Prediction Olympic Athletes’ performance using Anthropometric Measurements, Age, Gender and Nationality Data

**Assessment Cover Page¶**

**Student Full Name: Ilia Grishkin¶**

**Student Number: 2024139¶**

**Module Title: Strategic Thinking¶**

**Assessment Title: CA 2 Capstone Report¶**

**Assessment Due Date: 19th May 2024¶**

**Date of Submission: 19th May 2024¶**

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Project management methodology 4

Stages and Tasks 4

Project timeline 6

Introduction 7

Importance 8

Improvements in training and selection 8

Equality and fairness 9

Sport medicine 10

Objectives 10

Cross-Disciplinary Characteristics 10

Predictability Evaluation 10

Data source and Methods 10

EDA 11

Data Preparation 11

Machine Learning Models 12

Ethical consideration 13

Privacy and Data Protection 13

Transparency 13

Geographic bias 13

Competitive advantage 13

EDA 14

Summary Statistics 14

Distributions 15

Correlations 15

Scatter plots 16

Height and Weight vs. Medal 16

Height and Weight vs. Gender 17

Feature importance 17

Preparaton 18

Skewed data 18

Log Transformation 19

Missing values 20

1.1. Drop rows with missing values 21

1.2. Fill missing values with the mean 21

1.3. Fill missing values with ML models 21

Feature Engineering 22

Drop Irreverent features 22

Create new feature BMI 22

Label target variable 22

Encoding 22

Standart 22

PCA 22

Cross-Disciplinary Characteristics 24

Introduction 24

Modeling 26

Evaluation 28

Predictability Evaluation 30

Evaluation models 30

Confusion matrix 30

ROC-AUC Curve 31

Overfitting 31

Conclusion 33

Recommendation for future work 34

References 35

## Project management methodology

In our research, we will follow the CRISP-DM methodology. As was pointed out by Kurgan L. Musilek P (2006), the cross-industry process has already become a standard for the vast majority of industrial research and offers a very efficient research strategy.

### Stages and Tasks

Field understanding

In this section, we are going to analyse the ground of the problem the industry is currently facing, data source and plan our research. Additionally, we will take a look at which potential bias is inside of the data set and which precautions we should take to make results as objective as possible and prevent any harm to all people or organisations somehow related to the dataset and this research.

* **Tasks:**
  + Literature review
  + Data resources
  + Ethical consideration
  + Objectives
  + Methods

Data understanding

In this section, we will explore the dataset to set appropriate methods for preparation and modelling.

* **Tasks:**
  + Statistic summary
  + Distributions
  + Correlations
  + Feature importance
  + PCA

Data Preparation

This section aims to prepare the dataset for further analysis.

* **Tasks:**
  + Skewed features
  + Missing values
  + Feature engineering
  + Encoding
  + Splitting data
  + Scaling

Note: We will apply a few methods for each task to understand which method fits better for the dataset.

Data modeling

This section will be built based on the results from the first section after we analyse the industry background, current issues and studies which already have been done related to the issues. However, we can define the general process.

* **Tasks:**
  + Define ML models
  + Define Hyperparameters
  + Fit models
  + Define and fit DL model

Note: Important moment we need to keep in mind that the final report must briefly display the process of the model's selection. So. we need to create data frames for storing models' results.

Data evaluation

This section includes several evaluation methods and compares models' performance. The section aims to select the best model and hyperparameters. Additionally, we will need to extract information from the models' performance to form a conclusion.

* **Tasks:**
  + Compare accuracy scores
  + ROC-Auc scores
  + Confusion matrix

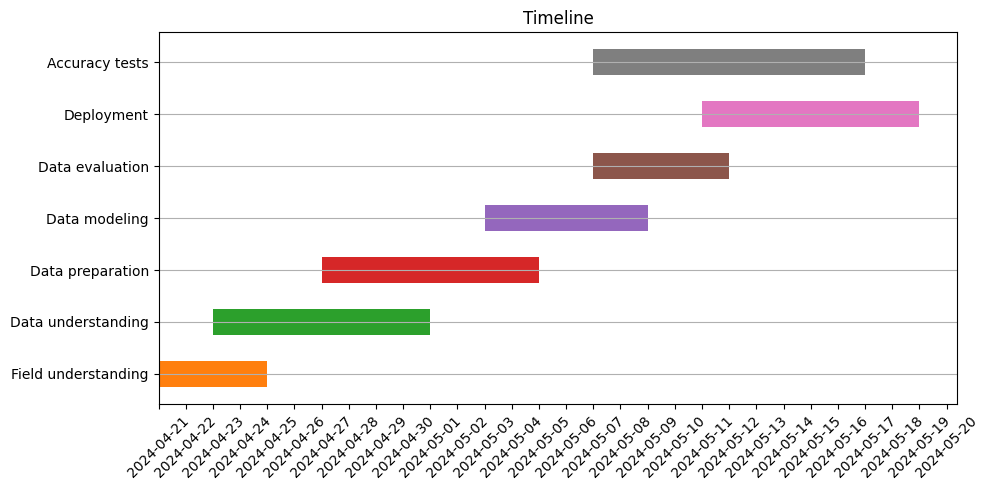
Deployment

This section is going to serve several proposes such as forming a conclusion, tracking accurate results for different combinations of methods, tracking report development, and forming a final report.

* **Tasks:**
  + Conclusion
  + Form final reports
  + Accuracy log
  + Project log

Note: The final report requires the total words used, so we need to add a column in our project log where we are going to track how many words we already have used.

### Project timeline



# Introduction

The vast majority of athletes consider the Olympic Games as the top of their sporting and physical achievements. A large number of advancements and discoveries can be associated with the drive to succeed in the Olympic games.

Olympic athletes represent the highest level of human physical abilities, as highlighted by Borms and Hebbelinck (1984) in their study. They assert that athletes competing in the Olympics or world-class athletes show the ideal composition of genetic predispositions and environmental factors, and as a result, they can reach a peak of their performance. Theoretically, those athletes who stand out the most in their specific events, possess the optimal physical structure for those types of physical activity, as discussed by Carter (1985).

Within the last century, the Olympic Games have been experiencing an enormous upward in competitiveness. In this evolving stage, coaches and support teams should collaborate closely with data scientists to create and develop predictive models for athletes' performance and identify outliers, patterns and trends. This collaboration is essential not only for improving current athletes' metrics but also for the selection of young talents for specific events. Numerous studies, including those by Cuk and Karácsony (2002), Arkaev and Suchilin (2004), pointed out the significance of anthropometric characteristics in influencing the success of athletes in achieving their goals.

The significance of elite athletes' age in their performance also must be mentioned. Sports researchers acknowledge that even the month of birth compared to other athletes may determine an athlete's likelihood of reaching high performance, a phenomenon known as the relative age effect, as highlighted by J. Musch (2001).

Another vital factor which has a significant impact on athletes’ successes is the nationality or the country they represent. It is very important to consider because each country which shows up at the Olympic Games has its methodologies for many disciplines. Identifying the correlations between countries which provide training and its success may be a key for mutually beneficial exchanges of methods, practices, approaches and databases between counties. Additionally, there is a strong correlation between a country and an athlete's performance directly related to the athlete selection process. As was mentioned by De Bosscher V. (2007), the larger population provides a bigger talent sample for recruitment opportunities for arranging training and competitions.

This study will specifically concentrate on predicting athletic performance by analysing anthropometric characteristics, including weight, height and body mass index. Also, relevant data such as age, gender and nationality for each athlete will be included in the analysis.

# Importance

Using advanced analytical methods for predicting athletes' results in world-class sporting competitions can be highly beneficial for many departments, organisations and teams involved in the sports industry.

### Improvements in training and selection

Sports performance analysis and techniques allow sports scientists, coaches, and athletes to analyze athlete performance objectively. Thus, performance analysis has become an important component of training Donoghue P. O. (2009), and it has a vital role in planning training and competition strategies Michalsik L.B. (2004) and Frantisek T. (2011). Machine learning uses previous experiences to make a connection with the future, and the success of the model is highly related to the characteristics of the dataset Sekeroglu B., Dimililer K. and Tuncal K. In recent years, several experiments using different ML models have been performed in sports.

