

POLITECNICO DI TORINO

Department of Management and Production Engineering – Class LM-31

Master's Degree in Engineering and Management

Path: Management of Sustainability and Technology



Master's Thesis

Predictive models for injury prevention in football: a machine learning approach using performance and injury data

Supervisor:

Prof. Giovanni Zenezini

Candidate:

Giorgio Tonizzo

Co-Supervisor:

Prof. Filippo Maria Ottaviani

Academic Year 2024-2025

Acknowledgments

I am profoundly grateful to my supervisor, Professor Giovanni Zenezini, and to my co-supervisor, Professor Filippo Maria Ottaviani, for their guidance and support during the development of this thesis. Their expertise and valuable feedback have been essential in defining both the focus and the excellence of this work.

Abstract

Injuries represent one of the most critical challenges in professional football, with significant implications for both on-field performance and the financial sustainability of clubs. This thesis explores the predictive potential of machine learning (ML) models in forecasting injury risk, while also identifying the most informative variables contributing to injury occurrence, based on match-level performance data and longitudinal injury histories. A comprehensive dataset was constructed for 100 outfield players in Italy's Serie A, covering the 2021/22, 2022/23, and 2023/24 seasons. Match statistics were collected from *FBref* and merged with injury records from *Transfermarkt*, restricted to time-loss events.

The entire methodological pipeline was implemented in Orange3 and included data preprocessing, feature selection, hyperparameter optimization, and model evaluation. Three target variables were defined: one categorical, classifying injury presence and type, and two binaries, aimed at predicting injury occurrence in the medium and short term. The models selected, Logistic Regression, Random Forest, and Gradient Boosting were trained on both original and rebalanced datasets using the oversampling technique called SMOTENC, and evaluated through F1-score, Matthews Correlation Coefficient (MCC), Confusion Matrix, and Log-loss.

The results underscore the complexity of the problem and emphasize the added value that more complete and accurate datasets, including parameters currently unavailable from public sources, to achieve satisfactory predictive performance. The study identifies a subset of variables with informative value, offering actionable insights for the future integration of ML-based tools into injury prevention and decision-support systems in professional football environments.

Table of contents

Acknowledgments	I
Abstract.....	II
1. Introduction.....	1
2. Literature Review	4
2.1. Design of the literature review	4
2.2. Injuries as a Strategic Cost Factor in the Football Industry	5
2.2.1. Football as Global Industry.....	5
2.2.2. Injury-Related Financial Costs and Club Revenue Losses	7
2.2.3. Impact on performance and team success.....	9
2.3. Injury definition and overview.....	11
2.3.1. Conceptualization and classification of injuries in football	12
2.3.2. Epidemiological evidence as a basis for injury impact	14
2.4. The role of Machine Learning.....	16
2.4.1. Introduction of Machine Learning.....	16
2.4.2. Machine learning use in soccer fields.....	19
2.4.3. Machine Learning in the Injuries Studies	21
2.5. Gaps in the literature, research contributions and future perspectives.....	24
3. Methods and Materials.....	27
3.1. Aims of the chapter	27
3.2. Data collection	27
3.3 Data preprocessing and features selection	38
3.4 Model selection and evaluation.....	56
3.5. Hyperparameters tuning.....	59
4. Results.....	62
4.1. Model Performance on the Training Set	62
4.1.1. Training-Set performance for Injury_Categorization_Next20Days.....	62
4.1.2. Training-set metrics for Injury_Next20Days	64
4.1.2.1. Training-Set Performance on original imbalanced data (Injury_Next20Days)	
.....	64
4.1.2.2. Training-Set performance on rebalanced data (Injury_Next20Days)	65
4.1.3. Training-Set metrics for Injury_Next3Days.....	67

4.1.3.1. Training-Set performance on original imbalanced data (Injury_Next3Days) ..	67
4.1.3.2 Training-Set performance on rebalanced data (Injury_Next3Days).....	69
4.2. Model Performance on the Test Set	71
4.2.1. Test-Set performance for Injury_Categorization_Next20Days.....	71
4.2.2. Test-Set performance of models trained on original and rebalanced data (Injury_Next20Days)	72
4.2.3. Test-Set performance of models trained on original and rebalanced data (Injury_Next3Days)	74
4.3. Features Importance	76
4.3.1. Feature importance with respect to Injury_Categorization_Next20Days	77
4.3.2 Feature importance with respect to Injury_Next20Days on the original and rebalanced training dataset	79
4.3.3 Feature importance with respect to Injury_Next3Days on the original and rebalanced training dataset	82
5. Discussion.....	86
6. Conclusion	90
<i>Bibliography</i>	92
<i>Web References</i>	96

1. Introduction

The football industry is one of the most valuable and globalised sectors in the sports economy, with top-tier clubs operating as complex businesses driven by commercial, media, and sporting performance (Deloitte, 2025).

Simultaneously, football remains a complex contact sport characterised by relatively high injury risks during both training and matches. Players perform at higher speeds and with greater physical intensity, demanding exceptional levels of fitness and increasingly rigorous training schedules (Pfirrmann et al., 2016). In this context, professional football clubs are increasingly exposed to both economic and performance-related challenges due to the growing incidence of player injuries, which represent a growing issue in terms of financial performance and competitive success (Howden's Professional Sport, 2024). International football organizations have expressed growing concern over the mounting physical and psychological pressures faced by elite athletes, which are contributing factors to the rising number of injuries (Pfirrmann et al., 2016).

Since 2020, a total of 14,292 injuries have been recorded across the top five European football leagues, excluding those related to COVID-19. Injury statistics show a consistent upward trend, largely driven by increasing fixture congestion due to the expansion of competitions. This culminated in 4,123 injuries during the 2023/24 season alone, resulting in an estimated financial loss of approximately €732 million (Howden's Professional Sport, 2024).

Given the relevance and complexity of the subject, it became essential to standardize the injury definition in professional football. For this purpose, an Injury Consensus Group was convened under the supporter of the Fédération Internationale de Football Association (FIFA) Medical Assessment and Research Centre. Using a nominal group consensus model, a panel reviewed and refined a working document outlining definitions, methodological standards, and implementation guidelines. As a result, injuries were categorised into two distinct types: a 'medical-attention injury', defined as any injury requiring medical evaluation or treatment, and a 'time-loss injury', referring to any injury that prevents a player from participating fully in future training sessions or match play (Fuller et al., 2006). In the context of this thesis, only the latter category, 'time-loss injuries', will be taken into consideration, as they are more directly linked to player availability and have clearer implications in terms of performance and economic impact. Thanks to the large amount of data collected during

training sessions and official matches, and to the adoption of electronic tracking and performance monitoring systems, research on injury prevention has become increasingly important.

In order to exploit the complex interaction between these various data sources, sports scientists have increasingly begun to apply Machine Learning (ML) techniques to predict potential injury risks (Leckey et al., 2024).

By definition:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ." (Hierons, 1999).

This definition highlights the core concept of machine learning: the ability of a system to improve its performance through exposure to increasing volumes of data. In the context of professional football, ML represents a powerful tool capable of uncovering complex and non-linear relationships among factors that precede injury occurrence, relationships that could otherwise remain unknown (Leckey et al., 2024).

Building on this theoretical framework, the present study applies machine learning techniques to develop predictive models aimed at supporting injury risk forecast. In particular, this thesis aims to answer the research question: *‘How machine learning models can improve the prediction of injury risk in football players, and which variables provide the most relevant information for forecasting purposes’*. To address this research question, a data collection was conducted on professional Serie A footballers. The resulting dataset encompasses detailed match-level metrics spanning three complete seasons. Concurrently, each athlete’s longitudinal injury record was reconstructed from the date of his first professional appearance. Following data acquisition, each phase of the analytical pipeline, from feature preprocessing to model comparison, was carried out within the Orange3 environment. Specifically, Orange3 was employed to execute feature preprocessing, during which injury records were merged with match-by-match performance data to align temporal and clinical events with performance metrics, composite metrics were generated to more effectively capture player load dynamics; feature selection, designed to reduce dimensionality and enhance model training efficiency; hyperparameter tuning for each candidate predictive algorithm; and subsequently, model comparison, to identify the approach exhibiting optimal predictive performance.

Finally, the algorithms results and features importance were thoroughly evaluated, examining model interpretability and performance metrics, and the inherent limitations of the study were critically discussed.

2. Literature Review

2.1. Design of the literature review

In the initial stage of this analysis, a review of the literature was conducted to assess whether existing studies demonstrate the impact of injuries on team performance and the economic outcomes of football clubs. This investigation aimed to evaluate the practical relevance of predictive approaches to injury prevention and forecasting, with the goal of supporting strategic decision-making in the management of players as key assets within the club structure. Building on this premise, the review proceeded to examine how current research has addressed injury prediction through the application of machine learning models. These guiding questions reflect the key dimensions relevant to the development of injury prediction models and inform the subsequent sections of the review:

- To what extent do injuries, in quantitative terms, affect a club's overall profitability through both direct costs and indirect costs ?
- What is the relationship between injury incidence and burden and team performance indicators, according to the available longitudinal evidence?
- Which operational definition of injury is most appropriate for predictive modelling and inter-study comparability?
- How does the categorization of injury severity influence the performance of machine learning models?
- Which sets of variables are most frequently used in existing models, and what is their relative importance?
- How does data segmentation affect the risk of information leakage and the external validity of the models?
- To what extent does sample size or specificity influence the accuracy, statistical robustness, and generalizability of predictive results?

The primary sources consulted were four major academic databases: Scopus, PubMed, Web of Science, and IEEE Xplore. In addition, non-scientific but institutionally relevant sources, such as websites of national football associations, players unions, and specialized sports analytics platforms, were consulted to complement the academic literature with sector specific data and context.

The temporal scope covered the period from 2010 to 2025, with most selected studies published within this range. A few earlier contributions were included when considered conceptually significant.

Keyword combinations were progressively refined and adapted to the syntax of each database. Examples of Boolean search strings include:

- "soccer*" AND "injury" AND "machine learning"
- "football" AND "injury risk" AND ("load monitoring" OR "training load")
- "sports injuries" AND ("prediction" OR "forecasting") AND ("data mining" OR "AI")
- "injury prevention" AND ("external load" OR "ACWR") AND "athlete"
- "deep learning" AND "injury prediction" AND "time series"

The search process was carried out iteratively, with titles and abstracts screened for relevance and full texts consulted to ensure alignment with the objectives and guiding questions of the review.

This design framework provides the conceptual and methodological foundation for the critical analysis presented in the following sections, where the literature is reviewed considering the guiding questions introduced above.

2.2. Injuries as a Strategic Cost Factor in the Football Industry

2.2.1. Football as Global Industry

Soccer is the most popular global sport, with 200,000 professional and 240 million amateur players, and with injury incidence higher than any other sport. (Majumdar et al., 2024a)

This exceptionally participation base underscores the game's unmatched capacity to attract worldwide interest and generate substantial revenues (Majumdar et al., 2024a; Schilde, 2025). It has long been observed that football occupies a central position within the

international sports economy (Schilde, 2017). Of the €45 billion generated by the global sports-events market in 2009, including ticket sales, media rights and marketing income, football alone accounted for €20 billion (Schilde, 2025). Also, annual revenue continued to rise, reaching €28.9 billion in Europe during the 2018/2019 season (Schilde, 2025).

This sustained expansion of the sport's income streams has been mirrored in club valuations (Football Benchmark, 2024). The aggregate enterprise value (EV) of the 32 leading European clubs rose from €26.3 billion in 2016 to €59.1 billion in 2024, an increase of 124 per cent in just eight seasons (Football Benchmark, 2024). Because EV reflects the debt-free price an investor would have to pay to acquire a club, it offers a comprehensive view for underlying corporate worth and shows how rising revenues translate directly into higher market valuations (Football Benchmark, 2024). To derive these estimates, It is employed an adjusted revenue-multiple model that weights five key factors:

(i) *Profitability* is measured by the staff costs-to-revenue ratio over the last two financial years. Since player and staff wages account for most total expenditures, a lower ratio reflects a stronger ability to generate profits (Football Benchmark, 2024).

(ii) *Popularity* is measured through the club's reach on major social media platforms, including Facebook, X, Instagram, YouTube, TikTok, and Weibo. This indicator serves as a view for brand strength and fan engagement (Football Benchmark, 2024).

(iii) *Sporting potential* is captured by the aggregate market value of the squad, based on Football Benchmark's Player Valuation tool. As the squad represents the club's core asset, its estimated value reflects the likelihood of on-field success and related revenue streams (Football Benchmark, 2024).

(iv) *Broadcasting rights* refer to the future income already secured at the league level and the related distribution mechanisms. This element plays a critical role in shaping the club's revenue-generating capacity over the medium term (Football Benchmark, 2024).

(v) *Stadium ownership* is also included in the model, as owning the home ground enables greater control over matchday revenues and offers additional commercial opportunities compared to leasing arrangements (Football Benchmark, 2024).

2.2.2. Injury-Related Financial Costs and Club Revenue Losses

As previously outlined, the overall value of a football club is determined by five key dimensions. A growing body of research suggests that injuries significantly affect several of these dimensions, both directly and indirectly, impacting not only sporting performance but also financial sustainability and the valuation of broadcasting rights (AIC, 2024; Eliakim et al., 2020; Pulici et al., 2023).

Focusing on the cost side of profitability, Pulici et al. (2023) quantify only the wages that become economically unproductive when players are unavailable, deliberately excluding variable match-bonuses, medical expenses and any loss of transfer market value, elements that will be addressed by subsequent studies in this chapter (Pulici et al., 2023). Drawing on injury-surveillance data from professional adult male players in the top five European leagues, the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, and French Ligue 1, and limiting the analysis to studies conducted from 2005 onward, the authors combine epidemiological injury data with the average daily salary of professional players (~€5,868) to convert the injury burden (days lost per 1,000 hours of exposure) into a direct salary cost (Pulici et al., 2023). Their analysis shows that knee injuries alone account around €204,000 per 1,000 hours, while thigh and joint/ligament injuries impose comparable burdens in the range of €180,000–€225,000 per 1,000 hours. Scaled to the injury profile of a top-tier club (around 50 injuries per season), these figures translate into more than €6 million per year in wages paid to players who cannot contribute on the pitch, before considering performance-linked revenues or additional medical outlays (Pulici et al., 2023).

A more recent contribution within the same analytical framework is provided by the *Men's European Football Injury Index 2023/24*, published by Howden. This report also focuses on the direct financial impact of injuries sustained by clubs competing in Europe's top five leagues (Howden's Professional Sport, 2024). Applied to a sample of 96 clubs, the report estimates a total injury-related cost of €732.02 million for the 2023/24 season, based on 4,123 reported injuries, a figure that reflects the ongoing upward trend (Howden's Professional Sport, 2024).

The “Injury Cost over Time” chart included in the report clearly illustrates a consistent increase in both the frequency and financial burden of injuries over the past four seasons (Howden's Professional Sport, 2024). *Figure 1* supports the trend described and contextualise the cumulative impact on club finances (Howden's Professional Sport, 2024).

This trend highlights the mounting economic pressure that injuries has on elite football clubs, reinforcing the strategic importance of targeted prevention programmes and efficient medical management protocols (Howden’s Professional Sport, 2024) .

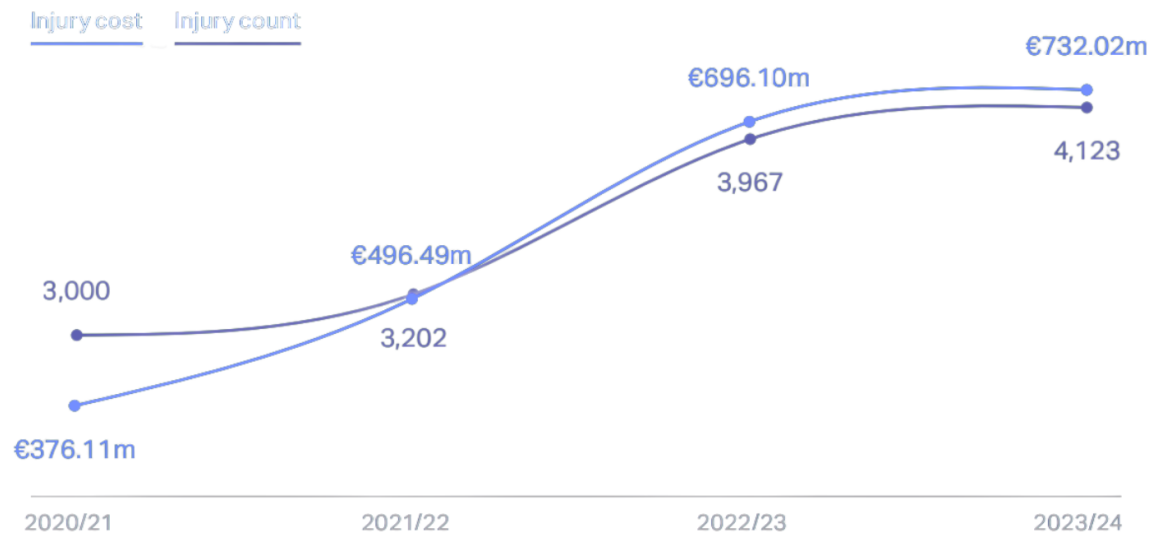


Figure 1 — Injury count and cost per season

Beyond the immediate financial outlay associated with paying the salaries of unavailable players, the economic repercussions of prolonged injuries extend to the capital valuation of the athletes themselves (AIC, 2024). Evidence from a recent study conducted across Serie A, the Premier League, and La Liga demonstrates that the depreciation in a player’s market value, calculated exclusively for injury-related absences exceeding the critical threshold of 90 cumulative days, consecutive or not, can surpass the corresponding wage loss, which is instead accounted from the first day of unavailability (AIC, 2024). In the analysis, the total estimated loss in market value reached €785 million, compared to €707 million in salary payments made to injured players (AIC, 2024). These findings underscoring how injuries affect not only short-term profitability but also the asset value of the squad. Consequently, their impact extends to at least two of the five dimensions that determine a football club’s enterprise value (AIC, 2024).

Extending the analysis to the broadcasting rights, evidence from a study conducted on Spain’s top-flight division highlights how injuries erode a club’s capacity to capture merit based broadcasting revenues (Torrejón et al., 2024).

Focusing specifically on lower limb muscle injuries, the authors analysed publicly available injury reports from all 20 La Liga clubs during the 2018/19 season, obtaining a total of 270 muscular injuries recorded (Torrejón et al., 2024).

To evaluate the impact on broadcasting revenue, the study compared each club's expected league position, with their actual end-of-season standing, taking in account the injuries for each club. Applying the redistribution rules set by *Real Decreto-ley 5/2015*, which allocates 25% of centrally pooled broadcasting income based on final league position, the authors estimated that nine clubs underperformed relative to expectations, resulting in a collective loss of €45.2 million in broadcasting revenues (Torrejón et al., 2024); *Real Decreto-ley 5/2015*).

Taken together, the studies reviewed highlight that the economic consequences of injuries in football extend well beyond the immediate costs associated with player salaries. Injuries negatively affect the asset valuation of players, resulting in potential losses in their market value (AIC, 2024). They also undermine overall club profitability by increasing the share of unproductive wages and reducing the ability to generate net profits (Pulici et al., 2023). Furthermore, injuries limit access to merit-based broadcasting revenues distributed according to final league standings (Torrejón et al., 2024). Collectively, these impacts reveal a multidimensional financial pressure that significantly undermines the enterprise value of football clubs, as defined by the five-pillar framework developed by (Football Benchmark, 2024).

2.2.3. Impact on performance and team success

The analysis of football injuries represents a critical aspect in the evaluation of team performance and competitive success. A growing number of studies have confirmed the strong correlation between injury rates and team performance indicators in professional football (Eliakim et al., 2020; Häggglund et al., 2013; Pulici et al., 2023) .

A demonstration of this relationship is provided by a longitudinal study conducted on 24 elite European football clubs (Häggglund et al., 2013). Over an observation period of 11 seasons, the researchers reported an average injury incidence of 7.7 injuries per 1,000 hours and an injury burden of 130 days lost per 1,000 hours (Häggglund et al., 2013; Pulici et al., 2023).

The results show that teams with a higher injury burden compared to the previous season experienced a significant decline in performance: every additional 100 days lost due to injury per 1,000 hours was associated, on average, with one position lower in the final league standings and 7.6 fewer points over the course of the season (Hägglund et al., 2013). Player availability, defined as the percentage of match opportunities in which players were fit to play, was also positively correlated with performance: a 5% increase in availability led on average, to 3.6 additional points and an improvement of 0.4 positions in the league table (Hägglund et al., 2013). A similar effect was observed at the international level, where both injury burden and match availability significantly influenced the UEFA Season Club Coefficient, an indicator of a team's European success (Hägglund et al., 2013).

A similar pattern emerges from a study that examined Premier League clubs over the 2012–2017 seasons: each additional 271 injury-days lost within a single campaign was associated, on average, with one place lower in the final league standings, whereas 136 injury-days corresponded to one league point fewer (Eliakim et al., 2020). The association between injury burden and the gap between expected and actual ranking proved statistically significant (Pearson's $r = -0.46$, $r^2 \approx 0.21$), denoting a moderate inverse linear relationship; the corresponding coefficient of determination indicates that approximately 21 % of the variance in final position is attributable to injury-related player unavailability (Eliakim et al., 2020). A comparable yet slightly weaker correlation was observed for total points gained ($r = -0.38$; $r^2 \approx 0.14$) (Eliakim et al., 2020). In overall, across the observation window, Premier League clubs recorded a seasonal mean of 1,410 injury-days and 58 discrete injuries, a level of absenteeism that corresponds to an estimated shortfall of roughly 4.8 league positions and nine points relative to the performance predicted by squad market value (Eliakim et al., 2020).

These findings collectively underscore the substantial impact of injury incidence and burden on team success, emphasizing the critical importance of effective injury-prevention and player-availability strategies in sustaining competitive performance at both national and international levels (Eliakim et al., 2020; Hägglund et al., 2013)

2.3. Injury definition and overview

As previously discussed in the introduction, the need to standardize terminology and methodological approaches in studies addressing football injuries led to the publication of the paper “Consensus statement on injury definitions and data collection procedures for studies of football (soccer) injuries” (Fuller et al., 2006). This definition provides a consistent framework for the classification and analysis of injuries in the context of professional football:

“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 et al., 2006).

An operational application of the definition proposed by Fuller et al. (2006) is found in the work of Ekstrand et al., (2011) developed within the framework of the UEFA Elite Club Injury Study, a longitudinal research programme carried out in collaboration with professional European clubs (Ekstrand et al., 2011). In this context, an injury is defined as any physical damage occurring during scheduled football activities (training sessions or matches) that prevents the player from fully participating in subsequent training sessions or matches (Ekstrand et al., 2011), exactly the definition of the time-loss injuries. The main goal is to objectively quantify the impact of injuries on player availability, thereby enabling standardised epidemiological comparisons across teams and seasons (Ekstrand et al., 2011).

