Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra

iscac
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

FootbAI – Football powered by Artificial Intelligence

Alberto Jorge Ventura Cortez

ISCAC | 2021

Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Alberto Jorge Ventura Cortez

FootbAI – Football powered by Artificial Intelligence

Coimbra, September of 2021

Instituto Superior
de Contabilidade
e Administração
Politécnico de Coimbra
COIMBRA BUSINESS SCHOOL
ISCAC.pt

Alberto Jorge Ventura Cortez

FootbAI – Football powered by Artificial Intelligence

Project work submitted to the Institute of Accounting and Administration of Coimbra in partial fulfilment of the requirements for the Master's Degree in Management Information Systems, held under the supervision of Professor António Trigo Ph.D. and the guidance of the Professor Nuno Loureiro Ph.D.

Coimbra, September of 2021

STATMENT OF RESPONSABILITY

I declare that I am the author of this project, which is an original and unpublished work, that has never been submitted to another Higher Education Institution for obtaining an academic degree or other qualification. I also attest that all citations are properly identified and that I am aware that plagiarism is a serious lack of ethics, which may result in the cancellation of this project.

"There's a storm inside of us. A burning. A river. A drive. An unrelenting desire to push yourself harder and further than anyone could think possible." – Marcus Luttrell

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

AKNOWLEDGMENTS

The completion of this master's Project had important support, without which, undoubtedly, it would not have been materialized, to them, my personal thanks.

To Professor António Trigo, who accompanied and guided me throughout this process, questioning the work done so that I could evolve along the way. For all the advice, which allowed me to be better at what I do and made me reflect on all the ideas so that I can find a way to make them operational. I would also like to thank for the availability and kindness that you always had towards me throughout this whole process.

To Professor Nuno Loureiro who was present in a previous work and help once again in this project. Thanks for being present when was important to clarify some concepts.

To my parents, without them I definitely would not be here. I am clearly the fruit of two people who always motivated me to be better and to seek what I aspire.

To Raquel Reis, who supported me and supports me in all moments, even in those that are difficult even for me. She is one of the people who makes me want to be better, both as a person and as a professional.

To my brother, André Cortez, for being an example as a person and as a researcher. He is undoubtedly the person who, in a more direct way have influenced me to be better, to chase different paths and aim for the continues improve of myself.

To my colleagues in the master's degree and all those who directly or indirectly were involved in this work, my thanks!

ABSTRACT

Modern football competition has the characteristics of fierce confrontation, long duration, intensity of play and large amounts of exercise, with high technical and tactical requirements. It is therefore a complex sport and one of the most important tasks is the selection of the most suitable players for the matches, which involves many factors, such as the players' physiological variables.

Football, as well as other sports and social areas, has suffered the influence of technologies and information systems that have contributed to the improvement of the sport, with the main emphasis on information systems to support the monitoring and analysis of players' performance, such as, for example, the use of information systems to record the players' physiological variables using GPS systems, and measurement of heart rates, among others.

Given the opportunity to analyse the data of a football team from the 2nd Regional Division of Football Association of Santarém, the idea arose to try to understand which are the most important physiological variables in the players' performance and, consequently, to improve the teams' performance, through a greater perception of the interconnection between physiological variables and sporting results.

In terms of results, it was possible to develop a model that grouped the physiological variables that had more influence on the victory, managing to have results of 79% of accuracy in predicting the victory with these variables. It was also possible to perform the same analysis performed on the games data, also on the training data, and the results pointed to different variables. In terms of the analysis of the athletes per position, despite the results being short, the application of a selection algorithm was able to rank the physiological variables per position, which are in line with the variables advocated by the scientific community in studies of these themes.

Keywords: Football; Management; Multi-platform integration; Machine Learning; Classification; Regression.

RESUMO

A competição futebolística moderna tem as características de confronto feroz, longa duração, intensidade de jogo e grandes quantidades de exercício, com elevados requisitos técnicos e táticos. É, portanto, um desporto complexo, sendo uma das tarefas mais importantes a seleção dos jogadores mais adequados para os jogos, o que envolve muitos fatores, como, por exemplo, as variáveis fisiológicas dos jogadores.

O futebol, bem como outras áreas desportivas e sociais, tem sofrido a influência de tecnologias e sistemas de informação que têm contribuído para a melhoria do desporto, com ênfase principal nos sistemas de informação para apoiar a monitorização e análise do desempenho dos jogadores, tais como, por exemplo, a utilização de sistemas de informação para registar as variáveis fisiológicas dos jogadores utilizando sistemas GPS, e a medição dos batimentos cardíacos, entre outros.

Dada a oportunidade de analisar os dados de uma equipa de futebol da 2ª Divisão Regional da Associação de Futebol de Santarém, surgiu a ideia de tentar compreender quais são as variáveis fisiológicas mais importantes no desempenho dos jogadores e, consequentemente, melhorar o desempenho das equipas, através de uma maior perceção da interligação entre as variáveis fisiológicas e os resultados desportivos.

Em termos de resultados, foi possível desenvolver um modelo que agrupou as variáveis fisiológicas que tiveram mais influência na vitória, conseguindo ter resultados de 79% de precisão na previsão da vitória com estas variáveis. Também foi possível realizar a mesma análise realizada sobre os dados dos jogos, também sobre os dados de treino, e os resultados apontaram para diferentes variáveis. Em termos da análise dos atletas por posição, apesar dos resultados serem curtos, a aplicação de um algoritmo de seleção foi capaz de classificar as variáveis fisiológicas por posição, que estão em linha com as variáveis defendidas pela comunidade científica em estudos sobre estes temas.

Palavras-chave: Futebol; Gestão; Integração multiplataformas; Machine Learning; Classificação; Regressão.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

INDEX

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

LIST OF FIGURES

LIST OF TABLES

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

LIST OF ACRONYMS

ADT: Alternating Decision Tree

AFS: Football Association of Santarém

AI: Artificial Intelligence

APSO-K: Accelerated Particle Swarm Optimization

CD: Central Defender

CI: Confidence Interval

CRISP-DM: Cross Industry Standard Process for Data Mining

CV: Cross-Validation

ETRFC: Extra Tree Random Forest Classifier

F: Forward

FB: Full Back

GPS: Global Positioning System

HR: Heart Rate

ICCSA 2021: The 21st International Conference on Computational Science and Applications

ID: Identity of the Player

J48con: J48 consolidated

KNN: K-Nearest Neighbours

Kohonen SOM: Kohonen Self-Organized Maps

LUp: Line-Up

MC: Midfielder

MD: Match Day

ML: Machine Learning

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

NB: Naïve Bayes

OM: Offensive Midfielder

REPTree: Reduced Error Pruning Tree

RF: Random Forest

RFE: Recursive Feature Elimination

RIPPER: Repeated Incremental Pruning Produce Error Reduction

RPE: Rate of Perceived Exertion

RPS: Ranked Probability Score

STCK: Spatio-Temporal Convolution Kernels

STCKdist: Spatio-Temporal Convolution Kernels Distance

StD: Standard Deviation

SVM: Support Vector Machines

U.A.: Absolute

UEFA: Union of European Football Associations

W: Winger

WRRatio: Work Ratio

XgB: Extreme Gradient Boosting

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1. INTRODUCTION

The use of GPS (Global Positioning Systems) devices has become common in professional football to follow the players' performance. This kind of systems generate huge amounts of data making difficult the analysis.

In order to solve this problem, systems based on Artificial Intelligence (AI) have been developed. These systems enable the treatment and analysis of large amounts of data, extracting interesting conclusions from the data (Kusmakar et al., 2020). AI has also made it possible to build predictive systems with unprecedented accuracy (Rossi, Perri, et al., 2016), with Rossi et al., (2016) defending that this amount of data provides a new opportunity for collaborations between data and sport sciences to maximize the AI potential for predicting football match performance.

In this sense, it becomes imperative to bring together data and sports scientists who can understand the various possibilities of analysing the data obtained from these GPS systems and use AI techniques to extract new insightful information that can lead to a better understanding and help coaching staff to make a more informed decision.

1.1 Background

The recent use of GPS systems in football brings the possibility to collect a large set of physiological variables from football players for further analysis. The analysis of this data, due to its large volume, requires sophisticated analysis techniques and information systems, both at the statistical level and at the level of AI, in particular machine learning (ML), to retrieve truly innovative information that will help in the understanding of the sports dynamics.

Regarding this, it's important to understand that AI can enhance the human intelligence, as García-Aliaga et al., (2020) affirms, human activities will have an advantage by using the support of computational capabilities. With these capabilities and "the development of society, artificial intelligence technology is gradually put into various fields" (Yang, 2020). So, it is imperative to understand what type of capabilities could help in this development, so, in this project, it was aimed to explore the capabilities of the AI, more specific, ML.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

It is important to understand that ML is not a new topic in other areas of research (Brandão et al., 2021; Pimenta et al., 2009; Seiça et al., 2019), but in sports there is a lacking in implementation/application. As Herold et al., (2019) defend in their study, computer scientists and sports scientists should join to obtain more accurate information with respect to individual and collective performance that may influence the outcome of football matches.

Most of the investigations using ML done in sports, more specifically in football, are related to the understanding the applied models or as a prediction method and few studies focused on the use of ML to improve tactical knowledge and performance (Herold et al., 2019).

It's well known that many people in academia and industry have addressed/tackled the problem of football match prediction, owing to both its interesting nature and its economic importance. Previous research regarding this problem can be divided into two major categories (Baboota & Kaur, 2019): result-based studies and goal-based studies.

Without doubt, we can say that AI will make prediction of outcomes in the sports industry reliable and accurate to a certain extent. But it is important to understand that if human element is involved in sports, there will always be unpredictability and uncertainty that makes it fascinating and surprising for its viewers (Keshav, 2020). Still, it's important to understand that unlike machines, humans can think and feel, which often guides their decision making and this could lead to different decisions than what a computer would choose (Keshav, 2020).

1.2 Motivation

The research work here presented has the objective to provide the knowledge, both in theoretical and practical, for the use of AI in sport, more specifically in football, linking the author's background, with the knowledge in information systems and machine learning provided throughout the master's course.

A multi-disciplinary approach, including the collaboration of big data technologies with football research may facilitate a comprehensive understanding of the tactical performance (Herold et al., 2019).

As Keshav, (2020) affirms, further studies may explore applications of AI on to predict the best possible line-up based on the physical as well as tactical attributes of the players and the fixture difficulty of a match. That opened the possibility of analysing the available data and apply ML techniques to provide new valuable knowledge.

With the ambition of doing something different in this project, and considering the works referred before, the authors tried to gather some information regarding football analysis and tried to contribute with different approaches to the analysis of this data. That originated important research questions, that will be addressed next.

1.3 Research questions

The importance of this project was to understand the possibilities involving ML techniques and how could the author improve his skills in data manipulation in Python in order to improve his knowledge and achieve better results when analysing different features in a football match.

In this line of thought, this research aimed at understanding how ML could help improve the decision-making ability from coaching staffs in a football team, making use of the variables collected by a GPS system.

The data collected was divided into two different datasets, one with the data regarding the matches, and other regarding the training sessions, this opens the possibility of analysing different aspects regarding these two moments.

Considering the two existing datasets and the goals of our work five research questions were proposed:

1. Players' physiological variables (match data) that most contribute to victory;
2. Players' physiological variables (match data) that most contribute to victory by position in the field;
3. Players' physiological variables (training sessions dataset) that most contribute to victory
4. Players' physiological variables (training sessions dataset) that most contribute to victory by training session
5. Predicting the starting line-up and chose the better prepare players;

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1.4 Research methodology

The methodology used for conducting the studies presented in this work was the Cross Industry Standard Process for Data Mining (CRISP-DM) commonly used in studies of this nature.

Laureano et al., (2014) used this methodology to better model the prediction of waiting time for admissions. This allowed them to identify the attributes of the clinic related to the length of stay.

Another study done in the medicine with this methodology was done by Morais et al., (2017). The study aimed to better understand which characteristics of a newborn baby brings the need for assistance to breathe at birth. This was done by analysing the characteristics of the mother and pregnancy to prevent neonatal mortality.

The CRISP-DM methodology defines six phases for its application Laureano et al., (2014): business understanding, data understanding, data preparation, modelling, evaluation, and deployment. The phases are shown in Figure 1.1.

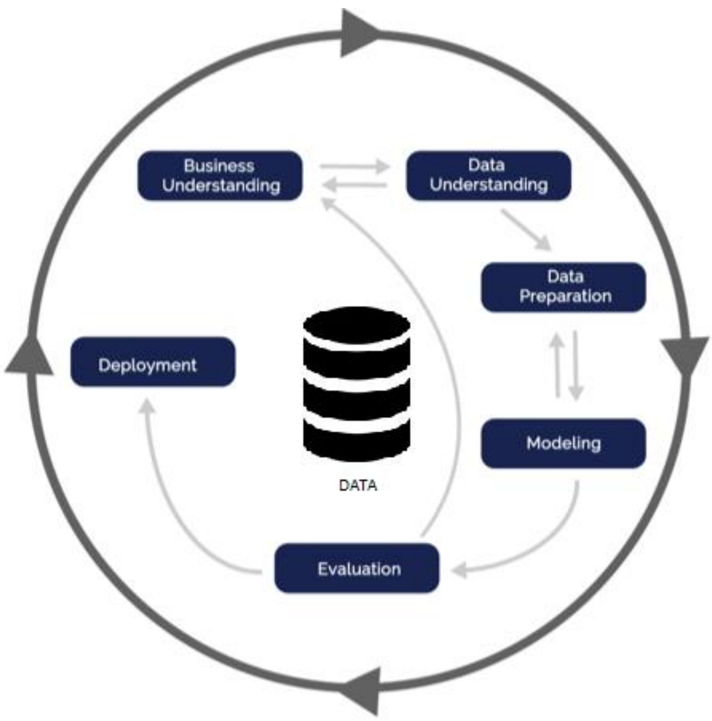


Figure 1.1. CRISP-DM methodology (adapted from Laureano et al., (2014)*)*

It is important to understand these six phases before advancing:

- Business Understanding – allows to understand the objectives and requirements of the study and should be used to plan the steps of the project.
- Data Understanding – it is the phase in the study that permits to identify, collect, and analyse the data, to understand the data and the quality of the data.
- Data Preparation – at this phase, the data are selected, cleaned and formatted so that they are ready to be used, and it is at this phase that the final dataset to be used is created.
- Modelling – it is the phase where it's determined the different algorithms and models to use, which can be improved through the different interaction to obtain better results.
- Evaluation – at this phase the results are assessed, and the processes review to find out if the model answers to the business needs and verify any overlooked, so that actions can be taken to improve the model.
- Deployment – it is when the model is put into production and customer can access the results. This work will not go through this phase.

