# PREDICTING THE RISK OF INJURY OF PROFESSIONAL FOOTBALL PLAYERS WITH MACHINE LEARNING

Bruno Gonçalo Pires Martins

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2018

Title: PREDICTING THE RISK OF INJURIES IN PROFESSIONAL FOOTBALL PLAYERS WITH MACHINE LEARNING

BRUNO GONÇALO PIRES MARTINS

MEGI

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# PREDICTING THE RISK OF INJURY OF PROFESSIONAL FOOTBALL PLAYERS WITH MACHINE LEARNING

by

Bruno Gonçalo Pires Martins

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, with a specialization in Information Analysis and Management

**Advisor:** Professor Roberto Henriques, PhD

July 2018

# ABSTRACT

Sports analytics is quickly changing the way sports are played. With the rise of sensor data and new tracking technologies, data is collected at an unprecedented degree which allows for a plethora of innovative analytics possibilities, with the goal of uncovering hidden trends and developing new knowledge from data sources.

This project creates a prediction model which predicts a player’s muscular injury in a professional football team using GPS and self-rating training data, by following a Data Mining methodology and applying machine learning algorithms. Different sampling techniques for imbalanced data are described and used. An analysis of the quality of the results of the different sampling techniques and machine learning algorithms are presented and discussed.

# KEYWORDS

Football; Injury Prediction; Sports Analytics; Data Mining; Predictive Analytics; Imbalanced Data.

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **CRISP-DM** | Cross Industry Standard Process for Data Mining |
| **DM** | Data Mining |
| **EDA** | Exploratory Data Analysis |
| **EPTS** | Electronic Performance and Tracking System |
| **GPS** | Global Positioning System |
| **IDE** | Integrated Development Environment |
| **LDA** | Linear Discriminant Analysis |
| **NCAA** | National Collegiate Athletic Association |
| **NN** | Neural Network |
| **PCA** | Principal Component Analysis |
| **SMOTE** | Synthetic Minority Over-Sampling Technique |
| **XGBoost** | Extreme Gradient Boosting |

# 1. INTRODUCTION

## 1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Sports is transforming from an area that relied exclusively on human knowledge to an area fueled by data analysis (Cintia, Giannotti, Pappalardo, Pedreschi, & Malvaldi, 2015). Sports analytics is the new field that models professional sports performance with a scientific approach using methods and techniques from different disciplines such as data mining, statistics, kinesiology, game theory, among others (Miller, 2015).

This field rapidly grew with the publication of Michael Lewis book on the use of statistical analysis in baseball player scouting (Lewis, 2004). A quantitative analyst working for the small Oakland Athletic baseball team used a statistical approach to discover and rank undervalued players in the market, and hence, compete with wealthier teams by spending a reduced amount of money. Since then, many of the strategies used by this small baseball team, have been used in some way by other Major League Baseball teams, as well as in other sports (Fry & Ohlmann, 2012).

Sports analytics is rapidly changing how sports are played. Before, it was common to rely on the "gut instinct" and "intuition" to make decisions regarding the game. The advances of sports science merged with the power of analytics, have made this type of decision making obsolete (Davenport, Thomas H.; Harris, 2013). Using data mining, machine learning, and statistical techniques, manager and coaches can find favorable approaches for an entire upcoming season (Schumaker, Solieman, & Chen, 2010). Hence, any sports team that can turn data into actionable knowledge has the potential to secure a competitive advantage.

In football, there's a marked trend of the increase of analytics use in United States professional sports leagues, and in European leagues (Ofoghi, Zeleznikow, MacMahon, & Raab, 2013). As the world's most popular sport, published research in football analytics has yet to attain the same level of sophistication as analytics in other professional sports (Cintia et al., 2015). These football analytics, which describe in detail training aspects, in-game statistics or the overall behavior of teams during the games, pave the road to understand, model and possibly predict the complex patterns underlying the performance of a football team. Due to the advances in different information systems, technology is now not only possible to use different and innovative data sources (like sensors and training data), but also employ analyses to the data gathered to obtain constructive insights that might change the team strategies (Davenport, Thomas H.; Harris, 2013).

One of the problems that can undermine any football team are injuries. The overall injury rate in the National Collegiate Athetic Association (NCAA) football is 7.7 per 1,000 athlete exposures, which combines games and training sessions (Dick et al., 2007). Injuries in football are common and harmful to any team. Consequently, it’s essential for any team to manage and try to prevent injuries from happening.

Sports analytics can play an essential role in this function by using data to anticipate if a player can get injured before the injury takes place. Monitoring and analyzing player data is critical to ensure that players received a suitable training schedule and given a suitable recovery time between matches and training sessions (Colby, Dawson, Heasman, Rogalski, & Gabbett, 2014). Sports scientists, coaches, and analytics professionals can work closely together to determine the appropriate training schedule to maximize the performance of a player, without increasing his injury rates (Simjanovic, Hooper, Leveritt, Kellmann, & Rynne, 2009). This project aims to create an injury prediction model that can be used to flag individuals at high risk of injury and plan intervention programs aimed at reducing its risk.

## 1.2. STUDY RELEVANCE AND IMPORTANCE

The importance of this project can be justified in four different dimensions: player performance, psychological impact, team performance and economic impact.

**Player performance:** In professional football, players run a risk of injury between 65% and 91% during a season (Hägglund, Waldén, & Ekstrand, 2005). The estimation is that each player will incur at least one injury per year, assuming he plays to a limit of 500 hours per year on a professional football team (Dvorak & Junge, 2000). The player’s activity can be interrupted for large periods of time (weeks or months) due to rehabilitation, surgical and medical treatments, making injuries very debilitating events for any football players career (Stubbe et al., 2015). The consequence of frequent injuries during career for any player is an increased difficulty to achieve maximum skill and performance levels because of lack of competition playing time and training (Pfirrmann, Herbst, Ingelfinger, Simon, & Tug, 2016).

**Player Psychological impact:** Top performing athletes are susceptible to problematic emotional responses after an injury. Even though they tend to be used to high levels of stress due to the competition element naturally present in sports, they are not exempt of increased depression and anxiety as well as diminished self-esteem following an injury (Leddy, Lambert, & Ogles, 1994). The emotional response to injury is different among different athletes, but often, the injury can trigger serious mental health problems like substance abuse and eating disorders (Putukian, 2016). Hence, Injury is a problem that has a psychological impact, as well as physical.

**Team performance:** Injuries have negative consequences not only for the player but also for the team. If any injuries are sustained, team results can suffer (Stubbe et al., 2015). In fact, previous research shows that successful football teams had significantly fewer injuries compared to non-successful football teams (Arnason et al., 2004). Also, the performance in male professional football is significantly affected by the number of injuries (Hägglund et al., 2013). Hence injury prevention is an important factor to increase a team's performance and its competitive advantage.

**Economic impact:** A general classification of the economic costs of sports injuries can be divided into direct costs and indirect costs (Öztürk, 2013). The direct costs are the costs of medical treatment, like diagnostic expenses, cost of medicine, rehabilitation, among others. The indirect costs are the potential lost earnings that come with the loss of playing time. In professional football, it's not uncommon to spend millions of dollars on a single player. Just one injury of this key player has the potential of having quite a significant economic impact in indirect costs (Hägglund et al., 2013).

## 1.3. STUDY OBJECTIVES

The prediction of injuries is an important and stimulating subject in sports and a vital component to prevent the risks and perils which accompany injuries.

In the proposed project we aim to create an accurate injury model to predict the likelihood of a player to get injured, for a professional football team using sensor and self-rating data, by following a Data Mining methodology and applying machine learning algorithms.

We can divide the primary goal of this project into several objectives:

1. Perform an Exploratory Data Analysis (EDA) to uncover data quality problems, reveal data insights and apply the necessary transformations;
2. Investigate different types of data aggregation operations to maximize the knowledge extraction and create a dataset that can be used for accurate injury prediction;
3. Evaluate different machine learning models and accuracy measures, comparing them against a test dataset;
4. Select the most accurate model for injury prediction, taking into consideration the accuracy measures.

## 2. SPORTS ANALYTICS AND INJURY PREDICTION

Sports analytics have been a growing field with increased academic and professional relevance. Not only has its use increased in sports organizations, but the number of published articles and books about this topic has grown exponentially. Figure 1 reveals the expanded evolution of sports analytics published articles on Google Scholar over the last 20 years. From a negligible research interest in 1997, we see its continued escalation until the year 2017.

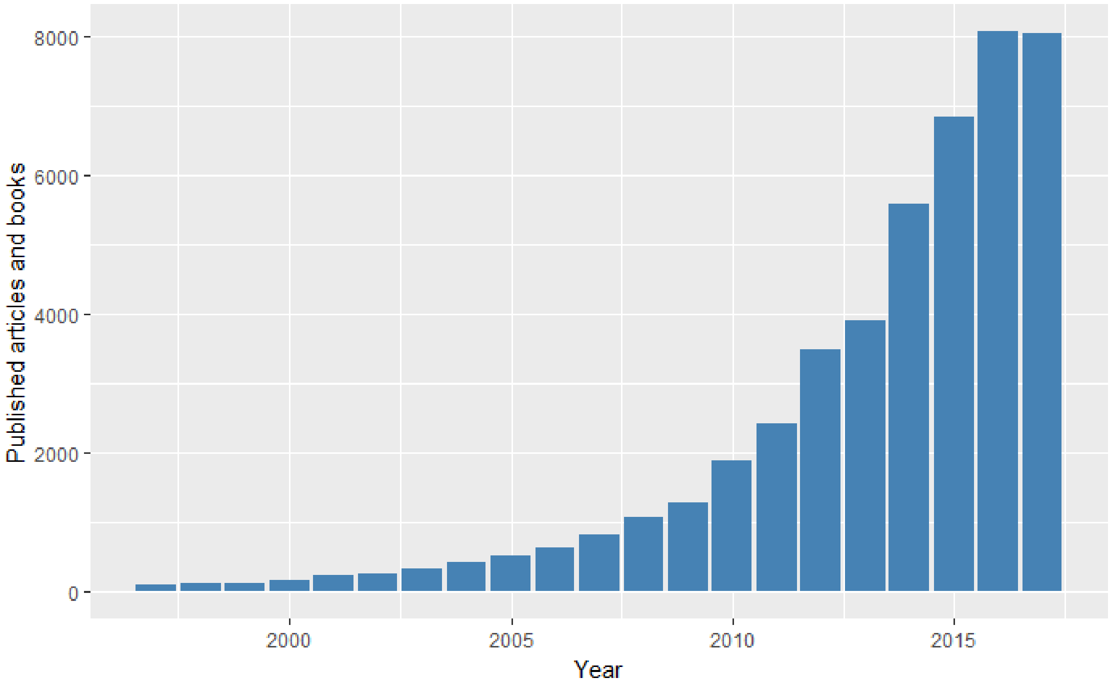

Figure 1 - Published articles and books with keyword “Sports Analytics” from 1997 to 2017 in Google Scholar.

The use of analytics has not only risen in academic research but most importantly, it has sparked interest in most sports organizations and fundamentally changed the way the sports are played (Ofoghi et al., 2013).

It’s being applied throughout many different areas in a sports organization: from scouting and talent identification (Bhandari et al., 1997; Coles, 2015;), to video analysis (Schumaker et al., 2010), to forecasting results (Rotshtein, Posner, & Rakityanskaya, 2005), statistical simulations (Kelley, Mureika, & Phillips, 2006), defining player ratings (McHale, Scarf, & Folker, 2012), performance analysis (Ofoghi et al., 2013), among others.

Currently, most of the discussion about sports analytics either concerns the performance analysis of players or teams or is indirectly related to injury prevention in players (Casals & Finch, 2016). Both require high-quality data, presented with robust statistical approaches.

Despite most football injuries being traumatic (Dvorak & Junge, 2000) - hence, not able to be predicted using this methodology), around one-third of all injuries are muscular and so liable to risk factors that contribute to their occurrence (Ekstrand, Hägglund, & Waldén, 2011). The problem of injury prediction has more commonly been studied by identifying these risk factors using typical statistical tools like logistic regression (Bittencourt et al., 2016). This methodology is used to identify linear relationships, but it has its limitations since injuries are a complex phenomenon and it's not easy to test variables with different weights and feedback loops as well as testing the different causal relationships (Quatman, Quatman, & Hewett, 2009). In the face of multiple factors possibly associated with injuries, it is also necessary to have control of confounding and multiple influences and knowledge.

One of the methods of monitoring different training factors in professional football is the use of sensor data with GPS technology (Aughey, 2011). It allows quantifying with precision the player's training load by measuring running distances, accelerations, impacts, among others. The use of this data for injury prediction is not yet fully explored (Colby et al., 2014).

Due to the big amount of data produced by these various tracking devices and the multifactorial complex and the dynamic nature of sports injuries, machine learning algorithms and data mining methodology are a particularly good fit. Machine learning in sports injuries has been used for diagnosing, where Bayesian classifiers and decision trees are a common approach for decision making support in the medical field. A number of methods for the diagnosis of sport injuries in football using machine learning have been developed and applied (I. Zelic, Kononenko, Lavrac, & Vuga, 1997) as well as motion capture analysis and assessment for injury prevention (Alderson, 2015) but machine learning models that can enable the identification of the risk factors of injury using training data are not so common in academic publications (Bittencourt et al., 2016) and are only now emerging in recent sports analytics conferences. One possible factor for the scarcity of research is the confidentiality of the data for each team since it can be considered a source of competitive advantage.

A relevant article presented in one of the most well-known sports analytics conferences: MIT Sloan Sports Analytics Conference, proposed to prevent the in-game injuries for NBA players using machine learning (Talukder et al., 2016). The authors have used game data, player workload, and data, and team schedules from two seasons with a sliding window approach where they aggregate the average data for a 14-day span and have a 7-day prediction window for the response variable of whether a player got injured. It's unlikely that the data of one training session can achieve accurate results, hence this aggregated approach. The results demonstrated strong accuracy in predicting whether a player would get injured in that 7-day prediction window.

In football, Kampakis (2016) used machine learning algorithms to attempt to predict injury, uniquely based on sensor data of training sessions of Tottenham Hotspur Football Club. The aggregation window approach was also used with the time frame of a week with promising results. It also compares using data only from injured players and using data from all players (injured and non-injured). The study verified that sensor training data could contain valuable information for the task of injury prediction.

Besides objective sensor data, the subjective player self-assessment tools, like Hooper's Index (Hooper & Mackinnon, 1995) or RPE (Haddad, Padulo, & Chamari, 2014), are also identified as useful tools for monitoring overtraining, quantifying exercise intensity and evaluating individual need for recovery, hence providing valuable information for the task of injury prediction (Kellmann, 2010; Simjanovic et al., 2009).