In response to the evolution of scientific standards, a new definition of injury has been proposed within the context of football. Unlike the 2006 formulation, which explicitly linked the occurrence of injury to participation in training sessions or matches, the updated version adopts a more general and physiologically grounded perspective (Waldén et al., 2023). Injury is now defined as:

“Tissue damage or other derangement of normal physical function, resulting from rapid or repetitive transfer of kinetic energy” (Waldén et al., 2023)

The removal of the explicit reference to football activity allows for a clearer distinction between the clinical definition of injury and its contextual classification. According to the authors, this choice aims to harmonise the definition with the separate categorisation of

health problems based on their relationship to sports activity, thereby contributing to greater methodological consistency in injury surveillance and reporting systems (Waldén et al., 2023)

2.3.1. Conceptualization and classification of injuries in football

In the early 2000s, a methodology was proposed to classify and assess injury risk in professional football by integrating more characteristics into a unified metric (Drawer & Fuller, 2002). This framework consider the injuries not only as countable events, but it is classified across multiple operational dimensions: severity (expressed as days of absence and classified into slight, minor, moderate, and major categories), anatomical location, clinical nature (e.g., muscle strains, contusions, sprains, fractures), and causal mechanism (e.g., receiving a tackle, running, shooting). Although not yet formalized as a theoretical framework, this structure represented a rigorous empirical application (Drawer & Fuller, 2002).

A structured and multidimensional classification of injuries is proposed articulates injury reporting across five distinct yet complementary dimensions:

- Causal mechanism, distinguishing between traumatic and overuse injuries (Fuller et al., 2006);
- Anatomical location, including the affected body part and side (Fuller et al., 2006);
- Clinical diagnosis, to be provided by a qualified healthcare professional when applicable (Fuller et al., 2006);
- Severity, expressed in terms of time loss from football activity (Fuller et al., 2006);
- Recurrence, identifying whether the injury is new or a repeated event (Fuller et al., 2006);

This classification system reflects a pragmatic approach to codifying injury data within the specific context of football, and has served as a reference model for numerous subsequent studies (Fuller et al., 2006).

Waldén et al., (2023) not only advanced an updated injury definition but also introduced a more granular classification framework tailored to professional football. While retaining

the core dimensions introduced by Fuller et al. (2006), the updated model introduces more categories of injury severity, ranging from 0 days to >180 days of time-loss, and expands definitions related to recurrence, tissue pathology, and return-to-play criteria (Waldén et al., 2023). Additionally, the framework formalizes football-specific mechanisms of injury, incorporates contextual variables such as player action at the time of onset, and provides detailed guidance on how to classify subsequent or exacerbated injuries (Waldén et al., 2023). This updated consensus serves as a comprehensive methodological reference for injury surveillance in modern football (Waldén et al., 2023).

In the broader literature, however, injury classification practices often deviate from the comprehensive structure proposed by Fuller et al. (2006). Among the approaches found in the literature, some studies adopt simplified categorization schemes focused primarily on the duration of absence from training or match play (Falese et al., 2016; Freitas et al., 2025). A commonly used framework, aligned with UEFA's operational standards, defines severity using four levels: light (1–3 days), minor (4–7 days), moderate (8–28 days), and severe (more than 28 days), without integrating diagnostic or anatomical specificity (Falese et al., 2016). Similarly, other empirical works classify injuries into broad macro-categories (e.g., muscle/tendon, ligament/joint) and group non-specific conditions under residual labels such as "other", limiting the resolution of injury typology and site (Freitas et al., 2025).

An additional specification, common in the context of injury prediction, concerns the exclusive focus on non-contact time-loss injuries (Rossi et al., 2018). Several recent studies in the field of injury prediction, particularly those focused on muscle injuries and employing machine learning algorithms, have further simplified the classification scheme by exclusively including non-contact time-loss injuries (Rossi et al., 2018). This choice, while aligned with the UEFA definition, is typically motivated by the need to reduce variability caused by unpredictable traumatic events (e.g., collisions), which may act as confounding factors in predictive analyses based on workload metrics and biometric data (Rossi et al., 2018).

2.3.2. Epidemiological evidence as a basis for injury impact

Following the conceptual definition and classification of injuries in football presented in the previous section, this chapter focuses on the epidemiological dimension. In professional football, decision-makers such as club owners, top managers, directors of health and performance, and technical staff increasingly rely on accurate and time-efficient methods to assess injury risk and monitor performance, with the dual objective of safeguarding athletes' health and enhancing team outcomes (Nassis et al., 2023).

Within this context, injury epidemiology is the discipline concerned with studying the incidence, severity, causes, and distribution of injuries within a defined athletic population (Palmer, 2015). Its main objective is to provide a robust quantitative basis for understanding the magnitude of the problem, identifying key risk factors, and informing the design of effective prevention strategies thanks to the implementation of well-structured observational studies and the use of standardized metrics (Palmer, 2015). Without systematic epidemiological analysis, the allocation of medical resources and the development of targeted interventions would remain largely speculative (Palmer, 2015).

To support the effective application of injury epidemiology in football, researchers commonly rely on a set of standardized metrics that enable the quantification of injury patterns and their impact on players and teams (López-Valenciano et al., 2020). A notable example of epidemiological research in football is the systematic review and meta-analysis conducted by López-Valenciano et al., (2020), which focused on professional male players. The authors reported an overall injury incidence of 8.1 per 1000 hours of exposure, with match-related injuries occurring at a substantially higher rate (36 per 1000 hours) compared to training-related injuries (3.7 per 1000 hours). Injuries most frequently affected the lower extremities (6.8 per 1000 hours), with muscle and tendon injuries being the most common types (4.6 per 1000 hours), often resulting from traumatic events. (López-Valenciano et al., 2020) Minor injuries, defined as those causing 1 to 3 days of absence, were the most prevalent severity category. The authors applied a random-effects model for the meta-analysis to account for between-study variability, and statistical heterogeneity was assessed using the I^2 statistic (López-Valenciano et al., 2020).

Another prominent example of epidemiological research in elite football is offered by a longitudinal surveillance study that systematically tracked injury and illness data across male and female squads of England's national teams over eight competitive seasons, providing a

robust empirical basis for understanding exposure-related risk differentials (Sprouse et al., 2024). As in the previous study, injury incidence was calculated per 1000 hours of exposure, distinguishing between training and match settings. Among senior male players, match-related injuries occurred at a rate of 31.1 ± 10.8 per 1000 hours, while training-related injuries were significantly lower at 4.0 ± 1.0 per 1000 hours. Similarly, the injury burden, expressed as days lost per 1000 hours of exposure, was markedly higher during matches (454 ± 196 days) than during training sessions (51 ± 22 days) (Sprouse et al., 2024). The statistical approach adopted in the study included two-way ANOVA, Bonferroni post-hoc comparisons, chi-square tests, and independent-samples t-tests, aimed at identifying significant differences in injury incidence, burden, and severity across sex, age groups, and exposure type (Sprouse et al., 2024).

Unlike the study by Lopez-Valenciano et al. (2020), which synthesised aggregated results from a broad set of published studies, the report written by Sprouse et al., (2024) was based on a consistent and standardised dataset collected prospectively within a single institutional context. Furthermore, while Lopez-Valenciano et al. (2020) focused primarily on injury types and anatomical locations, the study conducted by Sprouse et al., (2024) placed emphasis on the distinction between traumatic and overuse injuries, highlighting how the former predominated in match contexts, whereas the latter were more frequent during training. These findings underscore the role of stratified epidemiological surveillance and rigorous statistical analysis as essential sources of injury-related data. Such information is critical for the development of effective injury prevention or prediction strategies, as well as for ensuring the efficient allocation of medical resources within professional football settings.

Epidemiological research provides the quantitative backbone on which injury-prediction models are built: large-scale surveillance defines risk factors (Carey et al., 2018; Rossi et al., 2018; Van Eetvelde et al., 2021). Contemporary machine-learning studies consistently embed these epidemiological metrics as model features, selecting them as the most informative predictors of injury risk (Carey et al., 2018; Rossi et al., 2018; Van Eetvelde et al., 2021). The ability to identify, quantify and standardise population-level risk factors is essential for the selection of relevant input features and the definition of clinically meaningful prediction targets. (Rossi et al., 2018; Van Eetvelde et al., 2021).

2.4. The role of Machine Learning

2.4.1. Introduction of Machine Learning

Before exploring the many scientific applications of machine learning, it is essential to define a predictive algorithm: a set of rules or processes that enables a system to learn from data and estimate new outputs, and to outline the formal criteria by which such algorithms are classified (IBM, n.d.). The machine learning algorithm is a set of rules or processes used by an AI system to conduct tasks, most often to discover new data insights and patterns, or to predict output values from a given set of input variables (IBM, n.d.).

Predictive machine-learning algorithms can be formally classified according to the nature of the target variable, the learning paradigm adopted, the temporal structure of the data, and the possible aggregation of multiple models (Bai et al., 2018; Wakefield, n.d.). The first fundamental distinction among predictive algorithm types is based on the nature of the target variable. Problems are categorized as classification or regression tasks, depending on whether the target variable is discrete or continuous, respectively (Hu et al., 2020; Tizikara et al., 2022)

When the response variable Y takes real-valued measurements, such as physical, economic, or biological quantities, the task is defined as regression. Conversely, when Y represents a binary or categorical class, the task falls under the domain of classification. (Hu et al., 2020; Tizikara et al., 2022)

The distinction is clearly articulated in the literature, by definition: regression aims to “predict a continuous-valued output,” whereas classification is used to “predict a discrete-valued output” (Tizikara et al., 2022). This theoretical distinction is further illustrated by a range of practical applications. A spam-filtering model that assigns an email to either the “spam” or “not-spam” category is a classic example of binary classification (Hu et al., 2020; Tizikara et al., 2022). By contrast, a regression model can be employed to estimate the Optical Signal-to-Noise Ratio (OSNR) of an optical channel, producing a continuous value expressed in decibel (Tizikara et al., 2022).

A further classification of predictive algorithms concerns the learning paradigm they rely on (IBM, 2024). The first and most established form is supervised learning, a technique that relies on the use of labelled data, that is, input–output pairs in which the correct output is manually provided during training (IBM, 2024). The goal is to build a model capable of

learning the underlying relationships between inputs and outputs, and to generalize these relationships to unseen (IBM, 2024). The learning process is carried out across multiple stages: the available data is typically divided into training, validation, and test sets (IBM, 2024). Using optimization techniques the model's internal parameters are updated iteratively to minimize this error and enhance predictive accuracy (IBM, 2024). A concrete example is the development of an image classification model designed to automatically recognize vehicle types such as cars, motorcycles, or trucks (IBM, 2024). The model is trained on a labelled dataset containing images of vehicles along with their corresponding categories. Once trained, the system is able to accurately classify new, unseen images tests used to distinguish human users from automated bots (IBM, 2024).

Beyond supervised learning, a central role is played by unsupervised learning, a branch of machine learning that operates on unlabelled data, meaning that no explicit target variable is provided (IBM, 2024). In contrast to supervised learning, where each data point is associated with a known output, unsupervised algorithms are designed to discover hidden patterns, relationships, or intrinsic structures within the data, without any prior knowledge or human-provided labels (IBM, 2024; GeeksforGeeks, 2025).

This learning process is based on the analysis of similarities and differences among observations in the dataset, with the goal of uncovering implicit regularities (IBM, 2024). A representative application of unsupervised learning is the use of k-means clustering for market segmentation. This approach allows organizations to divide a customer base into distinct groups based on purchasing behaviour or preferences, without requiring predefined labels. The algorithm autonomously identifies shared characteristics among consumers, enabling the creation of meaningful segments to support targeted marketing strategies and personalized service delivery (IBM, 2024).

The third category of machine learning algorithms is represented by semi-supervised learning models (SSL), which lie conceptually between supervised and unsupervised learning paradigms (AltexSoft Editorial Team, 2024).

These algorithms leverage a small quantity of labelled data together with a large volume of unlabelled data. This approach mitigates the high costs and time demands of manual annotation while enhancing model performance beyond what unsupervised methods alone can achieve (Padmanabha Reddy et al., 2018). An illustrative example is the application of SSL in fraud detection. In scenarios where only a small fraction of transactions, say, 5%,

have been manually labelled as “fraud” or “non-fraud,” SSL can process the remaining 95% of the data by leveraging its structure and partial labels to identify potentially fraudulent behaviour, without requiring complete human annotation (AltexSoft Editorial Team, 2024).

Another key dimension for the classification of predictive algorithms concerns the temporal structure of the data and the model's ability to incorporate time-dependent information (Madhavan, 2016). A first distinction is made between static and dynamical models, based on whether or not the algorithm accounts for temporal dependencies when generating predictions (Madhavan, 2016).

Static models operate under the assumption that each observation is independent and identically distributed: the output is predicted solely from the input variables at the current time step with no reference to historical context. (Madhavan, 2016).

By contrast, dynamical models explicitly incorporate past observations or time-lagged features (Madhavan, 2016). They are designed to capture temporal evolution, making use of recursive real-time learning algorithms that update model parameters as new data becomes available (Madhavan, 2016). In these cases, input-output mappings are expressed in terms of both current and previous values, enabling the model to recognize trends (Madhavan, 2016).

Finally exist an architectural distinction that separates single-based algorithms from ensemble algorithms (Ganaie et al., 2022). Single-based models rely on a lone learner to generate predictions, whereas an ensemble combines the outputs of multiple base learners to achieve a level of generalization that surpasses any individual constituent, such as the Logistic Regression (Ganaie et al., 2022). On the other hand, an ensemble aggregates the diverse predictions produced by its base models, through strategies such as averaging, majority voting, or related fusion techniques, so that the collective outcome is demonstrably superior to that of each single component (Ganaie et al., 2022). Two notable examples of ensemble methods are Random Forest, which relies on bagging and decision trees, and Gradient Boosting, which builds sequential models by minimizing the prediction error of the previous ones (IBM, 2024).

2.4.2. Machine learning use in soccer fields.

Early machine-learning research on football focused on the RoboCup Soccer Server, a software simulator in which two teams of eleven virtual agents play matches whose every action is recorded in log files (Raines et al., 2000; Stone & Veloso, 1998). In 2000, Raines, Tambe and Marsella introduced, an offline analyst that turns RoboCup logs, time-ordered traces recorded by the simulation server for every robot action, into coach-oriented natural-language reports (Raines et al., 2000).

A representative log fragment for a shot stores variables such as ball velocity < 2.37 m/t and shot aim > 6.7 m from goal centre; the offline analyst exploits these values to learn rules explaining why the attempt fails (Raines et al., 2000).

Three complementary behavioural models: individual event, multi-agent interaction and global outcome are induced with decision-tree algorithm, which yields interpretable rules, augmented by pattern-matching procedures that detect recurrent action sequences (Raines et al., 2000).

On RoboCup '99 logs, the global model reproduces 87 % of training scores and 72 % of unseen matches, subset achieves 70 % pre-match accuracy in win-loss prediction (Raines et al., 2000).

Over the years the number of research papers on machine learning use in soccer has increased significantly, despite that it is still unclear what machine learning can offer to soccer clubs now and in the future, and how scientists and practitioners can prepare to take advantage of machine learning capabilities (Nassis et al., 2023).

In recent years, the application of machine learning techniques within the football domain has increased significantly, enabling new approaches for performance monitoring (M. Wang, 2014). One application of machine learning in the context of football was aimed at objectively and automatically evaluating the technical and tactical abilities of the teams participating in UEFA Euro 2012, with the goal of predicting each team's overall performance based on a set of statistical indicators (M. Wang, 2014). The core challenge addressed was a multi-criteria evaluation problem, namely, how to assign a comprehensive performance score to each team by simultaneously considering a variety of heterogeneous metrics (M. Wang, 2014). The resulting predictions were found to align with the actual

outcomes of the tournament, highlighting the potential of machine learning techniques for performance assessment in elite football settings. (M. Wang, 2014).

Building on this team-centred perspective, research shifted to an individual focus, seeking to characterise each player's contribution in finer detail (Pappalardo et al., 2019). PlayeRank, the studies carried out by Pappalardo et al., (2019), employs a supervised Linear Support Vector Classifier to learn the weights of seventy-six technical–tactical variables and assign an interpretable rating to every footballer; a soft k -means clustering routine identifies eight canonical roles, permitting cross-positional comparisons (Pappalardo et al., 2019). The model achieves 74 % unanimous agreement with professional talent scouts; in other words, whenever all three scouts chose the same player as better, PlayeRank made the same choice in 74 % of those comparisons, combining methodological transparency with external validity (Pappalardo et al., 2019).

Further advances leveraged high-frequency positional data: unsupervised algorithms automatically identified dynamic sub-groups of players during attacking phases, providing a context-sensitive classification of tactical patterns (Goes et al., 2021). A large-scale investigation based on 118 Eredivisie matches refined this approach by using a single k -means model to segment defenders, midfielders, and forwards frame by frame, achieving a mean silhouette score of 0.63 while deliberately eschewing ensemble techniques to preserve interpretability (Goes et al., 2021). The same analysis showed that successful attacks were preceded by a marked decline (Cohen $d \approx -0.41$) in longitudinal synchrony between defenders in possession and opposing forwards (Goes et al., 2021).

After the descriptive phase on sub-groups, research turned to complex tactical events capable of altering a match's outcome. Stein et al. (2022) address gegenpressing detection using Bundesliga event streams. Gegenpressing is defined as a team's attempt to regain possession within five seconds of losing the ball and as close as possible to the turnover location, through coordinated pressure by multiple players; only defensive transitions meeting these criteria are labelled positive, yielding a dataset of 11 108 instances (Bauer & Anzer, 2021). The primary algorithm, XGBoost, is described as a tree-boosting ensemble; two optimised configurations achieve, on the test set, an AUC of 0.874 and an F1-score of 0.67, outperforming Random Forest, Logistic Regression and naïve-rule baselines (Bauer & Anzer, 2021). The operational output consists of video playlists and reports importable into

analysis software, from which objective indicators of counter-pressing effectiveness can be derived across six Bundesliga seasons (Bauer & Anzer, 2021).

Finally, to close the loop from micro-tactical analysis to full-match prediction, a study in the *Decision Analytics Journal* (2024) forecast “home-win versus non-win” outcomes in the Premier League by integrating fifty-two performance, fatigue and meteorological variables (Wong et al., 2025). Six supervised models, Logistic Regression, Random Forest, Support Vector Machine, XGBoost, LightGBM and a Convolutional Neural Network, were combined in a stacking ensemble defined by the authors as a heterogeneous architecture; the system reached roughly 65 % accuracy, approaching bookmaker reliability without compromising stability (Wong et al., 2025).

Collectively, the evolution from the firsts model to the heterogeneous supervised ensembles demonstrates how machine learning has progressively expanded its descriptive and predictive capabilities in football.

2.4.3. Machine Learning in the Injuries Studies

The increasing complexity of modern football and the growing availability of high-frequency data have encouraged practitioners to adopt advanced technologies such as GPS tracking, inertial sensors, and psychophysiological monitoring to assess athletes’ condition and exposure. Although machine learning already holds a consolidated role in health sciences, its adoption in sports medicine, and specifically in football, remains emergent (Nassis et al., 2023).

In line with the emerging trend, two-season investigation on 36 professional footballers from the Polish Ekstraklasa club KKS Lech Poznań developed a multi-method framework for predicting weekly, non-contact lower-limb injuries by combining expert knowledge and machine-learning techniques (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023). GPS-derived external-load data: total distance, sprint and high-speed-running (HSR) distances, player-load, accelerations, decelerations and on-field time, were enriched with: their aggregates for the three preceding weeks, the acute-to-chronic ratios such as ACWR and finally six handcrafted expert rules capturing relative changes in sprint, HSR and player-load profiles; fuzzy linguistic variables further expanded the feature space

to model uncertainty (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

The full dataset comprised 1 064 micro-cycles (67 labelled injuries). It was partitioned into a training set of 693 observations and an independent test set of 371 (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023). Class imbalance in the training data (36 injuries) was mitigated with SMOTE, while missing values were median-imputed. Hyper-parameter tuning relied on 5×10 cross-validation, after which final evaluation was carried out exclusively on the untouched test subset containing the remaining 31 injury cases (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

Three modelling strategies were compared:

- Deterministic rule-based model: averaged the six expert-rule scores and raised an alert when the mean exceeded 6.5 (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).
- Fuzzy rule-based controller: generated a continuous risk score on the $[0, 10]$ scale, classifying risk as high when the score surpassed 0.6. Unlike crisp rules, fuzzy logic represents each input through linguistic labels (e.g., *low*, *medium*, *high*) with graded membership (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).
- Cost-sensitive Extreme Gradient Boosting (XGBoost): 500 trees, max depth 7, learning rate 0.05, trained on the full feature set and balanced with SMOTE (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

On the hold-out test set, XGBoost achieved the best performance (accuracy = 90.0 %, precision = 92.0 %, recall = 97.6 %, F1 = 94.7 %), clearly outperforming the rule-based approach (F1 = 53.0 %) and the fuzzy system (F1 = 18.7 %) (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023). Shapley value analysis identified total training time two weeks earlier, counts of decelerations and accelerations, and current-week HSR distance as the strongest predictors, highlighting the importance of rapid directional changes and cumulative load (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

Another study conducted on 34 professional football players from a Portuguese first-division club developed an automated system for daily prediction of non-contact injury risk, using GPS data collected over 36 weeks of the 2020–2021 season, covering a total of 255 sessions (Freitas et al., 2025). The initial dataset included 1,379 variables, which were reduced to 260 through a combination of zero-variance feature elimination (Freitas et al., 2025). The retained variables comprised 237 GPS metrics and 23 descriptive features, such as player position, day of the week, session type, and duration, with the addition of “dummy days” to capture sudden changes in training load. The binary target variable indicated the presence or absence of an injury on a given day (Freitas et al., 2025). To address the dataset’s high heterogeneity and imbalance, only 0.2% of records corresponded to injury events, the models were trained using both traditional and cost-sensitive approaches, coupled with stratified subject-independent cross-validation (Freitas et al., 2025). Three algorithms were compared: Support Vector Machines (SVM), three-layer Feedforward Neural Networks (FNN), and Adaptive Boosting (AdaBoost) with decision stumps (Freitas et al., 2025).

The best results were achieved by the cost-sensitive SVM, which reached a sensitivity of 71.43%, specificity of 74.19%, accuracy of 74.22%, and an AUC of 0.85, using only 20 features. AdaBoost exhibited a higher sensitivity of 78.57%, at the expense of specificity (65.02%) and required a larger number of features (Freitas et al., 2025). Despite the absence of physiological or clinical measurements, the study highlights how the integration of GPS and contextual variables can support short-term decision-making without requiring manual data collection (Freitas et al., 2025).

While the model proposed by Freitas et al. demonstrates the feasibility of injury prediction using only GPS and contextual data within a tightly controlled club-level environment, broader generalizability requires larger samples and more heterogeneous feature sets (Freitas et al., 2025). In this direction, a multi-season study explores a richer combination of physiological, psychological, historical, and workload indicators to model injury risk across a longer temporal horizon (Majumdar et al., 2024b).

This multi-season study analyses the association between training load and non-contact injuries, exploiting five years of data from thirty-five Premier League players for a total of 10 653 observations (Majumdar et al., 2024b). The information set comprises 106 variables: forty GPS metrics, fourteen physical measures (skinfolds, body mass, body-composition indices), four psychological indicators, six demographic characteristics, forty-two workload

ratios derived from ACWR, MSWR and EWMA, and two historical variables (*last injury area* and *days since last injury*) (Majumdar et al., 2024b). The binary target flags the onset of an injury within the subsequent seven-day window (Majumdar et al., 2024b)

The pronounced class imbalance, injuries represent 3.7 % of the sample, is mitigated through the Synthetic Minority Oversampling Technique and cost-sensitive weighting (Majumdar et al., 2024b). Model selection compares Extreme Gradient Boosting with a two-layer Artificial Neural Network (200–100 neurons, 0.5 dropout, Adam optimiser), both validated with ten-fold cross-validation; training incorporates the first four-and-a-half seasons, while testing involves the remaining half-season partitioned into three monthly subsets (Majumdar et al., 2024b).

