UNIVERSITÉ DE MONTPELLIER

Unité de Formation et de Recherche
en Sciences et Techniques des Activités Physiques et Sportives

Mémoire présenté en vue de l'obtention du Master 2 STAPS

Mention Entraînement et Optimisation de la Performance Sportive

Parcours Sciences et Techniques de l'Entraînement Physique

LA PRÉDICTION DES BLESSURES À PARTIR DES CHARGES D'ENTRAÎNEMENT

Présenté par

Emmanuel VALLANCE

Sous la direction de :

Nicolas SUTTON-CHARANI (MA, IMT Mines Alès)
Stéphane PERREY (PR, Univ. Montpellier)

Année universitaire 2018-2019

Table des matières

1. Introduction

De nos jours, le football est le sport le plus populaire au monde eu égard aux fans et au nombre de pratiquants qu'il regroupe à travers le monde. C'est une activité multifactorielle qui est intimement liée à plusieurs éléments de la performance qui sont d'ordres technique, tactique, psychologique, physique et physiologique (Hoff 2005). Sa pratique moderne actuelle descelle plusieurs spécificités à l'image de la rapidité du jeu et l'importance de remporter les différentes situations de duels. C'est en ce sens que Randers et al. en 2010 illustrent le fait que des outils actuels d'analyse tels que la vidéo et les GPS (i.e., un système de géolocalisation par satellite) peuvent fournir des données techniques et athlétiques précises sur l'activité des joueurs et des équipes en compétition. Ces outils permettent de vérifier la tendance énoncée ci-dessus : l'évolution vers un jeu plus rythmé, des joueurs de plus en plus « physiques ». Des changements notoires sur la possession de balle ont été flagrants, comme par exemple, le jeu du football Espagnol (Carney et al. 2010), celui-ci principalement basé sur la maîtrise de la possession en conservation de balle. En 2014, une étude de Wallace et Norton sur l'évolution du jeu lors des finales de Coupe du Monde entre 1966 et 2010 confirme que le jeu est en perpétuelle amélioration aussi bien dans la rapidité, la vitesse et l'efficacité. Fort de cette conclusion, il semble nécessaire de caractériser plus précisément l'activité physique du joueur de football professionnel.

1.1. Analyse de l'activité du footballeur

Il existe de nombreuses études qui ont déjà examinées l'activité Football ; plusieurs paramètres y sont mis en lumière comme les distances parcourues à différentes vitesses, les accélérations, les décélérations et la vitesse maximale (Vigne et al. 2010). Les sprints, en particulier, sont souvent considérés comme un élément majeur de la performance, mais finalement ils ne représentent que seulement 10 % de la distance totale parcourue au cours des matchs (Carling et al. 2008). Si l'on s'intéresse à la distance parcourue en match, en moyenne, les footballeurs professionnels réalisent une distance de 9 à 12 kilomètres (Di Salvo et al. 2007; Vigne et al. 2010). De plus, de nombreux paramètres vont venir influencer la performance comme par exemple : le lieu du match, la période, le niveau de l'adversaire, le

classement et bien d'autres (Felipe and Go 2018). Toutes ces données permettraient d'analyser de façon plus précise l'activité des joueurs en compétition et à l'entraînement de façon à influencer la gestion de la charge d'entraînement, l'entrainement individualisé, et la prévention des blessures au bénéfice de la performance (Pettersen et al. 2018).

1.2. Blessures dans le football

Selon une étude récente (Villa et al. 2019), l'incidence générale des blessures dans le football masculin élite varie de 2,48 à 9,4 blessures par 1 000 heures d'exposition. Ces auteurs ont également montré que le risque de blessure est plus grand lors des matchs. Les dernières études épidémiologiques ont mis en lumière l'augmentation des blessures sur les seize dernières années tout en soulignant que les incidents musculaires en étaient la principale cause (Jones et al., 2018). Les blessures étant omniprésentes dans ce sport de contact complexe (Gómez-Piqueras et al. 2017), il existe plusieurs facteurs de risque tels que le nombre de matchs joués, une accumulation de fatigue induite par la charge de travail suite aux séances d'entraînement, etc... Il est donc primordial d'estimer les charges de travail des entraînements et des matchs dans le football.

1.3. Contrôle de la charge d'entraînement

La quantification de la charge d'entraînement est indispensable de façon à évaluer la conformité entre les exercices proposés et ceux réalisés. En outre la charge d'entraînement varie en fonction des objectifs annuels. La réponse individuelle de l'entraînement (charge interne) à un programme imposé donné (charge externe) peut entraîner une différence entre les joueurs et, par conséquent, l'individualisation de l'entraînement peut être problématique (Casamichana et al. 2013). Concernant la charge interne, plusieurs méthodes sont utilisées pour la quantifier comme par exemple l'utilisation de questionnaire de perception de l'effort (Impellizzeri et al. 2004) ou bien la mesure de la fréquence cardiaque (Borresen, Ian Lambert, and Lambert 2009). Pour la charge externe, les outils modernes d'analyse tels que la vidéo et les GPS sont en mesure de fournir des données précises et reproductibles sur les déplacements et mouvements effectués (Randers et al. 2010). Par conséquent, l'utilisation de méthodes fiables

d'évaluation des charges est primordiale, car des variations lors de l'entraînement peuvent entraîner des inadaptations et des blessures chez le footballeur (Gabbett 2016). De plus, au fur et à mesure que le volume de données collectées par ces outils augmente, la complexité du problème de leur gestion l'est également. Il devient alors nécessaire de disposer de méthodes appropriées pour exploiter les données nécessaires à la quantification de la charge d'entraînement et des potentiels liens avec le risque de blessures.

1.4. Approches d'apprentissage automatique

Désormais, il est admis que les méthodes d'apprentissage automatique (machine learning) appliquées au sport constituent une possible aide au diagnostic au service de la performance mais sont pour le moment peu présentes dans les dernières études scientifiques. Pourtant, ces méthodes sont capables d'aider les chercheurs et les analystes de la performance dans le but d'étudier par exemple les équipes de football sur le plan tactique (Santos et al. 2018). Cette étude a permis de mettre en lumière des premiers résultats sur l'analyse de l'apprentissage du jeu et de ses conséquences en fonction des statistiques physiques, des relations inter-joueurs permettant une analyse spatio-temporelle précise et multivariée du jeu. Grâce à l'outil développé, l'analyste peut en temps réel prédire différents résultats en fonction de différentes combinaisons de variables. Ces auteurs proposent également comme perspective l'intégration de données supplémentaires comme par exemple le positionnement du ballon, les données biomécaniques de l'athlète de façon à mieux « nourrir » les modèles d'apprentissage. Ces nouvelles ouvertures permettent aujourd'hui de voir apparaître le « machine learning » dans le monde du sport, comme par exemple en natation (Męyk and Unold 2011) qui, à l'aide d'un algorithme complexe, modélise les charges d'entraînements et permet de prédire les éventuelles influences sur la performance des athlètes. L'objectif était de prédire le comportement d'un nageur en fonction d'une modification du microcycle effectué la veille grâce à une modélisation *a priori* de l'entraînement du sportif. De nombreuses études (Barros et al. 2007; Di Salvo et al. 2007, 2009) ont déjà examiné l'activité technique, tactique et physique des joueurs de football et leurs effets sur les risques de blessures, ainsi que les résultats suite à la modification des charges et des dynamiques de travail. Partant de ces différents constats, il semblerait que la

littérature scientifique ne se soit pas penchée sur la prédiction des blessures à partir des charges internes et externes dans les sports collectifs.

