CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA

Predição de Lesões no Futebol Profissional

Matheus Santos Melo
Matheus Maia Vieira

Prof. Orientador: Jorge de Abreu Soares
Co-Orientador: Lucas Giusti Tavares

Rio de Janeiro,
Dezembro de 2023

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA
CELSO SUCKOW DA FONSECA

Predição de Lesões no Futebol Profissional

Matheus Santos Melo
Matheus Maia Vieira

Projeto final apresentado em cumprimento às normas do Departamento de Educação Superior do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Prof. Orientador: Jorge de Abreu Soares
Co-Orientador: Lucas Giusti Tavares

Rio de Janeiro,
Dezembro de 2023

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

M528 Melo, Matheus Santos
Predição de lesões no futebol profissional / Matheus Santos Melo [e] Matheus Maia Vieira – 2023.
xii, 66f : il. (algumas color.) , enc.

Projeto Final (Graduação). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2023.
Bibliografia : f. 61-66.
Orientador: Jorge de Abreu Soares.
Co-orientador: Lucas Giusti Tavares.

1. Computação. 2. Futebol profissional. 3. Ferimentos e lesões – Análise de dados. I. Vieira, Matheus Maia. II. Soares, Jorge de Abreu (Orient.). III. Tavares, Lucas Giusti (Co-orient.). IV. Título.

CDD 004

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

AGRADECIMENTOS

Primeiramente, nossa gratidão por todo o suporte, ensinamentos e paciência do nosso Orientador, Jorge de Abreu Soares, e nosso Co-orientador, Lucas Giusti Tavares, ao decorrer do trabalho. Seus conhecimentos e orientações foram imprescindíveis para o êxito do projeto.

Agradecemos ao suporte emocional e incentivos durante essa trajetória por parte dos nossos familiares, amigos e todas as pessoas que contribuíram com esse trabalho de alguma forma. A jornada teria sido bem mais complicada sem vossas participações.

Por fim, somos gratos ao Fluminense Football Club e ao seu fisiologista, Juliano Spineti, pelo fornecimento dos dados dos jogadores do clube.

RESUMO

Dentre os diversos esportes existentes que trabalham com novas metodologias por meio da tecnologia, tem-se o futebol como um exemplo promissor no cenário. Devido à sua popularidade, o interesse em aliar tecnologia avançada a esse esporte é iminente. Um dos objetivos é a busca por medidas preventivas que visem diminuir a frequência de episódios lesivos, que proporciona um grande impacto na indústria esportiva, afetando tanto o desempenho da equipe quanto a situação econômica do clube. A fim de mitigar tal problema, um interessante caminho reside no uso da Ciência de Dados em informações referentes às diversas características dos atletas de uma equipe. Visto isso, o presente trabalho propõe, por meio de um conjunto de dados relacionados a atletas profissionais de futebol, prever a probabilidade de episódios lesivos sem contato que possam acometê-los em um microciclo, através da associação de algoritmos de aprendizado de máquina e variáveis-chave dos jogadores. Assim, os profissionais envolvidos com a gestão da equipe poderão adotar ações, tanto no campo médico quanto gerencial, potencialmente minimizando, ao final, os prejuízos futuros com as atividades.

Palavras-chave: Futebol profissional; Predição de lesões; Aprendizado de máquina; Lesões esportivas; Ciência de Dados

ABSTRACT

Among the many existing sports that work with new methodologies through technology, soccer is a promising example on the scene. Due to its popularity, the interest in combining advanced technology to this sport is imminent. One of the goals is the search for preventive measures that aim to reduce the frequency of harmful episodes, which has a great impact on the sports industry, affecting both the team's performance and the club's economic situation. In order to mitigate such a problem, an interesting path lies in the use of Data Science on data regarding the various characteristics of a team's athletes. With this in mind, the present work proposes, by means of a data set related to professional soccer athletes, to predict the likelihood of non-contact injury episodes that may affect them in a microcycle, through the association of machine learning algorithms and key variables of the players. Thus, the professionals involved with the team management will be able to take actions, both in the medical and managerial fields, potentially minimizing, in the end, future losses with the activities.

Keywords: Professional soccer; Injury prediction; Machine learning; Sports injuries; Data Science

Conteúdo

Lista de Figuras

Lista de Tabelas

LISTA DE ABREVIAÇÕES

ANN	Artificial Neural Network	27, 32
DT	Decision Tree	17, 24, 25, 27, 29, 30, 31, 32, 33, 43, 46, 48, 50, 51, 52, 53, 57, 59
ENET	Elastic Net	30, 32
GNB	Gaussian Naive Bayes	27, 32
GPS	Global Positioning System	3, 25, 26, 27, 29, 30, 32, 33
KNN	K-Nearest Neighbour	27, 32
LASSO	Least Absolute Shrinkage And Selection Operator	28, 30, 32
LCA	Ligamento Cruzado Anterior	29
LDA	Linear Discriminant Analysis	27, 32
LR	Logistic Regression	25, 27, 29, 32, 33, 48, 50, 52, 53, 57
MLP	Multi-Layer Perceptron	27, 32
OLS	Ordinary Least Square	30, 32
PCA	Análise De Componentes Principais	xii, 4, 13, 31, 41, 42, 43, 48, 52, 53, 54, 55, 57, 59, 60
RF	Random Forest	16, 17, 25, 27, 29, 32, 33, 44, 48, 50, 51, 52, 53, 55, 57, 59
SF	Stepwise Forward	30, 32
SVM	Support Vector Machine	27, 29, 32
XGB	EXtreme Gradient Boost	26, 27, 29, 30, 32

Capítulo 1

Introdução

O avanço tecnológico e crescente acesso da população aos meios de informação nas últimas décadas desencadeou um exponencial crescimento de dados gerados e armazenados [Services, 2015]. Essa explosão está criando oportunidades para novas formas de combinar e usar informações para encontrar valor, bem como fornecer desafios significativos devido ao tamanho dos dados que estão sendo gerenciados e analisados, beneficiando e otimizando diversos setores da sociedade, como por exemplo, o cenário esportivo [NIST Big Data Public Working Group, 2015].

Um ilustrativo exemplo de cenário de análise de dados nos esportes pôde ser observado no caso em que o gerente geral do time de beisebol Oakland Athletics, Billy Beane, usou-a para explorar, recrutar jogadores e construir uma equipe que ganhou o Campeonato Mundial, apesar de um escasso orçamento [Cullen et al., 2009]. Conforme exemplos como esse, observa-se que despertou-se o interesse mundial de unir ferramentas de análise e manipulação de dados com variáveis intrínsecas a um potencial sucesso no âmbito esportivo, possibilitando encontrar padrões e características cruciais que direcionam a diversos aspectos no jogo, conforme as informações presentes [Patel et al., 2020].

O trabalho realizado nas análises de desempenho no esporte envolve compreender como diversos fatores contribuem para o sucesso, tendo muitos parâmetros envolvidos, e começando a partir de medições em diversos setores relacionados, como: (i) estatura física, biofísica, saúde, aptidão e condicionamento; (ii) atletismo e medidas que lidam com velocidade, poder, força, flexibilidade e agilidade; (iii) medidas psicológicas de inteligência, personalidade e atitude; e (iv) medidas relativas à proficiência em conhecimento esportivo, habilidade e execução na prática e nos jogos [Miller, 2016].

Dentre os diversos esportes existentes que trabalham com novas metodologias por meio da tecnologia em busca de concluir objetivos, tem-se o futebol como um exemplo promissor no cenário. Devido à sua popularidade, o interesse em aliar tecnologia a esse esporte é iminente [Kirkendall and Dvorak, 2010]. Como exemplo, atualmente organizações esportivas investem recursos substanciais na procura de jogadores com potencial para se destacar; sendo assim,

são criados programas de identificação de talentos que visam detectar jogadores talentosos que demonstrem significativo desempenho, apresentando preditivos de sucesso na carreira futura do atleta [Bergkamp et al., 2019]. Uma grande variedade de características físicas e psicológicas desempenham um papel importante para o jogador atingir o nível de elite, promovendo assim uma identificação de talentos baseada em um conjunto versátil de variáveis a serem estudadas e trabalhadas como preditoras de sucesso [Jauhiainen et al., 2019]. Visto isso, paralelamente, a existência de lesões em cenários esportivos atraiu o interesse crescente dos investigadores, gestores e treinadores de investir em estudos e tecnologias que visem ações adequadas para evitá-las [Rossi et al., 2018].

Esses incidentes têm um grande impacto na indústria esportiva, afetando tanto o desempenho da equipe quanto a situação econômica do clube [Rossi et al., 2022]. Pesquisas demonstram que as lesões na Espanha, por exemplo, causam cerca de 16% das ausências na temporada de jogadores profissionais de futebol, correspondendo a um custo de cerca de 188 milhões de euros por temporada [Cuevas et al., 2021]. Jogadores de futebol profissional sofrem entre 2,5 a 9,4 lesões por 1000 horas de esforço [Pfirrmann et al., 2016], das quais cerca de um terço são por uso excessivo e consequentemente, conforme a pesquisa, potencialmente previsíveis e evitáveis. A maioria delas dura por volta de uma semana, entretanto, as mais recorrentes (correspondentes a 15% do total) demandam um tempo maior de repouso [Fiscutean, 2021]. Inclusive, a gravidade de um episódio danoso (também conhecido como severidade) pode ser potencializada conforme fatores como os custos totais e o tempo de trabalho perdido.

De acordo com o modelo UEFA [Hägglund et al., 2005], lesões com ou sem contato são definidas como qualquer dano tecidual sofrido por um jogador que proporcione a ausência nas atividades físicas por pelo menos um dia após o evento [Rossi et al., 2018]. Observou-se em Ekstrand et al. [2011] que as partes do corpo humano nesse esporte mais comumente acometidas por eventos lesivos são a coxa, joelho, tornozelo, quadril e virilha, contemplando principalmente quadros de distensão muscular, entorse ligamentar e contusões [Ekstrand et al., 2011].

No geral, a incidência de uma lesão no meio esportivo resulta em múltiplas repercussões. Hägglund et al. [2013] acompanhou o impacto de lesões no desempenho dos times da UEFA Champions League por 11 anos e apontou que um atleta, estando fora da equipe, pode significativamente influenciar de maneira negativa no desempenho do time. Por exemplo, um evento lesivo de um jogador importante durante uma partida pode acometer o resultado da partida, pois potencialmente afeta o psicológico dos outros participantes da equipe [Hägglund et al., 2013].

Somado a isso, a ausência também pode ampliar a carga de trabalho (treino e jogo) de seus companheiros, aumentando a probabilidade de lesões em todo o time [Fiscutean, 2021].

Em concordância com a incidência de lesões que acontecem no cenário futebolístico [Ekstrand et al., 2011] e as suas consequências negativas às equipes, medidas preventivas nesse quesito se tornaram um objetivo imprescindível em comum aos profissionais da medicina esportiva, a fim de obter, por meio de modelos práticos e úteis, um suporte na tomada de decisão dos treinadores e das equipes médicas em busca de evitar lesões inoportunas [Kirkendall and Dvorak, 2010]. Essa prática se tornou evidente nos últimos anos conforme a investigação de Ekstrand et al. [2021] em um escopo de 18 anos, que demonstrou decaimentos da incidência lesiva durante partidas e treinos e das taxas de reincidência, além do aumento da disponibilidade dos jogadores para jogar partidas e treinos [Ekstrand et al., 2021]. Além da incidência, estudar e definir a severidade (isto é, a análise da influência e magnitude ao redor das variáveis associadas a uma consequência danosa) é uma etapa crucial afim de construir medidas preventivas que possam diminuir lesões e até evitá-las [Kirkendall and Dvorak, 2010]. A severidade das lesões futebolísticas são descritas com base nos seguintes critérios: (i) natureza da lesão; (ii) duração e natureza do tratamento; (iii) tempo desportivo perdido; (iv) tempo de trabalho perdido; (v) danos permanentes; e (vi) custos associados [Inklaar, 1994].

Não só no futebol, como também em outros esportes, existem pesquisas e trabalhos já realizados ou em andamento que utilizam artifícios como base para a coleta e análise dos dados, assim como possíveis providências com os mesmos. Um dos exemplos mais consolidados evidencia-se pela empresa Catapult, a qual desenvolveu uma tecnologia vestível e um software para medir de forma confiável as informações de movimento para equipes inteiras em tempo real com o uso de Global Positioning System (GPS) (Sistema de Posicionamento Global), acelerômetro, giroscópio e magnetômetro [Sikka et al., 2019]. Esse tipo de tecnologia evidenciou resultados benéficos a times esportivos, como por exemplo, com o Toronto Raptors, que implementou dispositivos vestíveis e monitoramento de tecidos moles em busca de melhorar a situação do time, considerado o de maior número de lesões na NBA em 2012. Consequentemente, em 2014 obtiveram, aliado a essa tecnologia e uma melhor gestão dos dados coletados, o histórico de uma das menores taxas de lesões entre os times da NBA [Studnicka, 2020].

A literatura contribui atualmente com relevantes artigos que trabalham com o cenário esportivo e o estudo de possíveis variáveis imprescindíveis que potencializam uma posterior predição [Patel et al., 2020]. Dessa maneira, o objetivo principal deste trabalho visa prever a probabi-

lidade de que ocorram lesões traumáticas sem contato dentro de um microciclo, através da associação de algoritmos de aprendizado de máquina e a carga de treinos e jogos referentes a atletas profissionais masculinos de futebol do Fluminense Football Club. Assim, cria-se a possibilidade da comissão técnica do clube de monitorar os treinos e jogos da equipe, com o intuito de tomar medidas cabíveis para tentar evitar o quadro de lesão. Junto a isso, serão adotados diferentes métodos para melhora do desempenho dos modelos multidimensionais desenvolvidos, como balanceamnento de classes, Analise de Componentes Principais (PCA) e conceitos relacionados à multicolinearidade. De maneira paralela, o objetivo secundário do trabalho visa encontrar variáveis que tenham um potencial vínculo com os quadros lesivos, classificadas como potenciais fatores de risco, por meio dos atributos utilizados nos modelos multidimensionais e também através de análises unidimensionais utilizando gráficos exploratórios e o Teste U de Mann Whitnney.

Logo, este trabalho está organizado da seguinte forma: no Capítulo 2, abordaremos sobre definições que serão utilizadas no contexto do trabalho, referindo-se ao contexto de lesões e dos algoritmos aplicados que irão contemplar o proposto para o trabalho. Para o Capítulo 3, será feito um levantamento afim de buscar e refletir sobre artigos que contribuam cientificamente ao objetivo do trabalho e seu escopo. Já para o Capítulo 4, será abordada a proposta do trabalho, a composição do conjunto de dados que irão compor o modelo por meio de pré-processamento, além da explicação dos algoritmos metodológicas que irão originar os modelos preditivos. O Capítulo 5 consiste na aplicação das informações decorridas na etapa de metodologia conforme a proposta do trabalho, com o intuito de evidenciar as características que construíram os modelos de predição, além de uma análise dos resultados obtidos. Finalmente, o Capítulo 6 consolidará o que foi trabalhado nos tópicos supracitados, apresentará uma síntese dos resultados e contribuições obtidas pelos resultados experimentais, finalizando com a discussão sobre informações referentes a trabalhos futuros.

Capítulo 2

Fundamentação Teórica

A seguir, neste Capítulo, são descritos alguns conceitos importantes para compreendimento da pesquisa. São definições que caracterizam conceitos como o contexto relacionado a lesões e suas definições, bem como discernimento sobre aprendizado de máquina e métodos que serão utilizados para atingir o objetivo do trabalho.

2.1 Definições sobre Lesões

Em primeira instância, é necessário o entendimento dos conceitos referentes à lesão, para uma abordagem compreensiva da maneira que os atletas são acometidos negativamente a um evento lesivo. Visto isso, entende-se por lesão física como qualquer estresse no corpo que impeça o organismo de funcionar apropriadamente e resulte no emprego do corpo a um processo de recuperação [Walker, 2007].

No estudo de Ekstrand et al. [2011], durante sete anos consecutivos, em um tempo total de quinhentos e sessenta e seis mil horas, observou-se quatro mil, quatrocentas e oitenta e três lesões, obtendo-se um saldo de oito a cada mil horas. Dentro desse escopo, indicou-se apenas dois por cento de lesões na cabeça, enquanto que a distensão dos músculos presentes na parte posterior da coxa (isquiotibiais) foi a mais comum. Walker [2007] diz que, embora o termo lesão esportiva possa ser usado para definir qualquer trauma sofrido como resultado do esporte e do exercício, geralmente é levado em consideração, principalmente, os eventos que prejudicam o sistema músculo-esquelético (isto é, que inclui músculos, ossos, articulações, ligamentos e tendões), enquanto lesões mais severas, incluindo cabeça, pescoço e medula espinhal são tratadas separadamente. Em concordância com o autor Walker [2007], este trabalho irá focar nos cenários lesivos futebolísticos que prejudicam o sistema músculo-esquelético.

As definições encontradas para os tipos que constituem uma natureza lesiva podem ser classificadas em dois grupos de conceitos, abordados a seguir.

2.1.1 Lesões Traumáticas

Estas, como o próprio nome define, são originadas em decorrência de algum trauma súbito e de causa conhecida, ocorrendo por meio de contato com outro jogador ou sem interferência externa, possuindo incidência de 81% nas partidas e 59% nos treinos [Ekstrand et al., 2011]. Elas são conhecidas também como lesões agudas e podem resultar em dor, inchaço, sensibilidade, fraqueza e incapacidade de usar ou colocar peso na área lesionada [Walker, 2007].

Em concordância com Hägglund et al. [2005], elas podem ser divididas e definidas em seis tipos:

1. **Entorse**

   Consiste em uma lesão dos ligamentos e articulações por traumas referentes à torção (Figura 1). Ligamentos são conhecidos como tecidos conjuntivos fibrosos densos que conectam osso a osso e providenciam estabilidade para as articulações (responsáveis pela locomoção e, com os ossos, permitem ou limitam o movimento dos membros) [Walker, 2007].

Figura 1: Exemplo de entorse ligamentar no tornozelo. Fonte: ABTPé - Disponível em: `abtpe.org.br`. Acesso em: 23 mai, 2023.

2. **Distensão**

   Frequente nos grupos musculares, esta consiste na lesão dos tendões, que são os tecidos conjuntivos fibrosos que conectam o músculo ao osso e trabalham em conjunto com os músculos para exercer força nos ossos e produzir movimento. Isso causa danos aos músculos e outros tecidos moles (Figura 2) [Walker, 2007].

Figura 2: Exemplo de distensão muscular na região dos isquiotibiais e panturrilha. Fonte: Wikipedia - Disponível em: `bit.ly/4amkQ4V`. Acesso em: 25 mai, 2023.

Entorses e distensões musculares estão entre os tipos lesivos futebolísticos mais comuns [Ekstrand et al., 2011]. Suas classificações podem ser melhor definidas de acordo com a gravidade, que é categorizada em três graus (abordados na Tabela 1), conforme Walker [2007]. As ilustração dos graus da entorse e distensão muscular podem ser vistas, respectivamente nas Figuras 3 e 4.