1. Meayk E. and Unold O. (2011) combined fuzzy logic and machine learning to model swimming training. The classification accuracy of their proposed methodology reached 68.66%.
2. Ofoghi B. et al (2013), implemented machine learning techniques for selecting athletes in cycling as well as for strategic planning. They implemented a statistical approach, K-means and a naive Bayes classifier for inter-omnium analysis, and Bayesian belief networks with discrete optimization techniques for intra-omnium performance analysis.
3. Hore S. and Bhattacharya T. (2018) used naive Bayes, support vector machines (SVMs), multilayer perceptron feed-forward networks, and random forests to build a sustainability model for the National Basketball Association (NBA) players. They concluded that SVMs produced superior results than the other models, with an accuracy of 85.65%.
4. Muazu Musa R. et al. (2019), implemented a variation of k-nearest neighbour and linear regression to classify high-potential archers using physical fitness indicators.
5. Muazu Musa R. et al. (2019) conducted another study on the scouting of high-performance archers. They implemented artificial neural networks and k-nearest neighbour models. They used the selected performance parameters of 50 archers. The obtained results showed that the artificial neural network model achieved a higher accuracy (92%) than the other models.
6. Jesus K. et al. (2018) implemented and compared a linear model and artificial neural network to predict the backstroke start performances of ten male backstroke swimmers. They concluded that the artificial neural network outperformed the linear model.
7. Maanijou R. and Mirroshandel S. A. (2019) proposed a method to predict soccer player rankings based on an expert system and ensemble learning. Twenty features of soccer players were considered, and a comparison was performed by considering benchmark machine learning models such as multilayer perceptron, support vector machines, naive Bayes, logistic regression, etc. They concluded that the proposed method achieved the highest accuracy (60%).
8. Anik A.I. et al. (2018) proposed another method based on feature elimination and machine learning implementation to predict the performances of cricket players. The machine learning step of the proposed method consisted of linear regression and a support vector machine with linear and polynomial kernels. The highest prediction accuracies obtained for batsmen and bowlers were 91.5% and 75.3%, respectively.
9. Zhou Z. et al. (2017) conducted research to predict countermovement jump heights by using machine learning models. Decision trees, random forests, and linear regression models were implemented to train selected features of athletes. Evaluation was performed by considering three metrics; namely, the R2 score, root mean square error (RMSE) and mean absolute error (MAE). The obtained results demonstrated that linear regression outperformed the other considered machine learning models.

### Equality and fairness

The idea of regulating the impact of inequalities upon which individuals have little or no control and for which they cannot be claimed responsible has a long history in ethics and is included in various forms in most ethical theories (Arneson, 2015). A second kind of inequality is linked to what Heinilä (1982) calls system strength: the strength of the material, financial, technological, and scientific resources supporting an athlete or a team. Olympic national medal statistics state illustrate the point as it correlates with the ranking of nations according to gross national product (Flegl & Andrade, 2018).

### Sport medicine

As mentioned in research conducted by Bahr R, Clarsen B, Ekstrand J (2018), Injuries are very frequent issues in individual and team sports and can cause physical and financial problems. Discovering and learning about factors which may cause an injury in sports is a key component of prevention methods. Bahr R, Krosshaug T (2005).

# Objectives

#### Cross-Disciplinary Characteristics

We are going to discover if there is any chance to win a competition for an athlete with a different specialisation. - What is the probability of winning for the average athlete in their sport. - Which disciplines have the most similar requirements for physical characteristics. - Which disciplines have the least similar requirements for physical characteristics. - Compare the probability of winning in both disciplines.

#### Predictability Evaluation

In this section, we are going to check how given data behaves with different ML models and preparation phases. We will try to find out which models give us the most accurate prediction of athlete's performance. - For every stage of the preparation process we will create a few options to find which methods work better. For example, there are many different ways of performing imputations (mean, mode, ML models, dropping). So, we will try some methods for each stage. - Similarly, we will analyse how each selected model works with data. We will set a few different hyperparameters and use GridSearch to find the best tuning for each model - Every combination of preparation and best model with hyperparameters will be recorded in an accuracy log for concluding the section.

# Data source and Methods

***Data Source***

As a data source, we will use an available for public use dataset "120 years of Olympic history: athletes and results" published on Kaggle (<https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data>). The dataset contains 271116 rows and 15 columns which represent data about 134732 athletes performing in Olympic games from 1896 to 2016. The original source of the dataset we are working with is website (<https://www.sports-reference.com>)

## *E**DA*

Exploratory Data Analysis will help us to understand the structure of the dataset including the size, shape, properties and types of variables. Also, identify patterns and relationships between variables. Additionally, EDA allows us to select appropriate techniques and models for our further analysis.

**1. Statistical summary**

**2. Distributions**

* Bar plots

**3. Correlation**

* Heatmap
* Correlation matrix
* Scatter plots

**4. Feature importance**

* Gradient Boosting Classifier

#### Data Preparation

The data preparation process is a very important stage in the data analysis, it includes several steps to ensure that the data is in an appreciate format for analysis and modelling. Depending on the project stages of data preparation may vary. In this project we are going to perform the following stages:

**1. Normalising skewed features**

* Methods to test:
  + log Transformation
  + Cutting outliers

**2. Missing values**

* Methods to test:
  + Drop rows with missing values
  + Fill in the mean value
  + Fill with ML model

**3. Feature engineering**

* Methods to perform:
  + Drop irrelevant features
  + Create a new feature 'BMI'
  + Fill with ML model

**4. Encoding**

* Methods to test:
  + Standard encoder
  + One Hot encoder

#### Machine Learning Models

**1. PCA**

Our data set doesn't include a large number of features for applying Principal Components Analysis. However, it may be helpful to discover how this type of data can be compressed in case we get more detailed data.

**2. Logistic Regression for probability calculations**

This model mostly will be used for the Cross-Disciplinary Characteristics section to define the probability of successful performance.

**3. GridSearch with different classifiers**

Another section in our research "Predictability Evaluation" will have to define the best model and hyperparameters for predicting performance results. We are going to compare the performance of several models and each of them will be tuned as far as computing ability lets us to tune.

3.1. Machine Learning Models to test:

* Logistic Regression
* Decision Tree
* Random Forest
* Gradient Boosting
* K-Nearest Neighbors

3.2. Deep Learning Models to test:

* MLP Neural Network
* PyTorch Neural Network

# Ethical consideration

#### Privacy and Data Protection

To provide privacy during our analysis we will anonymise all records by removing personal information such as ID and Name of athletes. While doing our research we might discover information which can affect athletes' current lives even though they have already finished their careers. Many very dangerous zones in world-class media data analytics must be considered before creating and publishing a study. For instance, by analysing an athlete's performance during the training or previous performances and the main performance it is possible to find anomalies which can point to doping use. This unproven fact may have a negative impact on an athlete's life and career. To sum up, the information obtained from data may be misinterpreted and used against people related to it.

#### Transparency

All models, data preparation stages and conclusions are available for any reader. As well as comments with an explanation of why such a method was chosen. The reader can check how results were obtained and if there are any biases in the analysis.

#### Geographic bias

As was mentioned in the introduction section, geographical bias is a very big issue in world-class sports competitions. Athletes from different countries may be equipped differently and have different selection and training processes. Additionally, direct naming nationality of athletes may cause issues. Because nationality is a very important feature we are going to apply encoding techniques to anonymise it.

#### Competitive advantage

Because we use a public dataset and did not collect data using our resources, we should consider fair access to results for all people and organisations associated with the dataset, no matter how advanced the results are. The link to the repository is attached to this report.