Par conséquent, au vu des données recueillies dans le football moderne, il apparaît pertinent d'appliquer des méthodes d'apprentissage sur un ensemble de variables pour fournir davantage d'informations aux professionnels de l'entraînement. Nous émettons l'hypothèse principale que l'utilisation d'une méthode par apprentissage serait en mesure d'informer sur des risques de blessures en prenant en compte les évolutions des charges d'entraînement sur une semaine et un mésocycle de travail. Ces premières informations permettront de guider la programmation et l'individualisation des entrainements de façon à réduire le risque de blessure.

2. Méthodes et matériels

2.1. Participants

Vingt-cinq joueurs (âge : 28,7±6,2 ans ; taille : 178,1±4,2 cm ; masse corporelle : 76,9±9,2 kg) d'une même équipe de Ligue 2 française, ont été observés durant 245 séances d'entraînement, 38 matchs de Domino's Ligue 2, 2 matchs de Coupe de la Ligue et 3 matchs Coupe de France sur la saison 2017-2018.

2.2. Design expérimental

Plusieurs types de données d'un club professionnel de football ont été recueillies auprès des joueurs pendant les compétitions officielles, les matchs de préparation d'avant-saison, avant et après les séances d'entraînements. La saison complète a été analysée sur une période allant de Juin 2017 à Mai 2018 en prenant en compte les différentes coupures entre ces périodes, les trêves internationales et la trêve hivernale. Un premier jeu de données analysé concerne l'activité du joueur, récoltées à l'aide d'un système de « tracking GPS », traitées par ordinateur pour permettre une analyse en temps réel ou bien après la séance. Ce premier jeu de données reflète ce que l'on appelle la charge d'entraînement externe, c'est-à-dire qu'on analyse le travail physique objectif réalisé par l'athlète et sa capacité de performance d'un point de vue purement descriptif avec des valeurs telles que la vitesse, l'accélération, la distance parcourue etc. Un deuxième jeu de données concerne

deux questionnaires subjectifs remplis de façon journalière par les joueurs avant et après les séances d'entrainement uniquement. Le premier questionnaire, saisi avant la séance, comporte plusieurs questions sur la qualité du sommeil, la fatigue, le pourcentage de forme, l'humeur, la localisation ou non d'une douleur et l'éventuel degré d'inquiétude par rapport à celle-ci, enfin la présence d'une maladie. Le deuxième questionnaire, pour l'après séance collectait la satisfaction personnelle, le plaisir occasionné pendant l'entraînement et l'intensité de l'effort retranscrite par l'utilisation de l'échelle RPE pour rate of perceived exertion (Herman et al. 2006). Ce deuxième jeu de données reflète la charge interne, c'est-à-dire qu'on analyse la réponse psychologique et physiologique de la charge externe ainsi que d'autres facteurs environnementaux à l'aide de mesures subjectives. Il est important de noter que le monitoring de la charge d'entrainement est un enjeu majeur.

2.3. Analyse des données

L'évaluation de l'activité physique des joueurs a été effectuée grâce au système GPS Catapult (Optimeye S5, Catapult Innovations, Australie). Pendant chaque session, les joueurs portaient un appareil GPS 10 Hz contenant un accéléromètre triaxial 100 Hz et un gyroscope.
Les paramètres physiques individuels enregistrés étaient : la vitesse maximale, la distance totale parcourue, les accélérations à plus de 2 m/s/s, les décélérations à -2 m/s/s. Plus précisément, en se basant sur la littérature dédiée (Rampinini et al. 2007; Di Salvo et al. 2009) les variables suivantes ont été analysées : la distance totale parcourue dans chaque zone de vitesses spécifiques à l'activité (0-6 km/h, 6-15 km/h, 15-20 km/h, 20-25km/h, > 25 km/h). L'évolution du paramètre spécifique au GPS Catapult, le PlayerLoad (indice de fatigue mécanique de l'athlète) était également collectée (Barrett, Midgley, and Lovell 2014).

L'approche suivie pour cette étude a consisté (i) à construire un jeu de données mixtes agrégeant les données GPS, les données des questionnaires ainsi que les données de blessures, puis (ii) apprendre différents modèles prédictifs de manière à être en mesure de prédire la survenue de blessure (à 1 semaine et 1 mois) à partir des variables GPS et questionnaire. Les modèles ainsi construits peuvent donc servir d'alerte pour tout nouvel entraînement pour lequel le modèle prédirait une blessure et être exploités comme aide à la planification d'entraînement.

La jointure entre les 3 types de données a été réalisée grâce à la variable 'date' qui constitue la clé primaire du jeu de données liant les 3 types de données.
Les modèles d'apprentissage qui ont été mis en œuvre sont les suivants :

- L'analyse discriminante linéaire (*LDA*) (Fisher 1936; McLachlan 2004)
  Le classifieur de Bayes naïf (*naiveBayes*) (Maron 1961; Rish 2001)
- L'arbre de décision (*tree*) (Breiman, Friedman, & Stone 1984)
- La forêt aléatoire (*forest*) (Breiman 2001)
- Le Support Vecteur Machine (*SVM*) (Boser, Guyon, & Vapnik 1992)
- Le réseau de neurones à 1 couche cachée (*nnet*) (Werbos, 1975) (McCulloch and Pitts 1943; Rosenblatt 1958; Werbos 1974)

2.3.1. Performances prédictives

Différentes complexités mathématiques/calculatoires ont été considérées pour les modèles d'arbres et de forêts en faisant varier le paramètre de gain d'information 'cp' pour les arbres (*i.e.,* qui mesure à quel point chaque coupure de l'arbre considéré permet une séparation des classes entre occurrence de blessure et non-occurrence, et la taille, *i.e.* le nombre d'arbres des forêts. Les termes 'tree0', 'tree0.01' et 'tree0.02' désignent respectivement des arbres de décision pour cp = 0, cp = 0,01 et cp = 0,02. De la même manière les termes 'forest500' et 'forest2000' désignent des forêts aléatoires respectivement à 500 et 2000 arbres. Le terme 'inNode' désigne le nœud initial d'un arbre ; les prédictions seront dans ce cas systématiquement fidèles aux proportions de l'échantillon d'apprentissage (*e.g.,* si on a 75% de non-blessures et 25% de blessures, inNode prédira systématiquement des non-blessures). Ce modèle sert de référence à *dépasser*.
Une fois le jeu de données construit, tous ces modèles ont été évalués par 30 validations croisées à 10 couches à l'aide de 4 mesures d'efficacité prédictive (voir