1.5 Report outline

This thesis is comprised of six chapters. In the chapter one, the focus is the background, motivation, and goals. This chapter also addresses the methodology used (CRISP-DM).

In chapter two, the literature review provides the current state of the art on ML techniques and the studies done with the methodology used, with a focus in studies involving ML applied to team sports, more specifically in football.

Chapter three presents the characterization of the business where the project was conducted, that was a football team, and how the data was collected, the physiological variables of the players retrieved in the GPS systems. This chapter also focus on the data preparation and applied procedures.

The chapter four concerns the modelling and evaluation of the different algorithms. This part of the project it's divided into five different studies to maximize the depth of each study and understand how the different studies were influenced from the previous ones.

Fifth chapter is the discussion of results of the conducted studies in the light of the existing literature. The discussion was also divided by study.

Last chapter, chapter six, presents the final considerations, with a summary of the work developed, main contributions and limitations experienced during its execution, and some proposals for future work.

2. LITERATURE REVIEW

Football is considered to be within a set of sports that belongs to the so-called collective sports games, and is seen by several authors, as the most popular sport in the world (Behravan & Razavi, 2020; Baboota & Kaur, 2019). As Matesanz et al., (2018) say, professional football is regarded as the most popular sport in the world, famous for both its players and clubs.

Football has become one of the most analysed sports in terms of match analysis, and more recently, in terms of training analysis (Rossi, Savino, et al., 2016) allowing the gathering of more knowledge of the teams involved. As Marcelino et al. (2020) defend, there is a need for a holistic approach to sport performance, that could open up new ways of analysing and assessing team and individual performance.

2.1 Football

Sarmento et al., (2010) in their article, state that the study of the game by observing the behaviour of teams and players is not a recent phenomenon, and that "football has evolved over the years along with the development of computer systems that have allowed greater knowledge of this phenomenon" (Sarmento et al., 2014). In soccer "the high number of players, the complexity of tactical behaviours and the speed at which actions occur prevent the observational recording of behavioural interactions" (Sarmento et al., 2010). With the help of advanced statistical processes, some researchers have tried to find an association between cause and effect in different interaction contexts (Sarmento et al., 2014).

For da Costa et al., (2011), it seems important to create evaluation tools that allow recording reliable results on the tactical behaviour and progression of practitioners to improve the training and teaching/learning process.

According to Sarmento et al., (2010), the behaviours in collective games are more synchronized than the human eye can detect, so systematic observation is necessary, and the use of systematic observation can significantly increase the ability to understand and analyse the context of soccer players' behaviours, since the interaction between

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

teammates is inevitable and fundamental, or as defended by Clemente et al. (2015), these interactions are how a collective game is developed.

For Clemente et al., (2015) the ability to investigate the connections between team members determines the proximity of understanding the causes and factors associated with the group's relational evolution. “The identification of patterns of interaction in a working group enables the understanding of social processes that can enhance collective performance” (Clemente et al., 2015).

In addition, Clemente et al., (2015) state that quantification has irrefutable advantages regarding the validity and possibility of replication of research in different contexts, this work seeks to follow this path, as using mathematical methods can provide new insights on the players relationship and team dynamics (Clemente et al., 2015). However, it's not all roses, and “sometimes it is very hard to extract relevant features of the players manually from a structured dataset” (Behravan & Razavi, 2020).

2.2 Machine Learning in sports

According to Keshav, (2020) “the future of the sports industry lies in the hand of technology”. And, because of that, “sports activities have many opportunities for intelligent systems” (Maanijou & Mirroshandel, 2019).

For these opportunities to be followed, and because sports have too much information regarding the games and training sessions, it’s important to understand that “humans have a certain limitation when processing a large set of information” (Fialho et al., 2019). As Fialho et al., (2019) say, sports have a great amount of data and this could be perceived as a good example of AI problem. Another author who strengthens this idea is (Keshav, 2020) when he states that “whatever form of AI is used, it is evident that football is a sport that benefits from technical integration” (Keshav, 2020).

ML is a type of IA and will be presented later in this report works that are being done in the football domain, but first we will make a brief presentation of ML according to some authors. As Oliver et al., (2020) defends, ML offers a contemporary statistical approach where algorithms have been specifically designed to deal with imbalanced data sets and enable the modelling of interactions between many variables. Or as Herold et al.,

(2019) affirms in their study, this field of research, known as ML, is a form of AI that uses algorithms to detect meaningful patterns based on positional data.

"Machine learning is typically divided into two areas: supervised and unsupervised learning. In supervised learning, one aims to optimize a model on a set of labelled training data to fit to a given response. Case in point, the team tactic of penetrating passes can be learned by feeding the machine with examples of penetrating passes" (Herold et al., 2019).

"In unsupervised approaches, a model aims to uncover structures and patterns in unlabelled data. For complex problems with an unknown desired response, unsupervised machine learning approaches have been used to measure inter-player coordination, team–team interaction including the time preceding key game events such as shots on goal, and compactness" (Herold et al., 2019).

ML is a relatively new concept in football, and little is known about its usefulness in identifying performance metrics that determine match outcome. Therefore, it is important to try to join sports scientists with data scientists regarding a better understanding of this perspective, and try to take the best of it, or as Herold et al., (2019) suggest, that ML analysts/computer scientists, sports scientists and football coaches/analysts should form a symbiotic system to obtain more accurate information with respect to individual and collective performance that may influence the outcome of football matches.

Data science has emerged as a strategic area that, supported by the great possibility of data production for analysis, allows knowledge discovery in sport science with the aim of filling some gaps that traditional statistical methods could not achieve (García-Aliaga et al., 2020).

2.3 Machine Learning studies on football

This section presents some of the studies on the use of ML in football, with the presentation of the ML algorithms used and their datasets (see Table 2.1). These studies give an overview of the approaches used in this area, allowing the identification of the most interesting algorithms to be used in the study to be carried out in this work.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

To identify the relevant studies, a literature review was conducted by searching scientific databases like B-on, Scopus, SpringerLink and Google Scholar for recent papers (post 2016), having identified ten relevant papers presented in Table 2.1.

Table 2.1. Selected scientific papers on the use of ML in football

Author	Dataset	Algorithms
(Oliver et al., 2020)	355 Players were then tracked for a period of 10 months (August to June) during the 2014–2015 season to prospectively record all injuries sustained in training and competition	J48 consolidated (J48con), an alternating decision tree (ADT) Reduced error pruning tree (REPTree)
(Kusmakar et al., 2020)	Dataset from a season of Major League Soccer division of the United States and Canada. The dataset consists of the possession chain data from 13 matches. The interaction information comprises of time and duration of all ball passes and tackles between players. The dataset also includes the nature of the interaction which can be categorized as being between teammates or between opposing players. The positional information includes the x-y position of all individuals throughout the entire match (90 minutes).	Support vector machines (SVM)
(Baboota & Kaur, 2019)	Matches from 11 seasons (2005 to 2016) of the English Premier League. For statistics, they scraped the data from an online data base (https://www.fifaindex.com).	Gaussian naive Bayes SVM Random forest (RF) Gradient boosting (XgB)
(Knauf et al., 2016)	10 soccer games of the German Bundesliga from the 2011–2012 seasons.	Temporal kernel Gaussian kernel;
(Behravan & Razavi, 2020)	FIFA 20 dataset, containing the characteristics of 18,278 players.	Accelerated particle swarm optimization (APSO-K) SVM
(Matesanz et al., 2018)	The football player transfer market activities among European first leagues from 21countries between seasons 1996/1997 and 2015/2016. These include Austria (AUT; name of examined league: Bundesliga), Belgium (BEL; Jupiter Pro League), Croatia (CRO; 1. HNL), Denmark (DEN; Superligaen), England (ENG; Premier League), France (FRA; Ligue 1), Finland (FIN; Veikkausliiga), Germany (GER; Bundesliga), Greece (GRE; Super League), Hungary (HUN; NB I.), Italy (ITA; Serie A),	Kohonen Self-Organized Maps (Kohonen SOM)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

	The Netherlands (NED; Eredivisie), Norway (NOR; Eliteserien), Poland (POL; Ekstraklasa), Portugal (POR; Liga NOS), Russia (RUS; Premier Liga), Scotland (SCO; Premiership), Spain (ESP; La Liga), Switzerland (SUI; Super League), Turkey (TUR; Super Lig) and Ukraine (UKR; Premier Liga)	
(Rossi, Perri, et al., 2016)	80 training sessions of 26 Italian elite football players over 23 weeks with GPS (Global Positioning System)	Extra tree random forest classifier (ETRFC)
(García-Aliaga et al., 2020)	52 non-spatiotemporal descriptors including offensive, defensive and build-up variables that were computed from OPTA's on-ball event records of the matches for 18 national leagues between the 2012 and 2019 seasons.	Repeated Incremental Pruning Produce Error Reduction (RIPPER)
(Maanijou & Mirroshandel, 2019)	About 300 soccer players with 42 features from Iranian premier league were selected at the beginning of 2015–2016 season.	Multilayer perceptron J48 SVM Logistic Naive Bayes Voting algorithm
(Yang, 2020)	Two teams with 20 players each. One professional and one amateur team.	SVM

Next are presented the objectives and conclusions of each of the scientific papers presented in Table 2.1.

The study done by Oliver et al., (2020) aimed at understanding whether the use of ML improved the ability to predict and identify injury risk factors in elite male of youth football players. Oliver et al., (2020) compared a logistic regression model to ML techniques to understand if it these techniques could improve the overall analysis in injury prevention or the risk of injury. In the conclusion Oliver et al., (2020) reported that both statistical methods have similar accuracy and very low sensitivity. The ML model that they applied improved the identification of some important factors (asymmetries of knee valgus angle and body size) in injury profile in youth football players.

Kusmakar et al., (2020) aim to quantify player's interactions and connect that with the outcome using a ML approach. In conclusion, the ML approach in this study showed a mean sensitivity of 78.3% (95% confidence interval (CI): 70.3% - 85.3%), a specificity of 73.8% (95% CI: 69% - 80.2%) and an overall accuracy of 75.2% in predicting the outcome of segments (a phase of the match, e.g.: throw-in; pass; recovery; ball lost; challenge) in the matches, in this particular case, to predict the sequences of action that lead to a shot and the team that make that action.

Baboota & Kaur, (2019) presented a generalized predictive model for the results of the English Premier League. They created a feature set for determining the most important factors for predicting the results of a football match, and consequently created a highly accurate predictive system using ML. Their best model, using gradient boosting (XgB) achieved a "performance of 0.2156 on the Ranked Probability Score (RPS)[1] metric for game weeks 6 to 38 for the English Premier League aggregated over two seasons (2014–2015 and 2015–2016), whereas the betting organizations that their consider (Bet365 and Pinnacle Sports) obtained an RPS value of 0.2012 for the same period. Since a lower RPS value represents a higher predictive accuracy" (Baboota & Kaur, 2019), none the less, the model was not able to outperform the bookmaker's predictions, despite obtaining promising results.

Knauf et al., (2016) propose a novel class of spatio-temporal convolution kernels (STCK) to capture similarities in multi-object scenarios. They compare kernels and efficient approximations thereof to baseline techniques for clustering tasks using artificial and real-world data from team sports. They analyse two teams to understand their characteristics, the distance between kernels (STCKdist) capture was the characteristic of the best performing team. In theirs study, team A clearly acted with many short moves (long trajectories in cluster 1) and integrated many players in the playmaking (cluttered medians). By contrast, Team B acted with many long moves (short trajectories in all clusters) and preferred linear actions" (Knauf et al., 2016).

The authors Behravan & Razavi, (2020) propose a novel method for estimating the value of players in the transfer market, based on the FIFA 20 dataset. The dataset was clustered using an automatic clustering algorithm, the APSO-K (automatic particle swarm optimization) algorithm, which resulted in detecting four clusters: goalkeepers, strikers,

[1] RPS - is a measure of how good forecasts that are expressed as probability distributions are in the outcomes observed in a match.

defenders, and midfielders. Then, for each cluster, an automatic regression method, able to detect the relevant features, is trained. They were able to estimate the value of players with 74% accuracy.

Matesanz et al., (2018) explore the evolution of the football player's transfer network among 21 European first leagues between the seasons 1996/1997 and 2015/2016 and the season sports results from those transfers. In conclusions, first, the European transfer network seems to have reached an upper limit in both the number of clubs involved and in the number of players transferred. At the global financial crisis (2007/2008) the numbers stopped growing and the network became more connected and denser. Second, the relationship between transfer market activities and sportive performance is positive, i.e., transfer money spend is a key factor for success in UEFA (Union of European Football Associations) competitions.

The aim of the study performed by Rossi, et al. (2016) was to describe: "i) an in season short-term football training cycle; ii) the importance of the features provided by the GPS; iii) the overall periodization of the training sessions". The Extra tree random forest classifier algorithm was able to characterize the training inside the short-term cycle with a 63.6% (accuracy). This algorithm was able to classify the trainings in the two different class (short/long term cycle) with an accuracy of 90%.

The study of García-Aliaga et al., (2020) aimed to determine, using ML methods, whether the technical-tactical behaviours of players according to position are identified by their statistics, without including spatial-temporal descriptors. The results of the study revealed that it is possible to identify the position of players based on their statistics regarding of their technical-tactical behaviours. The study also sought to verify ML's ability to identify the most influential variables in each of the positions and to find groups of anomalous players, this was done by detecting the outliers in each group, and understanding what variables had more influence in each one.

