NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

Machine learning methods for sports injury

Athlete injury risk and Injury profiling

Luís Manuel Figueiredo Ribeiro

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Machine learning methods for sports injury

Athlete injury risk and Injury profiling

by

Luís Manuel Figueiredo Ribeiro

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management.

Supervised by

Vitor Santos, PhD, NOVA Information Management School

11, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 18 of November of 2024

Luís Manuel Figueiredo Ribeiro

DEDICATION

These pages tell the story of a man who dived into the unknown just for the thrill to see what's on the other side.

To my parents for the value driven education and affection.

To Mariana for your wisdom, love and support.

To Yola for your companionship.

Genuinely, couldn't have done it without you. The pack stands together!

ACKNOWLEDGEMENTS

To acknowledge everyone who help me along the process of writing this thesis would be as ambitious as having to re-write it in scientific Mandarin. Which makes me realise how lucky I am to have met you.

Daniel Moura, thank God for your adductor tendinopathy. If it wasn't for that we wouldn't have reunited, and you wouldn't have introduced Hugo.

Damm Hugo, you rock! The beekeeper of the Matrix.

Bruna and Falcão really appreciate your openness to help when times were tough.

João Pedro Figueira, you were such an honest and helpful guy. May I help you as you helped me.

My friend JA, thank you for understanding and supporting me whenever I had to focus on this project.

To Dr, Miguel Gouveia e Brito for receiving me and giving me the opportunity to talk and learn a bit more about the football business.

To Dr. António Martins I thank you for the sarcastic guidance and all the teaching along the process.

Francisco Tavares, your dedication and work precedes you. I am very grateful for the trust which allowed me to pursue my goals. Better late than never, right?

To the team of national padel managers, I thank you for being patient and understanding.

To Rawayana for getting me in the right mood and stimulating my creation.

To all my partners who lost padel games because of my lack of attention during the matches.

To the teachers in the Escola Superior de Saúde da Cruz Vermelha Portuguesa, especially Martinho, Ricardo and Rui who gave me all the conditions to use the school facilities to study and work on this. Thank you, guys!

To all my colleagues in Fisiogaspar, especially my colleagues who had to deal with my absence and my lack of sleep humor. Proud to share my days with all of you.

Prof. Vítor, I know you doubted this. I understand… but never doubt my resilience! Thank you for always standing by my side!

João Vaz, you give another meaning to the term "overwhelmed by work". Thank you for taking the time to help me. If it wasn't for our mentor Raul Oliveira we would have worked together.

To everyone who ever listened to me about the thrill of predicting injuries… Thank you very much!

ABSTRACT

This study analysed a dataset of elite youth football containing various variables, including injury-related data, demographic information, and performance metrics, with the goals of clarify the latest concepts and theories on sports injuries, understand the analytical approach of research in sports injuries and, most of all, explore the use of machine learning methods for injury prediction and injury risk profiling.

Traditional injury prevention models are limited and the potential of machine learning to improve injury prediction and prevention depends greatly on the understanding of the complex interactions between various determinants of injury and the need for a time-sensitive approach to injury prediction. Therefore, in this study we used a methodology and analysis approach that takes into consideration the cyclical nature of shifting risk factors to produce a dynamic, recursive picture of aetiology.

The methodological approach used to guide the application computer-based algorithms in the field of data mining, was the CRISP-DM, which involves six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

In this study we used a Stratified 10-fold cross-validation method to partition the data, and train several machine learning models, including Gradient Boost Classifier, Extreme Gradient Boosting, Decision Tree, K-Nearest Neighbours and Logistic Regression. For injury risk profiling we used k-means clustering algorithm.

In this study we found that muscle injuries are the most common type of injury among elite youth football players.

We found that the Gradient Boost Classifier and Extreme Gradient Boosting classifier had the highest accuracy scores in identifying injury risk, with F1-scores of 87% and 88%, respectively. We also found that total exposure time, training exposure, and match exposure were the features that most impacted an injured athlete.

With this study we concluded that machine learning a promising approach for injury prediction and prevention, and we highlight the importance of considering multiple factors and their interactions in injury risk profiling.

The main limitation of the study was the information contained in the dataset, which was limited in terms of time range and data dimensions.

For further studies we recommend expanding data dimensions, develop and refine machine learning algorithms to improve prediction and risk profiling for specific types of injury and explore the application of these algorithms in real-world sports situations.

KEYWORDS

Machine Learning; Sports Injury; Sports analytics; Complex system; Risk profile; Injury risk index; Injury prediction; Injury prevention.

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

ACL	Anterior Cruciate ligament
ADL	Ankle dorsiflexion lunge
AFL	Associação Futebol de Lisboa
AI	Artificial Intelligence
ASQZ	Adductor Squeeze test
AUC	Area under the curve
BMI	Body mass index
BW	Bodyweight
CMJ	Countermovement Jump
CRISP-DM	Cross Industry Standard Process for data mining
DT	Decision trees
FIFA	Fédération Internationale de Football Association
F-MARC	FIFA Medical assessment and research centre
FN	False Negatives
FP	False Positives
GK	Goalkeeper
GPS	Global positioning system
HROM	Hip Range of motion
HTML	Hypertext markup language
ID	Identification
IOC	International Olympic Committee
IT	Information Technology
KNN	K-nearest neighbours
LR	Logistic regression
MCC	Matthews correlation coefficient
MELCHIOR	Méthode d'Eimination et de Choix Includent les relations d'Ordre

ML	Machine Learning
MRST	Musculoskeletal Readiness Screening Tool
OLTP	Online transactional processing
OSICS	Orchad Sports Injury Classification System
PCA	Principal component analysis
PKFO	Passive Knee Fall out
ROC	Receiver operating charactheristic
SD	Standard Deviation
SHAP	Shapley Additive Explanations
SQL	Structured query language
SVM	Support Vector Machines
TN	True Negatives
TP	True Positives
WHO	World Health Organization
XML	Extensible markup language

1. INTRODUCTION

As a physiotherapist specialized in sports, I admit that having all information might seem as some sort of witchcraft if the reasoning behind the algorithm is not explained to coaches, medical staff and clubs managers. If we fail to turn to practice what scientific evidence shows us then we can´t understand "what were the things that were under our control that contributed to that injury, and what we can do to avoid those in the future" (Fiscutean, 2021).

1.1. CONTEXT

Injury is one of the biggest problems plaguing the sports industry. Although many consider injury simply normal at every sport and at every level. The abilility to predict injuries and avoid them has become the holly grail for sports organisations, no matter if you're the athlete, the coach, the medical staff or the CEO of the organization.

The athlete is eager to compete at high lever injury-free and therefore lengthening his/her career and for sports organizations means that millions may be saved.

In the high-performance sports industry there's a lot of skepticism on the applicability of predictive methods in injuries, and we must disclaim that indeed this happens due to the lack of success in the related research. There are a few reasons for this to happen which we will discuss further on this paper.

What's most interesting is that although sports community is not convinced of the accuracy of injury prediction and still is cautious to use them daily as a foundation for their work, the investment in predicting injuries is not slowing down.

Probably because it seems like predicting an injury is a very easy solvable problem, just get the data, run a machine learning model (or a statistical model) and use the conclusions as recommendations to apply on the athlete/team.

This simplistic vision of it all brings us more questions than answers, specially when it comes to the specificity of data used on the models. Our intention is to gather the updates on this subject and try to fill up the world of injury prediction with some more insights.

According to Van Eetvelde et al. (2021), sports injuries are a consequence of complex interactions of multiple risk factors and inciting events which makes it necessary to have a comprehensive model of their interaction, and even most outdated models only used data related to sport-specific risk factors, either intrinsic or extrinsic and that counts for most part of the unsuccessful attemps to model predictive methods.

In 1994 Meeuwisse (cited by van Eetvelde et al., 2021) noted, predictive modelling should not only focus on the prediction of the occurrence of an injury itself but, moreover, it should try to identify injury risk at an individual level and implement interventions to mitigate the level of risk.

On other hand Quarrie (2001) has stated his reverse opinion by saying that if there's a surveillance system implemented in a sports team and if "it's producing valid and reliable measures of injury experience among the participants" then "we can, at a group level, predict what will happen among those athletes in next series of exposures with reasonable confidence". As you will see, our data is gathered according to the surveillance system of FIFA, which is football specific.

If the question is to be made as "Will we be able to forecast which player will be injured, the type of injury and in which match?" then we believe we are a bit too far away from it. The type of information and the way the information is being analysed at this moment is still far from the reality of the game, either football, rugby or any other sport (Pereira, 2021).

Having this in mind one question arises, are we collecting the adequate data to predict injury Passfield (2017) and Pilka (2023) says that the inputs that we use aren´t necessarily related to the outcomes of interest, they're part of some athlete monitoring processes and may (or may not) be relevant to the outcome of injury prediction.

By definition, a sports injury is "any physical complaint sustained by a player that result from a match or training, irrespective of the need for medical attention or time loss from activities" (Franco et al., 2021) and we can classify them as overuse injuries (without a specific and identifiable event associated with its onset) or acute/direct injuries.

Academically this classification is used to distinguish injuries based on the provoking injury mechanism which is very useful but it doesn´t give you any information about what really matters on a daily sports setting, which is the severity of the injury and more importantly, the prognosis for full recovery.

As in any other sector of industry, the specific task to be done is always associated with some risk for occupational injury, whether it's spending long periods sitting, climbing light lamps or playing any sports. And within sports the predictive reasoning can´t be the same, as an example, rugby is sport where most injuries occur in contact situations so the risk for any player to sustain an injury in a particular match from a contact situation is absolutely random.

In this paper we will focus on the industry of professional football, within which "injury rates are several orders of magnitude higher than those reported in other occupations" (Drawer, 2002), this makes this sport the perfect setting for developing and testing theories of health and safety management.

A few financial models have been proposed to provide a basis for the understanding of the cost-benefits analysis of injury prevention strategies in professional football, which shows the awareness of the several sports organizations on how an injury affects the clubs' playing and financial performances (Hickey, 2014).
On linear thinking one may conclude that lowering the incidence rate of injury and its severity, the players will be able to fully participate in training or match play and the time-loss will also be lower.

The best way to achieve this is to prevent the occurrence of an injury, and therefore the implementation of prevention training program is crucial to minimize the time-loss injuries factor.

Even though the focus of this paper is not injury prevention we cannot overlook it. Nowadays, prevention is more than identifying and bridge the intrinsic and extrinsic risk factors, and it's more of understanding the relationship of several entities that affect the player on a daily basis. That's why prevention comes hand-in-hand with prediction (Loose, 2018).

One cannot prevent what he doesn't know or sees, that's why it is so important to have an accurate and constantly updated report on the players injury risk.

McCall et al. (2007) conducted literature research to know if the terminology used by authors matched their statistical methods and ultimately their interpretation. It's quite interesting to see that only 35% had reported statistical predictive analyses. All the others had used statistical approaches pertaining to association.

Even though association is important to identify a population that may be susceptible to injury and to provide evidence to implement a physical screening or a monitoring tool, prediction allows us to classify at an individual level if an athlete will incur an injury (hopefully with a good level of accuracy).

On basic terms, association refers to the explanatory power for a population and prediction for the predictive power at an individual level.

If seen by the sports science and medical staff then probably it would be more useful to have evidence that could support implementation of injury-prevention strategies (strong association measures) rather than a high classification capacity, at the individual level.

When it comes to sports there are several in which we can use data science, but the major goal is to get that extra competitive advantage, either by flagging when will an injury occur, by estimating an athlete's performance or by identifying talented individuals (Bahr, 2016).

Several studies have been conducted to try to find the "holy grail" of sports analytics and Claudino et al. (2019) reviewed the literature and realised that 74% of the AI (artificial intelligence) studies were related to sporting performance whereas only 26% were related to injury risk.

More, the study concluded that machine learning methods most appropriate for injury risk assessment are the artificial neural network, decision tree classifier, and support vector machine, and that most research refers to soccer, basketball, American football, Australian football, and handball (Claudino et al., 2019).

So where do we stand right now when it comes to predicting injury in sports right now, well it's a generalised opinion that indeed we do have something to work on but we're definitely not there!

By the end of the day there are decisions to be made like ruling the athlete out of training or competition, but for now that decision will still reside in an human factor whether it's a coach or medical practitioner.

1.2. MOTIVATION

A major concern for the sports community is not only to predict when it is more likely for a player to get injured but also what injury will it be and what may be the cause of the injury (Matheson et al., 2010).

Concerning the "what to do" to avoid injury, that's what the medical and the sports science team do the best, so as data analysis expert, we are supposed to tell them when (more likely) will it happen, what injury will it be and how will it happen.

Most practitioners and other stakeholders who deal with high performance athletes aren't aware of the strength of the predictive ability of an outcome variable nor do they know which variables (or variables) are best to predict an injury.

When data analysts refer to an input as "variable", sports health practitioners call it an injury risk factor. Along the course of time several authors have tried to identify the injury predictive value of, both, intrinsic and extrinsic risk factors (Mandorino, 2023), however, nowadays it is clear that the most valuable insight arises from the interaction of all those factors at a specific time and in a specific place.

It is clear, at this present time and for all the community of health practitioners in sports, that for non-contact injuries there's a matrix of factors related to etiology of injury and that they're dynamic, meaning that they change over time. More specifically, the relationship between these determinants of injury is constantly changing and that makes it almost impossible to keep up to date with injury predisposition unless we can find a model that is time sensitive to every change, which in fact is exactly the definition of machine learning- gather new data, learn from it and provide new outcomes.

Having this in mind we strongly believe that machine learning is the path that can lead us to the understanding of non-contact injury prediction, following the proposed by Huang, 2022.

Although the field of injury prediction is still a bit blur and frightening for some there are several reasons why we believe this path should be done and, ultimately, we can tell that these are main motivations:

- Injury prevention: One of the primary goals of injury prediction is to prevent injuries and act proactively by addressing the web of factors that contributed to their occurrence.
- Cost savings: It can help reduce the cost associated with treatment, rehabilitation and time off from competition.
- Productivity increase: reducing the number of non-competition/training days injuries will increase athletes' productivity.
- Improve safety: improve overall safety of athletes.

- Compliance with FIFA regulations of injury occurrence: FIFA is very much concerned on how much injuries affects the players and therefore we hope this may help to enlighten the path to create better health and safety regulations.
- Scientific knowledge: contribute to the development of scientific knowledge by adding new insights on the relationships between causes and mechanisms of injury.

1.3. OBJECTIVES

- The main purpose of this study is to help clarify the role of machine learning techniques in sport injury management. To achieve this main goal, we proposed the following intermediate objectives:
- Identify the best injury risk prediction model that will enable us to classify the athlete on the possibility of having an injury or not.
- Run an empirical experience by applying the model to extract knowledge out of a dataset related to football players in a way that can be used on a practical setting for the medical and sports performance teams as well for the coaches, about the risk of injury.
- Identify the injury risk profile for these group of athletes.

1.4. IMPORTANCE AND RELEVANCE

As in many other areas of the society the amount of data gathered every day in sports organizations is increasing and the reasoning behind it all is that the more, we know then the better we will be. This means that as athletes and organizations become more professional the need for information is bigger, and when that happens, sports science are pushed to find as much data (mislead with information) as possible about players performance, well-being and health. Clearly, the abundance of data does not deliver the same value on information (Pilka, 2023)

That's one of the reasons why the analysis of this automated data is fast developing field in the sports community. Not only by helping people deal with the huge data they get but also by helping them realize which data is most valuable and what information lies behind all that.

Even though this is not the goal for this paper it is important to understand that as data sets get bigger, then the need for a careful analysis is critical to enhance knowledge in sports science and also help "decision making of the practitioners who work on the optimization of training and competition strategies".

More and more the world of sports is evolving as an industry and in fact nowadays there are a lot of financial interests that affect all the stakeholders, especially the athletes

One of the major stakeholders involved in the sports business are the insurance companies, and the emergence of data science in sports (as in other areas like education, manufacturing, social networking services, streaming services, healthcare, financial modeling and marketing (Claudino et al., 2019) will help this industry to make a much more evidence-based decision whenever it will be needed.

An injury is perceived differently depending on your role in the sports industry. The medical staff sees it as an health problem, the coach sees it as a team tactical problem, the club sees it as an labor problem and ultimately, the insurance company sees it has an financial problem. The responsibility for the costs of a player when injured is on the insurance company's side and more if the rehabilitation is not fully successful (Carneiro, 2017).

Depending on the stakeholder point of view, the approach for data analysis of the variables related to the injury has to be different. For example, sports science and medical staff are much more interested in associating the risk factors so that they can understand what they could´ve done and should do to avoid injuries.

On the athlete's perspective, safety is crucial to a long-lasting injury-free career and to a good physical performance, therefore any study that aims to clear out the complex dynamic web of variables affecting the athlete is of great relevance (Franco, 2021).

As for the medical staff the relevance of this kind of studies is related to a more comprehensive and adequate injury management (in which it includes preventive strategies) and to help sports medicine to advance and create more knowledge of the causes and mechanisms of injury in sports.

Sports physiologists are mostly concerned about optimization of the athlete's performance (and finally the team), improve training programs and prevent injuries, therefore any activity that reduces injury rates is very relevant for them. That way they can plan the best strategy to optimize athletes' performance and put them to play at their best.

Organizations which its core business is sports are dependent on the athlete's performance either as a team or as individual. Therefore, the costs of treating/rehabilitating an athlete and the cost of his/her absence for competition are of extreme importance for the financial balance of the organization/club.

No matter the stakeholder involved, the process of decision making is more reliable and safer if it is evidence-based.

Hopefully, the great relevance of this study is to promote evidence-based decision making across every field in the sports industry when it comes to dealing with injuries.

For the matter of this paper there are some concepts with which we will base our research:

- Clinicians should be aware of how risk factors may interact, rather than list several isolated risk factors, to plan effective preventive intervention.
- Risk profile may include non-linear interaction between risk factors from different scales, such as biomechanical, training characteristics, psychological

and physiological. Additionally, risk profile should be continuously assessed throughout preseason and in season.

- The recognition of the web of determinants in clinical practice might include risk factors that strongly influence the outcome and interact in many different ways with several variables.

By the end of this study, we will be able to:

- Understand if the data collected in athlete monitoring and tracking tools used in football clubs (as suggested by FIFA) is relevant to injury aetology.
- Have an algorithm that can be dynamic and adaptable to specific circumstances along the competitive season to classify the athlete.
- Create a reliable and objective injury risk profiling that serves as basis for the implementation of tailored prevention strategies.

2. LITERATURE REVIEW

In this chapter we aim to clarify and what are the latest concepts and theories on sports injuries and how does that affect the analytical approach of research.

Over the last years a lot has changed in the field of sports injuries research and practical intervention, mainly we shifted from a reductionist theory of injury etiology focused on identifying injury risk factors into a more complex approach where the interaction in the web of determinants results in either injury or athlete adaptation.

This shifting has clear effects on the way we look at injury prediction and injury prevention. Therefore, we will show some evidence on latest injury prediction approaches.

After understanding sports injuries rationale we dive into sports analytics, focusing our research on the machine learning methods that will be used in our study, classification and clustering.

2.1. SPORTS INJURIES

"Injuries occur when energy is transferred to the body in amounts or at rates that exceed the threshold for human tissue damage" Baker (cited by Meeuwisse et al, 2007).

When we think about mechanical models of injury, we can conceptualize that when there's a transfer of energy that exceeds the ability of the body's tissues or structures to maintain their function and normal structure, an injury occurs! (Elsafty, 2024).

Usually, in sport injuries, we refer to mechanical energy transfer either it is on a single unique moment or by a cumulative overload of energy. A recent consensus statement on injury definitions and data collection procedures in football (soccer) suggested that injuries are:

> ''Any physical complaint sustained by a player that results from a football match or football training, irrespective of the need for medical attention or time-loss from football activities. An injury that results in a player receiving medical attention is referred to as a ''medical attention'' injury, and an injury that results in a player being unable to take a full part in future football training or match play as a ''time loss'' injury.'' (Fuller, 2006).

In a simplistic manner we can say that you can subject somebody to a transfer of energy so high that no matter how well conditioned and no matter what he/she's done so far, they will become injured.

However, if we could lower the energy transfer just a few Newtons, then probably, a well-conditioned person won't be injured and another similar person who's unconditioned will in fact be injured. The magnitude of any physical strain is related to the ability of the body to sustain such stress.

Before we dive deeper into the study of injury prediction, we must understand the different types of injury and their classification as it is used worldwide for academical studies related to football injuries.

In 2006, the Federation Internationale de Football Association Medical Assessment and Research Centre (F-MARC) agreed to host an injury study group with the purpose of adopting a consensus statement on definitions and methodological issues related to studies of football injuries (Fuller, 2006).

The report form presented by FIFA arose from the consensus and is presented in Appendix 1, 2, 3 and 4.

The federation definitions for injury classification is presented in Table 2.1.

Alongside the definition of injury, they came up with other definitions such as recurrent injury, injury severity, match exposure and training exposure (Fuller, 2006). The federation definitions for injury classification is presented in Table 2.1.

Fuller (2006) also proposes that injuries should also be classified according to their location, type of injury, side of body and injury mechanism (see Appendix 2).

These last concepts are very important when studying sports injuries because they'll enable us to understand the stress imposed on an athlete during the season. Nowadays, with the high demand for information there´s also a high demand to gather as much data as possible in everywhere and at every moment. Therefore, most recent studies use not only the time exposure to training/matches measured in "Hours" but also other metrics measured by innovative technology such as GPS (Pilka, 2023).

	Definition
Recurrent Injury	An injury of the same type and at the same site as an index injury and which occurs after a player's return to full participation from the index injury
Injury Severity	The number of days that have elapsed from the date of injury to the date of the player's return to full participation in team training and availability for match selection.
Match exposure	Play between teams from different clubs.
Training exposure	Team based and individual physical activities under the control or guidance of the team's coaching or fitness staff that are aimed at maintaining or improving players' football skills or physical condition.

Table 2.1 - Definitions for Injury Classification (Fuller, 2006)

Basically, there's a consensus for academic purposes for the definition of injury, recidive (recurrent injury) and injury severity.

The definitions stated above give us an overall characterization of injury in football.

The mechanism of injury can either be traumatic or overuse, and as stated by Fuller (2006), traumatic injuries "result from a specific, identifiable event" whereas an overuse injury is caused by "repeated microtrauma without a single, identifiable event responsible for the injury".

The discussion in the scientific community is that prediction might be of great utility when determining likelihood of traumatic non-contact or overuse injuries, leaving out the traumatic contact injuries (kampakis, 2016). This last one results in a contact energy that exceeds the body's ability to embrace it and therefore the result is an injury, whose severity will be determined by the magnitude of the impact. The prediction of such an event is almost impossible due to the uncertainty of the contact, particularly in a sport such as football, which despite being a non-contact sport allows contact between players (Singh, 2020).

2.1.1. From Risk Factors to Risk Profiles

A recent literature review conducted by Murphy (cited by Bahr,2003) on the risk factors for lower extremity injuries showed that generally the understanding of injury causation is very limited, mainly due to the large numbers of risk factors implicated. His findings show that there is very little agreement with respect to the findings.

Traditionally risk factors are divided into two categories, intrinsic (internal or athlete-related) and extrinsic (external or environmental), however this is not enough to completely understand the causation of injury. Therefore, to properly address the matter of injuries one must know which of the risk factors are modifiable (e.g. strength, balance, flexibility) and non-modifiable (e.g. age, gender), and the mechanism by which they occur (Bahr, 2003).

Even though we may be able to fully comprehend risk factors and their interactions it's still not enough to completely understand the events that lead to an injury, or any other outcome (adaptation /maladaptation). The multifactorial nature of sports injuries requires a complex and dynamic model approach as the one proposed by Meeuwise (1994).

Meeuwise (1994), names the intrinsic factors as predisposing factors and the extrinsic factors as enabling factors, both of which aren't capable of producing injury only by themselves. Moreover, the mere interaction between them is what prepares the athlete to react to an injury-inciting event (Meeuwise, 1994).

The same author highlighted the importance of examining intrinsic predisposing factors as well as those extrinsic factors that interact to make an athlete susceptible to injury before an injury-inciting event occurs. It looks like a contradiction but, nowadays most injury prevention models are outdated and are still based on injury surveillance, identification of risk factors, and related implementation of prevention strategies (Meeuwise, 1994).

For this Meeuwise (1994) defends that the inciting event is the final link for the causation of injury, yet he admits that despite the mechanism of injury what matters the most is the pattern of events that lead to an injury situation. This information is more important than the exact description of the biomechanical of joint motion or muscle action at the point of injury. And this is where we shift from an analytical perspective of risk factors (and their interaction) and we start to analyze risk profiles (Ekstrand, 2009).

Many investigations conducted over the years led to the conclusion that "there are strong limitations to this linear approach, where individuals who are exposed or not exposed to some (risk) factors are followed forward in time to measure outcome of disease /injury" (Meeuwisse, 2007).

This innovative approach to preventive intervention relies on the identification of regularities or risk profile moving from risk factors to risk pattern recognition (Meeuwisse, 2007).

In complex systems, such as these proposed by Meeuwise (2007) the risk factors are known as units and their interactions are frequently unknown and usually there isn't any direct relationship with the outcome (and if it there is it's very weak), so the only way to understand system dynamics is by observing its regularities (risk profile).

The occurrence of certain regularities helps us to predict the occurrence of certain events, and this is why predictive methods have a great deal of importance. If we focus injury prediction uniquely on the risk factors and their interaction, we may never get any satisfying and realistic results. Instead, if we forget about risk factors and focus on risk profiles and pattern recognition then our specific knowledge of predictive methods might be extremely groundbreaking.

2.1.2. Dynamic, Recursive Model of Etiology in Sports Injury

Although we have been focusing our efforts on the causation of injury, we cannot forget that adaptations will (and should) occur within the context of sport, either preceded by injury or not, and that this alters the risk and affects aetology in a dynamic, recursive way (Bittencourt, 2016).

For every chain of events triggered by the athlete there may be two main outcomes: injury or adaptation. If the matrix of units and events is of such nature that causes an injury then the result will be the withdrawal from further exposure, recovery will facilitate return to sport participation and most likely with a new set of intrinsic risk factors (Meeuwisse, 2007).

A post-injury status means that this athlete has a new level of predisposition to injury, and it can either be good or bad, depending on the post-injury body limitations. Yet, the new predisposal status of the athlete may very well be better than it was and, in this case, we say that there has been an adaptation (Meeuwisse, 2007).

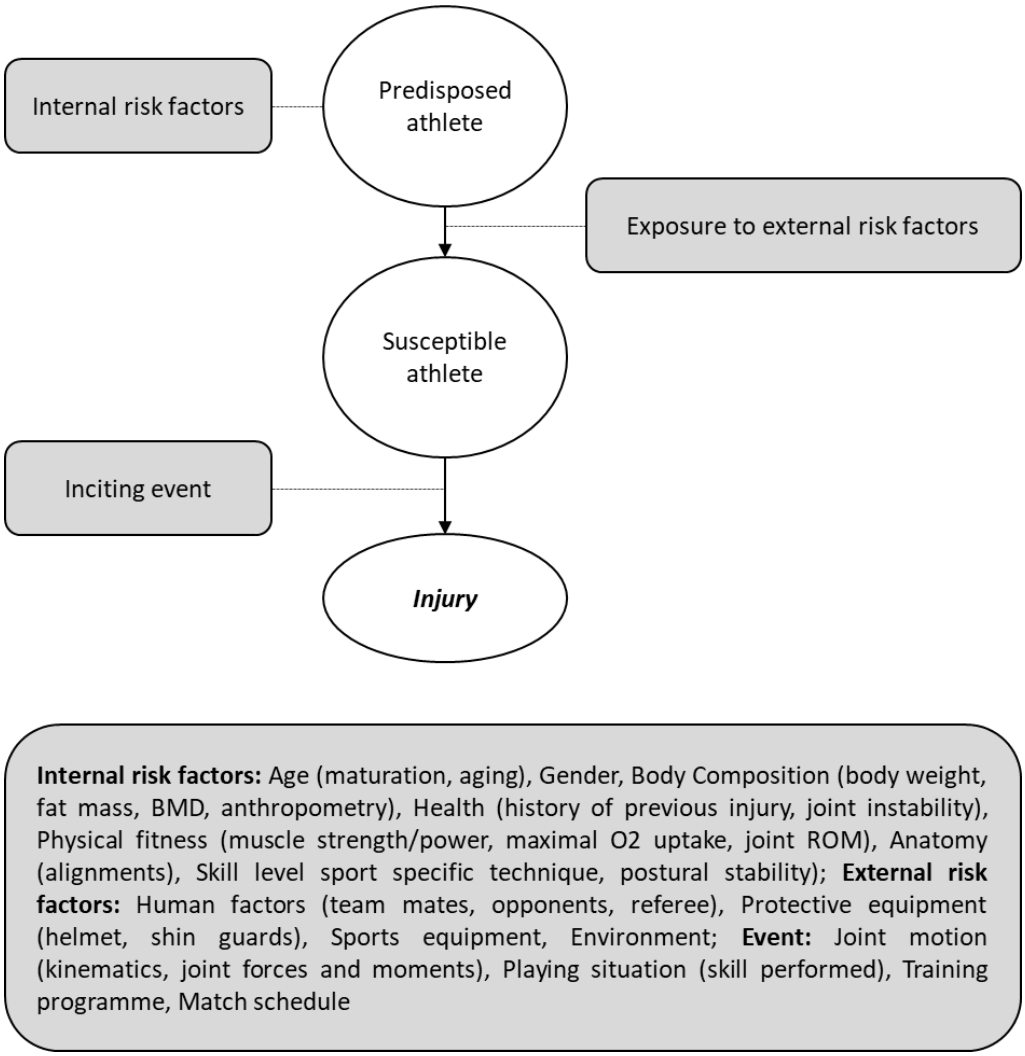


Figure 2.1 - Old Model for Injury Causation (Adapted from Meeuwise, 1993)

In this case the medical staff who is taking care of the athlete has a tremendous responsibility when designing the rehabilitation program, which will also be affected if the team has more information to base their intervention.