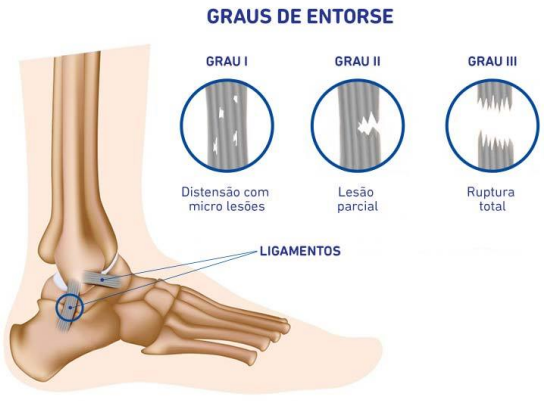

Figura 3: Ilustração dos três graus de um entorse ligamentar no tornozelo. Fonte: Ortopedista do Joelho - Disponível em: `ortopedistadojoelho.com.br/entorse-do-tornozelo/`. Acesso em: 25 mai, 2023.

Categoria	Definição
Primeiro Grau	Sendo a menos grave, é o resultado de um pequeno alongamento dos ligamentos, músculos ou tendões e é acompanhado por dor leve, algum inchaço e rigidez articular. Além disso, geralmente há muito pouca perda de estabilidade articular.
Segundo Grau	Com gravidade média, é o resultado do alongamento juntamente a alguma ruptura dos ligamentos, músculos ou tendões, apresenta um aumento do inchaço e da dor e uma perda moderada de estabilidade ao redor da articulação.
Terceiro Grau	Possuindo gravidade alta, é o resultado de uma ruptura completa ou ruptura de um ou mais ligamentos, músculos ou tendões e resultará em inchaço maciço, dor intensa e instabilidade grosseira.

Tabela 1: Definição de lesões referentes à entorse e distensão, de acordo com seu grau [Walker, 2007].

Figura 4: Ilustração dos três graus de uma distensão muscular. Fonte: Santibras Fisioterapia - Disponível em: `santibrasfisioterapia.com/distensao/`. Acesso em: 25 mai, 2023.

3. **Contusão**

   Além da entorse e distensão muscular, a contusão também é um tipo de lesão comum em esportes de impacto entre jogadores, como o caso do futebol [Ekstrand et al., 2011]. Resultado de um choque físico direto (com ou sem contato com outros jogadores), são entendidas como um impacto no músculo, tendão ou ligamento, causando hematomas e muitas vezes descoloração devido ao acúmulo de sangue ao redor do local do trauma (Figura 5) [Walker, 2007].

4. **Concussão**

   Entende-se como uma agitação violenta ou abalo do cérebro, resultando em comprometimento imediato ou transitório da função neurológica e é considerada a lesão esportiva mais grave (Figura 6) [Walker, 2007].

Figura 5: Hematoma resultante de uma contusão na região da canela e panturrilha. Fonte: Sports injury clinic - Disponível em: sportsinjuryclinic.net/sport-injuries/lower-leg/calf-pain/contusion-lower-leg. Acesso em: 26 mai, 2023.

Figura 6: Mancuello, ex-jogador do Clube de Regatas do Flamengo, disputando a bola momentos antes de sofrer uma concussão e cair no chão desacordado. Fonte: O Globo - Disponível em: oglobo.globo.com/esportes/choques-de-cabeca-preocupam-medicos-que-trabalham-no-futebol-21148196. Acesso em: 26 mai, 2023.

5. **Fratura**

   Associa-se a uma quebra ou rachadura traumática de um osso [Hägglund et al., 2005]. Entre jogadores profissionais de futebol, as fraturas equivalem a entre 10% a 12% de todas as lesões ocorridas nesse esporte [Court-Brown et al., 2008]. Apesar de sua representatividade relativamente baixa, as fraturas constituem uma das lesões mais graves sofridas pelos jogadores de futebol, respondendo pelo maior tempo de recuperação pós-lesão [Court-Brown et al., 2008]. Seus tipos podem ser vistos na Figura 7.

6. **Luxação**

   Entende-se como o deslocamento de um osso em relação ao osso oposto, resultando em perda parcial da articulação das extremidades da estrutura óssea oposta (Figura 8) [Marchiori, 2014].

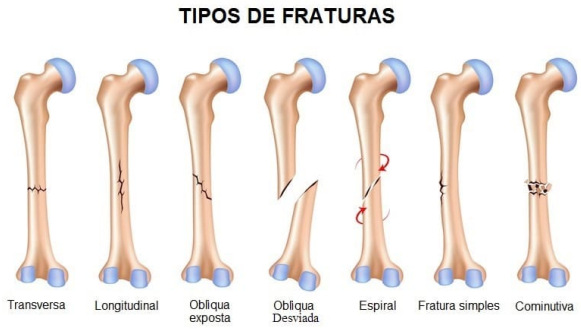

Figura 7: Tipos de fraturas do osso. Fonte: Sanarmed - Disponivel em: sanarmed.com/resumo-sobre-fraturas-sanarflix. Acesso em: 29 mai, 2023.

Figura 8: Exemplos de luxações no ombro. Fonte: Dr. Karina Levy - Disponivel em: drakarinalevy.com.br/luxacao-do-ombro/. Acesso em: 29 mai, 2023.

2.1.2 Lesões por uso excessivo

Também conhecidas como lesões crônicas, ao contrário das traumáticas, possuem origem dentro de um período maior de tempo, são insidiosas e sem trauma conhecido (por exemplo, tendinites, bursites e mini traumas por estresse). As suas consequências danosas são análogas às agudas, resultando em dor, inchaço, sensibilidade, fraqueza e incapacidade de usar ou colocar peso na área lesionada [Walker, 2007]. Lesões por uso excessivo representam 28% do total de lesões no futebol [Ekstrand et al., 2011].

Além das lesões serem classificadas como traumáticas e por uso excessivo, estas também são conceituadas a partir da sua severidade, sendo demonstrada na Tabela 2 [Walker, 2007].

2.1.3 Fatores de risco de lesão

Na natureza futebolística, a origem de lesões estão propicias a acometer os praticantes desse esporte a qualquer momento conforme um ramo complexo de características envolvidas, tanto

Nível	Classificação
Leve	Mínima dor e inchaço e não afetará a performance esportiva.
Moderado	Razoável dor e inchaço e limitará um pouco a performance esportiva.
Grave	Alta dor e inchaço e não só afetará a performance esportiva, como também as atividades rotineiras diárias.

Tabela 2: Classificação da lesão, de acordo com o nível da severidade [Walker, 2007].

nas sessões de treino quanto nas partidas [Vallance et al., 2020]. Segundo Rossi et al. [2018], no estado da arte há uma gama de abordagens que fornecem uma compreensão preliminar de quais variáveis podem influenciar no risco de lesão, enquanto ainda há uma necessidade de pesquisas práticas com o intuito de avaliar o potencial de modelos preditivos utilizando das mesmas. Apesar disso, o estudo dos potenciais fatores de riscos mostra-se imprescindível, visto que o uso de variáveis-chave associadas ao modelo preditivo predispõe uma possível melhor eficácia e performance do mesmo [Hughes et al., 2020].

Os fatores de risco podem ser classificados em dois tipos, observados abaixo.

1. **Fatores Intrínsecos**:

   São entendidos como as características individuais inertes a cada jogador, isto é, idade, sexo, histórico de lesões, tamanho do corpo, anatomia local e biomecânica, aptidão aeróbica, força, desproporcionalidade e rigidez musculares, frouxidão ligamentar, controle motor central, questões psicológicas e psicossociais, capacidade mental geral, dentre outros [Taimela et al., 1990].

2. **Fatores Extrínsecos**:

   Compreende-se como características do ambiente em que o atleta está exposto, isto é, carga de treinos e jogos disputados, fatores climáticos, superfície e condições do campo de jogo (seco, molhado, irregular), equipamentos (caneleiras, calçados), regras e condutas da partida dentre outros [Dvorak et al., 2000].

Algumas das causas para riscos de lesão, tanto intrínsecas quanto extrínsecas, podem possivelmente ser modificadas. Para Meeuwisse et al. [2007], um fator de risco pode ser minimizado conforme o atleta participa e se adapta ao ambiente ou a situações potencialmente lesivas, sem consequentemente sofrê-las. Como exemplo, em esportes de colisão, por conta do aprimoramento dos atributos intrínsecos relacionados à força (através de modificações e adaptações nas

cargas de treino e jogo do jogador), pode-se ter o risco de lesão diminuído no momento em que houver exposição a fatores e eventos externos contribuintes para um episódio danoso [Meeuwisse et al., 2007].

Não obstante, existem fatores de risco que não são modificáveis e podem salientar relevância a um potencial risco de lesão. Como exemplo, no modelo feito por Arnason et al. [2004], identificou-se que idade e histórico lesivo foram os protagonistas na ocorrência de lesões dos jogadores de um time de futebol profissional na Islândia.

2.2 Macrociclos, Mesociclos e Microciclos

Os dados relacionados à carga de treino e jogo dos atletas e as súbitas lesões podem ser coletados conforme processos organizacionais diferentes, nos quais Matveev destacou em três estruturas fundamentais. A primeira é a de **macroestrutura**, que são grandes ciclos de treinamento, chamados de *macrociclos*. A macroestrutura tem um planejamento anual, semestral e quadrimestral. A segunda, **mesoestrutura**, é de média duração de planejamento de treinamento e envolve os chamados *mesociclos*. Já a terceira, chamada **microestrutura**, representam sessões de treinamento de forma isolada - *microciclos* [Martins and de Oliveira, 2021].

O macrociclo normalmente se estende de um ano a um ano e meio de treinamento, e, para atletas, o início e o final desse período ocorrem normalmente após a última competição [Viru, 1991]. Esse tempo corresponde à planificação geral das atividades desenvolvidas pelo esportista.

O mesociclo, na sua origem, foi utilizado para descrever as principais fases de treinamento durante o ano (preparação, primeira transição, competição e segunda transição). Portanto, o mesociclo refere-se a um período de dois a três meses. Entretanto, para que ocorram grandes melhoras, as alterações devem ser feitas a cada quatro a seis semanas, podendo então o termo referido indicar esse período [Kraemer and Häkkinen, 2004].

Por fim, temos os microciclos, que são responsáveis por assegurar uma coordenação entre um regime de trabalho e a sua recuperação. Geralmente, referem-se a uma semana de treinamento; há, entretanto, desportistas que treinam três vezes por dia e necessitam de microciclos menores [Viru, 1991].

Em nosso trabalho, agrupamos em microciclos de um jogo ao outro, que engloba todas as atividades realizadas nesse período de tempo.

2.3 Métodos de Análise

Com o intuito de realizar uma análise exploratória dos atributos presentes no conjunto de dados utilizado no projeto, foram adotados os métodos de: (i) Correlação e multicolinearidade; (ii) Análise de Componentes Principais (PCA) e (iii) Teste U de Man Whitnney.

2.3.1 Correlação e multicolinearidade

A correlação pode ser descrita como o grau de associação entre duas variáveis [Asuero et al., 2006]. Paralelamente, a multicolinearidade refere-se à alta correlação com uma ou mais variáveis, sendo um problema que cause dificuldades a respeito da confiabilidade das estimativas dos parâmetros de um modelo [Alin, 2010]. Sendo assim, a multicolinearidade é um problema pois prejudica a significância estatística de uma variável independente [Allen, 1997].

2.3.2 Análise de Componentes Principais

PCA é um método estatístico versátil para reduzir uma tabela de dados de casos por variáveis às suas características essenciais, chamadas componentes principais. Elas consistem em combinações lineares das variáveis originais, explicando ao máximo a variância de todas as variáveis. No processo, o método fornece uma aproximação da tabela de dados original usando apenas poucas componentes principais [Greenacre et al., 2022].

2.3.3 Teste U de Man Whitnney

O teste de Mann-Whitney é um teste não paramétrico que busca determinar se dois grupos independentes pertencem à mesma população [MacFarland et al., 2016]. É usado quando há duas amostras provenientes de duas populações [Mann and Whitney, 1947]. Esse teste possui muitos usos apropriados e deve ser considerado ao usar: dados classificados, dados que se desviam dos padrões de distribuição aceitáveis, ou quando há diferenças perceptíveis no número de sujeitos nos dois grupos comparativos [MacFarland et al., 2016].

2.4 Métodos de Balanceamento de Classes

No processo de desenvolvimento de um modelo de aprendizado de máquina, normalmente ocorre um problema de desbalanceamento das classes, como ocorreu no conjunto de dados do trabalho desenvolvido. Para isso, existem estratégias de pré-processamento que equilibram e balanceiam os valores, com o intuito de tentar melhorar o desempenho do modelo. As estratégias mais comuns são de *Undersampling* e *Oversampling*, sendo a primeira utilizada como método principal para tratar o desequilíbrio de classes no projeto. As estratégias mencionadas podem ser descritas como:

1. *Undersampling*: Processo de diminuição do número de registros na classe majoritária. Isso significa que, durante a classificação, o tempo de treinamento também é consideravelmente diminuído [Liu, 2004].

2. *Oversampling*: Aumento do número de instâncias de classes minoritárias para equilibrar a distribuição das classes. O mais simples é o *Oversampling* aleatório, no qual apenas duplica instâncias minoritárias. O ponto negativo é que não é adicionada nenhuma informação nova ao conjunto de dados e pode causar ajuste excessivo dos classificadores [Zheng et al., 2015].

2.5 Aprendizado de Máquina

Aprendizado de Máquina (Machine Learning, em inglês) é uma subárea de Inteligência Artificial na qual um sistema pode aprender por meio do uso de dados em um modelo de previsão com viés matemático e estatístico, permitindo descoberta de padrões e providenciando resultados satisfatórios e precisos que induzem a conclusões analíticas, tendo aplicação, por exemplo, em diversas áreas da ciência, saúde, assim como, no futebol [Majumdar et al., 2022]. Neste último contexto, no qual o cerne deste trabalho se estabelece, técnicas modernas de aprendizado de máquina têm sido utilizadas com diversos objetivos preditivos, como por exemplo, monitoramento de cargas de treino e jogo, trajetórias na carreira dos jogadores, performance do clube, assim como predição de lesões [Nassis et al., 2023].

Os algoritmos de aprendizado de máquina podem ser agrupados de acordo com o tipo de aprendizado, podendo ser descritos como supervisionados e não-supervisionados:

- **Supervisionado** - Permite a obtenção de informações, baseadas em um conjunto de dados rotulados e com padrões definidos, possuindo informações de entrada e saída (isto é, para cada unidade da amostra há uma associação entre diversas entradas para uma saída). No cenário de previsão de lesões, por exemplo, os fatores de risco de lesão são considerados dados de entrada, enquanto que a ocorrência da própria é classificada como um dado de saída. Como exemplo, temos algoritmos referentes a regressão e classificação [Majumdar et al., 2022];

- **Não-supervisionado** - Nesta, também é permitida a obtenção de informações por base de dados, porém não apresentam rótulos específicos aos atributos. Como exemplo, pode-se citar técnicas de agrupamento e redução de dimensionalidade [Majumdar et al., 2022].

Dessa forma, conforme o objetivo do trabalho vigente, é necessário o discernimento dos algoritmos que serão utilizados. Nesse quesito, algoritmos de classificação com aprendizado de máquina para predição de lesões, por exemplo, possuem o intuito de prever corretamente quadros lesivos e não-lesivos, respectivamente consideradas como classes positivas e negativas, enquanto que a regressão serve para prever valores contínuos, como os resultados das métricas avaliativas [Majumdar et al., 2022].

2.5.1 Árvore de Decisão

A Árvore de decisão (em inglês, Decision Tree) estabelece regras para tomada de decisão, de forma simples e interpretável, conduzindo predições com relevante acurácia conforme a qualidade e precisão dos dados fornecidos [Kingsford and Salzberg, 2008]. Por meio de um conjunto de instruções de controle condicional, organizadas de forma hierárquica, os nós intermediários representam decisões e os nós das folhas podem ser rótulos de classe definidas (para problemas de classificação) [Belle and Papantonis, 2021].

Segundo Kingsford and Salzberg [2008], as decisões no modelo podem ser feitas a partir de perguntas simples binárias (isto é, sim ou não), também sendo respondidas em contextos mais complexos, levando em consideração combinações, lineares ou lógicas, que envolva muitos atributos de uma só vez. A árvore de decisão se torna cada vez maior conforme são incrementados nós intermediários com decisões a serem tomadas a partir da divisão dos dados de treino em conjuntos cada vez menores [Kingsford and Salzberg, 2008]. Porém, o principal problema é a possibilidade de "overfitting", no qual o modelo possui excelentes resultados por conta do

excesso de treino, porém não apresenta o mesmo com a mudança da entrada para dados de teste [Belle and Papantonis, 2021].

Para árvores de decisão, as duas métricas mais comuns adotadas para tomada de decisão binária (sim ou não) são a entropia e o índice Gini [Kingsford and Salzberg, 2008]. Concomitantemente, o índice Gini foi a medida aplicada na metodologia de árvores de decisão do presente trabalho.

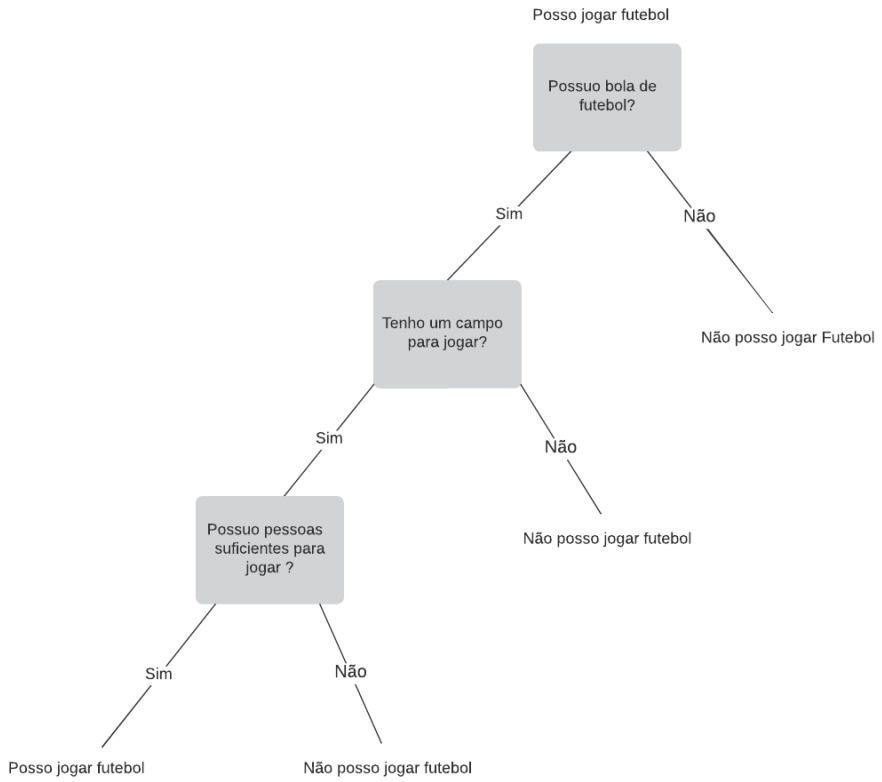

Figura 9: Exemplo de algortimo simples utilizando árvore de decisão. Fonte: Adaptado de Didática Tech - Disponível em: `didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/`. Acesso em: 16 jun, 2023.

2.5.2 Floresta Aleatória

O algoritmo de Floresta Aleatória (em inglês, Random Forest) é composto pela combinação de diversos modelos em árvore de decisão que, de forma randomizada, pode gerar uma maior acurácia dos resultados [Kingsford and Salzberg, 2008]. Random Forest (RF) é uma estratégia eficaz para evitar problemas de *overfitting*, por meio de uma boa generalização dos dados de treino e teste fornecidos [Breiman, 2001]. Com o intuito de realizar uma previsão

agregada, cada árvore individual é treinada em uma parte diferente do conjunto de dados de treino, que é selecionada aleatoriamente e composta por características dentro de uma pequena amostra [Belle and Papantonis, 2021]. Finalmente, cada execução pode resultar em uma árvore ligeiramente diferente. Assim, as previsões do conjunto resultante de árvores de decisão são combinadas com a previsão mais comum, gerando o resultado final [Kingsford and Salzberg, 2008]. Se os valores previstos forem rótulos de classe (categóricos), a árvore de decisão é chamada de árvore de classificação, enquanto se os valores previstos forem numéricos, a árvore é chamada de árvore de regressão [Johansson et al., 2014].

Um benefício importante na utilização de RF, em comparação ao Decision Tree (DT), é o fato de utilizar o resultado de diversas árvores de decisão para concluir o resultado, ao invés de apenas uma. Dessa forma, conforme a combinação de várias árvores, há uma redução da variância do modelo resultante, conduzindo-se a uma melhor generalização e maior acurácia [Belle and Papantonis, 2021]. Porém, o problema relacionado à utilização de RF é a falta de interpretabilidade, devido ao fato de que explicar o conjunto de árvores é bem mais complexo do que apenas uma [Belle and Papantonis, 2021].

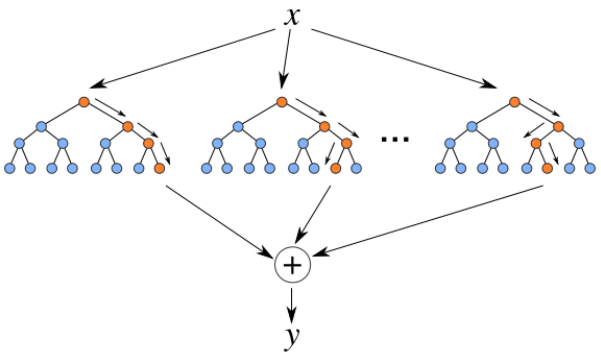

Figura 10: Ilustração de como funciona o algoritmo Random Forest, que combina um conjunto de modelos para atingir o resultado. Fonte: Carlos Baia - Disponível em: carlos baia.com/2016/12/24/decision-tree-e-random-forest/. Acesso em: 16 jun, 2023.

2.5.3 Regressão Logística

A regressão logística é uma abordagem de modelagem matemática que pode ser usada para descrever a relação de várias variáveis independentes (por exemplo, a altura e o peso) com uma variável dependente dicotômica (presença ou não de uma lesão) [Kleinbaum and Klein, 2010]. Normalmente, o resultado é binário, conferindo o nome de modelo logístico binário. Por outro lado, quando há apenas uma variável preditora, chama-se regressão logística simples. Com

vários preditores, incluindo variáveis categóricas e contínuas, o modelo é referido como uma regressão logística múltipla ou multivariável [Nick and Campbell, 2007].

A relação entre as variáveis independentes e a variável dependente se assemelha a uma curva em forma de S, conforme visto na Figura 12.

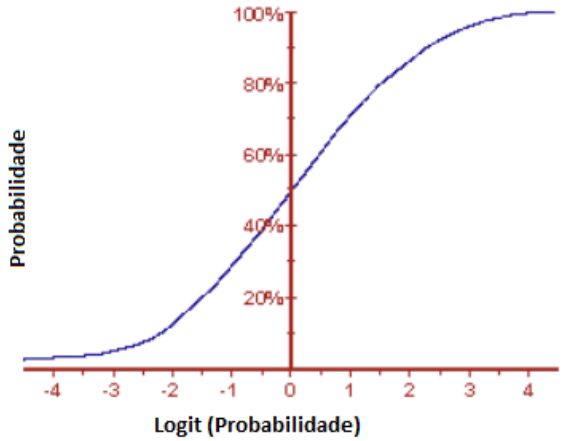

Figura 11: Ilustração da Curva Logística em formato de S. Fonte: Adaptado de Redalyc - Disponivel em: `redalyc.org/articulo.oa?id=76228118008/`. Acesso em: 16 jun, 2023.

2.6 Métodos de Validação e Avaliação

Para analisar a validação realizada no modelo e o desempenho obtido, é importante ter discernimento dos métodos que foram utilizados no projeto. Para validação, foi utilizada a estratégia de Validação Cruzada, enquanto que as métricas avaliativas foram divididas em seis conceitos, sendo eles: (i) Acurácia; (ii) Precisão; (iii) *Recall*; (iv) *F1-Score*; (v) *AUC* e (vi) Erro Percentual Absoluto Médio. Em seguida, descreve-se a explicação para cada conceito.

2.6.1 Validação Cruzada

A validação cruzada é um dos métodos de re-amostragem de dados mais amplamente utilizados para avaliar a capacidade de generalização de um modelo preditivo e para evitar *overfitting* [Berrar et al., 2019]. Este é classificado como o uso de modelos ou procedimentos que violam a par-simonia, ou seja, que incluem mais termos do que o necessário ou usam abordagens mais complicadas do que o necessário [Hawkins, 2004].

O objetivo da validação cruzada na fase de construção do modelo é de fornecer uma estimativa para o desempenho deste modelo final em novos dados[Berrar et al., 2019], também é

frequentemente usada para escolher hiper-parâmetros de um determinado algoritmo de aprendizagem [Arlot and Lerasle, 2016]

2.6.2 Métricas avaliativas

O desempenho de modelos que trabalham com classificação é comumente avaliado por meio de métricas compostas de formas diferentes [Majumdar et al., 2022], sendo contempladas no contexto do nosso trabalho. Ressalta-se que TP, TN, FP e FN referem-se às contagens de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos, respectivamente.

- **Acurácia**: Proporção de lesões e não lesões corretamente previstas para o número total de lesões e não lesões observadas.

$$\frac{TP+TN}{TP+TN+FP+FN}$$

- **Precisão**: Proporção de lesões previstas corretamente em relação ao número total de lesões previstas correta e incorretamente.

$$\frac{TP}{TP+FP}$$

- **Sensibilidade (*Recall*)**: Proporção de lesões previstas corretamente em relação ao total de lesões observadas.

$$\frac{TP}{TP+FN}$$

- ***F1-Score***: Média harmônica entre precisão e sensibilidade.

$$\frac{2*Precisao*Recall}{Precisao+Recall} = \frac{2*TP}{2*TP+FP+FN}$$

- ***AUC***: Área da curva de probabilidade da taxa de verdadeiros positivos e taxa de falsos positivos.

$$Specificity = \frac{TN}{FP+TN}$$

$$AUC = Recall \text{ - } (1 - Specificity)$$

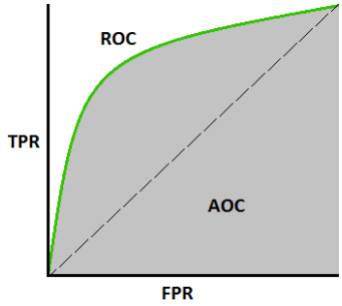

Figura 12: Exemplo de ilustração gráfica com aplicação do conceito de AUC-ROC. Fonte: https://48hours.ai/files/AUC.pdf. Acesso em: 20 novembro, 2023.

Por fim, como métrica avaliativa de modelos de regressão, usa-se o **Erro Percentual Absoluto Médio**, que é a média de todos os erros percentuais absolutos [Kim and Kim, 2016].

$$\sum_{i=1}^{D} |x_i - y_i|$$

Capítulo 3

Trabalhos Relacionados

A revisão literária sobre trabalhos e pesquisas referentes ao tema de predição de lesões no futebol profissional foi conduzida de acordo com as diretrizes do PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis) [Page et al., 2021].

3.1 Seleção dos Trabalhos

Inicialmente, foi construída uma chave de busca calibrada (em inglês) constituída de palavras-chave relacionadas ao tema do trabalho vigente e seus respectivos sinônimos, com o intuito de ampliar o escopo da presente pesquisa. Sendo assim, foram utilizados como descritores principais os termos "Soccer"/"Football"(referentes a futebol) e "Injury prediction"/"Injury risk prediction"(referentes à predição de lesões).

De maneira conjunta, mostrou-se necessária a aplicação de um filtro de exclusão para algumas expressões que pudessem encontrar artigos que não agregassem à pesquisa, pois, podem apresentar termos sinônimos, na língua inglesa, que se referem a outra modalidade esportiva, como o Futebol Australiano e o Futebol Americano. De maneira conclusiva, formou-se a chave de busca calibrada (em inglês): ("soccer"OR "football") AND ("injury prediction"OR "injury risk prediction") AND NOT ("Australian Football"OR "American Football"OR "NFL").

No escopo, foram considerados artigos na língua portuguesa e inglesa, e não houve restrição temporal.

Para a consulta dos trabalhos, foram selecionadas duas bases de dados: Pubmed [LM, 2012] e Scopus Elsevier [Elsevier, 2020].

- Devido ao fato de o tema deste trabalho estar relacionado à área de saúde, foi selecionada a base Pubmed, mediante sua especificidade de pesquisas nesta área.

- A base Scopus Elsevier foi utilizada devido à grande quantidade de material científico disponível em seu acervo, que também contempla a área de pesquisa destinada ao trabalho presente.

3.1.1 Critérios de Inclusão

Os critérios de inclusão possuem a função de direcionar o assunto escolhido e excluir pesquisas que não satisfazem o intuito do trabalho vigente.

Divididos em quatro critérios inclusivos diferentes, eles são:

- **Critério 1** - Trabalhos que abordem lesões classificadas como "sem contato", do tipo traumáticas e/ou uso excessivo e com foco principal nas regiões musculoesqueléticas;
- **Critério 2** - Trabalhos feitos utilizando dados de jogadores(as) no futebol profissional;
- **Critério 3** - Pesquisas que, mesmo com dados mistos de outros esportes, tenham enfoque principal no futebol;
- **Critério 4** - Enfoque na predição de lesões.
- **Critério 5** - Informações relevantes que foram utilizadas e reportadas no artigo abordadas com clareza.

3.1.2 Critérios de Exclusão

Os critérios de exclusão possuem a função de excluir pesquisas que não satisfazem ao intuito do trabalho vigente.

Divididos em quatro critérios exclusivos diferentes, eles são:

- **Critério 1** - Trabalhos que abordem lesões provindas do contato entre jogadores (Por exemplo, concussão).
- **Critério 2** - Trabalhos que não utilizam dados de jogadores(as) no futebol profissional (Futebol Universitário, Futebol Amador, entre outros);
- **Critério 3** - Pesquisas que possuam dados mistos que não se baseiam principalmente no futebol;
- **Critério 4** - Pesquisas encontradas na busca que não tenham relação com o contexto de predição de lesões.
- **Critério 5** - Trabalhos que não ofereçam informações relevantes que foram utilizadas na pesquisa realizada.

3.1.3 Critérios de Priorização

De maneira posterior aos artigos que foram incluídos na busca, são adotados critérios de prioridade que definem a relevância do material resultante.

Foram divididos em três níveis de prioridade, sendo eles:

- **Prioridade Alta** - Predição de Lesões através de estratégias de Aprendizado de Máquinas, utilizando no modelo dados dos(as) jogadores(as) referentes a treino e/ou jogo, coletados na temporada e/ou pré-temporada.
- **Prioridade Média** - Predição de Lesões através de estratégias que não envolvem o uso de Aprendizado de Máquina, utilizando no modelo dados dos(as) jogadores(as) referentes a treino e/ou jogo coletados na temporada ou pré-temporada.
- **Prioridade Baixa** - Não faz Predição de Lesões e menciona alguma(s) potencial(is) variável(is) analisada e estudada que possa potencialmente ter eficácia para Predição de Lesões.

3.1.4 Resultados

O resultado da seleção dos trabalhos buscados utilizando a metodologia PRISMA pode ser analisado em conformidade com a Figura 13. A quantidade de artigos identificados por meio da chave calibrada de busca utilizando palavras-chave selecionadas foi encontrada na execução desta na data de 05/03/2023.

3.2 Comparação entre os Trabalhos Relacionados

Diferentes estudos apresentam uma abordagem a respeito da prevenção de lesões futebolísticas e Aprendizado de Máquina, focando principalmente nos fatores de risco, também nomeadas características-chave [Majumdar et al., 2022]. Em concordância com Rossi et al. [2018], apesar de análises e estudos preliminares referentes à quais variáveis, intrínsecas e extrínsecas, que possam induzir a lesões ser uma etapa importante para a predição, a literatura limita-se a essas pesquisas enquanto que ainda é necessário trabalhos relacionados ao aprofundamento do potencial de algoritmos práticos de aprendizado de máquina que aliem o discernimento de potenciais

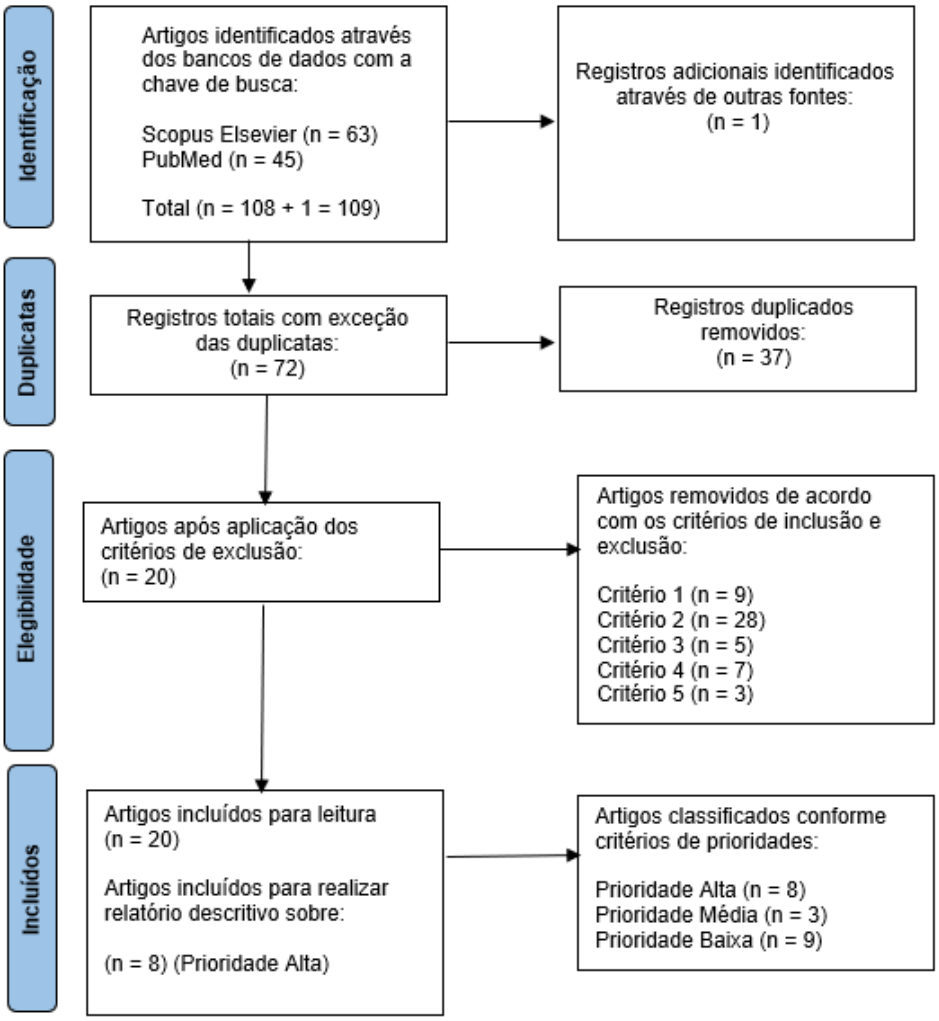

Figura 13: Fluxograma feito com PRISMA para seleção dos trabalhos [Page et al., 2021].

caraterísticas-chave com uma efetiva predição de lesões [Rossi et al., 2018]. O aprendizado de máquina demonstra relevância nesse contexto devido a sua capacidade de trabalhar de forma eficaz e flexível com uma gama vasta de atributos no conjunto de dados [Majumdar et al., 2022]. Em contrapartida, a predição de lesões ainda demonstra dificuldades devido às variedades distintas dos jogadores (por exemplo, diferenças biológicas individuais do corpo, predisposições físicas e condições psicofísicas) [Piłka et al., 2023].

Concomitantemente, recomenda-se a combinação de diversos potenciais fatores de risco em prol de mais acurácia à previsão do evento lesivo [Hughes et al., 2020]. Deste modo, evitando um modelo preditivo mono-dimensional (isto é, que trabalha apenas com uma variável referente ao jogador) e explorando inteiramente os dados e seus complexos padrões inerentes [Rossi et al., 2018], de natureza tanto intrínseca, quanto extrínseca [Majumdar et al., 2022].

A respeito dos algoritmos de aprendizado de máquina, DT foi o mais utilizado, presente nos artigos de Rossi et al. [2018], Vallance et al. [2020], Eetvelde et al. [2021] e Rossi et al. [2022],

que trabalham principalmente com variáveis em comum relacionadas a carga de treino e/ou jogo dos atletas extraídas com tecnologias GPS e algumas características subjetivas antropométricas. De acordo com a revisão sistemática de Eetvelde et al. [2021], o método DT é considerado um algoritmo de aprendizado de máquina popularmente utilizado no cenário de medicina esportiva pois evidencia uma fácil visualização e interpretabilidade do modelo produzido com este método preditivo.

A seguir, um resumo dos trabalhos e suas respectivas características descritivas, estudadas em questão.

1. **Effective injury forecasting in soccer with GPS training data and machine learning - [Rossi et al., 2018]**

   Nesta pesquisa de Rossi et al. [2018], o objetivo principal foi utilizar dados de treino dos jogadores (coletados com dispositivos GPS) para construir um modelo multi-dimensional de Aprendizado de Máquina com o intuito de prever que um jogador irá ter uma lesão sem contato na próxima sessão de treino ou jogo. Foram extraídas como base doze variáveis-chave das cargas de treino e seis características individuais (como idade, posição setorial e tempo total jogado) pertencentes a vinte e seis jogadores profissionais italianos na faixa de vinte e dois a trinta anos de idade durante vinte e três semanas ao longo da temporada 2013/2014. Os métodos principais aplicado de Aprendizado de Máquina incluíam: (i) DT; (ii) RF, (iii) Logistic Regression (LR).