The main objective of the study done by Maanijou & Mirroshandel, (2019) was to propose a solution to solve existing problems in identifying player's ability and ranking players based on existing information. The authors assume that automating the process of ranking soccer players is beneficial to managers who have limitations in their budget and time. They applied different classification algorithms on prepared data in order to choose the best model to rank new players. To improve their results, they used a new voting ensemble learning along with a genetic algorithm to combine all classification methods. Applying the voting algorithm (is an ensemble machine learning model that combines the predictions from multiple models) increased accuracy and F-measure of classification algorithm by 2.22% and 3.1%.

Yang, (2020) aim to create and test an AI model to evaluate of football training using the SVM algorithm together with fuzzy evaluation. To test the model, an experiment was conducted involving 20 players from two teams (a semi-professional team and a well-organized amateur team) and five experts (coaches). Although with many limitations, it was possible to conclude that this AI model for the evaluation of football team trainings is efficient (especially regarding the evaluation of players' athletic ability), having had better results with the semi-professional team.

The analysed studies showed different perspectives on the use of ML in football (*e.g.*, predicting match results, injury risk and classification of football players), allowing to verify the wide of the use of ML in football.

3. BUSINESS UNDERSTANDING AND DATA PREPARATION

This chapter explains the business environment of the project, with a summary of the data collected and stored. Subsequently, the structure of the data is explained, and a description of the data collected is presented, to which follows the data preparation stage. This stage covers all activities to construct the final dataset from the initial raw data (such as data cleaning and transformation) to prepare the data for use in the different models. Finally, some of the technological tools used in this work are presented.

3.1 Business understanding

Football game is within the so call collective games as one of the most important sports (Oliver et al., 2020), for that reason, it has been under a greater development in many areas within. To understand these developments, it's important to approach some of the concepts that are presented during this project and explain some of them before starting.

Football training sessions have evolved during the years, and new approaches started to become the norm. As the football becomes more and more professionalized for the players but also, for the coaching staff, the organization and demanding improves. These improvements help the planning organization, to gain advantage and be closer to success. The training sessions are organized in a microcycle that usually is in maximum of seven days and these training sessions are categorized regarding the previous match or the next match (Oliveira et al., 2018).

The dataset is composed by a team in the 2nd Regional Division of Football Association of Santarém in the 2018/2019 Season, and the team was composed by twenty-eight players with a mean age of 21.59 years.

The games in the Championship were divided in two phases. 1st Phase composed by 14 games and 2nd Phase by 10 games. The top three teams of the two series standings would qualify for the 2nd phase, where the division promotion would be disputed by six teams participating in this phase, the top three would be promoted to the 1st division of AF Santarém. For the Ribatejo Cup 4 games were played.

The dataset used for game analysis had a total of 33748 different episodes, regarding the different players, and games they played. The dataset regarding the training sessions had a total of 24360 different episodes, regarding the different players.

For the elaboration of this work, the dataset was composed by all players, who had worn GPS tracking devices from PlayerTek (Catapult Innovations, 2021), and the information was recorded for all the matches (26) and thirteen microcycles of three training sessions each, meaning thirty-nine training sessions (39).

Regarding the variables collected by the GPS system, they will be explained later in this work, so they can be associated with the acronyms defined.

3.2 Data understanding

This step of the process it is critical, as is it will be important to understand the data collected and explaining the variables in an easier and understandable way. In Table 3.1 are presented the original variables of the data collected in each dataset.

Table 3.1 Variables in each dataset

Games Dataset	Training Sessions Dataset
'Athlete', 'Game', 'Position', 'Home or Away', 'Pitch', 'Final Score', 'Minutes', 'Game Condition', 'RPE_J', 'sRPE_J', 'HR', '%HR', '<60%HR', '60-74,9%HR', '75-89,9%HR', '>90%HR', 'Player Load', 'Player Load.UA/min', 'Distance_m', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4'	*'Athlete', 'Game', 'Position', 'Home or Away', 'Pitch', 'Final Score', 'Minutes', 'RPE_J', 'Player Load', 'Player Load.UA/min', 'Distance Total', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4'*

Table 3.2 presents the description of all variables in the games and training sessions datasets.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 3.2. Description of variables in the datasets

Variable	Description
Game	Unit: Integer (1-26) Is the number of the game in the Championship games sequence.
Game condition	Unit: Integer (1-3) Identifies the players that played all the match (1), were sub-out (2) and sub-in(3).
Home or Away	Unit: Integer (0; 1) It's organized in a way that makes the understanding of the Home field advantage easier. 0 it's for the games at Home and 1 it's for away games.
Pitch	Unit: Integer (0; 1) It's to identify what type of pitch was the game. 0 it's for a natural grass and 1 for artificial grass.
Final Score	Unit: Integer (0; 1; 3) It's to understand how the final score relates to a victory, draw or loss. 0 it's for a lost game, 1 for a draw and 3 for a win.
Minutes	Unit: Minutes It's the number of minutes that the players are actively in the game.
RPE **(Rate of Perceived Exertion)**	Unit: Integer between (1-10) Stands for rate of perceived exertion, which is the numeric estimate of someone's exercise intensity. The ratings were originally based on those in the Borg scale, a way to measure how hard you're exercising, which ranges from 1 (no exertion) to 10 (extremely hard). Also represented in total per minutes (sRPE).
Heart Rate (HR)	Unit: Absolute and HR% It represents the HR during the game and it can be divided by the intensity, that was done by percentage of HRmax. The maximum heart rate is calculated as HRmax = 220 – age.
Player Load	Unit: Absolute (U.A.) or relative (U.A./min) values. Is calculated based on the acceleration data that are registered by the triaxial accelerometers. This variable, considered as a magnitude vector, represents the sum of

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

	the accelerations recorded in the anteroposterior, medio-lateral and vertical planes. Represented in Total and in Arbitrary Units (U.A.) per minutes.
Distance_m, Distance.m/min	Unit: Meters and meters per minute The total distance (Distance_m) and total distance per time (Distance_m/min) provide a good global representation of volume of exercise (walking, running) and is also a simple way to assess individual's contribution relative to a team effort.
Distance.0-3, Distance.3.4, Distance.4-5.5, Distance.5.5-7, Distance.>7	Unit: Meters per seconds These variables describe the distances at different speeds and it is divided in five different speed zones: "walking / jogging distance, 0.0 to 3.0 m/s; running speed distance, 3.0 to 4.0 m/s; high-speed running distance, 4.0 to 5.5 m/s; very high-speed running distance, 5.5 to 7.0 m/s; and sprint distance, a speed greater than 7.0 m/s" (Miguel et al., 2021).
WRRatio (Work Ratio)	Unit: Meters Are used to describe footballer's activity profiles. To calculate this ratio, one speed zone is defined as "pause/rest", and other(s) as "work/activity", through which the distances traveled in these zones are used to determine the ratio (division of the work quantity by the rest quantity). The data you have considers as "pause" the distance traveled at a speed < 1.5m/s and as "work" the distance traveled > 1.5m/s.
Acceleration	Unit: Meters per seconds Categorized based upon the acceleration of the movement, which is thought to represent the "intensity" of the action. It's divided in "low intensity", 0.0 to 2.0 m/s2; "moderate intensity", 2.0 to 4.0 m/s2; and "high intensity", greater than 4.0 m/s2 (Miguel et al., 2021).
Deacceleration	Unit: Meters per seconds Categorized based upon the deacceleration of the movement, which is thought to represent the "intensity" of the action. It's divided in "low intensity", 0.0 to -2.0 m/s2; "moderate intensity", -2.0 to -4.0 m/s2; and "high intensity", greater than -4.0 m/s2 (Miguel et al., 2021).

3.3 Data preparation

All the information collected by the GPS tracking system and context data (home/away; pitch; game; final score) in the games were integrated, so that they can be analysed, and conclusions drawn that can help in decision making. This information was loaded into an

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Excel file with the intention of becoming easier to explore the data and understand the effects of the different variables.

After all data collected and compiled in an Excel file, it was necessary to find out if there were some incongruences in the data to have valid datasets.

Missing and null values were found in the original datasets that would interfere with the use of the models to be applied. Thus, it was decided to replace these values by the mean values regarding the other players so that these values would not influence the outcomes. Although, if there was a variable that didn't had any value for the players in a particular game the values regarding that variable in the dataset was discarded.

As the classification algorithms need to have the outcome as a binary response, it was important to change the variable "Final Score" into a binary response. For this reason, was created a new variable in the dataset called "Win", which is the variable that classified the games into binary results (win/not win). This variable was based in the variable "Final-Score" (variable related to the game result which as the values win, draw, and lost).

In order to better understand the data collected, it was done a correlation plot for both datasets (Figure 3.2 and Figure 3.1) with the intention of understanding how the variables in the dataset related to each other, and how did they related to the target variable "Win".

Regarding the variable "Win", there weren't any physiological variables with strong relationship in both datasets. The only strong relationships were with contextual variables, like "Pitch" and "Home or Away", and as the work was aiming to analyze physiological variables, these were not important in variable selection for the models presented in the studies.

Analysing the plots, it is possible to see that there are strong correlations between the variables regarding the distances and between the variables regarding the accelerations and deaccelerations. This was to be expected, because these variables tend to be related to each other because they are in the same specific group.

There were also negative relationships between some variables in the match dataset that should be mentioned, such as the variables "Minutes" and the variables

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

"Distance.m/min" and "Player Load.UA/min", which had negative correlations (see Figure 3.2).

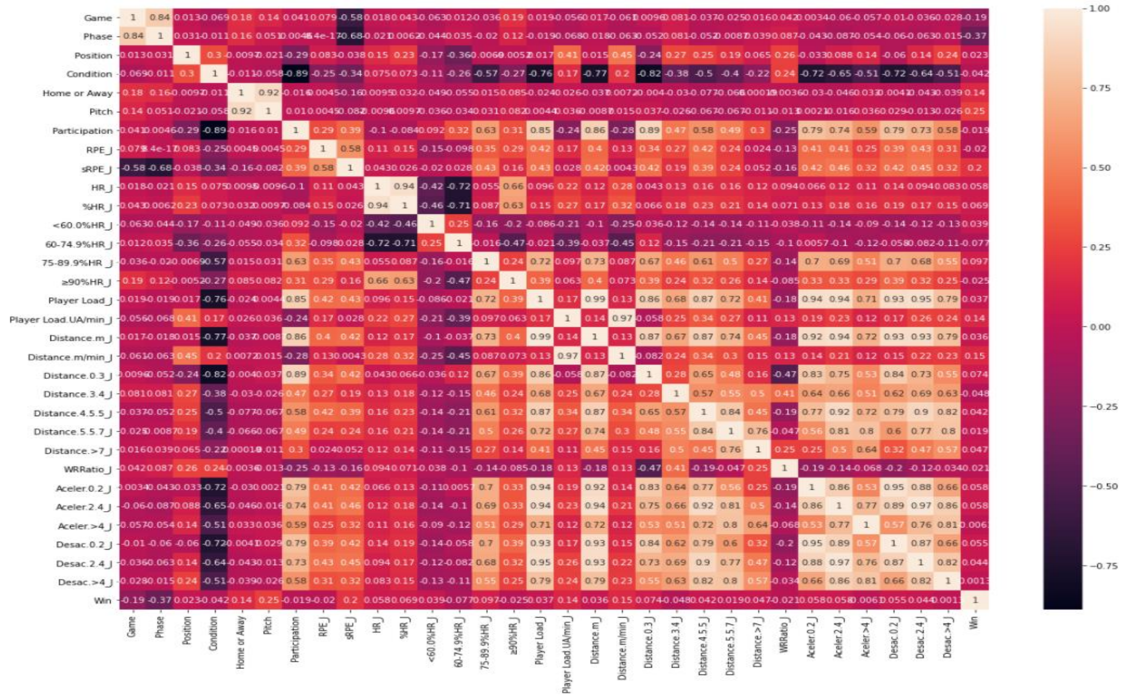


Figure 3.1. Correlation Matrix Match Dataset Plot

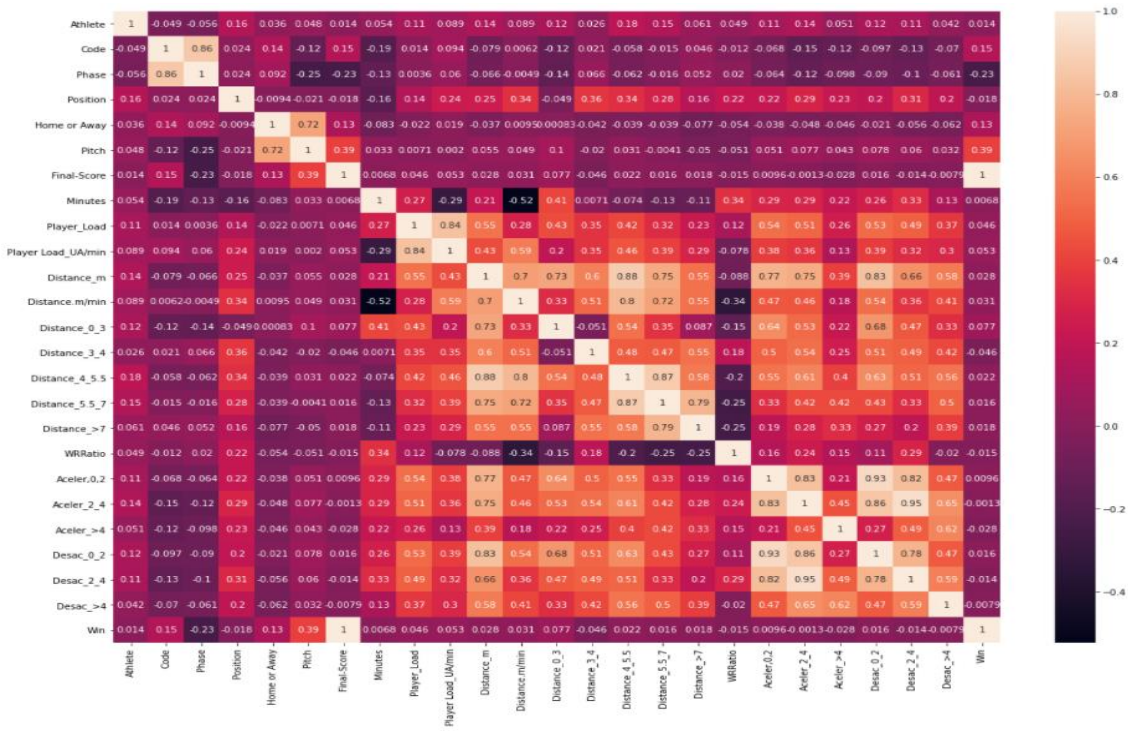


Figure 3.2. Correlation Matrix Training Dataset Plot

In the end of these procedures, the physiological variables were selected from each dataset and are summarized in Table 3.3.