Full recovery means adaptation and a novel predisposed athlete, but it can also result in changing the equipment or any other extrinsic risk factors, which means that the athlete's susceptible status will also be different (Franco, 2021).

If the athlete is inadequately recovered or, by choice, no longer accepts the risk of sports (injuries are sport-specific in nature) we must consider them to have been removed from sports participation (Meeuwisse, 2007).

On the other hand, as we discussed earlier, another possibility is that even though the athlete is exposed to potentially harmful situations he/she will not sustain an injury.

It means that the tissues and/or equipment adapted/maladapted and most likely the athlete will be better prepared to deal with a new set of events. This potential effect may be immediate or latent. (Christopher, 2021, Baker, 1992).

If we are to truly understand the etiology of injury and target appropriate prevention strategies, we must look beyond the initial set of risk factors that are thought to precede an injury and take into consideration how those risk factors may have changed through preceding cycles of participation, whether associated with prior injury or not (Meeuwisse, 2007).

The model presented previously considers the implications of repeated exposure, whether such exposure produces adaptation, maladaptation, injury or complete/incomplete recovery from injury. (Meeuwisse, 2007).

What is not reflected in simpler models is that there may also be recurrent changes in susceptibility to injury in the course of sports participation without injury and that these exposures can produce adaptation and continually change risk, on other words an athlete's exposure to potential harmful event alters the athlete's intrinsic risk factor and consequently its predisposition to injury. (Bahr, 2016).

2.1.3. Web of Determinants

If we must choose a concept to characterize and distinguish this complex system approach from others it would be by the concept of "understanding of interactions".

In 1969 Von Bertalanffy had already defined a complex system as a whole, having units that interact with each other. Von Bertanfaly was cited by Bittencourt (2016) with the affirmation that these interactions are modified by the presence of new and unpredictable units during the process. Later on Phillippe and Mansi (1998) called the "web of determinants" to the interaction between these units, and defined these same units as determinants.

Philipe (1998) says that "these determinants are linked to each other in a nonlinear manner" and that any small change in only a few determinants can cause huge and sometimes unexpected consequences. Bittencourt (2016) considers as determinants all the units related to biomechanical, behavioral, physiological and psychological as characteristics of the athlete, and the stable relationships between them are considered as "regularities".

By extrapolating these concepts beyond the field of sports injury we can say that disease (specially chronic illnesses) are caused by a web of causality that extends through the risk margins (Gigerenzer, 2011).

2.1.4. Paradigm Shift: From Reductionism to Complexity

Fonseca (2016) has written extensively on the way to look at injuries through complex systems eyes and in one of his papers he divides systems into three kinds. The simplest are the linear systems which are the ones that we deal with everyday and that we solve very fast and without too much thinking.

The other kinds of systems are the complicated ones which are those that we spend more time trying to understand but in the end they're just the linear problems with many parts. Fonseca and Bittencourt

(2016) say that “a simple problem and a complicated problem are the same thing, but the main difference is based on quantity”.

What differs linear/complicated problems from complex problems? According to Fonseca (2020) complex problems are the ones in which the solution doesn’t lie in the units that comprise the problem, but it does in the recursive, looping relationships between them.

The main characteristic of this kind of system is that these relationships are continuously changing, either by creation of new relationships or by disruption of other ones. So, at different levels that we try to assess the problem new things come up and some things appear/disappear.

Galea (cited by Bittencourt, 2016) characterizes complex systems by means of constant “feedbacks, interrelations among agents and discontinuous non-linear relations” and if we transfer that definition to sports injuries it means that no matter the determinants (eg.: risk factors) what really contributes to injury is the continuous spontaneous relationship between them rather than the factors.

It means that the interacting units are frequently unknown and/or not directly observable, so there’s a degree of uncertainty and spontaneity in these kinds of systems that makes it impossible to establish any sort of causation relationship (Plsek, 2001).

Even though problems are made up of chaotic relationships between its units there is still one that is common to all levels of the problem which is, the nature of the problem. Basically, we just have to apply the correct method according to the problem that we are trying to solve, either it’s a linear or a complex system. In this topic resides the secret for success when applying machine learning processes to predict sports injury (Chmait, 2021).

There are certain characteristics of a complex system that helps us understand its dynamics and functioning: *Open system (equifinality), Inherent non-linearity (emergence), Recursive Loop (Feedback), Self-organization (risk profile), Uncertainty* (Bittencourt, 2016).

Complex systems are open which means that they fully interact with the environment and evolve over time, producing several pathways to similar outcomes (Rickles, 2007), and this is called equifinality (diverse ways, same outcome).

Inherent nonlinearity refers to the fact that inputs are not proportional to the output, which means that a certain change on the input doesn’t imply the same change of the output and this is the reason for abrupt changes in the system’s configuration (emergence). If we transfer this definition to sports injury prediction we can say that this principle relates to the fact that even though a variable may be very important it may not be directly related to the cause of injury (Rickles, 2007).

The recursive loop (feedback) of a complex system is the characteristic by which the output is reprocessed and becomes new input for the system (Meeuwisse, 2007).

Complex systems are considered to be self-organized mainly due to the spontaneous non-linear interactions created by its units, which results in the emergence of properties that can’t be predicted only by observing the behavior of the individual units alone. The emerging patterns are the product of the systems self-organization related to the creation of a new order (Lartey, 2020).

Although we cannot predict any outcomes nor any relationship between the system's units, by means of self-organization there are some specific configurations of these interacting units that produce observable regularities. The challenge with complex systems is to understand if the continuous observation of regular patterns is related to the same final emerging outcome (in sports we consider the outcome as injury or adaptation) (Lartey, 2020).

It's not easy to predict when these emergent phenomena will happen so we have to rely on the mere probability that they will indeed happen. This uncertainty is the result of self-organization and existence of nonlinear relationships within the behavior of complex systems (Lartey, 2020).

2.1.5. Complex Model for Sports Injury

According to what was mentioned in the previous chapters the main question is "Are we using the right tool to solve this specific problem?", in other words, it means that we must use the right method for analysis of the system that we want to study.

If we look at the characteristics of a complex system and compare sport injuries specific features then we do identify a tremendous of similarities like, the interaction between units (determinants), the emergent outcome of injury or adaptation, the recursive loop of adaptation or recurrence and the produced regularities. Therefore, to study sports injuries we must embrace their complexity and explore adequate methodological procedures.

In the world of athletes' health and performance, regularities are called as risk profiles (pattern of interaction), and emergent phenomena are called injuries/adaptation and web of determinants are the intrinsic/extrinsic risk factors as well as all the other variables that affect the subject to be studied.

If by means of machine learning methods we are able to recognize risk profiles that prompt an injury then we will be able to conduct specific preventive strategies (Bittencourt, 2016).

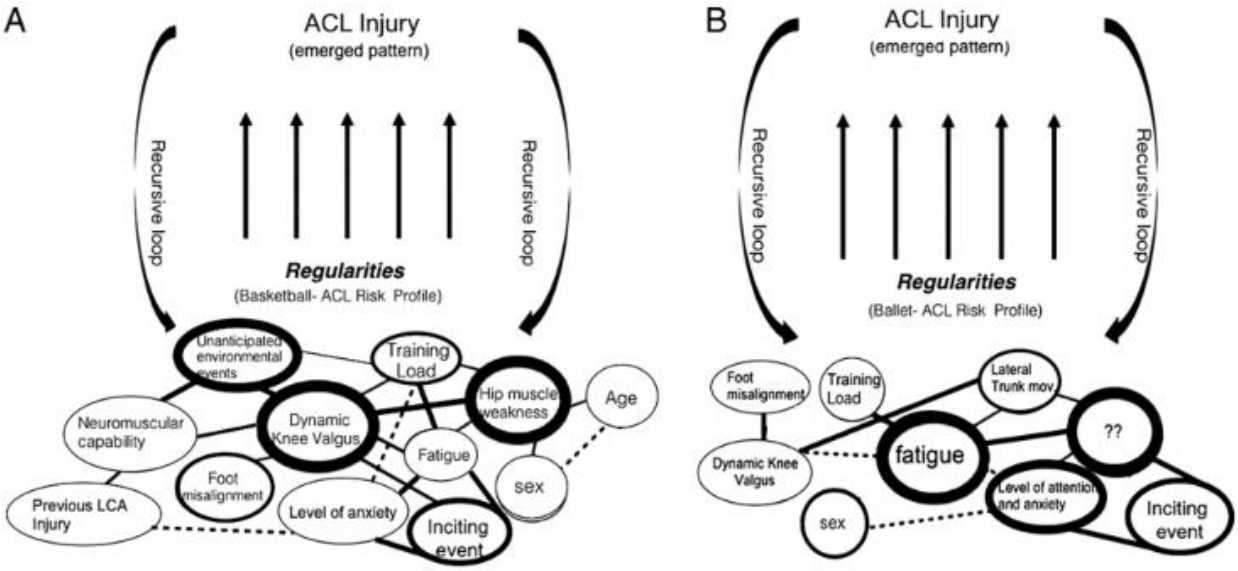


Figure 2.2 - (A) Web of Determinants for an ACL Injury in Basketball Athletes and (B) Web of Determinants for an ACL Injury in Ballet Dancer (Meeuwise, 2007).

The figure above shows us that risk factors interaction is far more important than the analysis of isolated risk factors, even though in this web of determinants we may find some that influence strongly the outcome than others (Bittencourt, 2016).

2.1.6. Injury Prediction

To predict an injury in a high-performance sport such as football, is a matter of great importance for everyone involved in the process making a winning team, and the focus will always be to have the best players healthy and available for every game and for every training session (Matheson, 2010).

On the internet of things the search term “injury prediction” leads us to what is the best predictor of injury, a previous injury! In a simplistic causal model such as the one described by Jovanovic (2017) this would make sense but what we now know about injuries is a completely different approach, and at should also be reflected on the statistical methods used to “predict” injuries. This is a discussion raised by Gigerenzer (2011), who says that in a world of complexity we are “pretty much confusing risk for uncertainty, using models that assume calculable probabilities”. Taleb (cited by Jovanovic, 2017) calls this “ludic fallacy”.

Sports injury prediction is more than just statistically calculate the likelihood of sustaining an injury, the complexity of the injury system is so deep that it makes it almost impossible (if not absolutely) to accurately predict an injury (McCall, 2017). According to Bittencourt (2016) we should rely on identifying risk profiles andembrace the complex nature of the sports injury by considering the multidirectional interaction between all factors.

This same author claims that to improve sports injuries prediction, as well as prevention, we should understand the philosophical paradigm of complex systems (both in movement and in injuries), identify the interaction among the phenomena of interest (sports injuries = emergent event, units or web of determinants) and choose a non-linear method of analysis.

2.1.7. Injury Prevention

In the world of high-performance sports such as football, the term “prevention” has been discussed and emphasized very often by physicians, physiotherapists and sports science teams. Matheson (2010) declared that over 3 decades we still haven’t made big progress and that the general knowledge on this is that prevention is not only repeat the same gesture over and over.

Matheson (2010) argue there wasn’t too much data gathered around the efficacy of prevention programs and there was an assumption that “sport injury rates were the same or greater than they were 3 decades ago”. Nowadays, prevention has been the focus of most governing bodies across different sports, as is the case of the IOC (International Olympic committee), World Rugby and FIFA (Ekstrand, 2009). These three institutions have been making great efforts to produce more knowledge about prevention in each specific sport. One curious fact that came out of all this is that, sport injury prevention research should not only be a matter of the sports science and medical team but it also should be thesubject of study for implementation of changes to the rules and regulations that govern sport (Ekstrand, 2009).

In 2006 FIFA developed a warm-up exercise program to help prevent and reduce the most common injuries in football, in the lower limb, and the conclusions were that teams who implemented these programs had a 30 to 70% reduction in number of injuries (Ekstrand, 2009). Studies also shown that there was increase in performance and motor skills. Even though theses results seem promising, they don't reflect the reality of professional football athletes in who is very important other variables such as training loading (Carey, 2017).

Having this in mind, the team working with the athlete should possess information on why that specific athlete may be at risk at any given situation and how injuries happen (injury mechanisms) (Bahr, 2003).

According to the complex systems approach, sports injury prevention relies on the identification of risk profiles, which means moving from analyzing risk factors to identifying risk pattern recognition. This way we should consider an "interconnected and multidirectional interaction between all factors, which embrace the complex nature of sports injury" (Bittencourt, 2016).

2.2. SPORTS ANALITICS

In the last decade, technological advancements have enabled the capture and collection of multiple types of data in the sport industry, specially in those whose business has a huge socio-economic impact, such has football, basketball, rugby and American football. In simplistic terms, is all about the application of data analysis and statistical techniques using sports-specific data to get insights and make an informed decision on different areas of the sports industry.

As in any other area of business, in sports there's data everywhere and that concerns all different kind of stakeholders, such as the teams directly involved with the athletes, club, media, bettors, and fans (Nevil, 2009).

Singh (2020) cites a couple of authors to claim that the data-driven approach to sports business and marketing is of immense value and data analytics is already an integral part of where clubs invest a lot of money. Not only by hiring specialized teams of data analysts but also investing in all sorts of data collecting tools, such as wearable devices, sensors, video and gathering of historical data.

For McCall (2017) when it comes to data analytics applied to sport science and medicine, there are three areas where decisions should be consistently data-driven: performance, scouting young talent and injury prediction/prevention. For the matter of this study we will focus on player health and injury analytics.

Data science has a tremendous impact on the sport science and medical approach to the athlete, either by tracking its physical condition and/or profiling the risk of injury. This makes it possible to apply preventive strategies, modify training loads and design recovery plans to maintain athletes at their top performance and minimize time away from the field (Wasserman, 2018; Singh, 2020; McCall, 2017).

Rein (2016) claimed "exciting times are emerging for team sports performance analysis as more and more data is going to become available" so the demand for stronger multi-disciplinary teams will increase, as "performance analysts, exercise scientists, biomechanists and medical staff will have to work together to make sense of these complex data sets" (Rein, 2016).

Later in 2020, Singh reviewed some literature and found that for fan-base marketing, consumer sentiments, player bidding, sport injury, player performance, promotions, bidding for games and others, a variety of analytical methods have been used, which include multivariate regression, descriptive analysis, optimization and even machine-learning methods (logistic regression, support vector machines, random forest, etc.).

2.2.1. Concepts

The feedstock of every machine learning algorithm is the specific industry data about its individuals, processes, events, transactions and so on. As in any other type of industry the availability of data is crucial for the construction of a real-world machine learning model (Sarker, 2021). There are several types of data, structured, semi-structured, unstructured and the metadata (Han, 2011).

Metadata is "data about data", and it can not be measured, nor it classifies / describes the properties of the subject. The importance of metadata is that it describes relevant information about the data itself making it more relevant and understandable for users (Chapman, 2000).

When the data has a well-defined structure, is highly organized and follows a standard order we call it structured data. Usually, it is stored in tabular format such as in relational databases in which you have a clear relationships between the rows and columns, like in Excel files or SQL databases. Some common examples of structured data are the names, dates, card numbers, stock information and addresses (Sarker, 2021). The data source for this kind of data are online forms, GPS sensors, network logs and OLTP systems (McCallum cited by Sarker, 2021).

In unstructured data there is no organization or specific format which means that it cannot be organized in relational databases and therefore it cannot be used by conventional data tools and methods (McCallum, 2005). In a simple manner we can say that unstructured data is the qualitative data captured in a form of text, audio or video files.

Semistrutured data relates to when data is not stored in relational databases but there is some form of organization that makes it easier to analyze, such as in HTML, XML and NoSQL databases.

As Sarker (2021) clearly points out, we can use different types of machine learning techniques according to their learning capabilities and the specific properties of the data set. Nowadays, in sports there is a lot of technology and records being used to collect data about those modifiable and non-modifiable risk factors and they range from medical reports, metrics derived from wearable technologies, self-reported data, test results and practitioner evaluations.

As much as possible this data needs to be of high quality and adding to that, if we really want to achieve a reliable injury prediction model than the data needs to be retrievable, accessible and continuously measured. The urgent need to be informed in real time also means that the metrics should be measured at the same rate which is not even near what we do right now. Yes, there are few wearable technologies that might be programmed to deliver information in real-time but, as we've seen before, there are a lot of other information about the functioning of the body (sleep, anxiety, etc) (Huang, 2021; Mason, 2023) that can only be measured by tests that can't be conducted in training or competition. If we are successful in measuring all of these variables, a number of issues arise, including

whether any data is redundant, if it fits the model, and whether it provides accurate information on the research topic.

Machine learning is an area of artificial intelligence that designs algorithms that allow computers to learn and enhance itself by means of finding statistical regularities or other patterns on data (Ayodele, 2021; Sarker, 2021; Susmita, 2019). The machine doesn't have to be continuously programmed and is able to make decisions, predict and forecast an event based on data. Any event or new data entering the system is processed and helps the algorithm to refine its output, or in other words it improves its performance as it gathers more information with no need for re-programing (Chmait, 2021).

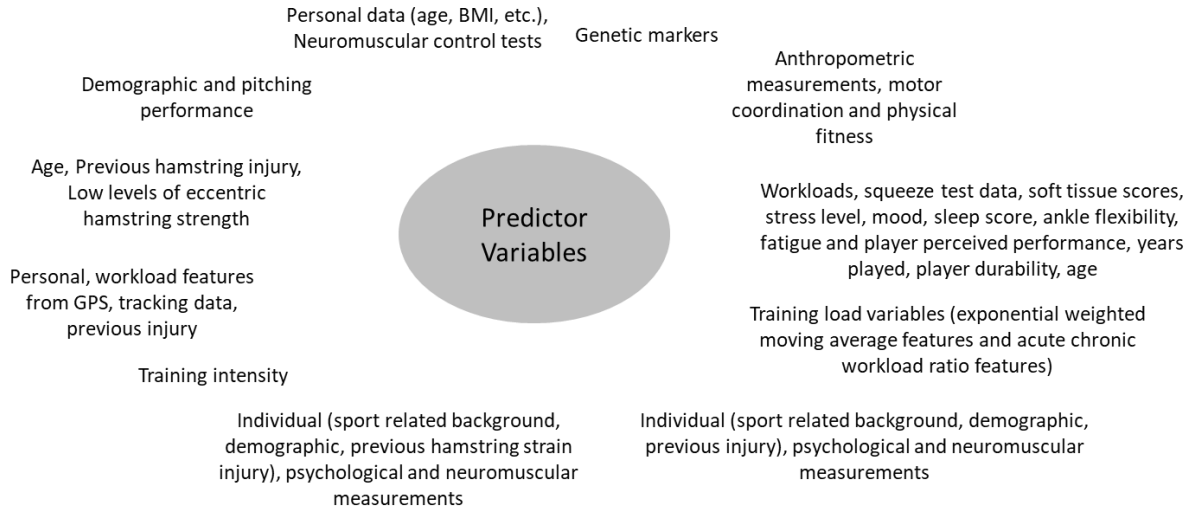


Figure 2.3 - Predictor Variables in the Included Studies of Eetvelde (2021) Literature Review

Nowadays machine learning is considered the most popular technology in the fourth industrial revolution which is characterized by the "ongoing automation of conventional manufacturing and industrial practices, including exploratory data processing using smart technologies" (Ślusarczyk cited by Sarker, 2021). The fields of application are very diversed such as in healthcare Lapham, 1995), sports, robotics, virtual personal assistant's computer games, marketing industries, data mining, traffic prediction, online transportation network, product recommendation, stock market prediction, online fraud prediction, agriculture, crime prediction, social media services, among others (Susmita, 2019).

There are several types of machine learning algorithms that can be applied to enhance the intelligence and the capabilities of an real-world application and each one of them must choosen according to the type of data that is collected by the industry (Susmita, 2019), which in this paper will be healthcare in sports. Mainly there are four categories: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

The main characteristic of a supervised learning is that it uses labeled datasets to train the algorithm and the outcome is either a classification or a prediction. It's considered to be the a "task-driven approach" (Sarker, 2021) because it's used when certain goals are to be accomplished from a set of inputs. Feeding the model with new input data makes it more fit to the goal purposed. It uses a training dataset that include inputs and related outputs and this relationship is being continuously checked by

the algorithm every time a new data enters the data set. Depending on the desired outcome supervised learning can be divided into two types: classification and regression.

Classification uses discrete variables and is considered to separate the data as it uses an algorithm to classify test data in specific categories, and the most common algorithms used are the linear classifiers, support vector machines (SVM), k-nearest neighbor (K-NN), random forest and decision trees. Regression uses continuous variables and is the type of learning that fits the data, which means that it stablishes the relationship between dependent and independent variables, in other words, to make predictions. The most popular regression models are linear regression and logistical regression. (Decker et al., 2023).

Unsupervised learning algorithms are created to extract structure from data samples, according to Ghahramani (2015), while Ayodele (2021) mentions that this method is all about getting the computer to "learn how to do something that we don't tell it how to, generally, it can be done by means of clustering or association".

In clustering, the training data is analyzed to create clusters of instances that are the most similar, and each time a new example enters the training data, it is allocated to one of the clusters. When there is sufficient data to construct clusters, as there would be in a real-world scenario of social information filtering, this strategy works quite well for (Ayodele, 2021).

By grouping data based on similar features and not on the output it is possible to learn something about the raw data that probably wouldn't be able to see otherwise (Iqbal, 2022).

The third type of machine learning – semi-supervised learning, is an hybridization (Mohamed and Sarker cited by Iqbal, 2021) of the two mentioned above because it deals with both labeled and unlabeled data. It is useful in real world contexts where there's numerous unlabeled data to add to the labeled data recorded by the company. Adding more data to a prediction model will obviously make it much more accurate, but the problem is that not all data is labeled and so it is necessary to use algorithms that may operate with non-labeled data (Han, 2011). According to Sarker (2021) "the ultimate goal is to provide a better outcome for prediction than that produced by using the labeled data alone" and this is done by treating new instances differently whether they're labeled or unlabeled.

Reinforcement learning is called the environment-driven approach (Iqbal, 2012) because the learning agents (software and machines) can perceive and interpret its environment, and therefore evaluate optimal behavior in a particular context (Sarker, 2020). Its core principles revolve around rewarding desired behavior and/or penalizing undesirable behavior by giving positive or negative values to the actions performed in a particular setting (Han, 2011; Iqbal, 2021). Mohammed (2016) claims that reinforcement learning is an effective method for enhancing the performance of AI systems used in a variety of applications, including robotics, gaming, autonomous vehicles, manufacturing, and supply chain logistics, among others (Mahdavinejad, 2018).

Various machine learning techniques may significantly contribute to the development of effective models in a variety of application areas, choosing the best one depends on the characteristics of the data and the desired output. Before the system can support intelligent decision-making, sophisticated

learning algorithms must first be taught using the real-world data and information relating to the target application that have been gathered (Iqbal, 2021).

Claudino et al. (2019) conducted a literature review in which they found that the sports that most make use of AI applications were soccer, basketball, handball and volleyball, and more interesting than that was that for purposes of predicting injury risk prediction and sporting performance the mostly used methods were artificial neural network, decision tree classifier, SVM and Markov process.

Until now almost every paper written on injury risk prediction only analyse both modifiable and non-modifiable risk factors (Eetvelde, 2021) in a very linear way and not so much as a complex mechanism. According to Eetvelde (2021) those papers who used ML algorithms such as decision trees (Ayala, 2019; López-Valenciano, 2018; Oliver, 2020; Thornton, 2017), Gini coefficient (Rommers, 2020) and SHAP summary plot (Majundar, 2022) there wasn't much consistency on the finding of the different studies because the variety of predictors was very wide. Besides that limitation, there were some features considered to have twice more impact on injuries, which were previous injury (López-Valenciano, 2020; Rossi, 2018, Ruddy, 2018), higher training load (Rossi, 2018; Thornton, 2017), higher body size in youth players (Oliver, 2020; Rommers, 2020).

2.2.2. Classifiers

In Machine learning, a classifier is an algorithm that assigns a class to a data point or observation. Although there are a lot of algorithms and its adaptations, choosing one of them depends on the type of problem, number of variables and, moreover, depends on the type of data (Rossi, 2018).

According to Mahesh (2019) the most commonly used classifiers are the decision Trees, Support Vector Machine (SVM) and Naïve Bayes.

Decision Tree classifier uses a tree-like graph to represent choices and their results, in which the leaves represent class labels and the branches represent the conjunction of all the decisions and sub-categories that lead to those class labels. The classification of data gets finer as the tree branches get bigger and this allows organic categorization. (Costa, 2021).

The SVM can be used to solve classification or regression problems, but it is mostly used for matter of classification. SVM categorizes data points by placing them on either side of a boundary line, known as the hyperplane. Using the training data, the algorithm creates the hyperplane, by means of some linear algebra calculations. It has been used in many applications specially in the early disease detection field (Acir et al., 2006 cited by Rejani, 2009)

The Naïve Bayes algorithm uses the probability of a specific data point to fall into one or more group of categories by using the Bayes Theorem (Mahesh, 2019).

The K-nearest Neihgbor (KNN) is a simple but very powerful classification algorithm that classifies instances based on a similarity measure. It can also be called as memory-based classification or lazy learning because it needs training examples to be present and that delays the run-time of learning (Cunningham et al., 2007).

Logistic Regression (LR) is a multivariate method for statistical analysis very common in the general health literature (Tetrault, Sauler, Wells, & Concato, 2008 cited by Park, 2013), along with linear

regression, discriminant analysis and proportional hazard regression. The difference between them is that for LR the outcome variable is binomial (Sperandei, 2013). This analyzes the relationship bet the independent variables and predicts the binary value of the dependent variable (Park, 2013).

For Arif Ali (2023) the XGBoost is the "superior machine learning algorithm in terms of prediction accuracy, interpretability, and classification versatility". Its power comes from the Gradient Boost, which is a concept that considers iterative improvements of poor learning methods and therefore reinforcement of a single weak model (Natekin, 2013). The XGBoost is an ensemble algorithm that uses parallel decisions trees boosting making them more capable of quickly and accurately solving a large range of data science problems (Zhang et al., 2022).

2.2.3. Classifier Evolution

According to Sweeney et al. (2022) the performance of a classification machine learning algorithm is commonly assessed by using four metrics: accuracy, precision, recall and F1.

The analysis of these metrics about the performance of each model provides insights "such as the trade-off between precision and recall, the ability to handle imbalanced datasets, and the ability to classify samples correctly" (Hernandez, 2020).

The Confusion Matrix is a tool used in ML to assess the performance of a classification model. It can be defined as a summarized table of the correct and incorrect prediction yielded by a classifier for binary classification tasks (Karimi, 2021). This table is in a form of a square matrix where the rows represent the predicted values given by the algorithm and the columns represent the real values. The number of correct and incorrect models predictions is displayed by means of a table and it can assume 4 values (Karimi, 2021).

- True positives (TP): occur when the model correctly predicts a positive data point;
- True negatives (TN): occur when the model correctly predicts a negative data point;
- False positives (FP): occur when the model incorrectly predicts a positive data point;
- False negatives (FN): occur when the modelincorrectly predicts a negative data point.

Accuracy is a machine learning metric very useful when classes are balanced, and it refers to the probability of success in identifying the right class of an observation (Zhongguo, 2017). Wu (2007) considered it the ratio between correct predictions and total number of predictions made. It is considered a correct prediction of all the true positives and the true negative observations. Incorrect predictions are all observations mistakenly observed, false positives and false negatives Boyd, 2012). The accuracy formula is (Wu, 2007):

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Equation 2.1 - Accuracy (Wu, 2007).

The measure ***Precision*** refers to the "exactness of a classifier" (Sweeney, 2022) and is also known in biomedicine as Positive predictive value (Sasaki, 2007). Lower precision values mean many false positives, and higher precision values indicate less false positives. In other words, precision indicates us how many of the predictive positive were actually positive (Wu, 2007).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Equation 2.2 - Precision (Wu, 2007).

A classifier **Recall** is about its sensitivity or completeness (Sweeney, 2022), and it refers to how many of the real positive observations were correctly predicted. A high recall score indicates less false negatives, and lower scores indicate more false negatives.

$$Recall = \frac{True\ Positive}{True\ Positive + True\ Negative}$$

Equation 2.3 - Recall (Wu, 2007).

F-measure (F1) is a score, ranging from 0 to 1, calculated by weighing the mean of precision and recall. For F1, 1 is the best score possible, and because it is a weighted score it means that both metrics have to be high in order to obtain a high value (Powers, 2011). In other words, if one of the metrics for precision or recall is low, than the F1 would never be close to 1. The F1 formula is this:

$$F1 = 2\ x \frac{precision\ x\ recall}{precision + recall}$$

Equation 2.4 - F-measure (F1) (Wu, 2007).

The Matthews correlation coefficient (MCC) is an alternative measure whose score doesn't get affected by the imbalanced datasets issue (Chicco, 2023). In statistic terms, the MCC is a "contingency matrix method of calculating the Pearson product moment correlation coefficient" (Powers, 2011) between actual and predicted values, and its scores range from -1 (perfect misclassification) to 1 (perfect classification).

For binary classifications, Jurman (2012) and Chicco (2017 and 2023) admit that MCC is the only rate capable of understanding if a model can correctly predict the majority of the positive and negative instances. In other words, a high score in the MCC is only possible if the model does well on both the negative and the positive elements.

By looking at its formula we can see that it considers similar proportion of each class of the confusion matrix (Chicco, 2020):

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)x(TP + FN)x(TN + FP)x(TN + FN)}}$$

Equation 2.5 - Matthews Correlation Coefficient (Chicco, 2020).

The receiver operator characteristic (ROC) curve is a common metric used since the early 70´s to evaluate binary classification methods, specially in the field of medicine, biostatistics and health care (Chicco, 2024). The ROC uses the True Positive Rate (sensitivity or recall) on the *Y* axis and False Positive Rate on the *X* axis, and the resulting curve is obtained by graphing the performance of a classification model at different thresholds. If the curve is close to the upper left corner of the graph then the classifier is considered to have a perfect performance (Chicco, 2024).