   Com base nos resultados da predição, o melhor foi referente ao método DT, que pôde detectar por volta de 80% das lesões com precisão de 50%. Também aponta-se que, analisando a influência do histórico de lesões associado às outras variáveis, os clubes precisam redobrar a atenção às sessões de treino dos jogadores que acabaram de retornar de lesão, pois estão mais vulneráveis nesse momento.

   Conclusivamente, o estudo apresenta um modelo preditivo útil para precaver os clubes futebolistas a respeito de gastos referentes a lesões. Rossi et al. [2018] aponta que, uma extensão da pesquisa pode ser obtida com a extração de dados de jogo dos jogadores, onde apresentam maiores índices de estresse físico e psicológico. Além disso, um estudo mais aprofundado pode ser realizado com dados intrínsecos e extrínsecos dos jogadores coletados durante mais temporadas.

2. **Predicting Injuries in Football Based on Data Collected from GPS-Based Wearable Sensors - [Piłka et al., 2023]**

   Nesta pesquisa de Piłka et al. [2023], o foco principal foi construir modelos de tomada de decisão que possam prever lesões sem contato dos membros inferiores do corpo resultadas de poucos treinos ou treinos em excesso. As variáveis coletadas foram baseadas nas cargas de treino e características intrínsecas de trinta e seis jogadores profissionais do time polonês PKO PB Ekstraklasa, durante duas temporadas consecutivas (2020/2021 e 2021/2022). Os métodos de tomada de decisão aplicados foram: (i) Expert Knowledge-Based Rules; (ii) Fuzzy Rule-Based Model, (iii) eXtreme Gradient Boost (XGB) (Aprendizado de Máquina).

   Com base nos resultados da predição com os métodos propostos, pôde-se notar a relevância de quatro atributos individuais: o tempo total que o atleta treinou, características dos jogadores duas semanas antes das análises serem feitas, o número de acelerações e desacelerações, e o número de metros corridos na velocidade entre 19.8-25.2km/h. Estas duas últimas são essenciais a serem analisadas em partidas de futebol, por serem atividades de alto impacto, com frequentes mudanças na dinâmica e direção de movimento. Além disso, o método de aprendizado de máquina XGB foi o mais eficaz, contribuindo em bons resultados referentes a uma acurácia de 90% e precisão de 92%.

   Conclusivamente, Piłka et al. [2023] aponta que apesar de resultados satisfatórios, algumas melhorias podem ser realizadas em estudos futuros, como por exemplo, um aumento da quantidade dados coletados, juntamente ao enriquecimento de fatores de risco intrínsecos dos jogadores nos treinos (batimentos cardíacos, taxa de esforço percebido, etc).

3. **Combining internal- and external-training-loads to predict non-contact injuries in soccer - [Vallance et al., 2020]**

   Nesta pesquisa de [Vallance et al., 2020], o intuito principal foi construir um modelo preditivo que sustente a hipótese de que o uso de variáveis intrínsecas e extrínsecas dos jogadores obtidas através de tecnologias de GPS e questionários subjetivos melhora a performance de modelos que preveja lesões sem contato. Foram consideradas duas vertentes temporais: uma semana e um mês. As cargas de treino e jogo, questionários perceptivos de bem-estar e lesões foram monitorados durante uma temporada completa (2017/2018), com quarenta participantes profissionais de idade entre vinte e três a trinta e cinco anos,

de times da segunda divisão francesa. No total, foram usados vinte e sete variáveis-chave. Foram utilizados diversos métodos de aprendizado de máquina, sendo eles: (i) K-Nearest Neighbour (KNN); (ii) Linear Discriminant Analysis (LDA); (iii) LR; (iv) Ridge classifier; (v) Gaussian Naive Bayes (GNB); (vi) DT; (vii) RF; (viii) Support Vector Machine (SVM); (ix) Multi-Layer Perceptron (MLP), (x) XGB.

Conforme os resultados obtidos, observou-se que os métodos com melhores performance foram usados classificadores KNN, DT, RF e XGB. Na maioria dos modelos, a combinação de atributos pessoais, histórico de lesões e dados obtidos de GPS e questionários ofereceu uma performance muito melhor, com métricas preditivas beirando 100%, dependendo da complexidade do modelo, especialmente na vertente temporal de um mês.

Conclusivamente, Vallance et al. [2020] enfatiza que variáveis subjetivas obtidas por meio de questionários relacionados ao bem-estar, como por exemplo, qualidade de sono, fadiga, estresse, entre outras, são fatores de risco determinantes na ocorrência de lesões dos jogadores profissionais. Acrescenta-se que, com um maior conjunto de dados e melhorias na coleta e qualidade das informações subjetivas e obtidas por tecnologia GPS, pode possivelmente agregar melhores performances aos modelos classificadores utilizados.

4. **Machine learning methods in sport injury prediction and prevention: a systematic review- [Eetvelde et al., 2021]**

   O artigo publicado [Eetvelde et al., 2021] é uma revisão sistemática realizada com o modelo PRISMA [Page et al., 2021], que analisa diversos métodos de aprendizado de máquina para realizar a previsão de lesões esportivas. Como resultado da busca sistemática, sobraram onze de duzentos e quarenta e nove estudos, após a aplicação dos critério de inclusão e exclusão. Os fatores de risco modificáveis presentes no modelo incluíam, por exemplo, carga de treino, características psicológicas e neuromusculares e nível de estresse, enquanto que os não-modificáveis envolviam, variáveis demográficas, marcadores genéticos, medidas antropométricas e história lesivo. Os métodos de aprendizado de máquina classificadores destacados eram referentes a modelos em árvore (principalmente DT), SVM e Artificial Neural Network (ANN).

   Com base nos resultados, artigos sobre futebol profissional foram os mais promissores para prever o risco de lesão, apontando que dados coletados na temporada e pré-temporada foram os mais úteis, contrapondo artigos que abordaram a ineficácia dessas

informações nesse contexto de lesões. Devido à grande variedade de recursos usados nos diferentes artigos, não foi encontrada muita consistência nos preditores importantes relatados, porém as características que foram relatadas de maneira repetida foram histórico lesivo, alta carga de treino e maior composição corporal (apenas em jogadores jovens).

Conclusivamente, Eetvelde et al. [2021] aponta que métodos de aprendizado de máquina são mecanismos úteis para identificar atletas em alto risco lesivo ou fatores de risco relevantes. No entanto, embora a maioria dos estudos analisados tenha aplicado adequadamente os métodos de aprendizado de máquina para prever lesões, a qualidade metodológica do estudo foi moderada a muito baixa.

5. **Predictive modeling of lower extremity injury risk in male elite youth soccer players using least absolute shrinkage and selection operator regression - [Kolodziej et al., 2023]**

   Nesta pesquisa de [Kolodziej et al., 2023], o foco principal foi dividido em duas vertentes: (i) identificar fatores de risco de lesão relacionados a características neuromusculares e biomecânicas jogadores de futebol juvenil profissional e (ii) avaliar a capacidade preditiva de uma abordagem de aprendizado de máquina usando um modelo Least Absolute Shrinkage and Selection Operator (LASSO) . Os parâmetros de performance biomecânica e neuromusculares foram coletadas de sessenta e dois jogadores de clubes alemães com idade entre dezesseis a dezoito anos na temporada 2018/2019.

   Com base nos resultados, pôde-se observar a relevância de alguns aspectos referentes a frequência lesiva por região do corpo e tipo de lesão. As partes mais comuns afetadas foram: tornozelo (36%), isquiotibiais (18%) e quadríceps (18%). Os tipos de lesão mais recorrentes foram entorses (48%) e distensões (39%). Já a performance do modelo preditivo com LASSO apresentou uma probabilidade de predição de 58% e dos fatores de risco biomecânicos e neuromusculares utilizados através de testes laboratoriais com os jogadores, três variáveis foram mais evidentes para a predição.

   De maneira conclusiva, Kolodziej et al. [2023] aponta que com a performance preditiva positiva do modelo, medidas biomecânicas e neuromusculares são variáveis-chave relevantes a serem utilizadas e trabalhadas no cenário preditivo de lesões. Um adendo limitante ressaltado é a abordagem de que os testes realizados foram obtidos na pré-temporada, podendo haver uma mudança relevante das características dos jogadores até

o momento da lesão. Assim, Kolodziej et al. [2023] indica o melhor monitoramento das variáveis através de tecnologias GPS, que coletam com precisão os dados dos jogadores em cada sessão de treino e/ ou jogo e possivelmente detectam com antecedência um episódio lesivo.

6. **Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes - [Jauhiainen et al., 2022]**

   Nesta pesquisa de [Jauhiainen et al., 2022], o intuito principal foi investigar a habilidade preditiva do modelo construído utilizando um conjunto de dados de jogadores profissionais femininas, afim de prever uma lesão ligamentar popular e específica do joelho conhecido como Ligamento Cruzado Anterior (LCA). O estudo trabalha com variáveis demográficas, neuromusculares, biomecânicas, anatômicas e genéticas que estejam relacionadas à fatores de risco de LCA. Foram utilizados quatro métodos de Aprendizado de Máquina: (i) LR ; (ii) RF ; (iii) SVM (Linear), (iv) SVM (Não-linear).

   Analisando conforme os resultados obtidos, demonstrou-se que, mesmo com um conjunto extenso de dados incluindo informações antropométricas, clínicas, neuromusculares, genéticas e medidas biomecânicas sofisticas em 3d, a predição de lesões de LCA teve um resultado ruim.

   De forma conclusiva, os resultados ruins obtidos a partir da avaliação preditiva com a construção de modelos de aprendizado de máquina indicam a necessidade de estudos futuros investigarem quais fatores de risco de lesão e modelos de aprendizado de máquina podem ser utilizados, focando na obtenção de uma predição de lesão mais precisa.

7. **Blood sample profile helps to injury forecasting in elite soccer players - [Rossi et al., 2022]**

   O artigo [Rossi et al., 2022] realiza um estudo abordando que a coleta de amostras sanguíneas pode ajudar na previsão de lesões, juntamente à utilização de dados coletados com GPS. Com esse intuito, o perfil dos jogadores, de acordo com a combinação das variáveis, pode ajudar a personalizar o modelo de aprendizado de máquina, aumentando sua capacidade de detectar o risco de lesões dos jogadores. A amostra é caracterizada por dezoito jogadores com idades entre vinte a vinte e nove anos na temporada de 2017/2018 e 2018/2019 e os métodos utilizados foram: (i) DT; (ii) XGB.

Conforme os resultados, o algoritmo XGB foi o melhor modelo para prever lesões com todo o dataset utilizado no estudo, apesar da classificação com DT também ter demonstrado boa performance em comparação ao modelo base usado no estudo. Atingindo a acurácia de 63%, a pesquisa demonstrou que a habilidade preditiva de lesão aumentou 15% em comparação a modelos que consideram apenas variáveis coletadas com GPS relacionadas à carga de treino.

Conclusivamente, Rossi et al. [2022] aponta que a análise com amostras de sangue é uma aproximação do estado de saúde dos jogadores de futebol, que permite traçar o perfil destes e personalizar as regras que preveem o risco de lesão individual. Visto isso, especialistas de campo em clubes de futebol não devem apenas monitorar as cargas de trabalho (treino e/ou jogo) para avaliar o status dos jogadores, mas também informações adicionais derivadas das características, ajudando a ter uma visão completa do bem-estar deles. Assim, consequentemente induzindo à criação de melhores cronogramas que visem maximizar o efeito dos treinos e minimize o risco de lesões.

8. **Predictive Modeling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players - [Martins et al., 2022]**

   Nesta pesquisa de Martins et al. [2022], o objetivo principal foi analisar a modelagem preditiva do risco de lesão com base em vinte e dois potenciais fatores de risco referentes a informações subjetivas dos jogadores (posição setorial, idade, experiência, histórico de lesões), parâmetros antropométricos de composição corporal (peso, altura, gordura corporal) e testes específicos de aptidão física (flexibilidade, força geral, força explosiva, velocidade, agilidade e resistência aeróbica). A variável alvo da previsão foi o número de lesões por temporada, podendo assim encontrar valores referentes a frequência lesiva da temporada. Foram coletados os dados de trinta e seis jogadores de futebol masculino do time CS Marítimo (Primeira Liga do Futebol Português) ao longo da temporada de 2020/2021. Os métodos regressores aplicados de Aprendizado de Máquina incluíam: (i) Ordinary Least Square (OLS); (ii) Ridge Model; (iii) LASSO; (iv) Elastic Net (ENET) e (v) Stepwise Forward (SF).

   Com base nos resultados, pôde-se observar a relevância de alguns aspectos referentes a frequência lesiva por região do corpo e tipo de lesão. As lesões nos membros inferiores foi a área mais comum (85.2%), com principalmente entorses (35.2%) e distensões mus-

culares (35.2%) nos tornozelos (29.4%), quadríceps (11.7%) e isquiotibiais (11.7%).

Conclusivamente, os autores enfatizam a necessidade de estudos referentes à identificação de fatores de risco para predição de lesões no futebol profissional afim de melhor performance de um modelo preditivo. Como contribuição, o estudo aponta que todas as variáveis dos jogadores relacionadas à posição em campo, composição corporal e obtidas por testes de aptidão física foram consideradas importantes potenciais preditores lesivos.

Na Tabela 3, observa-se a relação entre os métodos de aprendizado de máquina e as variáveis utilizadas, presentes nos artigos descritos.

Aliadamente às pesquisas apontadas, o trabalho atual busca conciliar informações de GPS e algoritmos de aprendizado de máquina, para prever lesões sem contato entre jogadores profissionais. Sendo assim, os artigos de Rossi et al. [2018]; Vallance et al. [2020]; Piłka et al. [2023] possuem maior proximidade a esse intuito. Principalmente, esses trabalhos consideram as sessões profissionais com as tecnologias GPS, enquanto que os outros utilizam de outras formas assíncronas a esse método. Para a modelagem, todos os três utilizam da Validação Cruzada para treinamento e validação dos modelos preditivos, além de se apresentarem semelhanças com os algoritmos de aprendizado de máquina utilizados, principalmente o de DT.

Particularmente, além da associação de variáveis-chave coletadas com GPS e algoritmos semelhantes para composição do modelo, o trabalho vigente busca utilizar de estratégias que possam incrementar a performance dos modelos de classificação gerados. Sendo assim, foram aplicadas ferramentas como *Undersampling* para lidar com o desbalanceamento das classes, enquanto que as pesquisam utilizam de *Oversampling* da variável dependente. Junto a isso, foram testados conceitos relacionados à multicolinearidade e divisão de componentes principais PCA ao conjunto de dados. Para o melhor modelo preditivo, foi analisada a relevância de cada atributo dos atletas que o compôs.

Por fim, com a combinação de parâmetros em uma função classificadora desenvolvida, resultou-se uma quantidade alta de combinações diferentes. Sendo assim, pôde-se analisar a importância e o impacto positivo ou negativo que os parâmetros da função evidenciaram, por meio de um modelo de regressão que realizou a predição dos valores de *F1*, classificada como a métrica avaliativa principal.

Estudo	Variáveis utilizadas	Algoritmos utilizados
[Rossi et al., 2018]	Subjetivas de composição corporal, Carga de treino com GPS, tempo jogado, histórico lesivo	DT, RF, LR
[Piłka et al., 2023]	Posição setorial, histórico lesivo, Carga de treino/jogo com GPS	XGB
[Vallance et al., 2020]	Subjetivas de composição corporal, Carga de treino com GPS e fatores intrínsecos através de questionários	KNN, LDA, LR, Ridge, GNB, DT, RF, SVM, MLP, XGB
([Eetvelde et al., 2021])	Histórico lesivo, carga de treino e composição corporal mais se repetem nos artigos da revisão	Modelos em árvore (DT principalmente), SVM e ANN
[Kolodziej et al., 2023]	Neuromusculares e biomecânicas	LASSO
[Jauhiainen et al., 2022]	Demográficas, neuromusculares, biomecânicas, anatômicas e genéticas	LR, RF, SVM
[Rossi et al., 2022]	Carga de treino/jogo com GPS e Amostras sanguíneas	DT e XGB
[Martins et al., 2022]	Subjetivas, composição corporal e testes de aptidão física	OLS, Ridge, LASSO, ENET, SF

Tabela 3: Características descritivas dos modelos feitos nos estudos selecionados.

Capítulo 4

Metodologia

Neste Capítulo, serão abordados os assuntos a respeito da metodologia que foi utilizada para a condução dos experimentos. As etapas contemplam: (i) Proposta da pesquisa; (ii) Obtenção do conjunto de dados por meio de Pré-processamento; (iii) Métodos de criação do modelo e avaliação;

4.1 Proposta

Este é um trabalho de viés exploratório com objetivo principal de construir modelos de aprendizado de máquina multidimensionais focados em prever, através das cargas de treino e jogo, a probabilidade de um jogador ter uma lesão aguda e sem contato no microciclo, dentro do escopo temporal de 2021 e 2022. Especificamente, para o objetivo proposto, foi utilizado como inspiração o artigo de Rossi et al. [2018]. Portanto, o algoritmo *baseline* do trabalho foi com DT, além de também ser realizada a comparação de desempenho preditivo usando os modelos RF e LR, analogamente ao artigo base [Rossi et al., 2018], que obteve relevantes resultados. Junto a isso, a pesquisa possui o intuito específico de analisar a importância e o impacto dos parâmetros utilizados entre os modelos preditivos, além de uma análise dos fatores de risco considerados na melhor predição obtida.

Previamente à criação dos modelos multidimensionais, o trabalho também possui o objetivo específico de encontrar potenciais fatores de risco, por meio de procedimentos estatísticos de análise bivariada.

4.2 Conjuntos de dados

Para o estudo, foram disponibilizadas informações privadas coletadas pelo Fluminense Football Club relacionadas aos seus atletas masculinos profissionais, entre os anos de 2017 a 2022. Os dados referentes aos valores da carga de treinos e jogos dos atletas foram coletados por meio de dispositivos GPS implantados em coletes justos vestíveis e registrados em um banco de da-

dos em nuvem disponíveis através da API Open Field da Catapult [Catapult, 2018] junto a uma chave de acesso (*token*) criptografada oferecida pelo clube, como forma de segurança dos dados.

Em paralelo, o clube ofertou uma planilha preenchida manualmente com o histórico lesivo dos atletas do time entre 2017 a 2023. Foram registradas 185 lesões. As colunas foram divididas em: (i) Data da lesão; (ii) Data da liberação do Departamento Médico; (iii) Quantidade de dias lesionado; (iv) Nome do atleta; (v) Segmento do corpo afetado; (vi) Tipo de lesão (Aguda ou Crônica), (vii) Subtipo de Lesão e (viii) Grau de severidade da lesão conforme a quantidade de dias lesionado. Visando o objetivo principal do trabalho, foram selecionadas apenas as lesões do tipo agudas e sem contato, nos anos de 2021 e 2022. Sendo assim, contabilizou-se 52 episódios lesivos agudos com subtipos referentes à articulações, músculos e ligamentos (como visto na Tabela 4, contemplando principalmente distensões musculares e entorses em diversos segmentos da parte inferior do corpo dos atletas.

Subtipo de Lesão	**Quantidade**
Muscular	34
Articular	4
Ligamentar	1

Tabela 4: Quantidade de lesões por subtipo, de acordo com as lesões do tipo aguda selecionadas.

4.3 Pré-processamento

No trabalho proposto, a primeira etapa necessária a ser executada foi a de pré-processamento do conjunto de dados obtido com tecnologias GPS, com o intuito de prepará-lo para utilização no modelo. Nesse processo, foram demonstrados todos os critérios adotados para extração da base pela API da Catapult [Catapult, 2018], seleção de filtros para realização da limpeza dos dados que não foram utilizados no modelo e remoção de colunas totalmente nulas e zeradas. Após a limpeza, foi evidenciado o esquema de agrupamento dos dados em microciclos, contemplando a criação de atributos derivados que seriam originados por meio da seleção de variáveis intrínsecas ao conjunto de dados disponibilizado.

4.3.1 Extração e Limpeza dos dados

Com o propósito de extrair os dados GPS por meio da API REST [Catapult, 2018], a Catapult disponibiliza *endpoints*, nos quais o usuário consegue fazer uma requisição e o *endpoint* retorna as informações desejadas, de acordo com a especificação da chamada e o uso da chave de acesso. O uso do *token* neste contexto serve para direcionar o usuário ao acesso do banco de dados correto no qual foram registradas as informações dos atletas, pelo clube. Dessa forma, foi criada uma função para acessar o *endpoint* que retorna todas as informações coletadas a partir das atividades de treino e jogo dos atletas.

Após a extração, a limpeza dos dados iniciou-se através da seleção das informações especificadas no escopo da proposta do trabalho, por meio de filtros contendo palavras-chave relacionadas a características de viés profissional, além de uma limitação do escopo para os anos de 2021/2022. Ao final do processo, resultou-se em um conjunto de dados inicial com 44.354 linhas e 1.715 colunas, no qual cada linha refere-se a um período, que é classificado como uma subdivisão dos dados coletados em uma sessão de treino ou jogo de cada atleta. Nesse sentido, por exemplo, uma atividade de jogo no conjunto de dados pode contemplar até três períodos, sendo eles: (i) Etapa de Aquecimento; (ii) Primeiro tempo da partida oficial e (iii) Segundo tempo da partida oficial. O número de colunas é o resultado da soma de todos os parâmetros (também chamados de métricas), com todos os nomes dos atributos registrados do atleta ao utilizar os coletes GPS e seus atributos derivados, criados pelo clube. A variável-alvo do modelo foi classificada por meio da planilha de registros de lesões e a data que ocorreu o evento, cada atividade foi rotulada com: 1, caso o jogador tenha tido uma lesão no dia em que foi registrada a lesão e 0, se o jogador não tiver tido um episódio lesivo. No gráfico da (Figura 14) pode-se notar a distribuição da classificação entre lesões e não lesões relacionadas aos dias de atividade.

Dentre os dias que ocorreram treinos e jogos ao longo dos dois anos, 182 jogadores tiveram participação de partes desse escopo, conforme analisado no gráfico de distribuição, na Figura 15. Devido ao fato de uma grande parte dos atletas não terem uma quantidade relevante de atividades para constituir dados relevantes ao estudo, foi definida uma faixa mínima de 12 dias, preenchida por pelo menos duas semanas de trabalho, incluindo duas folgas. Paralelamente, foram removidas as informações referentes aos goleiros, devido ao fato destes atletas terem treinos e métricas específicas para essa posição, além de não participarem de treinos feitos para

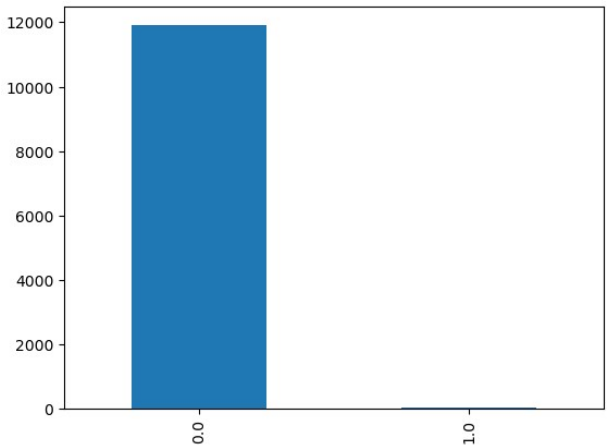

Figura 14: Gráfico de distribuição quantitativa entre as atividades com e sem lesão.

os outros companheiros em campo. Sendo assim, foram considerados os valores coletados de 79 jogadores, no total.

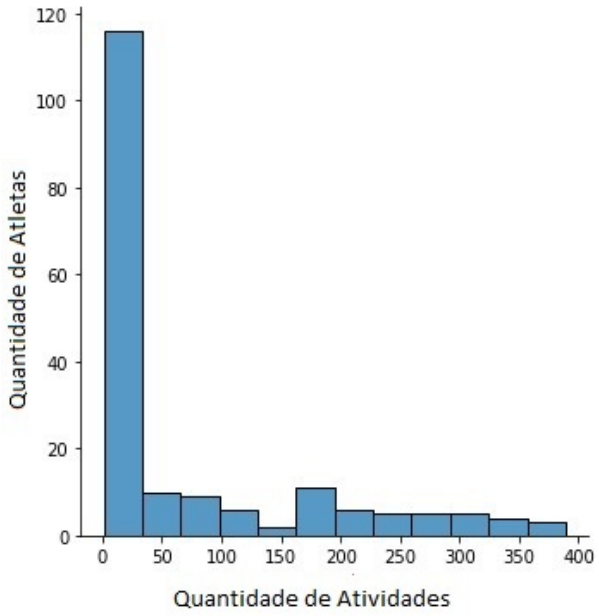

Figura 15: Gráfico de distribuição quantitativa entre atividades e atletas.

Para concluir a etapa de limpeza das informações, foram removidos 899 atributos constituídos inteiramente por valores zero ou nulo.

A gestão dos dados faltantes presentes em outros atributos não foi necessária pelo fato deles não terem sido selecionados para composição do modelo, conforme apresentado na próxima Subseção. Resumidamente, a atualização do conjunto de dados conforme o processo de limpeza e o saldo de períodos e atributos removidos pode ser observado na Tabela 5.

Descrição	Nº de Períodos	Nº de Atributos
Conjunto de dados pré limpeza	44.354	1.715
Remoção de atletas com < 12 atividades	43.618	1.715
Remoção de dados de goleiro	41.109	1.715
Remoção de colunas totalmente nulas/zeradas	41.109	801
Total removido	3.245	899

Tabela 5: Atualização do conjunto de dados conforme o processo de limpeza.

4.3.2 Microciclos e criação de Atributos derivados

Conforme visto na Figura 14, existe uma desproporcionalidade grande entre as sessões em que houveram lesões dos atletas, em comparação às que não ocorreram. Para um modelo classificador, um desbalanceamento de classes da variável-alvo pode ser prejudicial ao seu desempenho preditivo. Isso é explicado pelo fato do algoritmo não conseguir treinar resultados positivos quando é demonstrada uma quantidade tão baixa dos mesmos, afetando seu aprendizado mediante os dados utilizados. Assim, torna-se necessário o uso de medidas que reduzam a desproporcionalidade presente. Para o trabalho vigente, utilizaram-se duas estratégias com o intuito de amenizar o problema, sendo elas: (i) agrupamento dos períodos em microciclos e (ii) aplicação do método de *Undersampling*, aplicado na função que originou os modelos de classificação.

Inspirado no artigo do Piłka et al. [2023], que também utiliza da estratégia de microciclos, os períodos do conjunto de dados também foram agrupados dessa forma, diminuindo a quantidade de linhas no escopo e o desbalanceamento entre classes. Visto isso, os microciclos são preenchidos pela combinação de uma sequência de dias de treino e um dia de jogo de cada atleta, reiniciando a partir da atividade subsequente à partida oficial. Na Tabela 6, evidencia-se a proporcionalidade das classes conforme o agrupamento de períodos em atividades e microciclos, sendo este último o cerne do modelo classificador que será produzido. Os valores dos períodos foram destacados com N/A pois as lesões obtidas pela planilha de episódios lesivos correspondem ao dia de atividade em que ocorreram, não sendo possível saber especificamente qual período ocorreu a lesão.

Conjunto de dados	Sem lesão	Com lesão
Períodos	N/A	N/A
Atividades	11.908 (99.7%)	39 (0.3%)
Microciclos	4.287 (99.1%)	39 (0.9%)

Tabela 6: Proporção da quantidade de dados rotulados com e sem lesão, divididos em períodos, períodos agrupados pela atividade e períodos agrupados por microciclo.

Para compor as variáveis-chave do modelo, foram selecionados atributos específicos no conjunto de dados. A seleção (Tabela 7) foi feita baseada nos fatores de risco obtidos com GPS e utilizados na modelagem dos artigos de Vallance et al. 2020; Rossi et al. 2018; Piłka et al. 2023 (Seção 3.2), além das variáveis analisadas em um relatório oficial pós jogo disponibilizado pelo clube.

Nome	Descrição	Rossi et al. [2018]	Piłka et al. [2023]	Vallance et al. [2020]	Relatório
field_time	Tempo em min dentro de campo	X	X	X	X
total_distance	Distância total em m percorrida	X	X	X	X
meterage_per_minute	Distância total em m/min percorrida				X
velocity_band1_total_distance	Distância em m percorrida entre 0 a 1,1 km/h			X	
velocity_band2_total_distance	Distância em m percorrida entre 1,1 a 7,2 km/h			X	
velocity_band3_total_distance	Distância em m percorrida entre 7,2 a 14,4 km/h			X	
velocity_band4_total_distance	Distância em m percorrida entre 14,4 a 19,8 km/h			X	
velocity_band5_total_distance	Distância em m percorrida entre 19,8 a 25,2 km/h		X	X	
dist_>_19,8km/h	Distância em m percorrida acima de 19,8 km/h				X
dist_>_19,8km/h_m/min	Distância em m/min percorrida acima de 19,8 km/h				X
velocity_band6_total_distance	Distância em m percorrida acima de 25,2 km/h		X	X	X
dist_>_25,2km/h_m/min	Distância em m/min percorrida acima de 25,2 km/h				X
acel+desacel_ima_alta	Nº de Acelerações e Desacelerações com alta intensidade explosiva				X
acel+desacel_ima_alta_(min)	Nº de Acelerações e Desacelerações com alta intensidade explosiva por min				X
acel+desacel_qtd_>2ms	Nº de Acelerações e Desacelerações acima de 2m/s²	X	X	X	
acel+desacel_qtd_>_3ms	Nº de Acelerações e Desacelerações acima de 3m/s²	X	X		
total_player_load	Carga mecânica total realizada		X	X	
rhie_bout_count	Nº de Esforços Intensos Repetidos realizados				X
rhies_/_min	Nº de Esforços Intensos Repetidos realizados por minuto				X
mudanças_direção_totais	Nº de Mudanças de Direção do corpo				X
qtd_total_saltos	Nº de Saltos Verticais				X
max_vel	Velocidade Máxima atingida			X	X
max_effort_acceleration	Aceleração Máxima atingida				X

Tabela 7: Seleção dos atributos que darão origem às variáveis-chave utilizadas no modelo, conforme os artigos de Vallance et al. 2020; Rossi et al. 2018; Piłka et al. 2023 e o relatório pós jogo fornecido pelo clube.

Os atributos derivados são originados do resultado do agrupamento de variáveis especificamente selecionadas no conjunto de dados. Porém, como elas possuem classificações diferentes, o agrupamento deve respeitar o critério necessário, sendo divididos de três maneiras:

- **Critério 1**: O agrupamento é feito pela a soma de todos os valores do atributo no microciclo, para cada atleta.

$$Soma(atributo) \tag{1}$$

- **Critério 2**: O agrupamento é feito pela a soma de todos os valores do atributo no microciclo dividido pela tempo total em campo no microciclo, para cada atleta.

$$(Soma(atributo))/Soma(tempo\ em\ campo) \tag{2}$$

- **Critério 3**: O agrupamento é feito pelo valor máximo entre todos os valores do atributo no microciclo, para cada atleta.

$$Max(atributo) \tag{3}$$

Os atributos derivados obtidos podem ser observados na Tabela 8, conforme os atributos que foram utilizados para realizar o agrupamento (Variáveis de Cálculo) e o critério adotado para agrupá-los. Paralelamente, a variável-alvo tornou-se um episódio lesivo presente dentro de um microciclo, não mais sendo em um dia específico de atividade.

Posterior à criação dos atributos derivados pelo agrupamento dos períodos em microciclos (Tabela 8), criaram-se três novas variáveis-chave para compor o modelo, totalizando-se 26 atributos e uma variável-alvo. Elas foram:

1. **mc_Duração**: Contabiliza a quantidade de dias presentes em cada microciclo, como potencial fator de risco (apontado em Piłka et al. [2023]).

2. **Reincidencia_binario**: Atributo derivado da variável-alvo, evidenciando a categorização de uma reincidência lesiva (Rossi et al. 2018; Piłka et al. 2023).

3. **Reincidencia_soma**: Excepcionalmente, foi criada sem conformidade aos artigos e relatórios, na qual é composta pela soma acumulativa para cada vez que uma reincidência

Nome	Variáveis de Cálculo	Critério
mc_field_time	field_time	Eq. (1)
mc_tot_dist	total_distance	Eq. (1)
mc_tot_dist_min	total_distance e field_time	Eq. (2)
mc_vel1	velocity_band1_total_distance	Eq. (1)
mc_vel2	velocity_band2_total_distance	Eq. (1)
mc_vel3	velocity_band3_total_distance	Eq. (1)
mc_vel4	velocity_band4_total_distance	Eq. (1)
mc_vel5	velocity_band5_total_distance	Eq. (1)
mc_vel6	dist_>_19,8km/h	Eq. (1)
mc_vel6_min	dist_>_19,8km/h e field_time	Eq. (2)
mc_vel7	velocity_band6_total_distance	Eq. (1)
mc_vel7_min	velocity_band6_total_distance e field_time	Eq. (2)
mc_acel+desacel_alta	acel+desacel_ima_alta	Eq. (1)
mc_acel+desacel_alta_min	acel+desacel_ima_alta e field_time	Eq. (2)
mc_acel+desacel_>2ms	acel+desacel_qtd_>2ms	Eq. (1)
mc_acel+desacel_>3ms	acel+desacel_qtd_>_3ms	Eq. (1)
mc_carga_tot	total_player_load	Eq. (1)
mc_rhies	rhie_bout_count	Eq. (1)
mc_rhies_min	rhie_bout_count e field_time	Eq. (2)
mc_mud_dir	mudanças_direção_totais	Eq. (1)
mc_saltos	qtd_total_saltos	Eq. (1)
mc_max_vel	max_vel	Eq. (3)
mc_max_acel	max_effort_acceleration	Eq. (3)

Tabela 8: Agrupamento em microciclo das variáveis-chave que irão compor o modelo, conforme os atributos utilizados e o critério adotado.

lesiva acometer um atleta, ao longo do tempo. Dessa forma, paralelamente à variável de reincidência binária, há a possibilidade de avaliar a potencial influência de lesões acumuladas dos atletas como fator de risco para um novo episódio lesivo.

4.4 Modelagem e Avaliação

Primeiramente, com o intuito de encontrar potenciais fatores de risco, adotaram-se duas estratégias de análise bivariada, realizadas antes do desenvolvimento e aplicação dos modelos multi-dimensionais. As análises consistem em observar a relação entre cada atributo e a variável dependente, presentes no conjunto de dados. Para esse fim, aplicaram-se duas estratégias: (ii) Gráficos Exploratórios de distribuição; (i) Teste U de Man-Whitnney.

Por meio dos gráficos exploratórios, pode ser observado o comportamento das variáveis-chave dos atletas que tiveram lesões, pela distribuição de períodos com lesão e sem lesão. Já com os cálculos estatísticos obtidos pelo Teste U de Man Whitnney, a relevância dos atributos pode ser evidenciada pelo *p-value* resultante, conforme o intervalo de confiança (alfa) estabele-

cido.

Em seguida, foi feita uma função para produzir um modelo multi-dimensional de previsão, conforme a utilização dos parâmetros inseridos. O Algoritmo 1 descreve a sua aplicação, na qual requer sete parâmetros: *df*, *features*, *target*, *ml*, *n*, *cv* e *pca*. O parâmetro *df* corresponde ao conjunto de dados que será utilizado, enquanto que *features* são os atributos específicos que foram selecionados para o modelo e *target* se refere à variável-alvo. O parâmetro *ml* refere-se aos algoritmos de Aprendizado de Máquina que foram utilizados (inspirados no artigo de Rossi et al. [2018]). O valor *n* está relacionado ao valor resultante da redução do conjunto de dados, através da estratégia de *Undersampling*. O parâmetro *cv* atribui a quantidade de divisões feitas pela Validação Cruzada no processo de treinamento e validação do modelo. Finalmente, *pca* constitui o uso, ou não, da aplicação do algoritmo de PCA ao conjunto de dados. Caso o valor do parâmetro *pca* seja zero, significa que o algoritmo não é utilizado no conjunto de dados. No contrário, o PCA é aplicado, com o valor especificado equivalente ao número de componentes principais. A descrição de cada parâmetro pode ser observada na Tabela 9.

Inicialmente, a Função *classification_predictions* (Algoritmo 1) realiza o processamento do conjunto de dados *df* com os atributos (*features*) e variável-alvo (*target*) especificados por parâmetro (Linha 2), resultando no *df_processed*. Caso o valor de *pca* passado seja diferente de zero, o *df_processed* é obtido por meio da aplicação do PCA ao conjunto de dados (*df*), dividido pela quantidade de componentes principais especificada, visto nas Linhas 3 e 4.

Na Linha 7 é feito o procedimento de *Undersampling*, reduzindo o tamanho de linhas do conjunto de dados, de acordo com o valor inserido no parâmetro *n*. Assim, o resultado do *df_processed* após a redução do número de linhas é dividido entre as variáveis *X* e *target*, compostos pelos atributos e a variável-alvo que irão compor o modelo, respectivamente.