Table 3.3 Physiological variables used in each dataset

Games Dataset	Training Sessions Dataset
'RPE', 'HR', '%HR', '<60%HR', '60-74,9%HR', '75-89,9%HR', '>90%HR', 'Player Load', 'Player Load.UA/min', 'Distance_m', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4'	*'Player Load', 'Player Load.UA/min', 'Distance.m', 'Distance_m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4'*

Besides the selection of the variables for the datasets presented in the previous table (Table 3.3) and the creation of the target variable “win”, it is necessary to perform aggregation operations of the datasets' records.

In the case of the game’s dataset (with 33748 episodes), the study had two different objectives. The first was to understand which were the physiological variables of the team players that most influenced the victory in the game, and the second was to understand which were the physiological variables of the players per position (Central-Defender; Full-Backs; Midfielders; Offensive-Midfielders; Wingers; Fowards) that most influenced the victory. Thus, it was necessary to create two new datasets based on the first dataset. The first one was created by grouping all the players per game, in terms of the different variables, having obtained a dataset with 676 episodes. In the second one the information was grouped by the players' positions per game, having obtained a dataset with 10725 episodes.

The training dataset (with 24360 episodes) was used to perform two different analysis, the first one, similar to the first study regarding the match analysis, to verify which variables of the training most influence the victory (per microcycle, which corresponds to the set of training sessions in the week before the match), and a second analysis to verify which variables per training sessions (Tuesday, Thursday, and Friday) of the microcycle (the week) most influence the victory in the match. For this, two datasets were created based on the training dataset. The first was created by grouping the episodes per microcycle, having obtained a dataset with 1040 episodes and the second by grouping the episodes by training sessions (Tuesday, Thursday, and Friday) of the microcycle having obtained 18981 episodes.

Finally, a last dataset involving the two main datasets (see Table 3.3) was built for the study aiming at creating a new way to understand the selected line-up made by the coaching staff using the collected variables, and to try to create an index of preparedness of the players for the match. For this it was needed to create a new variable named, line-up, which identified in the training dataset the players that were in the starting eleven of each game. To create this new variable in the training set it was necessary to build a query (join) that related the two datasets using the variables "athlete" and "game" resulting in a new dataset with 16996 episodes and with different variables from the original datasets (see Table 3.4).

Table 3.4 Variables in Training Sessions Dataset with Line-up Identification

Training Sessions Dataset with Line-up Identification
'Athlete', 'Week', 'Position', 'Home or Away', 'Pitch', 'Final-Score', 'Minutes', 'Player_Load', 'Player Load_UA/min', 'Distance_m', 'Distance.m/min', 'Distance_0_3', 'Distance_3_4', 'Distance_4_5.5', 'Distance_5.5_7', 'Distance_>7', 'WRRatio', 'Aceler_0_2', 'Aceler_2_4', 'Aceler_>4', 'Desac_0_2', 'Desac_2_4', 'Desac_>4', 'Win', 'Line-up'

3.4 Tools Used

For collecting the GPS data, was use a Catapult device called PlayerTek, and it was worn by all the players, in every game and every training session. In the games, it was only recorded the data of the players that played, that mean, all the starters and the players who enter during the game, to have the most indicators possible. For the training sessions, it was recorded the information for all the players in the team. However, "even if there is a great volume of indicators it is very difficult for coaches and athletic trainers to periodize the trainings because of the multidimensional characteristics of football performance" (Rossi, Perri, et al., 2016).

For the data organization, it was used the Microsoft Excel, because all the data was easy to rename and put in chronological order. And for the data analysis, was use a free platform call Google Collaboratory.

4. MODEL AND EVALUATION

After understanding the data and being able to prepare the data going forward, the next step of the CRISP-DM process, was to create the different models to be used. Models with different characteristics were hypothesized, both to determine which variables best contribute to team wins, and to determine which variables most influence wins, by player position on the field. In this chapter, as well as in the following chapter, we have chosen to present the relative information per study carried out. In this way, it is easier to understand the different steps performed for each of the studies, in the case of this chapter the steps of modelling and model evaluation.

4.1 Players' physiological variables (match data) that most contribute to victory

This study aimed to identify, using ML algorithms and the match dataset, which physiological variables have the most influence on winning the game by analysing them in their entirety per match (the match dataset was grouped by match).

4.1.1 Modelling

To better evaluate the variables and understanding which were the most relevant in predicting a victory, four models with different sets of features (variables) were hypothesized (see Table 4.1).

Table 4.1. ML Models for the match dataset

Model	Features/Variables
Model 1	*['RPE', 'HR', '%HR', '<60%HR', '60-74,9%HR', '75-89,9%HR', '>90%HR', 'Player Load', 'Player Load.UA /min', 'Distance_m', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4', 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4']*
Model 2	*['60-74,9%HR', '75-89,9%HR', 'Player Load', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Accel.0-2', 'Accel.2-4', 'Deacc.0-2', 'Deacc.2-4']*
Model 3	*['>90% HR', 'Distance.>7', 'WRRatio', 'Acceler.>4', 'Deacc.>4']*
Model 4	*['Player Load.UA/min', 'Distance.m/min', 'Distance.0.3', 'Acceler.0.2']*

In the first model it consists of all the variables available in the dataset in order to understand how the impact of all variables as a whole on winning the game.

For the second model, all variables that had a high correlation between them (Figure 3.1) and those that showed high intensity for the game were removed.

The variables that were related to high intensity were used in the third model, to understand how these variables influence the outcome of the game.

For the last model, the variables that according to the literature review have the highest correlation with victory were chosen, for example, the variable Player Load/min (Soto et al., 2019).

Since the aim was to determine which variables contribute to the game win (a binary variable existing in the dataset described in the section 3.3) classification algorithms were selected. Thus, based on the conducted literature review, six ML algorithms were selected: Naïve Bayes (NB), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Extreme Gradient Boosting (XgB).

For each ML algorithm the dataset was divided into training and testing dataset. Training dataset was composed by 66% of all the data and the remaining was used for testing.

4.1.2 Evaluation

The performance of each ML algorithm was assessed through a confusion matrix, where it shows the present results of True Positives, False Positives, True Negatives and False Negatives. With these, it was possible to calculate the sensitivity, specificity and accuracy of each algorithm, in order to evaluate its performance regarding the four models.

Table 4.2, 4.3 and 4.4, and figure 4.2 present the results of the ML algorithms for the four models with and without Cross-Validation (CV).

The Cross Validation is a technique of resampling, and it is used to evaluate ML models with limited data (Baboota & Kaur, 2019). This procedure can use the sample of data used in validation and training, in different folds (Figure 4.1), or different samples. In this study, it was used in 7 folds, where all data was used for testing.

For the results with CV, the Standard Deviation (StD) was also measured. The best model was Model 4, which is presented with an image (see Figure 4.2) for better understanding of the results.

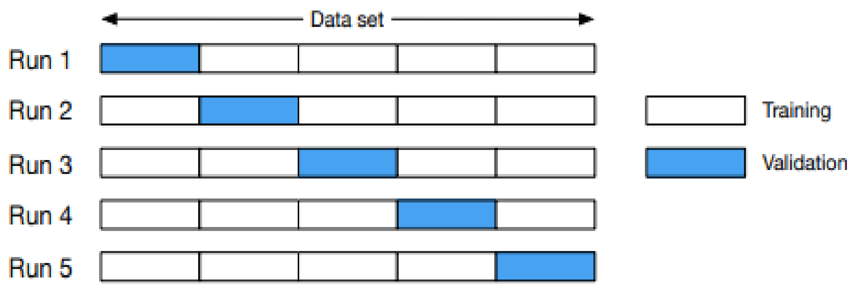


Figure 4.1. Example of Cross Validation with 5 folds

Model 1 (Table 4.2) as a mean accuracy of 0.476 without CV, and 0.560 with CV. The ML algorithms with the best accuracy without a CV were DT and NB with 57% and with CV were RF and SVM with 62% both.

Table 4.2. Results of ML algorithms for Model 1

Metrics / Algorithms	KNN	DT	RF	NB	SVM	XgB
Accuracy	**0.571**	**0.571**	0.429	**0.571**	0.286	0.429
Accuracy with CV	0.600	0.520	**0.620**	0.400	**0.620**	0.600
StD Accuracy with CV	0.330	0.140	0.080	0.330	0.280	0.220
Sensitivity	0.500	0.500	0.750	0.500	0.500	0.250
Specificity	0.333	0.667	0.667	0.667	0.667	0.667

Table 4.3 presents the results for Model 2 with a mean accuracy of 0.429 without CV, and 0.552 with CV. The best ML algorithm for this model regarding accuracy was KNN with a result of 57% without CV and XgB with a result of 76% with CV.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 4.3. Results of ML algorithms for Model 2

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	**0.571**	0.286	0.429	0.429	0.429	0.429
Accuracy with CV	0.600	0.620	0.400	0.450	0.480	**0.760**
StD Accuracy with CV	0.220	0.190	0.280	0.340	0.290	0.230
Sensitivity	0.500	0.000	0.750	0.250	0.250	0.500
Specificity	0.667	0.667	0.000	0.667	0.667	0.333

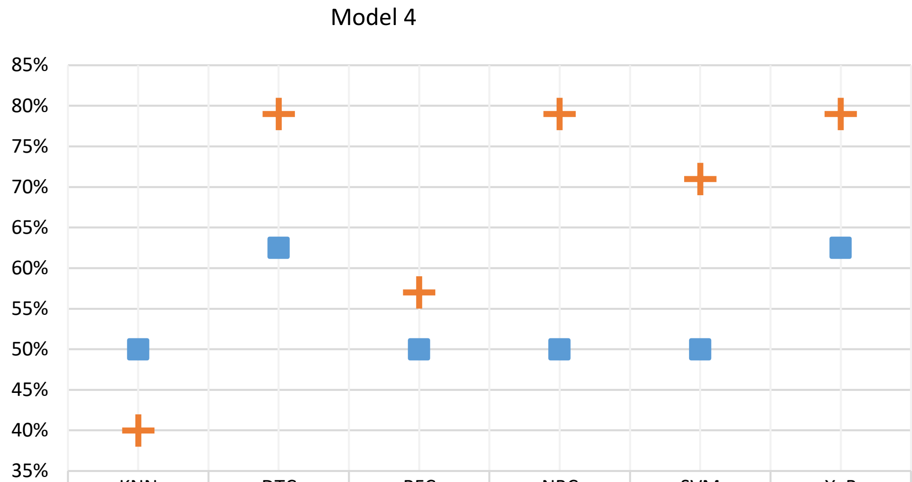
Model 3 as the worst mean results for accuracy (see Table 4.4) from all the models with 0.429 without CV and 0.340 with CV. The algorithms with the best accuracy without CV were RF and SVM 57% and with CV was NB with a result of 45%.

Table 4.4. Results of ML algorithms for Model 3

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	0.429	0.286	**0.571**	0.429	**0.571**	0.286
Accuracy with CV	0.290	0.210	0.360	**0.450**	0.400	0.330
StD Accuracy with CV	0.190	0.190	0.360	0.290	0.330	0.240
Sensitivity	0.200	0.000	0.800	0.600	0.400	0.000
Specificity	1.000	1.000	0.000	0.000	1.000	1.000

The last model used in this analysis was Model 4 (Figure 4.2), which held the best overall results, and the best individual technique results. This model has a mean accuracy of 0.542 without CV and 0.675 with CV. The two algorithms with the best results were DTC and XgB with a 63% of accuracy without CV and 79% of accuracy with CV.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt



	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	50%	63%	50%	50%	50%	63%
Accuracy with CV	40%	79%	57%	79%	71%	79%
StD Accuracy with CV	8%	19%	28%	19%	19%	19%
Sensitivity	43%	50%	57%	50%	50%	50%
Specificity	100%	100%	0%	50%	50%	100%

Figure 4.2. Plot of the ML algorithm results for Model 4

Table 4.5 presents a summary of the results of running all ML algorithms selected for the four models, with the accuracy mean values per ML algorithm. The best mean values were for the SVM (55%) and XgB (62%) algorithms.

Table 4.5 Results of the ML algorithms regarding all the models

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Model 1 Accuracy with CV	60%	52%	62%	40%	62%	60%
Model 2 Accuracy with CV	60%	62%	40%	45%	48%	76%
Model 3 Accuracy with CV	29%	21%	36%	45%	40%	33%
Model 4 Accuracy with CV	40%	79%	57%	79%	71%	79%
Mean Values	47%	54%	49%	52%	**55%**	**62%**

4.2 Players' physiological variables (match data) that most contribute to victory by position in the field

This study aimed at identifying which physiological variables per type of player (player's position on the field) influence the most the game outcome (victory). This study used the match dataset grouped by game and type of player.

4.2.1 Modelling

In this study, by position, it was decided to eliminate the Goalkeeper position from the study because it is a position where, due to its specificities, the variables do not have the same influence in winning the match as in the other positions and therefore it becomes difficult to understand its contribution to winning the match, at least with the models proposed in this study. So, the following positions for this study were used: Central Defender (CD), Full Back (FB), Midfielder (MC), Offensive Midfielder (OM), Winger (W) and Forward (F).

