Università degli studi di Torino

DIPARTIMENTO DI ECONOMIA E STATISTICA "COGNETTI DE MARTIIS"

Corso di Laurea Magistrale in Metodi Statistici ed Economici per le Decisioni

Tesi di Laurea Magistrale

Analysis of technical attributes of male and female national football teams: a comparison through a statistical machine learning approach

Relatore:
Prof.ssa Rosaria Ignaccolo

Correlatore:
Dr. Luca Pappalardo

Candidato:
Giuseppe Pontillo

Anno Accademico 2018/2019

"Up to five goals is journalism.
After that, it becomes statistics."

Author Unknown

Abstract

Too often women's football has been compared to men's football mainly on the basis of the players' physical attributes, offering an incomplete analysis of when the characteristics of *any* football team are studied analytically.

Thanks to the availability of an open *soccer-logs* data set provided by Wyscout, this thesis aims to statistically analyse and compare male and female national football teams based on their *technical qualities*, measured through the *event* data obtained from the last World Cup championships. An *event* could be defined as a certain action, such as a pass, a shot, a foul, a save attempt, and so on, made by a team's player in a match. First results show, for example, that there are significant differences in the number of key playing events, such as *passes*, percentage of *accurate passes* and *free kicks* made by the national teams during a match.

Through the use of particular methods and algorithms [Pappalardo et al., 2019a, Cintia et al., 2015], there were computed variables related to the technical characteristics of a team, such as the average *time between two passes* and the average *ball possession recovery time*, which can also define the *intensity* of a game, and variables that summarize and quantify the individual and collective performance of a team's players within a single value, such as the *H indicator* or the players' *ratings* aggregated for each team via mean and standard deviation. For example, the more the *ratings'* standard deviation, the more, in a particular match, the team was characterized by players that, individually, outperformed respect to their teammates.

Finally, all these features were used into advanced classification algorithms such as Decision Tree, Random Forest and AdaBoost with the task of classifying a team in a game as male (*class 0*) or female (*class 1*). All the classifiers were validated through a 10-fold Cross Validation on a *training* set and they all showed a good predictive performance, indicating that it is possible to distinct a male football team from a female one (and vice versa) on technical skills. Moreover, after fitting a Decision Tree on different versions of *training* set and looking at the importance that each variable had in the decision path every time, we find that the most important differences underlie in variables such as *players' individual performance variability*, *pass velocity*, *ball recovery time* and the *percentage of accurate passes* made by the teams.

Contents

1 Introduction 1

2 Event-Based Soccer Data 5

2.1 Data Description 5

2.2 Number of Events per game 6

2.3 *Passes* and *Shots* Focus 9

2.4 Pass-based Match Intensity features 12

3 Performance Synthetic Indicators 14

3.1 Players' Ratings 14

3.2 Teams' Passing Behaviour 16

3.2.1 The *H Indicator* 18

3.2.2 Team Flow Centrality 19

4 Team Gender Classification Problem 21

4.1 Predicting Performance Evaluation 23

4.2 Feature Importance definition in Decision Trees 28

5 Conclusions and Future Hints 36

Bibliography 40

Chapter 1

Introduction

Female football is said to have developed initially thanks to the independent women belonging to the *Kerr Ladies* team, who gave the greatest impetus to this sport since the early years of the twentieth century [Scardicchio, 2011].

As time passed, the *Kerr Ladies* intrigued the English crowds, not only for their exceptional attire characterized by long skirts and corsets, but also for their ability to stand up to male clubs in numerous charity competitions; the success and enthusiasm of these events led to the formation of about 150 women's teams in England in 1921 [Scardicchio, 2011].

The unexpected popularity of this movement aroused concerns within the English Football Association, which on December 5, 1921 decreed the following statement: "Complaints have been made as to football being played by women, the Council fell impelled to express their strong opinion that the game of football is quite unsuitable for females and ought not to be encouraged. . . the Council request the clubs belonging to the Association to refuse the use of their grounds for such matches" [Scardicchio, 2011].

This measure drastically slowed down the development of women's football, which, after a long period of stagnation, resurfaced in the first half of the 1960s in the Nordic countries of Europe, such as Norway, Sweden and Germany. From that moment on, its development was unstoppable, and from the courtyard of the Dick Kerr factory in Preston (UK), women's football moved to the stadiums of Europe and of the world, carving out, in small steps and not without sacrifices, an important showcase among the most popular sports in the world.

The *Kerr Ladies* English team remained alive for 58 years, winning 758 games out of 828, and scoring about 3,500 goals [Scardicchio, 2011].

Despite they were old times, the *Kerr Ladies* team won several matches against male soccer clubs, demonstrating that the clear physical differences were not enough to stop them.

From 2012 the number of women academies has doubled [Lange et al., 2018] and it is estimated that around 40 million girls and women play soccer worldwide [Pedersen et al., 2019]; the attention that women's football has acquired has stimulated the birth of statistical and mathematical comparisons with male football. Due to the low availability of data on professional women footballers, studies have been mainly developed taking samples from youth and university leagues [Lange et al., 2018, Sakamoto et al., 2012, Gioldasis et al., 2017].

The literature presents quite a lot researches that make a comparison based on physical features, especially because it was easier and cheaper to obtain data on players' physical attributes.

Bradley et al. [Bradley et al., 2014] compared, for instance, 111 soccer players (52 men and 59 women), drawn during a Champions League season (the only exception using a sample with *elite* footballers) with the objective of examining the differences in running performance in terms of distance travelled at high intensity. The players activities were coded in different speed thresholds, and the comparison was made in relation to the distance travelled, given the same threshold; in general, it has been noted that women tend to cover more distance than men at lower speeds, especially in the final minutes of the first half; at higher speed levels, however, men have better performances throughout the game. In particular, female central midfielders tend to cover less distance than their male counterpart, given the speed threshold between 15 and 23 km h^{-1}.

These findings suggest that women's training sessions should focus on muscle enhancement, in order to improve the running intensity, as well as the velocity of a game.

Parallel to the running performance, Sakamoto et al. [Sakamoto et al., 2012] examined the shooting performance, comparing the characteristics of impact on the ball and the body movement before an *instep kick* and an *inside kick*; the sample was made of 34 soccer players (17 boys and 17 girls), belonging to a university league. A *two-way Anova* highlighted women show lower average values than men, and between the two types of shooting, both on ball speed, on foot speed and on the so-called *ball-to-foot velocity ratio* (the latter only in relation to the *instep kick*). The *ball-to-foot velocity ratio* is such that, the longer the distance between the impact point of the ball and the foot's centre of gravity increases, the more it decreases. The female players, on average, make an *instep kick* having the foot's center of gravity further away from the ball than the men do; this is an aspect that could be trained and improved in order to bridge the lower shooting performance due to purely physical aspects of legs strength.

To complete the framework of studies developed on a physical comparison, Pedersen et al. [Pedersen et al., 2019] have hypothesized that the greatest performance differences between male and female soccer can be explained by the fact that women have to adapt to rules and regulations (such as the goal height and length, the size of the playing field, the game duration, and so on) built specifically for the physical characteristics of men. These rules were then scaled in relation to the differences in those physical attributes that are most affected by the rule itself. For example, taking into account the average height difference (±8%) between 20-25 years-old men and women extracted from the Norwegian Directorate for Health, the article supposed that the "fair" goal height should be 2.25 mt, instead of 2.44 mt. Although it goes beyond the goal of this thesis, it might be interesting to simulate a female competition using these new scaled rules and to analyze if there are differences in the team performance compared to reality (having first defined the meaning of *good performance* of a soccer team).

The comparisons between the two disciplines have also developed on the base of technical characteristics and behaviors, although with less interest and precision.

Gioldasis, Souglis and Christofilakis [Gioldasis et al., 2017], in particular, evaluated the technical skills between male and female players, based on their position on the field. Several skill tests were carried out on a sample of 27 female players and 37 male players taken from an amateur youth league; these tests provided information on the technical characteristics of the players, such as the quality of the shots, the quality of the long and short passes, the ability to dribbling and the ability to dribbling after a pass. They used a *one-way Anova* first, and a *post-hoc Tukey Test* then, to analyze between which couples roles the differences in the various abilities were significant. The statistical comparison, however, was made exclusively between the different roles in which the young players were classified, taking into account the male and female players separately; this is because the researchers' ultimate goal was to confirm that, even in women's football, there are differences in technical characteristics between roles.

An article by Sakellaris [Sakellaris, 2017], instead, focused on the average number of goals scored per game in the most important international football competitions (Fifa World Cup, Uefa Euro Cup, Copa America and Olympic Games). The average goals scored per game by the male and female national teams were compared (using the Fifa data source) first graphically, and then through hypothesis tests. In general, women's teams have a higher average number of goals scored per match than their male counterparts; such a result surprised the author on why the female football is still less followed than the male one, despite the greater number of goals, expected from a women's match, should increase the spectacle.

To answer this question, however, it seems inadequate to compare a male team with a female one using just the measurement of goals scored; apart from being a very variable measure, it cannot be an indicator of the spectacular nature of a football match on its own. The lack of a wider confrontation between the two disciplines has inspired the realization of this thesis.

Finally, the literature also presents works in which social attitudes expressed during a football match are compared; in the work of Lange et al. [Lange et al., 2018], for instance, 366 young Dutch footballers were followed in order to check which variables, and if among them the gender too, influence the tendency to help a partner or an opponent in difficulty. The *help* variable was defined as the propensity to stop the game to allow a player's care on the ground and it was quantified as a score (for score details look at [Lange et al., 2018, p. 5]). The results show how in *low-stake* situations, that is when your team is winning or the result is not in doubt, the willingness to help is greater than in *high-stake* situations, when your team is losing or the result is in doubt; moreover, there is a significant difference between men and women, with the latter showing, on average, a greater willingness to help.

All the works described above show a common weakness in having relatively small samples and, especially for women's teams, they have little information available on professional players. Such thesis, however, is fortunate to work on one of the largest *soccer-logs* data set; provided by Wyscout, it is unique in the breadth of sports competitions and registered players.

Therefore, these data make it possible to statistically analyse and compare male and female national football teams based on their technical qualities with more precision and accuracy.

A football technical skill can be quantify in several ways, such as the average number of events generated by a team during a match, the proportion of accurate passes made, the ability to quickly pass the ball between players, the number of play-phases, or actions, occurred in a match, and so on. An *event* could be defined as a certain action, such as a pass, a shot, a foul, a save attempt, and so on, made by a team's player in a match.

Moreover, it is also possible to build synthetic performance indicators that, for example, quantify the passing behaviour of a team in a certain match, such as the *H-Indicator* [Cintia et al., 2015], or highlight and measure the individual performance of the players [Pappalardo et al., 2019a]. These attributes can contribute to the comparison between male and female football teams, making it more accurate and interesting.

The processing of these data allows to answer questions such as: it is possible to distinguish a male team from a female one, based on their technical features? Do they have some technical characteristics in common that makes this distinction difficult to make? This work, at first, focused on the definition of particular technical attributes and statistically verifies the presence of differences between male and female football teams; finally, it used statistical machine learning classifiers that predict a football team as male and or female, in a certain match. It will be interesting to understand *on which particular features* the models will concentrate when they make the classification.

Chapter 2

Event-Based Soccer Data

2.1 Data Description

This work uses data related to the Fifa Men's World Cup 2018, with 101,759 events from the 64 games played and 736 players, and to the Fifa Women's World Cup 2019, with 71,636 events on the 44 matches available and 546 players.[1]

The data are presented in *json* format, characterized by *collections* and *documents*. A document describing a *match*, for example, presents the identities of the two challenging teams through the variable *teamsData*, which contains the variable *formation* (list of sub-documents) with the description of the players on the bench, on the lineup and information on any substitutions. For each player, the following features are described: *playerId*, *assists* (dummy), *goals*, *ownGoals* (dummy), *redCards* (dummy) e *yellowCards* (dummy). Finally, the document presents generic data on the match and its outcome (*winner*, *date*, *referees*, etc...).