2.2.4. Clustering

Clustering algorithms analyse the structural distribution of data and create rules for fitting new observations in groups with similar characteristics. It is a type of unsupervised learning and therefore it should be used whenever there is unlabelled data (data without defined categories or groups). The process of clustering implies dividing a dataset according to the clustering criteria without any prior knowledge about the dataset (Ahmed, 2020).

For Ahmed (2020) the ideal scenario would be to have each cluster consisting of similar data instances that are quite dissimilar from the instances in other clusters.

Among other machine learning methods, K-means has been listed among the top 10 clustering algorithms for data analysis (Wu, 2008) consisting of analysing the distance between different data points and grouping them according to their proximity. Susmita (2019) and Theng (2020) say that K-means is one of the most frequently used methods due to its simplicity and ability to reach near-optimal solutions quickly.

This algorithm seeks to identify groups within the data, with the variable *K* refering the number of groups found. Using the given features, the algorithm iteratively assigns each data point to one of the *K* groups. The similarity of features is used to cluster data points/observations. The centroids of the K clusters, which can be used to label fresh data, and labels for the training data—each data point is allocated to a particular cluster—are the outputs of the K-means clustering algorithm. A cluster's centroid is a set of characteristic values that together define the groups that are formed (Wu, 2020).

One of main limitations of the K-means algorithm is its random initialization of the centroids and consequently the unexpected convergence and other outlier effects (Ahmed, 2020). Modeling with different K values will surely produce different clusters with not so similar characteristics. Therefore, the k-means algorithm requires the calculation of the value of K beforehand.

One method to identify the optimal number of clusters for the dataset is the creation of an elbow plot. The elbow method is graphical way to identify the best value for K. It consists of training the clustering model for a few numbers of K (usually 10) and attributing a distortion/inertia score to all of them. The point where the curve inflects, known as the elbow point, indicates the optimal value for K (Horvat, 2021).

Both distortion and inertia are distance related metrics computed in the elbow method for each one of the K and their score defines the elbow shaped graph.

Bellinger (2023) states that the results of K-means clustering are “sensitive to the initial cluster centres generated during the initialisation step, which is why the algorithm is usually run several times”.

3. METHODOLOGY

The methodological aproach we will be using to conduct this research will allow us to evaluate the validity and reliability of this study, and more importantly will allow us to provide insight on how data was collected/generated, how was it analyzed and what conclusions can we take out of it.

Studies on preventing sports injuries should, whenever possible, use a methodology and analysis approach that takes into consideration the cyclical nature of shifting risk factors to produce a dynamic, recursive picture of etiology. (Meeuwise, 2007).

Galea (2008) says that modeling dynamic systems should involve "computer-based algorithms that capture dynamic interactions among different variables, with feedback loops".

3.1. OVERVIEW

To guide this study from scratch we opted to divide it into three phases, the exploration, analytical and conclusive phases. Already having in mind that the analytical phase would correspond to the data mining project itself, and therefore it would be guided by the CRISP-DM reference model. As seen in Figure 3.1 step 3, 4 and 5 comprehend all the six phases previously described.

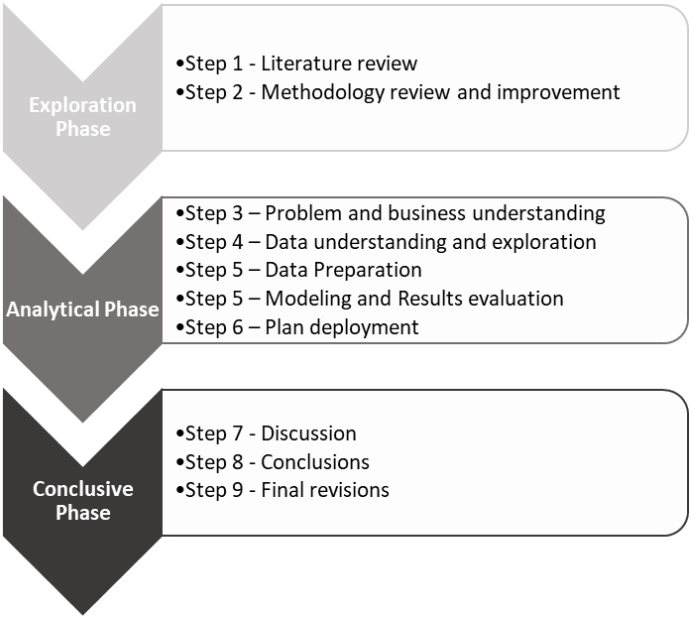


Figure 3.1 – Guide to This Study

Due to the academic nature of this study, the deployment phase won't be possible, yet we will describe some suggestions on how to implement these processes in a sports team.

For the Analytical Phase of this study we used used CRISP-DM process model to guide us through. which we correspond to the "Data Analysis" chapter, it used CRISP-DM process model. CRISP-DM

(acronym for Cross Industry Standard Process for Data Mining), a process model created in 1996 by a consortium of leading data mining users and suppliers, namely DaimlerChrysler AG, SPSS, NCR, and OHRA. Their main goal was to create a "non-proprietary and freely available" framework to address data mining issues. Its success arises due to the input from a wide range of practitioners and others such as data warehouses vendors and management consultancies, which made it "practical, real-world experience of how people conduct data mining projects" (Chapman, 2000).

One of the main limitations of any data mining project at that time was that the success/failure was very dependent of the person or team who was carrying it out and if it became successful couldn´t be repeated across the enterprise (Chapman, 2020). The industry needed a standardized approach that could help turn "business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience." (Wirth, 2000).

We decided to use this specific framework model due to three reasons:

- The methodology we used to guide our research included a phase of exploration which, basically, was all about understanding the problem of sports injury and how it affects the teams, and that is one of the main principles behind CRISP-DM framework and why it makes it so used in this sort of works;
- The CRISP-DM can be executed in a non-strict manner so we can move back and forth between phases in order to allow the recursive characteristics of a complex system of sports injury.

To approach the problem of sports injury prediction we will use computer based algorithms in the field of data mining, and as KDNuggets (cited by Laureano, 2014) shows, the CRISP-DM approach is the one used by 42% of the conducted research in this same area

The occurrence of an injury in an elite football athlete is a matter of great importance either for the athlete, the team, the sports science team, the medicine team, the club, the insurance company and the also for the fans. Every time an injury occurs an athlete will be out of the team and, depending on his/her skills the team may suffer great loss on its game results, and if so, will also suffer substantial economic draw down (Claudino et al, 2019).

Nowadays one of the main goals of any sports science and medicine team is to adopt a strategy to predict injuries, more specifically, how it will happen, why it will happen and more importantly when will it happen (Majundar, 2022). As demonstrated previously, it is becoming commonly known that the clear prediction of injury is absolutely impossible, yet, it is not impossible to identify when an athlete is at risk of suffering an injury.

The problem we will try put some light on is the knowledge on how sports science and medicine teams can identify if a football player is at risk of sustaining an injury (Owen, 2024).

3.2. CRISP-DM

In a CRISP-DM methodology is considered a life cycle of six phases for any data mining Project and according to Laureano (2014) there are some characteristics that should serve as a rule of thumb:

- The sequence of phases is flexible;
- It requires to move back and forth between phases;
- The task performed next is determined by the outcome of the previous phase;
- There is a cyclical nature of all the phases;
- The phase of deployment is not the end of the process, and it should trigger new business questions.

According to Chapman (2000) in the Consortium literature, the six phases of any CRISP-DM model are the “Business understanding”, “Data understanding”, “Data preparation”, “Modeling”, “Evaluation” and “Deployment”.

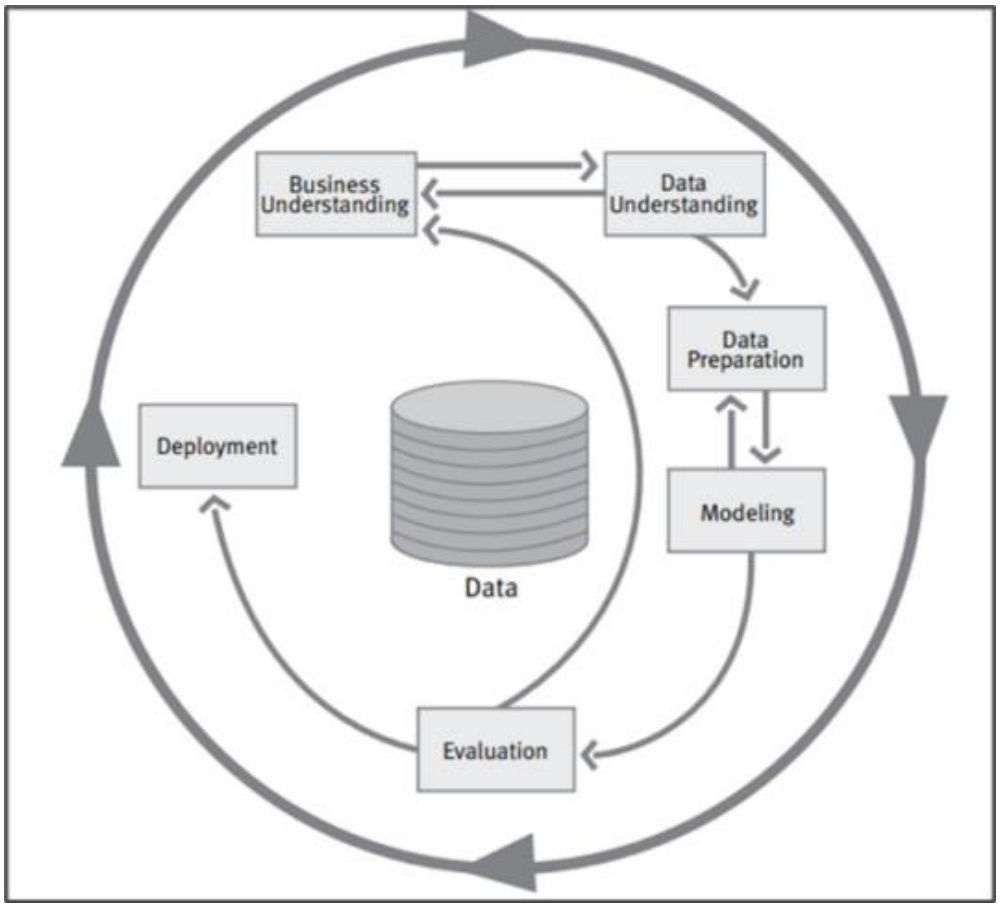


Figure 3.2 - Phases of the CRISP-DM reference model (Chapman, 2000).

3.2.1. Business Understanding

In the first phase we should identify the problem to be solved and specify in terms of context, determination of the data mining goal, data mining type (eg.: classification) and success criteria (Schröer et al., 2021).

3.2.2. Data Understanding

Understanding data involves its collection, exploration, description and checking its quality, or in other words "to become familiar with the data, identify data quality problems and discover first insights" (Chapman, 2000).

3.2.3. Data Preparation

It includes all activities needed to create the final dataset that will feed the modeling tool (Chapman, 2020). To ensure complete data preparation it should include the data selection, data cleaning and enrichment, and other data integration (which is not the case in this study).

3.2.4. Modeling

After the tasks around data, it's the phase where the most appropriate model of data mining is defined and applied, such as decision trees, cluster analysis, whatever it may be suitable to the solve the problem and to the dataset of the process. All techniques can be used but the most important is, how to explain the choice (Schröer et al., 2021).

3.2.5. Evaluation

After building the model and before deploying it is important to check the results against the defined objectives and evaluate thoroughly all the process in order to define further actions (Schröer et al., 2021).

3.2.6. Deployment

The creation of the most fitted model is not the end of the process, all of the tasks previously done and the knowledge gathered most be presented in a way that can be used in a real life/time situation, basically it has to create value for the decision making process (Schröer et al., 2021).

As Chapman (2000) described in the Consortium notebook, "deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise".

4. DATA ANALYSIS

This chapter corresponds to Steps 3, 4 and 5 of our Analytical Phase of research, which in a DM project are the first three steps of CRISP.

The data analysis comprehends not only the understanding of the data but also the problem and the domain in which we are working with.

4.1. BUSINESS UNDERSTANDING

In the previously, exploration phase we conducted a literature review on the subject we wanted to study which started by the searching these two keywords “injury” and “prediction”. We were far from knowing the philosophical trends behind what nowadays it’s called movement and sports injury, moreover, how does data science helps sports and health practitioners to improve their outcomes in high performance sports.

In 2015, David Joyce, a graduated physiotherapist expert in sport and conditioning released a book in which he claims that the best way to gain insight into future athletic performance or injury risk is to create a risk profile process based on generic warning index, specific warning index and determination of risk factors. Nowadays, the literature tells us that in fact we should look for an injury risk profile (Drawer, 2002), but we should also drop the terms risk factors for predictor variables. This is where it gets philosophical, new approaches to understand movement and sport injuries are based on the idea that both work as a complex system. For what injury concerns, the later literature reveals that there’s a web of determinants that’s affected by an inciting event, out of which it can result in injury or adaptation. This ever-changing interaction of all the web of determinants that produce different outcomes according to different behavior of the variables is what makes injury prediction a subject so hard to study and with so many different results.

One of the world’s great thinker on injury prediction is Fonseca and he claims that he changed his perspective completely after a few years of research, from strongly believe and trying to predict injury he now believes that it is impossible to predict a sports injury with enough detail to can objectively lead to the application of specific prevention strategies (Bittencourt, 2016).

The football team's medical staff carefully evaluated and documented every injury sustained during the competitive season, adhering to the guidelines provided by the Federation Internationale de Football Association (FIFA) - sponsored Medical Assessment and Research Centre (F-MARC) (see Appendix 2) who proposes a document to register injuries and their characteristics (Fuller, 2006. Fouasson-Chailloux, 2020).

4.2. DATA UNDERSTANDING

Before moving forward to prepare the data for modelling we should get acquainted with the dataset characteristics and became familiar with the terminology, values and their meaning. The potential insights derived from this step will be potentially very helpful in further steps.

The goal is to get a deep understanding of the dataset, clarify some specific domain measurement tools, identify variables relationships and assess the quality of the data.

4.2.1. Initial Data Collection

The database we will studying in this paper belongs to the club and it was buit to serve as a basis to Luis (2021) masters thesis on youth football injury epidemiology.

About data collection, the 226 records inserted in the dataset are from 184 male elite football players age ranging from 13 to 22 years old. Each input comprehends 40 specific features.

The data acquisition was conducted during the competitive season of 2019/2020, for eight months (July to March) 184 players were tracked for injuries. During the pre-season in July, all the players undertook full musculoskeletal screening tests according to the club's periodic health examination.

The research team has defined some exclusion criteria, as stated below (adapted from Luis, 2021):

- Trial athletes.
- Athletes that left the assessment period.
- Players who were injured at the time of the musculoskeletal.
- Athletes with no exposure or injury surveillance data recorded over the surveillance time and who did not want to participate in the study.

Besides the results derived from the screening tests, there were some other measurements for the bodyweight, countermovement jump, Passive Knee Fall Out, Active Dorsiflexion Lunge and the squeeze test, that were performed right before the study started. According to Luis (2021) the "participants were instructed about each test's protocol, and submaximal trials were used as familiarization and warm-up".

The research team decided to code each athlete with a whole number. There's a nuance to the coding of the athletes, every time an athlete got injured for the second time (recurrence or not) the input would be classifying the athlete´s whole number plus a letter, starting at b and so on.

For the ten attributes named below the research team used specific methods and devices:

- Data related to "Team", "Age", "Position" and "Dominant side" was collected in an interview.

- Height was measured by having the player standing straight against a wall with bare feet and align the top of the head to the tape measure mounted on the wall.
- Weight was measured by a digital scale with the player standing on it only wearing underwear and shorts.
- For the adductor squeeze test (ASQZ) the research team used a handheld dynamometer by Neuro Excellence, (Portugal). The measurement of maximal isometric hip adductor strength with this same portable dynamometer is dependable and with low values of error (Mesquita et al., cited by Moreno-Perez, 2019).
- The jump height collected in the counter movement Jump (CMJ) was measured by an optical measurement system consisting of a transmitting and a receiving bar with leds, OptoJump by Microgate (Italy). Glatthorn (2011) (Balsalobre-Fernandez, 2014) and concluded that "Opto jump photocell system demonstrated strong concurrent validity and excellent test-retest reliability for the estimation of vertical jump height.
- For the Passive knee Fall Out it was used a goniometer to identify the 90º of knee flexion and a measure tape to measure the distance of the fibula to the table.
- The Active dorsiflexion length was measured with a semirigid measure tape.

The football team's medical staff carefully evaluated and documented every injury sustained during the competitive season, adhering to the guidelines provided by the Federation Internationale de Football Association (FIFA) - sponsored Medical Assessment and Research Centre (F-MARC) (see Appendix 2) who proposes a document to register injuries and their characteristics.

4.3. DATA EXPLORATION

The dataset was presented in .xlsx format with 262 rows × 40 columns. It comprises 39 features, excluding the Key variable that codes each input and 262 inputs.

All the metadata related to the dataset is shown in the table 4.1.

The rules for coding each as the key variable ("ID_Código") were to attribute: - whole number to an athlete with no injury or only one injury; - whole *number* plus b means that the player had 2 injuries; - whole number plus c means that the player had 3 injuries; whole number plus d means that the player had 4 injuries.

#	Column	Dtype
0	ID_Código	object
1	Team	float64
2	Age	float64
3	Heigth	float64
4	Bodyweigth_1	float64
5	BMI	float64
6	Position	float64
7	Dominant Side	float64
8	Training_H	float64
9	Match_H	float64
10	Total_H	float64
11	Inj_Date	float64
12	Inj_Type	float64
13	Inj_Loss_days	float64
14	Inj_Severity	float64
15	Inj_Context	float64
16	Inj_Location	float64
17	Inj_Side	float64
18	Inj_Mechanism	float64
19	Inj_MuscleType	float64
20	Inj_LostTrainings	float64
21	Inj_LostMatches	float64
22	Inj_Contact	float64
23	Inj_Recidive	float64
24	Bodyweigth_2	float64
25	Squeeze_1	object
26	Squeeze_12	float64
27	Ratio_BW_Squeeze_1	float64
28	Ratio_BW_Squeeze_2	float64
29	CMJ_ESQ_1	object
30	CMJ_ESQ_2	object
31	CMJ_DTO_1	float64
32	CMJ_DTO_2	object
33	PKFO_ESQ_1	float64
34	PKFO_ESQ_2	float64
35	PKFO_DTO_1	float64
36	PKFO_DTO_2	float64
37	ADL_ESQ_1	float64
38	ADL_ESQ_2	float64
39	ADL_DTO_1	float64
40	ADL_DTO_2	float64

Figure 4.1 – Range Index in Jupyter Desktop

4.3.1. Team

The “Team” feature refers to the players are grouped by age. S23 stands for sub-23 (U23 in English) and it means that the players of that team are less than 23 years old. The same reasoning applies to the rest of the groups, meaning that S19 includes players under the age of 19 y/o and so on for the rest of the teams S17, S16, S15 and S14. The importance of this variable is not only related to the fact that they belong to the same age group but more to the fact that they have different coaches and different training methodologies. This study population respects the guidelines proposed by FIFA, there are more than one team of players, and the study lasted a period of one season (including preseason).

This variable has 226 entries cathegorized has 2,3,4,5,6 or 7 corresponding to the U23, U19, U17,U16,U15 and U14 teams, respectively.

4.3.2. Age

The "Age" attribute shows the current age of the player measured in years and can take any real whole number value, ranging from the youngest - 13 y/o to the oldest - 22 y/o.

There are 226 entries for this variable, with an average of 16 years old.

4.3.3. Height and Weight

Height (stature) and bodyweight (heaviness) are quantitative attributes that show anthropometric characteristics of a player, and they will be used to assess the Body Mass Index (BMI). The dataset has two different attributes for bodyweight. "Bodyweigth_1" was measured during pre-season at the tie of the screening tests by the clubs medical and sports team and "Bodyweigth_2" was measured by the time the study started.

There are 222 entries for height ranging from 137 to 196 cm with an average of 175cm.

Bodyweight variable has the same number of entries, with maximum weight of 91,3 and minimum of 34,1 (average of 63,5).

4.3.4. BMI

BMI stands for "Body Mass Index" and it is used to assess the relationship between weight and height. It's commonly used to categorise individuals into different classes of weight and is very useful to determine if a person has an healthy weight or not. BMI is calculated by using the formula: BMI = Weight (Kg) / Height (m2).

The BMI that characterizes our population was calculated with the values of first bodyweight measurement ("Bodyweight_1") and the height collected at the same moment. It has 220 entries, with an average of 20,67 Kg/m2, ranging from 30,57 and minimum 15,62.

4.3.5. Position

In a way to have some sort of characterization of the function of the player, the research team added the attribute "Position", which refers to the role and main area of operation on the pitch. With 226 entries this variable contains values of 0, 1, 2, 3, 4 and 5. Goalkeeper (GK) – code 0, is the one standing further in the back near to the goal. Centre-back – code 1 and wing – code 2 stand in the defense area right after the GK. Midfielders – code 3 and winger – code 4 stand in the middle on the pitch connecting defense and strike. The players who play closest to the opponents' goal are called forward – code 5.

4.3.6. Dominant Side

Another characteristic of the player is the dominant side, which can be described as the preference to use either the right or left side to perform a skilled or coordinated task. Of the 184 athletes, 52 are left-handed and 132 are right-handed.

There are 226 entries for this variable.

4.3.7. Training and Match Hours

One of the main variables used in injury prevalence studies is the exposure time (Calligeris, 2015), which is the time (in hours) players spend training ("Training_H") and on matches("Match_H") measured throughout the season. "Total_H" is the sum of the hours they spent on train and matches, is considered to be the total exposure time. This data was collected by the physical trainer of each team individually, all of them with 226 entries.

4.3.8. Squeeze Test and Countermovement Jump

For the matter of physical examination the research team chose the adductor squeeze test, the countermovement jump, the Passive knee Fall Out and the Active dorsiflexion lunge test. These were considered to provide the physical variables that would characterize the athlete. Both the ASQZ and the CMJ were used to measure peak force as it has been demonstrated its good intra and inter-tester reproducibility (Mesquita et al. cited by Moreno-Perez, 2019. Donahue, 2023)). The PKFO and ADL, were used to measure flexibility/mobility

The adductor squeeze test is used as a provocation test for hip/groin pain (Taylor et al., cited by Moreno-Perez, 2019) and screening tool to assess adductor muscle strength (O'Connor, 2023) and might help to identify the risk of groin injury. "Squeeze_1" means it was collected during pre-season for matter of screening physical tests and "Squeeze_2" means it was measured at the time the study was conducted.

To ensure reliability all participants performed the test following the same directions (as described by Moreno-Perez, 2019): 1) Lay supine; 2) Hips in neutral rotation position flexed to 45º, knees bent at 90ª; 3) Place dynamometer between the knees at the most prominent point of the medial femoral condyles; 4) Squeeze the cuff as hard as possible for 3 maximum contractions and held for 5 s with 3 min of passive recovery between contractions. The highest-pressure value displayed on the display was recorded for analysis.

For the first measurement of the squeeze test there were 215 entries and the second measurement had 223 entries.

CMJ stands for countermovement jump which can be used as a test or as an exercise (Glathorn, 2011). As test is used to assess the ability of the lower-body to generate force and power during a vertical jump by measuring jump height, power output and rate of force development (Balsalobre-Fernandez, 2014. Pérez Castilla, 2021). For this study the CMJ was adapted to a single leg, allowing arm swing and

no-pushup with the foot. Two measures of jump height were made. The terminology used indicates the test executed ("CMJ"), the side tested (left or right, "ESQ or "DTO") and if it is the first or second assessment. The directions to perform the CMJ were: 1) Start by standing with feet shoulder-width apart on a flat surface; 2) Transfer all the weight to a single leg; 3) Initiate jump with a quick downward motion by flexing the knees and hips to lower your body; 4) rapidly extend knees and hips (not the foot), pushing off the ground with maximum effort and jump as high possible (Petrigna, 2019).

The measurements of CMJ were made at the same moment as the Squeeze test so we have two measurements for each of the lower limbs.

There are 204 entries for the CMJ1 with left leg and 199 for the CMJ2. When testing the right leg, there are 205 and 199 entries for the CMJ1 and CMJ2 respectively.

4.3.9. Ratio_bw_squeeze

The ratio between the maximal isometric hip adductor strength and bodyweight ("Ratio_BW_Squeeze_1" and 2) was calculated for both measures and it refers to the normalization of the strength of the hip adductors by body mass.

The values were obtained by dividing the squeeze value for the bodyweight, for both moment of measurement, as shown below:

- Ratio_BW_Squeeze_1 = Squeeze_1 / Bodyweigth_1
- Ratio_BW_Squeeze_2 = Squeeze_2 / Bodyweigth_2

In both, Ratio_BW_Squeeze_1 and 2 we have 214 entries.

4.3.10. Passive Knee Fall Out and Active Dorsiflexion Lunge

The terminology PKFO stands for Passive Knee Fall out test, also named (as you can see in the literature) as Bent knee fall out Test is used to assess hip range of motion. It combines hip flexion, abduction and external rotation. This test is performed by having the patient laying supine with the knees bent to 90º. Then, the instructor passively moves the knees outwards while asking the patient to keep the soles of the feet together, and at the end of the movement applies a gentle pressure to ensure full range of motion in a relaxed position. The distance from the fibular head to the top of the table was then measured in centimeters with a rigid measuring tape (Van Klij, 2021. Terry, 2018).

Malliaras and Nevin (cited by Tak et al, 2015 and 2016) reported excellent reliability of the PKFO distance for assessing hip- and groin-related symptoms.

The PKFO test was realized in two distinct moments for both left and right lower limb, and the resulting entries were as follow:

- PKFO_ESQ_1: 213 entries
- PKFO_ESQ_2: 222entries
- PKFO_DTO_1: 213entries
- PKFO_DTO_2: 222entries

The active dorsiflexion lunge test (ADL) is a functional test used to assess ankle range of motion and was conducted with the patient standing on full weightbearing position with the tip of the toes and the kneecap equally touching the wall. Then the athlete was asked to, on one leg, slide the foot off the wall backwards while maintaining the knee in contact with the wall and the heel in contact with the floor. The result is then collected by measuring the maximum distance (in centimeters) between the tip of the toes and the wall, with a semirigid measuring tape (Terry, 2018).

Clanton (cited by Manoel et al, 2020) suggests that distances of more than 9 to 10 cm indicate the restriction of dorsiflexion and that the test is predictive of future football injuries (Halabchi, 2016).

As for the PKFO, the ADL test was realized in two distinct moments for both left and right lower limb:

- ADL_ESQ_1: 213 entries
- ADL_ESQ_2: 222 entries
- ADL_DTO_1: 214 entries
- ADL_DTO_2: 222 entries

4.3.11. Injury Related Variables

Now we will describe the variables related to the injury itself. For that we took in consideration the guidelines of FIFA as described by Fuller (2006), therefore we recorded as being an injury all the incidents that characterized as “any physical complaint sustained by a player that results from a football match or training, irrespective of the need for the medical attention or time loss from football activities”.

The feature “Inj_Type” (226 entries) refers to the type of injury sustained by the athlete which can be categorized has no injury (0) , fracture (1), other bone pathology (2), dislocation (3), ligament injury (4), meniscus injury (5), cartilage injury (6), muscle injury (7), tendon injury (8), contusion (9), abrasive injury (10), laceration wound (11), concussion (12), nerve trauma (13), synovitis (14), overload (15) and other type of injuries (16). Fuller (2006) recommends that if the number of injuries is too small “it may be necessary to combine individual type categories into the main groupings for analysis

purposes", which we did not because the percentage of records due to injury (57%, 129 out of 226 records) was higher than no-injury.

For "Inj_Date" the medical staff was instructed to record the day the injury took place as the day number after the first screening. Let's take an example to clarify input "20" has the attribute value of "3", which means that he sustained an injury three days after the season started. This feature has 129 entries corresponding to the number of injuries.

The number of days the player was out of training/match while injured was recorded by the medical staff as "Inj_Loss_days", counting the day of injury as day zero. The minimum days out was two which means that the player was out of train/match for two consecutive days. The average time of absence was 22 days.

The severity of the injury was defined as the extent of harm/damage caused by an injury and it was classified as Minimal (0), mild (1), moderate (2) and severe (3). For Fuller (2006) the severity of the injury should be defined as "The number of days that have elapsed from the date of injury to the date of the player's return to full participation in team training and availability for match selection", which in fact is what our research medical team defined as "Inj_Loss_Days". This nomenclature difference doesn't affect the purpose of this study.

Another item we should consider when addressing injuries is the context in which they happen ("Inj_Context"). The player can get injured while training in the club with his teammates (0), while playing a match against another team (1), while representing the national team (2) or in any other setting (3).

The "Inj_Location" refers to the exact place in the human body where the player sustained the injury. For this, the medical staff adopted the nomenclature proposed by Fuller (2006) and not the OSICS (Orchad Sports injury Classification System). It's important to note that all the categories are coded with an whole number: Face = 1; Cabeça ("head") = 2; Cervical = 3; ombro ("shoulder") = 4; cotovelo ("elbow") = 5; antebraço ("forearm") = 6; punho ("wrist") = 7; mãos/dedos ("hand/fingers") = 8; abdómen ("abdomen") = 9; Pélvis ("pelvis") = 10; lombar ("lumbar") = 11; Groin = 12; Anca ("hip") = 13; Coxa ("thigh") = 14; Joelho ("knee") = 15; Perna ("leg") = 16; Aquiles ("achilles") = 17; Tornozelo ("ankle") = 18; Pé/dedos ("foot/toes") = 19.
Another feature used to characterize injuries was the injury side ("Inj_Side"). The medical staff had to fill in the form by saying if the injury was on the dominant side (upper or lower limb) - 0, in the non-dominant side (upper or lower limb) - 1 or in the central part of the body (trunk and head) - 3.

The mechanism of injury which is identified in the data set as "Inj_Mechanism" refers to the specific actions or events that lead to the occurrence of an injury and usually is defined according to the sports in study.