Considering current academic research has not yet developed a machine learning model with the combination of the training sensor data and the subjective player self-ratings, the present project can amplify previous research by applying different machine learning algorithms using a data mining approach to create a predictive model for football injuries using not only the training sensor data but also the player's self-rating recovery assessment, which will enrich the available academic literature for injury prediction.

# 3. METHODS AND DATA

## 3.1. METHODOLOGY

A Data Mining (DM) approach will be conducted to analyze the data and create the injury prediction model. DM is a branch of computer science and artificial intelligence used for the discovery of hidden trends and patterns, as well as extracting knowledge from data sources (Witten, Frank, & Hall, 2011). In business setting, DM is frequently used to obtain new knowledge for decision making, but it can also be used to investigate and explore sports data, particularly elite football performance data (Schumaker et al., 2010). The most frequently used DM techniques are: classification, rule mining, clustering and relationship modeling (Ofoghi et al., 2013).

Data Mining projects don't have a standard framework to be used in its projects. Nevertheless, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is a popular methodology frequently used in DM projects to increase their success (Chapman et al., 2000). Even though CRISP-DM was created for business use, the general framework can be easily adapted to a sports data mining project.

It outlines a cycle with six phases (Figure 2):

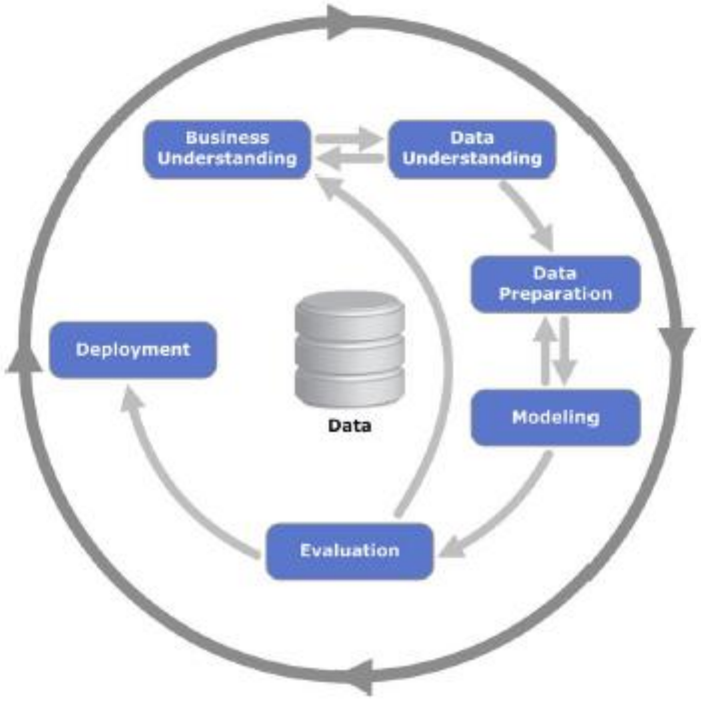

Figure 2 - CRISP-DM (Source: IBM SPSS Modeler CRISP-DM Guide)

In this project, the 6 phases will be composed of the following activities:

1. **Business understanding** – The first phase aims to define the project objectives as well as the requirements – done in Chapter 1. Considering this project is to be implemented on a specific football team, this phase was accomplished in collaboration with the team's objectives. Afterward, this understanding was converted into a data mining problem.

2. **Data understanding** – This second stage starts with the data collection, organization, and proceeds with EDA. The organization provides a large quantity of data from training sessions and matches, mostly GPS tracker data and self-evaluation questionnaires. An initial EDA will be able to discover insights in the data, identify data quality problems and discern relevant relationships between variables.

3. **Data preparation** – This third stage has the objective to use the raw data to construct the final dataset that is going to be used for modeling. Specific activities done at this phase can be fixing multiple data quality problems and variable transformations to maximize the amount of knowledge that can be extracted from this dataset considering our objective.

4. **Modeling** – At the fourth stage, different DM techniques are selected, applied and optimized. Different machine learning algorithms are going to be tested at this phase: Support Vector Machine, Random Forest, Extreme Gradient Boosting, Logistic Regression, among others.

5. **Evaluation** – At the fifth stage, the constructed model(s) are systematically assessed and compared to make sure they achieve the objectives defined in the first stage.

6. **Deployment** – The final stage aims to organize and present the knowledge in a way that can be used by the organization. The newly obtained knowledge is not useful if the team cannot use it. The IT department of the organization is responsable for the deployment and implementation phase.

## 3.2. SOFTWARE USED

We implement the machine learning models using the scripting language R (R Development Core Team, 2011) along with the Integrated Development Environment (IDE) R-Studio (RStudio team & RStudio, 2015). As an open-source statistical programming language, most of the machine learning algorithms are already implemented in R so that it can be used efficiently in this project.
The libraries used in this project were: “dplyr” (Wickham, Francois, Henry, & Müller, 2017), “VIM” (Kowarik & Templ, 2016) , “ggplot2” (Wickham, 2009), “corrplot” (Wei & Simko, 2017), “psych” (Revelle, 2017), caret (Jed Wing et al., 2017), “ROSE” (Lunardon, Menardi, & Torelli, 2014), “xgboost” (Chen & Guestrin, 2016).

## 3.3. SOURCE DATA DESCRIPTION

The data used in this project was collected over the course of one football season by a professional football team. The team collected the training session data of individual players using an Electronic Performance and Tracking System (EPTS). This system uses a device to track player positions using GPS in combination with microelectromechanical devices (accelerometer, gyroscope, digital compass) as well as heart-rate monitors and other devices to measure load and other physiological parameters. Apart from the tracking system, the team also collected self-rating data from the players before and after the training session to monitor self-perceived levels of fatigue and training load.

The original data used for this project was divided into four groups: identification data of each training session (8 variables), training sensor data (16 variables), self-ratings of fatigue, stress, muscle soreness and sleep (5 variables) and injury variables (7 variables) to a total of 36 variables. All the data refers to training sessions during the collected football season.

### 3.3.1. Identification Data

The training sessions are scheduled using a system of periodization that plans according to appropriate cycles and training phases in a way to prioritize objectives, thus creating precise features for each training session.

This data uses two different cycles: mesocycles and microcycles. The cycles vary in the amount of time. A mesocycle has a duration of roughly one month. A microcycle has a duration of 1 week. In general, one mesocycle has approximately 4 to 5 microcycles.

Each training day was also identified about the match day. If the match day is labeled MD, the next day after a match is MD+1, the next two days after a match is MD+2 and the previous day after a match is MD-1. It varies from MD+4 to MD-4. This type of identification allows a precise definition of the training load.

The list of 8 variables in this group described in Table 1.

Table 1 - Identification Variables List

| | Variable | Description | Example Data |
|---|---|---|---|
| 1. | **Mesocycle** | A period of training that corresponds to one month | 5; 6; 7; |
| 2. | **Microcycle** | A period of training that corresponds to one week | 8;10; 12; |
| 3. | **Training Unit** | The number of the individual training | 39;40;41; |
| 4. | **Match Day** | Number of days before or after a match | MD+1; MD+2; MD-1 |
| 5. | **Date** | Training Date | 2015.11.04; 2015.12.06; 2015.12.11 |
| 6. | **Hour** | Training Hour | 12:17; 10:34; 17:20 |
| 7. | **Player** | Individual Player Identification | PLA1;PLA14; PLA45 |
| 8. | **Position** | Position in the field | Forward; Center-Back; Wing Back; |

### 3.3.2. Sensor Training Data

The sensor data of the training sessions was gathered by the EPTS sports tracker StatSports Viper®. This tracker is used by leading teams across the world in multiple sports on top competitions including Premier League, NFL, NBA, among others. The StatSports Viper® streams live data in real-time through the Viper Live Streaming software as well as logging all data for post-session download.

The variables used in this project are not only a result from the GPS data, but they can also be the outcome of a formula using the GPS, accelerometer, gyroscope or magnetometer data provided by the device. All variables are an average of the training session for each player.

There is a total of 16 variables in this group, described in Table 2.

Table 2 - Training sensor variables list

| | Variable | Description | Example Data |
|---|---|---|---|
| 1. | **Training Time** | Number of minutes the training took | 86.67; 112.90; 63.08; |
| 2. | **Distance Total** | Total distance covered in a training session | 2917.13; 339.43; 2744.56 |
| 3. | **Distance Per Min** | Total distance covered per minute in a training session | 29.65; 32.35; 44.23; |
| 4. | **High-Speed Running** | Distance traveled by a player when their speed is in either Zone 5 or 6 | 7.24;35.32; 63.94; |
| 5. | **High Speed Running Per Min** | Distance traveled by a player when their speed is in either Zone 5 or 6 averaged per minute | 1.63; 0.11; 3.00; |
| 6. | **Heart Rate Exertion** | Total exertion of a session based on weighted heart rate values | 45.13; 18.79; 33.42; |
| 7. | **Speed Intensity (SI)** | Measure of total exertion of a player based on “time at speed” | 165.73; 279.94; 174.13; |
| 8. | **Dynamic Stress Load (DSL)** | Total weighted impacts | 21; 165; 4; |
| 9. | **Lower Speed Loading** | Load associated with the low speed (static) activity alone | 39.16; 71.76; 8.22; |
| 10. | **Impacts** | Number of impacts on a training session | 165; 75; 21; |
| 11. | **Accelerations** | Acceleration activity on a training session | 54; 36; 6; |
| 12. | **Decelerations** | Deceleration activity on a training session | 45; 16; 27; |
| 13. | **Sprints** | Number of sprints in a session | 0; 16; 7; |
| 14. | **Fatigue Index** | Accumulated DSL from the total session volume. DSL divided by Speed Intensity. | 0.56; 0.61; 0.80 |
| 15. | **Energy Expenditure** | Total energy associated with running (measured in kcal) | 311.90; 719.02;660.90; |
| 16. | **Total Load (TL)** | Gives the total of the forces on the player over the session | 67.53; 95.95; 68.57 |

### 3.3.3. Hooper’s Index

The player’s self-ratings were measured with the Hooper‘s Index Questionnaire (Hooper & Mackinnon, 1995), which were filled at the beginning of each training session or game with the goal of measuring the player’s recovery state between training and competitions, as well as to monitor overtraining. Each player answers four items on a 7-point scale where one is “Very, very low” and seven is “Very, very high”. The summation of these four ratings is the Hooper’s Index.

There is a total of 5 variables described in Table 3.

Table 3 - Hooper’s Index variable list

| | Variable | Description | Example Data |
|---|---|---|---|
| 1. | **Q1** | Item 1: Muscle Pain in inferior members (1-7 scale) | 4; 2; 7; |
| 2. | **Q2** | Item 2: Sleep quality of last night (1-7 scale) | 3; 2; 4; |
| 3. | **Q3** | Item 3: Fatigue Level (1-7 scale) | 5; 7; 3; |
| 4. | **Q4** | Item 4: Stress Level (1-7 scale) | 1; 4; 6; |
| 5. | **SUM** | Hooper’s Index (Sum of the 4 items) | 16; 12; 19; |

### 3.3.4. Injury Data

Injury data were manually collected by the coaches as the players got injured during the season into a table. The injuries were differentiated between overuse/muscular and traumatic injuries.

There is a total of 36 rows and seven variables described in Table 4.

Table 4 – Injury variable list

| | Variable | Description | Example Data |
|---|---|---|---|
| 1. | **Date** | Date when the injury took place | 2015.07.14; 2015.07.31 |
| 2. | **Player** | Identification of the player | PLA15; PLA34 |
| 3. | **Context** | Where the injury took place (training or match) | Training; Match |
| 4. | **Muscular** | Whether the injury was muscular or not | TRUE; FALSE |
| 5. | **Mechanism** | Traumatic or overuse injury | Overuse; Traumatic |
| 6. | **Leg** | Right or left leg (or both) | Right, Left, Both |
| 7. | **Days** | Total number of days the player was injured | 8;3;14; |

# 4. DATA UNDERSTANDING

## 4.1. SUMMARY STATISTICS

Summary statistics of the dataset can be observed in Table 5.

Table 5 – Summary statistics

| Variables | n | mean | sd | median | min | max | range | skew | kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| TrainingTime | 2943 | 77.56 | 14.19 | 78.58 | 11.73 | 112.9 | 101.17 | -1.11 | 2.98 |
| DistanceTotal | 2943 | 5265.42 | 1748.72 | 5291.47 | 617.83 | 9790.98 | 9173.15 | 0.05 | -0.75 |
| DistancePerMin | 2943 | 70.29 | 36.36 | 72.15 | 13.88 | 958.69 | 944.81 | 12.74 | 266.26 |
| HSR | 2943 | 129.83 | 137.04 | 83.77 | 0 | 723.02 | 723.02 | 1.4 | 1.66 |
| HSRMin | 2943 | 1.64 | 1.69 | 1.1 | 0 | 9.03 | 9.03 | 1.32 | 1.38 |
| HRE | 2712 | 116.95 | 72.37 | 98.98 | 5.14 | 424.43 | 419.29 | 0.89 | 0.34 |
| SI | 2943 | 249.1 | 89.4 | 249.72 | 19.83 | 497.29 | 477.46 | 0.09 | -0.72 |
| DSL | 2943 | 120.12 | 77.53 | 101.42 | 7.68 | 619.46 | 611.78 | 1.47 | 3.29 |
| LSL | 2943 | 36.15 | 10.22 | 35.84 | 0 | 92.5 | 92.5 | 0.14 | 1.19 |
| Impacts | 2943 | 31.77 | 37.59 | 18 | 0 | 339 | 339 | 2.57 | 8.91 |
| Accelerations | 2943 | 37.25 | 16.5 | 36 | 0 | 108 | 108 | 0.38 | 0.41 |
| Decelerations | 2943 | 33.22 | 18.05 | 32 | 0 | 97 | 97 | 0.47 | -0.11 |
| Sprints | 2943 | 3.03 | 3.76 | 2 | 0 | 24 | 24 | 1.7 | 2.95 |
| FatigueIndex | 2943 | 0.47 | 0.24 | 0.41 | 0.1 | 1.76 | 1.66 | 1.79 | 3.94 |
| EE | 2943 | 540.68 | 181.67 | 541.48 | 51.81 | 1072.71 | 1020.9 | 0.07 | -0.67 |
| TL | 2943 | 76.17 | 24.06 | 75.29 | 0 | 155.13 | 155.13 | 0.17 | -0.51 |
| Q1 | 3321 | 3.23 | 1.25 | 3 | 1 | 7 | 6 | 0.04 | 0.21 |
| Q2 | 3325 | 2.9 | 1.18 | 3 | 1 | 7 | 6 | 0.22 | 0.04 |
| Q3 | 3325 | 3.14 | 1.32 | 3 | 1 | 7 | 6 | -0.04 | -0.14 |
| Q4 | 3325 | 2.53 | 1.30 | 3 | 1 | 7 | 6 | -0.15 | 0.03 |
| Hooper_Sum | 3323 | 11.79 | 3.77 | 12 | 4 | 28 | 24 | -0.15 | 0.03 |

We can detect a difference on the number of the sensor data variables to Hooper's variables of approximately 380 observations, which means that in some training sessions where Hooper self-rating was collected, sensor data wasn't available.