Of particular importance are the documents describing an *event*, that is an action performed by a certain player during a particular game. Figure 2.1 shows an example: here we describe a *High pass* made by the Arab player Abdullah Ibrahim Otayf during game "Russia - Saudi Arabia", after about 4 seconds from the kick off. It is also indicated the field position from which the event started and towards which it ended; here it is a forward passage towards the left wing of the field.

Due to the variability of the field length and width from stadium to stadium, the event coordinates presented in the *positions* variable are defined as field percentages in the range [0,100] [Pappalardo and Cintia, 2018].
The female events' data set has the same structure.

The process of obtaining data is manual and, in particular, it "is performed by expert video analysts (the operators), who are trained and focused on data collection for soccer, through a proprietary software (the tagger). The tagger has been developed and improved over several years and it is constantly updated to always guarantee better and better performance at the highest standards. Based on the tagger and the videos

[1] We collect the Wyscout open *soccer logs* data set from [Pappalardo et al., 2019b].

of soccer games, to guarantee the accuracy of data collection, the tagging of events in a match is performed by three operators, one operator per team and one operator acting as responsible supervisor of the output of the whole match" [Pappalardo et al., 2019b].

Since this is not a fully automated process, it is sometimes possible to find some imputation error; fortunately these are very few and do not affect the accuracy of the results.

```
{'eventId' : 8,
 'subEventName' : 'High pass',
 'tags' : [{'id': 1801}],
 'playerId' : 139393,
 'positions' : [{'y' : 53, 'x' : 35}, {'y' : 19, 'x' : 75}],
 'matchId' : 2057954,
 'eventName' : 'Pass',
 'teamId' : 16521,
 'matchPeriod' : '1H',
 'eventSec' : 4.487814,
 'subEventId' : 83,
 'id' : 258612106}
```

Figure 2.1: *Event* document example

2.2 Number of Events per game

Giving a first glance at the data set, it seems useful to start the analysis evaluating if there was a difference in the total number of events occurred per game, regardless of the specific type of event.

In this phase, and even later, we consider only the events generated in the first and second half of a match, without considering the extra times and penalties, so as to make the comparison fairer and avoid any distortions.

Figure 2.2 shows the distribution of the total number of events over the World Cup matches. We observe that a male match presents an average number of events higher than a female one. Assuming *normality* distributions and different variances of the two populations, this statement could be statistically verified with a Welch t-test with Satterthwait degrees of freedom for independent samples. The *null* hypothesis H_0 is such that: $\mu_{Male} - \mu_{Female} = 0$, against the *alternative* H_1: $\mu_{Male} - \mu_{Female} > 0$.

After the implementation of the test, with a level of significance $\alpha = 0.1$, we can confirm that, on average, men players produce a greater number of events than women ($pval = 0.0802$), in a certain match.

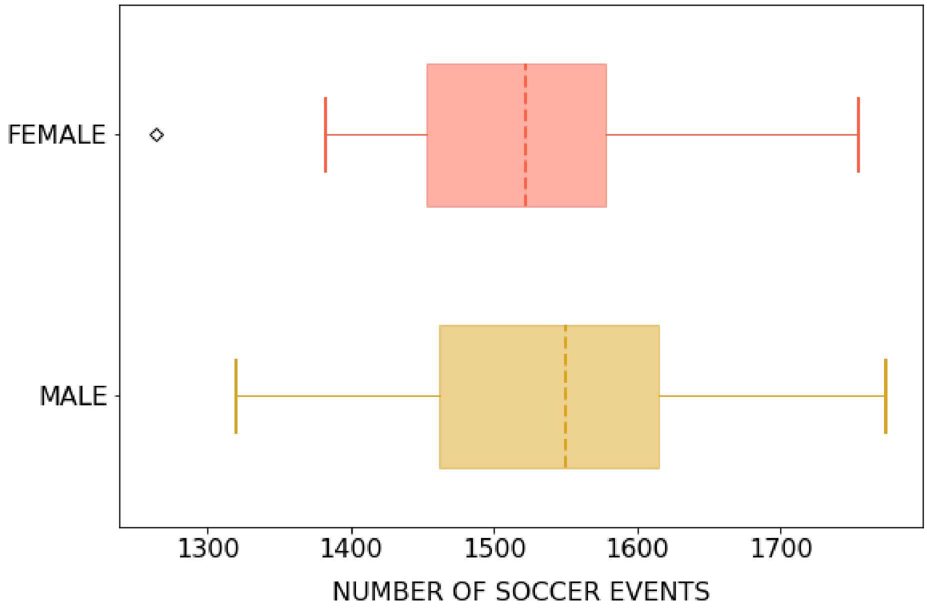


Figure 2.2: Boxplots of the total number of events occurred during the Men World Cup 2018 matches and the Women World Cup 2019 matches. The dashed line represents the averages. There are considered only the events occurred in the first and second half of a match.

Furthermore, we use a *two-way Anova* to verify whether there was also a difference between the two halves of a game and whether there was an interaction effect of sex and periods on the number of events occurred. With this design, the following hypotheses were being tested: $\mu_{Male} = \mu_{Female}$, $\mu_{1H} = \mu_{2H}$ and $\mu_{Gender} \times \mu_{Period} = 0$.

The results show that there is no difference between the *first half* and the *second half* in the number of events occurred in a World Cup match and the interaction term between the two variables has no impact as well ($pval = 0.7278$ and $pval = 0.5316$, respectively).

Based on these findings, we look at *where* this difference on the number of events occurred lies. For this purpose, we analyse the number of the specific events produced in a match. For each type of event, we compute its mean for every team and along the games, and we implement a *two-way Anova* to test the same hypotheses already seen previously (i.e., if the gender and the period variables have an independent significant effect on the occurrence of a particular event, and also if there is an interaction effect between the two factors).

Formally speaking, we are carrying out a so-called *factorial experiment* with two factors A (*gender*) and B (*match period*) with a and b levels respectively (two levels each, in our case), to test if they affect a response variable of interest (the number of a certain type of event). The underlying (linear) model can be written as:

$$Y_{ij} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{2.1}$$
$$\text{for } i = 1, 2, ...a, j = 1, 2, ...b, k = 1, 2, ...n_{ij},$$

"where μ is the overall mean response, τ_i is the effect due to the i-th level of factor A, β_j is the effect due to the j-th level of factor B and γ_{ij} is the effect due to any interaction between the i-th level of A and the j-th level of B" [Croarkin et al., 2006]; n_{ij} is the number of observations in each level combination.

Table 2.1 shows the main results. Since the interaction term was always statistically not significant, we omit it from the Table. We test the normality assumption of the residuals ϵ_{ijk} of the underlying linear model (with interaction term) through a *Jarque-Bera test* which, under the null hypothesis, assumes that the data is normally distributed. Generally speaking, the test statistic is based on two moments of the data, the skewness and the kurtosis, and has an asymptotic χ^2 distribution with 2 degrees of freedom. The test statistic is given by:

$$JB = n\left(\frac{S^2}{6} + \frac{(K-3)^3)}{24}\right) \tag{2.2}$$

where n is the number of data points, S is the sample skewness, and K is the sample kurtosis of the data.

Event	**Normality** ($\alpha = 0.05$)	**Gender**	**Period**
Pass	Yes	M >F (pval=0.0000)	1H >2H (pval=0.0046)
Shot	Yes	M = F (pval=0.6712)	2H >1H (pval=0.0003)
Free Kick	Yes	F >M (pval=0.0002)	1H = 2H (pval=0.5810)
Duel	Yes	F >M (pval=0.0064)	2H >1H (pval=0.0002)
Foul	Yes (but $\alpha = 0.1$)	M >F (pval=0.0003)	1H = 2H (pval=0.1732)
Others on the ball	Yes (but $\alpha = 0.1$)	F >M (pval=0.0332)	1H = 2H (pval=0.1519)

Table 2.1: Table of *two-way Anova* results; they are present just the results related to *gender* and *period* effects on the number of each type of event produced (this is because the *interaction term* was always statistically not significant).

On average, female national football teams produce higher number of *free kicks*, *duels* and *others on the ball* events, which contain accelerations, clearances and ball touches, than their male counterparts in a certain World Cup match; the opposite occurred with regard to the number of *passes* and *fouls*. For most types of events, we find no difference between the first and second half; the only exceptions are in the number of *passes*, *duels* and *shots*. The latter are meanly produced in greater numbers in the second half, when the players could be more tired and reaching the opponent's area could be easier.

Finally, we find no interaction effect of sex and game periods on the number of events produced.

Due to the small sample size, the analysis on *save attempts*, *goalkeeper leaving line* events and *offsides* show poor results, without indicating great differences neither between men and women football players nor between first and second half. It would therefore be necessary to expand the sample, perhaps including data on home leagues, and improve the significance of the results for these kind of events.

2.3 *Passes* and *Shots* Focus

Passes and *Shots* are two essential aspects in the analysis of a football game, and their peculiarities can be considered at the basis of the definition of *a technically strong* team. Going into details, we find interesting results concerning specific types of passes, such as *clearances*, *high passes*, *launches* and their *accuracy percentage* in a match, and in the shooting distance from the goal.

Starting from the *Pass* events analysis, Welch T-tests with Satterthwait degrees of freedom were used in order to statistically verify if there are any differences between male and female national football teams.

On average, we find no difference in the number of *High passes* and *Launches* produced during a match between men and women players ($pval = 0.2383$); this is also true in relation to the percentage of these pass types over the total number of passes occurred in a match ($pval = 0.9101$). We obtain the same result in relation to the number of *clearances* produced in a World Cup match ($pval = 0.8867$ and $pval = 0.6922$, respectively). So perhaps the difference presented in Table 2.1 in the number of *others on the ball* events could be due to differences in the number of accelerations and ball touches, rather than in the number of clearances.

Otherwise, we find no difference in the number of *accurate passes* occurred during a match, and especially in *the percentage* of these over the total number of passes made. In particular, men soccer players, on average, produce a significantly higher number of *accurate passes* than women ($pval = 0.0011$) and a higher percentage of these over the total number of passes ($pval = 0.0012$).

Moving now on the analysis of the *Shot* events, it can be interesting to analyze if, on average, there is a difference in the *distance* from which male and female national teams kick towards the center of the goal.

First of all, since each shot is geolocated in the field, they can be viewed as *spatial points* and it is possible to visualize them through a *Kernel estimate* of the *First Grade Intensity* function $\lambda(s)$. The $\lambda(s)$ function defines the average number of events s_i (the shots), with $i = 1, .., n$, occurred in an infinitesimal area containing the point s in the region of interest R, which in this case is the football field.

Its kernel estimate is defined as:

$$\lambda\hat{(s)} = \sum_{i=1}^{n} \frac{1}{\tau^2} k \left(\frac{s - s_i}{\tau} \right) \tag{2.3}$$

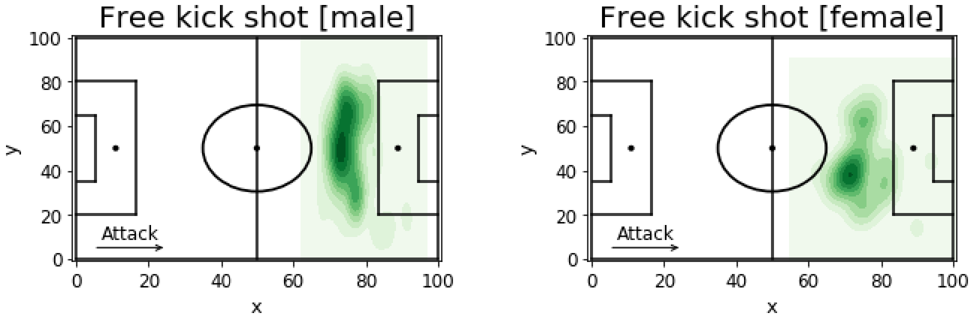
where τ is the *bandwidth* and it indicates the *grade of smoothness* of the estimated map and k is a kernel function.

Figures 2.3 and 2.4 show two examples of kernel estimates, and, in this case, we also consider the *free-kick shots* intensity function.