To achieve this, the dataset was divided by position on the field and a Recursive Feature Elimination (RFE) (Arndt & Brefeld, 2016) was used for each position, to rank the feature variables related to winning. The variables selected by the algorithm for each position are presented in Table 4.6.

Table 4.6 RFE selected feature variables for each position

Position	Features/Variables
Central Defender	*['Distance.m', 'Distance.0-3', 'Distance.3-4', 'Distance.>7']*
Full Back	*['<60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
Central Midfielder	*['Player Load.UA/min', 'Distance.m/min', 'WRRatio', 'Aceler.>4']*
Offensive Midfielder	*[<'60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
Winger	*['<60.0%HR', '60-74,9%HR', '75-89,9%HR', 'Player Load.UA/min']*
Forward	*['Distance.m', 'Distance.4-5.5', 'Distance.>7']*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Considering the results obtained in the previous study (see 4.1.2) it was decided to use, in this study, the two algorithms that had performed better: the SVM and XgB algorithms.

4.2.2 Evaluation

The selected variables as well as the accuracy of the ML algorithms for each player position on the field are presented in Table 4.7. The variables selected (using the RFE) allowed to understand the different needs for the different positions in a football game, and how these variables affect the match outcome. The variables are different for each position, and they are related to the specific demands. The two best results with the ML algorithms were to the Forwards with a 74% with SVM and 62% with XgB. All the positions achieved over 50% with at least one of the ML algorithms.

Table 4.7. Accuracy of the ML algorithms for the RFE selected features for each position

Position	SVM	XgB	Features/Variables
CD	66%	62%	*['Distance.m', 'Distance.0-3', 'Distance.3-4', 'Distance.>7']*
FB	63%	61%	*['<60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
MC	63%	59%	*['Player Load.UA/min', 'Distance.m/min', 'WRRatio', 'Aceler.>4']*
OM	52%	62%	*[<'60.0%HR', '60-74,9%HR', 'Player Load.UA/min', 'WRRatio']*
W	55%	56%	*['<60.0%HR', '60-74,9%HR', '75-89,9%HR', 'Player Load.UA/min']*
F	**74%**	**62%**	*['Distance.m', 'Distance.4-5.5', 'Distance.>7']*

4.3 Players' physiological variables (training sessions dataset) that most contribute to victory

This study aimed at understanding how training performance/sessions affect the match day, more specifically, the match outcome (or winning). For this study was used the training sessions dataset grouped by microcycle (week before the game).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4.3.1 Modelling

To better evaluate the variables and understand which were the most relevant in predicting a Win, the same four models (from the previous study) were implemented (see Table 4.8).

Table 4.8. ML Models for the training sessions dataset

Model	Features/Variables
Model 1	*['Player Load', 'Player Load.UA/min', 'Distance_m', 'Distance.m/min', 'Distance.0-3', 'Distance. 3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Distance.>7', 'WRRatio', 'Accel.0-2', 'Accel.2-4', 'Accel.>4' , 'Deacc.0-2', 'Deacc.2-4', 'Deacc.>4']*
Model 2	*['Player Load', 'Distance.m/min', 'Distance.0-3', 'Distance.3.4', 'Distance.4-5.5', 'Distance.5.5-7', 'Accel.0-2', 'Accel.2-4', 'Deacc.0-2', 'Deacc.2-4']*
Model 3	*['Distance.>7', 'WRRatio', 'Acceler.>4', 'Deacc.>4']*
Model 4	*['Player Load.UA/min', 'Distance.m/min', 'Distance.0.3', 'Acceler.0.2']*

For this study, the goal was also to determine which variables contribute to the win of the game. In this case, the classification algorithms from the previous study were selected, Naïve Bayes (NB), K-Nearest Neighbours (KNN), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM) and Extreme Gradient Boosting (XgB).

For each ML algorithm the dataset was divided into training and testing. Training dataset was composed by 66% of all the data, the remaining was used for testing.

4.3.2 Evaluation

Model 1 (Table 4.9) has a mean accuracy of 0.500 without CV, and 0.702 with CV. The ML algorithm with the best accuracy without a CV was NB with 70% and with CV was XgB with 76%.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 4.9. Results of ML algorithms for Model 1 (training sessions dataset)

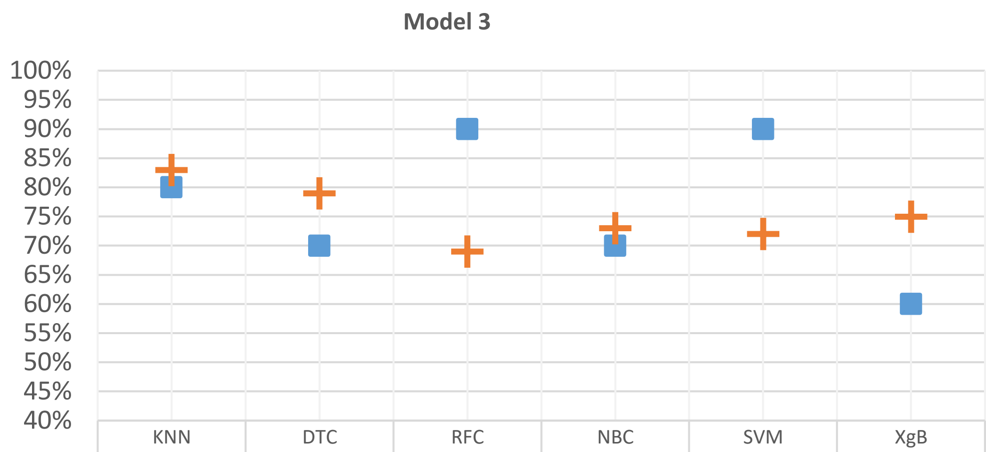
Metrics / Algorithms	KNN	DT	RF	NB	SVM	XgB
Accuracy	0,300	0,400	0,600	**0,700**	0,600	0,400
Accuracy with CV	0,690	0,730	0,610	0,730	0,690	**0,760**
StD Accuracy with CV	0,160	0,070	0,150	0,070	0,240	0,080
Sensitivity	0,000	0,000	0,833	1,000	0,833	0,666
Specificity	0,300	0,400	0,250	0,250	0,250	0,000

Table 4.10 presents the results for Model 2. This model has a mean accuracy of 0.433 without CV, and 0.752 with CV. The best ML algorithms for this model regarding accuracy without CV were SVM and XgB with a result of 60%. When CV is applied, the best result is for the XgB algorithm with a result of 76%.

Table 4.10. Results of ML algorithms for Model 2 (training sessions dataset)

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	0,200	0,300	0,500	0,400	**0,600**	**0,600**
Accuracy with CV	0,760	0,760	0,750	0,760	0,690	**0,790**
StD Accuracy with CV	0,080	0,080	0,100	0,170	0,190	0,120
Sensitivity	0,330	0,333	0,666	0,666	1,000	0,833
Specificity	0,000	0,250	0,250	0,000	0,000	0,250

Model 3 has the best mean results for accuracy (see Figure 4.3) for all the models with 0.767 without CV and 0.752 with CV. The algorithms with the best accuracy without CV were RFC and SVM with 90% and with CV was KNN with a result of 83%. This result with CV was the best performance in all the models run in this work.



	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	80%	70%	90%	70%	90%	60%
Accuracy CV	83%	79%	69%	73%	72%	75%
StD Accuracy CV	15%	9%	12%	7%	10%	10%
Sensitivity	100%	100%	100%	100%	100%	57%
Specificity	33%	0%	67%	0%	67%	67%

Figure 4.3. Plot of the ML algorithm results for Model 3 (training sessions dataset)

The last model used in this analysis was Model 4 (Table 4.11) that in contrast with the match data analysis did not perform as well when applying CV. This model has a mean accuracy of 0.783 without CV and 0.593 with CV. The three algorithms with the best results were DTC, RFC and SVM with a 90% of accuracy without CV. With CV the best accuracy was achieved with the DTC algorithm with an accuracy of 76%.

It would be expected that this model would have better results, as it happened in the first study (see section 4.1), since this model is based on the variables already identified in the literature as being the most important for winning the match, but this did not happen. Nevertheless, the values were not far from those obtained in model 3 (76% vs 83%).

Table 4.11 Results of ML algorithms for Model 4 (training sessions dataset)

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Accuracy	0,500	**0,900**	**0,900**	0,700	**0,900**	0,800
Accuracy with CV	0,510	**0,760**	0,580	0,650	0,550	0,510
StD Accuracy with CV	0,290	0,080	0,110	0,110	0,260	0,230
Sensitivity	0,500	1,000	1,000	0,770	1,000	0,880
Specificity	0,000	0,000	0,000	0,000	0,000	0,000

Table 4.12 presents a summary of the results of running all ML algorithms selected for the four models, with the accuracy mean values per ML algorithm. The best mean values were for the KNN (75%) and XgB (75%) algorithms.

Table 4.12 Results of the ML algorithms regarding all the models (training sessions dataset)

Metrics / Algorithms	KNN	DTC	RFC	NBC	SVM	XgB
Model 1 Accuracy with CV	69%	61%	61%	73%	69%	76%
Model 2 Accuracy with CV	76%	76%	75%	76%	69%	79%
Model 3 Accuracy with CV	83%	79%	69%	73%	72%	75%
Model 4 Accuracy with CV	72%	76%	58%	65%	55%	70%
Mean Values	**75%**	73%	66%	72%	66%	**75%**

4.4 Players' physiological variables (training sessions dataset) that most contribute to victory by training session

This study goal was to understand which variables in the different training sessions (Tuesday, Thursday, and Friday) have the most influence on winning the match in the weekend after those trainings. This is especially relevant as each training session has a specific program that has different effects on the variables above. By identifying which variables in each training session contribute more to the victory, the coaching staff can be more attentive to these specific variables in each corresponding training session.

4.4.1 Modelling

The aim was to identify relevant physiological variables in each training session, and how these variables influence the most the match outcome (winning). The dataset was divided for each training, and the division was regarding the days after the match (MD) or before the match as supported by the scientific community. So, the matches were always played at a Sunday, and the training days were, Tuesday, Thursday, and Friday, which means, MD+2, MD-3, MD-2 (Martín-García et al., 2018), that represents the days after a match day or the days before.

It was applied the XgB regressor algorithm to select the variables with more importance regarding winning by each training session and is presented in Table 4.13.

Table 4.13. Selected variables for the training sessions

Day of the week	Training	Features/Variables
Tuesday	MD + 2	*['Player_Load', 'Player Load_UA/min', 'Distance_0_3', 'Aceler_2_4']*
Thursday	MD - 3	*['Player Load_UA/min', ' Distance_0_3', ' Distance 3_4', 'Aceler_>4']*
Friday	MD - 2	*[' Distance_m', ' Distance_0_3', ' Distance_3_4', 'Aceler_>4']*

To perform this study, the two best performing ML algorithms regarding the training sessions studies were used: KNN and XgB (see Table 4.12).

4.4.2 Evaluation

The selected variables for each training are presented in Table 4.14 as well as the accuracy of the ML algorithms for each training session.

The selected variables by XgB regressor enabled to understand the different purposes of each training session, and how these affect a winning outcome in football games. The variables are different for each training as they are related to specific demands of each training session, in each microcycle, and could help characterizing each training session.

The best results with the ML algorithms were to the MD-2 with an 81% with XgB and MD-3 with 78% with XgB.

Table 4.14. Results of ML algorithms for training sessions

Training	KNN	XgB	Features/Variables
MD + 2	57%	72%	*['Player_Load', 'Player Load_UA/min', 'Distance_0_3', 'Aceler_2_4']*
MD - 3	66%	78%	*['Player Load_UA/min', ' Distance_0_3', ' Distance 3_4', 'Aceler_>4']*
MD - 2	72%	81%	*[' Distance_m', ' Distance_0_3', ' Distance_3_4', 'Aceler_>4']*

These models display a good performance with the use of the XgB algorithm to predict the victory in the game, with an accuracy always equal or higher than 72%, which reveals that the variables chosen for each training session are adequate and should be a concern of the coach.

Although, the results with KNN were not so good, in the case of MD-2, an accuracy of 72% was still obtained.

4.5 Predicting the starting line-up and chose the better prepare players

The objective of this study was to create a model to help football managers to define the starting players for each match. This study was divided in two parts: a classification study, which used the target variable line-up (which represents if a player was - value 1- or not - value 0 – in the line-up), and a regression study, which created an index to select or not a player as a starter.

In this study, we used the data from the training sessions merged with the data from the matches (see Table 3.4) grouped by player type, because the variables for choosing the players vary according to their position on the field. The idea of this study is to provide a tool to help the coach in the decision making process, based on the index created, which will allow improve the choice of starting players for each match.

4.5.1 Modelling

For this study two models were used, one with all the variables available in the dataset (see Table 3.4) and another with the variables (see Table 4.6) identified as most important by the RFE algorithm used in the study on physiological variables (match data) that contribute the most to the victory of the match by position in the field (see section 4.2).

The variables of this second model can help to understand how the players on the team are chosen and whether these variables have influence in selecting the line-up for the game in the weekend (presented in Table 4.15).

For this study, it was necessary to exclude the Heart Rate variables, in both models, because they had too many missing or null values. For this reason, our models, especially in Full-Backs and Wingers, had fewer variables in the study.

Table 4.15. Model 2 variables

Positions	**Features/Variables**
Central Defender (CD)	*'Distance_m', 'Distance_0_3', 'Distance_3_4', 'Distance_>7'*
Full-Back (FB)	*'Player Load_UA/min', 'WRRatio'*
Central-Midfielder (CM)	*'Player Load_UA/min', 'Distance_m/min', 'WRRatio','Aceler_>4'*
Winger (W)	*'Player Load_UA/min'*
Forward (F)	*'Distance_m', 'Distance_4_5.5', 'Distance_>7'*

To perform this study, the two best performing ML algorithms regarding the training sessions studies were used: KNN and XgB (see Table 4.12).

