INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Football Learning - Avaliação de Jogadores de Futebol usando Machine Learning

DOMINGOS BERNARDINO PEREIRA DA COSTA
Outubro de 2021

POLITÉCNICO
DO PORTO

Football Learning

Avaliação de Jogadores de Futebol usando Machine Learning

Domingos Bernardino Pereira da Costa

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas de Informação e Conhecimento

Orientador: Carlos Ferreira (cgf@isep.ipp.pt)

Júri:

Presidente: Fernando Jorge Duarte, Instituto Superior de Engenharia do Porto

Porto, Outubro de 2021

Resumo

Ao longo das últimas décadas, o futebol passou por uma evolução que transformou o que era um desporto para um negócio com enorme impacto social e financeiro. Essa transformação, aliada à constante necessidade de obter sucesso desportivo, criou a necessidade de inovação por parte de um clube desportivo de forma a distanciar-se dos seus adversários e de conquistar títulos e alcançar mais lucro financeiro.

O principal objetivo desta dissertação passa pela criação de um rating de qualidade de um jogador com recurso a um modelo de previsão de resultados de futebol com apenas duas possibilidades, jogo ganho ou jogo não ganho, em que serão retirados os pesos de estatísticas simples e avançadas, desde golos marcados a duelos aéreos ganhos.

Várias metodologias de machine learning foram estudadas para perceber quais as melhores para resolver os problemas de classificação e regressão presentes neste tema, tendo sido escolhidos aplicações de algoritmos de *support-vector machine* (SVM) e de regressão linear.

Os dados foram tratados de forma a retirar "lixo" que possa diminuir a performance da fase de modelação, que consistiu na criação de vários modelos divididos por competição, bem como um modelo com todos os jogos, aplicando os algoritmos estudados e comparando os resultados dos modelos e algoritmos entre si.

O modelo de previsão de resultados que utilizou regressão linear foi o escolhido para o cálculo dos *ratings* dos jogadores, visto que obteve uma taxa de acerto de 75.40% em comparação com o de SVM, que teve 75.00%. Com os pesos aplicados e os jogadores analisados, verificou-se que o modelo tem uma correspondência aceitável com bases de dados já existentes no mercado e com a opinião de conhecedores do negócio, alcançando uma percentagem de acerto de 72.22% num inquérito realizado a 11 pessoas ligadas ao futebol.

Palavras-chave: Machine Learning; Classificação; Regressão Linear; SVM; Futebol; Previsão de resultados

Abstract

In the last decades, football has gone through an evolution that transformed what was once a sport to a business with big social and financial impact. This transformation allied to the constant need of having sporting success has created the need of innovating by a football club so that it could distance itself from its opponents, win more titles and have bigger profits.

The main goal of this dissertation is to create a quality rating of a player using a football results prediction model with only two possibilities, game won or not, in which the weights of simple and advanced statistics will be taken and used, from goals scored to air duels won.

Several machine learning methodologies were studied to understand which are the best to solve the classification and regression problems present in this theme, having been chosen applications of support-vector machine (SVM) and linear regression algorithms.

The data was treated to remove any "garbage" that could diminish the modeling step performance, which consisted in the creation of various models divided by competition, as well as a model with all the games available, with the studied algorithms being applied and the results from the multiple models and algorithms being compared with each other.

The winning prediction model that used linear regression was chosen for the calculation of player ratings, since it obtained a 75.40% accuracy rate compared to SVM, which had 75.00%.

After the weights have been applied and the players analyzed, it was found that the model matches reasonably well with databases already in the market and with the opinion of experts in the field, reaching the percentage of 72.22% in a survey conducted with 11 people connected to football.

Keywords: Machine Learning; Classification; Linear Regression; SVM; Football; Results prediction

Agradecimentos

Agradeço aos meus pais por me terem educado e dado todos os instrumentos para que eu pudesse ter sucesso na minha vida e por me terem apoiado sempre.

Agradeço ao professor Carlos Ferreira pela orientação neste projeto e por me ter incentivado continuadamente na escrita e no desenvolvimento desta dissertação.

Agradeço todos os professores que me lecionaram desde que comecei a minha vida académica e se esforçaram imenso para instruir a mim e a muitos outros colegas meu.

Agradeço aos meus amigos por terem estado presentes nos bons e nos maus momentos, e pela ajuda indireta que me prestaram por discutir ideias comigo.

Por fim, agradeço ao Instituto Superior de Engenharia do Porto por me ter dado a chance de me formar, não só como engenheiro, mas também como pessoa.

Índice

Lista de Figuras

Lista de Equações

Lista de Tabelas

Lista de Excertos de Código

Acrónimos e Símbolos

Lista de Acrónimos

API	Application Programming Interface
CA	Capacidade Atual (Football Manager)
CP	Capacidade Potencial (Football Manager)
CRISP-DM	Cross Industry Standard Process for Data Mining
DCL	Data Control Language
DDL	Data Definition Language
DML	Data Manipulation Language
EWMA	Exponentially Weighted Moving Average
EWSA	Exponentially Weighted Smoothing Average
FFE	Fuzzy Front End
FIFA	Fédération Internationale de Football Association
FM	Football Manager
FURPS+	Functionality, Usability, Reliability, Performance, Supportability
GLM	Generalized Linear Models
HTML	HyperText Markup Language
IoT	Internet of Things
LSVC	Linear Support Vector Classification
NCD	New Concept Development
QFD	Quality Function Deployment
REST	Representational State Transfer
SAD	Sistema de Apoio à Decisão
SQL	Structured Query Language
SVM	Support-Vector Machines

UC	Use Case
UEFA	Union of European Football Associations

Lista de Símbolos

€	Euros
λ	Valor de suavização

1 Introdução

A presente dissertação está inserida na realização do Mestrado em Engenharia Informática, ramo de Sistemas de Informação e Conhecimento, do Instituto Superior de Engenharia do Porto. Este capítulo introdutório pretende descrever a contextualização do problema onde esta dissertação se insere, bem como apresentar o problema propriamente dito, os objetivos a alcançar e a abordagem a adotar para resolvê-lo. Por fim, a estrutura do documento é explicada.

1.1 Contexto

A evolução que o futebol teve nas últimas décadas, fruto do investimento massivo que tem sido feito no desporto, principalmente por parte de canais televisivos, transformou a modalidade em mais do que uma simples competição entre duas equipas de onze jogadores, tornando-se num negócio com implicações monetárias e sociais enormes. Os cinco principais campeonatos europeus, que correspondem a Inglaterra, Itália, Alemanha, Espanha e França, tiveram uma receita conjunta na época 1996/1997 de 2.497 milhões de euros, tendo subido para 18.100 milhões de euros, um crescimento de aproximadamente 625% ao longo dos últimos 24 anos (Lange, 2020). O Campeonato Europeu de 2020, a maior competição continental de seleções nacionais de países europeus, obteve uma audiência cumulativa de 5.23 mil milhões de espectadores e 7.5 mil milhões de visualizações e interações nas redes sociais, com a final a ser assistida por uma média de 328 milhões de pessoas globalmente (Jones, 2021).

Como tal, os clubes de futebol necessitam de ter uma harmonia entre o sucesso económico, visto que são efetivamente uma empresa com múltiplos funcionários, com salários para pagar e obrigações fiscais a cumprir, e o sucesso desportivo, já que o objetivo de uma equipa de futebol é conquistar títulos e atrair o maior número de fãs. Para alcançar esse sucesso, o caminho a seguir é fácil de perceber: ganhar partidas de futebol, que se resume a marcar mais golos que o adversário, com recurso aos melhores jogadores a jogar na equipa. A simplicidade, no entanto, é apenas na perceção do objetivo de se dirigir uma equipa de futebol. Sendo o

futebol um desporto constituído por um sistema com dois subsistemas (se ignorarmos a equipa de arbitragem) de onze elementos, que se modifica durante 90 minutos e sem turnos, a complexidade sobe exponencialmente e complica a procura por soluções e formas de descobrir os melhores jogadores, sistemas e formas de praticar futebol (Balagué et al., 2013).

Uma das formas de procurar as melhores soluções para se sobressair num jogo de futebol que tem vindo a ser mais explorada nos últimos anos é aplicar ferramentas de *data mining* e *machine learning* em dados estatísticos brutos de partidas de futebol, quer a tempo-real, quer dados históricos, para detetar padrões, agrupar jogadores semelhantes e avaliar jogadores da forma menos enviesada possível. Clive Humbly, matemático e criador do programa de recompensas da cadeira de supermercados Tesco, cunhou em 2006 a frase "Dados são o novo petróleo" (Arthur, 2013), de forma a simbolizar o seu valor para as empresas, e o desporto, sendo cada vez mais um negócio, acompanha esse pensamento.

O crescimento do uso de *data mining* no desporto em geral tem acompanhado o *boom* da utilização e do interesse em ferramentas de *machine learning*, que resulta do aumento exponencial da quantidade de dados disponíveis e da capacidade de processamento dos computadores e da descida no custo de armazenamento, conforme demonstrado no gráfico da Figura 1.

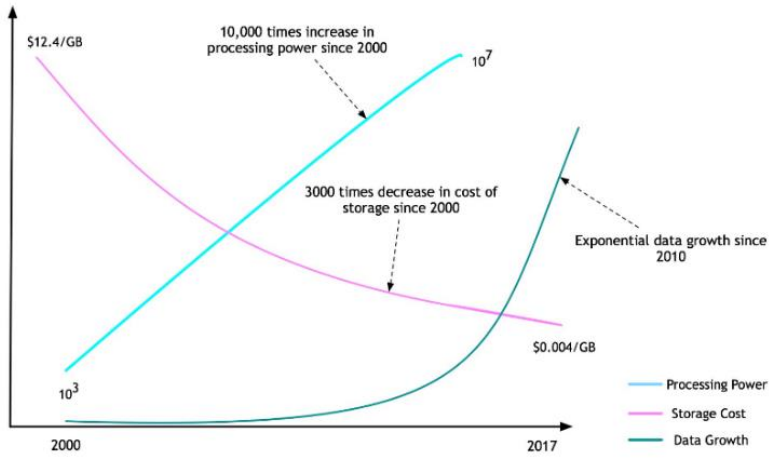


Figura 1 – Variação do preço de armazenamento, da capacidade de processamento e dos dados disponíveis (Menon, 2018)

Esta evolução tecnológica permitiu que a área de abordagem de *machine learning e data mining*, que antes faziam parte de um nicho de profissionais e que era impraticável efetuar com computadores "normais", se tornasse mais popular e acessível a qualquer pessoa curiosa sobre o tema, como se pode verificar no gráfico presente na Figura 2, que compara a popularidade de temas associados a este ramo, nomeadamente *machine learning*, *data science*, *big data*, *deep learning* e *data analysis*.

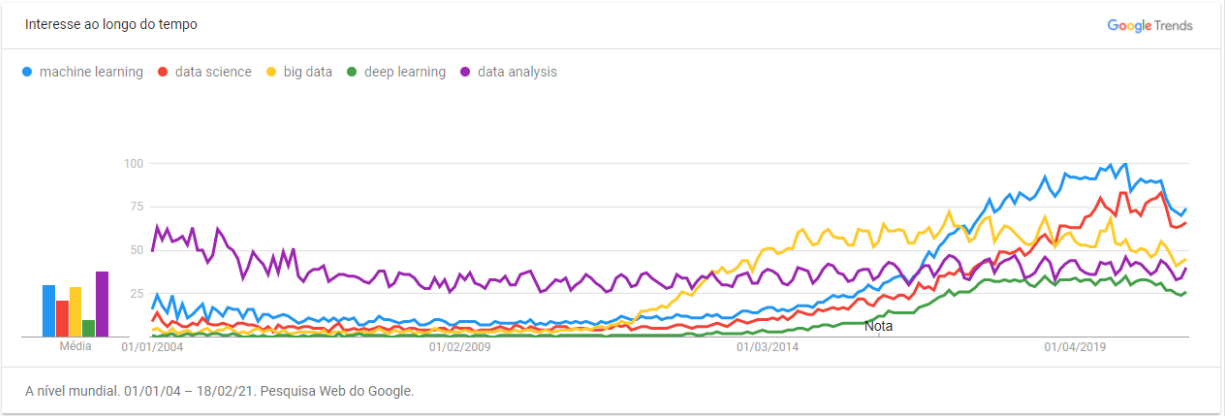


Figura 2 – Comparação de interesse de termos de *data mining*

1.2 Problema

O treinador de futebol e o responsável pelo recrutamento de novos jogadores vivem num mundo em que a corrida por novas tendências, quer de mercado, quer de esquemas e estratégias táticas, é apertada e repentina, o que significa que as entidades com poder de decisão precisam de ter, além do seu *know-how* técnico, ferramentas que lhes permitam prever quais serão os jogadores que melhor se encaixam no sistema da equipa. Além disso, uma equipa tem de estar sempre atenta ao mercado de forma a descobrir os próximos grandes atletas. Essa procura, tradicionalmente, é feita por profissionais de prospeção de jogadores, vulgarmente chamados de olheiros, e é um processo heurístico, ou seja, com base na experiência anterior, daí normalmente os olheiros serem antigos jogadores ou treinadores, que devido ao seu trajeto no desporto, conseguem detetar talento e interpretar a subjetividade do futebol melhor que uma pessoa sem essa passagem pela modalidade. No entanto, há a falta de dados objetivos que comprovem os juízos de qualidade e valor efetuados, pelo que se torna necessário haver uma ferramenta que auxilie a tarefa, com recurso a estatísticas e dados reais, de forma a consolidar a observação efetuada.

1.3 Objetivos

O objetivo deste projeto consiste no estudo aplicado ao futebol da utilização de ferramentas de *data mining* e *machine learning* de forma a obter mais informação auxiliar para as tarefas de gestão e prospeção de uma equipa profissional. Como tal, é idealizado o desenvolvimento de um sistema de apoio à decisão que auxilie na procura dos melhores jogadores, com recurso a ferramentas de *data mining* e *machine learning*, utilizando dados estatísticos reais dos jogadores, a duas entidades com poder de decisão no contexto de um clube de futebol: o treinador e o responsável da captação de novos jogadores, como o diretor desportivo ou o observador chefe.

Sendo assim, é possível dividir o projeto em dois objetivos principais:

1. Criar um modelo de previsão de resultados, com o objetivo de obter o peso calculado em cada uma das variáveis inseridas nesse modelo, com vista a utilizá-las nas estatísticas dos jogadores e calcular um valor absoluto do impacto do jogador em campo. Todos os dados a utilizar serão adquiridos a partir de APIs que capturam estatísticas de todos os jogos dos principais campeonatos europeus. Esse modelo será validado com bases de dados já existentes no mercado, bem como com a opinião de pessoas que percebam do negócio. Este objetivo servirá como prova de conceito para perceber se o objetivo seguinte pode avançar.
2. Desenvolver uma aplicação web alimentada pelo modelo, cuja principal função será fornecer os melhores candidatos, isto é, os jogadores, que se enquadrem nos filtros definidos, utilizando ferramentas gráficas e seletivas que permitam comparar estatísticas simples de jogadores, como passes completados ou desarmes efetuados, estatísticas mais complexas, como duelos ganhos, ou estatísticas mais subjetivas, como passes-chave. Além disso, a aplicação terá acesso, para cada jogador, a um valor correspondente a um *rating* obtido a partir da aplicações de ferramentas e abordagens de *machine learning* e *data mining* referidos no início desta secção.

O foco do documento será dirigido principalmente ao primeiro objetivo, dada a importância de criar e validar um modelo de dados que seja o "segredo" de uma aplicação como a descrita no ponto 2, e devido ao ramo de mestrado no qual a unidade curricular se insere.

1.4 Abordagem

A abordagem utilizada no documento, e por consequente no projeto final, é a metodologia CRISP-DM, sigla de *Cross-Industry Standard Process for Data Mining*, um processo standard de abordagem a projetos de *data mining* (Shearer, 2000). A metodologia tem seis passos, como se pode ver na Figura 3. O primeiro passo, *Business Understanding*, teve como tarefas uma pesquisa sobre o tema central, nomeadamente a evolução que o futebol foi tendo desde uma prática desportiva até ser um negócio com mais implicações económicas, além do crescimento da componente de análise estatística no contexto de futebol. Foi efetuada também uma pesquisa sobre as tecnologias que melhor se aplicam ao tema a ser estudado, e sobre trabalho relacionado, quer a nível de artigos e investigações científicas, quer a nível de projetos comerciais. Foram explorados algoritmos que criassem modelos de classificação para resolver o problema da previsão de resultado, como regressão linear ou máquinas vetoriais de suporte, e também abordagens para a retirada dos pesos obtidos nesses modelos. O segundo passo, *Data Understanding*, foi a fase em que o caso de estudo foi aprofundado, ou seja, os dados que irão ser usados são interpretados de forma a perceber o que cada um significa e que valor poderá ser retirado. Este passo, como o gráfico indica, implicou o regresso ocasional à primeira fase, de forma a procurar e adicionar mais valor para o projeto. O passo seguinte, *Data Preparation*, consiste no aprofundamento da análise referida anteriormente, em que os dados foram comparados entre si de forma a detetar correlações ou inconsistências, para que

os modelos criados contivessem o menor "lixo" possível. Na fase de *Modeling*, os modelos propriamente ditos são criados com base nos dados anteriormente tratados e, neste caso em concreto, as previsões obtidas e os pesos calculados nos modelos são retirados com vista a escolher o modelo que melhor se adequa para o problema proposto e para preparar a fase seguinte, a de *Evaluation*, em que os resultados do modelo escolhido são avaliados com base em comparações com valores de outros produtos no mercado, bem como com o ponto de vista de especialistas. A fase de *Deployment* considera-se o documento atual e a aplicação web onde os dados obtidos a partir do modelo poderão ser consultados.

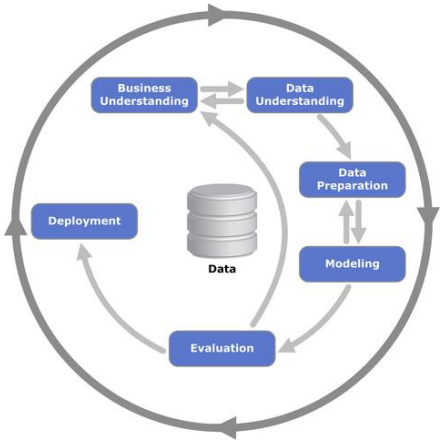


Figura 3 – Diagrama de CRISP-DM (Shearer, 2000)

1.5 Estrutura do Documento

O documento presente tem uma estrutura organizada em sete capítulos. O capítulo atual, que corresponde à Introdução, pretende expor o projeto realizado e contextualizar o problema identificado. O segundo capítulo diz respeito ao Contexto e ao Estado de Arte, onde é apresentada a pesquisa efetuada para perceber melhor o tema do documento, bem como descobrir as melhores ferramentas possíveis para usar. A Análise de Valor corresponde ao terceiro capítulo, que contém o valor identificado da aplicação para o cliente, os benefícios e sacrifícios, bem como uma comparação de soluções já existentes no mercado com a ideia de produto contemplada para este projeto. O quarto capítulo é Análise e Design, onde é desenhada a solução pensada para o problema exposto nos capítulos iniciais. O Caso de Estudo é o quinto capítulo, onde o conhecimento obtido nos capítulos anteriores será aplicado com o fim de resolver o problema exposto no início do documento, com recurso às melhores práticas de desenvolvimento de um modelo de *machine learning*. Os dados serão analisados, tratados, modelados e validados, seguindo a abordagem identificada na secção anterior. Por fim, um capítulo dedicado à Conclusão, em que se aborda o problema referido inicialmente e tiram-se conclusões sobre o projeto, com foco nos ganhos que se pode obter com soluções como a apresentada, bem como em melhorias futuras e abordagens diferentes para o projeto.

2 Contexto e Estado de Arte

Neste capítulo, pretende-se explorar o tema do futebol, com uma introdução à forma como passou de ser uma prática desportiva amadora para um negócio que movimenta imenso dinheiro e com fãs em todo o mundo, de forma a contextualizar o leitor na necessidade da inovação em qualquer que seja a área futebolística, passando depois para o ramo no qual o documento se enquadra, que é a análise de dados estatísticos com vista a retirar valor para os decisores num clube de futebol. De seguida, aborda-se as técnicas e ferramentas mais utilizadas de *machine learning* e *data mining* em soluções deste género, bem como uma breve explicação do que é um sistema de apoio à decisão e as normas a seguir. Por fim, é efetuado um levantamento de ferramentas usadas no mercado para resolver o problema exposto e de artigos científicos que exploram o tema.

2.1 Evolução do Futebol

Como falado no capítulo da introdução, os clubes de futebol necessitam ter uma harmonia entre o sucesso económico (secção 1.1), visto que são efetivamente uma empresa com múltiplos funcionários, com salários para pagar e obrigações fiscais a cumprir, e o sucesso desportivo, já que o objetivo de uma equipa de futebol é conquistar títulos e atrair o maior número de fãs. No entanto, esta visão económica e corporativa não é recente. William McGregor, o fundador da *Football League*, a primeira competição organizada de futebol, já descrevera o futebol como um grande negócio em 1905 (Szymanski and Kuypers, 1999). O Peñarol, a equipa com mais títulos conquistados no Uruguai, colocou um patrocínio na sua camisola de jogo na década de 1950, sendo considerado o primeiro clube no mundo a lucrar com publicidade na sua camisola, algo normal nos dias de hoje e que tem vindo a crescer (Kalt, 2015), como se pode ver na Figura 4. Fabio Chisari escreveu em 2006 que o Campeonato Mundial de Futebol de 1966 foi o ponto de viragem na ligação entre futebol e televisão, tendo sido a primeira grande competição a ser transmitida via satélite (Chisari, 2006). Martin Edwards, Irving Scholar e David Dein, três dirigentes desportivos ingleses da década de 1980,

foram revolucionários na sua visão de gestão de um clube de futebol, rejeitando "a visão tradicional do desporto como um negócio peculiar em que o clube é gerido como uma utilidade pública e não como uma organização com fins lucrativos" e tendo o objetivo de "procurar lucro a partir da comoditização do jogo" (Taylor, 2013).

Figura 4 – Crescimento do valor da publicidade na Europa (Noble, 2016)

Os dirigentes desportivos têm de pensar a gestão de um clube de uma forma empresarial e empreendedora. Com a globalização e enorme fluxo de dinheiro no futebol, é essencial ter sempre uma visão inovadora e de tentar sempre estar à frente da concorrência (Tjønndal, 2017). Uma empresa, segundo Peter Drucker, tem de se manter empreendedora para lá da sua fundação, com o risco de se tornar "tímida e retrógrada" e perder a sua qualidade essencial (Drucker, 2006), e aplicando isto para futebol, essa visão com foco na inovação torna-se ainda mais importante, visto ser um mundo volátil a nível de processos, subjetivo a nível de resultados e em que uma época desportiva má pode condenar um clube (Haugen, 2012). Existe o exemplo do clube inglês Leeds United, que efetuou no início do século XXI um investimento avultado, aumentando a sua dívida, para se colocar no topo do futebol inglês, e uma simples eliminação europeia e posterior não qualificação para a Liga dos Campeões da UEFA impediu o clube de controlar as suas finanças, resultando na despromoção para a segunda divisão inglesa em 2004 e depois para a terceira divisão em 2007, cunhando o termo "*doing a Leeds*", que é atribuído sempre que um clube arrisca imenso para ter sucesso e falha de forma retumbante (Merrick, 2010). O oposto também existe, em que um clube tem uma oportunidade única para se distanciar dos restantes e aproveita-a, como é o caso do Bayern de Munique. A equipa alemã, antes da década de 60, era uma equipa humilde e com poucos títulos, não participando sequer na primeira época da *Bundesliga,* a principal liga alemã. No entanto, com o surgimento de três dos melhores jogadores alemães da história nas suas camadas jovens, Sepp Maier, Franz Beckenbauer e Gerd Muller, o clube elevou o seu nível para ser o clube alemão com mais títulos. (FC Bayern, 2014)

Sendo o futebol um desporto, uma das formas de um clube inovar e se distanciar dos seus rivais e adversários acaba por ser no trabalho de campo, isto é, no desporto jogado. O

treinador do clube, devido ao seu papel como responsável pela forma como a equipa se apresenta dentro de campo, é um dos principais agentes de inovação, e a história do futebol é rica em treinadores que idealizaram novas estratégias táticas que revolucionaram o futebol. Nos primórdios do futebol, o desporto era baseado na capacidade de drible, em que os jogadores recebiam a bola e driblavam em direção à baliza adversária, enquanto que os restantes membros da equipa recuavam para proteger a sua baliza no caso de haver algum contra-ataque, reduzindo o passe a uma prática de último recurso (Cox, 2016). Em 1870, a equipa escocesa de futebol Queen's Park dominava o futebol escocês com o seu "jogo de combinação", baseado no trabalho de equipa e favorecendo a troca de bola entre os jogadores ao invés das capacidades individuais de drible (Garnham, 2004), o que resultou num domínio escocês no final do século XIX em relação à sua rival Inglaterra (Cox, 2016). O treinador do Arsenal na década de 1930 Herbert Chapman, aquando a modificação da regra de fora-de-jogo em 1925 que passou a permitir de três defesas para dois entre a bola e o recipiente do passe (LaBlanc and Henshaw, 1994), percebeu que era necessário adaptar a esta nova lei. O desporto, com esta mudança na lei, estava mais ofensivo, tendo subido de 2,65 golos por jogo na época 1925/26 para 3,69 na época seguinte, sendo a tática mais utilizada o 2-3-5 (Wilson, 2018). Herbert Chapman decidiu, então, recuar um dos médios para a defesa, de forma a ganhar consistência defensiva, passando a jogar num 3-2-5, ou na tática WM, como é vulgarmente conhecida. Esta mudança, incentivada pela oportunidade surgida com a alteração da lei de fora-de-jogo, transformou o desporto de uma prática ofensivamente vertiginosa para um jogo mais estratégico, e colocou o Arsenal de Chapman no topo do futebol britânico (Barclay, 2014). Mais recentemente, em 2008, Pep Guardiola, um pupilo de Johan Cruyff, que também revolucionou o futebol nos anos 70 como jogador e nos anos 90 como treinador com a sua forma de jogar de "Futebol Total", que consistia na fluidez dos onze jogadores, em que todos atacavam e defendiam, independentemente da sua posição original (Shetty, 2018), começou a treinar o FC Barcelona e popularizou o estilo e filosofia *tiki-taka*, que consistia no controlo massivo da posse-de-bola e no foco em passes curtos e movimentos constantes (Davies, 2013). O sucesso estrondoso do Barcelona nas épocas em que Guardiola treinou a equipa fez com que o futebol europeu se concentrasse na retenção da bola, como se pode ver na Figura 5, em que o número de passes aumentou desde 2015 para 2019 nas competições de maior nível e, consequentemente, teve também repercussões no crescimento da utilização de táticas de contra-ataque e futebol direto para contrariar equipas que usassem um estilo focado mais na posse de bola (Cox, 2016).

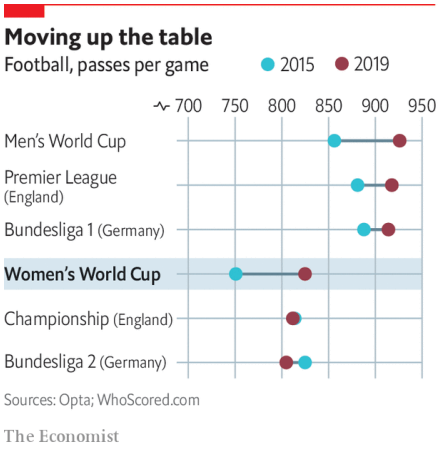


Figura 5 – Aumento no número de passes de 2015 para 2019. (Steinglass, 2019)

No entanto, não é só no trabalho tático que um treinador e equipa técnica poderão inovar. A vertente de treino e de análise pré e pós jogo revelam-se igualmente importantes, e o processo científico é usado nesses campos. Vítor Frade afirma que não há divisão entre o treino e o jogo de futebol, já que considera que o crescimento tático dos jogadores de uma equipa implica evoluções físicas, psicológicas e técnicas com vista a ir ao encontro de uma proposta de jogo pretendida, e para isso, os exercícios de treino apresentam uma importância elevada, visto que devem representar os acontecimentos que mais surgem num jogo (Lima, 2014). Um dos expoentes máximos da evolução de treino com recurso ao uso de metodologias científicas é Valeriy Lobanovskyi, o treinador com o maior número de trofeus ganhos no século XX (FourFourTwo, 2020). Além de ter sido um jogador de futebol antes de enveredar pelo cargo de treinador, Valeriy também era estudante de engenharia no Instituto Politécnico de Kiev, onde se apercebeu do potencial que os computadores tinham para resolver problemas com uma abordagem sistemática, e foi a partir de uma conversa com Anatoliy Zelentsov, um analista estatístico e cientista do departamento de educação física do Instituto Estatal de Kiev, que começou a ler o jogo como um sistema de 22 elementos, dois subsistemas de 11 elementos, que se movem numa área definida, o relvado, e sujeitos a uma série de restrições, as leis de jogo. Se os dois subsistemas fossem iguais, o resultado seria um empate e se um dos subsistemas fosse mais forte, seria esse o vencedor da partida. O que fascinou Lobanovskyi foi a peculiaridade de que a eficiência do subsistema, a equipa, era maior do que a soma das eficiências dos elementos que a constituíam, os jogadores, e isto levou-o a concluir que as técnicas cibernéticas ensinadas no Politécnico de Kiev poderiam ser aplicadas ao desporto (Wilson, 2011). Lobanobskyi é creditado por ser dos primeiros a focar-se fortemente na dieta dos jogadores, de forma a que a sua equipa alcance um pico de capacidade física, resultando numa grande capacidade das suas equipas em cumprirem as instruções táticas com grande precisão (Spedding, 2019). Foi também importante na evolução da componente estatística do jogo, que será detalhada na secção 2.2.

Além do trabalho tático e científico que uma equipa poderá fazer para inovar, existe a vertente extracampo, que vai desde a ter um modelo de negócio que consiste na compra e venda de jogadores de forma a fazer lucro, até à expansão da equipa como uma marca (Szabados, 2003). Szabados e, posteriormente, Krisztina Andras e Zsolt Havran dividiram os tipos de estratégia que uma equipa de futebol pode ter em cinco, com focos diferentes no Sucesso Desportivo (SD) e no Sucesso Financeiro (SF), como pode ser visto na Tabela 1 (Andras and Havran, 2015):

Tabela 1 – Tipos de Estratégia e Exemplos (Andras and Havran, 2015)

Tipo	**Transferência (SD + SF)**	**Comercial (SF > SD)**	**Círculo de Sucesso (SD + SF)**	**Sinergia (SD > SF)**	***L'art pour l'art* (só SD)**
Exemplos antigos	Grasshoppers	Clubes ingleses	Liverpool	Milan, Chelsea	Real Madrid, Barcelona
Exemplos recentes	Clubes belgas e portugueses	Clubes alemães	Basel, Rosenborg	Man City, PSG	Real Madrid, Barcelona

- Estratégia de Transferências, em que as equipas vendem os seus ativos com margens de lucros suficientes para continuarem a ser competitivas e operacionais
- Estratégia Comercial, em que o foco principal é tornar o clube uma empresa lucrativa, independentemente do sucesso dentro de campo,
- Estratégia do Círculo de Sucesso, em que o clube procura ter lucro com base em resultados desportivos positivos,
- Estratégia de Sinergia, em que o objetivo é ter sucesso desportivo a todo o custo, investindo imenso dinheiro, normalmente com recurso a um mecenas
- Estratégia *L'art pour l'art*, em que o clube tem uma gestão *Freestyle* em que a única preocupação é ganhar títulos, devido à facilidade tremenda em obter dinheiro

Outra área importante na evolução de um clube é o marketing da equipa, ou seja, a imagem do clube como marca global e a forma de alcançar adeptos que normalmente não seriam o alvo principal, principalmente devido à globalização económica do futebol (Andreff, 2008). O exemplo do Paris Saint-Germain (PSG) vai ao encontro dessa necessidade em inovar na comunicação com o adepto e na venda da sua imagem. Adquirido pela Qatar Sports Investiments (QSI) em 2011, com Nasser Al Khelaïfi a presidir o clube desde aí, o objetivo seria transformar a modesta equipa francesa com apenas 18 títulos, identificada como uma com enorme potencial devido à sua localização na maior área metropolitana europeia (Légetr-Moec, 2011), numa marca global como os New York Yankees, equipa de basebol, ou os Chicago Bulls, equipa de basquetebol (Mantoux, 2020). Juntamente com o investimento forte na contratação de jogadores como Neymar e Kylian Mbappé, os novos donos do clube decidiram capitalizar na cidade do clube, Paris, e modernizar a imagem do clube, começando uma remodelação do logótipo de forma a ser estilizado em torno da silhueta da Torre Eiffel e com um foco maior no nome da cidade, tal como notado na Figura 6. A justificação, segundo Fabien Allegre, o diretor de merchandising e diversificação da marca PSG, foi de que "Moda, estilo, design e elegância são os símbolos de Paris. Nós capitalizamos no ADN de Paris, as suas

raízes, a sua identidade e a sua originalidade. É isso que nos diferencia das outras cidades, o que nos torna diferentes, únicos; é isso que nos dá a capacidade de ser uma marca de estilo de vida, com um forte conteúdo criativo" (Mantoux, 2020). O resultado desta aposta é notório: 700 milhões de dólares de receita em 2019, quinto maior clube no ranking financeiro europeu, nomeado pela Forbes como a segunda maior marca desportiva do mundo a nível de crescimento e um retorno de investimento de 4185%, isto é, o clube foi adquirido por 70 milhões de dólares em 2011 e vale neste momento mais de 3 mil milhões de dólares (Mantoux, 2020).

Figura 6 – Antigo e novo logótipo (Footy Design, 2013)

2.2 Análise Estatística no Futebol

A vertente de análise estatística no desporto tem crescido ao longo do tempo, acompanhando a evolução que ocorre na tecnologia e nos algoritmos desenvolvidos para esses efeitos (secção 1.1), tendo sido os desportos denominados "americanos" (Basquetebol, Futebol Americano, Basebol e Hóquei no Gelo) a impulsionar a análise estatística aprofundada e a utilização de ferramentas de *machine learning* para identificar alvos no mercado e definir a estratégia a seguir na próxima partida (Stewart et al., 2007). No entanto, no futebol, existe uma negligência histórica por parte das entidades futebolísticas em armazenamento e obtenção de dados estatísticos que não sejam os mais elementares como golos marcados e golos sofridos, o que resulta num atraso do desporto-rei no continente europeu em comparação com os desportos norte-americanos, apesar de isso ter vindo a ser corrigido nos últimos anos (Luzum and Model, 2021).

As origens da análise estatística no futebol têm uma data: 18 de março de 1950. Charles Reep encontrava-se a ver um jogo caseiro do Swindon Town, uma equipa do terceiro escalão inglês, frente ao Bristol Rovers, e ficou frustrado com a quantidade de chances desperdiçadas pela equipa de Swindon durante a primeira parte. Na segunda parte, Reep começou a anotar estatísticas da partida e notou que o Swindon teve 147 ataques que resultaram num golo. Extrapolando esse valor, e assumindo 280 ataques por jogo e com uma média de dois golos marcados, apercebeu-se que havia uma taxa de acerto de 0.71%, o que significava que bastava uma pequena melhoria para que uma equipa obtivesse em média três golos por jogo, o que resultaria quase de certeza na promoção (Wilson, 2018). A partir daí, Reep começou a efetuar análises mais sofisticadas, publicadas no News Chronicle, que identificavam que a maior parte dos golos eram marcados em jogadas com uma sequência de menos de três

passes, pelo que propôs que o importante seria colocar a bola na frente o mais rápido possível e com o menor número de passes, cunhando o termo *long ball*, e isso chamou à atenção do Brentford, que o contratou como analista, e posteriormente do Wolverhampton Wanderers (Wilson, 2018). O seu trabalho continuou na *Royal Statistical Society*, juntamente com Bernard Benjamin. O duo publicou uma análise estatística de padrões de jogo no futebol entre 1953 e 1967, usando um conjunto de dados correspondentes a jogos da primeira divisão inglesa, de jogos específicos do Sheffield Wednesday e do Arsenal e de 11 jogos do Mundial de 1966. Reep e Benjamin descobriram que apenas 5% de todos os movimentos consistiam em quatro passes ou mais, e apenas 1% consistia em seis passes ou mais, concluindo que futebol de posse não era o mais desejado (Bray, 2008). Estas análises fizeram com que o futebol inglês adotasse o estilo de jogo *long ball*, ou *kick and rush*, focando-se num jogo mais direto e físico, de forma a ir ao encontro ao que se julgava ser mais efetivo (Sleight, 2010). No entanto, este estilo é fortemente criticado, tanto por intervenientes diretos do jogo, como Franz Beckenbauer que considera que o estilo "não se assemelha em nada a futebol" (Fifield, 2010), ou por Jonathan Wilson, autor do *Inverting The Pyramid*, um livro sobre a história das táticas de futebol, que critica o uso de estatísticas a favor de um determinado ponto de vista, sem ter em conta o contexto da competição em que elas foram retiradas, chegando a afirmar que "é horripilante que uma filosofia baseada numa má interpretação de dados tão elementar pudesse ter sido permitida a se tornar uma das bases do treino inglês" (Wilson, 2018).

Na década de 70 e 80, Valeriy Lobanovskyi, além de ter revolucionado o processo de treino, aplicando processos científicos (secção 2.1), impulsionou igualmente a análise estatística e o registo de dados que um jogo de futebol intrinsecamente possui (Wilson, 2011). O treinador soviético e a sua equipa técnica criaram uma metodologia de registo de eventos de jogo e posterior enquadramento aos seus jogadores, que permitia que cada jogador soubesse o objetivo da equipa para alcançar a vitória por via de ações por jogo, conforme se verifica na Tabela 2. No dia a seguir aos jogos, era afixado no balneário um quadro com os principais pontos estatísticos do encontro disputado, com vista a dar um resumo geral do que os seus jogadores fizeram. Lobanosvskyi justificara estas decisões com o caminho para se obter um grande poder de argumentação perante os jogadores, afirmando que "Quando eu era um jogador, era difícil avaliar os jogadores. O treinador podia dizer que o jogador não estava no sítio certo à hora certa e o jogador podia simplesmente discordar. Não havia vídeos, nenhum método real de análise, mas agora os jogadores não podem discordar. Eles sabem que na manhã a seguir ao jogo, a folha de papel será colocada, mostrando todos os dados característicos do seu jogo."(Wilson, 2018)

Tabela 2 – Tabela com ações a ter num jogo (Wilson, 2018)

Tipo de Ação	Objetivo da ação		
	Apertar (pressão no meio-campo adversário)	**Contra-atacar (pressão no próprio meio-campo)**	**Combinação de ambos os modelos**
Passe Curto:			
Para a frente	130	80	30-130
Para o lado	100	60	40-100
Para trás	70	40	20-70
Passe Médio:			
Para a frente	60	80	40-90
Para o lado	50	25	30-80
Para trás	25	15	10-30
Passe Longo:			
Para a frente	30	50	15-40
Para o lado	20	30	10-30
Para trás	0	0	0
Cabeceamentos	20-40	20-40	15-70
Corridas com a bola	140	80	70-150
Ultrapassar adv.	70	50	20-70
Interceções	80	110	70-140
Desarmes	50	70	30-80
Remates à baliza	10-20	15-35	10-35
Cabeceamentos à baliza	10-15	5-10	5-15
Reposições de bola	10-30	10-30	10-40
Percentagem de erros	20-35	15-30	25

Mais recentemente, existe o exemplo do FC Midtjylland, um clube dinamarquês que até 2015 conquistara apenas um título da segunda divisão da Dinamarca. Com um investimento de 6,2 milhões de libras por parte de Matthew Benham, empresário inglês de 53 anos dono da empresa Smartodds, que usa modelos matemáticos para prever resultados de futebol, a equipa dinamarquesa passou a ter um foco enorme na parte estatística do desporto. A equipa técnica passou a ter um especialista de remate, que duas vezes por mês analisa como cada jogador acerta na bola e prepara programas de treinos para os jogadores trabalharem no seu tempo livre, estatísticas de jogo para as palestras ao intervalo, como as chances, as meias-chances e os golos esperados de cada equipa, e um assistente com funções na análise de bolas paradas, em que escrutina estatísticas e excertos de vídeo de forma a criar rotinas para a equipa usar nas partidas. Além disso, o FC Midtjylland também utiliza modelos matemáticos para encontrar novos jogadores, disponibilizados por uma equipa sediada em Londres, ao invés de ter um único olheiro a utilizar uma simples base de dados que contenha todos os jogadores no planeta, sem qualquer valor acrescentado. Com estas inovações, o clube conquistou três campeonatos dinamarqueses e uma taça nacional desde a época 2014/15, conseguindo também qualificar-se pela primeira vez para a UEFA Champions League (Ingle, 2015).

2.3 Machine Learning

Machine Learning consiste na aplicação de abordagens matemáticas e estatísticas para que as máquinas aprendam a partir de dados recebidos, com vista a que efetuem previsões ou decisões sem serem programadas para tal, de forma a reproduzir a inteligência humana (El Naqa and Murphy, 2015). Existem quatro tipo de técnicas de *machine learning,* ou aprendizagem, cada uma com vários tipos de aplicações (Hajaj, 2020):

- Aprendizagem supervisionada — é a tarefa de fornecer dados históricos corretamente categorizados a um modelo, de forma a que esse modelo aprenda o que determina cada categoria e consiga atribuir a novos dados essas categorias (Russell and Norvig, 2009). Dependendo da situação, se é pretendido uma classe ou um valor numérico, utilizam-se dois tipos de aprendizagem, classificação ou regressão linear.
    - Classificação — É o processo de categorizar um conjunto de dados em classes ou grupos que podem ser ou não predefinidos, por exemplo, um professor a classificar as notas dos alunos numa escala de A a F, em que são usadas técnicas matemáticas como árvores de decisão, programação linear e análises estatísticas (Diwate, 2014)
    - Regressão — É uma técnica estatística em que estima a relação entre as variáveis de um objeto num valor numérico. Este conceito é explorado mais em detalhe na secção 2.3.2, visto que um dos objetivos é obter um valor de rating de um jogador com base noutras estatística, conforme referido na secção 1.3.
- Aprendizagem não supervisionada — é o processo da máquina aprender "sem um professor", ou seja, de captar padrões ou de extrair características que identifiquem dados de forma mais compacta em dados não classificados (Hinton, 1999). A abordagem mais utilizada é *clustering,* o processo de descobrir grupos de objetos em que os elementos num grupo serão similares ou relacionados entre si e diferentes ou não relacionados com objetos de outros grupos (Aparna and Nair, 2014).
- Aprendizagem semi-supervisionada — é uma abordagem em que se mistura conceitos da aprendizagem supervisionada e da aprendizagem não supervisionada, em que para um determinado *input* não existe uma categoria exata, mas sim um conjunto de potenciais categorias, por exemplo, na análise de uma imagem de um lince, a máquina pode não identificar como a categoria "lince", mas sim ao conjunto de categorias "felino" em que a categoria "lince" pertence (Cabannes et al., 2021).
- Aprendizagem reforçada — é um paradigma de aprendizagem em que agentes inteligentes aprendem a partir de estímulos, quer sejam eles recompensas ou castigos. Por exemplo, um modelo que esteja a aprender como jogar um jogo recebe uma recompensa se ganha mais pontos e recebe um castigo se não ganhar tantos pontos quanto o desejado (Hu et al., 2020).

2.3.1 Análise de Regressão

Análise de regressão é o método de investigação de relações entre variáveis na forma de uma equação ou de um modelo que liga uma variável de resposta com uma ou mais variáveis preditoras (Chatterjee and Hadi, 2015). As aplicações de regressão no mundo real vão desde estimar a probabilidade de um paciente sobreviver usando os resultados de um conjunto de diagnósticos (Fayyad et al., 1996), ou prever o preço de imóveis, a variável de resposta, com base nas características físicas do edifício e os impostos pagos, as variáveis preditoras (Chatterjee and Hadi, 2015). A Figura 7 demonstra um exemplo de regressão linear, em que se compara o vencimento (*income*) com o débito (*debt*), mostrando uma correlação fraca entre as duas variáveis, já que não há uma divisão visível efetuada pela linha desenhada entre os pontos circulares, que correspondem à classe *emprestar*, e os pontos cruzados, que correspondem à classe não *emprestar* (Fayyad et al., 1996).

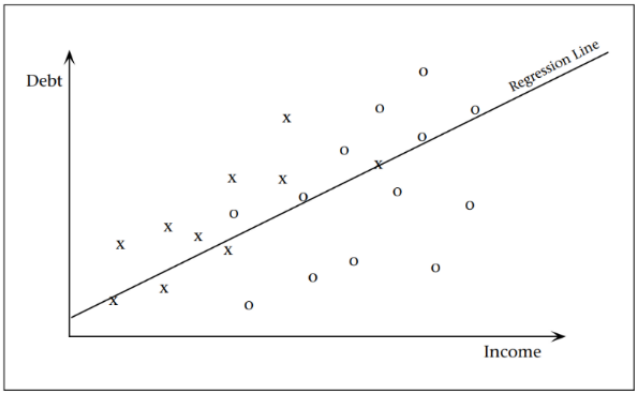


Figura 7 – Regressão linear entre débito e vencimento (Fayyad et al., 1996)

Abordagens que usem regressão, por norma, são técnicas de aprendizagem supervisionada, visto que é a procura de relações entre variáveis na definição de um tipo de valor contínuo já identificado pelo modelo de aprendizagem (Kaneko, 2018). No entanto, em vários conjuntos de dados usados no mundo real, existe uma quantidade vasta de dados por catalogar, o que dificulta a tarefa na aplicação de métodos tradicionais de aprendizagem supervisionada (Dyer et al., 2014). Desse modo, e considerando que o problema de regressão se define como um conjunto dado de l exemplos etiquetados $\mathrm{L_d} = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ em que $x_i \in \mathrm{R}^n, y_i \in \mathrm{R}, i = 1,2, \dots, l+u$, pretende-se prever o valor de y para cada novo exemplo de x, estende-se o problema para uma visão supervisionada, em que o pretendido é o mesmo, isto é, prever y para cada x, e dado um conjunto de l exemplos etiquetados $\mathrm{L_d} = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ e com a adição de um conjunto de exemplos não etiquetados $\mathrm{U_d} = \{x_{l+1}, x_{l+2}, \dots, x_{l+u}\}$ em que $x_i \in \mathrm{R}^n$ e $y_i \in \mathrm{R}$, mas com $i = 1,2, \dots, l+u$, ou seja, o conjunto não catalogado será também usado no algoritmo de aprendizagem (Kostopoulos et al., 2018). Uma das aplicações deste tipo de metodologia foi efetuado por Kostopoulos e a sua equipa em 2019, em que, baseados em dados de alunos de um curso de longa-distância, obtiveram previsões e interpretações sobre as notas dos alunos, em que a aplicação se

prendeu na interpretação de dados obtidos da performance dos variados alunos com vista a obter um valor de nota final de cada aluno (Kostopoulos et al., 2019).