In football, the categories are related to the most common biomechanical gestures (Elsafty, 2024). For the matter of the study the medical staff and physical trainer defined twenty-three actions that can be performed by the player while practicing football: Sprint = 1; Rodar ("turn") = 2; Rematar ("Kick") = 3; Passar/Cruzar ("Pass/cross") = 4; Driblar ("Dribble") = 5; Saltar ("Jump") = 6; Aterrar ("Land") = 7; Cair ("Fall down") = 8; Mergulhar ("Dive") = 9; Alongamento ("Stretching") = 10; Deslizar ("Slide") = 11;

Sobrecarga ("Overload") = 12; Compressão ("Compression") = 13; Atingido P/bola ("Hit by the ball") = 14; Colisão ("Collision") = 15; Cabecear ("Header") = 16; Interceptado ("Intercepted") = 17; Interceptar ("Interception") = 18; Pontapeado ("Kicked") = 19; Bloqueado ("Blocked") = 20; Uso de braço ("Arm usage") = 21; Acelerar ("Acceleration") = 22; Desacelerar ("Decceleration") = 23.

For all of the last injury related variables there are 129 entries which, as stated before, corresponds to the total amount of injuries.

Whenever there was a muscle injury (category 7 in the Inj_Type), the medical staff was instructed to add another level of information which was "Inj_MuscleType" to identify the injured muscle. To help specify and characterize the muscle injured there were created 6 categories: 1-Isquiotibiais (Hamstrings); 2-Adutores (Adductors); 3- Quadriceps; 4-Gémeo (Gastrocnemius); 5- Solear; 6-Rotadores Externos Anca (external rotators of the hip). The 43 entries in this feature correspond to the 43 muslce injuries registered.

The features "Inj_LostTrainings" and " Inj_LostMatches" are completed with the number of trainings and matches lost due to injury. All the entries on injuries were completed, 129.

The attribute "Inj_Contact" (129 entries) indicates if the injury sustained by the athlete was due to a direct trauma (contact injury) - 1 or indirect trauma (non-contact) - 0. The average was non-contact injuries.

The last attribute to characterize injury was "Inj_Recidive" (129 entries), which indicates if the injury is a recurrence of a previous consecutive injury - 1 or not - 0. On average the injuries are non-recidive.

4.3.12. Metadata

The table below shows the metadata of the dataset that will be studied.

Variable Name	*Type*	*Measurement Scale*	*Metric unit*	*Range or Categories*	*Source*
ID_Código					
Team	**Categorical Nominal**	Nominal Scale	**n.a**	**S23 = 2**; S19 = **3**; S17 = **4**; S16 = **5**; S15 = **6**; S14 = **7.**	Dataset
Age	**Numeric continuous**	Ratio scale in years	**whole number**	**Range: 13 to 22 years old.**	Dataset
Heigth	**Numeric continuous**	Ratio scale in centimeters (cm)	**whole number**	**Range: 137 to 196 cm**	Dataset
Bodyweigth_1	**Numeric continuous**	Ratio scale in Kilograms (Kg)	**Decimal number**	**Range: 34,1 to 91,3 Kg**	Dataset
Bodyweigth_2	**Numeric continuous**	Ratio scale in Kilograms (Kg)	**Decimal number**	**Range: 34 to 91 Kg**	Dataset

BMI	**Numeric continuous**	Ratio scale in Kg/m2	**Decimal number**	**Range: 15,62 to 30,57 Kg/m2**	Dataset
Position	**Categorical Nominal**	Nominal Scale	**n.a**	**GR ("Goalkeeper") = 0**; Defesa central ("Centre-back") = **1**; Defesa Lateral ("wing-back") = **2**; Médio ("Midfielder") = **3**; Extremo ("Winger") = **4**; Avançado ("Forward") = **5**.	Dataset
Dominant Side	**Binary**	Nominal Scale		**E ("Left")= 0**; D ("Right") = **1**	Dataset
Training_H	**Numeric continuous**	Interval scale in Hours	**Decimal number**	**Minimum = 0.00; maximum = 352**	Dataset
Match_H	**Numeric continuous**	Interval scale in Hours	**Decimal number**	**Minimum = 0.00; maximum = 42,25**	Dataset
Total_H	**Numeric continuous**	Interval scale in Hours	**Decimal number**	**Minimum = 0.00; maximum = 379,18**	Dataset
Squeeze_1	**Numeric continuous**	Ratio scale in Newtons (N)	**Whole number**	**Minimum = 13; maximum = 60**	Dataset
Squeeze_2	**Numeric continuous**	Ratio scale in Newtons (N)	**Whole number**	**Minimum = 19; maximum = 72**	Dataset
Ratio_BW_Squeeze_1	**Numeric continuous**	Ratio scale in Newtons per kilogram (N/kg)	**Decimal number**	**Minimum = 0.25; maximum = 0,85**	Dataset
Ratio_BW_Squeeze_2	**Numeric continuous**	Ratio scale in Newtons per kilogram (N/kg)	**Decimal number**	**Minimum = 0.29; maximum = 0,94**	Dataset
CMJ_ESQ_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 10; maximum = 28**	Dataset
CMJ_ESQ_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 10; maximum = 28**	Dataset
CMJ_DTO_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 8; maximum = 29**	Dataset
CMJ_DTO_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 7; maximum = 27**	Dataset

PKFO_ESQ_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 12; maximum = 44**	Dataset
PKFO_ESQ_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 12; maximum = 43**	Dataset
PKFO_DTO_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 13; maximum = 42**	Dataset
PKFO_DTO_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 13; maximum = 40**	Dataset
ADL_ESQ_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 1; maximum = 19**	Dataset
ADL_ESQ_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 2; maximum = 20**	Dataset
ADL_DTO_1	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 1; maximum = 20**	Dataset
ADL_DTO_2	**Numeric continuous**	Ratio scale in centimeters (cm)	**Whole number**	**Minimum = 1; maximum = 18**	Dataset
Inj_Type	**Categorical Nominal**	Nominal Scale	**n.a**	**Não Lesionado ("no injury") = 0**; Fratura ("fracture") = **1**; Outra Lesão óssea ("other bone pathology") = **2**; Luxação ("dislocation") = **3**; Ligamentar ("ligament injury") = **4**; Meniscal ("meniscus injury") = **5**; Cartilagem ("cartilage injury") = **6**; Lesão muscular ("muscle injury") = **7**; tendinosa ("tendon injury") = **8**; Contusão ("contusion") = **9**; Abrasão ("abrasive injury") = **10**; Laceração ("laceration wound") = **11**; Concussão ("concussion") = **12**; Nervo ("nerve trauma") = **13**; Sinovite ("synovitis") = **14**; Sobrecarga ("overload") = **15**; Outra Lesão ("other type") = **16**.	Dataset
Inj_Date	**Numeric continuous**	Interval scale	**whole number**	**Minimum = 3; maximum = 243**	Dataset
Inj_Loss_days	**Numeric continuous**	Interval scale	**whole number**	**Minimum = 2; maximum = 187**	Dataset
Inj_Severity	**Categorical Ordinal**	Ordinal scale with 4 levels	**n.a**	**Minimal = 0**; Mild = **1**; Moderate = **2**; Severe = **3.**	Dataset

Inj_Context	**Categorical Nominal**	Nominal Scale	**n.a**	**Treino ("Train") = 0**; Jogo ("Match") = **1**; Seleção ("national Team") = **2** Nacional; Outro ("Other") = **3**.	Dataset
Inj_Location	**Categorical Nominal**	Nominal Scale	**n.a**	**Face = 1**; Cabeça ("head") = **2**; Cervical = **3**; ombro ("shoulder") = **4**; cotovelo ("elbow") = **5**; antebraço ("forearm") = **6**; punho ("wrist") = **7**; mãos/dedos ("hand/fingers") = **8**; abdómen ("abdomen") = **9**; Pélvis ("pelvis") = **10**; lombar ("lumbar") = **11**; Groin = **12**; Anca ("hip") = **13**; Coxa ("thigh") = **14**; Joelho ("knee") = **15**; Perna ("leg") = **16**; Aquiles ("achilles") = **17**; Tornozelo ("ankle") = **18**; Pé/dedos ("foot/toes") = **19**.	Dataset
Inj_Side	**Categorical Nominal**	Nominal Scale	**n.a**	**Dominante ("Dominant side") = 0**; Não Dominante ("Non-dominant side") = **1**; Central = **3.**	Dataset
Inj_Mechanism	**Categorical Nominal**	Nominal Scale	**n.a**	**Sprint = 1**; Rodar ("turn") = **2**; Rematar ("Kick") = **3**; Passar/Cruzar ("Pass/cross") = **4**; Driblar ("Dribble") = **5**; Saltar ("Jump") = **6**; Aterrar ("Land") = **7**; Cair ("Fall down") = **8**; Mergulhar ("Dive") = **9**; Alongamento ("Stretching") = **10**; Deslizar ("Slide") = **11**; Sobrecarga ("Overload") = **12**; Compressão ("Compression") = **13**; Atingido P/bola ("Hit by the ball") = **14**; Colisão ("Collision") = 1**5**; Cabecear ("Header") = **16**; Interceptado ("Intercepted") = **17**; Interceptar ("Interception") = **18**; Pontapeado ("Kicked") = **19**; Bloqueado ("Blocked") = **20**; Uso de braço ("Arm usage") = **21**; Acelerar ("Acceleration") = **22**; Desacelerar ("Decceleration") = **23**.	Dataset

Inj_MuscleType	**Categorical Nominal**	Nominal Scale	**n.a**	**Isquiotibiais ("Hamstrings") = 1**; Adutores ("Adductors") = **2**; Quadriceps = **3**; Gémeo ("Gastrocnemius") = **4**; Solear ("Soleus") = **5**; Rotadores Externos Anca ("External rotators of the hip") = **6**	Dataset
Inj_LostTrainings	**Numeric continuous**	Interval scale	**whole number**	**Minimum = 1; maximum = 103**	Dataset
Inj_LostMatches	**Numeric continuous**	Interval scale	**whole number**	**Minimum = 0; maximum = 21**	Dataset
Inj_Contact	**Binary**	Nominal Scale	**n.a**	**Non Contact = 0**; Contact = **1**.	Dataset
Inj_Recidive	**Binary**	Nominal Scale	**n.a**	**N (No recidive) = 0**; Y (recidive) =**1**.	Dataset

Table 4.1 - Metadata

4.3.13. Dataset Analysis

The first analysis we did was on the demographic of the population of the study, elite youth football players.

The size of the sample corresponding to the U14 is almost the double of any other team with a mean age of 13,7 years old (SD = 4months).

The table below (table 4.2) summarises the demographic characteristics of the dataset.

Team	Players	Age (Y)	Height (cm)	Bodyweight (Kg)	BMI (Kg/m²)
	N	*Mean±SD*	*Mean±SD*	*Mean±SD*	*Mean±SD*
U23	25	19,9±1,3	181,4±7,6	73,4±7,8	22,3±1,6
U19	28	18,1±0,9	180,6±6,9	71,5±7,8	21,9±1,8
U17	25	16,7±0,5	179,4±7,3	68,2±6,0	21,2±1,3
U16	25	16,0±0,2	175,7±10,0	64,3±8,3	20,9±2,5
U15	32	14,9±0,2	173,7±8,33	61,1±7,5	20,2±1,8
U14	49	13,7±0,4	162,8±10,4	49,6±9,6	18,4±1.7

Table 4.2 - Demographic Characteristics (adapted from Luis, 2021)

In 2021, André Luis conducted an epidemiological study on workload and musculoskeletal risk factors in youth football using the same database as we did. Although this is not the goal for this study it is very valuable to understand the stress and strain players are exposed to as it also relates to injury.

According to the rules for youth football described in the regulations document of the Lisbon football Association (AFL, 2023), the U15 and the U14 play less 10 minutes than the other teams. This adjustment is necessary to comply with FIFA regulations for protecting the health of young football players. Therefore, the U23, U19, U17 and U16 play matches of 90 minutes divided in two parts of 45, separated by an break-time of 15 minutes.

By looking at the table below we realize that, on average, the team with higher training load was the U17 (241.7±52,1) and the one with the lowest average training times was the U14 (144,2±29,5). Coincidentally, U17's players were the most affected by injuries (72% out of 18), but on the other way they didn´t have any recurrence such as the U14's.

One interesting fact is that the standard deviation for the match load in the U14's is almost the same as the average hours played (8,3±7,9), this might be related to the lack of specialization of the players. At this age the players are on selection phase and there is not so much filter unlike other older teams. This may explain why there are players that didn't play any match and others that played full matches.

Total exposure time was an average of 211.04 minutes, and as it can be seen in the table 4.3, the team with the highest exposure time was the U17 and the U19. By looking at the standard deviation, we can hypothesize that these teams had great imbalances between the players of the squad. High SD values may tell us that there are players that play much more and others much less than the average.

Team	**Players**	**Training (H)**	**Match (H)**	**Exposure Time (H)**	Match Duration (H)
	N	*Mean±SD*	*Mean±SD*	*Mean±SD*	
U23	25	176,8±47,3	22,3±13.1	199,2 ±53,5	1,5
U19	28	234,8±69,5	17,9±13,2	252,8 ±74,9	1,5
U17	25	241.7±52,1	13,6±10,3	255,3 ±53,9	1,5
U16	25	175,6±66,5	16,1±7,4	191,7 ±69,6	1,5
U15	32	204,2±52,2	20,8±8,5	225,0 ±54,8	1,33
U14	49	144,2±29,5	8,3±7,9	152,6 ±31,3	1,33

Table 4.3 - Comparison of values for Training and Match load during the season for all ages (mean and standard deviation).

The player with the most exposure time (352,0 hours) is 15 years old and sustained two injuries, both on the knee of the dominant side (left), one was by direct contact and the other by non-contact (overload). Overall he lost 3 matches and 21 trainings.

On average the U17 players trained 161 times and played 9 matches, what puts them as the most exposed players. Coincidentally it's the team with most frequent overload injuries, so an hypothesis could be that total exposure time is a risk factor to be considered for overload injuries.

The U14 team has the players with least matches participation (Table 4.4). Two factors may contribute to this findings, one is the high number of players in the squad and the other is that the ain goal in this

ages is the education of the athlete and his development. Therefore, it's not about playing with the best to win but more about putting all the kids to play.

Team	Training (N)	Match (N)
	Mean	*Mean*
U23	118	15
U19	157	12
U17	161	9
U16	117	11
U15	154	16
U14	108	6

Table 4.4 - Values for the average number of trainings and matches during the season for all ages.

An analysis of the dataset shows that there was a total of 129 time-loss injuries that corresponded to a total of 2826 days of absence from training or any match participation. Of all theses injuries 23,5% occurred in club matches, 5,2% on national duty (training or match) and 71,3% occurred during training at the club. Although Luis (2021) found that "injuries occurring in match context presented higher incidence and burden".

The team with the most recidive injuries – 20% was the U15, this means that five players injured twice and one injured three times.

According to the side of the injury, out of the 129 injuries, 50,4% were on the dominant side, 37,2% were on the non-dominant and 12,4% was on the central part of the body.

For the non-contact injuries (74,4%), the most frequent were on the lower body, prevalently muscle injury (33,3% of total injuries). The most injured muscle was the quadriceps and, out of the 21 injuries 23,8% were considered moderate, 19% were considered mild and just 1% was considered severe. We should point out that hamstrings' injuries affected 14 players but no one had recurrence of this injury, which is an indicator of success in modern football medicine (Hickey, 2014).

The U23 team was the most affected by muscle injuries (n=12, 51,2% of total muscle injuries) "with both shooting and sprinting as the main injury mechanisms (n=6, 27.3% of quadriceps mechanisms of injury)" (Luis, 2021).

Luis (2021) realized that for lower body non-contact soft tissue injuries, the "Ratio_BW_Squeeze" was the only significantly different variable between groups with non-injured players showing higher values compared with injured players. The same study found a similar pattern for the non-contact groin injuries.

Interesting is to see that when comparing injured and non-injured groups for the muscle injuries, there were no differences in all the studied variables.

About burden, Luis (2021) found that muscle (Mean time-loss: 16.6 days, incidence: 0.9 injuries /1000h) and ligament injuries (Mean time-loss: 14.8 days; incidence: 0.61 injuries /1000h)" were the most burdensome, which is according to the findings of Calligeris (2015).

On average for all teams' players missed 2,4 matches and the team the had more players out of competition was the U15, with an average of 3,5 matches missed (see table 4.5). The players of the U17 and U15 had an average of approximately one month off (29, 6 and 29,9 days, respectively) which is in fact a very important information. As previously stated, the U17 team has the highest training load (241.7 hours), and data also shows that 87% of the injuries result of a non-contact mechanism, most frequently by overload (30%), sprinting (17%) and landing after a jump (10%).

Team	Players	Trainings Lost	Matches Lost	Days Lost
	N	*Mean±SD*	*Mean±SD*	*Mean±SD*
U23	25	10,6±9,3	2,8±2,4	16,1±13.6
U19	28	11,1±11,2	1,5 ±1,8	16,9 ±17,6
U17	25	17,1 ±23,9	2,9 ±4,7	29,6 ±40,2
U16	25	8,9 ±8,0	1,6 ±1,9	17,5 ±14,5
U15	32	16,7 ±21,7	3,5 ±4,7	29,9 ±39,9
U14	49	9,4 ±8,3	1,9 ±2,3	17,4 ±17,3

Table 4.5 - Comparison of absent days, trainings and matches loston all ages (mean and standard deviation)

Non-contact injuries represented 74.4% of total injuries with a total burden of 48 days lost /1000h.

Although the most severe type of injury was the meniscus with mean days lost of 104.5 ±74.9, it only represented 3.1% of total injuries. This may be explained by the fact that some meniscal injuries require surgical treatment which may lead to a longer period of absence.

The variables chosen to characterize the peak force were the squeeze test and the countermovement jump, and there fore we decided to compare the values measured among all the teams to check if the age variable influences i.

As we can see in the table 4.6 the ratio for the hip adductor strength and bodyweight is the same for all the teams (over 0,5N) which means that mostly all players are very well balanced between age, bodyweight and strength.

Van Klij (2021) identified thigh adductor squeeze test as a useful diagnostic tool in identifying groin pain.

Team	Players	Ratio BWSQZ (N/Kg)	CMJ ESQ (cm)	CMJ DTO (cm)
	N	*Mean±SD*	*Mean±SD*	*Mean±SD*
U23	25	0,6 ±0,1	20,5 ±2,6	21,5 ±3,2

U19	28	0,6 ±0,1	20,0 ±2,7	19,8 ±2,9
U17	25	0,7 ±0,1	20,4 ±2,8	19,8 ±3,2
U16	25	0,7 ±0,1	17,9 ±3,5	17,9 ±3,4
U15	32	0,7 ±0,1	17,2 ±3,5	16,9 ±2,8
U14	49	0,7 ±0,1	16,9 ±3,6	17,5 ±4,3

Table 4.6 - Comparison of values for hip adductor strength (ratio bodyweight-squeeze) and countermovement jump for the left and right sideon all ages (mean and standard deviation).

The CMJ is a very useful tool to assess the body power and there are some normative values that help us identify if a player has the strength he is supposed to according to his/her age (Petrigna, 2019. Siembida, 2021). Although, the way these tests were performed (on a single leg) don't allow us to compare with these normative values and therefore we can only assess that if a players get older, he also gets stronger or not. As we can see in the table above, this is true. Specially if we compare the strength of the U14 players with the strength of the U23. In fact, the information that may be considered more valuable for the CMJ is the standard variation in each sample.

With this in mind we observed that for the CMJ with left leg the intra-sample variation in the U23 is of 12,7% and gets higher as ages lower (21,3% for the U14). The same phenomenon happens for the CMJ of the right leg, the intra-sample variation is of 15% for the U23 and gets progressively higher as age decreases (25% for the U14). One reason for this is that the number of players for the U14 and U15 is higher so there is more diversity of athletic development among these players. Another reason is the obvious specialisation and the specific strengthening work accumulated over time that benefits older players.

So that we could have an ideia on how the variables do may relate to each other we decided to group them in two according to our expertise and other studies published on this domain.

4.3.13.1. Position and Inj_type

By looking at table 4.7 we can tell that ligament injuries are most common among wingers (n=8, 42%) and Centre-back (n=7, 26%). For Goakeepers (GK), centre-back and midfielders the most frequent injury are the muscle injuries.

Wing-back players are mostly affected by muscle and overload injuries.

Position	**3 Most Frequent Injury Type**						***Total Injuries***
	Ligament		**Muscle**		**Overload**		
	N	*%*	*N*	*%*	*N*	*%*	
GoalKeeper	0	*0,0%*	4	*66,7%*	0	0,0%	6
Centre-back	7	*25,9%*	8	*29,6%*	2	7,4%	27
Wing-back	2	*7,7%*	8	*30,8%*	6	23,1%	26
Midfielder	6	*18,2%*	14	*42,4%*	2	6,1%	33

Winger	8	*40,0%*	6	*30,0%*	3	15,0%	20
Forward	6	*35,3%*	3	*17,6%*	3	17,6%	17

Table 4.7 - Most frequent types of injury per position.

4.3.13.2. Position and Inj_mechanism

For goalkeepers (GK), the most frequent action that causes injuries is kicking the ball (33,3%). One of the specific GK gesture is kicking the ball far away to as closest as possible to the strikers and the magnitude of this action puts these players in higher risk of muscle injuries in the thigh of the dominant side.

The most common injury mechanism in all positions of the filed is overload, although there are differences on the type of injury and its location. In defense players overload affects them specially on the thigh of the dominant side. Strikers and wingers are mainly affected by overload on the groin on the non-dominant side. Midfielders mostly sustained muscle injuries on the thigh of the dominant side due to overload.

4.3.14. Position and Inj_context

For all positions the most common context of injury was during training with teammates (71,3%). Injuries sustained in matches corresponds to 23,3% and the positions that were least affected in this context were the forward and winger players (See Table 4.8).

While representing the national, one GK, one midfielder and one forward sustained injury, which corresponds to 2,3% of all injuries.

Position	**Training**		**Match**		**National Duty**		**Other**		**Total Injuries**
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*	
GoalKeeper	4	66,7%	1	16,7%	1	16,7%	0	0,0%	6
Centre-back	19	70,4%	8	29,6%	0	0,0%	0	0,0%	27
wing-back	18	69,2%	7	26,9%	0	0,0%	1	3,8%	26
Midfielder	22	66,7%	9	27,3%	1	3,0%	1	3,0%	33
Winger	16	80,0%	3	15,0%	0	0,0%	1	5,0%	20
Forward	13	76,5%	2	11,8%	1	5,9%	1	5,9%	17
Total	92	71,3%	30	23,3%	3	2,3%	4	3,1%	129

Table 4.8 - Most frequent types of injury per position.

4.3.14.1. Inj_type and Age

For the U14, U16 and U17 teams the most prevalent are the muscle injuries (25%, 23%, 29%, respectively), as for the older teams' ligaments sprains are the most common (U17, 17% and U19, 50%). One explanation may be that these are more experienced and are very specialised in the specific

movements of soccer. One of the main focus of the sports performance teams nowadays, either for performance and injury prevention, is specificity training. This means since young age, players are trained to be at their best in their specific position, this means that a midfielder won´t have exactly the same training as as goalkeeper, for example. Mandorino (2023) stated that "playing position, as well as sport specialization, exposes young soccer players to greater injury risk. Many factors (e.g., neuromuscular control, training load, maturity status) can modify the susceptibility to injury in young soccer players".

4.3.14.2. Inj_Recidive and Age

Half of the recidives happened in 15-year-old athletes, putting the U16 team as the one with the most recidivism with a total of 4. Out of the eight recurrent injuries, 75% were muscle injuries of the quadriceps,50% were considered severe injuries with an average of 36 days absent from training/match.

In 2012, Cloke ran a five-year period study on thigh muscles in youth soccer athletes (12306 players) and concluded that the quadriceps was the muscle with most prevalence, although the factors with poorest prognostic factors for recovery were hamstring injuries, contact injury, and older age.

The injury mechanism was different for each of the recidive which was sprinting (35%), overload (25%) and the rest was by decelerating after a sprinting, player intercepted while running and kicking the ball.

4.3.14.3. Inj_Severity and Inj_Loss_days

Usually, minimal severity injuries took 2 or 3 days to get the player back on the pitch, while for mild injuries the absence time was of most commonly of 6 days. For moderate injuries the most usual was that player was out for 15 days, and longest recoveries took 27 days. The most severe injuries took 187 days of absence but usually this severity means that the player will be out an average of 61 days off, most commonly 50.

4.3.14.4. Inj_Loss_days and Inj_Recidive

When analysing the graph we expected that the injuries with the most absent values were the ones that had the most recidives and that was not confirmed.

When recovering from a recurrent injury the average time off competition was 30 days. Although there were some recorrent injuries with recovery periods longer than 30 days, there were some other that happen after shorter period rehabilitation. We might wonder another explanation, which is that faster rehabilitation is less secure and reliable. Nevertheless, we don't have enough data to support us in this affirmation.

4.3.14.5. PKFO_ESQ, PKFO_DTO and Inj_Location

Our expectation with this analysis was that lower values of PKFO would correspond to higher number of injuries in the groin and hip. By looking at table 4.9 we can tell that even player with good hip range of motion (HROM) had groin and hip injuries. The average values for the PKFO on both sides is 27,7 cm. Out of the 25 injuries on the groin/hip, 11 were on players with HROM lower than the average and the other 14 players had HROM higher than the average. Our initial expectation does not relate to the findings on the data.

The table below shows that there isn't much difference on the mean values for HROM over all ages, and although joint mobility is part of the optimization goals for any athlete, it is in fact a characteristic that doesn't change much over time. Unless the mobility restriction is of such magnitude that puts any player at risk of injury. Therefore, and by analysing the standard deviation we realise that for all the samples (teams) there isn't a significant difference between them.

Team	Players	PKFO left	PKFO Right
	N	*Mean±SD*	*Mean±SD*
U23	25	26,2 ±5,2	26,4 ±5,2
U19	28	27,9 ±4,0	27,5 ±4,8
U17	25	25,6 ±4,5	26,2 ±4,5
U16	25	27,6 ±5,9	27,8 ±5,8
U15	32	28,3 ±4,6	28,1 ±4,7
U14	49	27,5 ±3,9	28,0 ±3,8

Table 4.9 - Comparison of values for hip range of motion for the left and right side for all ages (mean and standard deviation)

4.3.14.6. ADL_DTO, ADL_ESQ and Inj_Location

With this analysis we were curious to see if lower scores on the ADL would represent more injuries on the ankle and in fact that doesn't seem to have happened with the studied population. The average distance measured in the athletes who had ankle injuries was 8 cm and mostly occured on the non-dominant side, which expectedly is the side with which the player doesn't use to control the ball (see Table 4.10).

For all the population, the average distance for ADL is of 9 cm on the left and 8,5 on the right ankle.

Team	Players	ADL Left	ADL Right
	N	*Mean±SD*	*Mean±SD*
U23	25	9,3 ±3,5	9,2 ±3,5
U19	28	7,0 ±2,6	6,9 ±2,7
U17	25	9,6 ±3,2	9,2 ±3,2
U16	25	8,9 ±3,7	9,2 ±3,6
U15	32	9,0 ±3,9	8,5 ±3,5
U14	49	9,0 ±3,0	8,4 ±2,6

Table 4.10 - Comparison of values for ankle range of motion for the left and right side for all ages (mean and standard deviation)

The player who had the lowest scores for the ADL (specially on the right side, 1cm) sustained three injuries, two ligament injuries on the dominant and non-dominant side, and a non-contact muscle injury

on the thigh (non-dominant side). This player is 18 years old and was out for 25 days, lost 4 matches and 16 trainings.

4.3.15. Domain Understanding

By looking at the dataset and the features in it there are some domain-specific terminologies that we need to clarify.

One of them is the BMI, commonly used in medical and sports research. The World Health Organization (WHO) developed a classification that divides individuals in three categories: Normal (18.5 ≤ BMI < 25 kg/m2), overweight (25 ≤ BMI < 30 kg/m2) and obese (BMI ≥ 30 kg/m2). But high-performance athletes have more muscle and mineral bone density than a sedentary (non-athletic) person, which makes him/her weigh a few kg more (Walsh, 2018). In this case the BMI may indicate that an athlete is overweight but all he/her has is plenty of muscle. For this reason, BMI is an excellent indicator of potential health problems but not so useful to figure out how much muscle and fat the athlete has (Nevill, 2009). Although this is true, all the other methods to assess body composition are very dependent on the person who measures it, so it is not useful when doing prevalence studies in large populations.

The adductor squeeze test has become one of the most popular screening tests to assess adductor muscle strength in various sports due to its simplicity, low-cost, good reliability and validity (Delahunt et al., cited by Moreno-Perez, 2019). This test gives us the isometric adductor strength, but the most useful information is given by maximal hip adductor strength, which is basically the calculation of Hip adductor strength, normalized by body mass. In this study, the maximal hip adductor strength is named in the metadata as “Ratio_BW_Squeeze” and was calculated by dividing the squeeze test results by the weight.

For the countermovement jumps there are some normative values described in the literature. Yet we won’t take them into consideration because the method applied in this population wasn’t the standard as any other medical or sports study. Usually, the counter movement jump is performed on a bipodal position, so all the normative values found in the literature are specific for that method. These values were collected by having the player jump on one leg and this changes drastically the performance of this test. We believe that this adaptation doesn’t affect our focus and our goals. It would affect if we wanted to compare different populations which is not the case.

The term recidive or recurrence is used whenever the athlete experiences the same injury after having recovered from the initial one. Preventing recidives is one of the main concerns of the medical and physical trainer staff when implementing a recovery and return to play strategy.

4.4. DATA PREPARATION

For the previous chapter we used the complete raw data to explore as much information as we could and try to identify some patterns among the population. This process helped us understand better the

characteristics of the dataset and more importantly, group the features according to the type of information they provide.

