A Preventive Model for Muscle Injuries:
A Novel Approach based on Learning Algorithms

Alejandro López-Valenciano[1], Francisco Ayala[1], José Miguel Puerta[2], Mark De Ste Croix[3], Francisco Vera-García[1], Sergio Hernández-Sánchez[4], Iñaki Ruiz-Pérez [1], and Gregory Myer[5]

[1]Sports Research Centre, Miguel Hernandez University of Elche, Alicante, Spain; [2]Department of Computer Systems, University of Castilla-La Mancha, Albacete, Spain; [3]School of Physical Education, Faculty of Sport, Health and Social Care, University of Gloucestershire, Gloucester, United Kingdom; [4]Department of Pathology and Surgery, Physiotherapy Area, Miguel Hernandez University of Elche, Alicante, Spain; [5]Division of Sports Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

Accepted for Publication: 14 December 2017

Medicine & Science in Sports & Exercise,
DOI: 10.1249/MSS.0000000000001535

A Preventive Model for Muscle Injuries: A Novel Approach based on Learning Algorithms

Alejandro López-Valenciano[1], Francisco Ayala[1], José Miguel Puerta[2], Mark De Ste Croix[3], Francisco Vera-García[1], Sergio Hernández-Sánchez[4], Iñaki Ruiz-Pérez [1], and Gregory Myer[5]

[1]Sports Research Centre, Miguel Hernandez University of Elche, Alicante, Spain; [2]Department of Computer Systems, University of Castilla-La Mancha, Albacete, Spain; [3]School of Physical Education, Faculty of Sport, Health and Social Care, University of Gloucestershire, Gloucester, United Kingdom; [4]Department of Pathology and Surgery, Physiotherapy Area, Miguel Hernandez University of Elche, Alicante, Spain; [5]Division of Sports Medicine, Cincinnati Children‘s Hospital Medical Center, Cincinnati, OH

Corresponding Author. Francisco Ayala. Sports Research Centre, Miguel Hernandez University of Elche. Avda. de la Universidad s/n. 03202 Elche, Alicante, Spain. Email address: fayala@umh.es, Fax: +0034965222409.

Alejandro López-Valenciano were supported by predoctoral grant given by Ministerio de Educación, Cultura y Deporte (FPU) from Spain. We certify that no party having a direct interest in the results of the research supporting this article has or will confer a benefit on us or on any organization which we are associated, do not constitute endorsement by ACSM and they are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

ABSTRACT

Introduction: The application of contemporary statistical approaches coming from Machine Learning and Data Mining environments to build more robust predictive models to identify athletes at high risk of injury might support injury prevention strategies of the future.

Purpose: The purpose was to analyse and compare the behaviour of numerous machine learning methods in order to select the best performing injury risk factor model to identify athlete at risk of lower extremity muscle injuries (MUS$_{INJ}$).

Methods: A total of 132 male professional soccer and handball players underwent a pre-season screening evaluation which included personal, psychological and neuromuscular measures. Furthermore, injury surveillance was employed to capture all the MUS$_{INJ}$ occurring in the 2013/2014 seasons. The predictive ability of several models built by applying a range of learning techniques were analysed and compared.

Results: There were 32 MUS$_{INJ}$ over the follow up period, 21 (65.6%) of which corresponded to the hamstrings, three to the quadriceps (9.3%), four to the adductors (12.5%) and four to the triceps surae (12.5%). A total of 13 injures occurred during training and 19 during competition. Three players were injured twice during the observation period so the first injury was used leaving 29 MUS$_{INJ}$ that were used to develop the predictive models. The model generated by the SmooteBoost technique with a cost-sensitive ADTree as the base classifier reported the best evaluation criteria (area under the receiver operating characteristic curve score = 0.747, true positive rate = 65.9%, true negative rate = 79.1) and hence was considered the best for predicting MUS$_{INJ}$.

Conclusions: The prediction model showed moderate accuracy for identifying professional soccer and handball players at risk of MUS$_{INJ}$. Therefore, the model developed might help in the decision-making process for injury prevention.

Keywords: injury prevention, machine learning techniques, modelling, screening, soccer

INTRODUCTION

Lower extremity muscle injuries (MUS_{INJ}) are very common in professional sports, such as soccer (1), rugby (2) and handball (3). These sports require sudden acceleration and deceleration tasks with rapid changes of directions (4), as well as many situations in which players are required to repetitively kick a ball (5) and/or to be involved in tackling to keep possession of or to win the ball (6). Data have demonstrated that a typical professional soccer team with a 25-player squad could expect 15 MUS_{INJ} each season and MUS_{INJ} can account for more than a quarter of all lost time from injuries (1). In particular, injuries to four major muscle groups of the lower extremity (i.e. adductors, hamstrings, quadriceps, and triceps surae) comprise more than 90% of all MUS_{INJ} in soccer (1). Therefore, there is a clear necessity to develop and implement strategies aimed at preventing and reducing the number and severity of MUS_{INJ} in professional athletes.

Prior to establishing MUS_{INJ} prevention programmes, it is essential to identify athletes at high risk of MUS_{INJ} through a validated screening programme (7). Bahr (7), in a recently published thought-provoking critical review, suggested that prior to considering a screening programme as valid to predict and prevent sports injuries it should have successfully overcome three steps. The first step is to identify those potential risk factors that have demonstrated a strong relationship with injury in prospective studies and then define appropriate cut-off values. The second step is to determine the validity of the screening tests used to measure the risk factors to predict new injuries in a new athlete population. Finally, in the third step studies should document that an intervention programme targeting athletes identified as being at high risk, using the developed screen, must be more beneficial than the same intervention programme given to all athletes.

In recent years, a substantive effort has been made by the scientific community and medical practitioners to identify strong risk factors associated with the occurrence of muscle injuries. Thus, some prospective studies, but not all, have identified previous injury (8-10), older age (8, 10, 11), poor flexibility (8, 11, 12), fatigue (13) and decreased muscle strength or strength imbalances (4, 9, 12) as potential risk factors associated with MUS_{INJ}. Despite the fact that significant associations (causal relationship) were found between these risk factors and MUS_{INJ}, the ability of the cut-off scores proposed to predict injuries are not acceptable for screening purposes. In particular, most of the cut-off scores reported in previous studies show good true negative rates (e.g. how many individuals with a negative score were not injured), however the true positive rates were very low (e.g. how many individuals with a positive score were injured). The consequence of this has led Bahr (7) to conclude that: a) finding statistically significant associations between a test result and MUS_{INJ} is not sufficient evidence to use the test to predict who is at risk of injury; and b) there is no screening test available to predict sports injuries (including MUS_{INJ}) with adequate test properties and consequently the exercises included in intervention programmes are not evidence-based or supported as the link between risk factors and injury incidence remains to be established.

Perhaps one the main reasons behind the lack of available valid screening programmes to predict athletes at high risk of suffering a sport injury, including MUS_{INJ}, could be based on the use of statistical approaches, that in contrast to certain supervised learning algorithms (i.e. ensemble, class balance and cost-sensitive learning techniques), have not been specifically designed to deal with class imbalance problems, such as the MUS_{INJ} phenomenon, in which the number of injured players (minority class) prospectively reported is always much lower than the non-injured players (majority class) (14). Thus, in many scenarios including MUS_{INJ}, traditional multivariate analyses are often biased (for many reason) towards the majority class (known as the “negative” class) and therefore, there is a higher misclassification rate for the minority class instances (called the “positive” examples), which represent the most important concept (15). Another reason for the limited validation of screening programs might be due to the fact that most of the available studies have analysed the predictive ability of each risk factor in isolation or in conjunction with just two or three risk factors. However, the MUS_{INJ} phenomenon has been considered as being multifactorial, in which several factors have an influence on it, and in some cases interact among themselves (17). Therefore, it might be possible that the individual ability of each potential risk factor to impact on the likelihood to suffer a MUS_{INJ} could be very small and in most cases not statistically significant unless analysed in conjunction with other known factors simultaneously, as a complex component or factor.

The application of contemporary statistical approaches (e.g. supervised learning algorithms) coming from Machine Learning and Data Mining environments have been specifically designed to deal with class imbalance problems (14) and can manage a large number of variables in order to develop a robust predictive model, it might shed new light on this problematic area in sport medicine setting. In fact, these statistical approaches have been applied, among others, in several medical diagnosis studies reporting excellent results (18).

Therefore, the main purpose of the current prospective study was to analyse and compare the behaviour of some learning methods in order to select the best performing injury risk factor model to predict MUS_{INJ} in a cohort of professional athletes.

METHOD

Participants

A total of 132 male professional soccer (n = 98) and handball (n = 34) players took part in the current study. Soccer players were recruited from four different soccer teams that were engaged in the 1st (one team, n = 25) and 2nd B (three teams, n = 73) Spanish National

Soccer League divisions. Handball players were recruited from three different handball teams that were engaged in the 1st (one team, n = 11) and 3rd (two teams, n = 23) National Handball League divisions. The sample was homogeneous in potential confounding variables, such as body mass, stature, age, training regime (one game and 4–6 days of training per week), climatic conditions, level of play, resting periods and sport experience (at least 8 years).

Although football and handball are two team sports with different rules and physical demands, both have in common a high incidence rate of MUS_{INJ} associated with acute non-contact incidents (injuries with sudden onset and known cause) (1, 3). Bahr and Holme (19) stated that for prospective studies aimed at investigating potential risk factors for sports injury, a minimum of 20-50 injury cases should be recorded to detect moderate to strong associations. Therefore, 132 professional football and handball players were recruited to ensure that the appropriate number of MUS_{INJ} might be recorded, even with some attrition. Furthermore, another rationale behind the recruitment of players coming from two different sports was to carry out a preliminary exploration regarding the relevance of the feature sport as a personal or individual risk factor on the final predictive model selected. For example, the feature sport might be considered as relevant if it appears as a father node in the final model of a single decision tree structure or as a father or child node in numerous trees where the final model is based on a multiple decision trees structure (i.e. multiple classifiers).

The exclusion criteria were: a) presence of orthopaedic problems that prevented the proper execution of one or more of the neuromuscular tests selected for this study; and b) players who were transferred to other clubs and did not finish the 9-month follow up period. Only primary injuries we used for any player sustaining multiple $MUS_{INJ.}$

Prior to study participation, experimental procedures and potential risks were fully explained to the participants in verbal and written form, and written informed consent was obtained from them. An Institutional Research Ethics committee approved the study protocol prior to data collection, conforming to the recommendations of the Declaration of Helsinki.

Study design

A prospective cohort design was used to address the purposes of this study. In particular, all the MUS_{INJ} accounted for within the 9 months (2013/2014 season) following the initial testing session were prospectively collected for all players.

Players underwent a pre-season evaluation of a number of personal, psychological and neuromuscular measures, most of them considered potential sport-related injury risk factors.

For each soccer and handball team, the testing session was conducted at the pre-season phase of the year.

Testing procedure

The testing session had a total duration of approximately 120 min and was divided into three different parts (see Figure, Supplemental Digital Content 1, Graphical representation of testing procedure, http://links.lww.com/MSS/B167). The first part of the test session was used to obtain information related to the participants' personal or individual characteristics (5 min). The second part was designed to assess psychological measures related to sleep quality and athlete burnout (10 min). Finally, the third part of the session was used to assess a number of neuromuscular measures (105 min).

Each of the 8 testers who took part in this study conducted the same tests throughout all the testing sessions and they were blinded to the purposes of this study. All testers had more than 4 years of experience in neuromuscular assessment.

Personal or individual risk factors

The ad hoc questionnaire designed by Olmedilla, Laguna and Redondo (20) was used to record personal or individual features that have been defined as potential non-modifiable risk factors for sport injuries. Through this questionnaire sport-related background (sport, player position, current level of play, dominant leg [defined as the participant´s kicking leg]) and demographic (age, body mass, stature and body mass index) features were recorded. In addition, the presence within the last season (yes or no) of MUS_{INJ} with a total time taken to resume full training and competition > 8 days was also recorded (self-reported; see Table, Supplemental Digital Content 2, Personal injury risk factors recorded, http://links.lww.com/MSS/B168).

Psychological risk factors

Sleep quality and athlete burnout variables were measured through two validated and worldwide used likert scales. The Spanish version of the Pittsburgh Sleep Diary (21) was used to measure the sleep quality of the soccer and handball players. The final score of this scale was determined as the average of the scores obtained in each of its 7 items.

The Spanish version of the Athlete Burnout Questionnaire (22) was used to assess the three different dimensions that comprise athlete burnout: a) physical/emotional exhaustion; b) reduced sense of accomplishment; and c) sport devaluation. Specifically, it is a likert scale comprising 15 items, 5 per factor, which employs a response format in ordered categories, with five alternatives: almost never (1), not very often (2), sometimes (3), often (4) and almost always (5). (See Table, Supplemental Digital Content 3, description of the psychological risk factors recorded, http://links.lww.com/MSS/B169.)

Neuromuscular risk factors

Prior to the neuromuscular risk factor assessment, all participants performed the dynamic warm-up designed by Taylor, Sheppard, Lee and Plummer (23). This warm-up routine was chosen because it reflects the standard warm-up structure (aerobic exercises + dynamic stretching exercises + sport-specific movements executed at, or just below game intensity) that might be the most widely used in soccer and handball. In addition, the effects elicited by this dynamic warm-up routine have been demonstrated to be enough to optimise the subsequent physical performance in elite athletes (23). The overall duration of the entire warm-up was approximately 15-20 min. The assessment of the neuromuscular risk factors was carried out 3-5 min after the dynamic warm-up.

In the experimental session, participants were assessed from a number of neuromuscular performance measures obtained from 5 different testing manoeuvres: 1) dynamic postural control; 2) isometric hip abduction and adduction strength; 3) lower extremity joint ranges of motion; 4) core stability; and 5) isokinetic knee flexion and extension strength.

The order of the tests was consistent for all participants and was established with the intention of minimizing any possible negative influence among variables. A 5-min rest interval was given between consecutive testing manoeuvres.

Dynamic postural control

Dynamic postural control was evaluated using the Y-Balance device® and following the guidelines described by Shaffer et al. (24).

The distance reached in each direction (anterior, posteromedial and posterolateral) was normalized by dividing by the previously measured leg length to standardize the maximum reach distance ([excursion distance/leg length] x100 = % maximum reach distance) (24). The bilateral ratio (dominant / non dominant score) of each direction was also calculated. A bilateral ratio higher than 10% was considered as asymmetry. Finally, to obtain a global measure of the balance test for each leg, data from each direction were averaged to calculate a composite score.