# EDA

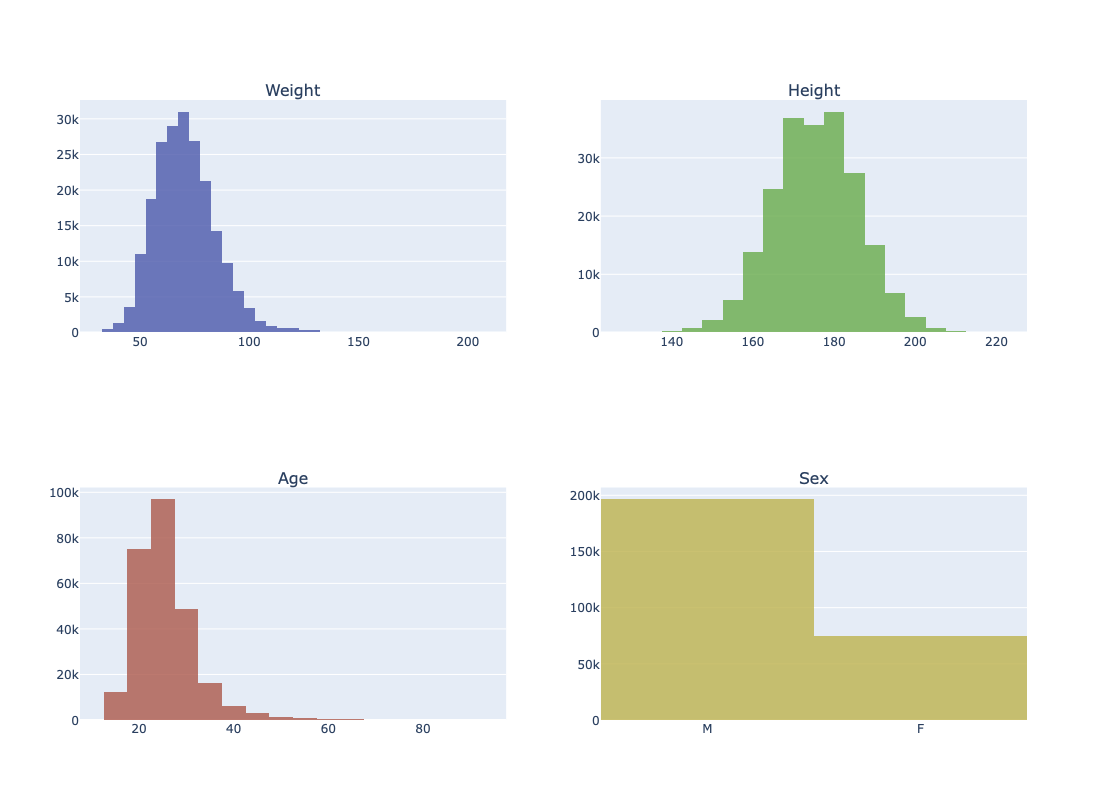
### Summary Statistics

count mean std min 25% 50% \  
ID 271116.0 68248.954396 39022.286345 1.0 34643.0 68205.0   
Age 261642.0 25.556898 6.393561 10.0 21.0 24.0   
Height 210945.0 175.338970 10.518462 127.0 168.0 175.0   
Weight 208241.0 70.702393 14.348020 25.0 60.0 70.0   
Year 271116.0 1978.378480 29.877632 1896.0 1960.0 1988.0   
  
 75% max   
ID 102097.25 135571.0   
Age 28.00 97.0   
Height 183.00 226.0   
Weight 79.00 214.0   
Year 2002.00 2016.0

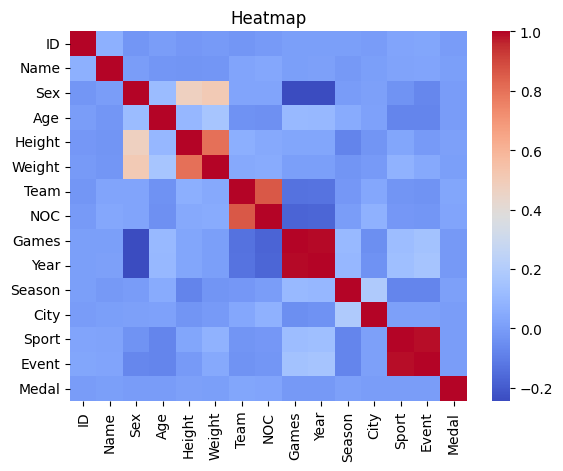
If we take a look at the statistic summary, we may get important information for our further analysis:

1. In the Age feature the mean value of 1.5 higher than the median, which tells us that this data is skewed.
2. Height and Weight features also a slight skewness, however, the size of it is not significant and shouldn't affect our analysis.
3. Standart deviation value tells us that:
   * 95% of athletes performing in the Olympic Games are aged approximately between 17 and 31 years
   * The range of height for the vast majority of athletes is between 165 and 185 centimetres
   * The weight feature follows a similar pattern as the height. The range of weight for 95% of athletes is approximately between 56 and 84 kg.

### Distributions



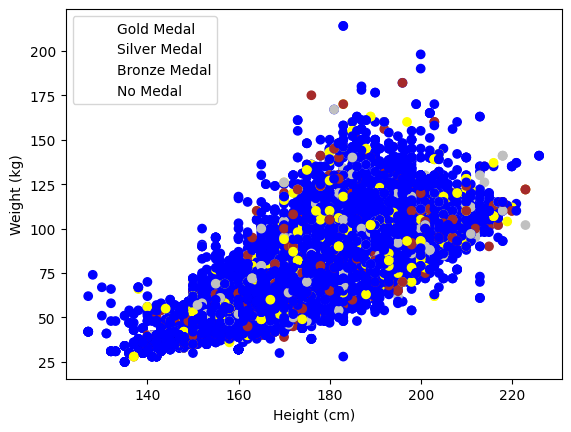
### Correlations



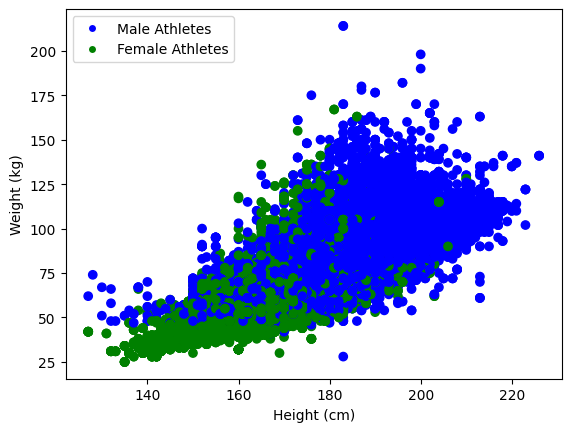
Correlations with Medal  
ID -0.004109  
Name 0.004151  
Sex -0.005118  
Age -0.002555  
Height 0.008385  
Weight 0.003913  
Team 0.028808  
NOC 0.025316  
Games -0.012009  
Year -0.013301  
Season 0.010670  
City -0.000774  
Sport -0.000114  
Event 0.000007  
Medal 1.000000  
Name: Medal, dtype: float64

### Scatter plots

#### Height and Weight vs. Medal



#### Height and Weight vs. Gender



### Feature importance

Let's take a look at feature importance as a part of our EDA/Correlation stage. To do this we are going to apply the Gradient Boosting (Classifier) model which is supposed to display a level of importance for each feature.

1. Define the target variable and split the dataset.
2. Define and train model
3. Get an importance score using the model
4. Set options to display the entire data frame with importantce scores
5. Create df for displaying scores and sort it by dis.

X=df\_corr\_encoded.drop(columns=['Medal'],axis = 1)  
y=df\_corr\_encoded['Medal']  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.25, random\_state=42)

gb\_model = GradientBoostingClassifier()  
gb\_model.fit(X\_train, y\_train)  
feature\_importances = gb\_model.feature\_importances\_  
pd.set\_option('display.max\_rows', None)   
pd.set\_option('display.max\_columns', None)   
importance\_df = pd.DataFrame({'Feature': X\_train.columns, 'Importance': feature\_importances})  
importance\_df = importance\_df.sort\_values(by='Importance', ascending=False)  
print(importance\_df)

Feature Importance  
7 NOC 0.268891  
13 Event 0.166690  
6 Team 0.130037  
11 City 0.100808  
1 Name 0.069405  
0 ID 0.046682  
8 Games 0.042121  
12 Sport 0.032468  
5 Weight 0.030729  
4 Height 0.030167  
9 Year 0.027288  
10 Season 0.026998  
3 Age 0.026186  
2 Sex 0.001529

# Preparaton

**Note:**

From this section and until the end we are going to test different approaches of processing and modeling data. We need to understand which method of dealing with data benefits the most. At the beginning of each subsection, I will list the options I am going to test.

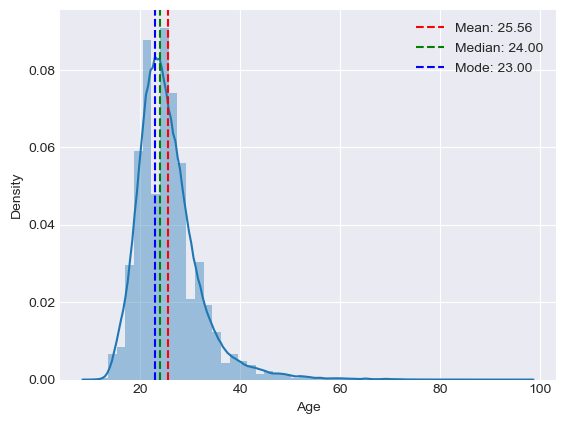
### Skewed data

Several authors highlighted the impacts of using skewed data and some solutions to deal with it. Kazerouni A., Zhao Q., Xie J., Tata S. and Najork M. (2020) and Monard M. (2002), discuss how skewed data may affect classification tasks.