It's also possible to observe a difference in the scaling of the different variables, the mean values range from 0.47 to 5265.42.

The skewness and kurtosis values indicate that we might have possible outliers on the variables: TrainingTime, DistancePerMin, HSR, HSRMin, HRE, DSL, Impacts, Sprints, and FatigueIndex.

## 4.2. CORRELATION MATRIX

The correlation coefficient measures the extent to which two variables change together and evaluate the monotonic relationship between the different pairs of variables. Spearman rank-order correlation was used to calculate the correlation matrix to uncover the correlation between the different variables to accurately select the machine learning algorithm to apply.

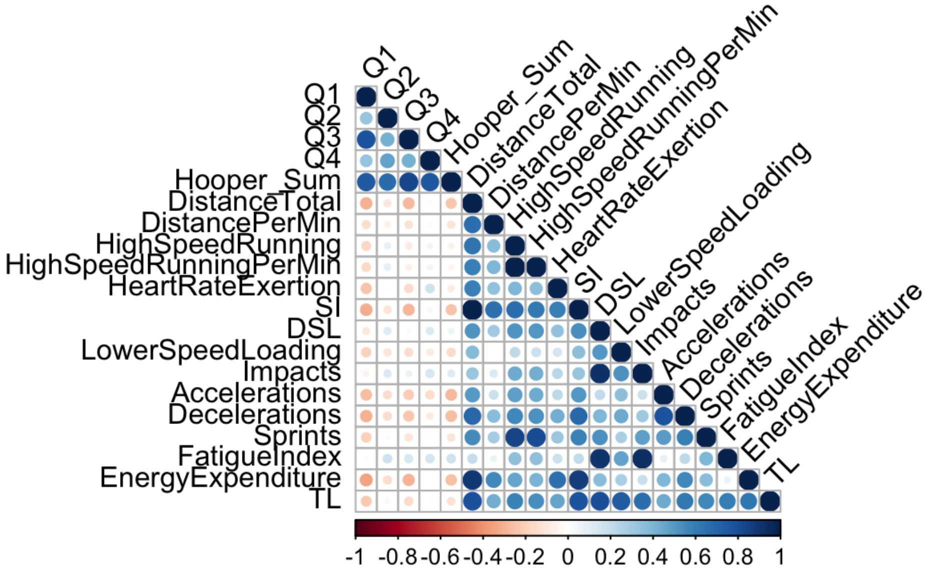

Figure 3 – Correlation plot

Figure 3 presents the correlation plot where – unsurprisingly - it's possible to observe a strong positive correlation between the Hooper question variables (Q1, Q2, Q3 and Q4) and "Hooper_Sum". The Hooper variables only have a weak negative correlation with the sensor training data.

Another unsurprising finding is the strong positive correlation of "Sprints" with "High-Speed Running" and "High Speed Running Per Min", since they measure the same type of activity.

We can also observe the strong positive correlation of "DistanceTotal" with SI ("Speed Intensity"), "Energy Expenditure" and TL ("Total Load"). The longer the distance ran in a training session, usually the higher the training load, this is a predictable discovery.

An interesting finding is a positive correlation of "Impacts" with "Fatigue Index". A possible interpretation might be that the more fatigued the players are, the more impacts they have. The positive correlation of "Impacts" with DSL ("Dynamic Stress Load") is only a natural consequence of the DSL formula which is the total of the weighted impacts.

## 4.3. MISSING VALUES

From Figure 4 we can observe that the variable with more missing values is the Heart Rate Exertion (HRE) with over 32% of the data being missing, which correlates to 1085 observations. Most sensor data variables have approximately 26% missing variables (approximately 1085 observations). The Hooper self-ratings variables have 17% missing records (approximately 707 observations). Considering most machine learning algorithms have difficulties handling missing data, this is a situation that will be handled on the data preparation step.

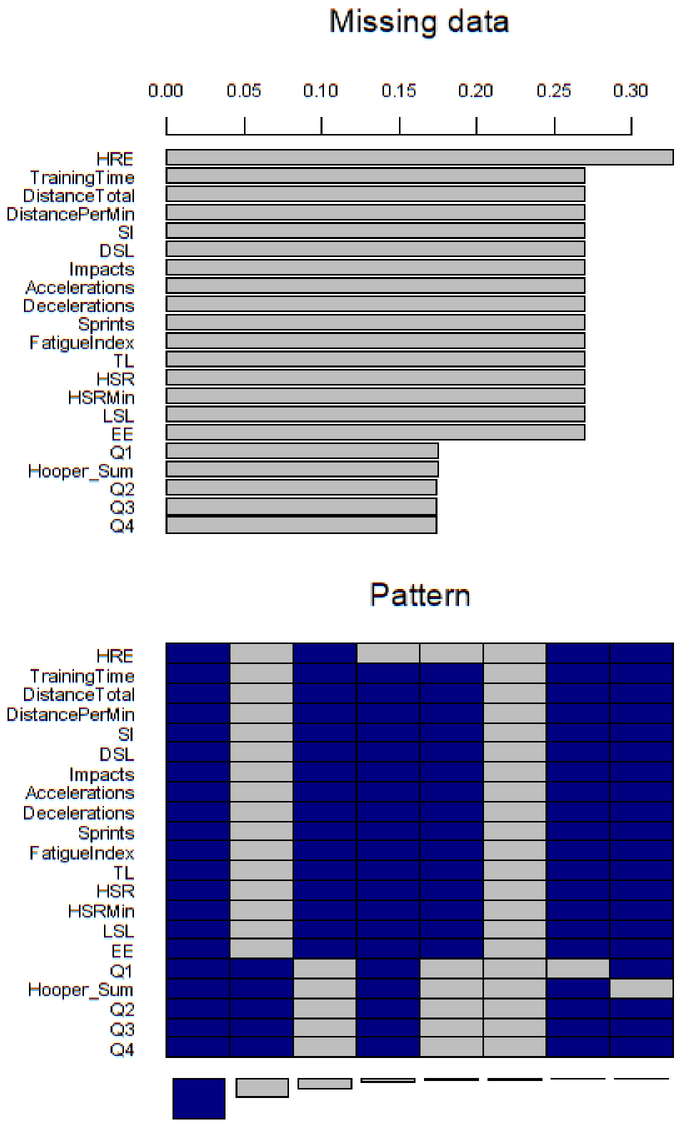

Figure 4 – Missing values plot

## 4.4. INJURY DESCRIPTION

On this project, only muscular injuries are used. Muscular injuries can be identified from the training and self-reporting data, unlike traumatic injuries, like the ones originated from a collision during training or a match, which cannot be predicted using this methodology. The injury severity and the time the player was in recovery were not considered for this article.

In 36 total injuries during the season, 18 were muscular injuries. From these 18 injuries, 5 were eliminated for happening outside of the normal season training of the team. The players were injured during the pre-season or during the break for playing with the national team when the training loads are different. Hence, 13 muscular injuries remained. From this observation, it’s possible to conclude that we have an unbalanced dataset.

## 4.5. VARIABLES COMPARISON

To better understand the dataset and the variable distribution, all the variables were compared with injured observations. This comparison allows recognizing the differences of non-injured to injured players without any machine learning algorithm. The plots were created using the variables with similar scaling for ease of understanding.

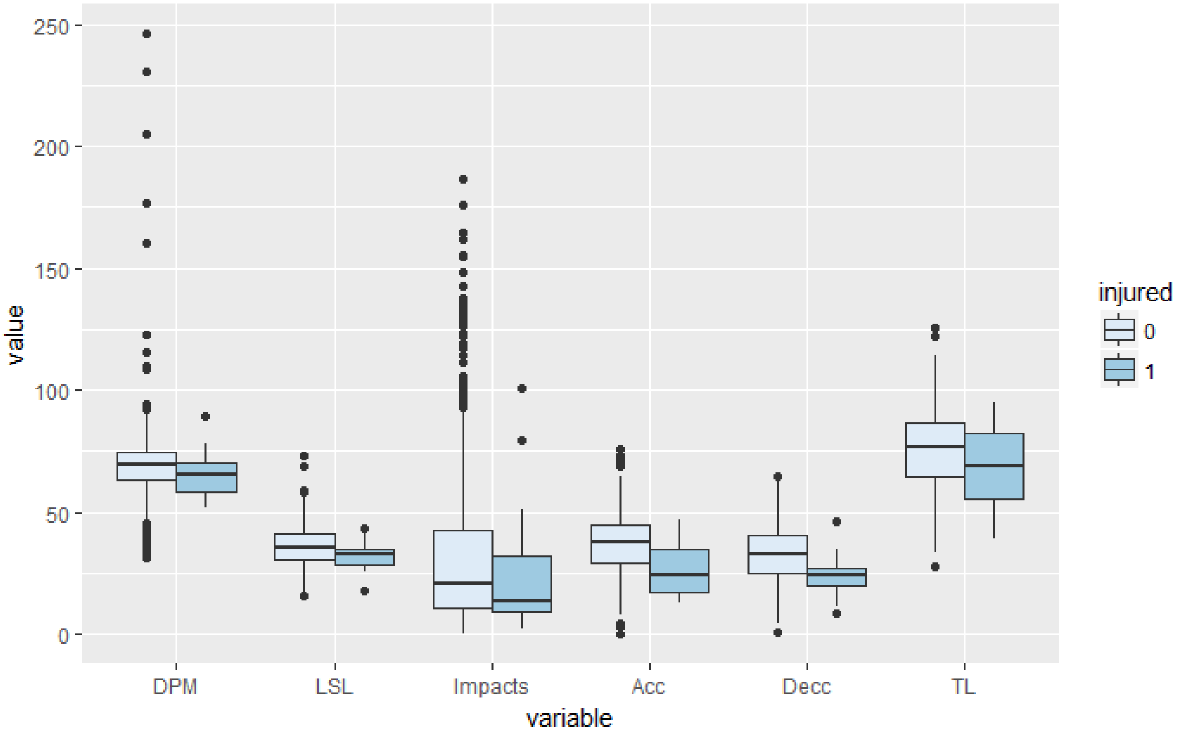

Figure 5 - DistancePerMinute (DPM), LowerSpeedLoading(LSL), Impacts, Accelerations (Acc), Decelerations (Decc) and Total Load (TL) variables splitted between injured and non-injured observations

By observing Figure 5, it's possible to immediately detect that all variables have a slightly lower mean value for the injured observations. On some variables, the difference is considerable higher like for example on Accelerations and Decelerations while other variables have a lower difference: DistancePerMinute (DPM) and LowerSpeedLoading (LSL). Outliers can also be detected on most variables.

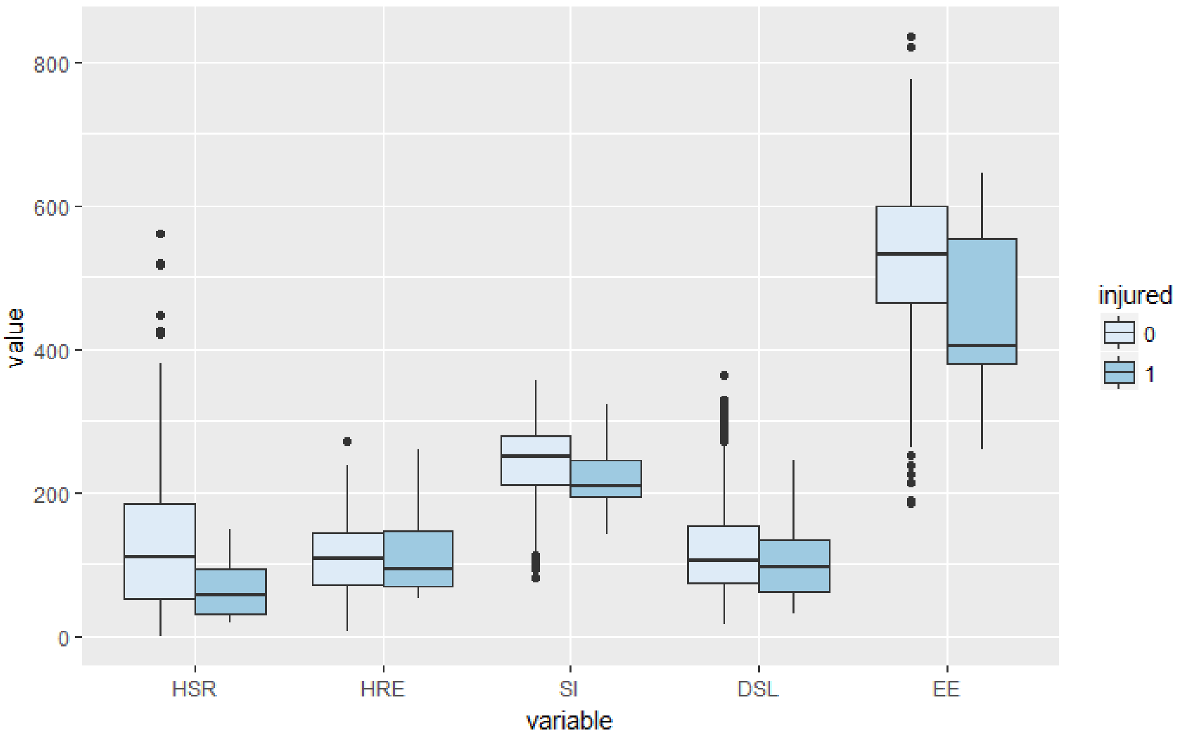

Figure 6 - High Speed Running (HSR), HeartRateExertion (HRE), Speed Intensity (SI), Dynamic Stress Load (DSL), Energy Expenditure (EE) variables split between injured and non-injured observations

Figure 6 shows the effect of injured players having a slightly lower mean on all variables, with some variables having a larger difference like Energy Expenditure (EE), High Speed Running (HSR) and Speed Intensity (SI).