Isometric hip abduction and adduction strength

Isometric hip abduction and adduction peak torques of the dominant and non dominant limb were assessed with a portable handheld dynamometer (Nicholas Manual Muscle Tester, Lafayette Indiana Instruments) in a supine lying position on a plinth with the participant‘s legs extended and following the methodology described by Thorborg, Petersen, Magnusson and Hölmich (25). Briefly, participants performed five trials of 5-second isometric maximal voluntary contraction for each hip movement. The mean of the three most closely related trials were used for the subsequent statistical analyses. Unilateral hip abductor/adductor peak torque ratio defined as the hip adductor peak torque divided by hip abductor peak torque was calculated for each leg. Furthermore, the hip abduction and adduction bilateral ratios were also determined as the quotient of the dominant hip mean isometric peak value by the non dominant hip mean isometric peak value. A side-to-side difference higher than 10% was defined as bilateral asymmetry.

Lower extremity joints range of motion

The passive hip flexion with knee flexed and extended, extension, abduction, external and internal rotation; knee flexion; and ankle dorsiflexion with knee flexed and extended ROMs of the dominant and non dominant legs were assessed following the methodology previously described (26). Furthermore, for each joint ROM measure, side-to-side differences were also calculated. In this sense, when side-to-side difference > 6º was found, players were categorised as showing bilateral asymmetries whereas scores ≤ 6º were accepted as normal (non bilateral asymmetries) (12).

Core stability

The unstable sitting protocol described by Barbado, Lopez-Valenciano, Juan-Recio, Montero-Carretero, van Dieen and Vera-Garcia (27) was used to assess participant‘s ability to control trunk posture and motion while sitting. Briefly, after a familiarization / practice period (2 minutes), participants performed different static and dynamic tasks while sitting on an unstable seat:

- One static stability task without visual feedback (test 1) and another with visual feedback (test 2). In test 1 participants were asked to sit still in their preferred seated position on the unstable seat, while in test 2 participants were requested to adjust their centre of pressure position to a target point located in the centre of a screen placed in front of them.

- Three dynamic stability tasks with visual feedback, in which participants were asked to track the target point, which moved along three possible trajectories (anterior-posterior, medial-lateral and circular).

All tasks were performed twice. The duration of each trial was 70 seconds and the rest period between trials was 1 minute. Participants performed each trial with arms crossed over the chest. All participants were able to maintain the sitting position without grasping a support rail.

The mean radial error was used as a global measure to quantify the trunk/core performance during the trials. This variable was calculated as the mean of vector distance magnitude of the centre of pressure from the target point trials (trials with visual feedback) or from the participant‘s own mean centre of pressure position (trials without visual feedback) (28).

Isokinetic knee flexion and extension strength

A Biodex System-4 isokinetic dynamometer (Biodex Corp., Shirley, NY, USA) and its respective manufacture software were used to determine isokinetic concentric and eccentric torques during knee extension and flexion actions in both limbs following the methodology described by Ayala et al. (29).

Two isokinetic gravity-corrected variables were extracted for each movement (flexion and extension), muscle action (concentric, eccentric) and velocity (60, 180 and 240º/s for concentric actions and 30, 60 and 180º/s for eccentric actions): peak torque (PT) and joint angle of peak torque (APT). In each of the three trials at each velocity, the PT and APT were reported as the single highest torque output and corresponding joint angle. For each isokinetic variable, the average of the 3 sets at each velocity was used for subsequent statistical analysis. When a variation >5% was found in the PT and APT values between the three trials, the mean of the two most closely related torque values were used for the subsequent statistical analyses. Reciprocal (conventional and functional) knee flexion to knee extension ratios as well as bilateral knee flexion and extension ratios were also calculated using peak torque values extracted for each velocity. Thus, the conventional knee flexion to knee extension ratios were calculated as the ratio between the PTs produced concentrically by knee flexor and knee extensor muscles during the isokinetic tests. Functional knee flexion to knee extension ratios were calculated as the ratio between the PTs produced eccentrically by the knee flexor muscles and concentrically by the knee extensor muscles. Bilateral knee flexion and extension ratios were calculated dividing the PT value of the dominant limb by the PT value of the non dominant leg.

Finally, the functional knee flexion to knee extension ratio proposed by Croisier, Ganteaume, Binet, Genty and Ferret (4) was also calculated as the ratio between the PTs produced eccentrically by the knee flexor at 30º/s and concentrically by the knee extensor muscles at 240º/s.

Injury Surveillance

Following the recommendations made by the International Injury Consensus Group (30), a MUS_{INJ} was defined as an acute pain in the muscle location that occurred during training or competition and resulted in the immediate termination of play and inability to participate in the next training session or match. These injuries were confirmed through a clinical examination (identifying pain on palpation, pain with isometric contraction, and pain with muscle lengthening) by team doctors. Players were considered injured until the club medical staff (medical doctor or physiotherapist) allowed full participation in training and availability for match selection. Only hamstrings, quadriceps, triceps surae and adductor muscles injuries were considered in this study.

The club medical staff of each club recorded $\mathrm{MUS}_{\mathrm{INJ}}$ on an injury form that was sent to the study group each month. For all $\mathrm{MUS}_{\mathrm{INJ}}$ that satisfied the inclusion criteria, team medical staffs provided the following details to investigators: muscle (hamstrings, quadriceps, triceps surae and adductors), leg injured (dominant/non dominant), injury severity based on lay off time from soccer or handball (slight/minimal [0-3 days], mild [4-7 days], moderate [8-28 days], and severe [>28 days]), date of injury, moment (training or match), whether it was a recurrence (defined as an $\mathrm{MUS}_{\mathrm{INJ}}$ that occurred in the same extremity and during the same season as the initial injury), and total time taken to resume full training and competition. At the conclusion of the 9 month follow up period, all data from the individual clubs were collated into a central database, and discrepancies were identified and followed up at the different clubs to be resolved. Some discrepancies among medical staff teams were found to diagnose minimal $\mathrm{MUS}_{\mathrm{INJ}}$ and to record their total time lost. To resolve these inconsistencies in the injury surveillance process (risk of misclassification of the players), only $\mathrm{MUS}_{\mathrm{INJ}}$ showing a time lost > 4 days (minor to severe) were selected for the subsequent statistical analysis.

Statistical analysis

The statistical analysis framework carried out in this study for analysing and comparing the behaviours of several machine learning techniques with the aim of finding the best model for predicting $\mathrm{MUS}_{\mathrm{INJ}}$ in professional soccer and handball players was based on a supervised learning perspective. From a statistical standpoint, the problem can be stated as follows: given a set of features F (in our case risk factors) and a target (discrete) variable (in our case $\mathrm{MUS}_{\mathrm{INJ}}$ [yes or no]), named class, C, we want to estimate/learn a mapping function $M{:}F \rightarrow C$.

Thus, the statistical analysis comprised two stages:

1. Data pre-processing. At this stage, the data set was prepared to apply the data mining techniques. To optimise this aspect, pre-processing methods such as data cleaning and data discretization were applied.
2. Data processing. At this stage, the taxonomy suggested by Galar, Fernandez, Barrenechea, Bustince and Herrera (14) to address learning with imbalanced data sets was applied. In particular, a study on the performance of some proposals for pre-processing, cost-sensitive learning and ensemble-based methods was carried out. In addition, the approach proposed by Elkarami, Alkhateeb and Rueda (31) for imbalanced data sets and based on the combination of a cost-sensitive classifier with class-balanced ensembles was also studied. Four classic decision tree algorithms were used as base classifiers in each method.

Data pre-processing

Data pre-processing is a crucial task, due to the quality and reliability of available information, which directly affects the results obtained. Thus, some specific pre-processing tasks were applied to prepare the data set so that the classification task could be performed appropriately.

Firstly, we deleted those players who did not complete all the neuromuscular tests for any reason (six soccer players) from the data set. This exclusion criterion was based on the fact that if a player had not completed a neuromuscular test a large number of features would be absent and this might have a negative impact on the performance of the models generated. In addition, four soccer players were also deleted because they left their respective teams before the follow up procedure was completed.

Secondly, we proceed to study the presence of outliers. In this study, an outlier was defined as a score or value that could not be classified as real or true due to the consequence of a human error or a machine failure. An example of an outlier was a hip adductor peak torque value of 1500 N because the measurement range of the hand-held dynamometer used was from 0 to 1335 N. In particular, we carried out an examination of the full data set using boxplots and the detected outliers were removed.

The third step consisted of looking for missing data. To address this issue, frequency tables and diagrams were built. Thus, missing data were replaced by the mean value of the corresponding variable of the specific sport modality (soccer or handball) of the players. For example, if a football player did not report his weight for any reason, then the average value of his counterpart soccer players was inputted. It should be noted that none of the variables reported a percentage of missing data and/or outliers higher than 3%. The SPSS 21.0 Statistical software was used to carry out this data cleaning process.

After having applied the above-mentioned data cleaning methods, we had to deal with an imbalance (showing an imbalance ratio of 0.34) and high dimensional data set comprised of 88 soccer and 34 handball players (instances) and 151 potential risk factors (features).

The final step comprised the discretization of the continuous features as this has shown to be an effective measure to improve the performance of some classifiers (32). Thus, continuous features were discretized according to the reference values previously reported to consider an athlete as being more prone to suffer an injury. In most features, the discretization reduced their dimensionality to three labels. In case no cut-off scores for detecting athletes at high risk of injury had been previously reported (e.g. stature, body weight, some isokinetic strength features), the unsupervised discretization algorithm available in the well-known Weka (Waikato Environment for Knowledge Analysis) Data Mining software was applied using the equal frequency binning approach (four cut point intervals). We selected four intervals in order to reflect taxonomy of low, low-moderate, moderate-high and high scores that might make the final model more comprehensible. For the discretization of the psychological features (see Table, Supplemental Digital Content 3, description of the psychological risk factors recorded, http://links.lww.com/MSS/B169) and the isokinetic APT features we used two and three intervals or labels respectively based on the authors´ extensive experience due to the fact that the range of possible scores were limited (i.e. from 0 to 5). Thus, lower extremity range of motion features (See Table, Supplemental Digital Content 4, description of the measures obtained from the lower extremity ROM, http://links.lww.com/MSS/B170) as well as both reciprocal knee flexion to knee extension ratios and bilateral knee flexion and extension ratios (See Table, Supplemental Digital Content 5, Description of the measures obtained from the isokinetic knee flexion and extension strength assessment, http://links.lww.com/MSS/B171) were discretised according to the previously suggested cut-off scores whereas dynamic postural control (See Table, Supplemental Digital Content 6, Description of the measures obtained from the dynamic postural control test, http://links.lww.com/MSS/B172), isometric hip abduction and adduction strength (See Table, Supplemental Digital Content 7, Description of the measures obtained from the isometric hip abduction and adduction strength test, http://links.lww.com/MSS/B173), core stability (See Table, Supplemental Digital Content 8, Description of the measures obtained from the core stability test, http://links.lww.com/MSS/B174) and isokinetic peak torque (See Table, Supplemental Digital Content 5, Description of the measured obtained from the isokinetic knee flexion and extension strength assessment, http://links.lww.com/MSS/B171) features were discretized using the Weka unsupervised discretization algorithm.

Data processing

Although in Data Mining and Machine Learning a wide range of paradigms have been used to tackle classification problem, only those that have been designed to deal with imbalance and high dimensional data sets were used. These paradigms might be categorized into three groups (14, 15):

a) External approaches that pre-process the data in order to reduce the effect of their class imbalance by resampling the data space.

b) Internal approaches that create new algorithms or modify existing ones to take the class imbalance problem into consideration (ensembles).

c) Cost-sensitive learning solutions incorporating both the data (external) and algorithmic level (internal) approaches assume higher misclassification costs for samples in the minority class and seek to minimize the high cost errors.

The taxonomy for external (oversampling), internal (ensembles) and cost-sensitive methods for learning with imbalanced data sets proposed by Galar et al. (14) and López et al. (15) was used to address the aim of this study. This taxonomy was implemented with the approach recently proposed by Elkarami, Alkhateeb and Rueda (31) due to the promising results showed to handle imbalanced data sets.

To achieve founded conclusions, four decision tree algorithms were selected to be used in the pre-processing, ensemble and cost sensitive learning methodologies: C4.5 (33), which is an algorithm for generating a pruned or unpruned decision tree; SimpleCart (34), which implements minimal cost-complexity pruning; ADTree (35), which is an alternating decision tree; and RandomTree (36), which considers K randomly chosen attributes at each node of the tree.

Hence a decision tree is a set of conditions organized in a hierarchical structure. An instance is classified by following the path of satisfied conditions from the root of the tree until a leaf is reached, which will correspond with a class label.

For the sake of brevity and the lack of space, we have not written here the code of the algorithms used in this study. Instead, we have only specified the names and refer the reader to their original sources. Furthermore, all the classification algorithms used are available in Weka Data Mining software.

Although there are several data balancing or rebalancing algorithms, we used three of the most popular methodologies which are the synthetic minority oversampling technique (SMOTE), random oversampling (ROS) and random undersampling (RUS). In brief, the main idea behind SMOTE is to create new minority class examples by interpolating several minority class instances that lie together for oversampling the training set. With these techniques, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k samples belonging to the minority class, nearest to the sample *i*. Regarding ROS, it duplicates some random minority instances until the total amount of minority instances reaches the percentage given and RUS, contrarily, removes some random majority samples. In our case, a level of balance in the training data near to the 40:60 was attempted. Additionally, the interpolations that are computed to generate new synthetic data are made considering the k-5-nearest neighbours of minority class instances using the Euclidean distance.

Regarding ensemble learning algorithms, classic ensembles such as Bagging, AdaBoost and AdaBoot.M1 were included in this study. Further, the algorithm families designed to deal with skewed class distributions in data sets were also included: Boosting-based and Bagging-based. The Boosting-based ensembles that were considered in the current study were SMOTEBoost and RUSBoost. Concerning Bagging-based ensembles, it was included from the OverBagging group, OverBagging (which uses random oversampling), UnderBagging (which uses random undersampling) and SMOTEBagging.

Concerning the cost-sensitive learning algorithms, two different approaches were used, namely MetaCost and the Cost Sensitive Classifier. We have only specified the names and refer the reader for further information to Galar et al. (14) and López et al. (15).

Regarding the number of internal classifiers used within each approach, all ensembles employed the same ten base classifiers (C4.5, SimpleCart, ADTree or RandomTree***)*** by default.

Finally, the behaviour of some specific combination of class-balanced ensembles with cost-sensitive base classifiers were also studied. The final cox matrix set up was based on the best performance reported after testing all the possibilities.

Supplemental Digital Content 9 summarizes the list of algorithms (n = 68) grouped by families and also shows the abbreviations that have been used along the experimental framework and a short description of them. (See Table, Supplemental Digital Content 9, Algorithms used in the data processing phase, http://links.lww.com/MSS/B175.)