At first glance, male football players tend to produce more *free-kick shots* from a greater distance than female players; otherwise there is no difference from where the *shots in motion* are made.

We measure every shooting distance in terms of *Euclidean distance* from the *real* origin position to the goal center, i.e., the [100, 50] field position. This is because, based on the data structure, the Euclidean distance from the starting point of a shot to the goal center is the same regardless of where a team is attacking to. Before implementing the tests, we convert the starting coordinates of the shots into meters, using 110 mt as pitch length and 65 mt as its width, for more interpretability.

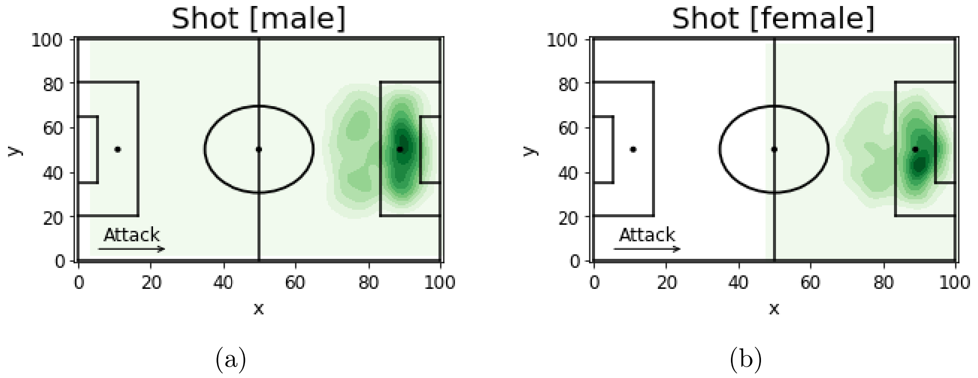
Because of the normality assumption in the *Euclidean distance* distribution was not verified, we use the non-parametric Mann-Whitney U-Test (for details see [Piccolo, 2010, Ch. 18]). Similarly to the T-test, it compares two independent samples obtained from two continuous random variables X and Y, with distribution functions $F_x(u)$ and $F_y(u)$. The test is based on the median and, under the null hypothesis, it assumes the equality of the two distributions: $F_x(u) = F_y(u)$.



(a)

(b)

Figure 2.3: Heatmaps showing the kernel estimate of the First Grade Intensity function $\lambda(s)$, where the *free-kick shots* are the event points s_i and the football field is the region of interest R. The darker is the green, the higher is the number of *free-kick shots* in a specific field zone. The pitch length (x) and width (y) are in the range [0,100], which indicates the percentage of the field starting from the left corner of the attacking team.



(a)

(b)

Figure 2.4: Same as in figure 2.3, but now the *shots in motion* are the event points s_i considered in the football field region R.

The results suggest that, on average, men players, in the 2018 World Cup matches, kick the ball from a greater distance than women players did during the 2019 World Cup, in terms of Euclidean distance from the center of the goal ($pval = 0.0048$).

This outcome, however, has been improved by also taking into account the possible different *perception* of distance that male and female football players may have.

In order to consider this aspect, the attacking midfield has been divided into *three zones* according to the two shooting distributions, i.e. looking at the minimum and the maximum starting position of a shot. The zones were defined in such a way that *Z1* is the area closest to the goal, *Z3* the furthest and *Z2* the middle zone. The female *personal* zones were found to be 1.1 meters closer to the goal.

The realization of a shot can be seen as distributed according to a binomial distribution, but when n is large enough, as in our case (1499 and 1023 shots, respectively for male and female players), we can approximate the binomial distribution to the normal distribution. So then, we use a Z-test for Proportions with two independent samples to verify if, in each *personal* zone, there is a difference in the shooting activity. Female soccer teams have a higher percentage of shots from *their* Z1 zone, with respect to their male counterparts ($pval = 0.014$); the opposite is true in the Z2 *personal* shooting area ($pval = 0.004$). Finally, female soccer teams have a higher percentage of shots from *their* Z3 zone ($pval = 0.022$).

To sum up, according to what we observe and focusing on these particular features, the only significant differences between male and female national football teams lie in the pass accuracy percentage and in the distance from where the shots are made, in a certain World Cup match.

2.4 Pass-based Match Intensity features

The possibility of using the *Actions Split* algorithm [Pappalardo et al., 2019b] and the good adaptability of the data structure, allowed us to further expand the comparison by building additional pass-based variables, which may define the *intensity* of a football match.

The *Actions Split* function splits a match into *ball possession phases*, i.e., given a list of soccer events, it splits them using the following rules: an action begins when a team gains the ball and ends if some of these cases occurs: end of first half or match, the ball goes out of the field, an offside occurs or there is a foul [Pappalardo et al., 2019b]. In women's matches, in particular, there was the presence of the so-called *cooling breaks*, i.e., pauses in the game due to excessive heat; the algorithm recognizes them and indicates them as an additional cause of end of action.

So, by implementing this algorithm on every available World Cup matches, it was possible to extract the following *match intensity* features for all the national teams: *pass velocity*, *ball recovery time* and *shooting time*. The first two variables measure how much time elapsed between two consecutive passes and how long before the same team regains the ball and starts the possession again, respectively; *shooting time* indicates how much time elapsed between *the same* team does two consecutive shots in a certain match. In addition, it has been defined a feature related to the average *passes length* made during a match.

Pass velocity has been defined in such a way that, in a particular match of interest, every time the *receiver* of a pass is *the same* that sent the next pass, the time elapsed between the two consecutive passes is recorded. If the *receiver* does not pass again the ball something happened (duel, foul, touch, etc...), and the time is not counted. In this way, it is also possible to differentiate between the passes of the two opposing teams.

Than, we set times on average for each team respectively, and we do it for every team, in every match.

Ball recovery time, moreover, considers the time elapsed between the last recorded pass of a team and the first new pass made by a player of the same team. Again, we average these times for each team in the same game, and we do it for every team, in every match.

A Mann-Whitney U-Test suggests that, on average, the time elapsed between two consecutive passes is lower for a female national team than a male one, in a certain World Cup match ($pval = 0.0130$); and also, according to the test, female teams regain the ball possession faster than male ones ($pval = 0.0014$).

The ball runs quickly between the feet of the female players, and a game is characterized by a high number of ball possession changes.

Finally, *shooting time* is obtained simply by using all the recorded times of the shots for each team in a match, and subtracting them chronologically two by two, the time elapsed between two consecutive shots was measured. Again, we take the average for each team, and we repeat the process on all the available football matches. For example, in the men's World Cup final, on average, approximately 345 seconds passed between two French shots, and about 281 seconds between two Croats.

However, a Mann-Whitney U-Test shows that there is no evidence to reject the null hypothesis of equal distributions of the time elapsed between two shots, between male and female national football teams ($pval = 0.6827$).

Thanks to the information contained in the data, it is also possible to measure the average *passes length* made by each team in a game. The length is measured in terms of *Euclidean distance* from where the pass started to the point where it ended. Assuming normal distribution of the passes' distance, the Welch T-test suggests that the distance travelled, on average, by the passes made by the male players is greater than that travelled by the female players, in a certain World Cup match ($pval = 0.0002$).

This first analysis of the technical characteristics of male and female national football teams reveals that some differences exist, and that they are mainly found in the number and in the distance from where certain types of event are produced, and in the so-called *match intensity features.*

All these aspects are linked to the collective performance of a team, which could affect the progress of a match and its spectacular nature. The presence of many shots from inside the penalty area may imply a lively and exciting game, or a team that shows a fast and precise style of play could enchant its fans. The results obtained may affirm that the differences between a male and a female football team balance in those characteristics that could define a *beautiful* game. Female players make shorter and faster passes, on average, but with less accuracy than male players. The latter, on the other hand, tend to kick from more prohibitive field positions where the chance to score is lower.

Based on these considerations, it is possible to know more about the playing style of a team by measuring some synthetic indicators that summarized its passing behaviour, such as the *H-indicator* [Cintia et al., 2015] and the *team Flow Centrality*, and that showed how relevant the individual performance of the players was, in a football match.

Chapter 3

Performance Synthetic Indicators

3.1 Players' Ratings

Defining the performance of a team in a match begins with the analysis of its individual parts, that is, the evaluation of its players. Pappalardo et al. [Pappalardo et al., 2019a] designed and implemented a data-driven framework, *PlayeRank*, that offers and quantifies the performance evaluation of soccer players in a match of interest. We use this algorithm to evaluate the performance of male and female football players who participated in the World Cup.

PlayeRank is based on data and, taking into account the different types of events made by the players and their role on the pitch, it aims to evaluate "the performance quality of a player u in a soccer match m" [Pappalardo et al., 2019a], by computing a numerical rating $r(u, m)$, called *performance rating.* So, given a single World Cup match m, the algorithm quantified the performance of a player u in m by averaging a *n-dimensional* feature vector $Q_u^m = [x_1, ..., x_p]$, where each x_j, with $j = 1, ..., p$, represents a feature that describes a certain aspect of u's behaviour that he or she shows in match m.

Some features are simply related to the number of a particular event produced by u in m (number of passes, number of shots, etc.), some others take into account the *outcome* of these events, e.g., if they are *accurate* or *not accurate.* Formally speaking, $r(u, m)$ could be defined as:

$$r(u,m) = \frac{1}{R}\sum_{i=1}^{p} w_j x_j \tag{3.1}$$

The w_j measure the importance of every x_j (in the winning goal) that occurred in a football match and they were estimated in the so-called *learning phase*; R is a normalization constant such that $r(u, m)$ is between 0 and 1.

The *learning phase* is a key phase of the algorithm and it consists in two main steps: feature weighting and role detector training. For the definition and the in-depth analysis of the machine learning models used for these two main goals, see pages 5-9 of [Pappalardo et al., 2019a].

Once the feature weights are estimated, the *performance scores* $r(u, m)$ for each player u in each match m are computed by applying Equation (3.1).

Moreover, in the study of Pappalardo et al. [Pappalardo et al., 2019a], every player is assigned to a role if he played at least 40% of the matches in that role; each role in the field is defined through a K-means clustering method implemented in the role detection phase of the *learning phase*. This is done because, considering that different roles need different technical characteristics, each player should be trained in those skills that most characterize that particular role, and for this they want to evaluate the players' performance based on different roles. However, in this context, we are primarily concerned with the definition of the ratings (also combined with the information of the goals scored) for each player in a game of interest m, without necessarily having information on the role in which they played; therefore, no ratings assessment is made based on the role played in the field.

So then, the *performance ratings* are computed by normalizing the scores and combining them with *goals scored*, given the goal weight assigned arbitrarily (it is set as 10% of total performance score).

For example, Harry Kane (Eng), in the match against Panama, scored three goals and, according to the *PlayeRank* result, showed an individual ratings of 0.99, demonstrating its centrality in the 6 to 1 victory. Similarly, the Australian champion Samantha Kerr, in the match against Jamaica, scored four times and presented a ratings of 0.80.

To obtain an aggregate performance measure for the whole team in a match m, we measure the *mean* and *standard deviation* of each $r(u, m)$ of the players belonging to the same national team. *High values* in the *ratings' standard deviation* suggest that, in a particular match, the team is characterized by players that, individually, outperform respect to their teammates; contrarily, *low values* imply that the performance level among the teammates is not so variable in that match.

For instance, in the match Portugal versus Spain (3-3), Portugal presented a ratings standard deviation of 0.22, which was quite high considering that the average value between just male teams was equal to 0.07; this value could have been due to the great individual performance of Cristiano Ronaldo that scored 3 goals. In the women's World Cup in France, the match United States versus Thailand, which USA won 13-0, was characterized by a great individual performances of Alex Morgan and Rose Lavelle that scored respectively 5 and 2 goals; the ratings standard deviation for United States was 0.24.

The measure of *ratings' standard deviation* is central in the performance evaluation of a team in a match m, as it captures the behavior of the players on the pitch with a single value.

