

# Contents

<b>Contents</b>	<b>xiii</b>
<b>1 Justification, Objectives and Contributions</b>	<b>1</b>
1.1 Sports Analytics . . . . .	1
1.2 Objectives of this thesis . . . . .	11
1.3 Contributions . . . . .	15
<b>2 Quality or chance? Application of machine learning and multivariate statistics techniques to improve the decision making process</b>	<b>17</b>
2.1 Introduction . . . . .	18
2.2 Material and methods . . . . .	20
2.3 Results . . . . .	30
2.4 Discussion . . . . .	40
2.5 Conclusion . . . . .	42
<b>3 Exploring the success of “Big Five” football teams with Multivariate Statistics techniques</b>	<b>43</b>
3.1 Introduction . . . . .	44
3.2 Material and methods . . . . .	46
3.3 Results . . . . .	51
3.4 Discussion . . . . .	67
3.5 Conclusion . . . . .	69
3.6 Appendix . . . . .	70

<b>4</b>	<b>Using the Skellam regression model in combination with the Random Forest algorithm to predict match results</b>	<b>77</b>
4.1	Introduction . . . . .	78
4.2	Material and methods . . . . .	79
4.3	Results . . . . .	84
4.4	Discussion . . . . .	97
4.5	Conclusion . . . . .	101
4.6	Appendix . . . . .	102
<b>5</b>	<b>Development of popularity indicators with Google Trends to measure popularity influence on the market value of players</b>	<b>113</b>
5.1	Introduction . . . . .	114
5.2	Material and methods . . . . .	117
5.3	Results . . . . .	124
5.4	Discussion . . . . .	130
5.5	Conclusion . . . . .	131
<b>6</b>	<b>General Conclusions</b>	<b>133</b>
6.1	Achievement of the objectives . . . . .	134
6.2	Future Lines . . . . .	136
	<b>Bibliography</b>	<b>139</b>
	<b>Abbreviations and acronyms</b>	<b>161</b>
	<b>Parameters and nomenclature</b>	<b>165</b>

# List of Figures

1.1	Categorisation of sports analytics studies as a function of two levels of analysis: the nature of the data available and the main objective of the studies.	3
2.1	Diagram of the double cross validation used to evaluate the classification models.	28
2.2	Cumulative explained variance ratio vs. the number of PCs.	31
2.3	SPE of the PCA model with nine PCs for teams.	32
2.4	Hotelling's $T^2$ chart of the PCA model with nine PCs for teams.	32
2.5	PCA scatterplot of team scores in the first two PCs (distribution of teams according to ranking; projected in PC1 / PC2) with indication of their position.	33
2.6	Multiple comparisons of the models (X-axis) vs. the MCC (Y-axis) as a function of the data balance. The dots indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences. The colour of the intervals indicates whether the MCC results correspond to a balance (blue) or unbalanced (yellow) data set.	34
2.7	Radar plot to compare the mean values of statistically significant game actions to differentiate between positions of the bottom, middle and top teams.	36
2.8	Radar plot for the comparison of teams misclassified as bottom with the mean values of the game actions (statistically significant to differentiate between positions) of the middle teams.	37
2.9	Radar plot for the comparison of the teams misclassified as middle with the mean values of the game actions (statistically significant to differentiate between positions) of the bottom teams.	38

2.10	Radar plot for the comparison of the teams poorly classified as top with the mean values of the game actions (statistically significant to differentiate between positions) of the middle teams. . . . .	38
2.11	Radar plot for the comparison of the teams poorly classified as middle with the mean values of the game actions (statistically significant to differentiate between positions) of the top teams. . . . .	39
3.1	PLS-DA regression coefficients with 95% jackknife confidence intervals for verifying no different behaviour on the top teams depending on the leagues	52
3.2	PLS-DA regression coefficients with 95% jackknife confidence intervals for verifying no different behaviour on the bottom teams depending on the leagues	53
3.3	Cumulative explained variance ratio vs. the number of PCs . . . . .	54
3.4	SPE of the PCA model with seven PCs for teams . . . . .	55
3.5	Hotelling's $T^2$ chart of the PCA model with seven PCs for teams . . . . .	55
3.6	PCA scores scatterplot of the teams and leagues projected in the PC1/PC2 space: top teams in blue and bottom teams in red . . . . .	56
3.7	PCA loadings scatterplot of the variables in the PC1/PC2 space sized by a variable's correlation strength to PC1. The colour of the dots indicates the negative (blue) or positive (red) correlation of the variables with PC1. Orange dotted arrow indicates the direction of the most discriminating PC	57
3.8	SPE of the PLS-DA model with two PCs for teams . . . . .	58
3.9	Hotelling's $T^2$ of the PLS-DA model with two PCs for teams . . . . .	58
3.10	PLS-DA scores scatterplot of the distribution of the teams and leagues projected in the PLS-DA1/PLS-DA2 space: top teams in blue and bottom teams in red . . . . .	59
3.11	PLS-DA weightings scatterplot showing the relationship between the explanatory variables and the response variables in the PLS1/PLS2 space . .	60
3.12	Importance of the variables in the model PLS-DA . . . . .	61
3.13	PLS-DA regression coefficients with 95% jackknife confidence intervals for the variables to predict the bottom teams . . . . .	62
3.14	Multiway importance plot with mean decrease accuracy (MDA) and mean decrease Gini (MDG) . . . . .	63
3.15	Multiple comparisons of the models (X-axis) vs. the AUC (Y-axis). The black points indicate the mean AUC for each model, and the intervals are based on 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences. .	66
A.1.	Boxplot with standardised values for the Top teams in each league . . . .	70
A.2.	Boxplot with standardised values for the bottom teams in each league . . .	71
A.3.	PCA scores scatterplot of the teams and leagues projected in the PC3/PC4 space: top teams in blue and bottom teams in red . . . . .	74