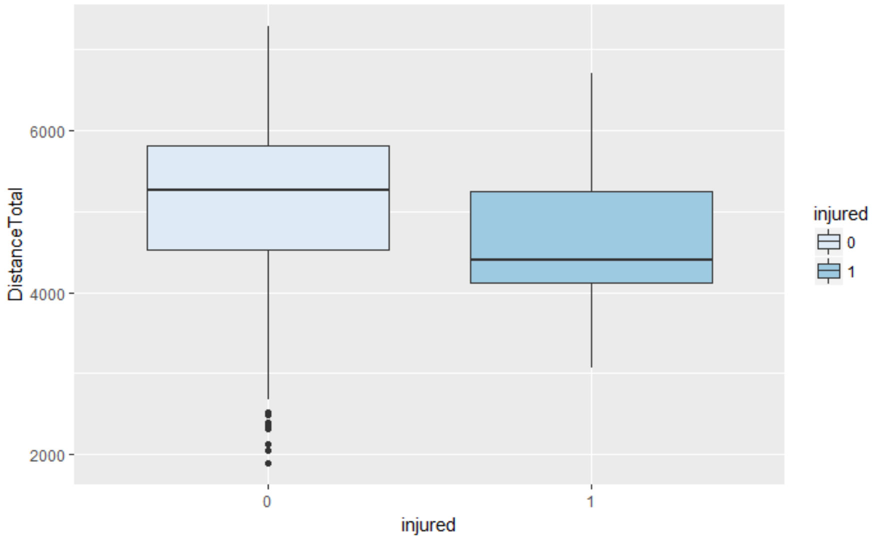

Figure 7 - Distance Total variable splitted between injured and non-injured observations

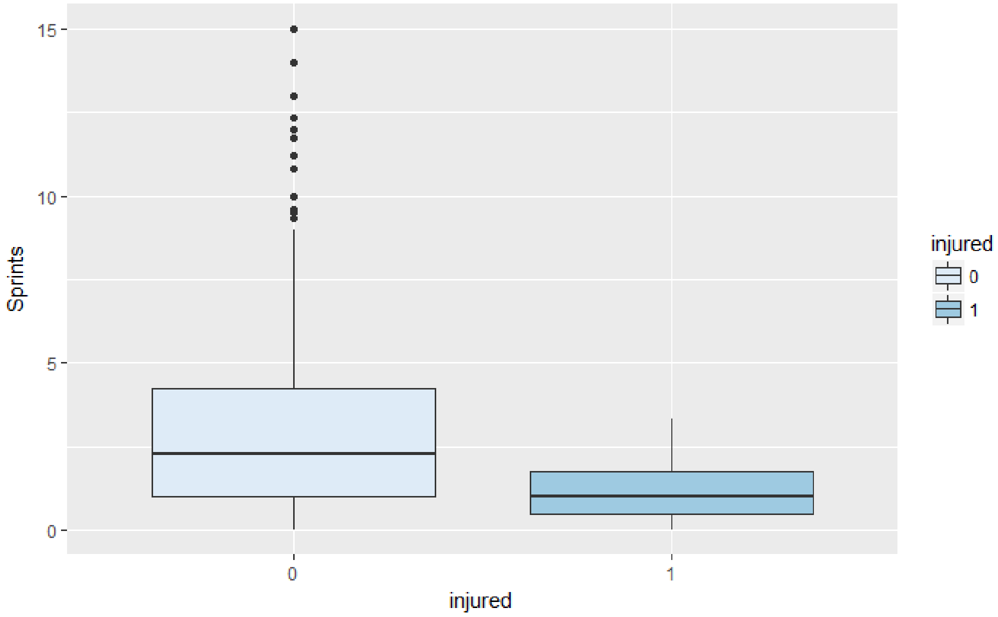

Figure 8 - Sprints variable splitted between injured and non-injured observations

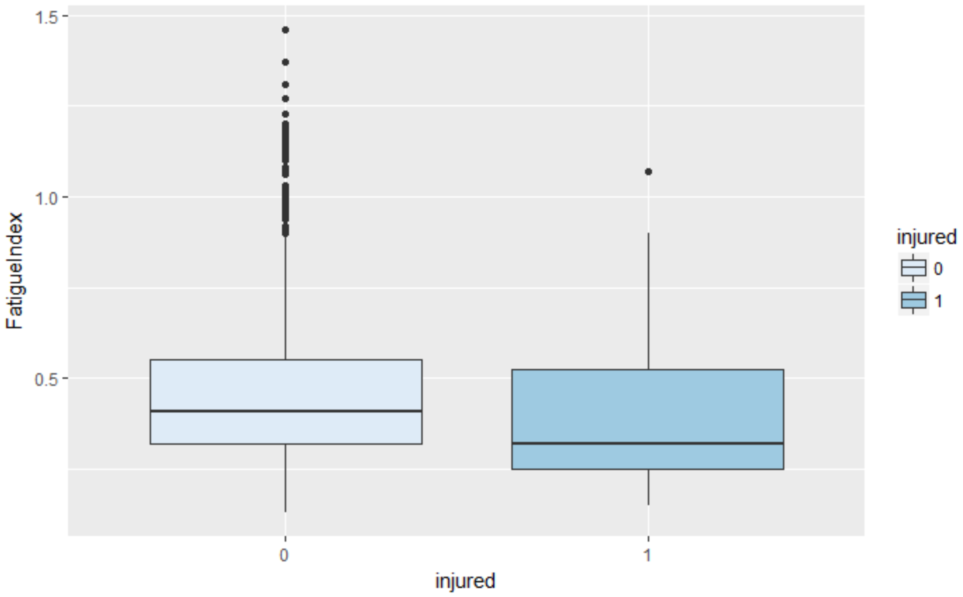

Figure 9 - Fatigue Index variable splitted between injured and non-injured observations

On the remaining three sensor data variables (Figure 7, Figure 8, Figure 9), it's possible to detect the same scenario of the variables having a slightly lower mean on injured observations. It's particularly surprising to observe Fatigue Index (Figure 9) where we might expect a higher mean on injured observations than on non-injured.

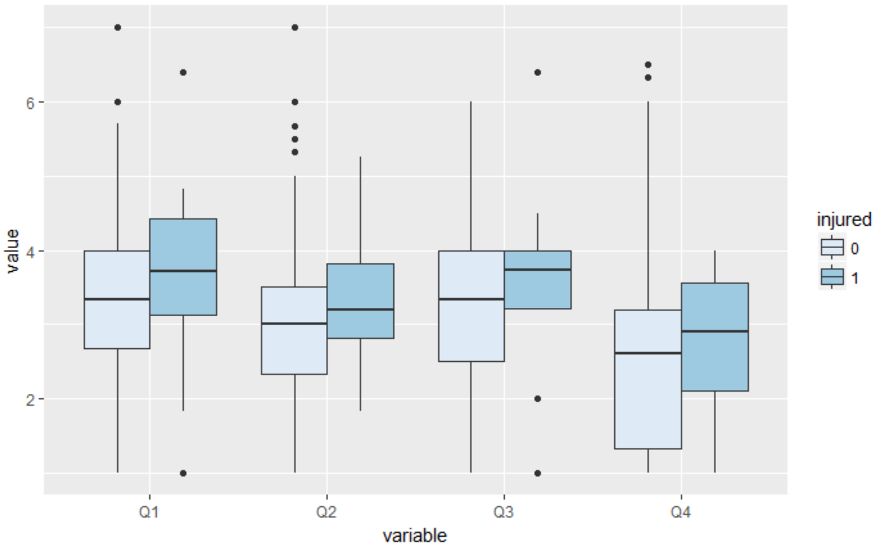

Figure 10 - Hooper variables splitted between injured and non-injured observations

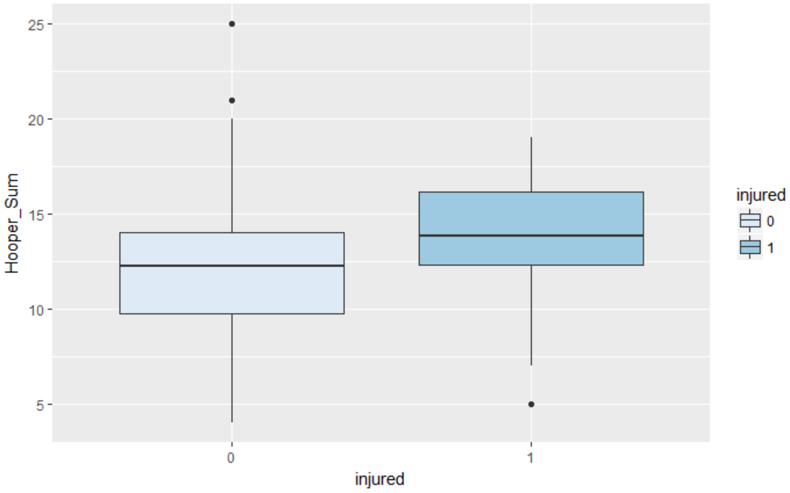

Figure 11 - Hooper Sum variable splitted between injured and non-injured observations

On the self-rating Hooper questionnaire (Figure 10, Figure 11), the relationship is reversed. A slightly higher mean on injured observations can be observed. This is not an unexpected outcome considering higher values of this scale can indicate an imbalance between stress and recovery.

# 5. DATA PREPARATION

## 5.1. DATA PRE-PROCESSING

Previous injury prediction models used aggregated data over a sliding window of time (chapter 2). The reason is that a muscular injury caused by overtraining is typically the result of a sequence of increase in training load and not of one individual training (Kellmann, 2010). Therefore, the model will be more likely to notice the difference in the data that leads to an injury if it's aggregated data from more than one training.

Considering it wouldn't be the most efficient approach to use the data of each training individually to predict the injury the data of one microcycle (one week) is, instead, averaged per player. Consequently, each observation is the average of the sensor data and self-ratings of one player in the training units contained within one microcycle. There are at least two training units aggregated for each player. A binary variable [0, 1] indicates whether the player became afflicted by an injury that week or not.

Other aggregation windows could have been used (for example: 2 days, 3 days, etc). We considered one week since it usually contains different training intensities and one match, as well as being a frequently used measure in football. Thus, using this window we can collect a more useful range of high-quality data and get a more accurate representation of the player's recovery and muscular state.

Observations that contained missing values were removed during this aggregation process. It could be argued that information is lost by deleting missing values instead of imputing them, but considering this is an unbalanced dataset, imputing them would only further add to the imbalance, as well as add more noise to the data. The methods for correcting the unbalanced dataset already add some noise – taking into account the artificial data methodology (see subsection 5.5.) –, hence the decision to delete the missing values.

Commonly, outliers are deleted (or treated) for better algorithmic performance, but on this dataset, even though outliers are present, it was decided not to delete them due to also deleting observations that belonged to the target injured class. Considering the already reduced number of injured observations, deleting them further would not aid the task of creating a prediction model.

From a 3470 rows dataset, after weekly aggregation and missing values deletion, 696 rows remained.

## 5.2. FEATURE SELECTION

The variable "Hooper_Sum" was removed considering the variable is nothing more than the sum of the four questions (Q1, Q2, Q3 and Q4), consequently doesn't add any value to the dataset. The variable "Training Time" was also deleted since it could be misrepresented as an injured player at the beginning of any training will necessarily have a lower training time.

On Table 6 is presented the complete list of variables used for the model with example data.

Table 6 - Complete list of variables and example data

| **Variables** | **Example data** |
|---|---|
| Q1 | 3.33; 5; 2.75; 2.5; 1; |
| Q2 | 4; 4; 3.5; 3.5; 2; |
| Q3 | 3.33; 4.67; 4; 2.75; 1; |
| Q4 | 3; 4; 3; 2.25; 1; |
| DistancePerMin | 69.8; 71.4; 65.4; 64; 58.5; |
| HighSpeedRunning | 220.8; 305.8; 276.5; 132.6; 27.4; |
| HighSpeedRunningPerMin | 2.64; 3.35; 3.11; 1.41; 0.3; |
| HeartRateExertion | 135.3; 164.4; 196; 161.2; 69.1; |
| SI | 287; 322; 275; 257; 244; |
| DSL | 168.6; 135.2; 205.9; 151.5; 75.1; |
| LowerSpeedLoading | 35.2; 35.3; 38.2; 32.1; 34.5; |
| Impacts | 43.8; 38; 75.8; 21.3; 13.5; |
| Accelerations | 41; 46.8; 45.5; 29.7; 36; |
| Decelerations | 41.8; 42.8; 55.2; 36; 24; |
| Sprints | 6; 6.25; 5.75; 0.67; 0; |
| FatigueIndex | 0.57; 0.41; 0.7; 0.56; 0.32; |
| EnergyExpenditure | 621; 608; 614; 506; 587; |
| TL | 88.2; 87.8; 84.3; 79.5; 68.9; |
| injured | 0; 0; 0; 0; 1; |

## 5.3. TESTING AND TRAINING DATASET

A cornerstone of the supervised learning algorithms is the division of data into a training and a testing dataset. The training dataset is used to build the classification model, and the testing dataset serves as the unknown data we wish to predict. 70% of the data is assigned to the training set and 30% to the testing set. The actual number of instances can be seen in Table 7.

This division on an already reduced dataset may raise some questions. If all examples are used in the training process, even if an evaluation technique like cross-validation is used, the algorithm can be optimistic compared to the performance of completely unseen samples. By doing this division, even though we are reducing the number of target variable observations, we gain in the reliability of the results.

Table 7 - Training and testing datasets number of observations

| Dataset Partition | Total Observations (N) | Not Injured (0) | Injured (1) |
|---|---|---|---|
| **Training** | 489 | 479 | 10 |
| **Testing** | 207 | 204 | 3 |

## 5.4. BALANCING THE DATASET

As discussed in Chapter 4 and as it can be observed from Figure 12, the target variable has a highly unbalanced distribution with only 1.9% of the observations of our dataset belonging to injured players.

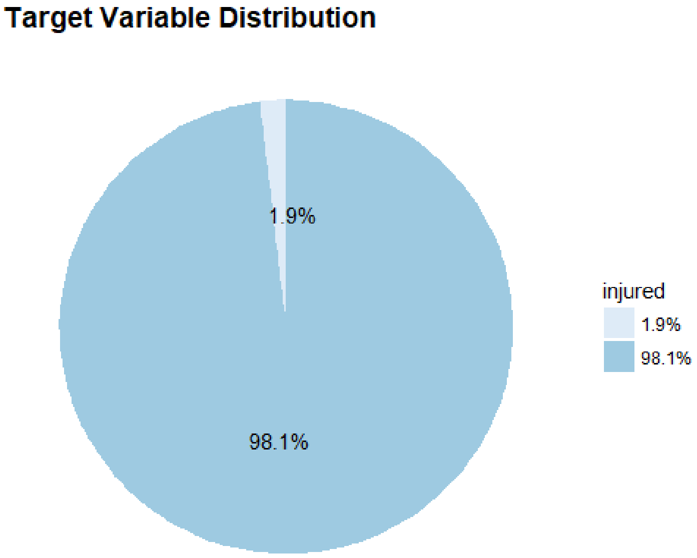

Figure 12 - Target Variable Distribution Plot

An unbalanced dataset with one prevailing class occurs in many different fields and circumstances. Most commonly, it is found in fraud detection where typically fraud transactions are a rare occurrence on a dataset. It can also be found in social sciences where the researchers might be attempting to identify anomalous behaviors. In these situations, the intrinsic nature of the problem generates an unbalanced dataset. Injuries in football are one of these circumstances where the number of training sessions during a whole season will typically be much larger than the number of injuries.