Obviously, not all teams had players that individually affect the performance of the whole team, but we observe that male national teams are more provided in this respect than female ones. For example, 50% of the male teams during the World Cup matches presented a ratings standard deviation value equal or greater than 0.07, and the 25% of them had a value greater than 0.10. As for women's national teams, however, the 50% showed a ratings standard deviation value equal or greater than 0.04, and the 75% of them had a value less than 0.06.

Figure 3.1 shows the histogram of the distribution of the ratings standard deviation values for national football teams during their respective World Cup games.

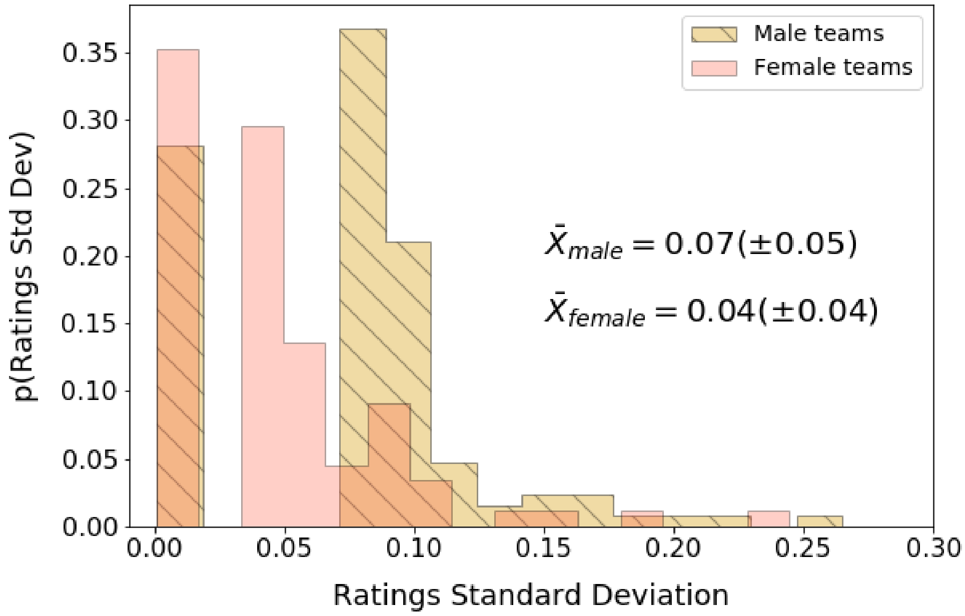


Figure 3.1: Histogram of the distribution of the *ratings' standard deviation* values for male and female national football teams during their respective World Cup games. It is also shown the sample average and standard deviation (in brackets).

This plot also confirms that the distribution of the *ratings' standard deviation* for the men's national football teams is shifted towards higher values than the women's national teams.

These measures that quantify how the individual performance of the players affect the overall performance of the team in a match, are certainly useful and informative in describing the technical aspects of a football team.

3.2 Teams' Passing Behaviour

Since *soccer-logs* include the location of all the passes made by each player during the matches, through the definition of *passing networks*, we can infer the passing behavior of a team within a single value. A team's passing network in a certain match is defined in such a way that a node was a football player and the edges were the passes made by the players. We compute the position of the nodes in the network considering the average players' positions from where they passed the ball. For instance, Figure 3.2 shows the passing network of France in the Russia World Cup final, where each node is labeled with the player identification code and the edges width is weighted with respect to how many times two players have passed the ball to each other.

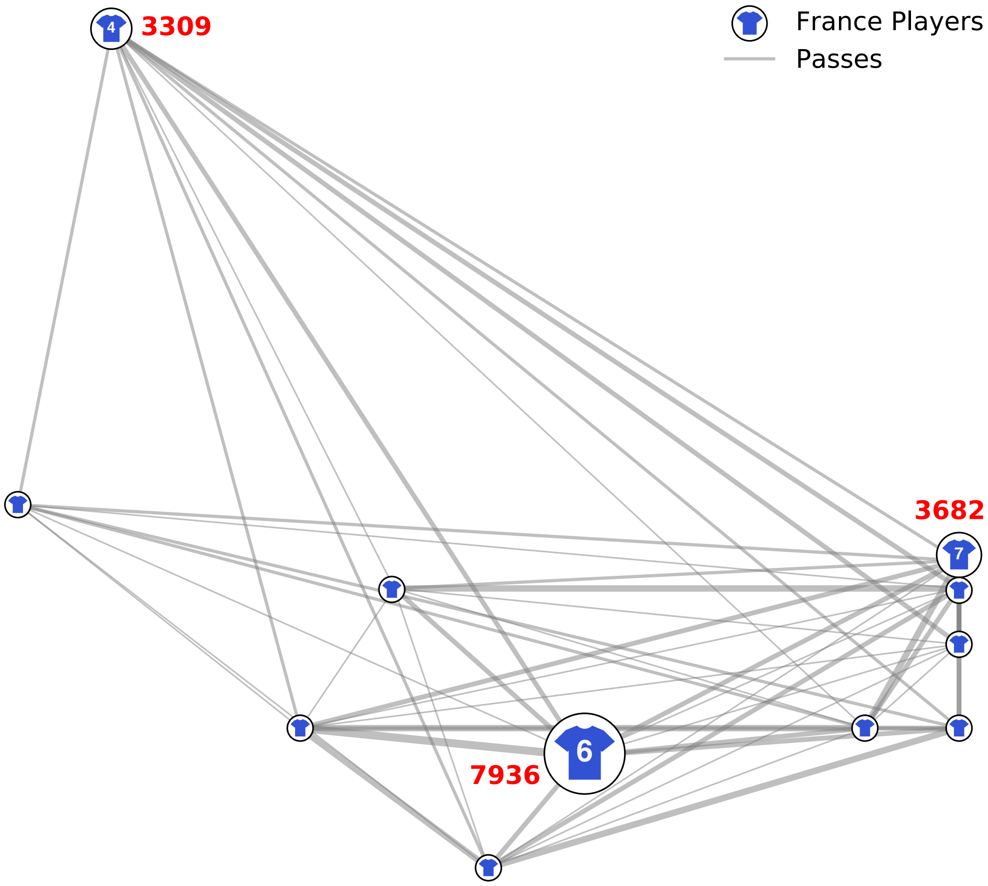


Figure 3.2: Passing network of France in the Russia World Cup final. Each node is labeled with the player identification code and the edges width is weighted with respect to how many times two players have passed the ball to each other. In red there are highlighted the players *Raphaël Varane* (id 3309), *Paul Pogba* (id 7936) and *Antoine Griezmann* (id 3682), who have been most involved in the passing network of the World Cup's finalist. The position of the nodes were computed considering the average players' positions from where they passed the ball. The algorithm used to draw the network was taken and modified as needed from the article [Pappalardo et al., 2019b].

In this work, we measure two indicators based on the passing networks of a team: the *H indicator*, theorized by Cintia et al. [Cintia et al., 2015], and the *team Flow Centrality*.

3.2.1 The *H Indicator*

The *H indicator* [Cintia et al., 2015] summarizes different aspects of the passing behaviour of a team into a single value. All these aspects are related to the pass-based performance features, which were measured using the passing network of a team in a certain match.

First of all, we compute the distribution of passes *over the players* considering the "average amount μ_p of passes managed by players in a team during a game [and] the variance σ_p of the amount of passes managed by players in a team during a game" [Cintia et al., 2015]. The higher σ_p values, the higher is the heterogeneity in the volume of passes managed by the players.

Moreover, we also consider the distribution of passes *over the zones of the pitch*, which is another key aspect of a team's passing behavior. To capture this aspect the football pitch is firstly split into 100 zones, each of size 11 mt x 6.5 mt, and than, a *zone passing network* is drawn, where now the nodes are zones of the pitch and an edge (Z_1,Z_2) represents all the passes performed by any player from zone Z_1 to zone Z_2 [Cintia et al., 2015].

Again, we take the "average amount μ_z of passes managed by zones of the pitch during the game [and] the variance σ_z of the amount of passes managed by zones of the pitch during the game" [Cintia et al., 2015]. High values of σ_z underlies the "coexistence of “hot” zones with high passing activity and “cold” zones with low pass activity during the game; low values of σ_z indicates, however, a more uniform distribution of the pass in game activity across the zones of the pitch" [Cintia et al., 2015].

Finally, we combine these indicators by their harmonic mean in order to summarize the passing behavior of a team into a single value; formally speaking, we have:

$$H_{team} = \frac{5}{(1/w + 1/\mu_p + 1/\sigma_p + 1/\mu_z + 1/\sigma_z)} \tag{3.2}$$

where w is simply the number of passes produced by the team in a match.

Tables 3.1 and 3.2 show the first five male and female national football teams with the highest average *H indicator*, computed across all the matches in the World Cup. According to the indicator, Spain was the strongest national team in the passing performance, so as Japan in the Women's World Cup.

Table 3.1: World Cup Russia 2018

National Team	**mean *H***
Spain	1.67
Egypt	1.60
Denmark	1.59
Australia	1.53
Iran	1.47

Table 3.2: World Cup France 2019

National Team	**mean *H***
Japan	1.56
England	1.53
Chile	1.52
Scotland	1.40
South Africa	1.34

Although it is not the purpose of this work, Cintia et al. demonstrate how "[...] the *H indicator* [italics added] of a team is better correlated with its success (goals, attempts, points) than the mere amount of passes (indicator w), highlighting the usefulness of the defined indicators in capturing important aspects of the performance of football teams" [Cintia et al., 2015]. In this context, the indicator will "feed" the final machine learning classifiers which, given a vector of characteristics of a team in a game, will classify it into a male or female, and it will be evaluated in the role it holds in this task.

3.2.2 Team Flow Centrality

By using team passing networks again, it is possible to measure another aspect of the passing behavior of a team: the centrality that each player has within the network of passes. The *team flow centrality* feature derives from the *player flow centrality* measure, which we compute (and modify as needed) using the algorithm taken from [Pappalardo et al., 2019b].

The *player flow centrality* ranks each player based on his centrality in the network of passes in a certain match. Formally speaking, it measures the current-flow-betweenness-centrality value for each node (remembering that each node is a football player). "*Betweenness centrality* is a widely used measure that captures a [node's] role in allowing information to pass from one part of the network to the other. [...] Technically, it measures the percentage of shortest paths that must go through the specific node. [...] The important thing to know is that betweenness is a measure of how important the node is to the flow of information through a network" [Golbeck, 2015, pp. 221-235]; in this context, it means quantifying how central a player was in passing the ball from one side of the field to the other.

The *team flow centrality* is then defined by setting on average the betweenness flow centrality values of players of the same team in all the matches they played. Moreover, it is computed a function to measure the *variability* σ_f in the passing flow centrality of a team in a match. High values of σ_f highlight that there are players that individually are at the center of a team passing behavior in a particular match; low values of σ_f, otherwise, shows an equilibrium between players of the same team in the flow passing centrality.

For instance, the Swedish women's national team, in the match against United States (which they lost 0-2), showed the minimum σ_f value among all the national teams considered, indicating that, maybe, every player on the field performed quite the same central passing role during the match.

Again in the Women's World Cup, the English team, in the match against Argentina which they beat 1-0, showed the maximum σ_f value among all the national teams, performing a passing centrality behavior more heterogeneous between players. However, no correlation was found between a team's flow centrality value and its success.

All these indicators defined so far are certainly able to give us more information about the playing style of a football team and the technical skills of its players; it will be interesting to understand if these characteristics differ between male and female national teams.

To do this, machine learning classifiers are used with the task of classifying a team as male or female, in a game of interest. In addition to assessing the predictive capacity of each classifier, it will be necessary to understand which variables they possibly focus on to differentiate a male team from a female one, and vice versa.

Chapter 4

Team Gender Classification Problem

To further investigate whether there are technical characteristics of a team that could distinguish between men's and women's soccer teams, we use classification algorithms such as Logistic classifier, Decision Tree, Random Forest and AdaBoost. The task was to classify a national team into male - *class 0* - or female - *class 1* -, in a certain World Cup game g, given the performance vector of Equation (4.1) for each team T:

$$\mathbf{P}_T^{(g)} = [x_1^{(g)}(T), ..., x_p^{(g)}(T)] \tag{4.1}$$

containing all the performance and technical variables x_j, with $j = 1, ..., p$, computed previously.