In order to evaluate the performance of the decision tree algorithms, the five fold stratified cross validation (SCV) technique was used (37). That is, we split the dataset into five stratified folds maintaining the class distribution, each one containing 20% of the patterns of the dataset. For each fold, the algorithm was trained with the examples contained in the remaining folds and then tested with the current fold. This value is set up with the aim of having enough positive class instances in the different folds, hence avoiding additional problems in the data distribution. A wide range of classification performance measures can be obtained from the SCV technique. A well-known approach to unify these measures and to produce an evaluation criterion is to use the Receiver Operating Characteristic (ROC) curve. In particular, the area under the ROC curve (AUC) corresponds to the probability of correctly identifying which one of the two stimuli is noise and which one is signal plus noise (15). Thus, the AUC was used as a single measure of a classifier‘s performance for evaluating which model is better on average and was interpreted as high (0.90- 1.00), moderate (0.70-0.90), low (0.70-0.50) and fail (>0.50) (38). Furthermore, two extra measures from the confusion matrix were also used as evaluation criteria: a) true positive rate (TPrate): TPrate = TP/(TP + FN) also called sensitivity or recall, is the proportion of actual positives which are predicted to be positive; and b) true negative rate (TNrate): TNrate = TN/(TN + FP) or specificity, that is the proportion of actual negatives which are predicted to be negative.

RESULTS

Muscle injuries epidemiology

There were 32 MUS_{INJ} over the follow up period, 21 (65.6%) of which corresponded to the hamstrings, three to the quadriceps (9.3%), four to the adductors (12.5%) and four to the triceps surae (12.5%). Injury distribution between the legs was 53.3% dominant leg and 46.7% non dominant leg. A total of 13 injures occurred during training and 19 during competition. In term of severity, most injures were categorized as moderate (n = 23) while only 9 cases were considered minor and no severe injuries were recorded. Three players were injured twice during the observation period, so their first injury was used as the index injury in the analyses. Consequently, 29 MUS_{INJ} were finally used to develop the predictive models.

Predictive model for lower extremity muscle injuries

Tables 1-3 show the average AUC, TPrate and TNrate results for all resampling, ensemble and cost-sensitive learning methods separately for each decision tree base classifier. The method that obtained the best performing result within each method is highlighted in bold. Furthermore, the model considered as the best for predicting MUS_{INJ} is highlighted in grey.

The ADTree base classifier showed the best performance in most of the methods analysed. In fact, the final model was built using the SmoteBagging ensemble method with the ADTree as base classifier using reweighted training instance (cost-sensitive).

Therefore, the final model selected to predict lower extremity MUS_{INJ} in professional soccer and handball players is comprised by 10 different cost sensitive classifiers (ADTrees) and 52 features. See, Supplemental Digital Content 10, First classifier, Graphical representation of the first classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B176; Supplemental Digital Content 11, Second classifier, Graphical representation of the second classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B177; Supplemental Digital Content 12, Third classifier, Graphical representation of the third classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B178; Supplemental Digital Content 13, Fourth classifier, Graphical representation of the fourth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B179; Supplemental Digital Content 14, Fifth classifier, Graphical representation of the fifth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B180; Supplemental Digital Content 15, Sixth classifier, Graphical representation of the sixth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B181; Supplemental Digital Content 16, Seventh classifier, Graphical representation of the seventh classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B182; Supplemental Digital Content 17, Eighth classifier, Graphical representation of the eighth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B183; Supplemental Digital Content 18, Ninth classifier, Graphical representation of the ninth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B184; Supplemental Digital Content 19, Tenth classifier, Graphical representation of the tenth classifier of the predictive model for muscle injuries, http://links.lww.com/MSS/B185; Supplemental Digital Content 20, Risk factor measures included in the model for predicting muscle injuries, http://links.lww.com/MSS/B186.

The cost matrix for cost-sensitive classifier was set to

$$C \left\{ \frac{0}{2} \middle| \frac{14}{0} \right\}$$

where a false negative had a cost of 14 and a false positive had a cost of 2. In our case, the false prediction of a non-injured athlete was penalized 7 times more with respect to the contrary error. This cost matrix was selected because it reported the best predictive performance in this particular scenario after having tested all the possible combinations.

The confusion matrix and the main cross validation results of the final model are shown in table 4. In terms of practical applications, each classifier has a vote (yes or no), and the final decision regarding whether or not a player might suffer an injury will be based on the combination of the votes of each individual classifier to each class (yes or no).

DISCUSSION

The main purpose of this study was to develop an injury risk factor-based model that would identify professional soccer and handball players at high risk of MUS_{INJ} by using learning methods from Machine Learning and Data Mining environments. With this aim in mind, a large number of personal, psychological and neuromuscular risk factors were assessed during the pre-season training periods and the MUS_{INJ} accounted within the following 9 months were also recorded. Thus, and after having run and compared the performance of several pre-processing, cost-sensitive learning and ensemble techniques to correctly classify players at high or low risk of MUS_{INJ}, the model generated by the SmooteBoost technique with a cost-sensitive ADTree as base classifier reported the best evaluation criteria (AUC score = 0.747; TPrate = 65.9; TNrate = 79.1).

Functioning of the predictive model to identify athletes at high risk of muscle injuries

The ADTree algorithm has the advantage of producing models that are easily represented as a tree with a limited number of nodes (less than 10 in our case). This property is achieved by constructing a tree that is a conjunction of rules which all contribute real-valued evidence toward a given instance being classified as either true (injured) or false (no injured). Unlike traditional tree models the classification of instances by ADTree is thus not determined by a single path traversed in the tree, but rather by the additive score of a collection of paths. The ADTree is graphically represented with two types of nodes: Elliptical *prediction nodes* and rectangular *splitter nodes* (Figure 1). Each splitter node is associated with a value indicating the rule condition: If the feature represented by the node satisfied the condition for a given instance, the prediction path will go through the left child node, otherwise the path will go through the right child node. The final classification score produced by the tree is found by summing the values from all the prediction nodes reached by the instance, with the root node being the precondition of the classifier. If the summed score is greater than zero, the instance is classified as false (no Injured).

To better explain how coaches and sport practitioners should use the model to predict MUS_{INJ}, we are going to explain the first classifier or ADTree using the fictional data displayed in figure 1. In addition, figure 1 represents in blue the paths followed by the selected instance or example.

In this classifier, we start with a baseline score of -1.252. The tree presents three father nodes placed up to the tree: APT_{ISOK}-$KE_{CON240°/s}$-Non Dominant Leg, YBalance-Anterior-Non Dominant Leg and History of MUS_{INJ} last season. Each father node represents a pathway that must be addressed.

Then, and if we start by the father node numbered as 1, placed on the left and represented by the feature named APT_{ISOK}-$KE_{CON240°/s}$-Non Dominant Leg, we realise that our player satisfies the rule condition, this is, he presents a score > 60º (Yes). Consequently, we must sum -0.497 to the initial score. Then, we have two different pathways that must be addressed. Thus, we first address the pathway that goes toward the node that contains the feature named PT_{ISOK}-$KF_{ECC30°/s}$-Non Dominant Leg. Our player satisfies again the rule condition (Yes) because he shows a score ranged from 158.3 to 198.1. Therefore, we sum -0.755 to the baseline score. Until here, we have reached an accumulative score of -2.504 (-1.252 + [-0.497] + [- 0.755]).

If we go back to the node number 1, and we follow the remaining pathway that goes toward the node number 3, we check that our player satisfies its rule condition, and then we add other -1.027 points to our scoreboard (-2.504 + [-1.027] = -3.531). As the path is not finished, we must continue through the Yes path and reach the last node, represented by the feature Core-

USNF. Here, our player satisfies again the rule condition and we must sum 0.939 point to our accumulate scoreboard. It should be noticed that this time the score summed is positive and hence, our accumulative score would be reduced. Therefore, by completing this first pathway started in the node 1 we have reached a total score of -2.592. Once we have completed this first path we must proceed with the other two primary paths, but taking into account that we have an accumulative scoreboard of -2.592.

Thus, and after completing the second main pathway, we must sum -0.246 (YBalance-Anterior-Non Dominant Leg = No) and + 0.689 (Sleep Quality = No) points to our scoreboard. Finally, we also have to sum 0.46 and 0.682 points coming from the third main pathway. All in all, our players have reached a global score of -1.007. The higher the global score is (in positive or negative way), the more confidence we are with the vote obtained.

Consequently, this classifier votes –Yes‖ and considers our athlete at high risk of injury. The final classification will be based on the combination of the votes of each individual classifier to each class (yes or no). In the very unlikely (but possible) case where a player ends with an equal amount of votes (i.e. five votes for no and five votes for yes), coaches and sport practitioners should adopt a conservative attitude and consider the athlete at high risk of $\mathrm{MUS}_{\mathrm{INJ}}$. The rationale behind this recommendation for the unlikely case of equal votes is based on the reported high incidence rate of muscle injuries in professional sports (1-3) and on the cost that a false negative diagnosis (low sensitivity) might have for team performance and player´s welfare as well as the economical cost for the club (39, 40).

Discussion of the predictive model results

The predictive ability of the current model to identify athletes at high risk of $\mathrm{MUS}_{\mathrm{INJ}}$ (AUC score = 0.747; TPrate = 65.9; TNrate = 79.1) is similar to the one reported by the only injury predictive model published to date (from the authors‘ knowledge) that was developed thanks to the application of a supervised learning algorithm (decision tress) and whose predictive properties were analysed using a resampling technique (i.e. 3-fold cross-validation) in a cohort of athletes different from those used for building it (16). Rossi, Pappalardo, Cintia, Iaia, Fernandez and Medina (16), after having collected (16 weeks) and pre-processed data about training workload (kinematic, metabolic and mechanical features) through GPS in professional soccer players, built a non-contact injury model with a tree-shape structure that reports a true positive and negative rates of 76% and 100%, respectively. In contrast to the model developed by Rossi et al. (16) that entails constant and individualised monitoring of each training session workload during the season in order to identify players at high risk of non-contact injury in the following game or training session, our model was conceived to be used as a single session pre-participation screening tool for the prevention of muscle injuries and hence, it is less time consuming and more injury-specific. On the other hand, the predictive properties (i.e. AUC, true positive and negative rates and false positive and negative rates) of the machine learning based predictive model built in the current study are higher than those reported in other models from previous studies to predict sport-related injuries in which traditional approaches and less exigent validation processes were applied (41-44). Thus, and for example, van Dyk et al. (44) after having carried out a pre-season assessment of the isokinetic hamstring and quadriceps strength in a large cohort of professional soccer players found that in spite of the fact that the regression analysis reported the presence of two independent predictors that were associated with the risk of hamstring strains (hamstring eccentric strength and quadriceps concentric strength), the ROC analysis demonstrated an AUC lower than 0.6. Likewise, Smith, Chimera and Warren (45) stated that those athletes showing unilateral dynamic balance asymmetries (determined through the Y-Balance test) higher than 4 cm had 2.3 times greater risk of a subsequent non-contact injury in comparison with more symmetrical players. However, the reported percentage of the true positive rate for this cut-off score was only 59%. Therefore, the application of contemporary statistical approaches from Machine Learning and Data Mining environments open an interesting perspective for the construction of injury prevention models that are both accurate and interpretable, helping coaches, physical trainers and medical practitioners in the decision-making process for injury prevention.

As it has been stated before, the model generated is comprised by 10 classifiers that contain the most relevant features (n = 52) for predicting MUS_{INJ}. In addition, each feature presented in the model shows a binary rule condition (yes or no) based on a specific cut-off score. Therefore, we consider that the model meets the two requirements (i.e. identifying relevant risk factors and defining cut-off scores) established in the first step suggested by Bahr (7) to be considered as a valid screening methodology.

Thus, the predictive model built considers the devaluation of the self-perceived benefits gained from sport involvement as being one of the main factors associated with an increased in the relative risk of MUS_{INJ} because it is presented in 5 of the 10 classifiers. This finding is in concordance with the results found by Cresswell and Eklund (46), who reported statistically significant correlations between sport-injuries and feelings of sport devaluation in a cohort of professional rugby players. Although the mechanisms behind the relationship between sport devaluation and injury have not been well defined yet, it might be possible that old professional athletes with a short term history of moderate to severe injuries would start questioning if the efforts made to achieve their current level of play is worth the benefits gained. These feelings of frustration might lead athletes to lose concentration and reduce the intensity of their actions during both training and match play, and thus increasing the risk of MUS_{INJ}. Therefore, psychological therapies aimed at reducing athlete burn out could help to reduce the risk of MUS_{INJ} in professional soccer and handball players.

Another strong risk factor reported by the model (presented in four classifiers) for MUS_{INJ} is having a history of MUS_{INJ} last season. Previous injury has been also identified in some prospective studies as one of the primary risk factors for $\mathrm{MUS_{INJ}}$ (8-10). A possible explanation for previous injury being such a consistent risk factor for re-injuries may be that the joints or muscles in question are not fully restored structurally and/or functionally (19). Consequently, more studies are needed in order to: a) design effective rehabilitation programmes after injury; and b) develop adequate return-to-play guidelines. Furthermore, evidence-based $\mathrm{MUS_{INJ}}$ prevention programs should be applied at the beginning of a player´s sport career in order to avoid or delay the first $\mathrm{MUS_{INJ}}$ as a high priority, in order to keep players from entering the vicious cycle of repeated injuries to the same muscle group.

Furthermore, the model built provides a main role to the isokinetic strength features measured through knee flexion and extension actions to predict future $\mathrm{MUS_{INJ}}$ (30 features up to 52). These results are not in agreement with the findings showed by van Dyk et al. (44) who reported that the use of isokinetic testing to determine the association between strength differences and hamstring muscle injuries was not supported. A possible reason behind the discrepancy between the finding reported by van Dyk et al. (44) and our results might be associated with the different statistical approach used. Thus, while van Dyk et al. (44) carried out a clustered multiple logistic regression analysis to identify isokinetic variables associated with the risk of hamstrings injuries, we used an analysis that included not only isokinetic variables but also a large number of personal, psychological and neuromuscular variables and took into account the different distribution presents in the class feature. It should be highlighted that our model endows a special protagonist for predicting future $\mathrm{MUS_{INJ}}$ to the APT measured through concentric (quadriceps) and eccentric (hamstrings) knee extension movements, as they are presented in 4 and 5 different classifiers respectively. This circumstance might support the hypothesis derived from the findings reported by Brockett, Morgan and Proske (47) so that where the players are able to achieve the PT this might be more relevant than the net PT value in order to prevent $\mathrm{MUS_{INJ}}$.

On the other hand, another relevant isokinetic feature for our predictive model is the conventional knee flexion and extension ratio measured at 60º/s. Surprisingly, no functional knee flexion and extension ratio feature were included in the final models despite being more conceptually relevant for muscle injuries than the conventional ratios (mainly hamstrings injuries). In this sense, we categorised the functional knee flexion and extension ratios using the cut-off scores reported in the literature. It is possible that these cut-off scores that were calculated using different isokinetic methodologies may not have been appropriate (very restrictive) for our model and hence, reduced its performance. Therefore, future studies should be conducted in order to explore if a potential reason for this circumstance and attempt to establish appropriate cut-off scores.