Com a possibilidade de se usar regressão em vários tipos de aprendizagem, a sua popularidade no mundo de *machine learning* é grande, com a KDNuggets, um dos principais blogs sobre *data science*, lançou um artigo sobre os principais algoritmos de *machine learning* para o ano de 2021, com destaque para vários que são úteis para resolver problemas de regressão (Shin, 2021), entre os quais:

- Regressão Linear — é usada para modelar a relação entre uma variável dependente y para uma ou mais variáveis independentes x numa forma linear (Schneider et al., 2010).
- Regressão Logística — funciona de forma similar à regressão linear mas devolve uma resposta binomial, isto é, usa-se para calcular a probabilidade de um número discreto de resultados, por norma dois (Sperandei, 2014).
- K-Nearest Neighbour — explicada com melhor detalhe na secção 2.3.1.1
- Gradient Boost —é uma técnica de *boosting*, isto é, com recurso a algoritmos que transformam agentes de aprendizagem fracos em agentes fortes (Zhou, 2012), usada tanto para problemas de regressão como de classificação em que se parte do pressuposto que é mais fácil minimizar uma função contínua suave do que uma função discreta, o que resulta na minimização da função de perda com a estimativa do impacto de pequenas variações nos parâmetros (Lecun et al., 1998).
- Modelos Lineares Generalizados – explicada com melhor detalhe na secção 2.3.1.2

2.3.1.1 Algoritmo K-nearest neighbors

Em 1951, Fix e Hodges escreveram um relatório no qual apresentavam um método não paramétrico para classificar pontos com base nos k pontos vizinhos mais próximos que ficou conhecido como a regra do *k-nearest neighbour* (Fix et al., 1951), tendo sido depois explorado por outros investigadores de forma a perceber as propriedades formais do algoritmo, como a descoberta de que para $k = 1$ e $n \to \infty$, o classificador dobra na taxa de erro de Bayes (Cover and Hart, 1967) e o seu posterior refinamento (Fukunaga and Hostetler, 1975), que colocou o classificador a obter metade da taxa de erro (Duda et al., 2000). Além disso, afinações usando abordagens de distância ponderada (Bicego and Loog, 2016) ou métodos *fuzzy* (Keller et al., 1985) tornaram o algoritmo um dos mais fundamentalmente sólidos e simples de aplicar (Pulabaigari and T, 2011). A Figura 8 tem um exemplo de como funcionaria na prática o algoritmo, em que o k — número de vizinhos — é igual a um. O ponto a cinzento, não identificado como azul ou vermelho, é classificado como de cor vermelha, devido ao vizinho mais próximo ser da cor vermelha.

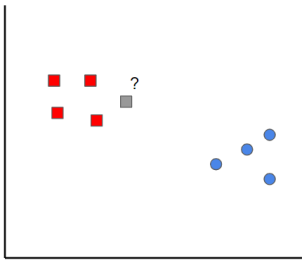


Figura 8 – Exemplo de k nearest (Shin, 2021)

Algumas aplicações do algoritmo consistem na previsão de eventos económicos com base em dados históricos de bancarrotas de bancos e de rácios financeiros, com recurso tanto a técnicas de classificação como de regressão (Imandoust and Bolandraftar, 2013) ou na classificação de doenças de coração, juntamente com algoritmos genéticos, de forma a melhorar a precisão de futuras previsões médicas (jabbar et al., 2013).

O algoritmo, sendo de aprendizagem supervisionada (Kramer, 2013), tem como objetivo descobrir a classe de um ponto, e a abordagem utilizada pode ser de classificação ou de regressão, em que no primeiro caso o valor retornado resultará do voto dos k vizinhos mais próximos do ponto a determinar, enquanto que no segundo caso o retorno será o valor da propriedade do objeto, ou seja, a média dos k vizinhos (Hastie and Tibshirani, 1995).

2.3.1.2 Modelo Linear Generalizado

John Nelder e Robert Wedderburn propuseram em 1972 um método de reponderar os mínimos quadrados de forma iterativa para obter uma estimativa da maior probabilidade dos parâmetros de um modelo. A essa proposta foi dado o nome de Modelo Linear Generalizado, que consistia na unificação de vários modelos estatísticos, como regressão linear ou regressão de Poisson (Nelder and Wedderburn, 1972).

O Modelo Linear Generalizado (GLM para *gradient linear models*) estende o modelo de forma que a variável objetivo possa ser qualquer membro da função exponencial. O modelo consiste em três componentes (Agresti, 2012):

1. A variável objetivo Y_i resultante da distribuição à qual faz parte (Normal, Exponencial, Poisson, etc.). A função de densidade de Y_i pode ser descrita como

$$f(y;\theta) = e^{yb(\theta)+c(\theta)+d(y)}$$

Equação 1 – Função de densidade da variável objetivo (Agresti, 2012)

em que b, c e d são funções conhecidas.

2. O indicador linear η_i que é uma função das variáveis previsoras $x_{i1}, \dots, x_{ip}$ com base nos parâmetros estimativos $\beta_i, \dots, \beta_p$. Com isso, resulta na Equação

$$\eta_i = \boldsymbol{x_i^T \beta}$$

Equação 2 – Indicador linear (Agresti, 2012)

3. A função de ligação monótona b que une o valor esperado da variável objetivo Y_i ao indicador linear η_i , originando na Equação

$$g(E[Y_i]) = \eta_i$$

Equação 3 – Função de ligação monótona (Agresti, 2012)

Devido ao tema da dissertação ter um grande foco na análise de estatísticas de forma a obter-se uma relação entre elas, identificou-se a aplicação de algoritmos que utilizem Modelos Lineares Generalizados como a que melhor permitirá obter o valor máximo de parâmetros do modelo de previsão de resultados, ajudando depois na obtenção dos pesos das estatísticas do modelo para calcular o valor de *rating* dos jogadores.

2.3.1.3 Média Móvel Exponencial com Suavização

A média móvel exponencial com suavização, traduzido do inglês *exponential weighted moving average* (EWMA) ou *exponential weighted smoothing average* (EWSA), é uma estatística cuja principal característica que a diferencia de uma média móvel normal é a menor importância dada ao longo do período de tempo analisado aos valores antigos de médias, ou seja, uma média móvel normal em que considera apenas cinco valores para o seu cálculo, no caso de ser registado um novo valor, descarta o valor mais antigo, enquanto a média móvel exponencial com um factor de suavização contabiliza todos os valores presentes no conjunto de dados, mas com maior foco nos valores mais recentes (Hunter, 1986).

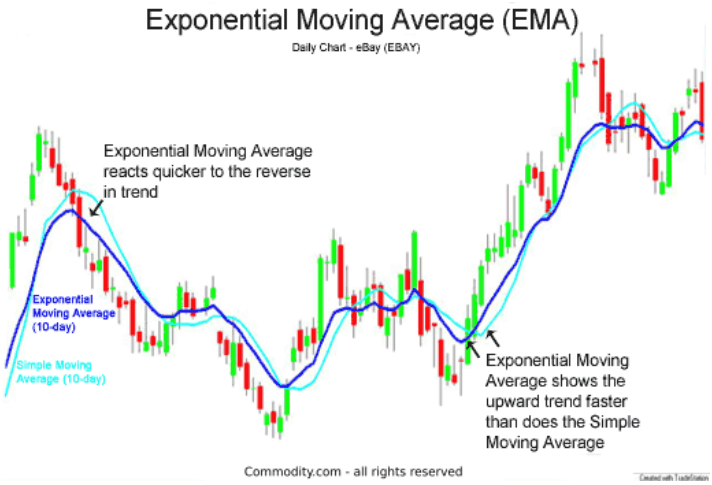


Figura 9 – Média móvel e móvel exponencial do valor de ações do eBay (Pines, 2020)

A Figura 9 compara a média móvel, a azul-claro, e a média móvel exponencial, a azul-escuro, e verifica-se que a média móvel exponencial reage mais rapidamente a tendências de subida e descida nos valores, enquanto a média móvel demora mais tempo a representar essas mudanças. No entanto, a média móvel exponencial acaba por ser mais propensa a falsos sinais (Pines, 2020).

Primeiramente exposta por Roberts em 1959 num artigo que comparava com outras de médias moveis simples (Roberts, 1959), a técnica tem como principal foco para retirar o ruído e suavizar os dados em análise numa sequência temporal de dados para melhor perceber

padrões e ciclos ao longo do tempo. A EWMA é aplicada regularmente em previsões de vendas e valores de ações na banca e no controlo de qualidade de processos (Perry, 2010), bem como na previsão de números de infeções da pandemia de COVID-19 (Oshinubi et al., 2021).

A fórmula que descreve o processo está demonstrado na Equação 4, em que quando a sequência de valores inicia em $t = 0$:

$$s_0 = x_0$$
$$s_t = \beta x_t + (1 - \beta) s_{t-1}, \qquad t > 0$$

Equação 4 – Fórmula de EWMA (Guthrie, 2003)

No qual β é o valor de suavização e está contido no intervalo $\beta \in]0,1[$.

Devido ao facto de que a média móvel exponencial permite aplicar um valor de suavização de forma a contabilizar todos os valores, mas dando mais importância aos mais recentes, esta técnica revela potencial para resolver alguns dos objetivos identificados neste problema. Tendo em conta que os jogadores de futebol podem ter picos de forma, quer sejam positivos ou negativos, a aplicação da EWMA permitirá identificar jogadores que estejam a jogar bem ou mal de forma sustentada, e evitar encontrar situações em que um baixo número de boas ou más exibições influenciem de forma incorreta a análise da sua qualidade.

2.3.2 Máquinas de Vetores Suporte

Uma máquina de vetores de suporte (SVM, do inglês *support vector machine*) é um algoritmo de *machine learning* que analisa dados para problemas de classificação e regressão. O algoritmo divide inicialmente o conjunto de dados em duas categorias e cria o modelo com base nessa divisão, tendo como objetivo determinar a que categoria uma nova entrada no conjunto pertence. Esta metodologia torna-o um classificador linear binário não probabilístico, visto que não atribui probabilidades de a nova entrada pertencer a uma categoria, mas sim catalogá-la (Suthaharan, 2015).

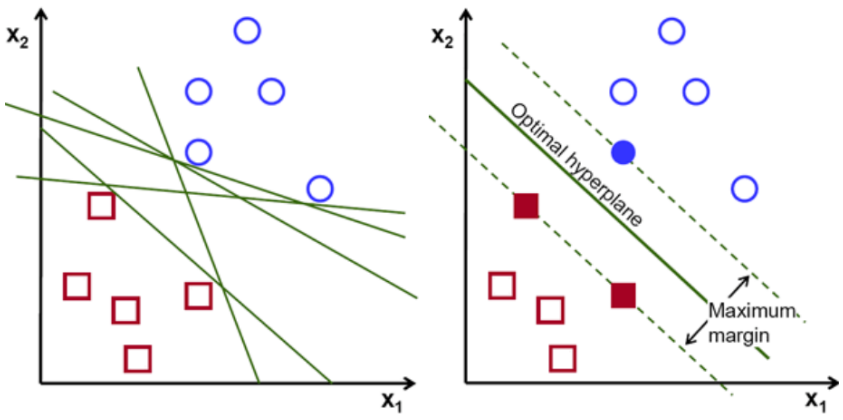


Figura 10 – Demonstração de identificação de hiperplanos (Ippolito, 2021)

As SVMs são usadas para obter o hiperplano de separação ideal com a ajuda de vetores de suporte, de modo a representar uma separação clara das diferentes classes no conjunto de dados. Os dados irão ser classificados com base na posição em que se encontram em relação ao hiperplano. Por exemplo, na Figura 10 pode-se observar essa divisão na classificação, em que os dados acima do hiperplano têm a classe azul circular, enquanto os abaixo têm a classe vermelha quadrada. Os pontos mais próximos do hiperplano são os Vetores de Suporte, que determinam a sua orientação e posição. de forma a maximizar a margem do classificador, estando identificados com cor preenchida (Ippolito, 2021).

Os algoritmos que aplicam SVMs, devido à sua característica de serem não probabilísticos, acabam por perder alguma subjetividade na identificação de resultados de futebol, e por consequência na atribuição de pesos das variáveis, em comparação com outros algoritmos. No entanto, e devido à sua boa capacidade de previsão, irão ser explorados no desenvolvimento deste caso de estudo.

2.4 Sistemas de Apoio à Decisão

A definição do que é um Sistema de Apoio à Decisão (*SAD*) tem vindo a evoluir ao longo dos últimos anos. Keen identificou num estudo de 1980 que um SAD teria de ser um sistema computacional integrado que adotasse uma das seguintes conceções (Keen, 1980):

- Que fosse definido com base na estrutura da tarefa que aborda
- Que requeresse uma estratégia de design distintiva baseado na evolução
- Que apoiasse os processos cognitivos dos tomadores de decisão
- Que refletisse uma estratégia que tornasse os computadores úteis para gerentes

Espinasse e Pascot, em 1987, definiram que os SAD deveriam assistir gestores em problemas semiestruturados, combinar o uso de modelos ou técnicas analíticas com acesso a bases-de-dados transicionais, ter um uso fácil e possuir a capacidade de perceber as particularidades do ambiente onde o sistema será usado e as características cognitivas do decisor (Espinasse and Pascot, 1987). Mais recentemente, Daniel Power, em 2002, identificou três características principais de um SAD: desenhados especificamente para facilitar processos de decisão, apoiar uma decisão e não efetuá-la automaticamente, e adaptar-se rapidamente às constantes alterações nas necessidades dos decisores (Power, 2002). Além disso, definiu uma taxonomia dividida em cinco tipos de Sistemas de Apoio à Decisão (Power, 2002). A tabela 4 terá um resumo da taxonomia de Power:

- Direcionado a dados — Foco na obtenção e análise de grandes quantidades de dados estruturados, como *data warehouses* ou sistemas de ficheiros. Direcionado a gestores ou fornecedores com recurso a uma página web ou a um servidor.
- Direcionado ao modelo — Foco no contexto do problema e em representar dados e parâmetros fornecidos pelos decisores de forma a analisar melhor as situações a

resolver. Direcionado a gestores ou clientes com recurso a um computador específico para o sistema.

- Direcionado ao conhecimento — Foco na sugestão de ações recomendadas com base em regras de negócio e bases de conhecimento, sendo *data mining* um dos termos relacionados. Direcionado a utilizadores conhecedores do negócio ou a clientes com recurso a páginas Web ou a um servidor.
- Direcionado a documentos — Foco na obtenção, classificação e gestão de documentos sem qualquer estruturação, como páginas web. O grupo de uso é mais vasto atualmente, encontrando-se na altura em expansão, e o uso é principalmente em páginas Web.
- Direcionado a comunicações — Foco na facilitação do processo de tomada de decisão em grupo. Direcionado a equipas, com recurso a páginas web ou a um sistema de chat.

Tabela 3 – Taxonomia de Power (Power, 2002)

Componente Dominante	**Grupos de Utilizadores**	**Propósito da Componente**	**Tecnologia usada**
Direcionado a dados	Gestores ou fornecedores	Obtenção e análise de dados	Web ou Servidor
Direcionado ao modelo	Gestores ou clientes	Análise de decisão	Computador preparado para o sistema
Direcionado ao conhecimento	Utilizadores internos ou clientes	Sugestões de Gestão e Escolha de Itens	Web ou Servidor
Direcionado a documentos	Grupo de utilização vasto	Pesquisa de páginas web e de documentos	Web
Direcionado a comunicações	Equipas internas	Conduzir uma reunião e auxiliar tomadas de decisão em grupo	Web ou Chat

Os objetivos propostos na secção 1.3 enquadram-se numa mistura de duas componentes de sistemas de apoio à decisão: a componente direcionada a dados, visto que um dos objetivos é a análise de dados estatísticos completos relacionados com jogadores, e a componente direcionada ao conhecimento, já que é pretendido obter valor baseado em conhecimento resultante dos modelos desenvolvidos.

Um sistema de suporte à decisão tem tradicionalmente quatro componentes, representados na Figura 11: uma interface para o utilizador interagir com o sistema, uma base de dados, um conjunto de ferramentas analíticas e modelos, e, por fim, a forma de comunicação entre todos os elementos, ou seja, a arquitetura do sistema (Sprague and Carlson, 1982).

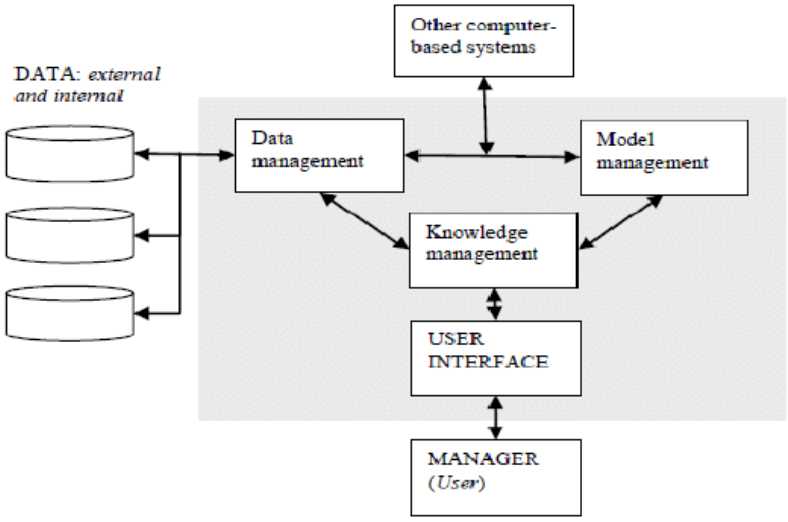


Figura 11 – Componentes de um SAD (Sprague and Carlson, 1982)

Para o desenvolvimento desta solução, e tendo em conta os componentes identificados, considerou-se o seguinte:

- *User Interface* — Aplicação Web com demonstração dos dados obtidos e gerados
- Gestor de Conhecimento — Motor de aprendizagem com base nos dados obtidos
- Gestor de Dados — Base de dados com todos os dados gerados pelo modelo de aprendizagem e pelas APIs externas.
- Gestor de Modelo — Conhecimento obtido a partir da investigação e do modelo criado

2.5 Trabalho Relacionado

Nesta secção, é realizado um estudo sobre os artigos científicos escritos sobre o tema desta dissertação, bem como sobre as soluções existentes no mercado para resolver o problema apresentado.

2.5.1 Artigos Científicos

2.5.1.1 *PlayeRank*: Data-driven Performance Evaluation and Player Ranking in Soccer

Este artigo científico tem como propósito criar uma aplicação que avalie a performance de jogadores com base num modelo calculado com dados estatísticos profundos, que vão desde passes e desarmes, até a uma lista de todas as ações e locais que um jogador efetuou e pisou em cada partida de 18 campeonatos ao longo de quatro épocas (Pappalardo et al., 2019).

A forma como a aplicação determina o valor de cada jogador, começando primeiro por identificar as estatísticas mais valiosas e influentes num resultado de uma partida por forma a

fundamentar o cálculo, revela ser uma metodologia que vai ao encontro dos padrões CRISP-DM referidos anteriormente, nomeadamente o passo *Data Understanding*.

A *framework* deste projeto está dividida em três etapas. Na primeira etapa, o modelo desenvolvido estuda todos os jogos obtidos durante o período indicado para retirar as estatísticas e eventos mais importantes na obtenção de uma vitória. Para isso, a cada registo de jogo é atribuído 1 à equipa que vence, e 0 à equipa que perde. No caso de empate, ambas as equipas ficam a 0. Depois, o peso de cada estatística para a obtenção é extraído, resolvendo um problema de classificação entre a performance estatística, composta pelos dados estatísticos de cada jogador, e o resultado da partida, de forma a ser usado no cálculo de performance, sendo usado o *Linear Support Vector Classifier* (LSVC). De seguida, na segunda etapa, é atribuído a cada jogador um papel, como defesa central ou avançado, com base nas coordenadas médias da sua posição em campo. Por fim, na terceira e última etapa, é efetuada a avaliação do jogador, usando os dados dos jogos em que o atleta participou, em que dois cálculos são efetuados:

$$r(u,m) = \frac{1}{R}\sum_{i=1}^{n} w_i \times x_i$$

Equação 5 – Cálculo do *rating* do jogador numa partida (Pappalardo et al., 2019)

Em que $r(u,m)$ corresponde ao *rating* do jogador u no jogo m e R é a constante de normalização para que valor do *rating* fique correspondido entre 0 e 1. w_i é o valor do peso calculado de cada estatística x_i recolhida. Neste cálculo, são retirados os golos, já que é inerente o valor dos golos marcados num jogo de futebol, visto que o objetivo passa por marcar mais que o adversário. Mesmo assim, os investigadores efetuaram também um cálculo em que os golos são considerados, de forma a dar essa possibilidade ao utilizador final. Finalmente, o valor do jogador é obtido calculando o *rating* ao longo de uma série de jogos com base na técnica *Exponential Weighted Moving Average* (EWMA). Sendo assim, o valor final é computado da seguinte forma:

$$\bar{r}(u,M) = \bar{r}\left(u,m_g\right) = \beta \times r\left(u,m_g\right) + (1-\beta) \times \bar{r}\left(u,m_{g-1}\right)$$

Equação 6 – Cálculo da média ponderada do *rating* do jogador (Pappalardo et al., 2019)

Sendo β o fator de suavização dentro do intervalo $[0,1]$, o que resulta no cálculo ponderado do *rating* $r\left(u,m_g\right)$ do jogador u na sua última partida m_g e os valores previamente suavizados $\bar{r}\left(u,m_{g-1}\right)$, o que resulta num maior peso dado às exibições mais recentes.

Na aplicação, toda esta informação é representada graficamente com acesso ao histórico das suas performances e a comparações com jogadores semelhantes. Além disso, a forma como o utilizador consegue filtrar dados e comparar perfis faz com que a tomada de decisão seja facilitada, tornando a experiência mais agradável e enriquecedora, conforme se vê nas Figuras 12 e 13.

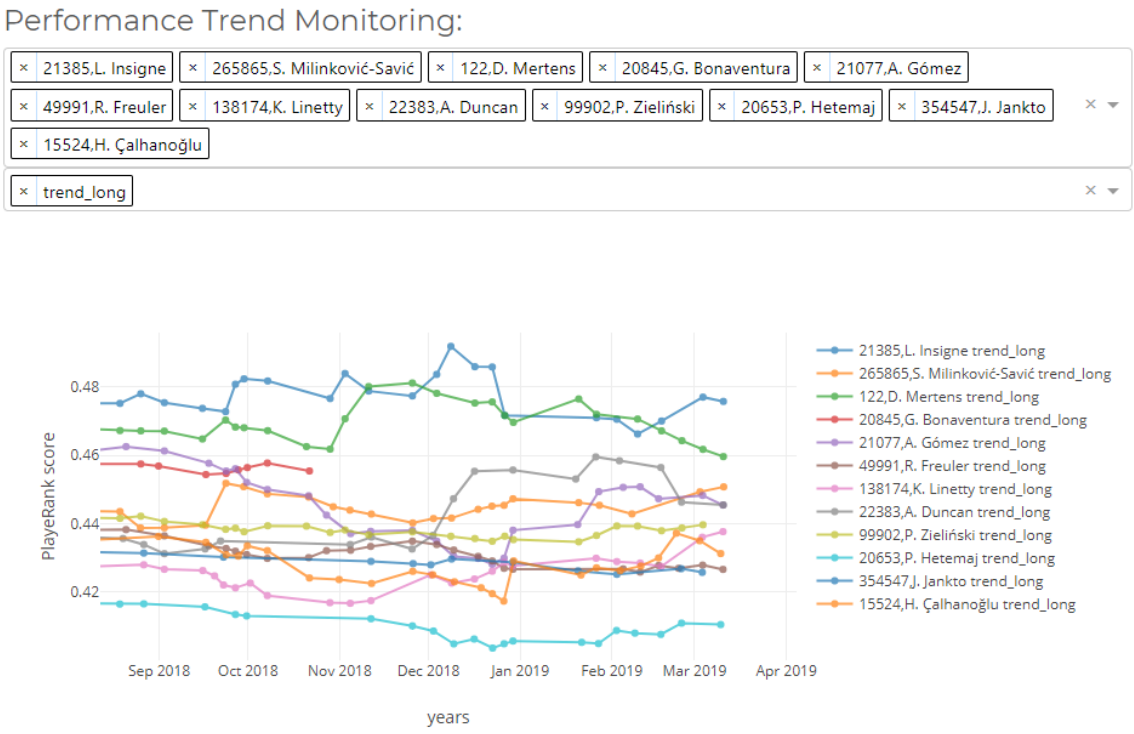


Figura 12 – Comparar a performance de jogadores escolhidos (Pappalardo et al., 2019)

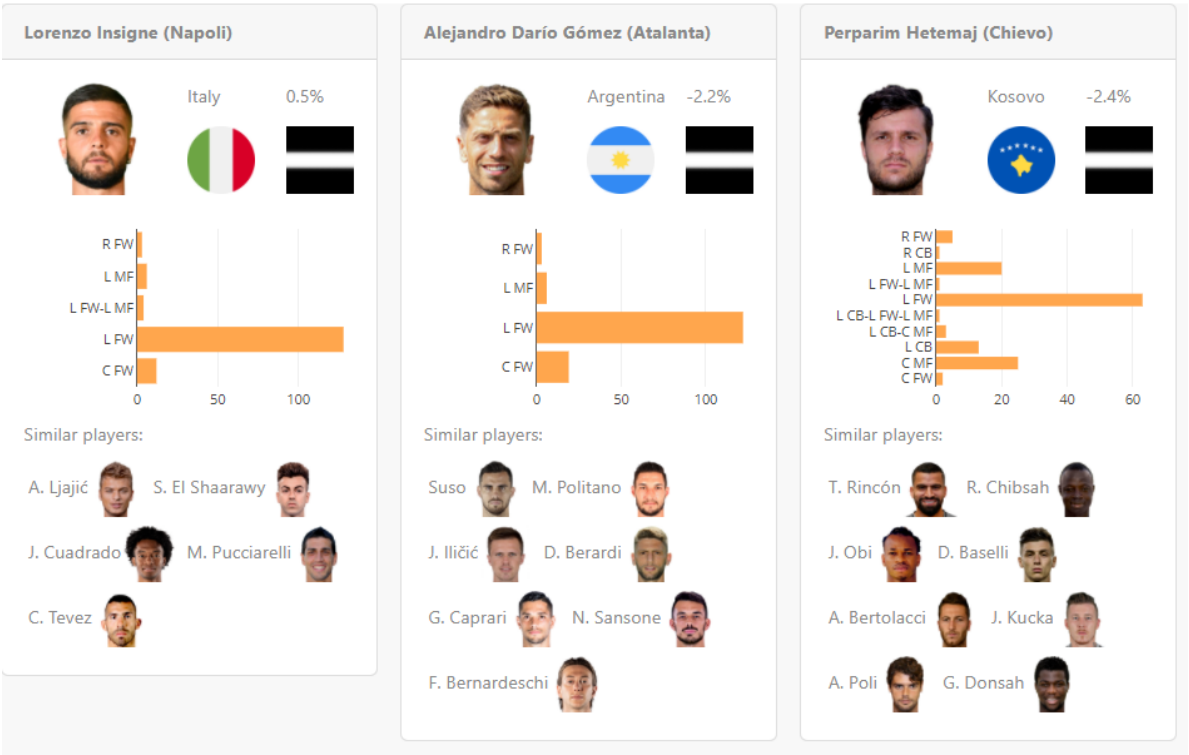


Figura 13 – Visualização de jogadores semelhantes e posições em que o jogador atuou (Pappalardo et al., 2019)

De forma a validar os resultados obtidos pela solução desenvolvida, os investigadores deste projeto efetuaram um inquérito dirigido a três olheiros de futebol, com a seguinte metodologia:

1. Organizar os jogadores por posição/*cluster* e pelo valor de qualidade obtido
2. Criar aleatoriamente um conjunto com 35% dos jogadores
3. Escolher um jogador de forma aleatória nesse novo subconjunto.
4. Para cada jogador selecionado no novo subconjunto, escolher um jogador x posições acima e x posições abaixo. Esse valor encontra-se num intervalo aleatório de [1,20] valores, ou seja, numa lista de 100 jogadores. Caso o valor de x obtido seja 10 e o primeiro jogador escolhido seja o 40º melhor jogador, o segundo jogador será o 30º melhor jogador e o terceiro jogador será o 50º melhor jogador.
5. Retirar os dois jogadores obtidos a partir do jogador original, ou seja, no exemplo dado no ponto 3, o 30º e o 50º melhor jogador, representado pela função $par(u_1, u_2)$
6. Repetir o passo 3 e 4 até obter uma lista de jogadores suficientes para criar um inquérito destinado a profissionais de futebol
7. No inquérito, solicitar a cada observador que respondesse para indicar qual o melhor jogador do par de jogadores da resposta, ou, no caso de os achar semelhantes, indicar que os acha de qualidade idêntica.
8. Tratar os dados obtidos depois de todas as respostas serem efetuadas, ou seja, retiravam os pares em que os três observadores identificavam como jogadores semelhantes e os pares em que dois observadores tinham opiniões contrárias, um achava o jogador u_1 melhor e outro achava o jogador u_2 melhor, enquanto o terceiro achava os dois semelhantes, retirando assim 8% dos pares de jogadores.
9. Calcular o valor de maioria de concordância c_{mai}, definida como a fração de pares em que a solução PlayerRank concorda com pelo menos dois dos observadores, e o valor de concordância unânime c_{una}, definida como a percentagem de pares em que os observadores concordam unanimemente com a solução.

Os resultados obtidos foram de 68% para o valor de c_{mai} e de 74% para o valor de c_{una}, demonstrando que, geralmente, a solução vai ao encontro da opinião de observadores especialistas na avaliação de jogadores.

No entanto, os resultados notam algumas limitações que precisam de ser tidas em conta:

- Todos os jogos são calculados de forma igual, quer seja um jogo da liga espanhola, a melhor liga segundo o ranking competitivo da UEFA, ou um jogo da liga grega, a décima-oitava melhor liga europeia segundo o mesmo ranking. O estudo indica que fez uma análise do peso de cada variável em todos os campeonatos e não foram notadas diferenças, porém, a exigência e o nível médio de cada competição são diferentes, bem como o contexto de um jogo, já que uma partida referente à 4ª jornada não tem o mesmo nível de pressão de uma partida da 31ª jornada, onde posições podem ser decididas, ou um jogo entre o 3º. e 4º, em que a competitividade acaba por ser maior do que um jogo entre o 2º e o 17º.

- O cálculo do valor de jogadores de posições mais defensivas não vai ao encontro do que seria esperado e tem *outliers* em relação às posições mais ofensivas. Por exemplo, num dos grupos identificados pelo projeto que corresponde à posição de avançado esquerdo, os dez melhores jogadores faziam parte de uma equipa colocada nos vinte-e-cinco melhores da UEFA, enquanto na posição de defesa centro direito apenas dois jogadores de dez faziam parte dos quadros de uma dessas equipas, e na posição de defesa centro esquerdo só um jogador do "top-25" de clubes europeus. Esta limitação pode ser devido ao facto de ser difícil ajuizar estatisticamente jogadores defensivos de equipas mais fortes, visto que normalmente irão ter menos ações defensivas que os defesas de equipas menos fortes.
- No cálculo do rating, os investigadores não consideram a estatística individual de golos para obter o valor do jogador, uma abordagem que permite descobrir outras estatísticas valiosas que não seja o número de golos. No entanto, consideram a estatística de assistência para golo, e isso acaba por ser incongruente com o indicado com a estatística de golos, já que uma assistência acaba por resultar sempre num golo, o que a torna inerentemente valiosa. Uma forma de resolver seria juntar as duas estatísticas numa só chamada "ação de golo", ou então efetuar a mesma separação de dados que se fez com a estatística de golos.

Este artigo científico serviu como inspiração para o tema desta dissertação, devido ao objetivo proposto ter sido alcançado, à aplicação de boas práticas quer de *data mining* quer de UI/UX.

2.5.1.2 Interactive exploratory soccer data analytics

A ferramenta desenvolvida por Delibas, Uzun, Fatih Inan, Guzey e Cakmak tem uma abordagem de "E se?", ou seja, é uma ferramenta gráfica cujo ponto forte é a possibilidade de criar situações hipotéticas de uma partida de futebol com base em dados reais. Permite rever situações que aconteceram numa partida e parar a qualquer momento para analisar sequências de passe, perceber o impacto da decisão do jogador, avaliar o risco tomado pelo jogador e detetar se a ação escolhida (passe, remate, drible) foi a mais correta, bem como criar situações alternativas, movendo os jogadores para outro sítio do campo ou explorando situações novas de passe ou remate, avaliando o seu grau de risco/sucesso ou a melhor posição de passe/remate, conforme se vê nas Figura 14 e 15 (Delibas et al., 2019).

Figura 14 – Repetição de jogo interativo, com teste de novas situações de passe (Delibas et al., 2019)

Figura 15 – Melhor posição para o jogador selecionado (#8) (Delibas et al., 2019)

Um dos pontos fortes desta investigação é o cálculo de um valor quantitativo da qualidade de um passe. A partir de um outro estudo (Cakmak et al., 2018) cujo objetivo era determinar uma equação que obtivesse o *Pass Efectivness Score*, calculado da seguinte forma:

$$\begin{aligned} \textit{Effectiveness}\ (\text{passe}(P_1, P_2))_{\text{ProxPass: passe }(P_2,P_3)} = {} & w_1 \times \text{Ganho}\ (\text{passe}(P_1, P_2)) + \\ & w_2 \times \text{Passe vantagem}\ (P_2) + \\ & w_3 \times \text{Chance Golo}\ (P_2) + \\ & w_4 \times \text{Tempo Decisão}\ (P_2) + \\ & w_5 \times \textit{Effectiveness}\ (\text{passe}(P_2, P_3)) \end{aligned}$$

Equação 7 – Cálculo de *Pass Efectivness Score* (Cakmak et al., 2018)

Dado um $\text{passe}(P_1, P_2)$ entre o jogador P_1 e o jogador P_2, pertencente a uma sequência de passes S, seguido de um passe (P_2, P_3) do jogador P_2 para o jogador P_3, obtém um cálculo de cinco métricas diferentes multiplicadas por pesos w_i's calculados anteriormente usando um algoritmo de otimização genérico. As cinco métricas dizem respeito ao seguinte:

- $\text{Ganho}\ (\text{passe}(P_1, P_2))$: Um passe efetivo diminui o número de jogadores entre a bola e a baliza adversária. No cálculo deste valor, subtrai-se o risco de perda de bola do jogador P_1 com o risco de perda de bola do jogador P_2 a partir da distância dos jogadores adversários ao raio de ação do passe.
- Passe vantagem (P_2): Quantifica a possibilidade que o jogador P_2 terá de continuar a jogada com qualidade, sendo um rácio entre o ganho e o risco do passe efetuado.
- $\text{Chance Golo}\ (P_2)$: Representa a probabilidade que o jogador P_2 em que os seguintes pontos são considerados:
    - Uma chance de golo aumenta assim que a distância à baliza $dist$ diminui. O valor de largura_baliza é 7,32 metros, usando as normas internacionais da FIFA.
    - Uma chance golo aumenta assim que ângulo α entre o jogador e os cantos da baliza aumenta, em que se assume que a marca de penalti é o local com maior chance de golo, logo, o ângulo do jogador na marca ângulo_penalti é usado como o melhor que um jogador consegue ter.
    - De forma a considerar os defesas que podem bloquear um remate, considera-se que um remate é um passe para o guarda-redes e calcula-se o risco de o "passe" ser intercetado, usando a mesma função utilizada nas funções Ganho e Passe Vantagem.
    - O cálculo está representado na Equação 8, em que se multiplica todos os valores obtidos de forma a obter o valor da chance de golo.
- $\text{Tempo Decisão}\ (P_2)$: Este valor traduz-se no tempo de duração que o jogador P_2 tem para decidir antes que o adversário mais próximo entre num duelo pela bola. É calculado usando a distância do jogador P_2 entre o adversário e a velocidade máxima que o adversário pode atingir. Caso essa velocidade não esteja nos dados, usa-se a velocidade máxima de 8.97 m/s, obtida a partir de um outro estudo (Rampinini et al., 2007).

- $w_5 \times$ *Effectiveness* $(\text{passe}(P_2, P_3))$: A função é recursiva, pelo que o score de *Effectiveness* do passe seguinte na sequência S será considerado no cálculo.

$$\text{ChanceGolo } (P) = \frac{\text{largura\_baliza}}{dist} * \frac{min(\alpha, \text{ângulo\_penalti })}{\text{ângulo\_penalti}} * \frac{1}{1 + \text{ Overall risk(pass } (P, \text{Goalkeeper }))}$$

Equação 8 – Cálculo da chance de golo (Delibas et al., 2019)

A ferramenta como instrumento de auxílio para um treinador e restante staff para detetar erros de decisão dentro de campo e criar situações hipotéticas para orientar os seus jogadores é bastante útil devido ao grande foco em calcular valores concretos de passe e de chances de golo. No entanto, no contexto de um observador de futebol, não revela grande utilidade, visto que não classifica os melhores jogadores nem tem funcionalidades de pesquisa. Porém, os cálculos e o tipo de abordagem científica efetuados demonstram grande utilidade para este projeto. O cálculo de *Effectivness* do passe é usado depois para criar um modelo que usa estatísticas históricas de passes concluídos de forma a perceber quando é que uma decisão acaba por ser boa ou não, e aplicar este cálculo para saber se um jogador efetua passes mais arriscados, mais conservadores e perceber a sua eficácia nesse tipo de lances, de forma a perceber melhor a sua qualidade como passador da bola. O cálculo de ChanceGolo também pode ser útil para perceber quais são os jogadores mais eficazes em frente à baliza.

2.5.2 Projetos e Produtos

2.5.2.1 GoalPoint Partners

A GoalPoint Partners é uma empresa de consultoria estatística de futebol que também presta serviços editoriais, de marketing e de publicidade com foco numa criativa análise estatística sobre futebol. Além do seu site e das suas páginas de rede social, onde são escritos artigos de análise de todos os jogos com participação de equipas portuguesas e sobre jogadores que se estejam a destacar pelo mundo fora, têm um serviço destinado a profissionais do mundo de futebol chamado GoalPointPro, que consiste na interpretação, transformação e análise de dados estatísticos de futebol através dos seus próprios métodos desenvolvidos, com vista a ajudar o proponente da tomada de decisão. (GoalPoint, 2017)

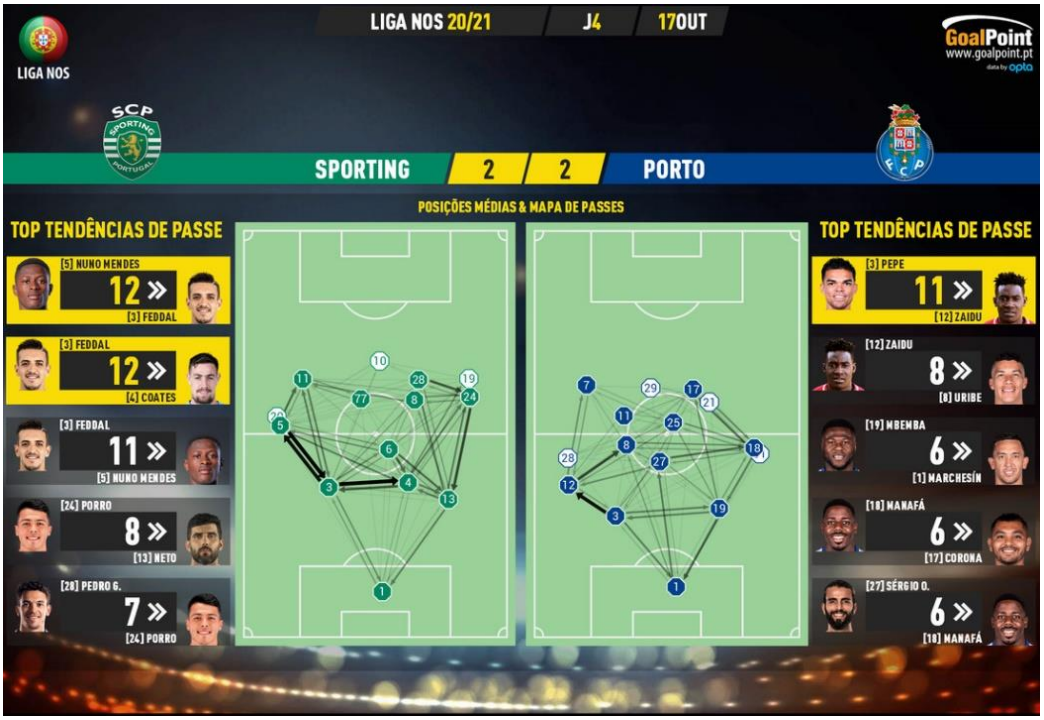


Figura 16 – Exemplo de análise efetuada pela GoalPoint (GoalPoint, 2017)

Os serviços oferecidos pela GoalPointPro são os seguintes:

- Player Scouting, que consiste na procura dos melhores jogadores disponíveis com base em critérios como idade e posição.
- Player Matching, uma funcionalidade que permite encontrar a partir de um perfil de jogador outros perfis semelhantes.
- Target Profiling, para o caso de o cliente já ter alvos previamente referenciados, é feita uma observação dos alvos, de forma a gerar relatórios de desempenho baseados na performance recente onde também são avaliados os pontos fortes e fracos.
- Psychometric Analysis, em que é criado um perfil de personalidade dos jogadores desejados, revelando informação sobre a sua inteligência competitiva, bem como a sua eventual integração e desenvolvimento.
- Player Promotion, onde são criados relatórios sobre um jogador do cliente e comparando-o com o desempenho médio de outros futebolistas da sua posição e com outros que sejam referências a nível mundial, com vista a promovê-los.
- Team Profiling, onde são criados relatórios de equipas com base num número de jogos definidos pelo cliente. Esses relatórios mostram os pontos fortes e fracos da equipa, quer a nível coletivo quer a nível individual.

O preço deste serviço não está disponível, sendo necessário um contacto com a equipa de vendas da empresa. A GoalPoint destaca-se dos restantes projetos estudados devido às componentes gráficas que permitem que o utilizador consiga rapidamente perceber os dados estatísticos apresentados, conforme se pode ver na Figura 16.

2.5.2.2 Wyscout

O Wyscout é uma empresa dedicada ao fornecimento de vídeos, estatísticas e informação sobre jogadores e equipas de futebol a profissionais de futebol, fornecendo uma plataforma composta por uma base de dados com mais de 800 competições mundiais de seleções e de clubes, abrangidas por análise de vídeo e estatística e por mais de 550.000 perfis referentes a jogadores e equipas (Wyscout, 2018).

A plataforma desenvolvida oferece soluções divididas em seis perfis: agentes de futebol, olheiros/observadores, treinadores, jogadores, jornalistas e árbitros, sendo que a análise será focada nas componentes oferecidas aos olheiros.

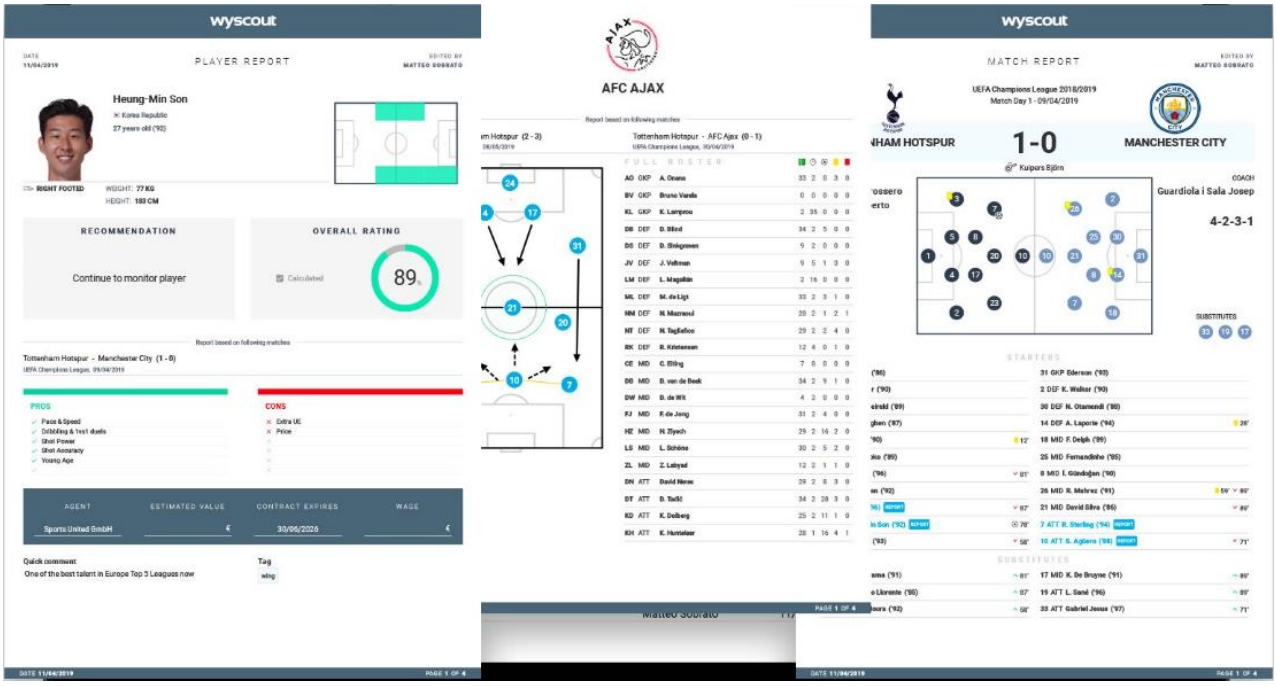


Figura 17 – Ecrãs da plataforma Wyscout (Wyscout, 2018)

Um observador tem acesso a dados completos dos jogadores referentes à equipa ou equipas das quais ficou responsável nos jogos que ele disputou, com uma deteção dos prós e contras, da posição média e de um *rating* geral para cada atleta e uma ficha geral de cada partida com estatísticas avançadas. A plataforma também fornece capacidades de filtragem e pesquisa para encontrar novos alvos, quer a partir de dados diretos de cada jogador como idade ou posição, quer a partir de dados calculados, como capacidade de drible acima da média. Além disso, é possível ver jogos de futebol de jogadores ou equipas à escolha de forma a observar *on-demand*, ou criar *clips* de forma a fundamentar e enriquecer opiniões a colocar num eventual relatório de observação, que pode ser gerado a partir de *templates*. Um exemplo da interface esperada na aplicação pode ser vista na Figura 17.

Esta solução, utilizada por profissionais de renome como Romelu Lukaku e Massimiliano Allegri, revela ser uma das grandes líderes do mercado, com funcionalidades que ajudam a tomada de decisão, fornecendo valor acrescentado a dados estatísticos brutos e com

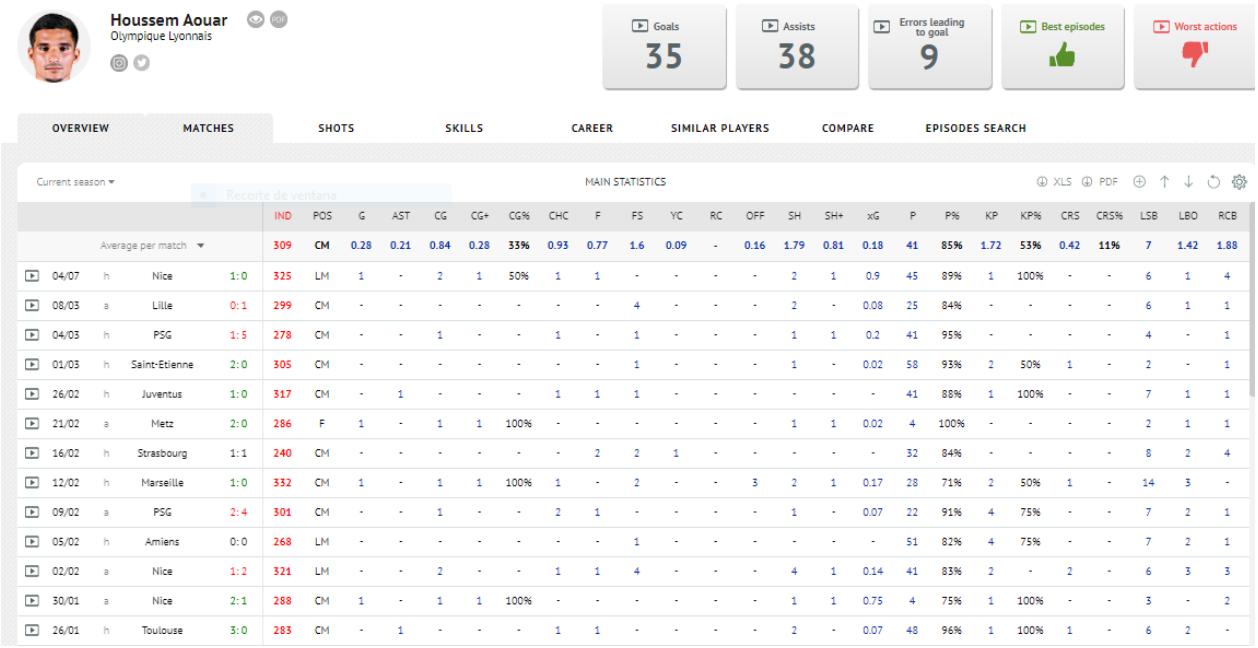
personalização à medida de cada perfil. Os valores de subscrição começam nos 200€ anuais, no entanto, as possibilidades são limitadas, com apenas a possibilidade de visualizar 50 minutos por mês de vídeos e com a criação de um único relatório mensal, pelo que uma equipa profissional com a necessidade de fazer inúmeros relatórios e de analisar imensos jogos por mês terá de subscrever um pacote personalizado. No contexto deste projeto, foi também efetuada uma comunicação com a Wyscout, de forma a perceber a possibilidade na disponibilização dos dados usados, sendo que a resposta recebida foi de que os preços do fornecimento de uma API apenas com os dados, ou seja, sem as ferramentas de análise e de vídeo, começariam nos 5.000€ anuais e poderiam escalar até aos 75.000€ (Wyscout, 2018).

