TÉCNICO LISBOA

Assessing Team Success in the Big-5 European Football Leagues

Thesis to obtain the Master of Science Degree in

INDUSTRIAL ENGINEERING AND MANAGEMENT

José Paulo Direito Fonseca

November 2022

TÉCNICO LISBOA

Contents

TÉCNICO LISBOA

01.

Introduction

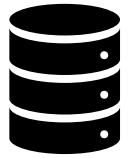
Context and Motivation

Competitive Advantage

Productivity

Profitability

European Football Market Size (2020/21) €27.6B



Team Performance

Team Management

Context and Motivation

TÉCNICO LISBOA

- ❑ Literature analysing game-related KPIs is considerable
- ❑ Insufficient emphasis placed on all aspects of the game
- ❑ Polarizing conclusions

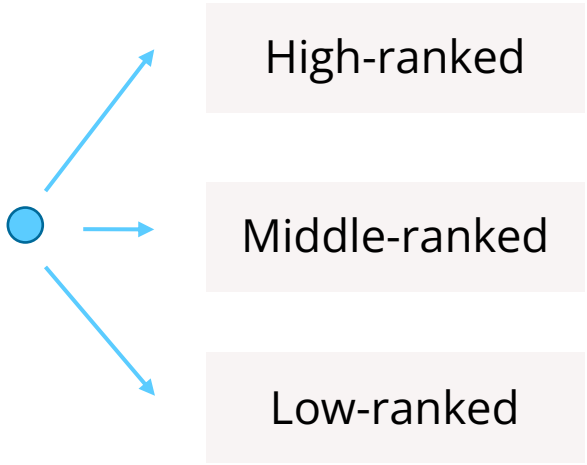
- ❑ Big-5 European Football Leagues Comparative Analysis are scarce

- ❑ Classification Analysis

Objectives

1. KPIs

1.1 Final ranking



1.2 League

PREMIER LEAGUE
LFP
BUNDESLIGA
SERIE A
LIGUE 1

2. Classification Models

2017/18
2018/19
2019/20
2020/21

→ ?

2021/22

TÉCNICO LISBOA

02.

Literature Review

Overview

TÉCNICO LISBOA

1.

The existing literature on game related KPIs linked to team success was studied and analysed

2.

The existing literature on game related KPIs with the purpose of characterizing the Big-5 European Football Leagues was studied and analysed

3.

The existing literature on the development and usage of Classification Models to predict Football related elements was studied and analysed

Key Performance Indicators

KPIs

Success

General Remarks

- ❑ Literature is considerable
- ❑ Focuses on 4 main areas:

Physical | Technical | Tactical | Mental

- ❑ Scope varied substantially:
  - ❑ Type of competition
  - ❑ Measure success
  - ❑ Number of leagues/games

Key Performance Indicators

TÉCNICO LISBOA

Key Takeaways

Successful

- Goals
- Shots; Shots on target
- High effectiveness in converting goals
- Increased action in the attacking 1/3 of the pitch
- Aerial advantage
- Ball Possession
- Lower number of passes

Unsuccessful

- Yellow cards
- Red Cards
- Dribbles
- Crosses

Big-5 European Football Leagues

TÉCNICO LISBOA

Key Takeaways

- Bundesliga players highest mean values in stature, body mass and BMI
- Italian clubs rigid tactical requirement for defensive organisation; best passing
- Spanish teams prioritise ball possession and players' individual technical ability to control the game; best quality players
- English teams ran a much longer total distance in high intensity running; quickest paced game and the hardest and most resilient approach to playing

Classification Models

TÉCNICO LISBOA

Key Takeaways

- Focus on predicting match outcomes rather than clubs final ranking
- Accuracy measures vary substantially
- Unable to generalize well to other data (leagues and season)

TÉCNICO LISBOA

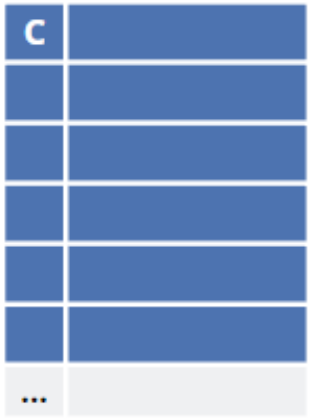
03.

Methods

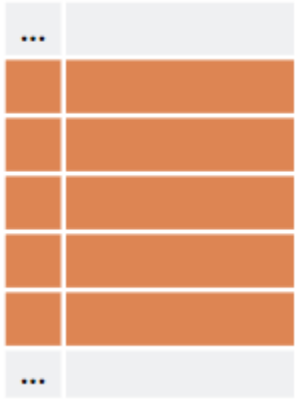
How to measure teams' performance?

TÉCNICO LISBOA

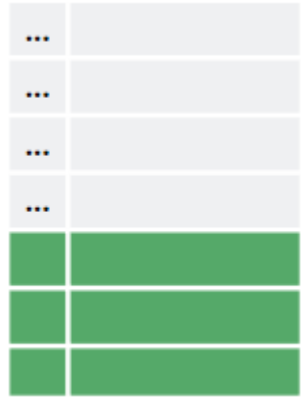
High-ranked



Middle-ranked



Low-ranked



UEFA CHAMPIONS LEAGUE

UEFA EUROPA LEAGUE

UEFA EUROPA CONFERENCE LEAGUE

Framework

Cross-Industry Standard Process for Data Mining

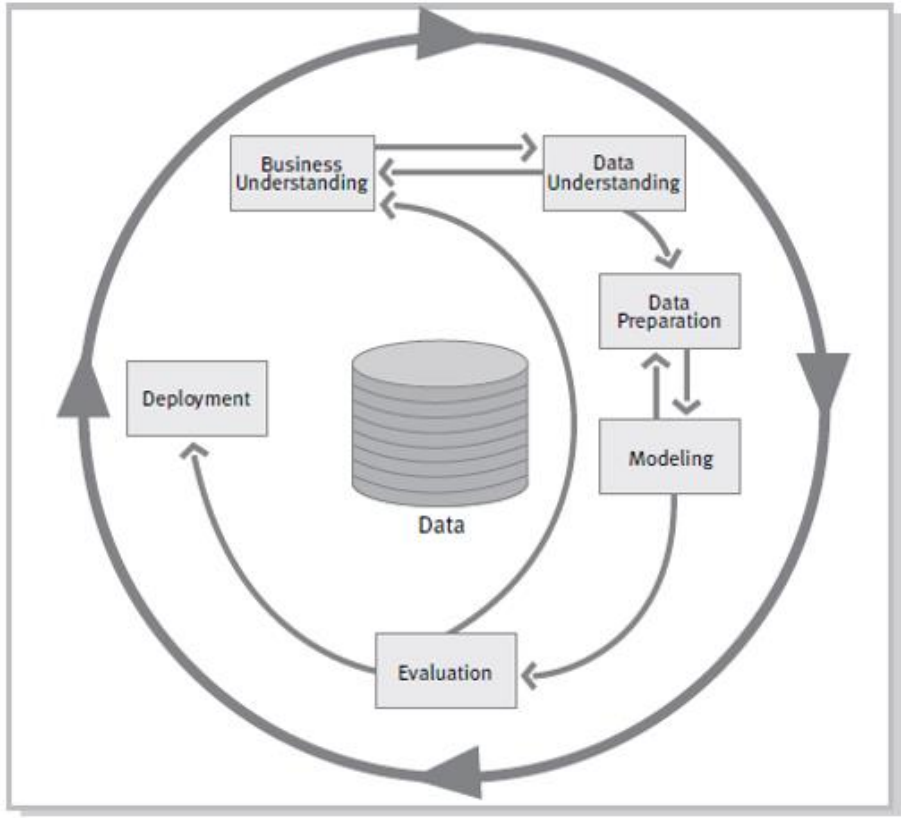


Figure 1 – Phases of the CRISP-DM reference model (from [4]).