Authors Ha T. N., Lubo-Robles D., Marfurt K. and Wallet B. (2021), highlighted the advantages of using log transform to deal with skewed data.

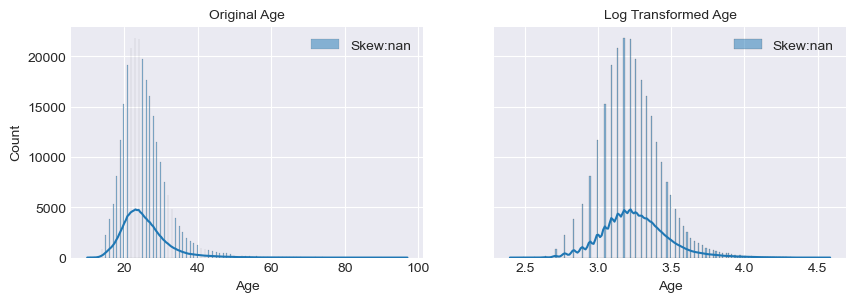
As we can see some independent features are skewed. This fact may affect the model's performance because certain models expect that both independent and dependent variables are evenly distributed. To handle this issue we may apply different techniques and compare their effectiveness.

1. Remove outliers
2. Log Transformation
3. Leave data hot as it is



#### Log Transformation

LogTranformation gets logarithm from each value and creates more smooth distrebution, and in some extend balances ratio between small and large values. Briefly, the logarithm can be defined as a power which we need to tranform log base to a number where we need to get logarithm from. The plot below illustrates how feature's distrebution will change if we apply logTransform.



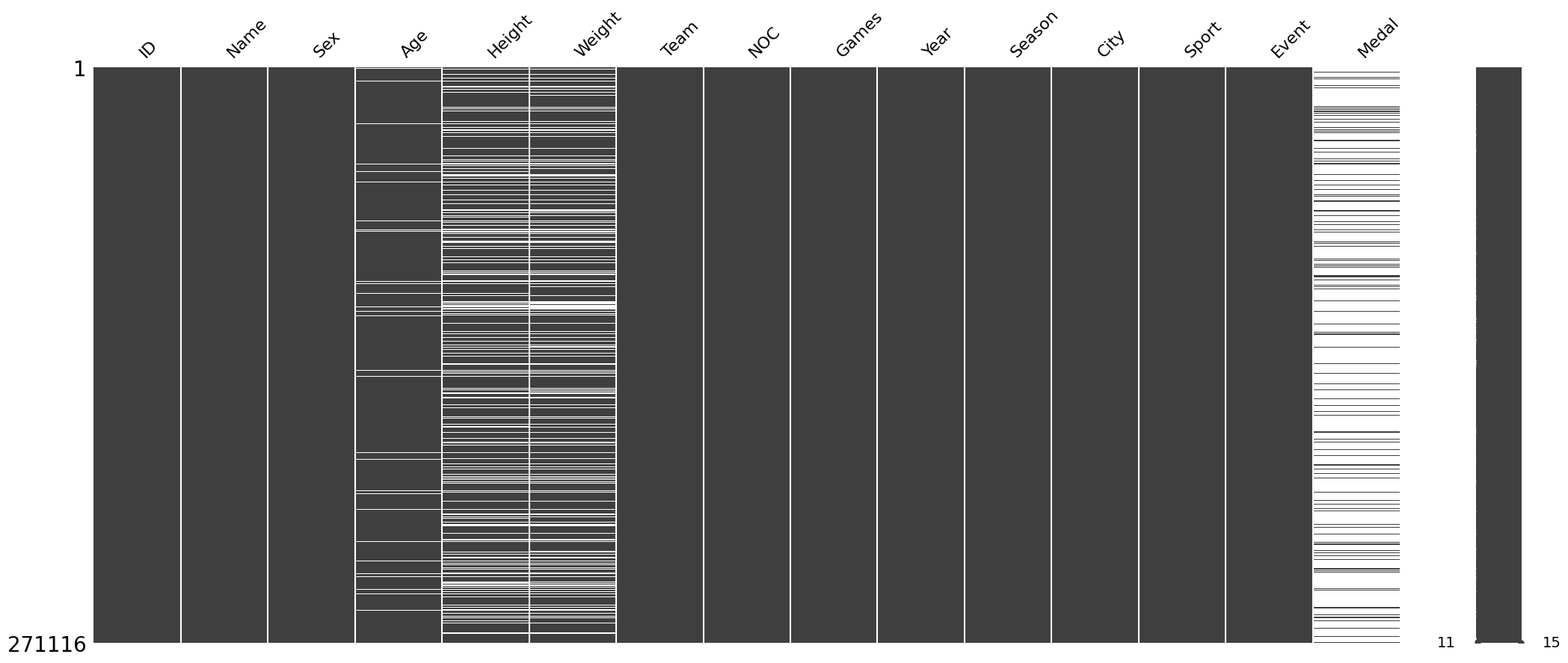
### Missing values

Several studies were reviewed dealing with missing values, and each of them focuses on specific techniques.

1. Palanivinayagam A. (2023), applies a machine learning method to impute missing values.
2. Raja P. (2020), uses unsupervised machine learning models to fill in missing values.
3. Makaba T. and Dogo E. (2019), compare different strategies of missing values imputation.

First of all, we need to understand what to do with missing values. In our case, there are 3 columns with missing data (Age, Height, Weight). So, we have several ways how to process this stage:

1. Drop rows with missing values
2. Fill them with the mean
3. Fill them with the ML model



#### 1.1. Drop rows with missing values

This approach has pros and cons. By removing rows which contain missing values we may lose significant part of the dataset, however the rest of dataset will contain very accurate data.

#### 1.2. Fill missing values with the mean

This method allows us to save shape of dataset but it may harm to final accuracy or make model overfitted.

#### 1.3. Fill missing values with ML models

Here we are going to fill missing values using KNN Regressor. But we have an obstacle as missing values in multiple features, what will make an imputation a bit challenging. If we would have only one feature, we could have split dataset where missing values are y\_test. However in our case we need temporary define missing all missing values but y\_test. ML models expect no NaN values in datasets. So ironically, to fill missing values we need to fill missing values first.

* The loop function below will help us with this problem.
  + we prepared data for modeling.
  + defined model, method for imputation and encoder.
  + define a list with columns contain missing values.
  + using loop define each column from the list as a test set, encode categorical value, and set mean for missing values.
  + train model and apply it for missing values.

### Feature Engineering

#### Drop Irreverent features

#### Create new feature BMI

#### Label target variable

At this stage we need to think about what we will define as a successful performance. We can include in our search only gold medalists or winners who got at least "Bronze". Considering level of competitiveness in Olympic Games the difference in performance between medalists is extremely small. In some events it can be 0.01 second. Considering that I think it is fair enough to assign all medalists in our database as winners. So, "1" in the column "Medal" will indicate medal and "0" no medal.

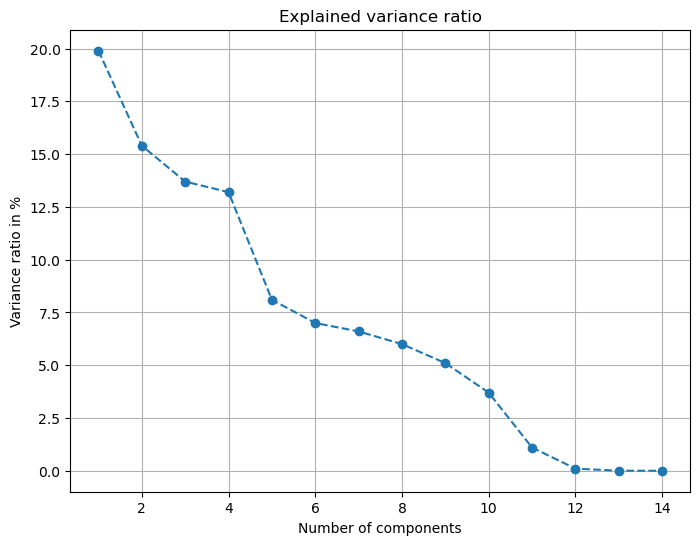
### Encoding

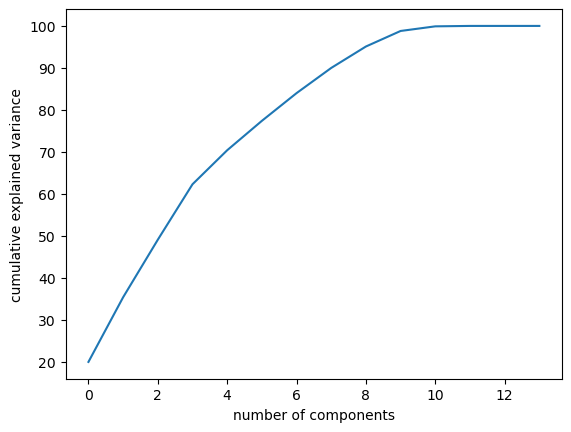
#### Standart

Here we need to encode data separately and store "Sport" keys in own encoder due to will need those keys later during decoding.