The problem of using an unbalanced dataset in machine learning is the tendency of the model to ignore the smaller class and focus on the dominant class, creating a skewed prediction (Menardi & Torelli, 2014). Hence, overlooking the issue of an unbalanced dataset leads to significant costs, both in model estimation and when the evaluation of the accuracy of the estimated model has to be measured (Batista, Prati, & Monard, 2004).

There are several methods for correcting the problem of the unbalanced dataset. Most research focuses on model estimation stage (Menardi & Torelli, 2014) and uses various techniques of data resampling (such as random oversampling or undersampling) as well the generation of artificial data (using specific algorithms to generate artificial examples similar to the rare observations) (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

For this article, five different data resampling techniques were used as an attempt to correct the unbalanced dataset:

1) **Random oversampling:** this method randomly replicates N samples from the minority class. The dataset can then become prone to overfitting since the new data is an exact copy of the data of the minority class (Menardi & Torelli, 2014).

2) **Random undersampling:** this method randomly eliminates N samples from the majority class as an attempt to balance the dataset. The most important problem is the elimination of potentially useful data for the training of the model (Menardi & Torelli, 2014).

3) **Random undersampling and oversampling combined:** the unified use of undersampling and oversampling with different over/undersampling rates was used with good results in previous studies (Kotsiantis et al., 2006).

4) **Synthetic Minority Over-Sampling Technique (SMOTE):** is an algorithm that aims to generate artificial examples of the minority class by interpolating between several examples that lie together. This way, it avoids the problem of overfitting that is present in oversampling (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)

5) **ROSE:** is an algorithm that generates new (artificial) data around the minority class according to a smoothed bootstrap method. Fundamentally, it generates a new synthetic dataset where the two classes have approximately the same number of examples (Menardi & Torelli, 2014).

To evaluate the effectiveness of the different data correction methods, they were applied uniquely to the training set, generating five different training sets, each with one different data resampling technique, leaving the testing dataset untouched and unbalanced. The new dataset sizes and the percentage of the injured classes are presented in Table 8. The R code for the dataset division with the different balancing methods is in Appendix 1.

Table 8 – Balanced datasets

| Data Correction Method | Total (N) | Not Injured (0) | Injured (1) | Percentage of Injured Class |
|---|---|---|---|---|
| **Random Oversampling** | 750 | 479 | 271 | 36% |
| **Random Undersampling** | 160 | 150 | 10 | 6% |
| **Random Oversampling & Undersampling Combined** | 410 | 220 | 190 | 46% |
| **ROSE** | 489 | 262 | 227 | 46% |
| **SMOTE** | 640 | 480 | 160 | 25% |

It can be argued that creating the resampled training dataset with these methods before model fitting may lead to optimistic estimates of performance since they may not reproduce the class imbalance that future predictions would most likely encounter. An alternative solution is to use the resampling during the cross-validation procedure (Kuhn, 2017). But considering the substantial increase in computing time and the decreased level of control of the training set, only the first approach was used.

## 5.5. REDUNDANCY REDUCTION

As an attempt to reduce redundancy in the dataset and improve training accuracy, Principal Component Analysis (PCA) was applied to the unbalanced (original) and the five balanced datasets. By reducing the number of features, using PCA, the original dataset is projected into a smaller space by using the *k* orthogonal non-correlated vectors to represent the data, resulting in dimensionality reduction (Agarwal, 2013). This way dimensionality reduction can be performed on the datasets, and then fit a machine learning algorithm to a smaller set of variables, while maintaining a big part of the variability of the original dataset (James, Witten, Hastie, & Tibshirani, 2013). One of the drawbacks of using this technique is the interpretability, especially after having used the components in a machine learning algorithm. Considering each principal component is a linear combination of the original variables, knowing which variable influenced more the target variable can be hard to explain (Bishop, 2007).

PCA always generates as many components as existing variables (Agarwal, 2013), to select how many of them to retain, there are three well-known heuristic methods (Berge & Kiers, 1996):

1) **Pearson criterion – or cumulative proportion of explained variance:** This criterion recommends retaining the components that explain approximately 80% to 90% of the total variance (Pearson, 1901).

2) **Kaiser's Rule:** this criterion recommends retaining as many components as are the eigenvalues larger than 1 (Kaiser, 1960).

3) **Scree Plot:** This graphical criterion where the curve of the scree plot is used to select the components (Cattell, 1966).

Considering Pearson's criterion is a popular and elegant method to PCA (Berge & Kiers, 1996) it was the chosen criterion in this project. The cumulative variance of the components on the different datasets can be seen in Table 9.

Table 9 - Cumulative percentage of variance on the six datasets. Emphasis added to the selected number of components.

| Components | Unbalanced | SMOTE | ROSE | Over | Under | Both |
|---|---|---|---|---|---|---|
| **comp 1** | 37.01 | 38.33 | 28.08 | 39.64 | 36.49 | 39.92 |
| **comp 2** | 55.19 | 56.46 | 41.62 | 57.56 | 54.85 | 57.84 |
| **comp 3** | 66.16 | 67.71 | 52.30 | 69.09 | 66.32 | 70.57 |
| **comp 4** | 73.83 | 75.31 | 59.13 | 76.70 | 74.41 | 77.82 |
| **comp 5** | **80.50** | **81.20** | 65.03 | **82.92** | **81.66** | **84.37** |
| **comp 6** | 85.13 | 85.64 | 69.78 | 87.08 | 86.85 | 88.14 |
| **comp 7** | 88.55 | 89.21 | 73.61 | 90.36 | 90.57 | 91.23 |
| **comp 8** | 91.46 | 92.14 | 76.88 | 92.93 | 93.17 | 93.58 |
| **comp 9** | 93.96 | 94.60 | **80.01** | 95.20 | 95.29 | 95.49 |
| **comp 10** | 95.99 | 96.47 | 82.96 | 96.60 | 96.70 | 96.77 |
| **comp 11** | 97.12 | 97.48 | 85.59 | 97.56 | 97.73 | 97.84 |
| **comp 12** | 98.09 | 98.38 | 87.98 | 98.49 | 98.43 | 98.72 |
| **comp 13** | 98.99 | 99.15 | 90.29 | 99.17 | 99.02 | 99.29 |
| **comp 14** | 99.52 | 99.57 | 92.46 | 99.60 | 99.56 | 99.66 |
| **comp 15** | 99.77 | 99.78 | 94.52 | 99.80 | 99.79 | 99.81 |
| **comp 16** | 99.87 | 99.88 | 96.51 | 99.89 | 99.89 | 99.92 |
| **comp 17** | 99.94 | 99.96 | 98.30 | 99.96 | 99.95 | 99.97 |
| **comp 18** | 100 | 100 | 100 | 100 | 100 | 100 |

# 6. MODELING & EVALUATION

## 6.1. MODEL SELECTION

Different types of models were used to find the most suitable one for the dataset used. Considering the small dataset size, different models can be tested without an exaggerated increase of computing time. Hence six different models were selected based on its characteristics of being able to handle the limitations present in our dataset.

### 6.1.1. Naïve Bayes

Naïve Bayes has been proven effective in many practical applications, including systems performance and medical diagnosis (Rish, 2001). Despite its shortcoming and even if its probability estimates are not accurate, it can work unexpectedly well in classification (Hilden, 1984), in fact, it has been used as a classifier to support the diagnostic of sports injuries using medical data (Igor Zelic, Kononenko, Lavra, & Vuga, 1997).

### 6.1.2. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis, originally proposed by Fisher (Fisher, 1936), has a similar approach to PCA, but not only does it find the component axes that maximize the variance, but it also uncovers the axes that maximize the separation between the projected means of the classes, hence, it can be used for supervised and unsupervised learning (Izenman, 2013). LDA was chosen by its ability to accurately work on smaller datasets in many research areas, not limited to cancer classification, facial recognition, and text classification (Sharma & Paliwal, 2015).

### 6.1.3. Support Vector Machines (SVM)

SVM is a classification machine learning algorithm that uses a maximization of the margin between two classes closest points to seek the optimal separating hyperplane between these two groups (Cortes & Vapnik, 1995). When a linear separator can't be found, observations are projected through kernel techniques, into a higher-dimensional space where they become linearly separable (Shmilovici, 2010). Even though SVM has been successfully used in a whole range of applications, it tends not to be the best choice when dealing with imbalanced datasets where the positive instances are rare compared to negative ones (Wu & Chang, 2003). However, when used in combination with balancing techniques like oversampling and undersampling, it can improve its effectiveness significantly by making SVM more sensitive to the positive class (Akbani, Kwek, & Japkowicz, 2004), hence making it a useful algorithm for the injury problem.

### 6.1.4. Artificial Neural Network (ANN)

An ANN uses mathematical models in a way to emulate the interactions of neurons in the brain. It learns from the training data the arbitrary nonlinear input-output connections by automatically adjust layer weights or even the network structure (James et al., 2013). It can be used to solve problems that are not solvable by conventional mathematical processes (Lang, Pitts, Damron, & Rutledge, 1997)

There has been an increasing interest in the use of ANNs in different types of domains. ANN have successfully been used in the past to characterize relationships in muscle activity and kinematic patterns (Hahn, 2007), performance modelling in sports (Silva et al., 2007), to predict relationship between perceived exertion and GPS training-load variable (Bartlett, O'Connor, Pitchford, Torres-Ronda, & Robertson, 2011) and a valid tool for modelling the training process (Shestakov, 2000). They can be used to improve our understanding of the relationship between the player injury and the GPS and self-rating features.

### 6.1.5. Ensemble Models

Two ensemble models were used to reduce bias and variance and overcome the limitations of the small size and unbalanced dataset: Random Forest (bagging) and Extreme Gradient Boosting (Boosting). Ensemble models generate and combine an elevated number of classifiers created on smaller individual subsets of data to produce an (expectantly) better estimator. This combination is made in two ways: bagging (e.g. random forest) and boosting (e.g. XGBoost) (Dietterich, 2000).

Bagging (which stands for Bootstrap Aggregating) algorithm randomly partitions the data into subsets (using bootstrap) and trains the base learners, compute the ensemble and use voting or averaging for the classification or regression, respectively (Dietterich, 2000). A boosting algorithm is sequential, instead of generating different classifiers – as bagging -, boosting turns weak learners into strong ones, by focusing on the examples previously misclassified to build the new classifiers (Friedman, 2001).

The bagging algorithm used is Random Forest, which uses an ensemble of classification trees (Breiman, 2001). Bootstrap is used to divide the data into subsets, build a classification tree and at, each split, uses a random selection of features. Each tree is grown to its full extent to reduce bias and the ensemble of trees then votes for the most popular class (Pal, 2005). It was chosen to use in this project by its predictive performance, not overfitting and incorporating the interactions among features (Díaz-Uriarte & Alvarez de Andrés, 2006).

The boosting algorithm used is Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). XGBoost was chosen for being an efficient implementation of the gradient boosting ensemble (Friedman, 2001), its respectable performance on machine learning competitions (Chen & He, 2015) and its reliant use on other domains (Tamayo et al., 2016; Torlay, Perrone-Bertolotti, Thomas, & Baciu, 2017). The boosted trees on XGBoost can be regression or classification-based trees.

## 6.2. TRAINING CONTROL PARAMETERS

For training control stratified, 4-fold Cross Validation is used. *K*-fold cross-validation is a resampling technique widely used in model validation (James et al., 2013). On this approach, the observations are drawn without replacement into four groups of approximately equal size and an equal proportion of classes. The first fold is used as a validation set, and the classifier is fitted on the remaining three folds. The excluded group is used as a statistically independent test sample for the model. Each group is excluded once. A smaller *k* (*k*=4) was used due to small sample size.

A parameters grid was used for training each model to find the optimal settings. The grid used is designated in Table 10.

Table 10 - Model parameter grid

| Model | Parameters |
|---|---|
| **Naïve Bayes** | - |
| **Linear Discriminant Analysis** | Discriminant Functions ∈ {1,2,3} |
| **Support Vector Machines** | Cost ∈ {0.25; 0.5; 1; 64}; Sigma ∈ {0.0357} |
| **Neural Networks** | Size ∈ {1…40}; Decay ∈ {0…0.1} |
| **Random Forest** | Randomly Selected Predictors ∈ {2…18} |
| **Extreme Gradient Boosting** | Boosting Iterations ∈ {1…30}; L2 Regularization ∈ {0…2}; L1 Regularization ∈ {1…4}; Learning Rate ∈ {0.01; 0.1…1} |

## 6.3. EVALUATION

The most popular and used evaluation metric to assess the performance of a machine learning algorithm is classifier accuracy (Provost, Fawcett, & Kohavi, 1997) which can be considered as the probability of success in identifying the right class of an observation (Maratea, Petrosino, & Manzo, 2014). However, on an unbalanced dataset where 99% of the observations belong to one class and only 1% to target class, accuracy can easily achieve 99% without correctly classifying any of the target examples. So, the conventional approach can produce misleading results (Weng & Poon, 2008).

The other issue with using accuracy as the evaluation metric is the problem with the misclassification costs. Especially on an unbalanced dataset, the target class (rarer examples) is – often – more important than the majority class, so that the cost of misclassing the target class is higher than to misclassify the majority class (Weng & Poon, 2008). In our dataset, misclassifying a muscular injury could have potentially higher adverse effects (as seen in Chapter 1) than misclassifying the majority class. So, we are willing to accept a potentially lower ability to predict the majority class (non-injured), as long as, the classifier accurately predicts the target events (injured players).

Therefore, the measures of recall and precision tend to raise its importance in an unbalanced dataset. In binary classification, it's possible to outline the observations into two classes: negative and positive, which leads to four possible outcomes: True Negatives, False Negatives, True Positives and False Positives. Different metrics are calculated based on these outcomes.