Table 1. Mesures binaires des performances prédictives) selon 2 problème prédictifs : la prédiction de blessure à 1 semaine et à 1 mois.
S'agissant d'un problème prédictif de classification binaire, différentes métriques sont envisageables pour mesurer les performances des modèles. La plupart d'entre elles tiennent compte de la notion de vrais/faux positifs/négatifs, *i.e.,* lorsque qu'une blessure est prédite (*positif*), correspond-t-elle à une blessure effective

(*vrai positif)* où à une non-blessure et donc à une erreur de prédiction (*faux positif*), et de même pour le cas de non-blessure prédite (*vrais/faux négatif*) car les performances des classifieurs binaires ne sont pas symétriques en général. Les 4 mesures d'efficacité prédictives suivantes ont été considérées :

Table 1. Mesures binaires des performances prédictives

accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
precision	$\frac{TP}{TP+FP}$
recall	$\frac{TP}{TP+FN}$
Area under the curve ROC (AUC)	rate (TP/FP)

où TP, TN, FP et FN désignent respectivement les vrais positifs, les vrais négatifs, les faux positifs et les vrais négatifs.

3. Résultats

Initialement les modèles SVM et LDA ont été testés pour obtenir le potentiel prédictif des données et pour les comparer aux modèles classiques. D'après les figures 1 et 2, les modèles *LDA*, *naiveBayes*, *nnet2* et *nnet10* obtiennent de mauvaises performances prédictives. Ils sont même souvent en dessous du niveau de *inNode* en termes d'accuracy. Les arbres et les forêts semblent globalement être de meilleurs classifieurs, surtout les arbres de complexité maximale *tree0*. Les SVM obtiennent des bons niveaux de précision mais des mauvais niveaux de rappels. Cela signifie que lorsque les modèles SVM prédisent des blessures ils sont considérés comme fiables, mais quand ils prédisent une absence de blessure il faut prendre du recul par rapport à leur prédiction. Les figures 1, 2 et 3, sont des boxplots (boîtes à moustache), i.e. représentations graphiques de la distribution des variables (en faisant apparaître le premier quartile, la médiane et le troisième quartile) avec aussi mise en relief de la formes des distributions au travers de l'option de graphique violons (violon plots).