Na Linha 8, é associada à uma variável *result*, o resultado do treinamento e validação do modelo, através de uma função específica chamada *cross_validation* (Algoritmo 2), na qual utiliza os parâmetros: *ml*, *X*, *target*, *cv*. A função *cross_validation* realiza a Validação Cruzada, na Linha 3, conforme a proporção de divisão do conjunto em treino e teste especificada em *cv*. Por fim, a função retorna o resultado da média dos testes por meio de quatro métricas de performance avaliativa, presentes na lista *scores* (Linha 2): (i) Acurácia; (ii) Precisão; (iii) *F1*; (iv) *Recall*; (v) *AUC*. Vale ressaltar que a medida de desempenho principal para o trabalho será com *F1*, pois consiste na média harmônica entre a Precisão e o *Recall*, sendo assim, uma forma mais robusta de análise dos resultados gerados. A partir da Figura 16, evidencia-se a estrutura

resumida da metodologia principal relacionada à criação do modelo preditivo (Algortimo 1).

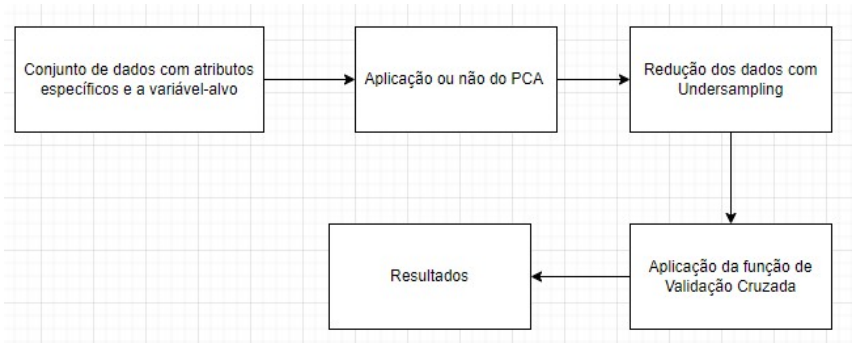

Figura 16: Resumo da metodologia principal aplicada para criação do modelo preditivo.

Algoritmo 1: Metodologia para Criação do Modelo Preditivo

function *classification_predictions(df, features, target, ml, n, cv, pca)*:
    $df\_processed \leftarrow copy(df, features, target)$
    if $pca \; != 0$ **then**
        $df\_processed \leftarrow get\_pca\_components(df, features, target)$
    end

    $X, target \leftarrow Undersampling(df\_processed, n)$
    $result \leftarrow cross\_validation(ml, X, target, cv)$
return *result*

Nome	Descrição [valores]
df	Conjunto de dados obtido ao fim da Seção 4.3
features	Lista de Atributos específicos escolhidos
target	Valor referente à Variável-Alvo
ml	Algoritmos de Aprendizado de Máquina utilizados, conforme Rossi et al. [2018]
n	Quantidade de amostras com aplicação do *Undersampling*
cv	Quantidade de divisões de treino-teste feita na Validação Cruzada
pca	Aplicação (ou não) do PCA e divisão em componentes principais, conforme a quantidade selecionada

Tabela 9: Parâmetros usados.

Algoritmo 2: Metodologia de Validação e Avaliação

function *cross_validation (ml, X, target, cv)*:
    $scores \leftarrow [accuracy, precision, f1, recall, AUC]$
    $result \leftarrow CrossValidation(ml, X, target, cv, scores)$
return *result*

Em primeira instância, o resultado *baseline* do projeto foi definido pela aplicação do Algoritmo 2 ao conjunto de dados obtido após o Pré-processamento, utilizando o algoritmo de

aprendizado de máquina *baseline* DT. Após esse processo, criaram-se modelos classificadores com a intenção de melhorar os resultados obtidos no *baseline*, conforme uma iteração do produto cruzado dos sete parâmetros da Função *classification_predictions*, demonstrado pela metodologia presente no Algoritmo 3. Na iteração, não foram apontados os parâmetros df e *target* por serem fixos como o conjunto de dados e a variável dependente, sendo apenas utilizados na chamada da função. Além disso, o número de divisões feitas apresentado pelo parâmetro *cv* também foi fixo, permitindo comparação entre o *baseline* e os modelos desenvolvidos. Paralelamente, os parâmetros restantes foram associados a um conjunto de possibilidades para cada, vistos como F, M, N, P referentes a $features$, ml, n e pca, respectivamente.

Algoritmo 3: Metodologia do Produto Cruzado

```
results ← ∅
foreach features ∈ F, ml ∈ M, n ∈ N, pca ∈ P do
    results ← results ∪ classification_predictions(df, features, target, ml, n, cv, pca)
end
```

Para ser feita a análise dos potenciais fatores de risco de lesão mais importantes entre os modelos preditivos produzidos, considerou-se o melhor resultado de todos. Dessa forma, o intuito foi observar o impacto positivo ou negativo de cada atributo em relação à lesão, ao decorrer do desenvolvimento do melhor modelo avaliado.

Por último, para analisar a relevância dos parâmetros combinados por produto cruzado no Algoritmo 3 que deram origem aos modelos preditivos, conduziu-se um modelo de regressão. O intuito da sua condução explica-se por conta da praticidade em analisar o impacto positivo ou negativo das combinações de parâmetros que originaram os modelos preditivos. Sendo assim, torna-se possível analisar o impacto real no desempenho dos modelos em um único escopo, ao invés de ser feita uma observação individual para cada classificação criada. A aplicação metodológica de análise individual foi adotada apenas para o melhor resultado, por conta da observação específica dos atributos parametrizados classificados como potenciais fatores de risco de lesão, como mencionado previamente.

Especificamente, os resultados dos modelos de classificação foram divididos em um conjunto de dados composto por quatro atributos categóricos e uma variável dependente. As quatro variáveis categóricas consistem nos valores presentes nos conjuntos F, M, N, P, chamados de Atributos, Algoritmo, Undersampling e PCA, respectivamente. De forma paralela, o *F1* obtido de cada modelo classificador foi a variável-alvo, definido como o indicador principal de perfor-

mance para o trabalho. A partir disso, cada categoria dos atributos categóricos foi considerada como um valor binário. Essa estratégia consiste na codificação *one-hot*, na qual transforma os valores categóricas em colunas para cada categoria, apresentando 1 quando tenha a presença da categoria ou caso contrário, 0.

A regressão foi desenvolvida com o algoritmo RF, no qual realizou a previsão da variável dependente do modelo classificador, *F1*. Dessa forma, a avaliação de performance entre os resultados de *F1* obtidos da classificação e o previsto com RF foi analisada por meio da métrica de Erro Percentual Absoluto Médio. Por fim, desenvolveu-se um gráfico que evidenciou as variáveis mais relevantes em questão de impacto, positivo ou negativo, aos resultados.

Capítulo 5

Avaliações Experimentais

Neste Capítulo, serão feitos testes práticos conforme as diretrizes metodológicas abordadas no Capítulo anterior. Sendo assim, haverá o direcionamento para duas vertentes: (i) potenciais fatores de risco de lesão resultados da análise bivariada entre os atributos e variável-alvo; (ii) Construção e performance preditiva dos modelos de aprendizado de máquina propostos. O projeto foi inteiramente implementado em Python 3.9.12 com Jupyter Notebooks foi feita a utilização das bibliotecas: pandas 2.1.0, numpy 1.25.2, seaborn 0.12.2, scipy 1.11.2, plotly 5.16.1, scikit-learn 1.3.0, matplotlib 3.7.2 e shap 0.43.0. A respeito dos dispositivos em que foram realizadas todas as etapas do projeto, utilizou-se: (i) um computador com sistema operacional Windows 10 Pro de 64 bits na versão 22H2 com processador Intel(R) Core(TM) i3-10100 de 3.60GHz e 12GB de memória RAM instalada; (ii) um computador com sistema operacional Windows 11 Pro de 64 bits na versão 22H2 com processador 12th Gen Intel(R) Core(TM) i5-12400F de 2.50GHz e 16GB de memória RAM instalada.

5.1 Análise Bivariada

Previamente ao desenvolvimento de modelos preditivos multivariados, serão abordados nessa Seção as análises bivariadas entre atributos e variável dependente obtidas através dos Gráficos Exploratórios e do Teste U de Mann Whitnney. Dessa forma, é possível ver os potenciais fatores de risco e o comportamento dos mesmos mediante episódios lesivos.

Inicialmente, evidencia-se pela Figura 17, o comportamento de cada um dos 26 atributos presentes no conjunto de dados, nas vertentes apresentando ou não lesão, ao longo dos microciclos existentes. O gráfico violino é uma união entre histograma e *boxplot*. Dessa forma, a linha tracejada representa a mediana dos valores numéricos de cada variável, que foram normalizados devido à distribuição dos dados.

Observa-se que, por conta do desbalanceamento de classes, o comportamento dos atributos em contextos lesivos não demonstra evidente alteração dentro do escopo de dados total. Apenas mc_vel7 e mc_vel7_min apresentam um uma redução considerável por conta de terem tido

valor 0 em três momentos em que a variável dependente foi 1. Enquanto isso, as variáveis com enfoque foram mc_rhies_min e mc_tot_dist_min devido à distância entre as medianas, apontando um aumento previamente a um quadro lesivo.

Em paralelo aos gráficos de distribuição, o Teste U de Mann Whitnney foi calculado para cada relação entre atributo e variável-alvo. O valor definido para o alfa foi de 0.05. Sendo assim, as variáveis com *p-value* menor ou igual à esse valor foram considerados potenciais fatores de risco. A partir do gráfico de dispersão com as 26 variáveis, foi traçada uma linha para determinar o valor mais aproximado à 0.05, evidenciado na Figura 18. Portanto, a única variável em destaque no gráfico foi mc_rhies_min, confirmando-se junto ao gráfico violino, a possibilidade dessa variável como um fator de risco de lesão.

5.2 Configuração experimental

Considerando as metodologias descritas nos Algoritmos 1 e 2 e parâmetros descritos na Tabela 8, é feito todo o processo de predição com o modelo *baseline* e as diversas combinações resultadas do produto cruzado entre todos os valores possíveis dos parâmetros.

Vale ressaltar que, conforme a criação dos atributos Reincidencia_binario e Reincidencia_soma, todos os modelos consistiram na utilização de apenas uma das duas variáveis por vez. Isso é explicado pelo fato delas serem fortemente correlacionadas, pois Reincidencia_soma é derivada de Reincidencia_binario. Dessa maneira, a utilização de ambas juntas é redundante e potencialmente impactaria na performance do modelo. Sendo assim, para cada modelo existe uma divisão em duas vertentes, sendo elas com Reincidencia_binario ou com Reincidencia_soma.

5.2.1 *Baseline*

A priori, para originar o *baseline*, utilizou-se a Metodologia do Algoritmo 2. A Validação Cruzada foi realizada no conjunto de dados com *X* igual à todos os atributos e *target* como a variável-alvo. Junto a isso, o parâmetro *ml* foi preenchido pelo algoritmo DT (*baseline* escolhido) e o número de divisões feitas entre conjuntos de treino e teste (*cv*) foi definido como 5.

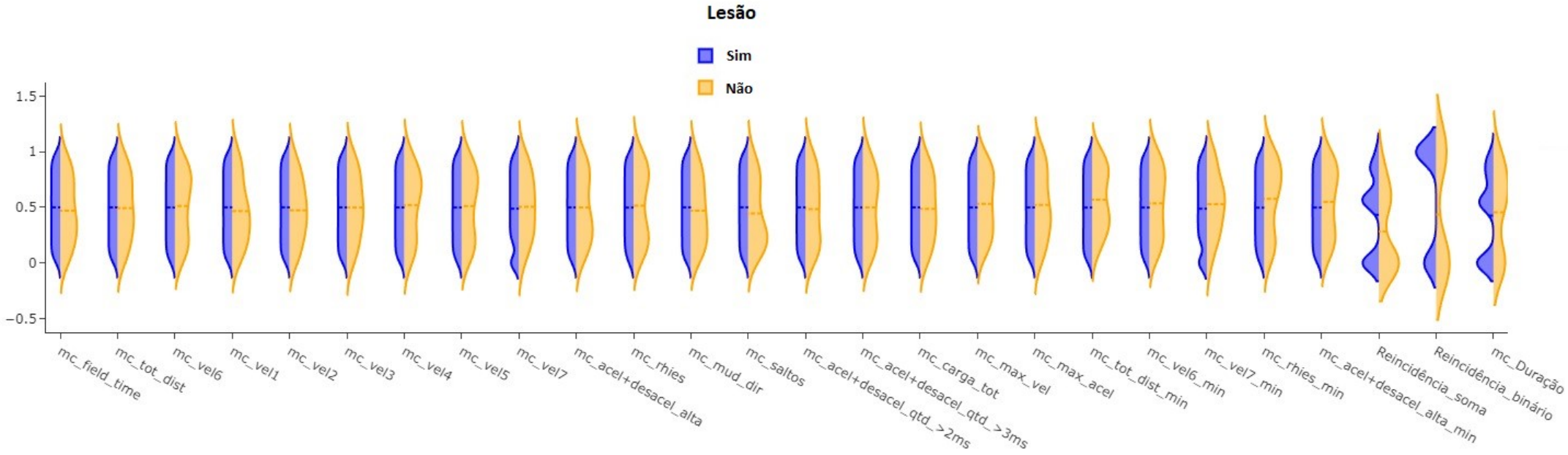

Figura 17: Gráfico violino do demonstrando o comportamento dos valores numéricos normalizados dos 26 atributos em contextos com lesão e sem lesão.

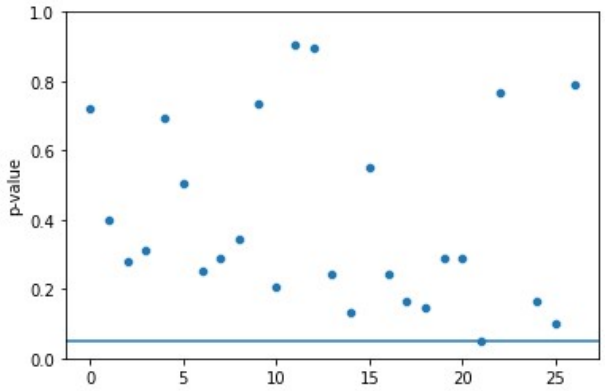

Figura 18: Gráfico de dispersão dos *p-value* obtidos dos 26 atributos através do teste estatístico de Man Whitnney U.

5.2.2 Modelagem Principal

Para a metodologia do Algoritmo 1, dentre os sete parâmetros presentes, *df* e *target* consistem em valores fixos. Isso se explica pois *df* é o escopo do conjunto de dados após o pré-processamento e com atributos especificados pelo parâmetro *features*, enquanto que *target* é a variável dependente. Intrinsecamente à proposta do trabalho, o parâmetro *ml* irá ser composto pelo *baseline* DT, além dos modelos LR e RF.

O parâmetro *pca* é composto por nove valores diferentes, no qual ditam a aplicação do PCA. Os dados são divididos entre o número de componentes principais escolhido, variando a possibilidade do modelo ser criado com uma a oito componentes principais, mesmo que a variância não mude significativamente após a divisão de duas variáveis, conforme visto na soma cumulativa da Tabela 10. Lembrando que, caso o valor seja zero, significa que o algoritmo do PCA não é aplicado ao conjunto de dados e são utilizados os atributos passados pelo parâmetro *features* no modelo.

Nº de Componentes	**Variância**
1	99.3%
2	99.9%
3	99.9%
4	99.9%
5	99.9%
6	99.9%
7	99.9%
8	99.9%

Tabela 10: Soma cumulativa da variância dos atributos do conjunto de dados, de acordo com o número de componentes.

Para amenizar o desbalanceamento de classes, o valor disponibilizado ao parâmetro *n* irá considerar a remoção das linhas em que a variável dependente não seja positiva, de forma aleatória, para essa quantidade especificada. Visto isso, foram consideradas três possibilidades para o *Undersampling*: (i) Redução para 130 amostras, com proporção de 70% sem lesões e 30% com lesões; (ii) Redução para 98 amostras, com proporção de 60% sem lesões e 40% com lesões; (iii) Redução para 130 amostras, com proporção de 50% sem lesões e 50% com lesões. Enquanto isso, o parâmetro *cv* irá respeitar um valor fixo de 5 divisões, assim como evidenciado no *baseline*.

Finalmente, $features$ foi preenchido por 30 grupos diferentes de atributos que foram considerados para cada modelo, sendo divididos em três partes e evidenciados na Tabela 12. Todos os grupos foram duplicados e classificados pela utilização dos atributos Reincidencia_Soma ou Reincidencia_binario. Na primeira parte, os dois primeiros grupos consistem na utilização de todos os atributos do conjunto de dados. A partir disso, como tentativa de melhora da performance dos resultados, foi considerada a multicolinearidade presente entre os atributos com correlação acima de 95% (Tabela 11), afim de remover variáveis que prejudiquem a performance. Dessa maneira, a segunda parte contempla 12 composições na qual apenas uma das variáveis correlacionadas é desconsiderada, realizando esse procedimento uma por uma. Como pode-se observar na Tabela 11, mc_carga_tot, mc_tot_dist, mc_field_time e mc_vel2 se repetem três vezes entre as correlações, enquanto que mc_vel6 e mc_vel5 apresentam uma correlação entre si. Sendo assim, seguindo o mesmo conceito de multicolinearidade, a terceira parte consiste em grupos com combinações de atributos em que considerou-se todos os atributos com apenas uma das quatro variáveis que se repetem e mc_vel6 ou mc_vel5, resultando em 16 possibilidades.

Atributo 1	**Atributo 2**	**Correlação**
mc_vel6	mc_vel5	99.1%
mc_carga_tot	mc_tot_dist	98.9%
mc_tot_dist	mc_vel2	96.8%
mc_vel2	mc_field_time	96.6%
mc_carga_tot	mc_vel2	95.7%
mc_tot_dist	mc_field_time	95.7%
mc_carga_tot	mc_field_time	95.7%

Tabela 11: Demonstração da multicolinearidade entre os atributos com correlação acima de 95%, ordenado pelo valor da correlação.

