INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Evolução dos jogadores jovens e o seu valor de mercado: O estudo da liga de futebol “Championship” na época 22/23

Pedro Miguel Simão da Silva

Master in Data Science

Supervisors:
Doctor *Catarina Maria Valente Antunes Marques*, *Associate* Professor,
Iscte – Instituto Universitário de Lisboa, Iscte Business School

Doctor *Rui Jorge Henriques Calado Lopes*, *Associate* Professor,
Iscte – Instituto Universitário de Lisboa, School of Technology and Architecture

September, 2024

BUSINESS SCHOOL

TECHNOLOGY AND ARCHITECTURE

Departamento de Métodos Quantitativos para Gestão e Economia
Departamento de Ciências e Tecnologia da Informação

Evolução dos jogadores jovens e o seu valor de mercado: O estudo da liga de futebol "Championship" na época 22/23

Pedro Miguel Simão da Silva

Mestrado em Ciências de Dados

Orientadores:
Doutora *Catarina Maria Valente Antunes Marques*, Professora Associada,
Iscte – Instituto Universitário de Lisboa, Iscte Business School

Doutor *Rui Jorge Henriques Calado Lopes*, Professor Associado,
Iscte – Instituto Universitário de Lisboa, School of Technology and Architecture

September, 2024

Agradecimemtos

Em primeiro lugar, gostaria de agradecer à escola Iscte Business School (IBS), e aos meus orientadores, os professores Doutores Catarina Marques e Rui Lopes, que me orientaram ao longo da dissertação, demonstrando constante disponibilidade para esclarecer as minhas dúvidas e fornecer opiniões e críticas essenciais para o progresso do trabalho.

Um agradecimento aos amigos que tive o privilégio de conhecer e com quem partilhei esta jornada académica. O vosso apoio e as experiências que partilhámos tornaram este percurso mais enriquecedor.

Agradeço a toda a minha família, em especial aos meus pais, pelo apoio, companheirismo e ajuda ao longo deste percurso académico.

Resumo

A presente dissertação tem como foco a análise da evolução dos jogadores jovens que participaram na segunda divisão inglesa de futebol, mais propriamente na *Championship*, na época 2022/2023, e apresenta como objetivos a análise da progressão na carreira e de valor de mercado.

Neste sentido esta investigação, foi dividida em dois estudos distintos baseados na metodologia CRISP-DM. O primeiro estudo concentra-se na progressão dos jogadores jovens, de modo a identificar os fatores que influenciam essa progressão ao longo das épocas. Para além disso, este estudo propõe um sistema de classificação da progressão dos jogadores numa escala de cinco níveis, desde "Progrediu Significativamente" a "Regrediu", no qual são baseados os modelos de classificação. O segundo estudo analisa o valor de mercado dos jogadores ao longo de diferentes épocas, utilizando séries temporais, e modelos de aprendizagem automática e investiga os fatores que o influenciam, com ênfase na previsão do valor de mercado no futuro.

Em cada estudo, foram usados e comparados vários modelos de aprendizagem automática, tendo sido escolhidos por apresentarem melhor desempenho, os modelos de Redes Neuronais Recorrentes (RNN) no estudo 1 e as Redes Neuronais e *Random Forest* no estudo 2. Ainda no estudo 2, na análise de evolução do valor de mercado, o modelo RNN mostrou-se o mais eficaz.

Os resultados obtidos revelaram padrões complexos na progressão dos jogadores ao longo das épocas, demonstrando que o desenvolvimento de um jovem jogador de futebol é um processo não linear e muitas vezes imprevisível. O sistema de classificação desenvolvido mostrou-se eficaz para avaliar a evolução dos atletas de forma categorizada. No que diz respeito ao valor de mercado, foram identificadas algumas tendências ao longo das épocas, destacando-se a variação de critérios de valorização entre avançados, médios e defesas.

Os resultados deste estudo oferecem informações importantes para clubes, jogadores e agentes, permitindo uma compreensão mais clara dos fatores que influenciam o desenvolvimento e valorização dos jogadores jovens. Além disso, apontam para a importância de otimizar estratégias de desenvolvimento de talentos e políticas de transferências no contexto do futebol profissional.

Palavras Chave: *Futebol*, *Jogadores jovens*, *Progressão*, *Championship*, *Valor de mercado*, *Aprendizagem automática*,

Abstract

This dissertation focuses on analysing the development of young players who played in the English second division, more specifically in Championship, during the 2022/2023 season. The main objective is to analyse the progression of their careers and their market value.

The research was divided into two separate studies, based on the CRISP-DM methodology. The first study focuses on player progression, identifying the factors that influence this evolution over the seasons. To this end, a classification system was proposed for player progression, categorised into five levels, from "Significant Progress" to "Regression", on which the classification models were based.

The second study analyses the market value of players over the seasons, using time series and machine learning models, with a focus on predicting future market value.

In both studies, various machine learning models were used and compared. In the first study, the Recurrent Neural Networks (RNN) model performed best, while in the second, Neural Networks and Random Forest proved to be more effective. However, in predicting the evolution of market value, the RNN model stood out as the most accurate.

The results revealed complex patterns in the players progression over the seasons, confirming that the development of a young player is a non-linear and sometimes unpredictable process. The classification system proved to be effective in assessing players progress. With regard to market value, different trends were identified, especially between forwards, midfielders and defenders.

The conclusions of this study provide valuable information for clubs, players and agents, allowing for a better understanding of the factors that influence the development and valuation of young players. In addition, the results underline the importance of optimising talent development strategies and transfer policies in professional football.

KEYWORDS: *Football*, *Young players*, *Progression*, *Championship*, *Market value*, *Machine Learning*

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AdaBoost: *Adaptive Boosting*

AI: *Artificial Intelligence*

API: *Application Programming Interface*

CRISP-DM: *Cross-industry standard process for data mining*

GPS: *Global Positioning System*

GRU: *Gated Recurrent Unit*

HTML: *HyperText Markup Language*

HTTP: *Hypertext Transfer Protocol*

JSON: *JavaScript Object Notation*

LSTM: *Long Short-Term Memory*

MAE: *Mean Absolute Error*

MAPE: *Mean Absolute Percentage Error*

RMSE: *Root Mean Square Error*

RNN: *Recurrent Neural Network*

SVM: *Support Vector Machine*

URL: *Uniform Resource Allocator*

CAPÍTULO 1

Introdução

1.1. Enquadramento e Objetivos

A transição de jogadores jovens para o escalão sénior é um momento crucial no desenvolvimento desportivo. De acordo com Drew (2020), durante essa fase, os atletas enfrentam desafios únicos que podem influenciar tanto o seu desempenho profissional, como a própria carreira a longo prazo. Compreender os fatores que influenciam essa transição e desenvolver intervenções eficazes para apoiar os jogadores jovens nesse processo, é fundamental para maximizar o seu potencial e promover o sucesso no escalão sénior.

Vários sistemas desportivos nacionais, ao redor do mundo, têm implementado Programas de Promoção de Talentos (PPTs) em diferentes níveis, abrangendo seleções juniores das federações desportivas e centros de formação de desporto juvenil. De acordo com Güllich e Barth (2024), estes programas visam identificar e selecionar jovens atletas de alto desempenho em idades precoces, muitas vezes durante ou antes da puberdade, com o objetivo de acelerar o seu desenvolvimento desportivo. Durante a fase júnior, fatores como maturidade, idade relativa em relação ao ano de nascimento e a prática desportiva intensificada na infância e adolescência podem impulsionar um rápido desempenho júnior. Entretanto, esses efeitos tendem a diminuir ou até mesmo se inverter na fase adulta. Além disso, os participantes dos PPTs enfrentam custos e riscos adicionais, como gestão de tempo, lesões e esgotamento, que podem prejudicar o seu desenvolvimento a longo prazo e até mesmo encerrar prematuramente as suas carreiras.

No futebol profissional inglês, há preocupações em relação à falta de jogadores que conseguem fazer a transição do escalão jovem para o profissional (Anderson e Miller, 2011). Apenas cerca de 10% dos jogadores jovens que iniciam uma carreira profissional conseguem atingir esse estatuto.

Mitchell et al. (2020) destacam que a falta de suporte adequado durante essa transição é um dos principais obstáculos, argumentando que, uma vez que os jogadores obtêm um contrato profissional, frequentemente enfrentam uma lacuna no *coaching* e no desenvolvimento, o que pode dificultar a sua adaptação ao nível exigido no futebol profissional. Além disso, a pressão para ter um desempenho imediato e a competição com jogadores mais experientes podem levar a um aumento de ansiedade, impactando negativamente o desempenho dos jovens atletas.

Paralelamente, o mercado de transferências no futebol apresenta dinâmicas complexas que merecem atenção. De acordo com Borland e Lye (1996), o mercado de transferências no futebol é influenciado por diversos fatores, incluindo a procura por jogadores em posições específicas e a intensa competição entre clubes. Esta dinâmica pode resultar numa

"corrida de licitações" por talentos, onde clubes que procuram reforçar as suas equipas oferecem salários mais elevados ou melhores condições para atrair jogadores. Esta competição não só eleva os preços dos jogadores, como também influencia diretamente o seu valor de mercado, refletindo a importância das estratégias de recrutamento e a gestão de talentos dentro do contexto desportivo.

O objetivo da dissertação é estudar a evolução dos jogadores jovens que participaram na segunda divisão inglesa de futebol, a denominada "*Championship*", na época 2022/2023 segundo a progressão na carreira e o valor de mercado. Neste sentido, pretende-se analisar a progressão dos jogadores jovens e o seu valor de mercado desde o início da carreira até à época atual (23/24) e identificar os fatores que influenciam a progressão (mais concretamente os fatores que levam à progressão dos jogadores, à regressão ou à manutenção na mesma posição), assim como os fatores que influenciam o valor de mercado. Recorre-se a técnicas de aprendizagem automática, mais especificamente técnicas de classificação, de regressão e de séries temporais, para realização das análises.

A dissertação está dividida em dois estudos para dar resposta às questões de investigação levantadas para cada um destes objetivos. No Estudo 1 propõe-se analisar a progressão na carreira dos profissionais do futebol, e perceber quais os fatores que influenciam a progressão na carreira com base em estatísticas de jogos por épocas. Este estudo possui as seguintes questões de investigação:

QI1: Como se caracteriza a evolução dos jogadores ao longo das diferentes épocas?

QI2: Como classificar a progressão de um jogador de época para época? Será desenvolvido um sistema de classificação que em cada época classifique um jogador em categorias diferentes (como progrediu, regrediu ou manteve-se).

QI3: Quais são os fatores que influenciam a progressão na carreira, regressão ou manutenção na mesma categoria?

O Estudo 2 tem como objetivo a análise do valor de mercado ao longo das diferentes épocas, mais especificamente, no Estudo 2 propõe-se investigar os fatores específicos que influenciam o valor de mercado dos jogadores, com particular atenção às diferentes posições de jogo. Esta análise visa proporcionar uma compreensão mais profunda das dinâmicas que regem a valorização dos atletas no mercado de transferências. Neste sentido, o Estudo 2 pretende dar resposta às seguintes questões de investigação:

QI4: Como se caracteriza o valor de mercado nos jogadores da "*Championship* 22/23" nas diferentes épocas?

QI5: Quais são os fatores que influenciam o valor de mercado (por posição de jogo)?

QI6: Como evoluiu o valor de mercado (desde 2017) e qual a previsão do valor de mercado de um jogador para o próximo semestre?

1.2. Motivação

Desde pequeno que tenho uma grande paixão pelo desporto, especialmente pelo futebol. Este gosto, que me acompanha há muito tempo, juntamente com o meu interesse crescente por tecnologias modernas, levou-me a escolher este tema para o meu projeto de

dissertação. A ideia de usar técnicas avançadas de *machine learning* no futebol pareceu-me uma excelente oportunidade para contribuir para o progresso da análise desportiva.

Acredito que juntar o conhecimento tradicional do futebol com as possibilidades da inteligência artificial pode revelar padrões e informações que ainda não foram descobertas. Esta nova abordagem não só pode melhorar a nossa compreensão sobre como os jovens talentos se desenvolvem, mas também fornecer instrumentos importantes para tomar decisões na gestão do desporto.

Este estudo pode esclarecer questões importantes sobre o progresso das carreiras no futebol e o que influencia o valor de mercado dos jogadores profissionais. Além disso, a oportunidade de trabalhar com dados do campeonato *Championship*, uma liga inglesa conhecida pela sua competitividade e por ser um bom local para o desenvolvimento de jovens talentos, aumenta o possível impacto deste estudo.

1.3. Estrutura do documento

A dissertação está estruturada em cinco capítulos, cuidadosamente organizados para garantir uma explicação clara e organizada de todo o processo. O próximo capítulo, Revisão de Literatura, estabelece uma base teórica que abrange o estado atual do conhecimento na área científica abordada, incluindo as áreas de aplicação e técnicas utilizadas. Em seguida, o capítulo Metodologia detalha o procedimento adotado no estudo, bem como as diversas experiências realizadas, destacando a análise dos dados em estudo. O capítulo Resultados e Discussão complementa a Metodologia, apresentando de forma crítica os resultados das experiências e a sua contribuição para o desenvolvimento da dissertação. Finalmente, no capítulo Conclusão e Recomendações, é realizada uma revisão do processo, avaliando a importância, relevância e utilidade dos resultados obtidos, além de discutir as dificuldades enfrentadas e as possíveis direções para futuras investigações.

CAPÍTULO 2

Revisão de literatura

Este capítulo apresenta uma revisão abrangente da literatura relacionada com os principais temas que fundamentam esta investigação. Em primeiro lugar, é explorada a transição dos atletas jovens para o futebol sénior, analisando os diversos fatores que influenciam este processo. De seguida, aborda-se o valor de mercado dos jogadores de futebol e os diferentes elementos que contribuem para a sua valorização. O capítulo prossegue com uma análise dos sistemas de classificação no contexto desportivo dos clubes, destacando diferentes metodologias de pontuação. Por fim, examina-se o papel crescente da tecnologia no desporto, com particular incidência no futebol, explorando a relevância dos dados e as aplicações de aprendizagem automática na análise de desempenho dos jogadores.

2.1. Transição de atletas jovens para o escalão de futebol sénior

O futebol é um dos desportos mais populares do mundo, contando com milhões de jogadores jovens registados. No entanto, apenas uma pequena percentagem destes jogadores (aproximadamente 0,5%) consegue atingir o estatuto profissional, assinando contratos com clubes de futebol de elite. De acordo com Morris et al. (2017), a transição dos jogadores jovens para o escalão sénior é um processo exigente, tanto para os jogadores como para os clubes, sobretudo ao nível das principais ligas. Neste contexto, os centros de formação de futebol desempenham um papel crucial na seleção e desenvolvimento de jovens talentos, com o objetivo primordial de fornecer jogadores de alta qualidade às equipas seniores.

Nos últimos anos, de acordo com os autores Carpels et al. (2021), os desportos coletivos possuem parâmetros mais complexos para definir o talento. Os atletas de sucesso possuem habilidades técnicas avançadas, consciência tática e apropriada capacidade física para o desporto praticado. Além disso, a seleção de jogadores também é influenciada por enviesamentos cognitivos exibidos por agentes sociais, como treinadores, que estão envolvidos no processo de seleção, o que leva a uma potencial falta de fiabilidade das decisões de seleção.

Embora as vantagens iniciais experimentadas por atletas mais velhos e mais maduros pareçam relativamente pequenas, podem surgir dois processos auto-sustentáveis que resultam em diferenças adicionais de desempenho (Helsen et al., 2005). Primeiro, o sucesso no desporto geralmente é acompanhado pelo *feedback* positivo de colegas, treinadores e pais, o que aumenta a motivação e autoconfiança do atleta e resulta em maior esforço e melhor desempenho. Segundo, o treino adicional com treinadores de elite com um grupo

de jogadores selecionados ajuda um jovem talento a atingir um nível de desempenho mais alto.

Jogadores mais velhos têm uma experiência maior do que os companheiros de equipa mais jovens, o que proporciona vantagens significativas em termos de desempenho. Sendo cronologicamente mais velhos, eles têm uma vantagem em relação aos que são jovens (Cobley et al., 2009). Vários estudos revelaram distribuições de datas de nascimento distorcidas, com uma super representação de atletas de nível juvenil e profissional nascidos na primeira parte do ano em vários desportos. Jovens atletas com maturação biológica avançada tendem a ter capacidades físicas superiores em comparação aos seus colegas mais jovens. Essa diferença na maturação biológica pode conferir uma vantagem no desempenho desportivo (Cobley et al., 2009). Além disso, estudos mostram que a seleção de jogadores para equipas de grande dimensão é influenciada pela idade relativa, com os jogadores mais velhos sendo selecionados com mais frequência do que os mais jovens.

Tem-se verificado que a data de nascimento dos atletas acarreta efeitos sistemáticos, conhecidos como efeito de idade relativa. Este fenómeno está amplamente documentado nos escalões inferiores, onde os jogadores nascidos nos primeiros meses do ano beneficiam de vantagens físicas face aos colegas nascidos mais tarde, podendo influenciar injustamente a seleção e exclusão precoce destes últimos (Bezuglov et al., 2023). Contudo, menos estudos analisaram se o efeito de idade relativa se mantém significativo ao longo da transição para o futebol profissional ou se atenua com a maturação psicofísica dos jogadores. Alguns trabalhos sugerem mesmo que os jogadores mais tardios atingem êxitos semelhantes ou superiores nos seniores (Bezuglov et al., 2023).

As organizações desportivas têm-se concentrado cada vez mais a identificar e desenvolver talentos precoces no desporto (Till e Baker, 2020). As federações nacionais, por exemplo, investem consideráveis recursos em programas de identificação e desenvolvimento de talentos, incluindo o apoio a competições internacionais juvenis (Chroepf e Lames, 2018). O desempenho das seleções nacionais juvenis é frequentemente visto como um indicador chave da produtividade do desenvolvimento de talentos e do potencial sucesso no escalão sénior.

No entanto, a relação entre o desempenho desportivo na juventude e o sucesso na carreira sénior ainda é incerta, tanto a nível individual como coletivo. De acordo com Herrebrøden e Bjørndal (2022), tal acontece pelo facto de que os sistemas de identificação e desenvolvimento de talentos variarem entre diferentes culturas desportivas e organizações desportivas. Além disso, os requisitos de desempenho também variam entre diferentes posições de jogo e formações táticas.

Christensen (2009) sugere que a identificação e o desenvolvimento de talentos no futebol baseiam-se particularmente em processos especulativos ou subjetivos. Existem preocupações significativas de que os programas de desenvolvimento de talentos no futebol inglês sejam ineficientes, considerando o investimento em recursos humanos, instalações e jogadores. Essas preocupações de longa data têm levado a um aumento da atividade

científica que examina os fatores que influenciam o desenvolvimento de talentos no futebol, com ênfase nas características psicossociais, comportamentais e ambientais que podem facilitar tanto o desenvolvimento de talentos quanto as transições de jovens para seniores (Mitchell et al., 2020). Características como disciplina, compromisso, resiliência, autorregulação e autoconsciência são frequentemente mencionadas na literatura como habilidades e características psicossociais-chave que podem facilitar a transição dos jovens para o futebol profissional. Essas características são consideradas importantes para o desenvolvimento do talento, pois podem aumentar as hipóteses de os jogadores tomarem decisões adequadas em relação ao seu estilo de vida, envolverem-se em comportamentos de treino e prática adequados e consequentemente, melhorarem os seus níveis de desempenho técnico e tático (Mitchell et al., 2020).

2.2. Valor de mercado dos jogadores de futebol

O valor de mercado de um jogador profissional de futebol é influenciado por diversos fatores intrínsecos ao próprio jogador e ao contexto em que atua (Prinz e Weimar, 2018). No âmbito individual, elementos como a idade, posição em campo, qualidade de jogo, produtividade em termos de golos e assistências, reputação, títulos conquistados e lesões sofridas influenciam diretamente a cotação de um atleta no mercado de transferências. Geralmente, jogadores mais jovens, versáteis taticamente e eficazes tendem a ter um valor de mercado mais elevado. Além disso, fatores externos relacionados à formação, clube e competições em que o jogador está inserido também desempenham um papel significativo na avaliação do seu valor. Aspetos como o tempo passado na formação do clube anterior, a qualidade e reputação da academia de formação, transferências anteriores e o nível competitivo do clube de origem, assim como, sair de uma formação reconhecida por revelar talentos ou de uma equipa que disputa competições europeias ou uma liga de topo pode aumentar substancialmente a cotação de um jogador (Prinz e Weimar, 2018).

Adicionalmente, o histórico de transferências lucrativas entre clubes contribui para a perceção do potencial de valorização do jogador, impactando positivamente quem vende e negativamente quem compra. A idade do jogador e o seu potencial de evolução também são fatores relevantes, sendo que atletas mais jovens têm uma projeção de desenvolvimento que aumenta exponencialmente o seu valor presente e futuro (Prinz e Weimar, 2018).

O valor de mercado dos atletas é um indicador-chave para os clubes, refletindo fatores como potencial, nível de desempenho e retorno financeiro. Dessa forma, na transição entre a formação e o futebol profissional, o valor de mercado de um jogador resulta de uma combinação complexa entre fatores individuais, desportivos e contextuais. Esses elementos variam ao longo da carreira do jogador, sendo influenciados pelo seu desempenho, lesões e pelas dinâmicas do mercado de transferências (Prinz e Weimar, 2018).

2.3. Ranking

No âmbito das classificações desportivas, os investigadores Vaziri et al. (2018) realizam uma análise detalhada de cinco métodos populares, para compreender e potencialmente

melhorar os sistemas de classificação no desporto, contribuindo para uma avaliação mais precisa e equitativa do desempenho das equipas. Cada método de avaliação contém características distintas e aplicações específicas. O método mais tradicional e amplamente utilizado é o das Vitórias-Derrotas, que simplesmente ordena as equipas com base no número total de vitórias. Apesar da sua simplicidade e do incentivo claro para vencer cada partida, este método apresenta limitações significativas, uma vez que não considera a força dos adversários nem as margens de vitória (Vaziri et al., 2018).

Em contraste, o método de Massey, desenvolvido por Kenneth Massey em 1997 para classificar equipas de futebol universitário, introduz uma abordagem mais sofisticada. Este método baseia-se no princípio de que a diferença nas classificações entre duas equipas deve refletir a diferença nas pontuações dos seus confrontos diretos. Embora considere as diferenças de pontuação, o que representa um avanço em relação ao método de Vitórias-Derrotas, ainda não leva em conta a força específica de cada adversário individual (Vaziri et al., 2018).

O método de Colley, criado por Wesley Colley em 2002, apresenta semelhanças com o de Massey, mas com uma diferença crucial: em vez de considerar o diferencial de pontos, foca-se no diferencial de vitórias-derrotas. Este método utiliza um sistema de equações lineares para determinar as classificações, oferecendo uma perspetiva alternativa que valoriza a consistência nas vitórias em detrimento das margens de vitória (Vaziri et al., 2018).

Uma abordagem mais complexa é apresentada pelo método de Markov, que utiliza cadeias de Markov para classificar as equipas. Este método interpreta cada jogo como um "voto" da equipa perdedora para a vencedora, criando assim uma rede interligada de resultados. A sua força reside na consideração da qualidade do adversário em cada vitória, proporcionando uma visão mais holística do desempenho da equipa. No entanto, os autores Vaziri et al. (2018) alertam para a sua sensibilidade a pequenas alterações nos dados, o que pode levar a flutuações significativas nas classificações.

Por fim, o método de Elo, originalmente concebido por Arpad Elo para classificações de xadrez, ganhou popularidade em diversos desportos. Este método destaca-se pela sua fórmula dinâmica de atualização das classificações após cada jogo, levando em conta não só o resultado, mas também a força relativa dos adversários. Embora ofereça incentivos claros para a vitória e considere a qualidade dos oponentes, os autores Vaziri et al. (2018) salientam que a sequência dos jogos pode ter um impacto substancial na classificação final de uma equipa, o que pode ser visto como uma limitação em certos contextos. Este método tem sido utilizado como base para modelos de previsão de resultados (Buchdahl, 2003).

No contexto do futebol, o sistema do método de Elo atribui uma pontuação inicial a cada equipa, e essa pontuação é ajustada de acordo com o resultado de cada jogo, considerando vitórias, empates e derrotas. O cálculo baseia-se na diferença entre as classificações das equipas adversárias, permitindo prever probabilidades futuras de vitória,

empate ou derrota. Como discutido por Hvattum e Arntzen (2010), este método pode ser modificado para incluir a diferença de golos, o que permite uma avaliação mais apurada do desempenho das equipas. Por exemplo, uma vitória por 3-0 é mais valorizada do que uma vitória por 2-1, conferindo assim uma maior precisão ao modelo de previsão.

Este sistema de pontuação tem a vantagem de ser dinâmico, ajustando-se continuamente com base nos resultados mais recentes, o que o torna particularmente útil em desportos como o futebol, onde as *performances* das equipas podem variar ao longo do tempo (Hvattum e Arntzen, 2010).

