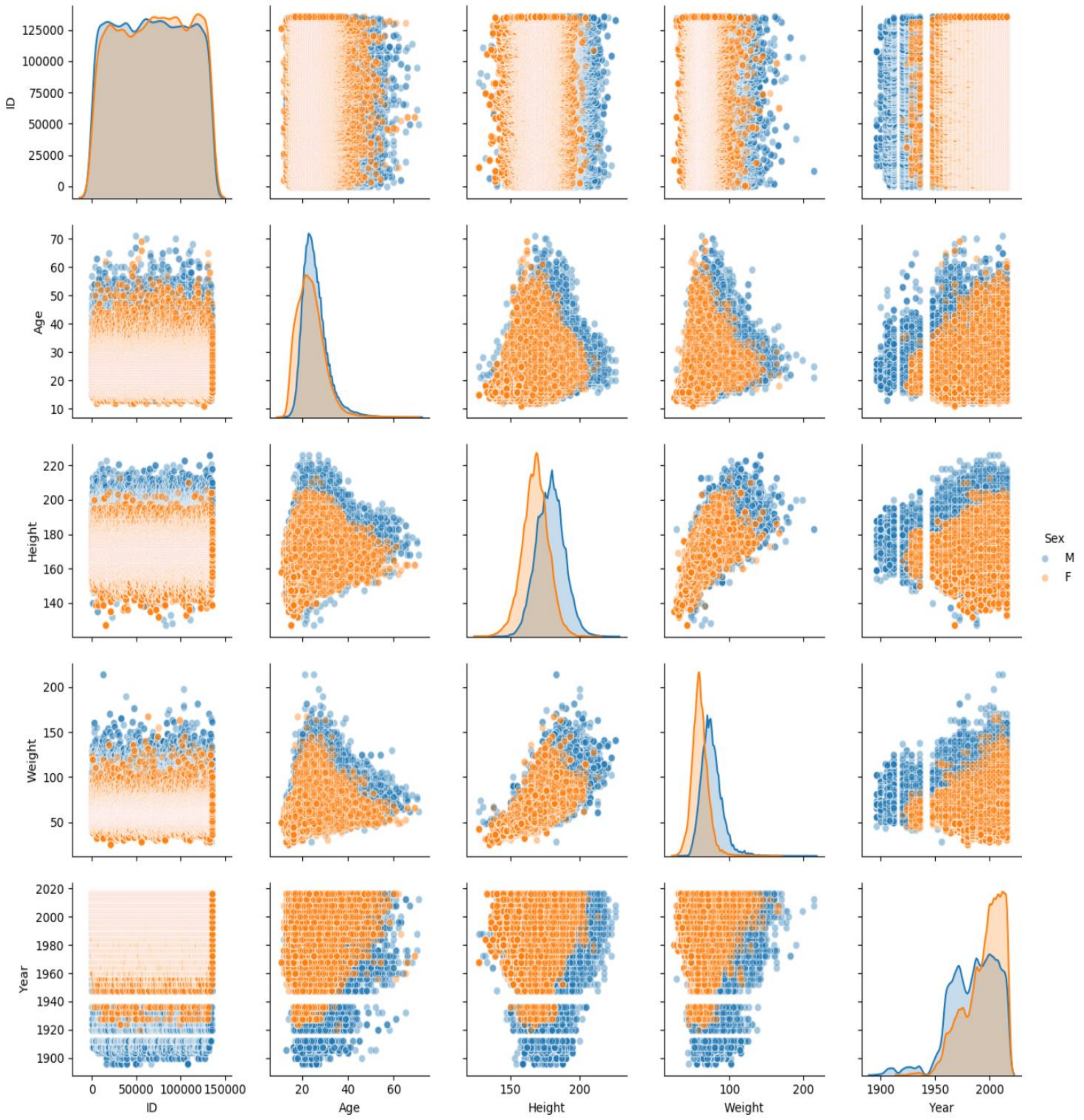
Pairplot for Summer Olympics Data, Colored by Medal