### PCA

In this particular case we are dealing with dataset which doesn't encount large number of features to make us use PCA compression. However, this data may be extended. On the world-class level of sport industry, scientists tent to collect hundreds or even thousands different anthropometric parameters for each athlete. In this case PCA transformation is essential tool for analyst. We are going to take a look which ratio of principle components data has, and how data can be compressed.





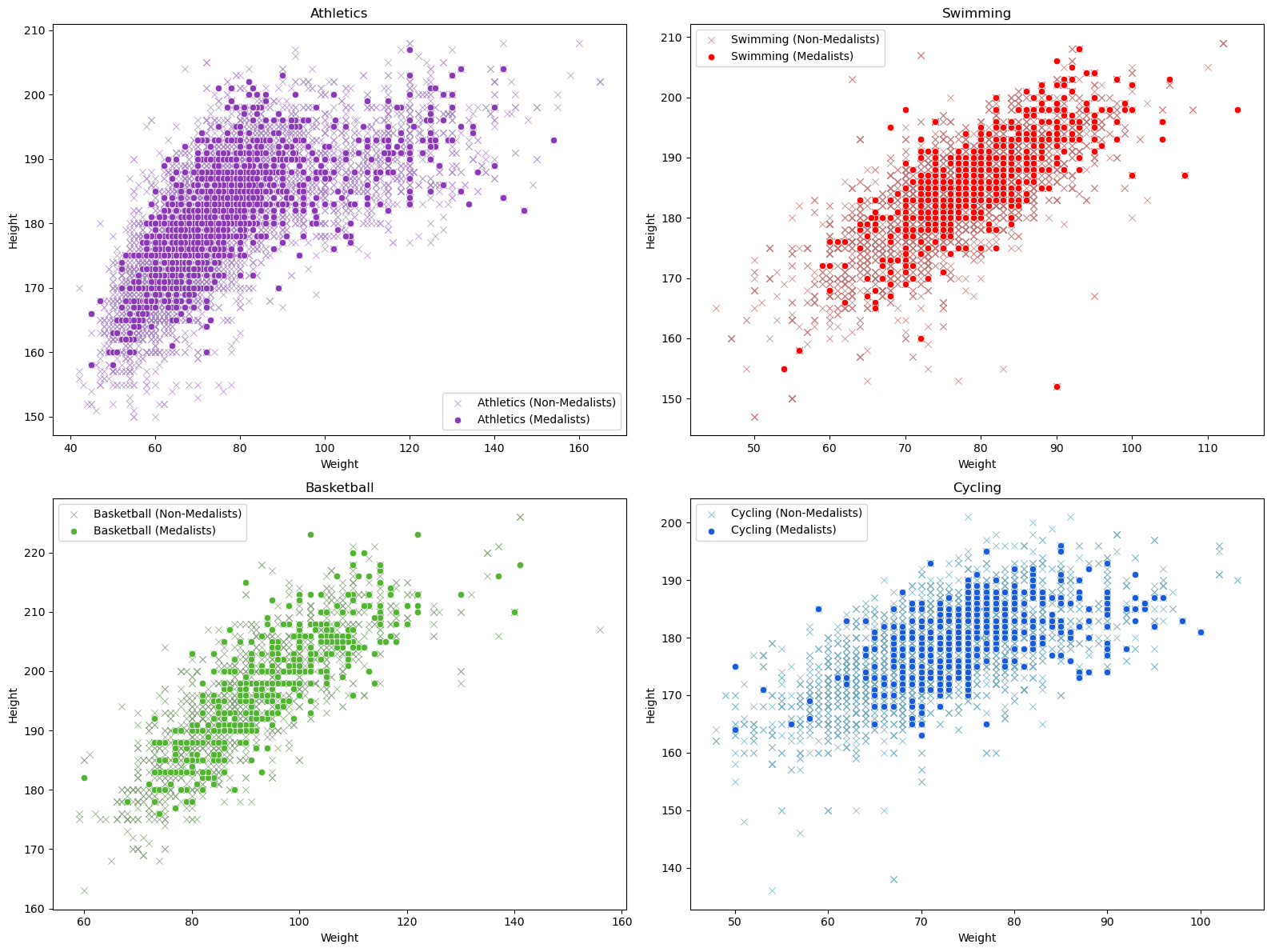
Percentage of variance: 95  
Number of components: 10

Index: 0, Feature: Sex, Variance Ratio: 19.9%  
Index: 1, Feature: Age, Variance Ratio: 15.4%  
Index: 2, Feature: Height, Variance Ratio: 13.8%  
Index: 3, Feature: Weight, Variance Ratio: 13.2%  
Index: 4, Feature: Team, Variance Ratio: 8.1%  
Index: 5, Feature: NOC, Variance Ratio: 7.0%  
Index: 6, Feature: Games, Variance Ratio: 6.5%  
Index: 7, Feature: Year, Variance Ratio: 6.0%  
Index: 8, Feature: Season, Variance Ratio: 5.1%  
Index: 9, Feature: City, Variance Ratio: 3.6%

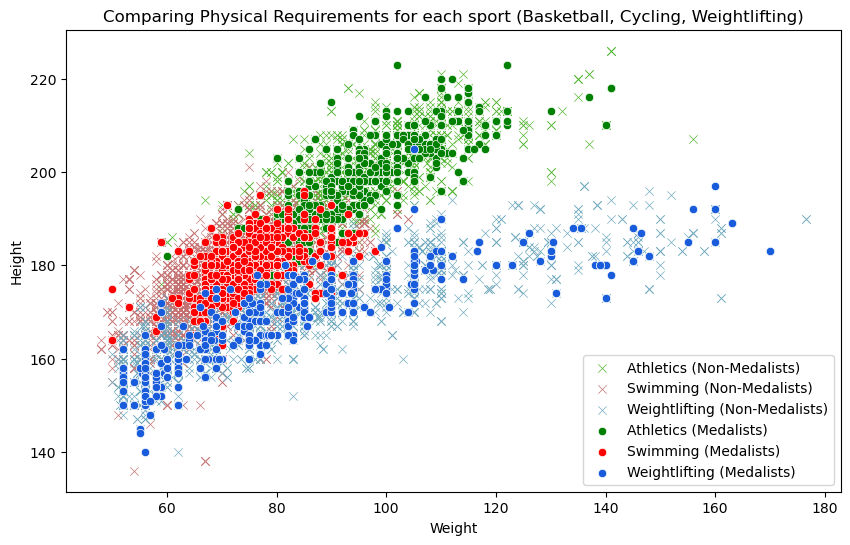
# Cross-Disciplinary Characteristics

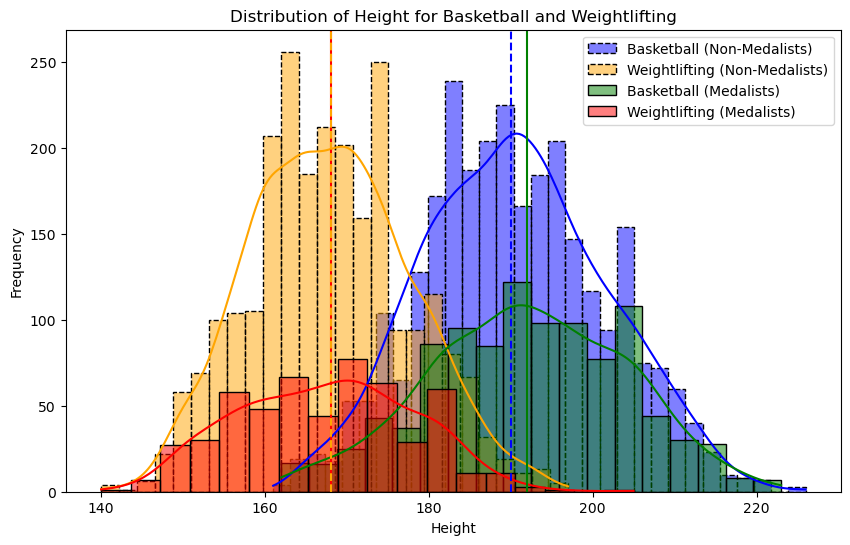
## Introduction

We will try to discover requariments of phisical condition for each sport and liklihhod of win for avarage athlete perfoming in a different sport compition. The chars below illustrates height and weight ration of athletes who perform in 4 different sports.