After this, it was thought that a Logistic Regression could be used to create an index that would help to choose the players best prepared for each game in relation to the selected physiological variables. This logistic regression used the variables from model 2 and the victory variable, with the objective of knowing if the chosen players were the best prepared, that is, if they won the game or not.

This could help the coaching staff in a team, create a less subjective analysis regarding the decision of the selection of players for the games, because he will have an index that indicates which players are the most suitable for each position, from the point of view of physiological variables.

4.5.2 Evaluation

First it was preferred to run both algorithms with all the variables available for all the positions (model 1), which results are presented in (Table 4.16).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 4.16. Model 1 (all variables without RFE) results

Metrics/Position	CD	FB	MC	W	F
KNN Accuracy	0,86	0,81	0,64	0,55	0,50
KNN Accuracy with CV	**0,98**	**0,81**	0,68	0,63	**0,80**
KNN StD Accuracy with CV	0,05	0,06	0,08	0,09	0,11
XgB Accuracy	0,86	0,90	0,74	0,70	0,56
XgB Accuracy with CV	**0,98**	**0,88**	0,67	0,66	**0,80**
XgB StD Accuracy with CV	0,05	0,09	0,08	0,17	0,14

This model as better results for the positions that have less players. This is to be expected as these positions have fewer players available, so it is easier to pick who played. It is important to highlight all the results achieved with CV were always higher than 63%, and in three player positions (CD, FB and F) the algorithms obtained an accuracy above 80%.

The results of the second model, an adaptation of the model that used RFE algorithm to identify the most important variables by player position in the field (see section 4.2), are presented in Table 4.17.

Table 4.17. Model 2 (RFE selected variables) results

Metrics / Position	CD	FB	MC	W	F
KNN Accuracy	0,86	0,66	**0,66**	0,70	0,39
KNN Accuracy with CV	**0,98**	**0,74**	0,55	**0,70**	**0,64**
KNN StD Accuracy with CV	0,05	0,18	0,07	0,15	0,12
XgB Accuracy	0,86	**0,90**	**0,74**	**0,78**	0,39
XgB Accuracy with CV	**0,98**	0,88	0,72	0,73	**0,74**
XgB StD Accuracy with CV	0,05	0,09	0,06	0,05	0,14

CD had the same results with model 1 and 2 with both algorithms, because of the small sample for this position. FB had the same results in the two models with XgB algorithm. F had better results with model 1 than with model 2. Although, for MC and W model 2 had better results achieving over 70% with the XgB algorithm.

After applying these models, it was important to understand if it was possible to create an index regarding the physical ability for each position. This could help the coaching staff in the decision-making process.

It was used a logistic regression, which is a useful way to create a model of probability of a certain class or event to exist. That model of probability will give an index. To create this, it was important to understand if the logistic regression could be applied to all the positions regarding the model created. So, the logistic regression was only applied to the positions were the variables achieved a P-value < 0,05. For that reason, only the results for FB (Figure 4.4), MC (Figure 4.5) and W (Figure 4.6) are presented.

	coef	std err	z	P>|z|	[0.025	0.975]
const	6.3733	2.382	2.676	0.007	1.705	11.041
Player Load_UA/min	-2.0795	0.905	-2.297	0.022	-3.854	-0.305
WRRatio	0.0998	0.133	0.749	0.454	-0.162	0.361

Figure 4.4. Logistic Regression for Full-Backs

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.0404	1.799	-1.690	0.091	-6.567	0.486
Player Load_UA/min	2.3050	0.784	2.939	0.003	0.768	3.842
Distância_m/min	-0.0994	0.064	-1.550	0.121	-0.225	0.026
WRRatio	0.1218	0.121	1.006	0.314	-0.116	0.359
Aceler_>4	0.0198	0.024	0.813	0.416	-0.028	0.068

Figure 4.5. Logistic Regression for Midfielders

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2661	1.978	-1.651	0.099	-7.143	0.611
Player Load_UA/min	3.1864	1.074	2.967	0.003	1.082	5.291
Distância_m/min	-0.0938	0.046	-2.040	0.041	-0.184	-0.004

Figure 4.6. Logistic Regression for Wingers

Based on logistic regression, an index was created to help select the players for the matches, based on the values of the physical variables of the week of training prior to the match, and understand if the player chosen for a particular match was the best prepared one. Using this index, the tables with the index values per player per microcycle (which includes the training sessions of the week) were created for the positions of the players where the regression was considered valid: Full-Backs (Figure 4.7), Wingers (Figure 4.8) and Midfielders (Figure 4.9).

The tables presented are divided into four columns and all are important for understanding the data and the proposed analysis. The first column represents the index in the training session dataset. The second column represents whether the match outcome in which the player participated was a win or not. The third column is the value of the logistic regression index. Finally, the fourth column has three values: "ID" that represents the identity of the player; "LUp" that indicates if this player was in the line-up; and "Week" that represents the number of the microcycle. To better understand the information related to each figure, two cases of correct decisions (with a green circle) and one decision that was incorrect (red circle) will be desecrated for each figure.

The table on Full-Back position is shown in (Figure 4.7), where good decisions are highlighted in green and bad ones in red.

	Victory	Index	ID/LUp/Week
44	1	0.784477	(8, 0, 2)
59	1	0.865374	(8, 0, 2)
79	1	0.807929	(8, 0, 6)
72	1	0.526791	(8, 0, 6)
67	1	0.662985	(8, 0, 6)
21	1	0.911756	(8, 1, 3)
95	1	0.911697	(8, 1, 3)
56	1	0.847468	(8, 1, 3)
30	1	0.726883	(8, 1, 4)
34	1	0.875619	(8, 1, 4)
52	1	0.917653	(8, 1, 7)
0	1	0.667928	(8, 1, 7)
75	1	0.672547	(8, 1, 7)
4	0	0.837648	(8, 1, 8)
36	0	0.783271	(8, 1, 9)
51	1	0.790515	(8, 1, 11)
60	1	0.947780	(8, 1, 12)
3	1	0.857310	(8, 1, 12)
23	1	0.792565	(8, 1, 12)
18	1	0.847074	(8, 1, 13)
55	0	0.735844	(10, 0, 8)
48	0	0.653187	(10, 0, 8)
20	0	0.833750	(10, 1, 1)
82	0	0.866102	(10, 1, 1)
80	0	0.411594	(10, 1, 1)
19	1	0.866759	(10, 1, 3)
38	1	0.865714	(10, 1, 3)
93	1	0.865563	(10, 1, 4)
61	1	0.909490	(10, 1, 4)
29	1	0.947054	(10, 1, 5)
5	1	0.787528	(10, 1, 5)
22	1	0.913458	(10, 1, 6)
70	1	0.903314	(10, 1, 6)
76	1	0.555153	(10, 1, 6)
94	1	0.776973	(10, 1, 7)
81	0	0.830641	(10, 1, 9)
46	1	0.841000	(10, 1, 10)
50	1	0.937167	(10, 1, 10)
63	1	0.864772	(10, 1, 11)
47	1	0.803693	(10, 1, 11)
91	1	0.927370	(10, 1, 12)
73	1	0.958043	(10, 1, 12)
62	1	0.598875	(10, 1, 12)

	Victory	Index	ID/LUp/Week
68	0	0.776008	(19, 0, 8)
45	1	0.511412	(19, 0, 13)
33	1	0.612893	(19, 0, 13)
83	0	0.406908	(19, 1, 9)
10	0	0.560884	(19, 1, 9)
31	0	0.442858	(19, 1, 9)
58	1	0.440744	(19, 1, 10)
53	1	0.493869	(19, 1, 10)
88	1	0.608787	(20, 0, 7)
71	1	0.568071	(20, 0, 7)
6	1	0.877902	(20, 0, 7)
77	1	0.675493	(20, 0, 11)
74	0	0.748045	(20, 1, 1)
2	0	0.621270	(20, 1, 1)
43	0	0.313663	(20, 1, 1)
57	1	0.673096	(20, 1, 2)
35	1	0.602226	(20, 1, 2)
25	1	0.595407	(20, 1, 2)
16	1	0.695457	(20, 1, 3)
66	1	0.654608	(20, 1, 4)
13	1	0.557984	(20, 1, 5)
86	0	0.656770	(20, 1, 8)
11	0	0.779103	(22, 0, 1)
9	0	0.282544	(22, 0, 1)
69	0	0.865051	(22, 0, 1)
27	1	0.746690	(22, 1, 2)
65	1	0.908037	(22, 1, 2)
15	0	0.468884	(24, 0, 8)
42	0	0.817034	(24, 0, 8)
8	1	0.583174	(24, 0, 12)
39	1	0.901443	(24, 0, 12)
12	1	0.674647	(24, 1, 5)
89	1	0.842830	(24, 1, 5)
26	1	0.933669	(24, 1, 5)
1	1	0.849880	(24, 1, 6)
24	1	0.880728	(24, 1, 6)
40	1	0.816811	(24, 1, 10)
90	1	0.774215	(24, 1, 10)
78	1	0.841953	(24, 1, 10)
92	1	0.795412	(24, 1, 13)
28	1	0.929237	(24, 1, 13)
32	1	0.529422	(25, 0, 3)

Figure 4.7. Full-backs Index analysis

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The first correct decision according to the model is from player with ID 8 (in the records 21, 95 and 56) in which the three training sessions concerning the third week, were training sessions where the index reached 91% in the first two sessions and 84% in the last training session of the week, that is, with values that indicate that he should be selected for the Line-Up, which in fact happened. It's also worth mentioning that the team won.

The second player analysed, ID number 20 (in the records 88, 71 and 6), had two training sessions with a low index value, achieving 60% and 56%, and a last training session with 87%. He wasn't selected for the Line-Up, which according to the model is correct.

The third player analysed was the same player, ID 20 (in the records 57, 35 and 25), but in the second week, where he was selected for the Line-Up but achieved 67% in the first training session, 60% in the second training session and 59% in the third session, so regarding the model proposed, this was a bad decision made by the coaching staff in comparison with the model proposed.

Figure 4.8 presents the table regarding the Winger position.

	Victory	Index	ID/LUp/Week		Victory	Index	ID/LUp/Week
44	1	0.985863	(1, 0, 13)	48	1	0.783514	(11, 1, 7)
32	0	0.600585	(4, 0, 1)	61	1	0.738240	(11, 1, 7)
20	0	0.480366	(4, 0, 1)	8	0	0.744943	(11, 1, 8)
60	0	0.515228	(4, 0, 8)	7	0	0.643265	(11, 1, 8)
11	0	0.663983	(4, 0, 9)	26	0	0.596132	(11, 1, 8)
3	1	0.657554	(6, 0, 10)	30	0	0.732079	(11, 1, 9)
53	1	0.708375	(6, 0, 11)	51	0	0.663642	(11, 1, 9)
22	1	0.937456	(9, 1, 5)	50	1	0.678929	(11, 1, 10)
5	1	0.863028	(9, 1, 5)	59	1	0.690924	(11, 1, 10)
24	1	0.784041	(9, 1, 6)	4	1	0.790619	(11, 1, 11)
21	1	0.933873	(9, 1, 6)	34	1	0.656216	(11, 1, 11)
54	1	0.873258	(9, 1, 6)	9	1	0.284175	(11, 1, 11)
39	0	0.904690	(9, 1, 8)	18	1	0.754153	(11, 1, 12)
31	1	0.848315	(9, 1, 11)	40	1	0.654352	(11, 1, 13)
45	1	0.472807	(9, 1, 11)	41	1	0.941015	(11, 1, 13)
0	1	0.955910	(9, 1, 12)	15	1	0.893009	(20, 0, 6)
36	1	0.929533	(9, 1, 13)	33	1	0.806206	(20, 1, 5)
57	1	0.952612	(9, 1, 13)	29	0	0.886793	(22, 0, 9)
38	1	0.679002	(11, 0, 5)	10	0	0.849515	(22, 0, 9)
28	1	0.668134	(11, 0, 5)	37	1	0.897781	(22, 1, 3)
12	0	0.527733	(11, 1, 1)	6	1	0.891282	(22, 1, 3)
2	0	0.668352	(11, 1, 1)	14	1	0.942967	(22, 1, 3)
56	1	0.677689	(11, 1, 2)	46	1	0.898490	(22, 1, 4)
13	1	0.840461	(11, 1, 3)	25	1	0.890783	(22, 1, 4)
35	1	0.724676	(11, 1, 3)	42	1	0.862396	(22, 1, 10)
43	1	0.782688	(11, 1, 4)	19	1	0.905162	(22, 1, 10)
55	1	0.783562	(11, 1, 4)	58	1	0.893174	(22, 1, 10)
16	1	0.815601	(11, 1, 6)	17	0	0.574255	(27, 1, 1)
49	1	0.555276	(11, 1, 6)	23	1	0.494040	(27, 1, 2)
47	1	0.855227	(11, 1, 6)	52	1	0.766669	(27, 1, 2)

Figure 4.8. Wingers Index analysis

First player analysed in this position, ID number 11 (in the records 38 and 28), was analysed in the two training sessions regarding the fifth week, where he achieved 67% in

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

the first training session and 66% in the last training of the week, and he wasn't selected for the Line-Up. This goes along with the model.

Another choice that is aligned with the model proposed, was the second player analysed, ID number 22 (in the records 37, 6 and 14), had three good sessions in the week three, achieving 89%, 89% and 94%, he was selected for the Line-Up and the team won the game in that week.

The third player analysed was the player with ID number 27 (in the records 23 and 52), in the second week was selected for the Line-Up but only achieved 49% in the first training and 76% in the last training session, which was a poorly decision regarding the model.

Finally, the table regarding the Midfielders is presented in Figure 4.9.