2.4. Tecnologia no desporto

A tecnologia desempenha um papel cada vez mais preponderante no universo desportivo, introduzindo inovações e melhorias em várias vertentes. Esta tem viabilizado a melhoria das competências técnicas dos atletas e a satisfação da comunidade desportiva, sendo tais competências refinadas por meio da análise de dados, onde é possível identificar tanto os pontos fortes quanto os fracos do atleta e implementar correções através de planos de treino específicos. Numerosos avanços tecnológicos englobam melhorias nos sistemas de monitorização e registo de tempos, vestuário e equipamentos desportivos, tecnologias de deteção de golo, recursos videográficos, monitorização de dados GPS (*Global Positioning System*), realidade virtual, assim como sistemas de deteção e cobertura de eventos desportivos através da *internet* (Hire Intelligence, 2024).

De acordo com o estudo de Araújo et al. (2020), os métodos de notação e recolha de dados espaço-temporais ganham particular relevância. Tecnologias como os sistemas de videoanálise, GPS ou rastreamento de movimento permitem caracterizar as relações espaciais e temporais entre os agentes durante a ação, fornecendo informação funcional sobre como se regulam mutuamente para alcançar objetivos estratégicos.

A nível individual, variáveis como a distância interpessoal ou sincronia de deslocamentos entre atletas fornecem dados essenciais sobre como se estruturam as coordenações em 1vs1 ou a nível de subgrupos. Ao nível coletivo, medidas como a entropia aproximada ou a vantagem numérica de jogadores por setor do terreno acrescentam compreensão sobre a estabilidade e organização das equipas (Araújo et al., 2020).

A importância da tecnologia, como a análise de vídeo, na recolha e análise de dados de desempenho, destaca a necessidade de comparação e normalização de dados para uma interpretação precisa dos resultados. O potencial da biomecânica em desportos coletivos sugere uma maior atenção a esse campo de estudo para entender e aprimorar a análise de desempenho nesses contextos. Os treinos das equipas podem ser complementados com recurso a simulações e computadores, permitindo melhorar o domínio de técnicas, identificação de erros e correções, algo que é frequentemente usado pelas agências espaciais (Hughes e Bartlett, 2002).

2.4.1. Tecnologia no futebol

Em todos os desportos, verifica-se a existência de uma tendência que os torna mais lógicos e racionais. No entanto, no futebol, a presença de momentos inesperados, especialmente em momentos cruciais como o golo, pode influenciar decisivamente o desfecho do jogo. Neste contexto, cabe ao treinador a responsabilidade de analisar de forma precisa todas as variáveis presentes em campo, procurando estabelecer correlações entre as ações de jogo e a subsequente sequência de eventos (Wang e Chen, 2022). Para que o treinador e a equipa técnica possam treinar e corrigir as falhas da equipa de forma eficaz, torna-se indispensável o acesso a dados que permitam avaliar o desempenho ao longo dos jogos. Isso inclui compreender o nível individual de cada jogador em relação à média da equipa e por sua vez, o nível global da equipa em comparação com os adversários (Dufour, 2003).

A análise de jogo no futebol profissional tem ganho importância crescente nos últimos anos, à medida que novos sistemas como o *Global Positioning System* (GPS) permitem recolher dados em tempo real sobre o desempenho dos jogadores. Contudo, é necessário sistematizar periodicamente a investigação nesta área de forma a organizar teoricamente os principais tópicos e conceitos estudados (Sarmento et al., 2018).

Neste sentido, os autores Sarmento et al. (2018) propõem um sistema abrangente baseado em tecnologias de ponta, visando proporcionar informações pertinentes através da gestão inteligente de dados e análise tática. Este sistema, dividido em módulos interligados, é concebido para atender às exigências específicas do universo futebolístico.

O módulo de Gestão de Informação Base assume a responsabilidade pela administração de dados cruciais, como informações sobre jogadores, equipas e competições. Utilizando funcionalidades CRUD (Criar, Ler, Atualizar, Eliminar), assegura a integridade dos dados por meio de regras de consistência. A Aquisição tática engloba a captação em tempo real da localização dos jogadores, utilizando sensorização como GPS, e o registo manual de eventos durante o jogo. A aplicação da filtragem de Kalman visa eliminar o ruído dos sinais dos sensores (Wang e Chen, 2022).

Segundo o site da FIFA (2022), para as competições internacionais de seleções, dentro dos estádios, foram disponibilizados *tablets* estrategicamente: um para a equipa de analistas, um para cada analista no banco de cada equipa e outro para a equipa médica. Fornecem dados visuais sobre a posição de cada jogador, incluindo passes, velocidade, ataques e outras estatísticas importantes do jogo.

2.4.2. Importância de dados no futebol

A análise de desempenho no futebol tem experimentado avanços significativos graças ao desenvolvimento da inteligência computacional. De acordo com Wang e Chen (2022) a criação de um sistema de análise de táticas futebolísticas, concebido com base em requisitos funcionais e de desempenho, representa um marco importante nesse cenário. O sistema, composto por módulos de gestão de informações básicas, recolha de dados táticos, deteção de padrões de passes, consulta de informações táticas e visualização, destaca-se pela sua abordagem inovadora. Uma das contribuições foi a proposta de um método

de deteção de padrões de passes, permitindo uma identificação inteligente de estratégias com base nas relações entre os jogadores. A *interface* foi meticulosamente projetada para proporcionar uma visualização simples e intuitiva de informações cruciais, como dados de atletas, equipas, competições, táticas e padrões de passes, utilizando inteligência computacional. Validações realizadas em conjunto com um treinador experiente demonstram a capacidade do sistema em analisar eventos como a mudança de estratégia defensiva numa fase decisiva do jogo, procurando obter melhores resultados escolhendo as suas estratégias através da análise do estado físico dos jogadores e do desempenho da equipa, com recurso a componentes de *software* (Wang e Chen, 2022).

Os avanços tecnológicos na captação e tratamento de dados têm sido fundamentais para a evolução da análise de desempenho, permitindo uma compreensão cada vez mais aprofundada das estratégias desportivas. De acordo com os autores Lord et al. (2022), a recolha precisa de grandes volumes de dados durante um jogo de futebol é essencial para replicar a complexidade de fatores em interação que influenciam cada ação. Para além da captação, o processamento e visualização dos dados também requerem atenção cuidadosa.

Após a normalização, análises estatísticas como *clustering* permitem identificar "estilos de jogo", reduzindo grandes conjuntos de dados a perfis práticos comunicáveis. Estes perfis holísticos, quando analisados em camadas de detalhe crescente e relacionados com variáveis técnico-táticas, fornecem valiosas perspetivas sobre as estratégias das equipas. Plataformas interativas como aplicações *Shiny* facilitam a comunicação destes *insights*, permitindo ao utilizador filtrar e explorar dinamicamente os dados através de visualizações. Diferentes tipos de gráficos contribuem para compreender as estratégias a diferentes níveis: perfis gerais de estilo de jogo, análise por componentes do jogo e eventos específicos (Lord et al., 2022).

A recolha de dados muitas vezes é realizada de forma manual, e a *Opta* é um exemplo de uma empresa que regista manualmente dados como por exemplo, passes completos, toques defensivos, percentagem de bolas ganhas, ataques e golos marcados. Esta empresa documenta cada evento do jogo, já que os mecanismos de processamento das câmaras não conseguem identificar todos esses dados. Normalmente, cada analista de jogo tem o seu próprio monitor, permitindo a visualização das ações de um jogador específico ou da equipa. A introdução manual de dados está sujeita à disposição, preferência e eventual erro humano. No entanto, os analistas aproveitam as escassas interrupções durante o jogo para rever golos e outros eventos (Bialik, 2014).

De acordo com Pappalardo et al. (2019), a deteção automática de padrões táticos facilita a análise pós-jogo dos treinadores. Além disso, a reconstrução das redes de passe entre jogadores permite identificar as linhas mestras e jogadores-chave em cada equipa. Outra vertente prende-se com a "ciência do sucesso". Estudar a relação entre *performance* e resultado desportivo possibilita compreender melhor os determinantes do êxito e do fracasso. Por outro lado, a riqueza espacial e temporal destes dados oferece múltiplas métricas para comparar o desempenho individual de jogadores, como a centralidade

ou a qualidade exibida ao longo de épocas. A análise científica proporciona ferramentas inestimáveis para treinadores, dirigentes e agentes desportivos tomarem decisões estratégicas cada vez mais fundamentadas. Permite igualmente desvendar novos aspetos sobre a dinâmica deste complexo sistema social que é o futebol de equipas. Resumindo, a ciência dos dados no desporto-rei revela-se essencial para potenciar o desempenho coletivo e individual na procura do sucesso (Pappalardo et al., 2019).

2.4.3. *Machine learning* no futebol

A *Artificial Intelligence* (AI) tem vindo a assumir um papel dominante na sociedade contemporânea, influenciando a nossa interação com o mundo e a resolução de problemas complexos. No âmbito desta revolução tecnológica encontra-se o campo da aprendizagem automática (*machine learning*), que permite aos computadores aprenderem com dados sem necessidade de programação explícita (Mahesh, 2020).

Mahesh (2020), oferece uma análise abrangente dos diversos algoritmos e técnicas de aprendizagem automática, realçando a sua relevância em diversas áreas. A capacidade de aprender com a experiência, sem a necessidade de instruções detalhadas, torna a aprendizagem automática uma ferramenta poderosa para extrair conhecimento e fazer previsões a partir de conjuntos de dados cada vez mais vastos.

A aprendizagem automática divide-se essencialmente em duas categorias: aprendizagem supervisionada e não supervisionada. Na aprendizagem supervisionada, é necessária a intervenção humana para fornecer exemplos de entrada e saída, o que pode ser interpretado como uma forma simplificada de orientação. Por outro lado, na aprendizagem não supervisionada, lidamos com dados não rotulados, desafiando-nos a descobrir padrões e estruturas ocultas nos dados sem a presença de um guia explícito (Mahesh, 2020).

Alguns dos algoritmos mais comuns em aprendizagem supervisionada incluem Árvores de Decisão, *Naive Bayes* e Máquinas de Vetor de Suporte (SVM). Por outro lado, em aprendizagem não supervisionada, destacam-se a Análise de Componentes Principais (PCA) e o agrupamento *K-Means*. A escolha entre aprendizagem supervisionada e não supervisionada depende da natureza do problema e dos dados disponíveis, com a primeira a ser mais adequada para dados rotulados e previsões precisas, enquanto a segunda é preferível para descobrir padrões em dados não rotulados (Mahesh, 2020).

De acordo com o site Sports (2024), o recurso à aprendizagem automática no futebol tem vindo a crescer significativamente, disponibilizando às equipas uma panóplia de ferramentas para aprimorar o desempenho dos jogadores, analisar padrões de jogo e otimizar estratégias táticas. Os algoritmos de aprendizagem automática conseguem processar grandes volumes de dados, desde estatísticas de jogadores, dados biométricos, padrões de desempenho até às condições meteorológicas, proporcionando às equipas uma compreensão mais profunda do desempenho dos jogadores e das condições do jogo.

2.4.4. *Machine learning* na análise de desempenho de jogadores

O uso de *machine learning* na análise de desempenho de jogadores de futebol tem ganhado destaque, fornecendo novas ferramentas para equipas técnicas e clubes. De acordo com Sun (2023), a avaliação de desempenho, que tradicionalmente era feita de forma subjetiva ou manual, agora pode ser conduzida de maneira automatizada, com base em grandes quantidades de dados gerados durante os jogos. Esses dados incluem métricas como o número de golos, assistências, passes completos, distância percorrida, entre outras estatísticas. Com isso, é possível obter uma análise mais precisa e imparcial do desempenho dos atletas. Ao analisar padrões táticos ao longo de diversos jogos, estes algoritmos identificam estratégias eficazes com base em estatísticas, habilidades específicas e características de jogo. Esta abordagem auxilia na identificação de fraquezas na abordagem adversária, permitindo que os treinadores ajustem as táticas e maximizem as probabilidades de sucesso. A avaliação do desempenho individual, que inclui movimentos, tomada de decisões e eficácia em diferentes situações de jogo, também pode ser refinada com o uso da aprendizagem automática. Isso possibilita ajustes personalizados nos treinos, otimizando o desenvolvimento de cada jogador (Sports, 2024).

Os algoritmos podem simular diversos cenários de jogo com base em dados históricos, auxiliando os treinadores na identificação das estratégias mais eficazes em diferentes situações. Este recurso revela-se particularmente útil no planeamento de jogadas específicas e na tomada de decisões táticas durante partidas importantes. (Sports, 2024)

A análise de desempenho de jogadores de futebol através de sistemas de pontuação automática baseados em *machine learning* tem sido uma aplicação crucial no campo desportivo. No entanto, segundo Sun (2023), a maioria dos modelos atuais tende a subvalorizar os jogadores defensivos, uma vez que se concentram principalmente em dados ofensivos mais facilmente quantificáveis, como golos e assistências (Sun, 2023).

Para superar essa limitação, algumas abordagens têm sido propostas. Por exemplo, um estudo proposto por Merhej et al. (2021), que desenvolveu um modelo que atribui pesos diferenciados às ações defensivas, como intercepções e desarmes, para melhorar a precisão na avaliação do desempenho dos jogadores nessa vertente. Outro estudo proposto por Cui et al. (2022), sugeriu um modelo de avaliação baseado em comportamentos com bola, o que pode não captar completamente o impacto de jogadores que contribuem de forma significativa sem a bola, como os que realizam deslocamentos estratégicos.

De acordo com investigador Sun (2023), os desafios enfrentados pelos modelos atuais incluem a complexidade dinâmica do jogo de futebol, que requer uma análise mais abrangente e integrada. Uma das abordagens mais promissoras consiste no uso de redes neuronais profundas, que conseguem processar dados temporais e espaciais em tempo real, possibilitando uma análise mais pormenorizada das interações entre os jogadores e das suas posições no campo. Estas redes têm a capacidade de identificar padrões complexos de comportamento e de antecipar o impacto de determinadas ações no desfecho final do jogo. Alguns estudos sugerem que a utilização de redes neuronais recorrentes (RNNs)

ou redes neuronais convolucionais (CNNs) pode ser uma solução eficaz para capturar a dinâmica do jogo e melhorar a precisão das avaliações de desempenho.

O estudo de Merzah et al. (2024), explora a aplicação inovadora de modelos de aprendizagem automática para a avaliação do desempenho de jogadores de futebol. O modelo PEMLM (*Performance Evaluation Machine Learning Model*), ao integrar dados abrangentes de treino e jogo, oferece uma análise minuciosa das habilidades e capacidades dos atletas. Nesse estudo, o desempenho dos jogadores de futebol é classificado em três categorias distintas: fraco, normal e ativo. A classe "fraco" refere-se a jogadores que apresentam um desempenho abaixo do esperado, indicando a necessidade de intervenções, como um aumento na carga de treino ou a sua substituição durante uma partida. A classe "normal" abrange aqueles que demonstram um desempenho aceitável, cumprindo as suas funções, mas que podem necessitar de ajustes na carga de trabalho para melhorar a sua eficácia. Por fim, a classe "ativo" é atribuída a jogadores que se destacam, contribuindo significativamente para a equipa e que devem continuar a ser utilizados tanto em jogos como em treinos (Merzah et al., 2024). As características utilizadas para detalhar as classes de desempenho refletem as habilidades e o desempenho físico dos jogadores. Entre estas, o género, a área de atuação do jogador em campo, que pode ser atacante, defensor ou jogador de meio-campo, também é uma característica relevante. Adicionalmente, são consideradas métricas como a distância cruzada percorrida pelo jogador, a velocidade, a *accuracy* de passes, o número de ações ofensivas e as ações defensivas, como desarmes e interseção (Merzah et al., 2024).

A avaliação do valor de jogadores profissionais no mercado de transferências tem sido tradicionalmente feita através de métodos como a regressão linear. No entanto, estudos recentes, como o de Li et al. (2022), têm introduzido modelos de *machine learning*, como o *Random Forest* e a regressão linear múltipla, para melhorar essa avaliação. Esses modelos conseguem processar grandes quantidades de dados e identificar padrões complexos, resultando em estimativas mais precisas do valor de mercado dos jogadores. Embora esses modelos tenham mostrado sucesso em muitos casos, ainda há espaço para melhorias, especialmente na consideração de fatores como histórico de lesões e condições de mercado (Li et al., 2022). A inclusão desses elementos, juntamente com dados de indicadores de *performance* física, pode aprimorar a precisão das estimativas de valor.

CAPÍTULO 3

Metodologia CRISP-DM e a sua aplicação

3.1. Contexto da metodologia

No presente capítulo, é essencial descrever a metodologia utilizada, destacando a importância de rever as etapas deste modelo e como a retropropagação impulsiona o desenvolvimento do projeto. Assim, de acordo com o autor Hotz (2024), o método *Cross-industry standard process for data mining* (CRISP-DM) é um modelo de estrutura de processos de padrão aberto para o planeamento ou desenvolvimento de projetos na área de prospeção de dados. O método CRISP-DM divide-se em seis etapas, tal como está presente na Figura 3.1. A primeira etapa corresponde ao *Business Understanding*, relativo ao entendimento do fenómeno em análise, a segunda etapa *Data Understanding*, que versa sobre a compreensão dos dados disponíveis e necessários para iluminar o fenómeno em análise, o *Data Preparation*, que visa a preparação e tratamento dos dados recolhidos, o *Modeling*, que trata a modelação realizada com base nos dados tratados, a etapa *Evaluation*, permite avaliar o processo, e por fim a etapa *Deployment*, que procura a implementação da solução alcançada, mas que não foi executada no presente projeto.

Esta metodologia é designada de padrão aberto por permitir o "vaivém" entre estas etapas donde o conhecimento alcançado é enriquecido pelas metas estabelecidas em etapas anteriores.

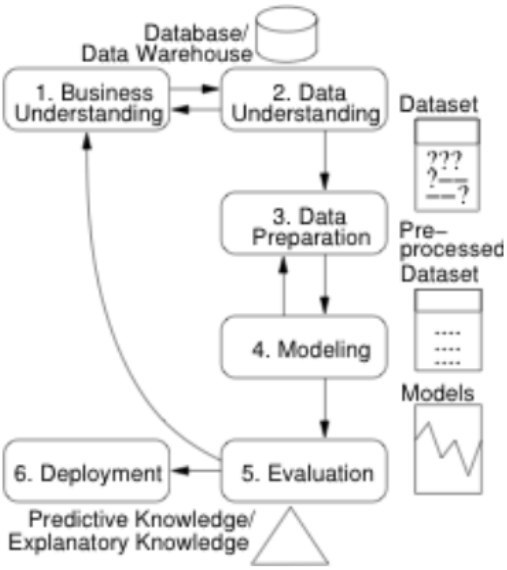


FIGURA 3.1. Metodologia CRISP-DM (adaptado de Chapman et al., 2000)

3.2. CRISP-DM Fase 1: Compreensão do Negócio

A análise de dados no futebol está a tornar-se cada vez mais crucial para clubes, fornecendo perceções sobre o desempenho e potencial dos jogadores. Este estudo foca-se num conjunto de dados que abrange futebolistas nascidos entre 2000 e 2006 que competiram na *Championship* (segunda divisão inglesa) na época 2022/2023.

A *Championship* é uma liga altamente competitiva e financeiramente relevante fora das principais divisões europeias, sendo um campo de estudo excelente. Muitos jogadores usam esta liga como trampolim para carreiras mais destacadas.

Os dados detalhados incluem estatísticas de desempenho ao longo de várias épocas, valores de mercado atuais e classificações dos clubes onde jogaram. Estas informações oferecem uma visão da evolução desses jovens talentos nas suas carreiras iniciais.

O principal objetivo deste estudo é classificar a evolução dos jogadores com base em parâmetros estatísticos e nas classificações dos clubes onde jogaram ao longo das épocas. Desenvolver um modelo preditivo para estimar o valor de mercado destes jovens jogadores e utilizar séries temporais para prever valores futuros de mercado dos jogadores a curto prazo, neste caso como a série temporal é semestral, o intuito é prever os valores de mercado dos jogadores para o semestre seguinte.

Compreender e antecipar o valor de mercado dos jogadores é crucial para os clubes, agentes e investidores no futebol moderno, permitindo decisões mais fundamentadas sobre contratações, vendas e desenvolvimento de talento. Além disso, a classificação da evolução dos jogadores pode ajudar a identificar talentos promissores precocemente, orientando estratégias de desenvolvimento e investimento a longo prazo.

Este estudo tem o potencial de fornecer perceções para várias partes interessadas no mundo do futebol. Pode ajudar os clubes a otimizar estratégias de contratação e desenvolvimento de jogadores, oferecer aos olheiros uma ferramenta adicional na identificação de talentos subvalorizados e auxiliar os investidores na avaliação do potencial de retorno em investimentos em jovens jogadores.

Ao analisar esses dados, espera-se identificar padrões de desenvolvimento, fatores-chave que influenciam o valor de mercado e indicadores precoces de sucesso futuro. Isso pode levar a uma compreensão da progressão de carreira de jovens futebolistas e dos fatores que contribuem para o seu sucesso e a valorização no mercado.

À medida que o futebol continua a evoluir com a crescente influência da análise de dados, estudos como este tornam-se cada vez mais relevantes. A aplicação correta dessas técnicas pode contribuir para uma tomada de decisão mais eficaz e para o desenvolvimento de estratégias inovadoras na gestão e desenvolvimento de talentos no futebol.

3.3. CRISP-DM Fase 2: Compreensão dos dados

Durante a fase de Compreensão dos Dados do processo CRISP-DM, realizou-se uma análise detalhada do esquema da base de dados e definiu-se precisamente o problema em estudo. Esta etapa foi crucial para determinar o conjunto de dados necessário, que

foi criado através da exportação dos dados obtidos, permitindo uma seleção e filtragem adequadas dos dados relevantes.

Os dados foram extraídos de três fontes distintas (*websites*):

- FBRef: Forneceu uma ampla gama de estatísticas de desempenho dos jogadores, incluindo dados estatísticas de jogo, métricas avançadas e informações posicionais.
- *Transfermarkt*: Contribuiu com informações sobre o valor de mercado dos jogadores, incluindo as datas de avaliação.
- *FootballDatabase*: Forneceu dados sobre o ranking e a pontuação dos clubes.

Para a extração dos dados, utilizou-se técnicas de *Web scraping*, um método automatizado de recolha de informações não estruturadas da *Web*. Este processo transformou os dados num conjunto estruturado, adequado para análise posterior.

Utilizou-se a linguagem de programação *Python* para o desenvolvimento de *scripts*, para a consolidação dos dados, para a análise de variáveis e utilização de modelos de *machine learning*,

3.3.1. FBref

O site FBref destaca-se no universo do futebol pela sua *interface* intuitiva e pela ampla gama de métricas e análises disponibilizadas. Esta plataforma permite aos utilizadores uma exploração profunda do desempenho de jogadores, equipas e ligas. Graças à sua sólida reputação no meio futebolístico, o FBref tornou-se uma ferramenta valiosa para aqueles que procuram informações detalhadas sobre este desporto.

3.3.1.1. Extração dos dados

A Figura 3.2 ilustra a página "2022-2023 *Championship Stats*". Nesta, é possível visualizar as estatísticas referentes à época selecionada através de um menu *dropdown*. As informações apresentadas abrangem diversos aspetos da competição, tais como: a equipa campeã da temporada, o melhor marcador da liga, a equipa que sofreu menos golos, entre outras estatísticas. Adicionalmente, a página oferece um botão de acesso direto aos resultados detalhados dos jogos desta liga, como se pode verificar na Figura (3.3), permitindo uma análise mais aprofundada das partidas individuais.

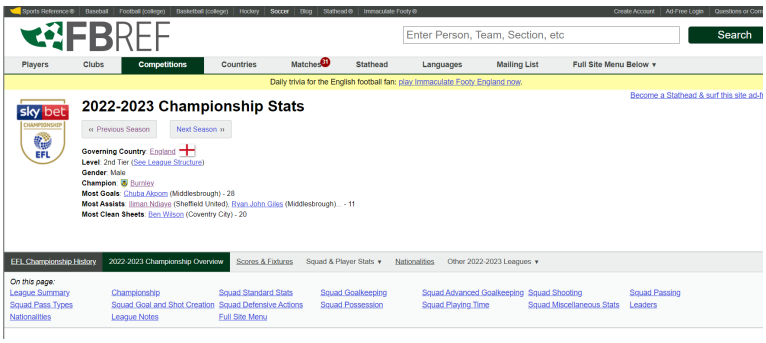
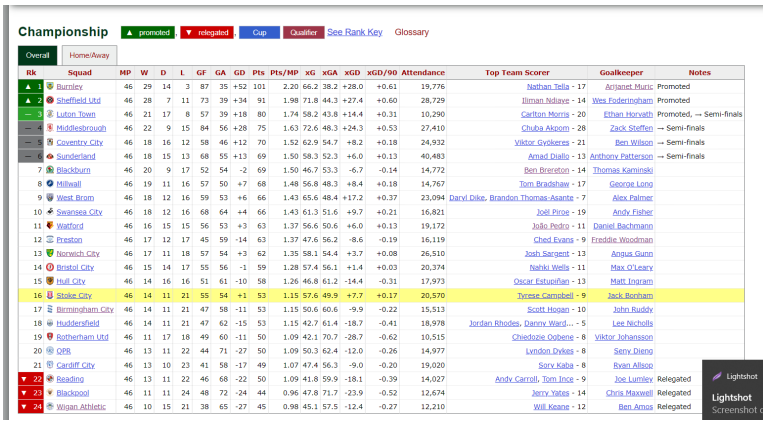


FIGURA 3.2. Página "*Championship* 2022-2023".