Framework

Business Understanding

What are the project goals from a business perspective?

Data Mining Question

Data Understanding

Earliest contact with the data.

Included:

Framework

TÉCNICO LISBOA

Data Preparation

Operations related to obtaining the final dataset.

- ❑ Selecting relevant features
  - → Not relevant to the study
  - → Exploratory Analysis
- ❑ Cleaning
- ❑ Constructing
- ❑ Integrating
- ❑ Formatting

Framework

Modelling and Evaluation

Selecting and implementing the appropriate techniques while ensuring optimum calibration for their parameters.

1. KPIs

Principal Component Analysis

- ❑ Data Standardization

Cluster Analysis

- ❑ Hierarchical (Agglomerative)
- ❑ Partitioning (K-means)

- ❑ Accuracy, Precision, Recall, F1-score
- ❑ Silhouette Coefficient

Framework

TÉCNICO LISBOA

Modelling and Evaluation

2. Classification Models

- Gaussian Naïve Bayes
- Adaptive Boosting
- Logistic Regression
- Random Forest
- K-Nearest Neighbors
- Extreme Gradient Boosting

- 3-fold Cross-validation mean Accuracy
- ROC curves

04.

Data

Data Understanding

TÉCNICO LISBOA

FBREF

CSV files

Category	until 2016/17	from 2017/18 on
Standard Stats	X	X
Goalkeeping	X	X
Advanced Goalkeeping		X
Shooting	X	X
Passing		X
Pass Types		X
Goal and Shot Creation		X
Defensive Actions		X
Possession		X
Playing Time	X	X
Miscellaneous	X	X