---

A.4.	PCA scores scatterplot of the teams and leagues projected in the PC5/PC6 space: top teams in blue and bottom teams in red . . . . .	75
A.5.	PLS-DA regression coefficients with 95% jackknife confidence intervals for the variables to predict the top teams . . . . .	75
4.1.	Twenty most important explanatory variables in each league, according to the RF, for predicting goal difference (Z) - Seasons 2019/2020 and 2020/2021	86
4.2.	Multiple comparisons of the leagues (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each league, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2019/2020 . . . . .	93
4.3.	Multiple comparisons of the leagues (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each league, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2020/2021 . . . . .	94
4.4.	Radar chart to compare the mean values of the main variables selected by PLS-DA and RF differentiating by season (2019/2020 (solid line) and 2020/2021 (dashed line)) and match result: win (green), loss (red) and draw (yellow) . . . . .	96
4.5.	Multiple comparisons of the models (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2019/2020 . . . . .	98
4.6.	Multiple comparisons of the models (X-axis) vs the MCC (Y-axis). The black points indicate the mean MCC for each model, and the intervals are based on the 95% Fisher's least significant difference (LSD) procedure. Models whose intervals do not overlap indicate statistically significant differences - Season 2020/2021 . . . . .	99
B.1.	Violin plot in combination with the box plot to compare the distribution of the MCC (Y-axis) depending on the league and model: PLS-DA (grey), RF (yellow) and SRM (blue) - Season 2019/2020 . . . . .	110
B.2.	Violin plot in combination with the box plot to compare the distribution of the MCC (Y-axis) depending on the league and model: PLS-DA (grey), RF (yellow) and SRM (blue) - Season 2020-2021 . . . . .	110
5.1.	Contribution of variables fitted by the RF method. . . . .	127
5.2.	Contribution of variables fitted by the GBM method. . . . .	128

# List of Tables

1.1	Empirical studies using eventing data . . . . .	5
1.2	Empirical studies using tracking data . . . . .	7
1.3	Empirical studies using global positioning systems (GPS) data . . . . .	9
1.4	Empirical studies on the causes of injury in elite football . . . . .	11
2.1	Variables classified by type of game actions . . . . .	21
2.2	Confusion matrix showing the distribution of predictions at TP, FN, FP and TN for a classification model . . . . .	29
2.3	Statistically significant variables ( $p$ -values $<0.05$ ) to differentiate among top, middle and bottom teams . . . . .	30
2.4	MCC values of the supervised learning models for unbalanced and balanced data. . . . .	33
2.5	General confusion matrix of the RF algorithm. . . . .	35
3.1	Comparison of the statistically significant variables ( $p$ -values $<0.05$ ) in the PLS-DA, RF and LR (thresholds 2.5, 5 and 10) models . . . . .	65
3.2	Statistically significant variables ( $p$ -values $<0.05$ ) for the two-sample test (top vs. bottom teams) . . . . .	67
A.1.	Mean and standard deviation of the variables for the top teams in the “Big Five” . . . . .	72
A.2.	Mean and standard deviation of the variables for the bottom teams in the “Big Five” . . . . .	73

4.1.	Most influential explanatory variables to predict the goal difference (Z) and the corresponding league and team they belong to, according to the RF, after discarding variables with a correlation higher than 0.7 in each league for both seasons . . . . .	87
4.2.	Regression coefficients and statistical significance of the most influential explanatory variables of the fitted SRM after discarding variables with a correlation higher than 0.7 - Seasons 2019/2020 and 2020/2021 . . . . .	89
4.3.	Goodness-of-fit statistics of the SRM for the “Big Five” - Seasons 2019/2020 and 2020/2021 . . . . .	91
4.4.	Sensitivity, Specificity, and MCC (Means and 95% Centred Intervals) of the SRM for the “Big Five” (75% training set and 25% testing set, 100 replications) - Season 2019/2020 . . . . .	92
4.5.	Sensitivity, Specificity, and MCC (Means and 95% Centred Intervals) of the SRM for the “Big Five” (75% training set and 25% testing set, 100 replications) - Season 2020/2021 . . . . .	92
B.1.	Variables classified by type of game actions and their corresponding description	102
B.2.	Comparison of the important and statistically significant variables ( $p$ -values<0.05) in the PLS-DA and RF, respectively, for the “Big Five” (75% training set and 25% testing set, 100 replications). The variables in bold indicate the top ten variables selected by the VIP and statistically significant for RF in most leagues and seasons - Season 2019/2020 . . . . .	109
B.3.	Comparison of the important and statistically significant variables ( $p$ -values<0.05) in the PLS-DA and RF, respectively, for the “Big Five” (75% training set and 25% testing set, 100 replications). The variables in bold indicate the top ten variables selected by the VIP and statistically significant for RF in most leagues and seasons - Season 2020/2021 . . . . .	111
5.1.	Reference players according to their popularity level and position . . . . .	118
5.2.	Variables grouped by class used to estimate players’ market value . . . . .	120
5.3.	Conversion factor (CF) and cumulative conversion factor (CCF) for the players according to their popularity level and position . . . . .	124
5.4.	Coefficients of the statistically significant variables ( $p$ -values<0.05) for the three models fitted by the MLR method. . . . .	125
5.5.	RMSE for all methods according to the three models (€). . . . .	126