	Victory	Index	ID/LUp/Week
44	1	0.784477	(8, 0, 2)
59	1	0.865374	(8, 0, 2)
79	1	0.807929	(8, 0, 6)
72	1	0.526791	(8, 0, 6)
67	1	0.662985	(8, 0, 6)
21	1	0.911756	(8, 1, 3)
95	1	0.911697	(8, 1, 3)
56	1	0.847468	(8, 1, 3)
30	1	0.726883	(8, 1, 4)
34	1	0.875619	(8, 1, 4)
52	1	0.917653	(8, 1, 7)
0	1	0.667928	(8, 1, 7)
75	1	0.672547	(8, 1, 7)
4	0	0.837648	(8, 1, 8)
36	0	0.783271	(8, 1, 9)
51	1	0.790515	(8, 1, 11)
60	1	0.947780	(8, 1, 12)
3	1	0.857310	(8, 1, 12)
23	1	0.792565	(8, 1, 12)
18	1	0.847074	(8, 1, 13)
55	0	0.735844	(10, 0, 8)
48	0	0.653187	(10, 0, 8)
20	0	0.833750	(10, 1, 1)
82	0	0.866102	(10, 1, 1)
80	0	0.411594	(10, 1, 1)
19	1	0.866759	(10, 1, 3)
38	1	0.865714	(10, 1, 3)
93	1	0.865563	(10, 1, 4)
61	1	0.909490	(10, 1, 4)
29	1	0.947054	(10, 1, 5)
5	1	0.787528	(10, 1, 5)
22	1	0.913458	(10, 1, 6)
70	1	0.903314	(10, 1, 6)
76	1	0.555153	(10, 1, 6)
94	1	0.776973	(10, 1, 7)
81	0	0.830641	(10, 1, 9)
46	1	0.841000	(10, 1, 10)
50	1	0.937167	(10, 1, 10)
63	1	0.864772	(10, 1, 11)
47	1	0.803693	(10, 1, 11)
91	1	0.927370	(10, 1, 12)
73	1	0.958043	(10, 1, 12)
62	1	0.598875	(10, 1, 12)
14	1	0.673258	(10, 1, 13)
37	1	0.871795	(10, 1, 13)
41	1	0.940116	(10, 1, 13)
84	1	0.629758	(14, 0, 4)
54	1	0.701712	(14, 0, 4)
87	1	0.603532	(14, 1, 5)

	Victory	Index	ID/LUp/Week
49	1	0.695287	(14, 1, 7)
17	0	0.432993	(14, 1, 8)
7	0	0.577226	(14, 1, 8)
64	0	0.786492	(14, 1, 9)
96	0	0.500672	(14, 1, 9)
85	0	0.583720	(19, 0, 8)
68	0	0.776008	(19, 0, 8)
45	1	0.511412	(19, 0, 13)
33	1	0.612893	(19, 0, 13)
83	0	0.406908	(19, 1, 9)
10	0	0.560884	(19, 1, 9)
31	0	0.442858	(19, 1, 9)
58	1	0.440744	(19, 1, 10)
53	1	0.493869	(19, 1, 10)
88	1	0.608787	(20, 0, 7)
71	1	0.568071	(20, 0, 7)
6	1	0.877902	(20, 0, 7)
77	1	0.675493	(20, 0, 11)
74	0	0.748045	(20, 1, 1)
2	0	0.621270	(20, 1, 1)
43	0	0.313663	(20, 1, 1)
57	1	0.673096	(20, 1, 2)
35	1	0.602226	(20, 1, 2)
25	1	0.595407	(20, 1, 2)
16	1	0.695457	(20, 1, 3)
66	1	0.654608	(20, 1, 4)
13	1	0.557984	(20, 1, 5)
86	0	0.656770	(20, 1, 8)
11	0	0.779103	(22, 0, 1)
9	0	0.282544	(22, 0, 1)
69	0	0.865051	(22, 0, 1)
27	1	0.746690	(22, 1, 2)
65	1	0.908037	(22, 1, 2)
15	0	0.468884	(24, 0, 8)
42	0	0.817034	(24, 0, 8)
8	1	0.583174	(24, 0, 12)
39	1	0.901443	(24, 0, 12)
12	1	0.674647	(24, 1, 5)
89	1	0.842830	(24, 1, 5)
26	1	0.933669	(24, 1, 5)
1	1	0.849880	(24, 1, 6)
24	1	0.880728	(24, 1, 6)
40	1	0.816811	(24, 1, 10)
90	1	0.774215	(24, 1, 10)
78	1	0.841953	(24, 1, 10)
92	1	0.795412	(24, 1, 13)
28	1	0.929237	(24, 1, 13)
32	1	0.529422	(25, 0, 3)

Figure 4.9. Midfielders Index analysis

First player analysed, ID number 10 (in the records 93 and 61), was analysed in the two training sessions regarding the fourth week, where he achieved 86% in the first training session and 90% in the last training session, in this week he was selected for the Line-Up and the team won the game. Regarding the model proposed, it was a good option.

The second player analysed, ID number 19 (in the records 45 and 33), was another correct decision made by the coaching staff, because he had two not so good sessions in the week thirteen, achieving 51% and 61%, he wasn't selected for the Line-Up but the team won the game in that week.

The third player analysed was the player ID number 20 (in the records 57, 35 and 25) in the second week, where he was selected for the Line-Up, and achieved 67% in the first training, 60% in the second training session, and 59% in the last training session, regarding the model, he did not achieve great results in the index, the team won the game.

5. DISCUSSION OF THE RESULTS

This chapter presents a discussion of the results of the five studies presented in the previous chapter, comparing them with the existing literature. This part of the project will explained the results and put some context and explaining why these results were important.

5.1 Players' physiological variables (match data) that most contribute to victory

The objective of this study was to identify which physiological variables have the most influence on winning matches (using ML algorithms), by analysing these physiological variables in each game. For this purpose four models were created with the following characteristics: the first model was conducted with all the variables measured; the second model was constructed with all variables, but those that had greater correlation between them were excluded, and those that were used in the third model; the third model was constructed to understand if high intensity had a greater correlation with the victory, and that was the model with the worst results; finally, the fourth model was related with some of the variables defend by some authors Altavilla et al., (2017) and those that had some correlation with wining.

The fourth model was the one that obtained the best results with an accuracy of 79% with three of the algorithms using CV. One of important facts discovered in this study was that the high intensity was not a key factor in game winning, and that was not corroborated in a study done by Altavilla et al., (2017) where they affirm that "high intensity covered distance has traditionally been identified as a key indicator of the physical performance during the matches".

This fourth model also revealed that the total distance travelled in a given period of time, represented in the dataset by the variable "Distance.m/min", is directly linked to team performance, which means that teams who kept their players more active, in the end will have better results.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5.2 Players' physiological variables (match data) that most contribute to victory by position in the field

In this second study, as previously mentioned, the goal was to identify which physiological variables have more influence in the victory of the game by position of the player on the field, and some of the results obtained (see Table 4.7) were consistent with the literature.

As was seen in this study, distance covered at max speed ("Distance.>7") from Forwards (F) was one of the variables that had influence in winning the game. This is in line with the studies of Almulla & Alam, (2020) and Baptista et al., (2018) who argue that the distance covered by Forwarders at very high speed from the winning team was higher than the losing side. The variables chosen for the Forward position presented the best results in predicting a win in this context, with a 74% in SVM and 62% in XgB (see Table 4.7).

Regarding Central Midfielders (CM) one of the results related with the selection of variable "distance covered per minutes" by the RFE algorithm is consistent with the studies of Borghi et al., (2020) and Altavilla et al., (2017) that refer that CM has the highest value in total distance covered during the games.

As for the Central Backs (CB) the "Distance 0.3" variable obtained by the RFE algorithm is also corroborated by the study of Borghi et al., (2020) who state in their study that the CB covered the shortest distances during the match.

Another variable chosen by the RFE algorithm, in this case for the Full Back (FB) and Winger (W) positions, was the "Player Load U.A./min". The choice of this variable for these positions is supported in the literature by (Baptista et al., 2018) who argue that CB had fewer turns per match than FB and W. The same authors also defended that FB covered more high intensity and sprinting distance than CB during the matches (Baptista et al., 2018), this was also well noted in our work because "'WRRatio'" was one of the variables selected for the FB. The variables selected for the FB position presented the second-best results in predicting a win in this context, with a 65% in KNN and 61% in XgB (see Table 4.7).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5.3 Players' physiological variables (training sessions dataset) that most contribute to victory

With this study it was aimed to better understand how training sessions will affect the match day. As it important as to understand the physiological variables involved in game winning it is also important to understand how the physiological variables involving the training session, will affect the game day, more specifically, in game winning.

Regarding this study, it was important to use the same models applied to the first part of this study, because it could give a better understanding when analysing the same football team. Although, the results were different than expected when the same models were used, it was not expected that the models would perform drastically different. As seen before Model 4, model with the best performance in former study, was the worst regarding the accuracy with a CV. And Model 3 that had the worst performance in the previous study, saw its performed improved significantly in this one. The fact that model 3, which includes the high intensity variables, was the best in this study may serve as a wake-up call to football coaches to have more intense training sessions. This study thus seems to recommend that football coaches pay more attention to these variables during training sessions to obtain a better results in matches.

5.4 Players' physiological variables (training sessions dataset) that most contribute to victory by training session

This study aimed to understand which variables are most important per training session and how the variables contribute to victory. After applying the XgB regressor algorithm on the dataset we obtained different variables for each training session (see Table 4.13).

The variable "Player Load", important in the players' recovery process, was selected for the first training session of the week (MD+2) which is in accordance with the existing literature where Swallow et al., (2021) refer that the variable "Player Load" had higher results in MD-5 (which is equal to our MD+2) than the MD-2. Both studies agree that the "Player Load" variable is most important in the recovery process, i.e., in the training session immediately after the match, i.e., the one farthest away from the next match, Although Swallow et al., (2021) defend that "the volume of accelerations efforts within training remained similar across all training days". This is consistent with our study because the variables related to acceleration are present in the various training sessions. However, it was not possible to validate the types of accelerations as the study present in Swallow et al., (2021) does not mention the intensities of the accelerations. Also, regarding this, Clemente et al., (2019) when analysed two teams from different countries, defended that the Portuguese team completed more sprints.

Regarding the variable of total distance ("Distance_m") it was referred as an important variable in the MD-2, and was not mentioned in the other training sessions, which is different than what is supported by the scientific community, which argues that the total distance is bigger in days before and progressive decreases as the week closes to the match (Clemente et al., 2019; Swallow et al., 2021; Martín-García et al., 2018; Oliveira et al., 2018).

5.5 Predicting the starting line-up and chose the better prepare players

This study aims to understand if the physiological variables, and the performance during the week affect the line-up in the game. As stated in Panduro et al., (2021), coaches should pay certain attention to individualized position-specific physical demands to secure a proper training stimulus related to the individual players.

First, the objective of this study was to use the variables and understand if the manager had chosen the better prepare players for the game, or if these variables did not have importance in choosing the line-up. It was also important to see if the algorithms used could predict the line-up. At this point, it was more difficult to predict with precision, due to the data set used, since it had a small sample for each player position, which makes it difficult to obtain reliable answers because there is not enough data. That said the results obtained were good, perhaps better than they probably are. The results were almost always better than 60% with CV, that means that the algorithm, based in the variables, could predict with some accuracy in the line-up, but it is very different from choosing from 8 players for 2 positions than to choose from 4 players to 2 positions as said by

Marcelino et al., (2020), being a good player implies more than the technical and physical capabilities, it's needed to be in the right place at the right time.

The second part of this study was to create an index or aptitude value to help coaching staff deciding in which player should be in the line-up, or which player is better prepared to help the team to win a game, regarding the physiological variables. Once again, the lack of sufficient data meant that the regression model built was only valid for three positions (Full-Backs, Wingers and Midfielders). However, for those three positions interesting results were obtained, which bring a good possibility to understand the choices made by the technical team, i.e., if the choices made by the technical team were in accordance with the index created (which represents the intensity of the value of the physiological variables selected for the different positions of the players), i.e., if the best prepared players by the coach were chosen. This study as the study done by Marcelino et al., (2020) where they defend that the study could be apply to different areas of the sport context, this framework could be improved and apply to recruitment, evaluation, or real time decisions.

It's obvious that these results are less than perfect, but with a bigger dataset, it could have a bigger impact in a team, and for scientific purposes, could bring a deeper analysis to discover new possibilities because, as defended by Stein et al., (2019), the future of team performance will be based on data insights and not on pure intuitions.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

6. CONCLUSIONS

This chapter presents the conclusions of the work carried out. Initially, a summary of the work developed is presented, followed by the main contributions of the work and the fulfilment of the proposed objectives. Finally, the limitations of the work done, and proposals of future work are presented.

6.1 Work Summary

This project was conducted aiming different possibilities, as the datasets provided were different: one was from training sessions and the other from all the matches of the season, which gave the possibility of performing different analysis.

First, was important to understand the main ideas and the objectives proposed. Then, it was important to analyse the literature related to the selected topic, which helped to understand how studies of machine learning have been done in this research area, as well as in other areas.

After the literature review, the business and the data were analysed. This was done to understand the information gather and how this data could be prepared to later apply the ML algorithms.

Next was the modelling and evaluation phase, where the studies proposed to answer the research questions formulated were presented along with a first analysis of the results. This chapter has been divided into different subchapters regarding each of the studies carried out for better understanding the process. To which followed the discussion of the results was presented, in which the results obtained in the previous chapter were compared with the existing literature. This discussion was also done by study.

6.2 Contributions

The main contributions of the work carried out include some of the results obtained in the studies and the dissemination of part of the results in the academic community.

In the first study, a Model with the variables selected from the authors (“Player Load.UA/min”, “Distance.m/min”, “Distance.0.3”, “Acceler.0.2”), to predict the win in a football match, was used. This model was the one that had better results with an

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

accuracy of 79%, meaning that with only these four variables, it could predict the wining of a football team with high probability.

The second study was related to the players positions on the field. In this case, a Recursive Feature Elimination method was chosen to rank the variables for each position regarding the variable Win. In this study the most important contribute was related to the Forward position, where it can predict the match wining using only three variables with an accuracy of 74%.