On the test set, the neural network attains a recall of 77 %, a precision of 13 % and an AUC of 0.69, marginally outperforming XGBoost, which records 73 % recall and 10 % precision (Majumdar et al., 2024b). Post-hoc interpretability via Shapley Additive Explanations assigns the highest contributions to *last injury area*, body mass, EWMA of meta-energy, daily meta-energy, and age, indicating that historical and anthropometric variables exert greater predictive influence than many detailed GPS metrics (Majumdar et al., 2024b).

Taken together, the three empirical studies show that machine-learning models, whether gradient-boosted trees, cost-sensitive support-vector machines or neural networks, outperform rule or threshold-based approaches and, by integrating GPS load metrics with contextual, historical and anthropometric information, offer a club-level decisions aimed at reducing non-contact injuries in professional football. (Freitas et al., 2025; Majumdar et al., 2024b; Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

2.5. Gaps in the literature, research contributions and future perspectives

Despite the growing scientific output on the application of machine-learning (ML) techniques for injury prediction in football, numerous gaps and limitations still hinder the operational implementation and generalizability of existing models. A recurring issue is the limited predictive performance of current approaches, particularly in terms of sensitivity and precision. Although some studies report high specificity, sensitivity often falls below

clinically acceptable thresholds, between 15.2 % and 55.6 %, with area-under-the-curve (AUC) values generally ranging from 0.66 to 0.83 (Nassis et al., 2023).

Low precision further highlights the need for recalibrating decision thresholds or adopting more sophisticated cost-sensitive learning techniques before operational deployment (Majumdar et al., 2024b). In addition, several studies are built and validated on small, highly imbalanced datasets, often with fewer than 1,000 observations, conditions that undermine statistical robustness and model reliability (L. Wang, 2024). The limited number of injury cases constrains inferential power, indicating the importance of multi-season or rolling-origin validation schemes to support long-term generalizability (Freitas et al., 2025).

A further methodological concern is the risk of information leakage during model development, especially when temporal or player-specific dependencies are not properly handled in data splitting, potentially leading to an overestimation of predictive performance (Freitas et al., 2025). Moreover, the interpretability of complex algorithms, particularly deep-learning models, remains a significant barrier to their integration into technical-staff decision-making because of their black-box nature and limited transparency in feature attribution (L. Wang, 2024).

From a data-perspective, most investigations rely exclusively on external-load metrics obtained from wearable GPS devices, such as distance covered, accelerations and player load. This narrow focus overlooks physiological and psychological factors that substantially influence injury risk, thereby limiting the comprehensiveness of the predictive framework (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023)

Input variables have emerged as a critical limitation in current injury prediction models (Nassis et al., 2023). The absence of real-time indicators reduces model accuracy. Furthermore, low data veracity and poor standardisation across studies further compromise predictive performance (Nassis et al., 2023). Enhancing model robustness therefore requires the inclusion of high-quality, context-aware, and physiologically relevant variables (Nassis et al., 2023).

External validity is further constrained by datasets drawn from a single club or competition, which restricts the applicability of findings to wider athlete populations and organisational contexts (Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska & Dyczkowski, 2023).

Future research should expand existing datasets in both breadth and depth by incorporating multi-season time windows, data from multiple clubs and a more heterogeneous set of predictors, including biomarkers, subjective fatigue indicators and contextual variables. Greater attention should also be devoted to the interpretability of ML models through explainable-AI techniques that enhance practitioner understanding and trust. Such developments are essential to ensure that predictive tools are not only accurate but also acceptable and usable within professional-football environments (Nassis et al., 2023; L. Wang, 2024).

3. Methods and Materials

3.1. Aims of the chapter

The purpose of this chapter is to delineate the research design and methodological decisions undertaken to answer the research question: “*How machine learning models can improve the prediction of injury risk in football players, and which variables provide the most relevant information for forecasting purposes ?*”. Specifically, it presents the sample-selection criteria, the data sources, the instruments used for data processing, and the analytical techniques applied. The chapter centres on the quantitative approach adopted, grounded in the use of machine-learning algorithms, with the aim of identifying the most informative features and evaluating the predictive performance of different models.

3.2. Data collection

The empirical foundation of this study is based on a dataset constructed from performance and injury data relating to a selected sample of 100 professional football players active in Italy’s top-tier league, Serie A, in all three consecutive seasons: 2021/22, 2022/23, and 2023/24.

In order to ensure the generalizability of the results and avoid distortions in the modelling process, goalkeepers were deliberately excluded from the sample. Their physical demands, playing style, and injury profiles differ substantially from outfield players, and their inclusion would have introduced outliers and imbalanced the distribution of relevant variables.

The dataset was stratified across five positional groups:

- 20 Center Backs
- 27 Central Midfielders
- 13 Forwards
- 25 Fullbacks
- 15 Wide Midfielders

The sample was defined through a structured selection process, guided by quantitative and qualitative criteria aimed at ensuring the robustness and internal consistency of the analysis. Firstly, all players who did not take part in at least one match in each of the three seasons were excluded, as such lack of continuity would have compromised the possibility of a temporally consistent observation. Secondly, a minimum participation threshold of 40 total appearances across the three-year period was applied, including substitute appearances. This requirement was introduced to guarantee a sufficient volume of individual-level data, thereby supporting meaningful model training and reducing statistical noise. Finally, players who, despite meeting the continuity and participation criteria, had no registered injury throughout their professional career were also excluded. This category included, for instance, some younger players who, while having reached the appearance threshold, had not yet sustained any documented injuries. The exclusion of such cases was necessary to avoid incorporating observations with no variation in the target variable, a condition that could have undermined the effectiveness of the predictive model training. The data used for this study were collected and organized into two distinct CSV tables. The first table contains the performance data of each selected player, recorded for every official match played during the three-season period. All competitions are included without exception: Serie A matches, Coppa Italia fixtures, any national team appearances (in case of international call-ups), and European club competitions such as the UEFA Champions League, Europa League, and Conference League.

The second table compiles the complete injury history for each player, adopting the definition of injury based on the concept of *time-loss injury*, in accordance with the conceptual framework established by Fuller et al. 2006 and discussed in the literature review chapter. In line with this definition, only injuries that resulted in the player's inability to take part in at least one subsequent training session or match were included. Conversely, episodes classified as *medical attention injuries*, that is, injuries requiring treatment but not causing any interruption of football activity, were excluded. This methodological choice ensures that the dataset reflects only injuries with a measurable impact on player availability, which is essential for developing a predictive model that is both consistent and analytically meaningful.

Performance data collection

For the collection of individual performance data, the website *FBref* was used, a public and widely recognized platform that provides detailed statistics on professional football players and international football competitions.

The choice of this source was driven by its reliability, historical consistency, and completeness, as well as by the availability of comprehensive, match-level data, a fundamental requirement for constructing a dataset suitable for predictive modelling. Moreover, *FBref* adopts standardized variable names and follows a data collection methodology based on the *StatsBomb* framework, one of the most authoritative international providers of advanced football analytics.

Specifically, for each of the 100 selected players, the *FBref*'s section titled *Match Logs* was consulted for each relevant season, allowing access to data on every official match played.

Within this section, the match data are organized across multiple sub-sections, each focusing on a specific area of player performance. In particular, the following statistical categories are provided separately for each match:

- *Summary*
- *Passing*
- *Pass Types*
- *Goal and Shot Creation*
- *Defensive Actions*
- *Possession*
- *Miscellaneous Stats*

All of these tables were accessed and their contents were systematically merged to form a single, unified dataset. During this integration process, the variables that appeared in more than one sub-section with identical values, were identified and removed, retaining only one unique version of each.

The complete list of variables extracted from *FBref* and included in the dataset, along with their respective definitions, is reported below in *Table 1*.

Table 1: *Perfomarce features*

Column	Description
Player	Player Name
Born	Date of birth of the player
Height	Player height
Weight	Player weight
Date	Date listed is local to the match
Day	Day of week
Competition	Competition, Number next to competition states which level in the country's league pyramid this league occupies
Round	Round or phase of competition
Venue	If the player play in the Home stadium or Away
Result	Match Result
Squad	Refers to the team that the player belongs to
Opponent	Refers to the opposing team in the match
Starter	Describe if the player is in the starting lineup * = squad captain
Position	Position most commonly played by the player (GK - Goalkeepers, DF - Defenders, MF - Midfielders, FW - Forwards, etc.)
Min	Minutes played
Goals	Number of Goals scored
Assists	Number of Assists
PenaltyMade	The number of penalty kicks the player successfully made.
PenaltyAttempted	The number of penalty kicks attempted by the player
TotalShot	Shots Total (Does not include penalty kicks)
TotalShotonTarget	Shots on Target (Note: Shots on target do not include penalty kicks)
YellowCards	Number of Yellow Cards received during the match
RedCard	Number of Red Cards received during the match
Tackles	The number of successful tackles made by the player (how many players they successfully tackled)
Interceptions	The number of interceptions made by the player (the number of times the player intercepted a pass from the opponent).
xG	Expected Goals. This is a statistical measure of the likelihood of a player scoring based on the quality of their chances, including penalty kicks but excluding penalty shootouts.

npxG	Non-Penalty Expected Goals. This is the same as xG, but it excludes penalty kicks from the calculation.
xAG	Expected Assisted Goals. This represents the expected goals that were created by a player's assist, based on the quality of the assist.
ShotCreatingActions	Shot-Creating Actions. These are two offensive actions that directly lead to a shot, such as passes, take-ons, or drawing fouls. A player can be credited with multiple actions if they lead to a shot.
GoalCreatingActions	Goal-Creating Actions. These are two offensive actions that directly lead to a goal, such as passes, take-ons, or drawing fouls. Like SCA, a player can be credited with multiple actions.
PassesCompl	Passes Completed. This refers to the total number of passes made by the player that reached their intended target. Includes live ball passes (including crosses) as well as corner kicks, throw-ins, free kicks and goal kicks.
PassesAttempt	Passes Attempted. The total number of passes attempted by the player Includes live ball passes (including crosses) as well as corner kicks, throw-ins, free kicks and goal kicks
PassComplPerc	Pass Completion Percentage. This is the percentage of successful passes out of the total passes attempted. Includes live ball passes (including crosses) as well as corner kicks, throw-ins, free kicks and goal kicks.
ProgressivePasses	Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch
TotPassDist	Total Passing Distance: Total distance, in yards, that completed passes have traveled in any direction
ProgressivePassDist	Progressive Passing Distance: Total distance, in yards, that completed passes have traveled towards the opponent's goal. Passes away from the opponent's goal are counted as zero progressive yards.
ShortPassesCompl	The total number of passes completed by a player that are between 5 and 15 yards in distance.
ShortPassesAttempt	The total number of passes attempted by a player that are between 5 and 15 yards in distance.

ShortPassesComplPerc	The percentage of successful passes completed (between 5 and 15 yards) out of the total passes attempted in that range.
MediumPassesCompl	The total number of passes completed by a player that are between 15 and 30 yards in distance
MediumPassesAttempt	The total number of passes attempted by a player that are between 15 and 30 yards in distance.
MediumPassesComplPerc	The percentage of successful passes completed (between 15 and 30 yards) out of the total passes attempted in that range
LongPassesCompl	The total number of passes completed by a player that are longer than 30 yards.
LongPassesAttempt	The total number of passes attempted by a player that are longer than 30 yards.
LongPassesComplPerc	The percentage of successful passes completed (longer than 30 yards) out of the total passes attempted in that range.
xA	The expected assists measure the likelihood that each completed pass will become a goal assist. It is determined by the pass type, location, distance, and the phase of play. It is also provided by Opta.
KeyPasses	A key pass is a pass that directly leads to a shot. This is a measure of the number of passes that have the potential to result in a goal (an assisted shot).
PassesinFinalThird	Passes into Final Third: Completed passes that enter the 1/3 of the pitch closest to the goal (not including set pieces)
PassesinPenaltyArea	The total number of completed passes into the 18-yard box (penalty area). Set-piece passes are excluded from this measure.
CrossesinPenaltyArea	The total number of completed crosses into the penalty area (18-yard box), excluding set-piece crosses.
LivePasses	Passes that are actively in play during normal gameplay.
DeadPasses	Passes made during a stoppage in play (e.g., free kicks, corner kicks, throw-ins, goal kicks).
FreeKicks	Passes attempted from a free kick situation, either aiming for a shot or a teammate.
ThroughPasses	A pass played through the opponent's defense into open space, typically for a teammate to run onto.
Switches	Passes that cover a distance of more than 40 yards across the width of the field to create space on the opposite side.

Crosses	Passes played from the wings into the penalty area to create scoring chances.
ThrowIns	The number of throw-ins a player has taken during a match after the ball goes out of bounds on the sideline.
CornersKicks	The number of corner kicks a player has taken when the ball goes out over the goal line off a defender.
InswingingCornerKicks	Corner kicks that curve inward toward the goal, making them more difficult to defend.
OutswingingCornerKicks	Corner kicks that curve outward, away from the goal, providing a different angle of attack.
StraightCornerKicks	Corner kicks that are kicked straight with minimal curvature, typically aiming for a teammate.
PassesOffside	Passes attempted when the receiving player is in an offside position, closer to the opponent's goal line than both the ball and second-to-last defender.
PassesFailed	Passes that are blocked by the opponent, typically by positioning themselves in the ball's path.
LivePassesShotCreat	Completed live-ball passes that lead to a shot attempt.
DeadPassesShotCreat	Completed dead-ball passes (such as free kicks, corner kicks, kick-offs, throw-ins, and goal kicks) that lead to a shot attempt.
TakeOnShotCreat	Successful take-ons that lead to a shot attempt.
ShotsleadShot	Shots that lead to another shot attempt.
FoulsDrawnShotCreat	Fouls drawn that lead to a shot attempt.
DefensiveActionsShotCreat	Defensive actions that lead to a shot attempt.
LivePassesGoalCreat	Completed live-ball passes that lead to a goal.
DeadPassesGoalCreat	Completed dead-ball passes (such as free kicks, corner kicks, kick-offs, throw-ins, and goal kicks) that lead to a goal.
TakeOnGoalCreat	Successful take-ons that lead to a goal.
ShotsleadGoalshot	Shots that lead to another goal-scoring shot.
FoulsDrawnGoalCreat	Fouls drawn that lead to a goal.
DefensiveActionsGoalCreat	Defensive actions that lead to a goal.
TacklesWin	The number of tackles where the player wins possession of the ball from the opponent.
TacklesInitialThird	Tackles made in the defensive third of the field, closest to the player's own goal.
TackelsMediumThird	Tackles made in the middle third of the field.
TackelsFinalThird	Tackles made in the attacking third of the field, closest to the opponent's goal.
DribblersTackledSucc	This refers to the number of times a player successfully tackles a dribbler (an opponent who is carrying the ball and attempting to get past the defender)

DribblesChallenged	This refers to the total number of attempts by a player to challenge a dribbler. It includes both successful tackles and failed challenges. A failed challenge occurs when the defender doesn't win possession or allows the dribbler to get past.
PercentageTackledWin	This measures the percentage of dribblers successfully tackled out of all the attempts to challenge dribblers. It is calculated as the number of dribblers tackled divided by the number of attempts to challenge them
ChallengesLost	This refers to the number of unsuccessful attempts made by a defender to challenge a dribbling player. A challenge is considered lost if the defender does not successfully stop the dribbler or if the dribbler gets past them.
Blocks	The number of times a player blocks the ball by positioning themselves in its path.
ShotsBlocked	This is the number of times a player blocks an opponent's shot on goal by standing in the path of the ball.
PassesBlocked	This refers to the number of times a player blocks a pass made by the opponent, often by positioning themselves in the way of the ball to stop it from reaching its target.
Interceptions	The number of times a player intercepts an opponent's pass.
Clearances	This refers to the number of times a player clears the ball away from the defensive area, usually by kicking the ball out of the defensive zone
Errors	This refers to mistakes made by the player that lead directly to an opponent's shot on goal
Touches	The number of times a player touches the ball during a match. For example, receiving a pass, dribbling, and sending a pass counts as one touch.
DefPenTouches	The number of times a player touches the ball within their own penalty area when defending.
TouchesInitialThird	The number of times a player touches the ball within their defensive third of the field (closest to their goal).
TouchesMediumThird	number of times a player touches the ball within the middle third of the field.
TouchesFinalThird	The number of times a player touches the ball within the attacking third of the field (closer to the opponent's goal).
AttPenTouches	The number of times a player touches the ball within the opponent's penalty area when attacking.

LiveTouches	The number of times a player touches the ball during active play, excluding corner kicks, free kicks, throw-ins, kick-offs, goal kicks, or penalty kicks.
TakeOnsAttempt	The number of times a player attempts to dribble past an opponent.
SuccessfulTakeOns	The number of times a player successfully dribbles past an opponent.
TakeOnSuccPercentage	The percentage of successful take-ons (successful dribbles past an opponent) out of all attempted take-ons.
TackledDuringTakeOn	The number of times a player is tackled while attempting to dribble past an opponent.
TackledDuringTakeOnPerc	The percentage of take-on attempts in which the player is tackled by the defender.
Carries	The number of times a player carries the ball with their feet, as opposed to passing or shooting.
TotalCarryDistance	The total distance, in yards, that a player moves the ball while controlling it with their feet, in any direction.
ProgressiveCarryDistance	The total distance, in yards, that a player moves the ball toward the opponent's goal. This counts as progressive when the ball moves at least 10 yards from its starting point.
ProgressiveCarries	Carries that move the ball at least 10 yards toward the opponent's goal or into the penalty area. Carries ending in the defending half of the pitch are excluded.
CarriesIntoFinalThird	The number of times a player carries the ball into the final third of the field, which is closest to the opponent's goal.
CarriesIntoPenaltyArea	The number of times a player carries the ball into the opponent's 18-yard penalty box.
Miscontrols	The number of times a player fails to control the ball after attempting to receive it, often resulting in a turnover.
Dispossessed	The number of times a player loses possession of the ball after being tackled by an opponent. This does not include failed take-ons.
PassesReceived	The number of successful passes received by a player. A successful pass is one that reaches the player without being intercepted or going out of bounds.
ProgressivePassesReceived	The number of completed passes that move the ball at least 10 yards toward the opponent's goal or into the penalty area.
SecondYellowCard	Second Yellow Card received
FoulsCommitted	Number of Fouls Committed

FoulsDrawn	Number of Fouls Drawn
Offside	Number of offside
PenaltyWon	Number of Penalty Won
PenaltyConceded	Number of Penalty Conceded
OwnGoals	Own Goals
BallRecoveries	Number of loose ball recovered
AerialWon	Refers to the number of times a player successfully wins an aerial duel. An aerial duel occurs when the ball is in the air, and the player competes with an opponent to gain possession of the ball
AerialLost	Refers to the number of times a player loses an aerial duel. In this case, the player competes for the ball in the air but does not win possession, and the opponent gains control instead.
ArealSuccPerc	This value is the percentage of aerial duels won by a player compared to the total number of aerial duels they attempted. It is calculated as the number of aerial duels won divided by the total number of aerial duels (won + lost) and then multiplied by 100 to give the percentage

The performance data collection process represents the first pillar of the dataset construction. Data was collected for all official matches played by the selected players during the previously defined three-season period, resulting in a dataset of 12,634 rows, each corresponding to a single match appearance by one of the 100 players.

It is important to highlight that some performance metrics were not available on *FBref* for competitions such as: international matches played with national teams, UEFA Europa League, UEFA Conference League, and Coppa Italia. In these cases, the corresponding values were recorded as missing data, in order to preserve the integrity and chronological continuity of the match logs.

Injuries data collection

This section focuses on the second component of the dataset: the collection and structuring of injury records, which define the target variable for the supervised learning framework.

For the construction of the dataset, the *Injury History* section of *Transfermarkt* website was consulted, an internationally recognized source for football statistics and data analysis.

For each of the selected players, records of all time-loss injuries sustained during their professional careers were extracted, excluding injuries that occurred during youth-level competition, in order to ensure a complete and individualized coverage of professional injury histories.

The key fields extracted from *Transfermarkt* for each injury are summarized below (*Table 2*)

Table 2: *Injuries features*

Column	Description
Player	Player name
Season	The football season in which the injury occurred (e.g., 2021/22).
Injury	Description of the injury type (e.g., muscle injury, cruciate ligament tear).
From	Start date of the injury.
Until	End date of the injury.
Days	Duration of the recovery period, calculated as the difference between <i>Until</i> and <i>From</i> .
Games Missed	Number of official matches missed by the player due to the injury, including the relevant team name.

The final dataset obtained from this process contains a total of 1,575 injury events, each meeting the criteria of a time-loss injury and linked to a specific match timeline. These injuries span across 176 distinct categories, ranging from muscular and ligament-related traumas (e.g., hamstring strain, cruciate ligament tear) to surgical conditions, joint inflammations, and systemic illnesses (e.g., pubalgia, eye injuries, concussion).

3.3 Data preprocessing and features selection

The preprocessing phase began with the two previously described CSV files. As a first step, a semantic restructuring of the injury dataset was carried out through the addition of a new column containing a classification of injury events into four clinical macro-categories:

- *MTI – Muscular and Tendon Injuries* (e.g., strains, muscle tears, tendinopathies)
- *ALI – Articular and Ligamentous Injuries* (e.g., sprains, ligament tears, joint instability)
- *BCI – Bone and Contusion Injuries* (e.g., fractures, bone bruises, impact-related microtraumas)
- *Other – Systemic Conditions and Miscellaneous Causes* (e.g., illnesses, viral infections, non-traumatic clinical conditions)

The complete mapping of the injury types included in the dataset into the four categories above is presented in *Table 3*.