Although with less presence than the isokinetic features, the classifiers that compose the predictive model include features from all the testing methodologies used, which might support the multifactorial character of the $\mathrm{MUS}_{\mathrm{INJ}}$ phenomenon. This characteristic of the model might support its congruence.

Finally, the feature sport (football or handball) was not included in any of the 10 classifiers that comprised the model for predicting $\mathrm{MUS}_{\mathrm{INJ}}$. Furthermore, the same statistical analysis framework that was conducted in the present study was carried out in a preliminary study for soccer players solely, showing a less favourable predictive performance score (AUC score = 0.646; TPrate = 56.0; TNrate = 70.5 [unpublished data from our laboratory]). Therefore, it may be that data from athletes from different sport modalities, but who have similar movement demands, $\mathrm{MUS}_{\mathrm{INJ}}$ incidence rates and injury mechanism, can be analysed all together in order to develop a more generalizable model. Future studies should explore this hypothesis by analysing and comparing the behaviour for predicting $\mathrm{MUS}_{\mathrm{INJ}}$ of models built using athletes from different sports, collectively and separately.

Using the cross-validation process, we consider that the model might have met the second step proposed by Bahr (7). However, due to the reduced sample size, we think more studies that re-evaluate the predictive performance of the model using data from new players are necessary.

LIMITATIONS

Although the model presented in this study shows moderate predictive scores, it should be acknowledged that more sophisticated algorithms (i.e. neural networks, genetic algorithms) might have developed models showing slightly better results than those found in the current study. However, the use of more complex algorithms would require sport medicine practitioners to carry out complex mathematical functions and operations, which might impact on the practical application of the model built. Thus, and in order to allow sport medicine practitioners to implement the model in their screening programmes, we decided to use decision trees algorithms as base classifiers because: a) they produce models that are easy to understand and carry out functioning for classifying instances (i.e.: simple rules) and can be used directly for decision making; and b) they have been widely used as base classifiers in some balancing, ensemble and cost sensitive learning techniques to deal with imbalance data sets.

The model developed in the present study was built with the goal of allowing sport medicine practitioners to accurately identify professional soccer and handball players at high risk of MUS_{INJ} during pre-season screenings. To address this issue, we used several predictors (risk factors) as well as external (oversampling) and internal (ensembles) methods and a decision tree (ADTree) as base classifier in order to build a model with moderate predictive accuracy. This set up allowed us to build a robust model (AUC score = 0.747; TPrate = 65.9; TNrate = 79.1) which was also very complex in nature (black box approach). Therefore, although the model fulfils the goal for which it was built (making predictions); its complexity (10 different classifiers and 52 predictors) does not afford the opportunity to answer the question concerning why MUS_{INJ} happen.

Another potential limitation of the current study is the population used. The sport background of participants was professional soccer and handball and the generalizability to other sport modalities and level of play cannot be ascertained. Furthermore the results reported in this study suggest that the feature _sport' does not influence the performance scores of the model selected, which might be due to the different sample size of both cohorts and the fact that only two different sports were analysed. Therefore from the current data set we cannot draw strong conclusions around how mixing players from differing sports will affect the classification performance of the models and more importantly, why and when we should mix players from differing sports.

Finally, it should also be noted that the model is dependent of the predictors used in the training process and hence, practitioners must follow the same assessment methodologies used in the current study in order to replicate the current results and gain the applicability in their populations.

CONCLUSION

The current study has used an injury risk factor model to identify professional soccer and handball players at high risk of MUS_{INJ} by applying a novel multifactorial approach and whose predictive ability has been determined through the exigent resampling technique called cross-validation. In this study the MUS_{INJ} risk model is comprised of 10 classifiers with a tree-shape structure and was developed thanks to the application of learning algorithms (on the training subsets) widely used in the Data Mining setting. Thus, the model reports an AUC score of 0.747 with true positive and negative rates of 65.9% and 79.1% respectively. We believe that the approach used here could replace the conventional statistical methods and can be used for coaches, physical trainers and medical practitioners to gain valuable information in the decision-making process aimed at reducing the number and severity of $\mathrm{MUS}_{\mathrm{INJ}}$ in professional soccer and handball players.

Acknowledgments

Alejandro López-Valenciano were supported by predoctoral grant given by Ministerio de Educación, Cultura y Deporte (FPU) from Spain.

Conflict of interest

We certify that no party having a direct interest in the results of the research supporting this article has or will confer a benefit on us or on any organization which we are associated, do not constitute endorsement by ACSM and they are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

REFERENCES

1. Ekstrand J, Hägglund M, Waldén M. Epidemiology of muscle injuries in professional football (soccer). *Am J Sports Med*. 2011;39(6):1226-32.
2. Brooks JH, Fuller CW, Kemp SP, Reddin DB. Incidence, risk, and prevention of hamstring muscle injuries in professional rugby union. *Am J Sports Med*. 2006;34(8):1297-306.
3. Langevoort G, Myklebust G, Dvorak J, Junge A. Handball injuries during major international tournaments. *Scand J Med Sci Sports*. 2007;17(4):400-7.
4. Croisier J-L, Ganteaume S, Binet J, Genty M, Ferret J-M. Strength imbalances and prevention of hamstring injury in professional soccer players a prospective study. *Am J Sports Med*. 2008;36(8):1469-75.
5. Mendiguchia J, Alentorn-Geli E, Idoate F, Myer GD. Rectus femoris muscle injuries in football: a clinically relevant review of mechanisms of injury, risk factors and preventive strategies. *Br J Sports Med*. 2013;47(6):359-66.
6. Faude O, Rößler R, Junge A. Football injuries in children and adolescent players: are there clues for prevention? *Sports Med*. 2013;43(9):819-37.
7. Bahr R. Why screening tests to predict injury do not work—and probably never will…: a critical review. *Br J Sports Med*. 2016;50:776-80.
8. Arnason A, Sigurdsson S, Gudmundsson A, Holme I, Engebretsen L, Bahr R. Risk factors for injuries in football. *Am J Sports Med*. 2004;32(1 Suppl):5S-16S.
9. Engebretsen A, Myklebust G, Holme I, Engebretsen L, Bahr R. Intrinsic risk factors for hamstring injuries among male soccer players a prospective cohort study. *Am J Sports Med*. 2010;38(6):1147-53.

10. Hägglund M, Waldén M, Ekstrand J. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *Br J Sports Med*. 2006;40(9):767-72.

11. Henderson G, Barnes CA, Portas MD. Factors associated with increased propensity for hamstring injury in English Premier League soccer players. *J Sci Med Sport*. 2010;13(4):397-402.

12. Fousekis K, Tsepis E, Poulmedis P, Athanasopoulos S, Vagenas G. Intrinsic risk factors of non-contact quadriceps and hamstring strains in soccer: a prospective study of 100 professional players. *Br J Sports Med*. 2011;45(9):709-14.

13. Hawkins RD, Fuller CW. A prospective epidemiological study of injuries in four English professional football clubs. *Br J Sports Med*. 1999;33(3):196-203.

14. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. 2012;42(4):463-84.

15. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sci*. 2013;250:113-41.

16. Rossi A, Pappalardo L, Cintia P, Iaia M, Fernandez J, Medina D. Effective injury prediction in professional soccer with GPS data and machine learning. *arXiv preprint arXiv:1705.08079*. 2017.

17. Mendiguchia J, Alentorn-Geli E, Brughelli M. Hamstring strain injuries: are we heading in the right direction? *Br J Sports Med*. 2012;46(2):81-5.

18. Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Netw*. 2008;21(2):427-36.

19. Bahr R, Holme I. Risk factors for sports injuries—a methodological approach. *Br J Sports Med*. 2003;37(5):384-92.

20. Olmedilla A, Laguna M, Redondo AB. Lesiones y características psicológicas en jugadores de balonmano. *Rev Andal Med Deporte*. 2011;4(1):6-12.

21. Macías J, Royuela A. La versión española del Índice de Calidad de Sueño de Pittsburgh. *Informaciones Psiquiátricas*. 1996;146(4):465-72.

22. Arce C, De Francisco C, Andrade E, Seoane G, Raedeke T. Adaptation of the Athlete Burnout Questionnaire in a Spanish sample of athletes. *Span J Psychol*. 2012;15(3):1529-36.

23. Taylor K-L, Sheppard JM, Lee H, Plummer N. Negative effect of static stretching restored when combined with a sport specific warm-up component. *J Sci Med Sport*. 2009;12(6):657-61.

24. Shaffer SW, Teyhen DS, Lorenson CL et al. Y-balance test: a reliability study involving multiple raters. *Mil Med*. 2013;178(11):1264-70.

25. Thorborg K, Petersen J, Magnusson S, Hölmich P. Clinical assessment of hip strength using a hand-held dynamometer is reliable. *Scand J Med Sci Sports*. 2010;20(3):493-501.

26. Cejudo A, Sainz de Baranda P, Ayala F, Santonja F. Perfil de flexibilidad de la extremidad inferior en jugadores de fútbol sala. *Rev Int Med Cienc Act Fís Deporte*. 2014;14(55):509-25.

27. Barbado D, Lopez-Valenciano A, Juan-Recio C, Montero-Carretero C, van Dieen JH, Vera-Garcia FJ. Trunk stability, trunk strength and sport performance level in judo. *PloS one*. 2016;11(5):e0156267.

28. Hancock GR, Butler MS, Fischman MG. On the problem of two-dimensional error scores: Measures and analyses of accuracy, bias, and consistency. *J Mot Behav*. 1995;27(3):241-50.

29. Ayala F, Puerta JM, Flores MJ et al. Análisis bayesiano de los principales factores de riesgo de lesión de la musculatura isquiosural. *Kronos*. 2016;15(1).

30. Fuller CW, Ekstrand J, Junge A et al. Consensus statement on injury definitions and data collection procedures in studies of football (soccer) injuries. *Scand J Med Sci Sports*. 2006;16(2):83-92.

31. Elkarami B, Alkhateeb A, Rueda L. Cost-sensitive classification on class-balanced ensembles for imbalanced non-coding RNA data. In: *Proceedings of the Student Conference (ISC), 2016 IEEE EMBS International*. 2016. p. 1-4.

32. Hacibeyoglu M, Arslan A, Kahramanli S. Improving Classification Accuracy with Discretization on Data Sets Including Continuous Valued Features. *Ionosphere*. 2011;34(351):2.

33. Quinlan JR. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*. 1996;28(1):71-2.

34. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Wadsworth & Brooks. *Monterey, CA*. 1984.

35. Freund Y, Mason L. The alternating decision tree learning algorithm. In: *Proceedings of the icml*. 1999. p. 124-33.

36. Aldous D. The continuum random tree. I. *Annals of Probability*. 1991:1-28.

37. Refaeilzadeh P, Tang L, Liu H. Cross-validation. In. *Encyclopedia of Database Systems*: Springer; 2009, pp. 532-8.

38. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ: BMJ*. 1994;309(6948):188.

39. Carling C, Le Gall F, McCall A, Nédélec M, Dupont G. Squad management, injury and match performance in a professional soccer team over a championship-winning season. *Eur J Sport Sci*. 2015;15(7):573-82.

40. Ekstrand J, Dvorak J, D'hooghe M. Sport medicine research needs funding: the International football federations are leading the way. *Br J Sports Med*. 2013;47(12):726-8.

41. Hewett TE, Myer GD, Ford KR et al. Biomechanical measures of neuromuscular control and valgus loading of the knee predict anterior cruciate ligament injury risk in female athletes. *Am J Sports Med*. 2005;33(4):492-501.

42. Krosshaug T, Steffen K, Kristianslund E et al. The vertical drop jump is a poor screening test for ACL injuries in female elite soccer and handball players: a prospective cohort study of 710 athletes. *Am J Sports Med*. 2016;44(4):874-83.

43. Timmins RG, Bourne MN, Shield AJ, Williams MD, Lorenzen C, Opar DA. Short biceps femoris fascicles and eccentric knee flexor weakness increase the risk of hamstring injury in elite football (soccer): a prospective cohort study. *Br J Sports Med*. 2016:50(24):1524-35.

44. van Dyk N, Bahr R, Whiteley R et al. Hamstring and quadriceps isokinetic strength deficits are weak risk factors for hamstring strain injuries: A 4-year cohort study. *Am J Sports Med*. 2016;44(7):1789-95.

45. Smith CA, Chimera NJ, Warren M. Association of y balance test reach asymmetry and injury in division I athletes. *Med Sci Sports Exerc*. 2015;47(1):136-41.

46. Cresswell SL, Eklund RC. The nature of player burnout in rugby: Key characteristics and attributions. *J Appl Sport Psychol*. 2006;18(3):219-39.

47. Brockett CL, Morgan DL, Proske U. Predicting hamstring strain injury in elite athletes. *Med Sci Sports Exerc*. 2004;36(3):379-87.

FIGURES LEGEND

Figure 1: Graphical representation of the first classifier. Prediction nodes are represented by ellipses and splitter nodes by rectangles. Each splitter node is associated with a real valued number indicating the rule condition, meaning: If the feature represented by the node satisfies the condition value the prediction path will go through the left child node, otherwise the path will go through the right child node. The numbers before the feature names in the prediction nodes indicate the order in which the different base rules were discovered. This ordering can to some extent indicate the relative importance of the base rules.

SUPPLEMENTAL DIGITAL CONTENT

- **SDC 1:** Graphical representation of testing procedure.

  The order of the different tests used to record the personal or individual, psychological and neuromuscular risk factors in the testing session is shown.

- **SDC 2:** Personal injury risk factors recorded**.**

  Description of the personal injury risk factors recorded (names and labels).

- **SDC 3:** Psychological risk factors recorded.

  Description of the psychological risk factors recorded (names and labels).

- **SDC 4:** Lower extremity joints ranges of motion measures recorded.

  Description of the measures obtained from the lower extremity joints (hip, knee and ankle) ranges of motion (names and labels).

- **SDC 5:** Isokinetic knee flexion and extension strength measures recorded.

  Description of the measures obtained from the isokinetic knee flexion and extension strength (concentric and eccentric) assessment (names and labels).

- **SDC 6:** Dynamic postural control measures recorded.

Description of the measures obtained from the Y-Balance test (names and labels).

- **SDC7:** Isometric hip abduction and adduction strength measures recorded

  Description of the measures obtained from the isometric hip abduction and adduction strength test (names and labels).

- **SDC 8:** Core stability measures recorded.

  Description of the measures obtained from the core stability test (names and labels).

- **SDC 9:** Algorithms used in the data processing phase.

  A list of algorithms (n = 68) grouped by families, the abbreviations that have been used along the experimental framework and a short description of them are displayed.

- **SDC 10**: First classifier.

  Graphical representation of the first classifier of the predictive model for muscle injuries.

- **SDC 11**: Second classifier.

  Graphical representation of the second classifier of the predictive model for muscle injuries.