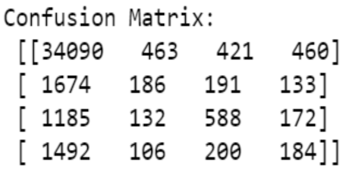
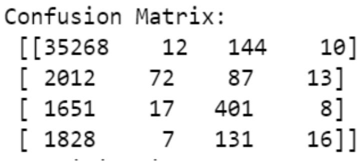
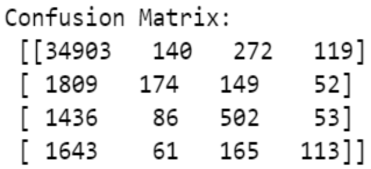
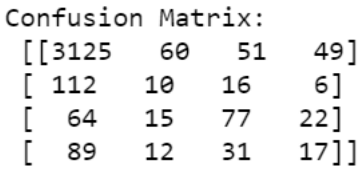
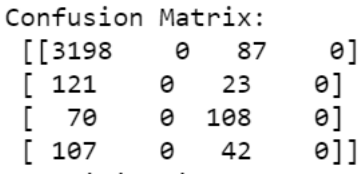
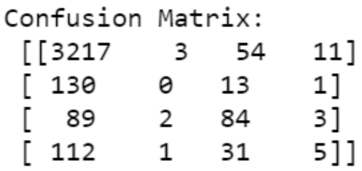
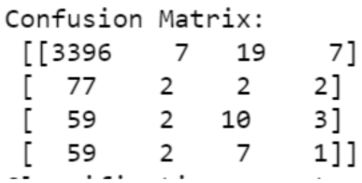
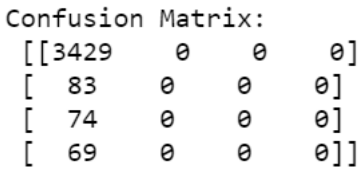
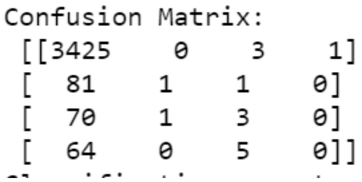
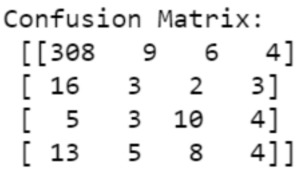
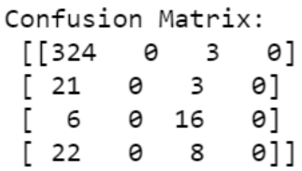
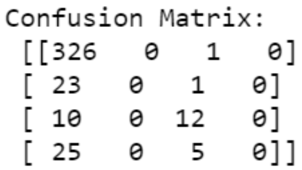
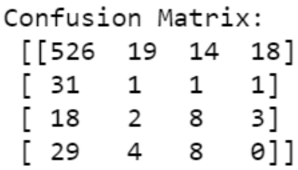
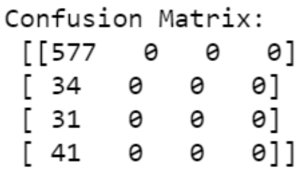
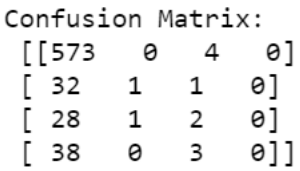
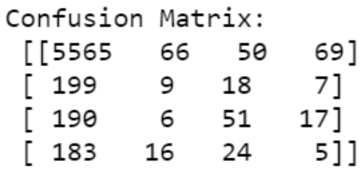
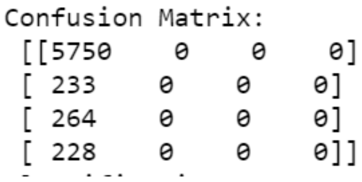
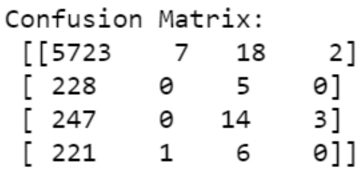
Pairplot for Summer Olympics Data, colored by Sport