We use twenty-two features x_j, including: nine counting variables, related to the number of each event type produced, one linked to the percentage of accurate passes, one related to the average Euclidean distance of shots from the goal center, six match intensity features obtained using the *Actions split* algorithm, including, in addition to the four already presented in the Section 2.4, also a feature linked to how much time, on average, a match stand stopped and the average time elapsed between two consecutive duels, and, finally, the team performance indicators described in Chapter 3.

The characteristics of each classifier are here briefly illustrated.

The Logistic model, in general, is used to estimate the relation between p independent variables X_p and one binary (or multi-class) dependent variable y; in particular, it estimates the probability of belonging to a certain class, given the vector of characteristics X, $p(y/X)$. For example, we have:

$$p(y =\hat{} 1/X) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + .. + \hat{\beta}_p X_p)} \tag{4.2}$$

where the $\hat{\beta}_j$ are estimated through *maximum likelihood* method. So each observation will be entrusted to the k-th class if its estimated probability of belonging to that class will be higher (or lower) than a certain probability *threshold* (default 0.5).

The Decision Tree classifier is a powerful method which divides the variable space into M different non-overlapping *regions* and for each observation in the same region R_m, the same forecast is made, given in this case by the most frequent class in the region. Each split is computed through the so-called *recursive binary splitting* method (Section 4.1 for Decision Tree details). The Random Forest "builds a large collection of *de-correlated* trees [on bootstrapped training samples]. [...] Before each split, [it] selects $m \leq p$ of the input variables at random as candidates for splitting." [Hastie et al., 2009, pp. 587-589]. Usually, in classification problems, the tuning parameter m is equal to $\sqrt{p}$ and the minimum region size is one. This method was built to reduce the variability of the Decision Tree classifier.

Finally, the AdaBoost classifier is another approach for improving the predictions from the Decision Tree. Here "[...] the trees are grown *sequentially*: each tree is grown using information from previously grown trees. [AdaBoost] does not involve bootstrap sampling; instead each tree is fit on a modified version of the original data set."[James et al., 2013, p. 321]. The modified version of the data consists in the fact that each *successive* tree is defined by concentrating more (through *weights*) on those observations wrongly classified by the *previous* tree. "The predictions from all of them are then combined through a weighted majority vote to produce the final prediction. [...] Their effect is to give higher influence to the more accurate classifiers in the sequence."[Hastie et al., 2009, p. 338]. Usually are used *stump* trees, i.e. trees with just one split.

We had information on 64 Russia World Cup matches and 44 France World Cup matches, analyzing in particular 32 male and 24 female national teams. Using the classifiers previously mentioned, we carry out two experiments, one in order to evaluate the predictive performance of each model, the other in order to select the variables that most influenced the choices.

We used the algorithms and functions present in the Python *scikit-learn* library, in the section dedicated to classification methods, to implement *all* the machine learning classifiers [Pedregosa et al., 2011]. In particular, we compute the following algorithms: *LogisticRegression*, *DecisionTreeClassifier*, *RandomForestClassifier* and *AdaBoostClassifier*.

4.1 Predicting Performance Evaluation

The first experiment involves dividing the entire data set into *training* (80%) and *validation* (20%) set. On the validation set we compute a 5-fold Cross-Validation to find the best *tuning parameters* of the classifiers (when needed) that maximize the $CV_{(5)}$ *mean accuracy.*

The k-fold Cross-Validation is a re-sampling method which "involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a [test] set, and the method is fit on the remaining k-1 folds" [James et al., 2013, p. 181]; the classification accuracy was then computed using the samples in the held-out fold. "This procedure is repeated k times; each time, a different group of observations is treated as a [test] set."[James et al., 2013, p. 181] The process defines k estimates of accuracy, $Acc_1, Acc_2, ..., Acc_k$, remembering that the accuracy of a classifier is the percentage of test observations correctly classified. The k-fold Cross-Validation estimate is computed by averaging these values:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Acc_i \tag{4.3}$$

For each classifier (with the exception of the Logistic classifier, without *tuning parameters*) and for every set of parameters arbitrarily chosen, we compute the $CV_{(5)}$ *mean accuracy* (4.3), and select the set that maximize it. We compute this procedure by using the *GridSearchCV()* function in the *scikit-learn* library [Pedregosa et al., 2011].

The choice of the *tuning parameters* of the various classifiers is fundamental to control the so-called *bias-variance* trade-off. The *variance* is related to how much $\hat{f}$ (i.e., the estimate of the function that describes a certain relationship between variables) would change if we estimated it using *another* training set; in general, the more a model is flexible, i.e., it fits well the training observations, the more the *variance* is high. The *bias* is the error introduced by approximating a real-life problem, complicated by definition; here the simpler a model, the more $\hat{f}$ is affected by *bias* [Hastie et al., 2009, Ch. 2].

So, the definition of the best *tuning parameters* through cross validation guarantees to obtain the best classifiers that simultaneously present low *variance* and low *bias*, and maximize their team predictive ability as male (*class 0*) or female (*class 1*).

Now, we validate each classifier, implemented with the *best parameters* selected on the validation set, in terms of *accuracy* and *F1-score* through a 10-fold Cross-Validation on the *training* data set. Table 4.1 shows the predictive performance results:

Classifier	Accuracy	F1-Score
Decision Tree	0.86 (±0.07)	0.84 (±0.08)
Random Forest	0.92 (±0.05)	0.89 (±0.08)
AdaBoost.M1	0.92 (±0.05)	0.90 (±0.07)
Logistic	0.89 (±0.08)	0.87(±0.09)
Baseline	0.55 (±0.12)	0.41 (±0.16)

Table 4.1: Table of $CV_{(10)}$ Accuracy and $CV_{(10)}$ F1-score computed on the training data set of each machine learning classifiers used to predict a football team in a game as male (*class 0*) or female (*class 1*). The *baseline classifier* always predicts by respecting the training set's class distribution, which is balanced. The *Accuracy* indicates the percentage of correctly classified teams over the total teams considered, while the *F1-score* is the weighted average of *precision* and *recall*. It is also shown the standard deviation (in brackets).

F1-score is the weighted average of *precision* and *recall*; *precision* is the ratio of correctly predicted *class 1* observations to the total predicted *class 1*; in our problem this is the percentage of teams labeled as female, which are actually female teams. *Recall* is the ratio of correctly predicted *class 1* observations to all the real observations in *class 1*, i.e., the percentage of real female teams identified as such by the classifier.

Therefore, this score takes both *false positives* (women's teams mistaken for men's) and *false negatives* (men's teams mistaken for women's) into account. If the cost of *false positives* and *false negatives* are very different, as in our case, it's better to look at both *precision* and *recall* together, i.e., the *F1-score*. Formally speaking, the *F1-score* is defined as:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.4}$$

Our first important conclusion is that all the classifiers show an overall predictive performance greater than 80% both for *accuracy* and *F1-score*, indicating that there is a difference in technical characteristics between male and female national football teams. Moreover, all the classifiers have a predictive performance significantly better than a classifier which always predicts by respecting the training set's class distribution ($CV_{(10)}Acc = 0.55$, $CV_{(10)}F1 = 0.41$). The models that present the best performance are certainly the Random Forest and the AdaBoost.M1 classifiers, with the latter having a greater $CV_{(10)}F1 = 0.90$ compared to 0.89 of the Random Forest.

The predictive strength of the Random Forest and of the AdaBoost.M1 classifiers compared to the other models is further confirmed when, dividing the *training* set into a *second training* set (70%) and into a *test* set (30%), the predictive performance is assessed on *test observations*, i.e., using observations that the classifiers did not consider during the training phase. In general, such operation is fundamental to understand if a classifier has a good ability in identifying the correct class of any "new" observation (a male or female team in our case) with vectors of characteristics different from those used in the training phase. In this case, we evaluate them through *ROC curves* representation, as shown in Figure 4.1.

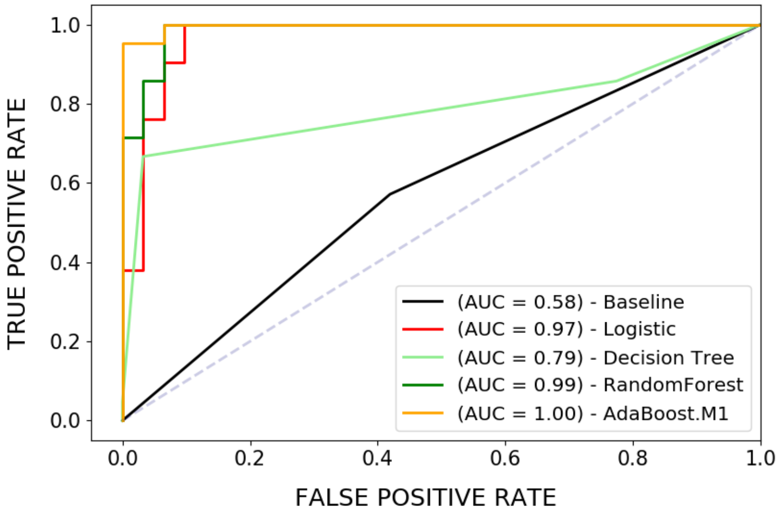


Figure 4.1: ROC curves for the classifiers considered on the *test* set. They trace out the *true positive rate* and the *false positive rate*, as the probability threshold changes. In this case, the *true positive rate* is the percentage of female teams correctly classified and the *false positive rate* is the percentage of male teams mistaken as female, using a given threshold. The actual thresholds are not shown. "The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate" [James et al., 2013, p. 148]. The AUC represents the *area under the curve*, and "[...] the larger the AUC the better the classifier" [James et al., 2013, p. 147]. Random Forest and Adaboost M1 show the best predictive performance using *test* observations.

The ROC curves are usually used to evaluate a classifier's predictive performance with "new" observations never used before by the classifier; in particular, they compare the *true positive rate* (i.e., the percentage of female teams correctly classified, in this case) and the *false positive rate* (i.e., the percentage of male teams mistaken as female), varying the probability classification threshold, i.e., the threshold beyond which an observation is assigned to *class 1* (female team). When the *true positive rate* and the *false positive rate* are both equal to 0, the threshold is equal to 1 (all the test observations are classified in *class 0*). "ROC curves are useful for comparing different classifiers, since they take into account all possible [classification] thresholds." [James et al., 2013, p. 147]. "The overall performance of a classifier, summarized over all possible threshold, is given by the *area under* the (ROC) curve (AUC). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier." [James et al., 2013, p. 147].

Random Forest and AdaBoost.M1 show again the best predictive performance in classifying a team in a game as male or female, also when they use *test* observations.

Although these models are good, they have poor interpretability in terms of identifying the main variables that they used to classify a team as male or female, in a certain game. They both have a *feature important* measure within them, but the Decision Tree, as well as the Logistic classifier, are undoubtedly more interpretable.

Decision Tree and Logistic classifier both have excellent abilities to identify a national football team as male or female in a game, $CV_{(10)}Acc = 0.86$ and $CV_{(10)}Acc = 0.89$, respectively; as in this context, it is interesting to understand which variables have the greatest influence on the decision path, we use the interpretation through the Decision Tree classifier, as it is very intuitive and simple and as it focuses on those variables also considered by AdaBoost classifier (which shows the best predictive performance) as the main ones.

A Decision Tree classifier divides the variable space into M different non-overlapping *regions* and for each training observation in the same region R_m, the same forecast is made, given in this case by the most frequent class in the region. The variable space is divided using the so-called *recursive binary splitting* approach. It consists in choosing that predictor X_j and its value s such that dividing the entire space of the predictors precisely by X_j and its value s, leads to the largest possible reduction in the *Gini index* [Hastie et al., 2009]. The *Gini index* is given by:

$$Gini = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.5}$$

where K is the number of classes (two in our case) and $\hat{p}_{mk}$ is the proportion of training data points that below to the same class k in a particular region m. It can be seen as the variance between the k classes, and the lower $Gini$ is, the more in that region R_m there is a dominance of observations from a certain k class.