PREMIER LEAGUE · LFP · BUNDESLIGA · SERIE A · LIGUE 1

Data Preparation

TÉCNICO LISBOA

≥ 0.9

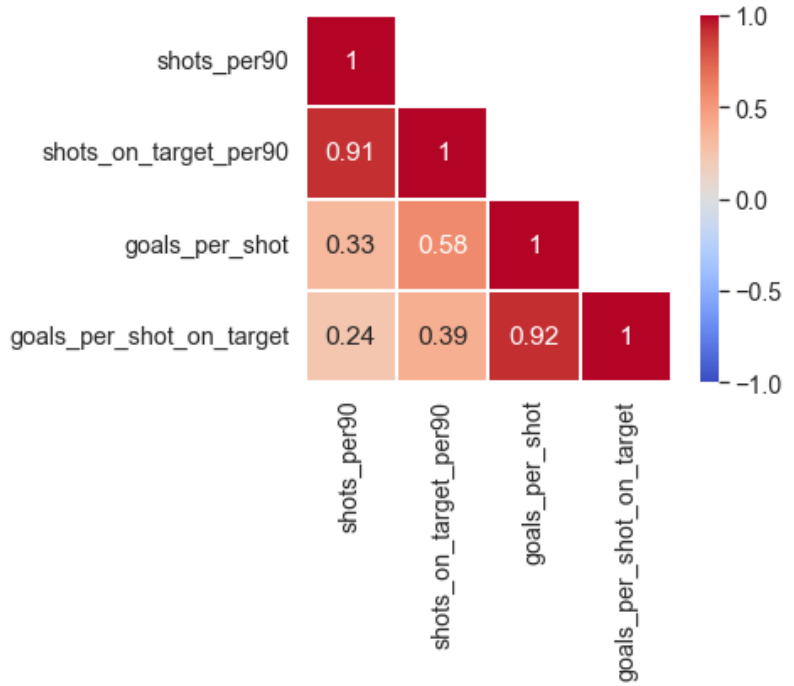


Figure 2 – Shooting variables correlation analysis

58 Missing Values

1 yard = 0.9144 meters

11 Categories ⟶ 6 Categories

98 Variables

490 Observations

Data Preparation

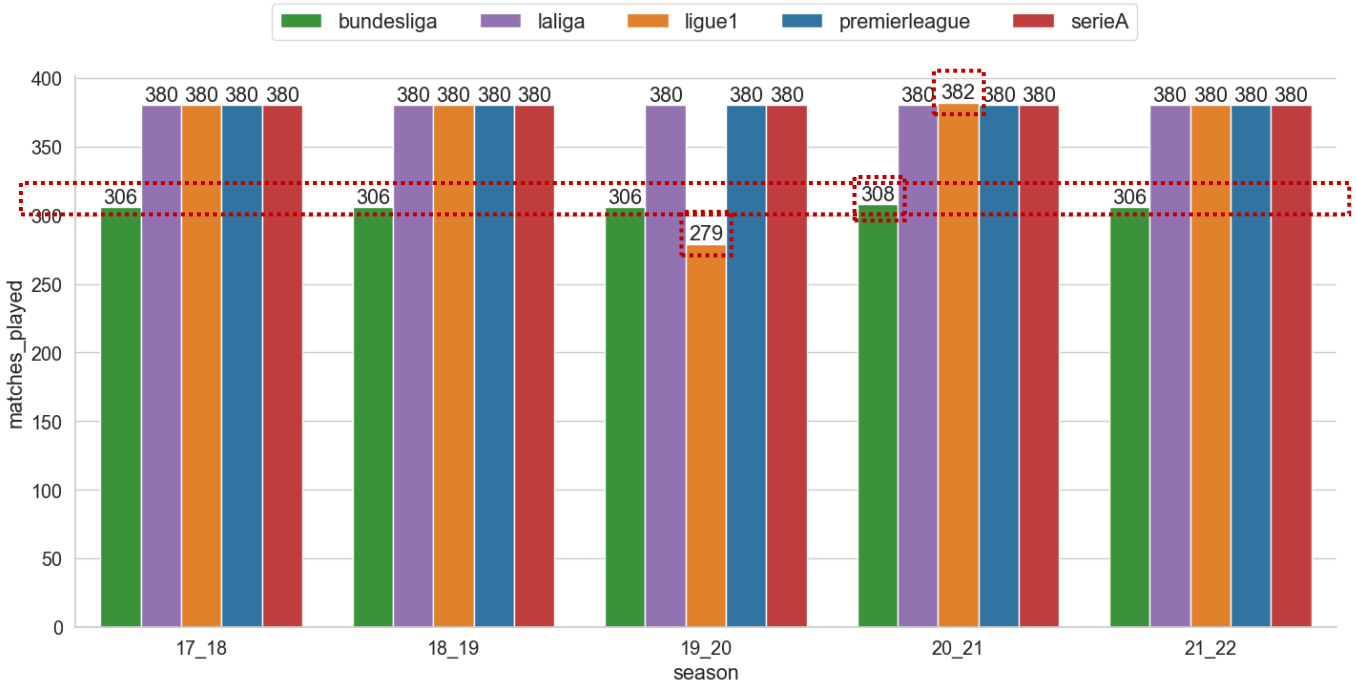


Figure 3 – Distribution of games in terms of leagues and seasons

18 Bundesliga Teams

$$UpdatedFeature = Feature * \frac{20}{18}$$

4 Additional matches 2020/21

Delete observations

279 Matches Ligue 1 - 2019/20

Don’t apply correction

TÉCNICO LISBOA

05.

Results

Principal Components Analysis

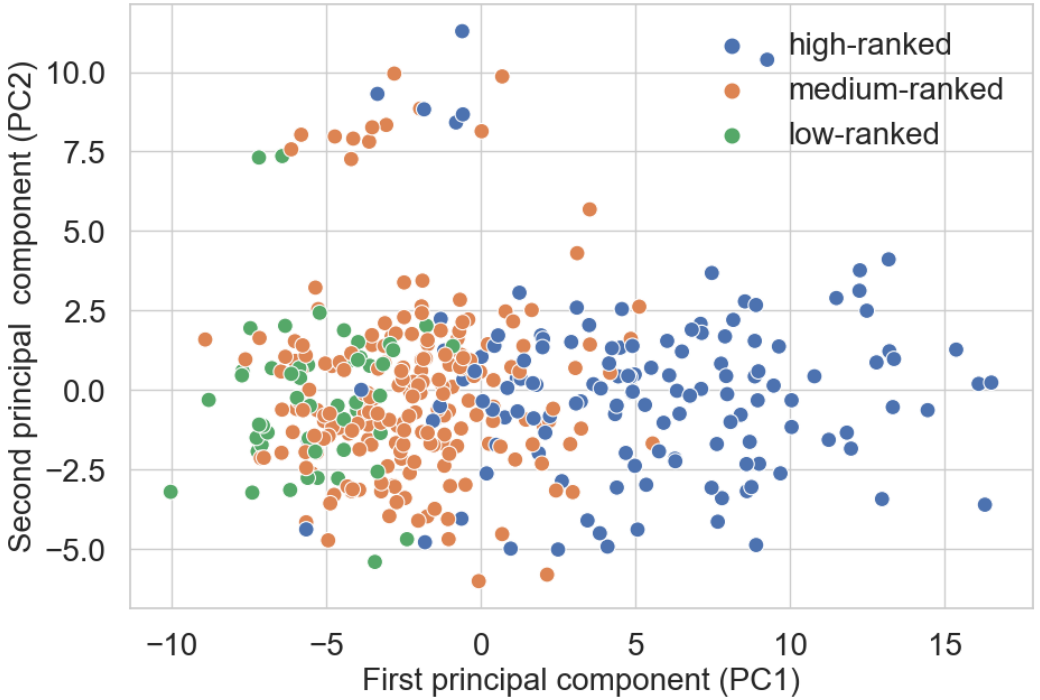


Figure 4 – Scores on PC1 vs PC2.

Category	Variable	Loadings	Pearson	p-value
Pss	num_times_ball_was_carried_towards_opponents_goal	0.175	0.934	$< 10^{-8}$
Pss	possession	0.172	0.917	$< 10^{-8}$
Pss	carries_in_attacking_1/3	0.167	0.886	$< 10^{-8}$
Pss	touches_in_offensive_1/3	0.166	0.883	$< 10^{-8}$
Pss	carries_into_goal_box	0.163	0.869	$< 10^{-8}$
P	live_ball_passes	0.173	0.923	$< 10^{-8}$
P	assists	0.165	0.876	$< 10^{-8}$
P	long_passes_completion%	0.161	0.854	$< 10^{-8}$
P	passes_total_distance_travelled_twoards_opponent	0.157	0.838	$< 10^{-8}$
P	pass_completion%	0.156	0.833	$< 10^{-8}$
GSC	shots_on_target_per90	0.171	0.909	$< 10^{-8}$
GSC	goals_scored_per90	0.170	0.907	$< 10^{-8}$
GSC	goal_creating_actions_per90	0.170	0.906	$< 10^{-8}$
GSC	passes_lead_to_goal	0.169	0.897	$< 10^{-8}$
GSC	shot_creating_actions_per90	0.167	0.889	$< 10^{-8}$
GSC	passes_lead_to_shot_attempt	0.166	0.884	$< 10^{-8}$
GS	wins	0.166	0.883	$< 10^{-8}$
GS	average_points_per_match	0.164	0.871	$< 10^{-8}$