2.5.2.3 InStat Scout

Fundada em 2007 e sediada atualmente em Dublin, mas originalmente de Moscovo, a empresa InStat tem como clientes clubes de renome como Barcelona, Real Madrid, Chelsea ou Bayern Munich, bem como uma relação de trabalho próxima com o futebol irlandês (InStat, 2021a). A plataforma InStat Scout é a solução apresentada pela InStat para observadores de futebol, e contém vídeos, estatísticas e gráficos interativos, com informação sobre mais de 960.000 jogadores de todo o mundo, e com mais de 6.000 jogos carregados para a sua base de dados todos os meses (InStat, 2021b).

A plataforma oferece as seguintes funcionalidades, com um exemplo da interface na Figura 18:

- 95 parâmetros no perfil de equipa e 70 parâmetros no perfil de jogador
- Cada parâmetro é clicável para gerar uma playlist de momentos correspondentes a esse parâmetro
- *Widgets* com bolas paradas e gráficos de remates
- Comparações entre equipas e jogadores de diferentes contextos
- Filtros customizáveis para pesquisa de jogadores
- Exportação de tabelas em formato PDF ou XLS.



				IND	POS	G	AST	CG	CG+	CG%	CHC	F	FS	YC	RC	OFF	SH	SH+	xG	P	P%	KP	KP%	CRS	CRS%	LSB	LBO	RCB
Average per match				309	CM	0.28	0.21	0.84	0.28	33%	0.93	0.77	1.6	0.09	-	0.16	1.79	0.81	0.18	41	85%	1.72	53%	0.42	11%	7	1.42	1.88
04/07	h	Nice	1:0	325	LM	1	-	2	1	50%	1	1	-	-	-	-	2	1	0.9	45	89%	1	100%	-	-	6	1	4
08/03	a	Lille	0:1	299	CM	-	-	-	-	-	-	-	4	-	-	-	2	-	0.08	25	84%	-	-	-	-	6	1	1
04/03	h	PSG	1:5	278	CM	-	-	1	-	-	1	-	1	-	-	-	1	1	0.2	41	95%	-	-	-	-	4	-	1
01/03	h	Saint-Etienne	2:0	305	CM	-	-	-	-	-	-	-	1	-	-	-	1	-	0.02	58	93%	2	50%	1	-	2	-	1
26/02	h	Juventus	1:0	317	CM	-	1	-	-	-	1	1	1	-	-	-	-	-	-	41	88%	1	100%	-	-	7	1	1
21/02	a	Metz	2:0	286	F	1	-	1	1	100%	-	-	-	-	-	-	1	1	0.02	4	100%	-	-	-	-	2	1	1
16/02	h	Strasbourg	1:1	240	CM	-	-	-	-	-	-	2	2	1	-	-	-	-	-	32	84%	-	-	-	-	8	2	4
12/02	h	Marseille	1:0	332	CM	1	-	1	1	100%	1	-	2	-	-	3	2	1	0.17	28	71%	2	50%	1	-	14	3	-
09/02	a	PSG	2:4	301	CM	-	-	1	-	-	2	1	-	-	-	-	1	-	0.07	22	91%	4	75%	-	-	7	2	1
05/02	h	Amiens	0:0	268	LM	-	-	-	-	-	-	-	1	-	-	-	-	-	-	51	82%	4	75%	-	-	7	2	1
02/02	a	Nice	1:2	321	LM	-	-	2	-	-	1	1	4	-	-	-	4	1	0.14	41	83%	2	-	2	-	6	3	3
30/01	a	Nice	2:1	288	CM	1	-	1	1	100%	-	-	-	-	-	-	1	1	0.75	4	75%	1	100%	-	-	3	-	2
26/01	h	Toulouse	3:0	283	CM	-	1	-	-	-	1	1	-	-	-	-	2	-	0.07	48	96%	1	100%	1	-	6	2	-

Figura 18 – Exemplo de ecrã da plataforma (InStat, 2021b)

A InStat oferece também um parâmetro *InStat Index*, um valor gerado por um algoritmo automático que considera a contribuição do jogador para o sucesso da equipa, a significância das suas ações, o nível do adversário e da competição onde o jogador compete. Além disso, cada posição tem entre 12 e 14 parâmetros chaves que irão ser mais preponderantes para o cálculo do *index*. Por exemplo, erros graves cometidos por defesas e a sua frequência irão ter um maior peso do que se forem feitos por avançados. O jogador terá de ter alcançado um determinado número de minutos e ações de forma a que o *index* seja calculado (InStat, 2021c). A Figura 19 demonstra uma listagem de guarda-redes da Serie A ordenados pelo *index* da InStat.

Goalkeepers (1-27)

		InStat Index	Nationality	Age	Date of birth	Weight	Height	Foot	Minutes played	Goals conceded	Shots saved
1	Buffon, Juventus	308	Italy	40	28.01.78	92	192	right	952	9	27 75%
2	Handanovic, Internazional	296	Slovenia	33	14.07.84	92	193	right	1 903	15	58 79%
3	Pepe Reina, Napoli	295	Spain	35	31.08.82	92	188	right	1 793	13	33 72%
4	Szczesny, Juventus	294	Poland	27	18.04.90	85	196	right	955	6	12 67%
5	Junior Costa, Bologna	293	Brazil	34	12.11.83	81	186	right	285	4	8 67%
6	Strakosha, Lazio	293	Albania	22	19.03.95	80	192		1 805	24	75 76%
7	Alisson, AS Roma	292	Brazil	25	02.10.92	91	192	right	1 800	14	55 80%
8	Donnarumma, AC Milan	285	Italy	18	25.02.99	90	196		1 924	27	57 68%
9	Sportiello, Fiorentina	284	Italy	25	10.05.92	87	192		1 909	21	58 73%
10	Puggioni, Sampdoria	281	Italy	37	17.01.81	81	187	right	852	12	40 77%
11	Bizzarri, Udinese	281	Argentina	40	09.11.77	89	193	right	1 240	16	44 73%
12	Gollini, Atalanta	281	Italy	22	18.03.95	80	188		288	3	6 67%
13	Rafael, Cagliari	280	Brazil	35	03.03.82	80	187	left	809	11	24 69%
14	Sirigu, Torino	280	Italy	31	12.01.87	80	192	right	1 911	27	66 71%
15	Cragno, Cagliari	278	Italy	23	28.06.94	78	184	right	1 028	17	49 74%

Figura 19 – InStat Index para Guarda-Redes da Liga Italiana em 2017/2018 (Truica, 2018)

A plataforma, juntamente com a Wyscout (secção 2.5.2.2), é das líderes de mercado. A possibilidade de visualizar ações de cada jogador em formato vídeo a partir da escolha de estatísticas, bem como o cálculo de um rating ponderado com base em parâmetros subjetivos como o nível da competição (ao contrário do artigo Playerank da secção 2.5.1.2), torna a solução como algo extremamente útil para um observador de futebol, que terá sempre informação contextualizada. A InStat não revela preços de subscrição para um serviço completo, tendo apenas disponibilizado o preço do acesso à aplicação móvel a partir das lojas Google Play e AppStore. Os valores estão divididos da seguinte forma:

- 35€ por mês, com acesso a relatórios e vídeos da própria equipa, mais os cinco principais campeonatos europeus, Liga dos Campeões, Liga Europa, Campeonato Mundial e Europeu de Seleções.
- 140€ por mês com acesso a relatórios e vídeos de todas as equipas da base de dados.

Ambos os planos têm um plafond de 3 horas de acesso por mês, sendo que quando terminar, os utilizadores ficam com acesso apenas a vídeos da própria equipa.

A plataforma demonstra algumas das mesmas limitações de outras plataformas estudadas: valores demasiado altos e restrições que impedem uma pesquisa mais aprofundada de jogadores.

2.5.2.4 Football Manager

Football Manager (FM) é uma série de jogos de computador de simulação de gestão futebolística de lançamento anual desenvolvida pela Sports Interactive e publicada pela SEGA (Sports Interactive, 2006). É um *bestseller* em países onde o futebol é culturalmente importante, e o objetivo é simular com a maior precisão possível a carreira de um treinador e, em particular, o jogo, as características dos jogadores e o relacionamento do utilizador com os intervenientes diretos, ou seja, jogadores, treinadores adversários, equipa técnica e direção, e indiretos, como adeptos e jornalistas (Hocquet, 2016). O realismo do jogo, portanto, depende, entre outras coisas, da confiabilidade de um banco de dados de centenas de milhares de jogadores de futebol reais, que é constantemente atualizado. A comunidade de jogadores, organizada em fóruns e redes de prospeção em todo o mundo, participa na construção da base de dados, sendo normalmente dividida por país ou região (Hocquet, 2016).

Devido à natureza da aplicação, um jogo de computador, a quantidade de funcionalidades e opções que um utilizador tem ao seu dispor é vasta, pelo que foram identificadas três vertentes principais a referir nesta secção:

- Acesso a estatísticas de jogos de futebol, e por consequência de jogadores
- Interface gráfica para o treinador ver informação sobre jogadores, equipas e competições
- Cálculo da qualidade de um jogador

NOME	JOGOS	GLS	AST	HDJ	% PASSE	DES	DRBPJ	% REMAT...	CL MED
Tammy Abraham	5 (14)	6	0	2	74%	0,81	1,61	59%	7.27
Jeremie Boga	10 (4)	1	2	1	77%	1,50	2,20	35%	6.90
Ruben Vargas	23 (1)	5	3	2	74%	0,65	2,23	49%	7.05
Kaio Jorge	23 (6)	12	4	2	74%	0,31	1,36	50%	7.11
Marco Asensio	10 (13)	2	4	0	74%	0,72	2,01	32%	6.93
Michał Karbownik	21 (7)	1	1	1	78%	2,05	0,77	33%	6.85
Ianis Hagi	27 (1)	6	1	1	83%	0,86	1,25	33%	6.84
Sean Longstaff	24 (2)	1	6	1	88%	3,13	0,67	33%	6.98
Giorgi Chakvetadze	21 (3)	5	5	1	81%	0,55	1,16	40%	6.98
Douglas Luiz	14 (3)	3	1	0	88%	2,81	0,36	38%	6.91
Ruben Loftus-Cheek	0 (12)	2	0	0	81%	3,58	0,83	36%	6.70
Yves Bissouma	25	1	0	0	84%	2,16	0,63	48%	6.87
Raphinha	3 (10)	3	2	1	70%	2,61	2,01	29%	7.10
Zeki Çelik	25 (1)	0	4	0	85%	3,26	0,77	0%	6.88
Valentino Lazaro	28	1	4	1	79%	2,76	0,67	60%	6.78
Abdou Diallo	27 (1)	2	0	1	83%	3,66	0,04	22%	6.87

Figura 20 – Listagem de jogadores com estatísticas (Sports Interactive and SEGA, 2021)

A Figura 20, retirada do jogo (Sports Interactive and SEGA, 2021), corresponde a um ecrã de pesquisa de jogadores, em que o utilizador poderá ver uma listagem com alvos identificados como “interessados”, que é o termo usado para descrever que a probabilidade de aceitarem uma proposta do clube que se está a treinar é alta. O utilizador poderá escolher que tipo de colunas é que a tabela de jogadores terá, desde informação biológica do jogador, como nacionalidade e idade, até informação contratual. Para o efeito deste projeto, selecionou-se a vista de estatísticas oferecidas por defeito pelo jogo, sendo que o jogo também permite que essas colunas sejam personalizadas, isto é, seria possível adicionar a esta tabela uma coluna correspondente à idade do jogador, por exemplo. As colunas correspondem a estatísticas geradas pelo motor de jogo da aplicação, e sendo valores ficcionais, a possibilidade de vermos qualquer estatística é vasta.

Figura 21 – Ecrã de filtro/pesquisa avançada (Sports Interactive and SEGA, 2021)

A Figura 21 demonstra um ecrã de filtro, onde o utilizador poderá filtrar a listagem inicial de jogadores. Neste caso, o utilizador definiu que pretendia encontrar jogadores que atuasse na posição de médio centro, com uma idade compreendida entre os 21 e os 25 anos, que jogue num clube da América do Sul e que efetua pelo menos dois passes decisivos a cada 90 minutos. Existem outras opções de filtragem, como a data de expiração de contrato ou a nacionalidade do jogador.

Sendo este software um jogo com uma vertente fortíssima na simulação de partidas de futebol, a quantidade de dados gerados e a disponibilidade imediata dos mesmos permite criar artefactos que numa situação real são quase impossíveis de gerar. No entanto, o jogo parte de uma configuração inicial que necessita de um contributo enorme de forma a que o jogo possua uma das base de dados de futebol mais ricas e completas no mundo, chegando a ser usada por clubes profissionais de futebol (Stuart, 2014). A Sports Interactive tem cerca de 86 coordenadores de pesquisa, cada um com uma equipa sob a sua coordenação, resultando num total de 1250 *researchers* (nos quais se insere o autor desta dissertação), que organizam uma base de dados com cerca de 350.000 pessoas, desde jogadores a treinadores, 60.000 equipas e milhares de competições (Hamilton and Karlsen, 2020).

Figura 22 – Perfil de jogador (Sports Interactive and SEGA, 2021)

Além de todos os dados pessoais e profissionais dos jogadores que são inseridos, existe também a categoria de dados futebolísticos, onde o jogador tem atributos de 0 a 20, sendo 0 o mínimo e 20 o máximo, que correspondem à qualidade do mesmo em ações de jogo como passe, desarme ou remate, que são preenchidos pela equipa indicada anteriormente, conforme se vê na Figura 22. Todos esses atributos resultam, ou têm de ir ao encontro, de um valor relativamente absoluto da qualidade do jogador, a Capacidade Atual (CA), um valor numérico de 0 a 200 (Hamilton and Karlsen, 2020). O valor da CA é relativo à posição em que o jogador joga, isto é, se um jogador tiver o valor máximo no atributo de remate, o seu peso relativo para o CA será maior se for um avançado ao invés de se for um defesa. A Capacidade Atual do jogador quanto maior for, melhor será o jogador. Por exemplo, o guarda-redes condecorado do Bayern, Manuel Neuer, tem um valor entre os 180 e os 190, enquanto que um jogador das divisões amadoras inglesas tem um valor na região dos 50 e 60 (Hamilton and Karlsen, 2020). O jogo também tem um valor correspondente à Capacidade Potencial (CP),

mas esse valor acaba por resultar de uma avaliação empírica e subjetiva dos *researchers* responsáveis pelo jogador, em que utilizam a sua experiência passada de análise de jogadores com características semelhantes para perceberem se o jogador em questão tem uma capacidade de evolução alta ou baixa (Hocquet, 2016).

Esta solução, sendo um jogo de computador cujo foco é a simulação da realidade, não vai ao encontro do problema proposto. No entanto, a interface gráfica, que permite ao jogador, que está a simular o papel de decisor, ter acesso a uma grande quantidade de informação personalizada, e a inspiração resultante de anos de utilização desta aplicação por parte do autor desta dissertação, demonstram uma grande importância para o contexto deste problema. Além disso, a base de dados, atestada por equipas especializadas em juízos de qualidade de jogadores de futebol, servirá como ferramenta-chave para a validação da implementação da solução.

2.5.3 Ferramentas e Tecnologias

2.5.3.1 R

R é uma linguagem e ambiente para programação estatística e foi desenvolvida pelos laboratórios Bell por uma equipa liderada por John Chambers. A linguagem R fornece uma grande variedade de técnicas estatísticas e gráficas, como modelação linear e não linear, testes estatísticos, análises temporais e agrupamentos, e é altamente extensível, sendo que um dos pontos fortes do R é a facilidade em desenhar gráficos de forma a irem ao encontro das necessidades científicas, incluindo símbolos e fórmulas matemáticas quando preciso, isto tudo graças à gama de ferramentas referidas anteriormente (Bell Laboratories, 2017).

Tendo em conta a natureza deste projeto, que tem uma grande preponderância na análise de dados estatísticos, esta linguagem revela-se como a mais adequada para esse uso (Bell Laboratories, 2017). As suas funcionalidades incluem:

- Manipulação e facilidade de armazenamento de dados;
- Operadores para cálculos em *arrays* e matrizes;
- Coleção vasta de ferramentas intermédias para análise de dados;
- Recursos gráficos para análise de dados;
- Possibilidade de utilização de conceitos elementares de programação, como ciclos, condições e funções recursivas, bem como ferramentas de leitura e escrita de ficheiros;
- Fácil integração com outras linguagens;
- Extensível via bibliotecas externas desenvolvidas por utilizadores ou pelos responsáveis da linguagem;
- Documentação online.

A linguagem R é poderosa para análise estatística, sendo por isso a mais indicada para responder ao proposto na secção 1.3. Serão utilizados algoritmos de regressão linear (secção

2.3.1) e de máquinas de vetores suporte (secção 2.3.2) disponibilizados gratuitamente pela comunidade.

2.5.3.2 SQL

Structured Query Language (SQL) é uma linguagem de base de dados composta por uma *data definition language* (DDL), que permite a especificação de esquemas de bases-de-dados; uma *data manipulation language* (DML), que oferece suporte a operações para recuperar, armazenar, modificar e excluir dados; e uma *data control language* (DCL), que permite aos administradores configurar os acessos à base de dados (Silva et al., 2016).

A popularidade do seu uso deve-se à sua sintaxe de alto nível, em que não se precisa de ter conhecimentos de programação para muitas das *queries* (pedidos à base de dados) e à sua implementação em todos os tipos de sistemas de gestão de base de dados, de *desktop* (Microsoft Access) a *open-source* (MySQL e PostgreSQL) e comercial (Oracle, IBM DB2, Microsoft SQL Servers). A sua ampla adoção tornou-se facilitada devido à estandardização do SQL pela ISO com a ANSI e várias agências nacionais (Fotache and Strimbei, 2015).

Devido a ser uma linguagem vastamente utilizada para armazenamento de dados e ao seu uso e ensino ao longo da licenciatura e mestrado, o SQL revela-se como a melhor opção para resolver o problema de como e onde registar os dados obtidos.

2.5.3.3 Framework .NET

.NET é uma plataforma de desenvolvimento que fornece uma interface de programação e integração entre serviços de Windows, APIS e tecnologias, direcionada ao desenvolvimento web (Thai and Lam, 2003). Com o seu desenvolvimento iniciado em 1997 de forma a responder ao surgimento da linguagem de programação Java, a Microsoft definiu quatro princípios fundamentais (Syme, 2020):

- Suporte a múltiplas linguagens de programação, como Visual Basic e C++.
- Suporte a *garbage collection*, inteiros sem sinal, compilação no ato de instalação e outras funcionalidades de *middleware*.
- Seria desenvolvida especificamente para desenvolvimento de aplicações em Windows, com interoperabilidade nativa com APIs baseadas em C e suporte com COM. No entanto, deveria ser genérica o suficiente para que um *porting* para outros sistemas fosse possível.
- O SDK seria gratuito e alinhado com os esforços emergentes à altura no estabelecimento de relações académicas.

Atualmente, transformou-se numa plataforma *open-source* que permite o desenvolvimento de aplicações de várias linguagens, como C#, F#, ou Visual Basic, e com recurso a várias bibliotecas para fins de web, mobile, desktop, jogos ou IoT (Microsoft, 2021). Essa versatilidade, e sendo um dos principais fins do seu uso o desenvolvimento de aplicações web, a tecnologia foi identificada para a vertente de demonstração de dados ao utilizador.

3 Análise de Valor

O capítulo da análise de valor relata o estudo efetuado em relação à mais-valia potencial do projeto a ser desenvolvido. Será dividido em seis subsecções, nomeadamente: *New Concept Development*, Definições de Valores, Benefícios e Sacrifícios, Proposta de Valor, Modelo de *Canvas* e Método de Análise Hierárquica para Comparação de Tecnologias. Esta divisão permitirá ao leitor acompanhar e perceber a descoberta de valor que foi efetuada para o projeto.

3.1 New Concept Development

Oito empresas-membro da *Process Effectiveness Network* pertencente à *Industrial Research Institute* tinham a "perceção clara de uma falta de ideias a entrar no processo de desenvolvimento" em comparação com a fase de *Product and Process Development* (Koen et al., 2001), e sentiram a necessidade de descobrir as melhores práticas do *Fuzzy Front End* (FFE) da inovação, com o objetivo de estruturar e generalizar o processo de desenvolvimento de uma nova ideia ou conceito, independentemente das diferentes metodologias de trabalho e negócios de cada empresa (Koen et al., 2001). O resultado foi a criação do modelo *New Concept Development* (NCD), demonstrado na Figura 23.

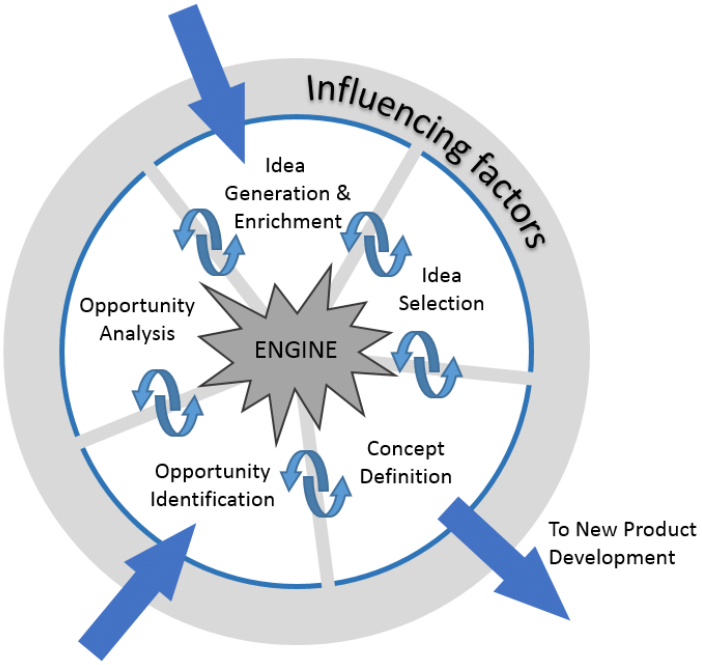


Figura 23 – Modelo de New Concept Development (Koen et al., 2001)

Resumidamente, o modelo NCD tem cinco elementos colocados numa forma circular e com setas que indicam a sugestão de que ideias flutuam entre os cinco membros em qualquer ordem ou combinação. Existem fatores influenciadores que podem ser identificados como as estratégias de negócio e as capacidades da organização, a evolução tecnológica das ferramentas e a restante competição. Todo este processo tem como fatores críticos a liderança e a cultura da organização. (Koen et al., 2001)

3.1.1 Identificação da Oportunidade

A organização identifica as oportunidades que se poderão explorar, sendo este passo motivado pelos objetivos de negócio, como uma resposta a uma ameaça competitiva ou um avanço científico ou tecnológico para se distanciar dos restantes. (Koen et al., 2001)

A oportunidade apresenta-se com o crescimento das ferramentas e tecnologias de data mining (secção 1.1), bem como do mercado ainda se encontrar aberto a novas ferramentas que auxiliem o processo de observação de jogadores e de uma melhor interpretação de dados estatísticos (secção 2.2), sendo também reforçada pela constante procura por inovar no que diz respeito a ferramentas e processos desportivos (secção 2.1) e pela necessidade em sustentar opiniões e juízos subjetivos (secção 1.2).

3.1.2 Análise da Oportunidade

Nesta fase, a informação adicional é usada para traduzir o passo da Identificação da Oportunidade num negócio específico e em oportunidades tecnológicas, analisando o mercado a partir de estudos ou efetuando experiências científicas. (Koen et al., 2001)

A partir da análise efetuada sobre o que existe no mercado (secção 2.5), a oportunidade revela-se como relevante. As soluções encontradas e estudadas apresentam um preço impeditivo para profissionais de clubes com menos recursos financeiros (ou freelancers) ou pouco equilíbrio na informação apresentada, ou então não há uma explicação do que significa cada dado estatístico, quer simples, quer calculado.

3.1.3 Génese da Ideia

Na fase da Génese da Ideia, uma ideia concreta nasce a partir da maturação e do desenvolvimento de uma oportunidade, sendo criada a partir de um processo formal com sessões de *brainstorming* ou bancos de ideia, ou então de uma forma informal e incomum, como um pedido inusitado de um cliente ou a chegada um novo material por parte dos fornecedores. (Koen et al., 2001)

Com vista no problema encontrado, surgiu a ideia de criar uma aplicação com dois potenciais focos identificados:

- Foco no trabalho de um observador, em que a aplicação teria uma interface dinâmica e ajustável que permita analisar o perfil estatístico e pessoal de variados jogadores de futebol, de forma a ajudar o observador a tomar as melhores decisões.
- Foco no trabalho de um treinador, em que a aplicação, com base em dados de jogos de futebol anteriores, sugira qual a melhor estratégia a seguir no próximo jogo de forma a contrariar os pontos fortes e atacar os pontos fracos da equipa adversária, bem como a apresentação de relatórios sobre como a sua equipa e os seus jogadores se encontram.

3.1.4 Seleção da Ideia

Esta fase corresponde à escolha da ideia, podendo ser um método de escolha simples, como escolher uma ideia de muitas que um individuo pensou, ou uma seleção a partir de um portefólio ou de um banco de ideias. (Koen et al., 2001)

Com as duas ideias identificadas na secção 3.1.3, a segunda ideia apresentada revela uma maior complexidade no que diz respeito à análise científica necessária para criar um sistema pericial que permitisse identificar pontos fortes e fracos de duas equipas de futebol para detetar a melhor forma de suceder numa situação futura, isto é, o jogo seguinte. Por outro lado, a primeira ideia consiste em analisar dados estatísticos já existentes e retirar informação

deles de forma a ajudar na criação de uma opinião ou de um relatório. Sendo assim, escolheu-se a primeira ideia identificada.

3.1.5 Desenvolvimento do Conceito e Tecnologia

No passo final do NCD, um *business case* é desenvolvido com base nas estimativas de mercado, necessidades do cliente, necessidades de investimento e risco geral do desenvolvimento do projeto. (Koen et al., 2001)

Com base na oportunidade encontrada e na ideia selecionada, inicialmente serão analisados os dados estatísticos possíveis de serem obtidos, de forma a se chegar a uma ideia do valor que se poderá retirar e demonstrar ao utilizador final na plataforma a desenvolver. Isto implicará também que a plataforma consista em vários módulos, desde a obtenção dos dados até à obtenção de conhecimento.

3.2 Valor, Valor para o Cliente, Valor Percecionado

A definição de valor é algo que tem vindo a ser estudado, com imensos autores a tentarem explicar diferentes tipos de valor, desde o valor laboral por Karl Marx e Frederich Engels, que definem que o valor de uma comodidade é determinada pelo tempo médio necessário para a produzir (Marx, 1942), até ao paradoxo de valor de Adam Smith, em que coloca a contradição de que a água, um bem intrinsecamente bastante valioso, tem um preço menor no mercado do que diamantes (Smith, 1817). Mais recentemente, Francis Buttle define que valor "é a combinação da perceção e utilidade que o cliente experienciou ao usar um produto ou serviço" (Buttle, 2019). Como tal, revela-se essencial explorar os conceitos relacionados com o valor para o cliente e valor percecionado.

O valor para o cliente, segundo Howard Butz e Leonard Goodstein, resulta da ligação estabelecida com o produto ou serviço depois de o ter utilizado e de ter considerado que providenciava valor acrescentado, dividindo em três subvalores: valor esperado, algo que o cliente naturalmente espera, valor desejado, algo que o cliente deseja mas não está presente nas restantes plataformas, e valor não antecipado, algo além do que o cliente espera e deseja (Butz and Goodstein, 1996). Aplicando estes conceitos ao âmbito deste projeto, o valor para o cliente é ter uma plataforma que lhe permita ter factos sobre jogadores que se encontra a observar, sendo depois a divisão efetuada a seguinte:

- Valor esperado: Sustentar opiniões sobre jogadores que se encontra a observar e da própria equipa na qual trabalha;
- Valor desejado: Encontrar jogadores de acordo com um perfil ou características selecionadas que tragam valor ao clube para o qual trabalha ou criar um portefólio de jogadores para sugerir a clubes como *freelancer;*
- Valor não antecipado: Obter factos que contraponham a sua opinião e detetar alvos em mercados e clubes alternativos.

O valor percecionado é a opinião que o cliente cria sobre a utilidade do serviço ou produto com base no que entende serem os benefícios obtidos em relação aos custos, além dos sacrifícios que terá de fazer (Zeithaml, 1988). No caso deste projeto, os benefícios da aplicação são o apoio que o utilizador irá ter na sua tarefa de observação de jogadores, suportando qualquer decisão que faça, e manter-se atualizado sobre os novos alvos mais desejados do desporto. A nível de sacrifícios, existe um tempo de adaptação à utilização do software, quer seja por nunca ter usado algo deste tipo ou por ter utilizado algum da concorrência, e existe também uma possibilidade de algumas opiniões do utilizador consideradas dogmáticas serem postas em causa pelos resultados de análise fornecidos pela aplicação.

3.3 Proposta de Valor

A proposta de valor descreve os benefícios que os clientes podem esperar dos produtos e serviços em junção com os ganhos que se podem adquirir e as dores que se podem aliviar (Osterwalder et al., 2014).

Foi utilizado o *Value Proposition Canvas*, disponibilizado pela Canvas Generation (Canvas Generation, 2021), de forma a estruturar as definições referidas, como se encontra ilustrado na Figura 24.

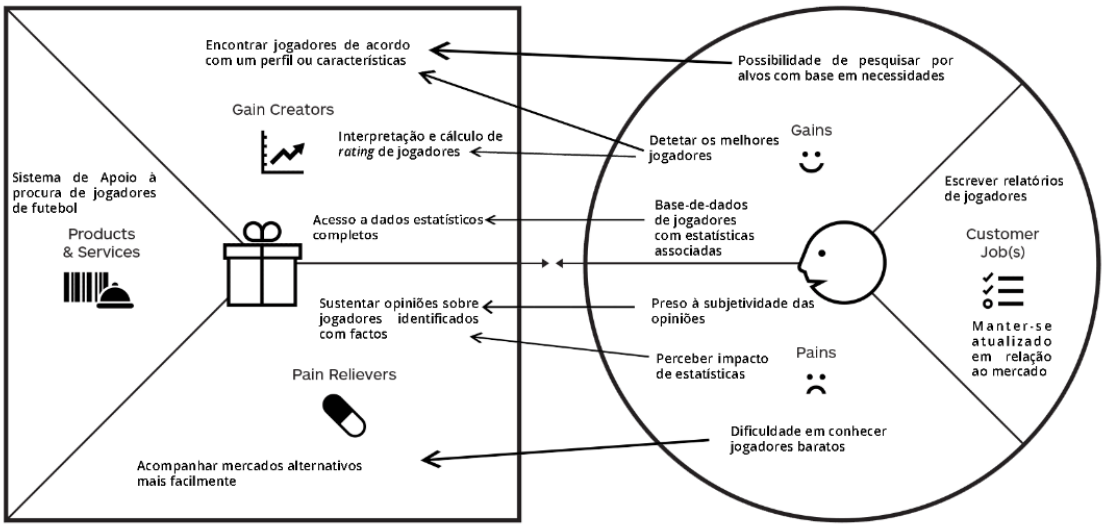


Figura 24 – Diagrama de Proposta de Valor (Canvas Generation, 2021)

A proposta de valor refere-se ao desenvolvimento de um sistema de apoio à procura de jogadores de futebol, com vista a encontrar jogadores de acordo com um perfil ou características, tendo acesso a dados estatísticos dos jogadores e com um cálculo de um *rating* baseado nessas mesmas estatísticas. Esta solução irá permitir que os observadores sustentem opiniões com recurso a factos e a acompanhar mercados alternativos mais facilmente.

A solução proposta irá trazer ganhos na pesquisa de alvos com base em necessidades que o observador precisa de satisfazer e na deteção de quais são os melhores jogadores estatisticamente, isto tudo com recurso a uma base de dados extensa. O observador conseguirá perceber o impacto real das ações que um jogador efetua e a objetivar as suas opiniões, bem como diminuir a dificuldade em detetar jogadores baratos. Tudo isto será com o propósito de escrever relatórios sobre jogadores e de se manter atualizado sobre o mercado de jogadores.

3.4 Método QFD para determinar prioridades

Quality Function Deployment (QFD) é uma abordagem orientada ao cliente para a inovação de produtos, que orienta tanto gestores como equipas de *design* de produto durante a conceptualização, criação e realização de novos produtos. O QFD estrutura o relacionamento entre as necessidades do mercado, com recurso a especificações técnicas e a especificações de produção, e as capacidades de operação, de forma a criar um plano de trabalhos (Govers, 1996).

O propósito do QFD possui três vertentes (Warwick Manufacturing Group, 2007):

- Oferecer melhor qualidade ao produto de modo a vender mais rapidamente ao menor custo possível;
- Assegurar o design do produto pretendido pelo cliente;
- Disponibilizar um sistema de controlo para futuros melhoramentos, tanto ao nível de design do produto, como ao nível do processo.

O QFD aborda o problema da inovação do produto dividindo os requisitos do cliente em segmentos e identificando meios para atingir cada segmento, envolvendo também todas as partes de uma empresa, facilitando o design simultâneo de produtos e processos (Warwick Manufacturing Group, 2007). O elemento utilizado para a avaliação QFD denomina-se *House Of Quality* (*HOQ*), sendo uma matriz que oferece uma relação entre a perspetiva do cliente e a perspetiva técnica do produto. A Figura 25 representa um esquema genérico desta matriz, onde se pode verificar uma estrutura semelhante à de uma casa, daí o seu nome (Kukhnavets, 2019).

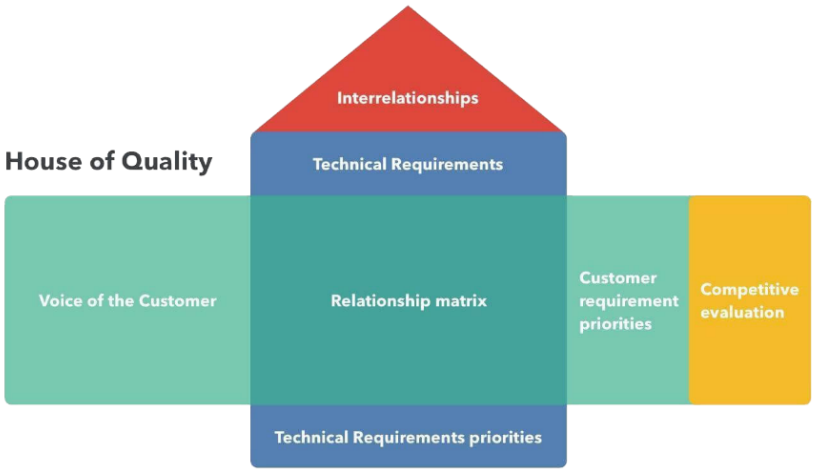


Figura 25 – House of Quality (Kukhnavets, 2019)

Observando a House of Quality fornecida pela Hygger e da autoria de Kukhnavets, pode-se retirar que esta mesma matriz se divide em sete partes. Começando pelo "telhado", representa a relação entre os vários requisitos técnicos, que permite perceber quais os que dependem uns dos outros, de maneira a adotar a melhor estratégica para uma maior compatibilidade.

A parte intermédia da casa contém uma matriz que se pode dividir em quatro partes:

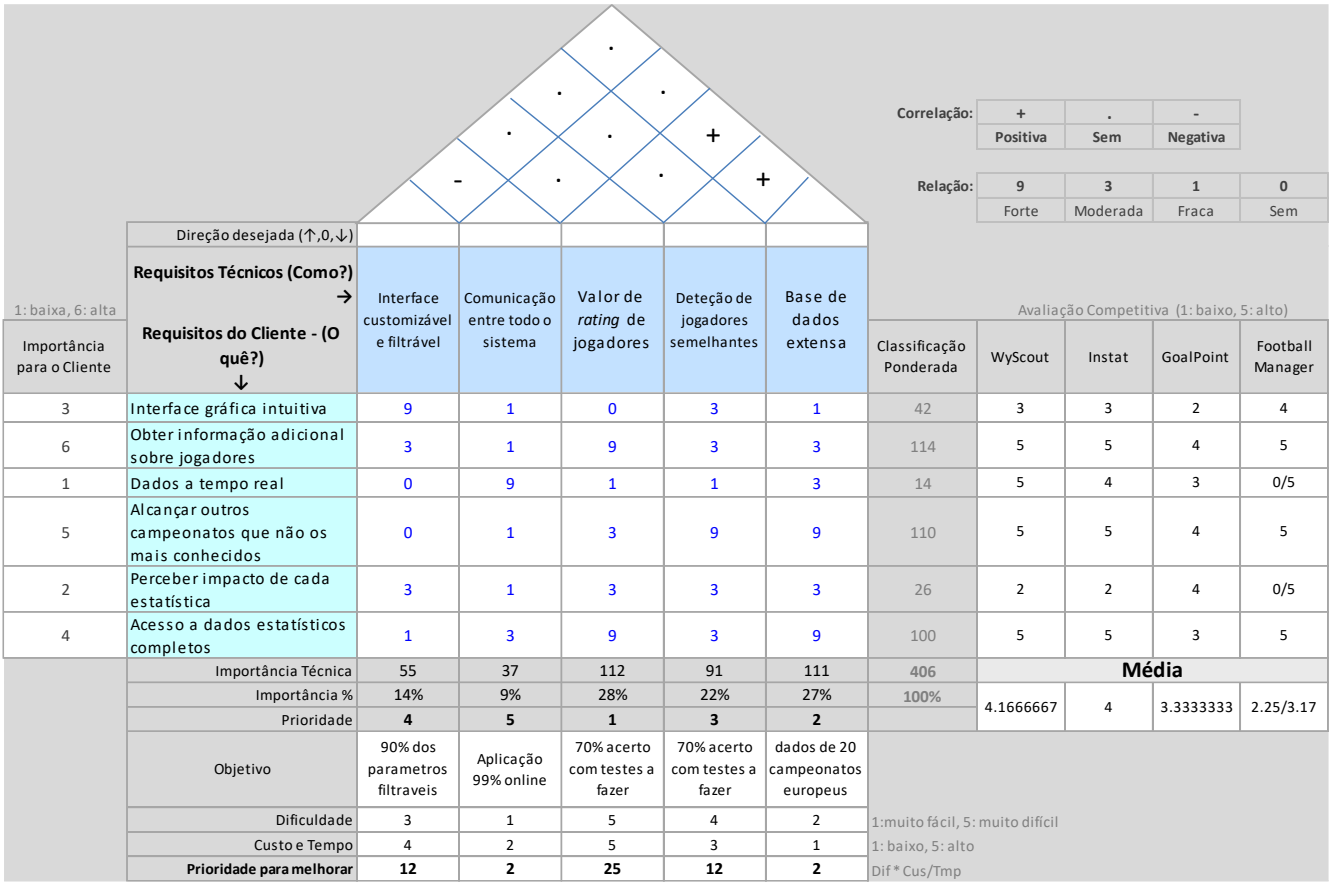
- Requisitos do cliente, representados do lado esquerdo, onde também se encontra a importância para o projeto.
- Requisitos técnicos, seguidos da relação entre os mesmos e os requisitos do cliente;
- Perspetiva da prioridade do cliente em relação aos seus requisitos;
- Perceção de como é que a competição se comporta nos requisitos do cliente.

Por último, na base desta matriz, está representada a prioridade a dar a cada requisito técnico.

Assim sendo e fazendo o paralelismo com o objetivo desta tese, a Figura 26 representa a matriz *HOQ* relativa ao desenvolvimento do software de gestão proposto.

Os requisitos do cliente identificados foram os seguintes, com as importâncias definidas para o cliente numa escala de 1 a 6, sendo 1 a importância mais baixa e 6 a importância mais alta:

- Interface gráfica intuitiva — importância 3
- Obter informação adicional sobre jogadores — importância 6
- Dados a tempo real — importância 1
- Alcançar outros campeonatos que não os mais conhecidos — importância 5
- Perceber impacto de cada estatística — importância 2
- Acesso a dados estatísticos completos — importância 4



							Correlação:	+	.	-	
								Positiva	Sem	Negativa	
							Relação:	9	3	1	0
								Forte	Moderada	Fraca	Sem
	Direção desejada (↑,0,↓)										
1: baixa, 6: alta	**Requisitos Técnicos (Como?)** →	Interface customizável e filtrável	Comunicação entre todo o sistema	Valor de *rating* de jogadores	Deteção de jogadores semelhantes	Base de dados extensa	Avaliação Competitiva (1: baixo, 5: alto)				
Importância para o Cliente	**Requisitos do Cliente - (O quê?)** ↓						Classificação Ponderada	WyScout	Instat	GoalPoint	Football Manager
3	Interface gráfica intuitiva	9	1	0	3	1	42	3	3	2	4
6	Obter informação adicional sobre jogadores	3	1	9	3	3	114	5	5	4	5
1	Dados a tempo real	0	9	1	1	3	14	5	4	3	0/5
5	Alcançar outros campeonatos que não os mais conhecidos	0	1	3	9	9	110	5	5	4	5
2	Perceber impacto de cada estatística	3	1	3	3	3	26	2	2	4	0/5
4	Acesso a dados estatísticos completos	1	3	9	3	9	100	5	5	3	5
	Importância Técnica	55	37	112	91	111	406	**Média**			
	Importância %	14%	9%	28%	22%	27%	100%	4.1666667	4	3.3333333	2.25/3.17
	Prioridade	**4**	**5**	**1**	**3**	**2**					
	Objetivo	90% dos parametros filtraveis	Aplicação 99% online	70% acerto com testes a fazer	70% acerto com testes a fazer	dados de 20 campeonatos europeus					
	Dificuldade	3	1	5	4	2	1:muito fácil, 5: muito difícil				
	Custo e Tempo	4	2	5	3	1	1: baixo, 5: alto				
	Prioridade para melhorar	**12**	**2**	**25**	**12**	**2**	Dif * Cus/Tmp				

Figura 26 – Diagrama QFD para a solução proposta (Saadeddin, 2021)

Os requisitos técnicos identificados foram os seguintes, com a importância técnica calculada a partir dos valores da matriz com os da importância para o cliente:

- Interface personalizável e filtrável — importância técnica 55
- Comunicação entre todo o sistema — importância técnica 37
- Valor de rating de jogadores — importância técnica 112
- Deteção de jogadores semelhantes — importância técnica 91
- Base de dados extensa — importância técnica 111

Seguindo a lógica aplicada na análise da figura 25, o telhado da HOQ apresenta três símbolos que correspondem ao poder dessa mesma relação, respetivamente, “.”, sem correlação; “+”, correlação positiva; e “-“, correlação negativa. Após esta descrição pode-se retirar as seguintes interpretações com base na análise do telhado da matriz:

- Base de dados extensa <> Deteção de jogadores semelhantes — Esta relação é positiva, visto que uma base de dados que contenha informação vasta e completa irá fazer com que a tarefa correspondente à identificação de jogadores semelhantes se torne mais fácil de efetuar, visto que o modelo de aprendizagem terá mais e melhores dados para efetuar os seus cálculos

- Base de dados extensa <> Valor de rating de jogadores — Esta relação é positiva, e segue uma lógica semelhante ao ponto anterior, isto é, quanto mais e melhores dados a base de dados tiver, o cálculo de obtenção do rating de jogador será melhor.
- Interface personalizável e filtrável <> Comunicação entre todo o sistema — Esta relação é negativa visto que, caso não haja uma comunicação forte entre o sistema ao ponto de os dados serem obtidos corretamente, a interface não terá capacidades de demonstrar os dados que o utilizador solicita.

Todas as restantes correlações são inexistentes ou sem grande impacto direto nas várias capacidades técnicas. De seguida, de algumas indicações mais gerais que se pode retirar deste diagrama, sobressaem-se as seguintes:

- O WyScout é a plataforma que mais ao encontro vai das necessidades do cliente. Destaca-se também o caso especial do Football Manager, que devido a ser um jogo, alguns dos campos podem ser 0 ou 5, dependendo da perspetiva que se quiser ter, sendo isso indicado pelo valor 0/5 e pelo cálculo da média para o caso de se considerar cada um dos valores.
- Por ordem, os requisitos técnicos mais importantes para o sistema são: Valor de rating de jogadores, base de dados extensa, Deteção de jogadores semelhantes-
- O requisito do Valor de rating de jogadores é igualmente identificado como a mais difícil e a que mais tempo irá necessitar. O requisito da base de dados extensa, apesar de ser das mais importantes, revela-se como uma das menos dispendiosas, visto que os dados não serão obtidos pela própria aplicação, mas sim a partir de uma API externa.

Em suma, pode-se concluir que a os requisitos referente às funcionalidades de *machine learning* são as mais importantes e que apresentam um maior custo e tempo necessário, sendo estes os requisitos que os clientes mais valorizam numa solução com vista a auxiliar a sua tomada de decisão.

4 Análise e Design

Neste capítulo é apresentada a ideia inicial de desenvolvimento do projeto com vista a resolver o problema encontrado, com base na inspiração de outras soluções e investigações existentes, bem como nas oportunidades encontradas e decisões tomadas no que diz respeito ao valor potencial da solução.

4.1 Requisitos

4.1.1 Requisitos Funcionais

Na Figura 27 está representado o diagrama de casos de uso para a solução proposta, de forma a representar os requisitos funcionais.

Os casos de uso identificados foram os seguintes:

UC1: Obter Dados da API

A Aplicação de Obtenção de Dados deverá conseguir obter os dados necessários a partir das APIs referidas no capítulo 4. Este caso de uso será executado automaticamente com uma frequência mínima de uma vez a cada três dias.

UC2: Pesquisar Jogadores com Filtros Personalizados

O observador terá a possibilidade de pesquisar jogadores com base em filtros e características definidas por si, como estatísticas presentes na base de dados, como golos ou duelos aéreos ganhos na época toda, ou dados do jogador, como nacionalidade ou idade.

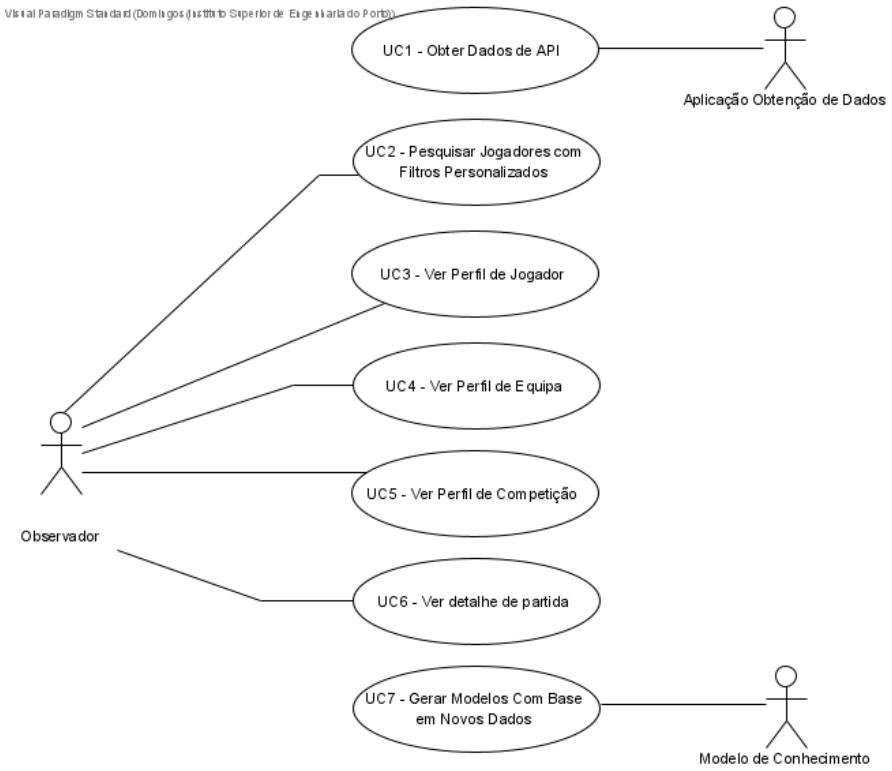


Figura 27 – Casos de Uso

UC3: Ver Perfil de Jogador

Neste caso de uso, o observador terá acesso ao perfil completo do jogador que escolheu. Poderá ver dados estatísticos completos e gerados no Modelo de Conhecimento, histórico de performance e jogadores semelhantes ao perfil selecionado.

UC4: Ver Perfil de Equipa

Semelhante ao UC2, neste caso a listagem será de jogadores da mesma equipa e com dados da equipa, como o onze mais vezes utilizado, forma da equipa, próximo jogo e posição do campeonato.

UC5: Ver Perfil de Competição

O caso de uso de Ver Perfil de Competição, o observador poderá ver a tabela da competição e os jogos disputados neste contexto, bem como os melhores executantes de vários parâmetros, como golos, passes chave e outras estatísticas presentes na base de dados, quer reais ou geradas.

UC6: Ver Detalhe de Partida

O observador poderá ver com detalhe cada partida presente na base de dados, visualizando os onzes de cada equipa e a forma como cada elemento da equipa contribuiu para o resultado do jogo disputado.

UC7: Gerar Modelos Com Base em Novos Dados

O Modelo de Conhecimento terá de incrementalmente aperfeiçoar os modelos existentes com recurso a novos dados existentes na base de dados. Este caso-de-uso está intrinsecamente ligado ao UC1, isto é, assim que o UC1 terminar a sua execução, o Modelo de Conhecimento deverá gerar os novos dados.

A Figura 28 representa o processo que irá ser o conjunto dos casos-de-uso de Obtenção de Dados e de Gerar Modelos com Base em Novos Dados. A Aplicação de Obtenção de Dados inicia o processo ao obter os dados brutos para análise e trata-os, atribuindo a nomenclatura correta para a solução. Depois, os dados são registados na base de dados central do sistema, que serão usados pelo Modelo de Conhecimento para iniciar o processo de aprendizagem. Após esse processo estar terminado, ou seja, os cálculos resultarem em novos dados, é efetuado um novo registo na base-dados, desta vez com os resultados da aprendizagem, terminando o processo.

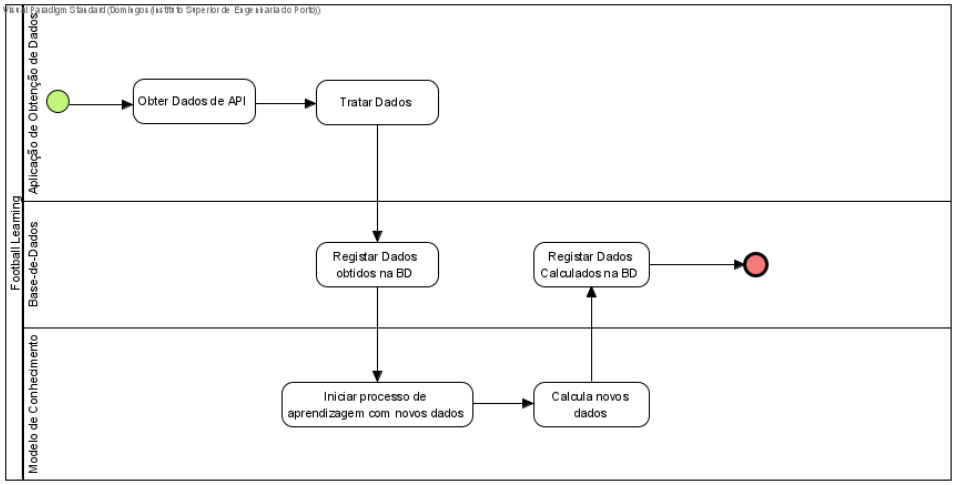


Figura 28 – Processo de Obtenção de Dados e Geração do Modelo

4.1.2 Requisitos Não Funcionais

De forma a representar os Requisitos Não Funcionais da solução, utilizou-se a classificação FURPS+, acrónimo para *Functionality* (Funcionalidade), *Usability* (Usabilidade), *Reliability* (Fiabilidade), *Performance* (Desempenho) e *Supportability* (Suportabilidade) (Grady and Caswell, 1987). Para este projeto, foram identificados os seguintes atributos usando a classificação referida:

- **Funcionalidade** — os modelos de aprendizagem usados pelo sistema deverão ser protegidos. Além disso, a aplicação deverá ser construída de forma que, no futuro, seja acessível adaptar a estrutura para que dados de novas APIs sejam usados.
- **Usabilidade** — a aplicação web do sistema deverá ser intuitiva e simples de usar, de forma a dar uma experiência de uso agradável. Os dados estatísticos mais complexos deverão também ser explicados ao utilizador

- **Fiabilidade** — os modelos gerados deverão ser testados numa fase inicial de desenvolvimento do projeto, conforme explicado na secção 4.1. A aplicação Web deverá também ser fiável em alturas de maior tráfego (por exemplo, jogos importantes) e independente da execução de obtenção de dados e de aprendizagem dos modelos.
- **Desempenho —** os dados deverão ser disponibilizados ao utilizador com um tempo de resposta mínimo.
- **Suportabilidade —** a aplicação web deverá poder ser acessível em vários dispositivos eletrónicos (computador, *tablet*, telemóvel) para que um observador consiga ter um suporte tecnológico sempre que possível.

4.2 Design de Alto Nível

Nesta secção são identificados os componentes idealizados para a aplicação a desenvolver, no sentido de responder ao problema referido anteriormente na secção 1.2.

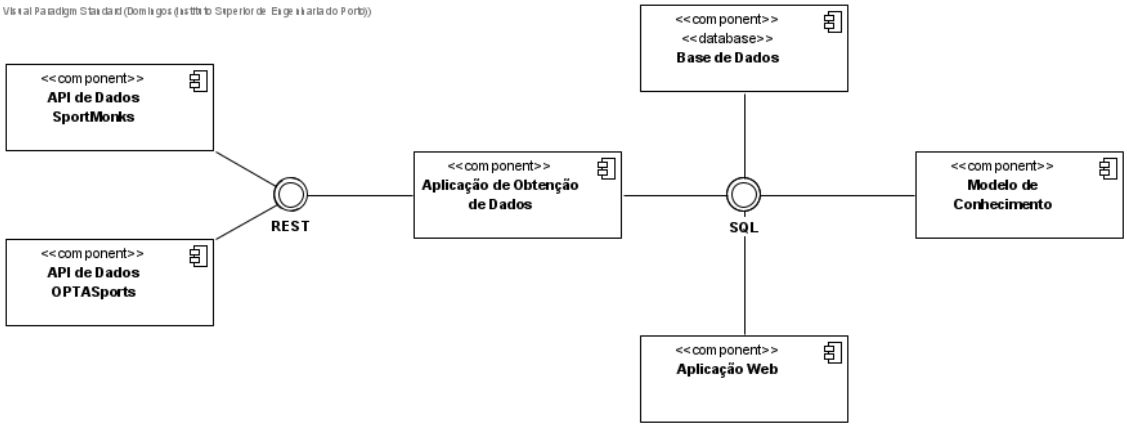


Figura 29 – Diagrama de Componentes de Alto Nível Proposto

No diagrama da Figura 29, identificam-se seis componentes, sendo estes:

- API de Dados SportMonks e API de Dados OPTASports — Representam os dados estatísticos brutos a obter. Serão acedidos usando uma API REST.
- Aplicação de Obtenção de Dados — Uma aplicação a desenvolver com o objetivo de aceder aos dados, limpá-los e gravá-los.
- Aplicação Web — Frontend do sistema onde o utilizador final poderá analisar os dados e tomar decisões.
- Base de Dados — Componente que irá conter toda a informação da aplicação, desde os dados limpos até ao modelo de conhecimento criado.
- Modelo de Conhecimento — Aplicação a desenvolver com recurso às técnicas de machine learning referidas na secção 2.3 de forma a interpretar os dados e obter novo conhecimento. Será desenvolvida em R.

4.2.1 Tecnologias Escolhidas

Com vista a responder às necessidades da solução, as seguintes tecnologias foram escolhidas para desenvolver os componentes do sistema:

- R — Para desenvolver o modelo de conhecimento a ser utilizado pelo sistema e para aplicar os algoritmos identificados na secção 2.3.
- .NET — Para desenvolver a aplicação Web de frontend e para a aplicação de obtenção de dados.
- SQL — Para armazenar os dados obtidos e calculados pelo modelo

4.3 Modelo de Domínio

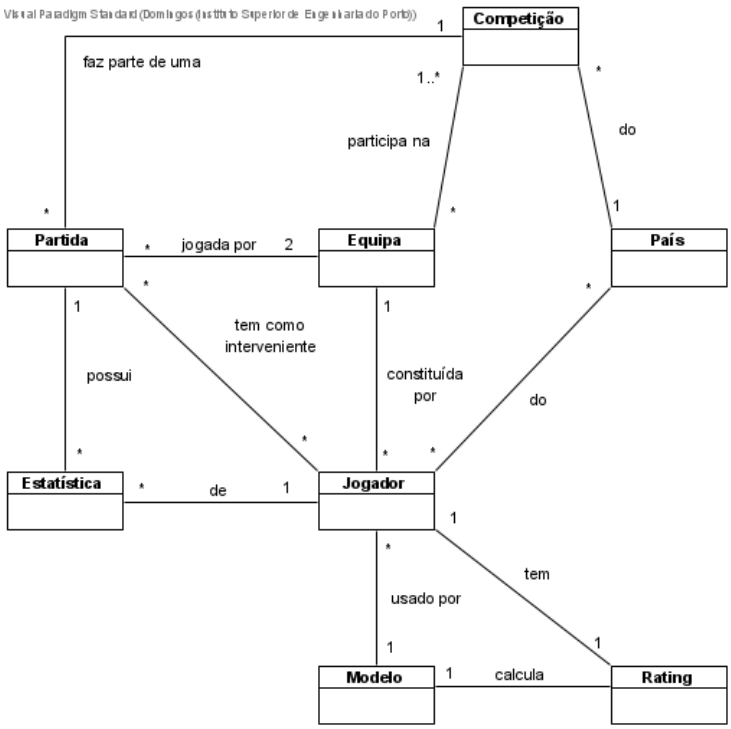


Figura 30 – Modelo de Domínio da Solução

O modelo de domínio representado na figura 33 consiste na ilustração das entidades e conceitos do problema e a forma como se relacionam entre si. De forma sucinta, a explicação para cada entidade:

- Jogador — Parte integral do sistema, é a entidade que será o alvo da utilização do observador
- Estatística — Representa todas as estatísticas que o sistema irá obter das base de dados.

- Partida — Elemento que consiste nos jogos aos quais as estatísticas dizem respeito.
- Equipa — Na sua essência é um conjunto de jogadores que joga partidas e participa em competições.
- Competição — Consiste em jogos por equipas de um determinado país ou região
- Modelo — Algoritmos e funções que usam dados de jogadores e estatísticas para calcular um rating
- Rating — Calculado a partir de um modelo que é associado a um jogador

5 Caso de Estudo

O caso de estudo, bem como o seu desenvolvimento do caso de estudo, serão apresentados nesta secção. O repositório de dados que alimenta os modelos criados será apresentado, com uma breve explicação sobre os dados presentes.

Tendo em conta que a abordagem utilizada foi a de CRISP-DM (secção 1.4), este capítulo irá ter subcapítulos referentes aos seis passos indicados anteriormente: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* e *Deployment*.

5.1 Dados em estudo

Os dados em estudo serão fornecidos pelo website FBRef.com, devido à sua grande especificidade a nível de estatísticas individuais. A base de dados será explicada detalhadamente na próxima secção.

5.1.1 FBRef.com

O website FBRef.com é um dos projetos relacionados com estatísticas de desporto da organização *Sports Reference*, que desde 2000 foca-se na publicação e edição de conteúdos em sites dedicados a vários desportos, nomeadamente às quatro *major leagues* norte-americanas, ao futebol americano e basquetebol universitários e ao futebol, com destaque para as cinco principais ligas nacionais da Europa, às três competições europeias e aos campeonatos do mundo e da europa de seleções, em que além de apresentar estatísticas simples como golos ou assistências, também contém estatísticas avançadas como passes rasteiros, transportes progressivos de bola ou pressões aplicadas no último terço do campo, sendo que a extensão temporal dos dados remonta à época 2017/2018. O *website* tem como fornecedor de dados para os jogos de futebol a StatsBomb (SportsReference, 2021).

Cada partida no *website*, dependendo da competição da qual faz parte, contém dados que vão desde os básicos, como data, local e resultado final, até aos mais avançados. A Figura 31 ilustra os dados disponibilizados em partidas com esses dados extra. No exemplo demonstrado, é possível ver que a plataforma permite escolher entre seis tabelas para ações de jogadores de campo, e uma sétima tabela para ações de guarda-redes:

Porto Player Stats Share & Export ▾ Glossary

Summary | Passing | Pass Types | Defensive Actions | Possession | Miscellaneous Stats

						Performance												Expected			SCA		Passes				Carries		Dribbles	
Player	#	Nation	Pos	Age	Min	Gls	Ast	PK	PKatt	Sh	SoT	CrdY	CrdR	Touches	Press	Tkl	Int	Blocks	xG	npxG	xA	SCA	GCA	Cmp	Att	Cmp%	Prog	Carries	Prog	Succ	Att
Mehdi Taremi	9	IRN	FW	28-214	90	1	0	0	0	2	2	0	0	24	17	0	0	1	0.6	0.6	0.0	1	0	8	17	47.1	0	13	1	0	0
Moussa Marega	11	MLI	FW	29-309	65	1	0	0	0	1	1	0	0	21	7	1	2	2	0.3	0.3	0.0	2	0	4	8	50.0	0	13	1	0	0
Marko Grujić	16	SRB	FW	24-310	25	0	0	0	0	1	0	0	0	17	1	0	1	1	0.0	0.0	0.0	0	0	9	14	64.3	2	8	1	0	0
Otávio	25	POR	LM	26-008	56	0	0	0	0	0	0	0	0	32	12	1	0	0	0.0	0.0	0.0	0	0	20	29	69.0	3	20	2	1	1
Luis Fernando Díaz	7	COL	LM	24-035	34	0	0	0	0	0	0	0	0	18	2	0	0	0	0.0	0.0	0.0	0	0	10	11	90.9	0	13	2	0	2
Jesús Corona	17	MEX	RM	28-042	89	0	0	0	0	0	0	0	0	35	13	2	2	2	0.0	0.0	0.0	2	1	16	24	66.7	1	20	3	2	5
Loum Ndiaye	6	SEN	RM	24-049	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.0	0.0	0.0	0	0	0	0		0	0	0	0	0
Mateus Uribe	8	COL	DM	29-333	90	0	0	0	0	0	0	0	0	31	17	2	3	1	0.0	0.0	0.0	1	0	20	25	80.0	0	17	2	0	0
Sérgio Oliveira	27	POR	DM	28-260	89	0	0	0	0	4	2	0	0	52	17	2	1	3	0.3	0.3	0.0	1	0	20	41	48.8	1	22	4	0	0
Francisco Conceição	85	POR	DM	18-065	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0.0	0.0	0.0	0	0	0	0		0	0	0	0	0
Zaidu Sanusi	12	NGA	LB	23-249	90	0	0	0	0	0	0	0	0	44	11	3	1	3	0.0	0.0	0.0	2	0	23	33	69.7	1	25	3	2	4
Pepe	3	POR	CB	37-357	90	0	0	0	0	0	0	0	0	38	7	2	0	1	0.0	0.0	0.0	0	0	24	34	70.6	0	20	0	0	0
Chancel Mbemba	19	COD	CB	26-193	90	0	0	0	0	0	0	0	0	36	4	0	1	2	0.0	0.0	0.0	0	0	24	27	88.9	1	20	0	0	0
Wilson Manafá	18	POR	RB	26-209	90	0	1	0	0	0	0	0	0	47	12	3	2	1	0.0	0.0	0.5	2	1	34	39	87.2	0	27	3	2	3
Agustín Marchesín	1	ARG	GK	32-338	90	0	0	0	0	0	0	0	0	50	0	0	0	0	0.0	0.0	0.0	0	0	28	47	59.6	0	21	0	0	0
15 Players					**990**	**2**	**1**	**0**	**0**	**8**	**5**	**0**	**0**	**445**	**121**	**16**	**13**	**17**	**1.2**	**1.2**	**0.6**	**11**	**2**	**240**	**349**	**68.8**	**9**	**239**	**22**	**7**	**15**

Porto Goalkeeper Stats Share & Export ▾ Glossary

				Shot Stopping					Launched			Passes				Goal Kicks			Crosses			Sweeper	
Player	Nation	Age	Min	SoTA	GA	Saves	Save%	PSxG	Cmp	Att	Cmp%	Att	Thr	Launch%	AvgLen	Att	Launch%	AvgLen	Opp	Stp	Stp%	#OPA	AvgDist
Agustín Marchesín	ARG	32-338	90	5	1	4	80.0	1.7	13	30	43.3	42	5	66.7	47.8	5	40.0	33.2	10	0	0.0	1	16.2

Figura 31 – Exemplo de uma tabela de dados de um jogo (FBRef.com, 2021)

- *Summary* – A primeira tabela contém dados que resumem a performance de um jogador na partida, com destaque para disponibilização de estatísticas como xG (golos esperados), xA (assistências esperadas) ou ações que tenham resultado em remates (SCA de *shot creating actions*) ou em golos (GCA de *goal creating actions*).
- *Passing* e *Pass Types* – Estas duas tabelas listam estatísticas relacionadas com a ação de passe de cada jogar da equipa. A primeira tabela contém informação sobre o sucesso, a qualidade e a finalidade do passe, isto é, se foi completado, se foi um passe valioso e para onde foi efetuado. A segunda tabela diz respeito à forma como os passes foram feitos, se foi numa situação de bola corrida ou de bola parada, se foi um passe rasteiro ou alto e com que parte do corpo a bola foi passada.
- *Defensive Actions* – Como o nome indica, a tabela contém dados sobre as ações defensivas de cada jogador, como desarmes e interceções, pressões aplicadas e duelos disputados.
- *Possession* – Dados que permitem perceber como um jogador atuou em relação aos seus dribles, transportes de bola, receções e em que zonas do campo o jogador tocou na bola.
- *Miscellaneous* – A última tabela diz respeito a dados que não se enquadram diretamente numa das outras cinco tabelas, como o registo disciplinar, as faltas cometidas e sofridas e os duelos aéreos de cada jogador.

- *Goalkeeper Stats* – As estatísticas de guarda-redes consistem numa única tabela onde se pode visualizar a performance a nível de defesa de remates, lançamentos e pontapés longos, passes, pontapés de baliza, cruzamentos que o guarda-redes teve de defender e agarrar e ações fora da sua grande área.

Além disso, a plataforma também tem páginas dedicadas a cada equipa, competição, jogador e país, com informações de cada uma das entidades e, caso se aplique, dados agregados das partidas efetuadas.

5.2 Business Understanding

Considera-se que a secção 2 referente ao capítulo de Contexto e Estado de Arte como a parte da implementação referente à investigação do negócio e das suas nuances, de forma que se alcance o objetivo pretendido. No entanto, um resumo dos principais pontos discutidos é apresentado na lista abaixo:

- O futebol tem evoluído ao longo das décadas de um desporto sem grande envolvimento científico para um desporto em que cada vez mais se utiliza ferramentas tecnológicas para auxiliar na tomada de decisões.
- O aumento significativo dos encaixes e encargos monetários fez com que os clubes passassem a ser geridos como empresas, com obrigações financeiras e a necessidade de melhorar o seu produto, que neste caso é a equipa de jogadores de futebol, seja com melhores profissionais ou com melhores ferramentas de apoio.
- A análise estatística de um jogo de futebol ajuda a conciliar a experiência na observação de pontos forte e fracos da equipa técnica de um clube com um suporte científico que permita detetar padrões ou dados que possam estar ocultos.

A finalidade do desenvolvimento desta ferramenta prende-se com o cálculo de um *rating* de qualidade de um jogador, aplicando algoritmos e ferramentas de *data mining* e *machine learning*, e com base no estudo referido na secção 2.5.1.1, um modelo de previsão de resultados de futebol com duas variáveis objetivo (vitória ou não vitória) será criado, retirando desse modelo os pesos das variáveis usadas para obter uma expressão matemática que calcule o rating de qualidade do jogador.

5.3 Data Understanding

Com base nos dados obtidos retirados do *website* referido na secção 5.1.1, o armazenamento pode ser dividido em três partes. A primeira, ilustrada na Figura 32, diz respeito à parte da base de dados que contém os dados informativos:

- Country — Registos de países existentes na plataforma (/en/countries/). É também inserido um registo correspondente à UEFA, mesmo não sendo um país, devido à forma como a base de dados se encontra desenhada.
- Competition — Tabela onde são armazenadas as competições, neste caso as cinco principais ligas nacionais europeias e as duas principais competições europeias
- Season — Contém registos de todas as épocas obtidas de cada competição (/en/comps/9/history/Premier-League-Seasons).
- Match — Registo de todos os jogos disputados em cada época retirada do *website* (/en/comps/9/1631/schedule/2017-2018-Premier-League-Scores-and-Fixtures). Esta tabela irá ter ligações com as tabelas das estatística, nomeadamente a tabela TeamStatsForMatch.
- Team — Tabela com as equipas que participaram nos jogos armazenados
- Player — Registos dos jogadores de cada equipa com pelo menos uma participação nas épocas retiradas.
- TeamInSeason — Uma equipa participa em várias épocas, e uma época tem várias equipas, pelo que se torna necessário criar uma tabela de apoio resultante desta ligação *many-to-many*. Esta tabela regista a equipa e a época, e efetua uma ligação com a entidade Match para que um jogo tenha a informação da equipa caseira e forasteira.
- PlayerInTeamInSeason — Da mesma forma que uma equipa participa em várias épocas, um jogador pode ter várias equipas ao longo das épocas da sua carreira, pelo que se tornou necessário criar uma tabela auxiliar com o registo de cada jogador que participou numa equipa em cada época. Tal como a tabela Match, a entidade PlayerInTeamInSeason irá ser referenciada nas tabelas referentes aos dados estatísticos, nomeadamente a tabela PlayerStatsFromMatch.

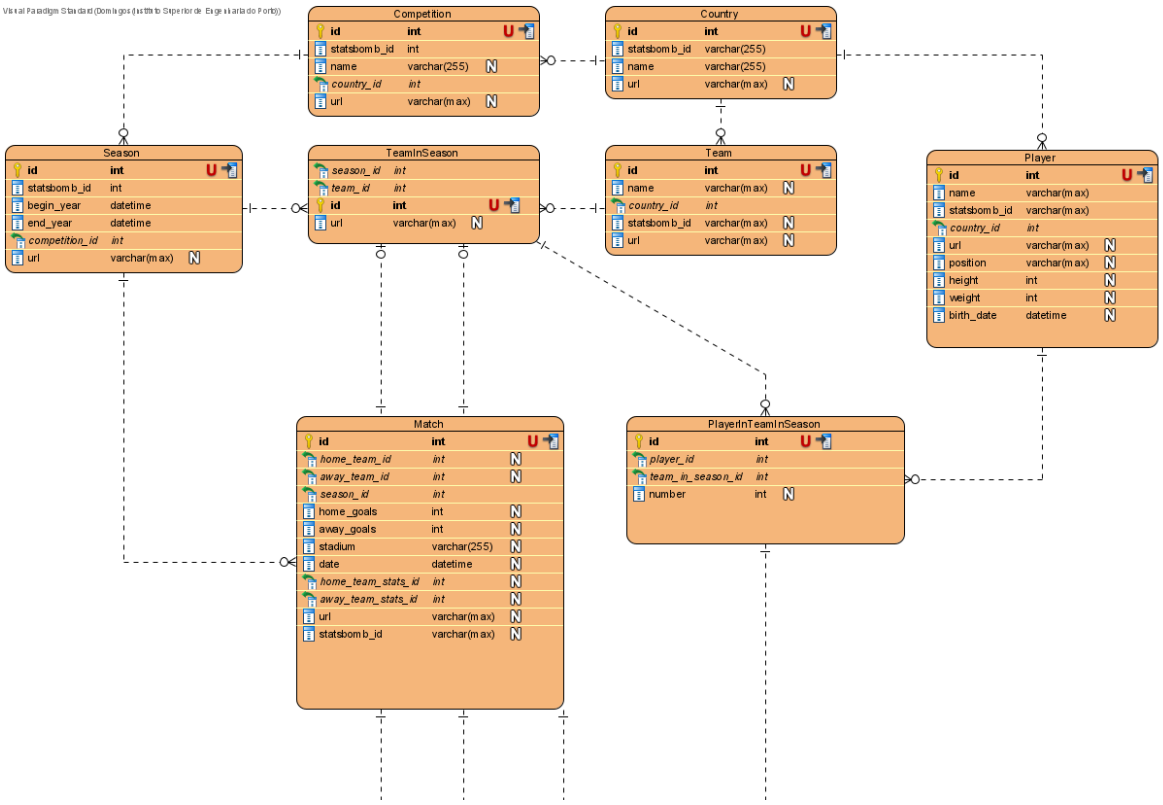


Figura 32 – Modelo Relacional do armazenamento de informação

A Figura 33 representa a segunda parte da estrutura da base de dados que contém os dados estatísticos de todas as partidas obtidas. A tabela de relação entre os nomes dos dados demonstrados no fbref.com e os existentes na base de dados pode ser consultada no Anexo 1:

- TeamStatsFromMatch — Entidade que liga as estatísticas dos jogadores às partidas guardadas na base de dados. A tabela contém um atributo que indica se a equipa jogou em casa ou fora de casa.
- PlayerStatsFromMatch — Registos que contêm as estatísticas gerais de cada jogador que tenha participado num jogo.
- DefensiveStatsFromPlayerStatsFromMatch — Dados estatísticos relacionados com a performance defensiva do jogador, tendo como base a tabela *Defensive Actions* e dados avulsos da tabela *Miscellaneous*.
- OffensiveStatsFromPlayerStatsFromMatch — Dados estatísticos relacionados com a performance ofensiva do jogador, tendo como base a tabela *Possession* e dados avulsos das tabela *Summary* e *Miscellaneous*.
- PassingStatsFromPlayerStatsFromMatch e PassingTypeStatsFromPlayerStatsFromMatch — Dados estatísticos relacionados com a performance de passe e com as suas características, com base nas tabelas *Passing* e *Pass Types*.

A divisão da base de dados em duas zonas permite analisar o que cada tabela representará para a solução final e como melhor os tratar na próxima fase da implementação. A primeira zona representa informação cujo tratamento analítico será reduzido, visto que são dados

informativos das equipas, competições e jogadores. Serão importantes, contudo, na vertente de demonstração ao utilizador final, bem como na criação de casos de estudo relevantes com base nos modelos criados.

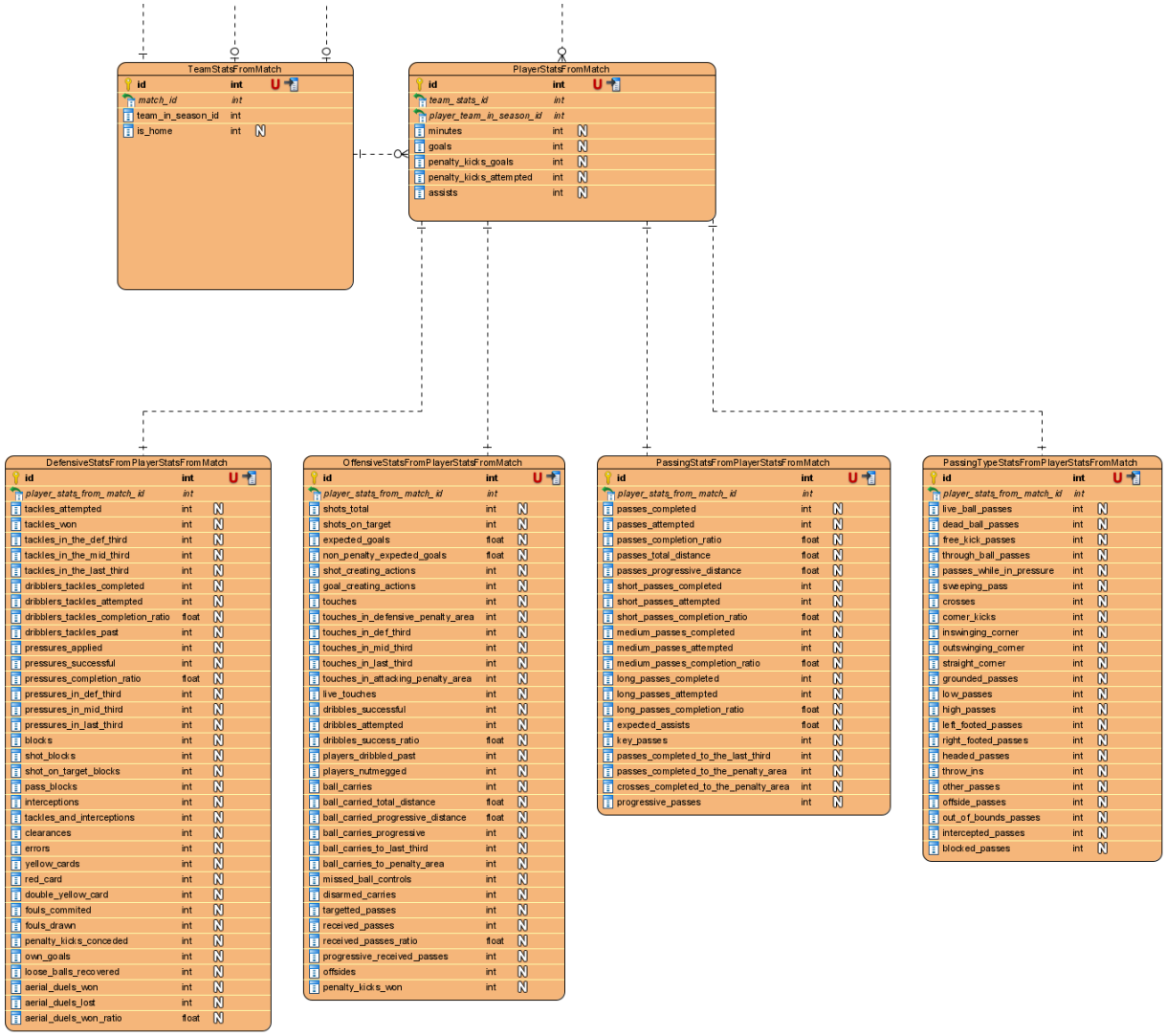


Figura 33 – Modelo Relacional do armazenamento de estatísticas

Os dados estatísticos das partidas de futebol estão armazenados na segunda zona, tornando esta zona importante para futuras análises. Tendo em conta que o objetivo inicial é prever o desfecho de jogos com base na exibição estatística de cada equipa, é necessário tratar os dados inicialmente, de forma a retirar variáveis que possam estar correlacionadas e que iriam poluir o conjunto de dados. Antes de aplicar métodos matemáticos, algumas variáveis podem ser retiradas imediatamente:

- Um jogo de futebol é ganho por se marcar mais golos que o adversário, logo qualquer tipo de ação que resulte em golo é intuitivamente mais valioso que outras ações. Por isso, as variáveis *goals*, *own_goals*, *assists*, *penalty_kicks_goals*,

penalty_kicks_conceded e *goal_creating_actions* são retiradas devido ao seu peso enorme na determinação de um resultado. A sua presença na aplicação de algoritmos iria tornar praticamente irrelevantes as restantes estatísticas. No entanto, estas estatísticas irão ser consideradas no cálculo do *rating* do jogador, pelo que um modelo auxiliar será criado com vista a identificar o peso real destas estatísticas.

- As estatísticas *penalty_kicks_attempted* e *penalty_kicks_won* são virtualmente iguais, visto que para uma equipa bater um penalti, precisa de sofrer uma falta dentro da grande área. Com isso, a estatística que se irá manter será a estatística de penaltis ganhos, já que como o penalti é o lance com maior percentagem de resultar em golo, com 83.14% dos penaltis a serem bem sucedidos (Ch and ler, 2021), a relação entre um golo de penalti e um penalti batido num jogador é elevadíssima. No entanto, o facto de um jogador ganhar o penalti não significa que será ele a marcar, tendo em conta que todas as equipas têm os seus batedores de penaltis definidos.
- As estatísticas acumuladas de ações independentes do seu sucesso, como passes (*passes_attempted*) ou dribles (*dribbles_attempted*), serão removidas de forma a criar novas estatísticas correspondentes a essas ações, mas falhadas, como passes falhados (*passes_failed*), efetuando uma simples subtração entre o número total de ações e o número de ações com sucesso. Com isto, permite-nos diferenciar as ações com e sem sucesso.
- Todas as estatísticas que representam um valor percentual (p.ex *passes_completion_ratio*) serão retiradas do conjunto de dados, visto que são dados relativos calculados a partir de outras estatísticas presentes.
- O modelo de dados tem estatísticas que correspondem à forma como o passe foi feito (*left_footed_passes*, *right_footed_passes*, *headed_passes*, *other_passes*) ou como o canto foi batido (*inswinging_corner*, *outswinging_corner*, *straight_corner*) são importantes para determinar o perfil do jogador, como a sua ambidesteridade, mas irrelevantes para obter a qualidade do jogador. Existe também a estatística de dribles pelo meio das pernas (*players_nutmegged*) que tem as mesmas características das estatísticas referidas anteriormente: relevantes para identificar o perfil do jogador, mas insignificante para perceber a qualidade do jogador.
- Existem duas estatísticas que no contexto global de uma equipa acabam por ter valores praticamente iguais: passes completados (passes_completed) e passes recebidos (received_passes). Quase sempre um passe completado por parte do Jogador A irá resultar num passe recebido pelo Jogador B, pelo que a sua correlação é grande. Com vista a não poluir os dados a serem usados pelos algoritmos, a estatística de passes recebidos irá ser retirada.
- As estatísticas *expected_goals* e *expected_assists* são retiradas visto serem estatísticas já calculadas com base em algoritmos que detetam a probabilidade de cada remate ou passe resultar num golo ou assistência (FBRef.com, 2020). A estatística irá ser apresentada no perfil de cada jogador, mas apenas como dado auxiliar.
- Por fim, as tabelas referentes às estatísticas de Guarda-Redes não serão importadas, bem como os dados estatísticos de "campo" por parte desses jogadores não serão

adicionados ao modelo a ser analisado. No entanto, a informação será carregada para implementações futuras com esses dados.

Finalmente, a terceira parte encontra-se demonstrada na Figura 34, em que se pode verificar a presença de quatro tabelas que auxiliaram o processo de modelação:

- AggregatedMatch — Tabela onde as estatísticas agregadas das partidas irão ser inseridas, de forma a obter rapidamente os dados, estando ligada aos registos presentes na tabela Match. Contém 132 colunas, pelo que se referencia o leitor ao Anexo 2 para visualizar a tabela completa.
- PlayerRatioHistory — Informação sobre os valores calculados para cada jogador num determinado jogo. Com esta decisão de registar para cada partida, será possível aplicar um cálculo ponderado, dando mais valor às exibições recentes. Os registos do *ratio* com e sem golo irão ser registados, para uma melhor comparação de modelos na etapa de avaliação. A tabela tem ligações com a tabela PlayerTeamInSeason e Match.
- PlayerRatio — Tabela com os registos para cada jogador do *ratio* acumulado e da média móvel das suas performances, para ser mais imediato o acesso a estes dados. Referencia em todos os registos o jogador a qual a informação sobre os *ratios* pertence, neste caso a um registo na tabela Player.
- VariableWeight — Registo dos pesos das variáveis ao longo to tempo a partir da coluna *date*, com vista a se poder ir registando as diferentes melhorias.

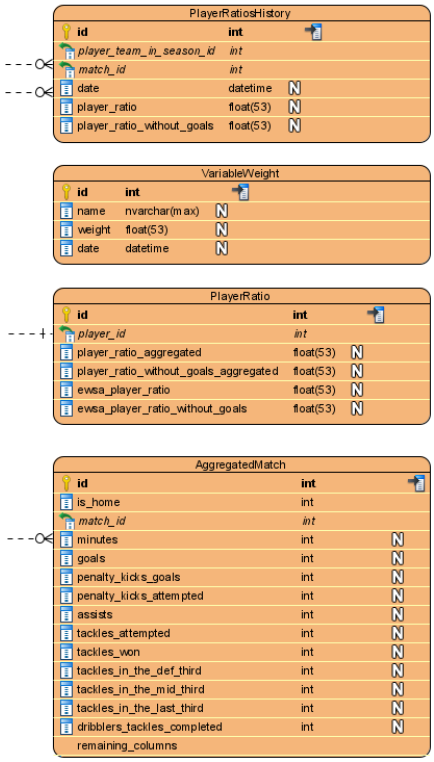


Figura 34 – Modelo Relacional dos dados do modelo

5.3.1 Análise preliminar dos dados

A informação sobre os dados obtidos a partir do website FBRef.com está listada nas duas tabelas presentes nesta secção. A Tabela 4 contém o número de registos em cada tabela cuja informação seja meramente ilustrativa ou para agrupar, em que se pode perceber que serão calculados os ratings para 6626 jogadores que participaram em 8175 jogos ao serviço de 208 equipas em 7 competições distintas, percorrendo quatro épocas distintas desde 2017/2018 para as competições nacionais, e três épocas desde 2018/2019 para as competições europeias.

Tabela 4 – Quantidade de dados presentes na base de dados

Dados	Quantidade
Equipas	208
Jogadores	6626
Competições	7
Jogos	8175

Os dados relativos a cada jogo e aos intervenientes e estatísticas geradas estão disponibilizados na Tabela 5. O conjunto de dados a ser utilizado no modelo de previsão consistirá nas estatísticas agrupadas dos jogadores numa partida, com 232653 registos, para 8175 jogos, num total de 16350 linhas, com duas por partida, uma linha para a equipa visitada e uma linha para a equipa visitante. 6134 desses registos coincidem com o de uma equipa vitoriosa, não sendo metade dos jogos em análise devido a situações em que um jogo termina empatado. Alguns dos jogos não foram considerados devido ao facto de terem sido cancelados, como por exemplo os jogos a partir da jornada 29 da época 2019/20 da Ligue 1 que foram anulados devido ao COVID-19. De forma a se fazer uma comparação entre diferentes contextos competitivos, oito modelos vão ser analisados: sete para cada competição e um com todos os dados indiscriminados. Os modelos com mais jogos são os da La Liga e da Premier League, ambos com 1520, e o com menos jogos é o da UEFA Champions League. Já a nível de jogadores, o modelo com mais jogadores distintos é o da UEFA Europa League, devido a serem a competição com mais equipas distintas. No entanto, é o modelo da Serie A que possui mais estatísticas de jogadores diferentes, o que pode ser explicado por serem um dos campeonatos em que o número de suplentes por jogo é de 12, permitindo ao treinador ter mais hipóteses de substituição, o que se traduz em mais jogadores a participar.

Tabela 5 – Quantidade de dados presentes relacionados com jogos na base de dados

Modelo	Dados Jogos	Dados de	Estatísticas da	Estatísticas da	Dados de	Estatísticas de

		Época de Equipa	Equipa numa Partida	Equipa numa Partida Vitoriosa	Época de Jogador	Jogador numa Partida
Bundesliga	1224	72	2448	916	1962	35032
La Liga	1520	80	3040	1101	2271	43744
Ligue 1	1419	80	2838	1045	2244	40304
Premier League	1520	80	3040	1173	2091	41939
Serie A	1519	80	3038	1136	2347	43754
UEFA Champions League	369	96	738	290	2130	10541
UEFA Europa League	604	168	1208	473	3799	17339
Total	8175	656	16350	6134	16844	232653

A Figura 35 apresenta um gráfico com 18 jogadores, nos quais estão presentes os 10 jogadores com mais assistências e os 10 jogadores com mais golos, de forma a representar o seu impacto ofensivo. Cada jogador tem no seu ponto respetivo a soma das duas estatísticas.

Três jogadores destacam-se dos restantes: Lionel Messi é o jogador com maior valor acumulado de estatísticas, estando em segundo na lista de melhores marcadores e melhores assistentes, o que revela um grande impacto no jogo ofensivo da equipa. Robert Lewandowski é o melhor marcador na base de dados com 154 golos, e Thomas Muller tem 67 assistências, o valor mais alto registado.

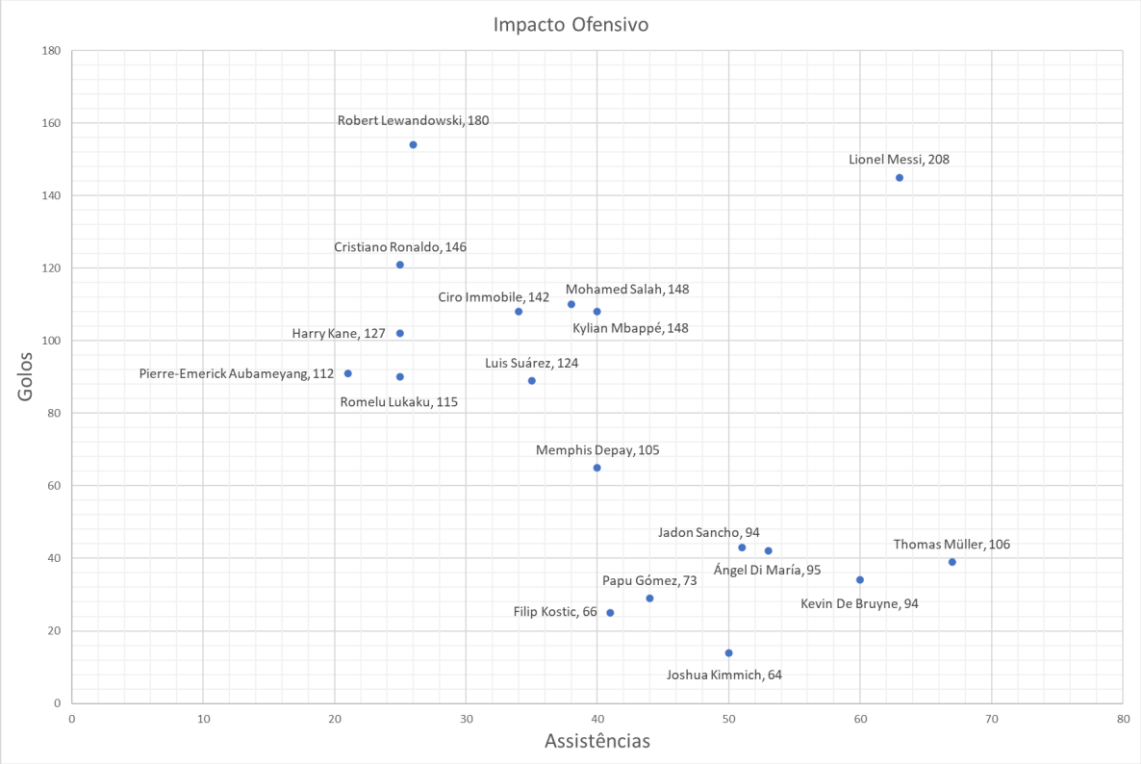


Figura 35 – Dez melhores marcadores e dez melhores assistentes

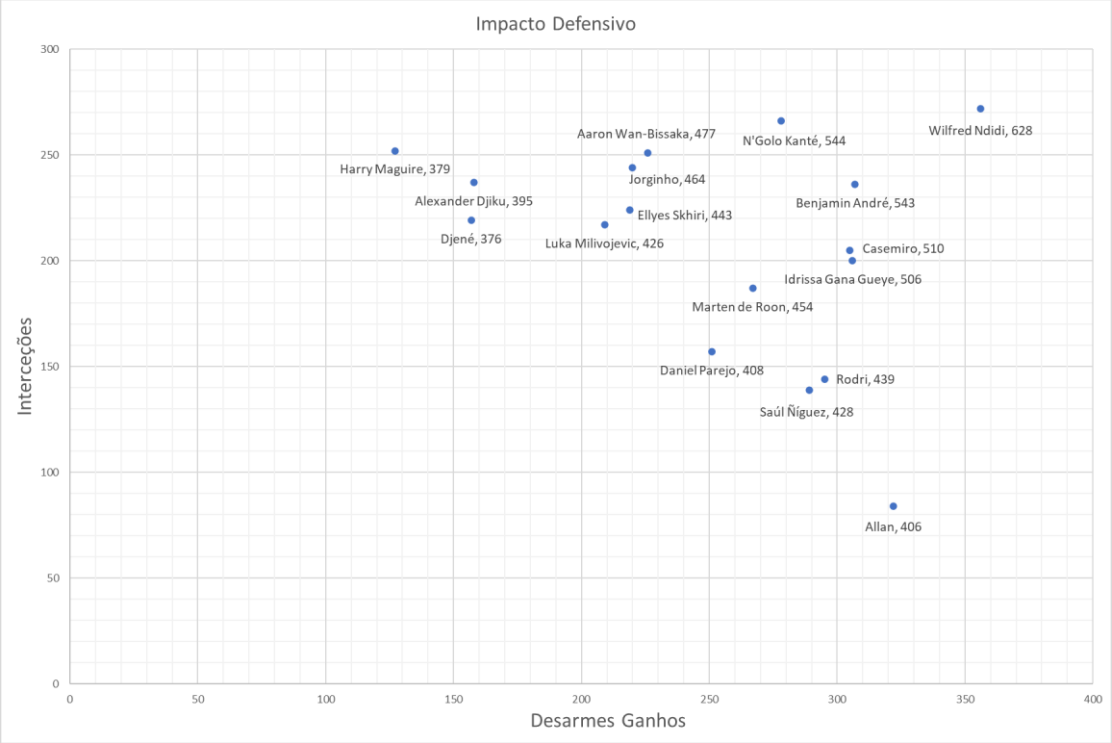


Figura 36 – Dez jogadores com mais desarmes e dez com mais interceções

A Figura 36 tem 17 jogadores que pertencem ao top 10 de desarmes ganhos ou ao top 10 de interceções ganhas. Os jogadores acabam por estar muito próximos uns dos outros, mas o destaque é Wilfred Ndidi, o jogador com mais interceções e desarmes ganhos na lista, mostrando ser dos jogadores com maior impacto defensivo no jogo.