After the first split, the division is repeated no longer on the entire space of the predictors, but within *only one* of the two regions chosen randomly, so as to have three regions instead of four. Here again only one of the three regions is divided and the process ends when all the created regions are as *pure* as possible, i.e. they contain observations all coming from the same class.

The tree thus created (T_0), however, may fit too well the training data, causing *overfitting*; than, in order to avoid it, the tree was pruned using the so-called *cost complexity pruning* method. It defines a sequence of sub-trees $T \subset T_0$ "to be any tree that can be obtained by pruning T_0", in function of a set of tuning parameters α, and for each α a tree is defined such that the following is minimized [Hastie et al., 2009]:

$$\sum_{m=1}^{|M|} \sum_{i=k}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) + \alpha \, |M| \tag{4.6}$$

where $|M|$ is the number of regions (or terminal nodes). "The *tuning parameter* [italics added] $\alpha \geq 0$ governs the tradeoff between tree size and its goodness of fit to the data." [Hastie et al., 2009, p. 308]. When $\alpha = 0$ it returns T_0, i.e. the original (overfitted) tree; as α grows, there is a 'cost' to be paid in terms of terminal nodes so that (4.6) is minimized, and for this reason a smaller tree will be defined.

As already seen, we select the best α, and therefore the best T_α, that has the maximum average accuracy on a 5-fold Cross-Validation on the *validation* set, and then we validate the tree on the *training* set through a 10-fold Cross-Validation.

"Trees can be displayed graphically, and are easily interpreted even by a non-expert (especially if they are small)" [James et al., 2013, p. 315] and, moreover, there is no need to scale variables.

So then, we carry out the extrapolation of the main variables used by the Decision Tree classifier through an additional experiment presented in the following section.

4.2 Feature Importance definition in Decision Trees

In the second experiment, we split again the *training* set into *another training* set (70%), which we call *fitting* set, and into a *test* set (30%). This split is repeated N times (twenty times in our case) based on different *random state* values, and we fit the Decision Trees with the *best parameters* on the *fitting* set each time.

Every time a Decision Tree is fitted, we take the *feature importance* measure for *each* variable and, considering two sets of *feature importance* values, $Set^{(i)}$ and $Set^{(j)}$, with $i, j = 0, ...19$ and $i \neq j$, the difference between each variable's values v_r is quantified via Normalized Root-Mean-Squared Error, given in Equation (4.7):

$$NRMSE(Set^{(i)}, Set^{(j)}) = \frac{\sqrt{\frac{1}{p}\sum_{r=1}^{p}(v_r^{(i)} - v_r^{(j)})^2}}{max\, v_r^{(i)} - min\, v_r^{(i)}} \tag{4.7}$$

where p is the total number of features used in the Decision Tree and $v_r^{(i)}$ and $v_r^{(j)}$ are the *feature importance* values for variable r in the "configurations" i and j, respectively.

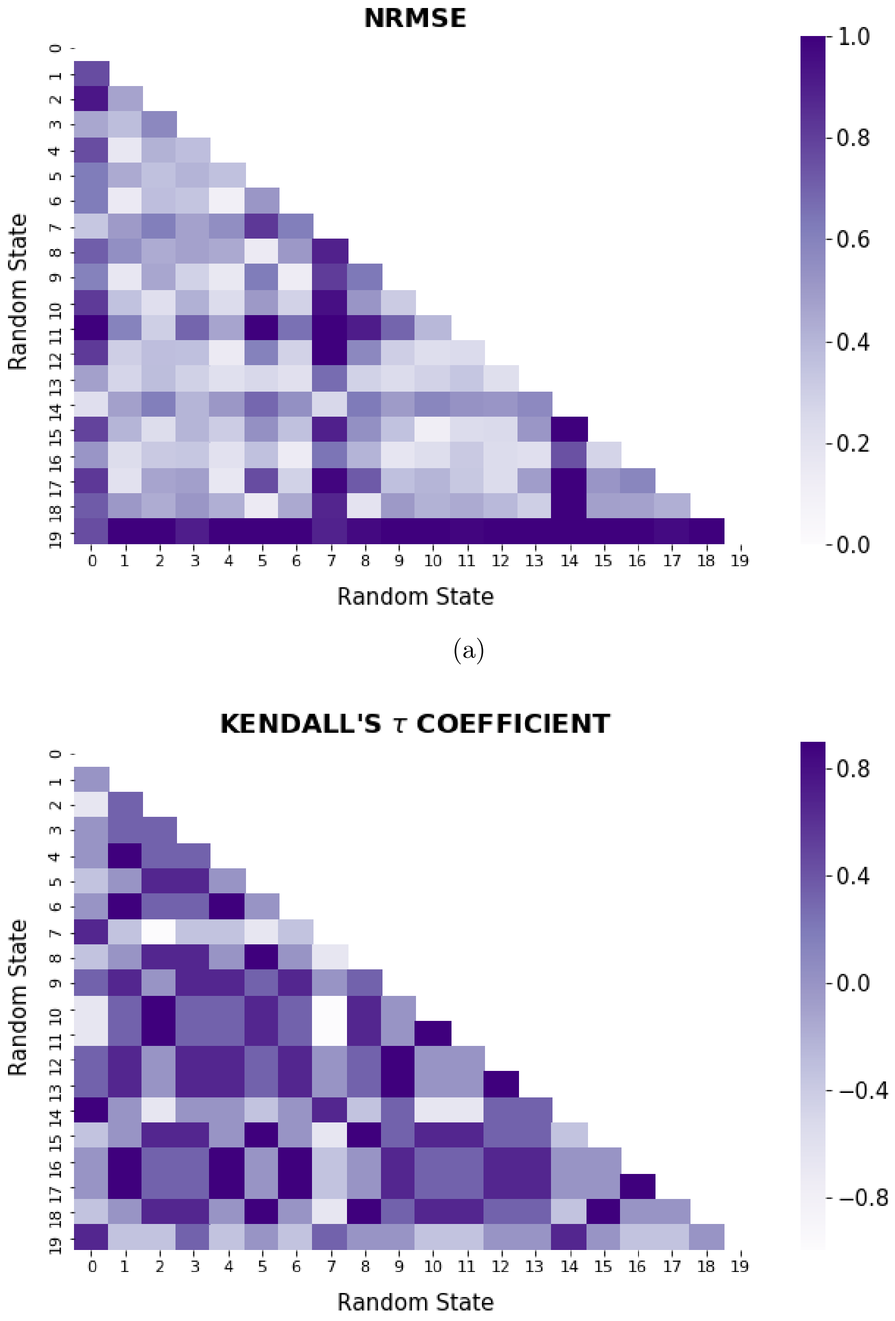
This is done to stabilize the *feature importance* measure of the classifier and reduce the potential variability in the choice of the *best features* that help the Tree in the football team classification task.

Moreover, we sort each *feature importance* set to rank each variable by importance, and we compute the *Kendall's τ correlation coefficient*. In general, the Kendall's τ is a measure of the correspondence between two rankings ($Set^{(i)}$ and $Set^{(j)}$ in our case); values close to 1 indicate strong agreement, values close to -1 indicate strong disagreement. It is given by the Equation (4.8):

$$\tau = \frac{(P - Q)}{\sqrt{(P + Q + T) \times (P + Q + U)}} \tag{4.8}$$

where P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in $Set^{(i)}$, and U the number of ties only in $Set^{(j)}$. If a tie occurs for the same pair in both $Set^{(i)}$ and $Set^{(j)}$, it is not added to either T or U [SciPy community, 2019, pp. 2599-2600].

The experiments show that the Decision Tree classifier, implemented on twenty different *fitting* set configurations, mainly concentrates on four variables to classify a national football team as male (*class 0*) or female (*class 1*), in a certain game: the *ratings'* standard deviation, the average pass velocity time (remembering that it measures the time between two consecutive passes), the average ball possession recovery time and the percentage of accurate passes that each team made in a World Cup match. To consolidate this statement, we measure the NRMSE and the Kendall's τ coefficient on the different Decision Trees, by considering just this four features (Figure 4.2):



(a)

(b)

Figure 4.2: Heatmaps indicating the Normalized Root-Mean-Squared Error (a) and the Kendall's τ coefficient (b) between the set of *feature importance* values and feature rankings $Set^{(i)}$ and $Set^{(j)}$, with $i, j = 0, ...19$ and $i \neq j$. We compute these indexes just considering the *ratings'* standard deviation, the average pass velocity time, the average ball possession recovery time and the percentage of accurate passes variables.

In total there are 190 *unique* pairs of trees (i, j), with $i, j = 0, ..19$ and $i \neq j$. For these four variables the 31,1% of couples (61 out of 190) have a Kendall's τ coefficient equal or greater than 0.6; that is, in 31% of cases there is a good agreement between the trees in believing that these variables have the same rank, which we know to be quite high (for example *ratings'* standard deviation presents an average rank of 1.95 (±1.8), the percentage of accurate passes has one of 2.9 (±1.5), the average ball possession recovery time has an average rank of 3.2 (±1.0) and the average pass velocity time has one of 4.95 (±3.9)). The percentage of concordant trees could be considered *low*, but if for example we add the *Foul* events, the percentage of pairs of trees with a Kendall's τ greater than 0.6 rises to 65%, *but* they agree on considering *Foul* less important (it has, in fact, an average rank of 7.2 (±1.9)).

Looking at the NRMSE, it presents an average of 22% over all the pairs; for these particular features 93 pairs out of 380 (around the 24% of the pairs) have a NRMSE equal or lower than 32% (here I have also considered mutual pairs since the denominator of Equation 4.7 considers the *minimum* and *maximum* only of the first set of values in the pair of sets). However, we notice how the most discordant Tree is that of configuration 19. $Set^{(19)}$ presents a *null feature importance* value both for the *ratings'* standard deviation and for the *average pass velocity*; instead, it admits variables such as the number of *fouls* and *goalkeeping leaving line* events among the most important variables in the classification. Without it, around the 25% of the pairs have a NRMSE equal or lower than 30%, confirming the central role of the four variables mentioned above.

Once we establish on which variables the Decision Tree concentrates the most, we choose one configuration to make predictions on the so-called *test* set, i.e., using games and teams that the classifier has "never seen".

Figures 4.4 and 4.5 graphically capture the test class predictions of the Decision Tree, and its mistakes. In the figures, we consider just the *ratings'* standard deviation, the average pass velocity time (remembering that low values mean faster pass) and the average ball possession recovery time. This is because, as shown in Figure 4.3, they are the first variables that the Tree has used in the classification task.

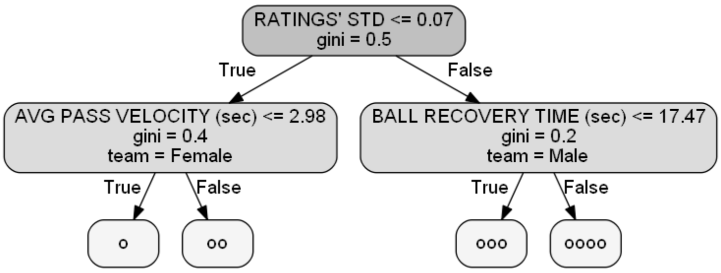


Figure 4.3: The first three main split of the Decision Tree; these could be considered also the most important variables used by the Tree to classify a team in a game as male or female.