Championship ▲ promoted ▼ relegated Cup Qualifier See Rank Key Glossary

Overall | Home/Away

Rk	Squad	MP	W	D	L	GF	GA	GD	Pts	Pts/MP	xG	xGA	xGD	xGD/90	Attendance	Top Team Scorer	Goalkeeper	Notes
▲ 1	Burnley	46	29	14	3	87	35	+52	101	2.20	66.2	38.2	+28.0	+0.61	19,776	Nathan Tella - 17	Arijanet Muric	Promoted
▲ 2	Sheffield Utd	46	28	7	11	73	39	+34	91	1.98	71.8	44.3	+27.4	+0.60	28,729	Iliman Ndiaye - 14	Wes Foderingham	Promoted
— 3	Luton Town	46	21	17	8	57	39	+18	80	1.74	58.2	43.8	+14.4	+0.31	10,290	Carlton Morris - 20	Ethan Horvath	Promoted, → Semi-finals
— 4	Middlesbrough	46	22	9	15	84	56	+28	75	1.63	72.6	48.3	+24.3	+0.53	27,410	Chuba Akpom - 28	Zack Steffen	→ Semi-finals
— 5	Coventry City	46	18	16	12	58	46	+12	70	1.52	62.9	54.7	+8.2	+0.18	24,932	Viktor Gyökeres - 21	Ben Wilson	→ Semi-finals
— 6	Sunderland	46	18	15	13	68	55	+13	69	1.50	58.3	52.3	+6.0	+0.13	40,483	Amad Diallo - 13	Anthony Patterson	→ Semi-finals
7	Blackburn	46	20	9	17	52	54	-2	69	1.50	46.7	53.3	-6.7	-0.14	14,772	Ben Brereton - 14	Thomas Kaminski	
8	Millwall	46	19	11	16	57	50	+7	68	1.48	56.8	48.3	+8.4	+0.18	14,767	Tom Bradshaw - 17	George Long	
9	West Brom	46	18	12	16	59	53	+6	66	1.43	65.6	48.4	+17.2	+0.37	23,094	Daryl Dike, Brandon Thomas-Asante - 7	Alex Palmer	
10	Swansea City	46	18	12	16	68	64	+4	66	1.43	61.3	51.6	+9.7	+0.21	16,821	Joël Piroe - 19	Andy Fisher	
11	Watford	46	16	15	15	56	53	+3	63	1.37	56.6	50.6	+6.0	+0.13	19,172	João Pedro - 11	Daniel Bachmann	
12	Preston	46	17	12	17	45	59	-14	63	1.37	47.6	56.2	-8.6	-0.19	16,119	Ched Evans - 9	Freddie Woodman	
13	Norwich City	46	17	11	18	57	54	+3	62	1.35	58.1	54.4	+3.7	+0.08	26,510	Josh Sargent - 13	Angus Gunn	
14	Bristol City	46	15	14	17	55	56	-1	59	1.28	57.4	56.1	+1.4	+0.03	20,374	Nahki Wells - 11	Max O'Leary	
15	Hull City	46	14	16	16	51	61	-10	58	1.26	46.8	61.2	-14.4	-0.31	17,973	Oscar Estupiñan - 13	Matt Ingram	
16	Stoke City	46	14	11	21	55	54	+1	53	1.15	57.6	49.9	+7.7	+0.17	20,570	Tyrese Campbell - 9	Jack Bonham	
17	Birmingham City	46	14	11	21	47	58	-11	53	1.15	50.6	60.6	-9.9	-0.22	15,513	Scott Hogan - 10	John Ruddy	
18	Huddersfield	46	14	11	21	47	62	-15	53	1.15	42.7	61.4	-18.7	-0.41	18,978	Jordan Rhodes, Danny Ward... - 5	Lee Nicholls	
19	Rotherham Utd	46	11	17	18	49	60	-11	50	1.09	42.1	70.7	-28.7	-0.62	10,515	Chiedozie Ogbene - 8	Viktor Johansson	
20	QPR	46	13	11	22	44	71	-27	50	1.09	50.3	62.4	-12.0	-0.26	14,977	Lyndon Dykes - 8	Seny Dieng	
21	Cardiff City	46	13	10	23	41	58	-17	49	1.07	47.4	56.3	-9.0	-0.20	19,020	Sory Kaba - 8	Ryan Allsop	
▼ 22	Reading	46	13	11	22	46	68	-22	50	1.09	41.8	59.9	-18.1	-0.39	14,027	Andy Carroll, Tom Ince - 9	Joe Lumley	Relegated
▼ 23	Blackpool	46	11	11	24	48	72	-24	44	0.96	47.8	71.7	-23.9	-0.52	12,674	Jerry Yates - 14	Chris Maxwell	Relegated
▼ 24	Wigan Athletic	46	10	15	21	38	65	-27	45	0.98	45.1	57.5	-12.4	-0.27	12,210	Will Keane - 12	Ben Amos	Relegated

Lightshot

Lightshot

Screenshot

FIGURA 3.3. Classificação da *Championship* 2022-2023.

Na página presente no Anexo A da Figura 5.1, encontram-se informações sobre o plantel de uma equipa numa certa época e os jogos que decorreram em todas as competições, incluindo detalhes específicos sobre cada jogo em cada competição.

A página dos jogadores, Figura 5.2 presente no Anexo A exibe informações como o nome, idade, data de nascimento, peso, altura, local de nascimento, pé preferido para jogar, clube atual, liga nacional em que joga e seleção nacional. Além disso, esta página inclui um resumo, biografia e estatísticas de todas as épocas, entre outras informações relevantes.

3.3.1.2. Definição de parâmetros e navegação no *website*

Foi desenvolvido um *script* para extrair e recolher dados detalhados sobre jogadores de futebol da liga *Championship*, com foco em jovens nascidos entre 2000 e 2006. Começou-se por definir as épocas de interesse (neste caso, 2022-2023) e estabeleceu-se uma estrutura para armazenar *links* de equipas e jogadores.

Uma sessão de pedidos *Hypertext Transfer Protocol* (HTTP) foi configurada com um *User-Agent* personalizado para evitar bloqueios do servidor. O *script* percorreu as páginas das equipas, extraindo *links* para os perfis dos jogadores que se enquadram no critério de idade (até 24 anos).

Diversas funções foram definidas para extrair informações específicas dos jogadores, como nome, data de nascimento, posição, pé dominante e país de origem. Uma função adicional utilizou o site *Transfermarkt* para obter um ID alternativo do jogador, permitindo uma verificação adicional dos dados, que foi utilizada mais tarde para fazer um *merge* com o valor de mercado dos jogadores. Criou-se uma estrutura de diretórios para armazenar os dados de cada jogador, organizando-os em pastas individuais. Implementou-se um sistema de gestão de erros e limpeza de dados, eliminando pastas de jogadores que não cumpriam os critérios de idade estabelecidos.

Os dados foram extraídos de tabelas *HyperText Markup Language* (HTML) e convertidos em *dataframes* da biblioteca pandas, aos quais foram adicionadas informações complementares sobre o jogador. Esses *dataframes* foram então exportados para formatos de ficheiros Excel, um para cada tabela relevante encontrada no perfil do jogador.

O código inclui atrasos aleatórios entre pedidos para evitar sobrecarregar os servidores de destino e utilizou-se a biblioteca *tqdm* para fornecer uma barra de progresso visual durante a execução.

Todos estes dados foram organizados em pastas individuais para cada jogador.

3.3.1.3. Consolidação dos dados

No final do *Web scraping*, obteve-se várias diretorias de ficheiros do tipo csv. A leitura e consolidação dos ficheiros de cada uma destas diretorias resultou num só *dataframe*. O processo de consolidação dos ficheiros, implementou uma abordagem sistemática para organizar, limpar e estruturar os dados obtidos.

Primeiramente, criou-se uma pasta principal para armazenar as informações de cada jogador. O programa iterou sobre uma lista de *Uniform Resource Allocator* (URL) dos jogadores, utilizando identificadores únicos para acessar às suas páginas individuais no site fbref.com.

Para cada jogador, foram extraídas várias informações, como o nome, ano e data de nascimento, posição em campo, pé dominante, país de origem, seleção nacional (se aplicável) e identificador no site *Transfermarkt*. Essas informações foram adicionadas às tabelas extraídas para cada jogador, que incluem dados sobre ações defensivas, criação de golos e remates, tipos de passes, tempo de jogo, posse de bola, remates e estatísticas padrão. Estas informações, estão apresentadas em várias tabelas no Anexo B.

O código utilizou funções especializadas, como *clean_playing()*, *clean_extra()* e *rename_wrong()*, para limpar e preparar os dados, removendo colunas redundantes, corrigindo nomes de colunas com o mesmo nome, garantindo a consistência dos dados. A consolidação dos dados foi realizada por meio de operações de *merge*. Cada tabela foi lida de um arquivo Excel, e posteriormente limpa juntando-se a um *dataframe* em crescimento. As colunas comuns usadas para o *merge* foram *'Season'*, *'Squad'*, *'Comp'* e *'Player_ID'*.

O código iterou por todas as subpastas na pasta principal, processando cada conjunto de tabelas de jogador. À medida que processou cada jogador, um *dataframe* final foi construído, contendo todas as estatísticas de todos os jogadores.

3.3.2. *Transfermarkt*

O site *Transfermarkt* destaca-se no universo do futebol pela sua *interface* e pela vasta gama de informações sobre transferências, valores de mercado e estatísticas de jogadores e equipas. Esta plataforma permite aos utilizadores uma exploração profunda do mercado de transferências e do desempenho de jogadores em várias ligas.

3.3.2.1. Extração dos dados

A Figura 3.4 ilustra a página "*Clubs Championship* 22/23". Nesta página, é possível visualizar as estatísticas referentes à época selecionada através de um menu *dropdown* para filtrar por época. As informações incluem os nomes das equipas, o número de jogadores por equipa, a média de idades por equipa, o valor de mercado por equipa e a média de valor de mercado por equipa, juntamente com outras estatísticas relevantes.

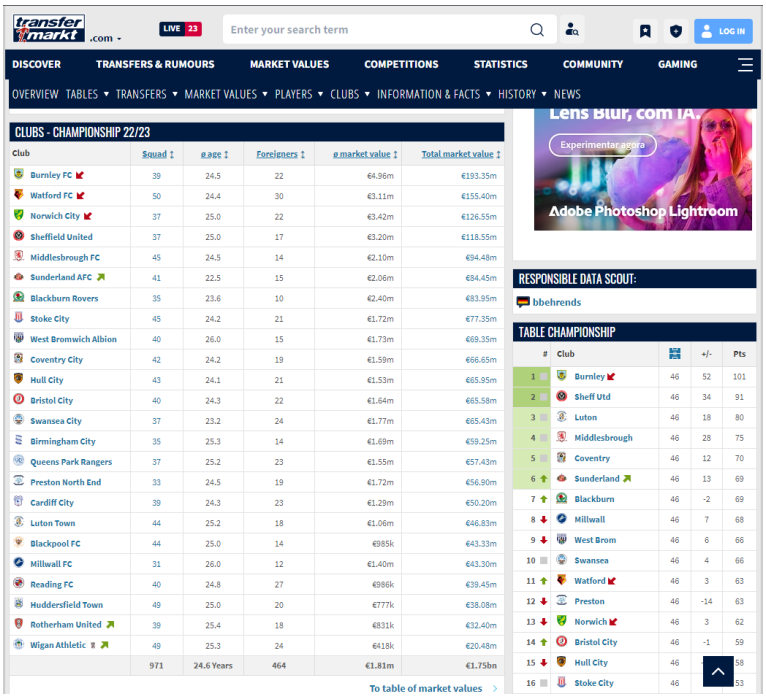


FIGURA 3.4. Página do *Transfermarkt* da *Championship* 2022-2023.

A página apresentada na Figura 5.3 presente no Anexo A contém informações sobre o plantel da equipa naquela época, incluindo detalhes específicos sobre cada jogador, como a data de nascimento, o clube atual onde joga e o seu valor de mercado atual.

Estas informações são também apresentadas na página dos jogadores 5.4 presente no Anexo A, que exibe para além destas, informações como o nome, idade, data de nascimento, posição, altura, nacionalidade, pé preferido para jogar, clube atual, liga nacional em que joga e um gráfico com o valor de mercado. Além disso, esta página inclui estatísticas da época atual, entre outras informações relevantes.

3.3.2.2. Definição de parâmetros e Navegação no *website*

Começou-se por definir uma lista vazia chamada "*data*" para armazenar os dados recolhidos e específica os anos de interesse, nomeadamente 2022 e 2023.

Uma sessão de requisições HTTP foi estabelecida utilizando a biblioteca *"requests"*, com um *User-Agent* configurado para simular um navegador *web* e evitar possíveis bloqueios do servidor. O *script* iterou sobre os anos especificados, construindo URLs dinâmicos para aceder às páginas das equipas do *Championship* em cada temporada.

Para cada equipa identificada, o *script* acedeu à página do plantel e extraiu informações sobre cada jogador. Utilizou-se a biblioteca *BeautifulSoup* para analisar o HTML das páginas e extrair os *links* relevantes. Para cada jogador, o código recolheu o nome, ID, URL do perfil e os dados do valor de mercado ao longo do tempo.

Utilizou-se uma *Application Programming Interface* (API) interna do *Transfermarkt* para obter os dados históricos do valor de mercado de cada jogador. Isto foi feito através de uma requisição separada a um *endpoint* específico, utilizando o ID do jogador.

Os dados recolhidos foram armazenados numa estrutura de dicionário, e guardados num ficheiro *JavaScript Object Notation* (JSON), que inclui o ano, nome do jogador, URL do perfil, ID do jogador e os dados do valor de mercado. Estes dados foram posteriormente convertidos num *dataframe* da biblioteca pandas para facilitar a análise.

3.3.2.3. Consolidação dos dados

O processo de consolidação de dados implicou uma transformação significativa da estrutura inicial, com o objetivo de criar um conjunto de dados mais detalhado e analiticamente útil. Essa reestruturação foi realizada através da criação de um novo *dataframe* que amplia o histórico de valores de mercado de cada jogador de futebol.

No *dataframe* original, cada linha representou um jogador único, com uma coluna *'RawMarketData'* que continha uma lista de dicionários. Cada dicionário nessa lista representava um ponto específico na história do valor de mercado do jogador. A reestruturação desdobrou essa informação complexa em múltiplas linhas, cada uma representando um momento específico na carreira do jogador.

O novo *dataframe* foi composto pelas seguintes colunas: Idade, Clube, Data, Preço, Jogador, URL do Jogador, e ID do Jogador. Essa estrutura possibilitou uma análise detalhada sobre a evolução temporal, pois foi possível traçar a trajetória do valor de mercado de um jogador ao longo do tempo, analisando por idade e clube, e comparar entre diferentes jogadores as tendências do mercado de transferências.

3.3.3. *FootballDatabase*

O site *FootballDatabase* é uma plataforma abrangente que oferece uma variedade de informações e estatísticas sobre futebol, incluindo *rankings* de clubes. De acordo com o site FootballDatabase (2024) o World Football Elo Ratings, adaptado do sistema de classificação Elo, proporciona um método alternativo para classificar as seleções nacionais de futebol. Este sistema emprega uma fórmula central:

$R_n = R_o + K * G * (W - W_e)$

Nesta fórmula, R_n representa a nova classificação da equipa, R_o a classificação anterior, K o índice de ponderação baseado na importância do torneio, G o fator de diferença de golos, W o resultado real do jogo e W_e o resultado esperado. Os aspetos fundamentais incluem a ponderação do torneio (K), que atribui pesos diferentes aos jogos consoante a sua importância, e o fator de diferença de golos (G), que ajusta o impacto de uma vitória de acordo com a margem de golos.

3.3.3.1. Extração dos dados

A Figura 5.5, que se encontra no Anexo A, apresenta a página do *ranking* mundial de clubes, onde os 10 principais clubes líderes do *ranking* atual são destacados. Nesta página, são exibidos o nome do clube, a sua classificação, o país de origem e a pontuação correspondente. Ao todo, 2959 equipas fazem parte deste *ranking*.

A Figura 3.5 apresenta a página que possui informações detalhadas sobre o histórico de classificação de cada clube, proporcionando uma visão abrangente do seu desempenho ao longo do tempo.

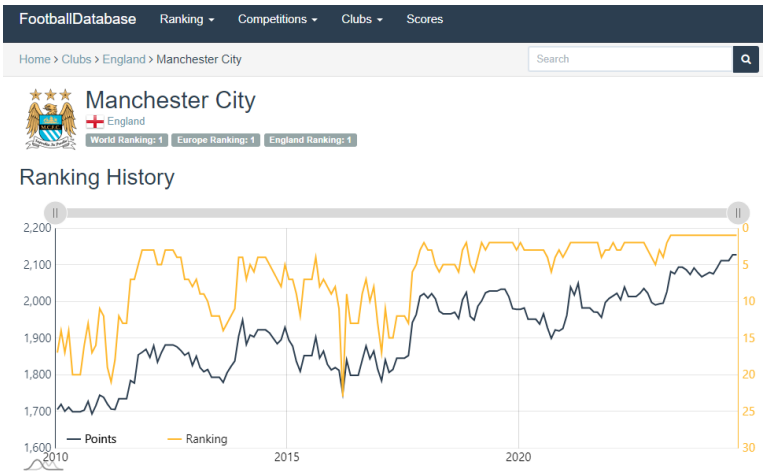


FIGURA 3.5. *Ranking* Mundial de um clube da página *FootballDatabase*.

3.3.3.2. Definição de parâmetros e Navegação no *website*

Foram estabelecidos diversos parâmetros e estruturas para a recolha e processamento de dados do *ranking* mundial de clubes de futebol do site *FootballDatabase*. O código começou com a definição da variável '*base_url*', que contém o endereço base do site, e a configuração dos cabeçalhos HTTP na variável '*headers*', incluindo um *User-Agent* para simular um navegador *web*.

Uma função, denominada por *'parse_footbase_ranking_row'*, foi desenvolvida para extrair informações detalhadas de cada linha da tabela de *ranking*. Esta função extrai para além dos dados básicos como posição, nome do clube e pontos, acedendo também à página individual de cada clube para extrair dados históricos de desempenho. Os dados recolhidos foram armazenados numa lista de dicionários chamada *'fdb_rankings'*, onde cada entrada representa um clube com as suas informações correspondentes.

O *script* foi concebido para iterar por várias páginas do *ranking*, desde a página 1 até à 58, permitindo uma recolha de dados abrangendo um grande número de clubes. Um tratamento específico para os dados históricos de cada clube, incluindo a manipulação de *strings* e datas utilizando expressões regulares, foi desenvolvido para garantir que os dados fossem limpos e formatados para análises futuras.

Para assegurar a robustez do *web scraping*, o código incluiu verificações de estado das requisições HTTP e tratamento de exceções. Após a recolha, os dados foram convertidos num *dataframe* da biblioteca pandas e exportados para um ficheiro Excel

3.3.3.3. Consolidação dos dados

O processo de consolidação dos dados neste código envolveu uma série de etapas específicas para transformar informações complexas sobre o *ranking* mundial de clubes de futebol num formato estruturado e passível de análise. Os dados foram submetidos a

um processo de limpeza e transformação, convertendo formatos de data. A informação foi então estruturada numa lista de dicionários, que posteriormente foi convertida num *dataframe* da biblioteca pandas e exportada para um ficheiro Excel.

Cada entrada do histórico de um clube foi transformada numa linha individual, criando um conjunto de dados longitudinal. Este processo incluiu a conversão de datas para um formato padrão, a eliminação de entradas com datas inválidas e a extração de componentes temporais (ano e mês) para facilitar análises cronológicas. O resultado final foi um *dataframe* detalhado e estruturado, com colunas como Clube, Ano, Mês, *Ranking* e Pontos.

3.3.4. Análise das variáveis

Após o desenvolvimento do *dataframe* obtido do *Web scraping*, foi efetuada a análise das variáveis, com o objetivo de perceber as suas distribuições e *outliers*.

O conjunto de dados resultante foi extenso e diversificado, abrangendo 210 jogadores e mais de 180 variáveis com 4382 observações. Algumas observações continham valores omissos.

O conjunto de dados pode ser organizado em 7 grupos:

- Informações gerais: Nome do jogador, idade, país de origem, posição e pé dominante.
- Estatísticas de jogo: Partidas jogadas, minutos em campo, golos e assistências.
- Métricas avançadas: *Expected Goals*, *Expected Assisted Goals* e progressões com a bola.
- Dados defensivos: *Tackles*, interseções e bloqueios.
- Estatísticas de posse: Toques na bola, passes completos e distância percorrida com a bola.
- Informações de mercado: Valor de mercado e datas de avaliação.
- Dados do clube: *Ranking* e pontuação da equipa.

Nesta etapa do CRISP-DM é aplicada a Análise Exploratória de Dados (EDA).

A compreensão deste conjunto de dados foi fundamental para as fases subsequentes, permitindo identificar padrões, relações entre variáveis que ajudou na modelação e em análises posteriores. Além disso, esta fase ajudou a identificar possíveis desafios, como dados em falta, valores extremos ou inconsistências que necessitaram de ser abordados na fase de preparação dos dados.

A análise à presença de valores em falta permitiu identificar que as variáveis "npxG/Sh", "Dist", "Succ%", e "Tkld%", possuem um maior número de valores nulos, totalizando 2733 valores omissos.

3.3.4.1. Ano de nascimento ("*Birthday_Year*") e Posição ("*Position*")

O ano de nascimento possui uma distribuição relativamente equilibrada entre os anos 2000 e 2003, havendo uma ligeira predominância de jogadores nascidos em 2002 (Ver Figura 3.6), o ano de 2002 destaca-se como o pico da distribuição, com pouco mais de 50

jogadores, seguido de perto pelos anos 2000 e 2003, que apresentam frequências próximas dos 45% jogadores cada. O ano de 2001 mostra uma pequena redução em comparação com os anos adjacentes, mas ainda mantém uma representação significativa. Há uma queda acentuada na frequência de jogadores nascidos em 2004 e 2005.

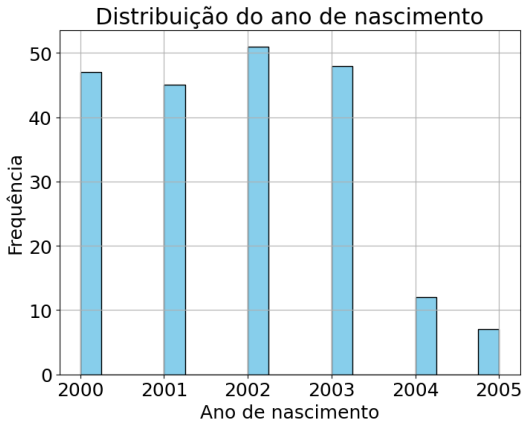


FIGURA 3.6. Distribuição do ano de nascimento.

Como se pode observar na Figura 3.7, os médios (MF) constituem o grupo mais numeroso, com aproximadamente 85 jogadores (40%), os defesas (DF) surgem como o segundo grupo mais representado, com cerca de 70 jogadores (33%), os avançados (FW) formam o terceiro maior grupo, com aproximadamente 55 jogadores (26%). Por fim, os guarda-redes (GK) são, como esperado, o grupo menos numeroso, com apenas cerca de 5 jogadores, o que é consistente com a natureza especializada desta posição e a necessidade limitada de rotação.

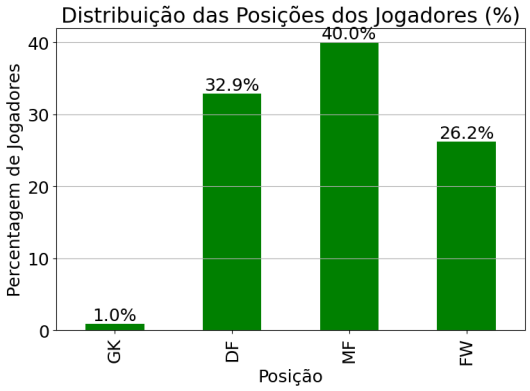


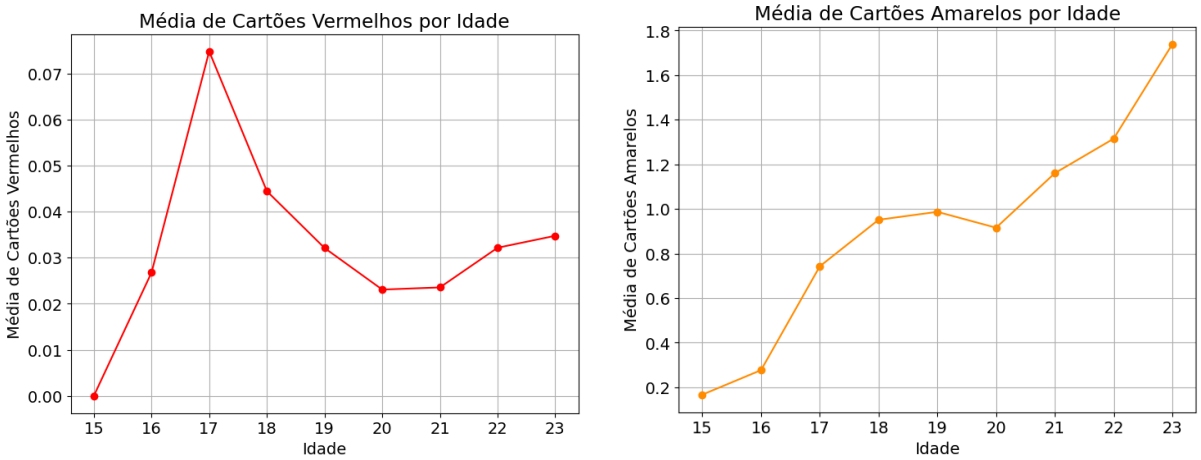
FIGURA 3.7. Distribuição das posições.