Table 3: Injuries categorization

Category	Description	Injuries reported
MTI: Muscular and Tendon Injuries	Tears, strains, ruptures, and inflammations of muscles and tendons	Abdominal muscle strain, Abdominal problems, Achilles Problems, Achilles heel problems, Achilles tendon contusion, Achilles tendon irritation, Achilles tendon rupture, Achilles tendon surgery, Adductor injury, Arch problems, Back problems, Calf muscle tear, Calf injuries, Calf strain, Chest injury, Contracture, Fatigue fracture, Hamstring injury, Hamstring strain, Hamstring muscle injury, Hip flexor problems, Hip Muscle Strain, Inflammation, Inflammation of the biceps tendon in the thigh, Leg injury, Muscle contusion, Injury to abdominal

		<p> muscles, Lumbago, Muscle fatigue, Muscle strain, Muscle tear, Muscle contusion, Muscle Fatigue, Muscle injury, Muscle problems, muscle stiffness, Muscle strain, Muscle tear, Muscular problems, overstretching, Patellar tendon dislocation, Patellar tendon irritation, Peroneus tendon injury, Pubalgia, Pubic irritation, Pubic stress, Right hip flexor problems, Sore Muscles, Strain in the thigh and gluteal muscles, Tendon irritation, Torn muscle fiber, Thigh injury, Thigh problems, Thigh strain, Torn thigh muscle, Tendon rupture, Torn muscle bundle, Torn muscle fiber, Torn muscle fiber in the adductor area, Torn thigh muscle </p>
<p>ALI: Articular and Ligamentous Injuries</p>	<p>Sprains, ligament injuries, meniscus lesions, and joint traumas</p>	<p> Ankle problems, Ankle injury, Ankle sprain, Ankle Surgery, Capsular tear of ankle joint, Collateral ligament injury, Cruciate ligament injury, Cruciate ligament Surgery, Cruciate ligament tear, Elbow injury, Groin injury, Groin problems, Hip tendon injury Inner knee ligament tear, , Inflammation in the head of fibula, Inflammation in the knee, Inflammation in the spine, Inflammation of ligaments in the knee, Injury to the ankle, Inner knee ligament tear, Inner ligament injury, Inner ligament stretch of the knee, Internal ligament strain, Knee collateral ligament strain, Knee medial ligament </p>

		tear, Knee inflammation, Knee injury, Knee problems, Ligament injury, Meniscus injury, Meniscus tear, Patellar tendon dislocation, Tear of the lateral meniscus, Syndesmosis ligament tear, Tendon irritation, Tendon rupture, Torn lateral knee ligament, Torn ligaments, Torn ligaments in the tarsus
BCI: Bone and Contusion Injuries	Fractures, contusions, hematomas, and impact-related injuries	Back injury, Broken arm, Broken cheekbone, Broken collarbone, Broken fibula, Broken foot, Broken nose bone, Bruise, Bruised on ankle, Bruised knee, Bruised ribs, Bone edema, Facial fracture, Finger injury, Foot bruise, Foot injury, Foot Surgery, Fracture, Head injury, Heel spur Rib fracture, Inflammation of pubic bone, Knee bruise, Knock, Laceration, Lumbar vertebra fracture, Metatarsal fracture, Minor knock, Nose injury, Nose surgery, Rib fracture, Scaphoid fracture, Scaphoid surgery, Shin bruise, Shoulder Injury, Stress reaction of the bone, Thumb injury, Tibia contusion, Toe injury, Wrist fracture,
Other: Systemic Conditions and Miscellaneous Causes	Fever, infections, viral illnesses, general non- traumatic conditions, and physical discomfort	Angina, Appendectomy, Bronchitis, Chickenpox, Cold, Concussion, Coronavirus, Covid-19, Dead leg, Eye injury, Fever, Fitness, Flu, Gastroenteritis, ill Influenza, Infection, Herniated disc, Inguinal hernia, Injury, ,

		Tonsillitis, , Intestinal virus, Lung contusion, , Fatigue fracture, Lack of fitness, Minor Injury, Physical discomfort, Pneumonia, Precautionary rest, Quarantine, Rest, sprain, Stomach flu, stomach problems, strain, Stress reaction of the bone, surgery, Testicular cancer, Tonsillitis, traffic accident, Virus, Unknown injury.
--	--	---

This aggregation was adopted with the aim of reducing the fragmentation of the original target variable, characterized by a high heterogeneity of textual descriptions, while retaining sufficient information to preserve the clinical and sporting relevance of the data. Moreover, this classification enables the extraction of features and the development of analyses adapted to specific injury categories, rather than treating injury as a generic and undifferentiated event.

Once the column containing the categorization of injuries was obtained, the two main datasets, one related to player performance and the other to injury records, were divided in two subsets: a training set, used to fit the predictive models, and a test set, reserved for the subsequent evaluation of their generalization ability.

The split was performed according to a chronologically ordered 80 percent (training) to 20 percent (test) ratio, a widely adopted practice in the machine learning literature that ensures a suitable trade-off between model learning and robust evaluation. Rather than being defined through random sampling, the division followed a temporal logic, considered more appropriate given the predictive goals of this study and the inherently sequential structure of time-stamped football data.

Specifically, the training set included all official matches played during the 2021/22 and 2022/23 seasons, as well as those up to December 5th, 2023 (part of the 2023/24 season). The remaining data from the current season were assigned to the test set. This same criterion was consistently applied to both datasets (performance and injury records), thus ensuring temporal coherence between predictors and target labels.

The adoption of a time-based split addresses two fundamental concerns. Firstly, it avoids data leakage, that is, the unintentional use of future information during the training phase, which would invalidate the statistical rigor of the evaluation. Secondly, it reproduces a realistic operational scenario, where predictions must be made based solely on historical data available at a given time.

Subsequently, the data preprocessing phase was carried out using *Orange3*, an open-source environment for data mining and machine learning, built around a visual interface based on interactive workflows. The environment enables the construction of analytical pipelines through connection between functional modules, known as *widgets*, each of which performs specific operations such as variable transformation, statistical analysis, feature selection, and model training. For the purposes of this study, version 3.38.1 of *Orange3* was employed, running within *Anaconda*, an integrated platform for managing *Python* environments and specialized data analysis packages. The integration between the tools allowed for the combined use of visual tools and custom *Python* scripts, ensuring both the transparency and traceability of all operations, and a high degree of flexibility and control over the transformations applied to the dataset.

The following section presents the main components of the *Orange3* workflow, each accompanied by a detailed description. The objective is to provide a transparent account of the data transformations performed, the reasoning behind each methodological decision, and the implications of the operations carried out.

The accompanying figures document the workflow, while the corresponding explanatory notes clarify the specific role and contribution of each step within the overall data processing pipeline.

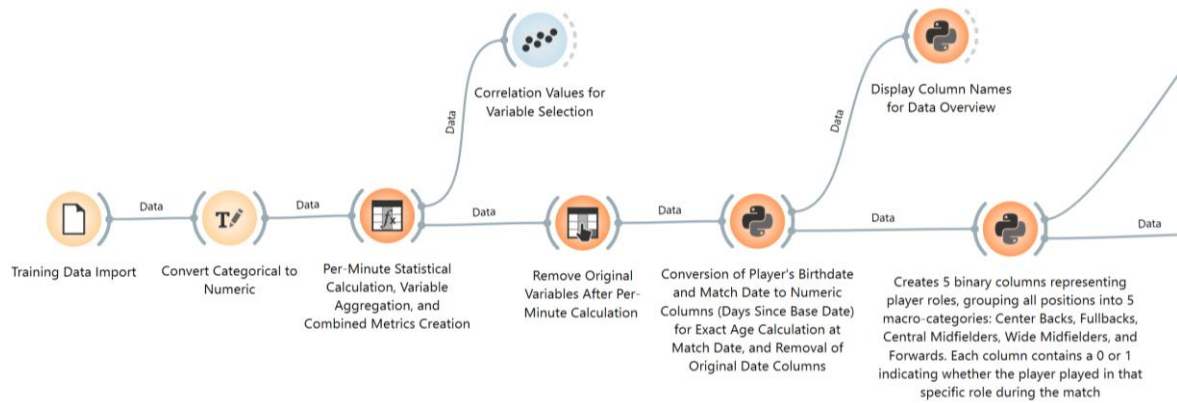


Figure 2a – Initial section of the preprocessing workflow in Orange3

Figure 2a illustrates the initial section of the preprocessing workflow implemented in Orange3. The workflow begins with the upload the CSV file containing the performance dataset.

Upon import, each variable in the dataset is automatically assigned a specific data type, chosen among the following:

- *Text*: non-numeric string values, typically used for identifiers, such as player names or team names.
- *Categorical*: discrete values representing qualitative information, such as match outcome, player position or match venue (home or away).
- *Numeric*: continuous or discrete numerical values, suitable for mathematical operations and statistical modelling. Examples include minutes played, number of passes, or shots taken.

In addition to the data type, each variable is also assigned a functional role within the Orange3 environment:

- *Feature*: a predictive variable used as input by machine learning algorithms.
- *Meta*: a contextual or descriptive variable, not used directly for training but retained for interpretation, filtering, or result annotation.

The second widget in the workflow was employed to correct the data type assigned to certain variables during the initial import process. Specifically, variables such as *Penalty Won*, *Penalty Conceded*, *Own Goals*, and *Penalty Attempted* were erroneously classified as

categorical, due to their low cardinality and discrete integer values (e.g., 0, 1, 2). While such classification may appear syntactically appropriate, it does not accurately reflect the semantic and functional nature of these variables within a predictive modelling context.

A *Formula* widget was subsequently employed to generate a set of normalized performance variables by dividing selected metrics by the number of minutes played in each match. This transformation allowed for the extraction of intensity-based indicators, capturing not only the volume of an action but also its frequency relative to the player's actual time on the field. By accounting for time exposure, the normalization mitigates variance associated with unequal match durations, and supports more reliable statistical learning.

Moreover, the decision to express variables on a per-minute basis ensures dimensional compatibility with widely adopted composite indicators in the sports science literature, such as the Acute:Chronic Workload Ratio (ACWR), which rely on temporally standardized units. This preprocessing step facilitates the integration of diverse performance metrics into cumulative workload models without structural inconsistencies.

The variables normalized through this transformation are summarized below (*Table 4*)

Table 4: Variables Normalized per Minute

Original Variable	Normalized Variable
Touches	Touches_min
TakeOnsAttemp	TakeOnsAttemp_min
TotalCarryDistance	TotalCarryDistance_min
PassesAttempted	PassesAttempted_min
TotalShot	TotalShot_min
DribblerChallenged	DribblerChallenged_min
Carries	Carries_min
ProgressiveCarryDistance	ProgressiveCarryDistance_min

In addition, using the same *Formula* widget, several new variables were created through the aggregation of related metrics and subsequent normalization by minutes played. This

process produced aggregated features that can capture multifactorial stress patterns and biomechanical demands that may not be evident from singular metrics.

Table 5: *Aggregated and Normalized Variables*

New Variable	Formula	Description
AerialDuelsTotal_min	$(\text{AerialLost} + \text{AerialWon}) / \text{Min}$	Frequency of aerial duels, indicative of player involvement in physical contacts.
FreeKicksStats_min	$(\text{FreeKicks} + \text{CornersKicks} + \text{PenaltyAttempted}) / \text{Min}$	Volume of set pieces taken, potentially stressing lower limbs through repetitive technical execution.
ActionInPenaltyArea_min	$(\text{DefPenTouches} + \text{AttPenTouches} + \text{CarriesIntoPenaltyArea}) / \text{Min}$	Actions in the penalty area, a zone of heightened collision risk.
LongPassesStats_min	$(\text{Crosses} + \text{LongPassesAttempt} + \text{Clearances}) / \text{Min}$	Long-range deliveries involving ballistic movements and muscular exertion.
DefensiveImpactActions_min	$(\text{Tackles} + \text{FoulsCommitted}) / \text{Min}$	Defensive challenges and fouls with high-intensity body impacts.
DefensiveBlocks_min	$(\text{ShotsBlocked} + \text{Interceptions}) / \text{Min}$	Defensive blocks, indicative of dynamic, high-load interventions.
OffensiveImpactActions_min	$(\text{TakeOnsAttempt} + \text{FoulsDrawn}) / \text{Min}$	Offensive 1v1 actions and fouls suffered, associated with muscular

		overload and sudden acceleration
--	--	----------------------------------

Finally, to capture individual morphological factors potentially associated with injury propensity, the Body Mass Index (BMI) was computed using the following formula:

$$BMI = \frac{Weight}{Height^2} \times 10000$$

The multiplicative factor of 10,000 was applied to account for the measurement units in the dataset, where height was expressed in centimetres, thereby converting the denominator to square meters. This calculated variable allowed for the inclusion of anthropometric characteristics in the predictive framework, supporting the hypothesis that physiological profile may modulate injury risk.

To further refine the feature space, a *Correlation* widget was connected to the *Formula* widget to evaluate the linear relationships among the variables. This step allowed for the identification and subsequent removal of features exhibiting excessively high correlation. The reasoning for this choice is grounded in a well-established principle of statistical learning: highly correlated predictors introduce redundancy and may compromise the performance of machine learning models. The presence of multicollinearity can lead to unstable parameter estimates, hinder the interpretability of feature importance, and increase the risk of overfitting, as the model may end up memorizing redundant patterns instead of learning the underlying data structure.

Moreover, the removal of redundant features improves computational efficiency and reduces the dimensionality of the dataset, thereby enhancing both the robustness and interpretability of the resulting models.

Following the feature selection performed through the *Select Columns* widget, which enabled the isolation of only those variables deemed relevant to the analysis pipeline, a subsequent *Python Script* widget was employed to convert two text-based columns: *Date* and *Born*, into numerical representations. Specifically, both variables were transformed into the number of days elapsed since January 1, 2000, resulting in the creation of two continuous variables: *Date_Days* and *Born_Days*. This transformation was a prerequisite for enabling dynamic calculations of each player's age at the time of every match. By aligning temporal and biographical data in a common numeric format, it became possible to compute match-

specific age values directly, which are essential in evaluating how age-related physiological changes may influence injury susceptibility throughout a player’s career.

To reduce the heterogeneity inherent in the original “Roles” column, where each match record described the specific positions covered by the player using role codes (e.g., “CM”, “AM”, “LW”), a *Python script* was employed to convert this categorical information into five binary macro-role indicators. Each of these new columns corresponds to a broader tactical category and takes value 1 if the player performed that macro-role in the given match, and 0 otherwise. Importantly, as players can occupy multiple tactical roles within a single match, the binary indicators are not mutually exclusive. This transformation served two key purposes. Firstly, it reduced dimensionality and increased interpretability by consolidating a large number of detailed roles into a manageable and analytically meaningful structure. Secondly, it enabled the model to capture role-specific injury risks while preserving positional flexibility, which is essential for reflecting the dynamic nature of player deployment in elite football.

The mapping applied for this transformation is summarized in *Table 6*.

Table 6: Role Categorization into Binary Macro-Positions

Macro-Role	Role Codes Included
Center Backs	CB
Fullbacks	FB, LB, RB, WB
Central Midfielders	DM, CM, AM
Wide Midfielders	LM, RM, LW, RW
Forwards	FW

To complement the initial preprocessing phase, *Figure 2b* presents the subsequent section of the *Orange3* workflow.

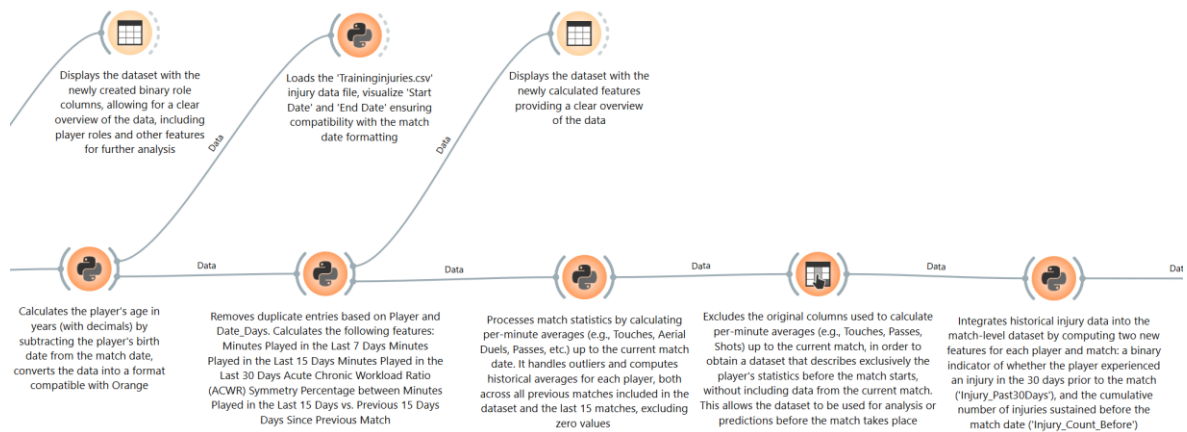


Figure 2b – Second section of the preprocessing workflow in Orange3

The first widget illustrated in *Figure 2b – Second section of the preprocessing workflow in Orange3* computes the player's age on the exact date of each match by applying the following formula:

$$Age = \frac{(Date_{Days} - Born_{Days})}{365.25}$$

The divisor 365.25 is used to account for leap years and convert the result into years with decimal precision. This transformation allows for the inclusion of a continuous variable capturing the player's precise age at the time of each appearance.

The subsequent *Python Script* widget extends the set of derived features and incorporates additional temporal and morphological indicators. These variables, described below, were derived by aggregating the minutes played by each athlete within predefined temporal windows preceding each match, using the chronological match dates as reference points. This allowed for a time-sensitive reconstruction of workload dynamics leading up to every game.

- *MinutesLast7days*, *MinutesLast15Days* and *MinutesLast30Days*: cumulative minutes played over the previous 7, 15 and 30 days, respectively. These serve as intermediate and chronic workload proxies, providing a broader picture of the player's physiological load.
- *ACWR (Acute Chronic Workload Ratio)*: calculated as the ratio between *MinutesLast7Days* and the average weekly load over the previous five weeks (i.e., *MinutesLast35Days* / 5). This metric, widely adopted in sports science, serves as an

indicator of training-load imbalance, with extreme values often associated with elevated injury risk.

- *SymPctMinutesLast15vsPrev15*: symmetric percentage change in minutes played between the most recent 15-day period and the preceding 15-day window. This feature captures fluctuations in workload, which may not be detected by raw accumulation metrics and are considered potential red flags for injury onset.
- *DaysSincePrevMatch*: number of days since the player's last official match. This indicator helps quantify recovery time and detect instances of fixture congestion, both of which may influence injury susceptibility.

Subsequently, all performance variables previously normalized on a per-minute basis, intended to capture the intensity of player actions during individual matches, were employed to transition the dataset from a post-match to a pre-match structure.

The aim of this transformation was to align the dataset with a realistic predictive scenario, in which the algorithm must rely exclusively on information available before the start of each match. To achieve these two additional types of features were computed:

- a cumulative historical average, calculated across all matches played by the player starting from their first appearance in the dataset up to, but excluding, the current match.
- a 15-day moving average, based solely on matches played during the 15 days preceding the match in question.

This approach enables the model to learn not only the player's typical workload levels over time but also short-term fluctuations in exposure, thereby introducing a dynamic and context-sensitive component. In particular, the ability to compare recent versus long-term averages enhances the model's capacity to detect acute spikes or sudden drops in load, conditions widely recognized in the sports science literature as strong predictors of injury risk.

In order to obtain a dataset that was entirely pre-match compliant, all original per-minute variable, those derived directly from post-match statistics, were subsequently removed using a Select Columns widget. This ensured that each remaining variable represented only information that would have been available prior to the match.

The result is a dataset in which each row constitutes a valid ex ante observation, meaning it is constructed entirely from information that would have been available prior to the match it refers to. This structure is essential to ensuring the logical and methodological integrity of any predictive application in a real-world sports context.

Integration between Performance and Injuries dataset

Once the construction of the performance dataset was completed, it was possible to proceed with the integration of the injury dataset. This operation aimed to enrich each row with historical data on injuries sustained by the same player up to that point, rigorously excluding any information pertaining to events occurring after the date of the observed match.

Following this integration, two additional variables were derived to enrich the dataset with synthetic indicators of each player's individual injury history. The first, labelled *Injury_Count_Before*, represents the cumulative number of documented injuries sustained by a player before the date of the match in question. This is a dynamic variable whose value increases over time in relation to prior injury events, allowing the model to assess whether a higher incidence of past injuries may be associated with an increased probability of future occurrences.

The second variable, *Injury_Past30Days*, is binary and takes the value 1 if the player had completed an injury spell within the 30 days preceding the observed match, and 0 otherwise. The purpose of this feature is to capture potential residual risk conditions, such as cases in which an athlete returns to play shortly after an injury episode, a situation that could be associated with an elevated likelihood of recurrence or overload.

The subsequent steps of the data preprocessing pipeline are depicted in the following figure which illustrates how the integration of the injury dataset is finalized and how the target variables are computed to support the supervised learning framework.

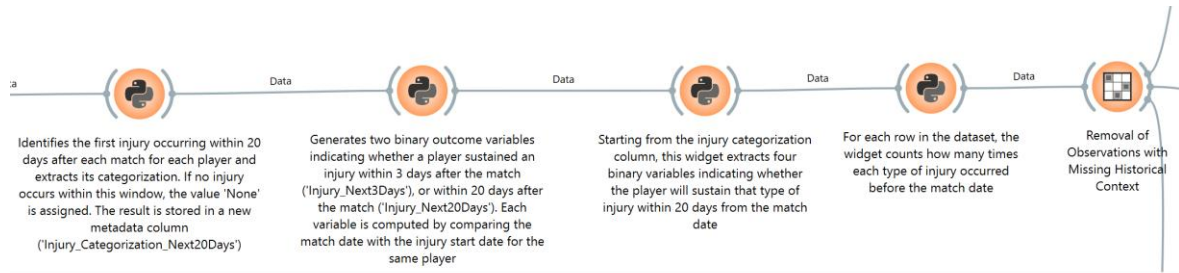


Figure 2c - Third section of the preprocessing workflow in Orange3

At this stage, for each observation, the player's injury history is queried to determine whether a new injury event begins within the 20 days following the date of the match. If an injuries occur within this timeframe, the earliest event in chronological order is identified, and its corresponding categorization (MTI, ALI, BCI, or OTHER) is assigned to the new field *Injury_Categorization_Next20Days*. In the absence of any injury, the variable is set to "None."

Subsequently, two additional binary indicators were computed to capture the presence or absence of injury events. These variables were derived by comparing the match date with each player's individual injury history, in order to determine whether a new episode occurred within a specific post-match time frame.

Specifically, the variable *Injury_Next20Days* takes the value 1 if the player sustained an injury within the 20 days following the match, and 0 otherwise. This feature is less specific than the first one, maintaining the observation window at medium term and serves as a proxy for assessing cumulative workload management and medium-term injury risk.

The variable *Injury_Next3Days* applies a more stringent threshold, identifying whether an injury occurred within three days of the match. This narrower window is particularly suited to capturing acute overload conditions or insufficient recovery, enabling the analysis of situations in which the player's performance or physical state in a short term may have directly contributed to the injury onset.

The combined definition of these three target variables was designed to support varying levels of predictive complexity. *Injury_Categorization_Next20Days* introduces a multi-class classification task, requiring the model not only to predict the occurrence of an injury but

also to distinguish among its main clinical categories. Although this approach increases the modelling complexity, it yields more informative and operationally actionable outputs for coaching and medical staff.

In contrast, the two binary variables (*Injury_Next20Days* and *Injury_Next3Days*) enable a more manageable exploration of the problem through standard binary classification tasks, aimed solely at predicting whether an injury will occur. These variables offer a valuable basis for comparative model evaluation, depending on the time horizon of interest.

A subsequent filtering step was undertaken to eliminate all observations that lacked an adequate historical context, namely, the first recorded match for each player. The absence of prior data prevented the computation of many engineered features, especially those based on temporal aggregations or injury history, thereby generating missing values flagged as “?”. Because these rows offered no predictive value and risked undermining both model training and evaluation.

All preprocessing steps detailed thus far, ranging from data cleaning to normalisation and feature engineering, were fitted on the training set and then identically reapplied to the test set using the parameters learned during training, thereby guaranteeing feature-space consistency and enabling an unbiased evaluation of the trained algorithms.

Finally, to tailor the analysis to each prediction task, a dedicated workflow was configured, complete with the selected learning algorithms and the corresponding evaluation widgets. The three pipelines are summarised in the Figures below, where *Figure 2d* depicts the workflow for *Injury_Categorization_Next20Days*, *Figure 2e* presents the workflow for *Injury_Next20Days* and *Figure 2f* illustrates the pipeline developed for *Injury_Next3Days*.