Pairplot for Summer Olympics Data, colored by Sex



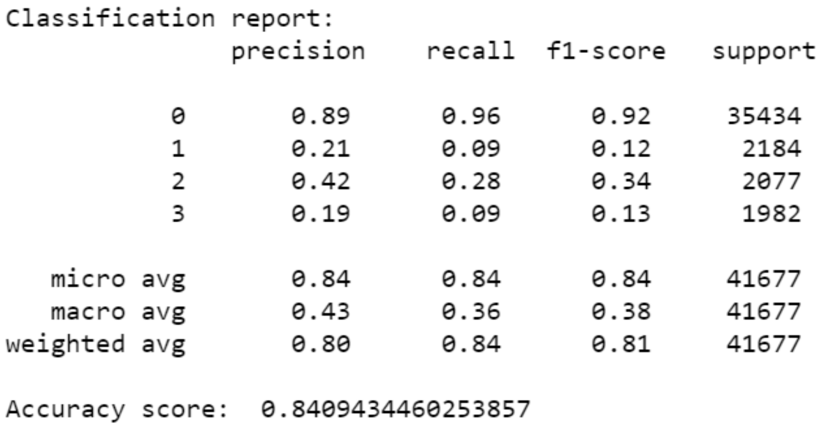
Confusion Matrices

	Random Forest	SVC	KNN
Full Summer Olympics Dataset	 Confusion Matrix: [[34090 463 421 460] [1674 186 191 133] [1185 132 588 172] [1492 106 200 184]]	 Confusion Matrix: [[35268 12 144 10] [2012 72 87 13] [1651 17 401 8] [1828 7 131 16]]	 Confusion Matrix: [[34903 140 272 119] [1809 174 149 52] [1436 86 502 53] [1643 61 165 113]]
Swimming Dataset	 Confusion Matrix: [[3125 60 51 49] [112 10 16 6] [64 15 77 22] [89 12 31 17]]	 Confusion Matrix: [[3198 0 87 0] [121 0 23 0] [70 0 108 0] [107 0 42 0]]	 Confusion Matrix: [[3217 3 54 11] [130 0 13 1] [89 2 84 3] [112 1 31 5]]
Gymnastics Dataset	 Confusion Matrix: [[3396 7 19 7] [77 2 2 2] [59 2 10 3] [59 2 7 1]]	 Confusion Matrix: [[3429 0 0 0] [83 0 0 0] [74 0 0 0] [69 0 0 0]]	 Confusion Matrix: [[3425 0 3 1] [81 1 1 0] [70 1 3 0] [64 0 5 0]]
Diving Dataset	 Confusion Matrix: [[308 9 6 4] [16 3 2 3] [5 3 10 4] [13 5 8 4]]	 Confusion Matrix: [[324 0 3 0] [21 0 3 0] [6 0 16 0] [22 0 8 0]]	 Confusion Matrix: [[326 0 1 0] [23 0 1 0] [10 0 12 0] [25 0 5 0]]
Equestrianism Dataset	 Confusion Matrix: [[526 19 14 18] [31 1 1 1] [18 2 8 3] [29 4 8 0]]	 Confusion Matrix: [[577 0 0 0] [34 0 0 0] [31 0 0 0] [41 0 0 0]]	 Confusion Matrix: [[573 0 4 0] [32 1 1 0] [28 1 2 0] [38 0 3 0]]
Athletics Dataset	 Confusion Matrix: [[5565 66 50 69] [199 9 18 7] [190 6 51 17] [183 16 24 5]]	 Confusion Matrix: [[5750 0 0 0] [233 0 0 0] [264 0 0 0] [228 0 0 0]]	 Confusion Matrix: [[5723 7 18 2] [228 0 5 0] [247 0 14 3] [221 1 6 0]]

Classification Reports & Accuracy Scores

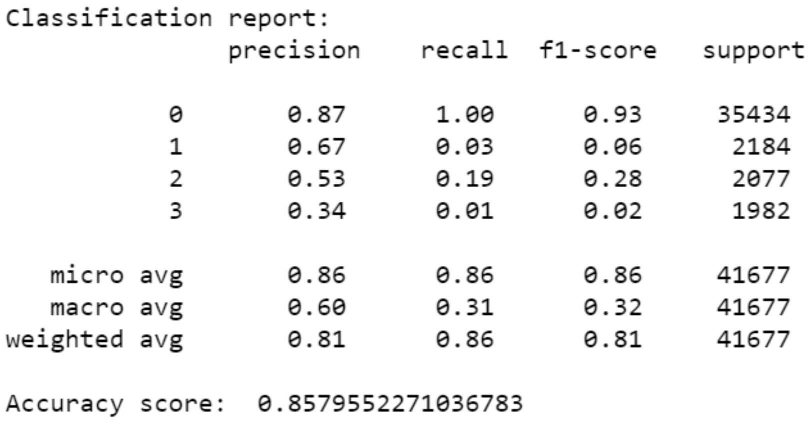
Full Summer Olympics Dataset

- Baseline accuracy score: 0.7293711159632411
- Random Forest Classifier



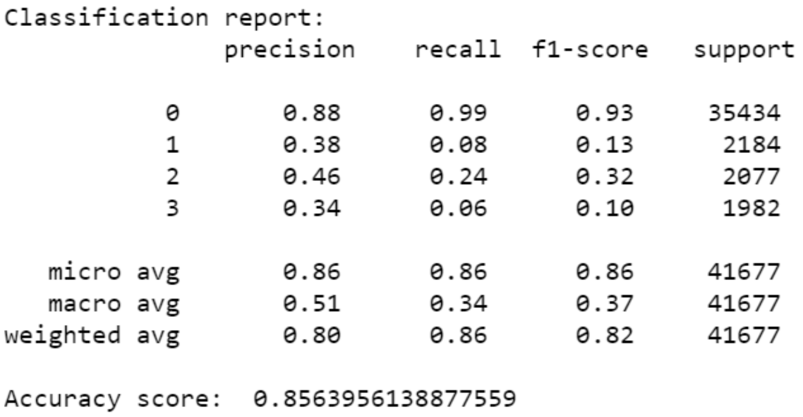
```
Classification report:
              precision    recall  f1-score   support

           0       0.89      0.96      0.92     35434
           1       0.21      0.09      0.12      2184
           2       0.42      0.28      0.34      2077
           3       0.19      0.09      0.13      1982

   micro avg       0.84      0.84      0.84     41677
   macro avg       0.43      0.36      0.38     41677
weighted avg       0.80      0.84      0.81     41677

Accuracy score:  0.8409434460253857
```

- Support Vector Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.87      1.00      0.93     35434
           1       0.67      0.03      0.06      2184
           2       0.53      0.19      0.28      2077
           3       0.34      0.01      0.02      1982

   micro avg       0.86      0.86      0.86     41677
   macro avg       0.60      0.31      0.32     41677
weighted avg       0.81      0.86      0.81     41677