- **SDC 12**: Third classifier.

  Graphical representation of the third classifier of the predictive model for muscle injuries.

- **SDC 13**: Fourth classifier.

  Graphical representation of the fourth classifier of the predictive model for muscle injuries.

- **SDC 14**: Fifth classifier.

  Graphical representation of the fifth classifier of the predictive model for muscle injuries.

- **SDC 15**: Sixth classifier.

Graphical representation of the sixth classifier of the predictive model for muscle injuries.

- **SDC 16**: Seventh classifier.

  Graphical representation of the seventh classifier of the predictive model for muscle injuries.

- **SDC 17**: Eighth classifier.

  Graphical representation of the eighth classifier of the predictive model for muscle injuries.

- **SDC 18**: Ninth classifier.

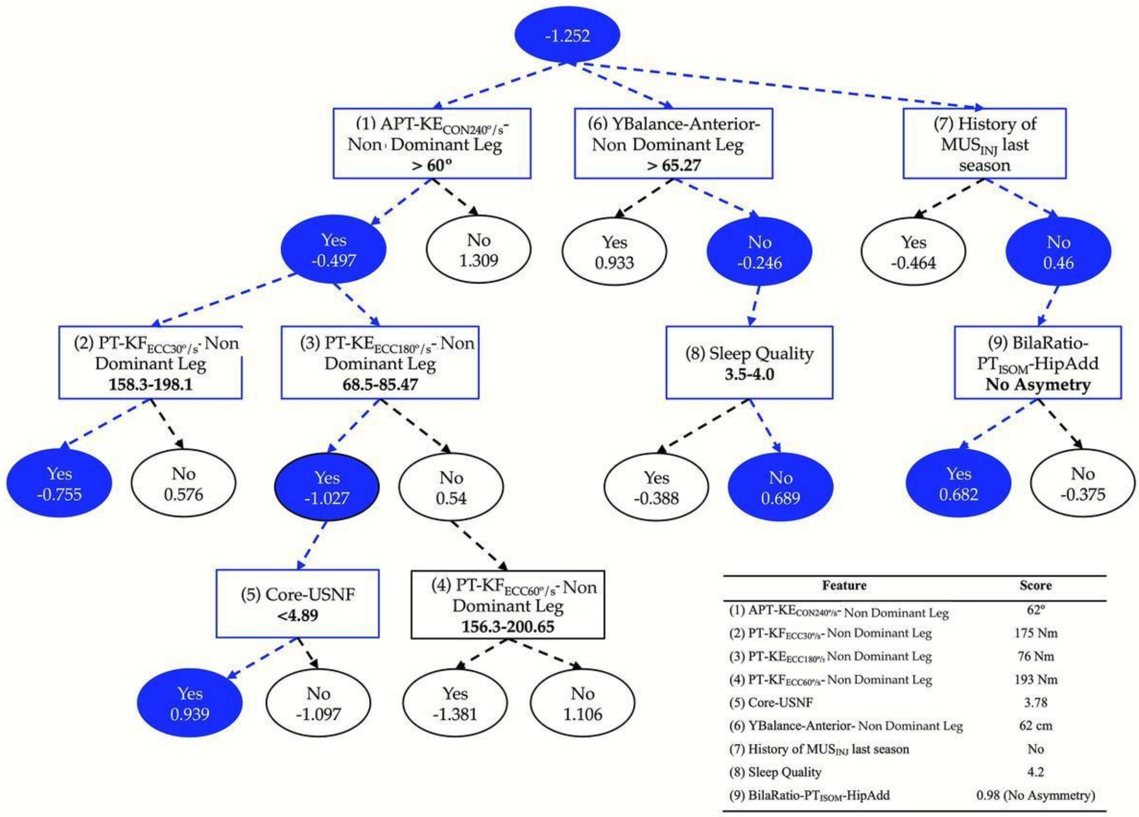
  Graphical representation of the ninth classifier of the predictive model for muscle injuries.

- **SDC 19**: Tenth classifier.

  Graphical representation of the tenth classifier of the predictive model for muscle injuries.

- **SDC 20:** Risk factor measures included in the model for predicting muscle injuries.

  Risk factor measures included in the model for predicting muscle injuries and the number of times that they appear in the classifiers, In bold are highlighted those that appear in four or more classifiers.


-1.252
(1) APT-KE$_{CON240°/s}$- Non Dominant Leg > 60º
(6) YBalance-Anterior- Non Dominant Leg > 65.27
(7) History of MUS$_{INJ}$ last season
Yes -0.497
No 1.309
Yes 0.933
No -0.246
Yes -0.464
No 0.46
(2) PT-KF$_{ECC30°/s}$- Non Dominant Leg 158.3-198.1
(3) PT-KE$_{ECC180°/s}$- Non Dominant Leg 68.5-85.47
(8) Sleep Quality 3.5-4.0
(9) BilaRatio-PT$_{ISOM}$-HipAdd No Asymetry
Yes -0.755
No 0.576
Yes -1.027
No 0.54
Yes -0.388
No 0.689
Yes 0.682
No -0.375
(5) Core-USNF <4.89
(4) PT-KF$_{ECC60°/s}$- Non Dominant Leg 156.3-200.65
Yes 0.939
No -1.097
Yes -1.381
No 1.106

Feature	Score
(1) APT-KE$_{CON240°/s}$- Non Dominant Leg	62º
(2) PT-KF$_{ECC30°/s}$- Non Dominant Leg	175 Nm
(3) PT-KE$_{ECC180°/s}$ Non Dominant Leg	76 Nm
(4) PT-KF$_{ECC60°/s}$- Non Dominant Leg	193 Nm
(5) Core-USNF	3.78
(6) YBalance-Anterior- Non Dominant Leg	62 cm
(7) History of MUS$_{INJ}$ last season	No
(8) Sleep Quality	4.2
(9) BilaRatio-PT$_{ISOM}$-HipAdd	0.98 (No Asymmetry)

Table 1: Average AUC, TPrate and TNrate results for all the decision tree methodologies in isolation and after having been applied in them the resampling techniques selected

Technique	**AUC**	TPrate	TNrate
Base classifiers			
J48	0.422	17.2	79.1
SCart	0.462	3.4	94.5
ADTree	**0.623**	**20.7**	**87.9**
RTree	0.609	51.7	65.9
Oversampling techniques			
SMT			
J48	0.452	31	78
SCart	0.489	34.5	71.4
ADTree	**0.608**	**31**	**76.9**
RTree	0.522	34.5	71.4
ROS			
J48	0.575	44	72.5
SCart	0.618	48.3	73.6
ADTree	**0.709**	**48.3**	**84.6**
RTree	0.711	55.2	82.4
Undersampling techniques			
RUS			
J48	0.607	55.2	62.4

SCart	0.574	13.8	93.4
ADTree	**0.662**	**62.1**	**70.3**
RTree	0.559	48.3	61.5

Table 2: Average AUC, TPrate and TNrate results for the ensembles techniques

Technique	**AUC**	TPrate	TNrate
Classic Ensembles			
ADB1			
J48	0.579	13.8	90.1
SCart	0.605	37.9	83.5
ADTree	**0.692**	**24.1**	**93.4**
RTree	0.594	10.3	98.9
M1			
J48	0.560	0	91.2
SCart	0.550	20.7	84.6
ADTree	**0.703**	**27.6**	**90.1**
RTree	0.517	20.7	85.7
BAG			
J48	0.544	6.9	93.4
SCart	0.669	3.4	97.8
ADTree	**0.722**	**10.3**	**98.9**
RTree	0.663	24.1	91.2
Boosting-based Ensembles			
SBO			
J48	0.494	24.1	76.9
SCart	**0.692**	**41.4**	**85.7**
ADTree	0.650	27.6	85.7

RTree	-	-	-
RUSB			
J48	0.610	37.9	75.8
SCart	0.649	51.7	78
ADTree	0.698	31	92
RTree	**0.717**	**48.3**	**84.6**
Bagging-based Ensembles			
OB			
J48	0.583	13.8	92.3
SCart	0.716	13.8	93.4
ADTree	**0.759**	**10.3**	**96.7**
RTree	0.633	13.8	89.0
UB			
J48	0.670	27.6	84.6
SCart	**0.708**	**31**	**87.9**
ADTree	0.624	41.4	73.6
RTree	0.570	27.6	82.4
SBAG			
J48	0.562	13.8	96.7
SCart	0.642	10.3	96.7
ADTree	**0.728**	**20.7**	**96.7**
RTree	0.547	24.1	93.4

Table 3: Average AUC, TPrate and TNrate results for the and cost-sensitive learning and class-balanced ensembles with a cost-sensitive classifier techniques

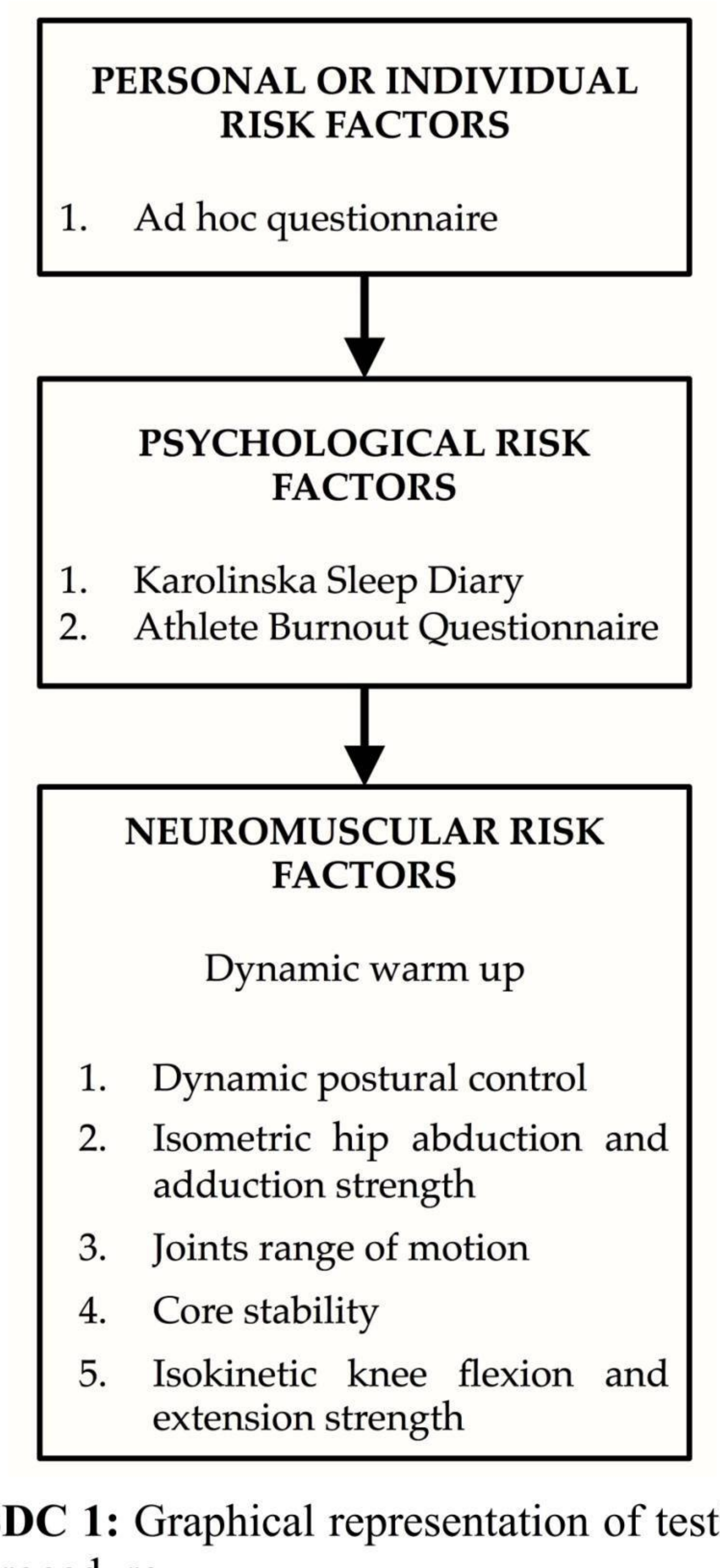
Technique	**AUC**	TPrate	TNrate
Cost-sensitive classification			
MetaCost			
J48	0.473	41.4	61.5
SCart	0.579	17.2	90.1
ADTree	**0.662**	**75.9**	**40.7**
RTree	0.561	48.3	63.7
CS-Classifier			
J48	0.526	51.7	57.1
SCart	0.543	44.0	52.7
ADTree	**0.642**	**51.7**	**70.3**
RTree	0.535	44.0	60.4
Class-balanced ensembles with a cost-sensitive classifier			
CS-SBAG			
J48	0.529	51.7	51.6
SCart	0.610	65.5	54.9
ADTree	**0.747**	**65.5**	**79.1**
RTree	0.541	6.9	86.8
CS-OBAG			

J48	0.514	41.4	72.5
SCart	0.606	55.2	63.7
ADTree	0.742	62.1	71.4
RTree	0.548	13.8	96.7
	CS-UBAG		
J48	0.553	41.4	67
SCart	0.649	51.7	69.2
ADTree	0.742	58.6	68.1
RTree	0.627	37.9	82.4

Table 4: Confusion Matrix and Cross validation results for the final prediction model

A	B	□ Classified as
19	10	A = Injured
19	72	B = Non Injured

Correctly classified instances	91 (75.8%)
Incorrectly Classified Instances	29 (24.1%)
Kappa statistic	0.401
Mean absolute error	0.405
AUC	0.747


PERSONAL OR INDIVIDUAL RISK FACTORS
1. Ad hoc questionnaire
PSYCHOLOGICAL RISK FACTORS
1. Karolinska Sleep Diary
2. Athlete Burnout Questionnaire
NEUROMUSCULAR RISK FACTORS
Dynamic warm up
1. Dynamic postural control
2. Isometric hip abduction and adduction strength
3. Joints range of motion
4. Core stability
5. Isokinetic knee flexion and extension strength

SDC 1: Graphical representation of testing procedure.

The order of the different tests used to record the personal or individual, psychological and neuromuscular risk factors in the testing session is shown.