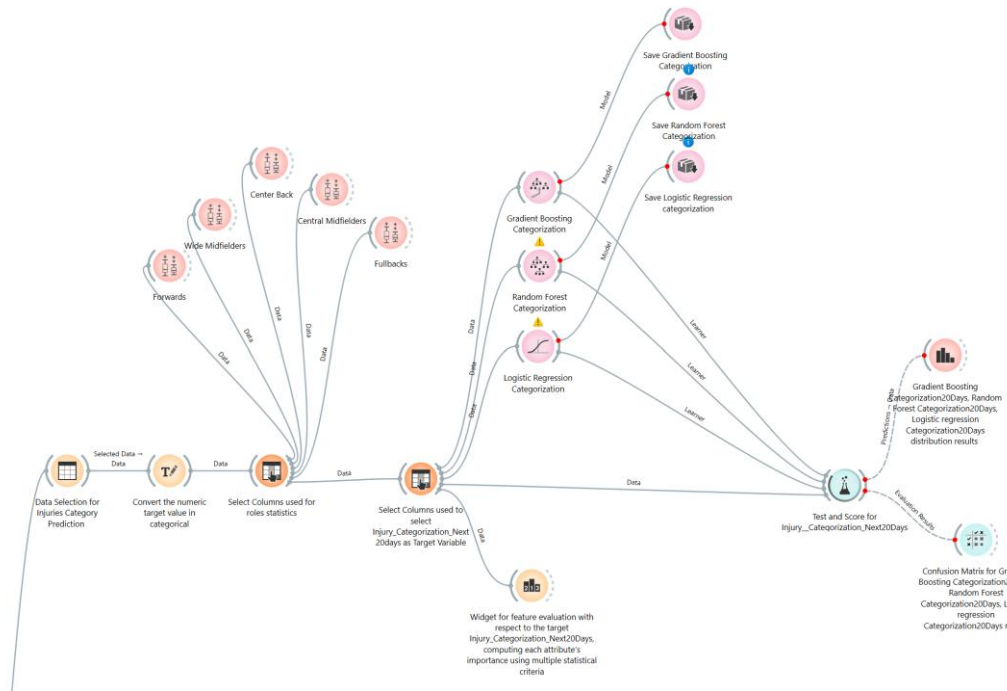


Figure 2d – Prediction workflow for Injury_Categorization_Next20Days

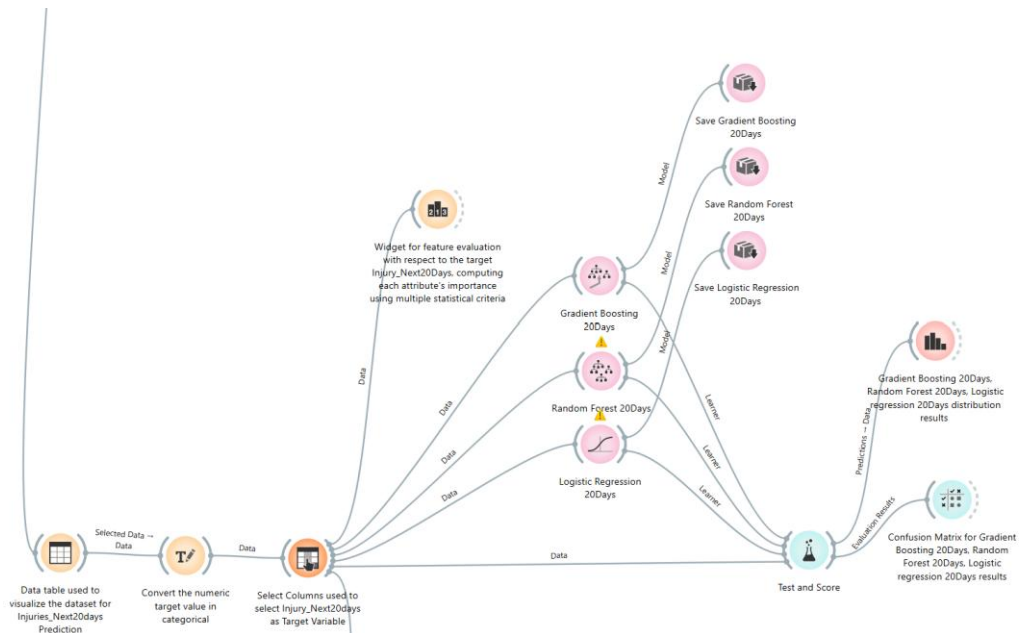


Figure 2e – Prediction workflow for Injury_Next20Days

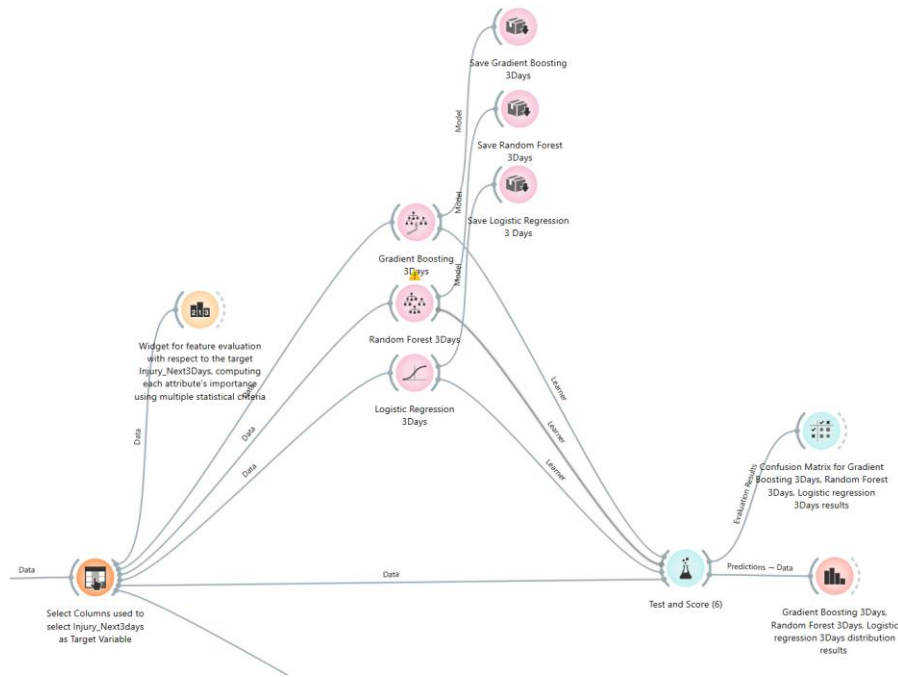


Figure 2f – Prediction workflow for *Injury_Next3Days*

To address the pronounced class imbalance in the target variables, *Injury_Next3Days* and *Injury_Next20Days*, a resampling strategy was adopted. The native dataset, designed to mirror the real-world incidence of injuries, exhibited a substantial disparity between positive and negative cases; while this asymmetry reflects clinical prevalence, it cause supervised models to favour the majority class at the expense of correctly identifying rare events.

As depicted in *Figure 2g* for *Injury_Next20Days* and *Figure 2h* for *Injury_Next3Days*, an alternative branch of the workflow was introduced to assess the effect of training on rebalanced datasets, employing the *Synthetic Minority Oversampling Technique for Nominal and Continuous* features (SMOTENC) via a *Python* widget. For each continuous numerical attribute, SMOTENC generates new minority instances by interpolating between a genuine minority case and one of its nearest minority neighbours. When the attribute is numeric but restricted to integer values, the synthetic value is subsequently rounded to the nearest integer and clipped to the empirical minimum–maximum range, preventing implausible outcomes. Categorical variables, which cannot be sensibly interpolated, are replicated by assigning the most frequent category observed among the reference instance and its neighbours, ensuring that every synthetic level already belongs to the attribute’s original domain.

Two distinct rebalancing scenarios were implemented. For *Injury_Next3Days*, a moderate oversampling increased the share of positive cases from 3 % to 10 % while preserving a credible class distribution; for *Injury_Next20Days*, a slightly more pronounced oversampling raised the positive proportion from 16 % to 25 %, obtaining a ratio of one positive instance for every four observations. The rebalanced datasets were subsequently employed to train predictive models in parallel with models fitted on the unaltered data, in order to verify whether a modest rebalancing could improve the detection of rare events.

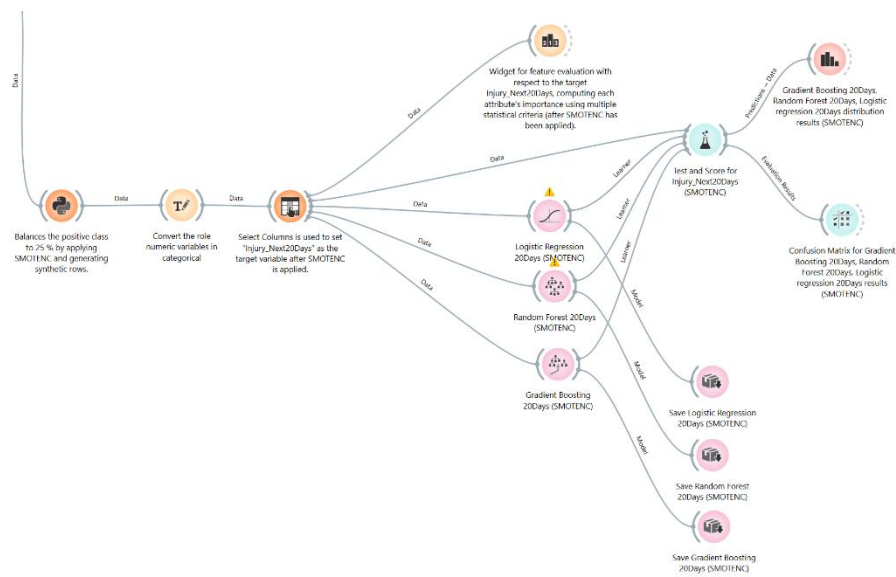


Figure 2g – Prediction workflow for *Injury_Next20Days* with SMOTE applied

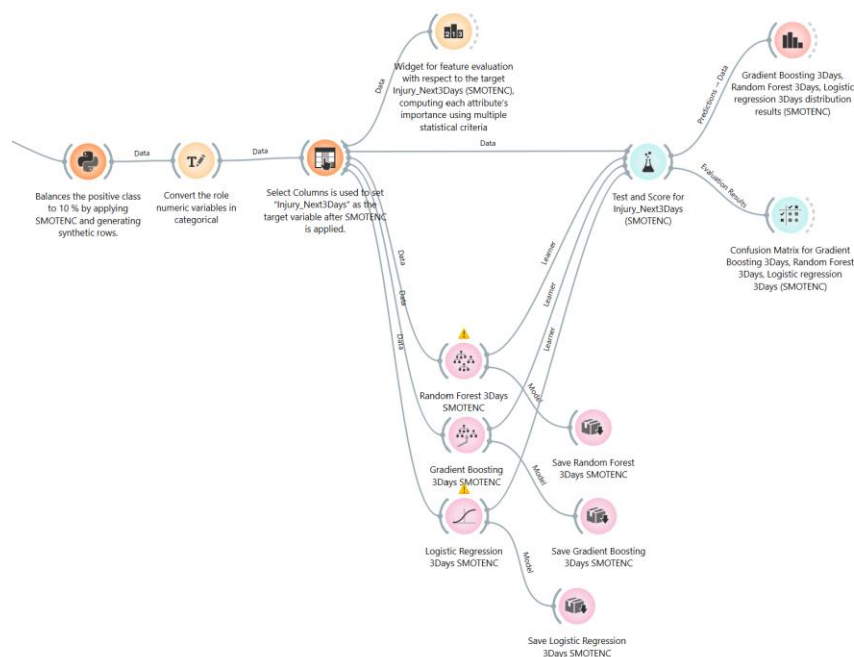


Figure 2h – Prediction workflow for *Injury_Next3Days* with SMOTE applied

3.4 Model selection and evaluation

All performance metrics discussed in this chapter were computed using a stratified 20-fold cross-validation applied exclusively to the training set. This technique ensures that each fold maintains the same class distribution as the original dataset, which is particularly important in settings with class imbalance.

Compared to more common configurations such as 10-fold cross-validation, the choice of 20 folds reflects a trade-off tailored to the specific characteristics of this study. Given the relatively large number of observations in the training set and the low presence of positive cases, a higher number of folds is an option that allows for more reliable validation without excessively reducing the number of observations for each training subset. This enhances the reliability of performance estimates, especially for rare-event detection.

As illustrated in the figures of the previous chapter, each of the three prediction features (*Injury_Categorization_Next20Days*, *Injury_Next20Days*, and *Injury_Next3Days*) was modelled and evaluated using the same trio of predictive algorithms: Gradient Boosting, Random Forest, and Logistic Regression.

- Gradient Boosting incrementally trains a sequence of weak decision trees, each one focused on correcting the residual error of the ensemble built so far (IBM, 2025). This sequential refinement allows the model to learn intricate non-linear interactions across heterogeneous features and to balance bias and variance effectively, an advantage when positive cases are scarce, though it increases computational cost and, if excessively deep or numerous trees are used, can heighten the risk of over-fitting (IBM, 2025).
- Random Forest aggregates a large number of decorrelated decision trees, each grown on a bootstrap sample, in example a training subset drawn with replacement, and on a random subset of predictors (IBM, 2025). The bootstrap procedure lowers model variance and confers robustness to noisy or partially missing data, while built-in estimates of feature importance support domain validation; the trade-offs are higher memory consumption and reduced immediate interpretability compared with a single tree (IBM, 2025).

- Logistic Regression provides a fast, linear baseline that outputs well-calibrated class probabilities and remains readily interpretable, making it suitable for communicating results to non-technical stakeholders (IBM, 2025). Its linear decision surface, however, may underperform when relationships among variables are strongly non-linear or when multicollinearity is present (IBM, 2025).

To rigorously assess classifier performance on the training set, four metrics were adopted: F1-score, Matthews Correlation Coefficient (MCC), Confusion Matrix, and Log-loss. Each parameter is selected to address a specific methodological need within the injury prediction context.

The F1-score is the harmonic mean of Precision and Recall:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}.$$

This formulation penalises imbalances between the two components and is therefore particularly suited to settings with rare positive cases, as is the case with imminent injuries. In this study, the primary objective is to maximise the model's ability to detect as many injuries as possible while keeping false alarms within acceptable bounds. Relying on Precision or Recall in isolation would mean favouring one aspect at the expense of the other, ultimately compromising the decision-making balance. The F1-score encapsulates this trade-off in a single scalar measure.

Matthews Correlation Coefficient (MCC)

The MCC combines all four entries of the confusion matrix and returns a value between -1 and +1 (with 0 indicating random classification). Unlike Accuracy or AUC, it remains informative even under severe class imbalance because it simultaneously balances true and false positives and negatives. In this research it is used as an indicator of overall robustness: a high MCC confirms that the model behaves consistently across the entire set of observations, not only within the minority class.

Confusion Matrix

The confusion matrix reports the counts of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Although it is not a scalar metric, it constitutes the most direct tool for operationally interpreting model outcomes, as it immediately quantifies the number of injuries correctly predicted and those that were missed.

Log-Loss

Finally, Log-loss was included as a metric to assess the probabilistic accuracy of predictions rather than simply the correctness of binary classifications. It evaluates how close the predicted probabilities are to the true outcomes, assigning heavy penalties to incorrect predictions made with high confidence. In other words, a classifier that assigns a high probability to the wrong class is penalised far more than one that makes the same mistake with uncertainty.

Within the scope of this research, Log-loss is used to assess model calibration, or the ability to provide reliable estimates of injury risk in probabilistic terms. This becomes particularly relevant in operational applications: for instance, two observations may both receive a negative binary prediction (no injury expected), yet the predicted probabilities might differ significantly: 0.10 in the first case, 0.40 in the second. Although both outputs result in the same classification, the latter indicates a much higher underlying risk, which may warrant a preventive medical intervention.

Therefore, the inclusion of Log-loss enables the evaluation to move beyond the binary outcome and towards a more nuanced understanding of the informational quality of the predicted probabilities. In a decision-making environment where the intervention threshold may vary depending on resource constraints or medical priorities, having access to well-calibrated risk estimates represents a tangible strategic advantage.

Taken together, these four metrics allow for a comprehensive, multi-level assessment of model performance.

3.5. Hyperparameters tuning

This chapter provides a discursive account of the hyper-parameter configurations adopted for the three algorithms under consideration, Logistic Regression, Random Forest, and Gradient Boosting, while maintaining a uniform setup across the various target variables, with the aim of ensuring a comparability of results.

Gradient Boosting

The Gradient Boosting model was configured with 100 trees and a learning rate of 0.10. In the initial configuration, the model was trained with a lower number of trees and reduced tree depth; however, under these conditions, it failed to effectively identify injury cases, producing nearly null sensitivity in all classification tasks. This limitation highlighted the need to enhance model capacity to capture the rare and multifactorial nature of injury events. As a result, both the number of trees and the maximum depth were increased in a controlled manner. This combination represents a balanced trade-off between predictive capacity and training stability: enough trees allow the model to gradually correct residual errors, while a moderate learning rate prevents each individual tree from having an excessive impact.

Each tree was limited to a maximum depth of 20 levels, enabling the model to capture complex interactions among predictors, such as those between workload, recovery time, and injury history. Greater depths would have produced overly specific trees, increasing the risk of replicating noise present in the training data. Conversely, a lower number of trees would have limited the model's ability to learn the underlying structure of the phenomenon to be predicted.

A minimum of five observations per internal node was also enforced to allow further splits. Lower values would make the model more sensitive to local fluctuations, while excessively high values would reduce the model's adaptability.

The sub-sampling parameter, which determines the fraction of observations used by each tree, was set to 1.00, meaning that all available data were used at every iteration.

Random Forest

The first hyperparameter that characterises the Random Forest model is the number of trees composing the ensemble. In general, increasing the number of trees helps reduce the variance of the model and improves its overall stability. However, beyond a certain threshold, the marginal gains become negligible. Conversely, setting this value too low may fail to provide sufficient coverage of the feature space, particularly in the presence of heterogeneous predictors. As with Gradient Boosting, the first experiments conducted with a lower number of trees led to unsatisfactory results, with the model showing limited ability to identify injury cases across all targets, this has led to a progressive increase in the number of trees. In this study, the number of trees was fixed at 100, a value which, according to both the literature and preliminary experimentation, represents a practical compromise, beyond which the error tends to stabilize and further increases yield only minimal improvements at the cost of longer training times.

A second relevant parameter is the number of candidate variables evaluated at each split. In a Random Forest, only a random subset of predictors is considered at each node to encourage diversification among the trees and to reduce the risk that dominant variables influence the entire ensemble. Increasing this number tends to reduce variance but may also raise the correlation between trees, while excessively low values can increase diversity at the cost of potentially overlooking relevant predictors. In this case, the value was set to 10, which approximates \sqrt{p} , where p is the total number of predictors (≈ 50). This choice allows the model to maintain low correlation between trees while ensuring that the most informative variables are not systematically excluded from the decision process.

Regardless of whether SMOTENC was applied or not, class balancing was enabled within the bootstrap samples used to train the individual trees. In this context, a bootstrap sample refers to a subset of the original dataset generated by randomly drawing observations with replacement, meaning that the same instance may appear multiple times while others may be excluded. Enabling class balancing within these samples ensures that positive cases are adequately represented in the training of each tree, thus reducing the risk that the model neglects the minority class during learning.

Finally, for the *Injury_Next3Days* target variable only, three additional constraints were introduced to mitigate overfitting risks: the maximum number of trees in the ensemble was

reduced to 8, the maximum depth of each individual tree was capped at 5 levels, and splitting was inhibited on subsets containing fewer than 10 samples.

Logistic regression

The Logistic Regression model was configured using L2 (Ridge) regularization, which introduces a penalty term into the loss function in order to constrain the magnitude of the coefficients. This helps control the model's variance and counteracts potential multicollinearity, particularly relevant in this study, which involves a large number of engineered and potentially correlated predictors.

The strength of the regularization is governed by the C parameter, set to 1.0. Preliminary analyses showed that this value represents a suitable compromise: it minimises the log-loss while maintaining a good level of recall. Stronger regularization led to underfitting, especially for the minority class, whereas weaker constraints increased variance without yielding meaningful improvements in F1-score or MCC.

To address the residual class imbalance, again automatic class weighting was enabled, regardless of whether SMOTENC was applied upstream. This option assigns greater penalties to misclassified instances of the minority class, enhancing the model's sensitivity to rare events.

Overall, the hyperparameter configurations adopted for each algorithm reflect a balance between model complexity, predictive performance, and robustness to class imbalance.

4. Results

To address the research question guiding this study, this chapter is structured into three distinct sections. The first section presents the results obtained during cross-validation on the training set, with the aim of evaluating the models in terms of accuracy, robustness, and predictive reliability. The second section reports the results achieved on the test set, enabling an objective assessment of the generalization capabilities of the selected models. Finally, the third section is dedicated to analysing the relative importance of predictors, in order to identify the variables that contributed most significantly to the algorithmic decisions, despite the absence of data related to training sessions, biometric indicators, or physiological measurements. In line with the quantitative approach adopted, all results are presented in tabular and graphical form, without subjective interpretation, which will be developed in the subsequent discussion chapter.

4.1. Model Performance on the Training Set

This section concisely and rigorously reports the results obtained by the three selected algorithms: Logistic Regression, Random Forest, and Gradient Boosting, during validation on the training set, performed by means of cross-validation. Performance is presented solely through the metrics defined in the model evaluation chapter (F1-score, Matthews Correlation Coefficient, Log-loss and the Confusion Matrix).

4.1.1. Training-Set performance for Injury_Categorization_Next20Days

Table 7, reported below summarizes the average evaluation metrics obtained during validation on the training set for the three algorithms considered. Gradient Boosting records the highest scores for both F1-score (0.898) and Matthews Correlation Coefficient (0.635), indicating superior ability to balance precision and recall while maintaining a strong overall correlation between predictions and observations. Random Forest follows with an F1 of 0.859 and an MCC of 0.502 yet achieves the lowest Log-loss (0.309), reflecting more accurate probabilistic calibration than the other models. Logistic Regression exhibits markedly lower performance (F1 = 0.383; MCC = 0.086) and a Log-loss of 1.528,

underscoring the limitations of a linear model when confronted with non-linear relationships and imbalanced classes.

Table 7: Training-set metrics for *Injury_Categorization_Next20Days*

Model	F1	MCC	LogLoss
Gradient Boosting Categorization	0.898	0.635	0.673
Random Forest Categorization	0.859	0.502	0.309
Logistic Regression Categorization	0.383	0.086	1.528

Building on these aggregate results, *Figure 3* shown below displays, for each algorithm, the class-normalized confusion matrix; absolute counts are provided at the bottom of each matrix.

- Gradient Boosting correctly classifies 40.4 % of ALI, 39.1 % of BCI, 54.9 % of MTI, 98.8 % of None and 45.5 % of Other. Misclassifications concentrate mainly on the None category: approximately 59 % of ALI and BCI instances are reassigned to this class.
- Random Forest correctly identifies 28.1 % of ALI, 22.5 % of BCI, 33.2 % of MTI, 99.6 % of None and 26.2 % of Other. Errors again gravitate towards None, which absorbs between 66 % and 78 % of injury-related observations.
- Logistic Regression yields lower class-level accuracies across the board: 37.9 % (ALI), 46.4 % (BCI), 26.6 % (MTI), 28.0 % (None) and 41.5 % (Other), and exhibits a more dispersed error pattern, without a single predominant sink class.