Precision is also known as Positive Predictive Value, can be considered as a measure of the classifier correctness. A low precision can indicate many false positives. It is calculated using the formula below:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

And recall (also known as sensitivity), which can be considered a measure of a classifier thoroughness. A low recall can indicate many false negatives. It's calculated using the formula below:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

These two metrics allow to grasp the relationship between the different classes and to get a more honest evaluation of an unbalanced dataset. A commonly used metric when evaluating unbalanced dataset, which combines recall and precision, is the ROC curve, which plots recall against the False Positive Rate (NOT precision) at various thresholds which allows to directly observe the tradeoff between them (Powers, 2011). The ROC curve, using a dynamic threshold approach, can transmit knowledge about the classifier performance in all possible combinations of the class distributions and the cost misclassification (Drummond & Holte, 2004). However, a possible complication is that, on our unbalanced dataset, the ROC curve can still present an optimistic assessment (Drummond & Holte, 2004) since they decouple the classifier performance from class skewness and the error costs (Fawcett, 2006).

An evaluation measure that can help probe the results of a dataset with a class unbalance is the F-measure (Powers, 2011). The F-measure, originated from the field of Information Retrieval, is calculated using the formula below (Provost et al., 1997):

$$F - Measure = F1 = 2 * \frac{precision * recall}{precision + recall}$$

F-measure is the harmonic mean of precision and recall (Sasaki, 2007). F-measure combines them into a single measure, usually with equal weights on both measures, and is also commonly used on unbalanced datasets (Powers, 2011; Weiss & Hirsh, 2000).

For the performance evaluation of this project, F-Measure is used for each classifier.

# 7. RESULTS AND DISCUSSION

## 7.1. RESULTS (WITHOUT PCA)

The results of the different models and the different sampling techniques (without PCA) are presented in Table 11 and Figure 13.

Table 11 - Training and testing evaluation measures of the models trained with the different datasets

| **Balancing** | **Model** | **F-Measure** | | **Precision** | | **Recall** | |
|---|---|---|---|---|---|---|---|
| | | **Training** | **Testing** | **Training** | **Testing** | **Training** | **Testing** |
| None | NB | 0.11 | 0.1 | 0.06 | 0.56 | 0.5 | 0.67 |
| None | LDA | 0 | 0 | 0 | 0 | 0 | 0 |
| None | SVM | 0 | 0 | 0 | 0 | 0 | 0 |
| None | NN | 0 | 0 | 0 | 0 | 0 | 0 |
| None | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| None | XGBoost | 0 | 0 | 0 | 0 | 0 | 0 |
| Over[r1] | NB | 0.73 | 0.33 | 0.61 | 0.29 | 0.9 | 0.67 |
| Over | LDA | 0.84 | 0.12 | 0.78 | 0.06 | 0.92 | 0.67 |
| Over | SVM | 0.99 | 0 | 0.99 | 0 | 1 | 0 |
| Over | NN | 1 | 0 | 1 | 0 | 1 | 0 |
| Over | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Over | XGBoost | 1 | 0 | 1 | 0 | 1 | 0 |
| Under | NB | 0.23 | 0.08 | 0.14 | 0.043 | 0.6 | 0.67 |
| Under | LDA | 0.14 | 0.12 | 0.25 | 0.07 | 0.1 | 0.33 |
| Under | SVM | 0 | 0 | 0 | 0 | 0 | 0 |
| Under | NN | 1 | 0 | 1 | 0 | 1 | 0 |
| Under | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Under | XGBoost | 0.67 | 0 | 1 | 0 | 0.5 | 0 |
| Both | NB | 0.75 | 0.05 | 0.65 | 0.03 | 0.89 | 0.67 |
| Both | LDA | 0.92 | 0.09 | 0.86 | 0.048 | 1 | 0.67 |
| Both | SVM | 1 | 0 | 1 | 0 | 1 | 0 |
| Both | NN | 1 | 0 | 1 | 0 | 1 | 0 |
| Both | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Both | XGBoost | 1 | 0.12 | 1 | 0.08 | 1 | 0.33 |
| SMOTE | NB | 0.65 | 0.04 | 0.53 | 0.02 | 0.84 | 0.33 |
| SMOTE | LDA | 0.72 | 0.13 | 0.69 | 0.07 | 0.76 | 0.67 |
| SMOTE | SVM | 1 | 0 | 1 | 0 | 1 | 0 |
| SMOTE | NN | 1 | 0.25 | 1 | 0.15 | 1 | 0.67 |
| SMOTE | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| SMOTE | XGBoost | 1 | 0.15 | 1 | 0.1 | 1 | 0.33 |
| ROSE | NB | 0.75 | 0.05 | 0.69 | 0.028 | 0.82 | 0.67 |
| ROSE | LDA | 0.72 | 0.08 | 0.70 | 0.044 | 0.74 | 0.67 |
| ROSE | SVM | 0.97 | 0.1 | 0.96 | 0.056 | 0.98 | 0.67 |
| ROSE | NN | 1 | 0.1 | 1 | 0.05 | 1 | 0.67 |
| ROSE | RF | 1 | 0.08 | 1 | 0.042 | 1 | 0.67 |
| **ROSE** | **XGBoost** | **0.50** | **0.40** | **0.44** | **0.29** | **0.59** | **0.67** |

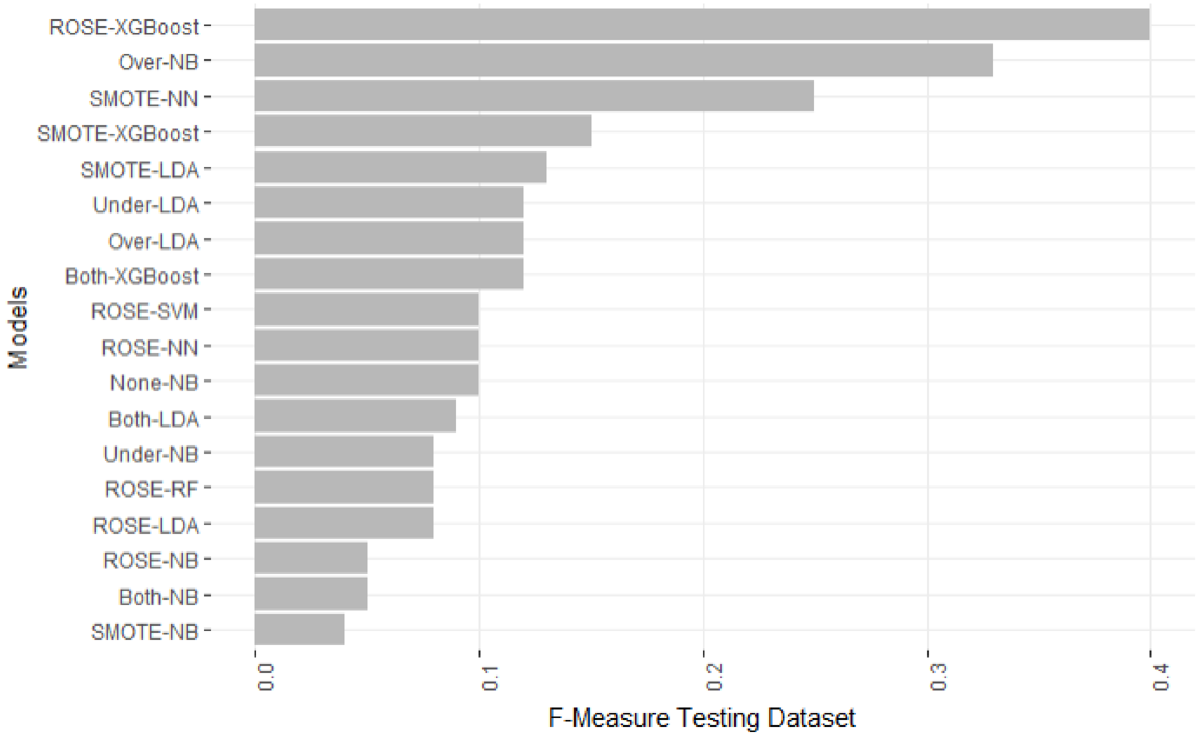

Figure 13 – Plot with F-Measure of the testing dataset (without PCA)

From Table 11, we can detect several results of overfitting to the data particularly on the more complex models like NN and RF where the F-Measure, Precision, and Recall often have the value of 1 on the training set - a perfect fit - yet, on the test set, the results are considerably worse.

It's also possible to observe that the training set without any balancing technique has the worst results with most of the algorithms not being able to make any accurate predictions on the testing dataset. The training sets where we applied algorithms that generate artificial data (SMOTE and ROSE) generally have better results.

It's particularly interesting to detect that a "simpler" algorithm like Naïve Bayes and LDA, often outperform the more complex ones, like NN, RF, and XGBoost in the F-Measure testing set. Such is the case on the no balancing, over and under balancing. Particularly Naïve Bayes has the second highest performance on all the datasets with an F-Measure of 0.33 on the underbalanced testing dataset as we can notice in Figure 13.

The overall problem of most of the used algorithms is in the precision results. The recall achieves a maximum of 0.67 on other ML algorithms, but it's usually at the cost of a lower precision result. So, in general, the algorithms tend to have little discriminatory value since they have too many false positives.

The best result in this table is with the ROSE balancing technique and the XGBoost algorithm (*eta=0.2; lambda=0.2; alpha=4; nrounds=20*) with an F-Measure on the testing dataset of 0.4. This result stands out due to the increased precision (0.29) since the recall (0.67) can also be found on the other algorithms. On this dataset, the ROSE balancing technique outperforms most of the other balancing techniques.

## 7.2. RESULTS (WITH PCA)

The results with PCA application are presented on Table 12 and Figure 14.

Table 12 - Training and testing evaluation measures of the models trained with the different datasets (with PCA).

| **Balancing** | **Model** | **F-Measure** | | **Precision** | | **Recall** | |
|---|---|---|---|---|---|---|---|
| | | **Training** | **Testing** | **Training** | **Testing** | **Training** | **Testing** |
| None | NB | 0.11 | 0.1 | 0.06 | 0.05 | 0.5 | 0.67 |
| None | LDA | 0 | 0 | 0 | 0 | 0 | 0 |
| None | SVM | 0 | 0 | 0 | 0 | 0 | 0 |
| None | NN | 0 | 0 | 0 | 0 | 0 | 0 |
| None | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| None | XGBoost | 0 | 0 | 0 | 0 | 0 | 0 |
| Over | NB | 0.71 | 0.06 | 0.61 | 0.03 | 0.86 | 0.67 |
| Over | LDA | 0.1 | 0 | 0.06 | 0 | 0.4 | 0 |
| Over | SVM | 0.12 | 0 | 0.07 | 0 | 0.4 | 0 |
| Over | NN | 0.18 | 0.06 | 0.1 | 0.03 | 0.9 | 0.33 |
| Over | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Over | XGBoost | 1 | 0 | 1 | 0 | 1 | 0 |
| Under | NB | 0.23 | 0.08 | 0.14 | 0.04 | 0.6 | 0.67 |
| Under | LDA | 0 | 0 | 0 | 0 | 0 | 0 |
| Under | SVM | 0 | 0 | 0 | 0 | 0 | 0 |
| Under | NN | 0 | 0 | 0 | 0 | 0 | 0 |
| Under | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Under | XGBoost | 0 | 0 | 0 | 0 | 0 | 0 |
| Both | NB | 0.75 | 0.05 | 0.65 | 0.02 | 0.89 | 0.67 |
| Both | LDA | 0.07 | 0.04 | 0.04 | 0.02 | 0.5 | 0.33 |
| Both | SVM | 0.09 | 0.04 | 0.05 | 0.02 | 0.5 | 0.33 |
| Both | NN | 0.15 | 0.08 | 0.08 | 0.04 | 1 | 0.67 |
| Both | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| Both | XGBoost | 0.98 | 0.11 | 0.96 | 0.06 | 1 | 0.33 |
| SMOTE | NB | 0.65 | 0.04 | 0.53 | 0.02 | 0.84 | 0.33 |
| SMOTE | LDA | 0.07 | 0 | 0.05 | 0 | 0.1 | 0 |
| SMOTE | SVM | 0 | 0 | 0 | 0 | 0 | 0 |
| SMOTE | NN | 0 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| SMOTE | RF | 1 | 0 | 1 | 0 | 1 | 0 |
| SMOTE | XGBoost | 0.55 | 0 | 0.75 | 0 | 0.43 | 0 |
| ROSE | NB | 0.75 | 0.05 | 0.69 | 0.03 | 0.82 | 0.67 |
| ROSE | LDA | 0.08 | 0.04 | 0.04 | 0.02 | 0.6 | 0.33 |
| ROSE | SVM | 0.09 | 0.04 | 0.05 | 0.02 | 0.5 | 0.33 |
| ROSE | NN | 0.09 | 0.04 | 0.05 | 0.02 | 0.5 | 0.33 |
| ROSE | RF | 1 | 0.08 | 1 | 0.04 | 1 | 0.67 |
| ROSE | XGBoost | 0.5 | 0.22 | 0.72 | 0.13 | 0.38 | 0.67 |

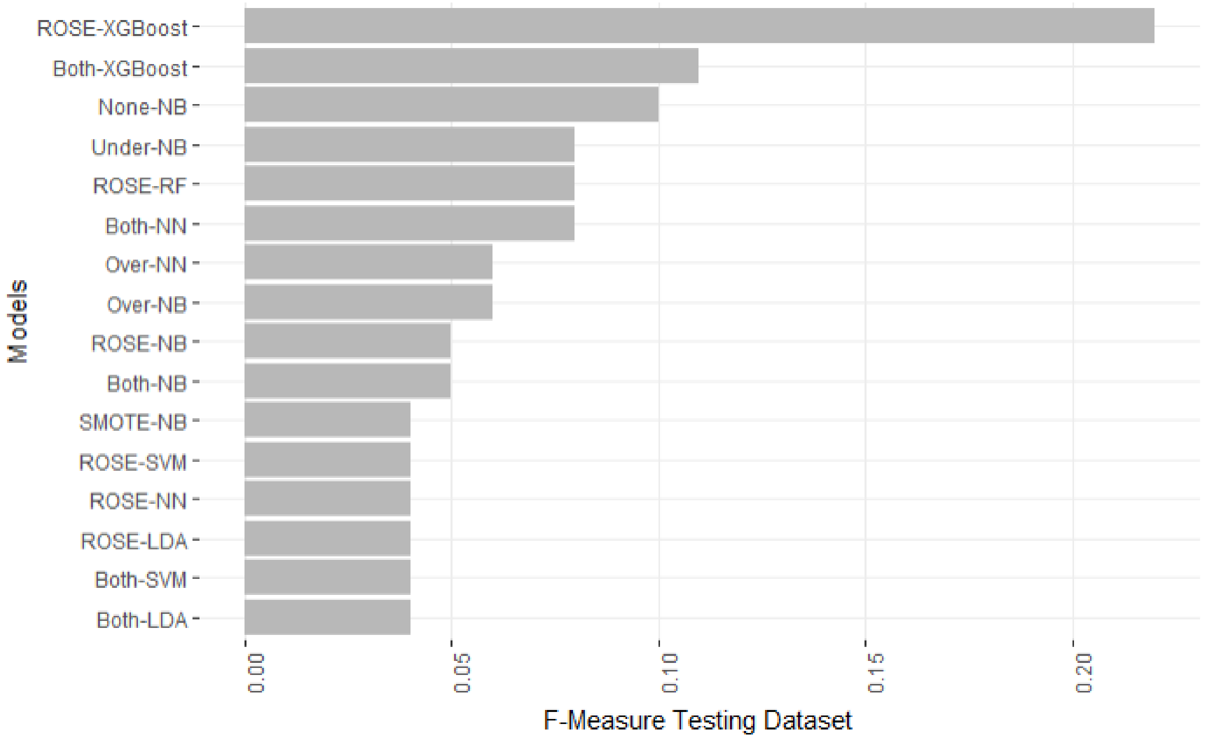

Figure 14 - Plot with F-Measure of the testing dataset (with PCA)

From Table 12, it’s possible to notice overall worse results with the application of PCA on the different datasets.