The features that are specific to characterize the athlete are: Age, Heigth, Bodyweigth, Body Mass Index, Position and Dominant Side.

The information on load/stress and burden is given by the features that register the number of hours in training and in matches and the total amount of hours in activity.

To get information and characterize the athlete according to his physical abilities we used some physical screening tests: Squeeze test (of which derived the ratio BWSQZ), Countermovement Jump, Passive Knee Fall Out and Dorsiflexion Lunge test.

For last, there were variables created together information about the injury: type of injury, date of injury, contact injury or not, injury severity, context in which injury occurred, location, side of the injury, injury mechanism, type of muscle injury, recidive injury.

There were also three variables that helped us understand the way the injuries influenced players availability to train and help the team in the match: trainings, matches and days loss due to injury.

Our goal in this chapter is to transform raw data previously analysed into a form that can be used to model with good precision the machine learning algorithm.

On the previous exploratory approach to data, we realised that there are a few attributes that provide the same information but were collected in different moments. The attributes that have repeated information are “Bodyweight”, “Squeeze”, “Ratio_BW_Squeeze”, “CMJ”, “PKFO” and “ADL”. Only one is a derived variable but all the others were collected in different moments and by different people. This happened because the screening tests that the medical and sports performance team conducted at the beginning of the season comprised tests that were also used by the research for the matter of this study. Knowing this we decided to keep the newest attributes which were collected by the research team by the time this study started, only for the purpose of modelling.

4.4.1. Data Cleaning

The dataset cleaning process included identifying and dealing with data errors (incorrectly imputed), and identification and dealing with missing values.

4.4.1.1. Data Errors

When analysing the dataset, we found three input errors on the attributes of CMJ_DTO_2, CMJ_ESQ_2 and CMJ_ESQ_1 for entries 64 and 19 respectively (both not injured athletes). The value found was “- and we don’t have any information about its meaning. Therefore, we will consider this as an error when completion the test. Most probable cause is that the player didn’t do the test and the research team filled the field with an “-“ instead of leaving it blank. We decided to handle this by changing the value and turning it into a Missing Value.

Similar phenomenon happens with entry 77 (not injured athlete), in variable Squeeze1 the value was “-“. We took the decision to consider it a missing value.

We didn't find any other data errors beside these two.

4.4.1.2. Missing Values

When analysing the missing values, we observed that some features had no missing values, those were: ID_Código; Team; Age; Position; Dominant Side; Training_H; Match_H; Total_H; Inj_Type.

Sixteen attributes had a percentage of missing values **below 10%**: Height; Bodyweigth_1; Bodyweigth_2; BMI, Squeeze_1; Squeeze_2; Ratio_BW_Squeeze_1; CMJ_DTO_1; PKFO_ESQ_1; PKFO_ESQ_2, PKFO_DTO_1, PKFO_DTO_2; ADL_ESQ_1; ADL_ESQ_2; ADL_DTO_1; ADL_DTO_2. These missing values are considered to be missing completely at random (MCaR). Three attributes had a percentage of missing values between 10 and 15%: Ratio_BW_Squeeze_2; CMJ_ESQ_2; CMJ_DTO_2, and as so they are missing completely at random.

Height; Bodyweigth_1; Bodyweigth_2; BMI were collected to characterize the demographics of the population and they are grouped according to the team they represent ("Age"), therefore we decided to impute values using forward fill. Exception was made for the BMI dimension, given that we can calculate it, if we have the MV's for height and bodyweight, which we now do (after the forward fill imputation). After handling these MV's we ended up with 226 entries for each of them (see table 4.11 below).

To the numerical variables of the dimensions that characterize the physical condition of the athlete we decided to impute the mean value. With this method we were able to clear the MV's for the squeeze tests, the CMJ's, the PKFO's and for the ADL's. Therefore, we could re-calculate the values for Ratio_BW_Squeeze_1 and 2 (RatioBWSQZ = Squeeze_/ Bodyweight) and lower the MV's on both dimensions, to zero (0).

Twelve attributes had a percentage of missing values of 42,9% (seen in yellow in Table 4.11): Inj_Date; Inj_Loss_days; Inj_Severity; Inj_Context; Inj_Location; Inj_Side; Inj_Mechanism; Inj_MuscleType; Inj_LostTrainings; Inj_LostMatches; Inj_Contact; Inj_Recidive. These features characterise an injury and if there is no injury then there'll be no values to fill them in. In this case the missing values refer to the inputs of athletes who didn't get injured and therefore we can consider them to be missing not at random (MNaR). In other words, 42,9% of inputs are of no injured athletes (code 0 , "No injury" of the "Inj_Type" variable).

The feature "Inj_MuscleType" had a 81% of missing values but (MNaR, seen in red in Table 4.11), in reality this is not true. As previously discussed, this item refers to a sub-population of athletes who suffered muscle injuries (n=43, approximately 19,02% of total injuries), and for the condition of having a muscle injury (code 7 in "Inj_Type") all inputs have a value, therefore there aren't any real missing values.

The table below depicts the method we chose to handle MV's in the different dimensions. For the dimensions with categorical variables (used to characterize injury) we imputed the value "25" as a new category for the non injured. We didn´t use the "0" because in some dimensions this value already corresponds to a category.

For the MV's on the trainings/matches lost we imputed the value 0 (zero) because indeed, the players didn´t lost any train/match. The same reasoning was applied to the injury date and days lost.

Missing Values

Variable Name	*Missing Values %*	*Method*	*Expected Inputs*	*Real Inputs*	*Missing values*	*Missing Values %*
ID_Código	0,0%	n.a.	226	226	**0**	0,0%
Team	0,0%	n.a.	**226**	226	**0**	0,0%
Age	0,0%	n.a.	**226**	226	**0**	0,0%
Heigth	2,2%	Forward Fill	**226**	226	**0**	0,0%
Bodyweigth_1	1,8%	Forward Fill	**226**	226	**0**	0,0%
Bodyweigth_2	1,8%	Forward Fill	**226**	226	**0**	0,0%
BMI	2,7%	Calculate	**226**	226	**0**	0,0%
Position	0,0%	n.a.	**226**	226	**0**	0,0%
Dominant Side	0,0%	n.a.	**226**	226	**0**	0,0%
Training_H	0,0%	n.a.	**226**	226	**0**	0,0%
Match_H	0,0%	n.a.	**226**	226	**0**	0,0%
Total_H	0,0%	n.a.	**226**	226	**0**	0,0%
Squeeze_1	4,9%	Impute Mean	**226**	226	**0**	0,0%
Squeeze_2	1,3%	Impute Mean	**226**	226	**0**	0,0%
Ratio_BW_Squeeze_1	5,3%	Calculate	**226**	226	**0**	0,0%
Ratio_BW_Squeeze_2	10,6%	Calculate	**226**	226	**0**	0,0%
CMJ_ESQ_1	9,7%	Impute Mean	**226**	226	**0**	0,0%
CMJ_ESQ_2	11,9%	Impute Mean	**226**	226	**0**	0,0%
CMJ_DTO_1	9,3%	Impute Mean	**226**	226	**0**	0,0%
CMJ_DTO_2	11,9%	Impute Mean	**226**	226	**0**	0,0%
PKFO_ESQ_1	5,8%	Impute Mean	**226**	226	**0**	0,0%
PKFO_ESQ_2	1,8%	Impute Mean	**226**	226	**0**	0,0%
PKFO_DTO_1	5,8%	Impute Mean	**226**	226	**0**	0,0%
PKFO_DTO_2	1,8%	Impute Mean	**226**	226	**0**	0,0%
ADL_ESQ_1	5,8%	Impute Mean	**226**	226	**0**	0,0%
ADL_ESQ_2	1,8%	Impute Mean	**226**	226	**0**	0,0%
ADL_DTO_1	5,3%	Impute Mean	**226**	226	**0**	0,0%
ADL_DTO_2	1,8%	Impute Mean	**226**	226	**0**	0,0%
Inj_Type	0,0%	n.a.	**226**	226	**0**	0,0%
Inj_Date	42,9%	Impute value "0"	**226**	226	**0**	0,0%

Inj_Loss_days	42,9%	Impute value "0"	**226**	226	**0**	0,0%
Inj_Severity	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_Context	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_Location	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_Side	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_Mechanism	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_MuscleType	81,0%	Impute value "25" - No Muscle Injury	**226**	226	**0**	0,0%
Inj_LostTrainings	42,9%	Impute value "0"	**226**	226	**0**	0,0%
Inj_LostMatches	42,9%	Impute value "0"	**226**	226	**0**	0,0%
Inj_Contact	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%
Inj_Recidive	42,9%	Impute value "25" - No Injury	**226**	226	**0**	0,0%

Table 4.11 - Handling Missing Values

4.4.2. Data Transformation

4.4.2.1. Encoding Variables

As shown in the metadata, most of the dimensions are characterized by a number. Although in some of them, the number relates to a category. As an example, for the dimension type of injury, “4” corresponds to the category of “Meniscus Injury”, this means that if a player sustains a meniscus strain it should be registered with the appropriate number (“4”).

So that the model can identify the numbers as categories we need to transform these numerical variables into categorical variables (see Table 4.12).

Variable Encoding

Variable Name	*Original Type*	*Transformed to*
Team	Numeric	Categorical Nominal
Position	Numeric	Categorical Nominal
Dominant Side	Numeric	Binary
Inj_Type	Numeric	Categorical Nominal
Inj_Severity	Numeric	Categorical Ordinal
Inj_Context	Numeric	Categorical Nominal
Inj_Location	Numeric	Categorical Nominal
Inj_Side	Numeric	Categorical Nominal
Inj_Mechanism	Numeric	Categorical Nominal
Inj_MuscleType	Numeric	Categorical Nominal

Inj_Contact	Binary	Categorical Nominal
Inj_Recidive	Binary	Categorical Nominal

Table 4.12 - Variable Encoding for Modeling

4.4.2.2. Variable Renaming

One other transformation we did on the dataset was to rename some variables in order to make it descriptive and therefore easier to visualize by the time we proceed with the analysis (see table 4.13).

Variables Renaming	
Original Name	*Re-name*
Training_H	**Training Exposure**
Match_H	**Match Exposure**
Total_H	**Total Exposure Time**
Inj_Date	**Injury Date**
Inj_Type	**Type of injury**
Inj_Loss_days	**Injury loss days**
Inj_Severity	**Injury Severity**
Inj_Context	**Injury Context**
Inj_Location	**Injury Location**
Inj_Side	**Injury Side**
Inj_Mechanism	**Injury Mechanism**
Inj_MuscleType	**Type of muscle Injury**
Inj_LostTrainings	**Trainings Lost**
Inj_LostMatches	**Matches Lost**
Inj_Contact	**Injury Contact**
Inj_Recidive	**Injury Recidive**
Ratio_BW_Squeeze_2	**Ratio BWSQZ**
CMJ_ESQ_2	**CMJ Left**
CMJ_DTO_2	**CMJ Right**
PKFO_ESQ_2	**PKFO Left**
PKFO_DTO_2	**PKFO Right**
ADL_ESQ_2	**ADL Left**
ADL_DTO_2	**ADL Right**

Table 4.13 - Original and renamed variables for modelling.

4.4.2.3. Metadata Post Data Preparation

After all the data preparation we ended up with the following dataset characteristics (Table 4.14).

Revised Metadata		
Variable Name	*Type*	*Range or Categories*
ID_Código		
Team	**Categorical Nominal**	S23 =**2**; S19 = **3**; S17 = **4**; S16 = **5**; S15 = **6**; S14 = **7.**
Age	**Numeric continuous**	Range: **13 to 22 years old.**
BMI	**Numeric continuous**	Range: **15,62 to 30,57 Kg/m2**
Position	**Categorical Nominal**	GR ("Goalkeeper") =**0**; Defesa central ("Centre-back") = **1**; Defesa Lateral ("wing-back") = **2**; Médio ("Midfielder") = **3**; Extremo ("Winger") = **4**; Avançado ("Forward") = **5**.
Dominant Side	**Binary**	E ("Left") =**0**; D ("Right") = **1**
Training Exposure	**Numeric continuous**	Minimum = **0.00;** maximum = **352**
Match Exposure	**Numeric continuous**	Minimum = **0.00;** maximum = **42,25**
Total Exposure Time	**Numeric continuous**	Minimum = **0.00;** maximum = **379,18**
Ratio BWSQZ	**Numeric continuous**	Minimum = **0.29;** maximum = **0,94**
CMJ Left	**Numeric continuous**	Minimum = **10;** maximum = **28**
CMJ Right	**Numeric continuous**	Minimum = **7;** maximum = **27**
PKFO Left	**Numeric continuous**	Minimum = **12;** maximum = **43**
PKFO Right	**Numeric continuous**	Minimum = **13;** maximum = **40**
ADL Left	**Numeric continuous**	Minimum = **2;** maximum = **20**
ADL Right	**Numeric continuous**	Minimum = **1;** maximum = **18**
Type of injury	**Categorical Nominal**	Não Lesionado ("no injury") =**0**; Fratura ("fracture") = **1**; Outra Lesão óssea ("other bone pathology") = **2**; Luxação ("dislocation") = **3**; Ligamentar ("ligament injury") = **4**; Meniscal ("meniscus injury") = **5**; Cartilagem ("cartilage injury") = **6**; Lesão muscular ("muscle injury") = **7**; tendinosa ("tendon injury") = **8**; Contusão ("contusion") = **9**; Abrasão ("abrasive injury") = **10**; Laceração ("laceration wound") = **11**; Concussão ("concussion") = **12**; Nervo ("nerve trauma") = **13**; Sinovite ("synovitis") = **14**; Sobrecarga

		("overload") = **15**; Outra Lesão other type") = **16**.
Injury Date	**Numeric continuous**	Minimum = **3;** maximum = **243**
Injury loss days	**Numeric continuous**	Minimum = **2;** maximum = **187**
Injury Severity	**Categorical Ordinal**	Minimal =**0**; Mild = **1**; Moderate = **2**; Severe = **3;** No Injury = **25**
Injury Context	**Categorical Nominal**	**Treino ("Train") = 0**; Jogo ("Match") = **1**; Seleção ("national Team") = **2** Nacional; Outro ("Other") = **3**.
Injury Location	**Categorical Nominal**	**Face = 1**; Cabeça ("head") = **2**; Cervical = **3**; ombro ("shoulder") = **4**; cotovelo ("elbow") = **5**; antebraço ("forearm") = **6**; punho ("wrist") = **7**; mãos/dedos ("hand/fingers") = **8**; abdómen ("abdomen") = **9**; Pélvis ("pelvis") = **10**; lombar ("lumbar") = **11**; Groin = **12**; Anca ("hip") = **13**; Coxa ("thigh") = **14**; Joelho ("knee") = **15**; Perna ("leg") = **16**; Aquiles ("achilles") = **17**; Tornozelo ("ankle") = **18**; Pé/dedos ("foot/toes") = **19;** No Injury = **25**
Injury Side	**Categorical Nominal**	Dominante ("Dominant side") =**0**; Não Dominante ("Non-dominant side") = **1**; Central = **3;** No Injury = **25**
Injury Mechanism	**Categorical Nominal**	Sprint =**1**; Rodar ("turn") = **2**; Rematar ("Kick") = **3**; Passar/Cruzar ("Pass/cross") = **4**; Driblar ("Dribble") = **5**; Saltar ("Jump") = **6**; Aterrar ("Land") = **7**; Cair ("Fall down") = **8**; Mergulhar ("Dive") = **9**; Alongamento ("Stretching") = **10**; Deslizar ("Slide") = **11**; Sobrecarga ("Overload") = **12**; Compressão ("Compression") = **13**; Atingido P/bola ("Hit by the ball") = **14**; Colisão ("Collision") = 1**5**; Cabecear ("Header") = **16**; Interceptado ("Intercepted") = **17**; Interceptar

		("Interception") = **18**; Pontapeado ("Kicked") = **19**; Bloqueado ("Blocked") = **20**; Uso de braço ("Arm usage") = **21**; Acelerar ("Acceleration") = **22**; Desacelerar ("Decceleration") = **23; No Injury = 25**
Type of muscle Injury	**Categorical Nominal**	Isquiotibiais ("Hamstrings") =**1**; Adutores ("Adductors") = **2**; Quadriceps = **3**; Gémeo ("Gastrocnemius") = **4**; Solear ("Soleus") = **5**; Rotadores Externos Anca ("External rotators of the hip") = **6**; No Injury = **25**
Trainings Lost	**Numeric continuous**	Minimum = **1;** maximum = **103**
Matches Lost	**Numeric continuous**	Minimum = **0;** maximum = **21**
Injury Contact	**Categorical Nominal**	Non Contact =**0**; Contact = **1;** No Injury = **25**
Injury Recidive	**Categorical Nominal**	N (No recidive) =**0**; Y (recidive) =**1;** No Injury = **25**

Table 4.14 - Revised Metadata

4.4.3. Data Reduction

To create a dataset that could improve the accuracy of the model to be created we selected some features that were irrelevant to the study.

Although each feature has its value, some only served as basis to compose other variables that will be present in the final dataset.

The BMI is composed by calculating the index between BW1 and Height, and so we will keep the BMI and eliminate the height and bodyweight (first measure).

There is another variable that reflects the bodyweight of the athlete which is BW_2. This one refers to exactly the same characteristic as BW_1 – body mass, but the difference is that it was recorded after the first one by the time the squeeze test was performed. This feature was useful to determine the ratio bodyweight-squeeze strength (ratio BWSQZ), therefore we can eliminate it and keep just the ratio BWSQZ.

Basically, BW1 was used to calculate BMI and BW2 was used to calculate the Ratio_BW_Squeeze_2.

BW1 is also used to calculate the Ratio_BW_Squeeze_1 which is a variable that we will eliminate, as well as the Squeeze_1 and Squeeze_2. These won't be part of the final dataset because what's more valuable is the ratio BWSQZ and so much the values for the test alone.

Another composed variable is the total hours played, which was calculated by the sum of hours the players spent training and in matches. But in this case, we decided to keep all variables, taking the risk of being redundant, which we will check when we have the correlation matrix.

The other variables we decided to eliminate were related to the measurements of the CMJ, PKFO and ADL acquired during pre-season as screening tests. As previously stated, the research team has collected this same metrics in a different moment, and this is the one we will keep for the matter of this study. Overall, at this stage we eliminated 12 variables and kept the other 28.

By analysing the Seaborn Heat map correlation matrix (Figure 4.2) we can identify some features very strongly related, and realized that all them were used to characterize an injury: 'Type of injury', 'Injury loss days', 'Injury Severity', 'Injury Context','Injury Location', 'Injury Side', 'Injury Mechanism','Type of muscle Injury','injury Contact', 'Injury Recidive','Injury Date', 'Trainings Lost'.

At first, we decided to keep these variables and compare the performance of a few models, but then we realized that the scores for each model's performance was 100%. An accuracy of 100% is perfect but shows that the model is overfitting. Therefore, we decided to go back and eliminate the variables identified by the correlation matrix as having a strong relationship. Although, if we did that then we would lose a lot of injury-related data and that wouldn't be advisable because injuries are our main subject. Another thing we found was that in the category "Type of Injury", there were some labels with very few entries and because of that we decided to divide this category into two groups – Injury and No Injury, and create a variable named Class. Class can take the value of 0 if the observation has no injury or the value of 1 if it has an injury of any type.

After this, we dropped the variables previously identified in the Heatmap (see Figure 4.2 below).

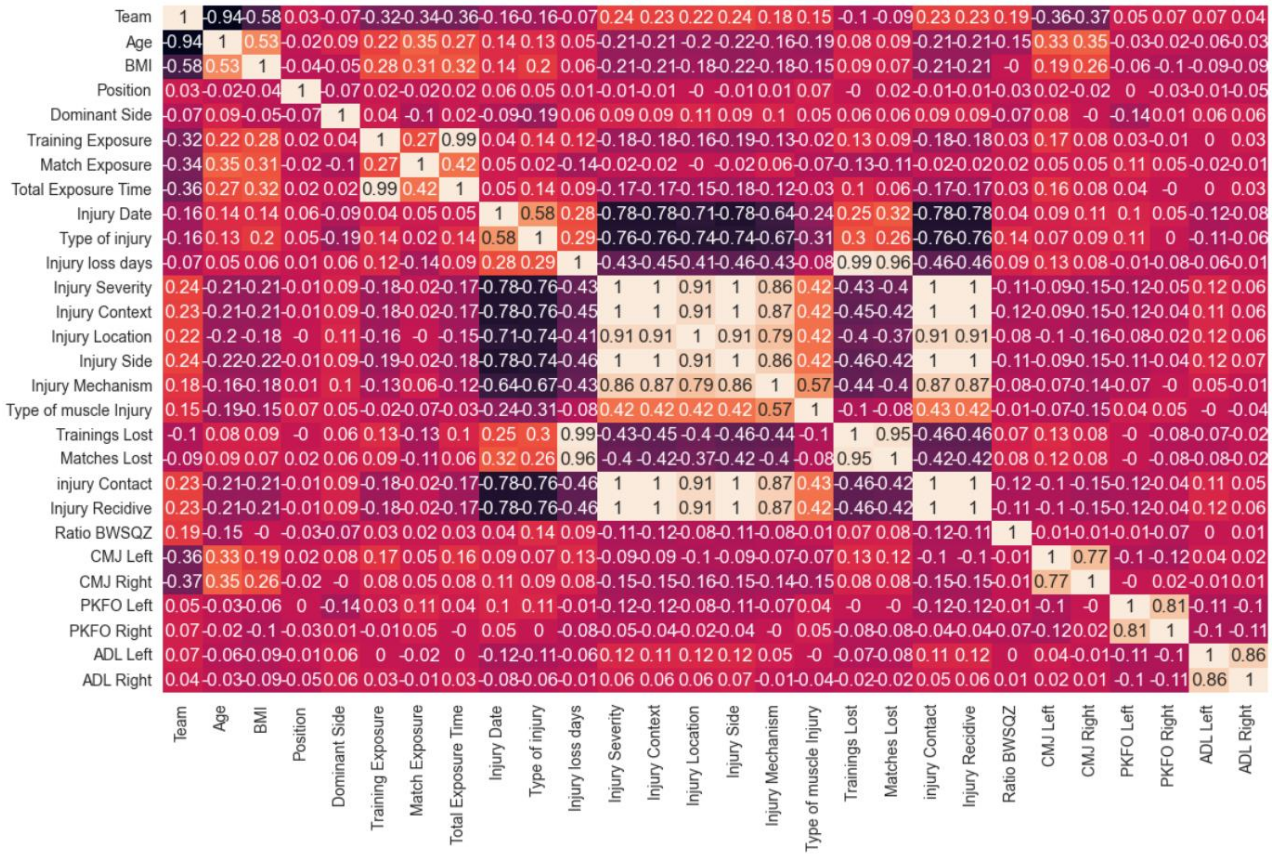


Figure 4.2 - Heatmap for the Correlation Matrix.

Below there is the table 4.15 with all the variables eliminated (seen as **X** in Table 4.15) and the ones kept for modelling (seen as ✓ in Table 4.15).

Variable Elimination

Original	*Eliminated*
Team	✓
Age	✓
Heigth	X
Bodyweigth_1	X
Bodyweigth_2	X
BMI	✓
Position	✓
Dominant Side	✓
Training_H	✓
Match_H	✓
Total_H	✓

Squeeze_1	X
Squeeze_2	X
Ratio_BW_Squeeze_1	X
Ratio_BW_Squeeze_2	✓
CMJ_ESQ_1	X
CMJ_ESQ_2	✓
CMJ_DTO_1	X
CMJ_DTO_2	✓
PKFO_ESQ_1	X
PKFO_ESQ_2	✓
PKFO_DTO_1	X
PKFO_DTO_2	✓
ADL_ESQ_1	X
ADL_ESQ_2	✓
ADL_DTO_1	X
ADL_DTO_2	✓
Inj_Type	X
Inj_Date	X
Inj_Loss_days	X
Inj_Severity	X
Inj_Context	X
Inj_Location	X
Inj_Side	X
Inj_Mechanism	X
Inj_MuscleType	X
Inj_LostTrainings	X
Inj_LostMatches	✓
Inj_Contact	X
Inj_Recidive	X
Class	✓

Table 4.15 – Final List of Variables for Modelling

5. MODELLING AND RESULTS EVALUATION

Claudino (et al., 2019) identified that for injury risk assessment, the artificial neural network, decision tree classifier, and support vector machine have been widely used in soccer, basketball, American football, Australian football, and handball.

5.1. TRAINING CONTROL PARAMETERS

First, we defined 'Class' as our target variable. 'Class' is a composed binary type of variable that was created by the grouping of the sub-categories of the 'Type of Injury', and so it takes the values we want to predict, which are: 'Injury' (1) or 'No Injury' (0).

A fundamental step to go through when training a model to be accurate and precise is the partition of the dataset into two samples: train and test (Rácz, 2021). The training dataset is used to train the algorithm, by which it will learn how to make predictions about the injury. The test dataset is a sample of the original dataset, and it is used to evaluate the model created. Although the magnitude of the partition may differ, the most common used is a 70/30 (Racz, 2021), and so we followed the guidelines provided by the literature (table 5.1).

	Observations
Train dataset	**158 (70%)**
Test dataset	**68 (30%)**

Table 5.1 - Train Test Split

Train test split

The partition of the data was achieved by using a Stratified 10-fold cross validation. The samples obtained would be imbalanced and by using this method of validation we assure that both train and test dataset are representative of each class (0 and 1). The use of 10-fold is the most common and assures lower prediction error (Prusty, 2022).

5.2. CLASSIFIER MODEL COMPARISON

For the classifier model evaluation we used the PyCaret classification module, it will train and evaluate the performance of all the models present in its library (PyCaret Docs). One of the advantages is the possibility to compare models within a short range of time. With this model we will be able to compare and evaluate our model's information collected and we will analyze and compare models' performance according to the problem that we want to solve.

Sweeney (2021) and Boyd (2012) stated that a good score for the accuracy measure would be between 70 and 90%. By looking at the model comparison (Figure 5.1) there are a few that show good accuracy, but the best accuracy scores were obtained by the ***Gradient Boost Classifier*** (87% probability in identifying the correct class of an observation). This classifier is considered to be an ensemble method where the "learning procedure consecutively fits new models to provide a more accurate estimate of the response variable" (Natekin et al., 2013).

As for precision scores, Ting (2011) indicates that scores over 70% are considered good. The classifier who scored best precision was ***Naïve Bayes*** *(98% accuracy in identifying positive occurrences)*.

The model with the best score for Recall (100% accuracy in making positive prediction) is the ***Dummy Classifier***, which is a simple classifier that makes predictions without identifying patterns in the data, it looks at the most frequent class of the training data set and predicts based on that.

The highest F1-score (88% of accuracy) was obtained by the ***Extreme Gradient Boosting*** classifier, which is an ensemble algorithm based on parallel decision trees using a gradient boosting method (Arif Ali, 2023), and is considered to be very powerful and scalable. It has become largely applied by data scientist to achieve breakthrough insights on a large number of ML issues.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8675	0.9317	0.8333	0.9357	0.8755	0.7346	0.7500	0.0170
xgboost	Extreme Gradient Boosting	0.8671	0.9336	0.8667	0.9071	0.8818	0.7297	0.7404	0.0370
et	Extra Trees Classifier	0.8612	0.9434	0.8778	0.8824	0.8779	0.7167	0.7205	0.0220
rf	Random Forest Classifier	0.8604	0.9354	0.8444	0.9121	0.8724	0.7179	0.7269	0.0260
nb	Naive Bayes	0.8421	0.9130	0.7333	0.9889	0.8369	0.6910	0.7251	0.0060
lightgbm	Light Gradient Boosting Machine	0.8358	0.9146	0.8556	0.8707	0.8549	0.6651	0.6802	0.1020
dt	Decision Tree Classifier	0.8171	0.8123	0.8556	0.8369	0.8420	0.6249	0.6354	0.0060
ada	Ada Boost Classifier	0.8167	0.8767	0.8444	0.8467	0.8406	0.6239	0.6334	0.0140
qda	Quadratic Discriminant Analysis	0.7921	0.8661	0.7889	0.8565	0.8132	0.5761	0.5930	0.0060
lr	Logistic Regression	0.7787	0.8971	0.7444	0.8568	0.7887	0.5570	0.5719	0.3370
ridge	Ridge Classifier	0.7475	0.7974	0.7889	0.7718	0.7745	0.4846	0.4960	0.0050
lda	Linear Discriminant Analysis	0.7096	0.8011	0.7444	0.7498	0.7391	0.4092	0.4241	0.0050
knn	K Neighbors Classifier	0.6188	0.6811	0.7000	0.6490	0.6642	0.2216	0.2354	0.2350
dummy	Dummy Classifier	0.5700	0.5000	1.0000	0.5700	0.7260	0.0000	0.0000	0.0080
svm	SVM - Linear Kernel	0.4863	0.5460	0.4889	0.2783	0.3547	-0.0123	-0.0228	0.0060

Figure 5.1 - PyCaretsscoregrid with Accuracy, Recall, Precision, F1 and Kappa scores for the classifiers model comparison.

By grouping all of the injury types in only one class with the label 1 ("Injury") we got a balanced dataset with nearly the same number of positive and negative observations, 43% no-injury (0) and 57% injury

(1). The problem we are dealing in this study is to prediction of injury on an athlete, in other words, prediction of True Positives, with a high cost for misjudging False Negatives. Therefore, the scores we will look at to evaluate these classifiers' performance is Accuracy and Recall.

5.3. Classifier Model Evaluation and Results

PyCaret's experiment workflow recommends the creation and evaluation of one or few specific models after the general model comparison, which gives us an overall view of models' performance but does not return any trained models.

We will create and analyze the following models: Decision Tree, K-Nearest Neighbor, Logistic Regression and Extreme Gradient Boosting.

5.3.1. Decision Tree

The ROC curve analysis and the scoregrid of the DT model presented in Figures 5.2 and 5.3 show that the rate of successful classification by the model was as high as 79% with a sensitivity 86%, a positive predictive value of 84% (precision), and an accuracy of 82%.

