Final Project: Rio 2016 Olympic Games

Omar El Nahhas, 214316IV

I. Introduction

This project will be defended online on January 17th 2022 through Microsoft Teams. The goal of this assignment is to apply several Data Mining methods on a real-world case. For this project, the participants of the Rio 2016 Olympic Games are being observed with the goal of analytically deriving useful insights into the athletes and their sport. The Rio 2016 Olympic Games are the last Summer Olympics that were held pre-pandemic. The dataset has been collected by Matt Riggot from Lund University based on official sources of the Olympics [2].

II. Methodology

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is used as a framework for the work flow of this project [4]. CRISP-DM consists of six stages, which are enumerated in the following sub-sections.

A. Business Understanding

Before digging into the data and trying to find valuable insights, the concept of what is valuable should be defined in the scope of this project. In other words, why is this data analysed and what are the questions which need to be answered by this analysis? The 'why' of this project's topic is fuelled by interest in athletics in general. More specifically, what are typical characteristics of winners, how much different athletes from different sports are physically similar to each other, and which sports are more skill-based, rather than physically-based sports? In order to get answers to these questions, three different Data Mining methods should be applied. As a result, the analytical questions are answered and the final project requirements are satisfied.

B. Data Understanding

This dataset includes the official statistics on the 11.538 athletes (6.333 men and 5.205 women) and 306 events at the 2016 Olympic Games in Rio de Janeiro. The athlete data is stored in athletes.csv; one athlete per row, and eleven columns. Empty cells are NA values. The event data is stored in events.csv; one event per row, and six columns. There are 306 events across 50 disciplines in 28 sports; 161 for men, 136 for women, and 9 mixed sex. A detailed overview of the dataset, its variables and their limits can be found in the GitLab repository's ReadME file and is excluded from this report due to size limitations. This overview describes the scope of every qualitative feature. Using descriptive statistics, the 6 quantitative features of the athletes dataset are described in Table I.

TABLE I: Summary of the quantitative features of the athletes dataset.

Measure	Age	Height	Weight	Gold	Silver	Bronze
Min.	20.00	1.21	31.00	0.00	0.00	0.00
1st Qu.	29.00	1.69	60.00	0.00	0.00	0.00
Median	32.00	1.76	70.00	0.00	0.00	0.00
Mean	32.70	1.77	72.07	0.06	0.06	0.06
3rd Qu.	36.00	1.84	81.00	0.00	0.00	0.00
Max.	68.00	2.18	170.00	5.00	2.00	2.00
NA's	0	330	659	0	0	0

C. Data Preparation

The first operation to prepare the data was to remove features which were not going to be useful for later analysis (based on experience). As a result, the athletes' ID, full name and info (containing human written qualitative text) were removed. Secondly, the date of birth entries were reformatted, moving from a 'yyyy-mm-dd' format and calculating the age, resulting in a simple integer. As a result of the descriptive statistics in Table I, it shows that the athlete dataset contains quite some NA values. It was considered to impute these values, however it was chosen to rather remove the NA's to keep the data as close to reality as possible. This decision was made after it was observed that the dataset contains no bias in the NA, meaning that there was no clear pattern in the data that was missing (for example, due to certain nationalities or sports). As one of the questions is related to the characteristics of a winner, it is not so important which medals of which colour an athlete has, but rather if the athlete has *any* medals. Therefore, the gold, silver and bronze features are merged into a boolean feature named 'podium', which is 1 if any medals have been obtained, and 0 if no medals have been obtained by the athlete. Two separate tables were created, one where everything was a factor (with age, height, and weight discretized) and one with mixed factor/numerical variables, with the aforementioned features as numerical rather than discretized factors. This is done for experimentation purposes during the modelling phase, to see which method works best given this data. The splitting of male/female athletes into two separate tables was initially considered, but was not necessary for the purpose of this project's goals and has therefore not been executed. Finally, as a result of the data understanding phase, the events.csv data is removed as it does not hold any valuable data for the purpose of this project. Only the athletes data processed according to the above-mentioned steps is used in the modelling phase. The feature selection is performed based on experience in the domain and through knowledge of good practice in the data mining field, rather than using numerical methods. This decision was made because of the nature of the data, which does not require numerical methods to perform feature selection.