In order to see how physical characteristics vary depending on the sport we can place data of 3 different sport on one chart. So, now we can clearly see the differences. Additionaly, we will specify the gender.





The bar chart demonstrate the goal for this section. As we can see medalists distrebutions have created an intersection as a cross area where medalists from both sports sharing the same height. However if we will take a look at median line for each distredution we can conclude that probability to win for avarage athlete from basketball significantly small and very likely in the rejection area. Similarly for an avarage athlete from Weightlifting. In this case 'Transfer score' will be very small or even zero, considering only height. Although there are many different sport in Olympic Games programm and each requests specific phisical characteristics, there are may sports with similar requirements where intersection area larger and median values closer.

## Modeling

For a train set we will define athlete's physical characteristics as well as discepline and gender. As a result we should get a probability score of win for an athlete with avarage characteristics in specific sport and gender. The probability score we are going to get won't display real probability for avarage athlete because the real number of factors impact to the victory much larger and number of participants varies every Game. However it mat show general phisical requirements for athletes performing in the sport and how specific phicical comdition impacts the result. Additionally, it will show us the level of competitiveness in each sport.

X = df\_encoded[['Height', 'Weight', 'Sex', 'Sport']]  
y = df\_encoded['Medal']  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LogisticRegression()  
model.fit(X\_train, y\_train)

LogisticRegression()

Probability of winning a medal for an athlete from sport (17) with average height and weight, gender (1) in sport (61): 0.16716188649289204

For another sport the probability is 17.8% that means it is less competitive that previos one.

To achive our objective for this section we need to apply avarage characteristics of athletes from one sport discipline (17) to another (61).

Probability of winning a medal for an athlete with average height and weight, specific gender, and sport: 0.1875744504751054

As we can see in this model we got probability 16.9%. This fugure means that if athlete from sport (17) where their probability to win equals 12.5%, performs in sport (61) they are more likely to win in 4.4%. However the same athlete has less chances to win that athletes from that sport (61) where their probability equals 17.8%.

Create a new df with all possible combinations of sports for specific gender. Note: Some sports include athletes with only one specific gender. So, before creating a list with unique combinations we will filter data by only one gender due to avoid any NaN values.

Now we need to create a loop using the code we created before. We do not need to rum ML model try each row right now because in this stage there will be only 63 unique values which we already have and we could complete this part just using placing values according the keys, like vlookup in excel. However, we will need to run the similar code any way when we will be calculating a probability with the second sport. So, this may be concidered as a checking on a loop before second run.

Note: a running the loop below may take 30 min, more or less. I am going to save results as csv file which will be attached to final submission. Code for reading csv will be placed after the loop.

**for** idx, sport **in** enumerate(tqdm(comb\_df['First Sport'])):  
 filtered\_data = df\_encoded[(df\_encoded['Sex'] == 1) & (df\_encoded['Sport'] == sport)]  
 **if** len(filtered\_data) > 0:   
 average\_height = filtered\_data['Height'].mean()  
 average\_weight = filtered\_data['Weight'].mean()  
 specific\_sport = str(sport)  
 specific\_gender = '1'  
 athlete\_data = [[average\_height, average\_weight, specific\_gender, specific\_sport]]  
 model = LogisticRegression()  
 model.fit(X\_train, y\_train)  
 probability = model.predict\_proba(athlete\_data)[0][1]  
 comb\_df.at[idx, 'Prob. First'] = probability  
 comb\_df.at[idx, 'Height'] = average\_height  
 comb\_df.at[idx, 'Weight'] = average\_weight  
print(comb\_df)

Here we have recorded phisical characteristics of avarage athlite in initual sport (that sport where the an atlite performs and probability to win in that sport). The next stage we need to calculate probability their probability to win in another sport.

Features "Weight" and "Height" have served their duties and may be removed.

comb\_df.drop(columns=['Height', 'Weight'], inplace=True)

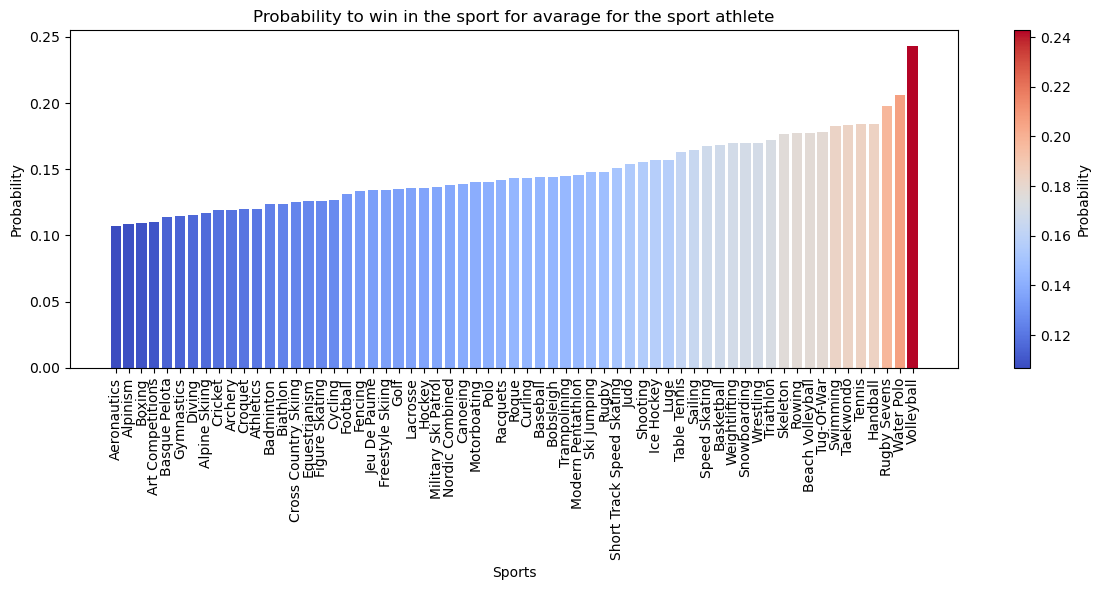
comb\_df.head()

First Sport Second Sport Prob. First Prob. Second  
0 8 8 0.168305 0.168305  
1 8 32 0.168305 0.197093  
2 8 24 0.168305 0.187104  
3 8 61 0.168305 0.236637  
4 8 17 0.168305 0.178687