The F1 score is 84% which makes it an elegible model to be used as a classifier.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8125	0.8016	0.8889	0.8000	0.8421	0.6129	0.6181
1	0.8750	0.8730	0.8889	0.8889	0.8889	0.7460	0.7460
2	0.8125	0.7857	1.0000	0.7500	0.8571	0.6000	0.6547
3	0.8125	0.8175	0.7778	0.8750	0.8235	0.6250	0.6299
4	0.7500	0.7302	0.8889	0.7273	0.8000	0.4754	0.4927
5	0.8125	0.8016	0.8889	0.8000	0.8421	0.6129	0.6181
6	0.8125	0.8175	0.7778	0.8750	0.8235	0.6250	0.6299
7	0.7500	0.7460	0.7778	0.7778	0.7778	0.4921	0.4921
8	0.8000	0.8056	0.7778	0.8750	0.8235	0.5946	0.6001
9	0.9333	0.9444	0.8889	1.0000	0.9412	0.8649	0.8729
Mean	0.8171	0.8123	0.8556	0.8369	0.8420	0.6249	0.6354
Std	0.0512	0.0577	0.0711	0.0771	0.0438	0.1068	0.1057

Figure 5.2 – Scoregrid for the DT model

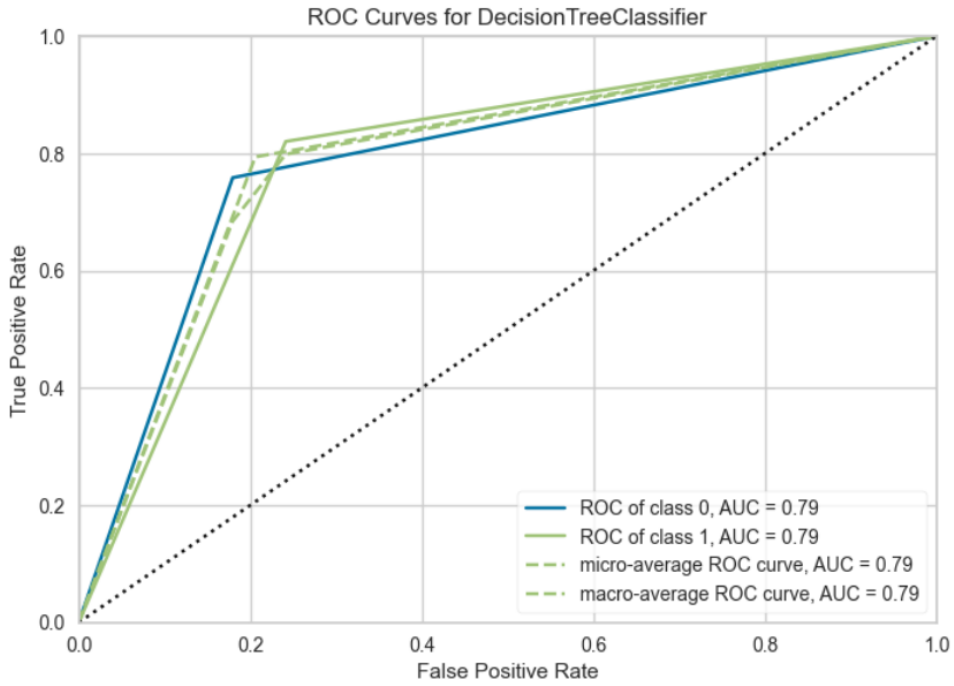


Figure 5.3 - ROC-AUC plot for the DT model

The feature importance plot shows the variables more relevant to the model. When a variable is altered, it is translated as a loss in the AUC, therefore, a variable is considered more relevant as the bigger is the loss in the AUC.

For the DT model, the most important variable is “Matches Lost”, followed by Match Exposure and Total Exposure Time (see Figure 5.4).

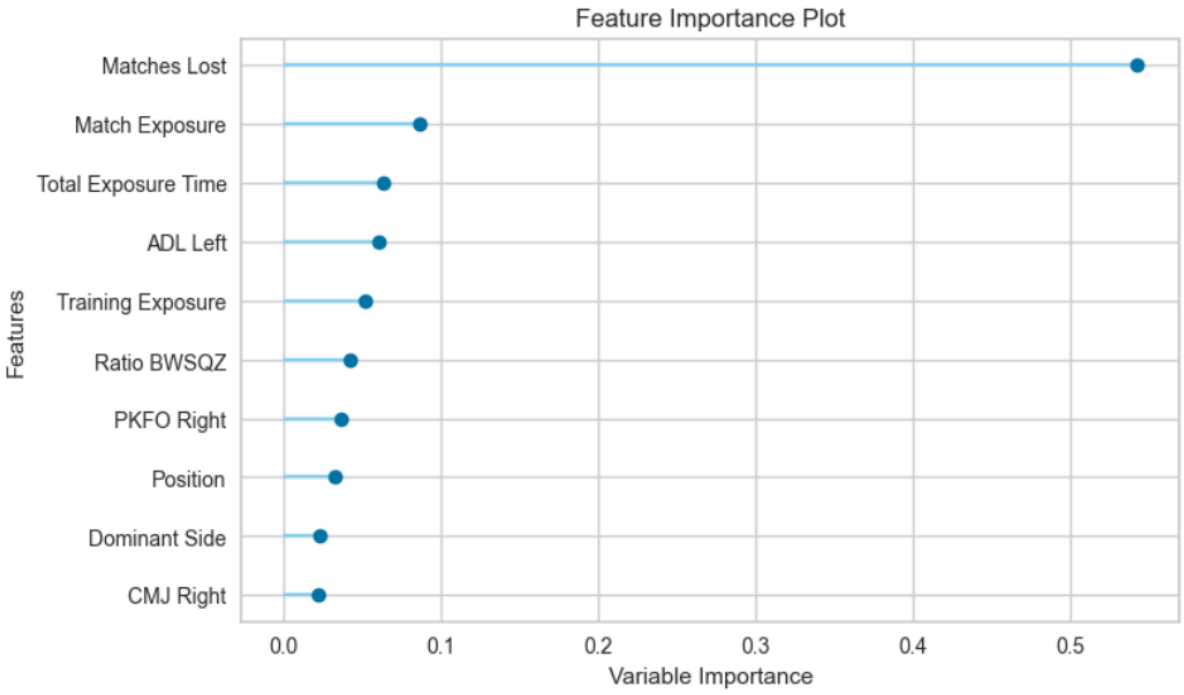


Figure 5.4 - Feature importance plot for the DT model.

5.3.2. K-Nearest Neighbour

Not surprisingly, the KNN model obtain unsatisfactory score values in all performance parameters (see Figure 5.5 and 5.6). For this specific dataset the KNN is neither good at predicting the number of positive and negative cases (only 62% success) nor it's precise enough not to identify false positives (65%). Only 65% of the injured athletes were actually predicted by this model, which is considered a low score for a classifier.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6250	0.5794	0.7778	0.6364	0.7000	0.2131	0.2208
1	0.6875	0.7302	0.7778	0.7000	0.7368	0.3548	0.3578
2	0.5625	0.5952	0.4444	0.6667	0.5333	0.1515	0.1627
3	0.6875	0.7460	0.8889	0.6667	0.7619	0.3333	0.3637
4	0.7500	0.9206	1.0000	0.6923	0.8182	0.4576	0.5447
5	0.6250	0.7619	0.7778	0.6364	0.7000	0.2131	0.2208
6	0.6875	0.7698	0.7778	0.7000	0.7368	0.3548	0.3578
7	0.5625	0.6429	0.5556	0.6250	0.5882	0.1250	0.1260
8	0.6000	0.6019	0.6667	0.6667	0.6667	0.1667	0.1667
9	0.4000	0.4630	0.3333	0.5000	0.4000	-0.1538	-0.1667
Mean	0.6188	0.6811	0.7000	0.6490	0.6642	0.2216	0.2354
Std	0.0927	0.1231	0.1928	0.0558	0.1179	0.1619	0.1806

Figure 5.5 – Scoregrid for the KNN model

The AUC scores show that the model is correct in 66% of its predictions.

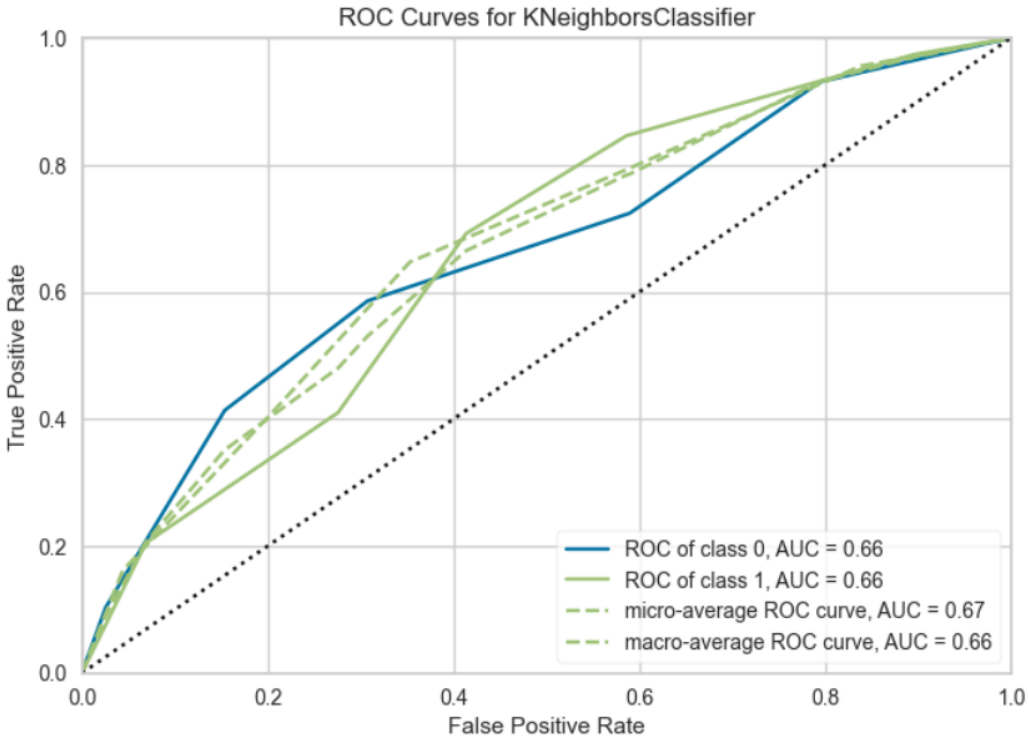


Figure 5.6 - ROC-AUC plot for the KNN model

According to the scores obtained previously we decided not to plot the feature importance, we wouldn't use that information for nothing else so there was no need to do it.

5.3.3. Logistic Regression

By using the LR algorithm we got a 78% accuracy in predicting if a player will get injured or won't get injured. However, its scores were better when analysing how many of the predicted injured players were in fact injured (precision of 86%). The sensitivity to identify an injured player among all the injured population is of 74% (see Figure 5.7).

In the setting that this study is focused on, the impact of having False Negatives is higher than having False Positives. And although we are working with a balanced dataset it is advisable to look at the F1 score and not only at the others.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.6875	0.8413	0.5556	0.8333	0.6667	0.3939	0.4229
1	0.7500	0.8571	0.5556	1.0000	0.7143	0.5224	0.5946
2	0.8125	0.9524	0.8889	0.8000	0.8421	0.6129	0.6181
3	0.6875	0.7937	0.6667	0.7500	0.7059	0.3750	0.3780
4	0.7500	0.8571	0.7778	0.7778	0.7778	0.4921	0.4921
5	0.8125	0.9683	0.7778	0.8750	0.8235	0.6250	0.6299
6	0.9375	0.9841	0.8889	1.0000	0.9412	0.8750	0.8819
7	0.7500	0.8095	0.6667	0.8571	0.7500	0.5077	0.5238
8	0.8000	0.9630	0.7778	0.8750	0.8235	0.5946	0.6001
9	0.8000	0.9444	0.8889	0.8000	0.8421	0.5714	0.5774
Mean	0.7787	0.8971	0.7444	0.8568	0.7887	0.5570	0.5719
Std	0.0689	0.0685	0.1222	0.0814	0.0776	0.1334	0.1309

Figure 5.7 – Scoregrid for the LR model

By analysing Figure 5.8 both the curves for class 1 and 0 are very close to the top left corner, where the sensitivity is high and specifity is low. The AUC scores for both classes is of 91% which is considered to be very good for a classifier model. It means that 91% of the predictions for the injured and non-injured class are correct.

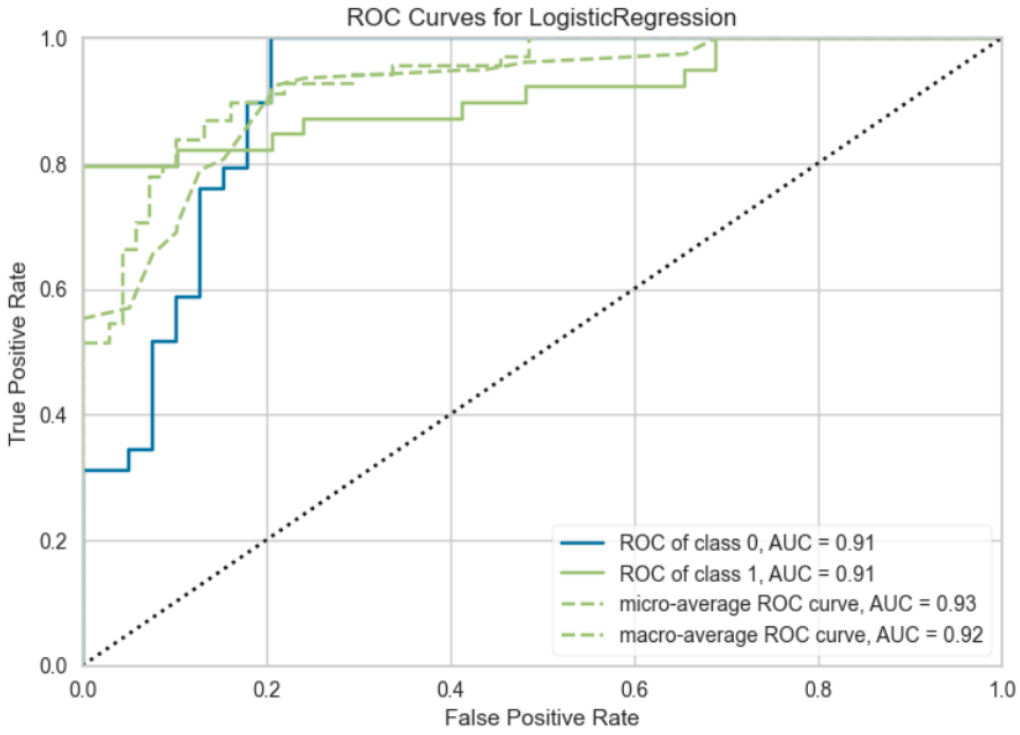


Figure 5.8 - ROC-AUC plot for the LR model.

For the LR model the features that most impact the prediction are Matches Lost, Ratio BWSQZ, Age, Dominant Side and Team (see Figure 5.9).

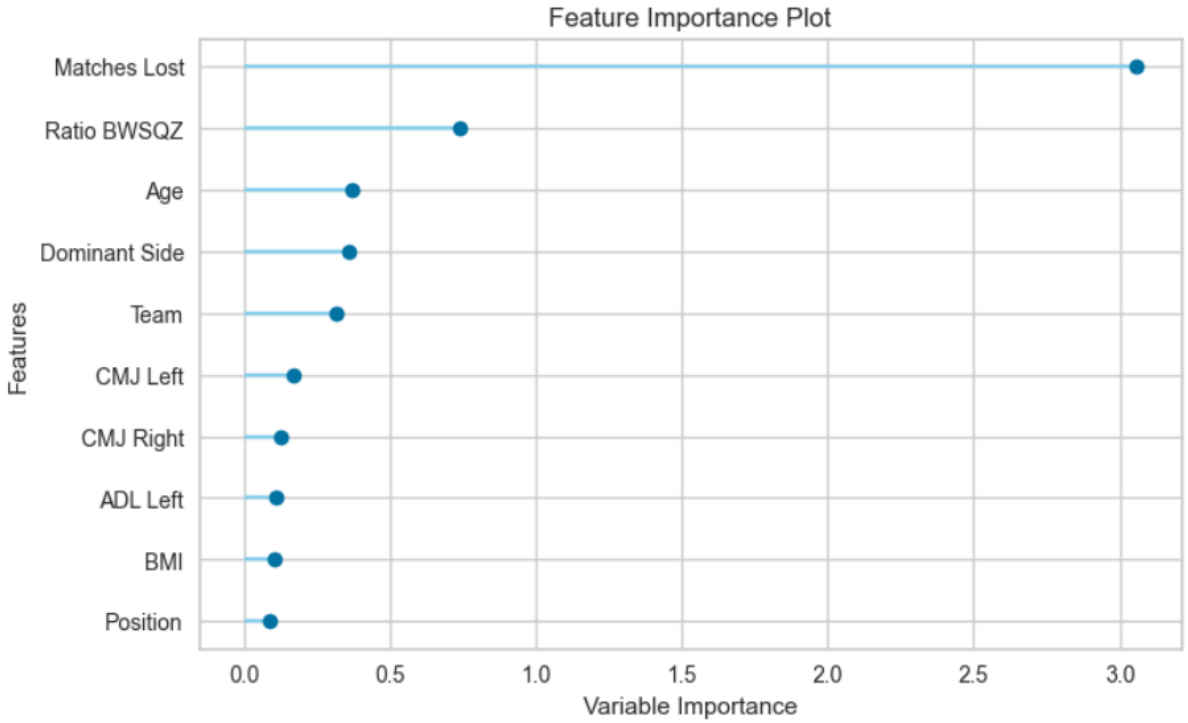


Figure 5.9 - Feature importance plot for the LR model

5.3.4. Extreme Gradient Boosting

The XGBoost classifier was the second top scorer of all the models tested for comparison and the best for the F1 score.

In our opinion the F1 score is of relative importance for this study because, even though the data set is balanced, the weight of having a False Negative (No-injury mistakenly predicted) is higher than having a False Positive.

When analysing the scoregrid for the XGBoost we can tell that it is 87% accurate in predicting if an athlete will be injured or not.

Of all the positive predictions it did, 91% of them were correct, and only 9% were identified as False Positives (identified as injured when in fact they weren't). The XGBoost has the sensitivity of 0.87, which means that it correctly identifies 87% of all injuries.

According to the rationale used for the previous models, the identification of a False Negative has a big impact and for that reason we must weight both precision and Recall with the F1-score. The F1 score is relatively high, 0.88, which means that this classifier is 88% accurate in identifying athletes with injury, and it minimizes the wrong predictions of athletes with injury/no-injury (see Figure 5.10).

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9375	0.9206	0.8889	1.0000	0.9412	0.8750	0.8819
1	0.9375	0.9048	0.8889	1.0000	0.9412	0.8750	0.8819
2	0.8750	0.9683	0.8889	0.8889	0.8889	0.7460	0.7460
3	0.8125	0.8413	0.7778	0.8750	0.8235	0.6250	0.6299
4	0.9375	0.9841	1.0000	0.9000	0.9474	0.8710	0.8783
5	0.8750	0.9524	1.0000	0.8182	0.9000	0.7377	0.7645
6	0.8750	0.9841	0.7778	1.0000	0.8750	0.7538	0.7778
7	0.6875	0.8730	0.7778	0.7000	0.7368	0.3548	0.3578
8	0.8667	1.0000	0.7778	1.0000	0.8750	0.7368	0.7638
9	0.8667	0.9074	0.8889	0.8889	0.8889	0.7222	0.7222
Mean	0.8671	0.9336	0.8667	0.9071	0.8818	0.7297	0.7404
Std	0.0709	0.0499	0.0831	0.0933	0.0602	0.1465	0.1485

Figure 5.10 - Scoregridfor the XGBoost model

By looking at the Figure of the ROC curve Figure 5.11) for the XGBoost classifier we realize that the curves for positive and negative predictions are close to the upper left corner, which means that the overall accuracy of the XGBoost is high. The AUC scores of 0.93 shows that the model is 91% accurate in predicting if a player will get injured or not.

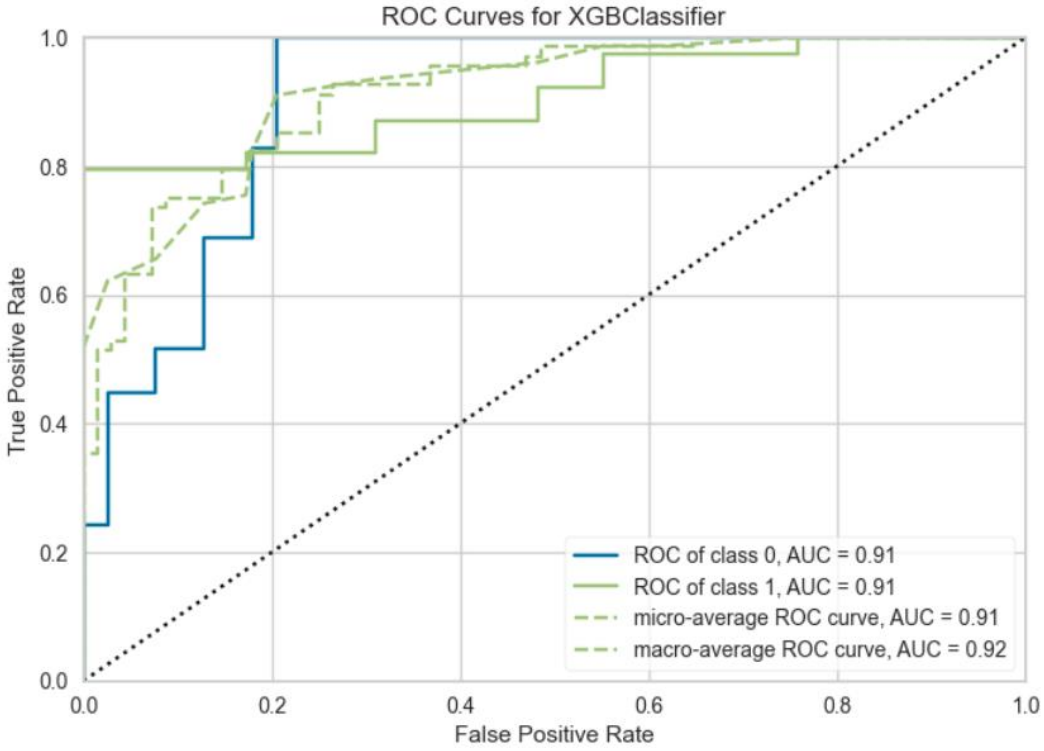


Figure 5.11 - ROC-AUC plot for the XGBoost model.

For the XGBoost algorithm the feature with the most impact is the Matches Lost, followed by ADL Left and Training Exposure (see Figure 5.12).

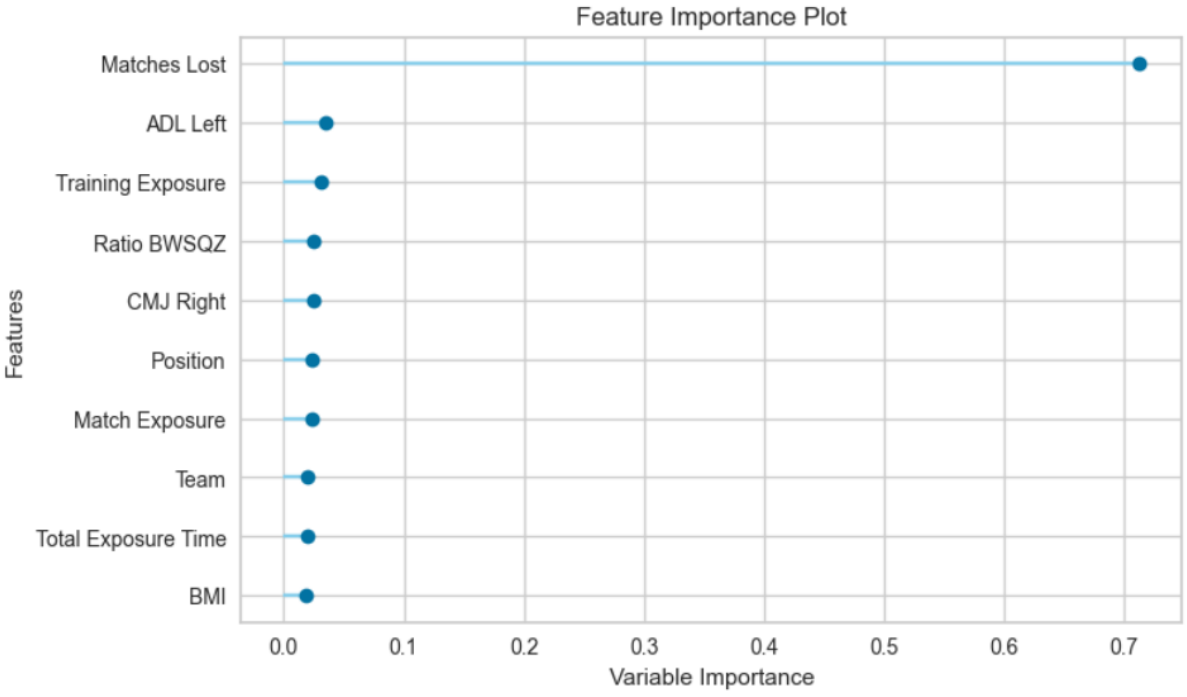


Figure 5.12 - Feature importance plot for the XGBoost model

This was the last model to be created for evaluation and, accordingly to the model comparison score grid it was the one who got the better scores for all categories.

Below (Figure 5.13) we show the overall view of the creation and evaluation of the classifiers DT, KNN, LR and XGBoost.

Classifier	**Evaluation Scores**						**Variable Importance**	
	Accuracy	*AUC*		*Recall*	*Precision*	*F1*	*MCC*	
Decision Trees	0.8171	Class 0 Class 1	0.78 0.79	0.8556	0.8369	0.8420	0.6354	1. Matches lost 2. Match exposure	3. Total Exposure Time 4. ADL Left
K-Nearest Neighbors	0.6188	Class 0 Class 1	0.66 0.67	0.7000	0.6490	0.6642	0.2354	n.a n.a	n.a n.a
Logistic Regression	0.7787	Class 0 Class 1	***0.91*** ***0.91***	0.7444	0.8568	0.7887	0.5719	1. Matches lost 2. Ratio BWSQZ	3. Age 4. Dominant Side
Extreme Gradient Boosting	***0.8671***	Class 0 Class 1	***0.91*** ***0.91***	***0.8667***	***0.9071***	***0.8818***	***0.7414***	1. Matches lost 2. ADL Left	3. Training Exposure 4. Ratio BWSQZ

Figure 5.13 - Scores for the evaluation of the classifiers DT, KNN, LR and XGBoost.

We opted to add another performance measure to the XGBoost and check once again its overall behavior. Because we are dealing with a binary class and because we are testing the model in a test data where we know the true observations, we thought that the Confusion Matrix would be the right scorer to choose.

For this case, there are 68 predictions (size of the test data sample), out of which 37 were predicted has having injury and 31 has not having injury.

According to the terminology used in Classification problems we can analyze the confusion matrix for this study in the following way:

- True Positives – 32, athletes correctly predicted to have injury.
- False Positives – 5, athletes that weren't injured but were predicted to have an injury.
- True Negatives – 24, athletes correctly predicted not to have injury.
- False negatives – 7, athletes that were injured but were predicted has not having injury.

For the matter of this study, which is predicting an injury, the main goal is to find a model that give us lower rates of False Negatives (Injured athletes identified as not having injury).

According to the Confusion Matrix (Figure 5.14), the XGBoost has a 10% probability of mistakenly identifying an injured athlete as non-injured.

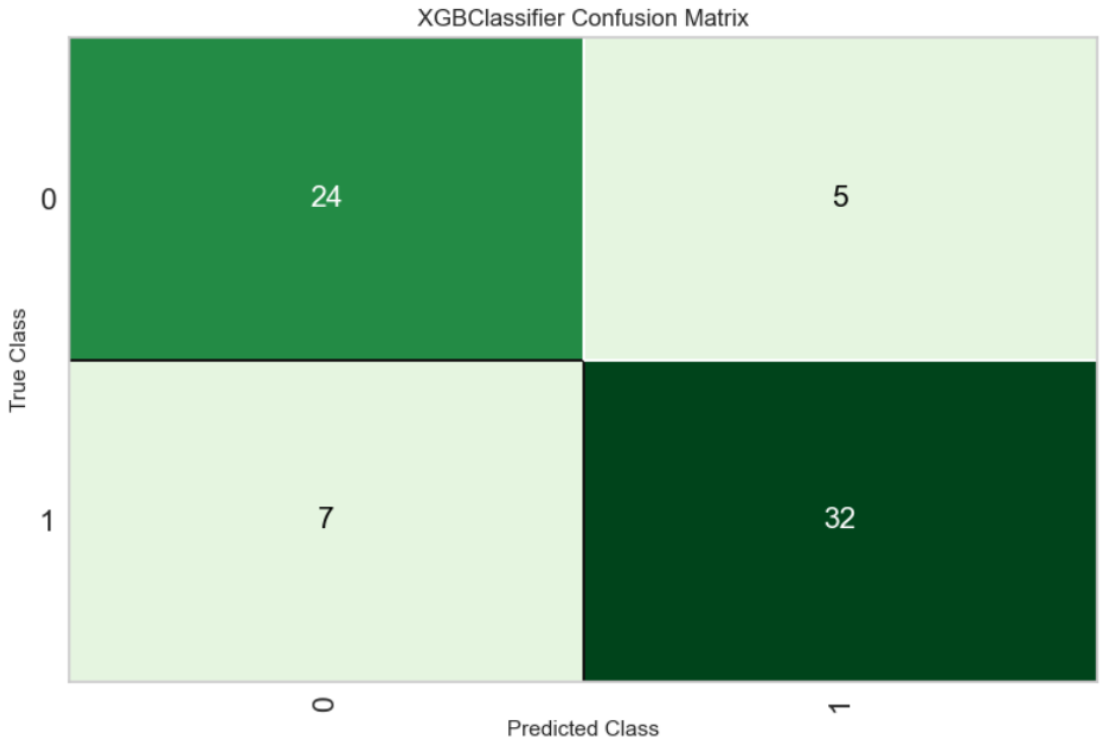


Figure 5.14 - Confusion Matrix plot for the XGBoost model

5.4. Clustering Model Evaluation and Results

For the matter of identifying an injury risk profile we used the k-means clustering which is one of the most powerful and popular unsupervised learning algorithms.