D. Modelling

Three separate Data Mining methods are applied to find answers to the questions. Firstly, classification is performed using decision trees to figure out which characteristics (features) of an athlete makes them a winner. Secondly, clustering is done with k-means and hierarchical clustering to find overlap in athlete's physical features from different sports. Finally, outlier detection using the local outlier factor is done to find which sports are more physical-based or skill-based when compared to each other. To slightly reduce the complexity of the analysis, 6 out of 28 sports for each method are selected based on suitability for the research question.

1) Classification: The goal of classification is to find out what typical characteristics of a winner are. I.e., can a winner be classified based on certain features, and if so, which features? The notion of required explainability for this type of question indicates that the decision tree approach would fit well for this classification problem. Initially, the sports that were selected were basketball, handball, volleyball, football, hockey and rugby. They were chosen due to their similarities in sports (based on personal experience), as well as having similar representation in the dataset to avoid bias. This resulted in a high 90%+ accuracy on the testing data. When diving deeper into the results, a clear bias was found in the nature of the sport. Namely, in a team sport of the same sex, either your whole team loses, or your whole team wins. Therefore, that relationship was uncovered quickly and debunks the high obtained accuracy. Moving forward, the sports were chosen more carefully, and were assured to be individually awarded sports. The new sports set were gymnastics, fencing, weightlifting, shooting, judo and cycling. Although there can be representatives from the same country, in these sports there are only podium places for individuals. The full data set for these sports contains 686 samples of athletes. The data is split into 80/20 (548/138) for the training and testing data respectively. Descriptive statistics have been performed to ensure that the test set has a similar underlying distribution compared to the training set. The confusion matrix of the decision tree on the test set is seen in Table II, having an accuracy of approximately 70% and F1-score of 70% as well. The accuracy is compared with a naive classifier (in order to get the No Information Rate), which the built model should exceed in accuracy in order to prove that the accuracy is statistically significant. This decision tree classifier has a p-value of 2.5e-6 and is statistically significant.

TABLE II: Decision tree confusion matrix.

		Predicted	
		[0]	**[1]**
True	**Losers [0]**	49	22
	Winners [1]	20	47

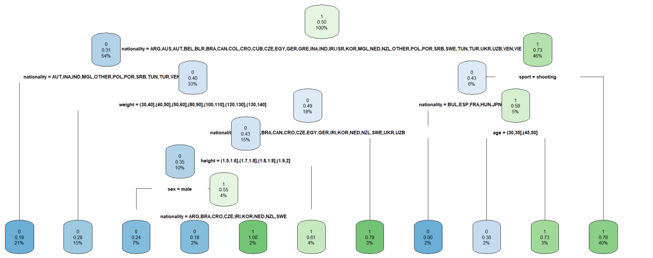
1
0.50
100%
0
0.31
54%
nationality = ARG,AUS,AUT,BEL,BLR,BRA,CAN,COL,CRO,CUB,CZE,EGY,GER,GRE,INA,IND,IRI,ISR,KOR,MGL,NED,NZL,OTHER,POL,POR,SRB,SWE,TUN,TUR,UKR,UZB,VEN,VIE
1
0.73
46%
nationality = AUT,INA,IND,MGL,OTHER,POL,POR,SRB,TUN,TUR,VEN
0
0.40
33%
0
0.43
6%
sport = shooting
weight = (30,40],(40,50],(50,60],(80,90],(100,110],(120,130],(130,140]
0
0.49
18%
nationality = BUL,ESP,FRA,HUN,JPN
1
0.58
5%
nationalit
0
0.43
15%
,BRA,CAN,CRO,CZE,EGY,GER,IRI,KOR,NED,NZL,SWE,UKR,UZB
age = (30,35],(45,50]
0
0.35
10%
height = (1.5,1.6],(1.7,1.8],(1.8,1.9],(1.9,2]
sex = male
1
0.55
4%
nationality = ARG,BRA,CRO,CZE,IRI,KOR,NED,NZL,SWE
0
0.18
21%
0
0.28
15%
0
0.24
7%
0
0.18
2%
1
1.00
2%
1
0.61
4%
1
0.79
3%
0
0.00
2%
0
0.36
2%
1
0.73
3%
1
0.78
40%

Fig. 1: Decision tree with Nationality as most important variable, based on mean decrease in Gini. Nationality (71), Sport (18), Weight (12), Age (8), Height (3), Sex (3).