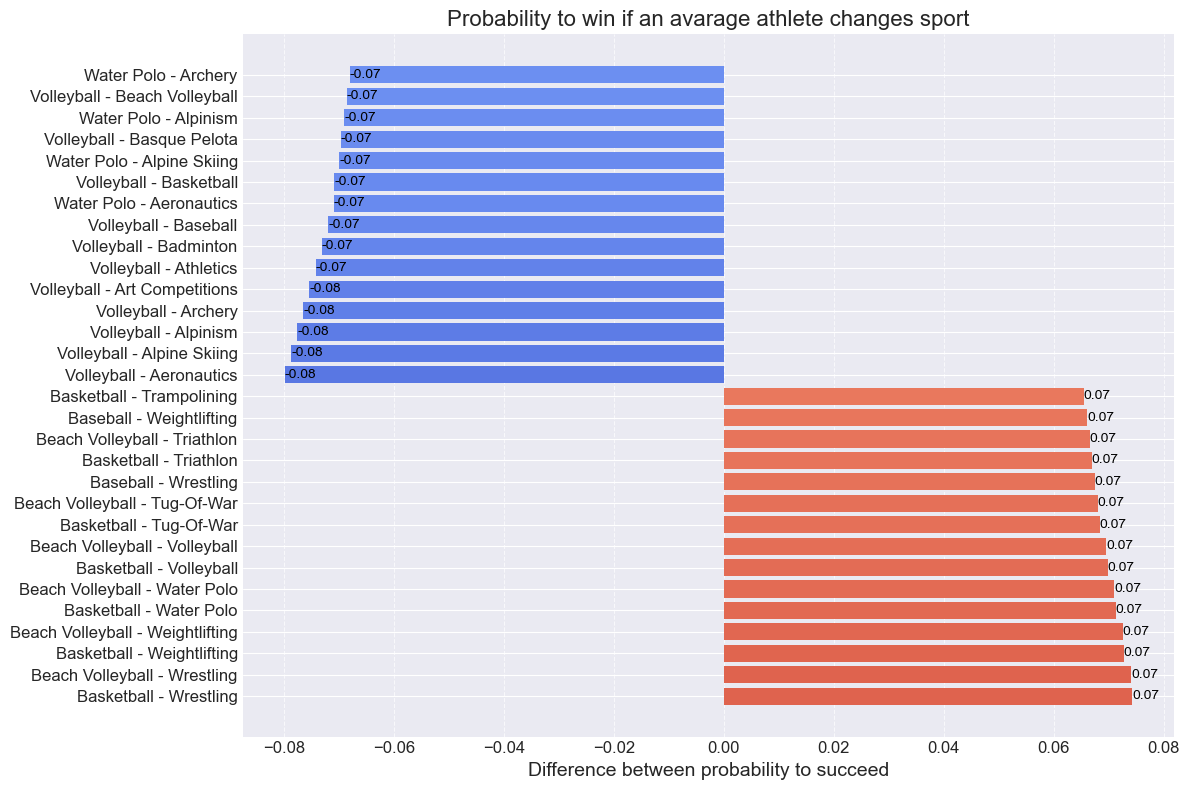
## Evaluation

result\_df.head()

First Sport Second Sport Prob. First Prob. Second Transfer score  
0 Basketball Wrestling 0.168305 0.242500 0.074196  
1 Beach Volleyball Wrestling 0.177687 0.251708 0.074021  
2 Basketball Weightlifting 0.168305 0.241025 0.072720  
3 Beach Volleyball Weightlifting 0.177687 0.250195 0.072509  
4 Basketball Water Polo 0.168305 0.239556 0.071252



Displayed chart illustrates pair of sports where else an avarage athlete can perform and how likely they will succeed. As it was mentioned before the given values do not fully display probality of an athlete's victory, there many different factors define athletes success. However the values sufficiently show the sport's requirements for physical condition. The higher probability lower avarage distance between avarage and predicting curves. In other words difference between medalists and non-medalist lower. Second remarkable note, how phicical requirements correlate across different sports. The section can be extended as long as new athletes' characteristics come. It may be general anthropometric mesurments such as a type of body, a lenght of limbs, ratio of body components, apaptation potential and many more. Additionaly, data about phisical measurements such as rapdity, endurance, strength, can be very helpful. Obtaining some of mentioned data will allow to add more features for analysis and get more accurate outcome.



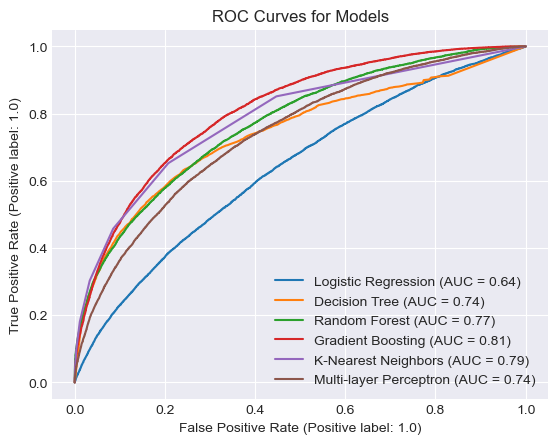
# Predictability Evaluation

## Evaluation models

### Confusion matrix

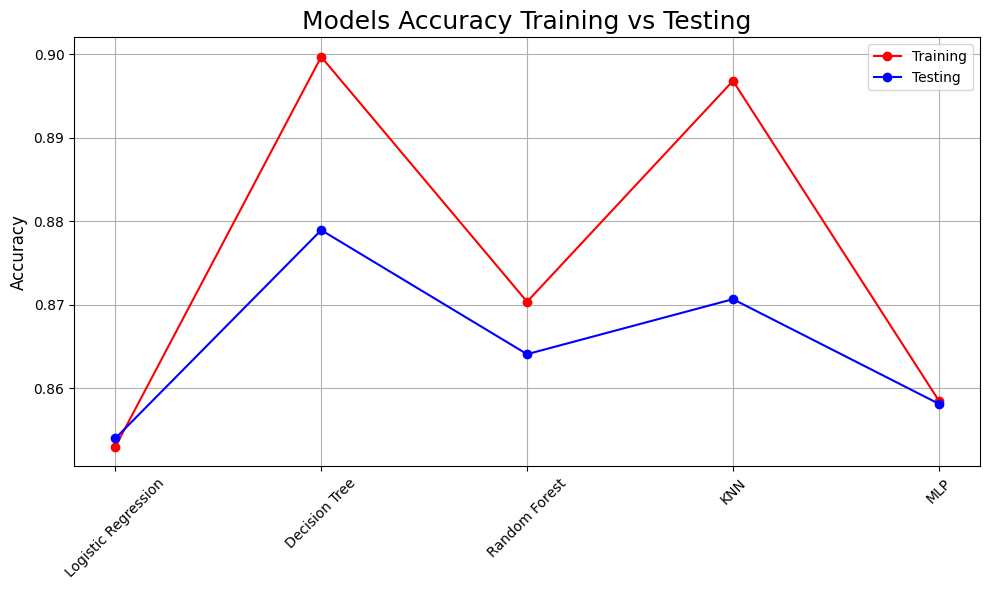
Confusion Matrix for Logistic Regression:  
[[57882 4]  
 [ 9893 0]]  
  
Confusion Matrix for Decision Tree:  
[[56440 1446]  
 [ 6759 3134]]  
  
Confusion Matrix for Random Forest:  
[[57776 110]  
 [ 9103 790]]  
  
Confusion Matrix for KNN:  
[[55985 1901]  
 [ 6865 3028]]  
  
Confusion Matrix for MLP:  
[[57261 625]  
 [ 8991 902]]

### ROC-AUC Curve



## Overfitting

Model Training Accuracy Testing Accuracy \  
0 Logistic Regression 0.852968 0.853981   
1 Decision Tree 0.899713 0.878945   
2 Random Forest 0.870358 0.864073   
3 KNN 0.896831 0.870668   
4 MLP 0.858511 0.858127   
  
 Mean Squared Error Mean Absolute Error \  
0 0.146019 0.146019   
1 0.121055 0.121055   
2 0.135927 0.135927   
3 0.129332 0.129332   
4 0.141873 0.141873   
  
 Best estimator \  
0 LogisticRegression(C=0.01, penalty='l1', solve...   
1 DecisionTreeClassifier(criterion='entropy', ma...   
2 (DecisionTreeClassifier(max\_depth=10, max\_feat...   
3 KNeighborsClassifier(n\_neighbors=7, p=1)   
4 MLPClassifier(activation='logistic', alpha=0.001)   
  
 Best parameters   
0 {'C': 0.01, 'penalty': 'l1', 'solver': 'saga'}   
1 {'criterion': 'entropy', 'max\_depth': 15, 'min...   
2 {'criterion': 'gini', 'max\_depth': 10, 'min\_sa...   
3 {'n\_neighbors': 7, 'p': 1, 'weights': 'uniform'}   
4 {'activation': 'logistic', 'alpha': 0.001, 'so...



# Conclusion

**Cross-Disciplinary Characteristics**

The of that section was to find the probability of winning for an average athlete from one sport in another. It was proven that each sport in the Olympic Games has its physical requirements for athletes, and athletes's performance relies to some extent on physical condition. Additionally, we investigated the competitive nature of each sport, and extracted probability coefficients of winning for an average athlete in each sport. The part of the section where we investigate the probability of winning for an athlete who changes sports shows major differences in sports requirements. This method is actively used in sports medicine to predict injuries athletes can get performing unusual physical activity. For example, Lovdal S., Hartigh R. and Azzopardi G. (2021) applied machine learning models to predict runners' injuries who specialise in different distances.