Nome do grupo	Descrição	RB	RS
Todos_RB	Todos os atributos	X	
Todos_RS	Todos os atributos		X
mcr1_RB	Todos menos mc_vel6	X	
mcr1_RS	Todos menos mc_vel6		X
mcr2_RB	Todos menos mc_carga_tot	X	
mcr2_RS	Todos menos mc_carga_tot		X
mcr3_RB	Todos menos mc_tot_dist	X	
mcr3_RS	Todos menos mc_tot_dist		X
mcr4_RB	Todos menos mc_vel2	X	
mcr4_RS	Todos menos mc_vel2		X
mcr5_RB	Todos menos mc_vel5	X	
mcr5_RS	Todos menos mc_vel5		X
mcr6_RB	Todos menos mc_field_time	X	
mcr6_RS	Todos menos mc_field_time		X
mcr7_RB	Todos menos mc_vel6, mc_carga_tot, mc_tot_dist, mc_vel2	X	
mcr7_RS	Todos menos mc_vel6, mc_carga_tot, mc_tot_dist, mc_vel2		X
mcr8_RB	Todos menos mc_vel5, mc_carga_tot, mc_tot_dist, mc_vel2	X	
mcr8_RS	Todos menos mc_vel5, mc_carga_tot, mc_tot_dist, mc_vel2		X
mcr9_RB	Todos menos mc_vel6, mc_field_time, mc_tot_dist, mc_vel2	X	
mcr9_RS	Todos menos mc_vel6, mc_field_time, mc_tot_dist, mc_vel2		X
mcr10_RB	Todos menos mc_vel5, mc_field_time, mc_tot_dist, mc_vel2	X	
mcr10_RS	Todos menos mc_vel5, mc_field_time, mc_tot_dist, mc_vel2		X
mcr11_RB	Todos menos mc_vel6, mc_field_time, mc_carga_tot, mc_vel2	X	
mcr11_RS	Todos menos mc_vel6, mc_field_time, mc_carga_tot, mc_vel2		X
mcr12_RB	Todos menos mc_vel5, mc_field_time, mc_carga_tot, mc_vel2	X	
mcr12_RS	Todos menos mc_vel5, mc_field_time, mc_carga_tot, mc_vel2		X
mcr13_RB	Todos menos mc_vel6, mc_field_time, mc_carga_tot, mc_tot_dist	X	
mcr13_RS	Todos menos mc_vel6, mc_field_time, mc_carga_tot, mc_tot_dist		X
mcr14_RB	Todos menos mc_vel5, mc_field_time, mc_carga_tot, mc_tot_dist	X	
mcr14_RS	Todos menos mc_vel5, mc_field_time, mc_carga_tot, mc_tot_dist		X

Tabela 12: Conjunto de grupos de atributos considerados no modelo através do parâmetro *features*. RB = Com Reincidencia_binario; RS = Com Reincidencia_Soma.

Algoritmo 4: Avaliação Experimental

$df \leftarrow df$
$target \leftarrow target$
$cv \leftarrow 5$
$F \leftarrow \{Todos\_RB, Todos\_RS, ..., mcr14\_RB, mcr14\_RS\}$
$M \leftarrow \{DT, RF, LR\}$
$N \leftarrow \{130, 98, 78\}$
$P \leftarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$
$results \leftarrow \emptyset$
foreach $features \in F, ml \in M, n \in N, pca \in P$ **do**
    $results \leftarrow results \cup classification\_predictions(df, features, target, ml, n, cv, pca)$
end

5.3 Resultados Experimentais

Com todos os parâmetros definidos para a modelagem principal, o produto cruzado resultou em 2430 combinações diferentes, composto pelos valores únicos (df, $target$, cv) e os conjuntos de valores (F, M, N, P). A aplicação da Metodologia do Produto Cruzado (Algoritmo 3) que definiu os resultados pode ser vista no Algoritmo 4. Com intuito analítico dos resultados principais, foram destacados na Tabela 14 os seis melhores desempenhos obtidos entre os 2430 modelos e, em seguida, os três melhores modelos de cada um dos algoritmos de aprendizado de máquina utilizados (Tabela 15). Vale ressaltar que, todos os resultados foram ordenados pela métrica avaliativa *F1*. Como base comparativa inicial, foi apontada na Tabela 13 o resultado da aplicação metodológica com Validação Cruzada para originar o *baseline*.

Algoritmo	Acurácia	Precisão	*Recall*	*F1*	*AUC*	RB	RS
DT	97.5%	2.4%	5.4%	3.2%	51.8%	X	
DT	97.4%	1.7%	2.9%	2.1%	50.6%		X

Tabela 13: Resultado *baseline*, com Validação Cruzada. RB = Com Reincidencia_binario; RS = Com Reincidencia_Soma.

Atributos	Algoritmo	Acurácia	Precisão	*Recall*	*F1*	*AUC*	*Under-sampling*	*PCA*
mcr3_RB	DT	71.9%	71.9%	79.6%	74.7%	71.8%	78	6
mcr3_RS	DT	71.9%	71.9%	79.6%	74.7%	71.8%	78	6
mcr8_RB	RF	70.7%	67.8%	80%	73.2%	71.5%	78	5
mcr8_RS	RF	70.7%	67.8%	80%	73.2%	71.5%	78	5
mcr8_RS	RF	68.2%	66.5%	79.6%	72.1%	67.7%	78	4
mcr8_RB	RF	68.2%	66.5%	79.6%	72.1%	67.7%	78	4

Tabela 14: Seis melhores modelos de classificação, ordenados por *F1*.

5.3.1 Análise dos Resultados

A partir da visualização do resultado *F1* do *baseline*, nota-se uma performance preditiva muito baixa, principalmente por conta do desbalanceamento de classe desproporcional entre episódios com e sem lesões. Dessa forma, comparativamente, o *F1* resultado do melhor modelo obtido evidenciou um aumento significativo de 71.5% com o atributo Reincidencia_binario e 72.6% com Reincidencia_soma, ambos tendo performance preditiva de 74.4%.

Analiticamente, dentre os seis melhores resultados ordenados por *F1*, podem ser destaca-

Atributos	Algoritmo	Acurácia	Precisão	*Recall*	*F1*	*AUC*	*Under-sampling*	*PCA*
mcr3_RB	DT	71.9%	71.9%	79.6%	74.7%	71.8%	78	6
mcr3_RS	DT	71.9%	71.9%	79.6%	74.7%	71.8%	78	6
mcr1_RB	DT	64.2%	60.2%	82.5%	69.1%	64.1%	78	7
mcr8_RB	RF	70.7%	67.8%	80%	73.2%	71.5%	78	5
mcr8_RS	RF	70.7%	67.8%	80%	73.2%	71.5%	78	5
mcr8_RS	RF	68.2%	66.5%	79.6%	72.1%	67.7%	78	4
mcr9_RS	LR	64.2%	77.3%	55.7%	59.8%	63.9%	78	1
mcr10_RB	LR	64.2%	77.3%	55.7%	59.8%	63.9%	78	1
mcr10_RS	LR	64.2%	77.3%	55.7%	59.8%	63.9%	78	1

Tabela 15: Três melhores modelos de classificação, divididos por cada algoritmo utilizado e ordenados por *F1*.

das informações relevantes a respeito dos parâmetros resultados. Primeiramente, conforme os 30 grupos de atributos utilizados na modelagem, o critério de multicolinearidade se mostrou necessário ao modelo, ao invés de apenas utilizar o grupo que consiste em todos os atributos do conjunto de dados. Junto a isso, a segregação de grupos de atributos com exclusão um a um, entre os fortemente correlacionados, se mostrou melhor principalmente com o mcr_3, no qual é removido apenas mc_tot_dist. Além disso, os últimos resultados foram destacados pela presença do grupo de atributos com mcr_8, que consiste na remoção específica dos valores mc_vel5, mc_carga_tot, mc_tot_dist e mc_vel2, mantendo apenas os atributos multi colineares mc_vel6 e mc_field_time. Vale ressaltar que, todos os modelos foram distribuídos com os atributos Reincidencia_binario ou Reincidencia_soma e não notou-se diferença no desempenho dos melhores resultados com essa estratégia.

Considerando a questão de desbalanceamento de classes, a aplicação da redução do conjunto de dados à uma proporção de 50:50 (78 amostras) foi dominante para os resultados mais promissores. Dessa forma, entende-se que o modelo realizou a etapa de treino e teste pela Validação Cruzada com uma distribuição bem balanceada dos dados classificados com lesões e não-lesões.

Analogamente ao *Undersampling*, o PCA também foi dominante em questão da sua aplicação ao conjunto de dados, visto pelas métricas mais relevantes. Especificamente, o valor do parâmetro *pca* igual à seis indicou uma melhor performance. Considerando que foram combinados oito valores diferentes para a divisão de componentes principais com a aplicação do PCA, de um a oito, os melhores resultados evidenciados apontaram que os melhores resultados não tiveram divisão de sete ou oito componentes em sua composição. Concomitantemente, nota-se

que os resultados melhoram consideravelmente com a utilização de um número de componentes principais maior que dois, apesar da variância cumulativa não diferenciar muito como visto na Tabela 10.

A respeito dos algoritmos utilizados, nota-se que o melhor modelo de todos foi através do algoritmo *baseline* DT, assim como destacado no artigo base de Rossi et al. [2018]. De maneira comparativa, as modelagens em árvore com DT e RF foram as de maior destaque, com valores avaliativos próximos a partir de diferentes distribuições de componentes principais e grupos de atributos. Sobre os melhores resultados obtidos com LR, apesar de menores que os obtidos com DT e RF, notaram-se desempenhos interessantes. Curiosamente, o algoritmo utilizou de apenas uma componente principal, na qual explica 99.3% dos dados do conjunto principal, como apontado na Tabela 10.

5.4 Importância dos Parâmetros

Conforme os resultados obtidos das 2430 combinações com o produto cruzado entre os parâmetros inseridos na Avaliação Experimental do Algoritmo 4, foi possível realizar análises referentes ao impacto de cada valor inserido. Para isso, foi aplicado o artificio de valores *SHAP*, no qual apresenta uma explicação nítida do impacto positivo ou negativo no resultado preditivo (*output*), conforme os valores dos atributos em cada linha do modelo. Os valores *SHAP* são medidos pela magnitude, que representa o nível de impacto.

Dessa maneira, a análise da importância e de impacto positivo ou negativo dos parâmetros foi dividida em duas vertentes. A primeira consiste na análise individual do modelo com melhor desempenho, observando especificamente a parametrização do grupo de atributos utilizado, no qual apresenta potenciais fatores de risco de lesão. Em segunda instância, foi desenvolvido um modelo de regressão com RF para analisar graficamente a relevância de cada parâmetro inserido na função preditiva entre as 2430 combinações.

5.4.1 Potenciais Fatores de Risco de Lesão

Para análise do parâmetro referente ao grupo de atributos utilizados no melhor desempenho, utilizou-se a metodologia de valores *SHAP* para observar a relevância de cada um. Porém, nota-se na Tabela 13 que o melhor modelo aplicou o algoritmo PCA aos atributos inseridos, dividindo-os em seis componentes principais. Dessa forma, em primeira instância não há como

obter conclusões diretas a respeito da relação entre o grupo de atributos inseridos por parâmetro e a influência nos quadros lesivos.

Visto isso, é possível analisar o impacto positivo ou negativo das componentes principais transformadas em valores *SHAP*. Evidenciado no gráfico da Figura 19, é feita uma divisão entre a relevância de cada componente principal, onde os pontos apresentam colorações diferentes referentes ao valor dos atributos. A coloração dos pontos varia entre vermelho representando os maiores valores até o azul retratando os menores valores. Junto a isso, quanto mais para a direita, maior impacto positivo no resultado enquanto que para a esquerda, maior impacto negativo, de acordo com os valores *SHAP* obtidos. As componentes principais estão ordenadas por ordem de relevância, ou seja, quais possuem maior impacto no modelo e distruibuição dos seus valores.

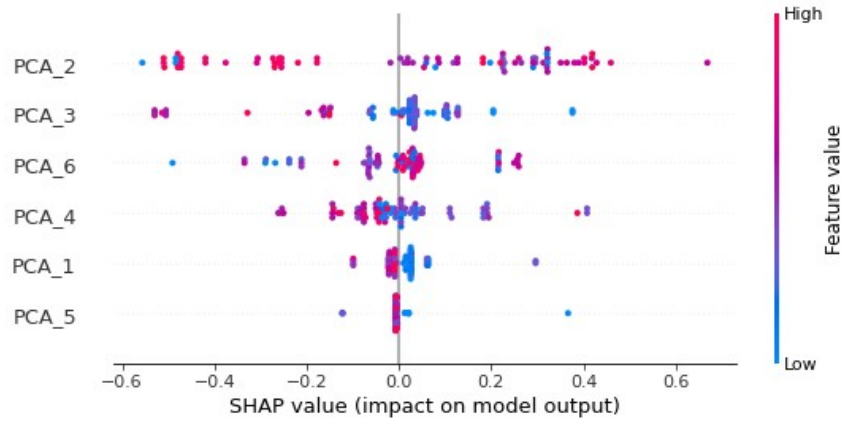

Figura 19: Gráfico de impacto das componentes principais do melhor modelo preditivo com SHAP.

Em seguida, por meio de um conjunto de dados preenchido pelos valores *SHAP* das componentes principais e o resultado preditivo (*output*), junto ao grupo de atributos do melhor modelo (antes de serem transformados com PCA), é possível calcular a correlação entre os valores. Assim, entende-se a potencial assimilação das variáveis à episódios lesivos.

Pela Tabela 16, nota-se que alguns atributos possuem correlação interessante com o resultado preditivo e pela quantidade de variáveis presentes no modelo, há uma boa distribuição das correlações. Como exemplo, baseando-se nos valores igual ou acima de 20%, evidencia-se que existe uma assimilação positiva entre as variáveis e o *output*, isto é, caso essas métricas aumentem seu valor, há um risco maior de lesão. Logo, Rencidencia_binario, mc_rhies_min, mc_tot_dist_min, mc_max_vel e mc_vel3 são considerados potenciais fatores de risco de lesão, de acordo com a análise feita entre as correlações.

Dentre os dois valores mais correlacionados com o *output*, pode-se ressaltar algumas informações relevantes. Primeiramente, o atributo Reincidencia_binario teve um maior destaque referente à sua correlação com quadros lesivos, assim como em textos de proximidade alta ao tema (mencionados na Seção 3), como os de Rossi et al. [2018], Piłka et al. [2023] e Vallance et al. [2020]. Além disso, o atributo mc_rhies_min foi destacado com uma potencial ligação à lesões, assim como observado na análise bivariada da Seção 5.1, com gráficos exploratórios e Teste U de Mann Whitnney.

Atributos	PCA1	PCA2	PCA3	PCA4	PCA5	PCA6	Output
Rencidencia_binario	-0.02	0.07	0.11	-0.11	-0.01	0.19	0.27
mc_rhies_min	-0.28	0.28	0.29	0.18	-0.23	-0.02	0.24
mc_tot_dist_min	-0.21	0.25	0.28	-0.25	-0.28	0.18	0.24
mc_max_vel	-0.38	0.34	0.22	0.07	-0.30	-0.17	0.22
mc_vel3	-0.54	0.30	0.28	-0.02	-0.45	-0.05	0.20
mc_rhies	-0.54	0.31	0.25	0.13	-0.33	-0.09	0.19
mc_vel4	-0.54	0.32	0.28	-0.02	-0.47	-0.08	0.19
mc_acel+desacel_>3ms	-0.41	0.32	0.17	0.16	-0.25	-0.06	0.17
mc_vel7	-0.48	0.28	0.20	0.03	-0.25	-0.23	0.17
mc_max_acel	-0.20	0.26	0.13	0.20	-0.14	-0.03	0.16
mc_vel6_min	-0.32	0.21	0.30	-0.06	-0.24	-0.28	0.15
mc_vel6	-0.55	0.27	0.28	0.01	-0.31	-0.29	0.15
mc_acel+desacel_>2ms	-0.54	0.31	0.15	0.09	-0.28	-0.07	0.15
mc_vel7_min	-0.28	0.25	0.19	0	-0.17	-0.25	0.14
mc_vel5	-0.55	0.27	0.27	0	-0.33	-0.27	0.13
mc_acel+desacel_alta_min	-0.28	0.05	0.18	0.09	-0.05	0.03	0.12
mc_tot_dist	-0.65	0.29	0.16	0	-0.29	-0.09	0.12
mc_vel2	-0.61	0.27	0.15	0.11	-0.27	-0.09	0.12
mc_acel+desacel_alta	-0.60	0.21	0.13	0.10	-0.21	-0.04	0.11
mc_field_time	-0.59	0.25	0.06	0.07	-0.22	-0.11	0.05
mc_mud_dir	-0.58	0.20	0.07	0.01	-0.24	-0.07	0.04
mc_vel1	-0.38	0.21	-0.25	0.06	0.21	-0.03	-0.03
mc_saltos	-0.59	0.09	0.03	0.08	-0.08	-0.14	0.01
mc_Duracao	-0.38	-0.02	0.05	-0.01	-0.28	-0.06	-0.01

Tabela 16: Correlação entre os atributos antes da transformação do PCA com valores *SHAP* das componentes principais e o resultado preditivo (*output*). Todos são decorrentes do melhor modelo preditivo obtido e evidenciado na Tabela 15.

5.4.2 Modelo de Regressão

Para o modelo de regressão, foi utilizado o algoritmo RF e foram selecionados a um novo conjunto de dados o resultado das combinações do produto cruzado entre os parâmetros inseridos na Avaliação Experimental do Algoritmo 4. Com as 2430 linhas e cinco colunas, o conjunto

foi dividido entre as variáveis categóricas referentes aos Atributos, Algoritmo, *Undersampling* e PCA, enquanto que o alvo preditivo foi a métrica *F1*. Antes da etapa de regressão, as categorias de cada variável categórica passaram por um pré-processamento por meio da estratégia de codificação *one-hot*. Dessa maneira, o Erro Percentual Absoluto Médio obtido resultante do modelo foi de 5.5%, evidenciando uma baixa porcentagem de erro do valor previsto, o que indica uma boa performance do modelo de regressão desenvolvido.

Pelo fato dos valores dos atributos serem binários, o gráfico na Figura 20 evidencia apenas duas cores, vermelho apontando o uso da variável no modelo e azul demonstrando a não utilização. Dessa maneira, pela lógica do gráfico observa-se que vermelho para direita aponta impacto positivo do seu uso e para esquerda, impacto negativo. Consequentemente, a cor azul para a direita significa impacto positivo na não utilização da variável, enquanto que para esquerda é o contrário.

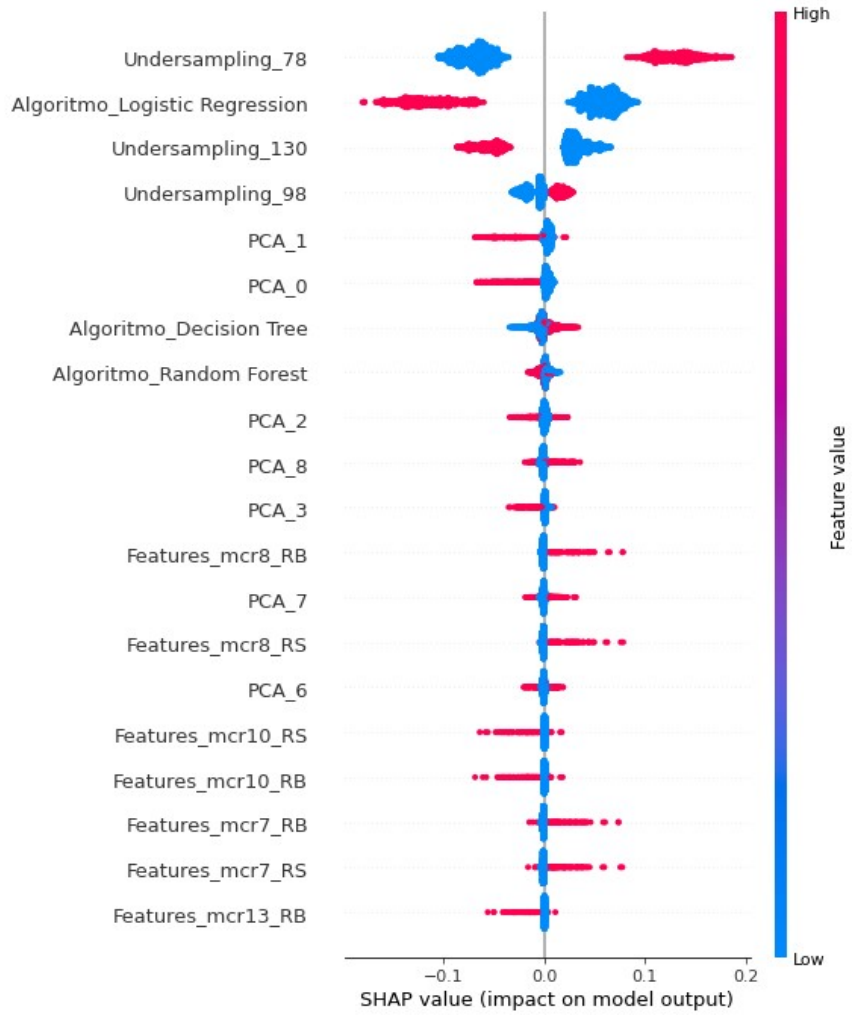

Figura 20: Gráfico de impacto dos valores parametrizados do modelo de regressão.