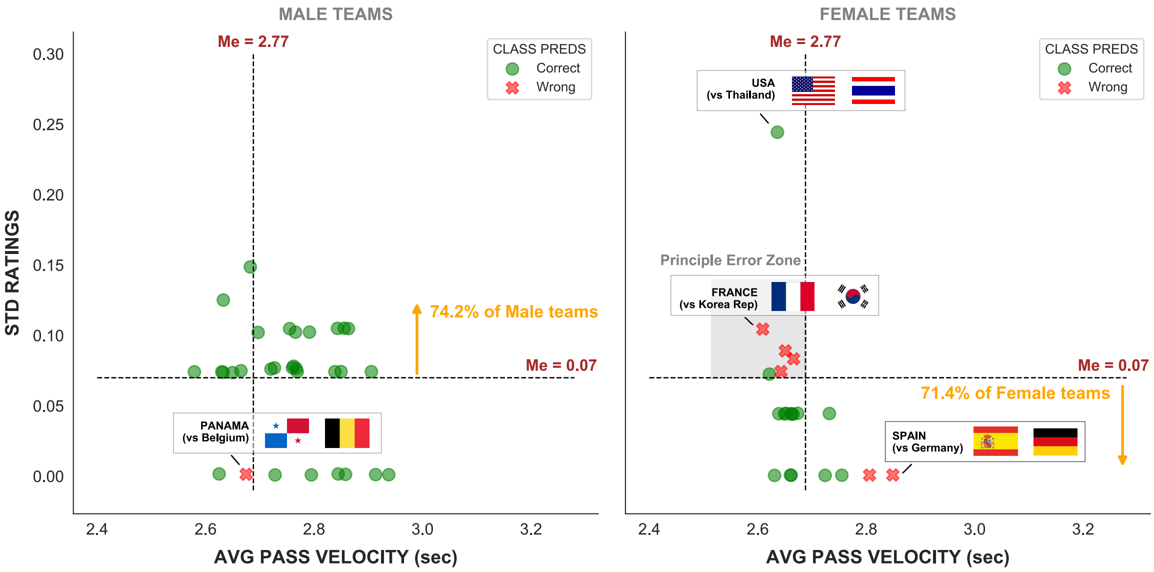


Figure 4.4: Scatter plots displaying *ratings' standard deviation* as a function of the *average pass velocity time* (where low values mean faster passes), among male national teams (left) and female national teams (right), in the *test* set. Information related to the class predicted by the Decision Tree is included: the *circle* indicates a team correctly classified, in a game; the *cross* indicates a mistake. The dashed lines are at the median values for the two variables over the entire *test* set. Note that 74.2% of male teams presented, during the matches played, a *ratings' std* value greater than the median value of 0.07; while 71.4% of female teams presented a value below the median.

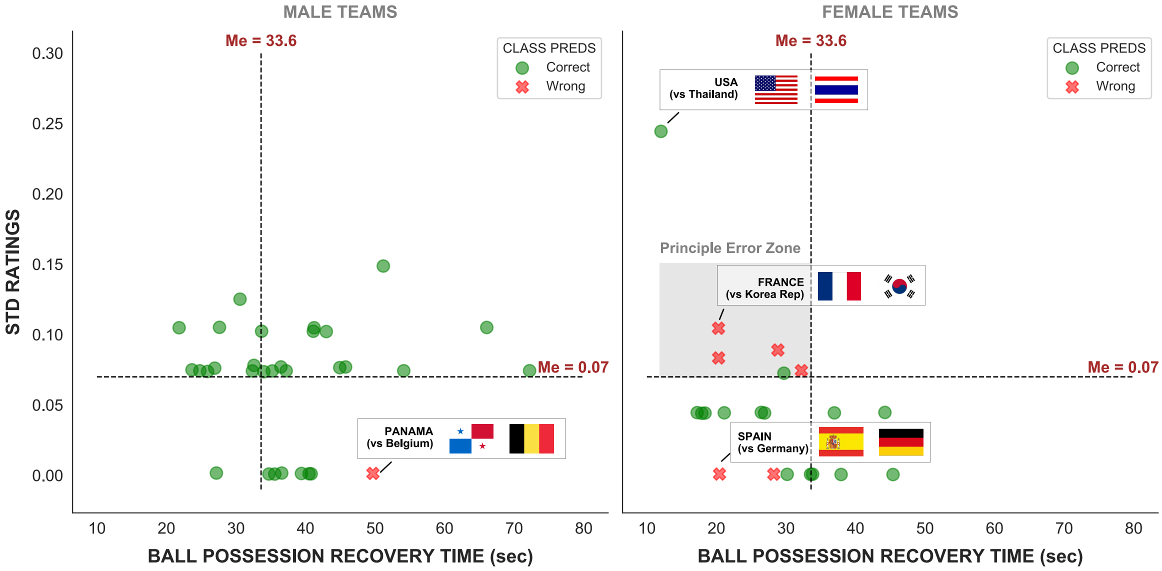


Figure 4.5: Scatter plots displaying *ratings' standard deviation* as a function of the *ball possession recovery time*, among male national teams (left) and female national teams (right), in the *test* set. As in figure 4.4, there is the information of the Tree's test class predictions. The dashed lines are at the median values for the two variables over the entire *test* set.

In six cases a female team has been mistaken as a male one, in a certain World Cup match; the opposite happened once (the red crosses). Comparing the characteristics of the correctly classified teams with those of the teams misclassified by the Tree, allows us to understand how it has behaved in classifying a team as male (*class 0*) or female (*class 1*), in a game.

For example, Uruguay, in its match against Russia which they won 3-0, is correctly classified as male national team. It presents a *ratings'* standard deviation value of 0.10 which, considering that the maximum value of it over all the national teams is 0.26 (shown by England against Panama), is quite high (which suggests that the team was characterized by players that, individually, outperformed respect to their teammates). Moreover, the ball possession recovery time is, on average, 41.2 seconds. So it seems that some few players have had a good individual performance, even if the team's overall ball recovery speed was quite low.

Panama, however, in its match against Belgium which they lost 0-3 (shown in Figures 4.4 and 4.5), is classified as a female team. They present a *ratings'* standard deviation value of 0.001, which is quite low and can indicate that maybe there was a bad team choral performance, with no player standing out from the rest; the average pass velocity time is of 2.68 seconds, and considering that the minimum value is of 2.08 seconds over all the national teams and in every match, they played quite fast (but perhaps without reaching the opposing area with danger); this particular aspect, combined with a percentage of accurate passes of 82.5%, according to the decision path of the Tree, conduces the team to be mistaken as a female one.

The Brazilian women's national football team, in its match against France, which they lost 2-1 at the extra time, presents a *ratings'* standard deviation value of 0.04, which is quite low and indicates that all the players had a similar performance; moreover, they passed the ball quite fast between each other, with an average pass velocity of 2.65 seconds with a great percentage of accurate passes of 97.1%. The latter result could have made the classifier wrong, but the very low *ratings'* standard deviation value allowed Brazil to be correctly indicated as female.

France, however, in its match against Korea Republic which they won 4-0 (highlighted in Figure 4.4 and 4.5), presents a *ratings'* standard deviation of 0.10, which is really high, considering that, just among the female national football teams, the average is 0.04. In particular, they presents an average ball possession recovery time of 20.29 seconds, which is precisely the characteristic for which the classifier wrongly classified it as a male team. Also the percentage of accurate passes was of 88%, making life more difficult for the Decision Tree classifier.

Therefore, from these four examples, it emerges that a national football team is classified by the Decision Tree as a female one if the performance level among the teammates is not so variable over a match (with the exception of USA and France), and if the ball runs quickly between the feet of the players. However, it seems that they are characterized by a lower percentage of accurate passes with respect to male football teams. The latter are instead characterized by high values in the average ball recovery time (in seconds) and by the presence of great individual performances in the various matches.

Of particular interest is the game United States versus Thailand (highlighted in Figures 4.4 and 4.5). Although there are some great individual performances by the US players ($ratings_{std} \approx 0.25$), the average ball possession recovery time is so low that the Decision Tree did not mistake them for male players. So the interaction between these three variables is fundamental for the tree in classifying a football team as male or female.

The last result on the French women's national team, moreover, is doubly interesting since in the *test* set there was also the vector of technical characteristics of Korea Republic in that game; therefore, despite the fact that they are two teams competing in the same game, the Tree indicated one as male and one as female.

Tracing the classifier's predictions for those teams that competed in the same game can be interesting to verify whether the results and the comments made previously are not simply due to chance.

To do this, we consider different *test* sets (again based on the random state value with which the first *training* set was divided); on each set we compute the class predictions, and we isolate the only games with both teams within the *test* set. In particular, we use thirty different *test* sets.

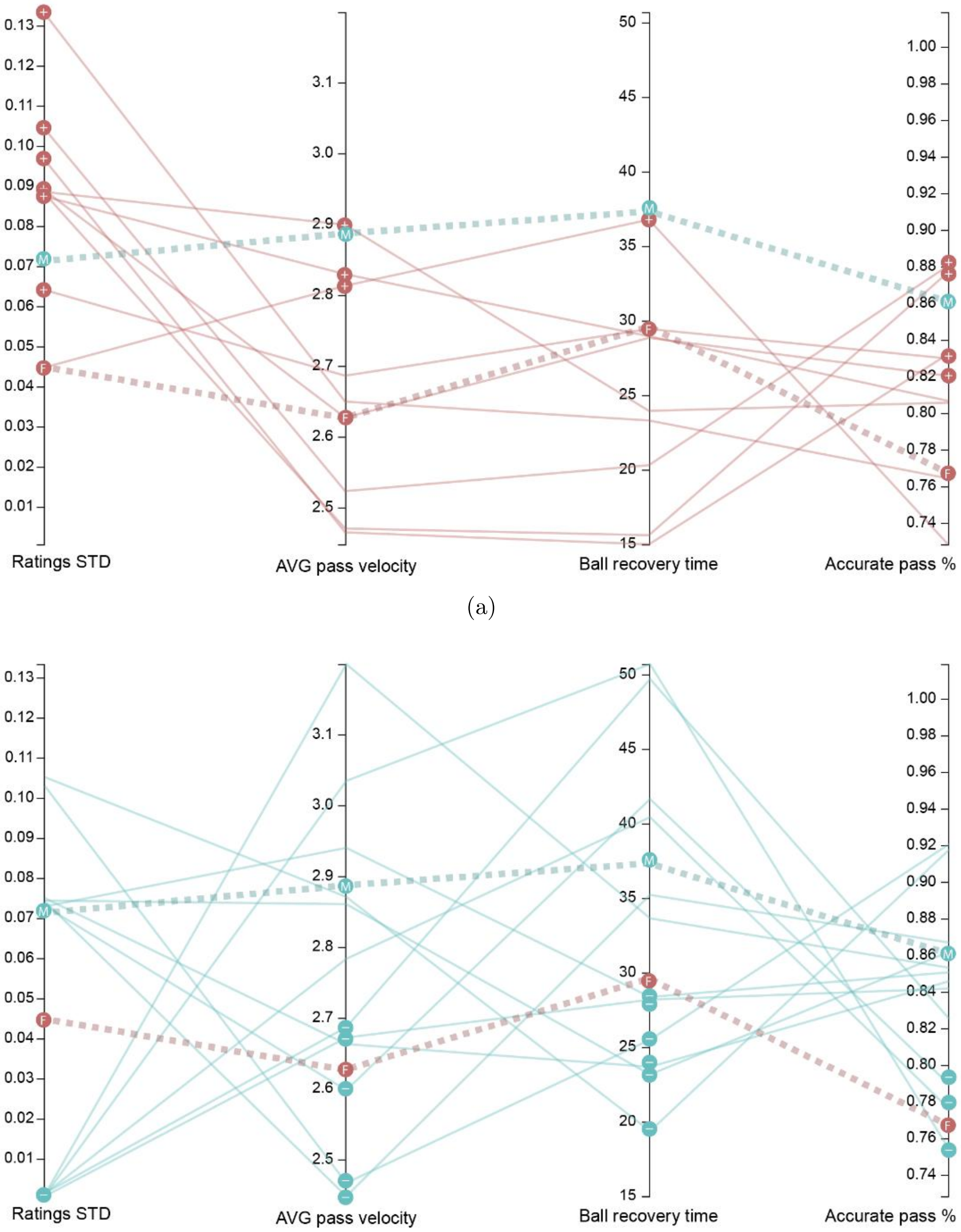
During the experiment, we draw 65 complete matches; of these, 22 games present mistakes in classifying one (or both) of the two opposing teams: 9 are women's football matches and 13 men's.