Accuracy score:  0.8579552271036783
```

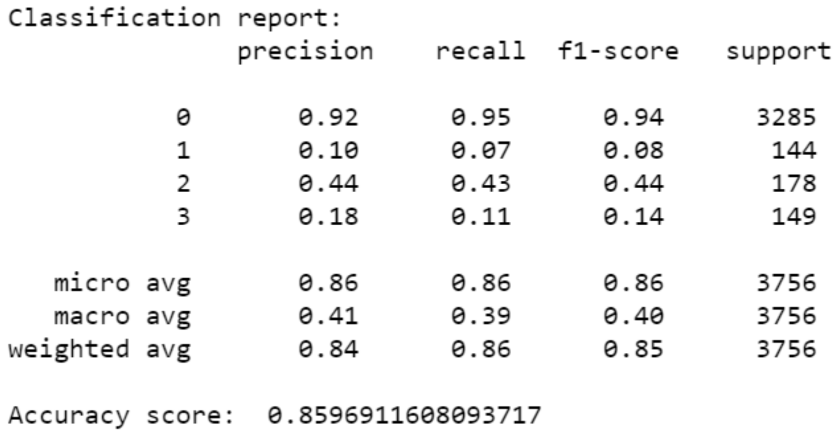
- K-Nearest Neighbors Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.88      0.99      0.93     35434
           1       0.38      0.08      0.13      2184
           2       0.46      0.24      0.32      2077
           3       0.34      0.06      0.10      1982

   micro avg       0.86      0.86      0.86     41677
   macro avg       0.51      0.34      0.37     41677
weighted avg       0.80      0.86      0.82     41677

Accuracy score:  0.8563956138877559
```

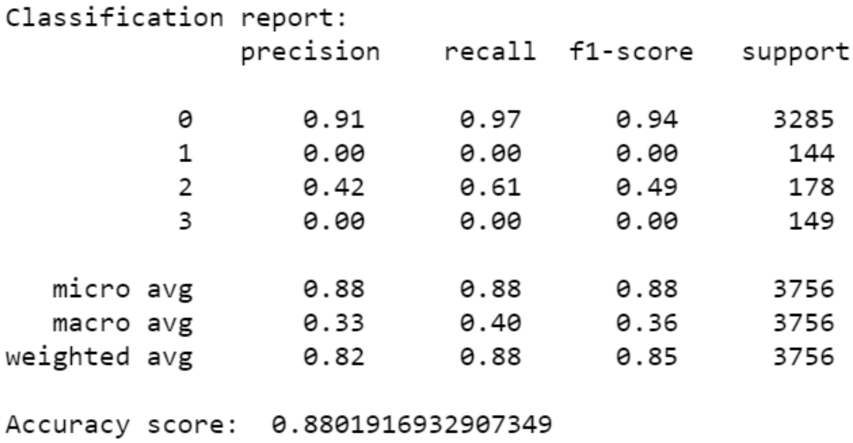
Swimming

- Baseline accuracy score: 0.7667731629392971
- Random Forest Classifier



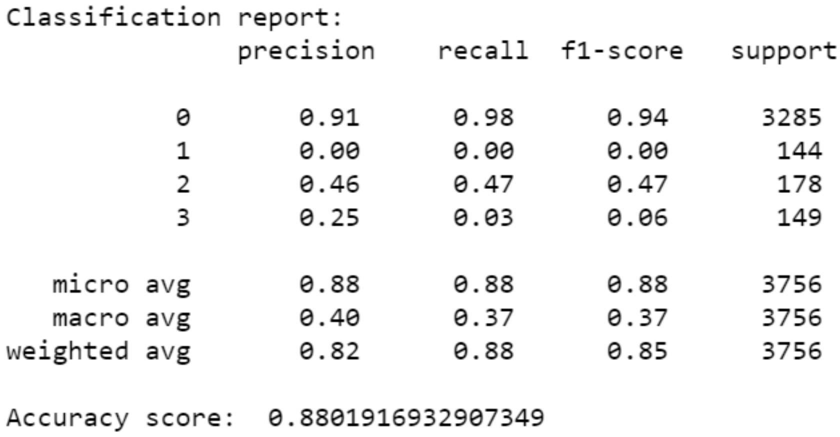
```
Classification report:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94      3285
           1       0.10      0.07      0.08       144
           2       0.44      0.43      0.44       178
           3       0.18      0.11      0.14       149

   micro avg       0.86      0.86      0.86      3756
   macro avg       0.41      0.39      0.40      3756
weighted avg       0.84      0.86      0.85      3756

Accuracy score:  0.8596911608093717
```

- Support Vector Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94      3285
           1       0.00      0.00      0.00       144
           2       0.42      0.61      0.49       178
           3       0.00      0.00      0.00       149

   micro avg       0.88      0.88      0.88      3756
   macro avg       0.33      0.40      0.36      3756
weighted avg       0.82      0.88      0.85      3756