Jain (cited by Ahmed, 2020) stated that "Clustering algorithms exploit the underlying structure of the data distribution and define rules for grouping the data with similar characteristics".

By using the plot_model(model, 'elbow')function we got a distortion score graph (Figure 5.15) that shows the optimal value for *K* in our study, *K* = 5.

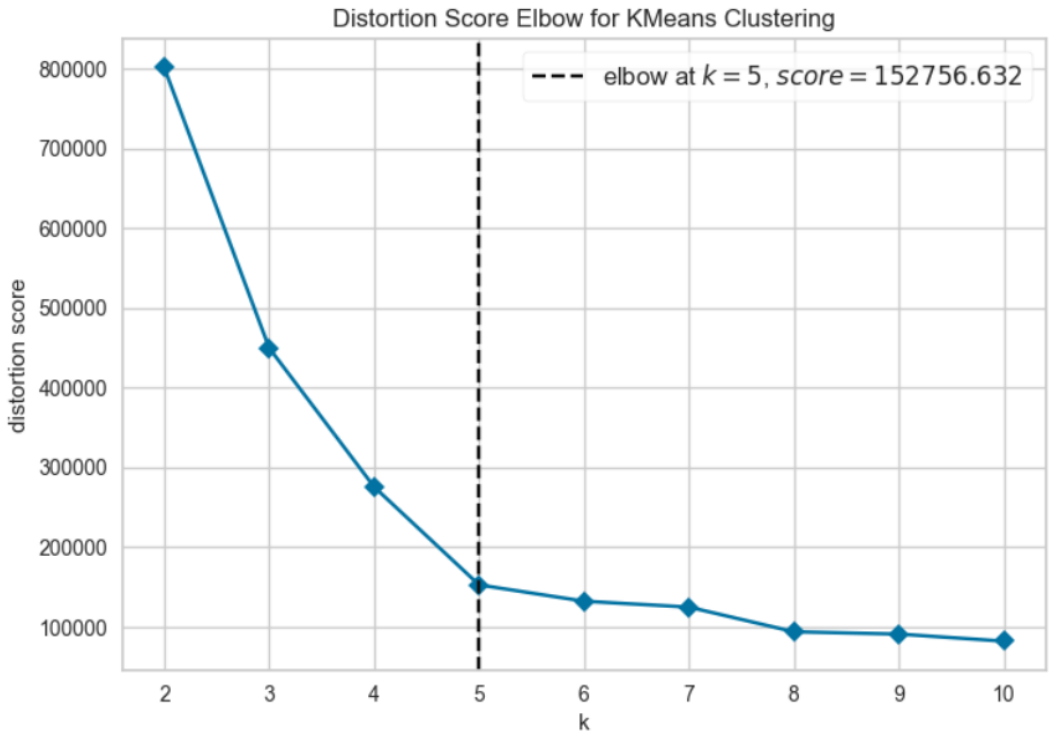


Figure 5.15 - Distortion Score Elbow

To create and evaluate the clustering algorithm we used the create_model ('kmeans', num_clusters=5) function and set the K parameter to 5. By executing this function a few performance metrics were calculated, such as the Silhouette, Calinski-Harabasz and Davies-Bouldin.

We will focus our model clustering analysis on the silhouette score, which is used to determine the quality of the clusters produced by the clustering algorithm. The silhouette score ranges from -1 to 1, and with the threshold commonly set at 0,5. If the score is above 0,5 then the cluster has a good quality, and if the score is below 0,5 then the cluster has low quality. It is worth saying that a score of 1 indicates a perfect fit and -1 a poor match between the data points and their corresponding cluster (Kaufman, 1990).

In our model, the Silhouette score for the K = 5 is 0.51 (see Figure 5.16). Although this score is not close to 1, it is above the threshold and therefore we can consider the model to have a good quality. One approach commonly used in these situations is to run the elbow plot and confirm the optimal number of clusters. We had already done it so the Silhouette had already been calculated with the apropriate number of clusters. According to the Silhouette score, the K = 5 clusters is a good partition.

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.5124	616.2198	0.5618	0	0	0

Figure 5.16 - Silhouette Score for Cluster Analysis

When analysing the Silhouette Plot (Figure 5.17) there are two criteria that should be present to consider the model competent, which are, the presence of clusters above the average score and a not

so wide variation of the cluster silhouette size. Concerning the first criteria the model is considered approved, and although the clusters aren't exactly equal in size, at least 3 are very similar and therefore we will consider aproval for the second criteria.

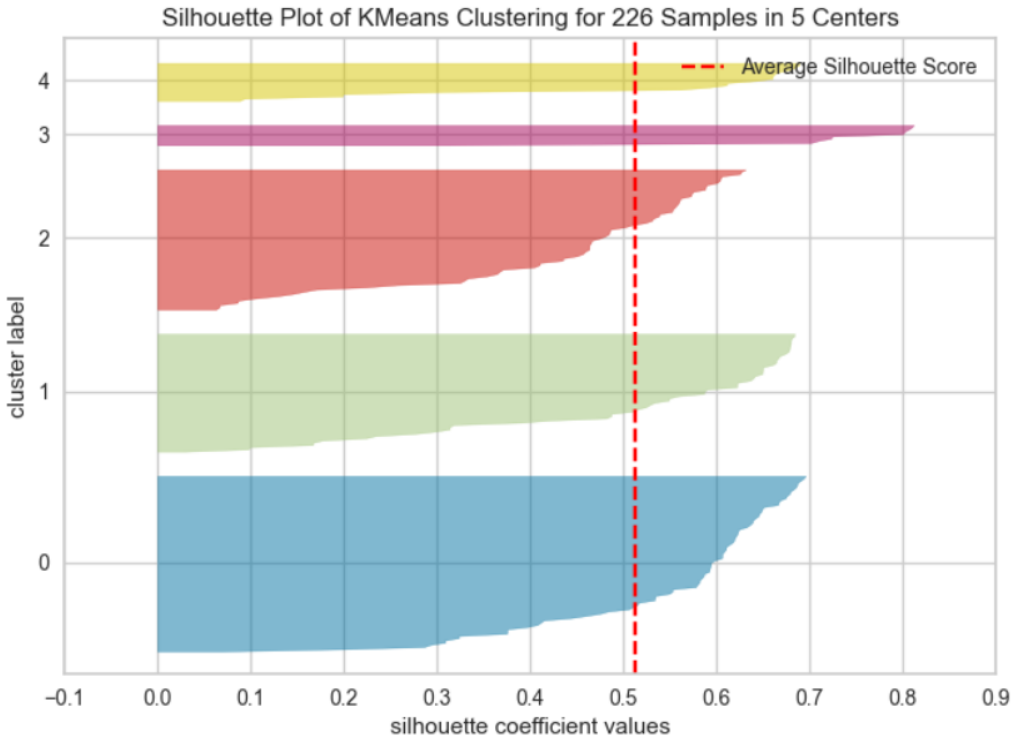


Figure 5.17 - Silhouette Plot for Cluster Analysis

The clusters identified by the algorithm differ in size and proportion of class 0 and 1 observations (Table 5.2 and Figure 5.18).

Cluster 0 is the biggest with 80 observations, out of which 53% are of non-injured athletes and 47% of injured athletes.

Cluster 1 has 54 entries, in which 72% are of injured athletes and the 28% left are non-injured players.

Cluster 2 includes 64 observations, of which 58% are related to injury and 42% to no-injury.

Cluster 3 is the smallest and only has 10 entries, 60% of them are related to no injury and 40% with injury inputs.

Cluster 4 includes 18 observations, 61% are injury related and 39% related to no injury.

Class Value Counts per Cluster

	Class 1 (Injury)	*Class 0 (No Injury)*	*Total*
Cluster 0	38	42	80
Cluster 1	39	15	54
Cluster 2	37	27	64

Cluster 3	4	6	10
Cluster 4	11	7	18

Table 5.2 - Class 0 and 1 Value Counts for the clusters

Figure 5.18 - Cluster distribution Plot

By plotting the cluster Interdistance map (Figure 5.19) we can visualize the average distance between cluster, and this can be seen as a measure of similarity. If the clusters are widely separated, it means that their composition is very different. However, if they are close then they will be similar.

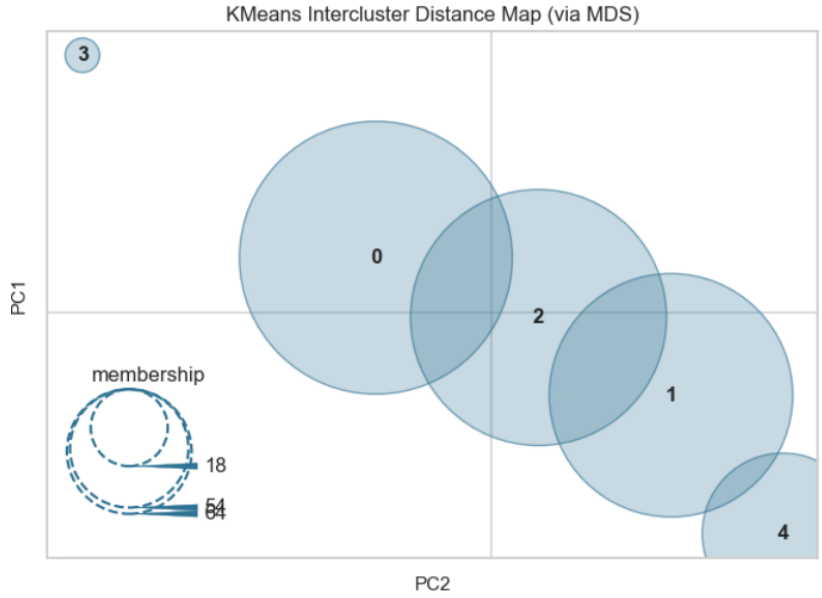


Figure 5.19 - K-Means InterclusterDistance Map

Due to the proximity and similarities among few clusters we opted to plot a PCA function to the K-means model. The purpose was to project the data in a lower dimensional space so that it could be easier to interpret and also to validate if we could lower the number of clusters by grouping the ones with most similarities.

By looking at the graph (Figure 5.20), we get the same information as previously seen on other graphs and therefore we decided not to conduct any other alteration.

Figure 5.20 - 2D Principal Component Analysis (PCA) plot for the K-Means model

6. DEPLOYMENT

The main purpose of any data science project is to extract information that can used on an daily basis to support decision making.

The data mining project described in this study has a very specific population and very diversified stakeholders. Therefore, if a data mining such as this would have to be implemented in a football team the deployment phase would have to be very well discussed among the medical team, sports performance team, coaches and executives, to understand who would have access to such sensitive information. According to our know-how on the way how football clubs proceed most likely we would have to grant different levels of access to information.

For the matter of this study we focus our deployment actions thinking about a real-world football team scenario where medical team, sports performance team and coaches work daily together on planning the trainings and the squad for the matches.

Our proposal includes the following principles:

1) Project Integration;

2) Decision Making Process;

3) Project Maintenance;

4) Project Improvement and growth.

The first two imperative actions that should follow any data mining **project integration** in a sport environment like this are:

- Creation of dashboards and reports;
- Creation of automatic alerts or notifications.

The dashboards with information such as it is demonstrated in this study should be accessible to the medical team, sports performance team, coaches and athletes. Therefore, its visualization should be easily readable and must ensure its usability. The information provided in dashboards should be concise and adequate to the daily basis decisions that will have to be made.

If possible, with any IT solution there should be implemented any sort of alert/notification to proactively inform the medical team, sports performance team and coaches about specific injury risk trends on their players.

The process of making decisions in any sports is influenced by several factors, some of them not so scientifically reliable. Therefore, the insights out of the model should be integrated smoothly in a pre-designed **data-driven decision making workflow**.

The **maintenance of the project** is a very important step because it possibilitate the discovery of other insights on the matter and will also provide the value needed to keep it running. In our opinion, it should be scheduled regularly collaborative opportunities (Eg.: meetings, workshops, brainstorming) between Data Scientists and Domain Experts to connect, communicate and share ideas. These moments of training sessions/workshops are not only reserved for the analysis of the project but also for the education of anyone that uses the information extracted out of the model to make decisions accurately.

Another action that shouldn't be forgotten is how to effectively communicate to the stakeholders responsible for financing the data mining project, how much valuable the insights provided are for the success of the club.

As time goes by, more data enters the through the pipe of data collection and with that **improvements** should be made to the project. It's very easy to keep the focus in sight so is is mandatory the creation of a feedback process to monitor the continuous integration of data into the model and the application of those discoveries into use.

Create a sports specialized data science knowledge center focused on scientific investigation and training.

7. DISCUSSION

Choosing the best model algorithm to predict if an athlete will get injured or not is a complex task and although the algorithms were designed to solve and answer specific questions this is not enough to understand how to choose the one that fits our problem.

In 2021, Costa attempted to use the MELCHIOR methodology (as proposed by Leclercq in 1984) to support the decision to choose a classifier algorithm. For his study he found that the best algorithm was Gradient Boosting Decision Tree and that the methodology used to assess it can "be used to solve problems of various types, considering that it presents a simple, flexible, reliable and fast methodology".

Zhongguo (2017) proposes a method to choose a classification algorithm based on Optimum parameters and dataset characteristics.

According to the (extensive) literature there isn't a reference methodology on which every ML project should support when it's about choosing the best classifier (Chicco, 2020). Therefore, we will present and discuss our choice for the most suitable algorithm for our prediction problem, based on the problem we want to solve, the characteristics of the dataset and the quantitative evaluation metrics obtained.

In this study, there are roughly almost equal number of positive and negative samples (43% no-injury and 57% injury), therefore is considered a balanced dataset.

The purpose of the case study is to choose a binary classification algorithm to predict whether an young football player will get injured (Positive Class) or not get injured (Negative class).

If our goal is to predict an injury, and if that injury is well predicted (*True Positive*) then we can take some action to prevent it. However, mistakenly identifying an athlete has having an injury when he does not, it is not that important because it doesn't cause any damage to the patient. Therefore, the cost of a *False Positive* is not that high. The prediction that may come with bigger costs for the team and the athlete is the identification of a *False Negative*. Before we identify the most suitable score to use when evaluating the classifier models, first we had to weight the importance of the predictions based on our goal.

Accuracy measures assertiveness when predicting positive instances and is considered to be a reasonably good performance metric, however it's is most suitable for multiclass cases and when the datasets are well balanced (Wang, Sokolova, Gu Q, Bekkar, Akosa cited by Chicco, 2020). In this case we want to evaluate a supervised binary classifier with balanced datasets, and therefore Accuracy is an appropriate metric.

The model with the best accuracy scores was the **XGBoost**, correctly predicting 88% of the times if an athlete will be injured or not.

For the Precision scores, the model with the best performance was also the **XGBoost**, with a 90% of exactness in identifying positive classes (injured athletes).

As Powers (2011) states, both Recall and Precision are measures focused on positive predictions, although they also provide some information about the rate and the nature of the errors of prediction that it is also very useful. The predictions errors are called False Negatives and False Positives, and as previously discussed the wight of having a False Negative in this study is higher than having a False Positive.

The score measure that is more affected by the presence or not of False Negatives is the Recall, meaning that if the models deliver a lot of False Negatives, the recall value will be low. Therefore, between Precision and Recall, the best metric to look at and evaluate our model would be Recall.

Again, the **XGBoost** is the model that performed better (sensitivity of 87%) on identifying injured athletes with low rate of mistakenly identifying non-injured athletes.

The perfect setting would be one where both Precision and recall would be of 1.0, although this is not possible because if we want to improve the positive classification, we will lose sensitivity in identifying negative instances and therefore the score for recall will lower.

According to this we would say that the best metrics to evaluate our model's performance would be Accuracy and Recall.

We can get over limitations on these two metrics by using the F1 score and adjusting the f-score to give more importance to recall. However, the lack of expertise by the research team could give rise to a mistaken conclusion. Nevertheless, just by analysing the F1 score we can tell that the **XGBoost** was the top scorer with 0.8818.

Wang (2021) refers to Chicco and Jurman (2020) by saying that "only relying on accuracy to evaluate the performance of machine learning models may lead to overoptimistic results, especially on imbalanced datasets".

Although we have a pretty well-balanced data set, there's a characteristic that imputates more importance in a label of the confusion matrix, which is the False Negatives. Therefore, we should consider a model that we know for sure will reduce the number of FN and the only way we can evaluate this is by using a metric that fully considers the size of the four classes of the confusion matrix.

Chicco (2017) said that "even if accuracy and F1 score are widely employed in statistics, both can be misleading" and the key to avoid these "dangerous misleading illusions", is to exploit the MCC performance score.

The latest study of Chicco (2023) suggests that for assessing binary classification, the MCC should replace the ROC-AUC as the standard metric. Moreover, a high MCC will always correspond to a high AUC but the opposite is not always true.

Nevertheless, all the metrics used for assessing the performance of the DT, KNN, LR and XGBoost were undoubletly in favor of this last classifier, the **XGBoost**.

This result follows Arif Ali's beliefs on the power of the XGBoost. In 2013 this author released a paper where he claims that XGBoost is the "most proeminent machine learning algorithm in a Python environment" for supervised and semi-supervised learning. Although there may be some algorithms specific to address problems in unstructured data (Articial Neural Networks) such as images and text, and others specific for small to medium structured data (decision trees), the Gradient Boosting concept is outperforming them by improving a single weak model by combining it with any other weak model to create a stronger model, in an iterative way (Arif Ali, 2013).

By analysing the Feature importance plot for all the classifiers evaluated, the variable that is present in all of them is the **Matches Lost**.

This variable is a feature included in the dataset that should characterize the consequences of an injury and it refers to the number of matches lost due to injury such as the Trainings Lost.

Recalling the reasoning for the data preparation stage, we dropped a few variables that were making every model tested overfitting the data. This happened because there were a lot of variables related to injury that the model couldn´t distinguish and put them in a sub-category of injury. Therefore, we decided to drop all the features that should characterize an injury and just leave the Type of Injury, matches lost and training lost. Following this reasoning we realized that for the type of injury there were some labels with only one instance and that was biasing the model. Due to this we grouped the Type of injury labels into just two classes, 1 – injury or 0 – no injury. This way we were able to work on cleaner data. However, the model only relates the features with the target variable and does not know whether it causes it or if it is a consequence. In our case, the model doesn't understand if any of the variables are responsible for having an injury or if they related because of the burden of the injury. Concluding, the feature that most impacts having an injury is the number of matches lost is a biased information that we, as domains specialist, should identify and interpret as not correct.

For what we said, the most advisable is to look at how much the other variables impact the model. As seen above, our preferred model is the XGBoost which shows us that what influences the most an injured athlete is the Training exposure and the ADL Left.

Also identified in DT model, the features that most impact an injured athlete are the ones related to the stress imposed on the athlete during training and matches (Total exposure time, training exposure and match exposure). These results are not in accordance to the findings of Luis (2021) about the workload as being an important factor in injury aetology. Still, we must state that the workload variables with which we dealt in this study only reflect the number of hours that the players are exposed to practicing football. In his study Luis found no differences between study groups, suggesting that workload alone cannot predict or explain injury. As domains specialists we totally agree with him when he says, "workload management is an important piece of the injury prevention puzzle that cannot be ignored nor dismissed".

By analysing the results of the K-Means Clustering we found some similarities between the clusters and realized that for the group of people that this dataset refers to, we can't differentiate it into

completely different groups. It would be considered perfect if these clusters had strict boundaries and rules that could aggregate each new entry clearly and with no doubt. Although, the findings don´t allows us to create an injury risk profile the characterization of the clusters can help us find some pattern and reasoning about a risk index.

According to the literature available about the sports injuries and screening tests, the amount of data needed to conduct a study of such nature is much bigger than what we had access to. On the other hand, some systematic reviews for injury prediction and prevention (Foreman et al. 2006; Maffey & Emery 2007; Shadle & Cacolice 2016; Van Beijsterveldt et al. 2013 cited by Christopher, 2021) concluded that there isn't a scientific valid agreement on the best musculoskeletal screening tools.

Nowadays it is accepted that screening tests do not predict injury by themselves due to multifactorial causation of injury, which is very complex and involves very diverse aspects of the athlete that cannot be assessed just by undertaking specific physical tests. Nevertheless, Bahr (2016) admits that screening tests may not predict injury, but they can help categorize athletes into different risk levels (cited by Luis, 2021).

Bittencourt (2016) and Fonseca (2012) are two of the main authors that defend the implementation of a model where the complex interaction between all the factors affecting the athlete (web of determinants) at a specific moment predisposes the athlete to an injury. This individual pattern is considered to be an injury risk profile (Galea, 2010).

Meeuwise (2007) describes perfectly what the science is trying objectivate

> "the nature injury in sport is different. First, exposure is a combination of both possessing a risk factor and then participating (to a greater or lesser degree) with the risk factor. An individual may be exposed to the same or different risk factors repeatedly through multiple participations. Injuries may or may not occur under similar conditions".

With the results obtained with this ML algorithm we will try to extract some knowledge about the dataset and enlighten what may be a risk profile for the athletes of this population. By what it has been described before it can't be our purpose to identify a risk profile that should be spread and applied to all new populations of youth football players. What we can is to define the injury risk profiles for the population that we are studying and assess if the methodology used is apropriate to be applied in different settings.

To create the specific injury profile, we opted to create a simple powerful clustering algorithm that could help us identify some clusters. By using the K-means clustering method we obtained five clusters that are not so dissimiliar among each other as we were hoping. Which, in fact makes sense according to the complexity of the aetology of injuries in sports.

So that we turn the findings into something more objective and applicable to the everyday practice we decided to rename the clusters according to the probability of having an injury (or not) in that group of athletes:

- For cluster 0, there are 42 non-injured athletes and 38 injured, which means that athletes with this profile have 47,5% of sustaining an injury.
- For cluster 1, we have 39 injured athletes that corresponds to a 72% likelihood that athletes with this profile will have an injury.
- Cluster 2 has 37 injured players out of 64 that corresponds to 58% of players with injury with this profile.
- In cluster 3 the total amount of observations is the lowest (n = 10) and 60% are players with injury.
- For cluster 4 there are 18 observations and 61% correspond to athletes that sustained an injury.

According to these scores we renamed the cluster as:

- Risk Level 1 > Cluster 3
- Risk Level2 > Cluster 0
- Risk Level3 > Cluster 2
- Risk Level4 > Cluster 4
- Risk Level5 > Cluster 1

We know that is not advisable to extrapolate these results to bigger populations because the behaviors of such algorithms are not linear, nevertheless, with the information that we have we can draw some injury risk profiles for this population.

The “Risk Level 1” group is composed of athletes with ranging ages from 14 to 22 years old with a mean BMI of 20,7 (normal), majority of players are midfielders, right-handed, with an average total exposure time of 38.95 hours.

The “Risk Level 2” group includes athletes with ages ranging from 13 to 21 years old, with a mean BMI of 19,39 (normal), most players are midfielders, right-handed with a total exposure time of 164,64 hours.

For the “Risk Level 3” the ages range between 15 and 22 years old, with a mean BMI of 21,27 (normal), mostly wingers and right-handed with a total exposure time of 216,64 hours.

“Risk Level 4” group has athletes with ages ranging from 14 to 19 years old, an average BMI of 21,39 (normal), most of them are midfielders, right-handed with a total exposure time of 337,10 hours.

The highest risk level group (level 5) includes players with an age range from 14 to 19 years old, an average BMI of 21,55 (normal), the majority of them are right-handed wingers with total exposure time of 262, 99 hours.

As we predicted the groups have similarities but the finding that is most prominent is that as the injury risk level increases so does the total exposure time. This is in accordance to the findings on the classification model where we observed that the features with most impact in the injury were related to the athletes exposure time in training and in matches.

The characterization of each cluster according to the average physical scores obtained in the screening is depicted in table 7.1, which surprisingly shows that the clusters we considered as having less risk of injury are the ones with overall lower values for the strength, power and range of motion tests.

Cluster	Ratio BWSQZ	CMJ Left	CMJ Right	PKFO Left	PKFO Right	ADL Left	ADL Right
	Mean	*Mean*	*Mean*	*Mean*	*Mean*	*Mean*	*Mean*
Risk Level 1	0.58	19.00	20.60	24.70	25.80	7.30	7.30
Risk Level2	0.68	18.02	18.35	28.30	28.58	9.11	8.81
Risk Level 3	0.65	18.37	18.50	28.06	27.57	8.82	8.39
Risk Level 4	0.64	18.38	18.83	28.55	28.50	8.55	8.94
Risk Level 5	0.67	20.85	20.59	26.61	26.68	8.68	8.25

Table 7.1 - Mean values for the Musculoskeletal screening tests present in the train and test samples, grouped by cluster

Wrona (2023) conducted a study in which found that, for the one leg CMJ, asymmetries at the 20% threshold are indicative of youth male soccer player having adduction asymmetry, and therefore increased risk of injury, specially in the hip and groin area. No differences over 20% were found intra-cluster nor inter-cluster.

According to Tak (2015), a “decreased HROM in professional soccer players is associated with more hip- and groin-related symptoms and with previous injuries, independent of the presence of a cam deformity” and with a linear causational thinking we would expect higher scores for the PKFO in the Risk Level 5 group and that is not the case.

Van Dyk conducted an impact factor of 6.1 study (2018) about hamstring injuries, and he concluded that deficits in both passive hamstrings and ankle dorsiflexion range are weak risk factors for hamstring injuries and suggests that any athletic screening tool should be used very cautiously due to its low clinical value.

In 2018 Terry examined if Musculoskeletal Readiness Screening Tool (MRST) scores were predictive of student athletes sustaining an injury. This MRST is a screening tool composed of six functional movements and results indicated that MRST scores were not predictive of injury. Although the screening tests and the population is not the same, this is also evidence that screening test by themselves are not predictors for sports injuries.

Talukder (et al., 2016 cited by Caparros 2018) proposed that average speed, number of games, number of games played, average distance, average game time, and average number of shots may be important variables in predicting injury risk for professional basketball.

One other study that confirms our findings about the workload influence on injuries is the one conducted by Rossi et al. (2018) where he constructed an injury prediction model based on GPS data. By using decision tree algorithms, he could successfully predict approximately 80% of the non-contact injuries

Most authors only study some factors of the injury aetiology due to limitations of conventional statistical methods (Huang, 2022) and insufficient data. Therefore, they may find some injury risk pattern but will never get the complete profiling of an injury risk athlete (Waterkamp, 2016; Bello et al., 2020; Isern-Kebschull et al., 2020, cited by Huang, 2022)

For Huang (2023) further studies should include information on training load, perceived well-being status, physiological response, physical performance, sleep (Huang, 2021) and nutrition (Mason et al., 2023), specially in adolescent athletes.

8. CONCLUSIONS

This study adds some knowledge to the field of screening instruments and the use of Machine Learning methods for injury prediction. As it is implicit in all the extensive literature on this matter, it also concludes that injury prediction is complex and multifactorial and may involve more than just physical components.

The screening tools proposed by FIFA to monitor athletes were created on the assumption that a longitudinal and repeated measure of data is enough to create a model capable of fully utilizing this information and revealing specific injury patterns. Much more than consecutive data measurements, what matters the most is that data can be derived from multiple sources that covers all aspects of the athlete's overall health and well-being.

This study created and evaluated a few algorithms and concluded that Extreme Gradient Bosting is the model with best performance for identifying athletes with injury.

By using a K-means method for clustering this specific dataset we were able to identify five clusters that served as the basis to create injury risk profiles. However, the profiles clustered didn't have as much disparity as we were hoping to, and we believe that the main reason for this was because of the sparse data information held by the dataset. Nevertheless, the K-means algorithm showed to be a good method to be considered for the matter of injury risk profiling.

The model wasn´t applied and tested in a real-world sports situation, however we believe this study offers valuable insights into future work on injury prediction and prevention due to its easy interpretation of injury risk profile. We believe that along with all the scientific research this study can promote some knowledge about the creation of an injury risk index that can be used commonly by all the stakeholders involved in the football sports industry.

Limitations and Future work

The main limitation to this study was the information contained in the dataset. As it has been demonstrated several times in this study, sports injuries aetology is multi-factorial and complex due to the various and changing over time relationships between injury determinants. It is recommended that further studies may use extensive multi-source data covering as many aspects as possible of the athlete's health and performance. As evidenced in this study the exposure time is a considerable injury risk factor, and we believe that if could have accessed more data on workload variables such as training and match intensity, weekly cumulative ratings of perceived exertion, acute: chronic workload ratio, monotony and strain, we could have had different results for the clustering conclusions.

Although the data set used in this study covered almost all youth athletes of the club, the time range of the observations was too short to comprise enough data for more valid injury risk profiling. Therefore, we suggest that for future studies the data could include several competitive seasons.

For future work we not only recommend expanding data dimensions, but we also recommend some research on classifying and profiling the risk of specific injuries. In our study we aggregated all the types of injuries and focused only on the athlete having injury or not. For the future we would apply the same reasoning and methods for specific injury types.

BIBLIOGRAPHY

Ahmed, M., Seraj, R, Islam, S. (2020). The k-means Algorithm: A Comprehensive Surveyand Performance Evaluation. Electronics 9(8), 1295.

Arif Ali, Z., Abduljabbar, ZH. (2023). Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. Academic Journal of Nawroz University (AJNU), Vol.12, No.2.

Associação Futebol de Lisboa. (2023). Regulamento de provas oficiais.

Ayala F., López-Valenciano, A., Gámez, J. (2019) A Preventive Model for Hamstring Injuries in Professional Soccer: Learning Algorithms. Int JSports Med 40:344–353.