3.3.4.2. Cartões vermelhos e cartões amarelos por idade

A distribuição do número de cartões vermelhos (média) por idade, está representada no gráfico da Figura 3.8a. Analisando o gráfico verifica-se uma tendência não linear. Os jogadores com 20 e 21 anos são os que tiveram menos cartões vermelhos, assim como os de

15 que não tiveram nenhum. Esta evolução sugere uma fase crítica aos 17 anos, possivelmente relacionada com a maturidade emocional e a adaptação a níveis mais competitivos do jogo.

Relativamente ao número de cartões amarelos por idade (Figura 3.8b) os jogadores com 22 e 24 anos são os que apresentam um número médio maior que chega a ser de quase 1,8 para os jogadores com 23 anos. Jogadores com 15 e 16 anos em 2005 registam a média mais baixa de todo o período analisado.



(A) Cartões vermelhos por idade. (B) Cartões amarelos por idade.

FIGURA 3.8. Distribuição de ações disciplinares por idade.

3.3.4.3. Idade por posição

A Figura 3.9 compara a distribuição da idade dos jogadores por posição. Os guarda-redes tendem a ser ligeiramente mais velhos, com uma mediana mais elevada e um intervalo interquartil mais compacto, sugerindo uma maior consistência etária nesta posição, apesar de ter *outliers* nos extremos inferiores, o que pode representar talentos precoces. Os defesas e médios mostram distribuições semelhantes, com medianas próximas dos 20 anos e amplitudes interquartis similares. Os avançados apresentam a mediana mais baixa, indicando que esta posição pode favorecer jogadores mais jovens.

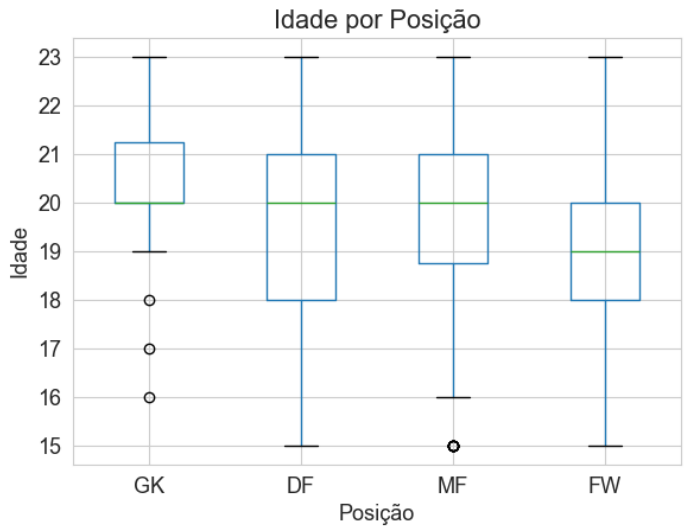


FIGURA 3.9. Distribuição da idade por posição

3.3.4.4. Minutos de jogo por idade

A distribuição dos minutos de jogo é comparada por idade no gráfico apresentado na Figura 3.10. Nos anos iniciais (15-16), as medianas são mais baixas e as caixas mais compactas, indicando menor tempo de jogo e menor variabilidade. A partir dos 17 anos, nota-se um aumento gradual nas medianas e um alargamento das caixas, sugerindo que os jogadores tendem a jogar mais minutos e existe uma maior diversidade no tempo de jogo dos jogadores. As idades entre 20 e 23 anos apresentam as medianas mais altas, próximas dos 60-65 minutos por jogo, e também as maiores amplitudes interquartis, refletindo uma maior variabilidade no tempo de jogo. Os bigodes (*whiskers*) estendem-se consistentemente de 0 a 120 minutos para todas as idades, indicando que em todas as faixas etárias há jogadores que não jogam e outros que jogam jogos completos.

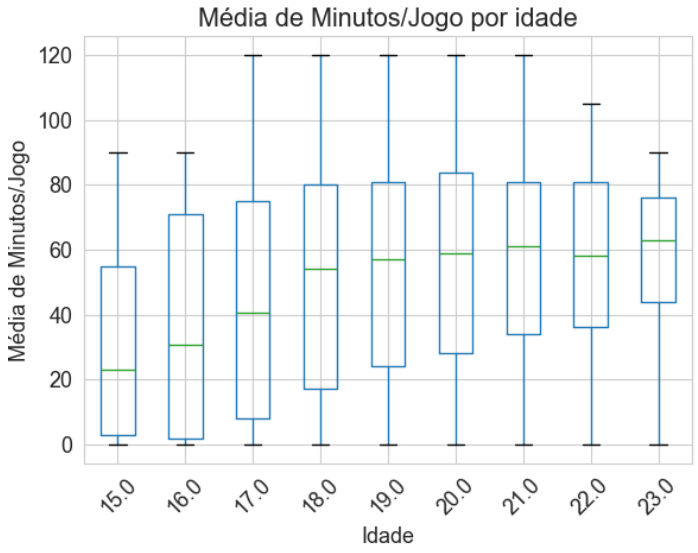
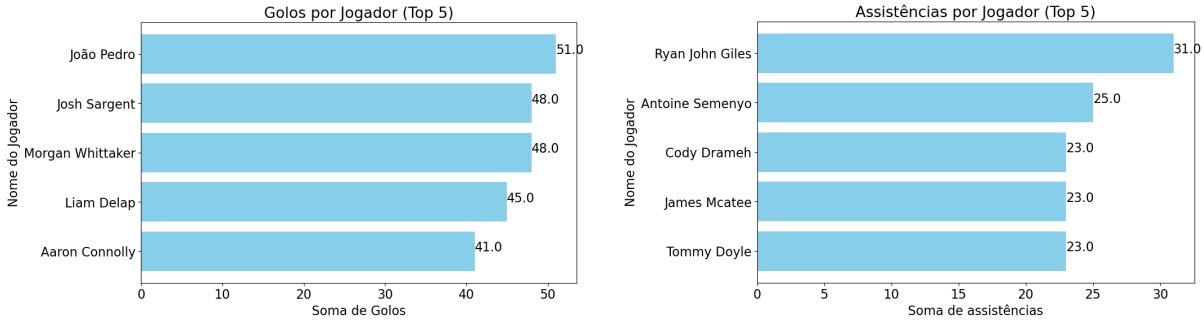


FIGURA 3.10. Média de minutos/jogo por idade.

3.3.4.5. Golos e Assistências por jogador (Top 5)

A Figura 3.11a apresenta o Top 5 dos jogadores com mais golos (em todas as temporadas). O jogador João Pedro lidera com 51 golos, seguido de Josh Sargent e Morgan Whittaker, ambos com 48 golos. Liam Delap ocupa a quarta posição com 45 golos, e Aaron Connolly fecha o top 5 com 41 golos.

A Figura 3.11b apresenta o Top 5 dos jogadores com mais assistências (em todas as temporadas). O jogador Ryan John Giles destaca-se claramente na liderança com 31 assistências, uma marca significativamente superior. Antoine Semenyo ocupa o segundo lugar com 25 assistências. Há um empate triplo na terceira posição entre Cody Drameh, James Mcatee e Tommy Doyle, todos com 23 assistências.



(A) Top 5 de jogadores com mais golos. (B) Top 5 de jogadores com mais assistências.

FIGURA 3.11. Jogadores no Top 5 em golos e assistências.

3.3.4.6. Evolução do valor de mercado dos jogadores (Top5)

O gráfico da Figura 3.12 ilustra a evolução do valor de mercado dos cinco jogadores mais valiosos entre 2017 e 2023, expressos em milhões de euros. As linhas coloridas representam as trajetórias únicas de cada jogador, sendo de realçar a linha vermelha que atinge o pico de 35 milhões em 2023 e corresponde ao jogador João Pedro, sendo este o mais valioso.

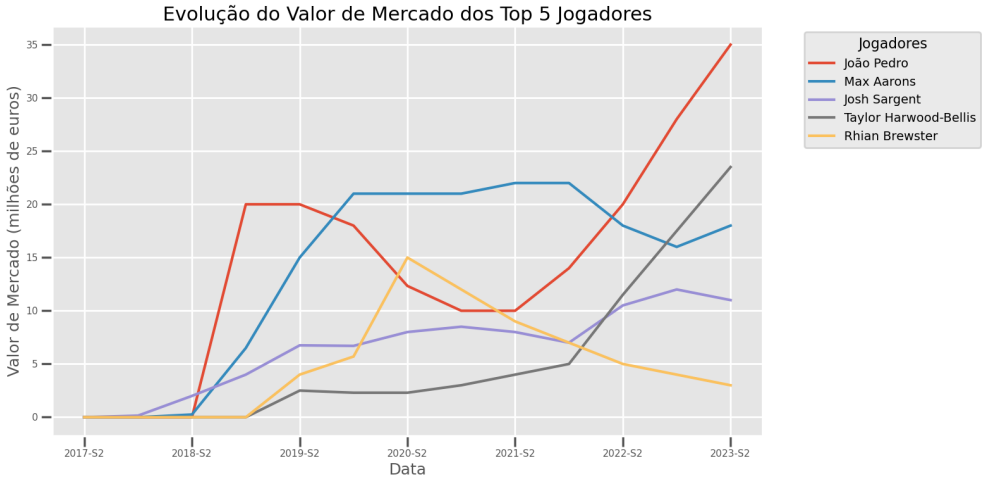


FIGURA 3.12. Top 5 de jogadores com maior alteração no mercado.

3.3.4.7. Valor médio por semestre

A Figura 3.13 retrata a evolução do valor médio de mercado dos jogadores entre 2017 e 2023, revelando uma tendência de crescimento acentuado. A linha vermelha parte de um nível próximo a zero em 2017, mantém-se estável até meados de 2018 e depois mostra um aumento constante e progressivo. O crescimento torna-se mais acentuado a partir de 2022, culminando num valor médio de cerca de 2,8 milhões em 2023, destacando uma valorização significativa dos jogadores ao longo do período analisado.

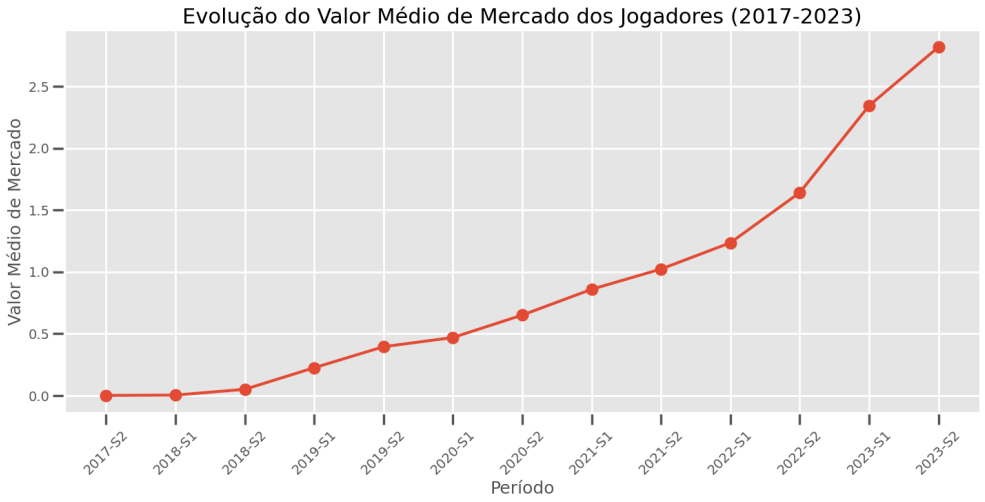


FIGURA 3.13. Distribuição da média de valor de mercado dos jogadores por semestre.

3.4. CRISP-DM Fase 3: Preparação dos dados

A etapa de preparação de dados é crucial para um cientista de dados, pois é o elo entre os dados originais e a aplicação de modelos. É o ponto de partida para diversas análises e experiências de dados, envolvendo análises detalhadas e aprofundadas de conjuntos de dados, revelando diversos padrões. O processo de tratamento e normalização dos dados constituiu um pilar fundamental na preparação do projeto, visando a criação de um conjunto de dados coeso, preciso e altamente informativo para análises aprofundadas no âmbito do futebol profissional. Este processo abrangeu quatro áreas cruciais: o *ranking* de clubes, os dados do *Transfermarkt*, a informação do FBref e a fusão integrada destas fontes diversas.

No que concerne ao *ranking* de clubes, a prioridade recaiu sobre a simplificação e uniformização das designações. Esta etapa foi para assegurar a consistência e comparabilidade das informações. Implementou-se uma função que normalizou as denominações, convertendo-as para minúsculas, eliminando pontuações e removendo prefixos comuns como 'FC', 'AFC' e 'SL'. Paralelamente, identificou-se sufixos indicativos de categorias específicas, como 'U21' ou 'Reservas', permitindo uma categorização precisa dos clubes em diferentes níveis competitivos. Esta abordagem não só resolveu discrepâncias na nomeação, mas também facilitou o agrupamento correto de dados referentes a um mesmo

clube, mesmo quando apresentados com variações nominais. A introdução de uma coluna adicional para o escalão do clube permitiu análises segmentadas, diferenciando claramente entre equipas principais e categorias de formação.

No tratamento dos dados do *Transfermarkt*, a normalização das posições dos jogadores assumiu um papel central. Diversas variantes e abreviações foram consolidadas em quatro categorias padrão: GK (guarda-redes), DF (defesa), MF (médio) e FW (avançado). Esta uniformização foi essencial para possibilitar comparações consistentes entre jogadores de diferentes equipas e ligas. Utilizou-se um dicionário predefinido para preencher valores em falta na coluna "*Position*", baseando-se em conhecimento prévio sobre os atletas, evitando assim métodos de imputação genéricos.

As épocas desportivas foram normalizadas para um formato consistente "AAAA-AAAA", crucial para análises temporais precisas e comparações entre épocas. Um aspeto significativo envolveu a conversão dos valores de mercado dos jogadores para um valor numérico padrão em milhões de euros. Este processo lidou com diversos formatos monetários, removendo símbolos e interpretando corretamente sufixos como 'k' para milhares e 'm' para milhões. O resultado foi uma coluna "*market_value_in_million_euros*" que facilitou comparações diretas e análises quantitativas precisas.

A transformação dos dados do FBref iniciou-se com uma análise da dimensão e diversidade do conjunto de dados. A normalização das posições dos jogadores foi realizada através de um dicionário de mapeamento elaborado, simplificando designações complexas como por exemplo 'DF-MF' para categorias básicas como 'DF'. O tratamento das épocas desportivas garantiu um formato consistente, essencial para análises temporais precisas e para evitar ambiguidades nas referências às temporadas.

Uma parte significativa do processo foi dedicada ao preenchimento de dados em falta, particularmente no que diz respeito às posições dos jogadores e aos identificadores de transferência. Utilizou-se dicionários predefinidos para preencher estas lacunas, melhorando substancialmente a integridade e a utilidade do conjunto de dados.

Esta etapa da fusão final de todas estas fontes, combinou informações de desempenho dos jogadores com as suas avaliações de mercado, utilizando várias colunas como chaves de correspondência, incluindo identificadores de chaves de jogadores, temporadas e informações demográficas. Após a junção, efetuaram-se operações de limpeza adicionais, removendo colunas duplicadas, renomeando colunas para manter a consistência e eliminando informações redundantes ou supérfluas.

O resultado final foi um conjunto de dados estruturado de forma lógica e pronto para suportar análises sofisticadas no domínio do futebol profissional. A atenção ao detalhe na limpeza e organização dos dados assegurou não apenas a fiabilidade das análises imediatas, mas também abriu caminhos para modelações mais sofisticadas, para previsões do valor de mercado e de desempenho dos jogadores.

3.4.1. Tratamento de Valores Omissos

O processo de limpeza e preparação dos dados iniciou-se com uma estratégia de substituição de valores NaN. Especificamente, nas colunas designadas como estatísticas de jogo do tipo numéricas, os NaN foram substituídos por zero, mas apenas quando o valor da coluna "MP" (Nº de jogos disputados) é nulo. Esta nuance foi crucial, pois pressupõe que a ausência de estatísticas para um jogador que não participou em nenhum jogo é logicamente representada por zeros, em vez de dados em falta.

A atenção aos detalhes continuou com o tratamento particular da coluna "*Country*". A decisão de preencher os valores ausentes com "*Europe Comp*" sugeriu uma categorização, para agrupar as competições europeias sob uma única designação, facilitando assim análises comparativas futuras. No que toca às colunas relacionadas com valores monetários, nomeadamente "*Price*" e "*market_value_in_million_euros*" decidiu-se a substituição de valores nulos por zero. Esta abordagem permitiu incluir jogadores sem valor de mercado definido nas análises, assumindo conservadoramente que a sua valorização é nula.

A função personalizada "*adjust_market_value*", ao utilizar a informação da época ("*Season*") para inferir e preencher dados ausentes relacionados com o valor de mercado, o código não só eliminou valores nulos, mas também adicionou consistência temporal aos dados. A escolha de julho como mês de referência para a atualização dos valores de mercado alinhou-se com os ciclos típicos de transferências no futebol europeu.

Identificou-se uma quantidade significativa de clubes sem correspondência de *ranking* e pontos. Para abordar essa questão, o código implementou uma estratégia de preenchimento focada em 80 clubes específicos listados em '*clubs_to_update*'. Para estes clubes, quando o *ranking* é ausente, atribui-se o valor 2906 (último lugar na classificação do ranking), e relativamente à ausência de pontos, preencheu-se com 1154 (maior número de pontos que existe no ranking do *FootballDatabase*). Essa técnica garante que esses clubes tenham valores consistentes, evitando problemas em análises posteriores.

A utilização da classe *SimpleImputer* da biblioteca *scikit-learn* representou uma abordagem para o preenchimento dos valores em falta remanescentes. A aplicação da média para dados numéricos e do valor mais frequente para dados categóricos assegurou que os valores imputados fossem representativos do conjunto de dados como um todo. O resultado final, condensado na estrutura "df_*all*", demonstrou ser um conjunto de dados livre de valores nulos, resolvendo os problemas associados a dados em falta.

3.4.2. Tratamento das variáveis

Após uma análise das variáveis, foi identificado um conjunto de colunas: '*Matches*', 'Ast_*pass*', 'xAG_*pass*', 'Gls_*shoot*', 'PK_*shoot*', 'PKatt_*shoot*', 'npxG_*shoot*', 'xG_s', 'PrgR_*poss*', 'PrgC_*poss*' e 'PrgP_*pass*'. Estas variáveis duplicadas foram removidas, a fim de aprimorar a qualidade e relevância dos dados para a análise em questão.

3.4.3. Criação de novas varáveis

3.4.3.1. Nova variável "*Progress*"

Uma nova variável foi criada, "*Progress*", de acordo com um conjunto de critérios de progressão, nomeadamente minutos jogados, o número de jogos, classificação do clube e estatísticas gerais por equipa e específicas por posição. Mediante as características de um jogador e em função de cada critério foram atribuídos pontos a cada jogador com o objetivo da sua classificação em termos de evolução na época relativamente à anterior.

Os critérios de progressão estão apresentados na tabela 3.1. A tabela 3.2 apresenta as estatísticas específicas por posições e a tabela 3.3 apresenta as classes de pontuação que permitem a classificação da progressão de um jogador em várias categorias, desde "Regrediu" a "Progrediu significativamente". A época inaugural de um jogador num clube é classificada como "Manteve-se".

Tabela 3.1. Critérios de Progressão

Critério	**Pontuação**
Minutos jogados e número de jogos	
Aumento > 10%	+1 ponto
Diminuição > 10%	-1 ponto
Classificação do clube	
Melhoria > 10%	+2 pontos
Qualquer melhoria	+1 ponto
Regressão > 10%	-2 pontos
Qualquer regressão	-1 ponto
Estatísticas gerais por equipa	
Golos marcados (onG) aumento > 10%	+1 ponto
Golos marcados (onG) diminuição > 10%	-1 ponto
Golos sofridos (onGA) diminuição > 10%	+1 ponto
Golos sofridos (onGA) aumento > 10%	-1 ponto
Estatísticas específicas por posição	
Melhoria > 10% em qualquer estatística	+1 ponto (cada)
Regressão > 10% em qualquer estatística	-1 ponto (cada)

Tabela 3.2. Estatísticas Específicas por Posição

Posição	**Estatísticas Específicas**
Avançados (FW)	Golos, assistências, penáltis
Médios (MF)	Passes completos, passes para dentro da área, passes progressivos
Defesas (DF)	Desarmes bem-sucedidos, percentagem de desarmes, bloqueios de remates

Tabela 3.3. Categorias de Progressão

Pontos Acumulados	**Categoria de Progressão**
6 ou mais	Progrediu Significativamente
3 a 5	Progrediu
1 a 2	Progrediu Ligeiramente
0	Manteve-se
Negativo	Regrediu

Analisemos um exemplo para melhor compreensão deste sistema de classificação de progressão de um jogador. A análise da progressão de Álvaro Carreras, de acordo com o sistema de pontuação, apresentou variações ao longo das épocas. Na temporada 2020-2021, atuou principalmente na equipa júnior do *Manchester United*, mantendo um desempenho estável. A época 2021-2022 registou uma regressão, pois apesar de manter presença na equipa júnior, não obteve minutos na equipa principal da *Premier League*. Em 2022-2023, Carreras transferiu-se para o *Preston* na *Championship*, onde disputou 39 partidas, resultando numa ligeira progressão devido ao aumento significativo de tempo de jogo e nível competitivo. A época 2023-2024 resultou numa avaliação de regressão. Esta classificação fundamenta-se em vários indicadores: diminuição significativa dos minutos jogados, transitando de 39 jogos na Championship para apenas 27 distribuídos por três equipas distintas (*Manchester United*, Granada e Benfica). Verificou-se um declínio nas estatísticas individuais, correlacionado com a redução do tempo de jogo. Adicionalmente, a mudança para clubes com classificações inferiores ao *Manchester United* influenciou negativamente a avaliação. Este cenário realça a relevância da estabilidade e da regularidade competitiva no desenvolvimento de um jogador, mesmo quando ocorrem transições para clubes de maior prestígio.

3.4.3.2. Variável "*Market_value_in_million_euro*"

Uma nova coluna, denominada '*market_value_in_million_euro*', foi criada para refletir o valor de mercado dos jogadores em milhões de euros. Essa coluna foi derivada da coluna '*Price*', que continha os valores já convertidos em números inteiros. O processo de criação envolveu a divisão de cada valor da coluna '*Price*' por 1.000.000, convertendo assim os valores de euros para milhões de euros. Essa abordagem proporcionou uma representação compreensível dos valores de mercado dos jogadores, facilitando comparações e análises subsequentes dos dados.

3.4.3.3. Nova Variável "*Semester*"

A criação da variável 'Semester' desempenhou um papel na etapa de preparação dos dados para análise de séries temporais. Essa variável transformou as datas específicas de avaliação de mercado dos jogadores em períodos semestrais, proporcionando um nível intermediário de detalhe temporal. A função 'obter_semestre' foi desenvolvida para converter cada data num formato semestral padronizado, como "2023-S1" para o primeiro semestre de 2023 e "2023-S2" para o segundo semestre. Essa abordagem possibilitou

capturar padrões sazonais nos valores de mercado dos jogadores, os quais podem ser influenciados por fatores como janelas de transferências ou desempenho em competições importantes que ocorrem em momentos específicos do ano. Ao agrupar os dados em semestres, conseguiu-se reduzir o ruído decorrente de flutuações de curto prazo, mantendo, ao mesmo tempo, uma resolução temporal adequada para identificar mudanças significativas ao longo do tempo. Essa variável facilitou a identificação de padrões cíclicos, tendências de longo prazo e possíveis pontos de viragem na valorização dos jogadores, fornecendo uma base sólida para análises comparativas e previsões futuras.

3.4.4. Criação de um novo *dataframe* para séries temporais

O novo *dataframe* representou uma transformação significativa dos dados originais, otimizando-os para análise de séries temporais. Esse *dataframe* contém 214 jogadores e foi construído agrupando os dados por jogador e semestre, calculando a média do valor de mercado para cada período. A estrutura resultante exibe os jogadores nas linhas e os semestres nas colunas, oferecendo uma visualização da evolução do valor de mercado de cada jogador ao longo do tempo. Uma característica relevante desse novo *dataframe* foi a gestão dos valores em falta. Criou-se uma função para lidar com lacunas nos dados, substituindo valores nulos (NaN) por zeros antes do primeiro valor registado e preenchendo lacunas intermédias com interpolação linear. Esse procedimento garantiu a continuidade das séries temporais, um aspeto crucial para muitas técnicas de análise. Ademais, o preenchimento dos valores em falta do segundo semestre de 2023 com os do primeiro semestre assegurou a integridade dos dados até ao final do ano. A exclusão da coluna referente ao primeiro semestre de 2024 refletiu a decisão de trabalhar apenas com dados completos e fiáveis. Essa estruturação dos dados facilitou a visualização de tendências individuais e comparações entre jogadores, ajudando na melhoria da aplicação de técnicas avançadas de análise de séries temporais.

Para aplicação dos modelos foi necessário que todas as variáveis em estudo fossem estacionárias, ou seja, possuam média constante, variância finita e constante, e covariância constante. A existência de tendência e/ou sazonalidade pode originar uma série não ser estacionária. Para testar a estacionariedade da variável recorreu-se aos testes de raiz unitária *Augmented Dickey–Fuller* (ADF) e *Phillips–Perron* (PP), sendo, $H_0 : \rho = 1$ (processo não estacionário), $H_1 : \rho < 1$ (processo estacionário). É desejável rejeitar a hipótese nula (H_0) que ocorre quando $p - \text{value} < \alpha$, sendo α o nível de significância, normalmente considera-se $\alpha = 5\%$.

Como os testes ADF e PP foram discordantes, recorreu-se ao teste de estacionariedade de *Kwiatkowski–Phillips–Schmidt–Shin* (KPSS), sendo, H_0 : série estacionária, H_1 : série não estacionária. É desejável não rejeitar a hipótese nula (H_0) que ocorre quando $p - \text{value} > \alpha$.

Foram realizados os testes a cada jogador para verificar se as séries eram estacionárias, os valores obtidos do p-value do jogador Aaron Rowe estão presentes na Tabela 3.4.

Tabela 3.4. Significância dos testes de raiz unitária (ADF e PP) e teste de estacionariedade (KPSS) para Aaron Rowe.

Jogador	ADF	PP	KPSS	Conclusão
Aaron Rowe	0.7	0.8	0.028167	Série não estacionária

Como a série temporal foi não estacionária, aplicou-se a primeira diferença para tentar estacionarizar a série. Recorreu-se novamente aos testes da raiz unitária para verificar se a série era estacionária. Os resultados obtidos do p-value do jogador Aaron Rowe apresentam-se na Tabela 3.5.

Tabela 3.5. Testes de raiz unitária (ADF e PP) e teste de estacionariedade (KPSS) com a primeira diferença.

Jogador	ADF	PP	Conclusão
Aaron Rowe	0.0253	0.006	Série estacionária

Como o teste ADF e PP não discordaram, não foi necessário realizar o teste de estacionariedade KPSS. A série do jogador Aaron Rowe foi considerada estacionária, e por isso, procedeu-se à modelação.

3.4.5. Relações entre varáveis

Com base no mapa de correlação apresentado na Figura 3.14, é possível observar um padrão complexo de relações entre as diversas variáveis numéricas analisadas. Este mapa, utilizando uma escala cromática que varia do azul-escuro (correlação positiva forte) ao amarelo-claro (correlação negativa forte), revelou informações significativas para a análise em questão.

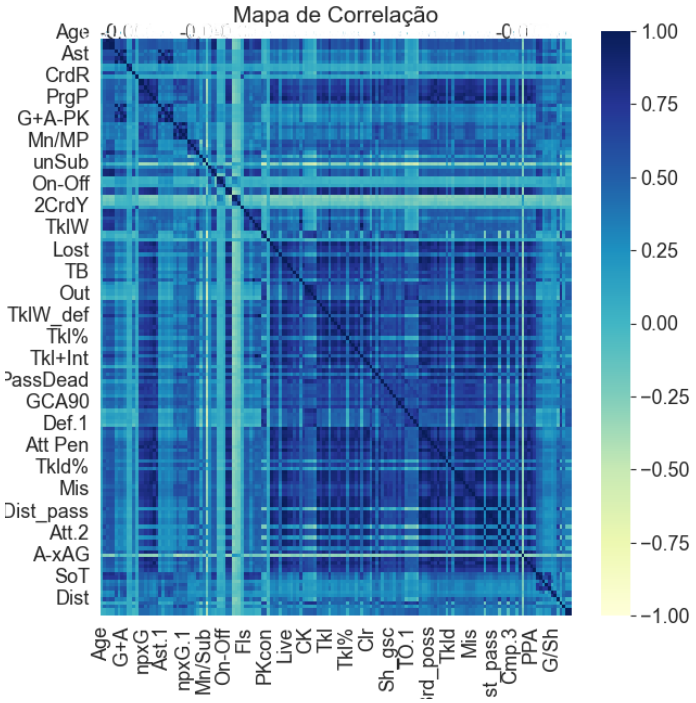


FIGURA 3.14. Mapa de correlação.

Notavelmente, observa-se *clusters* de variáveis com correlações fortes entre si, sugerindo a existência de grupos de métricas relacionadas. Por exemplo, verificou-se uma forte correlação positiva entre as variáveis "G+A-PK", "Ast" e "PrgP", indicando uma possível relação entre a idade dos jogadores, os golos marcados e as grandes penalidades atribuídas.

Adicionalmente, foi possível identificar variáveis que apresentam correlações negativas com outras, como é o caso de "Def.1" e "G+A", sugerindo uma relação inversa entre ações defensivas na grande área e ações ofensivas.

É importante salientar que algumas variáveis, como "On-Off" e "2CrdY", evidenciam correlações moderadas a fortes com diversas outras métricas, o que pode indicar a sua relevância como indicadores abrangentes do desempenho dos jogadores.

3.4.6. PCA

Na análise dos resultados da Análise de Componentes Principais (PCA), observou-se uma decomposição da variabilidade dos dados, para a determinação do número ideal de componentes a reter.

O gráfico da Figura 3.15 apresenta os resultados de uma análise de Componentes Principais (PCA), mostrando a dispersão dos dados num espaço bidimensional definido pelas duas primeiras componentes principais. A visualização revela uma concentração significativa de pontos na região inferior esquerda, aproximadamente entre -10 e 10 em ambos os eixos. A Componente Principal 1, representada no eixo horizontal, parece capturar uma maior variabilidade nos dados, com pontos a espalharem-se de -10 a 50.

Por outro lado, a Componente Principal 2, no eixo vertical, demonstra uma dispersão mais limitada, variando de cerca de -30 a 30. Observa-se a presença de vários valores atípicos, especialmente na direção positiva da Componente 1 e em ambas as direções da Componente 2.

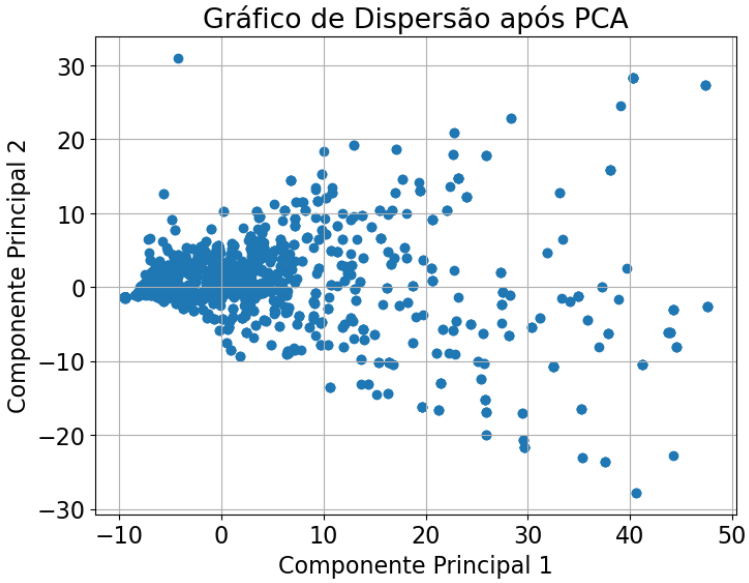


FIGURA 3.15. 2 componentes principais.

Os gráficos da Figura 3.16 e da Figura 3.17 apresentam o top 20 das variáveis que contribuíram para a primeira componente principal e segunda componente principal respetivamente. As 20 variáveis que mais contribuem para a primeira componente principal (PC1) estão ordenadas pela magnitude da sua contribuição, representada pelas barras azuis. As contribuições variam entre aproximadamente 0,11 e 0,12, indicando uma distribuição relativamente uniforme entre as variáveis apresentadas. As cinco variáveis com maior contribuição são "onxG", "Recov", "*Touches*", "*Live_pass*" e "*Carries*". Observa-se que as variáveis incluem diversos aspetos do jogo, como estatísticas ofensivas (por exemplo, "onxG"), defensivas (como "Recov") e de posse de bola (como "*Touches*" e "*Carries*"). Esta distribuição sugere que a PC1 captura uma visão abrangente do desempenho dos jogadores, incorporando múltiplos aspetos do jogo.

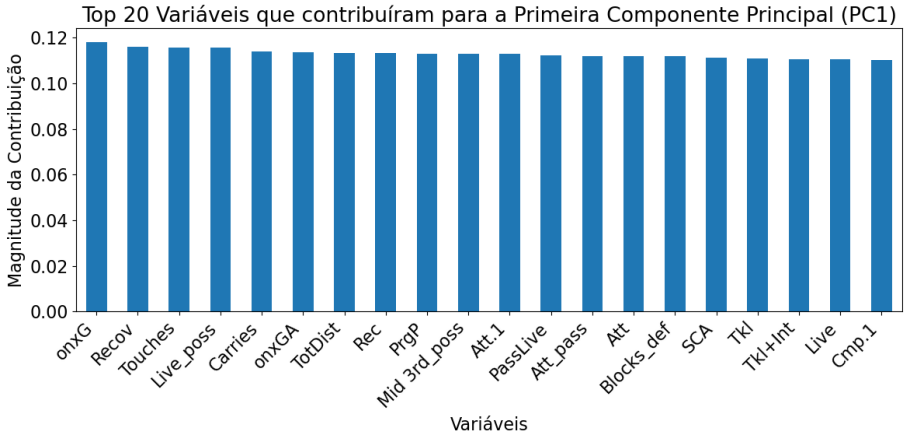


FIGURA 3.16. Variáveis que contribuíram para a primeira componente principal.

As 20 variáveis que mais contribuem para a segunda componente principal (PC2) variam entre aproximadamente 0,12 e 0,16, com uma diminuição gradual da esquerda para a direita. As cinco variáveis com maior contribuição são "xG", "npxG", "Def 3rd_pass", "Gls" e "G+PK". Nota-se uma predominância de variáveis relacionadas a golos e a ataques (como "xG", "npxG", "Gls"). Esta distribuição sugere que a PC2 pode estar a capturar principalmente aspetos ofensivos do jogo.

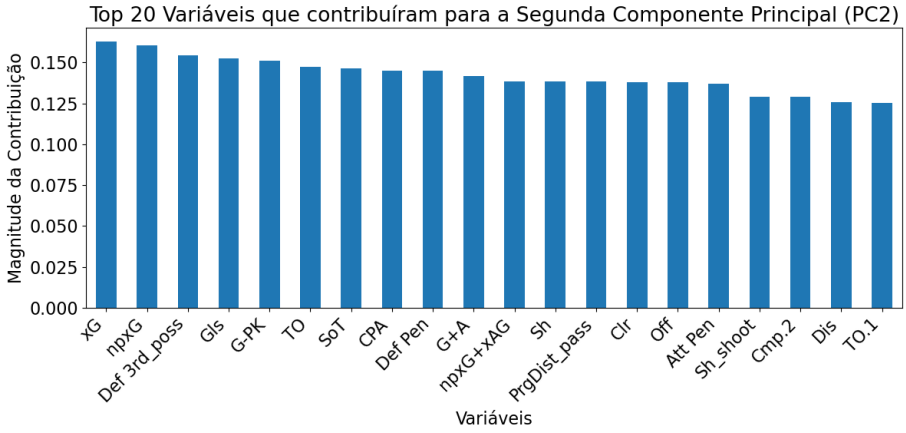


FIGURA 3.17. Variáveis que contribuíram para a segunda componente principal.

A Figura 3.18 apresenta cinco componentes principais, explicando um total de 62% da variância dos dados. A distribuição da variância explicada por cada componente é notavelmente desigual, com a primeira componente a capturar cerca de 42% da variância total, seguido por um decréscimo acentuado para as componentes subsequentes (aproximadamente 9%, 5%, 4% e 3%, respetivamente).

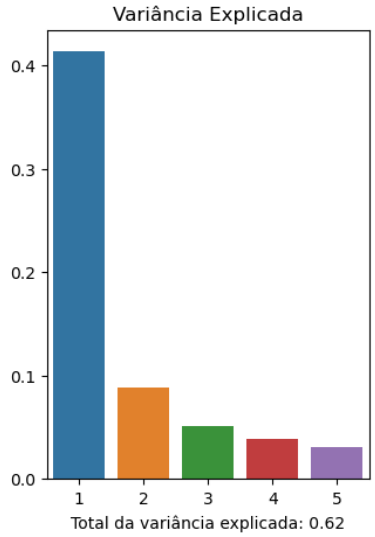


FIGURA 3.18. 5 componentes principais.

A determinação do número ideal de componentes principais a reter é uma decisão crucial que deve equilibrar a simplificação do modelo com a preservação de informação relevante. Considerando os resultados apresentados, chegou-se à conclusão que a retenção de quatro ou cinco componentes principais oferecia um compromisso adequado entre a redução de dimensionalidade e a preservação de informação significativa.

A escolha entre quatro ou cinco componentes pode ser justificada pela observação de que a quinta componente (Figura 3.18) contribuiu com um aumento na variância explicada (de 59% para 62%).

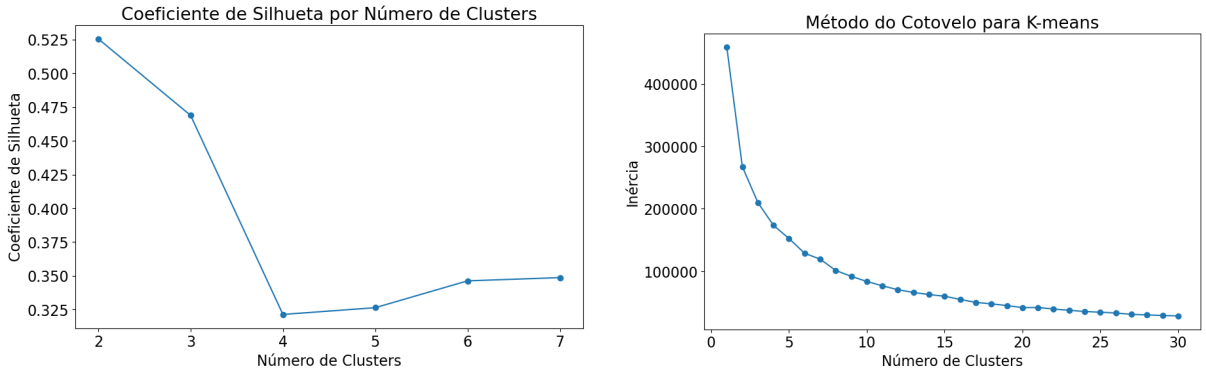
Adicionalmente, a aplicação da regra empírica de Kaiser, que sugere a retenção de componentes com valores próprios superiores a 1, corrobora a decisão de reter quatro ou cinco componentes. Considerando a análise apresentada e os resultados observados, a decisão de reter cinco componentes principais resultou como a escolha ideal para este estudo.

3.4.7. *Clusters*

Na análise dos resultados dos *clusters* apresentados utilizaram-se as componentes principais como variáveis de base.

As Figuras 3.19a e Figura 3.19b, apresentam dois métodos complementares, para determinar o número de *clusters*. O Coeficiente de Silhueta por número de *clusters*, uma métrica que avalia a qualidade da separação entre *clusters*. O gráfico revela um padrão interessante, com o coeficiente de silhueta atingindo o seu pico máximo de aproximadamente 0,525 quando são considerados 2 *clusters*. Observa-se um declínio acentuado do coeficiente à medida que o número de *clusters* aumenta, com uma estabilização relativa a partir de quatro *clusters*. Por outro lado, a Figura 3.19b ilustra o Método do Cotovelo para *K-means*, uma técnica que visa identificar o ponto onde o aumento do número de *clusters* deixa de proporcionar ganhos significativos na redução da inércia (variância

intra-*cluster*). O gráfico expressa uma curva característica, com uma queda acentuada na inércia para os primeiros *clusters*, seguida de uma estabilização gradual.



(A) Nº de *clusters* com coeficiente de Silhueta. (B) Nº de *clusters* com *K means*.

FIGURA 3.19. Avaliação do nº ideal de *clusters*.

A interpretação conjunta efetuada a estas figuras sugere que a escolha do número ideal de *clusters* deve equilibrar a coesão interna dos grupos com a separação entre eles. O Coeficiente de *Silhueta* favorece claramente uma solução com 2 *clusters*, indicando que esta configuração proporciona a melhor separação entre grupos. Contudo, é crucial considerar que uma solução tão parcimoniosa pode não capturar adequadamente a complexidade inerente aos dados, especialmente em contextos multidimensionais.

O Método do Cotovelo, por sua vez, não apresentou um "cotovelo" nítido, mas sugere que o ganho na redução da inércia diminui consideravelmente após 4-5 *clusters*. Esta observação alinha-se com o ponto de inflexão notado no gráfico do Coeficiente de Silhueta.

Tendo em consideração o Coeficiente de *Silhueta* e o Método do Cotovelo, conclui-se que uma solução entre 3 a 5 *clusters* seria a mais apropriada. Assim, com base nesta análise, o número de *clusters* escolhido para este estudo foi de 4 *clusters*, permitindo uma análise mais aprofundada e significativa dos agrupamentos identificados.

A análise dos mapas fatoriais apresentados na Figura 5.6 que se encontra no Anexo C, ofereceu uma perspetiva sobre a estruturação dos jogadores de futebol em quatro *clusters* distintos, utilizando a primeira componente principal (PC1) como base de comparação em relação ao número total de minutos jogados.

O *Cluster* 1 (rosa) e o *Cluster* 3 (verde) ocuparam predominantemente as posições superiores do eixo vertical, indicando jogadores com elevado número de minutos jogados. O *Cluster* 4 (laranja) concentrou-se na parte inferior do gráfico, representando jogadores com menor tempo de jogo. O *Cluster* 2 (azul-claro) apresentou uma distribuição mais dispersa, mas tende a ocupar posições intermediárias.

A consistência na distribuição dos *clusters* da Figura 5.6 sugere que a escolha de quatro grupos foi adequada para capturar as principais diferenças entre os jogadores em termos de minutos jogados.

3.5. CRISP-DM Fase 4: Modelação

3.5.1. Estudo sobre a análise de progressão de jogadores jovens

Modelos de Classificação

A fim de prever a progressão dos jogadores de futebol, foram testados vários modelos de aprendizagem automática, incluindo *Random Forest*, *Logistic Regression*, *Support Vector Machine* (SVM), *Gradient Boosting* e Redes Neuronais, com base no estudo do investigador Ayodele (2010).

O processo teve início com a preparação dos dados, onde as características foram divididas em numéricas e categóricas. Foi aplicado o *StandardScaler* e o *OneHotEncoder*, respetivamente. Os dados foram então divididos em conjuntos de treino e teste numa proporção de 80/20. A técnica de *SMOTE* foi utilizada para equilibrar as classes no conjunto de treino, e os dados foram ajustados conforme necessário para cada modelo.

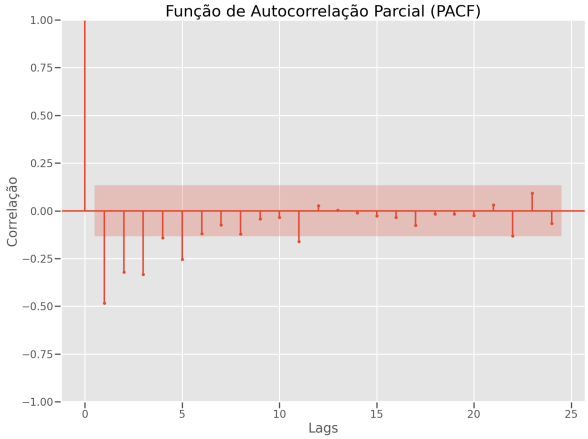
O modelo *Random forest*, que consiste num conjunto de árvores de decisão, foi configurado com 100 estimadores. É reconhecido pela sua capacidade de lidar com características não lineares e interações complexas. A *Logistic Regression*, um modelo linear para classificação binária, foi configurada com um máximo de 1000 iterações. O SVM, que procura o hiperplano ideal para separar as classes, foi incluído devido à sua eficácia em espaços de alta dimensionalidade. Este modelo pode ser especialmente útil quando as relações entre as características são complexas. O *Gradient Boosting*, que consiste num conjunto de árvores de decisão construídas sequencialmente, também foi utilizado. Seguidamente foram utilizados os modelos *Recurrent Neural Network* (RNN), *Gated Recurrent Unit* (GRU) e *Long Short-Term Memory* (LSTM). A *Recurrent Neural Network* (RNN) apresentou uma arquitetura com três camadas densas de 64, 32 e 5 neurónios, respetivamente, incorporando camadas de *dropout* com taxa de 0,3 e utilizando funções de ativação *ReLU* e *softmax*. O modelo foi otimizado com o algoritmo Adam, utilizando uma taxa de aprendizagem de 0,001 e implementando *early stopping* após 5 épocas sem melhoria. O treino foi realizado utilizando *batch size* de 32 amostras e um máximo de 100 épocas, reservando 20% dos dados para validação. A *Gated Recurrent Unit* (GRU) destaca-se pela sua capacidade de capturar dependências de longo prazo, esta apresentou uma arquitetura com duas camadas GRU de 64 e 32 unidades, intercaladas com camadas de *dropout*, seguidas de uma camada densa de saída. A otimização e o treino seguiram parâmetros similares à RNN, com a adição de um redimensionamento dos dados de entrada e um do *early stopping* para 10 épocas. Por fim, a *Long Short-Term Memory* (LSTM) apresentou uma estrutura semelhante à GRU, mas com a vantagem adicional de considerar a dimensão temporal dos dados e manter um erro mais constante, permitindo uma aprendizagem contínua ao longo de muitos dados futuros. Todos os modelos utilizaram a função de perda de entropia cruzada categórica e foram configurados para classificação multiclasse com 5 categorias. As técnicas de *dropout* e *early stopping* foram empregues em todos os modelos de redes neuronais para prevenir o sobreajustamento e otimizar o processo de treino.

3.5.2. Estudo sobre a análise do valor de mercado dos jogadores ao longo das diferentes épocas

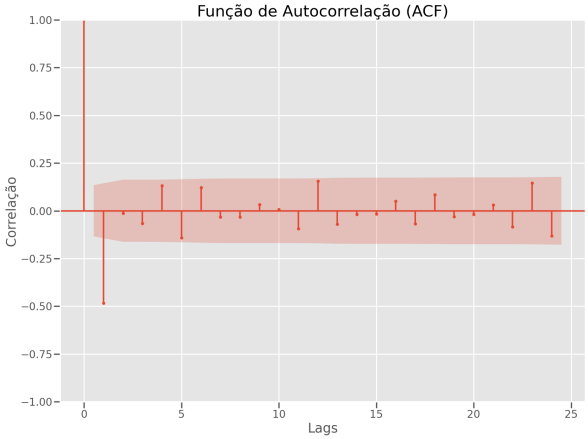
Séries temporais

Neste capítulo foram aplicados modelos aos dados para prever o valor de mercado dos jogadores nomeadamente, o *SARIMA*, LSTM, GRU, RNN e o modelo híbrido (*SARIMA* e LSTM) de acordo com o estudo dos autores Masini et al. (2023).

Os dados foram divididos em conjunto de treino e de teste. Os dados de treino tiveram observações compreendidas desde do segundo semestre de 2017 ao primeiro semestre de 2023. O conjunto de teste foi constituído com os dados do semestre de 2023. Para estimar o melhor modelo *SARIMA* foi necessário estimar os parâmetros (p,d,q)(P,D,Q)[S], para este efeito recorreu-se à função *auto_arima* onde se indica como parâmetros informação sobre a série como por exemplo, sazonalidade, diferenciação e número de *lags* significativos. Para indicar os valores que melhor se ajustam a p e q foram realizados os gráficos da função de autocorrelação (ACF) e autocorrelação parcial (PACF) a todos os jogadores. A título exemplificativo, as Figuras 3.20b e 3.20a apresentam ACF e PACF para o jogador Aaron Connolly, respetivamente.



(A) Autocorrelação Parcial (PACF) para a série diferenciada de ordem 1.



(B) Autocorrelação (ACF) para a série diferenciada de ordem 1.

FIGURA 3.20. Autocorrelação para a série diferenciada de ordem 1 do jogador Aaron Connolly

No gráfico da ACF (Figura 3.20b) observa-se um decaimento exponencial o que indica que $q = 0$. No gráfico da PACF (Figura 3.20a) existem um *lag* significativo, por isso, este é o número máximo de *lags* a considerar para p. Consideramos $d = 0$ visto que já foi aplicada a primeira diferença à série. Logo, para o jogador Aaron Connolly, ao aplicar a função *auto_arima* à série diferenciada com as características mencionadas, foi obtido como melhor modelo *SARIMA* (1, 0, 1)(0, 1, 0)[2].

No âmbito das redes neuronais, a preparação dos dados iniciou-se com a aplicação da transformação logarítmica (log1p) para abordar possíveis assimetrias na distribuição.

Aplicou-se regularização L2 em todas as camadas dos modelos para melhorar a sua capacidade de generalização.

A arquitetura da *Recurrent Neural Network* (RNN) começou com uma camada de entrada densa de 64 neurónios e função de ativação ReLU, seguida por uma camada de *dropout* de 20%. Segue-se uma segunda camada densa de 32 neurónios com ativação ReLU e outro *dropout* de 20%. Uma terceira camada densa de 16 neurónios precede a camada de saída com um único neurónio. Utilizou-se o Erro Quadrático Médio (*Mean Squared Error*, MSE) como função de custo e o otimizador Adam. O treino decorre durante 100 épocas com um *batch size* de 32, implementando-se um mecanismo de paragem antecipada para evitar o sobreajustamento.