Principal Components Analysis

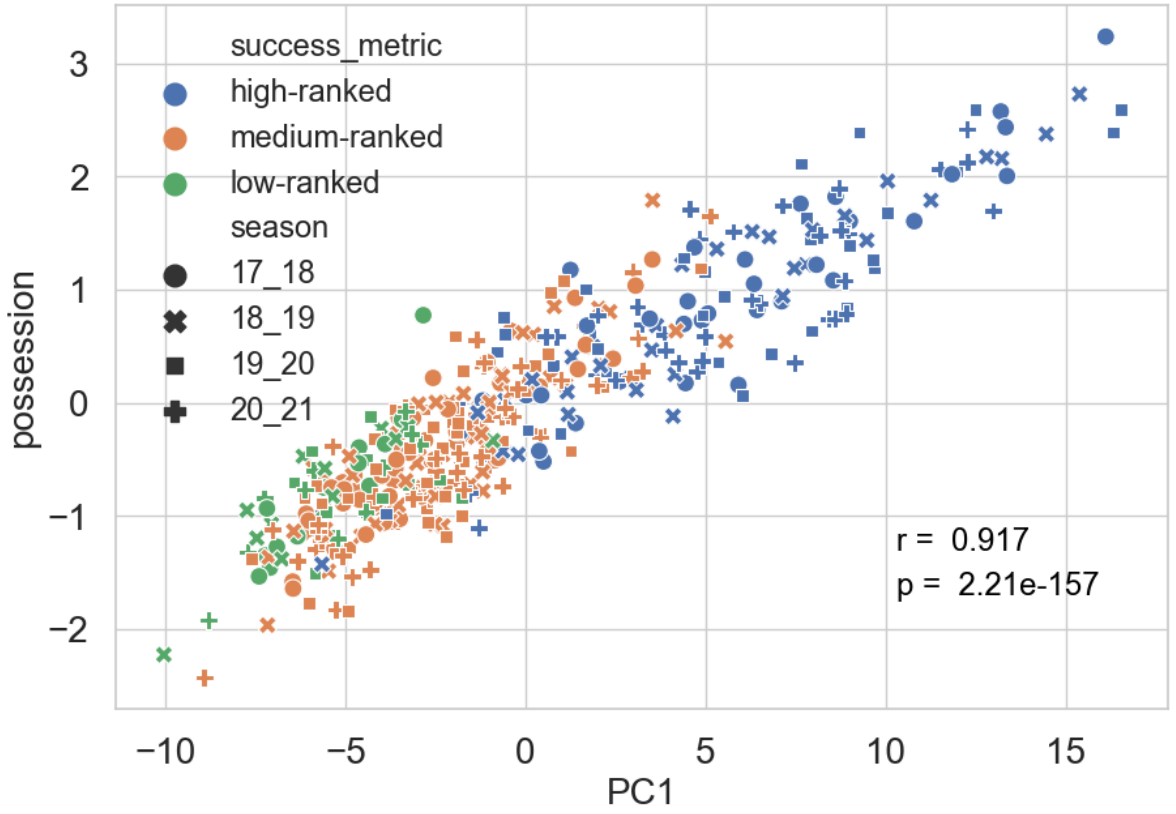


Figure 5 – PC1 scores vs possession

Principal Components Analysis

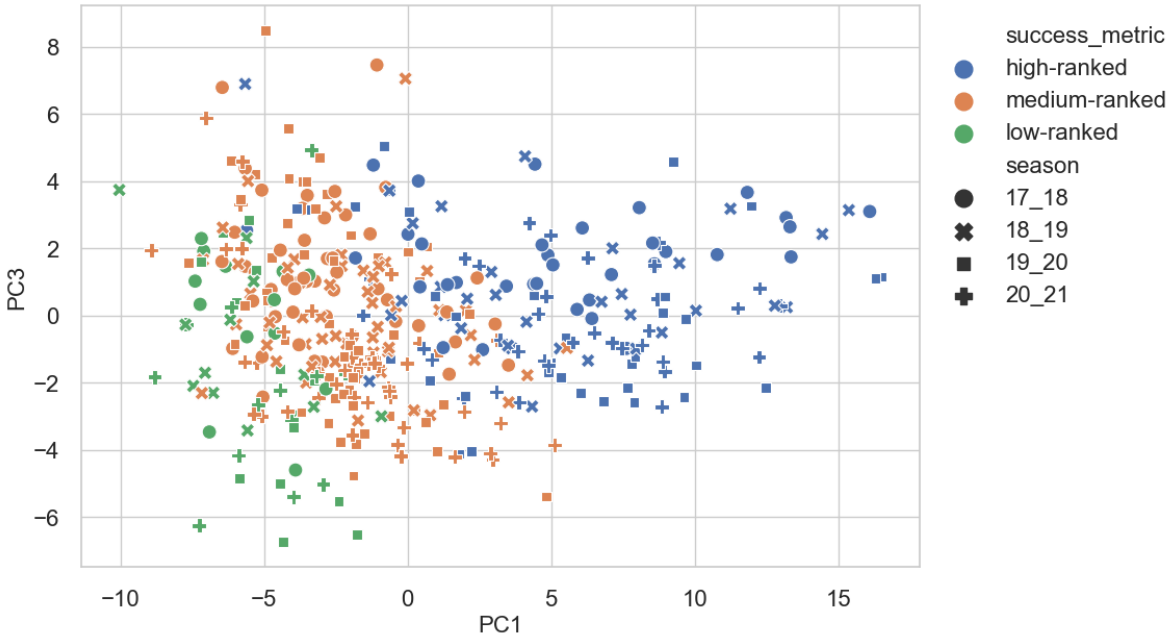
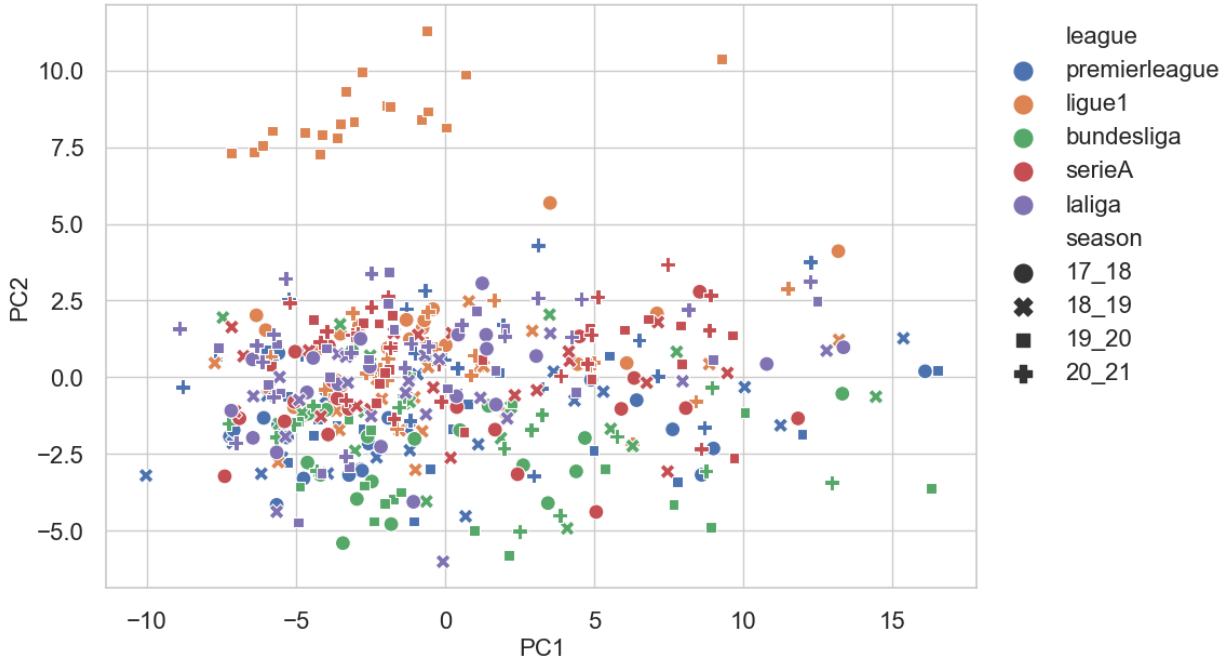


Figure 6 – Scores on PC1 vs PC3.