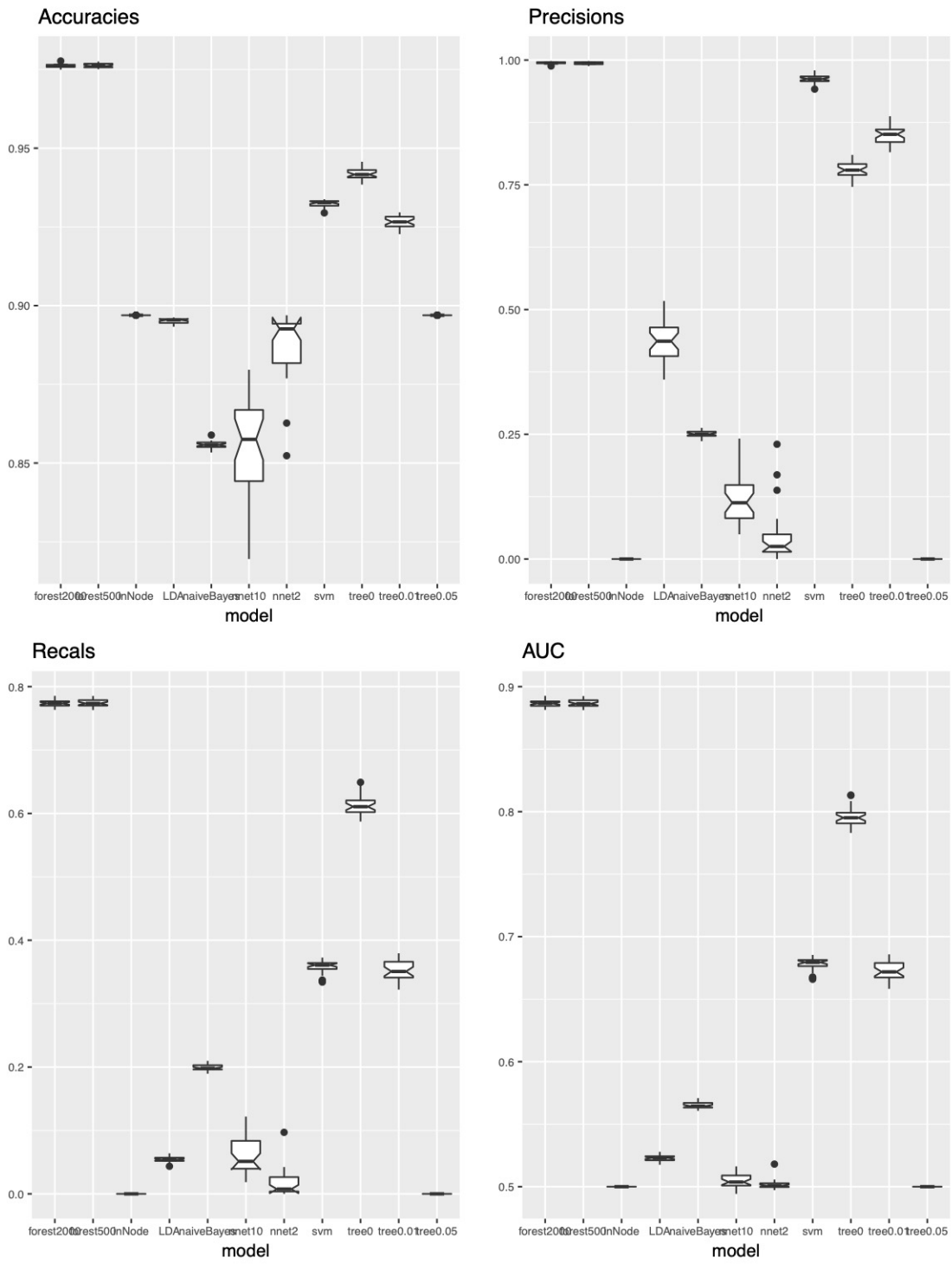

Figure 1. Evaluation des modèles de prédiction de blessure à une semaine

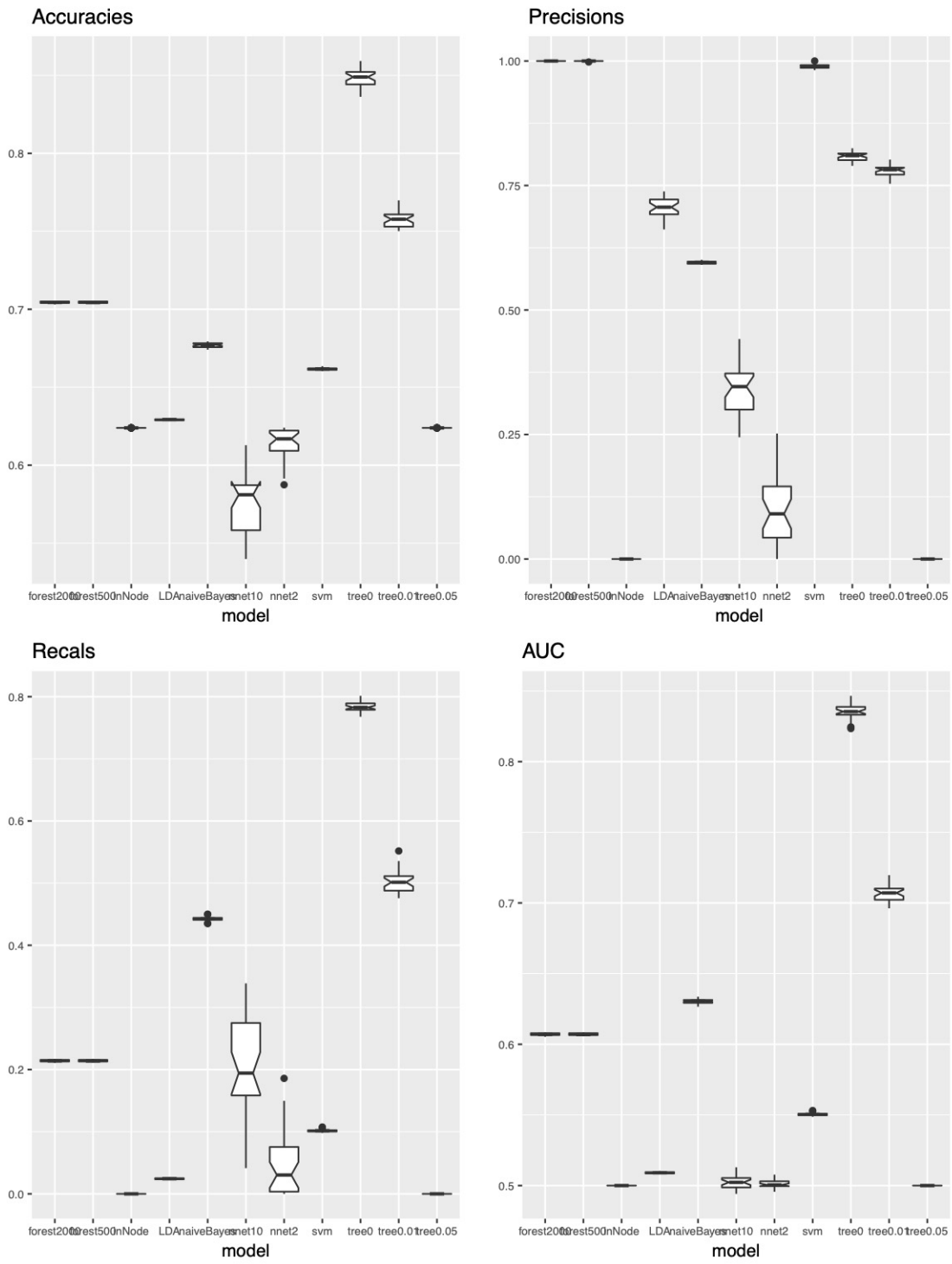

Figure 2. Evaluation des modèles de prédiction de blessure à un mois

La figure 3 représente la sensibilité des performances prédictives à la taille des forêts aléatoires. On observe que les forêts de grandes taille sont de meilleurs classifieurs que les forêts de petite taille mais que des forêts de 10 arbres semblent être un bon compromis efficacité/complexité (et donc en temps de calcul). On remarque aussi un phénomène assez surprenant : les forêts composées d'un nombre impair d'arbres obtiennent des meilleurs *recall* et *AUC* que celles contenant un nombre d'arbres pairs. Ceci s'explique probablement par les modèles d'agrégation des prédictions des arbres des forêts qui correspondent généralement à des votes où des égalités pouvant avoir des effets néfastes (car aboutissant à des prédictions finales plus ou moins aléatoires). Ce phénomène s'estompe néanmoins quand la taille des forêts augmente.

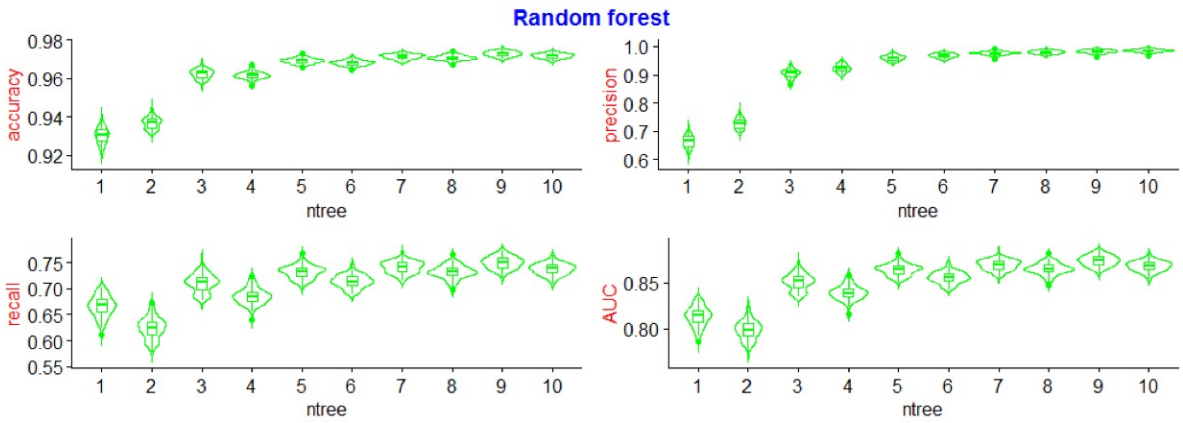

Figure 3. Effet de la taille des forêts aléatoires sur les performances prédictives

3.1.1. Explication prédictive

De manière à tirer le maximum d'informations possibles des modèles prédictifs mis en œuvre, 2 types de représentation sont ici présentées :

- Les graphiques correspondant aux arbres de décision
- Les poids des variables prédictives obtenus par arbres et forêts

Dans les 2 cas, les modèles ont été appris sur l'ensemble du jeu de données.

Les figures 5 et 6 représentent des arbres de décision obtenus pour la prédiction de blessures à 1 semaine et à 1 mois, avec des paramètres de complexité *cp* égaux à 0,01 et 0,014 (ces valeurs ont été choisies de manière à obtenir des arbres le plus lisible possible à l'œil nu), respectivement. On remarque que les variables possédant le meilleur potentiel prédictif (*i.e.* apparaissant le plus haut dans les arbres) sont les questions relatives à la bonne forme, la satisfaction et l'intensité ressenties en entraînement, ainsi que l'humeur. On note aussi que ces variables sont généra-

lement associées à des périodes de 2 semaines qui semblent donc correspondre à un bon laps de temps idéal pour calculer les moyennes des charges internes pour la prédiction de blessures.

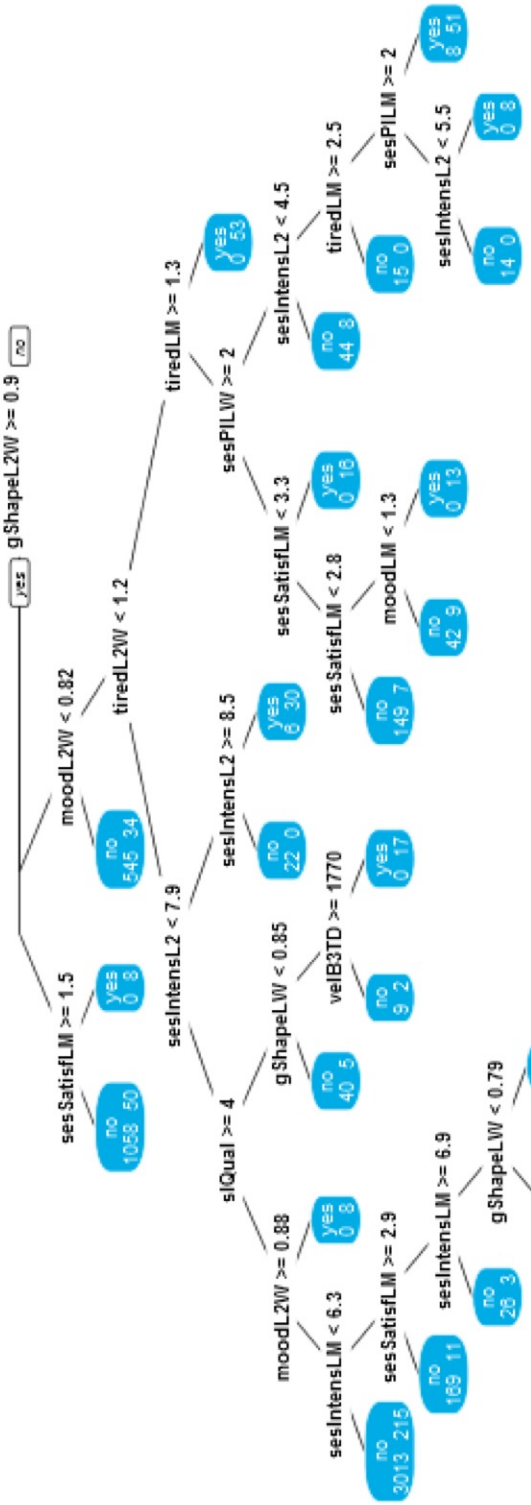

Figure 4. Arbre de décision pour la prédiction de blessure à une semaine

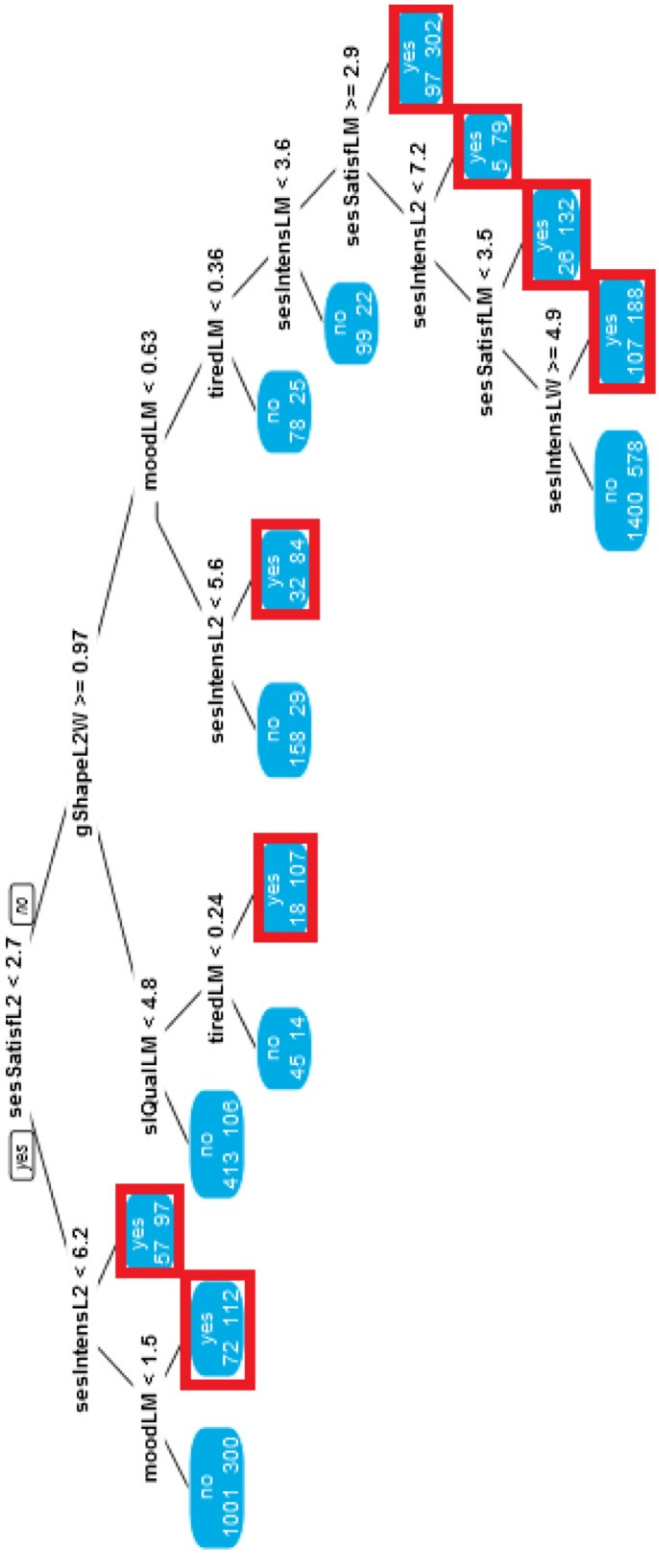

Figure 5. Arbre de décision pour la prédiction de blessure à un mois

D’après les figures 7 et 8, la période idéale à prendre en compte pour les calculs de charge interne moyenne est située entre 2 et 4 semaines. On observe également que la variable fatigue semble assez déterminante pour la prédiction de blessures. Enfin, seule l’exposition à vitesse maximale pendant un entraînement semble pertinent pour la prédiction de blessure.

Pour les figures 5 et 6 (les arbres), et les figures 7 et 8, le laps de temps idéal à utiliser pour prédire est 1 mois (Figures 7 et 8), mais que pour interpréter et comprendre les prédictions, des moyennes calculées sur 2 semaines sont préférables (Figures 5 et 6).

Les variables subjectives possèdent un potentiel prédictif/explicatif très important (en comparaison des variables objectives) mais elles sont plus coûteuse, i.e. c'est un peu compliqué de faire remplir des questionnaires à tous les joueurs à chaque entraînement.

Par exemple, si l’on s’intéresse aux feuilles des arbres (Figures 5 et 6), lorsque la satisfaction est > à 2,7, la forme moyenne des 2 dernières semaines est > 0,97, la qualité de sommeil du dernier mois est > à 4,8 et la fatigue du dernier mois est > 0,24, 107 joueurs sur 125 ont été blessés, on aura donc une probabilité de blessure égale à 107/125=0,856 dans cette configuration précise.

On observe aussi que pour des prédictions à 1 semaine (Figure 5, court terme), les forêts sont légèrement plus performantes que les arbres, alors qu'à 1 mois (Figure 6, moyen terme) les arbres ont un meilleur rappel (recall) et auront donc moins tendance à oublier des joueurs "en risque de blessure".

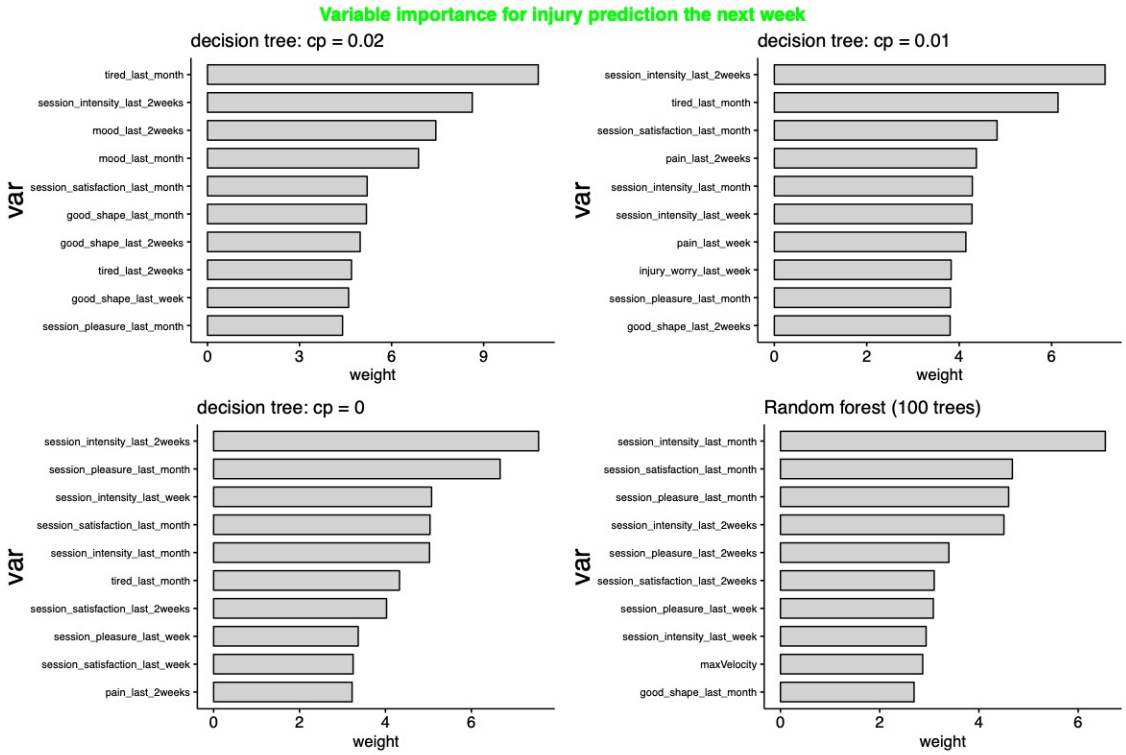

Figure 6. Importances relatives des variables prédictives à une semaine

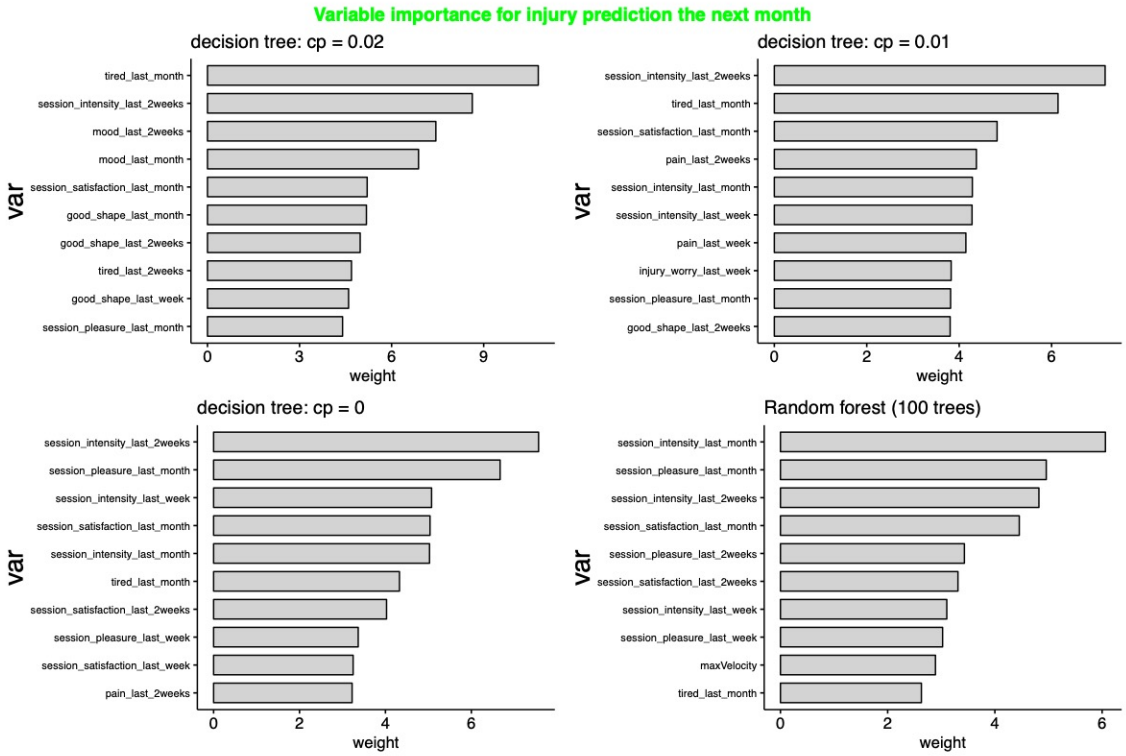

Figure 7. Importances relatives des variables prédictives à un mois

4. Discussion :

Au vu des résultats de cette étude pilote, quelques faits notoires sont à noter. Tout d'abord, avec l'utilisation de l'arbre de décision à 1 semaine et à 1 mois, on obtient une précision des résultats prédictifs de blessures très fiables, de l'ordre de 94% et de 74%, respectivement. A titre de comparaison, les arbres de décision exploités dans l'étude de Rossi et al. (2018) permettaient de déceler environ 80% des blessures sur l'échantillon analysé avec une précision approximative de 50%. De ce fait, l'algorithme présent dans notre méthode d'apprentissage automatique serait en mesure de classifier avec plus de précision les joueurs dits à risque concernant l'apparition de blessures et ainsi être capable de continuer de performer sans être perturbé par de « fausses alertes ». La précision de cette arbre, notamment à 1 semaine, qui diffère par rapport à Rossi et al. (2018), est permise par la mise en relation des datas GPS et des questionnaires subjectifs via l'algorithme, ce qui justifie l'apport de ce mémoire au vu de la littérature dans le monde sportif.

Un autre point qui valide le du choix des arbres, est que ce modèle gère naturellement les données manquantes (contrairement aux forêts), en partant du fait que les arbres proposent le meilleur support explicatif (un graphique apporte plus d'information que de simples pourcentages d'importance pour chaque variable car on peut y observer le sens d'action des variables et la manière dont elles influent ensemble la prédiction). Quand on veut comprendre pourquoi il vaut mieux planifier un entraînement de telle ou telle manière on utilisera plutôt des arbres. On notera que dans ce cas, les questionnaires subjectifs sont très précieux même lorsqu'ils ne sont remplis que par certains joueurs à quelques entraînements. Cependant, pour vraiment prédire et alerter, les forêts sont préférables à court terme mais nécessitent plus de prétraitement (notamment des imputations de valeurs manquantes pour les questionnaires non-remplis). A moyen terme les arbres seront donc privilégiés (en terme explicatif comme prédictif).

De plus, l'intérêt de cette étude réside dans le couplage de la méthode d'apprentissage automatique et des variations de la charge d'entraînement (interne et externe). On remarque ainsi que dès l'utilisation des arbres de décision prédictifs de blessures, les premiers nœuds sont presque toujours associés à des variables

subjectives. On peut donc émettre l'hypothèse qu'avec ces données et cet échantillon dans cette situation précise, la charge interne serait un facteur déterminant dans la prédiction de blessures. En d'autres termes plus pratiques, il serait primordial pour chaque entraineur de prêter une attention particulière envers les ressentis des athlètes avant et après les séances d'entraînement dans l'unique but de prévenir l'apparition des blessures.

Cependant dans une étude pilote certaines limites résident, mais sont en réalité de potentielles sources d'amélioration. De ce fait une taille de l'échantillon plus grande, s'étendant à plusieurs équipes avec des stratégies d'entraînement différentes, permettrait de tirer des conclusions plus générales concernant ce modèle prédictif de blessures. Également, les données GPS et questionnaires récoltés peuvent également être sources de progrès. En ce qui concerne les questionnaires remplis, l'influence d'une assiduité plus accrue dans l'utilisation de ces derniers par les joueurs serait fondamentale à observer. En ce qui concerne les données GPS, ces dernières sont présentes sous une forme moyennée par rapport à leur fréquence d'acquisition initiale de 10 Hz. Dans la course à la précision, il serait intéressant d'observer les conséquences en utilisant toutes les valeurs acquises sur cette fréquence. Aussi, de par la différence entre les joueurs il serait très intéressant d'individualiser les variables relatives à la charge externe (données extraites des GPS), en calculant par exemple des seuils de vitesse et d'accélération propres à chaque joueur au-delà desquels des blessures peuvent survenir. De cette manière le potentiel prédictif des variables GPS pourrait s'en trouver grandement augmenté, et pourrait avoir une influence sur les stratégies d'entraînement mises en place par les entraîneurs.

5. Conclusion :

L'objectif de cette étude pilote était de se pencher sur la question de l'utilisation des méthodes d'apprentissage automatique dans la prédiction des blessures en fonction des charges internes et externes de l'athlète. Les résultats de cette étude montrent qu'en fonction de la complexité des arbres de décisions choisis, la précision de la prédiction de la blessure est très proche des 100% notamment celle avec l'arbre de décision à 1 semaine.

De plus, il apparaît que les variables subjectives (i.e., charge interne) du questionnaire d'avant-séance (comme la qualité du sommeil, la fatigue, le pourcentage de forme, l'humeur) ainsi que le ressenti d'après-séance (RPE, satisfaction personnelle et le plaisir occasionné) se révèlent être un facteur déterminant dans l'apparition des blessures.

Enfin, bien que les résultats préliminaires de ce mémoire semblent encourageants et pertinents, des recherches futures avec l'augmentation de l'échantillon en touchant plusieurs équipes d'un même championnat peuvent fournir assez de données afin de passer de conclusions spécifiques à des conclusions générales concernant les méthodes d'apprentissage automatique.

6. Les points clés et applications pratiques :

Les intérêts des résultats apportés par ce mémoire sont :

- Illustrer que l'utilisation de la méthode d'apprentissage automatique peut être un nouvel outil d'analyse parmi ceux présents dans l'environnement de l'entraînement. La pertinence des résultats préliminaires de cette étude dans cet environnement pourrait être complémentaire à d'autres usages et routines actuels.

- Mettre en exergue que l'accouplage de données (subjectives et objectives) avec des algorithmes peut être considéré comme prédicteur de blessures. Les variables subjectives faisant référence à l'individu dans sa globalité (ressenti, bien-être physique et mental) semblent être prépondérantes à analyser dans le monde de l'entraînement.

7. Bibliographie

Barrett, Steve, Adrian Midgley, and Ric Lovell. 2014. "PlayerLoad™: Reliability, Convergent Validity, and Influence of Unit Position during Treadmill Running." *International Journal of Sports Physiology and Performance* 9(6): 945–52.

Barros, Ricardo M L et al. 2007. "Analysis of the Distances Covered by First Division Brazilian Soccer Players Obtained with an Automatic Tracking Method." *Journal of Sports Science and Medicine* 6(2): 233–42.

Borresen, Jill, Michael Ian Lambert, and Michael Ian Lambert. 2009. "The Quantification of Training Load, the Training Response and the Effect on Performance." *Sports Medicine* 39(9): 779–95.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. "A Training Algorithm for Optimal Margin Classifiers." Proceedings of the fifth annual workshop on Computational learning theory : 144–52.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1): 5–32.