The third study was aiming to understand if the training session could be analysed using the same models presented in the first study. The results were different than the ones of the first study, because the best Model was the model related with the high intensity variables. This model could predict the win with an 83% accuracy.

The fourth study meant to analyse the different training sessions and understand which variables could be more important in each training session to win in the match. To each training session was applied a XgB Regressor to rank the variables for their importance to the variable win. The training session with the best results was the MD-2, using the variables “Distance_m”, “Distance_0_3”, “Distance_3_4” and “Aceler_>4”, with 81% of accuracy using the XgB algorithm.

The last study aimed at a better understanding of the choices made by the coaching staff and propose a model to verify which players are better prepared in terms of their physiological variables. In this part of the study, the variables selected in the second study were used for the positions selected/chosen. All the weeks were analysed in order to understand if the choices were the correct ones and was created a physical ability index of the players, which could help the coaching staff in the choice of the team's starting players for each game.

The results obtained in these studies can be used by the coaching staff of a football team as they help to prepare the team, also helping to identify by player's position on the field which are the best prepared players who may provide a better performance in the game. Bottom line, the use of ML applied to training and team selection can represent a major step forward in the evolution of the sport, providing possibly critical information to the training staff that will allow in the end to help to better understand the demands of the positions on the field.

The work developed was also disseminated in the academic community, with the realization of two actions: a school seminar and the participation in an international conference with Scopus indexation:

- Investigar em TI, Roadmap para o Sucesso - Coimbra Business School;
- The 21st International Conference on Computational Science and Applications (ICCSA 2021), from the work done and presented in this conference, the authors published a paper in Lecture Notes in Computer Science with the title "Predicting physiological variables of players that make a winning football team: a machine learning approach" (Cortez et al., 2021).

6.3 Limitations.

This work can help to approach the gap between the sport science and computational sciences and to create knowledge and help practitioners in implementing new ways to engage the game.

However, this study lacks in depth because of the small sample analysed, because although the team had almost thirty players in their squad, the sample in the training sessions and the sample in the games, was too small and it could help the study to try to find different teams to use these approaches.

In this same perspective, it could also bring better results if the data from the training sessions was for the whole season, and not only for thirteen weeks, because it was a big problem to apply these techniques, when the sample for positions was too small, and it could not create a big difference between the players that had played the game, and those that did not played. It is easy to observe that if you have only three players for two positions on the field, this would not bring the results that it should be expected.

Due to the fact that the data collection used in these studies was not controlled by the authors, there was a need to adjust the study to the available data. This created some problems in data preparation such as the need to exclude the variable Heart-Rate for lack of observations. This created a problem, because it was important to have all the variables used in the matches dataset, so it can improve the knowledge and help understanding the different dynamics in the games and training sessions. This was also a problem in positional analysis, because it reduces the number of variables in same positions, that could help us had better results, and better insights.

6.4 Future Work

This study was conducted in only one team, and this was a reduce sample, so this was a large limitation, but it could serve as an example to try to involve more data from other teams, and permitted a better understanding of the game, using ML algorithms to help.

In the future, it would be interesting to do this kind of analysis with more teams, increasing the statistics to have a better understanding of what is involved in winning. The analysis should also be performed in different contexts, to verify if these variables still maintain the same relevance, if some of the contextual variables changed.

6.5 Final remarks

The studies conducted allowed to understand that the results obtain cannot express the reality involving a football game, because there are plenty factors important as well determining the outcome of a football match, in our studies - winning, with different situations having different effects, "the internal and external load could be affected by different situational variables" (Altavilla et al., 2017).

The results of the studies can be used by the coaching staff in a football team as they can help to prepare a team regarding the most important physiological variables. The studies also provide a better insight into the most important variables used to predict a win regarding the position of the players on the field. This can lead to a better understanding by the coaching staff of the demands associated with the different positions of the players on the field, and consequently to better decision making.

Without a doubt AI can be used to make predictions of matches outcome in the sports industry reliable and certain to an extent. But it is important to understand that if human element is involved in sports, there will always be unpredictability and uncertainty that

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

makes it fascinating and surprising for its viewers (Keshav, 2020). So, it is obvious that this work will always depend on other factors.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

REFERENCES

Almulla, J., & Alam, T. (2020). Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League. *IEEE Access*, *8*(December), 213695–213705. https://doi.org/10.1109/ACCESS.2020.3038601

Altavilla, G., Riela, L., di Tore, A. P., & Raiola, G. (2017). The physical effort required from professional football players in different playing positions. *Journal of Physical Education and Sport*, *17*(3), 2007–2012. https://doi.org/10.7752/jpes.2017.03200

Arndt, C., & Brefeld, U. (2016). Predicting the future performance of soccer players. *Statistical Analysis and Data Mining*, *9*(5), 373–382. https://doi.org/10.1002/sam.11321

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Baptista, I., Johansen, D., Seabra, A., & Pettersen, S. A. (2018). Position specific player load during matchplay in a professional football club. *PLoS ONE*, *13*(5), 1–11. https://doi.org/10.1371/journal.pone.0198115

Behravan, I., & Razavi, S. M. (2020). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, *3*(1). https://doi.org/10.1007/s00500-020-05319-3

Borghi, S., Colombo, D., la Torre, A., Banfi, G., Bonato, M., & Vitale, J. A. (2020). Differences in GPS variables according to playing formations and playing positions in U19 male soccer players. *Research in Sports Medicine*, *00*(00), 1–15. https://doi.org/10.1080/15438627.2020.1815201

Brandão, L., Belfo, F. P., & Silva, A. (2021). Wavelet-based cancer drug recommender system. *Procedia Computer Science, Communications in Computer and Information Science*, *181*, 487–494. https://doi.org/https://doi.org/10.1016/j.procs.2021.01.194

Catapult Innovations. (2021). *Playertek*. https://www.playertek.com/gb/

Clemente, F. M., Mendes, R. S., & Grafos, T. de. (2015). Social Network Analysis : Um Ensaio Sobre a Aplicabilidade Na Análise De Cooperação Em Social Network Analysis : an Overview About the Application on Cooperation Analysis in Sports Contexts Social Network Analysis : Una Visión General Sobre. *Revista Da Faculdade de Educação Física Da UNICAMP*, 175–194.

Clemente, F. M., Owen, A., Serra-Olivares, J., Nikolaidis, P. T., van der Linden, C. M. I., & Mendes, B. (2019). Characterization of the Weekly External Load Profile of Professional Soccer Teams from Portugal and the Netherlands. *Journal of Human Kinetics*, *66*(1), 155–164. https://doi.org/10.2478/hukin-2018-0054

Cortez, A., Trigo, A., & Loureiro, N. (2021). *Predicting Physiological Variables of Players that Make a Winning Football Team: A Machine Learning Approach* (pp. 3–15). https://doi.org/10.1007/978-3-030-86970-0_1

da Costa, I. T., Garganta, J., Greco, P. J., & Mesquita, I. (2011). Proposta de avaliação do comportamento tático de jogadores de Futebol baseada em princípios fundamentais do jogo. *Motriz. Revista de Educacao Fisica*, *17*(3), 511–524. https://doi.org/10.1590/S1980-65742011000300014

Fialho, G., Manhães, A., & Teixeira, J. P. (2019). Predicting Sports Results with Artificial Intelligence - A Proposal Framework for Soccer Games. *Procedia Computer Science*, *164*, 131–136. https://doi.org/10.1016/j.procs.2019.12.164

García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., & Luengo-Sánchez, S. (2020). In-game behaviour analysis of football players using machine learning techniques based on player statistics. *International Journal of Sports Science and Coaching*. https://doi.org/10.1177/1747954120959762

Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., & Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science and Coaching*, *14*(6), 798–817. https://doi.org/10.1177/1747954119879350

Keshav, R. (2020). *Applications of Artificial Intelligence in the Game of Football : The Global Applications of Artificial Intelligence in the Game of Football : The Global Perspective*. *September*. https://doi.org/10.18843/rwjasc/v11i2/03

Knauf, K., Memmert, D., & Brefeld, U. (2016). Spatio-temporal convolution kernels. *Machine Learning*, *102*(2), 247–273. https://doi.org/10.1007/s10994-015-5520-1

Kusmakar, S., Shelyag, S., Zhu, Y., Dwyer, D., Gastin, P., & Angelova, M. (2020). Machine Learning Enabled Team Performance Analysis in the Dynamical Environment of Soccer. *IEEE Access*, *8,* 90266–90279. https://doi.org/10.1109/ACCESS.2020.2992025

Laureano, R. M. S., Caetano, N., & Cortez, P. (2014). Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *13*, 83–98. https://doi.org/10.4304/risti.13.83-98

Maanijou, R., & Mirroshandel, S. A. (2019). Introducing an expert system for prediction of soccer player ranking using ensemble learning. *Neural Computing and Applications*, *31*(12), 9157–9174. https://doi.org/10.1007/s00521-019-04036-9

Marcelino, R., Sampaio, J., Amichay, G., Gonçalves, B., Couzin, I. D., & Nagy, M. (2020). Collective movement analysis reveals coordination tactics of team players in football matches. *Chaos, Solitons and Fractals, 138.* https://doi.org/10.1016/j.chaos.2020.109831

Martín-García, A., Gómez Díaz, A., Bradley, P. S., Morera, F., & Casamichana, D. (2018). Quantification of a professional football team's external load using a microcycle structure. *Journal of Strength and Conditioning Research*, *32*(12), 3511–3518. https://doi.org/10.1519/jsc.0000000000002816

Matesanz, D., Holzmayer, F., Torgler, B., Schmidt, S. L., & Ortega, G. J. (2018). Transfer market activities and sportive performance in European first football leagues: A dynamic network approach. *PLoS ONE, 13*(12), 1–17. https://doi.org/10.1371/journal.pone.0209362

Miguel, M., Oliveira, R., Loureiro, N., García-rubio, J., & Ibáñez, S. J. (2021). *Load Measures in Training / Match Monitoring in Soccer : A Systematic Review*. 1–26.

Morais, A., Peixoto, H., Coimbra, C., Abelha, A., & Machado, J. (2017). Predicting the need of Neonatal Resuscitation using Data Mining. *Procedia Computer Science*, *113*, 571–576. https://doi.org/10.1016/j.procs.2017.08.287

Oliveira, R., Brito, J. P., Martins, A., Mendes, B., Calvete, F., Carriço, S., Marinho, D. A., Ferraz, R., & Marques, M. C. (2018). In-season internal and external training load quantification of an elite European soccer team. *BioRxiv*, 1–19. https://doi.org/10.1101/489187

Oliver, J. L., Ayala, F., de Ste Croix, M. B. A., Lloyd, R. S., Myer, G. D., & Read, P. J. (2020). Using machine learning to improve our understanding of injury risk and prediction in elite male youth football players. *Journal of Science and Medicine in Sport*, *23*(11), 1044–1048. https://doi.org/10.1016/j.jsams.2020.04.021

Panduro, J., Ermidis, G., Røddik, L., Vigh-Larsen, J. F., Madsen, E. E., Larsen, M. N., Pettersen, S. A., Krustrup, P., & Randers, M. B. (2021). Physical performance and loading for six playing positions in elite female football: full-game, end-game, and peak periods. *Scandinavian Journal of Medicine and Science in Sports*, *October 2020*, 1–12. https://doi.org/10.1111/sms.13877

Pimenta, P., Belfo, F., & Trigo, A. (2009). Study the impact of Booking. com user scores and reviews in hotel management. *Book of Abstracts of the CENTERIS 2011–Conference on Enterprise Information Systems*, *30*, 8.

Rossi, A., Perri, E., Trecroci, A., Savino, M., Alberti, G., & Iaia, M. F. (2016). Characterization of In-season Elite Football Trainings by GPS Features: The Identity Card of a Short-Term Football Training Cycle. *IEEE International Conference on Data Mining Workshops, ICDMW*, *0*(December), 160–166. https://doi.org/10.1109/ICDMW.2016.0030

Rossi, A., Savino, M., Perri, E., & Iaia, M. F. (2016). by GPS features. *Icdm*, *December*. https://doi.org/10.1109/ICDMW.2016.106

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Sarmento, H., Anguera, T., Campaniço, J., & Leitão, J. (2010). Development and validation of a notational system to study the offensive process in football. *Medicina*, *46*(6), 401. https://doi.org/10.3390/medicina46060056

Sarmento, H., Marcelino, R., Anguera, M. T., CampaniÇo, J., Matos, N., & LeitÃo, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, *32*(20), 1831–1843. https://doi.org/10.1080/02640414.2014.898852

Seiça, A., Trigo, A., & Belfo, F. P. (2019). LexiNB - Uma Abordagem Bietápica de Classificação de Sentimentos em Tweets Relacionados com as Autoridades Fiscais Portuguesas. *Proceedings of the 19.ª Conferência Da Associação Portuguesa de Sistemas de Informação (CAPSI'2019) Held in Lisboa, Portugal, 11-12 October 2019. Paper 5.*

Soto, P. R., Nieto, D. C., Suarez, A. D., & Ortega, J. P. (2019). *Player Load and Metabolic Power Dynamics as Load Quantifiers in Soccer by*. *69*(September). https://doi.org/10.2478/hukin

Stein, M., Seebacher, D., Marcelino, R., Schreck, T., Grossniklaus, M., Keim, D. A., & Janetzko, H. (2019). Where to go: Computational and visual what-if analyses in soccer. *Journal of Sports Sciences*, *37*(24), 2774–2782. https://doi.org/10.1080/02640414.2019.1652541

Swallow, W. E., Skidmore, N., Page, R. M., & Malone, J. J. (2021). An examination of in-season external training load in semi-professional soccer players: considerations of one and two match weekly microcycles. *International Journal of Sports Science and Coaching*, *16*(1), 192–199. https://doi.org/10.1177/1747954120951762

Yang, Y. (2020). Evaluation Model of Soccer Training Technology Based on Artificial Intelligence. *Journal of Physics: Conference Series*, *1648*(4). https://doi.org/10.1088/1742-6596/1648/4/042085

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt