CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

ANÁLISE DE FATORES QUE IMPACTAM NO FUTEBOL BASEADA EM CIÊNCIA DE DADOS.

DANIEL DE OLIVEIRA CAPANEMA

Orientador: Adriano César Machado Pereira
Centro Federal de Educação Tecnológica de Minas Gerais

BELO HORIZONTE
FEVEREIRO DE 2023

DANIEL DE OLIVEIRA CAPANEMA

ANÁLISE DE FATORES QUE IMPACTAM NO FUTEBOL BASEADA EM CIÊNCIA DE DADOS.

Defesa de Tese apresentado ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito para a obtenção do título de Doutor em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional

Linha de pesquisa: Métodos Matemáticos Aplicados

Orientador: Adriano César Machado Pereira
Centro Federal de Educação Tecnológica de Minas Gerais

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL
BELO HORIZONTE
FEVEREIRO DE 2023

Capanema, Daniel de Oliveira
C236a Análise de fatores que impactam no futebol baseada em ciência de dados / Daniel de Oliveira Capanema. – 2023.
69 f.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional.
Orientador: Adriano César Machado Pereira.
Tese (doutorado) – Centro Federal de Educação Tecnológica de Minas Gerais.

1.Ciências do esporte – Processamento de dados – Teses. 2. Futebol – Treinamento técnico – Brasil – Teses. 3. Inteligência artificial – Teses. 4. Desempenho – Teses. 5. Ferimentos e lesões – Teses. I. Pereira, Adriano César Machado. II. Centro Federal de Educação Tecnológica de Minas Gerais. III. Título.

CDD 006.3081

Elaboração da ficha catalográfica pela bibliotecária Jane Marangon Duarte, CRB 6º 1592 / Cefet/MG

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
COORDENAÇÃO DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

"ANÁLISE DE FATORES QUE IMPACTAM NO FUTEBOL BASEADA EM CIÊNCIA DE DADOS".

Tese de Doutorado apresentada por **Daniel de Oliveira Capanema**, em 27 de fevereiro de 2022, ao Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG, e aprovada pela banca examinadora constituída pelos professores:

Documento assinado digitalmente
gov.br ADRIANO CESAR MACHADO PEREIRA
Data: 26/05/2023 16:34:42-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Adriano César Machado Pereira
Universidade Federal de Minas Gerais

Prof. Dr. Adriano Lima Alves
Fortaleza Esporte Clube

Documento assinado digitalmente
gov.br JOAO GUSTAVO DE OLIVEIRA CLAUDINO
Data: 29/05/2023 14:30:35-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. João Gustavo de Oliveira Claudino
Universidade Federal do Piauí

Documento assinado digitalmente
gov.br PAULO ROBERTO PEREIRA SANTIAGO
Data: 05/06/2023 14:28:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Paulo Roberto Pereira Santiago
Universidade de São Paulo

Profª. Drª. Elizabeth Fialho Wanner
Centro Federal de Educação Tecnológica de Minas Gerais

Documento assinado digitalmente
gov.br FLAVIO LUIS CARDEAL PADUA
Data: 29/05/2023 09:24:01-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Flávio Luis Cardeal Pádua
Centro Federal de Educação Tecnológica de Minas Gerais

Visto e permitida a impressão,

Prof. Dr. Thiago de Souza Rodrigues
Coordenador do Programa de Pós-Graduação Stricto Sensu em Modelagem Matemática e Computacional

Dedico esse trabalho a toda minha família, sobretudo meus pais, esposa e filha.

Agradecimentos

Agradeço a todos que incentivaram, torceram ou contribuíram com a elaboração desta tese. Agradecimento especial ao meu orientador Prof. Adriano Pereira.

Resumo

Esportes profissionais e de alto desempenho estão cada vez mais utilizando novas tecnologias para obter melhores resultados em treinamentos e competições. Estudos para prevenir lesões e melhorar o desempenho têm sido aplicados e bons resultados já foram alcançados. Para atingir os objetivos, técnicas de inteligência artificial são aplicadas aos dados para encontrar padrões, prever tendências, melhorar resultados, táticas, equipamentos e outros. Este trabalho analisa dados reais de clubes da primeira divisão do futebol brasileiro, aplicando métodos de inteligência artificial para classificar jogadores, encontrar padrões, variáveis ou características que influenciam nos resultados, definindo características que influenciam na ocorrência de lesões e no desempenho dos atletas. Jogadores foram classificados e comparações foram feitas em grupos que possuíam ou não jogadores lesionados. Também foram realizadas análises comparativas em grupos de jogadores com melhores e piores médias de eficiência física, mostrando que jogadores com melhores médias tendem a apresentar características de treino mais parecidas com as do jogo, enquanto jogadores com piores médias apresentam números no treinamento inferiores aos dos jogos. Também foram estudadas as principais características que influenciam nos resultados dos jogos, sendo uma delas a diferença de performance entre o primeiro e o segundo tempo, considerando variáveis relacionadas à distância. O classificador XGBoost conseguiu bons resultados ao prever esta diferença de performance, além de apresentar como e quais variáveis influenciam nesta diferença. O trabalho também apresenta revisões sistemáticas de inteligência artificial em esportes em equipes e individuais, trazendo informações relevantes e métricas sobre o assunto.

Palavras-chave: Inteligência Artificial, Ciência do Esporte, Ciência de Dados, Performance, Lesão, Treino.

Abstract

Professional and high-performance sports are increasingly using new technologies to obtain better results in training and competition. Studies to prevent injuries and improve performance have been applied and good results have already been achieved. To achieve the goals, artificial intelligence techniques are applied to the data to find patterns, predict trends, improve results, tactics, equipment and others. This work analyzes real data from clubs in the first division of Brazilian football, applying artificial intelligence methods to classify players, find patterns or characteristics that influence the results and define players who will perform better between the two game times. Players were classified and comparisons were made in groups with and without injured players. Comparative analyzes were also carried out in groups of players with better and worse averages of physical efficiency, showing that players with better averages tend to present characteristics of training more similar to those of the game, while players with worse averages present numbers in training lower than those of the games. . The main characteristics that influence the results of the games were also studied, one of them being the difference in performance between the first and second half, considering variables related to distance. The XGBoost classifier achieved good results in predicting this difference in performance, in addition to showing how and which variables influence this difference. The work also presents systematic reviews of artificial intelligence in team and individual sports, bringing relevant information and metrics on the subject.

Keywords: Artificial Intelligence, Sports Science, Data Science, Performance, Injury, Training.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A ciência de dados emergiu como uma área estratégica para explorar o conhecimento em ciência do esporte, com o objetivo de preencher algumas lacunas deixadas pelos métodos estatísticos tradicionais. Como uma área de conhecimento híbrido, a ciência de dados é mais do que a combinação de estatística e ciência da computação, pois exige como integrar técnicas estatísticas e computacionais em uma estrutura maior, problema por problema, e abordar questões específicas da disciplina (BLEI; SMYTH, 2017). Uma visão holística da ciência de dados exige uma compreensão do contexto dos dados, apreciando as responsabilidades envolvidas no uso de dados públicos e privados e comunicando claramente o que um conjunto de dados pode e não pode dizer sobre o mundo real (BLEI; SMYTH, 2017), no nosso caso, no mundo dos esportes. Nos modelos de aprendizagem, os algoritmos podem ser ajustados e otimizados para produzir melhores resultados para apoiar as decisões e fornecer conhecimentos aplicados aos atletas e profissionais do esporte.

Atrelada à ciência de dados, a inteligência artificial (IA) começa como um campo de pesquisa cujo objetivo era replicar o nível de inteligência humana em um sistema computacional, mas com o passar dos anos recuou para subproblemas especializados como maneiras de representar conhecimento, compreensão da linguagem natural, visão e outras áreas especializadas, como sistemas de verificação e validações (BROOKS, 1991). Atualmente, a inteligência artificial se mostra como importante instrumento social, impactando a vida das pessoas e a economia de países, desde assistentes de celulares à robôs que imitam o comportamento humano (LU et al., 2018).

As técnicas e métodos de inteligência artificial atraíram considerável atenção na indústria do esporte e na sociedade como um todo, devido ao grande número de dados e à iminente necessidade de transformar esses dados em conhecimento útil e soluções práticas (LAPHAM; BARTLETT, 1995a; PODHURSKYI, 2020; HASSAN; SCHRAPF; TILP, 2017a; RUSSELL; NORVIG, 2016; WITTEN et al., 2016; PASSFIELD; HOPKER, 2017; REIN; MEMMERT, 2016). O uso efetivo de dados em algumas áreas ainda está em desenvolvimento, como é o caso do esporte. Como na maioria das outras áreas da sociedade, um volume crescente de dados é coletado em todos os tipos de esportes e a análise automatizada de dados se tornou um campo importante e de rápido desenvolvimento (PHATAK; MEMMERT, 2021; CLAUDINO et al., 2021; CLAUDINO; CAPANEMA; SANTIAGO, 2021). Análises cuidadosas desses grandes conjuntos de dados podem aprimorar nosso conhecimento na ciência do esporte e, ao mesmo tempo, auxiliar na tomada de decisão dos profissionais que trabalham na otimização das estratégias de treinamento e competição (PASSFIELD; HOPKER, 2017; REIN; MEMMERT, 2016).

O uso de sistemas de localização em tempo real e o monitoramento de atletas 24 horas por dia, 7 dias por semana, são uma tendência crescente, tanto em esportes coletivos quanto individuais. Isso é feito por meios práticos, como o emprego de uma estrutura integrada de sensores vestíveis, aplicativos para *smartphones* e exames laboratoriais e coleta de dados durantes treinos e jogos (PHATAK; MEMMERT, 2021; DüKING et al., 2018; CLAUDINO; CAPANEMA; SANTIAGO, 2021).

Esta tese busca responder algumas questões utilizando a inteligência artificial aplicada ao futebol, com estudos de caso em times profissionais do futebol brasileiro. Seria possível determinar fatores ou características dos jogadores que influenciam na performance do atleta em jogo? A performance dos atletas tem ligação direta com o resultado da equipe? Busca-se, também, um entendimento mais aprofundado sobre variáveis relevantes e características que podem ajudar a tomada de decisão por parte de técnicos, atletas e comissão técnica. Além disso, a tese visa apresentar, através de revisões sistemáticas, características e análises dos estudos acerca da inteligência artificial aplicada ao esporte, abrindo perspectivas para novos estudos.

1.1 Justificativa

A inteligência artificial é uma área emergente no setor e nas formas acadêmicas, levando a mais tomadas de decisões baseadas em evidências em várias esferas da vida, incluindo serviços de redes sociais, serviços de *streaming*, assistência médica, fabricação, educação, modelagem financeira, policiamento e marketing (ISRANI; VERGHESE, 2019; SHORTLIFFE; SEPÚLVEDA, 2018; BAKER, 2015; JORDAN; MITCHELL, 2015). Além disso, vincular ciência e tecnologia ao crescimento da eficiência esportiva foi apontado como um caminho promissor. Para que isso ocorra, o trabalho voltado para a inovação, introdução e melhoria de processos realizados pelos departamentos de pesquisa e desenvolvimento (P&D) nas maiores empresas de tecnologia do mundo, também foi sugerido no campo do esporte. Devido ao seu ambiente veloz, os profissionais do esporte combinam dados (por exemplo, físico, técnico e tático) com a opinião de especialistas para informar decisões sobre os jogadores (MCCALL et al., 2016).

O estudo e aplicação de técnicas no esporte também teve uma rápida e crescente evolução nas últimas décadas em ligas de todo o mundo. As recentes implementações do *fair-play* financeiro e as apostas são áreas que terão grande crescimento nos estudos nos próximos anos. As técnicas desenvolvidas estão sendo usadas para apoiar a decisão em todos os aspectos do esporte profissional, incluindo:

- Estratégias, táticas e análises;
- Aquisição, avaliação e gastos com equipes;
- Regimes de treinamento e foco;

- Previsão e prevenção de lesões;
- Gerenciamento e previsão de desempenho;
- Resultados de jogos e previsão da tabela de classificação;
- Design e programação de torneios;
- Cálculo de probabilidades de apostas (BREFELD et al., 2019).

Uma questão importante na indústria do esporte é a possibilidade de melhorar o desempenho dos atletas. Historicamente, a capacidade da equipe de prescrever treinamentos para alcançar o desempenho atlético ideal pode ser atribuída a muitos anos de experiência pessoal. No entanto, são necessárias abordagens modernas com o objetivo de adotar métodos científicos para o desenvolvimento eficaz de programas de treinamento ideais (BORRESEN; LAMBERT, 2009). A aplicação de abordagens estatísticas contemporâneas da ciência de dados abre uma perspectiva interessante para lidar com os modelos de desempenho (LÓPEZ-VALENCIANO et al., 2018; NOVATCHKOV; BACA, 2013).

1.2 Objetivos

O objetivo principal deste trabalho é aplicar métodos de inteligência artificial para gerar uma base de conhecimento para auxiliar atletas e comissões técnicas nas tomadas de decisão além de apresentar variáveis relevantes e como elas influenciam na ocorrência de lesões e na performance dos atletas.

1.2.1 Objetivos específicos

O trabalho também visa alcançar os seguintes objetivos específicos, usando como base a ciência de dados e métodos de inteligência artificial:

- Classificar atletas com maior risco para ocorrência de lesões;
- Identificar possíveis fatores que contribuem para o melhor desempenho do atleta;
- Identificar atletas que terão melhor performance baseado em dados históricos;
- Compreender os aspectos mais relevantes para resultados positivos nos jogos;
- Gerar uma base de conhecimento para auxiliar atletas e comissões técnicas nas tomadas de decisões.

1.3 Contribuições

O trabalho apresenta estudos sobre as características de grupos de jogadores que se lesionaram na temporada, apresentando considerações que podem auxiliar na preparação de jogadores para evitar lesões. Também apresenta um comparativo entre atletas com altas e baixas médias de eficiência física, abrindo perspectivas sobre treinos com cargas mais próximas às executadas em jogos para melhoria do desempenho dos atletas. Os

estudos também fornecem um entendimento maior da influência de diversos fatores no desempenho de atletas de alto nível e também sobre resultados dos jogos, apresentando um modelo de apoio que irá reportar informações e previsões sobre o desempenho destes atletas entre o primeiro e segundo tempo, apoiando tomadas de decisão sobre cargas de treinamento, escalações de acordo com características do atleta, campeonato e rivais, além de decisões sobre substituições de atletas no intervalo do jogo. O trabalho portanto visa apresentar variáveis dentro e fora do campo que podem influenciar na preparação do atleta e no seu rendimento em jogo. Além disso, apresenta uma visão geral sobre os trabalhos sobre inteligência artificial aplicada ao esporte nos últimos anos.

1.4 Organização do trabalho

Este trabalho está organizado da seguinte forma: no Capítulo 2 são apresentados trabalhos sobre inteligência artificial no esporte e estudos que estão relacionados com o tema, além de duas revisões sistemáticas da literatura acerca da inteligência artificial no esporte. No Capítulo 3 são apresentados os conceitos utilizados nos demais capítulos deste trabalho. No Capítulo 4 é apresentada a metodologia utilizada para a realização das análises em dois estudos de caso apresentados neste trabalho. O Capítulo 5 apresenta as análises realizadas nos dados e os resultados obtidos com as respectivas discussões. Por fim, no Capítulo 6 é feita a conclusão do trabalho realizado e perspectivas de novos estudos.

2 Trabalhos relacionados

As pesquisas que tratam o uso de IA no esporte tiveram um grande crescimento nos últimos anos e hoje apoiam nas tomadas de decisão e explicam vários fatores relacionados aos esportes. Existem diversos trabalhos que aplicam ciência de dados e IA no esporte, com os mais diversos propósitos e métodos. Neste capítulo serão apresentados trabalhos relacionados ao tema da tese e ao final duas revisões sistemáticas sobre uso de IA no esporte em equipe e individual.

2.1 IA no esporte

O estudo de Lapham e Bartlett (1995b) foi um dos primeiros estudos utilizando IA na análise do desempenho no esporte. Em 1995, eles demonstraram que o crescente uso de computadores no processo de tomada de decisão, através do uso de técnicas de IA, seria um potencial norteador da disciplina e já apontou as redes neurais artificiais em experiências bem sucedidas juntamente com outros benefícios potenciais. Já o estudo de Zelič et al. (1997), publicado há mais de 20 anos, relatou o uso de IA com classificadores por árvore de decisão e classificação bayesiana no diagnóstico de lesões esportivas.

McCall et al. (2016) recomendaram que as equipes de futebol profissionais incluíssem P&D em suas atividades diárias para reduzir lesões e otimizar o desempenho. Assim como o presente trabalho, vários outros estudos já foram realizados abordando aplicações de IA no futebol. Estudos acerca da prevenção de lesões abordaram temas como “carga de treinamento” (ROSSI et al., 2018; JASPERS et al., 2018), “causas de lesões no joelho” (QILIN et al., 2016), “detecção de cardiopatia” (ADETIBA et al., 2017), “padrão de força de reação do solo” (ERTELT; SOLOMONOVS; GRONWALD, 2018), “fatores psicossociais de estresse” (PENSGAARD et al., 2018) e “lesões musculares” (LÓPEZ-VALENCIANO et al., 2018). Outros foram relacionados ao desempenho, tratando assuntos como “análise técnica e tática” (ABDULLAH et al., 2016; LINK; HOERNIG, 2017; COLÁS et al., 2015; CARPITA et al., 2015; HOCH et al., 2017; WANG, 2014), “comparecimento aos jogos” (STRNAD; NERAT; KOHEK, 2017) e “psicologia para trabalho em equipe colaborativo” (FUSTER-PARRA et al., 2016).

Estudos na literatura mostram que redes neurais artificiais são muito utilizadas para analisar e diminuir o risco de lesão dos atletas. Ge (2017) e Qilin et al. (2016) realizaram estudos que verificam os principais fatores para lesões no joelho e fazem sugestões para evitar estas lesões no futebol e vôlei de praia. Kautz et al. (2017) classificaram os atletas por vídeos e verificaram os mais propensos à lesões. Adetiba et al. (2017) também conseguiram bons resultados, identificando características de atletas propensos à problemas cardíacos.

Já Bartlett et al. (2017) encontraram formas de indicar cargas de treinamento melhores que as tradicionais para diminuir o risco de lesões.

Outros artigos utilizaram classificadores por árvore de decisão para diminuir o risco de lesão. Thornton et al. (2017) realizaram o monitoramento de carga de treinamento, com bons resultados, sugerindo diferenças pessoais baseadas em capacidades fisiológicas e demandas físicas. López-Valenciano et al. (2018) criaram um modelo preditivo para evitar lesões baseado em dados pessoais, psicológicos e neuromusculares.

Goswami et al. (2016) e Wu et al. (2017) realizaram estudos utilizando o método SVM em jogadores de futebol americano para identificar problemas causados pelos impactos no crânio. Whiteside et al. (2016) também utilizaram SVM, mas para identificar fatores significativos para risco de lesões e ajudar a diminuir o número de cirurgias em jogadores de *baseball*.

Outros estudos aplicados ao risco de lesões foram encontrados para o futebol. Jaspers et al. (2018) propõem cargas de treinamento baseadas no esforço e cargas externas, enquanto Pensgaard et al. (2018) analisam o efeito de situações estressantes e clima motivacional na ocorrência de lesões.

Existem diversos estudos utilizando RNA para previsão e análise de performance esportiva no basquete. Kempe, Grunz e Memmert (2015) e Wu (2013) analisam posicionamento de atletas e avaliações de ataque e defesa. Li (2013) propõem um modelo matemático para prever o resultado das partidas através da performance dos atletas. Bianchi, Facchinetti e Zuccolotto (2017) classificam os atletas baseado em diversos fatores para apoiar decisões de treinadores e compra de jogadores.

Performance esportiva utilizando RNA também é bastante utilizada no *handball*. Tilp e Schrapf (2015) e Schrapf, Alsaied e Tilp (2017) fazem análises estratégicas baseadas no posicionamento dos atletas. Hassan et al. (2017) também analisam o desempenho tático dos jogadores e Hassan, Schrapf e Tilp (2017b) preveem a jogada de ataque baseada em jogadas anteriores.

Também utilizaram RNA os estudos de Croft, Lamb e Middlemas (2015), que analisa variáveis significativas para explicar resultados de jogos e Tümer e Koçer (2017) que criou um modelo preditivo para tabelas de campeonato de vôlei. Já os estudos de Abdullah et al. (2016) e Strnad, Nerat e Kohek (2017) fazem previsões utilizando RNA para identificar habilidades técnicas e aspectos comportamentais, ambos no futebol.

Redes Bayesianas foram utilizadas no estudo de Healey (2017), que prevê o desempenho de atletas de *baseball* identificando padrões e nível dos jogadores. Também utilizaram redes bayesianas aplicadas ao futebol os estudos de Fuster-Parra et al. (2016), que analisa perfil colaborativo e trabalho em equipe para ajudar na formação de equipes e Link e Hoernig (2017) que propõe melhorias na performance através de posicionamento e

posse de bola.

A utilização de classificadores por árvore de decisão é bastante utilizada para analisar e prever performance no futebol americano, com estudos relacionados à retomada de bola (BOCK, 2017), previsão de performance baseada em testes fisiológicos (JELINEK et al., 2014) e relações entre performance e resultado dos jogos (ROBERTSON; BACK; BARTLETT, 2016). Fatores que relacionam performance com resultado de jogos também foram encontrados no basquete utilizando classificadores por árvores de decisão (ÇENE, 2018; LEICHT; GOMEZ; WOODS, 2017; LEICHT; GÓMEZ; WOODS, 2017).

Os estudos de Carpita et al. (2015) e Colás et al. (2015) utilizam classificadores por árvore de decisão aplicado ao futebol para determinar fatores que influenciam a vitória de uma equipe e indicadores para melhoria de performance baseada em análise em vídeos.

O método SVM aplicado no esporte para prever resultados de partidas foi utilizado em diversos estudos encontrados na literatura (VALERO, 2016; PAI; CHANGLIAO; LIN, 2017; GU; SAATY; WHITAKER, 2016). Também utilizaram SVM os estudos de Bock (2015) que utiliza dados históricos para prever o lançamento de jogadores de *baseball* e Wang et al. (2018) que utiliza sensores para classificar jogadores de vôlei por nível técnico.

Os estudos de Lu (2013) e Hoch et al. (2017) utilizaram o método *Fuzzy* para analisar desempenho físico e tático, no basquete e futebol, respectivamente. Sankaran (2014) faz uma classificação de jogadores de *cricket* utilizando *K-means* para verificar se existe relação entre a performance do jogador e seu valor de mercado, encontrando diferenças significativas que podem ajudar em contratações.

Outros artigos utilizaram métodos de IA para avaliar ou prever o resultados dos jogos, como os trabalhos de Wang (2014) aplicando *K-means* no futebol, Kolbush e Sokol (2017) aplicando regressão logística no futebol americano e Schulte et al. (2017) aplicando *Markov* no *hockey* no gelo.

Com relação aos estudos abordando a distância percorrida pelos jogadores nos jogos, vários estudos apresentam características que podem influenciar ou possuem relação com a distância percorrida. Bangsbo, Nørregaard e Thorsø (1991) mostram a relação do lactato sanguíneo com as distâncias percorridas por cada jogador, que também é relacionada à posição do atleta em campo. Os estudos de Mohr, Krustrup e Bangsbo (2003) apresentam, além de diferenças nos testes YoYo, que avaliam capacidade aeróbica, que a distância em alta velocidade tem alta relação com a posição do atleta, além de apresentar momentos onde a fadiga do atleta é mais comum. Outros trabalho também associam a posição do atleta ao desempenho nas distâncias percorridas no jogo, em diferentes níveis de velocidade, diferenciando, inclusive, as performances entre primeiro e segundo tempo (REILLY, 1976; RIENZI et al., 2000; TUMILTY, 1993; SALVO et al., 2007; PEñAS et al., 2010; DALEN et al., 2019).