5.4 Data Preparation

De modo que se possa começar a tratar os dados estatísticos, é preciso obtê-los a partir do website fbref.com. Para essa tarefa, foram usadas principalmente duas bibliotecas: *xml2*, usada para obter o documento HTML de uma determinada página web (Wickham et al., 2020, p. 2), e *rvest*, usada para pesquisar dentro de páginas HTML usando seletores CSS (Wickham and RStudio, 2021). Visto que não é uma API a devolver os dados limpos de qualquer formatação ou divisão de conteúdo devido à construção do website, alguns cuidados foram tidos em conta:

- Todos os links seguem o mesmo formato, nomeadamente domínio/idioma_de_apresentação/nome_da_entidade/identificador_da_entidade/restantes_parâmetros, como o exemplo do endereço para o jogo entre Porto e Juventus para a Liga dos Campeões (fbref.com/en/matches/dea17ee3/Porto-Juventus-February-17-2021-Champions-League), que tem a entidade a que diz respeito, o seu identificador e o nome do jogo, com este último parâmetro a ser apenas auxiliar para uma melhor navegação por parte do utilizador.

- Como tal, obter o endereço para cada entidade é uma tarefa que consegue ser resolvida de forma simples, excetuando perfis das competições e de equipas que estejam a participar na época atual. Exemplificando com o link para as estatísticas da época atual (2021-2022) do Manchester United (fbref.com/en/squads/19538871/Manchester-United-Stats), não existe um identificador que permita perceber qual a época a que diz respeito, ao contrário das épocas anteriores (fbref.com/en/squads/19538871/2019-2020/Manchester-United-Stats), em que a partir do endereço consegue-se retirar a época a que diz respeito. Isto implica um cuidado acrescido a retirar os dados.

Após esta preparação, os dados foram retirados seguindo a ordem abaixo listada:

1. Extrair todos os países presentes na base de dados do fbref.com e as sete competições a serem analisadas (Premier League, Ligue 1, Serie A, Bundesliga, La Liga, UEFA Champions League e UEFA Europa League).
2. Para as competições nacionais, retirar todas as épocas desde 2017/2018. Para as competições europeias, a primeira época passa para 2018/2019.
3. Percorrer a lista de jogos presente no endereço de cada época, já que contém todos os jogos realizados no decorrer da mesma.
4. O conteúdo de um endereço de uma partida tem informação referente à data do jogo, local, resultado e equipas participantes, bem como as dozes tabelas de estatísticas, divididas em seis por equipa. No excerto de Código 1, encontra-se a função que trata da importação das informações do jogo. O HTML da página é interpretado, usando diversos seletores e manipulações de dados, validando sempre preliminarmente se o registo já existe na base de dados. Aliás, esta validação é feita sempre que uma inserção é realizada.
5. No caso de ser uma instância de jogo que está por ser registada, a importação de dados avança para a criação de elementos referentes às estatísticas gerais da equipa, de cada jogador presente no jogo e da sua performance, efetuando primeiro uma importação das estatísticas mais gerais (golos e assistências para a tabela PlayerStatsFromMatch) e de seguida a importação das estatísticas mais aprofundadas, conforme se vê no Código 2.

Com a base de dados preenchida, a análise aos dados pode ser iniciada. Para esse efeito, a *query* presente no Código 3 é executada na base de dados de forma a retirar todos os dados individuais dos jogadores em todos os jogos, bem como com a informação adicional do jogo, como o nome da competição (*competition_name*), a época da partida (representada pelo início e final de época), o identificador do jogo, o número de golos de cada equipa da partida, da data em que se realizou, da informação de que se a estatística do jogador corresponde à equipa caseira e de identificadores de base de dados. As condições para a obtenção de dados, além das junções de base de dados de forma a obter toda a informação, consistem na remoção de qualquer registo de um guarda-redes e na seleção de jogos correspondentes a uma competição, identificada pelo texto "*COMPETITION_NAME*". Essa variável será

substituída à medida que a *query* será corrida para cada competição existente na base de dados, informação carregada anteriormente no algoritmo de carregamento de dados.

```
export_matches_to_database <- function(match_link, season_id) {
    #obter informação do jogo
    cat("Exporting match ", match_link, "\n", sep="")
    match_link_stripped <- unlist(strsplit(match_link, '/'))
    match_link_stripped <- match_link_stripped[lapply(match_link_stripped, function(x) x != "") > 0]
    statsbomb_id <- match_link_stripped[5]

    sql_expression_to_get_match <- paste("SELECT * FROM Match where statsbomb_id like '",
statsbomb_id, "'", sep="")
    resultset <- sqlQuery(connection_to_database, sql_expression_to_get_match)
    #se o jogo já tiver sido criado, obtém-se informação da partida
    if (length(resultset$id) > 0) {
        match_id <- resultset$id
        home_team_id <- resultset$home_team_id
        away_team_id <- resultset$away_team_id
        home_goals <- resultset$home_goals
        away_goals <- resultset$away_goals
        match_html <- NULL
    } else {
        match_html <- xml2::read_html(match_link)
        goals <- match_html %>% rvest::html_nodes(".scores > .score") %>% rvest::html_text()
        date_attributes <- unlist(match_html %>% rvest::html_nodes('span.venuetime') %>%
rvest::html_attrs())
        #a data obtém-se juntando o dia e a hora presentes no html
        datetime <- paste(date_attributes['data-venue-date'], date_attributes['data-venue-time'])
        #tratar das equipas
        teams_in_match <- match_html %>% rvest::html_nodes("div[itemprop='performer'] > strong > a")
%>% rvest::html_attr('href')
        teams_in_match_link <- paste(base_url, teams_in_match, sep ="")
        #exportar equipas para a bd. obtem equipa na epoca
        teams_in_season <- lapply(teams_in_match_link, export_team_to_database, season_id =
season_id) # nolint
        home_team_id <- teams_in_season[1]
        away_team_id <- teams_in_season[2]
        home_goals <- ifelse(!is.na(as.numeric(goals[1])), goals[1], 'NULL')
        away_goals <- ifelse(!is.na(as.numeric(goals[2])), goals[2], 'NULL')
        #inserir jogo na bd
        sql_expression_to_insert_match <- paste("INSERT INTO Match (home_team_id, away_team_id,
season_id, home_goals, away_goals, date, statsbomb_id, url) OUTPUT inserted.id VALUES (",
home_team_id,",", away_team_id,",", season_id, ",", home_goals,",", away_goals,",'",
datetime,"','",statsbomb_id, "','", match_link,"')", sep="")
        resultset <- sqlQuery(connection_to_database, sql_expression_to_insert_match)
        match_id <- resultset$id
    }
…
}
```

Código 1 – Função iterativa para obter todos os jogos das competições na base de dados

```
export_stat_for_table <- function(table_columns, stat_row, player_stats_from_match_id) {
    #obter tabela na qual os dados serão inseridos
    table_to_insert <- table_columns$table_in_football_learning[1]
    cat("Exporting      ",      table_to_insert,"      from      player_stats_from_match      ",
player_stats_from_match_id, "\n", sep="")
    sql_expression_to_get_player_stats_from_match     <-     paste("SELECT     id     FROM     ",
table_to_insert," where player_stats_from_match_id = ", player_stats_from_match_id,'',sep="")
    resultset                        <-                        sqlQuery(connection_to_database,
sql_expression_to_get_player_stats_from_match)
    #verificar se registo já existe
    if (length(resultset$id) > 0) {
        table_stats_id <- resultset$id
    } else {
        #obtem todas as colunas que serão inseridas
        columns_to_search <- paste(table_columns$column_name, sep=",", collapse=",")
        #valores a inserir na base-de-dados
        values_from_search  <-  paste(stat_row[table_columns$header_and_stats_name],  sep=",",
collapse=",")
        values_from_search <- gsub("NA", "NULL",values_from_search)
        #adicionar a variavel player_stats_from_match_id ao conjunto de colunas e valores
        columns_to_insert <- paste('player_stats_from_match_id', columns_to_search, sep=",")
        values_to_insert <- paste(player_stats_from_match_id, values_from_search, sep=",")
        sql_expression_to_insert_specific_stats <- paste("INSERT INTO ", table_to_insert," (",
columns_to_insert, ") OUTPUT inserted.id VALUES (", values_to_insert, ")", sep="")
        resultset <- sqlQuery(connection_to_database, sql_expression_to_insert_specific_stats)
        table_stats_id <- resultset$id
    }
    table_stats_id
}
```

Código 2 – Função de extração de estatísticas para cada tabela do FBRef.com

```
SELECT  Competition.name AS competition_name,
              YEAR(Season.begin_year)    AS    season_begin_year,    YEAR(Season.end_year)    AS
season_end_year,
              Match.id AS match_id, Match.home_goals, Match.away_goals, Match.date,
              Team.name AS team_name,
              TeamStatsFromMatch.is_home,
              PlayerStatsFromMatch.team_stats_id, PlayerStatsFromMatch.player_team_in_season_id,
PlayerStatsFromMatch.minutes,         ...         --         rest         of         statistics
DefensiveStatsFromPlayerStatsFromMatch.tackles_attempted,
DefensiveStatsFromPlayerStatsFromMatch.tackles_won,
DefensiveStatsFromPlayerStatsFromMatch.tackles_in_the_def_third,
DefensiveStatsFromPlayerStatsFromMatch.tackles_in_the_mid_third, ... -- rest of statistics
              OffensiveStatsFromPlayerStatsFromMatch.shots_total,
OffensiveStatsFromPlayerStatsFromMatch.shots_on_target,
OffensiveStatsFromPlayerStatsFromMatch.expected_goals, ... -- rest of statistics
              PassingStatsFromPlayerStatsFromMatch.passes_completed,
PassingStatsFromPlayerStatsFromMatch.passes_attempted,
PassingStatsFromPlayerStatsFromMatch.passes_completion_ratio,... -- rest of statistics
              PassingTypeStatsFromPlayerStatsFromMatch.live_ball_passes,
PassingTypeStatsFromPlayerStatsFromMatch.dead_ball_passes,
PassingTypeStatsFromPlayerStatsFromMatch.free_kick_passes,... -- rest of statistics
FROM   Match INNER JOIN
              Season ON Match.season_id = Season.id INNER JOIN
              Competition ON Season.competition_id = Competition.id INNER JOIN
              TeamStatsFromMatch   ON   Match.home_team_stats_id   =   TeamStatsFromMatch.id   OR
Match.away_team_stats_id  =  TeamStatsFromMatch.id  AND  Match.id  =  TeamStatsFromMatch.match_id
INNER JOIN
              TeamInSeason ON TeamStatsFromMatch.team_in_season_id = TeamInSeason.id INNER JOIN
              Team  ON  TeamInSeason.team_id  =  Team.id  INNER  JOIN  PlayerStatsFromMatch  ON
PlayerStatsFromMatch.team_stats_id = TeamStatsFromMatch.id INNER JOIN
              DefensiveStatsFromPlayerStatsFromMatch                                          ON
DefensiveStatsFromPlayerStatsFromMatch.player_stats_from_match_id    =    PlayerStatsFromMatch.id
INNER JOIN
              OffensiveStatsFromPlayerStatsFromMatch                                          ON
OffensiveStatsFromPlayerStatsFromMatch.player_stats_from_match_id    =    PlayerStatsFromMatch.id
INNER JOIN
              PassingStatsFromPlayerStatsFromMatch                                            ON
PassingStatsFromPlayerStatsFromMatch.player_stats_from_match_id  =  PlayerStatsFromMatch.id  INNER
JOIN
              PassingTypeStatsFromPlayerStatsFromMatch                                        ON
PassingTypeStatsFromPlayerStatsFromMatch.player_stats_from_match_id   =   PlayerStatsFromMatch.id
INNER JOIN
              PlayerInTeamInSeason  ON  TeamInSeason.id  =  PlayerInTeamInSeason.team_in_season_id
AND PlayerStatsFromMatch.player_team_in_season_id = PlayerInTeamInSeason.id INNER JOIN
              Player ON PlayerInTeamInSeason.player_id = Player.id
WHERE  (Competition.name LIKE 'COMPETITION_NAME') AND Player.position <> 'GK'
ORDER BY match_id DESC
```

Código 3 – *Query* de carregamento de dados

Posteriormente, os dados são divididos em conjuntos correspondentes a cada competição, bem como a criação de um que englobe todas as competições. Os dados são tratados seguindo os passos abaixo:

1. Para cada conjunto de dados, os valores numéricos são agregados numa única linha, seguindo a combinação *match_id* e *is_home*, que será sempre diferente.
2. A variável objetivo do modelo de previsão "*match_won*" é adicionada, que resulta de uma comparação simples entre os golos da equipa à qual a linha pertence e a equipa adversária.
3. As observações efetuadas na secção anterior são aplicadas, isto é, novas colunas são criadas, como a obtenção de passes falhados, e retiram-se outras colunas, como a coluna dos passes recebidos.
4. Por fim, esta informação é toda armazenada na tabela AggregatedMatches, referida na secção 5.2.

Com os dados devidamente carregados e tratados, o próximo passo consiste na verificação da correlação entre os dados, utilizando a função cor da biblioteca R *stats*, com a procura de relações com um coeficiente de significância maior que 0,90. Os resultados estão demonstrados na Tabela 6, com as colunas Estatística A e Estatística B, que indicam as variáveis comparadas, e o seu coeficiente de correlação. No Anexo 3, a mesma tabela encontra-se representada, mas com uma coluna adicional com a possível razão para haver uma correlação alta. Após uma análise dos resultados, concluiu-se que:

- Um jogador que é driblado (*dribblers_tackles_past*) é praticamente um lance em que falha o desarme (*dribblers_tackles_failed*), pelo que uma das estatísticas tem de ser retirada, tendo sido escolhida a estatística. De igual modo, se um jogador driblou com sucesso, quase sempre passou por um adversário, pelo que se retira a estatística *players_dribbled_past*.
- As estatísticas relacionadas com os toques na bola (contendo a palavra-chave *touches*) são irrelevantes e estão praticamente representadas nas estatísticas de passes ou de transportes de bola, pelo que serão retiradas. Igualmente, a estatística de toques a meio-campo (*touches_in_mid_third)* será retirada devido a ser natural que um jogador participe muito a meio-campo. No entanto, as estatísticas de toques no primeiro e último terço continuarão presentes, para verificar o impacto do número de ações nas zonas defensivas e ofensivas.
- Da mesma forma, as estatísticas que dizem respeito a ações em jogo corrido são praticamente semelhantes às estatísticas globais dessas mesmas ações, tornando-se necessário retirá-las. Devido a isso, as estatísticas de passes em jogo parados (*dead_ball_passes*) também são retiradas.
- A distância de passes normais (*passes_normal_distance*) encontra-se relacionada com várias estatísticas de passe e transporte de bola, pelo que será retirada.
- De forma semelhante às estatísticas de passes acertados e passes recebidos, a estatística de passes progressivos (*progressive_passes*) e passes progressivos recebidos (*progressive_passes_received*) possuem uma forte correlação entre si, sendo que a solução resulta em aplicar a mesma abordagem da estatísticas referidas no início deste ponto: considerar apenas uma das estatísticas para o modelo, neste caso a de passes efetuados, e retirar a outra.

Tabela 6 – Coeficientes de correlação (Primeira Execução)

Estatística A	Estatística B	Coeficiente
dribblers_tackles_past	*dribblers_tackles_failed*	1.000
touches	*live_touches*	0.998
passes_completed	*live_ball_passes*	0.993
live_touches	*live_ball_passes*	0.990
touches	*live_ball_passes*	0.986
live_touches	*passes_completed*	0.983
dribbles_successful	*players_dribbled_past*	0.981
touches	*passes_completed*	0.979
passes_completed	*grounded_passes*	0.979
passes_completed	*successful_ball_carries*	0.970
passes_completed	*passes_normal_distance*	0.969
live_ball_passes	*grounded_passes*	0.969
live_ball_passes	*successful_ball_carries*	0.966
medium_passes_completed	*passes_normal_distance*	0.965
live_ball_passes	*passes_normal_distance*	0.960
grounded_passes	*successful_ball_carries*	0.959
shot_creating_actions	*key_passes*	0.958
passes_completed	*medium_passes_completed*	0.957
live_touches	*successful_ball_carries*	0.955
live_touches	*grounded_passes*	0.954
medium_passes_completed	*live_ball_passes*	0.951
touches	*successful_ball_carries*	0.950
touches	*grounded_passes*	0.945
live_touches	*passes_normal_distance*	0.944
touches	*touches_in_mid_third*	0.944
grounded_passes	*passes_normal_distance*	0.943
touches_in_mid_third	*live_touches*	0.943
medium_passes_completed	*grounded_passes*	0.942
touches_in_mid_third	*live_ball_passes*	0.941
touches	*passes_normal_distance*	0.941
successful_ball_carries	*passes_normal_distance*	0.938
live_touches	*medium_passes_completed*	0.937
touches_in_mid_third	*passes_completed*	0.937
medium_passes_completed	*successful_ball_carries*	0.934
touches	*medium_passes_completed*	0.933
passes_completed	*short_passes_completed*	0.927
progressive_received_passes	*progressive_passes*	0.919
short_passes_completed	*live_ball_passes*	0.917
ball_carried_progressive_distance	*ball_carries_progressive*	0.917
short_passes_completed	*grounded_passes*	0.915
live_touches	*short_passes_completed*	0.913
touches	*passes_progressive_distance*	0.910
offsides	*offside_passes*	0.909
touches_in_mid_third	*passes_normal_distance*	0.909
touches_in_mid_third	*successful_ball_carries*	0.909
touches_in_mid_third	*grounded_passes*	0.907
touches	*short_passes_completed*	0.907
live_touches	*passes_progressive_distance*	0.906
passes_progressive_distance	*live_ball_passes*	0.905

- A estatística de passes rasteiros (*grounded_passes*) também tem uma correlação fortíssima com outras estatísticas, pelo que será retirada, bem como as estatísticas do mesmo tipo, isto é, que identifiquem a altura do passe: passes baixos (*low_passes*) e passes altos (*high_passes*).
- As estatísticas de passes chave (*key_passes*) e ações de criação de remate (*shot_creating_actions*) estão correlacionadas muito porque a primeira encontra-se englobada na segunda. Assim sendo, uma nova estatística será calculada que representará as ações de criação sem contar com os passes chave: *shot_creating_actions_without_key_passes*.
- As estatísticas de passe curto (*short_passes*) e (*médium_passes*) passe médio estão relacionadas com várias estatísticas, pelo que se torna importante criar uma nova estatística, *short_and_medium_passes*, que junte as duas, de forma a agrupá-las. Assim, haverá uma clara distinção entre passes de ligação de jogo (curtos e médios) e de alargamento de jogo (longos). No entanto, isto criará um problema com a estatística de total de passes completados. Assim sendo, a estatística *passes_completed* será retirada.
- Um passe para fora-de-jogo (*offside_passes*) naturalmente resulta num fora-de-jogo (*offsides*), no entanto, existem foras-de-jogo que não originaram de um passe. Dessa forma, a estatística de passes para fora-de-jogo será retirada do modelo, visto que já está contida na estatística de passes falhados.

Seguindo a remoção das variáveis consideradas excedentárias, a mesma função de verificação de correlação é aplicada, desta vez a estatísticas cujo coeficiente seja maior que 0.75. Os resultados, representados na Tabela 7, auxiliam a conclusão dos seguintes pontos:

- O número de relações entre estatísticas acima dos 90% de correlação passou de 49 para 2.
- Existem várias relações entre estatísticas em que uma das unidades é a ação em si (passe) e outra unidade é a distância resultante da ação ou de uma semelhante. Com isso, todas as estatísticas que estejam relacionadas com distância serão removidas.
- Grande parte dos bloqueios (*blocks*) efetuados correspondem a impedimentos de passes do adversário (*pass_blocks*), pelo que a relação entre as duas estatísticas é natural. Além disso, o modelo de dados já contempla a existência de outro tipo de bloqueios (*shot_blocks* e *shot_on_target_blocks*), que somados com os bloqueios de passe, resultam no valor da estatística de bloqueios, pelo que será retirada.
- Após a subtração dos passes chave na estatística de ações de criação de remate, a correlação desceu 9,7 valores percentuais, tornando as duas estatísticas mais independentes.
- A estatística de transportes de bola bem-sucedidos (*successful_ball_carries*) continua a ter várias relações com outras estatísticas. Além disso, a estatística pode ser irrelevante tendo em conta que já foi concluído que um passe é praticamente um transporte de bola bem-sucedido, e a estatística mais importante em relação a transportes seria a de transportes progressivos. Como tal, a estatística será retirada.

Tabela 7 – Coeficientes de correlação (Segunda Execução)

Estatística A	Estatística B	Coeficiente
successful_ball_carries	*short_and_medium_passes_completed*	0.966
ball_carried_progressive_distance	*ball_carries_progressive*	0.917
blocks	*pass_blocks*	0.897
pressures_in_mid_third	*pressures_failed*	0.875
passes_progressive_distance	*short_and_medium_passes_completed*	0.875
successful_ball_carries	*ball_carried_normal_distance*	0.874
passes_progressive_distance	*successful_ball_carries*	0.873
ball_carried_progressive_distance	*successful_ball_carries*	0.862
key_passes	*shot_creating_actions_without_key_passes*	0.861
touches_in_last_third	*passes_completed_to_the_last_third*	0.853
ball_carried_normal_distance	*short_and_medium_passes_completed*	0.853
ball_carried_progressive_distance	*short_and_medium_passes_completed*	0.846
passes_progressive_distance	*progressive_passes*	0.842
passes_completed_to_the_last_third	*progressive_passes*	0.833
touches_in_last_third	*ball_carries_progressive*	0.832
touches_in_attacking_penalty_area	*passes_completed_to_the_penalty_area*	0.830
ball_carries_progressive	*successful_ball_carries*	0.826
ball_carried_progressive_distance	*ball_carried_normal_distance*	0.822
ball_carries_progressive	*ball_carries_to_last_third*	0.819
passes_progressive_distance	*passes_completed_to_the_last_third*	0.817
ball_carries_progressive	*short_and_medium_passes_completed*	0.817
ball_carries_progressive	*passes_completed_to_the_last_third*	0.809
key_passes	*shots_missed*	0.808
passes_completed_to_the_last_third	*successful_ball_carries*	0.801
touches_in_last_third	*progressive_passes*	0.801
passes_completed_to_the_last_third	*short_and_medium_passes_completed*	0.797
shots_missed	*shot_creating_actions_without_key_passes*	0.791
ball_carried_progressive_distance	*ball_carries_to_last_third*	0.781
passes_progressive_distance	*ball_carried_normal_distance*	0.779
passes_progressive_distance	*long_passes_completed*	0.778
ball_carried_progressive_distance	*passes_progressive_distance*	0.778
ball_carries_progressive	*ball_carried_normal_distance*	0.772
touches_in_last_third	*touches_in_attacking_penalty_area*	0.768
pressures_in_def_third	*pressures_failed*	0.768
ball_carries_progressive	*progressive_passes*	0.758
touches_in_last_third	*ball_carries_to_last_third*	0.757
ball_carried_progressive_distance	*passes_completed_to_the_last_third*	0.756
loose_balls_recovered	*short_and_medium_passes_missed*	0.751

- A estatística de pressões falhadas (*pressures_failed*) tem duas correlações identificadas, nomeadamente com a estatística de pressões no primeiro terço (*pressures_in_def_third*) e no meio-campo (*pressures_in_mid_third*). Como existem mais falhadas do que com sucesso, a correlação acaba por ser natural.
- Existem outras correlações em que a sua presença acaba por ser intuitiva, nomeadamente entre ações dentro do último terço, como toques na bola (*touches_in_last_third*) ou transportes (*ball_carries_to_last_third*), e dentro da

grande área adversária (contendo a palavra-chave *to_the_penalty_area*). Os dados são exclusivos uns com os outros, isto é, as estatísticas contabilizadas dentro da grande área não estão incluídos nas de último terço.

Retiradas as estatísticas, a função de descoberta de correlações entre estatísticas uma terceira e última vez, sendo que o valor mínimo de correlação será de 0,70. Na Tabela 8 verifica-se que:

Tabela 8 – Coeficientes de correlação (Terceira Execução)

Estatística A	Estatística B	Coeficiente
pressures_in_mid_third	*pressures_failed*	0.875
key_passes	*shot_creating_actions_without_key_passes*	0.861
touches_in_last_third	*passes_completed_to_the_last_third*	0.853
passes_completed_to_the_last_third	*progressive_passes*	0.833
touches_in_last_third	*ball_carries_progressive*	0.832
touches_in_attacking_penalty_area	*passes_completed_to_the_penalty_area*	0.830
ball_carries_progressive	*ball_carries_to_last_third*	0.819
ball_carries_progressive	*short_and_medium_passes_completed*	0.817
ball_carries_progressive	*passes_completed_to_the_last_third*	0.809
key_passes	*shots_missed*	0.808
touches_in_last_third	*progressive_passes*	0.801
passes_completed_to_the_last_third	*short_and_medium_passes_completed*	0.797
shots_missed	*shot_creating_actions_without_key_passes*	0.791
touches_in_last_third	*touches_in_attacking_penalty_area*	0.768
pressures_in_def_third	*pressures_failed*	0.768
ball_carries_progressive	*progressive_passes*	0.758
touches_in_last_third	*ball_carries_to_last_third*	0.757
loose_balls_recovered	*short_and_medium_passes_missed*	0.751
touches_in_last_third	*passes_completed_to_the_penalty_area*	0.749
fouls_drawn	*free_kick_passes*	0.747
touches_in_last_third	*short_and_medium_passes_completed*	0.746
interceptions	*tackles_and_interceptions*	0.741
passes_completed_to_the_penalty_area	*progressive_passes*	0.741
touches_in_attacking_penalty_area	*key_passes*	0.736
progressive_passes	*short_and_medium_passes_completed*	0.733
touches_in_attacking_penalty_area	*shot_creating_actions_without_key_passes*	0.717
long_passes_completed	*sweeping_pass*	0.716
touches_in_attacking_penalty_area	*progressive_passes*	0.701

- Já não existe nenhuma correlação acima de 0,90.
- Existem correlações com um valor acima de 0,70, mas com uma justificação direta para o seu coeficiente, como a ligação entre faltas sofridas (*fouls_drawn*) e passes originados de livres (*free_kick_passes*), em que é natural o valor de passes vindo de livres subirem à medida que mais faltas são sofridas, e entre toques na grande área adversária (*touches_in_attacking_penalty_area*) e passes chave (*key_passes*), já que um passe chave é identificado como um passe que resultou num remate.

- Duas correlações encontradas consistem na identificação de estatísticas em que uma se encontra agregada na outra. A primeira relação é a entre desarmes (*tackles*) e desarmes e intercepções (*tackles_and_interceptions*). A segunda estatística será removida, visto ser a soma de duas outras estatísticas que já existem. A última relação é entre passes extensos (*sweeping_passes*) e passes longos completos (*long_passes_completed*), em que os passes extensos são identificados como passes com mais de 40 jardas de distância, e passes longos são passes com mais de 30 jardas. Como tal, o valor dos passes extensos será retirado do valor dos passes longos completos.

Após a análise de correlação, as alterações efetuadas no modelo principal são aplicadas a cada conjunto criado. Além disso, são criados conjuntos adicionados em que os dados são normalizados usando a função *minimax*, representada na Equação 9, visto que preserva a relação entre as variáveis enquanto as converte para a mesma escala (Ciaburro, 2018). Estes dados irão ser utilizados para a aplicação de algoritmos SVM, com vista a ter dois tipos de previsões diferentes e comparar resultados.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Equação 9 – Função de normalização min-max (Ciaburro, 2018)

5.5 Modeling

A modelação deste caso de estudo tem como passos os seguintes quatro pontos:

1. Obter os dados agregados na base de dados e aplicar o tratamento de dados referido na secção anterior.
2. Aplicar dois algoritmos diferentes de classificação, regressão linear e SVM, a vários modelos de cada campeonato e a um modelo com todos os dados registados.
3. Analisar os modelos criados e escolher os melhores pesos.
4. Obter os pesos decididos como os mais adequado e aplicar aos jogadores, registando todos estes dados.

No primeiro passo, a *query* no Código 3 da secção anterior é executada para cada competição registada e um *data frame* é criado para facilidade de navegação. Depois desse carregamento, é feito um novo pedido, desta vez para obter os dados relacionados com as estatísticas agregadas. Uma comparação entre o tamanho de cada conjunto, isto é, o número de jogos com estatísticas agregadas e o número de jogos com estatísticas ainda por tratar, é efetuada, sendo que caso os dados sem tratamento existam em maior número, um novo modelo é criado com os dados agregados, sendo que caso um dos novos jogos calculados já esteja inserido na base de dados, a execução avança para a próxima iteração.

```
#obter competições
competitions_in_database <- sqlQuery(connection_to_database, sqlQueryToGetAllCompetitions)
#obter jogos usando query que obtem todas as stats
games_with_statistics <- by(competitions_in_database, seq_len(nrow(competitions_in_database)),
                        function(competition_in_database)
get_all_matches_from_competition(competition_in_database))
#criar data frame do resultado
games_with_statistics <- do.call(rbind.data.frame, games_with_statistics)
games_with_statistics[is.na(games_with_statistics)] <- 0
#obter todos os jogos agregados na bd
sql_expression_to_select_matches <- paste("SELECT * FROM AggregatedMatch", sep="")
aggregated_matches <- sqlQuery(connection_to_database, sql_expression_to_select_matches)

#se os jogos sem tratamento estiverem em maior número, significa que há dados por adicionar à tabela
AggregatedMatch
if (length(unique(games_with_statistics$match_id)) * 2 > length(aggregated_matches[,1])) {
    games_with_statistics_complete <- get_aggregated_matches_data_model(games_with_statistics)
} else {
    #remover coluna de id dos resultados
    games_with_statistics_complete <- aggregated_matches[-grep("id", colnames(aggregated_matches))[1]]
}

#primeiro tratamento de dados
games_with_statistics_model <- treat_data(games_with_statistics_complete)
#tratar dados correlacionados estudados anteriormente
games_with_statistics_model_2 <- remove_correlated_data_1(games_with_statistics_model)
games_with_statistics_model_3 <- remove_correlated_data_2(games_with_statistics_model_2)
games_with_statistics_model_final <- remove_correlated_data_3(games_with_statistics_model_3)
games_with_statistics_model_final_normalized   <-   as.data.frame(sapply(games_with_statistics_model_final,
minmaxnorm))

#dividir frame por competições
games_divided_by_competition <- games_with_statistics_complete %>%
                              group_split(games_with_statistics_complete$competition_name, keep = FALSE)
#lista de variáveis para cada competição obtida
list_of_data_frame_variables <-  vector(mode = "list", length = length(games_divided_by_competition))
for (row in 1:length(list_of_data_frame_variables)) {
    competition_name <- games_divided_by_competition[[row]]$competition_name[[row]]
    variable_name_for_data_frame <- paste(to_snake_case(competition_name), 'data_frame', sep="_")
    data_frame_to_call <- do.call(rbind.data.frame, games_divided_by_competition[row])
    list_of_data_frame_variables[row] <- variable_name_for_data_frame
    data_frame_to_call <- create_data_model(data_frame_to_call)
    #primeiro tratamento de dados
    assign(variable_name_for_data_frame, treat_data(data_frame_to_call))
    #tratar dados correlacionados estudados anteriormente
    assign(variable_name_for_data_frame, remove_correlated_data_1(get(variable_name_for_data_frame)))
    assign(variable_name_for_data_frame, remove_correlated_data_2(get(variable_name_for_data_frame)))
    assign(variable_name_for_data_frame, remove_correlated_data_3(get(variable_name_for_data_frame)))
}
```

Código 4 – Criação do modelo de dados

Com o novo *data frame* criado, o tratamento e remoção de estatísticas irrelevantes ou correlacionadas é executado, bem como um novo conjunto de dados normalizados. Por fim, todos estes passos voltam a ser percorridos para obter diferentes modelos de dados de cada campeonato. O Código 4 tem as instruções em R executadas para ir ao encontro do primeiro passo.

O segundo passo inicia com a criação de modelos de previsão de resultados de um jogo de futebol baseado nas estatísticas globais de uma equipa. De forma à comparação entre os dois tipos de algoritmos de classificação, regressão linear e *support-vector machines* (SVM), ser o mais justa possível, os dados foram divididos em dois subconjuntos: um conjunto de treino com 70% dos dados e 30% no conjunto de teste. Os dois algoritmos serão validados usando o método de *cross-validation* com 10 *folds.*

No caso da aplicação de regressão linear, utiliza-se o método *glm*() da biblioteca stats (DataCamp, 2021), que requere como parâmetros a variável objetivo, definida como a coluna *match_won*, os dados a percorrer, e o tipo de distribuição que existe, que neste caso será uma distribuição binomial, devido a ser essencialmente uma pergunta de sim ou não (Weisberg, 2005). Depois da configuração, o modelo é criado e treinado, sendo validado de seguida com recurso à biblioteca *boot* e ao seu método *cv.glm*(), que especificamente aplica validações cruzadas em modelos lineares generalizados (Canty and Ripley, 2021). Por fim, são registados os pesos obtidos em cada execução, bem como a eficácia e o erro obtido no processo de *cross-validation*.

Para SVM, a biblioteca *caret* (Kuhn, 2019) é utilizada, devido à sua disponibilidade de métodos auxiliares ao procedimento. As parametrizações iniciais da execução do algoritmo implicam logo a validação do modelo usando *cross-validation*, sendo então três repetições em dez *folds* de dados, com base no método de reamostragem *repeatedcv*. O tipo de modelo configurado é linear *support-vector machines*, que é disponibilizado pela biblioteca com a parametrização *svmLinear*, efetuando também um pré-processamento utilizando métodos de centralização e dimensionamento. Depois das configurações, dois modelos distintos são criados, o primeiro utilizando a parametrização *tuneLength*, que efetua uma afinação utilizando *n* valores por defeito da biblioteca, e o segundo utilizando a parametrização *tuneGrid*, em que é o próprio utilizador que define os valores a executar. Para estes dois casos, é 10 a quantidade de valores para a afinação, e os valores personalizados são os seguintes: 0, 0.01, 0.05, 0.1, 0.25, 0.5, 1, 1.5, 2, 10 e 50.

Com as execuções terminadas, o terceiro passo começa por analisar os modelos criados e perceber qual será o melhor para o caso de estudo, e para isso, dois dados são analisados:

- Precisão do modelo de previsão de resultados
- Pesos atribuídos às variáveis do modelo

O motivo pelo qual estes dois dados são tidos em conta em vez de apenas o da precisão e do desvio padrão do modelo prende-se com o objetivo principal do caso de estudo. Como o objetivo é atribuir um valor objetivo de qualidade de jogador (um rating), a decisão não pode ser apenas escolher o modelo que melhor acerta nos resultados, mas sim o que, aliada a essa precisão, obtenha pesos de variáveis que façam sentido no contexto do negócio.

Tabela 9 – Valores de precisão e erro obtidos nas execuções dos algoritmos

Tabela de Precisões e Desvios Padrões Obtidas			
Algoritmo / Modelo	**Regressão Linear Precisão**	***SVMLinear* Precisão**	***SVMLinear TuneGrid* Precisão**
Bundesliga	0.771	0.762	0.775
La Liga	0.758	0.735	0.729
Ligue 1	0.783	0.753	0.754
Premier League	0.759	0.772	0.772
Serie A	0.770	0.748	0.748
UEFA Champions League	0.715	0.701	0.719
UEFA Europa League	0.713	0.735	0.735
Modelo com todos os jogos	0.767	0.764	0.764
Média	**0.754**	**0.746**	**0.750**

A Tabela 9 contém as precisões obtidas nas múltiplas execuções efetuadas, de forma a comparar as precisões. Os valores demonstram que o algoritmo que melhor previu os resultados foi o de regressão linear, obtendo os melhores valores em todos os modelos. Os indicadores de validação mantiveram-se dentro da mesma gama de valores.

Na comparação entre os dois tipos de execução do algoritmo *SVMLinear*, a aplicação de valores de afinação personalizados permitiu obter melhores resultados em quase todos os modelos, especialmente no modelo de jogos pertencentes à UEFA Champions League. Com isso, o modelo de *SVMLinear* irá ser descartado.

Em ambas as aplicações, o modelo com todos os jogos acabou por fazer parte dos melhores valores de precisão, o que abre a discussão para a hipótese de que o comportamento melhora à medida que o número de registos no conjunto aumenta, bem como que as diferenças entre competições não fazem com que as previsões piorem ou melhorem, isto é, o modelo não aparenta ter uma melhor eficácia se os jogos fizerem apenas parte de uma única competição. No lado oposto, os piores valores de eficácia pertencem ao conjunto de dados referente às competições da UEFA, o que se justifica com o grau de competitividade, em que jogam as melhores equipas europeias, e com terem o menor número de jogos.

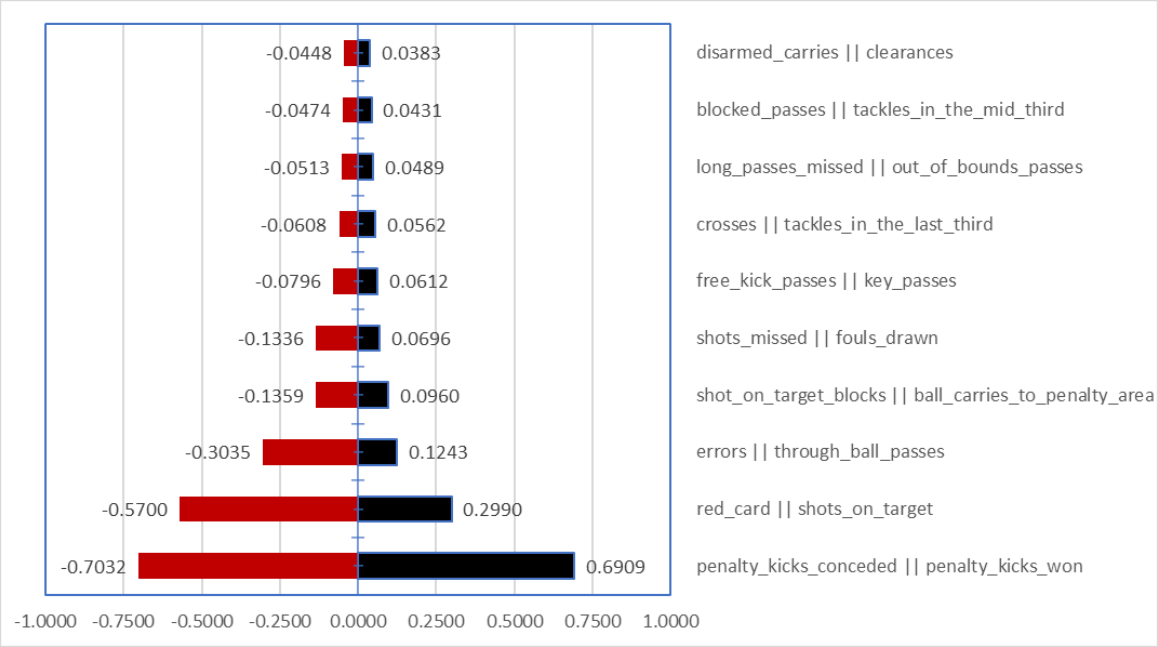


Figura 37 – Gráfico de pesos obtidos em regressão linear no modelo com todos os jogos

A Figura 37 demonstra os dez pesos com maior importância para a previsão uma vitória e os dez com maior impacto para uma não vitória obtidos na aplicação de regressão linear no modelo com todos os jogos. Aplicando o conhecimento obtido do negócio a partir do Contexto e Estado de Arte, é possível verificar que os pesos obtidos fazem sentido dentro do possível. As estatísticas com maior impacto têm como consequência alguma ação que coloca a equipa mais perto de marcar um golo, com destaque para as cinco principais: penaltis ganhos (*penalty_kicks_won*), remates à baliza (*shots_on_target*), passes em desmarcação (*through_ball_passes*), transportes de bola para a grande área (*ball_carries_to_penalty_area*) e faltas ganhas (*fouls_drawn*). De igual modo, as cinco com maior impacto negativo são dados que indiciam ações que resultam em oportunidades adversárias ou em piorar as chances de marcar golos da própria equipa, como penaltis concedidos (*penalty_kicks_conceded*), cartões vermelho (*red_card*), erros (*errors*), bloqueios defensivos de remates à baliza (*shot_on_target_blocks*) ou remates falhados (*shots_missed*). Estas duas últimas estatísticas, no entanto, permitem abrir uma discussão sobre o motivo pela sua inclusão. A primeira estatística teria hipoteticamente mais valor, porque bloquear um remate adversário que se dirige à baliza para impedir um golo parece ser algo positivo. No entanto, o problema aqui seria a continuidade dessa ação. Um bloqueio não significa uma bola recuperada, e se o remate se dirigia à baliza e teve de ser bloqueado por um defesa (imaginando uma situação em que o guarda-redes não esteja na baliza), é bastante provável que o bloqueio resulte numa recarga do adversário que possa resulta num golo. Já a outra estatística de remates falhados, tendo em conta que a estatística de natureza semelhante, mas resultado oposto (remates à baliza) é das com maior peso para a vitória de uma equipa, pelo que intuitivamente se esperaria um menor impacto no peso dos remates falhados, mas não um valor quase simetricamente oposto e com peso negativo. A explicação para esta diferença pode ser as inúmeras finalidades de um remate à baliza em oposição a um remate falhado:

um remate falhado irá sempre resultar numa reposição de bola adversária (pontapé de baliza ou lançamento) ou numa recuperação da equipa contrária, enquanto um remate à baliza pode resultar em golo, numa defesa do guarda-redes completa e incompleta, que por sua vez pode resultar noutra chance de golo, como um canto ou uma recarga (novo remate). Por fim, seria de estranhar a presença da estatística de cruzamentos nas variáveis com pior peso, visto que é um tipo de passe que, por definição, tem como zona de destino a grande área adversária, mas o seu impacto negativo deve-se à não distinção nos dados originais entre um cruzamento bem e malsucedido. Sem esta divisão, os jogadores que mais acertam cruzamentos não são valorizados, especialmente numa estatística em que a percentagem de acerto é de 23.5% (Soccerment Research, 2017).

Tabela 10 – Dez maiores e menores pesos obtidos no modelo linear com todos os jogos e os seus valores nos restantes modelos lineares

Variável	Bundesliga	La Liga	Ligue 1	Premier League	Serie A	UEFA Champions League	UEFA Europa League	Modelo com Todos os Jogos
penalty_kicks_won	0.7059	0.7957	0.6004	0.7261	0.5509	0.9268	0.9025	0.6909
shots_on_target	0.2492	0.3419	0.2730	0.2614	0.3007	0.3360	0.4254	0.2990
through_ball_passes	0.1761	0.1446	0.1537	0.1905	0.0417	0.1432	0.0422	0.1243
ball_carries_to_penalty_area	0.1662	0.1820	0.1347	0.0837	0.1235	0.0211	0.0365	0.0960
fouls_drawn	0.1011	0.0429	0.0749	0.0422	0.0946	0.0093	0.0886	0.0696
key_passes	0.1296	0.0829	0.0538	0.0230	0.0200	0.0493	0.0432	0.0612
tackles_in_the_last_third	0.0611	0.0024	0.0291	0.1676	0.0400	0.0517	0.0932	0.0562
out_of_bounds_passes	0.0564	0.0683	0.0494	0.0273	0.0061	0.0146	0.0213	0.0489
tackles_in_the_mid_third	0.0742	0.0055	0.0579	0.0813	0.0353	0.0472	0.0341	0.0431
clearances	0.0350	0.0233	0.0393	0.0502	0.0487	0.0492	0.0555	0.0383
disarmed_carries	0.0701	0.0583	0.0712	0.0164	0.0481	0.0557	0.1000	0.0448
blocked_passes	0.0411	0.0694	0.0649	0.0339	0.0536	0.0163	0.0236	0.0474
long_passes_missed	0.0536	0.0648	0.0674	0.0421	0.0694	0.0628	0.0734	0.0513
crosses	0.0153	0.0941	0.0579	0.0807	0.0645	0.0525	0.0395	0.0608
free_kick_passes	0.1105	0.0692	0.0761	0.0397	0.0752	0.0171	0.1147	0.0796
shots_missed	0.2688	0.1074	0.1625	0.2012	0.0963	0.0779	0.0768	0.1336
shot_on_target_blocks	0.0606	0.4916	0.1345	0.2207	0.3365	0.0737	1.0484	0.1359
errors	0.4062	0.3496	0.2775	0.3947	0.3973	0.7891	0.2611	0.3035
red_card	0.7448	0.7108	0.6162	0.5362	0.8962	0.7768	0.0311	0.5700
penalty_kicks_conceded	0.7473	0.6176	0.9322	1.1488	0.6709	0.3805	1.0027	0.7032

Encontram-se listados na Tabela 10 os dez maiores (com cor preta) e menores pesos (com cor vermelha) referidos anteriormente, só que com a comparação do valor obtido nos restantes modelos. A partir dos dados presentes na tabela, pode-se verificar que:

- Os pesos negativos do modelo com todos jogos mantém a mesma propriedade ao longo de todos os modelos, à exceção de dois casos, em que na Série A o peso de um bloqueio defensivo de remate que ia à baliza (*shot_on_target_blocks*) tem uma subida para 0.3655, e na UEFA Champions League os passes a partir de livres (*free_kick_passes*) sobe para 0.0171.

- À exceção do modelo da UEFA Champions League e da variável de desarmes no meio-campo (*tackles_in_the_mid_third*) nos modelos da Serie A e da UEFA Europa League, todos os pesos positivos obtidos no modelo com todos os jogos são também positivos nos restantes modelos, variando apenas no valor de impacto.
- Os dois maiores pesos em todos os modelos correspondem sempre a penáltis ganhos (*penalty_kicks_won*) em primeiro lugar e a remates à baliza (*shots_on_target*) em segundo lugar, à exceção do modelo da Serie A, em que a variável de remates à baliza passa para terceiro lugar. Isto permite concluir que, de facto, as estatísticas mais importantes são as que permitem de forma mais direta obter golos: rematar à baliza e ganhar o lance com maior probabilidade de resultar em golo.
- No lado oposto, os dois menores pesos, penaltis concedidos (*penalty_kicks_conceded*) e cartões vermelhos (*red_card*), também permanecem nas duas posições inferiores em todos os modelos relacionados com campeonatos nacionais, mas não mantendo sempre a sua posição, isto é, em alguns casos a estatística com menor peso é o cartão vermelho, enquanto noutros é cometer um penalti. No entanto, nas competições europeias, isso já não acontece, sendo que na UEFA Champions League a estatística de erros cometidos (*errors*) aparece em primeiro lugar das com maior impacto negativo, passando os cartões vermelhos para segundo e os penaltis concedidos para terceiro, e na UEFA Europa League a estatística de expulsão do jogo passa para décimo oitavo lugar, com a estatística de bloqueios defensivos de remates à baliza a subir para primeiro lugar. Esta diferença pode ser justificada com a natureza competitiva diferente entre competições europeias e competições nacionais. O grau de dificuldade teórica e do contexto de um jogo difere do campeonato nacional para competições continentais, em que tanto se pode defrontar uma das melhores equipas europeias, como uma equipa de um campeonato periférico de menor valor.

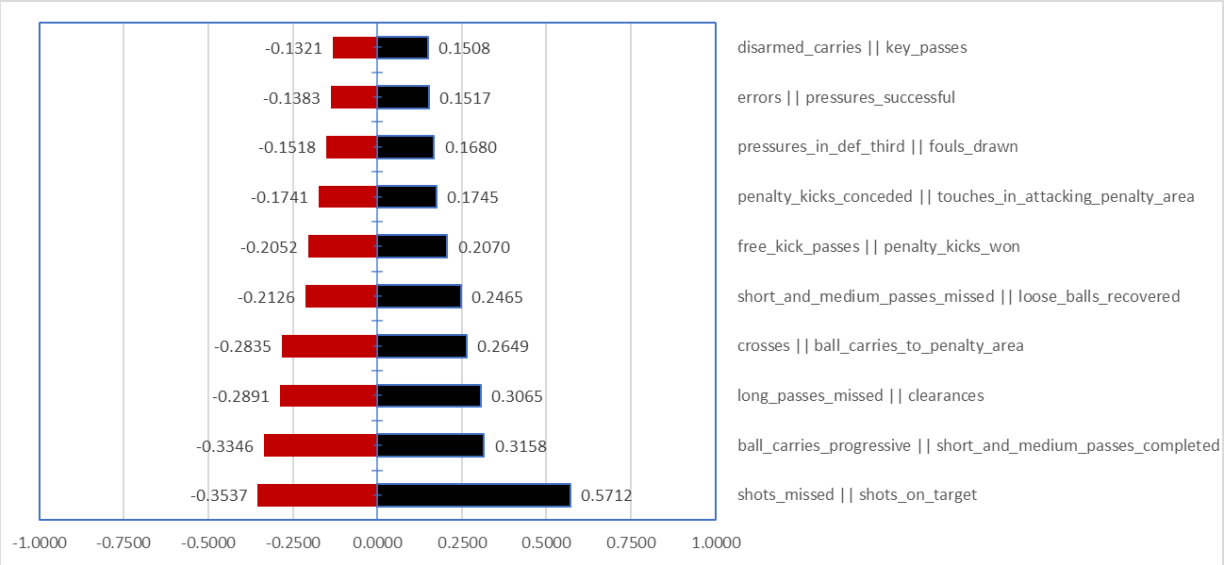


Figura 38 – Gráfico de pesos obtidos em *SVMLinear* com *TuneGrid* no modelo com todos os jogos

As dez variáveis do modelo de *SVMLinear* com *TuneGrid* com maior e menor peso são apresentadas no gráfico presente na Figura 38. Ao contrário do que aconteceu no modelo criado usando regressão linear, os pesos acabam por ter intervalos mais pequenos entre si, já que a diferença entre o primeiro e o décimo valor é de 0.4204 para os impactos positivos e de 0.2216 para os impactos negativos, enquanto na regressão linear, essa diferença aumenta para 0.6526 nos positivos e 0.6585 nos negativos. As variáveis presentes no top 10 também são diferentes, com o maior destaque a ser a presença da estatística de passes curtos e médios efetuados com sucesso, em que no modelo de regressão linear a variável tinha um valor de 0.0046, quase cem vezes menor. No entanto, e novamente com base no conhecimento obtido do negócio a partir do Contexto e Estado de Arte, os valores obtidos são coerentes com o que se espera das estatísticas mais importantes para alcançar uma vitória, só que desta vez com uma particularidade que difere na abordagem ao problema: os valores foram todos normalizados, o que faz com que estatísticas que anteriormente eram de grande volume, como o exemplo das estatísticas de passe, possam ter uma maior ou menor importância, dependendo de se o seu peso é altamente negativo ou positivo, para a previsão de resultados.

Tabela 11 – Dez maiores e menores pesos obtidos no modelo SVM Linear com Tune Grid com todos os jogos e os seus valores nos restantes modelos

Variável	Bundesliga	La Liga	Ligue 1	Premier League	Serie A	UEFA Champions League	UEFA Europa League	Modelo com Todos os Jogos
shots_on_target	0.5625	0.4745	0.5574	0.5336	0.5567	0.4367	0.9693	0.5712
short_and_medium_passes_completed	0.0943	0.1254	0.3587	0.7869	0.3002	0.5113	0.4835	0.3158
clearances	0.1750	0.1708	0.2713	0.3980	0.4209	0.2256	0.3473	0.3065
ball_carries_to_penalty_area	0.2241	0.2728	0.2713	0.2035	0.2531	0.1970	0.1247	0.2649
loose_balls_recovered	0.1368	0.2094	0.2190	0.2318	0.4027	0.2230	0.1074	0.2465
penalty_kicks_won	0.1218	0.2335	0.1620	0.2178	0.1700	0.1819	0.2748	0.2070
touches_in_attacking_penalty_area	0.0833	0.2118	0.1406	0.4314	0.2659	0.1617	0.1843	0.1745
fouls_drawn	0.1527	0.0960	0.2333	0.0467	0.3852	0.0936	0.3749	0.1680
pressures_successful	0.1595	0.0750	0.1838	0.0462	0.1051	0.0168	0.4401	0.1517
key_passes	0.1134	0.0776	0.1636	0.2584	0.1974	0.1544	0.0248	0.1508
disarmed_carries	0.0868	0.0506	0.1657	0.0301	0.1230	0.0571	0.2535	0.1321
errors	0.1442	0.1284	0.0701	0.1704	0.1994	0.2247	0.0535	0.1383
pressures_in_def_third	0.0714	0.1605	0.2412	0.0844	0.2518	0.0712	0.3529	0.1518
penalty_kicks_conceded	0.2457	0.1394	0.2263	0.2240	0.2074	0.0454	0.2384	0.1741
free_kick_passes	0.1538	0.1549	0.2003	0.1285	0.3764	0.0502	0.5043	0.2052
short_and_medium_passes_missed	0.2194	0.1465	0.1136	0.2714	0.2722	0.1455	0.0818	0.2126
crosses	0.1950	0.3027	0.2934	0.3644	0.3511	0.4074	0.1489	0.2835
long_passes_missed	0.2323	0.2575	0.3025	0.2352	0.3892	0.2524	0.4396	0.2891
ball_carries_progressive	0.2098	0.2103	0.2627	0.5682	0.3530	0.0862	0.3726	0.3346
shots_missed	0.3613	0.2184	0.3874	0.5646	0.5565	0.1928	0.0647	0.3537

A Tabela 11 lista as vinte variáveis representadas na Figura 39, divididas em dois grupos: as dez variáveis com maior impacto (com cor preta) e as dez com menor impacto (com cor vermelha) do modelo com todos os jogos e os valores correspondentes nos restantes modelos *SVMLinear TuneGrid* criados. A partir da informação presente, pode-se verificar que:

- Os pesos positivos do modelo com todos jogos mantém a mesma propriedade ao longo de todos os modelos, à exceção da estatística de passes chave (*key_passes*) no modelo da UEFA Europa League. No entanto, os pesos negativos acabam por ter mais discrepâncias nos restantes, especialmente a variável de controlos de bola desarmados (*disarmed_carries*), que passa a ter um peso positivo no modelo da Premier League, o modelo da UEFA Champions League tem as variáveis de passes de livres diretos (*free_kick_passes*) e pressões aplicadas no primeiro terço/terço defensivo (*pressures_in_def_third*) com impacto positivo, e o modelo da UEFA Europa League tem a variável de remates falhados (*shots_missed*) com valor positivo.
- A variável com maior peso em seis dois oito modelos é a variável de remates à baliza (*shots_on_target*), o que acaba por ser natural, visto ser a estatística que intuitivamente mais contribui para marcar golos. Nos dois modelos que não é a com mais impacto, é a variável de passes curtos e médios completados (*short_and_medium_passes_completed*) que fica em primeiro lugar, passando a variável de remates para segundo lugar.
- Nas variáveis com maior impacto negativo, a com mais peso é diferente ao longo dos modelos, normalmente sendo uma entre remates falhados (*shots_missed*), passes longos falhados (*long_passes_missed*), transportes de bola progressivos (*ball_carries_progressive*) e cruzamentos (*crosses*) A inclusão da primeira acaba por ser percetível visto que a conclusão do lance acaba sempre por ser uma reposição de bola adversária, no entanto, a segunda abre a discussão de quão negativo o seu impacto realmente deveria ser. Um passe falhado, seja qual for o seu tipo, é uma ação negativa, no entanto, um passe longo, se for de trás para a frente, pode permitir à equipa subir no terreno e colocar-se numa melhor posição de efetuar ações defensivas e ofensivas mais perto da baliza adversária. Porém, é também possível perceber que uma equipa que faça mais passes longos falhados acabe por ter menos posse-de-bola, visto que um passe longo acaba por ter uma mais difícil execução do que um passe curto, o que fará com que a equipa não consiga ter bola tantas vezes como se efetuasse mais passes curtos. As terceiras e quartas estatísticas também representam, à partida, ações ofensivas de qualidade, mas também são estatísticas de alto risco ou de baixo ganho. A estatística de cruzamento já foi referida na análise do modelo de regressão linear, onde se verificou que um cruzamento tem apenas 23.5% de acerto, e a estatística de transportes de bola pode simbolizar a velha máxima do futebol: para passar pelo adversário, é mais fácil passar a bola ao colega do que tentar driblar a bola (Keel, 2016).
- Os pesos de penaltis ganhos e concedidos acabam por ter um peso menor do que o esperado. Isso pode ser explicado com a ação de normalização de dados efetuada, já que o número de penaltis por jogo acaba por ser de pouquíssimo volume, com a média de penaltis por jogo no futebol europeu ser de 0.36, um valor que se traduz em 1 penalti a cada 2.78 jogos (Durán, 2020).

Com estas observações, o modelo a ser utilizado será o de regressão linear. Além da previsão de resultados ser melhor nesse modelo, os pesos das variáveis acabam por ter uma

interpretação mais simples de explicar e aplicar, visto que o conjunto de dados usado para testar não recebeu tratamentos de modificação. O modelo de SVM, apesar de ter uma precisão semelhante à do modelo de regressão linear, a interpretação e aplicação dos pesos obtidos vai mais ao encontro de uma análise de equipa, e não individual, e isso pode ser exemplificado com o peso alto dado às variáveis de passes curtos e médios, ou seja, para saber se uma equipa no seu todo tem performances que a colocam mais perto da vitória, o modelo de SVM poderia ser aplicado e mais fácil de perceber. Já o modelo de regressão linear tem uma interpretação mais direta para a performance individual de um jogador. Por fim, o modelo a utilizar será o modelo com todos os jogos, já que a diferença quer a nível de previsão, quer a nível dos pesos obtidos foi considerada insignificante, além de que o objetivo é perceber quais são os melhores jogadores independentemente da competição em que participam.

Com o modelo de regressão linear validado, decidiu-se fazer uma nova execução do modelo de regressão linear, desta vez num conjunto com todos os dados presentes, ignorando qualquer divisão de dados, seja por competição, seja de forma aleatória para validações e comparações. Com isto, obteve-se os pesos finais a utilizar na próxima fase do modelo, com os dez maiores e os dez menores a poderem ser verificados na Tabela 12. No Anexo 4, todos os pesos poderão ser verificados.

Tabela 12 – Pesos finais do modelo linear de previsão

Variável	Peso
penalty_kicks_won	0.7185
shots_on_target	0.3083
through_ball_passes	0.1343
ball_carries_to_penalty_area	0.1115
fouls_drawn	0.0703
tackles_in_the_last_third	0.0637
out_of_bounds_passes	0.0486
key_passes	0.0472
clearances	0.0387
touches_in_attacking_penalty_area	0.0303
blocked_passes	0.0430
yellow_cards	0.0494
long_passes_missed	0.0506
crosses	0.0611
free_kick_passes	0.0796
shots_missed	0.1286
shot_on_target_blocks	0.2029
errors	0.3341
red_card	0.6054
penalty_kicks_conceded	0.7270

De seguida, criaram-se três novos modelos de regressão com todos os jogos, mas com a novidade de em cada modelo ter uma nova estatística das três relacionadas com a equipa

marcar golos: golos (*goals*), assistências (*assists*) e ações de criação de golo (*goal_creating_actions*). As mesmas parametrizações do algoritmo de regressão linear foram aplicadas e os novos pesos, bem como as previsões alcançadas, podem ser vistas na Tabela 13.

Tabela 13 – Pesos e Precisões dos modelos com as variáveis relacionadas com golo

Tabela de Pesos e Precisões dos novos modelos Dados / Modelo	Precisão	Peso da Variável em Estudo	Segunda variável	Segundo maior peso
Modelo com a variável *goals*	0.8377	1.8713	*ball_carries_ to_penalty_area*	0.0967
Modelo com a variável *assists*	0.8146	1.4457	*penalty_kicks_ won*	1.0487
Modelo com a variável *goal_creating_actions*	0.8232	0.8531	*shots_on_ target*	0.1759

As precisões aumentaram substancialmente com a introdução destas variáveis, o normal visto que são estatísticas que têm sempre como ação final a equipa marcar golo. Além disso, os pesos obtidos são substancialmente maiores aos verificados na Tabela 12 do modelo sem estas estatísticas presentes. Além disso, as segundas maiores variáveis acabam por ficar com um peso muito menor do que a variável em estudo, à exceção da estatística de penaltis ganhos no modelo de assistências, que teve o seu peso aumentado. No entanto, é de notar que a precisão do modelo com a variável de ações de criação de golo foi superior à do modelo com assistências, mas o peso acabou por ser bastante menor. Há várias explicações possíveis, desde as várias ligações que o modelo possa ter encontrado entre a estatística de ações de criação de golo e outros dados, até ao impacto das assistências ser tão grande que acaba por denegrir outras variáveis importantes.

De notar também que a variável de golos na própria baliza (*own_goals*) irá ser adicionada, obtendo o valor simétrico do peso atribuído ao golo, isto porque como um autogolo acaba por prejudicar de forma tão direta como um golo na baliza adversária, decidiu-se dar o mesmo valor, mas com o sinal negativo.

Com o modelo escolhido e os pesos registados na base de dados, a última fase da modelação consiste no registo dos *ratios* de todos os jogadores. De acordo com o modelo relacional demonstrado na secção 5.2, a tabela que contém essa informação terá registos para todos os jogos de cada jogador identificado na base de dados, sendo os valores calculados os seguintes:

- O valor de qualidade de jogador;
- O valor de qualidade de jogador sem contabilizar as estatística de golos e autogolos;
- Os valores referidos anteriormente, mas divididos pelos minutos efetuados na partida associada ao registo. Estes valores só serão preenchidos caso o jogador tenha efetuado pelo menos 10 minutos na partida, de forma a evitar *outliers*.

Cada registo na tabela tem também o valor da data a que foi efetuada a partida, uma coluna que será aplicada num algoritmo de média móvel em todos os registos de cada jogador, com vista a privilegiar as performances mais recentes de cada jogador, ao invés de fazer uma

simples média móvel. Para o efeito, será utilizada a técnica de *Exponential Weighted Moving Average* (EWMA), referida na secção 2.3.1.3, com recurso à função *ewmaSmooth* da biblioteca de R *qcc* (Scrucca et al., 2017). A utilização desta técnica deve-se à facilidade de aplicação da tecnologia disponibilizada, a obter-se novos valores para serem comparados, bem como à inspiração retirada do artigo científico Playerank referido na secção 2.5.1.1.

Para se determinar a escolha do valor de suavização, definido como *lambda* λ, escolheu-se o jogador argentino Lionel Messi para verificar o comportamento de um valor mais perto de 0 e de outro mais perto de 1. O jogador tem 162 jogos na base de dados, com a melhor exibição a ter um valor de *rating* de 12.497 e a pior exibição a ter -1.083. A média das suas exibições todas é de 1.748. Neste caso, os valores escolhidos foram 0.05 e 0.65.

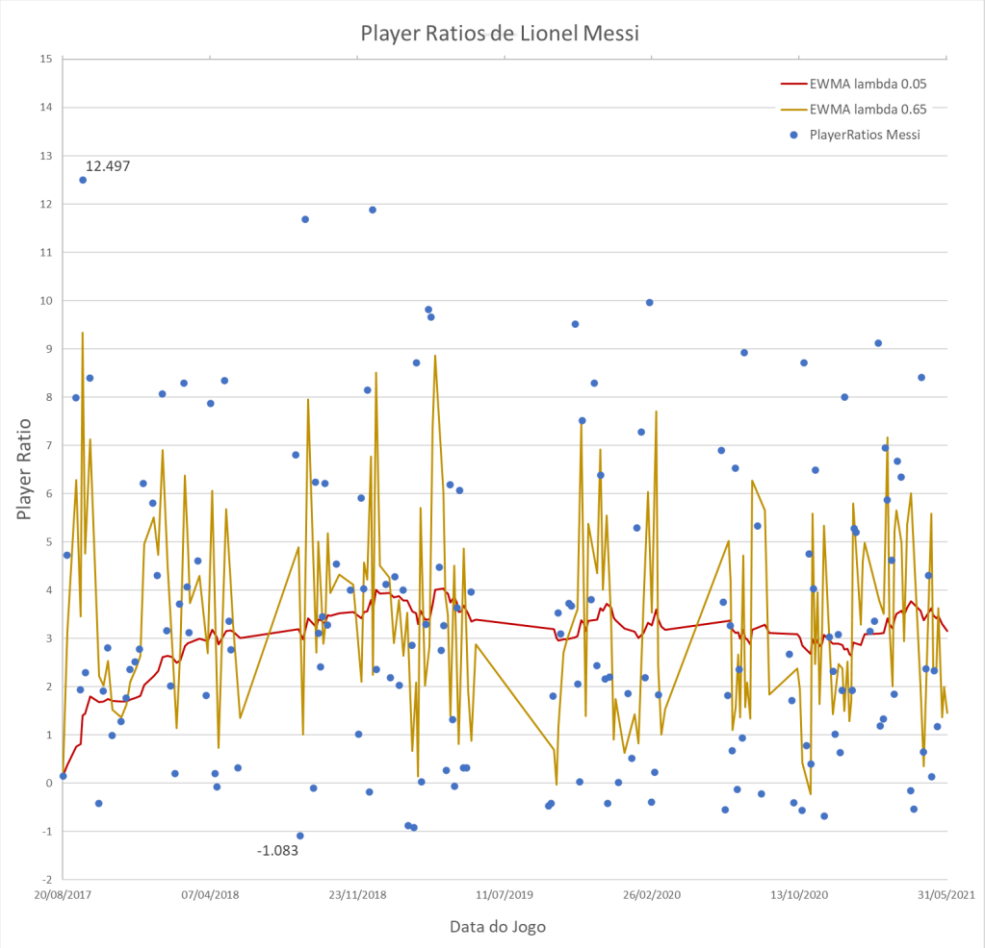


Figura 39 – Comparação de valores de suavização no EWMA

A Figura 39 demonstra três dados: todas as exibições de Lionel Messi avaliadas pelo modelo, representadas a azul e como pontos no gráfico, a média móvel usando 0.05 como o valor de λ a vermelho, e a média móvel usando 0.65 como o valor de λ a amarelo. O gráfico demonstra que os valores das médias móveis usando 0.65 como valor de suavização criam muitos picos negativos e positivos, o que pode ser um problema visto que privilegia exibições muito boas ou muito más, ao contrário de beneficiar um momento sustentado de boa ou má forma.

Como tal, o valor de suavização a utilizar será de 0.05, com vista a eliminar oscilações e a privilegiar momentos de forma sustentados e não performances espontâneas de grande ou baixa qualidade.

Para o passo seguinte de avaliação, além dos valores calculados aplicando a técnica de EWMA, será também calculado um valor de rating global, ou seja, as estatísticas totais de cada jogador ao longo da carreira serão somadas para serem calculadas com recurso aos pesos registados. Desta forma, dois tipos de *ratings* serão comparados entre si para perceber qual o melhor se adequa ao caso de estudo.

5.6 Model Evaluation

A avaliação do modelo será determinada seguindo os dois passos abaixo:

- Comparação dos valores obtidos com outros existentes no mercado.
- Inquérito realizado a intervenientes no desporto de forma ativa, seja como jogadores, membros de equipas técnicas ou estudantes em cursos de treinador.

De forma a comparar os valores obtidos nos cálculos dos modelos, seis índices calculados no modelo vão ser colocados à prova:

- O valor de *ratio* acumulado com e sem golos.
- O valor de EWSA do *ratio* com e sem golos ao longo da carreira do jogador com e sem golos, sendo escolhido o último valor da lista de médias móveis.
- A multiplicação entre o número de minutos e o valor de EWSA do *ratio* com e sem golos.

Os 20 melhores jogadores divididos por posição em cada índice que tenham pelo menos 30 jogos disputados e no mínimo 1350 minutos efetuados foram escolhidos para serem alvo de comparações. Alguns jogadores estão presentes no top 20 de todos os ratings, pelo que o número de jogadores em cada tabela de posição será diferente. Os valores registados na base de dados do FIFA 22 e do Football Manager (FM) 2021, dois videojogos dedicados à simulação de futebol, serão usados como comparação, sendo também escolhidos os 20 melhores em cada posição na base de dados. De realçar uma diferença entre os dois jogos: a base de dados do FIFA 22 diz respeito ao início da época 2021/2022, e a base de dados do FM 2021 refere-se ao fecho de mercado de inverno da época 2020/2021. Isto significa que os valores dados nos dois jogos têm como pontos de partida diferentes, o que pode influenciar algumas comparações. O objetivo seria utilizar a base de dados da versão de 2022 do Football Manager, mas até à data de escrita do documento, o jogo ainda não tinha sido lançado para o público geral. No entanto, e devido à forma como os dados pertencentes ao Football Manager são alvos de um forte escrutínio por parte da equipa responsável pela gestão dos dados, conforme referido na secção 2.5.2.4, será a principal métrica de comparação a ser utilizada.

Antes de ser feita a avaliação dos valores obtidos, analisaram-se os resultados e observou-se o seguinte:

- O avançado argentino Lionel Messi foi o jogador com o valor mais alto de *ratio* acumulado com e sem golos, 554.46 e 283.12 respetivamente. Nos valores das médias móveis, em ambos os dados ficou em terceiro lugar, com o avançado polaco Robert Lewandowski a obter a melhor média móvel contabilizando os golos, com o valor de 3.86, e o médio alemão Thomas Muller a ser o jogador com a melhor média móvel sem golos, com 1.53 como média. O jogador com os valores mais baixos quer nos dois *ratios* acumulados, quer nas duas médias móveis, é o defesa brasileiro Lucas Lima, com -0.34 e -0.36 como médias.
- Nos valores da média móvel multiplicada pelos minutos disputados por cada jogador, o com valor mais alto é Robert Lewandowski. No entanto, sem contabilizar os golos, Lionel Messi volta a ser o melhor. Nos valores negativos do mesmo comparativo, é o defesa uruguaio Damián Suárez que aparece como o jogador com o pior valor.
- A melhor exibição registada na base de dados pertence ao avançado brasileiro Neymar ao serviço do Paris Saint-Germain frente ao Dijon, no dia 17 de janeiro de 2018, com 16.24 como *rating*. Marcou quatro golos e fez duas assistências, tendo inclusive recebido uma nota perfeita de 10 pelo jornal L'Equipe, sendo a oitava desde 1988 (Assouline, 2018). A melhor exibição sem golos contabilizados coube ao avançado holandês Memphis Depay, realizada ao serviço do Lyon no terreno do Metz a 8 de abril de 2018, em que assistiu para quatro golos, obtendo 10.53 no valor de *rating*.
- Já a pior exibição foi realizada pelo defesa espanhol Mikel San José, que a 28 de abril num jogo frente à Real Sociedad, jogando pelo Athletic Club, marcou dois golos na própria baliza e ainda cometeu mais dois erros que resultaram em remates do adversário, com -4.34 a ser o *rating* da sua performance. Se não contabilizarmos os golos na própria, a pior exibição foi efetuada pelo avançado costa-marfinense Max Gradel, a 18 de novembro de 2017, enquanto era membro do Toulouse numa partida frente ao Metz. Num rating de -2.64, o jogador rematou três vezes sem nunca acertar na baliza, e efetuou dez cruzamentos, além de ter trinta passes falhados.

Posto isto, os jogadores foram agrupados em três grupos: avançados, médios e defesas, de forma a ser efetuada a comparação dos valores calculados e dos valores presentes no mercado. Os valores a comparar são o PRA (*Player Ratio* Acumulado), PRSGA (*Player Ratio* Sem Golos Acumulado), EWSAPR (*Exponential Weighted Smoothing Average Player Ratio*), EWSAPRSG (*Exponential Weighted Smoothing Average Player Ratio* Sem Golos), EWSAPRXM (*Exponential Weighted Smoothing Average Player Ratio* multiplicado por minutos) e EWSAPRSGXM (*Exponential Weighted Smoothing Average Player Ratio* Sem Golos multiplicado por minutos).

A Tabela 14 contém os dados calculados usando o modelo de regressão linear e a sua posição relativa naquele parâmetro, bem como os valores e posições correspondentes de cada jogador nas bases-de-dados do FIFA e do FM. Como nos modelos de comparação, a escala já

está definida e não contempla a presença de valores decimais, valores iguais correspondem a posições relativas iguais, por exemplo, Sadio Mané e Kylian Mbappé têm o mesmo valor de qualidade na base de dados do FM, pelo que a sua posição relativa é de sextos classificados, enquanto Mohamed Salah, que tem um valor abaixo dos dois jogadores anteriores, fica em oitavo, mesmo tendo o sétimo índice mais alto. Todos os dados foram convertidos para a mesma escala dos dados presentes no FM, de forma a ser mais fácil comparar os dados. No caso dos dados do FIFA, foi efetuada um cálculo de proporcionalidade direta, multiplicando o valor presente pelo coeficiente $\alpha = \frac{99}{200}$, sendo 99 o máximo estipulado no FIFA e 200 o máximo no FM. Para os dados do modelo em estudo, devido a não haver um máximo ou mínimo conhecido, a seguinte equação inspirada no cálculo de normalização min-max (Ciaburro, 2018) foi utilizada:

$$R_{jogadorescalax} = \frac{x - x_{min}}{x_{max} - x_{min}} \times (y_{max} - y_{min}) + y_{min}$$

Equação 10 – Equação de normalização para comparação

Em que $R_{jogadorescalax}$ é o valor calculado resultante da normalização do *rating* x usando os valores mínimos e máximos de cada elemento presente na tabela, multiplicada subtração entre o máximo e mínimo dos valores do FM y, somando por fim o valor mínimo de FM, para que fiquem todos entre o intervalo máximo de valores pertencentes ao FM. Depois desse tratamento, os dados foram analisados e retiraram-se as seguintes observações:

- Dos vinte melhores jogadores da base de dados do Football Manager, quatro não estão presentes no top 20 de qualquer um dos *ratios* calculados: Paulo Dybala, Leroy Sané, Eden Hazard e Sergio Aguero. A não presença destes jogadores justifica-se com lesões que os jogadores tiveram que prejudicaram a sua performance. O alemão Leroy Sané não jogou praticamente a época toda de 2019/2020 devido a lesão, Eden Hazard sofreu 12 lesões desde 2019/2020 até ao final de 2020/2021, Sergio Aguero é um jogador propenso a lesões e desde 2017/2018 lesionou-se 14 vezes que o fizeram perder 61 jogos, e Paulo Dybala teve uma lesão grave em 2020/2021 que o fez perder 18 jogos. Com isto, a discrepância acaba por ser natural, visto que o modelo é baseado em estatísticas que são prejudicadas por situações destas e os valores do FM são baseados na apreciação que uma pessoa tem da qualidade do jogador.
- A diferença entre os *ratios* com e sem golos acaba por ser uma conclusão óbvia: os jogadores com mais golos acabam por sair mais destacados nos *ratios* que considerem essa estatística, já os jogadores com perfil mais criativo, isto é, mais assistências e passes chave, sobem posições nos *ratios* sem golos. No entanto, nos melhores jogadores, a presença de golos ou não acaba por não modificar muito a posição. Lewandowski desce sempre de posições relativas entre os dois ratios, mas não mais que 6 posições, Cristiano Ronaldo mantém-se sempre no top 20 seja qual for o *ratio*, e Ciro Immobile em nenhum dos casos sai do top 10. Isto ajuda a perceber que os cálculos sem contabilizar diretamente os golos, acabam por contabilizá-los indiretamente, seja a partir de remates ou de ações na grande área adversária.

Tabela 14 – Valores de avançados calculados

Jogador	PRA		PRSGA		EWSAPR		EWSAPRSG		EWSAPRXM		EWSAPRSGXM		FIFA 22		FM 2021	
Nome	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank
L. Messi	195.00	1	195.00	1	179.67	3	194.65	3	187.36	2	195.00	1	187.88	1	195	1
R. Lewandowski	185.25	2	169.40	8	195.00	1	190.50	6	195.00	1	184.87	3	185.86	2	184	3
K. Mbappé	176.59	3	175.10	2	176.32	4	185.25	7	170.81	4	171.44	8	183.84	3	180	6
M. Salah	175.93	4	172.62	5	152.47	22	159.39	42	164.39	10	169.85	9	179.80	7	179	8
C. Immobile	173.98	5	169.50	7	162.00	9	181.66	9	167.04	7	179.15	5	175.76	15	163	24
C. Ronaldo	173.26	6	161.13	13	165.76	7	166.88	20	170.67	5	168.30	12	183.84	3	194	2
L. Suárez	171.17	7	173.51	4	151.31	28	155.44	51	156.36	15	157.58	22	177.78	11	170	15
K. Benzema	168.50	8	169.91	6	162.56	8	175.00	14	169.54	6	176.76	6	179.80	7	178	9
R. Sterling	168.22	9	173.51	3	144.51	55	164.84	24	151.75	24	166.50	15	177.78	11	171	13
H. Kane	164.94	10	153.33	26	159.92	13	162.60	34	165.47	9	164.98	17	181.82	6	183	4
R. Lukaku	163.35	11	156.77	16	167.38	6	192.87	5	172.04	3	188.07	2	177.78	11	170	15
T. Werner	162.04	12	163.46	10	144.42	56	172.85	15	149.91	27	169.01	11	169.70	35	155	53
Neymar	159.88	13	159.35	14	160.15	12	175.56	13	149.37	31	151.90	34	183.84	3	181	5
Son H.	159.41	14	161.51	11	143.73	65	150.92	68	148.57	33	152.23	30	179.80	7	172	12
S. Mané	158.88	15	157.21	15	144.32	58	154.62	56	153.05	17	160.97	20	179.80	7	180	6
D. Zapata	158.12	16	155.07	20	149.47	33	162.72	32	149.89	28	155.44	26	167.68	48	157	38
P. Aubameyang	157.84	17	144.46	53	137.60	117	134.00	164	145.35	41	142.62	57	171.72	25	158	35
I. Aspas	157.41	18	154.06	23	159.88	14	194.96	2	160.81	11	181.51	4	169.70	35	152	67
J. Vardy	156.91	19	148.78	38	152.46	23	176.63	12	157.17	14	172.78	7	173.74	20	160	29
W. Ben Yedder	156.60	20	150.76	33	154.76	19	157.94	45	152.61	19	151.20	38	169.70	35	150	89
A. Griezmann	156.29	21	154.84	21	146.18	40	156.10	50	153.42	16	160.13	21	171.72	25	171	13
D. Mertens	156.20	22	155.69	18	135.78	141	138.67	125	139.89	72	140.25	70	169.70	35	155	53
E. Džeko	156.07	23	155.42	19	142.29	77	150.96	67	148.61	32	154.11	27	167.68	48	155	53
G. Moreno	155.44	24	146.70	44	160.99	11	168.30	18	166.51	8	169.45	10	173.74	20	152	67
J. Sancho	155.30	25	163.77	9	158.13	15	193.14	4	151.94	22	166.76	14	175.76	15	155	53
R. Mahrez	154.73	26	161.50	12	154.17	20	182.36	8	152.37	20	166.02	16	173.74	20	156	44
M.Rashford	153.98	27	154.09	22	144.57	54	163.06	30	149.62	29	161.32	19	171.72	25	157	38
S. Aguero	153.97	28	145.66	47	143.55	66	140.38	116	139.20	75	134.00	103	175.76	15	178	9
R. Firmino	153.61	30	156.48	17	137.47	118	147.75	81	145.90	37	154.07	28	171.72	25	166	19
E. Cavani	153.44	32	146.10	46	154.86	18	160.90	37	145.12	43	143.23	54	171.72	25	152	67
M. Depay	152.98	34	147.21	43	161.08	10	179.91	10	158.79	12	166.94	13	171.72	25	153	63
A. Kramaric	150.15	41	139.06	72	155.66	17	150.58	70	150.65	26	144.09	52	167.68	48	146	139
Á. Di Maria	149.93	42	152.81	28	136.61	126	156.32	49	140.93	66	151.93	33	175.76	15	164	22
L. Insigne	149.66	45	147.48	42	146.07	41	146.51	87	152.67	18	151.97	32	173.74	20	160	29
W. Weghorst	148.06	49	141.96	65	156.00	16	169.54	17	151.82	23	155.77	25	167.68	48	140	239
E. Hazard	148.00	50	152.36	30	136.10	136	150.51	71	136.83	93	141.14	65	171.72	25	175	11
P. Dybala	147.22	51	142.28	62	134.00	166	135.52	147	136.69	95	135.39	94	175.76	15	168	18
E. Haaland	146.79	53	136.62	82	179.74	2	195.00	1	145.53	40	142.93	56	177.78	11	166	19
F. Chiesa	146.58	54	148.81	37	146.71	39	163.27	29	151.35	25	161.43	18	167.68	48	153	63
L. Sané	146.47	56	150.14	34	136.31	131	146.57	86	135.18	110	136.09	91	169.70	35	169	17
A. Silva	145.69	60	137.96	77	170.72	5	178.94	11	157.73	13	156.63	23	169.70	35	145	152
K. Coman	141.00	80	142.67	58	141.68	83	163.36	28	137.69	85	142.98	55	173.74	20	151	77
M. Diaby	136.58	128	137.95	78	141.35	87	167.64	19	136.21	99	142.03	60	163.64	87	145	152
D. Kamada	134.00	176	134.00	104	141.72	81	169.73	16	134.00	127	137.80	83	159.60	152	134	449

- Os dez melhores jogadores na base de dados do FIFA 22 estão todos englobados no top 10 de uma das estatísticas calculadas, à exceção do senegalês Sadio Mané e do sul-coreano Son Heung-Min, que estão presentes apenas no top 15. Isto demonstra que para avaliar a qualidade de um jogador há três perguntas que se precisam de fazer: o que é que o jogador já demonstrou, o que é que o jogador atualmente demonstra, e o que é que ele pode vir a demonstrar. Para esta comparação, a última pergunta não é feita, visto isso se tratar de previsões da sua qualidade, mas as outras perguntas estão representadas nos *ratings* calculados, ou seja, há jogadores que são considerados de alto nível porque se encontram numa forma recente muito boa, que é representada pelo seu valor de média móvel, e há jogadores cujo histórico de exibições os coloca na lista de melhores jogadores.
- Os resultados demonstram a diferença no significado do valor acumulado e do valor da média móvel, e isso está representado nos valor do norueguês Erling Haaland. Assinou a meio da época de 2019/2020 pelo Borussia Dortmund, pelo que os seus registos só começam a partir dessa época, o que coloca os seus dados acumulados em posições baixas em relação aos restantes. No entanto, as suas exibições têm sido bastante boas, sendo considerado o melhor jogador sub-21 do mundo em 2020 (Pisani, 2020), o que faz com que as suas exibições mais recentes ganhem um peso adicional e o tornem como um dos melhores atualmente. Um exemplo oposto é o do gabonês Pierre-Emerick Aubameyang, que tem 91 golos marcados no conjunto de dados, sendo inclusive o oitavo melhor marcador na lista inteira, mas como a sua última época foi a pior das quatro estudadas, e tendo marcado apenas 4 golos nos últimos 15 jogos, faz com que a sua média móvel de performance seja das mais baixas, sendo mesmo a pior se for a média sem golos.

Tabela 15 – Diferenças para os avançados entre os valores calculados e os valores do FM

Indicador	PRA Escala FM	PRSGA Escala FM	EWSAPR Escala FM	EWSAPRSG Escala FM	EWSAPRXM Escala FM	EWSAPRSGXM Escala FM
Soma das diferenças	370.43	464.29	711.67	803.92	598.16	600.83
Média das diferenças	8.42	10.55	16.17	18.27	13.59	13.66
Desvio padrão	7.86	9.11	10.63	10.79	10.08	10.23

Na Tabela 15 encontram-se listadas a soma e a média das diferenças absolutas, bem como o desvio padrão, da relação entre os *ratios* calculados e os dados representados no FM. O *ratio* com menores valores nos indicadores acabou por ser o PRA, seguido do PRSGA. Os piores foram os resultados das médias móveis. Com isto, pode-se concluir que, no caso de avançados, a performance aglomerada das últimas épocas acaba por influenciar na ideia que se tem da qualidade do jogador, enquanto para melhor determinar a forma atual de um jogador, a média móvel é o melhor fator.

A Tabela 16 apresenta os dados calculados para os médios, utilizando a mesma metodologia usada para os avançados e a mesma equação de normalização. Os resultados podem ser analisados da seguinte forma:

Tabela 16 – Valores de médios calculados

Jogador	PRA		PRSGA		EWSAPR		EWSAPRSG		EWSAPRXM		EWSAPRSGXM		FIFA 22		FM 2021	
Nome	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor1	Rank	Valor	Rank
T. Müller	188.00	1	188.00	1	188.00	1	188.00	1	188.00	1	188.00	1	175.76	6	164	10
N.Fekir	173.99	2	165.13	4	162.87	16	159.51	14	160.84	9	157.74	10	169.70	26	153	54
K. Bruyne	173.36	3	169.67	3	178.65	4	169.51	4	175.84	2	167.31	3	183.84	1	188	1
D. Silva	169.89	4	170.00	2	162.60	17	164.83	10	157.85	14	159.68	9	171.72	19	168	4
P. Gómez	166.18	5	162.43	5	155.40	36	151.53	44	156.53	17	152.63	16	171.72	19	157	33
P. Coutinho	163.27	6	152.63	10	161.63	19	153.39	35	153.12	24	147.05	41	165.66	51	163	15
B. Fernandes	162.17	7	144.82	20	180.58	2	165.30	8	162.09	7	152.28	18	177.78	4	165	8
B. Silva	160.87	8	156.59	6	159.56	24	163.45	11	158.33	13	162.00	4	173.74	12	171	2
H. Aouar	159.72	9	156.25	7	163.19	15	157.45	19	162.20	6	156.73	12	163.64	78	142	180
D. Alli	158.43	10	153.66	8	156.07	34	156.13	22	151.37	29	151.36	21	161.62	107	144	157
L. Goretzka	157.98	11	152.98	9	171.08	7	169.25	5	161.89	8	160.39	7	175.76	6	161	20
P. Pogba	157.59	12	151.04	11	153.88	42	151.29	48	151.64	28	149.32	27	175.76	6	167	6
P. Sarabia	156.07	13	144.67	21	152.88	51	144.77	106	148.09	44	141.72	72	161.62	107	148	105
S. Milinkovic-Savic	155.50	14	143.82	26	156.25	33	149.76	55	159.59	11	152.54	17	171.72	19	162	18
I. Gündogan	154.15	15	143.88	24	165.68	12	154.59	30	162.94	5	152.83	15	171.72	19	160	25
L. Alberto	153.79	16	142.66	29	147.44	89	138.52	241	148.34	42	139.30	103	169.70	26	161	20
P. Zielinski	152.77	17	146.62	15	169.21	10	166.37	7	171.37	3	168.47	2	163.64	78	154	49
C. Nkunku	152.28	18	148.63	12	166.53	11	166.52	6	152.27	26	152.09	20	163.64	78	138	283
A. Ramsey	150.18	19	145.51	18	148.99	70	148.83	64	141.03	89	140.75	82	161.62	107	151	76
F. Kessié	150.02	20	137.05	46	164.20	13	151.04	50	171.27	4	155.80	14	169.70	26	153	54
Casemiro	149.97	21	143.82	25	155.56	35	154.82	28	160.52	10	159.80	8	179.80	3	166	7
M. Sabitzer	149.49	22	139.01	37	155.30	37	144.67	108	154.04	21	144.15	57	169.70	26	149	95
L. Pellegrini	149.11	23	144.60	22	160.81	21	151.13	49	156.95	16	148.58	32	163.64	78	153	54
H. Çalhanoglu	148.89	24	140.16	34	153.52	46	148.37	71	154.00	22	148.90	30	165.66	51	152	63
J. Lingard	148.78	25	140.01	36	171.24	6	160.69	13	154.56	20	148.13	35	159.60	143	136	303
L. Modric	148.43	26	146.33	16	146.68	98	141.95	156	147.21	50	142.55	68	175.76	6	168	5
Fabián	148.34	27	144.84	19	146.45	104	142.79	142	148.41	41	144.55	54	165.66	51	155	41
P. Foden	147.80	30	140.60	33	179.18	3	172.31	3	149.26	38	146.26	48	169.70	26	152	63
J. Draxler	147.03	32	146.31	17	152.41	53	153.67	33	142.51	78	143.01	63	161.62	107	145	136
Rodri	146.41	33	146.76	14	152.31	54	155.03	26	157.33	15	160.74	5	173.74	12	156	37
T. Ndombele	146.17	34	147.08	13	144.02	141	145.07	101	142.65	76	143.55	61	165.66	51	149	95
D. Parejo	144.82	35	130.70	103	142.77	157	143.78	119	147.99	45	149.61	23	173.74	12	155	41
M. Llorente	144.69	36	138.04	43	170.48	9	165.17	9	153.63	23	150.35	22	173.74	12	152	63
J. Grealish	144.20	42	140.05	35	173.05	5	175.72	2	151.14	31	152.20	19	169.70	26	163	15
Fabinho	143.63	44	141.95	30	139.96	226	142.44	150	140.89	91	143.56	60	173.74	12	164	10
T. Kroos	143.43	46	138.90	39	153.83	43	154.99	27	155.10	18	156.36	13	177.78	4	170	3
C. Soler	143.25	47	135.90	51	158.54	27	146.20	91	158.68	12	146.54	45	163.64	78	150	85
Jorginho	142.50	48	134.78	64	146.49	103	140.45	195	149.39	37	142.66	67	171.72	19	146	128
S. Busquets	141.79	50	144.01	23	148.73	74	154.54	31	151.24	30	157.73	11	173.74	12	162	18
M. Mount	141.55	55	135.43	58	159.85	23	156.17	21	148.96	39	146.49	46	167.68	39	154	49
D. Klaassen	141.29	59	135.73	53	161.84	18	155.95	24	150.98	32	146.98	42	159.60	143	141	199
T. Alcántara	140.03	63	137.96	44	143.82	143	143.82	117	141.20	87	141.14	77	173.74	12	161	20
B. Saka	138.16	76	134.50	68	157.06	29	157.22	20	143.94	68	143.79	59	161.62	107	145	136
Koke	137.96	77	130.77	101	135.54	381	134.89	374	137.22	145	136.64	148	171.72	19	163	15
N. Kanté	137.89	78	134.78	65	139.37	243	141.72	166	140.40	95	142.96	65	181.82	2	165	8
W. Ndidi	137.61	80	136.43	50	153.57	44	159.40	15	154.70	19	160.74	6	171.72	19	152	63
M. Verratti	136.97	90	138.90	38	149.58	67	155.97	23	144.90	63	149.58	24	175.76	6	161	20
A. Golovin	136.27	98	130.33	106	161.17	20	158.45	18	145.98	57	144.36	55	159.60	143	147	116
M. Pjanic	135.66	107	129.41	112	133.00	486	133.00	437	133.00	305	133.00	270	165.66	51	164	10
F. Jong	135.40	112	133.13	81	148.80	72	151.30	47	143.12	70	144.71	51	175.76	6	164	10
Fernando	135.07	114	132.65	86	158.58	26	159.06	17	147.41	48	147.52	37	169.70	26	144	157
K. Laimer	134.86	119	132.34	90	154.20	41	159.08	16	144.73	64	147.49	38	163.64	78	136	303
Fernandinho	134.67	121	132.05	95	138.82	257	141.88	158	138.83	112	141.76	71	167.68	39	164	10
L. Paquetá	134.36	125	128.17	127	170.55	8	161.40	12	145.42	60	141.62	74	163.64	78	140	218
J. Willock	133.00	134	122.72	229	163.65	14	146.34	87	139.95	101	134.42	200	151.52	391	133	428

- O alemão Thomas Muller é o melhor médio em todos os *ratios* calculados, o que comprova a sua polivalência ofensiva e o seu estilo de jogo, que é a de procurar de forma mais eficaz o golo, tendo até cunhado o termo *Raumdeuter* (alemão para "analisador de espaços") para explicar a sua peculiar forma de jogar (Harding, 2021). Além disso, tem sido extremamente regular e sem picos de forma, fazendo com que também a sua média móvel esteja sempre nos valores mais altos.
- O melhor médio em ambos as bases-de-dados de comparação é o belga Kevin de Bruyne, e os seus *ratios* figuram sempre nos cinco melhores. Uma justificação possível é o número de lesões que De Bruyne teve nas últimas épocas que o impediram de manter o alto nível de performance.
- Nabil Fekir, natural de França, é dos melhores jogadores identificados por todos os *ratios*, fazendo parte do top 20 de todos e considerado o segundo melhor no primeiro índice, em comparação com as bases-de-dados, em que aparece em posições mais modestas. O jogador francês é o segundo melhor marcador da lista em estudo com 42 golos e 10 penaltis sofridos, o dobro do segundo melhor.
- O modelo tem preferência por jogadores de índole mais ofensiva, e isso percebe-se com a presença nas posições mais altas de médios com muitas participações atacantes, contrapondo com médios mais de equilíbrio ou puramente defensivos, que aparecem em posições mais baixas, exemplificado com casos como o de Jorginho, vencedor do prémio de melhor jogador do ano de 2021 (Shread, 2021), ou de N'Golo Kanté, considerado o melhor médio da edição de 2020/2021 da UEFA Champions League (UEFA.com, 2021), que não estão em nenhum top 40.
- O inglês Joe Willock e o bósnio Miralem Pjanic são dois exemplos que representam pontos extremos das avaliações analíticas e empíricas, respetivamente. O médio inglês tem valores modestos em todos os registos comparados, excetuando o da média móvel com golos, em que alcançou o 14º lugar. Essa posição alta é explicada pelo seu final de época de 2020/2021, em que marcou sete golos em sete jogos seguidos, alcançando uma média de um golo a cada 55.6 minutos. Já o médio bósnio era considerado dos melhores na base de dados do Football Manager 2021, tendo como base os anos como jogador influente na equipa italiana Juventus, mas a sua forma mais recente, em que não fez qualquer golo ou assistência ao serviço do Barcelona em 2020/2021, coloca-o como um dos piores em análise em todos os *ratios*. O jogador, entretanto, é dos três desta lista, juntamente com Pablo Sarabia e Davy Klaassen, que já não se encontra num clube dos cinco campeonatos principais europeus.

Tabela 17 – Diferenças para os médios entre os valores calculados e os valores do FM

Indicador	PRA Escala FM	PRSGA Escala FM	EWSAPR Escala FM	EWSAPRSG Escala FM	EWSAPRXM Escala FM	EWSAPRSGXM Escala FM
Soma dos valores	614.17	828.65	761.15	683.84	566.20	570.22
Média dos valores	11.17	15.07	13.84	12.43	10.29	10.37
Desvio padrão	8.73	8.41	8.88	8.08	7.97	7.62

Na Tabela 18 estão tabelados os indicadores para validar as diferenças entre os valores calculados e os disponibilizados pela base de dados do FM. Os melhores indicadores foram a multiplicação das médias móveis pelos minutos jogados, ao contrário do acumulado do PRA, que tem a pior média de diferenças. Os desvios padrões, no entanto, revelam que os valores calculados ficam relativamente perto dos valores em comparação.

Os dados calculados para os defesas estão apresentados na Tabela 18, com recurso à mesma metodologia utilizada na análise dos médios e avançados e a mesma equação de normalização. Os resultados podem ser analisados da seguinte forma:

- O alemão Joshua Kimmich é o melhor em ambos os *ratios* acumulados, o que condiz com as métricas de comparação, em que ele está nos melhores três em cada uma.
- O holandês Virgil Van Dijk, o melhor quer no FM 2021, quer no FIFA 22, e o segundo melhor jogador do mundo em 2019 para a revista L'Equipe (Dutton, 2019), encontra-se no top 12 de todos os *ratios*, mas em nenhum deles chega aos três melhores. Há duas justificações para o defesa holandês não estar mais acima na tabela, sendo a primeira o facto de ele só ter oito jogos em 2020/2021 devido a uma lesão grave, e a segunda a ser a sua falta de números ofensivos em comparação com outros nomes.
- O português Raphael Guerreiro e o colombiano Juan Cuadrado surgem como os melhores nos *ratios* de média móvel com e sem golos, respetivamente. Ambos os defesas fizeram a sua melhor época ao serviço dos seus clubes na última época em análise, com 6 golos e 10 assistências para o português, e 2 golos e 16 assistências para o colombiano. Aliás, Juan Cuadrado esteve em 8 contribuições nos últimos 15 jogos da época 2020/2021, terminando assim a temporada em boa forma.
- Conforme verificado na análise dos médios, os pesos calculados e utilizados para a avaliação dos jogadores valorizam mais o contributo ofensivo do que defensivo, e isso verifica-se na baixa posição nos rankings de defesas que têm posições altas no FM e FIFA, e que foram reconhecidos como dos melhores na sua posição pela crítica. O português Rúben Dias não passa dos 350 melhores em todos os rankings e foi considerado o melhor jogador da época 2020/2021 da Premier League (Kerai, 2021), os italianos Leonardo Bonucci e Giorgio Chiellini e o senegalês Kalidou Koulibaly fazem parte dos trinta nomeados para melhor jogador do mundo de 2021 da revista L'Equipe e nenhum figura nos cem melhores de qualquer ranking (Lara, 2021).
- O inglês Trent Alexander-Arnold não chega a nenhum dos top 100 de qualquer *rating*, o que não vai ao encontro aos seus valores nas bases de dados de comparação e ao seu palmarés individual, tendo sido o melhor jogador jovem e melhor defesa direito da liga inglesa em 2019/2020 com 13 assistências para golo (BBC, 2020). No entanto, o defesa direito é alvo de um dos problemas identificados na secção de avaliação do modelo: não haver uma divisão nos dados de cruzamentos acertados e falhados. Encontra-se em terceiro lugar na lista de jogadores com mais cruzamentos realizados, com 647, e isso reduz os seus *ratios* drasticamente. O jogador com mais cruzamentos é o sérvio Filip Kostic, o melhor nos dois índice que multiplicam a média móvel pelos minutos, só que o seu número elevado de cruzamentos é compensada com 25 golos e 41 assistências, enquanto Trent Alexander-Arnold tem 8 golos e 38 assistências.

Tabela 18 – Valores de defesas calculados

Jogador	PRA		PRSGA		EWSAPR		EWSAPRSG		EWSAPRXM		EWSAPRSGXM		FIFA 22		FM 2021	
Nome	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank	Valor	Rank
J. Kimmich	180.00	1	180.00	1	166.52	7	157.45	10	178.63	2	175.09	3	179.80	1	171	3
A. Hakimi	164.05	2	155.43	4	173.90	4	169.70	4	166.25	4	170.32	4	171.72	16	157	23
J. Alba	161.06	3	163.39	2	135.66	93	136.92	82	141.76	26	147.63	20	173.74	8	161	12
J. Cuadrado	157.20	4	156.61	3	179.98	2	180.00	1	167.87	3	176.62	2	167.68	39	159	20
M. Cornet	156.63	5	144.34	8	151.17	26	149.55	21	135.99	45	139.92	45	151.52	428	136	282
R. Guerreiro	155.84	6	143.38	9	180.00	1	166.47	6	163.24	6	159.56	9	169.70	23	154	37
R. Gosens	154.11	7	138.12	15	161.41	11	139.73	51	157.33	12	141.30	39	167.68	39	150	56
F. Kostic	153.01	8	135.73	19	166.55	6	160.09	9	180.00	1	180.00	1	169.70	23	141	172
V. van Dijk	152.61	9	148.94	7	159.63	12	153.14	12	163.50	5	162.90	7	179.80	1	180	1
S. Ramos	151.64	10	133.02	31	162.80	10	139.71	52	161.88	7	143.27	32	177.78	3	174	2
S. de Vrij	148.01	11	143.38	10	135.19	100	138.14	68	142.58	21	150.91	15	171.72	16	162	10
A. Robertson	146.92	12	149.43	5	122.36	359	126.51	294	127.35	99	135.59	66	175.76	4	160	16
S. Roberto	145.12	13	149.36	6	139.63	65	143.02	37	133.98	54	141.61	38	163.64	81	158	21
T. Hernández	143.86	14	136.24	18	169.25	5	152.13	15	161.60	8	152.09	12	169.70	23	155	31
Marquinhos	143.63	15	134.91	23	157.81	15	138.27	65	160.09	9	143.69	30	175.76	4	163	7
L. Vázquez	143.51	16	140.04	14	157.36	16	160.26	8	140.24	29	148.21	18	163.64	81	143	141
N. Elvedi	143.17	17	140.68	13	152.97	22	151.32	16	158.53	11	162.99	6	159.60	138	138	227
H. Maguire	142.64	18	141.09	11	133.04	130	136.66	84	142.50	22	151.88	13	169.70	23	152	46
C. Azpilicueta	141.67	19	140.97	12	125.24	277	127.50	259	131.95	63	137.84	57	167.68	39	143	141
M. Alonso	141.12	20	129.76	57	153.14	21	141.81	42	145.69	17	140.50	44	159.60	138	144	126
J. Tarkowski	138.10	24	137.70	16	130.50	168	131.01	173	134.80	50	138.79	49	163.64	81	135	320
H. Hateboer	137.64	25	133.98	28	141.84	49	145.00	29	141.84	25	150.32	16	159.60	138	145	112
G. Piqué	137.52	26	132.78	34	132.17	143	135.10	107	135.99	46	143.45	31	169.70	23	162	10
M. Hummels	135.81	30	132.82	33	141.75	51	136.30	89	144.61	18	142.71	35	173.74	8	156	27
G. Di Lorenzo	135.57	31	129.36	64	157.11	17	152.85	13	159.09	10	160.92	8	161.62	102	148	90
J. Evans	135.01	35	137.12	17	151.13	27	150.49	18	151.19	13	156.48	10	161.62	102	149	76
C. Romero	134.17	41	132.94	32	153.80	20	155.89	11	142.41	23	149.93	17	165.66	58	144	126
L. Dunk	133.95	43	132.68	36	143.20	42	132.17	147	150.93	14	140.87	41	157.58	184	144	126
F. Acerbi	133.84	45	135.71	20	120.66	411	128.91	222	125.94	108	140.64	42	167.68	39	155	31
R. Baku	131.56	54	126.62	96	174.12	3	165.24	7	147.04	16	147.67	19	161.62	102	135	320
J. Bernat	131.31	57	129.48	62	154.35	18	151.31	17	132.36	61	135.52	67	165.66	58	150	56
K. Walker	130.34	67	132.52	39	120.51	421	123.70	384	122.01	162	128.02	123	171.72	16	155	31
M. Škriniar	129.64	74	129.57	60	124.48	299	121.65	453	129.39	85	127.45	136	173.74	8	161	12
A. W. Bissaka	129.34	77	131.30	44	145.86	35	152.48	14	149.55	15	163.38	5	167.68	39	145	112
D. Alaba	129.28	80	125.96	107	141.22	55	146.22	26	143.83	19	154.86	11	169.70	23	171	3
D. Carvajal	129.28	81	131.23	45	133.52	122	139.78	50	129.58	83	139.57	47	171.72	16	153	41
R. Varane	128.09	99	126.64	95	127.73	225	129.03	219	131.36	71	136.00	63	173.74	8	165	6
T. Silva	127.31	110	127.80	79	126.56	253	127.97	245	124.30	123	128.65	120	171.72	16	156	27
M. de Ligt	127.11	113	123.34	135	145.27	36	145.84	28	130.08	79	135.07	68	171.72	16	163	7
J. Cancelo	126.29	121	124.46	117	146.36	33	144.69	31	140.44	28	143.91	27	173.74	8	157	23
R. Bensebaini	126.07	127	117.87	251	149.24	30	129.33	208	143.45	20	128.95	116	159.60	138	141	172
T. A. Arnold	125.51	138	120.35	187	114.21	634	110.00	822	114.96	321	110.00	750	175.76	4	160	16
Djené	125.07	144	128.25	70	131.57	153	139.42	55	136.87	42	151.35	14	163.64	81	147	100
D.Alves	124.82	148	125.83	109	165.46	8	171.35	2	127.32	100	134.90	69			147	100
K. Koulibaly	123.77	165	121.41	171	110.00	773	113.62	723	110.00	549	115.09	461	173.74	8	161	12
A. Laporte	121.77	204	118.98	227	116.73	541	117.26	587	114.76	328	117.02	382	173.74	8	170	5
G. Chiellini	121.25	214	122.57	150	119.66	448	128.51	227	113.45	382	122.16	230	173.74	8	160	16
L. Bonucci	119.79	245	112.38	482	115.52	589	115.14	665	116.75	272	117.16	371	171.72	16	163	7
W. Endo	118.41	288	117.02	279	154.17	19	149.58	20	118.69	227	121.47	242	157.58	184	132	453
J. Giménez	117.83	304	115.27	341	120.21	429	121.95	443	115.96	293	119.80	288	169.70	23	161	12
Ó. Mingueza	114.21	450	112.99	450	165.35	9	168.40	5	117.21	256	122.26	227	151.52	428	130	570
J. Justin	113.21	507	112.83	462	147.10	31	149.83	19	119.01	218	124.19	191	155.56	240	143	141
R. Dias	113.03	517	113.67	411	119.92	441	123.54	386	112.36	427	117.19	367	175.76	4	160	16
J. Mæhle	112.86	524	113.47	421	159.12	13	170.71	3	118.53	231	126.23	154	155.56	240	137	260
A. Truffert	110.00	694	108.95	698	157.89	14	143.52	35	115.91	294	116.62	392	151.52	428	110	3789

Tabela 19 – Diferenças para os defesas entre os valores calculados e os valores do FM

Indicador	PRA Escala FM	PRSGA Escala FM	EWSAPR Escala FM	EWSAPRSG Escala FM	EWSAPRXM Escala FM	EWSAPRSGXM Escala FM
Soma dos valores	924.15	999.43	1010.78	966.20	911.05	819.32
Média dos valores	21.00	22.71	22.97	21.96	20.71	18.62
Desvio padrão	13.46	13.84	15.33	14.44	15.62	14.09

A Tabela 19 contém os indicadores igualmente utilizados na análise de médios e avançados. Os valores demonstram que o modelo tem a pior eficácia nos defesas, em comparação com as restantes posições, o que ajuda a sustentar a hipótese de que o modelo valoriza e identifica melhor os jogadores que contribuem ofensivamente, ao invés dos que contribuem defensivamente. A média de diferença ronda as 20 unidades e o desvio as 14 unidades, com a melhor média a ser a do EWSAPRSGXM e o menor desvio a ser do PRA.

Feita esta análise por posição, e verificando que os índices acabam por ter diferenças para os valores de comparação semelhantes entre si, decidiu-se passar a utilizar apenas o PRA e o EWSAPR para os próximos testes de validação e para a implementação.

5.7 Análise de Resultados

Com vista a validar os resultados obtidos a partir da aplicação do modelo nos dados registado, um questionário foi redigido e enviado a 11 pessoas com participação ativa no futebol de escalão sénior. Desses 11, participaram membros de equipas profissionais como Leixões ou Arouca e observadores profissionais de agências de carreiras de jogadores, como a Team of Future. A nível de cargo, responderam quatro observadores, três jogadores federados, dois estudantes em cursos de treinador de associações de futebol, um treinador e um membro de equipa técnica, conforme se vê na Figura 40. Uma cópia do questionário pode ser vista no Anexo 5.

O questionário consiste em 18 perguntas de escolha múltipla, em que é solicitada a escolha do melhor jogador de um par de jogadores ou indicar que são de qualidade semelhante, caso seja essa a opinião da pessoa a responder. Os 18 pares de jogadores foram escolhidos de acordo com o seguinte procedimento, inspirado no processo de validação usado pelo artigo científico referido na secção 2.5.1.1:

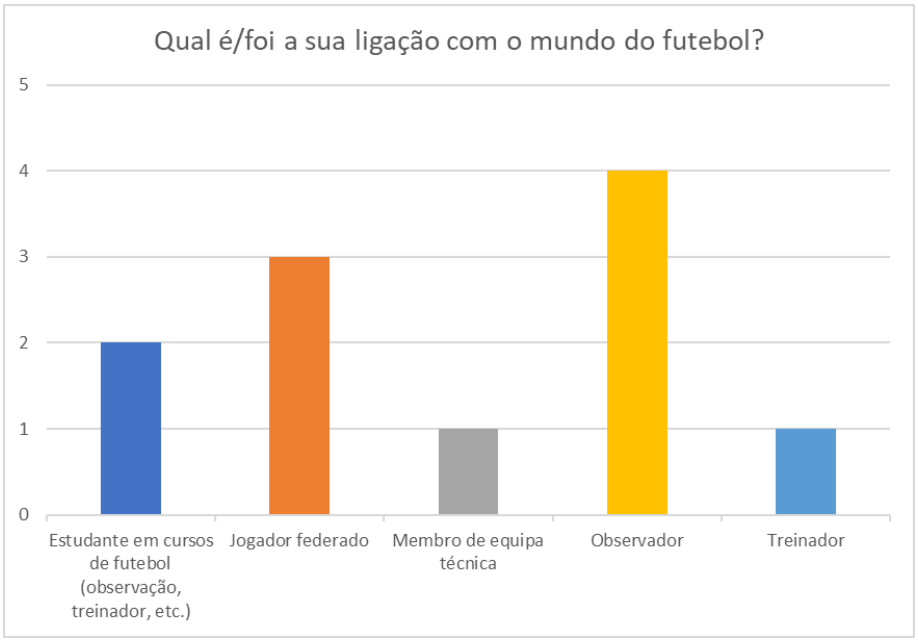


Figura 40 – Inquiridos do questionário por cargo

1. Os dados são divididos em três subconjuntos: defesas, médios e avançados. Os jogadores em cada conjunto devem ter pelo menos 1350 minutos e 30 jogos disputados, e com os ratios a analisar, PRA e EWSAPR, acima de 0, para evitar situações de jogadores com poucos jogos e hipoteticamente de baixa qualidade.
2. Para cada conjunto, são escolhidos 50% dos jogadores de forma aleatória, criando um subconjunto da posição.
3. Criar dois novos subconjuntos, um ordenado pelo índice de PRA, e outro pelo índice de EWSAPR.
4. Para cada novo subconjunto, é escolhido um elemento aleatoriamente da posição 21 até à posição n-21, sendo n o número total de elementos no subconjunto.
5. Um número aleatório *index_gap* de 5 a 15 é escolhido.
6. São escolhidos dois jogadores, um que se encontre abaixo o número de posições equivalente a *index_gap*, e outro acima.
7. Repetir o processo até se obter 18 pares, 6 por posição e, em cada posição, 3 por índice a avaliar.

Depois de as 11 pessoas comunicadas terem respondido ao questionário, as respostas foram compiladas na Tabela 20, com cada jogador a ter o seu valor no FM ao lado numa coluna cinzenta. Os jogadores listados na coluna Jogador 1 foram os jogadores identificados pelo modelo como os melhores do par e os listados na coluna Jogador 2 como os piores.

Tabela 20 – Respostas ao questionário

P	Jogador 1	FM	Jogador 2	FM	Índice a Comparar	Posição	Votos 1	Igual	Votos 2	Diferença	Melhor FM 2021
1	David García	129	Scott Dann	126	PRA	Defesa	7	2	2	5	David Garcia
2	Rodrigo Becão	133	Matthias Zimmermann	126	PRA	Defesa	7	2	2	5	Rodrigo Becão
3	Renato Tapia	132	Julian Chabot	128	PRA	Defesa	10	1	0	10	Renato Tapia
4	Roger Ibanez	145	José Fonte	139	EWSAPR	Defesa	3	1	7	-4	Roger Ibanez
5	Joris Gnagnon	135	Leander Dendoncker	141	EWSAPR	Defesa	0	0	11	-11	Leander Dendoncker
6	Christian Kabasele	133	Tin Jedvaj	127	EWSAPR	Defesa	4	3	4	0	Christian Kabasele
7	Matteo Scozzarella	126	Nicolás Domínguez	131	PRA	Médio	0	0	11	-11	Nicolás Domínguez
8	Bruno Guimarães	139	Renato Sanches	143	PRA	Médio	6	1	4	2	Renato Sanches
9	Axel Witsel	152	Antonín Barák	136	PRA	Médio	11	0	0	11	Axel Witsel
10	Kerem Demirbay	145	Steven Davis	134	EWSAPR	Médio	9	0	2	7	Kerem Demirbay
11	Koke	163	Jozabed	125	EWSAPR	Médio	11	0	0	11	Koke
12	Sebastián Cristóforo	121	Samuel Moutoussamy	117	EWSAPR	Médio	7	3	1	6	Sebastián Cristóforo
13	Raúl García	156	Dodi Lukebakio	134	PRA	Avançado	8	1	2	6	Raúl García
14	Patrik Schick	140	Wahbi Khazri	135	PRA	Avançado	9	0	2	7	Patrik Schick
15	Marcus Rashford	157	Ángel Correa	154	PRA	Avançado	5	5	1	4	Marcus Rashford
16	Edin Džeko	155	Amine Gouiri	130	EWSAPR	Avançado	8	0	3	5	Edin Džeko
17	Shon Weissman	122	Angelo Fulgini	130	EWSAPR	Avançado	4	4	3	1	Angelo Fulgini
18	Riccardo Meggiorini	121	Lucas Pérez	139	EWSAPR	Avançado	1	1	9	-8	Lucas Pérez
						Total	110	24	64	46	

Numa primeira análise, é possível verificar que a maioria das respostas foram ao encontro dos jogadores cujos ratings eram os maiores do par, com 110 votos de 198 de possíveis. 64 votos foram para os jogadores que não foram identificados como os melhores e 24 votos foram para situações em que o inquirido considerou o par de qualidade semelhante.

A Tabela 21 apresenta as percentagens de acerto entre as respostas dos inquiridos, os resultados do modelo e os valores presentes no Football Manager 2021. Quatro tipos de percentagens foram identificados:

- Percentagens Globais – Efetuada comparações entre as respostas e os valores globais sem o tipo de indicador diferenciado
- Percentagens PRA e Percentages EWSAPR – As respostas são comparadas com os indicadores de forma distinta, com o objetivo de perceber qual deles foi o melhor.
- Percentagens FM – Validar as respostas dos inquiridos e os valores do modelo

Tabela 21 – Percentagens de acerto

Percentagens Globais	
Correspondência entre resposta e índices	55.56%
Correspondência sem “Iguais”	63.22%
Correspondência com "Iguais" para modelo	67.68%
Correspondência com jogador com mais votos	72.22%
Percentagens PRA	
Correspondência entre resposta e índice PRA	63.64%
Correspondência PRA sem "Iguais"	72.41%
Correspondência PRA com "Iguais" para modelo	75.76%
Correspondência com jogador com mais votos	88.89%
Percentagens EWSAPR	
Correspondência EWSAPR	47.47%
Correspondência EWSAPR sem "Iguais"	54.02%
Correspondência EWSAPR com "Iguais" para modelo	59.60%
Correspondência com jogador com mais votos	55.56%
Percentagens com FM	
Correspondência de modelo com FM	72.22%
Correspondência de respostas com o FM	64.14%
Correspondência de respostas com iguais com o FM	73.23%
Correspondência com jogador com mais votos	77.78%

Verifica-se que, globalmente, os inquiridos concordaram em 72.22% com os valores do modelo, o que revela a precisão dos índices em relação à opinião generalizada de quem participa no mundo do futebol. Se contabilizarmos os votos totais e não apenas uma escolha de quem teve mais votos dos dois, a correspondência desce para 55.56%, mas se forem retirados os votos na opção de que os jogadores têm qualidade semelhante, a percentagem sobre para 63.22%, continuando a subir para 67.68% se os votos de qualidade semelhante forem contabilizados como votos a favor do modelo.

Na divisão por tipo de índice, o com melhor resultados é claramente o PRA, com 88.89% dos jogadores com mais votos a corresponderem com o jogador identificado como melhor, verificando-se uma diferença grande para o índice EWSAPR, em que a correspondência esteve apenas nos 55.56%.

Por fim, quer o modelo, quer as respostas dos inquiridos apresentam percentagens semelhantes de relação com os valores presentes no FM 2021. 72.22% dos melhores jogadores identificados pelo modelo e 77.78% dos jogadores escolhidos pelos inquiridos são considerados de qualidade superior no FM, sendo que no caso dos inquiridos a percentagem de correspondência desce caso sejam considerados os votos todos e não apenas a escolha do com o maior número de votação.

A avaliação efetuada na secção 5.6, bem como a validação com a opinião de pessoas ligadas ao negócio, permitiu verificar que o modelo escolhidos e consequentes pesos ajudam a decifrar os melhores jogadores de forma global.

A divisão por posições ajudou a perceber as atuais limitações, já que nos avançados os valores calculados são os mais próximos verificados, inclusive com algumas situações em que se verificou a desatualização dos dados em comparação, nomeadamente do FM 2021. No entanto, os índices de qualidade não mantiveram a qualidade nas posições de médios e defesas, especialmente nos últimos, em que jogadores considerados pela especialidade como dos melhores do mundo não figuravam sequer nos melhores 100 de todos os índices analisados. Foi também possível verificar a diferença de significado que cada índice representava, com o PRA a indicar jogadores que foram consistentemente bons ao longo de vários jogos e épocas efetuadas, e o EWSAPR a valorizar jogadores em bons momentos de forma, mesmo com menos jogos que outros jogadores de qualidade semelhante. A importância dos golos acabou por apenas demonstrar que a estatística, apesar de ser importante e poder ajudar a distinguir um jogador bom de um jogador muito bom, não altera de forma significativa os resultados.

Por sua vez, o inquérito validou de forma efetiva os índices calculados e que podem funcionar de forma bastante satisfatória numa tarefa de comparação entre dois jogadores de valor equivalente.

5.8 Deployment

A fase de *deployment*, além da escrita do presente documento, corresponde ao desenvolvimento de uma aplicação web alimentada pelos dados obtidos e pelo modelo criado nas secções anteriores que vá ao encontro dos requisitos indicados na secção 4.1.

Com base no discutido no capítulo de Model Evaluation e no de Análise de Resultados, escolheu-se para implementar os dois índices Player Ratio Acumulado e EWSA Player Ratio, que irão servir dois propósitos:

- O PRA servirá como indicador da qualidade geral do jogador;
- O EWSAPR terá como propósito indicar os jogadores em melhor forma.

Conforme referido na secção 1.3, o principal foco do documento prende-se com criação e a avaliação de um modelo que alimentasse uma aplicação web, servindo como prova de conceito para um desenvolvimento mais detalhado futuramente. Como tal, a atual aplicação web tem apenas dois ecrãs simples onde se mostra a informação puramente de jogadores. Foi utilizada uma interface gráfica baseada nas soluções já existentes da Microsoft (Microsoft, 2020).

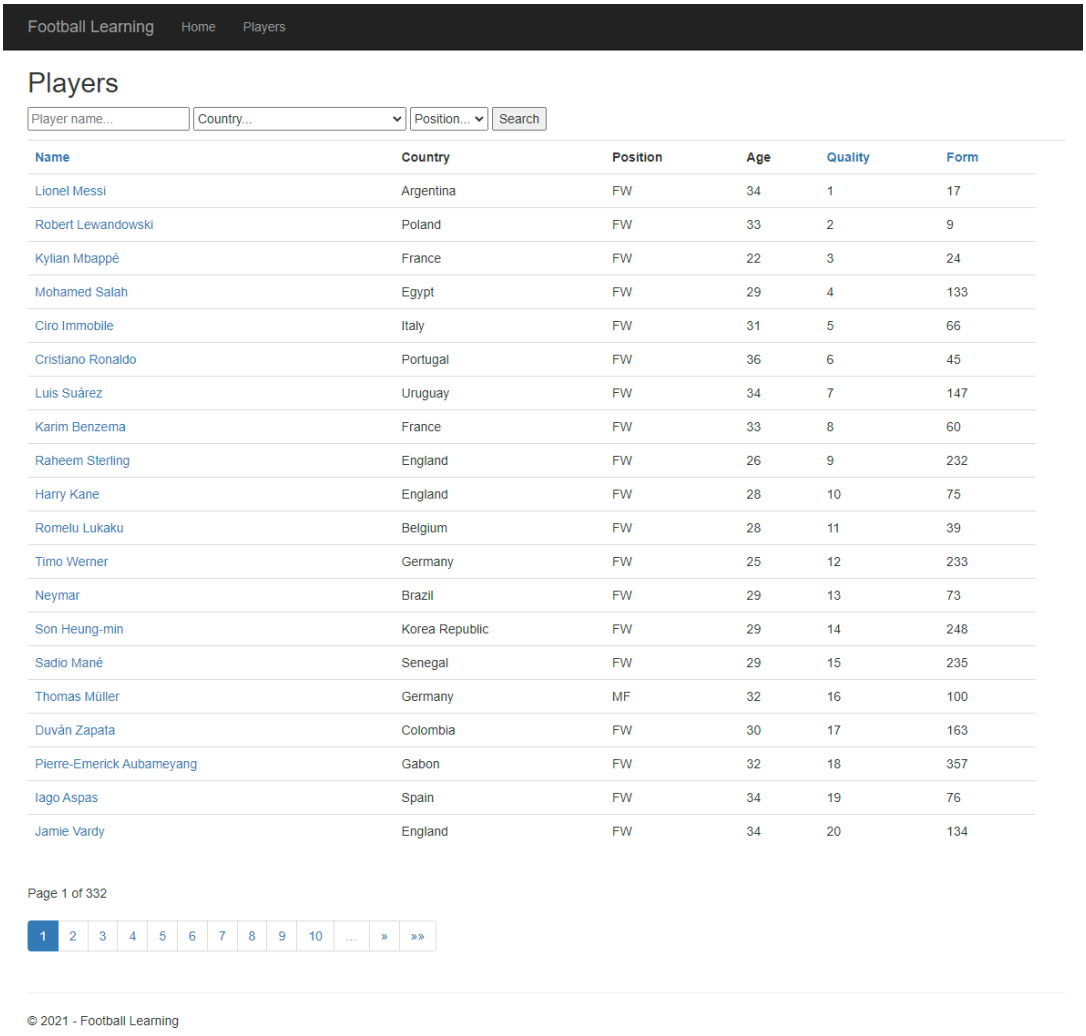


Figura 41 – Ecrã de listagem de jogadores

A Figura 41 apresenta a listagem de jogadores. A informação listada em cada jogador corresponde ao nome, país, posição, idade, posição no ranking de qualidade e posição no ranking da forma atual. É possível também ordenar a lista por nome e pelos dois índices de qualidade, além de ser possível filtrar por nome, país e posição.

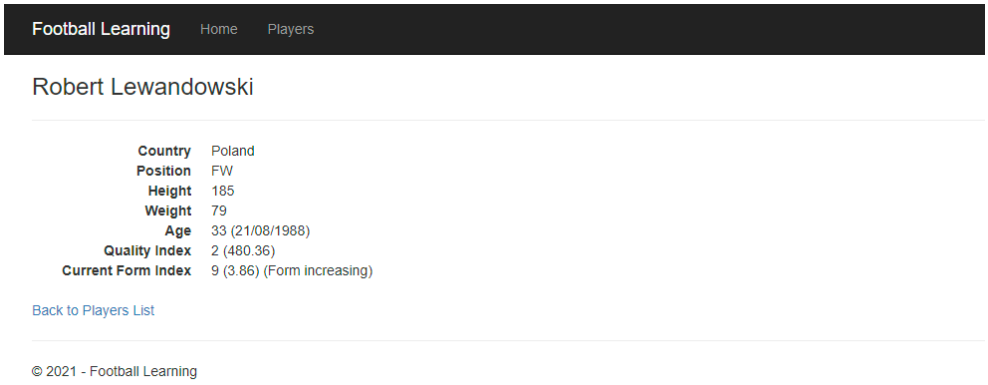


Figura 42 – Página de detalhe de um jogador

Na Figura 42 encontra-se demonstrado a página associada ao perfil de um jogador, neste caso do jogador polaco Robert Lewandowski. Além da informação indicada na listagem, a página de detalhe apresenta também o peso, a altura e a idade com a data de nascimento. Para os dois índices vindos do modelo, é colocado entre parênteses o valor bruto, e no caso do índice de forma, é indicado se o jogador está a subir de forma ou não. Para isso, é calculada a média de performance dos últimos cinco jogos do jogador e, caso essa média seja igual ou maior que o valor bruto da forma, é apresentada uma mensagem a indicar que a forma está a subir. Se a média for menor, a mensagem indica o oposto, que a forma está a descer.

6 Conclusão

O principal objetivo deste projeto consistiu no estudo aplicado ao futebol da utilização de ferramentas de data mining e *machine learning* de forma a obter mais informação auxiliar para as tarefas de gestão e prospeção de uma equipa profissional na escolha dos melhores jogadores. Esse objetivo pode ser dividido em dois, no qual o primeiro foi a criação e validação de um modelo de previsão de resultados com o propósito de retirar os pesos calculados e aplicá-los aos jogadores presentes na base de dados, e o segundo criar uma aplicação web alimentada pelo modelo e pelos cálculos obtidos.

Foi escolhido o modelo de regressão linear de previsão de resultados, que obteve uma percentagem de acerto de 76.70%, e os pesos retirados desse modelo eram compreensíveis e iam ao encontro do que foi estudado no capítulo do Contexto e Estado de Arte. Os índices de qualidade calculados com base nesses pesos foram comparados com bases de dados externas, no qual se verificaram resultados prometedores, e com opiniões de pessoas ligadas ao futebol, com resultados bastante positivos no inquérito realizado, em que se obteve uma taxa de correspondência de 72.22% entre os jogadores mais votados e os com mais respostas.

A aplicação web funciona adequadamente para uma primeira fase, com a possibilidade de procurar pelos melhores jogadores e pelos que se encontram em melhor forma. A página de perfil também apresenta a informação resultante dos modelos estudados.

6.1 Limitações e melhorias futuras

O trabalho realizado, apesar de cumprir os objetivos propostos e demonstrar potencial, deixa algumas limitações e eventuais melhorias que podem ser exploradas no futuro. A seguinte lista enumera alguma das identificadas ao longo do projeto:

- Criar um modelo de previsão cujo foco seja a parte defensiva do jogo. O modelo criado tinha como objetivo prever a equipa vencedora, no entanto isso resultou em

pesos de variáveis com tendência para valorizar mais as estatísticas ofensivas. Como tal, um modelo que tentasse prever o número de golos sofridos ou se a equipa não sofreu golos poderia resolver esse problema.

- Os guarda-redes foram ignorados devido às estatísticas serem vastamente diferentes das dos jogadores de campo. Seria preciso estudar a melhor forma de criar um modelo que consiga prever da melhor forma possível o impacto de um guarda-redes no resultado da sua equipa.
- O modelo de previsão de resultados, apesar de ter resultados bastante adequados, pode ser melhorado. Atualmente, o modelo prevê apenas se uma equipa venceu ou se não venceu o jogo, pelo que modificar o modelo para prever o desfecho real do jogo (vitória, empate ou derrota) ou, em associação ao modelo referido no ponto anterior de previsão de golos sofridos, modificar o modelo para prever o número de golos marcados.
- Utilizar mais dados de jogadores e competições. O atual repositório de dados já é vasto o suficiente para provar que o conceito funciona, no entanto, numa expectativa de tornar isto num produto comercial, os dados utilizados precisam de ser mais vastos, de forma a abranger outras competições e jogadores. Um dos *pain relievers* referidos na secção 3.3 referia o acompanhamento de campeonatos periféricos, e para satisfazer essa necessidade, o modelo e a aplicação precisam de ter muitos mais dados.
- Os algoritmos usados permitiram analisar duas abordagens diferentes para resolver o objetivo e de interpretar melhor os resultados obtidos. Contudo, é possível que alguns dos algoritmos não abordados neste documento possam ter melhores performances que o modelo final, quer seja a nível do valor da precisão, quer seja a nível dos pesos obtidos, pelo que testar outros algoritmos de *machine learning* é algo que seria interessante para o futuro, nem que sirva apenas como outra validação para o modelo atual.
- A aplicação web cumpre os requisitos mínimos de listagem e visualização de perfis de jogadores, mas não vai ao encontro dos requisitos levantados na secção 4 de Análise e Design. Como tal, torna-se fulcral no futuro melhorar a aplicação para responder aos requisitos, bem como criar novas funcionalidades e afinar a arquitetura e o design estrutural da aplicação.

Referências

Agresti, A., 2012. Categorical Data Analysis. John Wiley & Sons.

Andras, K., Havran, Z., 2015. New business strategies Of football clubs. Appl. Stud. Agribus. Commer. 9, 67–74. https://doi.org/10.19041/APSTRACT/2015/1-2/13

Andreff, W., 2008. Globalisation of the sports economy, Rivista di diritto ed Economia dello Sport, vol. IV, fasc. 3, 2008, pp. 13-32. Riv. Dirit. Ed Econ. Dello Sport 4, 13–32.

Arthur, C., 2013. Tech giants may be huge, but nothing matches big data [WWW Document]. The Guardian. URL http://www.theguardian.com/technology/2013/aug/23/tech-giants-data (accessed 2.19.21).

Assouline, G., 2018. Noté 10/10 dans "L'Équipe", Neymar sifflé par une partie du Parc des Princes [WWW Document]. Le HuffPost. URL https://www.huffingtonpost.fr/2018/01/18/neymar-note-10-10-dans-lequipe-siffle-par-une-partie-du-parc-des-princes-pendant-psg-dijon_a_23336593/ (accessed 10.10.21).

Balagué, N., Torrents, C., Hristovski, R., Davids, K., Araujo, D., 2013. Overview of Complex Systems in Sport. J. Syst. Sci. Complex. 26, 4–13. https://doi.org/10.1007/s11424-013-2285-0

Barclay, P., 2014. The Life and Times of Herbert Chapman: The Story of One of Football's Most Influential Figures. Hachette UK.

BBC, 2020. De Bruyne named PFA Player of the Year. BBC Sport.

Bell Laboratories, R.W., 2017. R: What is R? [WWW Document]. URL https://www.r-project.org/about.html (accessed 10.10.17).

Bicego, M., Loog, M., 2016. Weighted K-Nearest Neighbor revisited. pp. 1642–1647. https://doi.org/10.1109/ICPR.2016.7899872

Bray, K., 2008. How to Score, Reprint edition. ed. Granta UK, London.

Buttle, F., 2019. Customer Relationship Management, 4ª edição. ed. Routledge, London ; New York.

Butz, H.E., Goodstein, L.D., 1996. Measuring customer value: Gaining the strategic advantage. Organ. Dyn. 24, 63–77. https://doi.org/10.1016/S0090-2616(96)90006-6

Cabannes, V., Bach, F., Rudi, A., 2021. Disambiguation of weak supervision with exponential convergence rates. ArXiv210202789 Cs Stat.

Cakmak, A., Uzun, A., Delibas, E., 2018. Computational Modeling of Pass Effectiveness in Soccer. Adv. Complex Syst. 21. https://doi.org/10.1142/S0219525918500108

Canty, A., Ripley, B., 2021. boot: Bootstrap Functions (Originally by Angelo Canty for S).

Canvas Generation, 2021. The Value Proposition Canvas – Canvas Generation. URL https://www.canvasgeneration.com/canvas/the-value-proposition-canvas/ (accessed 3.6.21).

Ch, M., ler, 2021. What Percentage of Penalties are Scored? [Statistical Breakdown]. SQaF. URL https://sqaf.club/what-percentage-of-penalties-are-scored-stats/ (accessed 9.21.21).
Chatterjee, S., Hadi, A.S., 2015. Regression Analysis by Example. John Wiley & Sons.
Chisari, F., 2006. When Football Went Global: Televising the 1966 World Cup. Hist. Soc. Res. Hist. Sozialforschung 31, 42–54.
Ciaburro, G., 2018. Regression Analysis with R: Design and develop statistical nodes to identify unique relationships within data at scale. Packt Publishing, Birmingham Mumbai.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans Inf Theory. https://doi.org/10.1109/TIT.1967.1053964
Cox, M., 2016. 6 football tactics that changed the game as we know it [WWW Document]. fourfourtwo.com. URL https://www.fourfourtwo.com/features/6-football-tactics-changed-game-we-know-it (accessed 2.20.21).
DataCamp, 2021. glm function - RDocumentation [WWW Document]. URL https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm (accessed 9.29.21).
Davies, J.C., 2013. Coaching the Tiki Taka Style of Play. SoccerTutor.com.
Delibas, E., Uzun, A., Inan, M., Guzey, O., Cakmak, A., 2019. Interactive Exploratory Soccer Data Analytics. INFOR Inf. Syst. Oper. Res. 57, 141–164. https://doi.org/10.1080/03155986.2018.1533204
Diwate, R., 2014. Data Mining Techniques in Association Rule : A Review.
Drucker, P.F., 2006. Innovation and Entrepreneurship, Reprint edição. ed. Harper Business, New York, NY.
Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, 2º edição. ed. Wiley-Interscience, New York.
Durán, A., 2020. How significant are penalties in European football? [WWW Document]. Driblab Footb. Powered Data. URL https://www.driblab.com/driblab-en/how-significant-are-penalties-in-european-football/ (accessed 10.5.21).
Dutton, T., 2019. Van Dijk missed out on Ballon d'Or to Messi "by just two votes" [WWW Document]. URL https://www.standard.co.uk/sport/football/virgil-van-dijk-missed-out-on-ballon-d-or-to-lionel-messi-by-just-two-votes-a4302841.html (accessed 10.12.21).
Dyer, K., Capo, R., Polikar, R., 2014. COMPOSE: A Semisupervised Learning Framework for Initially Labeled Nonstationary Streaming Data. Neural Netw. Learn. Syst. IEEE Trans. On 25, 12–26. https://doi.org/10.1109/TNNLS.2013.2277712
El Naqa, I., Murphy, M.J., 2015. What Is Machine Learning?, in: El Naqa, I., Li, R., Murphy, M.J. (Eds.), Machine Learning in Radiation Oncology: Theory and Applications. Springer International Publishing, Cham, pp. 3–11. https://doi.org/10.1007/978-3-319-18305-3_1
Espinasse, B., Pascot, D., 1987. DECISION SUPPORT SYSTEMS (DSS): A KNOWLEDGE ORIENTED APPROACH. pp. 105–108. https://doi.org/10.1016/B978-0-08-034350-1.50026-7
Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. AI Mag. 17, 37. https://doi.org/10.1609/aimag.v17i3.1230

FBRef.com, 2021. Porto vs. Juventus Match Report – Wednesday February 17, 2021 (Leg 1) [WWW Document]. FBref.com. URL https://fbref.com/en/matches/dea17ee3/Porto-Juventus-February-17-2021-Champions-League (accessed 10.8.21).
FBRef.com, 2020. xG Explained [WWW Document]. FBref.com. URL https://fbref.com/en/expected-goals-model-explained/ (accessed 9.15.21).
FC Bayern, 2014. The Golden Years [WWW Document]. FC Bayern Munich. URL https://fcbayern.com/us/news/2014/09/fc-bayern-munich-history---part-iv---the-golden-years (accessed 2.20.21).
Fifield, D., 2010. World Cup 2010: Franz Beckenbauer attacks England's "kick and rush" [WWW Document]. the Guardian. URL http://www.theguardian.com/football/2010/jun/15/world-cup-franz-beckenbauer-england (accessed 3.6.21).
Fix, E., Hodges, J.L., USAF School of Aviation Medicine, 1951. Discriminatory analysis: nonparametric discrimination, consistency properties. USAF School of Aviation Medicine, Randolph Field, Tex.
Footy Design, 2013. PSG NEW LOOK. Footy Des. URL https://footydesign.wordpress.com/2013/02/22/psg-new-look/ (accessed 2.26.21).
Fotache, M., Strimbei, C., 2015. SQL and Data Analysis. Some Implications for Data Analysits and Higher Education. Procedia Econ. Finance, Globalization and Higher Education in Economics and Business Administration - GEBA 2013 20, 243–251. https://doi.org/10.1016/S2212-5671(15)00071-4
FourFourTwo, 2020. The 100 greatest football managers of all time [WWW Document]. fourfourtwo.com. URL https://www.fourfourtwo.com/features/best-greatest-football-managers-ever-all-time (accessed 3.6.21).
Fukunaga, K., Hostetler, L., 1975. k-nearest-neighbor Bayes-risk estimation. IEEE Trans. Inf. Theory 21, 285–293. https://doi.org/10.1109/TIT.1975.1055373
Garnham, N., 2004. Association Football and Society in Pre-partition Ireland. Ulster Historical Foundation.
GoalPoint, 2017. GoalPoint | Serviços profissionais GoalPointPro | GoalPoint. URL https://goalpoint.pt/pro (accessed 2.19.21).
Govers, C.P.M., 1996. What and how about quality function deployment (QFD). Int. J. Prod. Econ., Proceedings of the 8th International Working Seminar on Production Economics 46–47, 575–585. https://doi.org/10.1016/0925-5273(95)00113-1
Grady, R.B., Caswell, D.L., 1987. Software metrics: establishing a company-wide program. Prentice-Hall, Inc., USA.
Guthrie, W.F., 2003. NIST/SEMATECH e-Handbook of Statistical Methods (NIST Handbook 151). https://doi.org/10.18434/M32189
Hajaj, Y., 2020. Introduction to Supervised, Semi-supervised, Unsupervised and Reinforcement Learning | Baeldung on Computer Science [WWW Document]. URL https://www.baeldung.com/cs/machine-learning-intro (accessed 2.28.21).
Hamilton, T., Karlsen, T.-K., 2020. Football Manager 2021 could find next Messi, but would club scouts use it? [WWW Document]. ESPN.com. URL https://www.espn.com/soccer/soccer-transfers/story/4242573/football-manager-2021-could-find-next-messibut-would-club-scouts-use-it (accessed 3.6.21).

Harding, J., 2021. Thomas Müller: Germany's resurgent "Raumdeuter" urges Harry Kane to be patient | DW | 26.06.2021 [WWW Document]. DW.COM. URL https://www.dw.com/en/thomas-m%C3%BCller-germanys-resurgent-raumdeuter-urges-harry-kane-to-be-patient/a-58056823 (accessed 10.11.21).

Hastie, T., Tibshirani, R., 1995. Discriminant adaptive nearest neighbor classification and regression, in: Proceedings of the 8th International Conference on Neural Information Processing Systems, NIPS'95. MIT Press, Cambridge, MA, USA, pp. 409–415.

Haugen, K., 2012. Always change a winning team, 2nd ed.

Hinton, G. (Ed.), 1999. Unsupervised Learning: Foundations of Neural Computation, 1st edição. ed. Bradford Books, Cambridge, Mass.

Hocquet, A., 2016. Football Manager: Mutual Shaping between Game, Sport, and Community. J. Media Stud. Pop. Cult. Rev. Détudes Médias Cult. Pop. 6.

Hu, J., Niu, H., Carrasco, J., Lennox, B., Arvin, F., 2020. Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning. IEEE Trans. Veh. Technol. 69, 14413–14423. https://doi.org/10.1109/TVT.2020.3034800

Hunter, J.S., 1986. The Exponentially Weighted Moving Average. J. Qual. Technol. 18, 203–210. https://doi.org/10.1080/00224065.1986.11979014

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Imandoust, S.B., Bolandraftar, M., 2013. Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. Int J Eng Res Appl 3, 605–610.

Ingle, S., 2015. How Midtjylland took the analytical route towards the Champions League [WWW Document]. the Guardian. URL http://www.theguardian.com/football/2015/jul/27/how-fc-midtjylland-analytical-route-champions-league-brentford-matthew-benham (accessed 2.18.21).

InStat, 2021a. InStat Sport: About Us | LinkedIn [WWW Document]. URL https://www.linkedin.com/company/instat-sport/about/ (accessed 2.27.21).

InStat, 2021b. INSTAT SCOUT - InStat [WWW Document]. URL https://instatsport.com/football/instat_scout (accessed 2.19.21).

InStat, 2021c. INSTAT INDEX - InStat [WWW Document]. URL https://instatsport.com/football/instat_index (accessed 2.27.21).

Ippolito, P.P., 2021. SVM: Feature Selection and Kernels [WWW Document]. Medium. URL https://towardsdatascience.com/svm-feature-selection-and-kernels-840781cc1a6c (accessed 10.3.21).

jabbar, M.A., Deekshatulu, B.L., Chandra, P., 2013. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technol., First International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013 10, 85–94. https://doi.org/10.1016/j.protcy.2013.12.340

Jones, R., 2021. Euro 2020 reaches cumulative global audience of 5.2bn. SportsPro. URL https://www.sportspromedia.com/news/euro-2020-tv-audience-ratings-viewership-social-media-figures/ (accessed 10.8.21).

Kalt, H., 2015. Soccer and Sponsorships. Soccer Polit. Polit. Footb. URL https://sites.duke.edu/wcwp/2015/03/03/soccer-and-sponsorships/ (accessed 2.19.21).

Kaneko, H., 2018. Illustration of merits of semi-supervised learning in regression analysis. Chemom. Intell. Lab. Syst. 182, 47–56. https://doi.org/10.1016/j.chemolab.2018.08.015

Keel, T., 2016. Johan Cruyff's best quotes: The game-changing wisdom of a true football legend [WWW Document]. Eurosport. URL https://www.eurosport.com/football/johan-cruyff-s-best-quotes-the-game-changing-wisdom-of-a-true-football-legend_sto5366190/story.shtml (accessed 10.15.21).

Keen, P.G.W., 1980. Decision support systems : a research perspective (Working Paper). Cambridge, Mass. : Center for Information Systems Research, Afred P. Sloan School of Management.

Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy k-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. 580–585.

Kerai, H., 2021. Ruben Dias: Manchester City defender wins Premier League Player of the Season award [WWW Document]. Sky Sports. URL https://www.skysports.com/football/news/11679/12325328/ruben-dias-manchester-city-defender-wins-premier-league-player-of-the-season-award (accessed 10.12.21).

Koen, P., Ajamian, G., Burkart, R., Clamen, A., Davidson, J., D'Amore, R., Elkins, C., Herald, K., Incorvia, M., Johnson, A., Karol, R., Seibert, R., Slavejkov, A., Wagner, K., 2001. Providing Clarity and A Common Language to the "Fuzzy Front End." Res.-Technol. Manag. 44, 46–55. https://doi.org/10.1080/08956308.2001.11671418

Kostopoulos, G., Karlos, S., Kotsiantis, S., Ragos, O., 2018. Semi-supervised regression: A recent review. J. Intell. Fuzzy Syst. 35, 1–18. https://doi.org/10.3233/JIFS-169689

Kostopoulos, G., Kotsiantis, S., Fazakis, N., Koutsonikos, G., Pierrakeas, C., 2019. A Semi-Supervised Regression Algorithm for Grade Prediction of Students in Distance Learning Courses. Int. J. Artif. Intell. Tools. https://doi.org/10.1142/S0218213019400013

Kramer, O., 2013. K-Nearest Neighbors, in: Kramer, O. (Ed.), Dimensionality Reduction with Unsupervised Nearest Neighbors, Intelligent Systems Reference Library. Springer, Berlin, Heidelberg, pp. 13–23. https://doi.org/10.1007/978-3-642-38652-7_2

Kuhn, M., 2019. The caret Package.