Accuracy score:  0.8801916932907349
```

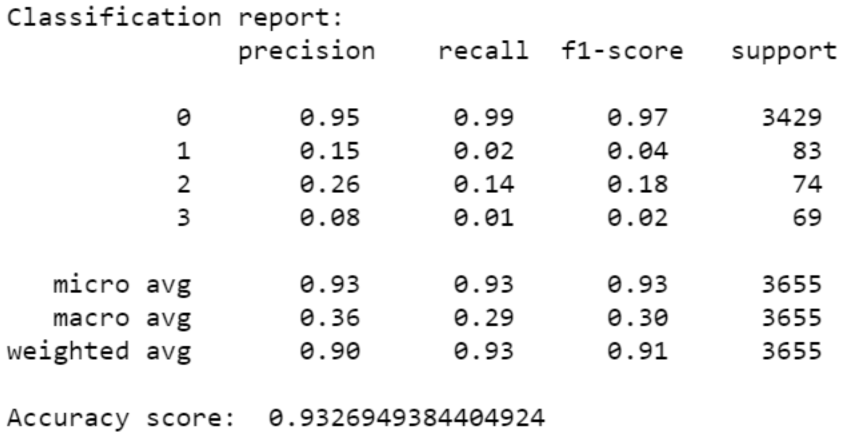
- K-Nearest Neighbors Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94      3285
           1       0.00      0.00      0.00       144
           2       0.46      0.47      0.47       178
           3       0.25      0.03      0.06       149

   micro avg       0.88      0.88      0.88      3756
   macro avg       0.40      0.37      0.37      3756
weighted avg       0.82      0.88      0.85      3756

Accuracy score:  0.8801916932907349
```

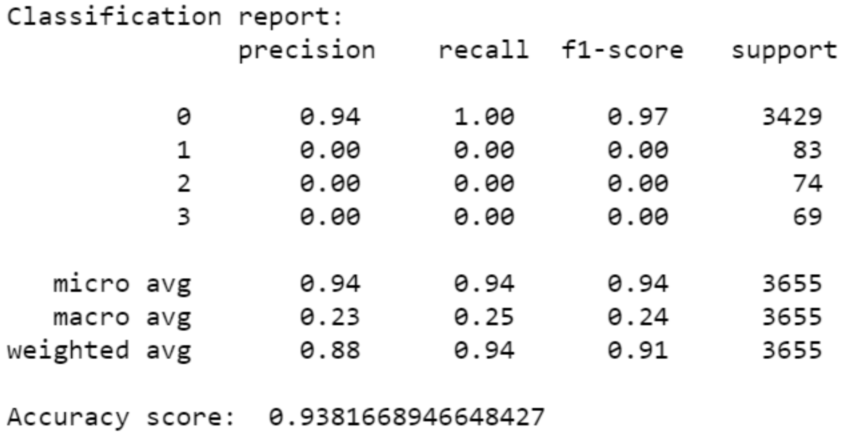
Gymnastics

- Baseline accuracy score: 0.8774281805745554
- Random Forest Classifier



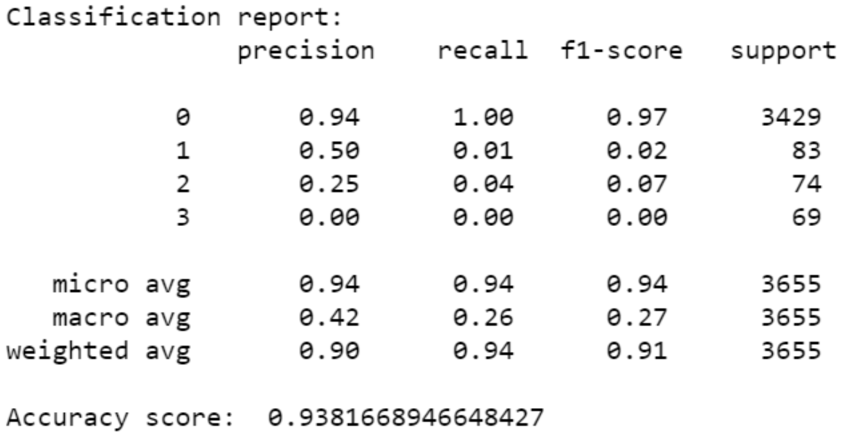
```
Classification report:
              precision    recall  f1-score   support

           0       0.95      0.99      0.97      3429
           1       0.15      0.02      0.04        83
           2       0.26      0.14      0.18        74
           3       0.08      0.01      0.02        69

   micro avg       0.93      0.93      0.93      3655
   macro avg       0.36      0.29      0.30      3655
weighted avg       0.90      0.93      0.91      3655

Accuracy score:  0.9326949384404924
```

- Support Vector Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      3429
           1       0.00      0.00      0.00        83
           2       0.00      0.00      0.00        74
           3       0.00      0.00      0.00        69

   micro avg       0.94      0.94      0.94      3655
   macro avg       0.23      0.25      0.24      3655
weighted avg       0.88      0.94      0.91      3655