	Gradient Boosting					Random Forest					Logistic Regression					Σ
	ALI	BCI	MTI	None	Other	ALI	BCI	MTI	None	Other	ALI	BCI	MTI	None	Other	
Actual ALI	40.4 %	0.5 %	0.5 %	58.6 %	0.0 %	28.1 %	0.0 %	0.0 %	71.9 %	0.0 %	37.9 %	16.7 %	16.3 %	16.3 %	12.8 %	203
Actual BCI	0.7 %	39.1 %	1.4 %	58.7 %	0.0 %	0.0 %	22.5 %	0.0 %	77.5 %	0.0 %	13.0 %	46.4 %	8.0 %	15.2 %	17.4 %	138
Actual MTI	0.0 %	0.0 %	54.9 %	45.0 %	0.1 %	0.0 %	0.1 %	33.2 %	66.7 %	0.0 %	21.2 %	17.2 %	26.6 %	16.7 %	18.3 %	831
Actual None	0.1 %	0.1 %	0.7 %	98.8 %	0.2 %	0.0 %	0.0 %	0.3 %	99.6 %	0.1 %	15.3 %	20.2 %	14.3 %	28.0 %	22.1 %	7902
Actual Other	0.3 %	0.3 %	1.2 %	52.6 %	45.5 %	0.0 %	0.0 %	0.3 %	73.5 %	26.2 %	13.2 %	15.7 %	12.6 %	16.9 %	41.5 %	325
Σ	93	65	522	8554	165	60	33	302	8915	89	1526	1891	1436	2464	2082	9399

Figure 3 - Confusion matrices for *Injury_Categorization_Next20Days*

4.1.2. Training-set metrics for Injury_Next20Days

To assess the impact of class rebalancing in a comprehensive manner, this section is divided into two complementary subsections: the first examines model performance on the non-rebalanced dataset, which preserves the original injury distribution, whereas the second presents the same metrics computed on the dataset rebalanced with SMOTENC, thereby enabling a direct comparison between scenarios with and without minority-class oversampling.

4.1.2.1. Training-Set Performance on Original Imbalanced Data (Injury_Next20Days)

Table 8 reports the average performance metrics obtained on the training dataset, without class rebalancing, for the prediction of *Injury_Next20Days*. Gradient Boosting emerges as the best-performing model in terms of both F1-score (0.902) and Matthews Correlation Coefficient (0.628), confirming its ability to capture a balanced trade-off between sensitivity and precision while maintaining overall predictive coherence. Random Forest follows closely with an F1-score of 0.870 and an MCC of 0.527; despite a slightly lower discriminative power, it achieves the lowest Log-loss (0.252), indicating a more calibrated probabilistic output. Logistic Regression, performs considerably worse (F1 = 0.656; MCC = 0.129; Log-loss = 0.664), reaffirming its structural limitations in modelling complex and non-linear relationships, especially in the presence of an imbalanced outcome variable.

Table 8 - Training-set metrics for Injury_Next20Days

Model	F1	MCC	LogLoss
Gradient Boosting 20Days	0.902	0.628	0.394
Random Forest 20Days	0.870	0.527	0.252
Logistic Regression 20Days	0.656	0.129	0.664

Figure 4 provides further insight by showing, for each model, the normalized confusion matrices. These illustrate how predictions are distributed across the actual classes, highlighting the internal classification dynamics underlying the metrics presented above.

- Gradient Boosting achieves a recall of 53.2% for the positive, while maintaining a specificity of 98.2%. This configuration results in a well-balanced classifier, capable of detecting more than half of all injuries without compromising performance on the majority class.
- Random Forest yields slightly lower performance for the minority class, with 35.3% of true positives correctly identified and 99.4% of true negatives retained. The increased specificity comes at the expense of missed.
- Logistic Regression correctly classifies 56.3% of positive instances but at the cost of substantial misclassification of negatives (only 61.0% accuracy on class 0). The relatively high false positive rate contributes to the model's poor Log-loss and MCC.

		Gradient Boosting		Random Forest		Logistic Regression		Σ
		0.0	1.0	0.0	1.0	0.0	1.0	
Actual	0.0	98.2 %	1.8 %	99.4 %	0.6 %	61.0 %	39.0 %	7902
	1.0	46.8 %	53.2 %	64.7 %	35.3 %	43.7 %	56.3 %	1497
Σ		8460	939	8820	579	5477	3922	9399

Figure 4 - Confusion matrices for *Injury_Next20Days*

4.1.2.2. Training-Set performance on rebalanced data (*Injury_Next20Days*)

Table 9 presents the average performance metrics obtained on the training set for the *Injury_Next20Days* target after applying class rebalancing via SMOTENC. Among the models evaluated, Random Forest achieves the best overall performance with an F1-score of 0.889 and the highest Matthews Correlation Coefficient (MCC) at 0.714, alongside the lowest Log-loss (0.282). Gradient Boosting closely follows with an F1-score of 0.885 and an MCC of 0.691, while showing slightly less accurate calibration (Log-loss = 0.586).

Logistic Regression remains clearly inferior, with an F1-score of 0.626, an MCC of 0.162, and a Log-loss of 0.664, confirming the difficulties faced from classifiers in capturing complex relationships, even after oversampling.

Table 9: Training-set metrics for Injury _Next20Days on rebalanced data

Model	F1	MCC	LogLoss
Random Forest 20Days (SMOTENC)	0.889	0.714	0.282
Gradient Boosting 20Days (SMOTENC)	0.885	0.691	0.586
Logistic Regression 20Days (SMOTENC)	0.626	0.162	0.664

Building on these summary metrics, *Figure 5* shows the class-normalized confusion matrices for each algorithm, allowing for a more inspection of prediction behaviour. Absolute counts are provided below each matrix.

- Gradient Boosting correctly identifies 69.6% of injured players and 95.3% of non-injured players. The false negative rate remains non-negligible (30.4%), but is substantially reduced compared to the non-SMOTENC scenario.
- Random Forest achieves a true positive rate of 62.5% and a true negative rate of 98.8%, leading to the best Log-loss among the three algorithms. Its F1-score also slightly exceeds that of Gradient Boosting, confirming its overall superior balance between classification performance and probabilistic calibration.
- Logistic Regression continues to show weaker results, with a true positive rate of 57.8% and a true negative rate of 60.8%. The model struggles to learn the minority class despite oversampling.

		Gradient Boosting		Random Forest		Logistic Regression		Σ
		0.0	1.0	0.0	1.0	0.0	1.0	
Actual	0.0	95.3 %	4.7 %	98.8 %	1.2 %	60.8 %	39.2 %	7902
	1.0	30.4 %	69.6 %	37.5 %	62.5 %	42.2 %	57.8 %	2634
Σ		8333	2203	8794	1742	5917	4619	10536

Figure 5 - Confusion matrices for *Injury_Next20Days* on rebalanced data

Comparing these outcomes with those obtained without class rebalancing reveals several important trends. Both Gradient Boosting and Random Forest show improvements in true positive rates and MCC when SMOTENC is applied. Specifically, Random Forest benefits most in terms of F1-score (+0.019) and MCC (+0.187), while Gradient Boosting shows a marginal decrease in F1 but a notable improvement in MCC (+0.063). On the other hand, Logistic Regression experiences an increase in F1 but maintains a relatively low MCC, indicating that oversampling alone is insufficient to compensate for its lack of non-linear modelling capacity

4.1.3. Training-set metrics for *Injury_Next3Days*

Consistent with the structure adopted for the *Injury_Next20Days* variable, the analysis of *Injury_Next3Days* is likewise divided into two distinct subsections. The first presents the results obtained on the original data, preserving the real-world distribution of injuries; the second reports the model performance following the application of the SMOTENC rebalancing technique.

4.1.3.1. Training-Set performance on original imbalanced data (*Injury_Next3Days*)

Table 10 reports the average performance metrics obtained on the training set for the *Injury_Next3Days* target in the absence of class rebalancing. As expected, the prediction task proves considerably more challenging given the huge imbalance of the target variable, positive cases account for less than 4% of the total observations. The Gradient Boosting algorithm achieves a F1-score (0.948) and the lowest Log-loss (0.347), although its Matthews Correlation Coefficient (0.054) remains low. Random Forest yields a slightly

lower F1-score (0.896) and higher Log-loss (0.525), but records the highest MCC (0.077), indicating a marginally better balance between true and false classifications under these extreme conditions. Logistic Regression, by contrast, performs poorly across all metrics, with an F1-score of 0.731, an MCC of 0.051, and the highest Log-loss (0.661), underscoring the model's structural limitations

Table 10 - Training-set metrics for Injury_Next3Days

Model	F1	MCC	LogLoss
Gradient Boosting 3Days	0.948	0.054	0.347
Random Forest 3Days	0.896	0.077	0.525
Logistic Regression 3Days	0.731	0.051	0.661

These results are further clarified by the class-normalised confusion matrices reported in *Figure 6*. Absolute values are indicated at the base of each matrix.

- Gradient Boosting classifies 99.7% of non-injury instances correctly, but fails to detect the majority of positive cases, with a true positive rate of only 2.1%.
- Random Forest shows a better balance, identifying 24.2% of injury instances while maintaining a high specificity (88.2%). Its improved MCC reflects a more equitable distribution of classification errors, though the true positive rate remains low in absolute terms.
- Logistic Regression, as anticipated, exhibits the weakest performance. It correctly identifies 51.8% of injuries but misclassifies 38.3% of non-injury cases, leading to significant loss in both discrimination and calibration, as evidenced by the high Log-loss.

		Gradient Boosting		Random Forest		Logistic Regression		Σ
		0.0	1.0	0.0	1.0	0.0	1.0	
Actual	0.0	99.7 %	0.3 %	88.2 %	11.8 %	61.7 %	38.3 %	9069
	1.0	97.9 %	2.1 %	75.8 %	24.2 %	48.2 %	51.8 %	330
Σ		9363	36	8252	1147	5753	3646	9399

Figure 6 - Confusion matrices for *Injury_Next3Days*

4.1.3.2 Training-Set performance on rebalanced data (*Injury_Next3Days*)

Table 4.6 presents the average performance metrics obtained on the training set for the *Injury_Next3Days* target after applying class rebalancing via SMOTENC. Among the models evaluated, Gradient Boosting achieves the highest F1-score (0.939) and the best Matthews Correlation Coefficient (MCC = 0.659), reflecting a strong trade-off between precision and recall and a consistent alignment between predicted and actual outcomes. Random Forest follows closely with an F1-score of 0.932 and an MCC of 0.628, and outperforms the other models in terms of probabilistic calibration, as indicated by the lowest Log-loss value (0.165). Logistic Regression remains less effective, with an F1-score of 0.693, an MCC of 0.121, and a Log-loss of 0.660, confirming its limited capacity to model complex, non-linear interactions even after oversampling.

Table 11 - Training-set metrics for *Injury_Next20Days* on rebalanced data

Model	F1	MCC	LogLoss
Gradient Boosting 3Days SMOTENC	0.939	0.659	0.337
Random Forest 3Days SMOTENC	0.932	0.628	0.165
Logistic Regression 3Days SMOTENC	0.693	0.121	0.660

Building on these summary metrics, Figure 7 shows the class-normalized confusion matrices for each algorithm, providing a detailed overview of classification performance across the binary outcome. Absolute instance counts are reported below each matrix.

- Gradient Boosting correctly identifies 51.8% of injured players and 99.4% of non-injured players. The model displays a relatively balanced distribution of errors and high classification accuracy for the majority class.
- Random Forest achieves the highest true negative rate (99.9%) among all algorithms and identifies 43.1% of the injured class correctly. While its recall is lower than that of Gradient Boosting.
- Logistic Regression classifies 57.6% of injured players correctly and achieves a true negative rate of 62.1%. Despite improvements from the original dataset, the model still struggles to reach competitive levels of predictive reliability.

		Gradient Boosting		Random Forest		Logistic Regression		Σ
		0	1	0	1	0	1	
Actual	0	99.4 %	0.6 %	99.9 %	0.1 %	62.1 %	37.9 %	9069
	1	48.2 %	51.8 %	56.9 %	43.1 %	42.4 %	57.6 %	1007
Σ		9496	580	9632	444	6057	4019	10076

Figure 7 - Confusion matrices for Injury_Next20Days on rebalanced data

A comparison with the performance obtained without class rebalancing reveals clear improvements, particularly for tree-based models. Gradient Boosting maintains similar F1 performance while improving in MCC (+0.605), confirming a more stable predictive structure. Random Forest also shows marked improvements across all metrics: F1 increases by +0.036, MCC by +0.551, and Log-loss drops from 0.525 to 0.165. Conversely, Logistic Regression sees a decline in F1-score (−0.038), a marginal increase in MCC (+0.070), and a nearly unchanged Log-loss. These results confirm that oversampling via SMOTENC enhances model robustness and reliability, particularly when combined with ensemble methods capable of capturing non-linear patterns.

4.2. Model Performance on the Test Set

This section rigorously reports the performance of the three selected algorithms when evaluated on the independent test set. As the previous chapter, also in this one the results are presented exclusively through the metrics established in the *Model Evaluation* chapter, without any interpretative commentary, which is deferred to the subsequent *Discussion* chapter.

4.2.1. Test-Set performance for Injury_Categorization_Next20Days

Table 12 reports the evaluation metrics obtained on the independent test set for the multiclass target *Injury_Categorization_Next20Days*. The three algorithms produce closely aligned F1-scores, all below 0.19. Random Forest obtains the highest F1-score (0.186) and a Matthews Correlation Coefficient equal to zero, together with the lowest Log-loss (0.865). Gradient Boosting reaches an F1-score of 0.186 and an MCC of -0.006 , while displaying the highest Log-loss (2.403). Logistic Regression records an F1-score of 0.179, an MCC of -0.009 , and an intermediate Log-loss (1.382).

Table 12 - Test-set metrics for Injury_Categorization_Next20Days

Model	F1	MCC	Log_loss
Gradient Boosting	0.1856	-0.0059	2.4027
Random Forest	0.1857	0	0.8649
Logistic Regression	0.1785	-0.0089	1.3818

Figure 8 - Confusion matrices for Injury_Categorization_Next20Days Figure 8 depicts the class-normalized confusion matrices for the three models; absolute counts are indicated below each matrix.

- Gradient Boosting assigns 100 % of all instances to the class *None*.
- Random Forest exhibits the same behaviour, predicting exclusively the class *None*.
- Logistic Regression exhibits limited performance in identifying most injury categories, with only the 'None' class being correctly classified in more than 50% of cases (61.3%). All other classes show low correct classification rates, ranging from 2.6% to 22.0%.

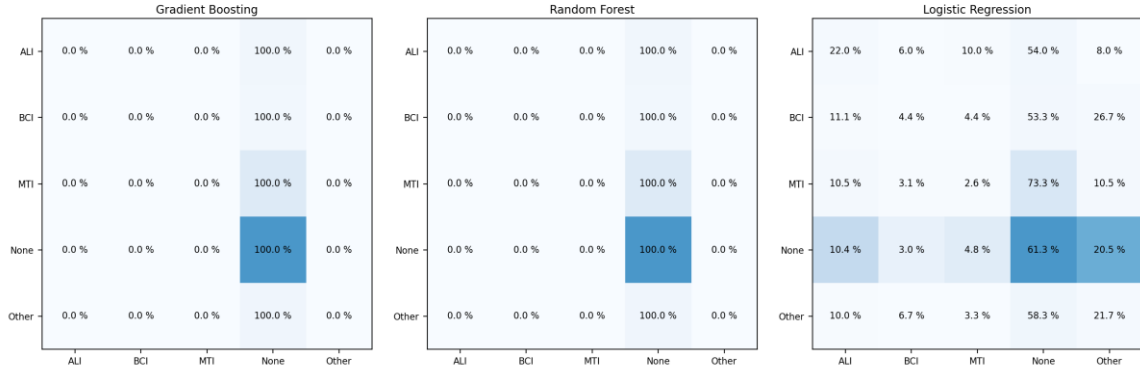


Figure 8 - Confusion matrices for Injury_Categorization_Next20Days

4.2.2. Test-Set performance of models trained on original and rebalanced data (Injury_Next20Days)

The present subsection reports the results achieved by the algorithms on both the original dataset and the rebalanced one. Performance is summarised in *Table 13* through F1-score, Matthews Correlation Coefficient (MCC), and Log-loss; the associated confusion matrices are shown in *Figure 9*.

Table 13 summarizes the test-set performance in both of dataset. In the rebalanced configuration, Random Forest achieves the highest F1-score (0.240) and the greatest proportion of correctly classified positive and negative cases, although with the highest Log-loss among the tree-based approaches (1.288). Gradient Boosting, evaluated on the SMOTENC dataset, records an F1-score of 0.236 and an MCC of -0.017 , accompanied by the largest Log-loss overall (3.742), indicating limited calibration accuracy. Logistic Regression exhibits marginal improvements after oversampling, F1 rises to 0.179 and MCC to 0.004, while its Log-loss (0.633) remains essentially unchanged relative to the Plain variant (0.628). Under the original class distribution, both Gradient Boosting and Random Forest assign every observation to the non-injury class, yielding F1-scores of zero; Random Forest nonetheless attains the lowest overall Log-loss (0.447).

Table 13 - Test-Set metrics for Injury_Next3Days: Models trained on original and rebalanced data

Algorithm	Variant	F1	MCC	Log_loss
Random Forest 20Days	Plain	0	0	0.4471
Logistic Regression 20Days	Plain	0.1695	-0.0023	0.6283
Gradient Boosting 20Days	Plain	0	0	1.1042
Random Forest 20Days	SMOTENC	0.2401	-0.0156	1.2878
Logistic Regression 20Days	SMOTENC	0.1792	0.0044	0.6329
Gradient Boosting 20Days	SMOTENC	0.2359	-0.017	3.7416

These results are further clarified by the Confusion matrices reported in *Figure 9*. The top row refers to the model performance on the original dataset, while the bottom row corresponds to results obtained after applying a rebalancing strategy. Percentages are normalized by the actual class.

- Gradient Boosting (original) classifies 100.0 % of non-injury instances correctly but fails to detect any injury cases, with a true positive rate of 0.0 %. Trained after the rebalancing, the same model shows an inverted trend, assigning 83.9 % of class 0 and 82.1 % of class 1 instances to the positive class.
- Random Forest (original) behaves identically to Gradient Boosting, predicting all instances as non-injury and resulting in a null sensitivity. When trained on the rebalanced dataset, it reassigns the majority of samples to the positive class, with 91.1 % of non-injury cases and 89.8 % of injury cases classified as such.
- Logistic Regression (original) detects 21.8 % of injury cases while maintaining 78.0 % specificity. With rebalanced data, it displays a comparable pattern, correctly classifying 24.0 % of injuries and 76.6 % of non-injury instances.



Figure 9 - Confusion matrices on Test Set: Models trained on original vs rebalanced data

4.2.3. Test-Set performance of models trained on original and rebalanced data (*Injury_Next3Days*)

The present subsection reports the results achieved by the algorithms on both the original dataset and the rebalanced one, considering a prediction horizon of 3 days.

Under the original class distribution, both Logistic Regression and Random Forest assign all instances to the non-injury class, yielding F1-scores and MCC values equal to zero. Gradient Boosting slightly deviates from this trend, reaching a minimal true positive rate of 8.5 %, with a corresponding F1-score of 0.0409 and MCC of -0.0091 . Among the three, Logistic Regression records the lowest Log-loss (0.1425), while Gradient Boosting exhibits the highest (0.6937), reflecting poorer calibration.

After applying the SMOTENC rebalancing strategy, all models improve in terms of positive class detection. Random Forest achieves the best F1-score (0.0589) and the highest true positive rate (67.1 %) but also registers a high Log-loss (1.0631). Gradient Boosting performs comparably in terms of F1 (0.0567), although with the highest Log-loss overall (1.5115). Logistic Regression shows limited change, with an F1-score of 0.0413 and a relatively low Log-loss (0.5996).

Table 14 - Test-Set metrics for Injury_Next3Days: Models trained on original and rebalanced data

Algorithm	Variant	F1	MCC	Log_loss
Logistic Regression 3Days	Plain	0	0	0.1425
Random Forest 3Days	Plain	0	0	0.2073
Gradient Boosting 3Days	Plain	0.0409	-0.0091	0.6937
Logistic Regression 3Days	SMOTENC	0.0413	-0.0213	0.5996
Gradient Boosting 3Days	SMOTENC	0.0567	-0.012	1.5115
Random Forest 3Days	SMOTENC	0.0589	-0.0077	1.0631

These results are clarified by the confusion matrices reported in *Figure 10*. As the previous paragraph the top row refers to the model performance on the original dataset, while the bottom row corresponds to the rebalanced configuration.

- Gradient Boosting (original) classifies 89.9 % of non-injury instances correctly and detects 8.5 % of injury cases. After rebalancing, it correctly identifies 54.9 % of positive cases, but misclassifies 58.2 % of non-injury instances.
- Random Forest (original) mirrors the behaviour of Logistic Regression, with 100.0 % of instances predicted as non-injury. Post-rebalancing, it reaches a true positive rate of 67.1 %, while classifying 69.1 % of non-injury instances correctly.
- Logistic Regression (original) assigns all instances to the non-injury class, yielding 100.0 % specificity and 0.0 % sensitivity. With rebalanced data, sensitivity increases to 14.6 %, while specificity drops to 80.6 %.

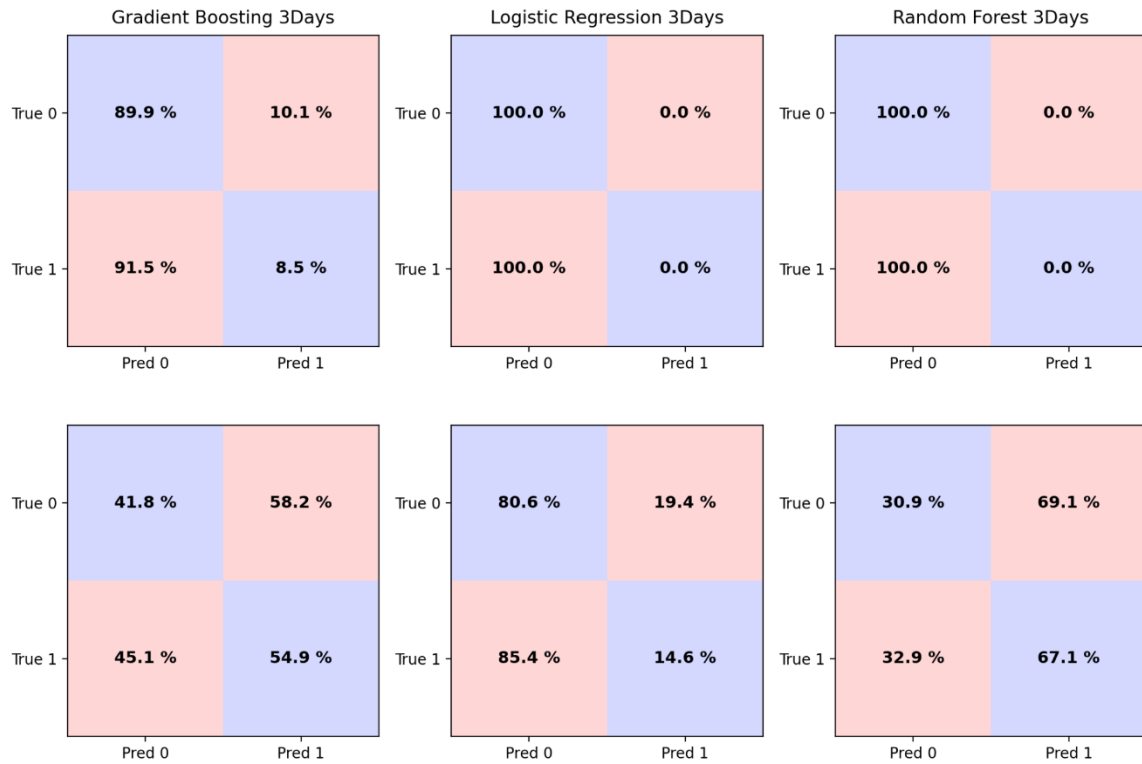


Figure 10 - Confusion matrices on Test Set: Models trained on original vs rebalanced data

4.3. Features Importance

This section presents the results of the feature importance analysis, conducted to quantify the contribution of each predictor to the injury risk classification task. Feature importance scores are reported as computed with respect to the target variables, based on the initial training dataset. The analysis aims to assess the statistical association between each predictor and the outcome labels, and results are presented separately for each of the three classification tasks: injury categorization into five classes, injury occurrence within the next 20 days, and injury occurrence within the next 3 days.