Category	Variable	Loadings	Pearson	p-value
G	clean_sheets%	0.196	0.482	$< 10^{-8}$
G	launched_goalkicks%	0.190	0.468	$< 10^{-8}$
G	passes_average_length	0.174	0.429	$< 10^{-8}$
G	defensive_actions_average_distance	0.165	0.406	$< 10^{-8}$
Pss	touches_in_defensive_1/3	-0.296	-0.729	$< 10^{-8}$
Pss	touches_in_defensive_penalty_area	-0.333	-0.819	$< 10^{-8}$
P	medium_passes_completion%	-0.169	-0.416	$< 10^{-8}$
G	passes_attempted_by_goalkeeper	-0.153	-0.375	$< 10^{-8}$
G	corner_kicks_goals_against	-0.169	-0.417	$< 10^{-8}$
G	penalty_kicks_allowed_against	-0.170	-0.418	$< 10^{-8}$
G	num_saves	-0.192	-0.472	$< 10^{-8}$
G	goals_against_per90	-0.202	-0.497	$< 10^{-8}$
G	throws_attempted	-0.230	-0.566	$< 10^{-8}$
DA	penalty_kicks_conceded	-0.161	-0.395	$< 10^{-8}$
DA	blocked_shots	-0.201	-0.493	$< 10^{-8}$

Principal Components Analysis



Category	Variable	Loadings	Pearson	p-value
Pss	ball_losses	-0.167	-0.465	$< 10^{-8}$
Pss	failed_attempts_to_regain_the_ball	-0.224	-0.626	$< 10^{-8}$
P	offsides	-0.150	-0.419	$< 10^{-8}$
P	low_passes	-0.159	-0.445	$< 10^{-8}$
P	high_passes	-0.239	-0.667	$< 10^{-8}$
P	blocked_passes_by_opponent	-0.244	-0.681	$< 10^{-8}$
GS	num_loose_balls_recovered	-0.315	-0.880	$< 10^{-8}$
DA	pressure_attacking_1/3	-0.189	-0.528	$< 10^{-8}$
DA	tackles_middle_1/3	-0.205	-0.572	$< 10^{-8}$
DA	winned_tackles	-0.213	-0.593	$< 10^{-8}$
DA	pressure_middle_1/3	-0.219	-0.611	$< 10^{-8}$
DA	blocked_passes	-0.255	-0.711	$< 10^{-8}$

Figure 6 – Scores on PC1 vs PC2, coloured according with the league.