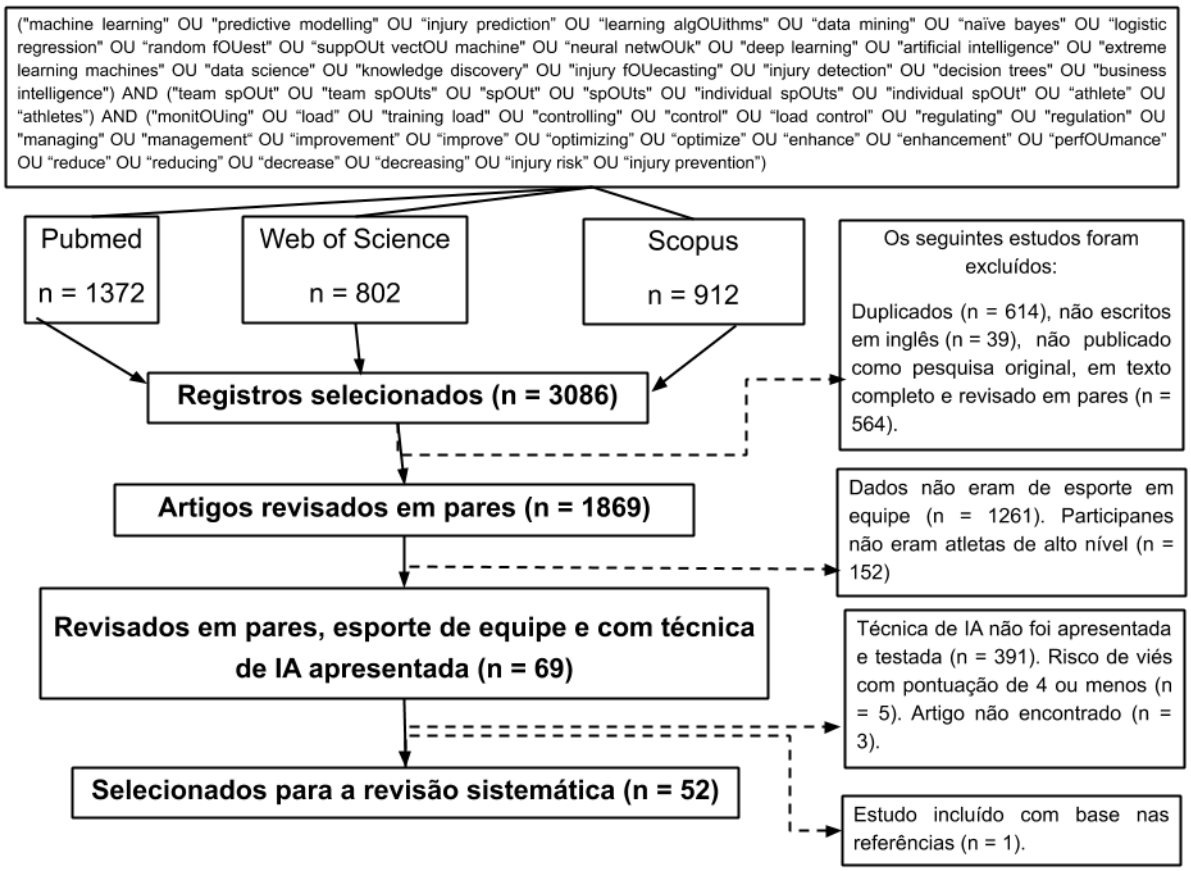
Esta tese também aplica técnicas para identificar fatores que contribuem com lesões em atletas, utilizando técnicas diferentes para classificar os jogadores em grupos baseados em variáveis correlacionadas e posteriormente observar a presença de jogadores lesionados nestes grupos. O maior número de análise foi realizado sobre a performance esportiva, diferenciando dos trabalhos relacionados apresentados anteriormente por identificar os fatores de performance que influenciam nos resultados da equipe e por propor prever o desempenho do atleta no segundo tempo, no que diz respeito as distâncias percorridas, baseado em dados históricos.

2.2 Revisão sistemática da literatura

2.2.1 Esportes de equipe

Claudino et al. (2019) realizaram uma revisão sistemática da literatura entre os anos de 2013 e 2018 em 3 repositórios de dados, que retornaram 3086 estudos. Após removerem os duplicados, que não estavam em inglês e que não foram revisados em pares, obtiveram um total de 1869 artigos, sendo 456 relacionados à esportes coletivos e com jogadores profissionais ou de alto rendimento. Destes, 58 estudos apresentavam a técnica de inteligência artificial descrita e testada além de passar por um filtro que analisava a qualidade do estudo conforme apresentado na Figura 1. A revisão apresentou que nestes 58 artigos, a amostra era formada majoritariamente por atletas do sexo masculino (97%) e os esportes onde houveram mais estudos foram futebol (26%), basquete (22%) e *handball* (10%).

Figura 1 – Revisão sistemática da literatura para artigos relacionados com inteligência artificial no esporte



Fonte: Adaptação de Claudino et al. (2019)

O estudo de Claudino et al. (2019) dividou os artigos em dois grupos, os que tratavam de performance esportiva (74%) e os que tratavam de risco de lesões (26%). Considerando especificamente os artigos relacionados à performance esportiva, e que é apresentado na Figura 2, o método de IA mais utilizado foi redes neurais artificiais (RNA) (34%), seguida por árvore de decisão (23%) e em terceiro lugar processos de markov e máquina de vetores de suporte (SVM) (12% cada).

Figura 2 – Estudos abordando Performance esportiva

Fonte: Adaptação de Claudino et al. (2019)

Considerando artigos que trataram de risco de lesão, apresentado na Figura 3, o método de IA mais utilizado foi redes neurais artificias (47%), seguida por árvore de decisão (21%) e máquina de vetores de suporte (15%). Em ambas abordagens, performance esportiva e risco de lesão, é possível ver que redes neurais e árvore de decisão são os métodos largamente utilizados, e que os esportes mais estudados, futebol e basquete, não estão concentrados em métodos específicos, apresentando variações de métodos e abordagens.

Figura 3 – Estudos abordando risco de lesão no esporte

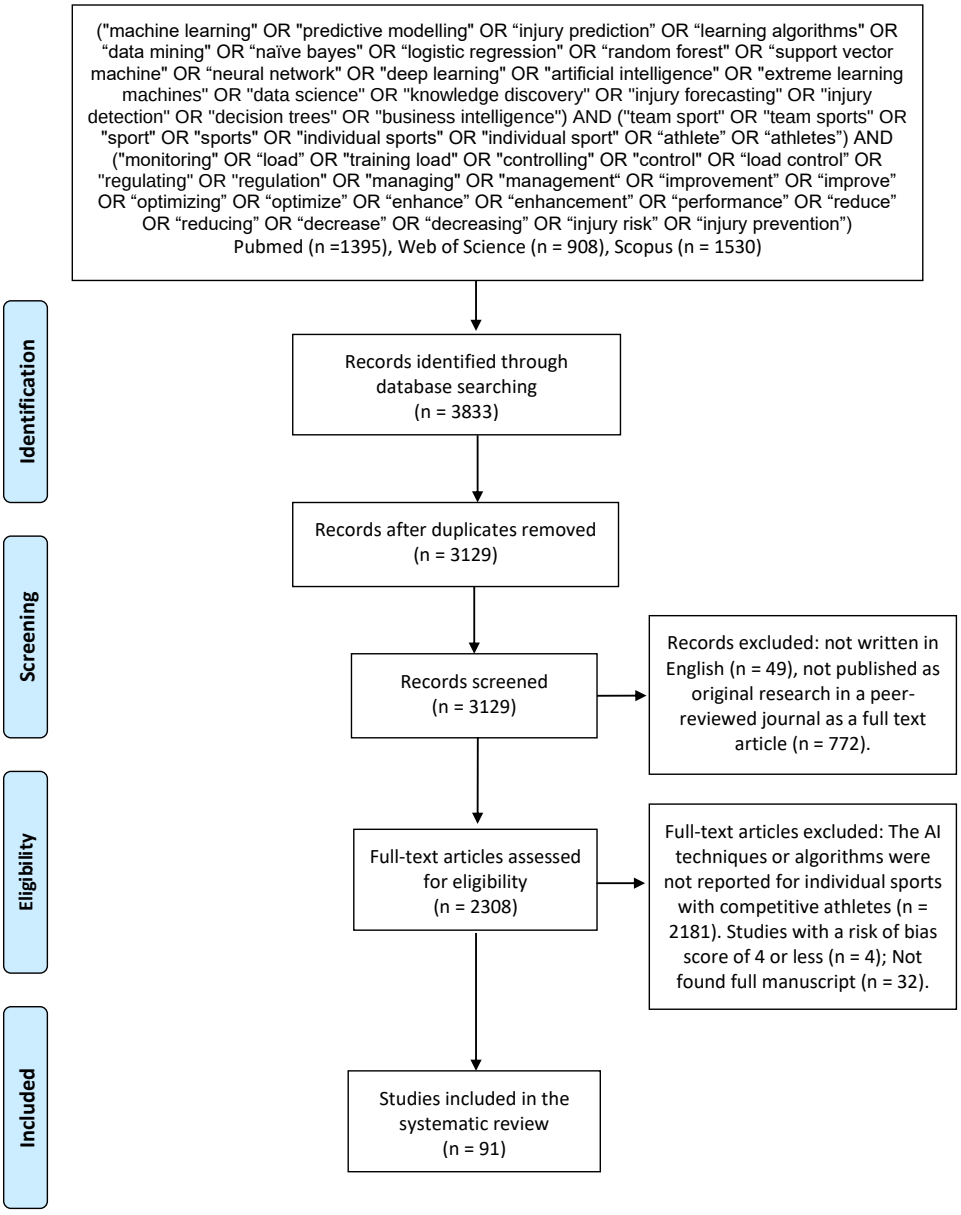
Fonte: Adaptação de Claudino et al. (2019)

As análises realizadas nesta revisão sistemática mostraram que as técnicas ou métodos de IA para prever o risco de lesões e desempenho esportivo mais usados em esportes coletivos eram redes neurais artificiais, classificador por árvore de decisão. Os esportes coletivos com a maioria das aplicações de IA eram futebol, basquete, handebol e vôlei. O estado atual de desenvolvimento na área propõe um futuro promissor no que diz respeito ao uso de IA em esportes de equipe.

2.2.2 Esportes individuais

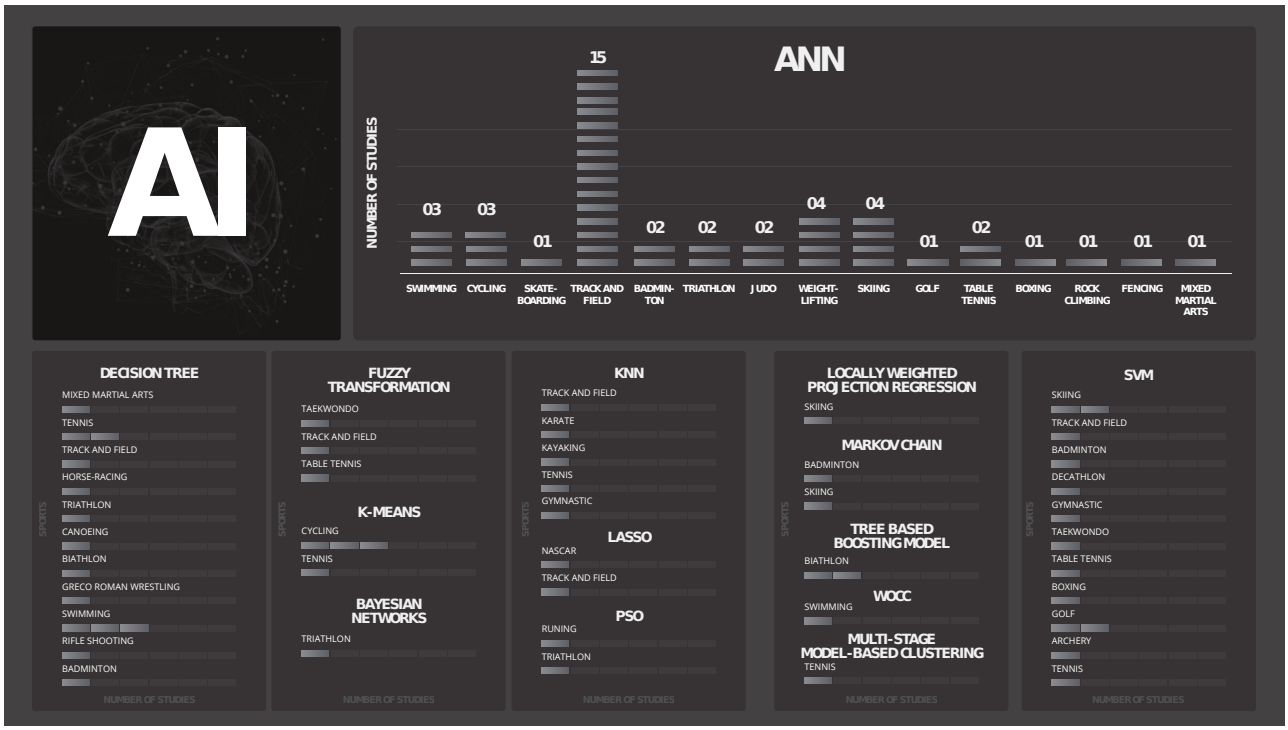
Uma revisão sistemática da literatura, desta vez abordando esportes individuais, foi realizada considerando trabalhos entre os anos de 2013 e 2021 em 3 repositórios de dados, que retornaram 3833 estudos. Após removerem os duplicados, que não estavam em inglês, que não foram revisados em pares e com jogadores profissionais ou de alto rendimento, foram obtidos 91 estudos que atendiam aos critérios da pesquisa, com técnica de inteligência artificial descrita e testada além de passar por um filtro que analisava a qualidade do estudo conforme apresentado na Figura 4. A revisão apresentou que nestes 91 artigos, a amostra era formada majoritariamente por atletas do sexo masculino (90%) e os esportes onde houveram mais estudos foram atletismo (22%), esqui (9%) e natação (8%). Assim como no trabalho de esportes coletivos, os métodos mais utilizados foram redes neurais artificiais e classificador por árvore de decisão, como mostra a Figura 5.

Figura 4 – Revisão sistemática da literatura para artigos relacionados com inteligência artificial no esporte



Fonte: Autores

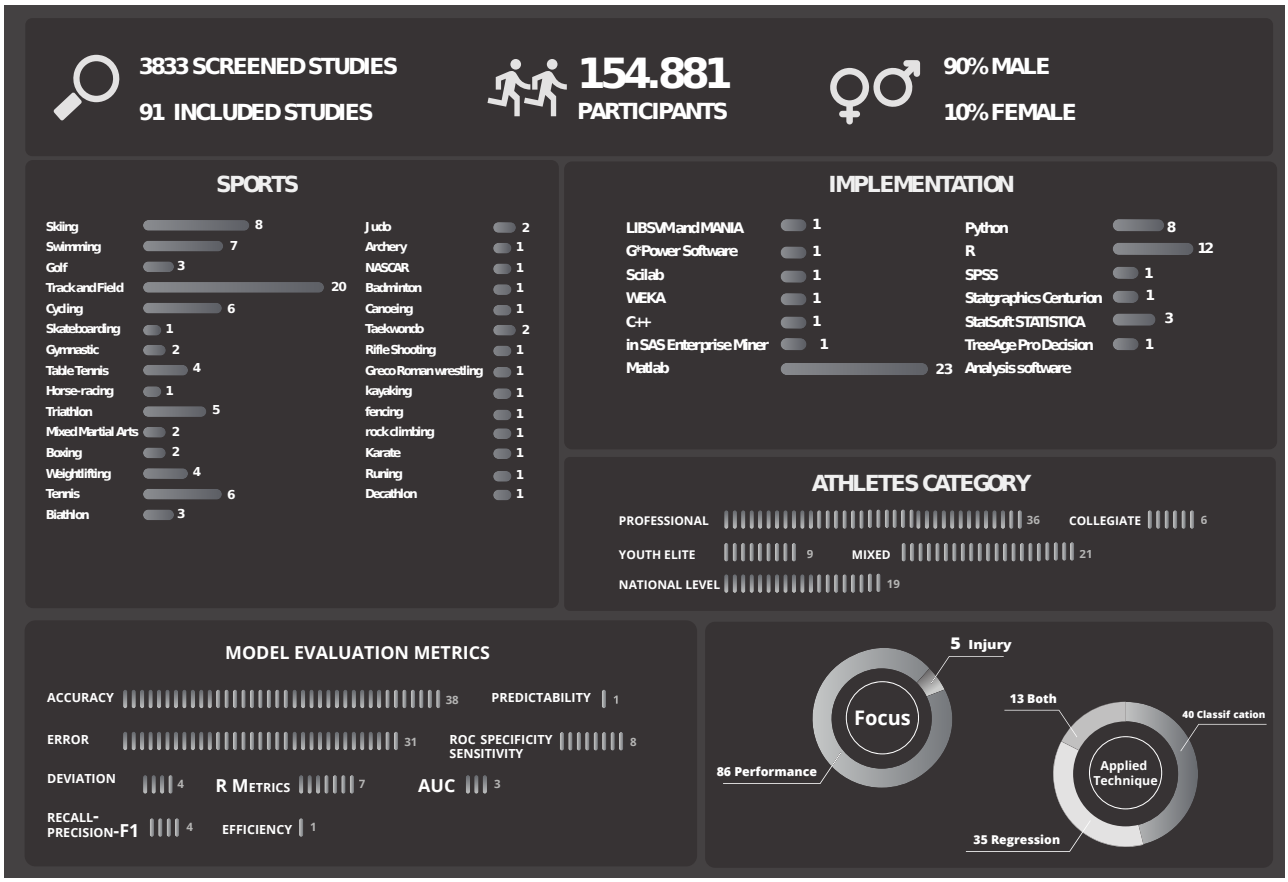
Figura 5 – Métodos e esportes utilizando IA em esportes individuais



Fonte: Autores

Outra característica em comum com o estudo de esportes coletivos foi a abordagem dos estudos, onde 94% tratava de performance e apenas 6% risco de lesão. As ferramentas mais utilizadas nos estudos não foram obvias, visto que Python ficou em terceiro lugar, atrás de Matlab e R. Outra característica que chamou atenção nos estudos foram as métricas de avaliação do modelo, onde vários estudos sequer apresentavam ou apenas apresentavam acurácia, como mostra a Figura 6.

Figura 6 – Características dos estudos utilizando IA em esportes individuais



Fonte: Autores

As análises deste estudo mostraram, além das características descritas anteriormente, a necessidade de melhores métricas de avaliação dos modelos, e possíveis razões para a escolha por porte dos autores por determinados esportes e algoritmos. Mostrou a necessidade de mais estudos envolvendo atletas do sexo feminino e em outros esportes mais populares. Além destas análises, o estudo também apresentou os eventos abordando IA no esporte pelo mundo, sendo que apenas um acontece no Brasil, como mostra a Figura 7.

Figura 7 – Eventos de IA em esportes no mundo.

Fonte: Autores

Neste capítulo foram apresentados vários trabalhos e revisões acerca da aplicação de inteligência artificial no esporte. No Capítulo 3 serão apresentados conceitos e metodologias aplicados neste trabalho.

3 Fundamentação teórica

Neste capítulo será apresentado uma descrição formal dos conceitos de inteligência artificial e uma descrição sintética dos métodos de IA utilizados nos trabalhos apresentados no Capítulo 5 e também nos propostos e aplicados no Capítulo 4.

A IA é um campo de pesquisa relativamente recente, com trabalhos relacionados à partir da década de 50. Atualmente possui aplicações nas mais diversas áreas de estudo como matemática, jogos, veículos automotivos, medicina, dentre outros, sendo considerada hoje de grande relevância e um campo universal (RUSSELL; NORVIG, 2016).

Não existe uma definição única para o termo IA, mas pode-se dizer que é um campo de pesquisa que faz com que máquinas simulem o pensamento humano. Na Tabela 1 são apresentadas definições de IA segundo quatro grupos de inteligência.

3.1 Algoritmos de inteligência artificial

Atualmente existem vários métodos ou algoritmos de inteligência artificial que podem ser aplicados para os mais diversos propósitos. Geralmente estes métodos são aplicados para problemas de classificação ou regressão, que estão entre as técnicas mais úteis na estatística moderna aplicada. Estas técnicas, ou melhor, classes de técnicas, são usadas nas mais diversas áreas e aplicações, seja em química, medicina, ciências sociais e várias outras (NæS; MEVIK, 2001).

Classificar uma observação em uma das várias populações é conhecido como análise discriminante ou classificação. Relacionar variáveis qualitativas a outras variáveis

Tabela 1 – Definições de Inteligência Artificial

Pensando como um humano	**Pensando racionalmente**
Um esforço para fazer os computadores pensarem; máquinas com mentes, no sentido total e literal. Atividades associadas ao pensamento humanos; atividades como tomada de decisões, resolução de problemas e aprendizado .	Estudo das faculdades mentais pelo uso de modelos computacionais. Estudo das computações que possibilitam perceber, raciocinar e agir.
Agindo como seres humanos	**Agindo racionalmente**
Criar máquinas que executam ações que exigem inteligência humana. Computador realizar tarefas que são bem executadas por humanos.	Estudo de agentes inteligentes. Desempenho inteligente de artefatos

Fonte: Adaptação de Russell e Norvig (2016)

através de uma forma funcional logística é conhecido como regressão logística (PRESS; WILSON, 1978). O objetivo da regressão logística é encontrar o modelo mais adequado para descrever a relação entre o resultado (variável dependente ou resposta) e um conjunto de variáveis independentes (preditoras ou explicativas) (PERME; BLAS; TURK, 2004). A análise discriminante linear pode ser usada para determinar qual variável discrimina entre duas ou mais classes e derivar um modelo de classificação para prever a associação ao grupo de novas observações (WORTH; CRONIN, 2003).

Seja para classificação ou regressão, existem diversos métodos ou algoritmos clássicos da literatura que podem ser utilizados. Os métodos apresentados neste trabalho estão descritos na Tabela 2.

Tabela 2 – Descrição dos métodos de Inteligência Artificial

IA	**Descrição**
Árvore de decisão	O classificador por árvore de decisão é uma ferramenta de suporte à decisão que usa um gráfico ou modelo de decisões em forma de árvore e suas possíveis consequências, incluindo resultados de eventos aleatórios, custos de recursos e utilidade. É uma maneira de exibir um algoritmo que contém apenas instruções de controle condicionais. As árvores de decisão são comumente usadas na pesquisa operacional, especificamente na análise de decisão, para ajudar a identificar uma estratégia com maior probabilidade de atingir uma meta, mas também são uma ferramenta popular no aprendizado de máquina. O aprendizado da árvore de decisão usa uma árvore de decisão (como modelo preditivo) para passar de observações sobre um item (representado nos ramos) a conclusões sobre o valor alvo do item (representado nas folhas). É uma das abordagens de modelagem preditiva usadas em estatística, mineração de dados e aprendizado de máquina. Modelos de árvore nos quais a variável de destino pode assumir um conjunto discreto de valores são chamados de árvores de classificação; nessas estruturas de árvore, folhas representam rótulos de classe e ramos representam conjunções de recursos que levam a esses rótulos de classe.

Continua na próxima página

Tabela 2 – continuação da página anterior

IA	Descrição
K-means	Agrupamento *K-Means* é um tipo de aprendizado não supervisionado, usado quando há dados não rotulados (ou seja, dados sem categorias ou grupos definidos). O objetivo desse algoritmo é encontrar grupos nos dados, com o número de grupos representados pela variável K. O algoritmo trabalha iterativamente para atribuir cada ponto de dados a um dos grupos K com base nos recursos fornecidos. Os pontos de dados são agrupados com base na similaridade de recursos. Os resultados do algoritmo *K-Means Clustering* são: os centróides dos *clusters* K, que podem ser usados para rotular novos dados; rótulos para os dados de treinamento (cada ponto de dados é atribuído a um único *cluster*). Em vez de definir grupos antes de analisar os dados, o agrupamento permite encontrar e analisar os grupos que se formaram organicamente. Cada centróide de um *cluster* é uma coleção de valores de recursos que definem os grupos resultantes. Examinar os pesos dos recursos do centróide pode ser usado para interpretar qualitativamente que tipo de grupo cada *cluster* representa.
Tree based boosting model	Um modelo baseado em árvore com reforço, do inglês *Tree based boosting model* (XGB), tem se destacado em competições recentes de aprendizado de máquina. Ele é capaz de modelar padrões complexos usando uma combinação de árvores de decisão construídas sequencialmente.

Fonte: Adaptação de Claudino et al. (2019)

Para validação dos modelos utilizados utilizados e avaliação da performance foram utilizados duas técnicas, k-*fold cross-validation e holdout*. A ideia básica do k-fold cross-validation é dividir os dados disponíveis em k partes iguais, em que uma dessas partes é usada como conjunto de teste e as outras k-1 partes são usadas como conjunto de treinamento. O modelo é treinado k vezes, cada vez usando uma das partes como conjunto de teste e as outras k-1 partes como conjunto de treinamento. Após cada treinamento, a performance do modelo é medida usando as amostras do conjunto de teste e os resultados são combinados para obter uma avaliação geral da performance do modelo. Em outras palavras, a técnica de k-fold cross-validation calcula a média e a variância das métricas de performance (como precisão, recall, F1-score, etc.) em k iterações (JAMES et al., 2013).

O teste holdout é uma técnica comum de validação de modelo em aprendizado de máquina, em que o conjunto de dados é dividido em dois subconjuntos mutuamente exclusivos: um subconjunto de treinamento e um subconjunto de teste. O subconjunto de treinamento é utilizado para ajustar o modelo e o subconjunto de teste é usado para avaliar a capacidade do modelo de generalização em dados não vistos. Essa avaliação é realizada calculando-se as métricas de desempenho, como acurácia, precisão, recall e F1-score, para os dados de teste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Os métodos de inteligência artificial podem ser aplicados com o apoio de ferramentas de implementação, que devem ser escolhidas dependendo da natureza do problema, habilidade do pesquisador, disponibilidade, bibliotecas, etc. Neste trabalho, foi utilizado o Python, que é uma excelente linguagem de programação que, com ferramentas básicas adicionais, se transforma em uma linguagem de alto nível adequada para código científico e de engenharia, que é rápida o suficiente para ser processada, mas também flexível o suficiente para ser acelerada com extensões adicionais (OLIPHANT, 2007). Nas últimas décadas, o Python se tornou uma das linguagens mais populares entre os pesquisadores nos mais diversos campos da ciência de dados, com suas diversas estruturas e bibliotecas, geralmente com uma linguagem clara e simples (MCKINNEY, 2012; DEMŠAR et al., 2013; PEDREGOSA et al., 2011).

3.2 CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) (WIRTH; HIPP, 2000) é um dos processos atualmente utilizados para definir e guiar as iterações no desenvolvimento e aplicação de diferentes abordagens de Ciência de Dados no mercado (SCHRÖER; KRUSE; GÓMEZ, 2021).

Basicamente, o CRISP-DM é um processo iterativo e cíclico em que o desenvolvimento do projeto pode ser enquadrado em 6 fases: Entendimento do problema, Entendimento dos Dados, Preparação de Dados, Modelagem, Avaliação e Implantação.

A Figura 8 exemplifica as fases do processo assim como as diferentes interações, e caminhos que são possíveis de se percorrer durante a aplicação do método.

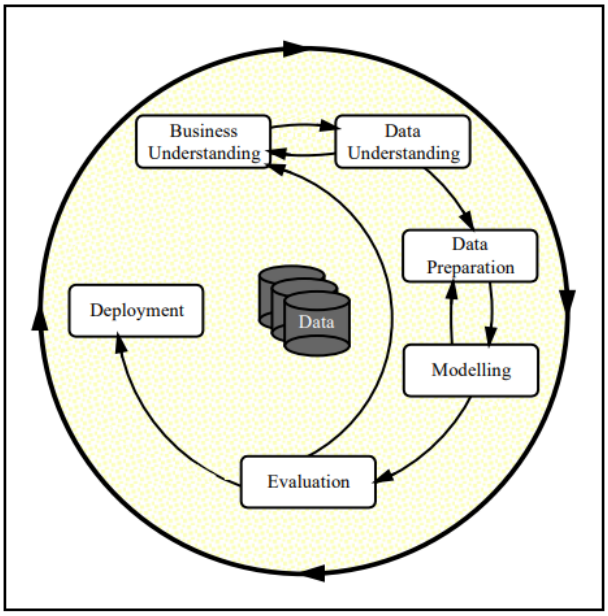
3.2.1 Processo

Inspirado no CRISP-DM a Figura 9 apresenta um diagrama do processo aplicado a esse trabalho.

No diagrama é apresentada as seguintes fases:

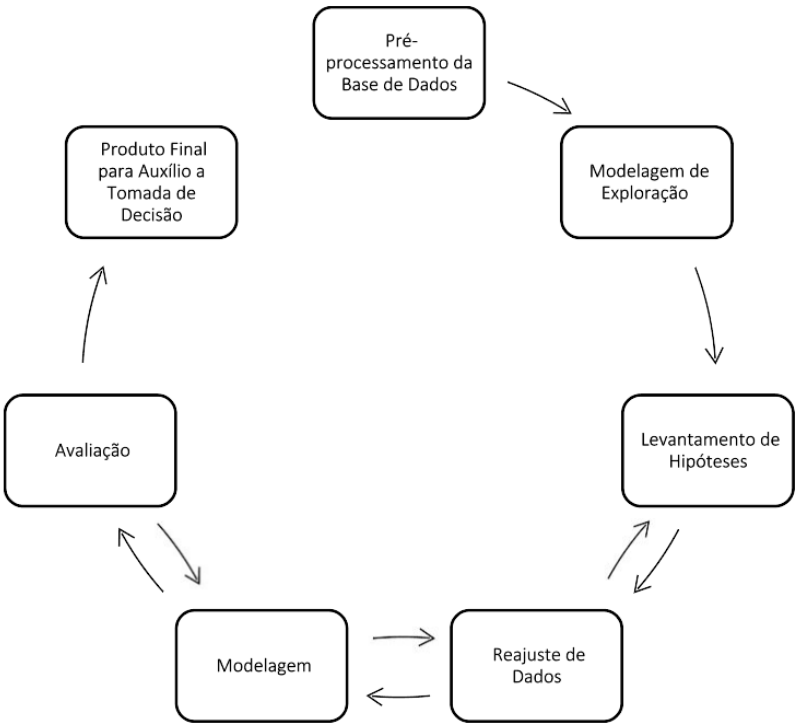
- Pré-processamento de dados: Processamento inicial da base obtida, com diferentes tipos de ajustes como limpeza de dados, padronização e aplicação de técnicas iniciais

Figura 8 – Fases CRISP-DM



Fonte: Wirth e Hipp (2000)

Figura 9 – Processo Metodológico



Fonte: Wirth e Hipp (2000)

para tratamento de dados.

- Modelagem e Exploração: Modelagem inicial utilizando técnicas mais tradicionais para exploração do comportamento dos dados e obtenção de *baseline*.

- Levantamento de Hipóteses: Com base na etapa anterior e nos *baselines* obtidos, hipóteses acerca do comportamento dos dados e de como explicá-los são levantadas. Pode-se voltar nessa etapa várias vezes, uma vez que a cada nova avaliação da modelagem, novos comportamentos podem surgir.
- Reajuste de Dados: Conforme as hipóteses levantadas, os dados são reajustados buscando obter cada vez mais resultado melhores e mais significativos. Essa também é uma etapa que é acionada constantemente.
- Modelagem: Aplicação de técnicas cada vez mais otimizadas e mais ajustadas àquilo que se almeja obter de resultados. O processo de modelagem acontece várias vezes dependendo dos objetivos finais.
- Avaliação: A seguir ao processo de modelagem é sempre necessário avaliar os resultados obtidos, se são condizentes e pertinentes aos objetivos. Em caso de resultados ainda não suficientes ou questões em aberto, pode se sempre voltar as etapas anteriores.
- Produto Final: Por fim, após todo o processo, busca-se a construção de um produto final, como um modelo implantado para o auxílio a tomada de decisão.

No Capítulo 4 será apresentada a metodologia utilizada para realização de dois estudos de caso utilizando dados de dois times de futebol da primeira divisão do campeonato brasileiro.

4 Metodologia

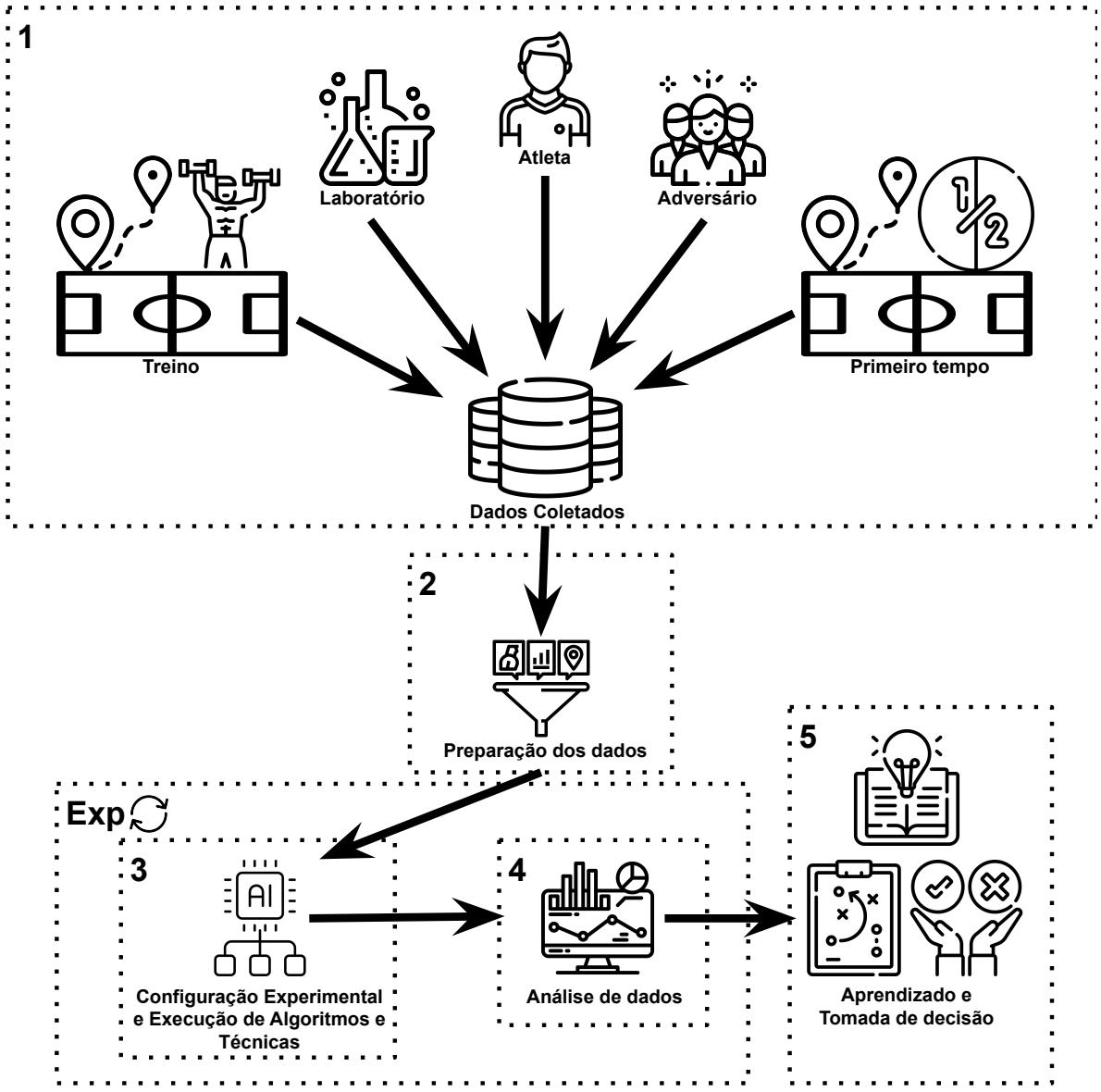
Neste capítulo serão apresentadas as etapas seguidas para realização do trabalho proposto com suas respectivas metodologias, ferramentas e processos utilizados. A Figura 10 apresenta estas etapas, que consistem na coleta dos dados de diversas fontes, um filtro onde são realizados vários tratamentos nos dados. Na sequência existem as fases de execução dos algoritmos e análise dos dados, que juntas fazem parte de uma etapa cíclica de experimentação para melhoria do modelo. Por fim temos a fase de tomada de decisão baseada nas informações geradas.

A metodologia apresentada na Figura 10 é inspirada no CRISP-DM e foi aplicada nos dois estudos de caso que foram realizados neste trabalho, que seguiram etapas semelhantes, mas com algumas características específicas. O primeiro estudo de caso utilizou dados de um grande clube da série A do futebol brasileiro, analisando dados de laboratório, treinos e jogos. Neste primeiro estudo de caso, foram realizadas análises sobre lesões e eficiência física dos atletas. No segundo estudo de caso, que também analisou dados de um grande clube da série A do futebol brasileiro, foram utilizados dados de treino, jogo, dados de atletas e adversários para entender a performance dos atletas e as influências nos resultados das partidas. Os estudos de caso serão detalhados no Capítulo 5. As etapas da metodologia proposta serão descritas nas próximas seções.

4.1 Etapa 1: coleta de dados

Nesta etapa os dados são coletados de diversas fontes, em diversos formatos e em períodos distintos. Dados dos atletas são informações referentes à idade, peso, tempo sem realizar partidas e outras informações pessoais. Os dados do adversário são referentes à qualidade do time, posse de bola, local do jogo, dentre outros. Os dados de laboratório são coletados em intervalos distintos e apresentam dados acerca da condição física do jogador, frequência cardíaca e outros dados coletados em exames. Os dados de treino são coletados nos dias anteriores aos jogos e apresentam duração, período do dia, distâncias e cargas dos treinamentos. Os dados de jogos são fornecidos com informações sobre resultado, tempo de posse de bola, local da partida, distâncias percorridas em diferentes velocidades, dentre outros. Os dados dos jogos são separados por primeiro tempo, segundo tempo e jogo total. Na metodologia é destacada a coleta do primeiro tempo para representar que estas informações são utilizadas para definir e prever as informações da segunda etapa da partida, mas na prática os dados do segundo tempo também são considerados para testar e validar as análises.

Figura 10 – Metodologia aplicada



Fonte: Autores

4.2 Etapa 2: preparação dos dados

Nesta etapa são realizados vários tratamentos nos dados, limpeza de dados ruins ou incorretos, geração de novos campos, padronização dos dados e inclusão de dados relevantes que eventualmente não estão nas bases fornecidas. Utilizando a linguagem Python, os dados são mesclados em *dataframes* onde recebem os tratamentos. Esta é uma fase extremamente importante pois inconsistências nos dados podem levar a análises incorretas, que certamente prejudicarão a qualidade dos estudos.

4.3 Etapa 3: configuração experimental e execução de algoritmos e técnicas

A fase de execução é caracterizada pela elaboração e execução de algoritmos e técnicas específicos para ciência de dados. Primeiramente são observadas correlações entre variáveis e definição do melhor conjunto de dados a ser utilizado. De acordo com a necessidade e os objetivos do fisiologista e demais interessados, são levantados o tipos de análise, se será um agrupamento, se será um modelo preditivo, se será verificada a influência variáveis em objetivos definidos, dentre outros. Nesta fase também são gerados os resultados da execução dos algoritmos e gráficos para auxiliar, na próxima etapa, a análise dos dados.

4.4 Etapa 4: análise de dados

Com os resultados obtidos na fase de execução dos algoritmos, é necessário analisar a qualidade dos resultados e se os dados obtidos satisfazem as necessidades levantadas anteriormente. Existem diversas métricas a serem utilizadas, que vão depender do método utilizado e características da observação. Neste momento, dependendo dos resultados obtidos, novas demandas são levantadas, outras descartadas, e com auxílio dos especialistas no negócio, o processo volta para a etapa anterior para novas execuções.

4.5 Etapa Exp: experimentos

A etapa de experimentos é, na verdade, um agrupamento das etapas de execução de algoritmos e análise de dados. Aqui o importante é representar que estas duas etapas são cíclicas, e são executadas enquanto houverem resultados insatisfatórios ou novas demandas forem levantadas.

4.6 Etapa 5: aprendizado e tomada de decisão

Com os dados consistentes e resultados satisfatórios, gerados nas etapas anteriores, é o momento de interpretar os dados, aprender com as informações geradas e tomar decisões baseadas nos fatos observados. Aqui é o momento onde os usuários irão identificar características importantes de cada jogador, tendências de jogo, situações que favorecem ou prejudicam os resultados dos jogos, cargas de treinamento, dentre outros.

A metodologia apresentada na Figura 10 foi inspirada no modelo CRISP-DM, apresentado no Capítulo 3, e representa o processo utilizado neste trabalho e aplicado nos dois estudos de caso que serão vistos no Capítulo 5.

5 Estudos de caso

Neste capítulo serão apresentados dois estudos de caso, utilizando a metodologia proposta no Capítulo 4. As análises foram realizadas baseadas em dados de dois times profissionais da primeira divisão do campeonato brasileiro de futebol. Como são duas análises distintas iremos tratar como estudo de caso 1 e estudo de caso 2. O estudo de caso 1 compreende a temporada de 2018 e os estudos foram realizados com base em planilhas com informações de treinos, jogos, resultados laboratoriais e dados dos jogadores. O estudo de caso 2 compreende as temporadas de 2021 e 2022, com estudos realizados com base em planilhas com informações de treinos e jogos. Os dados dos dois estudos de caso foram fornecidos pelo fisiologista do clube.

5.1 Estudo de caso 1: análise de lesões e eficiência física

Primeiramente, utilizando a linguagem Python, foram selecionadas 3 guias das planilhas para análises iniciais que passaram por uma série de tratamentos em campos inconsistentes, remoção de jogadores teste, correção de erros de digitação, padronização de definições, substituição do nome para preservar identidades, etc. Estas tabelas são apresentadas a seguir:

- Arquivo “A-C TEMPORADA 2018 FORMULAS”, guia “Dados Gerais”, contendo 7.190 linhas e 13 colunas, apresentando dados acerca dos treinos realizados pelos atletas no intervalo de tempo, como distâncias percorridas a cada treino, PSE, etc;
- Arquivo “A-C TEMPORADA 2018 FORMULAS”, guia “Alt Int”, contendo 6.236 linhas e 3 colunas, apresentando dados acerca dos treinos realizados pelos atletas no intervalo de tempo, considerando os valores obtidos com carga em alta intensidade;
- Arquivo “CRONICA FORMATADA PARA ANÁLISE”, guia “DADOS TRATADOS VAR IND E DEN”, contendo 6.293 linhas e 51 colunas, apresentando dados acerca dos exames realizados pelos atletas em determinados períodos, como CK, HRV, dor, etc;

As 3 tabelas foram mescladas em uma única fonte de dados utilizando como índice o dia da observação e o atleta, resultando em uma base com 8300 linhas e 63 colunas, que será chamada de tabela concatenada.

Para um melhor entendimento das informações existentes na base estudada, os dados foram caracterizados e serão realizadas observações sob 3 pontos principais: Jogadores, Treino e Análises Laboratoriais.

5.1.1 Jogadores

As observações foram feitas em 46 jogadores distintos, distribuídos em 6 posições no campo. Existe a informação da idade do jogador, porém ela só foi reportada em 13 dos 46 jogadores, desta forma não é possível fazer considerações relevantes acerca das idades dos jogadores. Não existem muitos dados sobre os jogadores que poderiam ser importantes para as análises como peso, altura e outras características pessoais.

5.1.2 Treinos

A análise dos dados do treino foi realizada baseada nos dados gerais apresentados na planilha "A-C TEMPORADA 2018 FORMULAS". Como dito, esta tabela após tratamento resultou em 7190 observações. A tabela possui bastantes campos nulos e valores zeros, que são resultados da não coleta, ausência do jogador ou outro fator. Na Tabela 3 são apresentadas a definição de cada campo da tabela e a quantidade de valores nulos ou zeros que cada campo possui.

Tabela 3 – Dados da tabela de Treino

Campo	**Definição**	**Nulos**	**Zeros**
index	Índice autonumerado	0	1
Posicao	Posição em campo que o atleta realizou a atividade	1.127	0
Nick	Nome do atleta substituído por um apelido	0	0
Presenca	Tipo da atividade: Presente, ausente, jogo, treino, Jogo Sub23, etc	715	0
DATA	Data da atividade	0	0
EF	Estado Físico: Como o atleta estava se sentindo antes da atividade (0-10)	3.122	41
Disttotalm	Distância total percorrida na atividade	9	2.995
Distaltaintensidade(m)	Distância total percorrida em alta intensidade	8	3.451
MinutosTotais	Tempo gasto para realizar a atividade	14	1.896
PSE	Percepção do atleta sobre o treinamento após realizá-lo (0-10)	1.907	39
PSEXMIN	Produto do PSE e Minutos	1	2.085
DES	Desaceleração	1	3.363
ACE	Aceleração	1	3.325
Trimp	Carga baseada na frequência cardíaca	8	3.123
NumTreino	Número do treino no dia (1-Primeiro treino, 2-Segundo treino)	0	0

Acerca da presença dos jogadores nos treinos, cada participação foi caracterizada conforme a Tabela 4 que, além da definição de cada status de participação, também apresenta a quantidade de ocorrência de cada uma delas.

Tabela 4 – Participação dos Atletas nas Atividades

Status de Presença	Definição	Ocorrências
AUSENTE	Faltou à atividade	1
BASE	Atividade junto com a categoria de base	36
DM	No Departamento Médico no dia da atividade	698
FOLGA	Em folga no dia da atividade	278
G1	Iniciou o jogo como titular	338
G2	Iniciou como reserva	322
G3	Não jogou mas fez alguma atividade	36
JUSTIFICADO	Falta justificada	28
PRESENTE	Presente na atividade	4.570
SUB20	Atividade com a equipe sub-20	3
SUB23	Atividade com a equipe sub-23	40
TRANSIÇÃO	Treino Individual após DM	125

Na Tabela 5 é apresentada a quantidade de atividades que cada jogador realizou por tipo de presença.

Tabela 5 – Atividades dos Jogadores por Presença

ATLETA	AUS	BAS	DM	FOL	G1	G2	G3	JUST	PRES	S20	S23	TRAN
Jogador0			4	11	3	12	2		153		9	
Jogador1			4	9	18	7	1		155			
Jogador2			60	9	9	11		3	99			1
Jogador3			3	8		9	1		97			
Jogador4			9	8	20	4			151			2
Jogador5			5	10	5	14	2		149		9	
Jogador6			34	9	5	13	1	12	115			
Jogador7		34		8		1	4		57	1		3
Jogador8			9	8	1	24			151			
Jogador9			11	9	16	10			146			
Jogador10			19	10	17	5	2		140			5
Jogador11		1	3	9		3	3	4	147		8	
Jogador12			105	9	6	3			65		2	
Jogador13			42	5	21				123			
Jogador14			1	9	5	9	2		96			
Jogador15				9	24	6			152			
Jogador16			11	10	20	4	2	2	143			
Jogador17			116	5				1	39			33
Jogador18			49	9	20	3	1		96			14
Jogador19			13	9	7	16	1	1	142			

Continua na próxima página

Tabela 5 – continuação da página anterior

ATLETA	AUS	BAS	DM	FOL	G1	G2	G3	JUST	PRES	S20	S23	TRAN
Jogador20				9	27	3	1		152			
Jogador21			40	4	6	12			77			52
Jogador22			5	9	30		1		147			
Jogador23			40	9	16	2	3		121			4
Jogador24			12	5	3	4			42			3
Jogador25	1		12	9	12	13			146			
Jogador26			14	8		12	2		83			
Jogador27			5	7		15	3		160		4	
Jogador28			34	9	10	15			121			3
Jogador29			1	5	12			3	54			
Jogador30				9	9	17	1		155			
Jogador31			6	8	10	17	2		149		1	
Jogador32						10		2	61			
Jogador33				5		4	1		53	2	7	
Jogador34				3	2	14			88			
Jogador35				3	3	15			89			
Jogador36				3	1	14			90			
Jogador37									8			
Jogador38						1			31			
Jogador39			2						22			
Jogador40									31			
Jogador41									31			
Jogador42									30			
Jogador43									24			
Jogador44									3			
Jogador45									1			

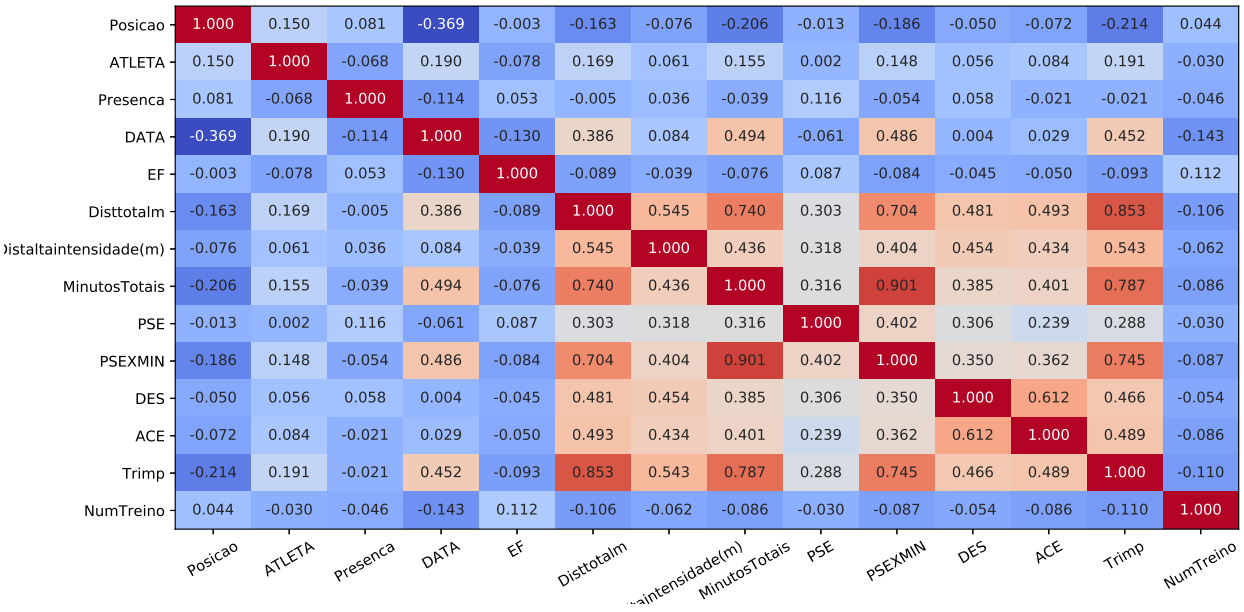
Considerando a posição que o atleta realizou a atividade, apresentamos na Tabela 6 o número de atividades por posição e a quantidade de atletas distintos que treinou em cada posição. Existem mais atletas por posição que o total de atletas, pois vários deles treinaram em mais de uma posição durante o período de análise.

Tabela 6 – Análise por posição

Posição	Atividades	ATLETAS
ATACANTE	1394	10
GOLEIRO	755	5
LATERAL	885	6
MEIA	870	34
VOLANTE	1129	7
ZAGUEIRO	1030	8

Os demais campos são numéricos e análises mais específicas serão realizadas através de métodos estatísticos e computacionais apresentados posteriormente. Para entender melhor a correlação entre as variáveis foi aplicada uma função do Python para obter essas correlações e o resultado obtido é apresentado na Figura 11. Nesta matriz, os valores possuem um intervalo entre -1 e 1, sendo que valores próximos de -1 e 1 apresentam alta correlação e próximos de 0 mostram que as variáveis possuem baixa correlação.

Figura 11 – Correlação entre as variáveis da tabela de atividades realizadas pelos atletas



Observando a Figura 11 é possível perceber que diversas variáveis possuem uma alta correlação entre elas, porém, algumas são valores derivados de outros campos, portanto a correlação é esperada. A variável TRIMP, que é baseada na frequência cardíaca do atleta, possui alta correlação com as distâncias e tempo de treino, o que também é compreensível. Um fato que chama atenção é que variáveis que são baseadas na percepção do jogador, como EF e PSE, não possuem relação com as demais variáveis, o que mostra que estado físico que o atleta relata estar sentindo antes e depois dos treinos pode não ter relação com as demais variáveis observadas.

5.1.3 Resultados laboratoriais

A análise dos dados dos resultados de exames laboratoriais foi realizada baseada nos dados gerais apresentados na planilha “A-C CRONICA FORMATADA PARA ANALISE”. Como dito, esta base, após tratamento, resultou em 6293 observações. A base possui bastantes campos nulos, que são resultados da não coleta, ausência do jogador ou outro fator. Na Tabela 7 são apresentados a definição de cada campo da tabela e a quantidade de valores nulos ou zeros que cada campo possui.

Tabela 7 – Dados da tabela Resultado.

Coluna	Definição	Nulos	Zeros
index	Índice autonumerado	0	1
DATA	Data Observação	0	0
ATLETA	Nome do atleta	0	0
POSIÇÃO	Posição	0	0
CK34-46HAPOS	Creatina quinase	6.154	0
PERCCKMAX	Percentual de Creatina sobre o máximo	6.156	0
CKMAX	Maior creatina medida na temporada	5.525	0
EficFis	Eficiência Física	6.179	0
DOR	Nível de Dor	6.181	20
HRV	Estado de Recuperação	5.581	0
Eficiencia1tRES	Eficiência Física no primeiro tempo	6.100	0
Eficiência2tRES	Eficiência Física no segundo tempo	6.100	0
EficiênciaTOTAL	Média das Eficiências	6.100	0
Idade	Idade do Atleta	6.280	0
Tipodelesão	Tipo da Lesão	6.280	0
Localdalesão	Local da Lesão	6.281	0
Lesãodescritiva	Descrição da Lesão	6.280	0
Exame	Tipo de Exame realizado	6.283	0
Severidadelesão	Grau da Lesão	6.291	0
Membrolesionado	Membro Lesionado	6.281	0
TreinooucompetiCAo	Ocorrência em jogo ou treino	6.280	0
TempoAfastamento(dias)	Tempo que deverá ficar afastado	6.281	0

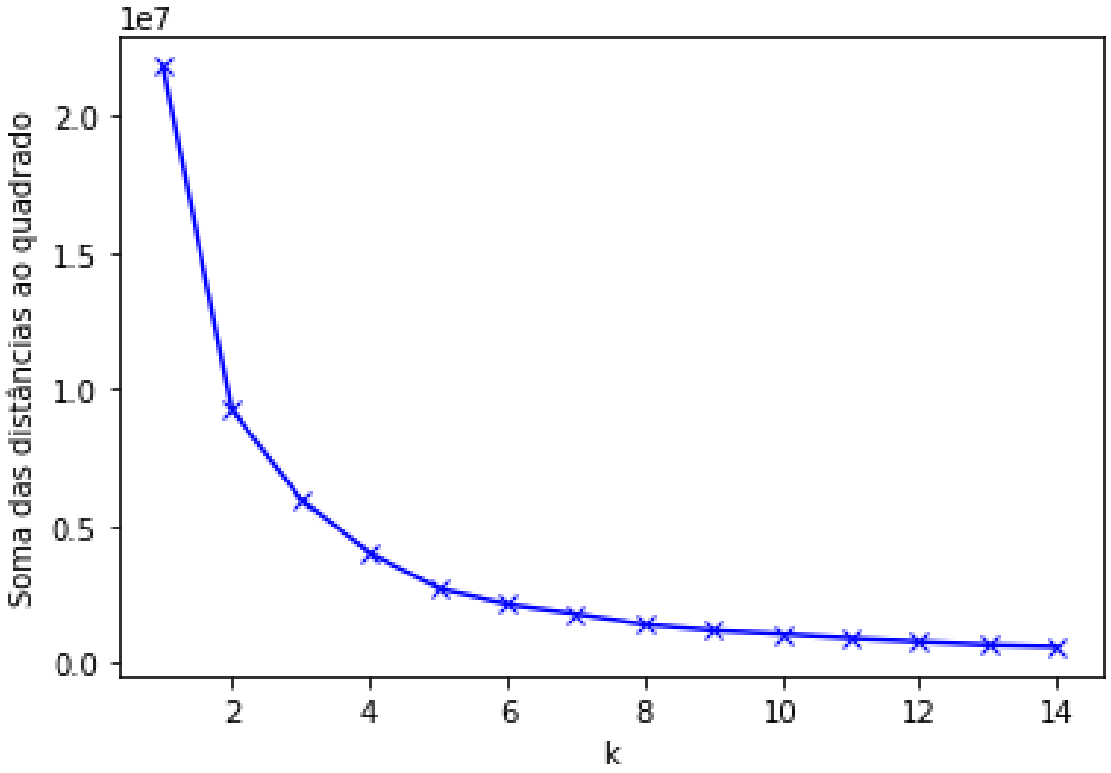
5.1.4 Agrupamento e lesões

Nesta seção será apresentado o agrupamento dos jogadores segundo características observadas nos treinamentos.

Para esta análise foi utilizado o método *K-means*, onde os jogadores foram classi-

ficados em grupos segundo as características coletadas na tabela de treino dos atletas. Para determinar o número de grupos que seriam considerados no algoritmo foi utilizado o método *Elbow* no Python, que calcula uma função de custo, que é a soma dos quadrados das distâncias internas dos grupos, que é determinado por K. O melhor número para a quantidade de grupos é quando a adição de um novo grupo não muda significativamente a função de custo. Neste caso, foi possível observar que o número ideal de grupos para classificar os jogadores seria entre 3 e 5, conforme é apresentado na Figura 12, sendo escolhida a utilização de 5 grupos.

Figura 12 – Método *Elbow* para determinar o número de grupos ideal



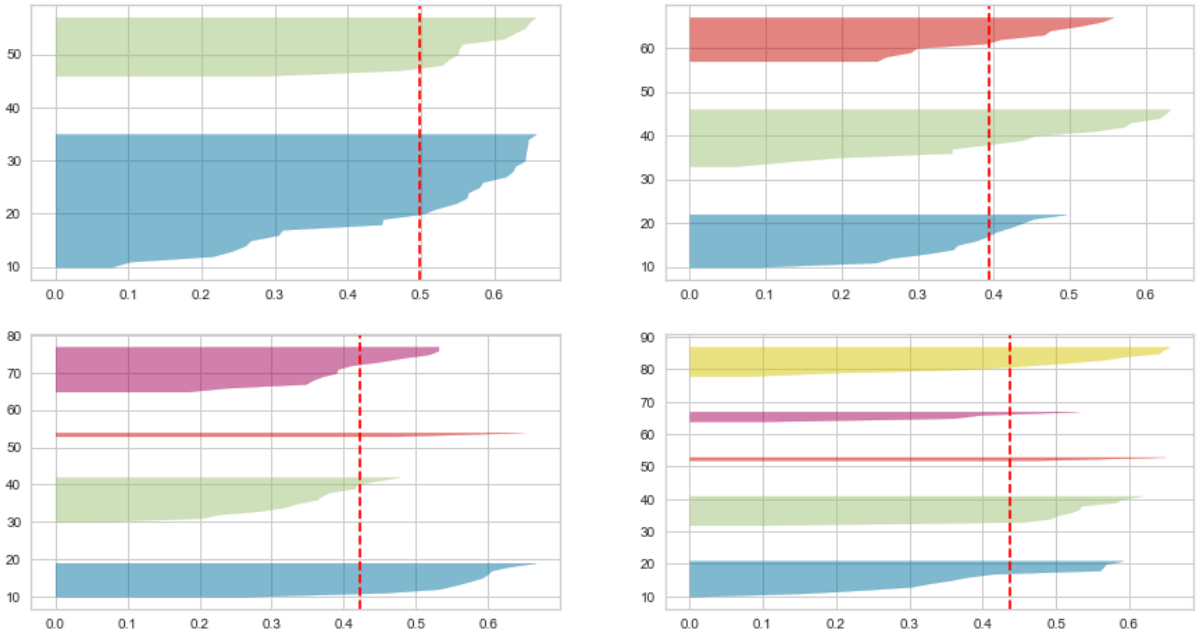
Ao aplicar o método *K-means* foram removidos jogadores que tinham menos que 10 observações durante a temporada e foram considerados os seguintes parâmetros coletados a cada treino do atleta: Distância total percorrida, TRIMP, EF, Distância em alta velocidade, desaceleração, aceleração, minutos totais do treino e PSE. Para cada um destes parâmetros foram calculados o valor médio e o desvio padrão por jogador e após executado o algoritmo *K-means* na ferramenta Python, os jogadores foram classificados em grupos que foram identificados de 0 a 4. Para entender melhor as características dos jogadores de cada grupo foram calculados a quantidade de jogadores lesionados que constava em cada um dos grupos e o resultado é apresentado na Tabela 8. Como é possível observar somente em dois grupos foram alocados jogadores lesionados, e no grupo com maior número de jogadores nenhuma ocorrência de lesão foi observada.

Tabela 8 – Atletas por grupos *K-mean*

Grupo	Num Atletas	Num Lesionados
0	14	0
1	9	4
2	4	0
3	9	3
4	2	0

Como os grupos 2 e 4 tiveram poucos jogadores e nenhum lesionado, foi aplicado o método *Silhoutte* para validar a escola da quantidade de grupos. A Figura 13 mostra que, as escolhas com 3 ou 5 grupos são adequadas, mas os 5 grupos como proposto anteriormente ainda é uma melhor opção. Sendo assim, será mantida a análise com 5 grupos.

Figura 13 – Método *Silhoutte* para validar o número de grupos ideal



O grupo ao qual cada jogador pertencia foi incluído na base de dados para que fossem analisadas características dos jogadores de cada grupo à partir das informações coletadas em todos treinos e jogos da temporada. Os detalhes sobre cada grupo são apresentados na Tabela 9. Fazendo uma análise comparativa entre os grupos que possuem jogadores que foram lesionados, 1 e 3, com os demais grupos que não possuem jogadores lesionados é possível observar que os atributo EF e PSE, que são valores que os jogados expressam sua percepção do estado físico antes e depois dos treinos, nos grupos que possuem lesionados a média e desvio padrão são maiores que os outros grupos, além de serem grupos onde algum jogador relatou EF com valor 9 e PSE com valor 10, que levanta uma questão importante e que merece ser melhor analisada, haveria uma relação entre o jogador estar se sentindo melhor e se lesionar? Outros atributos também se diferenciam

Tabela 9 – Características dos Grupos dos Jogadores

Medida		**Grupo** 0	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**
Número de Atletas		14	9	4	9	2
Número de Lesionados		0	4	0	3	0
EF	Média	3,61	3,73	3,49	3,81	3,63
	Desvio Padrão	0,93	1,07	0,95	1,0	0,99
	Máximo	7,0	9,0	6,0	9,0	8,0
Distância total	Média	3.628,01	3.811,96	2.819,50	3.623,84	4.418,96
	Desvio Padrão	1.950,73	2.865,22	1.940,62	2.435,86	3.433,99
	Máximo	10.235,0	11.694,0	8.945,0	11.122,0	11.818,0
Distância Alta Intensidade	Média	134,10	141,23	86,50	147,78	218,63
	Desvio Padrão	146,53	186,93	111,73	181,01	242,77
	Máximo	1136,0	1371,0	733,0	1215,0	1069,0
Minutos Totais	Média	75,07	71,83	67,02	71,32	72,06
	Desvio Padrão	26,32	28,35	25,03	27,41	30,18
	Máximo	190,37	190,37	132,0	163,37	190,37
PSE	Média	4,99	4,93	4,48	5,09	4,80
	Desvio Padrão	1,51	2,03	1,51	1,71	2,19
	Máximo	9,0	10,0	9,0	10,0	10,0
DES	Média	43,26	41,77	34,16	37,63	48,03
	Desvio Padrão	27,92	35,93	27,48	30,09	42,23
	Máximo	135,0	180,0	111,0	143,0	154,0
ACE	Média	15,02	12,97	13,64	12,86	13,98
	Desvio Padrão	12,02	11,33	11,25	10,66	12,73
	Máximo	210,0	69,0	51,0	53,0	58,0
Trimp	Média	136,01	140,47	124,50	125,84	135,47
	Desvio Padrão	83,43	118,32	93,08	96,37	120,35
	Máximo	402,96	461,75	389,1	437,7	437,3

entre os grupos com lesionados e não lesionados, como as distâncias e o tempo. Os grupos com jogadores que tendem a se lesionar aparentemente percorrem maiores distancias, em alta intensidade ou não, em períodos de tempo até menores que em outros grupos, o que pode levantar hipóteses que este esforço maior pode tender à lesões. Não foram observadas grandes diferenças para os parâmetros desaceleração, aceleração e TRIMP.

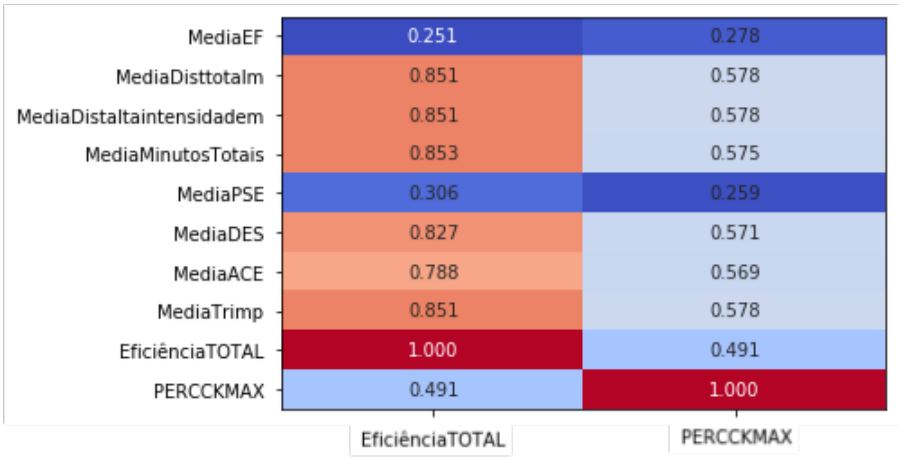
5.1.5 Análises de eficiência física

Duas informações importantes apresentadas na base de dados laboratoriais do atleta, coletadas através de exames e análises durante e após os jogos são a Eficiência Física do jogador e a creatina quinase (CK), que são informações sobre como o jogador rendeu fisicamente em campo, com base em frequência cardíaca, e como foi a recuperação do atleta dois dias após o jogo. Conhecer melhor o que influencia o jogador a obter melhores resultados neste parâmetro seriam de extrema importância para melhorar a performance dos atletas e fazer com que eles se recuperem mais rapidamente.

Um problema encontrado na base de dados laboratoriais é que, diferente dos dados dos treinos, como distância percorrida e estado físico, os dados relativos à eficiência física e CK não são coletados em todas atividades, e também não possuem um intervalo definido para serem coletados, ocorrendo semanalmente, quinzenalmente e até com intervalos maiores. Considerando isto, foi criada uma base de dados contendo a Eficiência Física, o CK e a média de todos dados dos treinos dos 10 dias anteriores ao observado nos dados laboratoriais, gerando uma tabela com 193 linhas.

Para uma primeira avaliação da influência do treino do atleta com sua eficiência e recuperação foi gerado um gráfico de correlacionamento entre estas variáveis, conforme apresentado na Figura 14. É possível observar que a correlação das variáveis sobre percepção do atleta sobre seu estado físico, seja antes ou após as atividades, EF e PSE, possuem baixa correlação com a eficiência e a recuperação, que deve ser melhor estudado para verificar se o atleta informa valores aleatórios sem uma análise do que ele realmente está sentindo, se a sensação do atleta realmente não tem relação com as variáveis ou algum outro fator psicológico. As demais variáveis apresentaram uma boa correlação com a variável de Eficiência Física e uma correlação média com o CK, mas que não pode ser descartada, pois mesmo os valores próximos de 0.5 podem ser significativos, mas outras análises são necessárias.

Figura 14 – Correlação entre Eficiência Física e CK com médias dos últimos treinos

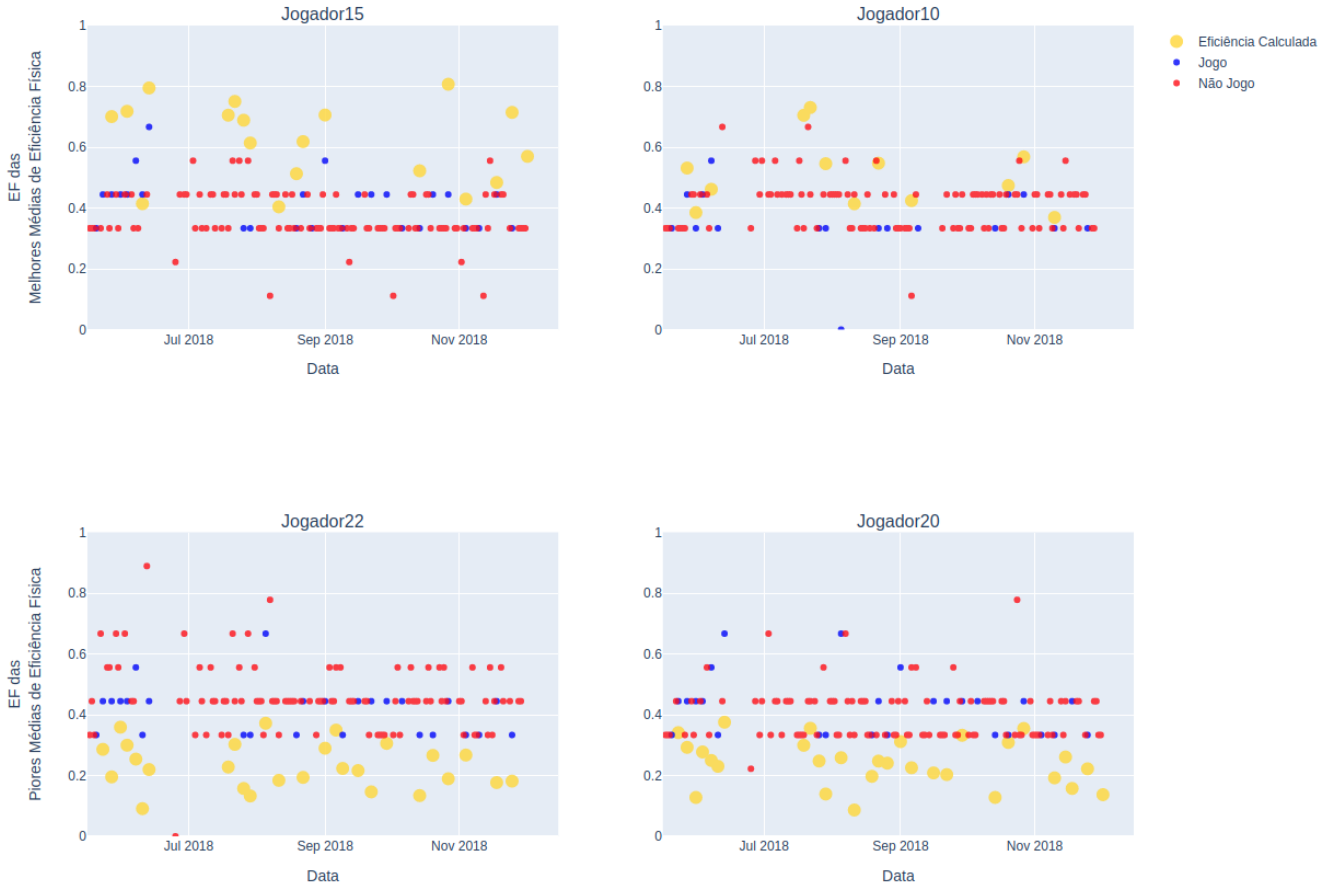


Diante da alta correlação entre a eficiência física e demais variáveis, foram verificados quais eram os atletas com melhores e piores médias de eficiência física, sendo desconsiderados goleiros e jogadores que haviam realizado menos que 3 observações desta eficiência. As melhores médias foram dos jogadores 15 e 10, já as piores médias foram dos jogadores 22 e 20. Os dados então foram plotados em figuras, com os jogadores com melhores médias na primeira linha e os com piores médias na segunda linha,

relacionando a eficiência física com as demais variáveis. As atividades realizadas foram diferenciadas entre jogos e não jogos para auxiliar a análise. Como os valores das variáveis possuem escalas diferentes, todas variáveis foram normalizadas para valores entre 0 e 1, facilitando assim as análises nos gráficos.

A Figura 15 apresenta pontos referentes às percepções dos jogadores acerca do estado físico antes das atividades e as eficiências físicas examinadas. É possível perceber que não existem diferenças visíveis entre os jogadores das melhores e piores médias de eficiência, estando ambos os grupos com valores predominantemente entre 0,4 e 0,6, sendo que os jogadores das piores médias aparentemente sinalizam estar com estado melhor que os das melhores médias, porém, majoritariamente os jogadores tendem a dar a mesma resposta em praticamente todas atividades.

Figura 15 – Melhores e Piores Médias de Eficiência por Percepção do Estado Físico

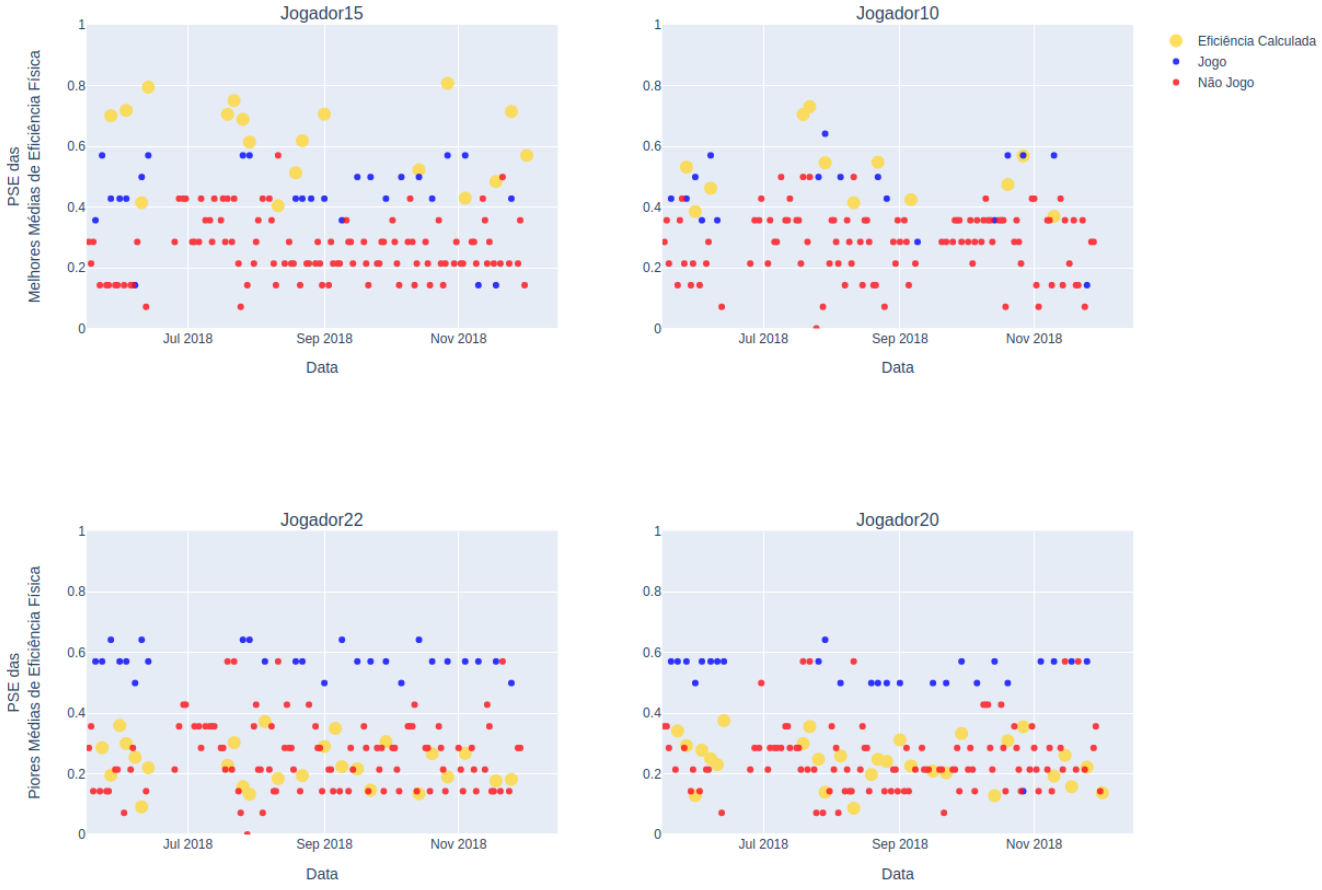


Diferente da percepção do estado físico antes das atividades, na percepção após as atividades é possível perceber uma diferença entre os grupos dos jogadores com piores e melhores médias, como é apresentado na Figura 16. É possível perceber que os jogadores com melhores médias, além de possuir uma dispersão maior dos dados, não diferem tanto a percepção em atividades relacionadas à jogos e treinos, enquanto os com piores médias visivelmente tendem a relatar um estado físico melhor pós atividades de jogos do que em

treinos.

Entender melhor este comportamento pode ser importante para descobrir o motivo destes jogadores com piores médias queixarem mais em treinos, se isto influência na carga de treinamento que é indicada pelos preparadores, se eles tentam com isso aumentar o descanso, entre outros.

Figura 16 – Melhores e Piores Médias de Eficiência por Percepção do Estado Físico pós atividade

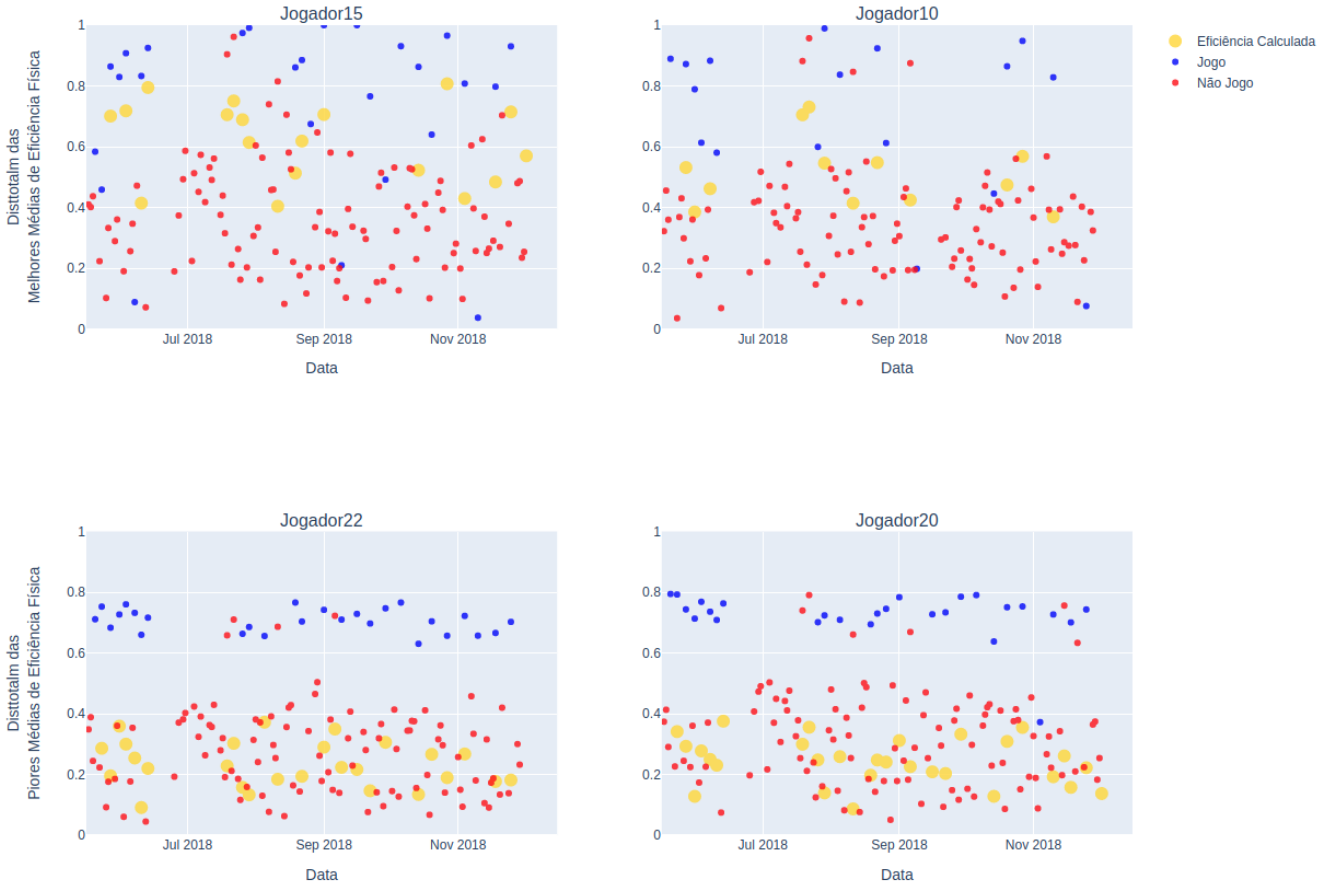


A Figura 17 apresenta a distância total percorrida pelos jogadores dos dois grupos, sendo que em ambos a distância percorrida em jogos é maior que a percorrida em atividades que não são jogos. Uma diferença entre os grupos é que os jogadores com melhores médias aparentemente percorrem distâncias maiores, correndo em dias de jogos entre 0,7 e 1,0 e em atividades que não são jogos variando até 0,6, enquanto o grupo com jogadores com piores médias variam a distância entre 0,6 e 0,8 em jogos e com variações até 0,5 quando não estão em jogos.

Aqui podemos observar que jogadores com melhores médias são os que percorrem as maiores distâncias nos jogos, mas que essa distância varia entre os jogos, enquanto os jogadores com piores médias tendem a percorrer a mesma distância nos jogos, com pouca variação. Nos atletas com melhores existe uma grande dispersão dos dados, com jogos

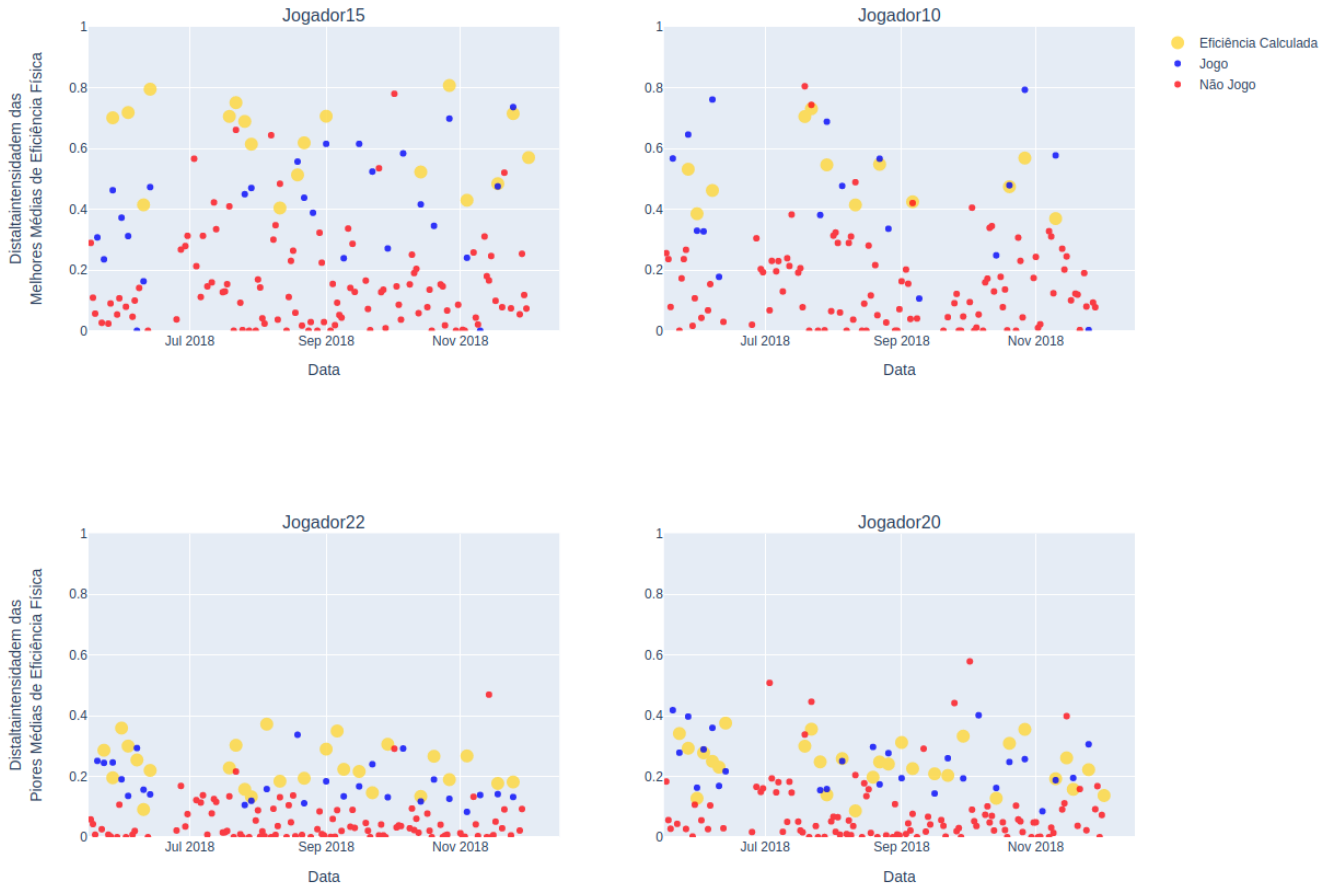
e treinos com distâncias parecidas, enquanto jogadores com piores médias claramente percorrem distâncias maiores nos jogos, necessitando analisar se esta diferença entre treino e jogo influência na eficiência física, auxiliando os preparadores a adequarem as distâncias percorridas nos treinos.

Figura 17 – Melhores e Piores Médias de Eficiência por Distância percorrida



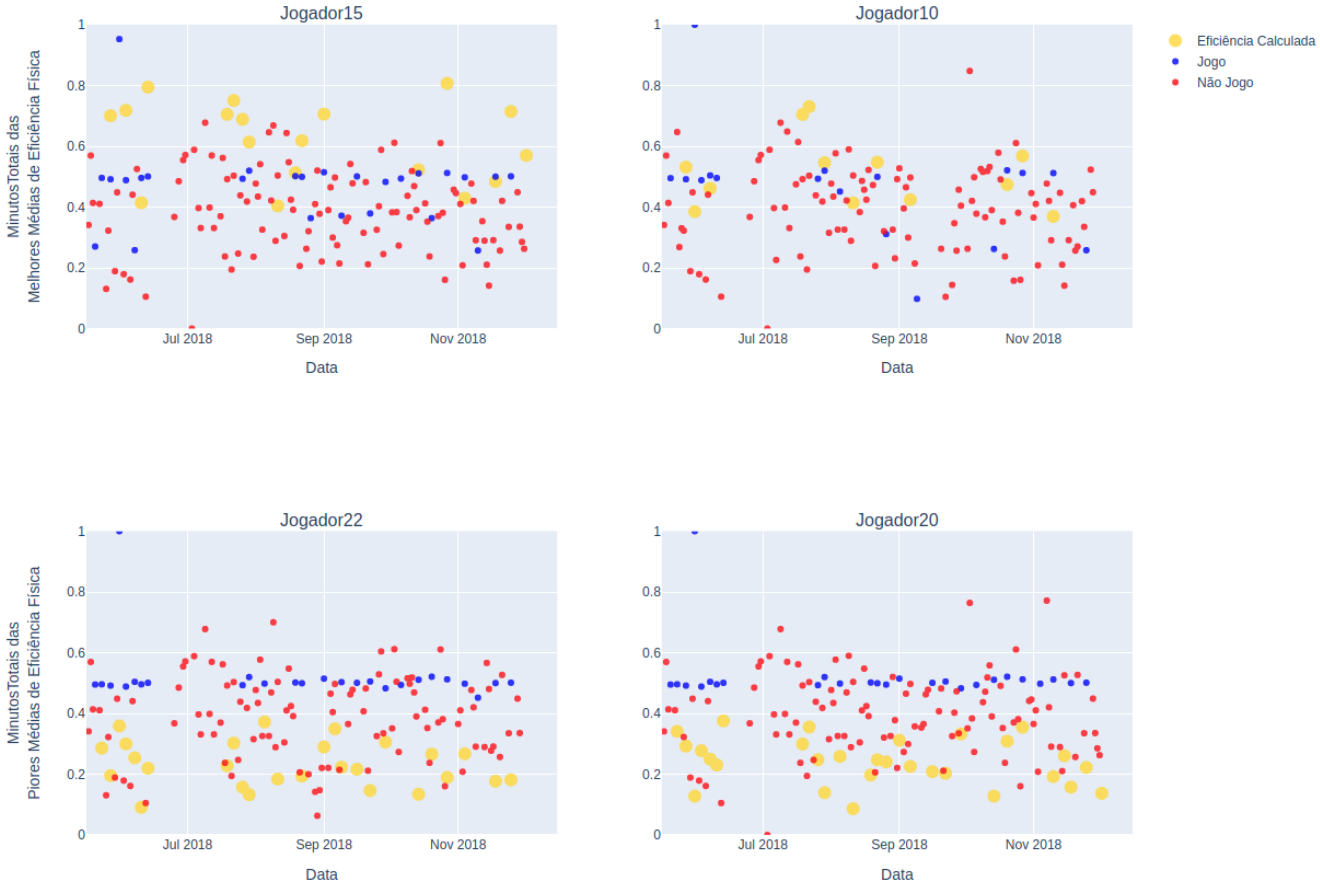
Assim como na distância total percorrida, quando são analisadas as distâncias em alta intensidade, os dois grupos apresentam desempenho maior em dias de jogos, como é mostrado na Figura 18. Os grupos também se diferem nesta variável pela dispersão dos dados, tendo os de melhores médias variações chegando aos 0,7, enquanto os piores variam até 0,4. Outra vez, os jogadores com melhores médias, possuem uma dispersão maior dos dados, o que indica que os treinos destes jogadores possuem cargas mais variadas que as dos jogadores com piores médias.

Figura 18 – Melhores e Piores Médias de Eficiência por Distância percorrida em alta intensidade



A Figura 19 apresenta o tempo total de atividades dos jogadores, e como é esperado não existem grandes diferenças entre os grupos com melhores e piores médias, visto que os jogadores treinam juntos e jogam os mesmos jogos.

Figura 19 – Melhores e Piores Médias de Eficiência por Minutos Totais



As figuras 20 e 21 apresentam, respectivamente, a aceleração e desaceleração dos jogadores. Enquanto na aceleração os dois grupos apresentam graficamente as mesmas características, na desaceleração é possível perceber que o grupo com melhores médias possui uma maior dispersão dos dados, além de não apresentar grandes diferenças na medição em dias de jogo ou não.

Figura 20 – Melhores e Piores Médias de Eficiência por Aceleração

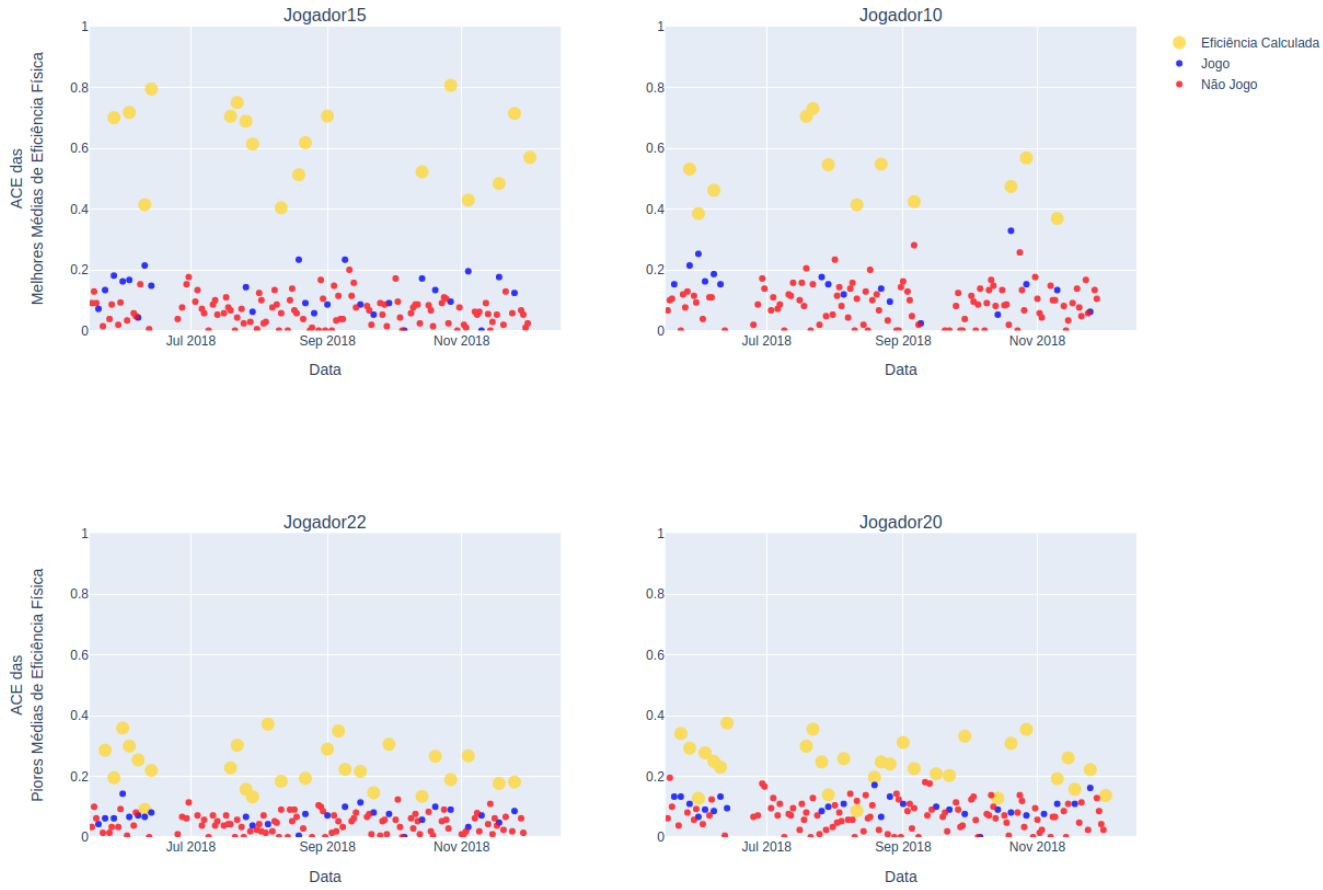
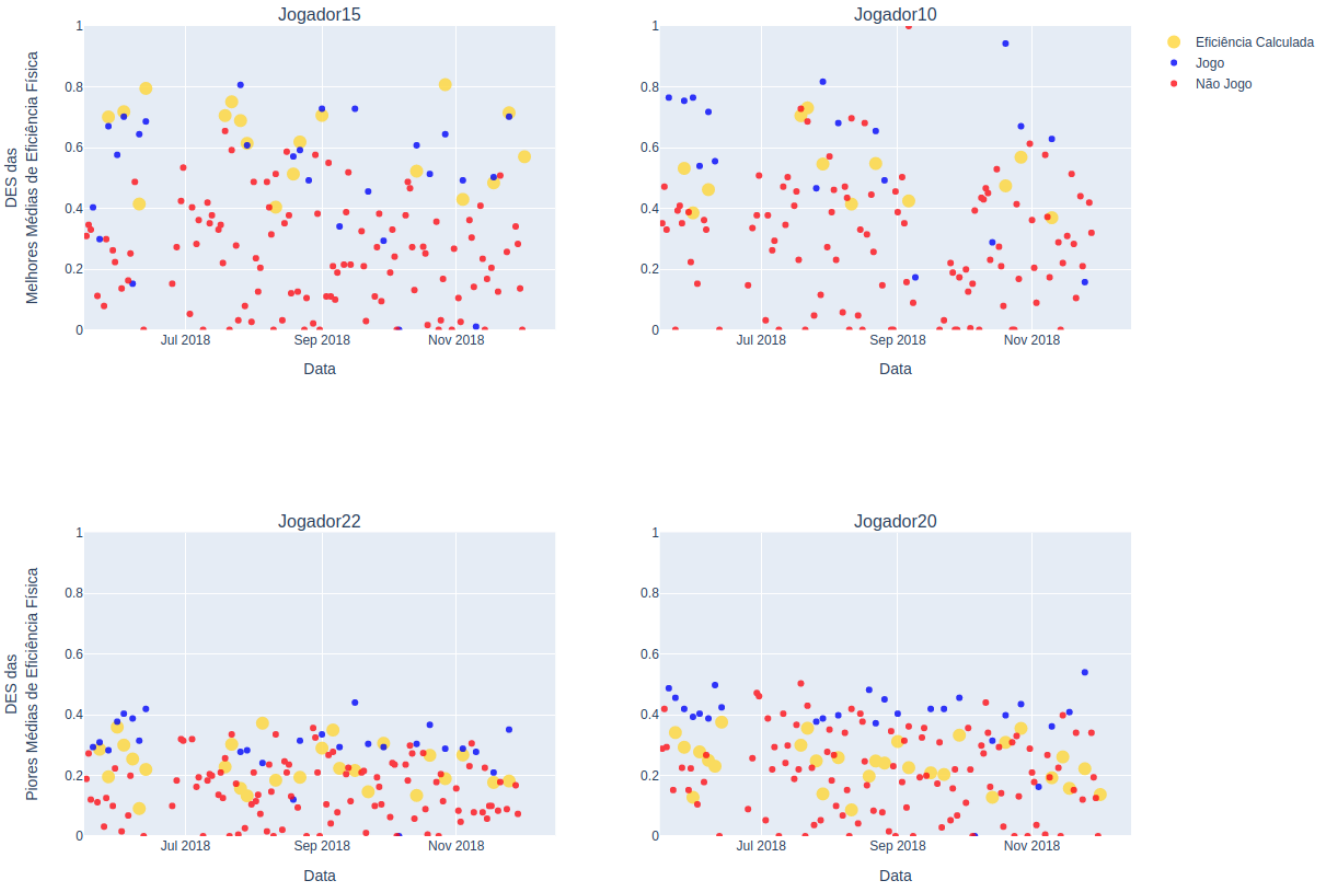
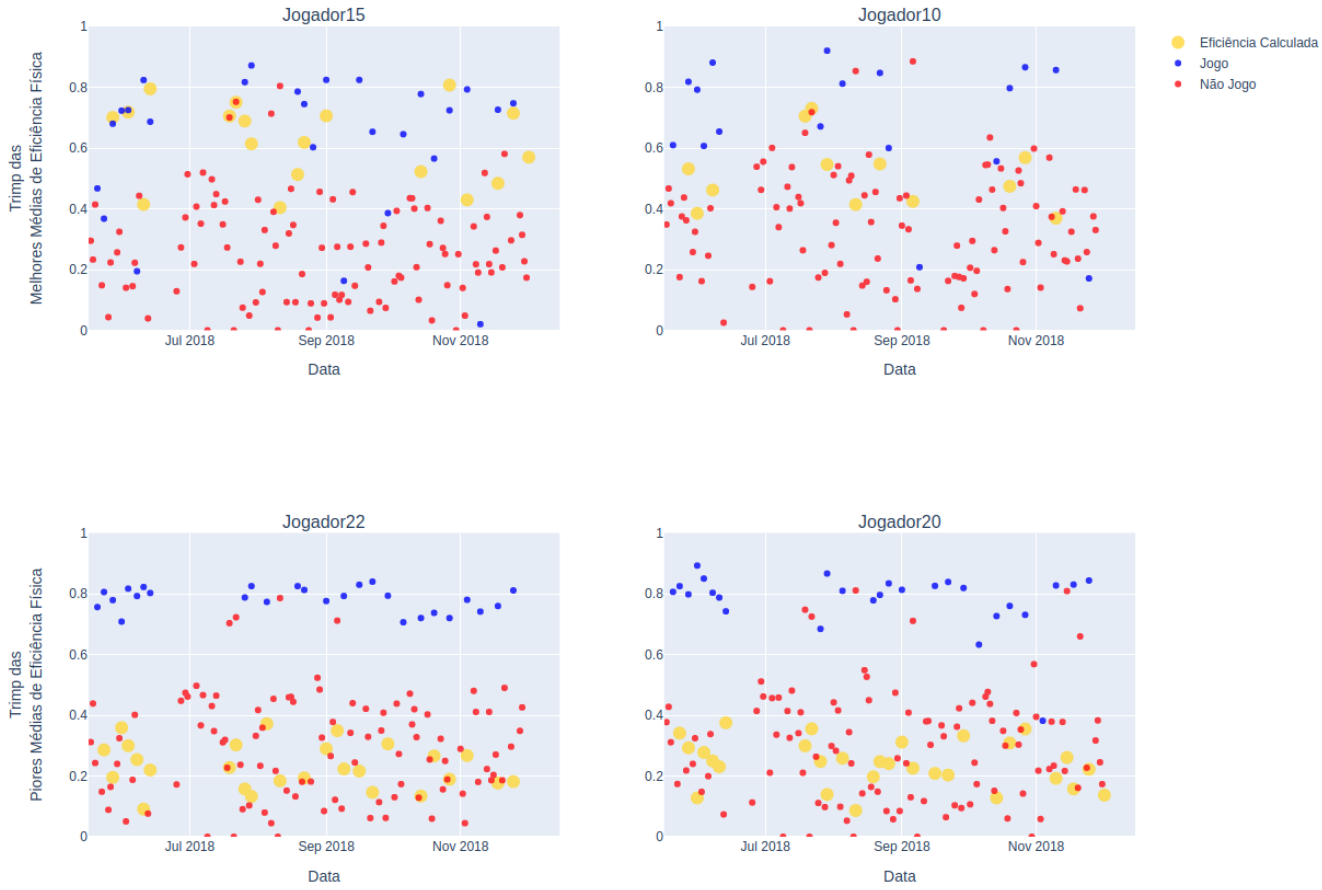


Figura 21 – Melhores e Piores Médias de Eficiência por Desaceleração



Na medição do TRIMP, que é uma variável baseada na frequência cardíaca e é apresentada na Figura 22, ambos os grupos possuem uma boa dispersão dos dados, mas a variação das medições em atividades de jogos e não jogos é maior no grupo com jogadores com piores médias de eficiência física, ou seja, os jogadores com melhores médias possuem o TRIMP nos dois tipos de atividades aproximados, enquanto os jogadores com piores médias apresentam em dias de jogos medições acima de 0,8, enquanto em atividade que não são jogos próximas de 0,4.

Figura 22 – Melhores e Piores Médias de Eficiência por Trimp



É possível perceber diferenças significativas nos dados dos jogadores com piores e melhores médias de eficiência física, principalmente no que diz respeito à dispersão dos dados, sendo que os jogadores com melhores médias variam mais as medições das variáveis durante a temporada, além disso, os jogadores com melhores médias variam menos entre atividades de jogos e treino, enquanto o grupo de jogadores com piores médias, aparentemente apresentam medições maiores em atividades de jogos.

5.2 Estudo de caso 2: análise de resultados e performance

Nesta base, também utilizando a linguagem Python, foram selecionados os dados de duas planilhas do excel, uma com dados de treino e outra com dados de jogos, ambas possuíam dados pessoais dos atletas, dados dos jogos e dos adversários, além dos dados coletados por GPS. Ambas planilhas possuíam dados das temporadas 2021 e 2022. Para realizar as análises iniciai foram necessárias uma série de tratamentos em campos inconsistentes, remoção de linhas com dados agrupados, correção de erros de digitação, padronização de definições, substituição de nomes para preservar identidade, etc. Estas tabelas são apresentadas a seguir:

- Arquivo “Dados Treino”, guia “BD GPS”, contendo 17.757 linhas e 49 colunas, apresentando dados acerca dos treinos realizados pelos atletas, como data, período, intensidade, distâncias percorridas, etc. Várias colunas possuem dados derivados de outras colunas;
- Arquivo “Dados Jogo”, guia “Data GPS Partida”, contendo 4.141 linhas e 83 colunas, apresentando dados acerca dos jogos realizados, como dados do time, dados do adversário, período sem jogar, distâncias percorridas com variações das velocidades, etc. Várias colunas possuem dados derivados de outras colunas.

As 2 tabelas foram mescladas em uma tabela com dados dos jogos e treinos. A nova tabela gerada continha vários campos repetidos ou derivados de outros campos que foram removidos. As informações consideradas mais relevantes pelos autores e fisiologista do clube foram mantidas e apresentadas na Tabela 10, com a descrição das colunas das tabelas mescladas.

Tabela 10 – Dados das tabelas de Jogos e Treinos

Coluna	**Descrição**
Item	Tipo de atividade, podendo ser treino, jogo, primeiro ou segundo tempo.
Data	Data da atividade.
Hora	Hora da atividade.
Local	Dentro ou fora de casa.
GF	Gols à favor.
GC	Gols contra.
Posse Equipe	Tempo de posse de bola da equipe.
Posse Rival	Tempo de posse de bola do rival.
Rival	Time rival.
Jogador	Nome do atleta.
Posição	Posição do atleta em campo.
Idade	Idade do atleta.
Última Partida	Quantidade em horas da última partida do jogador.
Tempo	Duração da atividade.
Distância Total	Distância total percorrida pelo atleta na atividade.
Distância em Speed Zone 3	Distância percorrida pelo atleta em velocidade entre 14.40 e 19.79 Km/h.
Distância em Speed Zone 4	Distância percorrida pelo atleta em velocidade entre 19.80 e 25.19 Km/h.
Distância em Speed Zone 5	Distância percorrida pelo atleta em velocidade acima de 25.20 Km/h.
Número de acelerações (2.00 - 2.99)	Quantidade de acelerações do atleta no intervalo entre 2.00 e 2.99 m/s2.
Número de acelerações (-2.99 - -2.00)	Quantidade de acelerações do atleta no intervalo entre 2.00 e 2.99 m/s2 (mudança de direção).
Número de acelerações (3.00 - 50.00)	Quantidade de acelerações do atleta no intervalo entre 3.00 e 50.00 m/s2.
Número de acelerações (-50.00 - -3.00)	Quantidade de acelerações do atleta no intervalo entre 3.00 e 50.00 m/s2 (mudança de direção).

Além dos dados fornecidos nas planilhas, dois campos foram criados para auxiliar nas análises, um referente ao nível do time rival e outro referente ao resultado da partida. O campo nível do rival foi gerado baseado na posição que o time ficou no campeonato, ou seja, o time campeão é nível 1, o segundo colocado nível 2 e assim por diante. O nível do rival também foi classificado com uma variável categórica, sendo considerado um time "BOM"quando terminou o campeonato na primeira metade da tabela e considerado um time "RUIM"quando terminou o campeonato na segunda metade da tabela. Já o campo referente ao resultado da partida foi gerado baseado em informações repassadas pelo fisiologista do clube, que classificou o que era considerado por eles como um resultado positivo, negativo ou neutro. Este valor leva em consideração, além do resultado da partida, o nível do rival, se o jogo foi fora de casa e se o resultado do jogo inverteu do primeiro para o segundo tempo.

A Tabela 11 apresenta estes critérios.

Tabela 11 – Classificação do resultado da partida

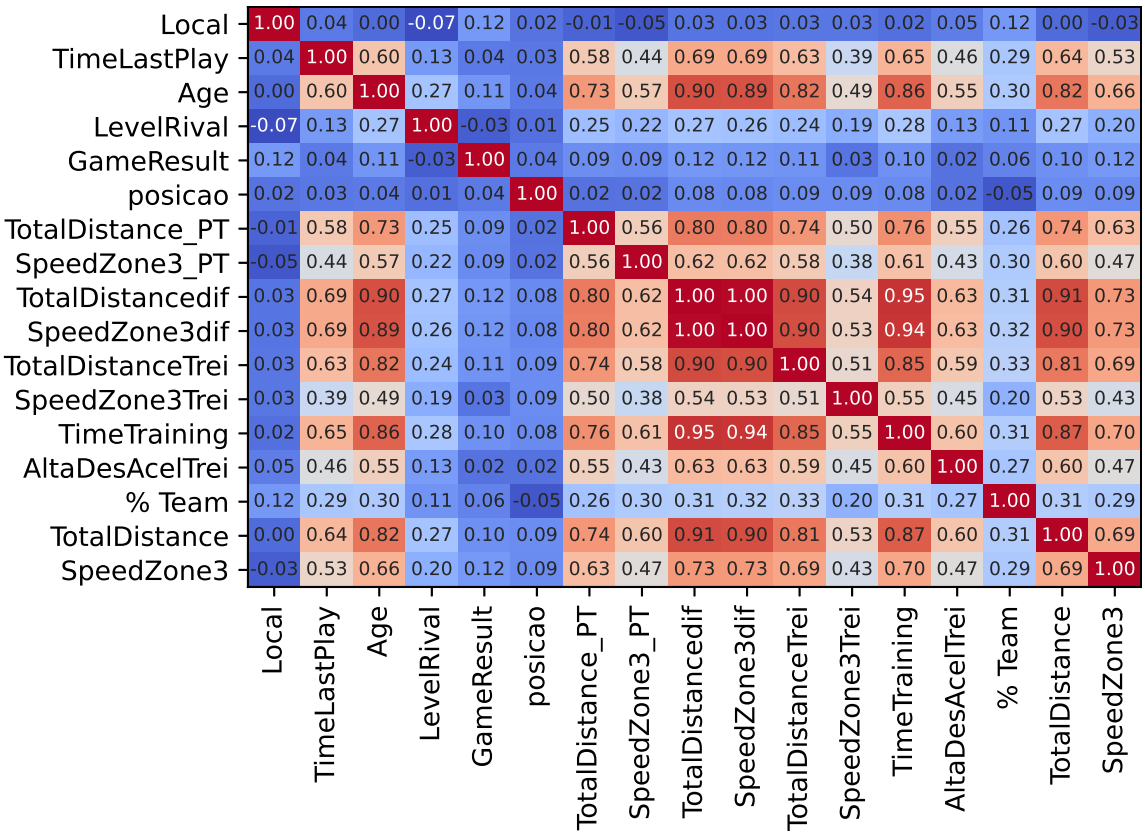
Critérios	**Resultado**
Terminou o primeiro tempo perdendo e terminou o segundo tempo empatado ou ganhando	Positivo.
Terminou o primeiro tempo ganhando e terminou o segundo tempo perdendo	Negativo.
Rival Bom, jogo dentro de casa e venceu	Positivo.
Rival Bom, jogo dentro de casa e empatou	Neutro.
Rival Bom, jogo dentro de casa e perdeu	Negativo.
Rival Bom, jogo fora de casa e ganhou	Positivo.
Rival Bom, jogo fora de casa e empatou	Positivo.
Rival Bom, jogo fora de casa e perdeu	Neutro.
Rival Ruim, jogo dentro de casa e ganhou	Neutro.
Rival Ruim, jogo dentro de casa e empatou	Negativo.
Rival Ruim, jogo dentro de casa e perdeu	Negativo.
Rival Ruim, jogo fora de casa e ganhou	Positivo.
Rival Ruim, jogo fora de casa e empatou	Neutro.
Rival Ruim, jogo fora de casa e perdeu	Negativo.

Com os dados separados do primeiro e segundo tempo de jogo, foram geradas métricas das diferenças de performance entre o primeiro e segundo tempo, com os mesmos nomes dos campos originais acrescidos do sufixo "dif", que basicamente é a diferença de performance apresentada pelos atletas no segundo tempo, ou seja, este campo é um percentual de quanto o atleta ganhou ou perdeu de performance entre o primeiro e o segundo tempo.

Para garantir o cálculo correto da diferença entre os dois tempos de jogos, apenas jogadores que jogaram completamente primeiro e segundo tempos foram mantidos na base. Posteriormente, buscando manter um pouco mais de volume na base de dados, optou-se por alterar essa regra e manter jogadores que jogaram pelo menos 80 minutos.

Para entender melhor as variáveis e a correlação entre elas algumas análises foram feitas, na Figura 23 é apresentado um gráfico de correlação com dados gerais do atleta e do jogo, além das informações de distâncias percorridas e distâncias percorridas na speedzone3.

Figura 23 – Correlação entre as variáveis gerais, treino, primeiro tempo e diferenças



Observando a Figura 23 é possível perceber que a idade do atleta e tempo desde o último jogo possuem boas correlações com as distâncias, além disso, as distâncias medidas no primeiro tempo, no último dia de treino, no jogo total e a diferença entre o primeiro e segundo tempo possuem altas correlações entre si, o que nos leva a analisar se a distância percorrida no treino pode influenciar na performance do atleta no jogo, ou se dados do treino e do primeiro tempo podem ter relação com a diferença de performance do atleta no segundo tempo, dentre outras.

Para melhorar o entendimento sobre os treinos dos atletas, foram criados novos campos sobre os dados dos treinos, como médias, mínimos, máximos, informações sobre os treinos dos 3 dias anteriores aos jogos, dentre outros.

Com a inclusão de novos, geração de dados separados por treino, primeiro tempo, segundo tempo, diferença entre os tempos, jogo total, colunas sumarizadas, dentre outros, foi necessário realizar estratégias para reduzir a quantidade de variáveis para facilitar o estudo através dos algoritmos. Primeiramente foi aplicado o método PCA, para definir componentes principais e reduzir a dimensionalidade da base. Utilizando 15 componentes principais foi possível explicar mais de 76% dos dados, como mostra a Figura 5.2.

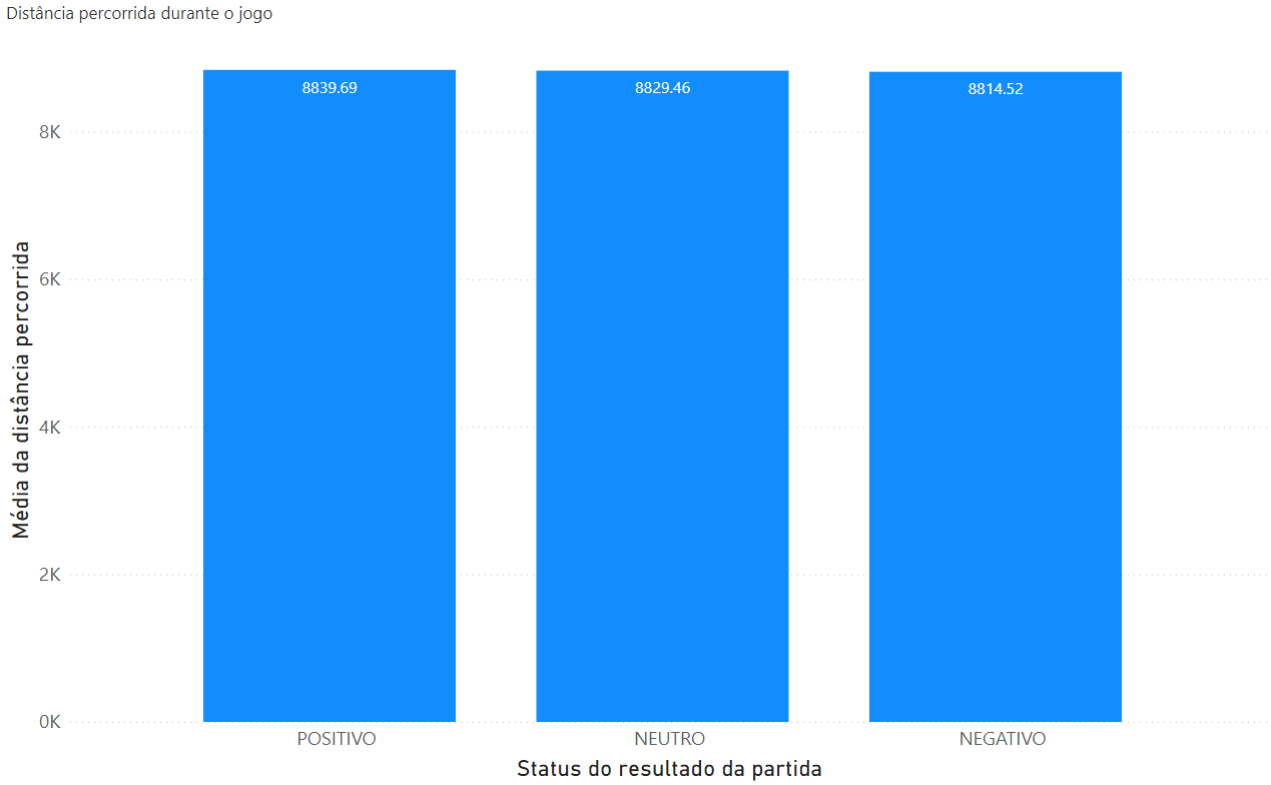
PCA	**Percentual**
1	0.15591224
2	0.12812059
3	0.091817
4	0.0548753
5	0.04784123
6	0.04095924
7	0.03852301
8	0.03351885
9	0.0313237
10	0.02876787
11	0.02527338
12	0.02373374
13	0.02134642
14	0.02041676
15	0.01960325

Tabela 12 – Componentes principais do método PCA

5.2.1 Análises

Com as primeiras observações realizadas na fase de tratamento dos dados, análises das distâncias começaram a ser estudadas para entender sua influência nos resultados do atleta e da equipe. Utilizando as variáveis de status do resultado da partida e as distâncias percorridas pelo atleta foram gerados alguns gráficos que serão apresentados a seguir.

Figura 24 – Média da distância total percorrida pelos atletas nos jogos.



A Figura 24 mostra a distância total percorrida pelos atletas e o status do resultado da partida, nela podemos perceber que de uma forma geral a distância total percorrida não é fator determinante para o resultado, visto que a média das distâncias em jogos com resultados negativos ou positivos são praticamente os mesmos.

Figura 25 – Média da diferença da distância percorrida no segundo tempo em relação ao primeiro tempo.

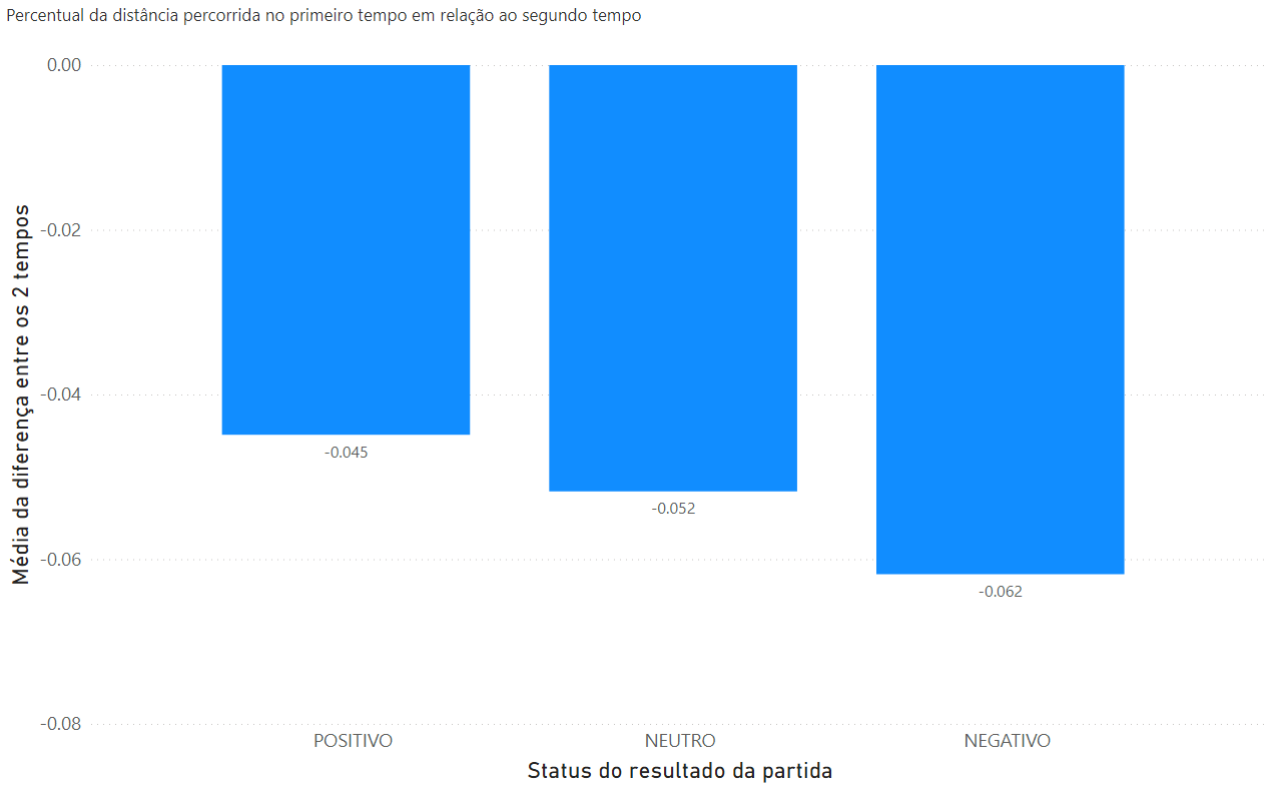
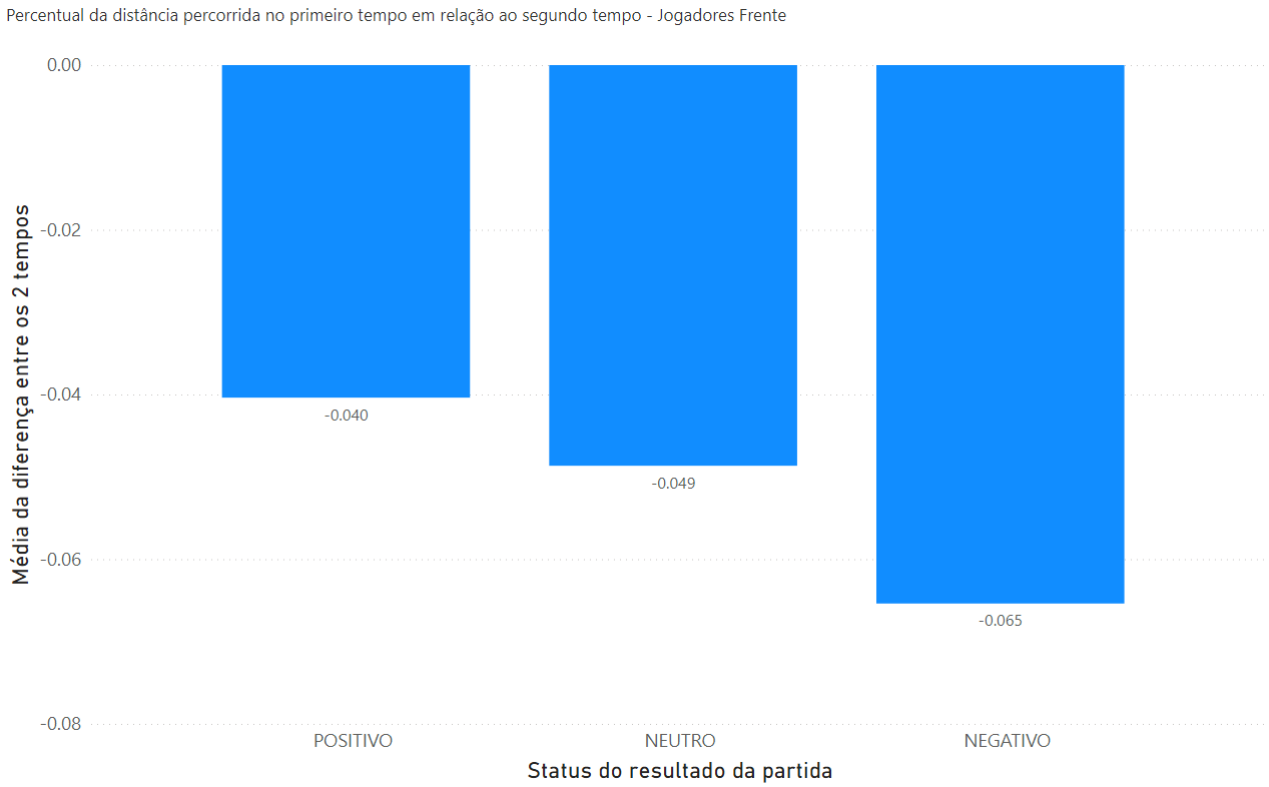


Figura 26 – Média da diferença da distância percorrida pelos jogadores de frente no segundo tempo em relação ao primeiro tempo.



A Figura 25 mostra a diferença, em percentual, da distância percorrida no segundo tempo em relação ao primeiro tempo, podemos ver que os valores são negativos, pois de uma forma geral os atletas perdem performance na segunda metade do jogo, porém, é possível observar que nos resultados negativos a perda é maior que em resultados positivos. Na Figura 26, que mostra os mesmos dados porém apenas com atletas de meio de campo e ataque, a diferença é ainda maior, e esta diferença aumenta analisando posições ou atletas específicos, o que nos mostra que quando os jogadores perdem muita performance entre os dois tempos o resultado tende a ser negativo, nos levando a analisar melhor os fatores que influenciam esta perda de performance e tentar prever este valor no intervalo dos jogos para auxiliar nas decisões tomadas.

Para compreender melhor esta relação, o percentual de variação entre as distâncias percorridas nos primeiro e segundo tempos foram calculados para cada observação na base de dados, este valor será definido daqui em diante como $\Delta dist$, que pode ser definido por:

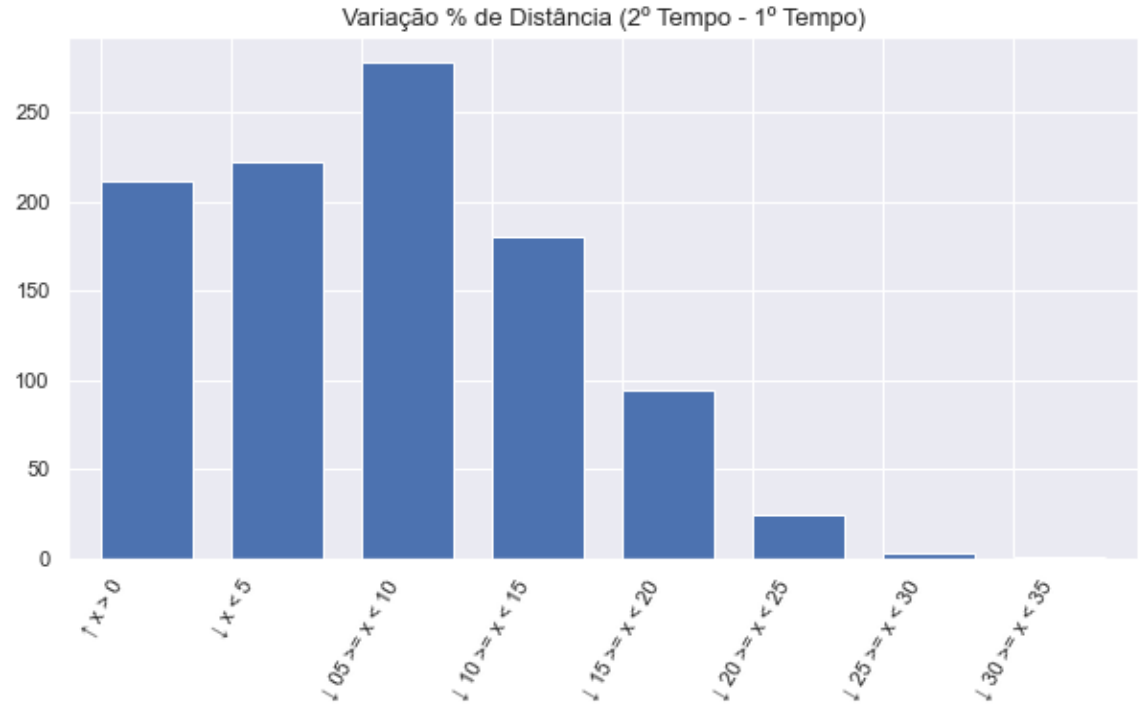
$$\Delta dist = \frac{Distancia2tempo - Distancia1tempo}{Distancia1tempo}.$$

Como o *target* $\Delta dist$ corresponde a valores contínuos optou-se por iniciar a modelagem através da aplicação de técnicas de regressão, como Random Forest Regressor e XGBoost Regressor, não obtendo bons resultados. Assim, buscando alternativas de melhoria nos resultados, foram observadas a importância de cada variável na execução dos algoritmos, percebendo que o grande número de variáveis não traziam ganho de informação e que poderiam estar prejudicando o comportamento dos modelos. Para diminuir o número de variáveis, selecionando as mais adequadas, foi utilizada a biblioteca *featurewiz*, que utiliza uma aplicação recursiva do XGBoost para a seleção das variáveis. Foram removidas variáveis com baixa representatividade ou altamente correlacionadas, reduzindo de 87 para 25 variáveis, mas ainda não apresentando resultados satisfatórios para os métodos de regressão.

Como os resultados não foram bem descritos pelos modelos de regressão, optou-se por transformar o $\Delta dist$ em classes. Transformar o *target* de valores contínuos para categorização em um número menor de classes pode gerar melhores resultados, dado que o volume de dados não é significativamente grande para explorar as diferentes combinações e cenários, podendo dificultar o reconhecimento de padrões mais bem definidos. A Figura 27 apresenta o o $\Delta dist$ em classes considerando situações onde o atleta melhorou a performance no segundo tempo, e intervalos de 5% para situações onde o atleta piorou a performance entre os tempos.

A Figura 27 mostra a distribuição das variações percentuais de $\Delta dist$, onde $\uparrow$ representa os casos em que o valor aumentou do segundo para o primeiro tempo e $\downarrow$ os casos em que houve queda de desempenho. Observa-se que os valores estão altamente

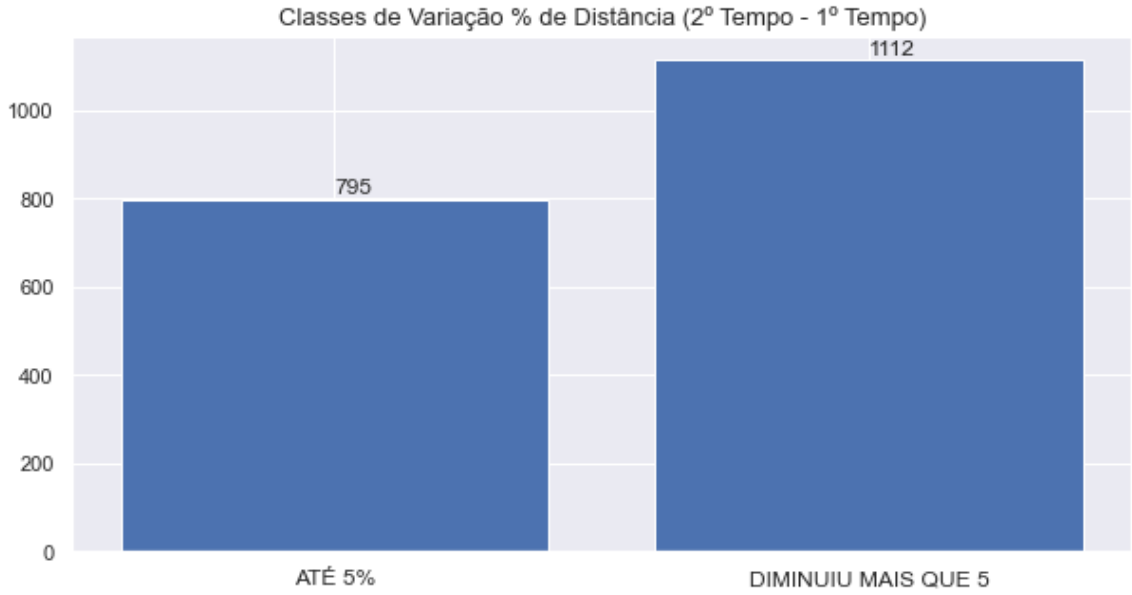
Figura 27 – Variação percentual de distância percorrida $\Delta dist$.



concentrados nos primeiros grupos. De maneira empírica e considerando que, na prática, identificar se o jogador irá perder muito ou pouco rendimento já seria suficiente para auxiliar na tomada de decisão, uma nova divisão foi feita em duas classes, buscando manter coerência e balanceamento na quantidade de observações. As classes definidas foram os casos em que a variação da distância percorrida aumentou ou diminuiu até 5% e outra com valores onde a distância percorrida no segundo tempo diminui mais que 5% em relação ao primeiro tempo. Esses valores também fazem sentido em termos de negócio, já que jogadores que possuem uma queda de rendimento maior poderiam ser separadamente avaliados. O resultado final dessa divisão pode ser visualizado na Figura 28.

Após vários experimentos, foi percebido que fazer um teste com a base completa não trazia bons resultados visto que observações muito antigas não faziam tanto sentido nas análises, por exemplo, um jogo realizado em fevereiro não poderia servir de parâmetro para um jogo realizado em novembro, visto que o time mudou, a performance do atleta, e outras variáveis. Dessa forma, foi realizado um *split* na base, dividindo-a em 4 partes, e cada parte foi analisada separadamente. É importante considerar que os dados estão ordenados por data, para preservar as questões temporais. A partir dos dados categorizados torna-se possível a aplicação de técnicas de classificação. Cada *split* base de dados inicial foi divida em dois subconjuntos, o primeiro com 80% dos dados, chamado de Treinamento e o segundo com 20% denominado Teste. Em uma avaliação simples, o modelo foi treinado com o conjunto Treinamento e avaliado através do conjunto Teste, os resultados dessa avaliação são aqui denominados como Teste Hold-out.

Figura 28 – Categorização distância percorrida $\Delta dist$.



Teste Holdout Baseline	**Média**				**Melhor**			
	Prec	Rec	F1	Accur	Prec	Rec	F1	Accur
Decision Tree Classifier	0.56	0.56	0.55	0.57	0.63	0.64	0.61	0.61
Random Forest Classifier	0.61	0.60	0.58	0.60	0.69	0.67	0.61	0.61
XGBoost Classifier	0.56	0.56	0.55	0.57	0.58	0.59	0.57	0.57

Tabela 13 – Resultados dos modelos de classificação $\Delta dist$ *baseline*.

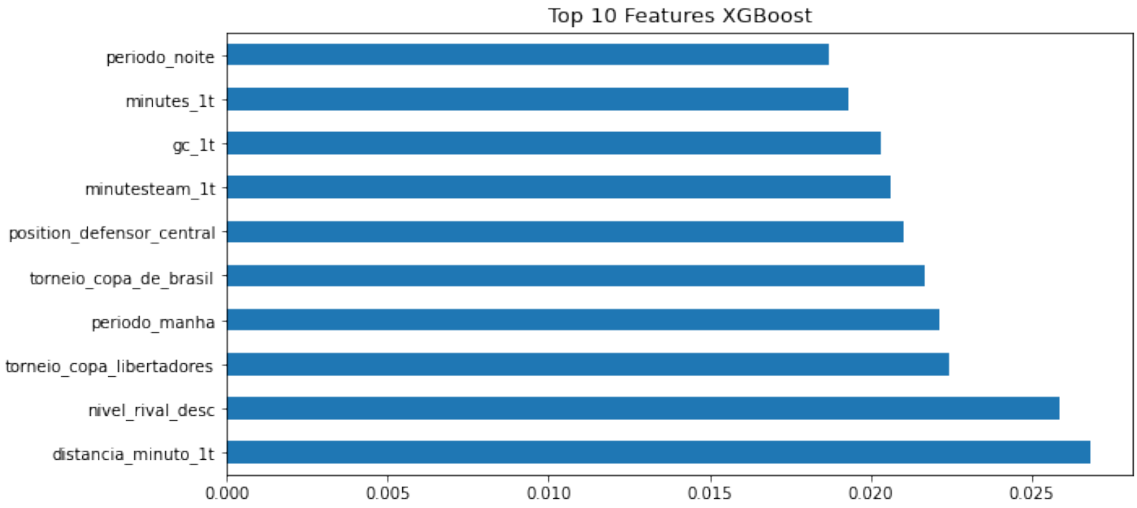
Teste Holdout Tunned	**Média**				**Melhor**			
	Prec	Rec	F1	Accur	Prec	Rec	F1	Accur
Decision Tree Classifier	0.61	0.61	0.61	0.61	0.70	0.71	0.69	0.69
Random Forest Classifier	0.65	0.64	0.62	0.62	0.72	0.67	0.59	0.60
XGBoost Classifier	0.62	0.62	0.62	0.63	0.66	0.67	0.65	0.65

Tabela 14 – Resultados dos modelos de classificação $\Delta dist$ *tunned*.

A Tabela 13 mostra os resultados dos métodos adotados para o Teste Holdout. Alguns *splits* tiveram resultados abaixo dos demais, isso é devido ao fato de alguns *aplits* considerarem períodos de descanso, férias, copa do mundo e outras paralisações. Assim, para todos resultados apresentados, são exibidos a média e o melhor resultado. Apesar de todos métodos apresentarem bons resultados, não foram satisfatórios os resultados do método XGBoost, que nesse momento era o pior dos 3 métodos testados. Dessa forma, foi realizado um *tunning* na base, selecionando alguns parâmetros de cada modelo, variando valores de parâmetros, e utilizando a ferramenta *GridSearch*. A Tabela 14 mostra os resultados dos métodos adotados para o Teste Holdout após o *tunning*. Desta vez o método XgBoost apresentou a melhor média considerando a acurácia, que ficou em 0.63 e bons resultados para outras métricas, com 0.62 para o *Recall*, *Precision e F1-Score*.

A partir do classificador XGBoost, que apresentou melhores resultados, são apresentadas na Figura 29 as variáveis mais significativas na geração dos resultados. Chama a atenção o quanto a posição do atleta em campo e período da atividade é um fator influente de diversas formas para a variação de desempenho. Além disso, o campeonato, rival e gols feitos no primeiro tempo também estão em destaque. Apesar de poder confirmar que esses são fatores de impacto, a forma como eles interagem com o resultado $\Delta dist$ pode ser das mais variadas formas e portanto algum tipo de explicabilidade também é necessária, visando buscar cada vez mais detalhes de como essas variáveis podem se relacionar ao *target*.

Figura 29 – Variáveis mais relevantes XGBoostClassifier.

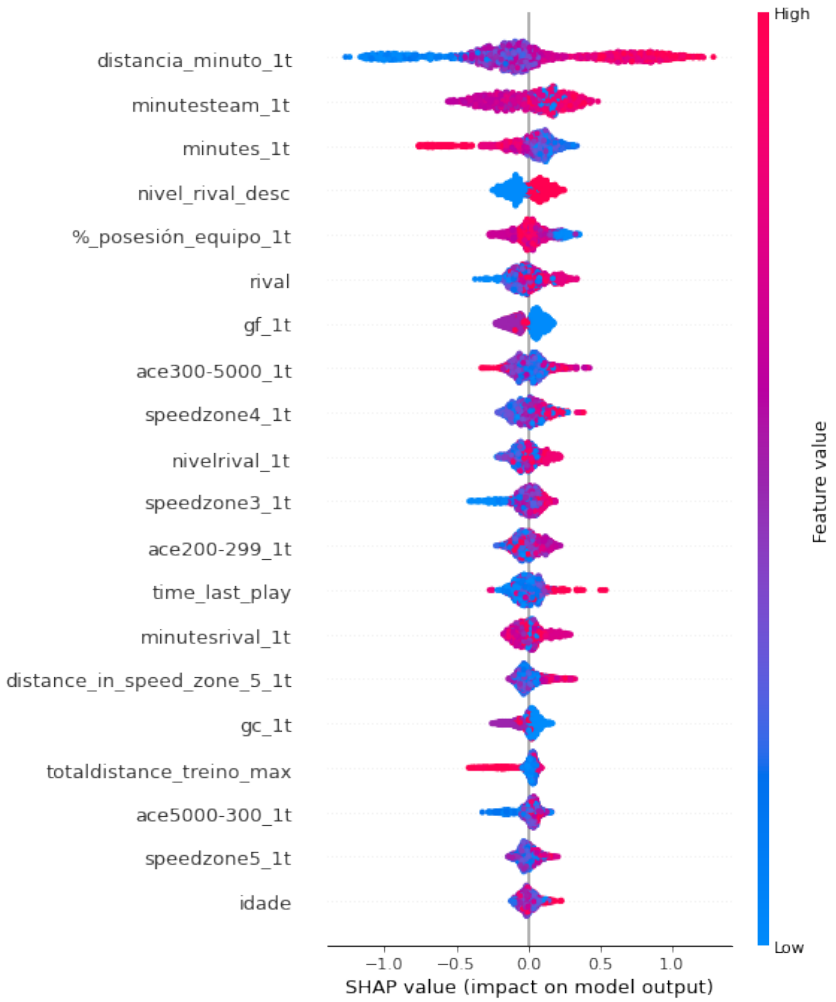


Para complementar a análise e explicar as decisões do modelo de uma forma mais intuitiva, foi utilizado o método SHAP. A Figura 30 mostra a influência das variáveis por classes específica, sendo que o eixo x mostra como cada variável influencia para o resultado estar em cada uma das duas classes. A cor dos pontos representam o aumento/diminuição do valor da variável. Tons vermelhos são valores mais altos e tons azulados são valores mais baixos.

A Figura 30 mostra que algumas variáveis relacionadas à distância percorrida no primeiro tempo, quando altas, contribuem para o jogador estar na classe onde se perde performance entre os tempos, o que é compreensível, visto que se o jogador percorre altas distâncias em alta velocidade, tende a cansar e render menos no segundo tempo. Uma variável muito interessante é a que indica os gols a favor no primeiro tempo, que quando mais altas, contribuem para o jogador estar na classe que indica que o jogador não irá perder performance, o que pode sugerir que a característica deste time não é diminuir o ritmo no segundo mesmo tendo feito gols. No nível do rival, é possível perceber que quando o adversário é melhor, os jogadores perdem performance para o segundo tempo. Outra variável que chama atenção é a distância percorrida nos treinos, que contribuem

para a classificação, indicando que o treino de fato contribui para o preparo do atleta e a manutenção de sua performance no segundo tempo.

Figura 30 – SHAP - Impacto nas Classes.



5.3 Considerações finais

O primeiro estudo de caso abordou as características de jogadores lesionados e aspectos de jogadores com baixa ou alta eficiência física. Para entender as lesões, os jogadores foram classificados de acordo com variáveis relacionadas à distâncias, percepções e resultados laboratoriais. Estes jogadores foram divididos em 5 grupos, sendo 3 deles com um número considerável de atletas. Ao verificar em quais grupos estariam os jogadores que tiveram lesões durante a temporada, eles foram distribuídos no segundo e terceiro grupo com mais atletas, e nenhum no grupo com maior número de atletas. Ao comparar os grupos que possuíam jogadores lesionados com os grupos que não possuíam, verificou-se que os lesionados possuíam maiores médias de eficiência física nos jogos, percorriam maiores distâncias, em alta intensidade ou não e número de desacelerações. Estas variáveis representam informações coletadas em jogos e treinos, e podem indicar

que as lesões podem ser resultado de uma carga excessiva de treinamento. Como a base de dados é relativamente pequena e não existem informações suficientes sobre a origem das lesões, não é possível fazer análises conclusivas, mas abrem perspectivas para novos estudos e análises.

A eficiência física do atleta nas atividades também foi analisada no primeiro estudo de caso, onde foi encontrada uma alta correlação com outras variáveis significativas. Neste estudo, os dois jogadores com as melhores eficiências físicas foram comparados com os dois jogadores com as piores eficiências físicas e características destes jogadores foram analisadas. Os jogadores com piores eficiências físicas tem uma variação negativa maior da percepção do estado físico do início para o fim do treino. Os jogadores com melhores eficiências possuem características de treino mais parecidas com as do jogo quando observadas as variáveis relacionadas à distância, tempo de atividade e medições cardíacas, enquanto os jogadores com piores eficiências possuem índices menores destas variáveis nos treinos e maiores nos jogos. Isto leva a conclusões que os treinos devem refletir ao máximo os números apresentados durante os jogos, para que o jogador consiga alcançar a maior eficiência física possível.

O segundo estudo de caso abordou características que influenciam nos resultados e desempenho dos atletas nos jogos. O resultado final de um jogo nem sempre reflete o desempenho da equipe na partida, sendo assim, foi criada, consultando o fisiologista do clube, uma variável que informava se o resultado foi positivo, negativo ou neutro, baseado em diversas variáveis. Após análises diversas foi observado que, dentre outras características, as variáveis relacionadas à distância tinha grande influência com o resultado da partida. A distância total percorrida não apresenta tanta relevância, pois o time pode ter resultados positivos ou negativos, seja em jogos mais táticos, onde os jogadores jogam mais fechado, ou não. Entretanto, um fator com alta influência no resultado é a diferença das variáveis relacionadas à distância entre o primeiro e o segundo tempo, principalmente quando estes valores diminuem, indicando que o jogador perdeu desempenho entre as duas etapas. Para entender esta variação no desempenho dos atletas foi criado um *target* $\Delta dist$, que indica o quanto o atleta perdeu de desempenho entre as duas etapas. Foram realizados experimentos com 3 classificadores, sendo que o XGBoost obteve os melhores resultados para prever esta variação do desempenho do atleta entre o primeiro e segundo tempo, com acurácia de 0.65, baseado em dados de treino, pessoais, do rival e do primeiro tempo. Os resultados obtidos permitem que a comissão técnica entenda quais jogadores tendem a perder desempenho e tomem decisões já no intervalo do jogo, substituindo o atleta ou mudando a estratégia do jogo. Também foram analisadas quais as variáveis tinham maior influência neste *target* e os resultados mostram que algumas posições tendem a ter maior variação do desempenho entre os tempos, o mesmo acontece com o campeonato que está sendo disputado, o que pode influenciar na carga de treinamento de ou definir melhor a característica física de jogadores para determinadas posições ou torneios. O tempo que

o jogador está sem jogar uma partida oficial e o nível do rival também influenciam nesta variação, o que contribui para decisões como não deixar jogadores que estão parados jogar toda a partida e na melhor escolha do time dependendo do rival. Variáveis sobre quantidade e distâncias percorridas em alta velocidade também tiveram grande influência, principalmente as do primeiro tempo, que também podem sugerir substituições, além das distâncias percorridas em treino que melhoram a manutenção da performance do atleta na segunda metade da partida, sugerindo adaptações de cargas de treinamento nos treinos para aqueles atletas que apresentam diminuição de performance.

6 Conclusão

Este trabalho apresentou estudos com aplicação de ciência de dados e inteligência artificial no futebol, apresentando características de grupos de jogadores lesionados, jogadores com melhores eficiências físicas, aspectos que influenciam em resultados de jogos e previsão da diferença de desempenho de atletas entre os dois tempos do jogo.

No primeiro estudo de caso, foi abordado as características de jogadores lesionados e aspectos de jogadores com baixa ou alta média de eficiência física. Para entender as lesões, os jogadores foram classificados de acordo com variáveis relacionadas à distâncias, percepções e resultados laboratoriais. Os grupos que possuíam jogadores lesionados possuem médias maiores para as percepções dos estado físico antes e após o jogo, distância total e em alta intensidade, além das desacelerações. Estas variáveis representam informações coletadas em jogos e treinos, e podem indicar que as lesões podem ser resultado de uma carga excessiva de treinamento e que a percepção passada pelo jogador pode não refletir a realidade.

Também no primeiro estudo de caso, foram analisados os atletas com melhores e piores médias de eficiência física em jogos e treinos. Os jogadores com melhores eficiências possuem características de treino mais parecidas com as do jogo quando observadas as variáveis relacionadas à distância, tempo de atividade e medições cardíacas, enquanto os jogadores com piores eficiências possuem índices menores destas variáveis nos treinos. Isto leva a conclusões que os treinos devem refletir ao máximo os números apresentados durante os jogos, para que o jogador consiga alcançar uma maior eficiência física.

O segundo estudo de caso abordou características que influenciam nos resultados e desempenho dos atletas nos jogos. Foi observado que a diferença das distâncias percorridas pelo atleta entre os dois tempos de jogo influência em jogos com resultado positivos, negativos ou neutro. Aplicando o método XGBoost para prever essa diferença entre as distâncias, baseado em dados históricos, e foram obtidos resultados satisfatórios para a classificação dos atletas, com 0.62 de acurácia. Também foram analisadas quais as variáveis tinham maior influência neste variação, sendo as mais relevantes as que tratavam das distâncias percorridas no primeiro tempo, mas também outras menos óbvias como a posição do atleta, o campeonato disputado, nível do rival, gols no primeiro tempo e distâncias percorridas em treinos.

Os resultados dos dois estudos de caso mostram que as distâncias percorridas em treinos e jogos é fator extremamente relevante para o desempenho dos atletas e os resultados dos jogos. As informações geradas nos estudos, formam uma base de conhecimento que pode auxiliar em diversas tomadas de decisões sobre fatores para

escalar os times de acordo com o rival, campeonato e dados de treino, além de sugerir possíveis atletas a serem substituídos. Também fornece informações para definir cargas de treinamento mais adequadas.

Também foram realizadas duas revisões sistemáticas completas sobre inteligência artificial aplicada ao esporte. A primeira revisão, abordando esportes em equipe, apresentou que nos 58 artigos selecionados pelos critérios estabelecidos, a amostra era formada majoritariamente por atletas do sexo masculino (97%) e os esportes onde houveram mais estudos foram futebol (26%), basquete (22%) e *handball* (10%). A maioria dos estudos tratavam de performance esportiva (74%), onde o método de IA mais utilizado foi redes neurais artificiais (RNA) (34%), seguida por árvore de decisão (23%) e em terceiro lugar processos de markov e máquina de vetores de suporte (SVM) (12% cada). Considerando os artigos que tratavam de risco de lesão (26%), o método de IA mais utilizado foi redes neurais artificias (47%), seguida por árvore de decisão (21%) e máquina de vetores de suporte (15%).

As análises realizadas nesta revisão sistemática mostraram que as técnicas ou métodos de IA para prever o risco de lesões e desempenho esportivo mais usados em esportes coletivos eram redes neurais artificiais, classificador por árvore de decisão. Os esportes coletivos com a maioria das aplicações de IA eram futebol, basquete, handebol e vôlei. O estado atual de desenvolvimento na área propõe um futuro promissor no que diz respeito ao uso de IA em esportes de equipe.

A revisão sistemática realizada com esportes individuais, que selecionou 91 artigos, apresentou uma amostra formada majoritariamente por atletas do sexo masculino (90%) e os esportes onde houveram mais estudos foram atletismo (22%), esqui (9%) e natação (8%). Assim como no trabalho de esportes coletivos, os métodos mais utilizados foram redes neurais artificiais e classificador por árvore de decisão. Outra característica em comum com o estudo de esportes coletivos foi a abordagem dos estudos, onde 94% tratava de performance e apenas 6% risco de lesão. As ferramentas mais utilizadas nos estudos não foram obvias, visto que Python ficou em terceiro lugar, atrás de Matlab e R. Outra característica que chamou atenção nos estudos foram as métricas de avaliação do modelo, onde vários estudos sequer apresentavam ou apenas apresentavam acurácia

As duas revisões sistemáticas mostraram características importantes acerca dos estudos em esportes utilizando inteligência artificial. A grande maioria dos estudos aborda performance esportiva e poucos lesão, também mostrou a preferência pelos métodos de redes neurais e árvore de decisão, e a necessidade de estudos envolvendo mulheres e esportes mais populares.

6.1 Trabalhos futuros

Para a continuidade dos estudos são propostas seguintes iniciativas:

- Novas análises sobre as diferenças de performance entre os tempos de jogos, acrescentando dados de laboratório;
- Coleta de dados utilizando outras fontes, como análises de imagens, para a geração de uma quantidade maior de dados;
- Coleta dos dados do primeiro tempo através do próprio sistema utilizado pelo clube e não por meio de planilhas, para que possibilite a entrega das análises ainda no intervalo;
- Elaboração de novos artigos para divulgação científica dos resultados obtidos no trabalho.

Referências

ABDULLAH, M.; MALIKI, A.; MUSA, R.; KOSNI, N.; JUAHIR, H. Intelligent prediction of soccer technical skill on youth soccer player's relative performance using multivariate analysis and artificial neural network techniques. **International Journal on Advanced Science, Engineering and Information Technology**, v. 6, n. 5, p. 668–674, 2016. Citado 2 vezes nas páginas 5 e 6.

ADETIBA, E.; IWEANYA, V. C.; POPOOLA, S. I.; ADETIBA, J. N.; MENON, C. Automated detection of heart defects in athletes based on electrocardiography and artificial neural network. **Cogent Engineering**, Taylor & Francis, v. 4, n. 1, p. 1411220, 2017. Citado na página 5.

BAKER, M. Data science: Industry allure. **Nature**, Nature Research, v. 520, n. 7546, p. 253–255, 2015. Citado na página 2.

BANGSBO, J.; NøRREGAARD, L.; THORSø, F. Activity profile of competition soccer. **Canadian journal of sport sciences = Journal canadien des sciences du sport**, v. 16, n. 2, p. 110—116, June 1991. ISSN 0833-1235. Citado na página 7.

BARTLETT, J. D.; O'CONNOR, F.; PITCHFORD, N.; TORRES-RONDA, L.; ROBERTSON, S. J. Relationships between internal and external training load in team-sport athletes: evidence for an individualized approach. **International journal of sports physiology and performance**, Human Kinetics, Inc., v. 12, n. 2, p. 230–234, 2017. Citado na página 6.

BIANCHI, F.; FACCHINETTI, T.; ZUCCOLOTTO, P. Role revolution: towards a new meaning of positions in basketball. **Electronic Journal of Applied Statistical Analysis**, v. 10, n. 3, p. 712–734, 2017. Citado na página 6.

BLEI, D. M.; SMYTH, P. Science and data science. **Proceedings of the National Academy of Sciences**, v. 114, n. 33, p. 8689–8692, Jul 2017. Citado na página 1.

BOCK, J. R. Pitch sequence complexity and long-term pitcher performance. **Sports**, Multidisciplinary Digital Publishing Institute, v. 3, n. 1, p. 40–55, 2015. Citado na página 7.

BOCK, J. R. Empirical prediction of turnovers in nfl football. **Sports**, Multidisciplinary Digital Publishing Institute, v. 5, n. 1, p. 1, 2017. Citado na página 7.

BORRESEN, J.; LAMBERT, M. I. The quantification of training load, the training response and the effect on performance. **Sports medicine**, Springer, v. 39, n. 9, p. 779–795, 2009. Citado na página 3.

BREFELD, U.; DAVIS, J.; HAAREN, J. V.; ZIMMERMANN, A. **Machine Learning and Data Mining for Sports Analytics: 5th International Workshop, MLSA 2018, Co-located with ECML/PKDD 2018, Dublin, Ireland, September 10, 2018, Proceedings**. Springer International Publishing, 2019. (Lecture Notes in Computer Science). ISBN 9783030172749. Disponível em: <https://books.google.com.br/books?id=0G2QDwAAQBAJ>. Citado na página 3.

BROOKS, R. A. Intelligence without representation. **Artificial intelligence**, Elsevier, v. 47, n. 1-3, p. 139–159, 1991. Citado na página 1.

CARPITA, M.; SANDRI, M.; SIMONETTO, A.; ZUCCOLOTTO, P. Discovering the drivers of football match outcomes with data mining. **Quality Technology & Quantitative Management**, Taylor & Francis, v. 12, n. 4, p. 561–577, 2015. Citado 2 vezes nas páginas 5 e 7.

ÇENE, E. What is the difference between a winning and a losing team: insights from euroleague basketball. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 18, n. 1, p. 55–68, 2018. Citado na página 7.

CLAUDINO, J.; FILHO, C. C.; BOULLOSA, D.; ALVES, A.; CARRION, G.; GIANONI, R.; GUIMARãES, R.; VENTURA, F.; ARAúJO, A.; ROSSO, S. D.; AFONSO, J.; SERRAO, J. The role of veracity on the load monitoring of professional soccer players: A systematic review in the face of the big data era. **Applied Sciences**, v. 11, p. Article 6479, 07 2021. Citado na página 1.

CLAUDINO, J. G.; CAPANEMA, D. de O.; SANTIAGO, P. R. P. AIM in sports medicine. In: . [S.l.]: Springer International Publishing, 2021. p. 1–6. Citado 2 vezes nas páginas 1 e 2.

CLAUDINO, J. G.; CAPANEMA, D. de O.; SOUZA, T. V. de; SERRÃO, J. C.; PEREIRA, A. C. M.; NASSIS, G. P. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. **Sports medicine-open**, Springer, v. 5, n. 1, p. 28, 2019. Citado 5 vezes nas páginas 8, 9, 10, 11 e 19.

COLÁS, R. M.; FÉLEZ, R. M.; TORRES-SOSPEDRA, J.; USÓ, A. M. Team activity recognition in association football using a bag-of-words-based method. Elsevier, 2015. Citado 2 vezes nas páginas 5 e 7.

CROFT, H.; LAMB, P.; MIDDLEMAS, S. The application of self-organising maps to performance analysis data in rugby union. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 15, n. 3, p. 1037–1046, 2015. Citado na página 6.

DALEN, T.; LORåS, H.; HJELDE, G. H.; KJøSNES, T. N.; WISLøFF, U. Accelerations – a new approach to quantify physical performance decline in male elite soccer? **European Journal of Sport Science**, Routledge, v. 19, n. 8, p. 1015–1023, 2019. Citado na página 7.

DEMŠAR, J.; CURK, T.; ERJAVEC, A.; GORUP, Č.; HOČEVAR, T.; MILUTINOVIČ, M.; MOŽINA, M.; POLAJNAR, M.; TOPLAK, M.; STARIČ, A. et al. Orange: data mining toolbox in python. **The Journal of Machine Learning Research**, JMLR. org, v. 14, n. 1, p. 2349–2353, 2013. Citado na página 20.

DüKING, P.; ACHTZEHN, S.; HOLMBERG, H.-C.; SPERLICH, B. Integrated framework of load monitoring by a combination of smartphone applications, wearables and point-of-care testing provides feedback that allows individual responsive adjustments to activities of daily living. **Sensors**, MDPI AG, v. 18, n. 5, p. 1632, maio 2018. Citado na página 2.

ERTELT, T.; SOLOMONOVS, I.; GRONWALD, T. Enhancement of force patterns classification based on gaussian distributions. **Journal of biomechanics**, Elsevier, v. 67, p. 144–149, 2018. Citado na página 5.

FUSTER-PARRA, P.; GARCÍA-MAS, A.; CANTALLOPS, J.; PONSETI, F. J.; LUO, Y. Ranking features on psychological dynamics of cooperative team work through bayesian networks. **Symmetry**, Multidisciplinary Digital Publishing Institute, v. 8, n. 5, p. 34, 2016. Citado 2 vezes nas páginas 5 e 6.

GE, S. Research on the factors of basketball injury in physical teaching based on artificial neural network. **Rev Facultad Ingen**, v. 32, n. 3, p. 415–422, 2017. Citado na página 5.

GOSWAMI, R.; DUFORT, P.; TARTAGLIA, M.; GREEN, R.; CRAWLEY, A.; TATOR, C.; WENNBERG, R.; MIKULIS, D.; KEIGHTLEY, M.; DAVIS, K. D. Frontotemporal correlates of impulsivity and machine learning in retired professional athletes with a history of multiple concussions. **Brain structure and function**, Springer, v. 221, n. 4, p. 1911–1925, 2016. Citado na página 6.

GU, W.; SAATY, T. L.; WHITAKER, R. Expert system for ice hockey game prediction: data mining with human judgment. **International Journal of Information Technology & Decision Making**, World Scientific, v. 15, n. 04, p. 763–789, 2016. Citado na página 7.

HASSAN, A.; SCHRAPF, N.; RAMADAN, W.; TILP, M. Evaluation of tactical training in team handball by means of artificial neural networks. **Journal of sports sciences**, Routledge, v. 35, n. 7, p. 642–647, 2017. Citado na página 6.

HASSAN, A.; SCHRAPF, N.; TILP, M. The prediction of action positions in team handball by non-linear hybrid neural networks. **International Journal of Performance Analysis in Sport**, Routledge, v. 17, n. 3, p. 293–302, 2017. Citado na página 1.

HASSAN, A.; SCHRAPF, N.; TILP, M. The prediction of action positions in team handball by non-linear hybrid neural networks. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 17, n. 3, p. 293–302, 2017. Citado na página 6.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. Springer, 2009. Disponível em: <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>. Citado na página 20.

HEALEY, G. Learning, visualizing, and assessing a model for the intrinsic value of a batted ball. **IEEE Access**, IEEE, v. 5, p. 13811–13822, 2017. Citado na página 6.

HOCH, T.; TAN, X.; LESER, R.; BACA, A.; MOSER, B. A knowledge discovery framework for the assessment of tactical behaviour in soccer based on spatiotemporal data. **Mathematical and Computer Modelling of Dynamical Systems**, Taylor & Francis, v. 23, n. 4, p. 384–398, 2017. Citado 2 vezes nas páginas 5 e 7.

ISRANI, S. T.; VERGHESE, A. Humanizing artificial intelligence. **Jama**, American Medical Association, v. 321, n. 1, p. 29–30, 2019. Citado na página 2.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: <https://faculty.marshall.usc.edu/gareth-james/ISL/>. Citado na página 19.

JASPERS, A.; BEÉCK, T. O. D.; BRINK, M. S.; FRENCKEN, W. G.; STAES, F.; DAVIS, J. J.; HELSEN, W. F. Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? **International journal of sports physiology and performance**, Human Kinetics, v. 13, n. 5, p. 625–630, 2018. Citado 2 vezes nas páginas 5 e 6.

JELINEK, H. F.; KELAREV, A.; ROBINSON, D. J.; STRANIERI, A.; CORNFORTH, D. J. Using meta-regression data mining to improve predictions of performance based on heart rate dynamics for australian football. **Applied Soft Computing**, Elsevier, v. 14, p. 81–87, 2014. Citado na página 7.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 2.

KAUTZ, T.; GROH, B. H.; HANNINK, J.; JENSEN, U.; STRUBBERG, H.; ESKOFIER, B. M. Activity recognition in beach volleyball using a deep convolutional neural network. **Data Mining and Knowledge Discovery**, Springer, v. 31, n. 6, p. 1678–1705, 2017. Citado na página 5.

KEMPE, M.; GRUNZ, A.; MEMMERT, D. Detecting tactical patterns in basketball: comparison of merge self-organising maps and dynamic controlled neural networks. **European journal of sport science**, Taylor & Francis, v. 15, n. 4, p. 249–255, 2015. Citado na página 6.

KOLBUSH, J.; SOKOL, J. A logistic regression/markov chain model for american college football. **International Journal of Computer Science in Sport**, Sciendo, v. 16, n. 3, p. 185–196, 2017. Citado na página 7.

LAPHAM, A.; BARTLETT, R. The use of artificial intelligence in the analysis of sports performance: A review of applications in human gait analysis and future directions for sports biomechanics. **Journal of Sports Sciences**, Routledge, v. 13, n. 3, p. 229–237, 1995. Citado na página 1.

LAPHAM, A.; BARTLETT, R. The use of artificial intelligence in the analysis of sports performance: A review of applications in human gait analysis and future directions for sports biomechanics. **Journal of Sports Sciences**, Taylor & Francis Group, v. 13, n. 3, p. 229–237, 1995. Citado na página 5.

LEICHT, A. S.; GÓMEZ, M. A.; WOODS, C. T. Explaining match outcome during the men's basketball tournament at the olympic games. **Journal of sports science & medicine**, Dept. of Sports Medicine, Medical Faculty of Uludag University, v. 16, n. 4, p. 468, 2017. Citado na página 7.

LEICHT, A. S.; GOMEZ, M. A.; WOODS, C. T. Team performance indicators explain outcome during women's basketball matches at the olympic games. **Sports**, Multidisciplinary Digital Publishing Institute, v. 5, n. 4, p. 96, 2017. Citado na página 7.

LI, C. Predict the neural network mathematical model of basketball team scores based on improved bp algorithm. **BioTechnol Indian J**, v. 8, n. 5, p. 628–633, 2013. Citado na página 6.

LINK, D.; HOERNIG, M. Individual ball possession in soccer. **PloS one**, Public Library of Science, v. 12, n. 7, 2017. Citado 2 vezes nas páginas 5 e 6.

LÓPEZ-VALENCIANO, A.; AYALA, F.; PUERTA, J. M.; CROIX, M. D. S.; VERA-GARCÍA, F.; HERNÁNDEZ-SÁNCHEZ, S.; RUIZ-PÉREZ, I.; MYER, G. A preventive model for muscle injuries: A novel approach based on learning algorithms. **Medicine and science in sports and exercise**, NIH Public Access, v. 50, n. 5, p. 915, 2018. Citado 3 vezes nas páginas 3, 5 e 6.

LU, G. Evaluation model of young basketball players' physical quality and basic technique based on rbf neural network. **BioTechnol Indian J**, v. 8, n. 9, p. 1193–1198, 2013. Citado na página 7.

LU, H.; LI, Y.; CHEN, M.; KIM, H.; SERIKAWA, S. Brain intelligence: go beyond artificial intelligence. **Mobile Networks and Applications**, Springer, v. 23, n. 2, p. 368–375, 2018. Citado na página 1.

MCCALL, A.; DAVISON, M.; CARLING, C.; BUCKTHORPE, M.; COUTTS, A. J.; DUPONT, G. **Can off-field 'brains' provide a competitive advantage in professional football?** [S.l.]: BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine, 2016. Citado 2 vezes nas páginas 2 e 5.

MCKINNEY, W. **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython**. [S.l.]: "O'Reilly Media, Inc.", 2012. Citado na página 20.

MOHR, M.; KRUSTRUP, P.; BANGSBO, J. Match performance of high-standard soccer players with special reference to development of fatigue. **Journal of sports sciences**, v. 21, p. 519–28, 08 2003. Citado na página 7.

NOVATCHKOV, H.; BACA, A. Artificial intelligence in sports on the example of weight training. **Journal of sports science & medicine**, Dept. of Sports Medicine, Medical Faculty of Uludag University, v. 12, n. 1, p. 27, 2013. Citado na página 3.

NæS, T.; MEVIK, B.-H. Understanding the collinearity problem in regression and discriminant analysis. **Journal of Chemometrics**, v. 15, p. 413 – 426, 05 2001. Citado na página 17.

OLIPHANT, T. E. Python for scientific computing. **Computing in Science & Engineering**, IEEE, v. 9, n. 3, p. 10–20, 2007. Citado na página 20.

PAI, P.-F.; CHANGLIAO, L.-H.; LIN, K.-P. Analyzing basketball games by a support vector machines with decision tree model. **Neural Computing and Applications**, Springer, v. 28, n. 12, p. 4159–4167, 2017. Citado na página 7.

PASSFIELD, L.; HOPKER, J. G. A mine of information: can sports analytics provide wisdom from your data? **International journal of sports physiology and performance**, Human Kinetics, Inc., v. 12, n. 7, p. 851–855, 2017. Citado na página 1.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 20.

PENSGAARD, A. M.; IVARSSON, A.; NILSTAD, A.; SOLSTAD, B. E.; STEFFEN, K. Psychosocial stress factors, including the relationship with the coach, and their influence on acute and overuse injury risk in elite female football players. **BMJ open sport & exercise medicine**, BMJ Specialist Journals, v. 4, n. 1, p. e000317, 2018. Citado 2 vezes nas páginas 5 e 6.

PERME, M.; BLAS, M.; TURK, S. Comparison of logistic regression and linear discriminant analysis: A simulation study. **Metodološki Zvezki**, v. 1, p. 143–161, 01 2004. Citado na página 18.

PEñAS, C.; CASAIS, L.; DOMINGUEZ, E.; SAMPAIO, J. The effects of situational variables on distance covered at various speeds. **European Journal of Sport Science**, v. 10, p. 103–109, 03 2010. Citado na página 7.

PHATAK, A.; MEMMERT, D. Artificial intelligence based body sensor network framework—narrative review: Proposing an end-to-end framework using wearable sensors, real-time location systems and artificial intelligence/machine learning algorithms for data collection, data mining and knowledge discovery in sports and healthcare. **Sports Medicine - Open**, v. 7, 12 2021. Citado 2 vezes nas páginas 1 e 2.

PODHURSKYI, S. E. Performance of striking techniques among qualified muay thai athletes of different weight classes. **International Journal of Performance Analysis in Sport**, Routledge, v. 20, n. 2, p. 294–304, 2020. Citado na página 1.

PRESS, S. J.; WILSON, S. Choosing between logistic regression and discriminant analysis. **Journal of the American Statistical Association**, Taylor & Francis, v. 73, n. 364, p. 699–705, 1978. Citado na página 18.

QILIN, S.; XIAOMEI, W.; XIAOLING, F.; YUANPING, C.; SHAOYONG, W. Study on knee joint injury in college football training based on artificial neural network. **RISTI (Revista Iberica de Sistemas e Tecnologias de Informacao)**, AISTI (Iberian Association for Information Systems and Technologies), n. E10, p. 197–211, 2016. Citado na página 5.

REILLY, T. A motion analysis of work-rate in different positional roles in professional football match-play. In: . [S.l.: s.n.], 1976. Citado na página 7.

REIN, R.; MEMMERT, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. **SpringerPlus**, SpringerOpen, v. 5, n. 1, p. 1410, 2016. Citado na página 1.

RIENZI, E.; DRUST, B.; REILLY, T.; CARTER, J.; MARTIN, A. Investigation of anthropometric and work-rate profiles of elite south american international soccer players. **The Journal of sports medicine and physical fitness**, v. 40, n. 2, p. 162—169, June 2000. ISSN 0022-4707. Citado na página 7.

ROBERTSON, S.; BACK, N.; BARTLETT, J. D. Explaining match outcome in elite australian rules football using team performance indicators. **Journal of sports sciences**, Routledge, v. 34, n. 7, p. 637–644, 2016. Citado na página 7.

ROSSI, A.; PAPPALARDO, L.; CINTIA, P.; IAIA, F. M.; FERNÁNDEZ, J.; MEDINA, D. Effective injury forecasting in soccer with gps training data and machine learning. **PloS one**, Public Library of Science, v. 13, n. 7, 2018. Citado na página 5.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Malaysia; Pearson Education Limited„ 2016. Citado 2 vezes nas páginas 1 e 17.

SALVO, V. D.; BARON, R.; TSCHAN, H.; MONTERO, F.; BACHL, N.; PIGOZZI, F. Performance characteristics according to playing position in elite soccer. **International journal of sports medicine**, v. 28, p. 222–7, 03 2007. Citado na página 7.

SANKARAN, S. Comparing pay versus performance of ipl bowlers: an application of cluster analysis. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 14, n. 1, p. 174–187, 2014. Citado na página 7.

SCHRAPF, N.; ALSAIED, S.; TILP, M. Tactical interaction of offensive and defensive teams in team handball analysed by artificial neural networks. **Mathematical and Computer Modelling of Dynamical Systems**, Taylor & Francis, v. 23, n. 4, p. 363–371, 2017. Citado na página 6.

SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A systematic literature review on applying crisp-dm process model. **Procedia Computer Science**, Elsevier, v. 181, p. 526–534, 2021. Citado na página 20.

SCHULTE, O.; KHADEMI, M.; GHOLAMI, S.; ZHAO, Z.; JAVAN, M.; DESAULNIERS, P. A markov game model for valuing actions, locations, and team performance in ice hockey. **Data Mining and Knowledge Discovery**, Springer, v. 31, n. 6, p. 1735–1757, 2017. Citado na página 7.

SHORTLIFFE, E. H.; SEPÚLVEDA, M. J. Clinical decision support in the era of artificial intelligence. **Jama**, American Medical Association, v. 320, n. 21, p. 2199–2200, 2018. Citado na página 2.

STRNAD, D.; NERAT, A.; KOHEK, Š. Neural network models for group behavior prediction: a case of soccer match attendance. **Neural Computing and Applications**, Springer, v. 28, n. 2, p. 287–300, 2017. Citado 2 vezes nas páginas 5 e 6.

THORNTON, H. R.; DELANEY, J. A.; DUTHIE, G. M.; DASCOMBE, B. J. Importance of various training-load measures in injury incidence of professional rugby league athletes. **International journal of sports physiology and performance**, Human Kinetics, Inc., v. 12, n. 6, p. 819–824, 2017. Citado na página 6.

TILP, M.; SCHRAPF, N. Analysis of tactical defensive behavior in team handball by means of artificial neural networks. **IFAC-PapersOnLine**, Elsevier, v. 28, n. 1, p. 784–785, 2015. Citado na página 6.

TÜMER, A. E.; KOÇER, S. Prediction of team league's rankings in volleyball by artificial neural network method. **International Journal of Performance Analysis in Sport**, Taylor & Francis, v. 17, n. 3, p. 202–211, 2017. Citado na página 6.

TUMILTY, D. Physiological characteristics of elite soccer players. **Sports Medicine**, v. 16, p. 80–96, 1993. Citado na página 7.

VALERO, C. S. Predicting win-loss outcomes in mlb regular season games–a comparative study using data mining methods. **International Journal of Computer Science in Sport**, De Gruyter Open, v. 15, n. 2, p. 91–112, 2016. Citado na página 7.

WANG, M. Evaluating technical and tactical abilities of football teams in euro 2012 based on improved information entropy model and som neural networks. **International Journal of Multimedia and Ubiquitous Engineering**, v. 9, n. 11, p. 293–302, 2014. Citado 2 vezes nas páginas 5 e 7.

WANG, Y.; ZHAO, Y.; CHAN, R. H.; LI, W. J. Volleyball skill assessment using a single wearable micro inertial measurement unit at wrist. **IEEE Access**, IEEE, v. 6, p. 13758–13765, 2018. Citado na página 7.

WHITESIDE, D.; MARTINI, D. N.; LEPLEY, A. S.; ZERNICKE, R. F.; GOULET, G. C. Predictors of ulnar collateral ligament reconstruction in major league baseball pitchers. **The American journal of sports medicine**, SAGE Publications Sage CA: Los Angeles, CA, v. 44, n. 9, p. 2202–2209, 2016. Citado na página 6.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: MANCHESTER. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [S.l.], 2000. v. 1, p. 29–39. Citado 2 vezes nas páginas 20 e 21.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016. Citado na página 1.

WORTH, A.; CRONIN, M. The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. **Journal of Molecular Structure: THEOCHEM**, v. 622, p. 97–111, 03 2003. Citado na página 18.

WU, L. The participating team's technical analysis of women's basketball in the 30th olympic games based on neural network. **J Chem Pharma Res**, v. 5, n. 11, p. 152–158, 2013. Citado na página 6.

WU, L. C.; KUO, C.; LOZA, J.; KURT, M.; LAKSARI, K.; YANEZ, L. Z.; SENIF, D.; ANDERSON, S. C.; MILLER, L. E.; URBAN, J. E. et al. Detection of american football head impacts using biomechanical features and support vector machine classification. **Scientific reports**, Nature Publishing Group, v. 8, n. 1, p. 1–14, 2017. Citado na página 6.

ZELIČ, I.; KONONENKO, I.; LAVRAČ, N.; VUGA, V. Induction of decision trees and bayesian classification applied to diagnosis of sport injuries. **Journal of Medical Systems**, Springer, v. 21, n. 6, p. 429–444, 1997. Citado na página 5.