To estimate the relevance of each predictive variable in the classification process, three filter methods, recognized in the literature, were employed: ReliefF, Information Gain and Chi-Square. Each method assigns a score that quantifies the extent to which a given predictor contributes to the differentiation between the classes of the target variable.

Information Gain measures the reduction in entropy: that is, the uncertainty associated with the distribution of the target variable, resulting from splitting the dataset based on the values of a specific predictor. This metric quantifies how much information a feature

provides about the output to be predicted: a high Information Gain indicates that the variable effectively distinguishes between output classes, making it particularly valuable for predictive models.

The Chi-Square test, used to assess the statistical association between an independent variable and a categorical target, is based on comparing the observed frequencies with those expected under the assumption of independence. Although this method is formally applicable only to categorical variables, the *Orange3* software automatically applies a discretization procedure when the test is used on continuous variables. The resulting score reflects the extent to which the probability distribution of the target classes differs across the intervals created. A high Chi-Square value thus indicates a statistically significant dependency between the predictor and the target variable.





























































Unlike methods based on global statistics, ReliefF evaluates each observation by considering its specific characteristics and identifying the most similar instances in the dataset. For each instance, the algorithm compares the values of each predictor with those of its nearest neighbours belonging to the same and different target classes. A predictor is considered important if, within similar contexts, its variation is consistently associated with changes in the target class. This local analysis enables the detection of patterns that vary across subgroups and allows the algorithm to capture complex relationships, including interactions (e.g., between workload and player role) and non-linear effects that would not be detected by globally aggregated methods.

4.3.1. Feature importance with respect to Injury_Categorization_Next20Days

The importance of each predictor was computed with respect to the five-class injury categorization target (MTI, ALI, BCI, Other, None). To facilitate a clear and independent evaluation of each relevance metric, the results are reported in three separate columns, one for each method. This structure enables a direct comparison of the relative contribution of each feature to the classification task, according to different statistical criteria.

Table 15 shows the top 20 features for each metric, allowing a direct comparison of the variables identified as most informative by each statistical criterion.

Table 15 - features ranked by Relief, Info. Gain and Chi-square with respect to Injury_Categorization_Next20Days

	#	ReliefF		#	Info. gain		#	χ^2
 BCL_Before		0.073	1	 Other_Before	0.012	 ALI_Before		97.026
 MTI_Before		0.071	2	 MTI_Before	0.011	 MTI_Before		89.799
 Other_Before		0.070	3	 ALI_Before	0.010	 Injury_Count_Before		83.314
 ALI_Before		0.066	4	 Injury_Count_Before	0.009	 Forwards	2	74.041
 Age		0.061	5	 BMI	0.009	 BCL_Before		73.351
 BMI		0.060	6	 TotalShotPerMin_Avg	0.009	 TotalShotPerMin_Avg		59.834
 Injury_Count_Before		0.054	7	 ActionInPenaltyAreaPerMin_Avg	0.007	 TakeOnsAttempPerMin_Avg15		54.506
 ProgressiveCarryDistancePerMin_Avg		0.050	8	 OffensiveImpactActionsPerMin_Avg	0.007	 TotalShotPerMin_Avg15		48.312
 MinutesLast30Days		0.048	9	 BCL_Before	0.007	 OffensiveImpactActionsPerMin_Avg		45.034
 MinutesLast15Days		0.047	10	 TakeOnsAttempPerMin_Avg	0.006	 TakeOnsAttempPerMin_Avg		42.719
 SymPctMinutesLast15vsPrev15		0.045	11	 PassesAttemptedPerMin_Avg	0.006	 MinutesLast30Days		41.052
 ACWR		0.043	12	 AerialDuelsTotalPerMin_Avg	0.006	 OffensiveImpactActionsPerMin_Avg15		40.408
 MinutesLast7Days		0.042	13	 Forwards	0.006	 Injury_Past30Days		37.799
 TotalCarryDistancePerMin_Avg		0.041	14	 TouchesPerMin_Avg	0.006	 Other_Before		33.362
 ProgressiveCarryDistancePerMin_Avg15		0.035	15	 LongPassesStatsPerMin_Avg	0.006	 PassesAttemptedPerMin_Avg		30.684
 Injury_Past30Days		0.033	16	 TakeOnsAttempPerMin_Avg15	0.005	 TouchesPerMin_Avg		26.851
 TotalCarryDistancePerMin_Avg15		0.031	17	 CarriesPerMin_Avg	0.005	 ActionInPenaltyAreaPerMin_Avg		25.477
 DribblerChallengedPerMin_Avg		0.025	18	 TotalShotPerMin_Avg15	0.005	 DribblerChallengedPerMin_Avg		23.955
 LongPassesStatsPerMin_Avg		0.025	19	 OffensiveImpactActionsPerMin_Avg15	0.005	 ProgressiveCarryDistancePerMin_Avg15		23.067
 PassesAttemptedPerMin_Avg		0.022	20	 MinutesLast30Days	0.005	 PassesAttemptedPerMin_Avg15		22.112

The magnitude of the importance scores varied across the three evaluation criteria. For Information Gain, the highest values ranged between 0.012 and 0.009, with a slow decline beyond the top five predictors. Chi-square scores were generally higher in absolute terms, with the top three variables exceeding 80 and a maximum value of 97.026. ReliefF scores showed a more gradual distribution, with top values between 0.073 and 0.060, and several additional features scoring in the 0.045–0.050 range. Overall, the strongest signals were concentrated among a limited number of features, particularly those related to previous injury history.

When considering each metric individually:

- Information Gain: the top-ranked features were Other_Before (0.012), MTI_Before (0.011), and ALI_Before (0.010). All these predictors are related to previous injury history. The highest-scoring non-injury feature was BMI (0.009), followed by TotalShotPerMin_Avg (0.009).
- Chi-square: the most relevant features were ALI_Before ($\chi^2 = 97.026$), MTI_Before (89.799), and Injury_Count_Before (83.314), all significantly associated with the injury categorization target. Among non-injury variables, excluding the binary variable ‘Forwards’, the highest ReliefF scores were obtained by TotalShotPerMin_Avg (59.834) and TakeOnsAttempPerMin_Avg15 (54.506).





























































- ReliefF: the highest relevance scores were assigned to BCI_Before (0.073), MTI_Before (0.071), and Other_Before (0.070). BMI and ALI_Before also obtained relatively high values (0.060 and 0.066, respectively). Among match-related statistics, ProgressiveCarryDistancePerMin_Avg and MinutesLast30Days scored above 0.045.

4.3.2 Feature importance with respect to Injury_Next20Days on the original and rebalanced training dataset

This subsection reports the feature importance values obtained through two separate analysis: the first conducted on the original training dataset, and the second on a rebalanced version obtained by applying the Synthetic Minority Over-sampling Technique. The results are presented separately for each case.

Table 16 shows the top 20 features ranked by Information Gain, Chi-square, and ReliefF with respect to the Injury_Next20days features, as computed during the training phase on the original dataset.

Table 16 - features ranked by Relief, Info. Gain and Chi-square with respect to Injury_Next20Days on original dataset

		#	ReliefF		#	Info. gain		#	χ^2
1	 Injury_Past30Days		<div><div></div>0.058</div>	1	 Other_Before	<div><div></div>0.007</div>	1	 Injury_Count_Before	<div><div></div>56.866</div>
2	 BCI_Before		<div><div></div>0.039</div>	2	 Injury_Count_Before	<div><div></div>0.006</div>	2	 ALI_Before	<div><div></div>55.549</div>
3	 Other_Before		<div><div></div>0.037</div>	3	 BMI	<div><div></div>0.006</div>	3	 MTI_Before	<div><div></div>55.140</div>
4	 BMI		<div><div></div>0.036</div>	4	 MTI_Before	<div><div></div>0.006</div>	4	 TotalShotPerMin_Avg	<div><div></div>44.989</div>
5	 SymPctMinutesLast15vsPrev15		<div><div></div>0.035</div>	5	 ALI_Before	<div><div></div>0.006</div>	5	 Forwards	<div><div></div>44.232</div>
6	 ALI_Before		<div><div></div>0.034</div>	6	 OffensiveImpactActionsPerMin_Avg	<div><div></div>0.004</div>	6	 TakeOnsAttemptPerMin_Avg15	<div><div></div>42.097</div>
7	 Age		<div><div></div>0.032</div>	7	 TotalShotPerMin_Avg	<div><div></div>0.004</div>	7	 TotalShotPerMin_Avg15	<div><div></div>36.887</div>
8	 MinutesLast30Days		<div><div></div>0.031</div>	8	 TakeOnsAttemptPerMin_Avg15	<div><div></div>0.004</div>	8	 OffensiveImpactActionsPerMin_Avg	<div><div></div>34.377</div>
9	 MinutesLast15Days		<div><div></div>0.026</div>	9	 Forwards	<div><div></div>0.004</div>	9	 OffensiveImpactActionsPerMin_Avg15	<div><div></div>32.683</div>
10	 Injury_Count_Before		<div><div></div>0.024</div>	10	 TotalShotPerMin_Avg15	<div><div></div>0.004</div>	10	 PassesAttemptedPerMin_Avg	<div><div></div>28.825</div>
11	 MTI_Before		<div><div></div>0.023</div>	11	 OffensiveImpactActionsPerMin_Avg15	<div><div></div>0.003</div>	11	 BCI_Before	<div><div></div>24.710</div>
12	 TotalCarryDistancePerMin_Avg15		<div><div></div>0.022</div>	12	 PassesAttemptedPerMin_Avg	<div><div></div>0.003</div>	12	 TakeOnsAttemptPerMin_Avg	<div><div></div>23.825</div>
13	 TotalCarryDistancePerMin_Avg		<div><div></div>0.021</div>	13	 BCI_Before	<div><div></div>0.003</div>	13	 MinutesLast30Days	<div><div></div>23.750</div>
14	 ProgressiveCarryDistancePerMin_Avg15		<div><div></div>0.021</div>	14	 LongPassesStatsPerMin_Avg	<div><div></div>0.003</div>	14	 TouchesPerMin_Avg	<div><div></div>21.066</div>
15	 ProgressiveCarryDistancePerMin_Avg		<div><div></div>0.021</div>	15	 MinutesLast30Days	<div><div></div>0.003</div>	15	 DefensiveBlocksPerMin_Avg	<div><div></div>19.679</div>
16	 DribblerChallengedPerMin_Avg		<div><div></div>0.016</div>	16	 TouchesPerMin_Avg	<div><div></div>0.002</div>	16	 PassesAttemptedPerMin_Avg15	<div><div></div>18.776</div>
17	 ACWR		<div><div></div>0.011</div>	17	 TakeOnsAttemptPerMin_Avg	<div><div></div>0.002</div>	17	 DefensiveBlocksPerMin_Avg15	<div><div></div>18.371</div>
18	 LongPassesStatsPerMin_Avg		<div><div></div>0.011</div>	18	 DefensiveBlocksPerMin_Avg	<div><div></div>0.002</div>	18	 Other_Before	<div><div></div>17.543</div>
19	 FreeKickStatsPerMin_Avg		<div><div></div>0.011</div>	19	 DefensiveImpactActionsPerMin_Avg	<div><div></div>0.002</div>	19	 Injury_Past30Days	<div><div></div>14.828</div>
20	 CarriesPerMin_Avg		<div><div></div>0.011</div>	20	 ActionInPenaltyAreaPerMin_Avg	<div><div></div>0.002</div>	20	 CarriesPerMin_Avg	<div><div></div>14.468</div>

The magnitude of the importance scores varied across the three evaluation criteria. Information Gain values were generally low in absolute terms, with the top-ranked features scoring between 0.007 and 0.006, followed by a gradual decline below. Chi-square scores

reached higher values, with the three most significant predictors exceeding 55 and the highest reaching 56.866. ReliefF values showed a broader distribution, with a maximum of 0.058 and several features scoring between 0.030 and 0.039. Across all metrics, a limited number of variables contributed the most to the classification task, many of which were related to past injuries or short-term physical exposure.

When considering each metric individually:

- Information Gain: the top-ranked features were Other_Before (0.007), Injury_Count_Before (0.006), and BMI (0.006), followed by MTI_Before and ALI_Before (both 0.006). Among non-injury-related features, OffensiveImpactActionsPerMin_Avg and TotalShotPerMin_Avg reached 0.006 and 0.004, respectively.
- Chi-square: the most relevant features were Injury_Count_Before ($\chi^2 = 56.866$), ALI_Before (55.549), and MTI_Before (55.140). Other significant variables included TotalShotPerMin_Avg (44.989) and Forwards (44.232), followed by several performance metrics such as TakeOnsAttempPerMin_Avg15 and OffensiveImpactActionsPerMin_Avg.
- ReliefF: the top scores were assigned to Injury_Past30Days (0.058), BCI_Before (0.039), and Other_Before (0.037). BMI, SymPctMinutesLast15vsPrev15, and ALI_Before followed closely, each scoring above 0.034. Exposure and workload indicators such as MinutesLast30Days, MinutesLast15Days, and Injury_Count_Before were also present in the top 10.

The analysis now focuses on the the binary classification target *Injury_Next20Days*, evaluated during the training phase on the SMOTE-rebalanced dataset. *Table 17* shows the top 20 features ranked by Information Gain, Chi-square, and ReliefF with respect to this target.

Table 17 - features ranked by Relief, Info. Gain and Chi-square with respect to Injury_Next20Days on rebalanced dataset

	#	ReliefF	#	Info. gain	#	χ^2
1 BCI_Before		0.045	1 TakeOnsAttempPerMin_Avg15	0.014	1 TakeOnsAttempPerMin_Avg15	140.744
2 Age		0.041	2 Other_Before	0.013	2 Injury_Count_Before	109.974
3 ALI_Before		0.036	3 ALI_Before	0.011	3 OffensiveImpactActionsPerMin_Avg15	109.500
4 MTI_Before		0.033	4 OffensiveImpactActionsPerMin_Avg15	0.009	4 MTI_Before	103.477
5 BMI		0.029	5 Injury_Count_Before	0.009	5 ALI_Before	100.703
6 Injury_Count_Before		0.028	6 MTI_Before	0.009	6 TotalShotPerMin_Avg15	90.691
7 MinutesLast30Days		0.027	7 DribblerChallengedPerMin_Avg15	0.009	7 OffensiveImpactActionsPerMin_Avg	81.689
8 ProgressiveCarryDistancePerMin_Avg15		0.024	8 BMI	0.009	8 TotalShotPerMin_Avg	76.319
9 ProgressiveCarryDistancePerMin_Avg		0.023	9 TotalShotPerMin_Avg15	0.008	9 BCI_Before	73.826
10 TotalCarryDistancePerMin_Avg		0.022	10 DefensiveBlocksPerMin_Avg15	0.008	10 TakeOnsAttempPerMin_Avg	61.031
11 Injury_Past30Days		0.022	11 OffensiveImpactActionsPerMin_Avg	0.007	11 Forwards	57.444
12 Other_Before		0.020	12 TotalShotPerMin_Avg	0.007	12 PassesAttemptedPerMin_Avg	55.001
13 Center Backs	2	0.018	13 ActionInPenaltyAreaPerMin_Avg	0.006	13 MinutesLast30Days	43.534
14 AerialDuelsTotalPerMin_Avg		0.014	14 BCI_Before	0.006	14 TouchesPerMin_Avg	39.416
15 DribblerChallengedPerMin_Avg		0.013	15 TakeOnsAttempPerMin_Avg	0.006	15 Other_Before	39.087
16 MinutesLast15Days		0.013	16 FreeKickStatsPerMin_Avg15	0.005	16 MinutesLast15Days	30.484
17 TotalCarryDistancePerMin_Avg15		0.013	17 PassesAttemptedPerMin_Avg	0.005	17 DefensiveBlocksPerMin_Avg	24.405
18 LongPassesStatsPerMin_Avg		0.013	18 MinutesLast30Days	0.005	18 PassesAttemptedPerMin_Avg15	23.959
19 ACWR		0.012	19 Forwards	0.004	19 ActionInPenaltyAreaPerMin_Avg15	23.262
20 PassesAttemptedPerMin_Avg		0.010	20 ActionInPenaltyAreaPerMin_Avg15	0.004	20 Injury_Past30Days	22.568

Overall, the magnitude of the importance scores remained modest across all three metrics. Information Gain values peaked at 0.014 and declined gradually below 0.010. Chi-square scores were substantially higher, with several predictors exceeding 100 and a maximum value of 140.744. ReliefF values ranged from 0.045 to 0.020, with only a few variables scoring above 0.030.

When considering each metric individually:

- Information Gain: the most relevant predictors were TakeOnsAttempPerMin_Avg15 (0.014), Other_Before (0.013), and ALI_Before (0.011). Other notable features included OffensiveImpactActionsPerMin_Avg15, Injury_Count_Before, and MTI_Before, each scoring between 0.009 and 0.006.
- Chi-square: the top three variables were TakeOnsAttempPerMin_Avg15 ($\chi^2 = 140.744$), Injury_Count_Before (109.974), and OffensiveImpactActionsPerMin_Avg15 (109.500). Additional relevant features included MTI_Before, ALI_Before, and both TotalShotPerMin_Avg15 and OffensiveImpactActionsPerMin_Avg, with scores ranging from 90.691 to 81.689.
- ReliefF: the highest-ranked features were BCI_Before (0.045), Age (0.041), and ALI_Before (0.036), followed by MTI_Before (0.033), BMI (0.029), and

Injury_Count_Before (0.028). Other variables such as MinutesLast30Days and ProgressiveCarryDistancePerMin_Avg15 also appeared in the top positions.

A comparison between the two training configurations highlights both consistencies and shifts in feature relevance. In both cases, prior injury history variables (e.g., Other_Before, ALI_Before, MTI_Before, Injury_Count_Before) consistently ranked among the top features across all metrics, confirming their stable predictive contribution. However, the application of SMOTENC notably increased the relative importance of technical and match-derived variables such as TakeOnsAttempPerMin_Avg15, which became the top-ranked predictor in both Information Gain and Chi-square. Similarly, ReliefF scores revealed a greater dispersion in the rebalanced dataset, with a wider set of moderately relevant features.

4.3.3 Feature importance with respect to Injury_Next3Days on the original and rebalanced training dataset

As in the previous analysis, this subsection reports the feature importance values computed with respect to the *Injury_Next3Days* target. Two separate evaluations were performed: one on the original, and one on the rebalanced version.

Table 18 presents the highest-ranked predictors for the *Injury_Next3Days* classification task, as derived from the feature relevance analysis performed during the training phase on the original dataset.

Table 18 - features ranked by Relief, Info. Gain and Chi-square with respect to Injury_Next3Days on original dataset

	#	ReliefF	#	Info. gain	#	χ^2
1		Age 0.094	1	ALI_Before 0.002	1	ALI_Before 24.896
2		BCI_Before 0.092	2	Injury_Count_Before 0.002	2	Injury_Count_Before 17.576
3		ALI_Before 0.092	3	MTI_Before 0.002	3	TotalShotPerMin_Avg15 16.628
4		MTI_Before 0.080	4	TotalShotPerMin_Avg15 0.002	4	MTI_Before 16.565
5		BMI 0.079	5	TotalShotPerMin_Avg 0.001	5	Forwards 14.944
6		Other_Before 0.074	6	Forwards 0.001	2	TotalShotPerMin_Avg 14.496
7		Injury_Count_Before 0.065	7	Other_Before 0.001	7	BCI_Before 10.547
8		SymPctMinutesLast15vsPrev15 0.063	8	SymPctMinutesLast15vsPrev15 0.001	8	OffensiveImpactActionsPerMin_Avg 8.861
9		ProgressiveCarryDistancePerMin_Avg 0.060	9	MinutesLast30Days 0.001	9	PassesAttemptedPerMin_Avg 8.833
10		MinutesLast30Days 0.058	10	PassesAttemptedPerMin_Avg15 0.001	10	PassesAttemptedPerMin_Avg15 8.833
11		TotalCarryDistancePerMin_Avg 0.056	11	AerialDuelsTotalPerMin_Avg15 0.001	11	TouchesPerMin_Avg 8.306
12		MinutesLast7Days 0.050	12	TouchesPerMin_Avg 0.001	12	MinutesLast30Days 8.293
13		MinutesLast15Days 0.048	13	BMI 0.001	13	TouchesPerMin_Avg15 7.150
14		ACWR 0.043	14	PassesAttemptedPerMin_Avg 0.001	14	OffensiveImpactActionsPerMin_Avg15 7.077
15		Injury_Past30Days 0.042	15	BCI_Before 0.001	15	TakeOnsAttempPerMin_Avg15 6.281
16		TotalCarryDistancePerMin_Avg15 0.036	16	OffensiveImpactActionsPerMin_Avg 0.001	16	TakeOnsAttempPerMin_Avg 5.593
17		DribblerChallengedPerMin_Avg 0.035	17	OffensiveImpactActionsPerMin_Avg15 0.001	17	DefensiveBlocksPerMin_Avg 5.578
18		AerialDuelsTotalPerMin_Avg 0.034	18	TouchesPerMin_Avg15 0.001	18	CarriesPerMin_Avg 5.224
19		ActionInPenaltyAreaPerMin_Avg 0.032	19	ActionInPenaltyAreaPerMin_Avg 0.001	19	Fullbacks 4.479
20		PassesAttemptedPerMin_Avg 0.029	20	DefensiveBlocksPerMin_Avg 0.001	20	Other_Before 4.218

The results highlight differences in the scale and distribution of scores across the selected metrics. Information Gain values were extremely low in absolute terms. Chi-square scores were slightly more dispersed, with a maximum of 24.896 and only five variables exceeding the threshold of 14. ReliefF, on the contrary, showed a broader and more structured distribution: the top features reached values above 0.090, with at least ten predictors exceeding 0.050, suggesting a stronger local discriminative power for this metric.

When analyzing the results by metric:

- Information Gain identified ALI_Before (0.002), Injury_Count_Before (0.002), and MTI_Before (0.002) as the most informative features.
- Chi-square highlighted ALI_Before ($\chi^2 = 24.896$), Injury_Count_Before (17.576), and TotalShotPerMin_Avg15 (16.628) as the most statistically associated with the target, with MTI_Before, Forwards, and BCI_Before also showing moderate relevance.
- ReliefF yielded considerably higher scores overall, with Age (0.094), BCI_Before (0.092), ALI_Before (0.092), and MTI_Before (0.080) emerging as the top contributors.

The analysis now focuses on the the binary classification target *Injury_Next3Days*, evaluated during the training phase on the rebalanced training dataset.