**Predictability Evaluation**

Based on the models' evaluation was concluded that the KNN and Gradient Boost can predict medalists with the highest accuracy compared with other models tested. Models such as KNN, Gradient Boost and Random Forest may reach an accuracy close to 90%, however exact result depends on the preparation process. A similar study was conducted by Musa, R. (2019), where the author analysed archers' data to predict their performance, highlighting the KNN classifier as the most effective model.

Another interesting study made by Bunker, R. and Thabtah F. (2019) concluded that ANN models are the most effective methods for predicting sports data. This study focused mostly on general predictability, however, the mentioned studies can be considered as starting points in deeper analysis

## Recommendation for future work

The study has many different ways for improvements, in this section, I would like to highlight 2 main ways to develop this study.

**1. Qualitatively, using the same dataset but different methods**

* Many methods of data preparation were not tested. So there is a big scope to investigate how each method applied affects the Machine learning model.
  + Investigate more relationships with the database.
  + Calculation probabilities for each specific class.

**2. Quantitatively, using the same or similar method with extended or additional dataset**

Any additional information may add complicity and make predictions and probability more accurate. There are many different athlete metrics which may significantly improve the outcome such as:

1. Physical data:
   * Detailed anthropometrical measurements.
   * Results of physical testing.
   * Biochemical data such as blood testing.
   * Genetic data.
2. Detailed environmental data such as:
   * Couches, methodologies, physiotherapist.
   * Establishment which provides an athlete's preparation.
   * Amount of investment, funds, grants, and salaries.
   * Athlete's family.

# References

Anik, A. I., Yeaser, S., Hossain, A. I., and Chakrabarty, A. (2018) Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms, Proc. 4th Int. Conf. Electr. Eng. Inf. Commun. Technol. (iCEEiCT), pp. 500-505.

Arkaev, L. I., and Suchilin, N. G. (2004) How to Create Champions. Oxford, Meyer & Meyer Sport.

Arneson, R. (2015). Equality of opportunity. In E. Zalta (Ed.), The Stanford encyclopedia of philosophy (Summer 2015 ed.). Retrieved from <https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/>

Bahr, R., Clarsen, B., and Ekstrand, J. (2018). Why we should focus on the burden of injuries and illnesses, not just their incidence. Br J Sports Med, 52, 1018–1021. <https://doi.org/10.1136/bjsports-2017-098160>

Bahr, R., and Krosshaug, T. (2005). Understanding injury mechanisms: a key component of preventing injuries in sport. Br J Sports Med, 39, 324–329. <https://doi.org/10.1136/bjsm.2005.018341>

Bunker, R. and Thabtah F. (2019) A machine learning framework for sport result prediction. Applied computing and informatics 15 (1), 27-33, 2019

Borms, J., and Hebbelinck, M. (1984). Review of Studies on Olympic Athletes. In J. E. L. Carter (Ed.), Physical Structure of Olympic Athletes Part 11: Kinanthropometry of Olympic Athletes (pp. 7-27). New York: Basel-Karger.

Carter, J. E. L. (1985). Morphological Factors Limiting Human Performance. In D. H. Clarke & H. M. Eckert (Eds.), Limits of Human Performance. American Academy of Physical Education Papers, No. 18. Champaign: Human Kinetics.

Cuk, I., and Karácsony, I. (2002). Rings. Methods, Ideas, Curiosities, History. Norman (OK), Paul Ziert & Associates.

De Bosscher, V. (2007). Sports Policy Factors Leading to International Sporting success. Published Doctoral thesis, Vrije Universiteit Brussel, Brussel, VUBPRESS.

De Jesus, K., Ayala, H. V. H., de Jesus, K., Coelho, L. D. S., Medeiros, A. I. A., & Abraldes, J. A. (2018). Modelling and predicting backstroke start performance using non-linear and linear models. J. Hum. Kinetics, 61(1), 29-38.

Derrick, B., Toher, D., & White, P. (2017). "How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017)" (PDF). The Quantitative Methods for Psychology, 13(2), 120–126.

Donoghue, P. O. (2009). Research Methods for Sports Performance Analysis. Evanston, IL, USA: Routledge.

Fisher, R. A. (1922). "On the Interpretation of χ2 from Contingency Tables, and the Calculation of P". Journal of the Royal Statistical Society, 85(1), 87–94.

Flegl, M., and Andrade, L. A. (2018). Measuring countries’ performance at the Summer Olympic Games in Rio 2016. OPSEARCH, 55, 823–846. <https://doi.org/10.1007/s12597-018-0347-8>

Frantisek, T. (2011). Competitive Loading in Top Team Handball. Vienna, Austria: European Handball Federation WEB PERIODICAL.

Ha, T. N., Lubo-Robles, D., Marfurt, K., and Wallet, B. (2021). In-depth analysis of logarithmic data transformation and per-class normalization in machine learning. Interpretation, 9(3), T685-T710.

Heinilä, K. (1982). The totalization process in international sport. Sportwissenschaft, 12(3), 235–254.

Hore, S., and Bhattacharya, T. (2018). A machine learning based approach towards building a Sustainability Model for NBA players. Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT), pp. 1690-1694.

Johnson, E. (Ed.). (2000). Essentials of Exercise Physiology (pp. 500-527). Lippincott Williams and Wilkins, USA.

Kazerouni, A., Zhao, Q., Xie, J., Tata, S., and Najork, M. (2020). Active learning for skewed data sets. arXiv preprint arXiv:2005.11442.

Kendall, M. G., and Stuart, A. (1973). The Advanced Theory of Statistics, Volume 2: Inference and Relationship. Griffin.

Kurgan, L., and Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. Knowl. Eng. Rev., 21, 1–24.

Lovdal, S., Hartigh, R., and Azzopardi, G. (2021). Injury prediction in competitive runners with machine learning. International Journal of Sports Physiology and Performance, 16(10), 1522-1531.

Maanijou, R., and Mirroshandel, S. A. (2019). Introducing an expert system for prediction of soccer player ranking using ensemble learning. Neural Comput. Appl., 31(12), 9157-9174.

Makaba, T., and Dogo, E. (2019). A comparison of strategies for missing values in data on machine learning classification algorithms. International multidisciplinary information technology and engineering conference (IMITEC), 1-7.

Meayk, E., and Unold, O. (2011). Machine learning approach to model sport training. Comput. Hum. Behav., 27(5), 1499-1506.

Michalsik, L. B. (2004). Analysis of working demands of Danish handball players. In What’s Going Gym. Copenhagen, Denmark: Forlaget Underskoven, pp. 321-330.

Monard, M. C. (2002). Learning with skewed class distributions. GEAPA Batista, 85, 173-180.

Musa, R. M., Majeed, A. P. P. A., Taha, Z., Abdullah, M. R., Maliki, A. B. H. M., and Kosni, N. A. (2019). The application of artificial neural network and k-nearest neighbour classification models in the scouting of high-performance archers from a selected fitness and motor skill performance parameters. Sci. Sports, 34, 241-249.

Musa, R. M., Majeed, A. P. P. A., Taha, Z., Chang, S. W., Nasir, A. F. A., and Abdullah, M. R. (2019). A machine learning approach of predicting high potential archers by means of physical fitness indicators. PLoS ONE, 14(1).

Musch, J., and Grondin, S. (2001). Unequal competition as an impediment to personal development: a review of the relative age effect in sport.

Nikulin, M. S. (1996). A guide to chi-squared testing. New York: Wiley.

Ofoghi, B., Zeleznikow, J., Macmahon, C., and Dwyer, D. (2013). Supporting athlete selection and strategic planning in track cycling omnium: A statistical and machine learning approach. Inf. Sci., 233, 200-213.

Palanivinayagam, A. (2023). Effective handling of missing values in datasets for classification using machine learning methods. Information, 14(2), 92.

Raja, P. (2020). Missing value imputation using unsupervised machine learning techniques. KJSC Thangavel Soft Computing, 24(6), 4361-4392.

Sekeroglu, B., Dimililer, K., and Tuncal, K. (2019). Artificial Intelligence in Education: Application in student performance evaluation. Dilemas Contemporaneos Educacion Politica Valores, 7(1), 1-21.

Sprinthall, R. C. (2011). Basic statistical analysis (9th ed.). Pearson Education.

Zhou, Z., Shakya, S., and Sha, Z. (2017). Predicting countermovement jump heights by time domain frequency domain and machine learning algorithms. Proc. 10th Int. Symp. Comput. Intell. Des. (ISCID), pp. 167-170.

Author: Ilia Grishkin

<https://github.com/CCT-Dublin/capstone-project-feb-2024-ft-Ilia-Grishkin>