Accuracy score:  0.9381668946648427
```

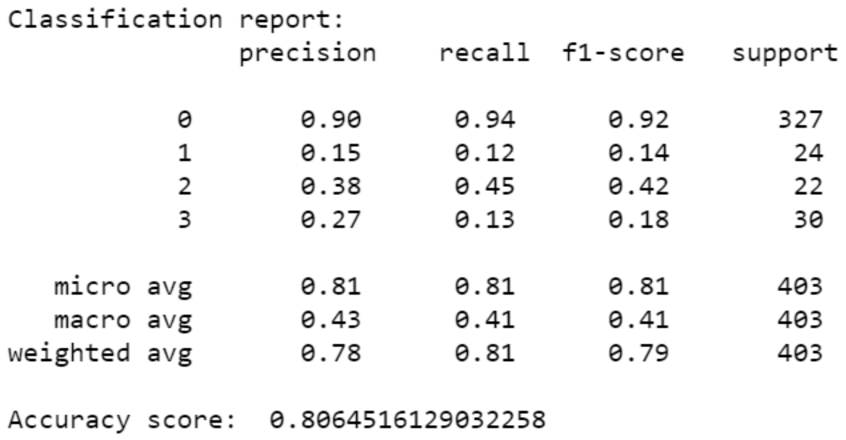
- K-Nearest Neighbors Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      3429
           1       0.50      0.01      0.02        83
           2       0.25      0.04      0.07        74
           3       0.00      0.00      0.00        69

   micro avg       0.94      0.94      0.94      3655
   macro avg       0.42      0.26      0.27      3655
weighted avg       0.90      0.94      0.91      3655

Accuracy score:  0.9381668946648427
```

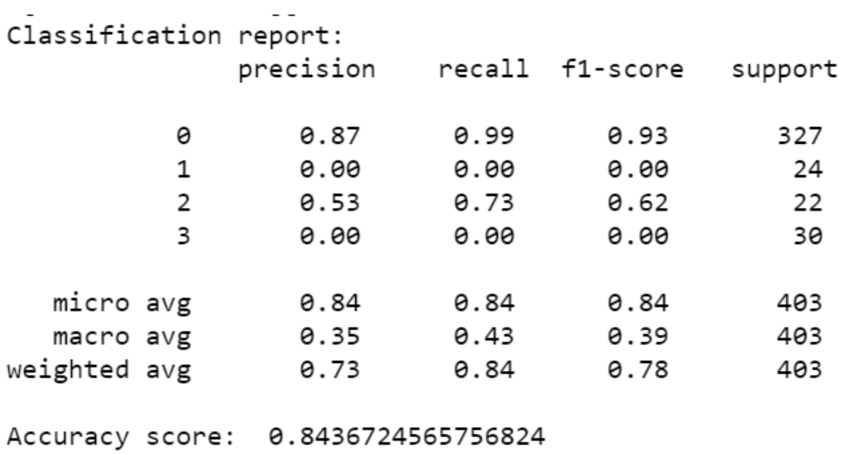
Diving

- Baseline Accuracy Score: 0.6823821339950372
- Random Forest Classifier



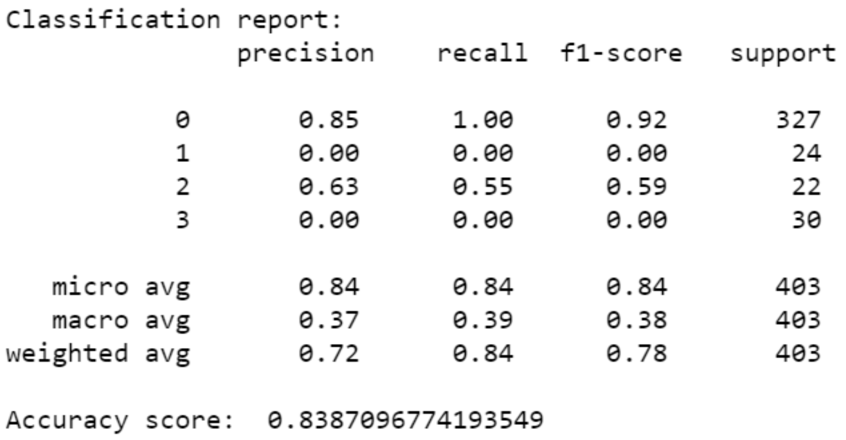
```
Classification report:
              precision    recall  f1-score   support

           0       0.90      0.94      0.92       327
           1       0.15      0.12      0.14        24
           2       0.38      0.45      0.42        22
           3       0.27      0.13      0.18        30

   micro avg       0.81      0.81      0.81       403
   macro avg       0.43      0.41      0.41       403
weighted avg       0.78      0.81      0.79       403

Accuracy score:  0.8064516129032258
```

- Support Vector Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.87      0.99      0.93       327
           1       0.00      0.00      0.00        24
           2       0.53      0.73      0.62        22
           3       0.00      0.00      0.00        30

   micro avg       0.84      0.84      0.84       403
   macro avg       0.35      0.43      0.39       403
weighted avg       0.73      0.84      0.78       403