Appendix 2: Description of the personal injury risk factors recorded

Name	Labels
Sport	Soccer or handball
Player position	Goalkeeper, defender, midfielder or striker
Current level of play	1st division, 2nd B division, or 3rd division
Dominant leg	Right, left or two-footed
Age	Sub21, sub23, senior [23-30 y] or veteran [> 30y]
Body mass (kg)	<71.65, 71.65-76.55, >76.55-82.8 or >82.8
Stature (cm)	<1.76, 1.76-1.81, >1.81-1.84 or >1.84
BMI (kg/m^2)	<22.75, 22.75-23.55, >23.55-24.75 or >24.75
History of MUS$_{INJ}$ last season	Yes or no

MUS$_{INJ}$: Lower extremity muscle injury; BMI: body mass index

Appendix 3: Description of the psychological risk factors recorded

Name	Labels
Sleep quality	<3.5, 3.5-4.0 or >4.0
Athlete Burnout Questionnaire	
a) Physical/emotional exhaustion	<2.5 or ≥2.5
b) Reduced sense of accomplishment	≤ 2.5 or >2.5
c) Sport devaluation	<1.1, 1.1-1.49, >1.49-1.9 or >1.9

Appendix 4: Description of the measures obtained from the lower extremity range of motion assessment tests

Name	Labels
ROM-HF$_{KF}$	≤150 or >150 (1)
ROM-HF$_{KE}$	<80, 80-100 or >100 (2)
ROM-HE	<5, 5.0-15 or >15 (5)
ROM-HABD	<50, 50-70 or >70 (3)
ROM-HIR	<45, 45-60 or >60 (1)
ROM-HER	<40, 40-55 or >55 (1)
ROM-KF	<110, 110-130 or >130
ROM-AKDF$_{KE}$	<30, 30-40 or >40 (5)
ROM- AKDF$_{KF}$	<30, 30-40 or >40 (4)

ROM: range of motion; HF$_{KF}$: hip flexion with the knee flexed; HF$_{KE}$: hip flexion with the knee extended; HE: Hip extension; HABD: hip abduction at 90º of hip flexion; HIR: hip internal rotation; HER: hip external rotation; KF: knee flexion; AKDF$_{KE}$: ankle dorsi-flexion with the knee extended; AKDF$_{KF}$: ankle dorsi-flexion with the knee flexed.

(1): American Academy of Orthopaedic Association, 1975; (2): Palmer & Epler, 2002; (3): Gerhardt, 1994; (4) Pope, Herbert & Kirwan (1998); (5) Cejudo, 2016.

Appendix 5: Description of the measures obtained from the isokinetic knee flexion and extension strength assessment

Measure	Labels	
	Dominant Leg	Non Dominant Leg
	Concentric Muscle Actions	
PT-KE$_{60}$	<163.1, 163.1-184.605, >184.605-211.05 or >211.05	<158.3, 158.3-179.14, >179.14-197.3 or >197.3
PT-KF$_{60}$	<74.6, 74.6-87.505, >87.505-104.65 or >104.65	<68.7, 68.7-84.9, >84.9-98.2 or >98.2
PT-KE$_{180}$	<112.05, 112.05-129.3, >129.3-146.3 or >146.3	<113.6, 113.6-128.495, >128.495-146.55 or >146.55
PT-KF$_{180}$	<59.55, 59.55-70.4, >70.4-81.4 or >81.4-	<60.1, 60.1-68.35, >68.35-79.75 or >79.75
PT-KE$_{240}$	<98.05, 98.05-114.55, >114.55-129.3 or >129.3	<95.45, 95.45-113.9, >113.9-130.65 or >130.65
PT-KF$_{240}$	<57.8, 57.8-65.86, >65.86-78.75 or >78.75	<55.7, 55.7-64.095, >64.095-75.75 or >75.75
PT-KE$_{300}$	<90.75, 90.75-104.15, >104.15-117.45 or >117.45	<85.45, 85.45-103.45, >103.45-115.2 or >115.2
PT-KF$_{300}$	<54.55, 54.55-61.9, >61.9-74.3 or >74.3	<48.2, 48.2-58.55, >58.55-69.1 or >69.1
APT-KE	<45, 45-60 or >60	
APT-KF	<25, 25-35 or >35	
	Eccentric Muscle Actions	

PT-KE$_{30}$	<72.75, 72.75-90.105, >90.105-109.15 or >109.15	<70.65, 70.65-84.12, >84.12-95.75 or >95.75
PT-KF$_{30}$	<169.2, 169.2-207.42, >207.42-242.2 or >242.2	<158.3, 158.3-198.1, >198.1-236.9 or >236.9
PT-KE$_{60}$	<74.4, 74.4-91.14, >91.14-109 or >109	<68.85, 68.85-86.3, 86.3-101.65 or >101.65
PT-KF$_{60}$	<175.6, 175.6-211.28, >211.28-244.9 or >244.9	<156.3, 156.3-200.65, >200.65-239.95 or >239.95
PT-KE$_{180}$	<73.6, 73.6-89.95, >89.95-106 or >106	<68.5, 68.5-85.475, >85.475-96.45 or >96.45
PT-KF$_{180}$	<155.35, 155.35-192.65, >192.65-221.3 or >221.3	<157.2, 157.2-187.99, >187.99-216.05 or >216.05
APT-KE	<25, 25-35 or >35	
APT-KF	<50, 50-65 or >65	
Unilateral Conventional Ratios		
(1) KF/KE$_{\text{CONV60}}$	<0.47, 0.47-0.60 or >0.60	
(2) KF/KE$_{\text{CONV180}}$	≤0.60 or >0.60	
(3) KF/KE$_{\text{CONV240}}$	≤0.60 or >0.60	
KF/KE$_{\text{CONV300}}$	<0.6 0.6-0.8 or >0.8	
Unilateral Functional Ratios		
(4) KF/KE$_{\text{FUNC60}}$	<0.6, 0.6-0.7 or >0.7	
KF/KE$_{\text{FUNC180}}$	≤0.80 or >0.80	
(5) KF$_{30}$/KE$_{240}$	<0.8, 0.8-1.0 or >1.0	
Bilateral Ratios		

KF/KF$_{CON60}$	No Asymmetry or Asymmetry
KF/KF$_{CON180}$	No Asymmetry or Asymmetry
KF/KF$_{CON240}$	No Asymmetry or Asymmetry
KE/KE$_{CON60}$	No Asymmetry or Asymmetry
KE/KE$_{CON180}$	No Asymmetry or Asymmetry
KE/KE$_{CON240}$	No Asymmetry or Asymmetry
KF/KF$_{ECC60}$	No Asymmetry or Asymmetry
KF/KF$_{ECC180}$	No Asymmetry or Asymmetry
KF/KF$_{ECC240}$	No Asymmetry or Asymmetry
KE/KE$_{ECC60}$	No Asymmetry or Asymmetry

PT: peak torque; KE: knee extension; KF: knee flexion; CON: concentric; ECC: eccentric; APT: angle of peak torque; (1) Croisier et al. (2003); (2): Yeung et al. (2009); (3): Devan et al. (2004); (4): Dauty et al. (2003); (5) Croisier et al. (2002)

Appendix 6: Description of the measures obtained from the dynamic postural control test

Name	Labels	
	Dominant Leg	**No Dominant Leg**
YBalance-Anterior	<56.48, 56.48-60.055, >60.055-63.86 or >63.86	<57.3, 57.3-60.895, >60.895-65.27 or >65.27
YBalance-PosteroMedial	<97.535, 97.535-104.055, >104.055-108.885 or >108.885	<100.42, 100.42-104.905, >104.905-108.8 or >108.8
YBalance-PosteroLateral	<94.35, 94.35-99.485, >99.485-106.79 or >106.79	<93.625, 93.625-99.175, >99.175-104.48 or >104.48
BilaRatio-YBalance-Anterior	No Asymmetry or Asymmetry	
BilaRatio-YBalance-PosteroMedial	No Asymmetry or Asymmetry	
BilaRatio-YBalance-PosteroLateral	No Asymmetry or Asymmetry	
YBalance-Composite	<83.245, 83.245-87.86, >87.86-92.035 or >92.035	<84.185, 84.185-87.985, >87.985-91.84 or >91.84

Appendix 7: Description of the measures obtained from the isometric hip abduction and adduction strength test

Name	Labels	
	Dominant Leg	Non Dominant Leg
PT_{ISOM}-HipAbd	<182.225, 182.225-204.09, >204.09-221.17 or >221.17	<188.575, 188.575-208.9, >208.9-227 or >227
PT_{ISOM}-HipAbd-Normalice	<2.39, 2.39-2.65, >2.65-2.945 or >2.945	<2.485, 2.485-2.705, >2.705-2.935 or >2.935
PT_{ISOM}-HipAdd	<187.75, 187.75-205.335, >205.335-224.54 or >224.54	<181.975, 181.975-199.9, >199.9-224.2 or >224.2
PT_{ISOM}-HipAdd-Normalise	<2.385, 2.385-2.735, >2.735-2.99 or >2.99	<2.355, 2.355-2.655, >2.655-2.945 or >2.945
UnRatio-ISOM-HipAbd/HipAdd	<0.936, 0.936-1.045, >1.045-1.17 or >1.17	<0.905, 0.905-0.973, >0.973.065 or >1.065
BilaRatio-PT_{ISOM}-HipAbd/HipAdd	No Asymmetry or Asymmetry	

Bila: bilateral; Uni: unilateral; ISOM: isometric; PT: peak torque; Abd: abduction; Add: adduction.

Appendix 8: Description of the measures obtained from the core stability test

Name	Labels
USNF	<4.895, 4.895-6.14, >6.14-7.83 or >7.83
USWF	<4.335, 4.335-5.475, >5.475-6.84 or >6.84
USML	<6.915, 6.915-8.47, >8.47-9.62 or >9.62
USAP	<7.19, 7.19-8.33, >8.33-9.865 or >9.865
USCD	<9.01, 9.01-10.555, >10.555-12.375 or >12.375

USNF: unstable sitting without feedback; USWF: unstable sitting with feedback; USML: unstable sitting while performing medial-lateral displacements with feedback; USAP: unstable sitting while performing anterior-posterior displacements with feedback; USCD: unstable sitting while performing circular displacements with feedback.

Appendix 9: Algorithms used in the data processing phase

Base classifiers		
Abbr.	*Method*	*Short Description*
J48	J48	Algorithm for generating a pruned or unpruned C4.5 decision tree
SCart	SimpleCart	Algorithm for implementing minimal cost-complexity pruning
ADTree	ADTree	Alternating decision tree
RTree	RandomTree	Algorithm that considers K randomly chosen attributes at each node of the tree

Resampling techniques		
Abbr.	*Method*	*Short Description*
SMT	SMOTE	Each decision tree applied on data set previously pre-processed with Smote
ROS	Random over sampling	Each decision tree applied on data set previously pre-processed with random over sampling
RUS	Random under sampling	Each decision tree applied on data set previously pre-processed with random under sampling

Classis Ensembles		
Abbr.	*Method*	*Short Description*
ADAB	AdaBoost	Classic AdaBoost, without using confidences
M1	AdaBoost.M1	Multi-class AdaBoost, slightly different

		weight update
BAG	Bagging	Classic Bagging, resampling with replacement, bag size equal to original data set size.

Boosting-based Ensembles

Abbr.	*Method*	*Short Description*
SBO	SmoteBoost	AdaBoost.M2 with Smote in each iteration
RUS	RusBoost	AdaBoost.M2 with random undersampling in each iteration

Cost-sensitive learning

Abbr.	*Method*	*Short Description*
MetaCost	MetaCost	Makes base classifier cost-sensitive by passing it to Bagging
CS-Classifier	Cost Sensitive Classifier	Makes base classifier cost-sensitive.

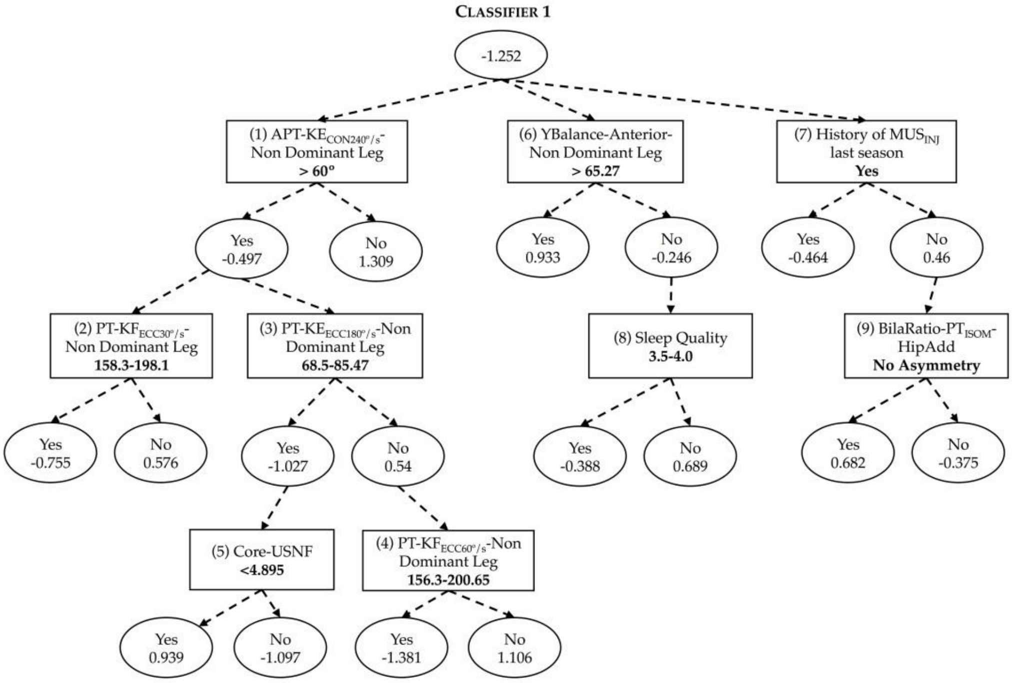
Bagging-based Ensembles

Abbr.	*Method*	*Short Description*
OBAG	OverBagging	Bagging with oversampling of the minority class.
UBAG	Underbagging	Bagging with undersampling of the majority class.
SBAG	SmoteBagging	Bagging where each bag´s Smote quantity varies

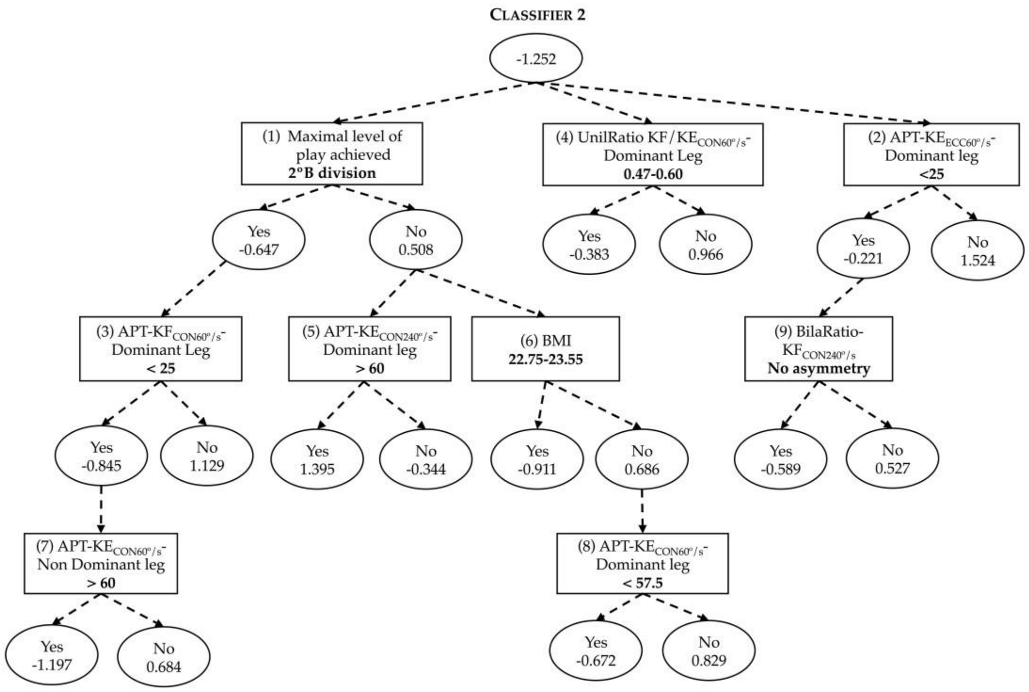
Ensembles with a cost-sensitive based classifier