Carling, Christopher, Jonathan Bloomfield, Lee Nelsen, and Thomas Reilly. 2008. "The Role of Motion Analysis in Elite Soccer: Contemporary Performance Measurement Techniques and Work Rate Data." *Sports Medicine* 38(10): 839–62.

Carney, Dana R. et al. 2010. "Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer." *Journal of Nonverbal Behavior* 9(1): 278–90.

Casamichana, David et al. 2013. "Relationship between Indicators of Training Load in Soccer Players." *Journal of Strength and Conditioning Research* 27(2): 369–74.

Felipe, Luis, and Santiago Go. 2018. "Influence of Contextual Variables and the Pressure to Keep Category on Physical Match Performance in Soccer Players."*PloS one*: 1–10.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7(2): 179–88.

Gabbett, Tim J. 2016. "The Training-Injury Prevention Paradox: Should Athletes Be Training Smarter and Harder?" *British Journal of Sports Medicine* 50(5): 273–80.

Gómez-Piqueras, Pedro, Sixto González-Víllora, María Sainz de Baranda Andújar, and Onofre Contreras-Jordán. 2017. "Functional Assessment and Injury Risk in a Professional Soccer Team." *Sports* 5(1): 9.

Herman, L et al. 2006. "Validity and Reliability of the Session RPE Method for Monitoring Exercise Training Intensity." *South African Journal of Sports Medicine* 18(1): 14.

Hoff, Jan. 2005. "Training and Testing Physical Capacities for Elite Soccer Players." *Journal of Sports Sciences* 23(6): 573–82.

Impellizzeri, Franco M. et al. 2004. "Use of RPE-Based Training Load in Soccer." *Medicine and Science in Sports and Exercise* 36(6): 1042–47.

L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. 1984. "Classification Algorithms and Regression Trees." *Classification and Regression Trees*: 246–80.

Maron, M. E. 1961. "Automatic Indexing: An Experimental Inquiry." *Journal of the ACM* 8(3): 404–17.

McCulloch, W.S., and W. Pitts. 1943. "A Logical Calculus Nervous Activity." *Bulletin of Mathematical Biology* 52(l): 99–115.

McLachlan, G. 2004. Wiley-Interscience *Discriminant Analysis and Statistical Pattern Recognition*.

Męyk, Edward, and Olgierd Unold. 2011. "Machine Learning Approach to Model Sport Training." *Computers in Human Behavior* 27(5): 1499–1506.

Pettersen, Svein A, Håvard D Johansen, Ivan A M Baptista, and Pål Halvorsen. 2018. "Quantified Soccer Using Positional Data : A Case Study." 9(July): 1–6.

Rampinini, E. et al. 2007. "Variation in Top Level Soccer Match Performance." *International Journal of Sports Medicine* 28(12): 1018–24.

Randers, Morten B. et al. 2010. "Application of Four Different Football Match Analysis Systems: A Comparative Study." *Journal of Sports Sciences* 28(2): 171–82.

Rish, Irina. 2001. "An Empirical Study of the Naive Bayes Classifier." *International Joint Conferences on Artificial Intelligence 2001 Workshop on Empirical Methods in Artificial Intelligence* (January 2001): 41–46.

Rosenblatt, F. 1958. "Frosenblatt." *Psychological Review* 65(6): 1–23.

Rossi, Alessio et al. 2018. "Effective Injury Forecasting in Soccer with GPS Training Data and Machine Learning." *PLoS ONE* 13(7): 1–15.

Di Salvo, V. et al. 2007. “Performance Characteristics According to Playing Position in Elite Soccer.” *International Journal of Sports Medicine* 28(3): 222–27.

Di Salvo, V. et al. 2009. “Analysis of High Intensity Activity in Premier League Soccer.” *International Journal of Sports Medicine* 30(3): 205–12.

Santos, Alejandro Benito, Roberto Theron, Antonio Losada, and Jaime E Sampaio. 2018. “Data-Driven Visual Performance Analysis in Soccer: An Exploratory Prototype.”*Frontiers in Psychology 9: 2416.*

Vigne, G. et al. 2010. “Activity Profile in Elite Italian Soccer Team.” *International Journal of Sports Medicine* 31(5): 304–10.

Villa, Francesco Della et al. 2019. “The Effect of Playing Position on Injury Risk in Male Soccer Players : Systematic Review of the Literature and Risk Considerations for Each Playing Position Take-Home Points.”*American journal of orthopedics (Belle Mead, N.J.)*: 1–11.

Werbos, P. 1974. “‘Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences,’ Unpublished Ph.D. Dissertation, Harvard University, Department of Applied Mathematics.” *Ci.Nii.Ac.Jp* (June).

8. Table des illustrations

9. Annexe : Dictionnaire des variables

Nom abrégé	type	nom complet
date	numeric	
playerName	categorical	
duration	numeric	
numPlayers	numeric	
numPeriods	numeric	
numParam	numeric	Num Parameters
sessiontype	categorical	
positionName	categorical	
maxVel	numeric	Maximum Velocity
velB1TD	numeric	Velocity Band 1 Total Distance
velB2TD	numeric	
velB3TD	numeric	
velB4TD	numeric	
velB5TD	numeric	
velB6TD	numeric	
totPL	numeric	Total Player Load
accB1TEC	numeric	Acceleration Band 1 Total Effort Count
accB2TEC	numeric	
accB3TEC	numeric	
accB4TEC	numeric	
injury	categorical	
slQual	numeric	sleep quality
tired	numeric	
gShape	numeric	good shape
mood	numeric	
pain	numeric	
injWorry	numeric	injury worry
ill	numeric	
sesIntens	numeric	session intensity
sesSatisf	numeric	session satisfaction
sesPl	numeric	session pleasure
slQualLW	numeric	sleep quality Last Week (LW)
tiredLW	numeric	
gShapeLW	numeric	
moodLW	numeric	
painLW	numeric	
injWorryLW	numeric	
illLW	numeric	

sesIntensLW	numeric	
sesSatisfLW	numeric	
sesPlLW	numeric	
slQualL2W	numeric	sleep quality Last 2 Weeks (L2W)
tiredL2W	numeric	
gShapeL2W	numeric	
moodL2W	numeric	
painL2W	numeric	
injWorryL2W	numeric	
illL2W	numeric	
sesIntensL2W	numeric	
sesSatisfL2W	numeric	
sesPlL2W	numeric	
slQualLM	numeric	sleep quality Last Month (LM)
tiredLM	numeric	
gShapeLM	numeric	
moodLM	numeric	
painLM	numeric	
injWorryLM	numeric	
illLM	numeric	
sesIntensLM	numeric	
sesSatisfLM	numeric	
sesPlLM	numeric	
injuryNextWeak	categorical	
injuryNextMonth	categorical	