Among the misclassified female matches, the teams mistaken present a *ratings'* standard deviation of 0.08, on average, which is almost double the average *ratings'* standard deviation value among female teams of 0.04. Moreover, they show, on average, a percentage of accurate passes around 82% (±4%), which is far from the average value of 76.7% (±9%) among female teams.

The wrong male matches, on the other hand, are played by teams with an average *ratings'* standard deviation of 0.05, that is quite lower than the average *ratings'* standard deviation value among male teams of 0.07; they also show a ball possession recovery time value of 32.04 (±10.4) seconds, on average, which differs from the average value of 37.32 seconds (±10.1) among male national teams.

In both cases, the *ratings'* standard deviation (mainly), together with the other main variables described above, have played a fundamental role in confusing the Decision Tree classifier.

Figure 4.6 visualizes the main variables values (i.e., *ratings'* standard deviation, average pass velocity, ball possession recovery time and percentage of accurate passes) of the *wrongly classified* teams; we compare them with the average values of the teams in the class to which they *really* belong.



(a)

(b)

Figure 4.6: *Parallel coordinates* plots showing the values of the main variables of the *wrongly classified* teams. (a) shows the female teams mistaken for male, while (b) shows the male teams wrongly classified. The dotted lines represent the average values, considering the teams in the class to which they really belong, respectively for male (blue) and female (red) teams. The circles highlight those particular values of the variables that may have confused the Decision Tree in the classification between male and female teams. For example, most of the wrong male teams have *ratings'* standard deviation much lower than the male average; while female teams mistaken for male, have values far above the average both for the *ratings'* standard deviation and for the pass accuracy.

Mistaken female teams show values close to the male averages, especially for *ratings'* standard deviation and percentage of pass accuracy. Mistaken male teams, however, show low *ratings'* standard deviation, and are characterized by an average pass velocity and ball possession recovery time closer to those values that the Decision Tree believes to be common for a female team.

So, we observe again the central role of these technical characteristics in distinguishing between male and female teams.

Table 4.2: Summary of results

Question	Answer
It is possible to distinguish a team in male or female, based on technical skills?	Yes it is. All the classifiers show an overall predictive performance greater than 80% both for $CV_{(10)}$ Accuracy and $CV_{(10)}$ F1-score.
Given its ease of interpretation, on which variables does the Decision Tree classifier primarily focus on?	It mainly focuses on four variables: the *ratings'* standard deviation (which also capture the individual player performance), the average pass velocity time, the average ball possession recovery time and the percentage of accurate passes.
Is the importance of these variables stable?	Yes it is. The NRMSE and the Kendall's τ coefficient, computed between different Tree implementations, confirm the central role of these variables.
When is a *male* team mistaken for a female team?	When the performance level among the teammates is not so variable, and the velocity between two consecutive passes, as well as the ball recovery time are low.
When is a *female* team mistaken for a male team?	When the percentage of accurate passes is high, but the recovery of ball possession is slower; and the team is characterized by the presence of great individual performances.

Chapter 5

Conclusions and Future Hints

In this thesis, the availability of event-based *soccer-logs*, provided by Wyscout and related to the Russia 2018 and France 2019 World Cup championships, allowed us to analyse national football team *technical* and *performance evaluation* characteristics, which were then used to compare male and female teams. The main objective was to statistically verify whether a comparison between the two disciplines on these variables was possible and whether it led to some interesting results. All functions and results are also *reproducible*, since the data provided by Wyscout are *open.* We storage the data in a *MongoDB* database, and we use the *pymongo* library as driver.[1]

We obtain many surprising results, observing differences in practically all those variables that can summarize the technical quality of a team. For instance, a Mann-Whitney U-Test suggested that, on average, the time elapsed between two consecutive passes is lower for a female national team with respect to a male one, in a certain World Cup match ($pval = 0.0130$); and also, according to the test, female teams tend to regain the ball possession faster than male ones ($pval = 0.0014$).

Male national teams, however, produced a significantly higher number of *accurate passes* ($pval = 0.0011$) and a higher percentage of these over the total number of passes ($pval = 0.0012$). Moreover, in the 2018 World Cup matches, men players kicked the ball from a greater distance than women players did during the 2019 World Cup, in terms of euclidean distance from the center of the goal ($pval = 0.0048$).

All these results are novel and not yet present in literature, which instead focuses mainly on a physical comparison between the two disciplines.

The use of machine learning classifiers further enriched the analysis: firstly by confirming that differences in technical characteristics between male and female national teams could be identified during their respective world championships; then, thanks to the use of Decision Trees, it became evident *where* these differences were more pronounced. In particular, by training the Trees on different *training* sets, we found that, based on the variable rankings, given by the *feature important* measure, the *ratings'*

[1] The following link `https://github.com/beppontillo/SoccerAnalytics` is associated with the GitHub repository containing *all* the codes and functions used in the thesis.

standard deviation, the average pass velocity time, the ball possession recovery time and the percentage of accurate passes were the main distinguishing variables. In fact, for these four variables, the 51,5% of pairs of trees had a Kendall's τ coefficient equal or greater than 0.6; that is, in 51% of cases there is a good agreement between the trees in believing that these variables have the same rank, which we know to be quite high (Section 4.2 for details).

Concentrating on a particular Tree's decision path, it was found that women's national teams, on average, seemed to be characterized by players of equal technical values (low *ratings'* standard deviation values), contrary to what happens for men's teams; the latter, in fact, were usually characterized by the presence of players who individually lead their team to victory, or who, in any case, influenced the collective performance of the team.

Generally speaking, supporters are often attracted by such great players, who can determine the following for an entire sport. Sports competitions of this kind always concentrate the presence of the best players in the world, and this has happened no less for the women's France 2019 World Cup championship, especially in countries where women's football is usually less followed.

Finally, no consideration has been made in this thesis with respect to any *correlations* that these variables may have with the *success* of a team, and if there may be any further differences between male and female teams in this aspect. This, combined with the possible use of larger data sets, could be a good starting point from which to continue the comparison between the two disciplines, which increasingly fascinates soccer analysts.

List of Figures

2.1 *Event* document example 6

2.2 Boxplots of the total number of events occurred during the Men World Cup 2018 matches and the Women World Cup 2019 matches. 7

2.3 Heatmaps describing the pitch zones from where the *free-kick shots* were more frequently made by male and female players, during their respective World Cup championships. 10

2.4 Heatmaps describing the pitch zones from where the *shots in motion* were more frequently made by male and female players, during their respective World Cup championships. 11

3.1 Histogram of the distribution of the *ratings' standard deviation* values for male and female national football teams during their respective World Cup games. 16

3.2 Passing network of France in the Russia World Cup final. 17

4.1 ROC curves for the Baseline classifier, Logistic classifier, Decision Tree, Random Forest and Adaboost classifier on the *test* set. 25

4.2 Heatmaps indicating the Normalized Root-Mean-Squared Error (a) and the Kendall's τ coefficient (b) between the set of *feature importance* values and feature rankings $Set^{(i)}$ and $Set^{(j)}$, with $i, j = 0, ...19$ and $i \neq j$. 29

4.3 The first three main split of the Decision Tree classifier. 30

4.4 Scatter plots displaying *ratings' standard deviation* as a function of the *average pass velocity time*, among male national teams (left) and female national teams (right), in the *test* set. 31

4.5 Scatter plots displaying *ratings' standard deviation* as a function of the *ball possession recovery time*, among male national teams (left) and female national teams (right), in the *test* set. 31

4.6 *Parallel coordinates* plots showing the values of the main variables of the *wrongly classified* teams. 34

List of Tables

2.1 Table of *two-way Anova* results. 8

3.1 List of the first five male national football teams with the highest average *H indicator* during the World Cup Russia 2018 games. 19

3.2 List of the first five female national football teams with the highest average *H indicator* during the World Cup France 2019 games. 19

4.1 Table of $CV_{(10)}$ Accuracy and $CV_{(10)}$ F1-score computed on the training data set of each machine learning classifiers used to predict a football team in a game as male (*class 0*) or female (*class 1*). 24

4.2 Summary of results 35

Bibliography

[Bradley et al., 2014] Bradley, P. S., Dellal, A., Mohr, M., Castellano, J., and Wilkie, A. (2014). Gender differences in match performance characteristics of soccer players competing in the uefa champions league. *Human Movement Science*, 33:159 – 171.

[Cintia et al., 2015] Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., and Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

[Croarkin et al., 2006] Croarkin, C., Tobias, P., Filliben, J., Hembree, B., Guthrie, W., et al. (2006). Nist/sematech e-handbook of statistical methods. *NIST/SEMATECH, July. Available online: http://www.itl.nist.gov/div898/handbook.*

[Gioldasis et al., 2017] Gioldasis, A., Souglis, A., and Christofilakis, O. (2017). Technical skills according to playing position of male and female soccer players. *International Journal of Sport Culture and Science*, 5:293 – 301.

[Golbeck, 2015] Golbeck, J. (2015). *Introduction to Social Media Investigation: A Hands-on Approach.* Syngress, first edition.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Series in Statistics, second edition.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[Lange et al., 2018] Lange, P. A. M. V., Manesi, Z., Meershoek, R. W. J., Yuan, M., Dong, M., and Doesum, N. J. V. (2018). Do male and female soccer players differ in helping? A study on prosocial behavior among young players. *PloS one*, 13(12)(e0209168).

[Pappalardo and Cintia, 2018] Pappalardo, L. and Cintia, P. (2018). Quantifying the relation between performance and success in soccer. *Advances in Complex Systems*, 21(03n04):1750014.

[Pappalardo et al., 2019a] Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., and Giannotti, F. (2019a). Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5).

[Pappalardo et al., 2019b] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019b). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(236):1–15.

[Pedersen et al., 2019] Pedersen, A. V., Aksdal, I. M., and Stalsberg, R. (2019). Scaling demands of soccer according to anthropometric and physiological sex differences: a fairer comparison of men's and women's soccer. *frontiers in Psychology*, 10:762.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Piccolo, 2010] Piccolo, D. (2010). *Statistica*. Il Mulino, Bologna, third edition.

[Sakamoto et al., 2012] Sakamoto, K., Hong, S., Tabei, Y., and Asai, T. (2012). Comparative study of female and male soccer players in kicking motion. *Procedia Engineering*, 34:206 – 211. Engineering of sport conference 2012.

[Sakellaris, 2017] Sakellaris, D. (2017). The In-Game Comparison Between Male and Female Footballers. *Statathlon*.

[Scardicchio, 2011] Scardicchio, A. (2011). *Storia e storie del calcio femminile*. Lampi di Stampa, Milano.

[SciPy community, 2019] SciPy community (2019). Scipy reference guide, release 1.4.1. *Available online: https://docs.scipy.org/doc/scipy/scipy-ref-1.4.1.pdf*.

Dedication

Alla prof.ssa Rosaria Ignaccolo e al dr. Luca Pappalardo, che mi hanno dato l'opportunità di portare avanti questo progetto di tesi tra l'università di Torino e il CNR di Pisa. Ricevere i vostri insegnamenti e consigli è stata una fortuna e li porterò con me tutta la vita.

A Michela, Paolo, Alessio e Daniele, che in questi mesi mi hanno accolto e sono stati sempre disponibili ad aiutarmi e ad insegnarmi l'utilizzo di tutti gli strumenti di cui avevo bisogno.

Ai miei genitori e a mio fratello, grazie ai quali sono la persona che sono e ai quali posso dire solo un sincero grazie. Un grazie anche ad Amelia, Peppe, Elisa, Pasquale e Riccardo, che mi hanno ospitato e permesso di conoscere il calore accogliente della città di Livorno.

Alla mia P, per cui rivolgerle un semplice grazie non sarà mai abbastanza. Ti sei dedicata a questa esperienza quasi quanto me e, seppur lontani, non mi hai mai fatto sentire solo.

A Mike, ormai più un fratello che un amico e alle nostre chiamate del Giovedì. E infine, sicuramente non per importanza, a tutti gli Scarra, che dovunque mi trovi, mi fanno sempre sentire con loro, a casa.