Abbr.	*Method*	*Short Description*

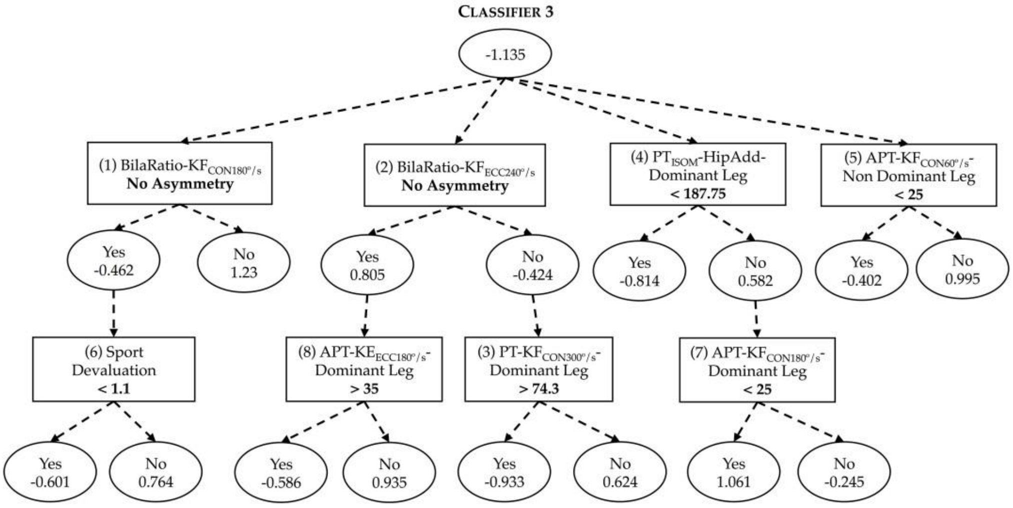
CS-SBAG	Cost sensitive SmoteBagging	SmoteBagging with an asymmetric classification cost matrix in the base classifier
CS.OBAG	Cost sensitive OverBagging	OverBagging with an asymmetric classification cost matrix in the base classifier
CS- UBAG	Cost sensitive UnderBagging	UnderBagging with an asymmetric classification cost matrix in the base classifier


CLASSIFIER 1
-1.252
(1) APT-KE$_{CON240°/s}$-Non Dominant Leg > 60º
Yes -0.497
No 1.309
(2) PT-KF$_{ECC30°/s}$-Non Dominant Leg 158.3-198.1
Yes -0.755
No 0.576
(3) PT-KE$_{ECC180°/s}$-Non Dominant Leg 68.5-85.47
Yes -1.027
No 0.54
(5) Core-USNF <4.895
Yes 0.939
No -1.097
(4) PT-KF$_{ECC60°/s}$-Non Dominant Leg 156.3-200.65
Yes -1.381
No 1.106
(6) YBalance-Anterior-Non Dominant Leg > 65.27
Yes 0.933
No -0.246
(8) Sleep Quality 3.5-4.0
Yes -0.388
No 0.689
(7) History of MUS$_{INJ}$ last season Yes
Yes -0.464
No 0.46
(9) BilaRatio-PT$_{ISOM}$-HipAdd No Asymmetry
Yes 0.682
No -0.375

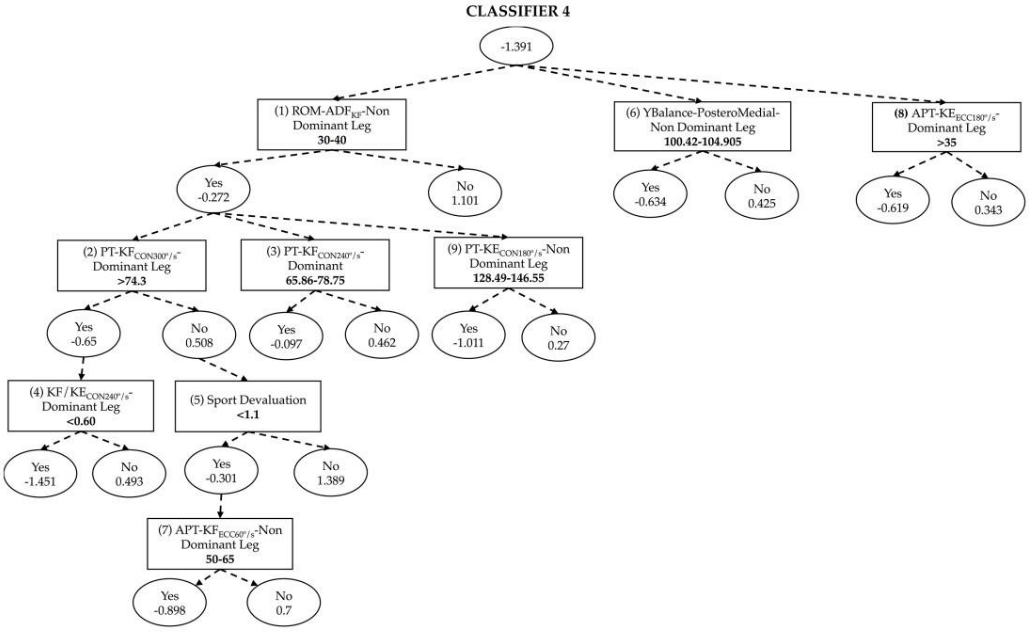
SDC 10: First classifier.
Graphical representation of the first classifier of the predictive model for muscle injuries.


CLASSIFIER 2
-1.252
(1) Maximal level of play achieved **2°B division**
Yes -0.647
No 0.508
(4) UnilRatio KF/KE$_{CON60°/s}$- Dominant Leg **0.47-0.60**
Yes -0.383
No 0.966
(2) APT-KE$_{ECC60°/s}$- Dominant leg **<25**
Yes -0.221
No 1.524
(3) APT-KF$_{CON60°/s}$- Dominant Leg **< 25**
Yes -0.845
No 1.129
(5) APT-KE$_{CON240°/s}$- Dominant leg **> 60**
Yes 1.395
No -0.344
(6) BMI **22.75-23.55**
Yes -0.911
No 0.686
(9) BilaRatio-KF$_{CON240°/s}$ **No asymmetry**
Yes -0.589
No 0.527
(7) APT-KE$_{CON60°/s}$- Non Dominant leg **> 60**
Yes -1.197
No 0.684
(8) APT-KE$_{CON60°/s}$- Dominant leg **< 57.5**
Yes -0.672
No 0.829

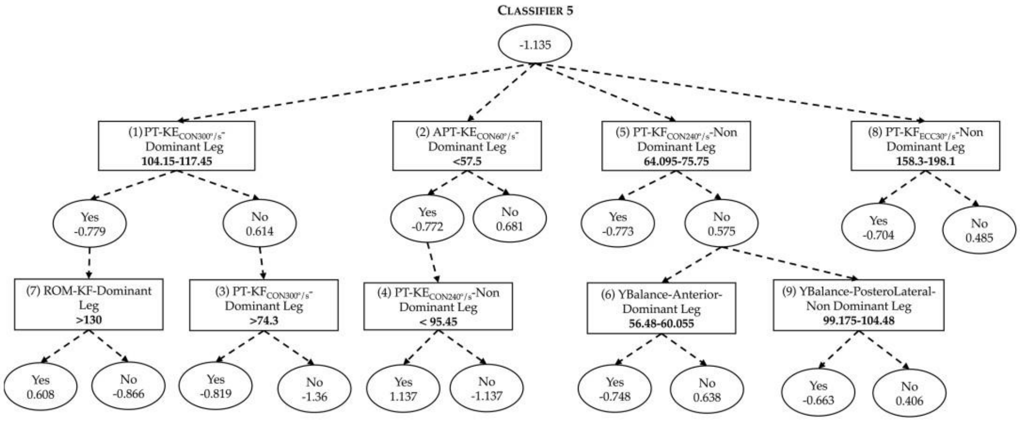
SDC 11: Second classifier.
Graphical representation of the second classifier of the predictive model for muscle injuries.


CLASSIFIER 3
-1.135
(1) BilaRatio-KF$_{CON180°/s}$ No Asymmetry
(2) BilaRatio-KF$_{ECC240°/s}$ No Asymmetry
(4) PT$_{ISOM}$-HipAdd-Dominant Leg < 187.75
(5) APT-KF$_{CON60°/s}$- Non Dominant Leg < 25
Yes -0.462
No 1.23
Yes 0.805
No -0.424
Yes -0.814
No 0.582
Yes -0.402
No 0.995
(6) Sport Devaluation < 1.1
(8) APT-KE$_{ECC180°/s}$- Dominant Leg > 35
(3) PT-KF$_{CON300°/s}$- Dominant Leg > 74.3
(7) APT-KF$_{CON180°/s}$- Dominant Leg < 25
Yes -0.601
No 0.764
Yes -0.586
No 0.935
Yes -0.933
No 0.624
Yes 1.061
No -0.245

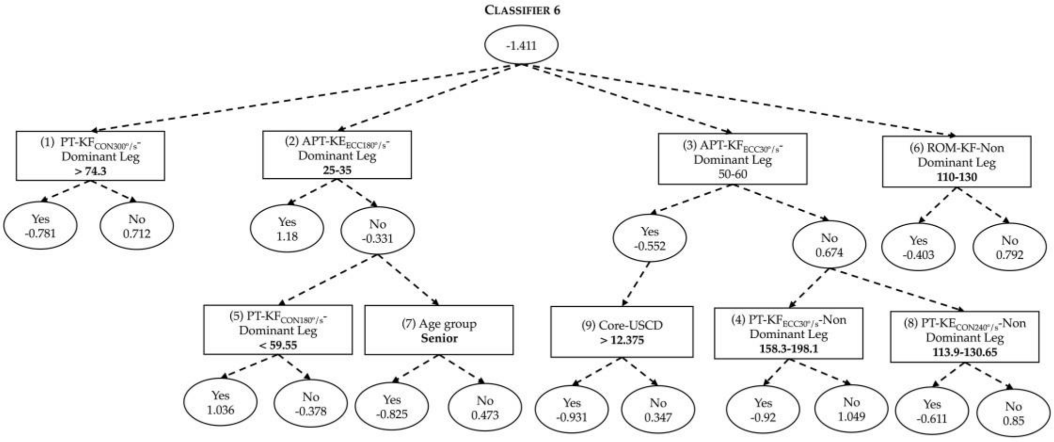
SDC 12: Third classifier.
Graphical representation of the third classifier of the predictive model for muscle injuries.


CLASSIFIER 4
-1.391
(1) ROM-ADF$_{KF}$-Non Dominant Leg 30-40
Yes -0.272
No 1.101
(6) YBalance-PosteroMedial-Non Dominant Leg 100.42-104.905
Yes -0.634
No 0.425
(8) APT-KE$_{ECC180°/s}$-Dominant Leg >35
Yes -0.619
No 0.343
(2) PT-KF$_{CON300°/s}$-Dominant Leg >74.3
Yes -0.65
No 0.508
(3) PT-KF$_{CON240°/s}$-Dominant 65.86-78.75
Yes -0.097
No 0.462
(9) PT-KE$_{CON180°/s}$-Non Dominant Leg 128.49-146.55
Yes -1.011
No 0.27
(4) KF/KE$_{CON240°/s}$-Dominant Leg <0.60
Yes -1.451
No 0.493
(5) Sport Devaluation <1.1
Yes -0.301
No 1.389
(7) APT-KF$_{ECC60°/s}$-Non Dominant Leg 50-65
Yes -0.898
No 0.7

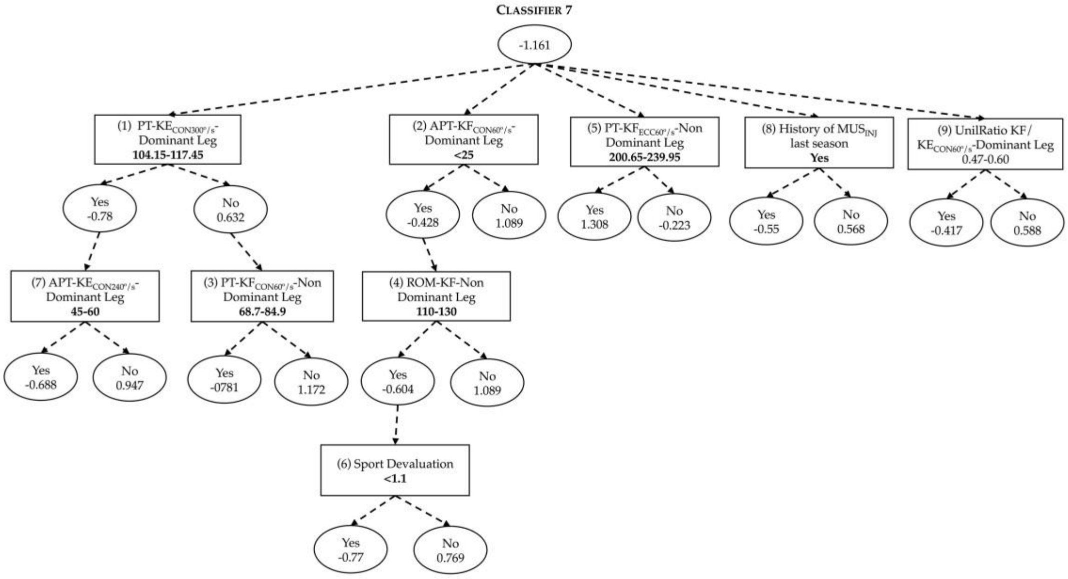
SDC 13: Fourth classifier.
Graphical representation of the fourth classifier of the predictive model for muscle injuries.