Principal Components Analysis

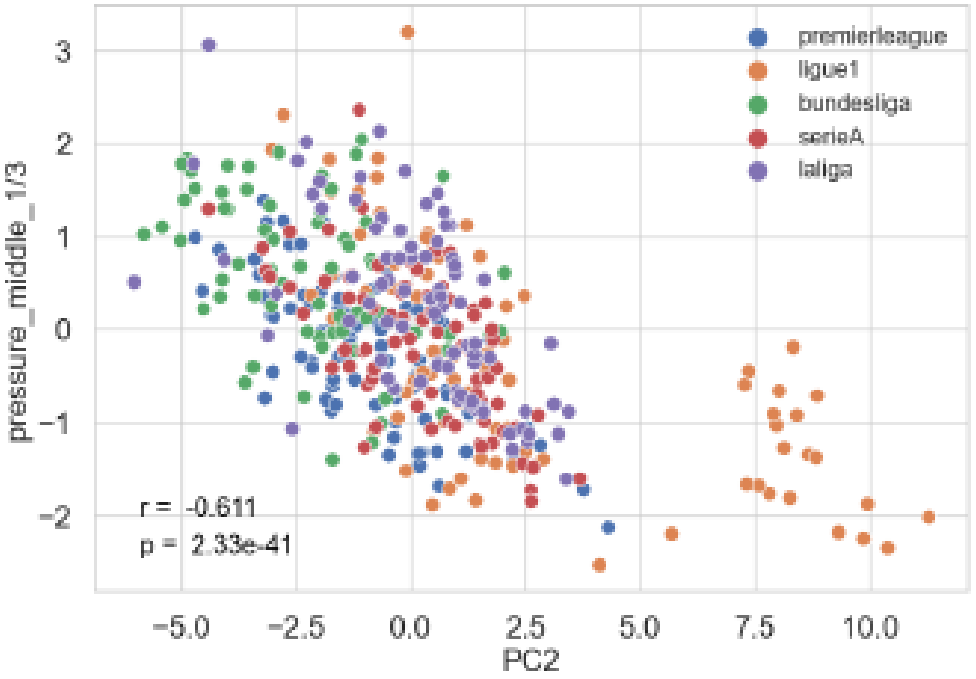


Figure 6 – PC2 scores vs pressure_middle_third

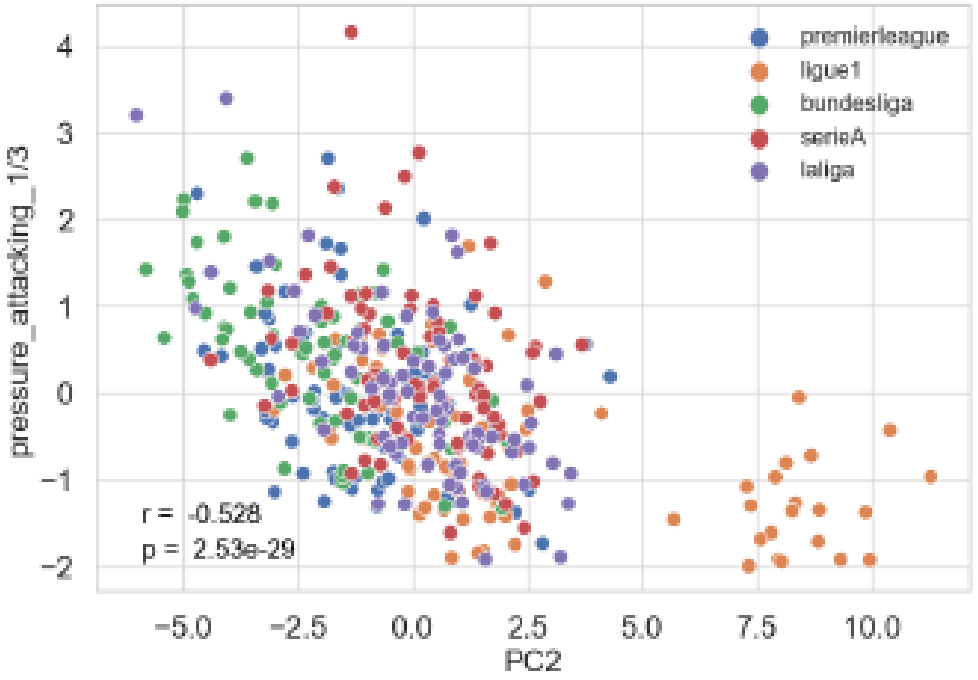


Figure 7 – PC2 scores vs pressure_attacking_third

Principal Components Analysis

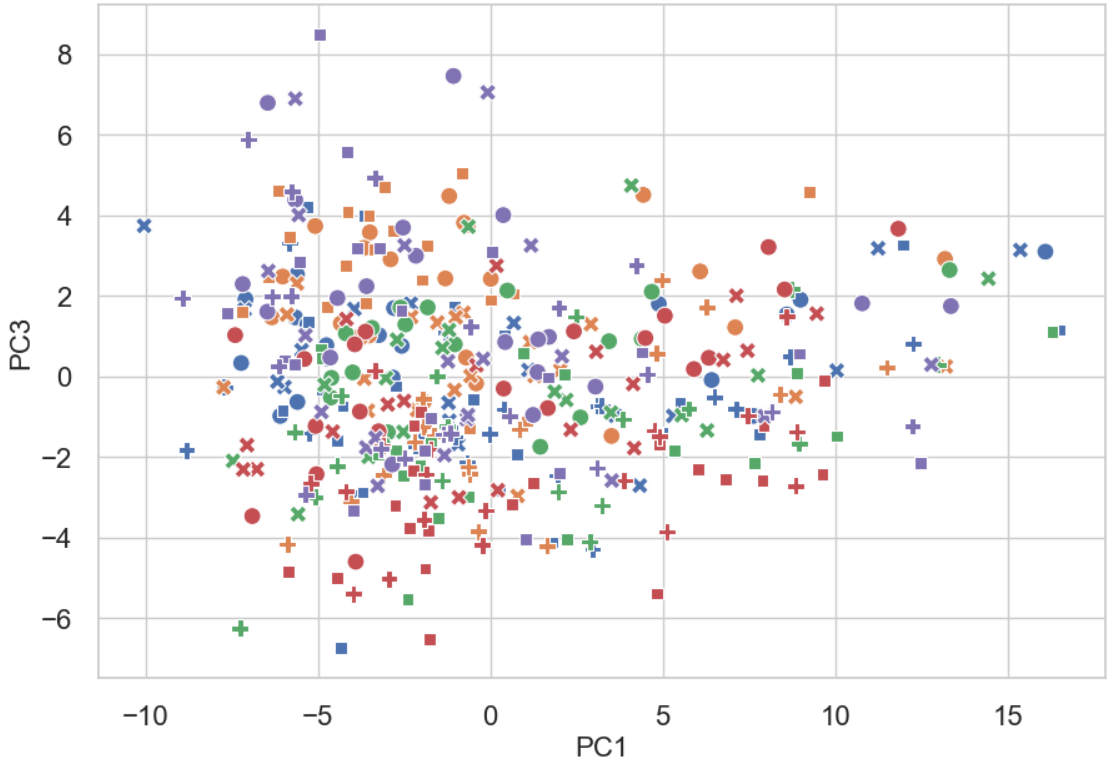


Figure 8 – Scores on PC1 vs PC3, coloured according with the league.

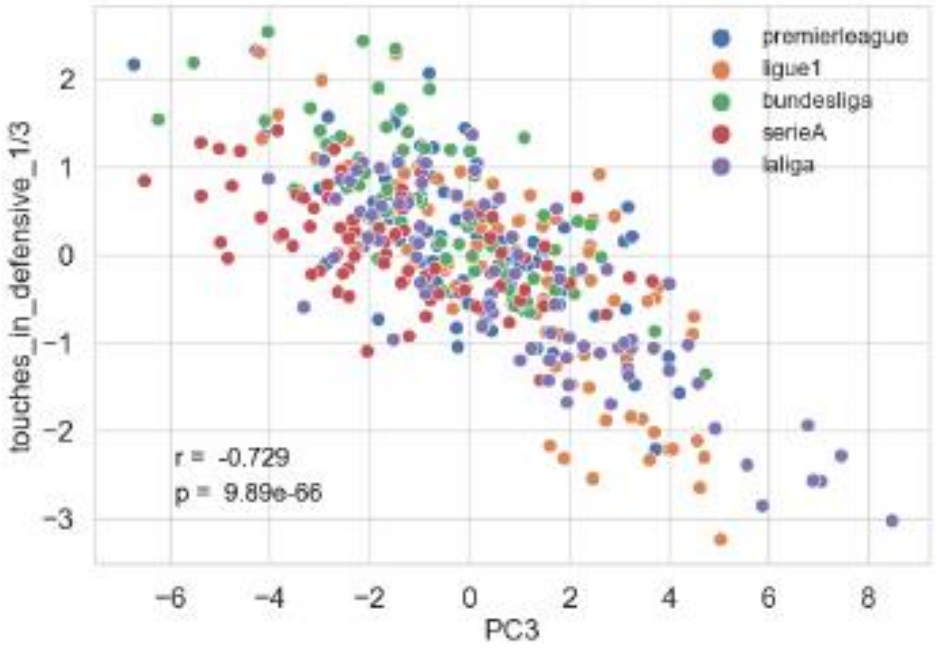


Figure 9 – PC3 scores vs touches_in_defensive_1/3

Principal Components Analysis

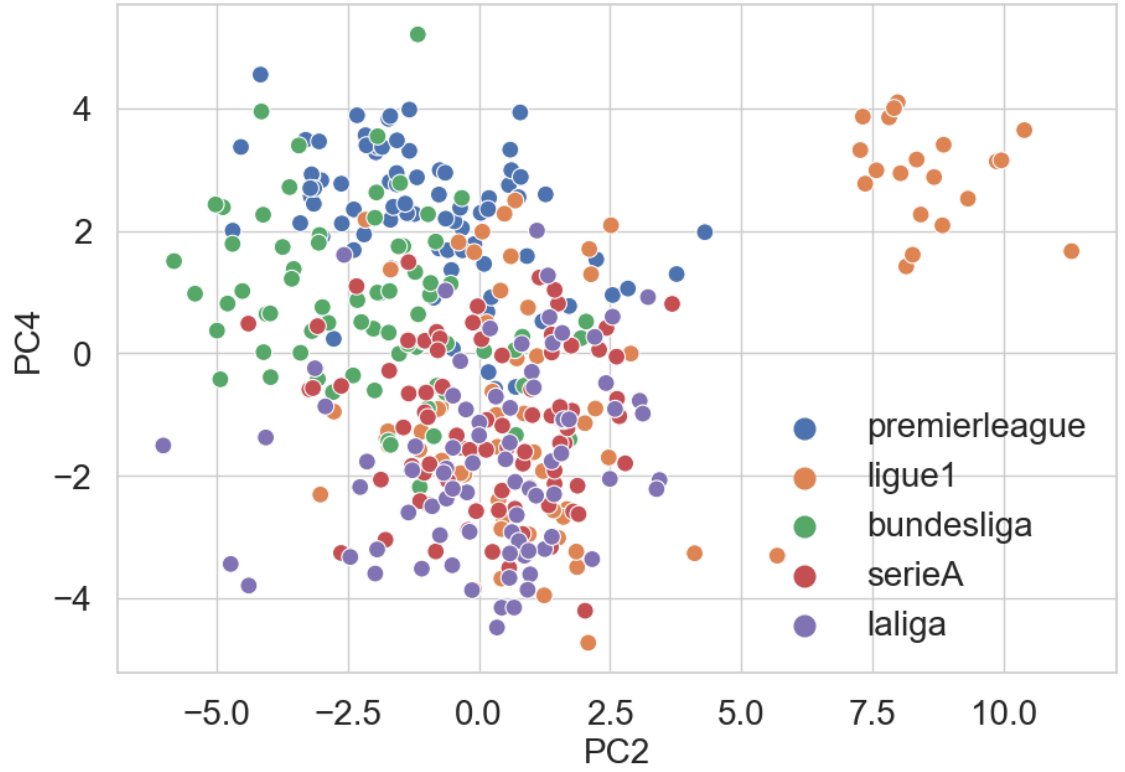


Figure 10 – Scores on PC2 vs PC4, coloured according with the league.