2) Clustering: The goal of clustering is to find out how much different athletes from different sports are physically similar to each other. If the clustering can make 6 perfect clusters, every sport has very specific physical characteristics. If this is not the case, it will be observed which particular sports have overlap by comparing the cluster labels with the ground-truth. Therefore, the clustering is performed with 6 clusters, as there are a total of 6 sports under analysis: cycling, fencing, gymnastics, judo, shooting and weightlifting. K-means and hierarchical clustering are used to create the clusters. In order to accomplish the goal, the features of the athlete's height, weight and age are used to find physical similarities. The 3D plot in Fig. 2 shows the physical properties of the athletes per sport. 6-Means clustering resulted in a 16% accuracy, while 6-Hierarchical clustering resulted in a 20% accuracy. Due to similar performance and visually indistinguishable results, only the 6-Hierarchical clustering plot (Fig. 3) and confusion matrix (Table III) are shown. Accuracy in this context is defined as overlap between the ground-truth 6 sports and the given cluster as a result of 6-Hierarchical clustering.

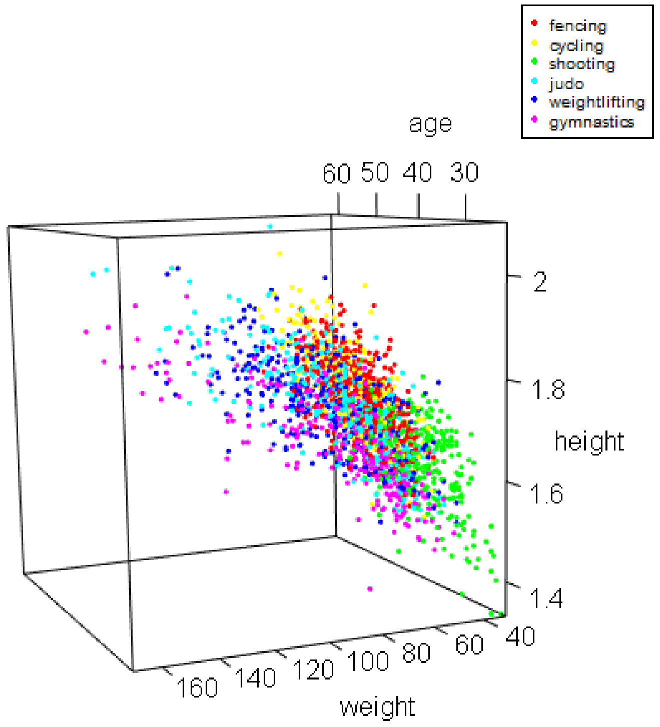

Fig. 2: Ground-truth 3D plot of athlete's physical properties.

3) Outlier analysis: The goal of classification is to find out which sports are more skill-based, rather than physically-based. Based on the nature of this question, 6