Kukhnavets, P., 2019. What is the Key Value of Quality Function Deployment in Product Development? | Hygger.io [WWW Document]. Hygger Proj. Manag. Softw. Tools Co. URL https://hygger.io/blog/quality-function-deployment-qfd-in-product-development/ (accessed 3.7.21).

LaBlanc, M.L., Henshaw, R., 1994. The World Encyclopedia of Soccer. Gale Research.

Lange, D., 2020. "Big Five" European football leagues revenue 1996-2021 [WWW Document]. Statista. URL https://www.statista.com/statistics/261218/big-five-european-soccer-leagues-revenue/ (accessed 2.19.21).

Lara, L., 2021. The nominees for the 2021 Ballon d'Or [WWW Document]. MARCA. URL https://www.marca.com/en/football/international-football/2021/10/08/61608d7c268e3e0c328b4641.html (accessed 10.12.21).

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324. https://doi.org/10.1109/5.726791

Légetr-Moec, A., 2011. Al-Khelaïfi détaille son plan pour le PSG et le mercato [WWW Document]. Foot Mercato Info Transf. Footb. - Actu Foot Transf. URL https://www.footmercato.net/a8884924937910148563-al-khelaifi-detaille-son-plan-pour-le-psg-et-le-mercato (accessed 2.26.21).

Lima, L.F. dos S., 2014. O jovem futebolista : uma proposta metodológica para o futebol de 11.

Luzum, N., Model, M., 2021. The Soccer Analytics Revolution. URL https://sites.duke.edu/socceranalyticsrevolution/ (accessed 3.6.21).

Mantoux, A., 2020. How Paris Saint-Germain became one of the most desirable brands in the world [WWW Document]. Lux. Trib. URL https://www.luxurytribune.com/en/how-paris-saint-germain-became-one-of-the-most-desirable-brands-in-the-world/ (accessed 2.26.21).

Marx, K., 1942. Karl Marx and Frederick Engels: Selected Correspondence, 1846-1895. International Publishers.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochem. Medica 22, 276–282.

Menon, P., 2018. An Executive Primer to Deep Learning [WWW Document]. URL https://www.datasciencecentral.com/profiles/blogs/an-executive-primer-to-deep-learning (accessed 2.18.21).

Merrick, P., 2010. Bleed White: The fall and rise of Leeds United... to be continued. AuthorHouse.

Microsoft, 2021. What is .NET? An open-source developer platform. [WWW Document]. Microsoft. URL https://dotnet.microsoft.com/learn/dotnet/what-is-dotnet (accessed 3.7.21).

Microsoft, 2020. Tutorial: Get Started with Entity Framework 6 Code First using MVC 5 [WWW Document]. URL https://docs.microsoft.com/en-us/aspnet/mvc/overview/getting-started/getting-started-with-ef-using-mvc/creating-an-entity-framework-data-model-for-an-asp-net-mvc-application (accessed 10.14.21).

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models.

Noble, J., 2016. Sponsors put their shirts on top-flight football [WWW Document]. URL https://www.ft.com/content/0817b946-2326-11e6-9d4d-c11776a5124d (accessed 2.26.21).

Oshinubi, K., Al-Awadhi, F., Rachdi, M., Demongeot, J., 2021. Data Analysis and Forecasting of COVID-19 Pandemic in Kuwait. https://doi.org/10.1101/2021.07.24.21261059

Osterwalder, A., Pigneur, Y., Bernarda, G., Smith, A., 2014. Value Proposition Design: How to Create Products and Services Customers Want. John Wiley & Sons.

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., Giannotti, F., 2019. PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Trans. Intell. Syst. Technol. 10, 59:1-59:27. https://doi.org/10.1145/3343172

Perry, M., 2010. The Exponentially Weighted Moving Average. https://doi.org/10.1002/9780470400531.eorms0314
Pines, L., 2020. Here's Why The Exponential Moving Average Will Give You Quick Results [WWW Document]. Commodity.com. URL https://commodity.com/technical-analysis/ma-exponential/ (accessed 10.13.21).
Pisani, S., 2020. Dortmund star Haaland wins 2020 Golden Boy award | Goal.com [WWW Document]. URL https://www.goal.com/en/news/dortmund-star-haaland-wins-2020-golden-boy-award/1j42qccuot1o51w1wadqwymq9v (accessed 10.10.21).
Power, D.J., 2002. Decision Support Systems: Concepts and Resources for Managers. Greenwood Publishing Group.
Pulabaigari, V., T, H.S., 2011. An Improvement to k-Nearest Neighbor Classifier. Presented at the 2011 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2011. https://doi.org/10.1109/RAICS.2011.6069307
Rampinini, E., Bishop, D.J., Marcora, S., Bravo, D., Sassi, R., Impellizzeri, F., 2007. Validity of Simple Field Tests as Indicators of Match-Related Physical Performance in Top-Level Professional Soccer Players. Int. J. Sports Med. 28, 228–35. https://doi.org/10.1055/s-2006-924340
Roberts, S.W., 1959. Control Chart Tests Based on Geometric Moving Averages. Technometrics 1, 239–250. https://doi.org/10.1080/00401706.1959.10489860
Russell, S., Norvig, P., 2009. Artificial Intelligence: A Modern Approach, 3ª edição. ed. Pearson, Upper Saddle River.
Saadeddin, D., 2021. Quality Function Deployment Template | Continuous Improvement Toolkit. URL https://citoolkit.com/templates/quality-function-deployment-template/ (accessed 3.3.21).
Schneider, A., Hommel, G., Blettner, M., 2010. Linear Regression Analysis. Dtsch. Ärztebl. Int. 107, 776–782. https://doi.org/10.3238/arztebl.2010.0776
Scrucca, L., Snow, G., Bloomfield, P., 2017. qcc: Quality Control Charts.
Shearer, C., 2000. The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. 5, 13–22.
Shetty, S., 2018. Total Football - A graphic history of the world's most iconic soccer tactics: The evolution of football formations and plays. Aurum.
Shin, T., 2021. All Machine Learning Algorithms You Should Know in 2021. KDnuggets. URL https://www.kdnuggets.com/all-machine-learning-algorithms-you-should-know-in-2021.html/ (accessed 3.2.21).
Shread, J., 2021. Jorginho wins UEFA men's player of the year, while Chelsea boss Thomas Tuchel named coach of the year [WWW Document]. Sky Sports. URL https://www.skysports.com/football/news/11095/12391488/jorginho-wins-uefa-mens-player-of-the-year-while-chelsea-boss-thomas-tuchel-named-coach-of-the-year (accessed 10.11.21).
Silva, Y., Almeida, I., Queiroz, M., 2016. SQL: From Traditional Databases to Big Data. pp. 413–418. https://doi.org/10.1145/2839509.2844560
Sleight, H., 2010. Graham Taylor: playing the long ball [WWW Document]. fourfourtwo.com. URL https://www.fourfourtwo.com/performance/tactics/graham-taylor-playing-long-ball (accessed 3.6.21).
Smith, A., 1817. An Inquiry Into the Nature and Causes of the Wealth of Nations. Oliphant, Waugh & Innes.

Soccerment Research, 2017. Crossing: An effective Strategy? | Stats & Analysis. Soccerment. URL https://soccerment.com/crossing-effective-strategy/ (accessed 10.12.21).
Spedding, J., 2019. Valeriy Lobanovskyi: The Scientist Who Dominated Football in the Soviet Union [WWW Document]. 90min.com. URL https://www.90min.com/posts/6413281-valeriy-lobanovskyi-the-scientist-who-dominated-football-in-the-soviet-union (accessed 3.6.21).
Sperandei, S., 2014. Understanding logistic regression analysis. Biochem. Medica 24, 12–8. https://doi.org/10.11613/BM.2014.003
Sports Interactive, 2006. Studio Timeline | 2006 | SI Games [WWW Document]. URL http://www.sigames.com/history (accessed 3.6.21).
Sports Interactive, SEGA, 2021. Football Manager 2021, Football Manager. Sports Interactive.
SportsReference, 2021. About Sports Reference [WWW Document]. Sports-Ref. URL https://www.sports-reference.com/about.html (accessed 9.14.21).
Sprague, R.H., Carlson, E., 1982. Building Effective Decision Support Systems. Prentice Hall, Englewood Cliffs, N.J.
Steinglass, M., 2019. Women's football is flourishing, on the pitch and off it. The Economist.
Stewart, M., Mitchell, H., Stavros, C., 2007. Moneyball Applied: Econometrics and the Identification and Recruitment of Elite Australian Footballers. Int. J. Sport Finance 2, 231–248.
Stuart, K., 2014. Why clubs are using Football Manager as a real-life scouting tool [WWW Document]. the Guardian. URL http://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool (accessed 3.6.21).
Suthaharan, S., 2015. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 1st ed. 2016 edition. ed. Springer, New York Heidelberg.
Syme, D., 2020. The early history of F#. Proc. ACM Program. Lang. 4, 75:1-75:58. https://doi.org/10.1145/3386325
Szabados G., 2003. Labdarúgóklubok stratégiái. Veztud. - Bp. Manag. Rev. 34, 32–43.
Szymanski, S., Kuypers, T., 1999. Winners and Losers. Viking.
Taylor, M., 2013. The Association Game: A History of British Football. Routledge.
Thai, T., Lam, H., 2003. .NET Framework Essentials. O'Reilly Media, Inc.
Tjønndal, A., 2017. Sport innovation: developing a typology. Eur. J. Sport Soc. 14, 291–310. https://doi.org/10.1080/16138171.2017.1421504
Truica, A., 2018. Note: *InStat Index is an index used by the professional scouting & analysis tool @InStatFootball to evaluate player's performances. InStat Index is automatically generated through a strict algorithm. @AlexTruica. URL https://twitter.com/AlexTruica/status/953658558964383744/photo/3 (accessed 2.27.21).
UEFA.com, 2021. Champions League Midfielder of the Season: N'Golo Kanté [WWW Document]. UEFA.com. URL https://www.uefa.com/uefachampionsleague/news/026c-1317baf0a7ae-4e35592f18ad-1000--champions-league-midfielder-of-the-season-n-golo-kante/ (accessed 10.11.21).
Warwick Manufacturing Group, 2007. Product Excellence using Six Sigma.

Weisberg, S., 2005. Applied linear regression. Hoboken, N.J. : Wiley-Interscience.
Wickham, H., Hester, J., Ooms, J., RStudio, example), R.F. (Copy of R. -project homepage cached as, 2020. xml2: Parse XML.
Wickham, H., RStudio, 2021. rvest: Easily Harvest (Scrape) Web Pages.
Wilson, J., 2018. Inverting The Pyramid: The History of Soccer Tactics. PublicAffairs.
Wilson, J., 2011. How Valeriy Lobanovskyi's appliance of science won hearts and trophies | Jonathan Wilson [WWW Document]. the Guardian. URL http://www.theguardian.com/football/blog/2011/may/12/valeriy-lobanovskyi-dynamo-kyiv (accessed 3.6.21).
Wyscout, 2018. Company. Wyscout. URL https://wyscout.com/company/ (accessed 2.23.21).
Zeithaml, V., 1988. Consumer Perceptions of Price, Quality and Value: A Means-End Model and Synthesis of Evidence. J. Mark. 52, 2–22. https://doi.org/10.1177/002224298805200302
Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms, 1st ed. Chapman & Hall/CRC.

Anexos

Anexo 1 — Ligação entre dados FBRef e tabelas de base de dados

Tabela no fbref.com	Tabela na Base de dados	Header na tabela do fbref.com	Nome da Estatística	Junção de header e nome	Nome da coluna na base de dados
summary	Player	none	Player	none.Player	player_name/player_id
summary	PlayerInTeamInSeason	none	Number	none.Number	player_number/player_id
summary	Player	none	Position	none.Position	player_position/player_id
summary	PlayerStatsFromMatch	none	Mins	none.Mins	minutes
summary	PlayerStatsFromMatch	Performance	Gls	Performance.Gls	goals
summary	PlayerStatsFromMatch	Performance	PK	Performance.PK	penalty_kicks_goals
summary	PlayerStatsFromMatch	Performance	PKatt	Performance.PKatt	penalty_kicks_attempted
summary	OffensiveStatsFromPlayerStatsFromMatch	Performance	Sh	Performance.Sh	shots_total
summary	OffensiveStatsFromPlayerStatsFromMatch	Performance	SoT	Performance.SoT	shots_on_target
summary	OffensiveStatsFromPlayerStatsFromMatch	Expected	xG	Expected.xG	expected_goals
summary	OffensiveStatsFromPlayerStatsFromMatch	Expected	npxG	Expected.npxG	non_penalty_expected_goals
summary	OffensiveStatsFromPlayerStatsFromMatch	SCA	SCA	SCA.SCA	shot_creating_actions
summary	OffensiveStatsFromPlayerStatsFromMatch	SCA	GCA	SCA.GCA	goal_creating_actions
passing	PassingStatsFromPlayerStatsFromMatch	Total	Cmp	Total.Cmp	passes_completed
passing	PassingStatsFromPlayerStatsFromMatch	Total	Att	Total.Att	passes_attempted
passing	PassingStatsFromPlayerStatsFromMatch	Total	Cmp%	Total.Cmp%	passes_completion
passing	PassingStatsFromPlayerStatsFromMatch	Total	TotDist	Total.TotDist	passes_total_distance
passing	PassingStatsFromPlayerStatsFromMatch	Total	PrgDist	Total.PrgDist	passes_progressive_distance
passing	PassingStatsFromPlayerStatsFromMatch	Short	Cmp	Short.Cmp	short_passes_completed

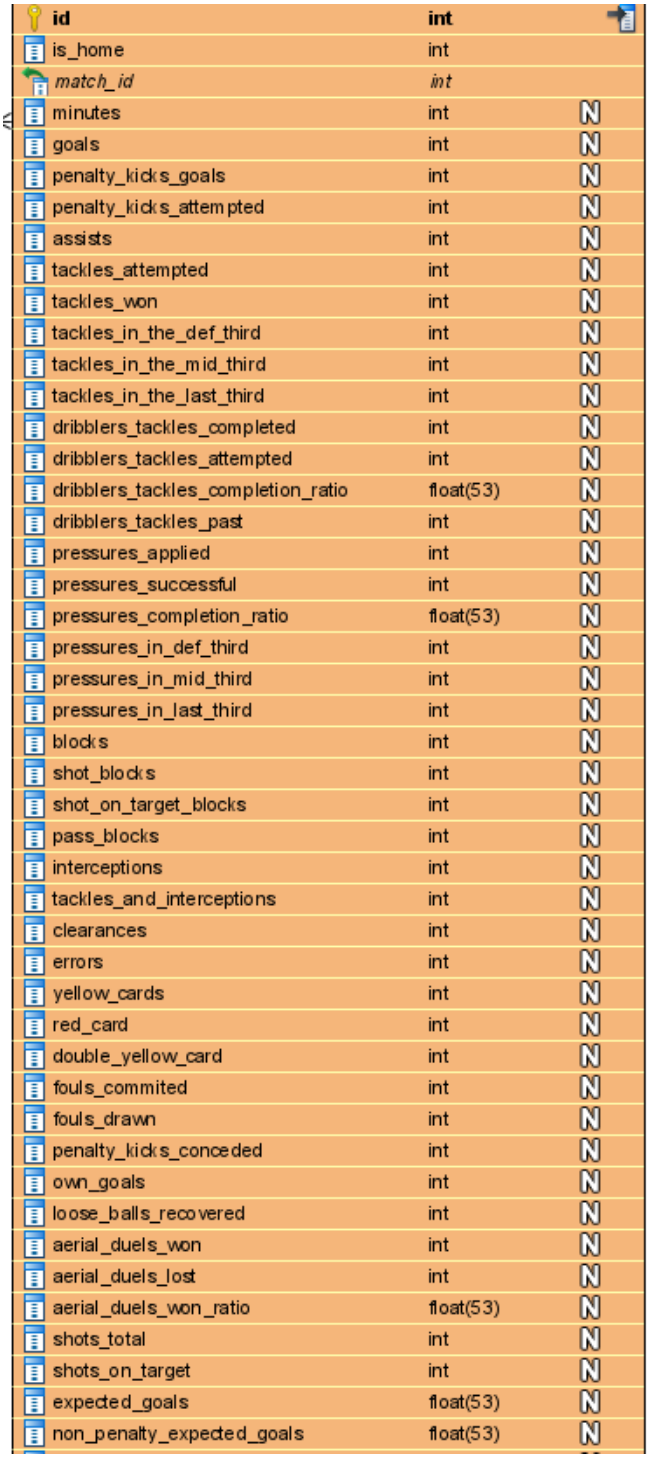
passing	PassingStatsFromPlayerStatsFromMatch	Short	Att	Short.Att	short_passes_attempted
passing	PassingStatsFromPlayerStatsFromMatch	Short	Cmp%	Short.Cmp%	short_passes_completion
passing	PassingStatsFromPlayerStatsFromMatch	Medium	Cmp	Medium.Cmp	medium_passes_completed
passing	PassingStatsFromPlayerStatsFromMatch	Medium	Att	Medium.Att	medium_passes_attempted
passing	PassingStatsFromPlayerStatsFromMatch	Medium	Cmp%	Medium.Cmp%	medium_passes_completion
passing	PassingStatsFromPlayerStatsFromMatch	Long	Cmp	Long.Cmp	long_passes_completed
passing	PassingStatsFromPlayerStatsFromMatch	Long	Att	Long.Att	long_passes_attempted
passing	PassingStatsFromPlayerStatsFromMatch	Long	Cmp%	Long.Cmp%	long_passes_completion
passing	PlayerStatsFromMatch	none	Ast	none.Ast	assists
passing	PassingStatsFromPlayerStatsFromMatch	none	xA	none.xA	expected_assists
passing	PassingStatsFromPlayerStatsFromMatch	none	KP	none.KP	key_passes
passing	PassingStatsFromPlayerStatsFromMatch	none	1/3	none.1/3	passes_completed_to_the_last_third
passing	PassingStatsFromPlayerStatsFromMatch	none	PPA	none.PPA	passes_completed_to_the_penalty_area
passing	PassingStatsFromPlayerStatsFromMatch	none	CrsPA	none.CrsPA	crosses_completed_to_the_penalty_area
passing	PassingStatsFromPlayerStatsFromMatch	none	Prog	none.Prog	progressive_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	Live	Pass Types.Live	live_ball_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	Dead	Pass Types.Dead	dead_ball_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	FK	Pass Types.FK	free_kick_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	TB	Pass Types.TB	through_ball_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	Press	Pass Types.Press	passes_while_in_pressure
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	Sw	Pass Types.Sw	sweeping_pass
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	Crs	Pass Types.Crs	crosses
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Pass Types	CK	Pass Types.CK	corner_kicks
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Corner Kicks	In	Corner Kicks.In	inswinging_corner
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Corner Kicks	Out	Corner Kicks.Out	outswinging_corner
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Corner Kicks	Str	Corner Kicks.Str	straight_corner
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Height	Ground	Height.Ground	grounded_passes

pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Height	Low	Height.Low	low_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Height	High	Height.High	high_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Body Parts	Left	Body Parts.Left	left_footed_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Body Parts	Right	Body Parts.Right	right_footed_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Body Parts	Head	Body Parts.Head	headed_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Body Parts	TI	Body Parts.TI	throw_ins
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Body Parts	Other	Body Parts.Other	other_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Outcomes	Off	Outcomes.Off	offside_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Outcomes	Out	Outcomes.Out	out_of_bounds_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Outcomes	Int	Outcomes.Int	intercepted_passes
pass_types	PassingTypeStatsFromPlayerStatsFromMatch	Outcomes	Blocks	Outcomes.Blocks	blocked_passes
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Tackles	Tkl	Tackles.Tkl	tackles_attempted
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Tackles	TklW	Tackles.TklW	tackles_won
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Tackles	Def 3rd	Tackles.Def 3rd	tackles_in_the_def_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Tackles	Mid 3rd	Tackles.Mid 3rd	tackles_in_the_mid_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Tackles	Att 3rd	Tackles.Att 3rd	tackles_in_the_last_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Vs Dribbles	Tkl	Vs Dribbles.Tkl	dribblers_tackles_completed
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Vs Dribbles	Att	Vs Dribbles.Att	dribblers_tackles_attempted
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Vs Dribbles	Tkl%	Vs Dribbles.Tkl%	dribblers_tackles_completion
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Vs Dribbles	Past	Vs Dribbles.Past	dribblers_tackles_past
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	Press	Pressures.Press	pressures_applied
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	Succ	Pressures.Succ	pressures_successful
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	%	Pressures.%	pressures_completed
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	Def 3rd	Pressures.Def 3rd	pressures_in_def_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	Mid 3rd	Pressures.Mid 3rd	pressures_in_mid_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Pressures	Att 3rd	Pressures.Att 3rd	pressures_in_last_third
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Blocks	Blocks	Blocks.Blocks	blocks

defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Blocks	Sh	Blocks.Sh	shot_blocks
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Blocks	ShSv	Blocks.ShSv	shot_on_target_blocks
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	Blocks	Pass	Blocks.Pass	pass_blocks
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	none	Int	none.Int	interceptions
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	none	Tkl+Int	none.Tkl+Int	tackles_and_interceptions
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	none	Clr	none.Clr	clearances
defensive_actions	DefensiveStatsFromPlayerStatsFromMatch	none	Err	none.Err	errors
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Touches	Touches.Touches	touches
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Def Pen	Touches.Def Pen	touches_in_defensive_penalty_area
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Def 3rd	Touches.Def 3rd	touches_in_def_third
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Mid 3rd	Touches.Mid 3rd	touches_in_mid_third
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Att 3rd	Touches.Att 3rd	touches_in_last_third
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Att Pen	Touches.Att Pen	touches_in_attacking_penalty_area
possesion	OffensiveStatsFromPlayerStatsFromMatch	Touches	Live	Touches.Live	live_touches
possesion	OffensiveStatsFromPlayerStatsFromMatch	Dribbles	Succ	Dribbles.Succ	dribbles_successful
possesion	OffensiveStatsFromPlayerStatsFromMatch	Dribbles	Att	Dribbles.Att	dribbles_attempted
possesion	OffensiveStatsFromPlayerStatsFromMatch	Dribbles	Succ%	Dribbles.Succ%	dribbles_success_ratio
possesion	OffensiveStatsFromPlayerStatsFromMatch	Dribbles	#Pl	Dribbles.#Pl	players_dribbled_past
possesion	OffensiveStatsFromPlayerStatsFromMatch	Dribbles	Megs	Dribbles.Megs	players_nutmegged
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	Carries	Carries.Carries	ball_carries
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	TotDist	Carries.TotDist	ball_carried_total_distance
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	PrgDist	Carries.PrgDist	ball_carried_progressive_distance
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	Prog	Carries.Prog	ball_carries_progressive
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	1/3	Carries.1/3	ball_carries_to_last_third
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	CPA	Carries.CPA	ball_carries_to_penalty_area
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	Mis	Carries.Mis	missed_ball_controls
possesion	OffensiveStatsFromPlayerStatsFromMatch	Carries	Dis	Carries.Dis	disarmed_carries

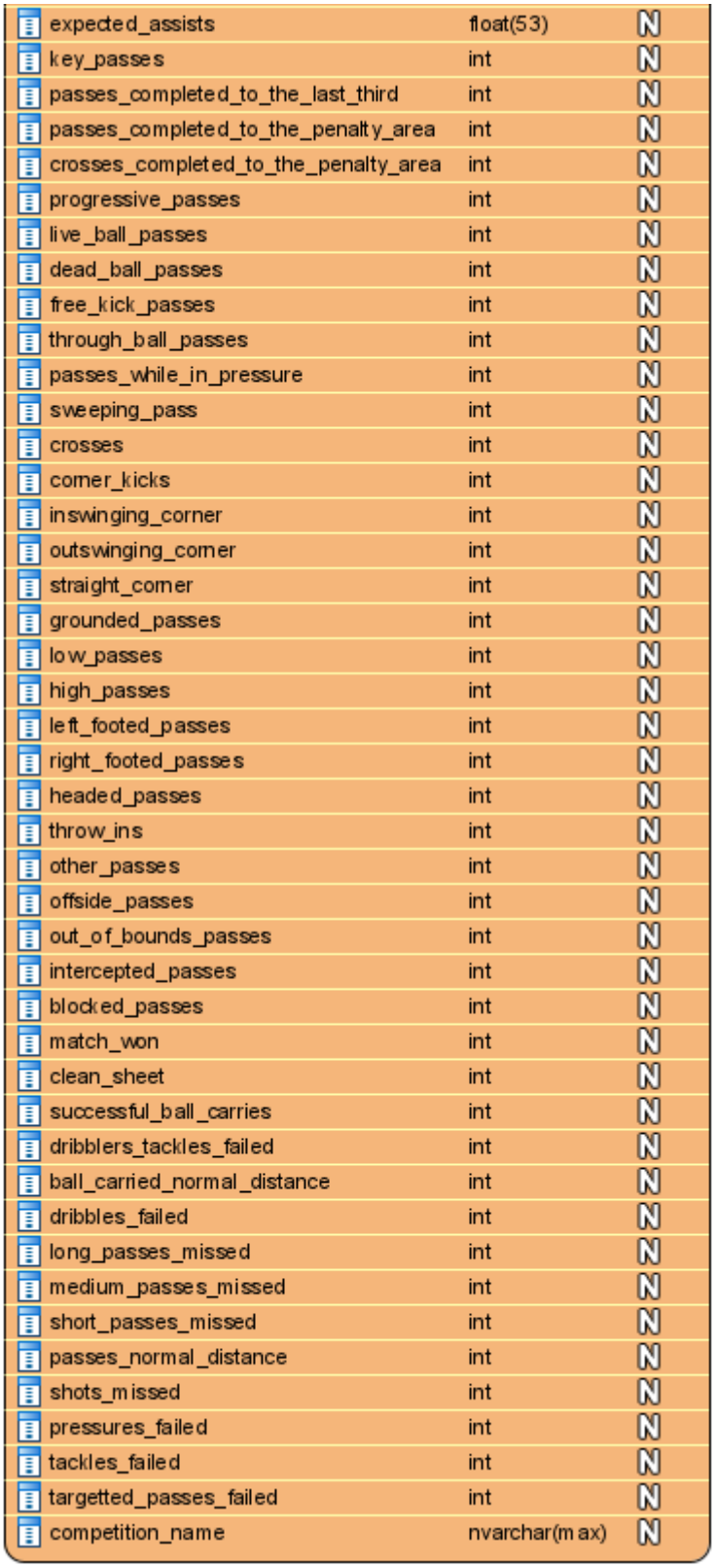
possesion	OffensiveStatsFromPlayerStatsFromMatch	Receiving	Targ	Receiving.Targ	targetted_passes
possesion	OffensiveStatsFromPlayerStatsFromMatch	Receiving	Rec	Receiving.Rec	received_passes
possesion	OffensiveStatsFromPlayerStatsFromMatch	Receiving	Rec%	Receiving.Rec%	received_passes_ratio
possesion	OffensiveStatsFromPlayerStatsFromMatch	Receiving	Prog	Receiving.Prog	progressive_received_passes
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	CrdY	Performance.CrdY	yellow_cards
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	CrdR	Performance.CrdR	red_card
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	2CrdY	Performance.2CrdY	double_yellow_card
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	Fls	Performance.Fls	fouls_commited
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	Fld	Performance.Fld	fouls_drawn
miscellaneous	OffensiveStatsFromPlayerStatsFromMatch	Performance	Off	Performance.Off	offsides
miscellaneous	OffensiveStatsFromPlayerStatsFromMatch	Performance	PKwon	Performance.PKwon	penalty_kicks_won
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	PKcon	Performance.PKcon	penalty_kicks_conceded
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	OG	Performance.OG	own_goals
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Performance	Recov	Performance.Recov	loose_balls_recovered
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Aerial Duels	Won	Aerial Duels.Won	aerial_duels_won
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Aerial Duels	Lost	Aerial Duels.Lost	aerial_duels_lost
miscellaneous	DefensiveStatsFromPlayerStatsFromMatch	Aerial Duels	Won%	Aerial Duels.Won%	aerial_duels_won_ratio

Anexo 2 – Tabela AggregatedMatch da Base de Dados



id	**int**	
is_home	int	
match_id	*int*	
minutes	int	N
goals	int	N
penalty_kicks_goals	int	N
penalty_kicks_attempted	int	N
assists	int	N
tackles_attempted	int	N
tackles_won	int	N
tackles_in_the_def_third	int	N
tackles_in_the_mid_third	int	N
tackles_in_the_last_third	int	N
dribblers_tackles_completed	int	N
dribblers_tackles_attempted	int	N
dribblers_tackles_completion_ratio	float(53)	N
dribblers_tackles_past	int	N
pressures_applied	int	N
pressures_successful	int	N
pressures_completion_ratio	float(53)	N
pressures_in_def_third	int	N
pressures_in_mid_third	int	N
pressures_in_last_third	int	N
blocks	int	N
shot_blocks	int	N
shot_on_target_blocks	int	N
pass_blocks	int	N
interceptions	int	N
tackles_and_interceptions	int	N
clearances	int	N
errors	int	N
yellow_cards	int	N
red_card	int	N
double_yellow_card	int	N
fouls_commited	int	N
fouls_drawn	int	N
penalty_kicks_conceded	int	N
own_goals	int	N
loose_balls_recovered	int	N
aerial_duels_won	int	N
aerial_duels_lost	int	N
aerial_duels_won_ratio	float(53)	N
shots_total	int	N
shots_on_target	int	N
expected_goals	float(53)	N
non_penalty_expected_goals	float(53)	N

shot_creating_actions	int	N
goal_creating_actions	int	N
touches	int	N
touches_in_defensive_penalty_area	int	N
touches_in_def_third	int	N
touches_in_mid_third	int	N
touches_in_last_third	int	N
touches_in_attacking_penalty_area	int	N
live_touches	int	N
dribbles_successful	int	N
dribbles_attempted	int	N
dribbles_success_ratio	float(53)	N
players_dribbled_past	int	N
players_nutmegged	int	N
ball_carries	int	N
ball_carried_total_distance	int	N
ball_carried_progressive_distance	int	N
ball_carries_progressive	int	N
ball_carries_to_last_third	int	N
ball_carries_to_penalty_area	int	N
missed_ball_controls	int	N
disarmed_carries	int	N
targetted_passes	int	N
received_passes	int	N
received_passes_ratio	float(53)	N
progressive_received_passes	int	N
offsides	int	N
penalty_kicks_won	int	N
passes_completed	int	N
passes_attempted	int	N
passes_completion_ratio	float(53)	N
passes_total_distance	int	N
passes_progressive_distance	int	N
short_passes_completed	int	N
short_passes_attempted	int	N
short_passes_completion_ratio	float(53)	N
medium_passes_completed	int	N
medium_passes_attempted	int	N
medium_passes_completion_ratio	float(53)	N
long_passes_completed	int	N
long_passes_attempted	int	N
long_passes_completion_ratio	float(53)	N



expected_assists	float(53)	N
key_passes	int	N
passes_completed_to_the_last_third	int	N
passes_completed_to_the_penalty_area	int	N
crosses_completed_to_the_penalty_area	int	N
progressive_passes	int	N
live_ball_passes	int	N
dead_ball_passes	int	N
free_kick_passes	int	N
through_ball_passes	int	N
passes_while_in_pressure	int	N
sweeping_pass	int	N
crosses	int	N
corner_kicks	int	N
inswinging_corner	int	N
outswinging_corner	int	N
straight_corner	int	N
grounded_passes	int	N
low_passes	int	N
high_passes	int	N
left_footed_passes	int	N
right_footed_passes	int	N
headed_passes	int	N
throw_ins	int	N
other_passes	int	N
offside_passes	int	N
out_of_bounds_passes	int	N
intercepted_passes	int	N
blocked_passes	int	N
match_won	int	N
clean_sheet	int	N
successful_ball_carries	int	N
dribblers_tackles_failed	int	N
ball_carried_normal_distance	int	N
dribbles_failed	int	N
long_passes_missed	int	N
medium_passes_missed	int	N
short_passes_missed	int	N
passes_normal_distance	int	N
shots_missed	int	N
pressures_failed	int	N
tackles_failed	int	N
targetted_passes_failed	int	N
competition_name	nvarchar(max)	N

Anexo 3 – Primeira correlações e razões

Estatística A	Estatística B	Coef.	Possível Razão
dribblers_tackles_past	dribblers_tackles_failed	1.000	Um jogador que é driblado é praticamente um lance em que falha o desarme.
touches	live_touches	0.998	Existem poucas situações em que um toque de bola não seja numa situação de jogo corrido (live_touches).
passes_completed	live_ball_passes	0.993	Quase sempre um passe completado corresponde a um passe em lance corrido.
live_touches	live_ball_passes	0.990	Um jogador de futebol não dá mais que dois ou três toques sem depois efetuar um passe.
touches	live_ball_passes	0.986	
live_touches	passes_completed	0.983	
dribbles_successful	players_dribbled_past	0.981	Se um jogador efetuou um drible com sucesso, intuitivamente se pressupõem que passou por um adversário.
touches	passes_completed	0.979	Um jogador de futebol não dá mais que dois ou três toques sem depois efetuar um passe.
passes_completed	grounded_passes	0.979	Grande parte dos passes completados são passes rasteiros.
passes_completed	successful_ball_carries	0.970	Excetuando passes de primeira, um passe completado corresponde praticamente em transportes de bola bem sucedidos.
passes_completed	passes_normal_distance	0.969	Quão maior for o número de passes, maior será a distância efetuada
live_ball_passes	grounded_passes	0.969	Como quase sempre um passe completado corresponde a um passe em lance corrido, a mesma razão da relação pode ser usada.
live_ball_passes	successful_ball_carries	0.966	Excetuando passes de primeira, um passe completado corresponde praticamente em transportes de bola bem sucedidos.

medium_passes_completed	passes_normal_distance	0.965	Quão maior for o número de passes, maior será a distância efetuada. Neste caso, o tipo de passe é um passe de meia distância.
live_ball_passes	passes_normal_distance	0.960	Quase sempre um passe completado corresponde a um passe em lance corrido, pelo que a correlação entre distância e número de passes se mantém.
grounded_passes	successful_ball_carries	0.959	Excetuando passes de primeira, um passe completado corresponde praticamente em transportes de bola bem sucedidos. O mesmo acontece com passes rasteiros
shot_creating_actions	key_passes	0.958	Um passe chave quase sempre resulta num remate, pelo que acaba por ser natural a sua forte correlação com a estatística referente a ações criadoras de remates.
passes_completed	medium_passes_completed	0.957	À medida que o número de passes aumenta, o número de passes de média distância aumenta.
live_touches	successful_ball_carries	0.955	Um transporte de bola bem sucedido terá sempre de ter associado um toque em lance corrido.
live_touches	grounded_passes	0.954	Um jogador de futebol não dá mais que dois ou três toques sem depois efetuar um passe, sendo que a correlação também afeta os passes rasteiros.
medium_passes_completed	live_ball_passes	0.951	Quase sempre um passe completado corresponde a um passe em lance corrido, no entanto a correlação desce quando a distância começa a aumentar.
touches	successful_ball_carries	0.950	Um transporte de bola bem sucedido terá sempre de ter associado um toque.

touches	grounded_passes	0.945	Um jogador de futebol não dá mais que dois ou três toques sem depois efetuar um passe, sendo que a correlação também funciona entre toques e passes rasteiros.
live_touches	passes_normal_distance	0.944	À medida que os toques de bola aumentam, os passes aumentam e, por sua vez, a distância também aumenta.
touches	touches_in_mid_third	0.944	A zona de campo com maior número de ações é o meio campo, pelo que é natural o número de toques no meio campo aumentar à medida que os toques no total aumentam.
grounded_passes	passes_normal_distance	0.943	Com o aumento dos passes rasteiros, a distância de passes também aumenta.
touches_in_mid_third	live_touches	0.943	A zona de campo com maior número de ações é o meio campo, pelo que é natural o número de toques no meio campo aumentar à medida que os toques em jogo corrido no total aumentam.
medium_passes_completed	grounded_passes	0.942	Grande parte dos passes completados são passes rasteiros, pelo que à medida que aumentam, os passes de média distância também aumentam.
touches_in_mid_third	live_ball_passes	0.941	Como a zona de campo com maior número de ações é o meio-campo, o número de passes em lance corrido tem uma grande correlação com esta estatística.
touches	passes_normal_distance	0.941	A distância de passes aumenta assim que o número de toques e transportes de bola aumentam.
successful_ball_carries	passes_normal_distance	0.938	
live_touches	medium_passes_completed	0.937	Continuação de uma grande correlação entre as estatísticas de passes, toques de bola e transportes.
touches_in_mid_third	passes_completed	0.937	
medium_passes_completed	successful_ball_carries	0.934	
touches	medium_passes_completed	0.933	

passes_completed	short_passes_completed	0.927	
progressive_received_passes	progressive_passes	0.919	Da mesma forma que os passes recebidos têm uma correlação direta com os passes completados, os passes progressivos tentados têm uma forte correlação com os passes progressivos recebidos.
ball_carried_progressive_distance	ball_carries_progressive	0.917	À medida que a distância de transportes progressivos aumenta, o número de transportes naturalmente aumenta.
short_passes_completed	live_ball_passes	0.917	Continuação de uma grande correlação entre as estatísticas de passes, toques de bola e transportes.
short_passes_completed	grounded_passes	0.915	
live_touches	short_passes_completed	0.913	
touches	passes_progressive_distance	0.910	
offsides	offside_passes	0.909	Um passe para fora-de-jogo naturalmente resulta em foras-de-jogo, pelo que é natural a forte correlação.
touches_in_mid_third	passes_normal_distance	0.909	Continuação de uma grande correlação entre as estatísticas de passes, toques de bola e transportes.
touches_in_mid_third	successful_ball_carries	0.909	
touches_in_mid_third	grounded_passes	0.907	
touches	short_passes_completed	0.907	
live_touches	passes_progressive_distance	0.906	Os toques na bola têm uma correlação forte com as estatísticas de passes, pelo que se mantém também com a distância de passes progressivos.
passes_progressive_distance	live_ball_passes	0.905	Quase sempre um passe completado corresponde a um passe em lance corrido, pelo que a distância de passes progressivos aumenta à medida que o número de passes em lance corrido aumenta.

Anexo 4 – Pesos retirados do Modelo Linear

Variável	Peso
aerial_duels_lost	0.0192
aerial_duels_won	0.0036
ball_carries_progressive	0.0252
ball_carries_to_last_third	0.0146
ball_carries_to_penalty_area	0.1115
blocked_passes	0.0430
clearances	0.0387
corner_kicks	0.0004
crosses	0.0611
crosses_completed_to_the_penalty_area	0.0138
disarmed_carries	0.0419
dribblers_tackles_completed	0.0246
dribblers_tackles_failed	0.0034
dribbles_failed	0.0326
dribbles_successful	0.0088
errors	0.3341
fouls_commited	0.0066
fouls_drawn	0.0703
free_kick_passes	0.0796
intercepted_passes	0.0172
interceptions	0.0183
key_passes	0.0472
long_passes_completed	0.0113
long_passes_missed	0.0506
loose_balls_recovered	0.0254
missed_ball_controls	0.0332
offsides	0.0127
out_of_bounds_passes	0.0486
pass_blocks	0.0203
passes_completed_to_the_last_third	0.0072
passes_completed_to_the_penalty_area	0.0014
passes_while_in_pressure	0.0003
penalty_kicks_conceded	0.7270
penalty_kicks_won	0.7185
pressures_failed	0.0000
pressures_in_def_third	0.0131
pressures_in_last_third	0.0068

pressures_in_mid_third	0.0012
pressures_successful	0.0192
progressive_passes	0.0140
red_card	0.6054
short_and_medium_passes_completed	0.0047
short_and_medium_passes_missed	0.0271
shot_blocks	0.0088
shot_creating_actions_without_key_passes	0.0279
shot_on_target_blocks	0.2029
shots_missed	0.1286
shots_on_target	0.3083
sweeping_pass	0.0001
tackles_failed	0.0000
tackles_in_the_def_third	0.0145
tackles_in_the_last_third	0.0637
tackles_in_the_mid_third	0.0301
tackles_won	0.0044
targetted_passes_failed	0.0020
through_ball_passes	0.1343
throw_ins	0.0121
touches_in_attacking_penalty_area	0.0303
touches_in_def_third	0.0026
touches_in_defensive_penalty_area	0.0011
touches_in_last_third	0.0001
yellow_cards	0.0494

Anexo 5 – Inquérito

Questionário de escolha de melhores jogadores

Este questionário foi redigido no âmbito da unidade curricular Tese/Dissertação/Estágio do Mestrado em Engenharia Informática, na área de Especialização em Sistemas de Informação e Conhecimento, do Instituto Superior de Engenharia do Porto, e destina-se a pessoas que estejam envolvidas no mundo do futebol, como jogadores, membros de equipa técnica, observadores ou estudantes de cursos relacionados diretamente com a área.

O questionário contém três secções referentes a três zonas de posição: defesas, médios e avançados. Cada secção terá seis perguntas a questionar qual o melhor jogador das opções disponibilizadas.

Os jogadores alvos do questionário fizeram pelo menos 30 jogos e 1350 minutos nas últimas quatro épocas (desde 2017/2018) nas cinco principais ligas europeias (Alemanha, Espanha, França, Inglaterra e Itália) e nas duas competições europeias de clubes (Liga dos Campeões e Liga Europa).

O inquirido deverá escolher qual o melhor jogador a partir da sua experiência como participante no desporto ou, no caso de achar os jogadores de qualidade semelhante, selecionar essa opção. Cada jogador terá um link para uma base de dados de registos futebolísticos, neste caso o ZeroZero, para auxiliar o processo de decisão.

Inicie sessão no Google para guardar o seu progresso. Saiba mais

*Obrigatório

Qual é/foi a sua ligação com o mundo do futebol? *

- ◯ Treinador
- ◯ Membro de equipa técnica
- ◯ Observador
- ◯ Diretor ou responsável administrativo
- ◯ Jogador federado
- ◯ Estudante em cursos de futebol (observação, treinador, etc.)

Seguinte Limpar formulário

Questionário de escolha de melhores jogadores

Inicie sessão no Google para guardar o seu progresso. Saiba mais

*Obrigatório

Defesas

Serão listados seis pares de jogadores considerados como defesas. Estão contemplados jogadores que atuem predominantemente a defesa central, defesa lateral e ala.

Qual é o melhor jogador entre David García (https://www.zerozero.pt/player.php?id=421648) e Scott Dann (https://www.zerozero.pt/player.php?id=50126)? *

- ◯ David Garcia
- ◯ Scott Dann
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Rodrigo Becão (https://www.zerozero.pt/player.php?id=491949) e Matthias Zimmermann (https://www.zerozero.pt/jogador.php?id=110235)? *

- ◯ Rodrigo Becão
- ◯ Matthias Zimmermann
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Julian Chabot (https://www.zerozero.pt/player.php?id=330705) e Renato Tapia (https://www.zerozero.pt/player.php?id=187910)? *

- ◯ Julian Chabot
- ◯ Renato Tapia
- ◯ Qualidade semelhante

Qual é o melhor jogador entre José Fonte (https://www.zerozero.pt/player.php?id=15537) e Roger Ibanez (https://www.zerozero.pt/player.php?id=554363)? *

- ◯ José Fonte
- ◯ Roger Ibanez
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Leander Dendoncker (https://www.zerozero.pt/player.php?id=149512) e Joris Gnagnon (https://www.zerozero.pt/player.php?id=485819)? *

- ◯ Leander Dendoncker
- ◯ Joris Gnagnon
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Christian Kabasele (https://www.zerozero.pt/player.php?id=70734) e Tin Jedvaj (https://www.zerozero.pt/player.php?id=316233)? *

- ◯ Christian Kabasele
- ◯ Tin Jedvaj
- ◯ Qualidade semelhante

Anterior Seguinte Limpar formulário

Questionário de escolha de melhores jogadores

Inicie sessão no Google para guardar o seu progresso. Saiba mais

*Obrigatório

Médios

Serão listados seis pares de jogadores considerados como médios. Estão contemplados jogadores que atuem predominantemente a médio defensivo, médio centro, médio ofensivo e médio ala.

Qual é o melhor jogador entre Nicolás Domínguez (https://www.zerozero.pt/player.php?id=557440) e Matteo Scozzarella (https://www.zerozero.pt/player.php?id=88591)? *

- ◯ Nicolás Domínguez
- ◯ Matteo Scozzarella
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Renato Sanches (https://www.zerozero.pt/player.php?id=155284) e Bruno Guimarães (https://www.zerozero.pt/player.php?id=463478)? *

- ◯ Renato Sanches
- ◯ Bruno Guimarães
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Axel Witsel (https://www.zerozero.pt/player.php?id=34169) e Antonín Barák (https://www.zerozero.pt/player.php?id=350163)? *

- ◯ Axel Witsel
- ◯ Antonín Barák
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Kerem Demirbay (https://www.zerozero.pt/player.php?id=237558) e Steven Davis (https://www.zerozero.pt/player.php?id=1216)? *

- ◯ Kerem Demirbay
- ◯ Steven Davis
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Koke (https://www.zerozero.pt/player.php?id=113787) e Jozabed (https://www.zerozero.pt/player.php?id=220248)? *

- ◯ Koke
- ◯ Jozabed
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Samuel Moutoussamy (https://www.zerozero.pt/player.php?id=236646) e Sebastián Cristóforo (https://www.zerozero.pt/player.php?id=206820)? *

- ◯ Samuel Moutoussamy
- ◯ Sebastián Cristóforo
- ◯ Qualidade semelhante

Anterior Seguinte Limpar formulário

Questionário de escolha de melhores jogadores

Inicie sessão no Google para guardar o seu progresso. Saiba mais

*Obrigatório

Avançados

Serão listados seis pares de jogadores considerados como avançados. Estão contemplados jogadores que atuem predominantemente a extremo, segundo avançado e ponta-de-lança.

Qual é o melhor jogador entre Raúl García (https://www.zerozero.pt/player.php?id=10882) e Dodi Lukebakio (https://www.zerozero.pt/player.php?id=386235)? *

- ◯ Raúl García
- ◯ Dodi Lukebakio
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Wahbi Khazri (https://www.zerozero.pt/player.php?id=232110) e Patrik Schick (https://www.zerozero.pt/player.php?id=406221)? *

- ◯ Wahbi Khazri
- ◯ Patrik Schick
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Marcus Rashford (https://www.zerozero.pt/player.php?id=434080) e Ángel Correa (https://www.zerozero.pt/player.php?id=332284)? *

- ◯ Marcus Rashford
- ◯ Ángel Correa
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Amine Gouiri (https://www.zerozero.pt/player.php?id=529101) e Edin Džeko (https://www.zerozero.pt/player.php?id=38299)? *

- ◯ Amine Gouiri
- ◯ Edin Džeko
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Shon Weissman (https://www.zerozero.pt/player.php?id=414409) e Angelo Fulgini (https://www.zerozero.pt/player.php?id=404530)? *

- ◯ Shon Weissman
- ◯ Angelo Fulgini
- ◯ Qualidade semelhante

Qual é o melhor jogador entre Lucas Pérez (https://www.zerozero.pt/player.php?id=166064) e Riccardo Meggiorini (https://www.zerozero.pt/player.php?id=11123)? *

- ◯ Lucas Pérez
- ◯ Riccardo Meggiorini
- ◯ Qualidade semelhante

Anterior | Submeter | Limpar formulário