The patterns detected on Table 11, can also be observed here. The dataset with no balancing technique can’t generate any predictions (except for Naïve Bayes). NN and RF often overfit with training values of 1 and worse values on the testing dataset.

It’s particularly interesting to note that the SMOTE balancing technique only has prediction with Naïve Bayes, not generating any predictions with the other algorithms. Since both ROSE and SMOTE generate artificial data, we would expect better values on both.

Similarly to Table 11, the best model is also XGBoost using the ROSE balancing technique (*eta=0.27; lamba=0.3; alpha=1; nrounds=19*) with an F-Measure of 0.22, precision of 0.13 and recall of 0.67. Naïve Bayes also achieved better results than the remaining algorithms by having the third and fourth best result (Figure 14).

### 7.3. VARIABLE IMPORTANCE

Figure 15 displays the contributions of the different features to the winning XGBoost algorithm. The variable importance plot was gathered using the *"xgb.plot.importance"* function from the xgboost library (Chen & Guestrin, 2016). R code for the winning model and variable importance plot can be consulted on Appendix 2.

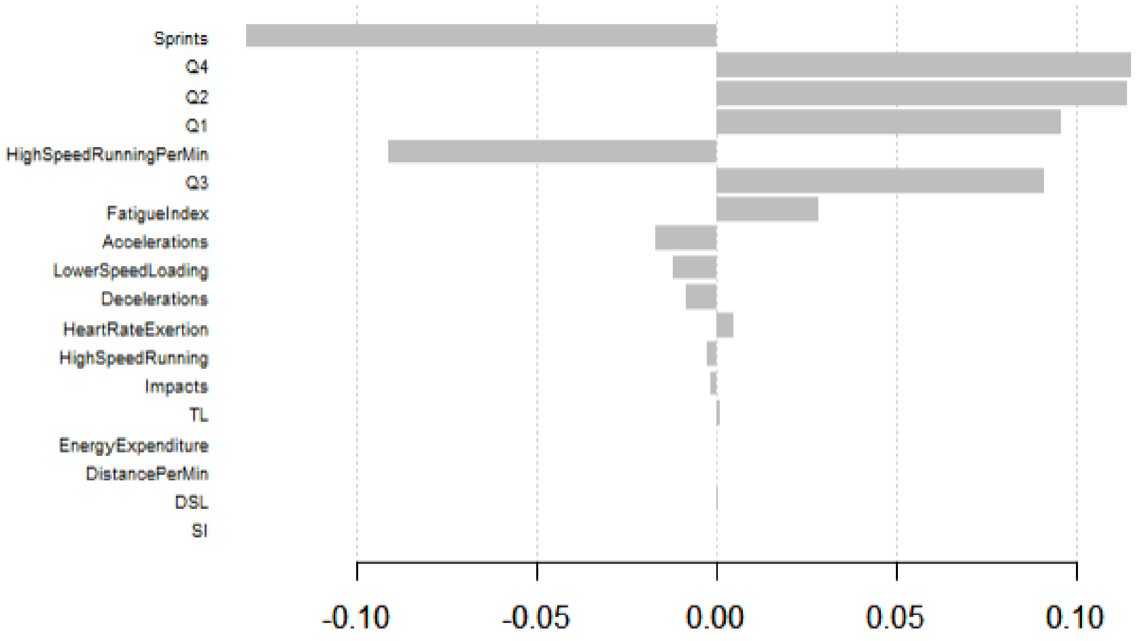

Figure 15 - Variable contribution plot of the winning model

When analyzing the eight most important features of this model, we observe that "Sprints" is the feature with the highest (negative) contribution to the injury prediction model. And especially when combined with "HighSpeedRunningPerMin" and "Accelerations" features, which measure a similar occurrence, we can perceive that lower values on these variables, increase the risk of injury. Players that have higher risks of injury tend to have less high-speed running, sprints and accelerations during training.

We can also observe that the self-ratings features (Q1, Q2, Q3, Q4) have a large contribution to the detection of injuries with Q4 (Stress Level) and Q2 (Sleep Quality) having the largest impact. The higher the values of all these questions, the higher the likelihood the player has of contracting a muscular injury. We can recall from Chapter 3, higher values on these four questions indicates higher subjective tiredness.

The “Fatigue Index” also has a relevant negative contribution to the model. Players with a higher likelihood to get injured will tend to have a higher “Fatigue Index”.

## 8. DISCUSSION AND CONCLUSION

Predicting football injuries is far from being a trivial or minor task. The human body is a complicated organism, and injuries can be caused by an infinite number of causes, some of them, impossible to predict with current methods, as outside agents cause them. Understanding and being able to predict, with reasonable accuracy, a certain type of injuries is a step forward to adjust training loads and training schedules to maximize the likelihood for the coaching team to prevent an injury to take place. Machine learning techniques have been proposed and applied to address this problem with some degree of success.

In this project we used training EPTS data, as well as self-rating subjective data, to create an injury prediction model. The four proposed objectives in Chapter 1 were accomplished.

First, the EDA uncovered missing data in different variables and an unbalanced dataset with the target variable having only 1.9% of the total observations. Secondly, different types of data aggregation and transformations were considered. We decided to aggregate the data by weekly microcycles since it usually contains different training intensities and one match, as well as being a frequently used measure in football. The data was balanced using various sampling techniques and the redundancy was reduced with PCA. Thirdly, different machine learning algorithms were applied and evaluated using F-Measure with 4-fold cross validation. For the final objective, the ROSE sampling technique with the XGBoost algorithm was considered the most accurate model.

The final model can be deployed on any training dashboard to monitor each player's risk of injury. This injury prediction algorithm can be used to determine training load thresholds for individual players, above which, injury risk substantially increases. This could be a practical and useful application of the algorithm for coaches and trainers.

This injury prediction algorithm provides a useful way to combine the different features and swiftly flag the players in the risk of injury. It's particularly revealing that the features with the biggest contribution for the results of this algorithm merge subjective and objective measures.

Any sports training program enforces some amount of stress on athlete, which may shift their performance on a continuum that can evolve to overtraining syndrome, where muscular injuries are more prone to happening (Margonis et al., 2007). The initial stages of overtraining are generally associated with subjective feelings of fatigue and exhaustion but aren't necessarily accompanied by visible decrements of performance. As the player continues with the same training load, these subjective feelings of fatigue become associated with noticeable decreases in performance (Budgett, 1990). It’s interesting to notice that the most relevant objective features of the model are related to a decrease of sprinting frequency and velocity. Sprinting is recognized as being one of the primary mechanisms for hamstring strains, responsible for 57% of all hamstring injuries, accounting for 12-16% of total injuries (Woods et al., 2004).

It’s often difficult for the human brain to observe these different data sources (objective and subjective) and take useful information in a prompt manner, to avoid further navigating down into the overtraining continuum until the danger zone. Hence, the combination of the Hooper’s subjective rating scale, as well as, the relevant objective measures of the EPTS sensor, injury identification can become easier and more accurate as it can provide important information, not only to possibly identify overtraining but also to address the players training.

## 8.1. LIMITATIONS AND FUTURE WORK

However, this study has important limitations that can be addressed in future articles. The first limitation is evident. The data from only one season and one team is – unsurprisingly - too small since the number of injuries in one season, especially considering only muscular injuries, tend to be reduced. Training a machine learning model with an unbalanced and small dataset creates a less reliable model. Combining the data of different teams and/or different seasons can be a possibility to achieve a more accurate model.

The second limitation is only using data from one football team. The generalization of the results can questioned. Data from other professional football teams is needed to be able to generalize the results.

The data sources used in this study can also have many other potential uses within the machine learning domain. They can be a priceless source of performance information that is available near real time for coaches and support staff, which, despite access to this information, tend not to utilize its full potential without decision support models that automatically evaluate the dozens of data features of the whole team (Ofoghi et al., 2013).

This opens a fruitful path for machine learning applications which can be used not only for injury prediction but also for many other different football relevant purposes. It could be adopted to identify different activity profiles by position or competition level, as well as to evaluate performance in game specific tasks. It can be used even earlier to identify talented players based on their psychological characteristics and practice history, as it was done in an earlier research (Gonçalves, e Silva, Carvalho, & Gonçalves, 2011) and can be expanded to different sports. Although, the data on this project was used for the prediction of injuries, it can also be employed to help predict the time for a player to return to play after an injury.

Nonetheless, regarding future work in injury prediction, three objective suggestions are proposed:

1) A possible adaptation to other sports can be tested. Many other team sports have the same problem of overtraining and muscular fatigue and use a similar type of sensors in their training; hence, machine learning prediction could be applied the same way.
2) Using different sliding windows for injury prediction. On this project, a sliding window of one week (or one microcycle) was used to preserve the natural training cycles of the team, but different sliding windows can be used (2 days up to two weeks) and may provide better results.
3) Adding more features to the model. This project used sensor data and self-ratings, but other features can be added as more data is gathered from the players, such as age, nutrition, sleep, among others. Combining the data from matches could be particularly useful considering that matches can be stressful and strenuous for the players and it can be considered a risk factor for injuries (Ekstrand et al., 2011).

## REFERENCES

Agarwal, S. (2013). Data Mining: Data Mining Concepts and Techniques. In *2013 International Conference on Machine Intelligence and Research Advancement* (pp. 203–207). IEEE. https://doi.org/10.1109/ICMIRA.2013.45

Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying Support Vector Machine to Imbalanced Datasets. *Machine Learning: ECML 2004*, 39–50. https://doi.org/10.1.1.102.5233

Alderson, J. (2015). A markerless motion capture technique for sport performance analysis and injury prevention: Toward a “big data”, machine learning future. *Journal of Science and Medicine in Sport*, *19*, e79. https://doi.org/10.1016/j.jsams.2015.12.192

Arnason, A., Sigurdsson, S. B., Gudmundsson, A., Holme, I., Engebretsen, L., & Bahr, R. (2004). Physical fitness, injuries, and team performance in soccer. *Medicine and Science in Sports and Exercise*, *36*(2), 278–85. https://doi.org/10.1249/01.MSS.0000113478.92945.CA

Aughey, R. J. (2011). Applications of GPS Technologies to Field Sports. *International Journal of Sports Physiology and Performance*, *6*(3), 295–310. https://doi.org/10.1123/ijspp.6.3.295

Bartlett, J. D., O’Connor, F., Pitchford, N., Torres-Ronda, L., & Robertson, S. (2011). Relationships Between Internal and External Training Load in Team Sport Athletes: Evidence for an Individualised Approach. *International Journal of Sport Nutrition and Exercise Metabolism*, *32*, 1–44. https://doi.org/10.1123/ijspp.2015-0012

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explor. Newsl.*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Berge, J. M. F., & Kiers, H. A. L. (1996). Optimality criteria for principal component analysis and generalizations. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 335–345. https://doi.org/10.1111/j.2044-8317.1996.tb01092.x

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*, *1*, 121–125. https://doi.org/10.1023/A:1009782106822

Bishop, C. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4), 49901. https://doi.org/10.1117/1.2819119

Bittencourt, N. F. N., Meeuwisse, W. H., Mendonça, L. D., Nettel-Aguirre, A., Ocarino, J. M., & Fonseca, S. T. (2016). Complex systems approach for sports injuries: moving from risk factor identification to injury pattern recognition—narrative review and new concept. *British Journal of Sports Medicine*, *50*(21), 1309–1314. https://doi.org/10.1136/bjsports-2015-095850

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Budgett, R. (1990). Overtraining syndrome. *British Journal of Sports Medicine*, *24*(4), 231–236. https://doi.org/10.1136/bjsm.24.4.231

Casals, M., & Finch, C. F. (2016). Sports Biostatistician: a critical member of all sports science and medicine teams for injury prevention. *Injury Prevention*, injuryprev-2016-042211. https://doi.org/10.1136/injuryprev-2016-042211

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76. https://doi.org/10.1109/ICETET.2008.239

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2015). Higgs Boson Discovery with Boosted Trees. *JMLR: Workshop and Conference Proceedings*, *42*(May 2014), 69–80.