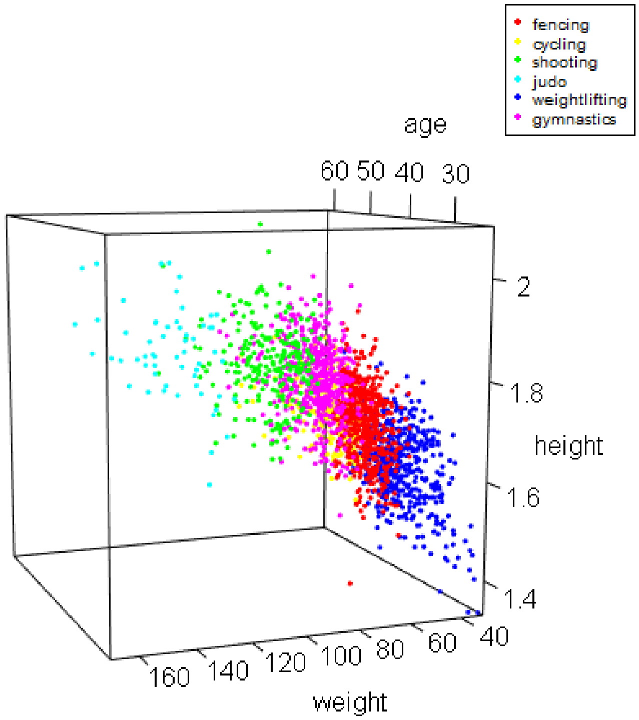

Fig. 3: 6-Hierarchical clustering 3D plot of athlete's physical properties.

TABLE III: 6-Hierarchical clustering confusion matrix.

		Predicted					
		[1]	**[2]**	**[3**	**[4]**	**[5]**	**[6]**
True	**cycling [1]**	275	203	20	0	0	4
	fencing [2]	104	128	10	0	0	3
	gymnastics [3]	154	30	0	0	0	135
	judo [4]	140	133	71	20	5	8
	shooting [5]	140	164	59	6	1	10
	weightlifting [6]	87	88	43	13	17	10

distinct sports were chosen. Some sports are stereotypically known as physically intense, while others require more skill. Consequently, gymnastics, canoe, weightlifting, shooting, basketball and table tennis were chosen as the 6 sports for this analysis. By combining these sports with features regarding only the height and weight, the most frequently occurring physical properties will be seen as the dominant set. Outliers in this case would be the athletes who are not in a 'normal' range of physical properties, and are then assumed to be using more of their skill rather than their physical properties. Applying the local outlier factor (k = 3) on the height and weight of the athletes in the aforementioned sports results in the density plot of the local outlier factors in Fig. 4, where all points beyond the red vertical line are labeled as outliers.

Using the frequency of the outliers per particular sport and dividing it by the total occurences of the sport, the relative outliers frequency of the sport's physicality importance is found. The higher this percentage, the higher the ratio of outliers within that sport. A high outlier ratio indicates that a significant amount of athletes within that Olympic sport deviate from the average physical features of the other Olympic sports. A deviation in physical features can be both at the upper tail (elite) and lower tail (very poor). However, in the context of this analysis, a deviation is seen as below average, meaning that the athlete within that sport is more reliant on skill versus their physical dominance. The outliers of sport physicality importance per sport is seen in Fig. 5.

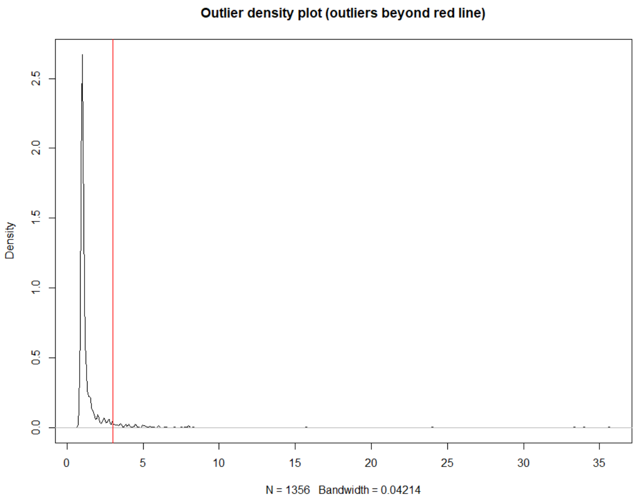

Fig. 4: Density plot of the local outlier factors.

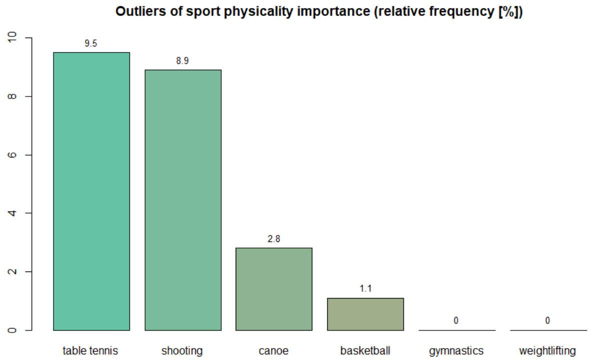

Fig. 5: Barplot of relative outliers regarding physical importance in the sport [%].