A partir do gráfico da Figura 20, nota-se a influência na performance do modelo à alguns

atributos. A utilização do uso de estratégias de *Undersampling* relacionados à redução das amostras a 130 ou 98 indicam majoritariamente um impacto negativo ao modelo, enquanto que não utilizá-los implica em uma influência positiva. Já com 78 amostras, a utilização do *Undersampling* se comportou de forma contrária, como visto nos melhores resultados da Tabela 14. Isso aponta a necessidade do balanceamento de classes em prol de melhores resultados, como foi foi feito no trabalho.

Sobre o algoritmo PCA, evidencia-se que a melhor escolha é de aplicá-los ao modelo em prol de melhor desempenho avaliativo, como visto no resultado negativo quando o parâmetro equivale a zero. Junto a isso, o acréscimo de divisão dos dados em mais componentes principais influenciou em melhores resultados.

Considerando os algoritmos de predição DT, RF e LR, o ponto principal extraído no gráfico foi o grande impacto negativo da performance com a utilização da LR. Apesar disso, LR não teve porcentagens ruins em suas melhores métricas. Em seguida, pode-se notar que a utilização do algoritmo RF também demonstrou um impacto negativo no desempenho, de forma consideravelmente menor que com LR. Enquanto isso, o algoritmo DT teve uma relação positiva com a melhora de performance. Dessa forma, principalmente, confirma-se a relação de DT à bons resultados como apontado pelo artigo base de Rossi et al. [2018], que também o usou como *baseline*.

Por fim, analisando modelos que considerem multicolinearidade aos atributos utilizados, nota-se principalmente o impacto negativo dos modelos que utilizem dos grupos de atributos mcr10_RB, mcr10_RS e mcr13_RB. De forma contrária, os grupos mcr8_RB, mcr8_RS, mcr7_RB e mcr7_RS foram destacados por terem relevante impacto positivo, a partir do seu uso nos modelos.

Capítulo 6

Considerações Finais

6.1 Resumo dos Capítulos

No Capítulo 1, foram abordadas perspectivas sobre o cenário esportivo e sua aliança a ferramentas de Ciência de Dados. Ressalta-se também o contexto lesivo no futebol e a necessidade de adotar medidas preventivas para melhorá-lo nesse aspecto. Assim, ao final do Capítulo introdutório, apontou-se que o objetivo principal deste trabalho foca-se na utilização de um conjunto de dados relacionados a atletas profissionais de futebol, com o propósito de prever a probabilidade de episódios lesivos sem contato que possam acometê-los em um microciclo, através da associação de algoritmos de aprendizado de máquina e variáveis-chave dos jogadores - possibilitando, assim, que o clube adote medidas necessárias que possibilitem evitar uma lesão.

No Capítulo 2, foram explicados conceitos imprescindíveis ao discernimento e compreensão de todo o escopo que o trabalho e seus objetivos possuem em seu cerne. Nesse contexto, abordou-se as definições conceituais de lesão, abrangendo seus tipos e uma explicação relevante sobre os fatores de risco de lesão pertencentes aos atletas e ao ambiente que estes estão inseridos. Paralelamente, foram ponderadas informações sobre aprendizado de máquina e os algoritmos que serão utilizados no presente trabalho, além da contextualização de conceitos importantes para compreender suas aplicações mediante o objetivo do trabalho.

No Capítulo 3, realizou-se uma busca sistemática com a ferramenta PRISMA [Page et al., 2021], possibilitando o levantamento de artigos específicos relevantes para o trabalho vigente. Para isso, foram definidos diversos critérios de inclusão e exclusão, além de serem definidas prioridades, conforme o resultado final da busca. Os artigos de prioridade alta foram destacados como os estudos mais relevantes por assemelharem-se ao objetivo do trabalho. Através de uma análise e comparação dos artigos relacionados, pôde-se extrair informações descritivas e conclusivas importantes.

No Capítulo 4, evidenciou-se a proposta do trabalho abordada de maneira específica e objetiva. Em seguida, foi apresentada uma descrição a respeito dos conjuntos de dados disponibilizados para o pré-processamento, compondo as informações que preencheram os modelos

preditivos desenvolvidos. Após essa etapa, foram apresentadas as estratégias metodológicas aplicadas nos algoritmos 1, 2 e 3, evidenciando todas as etapas relacionadas à modelagem e avaliações.

Finalmente, no Capítulo 5, todas as informações foram primeiramente instanciadas e depois aplicadas aos algoritmos informados na metodologia. Dessa maneira, foi desenvolvido o modelo *baseline* do projeto, além de diversas combinações diferentes de modelos preditivos, por meio do produto cruzado entre conjuntos de parâmetros. Sequencialmente, foi possível analisar os resultados obtidos e extrair informações importantes relacionadas à performance e a importância dos parâmetros utilizados.

6.2 Contribuições

Desde o princípio, o texto possui informações imprescindíveis para o discernimento do contexto de lesões. Junto a isso, são consideradas contribuições científicas da literatura que providenciam conteúdos relevantes ao cenário, a partir de uma busca sistemática de trabalhos relacionados. Adicionalmente, é adotada uma metodologia que possibilita a aplicação de diversas combinações de parâmetros para compor uma grande quantidade de modelos distintos, com o intuito de aprimorar o desempenho preditivo. Após isso, ainda foi possível apontar os principais parâmetros que impactaram em um melhora ou piora dos resultados gerados.

No total, foram originados 2430 modelos preditivos, oriundos do produto cruzado entre as possibilidades com sete parâmetros principais. O melhor resultado obtido foi o *F1* de 74.4% com o algoritmo de DT. Com isso, nota-se a relevância do uso do algoritmo de árvore de decisão em prol de métricas interessantes, assim como o artigo base de Rossi et al. [2018], que obteve uma precisão por volta de 50%. Paralelamente, o modelo RF também é um algoritmo de aprendizado de máquina interessante a ser utilizado.

Ressalta-se que os modelos foram desenvolvidos em duas vertentes de atributos, sendo consideradas as reincidências lesivas de duas maneiras diferentes para cada caso, no qual uma delas parte da literatura a respeito da existência ou não de uma reincidência lesiva do atleta, enquanto a outra é uma nova proposta relacionada à soma de reincidências de cada atleta. Porém, para casos em que foi utilizado o algoritmo PCA, não se notou evidente diferença no desempenho por meio da utilização de cada uma, enquanto que nos casos em que não houve aplicação do PCA, curiosamente houve diferença.

Além disso, evidenciou-se a aplicação de conceitos interessantes à melhora de performance

dos modelos preditivos, como o uso de PCA com diversas componentes e balanceamento de classes por meio do *Undersampling*. Vale ressaltar também a importância de considerar grupos diferentes de atributos conforme a multicolinearidade.

Finalmente, por meio da aplicação metodológica dos valores *SHAP*, pôde-se entender potenciais fatores de risco de lesão, obtidos conforme os parâmetros do melhor modelo obtido. Para esse fim, foi feito o cálculo de correlação entre os valores do grupo de atributos presentes na melhor performance com o impacto positivo ou negativo das componentes principais e o resultado preditivo. Dessa forma, observou-se que Rencidencia_binario, mc_rhies_min, mc_tot_dist_min, mc_max_vel e mc_vel3 são considerados potenciais fatores de risco de lesão, de acordo com a análise feita entre as correlações.

6.3 Trabalhos futuros

Este trabalho pode ser estendido em diversas vertentes possíveis de melhorias ou novas perspectivas. Primeiramente, podem ser incluídos fatores de risco intrínsecos dos atletas, que foram coletados pelo clube, além dos dados extrínsecos de treino e jogo com GPS. Como exemplo, existem testes bioquímicos obtidos por meio de amostras sanguíneas dos jogadores, assim como evidenciado no trabalho de Rossi et al. [2022] e dados de Percepção Subjetiva de Esforço (PSE). Para o escopo temporal, seria interessante adicionar o ano de 2023 em que o clube apresentou participações em mais atividades de treino e teste no ano.

Apesar da quantidade relativamente baixa de episódios lesivos presentes no conjunto de dados disponibilizado, também é possível explorar e estudar relações dos dados com diferentes tipos de lesões. Além disso, ressalta-se que caso os episódios lesivos fossem pontuados precisamente conforme o período que ocorreu ao invés de classificá-los em um dia de atividade, seria possível utilizar os dados sem precisar agrupá-los por atividades ou microciclos. Dessa maneira, além de não precisar criar atributos derivados com critérios específicos de agrupamento, poderia haver a aplicação do algoritmo de PCA à todos os atributos presentes no conjunto de dados original para compor um modelo preditivo experimental.

Finalmente, para analisar a performance dos modelos classificadores de maneira conjunta, por meio de um modelo de regressão (como visto no trabalho), podem ser consideradas como variável-alvo base outras métricas além do *F1*, como a Precisão, *Recall* ou *AUC*.

Bibliografia

Alin, A. (2010). Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, 2(3):370–374.

Allen, M. P. (1997). The problem of multicollinearity. *Understanding regression analysis*, pages 176–180.

Arlot, S. and Lerasle, M. (2016). Choice of v for v-fold cross-validation in least-squares density estimation. *The Journal of Machine Learning Research*, 17(1):7256–7305.

Arnason, A., Sigurdsson, S. B., Gudmundsson, A., Holme, I., Engebretsen, L., and Bahr, R. (2004). Risk Factors for Injuries in Football. *American Journal of Sports Medicine*, 32(SUPPL. 1):5–16.

Asuero, A. G., Sayago, A., and González, A. (2006). The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59.

Belle, V. and Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4(July):1–25.

Bergkamp, T. L., Niessen, A. S. M., den Hartigh, R. J., Frencken, W. G., and Meijer, R. R. (2019). Methodological Issues in Soccer Talent Identification Research. *Sports Medicine*, 49(9):1317–1335.

Berrar, D. et al. (2019). Cross-validation.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Catapult (2018). Catapult api overview. https://docs.connect.catapultsports.com/reference/introduction.

Court-Brown, C. M., Wood, A. M., and Aitken, S. (2008). The epidemiology of acute sports-related fractures in adults. *Injury*, 39.

Cuevas, I. F., Carmona, P. G., Quintana, M. S., and Arnaiz-Lastras, J. (2021). Economic costs estimation of soccer injuries in first and second spanish division professional teams View project ANTROPOMETRY View project. (May 2014):1–4.

Cullen, F. T., Myer, A. J., and Latessa, E. J. (2009). Eight Lessons from moneyball: The high cost of ignoring evidence-based corrections. *Victims and Offenders*, 4(2):197–213.

Dvorak, J., Junge, A., Chomiak, J., Graf-Baumann, T., Peterson, L., Rösch, D., and Hodgson, R. (2000). Risk factor analysis for injuries in football players: Possibilities for a prevention program. *American Journal of Sports Medicine*, 28(5 SUPPL.).

Eetvelde, H. V., Mendonça, L. D., Ley, C., Seil, R., and Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, 8(1).

Ekstrand, J., Hägglund, M., and Waldén, M. (2011). Injury incidence and injury patterns in professional football: The UEFA injury study. *British Journal of Sports Medicine*, 45(7):553–558.

Ekstrand, J., Spreco, A., Bengtsson, H., and Bahr, R. (2021). Injury rates decreased in men's professional football: An 18-year prospective cohort study of almost 12 000 injuries sustained during 1.8 million hours of play. *British Journal of Sports Medicine*, 55(19):1084–1091.

Elsevier (2020). Scopus: Guia de Referência Rápida. pages 1–16.

Fiscutean, A. (2021). Data scientists are predicting sports injuries with an algorithm. *Nature*, 592(7852):S10–S11.

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., and Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100.

Hägglund, M., Waldén, M., Bahr, R., and Ekstrand, J. (2005). Methods for epidemiological study of injuries to professional football players: Developing the UEFA model. *British Journal of Sports Medicine*, 39(6):340–346.

Hägglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson, H., and Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: An 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, 47(12):738–742.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

Hughes, T., Riley, R., Callaghan, M., and Sergeant, J. (2020). The Value of Preseason Screening for Injury Prediction: The Development and Internal Validation of a Multivariable Prognostic Model to Predict Indirect Muscle Injury Risk in Elite Football (Soccer) Players. *Sports Medicine - Open*, 6(1).

Inklaar, H. (1994). Soccer Injuries: I: Incidence and Severity. *Sports Medicine: Evaluations of Research in Exercise Science and Sports Medicine*, 18(1):55–73.

Jauhiainen, S., Aÿrämö, S., Forsman, H., and Kauppi, J. P. (2019). Talent identification in soccer using a one-class support vector machine. *International Journal of Computer Science in Sport*, 18(3):125–136.

Jauhiainen, S., Kauppi, J.-P., Krosshaug, T., Bahr, R., Bartsch, J., and Äyrämö, S. (2022). Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes. *American Journal of Sports Medicine*, 50(11):2917–2924.

Johansson, U., Bostrom, H., Lofstrom, T., and Linusson, H. (2014). Regression conformal prediction with random forests. *Machine learning*, 97:155–176.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Kingsford, C. and Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, 26(9):1011–1012.

Kirkendall, D. T. and Dvorak, J. (2010). Effective injury prevention in soccer. *Physician and Sportsmedicine*, 38(1):147–157.

Kleinbaum, D. G. and Klein, M. (2010). *Logistic Regression*. Springer New York.

Kolodziej, M., Groll, A., Nolte, K., Willwacher, S., Alt, T., Schmidt, M., and Jaitner, T. (2023). Predictive modeling of lower extremity injury risk in male elite youth soccer players using least absolute shrinkage and selection operator regression. *Scandinavian Journal of Medicine and Science in Sports*, (February 2022):1–13.

Kraemer, W. J. and Häkkinen, K. (2004). *Treinamento de força para o esporte*. Artmed.

Liu, A. Y.-c. (2004). The effect of oversampling and undersampling on classifying imbalanced text datasets.

LM, N. N. (2012). PubMed Basics. (December):1–8.

MacFarland, T. W., Yates, J. M., MacFarland, T. W., and Yates, J. M. (2016). Mann–whitney u test. *Introduction to nonparametric statistics for the biological sciences using R*, pages 103–132.

Majumdar, A., Bakirov, R., Hodges, D., Scott, S., and Rees, T. (2022). Machine Learning for Understanding and Predicting Injuries in Football. *Sports Medicine - Open*, 8(1).

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Marchiori, D. M. (2014). Chapter 10 - trauma. In Marchiori, D. M., editor, *Clinical Imaging (Third Edition)*, pages 625–765. Mosby, Saint Louis, third edition edition.

Martins, A. S. and de Oliveira, A. L. (2021). A periodização do treinamento desportivo. *Anais da Jornada Científica dos Campos Gerais*, 19(1).

Martins, F., Przednowek, K., França, C., Lopes, et al. (2022). Predictive Modeling of Injury Risk Based on Body Composition and Selected Physical Fitness Tests for Elite Football Players. *Journal of Clinical Medicine*, 11(16).

Meeuwisse, W. H., Tyreman, H., Hagel, B., and Emery, C. (2007). A dynamic model of etiology in sport injury: The recursive nature of risk and causation. *Clinical Journal of Sport Medicine*, 17(3):215–219.

Miller, T. W. (2016). *Sports Analytics and Data Science*.

Nassis, G., Verhagen, E., Brito, J., Figueiredo, P., and Krustrup, P. (2023). A review of machine learning applications in soccer with an emphasis on injury risk. *Biology of Sport*, 40(1):233–239.

Nick, T. G. and Campbell, K. M. (2007). *Logistic Regression*, pages 273–301. Humana Press, Totowa, NJ.

NIST Big Data Public Working Group (2015). NIST Special Publication 1500-1 - NIST Big Data Interoperability Framework: Volume 1, Definitions. *NIST Special Publication*, 1:32.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, 372.

Patel, D., Shah, D., and Shah, M. (2020). The Intertwine of Brain and Body: A Quantitative Analysis on How Big Data Influences the System of Sports. *Annals of Data Science*, 7(1):1–16.

Pfirrmann, D., Herbst, M., Ingelfinger, P., Simon, P., and Tug, S. (2016). Analysis of injury incidences in male professional adult and elite youth soccer players: A systematic review. *Journal of Athletic Training*, 51(5):410–424.

Piłka, T., Grzelak, B., Sadurska, A., Górecki, T., and Dyczkowski, K. (2023). Predicting injuries in football based on data collected from gps-based wearable sensors. *Sensors*, 23(3).

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., and Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264.

Rossi, A., Pappalardo, L., Filetti, C., and Cintia, P. (2022). Blood sample profile helps to injury forecasting in elite soccer players. *Sport Sciences for Health*, 19(1):285–296.

Services, E. E. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley.

Sikka, R. S., Baer, M., Raja, A., Stuart, M., and Tompkins, M. (2019). Analytics in sports medicine: implications and responsibilities that accompany the era of big data. *jbjs*, 101(3):276–283.

Studnicka, A. (2020). The emergence of wearable technology and the legal implications for athletes, teams, leagues and other sports organizations across amateur and professional athletics. *DePaul J. Sports L.*, 16:i.

Taimela, S., Kujala, U. M., and Osterman, K. (1990). Intrinsic Risk Factors and Athletic Injuries. *Sports Medicine*, 9(4):205–215.

Vallance, E., Sutton-Charani, N., Imoussaten, A., Montmain, J., and Perrey, S. (2020). Combining internal- and external-training-loads to predict non-contact injuries in soccer. *Applied Sciences (Switzerland)*, 10(15).

Viru, A. (1991). Principios básicos aplicables a la construccción de macrociclos. *Buenos Aires*.

Walker, B. (2007). *The anatomy of sports injuries : your illustrated guide to prevention, diagnosis and treatment*. North Atlantic Books.

Zheng, Z., Cai, Y., and Li, Y. (2015). Oversampling method for imbalanced classification. *Computing and Informatics*, 34(5):1017–1037.