CLASSIFIER 5
-1.135
(1) PT-KE$_{CON300°/s}$- Dominant Leg 104.15-117.45
(2) APT-KE$_{CON60°/s}$- Dominant Leg <57.5
(5) PT-KF$_{CON240°/s}$-Non Dominant Leg 64.095-75.75
(8) PT-KF$_{ECC30°/s}$-Non Dominant Leg 158.3-198.1
Yes -0.779
No 0.614
Yes -0.772
No 0.681
Yes -0.773
No 0.575
Yes -0.704
No 0.485
(7) ROM-KF-Dominant Leg >130
(3) PT-KF$_{CON300°/s}$- Dominant Leg >74.3
(4) PT-KE$_{CON240°/s}$-Non Dominant Leg < 95.45
(6) YBalance-Anterior-Dominant Leg 56.48-60.055
(9) YBalance-PosteroLateral-Non Dominant Leg 99.175-104.48
Yes 0.608
No -0.866
Yes -0.819
No -1.36
Yes 1.137
No -1.137
Yes -0.748
No 0.638
Yes -0.663
No 0.406

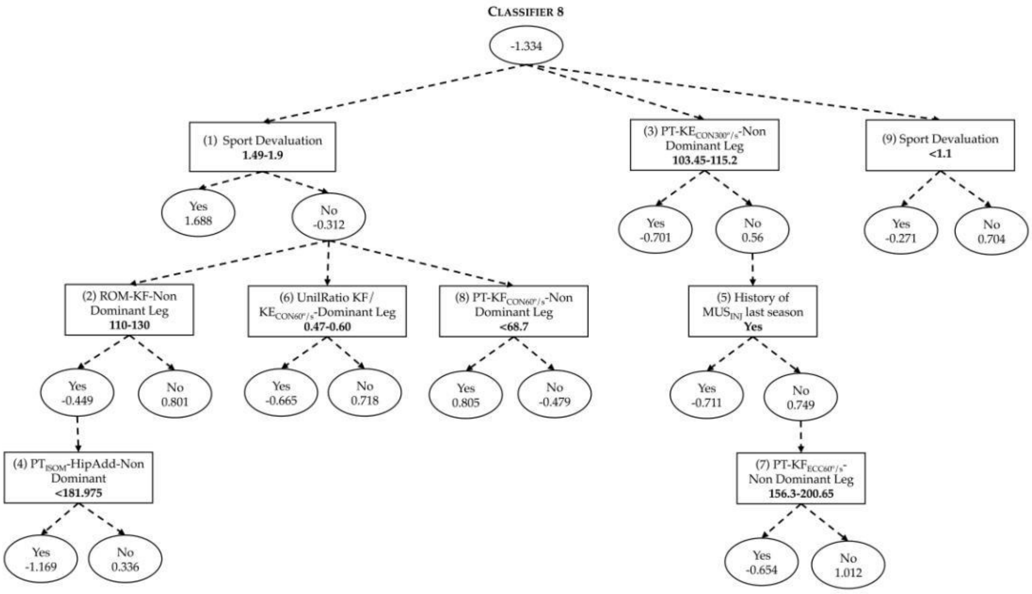
SDC 14: Fifth classifier.
Graphical representation of the fifth classifier of the predictive model for muscle injuries.



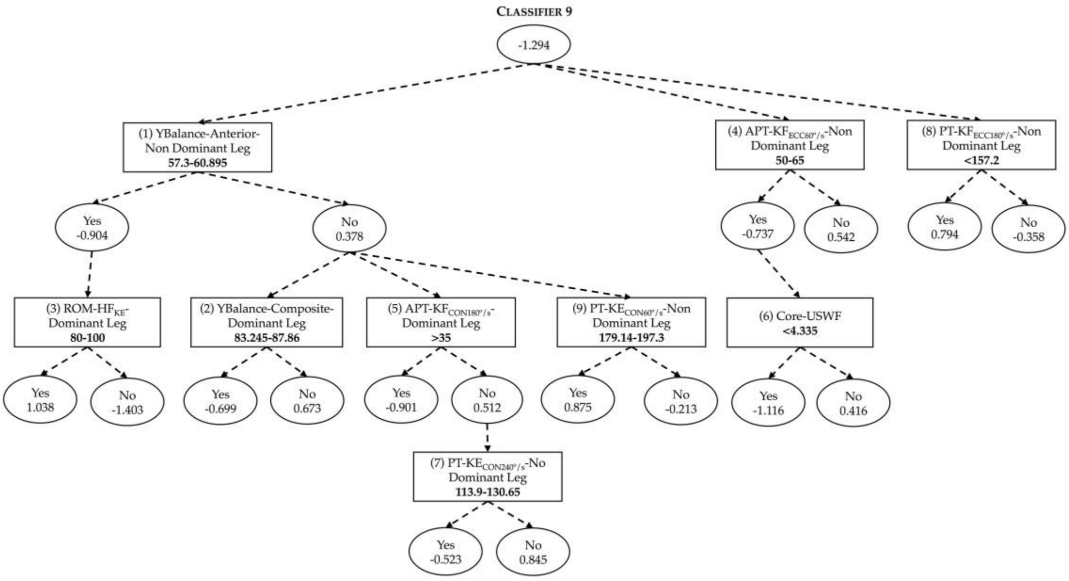
SDC 15: Sixth classifier.
Graphical representation of the sixth classifier of the predictive model for muscle injuries.


CLASSIFIER 7
-1.161
(1) PT-KE$_{CON300°/s}$-Dominant Leg 104.15-117.45
Yes -0.78
No 0.632
(7) APT-KE$_{CON240°/s}$-Dominant Leg 45-60
Yes -0.688
No 0.947
(3) PT-KF$_{CON60°/s}$-Non Dominant Leg 68.7-84.9
Yes -0781
No 1.172
(2) APT-KF$_{CON60°/s}$-Dominant Leg <25
Yes -0.428
No 1.089
(4) ROM-KF-Non Dominant Leg 110-130
Yes -0.604
No 1.089
(6) Sport Devaluation <1.1
Yes -0.77
No 0.769
(5) PT-KF$_{ECC60°/s}$-Non Dominant Leg 200.65-239.95
Yes 1.308
No -0.223
(8) History of MUS$_{INJ}$ last season Yes
Yes -0.55
No 0.568
(9) UnilRatio KF/ KE$_{CON60°/s}$-Dominant Leg 0.47-0.60
Yes -0.417
No 0.588

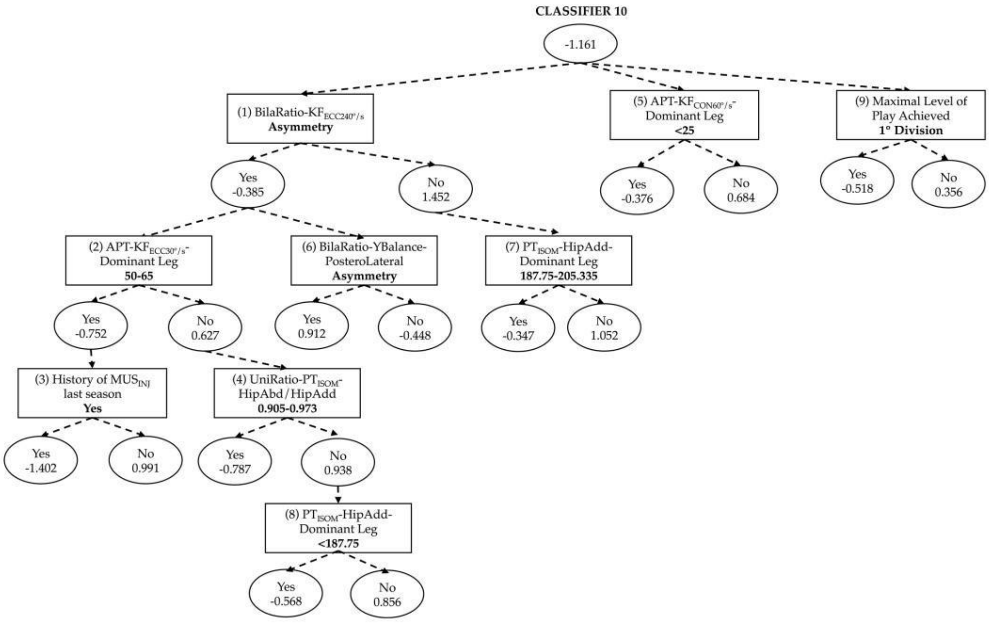
SDC 16: Seventh classifier.
Graphical representation of the seventh classifier of the predictive model for muscle injuries.


CLASSIFIER 8
-1.334
(1) Sport Devaluation 1.49-1.9
Yes 1.688
No -0.312
(2) ROM-KF-Non Dominant Leg 110-130
Yes -0.449
No 0.801
(4) PT_{ISOM}-HipAdd-Non Dominant <181.975
Yes -1.169
No 0.336
(6) UnilRatio KF/ $KE_{CON60°/s}$-Dominant Leg 0.47-0.60
Yes -0.665
No 0.718
(8) PT-$KF_{CON60°/s}$-Non Dominant Leg <68.7
Yes 0.805
No -0.479
(3) PT-$KE_{CON300°/s}$-Non Dominant Leg 103.45-115.2
Yes -0.701
No 0.56
(5) History of MUS_{INJ} last season Yes
Yes -0.711
No 0.749
(7) PT-$KF_{ECC60°/s}$- Non Dominant Leg 156.3-200.65
Yes -0.654
No 1.012
(9) Sport Devaluation <1.1
Yes -0.271
No 0.704

SDC 17: Eighth classifier.
Graphical representation of the eighth classifier of the predictive model for muscle injuries.


CLASSIFIER 9
-1.294
(1) YBalance-Anterior-Non Dominant Leg 57.3-60.895
Yes -0.904
No 0.378
(4) APT-KF$_{ECC60°/s}$-Non Dominant Leg 50-65
Yes -0.737
No 0.542
(8) PT-KF$_{ECC180°/s}$-Non Dominant Leg <157.2
Yes 0.794
No -0.358
(3) ROM-HF$_{KE}$-Dominant Leg 80-100
Yes 1.038
No -1.403
(2) YBalance-Composite-Dominant Leg 83.245-87.86
Yes -0.699
No 0.673
(5) APT-KF$_{CON180°/s}$-Dominant Leg >35
Yes -0.901
No 0.512
(7) PT-KE$_{CON240°/s}$-No Dominant Leg 113.9-130.65
Yes -0.523
No 0.845
(9) PT-KE$_{CON60°/s}$-Non Dominant Leg 179.14-197.3
Yes 0.875
No -0.213
(6) Core-USWF <4.335
Yes -1.116
No 0.416

SDC 18: Ninth classifier.
Graphical representation of the ninth classifier of the predictive model for muscle injuries.



SDC 19: Tenth classifier.
Graphical representation of the tenth classifier of the predictive model for muscle injuries.

Appendix 10: Risk factor measures included in the model for predicting MUS_{INJ} and the number of times that they appear in the classifiers, In bold are highlighted those that appear in four or more classifiers

Risk Factor	Nº of Classifiers
Personal measures	
Age group	1
History of MUS_{INJ} last season	4
Maximal level of play achieved	2
BMI	1
Psychological measures	
Sleep Quality	1
Sport Devaluation	5
Dynamic postural control measures	
YBalance-Anterior- Dominant Leg	1
YBalance-Anterior-Non Dominant Leg	2
YBalance-Composite-Dominant Leg	1
YBalance-PosteroLateral-Non Dominant Leg	1
YBalance-PosteroMedial-Non Dominant Leg	1
BilaRatio-YBalance-PosteroLateral	1
Isometric hip abduction and adduction strength measures	
BilaRatio-PT_{ISOM}-HipAdd	1
PT_{ISOM}-HipAdd-Dominant Leg	2
PT_{ISOM}-HipAdd-No Dominant	1
UniRatio-PT_{ISOM}-HipAbd/HipAdd	1

Lower extremity joints range of motion measures	
ROM-ADF$_{KF}$-Non Dominant Leg	1
ROM-HF$_{KE}$-Dominant Leg	1
ROM-KF-Dominant Leg	1
ROM-KF-Non Dominant Leg	3
Core stability measures	
Core-USNF	1
Core-USWF	1
Core-USCD	1
Isokinetic knee flexion and extension strength measures	
APT-KE$_{CON240°/s}$-Dominant leg	2
APT-KE$_{CON240°/s}$-Non Dominant Leg	1
APT-KE$_{CON60°/s}$-Dominant leg	2
APT-KE$_{CON60°/s}$-Non Dominant leg	1
APT-KE$_{ECC180°/s}$-Dominant Leg	3
APT-KE$_{ECC60°/s}$-Dominant leg	1
APT-KF$_{CON180°/s}$-Dominant Leg	2
APT-KF$_{CON60°/s}$-Dominant Leg	3
APT-KF$_{CON60°/s}$-Non Dominant Leg	1
APT-KF$_{ECC30°/s}$-Dominant Leg	2
APT-KF$_{ECC60°/s}$-Non Dominant Leg	2
BilaRatio-KF$_{CON180°/s}$	1
BilaRatio-KF$_{CON240°/s}$	1
BilaRatio-KF$_{ECC240°/s}$	2
PT-KE$_{CON180°/s}$-Non Dominant Leg	1

PT-KE$_{\mathrm{CON240°/s}}$-Non Dominant Leg	3
PT-KE$_{\mathrm{CON300°/s}}$-Dominant Leg	2
PT-KE$_{\mathrm{CON300°/s}}$-Non Dominant Leg	1
PT-KE$_{\mathrm{CON60°/s}}$-Non Dominant Leg	1
PT-KE$_{\mathrm{ECC180°/s}}$-Non Dominant Leg	1
PT-KF$_{\mathrm{CON180°/s}}$-Dominant Leg	1
PT-KF$_{\mathrm{CON240°/s}}$- Dominant	1
PT-KF$_{\mathrm{CON240°/s}}$-Non Dominant Leg	1
PT-KF$_{\mathrm{CON300°/s}}$-Dominant Leg	4
PT-KF$_{\mathrm{CON60°/s}}$-Non Dominant Leg	2
PT-KF$_{\mathrm{ECC180°/s}}$-Non Dominant Leg	1
PT-KF$_{\mathrm{ECC30°/s}}$-Non Dominant Leg	3
PT-KF$_{\mathrm{ECC60°/s}}$-Non Dominant Leg	3
UnilRatio KF/KE$_{\mathrm{CON60°/s}}$-Dominant Leg	3
UniRatio-KF/KE$_{\mathrm{CON240}}$-Dominant Leg	1

MUS$_{\mathrm{IN}}$: Muscle injury; BMI: body mass index; Bila: bilateral; Uni: unilateral; ISOM. Isometric; Add: adduction; Abd: abduction; ROM: range of motion; ADF: ankle dorsi-flexion; KE: knee extension; KF: knee flexion; HF: hip flexion; APT: angle of peak torque; ECC: eccentric; CON: concentric; PT: peak torque; s: seconds; $^{\circ}$: degree; USNF: unstable sitting without feedback; USWF: unstable sitting with feedback; USCD: unstable sitting while performing circular displacements with feedback