E. Evaluation

The decision tree depicted in Fig. 1 gives high importance to the nationality of the athlete, which could be a logical consequence of some countries being represented by more athletes than others. In turn, countries being represented by more athletes than others means that there are better facilities for those athletes based on their country of origin, and therefore assumingly perform better on the Olympics. The decision tree is trained again without taking into account the nationality as feature, resulting in only a 51% accuracy and 55% F1-score. This decision tree classifier without the nationality feature has a p-value of 0.94 and is statistically insignificant. The confusion matrix of the decision tree in

Table II showing an accuracy and F1-score of 70% on individual sports indicates that **the decision tree is capable of classifying what typical characteristics of winners are**, giving the most importance to the athlete's nationality, sport, weight, age, height and sex, respectively.

The confusion matrix in Table III is quite literally showing the confusion of this clustering method to predict a sport based on physical features of the athlete. For example, gymnastics and weightlifting on the right side of the table seem to have quite the overlap. This makes a lot of sense when observing preparation; many Olympic sports (such as gymnastics) perform weightlifting as part of their training. Concluding, **hierarchical clustering shows that the physical similarity between athletes of different Olympic sports is very high** within this dataset, as the clusters cannot tell them apart with a high accuracy. This makes sense as these are elite athletes with physical features which are cross-applicable in different Olympic sports.

The barplot in Fig. 5 shows a high relative frequency to sports like table tennis and shooting, while gymnastics and weightlifting have a relative frequency of 0. Table tennis and shooting have many physically outlying athletes, indicating that this Olympic sport is more skill-based than physically based relative to the other Olympic sports it is compared to. On the contrary, gymnastics and weightlifting are relatively more physically-based than skill-based when compared with the other Olympic sports in this analysis. Basketball and canoe are inbetween physical and skill-based, although the low relative frequency indicates they tend to belong to the more physically-based Olympic sports according to this analysis. It can be concluded that **the local outlier factor can successfully distinguish between skill-based and physically-based sports** when compared to each other (within the scope of this chosen Olympic sports sample). Do note that in real life all Olympic sports require a high level of skill and physical performance, and that the outlier analysis is merely comparing between a selection of Olympic sports.

F. Deployment

Due to the nature of this project and the data, the actual deployment of the model is out of scope. The decision tree model in the shape presented in this report has a total size of approximately 135 kB without any optimization or compression steps involved. As a result, the model would be small enough to be ran on an embedded micro-controller, such as a board from the STM32F4 family [3]. Besides the actual model deployment, the deployment phase also consists out of documenting the modelling process and reporting the conclusions and interpretations of the results. The exact methodology is described in this report and can be observed in the code and comments to explain certain operations [1]. In overall review of the project, the analytical research questions mentioned in the business understanding phase have been successfully answered during the evaluation phase by applying several Data Mining methods, and the results have been evaluated and interpreted in this report. Therefore, this project has been executed and concluded successfully. In terms of improvement for further research, different sets of sports should be tested per method in order to get a global overview of the results across all Olympics sports, rather than just a manually selected set. Furthermore, no numerical methods have been used for feature selection. Rather, domain knowledge, choice of modelling methods and the fact that there were little features to choose from, presumably removed the need for feature selection through numerical methods. Further research should review the modelling results of experience based feature selection in this report with modelling results using numerical method based feature selection.

REFERENCES

[1] Omar El Nahhas. data-mining-iti8730-finalproject. https://gitlab.cs.ttu.ee/omarel/data-mining-iti8730-finalproject, 2022.

[2] Matt Riggott. rio2016. https://github.com/flother/rio2016, 2016.

[3] STMicroelectronics. *The STM32F4 family*. Available at https://www.st.com/resource/en/datasheet/dm00037051.pdf.

[4] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Springer-Verlag London, UK, 2000.