Category	Variable	Loadings	Pearson	p-value
P	inswing_corner_kicks	0.278	0.600	$< 10^{-8}$
P	outswing_corner_kicks	0.226	0.486	$< 10^{-8}$
P	straight_corner_kicks	0.191	0.411	$< 10^{-8}$
DA	pressure_to_opponent_completed%	0.217	0.467	$< 10^{-8}$
DA	tackles_completed%	0.169	0.364	$< 10^{-8}$
G	penalty_kicks_allowed_against	-0.160	-0.345	$< 10^{-8}$
GSC	free_kicks_shots	-0.194	-0.418	$< 10^{-8}$
GSC	fouls_lead_to_shoot_attempt	-0.244	-0.525	$< 10^{-8}$
GS	num_red_cards	-0.170	-0.366	$< 10^{-8}$
GS	fouls_commited	-0.253	-0.545	$< 10^{-8}$
GS	num_yellow_cards	-0.283	-0.610	$< 10^{-8}$
GS	fouls_drawn	-0.300	-0.646	$< 10^{-8}$

Principal Components Analysis

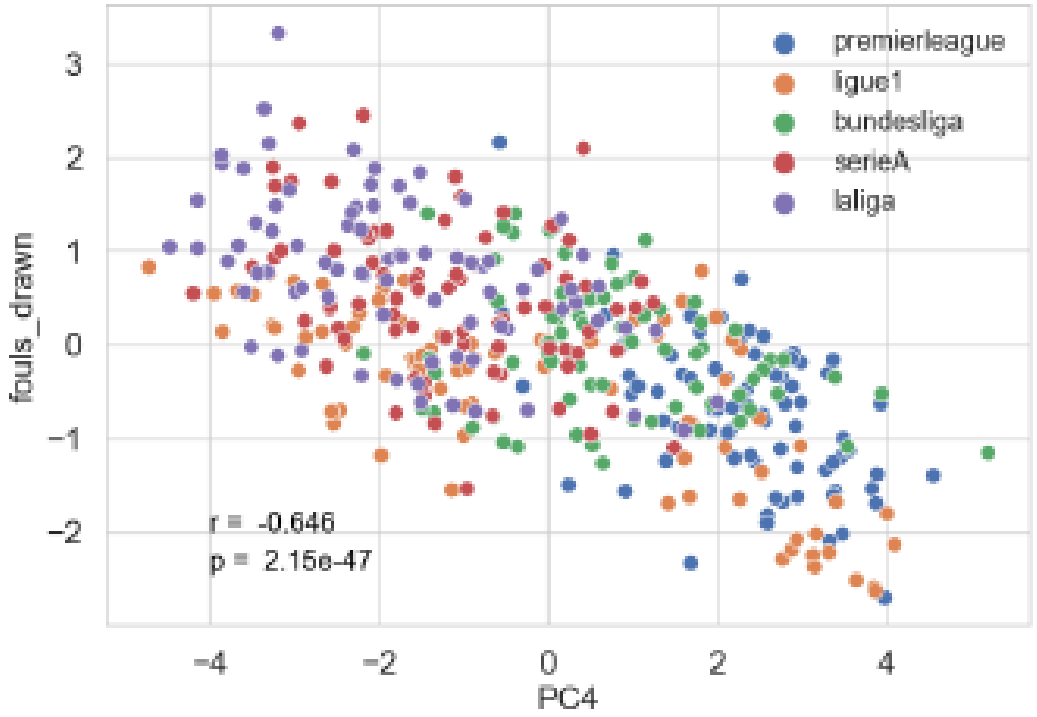


Figure 11 – PC2 scores vs fauls_drawn

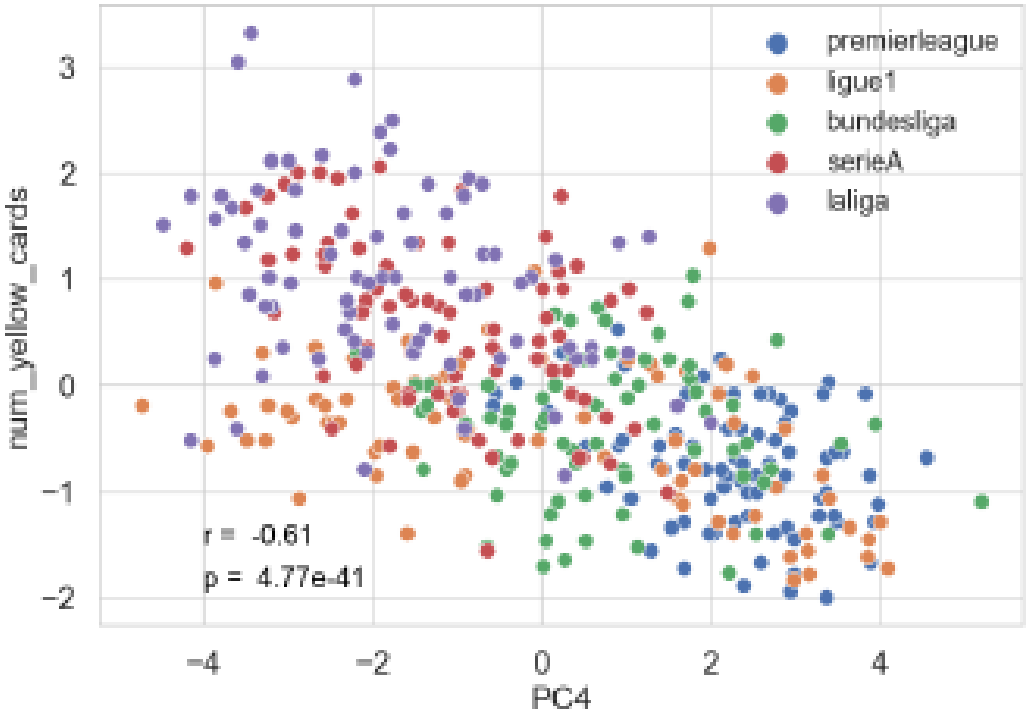


Figure 12 – PC2 scores vs num_yellow_cards

Cluster Analysis

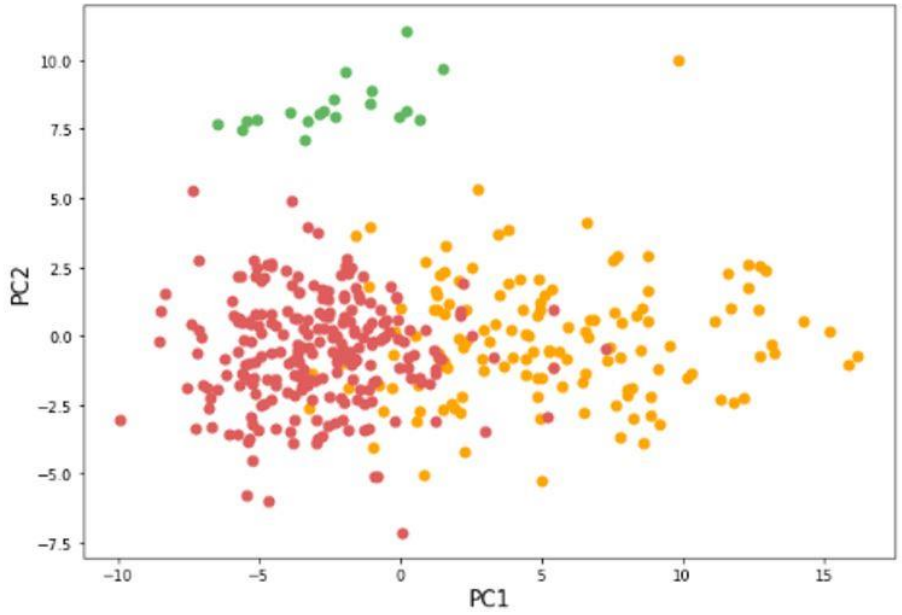


Figure 13 - PC1 vs PC2 scores coloured according to the clusters obtained from hierarchical Clustering, all input variables, and observations with ward linkage and Euclidian distance.

Table 7.9: K-Means Clustering Results

	Original Data			PCA Data		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
$K = 2$						
high-ranked	0.75	0.65	0.70	0.90	0.68	0.77
middle + low-ranked	0.83	0.88	0.85	0.85	0.96	0.90
$K = 3$						
high-ranked	0.90	0.51	0.65	0.98	0.60	0.74
middle-ranked	0.59	0.58	0.59	0.62	0.47	0.53
low-ranked	0.24	0.52	0.33	0.24	0.71	0.36
Accuracy						
$K = 2$	0.80			0.86		
$K = 3$	0.55			0.55		

Classification Models

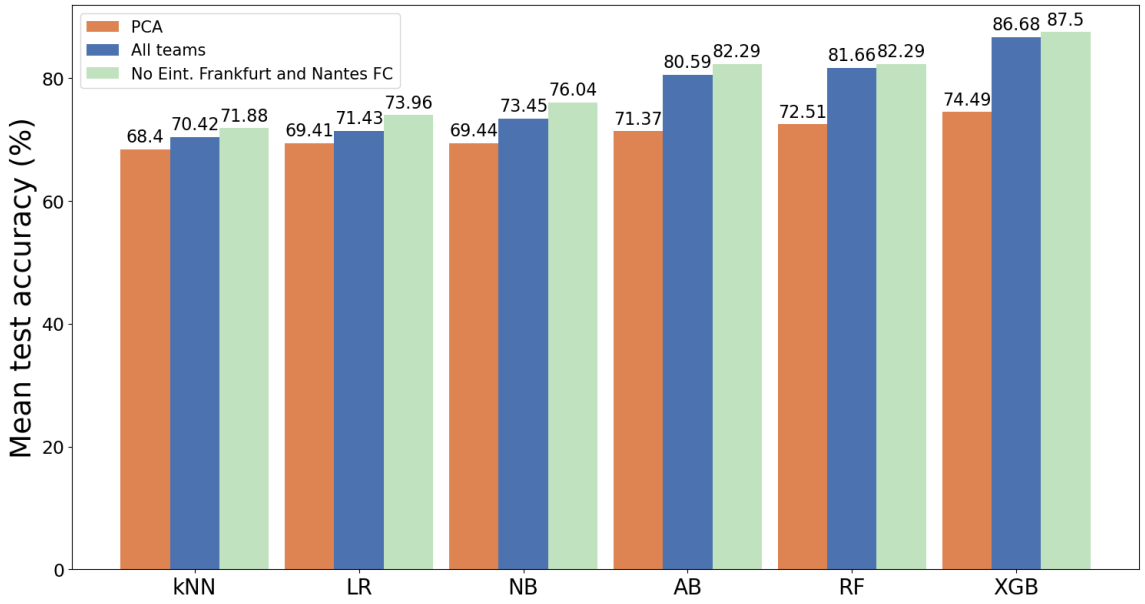


Figure 14 - Mean test accuracy for the applied classifiers before (blue) and after (green) removing the Eintracht Frankfurt e. V. and Nantes and when using data projected in the first 22 PCs (orange).