Ayodele, T. (2010). Types of Machine Learning Algorithms. University of Portsmouth United Kingdom.

Bahr, R. (2016). Why screening tests to predict injury do not work - and probably never will…: a critical review. Br J SportsMed. 50:776–780.

Bahr, R., Holme, I. (2003). Risk factors for sports injuries - a methodological approach. Br J Sports Med. 37:384–392.

Baker, S., O'Neill, B., Karpf, R. (1992). The Injury Fact Book. New York, NY: Oxford University Press.

Balsalobre-Fernandez, C. (2014). The concurrent validity and reliability of a low-cost, high-speed camera-based method for measuring the flight time of vertical jumps. Journal of Strength and Conditioning Research. 28(2)/528–533.

Bertanffaly, V. (1969). General system theory: foundations, development, applications. Penguin university books. New York, USA: George Braziller Inc.

Bittencourt, N., Fonseca, S.(2016). Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. Br JSports Med;50:1309–1314

Boyd, S., Mohri, M. (2012). Accuracy at the top. Advances in Neural Information Processing Systems 2. January.

Calligeris, T., Burgess, T. (2015). The incidence of injuries and exposure time of professional football club players in the Premier Soccer League during football season. SAJSM Vol. 27, N. 1.

Caparrós, T., Casals, M. (2018). Low External Workloads Are Related to Higher Injury Risk in Professional Male Basketball Games. Journal of Sports Science and Medicine. 17, 289-297

Carey, D., Ong, K. (2017). Predictive Modelling of Training Loads and Injury in Australian Football. International Journal of Computer Science in Sport. 17.

Carneiro, J. (2017). Acidentes de Trabalho dos Jogadores de Futebol. Publicações AJJ.

Chapman, P., Clinton, J. (2000). CRISP-DM 1.0. Step-by-step data mining guide.

Chicco, D. (2017). Ten quick tips for machine learning incomputational biology. Biodata Mining. 10:35.

Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 scoreand accuracy in binary classificationevaluation. BMC Genomics. 21:6.

Chicco, D., Jurman, G. (2023). The Matthews correlation coeficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. Biodata Mining. 16:4.

Chmait, N. Westerbeek, H. (2021) Artificial Intelligence and Machine Learning in Sport Research: An Introduction for Non-data Scientists. Front. Sports Act. Living 3:682287.

Christopher, R., Brandt, C., Benjamin-Damon, N., (2021). Systematic review of screening tools for common soccer injuries and their risk factors. South African Journal of Physiotherapy.77(1).

Claudino, J., Capanema, D. (2019). Current Approaches to the Use of Artifcial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. Sports Med - Open 5:28.

Costa, I., Basilio, M., Maeda, S., Rodrigues, M. (2021). Algorithm Selection for Machine Learning Classification: An Application of the MELCHIOR Multicriteria Method. Modern Management based on Big Data II and Machine Learning and Intelligent Systems III. A.J. Tallón Ballesteros (Ed.)

Cunningham, P., Delany, S. (2007). k-Nearest neighbour classifiers. Mult Classif Syst. 54. 10.1145/3459665.

Decker, T., Gross, R. (2023). The Thousand Faces of Explainable AI Along the Machine Learning Life Cycle: Industrial Reality and Current State of Research.

Donahue, P., McInnis, A. (2023). Examination of Countermovement Jump Performance Changes in Collegiate Female Volleyball in Fatigued Conditions. J. Funct. Morphol. Kinesiol. 8, 137.

Drawer, S., Fuller, CW. (2002). Evaluating the level of injury in English professional football using a risk based assessment process. Br J Sports Med. Dec;36(6):446-51

Eetvelde, V. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. J Exp Ortop. 8:27.

Eetvelde, V. Mendonça, L. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. J EXP ORTOP. 8:27.

Ekstrand, J., Hagglund, M., Walden, M. (2009). Injury Incidence and Injury Patterns in Professional Football: The UEFA Injury Study. British Journal of Sports Medicine, 45, 553-538.

Ekstrand,J., Hägglund, M., Waldén, M. (2009). Injury incidence and injury patterns in professional football - the UEFA injury study. British journal of sports medicine, 060582.

Elsafty, O., Berkey, C., Dauskardt, R. (2024). Insights and mechanics-driven modeling of human cutaneous impact injuries. Journal of the Mechanical Behavior of Biomedical Materials. 153.

Fiscutean, A. (2021). Could an algorithm predict an injury? Nature, Vol 592, 10–11.

Fouasson-Chailloux, A., Mesland, O., Menu, P., Dauty. M. (2020). Soccer injuries documented by F-MARC guidelines in 13- and 14-year-old national elite players: A 5-year cohort study. Science & Sports, 35, pp.145 - 153.

Franco, M., Madaleno, F. (2021). Prevalence of overuse injuries in athletes from individual and team sports: A systematic review with meta-analysis and GRADE recommendations. In Brazilian Journal of Physical Therapy (Vol. 25, Issue 5, pp. 500–513).

Fuller, C., Ekstrand, J. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. Br J Sports Med;40:193–201.

Galea, S., Riddle, M. (2010). Causal thinking and complex system approaches in epidemiology. International Journal of Epidemiology. 39:97–106.

Ghahramani, Z. (2015) Probabilistic machine learning and artificial intelligence. Nature 521:452–459.

Gigerenzer, G., Gaissmaier, W. (2011). Heuristic decision making. Annu Rev Psychol. 62:451-482

Glathorn, J., Gouge, S. (2011). Validity and reliability of optojump photoelectric cells for estimating vertical jump height. Journal of Strength and Conditioning Research - Technical Report.556-560.

Halabchi, F., Angoorani, H., Mirshahi, M. (2016). The Prevalence of Selected Intrinsic Risk Factors for Ankle Sprain Among Elite Football and Basketball Players. Asian journal of sports medicine, 7(3).

Han, J., Pei, J., (2011). Data mining: concepts and techniques. Amsterdam: Elsevier.

Hernández, A., Hidalgo, D. (2020). Detection of Fraud Patterns in Accounting Accounts Using Data Mining Techniques. Open Journal of Business and Management, 8, 1609-1618.

Hickey, J., Shield, A., (2014). The financial cost of hamstring strain injuries in the Australian Football LeagueBritish Journal of Sports Medicine; 48:729-730.

Horvat, M., Jovic, A., Burnik, K. (2021). "Assessing the Robustness of Cluster Solutions in Emotionally-Annotated Pictures Using Monte-Carlo Simulation Stabilized K-Means Algorithm". Machine Learning and Knowledge extraction 3, n2:435-452.

Huang, K., Ihm, J. (2021). Sleep and Injury Risk. American College of Sports Medicine. Volume 20, number 6.

Huang, Y., Bai, Z., Wang, Y., Ye, X. (2023). A novel approach for sports injury risk prediction: based on time-series image encoding and deep learning. Front. Physiol. 14:1174525

Huang, Y., Huang, S., Wang, Y., Li, Y., Gui, Y. (2022), A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. Front. Physiol. 13:937546.

Iqbal, H. (2021). Machine Learning: Algorithms, Real-World Applications and Research. SN Computer Science 2:160.

Iqbal, T., Elahi, A. (2022). Exploring Unsupervised Machine Learning Classification Methods for Physiological Stress Detection. Front. Med. Technol. 4:782756

Jovanovic, M. (2017). Uncertainty, heuristics and injury prediction. Aspetar Sports Medicine Journal - Screening and Prevention.

Joyce, D. Lewindon, D. (2015). Injury risk profiling process. Sports Injury Prevention and rehabilitation. Chapter 6.

Jurman, G., Riccadonna, S., Furlanello, C. (2012). A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. PLoS ONE 7(8).

Kampakis, S. (2016). Predictive modeling of football injuries. Department of Computer Science University College London

Karimi, Z. (2021). Confusion Matrix. Counseling Psychology.

Lapham, A., Bartlett, R. (1995) The use of artificial intelligence in the analysis of sports performance: A review of applications in human gait analysis and future directions for sports biomechanics, Journal of Sports Sciences, 13:3, 229-237.

Lartey, Franklin. (2020). Chaos, Complexity, and Contingency Theories: A Comparative Analysis and Application to the 21st Century Organization. Journal of Business Administration Research. 9. 44.

Laureano, R., Caetano, N. (2014). Previsão de tempos de internamento num hospital português: aplicação da metodologia CRISP-DM. RISTI, N.º 13.

Loose, O., Achenbach, L., Fellner, B., Lehmann, J., Jansen, P., Nerlich, M., Angele, P., Krutsch, W. (2018). Injury prevention and return to play strategies in elite football: no consent between players and team coaches. Archives of Orthopaedic and Trauma Surgery. 138.

López-Valenciano, A., Ruiz-Pérez, I., Garcia-Gómez, A., Vera-Garcia, FJ., De Ste Croix, M., Myer, GD., Ayala, F. (2020). Epidemiology of injuries in professional football: a systematic review and meta-analysis. Br J Sports Med. Jun;54(12):711-718.

Luis, A. (2021). Youth Football Injury Epidemiology: a prospective study on workload and musucloskeletal Risk Factors. Master Thesis, Faculdade de Motricidade humana de Lisboa.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., & Sheth, A. P. (2018). Machine learning for internet of things data analysis: A survey. Digital Communications and Networks, 4(3), 161–175.

Mahdavinejad, MS., Rezvan, M., Barekatain, M., Adibi, P., Barnagh, P., Sheth, AP. (2018). Machine learning for internet of things data analysis: a survey. Digit Commun Netw. 4(3):161–75.

Mahesh, B. (2019). Machine Learning Algorithms -A Review. International Journal of Science and Research (IJSR). 9.

Majundar, A., Bakirov, R. (2022). Machine Learning for Understanding and Predicting Injuries in Football. Sports Medicine - Open. 8:73.

Mandorino, M., Figueiredo, AJ., Gjak,a M., Tessitore A. (2023). Injury incidence and risk factors in youth soccer players: a systematic literature review. Part II: Intrinsic and extrinsic risk factors. Biol Sport. 40(1):27–49.

Mandorino, M., Figueiredo, AJ., Gjak,a M., Tessitore A. (2023). Injury incidence and risk factors in youth soccer players: a systematic literature review. Part I: epidemiological analysis. Biol Sport. 40(1):3–25

Manoel, Lucas & Xixirry, Marcela & Saad, Marcelo & Riberto, Marcelo. (2020). Identification of Ankle Injury Risk Factors in Professional Soccer Players Through a Preseason Functional Assessment. Journal of Sport and Exercise Psychology. 8.

Mason, L., Connolly, J., Devenney, L.E. Lacey, K. (2023). Sleep, Nutrition, and Injury Risk in Adolescent Athletes: A Narrative Review. Nutrients, 15, 5101.

Matheson, G. O., Mohtadi, N. G., Safran, M., & Meeuwisse, W. H. (2010). Sport injury prevention: Time for an intervention? Clinical Journal of Sport Medicine, 20(6), 399–401.

McCall, A., Fanchini, M., & Coutts, A. J. (2017). Prediction: The Modern-Day Sport-Science and Sports-Medicine “Quest for the Holy Grail”. International Journal of Sports Physiology and Performance, 12(5), 704-706.

McCallum A. (2005). Information extraction: distilling structured data from unstructured text. Queue. 3(9):48–57.

Meeuwisse, W. (1994). Assessing causation in Sport Injury: a multifactorial model. Clinical Journal of Sports Medicine, 4:166-170. Raven Press., New York

Meeuwisse, W., Tyreman, H. (2007). A Dynamic Model of Etiology in Sport Injury: The Recursive Nature of Risk and Causation. Clin J Sport Med. Volume 17, Number 3.

Mohammed, M., Khan, MB., Bashier, M. (2016). Machine learning: algorithms and applications. CRC Press.

Moreno-Perez, V., Travassos, B. (2019). Adductor Squeeze Test and Groin Injuries in elite Football players: A prospective study, Physical Therapy in Sport (2019).

Natekin, A., Knoll, A. (2013). Gradient Boosting machines, a tutorial. Frontiers in neurorobotics. Volume 7:21.

Nevill, A., Holder, R., Watts, A. (2009). The changing shape of “successful” professional footballers. Journal of sports sciences. 27. 419-26.

O’ Connor, C., McIntyre, M., Delahunt, E. (2023). Reliability and validity of common hip adduction strength measures: The ForceFrame strength testing system versus the sphygmomanometer, Physical Therapy in Sport, Volume 59, Pages 162-167.

Oliver, JL., Ayala, F., Lloyd, RS., Myer, GD., Read, PJ. (2020). Using machine learning to improve our understanding of injury risk andprediction in elite male youth football players. J Sci Med Sport.

Park, H. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. Journal of Korean Academy of Nursing. 43. 154-164.

Passfield, L., Hopker, J. G. (2017). A Mine of Information: Can Sports Analytics Provide Wisdom From Your Data?. International Journal of Sports Physiology and Performance, 12(7), 851-855.

Pereira, L., Lopes. R. Louçã, J., Araújo, D., Ramos, J.(2021). The soccer game, bit by bit: An information-theoretic analysis, Chaos, Solitons & Fractals, Volume 152, 111356.

Pérez Castilla, A., Rojas, F., Fernandes, J. (2021). Reliability and Magnitude of Countermovement Jump Performance Variables: Influence of the Take-off Threshold. Measurement in Physical Education and Exercise Science. 25.

Petrigna, L., Karsten, B., Marcolin, G., Paoli, A., D’Antona, G., Palma, A. Bianco, A. (2019) A Review of Countermovement and Squat Jump Testing Methods in the Context of Public Health Examination in Adolescence: Reliability and Feasibility of Current Testing Procedures. Front. Physiol. 10:1384

Philippe, P., Mansi, O. (1998). Nonlinearity in the epidemiology of complex health and disease processes. Theor Med Bioeth. Dec;19(6):591-607

Piłka, T., Grzelak, B., Sadurska, A, Górecki T, Dyczkowski K. (2023). Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors. Sensors (Basel). Jan 20;23(3):1227.

Plsek, PE., Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. BMJ. Sep 15;323(7313):625-8.

Powers, D., Ailab, G. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981.

Prusty, S., Patnaik, S. (2022). SKCV: Stratified K-fold cross-validationon ML classifiers for predicting cervical cancer. Front. Nanotechnol. 4:972421.

Quarrie, K., Alsop, J., Waller, AE., Bird, YN., Marshall, S., Chalmers, D. (2001). The New Zealand rugby injury and performance project. VI. A prospective cohort study of risk factors for injury in rugby union football. British journal of sports medicine. 35. 157-66. 10.1136/bjsm.35.3.157.

Rácz, A., Bajusz, D., Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. Molecules, 26.

Rein, R., Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus, 5:1410

Rejani, Y., Thamarai, S. (2009). Early Detection of Breast Cancer using SVM Classifier Technique. International Journal on Computer Science and Engineering. 1.

Rickles, D., Hawe, P., & Shiell, A. (2007). A simple guide to chaos and complexity. Journal of Epidemiology and Community Health, 61(11), 933-937.

Rommers, N., Rössler, R., Verhagen, E., Vandecasteele, F., Verstockt, S., Vaeyens, R,. Lenoir, M., D'Hondt, E., Witvrouw, E. (2020) A Machine Learning Approachto Assess Injury Risk in Elite Youth Football Players. Med Sci Sports Exerc.

Rossi, A., Pappalardo, J. (2018). Efective injury forecasting in soccer with GPS training data and machinelearning. PLoS ONE 13:e0201264.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, FM., Fernàndez, J., Medina, D. (2018) Effective injury forecasting in soccer with GPS training data and machine learning. PLoS ONE 13(7): e0201264.

Ruddy, J., Cormack, S.J, Whiteley, R., Williams, MD., Timmins, RG. Opar, DA. (2019) Modeling the Risk of Team Sport Injuries: A Narrative Review of Different Statistical Approaches. Front. Physiol. 10:829

Ruddy, J., Shield, A. (2018). Predictive modeling of hamstring strain injuries in elite Australian footballers. School of Exercise Science.

Sarker, IH., Kayes, ASM., Badsha, S., Alqahtani, H., Watters, P. (2020). Cybersecurity data science: an overview from machine learning perspective. J Big Data. 7(1):1–29.

Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor Mater.

Schröer, C. (2021). A systematic literature review on Applying CRIS-DM Process Model. Procedia Computer Science 181. 526–534.

Siembida, P., Zawadka, M., Gawda, P. (2021). Relationship between vertical jump performance during different training periods and results of 200m-sprint. Balt J Health Phys Act. ;13(3):1-9

Singh, N. (2020). Sports Analytics: a review. The International Technology Management Review. Vol. 9(1); 2020, pp. 64–69

Ślusarczyk, B. (2018). Industry 4.0: Are we ready? Polish J Manag Stud. 17.

Sperandei, S. (2013). Understanding logistic regression analysis. Biochemia Medica 24(1):12-8.

Susmita, R. (2019). A Quick Review of Machine Learning Algorithms. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)

Sweeney, C., Ennis, E., Mulvenna, M., Bond, R., O'Neill, S. (2022). How Machine Learning Classification Accuracy Changes in a Happiness Dataset with Different Demographic Groups. Computers, 11, 83.

Tak, I., Glasgow, P., Langhout, R., Weir, A., Kerkhoffs, G., Agricola, R. (2015). Hip Range of Motion Is Lower in Professional Soccer Players With Hip and Groin Symptoms or Previous Injuries, Independent of Cam Deformities. The American Journal of Sports Medicine. 44.

Tak, I., Glasgow, P., Langhout, R., Weir, A. (2016). Hip Range of Motion Is Lower in Professional Soccer Players With Hip and Groin Symptoms or Previous Injuries, Independent of Cam Deformities. The American Journal of Sports Medicine. 44(3):682-688

Terry, AC., Thelen, MD., Crowell, M., Goss, DL. (2018). The musculoskeletal readiness screening tool- athlete concern for injury & prior injury associated with future injury. Int J Sports Phys Ther. 13(4):595-604.

Theng, M., Theng, D.. (2020). Machine Learning Algorithms for Predictive Analytics: A Review and New Perspectives.

Thornton, HR., Delaney, JA., Duthie, GM., Dascombe, BJ. (2017) Importance ofVarious Training-Load Measures in Injury Incidence of Professional RugbyLeague Athletes. Int J Sports Physiol Perform 12:819–824.

Ting, S.L. & Ip, W.H. & Tsang, Albert. (2011). Is Naïve Bayes a Good Classifier for Document Classification?. International Journal of Software Engineering and its Applications. 5.

Valle, X., Mechó, S. (2022). Return to Play Prediction Accuracy of the MLG-R Classification System for Hamstring Injuries in Football Players: A Machine Learning Approach. Sports Medicine

Van Klij, P. (2021). Do hip and groin muscle strength and symptoms change throughout a football season in professional male football players? A prospective cohort study with repeated measures. J Sci Med Sport.

Walsh, J., Heazlewood, I. (2018). Body Mass Index in Master Athletes: Review of the Literature. Journal of Lifestyle Medicine. Vol. 8, No. 2, 79-98.

Wasserman, E., Herzog, M. (2018). Fundamentals of Sports Analytics. Clin Sports Med 37; 387–400.

Wirth, R. Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Wrona, H., Zerega, R., King, V., Reiter, R. (2023). "Ability of Countermovement Jumps to Detect Bilateral Asymmetry in Hip and Knee Strength in Elite Youth Soccer Players" Sports 11, no. 4: 77.

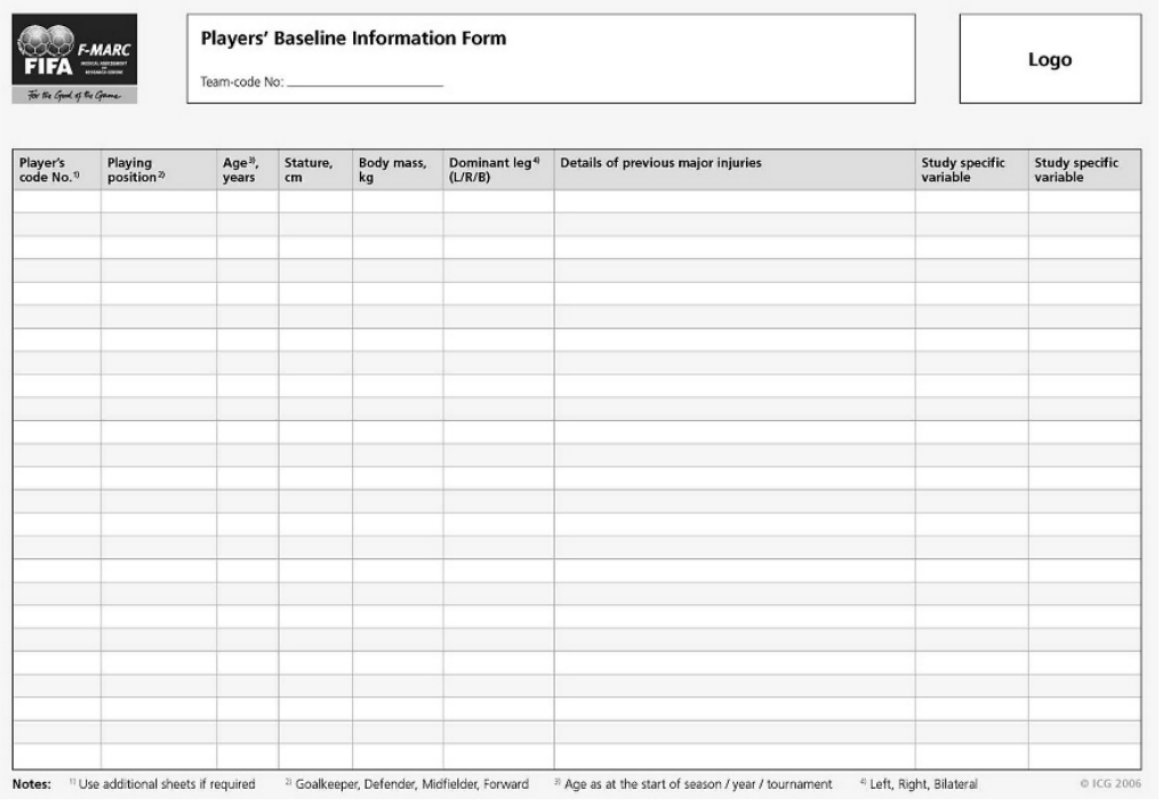
Wrona, H.L., Zerega, R., King, V.G., Reiter, C.R., Odum, S., Manifold, D. (2023). Ability of Countermovement Jumps to Detect Bilateral Asymmetry in Hip and Knee Strength in Elite Youth Soccer Players. Sports,11, 77.

Wu, X., Kumar, V., Quinlan, R. (2007). Top 10 algorithms in data mining. Knowledge and Information Systems. 14.

Zhang, P., Jia, Y., Shang, Y. (2022). Research and application of XGBoost in imbalanced data. International Journal of Distributed Sensor Networks. 18(6).

Zhongguo, Y., Li, H., Ali, S. (2017). Choosing Classification Algorithms and Its Optimum Parameters based on Data Set Characteristics. Journal of Computers. 28. 26-38.

APPENDIXES



Players' Baseline Information Form

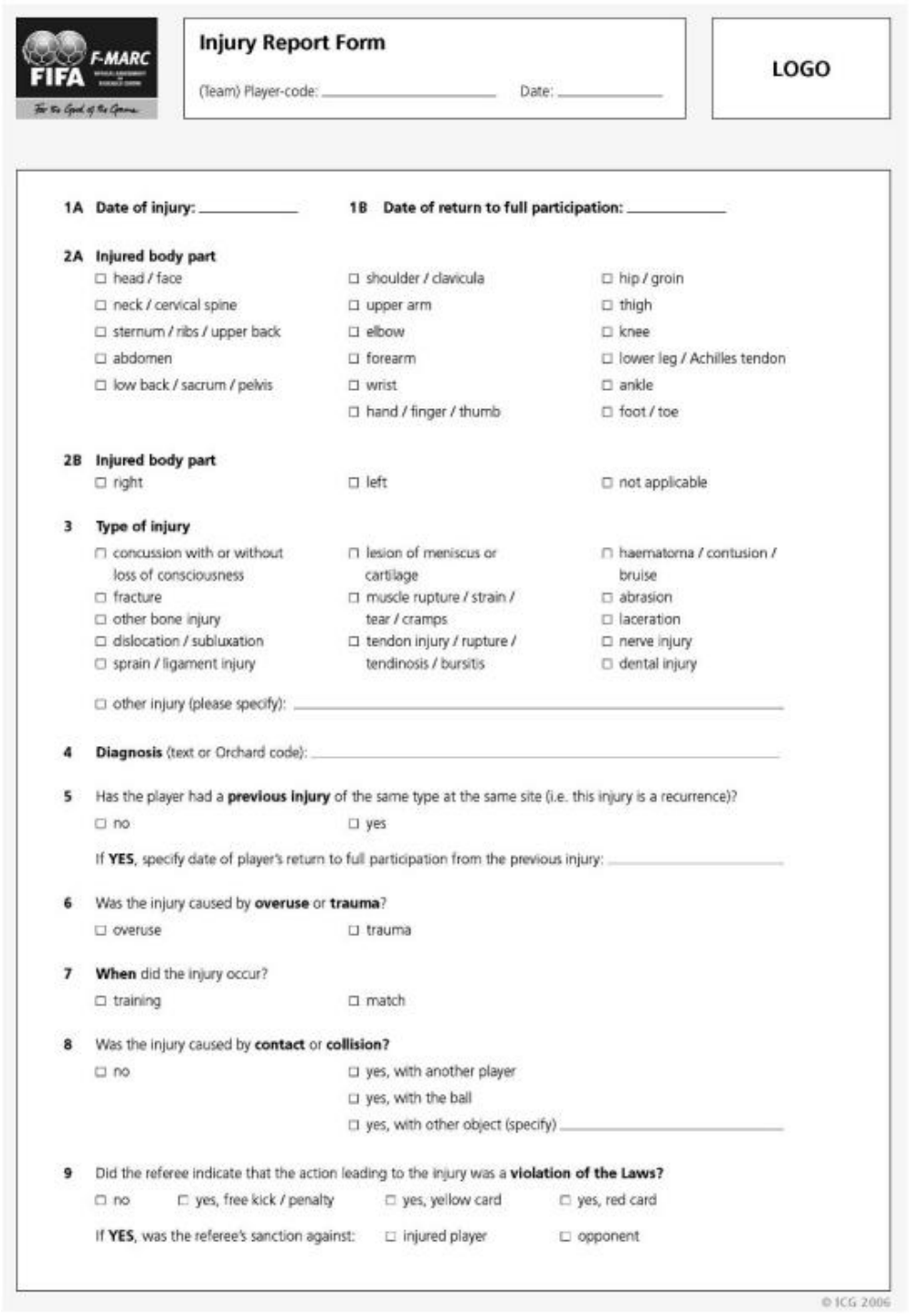
Team-code No: _____

Logo

Player's code No.[1]	Playing position[2]	Age[3], years	Stature, cm	Body mass, kg	Dominant leg[4] (L/R/B)	Details of previous major injuries	Study specific variable	Study specific variable

Notes: [1] Use additional sheets if required [2] Goalkeeper, Defender, Midfielder, Forward [3] Age as at the start of season / year / tournament [4] Left, Right, Bilateral © ICG 2006

Appendix A – Players' Baseline Information Form



Injury Report Form

(Team) Player-code: _____ Date: _____

LOGO

1A Date of injury: _____ **1B Date of return to full participation:** _____

2A Injured body part

- ☐ head / face
- ☐ neck / cervical spine
- ☐ sternum / ribs / upper back
- ☐ abdomen
- ☐ low back / sacrum / pelvis
- ☐ shoulder / clavicula
- ☐ upper arm
- ☐ elbow
- ☐ forearm
- ☐ wrist
- ☐ hand / finger / thumb
- ☐ hip / groin
- ☐ thigh
- ☐ knee
- ☐ lower leg / Achilles tendon
- ☐ ankle
- ☐ foot / toe

2B Injured body part

- ☐ right
- ☐ left
- ☐ not applicable

3 Type of injury

- ☐ concussion with or without loss of consciousness
- ☐ fracture
- ☐ other bone injury
- ☐ dislocation / subluxation
- ☐ sprain / ligament injury
- ☐ lesion of meniscus or cartilage
- ☐ muscle rupture / strain / tear / cramps
- ☐ tendon injury / rupture / tendinosis / bursitis
- ☐ haematoma / contusion / bruise
- ☐ abrasion
- ☐ laceration
- ☐ nerve injury
- ☐ dental injury
- ☐ other injury (please specify): _____

4 Diagnosis (text or Orchard code): _____

5 Has the player had a **previous injury** of the same type at the same site (i.e. this injury is a recurrence)?

- ☐ no
- ☐ yes

If **YES**, specify date of player's return to full participation from the previous injury: _____

6 Was the injury caused by **overuse** or **trauma**?

- ☐ overuse
- ☐ trauma

7 When did the injury occur?

- ☐ training
- ☐ match

8 Was the injury caused by **contact** or **collision?**

- ☐ no
- ☐ yes, with another player
- ☐ yes, with the ball
- ☐ yes, with other object (specify) _____

9 Did the referee indicate that the action leading to the injury was a **violation of the Laws?**

- ☐ no
- ☐ yes, free kick / penalty
- ☐ yes, yellow card
- ☐ yes, red card

If **YES**, was the referee's sanction against: ☐ injured player ☐ opponent

Appendix B - Injury Report Form

Exposure Report Form

(for the documentation of individual players' exposures)

(Team) Player-code: _____

LOGO

Date										
Training / Match										
Study specific variable										
Study specific variable										
Player Code No.	Report the duration of training and match play for each player (minutes)									

© ICG 2006

Appendix C - Exposure Report Form

Exposure Report Form
(for the documentation of team exposures)

(Team) Player-code: _____

LOGO

Date	Match / Training	No. of players (fully participating in training)	Duration of training session (minutes)	Study specific variable	Study specific variable

Appendix D - Exposure Report Form (for documentation of team exposure)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa