# From Data to Victory

## Predicting Olympic Athletes  Success and Exploring Cross-Disciplinary Traits with Machine Learning

Author: Ilia Grishkin

cct | College Dublin
Computing ▪ IT ▪ Business

## Introduction

The Olympic Games represent the pinnacle of athletic achievement, with athletes striving to reach extraordinary physical and mental performance levels. Olympic competitors embody a unique blend of genetic predisposition, rigorous training, and environmental factors that collectively foster peak performance. However, as the competition intensifies, the role of data science and predictive modelling in sports has become essential. Coaches and analysts increasingly collaborate with data scientists to develop models that forecast performance, reveal patterns, and enhance talent selection.

## Objectives

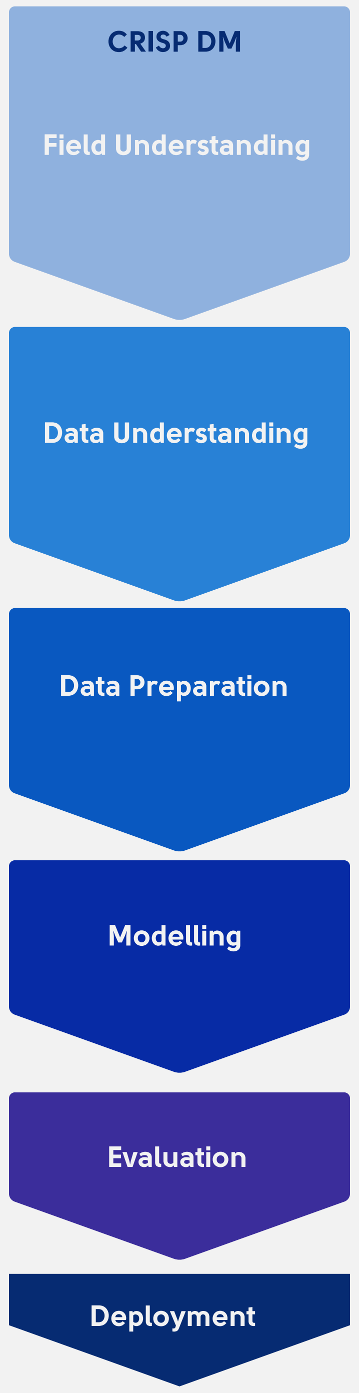
### Cross-Disciplinary Characteristics

- Identify similarities between all sport pairs presented in the dataset and which sport types are the most suitable for athlete's transfer or additional sport.
- Identify the probability of winning for an athlete with certain physical characteristics in each sport presented in the dataset.

### Predictability Evaluation

- Trace a model and parameters with the best performance in classifying athletes as medalists or non-medalists based on physical characteristics. Additionally, ascertain which data preparation processes such as SMOTE and PCA transformations contribute the most to the models' performance.

## Data source and Methods



- Dataset about 134732 athletes performing in the Olympic games from 1896 to 2016
- Ethical consideration
- Literature review
- Data resources
- Objectives
- Methods

- Feature importance
- Statistic summary
- Distributions
- Correlations
- PCA

- Feature engineering
- Skewed features
- Missing values
- Encoding
- Scaling
- SMOTE

- Multi-Layer Perceptron (MLP)
- Gradient Boosting Classifier
- Random Forest Classifier
- Logistic Regression
- GridSearch
- XGBoost

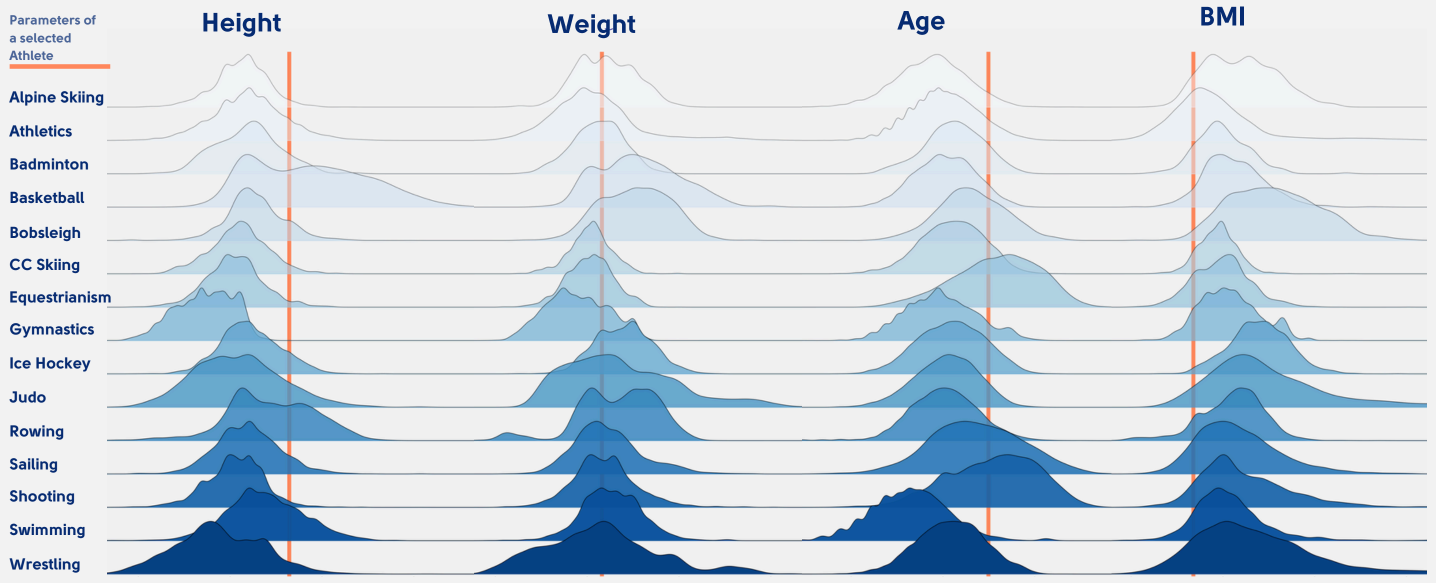
- Train/Test Accuracy,
- Precision
- F1 score
- Recall

- Visualisation
- Conclusion
- Poster

## Probability of winning

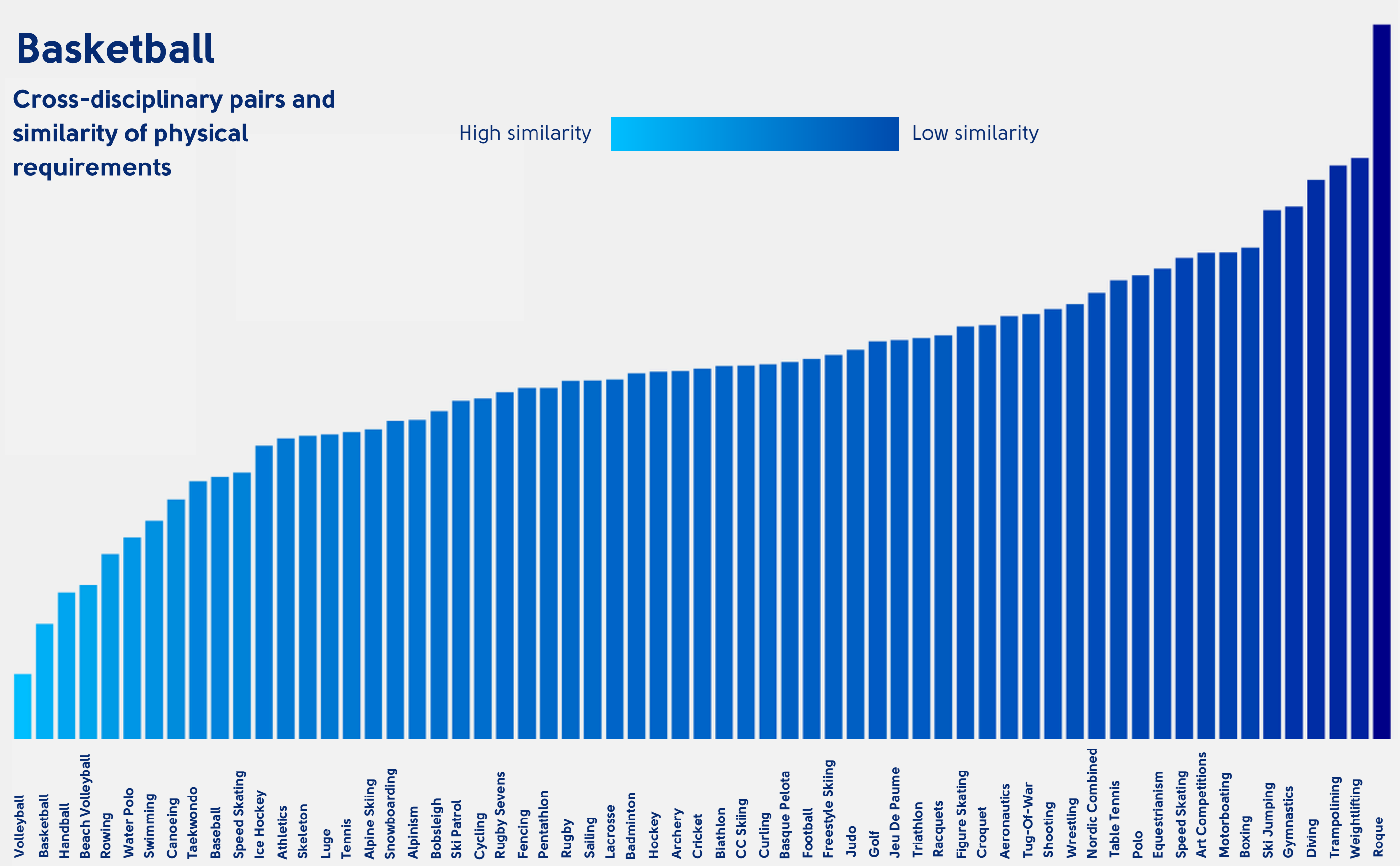
In this section, we have determined the probability of winning for an athlete with certain physical characteristics. This probability was calculated by measuring the Euclidean distance between the parameters of the selected athlete and the mean values of winners in each sport. In the plot below, the orange line represents the parameters of the selected athlete compared with the distribution of the same parameters among winners in each sport. The closer an athlete's parameters are to the mean parameters of winners, the higher the probability of winning, given the selected features. After calculating the Euclidean distance between the athlete's parameters and each sport's mean values, we transformed that distance into a probabilistic framework using the sigmoid function.

Additionally, if we want to assess a person's predisposition for a specific sport, we can use the same procedure, except for the final step. In this case, the Euclidean distance alone can serve as a predisposition score.



## Similarities between sport pairs

This section is dedicated to the similarity of physical requirements in all sports presented in the dataset. Similarly to the previous stage, we have calculated Euclidean distance. However, in this case, instead of the parameters of a certain athlete, we applied the mean value of all athletes in each sport. As a result, we have a data frame with the similarity score for each pair of sports. The plot below illustrates 1 out of 227 rows in the data frame, where we visualised all similarity scores for selected sports (Basketball) in a specific gender group. This insight may be helpful for athletes and coaches with planning cross-disciplinary transfers.



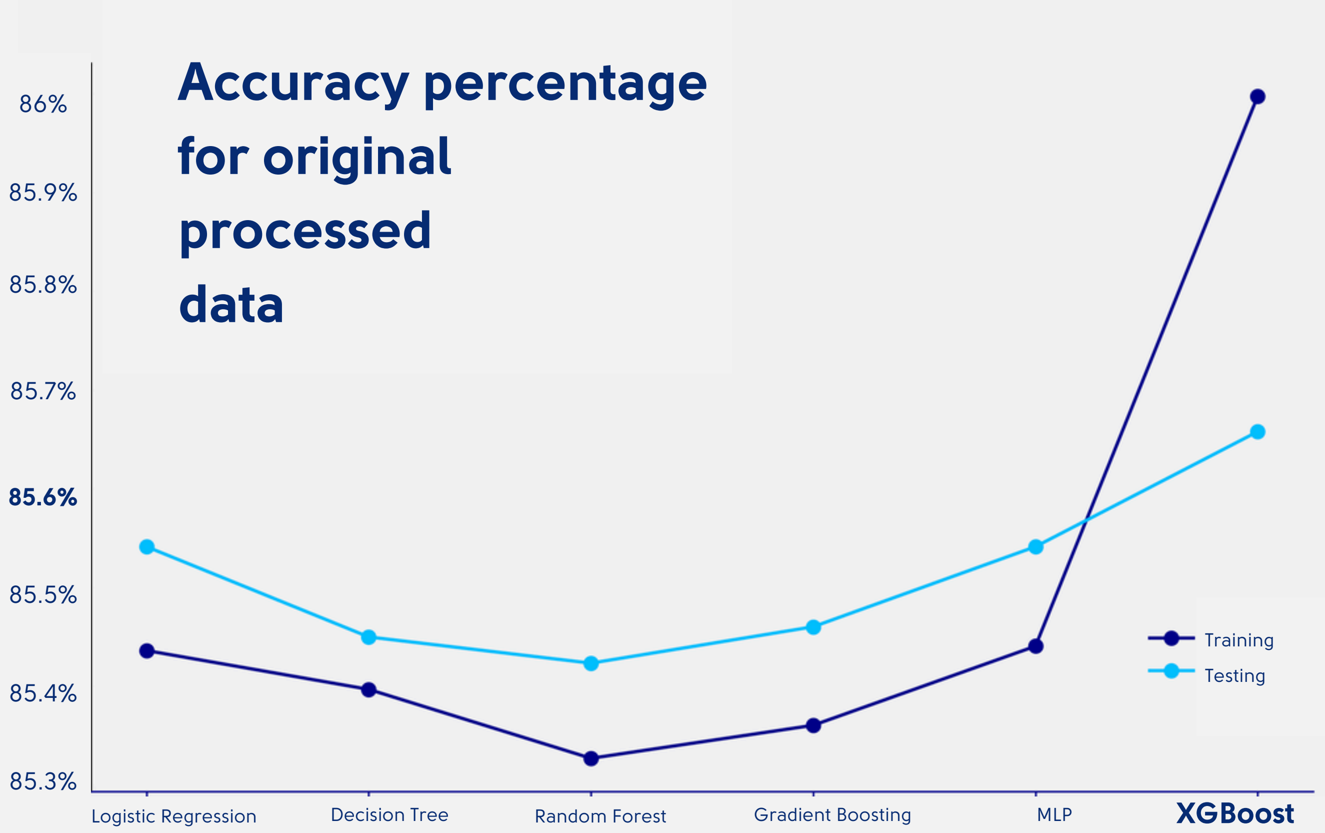
## Predictability Evaluation Results

We can conclude that XGBoost is the top model in this project. It consistently achieves high test accuracy and balanced performance metrics across different datasets.

PCA and SMOTE did not significantly improve the models' performance. In some cases, they even reduced accuracy, suggesting that the original data might already capture the essential features well enough.

Using the original processed data with XGBoost seems to be the most effective approach for predicting medalists based on physical characteristics, providing a strong balance between precision, recall, and F1 scores.

XGBoost showed the highest test accuracy score (85.66%) and F1 score (0.798)



## Scope

### Qualitative Improvements

Include and test the impact of other data preparation techniques such as dimensional reduction using LDA, different SMOTE transformations for balancing train and test sets and other stages related to data preparation. Additionally, there is big scope to improve Modeling by testing other models such as KNN, SVM, and ANN.

### Quantitative Extenuation

It is very likely that including more features about athletes' physical and mental abilities can improve a model's performance. Modern sports science considers thousands of parameters for tracking an athlete's condition. Including, some of those parameters may improve the models' performance.

### Creating a Methodology

Obtained insights about sports-type similarities and how each sport is suitable for athletes with certain physical characteristics, may serve as a baseline for a methodology for couches. It may support the decision about transferring a performing athlete to another sports event or choosing a sport for young athletes.

### Building a web Application

Both parts of this project may be transformed into a web application. For the first part, it would be able to recommend a sport based on the physical parameters a user provides. The second part may predict if an athlete with certain characteristics has a chance to win a medal in their sport.