O modelo *Gated Recurrent Unit* (GRU) foi composto por três camadas GRU (64, 32 e 16 unidades, respetivamente), cada uma seguida por uma camada de *dropout* de 20%. A arquitetura do modelo *Long Short-Term Memory* (LSTM) foi semelhante, substituindo as camadas GRU por LSTM. Ambos os modelos utilizaram MSE como função de custo, o otimizador Adam, e foram treinados durante 100 épocas com um *batch size* de 32, incluindo um mecanismo de paragem antecipada, o *early stopping*.

A convergência da função de custo sugere que não foi necessário aumentar o número de épocas para os modelos apresentados.

Para melhorar os resultados, desenvolveu-se um modelo híbrido que combina o modelo *SARIMA* com o modelo LSTM. O modelo *SARIMA* foi escolhido devido à sua habilidade em capturar tendências e sazonalidades de longo prazo presentes nos dados, proporcionando previsões mais precisas. Utilizaram-se os mesmos parâmetros do modelo *SARIMA* mencionado anteriormente. Da mesma forma, implementou-se um modelo de LSTM, com a mesma arquitetura definida anteriormente. Neste caso a LSTM recebeu como *input* os valores dos resíduos de cada previsão do modelo *SARIMA* e foi treinada para prever o valor real subsequente. Essa abordagem permite aproveitar a capacidade das redes neuronais em aprender padrões complexos e capturar dependências de curto prazo nos dados. Procurou-se obter a melhor capacidade do *SARIMA* em lidar com tendências de longo prazo e sazonalidades, juntamente com a capacidade da LSTM em capturar padrões mais complexos e dinâmicos.

Modelos de Previsão

O processo de modelação teve início com uma preparação dos dados. Os jogadores foram agrupados em três posições principais: avançados (FW), médios (MF) e defesas (DF), permitindo uma análise diferenciada por função no campo. As características relevantes foram selecionadas, eliminando informações não essenciais para a análise, como nomes dos jogadores e detalhes de identificação, concentrando-se nos atributos mais significativos para a previsão. A variável alvo foi definida como o valor de mercado dos jogadores em milhões de euros.

Para aprimorar a qualidade dos dados e mitigar possíveis distorções, foi implementada uma transformação robusta. Esta abordagem envolveu o uso de uma função logarítmica

com sinal, capaz de lidar eficazmente com valores tanto positivos quanto negativos, preservando assim a integridade da informação e normalizando a distribuição dos dados. Em seguida, foram identificados e removidos *outliers* extremos através da aplicação de *z-scores*, um método estatístico que permite identificar com precisão valores atípicos. O conjunto de dados resultante foi então submetido a um processo de escalonamento utilizando o *RobustScaler*, uma técnica avançada conhecida pela sua capacidade de lidar com *outliers*, garantindo que todas as características estejam na mesma escala sem serem indevidamente influenciadas por valores extremos.

Após essas etapas de pré-processamento, os dados foram divididos em conjuntos de treino e teste numa proporção de 80/20.

No âmbito da previsão do valor de mercado dos jogadores de futebol, foram implementados diversos modelos de aprendizagem automática. A *Linear Regression* pressupõe uma relação linear entre as características dos jogadores e o seu valor, permitindo uma análise direta da influência de cada variável. O modelo *Random Forest*, baseado em múltiplas árvores de decisão, ofereceu robustez e capacidade de generalização superiores. O *Adaptive Boosting* (AdaBoost) foi implementado como uma técnica avançada de *ensemble learning*, combinando múltiplos modelos para criar um modelo final robusto e preciso. O *Multi-layer Perceptron Regressor* (MLPRegressor) representa uma abordagem não linear, eficaz na captura de relações complexas nos dados.

Para melhorar o desempenho dos modelos, foram selecionadas as dez variáveis com os maiores coeficientes em termos absolutos. Estes modelos visam identificar e utilizar as características e estatísticas mais relevantes dos jogadores, proporcionando uma compreensão aprofundada dos fatores que influenciam o valor de mercado no futebol profissional.

3.6. CRISP-DM Fase 5: Avaliação

Nesta etapa do estudo, analisaram-se os vários modelos e especialmente, os resultados alcançados em resposta às questões de investigação e problemas iniciais identificados. Este momento foi caracterizado principalmente por momentos de análise crítica e reflexão sobre os resultados obtidos.

De acordo com a metodologia CRISP-DM, esta fase é designada por avaliar os resultados, rever o procedimento e determinar os próximos passos, e foi exatamente nesse sentido que a análise foi realizada. Foi feita uma avaliação do desempenho dos modelos, com especial atenção à sua capacidade de prever com precisão as várias categorias de progressão dos jogadores e a respetiva previsão do valor de mercado.

A natureza cíclica da metodologia CRISP-DM foi evidenciada nesta fase, já que os resultados obtidos durante a avaliação permitiram identificar áreas com potencial de melhoria. possibilitando regressar a fases anteriores, nomeadamente a fase de *Data Preparation* e *Modeling*, para que o estudo possa ser alinhado.

CAPÍTULO 4

Resultados e Discussão

4.1. Estudo sobre a análise de progressão de jogadores jovens

A avaliação dos modelos foi feita através de validação cruzada estratificada com 5 partições. Posteriormente, os modelos foram treinados utilizando o conjunto completo de dados de treino equilibrado e em seguida, avaliados no conjunto de teste.

Foram utilizadas as seguintes métricas para avaliar os modelos de classificação:

accuracy : para avaliar a proporção geral de predições corretas do modelo;

precision: para medir a proporção de predições positivas corretas;

recall: para avaliar a proporção de casos positivos reais que foram corretamente identificados;

F1-score: para fornecer um equilíbrio entre precisão e *recall*, sendo particularmente útil em conjuntos de dados desbalanceados.

Após treinar os modelos, foram calculadas estas métricas para o conjunto de teste, utilizando a média ponderada (*weighted average*) para precisão, *recall* e *F1-score*, a fim de considerar o desbalanceamento entre as classes. Os resultados obtidos estão apresentados na Tabela 4.1:

Tabela 4.1. Comparação de desempenho entre diferentes modelos

Modelo	***Accuracy***	***Precision***	***Recall***	***F1-score***
Random forest	0,5615	0,57	0,56	0,56
Logistic Regression	0,7538	0,75	0,75	0,75
SVM	0,4904	0,49	0,49	0,49
Gradient Boosting	0,5808	0,59	0,58	0,57
RNN	0,8154	0,82	0,82	0,81
GRU	0,8017	0,80	0,79	0,81
LSTM	0,8136	0,82	0,81	0,81

Através da Tabela 4.1, é possível confirmar que os modelos de redes neuronais recorrentes (RNN, LSTM, GRU) demonstraram um desempenho significativamente superior aos modelos tradicionais de aprendizagem automática. O modelo RNN foi o que apresentou a *accuracy* mais alta, indicando um melhor desempenho em relação aos outros modelos. Após decidir qual o modelo a ser utilizado na análise, procedeu-se ao ajuste de hiperparâmetros específicos para a parte RNN. Foram testadas várias combinações de hiperparâmetros, incluindo:

- Número de neurónios na primeira camada: 32, 64 e 128;

- Número de neurónios na segunda camada: 16, 32 e 64;
- Taxa de descarte (*dropout Rate*): 0.2 e 0.4;
- Taxa de aprendizagem (*Learning Rate*): 0.01 e 0.0001;
- *Batch size*: 32 e 64;
- Épocas: 50 e 100

Após testar todas as combinações, o modelo que obteve os melhores resultados foi aquele que utilizou a primeira camada com 64 neurónios, a segunda camada com 16 neurónios, um *dropout* de 0.2, a taxa de aprendizagem de 0.01, o *batch size* com 64 e 50 como número de épocas. Este processo de "*Hiperparameter tuning*" é fundamental para encontrar a configuração de hiperparâmetros que melhor se adequa ao problema em questão. Através da otimização dos hiperparâmetros, foi possível melhorar o desempenho do modelo e obter resultados mais precisos e confiáveis.

Tabela 4.2. Relatório de Classificação do Modelo RNN com os hiperparâmetros

Categoria	***Precision***	***Recall***	***F1-score***	***Support***
Manteve-se	0.79	0.66	0.72	107
Progrediu	0.84	0.94	0.89	136
Progrediu Ligeiramente	0.78	0.79	0.78	94
Progrediu Significativamente	0.92	0.86	0.89	28
Regrediu	0.85	0.85	0.85	155
Accuracy	0.82			520
Macro avg	0.83	0.82	0.83	520
Weighted avg	0.82	0.82	0.82	520

Através da tabela 4.2, pode-se observar que o modelo RNN otimizado demonstrou um bom desempenho na classificação da progressão dos jogadores de futebol, demonstrando uma *accuracy* de 0,82.

O modelo revelou uma eficácia em reconhecer jogadores que progrediram, com um *recall* de 0,94 para a categoria "Progrediu". Esta *performance* indicou que o sistema é particularmente habilidoso em detetar padrões de melhoria no desempenho dos atletas, capturando com precisão 94% dos casos de progressão. Da mesma forma, a categoria "Progrediu Significativamente" apresentou a maior precisão (0,92), sugerindo que quando o modelo prevê uma progressão acentuada, a sua fiabilidade é alta.

É interessante notar que o modelo manteve um equilíbrio sólido entre precisão e *recall* na maioria das categorias, refletido nos *F1-scores* consistentes. Por exemplo, a categoria "Regrediu" demonstra uma harmonia com valores idênticos de 0,85 para precisão, *recall* e *F1-score*. Essa consistência sugeriu que o modelo é igualmente eficaz em identificar tanto os casos positivos quanto os negativos de regressão no desempenho dos jogadores.

No entanto, houve uma pequena dificuldade do modelo em identificar corretamente os jogadores que mantiveram um desempenho estável. A categoria "Manteve-se" apresentou

o menor *recall* (0,66), indicando que o sistema tem uma tendência a classificar incorretamente alguns desses casos em outras categorias. Essa observação pode sugerir que a estabilidade no desempenho possui características mais complexas, e que consequentemente o modelo tenha mais dificuldade em capturar.

É importante considerar a variação no número de casos (*support*) entre as diferentes categorias. A categoria "Progrediu Significativamente", por exemplo, tem o menor número de casos (28), o que pode contribuir para sua alta precisão.

Esta análise detalhada do desempenho do modelo RNN otimizado ofereceu uma visão aprofundada sobre os padrões de desenvolvimento dos jogadores de futebol. O modelo demonstra uma capacidade de discernir diferentes níveis de progressão, regressão e estabilidade no desempenho atlético.

Através da Figura 4.1 identificaram-se fatores que influenciam a progressão significativa na carreira. O número de golos marcados pela equipa com o jogador em campo (*num_onG*). O número de jogos disputados (*num_MP*), foi o segundo fator mais importante, seguido dos golos sofridos pela equipa enquanto esteve em campo (*num_onGA*). Os pontos de classificação do *ranking* da equipa (*num_Points*) também influenciam. Jogar por clubes como *Burnley*, *Sunderland*, *Watford* e *Millwall* pareceu contribuir para o progresso. A idade do jogador (*num_Age*) e as assistências (*num_Ast*) tiveram uma influência menor, mas ainda assim relevante.

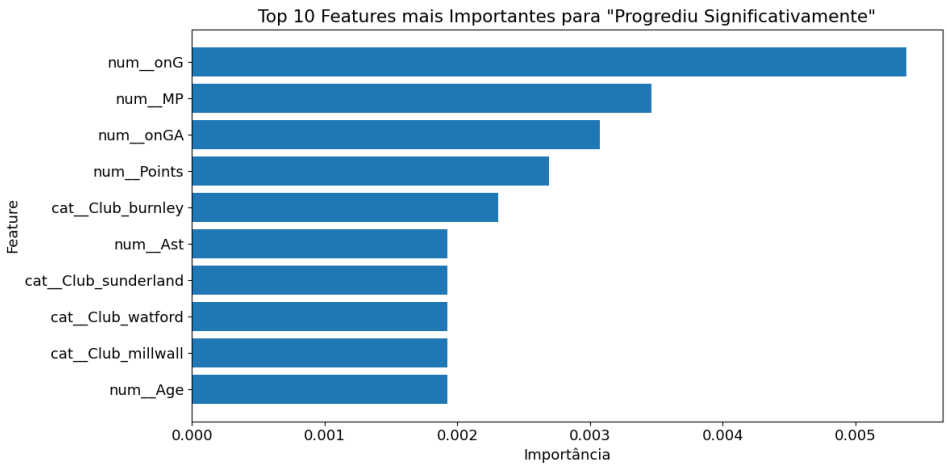


FIGURA 4.1. Top 10 de *features* mais importantes para progrediu significativamente.

Através da Figura 4.2 identificaram-se os fatores que influenciam a progressão na carreira, como o número de golos marcados pela equipa enquanto esteve em campo (*num_onG*) que se manteve como o fator principal. A idade do jogador (*num_Age*) surgiu como o segundo fator mais importante, indicando a relevância da fase de desenvolvimento do atleta. O *ranking* do clube (*num_Club_Ranking*) foi o terceiro fator mais influente. O número total de golos (*num_Gls*) e o ano em que termina a temporada (*num_Season_Y_End*) também foram fatores importantes. A participação em grandes

penalidades (*num_PK*) influencia o progresso. Jogar por clubes como *Charleroi*, *Cardiff City* e *Colchester United* é benéfico para profredir. As assistências (*num_Ast*) continuaram a ser um fator importante, embora menos relevante que na progressão significativa.

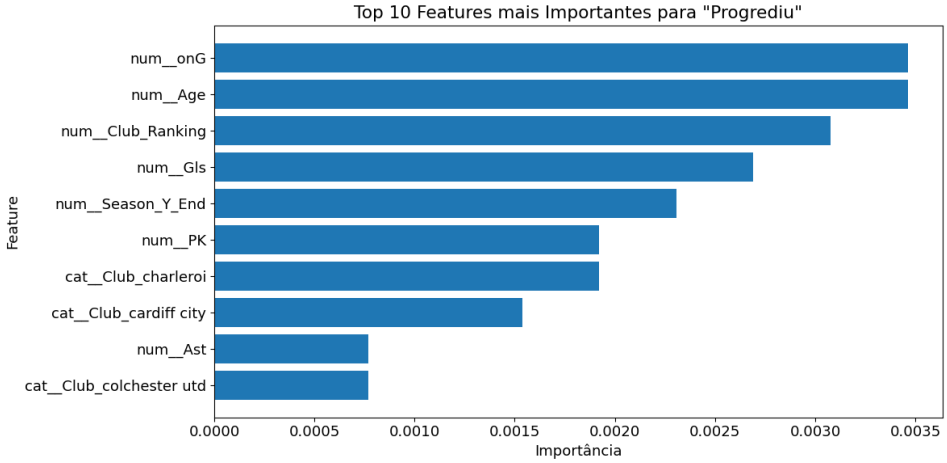


FIGURA 4.2. Top 10 de *features* mais importantes para progrediu.

No que diz respeito à regressão na carreira de um jogador os fatores que influenciaram mais foram a idade (*num_Age*), o nº de minutos jogados (*num_Min*) e o ano em que termina a época (*num_Season_Y_End*).

4.2. Estudo sobre a análise do valor de mercado dos jogadores ao longo das diferentes épocas

No que respeita à análise do valor de mercado, foram utilizadas as seguintes métricas para avaliar os modelos de previsão:

Mean Absolute Error (MAE): para medir o erro médio absoluto das previsões;

Root Mean Square Error (RMSE): para avaliar o erro quadrático médio, dando mais peso a erros maiores;

Mean Absolute Percentage Error (MAPE): para medir o erro percentual médio absoluto, permitindo uma comparação relativa do desempenho;

Séries Temporais

Após treinar os modelos, foram calculadas as métricas atrás mencionadas para o conjunto de teste. Os resultados obtidos estão apresentados na tabela a seguir:

Modelo	**MAE**	**RMSE**	**MAPE**
SARIMA	1,21	2,72	0,57%
GRU	0,93	2,14	0,67%
RNN	0,79	1,73	0,81%
LSTM	1,12	2,70	0,79%
Hybrid	1,77	3,00	4,44%

TABELA 4.3. Comparação de desempenho entre diferentes modelos de previsão

Através da Tabela 4.3, é possível confirmar que o modelo RNN foi o que apresentou as métricas mais baixas, indicando um melhor desempenho em relação aos outros modelos. O modelo RNN, em particular, demonstrou uma redução significativa no MAE e RMSE em comparação com os outros modelos. Após decidir qual o modelo a ser utilizado na análise, procedeu-se ao ajuste de hiperparâmetros específicos para a parte RNN. Foram testadas várias combinações de hiperparâmetros, incluindo:

- Taxa de descarte (*dropout* Rate): 0.2, 0.4 e 0.5;
- Otimizadores: Adam e *RMSprop*;
- Taxa de aprendizagem (*Learning Rate*): 0.01 e 0.0001;
- Número de neurónios em cada camada: 128, 64, 16, 4 e 2.

Após testar todas as combinações, o modelo que obteve os melhores resultados foi aquele que utilizou um *dropout* de 0,2 e o otimizador Adam com uma taxa de aprendizagem de 0,01 e 128 neurónios em cada camada.

Após encontrar o melhor modelo, foi criado um modelo RNN com todos os dados, para conseguir criar previsões futuras, neste caso foi feita a previsão para 1 semestre no futuro, cujos resultados se encontram na Figura 4.3

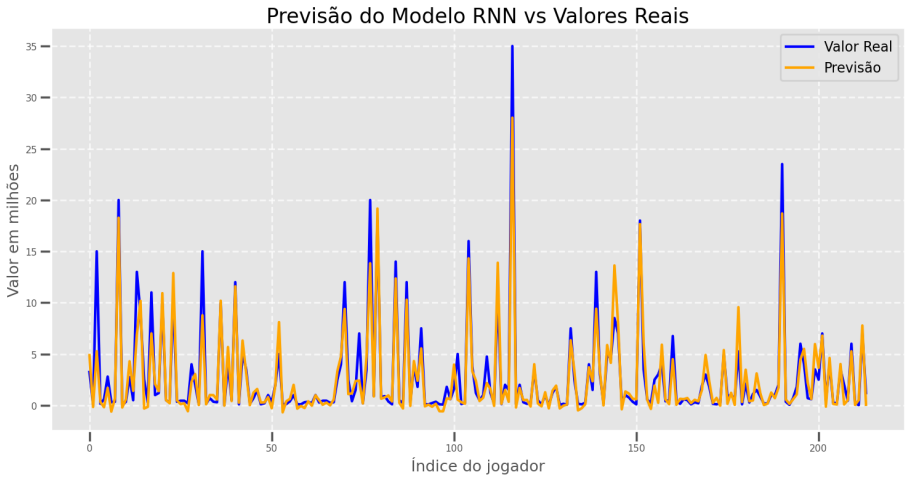


FIGURA 4.3. *Out-of-sample* para o modelo RNN.

Um ponto de interesse na previsão de séries temporais prende-se com o facto de a previsão poder ser feita em *out-of-sample*. Em termos muito práticos, é utilizada para a obtenção de valores futuros para uma data "fora" da janela temporal da série. Ao analisar a Figura 4.3, observou-se esta abordagem utilizando o modelo RNN para fazer previsões sobre valores futuros dos dados de validação. Ao comparar as previsões com os valores reais, foi possível avaliar que o modelo teve um bom desempenho.

Modelos de Previsão

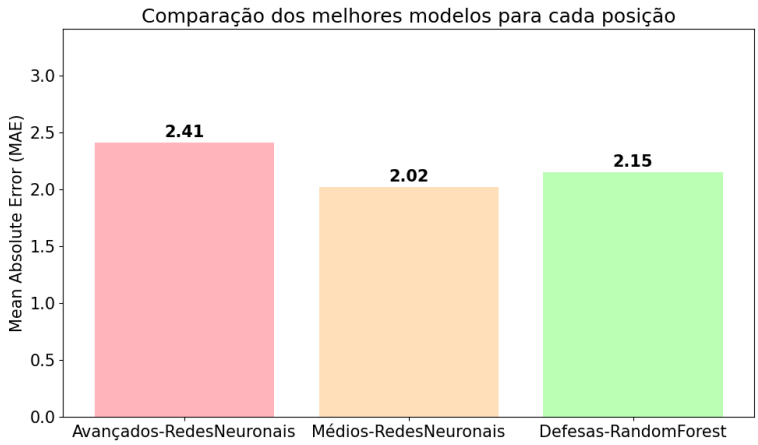


FIGURA 4.4. *Performance* do modelo para as 3 posições de campo (MAE)

TABELA 4.4. Resultados dos Modelos

Posição	**Modelo**	**MAE**	**RMSE**	**MAPE**
Avançados	*Linear Regression*	2.50	3.72	15.15
	Random Forest	2.41	3.86	16.89
	AdaBoost	3.19	4.34	21.36
	Redes Neuronais	2.18	3.18	10.97
Médios	*Linear Regression*	2.24	3.21	10.45
	Random Forest	2.06	3.07	9.64
	AdaBoost	2.71	3.40	11.63
	Redes Neuronais	2.02	2.97	8.95
Defesas	*Linear Regression*	2.35	3.53	14.90
	Random Forest	2.15	3.12	14.12
	AdaBoost	2.87	3.93	15.43
	Redes Neuronais	2.76	4.29	18.81

Através da Tabela 4.4, que apresenta os resultados dos diversos modelos aplicados para prever o valor de mercado dos jogadores por posição, conseguiu-se concluir o seguinte:

Para os avançados, o modelo de Redes Neuronais demonstrou o melhor desempenho global, apresentando o menor Erro Absoluto Médio (MAE) de 2,18, o menor Erro Quadrático Médio (RMSE) de 3,18, e a menor Percentagem Média de Erro Absoluto (MAPE) de 10,97%.

No caso dos médios, observou-se um padrão semelhante, com o modelo de Redes Neuronais a exibir o melhor desempenho em termos de MAE (2,02), RMSE (2,97) e MAPE (8,95%). O modelo *Random Forest* para esta posição também apresentou resultados competitivos.

Quanto aos defesas, o modelo *Random Forest* destacou-se com o menor MAE (2,15) e o melhor RMSE (3,12). No entanto, verificou-se que todos os modelos para esta posição apresentaram desempenhos mais modestos em comparação com as outras posições.

O modelo *AdaBoost* consistentemente apresentou o pior desempenho em todas as métricas para todas as posições, indicando que esta abordagem pode não ser a mais adequada para este tipo específico de previsão.

Para otimizar as Redes Neuronais nos casos dos avançados e médios, procedeu-se ao ajuste de hiperparâmetros. No caso das Redes Neuronais, os principais hiperparâmetros a serem ajustados incluíram:

- *hidden_layer_sizes*: (50,), (100,), (50, 50) e (100, 50);
- *activation*: *relu* e *tanh*;
- *alpha*: 0.0001, 0.001 e 0.01;
- *learning_rate*: *constant* e *adaptive*.

Os melhores hiperparâmetros encontrados para os avançados foram: *activation*: *tanh*, *alpha*: 0,01, *hidden_laye_sizes*: (100, 50), *learning_rate*: *constant*, enquanto que para os médios foram: *activation*: *tanh*, *alpha*: 0.001, *hidden_layer_sizes*: (50,), *learning_rate*: *constant*.

Para os defesas, procedeu-se ao ajuste de hiperparâmetros. No caso do modelo *Random Forest*, os príncipais hiperparâmetros a serem ajustados incluíram:

- *bootstrap*: *True*;
- *max_depth*: 80, 90, 100 e 110;
- *max_features*: 2 e 3;
- *min_samples_leaf*: 3, 4 e 5;
- *min_samples_split*: 8, 10 e 12;
- *n_estimators*: 100, 200, 300, 1000.

Os melhores hiperparâmetros encontrados para os defesas foram: *bootstrap*: *True*, *max_depth*: 110, *max_features*: 3, *min_samples_leaf*: 3, *min_samples_split*: 8, *n_estimators*: 100.

Após encontrar os melhores hiperparamêtros, criaram-se os modelos e conseguiu-se perceber as características que contribuem mais para cada modelo e posição (ver Figura 4.5, Figura 4.6 e Figura 4.7).

Avançados

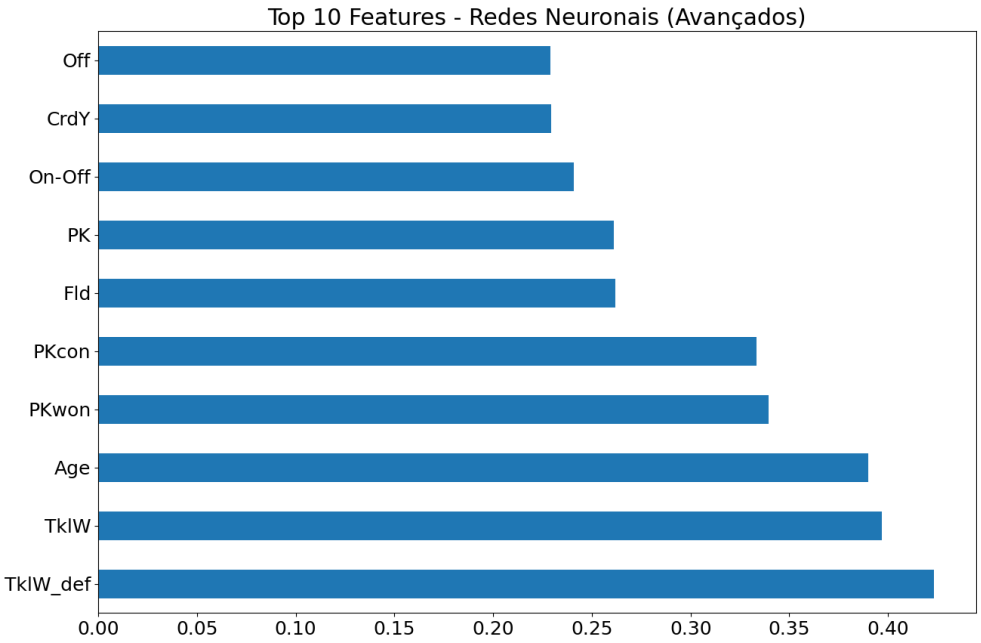


FIGURA 4.5. Top 10 *features* mais importantes para os avançados

A análise da Figura 4.5, referente à importância das características no modelo de Redes Neuronais para avançados, revelou um conjunto de fatores diversificado e em alguns aspetos, contraintuitivo.

Destacou-se, de forma inesperada, a preponderância de habilidades defensivas (TkIW_def e TkIW) no topo da hierarquia, sugerindo que a versatilidade e a capacidade de contribuir defensivamente foram altamente valorizadas nos avançados modernos. A idade surgiu como o terceiro fator mais influente, demonstrando a importância do potencial de desenvolvimento e longevidade na carreira.

A eficácia em situações de grande penalidade (PKwon, PKcon e PK) emergiu como um conjunto de características significativas, evidenciando a valorização de jogadores decisivos em momentos críticos. A capacidade de atrair faltas (Fld) e o impacto global no desempenho da equipa (On-Off) também se afiguraram como atributos relevantes.

Apesar dos foras de jogo (Off), curiosamente, métricas tradicionalmente associadas ao desempenho ofensivo, como golos marcados ou assistências, não contribuíram diretamente neste top 10. Esta ausência revela que o modelo está a captar nuances menos evidentes, mas potencialmente cruciais, no contributo dos avançados para o jogo coletivo.

Médios

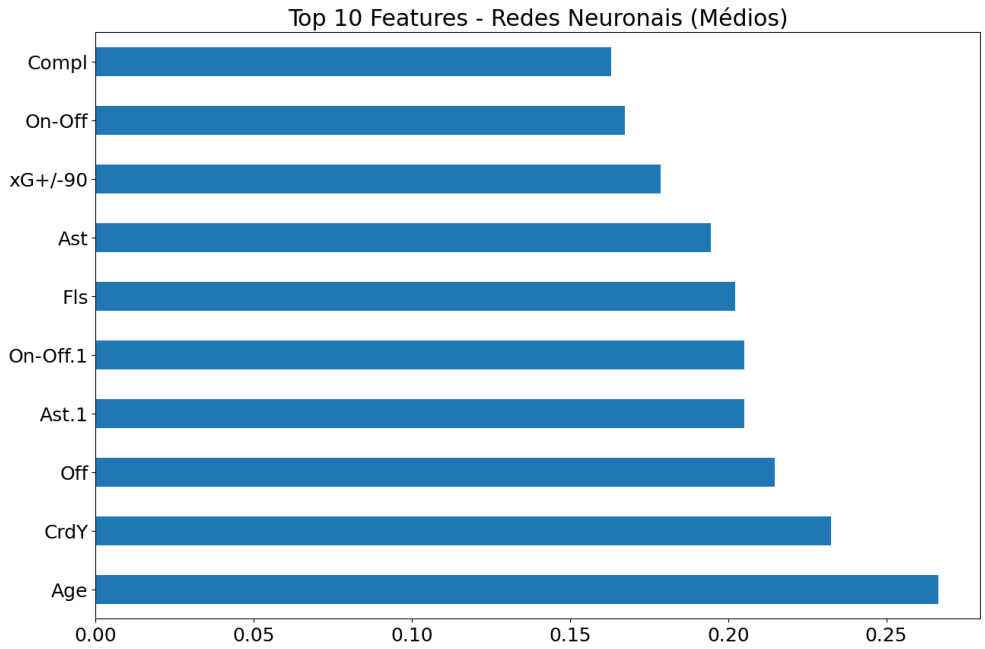


FIGURA 4.6. Top 10 *features* mais importantes para os médios

A análise do gráfico da Figura 4.6 relativo à importância de características para o modelo de Redes Neuronais aplicado aos médios revelou um conjunto de fatores que influenciam o seu valor de mercado. A idade emergiu como o elemento preponderante, sublinhando a importância crucial do equilíbrio entre experiência e potencial de desenvolvimento na valorização destes jogadores. Surpreendentemente, os cartões amarelos (CrdY) surgiram como a segunda característica mais relevante, sugerindo que a disciplina tática foi valorizada nesta posição.

A capacidade de criar oportunidades, representada por métricas de assistências (Ast), destacaram-se como atributos significativos. A influência global do jogador no desempenho da equipa, medida através de várias métricas "On-Off", também se revela crucial. Características como faltas cometidas e contribuição para golos esperados por 90 minutos completaram o quadro, indicando a valorização de uma agressividade controlada e um impacto ofensivo consistente.

Observou-se a ausência de características puramente defensivas no top 10, o que pode indicar uma tendência para privilegiar aspetos ofensivos e de construção de jogo nos médios. Esta configuração de características importantes revela uma apreciação multifacetada do valor dos médios, alinhada com a evolução do futebol moderno, onde estes jogadores são frequentemente avaliados pela sua capacidade de influenciar o jogo em múltiplas decisões de jogo.

Defesas

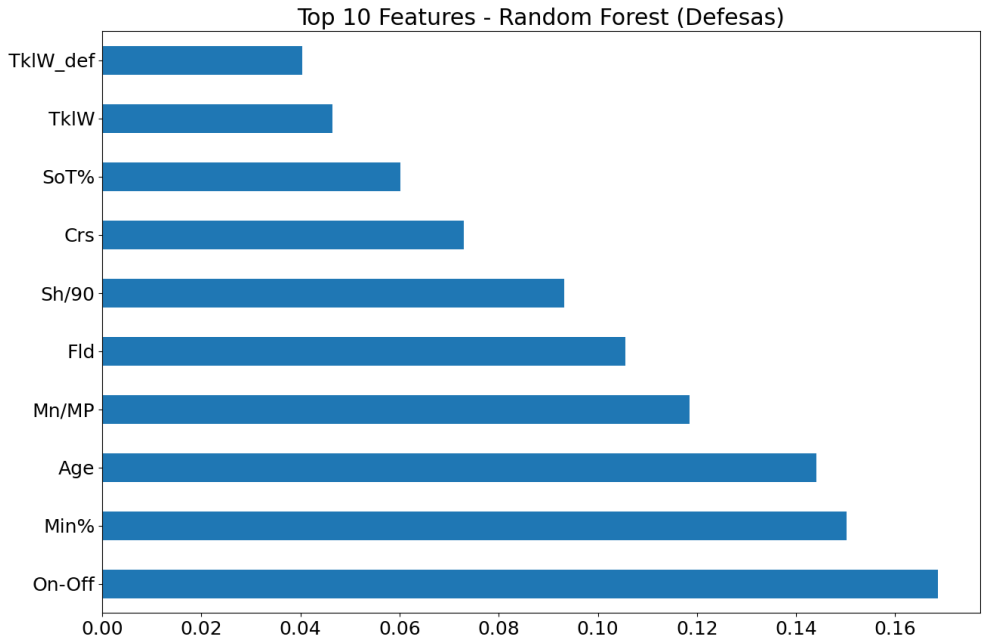


FIGURA 4.7. Top 10 *features* mais importantes para os defesas

A análise do gráfico da Figura 4.7, relativo à importância de características para o modelo *Random Forest* aplicado aos defesas, a métrica "On-Off" surgiu como o elemento preponderante, sugerindo que o impacto global do defesa no desempenho da equipa é crucial para a sua valorização. Esta métrica provavelmente captura a diferença no rendimento coletivo com e sem a presença do jogador em campo, sublinhando a importância da influência individual no contexto da equipa.

A percentagem de minutos jogados (Min%) e a relação entre minutos jogados e partidas disputadas (Min/Mp) surgiram como o segundo e quarto fatores mais relevantes, respetivamente. Estas métricas enfatizam a valorização da regularidade, fiabilidade e consistência na utilização do jogador, aspetos fundamentais para a estabilidade defensiva de uma equipa.

Curiosamente, o número de faltas sofridas (Fld) apareceu como um fator significativo, sugerindo que defesas capazes de atrair faltas dos adversários, seja por posicionamento inteligente ou por provocar situações de pressão, são particularmente apreciados. A métrica de remates por 90 minutos indicou que defesas com capacidade ofensiva também são valorizados, refletindo a tendência moderna de defesas que contribuem no ataque. A idade surgiu como um fator relevante, sublinhando o delicado equilíbrio entre experiência e potencial de desenvolvimento na avaliação dos defesas. A presença de métricas como cruzamentos e desarmes (tanto defensivos como em posse de bola) no top 10 realçou a importância de uma gama diversificada de habilidades, desde a contribuição ofensiva até às competências defensivas tradicionais.

CAPÍTULO 5

Conclusão e Recomendações

O presente estudo, centrado na progressão dos jovens jogadores da *Championship* na época 2022/2023, apresenta uma análise sobre a evolução na carreira e o valor de mercado desses atletas, respondendo às questões de investigação propostas e contribuindo significativamente para a compreensão das dinâmicas do futebol profissional.

O primeiro estudo, focado na análise da progressão dos jogadores ao longo das diferentes épocas revelou padrões complexos, indicando que o percurso de um jovem jogador de futebol não é linear nem previsível.

A classificação da progressão de um jogador caracteriza-se através de um sistema de avaliação quantitativo. Este sistema, proposto no âmbito da dissertação, baseia-se em critérios que incluem minutos jogados, classificação do clube, estatísticas gerais e específicas por posição. A cada jogador são atribuídos pontos positivos ou negativos em cada categoria, mediante da percentagem de melhoria ou regressão, respetivamente. A soma destes pontos determina a categoria de progressão do jogador, variando entre "Progrediu Significativamente" e "Regrediu". Este método permite uma análise da evolução dos atletas, considerando fatores do seu desempenho e contexto. A abordagem facilita comparações entre épocas e entre posições de jogadores, fornecendo uma visão do desenvolvimento dos atletas no futebol profissional. Esta ferramenta, revelou-se eficaz para avaliar a evolução dos atletas de época em época.

Um modelo baseado em Redes Neuronais Recorrentes (RNN), demonstrou uma precisão global de 82% na categorização da evolução dos jogadores. Este foi particularmente eficaz na identificação de jogadores que progrediram, com um *recall* de 94% para esta categoria. Isso sugere que o sistema é capaz de capturar padrões complexos de desenvolvimento para a progressão. O estudo revela ainda que a progressão na carreira de um futebolista depende principalmente do desempenho ofensivo, medido pelos golos marcados pela equipa enquanto o jogador esteve em campo e assistências, bem como do tempo de jogo. A idade do jogador influência tanto a progressão como a regressão, indicando a importância da fase de carreira. O *ranking* e a identidade do clube também afetam o progresso, sugerindo que certas equipas oferecem melhores condições de desenvolvimento. No caso de jogadores que regrediram, os fatores determinantes da regressão, além da idade, são os minutos jogados e o ano em que termina a época. Relativamente à manutenção na mesma categoria pode-se inferir que resulta de um equilíbrio entre os fatores de progressão e regressão.

O segundo estudo, centrado na análise do valor de mercado, revelou tendências ao longo das diferentes épocas para os jogadores da *Championship* 2022/2023. A caracterização pormenorizada desses valores oferece uma visão única sobre como o mercado avalia o potencial e o desempenho dos jovens jogadores ao longo do tempo.

O valor de mercado dos jogadores da *Championship* 22/23 apresenta uma tendência de crescimento ao longo das épocas analisadas. A evolução do valor médio de mercado dos jogadores de 2017 a 2023 tem sofrido um aumento constante, com uma aceleração notável a partir de 2021. Na análise dos top 5 jogadores, observa-se a existência de variabilidade individual nas trajetórias do valor de mercado. O jogador João Pedro destaca-se com um aumento acentuado no valor de mercado nas últimas duas épocas, enquanto outros jogadores como, o Max Aarons e o Josh Sargent, mostram padrões de crescimento diferentes, esta variação reflete a dinâmica do mercado de transferências.

O estudo revelou que diferentes modelos de aprendizagem automática são mais adequados para diferentes posições em campo. Para os avançados e médios, as Redes Neuronais apresentaram o melhor desempenho, enquanto que para os defesas, o modelo *Random Forest* mostrou-se mais eficaz. Esta diferenciação ressalta a complexidade e a especificidade dos fatores que influenciam o valor de mercado em diferentes posições.

A identificação dos fatores que influenciam o valor de mercado, especialmente quando analisados por posição de jogo, proporcionou informações valiosas. Esta abordagem diferenciada por posição demonstrou que os critérios de valorização podem variar significativamente entre avançados, médios e defesas, refletindo as diferentes expectativas e exigências para cada função no campo. A análise das características mais importantes para cada posição revelou informações contraintuitivas. Para os avançados, por exemplo, habilidades defensivas surgiram como fatores significativos, desafiando noções tradicionais sobre o que torna um atacante valioso. Para os médios, a idade e a disciplina (representada pelos cartões amarelos) emergiram como fatores cruciais. Já para os defesas, o impacto global no desempenho da equipa e a regularidade de jogo foram os aspetos mais relevantes.

A análise da evolução do valor de mercado desde 2017 e a subsequente previsão para o próximo semestre representam uma contribuição significativa para a compreensão das tendências do mercado de transferências. Esta análise foi realizada recorrendo a modelos de séries temporais. O estudo comparou diferentes modelos, incluindo *SARIMA*, GRU, RNN, LSTM e um modelo híbrido. O modelo RNN apresentou o melhor desempenho, com os menores valores de MAE (0,79), RMSE (1,73) e MAPE (0,81%). Os resultados, indicam que o modelo captura a tendência dos valores de mercado, com algumas discrepâncias em picos de valor. Esta análise permite prever quantitativamente a evolução do valor de mercado dos jogadores, contribuindo para decisões de transferências e gestão de ativos no futebol.

Esta investigação contribui de forma significativa para o corpo de conhecimento sobre o desenvolvimento de jovens jogadores no futebol profissional. Ao abordar tanto a progressão na carreira quanto as dinâmicas do valor de mercado, o estudo oferece uma visão abrangente e composta das trajetórias dos jovens talentos no futebol inglês.

Este estudo apresenta algumas limitações que devem ser consideradas. A recolha e tratamento de dados através de *web scraping* constituiu um desafio significativo durante a investigação. Uma das principais dificuldades encontradas consistiu na gestão de dados em falta, exigindo a implementação de técnicas de imputação e a validação entre três fontes distintas de informação. A interpretação dos modelos de aprendizagem automática revelou-se igualmente desafiante, para a sua otimização e validação dos resultados.

Este estudo respondeu às questões de investigação propostas e abre caminho para novas investigações. Futuras pesquisas poderão considerar um período mais longo, comparar diferentes ligas ou incorporar análises de fatores identificados como cruciais para o desenvolvimento e valorização dos jogadores.

Nessas investigações, recomenda-se acrescentar variáveis exógenas, como estatísticas de jogo, ao conjunto de dados das séries temporais. Outra recomendação é extrair informações adicionais, como o salário do jogador por época e o histórico de lesões, considerados elementos relevantes na progressão de um jogador jovem. Uma abordagem interessante seria também explorar a incorporação de análises de sentimento das redes sociais. A perceção pública e a reputação *online* dos jogadores podem ter um impacto no seu valor de mercado e na progressão da carreira.

Bibliografia

Drew, K. L. (2020). *Investigating the Junior-to-Senior Transition in Sport: Interventions to Support the Transitional Process*.

Güllich, A. e Barth, M. (2024). Effects of Early Talent Promotion on Junior and Senior Performance: A Systematic Review and Meta-Analysis. *Sports Medicine* 54 (3), 697–710. ISSN: 11792035. DOI: 10.1007/s40279-023-01957-3.

Anderson, G. e Miller, R. M. (2011). The academy system in English professional football: Business value or following the herd? *Management School Research Paper Series*.

Mitchell, T., Gledhill, A., Nesti, M., Richardson, D. e Littlewood, M. (2020). Practitioner Perspectives on the Barriers Associated With Youth-to-Senior Transition in Elite Youth Soccer Academy Players. *International Sport Coaching Journal* 7 (3), 273–282. ISSN: 23289198. DOI: 10.1123/iscj.2019-0015.

Borland, J. e Lye, J. (1996). Matching and Mobility in the Market for Australian Rules Soccer Coaches. *Industrial and Labor Relations Review* 50.(1), 143–158.

Morris, R., Tod, D. e Eubank, M. (2017). From youth team to first team: An investigation into the transition experiences of young professional athletes in soccer. *International journal of sport and exercise psychology* 15.(5), 523–539.

Carpels, T., Scobie, N., Macfarlane, N. G. e Kemi, O. J. (2021). Youth-to-senior transition in elite European club soccer. *International journal of exercise science* 14 (6), 1192–1203.

Helsen, W. F., Van Winckel, J. e Williams, A. M. (2005). The relative age effect in youth soccer across Europe. *Journal of sports sciences* 23.(6), 629–636.

Cobley, S., Baker, J., Wattie, N. e McKenna, J. (2009). Annual age-grouping and athlete development: a meta-analytical review of relative age effects in sport. *Sports Medicine* 39, 235–256.

Bezuglov, E., Morgans, R., Butovskiy, M., Emanov, A., Shagiakhmetova, L., Pirmakhanov, B. e Lazarev, A. (2023). The relative age effect is widespread among European adult professional soccer players but does not affect their market value. *PLoS ONE* 18.(3), e0283390. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0283390.

Till, K. e Baker, J. (2020). Challenges and [possible] solutions to optimizing talent identification and development in sport. *Frontiers in Psychology* 11, 664.

Chroepf, B. e Lames, M. (2018). Career patterns in German football youth national teams–a longitudinal study. *International journal of sports science & coaching* 13 (3), 405–414.

Herrebrøden, H. e Bjørndal, C. Thue (2022). Youth International Experience Is a Limited Predictor of Senior Success in Football: The Relationship Between U17, U19, and U21 Experience and Senior Elite Participation Across Nations and Playing Positions. *Frontiers in Sports and Active Living* 4. ISSN: 26249367. DOI: 10.3389/fspor.2022.875530.

Christensen, M. K. (2009). "An Eye for Talent": Talent Identification and the "Practical Sense" of Top-Level Soccer Coaches. *Sociology of Sport Journal* 26 (3), 365–382.

Prinz, Joachim e Weimar, Daniel (2018). «The golden generation: the personnel economics of youth recruitment in European professional soccer». Edward Elgar Publishing Ltd., pp. 47–71. ISBN: 9781786430915. DOI: 10.4337/9781786430915.00010.

Vaziri, B., Dabadghao, S., Yih, Y. e Morin, T. L. (2018). Properties of sports ranking methods. *Journal of the Operational Research Society* 69 (5), 776–787. ISSN: 14769360. DOI: 10.1057/s41274-017-0266-8.

Buchdahl, J. (2003). Fixed odds sports betting: Statistical forecasting and risk management. *London: High Stakes*.

Hvattum, L. M. e Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting* 26 (3), 460–470. ISSN: 01692070. DOI: 10.1016/j.ijforecast.2009.10.002.

Hire Intelligence (2024). *The Evolution of Technology in Sport*. https://www.hireintelligence.co.uk/evolution-of-technology-in-sport//. Último acesso em: 20 de outubro de 2024.

Araújo, D., Davids, K. e Renshaw, I. (2020). *Cognition, emotion and action in sport: an ecological dynamics perspective*. Handbook of sport psychology, 535-555.

Hughes, Mike D. e Bartlett, Roger M. (2002). *The use of performance indicators in performance analysis*. DOI: 10.1080/026404102320675602.

Wang, J. e Chen, J. (2022). Design and Research of Dynamic Evolution System in Football Tactics under Computational Intelligence. *Mathematical Problems in Engineering* 2022. ISSN: 15635147. DOI: 10.1155/2022/3772236.

Dufour, W. (2003). *Computer-assisted scouting in soccer*. In: Science and Football II, pp. 160–164. Taylor & Francis.

Sarmento, H., Clemente, F. M., Araújo, D., Davids, K., McRobert, A. e Figueiredo, A. (2018). What Performance Analysts Need to Know About Research Trends in Association Football (2012–2016): A Systematic Review. *Sports Medicine* 48 (4), 799–836. ISSN: 11792035. DOI: 10.1007/s40279-017-0836-6.

FIFA (2022). https://www.fifa.com/pt/articles/copa-do-mundo-2022-tecnologia-semiautomatizada-de-impedimento. Último acesso em: 20 de outubro de 2024.

Lord, F., Pyne, D. B., Welvaert, M. e Mara, J. K. (2022). Capture, analyse, visualise: An exemplar of performance analysis in practice in field hockey. *PLoS ONE* 17 (5). ISSN: 19326203. DOI: 10.1371/journal.pone.0268171.

Bialik, C. (2014). *The People Tracking Every Touch, Pass And Tackle in the World Cup*. https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/. Último acesso em: 20 de outubro de 2024.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. e Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data* 6 (1). ISSN: 20524463. DOI: 10.1038/s41597-019-0247-7.

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)* 9 (1), 381–386. DOI: 10.21275/art20203995.

Sports, B. (2024). *The impact of data analytics on football team performance*. https://www.linkedin.com/pulse/impact-data-analytics-football-team-performance-brandsen-sports/. Último acesso em: 20 de outubro de 2024.

Sun, D. (2023). An Overview of Machine Learning Applications in the Football Field. *Applied and Computational Engineering* 8 (1), 318–322. ISSN: 2755-2721. DOI: 10.54254/2755-2721/8/20230178.

Merhej, C., Beal, R. J., Matthews, T. e Ramchurn, S. (2021). «What happened next: using deep learning to value defensive actions in football event-data». *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. New York, NY, USA: Association for Computing Machinery, pp. 3394–3403. DOI: 10.1145/3447548.3467194.

Cui, Y. X., Li, Y. S. e Shen, Y. F. (2022). «Research status and prospect of football match performance evaluation model». *Compilation of Abstracts of the 12th National Sports Science Conference - Poster Exchange (Sports Engineering Branch)*, pp. 110–111.

Merzah, B. M., Croock, M. S. e Rashid, A. N. (2024). Intelligent Classifiers for Football Player Performance Based on Machine Learning Models. *International journal of electrical and computer engineering systems* 15.(2), 173–183.

Li, C., Kampakis, S. e Treleaven, P. (2022). Machine learning modeling to evaluate the value of football players. *arXiv preprint arXiv:2207.11361*.

Hotz, N. (2024). *What is CRISP DM*. https://www.datascience-pm.com/crisp-dm-2/. Último acesso em: 20 de outubro de 2024.

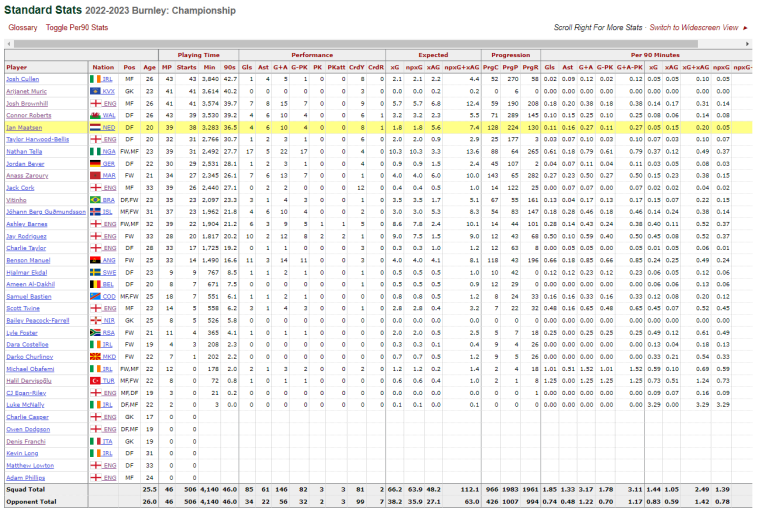
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. e Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc* 9 (13), 1–73.

FootballDatabase (2024). *Methodology for calculating FootballDatabase's world football clubs ranking*. https://www.footballdatabase.com/methodology. Último acesso em: 20 de outubro de 2024.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning* 3.(19-48), 5–1.

Masini, R. P., Medeiros, M. C e Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys* 37.(1), 76–111.

Anexo A – *Web Scraping*

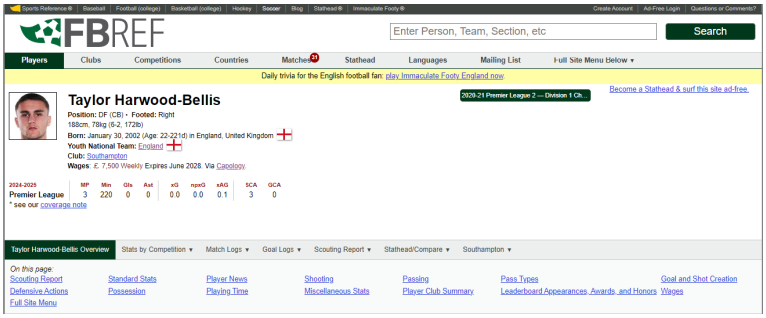


Standard Stats 2022-2023 Burnley: Championship

Glossary · Toggle Per90 Stats · Scroll Right For More Stats · Switch to Widescreen View

				Playing Time				Performance							Expected			Progression		Per 90 Minutes									
Player	Nation	Pos	Age	MP	Starts	Min	90s	Gls	Ast	G+A	G-PK	PK	PKatt	CrdY	CrdR	xG	npxG	xAG	npxG+xAG	PrgC	PrgP	PrgR	Gls	Ast	G+A	G-PK	G+A-PK	xG	xAG	xG+xAG	npxG	npxG+
Josh Cullen	IRL	MF	26	43	43	3,840	42.7	1	4	5	1	0	0	8	0	2.1	2.1	2.2	4.4	52	270	58	0.02	0.09	0.12	0.02	0.12	0.05	0.05	0.10	0.05	
Arijanet Muric	KVX	GK	23	41	41	3,614	40.2	0	0	0	0	0	0	3	0	0.0	0.0	0.2	0.2	0	6	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Josh Brownhill	ENG	MF	26	41	41	3,574	39.7	7	8	15	7	0	0	9	0	5.7	5.7	6.8	12.4	59	190	208	0.18	0.20	0.38	0.18	0.38	0.14	0.17	0.31	0.14	
Connor Roberts	WAL	DF	26	43	39	3,530	39.2	4	6	10	4	0	0	6	1	3.2	3.2	2.3	5.5	71	289	145	0.10	0.15	0.25	0.10	0.25	0.08	0.06	0.14	0.08	
Ian Maatsen	NED	DF	20	39	38	3,283	36.5	4	6	10	4	0	0	8	1	1.8	1.8	5.6	7.4	128	224	130	0.11	0.16	0.27	0.11	0.27	0.05	0.15	0.20	0.05	
Taylor Harwood-Bellis	ENG	DF	20	32	31	2,766	30.7	1	2	3	1	0	0	6	0	2.0	2.0	0.9	2.9	25	177	3	0.03	0.07	0.10	0.03	0.10	0.07	0.03	0.10	0.07	
Nathan Tella	NGA	FW,MF	23	39	31	2,492	27.7	17	5	22	17	0	0	4	0	10.3	10.3	3.3	13.6	88	64	265	0.61	0.18	0.79	0.61	0.79	0.37	0.12	0.49	0.37	
Jordan Beyer	GER	DF	22	30	29	2,531	28.1	1	2	3	1	0	0	4	0	0.9	0.9	1.5	2.4	45	107	2	0.04	0.07	0.11	0.04	0.11	0.03	0.05	0.08	0.03	
Anass Zaroury	MAR	FW	21	34	27	2,345	26.1	7	6	13	7	0	0	1	0	4.0	4.0	6.0	10.0	143	65	282	0.27	0.23	0.50	0.27	0.50	0.15	0.23	0.38	0.15	
Jack Cork	ENG	MF	33	39	26	2,440	27.1	0	2	2	0	0	0	12	0	0.4	0.4	0.5	1.0	14	122	25	0.00	0.07	0.07	0.00	0.07	0.02	0.02	0.04	0.02	
Vitinho	BRA	DF,FW	23	35	23	2,097	23.3	3	1	4	3	0	0	1	0	3.5	3.5	1.7	5.1	67	55	161	0.13	0.04	0.17	0.13	0.17	0.15	0.07	0.22	0.15	
Johann Berg Gudmundsson	ISL	MF,FW	31	37	23	1,962	21.8	4	6	10	4	0	0	2	0	3.0	3.0	5.3	8.3	54	83	147	0.18	0.28	0.46	0.18	0.46	0.14	0.24	0.38	0.14	
Ashley Barnes	ENG	FW,MF	32	39	22	1,904	21.2	6	3	9	5	1	1	5	0	6.6	7.8	2.4	10.1	14	44	101	0.28	0.14	0.43	0.24	0.38	0.40	0.11	0.52	0.37	
Jay Rodriguez	ENG	FW	33	28	20	1,817	20.2	10	2	12	8	2	2	1	0	9.0	7.5	1.5	9.0	12	43	68	0.50	0.10	0.59	0.40	0.50	0.45	0.08	0.52	0.37	
Charlie Taylor	ENG	DF	28	33	17	1,725	19.2	0	1	1	0	0	0	3	0	0.3	0.3	1.0	1.2	12	63	8	0.00	0.05	0.05	0.00	0.05	0.01	0.05	0.06	0.01	
Benson Manuel	ANG	FW	25	33	14	1,490	16.6	11	3	14	11	0	0	3	0	4.0	4.0	4.1	8.1	118	43	196	0.66	0.18	0.85	0.66	0.85	0.24	0.25	0.49	0.24	
Hjalmar Ekdal	SWE	DF	23	9	9	767	8.5	1	1	2	1	0	0	1	0	0.5	0.5	0.5	1.0	10	42	0	0.12	0.12	0.23	0.12	0.23	0.06	0.05	0.12	0.06	
Ameen Al-Dakhil	BEL	DF	20	8	7	671	7.5	0	0	0	0	0	0	1	0	0.5	0.5	0.5	0.9	12	29	0	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.13	0.06	
Samuel Bastien	COD	MF,FW	25	10	7	551	6.1	1	1	2	1	0	0	0	0	0.8	0.8	0.5	1.2	8	24	33	0.16	0.16	0.33	0.16	0.33	0.12	0.08	0.20	0.12	
Scott Twine	ENG	MF	23	14	5	558	6.2	3	1	4	3	0	0	1	0	2.8	2.8	0.4	3.2	7	22	32	0.48	0.16	0.65	0.48	0.65	0.45	0.07	0.52	0.45	
Bailey Peacock-Farrell	NIR	GK	25	8	5	526	5.8	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Lyle Foster	RSA	FW	21	11	4	365	4.1	1	0	1	1	0	0	0	0	2.0	2.0	0.5	2.5	5	7	18	0.25	0.00	0.25	0.25	0.25	0.49	0.12	0.61	0.49	
Dara Costelloe	IRL	FW	19	4	3	208	2.3	0	0	0	0	0	0	0	0	0.3	0.3	0.1	0.4	9	4	26	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.18	0.13	
Darko Churlinov	MKD	FW	22	7	1	202	2.2	0	0	0	0	0	0	0	0	0.7	0.7	0.5	1.2	9	5	26	0.00	0.00	0.00	0.00	0.00	0.33	0.21	0.54	0.33	
Michael Obafemi	IRL	FW,MF	22	12	0	178	2.0	2	1	3	2	0	0	2	0	1.2	1.2	0.2	1.4	2	4	18	1.01	0.51	1.52	1.01	1.52	0.59	0.10	0.69	0.59	
Halil Dervişoğlu	TUR	MF,FW	22	8	0	72	0.8	1	0	1	1	0	0	0	0	0.6	0.6	0.4	1.0	2	1	8	1.25	0.00	1.25	1.25	1.25	0.73	0.51	1.24	0.73	
CJ Egan-Riley	ENG	MF,DF	19	3	0	21	0.2	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0.0	0	0	1	0.00	0.00	0.00	0.00	0.00	0.09	0.07	0.16	0.09	
Luke McNally	IRL	DF,MF	22	2	0	3	0.0	0	0	0	0	0	0	0	0	0.1	0.1	0.0	0.1	0	0	0	0.00	0.00	0.00	0.00	0.00	3.29	0.00	3.29	3.29	
Charlie Cresser	ENG	GK	17	0	0																											
Owen Dodgson	ENG	DF,MF	19	0	0																											
Denis Franchi	ITA	GK	19	0	0																											
Kevin Long	IRL	DF	31	0	0																											
Matthew Lowton	ENG	DF	33	0	0																											
Adam Phillips	ENG	MF	24	0	0																											
Squad Total			**25.5**	**46**	**506**	**4,140**	**46.0**	**85**	**61**	**146**	**82**	**3**	**3**	**81**	**2**	**66.2**	**63.9**	**48.2**	**112.1**	**966**	**1983**	**1961**	**1.85**	**1.33**	**3.17**	**1.78**	**3.11**	**1.44**	**1.05**	**2.49**	**1.39**	
Opponent Total			**26.0**	**46**	**506**	**4,140**	**46.0**	**34**	**22**	**56**	**32**	**2**	**3**	**99**	**7**	**38.2**	**35.9**	**27.1**	**63.0**	**426**	**1007**	**994**	**0.74**	**0.48**	**1.22**	**0.70**	**1.17**	**0.83**	**0.59**	**1.42**	**0.78**	

FIGURA 5.1. Jogadores da *Championship* 2022-2023.



FBREF

Enter Person, Team, Section, etc · Search

Players · Clubs · Competitions · Countries · Matches · Stathead · Languages · Mailing List · Full Site Menu below

Daily trivia for the English football fan: play Immaculate Footy England now

Become a Stathead & surf this site ad-free

2023-24 Premier League 2 — Division 1 Ch...

Taylor Harwood-Bellis

Position: DF (CB) • **Footed:** Right

188cm, 78kg (6-2, 172lb)

Born: January 30, 2002 (Age: 22-221d) in England, United Kingdom

Youth National Team: England

Club: Southampton

Wages: £ 7,500 Weekly Expires June 2028. Via Capology

2024-2025	MP	Min	Gls	Ast	xG	npxG	xAG	SCA	GCA
Premier League	3	220	0	0	0.0	0.0	0.1	3	0

* see our coverage note

Taylor Harwood-Bellis Overview · Stats by Competition · Match Logs · Goal Logs · Scouting Report · Stathead/Compare · Southampton

On this page:

Scouting Report · Standard Stats · Player News · Shooting · Passing · Pass Types · Goal and Shot Creation

Defensive Actions · Possession · Playing Time · Miscellaneous Stats · Player Club Summary · Leaderboard Appearances, Awards, and Honors · Wages

Full Site Menu

FIGURA 5.2. Página do perfil de um jogador da *Championship* 2022-2023.

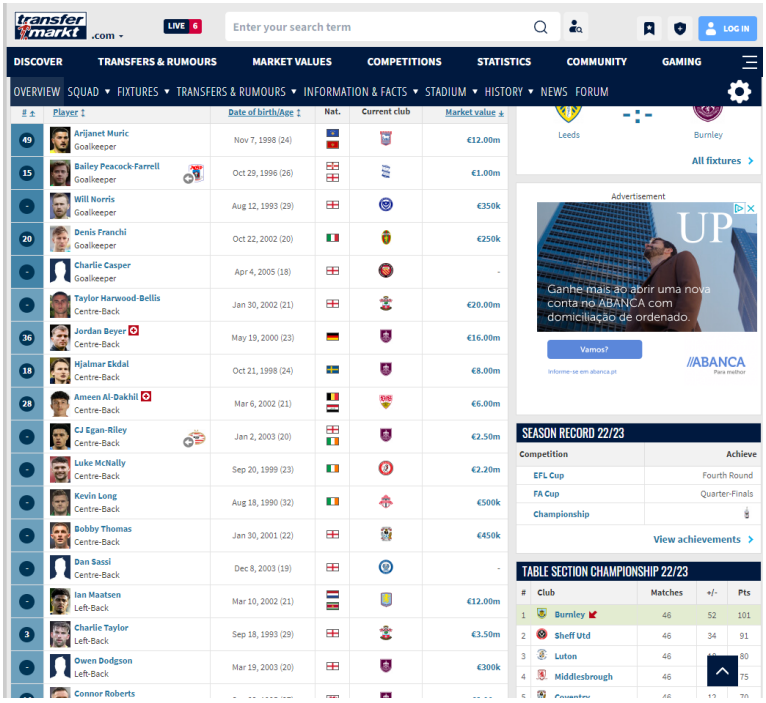


FIGURA 5.3. Página do *Transfermarkt* dos jogadores da *Championship* 2022-2023.

FIGURA 5.4. Perfil de um jogador do *Transfermarkt* da *Championship* 2022-2023.

FIGURA 5.5. *Ranking* Mundial da página *FootballDatabase*.

Anexo B – Variáveis

Tabela 5.1. *Transfermarkt*

Nome da Coluna	**Descrição da Coluna**
Date_Market_Value	Data do valor de mercado
Year_Market_Value	Ano do valor de mercado
Month_Market_Value	Mês do valor de mercado
Price	Preço

Tabela 5.2. *FootballDatabase*

Nome da Coluna	**Descrição da Coluna**
Year	Ano da classificação do *ranking*
Month_Market_Value	Mês da classificação do *ranking*
Ranking	Classificação do clube
Points	Pontos do Clube
Club	Nome do clube

Tabela 5.3. FBRef Dados dos Jogadores

Nome da Coluna	**Descrição da Coluna**
Player_Name	Nome do jogador
Player_ID_Transfer	ID do jogador para transferências
Position	Posição do jogador
Birth_Year	Ano de nascimento
Birth_Date	Data de nascimento
Squad	Equipa
Level	Nível em que joga
Age	Idade
Footed	Pé dominante
Country_born	País de nascimento
National_Team	Seleção nacional

Tabela 5.4. FBRef *Standard Stats*

Nome da Coluna	Descrição da Coluna
Season	Temporada
Squad	Equipa
Comp	Competição
Player_ID	ID do jogador
Country	País
MP	Partidas jogadas
Starts	Partidas iniciadas como titular
Min	Minutos jogados
90s	Número de partidas completas (90 minutos)
Gls	Golos marcados
Ast	Assistências
G+A	Golos + Assistências
G-PK	Golos, excluindo os de penáltis
PK	penáltis convertidos
PKatt	Tentativas de penáltis
CrdY	Cartões amarelos
CrdR	Cartões vermelhos
xG	Golos esperados
npxG	Golos esperados excluindo os de penáltis
xAG	Assistências esperadas
npxG+xAG	Golos esperados (excluindo penáltis) + Assistências esperadas
PrgC	Conduções progressivas
PrgP	Passes progressivos
PrgR	Passes progressivos recebidos
Gls.1	Golos marcados por 90 minutos
Ast.1	Assistências por 90 minutos
G+A.1	Golos + Assistências por 90 minutos
G-PK.1	Golos (excluindo penáltis) por 90 minutos
G+A-PK	Golos + Assistências (excluindo penáltis) por 90 minutos
xG.1	Golos esperados por 90 minutos
xAG.1	Assistências esperadas por 90 minutos
xG+xAG	Golos esperados + Assistências esperadas por 90 minutos
npxG.1	Golos esperados (excluindo penáltis) por 90 minutos
npxG+xAG.1	Golos esperados (excluindo penáltis) + Assistências esperadas por 90 minutos
Matches	Partidas

Tabela 5.5. FBRef *Playing Time*

Nome da Coluna	**Descrição da Coluna**
Mn/MP	Minutos por partida
Min%	Percentagem de minutos jogados
Mn/*Start*	Minutos por partida iniciada
Compl	Partidas completas
Subs	Substituições
Mn/Sub	Minutos por substituição
unSub	Não utilizado (no banco)
PPM	Pontos por partida
onG	Golos marcados pela equipa com o jogador em campo
onGA	Golos sofridos pela equipa com o jogador em campo
+/-	Saldo de golos com o jogador em campo
+/-90	Saldo de golos por 90 minutos
On-Off	Diferença de desempenho da equipa com e sem o jogador
onxG	Golos previstos da equipa com o jogador em campo
onxGA	Golos sofridos previstos da equipa com o jogador em campo
xG+/-	Saldo de xG
xG+/-90	Saldo de xG por 90 minutos
On-Off.1	Diferença de desempenho da equipa por 90 minutos

Tabela 5.6. FBRef *Miscellaneous Stats*

Nome da Coluna	**Descrição da Coluna**
2CrdY	Dois cartões amarelos (expulsão)
Fls	Faltas cometidas
Fld	Faltas sofridas
Off	Foras de jogo
Crs	Cruzamentos
Int	Interceptações
TklW	Desarmes bem-sucedidos
PKwon	penáltis ganhos
PKcon	penáltis concedidos
OG	Autogolos
Recov	Recuperações de bola
Won	Duelos aéreos ganhos
Lost	Duelos aéreos perdidos
Won%	Percentagem de duelos ganhos

Tabela 5.7. FBRef *Pass Types*

Nome da Coluna	Descrição da Coluna
Att	Tentativas de passes
Live	Bolas vivas
Dead	Bolas paradas
FK	Tiros livres
TB	Passes em profundidade
Sw	Trocas de jogo (passes longos laterais)
Crs_tpass	Cruzamentos (passes)
TI	Arremessos laterais
CK	Cantos
In	Passes para dentro da área (cantos)
Out	Passes para fora da área (cantos)
Str	Passes diretos (cantos)
Cmp	Passes completos
Off_tpass	Passes que resultaram em fora de jogo
Blocks	Passes bloqueados

Tabela 5.8. FBRef *Defensive Actions*

Nome da Coluna	Descrição da Coluna
Tkl	Tentativas de desarme
TklW_def	Desarmes bem-sucedidos (defesa)
Def 3rd	Ações de desarme no terço defensivo
Mid 3rd	Ações de desarme no terço médio
Att 3rd	Ações de desarme no terço ofensivo
Tkl.1	Número de dribladores desafiados
Att_def	Tentativas de desarme (defesa)
Tkl%	Percentagem de desarmes bem-sucedidos
Lost_def	Duelos defensivos perdidos
Blocks_def	Bloqueios defensivos
Sh	Número de remates que bloqueou
Pass	Passes bloqueados
Int_def	Interceptações (defesa)
Tkl+Int	Desarmes + Interceptações
Clr	Defesas
Err	Erros que levaram a um remate

Tabela 5.9. FBRef *Goal and Shot Creation*

Nome da Coluna	**Descrição da Coluna**
SCA	Ações que criaram remates
SCA90	Ações que criaram remates por 90 minutos
PassLive	Passes completos em jogo que levaram à criação de um remate
PassDead	Passes de bola parada que levaram à criação de um remate
TO	Dribles bem sucedidos que levaram à criação de um remate
Sh_gsc	Remates que levaram à criação de um remate
Fld_gsc	Faltas que levaram à criação de um remate
Def	Ações defensivas que levaram à criação de um remate
GCA	Ações que criaram golos
GCA90	Ações que criaram golos por 90 minutos
***PassLive*.1**	Passes completos em jogo que levaram à criação de golo
***PassDead*.1**	Passes de bola parada que levaram à criação de golo
TO.1	Dribles bem sucedidos que levaram à criação de golo
Sh.1	Remates que levaram à criação de golo
Fld.1	Faltas que levaram à criação de golo
Def.1	Ações defensivas que levaram à criação de golo

Tabela 5.10. FBRef *Possession*

Nome da Coluna	Descrição da Coluna
Touches	Toques na bola
Def Pen	Toques na área defensiva
Def 3rd_poss	Posse no terço defensivo
Mid 3rd_poss	Posse no terço médio
Att 3rd_poss	Posse no terço ofensivo
Att Pen	Toques na área ofensiva
Live_poss	Toques em bola viva
Att_poss	Tentativas de drible
Succ	Dribles bem-sucedidos
Succ%	Percentagem de dribles bem-sucedidos
Tkld	Número de vezes em que foi abordado durante uma tentativa de ataque
Tkld%	Percentagem do número de vezes em que foi abordado durante uma tentativa de ataque
Carries	Conduções de bola
TotDist	Distância total de condução de bola
PrgDist	Distância progressiva de condução
PrgC_poss	Conduções progressivas (posse)
1/3	Conduções para o último terço
CPA	Conduções para a área de penálti
Mis	Controlos de bola errados
Dis	Desarmado durante a condução de bola
Rec	Passes recebidos
PrgR_poss	Receção de passes progressivos (posse)

Tabela 5.11. FBRef *Passing*

Nome da Coluna	Descrição da Coluna
Cmp_pass	Passes concluídos
Att_pass	Passes tentados
Cmp%	Percentagem de passes concluídos
TotDist_pass	Distância total dos passes
PrgDist_pass	Distância progressiva dos passes
Cmp.1	Passes curtos concluídos
Att.1	Passes curtos tentados
Cmp%.1	Percentagem de passes curtos concluídos
Cmp.2	Passes médios concluídos
Att.2	Passes médios tentados
Cmp%.2	Percentagem de passes médios concluídos
Cmp.3	Passes longos concluídos
Att.3	Passes longos tentados
Cmp%.3	Percentagem de passes longos concluídos
Ast_pass	Assistências (passes)
xAG_pass	Golos com Assistências esperadas (passes)
xA	Assistências esperadas
A-xAG	Assistências menos golos esperados
KP	Passes-chave
1/3_pass	Passes para o último terço
PPA	Passes para a área de penálti
CrsPA	Cruzamentos para a área de penálti
PrgP_pass	Passes progressivos

Tabela 5.12. FBRef *Shooting*

Nome da Coluna	**Descrição da Coluna**
SoT	Remates enquadrados (*Shots on Target*)
SoT%	Percentagem de remates enquadrados
Sh/90	Remates por 90 minutos
SoT/90	Remates enquadrados por 90 minutos
G/Sh	Golos por remate
G/SoT	Golos por remate enquadrado
Dist	Distância média dos remates
FK	Golos de livre direto
xG	Golos esperados
npxG	Golos esperados excluindo penáltis
npxG/Sh	Golos esperados excluindo penáltis por remate
G-xG	Diferença entre golos marcados e golos esperados
np:G-xG	Diferença entre golos marcados (excluindo penáltis) e golos esperados (excluindo penáltis)

Anexo C – Mapa fatorial

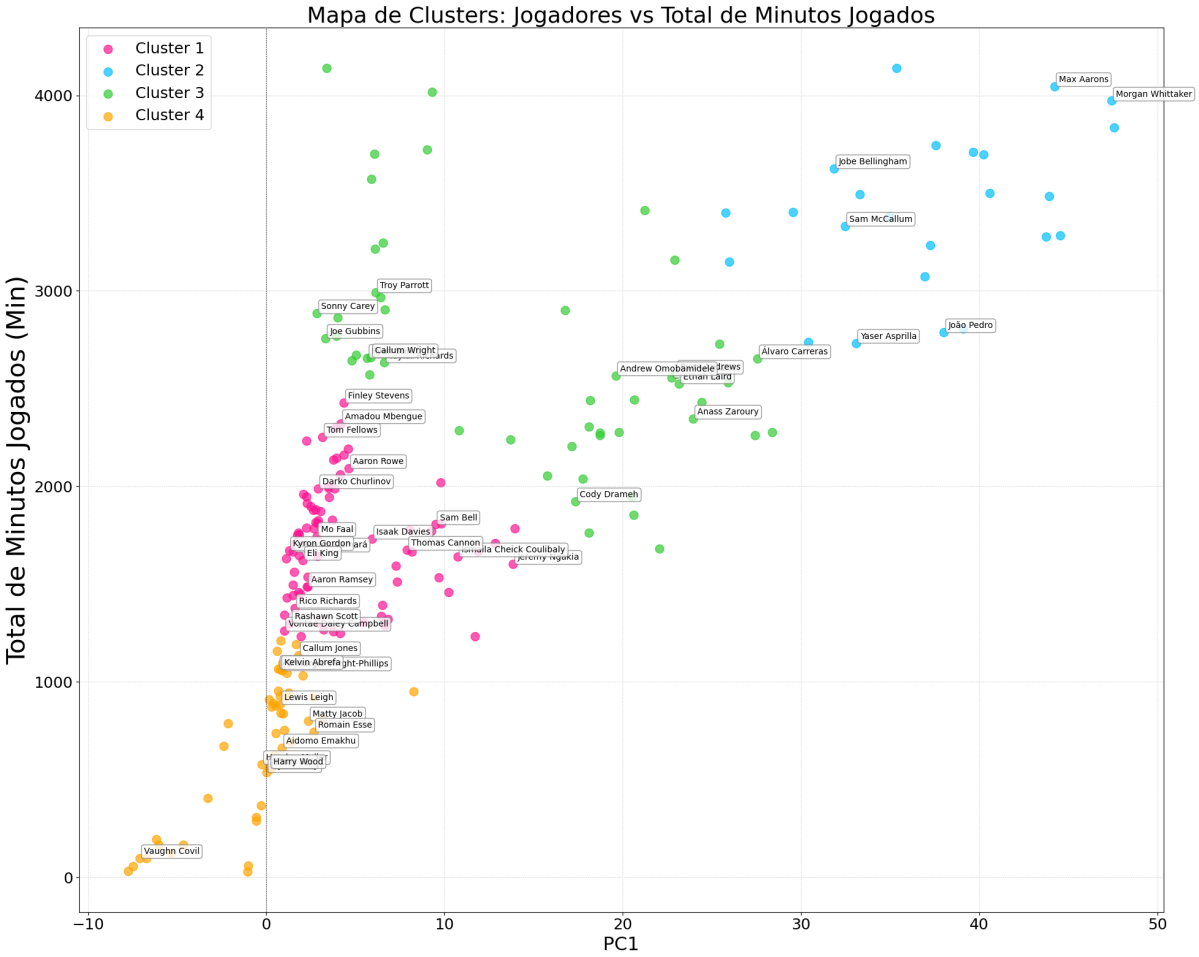


FIGURA 5.6. Mapa fatorial de *clusters* (Jogadores vs Total de minutos jogados)