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). xgboost: Extreme Gradient Boosting. Retrieved from https://cran.r-project.org/package=xgboost

Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., & Malvaldi, M. (2015). The harsh rule of the goals: Data-driven performance indicators for football teams. *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*. https://doi.org/10.1109/DSAA.2015.7344823

Colby, M. J., Dawson, B., Heasman, J., Rogalski, B., & Gabbett, T. J. (2014). Accelerometer and GPS-Derived Running Loads and Injury Risk in Elite Australian Footballers. *Journal of Strength and Conditioning Research*, *28*(8), 2244–2252. https://doi.org/10.1519/JSC.0000000000000362

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Davenport, Thomas H.; Harris, J. (2013). Competing on Analytics: The New Science of Winning Davenport. *Journal of Information Technology Case and Application Research*, *15*(4), 59–61. https://doi.org/10.1080/15228053.2013.10845729

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 3. https://doi.org/10.1186/1471-2105-7-3

Dick, R., Ferrara, M. S., Agel, J., Courson, R., Marshall, S. W., Hanley, M. J., & Reifsteck, F. (2007). Descriptive epidemiology of collegiate men's football injuries: National Collegiate Athletic Association Injury Surveillance System, 1988-1989 through 2003-2004. *Journal of Athletic Training*, *42*(2), 221–33. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17710170

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Drummond, C., & Holte, R. C. (2004). What ROC Curves Can't Do (and Cost Curves Can). *Workshop ROC Analysis in AI: ROCAI*, 19–26. https://doi.org/10.1007/s10994-006-8199-5

Dvorak, J., & Junge, a. (2000). Football injuries and physical symptoms. A review of the literature. *The American Journal of Sports Medicine*, *28*(5), S3–S9. https://doi.org/10.1177/28.suppl_5.S-3

Ekstrand, J., Hägglund, M., & Waldén, M. (2011). Epidemiology of Muscle Injuries in Professional Football (Soccer). *The American Journal of Sports Medicine*, *39*(6), 1226–1232.

https://doi.org/10.1177/0363546510395879

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fisher, R. A. (1936). The Use of Multiple Measurements In Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/DOI 10.1214/aos/1013203451

Fry, M. J., & Ohlmann, J. W. (2012). Introduction to the special issue on analytics in sports, part I: General sports applications. *Interfaces*, *42*(2), 105–108. https://doi.org/10.1287/inte.1120.0633

Gonçalves, C. E., e Silva, M. J. C., Carvalho, H. M., & Gonçalves, Â. (2011). Why do they engage in such hard programs? the search for excellence in youth basketball. *Journal of Sports Science and Medicine*, *10*(3), 458–464.

Haddad, M., Padulo, J., & Chamari, K. (2014). The Usefulness of Session Rating of Perceived Exertion for Monitoring Training Load Despite Several Influences on Perceived Exertion. *International Journal of Sports Physiology and Performance*, *9*(5), 882–883. https://doi.org/10.1123/ijspp.2014-0010

Hägglund, M., Waldén, M., & Ekstrand, J. (2005). Injury incidence and distribution in elite football - A prospective study of the Danish and the Swedish top divisions. *Scandinavian Journal of Medicine and Science in Sports*, *15*(1), 21–28. https://doi.org/10.1111/j.1600-0838.2004.00395.x

Hägglund, 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), 738–42. https://doi.org/10.1136/bjsports-2013-092215

Hahn, M. E. (2007). Feasibility of estimating isokinetic knee torque using a neural network model. *Journal of Biomechanics*, *40*(5), 1107–1114. https://doi.org/10.1016/j.jbiomech.2006.04.014

Hilden, J. (1984). Statistical diagnosis based on conditional independence does not require it. *Computers in Biology and Medicine*, *14*(4), 429–435. https://doi.org/10.1016/0010-4825(84)90043-X

Hooper, S. L., & Mackinnon, L. T. (1995). Monitoring Overtraining in Athletes: Recommendations. *Sports Medicine*, *20*(5), 321–327. https://doi.org/10.2165/00007256-199520050-00003

Izenman, A. J. (2013). Linear Discriminant Analysis. In *Modern Multivariate Statistical Techniques* (pp. 237–280). Springer New York. https://doi.org/10.1007/978-0-387-78189-1_8

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., … Hunt., T. (2017). caret: Classification and Regression Training. Retrieved from https://cran.r-project.org/package=caret

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kampakis, S. (2016). Predictive modelling of football injuries, (April). Retrieved from http://arxiv.org/abs/1609.07480

Kelley, D. J., Mureika, J. R., & Phillips, J. A. (2006). Predicting Baseball Home Run Records Using Exponential Frequency Distributions. *arXiv Preprint Physics*. Retrieved from http://arxiv.org/abs/physics/0608228

Kellmann, M. (2010). Preventing overtraining in athletes in high-intensity sports and stress/recovery monitoring. *Scandinavian Journal of Medicine & Science in Sports*, *20*(SUPPL. 3), 95–102. https://doi.org/10.1111/j.1600-0838.2010.01192.x

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. *Science*, *30*(1), 25–36. https://doi.org/10.1007/978-0-387-09823-4_45

Kowarik, A., & Templ, M. (2016). Imputation with the {R} Package {VIM}. *Journal of Statistical Software*, *74*(7), 1–16. https://doi.org/10.18637/jss.v074.i07

Kuhn, M. (2017). Caret package. *Journal of Statistical Software*, *28*(5).

Lang, E. W., Pitts, L. H., Damron, S. L., & Rutledge, R. (1997). Outcome after severe head injury: An analysis of prediction based upon comparison of neural network versus logistic regression analysis. *Neurological Research*, *19*(3), 274–280. https://doi.org/10.1080/01616412.1997.11740813

Leddy, M. H., Lambert, M. J., & Ogles, B. M. (1994). Psychological consequences of athletic injury among high-level competitors. *Research Quarterly for Exercise and Sport*, *65*(4), 347–354. https://doi.org/10.1080/02701367.1994.10607639

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. New York and London: Norton. https://doi.org/10.1002/mde.1220

Lunardon, N., Menardi, G., & Torelli, N. (2014). {ROSE}: a {P}ackage for {B}inary {I}mbalanced {L}earning. *{R} Journal*, *6*(1), 82–92.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, *257*, 331–341. https://doi.org/10.1016/j.ins.2013.04.016

Margonis, K., Fatouros, I. G., Jamurtas, A. Z., Nikolaidis, M. G., Douroudos, I., Chatzinikolaou, A., … Kouretas, D. (2007). Oxidative stress biomarkers responses to physical overtraining: Implications for diagnosis. *Free Radical Biology and Medicine*, *43*(6), 901–910. https://doi.org/10.1016/j.freeradbiomed.2007.05.022

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). On the development of a soccer player performance rating system for the english Premier League. *Interfaces*, *42*(4), 339–351. https://doi.org/10.1287/inte.1110.0589

Menardi, G., & Torelli, N. (2014). *Training and assessing classification rules with imbalanced data*. *Data Mining and Knowledge Discovery* (Vol. 28). https://doi.org/10.1007/s10618-012-0295-5

Miller, T. W. (2015). *Sports Analytics and Data Science Winning the Game with Methods and Models*. FT Press.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data Mining in Elite Sports: A Review and a Framework. *Measurement in Physical Education and Exercise Science*, *17*(3), 171–186. https://doi.org/10.1080/1091367X.2013.805137

Öztürk, S. (2013). What is the economic burden of sports injuries? *Joint Diseases and Related Surgery*, *24*(2), 108–111. https://doi.org/10.5606/ehc.2013.24

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pfirrmann, 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. *Journal of Athletic Training*, *51*(5), 410–424. https://doi.org/10.4085/1062-6050-51.6.03

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. https://doi.org/10.1.1.214.9232

Pozzolo, A. D., Caelen, O., & Bontempi, G. (2015). unbalanced: Racing for Unbalanced Methods Selection. Retrieved from https://cran.r-project.org/package=unbalanced

Provost, F., Fawcett, T., & Kohavi, R. (1997). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning*, 445–453. https://doi.org/10.1.1.49.5218

Putukian, M. (2016). The psychological response to injury in student athletes: A narrative review with a focus on mental health. *British Journal of Sports Medicine*, *50*(3), 145–148. https://doi.org/10.1136/bjsports-2015-095586

Quatman, C. E., Quatman, C. C., & Hewett, T. E. (2009). Prediction and prevention of musculoskeletal injury: a paradigm shift in methodology. *British Journal of Sports Medicine*, *43*(14), 1100–1107. https://doi.org/10.1136/bjsm.2009.065482

R Development Core Team. (2011). R Language Definition. *Web*, *0*, 62. https://doi.org/10.1016/0164-1212(87)90019-7

Revelle, W. (2017). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych

Rish, I. (2001). An empirical study of the naive Bayes classifier. *Empirical Methods in Artificial Intelligence Workshop, IJCAI*, *22230*(JANUARY 2001), 41–46. https://doi.org/10.1039/b104835j

Rotshtein, A. P., Posner, M., & Rakityanskaya, A. B. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, *41*(4), 619–630. https://doi.org/10.1007/s10559-005-0098-4

RStudio team, & RStudio. (2015). RStudio: Integrated development environment for R. *The Journal of Wildlife Management*, *75*(8), 1. https://doi.org/10.1002/jwmg.232

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1–5. Retrieved from http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). *Sports Data Mining*. *Information Systems Journal* (Vol. 26). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-6730-5

Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics*, *6*(3), 443–454.

https://doi.org/10.1007/s13042-013-0226-9

Shestakov, M. . (2000). Artificial intelligence in sport science of the 21st century. *Theory and Practice of Physical Education*, (7), 8–14.

Shmilovici, A. (2010). Data Mining and Knowledge Discovery Handbook. In *Data Mining and Knowledge Discovery Handbook* (pp. 231–247). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4

Silva, A. J., Costa, A. M., Oliveira, P. M., Reis, V. M., Saavedra, J., Perl, J., … Marinho, D. A. (2007). The use of neural network technology to model swimming performance. *Journal of Sports Science & Medicine*, *6*(1), 117–25. https://doi.org/Article

Simjanovic, M., Hooper, S., Leveritt, M., Kellmann, M., & Rynne, S. (2009). The use and perceived effectiveness of recovery modalities and monitoring techniques in elite sport. *Journal of Science and Medicine in Sport*, *12*, S22. https://doi.org/10.1016/j.jsams.2008.12.057

Stubbe, J. H., Van Beijsterveldt, A. M. M. C., Van Der Knaap, S., Stege, J., Verhagen, E. A., Van Mechelen, W., & Backx, F. J. G. (2015). Injuries in professional male soccer players in the Netherlands: A prospective cohort study. *Journal of Athletic Training*, *50*(2), 211–216. https://doi.org/10.4085/1062-6050-49.3.64

Talukder, H., Vincent, T., Foster, G., Hu, C., Huerta, J., Kumar, A., … Simpson, S. (2016). Preventing in-game injuries for NBA players Paper ID : 1590. *MIT Sloan Sports Analytics Conference*, *2015*, 1–13.

Tamayo, D., Silburt, A., Valencia, D., Menou, K., Ali-Dib, M., Petrovich, C., … Murray, N. (2016). A Machine Learns to Predict the Stability of Tightly Packed Planetary Systems, 1–7. https://doi.org/10.3847/2041-8205/832/2/L22

Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, *4*(3), 159–169. https://doi.org/10.1007/s40708-017-0065-7

Wei, T., & Simko, V. (2017). R package “corrplot”: Visualization of a Correlation Matrix. Retrieved from https://github.com/taiyun/corrplot

Weiss, G. M., & Hirsh, H. (2000). Learning to predict extremely rare events. *AAAI Workshop on Learning from Imbalanced Data Sets*, 00–05. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Learning+to+Predict+Extremely+Rare+Events

Weng, C. G., & Poon, J. (2008). A new evaluation measure for imbalanced datasets. *Conferences in Research and Practice in Information Technology Series*, *87*, 27–32. https://doi.org/xd AQaqqqAAAAAAAAAAAAE4E44

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). dplyr: A Grammar of Data Manipulation. Retrieved from https://cran.r-project.org/package=dplyr

Witten, I. H., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. https://doi.org/0120884070, 9780120884070

Woods, C., Hawkins, R. D., Maltby, S., Hulse, M., Thomas, A., & Hodson, A. (2004). The Football Association Medical Research Programme: an audit of injuries in professional football—analysis of hamstring injuries. *British Journal of Sports Medicine*, *38*(1), 36 LP-41. Retrieved from http://bjsm.bmj.com/content/38/1/36.abstract

Wu, G., & Chang, E. Y. E. (2003). Class-boundary alignment for imbalanced dataset learning. *The Twentieth International Conference on Machine Learning (ICML), Workshop on Imbalanced Data Sets*, (1), 49–56. Retrieved from http://www.site.uottawa.ca/~nat/Workshop2003/Wu-final.pdf

Zelic, I., Kononenko, I., Lavra, N., & Vuga, V. (1997). Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. *Journal of Medical Systems*, *21*(6), 429–444. https://doi.org/10.1023/A:1022880431298

Zelic, I., Kononenko, I., Lavrac, N., & Vuga, V. (1997). Diagnosis of sport injuries with machine learning: explanation of induced decisions. *Proceedings of Computer Based Medical Systems*, 195–199. https://doi.org/10.1109/CBMS.1997.596433

## APPENDIX 1 – R CODE FOR BALANCING THE DATASET

R Code for using the different dataset balancing methods using the packages: “unbalanced” (Pozzolo, Caelen, & Bontempi, 2015) and ROSE(Lunardon et al., 2014).

```
library(unbalanced)
library(ROSE)

#Oversampling - Randomly replicate instances in the minority class
df.balanced.over <- ovun.sample(injured ~ ., data = training, method =
                    "over",N = 750, sedd=456)$data
table(df.balanced.over$injured)


#Undersampling - Randomly remove instances in the majority class
df.balanced.under <- ovun.sample(injured ~ ., data = training, method =
                     "under", N = 160, seed = 456)$data
table(df.balanced.under$injured)


#Using Both Methods
df.balanced.both <- ovun.sample(injured ~ ., data = training, method =
                    "both", p=0.5, N=410, seed = 456)$data
table(df.balanced.both$injured)


#Using Artificial Data with ROSE
df.balanced.rose <- ROSE(injured ~ ., data = training, seed = 456)$data
table(df.balanced.rose$injured)


#Using SMOTE

data<-ubBalance(X=training[,-19], Y=training$injured, positive="Y",
      type="ubSMOTE", percOver=1550, percUnder=320, verbose=TRUE)

df.balanced.smote<-cbind(data$X,data$Y)
colnames(df.balanced.smote)[colnames(df.balanced.smote)=="data$Y"] <-
"injured"
table(df.balanced.smote$injured)
```

## APPENDIX 2 – R CODE FOR THE WINNING MODEL

R Code for the winning model using the packages: “xgboost” (Chen, He, Benesty, Khotilovich, & Tang, 2018) and “caret” (Jed Wing et al., 2017).

```
library(xgboost)

levels(df.training$injured) <- c(0, 1)
levels(df.testing$injured) <- c(0, 1)

xgb.training <- as.matrix(df.training[,-19])
xgb.label <- as.matrix(df.training$injured)

xgb.model <- xgboost(data = xgb.training,
 label = xgb.label,
 booster = "gblinear",
 eta = 0.2,
 lambda=0.2,
 alpha=4,
 seed = 456,
 eval_metric = "F",
 objective = "binary:logistic",
 nrounds = 20
)

xgb.predict <- predict(xgb.model, data.matrix(df.testing))

xgb.cm <- confusionMatrix(xgb.predict, df.testing$injured, positive="1")

importance_matrix <- xgb.importance(feature_names = colnames(df.training[,-
                     19]), model = xgb.model)

xgb.plot.importance(importance_matrix = importance_matrix)
```