Table 7.15: Extreme Gradient Boosting results (overall accuracy of 0.91).

(a) Confusion matrix

	Predicted		
	high-ranked	middle-ranked	low-ranked
Actual high-ranked	33	2	0
Actual middle-ranked	4	43	1
Actual low-ranked	0	2	13

(b) Measures of performance

	Precision	Recall	F1-score
high-ranked	0.89	0.94	0.92
middle-ranked	0.91	0.90	0.91
low-ranked	0.93	0.87	0.90

TÉCNICO LISBOA

06.

Conclusions

Final Remarks

TÉCNICO LISBOA

Successful

- Goals
- Shots; Shots on target
- Increased action in the attacking 1/3 of the pitch
- Ball Possession
- Passing accuracy
- Assists
- Passes that moved the ball towards the opponent's goal

Unsuccessful

- Overall salient patterns in all defensive statistics
- Increased action in the defending 1/3 of the pitch
- Goals conceded
- Number of saves

The COVID-19 pandemic effect was successfully captured by the Analysis

The Cluster Analysis showed increased performance when the middle and low-ranked teams were treated as part of the same group

Final Remarks

Premier League and Bundesliga strong **overall presence in all areas of the pitch**

Serie A clubs prominent use of the **defensive patterns**

La Liga and Serie A **highest number of infringements** and yellow and red cards received

Premier League lowest reported number of fouls drawn and committed

Ensemble techniques achieved the best results

XGBoost accuracy **91%**

Future Avenues of Research

TÉCNICO LISBOA

Impact of French Ligue 1 2019/20 observations in the analysis

Applying the classifiers to only a partially complete data set (for example mid-season)

TÉCNICO LISBOA

Thank You.