Accuracy score:  0.8436724565756824
```

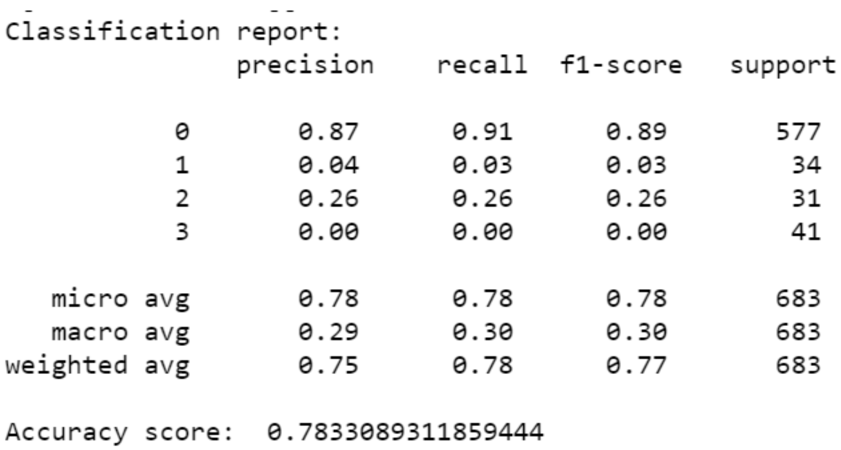
- K-Nearest Neighbors Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.85      1.00      0.92       327
           1       0.00      0.00      0.00        24
           2       0.63      0.55      0.59        22
           3       0.00      0.00      0.00        30

   micro avg       0.84      0.84      0.84       403
   macro avg       0.37      0.39      0.38       403
weighted avg       0.72      0.84      0.78       403

Accuracy score:  0.8387096774193549
```

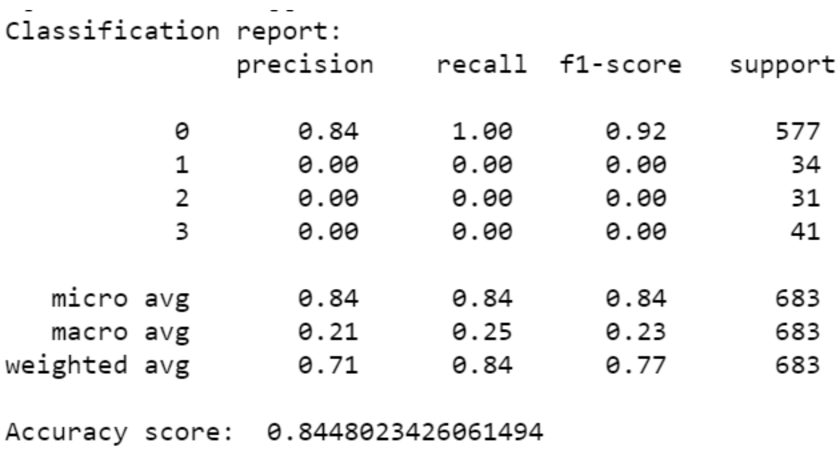
Equestrianism

- Baseline Accuracy Score: 0.7232796486090776
- Random Forest Classifier



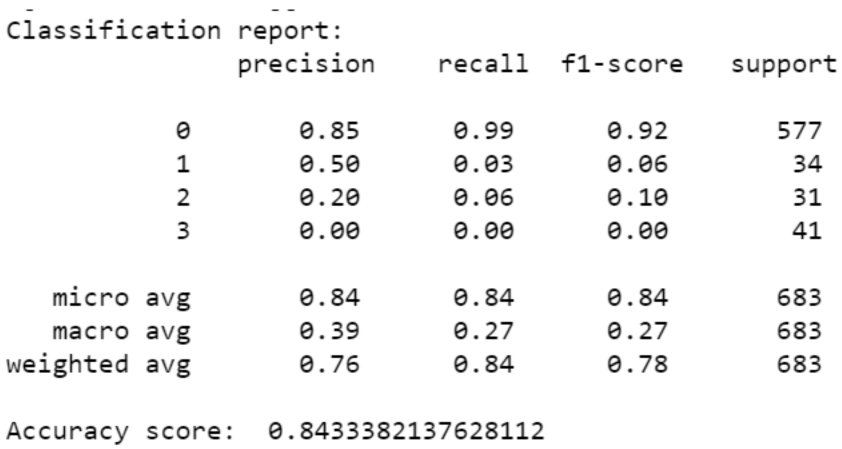
```
Classification report:
                precision    recall  f1-score   support

           0       0.87      0.91      0.89       577
           1       0.04      0.03      0.03        34
           2       0.26      0.26      0.26        31
           3       0.00      0.00      0.00        41

   micro avg       0.78      0.78      0.78       683
   macro avg       0.29      0.30      0.30       683
weighted avg       0.75      0.78      0.77       683

Accuracy score:  0.7833089311859444
```

- Support Vector Classifier



```
Classification report:
                precision    recall  f1-score   support

           0       0.84      1.00      0.92       577
           1       0.00      0.00      0.00        34
           2       0.00      0.00      0.00        31
           3       0.00      0.00      0.00        41

   micro avg       0.84      0.84      0.84       683
   macro avg       0.21      0.25      0.23       683
weighted avg       0.71      0.84      0.77       683

Accuracy score:  0.8448023426061494
```