Table 19 - features ranked by Relief, Info. Gain and Chi-square with respect to *Injury_Next3Days* on rebalanced dataset

		#	ReliefF		#	Info. gain		#	χ^2
1	N BCI_Before		0.079	1	N TakeOnsAttempPerMin_Avg15	0.011	1	N TotalShotPerMin_Avg15	92.113
2	N BMI		0.067	2	N FreeKickStatsPerMin_Avg15	0.011	2	N TakeOnsAttempPerMin_Avg15	90.749
3	N MTI_Before		0.060	3	N TotalShotPerMin_Avg15	0.010	3	N OffensiveImpactActionsPerMin_Avg	81.373
4	N ALI_Before		0.060	4	N DribblerChallengedPerMin_Avg15	0.010	4	N OffensiveImpactActionsPerMin_Avg15	76.683
5	N Other_Before		0.054	5	N DefensiveBlocksPerMin_Avg15	0.009	5	N ALI_Before	76.138
6	N Age		0.046	6	N OffensiveImpactActionsPerMin_Avg	0.009	6	N MTI_Before	73.464
7	N Injury_Count_Before		0.039	7	N ALI_Before	0.009	7	N Injury_Count_Before	73.169
8	N MinutesLast30Days		0.039	8	N OffensiveImpactActionsPerMin_Avg15	0.007	8	G Forwards	71.644
9	N ProgressiveCarryDistancePerMin_Avg		0.033	9	N TotalShotPerMin_Avg	0.007	9	N TotalShotPerMin_Avg	61.407
10	N TotalCarryDistancePerMin_Avg		0.030	10	N MTI_Before	0.007	10	N TakeOnsAttempPerMin_Avg	56.138
11	N DribblerChallengedPerMin_Avg		0.027	11	N Injury_Count_Before	0.006	11	N PassesAttemptedPerMin_Avg	49.636
12	N MinutesLast15Days		0.026	12	N TakeOnsAttempPerMin_Avg	0.006	12	N TouchesPerMin_Avg	42.642
13	G Fullbacks	2	0.026	13	N Other_Before	0.006	13	N BCI_Before	26.377
14	N SymPctMinutesLast15vsPrev15		0.025	14	G Forwards	0.005	14	N MinutesLast30Days	25.813
15	N FinalScore		0.023	15	N ActionInPenaltyAreaPerMin_Avg	0.005	15	N PassesAttemptedPerMin_Avg15	23.535
16	N LongPassesStatsPerMin_Avg		0.022	16	N PassesAttemptedPerMin_Avg	0.005	16	N Other_Before	22.854
17	G Central Midfielders	2	0.020	17	N DefensiveImpactActionsPerMin_Avg15	0.004	17	N FreeKickStatsPerMin_Avg15	18.999
18	N PassesAttemptedPerMin_Avg		0.020	18	N ActionInPenaltyAreaPerMin_Avg15	0.004	18	N CarriesPerMin_Avg	18.948
19	N ActionInPenaltyAreaPerMin_Avg		0.019	19	N TouchesPerMin_Avg	0.004	19	G Fullbacks	18.569
20	N Injury_Past30Days		0.018	20	N BMI	0.004	20	N TouchesPerMin_Avg15	18.246

The magnitude of the relevance scores varied across the three evaluation criteria. Information Gain values peaked at 0.011. Chi-square scores showed dispersion, with the most significant variable reaching 92.113 and several others exceeding 70. ReliefF has the highest-ranked predictors scoring between 0.079 and 0.060, and additional features above 0.030.

When examining each criterion individually:

- Information Gain ranked TakeOnsAttempPerMin_Avg15 and FreeKickStatsPerMin_Avg15 as the most relevant (both 0.011), followed by TotalShotPerMin_Avg15 (0.010). Injury history variables such as ALI_Before, MTI_Before, and Injury_Count_Before followed closely.
- Chi-square highlighted TotalShotPerMin_Avg15 ($\chi^2 = 92.113$), TakeOnsAttempPerMin_Avg15 (90.749), and OffensiveImpactActionsPerMin_Avg (81.373) as the most statistically associated with the target. Additional variables exceeding 70 included

OffensiveImpactActionsPerMin_Avg15, ALI_Before, MTI_Before, and Injury_Count_Before.

- ReliefF showed the highest scores for BCI_Before (0.079), followed by BMI (0.067), MTI_Before and ALI_Before (both 0.060). Several exposure and workload indicators such as MinutesLast30Days, ProgressiveCarryDistancePerMin_Avg, and DribblerChallengedPerMin_Avg also appeared among the top contributors.

A comparative analysis between the original and rebalanced training datasets for the prediction of *Injury_Next3Days* reveals a shifts in the most relevant features. While history injury variables consistently appeared among the top-ranked predictors across, their relative importance decreased slightly in the rebalanced dataset, particularly in the Information Gain and Chi-square rankings. In contrast, match-related performance metrics showed a marked increase in relevance, occupying leading positions in the rebalanced scenario. ReliefF scores also highlighted a broader range of physical exposure and workload indicators, including ProgressiveCarryDistancePerMin_Avg, MinutesLast30Days, and SymPctMinutesLast15vsPrev15, which were less prominent in the unbalanced configuration.

5. Discussion

This section critically discusses the key findings of this study.

The results obtained during the development phase highlight a superiority of tree-based algorithms over the linear model. For the variable *Injury_Categorization_Next20Days*, both Gradient Boosting and Random Forest achieved very high macro-F1 scores, however, this apparent effectiveness is largely sustained by a strong tendency to classify instances as “None”, as illustrated by the confusion matrices in the previous chapter, suggesting an optimistic bias driven by class imbalance.

A similar pattern emerges for the variable *Injury_Next20Days*. During training, the ensemble models correctly identify situations with a high risk of injury when injuries do occur, while keeping the false-positive rate low. After applying SMOTENC, the precision to detect an injury increases, and the number of false-positive remain low. In contrast, Logistic Regression, limited by its linear structure and inability to capture nonlinear interactions, performs consistently worse already in the development phase.

The variable *Injury_Next3Days*, affected by extreme class imbalance ($\approx 3\%$ positive cases), presents additional challenges, both ensemble approaches show low performance with near-zero recall; similarly, the results of logistic regression are unreliable, as the model tends to behave randomly when predicting whether an injury will occur. In this case, the application of SMOTENC, restores sensitivity for the tree-based models without increasing false alarms, confirming the utility of rebalancing.

However, during the testing phase on data from the second half of the 2023/24 season, the performance observed in training is not replicated. The ensemble models trained on the original dataset fail to detect any injuries, highlighting their sensitivity to class imbalance. On the other hand, when trained on rebalanced data, the models tend to overcompensate, classifying most cases as "injuries" due to overly shifted decision boundaries.

Increasing number and depth of trees allowed the ensemble models to learn specific decision rules within the training set, resulting in improved performance metrics during training. However, these rules were overly adapted to the training data, leading to poor generalization on new samples. Attempts to reduce model complexity by lowering the number and depth of trees did not lead to meaningful improvements. Logistic Regression

showed no clear signs of overfitting during training, also failed to perform satisfactorily on the test set.

Overall, the comparison between simple and more complex models suggests that the prediction difficulties encountered are not attributable to the learning algorithm itself. Rather, they reflect the intrinsic complexity of the phenomenon under study, which is influenced by multiple external factors that are either unobservable or not represented in the dataset. These factors limit the ability to produce replicable predictions under realistic conditions, regardless of the level of sophistication of the model employed.

Despite the intrinsic limitations of the problem, the analysis still made it possible to identify which variables contribute the most to predicting injury risk, both in the short term (3 days) and in the medium term (20 days). In the case of short-term prediction, the main challenge lies in the rarity of positive events. Medium-term prediction slightly benefits from a higher number of positive cases and from greater temporal stability, which allows the models to detect patterns associated with workload accumulation and injury history.

As reported in the *Table 18* from the Results chapter, the features associated with short-term injury prediction showed very low *Information Gain* values. This indicates that, when considered individually, these variables have limited informational power in distinguishing between injured and non-injured cases.

However, the same variables obtained higher scores when evaluated using *ReliefF* and *Chi-square* metrics. These indicators account not only for the distribution of the target variable, but also for local feature interactions (*ReliefF*) and statistical dependence between categorical variables and the output (*Chi-square*). This suggests that some features, while weak in isolation, may still offer relevant information when combined with others.

Across all importance measures, injury history emerged as the most informative group of features. In particular, the total number of previous injuries (Injury_Count_Before) and the breakdown by type (e.g., MTI, ALI, BCI) ranked among the top predictors, as they reflect accumulated player vulnerability over time, an established risk factor in the literature.

Following this, some variables related to the athlete's individual profile, such as Age and BMI, also demonstrated significant informational value, confirming the role of structural predispositions in injury risk modelling.

The recurring presence, among the most informative variables, of offensive indicators: as TotalShotPerMin, TakeOnsAttemptPerMin, OffensiveImpactActionsPerMin, and the categorical variable Forwards suggests that players with a clearly offensive role and active involvement in attacking phases are exposed to a higher risk of injury. These variables, along with ProgressiveCarryDistancePerMin_Avg, reflect not only the physical intensity sustained during matches but also the number of high muscular load actions (accelerations, changes of direction, sprints) that typically define the functional profile of attacking players.

Lastly, among workload-related indicators, variables such as SymPctMinutesLast15vsPrev15 and MinutesLast30Days proved particularly informative. Notably, the model assigned greater relevance to cumulative workload over the last 30 days than to shorter periods (7 or 15 days), suggesting that long-term fatigue accumulation may play a more prominent role in injury risk than acute exposure. At the same time, quickly variations in recent workload, captured by the SymPct variable may indicate an increased risk due to a mismatch between physical demand and the player's physiological adaptation capacity.

In the case of medium-term injury prediction, the average *Information Gain* values are higher than those observed for short-term prediction. This suggests that the input features, when considered individually, contain a greater amount of useful information to distinguish between injured and non-injured cases. As a result, the predictive signal appears more stable, allowing the model to identify more robust patterns.

As already observed in the three-day prediction model, injury history remains the most influential group of predictors across all evaluation metrics. This further reinforces the evidence that chronic vulnerability represents a central risk factor.

Also features such as Age and BMI also maintain high informational value, as previously noted. One of the most significant differences concerns the variable *Injury_Past30Days*, which played a marginal role in the three-day model but becomes the top predictor in the

20-day prediction according to *ReliefF*. This highlights how recent injury events have a delayed but meaningful impact on the athlete's vulnerability over the medium term.

The study therefore highlights that injury prediction in football can hardly rely on linear models or the isolated analysis of individual variables, as the phenomenon is driven by complex and non-linear dynamics. The interaction between factors is crucial and must be properly captured to enhance the predictive performance of the algorithms.

Despite the structural limitations of the available data, the analysis made it possible to identify a set of particularly informative variables that provide a concrete foundation for risk modelling. The results suggest paying special attention to athletes with a high history of injuries: an injury should not be treated as an isolated, time-bound event, but rather as a persistent indicator of vulnerability, capable of increasing the likelihood of recurrence.

Additional relevant factors include the player's age and physical condition, a high workload over the past 30 days, significant workload variation over the previous 15 days, and a role-specific predisposition to engage in numerous high-intensity actions during matches. When combined, these variables can help guide player management throughout the season.

If integrated into an effective monitoring system, these elements can assist clubs in reducing prolonged absences, containing the financial losses associated with injuries, and maintaining the expected level of sporting performance, as discussed in the previous chapters.

6. Conclusion

This study shows that injuries are a critical issue for Football clubs, affecting both finances, through direct and indirect costs, and on-field performance. It evaluates how well machine-learning models can predict injuries in the short term (3 days) and medium term (20 days), analysing data from more than 12 000 matches and a heterogeneous sample of 100 Serie A players across three seasons.

Three algorithms: Logistic Regression, Random Forest and Gradient Boosting, were trained, and class imbalance was addressed with the SMOTENC oversampling technique. Although the models achieved good precision during training, their external F1-score settled at about 0.20, underlining the complexity of injury prediction and the need for richer data. Feature-importance analysis identified injury history, age, BMI and several high-intensity offensive indicators (e.g. take-ons, shots per minute, progressive carries) as key variables. This suggests that future studies should include these features and that decision-makers in sports contexts should closely monitor them when managing player workload and match availability

While the study offers theoretical and practical contributions, it presents several limitations that should be considered. To improve clinical reliability, it should incorporate data not currently available in public repositories: GPS metrics (distance covered, sprint counts, accelerations, changes of direction); perceived-fatigue scores; physiological indicators such as heart-rate variability, sleep quality, hydration status and blood biomarkers (haemoglobin, creatine kinase, cortisol). Match-context variables, such as opponents' high-intensity defensive actions and weather conditions that affect pitch quality, could also add value.

Further research could include economic variables such as salary, current market value and projected end-of-season value for each player, so that injury probability can be combined with indices estimating a club's financial risk.

Finally, this thesis demonstrates that injury prediction is a multifactorial challenge. Publicly available data alone cannot provide fully reliable medical support, but they offer a useful starting point for identifying risk profiles and informing management decisions during the season.

Bibliography

- AIC. (2024). “Injury Time” – Il peso economico degli infortuni. *Sustainability (Switzerland)*, 11(1), 1–14.
[https://www.assocalcatori.it/sites/default/files/attachment/news/INJURY TIME AIC.pdf](https://www.assocalcatori.it/sites/default/files/attachment/news/INJURY%20TIME%20AIC.pdf)
- Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*.
- Bauer, P., & Anzer, G. (2021). Data-driven detection of counterpressing in professional football: A supervised machine learning task based on synchronized positional and event data with expert-based feature extraction. *Data Mining and Knowledge Discovery*, 35(5), 2009–2049. <https://doi.org/10.1007/s10618-021-00763-7>
- Carey, D. L., Ong, K., Whiteley, R., Crossley, K. M., Crow, J., & Morris, M. E. (2018). Predictive modelling of training loads and injury in Australian football. *International Journal of Computer Science in Sport*, 17(1), 49–66.
<https://doi.org/10.2478/ijcss-2018-0002>
- Drawer, S., & Fuller, C. W. (2002). Evaluating the level of injury in English professional. *British Journal of Sports Medicine*, 36(6), 446–451.
- Ekstrand, J., Häggglund, M., & Waldén, M. (2011). Injury incidence and injury patterns in professional football: The UEFA injury study. *British Journal of Sports Medicine*, 45(7), 553–558. <https://doi.org/10.1136/bjsm.2009.060582>
- Eliakim, E., Morgulev, E., Lidor, R., & Meckel, Y. (2020). Estimation of injury costs: Financial damage of English Premier League teams’ underachievement due to injuries. *BMJ Open Sport and Exercise Medicine*, 6(1), 1–5.
<https://doi.org/10.1136/bmjsem-2019-000675>
- Falese, L., Della Valle, P., & Federico, B. (2016). Epidemiology of football (soccer) injuries in the 2012/2013 and 2013/2014 seasons of the Italian Serie A. *Research in Sports Medicine*, 24(4), 426–432. <https://doi.org/10.1080/15438627.2016.1239105>
- Football Benchmark. (2024). *The European Elite 2024 - Football Clubs’ Valuation*. 1–23.
- Freitas, D. N., Mostafa, S. S., Caldeira, R., Santos, F., Fermé, E., Gouveia, É. R., & Morgado-Dias, F. (2025). Predicting noncontact injuries of professional football players using machine learning. *PLoS ONE*, 20(1), 1–21.
<https://doi.org/10.1371/journal.pone.0315481>
- Fuller, C. W., Ekstrand, J., Junge, A., Andersen, T. E., Bahr, R., Dvorak, J., Häggglund, M., McCrory, P., & Meeuwisse, W. H. (2006). Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *British Journal of Sports Medicine*, 40(3), 193–201. <https://doi.org/10.1136/bjsm.2005.025270>

- Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115(March), 105151. <https://doi.org/10.1016/j.engappai.2022.105151>
- Goes, F. R., Brink, M. S., Elferink-Gemser, M. T., Kempe, M., & Lemmink, K. A. P. M. (2021). The tactics of successful attacks in professional association football: large-scale spatiotemporal analysis of dynamic subgroups using position tracking data. *Journal of Sports Sciences*, 39(5), 523–532. <https://doi.org/10.1080/02640414.2020.1834689>
- Häggglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., & Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12), 738–742. <https://doi.org/10.1136/bjsports-2013-092215>
- Hierons, R. (1999). Machine learning. Tom M. Mitchell. Published by McGraw-Hill, Maidenhead, U.K., International Student Edition, 1997. ISBN: 0-07-115467-1, 414 pages. Price: U.K. £22.99, soft cover. In *Software Testing, Verification and Reliability* (Vol. 9, Issue 3). [https://doi.org/10.1002/\(sici\)1099-1689\(199909\)9:3<191::aid-stvr184>3.0.co;2-e](https://doi.org/10.1002/(sici)1099-1689(199909)9:3<191::aid-stvr184>3.0.co;2-e)
- Howden's Professional Sport. (2024). *Men's European football injury index 2023/24*. <https://www.howdengroupholdings.com/news/howden-2022-23-mens-european-football-injury-index>
- Hu, L., Chen, J., Vaughan, J., Yang, H., Wang, K., Sudjianto, A., & Nair, V. N. (2020). *Supervised Machine Learning Techniques : An Overview with Applications to Banking*.
- Leckey, C., Van Dyk, N., Doherty, C., Lawlor, A., & Delahunt, E. (2024). Machine learning approaches to injury risk prediction in sport: A scoping review with evidence synthesis. *British Journal of Sports Medicine*. <https://doi.org/10.1136/bjsports-2024-108576>
- López-Valenciano, A., Ruiz-Pérez, I., Garcia-Gómez, A., Vera-Garcia, F. J., De Ste Croix, M., Myer, G. D., & Ayala, F. (2020). Epidemiology of injuries in professional football: A systematic review and meta-analysis. *British Journal of Sports Medicine*, 54(12), 711–718. <https://doi.org/10.1136/bjsports-2018-099577>
- Majumdar, A., Bakirov, R., Hodges, D., McCullagh, S., & Rees, T. (2024a). A multi-season machine learning approach to examine the training load and injury relationship in professional soccer. *Journal of Sports Analytics*, 10(1), 47–65. <https://doi.org/10.3233/jsa-240718>
- Majumdar, A., Bakirov, R., Hodges, D., McCullagh, S., & Rees, T. (2024b). A multi-season machine learning approach to examine the training load and injury relationship in professional soccer. *Journal of Sports Analytics*, 10(1), 47–65. <https://doi.org/10.3233/jsa-240718>

- Nassis, G. P., Verhagen, E., Brito, J., Figueiredo, P., & Krstrup, P. (2023). A review of machine learning applications in soccer with an emphasis on injury risk. *Biology of Sport*, 40(1), 233–239. <https://doi.org/10.5114/biolSport.2023.114283>
- Padmanabha Reddy, Y., Viswanath, P., & Eswara Reddy, B. (2018). Semi-supervised learning: a brief review. *International Journal of Engineering & Technology*, 7(1.8), 81. <https://doi.org/10.14419/ijet.v7i1.8.9977>
- Palmer, D. (2015). Epidemiology of Sports Injuries and Illnesses CO. *ABC of Sports and Exercise Medicine*, July, 1–4.
- Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, 10(5). <https://doi.org/10.1145/3343172>
- Pfaffmann, D., Herbst, M., Ingelfinger, P., Simon, P., & Tug, S. (2016). *Analysis of Injury Incidences in Male Professional Adult and Elite Youth Soccer Players: A Systematic Review*. 51(5), 410–424. <https://doi.org/10.4085/1062-6050-51.6.03>
- Pulici, L., Certa, D., Zago, M., Volpi, P., & Esposito, F. (2023). Injury Burden in Professional European Football. *Clinical Journal of Sport Medicine*, 00(00), 1–8.
- Raines, T., Tambe, M., & Marsella, S. (2000). Automated assistants to aid humans in understanding team behaviors. *Proceedings of the International Conference on Autonomous Agents*, 419–426. <https://doi.org/10.1145/336595.337558>
- Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., & Medina, D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. *PLoS ONE*, 13(7), 1–15. <https://doi.org/10.1371/journal.pone.0201264>
- Schilde, K. (2025). The political economy of European security. In *The Political Economy of European Security*. <https://doi.org/10.1017/9781108182492>
- Sprouse, B., Alty, J., Kemp, S., Cowie, C., Mehta, R., Tang, A., Morris, J., Cooper, S., & Varley, I. (2024). The Football Association Injury and Illness Surveillance Study: The Incidence, Burden and Severity of Injuries and Illness in Men's and Women's International Football. *Sports Medicine*, 54(1), 213–232. <https://doi.org/10.1007/s40279-020-01411-8>
- Stone, P., & Veloso, M. (1998). Layered approach to learning client behaviors in the robocup soccer server. *Applied Artificial Intelligence*, 12(2–3), 165–188. <https://doi.org/10.1080/088395198117811>
- Tizikara, D. K., Serugunda, J., & Katumba, A. (2022). *Machine Learning-Aided Optical Performance Monitoring Techniques* : 2(January), 1–21. <https://doi.org/10.3389/frcmn.2021.756513>

- Tomasz Piłka, Bartłomiej Grzelak, Aleksandra Sadurska, T. G., & Dyczkowski, and K. (2023). *Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors*. 1–15.
- Torrejón, L. N., Martínez-Serrano, A., Villalón, J. M., & Alcaraz, P. E. (2024). Economic impact of muscle injury rate and hamstring strain injuries in professional football clubs. Evidence from LaLiga. *PLoS ONE*, *19*(6), 1–13. <https://doi.org/10.1371/journal.pone.0301498>
- Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., & Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, *8*(1). <https://doi.org/10.1186/s40634-021-00346-x>
- Waldén, M., Mountjoy, M., McCall, A., Serner, A., Massey, A., Tol, J. L., Bahr, R., D’Hooghe, M., Bittencourt, N., Della Villa, F., Dohi, M., Dupont, G., Fulcher, M., Janse Van Rensburg, D. C. (Christa), Lu, D., & Andersen, T. E. (2023). Football-specific extension of the IOC consensus statement: Methods for recording and reporting of epidemiological data on injury and illness in sport 2020. *British Journal of Sports Medicine*, *57*(21), 1341–1350. <https://doi.org/10.1136/bjsports-2022-106405>
- Wang, L. (2024). *Injury Prediction in Sports : A survey on machine learning methods*. March 2020, 1–18.
- Wang, M. (2014). *Evaluating Technical and Tactical Abilities of Football Teams in Euro 2012 Based on Improved Information Entropy Model and SOM Neural Network*. *9*(11), 293–302.
- Wong, A., Li, E., Le, H., Bhangu, G., & Bhatia, S. (2025). A predictive analytics framework for forecasting soccer match outcomes using machine learning models. *Decision Analytics Journal*, *14*(December 2024), 100537. <https://doi.org/10.1016/j.dajour.2024.100537>

Web References

<https://www.deloitte.com/uk/en/services/consulting-financial/analysis/deloitte-football-money-league.html>

Retrieved June 3, 2025

<https://www.assocalcatori.it/news/il-peso-economico-degli-infortuni>

Retrieved June 3, 2025

<https://www.ibm.com/think/topics/machine-learning-algorithms>

Retrieved June 3, 2025

<https://www.boe.es/eli/es/rdl/2015/04/30/5>

Retrieved June 17, 2025

https://www.sas.com/en_ie/insights/articles/analytics/machine-learning-algorithms.html

Retrieved June 15, 2025

<https://www.ibm.com/think/topics/boosting>

Retrieved June 7, 2025

<https://www.ibm.com/think/topics/unsupervised-learning>

Retrieved June 8, 2025

<https://www.geeksforgeeks.org/machine-learning/unsupervised-learning/>

Retrieved June 8, 2025

<https://www.altexsoft.com/blog/semi-supervised-learning/>

Retrieved June 9, 2025

<https://www.ibm.com/think/topics/k-means-clustering>

Retrieved June 19, 2025

<https://fbref.com/en/>

Retrieved April 27, 2025

<https://www.transfermarkt.com/>

Retrieved April 28, 2025

<https://www.ibm.com/think/topics/gradient-boosting>

Retrieved July 4, 2025

<https://www.ibm.com/think/topics/random-forest>

Retrieved July 4, 2025

<https://www.ibm.com/think/topics/logistic-regression>

Retrieved July 4, 2025

ChatGPT Plus and Deepl, AI-based tools, were used to improve the readability and language of the work.