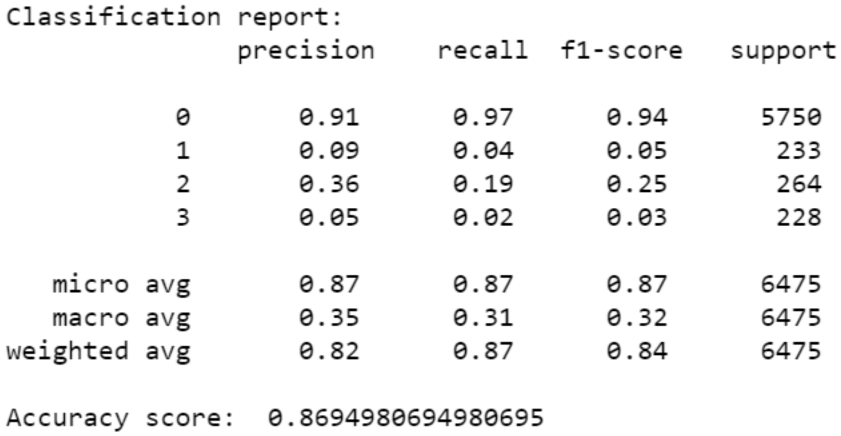
- K-Nearest Neighbors Classifier



```
Classification report:
                precision    recall  f1-score   support

           0       0.85      0.99      0.92       577
           1       0.50      0.03      0.06        34
           2       0.20      0.06      0.10        31
           3       0.00      0.00      0.00        41

   micro avg       0.84      0.84      0.84       683
   macro avg       0.39      0.27      0.27       683
weighted avg       0.76      0.84      0.78       683

Accuracy score:  0.8433382137628112
```

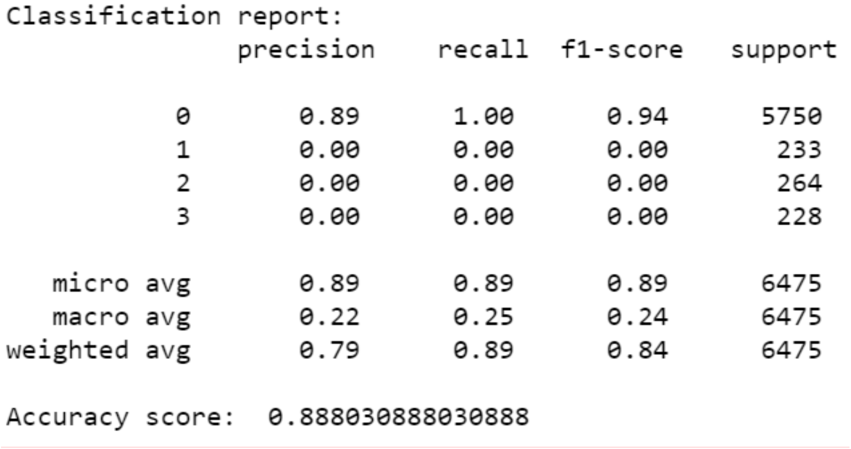
Athletics

- Baseline Accuracy Score: 0.7913513513513514
- Random Forest Classifier



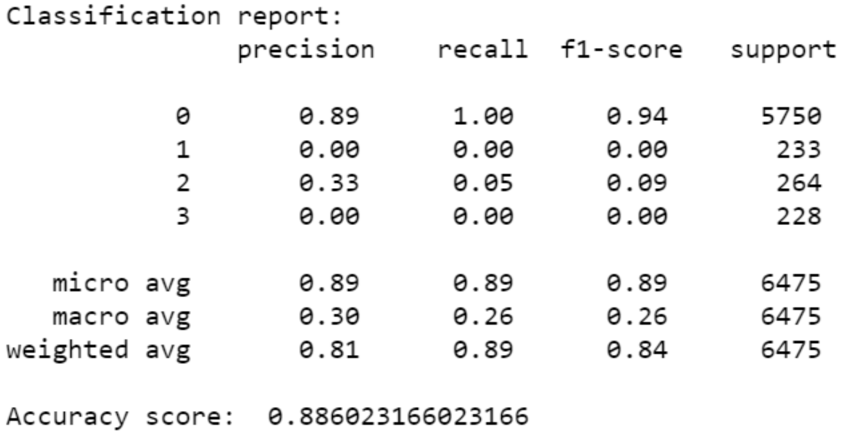
```
Classification report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94      5750
           1       0.09      0.04      0.05       233
           2       0.36      0.19      0.25       264
           3       0.05      0.02      0.03       228

   micro avg       0.87      0.87      0.87      6475
   macro avg       0.35      0.31      0.32      6475
weighted avg       0.82      0.87      0.84      6475

Accuracy score:  0.8694980694980695
```

- Support Vector Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      5750
           1       0.00      0.00      0.00       233
           2       0.00      0.00      0.00       264
           3       0.00      0.00      0.00       228

   micro avg       0.89      0.89      0.89      6475
   macro avg       0.22      0.25      0.24      6475
weighted avg       0.79      0.89      0.84      6475

Accuracy score:  0.888030888030888
```

- K-Nearest Neighbors Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94      5750
           1       0.00      0.00      0.00       233
           2       0.33      0.05      0.09       264
           3       0.00      0.00      0.00       228

   micro avg       0.89      0.89      0.89      6475
   macro avg       0.30      0.26      0.26      6475
weighted avg       0.81      0.89      0.84      6475

Accuracy score:  0.886023166023166
```